stica

metodi

zione

e di apprendimento del corso è proposta all'interno di un ambiente
dio, che ha l'obiettivo di completare il libro offrendo risorse didattiche
utonomo o per assegnazione del docente.

monouso presente sulla copertina di questo libro consente l'accesso
ab, una piattaforma digitale interattiva specificamente pensata
e verificare i progressi durante lo studio.

ssibilità di accedere al manuale online: l'edizione digitale del testo
onalità che permettono di personalizzarne la fruizione, attivare
gitalizzata, inserire segnalibri anche su tablet e smartphone.

ve e valutative sono dettagliate nella pagina di catalogo dedicata
e all'indirizzo linkpearson.it/TBJQNL450S o tramite QR code.

Gli autori

Giuseppe Cicchitelli, è stato professore ordinario di statistica presso l'Università degli Studi di Perugia e direttore del Dipartimento di Scienze statistiche della stessa Università. È autore o coautore di diversi libri di statistica. È membro dell'International Statistical Institute (ISI). I suoi interessi di ricerca prevalenti riguardano il campionamento su popolazioni finite e l'analisi dei dati spaziali.

Pierpaolo D'Urso è professore ordinario di Statistica presso la Sapienza – Università di Roma, dove è direttore del dipartimento di Scienze sociali ed economiche e prorettore per la formazione del personale. È titolare degli insegnamenti di Statistica e Data science e statistica multivariata. È componente del Nucleo di Valutazione dell'Università Luiss. È associate editor di diverse riviste internazionali e autore di numerosi articoli concernenti principalmente la statistica multivariata. È classificato nella World's Top 2% Scientists della Stanford University.

Marco Minozzo è professore associato di Statistica presso l'Università di Verona, dove è presidente del corso di laurea in Economia e Commercio e membro del Presidio della Qualità per la Didattica. È stato titolare di insegnamenti di probabilità e statistica in corsi di laurea, master e dottorato, in Italia e all'estero. È autore di numerosi lavori scientifici presentati e pubblicati a livello internazionale.

attaforma **MyLab**,
lente universitario
.it/catalogo
tudente monouso
tina del libro.
edere l'accesso
Lab al consulente
:
ntatti/
itari.html

9788891915382A

Pearson

Statistica
Principi e metodi
Quarta edizione

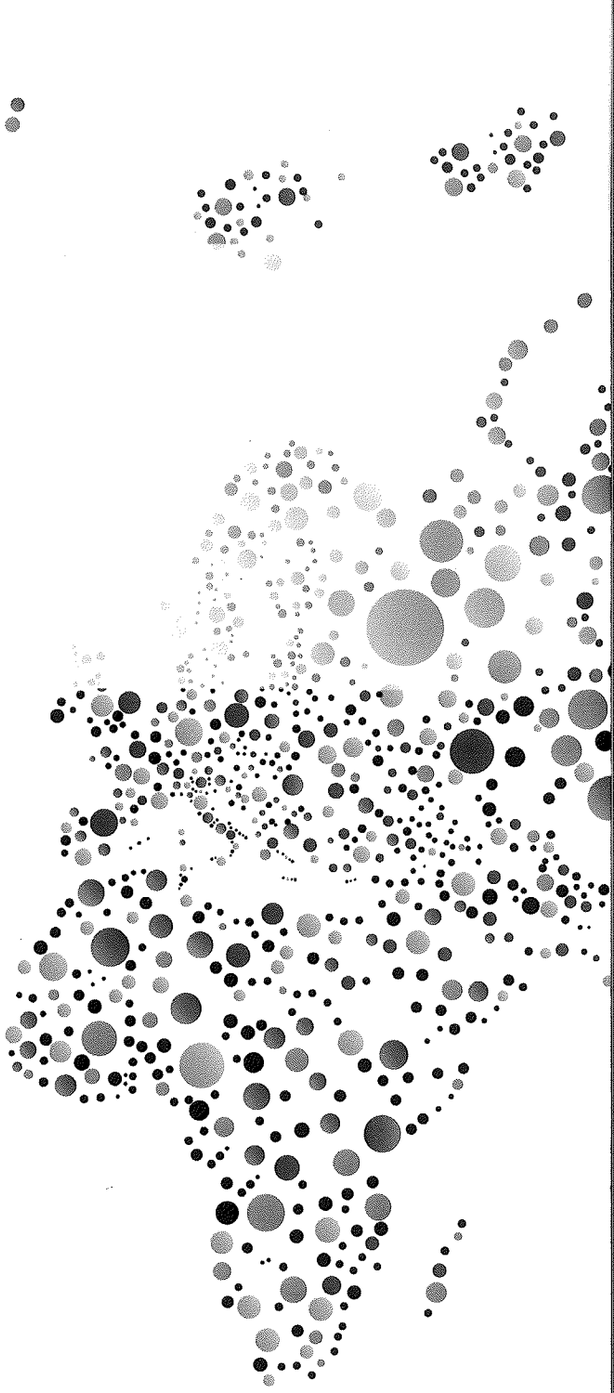
Giuseppe Cicchitelli
Pierpaolo D'Urso
Marco Minozzo

Statistica

Principi e metodi

Quarta edizione

Giuseppe Cicchitelli
Pierpaolo D'Urso
Marco Minozzo



Pearson

MyLab Codice studente

Statistica

Principi e metodi

Quarta edizione

Giuseppe Cicchitelli
Pierpaolo D'Urso
Marco Minozzo

Pearson

Pearson

© 2022 Pearson Italia, Milano-Torino

Il presente testo è di proprietà di Pearson Italia la quale non è associata, né direttamente né indirettamente, a eventuali marchi di terzi che venissero richiamati per gli scopi illustrativi ed educativi che ha la pubblicazione.
Per i passi antologici, per le citazioni, per le riproduzioni grafiche, cartografiche e fotografiche appartenenti alla proprietà di terzi, inseriti in quest'opera, l'editore è a disposizione degli aventi diritto non potuti reperire nonché per eventuali non volute omissioni e/o errori di attribuzione nei riferimenti.
È vietata la riproduzione, anche parziale o ad uso interno didattico, con qualsiasi mezzo, non autorizzata.
Le fotocopie per uso personale del lettore possono essere effettuate nei limiti del 15% di ciascun volume dietro pagamento alla SIAE del compenso previsto dall'art. 68, commi 4 e 5, della legge 22 aprile 1941, n. 633.
Le riproduzioni effettuate per finalità di carattere professionale, economico o commerciale o comunque per uso diverso da quello personale possono essere effettuate a seguito di specifica autorizzazione rilasciata da CLEARedi, Corso di Porta Romana 108, 20122 Milano, e-mail autorizzazioni@clearedi.org e sito web www.clearedi.org.
Pearson non si assume alcuna responsabilità per i Materiali pubblicati da terze parti sui propri siti Web e/o piattaforme o accessibili, tramite collegamenti ipertestuali o altri "collegamenti" digitali, a siti ospitati da terze parti non controllati direttamente da Pearson ("sito di terze parti"). Per approfondimenti si invita a consultare il sito pearson.it.

I nostri libri sono ecosostenibili: la carta è prodotta sostenendo il ciclo naturale e per ogni albero tagliato ne viene piantato un altro; il cellofan è realizzato con plastiche da recupero ambientale o riciclate; gli inchiostri sono naturali e atossici; i libri sono prodotti in Italia e l'impatto del trasporto è ridotto al minimo.

Realizzazione editoriale: Carmelo Giarratana
Grafica di copertina: Simone Tartaglia
Immagine di copertina: RoseGarden/Shutterstock
Stampa: Rotolito S.p.A. - Pioltello (MI)

Microsoft Excel © Microsoft. Used with permission from Microsoft
Tutti i marchi citati nel testo sono di proprietà dei loro detentori.

9788891915382

Printed in Italy

1ª edizione: settembre 2022

Ristampa
02 03 04 05 06

Anno
24 25 26 27 28

LIBRI DI TESTO E SUPPORTI DIDATTICI
Il sistema di gestione per la qualità della Casa Editrice è certificato in conformità alla norma **UNI EN ISO 9001:2015** per l'attività di **progettazione, realizzazione e commercializzazione di: • prodotti editoriali scolastici, dizionari lessicografici, prodotti per l'editoria di varia ed università • materiali didattici multimediali off-line • corsi di formazione e specializzazione in aula, a distanza, e-learning.**
Member of CISQ Federation
CERTIFIED MANAGEMENT SYSTEM
ISO 9001

a Nicola
a Giordana
a Ivana e Agostino

Sommario

Prefazione xiii
Pearson MyLab xv

Capitolo 1 Nozioni introduttive **1**

1.1 Introduzione 1
1.2 Cenni storici sullo sviluppo della statistica 2
1.3 La statistica nel processo di acquisizione di nuova conoscenza . . 5
1.4 La statistica nelle attività operative e nella vita quotidiana 6
1.5 Fonti di dati statistici nazionali e internazionali 7
1.5.1 L'Istat e il Sistema statistico nazionale 8
1.5.2 L'Eurostat e il Sistema statistico europeo 9
1.5.3 Altre fonti nazionali e internazionali 10
1.6 Terminologia statistica essenziale 11
1.7 Scale di misurazione e tipologie di variabili 13
1.8 Genesi dei dati statistici 15
1.8.1 Indagini statistiche 16
1.8.2 Esperimenti 21
1.8.3 Studi osservazionali o sul campo 23
1.9 Matrice dei dati 24
1.10 Statistica descrittiva e inferenza statistica 25
1.11 Alcuni calcoli statistici elementari: differenze relative e rapporti statistici 26
Riassunto 28
Commenti e curiosità 29
Esercizi 32
Appendice 35

PARTE I STATISTICA DESCRITTIVA

Capitolo 2 Serie di osservazioni e distribuzioni di frequenza **41**

2.1 Introduzione 41
2.2 Dalle serie di osservazioni alle distribuzioni di frequenza 43
2.2.1 Distribuzioni di frequenza in classi 46
2.3 Distribuzioni doppie di frequenza 53
2.4 Serie storiche e serie territoriali 55
2.5 Sulle imperfezioni nei dati 56
Riassunto 59
Commenti e curiosità 59
Esercizi 61

Capitolo 3 Rappresentazioni grafiche 67

3.1 Introduzione 67
3.2 Distribuzioni semplici di frequenza 67
3.2.1 Distribuzioni di frequenza di variabili discrete 67
3.2.2 Distribuzioni di frequenza in classi 71
3.2.3 Distribuzioni di frequenza di variabili qualitative 75
3.3 Distribuzioni doppie di frequenza 78
3.4 Serie storiche 79
3.5 Serie territoriali 82
3.6 Problemi di scala 83
Riassunto 84
Commenti e curiosità 84
Esercizi 89

Capitolo 4 Medie 93

4.1 Introduzione 93
4.2 Media aritmetica 93
4.3 Media armonica 96
4.4 Media geometrica 97
4.5 Media quadratica 98
4.6 Medie di potenza 99
4.7 Medie analitiche delle distribuzioni di frequenza 100
4.7.1 Il caso delle distribuzioni di frequenza in classi 102
4.8 Medie ponderate 105
4.9 Mediana 107
4.10 Quartili e quantili 109
4.11 Mediana, quartili e quantili delle distribuzioni di frequenza 111
4.11.1 Il caso delle distribuzioni di frequenza in classi 113
4.12 Valore centrale 115
4.13 Moda 115
4.14 Medie e dati anomali 116
4.15 La scelta della media 117
Riassunto 118
Commenti e curiosità 119
Esercizi 122

Capitolo 5 Variabilità 127

5.1 Introduzione 127
5.2 Scostamenti medi 128
5.2.1 Il caso delle distribuzioni di frequenza 131
5.2.2 Il caso delle distribuzioni di frequenza in classi 132
5.3 Differenze medie 134
5.3.1 Il caso delle distribuzioni di frequenza 136
5.4 Campo di variazione e differenza interquartile 136
5.5 Indici percentuali di variabilità 137
5.6 Indici di disuguaglianza o di concentrazione 139
5.6.1 Rapporto di concentrazione di Gini 140
5.6.2 Rapporto di concentrazione delle distribuzioni di frequenza 144
5.6.3 Il caso delle distribuzioni di frequenza in classi 145

5.7 Eterogeneità 146
5.8 Quale indice impiegare 148
Riassunto 148
Commenti e curiosità 149
Esercizi 152

Capitolo 6 Forma delle distribuzioni di frequenza 157

6.1 Introduzione 157
6.2 Asimmetria 157
6.3 Curtosi 163
Riassunto 166
Commenti e curiosità 166
Esercizi 167

Capitolo 7 Uno sguardo d'insieme alle costanti caratteristiche 169

7.1 Introduzione 169
7.2 Grafici e costanti caratteristiche 170
7.2.1 Grafico a scatola 171
7.3 Disuguaglianza di Chebyshev 172
Riassunto 174
Commenti e curiosità 174
Esercizi 175

Capitolo 8 Numeri indici e indicatori compositi 177

8.1 Introduzione 177
8.2 Serie storiche di numeri indici 178
8.3 Numeri indici sintetici 183
8.4 Indicatori compositi 187
Riassunto 191
Commenti e curiosità 191
Esercizi 192

Capitolo 9 Analisi della dipendenza nelle tabelle di contingenza 195

9.1 Introduzione 195
9.2 Dipendenza tra le variabili di una tabella di contingenza 198
9.2.1 Misura della dipendenza 201
9.3 Dipendenza in media 204
9.3.1 Misura della dipendenza in media 205
Riassunto 207
Parole chiave 207
Esercizi 209

Capitolo 10 Regressione lineare semplice 213

10.1 Introduzione 213
10.2 Regressione lineare semplice 215
10.3 Adattamento della retta di regressione ai punti osservati 220
Riassunto 223

Commenti e curiosità 224
Esercizi 227

Capitolo 11 Correlazione **233**

11.1 Introduzione 233
11.2 Misura della correlazione 235
11.2.1 Correlazione tra graduatorie 238
Riassunto 239
Commenti e curiosità 239
Esercizi 241

PARTE II FONDAMENTI DI PROBABILITÀ

Capitolo 12 Probabilità **245**

12.1 Introduzione 245
12.2 Esperimenti casuali, spazio campionario ed eventi 246
12.3 Probabilità 251
12.3.1 Interpretazione della probabilità 252
12.4 Assegnazione delle probabilità agli eventi 254
12.5 Probabilità condizionata 257
12.6 Indipendenza tra eventi 259
12.7 Formula di Bayes 262
Riassunto 265
Commenti e curiosità 265
Esercizi 268

Capitolo 13 Variabili casuali **273**

13.1 Introduzione 273
13.2 Variabili casuali discrete 274
13.2.1 Media e varianza 278
13.3 Variabili casuali continue 280
13.3.1 Media e varianza 284
13.3.2 Quantili 285
13.4 Ulteriori proprietà delle variabili casuali 286
Riassunto 289
Commenti e curiosità 289
Esercizi 292

Capitolo 14 Alcune distribuzioni di probabilità parametriche **295**

14.1 Introduzione 295
14.2 Distribuzione uniforme discreta 295
14.3 Distribuzione di Bernoulli 296
14.4 Distribuzione binomiale 297
14.5 Distribuzione di Poisson 301
14.6 Distribuzione uniforme continua 304
14.7 Distribuzione esponenziale 305
14.8 Distribuzione normale 306

14.8.1 Distribuzione normale standard 309
14.8.2 Probabilità e quantili tramite la tavola della distribuzione normale standard 311
14.9 Distribuzione chi-quadrato 314
Riassunto 316
Commenti e curiosità 316
Esercizi 319

Capitolo 15 Distribuzioni di probabilità congiunte **325**

15.1 Introduzione 325
15.2 Variabili casuali doppie 325
15.3 Distribuzione di probabilità congiunta di due variabili casuali discrete 327
15.3.1 Distribuzioni di probabilità marginali 328
15.3.2 Distribuzioni di probabilità condizionate 329
15.3.3 Indipendenza tra due variabili casuali 331
15.3.4 Covarianza e correlazione 333
15.4 Distribuzione di probabilità congiunta di due variabili casuali continue 334
15.5 Distribuzione di probabilità congiunta di n variabili casuali 336
15.5.1 Combinazioni lineari di n variabili casuali 337
Riassunto 339
Commenti e curiosità 339
Esercizi 340

Capitolo 16 Legge dei grandi numeri e teorema del limite centrale **345**

16.1 Introduzione 345
16.2 Legge dei grandi numeri 346
16.3 Teorema del limite centrale 348
16.3.1 Approssimazione della distribuzione binomiale con la normale 351
Riassunto 353
Commenti e curiosità 353
Esercizi 354

Capitolo 17 Campioni casuali e distribuzioni campionarie **357**

17.1 Introduzione 357
17.2 Campione casuale 357
17.3 Distribuzione di probabilità del campione casuale 360
17.4 Statistiche e distribuzioni campionarie 362
17.5 Distribuzione campionaria della media 365
17.5.1 Popolazione normale 366
17.5.2 Grandi campioni 367
17.6 Distribuzione campionaria della varianza 368
17.6.1 Popolazione normale 369
17.7 Distribuzione campionaria della media quando la varianza σ^2 non è nota 370
17.7.1 Popolazione normale 370
17.7.2 Grandi campioni 372

17.8 Stima dei parametri e verifica delle ipotesi 373
Riassunto 374
Commenti e curiosità 375
Esercizi 376

PARTE III INFERENZA STATISTICA

Capitolo 18 Stima puntuale **383**

18.1 Introduzione 383
18.2 Proprietà degli stimatori 384
18.2.1 Non distorsione 384
18.2.2 Efficienza 387
18.2.3 Proprietà asintotiche 392
18.3 Ricerca dello stimatore ottimo 394
18.3.1 Metodo della massima verosimiglianza 395
18.4 Margine di errore nella stima della media 396
18.4.1 Ampiezza del campione e margine di errore 398
Riassunto 399
Commenti e curiosità 399
Esercizi 402

Capitolo 19 Stima intervallare **405**

19.1 Introduzione 405
19.2 Stima intervallare della media di una popolazione normale con varianza nota 406
19.3 Stima intervallare della media di una popolazione normale con varianza non nota 411
19.4 Stima intervallare della media nel caso di grandi campioni 413
19.5 Stima intervallare della varianza di una popolazione normale . . . 416
Riassunto 418
Commenti e curiosità 418
Esercizi 419

Capitolo 20 Verifica delle ipotesi **421**

20.1 Introduzione 421
20.2 Verifica delle ipotesi 422
20.3 Test di ipotesi sulla media di una popolazione normale con varianza nota 429
20.3.1 Livello di significatività osservato 433
20.4 Test di ipotesi sulla media di una popolazione normale con varianza non nota 435
20.4.1 Livello di significatività osservato 437
20.5 Test di ipotesi sulla media di una popolazione nel caso di grandi campioni 438
20.5.1 Popolazioni Bernoulliane 439
20.6 Test di ipotesi sulla varianza di una popolazione normale 440
20.6.1 Livello di significatività osservato 443
20.7 Criteri di ottimizzazione nella verifica delle ipotesi 443

Riassunto 449
Commenti e curiosità 450
Esercizi 455

Capitolo 21 Confronti tra due popolazioni **459**

21.1 Introduzione 459
21.2 Inferenza sulle medie di due popolazioni normali 459
21.2.1 Varianze non note e piccoli campioni 463
21.3 Inferenza sulle medie di due popolazioni nel caso di grandi campioni 466
21.3.1 Popolazioni Bernoulliane 468
21.4 Verifica dell'ipotesi di uguaglianza delle varianze di due popolazioni normali 471
Riassunto 473
Commenti e curiosità 473
Esercizi 474

Capitolo 22 Verifica di ipotesi per dati di frequenza **479**

22.1 Introduzione 479
22.2 La statistica test chi-quadrato 481
22.2.1 Verifica dell'ipotesi nulla di indipendenza 484
22.2.2 Verifica dell'ipotesi di omogeneità 486
Riassunto 488
Commenti e curiosità 489
Esercizi 490

Capitolo 23 Inferenza sul modello di regressione lineare **493**

23.1 Introduzione 493
23.2 Stima di β_0 e β_1 495
23.3 Stima della varianza del termine di errore 499
23.4 Stima intervallare e verifica di ipotesi su β_0 e β_1 500
23.5 Stima intervallare del valore atteso di Y 504
23.6 Intervallo predittivo per Y 506
Riassunto 507
Commenti e curiosità 507
Esercizi 509

Capitolo 24 Inferenza su popolazioni finite **513**

24.1 Introduzione 513
24.2 Piani di campionamento, parametri e stimatori 514
24.3 Campionamento casuale semplice 518
24.3.1 Popolazioni Bernoulliane 521
24.4 Campionamento stratificato 523
24.4.1 Allocazione delle unità negli strati 525
24.5 Campionamento a due stadi semplici 527
24.6 Considerazioni finali 529
Riassunto 531
Commenti e curiosità 532
Esercizi 534

Capitolo 25 Regressione lineare multipla online

Capitolo 26 Analisi della varianza online

Appendice A Statistica descrittiva: dimostrazioni 539

A.1 L'operatore di sommatoria 539
A.2 Medie 540
A.2.1 Media aritmetica 540
A.2.2 Mediana 542
A.3 Concentrazione 542
A.4 Forma delle distribuzioni di frequenza 545
A.5 Disuguaglianza di Chebyshev 546
A.6 Analisi della dipendenza nelle tabelle di contingenza 547
A.7 Regressione lineare semplice 548
A.8 Correlazione 551

Appendice B Probabilità e inferenza: dimostrazioni 553

B.1 Famiglie di eventi e misure di probabilità 553
B.2 Regola della somma 554
B.3 Elementi di calcolo combinatorio 554
B.4 Funzione di ripartizione e funzione di densità 558
B.5 Valore atteso di trasformazioni di variabili casuali 560
B.6 Disuguaglianza di Chebyshev 562
B.7 Disuguaglianza di Markov 563
B.8 Distribuzioni di probabilità 563
B.9 Valore atteso di combinazioni lineari di funzioni di due o più variabili casuali 566
B.10 Argomenti di stima puntuale 568
B.11 Inferenza su popolazioni finite 571

Appendice C Tavole statistiche 577

Bibliografia 587

Indice analitico 589

Soluzione degli esercizi 593

Prefazione

Statistica. Principi e metodi è stato progettato per gli insegnamenti di statistica presenti nei corsi di studio di: statistica, economia, scienze politiche e scienze sociali. Ma può essere un utile strumento di studio anche per studenti di statistica applicata nei corsi di laurea di matematica, di scienze e di ingegneria.

I prerequisiti necessari per la lettura del testo non vanno oltre le conoscenze di algebra fornite dalla generalità delle scuole superiori. Gli strumenti di analisi matematica sono utilizzati raramente. Quando non sono indispensabili, sono menzionati in nota o contenuti nelle appendici.

Gli aspetti concettuali hanno la priorità rispetto alle procedure, alle dimostrazioni e ai calcoli. Gli argomenti vengono introdotti in modo discorsivo e con molti esempi. Tuttavia, il perseguimento della chiarezza non va a scapito del rigore: i concetti di base e le proprietà delle metodologie via via proposte vengono delineati in definizioni e proposizioni con accorgimenti grafici mirati a rendere agevole lo studio. Le dimostrazioni di queste ultime vengono date immediatamente, a meno che non richiedano calcoli di qualche complessità, nel qual caso sono rinviate alle appendici del volume. Ad ogni modo, la struttura del testo consente di saltare le dimostrazioni senza perdere il filo del ragionamento.

La statistica descrittiva occupa uno spazio più ampio rispetto ad altri testi universitari di statistica di base. Questa scelta risponde a due esigenze: quella di recuperare una base comune di conoscenze di statistica elementare in studenti provenienti da scuole superiori eterogenee, dove il programma di matematica non include talvolta che qualche cenno di statistica, e quella di offrire un insieme coordinato di strumenti di analisi dei dati, importanti in sé, ma utili anche per lo studio della probabilità e dell'inferenza statistica, in considerazione dei parallelismi concettuali che esistono tra i due comparti.

Anche se l'enfasi non è sui calcoli, alla fine dei capitoli, sono previsti spazi per l'illustrazione delle principali funzioni Excel per il calcolo statistico. La scelta di questo foglio elettronico è legata soprattutto alla sua immediata e generale disponibilità.

Principali innovazioni

In questa quarta edizione di *Statistica. Principi e metodi* rimane inalterato l'impianto complessivo, in considerazione dell'ottimo riscontro che il testo ha trovato in ambito universitario. Di seguito, sono indicate le principali innovazioni.

- Nel *capitolo introduttivo*, è stato dato ulteriore risalto alle indagini condotte dall'Istat.
- Nel *capitolo sulle rappresentazioni grafiche* è stata estesa la parte aneddotica sui precursori delle moderne tecniche di visualizzazione dei dati.
- Il *capitolo sui numeri indici* è stato arricchito di un paragrafo sui cosiddetti indicatori compositi, strumenti assai utili per il confronto di unità territoriali riguardo a fenomeni socio-economici complessi.

- Il *capitolo dedicato al calcolo delle probabilità* è stato parzialmente riscritto, soprattutto nella parte iniziale, per rendere più graduali e chiari la definizione degli eventi aleatori e il meccanismo per assegnare loro la misura di probabilità.
- Il *capitolo sulla verifica delle ipotesi* è stato ristrutturato per ampi tratti, esplicitando e chiarendo il percorso logico-metodologico con cui si arriva alla definizione della zona di rifiuto nei singoli problemi di inferenza.
- Il *capitolo dedicato all'inferenza su popolazioni finite* – che nella precedente edizione era disponibile solo in formato elettronico –, sostanzialmente modificato e ampliato, fa ora parte integrante del volume.
- Le sezioni *Commenti e curiosità* di diversi capitoli sono state arricchite con l'obiettivo di offrire ulteriori spunti di riflessione sull'evoluzione storica e sulle idee fondanti della statistica.
- In vari *capitoli*, in particolare nel primo, nel decimo e nell'undicesimo (dedicati agli aspetti introduttivi, alla regressione lineare semplice e alla correlazione), sono state eliminate alcune ridondanze e usata una terminologia più in linea con gli standard internazionali. Si è scelto anche di adottare come separatore decimale il punto e come separatore delle migliaia uno spazio, limitatamente ai numeri da 10 000 in avanti.

Ringraziamenti

Gli autori esprimono gratitudine a Mauro Costantini, Fulvio De Santis e Vincenzina Vitale per aver contribuito alla stesura di alcune note di *Commenti e curiosità*.

Un grazie speciale va a Daniela Marella per l'impegno profuso nel controllo delle bozze di tutti i capitoli e in particolare di quelle del Capitolo 24. Grazie anche a Leonardo Alaimo, Franco Bressan, Livia De Giovanni e Raffaele Mattera, che hanno controllato parte delle bozze.

Gli autori sono grati infine a Michele Baldassarri, che ha svolto un delicato e prezioso lavoro, curando l'impaginazione preliminare del volume, con ripetuti interventi sul latex dei singoli capitoli, e il controllo minuzioso delle soluzioni degli esercizi con l'aggiornamento del relativo file Excel.

Piattaforma MyLab

Sulla piattaforma MyLab del corso sono disponibili:

- Le soluzioni di tutti gli esercizi proposti, ad eccezione di quelli teorici contrassegnati con un asterisco. Per gli esercizi di inferenza più complicati, la soluzione non si limita al risultato, ma illustra il procedimento per ottenerlo.
- Le soluzioni in Excel. Si tratta di uno strumento molto utile, che si raccomanda, perché il foglio elettronico consente di comprendere, tramite la lettura delle formule nelle caselle, le operazioni matematiche che portano al risultato finale.
- Insiemi di dati statistici, che in parte vengono utilizzati per lo svolgimento degli esercizi di fine capitolo contrassegnati dall'icona del computer 💻.
- Nel laboratorio interattivo MyMathLab sono disponibili infine strumenti per l'autovalutazione del livello di apprendimento: si tratta di test a risposta multipla e di esercizi interattivi di statistica descrittiva, probabilità e inferenza statistica con soluzioni guidate.

Pearson MyLab

UN AMBIENTE PER LO STUDIO

L'attività di apprendimento di questo corso continua in **MyLab**, l'**ambiente digitale per lo studio** che completa il libro offrendo **risorse didattiche** fruibili sia in **modo autonomo** sia per **assegnazione del docente**. Il **codice studente monouso sulla copertina** di questo libro consente l'**accesso per 18 mesi a MyLab.**

COME ACCEDERE

1. Registrati o effettua il **login** alla tua MyPearsonPlace come **studente universitario** all'indirizzo he.pearson.it/catalogo.

2. Attiva il prodotto all'indirizzo he.pearson.it/attiva-prodotto inserendo il codice studente monouso presente in copertina.

3. Entra nella sezione MyPearsonPlace (he.pearson.it/my-home) dove troverai i tuoi prodotti attivi.

MyLab CODICE STUDENTE MONOUSO
xxxxx-xxxxx-xxxxx-xxxxx-xxxxx-xxxxx

CHE COSA CONTIENE

MyLab offre la possibilità di accedere al Manuale online: l'**edizione digitale del testo** arricchita da funzionalità che permettono di personalizzarne la fruizione, attivare la sintesi vocale, inserire segnalibri.
Inoltre la **piattaforma digitale MyLab** integra e monitora il percorso individuale di studio con **attività formative e valutative specifiche**. La loro descrizione dettagliata è consultabile nella pagina di catalogo dedicata al libro, all'indirizzo linkpearson.it/TBJQNL450S oppure tramite il presente QR code.

CAPITOLO

1

Nozioni introduttive

mmario del capitolo

.1 Introduzione

.2 Cenni storici sullo sviluppo della statistica

.3 La statistica nel processo di acquisizione di nuova conoscenza

.4 La statistica nelle attività operative e nella vita quotidiana

.5 Fonti di dati statistici nazionali e internazionali

.6 Terminologia statistica essenziale

.7 Scale di misurazione e tipologie di variabili

.8 Genesi dei dati statistici

.9 Matrice dei dati

10 Statistica descrittiva e inferenza statistica

11 Alcuni calcoli statistici elementari: differenze relative e rapporti statistici

Riassunto

Commenti e curiosità

Esercizi

Appendice

1.1 Introduzione

Prima di descrivere il contenuto della statistica intesa come disciplina scientifica, è bene rilevare che il plurale "statistiche" è correntemente usato nel senso di dati statistici o di indicatori sintetici – tabelle, grafici, medie, percentuali ecc. – che servono a trasmettere informazione e conoscenza intorno a fenomeni naturali, economici e sociali. Così sono statistiche le informazioni diffuse dal servizio meteorologico nazionale, quelle fornite periodicamente dall'Istituto nazionale di statistica (Istat) sull'occupazione e la disoccupazione, quelle contenute nei bollettini dell'Istituto superiore di sanità sulla dinamica di un'epidemia. Naturalmente, sono statistiche anche i dati prodotti da altri enti, ad esempio, singole società commerciali che possono diffondere dati relativi alle loro attività.

La statistica – al singolare – è associata in genere all'attività di produzione e diffusione di dati statistici, tipicamente a quella dell'Istat. Il riferimento è pertinente. Non vi è dubbio infatti che la produzione e la diffusione dei dati di carattere socio-economico da parte dell'Istat è un campo importante di applicazione del metodo statistico, in quanto l'informazione statistica che ne scaturisce è utile per orientare le politiche pubbliche e costituisce una fonte preziosa di conoscenza e coscienza per i cittadini.

Per questo motivo, nelle pagine che seguono, sarà dedicata particolare attenzione all'organizzazione dell'Istituto e alle principali indagini da esso condotte, tra cui il censimento della popolazione. Tuttavia, la statistica intesa come metodologia ha un ruolo ben più ampio e pervasivo: nella ricerca scientifica, nelle attività economico-produttive, nelle attività professionali e perfino nella vita quotidiana.

Si pensi agli studi clinici controllati, in cui la metodologia statistica guida i ricercatori nella pianificazione degli esperimenti, nell'analisi dei dati e nella generalizzazione delle evidenze tratte dai dati osservati circa l'efficacia e la sicurezza del farmaco (ad esempio un vaccino) oggetto della sperimentazione.

Si pensi al controllo di processo in un'azienda industriale. Per stabilire se il processo è "in controllo", si ricorre all'osservazione di un campione di pezzi prelevati dalla linea di produzione, con cui si quantifica in modo approssimato il tasso di difettosità del processo. In questa attività, si applica la statistica per la formazione del campione e per l'estensione all'intero processo del dato campionario.

E ancora, i sondaggi elettorali o di altro tipo diffusi tramite i mezzi di comunicazione di massa sono il frutto dell'applicazione della statistica, da cui si traggono le regole per la formazione del campione degli intervistati e per l'uso appropriato dei dati campionari. Regole che devono

essere conosciute, in linea di massima, anche dal lettore fruitore di queste indagini per poterne cogliere l'autentico significato.

Alla luce degli esempi proposti, si può definire la *statistica* come segue.

DEFINIZIONE 1.1

La statistica è la disciplina che elabora i principi e le metodologie che presiedono al processo di rilevazione e raccolta dei dati, all'analisi e interpretazione dei dati stessi e, sotto certe condizioni, alla generalizzazione dell'evidenza empirica tratta dal campione.

La definizione apparirà più chiara via via che si procederà nell'esposizione degli argomenti. Con l'espressione "sotto certe condizioni", ci si riferisce alla natura dei dati osservati: la generalizzazione dei risultati campionari all'intera popolazione richiede che il campione sia generato sulla base di un meccanismo casuale.

1.2 Cenni storici sullo sviluppo della statistica

La statistica è una scienza relativamente giovane: il suo sviluppo è avvenuto in gran parte nei secoli XIX e XX; le sue origini come disciplina autonoma risalgono al XVII secolo, quando in Inghilterra si sviluppò, a opera di John Graunt (1620-1674) e William Petty (1623-1687), un indirizzo di ricerca, che prese il nome di *Political Arithmetic*, caratterizzato dall'uso del metodo empirico induttivo, proprio delle scienze naturali, nell'investigazione dei fenomeni demografici e sociali[1]. Il merito di queste ricerche sta nell'aver messo in risalto, tra i contemporanei, l'importanza della raccolta e dell'appropriato uso dei dati, più che nell'aver introdotto specifiche metodologie di analisi.

Nacque così una nuova disciplina che però non assunse subito il nome di statistica. Secondo l'opinione prevalente, la parola "statistica" venne impiegata per la prima volta per designare l'indirizzo scientifico iniziato in Germania da Hermann Conring (1606-1681), con l'avvio di un insegnamento universitario di scienza politica avente per obiettivo la descrizione delle "cose notevoli di uno Stato". Dovettero trascorrere circa due secoli prima che l'indirizzo investigativo inaugurato dalla *Political Arithmetic* potesse assumere alcuni dei connotati della statistica moderna: infatti, occorsero due secoli perché si sviluppasse quella branca della matematica che ha a che fare con la logica dell'incerto, il calcolo delle probabilità, e perché fosse affrontato il problema della "misura dell'incertezza".

Per misura dell'incertezza si intende qui la quantificazione del grado di credibilità da attribuire ai risultati dell'indagine empirica: Graunt poteva stabilire, sulla base delle registrazioni ecclesiastiche (vedi Nota 1), che il tasso di sopravvivenza dei maschi dai 50 ai 70 anni fosse pari al 40%, ma non era in grado di effettuare una valutazione del grado di errore o della validità di questa stima, valutazione indispensabile per la generalizzazione del risultato.

[1] Un esempio di applicazione del nuovo approccio allo studio dei fenomeni demografici riguarda la determinazione del numero delle famiglie di Londra nel 1660 (esposta nell'opera di Graunt dal titolo *Natural and Political Observations upon the Bills of Mortality* ... pubblicata nel 1662 e basata sulle risultanze dei registri dei battesimi e dei defunti istituiti dalla Chiesa d'Inghilterra): osservando che durante un anno si registravano 3 morti ogni 11 famiglie, Graunt calcolò che vi dovessero essere circa 84 000 famiglie, dato che i decessi in quell'anno erano stati circa 23 000. Dall'osservazione dei dati con metodo scientifico, Graunt pervenne alla scoperta di vere e proprie leggi che governano i fenomeni demografici e sociali, quali l'eccedenza delle nascite maschili su quelle femminili, l'inurbamento delle popolazioni rurali ecc.

Il calcolo delle probabilità è lo strumento per la soluzione del problema diretto: usando una metafora, consente di calcolare la probabilità di estrarre una pallina di un dato colore, nota la composizione dell'urna. La statistica inferenziale ha il compito di risolvere quello che si chiama il "problema inverso"; di rispondere cioè a domande del tipo: "qual è la percentuale di palline bianche nell'urna, avendo osservato x palline bianche in n prove?".

Lo sviluppo del calcolo delle probabilità va ascritto a grandi matematici, tra i quali meritano menzione: Blaise Pascal (1623-1662), Pierre Fermat (1601-1665), Jacob Bernoulli (1654-1705), Abraham De Moivre (1667-1754), Thomas Bayes (1702-1761), Pierre Simon Laplace (1749-1827), Adrien Marie Legendre (1752-1833), Carl Friedrich Gauss (1777-1855). In verità, i contributi scientifici degli ultimi quattro studiosi non sono di pura probabilità: viene affrontato, sotto diversi aspetti, il problema inverso e vengono ottenuti risultati essenziali per lo sviluppo della statistica.

A Bayes va il merito di avere affrontato, se non risolto (a causa di calcoli matematici troppo complessi per quei tempi), il seguente problema. Sia θ la probabilità che un evento aleatorio E si verifichi in una prova: qual è la probabilità che θ assuma un valore compreso nell'intervallo (a, b) sapendo che in n prove l'evento E si è verificato x volte? Si tratta evidentemente di una questione importante: poter dire, alla luce delle osservazioni empiriche, che la quantità incognita θ (per esempio la probabilità di sopravvivenza a 70 anni di un cinquantenne, quantità evocata in precedenza) è compresa con un elevato livello di probabilità entro un determinato intervallo è un genuino processo induttivo.

Il campo di ricerca di Laplace fu molto vasto. Con riferimento agli argomenti più attinenti alla statistica, vanno ricordati: il problema induttivo già affrontato da Bayes[2], la scelta della media aritmetica per sintetizzare un insieme di misure ripetute di una stessa grandezza allo scopo di stimarla al meglio e il cosiddetto teorema del limite centrale[3].

Un importante avanzamento, che avrà un impatto fondamentale sullo sviluppo della statistica, è rappresentato dalla formulazione del cosiddetto metodo dei minimi quadrati. Nella sua applicazione più semplice, il metodo risolve il problema seguente. Date n misure di una stessa grandezza, qual è la sintesi di tali misure che meglio approssima la grandezza incognita? Secondo il metodo dei minimi quadrati, la sintesi migliore è data dalla media aritmetica delle n misure. L'introduzione del metodo viene attribuita congiuntamente a Legendre e a Gauss[4] e l'importanza che ebbe all'epoca era legata ad alcuni problemi scientifici in campo astronomico, come quello di determinare e rappresentare matematicamente il moto lunare.

Anche se l'applicazione della statistica al campo dei fenomeni sociali era stata adombrata negli scritti di Bernoulli e di Laplace, un passo decisivo in questa

[2] Come applicazione del suo metodo, Laplace considerò il fenomeno dell'eccedenza delle nascite maschili (fenomeno già osservato da Graunt, vedi Nota 1). Indicando con θ la probabilità di una nascita maschile, trovò che la probabilità che $\theta \leq 0.5$ era pari a $1.1521 \cdot 10^{-42}$, ciò alla luce dell'osservazione empirica che a Parigi, nel periodo 1745-1770, erano nati 251 527 maschi e 241 945 femmine. Da ciò trasse la conclusione che doveva essere $\theta > 0.5$.

[3] In base a questo teorema, sotto certe condizioni, la probabilità che la media di un campione di osservazioni (come il peso medio di un campione di pezzi di un processo produttivo) cada in un dato intervallo può essere calcolata utilizzando la cosiddetta curva di Gauss.

[4] In realtà Gauss rivendicò nei suoi scritti di aver usato il metodo prima del 1805, anno della pubblicazione di Legendre contenente l'esposizione del metodo. In ogni caso, Gauss seguì un'impostazione del problema in termini probabilistici che lo portò alla riscoperta della curva di Gauss come distribuzione di probabilità degli errori accidentali. Si veda al riguardo la nota 1 di "Commenti e curiosità" del Capitolo 23.

direzione fu compiuto grazie ai contributi di Adolphe Quetelet (1796-1874) e, soprattutto, di Francis Galton (1822-1911).

Quetelet fu un uomo dai molteplici interessi: matematico come formazione, si occupò, infatti, di astronomia, di fisica, di meteorologia e di sociologia. Fu anche un grande organizzatore: organizzò le statistiche ufficiali del Belgio e fondò numerose associazioni scientifiche, in patria e all'estero. Per quanto riguarda l'analisi dei dati sociali, il contributo fondamentale dello studioso fu la formulazione del concetto di uomo medio. Nello studio dei dati sulla popolazione, Quetelet esaminò una molteplicità di possibili relazioni tra fenomeni attraverso la costruzione di tabelle e grafici. Esaminò tassi di natalità e di mortalità per mese, per città, per temperatura e per ora del giorno. Studiò i caratteri antropometrici (quali peso, statura, circonferenza toracica ecc.), ma anche le qualità morali degli individui, tramite le statistiche su alcolismo, suicidi, malattie mentali ecc. La finalità di queste ricerche era quella di scoprire le leggi che regolano la società umana, similmente a quanto gli astronomi avevano fatto per le leggi dell'universo nel secolo precedente.

L'idea di uomo medio fu probabilmente concepita nel sintetizzare i dati antropometrici allo scopo di effettuare confronti tra gruppi di persone. Per esempio, disponendo dei dati sulla statura e il peso di un grande numero di coscritti francesi, la statura media e il peso medio furono assunti come statura e peso del "coscritto medio" francese. La stessa cosa fu fatta per un gruppo numeroso di coscritti belgi, allo scopo di effettuare confronti tra i due gruppi. Questo approccio, potenzialmente applicabile a qualsiasi caratteristica misurabile, doveva servire, nell'idea dello studioso, per effettuare confronti nel tempo, nello spazio e per categorie di persone: l'uomo medio era un espediente per neutralizzare le variazioni casuali presenti a livello individuale e far emergere le regolarità, le leggi che regolano la società.

Un secondo filone di ricerca vide l'applicazione della curva di Gauss allo studio dei dati antropometrici come strumento per stabilire se la popolazione studiata presentasse omogeneità, condizione necessaria per poter effettuare confronti tramite le medie. Il ragionamento era il seguente: se un insieme di misure di una data variabile sono omogenee (nel senso che sono influenzate da fattori comuni dominanti, differendo soltanto per cause accidentali), i dati si devono disporre secondo la curva degli errori accidentali. Va osservato che l'importanza dell'opera di Quetelet sta, più che nelle indicazioni metodologiche, nell'impulso che diede all'applicazione dei modelli probabilistici nello studio dei fenomeni sociali.

Si deve senza dubbio a Galton una vera e propria svolta per la costruzione di una metodologia empirica e concettuale nello studio dei fenomeni sociali. Galton studiò medicina a Cambridge, ma, grazie a una ricca eredità, potè abbandonare la carriera di medico. Ebbe molti interessi: esploratore in Africa per due anni, si occupò, successivamente, di meteorologia, psicologia, antropologia e sociologia. Tra i risultati di Quetelet, Galton fece tesoro, in particolare, della tecnica di rappresentare con la curva normale i dati relativi a casi omogenei (come gli individui di uno stesso gruppo etnico), ma le sue analisi furono molto più profonde: cercò infatti di spiegare i meccanismi che fanno sì che i dati tendano a disporsi nella forma a campana tipica della curva di Gauss. Egli si interrogò in particolare su come la trasmissione ereditaria dei caratteri da una generazione all'altra potesse essere compatibile con tale caratteristica dei dati. E fu proprio nello studio della trasmissione ereditaria dei caratteri che Galton diede il contributo principale, l'introduzione del concetto e della tecnica della regressione. Osservò infatti, dall'esame di numerosi dati sulle stature di padri e figli, che da padri più alti della media nascevano figli più alti della media, e che da padri più bassi della media nascevano figli più bassi della media; tuttavia, osservò che le stature dei figli

si discostavano dalla media meno di quelle dei loro padri. Questo fenomeno fu chiamato "regressione", intesa come tendenza della razza a tornare verso i valori medi. Con Galton si approda al XX secolo, allorché – nel corso dei primi quattro decenni – si può dire che la statistica assunse una fisionomia molto prossima a quella di oggi. Due sono le figure preminenti di questo periodo: Karl Pearson (1857-1936) e Ronald Aylmer Fisher (1890-1962).

Il primo, inglese di nascita, ma fervente ammiratore della cultura tedesca e di Marx, in onore del quale cambiò il suo nome da Charles a Karl, fu fautore dell'uso della probabilità nella generalizzazione dei risultati empirici. Introdusse una importantissima tecnica (nota come test del chi-quadrato) per la verifica dell'adattamento dei dati osservati a una curva teorica. Introdusse un'ampia classe di curve asimmetriche, utili nella descrizione dei dati relativi a situazioni in cui non sussistono le condizioni per l'applicazione della curva normale. Pearson fu un grande organizzatore: fondò riviste scientifiche di statistica e creò una scuola di grande rilievo.

Fisher è stato, forse, la figura di statistico più eminente del secolo scorso: i suoi contributi permeano gran parte della statistica odierna che si occupa dell'analisi dei dati di natura sperimentale (i dati provenienti da esperimenti controllati condotti negli studi di medicina, biometria, psicologia ecc.). Fondamentale è stato anche il suo apporto per la definizione e la costruzione dei piani degli esperimenti.

La rassegna precedente è sicuramente incompleta: si sono volute tratteggiare le figure degli studiosi che hanno avuto un maggiore influsso sullo sviluppo del metodo statistico. Per ogni ulteriore approfondimento si rinvia, tra gli altri, a Stigler (1986).

1.3 La statistica nel processo di acquisizione di nuova conoscenza

Come è emerso nella panoramica sullo sviluppo storico del metodo statistico, la statistica è nata proprio come mezzo per stabilire se, in una determinata situazione, l'evidenza empirica, i dati osservati, sono in accordo con un'ipotesi o una teoria scientifica. In questo senso, Laplace, Gauss e altri uomini di scienza operarono nell'ambito dell'astronomia. Questo stesso paradigma venne applicato nelle scienze sociali. Come caso emblematico, si ricorda il fenomeno dell'eccedenza delle nascite maschili (si veda la Nota 2) trattato da Laplace: era stata accertata, come evidenza empirica, un'eccedenza di nascite maschili (il fenomeno era stato messo in luce da Graunt nel 1662; si veda la Nota 1). Nell'analisi di Laplace, questa osservazione venne assunta come ipotesi da verificare; l'ipotesi venne confrontata con i dati reali relativi alle nascite registrate a Parigi in un certo periodo e non venne rigettata.

Dunque, nelle scienze empiriche, intendendo con ciò sia le scienze naturali (astronomia, fisica, biologia ecc.) che le scienze sociali (economia, psicologia, sociologia ecc.), la ricerca scientifica procede, almeno idealmente, secondo le fasi sotto indicate:

(a) formulazione del problema, eventualmente sotto forma di ipotesi;

(b) individuazione dei dati pertinenti;

(c) rilevazione dei dati;

(d) analisi dei dati;

(e) interpretazione dei risultati, ossia trasformazione dell'evidenza tratta dai dati in risposte alle domande da cui la ricerca ha preso avvio.

La statistica è uno strumento indispensabile nelle fasi (c) e (d). In particolare, nella fase (c) essa presiede alla pianificazione dell'indagine, dell'esperimento o dell'osservazione sul campo (si veda il Paragrafo 1.8); nella fase (d) vengono utilizzati i metodi statistici per l'analisi dei dati, sia i metodi per la descrizione ed esplorazione dei dati, sia i metodi della cosiddetta inferenza statistica, che consentono di trarre dai dati osservati indicazioni generali, mediante le tecniche della stima e della verifica delle ipotesi su specifiche grandezze concernenti il fenomeno studiato.

Va precisato che nella fase (b) la statistica si limita a fornire i criteri e gli schemi con cui acquisire i dati: la selezione dei dati utili ai fini della ricerca è compito della disciplina specifica in cui si opera (astronomia, economia ecc.). Sono pure di pertinenza della disciplina specifica l'interpretazione dei risultati.

Nella Figura 1.1 è rappresentato il processo che dalla formulazione del problema conoscitivo porta all'acquisizione di nuova conoscenza, passando attraverso l'applicazione dei metodi statistici pre-osservazionali (piano di campionamento e piano degli esperimenti), la raccolta dei dati, l'applicazione ai dati raccolti dei metodi statistici post-osservazionali (metodi di analisi dei dati).

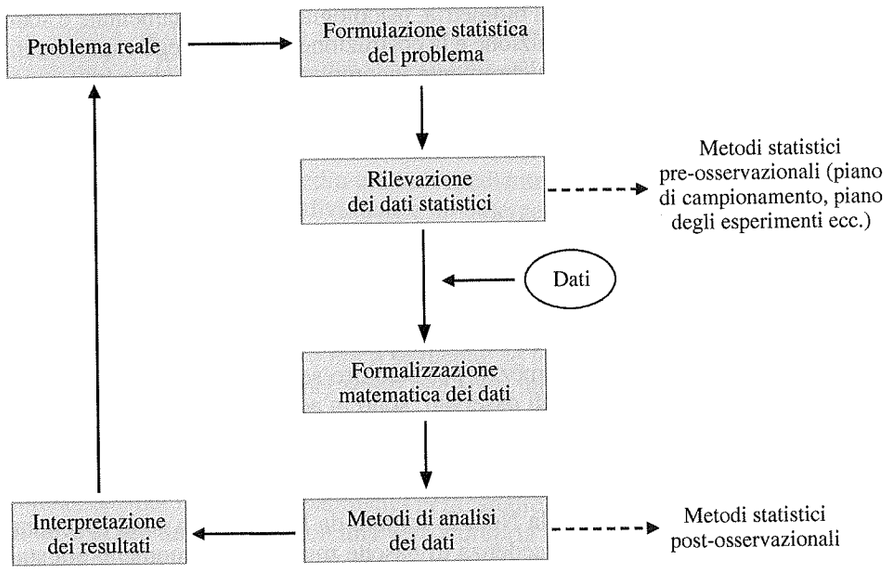


Figura 1.1
Il processo di acquisizione di nuova conoscenza con la statistica.

1.4 La statistica nelle attività operative e nella vita quotidiana

Per attività operative si intendono tutte le attività pratiche finalizzate al raggiungimento di uno scopo. Quando l'attività operativa presenta una certa complessità, la decisione di realizzarla viene presa seguendo, almeno idealmente, le fasi qui indicate:

(a) individuazione e descrizione dello scopo dell'attività;

(b) utilizzo dei metodi statistici per analizzare dati relativi, ad esempio, alle strutture commerciali concorrenti in una data area, o per raccogliere e analizzare informazioni sui potenziali clienti al fine di venire incontro ai loro bisogni e aspettative;

(c) esame dei mezzi e degli strumenti da utilizzare;

(d) previsione dei risultati conseguibili;

(e) decisione finale.

È facile comprendere come, in alcune situazioni, nelle fasi (b) e (d), possa essere utile l'impiego della statistica. Si pensi, per esempio, alla decisione da parte di una catena di supermercati di aprire un nuovo punto vendita. Nella fase (b), la statistica si rende necessaria per il reperimento e l'analisi dei dati concernenti gli esercizi concorrenti localizzati nell'area prescelta, per l'acquisizione e l'analisi delle informazioni sulla clientela potenziale onde adattare l'offerta alle caratteristiche della clientela stessa e così via. Nella fase (d), la statistica può intervenire con le tecniche di analisi di mercato volte a prevedere il comportamento di acquisto del consumatore; a questo fine, può essere utile effettuare un'indagine sul campo con apposite interviste ai potenziali clienti.

Come altro esempio, si consideri la formulazione di un provvedimento governativo volto a incentivare il mercato dell'auto. Un comportamento razionale e rigoroso richiede l'intervento della statistica nelle due fasi anzidette. Nella fase (b), per analizzare le caratteristiche del parco auto potenzialmente interessato al provvedimento; per reperire questi dati, non si deve ricorrere a un'indagine *ad hoc*, potendosi utilizzare i dati del registro automobilistico. Nella fase (d), la statistica consente di valutare il costo dell'operazione con l'impiego di modelli per la previsione del numero e delle caratteristiche dei fruitori del provvedimento.

E ancora, l'operatore di borsa che deve decidere un'operazione finanziaria impiegherà la statistica, prima di tutto, per la valutazione dello stato dell'arte nel mercato dei titoli, poi per prevedere il futuro andamento dei titoli che intende trattare.

In definitiva, in ogni attività operativa di qualche rilevanza e complessità, la statistica può essere di ausilio per assumere decisioni ponderate e consapevoli, fondate sulla osservazione e sulla previsione dei fatti.

Quando si parla del ruolo della statistica nella ricerca scientifica o nelle attività operative, si pensa a un impiego, per così dire, professionale della disciplina. Va però considerato che un bagaglio minimo di conoscenze statistiche dovrebbe essere posseduto da ogni individuo per comprendere e interpretare le informazioni di natura quantitativa che sempre più copiosamente i mezzi di comunicazione diffondono: si tratti di grafici, tabelle, statistiche sintetiche sull'inflazione, sull'occupazione, sugli incidenti stradali ecc. Non solo: è auspicabile che il cittadino sia in grado di esprimere, di comunicare, informazioni quantitative ricorrendo agli strumenti basilari della statistica. Quello della statistica è un vero e proprio linguaggio, con la propria sintassi, che amplia fortemente le capacità di comunicare e interpretare. Con un neologismo inglese, si chiama *numeracy* questa abilità che, come la *literacy* (saper leggere e scrivere), ormai conquistata dai più, è un requisito indispensabile per il cittadino moderno.

1.5 Fonti di dati statistici nazionali e internazionali

Ogni Paese è dotato di un proprio organismo ufficiale, un istituto nazionale di statistica, che ha il compito di raccogliere, elaborare e pubblicare i dati nazionali ufficiali su un'ampia varietà di aspetti: popolazione, economia, occupazione ecc. Ad esempio, esiste il Census Bureau negli Stati Uniti, lo Statistics Canada in Canada, l'INSEE in Francia, l'Istat in Italia e così via. Queste enti raccolgono dati statistici – attraverso censimenti (rilevazioni sull'intera popolazione) e inda-

gini campionarie (rilevazioni su un opportuno sottoinsieme della popolazione) – e ricevono dati dagli uffici amministrativi centrali e locali.

In generale, sono molti gli organismi e le istituzioni in ogni Paese che si occupano di produrre, elaborare e rendere disponibili dati statistici, ciascuno con riferimento al proprio ambito di attività. Oltre agli Istituti nazionali di statistica, vi sono le Banche centrali, le Camere di commercio, le associazioni di categoria e così via, che costituiscono, tutte, fonti di dati.

1.5.1 L'Istat e il Sistema statistico nazionale

In Italia, alla produzione e diffusione delle statistiche relative a tutti gli aspetti della vita del paese è preposto l'Istat, l'Istituto nazionale di statistica, un ente di ricerca pubblico con ordinamento autonomo. Nel 2016, L'Istat ha compiuto 90 anni dalla sua istituzione avvenuta nel 1926. Gli organi di governo dell'Istat sono il Presidente e il Consiglio, che esercitano le funzioni di indirizzo tecnico, scientifico e amministrativo. L'offerta informativa dell'Istat deve soddisfare rigidi protocolli di qualità: le statistiche ufficiali prodotte devono essere pertinenti, tempestive, coerenti, accurate, attendibili e comparabili tra le diverse regioni e i diversi Paesi dell'UE; in particolare esse devono soddisfare gli standard qualitativi indicati dall'Eurostat. Sulla qualità di tale offerta informativa vigila la Commissione per la garanzia dell'informazione statistica, istituita presso la Presidenza del Consiglio dei ministri.

Dal 1989 l'Istat svolge un ruolo di indirizzo, coordinamento, assistenza tecnica e formazione all'interno del Sistema statistico nazionale (Sistan), che è la rete degli Uffici statistici di tutte le amministrazioni pubbliche centrali e territoriali e degli enti di rilevanza nazionale, rete che ha come obiettivo principale la razionalizzazione della produzione e diffusione delle informazioni e l'ottimizzazione delle risorse destinate alle statistiche ufficiali, ossia le statistiche prodotte dall'Istat e dai sistemi pubblici a esso collegati. Oltre all'Istat, il Sistan comprende: gli enti e organismi pubblici d'informazione statistica (Inea, Isfol), gli uffici di statistica delle amministrazioni dello Stato e di altri enti pubblici, degli Uffici territoriali del Governo, delle Regioni e delle Province autonome, delle Province, delle Camere di commercio, dei Comuni, singoli o associati, e gli uffici di statistica di altre istituzioni pubbliche e private che svolgono funzioni di interesse pubblico.

Per l'esercizio delle sue funzioni, l'Istat si avvale del supporto del Comitato di indirizzo e coordinamento dell'informazione statistica, che, come organo di governo del Sistema statistico nazionale, delibera il programma statistico nazionale, ossia l'atto normativo che stabilisce le rilevazioni statistiche di interesse pubblico affidate al Sistema statistico nazionale e i relativi obiettivi.

Tra le molteplici indagini condotte dall'Istat si ricordano qui quelle sulle famiglie, che hanno come oggetto: i consumi, il reddito e le condizioni di vita, l'occupazione e la disoccupazione ecc. Per qualche maggiore dettaglio al riguardo, si rinvia alla nota 4 di "Commenti e curiosità".

Sono molteplici le statistiche che l'Istat mette a disposizione sotto forma di pubblicazioni mensili, annuali o con altra periodicità (quali indagini speciali e censimenti). Fra le pubblicazioni periodiche si segnala, in particolare, l'Annuario statistico italiano, pubblicazione che contiene le statistiche relative ai principali fenomeni demografici, sociali ed economici.

Recentemente, l'Istat ha sostituito i censimenti tradizionali a cadenza decennale con i censimenti permanenti effettuati attraverso rilevazioni campionarie con cadenza annuale, biennale o triennale. Le informazioni fornite dalle rilevazioni campionarie vengono integrate con quelle provenienti dalle fonti amministrative. L'integrazione dei dati campionari con i dati amministrativi garantisce l'esaustività dell'informazione e anche altri vantaggi come: l'aumento della qualità e quantità dell'offerta informativa, il contenimento del "fastidio statistico" sui cittadini e operatori economici e la riduzione dei costi complessivi. La strategia dei censimenti permanenti, coerentemente con le politiche di sviluppo europee e con il programma di modernizzazione dell'Istat, è estesa a tutte le aree tematiche: popolazione e abitazioni, imprese, istituzioni non profit e istituzioni pubbliche e agricoltura.

Nel 2018 è stato avviato il primo censimento permanente della popolazione e delle abitazioni, a cadenza annuale, in cui vengono considerate due rilevazioni campionarie sul territorio: una basata sulle liste anagrafiche e l'altra su un campione areale di indirizzi. Nel 2020 le due rilevazioni sono state sospese a causa della pandemia. Nel 2019 è stato effettuato il primo censimento permanente delle imprese, a cadenza triennale. Nel 2016 sono stati avviati il primo censimento permanente delle istituzioni non profit, a cadenza triennale, e il primo censimento permanente delle istituzioni pubbliche, a cadenza biennale.

Il primo censimento permanente dell'agricoltura è stato effettuato nel 2022. Nel 2021 è stato effettuato l'ultimo censimento generale dell'agricoltura a cadenza decennale, con cui si chiude definitivamente la lunga storia dei censimenti generali dell'Istat (si veda al riguardo la nota 3 di "Commenti e curiosità"). I risultati di tutti i censimenti permanenti vengono diffusi e sono consultabili *online* sul sito dell'Istat.

Recentemente, l'Istat ha realizzato la nuova versione di *datawarehouse*, denominata I.Stat (http://www.istat.it/dati/i_stat.html), dove i dati statistici sono organizzati per temi ed è data la possibilità di comporre grafici e tabelle personalizzati.

Inoltre, nel 2016, in occasione dei 90 anni dalla sua fondazione, l'Istat ha rinnovato e aggiornato la banca dati delle serie storiche. Essa rappresenta un notevole patrimonio informativo, costituito da circa 1500 serie relative a numerosi aspetti della realtà socio-economica del nostro Paese. Per alcune aree tematiche – salute, giustizia, aspetti demografici, commercio estero – le statistiche partono dal 1861 e arrivano a oggi. Le 1500 serie storiche sono articolate in sei macro aree: Popolazione e Società, Istruzione e Lavoro, Salute e Welfare, Industria e Servizi, Ambiente e Agricoltura, Economia e Finanza locale. Da segnalare, in particolare, la costruzione della serie storica del Pil dal 1861 al 2015 (e le relative variazioni su base annua), frutto della collaborazione tra Banca d'Italia e Istat.

1.5.2 L'Eurostat e il Sistema statistico europeo

L'Eurostat è l'ufficio statistico dell'Unione europea. È una Direzione generale della Commissione europea che raccoglie ed elabora i dati provenienti dagli Istituti di statistica degli Stati membri dell'Unione europea, promuovendo il processo di armonizzazione delle metodologie statistiche adottate dagli Istituti di statistica degli Stati membri al fine di produrre, a beneficio dell'UE, dati statistici di qualità comparabili tra Paesi e regioni. Tra le sue attività principali vi è la definizione di dati macroeconomici a supporto delle decisioni della Banca centrale europea relative alla definizione delle politiche monetarie per l'euro.

L'Eurostat coopera con altre organizzazioni internazionali come le Nazioni Unite e l'OECD nella definizione degli standard statistici internazionali. È di particolare importanza inoltre la sua attività di cooperazione per il miglioramento della capacità statistica dei paesi in via di sviluppo.

L'Eurostat opera all'interno del Sistema statistico europeo, che è la rete costituita, oltre che dall'Eurostat, dagli Istituti nazionali di statistica. L'organo di

supporto tecnico del Sistema statistico europeo è il Comitato formato dall'Eurostat e dai presidenti degli Istituti nazionali di statistica degli Stati membri e dei paesi dell'EFTA (European Free Trade Association). Tra i compiti del Comitato, fondamentale è quello della formulazione del programma statistico europeo avente a oggetto lo sviluppo, la produzione e la diffusione delle statistiche europee, tramite l'individuazione dei principali settori e obiettivi delle attività statistiche nell'orizzonte temporale di un quinquennio.

Con la nuova legge statistica europea, approvata nel 2015, viene rafforzato il ruolo degli Istituti nazionali di statistica di coordinamento dei sistemi statistici nazionali e di sostegno del principio di indipendenza professionale degli Istituti stessi. Tale normativa garantisce che le statistiche europee prodotte dai Paesi membri siano affidabili, seguano criteri e definizioni comuni, in modo che le informazioni possano essere sempre comparabili e armonizzate.

1.5.3 Altre fonti nazionali e internazionali

Oltre all'Istat e ai differenti enti che fanno parte del Sistan, vari altri enti, pubblici e privati, producono e rendono disponibili statistiche settoriali, tra cui si ricordano, a livello nazionale, quelli indicati di seguito.

- Banca d'Italia (https://www.bancaditalia.it). Le informazioni statistiche prodotte dalla Banca d'Italia riguardano le seguenti tematiche: moneta, intermediari e finanza; finanza pubblica; sistema dei pagamenti; principali indicatori dell'economia italiana; statistiche diffuse nelle pubblicazioni economiche; conti finanziari e ricchezza delle famiglie; rapporti con l'estero; indagini su famiglie e imprese; statistiche dell'eurosistema; statistiche storiche.
- Camere di commercio (http://www.camcom.gov.it).
- INPS (http://www.inps.it).
- INAIL (https://www.inail.it).
- ACI (http://www.aci.it).

A livello internazionale, oltre all'Eurostat (http://ec.europa.eu/eurostat), si segnalano gli enti sottoindicati.

- Banca centrale europea (BCE – https://www.ecb.europa.eu). È la banca centrale incaricata dell'attuazione della politica monetaria per i Paesi dell'Unione europea che hanno aderito alla moneta unica e che formano la cosiddetta zona euro.
- Commissione statistica delle Nazioni Unite (http://unstats.un.org). È una commissione del Consiglio economico e sociale delle Nazioni Unite che produce le varie statistiche demografiche, economiche e commerciali necessarie alle agenzie ONU.
- Fondo monetario internazionale (IMF – http://www.imf.org). È un'organizzazione costituita dai governi nazionali di 189 Paesi che ha come obiettivi: la promozione della cooperazione monetaria internazionale e della stabilizzazione dei cambi, la facilitazione dell'espansione e della crescita equilibrata del commercio mondiale e l'aiuto agli Stati membri a correggere temporanei squilibri nelle bilance dei pagamenti.
- Banca mondiale (http://www.worldbank.org). È un'istituzione internazionale che comprende a sua volta la Banca internazionale per la ricostruzione e lo sviluppo e l'Agenzia internazionale per lo sviluppo, che hanno come obiettivo la lotta alla povertà e l'organizzazione degli aiuti e dei finanziamenti agli Stati in difficoltà.

- Organizzazione per la cooperazione e lo sviluppo economico (OECD – https://www.oecd.org). È un'organizzazione internazionale di studi economici per i paesi membri, paesi sviluppati aventi in comune un sistema di governo di tipo democratico e un'economia di mercato. In particolare, l'OECD ha come obiettivi la promozione di politiche per: realizzare più alti livelli di crescita economica sostenibile e di occupazione nei Paesi membri, favorendo gli investimenti e la competitività e mantenendo la stabilità finanziaria; contribuire allo sviluppo dei Paesi non membri; contribuire all'espansione del commercio mondiale su base non discriminatoria in linea con gli obblighi internazionali.
- Organizzazione mondiale della sanità (WHO – http://www.who.int). È un'agenzia speciale dell'ONU per la salute.
- Organizzazione delle Nazioni Unite per l'alimentazione e l'agricoltura (FAO – http://www.fao.org). È un'agenzia specializzata delle Nazioni Unite con lo scopo di contribuire ad accrescere i livelli di nutrizione, aumentare la produttività agricola, migliorare la vita delle popolazioni rurali e contribuire alla crescita economica mondiale.

1.6 Terminologia statistica essenziale

I numeri assumono la qualità di statistiche, cioè di dati statistici, se sono il risultato dell'osservazione intenzionale di una molteplicità di casi individuali (persone, oggetti, luoghi ecc.) finalizzata alla conoscenza e alla comprensione di un fenomeno. In questo senso, la temperatura osservata in una certa stazione meteorologica alle ore 13 del 21 giugno non è un dato statistico; diviene tale se è considerata nel quadro delle temperature rilevate nello stesso istante in altre stazioni meteorologiche della regione o del Paese, perché in tal caso i dati consentono di effettuare valutazioni e confronti di possibile interesse. E ancora: il livello di colesterolo nel sangue di una persona, quale risulta dalle analisi cliniche, non è un dato statistico; lo è invece se fa parte di un insieme di osservazioni su soggetti – che si trovano in condizioni di salute simili – i quali sono sottoposti a qualche terapia; potrebbe interessare infatti studiare l'effetto della terapia sul livello di colesterolo nel sangue.

DEFINIZIONE 1.2

Una popolazione è l'insieme dei casi individuali in cui si manifesta il fenomeno oggetto di studio.

Esempi di popolazione sono un insieme di oggetti, persone, eventi, transazioni ecc.

DEFINIZIONE 1.3

Un campione è un sottoinsieme delle unità della popolazione ottenuto mediante un processo di selezione con l'obiettivo di studiare le caratteristiche della popolazione.

Si considerino gli esempi che seguono:

(a) indagine su tutte le aziende del comparto meccanico di una data regione per studiarne le caratteristiche e, in particolare, per determinare il totale degli addetti;

(b) sondaggio sull'opinione degli italiani sull'euro basato sulle interviste a 1500 persone;

(c) valutazione del grado di difettosità dei pezzi prodotti da una linea di produzione in un'azienda industriale mediante l'osservazione di 500 pezzi scelti secondo qualche criterio durante una giornata lavorativa.

Nel caso (a) la popolazione è l'insieme di tutte le aziende meccaniche della regione. Nell'esempio (b), la popolazione è l'insieme degli italiani adulti. Nell'esempio (c), non è immediatamente chiaro quale sia la popolazione. I 500 pezzi in osservazione sono senza dubbio un campione; questa volta la popolazione non esiste realmente, ma è solo ideale: è l'insieme virtuale (teoricamente illimitato) di tutte le osservazioni che si possono effettuare se si immagina che la linea di produzione continui a funzionare senza limiti sotto lo medesime condizioni operative. Questo argomento verrà ripreso e discusso più avanti, nel Capitolo 17.

DEFINIZIONE 1.4

Le unità statistiche sono i casi individuali su cui si rilevano i dati.

DEFINIZIONE 1.5

Una variabile è ogni aspetto elementare, ogni caratteristica oggetto di rilevazione sulle unità statistiche.

Nell'esempio (a), una delle variabili considerate è il numero di dipendenti della singola azienda. Nell'esempio (b), la variabile è l'opinione del cittadino italiano adulto sull'euro. Nell'esempio (c), la variabile è la caratteristica esaminata per valutare se un pezzo sia difettoso o meno: può essere il peso, la lunghezza ecc. Nella realtà, la rilevazione riguarda sempre una pluralità di variabili che possono essere studiate singolarmente o in un modo congiunto.

DEFINIZIONE 1.6

Le modalità di una variabile sono le sue possibili realizzazioni, ossia i diversi modi in cui essa può manifestarsi nelle unità statistiche.

Nel caso (a), le modalità della variabile sono i numeri $1, 2, \ldots$ Nel caso (b), le modalità possono essere, ad esempio, favorevole, non favorevole, indifferente. Nell'esempio (c), le modalità sono espresse da numeri che rappresentano il peso, la lunghezza o altra grandezza.

Le modalità di una variabile devono essere esaustive, ossia devono rappresentare tutti i possibili modi di essere della variabile e non sovrapposte, nel senso che a ogni unità statistica deve corrispondere una e una sola modalità.

Una variabile può essere qualitativa o quantitativa a seconda che le sue modalità siano costituite da espressioni verbali o da numeri.

Le variabili qualitative possono essere nominali o ordinali a seconda che le loro modalità non presentino o presentino un ordinamento naturale. Il sesso, l'attività economica e il luogo di nascita sono esempi di variabili del primo tipo. Sono esempi di variabili del secondo tipo il grado di soddisfazione per un servizio erogato alla comunità (poco, abbastanza, molto), il titolo di studio (senza titolo, licenza elementare, licenza media, diploma, laurea triennale, laurea magistrale, dottorato), il grado militare (soldato semplice, caporale, sergente ecc.).

Una variabile quantitativa può essere discreta o continua. È discreta se le sue modalità costituiscono un insieme numerabile. È continua se le sue modalità sono tutti i numeri di un dato intervallo.

L'ampiezza della famiglia, potendo assumere i valori $1, 2, 3, \ldots$, è un esempio di variabile discreta. Un altro esempio è il numero di studenti che frequentano le lezioni di un corso universitario. Se il numero di studenti iscritti al corso sono 152, la variabile numero di studenti frequentanti ha come modalità un numero intero appartenente all'intervallo $[1, 152]$.

Esempi di variabili continue sono l'altezza, l'età e la temperatura. A causa della limitata precisione degli strumenti di misurazione, in pratica anche le variabili continue diventano discrete. Ad esempio, l'età di una persona può essere misurata fino a una certa frazione di secondo, ma non con una precisione infinitesima.

Un'altra distinzione utile è quella tra variabili quantitative trasferibili e non trasferibili a seconda che abbia o non abbia senso ipotizzare la diminuzione della modalità di una data unità a vantaggio di un pari aumento della modalità di un'altra unità. Sono esempi di variabili trasferibili il reddito e il patrimonio delle famiglie.

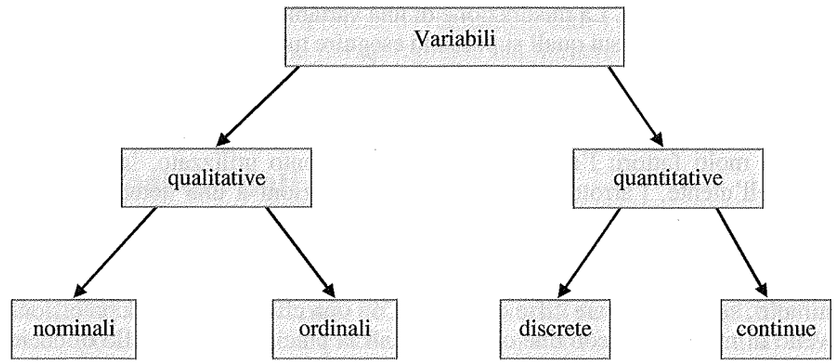


Figura 1.2
Schema di classificazione delle variabili.

1.7 Scale di misurazione e tipologie di variabili

Con riferimento a un singolo elemento dell'insieme di unità statistiche esaminato (popolazione o campione), misurare una variabile consiste nell'assegnare ad essa la modalità appropriata. Questa operazione, condotta per tutte le unità dell'insieme, produce le osservazioni, ossia i dati statistici, che sono l'oggetto delle analisi statistiche. La misurazione si declina in modi diversi per i differenti tipi di variabili.

Variabili nominali

Misurare una variabile nominale equivale a riconoscere che l'unità statistica osservata presenta l'una o l'altra delle modalità della variabile. Ad esempio, nel caso dello stato occupazionale delle persone in età lavorativa, misurare la variabile consiste nello stabilire quale delle tre modalità, occupato, disoccupato e altro stato, è presente nell'unità stessa.

Variabili ordinali

Misurare una variabile ordinale equivale a riconoscere che la singola unità presenta l'una o l'altra delle modalità ordinate della variabile stessa. Tali modalità possono essere etichettate con nomi, numeri o altri simboli ordinali (ad esempio, le lettere alfabetiche). Nell'esempio dei gradi militari, la misurazione consente di assegnare alla singola unità una delle possibili modalità ordinate (soldato semplice, caporale ecc.).

Variabili discrete

Se la variabile è quantitativa discreta, la misurazione consiste in un conteggio sulla singola unità statistica e nell'assegnazione ad essa del numero corrispondente. Ad esempio, in un'indagine sulle famiglie, la variabile dimensione della famiglia viene misurata tramite il conteggio dei membri della singola famiglia osservata. In questo modo, a ogni unità statistica viene associato un numero ed è possibile eseguire operazioni aritmetiche (addizione, sottrazione, moltiplicazione o divisione) e operazioni logiche (come "uguale a" o "maggiore di") sui dati numerici così ottenuti.

Variabili continue

In questo caso, la misurazione consiste nell'osservare i valori forniti da un apposito strumento di misura. La misurazione di una variabile per tutte le unità di studio fornisce dati numerici sui quali si possono eseguire tutte le operazioni aritmetiche e logiche già menzionate per il caso discreto.[5]

In pratica, la misurazione di una variabile continua produce sempre un'approssimazione del valore vero della grandezza. Il livello di approssimazione dipende da molti fattori: l'accuratezza dello strumento utilizzato, la capacità di lettura dell'utente, l'arrotondamento delle misurazioni a una determinata cifra decimale e così via.

La prima scelta da compiere è quella relativa al livello di precisione che si vuole raggiungere. Per esempio, l'altezza di un individuo può essere misurata in centimetri, senza utilizzare cifre decimali. Con una cifra decimale la precisione è a livello di millimetro, con due cifre decimali la precisione è a livello di decimo di millimetro e così via. Il livello di precisione è legato all'ordine di grandezza del fenomeno in studio. Così, per la statura delle persone, l'approssimazione al centimetro è ordinariamente sufficiente; non lo è per la lunghezza dei petali di una specie di fiori, per la quale una precisione adeguata è il millimetro, se non il decimo di millimetro. È anche evidente che dal livello di precisione desiderato dipende la scelta dello strumento più appropriato (il metro per la statura, il calibro per la lunghezza dei petali ecc.).

[5] Chiaramente, si può stabilire una gerarchia tra i livelli di misurazione, basata sulle operazioni logiche e aritmetiche consentite. La posizione più bassa è occupata dalla scala nominale, rispetto alla quale le unità statistiche si possono classificare solo in categorie. In seconda posizione, vi è la scala ordinale, rispetto alla quale le unità statistiche si possono ripartire tra modalità che presentano un ordinamento. In alto, si trova la scala di misura per le variabili discrete e per quasi tutte le variabili continue; in questo caso, sui dati si possono eseguire tutte e quattro le operazioni aritmetiche (–, +, ×, /) e le operazioni logiche ("uguale a" o "maggiore di").

In molti manuali di statistica, si dice che queste variabili si misurano su una scala di rapporti, per distinguerle dalle variabili quantitative per le quali il numero 0 non significa assenza della variabile, e quindi il rapporto tra due modalità non è indicativo ai fini del confronto. Ad esempio, se la temperatura (in gradi Celsius) è pari a 20 a Roma e 10 a Parigi, non si può dire che a Roma sia due volte più caldo che a Parigi. Infatti, 0 sulla scala Celsius non significa assenza di temperatura, è semplicemente un punto della scala (il punto di congelamento dell'acqua).

Per ogni misurazione "approssimata", l'ultima cifra (spesso chiamata cifra incerta) dipende dalla sensibilità del dispositivo di misurazione. Ad esempio, quando si misura l'altezza, una misura registrata di 174 cm indica che è stato osservato un valore compreso tra 173.5 e 174.5; pertanto, la registrazione implica un errore rispetto alla misurazione effettiva; l'errore massimo è di mezzo centimetro, e questo si verifica quando la misura effettiva è esattamente 173.5 o 174.5. Se il grado di precisione è il millimetro, la misura di 174.6 cm indica che è stato osservato un numero compreso tra 174.55 e 174.65; in questo caso, l'errore dovuto all'arrotondamento sarà al massimo di mezzo millimetro.

Quanto detto vale per una qualsiasi variabile continua: la misura che il rilevatore annota sottende sempre un intervallo di valori possibili, detto intervallo di tolleranza, e presenta una deviazione rispetto alla misura effettiva non superiore alla metà dell'unità che esprime il livello di precisione prescelto.

È opportuno segnalare un'apparente eccezione. Si consideri, ad esempio, l'età delle persone espressa in anni compiuti. L'annotazione 52 anni indicherà che l'età osservata è compresa nell'intervallo chiuso a sinistra [52, 53), non già, come in precedenza, nell'intervallo (51.5, 52.5). Ciò perché l'unità di misura "anni compiuti" implica un'approssimazione solo per difetto: si prende infatti la parte intera del numero che esprime l'età reale del soggetto (in anni e frazioni); così alle età effettive 52 e 2 mesi e 52 e 11 mesi corrisponderà lo stesso numero di anni compiuti, 52. Ne segue che l'errore massimo che si commette è vicino a un anno. Questa osservazione va estesa ovviamente a tutti i casi in cui la variabile è il tempo e le modalità sono durate espresse dal numero di intervalli temporali (anni, mesi ecc.) completamente trascorsi.

1.8 Genesi dei dati statistici

I dati statistici traggono origine da un'attività intenzionale rivolta all'acquisizione di informazioni sul fenomeno o sui fenomeni di interesse. Una parte notevole delle statistiche di cui si occupano gli Istituti nazionali di statistica o gli altri organismi deputati alla produzione e diffusione di dati statistici sono legati all'attività amministrativa o istituzionale di enti che hanno le più diverse sfere di competenza. Per esempio, le attività degli uffici anagrafici dei comuni, le cui registrazioni hanno finalità civili e amministrative, producono dati statistici che sono alla base delle statistiche demografiche dell'Istat relative alla popolazione residente. E ancora, l'anagrafe delle imprese delle Camere di commercio, dove vengono registrate le iscrizioni e le cancellazioni delle imprese a livello di provincia, ha una funzione rivolta agli operatori economici, ma produce dati statistici di interesse più generale. Considerazioni simili valgono per il Pubblico registro automobilistico dell'ACI, per i dati provenienti dall'attività istituzionale della Banca d'Italia nel comparto del credito e così via.

Dal punto di vista dell'utilizzatore, le fonti di dati si distinguono in primarie e secondarie a seconda che l'utilizzatore coincida o meno con chi ha raccolto i dati. Per esempio, i dati sulla nati/mortalità delle imprese costituiscono una fonte primaria per le Camere di commercio e una fonte secondaria per gli studiosi di economia che ne fanno uso.

In questo paragrafo, assumendo il punto di vista di chi vuole effettuare uno studio, una investigazione, si descrivono sinteticamente i diversi processi che danno origine ai dati statistici: l'indagine statistica, l'esperimento, lo studio osservazionale o sul campo.

1.8.1 Indagini statistiche

Si parla di indagine statistica quando lo studio riguarda una popolazione, le cui unità sono entità (persone, imprese, abitazioni ecc.) individuabili e osservabili, che va sotto il nome di popolazione finita. Rientra in questa tipologia l'indagine Istat sulle forze di lavoro, volta a stimare il numero di occupati e disoccupati e a raccogliere informazioni sui principali aggregati dell'offerta di lavoro, come il tipo di occupazione, l'area di attività economica ecc. Un altro esempio è dato dall'indagine sulle famiglie, che raccoglie dati socio-demografici relativi alle condizioni di vita delle persone (si veda la nota 4 di "Commenti e curiosità" di questo capitolo).

Se la rilevazione è limitata a una parte delle unità statistiche della popolazione, si parla di indagine campionaria. Quando invece la rilevazione prevede l'osservazione della totalità delle unità statistiche della popolazione, si parla di indagine censuaria o totale.

Sia l'indagine censuaria sia l'indagine campionaria esigono che sia disponibile la lista o base di campionamento delle unità statistiche della popolazione, ossia l'elenco delle unità statistiche che formano la popolazione, comprensivo delle indicazioni per l'individuazione e la localizzazione delle unità. Talvolta, la lista è desumibile dagli archivi tenuti da enti pubblici o privati, come nel caso dei veicoli in circolazione (la cui lista è estraibile dal Pubblico registro automobilistico). Altre volte, la lista va costruita *ex-novo* con tutte le difficoltà e i problemi di qualità per i risultati che si possono raggiungere. Tra i rischi da fronteggiare, si ricordano quello della possibile incompletezza (assenza di una parte delle unità della popolazione) e quello della possibile duplicazione di unità (unità statistiche presenti più volte).

La generalità delle indagini condotte dall'Istat o da altri enti sono di tipo campionario. La preferenza data alle indagini campionarie è dovuta al fatto che queste hanno un costo incomparabilmente più basso e possono realizzarsi in tempi molto più brevi rispetto alle indagini censuarie. Naturalmente, le elaborazioni statistiche che si effettuano sui dati del campione – le cosiddette statistiche campionarie – sono delle approssimazioni di quelle che si otterrebbero se si facesse una rilevazione censuaria. La metodologia statistica consente però di misurare il grado di approssimazione e, in un certo senso, di controllarlo, a patto che il campione sia probabilistico, cioè estratto con un meccanismo di casualità.

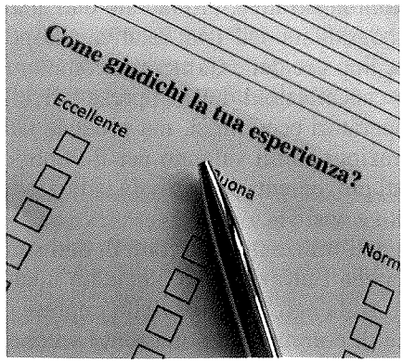


Figura 1.3
Emblema di indagine statistica.

Alcune tecniche di campionamento

Le modalità per formare un campione casuale sono molteplici; di seguito si elencano le più comuni.

- Campione casuale semplice. Le unità del campione sono selezionate con un meccanismo aleatorio tale da assicurare a tutte le unità della popolazione la stessa probabilità di essere inserite nel campione. Se si indica con N il numero delle unità della popolazione e con n la numerosità del campione, l'estrazione di un campione casuale semplice equivale a estrarre a sorte n palline da un'urna contenente N palline recanti i numeri da 1 a N che contraddistinguono le unità della popolazione (l'estrazione va intesa senza ripetizione, cioè senza rimettere la pallina nell'urna dopo ogni estrazione).
- Campione sistematico. Supponendo che il numero delle unità della popolazione, N, sia multiplo della dimensione del campione, n, e assumendo che le unità siano elencate in una lista, si definisce il passo di campionamento $p = N/n$ e si seleziona casualmente un numero intero minore o uguale a p. Se r è il numero casuale estratto, si definisce campione sistematico l'insieme delle unità contraddistinte dai numeri $[r, r + p, r + 2p, \ldots, r + (n - 1)p]$. Se N non è multiplo di n si assume come passo di campionamento l'intero più vicino a N/n. Sotto certe condizioni, il campionamento sistematico porta alla formazione di un campione molto simile al campione casuale semplice.
- Campione stratificato. Supponendo di avere informazioni aggiuntive sulla popolazione, si suddivide in base a queste la popolazione stessa in sottopopolazioni o strati, ognuno dei quali contiene unità tra loro omogenee secondo un certo criterio. Poi, da ciascuno strato si estrae un campione casuale semplice di opportuna ampiezza.
- Campione a grappoli. Si suppone che la popolazione sia suddivisa in un determinato numero di sottoinsiemi chiamati grappoli; di tali grappoli si prende un campione casuale e si considerano tutte le unità appartenenti ai grappoli prescelti.
- Campione a due stadi. Può essere visto come un campione a grappoli in cui, anziché rilevare la totalità delle unità all'interno del singolo grappolo selezionato, si limita la rilevazione a un campione casuale di tale grappolo. In questo modo, si hanno due stadi di campionamento: al primo stadio si estraggono i grappoli, al secondo le unità statistiche oggetto di studio.

ESEMPIO 1.1

Si estragga un campione casuale semplice di 10 unità dalla popolazione di cui all'Esercizio 1.6.

Soluzione

La popolazione è composta da 92 unità numerate da 1 a 92. Occorre selezionare a caso 10 numeri interi compresi tra 1 e 92. Questo può essere fatto con il computer utilizzando uno dei tanti *software* disponibili (ad esempio Excel). Si supponga di aver estratto i seguenti numeri:

$$28, 62, 24, 49, 59, 72, 5, 16, 83, 11.$$

Il campione casuale sarà composto dalle unità della popolazione contrassegnate da tali numeri. ■

ESEMPIO 1.2

Considerati $N = 20$ studenti universitari iscritti a un corso avanzato, si supponga di dover estrarre un campione sistematico di numerosità $n = 5$.

Soluzione

Bisogna innanzitutto formare un elenco numerato degli studenti (ponendoli, ad esempio, in ordine alfabetico). Poiché il passo di campionamento è $p = 20/5 = 4$, si estrae casualmente un numero intero compreso tra 1 e 4. Se il numero casuale estratto è $r = 3$, il campione sistematico sarà formato dalle unità della popolazione contraddistinte dai numeri: 3, 7, 11, 15, 19. ■

ESEMPIO 1.3

Si selezioni un campione stratificato di dimensione 15 dalla popolazione dell'Esercizio 1.5, stratificando per sesso.

Soluzione

La popolazione viene suddivisa in due strati, quello dei maschi e quello delle femmine. Poi, da ogni strato, viene selezionato un campione casuale semplice. Ciò richiede la creazione di due elenchi separati e la numerazione progressiva delle unità in ogni strato. Supponendo che l'intera dimensione del campione sia divisa in due parti proporzionali alla dimensione degli strati, il campione stratificato si ottiene selezionando un campione casuale semplice di dimensione 11 dallo strato dei maschi e un campione casuale semplice di dimensione 4 dallo strato delle femmine. ■

ESEMPIO 1.4

Un esempio di campionamento a due stadi è rappresentato dall'indagine sulle forze di lavoro condotta dall'Istat. Lo scopo dell'indagine è quello di raccogliere informazioni sul mercato del lavoro, come il numero di occupati e disoccupati, la distribuzione dei dipendenti per settore di impiego, la condizione lavorativa (occupato o autonomo), le qualifiche dei lavoratori ecc. La popolazione è composta dalle persone di almeno 15 anni. Al primo stadio, viene estratto un campione casuale di comuni; al secondo stadio, viene estratto un campione casuale di famiglie da ciascuno dei comuni selezionati al primo stadio. ■

Quando il campione è formato con un meccanismo aleatorio, del tipo di quelli descritti sopra, i metodi della statistica inferenziale consentono di estendere all'intera popolazione le conclusioni tratte dalle elaborazioni statistiche dei dati del campione. In generale, le indagini campionarie fatte dagli Istituti nazionali di statistica o da altre istituzioni pubbliche sono di tipo probabilistico. Diverse altre indagini, svolte soprattutto da organismi privati, sono basate per varie ragioni (la ragione prevalente è la necessità di limitare i costi) su campioni non probabilistici. Qui ci si limita a ricordare brevemente alcune tecniche di formazione di campioni non casuali.

- Campione per quote. La popolazione viene divisa in strati sulla base di alcune caratteristiche strutturali della popolazione. Vengono poi definite le quote, ossia il numero di persone da intervistare per ogni strato. È demandata infine all'intervistatore la scelta delle persone da intervistare all'interno di ciascuno strato.
- Campione a valanga. Si individua un primo gruppo di persone da intervistare che possiedano le caratteristiche di interesse. Queste persone a loro volta ne individuano altre che presentano le medesime caratteristiche. È una tecnica molto utilizzata per popolazioni rare o quando non si dispone di una lista.

- Campione per testimoni privilegiati. Si tratta di individuare le persone che, per gli studi svolti o per la professione esercitata, sono molto esperte sul fenomeno oggetto di studio (testimoni privilegiati). Questo tipo di campione è utilizzato in indagini che riguardano argomenti complessi e delicati. Un esempio di campione per testimoni privilegiati è il cosiddetto metodo Delphi, che consiste nella raccolta di informazioni da un gruppo di esperti chiamati a esprimere un giudizio su determinate questioni o su eventi.

L'argomento del campionamento probabilistico e del connesso problema di riporto alla popolazione dei risultati campionari sarà sviluppato nel Capitolo 24.

Va sottolineato che alla base di ogni indagine statistica vi è il cosiddetto piano dell'indagine, che specifica tutti gli aspetti relativi allo svolgimento dello studio: definizione della popolazione di riferimento, individuazione delle variabili da rilevare, definizione degli strumenti di rilevazione, determinazione dei costi della rilevazione e della relativa copertura, scelta del piano di campionamento, se si opta per l'indagine campionaria, individuazione della modalità di analisi dei dati ecc.

Il questionario

Il questionario è un insieme coordinato di domande da sottoporre alle unità statistiche (un campione o una popolazione). La costruzione del questionario è compito assai impegnativo e delicato se si considera che la qualità del questionario contribuisce in modo rilevante a determinare la qualità dell'indagine. I singoli quesiti possono essere suddivisi nelle categorie seguenti:

- domande aperte: sono domande a risposta libera;
- domande chiuse: è previsto un elenco di risposte tra cui l'intervistato è chiamato a scegliere;
- domande miste: sono domande in cui c'è un elenco delle risposte principali, ma si lascia all'intervistato la possibilità di dare una risposta diversa da quelle previste;
- domande filtro: consentono di passare direttamente da una batteria di domande a un'altra, evitando di sottoporre all'intervistato domande non pertinenti;
- domande strutturate: sono domande che ammettono risposte ad alternative prefissate e consentono all'intervistato di scegliere tra varie possibili combinazioni di risposte, spesso presentate in forma tabellare.

Particolare attenzione deve essere posta sulla formulazione delle domande chiuse. Il numero di alternative e l'ordine con cui esse vengono presentate possono influire sui risultati. Spesso alle diverse modalità di risposta sono associate delle scale di valutazione. Al riguardo, un esempio importante è la scala Likert, la più utilizzata per la sua semplicità. L'intervistato indica il suo grado di accordo o disaccordo rispetto a una sequenza di affermazioni, per esempio concordo pienamente, concordo, non concordo, dissento, dissento fortemente. Alle risposte associate alla singola domanda vengono attribuiti dei punteggi, per esempio $(1, 2, 3, 4, 5)$ o $(2, 1, 0, -1, -2)$. La somma dei punteggi ottenuti da ciascun intervistato sull'intera batteria di domande rappresenta la posizione dell'intervistato riguardo al fenomeno in esame.

Stante l'impatto del questionario sulla qualità dei risultati dell'indagine, sono previste alcune operazioni preliminari mirate a una sorta di collaudo del questionario. Al riguardo, si parla di pre-test, quando si somministra una versione prov-

visoria del questionario a un piccolo campione di persone allo scopo di rilevare eventuali problemi che gli intervistati potrebbero incontrare nel rispondere ai quesiti. Un altro accorgimento possibile è quello di sottoporre a coppie di campioni simili due versioni alternative del questionario, che differiscono per un singolo aspetto, per esempio la formulazione delle domande. Ciò al fine di individuare eventuali differenti esiti delle due formulazioni. E ancora, si può scegliere la via dell'indagine pilota con cui viene somministrato il questionario a un campione di dimensione ridotta.

Nell'Appendice a questo capitolo è riportato, a titolo di esempio, una parte del questionario che l'Istat utilizza per la conduzione del censimento permanente della popolazione e delle abitazioni.

Tecniche di somministrazione del questionario

La somministrazione del questionario si realizza con l'intervista, che può essere diretta (*face to face*), telefonica, postale, via Internet.

L'intervista diretta è stata per molti anni la tecnica di somministrazione del questionario più utilizzata. Con essa si ha un contatto diretto tra intervistato e intervistatore: l'intervistatore si reca a casa o presso il luogo di lavoro dell'intervistato e somministra il questionario cartaceo leggendo le domande, mostrando eventuale materiale illustrato e registrando le risposte sul questionario stesso. Da qualche tempo gli intervistatori utilizzano un portatile che consente la somministrazione del questionario in versione elettronica. In questo caso, l'indagine diretta è condotta col metodo CAPI (*Computer Aided Personal Interview*). Nell'intervista diretta, al fine di far comprendere meglio il significato delle domande agli intervistati, l'intervistatore utilizza domande addizionali non previste nel questionario. Sono molto diffuse anche le interviste dirette con auto-compilazione del questionario.

Nell'intervista telefonica di tipo tradizionale, l'intervistatore chiama la persona da intervistare al telefono di casa o al numero di cellulare (ciò avviene sempre più spesso con l'espansione dell'uso di questo strumento), sottopone le domande e registra le risposte sul questionario cartaceo. Oggi, nella maggior parte dei casi, viene adottato il cosiddetto metodo CATI (*Computer Assisted Telephone Interviewing*): l'intervista avviene utilizzando un questionario elettronico, che consente la codifica e l'archiviazione diretta su supporto informatico delle informazioni raccolte. Uno dei vantaggi di questa modalità di intervista è quello che via via che le risposte sono registrate esse vengono anche elaborate e sottoposte a verifiche di coerenza, cosicché l'intervistatore è in grado di correggere le eventuali risposte errate o incoerenti.

L'intervista postale si realizza attraverso l'invio presso il domicilio o il luogo di lavoro dell'intervistato di una busta contenente: una lettera di accompagnamento, il questionario e una busta per il ritorno, solitamente pre-affrancata e pre-indirizzata. Spesso si segue un approccio misto: invio del questionario per posta e sollecito telefonico (talvolta in occasione del sollecito si procede all'eventuale completamento del questionario o alla sua totale somministrazione). Si stanno sperimentando autointerviste effettuate con l'aiuto del PC, attraverso il cosiddetto metodo CASI (*Computer Assisted Self Interview*), in cui l'intervistato compila il questionario, per esempio utilizzando il proprio PC.

L'intervista via Internet può essere effettuata in due modi diversi. Si può inviare il questionario per posta elettronica dopo aver costruito una *mailing list*; il questionario viene poi restituito dai rispondenti sempre per posta elettronica. In alternativa, il questionario può essere predisposto in un sito web, nel qual caso, si invia un messaggio di posta elettronica alle persone da intervistare con l'indicazione dell'indirizzo del sito e delle credenziali per accedervi per la compilazione del questionario.

Nella Figura 1.4 sono schematicamente messe a confronto le diverse tecniche di somministrazione del questionario, elencando vantaggi e svantaggi di ciascuna di esse.

	Intervista diretta	Intervista telefonica	Intervista postale	Intervista via Internet
Vantaggi	- tasso di rifiuto basso; - elevata qualità dell'informazione raccolta, anche grazie all'uso di stimoli visivi e multimediali (nel caso del metodo CAPI) e alla possibilità di usare domande complesse e aperte; - possibilità di controllare la composizione del campione e l'identità dell'intervistato.	- tempi e costi ridotti rispetto all'intervista diretta; - facile e ampia raggiungibilità dell'intervistato; - controllo progressivo della composizione del campione; - È possibile elaborare le informazioni raccolte in tempo reale (col metodo CATI).	- costi bassi.	*posta elettronica* - tempi e costi bassi. *sito web* - tempi e costi bassi; - presentazione di stimoli multimediali, link ad altre pagine web; - dati immediatamente disponibili per l'analisi.
Svantaggi	- tempi e costi elevati; - difficile monitoraggio dell'attività degli intervistatori, che tra l'altro, possono condizionare gli intervistati; - difficilmente si ottengono informazioni su temi delicati.	- impossibilità di sottoporre stimoli visivi; - necessità di limitare il tempo di somministrazione del questionario (questionario breve); - possibile distorsione dovuta al contatto con l'intervistatore; - difficoltà nell'identificazione del rispondente; - impossibilità di contattare persone senza telefono.	- tasso di risposta generalmente basso; - problemi sulla qualità connessi all'autocompilazione; - non è possibile controllare l'identità dell'intervistato.	*posta elettronica* e *sito web* - tasso di risposta basso; - problemi di rappresentatività del campione (la diffusione di Internet non è ancora molto alta). *sito web* - tasso di risposta basso; - problemi di rappresentatività del campione.

Figura 1.4 Vantaggi e svantaggi delle diverse tecniche di somministrazione del questionario.

1.8.2 Esperimenti

Si parla di esperimento quando persone, animali o oggetti vengono sottoposti a un trattamento per osservare su di essi la risposta, ossia la reazione al trattamento. Per trattamento si intende una specifica condizione sperimentale nella quale le unità statistiche vengono osservate. La condizione sperimentale è determinata dal livello assunto da uno o più variabili, dette variabili esplicative o fattori.

ESEMPIO 1.5

Un ingegnere addetto allo sviluppo del prodotto intende studiare la resistenza alla trazione di una nuova fibra sintetica che si vuole impiegare nella produzione di camicie. Da precedenti esperienze è noto che sulla resistenza influisce la percentuale di cotone presente nella mescola della fibra. È noto anche che la percentuale di cotone potrà variare tra il 10 e il 40 per cento del peso totale della mescola se si vuole che il prodotto possegga alcune caratteristiche, come il mantenimento della stiratura. L'ingegnere fissa come livelli del fattore sperimentale le percentuali 15, 20, 25, 30, 35, e decide di effettuare 10 osservazioni per ciascun livello. (Esempio tratto da Montgomery, 2005, p. 69).

Discussione

Si tratta di un esperimento a un solo fattore, la percentuale di cotone. I trattamenti sono i livelli del fattore sperimentale, cioè le cinque percentuali di cotone.

La risposta è la resistenza alla trazione. Nell'effettuare l'esperimento potrebbe essere opportuno rendere casuale l'ordine di esecuzione delle prove, se si vuole neutralizzare l'eventuale effetto sulla risposta dell'ordine di esecuzione (si pensi all'effetto che potrebbe produrre il riscaldamento graduale della macchina). ■

ESEMPIO 1.6

Nel progettare una batteria da utilizzare in situazioni soggette a forti variazioni di temperatura, un ingegnere vuole studiare l'effetto sulla durata della batteria di due fattori, il materiale utilizzato per la piastra e la temperatura. Egli decide di sperimentare tre diversi materiali, A, B e C, e di effettuare l'osservazione della durata a tre livelli di temperatura, 15, 70 e 125; decide inoltre di effettuare quattro prove per ogni combinazione dei due fattori (esempio tratto da Montgomery, 2005, p. 199).

Discussione

I fattori sperimentali sono due, il materiale e la temperatura. I trattamenti sono nove, perché nove sono le combinazioni dei livelli dei due fattori. La variabile risposta è la durata della batteria. ■

ESEMPIO 1.7

Per studiare l'effetto di una campagna pubblicitaria, un'agenzia ha condotto un esperimento su un campione di 200 studenti universitari. Il messaggio pubblicitario è stato inserito in un programma televisivo di 50 minuti ed era caratterizzato dalla durata, 40 secondi in una versione ridotta e 80 secondi nella versione integrale, e dal numero di ripetizioni del messaggio, due volte e quattro volte. Al termine del programma televisivo, i soggetti sono stati intervistati ponendo loro domande sui contenuti del messaggio pubblicitario e sul giudizio riguardante l'oggetto reclamizzato.

Discussione

I fattori sperimentali sono due, la durata e la frequenza dei messaggi pubblicitari. I trattamenti sono quattro, numero degli abbinamenti possibili delle due durate con le due frequenze. Vi sono almeno due variabili risposta: il grado di attenzione e il grado di interesse nel prodotto suscitati dal messaggio pubblicitario. È evidente che si pone anche qui il problema dell'assegnazione dei soggetti ai trattamenti. È chiaro che l'assegnazione casuale di 50 soggetti a ciascun trattamento potrebbe limitare l'effetto dei cosiddetti fattori sub-sperimentali (ossia tutte quelle circostanze e quegli elementi che, oltre ai fattori sperimentali, influenzano la risposta) come il sesso, l'età, la classe sociale di appartenenza ecc. ■

ESEMPIO 1.8

In uno studio clinico, un gruppo di pazienti, detto gruppo sperimentale, assume il nuovo farmaco oggetto della sperimentazione, mentre un secondo gruppo, denominato gruppo di controllo, riceve una terapia standard, cioè un farmaco già utilizzato per la patologia in questione.

Discussione

Il fattore sperimentale è il nuovo farmaco. I trattamenti sono due, nuovo farmaco e vecchio farmaco. La variabile risposta è l'esito della terapia, cioè guarigione o non guarigione. È ragionevole rendere casuale l'assegnazione dei soggetti ai due trattamenti, allo scopo di neutralizzare l'impatto sulla risposta dei fattori sub-sperimentali. Al di là della schematizzazione qui effettuata, la questione della sperimentazione dei farmaci è assai complessa: è soggetta a una normativa rigorosa, volta alla tutela della salute dei cittadini, e segue procedure molto ben delineate nel cosiddetto protocollo sperimentale. ■

Copyright Rothamsted Research

Figura 1.5
Emblema di esperimento (immagine relativa alla storica stazione sperimentale inglese di Rothamstead, specializzata nella ricerche in agricoltura).

1.8.3 Studi osservazionali o sul campo

Negli studi osservazionali o sul campo, non esiste una popolazione finita da indagare, né vi sono unità statistiche che il ricercatore decide di assegnare ai diversi trattamenti: all'opposto, sono le unità stesse che, in un certo senso, si assegnano all'uno o all'altro trattamento.

ESEMPIO 1.9

Un esempio classico di studio osservazionale è quello riguardante i danni alla salute prodotti dal fumo di sigarette. Si immagini che un gruppo di ricercatori abbia organizzato un'attività di monitoraggio su 500 soggetti, di cui 150 fumatori e 350 non fumatori, in età compresa tra 30 e 40 anni, resisi disponibili a essere seguiti nel corso di 20 anni per la rilevazione di eventuali patologie tradizionalmente associate al fumo, come l'infarto, il cancro al polmone ecc.

Discussione

La variabile esplicativa o fattore è l'appartenenza o non appartenenza alla categoria di fumatore: in questo modo, si configurano due trattamenti a cui i soggetti si sono autoassegnati. La variabile risposta è l'insorgenza di una delle patologie associate al fumo. I fumatori costituiscono una sorta di gruppo sperimentale, mentre i non fumatori formano, per così dire, il gruppo di controllo. La finalità dello studio è quella di stabilire se la frequenza delle patologie indicate sia significativamente maggiore[6] tra i fumatori che non tra i non fumatori. Anche in questo tipo

[6] L'espressione "significativamente maggiore" è qui utilizzata nell'accezione corrente di "molto più grande". Nella parte del volume dedicata all'inferenza statistica, si darà all'espressione un significato tecnico preciso.

di studi, notevole cura va dedicata al controllo di eventuali fattori di disturbo, che potrebbero provocare un aumento della frequenza delle patologie in questione: è importante che il gruppo sperimentale e quello di controllo abbiano una composizione il più possibile simile per quanto riguarda il sesso, l'età e ogni altra caratteristica che potrebbe influire sulla variabile risposta. ■

ESEMPIO 1.10

Un esempio di come i risultati di studi osservazionali possano dar luogo a conclusioni distorte è riportato in Freedman *et al.* (1998) e riguarda la pellagra. Questa malattia venne identificata per la prima volta nel 1700 dal medico spagnolo Gaspar Casal tra i poveri che abitavano nelle Asturie. All'inizio del 1800 la malattia si diffuse in Europa colpendo migliaia di persone soprattutto in Francia, Austria e Romania. La pellagra fu considerata dagli epidemiologi di quel tempo come una malattia infettiva associata alle inadeguate condizioni igieniche delle famiglie colpite e diffusa da una particolare specie di mosca, chiamata *simulium*, che si nutre di sangue. Tale conclusione fu dovuta all'osservazione che questo tipo di mosca aveva la stessa diffusione territoriale della malattia ed era particolarmente attiva in primavera, stagione in cui si osservava in modo più accentuato l'insorgenza di nuovi casi di pellagra.

Discussione

L'associazione tra la diffusione della pellagra e le condizioni ambientali sopra indicate (condizioni igieniche sfavorevoli e presenza della mosca *simulium*) fece pensare che queste fossero la causa della malattia e della sua diffusione. Agli inizi del '900, grazie agli studi dell'epidemiologo americano Joseph Golberg, si scoprì che la pellagra è causata da una cattiva alimentazione (consumo di cibi, come il mais, poveri di una particolare vitamina, la niacina) e non è infettiva. L'esempio suggerisce come occorra essere cauti nell'interpretare un'associazione osservata tra variabili esplicative e variabile risposta come relazione di causa a effetto. ■

1.9 Matrice dei dati

Assumendo il punto di vista di chi ha "prodotto" i dati, mediante un'indagine, un esperimento o uno studio sul campo, ci si pone ora il problema di come organizzare i dati per poterli elaborare o per metterli a disposizione di altri soggetti per studi e analisi.

L'organizzazione più naturale consiste nel disporre i dati statistici secondo lo schema di una matrice, denominata matrice dei dati dove ogni riga corrisponde a una singola unità statistica e ogni colonna a una specifica variabile. In tal modo, si hanno tante righe quante sono le unità rilevate e tante colonne quante sono le variabili osservate.

ESEMPIO 1.11

Un esempio di matrice di dati è riportato nell'Esercizio 1.6. I dati si riferiscono a nove variabili quantitative rilevate su 92 persone. A titolo illustrativo, qui si considerano le prime dieci righe della matrice (Tabella 1.1).

Discussione

La lettura della tavola per riga informa sui valori assunti dalle variabili osservate nella singola unità; la lettura per colonna dà conto invece delle modalità che la

singola variabile assume nelle unità statistiche dell'insieme considerato. Così, osservando la prima riga si desume che l'unità individuata dal numero 1 ha 22 anni, pesa 79 kg e così via, mentre l'esame, per esempio della quarta colonna, dà l'informazione sulla statura di tutte le unità dell'insieme: si legge allora che la prima unità ha statura 184 cm, la seconda 168 e così via. ■

Unità	Età (anni)	Peso (kg)	Altezza (cm)	Collo (cm)	Torace (cm)	Addome (cm)	Coscia (cm)	Avamb. (cm)	Polso (cm)
1	22	79	184	38.5	94	83	59	28.9	18.2
2	22	70	168	34.0	96	88	60	25.2	16.6
3	23	70	172	36.2	93	85	59	27.4	17.1
4	23	90	187	42.1	100	89	63	30.0	19.2
5	23	73	184	35.5	92	77	56	27.2	18.2
6	23	85	197	38.0	97	85	59	29.7	18.3
7	24	84	181	34.4	97	100	63	27.7	17.7
8	24	95	190	39.0	105	94	66	30.6	18.8
9	24	71	180	35.7	93	82	56	28.3	17.3
10	24	95	185	39.2	102	99	71	30.3	18.7

Tabella 1.1 Porzione della matrice dei dati descritta nell'Esercizio 1.6.

Non sempre è possibile ottenere dalla rilevazione la totalità dei dati previsti dal piano dell'indagine. Può accadere che vengano a mancare, in tutto o in parte, le informazioni relative ad alcune unità statistiche. Questo problema si presenta anche nel caso degli esperimenti: si pensi alle persone che escono da uno studio clinico controllato (per volontà o per altra causa). Tutto ciò comporta che intere righe e/o singole caselle della matrice dei dati siano vuote.

Ai fini dell'analisi, è opportuno disporre di una matrice completa. Ma come riempire le caselle vuote? Vi sono tecniche statistiche che consentono di risolvere il problema senza che siano inficiati i risultati dell'analisi statistica. Di questi problemi e di altri aspetti attinenti alla qualità dei dati si tornerà a discutere nel prossimo capitolo dopo che si sarà introdotta una prima formalizzazione per le variabili oggetto della rilevazione e per le relative osservazioni.

1.10 Statistica descrittiva e inferenza statistica

Tradizionalmente, si usa suddividere il campo della statistica metodologica in due settori, denominati statistica descrittiva e inferenza statistica. I principi e i metodi della statistica descrittiva riguardano l'organizzazione, la visualizzazione e le elaborazioni statistiche mirate alla sintesi dei dati. Sono incluse quindi tutte quelle attività concernenti la costruzione di tabelle, la rappresentazione grafica, il calcolo di medie o qualsiasi altra misura di sintesi opportuna che possa aiutare a comprendere il fenomeno oggetto di studio in maniera semplificata.

La statistica inferenziale invece include l'insieme dei metodi che consentono di generalizzare le conclusioni tratte dai dati di un campione, come avviene nelle indagini parziali; in esse, il campione casuale viene analizzato con il fine di raccogliere informazioni sulla popolazione da cui proviene.

La statistica inferenziale è suddivisa in: stima dei parametri e verifica delle ipotesi. Per entrambe le metodologie, l'obiettivo è acquisire informazioni su costanti numeriche (ad esempio, la media della popolazione) che descrivono alcuni

aspetti della popolazione oggetto di studio. Queste quantità sono chiamate parametri. Nella stima dei parametri, i dati del campione e la teoria della probabilità vengono utilizzati per assegnare un valore (stima puntuale) o un insieme di valori (stima intervallare) a un dato parametro. Nella verifica d'ipotesi, invece, i dati campionari e la teoria delle probabilità vengono utilizzati per decidere se rifiutare o meno un'ipotesi su tale parametro, ad esempio, l'ipotesi che la media della popolazione sia inferiore o uguale a un certo valore.

Va detto che, nelle applicazioni reali, non esiste una rigida demarcazione tra i due settori della statistica: il ricercatore potrà usare i metodi della statistica descrittiva (rappresentazioni grafiche, tabelle di frequenza ecc.) come strumento preliminare per la formulazione di ipotesi o per stima del parametro di interesse.

I capitoli successivi al primo sono suddivisi in tre parti. La prima include i Capitoli 2-11 ed è dedicata alla statistica descrittiva. La seconda, costituita dai Capitoli 12-16, contiene gli elementi essenziali di calcolo delle probabilità. La terza comprende i Capitoli 17-24 ed ha come oggetto i metodi dell'inferenza statistica. Il Capitolo 24 riguarda esclusivamente l'inferenza sulle popolazioni finite definite nel Paragrafo 1.8.1.

1.11 Alcuni calcoli statistici elementari: differenze relative e rapporti statistici

È opportuno ora introdurre alcune operazioni elementari utili per confrontare grandezze statistiche. Siano a e b i livelli di uno stesso fenomeno, espressi nella stessa unità di misura. La loro differenza relativa (assumendo, ad esempio, che a sia la quantità di riferimento) è definita come il rapporto

$$\frac{b-a}{a}.$$

Questo rapporto è un numero adimensionale (un numero puro). Moltiplicando la differenza relativa per 100, si ottiene la differenza percentuale:

$$\text{differenza percentuale} = \frac{b-a}{a} \cdot 100.$$

ESEMPIO 1.12

Al 31 dicembre 2015 la popolazione residente in Italia e in Spagna risultava rispettivamente pari a 60.6 milioni e 46.5 milioni. La differenza percentuale tra le due popolazioni, prendendo come riferimento la popolazione residente in Spagna, è data da

$$\frac{60.6-46.5}{46.5} \cdot 100 = 30.3.$$

Ciò significa che, rispetto alla Spagna, la popolazione italiana è più grande del 30.3%.

■

Il confronto tra le grandezze a e b può essere effettuato anche tramite uno dei rapporti: a/b e b/a. Il rapporto a/b, per esempio, indica quante volte a contiene b: se esso è uguale 1.5, si deduce che a è pari a 1.5 volte b; se è uguale a 0.8, ciò significa che a è 0.8 volte b e così via. Moltiplicando tali rapporti per 100, si hanno i rapporti percentuali.

Dato un insieme di grandezze omogenee (cioè espresse nella stessa unità di misura), $a_1, a_2, \ldots, a_k$, le singole grandezze si possono confrontare con il totale

$A = a_1 + a_2 + \cdots + a_k$ mediante i seguenti rapporti percentuali:

$$\frac{a_1}{A} \cdot 100, \ \frac{a_2}{A} \cdot 100, \ldots, \frac{a_k}{A} \cdot 100,$$

la cui somma è ovviamente pari a 100.

I rapporti appena calcolati – anche definiti come rapporti di composizione – trasformano le quantità $a_1, a_2, \ldots, a_k$ in percentuali del totale A e sono utili per mettere meglio in luce l'importanza delle singole quantità rispetto al totale.

ESEMPIO 1.13

Un esempio di rapporti percentuali è contenuto nella Tabella 1.2. dove si riporta la frequenza delle attività artistiche nei 28 Paesi europei.

Frequenza	Percentuale
Ogni giorno	6
Almeno una volta alla settimana (non tutti i giorni)	12
Almeno una volta al mese (non tutte le settimane)	12
Almeno una volta all'anno (non tutti i mesi)	5
No negli ultimi 12 mesi	65
Totale	100

Tabella 1.2
Frequenza delle attività artistiche (EU-28, 2015. Fonte: Eurostat).

Ciascuna delle suddette percentuali è data dal rapporto (moltiplicato per 100) tra il numero di persone con 16 anni e più, appartenenti a una delle categorie considerate, e la popolazione totale con 16 anni e più.

■

Se le grandezze $a_1, a_2, \ldots, a_k$ vengono messe a confronto tra di loro, si ottengono i cosiddetti rapporti di coesistenza. Questi rapporti trovano larga applicazione nell'analisi dei fenomeni demografici. Un esempio è rappresentato dal cosiddetto rapporto di mascolinità, rapporto, moltiplicato per 100, tra la popolazione residente di sesso maschile e quella di sesso femminile.

ESEMPIO 1.14

In Europa, al 31 dicembre 2015, la popolazione residente era composta da 249.4 milioni di maschi e da 260.9 milioni di femmine (Fonte: Eurostat). Il rapporto di mascolinità era

$$\frac{249.4}{260.9} \cdot 100 = 95.6.$$

Se ne deduce che vi erano 95.6 maschi ogni 100 femmine.

■

Altri esempi di rapporti di coesistenza sono gli indicatori della struttura per età della popolazione, quali l'indice di vecchiaia, dato dal rapporto, moltiplicato per 100, tra i residenti con età di 65 anni e più e i residenti con meno di 15 anni; l'indice di dipendenza degli anziani, definito come il rapporto percentuale tra i residenti con età di 65 anni e oltre e i residenti di età compresa tra i 15 e i 64 anni.

Vale la pena menzionare ora i cosiddetti tassi, dove il numeratore è il numero di occorrenze di un evento durante un periodo di tempo e il denominatore è il numero di persone che sono esposte a tale evento in quel periodo di tempo.

Un esempio è dato dal tasso di natalità, definito come il rapporto, moltiplicato per 1000, tra il numero dei nati vivi in un certo anno e la popolazione residente media dell'anno.[7] Questo rapporto mostra il numero di nascite di un dato anno per 1000 abitanti.

ESEMPIO 1.15

La tabella dell'Esercizio 1.8 riporta i nati vivi e la popolazione media dell'anno di alcuni Paesi dell'UE nel 2018. Con questi dati è possibile calcolare i tassi di natalità. Ad esempio, il tasso di natalità dell'Austria è 9.66 (per 1000 residenti). ■

Sembra utile infine definire i rapporti di densità dove il numeratore è la misura di una data variabile e il denominatore è la dimensione (lunghezza, superficie, estensione temporale ecc.) di una seconda variabile a cui la prima è connessa. Rientrano in questa categoria di rapporti la densità di abitanti per km^2, il numero di veicoli circolanti per km di rete viaria, il reddito pro-capite.

Riassunto

Questo capitolo definisce i contenuti della statistica e presenta una breve panoramica dello sviluppo di questa disciplina dal XVI al XX secolo, delineando il ruolo che essa svolge nella ricerca scientifica e nella vita quotidiana. Vengono poi presentati e spiegati i principali termini tecnici (popolazione, campione, unità statistica, variabile). Segue una discussione sui diversi tipi di variabili (nominale, ordinale, discreta, continua) e su cosa si intenda per misurazione con riferimento a ciascun tipo di variabile.

Particolare enfasi è data alle tecniche di produzione dei dati, distinguendo tra indagini statistiche, esperimenti e studi osservazionali.

Per quanto riguarda le indagini, vengono illustrate le principali tecniche di costruzione di un campione casuale, le problematiche inerenti alla formulazione del questionario e le diverse modalità di somministrarlo.

La definizione generale di esperimento è fornita mediante diversi esempi tratti da applicazioni reali. Il tema degli studi osservazionali o sul campo è organizzato in modo simile.

L'ultima parte del capitolo è dedicata alla definizione di alcuni strumenti statistici elementari (rapporti, tassi, indici) utili per effettuare in modo appropriato confronti tra grandezze associate a fenomeni economici e sociali.

Parole chiave

base di campionamento, 16
campione, 11
campione a due stadi, 17
campione a grappoli, 17
campione a valanga, 18
campione casuale semplice, 17
campione per quote, 18
campione per testimoni privilegiati, 19
campione sistematico, 17
campione stratificato, 17
differenza percentuale, 26
differenza relativa, 26
domande aperte, 19
domande chiuse, 19
esperimenti, 21
fattori, 21
fattori sub-sperimentali, 22

[7] Si noti che il fattore moltiplicativo non sempre è 1000. Quando la proporzione tra le grandezze che si confrontano è minore di un millesimo, si prende come moltiplicatore 10 000 o una potenza anche maggiore di 10. Se il confronto è tra un numero piccolo (riguardante eventi rari, crimini o malattie rare) e un numero grande come la popolazione di un Paese, il moltiplicatore usato è 100 000.

fonti primarie, 15
fonti secondarie, 15
gruppo di controllo, 22
gruppo sperimentale, 22
indagini campionarie, 16
indagini censuarie, 16
indagini statistiche, 15
indagini totali, 16
indice di dipendenza degli anziani, 27
indice di vecchiaia, 27
inferenza statistica, 25
intervallo di tolleranza, 15
intervista diretta, 20
intervista postale, 20
intervista telefonica, 20
intervista via Internet, 20
Istat, 8
lista, 16
matrice dei dati, 24
modalità 12
numeracy, 7
osservazioni, 13
piano dell'indagine, 19
popolazione, 11
popolazioni finite 16
protocollo sperimentale, 23
questionario, 19
rapporti di coesistenza, 27
rapporti di composizione, 27
rapporti di densità, 28
rapporto di mascolinità, 27
scala di rapporti, 14
scala di Likert, 19
statistica, 2
statistica descrittiva, 25
statistica inferenziale, 25
statistiche campionarie, 16
stima dei parametri, 25
studi osservazionali, 15
studi sul campo, 23
tassi, 28
tasso di natalità, 28
trattamento, 21
unità statistica, 12
uomo medio, 4
variabile, 12
variabili esplicative, 21
variabili qualitative, 12
variabili qualitative nominali, 12
variabili qualitative ordinali, 12
variabili quantitative, 12
variabili quantitative continue, 13
variabili quantitative discrete, 13
variabili quantitative non trasferibili, 13
variabili quantitative trasferibili, 13
verifica d'ipotesi, 25

Commenti e curiosità

1. Sulla definizione di statistica

Come per ogni disciplina, è problematico dare una definizione completamente illuminante della statistica. Ne è testimonianza il fatto che W. F. Wilcox, in uno scritto del 1935, riuscì a elencare 115 definizioni diverse della statistica.

Talvolta si percepisce tra i non addetti ai lavori un atteggiamento di scetticismo verso la statistica o meglio verso le statistiche, colpevoli di poter essere costruite o utilizzate per trasmettere falsi messaggi. Un tale atteggiamento è suggellato nella caustica frase attribuita a B. Disraeli (1804-1881): "Ci sono tre tipi di bugie: le semplici bugie, le menzogne e le statistiche". La stessa ironia e disincanto verso le statistiche (identificate nella media aritmetica) sono espressi umoristicamente nel sonetto "La Statistica" di Trilussa (1871-1950):

LA STATISTICA

Sai ched'è la statistica? È 'na cosa
che serve pe' fa' un conto in generale
de la gente che nasce, che sta male,
che more, che va in carcere e che sposa.

Ma pe' me la statistica curiosa
è dove c'entra la percentuale,
pe' via che, lì, la media è sempre eguale
puro co' la persona bisognosa.

Me spiego: da li conti che se fanno
seconno le statistiche d'adesso
risurta che te tocca un pollo all'anno:

e, se nun entra nelle spese tue,
t'entra ne la statistica lo stesso
perché c'è un antro che ne magna due.

Molte arguzie e frasi a effetto sulla statistica si possono trovare all'indirizzo: www.istat.it/it/archivio/16448.

2. Cavour: non solo padre della patria, ma anche statistico della patria

The History Collection / Alamy Stock Photo

Camillo Benso conte di Cavour, uno dei principali protagonisti dell'Unità d'Italia, fu anche il promotore più illustre della statistica ufficiale. Pietro Maestri (medico e valoroso patriota), nella sua relazione al V Congresso internazionale di statistica del 1863 a Berlino, nella sua veste di direttore della Giunta centrale di statistica del Regno d'Italia, annoverò Cavour, "[...] che dagli studi statistici ed economici pigliò le mosse alla sua nobile carriera troppo presto, ma tanto utilmente compiuta [...]", tra gli antesignani della statistica ufficiale italiana (Fracassi, 1961).

Invero, Cavour, studioso in gioventù delle discipline matematiche, dell'economia pubblica, delle istituzioni politiche e dei problemi sociali, fu pienamente consapevole del fatto che un'efficiente organizzazione del servizio statistico dello Stato potesse illuminare i processi decisionali dei governi, accrescere l'informazione utile agli scienziati, indirizzare la pubblica opinione e favorire un ordinato sviluppo della società civile.

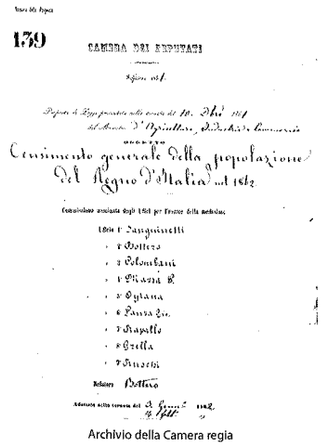
Cavour fu membro della Commissione superiore di statistica istituita a Torino nel 1836 dal re Carlo Alberto con il compito di riunire i documenti raccolti dalle Commissioni di statistica stabilite in ciascuna Provincia del Regno di Sardegna e renderli pubblici. Successivamente, dall'ottobre del 1850 al febbraio del 1852, fu chiamato da Massimo d'Azeglio a dirigere il nuovo Ministero di marina, agricoltura e commercio. In questa veste, si occupò di tutti i problemi concernenti la statistica ufficiale, compreso quello del censimento della popolazione.

Nel 1852 divenne capo del Governo del Regno di Sardegna. In questo ruolo, continuò a manifestare la sua alta sensibilità al tema dell'organizzazione della statistica ufficiale. Su sua proposta fu istituito, nel luglio 1860, il nuovo Ministero di agricoltura, industria e commercio, nel cui ambito furono riorganizzati tutti i servizi statistici, compreso quello relativo al censimento della popolazione. In tal modo Cavour attuò la sua idea di realizzare una statistica generale, che unificasse, fra l'altro, i metodi di rilevazione e le operazioni di spoglio e di pubblicazione dei dati.

Occorre infine ricordare che, anteriormente al 17 marzo 1861, data di proclamazione del Regno d'Italia, il governo presieduto da Cavour aveva già deciso di realizzare il primo Censimento generale della popolazione del Regno d'Italia. La prematura morte di Cavour, avvenuta il 6 giugno 1861, non consentì di far tesoro delle sue visioni e delle sue doti intellettuali nel processo di crescita organizzativa della statistica pubblica italiana.

3. 160 anni di storia della popolazione italiana raccontati dai censimenti

Per gli statistici del Risorgimento, la statistica era uno strumento fondamentale per la realizzazione delle basi di un futuro Stato indipendente. Come sottolineavano Cesare Correnti (patriota italiano, senatore e ministro dell'Istruzione pubblica del Regno d'Italia) e Pietro Maestri, la conoscenza della statistica e la diffusione delle statistiche pubbliche rappresentavano uno strumento fondamentale per il trasparente funzionamento di un potere legittimato dal consenso popolare e quindi da una società libera (Patriarca, 2011). In questa ottica – come scrissero Correnti e Maestri nella prefazione del primo Annuario statistico italiano del 1864 – la statistica rappresentava la "disciplina della democrazia" e "la coscienza riflessiva e sperimentale dell'umanità".


139
CAMERA DEI DEPUTATI
Censimento generale della popolazione del Regno d'Italia
Archivio della Camera regia

La consapevolezza che l'informazione statistica potesse avere un ruolo importante nella crescita dello Stato unitario indusse i governanti del neo Regno d'Italia a realizzare nel 1861 il primo Censimento generale della popolazione, utile strumento per conoscere lo stato socio-demografico della Nazione, che si rivelò in seguito fondamentale per comprendere la trasformazione nel tempo dell'Italia, in termini demografici, sociali ed economici e a diversi livelli territoriali. Informazioni indispensabili per chi riveste ruoli di governo e programmazione territoriale, per i cittadini e per le istituzioni che hanno la necessità di valutare le politiche realizzate e di avviare i processi decisionali ad esse connessi.

Il primo Censimento della popolazione italiana, come già detto, risale al 1861, anno di nascita del Regno d'Italia. Nella notte tra il 31 dicembre del 1861 e l'1 gennaio del 1862, a nove mesi dall'Unità d'Italia, gli italiani si contarono per tracciare la prima fotografia della popolazione, suddivisa per sesso, età e stato civile. La popolazione del Regno all'1 gennaio 1862 fu di 21 777 334 abitanti, che vivevano in 59 provincie, ripartite in 7720 comuni. La popolazione era distribuita in base al sesso in 10 897 236 maschi (50.04%) e 10 990 098 femmine (49.96%). L'età media fu di 27 anni, mentre la percentuale degli ultrasettantenni risultò ridottissima. Una popolazione quindi molto più giovane di quella attuale, con un alto tasso di natalità (i bambini con meno di 10 anni rappresentano il 24% del totale) e nuclei familiari numerosi (con 4 componenti in media). Il dato più allarmante fu legato all'analfabetismo: il 78 per cento degli italiani non era in grado di leggere e scrivere, con picchi del 90 per cento in Calabria, Sicilia e Sardegna.

Proposta di legge del 1861 (Archivio della Camera regia); **Decreto Regio, Gazzetta Ufficiale** (Biblioteca e Archivio Storico Istat).

La legge 20 giugno 1871, che dispose l'esecuzione del censimento generale della popolazione per il 31 dicembre 1871, ebbe grande rilevanza per vari motivi e in particolare perché istituì l'obbligo della cadenza decennale dei censimenti della popolazione (art. 1). Dal 1861 tale cadenza è stata sempre rispettata. Le uniche eccezioni furono il 1891, quando il censimento non venne svolto per difficoltà finanziarie, e il 1941, a causa della seconda guerra mondiale. Nei censimenti del 1861 e 1871 i concetti di convivenza e famiglia coincidevano; furono distinti dal 1881. Nei censimenti del 1861, 1871 e 1881 la popolazione fu classificata in "popolazione dei centri, dei casali e delle case sparse", definizione che sparì dal 1901.

Nel 1901 la data di riferimento della rilevazione venne spostata a febbraio e vennero introdotte schede individuali per ogni componente della famiglia. Nel 1911, per la prima volta, venne previsto il limite di età di 10 anni per rispondere alle domande sul lavoro, mentre nel 1921 venne realizzato l'ultimo censimento gestito dai comuni per poi essere affidato all'Istat. Nel 1931 i dati raccolti per il censimento furono per la prima volta elaborati con macchine perforatrici.

Nel 1936 il censimento venne anticipato, secondo un programma (non realizzato in seguito), di rendere le rilevazioni a cadenza quinquennale. Fu questo l'ultimo censimento della popolazione prima della seconda guerra mondiale. Il successivo venne realizzato dopo l'avvento della Repubblica, nel 1951, insieme al primo censimento delle abitazioni. Nel 1971, venne realizzato il primo censimento dei gruppi linguistici di Trieste e Bolzano con questionario tradotto anche in lingua tedesca. Nel 1991, il questionario venne tradotto in sei lingue, oltre all'italiano, e corredato da un foglio individuale per straniero non residente in Italia. Nel 2001, per la prima volta, furono censiti anche gli edifici. Quello del 2011 è stato l'ultimo censimento generale della popolazione e delle abitazioni a cadenza decennale. In esso, per la prima volta, i dati furono acquisiti anche tramite Internet.

Dal 2018 l'Istat effettua il censimento permanente della popolazione a cadenza annuale che coinvolge non più tutte le famiglie italiane, ma un campione rappresentativo di famiglie. La validità della rilevazione è garantita dal fatto che l'informazione campionaria è integrata da quella proveniente dalle fonti amministrative.

4. Su alcune indagini dell'Istat

Indagine sulle spese delle famiglie

L'obiettivo dell'indagine è quello di rilevare i comportamenti di spesa delle famiglie residenti in Italia: vengono rilevate tutte le spese sostenute dalle famiglie residenti per acquistare beni e servizi destinati al consumo familiare (inclusi gli autoconsumi, il fitto figurativo e i regali); ogni altra spesa effettuata dalla famiglia per scopo diverso dal consumo è esclusa dalla rilevazione (ad esempio il pagamento delle imposte o le spese connesse con l'attività professionale). La rilevazione, effettuata sulla base di una classificazione armonizzata (a livello EU) delle voci di spesa, è di tipo campionario ed è continua (dall'1 gennaio al 31 dicembre di ogni anno). Le unità di rilevazione sono le famiglie, che sono estratte casualmente dalle liste anagrafiche comunali. Per ulteriori informazioni si rimanda ai *link*: https://www.istat.it/it/archivio/71980; https://www.istat.it/it/archivio/180341.

Indagine sul reddito e le condizioni di vita delle famiglie

L'indagine è inserita all'interno di un più ampio sistema informativo europeo coordinato da Eurostat. Ha come obiettivo prioritario quello di fornire statistiche per l'analisi della distribuzione dei redditi, della diseguaglianza e della povertà, della deprivazione, dell'esclusione sociale e della qualità della vita delle famiglie. L'indagine fornisce due tipi di dati: dati trasversali, relativi a un determinato anno di indagine; dati longitudinali relativi a quattro anni di indagine. Le unità di rilevazione sono le famiglie che vengono estratte casualmente, nell'ambito dei comuni campione, dagli archivi anagrafici nazionali oppure dall'elenco dei nominativi già coinvolti nel censimento permanente della popolazione e delle abitazioni, secondo un piano di campionamento che le rende statisticamente rappresentative della popolazione residente in Italia. Per ulteriori informazioni si rimanda ai *link*: https://www.istat.it/it/archivio/5663; https://www.istat.it/it/archivio/216947.

Indagine sulle forze di lavoro

La rilevazione campionaria sulle forze di lavoro rappresenta la principale fonte di informazione statistica sul mercato del lavoro italiano. Le informazioni rilevate presso la popolazione rappresentano la base sulla quale vengono ottenute le stime ufficiali degli occupati e dei disoccupati e le informazioni sui principali aggregati dell'offerta di lavoro – professione, settore di attività economica, ore lavorate, tipologia e durata dei contratti, formazione. La rilevazione sulle forze di lavoro è armonizzata a livello europeo. Le unità di rilevazione sono le famiglie estratte casualmente dalle liste anagrafiche comunali in base a un piano di campionamento che consente la formazione di un campione rappresentativo della popolazione residente in Italia. La rilevazione è continua, cioè le informazioni sono rilevate con riferimento a tutte le settimane di ciascun trimestre, mediante una distribuzione uniforme del campione nelle settimane. Le famiglie estratte che fanno parte nel campione vengono intervistate quattro volte nell'arco di 15 mesi. Ogni famiglia viene intervistata per due trimestri consecutivi; subito dopo vi è un'interruzione per i due trimestri; poi, viene nuovamente intervistata per altri due trimestri. Per ulteriori informazioni si rimanda al *link*: https://www.istat.it/it/archivio/8263.

Indagine multiscopo sulle famiglie: aspetti della vita quotidiana

L'indagine campionaria "Aspetti della vita quotidiana" fa parte di un sistema integrato di indagini sociali – le indagini multiscopo sulle famiglie – e raccoglie informazioni relative alla vita quotidiana degli individui e delle famiglie. A partire dal 1993, l'indagine viene svolta con cadenza annuale. Le informazioni raccolte consentono di conoscere le abitudini dei cittadini e i problemi che essi affrontano ogni giorno e se sono soddisfatti del funzionamento di quei servizi di pubblica utilità, che dovrebbero contribuire al miglioramento della qualità della vita. Gli ambiti tematici analizzati sono: scuola, lavoro, vita familiare e di relazione, abitazione e zona in cui si vive, tempo libero, partecipazione politica e sociale, salute, stili di vita. Il campione di famiglie su cui avviene la rilevazione è estratto casualmente dall'elenco dei nominativi coinvolti nelle rilevazioni censuarie, secondo un piano di campionamento che lo rende rappresentativo della popolazione residente in Italia. Per ulteriori informazioni si rimanda al *link*: https://www.istat.it/it/archivio/91926, https://www.istat.it/it/archivio/217037.

Esercizi

1.1 Per ciascuna delle variabili di seguito indicate, si individuino la popolazione e l'unità statistica:

a) livello di informatizzazione delle imprese italiane;
b) disoccupazione negli Stati dell'Unione europea;
c) consenso degli italiani rispetto a un determinato provvedimento legislativo;
d) spesa delle famiglie italiane per l'acquisto di generi alimentari.

1.2 Si voglia condurre un'indagine su come gli studenti che frequentano un certo corso universitario raggiungono l'università. Si indichino le possibili variabili di interesse e, per ciascuna di esse, la tipologia e le modalità che può assumere.

1.3 Si consideri il *file* "Comuni", dove sono elencati i comuni di una provincia italiana e le relative popolazioni residenti (Fonte: Istat):

Unità	Popolazione residente
1	28 143
2	21 800
⋮	⋮
59	1465

Si individui l'unità statistica e si indichi il tipo di variabile.

1.4 Si consideri il *file* "Cedro nero", dove sono riportati – per un campione di 31 esemplari di cedro nero – l'altezza, il diametro e il volume (Hand *et al.*, 1994, con adattamenti):

Unità	Volume	Diametro	Altezza
1	0.292	21.08	21.34
2	0.292	21.84	19.81
⋮	⋮	⋮	⋮
31	2.180	52.32	26.52

Si indichino l'unità statistica e la tipologia delle variabili.

1.5 Si consideri il *file* "Corso di statistica", che contiene i dati di un corso universitario di statistica (Bernstein e Bernstein, 1999):

Unità	Sesso	Esame	Altezza	Peso	Colore dei capelli	Voto
1	M	90	176	82	castano	B
2	M	94	166	63	castano	A
⋮	⋮	⋮	⋮	⋮	⋮	⋮
64	F	94	164	55	biondo	A

Si identifichino le variabili considerate indicandone la tipologia.

1.6 Si consideri il *file* "Dati antropometrici" contenente i dati relativi alle caratteristiche antropometriche di un campione di individui di sesso maschile (http://lib.stat.cmu.edu/datasets, con adattamenti):

Unità	Età (anni)	Peso (kg)	Altezza (cm)	Collo (cm)	Torace (cm)	Addome (cm)	Coscia (cm)	Avamb. (cm)	Polso (cm)
1	22	79	184	38.5	94	83	59	28.9	18.2
2	22	70	168	34.0	96	88	60	25.2	16.6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
92	40	76	182	39.4	90	84	57	29.3	18.2

Si indichi, per ciascuna variabile, l'ampiezza dell'intervallo di tolleranza sotteso alle misure esposte.

1.7 Si prenda il *file* "Diamanti", che contiene i dati relativi a 308 diamanti (dimensione in carati, colore, lucentezza, tipo di certificazione, prezzo in dollari di Singapore (www.amstat.org/publications/jse/v9n2/4c.dat). Si indichi la tipologia di ciascuna variabile.

Unità	Dimensione in carati	Colore	Lucentezza	Tipo di Certificazione	Prezzo
1	0.3	D	VS2	GIA	1302
2	0.3	E	VS1	GIA	1510
⋮	⋮	⋮	⋮	⋮	⋮
308	1.09	I	VVS2	HRD	9107

1.8 La tabella seguente riporta la superficie totale, la popolazione a metà anno e i nati vivi di alcuni Paesi europei nel 2018 (Fonte: Eurostat):

Paese	Superficie totale (migliaia di km quadrati)	Popolazione a metà anno (migliaia di abitanti)	Nati vivi
Austria	83.9	8 858	85 535
Danimarca	42.9	5 806	61 476
Francia	632.8	65 018	717 795
Germania	357.3	83 019	954 874
Ungheria	93.0	9778	89 807
Italia	302.9	60 360	439 747
Polonia	312.7	38 411	388 178
Portogallo	92.2	10 277	87 020
Spagna	506.0	46 658	369 302
Regno Unito	284.5	66 436	731 213

a) Si calcoli la densità di popolazione per ciascun Paese.

b) Si calcoli il tasso di natalità per ciascun Paese (la definizione di tasso di natalità si trova nel Paragrafo 1.11).

c) Si determini la popolazione a metà anno della Romania nel 2018 sapendo che il tasso di natalità e i nati vivi di quell'anno furono 9.6 e 188 755.

Appendice

Censimento permanente popolazione e abitazioni

Questionario della sezione Abitazioni

ALLOGGIO

1. TIPO DI ALLOGGIO

* **1.1 Indicare il tipo di alloggio**

1 Abitazione ➡ *** 1.2 Indicare il tipo di abitazione**
1 Abitazione, villa/villetta monofamiliare
2 Abitazione, villa/villetta bifamiliare/plurifamiliare
3 Appartamento (ad esempio, in un condominio)

2 Altro tipo di alloggio (container, baracca, roulotte, camper, ecc.)

3 Struttura residenziale collettiva (hotel, casa di riposo, ecc.) ➡ **andare a dom. 1.5**

* **1.3 L'alloggio è occupato da**

1 Una sola famiglia ➡ **andare a dom. 1.5**

2 Più famiglie che coabitano ➡ **specificare il numero |_|_|**

* **1.4 Indicare il numero complessivo dei componenti di tutte le famiglie coabitanti**

Numero di componenti |_|_|

* **1.5 A che titolo la Sua famiglia occupa l'alloggio?**

1 Proprietà (totale o parziale) ➡ per chi ha risposto Abitazione alla domanda 1.1 andare a dom. 2.1, per chi ha risposto Altro tipo di alloggio alla domanda 1.1 andare a dom. 5.1, per chi ha risposto Struttura residenziale collettiva alla domanda 1.1 andare a dom. 5.1

2 Affitto ➡ per chi ha risposto Struttura residenziale collettiva alla domanda 1.1 andare a dom. 5.1

3 Usufrutto o riscatto ➡ per chi ha risposto Abitazione alla domanda 1.1 andare a dom. 2.1, per chi ha risposto Altro tipo di alloggio alla domanda 1.1 andare a dom. 5.1, per chi ha risposto Struttura residenziale collettiva alla domanda 1.1 andare a dom. 5.1

4 Altro titolo (gratuito, prestazioni di servizio, ecc.) ➡ per chi ha risposto Abitazione alla domanda 1.1 andare a dom. 2.1, per chi ha risposto Altro tipo di alloggio alla domanda 1.1 andare a dom. 5.1, per chi ha risposto Struttura residenziale collettiva alla domanda 1.1 andare a dom. 5.1

1.6 L'alloggio è stato preso in affitto già arredato?

1 Sì ➡ per chi ha risposto Abitazione alla domanda 1.1 andare a dom. 2.1, per chi ha risposto Altro tipo di alloggio alla domanda 1.1 andare a dom. 5.1

2 No ➡ per chi ha risposto Abitazione alla domanda 1.1 andare a dom. 2.1, per chi ha risposto Altro tipo di alloggio alla domanda 1.1 andare a dom. 5.1

Fonte: Istat

2. CARATTERISTICHE DELL'ABITAZIONE

*** 2.1 Chi è il proprietario dell'abitazione?**

1 Persona fisica (o più persone in comproprietà)
2 Impresa o società (assicuratrice, bancaria, immobiliare, di costruzione, commerciale, ecc.)
3 Cooperativa edilizia di abitazione
4 Stato, Regione, Provincia
5 Comune
6 Ente previdenziale (INPS, INARCASSA, ecc.)
7 Aziende Territoriali per l'Edilizia Residenziale (ATER) o Istituto Autonomo Case Popolari (IACP) e simili
8 Altro

*** 2.2 Qual è la superficie dell'abitazione?**

Considerare la superficie interna e calpestabile dell'abitazione, escludendo balconi, terrazze e pertinenze quali cantine, soffitte, garage, box auto, ecc.

Metri quadrati (arrotondati senza decimali) |_|_|_|

*** 2.3 Di quante stanze è composta l'abitazione?**

Esclusi i bagni, i vani accessori e le pertinenze quali cantine, soffitte, garage, box auto, ecc.

Una stanza è un locale che riceve aria e luce diretta dall'esterno e che ha dimensioni tali da consentire la collocazione di un letto lasciando lo spazio sufficiente per muoversi. La cucina con caratteristiche di stanza va inclusa nel conteggio.

Numero di stanze |_|_|

3. IMPIANTO DI CLIMATIZZAZIONE (riscaldamento, aria condizionata)

*** 3.1 Indicare l'impianto di riscaldamento dell'abitazione**

1 Impianto centralizzato ad uso di più abitazioni
2 Impianto autonomo ad uso esclusivo dell'abitazione
3 Apparecchi singoli fissi (camino, stufa, radiatore, pompe di calore, ecc.)
4 Nessun tipo di impianto ➡ **andare a dom. 3.3**

Se l'abitazione dispone di due o più impianti, selezionare il tasto Aggiungi impianto

*** 3.2 Indicare quale combustibile o energia alimenta l'impianto**
(sono possibili più risposte)

1 Metano, gas naturale
2 Gasolio
3 GPL (Gas Petrolio Liquefatto)
4 Carbone
5 Olio combustibile
6 Energia elettrica
7 Legna, pellets, cippato o altre biomasse
8 Energia solare
9 Altro combustibile o energia

3.3 L'abitazione ha un impianto fisso di aria condizionata?

1 Sì
2 No

Fonte: Istat

4. CARATTERISTICHE DELL'EDIFICIO

*** 4.1 In che tipo di edificio si trova la Sua abitazione?**

1 Residenziale (edificio per uso abitativo o prevalentemente per uso abitativo)
2 Non residenziale (edificio per uso NON abitativo o prevalentemente per uso NON abitativo)

*** 4.2 In che periodo è stato costruito l'edificio?**

1 Prima del 1919
2 Dal 1919 al 1945
3 Dal 1946 al 1960
4 Dal 1961 al 1970
5 Dal 1971 al 1980
6 Dal 1981 al 1990
7 Dal 1991 al 2000
8 Dal 2001 al 2005
9 Dal 2006 al 2010
10 Dal 2011 al 2015
11 Dal 2016 in poi

*** 4.3 Quanti interni ci sono nell'edificio?**

Considerare l'insieme delle unità immobiliari (abitazioni, studi professionali, agenzie di servizi e simili) presenti nell'edificio.

Numero di interni |_|_|_|

4.4 A quale piano dell'edificio si trova la Sua abitazione?

Indicare il piano riferendosi al lato dell'ingresso principale dell'edificio.

Per abitazioni su più livelli, indicare il piano dell'ingresso principale dell'abitazione.

1 Piano –1 o inferiore (seminterrato o interrato)
2 Piano 0 (piano terra o piano rialzato)
3 1° piano
4 2° piano
5 3° piano
6 4° piano
7 5° piano
8 6° piano
9 7° piano
10 8° piano
11 9° piano
12 10° piano o superiore ➡ **specificare il piano** |_|_|

4.5 Indicare il numero di piani fuori terra dell'edificio

Inserire nel conteggio anche il piano terra: ad esempio, in caso di edificio composto da piano terra e primo piano, indicare 2.

Se un edificio è dotato di più entrate a livelli diversi contare il numero di piani partendo dall'entrata posta al livello più basso.

Escludere dal conteggio un eventuale ultimo piano a uso NON abitativo, ad esempio con lavatoio o soffitte.

Numero di piani fuori terra |_|_|

Fonte: Istat

4.5 bis L'edificio ha un'entrata in piano oppure tramite rampa, servoscala e/o piattaforma elevatrice in modo da essere accessibile anche a persone con ridotta capacità motoria?

1 Sì
2 No

* **4.6 Nell'edificio è presente l'ascensore?**

1 Sì
2 No ➡ **andare a dom. 5.1**

Se ha risposto Sì alla dom. 4.6

4.7 L'ascensore è idoneo al trasporto di persone con ridotta capacità motoria?

1 Sì
2 No
9 Non so

5. AUTOMOBILI E POSTI AUTO

5.1 La Sua famiglia dispone di automobili?

1 Sì ➡ **specificare il numero** |_|_|
2 No

5.2 La Sua famiglia dispone di uno o più posti auto privati?

Considerare la disponibilità di almeno un posto auto, personale e garantito (come ad esempio box, posto auto in garage, posto auto in cortile ad uso riservato, ecc.), all'interno dell'edificio in cui dimora la famiglia o in prossimità di esso.

1 Sì
2 No

Se ha risposto Sì alla dom. 5.2

5.2.1 Si tratta di:

(sono possibili più risposte)

1 Box auto
2 Posto auto al chiuso (in autorimessa collettiva, garage e simili)
3 Posto auto all'aperto riservato (condominiali, cortili e simili)

Se ha risposto Box auto alla dom. 5.2.1

5.2.2 Il box auto è:

1 In affitto
2 Di proprietà (o ad altro titolo)

Fonte: Istat

PARTE I

Statistica descrittiva

Capitolo 2
Serie di osservazioni e distribuzioni di frequenza

Capitolo 3
Rappresentazioni grafiche

Capitolo 4
Medie

Capitolo 5
Variabilità

Capitolo 6
Forma delle distribuzioni di frequenza

Capitolo 7
Uno sguardo d'insieme alle costanti caratteristiche

Capitolo 8
Numeri indici e indicatori compositi

Capitolo 9
Analisi della dipendenza nelle tabelle di contingenza

Capitolo 10
Regressione lineare semplice

Capitolo 11
Correlazione

CAPITOLO

2

Serie di osservazioni e distribuzioni di frequenza

Sommario del capitolo

2.1 Introduzione
2.2 Dalle serie di osservazioni alle distribuzioni di frequenza
2.3 Distribuzioni doppie di frequenza
2.4 Serie storiche e serie territoriali
2.5 Sulle imperfezioni nei dati
Riassunto
Commenti e curiosità
Esercizi

2.1 Introduzione

Si consideri un insieme di n unità statistiche. Siano X, Y, Z, ... le variabili rilevate in un'indagine, in un esperimento o in uno studio sul campo.[1] Si indichino con $x_i, y_i, z_i, \ldots$, le modalità che queste variabili assumono nell'unità statistica identificata con il numero i ($i = 1, 2, \ldots, n$). Ad esempio, in un'indagine sull'opinione degli italiani su un provvedimento governativo, condotta con un campione di ampiezza 1500, le variabili X, Y e Z possono essere il giudizio sul provvedimento, il sesso e l'età, cosicché x_i, y_i, z_i ($i = 1, 2, \ldots, 1500$) indicheranno il giudizio (ad esempio "favorevole"), il sesso (ad esempio "maschio") e l'età (ad esempio 36 anni) osservati sull'unità i.

Come è noto, la forma tipica con cui vengono presentati i dati relativi a due o più variabili osservate su un insieme di unità statistiche è quella della matrice dei dati, dove ciascuna riga corrisponde a una unità statistica e contiene le modalità delle variabili osservate sull'unità stessa, mentre ciascuna colonna corrisponde a una variabile e contiene le modalità di quest'ultima, osservate su tutte le unità dell'insieme.

Un esempio di matrice di dati è riportato nella Figura 2.1.

Unità	Sesso	Età	Peso	Unità	Sesso	Età	Peso
1	M	A	12	19	F	A	10
2	F	A	15	20	F	C	23
3	M	A	15	21	F	A	7
4	M	A	15	22	M	A	8
5	M	A	13	23	M	C	19
6	M	A	13	24	F	C	22
7	F	A	5	25	F	A	10
8	F	C	20	26	F	A	10
9	M	A	12	27	F	C	19
10	F	A	6	28	F	C	20
11	M	C	21	29	M	B	16
12	M	A	11	30	F	A	11
13	M	B	15	31	M	A	14
14	M	A	14	32	F	A	11
15	M	A	13	33	M	C	24
16	F	C	22	34	M	A	9
17	M	A	14	35	M	B	16
18	M	A	11	36	F	C	19

Figura 2.1 Estratto dell'insieme di dati "Captures Experiments" (Hand *et al.*, 1994).

[1] In genere, le variabili vengono indicate con le lettere maiuscole dell'ultima parte dell'alfabeto latino. Se sono numerose, si usa una stessa lettera con deponenti numerici; ad esempio, quattro variabili possono essere indicate con X_1, X_2, X_3 e X_4.

Per economia di spazio, la matrice di tre colonne e 36 righe è suddivisa in due parti messe una a fianco all'altra, ma la si descrive come se fosse scritta per esteso. I dati si riferiscono a un insieme di 36 unità statistiche, 36 roditori (*peromyscus maniculatus*), su cui sono state osservate tre variabili, il sesso (M o F), l'età con tre modalità, A, B e C (A: giovane; B: semiadulto; C: adulto) e il peso (in grammi). Si osservi che i numeri posti nella prima colonna servono a identificare le unità statistiche. In senso stretto, la matrice dei dati è composta da tante colonne quante sono le variabili considerate: la numerazione progressiva delle unità è in genere omessa, essendo determinata implicitamente dalla posizione delle righe: la prima unità è quella i cui dati sono nella prima riga (procedendo dall'alto in basso), la seconda è quella i cui dati sono nella seconda riga e così via.

Si consideri ora una specifica colonna, ad esempio la prima, e si indichi con X la variabile corrispondente, in questo caso il sesso: la sequenza delle osservazioni M, F, ..., F che si leggono nella colonna, procedendo dall'alto in basso, possono essere indicate in simboli con $x_1, x_2, \ldots, x_{36}$, dove x_i $(i = 1, 2, \ldots, n)$ è la modalità di X osservata sull'unità i.

Per un insieme di n unità statistiche, le osservazioni di una generica variabile X sulle n unità dell'insieme vengono indicate con la sequenza

$$x_1, x_2, \ldots, x_n,$$

denominata serie di osservazioni della variabile X. Per segnalare che le osservazioni si riferiscono a una sola variabile, si usa talvolta l'espressione serie semplice di osservazioni.

Si considerino ora due variabili, ad esempio quelle corrispondenti alla prima e alla seconda colonna della matrice in Figura 2.1, e le si indichino con X e Y. La sequenza delle coppie di osservazioni che si leggono nelle due colonne procedendo dall'alto in basso (A, 12), (A, 15), ..., (C, 19) possono essere indicate in termini generali come: $(x_1, y_1), (x_2, y_2), \ldots, (x_{36}, y_{36})$, dove (x_i, y_i) sono le modalità che X e Y assumono nell'unità i.

Per un insieme di n unità statistiche, le osservazioni delle variabili X e Y sulle unità dell'insieme vengono indicate con la sequenza di coppie ordinate

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

denominata serie doppia di osservazioni delle variabili X e Y.

Nel caso di tre variabili, X, Y e Z, la sequenza di terne ordinate

$$(x_1, y_1, z_1), (x_2, y_2, z_2), \ldots, (x_n, y_n, z_n),$$

dove (x_i, y_i, z_i) sono le osservazioni di X, Y e Z sull'unità i, è denominata serie tripla di osservazioni.

Naturalmente, l'idea di serie di osservazioni può essere generalizzata a un numero qualsiasi, p, di variabili (si veda il Paragrafo 2.5). L'espressione generale per indicare una sequenza di osservazioni su due o più variabili è quella di serie multipla di osservazioni.

La statistica descrittiva ha come oggetto prevalente la rappresentazione e l'analisi tramite indici sintetici (medie, indici di variabilità ecc.) di serie di osservazioni relative a singole variabili. Al riguardo, con un termine di derivazione anglosassone, si parla di analisi statistica univariata. Rientra tradizionalmente nella statistica descrittiva anche l'analisi esplorativa delle relazioni tra due variabili – dipendenza, regressione e correlazione – che opera su serie doppie di osservazioni. Questo tipo di studi è indicato anche come analisi statistica bivariata.

2.2 Dalle serie di osservazioni alle distribuzioni di frequenza

È opportuno premettere che la totalità delle analisi statistiche descrittive su singole variabili quantitative (medie, indici di variabilità e indici di forma), oggetto dei Capitoli 4, 5 e 6, possono essere condotte direttamente sulle serie di osservazioni. Similmente, le analisi statistiche su coppie di variabili quantitative (regressione e correlazione) possono essere svolte direttamente sulle serie doppie di osservazioni. Se invece la variabile è qualitativa, è necessario organizzare preliminarmente i dati in una distribuzione di frequenza.

Costruire una distribuzione di frequenza da una serie di osservazioni di una variabile qualitativa è semplice: è sufficiente contare quante volte la singola modalità della variabile è presente nei dati. Quando la serie è composta da un numero limitato di osservazioni, ciò avviene utilizzando una tabella di spoglio, come descritto nell'esempio che segue.

ESEMPIO 2.1

Sulla base della matrice dei dati presentata nella Figura 2.1, si costruisca la distribuzione di frequenza per la variabile età.

Soluzione

Le modalità della variabile età sono: A, B e C. Occorre costruire una tabella di spoglio avente due colonne: la prima conterrà le modalità, la seconda accoglierà i risultati dello spoglio nel modo seguente. Dalla colonna della matrice dei dati corrispondente alla variabile età, si leggono, procedendo dall'alto in basso, le osservazioni e per ciascuna di esse si colloca una lineetta verticale nella seconda colonna della tabella di spoglio accanto alla modalità di volta in volta osservata. Per comodità di conteggio, il quinto caso viene generalmente indicato con una linea obliqua sovrapposta alle precedenti quattro linee verticali.

La prima modalità che si incontra è A, si pone allora una lineetta verticale accanto ad A; la seconda modalità è ancora A, quindi si aggiunge un'altra lineetta verticale accanto ad A e così via. Continuando questa operazione per tutte le osservazioni, si ottiene la tabella che segue:

Età	Frequenza
A	𝍸 𝍸 𝍸 𝍸 |||
B	|||
C	𝍸 𝍸

Il conteggio delle lineette verticali accanto alla singola modalità consente di stabilire quante volte la modalità è presente nei dati:

Età	Frequenza
A	23
B	3
C	10
Totale	36

Tabella 2.1
Tabella di frequenza per la variabile età, le cui osservazioni si deducono dalla Figura 2.1.

■

La Tabella 2.1 è un esempio di *distribuzione di frequenza*. Naturalmente, con il procedimento illustrato è possibile costruire distribuzioni di frequenza a partire da serie di osservazioni di variabili quantitative qualsiasi. L'Esempio 2.2 presenta la distribuzione di frequenza per una variabile quantitativa discreta.

È appena il caso di segnalare che lo spoglio manuale – così si chiama l'operazione descritta nell'esempio precedente – è solo un espediente didattico: nella realtà, lo spoglio viene realizzato generalmente con programmi di calcolo statistico.

DEFINIZIONE 2.1

Sia X una variabile, qualitativa o quantitativa, e siano $x_1, x_2, \ldots, x_k$ le sue modalità. Una distribuzione di frequenza è costituita dalla lista delle modalità di X e dalle rispettive frequenze:

Modalità	Frequenza
x_1	n_1
x_2	n_2
$\vdots$	$\vdots$
x_k	n_k
Totale	n

dove $n_1, n_2, \ldots, n_k$ sono le frequenze assolute di $x_1, x_2, \ldots, x_k$, rispettivamente.

Con la costruzione della distribuzione di frequenza, si passa da una serie di osservazioni o, per così dire, dai dati grezzi a una prima forma di sintesi, utile per una più efficace presentazione dei dati, ma anche per la comprensione del fenomeno oggetto di studio.

Nel linguaggio corrente, come sinonimo di distribuzione di frequenza si usa talora, con una metonimia, l'espressione tabella di frequenza. È opportuno sottolineare che l'organizzazione dei dati nella forma di distribuzione di frequenza è un'operazione che, oltre a essere necessaria per una efficace presentazione dei dati, si rivela utile per una prima analisi volta all'individuazione delle caratteristiche principali del fenomeno oggetto di studio.

Spesso, accanto alle frequenze assolute, vengono presentate le frequenze relative o proporzioni, ottenute rapportando le frequenze assolute al totale delle unità n:

$$f_i = \frac{\text{frequenza assoluta di } x_i}{\text{numero totale di unità}} = \frac{n_i}{n}, \quad i = 1, 2, \ldots, k.$$

Le frequenze relative sono dei particolari rapporti di composizione (si veda il Paragrafo 1.11) e come tali hanno il pregio di consentire una valutazione rapida del peso, dell'importanza, della singola modalità all'interno della distribuzione di frequenza.

Naturalmente, valgono le proprietà:

$$f_1 + f_2 + \cdots + f_k = 1, \quad 0 \le f_i \le 1, \quad i = 1, 2, \ldots, k.$$

In pratica, si utilizzano spesso le frequenze percentuali, che si ottengono moltiplicando per 100 le frequenze relative:

$$p_i = f_i \cdot 100, \quad i = 1, 2, \ldots, k.$$

In questo caso, si ha

$$p_1 + p_2 + \cdots + p_k = 100, \quad 0 \le p_i \le 100, \quad i = 1, 2, \ldots, k.$$

Quando si associano alle modalità le frequenze relative o quelle percentuali, si parla di distribuzioni di frequenza relativa o di distribuzioni di frequenza percentuale.

ESEMPIO 2.2

Si faccia riferimento al punto 4 della sezione "Commenti e curiosità" di questo capitolo, in particolare alla serie di osservazioni contenuta nella colonna relativa alla XIV esima armata, che si riproduce qui sotto per comodità:

1, 1, 2, 1, 1, 3, 0, 4, 0, 1, 0, 3, 2, 1, 0, 2, 1, 1, 0, 0.

Si costruisca la distribuzione di frequenza.

Soluzione

Dallo spoglio dei dati si ottiene la distribuzione di frequenza riportata nella tabella che segue:

Morti per calcio di cavallo	Frequenza	Frequenza relativa
0	6	0.30
1	8	0.40
2	3	0.15
3	2	0.10
4	1	0.05
Totale	20	1.00

Tabella 2.2
Distribuzione di frequenza ricavata dalle serie di osservazioni considerata.

Quando le modalità della variabile sono ordinabili, ha senso definire le cosiddette *frequenze cumulate*.

DEFINIZIONE 2.2

Data la distribuzione di frequenza della variabile X a modalità ordinabili, la frequenza cumulata della modalità x_i $(i = 1, 2, \ldots, k)$ è definita come il numero di unità che presentano modalità inferiori o uguali x_i:

$$N_i = n_1 + n_2 + \cdots + n_i, \quad i = 1, 2, \ldots, k.$$

Associando alle modalità di X le rispettive frequenze cumulate, si ottiene la distribuzione di frequenza cumulata:

Modalità	Frequenza cumulata
x_1	$N_1 = n_1$
x_2	$N_2 = n_1 + n_2$
$\vdots$	$\vdots$
x_k	$N_k = \sum_{i=1}^{k} n_i = n$

Le quantità

$$F_i = \frac{N_i}{n}, \quad i = 1, 2, \ldots, k.$$

sono chiamate frequenze relative cumulate. Da queste, moltiplicando per 100, si ottengono le frequenze cumulate percentuali:

$$P_i = F_i \cdot 100 = \frac{N_i}{n} \cdot 100, \quad i = 1, 2, \ldots, k.$$

ESEMPIO 2.3

Si calcoli la distribuzione di frequenza cumulata (assoluta e relativa) per la distribuzione di frequenza dell'Esempio 2.2.

Soluzione

Le frequenze cumulate, assolute e relative, sono riportate nella terza e quarta colonna della Tabella 2.3. Le prime si ottengono prendendo la prima frequenza assoluta da sola, la somma delle prime due, la somma delle prime tre e così via. Le frequenze relative cumulate si ricavano dividendo per 20 le frequenze assolute cumulate. Il significato dei numeri ottenuti è evidente: per esempio, 6 è il numero di unità in cui $X \leq 0$; 14 è il numero di unità in cui $X \leq 1$ e così via.

Tabella 2.3
Frequenze cumulate per la distribuzione di frequenza in Tabella 2.2.

Morti per calcio di cavallo	Frequenza	Frequenza cumulata	Frequenza rel. cumulata
0	6	6	0.30
1	8	14	0.70
2	3	17	0.85
3	2	19	0.95
4	1	20	1.00
Totale	20	–	–

■

Le frequenze cumulate si utilizzano in alcune analisi statistiche, ad esempio nella determinazione delle medie di posizione delle distribuzioni di frequenza (Capitolo 4).

2.2.1 Distribuzioni di frequenza in classi

L'osservazione di una variabile continua può dar luogo a un insieme di misure in gran parte diverse le une dalle altre, per le quali la costruzione delle distribuzioni di frequenza con il procedimento prima illustrato può non dare risultati soddisfacenti. Si pensi, ad esempio, alla statura (in cm) di 100 individui: supponendo che le modalità, cioè le misure distinte, siano 25, la classificazione delle unità secondo queste modalità costituirebbe una sintesi poco efficace. In situazioni del genere, ha senso formare delle classi raggruppando le misure vicine tra loro in intervalli numerici. Nel caso delle stature, si potrebbe considerare, in luogo delle misure singole, ad esempio 170, 171, 172, 173 e 174, l'intervallo [170, 174], indicato alternativamente con 170-174. Una volta scelti, questi intervalli costituiscono le nuove modalità: con esse si effettua lo spoglio, procedendo come illustrato in precedenza, per poi ottenere la distribuzione di frequenza in classi.

ESEMPIO 2.4

Si vogliano raggruppare in classi i dati riportati qui sotto, che si riferiscono alla lunghezza in cm dell'avambraccio di 140 maschi adulti (Hand *et al.*, 1994, p. 107; si tratta di dati utilizzati da K. Pearson all'inizio del 1900 per uno studio sulla ereditarietà).

44	47	53	43	47	52	45	52	46	52
48	44	46	43	48	51	49	49	45	46
46	48	49	48	49	53	44	47	46	49
48	48	52	50	47	45	49	46	50	54
48	52	52	50	47	51	46	50	50	48
52	44	41	49	50	48	49	49	53	47
46	46	47	52	47	42	44	44	46	50
51	47	48	51	51	47	44	47	45	44
47	47	44	48	45	48	50	49	46	47
53	50	46	43	50	52	45	49	50	47
45	51	51	47	42	49	51	44	43	46
49	47	50	46	49	43	51	41	48	53
50	47	47	47	49	41	43	46	47	49
46	46	50	52	51	44	50	48	49	45

Soluzione

Procedendo dall'alto in basso, per ciascuna unità si legge il valore assunto dalla variabile e lo si assegna alla classe di appartenenza con una lineetta verticale. Al termine dell'operazione di spoglio, vengono conteggiati i casi appartenenti alla singola classe e viene costruita la distribuzione di frequenza associando a ciascuna classe la rispettiva frequenza. Nella Figura 2.2, viene presentato lo spoglio dei dati limitatamente alle prime 14 osservazioni (quelle riportate nella prima colonna della tabella presentata sopra), avendo assunto come classi gli intervalli: 41-42, 43-44, 45-46, 47-48, 49-50, 51-52, 53-54.

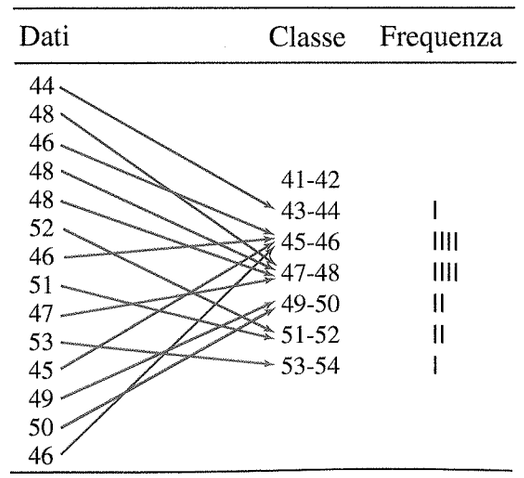


Figura 2.2
Raggruppamento in classi.

Lo spoglio completo delle 140 osservazioni dà luogo alla distribuzione di frequenza in classi contenuta nella Tabella 2.4:

Tabella 2.4
Distribuzione di frequenza in classi ottenuta dalla serie di osservazioni considerata.

Lunghezza avambraccio (cm)	Frequenza
41-42	5
43-44	17
45-46	25
47-48	35
49-50	32
51-52	20
53-54	6
Totale	140

In definitiva, le 140 osservazioni vengono così raggruppate: 5 appartengono alla classe 41-42, 17 alla classe 43-44 e così via. ■

Nel Paragrafo 1.7, è stato chiarito come a ogni osservazione di una variabile continua, così come viene annotata dal rilevatore, corrisponda un intervallo di valori possibili, l'intervallo di tolleranza. Così, nell'Esempio 2.4, a fronte della quantità 44 (primo valore in alto a sinistra), la gamma dei valori reali possibili è rappresentata dall'intervallo 43.5-44.5, a fronte di 47 (secondo valore nella prima riga) vi è l'intervallo 46.5-47.5 e così via. Ne segue che le classi di cui all'Esempio 2.4 si debbono scrivere a rigore nel modo seguente: 40.5-42.5, 42.5-44.5, ..., 52.5-54.5. Sono queste le cosiddette classi reali ottenute sottraendo 0.5 all'estremo sinistro e aggiungendo 0.5 all'estremo destro di ciascun intervallo. La costante diventa 0.05 se gli estremi di classe presentano una cifra decimale; diventa 0.005 se gli estremi di classe presentano due cifre decimali e così via.

Con il procedimento esposto, gli estremi di classe reali presentano una cifra decimale in più rispetto ai dati oggetto di classificazione: non sussiste perciò la possibilità che una osservazione coincida con un estremo di classe.

Come già accennato nel Capitolo 1, si deroga a questo criterio quando la variabile è una durata espressa in intervalli temporali (anni, mesi, giorni), come l'età in anni compiuti. In questo caso, infatti il numero intero che esprime la durata è il risultato di un'approssimazione per difetto della durata effettiva. Così per l'età, la quantità 20 (anni compiuti) è compatibile con un'età effettiva (in anni e frazione) compresa nell'intervallo [20, 21). In questi casi, le classi reali si ottengono aggiungendo 1 all'estremo destro di ciascuna classe.

Quando gli estremi di classe sono numeri che presentano la stessa precisione (cioè lo stesso numero di cifre decimali) delle osservazioni da classificare, è necessario stabilire la cosiddetta chiusura, prima di procedere allo spoglio. Volendo, per esempio, raggruppare nelle classi 1.0-2.0, 2.0-3.0 le osservazioni:

$$1.2,\ 1.8,\ 2.4,\ 2.0,\ 2.9,\ 1.5, \ldots,$$

bisogna decidere in quale classe collocare l'osservazione 2.0; se si decide di assegnarlo alla prima classe, vorrà dire che ciascuna delle due classi è chiusa a destra, nel senso che contiene le modalità della variabile coincidenti con il limite destro della classe; al contrario, se si decide di assegnarlo alla seconda classe, le due classi saranno chiuse a sinistra.

In termini generali, una classe è "chiusa a destra" se include le unità che presentano modalità esattamente uguali all'estremo destro della classe; mentre è "chiusa a sinistra" se include le unità che presentano modalità esattamente uguali all'estremo sinistro. Nel seguito, quando sarà necessario, si adotterà la notazione

matematica utilizzata per indicare la chiusura di un intervallo:[2]

- classe chiusa a sinistra: $[a, b)$;
- classe chiusa a destra: $(a, b]$;
- classe chiusa a sinistra e a destra: $[a, b]$.

Chiaramente, la doppia chiusura, a destra e a sinistra, ha senso solo per la prima e l'ultima classe. Per esempio, nel caso di classi chiuse a destra avrebbe senso una prima classe del tipo [15, 20].

Le k classi in cui è suddivisa la variabile X vengono indicate con:

$$c_0\text{-}c_1, c_1\text{-}c_2, \ldots, c_{k-1}\text{-}c_k,$$

in cui il limite superiore di ogni classe è uguale al limite inferiore della classe successiva. Le frequenze corrispondenti sono indicate con $n_1, n_2, \ldots, n_k$.

DEFINIZIONE 2.3

La distribuzione di frequenza di una variabile suddivisa in classi, denominata per brevità distribuzione di frequenza in classi, è costituita dalle classi con le rispettive frequenze:

Classe	Frequenza
$c_0\text{-}c_1$	n_1
$c_1\text{-}c_2$	n_2
⋮	⋮
$c_{k-1}\text{-}c_k$	n_k
Totale	n

Con il raggruppamento in classi si produce una perdita di informazioni: le osservazioni appartenenti alla singola classe non si conoscono più individualmente, ma se ne conosce solo l'appartenenza alla classe. Così, nell'Esempio 2.4, dalle cinque osservazioni appartenenti alla prima classe, che sono

$$41, 41, 41, 42, 42,$$

si passa all'indicazione, più imprecisa, che 5 unità presentano modalità comprese nell'intervallo 41-42. Tutto ciò si riflette negativamente sulla precisione delle costanti caratteristiche che si usano nell'analisi dei dati. Ad esempio, la media aritmetica di una distribuzione di frequenza in classi non può essere calcolata esattamente: il suo valore esatto si può ottenere solo utilizzando i dati individuali.

Il raggruppamento in classi è una forma di sintesi e di semplificazione dei dati grezzi, che si rende necessaria quando le modalità assunte dalla variabile sono molto numerose.

Se ci si domanda quali debbano essere il numero e l'ampiezza delle classi, va sottolineato che non ci sono risposte generali a queste domande:[3] un criterio guida è che gli intervalli devono essere di numero e ampiezza tali da minimizzare

[2] In alcuni testi di statistica, vengono usati i simboli $a \vdash b$, $a \dashv b$ e $a \vdash\!\dashv b$ per indicare una classe chiusa a sinistra, chiusa a destra e chiusa a sinistra e a destra.

[3] Sono state proposte nel tempo varie regole empiriche per l'individuazione del numero di classi: quella più frequentemente citata è sintetizzata nella cosiddetta formula di Sturges data da

$$C = 1 + \frac{10}{3} \cdot \log(n),$$

dove $\log(n)$ è il logaritmo in base 10 di n.

la perdita di informazione. Considerando che il numero delle classi è in proporzione inversa alle loro ampiezze, il criterio indicato indurrebbe a formare un numero elevato di piccoli intervalli. A tale proposito, va tenuto presente che troppe classi possono ridurre l'efficacia della presentazione dei dati in forma tabellare. Il problema è quindi quello di bilanciare queste esigenze contrastanti. La regola generale di definire classi di uguale ampiezza può essere ignorata se le osservazioni presentano un addensamento differenziato all'interno del campo di variazione della variabile; in tal caso, può essere conveniente formare classi di ampiezza diversa, ossia classi più piccole nella parte dove le osservazioni sono più addensate e classi più ampie altrove. Quando l'insieme di dati contiene modalità estremamente piccole o grandi, è conveniente utilizzare classi aperte (si veda l'Esempio 2.5).

Naturalmente, anche per le distribuzioni di frequenza in classi, è possibile definire la distribuzione di frequenza relativa o percentuale e la distribuzione di frequenza cumulata nelle diverse accezioni (assoluta, relativa o percentuale).

ESEMPIO 2.5

Si consideri la distribuzione di frequenza presentata nella Tabella 2.5 (dati Istat), dove le frequenze sono espresse in migliaia:

Tabella 2.5
Consumatori di farmaci per età.

Età (anni compiuti)	≤ 14	15-34	35-64	≥ 65	Totale
N. di consumatori	1233	2314	7710	7055	18 312

Si individuino le classi reali e si illustri il significato delle classi estreme.

Soluzione

In base a quanto osservato in merito alle classi che indicano intervalli temporali, le classi reali sono:

$$< 15,\ 15\text{-}35,\ 35\text{-}65,\ \geq 65.$$

Quanto alle classi estreme, la prima contiene tutti i soggetti con età minore di 15 anni. L'ultima accoglie tutti i soggetti con età maggiore o uguale a 65 anni. ■

Si fa ricorso al raggruppamento in classi anche nel caso delle variabili discrete che presentano un numero elevato di modalità (si veda l'Esempio 2.6).

ESEMPIO 2.6

Si faccia riferimento all'Esercizio 1.5 e si consideri la serie di osservazioni relative al punteggio ottenuto all'esame (colonna tre della matrice dei dati). Si costruisca la distribuzione di frequenza assumendo le seguenti classi della stessa ampiezza:

$$44\text{-}49, 50\text{-}54, \ldots, 95\text{-}99.$$

Si calcolino le frequenze cumulate (assolute e relative).

Soluzione

La distribuzione di frequenza in classi mostrata nella Tabella 2.6 si ottiene contando le osservazioni che ricadono nella singola classe. La frequenza cumulata di ciascuna classe si calcola aggiungendo alla frequenza della classe le frequenze delle classi precedenti.

Tabella 2.6
Distribuzione di frequenza in classi per la serie di osservazioni considerata.

Voto	Frequenza	Frequenza cumulata	Frequenza rel. cumulata
45-49	1	1	0.016
50-54	0	1	0.016
55-59	4	5	0.078
60-64	3	8	0.125
65-69	5	13	0.203
70-74	3	16	0.250
75-79	6	22	0.344
80-84	11	33	0.516
85-89	9	42	0.656
90-94	17	59	0.922
95-99	5	64	1.000
Totale	64	–	–

L'interpretazione delle frequenze cumulate è immediata: ad esempio, vi sono 5 unità che presentano modalità non superiori a 59, 13 unità con modalità non superiori a 69 e così via. ■

Anche nel caso discreto, si pone il problema della correzione degli estremi di classe: le classi reali si ottengono diminuendo di 0.5 l'estremo sinistro e aumentando della stessa quantità l'estremo destro. Nell'esempio precedente, le classi reali sono 44.5-49.5, 49.5-54.5 e così via.

Vengono ora date le definizioni di valore centrale di classe, ampiezza di classe e densità di frequenza, nozioni che saranno utilizzate prevalentemente nei prossimi tre capitoli.

DEFINIZIONE 2.4

Data una distribuzione di frequenza in classi, i valori centrali di classe si ottengono determinando, per ogni classe, la semisomma degli estremi:

$$\bar{x}_i = \frac{\text{estremo sinistro} + \text{estremo destro}}{2} = \frac{c_{i-1} + c_i}{2},\quad i = 1, 2, \ldots, k.$$

Le ampiezze di classe si ottengono determinando, per ogni classe, la differenza tra estremo destro ed estremo sinistro:

$$d_i = c_i - c_{i-1},\quad i = 1, 2, \ldots, k.$$

Le densità di frequenza si ottengono determinando, per ogni classe, il rapporto tra frequenza e ampiezza della classe:

$$h_i = \frac{\text{frequenza della } i\text{-esima classe}}{\text{ampiezza della } i\text{-esima classe}} = \frac{n_i}{d_i},\quad i = 1, 2, \ldots, k.$$

I valori centrali serviranno per il calcolo delle medie e degli indici di variabilità; le densità di frequenza saranno utili per le rappresentazioni grafiche.

Quando le classi hanno ampiezze diverse, le frequenze non sono comparabili, mentre lo sono le densità di frequenza.

Le densità di frequenza sono particolari rapporti di densità (si veda il Paragrafo 1.11): la frequenza della classe i viene divisa per l'ampiezza di questa, cosicché il numero che ne risulta indica quante sono in media le unità per unità di ampiezza della classe. Tali rapporti possono essere confrontati tra loro, diversamente dalle frequenze, che lo sono solo se le classi hanno la stessa ampiezza.

Ad esempio, in una ipotetica distribuzione di frequenza, dove le classi di età 21-30 e 31-50 hanno frequenze 72 e 106, l'affermazione che la seconda classe presenta una frequenza maggiore della prima non ha senso dal punto di vista statistico, perché non tiene conto del fatto che le classi hanno ampiezza diversa. Ha senso invece dire che la densità di frequenza della prima classe, 7.2, è superiore a quella della seconda, pari a 5.3, ossia che nella prima classe il numero medio di unità statistiche per unità di ampiezza è superiore a quello della seconda.

ESEMPIO 2.7

Si determinino le densità di frequenza per la distribuzione di cui all'Esempio 2.4.

Soluzione

Le densità di frequenza sono collocate nella quarta colonna della Tabella 2.7.

Tabella 2.7
Densità di frequenza per la distribuzione di cui alla Tabella 2.4.

Lunghezza avambraccio (cm)	Frequenza	Ampiezza della classe	Densità di frequenza
40.5-42.5	5	2.0	2.5
42.5-44.5	17	2.0	8.5
44.5-46.5	25	2.0	12.5
46.5-48.5	35	2.0	17.5
48.5-50.5	32	2.0	16.0
50.5-52.5	20	2.0	10.0
52.5-54.5	6	2.0	3.0
Totale	140	–	–

Ciascuna densità di frequenza rappresenta il numero medio di unità per ogni centimetro di ampiezza della classe. Per esempio, la prima, pari a 2.5, indica che, mediamente, all'interno della classe 40.5-42.5 vi sono 2.5 unità per ogni centimetro. ■

Per calcolare alcune costanti caratteristiche, come la mediana delle distribuzioni di frequenza in classi, l'assunzione a cui si fa spesso ricorso è che le unità siano distribuite uniformemente all'interno delle classi. Questa ipotesi ha il significato indicato di seguito. Si consideri la classe c_{i-1}-c_i avente frequenza n_i e ampiezza $d_i = c_i - c_{i-1}$. Si suddivida la classe in n_i intervallini della stessa ampiezza d_i/n_i. Si supponga che il singolo intervallino rappresenti con la sua lunghezza un'unità statistica e con il suo valore centrale la modalità assunta dalla variabile su tale unità. Per distribuzione uniforme delle unità nelle classi si intende questa particolare disposizione delle n_i unità nella classe. La conseguenza di questa assunzione è che, se si prende un sottointervallo di ampiezza δ all'interno della classe i-esima, il numero di unità, $n(\delta;\, i)$, appartenenti ad esso è proporzionale a δ, con coefficiente di proporzionalità pari alla densità di frequenza della classe:

$$n(\delta;\, i) = \frac{n_i}{d_i} \cdot \delta. \tag{2.1}$$

ESEMPIO 2.8

Si riprenda la distribuzione di frequenza dell'Esempio 2.7 e si calcoli il numero di unità appartenenti all'intervallo 42.7-46.9, nell'ipotesi di uniforme distribuzione.

Soluzione

Applicando la (2.1) ai sottointervalli 42.7-44.5 e 46.5-46.9, il numero teorico di osservazioni che rientrano nell'intervallo 42.7-46.9 è dato da

$$\frac{17}{2}(44.5 - 42.7) + 25 + \frac{35}{2}(46.9 - 46.5) = 47.3.$$

■

2.3 Distribuzioni doppie di frequenza

Analogamente a quanto visto nel Paragrafo 2.2, si illustra ora come costruire una distribuzione doppia di frequenza a partire da una serie doppia di osservazioni, (x_1, y_1), (x_2, y_2), …, (x_n, y_n). Con qualche abuso di notazione, si indichino con $x_1, x_2, \ldots, x_s$ le modalità distinte di X e con $y_1, y_2, \ldots, y_t$ le modalità distinte di Y. Si dispongano $x_1, x_2, \ldots, x_s$ su di una colonna e $y_1, y_2, \ldots, y_t$ lungo una riga in modo da avere una tabella con $s \cdot t$ caselle, dove ogni casella corrisponde a una specifica coppia di modalità di X e di Y, (x_i, y_j). Costruire una distribuzione doppia di frequenza vuol dire contare quante volte la coppia (x_i, y_j) è presente nei dati. Il conteggio avviene idealmente esaminando la singola coppia (x_i, y_i) della serie doppia di osservazioni e inserendo una lineetta verticale nell'appropriata casella di una tabella di spoglio doppia (si veda l'Esempio 2.9).

Il procedimento appena descritto continua a valere anche quando una o entrambe le variabili sono quantitative e divise in classi.

ESEMPIO 2.9

Si consideri nuovamente l'insieme di dati in Figura 2.1 e si costruisca la distribuzione doppia di frequenza per le coppie di osservazioni contenute nella seconda e terza colonna riportate qui sotto per comodità.

(A, 12), (A, 15), (A, 15), (A, 15), (A, 13), (A, 13), (A, 5), (C, 20), (A, 12), (A, 6), (C, 21), (A, 11), (B, 15), (A, 14), (A, 13), (C, 22), (A, 14), (A, 11), (A, 10), (C, 23), (A, 7), (A, 8), (C, 19), (C, 22), (A, 10), (A, 10), (C, 19), (C, 20), (B, 16), (A, 11), (A, 14), (A, 11), (C, 24), (A, 9), (B, 16), (C, 19).

Soluzione

Nella tabella doppia di spoglio riportata di seguito, è descritto il processo di assegnazione delle unità statistiche alle caselle, avendo scelto per la variabile peso le classi: 4.5-10.5, 10.5-15.5, 15.5-20.5, 20.5-25.5. Si tratta di assegnare la singola coppia di osservazioni della serie doppia alla casella pertinente inserendo in questa una lineetta verticale (con l'avvertenza di usare una linea diagonale per indicare il quinto caso).

Età	Peso			
	4.5-10.5	10.5-15.5	15.5-20.5	20.5-25.5
A	𝍸 ||	𝍸 𝍸 𝍸 |	-	-
B	-	|||	-	-
C	-	-	𝍸	𝍸

Si perviene poi alla distribuzione doppia di frequenza inserendo in ciascuna casella la frequenza espressa numericamente (Tabella 2.8).

Tabella 2.8
Distribuzione doppia di frequenza per i dati in Figura 2.1 per età e peso.

Età	Peso				Totale
	4.5-10.5	10.5-15.5	15.5-20.5	20.5-25.5	
A	7	16	0	0	23
B	0	3	0	0	3
C	0	0	5	5	10
Totale	7	19	5	5	36

I dati riportati nella Tabella 2.8 sono di immediata interpretazione. Ad esempio, 7 è il numero di roditori con età A (giovane) e peso compreso tra 4.5 e 10.5 g; 3 è il numero di roditori con età B (semi adulto) e peso tra 10.5 e 15.5 g, e così via. I numeri posti nell'ultima riga e nell'ultima colonna della tabella sono, rispettivamente, i totali di colonna e di riga. I totali di colonna sono le frequenze delle classi della variabile peso, mentre i totali di riga sono le frequenze delle modalità assunte dalla variabile età. Ad esempio, 23 è il numero di roditori di età A, mentre 19 è il numero di roditori con peso compreso tra 10.5-15.5 g. ■

Il procedimento appena descritto può essere applicato anche per costruire distribuzioni doppie di frequenza in cui le due variabili sono entrambe qualitative o entrambe quantitative. Quando una o tutte e due le variabili sono quantitative, sorge il problema della formazione delle classi con le regole discusse nel Paragrafo 2.2.1.

La distribuzione doppia di frequenza è anche denominata tabella di contingenza.

ESEMPIO 2.10

Con riferimento alla Tabella 2.9 (Fonte: Istat; frequenze in migliaia), si illustri il significato delle frequenze: 1093, 714, 7055 e 16 779.

Tabella 2.9
Consumatori di farmaci per età e fonte di prescrizione.

Età	Fonte di prescrizione		Totale
	Medico	Autoprescrizione	
≤ 14	1 093	140	1 233
15-34	1 746	568	2 314
35-64	6 996	714	7 710
≥ 65	6 944	111	7 055
Totale	16 779	1533	18 312

Soluzione

Il numero 1093 è la frequenza dei consumatori di farmaci con età fino a 14 anni che si avvalgono della prescrizione medica; 714 è il numero dei consumatori di farmaci con età tra 35 e 64 anni che non si avvalgono della prescrizione medica; 7055 sono i consumatori di farmaci con età di 65 anni e oltre; 16 799 sono i consumatori di farmaci che si avvalgono della prescrizione medica. ■

Per dare generalità a quanto esposto, la distribuzione doppia di frequenza viene rappresentata in simboli nella Tabella 2.10. In essa, n_{ij} è il numero di unità in cui X è uguale a x_i e Y è uguale a y_j. Il simbolo n_{i0} è la somma delle frequenze nella riga corrispondente alla modalità x_i di X e come tale rappresenta la frequenza di x_i. Analogamente, n_{0j} è la somma delle frequenze nella colonna corrispondente alla modalità y_j di Y, ossia la frequenza di y_j.

Tabella 2.10
Distribuzione doppia di frequenza.

Variabile X	Variabile Y						Totale
	y_1	y_2	$\cdots$	y_j	$\cdots$	y_t	
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1t}	n_{10}
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2t}	n_{20}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{it}	n_{i0}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_s	n_{s1}	n_{s2}	$\cdots$	n_{sj}	$\cdots$	n_{st}	n_{s0}
Totale	n_{01}	n_{02}	$\cdots$	n_{0j}	$\cdots$	n_{0t}	n

Quanto detto per le distribuzioni doppie di frequenza può essere esteso al caso di più di due variabili. Ad esempio, se si considerano congiuntamente tutte e tre le colonne dei dati in Figura 2.1, è possibile costruire una distribuzione tripla di frequenza, applicando un processo di conteggio simile a quello descritto in precedenza. Nel prosieguo di questo volume, saranno oggetto di studio solo le distribuzioni di frequenza semplici (cioè di una sola variabile) e le distribuzioni di frequenza doppie, che saranno riprese nel Capitolo 9.

2.4 Serie storiche e serie territoriali

Si considerano ora due tipologie speciali di dati: le serie storiche e le serie territoriali. Una serie storica può essere definita come la sequenza ordinata delle modalità di una variabile quantitativa osservati a intervalli temporali equidistanti. Il termine "tempo" può indicare un intervallo temporale o un preciso istante. Nell'Esempio 2.11, l'anno è un intervallo temporale, mentre nell'Esempio 2.12 l'anno indica un istante (l'inizio dell'anno).

ESEMPIO 2.11

Si osservi la serie storica riportata nella tabella che segue (Fonte: Istat; dati in migliaia):

Tabella 2.11
Nati vivi in Italia dal 2011 al 2018.

Anno	2011	2012	2013	2014	2015	2016	2017	2018
Nati vivi	531	522	498	495	480	473	458	440

■

ESEMPIO 2.12

Nella tabella che segue, è riportata la popolazione residente in Italia all'inizio di ogni anno dal 2012 al 2019 (Fonte: Istat; dati in migliaia):

Tabella 2.12
Popolazione residente in Italia dal 2012 al 2019.

Anno	2012	2013	2014	2015	2016	2017	2018	2019
Pop. residente	59 394	59 685	60 783	60 796	60 666	60 585	60 484	60 360

■

Si ha una serie territoriale quando le modalità della variabile di interesse sono riferite a entità territoriali.

ESEMPIO 2.13

Nella Tabella 2.13 si riportano gli indici di dipendenza degli anziani in Italia, a inizio 2019 (per la definizione dell'indice, si veda il Paragrafo 1.11) per area geografica (Fonte: Istat).

Tabella 2.13
Indice di dipendenza degli anziani in Italia nel 2019 per area geografica.

Area geografica	Indice di dipendenza degli anziani
Nord-Ovest	38.1
Nord-Est	37.0
Centro	37.0
Sud	31.9
Isole	33.6

■

2.5 Sulle imperfezioni nei dati

Indipendentemente dal processo che ha portato alla rilevazione statistica (indagine statistica, esperimento, studio sul campo), i dati raccolti non sono indenni da imperfezioni dovute a svariati possibili fattori. Per facilitare l'esame di questo problema, conviene prima riprendere la nozione di matrice dei dati per definirla in termini più generali di quanto fatto fin qui.

La nozione è stata introdotta nel Paragrafo 1.9. Se ne è fatto uso poi nell'esempio iniziale del paragrafo introduttivo di questo capitolo e, successivamente, in modo implicito, quando si è affermato che una serie semplice di osservazioni può essere riguardata come una singola colonna della matrice dei dati, una serie doppia o tripla di osservazioni come una coppia o una terna di colonne della medesima matrice.

Si supponga di aver osservato p variabili, $X_1, X_2, \ldots, X_p$, su un insieme di n unità statistiche. La rappresentazione in simboli della matrice dei dati è lo schema, costituito da n righe e p colonne, che raccoglie le $n \times p$ osservazioni:

$$\begin{array}{cccccc} x_{11} & x_{12} & \ldots & x_{1l} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2l} & \ldots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \ldots & x_{il} & \ldots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nl} & \ldots & x_{np} \end{array}$$

Dunque, poiché nelle righe si rappresentano le n unità statistiche e nelle colonne le p variabili osservate, la matrice dei dati ha dimensione $n \times p$ e i suoi elementi sono caratterizzati come segue.

- Il generico elemento x_{il} ($i = 1, 2, \ldots, n$, $l = 1, 2, \ldots, p$) rappresenta la modalità della variabile l-esima osservata sull'unità i-esima.
- Gli elementi che si trovano nella generica riga i-esima rappresentano le modalità delle p variabili osservate sulla stessa i-esima unità; l'intera riga rappresenta quindi l'intero corredo informativo relativo all'i-esima unità. Nel caso, per esempio, di un questionario con p domande (variabili) somministrato a n individui, la riga considerata rappresenta tutte le risposte alle diverse p domande date dallo stesso i-esimo individuo intervistato.
- Gli elementi che si trovano nella generica colonna l-esima rappresentano le modalità della variabile l-esima osservate su tutte le n unità. Nel caso del questionario di cui sopra, la colonna considerata rappresenta tutte le risposte alla stessa l-esima domanda date dagli n individui intervistati.

Prima di passare alle analisi statistiche, la matrice dei dati va sottoposta a un controllo accurato per accertare le possibili imperfezioni e, se possibile, per sanarle.

Dati mancanti

Si distinguono due tipi di dati mancanti o mancate risposte: mancate risposte totali e mancate risposte parziali. Le prime si hanno quando sono assenti tutte le informazioni riguardanti una o più unità: nella matrice dei dati vi sono quindi una o più righe in meno rispetto al numero di unità statistiche rilevate, a causa della mancata partecipazione all'indagine di una o più unità.

Quello delle mancate risposte totali è un problema comune a tutte le rilevazioni statistiche, problema che non si riesce quasi mai a scongiurare del tutto. La situazione più delicata si presenta quando si ha motivo di ritenere che a non rispondere siano gli individui con caratteristiche particolari. In questo caso, il rischio è che l'insieme dei rispondenti non sia "rappresentativo" della popolazione di riferimento, con ricadute negative sulla qualità dei risultati dell'indagine.

Si supponga, ad esempio, che in un'ipotetica indagine campionaria sui redditi delle famiglie si sospetti che le mancate risposte riguardino prevalentemente le famiglie più abbienti e, nonostante ciò, si stimi il reddito medio con la media aritmetica semplice dei redditi dei rispondenti. In questo caso, si rischia di sottostimare fortemente il reddito medio della popolazione. Per queste situazioni, la metodologia statistica mette a disposizione metodi e tecniche per fronteggiare il fenomeno o per attenuarne gli effetti.

L'idea guida è quella di attribuire maggiore importanza, di "far contare di più", nel calcolo della media, i rispondenti che appartengono alle categorie più

coinvolte nelle mancate risposte. Tornando all'esempio, nel calcolare la media dei redditi rilevati nel campione, occorre assegnare pesi più elevati ai redditi dei più abbienti e pesi più bassi agli altri. Per fare ciò è necessario disporre di informazioni supplementari sulla popolazione. Nell'esempio, si ha bisogno di sapere quante sono le famiglie che si possono considerare abbienti, cioè che superano un prefissato livello di reddito familiare.

Si parla di mancate risposte parziali quando uno o più intervistati non rispondono a qualche domanda. In questo caso, i dati mancanti non si polarizzano su una o più righe (unità), ma sono sparpagliati nella matrice dei dati. Per far fronte a questo problema, occorre capire qual è il meccanismo alla base delle mancate risposte. In altri termini, la strategia da adottare dipende dal legame associativo tra le mancate risposte e la variabile studiata. Se questo legame non esiste (come nel caso in cui l'intervistato affidasse la decisione di rispondere alla singola domanda al lancio di una moneta), l'omissione delle unità non rispondenti non pregiudica la qualità dell'analisi statistica. All'opposto, quanto più questo legame associativo è forte tanto maggiore è il rischio che i risultati dell'indagine deviino dal vero.

È evidente che nelle indagini complesse – tipicamente quelle condotte dagli Istituti nazionali di statistica – il numero delle variabili rilevate è in genere alto e il meccanismo sotteso alle mancate risposte può variare da variabile a variabile, cosicché sarebbe oltremodo complicato adottare criteri di analisi differenti per le diverse variabili. Si usano pertanto criteri di "pulitura" dei dati standard che restituiscono una matrice senza lacune. Ciò viene realizzato tramite tecniche di assegnazione del dato alle caselle vuote della matrice, che qui non è possibile illustrare: ci si limita solo a precisare che il meccanismo può essere casuale (all'interno di regole di coerenza, che inibiscano, ad esempio, che venga assegnato un dato incompatibile con le risposte date a una o più altre variabili), oppure basato su criteri di similarità, come quello di assegnare alla casella vuota il dato del rispondente il cui profilo è più simile a quello dell'unità statistica considerata.

Dati anomali ed errori di misura

Se si fa riferimento a una colonna della matrice dei dati (variabile), i dati anomali sono rappresentati da quelle modalità che differiscono sensibilmente dalle restanti o *outlier*. Essi possono essere generati da un'elevata variabilità intrinseca del fenomeno a cui la variabile si riferisce o da errori di rilevazione o di registrazione dei dati. Sono disponibili tecniche esplorative per individuare i dati anomali, che operano su singole colonne o su due o più colonne della matrice dei dati.

Per quanto riguarda il trattamento dei dati anomali, bisogna distinguere due casi. Se il dato anomalo è dovuto a un errore di rilevazione non correggibile, è preferibile eliminarlo e trattarlo come dato mancante. Se viceversa è dovuto alla variabilità intrinseca della variabile, è preferibile conservarlo, effettuando eventualmente un'appropriata trasformazione (ad esempio, logaritmica) per attenuarne l'effetto sull'elaborazione dei dati.

Gli errori di misura sono definiti come le differenze tra le modalità reali assunte dalle variabili e quelle rilevate sulla singola unità statistica. Gli errori di misura possono essere attribuiti al rispondente o al rilevatore (ma non è escluso che possano originarsi durante il trattamento e l'elaborazione dei dati). Nel primo caso, possono essere causati da: risposte non corrette per negligenza, erronea interpretazione del questionario, memoria imprecisa di eventi passati, risposte mendaci ecc. Gli errori di misura possono non palesarsi (nel senso che non sono riconoscibili) o produrre dati inammissibili (ad esempio, esterni al campo di variazione della variabile) o dati anomali. I dati inammissibili vengono cancellati e trattati come mancanti.

I dati mancanti e gli errori di misura sono qualificati come errori non campionari, per differenziarli dai cosiddetti errori campionari, che peraltro si riferiscono non alle singole unità statistiche, ma all'intero campione: si parla di errore campionario infatti per intendere la differenza tra la stima di una grandezza (ad esempio, il reddito medio delle famiglie di un Paese) ottenuta con i dati di un campione e il valore reale della grandezza.

Riassunto

Questo capitolo è dedicato principalmente alla descrizione del procedimento che consente di organizzare e riassumere i dati mediante tabelle, molto utili ai fini della presentazione e dell'interpretazione dei dati. Viene illustrato come si costruiscono le distribuzioni di frequenza a partire da serie di osservazioni di variabili qualitative o quantitative. Viene poi introdotta la distribuzione di frequenza cumulata per le variabili a modalità ordinabili.

Nel caso di variabili continue o discrete, ma con un numero elevato di modalità, la costruzione delle distribuzioni di frequenza richiede che siano preliminarmente definite le classi, costituite da insiemi di modalità vicine.

Viene pure illustrato il procedimento con cui si passa da serie doppie di osservazioni a distribuzioni doppie di frequenza.

Infine, viene esaminato in breve il problema delle imperfezioni nei dati, dovute alle mancate risposte, ai valori anomali e agli errori di misura, facendo cenno ai possibili rimedi con cui si può neutralizzare o attenuare il loro effetto sui risultati delle analisi statistiche.

Parole chiave

ampiezze di classe, 51
analisi statistica bivariata, 42
analisi statistica univariata, 42
chiusura delle classi, 48
classi reali, 48
dati anomali, 58
dati mancanti, 57
densità di frequenza, 51
distribuzioni di frequenza, 44
distribuzioni di frequenza cumulata, 45
distribuzioni di frequenza in classi, 49
distribuzioni di frequenza percentuale, 45
distribuzioni di frequenza relativa, 45
distribuzioni doppie di frequenza, 53
frequenze, 44
frequenze cumulate, 45
frequenze cumulate percentuali, 46
frequenze percentuali, 44
frequenze relative, 44
frequenze relative cumulate, 46
errori campionari, 59
errori di misura, 58
errori non campionari, 59
intervalli di tolleranza, 48
mancate risposte parziali, 57
mancate risposte totali, 57
perdita di informazione, 49
proporzioni, 44
serie di osservazioni, 42
serie doppie di osservazioni, 42
serie multiple di osservazioni, 42
serie semplici di osservazioni, 42
serie storiche, 55
serie territoriali, 56
serie triple di osservazioni, 42
tabella di contingenza, 54
tabella di frequenze, 44
tabella di spoglio, 43
uniforme distribuzione delle unità nelle classi, 52
valori centrali di classe, 51

Commenti e curiosità

1. Soldati scozzesi per circonferenza toracica

Nella tabella che segue si riporta la distribuzione di frequenza della circonferenza toracica di 5732 soldati scozzesi. Questi dati sono legati al nome dello statistico belga Quetelet (1796-1874) che li utilizzò nell'ambito dello studio dell'adattamento di modelli statistici ai dati reali (Hand *et al.*, 1994, pp. 123-124).

Misura (in pollici)	Fequenza	Misura (in pollici)	Frequenza
33	3	41	935
34	19	42	646
35	81	43	313
36	189	44	168
37	409	45	50
38	753	46	18
39	1062	47	3
40	1082	48	1

2. Altezza di genitori e figli

Nella tabella che segue si fa cenno ai dati sulle stature dei figli e dei genitori utilizzati da Galton (1822-1911) nello studio sull'ereditarietà dei caratteri antropometrici (il *file* completo in formato Excel si trova nel sito del volume sotto il nome di "Stature"). Si tratta di 928 coppie di dati espressi in pollici: accanto alla statura di ogni ragazzo(a), posto nella prima colonna, è indicata la media delle stature dei genitori, posta nella seconda colonna. Per ragioni di omogeneità, sia le stature delle ragazze che quelle delle rispettive madri risultano moltiplicate per 1.08.

Statura figlio/a	Statura media dei genitori
61.7	70.5
61.7	68.5
⋮	⋮
73.7	69.5

3. Sepali e petali di iris

Qui di seguito, si richiamano i dati sulla lunghezza e larghezza dei sepali e dei petali di tre specie di iris che furono utilizzati da R.A. Fisher (1890-1962) per l'introduzione di una metodologia statistica per la discriminazione tra specie (si veda Hand *et al.*, 1994, pp. 25-26). I dati completi sono disponibili in formato Excel nel sito del volume (*file* "Iris").

Iris setosa				Iris versicolor				Iris virginica			
Lunghezza sepali	Larghezza sepali	Lunghezza petali	Larghezza petali	Lunghezza sepali	Larghezza sepali	Lunghezza petali	Larghezza petali	Lunghezza sepali	Larghezza sepali	Lunghezza petali	Larghezza petali
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

4. Morti per calcio di cavallo

Nella tabella sottostante sono riportati dati molto particolari: si riferiscono ai decessi di soldati prussiani causati da calcio di cavallo, osservati in 14 corpi d'armata durante 20 anni. Questi dati sono entrati nella storia della statistica perché sono associati da Bortkiewicz (1868-1931) a una particolare distribuzione di probabilità, la cosiddetta "legge degli eventi rari" che sarà illustrata nel Capitolo 14.

Anno	Corpo d'armata														Totale
	G	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XIV	XV	
1875	0	0	0	0	0	0	0	1	1	0	0	0	1	0	3
1876	2	0	0	0	1	0	0	0	0	0	0	0	1	1	5
1877	2	0	0	0	0	0	1	1	0	0	1	0	2	0	7
1878	1	2	2	1	1	0	0	0	0	0	1	0	1	0	9
1879	0	0	0	1	1	2	2	0	1	0	0	2	1	0	10
1880	0	3	2	1	1	1	0	0	0	2	1	4	3	0	18
1881	1	0	0	2	1	0	0	1	0	1	0	0	0	0	6
1882	1	2	0	0	0	0	1	0	1	1	2	1	4	1	14
1883	0	0	1	2	0	1	2	1	0	1	0	3	0	0	11
1884	3	0	1	0	0	0	0	1	0	0	2	0	1	1	9
1885	0	0	0	0	0	0	1	0	0	2	0	1	0	1	5
1886	2	1	0	0	1	1	1	0	0	1	0	1	3	0	11
1887	1	1	2	1	0	0	3	2	1	1	0	1	2	0	15
1888	0	1	1	0	0	1	1	0	0	0	0	1	1	0	6
1889	0	0	1	1	0	1	1	0	0	1	2	2	0	2	11
1890	1	2	0	2	0	1	1	2	0	2	1	1	2	2	17
1891	0	0	0	1	1	1	0	1	1	0	3	3	1	0	12
1892	1	3	2	0	1	1	3	0	1	1	0	1	1	0	15
1893	0	1	0	0	0	1	0	2	0	0	1	3	0	0	8
1894	1	0	0	0	0	0	0	0	1	0	1	1	0	0	4
Totale	16	16	12	12	8	11	17	12	7	13	15	25	24	8	196

Esercizi

2.1 Si consideri il *file* "Corso di statistica" descritto nell'Esercizio 1.5. Si consideri la variabile "Voto", la cui serie di osservazioni è riprodotta qui sotto per comodità:

B, A, B, D, B, B, A, B, B, C, B, B, C, A, C, B, F, A, B, C, B, B, C, B, A, A, B, A, B, B, D, B,
A, B, B, C, B, D, C, C, C, A, A, B, D, A, B, B, C, F, A, B, C, C, A, A, C, D, F, B, A, A, B, A

a) Si costruisca la distribuzione di frequenza.

b) Si determinino le frequenze percentuali e si commentino i risultati.

2.2 Le seguenti osservazioni si riferiscono al numero di tabelle statistiche contenute nelle pagine che vanno dalla 213 alla 232 di un manuale di statistica:

1, 2, 1, 0, 4, 5, 4, 1, 1, 0, 1, 1, 0, 0, 0, 0, 4, 2, 2, 3

a) Si costruisca la distribuzione di frequenza.

b) Si determini la distribuzione di frequenza cumulata (assoluta e relativa).

2.3 Si considerino le osservazioni nella prima colonna del *file* "Iris" riguardanti la lunghezza dei sepali dell'iris setosa.

a) Si costruisca la distribuzione di frequenza in classi, adottando le seguenti classi di uguale ampiezza: 4.25-4.45, 4.45-4.65, . . . , 5.65-5.85.

b) Si determini la distribuzione di frequenza cumulata (assoluta e relativa).

2.4 Si consideri la distribuzione di frequenza relativa delle famiglie italiane per numero di componenti nel 2017 (Fonte: Istat):

Numero di componenti	1	2	3	4	5	≥ 6	Totale
Famiglie (freq. rel.)	0.319	0.275	0.196	0.157	0.041	0.012	1.000

a) Si determini la distribuzione di frequenza relativa cumulata.

b) Si determini la distribuzione di frequenza assoluta (con le frequenze in migliaia), sapendo che il numero totale di famiglie era 25 865 (in migliaia).

2.5 Si considerino le prime 38 osservazioni della variabile peso nel *file* di dati dell'Esercizio 1.6.

a) Si costruisca la distribuzione di frequenza usando le classi: fino a 70, 71-80, 81-90, 91 e oltre.

b) Si calcolino le frequenze relative cumulate e le si usino per descrivere la distribuzione di frequenza.

2.6 Coppie che hanno fatto richiesta di adozione per durata del matrimonio (Fonte: Istat):

Durata del matrimonio	0-2	3-5	6-10	11-15	16-20	≥ 20	Totale
N. di coppie	386	1800	2844	1573	569	315	7487

a) Quali sono le classi reali?

b) Si descriva il significato del numero 1800.

c) Si calcolino le frequenze percentuali.

d) Si calcolino le densità di frequenza (per l'ultima classe si assuma come estremo destro 26 anni).

e) Si calcoli il numero approssimato di coppie con una durata del matrimonio compresa tra 7 e 18 anni, assumendo l'uniforme distribuzione delle unità nelle classi.

2.7 Distribuzione di frequenza delle precipitazioni giornaliere (mm) registrate in un anno in una stazione meteorologica:

Precipitazioni giornaliere (mm)	0-1	1-10	10-30	30-100	Totale
Frequenza (giorni)	36	55	24	5	120

a) Si calcoli la percentuale di giornate piovose con precipitazioni fino a 10 mm.

b) Assumendo l'uniforme distribuzione delle unità nelle classi, si calcoli la percentuale di giornate piovose con precipitazioni comprese tra 5 e 25 mm.

c) Con riferimento all'intero anno, si determini la percentuale di giornate senza pioggia.

2.8 Si costruisca la distribuzione di frequenza in classi dei rendimenti giornalieri delle azioni Campari dall'1 settembre al 31 dicembre 2019, utilizzando le classi indicate di seguito: $(-3,-2], (-2,-1], (-1,0], (0,1], (1,2], (2,3]$. (*Suggerimento*: si utilizzino i dati contenuti nel *file* "Quotazioni" per calcolare i rendimenti giornalieri).

2.9 Si considerino di nuovo i dati dell'Esercizio 1.5 e si costruisca la distribuzione doppia di frequenza per le variabili esame e voto. Per la prima variabile si utilizzino le seguenti classi della stessa ampiezza: 45-49, 50-54, . . . , 95-99.

2.10 Si consideri la distribuzione dei comuni per ampiezza demografica ricavata dai dati riportati nell'Esercizio 1.3:

Ampiezza demografica (in migliaia)	Numero di comuni
≤ 0.5	4
0.5-1	6
1-2	24
2-3	11
3-4	10
4-5	5
5-10	13
10-15	3
15-20	5
20-30	5
30-40	2
40-50	1
50-65	1
≥ 65	2
Totale	92

Si usi la (2.1) per calcolare il numero di comuni con popolazione compresa tra 2500 e 17 000.

2.11 Si consideri la matrice dei dati dell'Esercizio 1.6, limitatamente agli individui di età minore o uguale a 35 anni. Si costruisca la distribuzione doppia di frequenza per le variabili circonferenza del torace e circonferenza dell'addome. Si adottino le classi che seguono: fino a 90, 91-100, 101-110, oltre 110, per la prima variabile; fino a 80, 81-90, 91-100, oltre 100, per la seconda.

2.12 Colore degli occhi e dei capelli di 5387 bambini nel Caithness, Scozia (si veda Hand *et al.*, 1994, p. 146):

Colore degli occhi	Colore dei capelli				
	Chiaro	Rosso	Bruno	Scuro	Nero
Azzurro	326	38	241	110	3
Chiaro	688	116	584	188	4
Bruno	343	84	909	412	26
Scuro	98	48	403	681	85

a) Si confronti la distribuzione del colore dei capelli dei bambini con gli occhi azzurri con quella dei bambini con gli occhi scuri (si usino le frequenze percentuali).

b) Si calcoli la percentuale di bambini con i capelli rossi tra quelli che hanno gli occhi chiari.

c) Si calcoli la percentuale di bambini con occhi azzurri o bruni tra quelli che hanno i capelli chiari.

d) Si calcoli la percentuale di bambini con gli occhi azzurri.

2.13 La tabella seguente mostra la distribuzione dei comuni italiani per area geografica e ampiezza demografica nel 2018 (Fonte: Istat):

Area geografica	Ampiezza demografica						Totale
	≤ 2000	2001-5000	5001-20 000	20 001-50 000	50 001-100 000	≥ 100 001	
Nord-Ovest	1609	711	536	92	18	7	2973
Nord-Est	404	413	480	60	8	16	1381
Centro	346	247	267	77	22	8	967
Sud	776	444	418	99	38	8	1783
Isole	306	215	176	51	13	6	767
Totale	3441	2030	1877	379	99	45	7871

a) Si sostituiscano alle frequenze di riga le corrispondenti frequenze percentuali (ponendo uguale a 100 il totale della singola riga).

b) Sulla base delle elaborazioni di cui al punto precedente, si confrontino le distribuzioni secondo l'ampiezza demografica delle aree con quella relativa all'Italia.

c) In quale regione si registra la maggiore incidenza dei comuni piccoli (fino a 2000 abitanti)?

2.14 Si consideri la seguente distribuzione doppia:

Lunghezza medio (mm)	Lunghezza piede (mm)							Totale
	A	B	C	D	E	F	G	
95-99	8	7	0	0	0	0	0	15
100-104	20	96	71	20	0	0	0	207
105-109	13	106	395	253	40	0	0	807
110-114	0	28	306	693	332	55	0	1414
115-119	0	0	41	282	474	197	24	1018
120-124	0	0	0	30	155	158	61	404
125-129	0	0	0	0	0	32	42	74
Totale	41	237	813	1278	1001	442	127	3939

Legenda: A: 220-229, B: 230-239, C: 240-249, D: 250-259, E: 260-269, F: 270-279, G: 280-289.

a) Come si interpretano i numeri che si trovano nell'ultima colonna della tabella? E quelli che si trovano nell'ultima riga?

b) Che cosa suggerisce la disposizione delle frequenze interne alla tabella?

c) Si determini la frequenza percentuale delle unità in cui la lunghezza del medio è tra 105 e 114 e quella del piede è maggiore o uguale a 260.

d) È possibile determinare il numero delle unità in cui la lunghezza del medio è minore o uguale a 107? Si motivi la risposta.

Distribuzioni di frequenza con Excel

Da una serie di osservazioni secondo una sola variabile, è possibile passare alla distribuzione di frequenza utilizzando la procedura "Tabella pivot". Qui di seguito è illustrata la sequenza delle operazioni da compiere utilizzando i dati della Figura 2.1.

- Si inseriscono i dati nel foglio di Excel e si sceglie la voce "Tabella pivot" nel menu "Inserisci" (si veda la Figura 2.3).

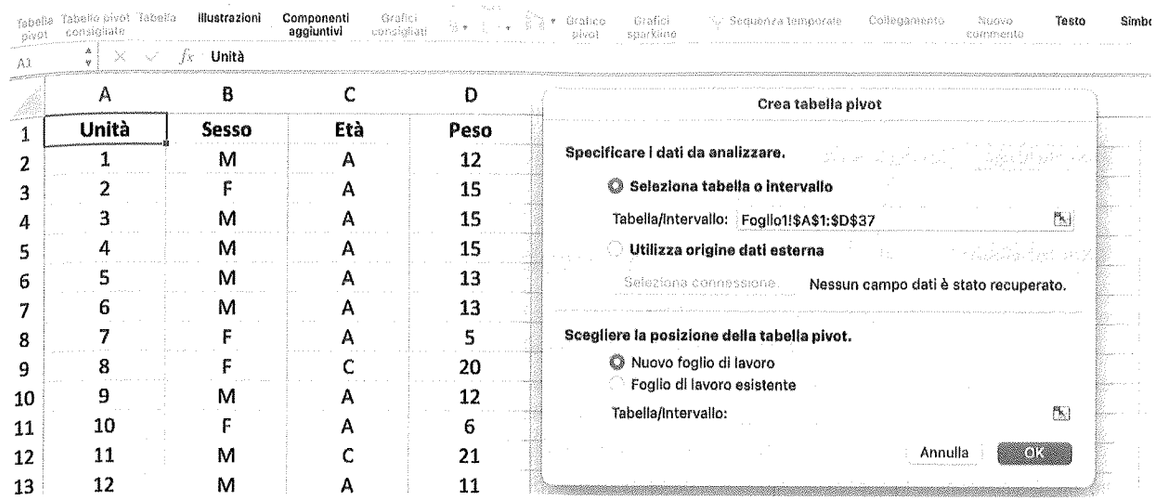


Figure 2.3 Avvio della procedura tabella pivot.

- Nella finestra di dialogo si seleziona l'area del foglio che contiene i dati includendo il nome delle variabili e si specifica dove deve essere collocato il rapporto di tabella pivot (la posizione predefinita per una nuova tabella pivot è un nuovo foglio di lavoro). Selezionato il tasto OK, si visualizza il riquadro "Campi tabella pivot" (si veda la Figura 2.4).

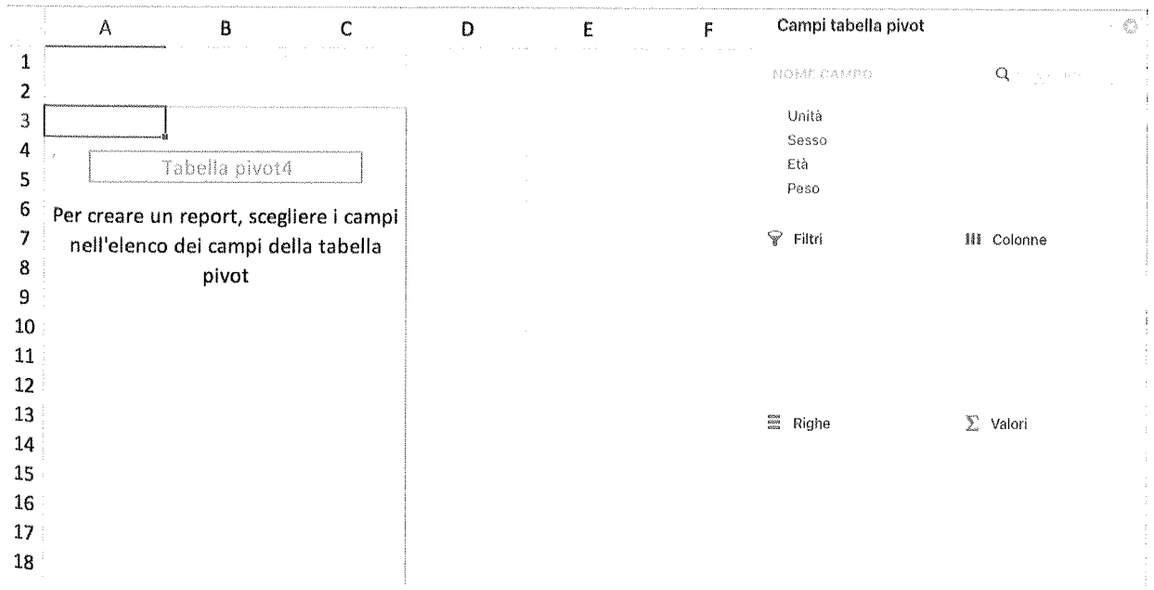


Figure 2.4 Riquadro "Campi" della procedura tabella pivot.

- Si trascina il nome del campo che interessa (età, nell'esempio) nell'area "Righe" o "Colonne" a seconda che si voglia che la tabella sia disposta in verticale o in orizzontale. Lo stesso nome va poi trascinato nell'area "Valori". Se in questa appare la dizione "Conteggio", la procedura è terminata; altrimenti, va aperto il menu a tendina, selezionando "Conteggio" (si veda la Figura 2.5).

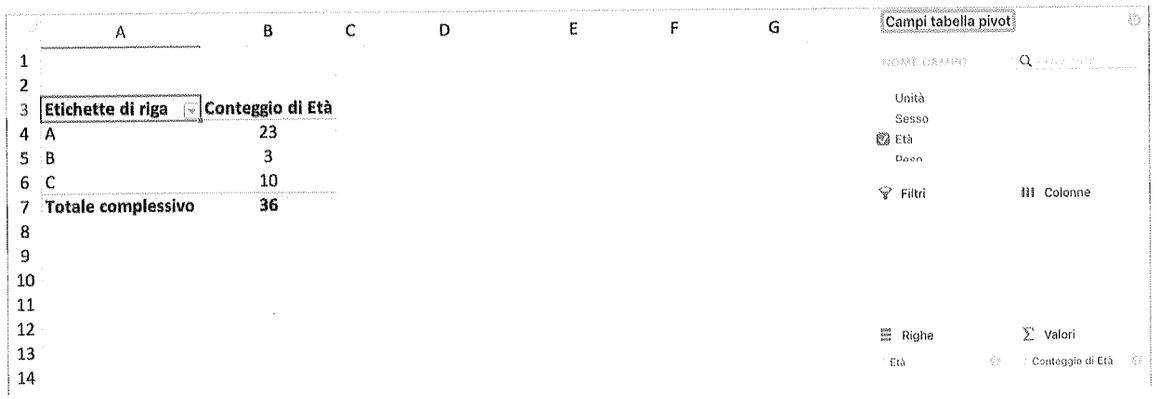


Figure 2.5 Distribuzione di frequenza ottenuta con la procedura tabella pivot.

Quando si vogliono raggruppare i dati in classi, la via più conveniente è quella di ricodificare i dati, per esempio con la funzione SE( ), assegnando ai termini della prima classe il valore 1, a quelli della seconda classe il valore 2 e così via. Poi, si effettua la procedura sui dati ricodificati per ottenere la distribuzione di frequenza desiderata. Per migliorarne la leggibilità, si possono modificare le caselle della tabella contenente i codici sostituendo a questi le classi corrispondenti.

La procedura tabella pivot consente anche di costruire distribuzioni doppie di frequenza. A questo proposito, è necessario specificare l'area del foglio contenente i dati, includendo le caselle che contengono i nomi delle variabili. Successivamente, nel riquadro "Campi della tabella pivot", compaiono i pulsanti con i nomi delle due variabili di interesse (Età e Sesso, nell'esempio): uno di essi va trascinato nell'area "Righe" e l'altro nell'area "Colonne". Va ricordato che nell'area "Valori" si deve leggere "Conteggio".

La distribuzione doppia di frequenza ottenuta mediante la procedura appena descritta è mostrata nella Figura 2.6.

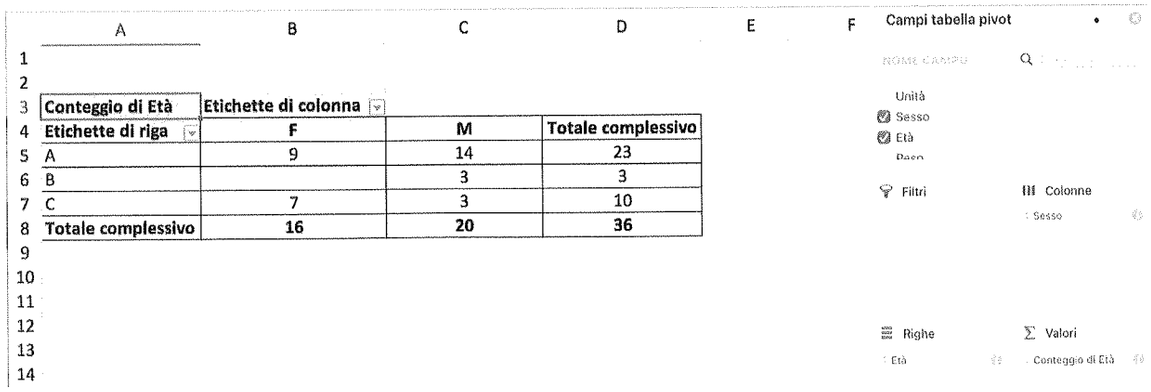


Figure 2.6 Distribuzione doppia di frequenza ottenuta con la procedura tabella pivot.

CAPITOLO

3

Rappresentazioni grafiche

Sommario del capitolo

3.1 Introduzione
3.2 Distribuzioni semplici di frequenza
3.3 Distribuzioni doppie di frequenza
3.4 Serie storiche
3.5 Serie territoriali
3.6 Problemi di scala
Riassunto
Commenti e curiosità
Esercizi

3.1 Introduzione

La forza dei grafici sta nella loro capacità di rappresentare i dati in modo sintetico ed efficace. In particolare, essi:

- rendono visibili le caratteristiche essenziali dei dati – strutture, tendenze, relazioni – facilitando l'interpretazione dei fenomeni;
- semplificano la complessità dei dati;
- consentono il confronto fra due o più insiemi di dati.

In questo capitolo, si concentrerà l'attenzione sulla rappresentazione grafica delle distribuzioni di frequenza, semplici e doppie, delle serie storiche e delle serie territoriali. Si tratterà solo di una guida alla scelta del grafico più appropriato nelle situazioni concrete. È chiaro che la realizzazione dei grafici dovrà poi avvalersi di un programma di calcolo statistico.

3.2 Distribuzioni semplici di frequenza

Gli argomenti di questo paragrafo sono suddivisi in tre parti distinte per tipo di variabile. La prima è dedicata alle distribuzioni di frequenza di variabili discrete a modalità singole; la seconda alle distribuzioni di frequenza in classi, tipicamente di variabili continue; la terza alle distribuzioni di frequenza di variabili qualitative.

3.2.1 Distribuzioni di frequenza di variabili discrete

Si consideri la distribuzione di frequenza di una variabile discreta:

Modalità	x_1	x_2	...	x_k	Totale
Frequenza	n_1	n_2	...	n_k	n

dove $x_1, x_2, \ldots, x_k$ sono le modalità della variabile e $n_1, n_2, \ldots, n_k$ le rispettive frequenze.

Grafico a punti

Se il numero n delle unità non è grande, una rappresentazione semplice è il grafico a punti, dove la modalità x_i viene rappresentata con un punto (riferito a una scala graduata) ripetuto tante volte quant'è la sua frequenza n_i; i punti vengono disposti in colonna, opportunamente spaziati fra loro.

La Figura 3.1 mostra il grafico a punti della distribuzione di frequenza della variabile esame, dedotta dal *file* "Corso di statistica" descritto nell'Esercizio 1.5.

Figura 3.1
Grafico a punti della distribuzione considerata.

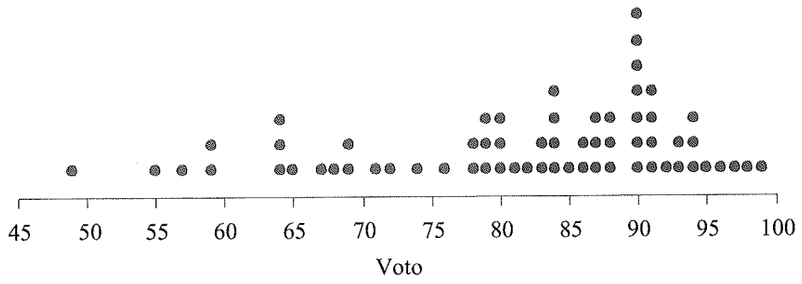


Grafico della funzione di frequenza

Per descrivere la relazione tra frequenza e modalità in una distribuzione di frequenza di una variabile discreta, si può definire la funzione di frequenza come segue:

$$n(x) = \begin{cases} n_i, & \text{se } x = x_i,\ i = 1, 2, \ldots, k, \\ 0, & \text{altrove.} \end{cases}$$

La rappresentazione grafica di questa funzione consiste nel riportare i valori della variabile sull'asse orizzontale, le frequenze sull'asse verticale e nel disegnare per ogni x_i una linea verticale di altezza n_i che collega i punti di coordinate $(x_i, 0)$ e (x_i, n_i). Si ottiene così il grafico ad aste.

ESEMPIO 3.1

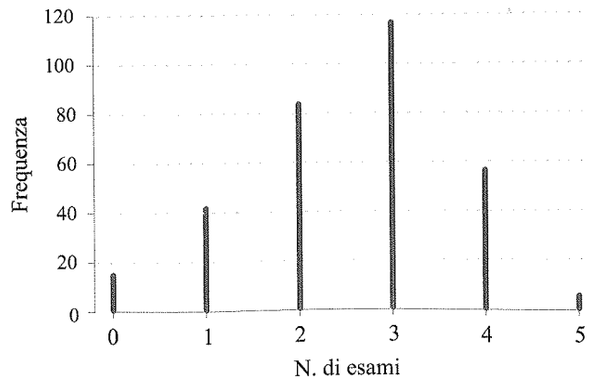
Si consideri la distribuzione, riportata in Tabella 3.1, del numero di esami superati nel primo anno di corso da un campione di studenti universitari di Economia:

Tabella 3.1
Distribuzione del numero di esami superati nel primo anno di corso da un campione di studenti di Economia.

N. di esami	0	1	2	3	4	5	Totale
Frequenza	14	41	83	116	56	5	315

La Figura 3.2 mostra il grafico ad aste relativo alla distribuzione considerata.

Figura 3.2
Grafico ad aste della distribuzione di cui alla Tabella 3.1.



In modo equivalente, si può rappresentare la funzione di frequenza relativa:

$$f(x) = \begin{cases} f_i, & \text{se } x = x_i,\ i = 1, 2, \ldots, k, \\ 0, & \text{altrove,} \end{cases}$$

essendo f_i la frequenza relativa della modalità x_i. La sostituzione di n_i con f_i non influisce sulla forma del grafico.

Grafico a nastri

Qualche volta, le linee verticali associate alle modalità della distribuzione vengono sostituite con nastri di altezze (o larghezze) corrispondenti alle frequenze; si ottiene così un grafico a nastri. Questo tipo di grafico viene illustrato nel prossimo esempio, dove tre distribuzioni di frequenza aventi le stesse modalità sono rappresentate in un unico grafico per consentirne il confronto.

ESEMPIO 3.2

La seguente tabella contiene la distribuzione percentuale delle famiglie per dimensione in Italia, Francia e Germania nel 2018 (Fonte: Eurostat).

Tabella 3.2
Distribuzione delle famiglie per dimensione in Italia, Francia e Germania nel 2018.

Dimensione	Italia	Francia	Germania
1	32.6	35.9	41.5
2	27.9	32.8	34.0
3	19.4	13.3	12.0
4	15.8	12.6	9,1
5 o più	4.3	5.4	3.4
Totale	100.0	100.0	100.0

Queste distribuzioni sono rappresentate in Figura 3.3.

Figura 3.3
Grafico a nastri multiplo per le distribuzioni di frequenza della Tabella 3.2.

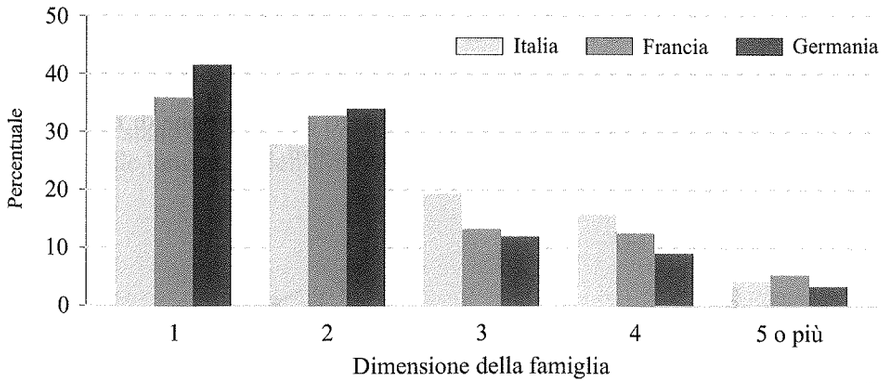


Grafico della funzione di ripartizione

DEFINIZIONE 3.1

Data una distribuzione di frequenza della variabile discreta X, la funzione di ripartizione, $F(x)$, è la funzione che associa a ogni numero reale x la frequenza relativa delle unità in cui $X \leq x$.

Per chiarire la definizione, si consideri la distribuzione di frequenza ipotetica del numero di figli in un campione di 50 famiglie:

N. di figli	0	1	2
N. di famiglie	12	24	14
F_i	0.24	0.72	1

dove F_i è la frequenza relativa cumulata per $x = x_i$ ($x_1 = 0, x_2 = 1, x_3 = 2$).

La funzione di ripartizione è definita per qualsiasi numero reale. Ad esempio, nei punti

$$x = -1.5,\ 0.7,\ 1.2,\ 2,\ 3.6,$$

si ha:

$$F(-1.5) = 0,\ F(0.7) = 0.24,\ F(1) = 0.72,\ F(1.2) = 0.72,\ F(2) = 1,\ F(3.6) = 1.$$

Ciò perché: la frequenza relativa delle famiglie con un numero di figli minore o uguale a -1.5 è 0; la frequenza relativa delle famiglie con un numero di figli minore o uguale a 0.7 è 0.24 e così via.

In generale, si può scrivere:

$$F(x) = \begin{cases} 0, & \text{se } x < x_1, \\ F_i, & \text{se } x_i \le x < x_{i+1},\ i = 1, 2, \ldots, k-1, \\ 1, & \text{se } x \ge x_k. \end{cases}$$

$F(x)$ è uguale a 0 per $x < x_1$, perché 0 è la frequenza delle unità che presentano una modalità minore di x_1 (essendo x_1 la modalità più piccola); $F(x) = 1$ per $x \ge x_k$, perché tutte le unità presentano una modalità minore o uguale a x_k (essendo x_k la modalità più grande); $F(x) = F_i$ per $x_i \le x < x_{i+1}$, perché non ci sono unità che presentano modalità maggiore di x_i e minore di x_{i+1}. Chiaramente, $F(x)$ è una funzione costante a tratti che presenta dei "salti" pari a $f_1, f_2, \ldots, f_k$ nei punti $x_1, x_2, \ldots, x_k$.

Nella Figura 3.4, viene mostrato l'andamento della funzione di ripartizione in corrispondenza di tre generiche modalità successive.

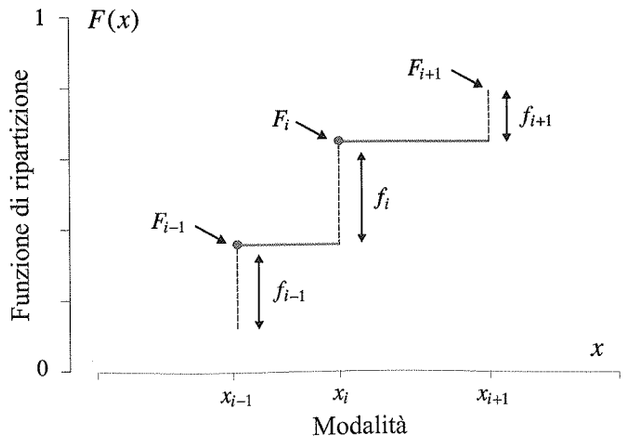


Figura 3.4
Funzione di ripartizione per una distribuzione di frequenza di una variabile discreta.

ESEMPIO 3.3

La Figura 3.5 mostra il grafico della funzione di ripartizione della distribuzione di frequenza data nella Tabella 3.1, in cui le frequenze relative cumulate corrispondenti alle modalità: 0, 1, 2, 3, 4, 5, sono: 0.044, 0.175, 0.438, 0.806, 0.984, 1.000.

Come si può osservare, la funzione di ripartizione è una funzione costante a tratti che assume il valore 0.044 nell'intervallo $[0, 1)$, 0.175 nell'intervallo $[1, 2)$ e così via: la funzione cresce solo in corrispondenza delle singole modalità $0, 1, \ldots, 5$.

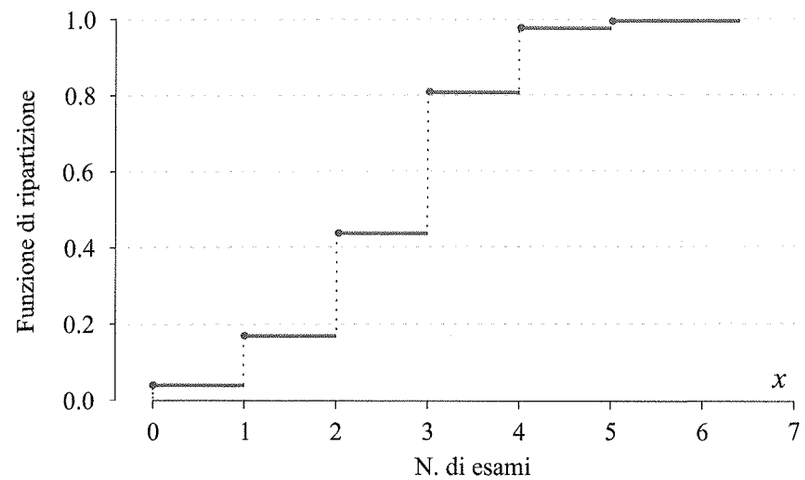


Figura 3.5
Funzione di ripartizione per la distribuzione di cui alla Tabella 3.1.

■

3.2.2 Distribuzioni di frequenza in classi

Grafico a punti

Vale la pena notare che il grafico a punti può essere utilizzato anche per rappresentare una serie di osservazioni di una variabile continua, naturalmente prima che queste siano raggruppate in classi. La Figura 3.6 mostra il grafico a punti per la serie delle lunghezze dell'avambraccio esaminata nell'Esempio 2.4.

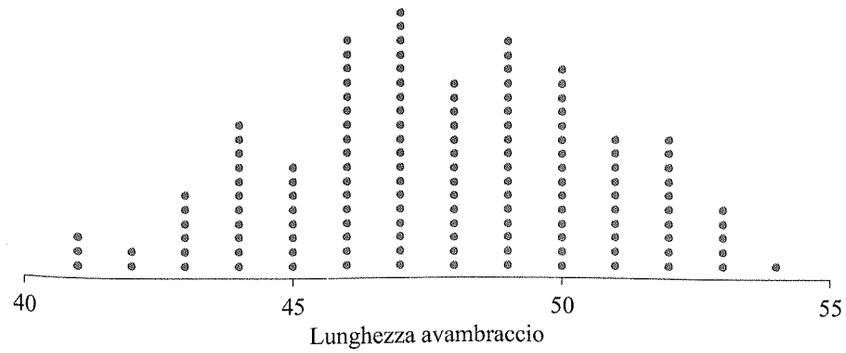


Figura 3.6
Grafico a punti dell'insieme di dati esaminato nell'Esempio 2.4.

Si consideri ora una distribuzione di frequenza in classi:

Classe	Frequenza
c_0-c_1	n_1
c_1-c_2	n_2
⋮	⋮
c_{k-1}-c_k	n_k
Totale	n

La distribuzione descritta nella tabella è composta da: n_1 osservazioni appartenenti alla classe c_0-c_1, n_2 osservazioni appartenenti alla classe c_1-c_2 e così via.

Come si è visto nel Capitolo 2, è possibile associare a ciascuna classe una densità di frequenza, data dal rapporto tra la frequenza e l'ampiezza della classe. In simboli, la densità di frequenza della classe i-esima è n_i/d_i, dove n_i è la frequenza e $d_i = c_i - c_{i-1}$ l'ampiezza della classe.

Istogramma

Il grafico che ne risulta è l'*istogramma di frequenza*, che viene definito qui di seguito.

DEFINIZIONE 3.2

L'istogramma di frequenza è un grafico adatto a rappresentare una distribuzione di frequenza in classi. Per la sua costruzione si riportano gli estremi di classe sull'asse orizzontale, le densità di frequenza sull'asse verticale; si disegna poi sopra ogni classe un rettangolo di altezza pari alla densità di frequenza e, di conseguenza, di area pari alla frequenza della classe.

L'operazione descritta nella definizione è rappresentata nella Figura 3.7 con riferimento a una generica classe.

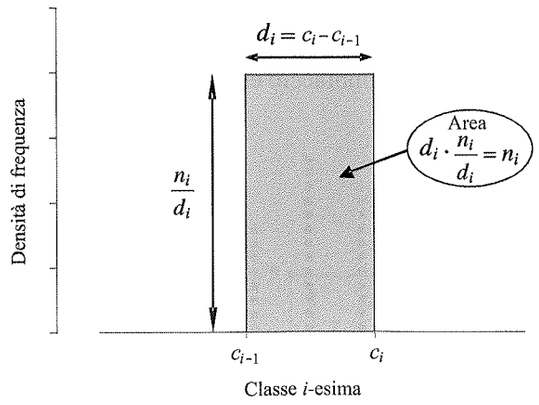


Figura 3.7
Istogramma: rettangolo relativo alla i-esima classe.

ESEMPIO 3.4

La Figura 3.8 mostra l'istogramma della distribuzione di frequenza discussa nell'Esempio 2.7, che si riporta qui per comodità.

Lunghezza avanbraccio (cm)	Frequenza	Ampiezza della classe	Densità di frequenza
40.5-42.5	5	2.0	2.5
42.5-44.5	17	2.0	8.5
44.5-46.5	25	2.0	12.5
46.5-48.5	35	2.0	17.5
48.5-50.5	32	2.0	16.0
50.5-52.5	20	2.0	10.0
52.5-54.5	6	2.0	3.0
Totale	140	–	–

Tabella 3.3
Densità di frequenza per la distribuzione di cui alla Tabella 2.4.

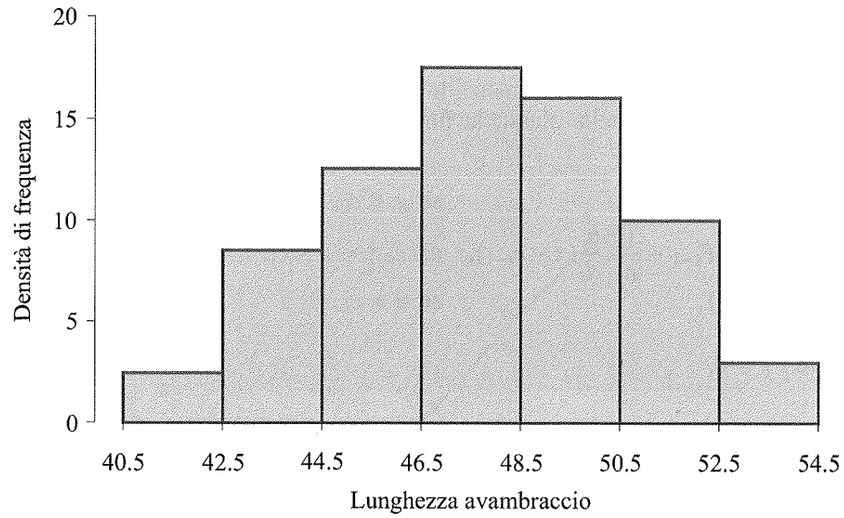


Figura 3.8
Istogramma per la distribuzione di cui alla Tabella 3.3.

Discussione

La Figura 3.8 aiuta a chiarire che l'altezza di ogni rettangolo è la densità di frequenza della classe a cui si riferisce; le altezze dei rettangoli sono: 2.5 per la prima classe, 8.5 per la seconda e così via. Per ogni rettangolo, l'area rappresenta la frequenza assoluta della classe: ad esempio, il primo rettangolo ha area 5, il secondo 17 e così via. ■

Se si mettono sull'asse verticale le frequenze assolute o relative, anziché le densità di frequenza, si possono produrre effetti distorsivi se le classi hanno ampiezza diversa. Ad esempio, in un'ipotetica distribuzione di frequenza costituita da due classi, 170-175 e 175-185, con frequenze 50 e 30, se sull'asse verticale fossero poste le frequenze, queste sarebbero rappresentate da due rettangoli aventi aree pari a 250 e a 300. In questo modo, si produrrebbe l'immagine che la frequenza (area) sia maggiore in 175-185 che non in 170-175.

In conclusione, una buona regola generale è quella di utilizzare sempre le densità di frequenza come altezze dei rettangoli che formano l'istogramma di frequenza.

Si noti che il grafico può anche essere costruito utilizzando le frequenze relative anziché quelle assolute. In tal caso, sull'asse verticale si pongono le densità di frequenza relative f_i/d_i $(i = 1, 2, \ldots, k)$, cosicché le aree dei rettangoli sono uguali alle frequenze relative.

L'istogramma può essere impiegato anche quando la variabile è discreta. Si ricorda che vanno considerati gli estremi di classe reali (si vedano il Paragrafo 2.2.1 e l'Esercizio 3.3).

Grafico della funzione di ripartizione

La definizione di funzione di ripartizione data in precedenza continua a essere valida anche per le distribuzioni di frequenza in classi. Il problema è quello di dare a $F(x)$ un'espressione analitica, sia pure approssimata. Una soluzione viene indicata nella proposizione che segue, nell'ipotesi che le unità statistiche siano uniformemente distribuite all'interno delle classi.

PROPOSIZIONE 3.1

Data una distribuzione di frequenza in classi, nell'ipotesi di uniforme distribuzione delle unità nelle classi, la funzione di ripartizione $F(x)$ assume la forma

$$F(x) = \begin{cases} 0, & \text{se } x < c_0, \\ F_{i-1} + \dfrac{1}{n} \cdot \dfrac{n_i}{d_i}(x - c_{i-1}), & \text{se } c_{i-1} < x \le c_i,\ i = 1, 2, \ldots, k, \\ 1, & \text{se } x > c_k, \end{cases} \qquad (3.1)$$

dove $F_0 = 0$.

Per ogni x appartenente all'intervallo c_{i-1}-c_i, in forza della (2.1), il numero di unità comprese tra c_{i-1} e x è dato da

$$\frac{n_i}{d_i}(x - c_{i-1}).$$

Aggiungendo a questa quantità il numero di unità con modalità minori o uguali a c_{i-1}, pari a N_{i-1}, e dividendo il risultato per n si perviene immediatamente alla seconda delle (3.1). È chiaro poi che $F(x) = 0$ per $x < c_0$ e $F(x) = 1$ per $x > c_k$, essendo nulla la frequenza delle unità aventi modalità inferiori a c_0 o superiori a c_k.

■

Si osservi che in $x = c_i$ la funzione di ripartizione è uguale a F_i, frequenza relativa cumulata della classe i-esima. Si ha allora:

$$F(c_1) = F_1,\ F(c_2) = F_2, \ldots, F(c_k) = 1.$$

La seconda delle (3.1) è l'equazione del segmento di retta che congiunge i punti di coordinate (c_{i-1}, F_{i-1}) e (c_i, F_i). La Figura 3.9 mostra questo segmento, segnalando che la sua pendenza è determinata dalla densità di frequenza della classe c_{i-1}-c_i.

In definitiva, il grafico della funzione di ripartizione si costruisce individuando i punti di coordinate $(c_0, 0), (c_1, F_1), (c_2, F_2), \ldots, (c_k, 1)$ e unendo le coppie di punti consecutivi con segmenti di retta.

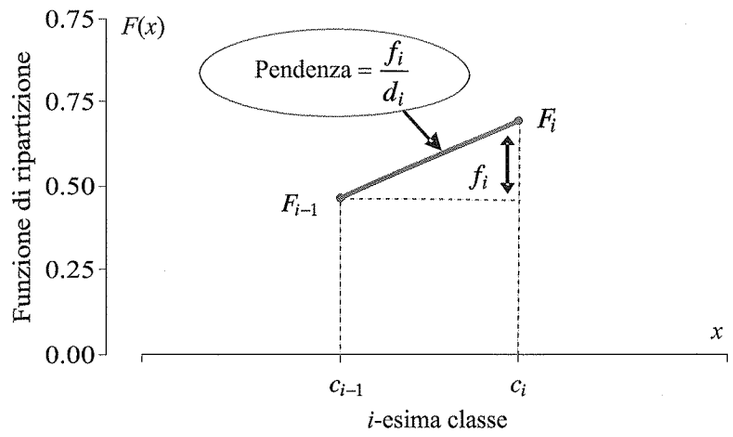


Figura 3.9
Funzione di ripartizione nel caso di una distribuzione di frequenza in classi.

ESEMPIO 3.5

La Figura 3.10 mostra la funzione di ripartizione della distribuzione di frequenza in classi data nella Tabella 3.3 (le cui frequenze relative cumulative sono 0.04, 0.16, 0.34, 0.59, 0.81, 0.96, 1.00).

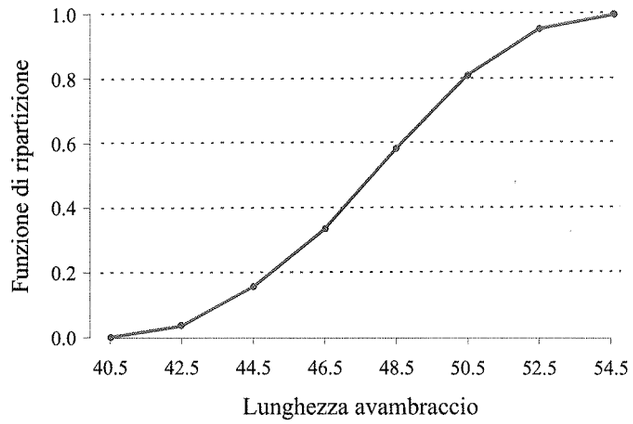


Figura 3.10
Funzione di ripartizione della distribuzione di cui alla Tabella 3.3.

Discussione

Come si vede, la funzione di ripartizione mette in luce il ritmo di accrescimento delle frequenze cumulate delle varie classi. I segmenti di retta che presentano pendenza maggiore sono quelli associati alla terza, quarta e quinta classe, che sono le classi con la più alta densità di frequenza.

■

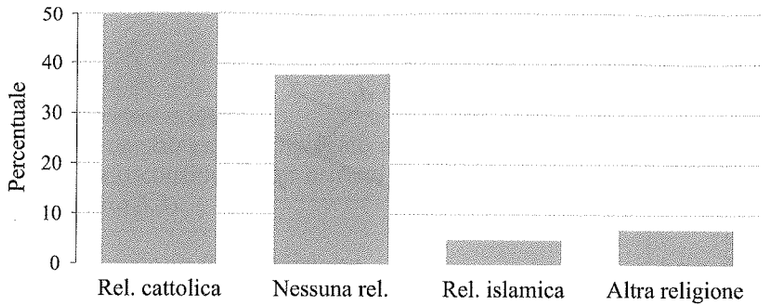
3.2.3 Distribuzioni di frequenza di variabili qualitative

Le distribuzioni di frequenza di variabili qualitative vengono in genere rappresentate con grafici a nastri, in cui la singola modalità è rappresentata da un rettangolo di altezza (o lunghezza) uguale alla frequenza della modalità.

ESEMPIO 3.6

Nella Figura 3.11, si utilizza il grafico a nastri per rappresentare la distribuzione di frequenza percentuale della popolazione francese per religione (Fonte: Eurobarometer, 2018).

Figura 3.11
Grafico a nastri della distribuzione di frequenza percentuale della popolazione francese per religione.



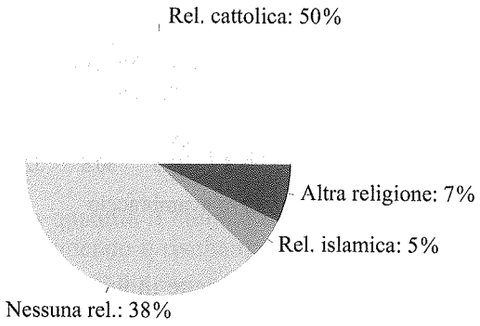
Un'alternativa al grafico a nastri è il cosiddetto grafico a torta, in cui un cerchio è suddiviso in settori circolari di dimensioni proporzionali alle frequenze delle singole modalità. I settori circolari vengono poi colorati diversamente per tenere distinte le modalità. Va osservato che tale grafico è di lettura meno immediata rispetto al grafico a nastri, essendo più difficile confrontare le dimensioni dei settori circolari che non le altezze (o lunghezze) dei nastri.

Grazie alla diffusa disponibilità di pacchetti informatici per la costruzione dei grafici, si possono realizzare figure tridimensionali anche se non è consigliabile utilizzare più dimensioni di quelle che hanno effettivamente i dati. Nell'esempio precedente, i dati hanno due dimensioni; l'uso di una terza dimensione risulta essere ridondante.

ESEMPIO 3.7

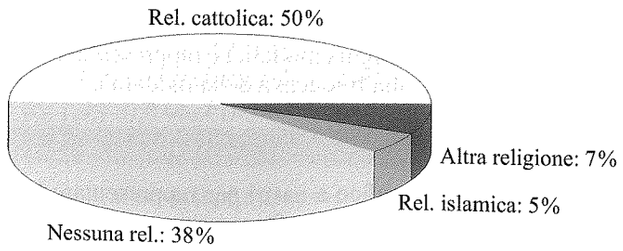
Nella Figura 3.12 è rappresentato un grafico a torta per la distribuzione della popolazione francese per religione rappresentata sopra con un grafico a nastri.

Figura 3.12
Grafico a torta per la distribuzione di frequenza percentuale della popolazione francese per religione.



La Figura 3.13 mostra la versione tridimensionale del grafico 3.12. Al di là dell'aspetto estetico, la terza dimensione complica la lettura dei dati.

Figura 3.13
Grafico 3D corrispondente alla Figura 3.12.



Una distribuzione di frequenza di una variabile qualitativa con k modalità può essere anche rappresentata ponendo uno sopra l'altro i k nastri, ciascuno avente altezza pari alla frequenza della singola modalità. In questo modo, si ottiene il grafico a nastri sovrapposti. Questo tipo di grafico è utile soprattutto se si vogliono confrontare diverse distribuzioni di frequenza percentuali.

Lo stesso confronto può anche essere effettuato usando un grafico a nastri multiplo, in cui le percentuali di ogni modalità sono rappresentate da più nastri con le opportune altezze (o lunghezze), come illustrato nell'Esempio 3.8.

ESEMPIO 3.8

Le tre distribuzioni di frequenza percentuali contenute nella Tabella 3.4 sono rappresentate, nella Figura 3.14, con un grafico a nastri sovrapposti e, nella Figura 3.15, con un grafico a nastri multiplo.

Tabella 3.4
Distribuzione di frequenza percentuale della variabile voto per sesso tratta dal *file* descritto nell'Esercizio 1.5.

Voto	Frequenza percentuale		
	Femmina	Maschio	Totale
A	31.3	25.0	26.6
B	37.5	41.7	40.6
C	18.8	20.8	20.3
D	6.2	8.3	7.8
F	6.2	4.2	4.7
Totale	100.0	100.0	100.0

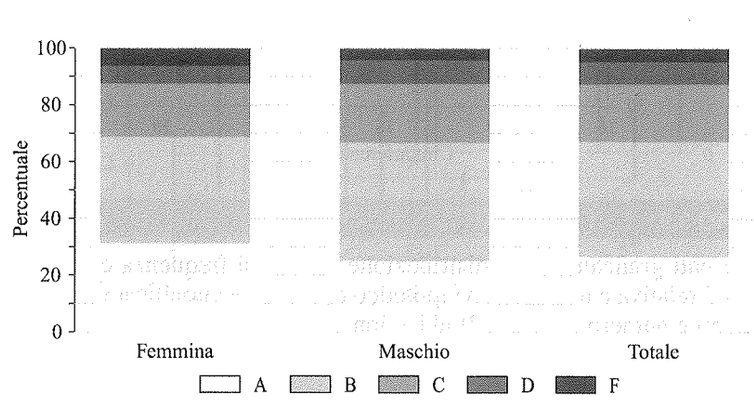


Figura 3.14
Grafico a nastri sovrapposti per la distribuzione data nella Tabella 3.4.

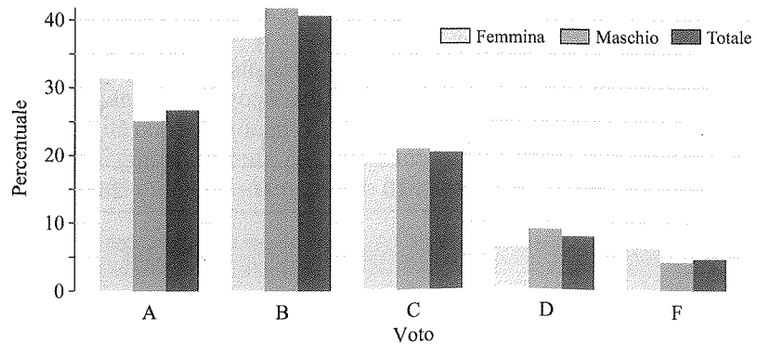


Figura 3.15
Grafico a nastri multiplo per la distribuzione data nella Tabella 3.4.

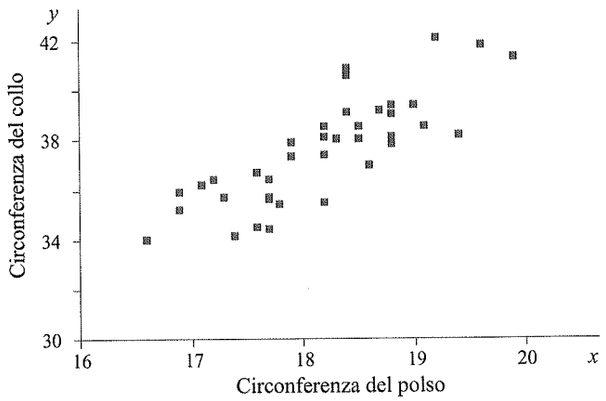
Vale la pena osservare che i nastri sovrapposti, non poggiando su una base comune, rendono poco agevole la comparazione delle rispettive altezze.

3.3 Distribuzioni doppie di frequenza

Si inizia con l'osservare che la rappresentazione grafica naturale di una serie doppia di osservazioni di due variabili quantitative, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, è il cosiddetto grafico di dispersione, che si ottiene ponendo sull'asse delle ascisse le modalità di X e su quello delle ordinate le modalità di Y, di modo che la singola coppia (x_i, y_i), $i = 1, 2, \ldots, n$, sia rappresentata da un punto del piano cartesiano.

La Figura 3.16 mostra il grafico di dispersione della serie doppia di osservazioni delle variabili circonferenza del polso e circonferenza del collo di 38 maschi adulti di età tra 22 e 30 anni (contenuta nel *file* "Dati antropometrici"): $(18.2, 38.5), (16.6, 34.0), \ldots, (16.9, 35.9)$. Il grafico è un insieme di punti del piano cartesiano: un primo punto rappresenterà la coppia ordinata $(18.2, 38.5)$, un secondo punto la coppia $(16.9, 35.9)$ e così via.

Figura 3.16
Grafico di dispersione della serie doppia di osservazioni descritta sopra.



ESEMPIO 3.9

Si rappresenti graficamente la distribuzione doppia di frequenza contenuta nella Tabella 3.5 relativa a un campione ipotetico di famiglie classificate per numero di componenti e numero di vani dell'abitazione.

Tabella 3.5
Famiglie per numero di componenti e numero di vani dell'abitazione.

N. di componenti	N. di vani					Totale
	2	3	4	5	6	
1	2	3	0	0	0	5
2	0	4	3	0	1	8
3	0	2	2	4	0	8
4	0	1	0	3	1	5
5	0	0	0	1	1	2
Totale	2	10	5	8	3	28

Discussione

L'uso del grafico di dispersione per rappresentare le coppie di modalità $(1, 2)$, $(1, 3)$ ecc. della distribuzione doppia di frequenza è inappropriato perché fa perdere l'informazione sulle frequenze associate a tali coppie. Due possibili soluzioni sono il grafico a bolle e il grafico ad aste in un sistema di tre assi cartesiani. Nel primo, la frequenza di ciascuna coppia di modalità viene rappresentata con cerchi aventi aree proporzionali alle frequenze. Nel secondo, la frequenza viene rappresentata con il terzo asse cartesiano.

Le Figure 3.17 e 3.18 presentano il grafico a bolle e quello ad aste della distribuzione doppia di frequenza.

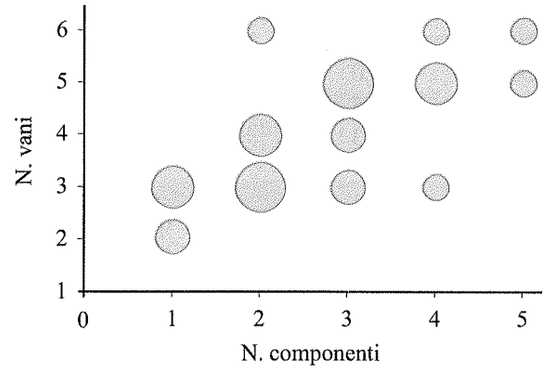


Figura 3.17
Grafico a bolle della distribuzione doppia di frequenza di cui alla Tabella 3.5.

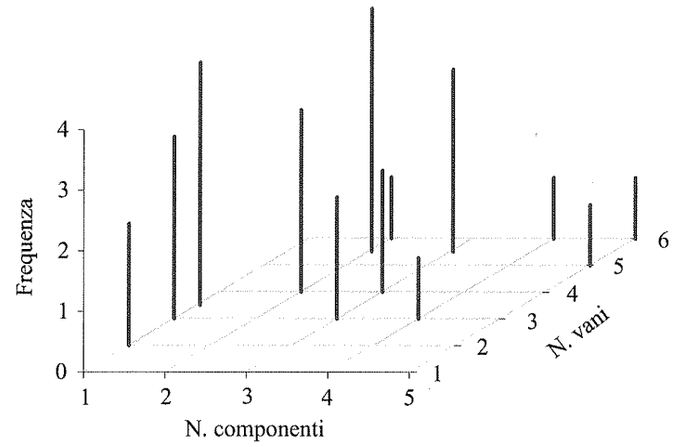


Figura 3.18
Grafico ad aste della distribuzione doppia di frequenza di cui alla Tabella 3.5.

■

3.4 Serie storiche

Per la rappresentazione grafica delle serie storiche si ricorre generalmente ai diagrammi cartesiani. Si pongono sull'asse delle ascisse i tempi e su quello delle ordinate le quantità corrispondenti; i conseguenti punti del piano cartesiano vengono poi uniti con segmenti di rette per facilitare la percezione visiva dell'andamento del fenomeno. Si ottiene così un grafico a linee.

ESEMPIO 3.10

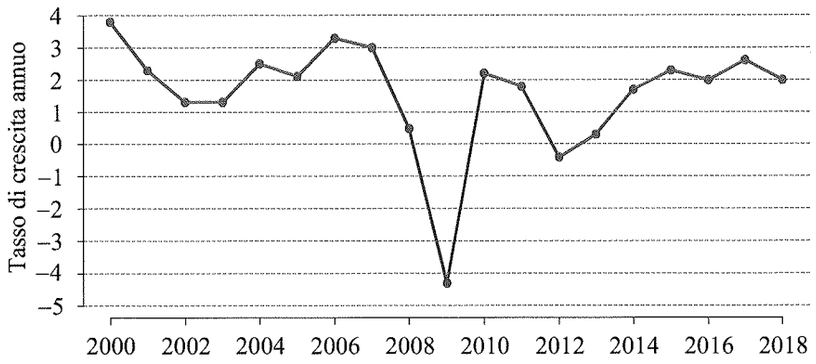
Nella Tabella 3.6 sono riportati i tassi di crescita reale annui del prodotto interno lordo dei paesi dell'UE-28 dal 2000 al 2018 (Fonte: Eurostat).

Tabella 3.6
Tassi di crescita reale annui del prodotto interno lordo dei paesi dell'UE-28 dal 2000 al 2018.

Anno	Tasso di crescita annuo	Anno	Tasso di crescita annuo
2000	3.8	2010	2.2
2001	2.3	2011	1.8
2002	1.3	2012	–0.4
2003	1.3	2013	0.3
2004	2.5	2014	1.7
2005	2.1	2015	2.3
2006	3.3	2016	2.0
2007	3.0	2017	2.6
2008	0.5	2018	2.0
2009	–4.3		

Il grafico a linee di questa serie storica è riportato qui di seguito.

Figura 3.19
Tassi di crescita reale annui del prodotto interno lordo dei paesi dell'UE-28 dal 2000 al 2018.

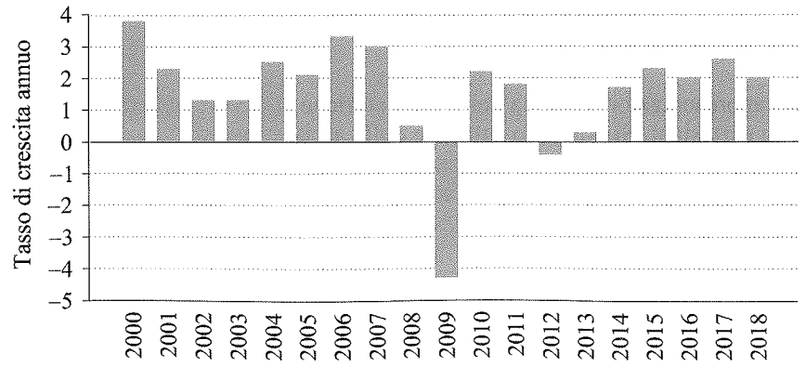


Alternativamente, una serie storica può essere rappresentata con un grafico a nastri come illustrato nell'Esempio 3.11.

ESEMPIO 3.11

Nella Figura 3.20 viene rappresentata la stessa serie storica dell'Esempio 3.10 con un grafico a nastri.

Figura 3.20
Tassi di crescita reale annui del prodotto interno lordo dei paesi dell'UE-28 dal 2000 al 2018.



Vale la pena segnalare che si distingue tra variabili di stato e variabili di flusso. Il criterio distintivo è legato al processo di misurazione della variabile. Nel primo caso, la misura è riferita a un determinato istante, ad esempio il 31 dicembre di un certo anno, nel secondo, la misura è riferita a un intervallo di tempo, ad esempio un anno. La popolazione di un Paese è un esempio di variabile di stato, il PIL è un esempio di variabile di flusso. I tempi sono propriamente rappresentati da punti, nel primo caso, e da intervalli, nel secondo. A rigore, i grafici a linee dovrebbero essere utilizzati per le variabili di stato e i grafici a nastri per le variabili di flusso. Nelle applicazioni pratiche tuttavia questa distinzione è generalmente disattesa a favore di grafici a linee che rendono più agevole la percezione visiva dell'andamento del fenomeno.

Per finalità di confronto, due o più serie possono essere rappresentate su uno stesso grafico, come mostrato nella Figura 3.21.

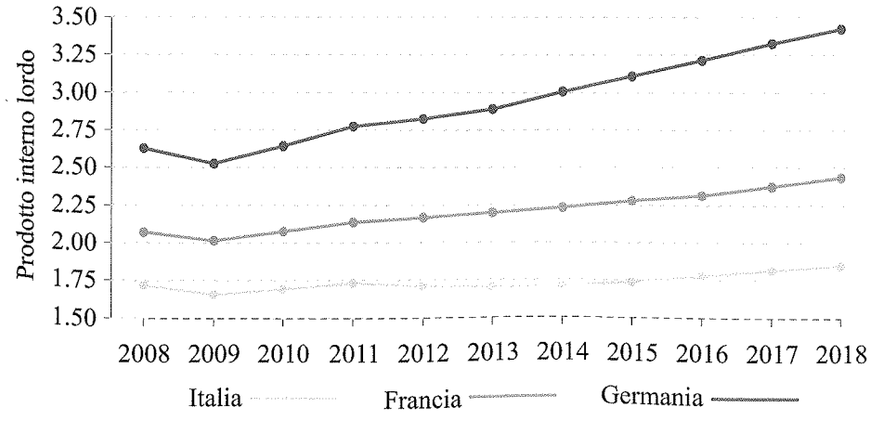
ESEMPIO 3.12

La Figura 3.21 contiene i grafici a linee del prodotto interno lordo (PIL) ai prezzi di mercato di Italia, Francia e Germania dal 2008 al 2018, riportati nella Tabella 3.7 (Fonte: Eurostat; valori in migliaia di miliardi di euro).

Tabella 3.7
Prodotto interno lordo (PIL) ai prezzi di mercato di Italia, Francia e Germania dal 2008 al 2018.

Anno	Italia	Francia	Germania
2008	1638	1992	2546
2009	1577	1936	2446
2010	1611	1995	2564
2011	1649	2058	2694
2012	1624	2089	2745
2013	1613	2117	2811
2014	1627	2150	2927
2015	1655	2198	3030
2016	1696	2234	3134
2017	1737	2295	3245
2018	1765	2353	3344

Figura 3.21
Prodotto interno lordo (PIL) ai prezzi di mercato di Italia, Francia e Germania dal 2008 al 2018.



3.5 Serie territoriali

Le serie territoriali si rappresentano generalmente con grafici a nastri o con cartogrammi. In questi ultimi, le aree geografiche in cui il territorio è suddiviso sono contraddistinte da diverse gradazioni di uno stesso colore a seconda delle modalità che la variabile di interesse assume nelle aree stesse. Al riguardo, si parla anche di mappe tematiche.

ESEMPIO 3.13

Nelle figure che seguono sono rappresentate due serie territoriali. Nella prima, viene usato il grafico a nastri orizzontali per rappresentare i pernottamenti nelle strutture ricettive nelle prime 20 regioni dei Paesi dell'UE-28 (Fonte: Eurostat; anno 2017; dati in milioni). Nella seconda, viene usato il cartogramma per rappresentare l'indice di vecchiaia in Italia per Regione (Fonte: Istat, 2019).

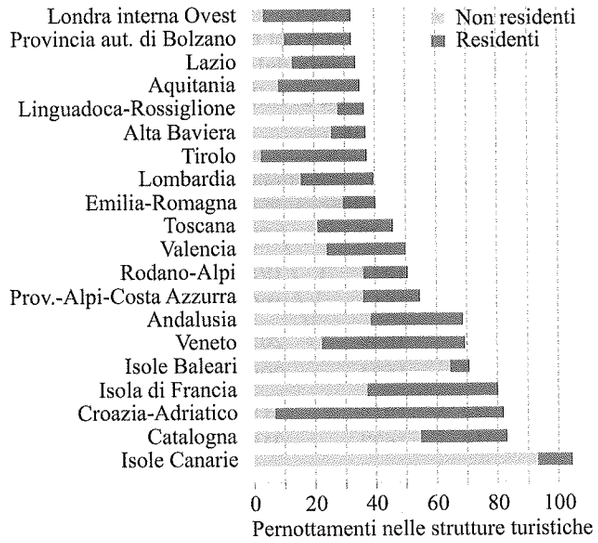


Figura 3.22
Prime 20 regioni dei paese dell'UE-28 in termini di pernottamenti nelle strutture ricettive.

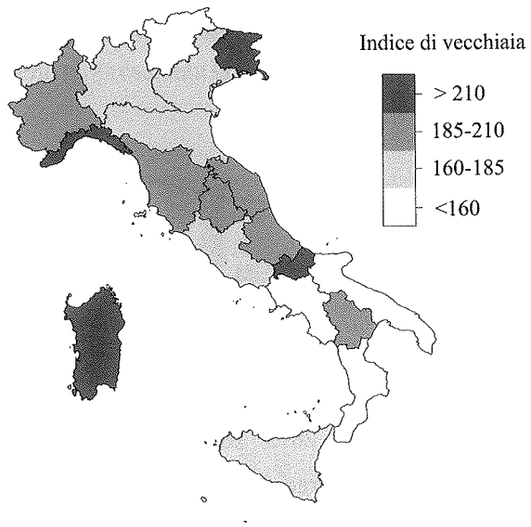


Figura 3.23
Indice di vecchiaia in Italia per regione.

3.6 Problemi di scala

Quando le grandezze da rappresentare sono caratterizzate da una forte dispersione (per esempio, alcuni valori sono minori di 1, mentre altri nell'ordine di centinaia o migliaia), il problema è quello di rendere visibili sia i valori piccoli sia quelli grandi.

La soluzione più comune del problema è quella dell'interruzione della scala con un simbolo appropriato, come indicato nei grafici (a) e (b) della Figura 3.24.

Una seconda possibilità è quella di effettuare una trasformazione logaritmica delle quantità da rappresentare. Il logaritmo ha infatti la proprietà di comprimere la scala lineare. Per esempio, i logaritmi in base 10 di 0.1, 1, 10, 100 e 1000 sono, rispettivamente, −1, 0, 1, 2 e 3. Come si vede, nella scala logaritmica, la distanza tra i numeri 0.1 e 1, 1 e 10, 10 e 100, 100 e 1000 è costante e uguale a 1 (si veda il grafico (c) della Figura 3.24).

ESEMPIO 3.14

Nella Figura 3.24 vengono illustrate le tecniche per fronteggiare i problemi di scala. Il grafico (a), si riferisce agli impieghi dei risparmi delle famiglie italiane (Fonte: Banca d'Italia), viene "tagliato" un nastro per segnalare che per esso non vale la scala di riferimento. Nel grafico (b), si fanno coesistere due scale per le ascisse, una per valori da 0 a 40, l'altra per valori da 40 a 65. Questo accorgimento consente di rendere più visibili gli istogrammi relativi alle prime classi. Nel grafico (c), che rappresenta la frequenza relativa di testa nel lancio ripetuto di una moneta, viene adottata la scala logaritmica per le ascisse.

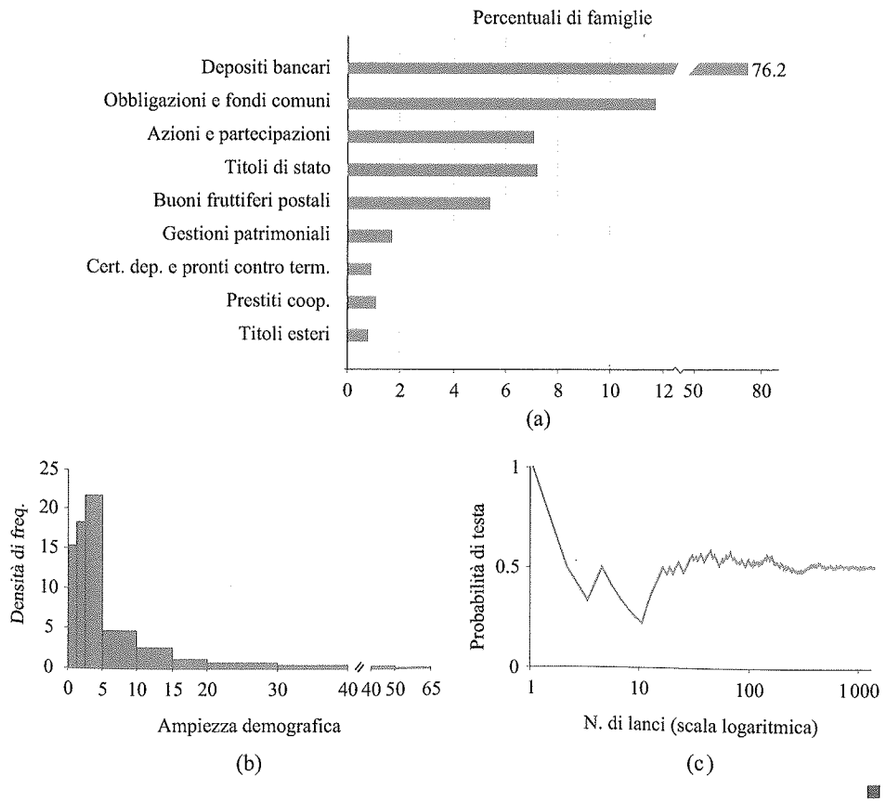


Figura 3.24
Soluzioni *ad hoc* di specifici problemi di scala.

Riassunto

In questo capitolo, viene trattato il tema della rappresentazioni grafiche con particolare riferimento alle distribuzioni di frequenza, semplici e doppie, alle serie storiche e alle serie territoriali.

Le distribuzioni di frequenza delle variabili quantitative sono rappresentate con diagrammi cartesiani: il grafico ad aste per le variabili discrete e l'istogramma di frequenza per le distribuzioni in classi. In entrambi i casi, si può fare ricorso al grafico della funzione di ripartizione. Le distribuzioni di frequenza di variabili qualitative sono generalmente rappresentate con grafici a nastri. Anche i grafici a torta sono abbastanza utilizzati, soprattutto per le distribuzioni di frequenza percentuali.

Le distribuzioni doppie di frequenza di due variabili quantitative possono essere rappresentate su un piano cartesiano con un grafico a bolle oppure su un sistema di tre assi cartesiani con un diagramma tridimensionale.

Per le serie storiche, la rappresentazione più comunemente usata è il grafico a linee: sull'asse orizzontale si riportano i tempi e su quello verticale le corrispondenti modalità della variabile di interesse.

Per le serie territoriali le rappresentazioni più appropriate sono i grafici a nastri e i cartogrammi.

Parole chiave

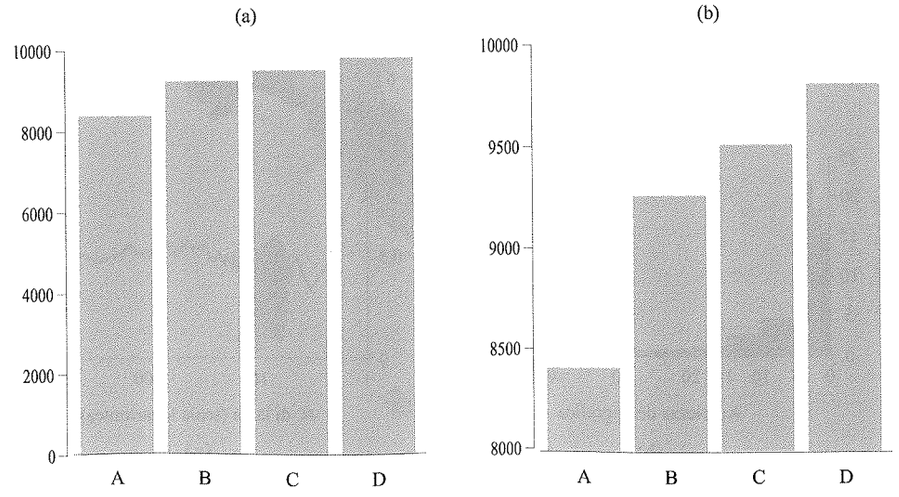
cartogramma, 82
funzione di frequenza, 68
funzione di ripartizione, 69
grafico a bolle, 79
grafico ad aste, 68
grafico a linee, 79
grafico a nastri, 69
grafico a nastri multiplo, 77
grafico a nastri sovrapposti, 77
grafico a punti, 67
grafico a torta, 76
istogramma di frequenza, 72
mappe tematiche, 82
variabili di flusso, 81
variabili di stato, 81

Commenti e curiosità

1. Grafici ingannevoli

I due grafici nella Figura 3.25 rappresentano la stessa serie storica. Il grafico (b) è ricavato dal grafico (a) modificando l'origine della scala, fissandola a 8000. In questo modo, le differenze di altezza dei nastri appaiono molto più accentuate di quanto non lo siano in realtà. Ne segue che il grafico (b) trasmette al lettore un'impressione fuorviante.

Figura 3.25
Grafici relativi a una stessa serie con scale differenti.



2. Infografica: quando i grafici raccontano le notizie

La rappresentazione grafica dei dati sta avendo un ruolo sempre più rilevante nella comunicazione e diffusione dell'informazione – sui giornali, sulle riviste specializzate, in televisione e sulla rete – intorno ai più svariati fenomeni: socio-demografici, economici, finanziari e anche epidemiologici (come nel caso della pandemia di covid-19).

Tutto ciò è agevolato, da un lato, dalla facilità di accesso in tempo reale ai dati di una vasta molteplicità di fonti, dall'altro, dalla disponibilità di programmi di grafica avanzata con cui si possono realizzare in tempi rapidi suggestive ed efficaci rappresentazioni grafiche, anche di tipo interattivo.

Questa modalità di diffusione dell'informazione va sotto il nome di infografica, termine che deriva dalla fusione delle parole *information* e *graphic*, e che indica una visualizzazione dell'informazione in cui la componente grafica, il "colpo d'occhio", prevale su quella testuale.

Un esempio di infografica giornalistica viene presentato di seguito: con una combinazione di grafici viene raccontata la storia della guerra in Afganistan nel ventennio 2002-2021.

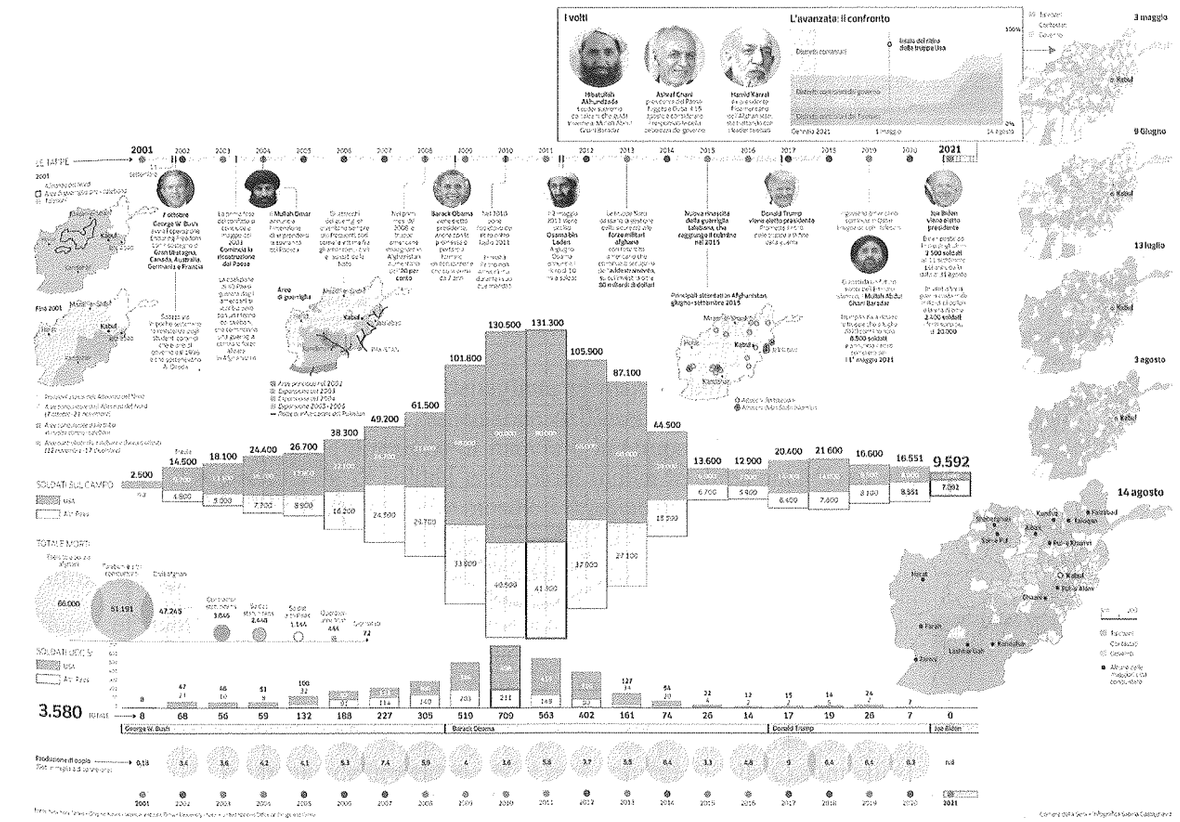


Immagine a colori

3. I precursori dell'infografica

Si potrebbero scrivere pagine e pagine su questo tema. Per ragioni di spazio, ci si limiterà a descrivere in breve alcuni contributi di tre protagonisti: William Playfair, Charles Joseph Minard e Florence Nightingale.

William Playfair

William Playfair (1759-1823) – ingegnere, disegnatore, statistico, scrittore e agente segreto scozzese – è considerato l'inventore delle rappresentazioni grafiche in statistica. Si devono a lui i primi grafici a nastri, a linee e a torta. Qui sotto vengono presentati tre suoi grafici piuttosto creativi per l'epoca.

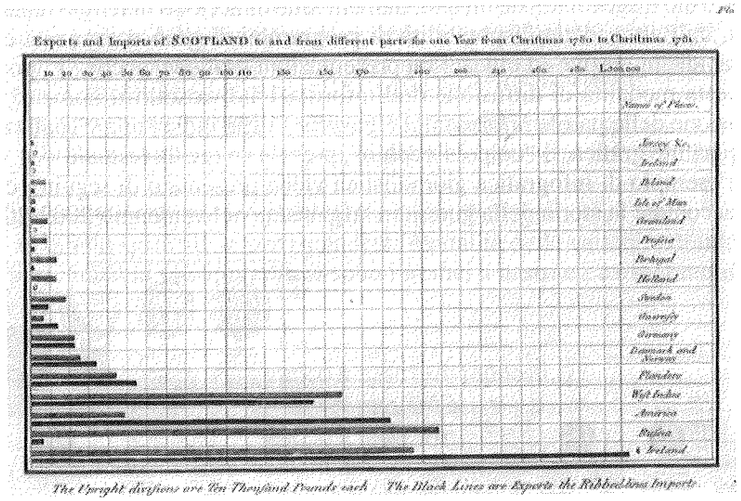


Figura 3.26
Grafico a nastri delle esportazioni e delle importazioni tra la Scozia e altri Paesi dal 25 dicembre 1780 al 25 dicembre 1781.

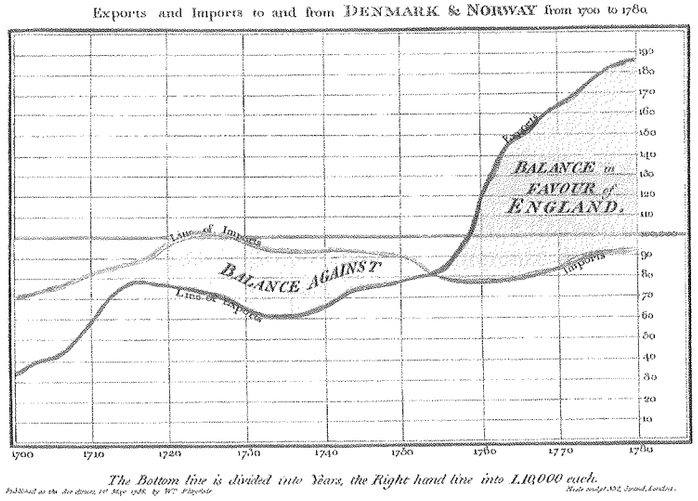


Figura 3.27
Grafico a linee delle importazioni ed esportazioni tra Inghilterra e Danimarca e Norvegia nel periodo 1700-1780.

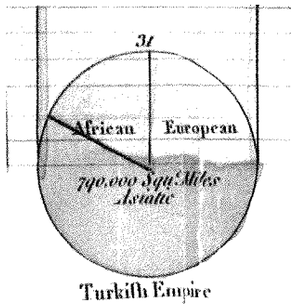


Figura 3.28
Grafico a torta dell'estensione dell'impero turco in Asia, Europa e Africa prima del 1789.

Charles Joseph Minard

Charles Joseph Minard (1781-1870) – ingegnere civile francese – è stato reso famoso dal suggestivo grafico che illustra le forti perdite subite dall'esercito di Napoleone negli anni 1812-1813, durante la catastrofica campagna di Russia.

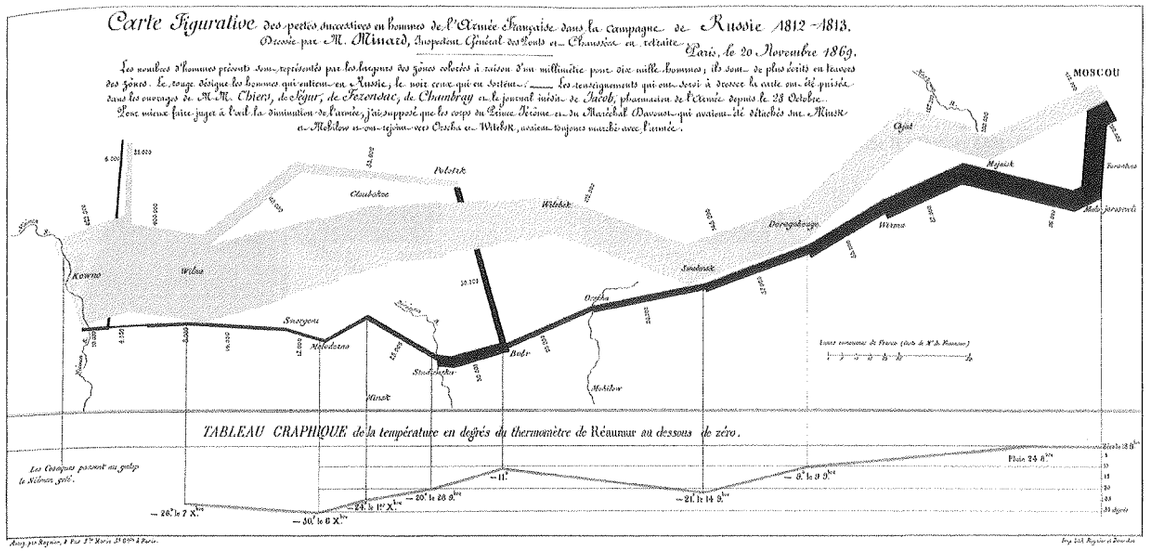
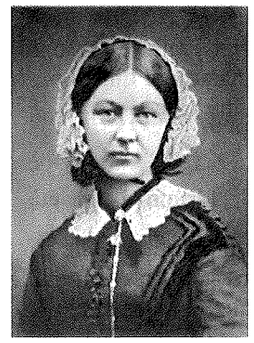


Figura 3.29 Grafico della campagna di Russia dell'esercito di Napoleone – 1812-1813.

Il grafico descrive in maniera molto dettagliata la marcia delle truppe napoleoniche verso Mosca e la ritirata. Vengono rappresentate: la dimensione dell'esercito, i nomi di alcune posizioni geografiche, la strada percorsa, il tempo, la direzione di marcia e la temperatura. In particolare, l'ampiezza della linea rappresenta il numero totale di soldati e il colore indica la direzione. Sotto al grafico principale viene disegnata la linea della temperatura che mostra il drastico crollo termico durante l'inverno.

Florence Nightingale



Florence Nightingale (1820-1910) – infermiera e statistica inglese nata a Firenze (fu chiamata Florence in onore di Firenze; la sorella maggiore, nata a Napoli, era stata chiamata Parthenope perché nata a Napoli) – fu la fondatrice dell'assistenza infermieristica dei soldati in guerra ed ebbe un ruolo centrale nella costituzione della *Royal Commission on the Health of the Army*. Non potendo farne parte, in quanto donna, fu incaricata della redazione del rapporto finale della Commissione, che arricchì di statistiche molto dettagliate. Ebbe un ruolo importante anche nella successiva attuazione delle raccomandazioni finali del rapporto, che rivoluzionarono la sanità militare: fu costituito l'*Army Medical College* e molti ospedali, soprattutto militari, vennero costruiti seguendo le sue indicazioni.

Era nota come la "signora con la lanterna" grazie a un articolo del Times, che ne lodava l'abnegazione durante l'assistenza ai soldati feriti, e alla lettera di un paziente soldato che descriveva la sua attività di assistenza infermieristica. Infatti, durante la guerra di Crimea (1853-56), Nightingale, una volta che i medici avevano terminato il loro turno, si recava di notte in ospedale a controllare le condizioni dei malati, aggirandosi tra i letti con una lanterna. Fu proprio durante questa guerra che, per convincere la regina Vittoria a migliorare le condizioni degli ospedali da campo, costruì il celebre istogramma circolare con cui diede evidenza delle cause di morte dei soldati nei vari mesi del conflitto.

Durante la guerra di Crimea, il tasso di mortalità dei soldati raggiunse livelli molto elevati, ma non furono le battaglie le uniche responsabili di queste morti: all'ospedale militare allestito nella caserma Selimiye di Scutari, Florence Nightingale e le sue infermiere scoprirono che i soldati feriti erano mal curati nell'indifferenza delle autorità. L'ospedale era sovraffollato e il personale medico sovraccarico, le medicine erano scarse, l'igiene trascurata, gli ambienti scarsamente ventilati, le infezioni di massa comuni e spesso fatali, la cucina non attrezzata. Nightingale utilizzò il suo grafico per dimostrare che la maggior parte dei decessi era dovuta a pratiche ospedaliere scadenti.

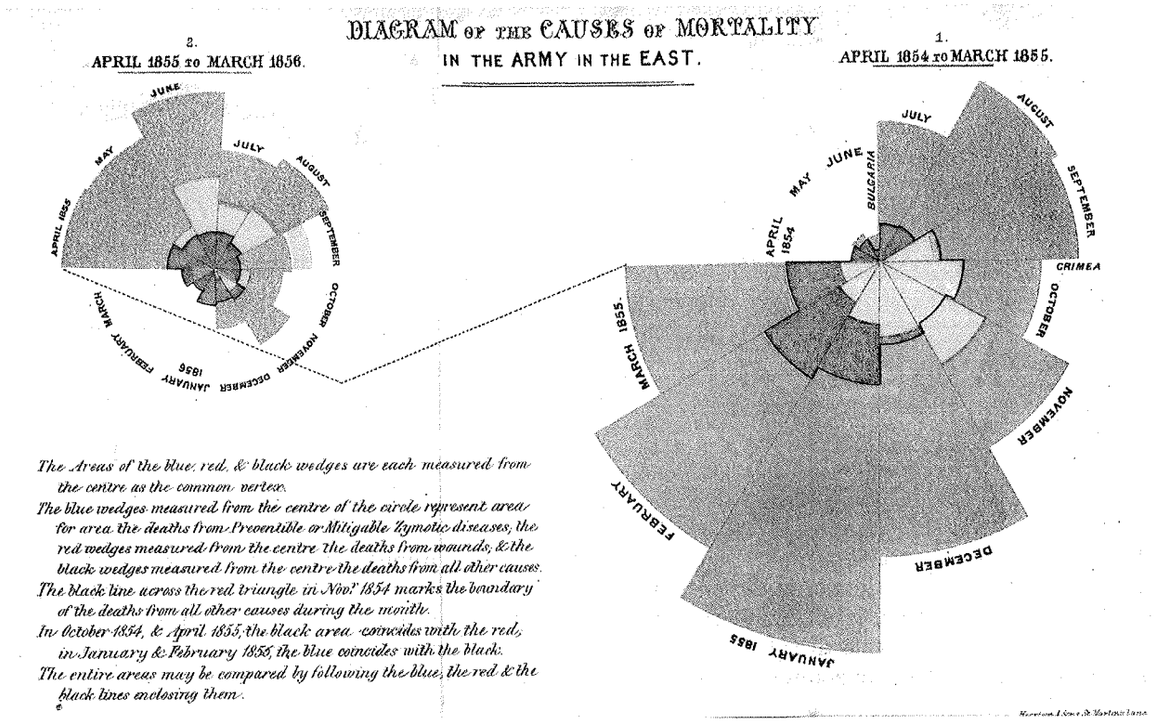


Figura 3.30 Grafico delle cause di mortalità nell'esercito d'Oriente.

Nel grafico, per ogni mese, lo spicchio di colore rappresenta i morti per "cause prevenibili", cioè malattie e infezioni; lo spicchio di colore i morti in battaglia; lo spicchio di colore i deceduti per altre cause.

4. Piramide delle età

La piramide delle età è il grafico con cui si rappresenta la distribuzione della popolazione di un Paese per sesso e classi di età. Sull'asse verticale vengono poste le classi di età (generalmente di ampiezza 5), su quello orizzontale le frequenze di classe (si possono usare le frequenze invece delle densità di frequenza perché le classi hanno la stessa ampiezza). Il grafico è la combinazione di due istogrammi di frequenza (orizzontali) disposti simmetricamente intorno all'asse verticale.

Nella Figura 3.31 sono rappresentate le piramidi delle età della popolazione italiana nel 1971 e nel 2021 (la scala dell'asse orizzontale è espressa in migliaia). Il nome attribuito al grafico è abbastanza plausibile per il grafico di sinistra; non lo è per quello di destra: negli ultimi decenni la struttura per età della popolazione è cambiata notevolmente, data la minore incidenza delle classi giovanili, a cui si contrappone il peso molto più rilevante delle classi di età più avanzate.

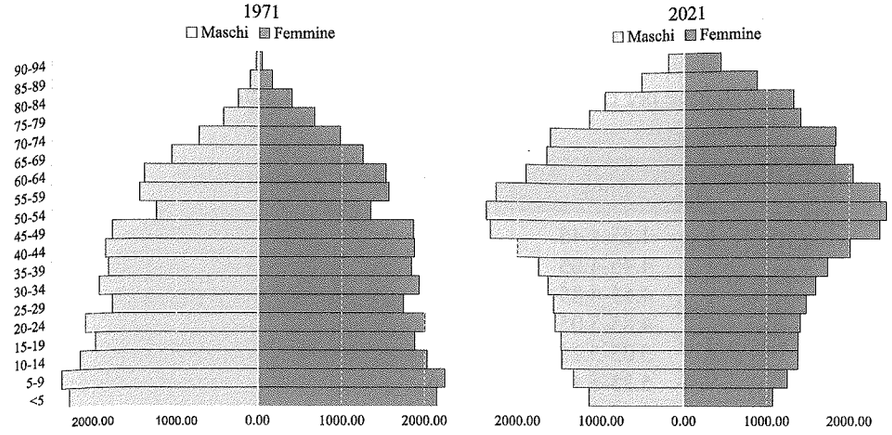


Figura 3.31
Piramidi dell'età della popolazione italiana negli anni 1971 e 2021.

Esercizi

3.1 Si costruisca il grafico a punti per i dati sulla lunghezza dei sepali dell'iris setosa contenuti nella prima colonna del *file* "Iris" .

3.2 Si consideri la distribuzione di frequenza relativa a 1534 biologi secondo il numero di articoli di ricerca pubblicati in una certa rivista (Hand *et al.*, 1994, pp. 141-142):

N. articoli	Frequenza
1	1062
2	263
3	120
4	50
5	22
6	7
7	6
8	2
9	0
10	1
11	1

a) Si disegni il grafico ad aste.
b) Si determini la distribuzione di frequenza relativa cumulata e la si rappresenti graficamente.

3.3 La distribuzione di frequenza riportata nella tabella sottostante è stata ottenuta dal *file* "Corso di statistica" descritto nell'Esercizio 1.5.

Punteggio	Frequenza
49-59	5
60-69	8
70-79	9
80-84	11
85-89	9
90-94	17
95-99	5
Totale	64

Si rappresenti la distribuzione con un istogramma di frequenza. (*Suggerimento*: si utilizzino gli estremi di classe reali: 48.5-59.5, 59.5-69.5 ecc.)

3.4 Si consideri la seguente distribuzione per classi ottenuta dai dati descritti nell'Esercizio 3.1:

Lunghezza sepali (cm)	Frequenza
4.25-4.45	4
4.45-4.65	5
4.65-4.85	7
4.85-5.05	12
5.05-5.25	11
5.25-5.45	6
5.45-5.65	2
5.65-5.85	3
Totale	50

a) Si costruisca l'istogramma di frequenza.

b) Per la distribuzione data, si possono usare le frequenze al posto delle densità di frequenza? Si motivi la risposta.

c) Si disegni la funzione di ripartizione e in base a questa si calcoli la percentuale di osservazioni minori o uguali a 5.2.

3.5 Nei paesi dell'UE-28, la spesa media delle famiglie in beni e servizi culturali per destinazione di spesa nell'anno 2015 è riportata nella tabella che segue:

Destinazione di spesa	Percentuale
Materiale per l'espressione e la creazione artistica	11.5
Libri e riviste	25.0
Attrezzature	28.5
Tariffe abbonamenti	19.7
Intrattenimento	15.3

Si rappresenti la distribuzione con un grafico a torta.

3.6 La tabella che segue riporta il tasso di disoccupazione giovanile per sesso nei paesi dell'UE-28 dal 2010 al 2018:

Anno	Percentuale		
	Femmine	Maschi	Totale
2010	20.3	22.0	21.2
2011	21.0	22.4	21.7
2012	22.4	23.9	23.2
2013	22.9	24.4	23.7
2014	21.5	22.9	22.2
2015	19.5	21.1	20.4
2016	17.9	19.4	18.7
2017	16.1	17.5	16.9
2018	14.6	15.8	15.2

Si disegni il grafico a linee per le tre serie storiche.

3.7 La tabella che segue contiene la distribuzione percentuale del livello di padronanza della lingua straniera più conosciuta nei sei paesi dell'UE-28 più popolati (Fonte: Eurostat; anno 2017):

Paese	Percentuale		
	Base	Buono	Ottimo
Francia	51.1	29.0	19.9
Germania	39.6	32.8	27.6
Italia	63.7	25.5	10.8
Polonia	59.1	26.1	14.8
Spagna	38.8	31.1	30.1
Regno Unito	22.0	25.0	53.0

Si rappresenti la distribuzione percentuale utilizzando un grafico a nastri sovrapposti.

3.8 La seguente tabella riporta i tassi di occupazione per sesso (fascia di età 20-64 anni) nelle cinque maggiori economie dell'UE-28 nel 2018 (Fonte: Eurostat):

Paese	Percentuale		
	Femmine	Maschi	Totale
Francia	67.6	75.2	71.3
Germania	75.8	83.9	79.9
Italia	53.1	72.9	63.0
Spagna	61.0	73.1	67.0
Regno Unito	73.8	83.7	78.7

Si rappresentino con grafico a nastri multiplo i tassi di occupazione per sesso nei cinque paesi.

3.9 Nella seguente tabella è riportato il PIL reale pro capite dei paesi dell'UE-27 nel 2018 (Fonte: Eurostat):

Paese	PIL	Paese	PIL	Paese	PIL
Austria	38	Francia	33	Malta	22
Belgio	36	Germania	36	Olanda	42
Bulgaria	7	Grecia	18	Polonia	12
Croazia	12	Ungheria	13	Portogallo	18
Cipro	24	Irlanda	58	Romania	9
Rep. Ceca	18	Italia	27	Slovenia	20
Danimarca	48	Lettonia	12	Spagna	25
Estonia	15	Lituania	13	Svezia	44
Finlandia	37	Lussemburgo	83	Regno Unito	33

Si rappresenti la serie spaziale utilizzando un grafico a nastri orizzontali.

3.10 Consumo energetico (miliardi di tonnellate di petrolio equivalente) dei primi dieci consumatori di energia dell'UE nel 2018 (Fonte: Eurostat):

Paese	Energia consumata	
	Gas naturale	Altri prodotti
Austria	5.0	21.0
Belgio	10.0	23.0
Francia	28.0	112.0
Germania	53.0	148.0
Italia	34.0	80.0
Olanda	17.0	28.0
Polonia	9.0	61.0
Spagna	14.0	68.0
Svezia	0.5	31.5
Regno Unito	22.0	25.0

Si rappresenti la serie territoriale con un grafico a nastri sovrapposti.

Grafici con Excel

Microsoft Excel offre una vasta gamma di grafici tra cui è possibile scegliere la soluzione più appropriata per la situazione concreta. Di seguito viene presentato il menu di scelta del tipo di grafico.

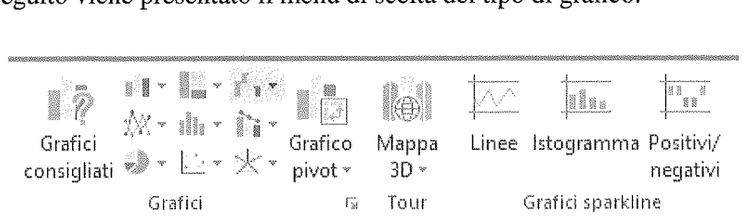


Figura 3.32
Tipologie di grafici con Excel.

CAPITOLO

4

Medie

Sommario del capitolo

4.1 Introduzione
4.2 Media aritmetica
4.3 Media armonica
4.4 Media geometrica
4.5 Media quadratica
4.6 Medie di potenza
4.7 Medie analitiche delle distribuzioni di frequenza
4.8 Medie ponderate
4.9 Mediana
4.10 Quartili e quantili
4.11 Mediana, quartili e quantili delle distribuzioni di frequenza
4.12 Valore centrale
4.13 Moda
4.14 Medie e dati anomali
4.15 La scelta della media
Riassunto
Commenti e curiosità
Esercizi

4.1 Introduzione

Le medie servono a rappresentare in modo sintetico un insieme di dati e misurano, in qualche modo, l'ordine di grandezza del fenomeno di cui i dati sono espressione. Le medie sono anche chiamate misure di tendenza centrale perché il loro valore rappresenta il "centro" attorno a cui tendono a collocarsi i dati. Ad eccezione della moda, le medie che verranno illustrate in questo capitolo possono essere raggruppate in due categorie: medie analitiche e medie di posizione. Il primo gruppo comprende le medie: aritmetica, armonica, geometrica e quadratica. Esse sono così chiamate perché si ottengono dall'applicazione di operazioni algebriche alle osservazioni. Appartengono al secondo gruppo: la mediana, i quartili, i decili (più in generale i quantili) e il valore centrale. Ciascuna di esse corrisponde a un elemento o alla media di due elementi che occupano particolari posizioni nella graduatoria non decrescente delle osservazioni.

Le medie analitiche possono essere calcolate solo su dati quantitativi. Le medie di posizione possono essere calcolate sia su dati quantitativi sia su dati ordinali. Sui dati nominali è possibile calcolare solo la moda.

Verrà prima trattato il caso in cui i dati sono nella forma di serie di osservazioni, $x_1, x_2, \ldots, x_n$. Le formule o le procedure di calcolo verranno poi estese al caso delle distribuzioni di frequenza.

4.2 Media aritmetica

La media aritmetica è la media per antonomasia: nel seguito, quando si userà il termine media senza altra qualificazione, si farà riferimento ad essa. Insieme alle percentuali e ai grafici, è lo strumento statistico più largamente conosciuto e utilizzato.

DEFINIZIONE 4.1

La media aritmetica di una serie di n osservazioni è definita come la somma[1] delle osservazioni divisa per il loro numero:

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i. \qquad (4.1)$$

[1] Per l'operatore di sommatoria Σ si veda il Paragrafo A.1 dell'Appendice A.

PROPOSIZIONE 4.1

La media gode delle proprietà di seguito indicate.

(a) È interna, nel senso che è maggiore o uguale all'osservazione più piccola e minore o uguale all'osservazione più grande.

(b) È quel valore che sostituito a ciascuna delle osservazioni della serie lascia invariata la somma:

$$n\mu = \sum_{i=1}^{n} x_i.$$

(c) La somma algebrica degli scarti dalla media, $x_i - \mu$, è nulla:

$$\sum_{i=1}^{n}(x_i - \mu) = 0.$$

(d) La somma dei quadrati degli scarti da una costante c è minima quando c è uguale alla media aritmetica:

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \min_{c} \sum_{i=1}^{n}(x_i - c)^2.$$

(e) Se dalla serie di osservazioni $x_1, x_2, \ldots, x_n$, avente media μ, si passa alla nuova serie $y_1, y_2, \ldots, y_n$ mediante la trasformazione lineare $y_i = a + bx_i$ $(i = 1, 2, \ldots, n)$, dove a e b sono costanti arbitrarie, la media aritmetica della nuova serie è data da $a + b\mu$. Al riguardo si parla di linearità della media aritmetica.

Le dimostrazioni si trovano nel Paragrafo A.2.1 dell'Appendice A, dove è illustrata un'ulteriore proprietà della media. ■

- La proprietà (a) è ovvia. Va notato che il segno di uguaglianza sussiste solo se tutte le osservazioni sono tra loro uguali.

- La proprietà (b) deriva direttamente dalla definizione di media aritmetica: moltiplicando la media aritmetica per il numero delle osservazioni, si ottiene il numeratore della (4.1), che è la somma delle osservazioni.

- La proprietà (c) conferma che μ è un valore intermedio rispetto alle osservazioni minima e massima: una parte delle osservazioni sono minori di μ e danno scarti negativi, le restanti sono maggiori di μ e danno scarti positivi. La particolarità della media è che la somma degli scarti positivi è uguale alla somma degli scarti negativi.

- La proprietà (d) può essere espressa tramite la seguente relazione di disuguaglianza

  $$(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2 \leq (x_1 - c)^2 + (x_2 - c)^2 + \cdots + (x_n - c)^2.$$

  Se si assume come distanza tra due numeri il quadrato della loro differenza, la somma dei quadrati a sinistra e quella a destra della disuguaglianza possono essere interpretate, rispettivamente, come la distanza delle osservazioni dalla media aritmetica e dalla costante c. Ne segue che la media è il valore "più vicino" alle n osservazioni rispetto a ogni altro possibile numero e pertanto le rappresenta al meglio.

- La proprietà (e) diviene più chiara se la si applica a un problema reale. Il nuovo salario mensile dei lavoratori di un'azienda si ottiene moltiplicando per 1.1 il precedente salario e sommando 20 al risultato. Ci si chiede quale sia il nuovo salario medio mensile sapendo che il vecchio era 1200 euro. In base alla proprietà (e), la nuova media è $20 + 1.1 \cdot 1200 = 1340$.

ESEMPIO 4.1

In un campione di cinque famiglie, sono state rilevate le seguenti ampiezze: 2, 3, 3, 4, 5. Si calcoli la media aritmetica e si verifichino numericamente le proprietà enunciate nella Proposizione 4.1.

Soluzione

La media aritmetica è data da

$$\mu = \frac{2+3+3+4+5}{5} = \frac{17}{5} = 3.4.$$

- Proprietà (a). La media aritmetica, 3.4, è maggiore di 2 e minore di 5, essendo 2 e 5 il minimo e il massimo della serie.

- Proprietà (b). Il prodotto $5 \cdot 3.4 = 17$ è uguale alla somma delle osservazioni $2 + 3 + 3 + 4 + 5 = 17$.

- Proprietà (c). Vale l'identità

  $$(2 - 3.4) + (3 - 3.4) + (3 - 3.4) + (4 - 3.4) + (5 - 3.4)$$
  $$= -1.4 - 0.4 - 0.4 + 0.6 + 1.6 = 0.$$

- Proprietà (d). La somma dei quadrati degli scarti da $\mu = 3.4$ è pari a

  $$(2 - 3.4)^2 + (3 - 3.4)^2 + (3 - 3.4)^2 + (4 - 3.4)^2 + (5 - 3.4)^2 = 5.2,$$

  mentre la somma dei quadrati degli scarti da $c = 2$ è

  $$(2 - 2)^2 + (3 - 2)^2 + (3 - 2)^2 + (4 - 2)^2 + (5 - 2)^2 = 15,$$

  da cui risulta che la prima somma è minore della seconda. Si osserverebbe la medesima disuguaglianza se si assegnasse a c un qualsiasi altro valore diverso da 3.4.

- Proprietà (e). Applicando, per esempio, la trasformazione $y_i = -1 + 2x_i$ ai dati iniziali, si ottiene

  $$3, 5, 5, 7, 9.$$

  La media aritmetica della nuova serie di osservazioni è 5.8. Tale quantità può essere calcolata applicando la trasformazione lineare alla media aritmetica della serie iniziale:

  $$-1 + 2 \cdot 3.4 = 5.8.$$

■

4.3 Media armonica

Un'automobile percorre il tragitto da A a B alla velocità di 80 km/h e poi il tragitto inverso alla velocità di 40 km/h. Si vuole calcolare la velocità media. Indicando con s la distanza tra A e B, con t_1 e t_2 i tempi di percorrenza, con v_1 e v_2 le velocità tenute nel percorso di andata e di ritorno, la velocità media è data dal rapporto tra la distanza complessiva percorsa e il tempo complessivamente impiegato:

$$\frac{2s}{t_1 + t_2} = \frac{2s}{\frac{s}{v_1} + \frac{s}{v_2}} = \frac{2}{\frac{1}{v_1} + \frac{1}{v_2}}.$$

Dunque, la velocità media è il reciproco della media aritmetica dei reciproci delle due velocità. Nell'esempio, si ha

$$\frac{2}{\frac{1}{80} + \frac{1}{40}} = 53.3,$$

un numero ben diverso da 60, che è il valore assunto dalla media aritmetica.

DEFINIZIONE 4.2

La media armonica di una serie di n osservazioni, che si assumono tutte diverse da zero, è data dal rapporto tra n e la somma dei reciproci delle osservazioni:

$$\mu_a = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}. \tag{4.2}$$

Dall'esame della (4.2) si comprende che la media armonica è il reciproco della media dei reciproci delle osservazioni.

PROPOSIZIONE 4.2

La media armonica gode delle proprietà di seguito indicate.

(a) È interna, cioè è maggiore o uguale all'osservazione più piccola e minore o uguale all'osservazione più grande.

(b) È il numero che sostituito a ciascuna delle osservazioni lascia invariata la somma dei reciproci:

$$n\frac{1}{\mu_a} = \sum_{i=1}^{n} \frac{1}{x_i}.$$

(c) Se ogni osservazione della serie è moltiplicata per una costante diversa da zero, la media armonica della nuova serie è uguale alla media armonica delle osservazioni iniziali moltiplicata per la costante.

Le semplici dimostrazioni sono lasciate per esercizio. ■

ESEMPIO 4.2

I tempi (in ore) impiegati dai cinque addetti di un ufficio pubblico per l'espletamento di una pratica sono:

$$4,\ 4.5,\ 5,\ 5.2,\ 6.$$

Si calcoli il tempo medio di esecuzione della pratica.

Soluzione

Applicando la (4.2), si ha

$$\mu_a = \frac{5}{\frac{1}{4} + \frac{1}{4.5} + \frac{1}{5} + \frac{1}{5.2} + \frac{1}{6}} = 4.85.$$

Vale la pena di osservare che 4.85 è il tempo ipotetico che, se fosse osservato dai 5 lavoratori, lascerebbe inalterata la produttività oraria complessiva (la produttività oraria è il reciproco del tempo impiegato per svolgere l'operazione considerata). Invero, per la proprietà (b) della media armonica, vale la relazione

$$\frac{5}{4.85} = \frac{1}{4} + \frac{1}{4.5} + \frac{1}{5} + \frac{1}{5.2} + \frac{1}{6}.$$

■

4.4 Media geometrica

Si immagini di avere un investimento che rende il 10% nel primo anno, il 50% nel secondo e il 25% nel terzo. Qual è il rendimento medio nei tre anni? Non è la media aritmetica dei tre rendimenti. Infatti, dopo il primo anno ogni euro investito diviene 1.10 euro, quantità che, al termine del secondo anno, passa a $1.10 \cdot 1.50$ euro e questa, alla fine del terzo anno, diviene $1.10 \cdot 1.50 \cdot 1.25 = 2.0625$ euro. Ne segue che la risposta appropriata al quesito consiste nel determinare il fattore costante c che sostituito a ciascuno dei fattori del prodotto dà lo stesso risultato. Si tratta di risolvere rispetto a c l'equazione

$$c^3 = 1.10 \cdot 1.50 \cdot 1.25 = 2.0625,$$

da cui $c = \sqrt[3]{2.0625} = 1.273$. In definitiva, il tasso di rendimento medio è 27.3% e non 28.3% come risulterebbe dall'uso della media aritmetica. Ebbene, c è la media geometrica delle quantità 1.10, 1.50 e 1.25.

DEFINIZIONE 4.3

La media geometrica di una serie di n osservazioni, tutte maggiori di 0, è data dalla radice n-esima del prodotto delle osservazioni:

$$\mu_g = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}, \tag{4.3}$$

dove $\prod_{i=1}^{n} x_i = x_1 \cdot x_2 \cdot \cdots \cdot x_n$.

PROPOSIZIONE 4.3

La media geometrica gode delle proprietà di seguito indicate.

(a) È interna, nel senso indicato in precedenza.

(b) È il numero che sostituito a ciascuna delle n osservazioni della serie lascia invariato il prodotto:

$$\mu_g^n = \prod_{i=1}^{n} x_i.$$

(c) Se ciascuna osservazione della serie viene moltiplicata per una costante positiva, la media geometrica della nuova serie è uguale alla media geometrica delle osservazioni iniziali moltiplicata per la costante.

Le semplici dimostrazioni sono lasciate per esercizio. ■

ESEMPIO 4.3

La Tabella 4.1 riporta il prodotto interno lordo (PIL) dell'Italia dal 2014 al 2018 (Fonte: ISTAT, dati in miliardi di euro, prezzi costanti 2010).

Tabella 4.1
Prodotto interno lordo ai prezzi di mercato in Italia, anni 2014-2018.

Anno	2014	2015	2016	2017	2018
PIL	1543	1557	1575	1601	1615

Si calcoli la variazione media percentuale annua dal 2014 al 2018.

Soluzione

Si confronti il prodotto interno lordo di ciascun anno con quello dell'anno precedente mediante i rapporti: $1557/1543 = 1.0091$, $1575/1557 = 1.0116$, $1601/1575 = 1.0165$, $1615/1601 = 1.0087$. La media geometrica dei quattro valori è 1.01. Ciò significa che la variazione percentuale media del prodotto interno lordo dal 2014 al 2018 è stata pari all'1% (si rinvia al Capitolo 8 per i dettagli). ■

Rispetto alla media aritmetica, la media geometrica è meno sensibile alla presenza nella serie di dati anomali (valori eccezionalmente piccoli o eccezionalmente grandi). Così, date le due distribuzioni: A: 1, 5, 10, 20, B: 1, 5, 10, 60, che differiscono solo per il quarto termine, che in B è il triplo del corrispondente termine in A, la media geometrica è 5.6 in A e 7.4 in B, mentre la media aritmetica è 9 in A e 19 in B.

4.5 Media quadratica

In alcuni casi, in presenza di una o più osservazioni con valori negativi, può avere senso non considerarne il segno elevandole al quadrato. La media aritmetica dei valori così trasformati è espressa nel quadrato dell'unità di misura della variabile. Volendo poi ripristinare la scala iniziale, occorre effettuare la trasformazione inversa (estrazione di radice). In questo modo, si ottiene la *media quadratica*.

DEFINIZIONE 4.4

La media quadratica di una serie di n osservazioni è data da

$$\mu_q = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}. \tag{4.4}$$

PROPOSIZIONE 4.4

La media quadratica gode delle proprietà di seguito indicate.

(a) È interna, se tutte le osservazioni della serie sono positive.

(b) È il numero che sostituito a ciascuna delle osservazioni della serie lascia invariata la somma dei quadrati:

$$n\mu_q^2 = \sum_{i=1}^{n} x_i^2.$$

(c) Se ciascuna osservazione della serie viene moltiplicata per una costante positiva, la media quadratica della nuova serie è uguale alla media quadratica delle osservazioni iniziali moltiplicata per la costante.

Le semplici dimostrazioni sono lasciate per esercizio. ■

ESEMPIO 4.4

L'osservazione di quattro pezzi prodotti in un determinato processo industriale ha mostrato le seguenti deviazioni dal peso standard (in chilogrammi): 0.02, −0.01, 0.03, −0.03. L'ufficio qualità dell'azienda vuole determinare una media degli scostamenti dallo standard che prescinda dal loro segno.

Soluzione

Una soluzione del problema consiste nel calcolo della media quadratica:

$$\mu_q = \sqrt{[0.02^2 + (-0.01)^2 + 0.03^2 + (-0.03)^2]/4} = 0.024.$$

■

4.6 Medie di potenza

Le medie analitiche finora descritte sono casi particolari di una classe generale di medie dette *medie di potenza.*

DEFINIZIONE 4.5

La media di potenza di ordine r (con r reale positivo) di una serie di n osservazioni, che si assumono tutte positive, è data da

$$\mu^{(r)} = \sqrt[r]{\frac{1}{n}\sum_{i=1}^{n} x_i^r}.$$

È facile riconoscere che:

$$\begin{array}{lll} \text{per } r = -1, & \mu^{(r)} = \mu_a & \text{(media armonica);} \\ \text{per } r = 1, & \mu^{(r)} = \mu & \text{(media aritmetica);} \\ \text{per } r = 2, & \mu^{(r)} = \mu_q & \text{(media quadratica).} \end{array}$$

Si può dimostrare che:

$$\begin{array}{lll} \text{per } r \to -\infty, & \mu^{(r)} = x_{(1)} & \text{(osservazione minima);} \\ \text{per } r \to 0, & \mu^{(r)} = \mu_g & \text{(media geometrica);} \\ \text{per } r \to +\infty, & \mu^{(r)} = x_{(n)} & \text{(osservazione massima).} \end{array}$$

Si può dimostrare inoltre che $\mu^{(r)}$ è una funzione crescente di r, cosicché, in particolare, sussiste l'ordinamento (per la dimostrazione si veda Pace e Salvan, 1996, pp. 170-171)

$$x_{(1)} \leq \cdots \leq \mu^{(-1)} \leq \mu^{(0)} \leq \mu^{(1)} \leq \mu^{(2)} \leq \cdots \leq x_{(n)},$$

da cui si deduce che:

- le medie di potenza assumono tutte un valore compreso tra l'osservazione più piccola e quella più grande;
- la media armonica è minore o uguale alla media geometrica, la media geometrica è minore o uguale alla media aritmetica, quest'ultima è minore o uguale alla media quadratica.

4.7 Medie analitiche delle distribuzioni di frequenza

Le formule delle medie analitiche finora considerate si riferiscono a serie di osservazioni. È necessario ora adattarle al caso delle distribuzioni di frequenza. A tal fine, si riprenda la tabella contenente la distribuzione di frequenza esaminata nel Capitolo 2.

Tabella 4.2
Distribuzione di frequenza

Modalità	x_1	x_2	...	x_k	Totale
Frequenza	n_1	n_2	...	n_k	n

La Tabella 4.2 indica che ci sono n_1 osservazioni uguali a x_1, n_2 osservazioni uguali a x_2 e così via. Diventa allora semplice adattare le formule introdotte in precedenza alla nuova situazione.

Media aritmetica

La somma di tutte le osservazioni presenti in una distribuzione di frequenza può essere ottenuta sommando gli n_1 numeri uguali a x_1, gli n_2 numeri uguali a x_2 e così via. Ma questo equivale a sommare le quantità $x_1 \cdot n_1, x_2 \cdot n_2, \ldots, x_k \cdot n_k$. La (4.1) diventa allora

$$\mu = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + \cdots + x_k \cdot n_k}{n} = \frac{1}{n} \sum_{i=1}^{k} x_i n_i. \qquad (4.5)$$

La (4.5) può essere scritta in modo equivalente nella forma

$$\mu = x_1 \cdot f_1 + x_2 \cdot f_2 + \cdots + x_k \cdot f_k = \sum_{i=1}^{k} x_i f_i, \qquad (4.6)$$

essendo $f_i = n_i/n$ la frequenza relativa della modalità x_i. In base alla (4.6), per calcolare la media aritmetica è sufficiente conoscere le modalità x_i e le rispettive frequenze relative o percentuali.

ESEMPIO 4.5

Si calcoli la media aritmetica della distribuzione di cui all'Esempio 3.1.

Soluzione

Conviene impostare i calcoli nella forma indicata nella Tabella 4.3:

Tabella 4.3
Calcolo di μ nelle distribuzioni di frequenza.

N. di esami x_i	Frequenza n_i	$x_i \cdot n_i$
0	14	0
1	41	41
2	83	166
3	116	348
4	56	224
5	5	25
Totale	315	804

La somma dei numeri dell'ultima colonna della tabella è il numeratore della media aritmetica. Si ha allora $\mu = 804/315 = 2.55$. Il numero ottenuto va così interpretato: i 315 laureati del gruppo a cui si riferisce la distribuzione hanno sostenuto in media 2.55 esami nel primo anno di università.

■

Medie armonica, geometrica e quadratica

In analogia con quanto visto per la media aritmetica, è facile stabilire che le formule (4.2), (4.3) e (4.4), nel caso delle distribuzioni di frequenza, assumono la forma:

$$\mu_a = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}}, \quad \mu_g = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \cdots \cdot x_k^{n_k}}, \quad \mu_q = \sqrt{\frac{1}{n} \sum_{i=1}^{k} x_i^2 n_i}. \qquad (4.7)$$

Per una proprietà dei logaritmi, la formula della media geometrica può anche essere espressa come

$$\mu_g = \exp\left[\frac{1}{n} \sum_{i=1}^{k} n_i \ln(x_i)\right].$$

ESEMPIO 4.6

Nella Tabella 4.4 è illustrato il procedimento di calcolo delle medie armonica, geometrica e quadratica della distribuzione ipotetica di un campione di famiglie secondo il numero dei componenti.

Tabella 4.4
Calcolo di μ_a, μ_g e μ_q nelle distribuzioni di frequenza.

N. di comp. x_i	Frequenza n_i	n_i/x_i	$n_i \ln(x_i)$	$x_i^2 \cdot n_i$
2	2	1.00	1.39	8
3	4	1.33	4.39	36
4	4	1.00	5.55	64
5	2	0.40	3.22	50
6	2	0.33	3.58	72
7	1	0.14	1.95	49
Totale	15	4.20	20.08	279

Le medie assumono i valori:

$$\mu_a = \frac{15}{4.2} = 3.57, \quad \mu_g = \exp\left(\frac{20.08}{15}\right) = 3.81, \quad \mu_q = \sqrt{\frac{279}{15}} = 4.31.$$

■

È appena il caso di osservare che le espressioni matematiche precedentemente usate per descrivere le proprietà delle medie devono essere adattate al caso delle distribuzioni di frequenza. Ad esempio, le proprietà (b), (c) e (d) della media aritmetica divengono:

$$n\mu = \sum_{i=1}^{k} x_i n_i, \quad \sum_{i=1}^{k} (x_i - \mu) n_i = 0, \quad \sum_{i=1}^{k} (x_i - \mu)^2 n_i = \min_c \sum_{i=1}^{k} (x_i - c)^2 n_i.$$

4.7.1 Il caso delle distribuzioni di frequenza in classi

Ordinariamente, il calcolo di una media analitica in una distribuzione di frequenza con modalità raggruppate in classi fornisce solo un'approssimazione del valore esatto che si otterrebbe se si utilizzassero i dati non raggruppati (ammesso che siano disponibili). Si consideri la media aritmetica e si esamini una distribuzione di frequenza in classi:

Tabella 4.5
Distribuzione di frequenza in classi.

Classe	c_0-c_1	c_1-c_2	...	c_{k-1}-c_k	Totale
Frequenza	n_1	n_2	...	n_k	n

Sia μ_1 la media aritmetica delle n_1 osservazioni appartenenti alla classe c_0-c_1, μ_2 quella delle n_2 osservazioni appartenenti alla classe c_1-c_2 e così via. Se conosciute, queste quantità, chiamate medie di classe, consentono di calcolare senza errore la media aritmetica della distribuzione. Ciò perché il prodotto $\mu_i \cdot n_i$ è esattamente uguale alla somma delle osservazioni appartenenti alla classe c_{i-1}-c_i, $i = 1, 2, \ldots, k$, e quindi

$$\mu_1 \cdot n_1 + \mu_2 \cdot n_2 + \cdots + \mu_k \cdot n_k = \sum_{i=1}^{k} \mu_i n_i$$

è la somma esatta delle osservazioni dell'intera distribuzione. Considerato che le quantità μ_i non sono generalmente conosciute, si assume che esse siano approssimativamente uguali ai valori centrali delle classi:

$$\mu_i \approx \frac{c_{i-1} + c_i}{2}.$$

Posto $\bar{x}_i = (c_{i-1} + c_i)/2$, il valore approssimato della media aritmetica è dato da

$$\mu = \frac{1}{n} \sum_{i=1}^{k} \bar{x}_i n_i.$$

Si può dimostrare che $\bar{x}_i = \mu_i$ se le unità sono uniformemente distribuite all'interno della classe (si veda il Paragrafo 4.12).

ESEMPIO 4.7

Si riprenda la distribuzione in Tabella 2.4 dell'Esempio 2.4 e si calcoli la media aritmetica utilizzando la distribuzione di frequenza supponendo di non disporre più dei dati individuali con cui è stata costruita.

Soluzione

In base alla Tabella 4.6, che contiene il dettaglio dei calcoli, la media aritmetica è

$$\mu = \frac{6682}{140} = 47.73.$$

Tabella 4.6
Calcolo di μ in una distribuzione di frequenza in classi.

Lunghezza avambraccio	Frequenza n_i	Val. centrale $\bar{x}_i$	$\bar{x}_i \cdot n_i$
41-42	5	41.5	207.5
43-44	17	43.5	739.5
45-46	25	45.5	1137.5
47-48	35	47.5	1662.5
49-50	32	49.5	1584.0
51-52	20	51.5	1030.0
53-54	6	53.5	321.0
Totale	140		6682.0

La media così ottenuta è un'approssimazione della media esatta, che può essere calcolata utilizzando direttamente la serie di osservazioni data nell'Esempio 2.4. Ciò dipende dal fatto che le medie di classe vengono approssimate con i valori centrali. Si consideri, ad esempio, che la media dei cinque valori appartenenti alla classe [41, 42],

41, 41, 41, 42, 42,

è 41.4, e che questa quantità viene approssimata con il valore centrale della classe, 41.5. Considerazioni analoghe valgono per le altre classi. ■

È immediato estendere quanto detto per la media aritmetica alle altre medie analitiche: basta sostituire nelle (4.7) a x_i i valori centrali di classe $\bar{x}_i$.

ESEMPIO 4.8

Nella Tabella 4.7 viene presentato il calcolo della media armonica e della media geometrica per la distribuzione di cui all'Esempio 4.7. Per la media quadratica l'impostazione dei calcoli è analoga.

Tabella 4.7
Calcolo di μ_a e μ_g nelle distribuzioni di frequenza in classi.

Lunghezza avambraccio	Frequenza n_i	Val. centrale $\bar{x}_i$	$n_i/\bar{x}_i$	$n_i \ln(\bar{x}_i)$
41-42	5	41.5	0.120	18 628
43-44	17	43.5	0.391	64 137
45-46	25	45.5	0.549	95 443
47-48	35	47.5	0.737	135 126
49-50	32	49.5	0.646	124 863
51-52	20	51.5	0.388	78 832
53-54	6	53.5	0.112	23 878
Totale	140		2.943	540 907

Si ottiene perciò:

$$\mu_a = 140/2.943 = 47.54, \quad \mu_g = \exp(540.907/140) = 47.64.$$

■

ESEMPIO 4.9

Si calcoli la media aritmetica della distribuzione di frequenza dei comuni di una certa regione per superficie esposta nella Tabella 4.8. La terza colonna della tabella contiene la superficie totale dei comuni appartenenti alla singola classe: ad esempio, la quantità 6029 è la superficie totale dei comuni della prima classe.

Tabella 4.8
Distribuzione di frequenza in classi con totali di classe noti.

Superificie (ettari)	N. di comuni n_i	Superficie totale t_i
≤ 1000	8	6 029
1001-2000	14	21 024
2001-4000	27	78 445
4001-6000	9	44 871
6001-10 000	8	57 034
10 001-25 000	8	118 944
Totale	74	326 347

Soluzione
Dato che la somma dei valori della terza colonna della Tabella 4.8 è la superficie complessiva dei comuni, la media aritmetica si ottiene immediatamente dal rapporto

$$\mu = \frac{326\,347}{74} = 4410.1.$$

■

4.8 Medie ponderate

Vi sono situazioni in cui è ragionevole stabilire che i dati contribuiscano in modo differenziato a determinare il valore della media. Si pensi, ad esempio, a un procedimento di selezione del personale in cui i candidati sono valutati sulla base di tre prove: un tema su un argomento specifico, un test a risposta multipla e un colloquio motivazionale. La commissione di valutazione può ritenere di dare maggiore importanza al tema e al colloquio motivazionale che non al test a risposta multipla. Il punteggio medio del singolo candidato non potrà essere allora calcolato con la media aritmetica semplice dei punteggi ottenuti nelle tre prove (somma dei punteggi divisa per tre): occorrerà utilizzare la *media aritmetica ponderata* definita come segue.

Media aritmetica ponderata

DEFINIZIONE 4.6

Siano $x_1, x_2, \ldots, x_k$ le modalità di una variabile quantitativa e $w_1, w_2, \ldots, w_k$ i rispettivi pesi. La media aritmetica ponderata è allora data da

$$\mu = \frac{x_1 \cdot w_1 + x_2 \cdot w_2 + \cdots + x_k \cdot w_k}{w_1 + w_2 + \cdots + w_k} = \frac{1}{W}\sum_{i=1}^{k} x_i w_i, \qquad (4.8)$$

dove

$$W = w_1 + w_2 + \cdots + w_k.$$

Le proprietà indicate nella Proposizione 4.1 valgono, con i necessari adattamenti, anche per la media aritmetica ponderata. Ad esempio, le proprietà (c) e (d) divengono:

$$\sum_{i=1}^{k}(x_i - \mu)w_i = 0, \quad \sum_{i=1}^{k}(x_i - \mu)^2 w_i = \min_c \sum_{i=1}^{k}(x_i - c)^2 w_i.$$

La media aritmetica ponderata assume lo stesso valore della media aritmetica semplice se i pesi w_i sono uguali tra loro (si veda l'Esercizio 4.11*).

ESEMPIO 4.10

Nella Tabella 4.9 sono esposti i voti riportati da uno studente universitario in cinque discipline nel primo anno del corso di laurea in economia. Per ogni disciplina è dato anche il numero dei crediti.

Tabella 4.9
Voti riportati da uno studente in cinque discipline e relativi crediti.

Disciplina	A	B	C	D	E
Voto	25	26	28	28	30
N. di crediti	5	5	10	5	12

Si calcoli la media aritmetica ponderata dei voti utilizzando come pesi i crediti delle discipline.

Soluzione

In forza della (4.8), si ha

$$\mu = \frac{25 \cdot 5 + 26 \cdot 5 + 28 \cdot 10 + 28 \cdot 5 + 30 \cdot 12}{5 + 5 + 10 + 5 + 12} = 27.97.$$

La media aritmetica semplice è pari a 27.4, quantità che rappresenta in modo erroneo la distribuzione data, perché non tiene conto del fatto che i voti maggiori sono stati ottenuti nelle discipline con i crediti più elevati (discipline C ed E). ■

La Figura 4.1 contiene sei diagrammi in cui alle cinque modalità (i valori riportati sull'asse x) è assegnato un diverso sistema di pesi, rappresentati dalle altezze delle linee verticali. I grafici (b), (c) e (d) mostrano che, maggiore è il peso di una data modalità, più la media ponderata tende a essere vicina a tale modalità. Quando invece i pesi differiscono solo leggermente, come in (a), la media ponderata si avvicina alla media aritmetica semplice, che è pari a 170. Interessante è anche il diagramma (e), dove i pesi differiscono notevolmente, ma quelli delle due osservazioni estreme si bilanciano a vicenda.

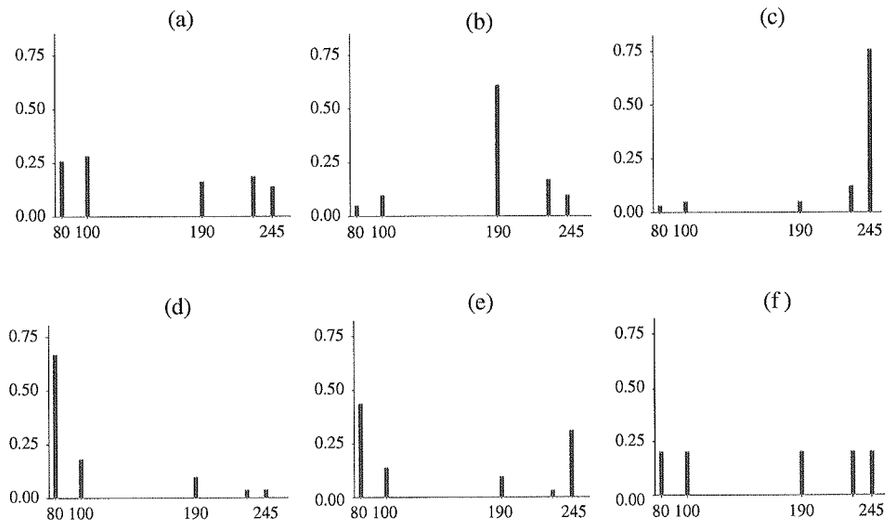


Figura 4.1
Effetto dei pesi sul valore della media ponderata. La media ponderata assume i valori: (a) $\mu = 153$, (b) $\mu = 190$, (c) $\mu = 230$, (d) $\mu = 105$, (e) $\mu = 150$, (f) $\mu = 170$.

Altre medie analitiche ponderate

Per completezza, si riportano di seguito le versioni ponderate delle formule delle medie armonica, geometrica e quadratica:

$$\mu_a = \frac{W}{\sum_{i=1}^{k} \frac{w_i}{x_i}}, \quad \mu_g = \sqrt[W]{x_1^{w_1} \cdot x_2^{w_2} \cdot \cdots \cdot x_k^{w_k}}, \quad \mu_q = \sqrt{\frac{1}{W} \sum_{i=1}^{k} x_i^2 w_i},$$

dove $W = w_1 + w_2 + \cdots + w_k$. Queste formule si ottengono dalle (4.7) sostituendo n_i con il peso w_i ed n con W.

ESEMPIO 4.11

Un'automobile percorre tre tratti di autostrada, della lunghezza di 50, 70 e 85 km, alle velocità di 100, 120 e 130 km/h. Si determini la velocità media.

Soluzione

Bisogna usare la media armonica ponderata:

$$\mu_a = \frac{50 + 70 + 85}{\frac{50}{100} + \frac{70}{120} + \frac{85}{130}} = 118.01.$$

■

4.9 Mediana

La *mediana* è una media largamente utilizzata. Diversamente da quanto visto per le medie analitiche, il suo calcolo non viene descritto da una formula, ma da una procedura.

DEFINIZIONE 4.7

Data la serie di osservazioni $x_1, x_2, \ldots, x_n$, sia $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, dove $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, la corrispondente sequenza non decrescente. Se n è dispari, la mediana, m, è il termine che occupa la posizione centrale, $(n+1)/2$, di tale sequenza. Se n è pari, la mediana è definita come la media aritmetica dei termini che occupano le due posizioni centrali, $n/2$ e $n/2 + 1$, della stessa sequenza. In simboli, si ha

$$m = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ è dispari,} \\ \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right], & \text{se } n \text{ è pari.} \end{cases} \tag{4.9}$$

ESEMPIO 4.12

Si consideri il numero di posti letto disponibili nei sette ospedali di una grande città:

212, 320, 180, 410, 374, 215, 538.

Si calcoli la mediana.

Soluzione

Per prima cosa, le osservazioni vanno ordinate in senso non decrescente:

180, 212, 215, 320, 374, 410, 538.

La posizione centrale della sequenza non decrescente è $(n+1)/2 = (7+1)/2 = 4$. L'osservazione che occupa la quarta posizione nella graduatoria è 320. Per la (4.9), la mediana è allora

$$m = x_{(4)} = 320.$$

Se gli ospedali fossero otto e l'ottava osservazione fosse 290, la sequenza non decrescente diverrebbe:

$$180, 212, 215, 290, 320, 374, 410, 538,$$

in cui le posizioni centrali sono

$$\frac{n}{2} = \frac{8}{2} = 4 \quad \text{e} \quad \frac{n}{2} + 1 = \frac{8}{2} + 1 = 5.$$

In definitiva,

$$m = \frac{x_{(4)} + x_{(5)}}{2} = \frac{290 + 320}{2} = 305.$$

■

La mediana, come termine centrale o media dei due termini centrali della sequenza non decrescente delle osservazioni, è una misura dell'ordine di grandezza di queste ultime.

PROPOSIZIONE 4.5

La mediana gode delle proprietà di seguito indicate.

(a) È interna.

(b) La somma dei valori assoluti degli scarti delle osservazioni da una costante c è minima quando c è uguale alla mediana:

$$\sum_{i=1}^{n} |x_i - m| = \min_c \sum_{i=1}^{n} |x_i - c|. \tag{4.10}$$

La dimostrazione della proprietà (b) è contenuta nel Paragrafo A.2.2 dell'Appendice A.

■

Se si assume come misura della distanza tra due numeri la differenza in valore assoluto, la (4.10) indica che la mediana è il valore "più vicino" alle osservazioni rispetto a qualsiasi altro numero. Analogamente alla media aritmetica, la mediana costituisce il "centro" delle osservazioni e, come tale, le rappresenta sinteticamente.

ESEMPIO 4.13

Si verifichi empiricamente la proprietà (b) della mediana utilizzando i dati dell'Esempio 4.12 che, per comodità, vengono riscritti di seguito:

$$212,\ 320,\ 180,\ 410,\ 374,\ 215,\ 538.$$

Soluzione

La somma dei valori assoluti degli scarti dalla mediana assume il valore

$$|212 - 320| + |320 - 320| + \cdots + |538 - 320| = 715.$$

Posto, ad esempio, $c = 200$, la somma dei valori assoluti degli scarti da questo numero è uguale a

$$|212 - 200| + |320 - 200| + \cdots + |538 - 200| = 889.$$

Come si vede, la prima somma è minore della seconda, e la medesima disuguaglianza si realizzerebbe se si assegnasse a c un qualsiasi altro valore diverso da 320.

■

4.10 Quartili e quantili

Accanto alla mediana, si possono introdurre altri valori medi, legati ai termini che occupano particolari posizioni nella sequenza non decrescente delle osservazioni. Si inizierà con i quartili, le tre quantità che suddividono la sequenza ordinata delle osservazioni in quattro parti nel modo descritto di seguito.

DEFINIZIONE 4.8

Data la serie di osservazioni $x_1, x_2, \ldots, x_n$, sia $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, dove $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, la corrispondente sequenza non decrescente. L'l–esimo quartile, $l = 1, 2, 3$, è definito nel modo seguente. Siano $h - 1$ e h due interi consecutivi tali che

$$h - 1 \leq n \cdot (l/4) < h.$$

Allora l'l-esimo quartile è dato da

$$q_l = \begin{cases} x_{(h)}, & \text{se } n \cdot (l/4) > h - 1, \\ \dfrac{1}{2}[x_{(h-1)} + x_{(h)}], & \text{se } n \cdot (l/4) = h - 1. \end{cases} \tag{4.11}$$

Naturalmente, il secondo quartile coincide con la mediana.

Quando n è multiplo di quattro, il primo quartile è l'elemento separatore tra il primo quarto delle osservazioni ordinate e i restanti tre quarti; il secondo quartile è l'elemento separatore tra la prima metà delle osservazioni ordinate e la restante metà; il terzo quartile è l'elemento separatore tra i primi tre quarti delle osservazioni ordinate e il restante quarto. Di conseguenza, esattamente un quarto delle osservazioni assume un valore minore o uguale a q_1, esattamente la metà delle osservazioni presenta un valore minore o uguale a q_2 ecc.[2]

ESEMPIO 4.14

Si determinino i quartili delle osservazioni che seguono:

$$3,\ 6,\ 7,\ 9,\ 12,\ 13,\ 16,\ 18.$$

Soluzione

In base alla definizione (4.11), si ottiene:

- primo quartile: poiché $n \cdot (1/4) = 8 \cdot (1/4) = 2$, si ha

$$q_1 = \frac{x_{(2)} + x_{(3)}}{2} = \frac{6 + 7}{2} = 6.5;$$

- secondo quartile: poiché $n \cdot (2/4) = 8 \cdot (2/4) = 4$, si ha

$$q_2 = \frac{x_{(4)} + x_{(5)}}{2} = \frac{9 + 12}{2} = 10.5;$$

[2] Se n non è multiplo di 4, le proposizioni corrette sono: almeno un quarto dei termini della serie assume un valore minore o uguale a q_1, almeno la metà dei termini della serie assume un valore minore o uguale a q_2 ecc.

- terzo quartile: poiché $n \cdot (3/4) = 8 \cdot (3/4) = 6$, si ha

$$q_3 = \frac{x_{(6)} + x_{(7)}}{2} = \frac{13 + 16}{2} = 14.5.$$

Come si può osservare, i tre quartili suddividono la sequenza non decrescente delle osservazioni esattamente in quattro parti di uguale numerosità. ■

ESEMPIO 4.15

Le quotazioni alla Borsa di Milano di un titolo azionario in 10 sedute successive sono state:

12.8, 13.0, 13.4, 13.4, 13.6, 13.5, 13.6, 13.7, 14.1, 13.8.

Si calcolino i quartili.

Soluzione

L'ordinamento delle osservazioni in senso non decrescente dà:

12.8, 13.0, 13.4, 13.4, 13.5, 13.6, 13.6, 13.7, 13.8, 14.1.

In base alla (4.11), si ottiene:

- primo quartile: $n \cdot (1/4) = 10 \cdot (1/4) = 2.5$, da cui

$$q_1 = x_{(3)} = 13.4;$$

- secondo quartile: $n \cdot (1/4) = 10 \cdot (2/4) = 5$, da cui

$$q_2 = m = \frac{x_{(5)} + x_{(6)}}{2} = \frac{13.5 + 13.6}{2} = 13.55;$$

- terzo quartile: $n \cdot (3/4) = 10 \cdot (3/4) = 7.5$, da cui

$$q_3 = x_{(8)} = 13.7.$$

■

L'idea dei quartili può essere generalizzata senza difficoltà. Ad esempio, si possono definire i nove decili come le quantità che suddividono la sequenza non decrescente delle osservazioni in dieci parti, oppure i 99 centili (o percentili) e così via. Queste quantità appartengono alla categoria generale dei quantili.

La procedura per il calcolo dei quantili è analoga a quella illustrata in dettaglio per i quartili. Per esempio, per calcolare l'l-esimo decile ($l = 1, 2, \ldots, 9$), d_l, si devono individuare due interi consecutivi, $h - 1$ e h, tali che

$$h - 1 \leq n \cdot \frac{l}{10} < h.$$

Dopodiché, d_l è uguale a $x_{(h)}$ o a $[x_{(h-1)} + x_{(h)}]/2$ a seconda che $n \cdot (l/10)$ sia maggiore o uguale a $h - 1$.

ESEMPIO 4.16

Si determini il settimo decile di una serie di osservazioni composta da 36 termini.

Soluzione

Poiché

$$n \cdot \frac{7}{10} = 36 \cdot \frac{7}{10} = 25.2$$

è maggiore di 25 e minore di 26, il settimo decile è dato dal termine che occupa il posto 26 nella sequenza non decrescente delle osservazioni. ■

4.11 Mediana, quartili e quantili delle distribuzioni di frequenza

Come è stato fatto per le medie analitiche, è utile illustrare il procedimento di calcolo della mediana, dei quartili e dei quantili quando i dati sono nella forma di distribuzioni di frequenza.

Mediana

Si determinano le frequenze cumulate $N_1, N_2, \ldots, N_k$ e si identificano le due frequenze cumulate consecutive, N_{h-1} e N_h, tali che

$$N_{h-1} \leq n \cdot \frac{1}{2} < N_h.$$

La mediana è data da[3]

$$m = \begin{cases} x_h, & \text{se } n/2 > N_{h-1}, \\ (x_{h-1} + x_h)/2, & \text{se } n/2 = N_{h-1}, \end{cases} \tag{4.12}$$

dove x_{h-1} e x_h sono le modalità di X in corrispondenza delle quali le frequenze cumulate sono N_{h-1} e N_h, rispettivamente. È chiaro che N_{h-1} va posta uguale a 0 quando $h = 1$.

ESEMPIO 4.17

Si calcoli la mediana della distribuzione di frequenza che segue:

N. di esami x_i	Frequenza n_i	Frequenza cum. N_i
0	14	14
1	41	55
2	83	138
3	116	254
4	56	310
5	5	315
Totale	315	–

Tabella 4.10 Distribuzione di frequenza cumulata tratta dall'Esempio 4.5.

[3] Questa definizione deriva dall'applicazione della (4.11) a una distribuzione di frequenza. Infatti, N_{h-1} e N_h indicano la massima posizione occupata, rispettivamente, da x_{h-1} e x_h nella sequenza non decrescente che si ottiene riscrivendo le modalità in forma di serie di osservazioni, cioè disponendo in successione n_1 numeri pari a x_1, n_2 numeri pari a x_2 ecc.

Soluzione

Poiché $n \cdot (1/2) = 315/2 = 157.5$, si ha

$$N_3 < 157.5 < N_4,$$

da cui, in forza della (4.12), si ottiene

$$m = x_4 = 3.$$

■

Quartili

Il calcolo dei quartili segue una procedura analoga. Una volta determinate le frequenze cumulate, vengono individuate le due frequenze cumulate consecutive, N_{h-1} e N_h, tali che

$$N_{h-1} \leq n \cdot \frac{l}{4} < N_h,$$

con $l = 1, 2, 3$; dopodiché

$$q_l = \begin{cases} x_h, & \text{se } n \cdot (l/4) > N_{h-1}, \\ (x_{h-1} + x_h)/2, & \text{se } n \cdot (l/4) = N_{h-1}. \end{cases} \tag{4.13}$$

ESEMPIO 4.18

Si calcolino il primo e il terzo quartile della distribuzione di frequenza considerata nell'esempio precedente.

Soluzione

- Primo quartile: la quantità $n \cdot (1/4) = 315 \cdot (1/4) = 78.75$ soddisfa la disuguaglianza

  $$N_2 < 78.75 < N_3,$$

  da cui $q_1 = x_3 = 2$.
- Terzo quartile: la quantità $n \cdot (3/4) = 315 \cdot (3/4) = 236.25$ è tale che

  $$N_3 < 236.25 < N_4,$$

  da cui $q_3 = x_4 = 3$.

■

I decili, i centili e più in generale i quantili di una distribuzione di frequenza si definiscono in modo analogo sostituendo nella (4.13) a $l/4$ la frazione appropriata: $l/10$ per i decili, $l/100$ per i centili e così via.

Con la Figura 4.2 si vuole mettere in luce la relazione tra i quantili e la funzione di ripartizione. Il grafico è basato sulla distribuzione di frequenza esaminata nell'Esempio 4.5 e mostra come i quartili possano essere determinati a partire dalla funzione di ripartizione.

Ad esempio, per trovare q_1 si traccia la linea orizzontale che interseca l'asse delle ordinate in $F(x) = 0.25$ e si individua il valore sull'asse delle x in corrispondenza del quale la linea interseca la funzione di ripartizione, 2 in questo caso. Se la linea orizzontale interseca la funzione di ripartizione in un tratto in cui essa è costante, il quartile deve essere definito come il punto medio dell'intervallo corrispondente in x. Procedendo in modo simile, si trova che $q_2 = q_3 = 3$.

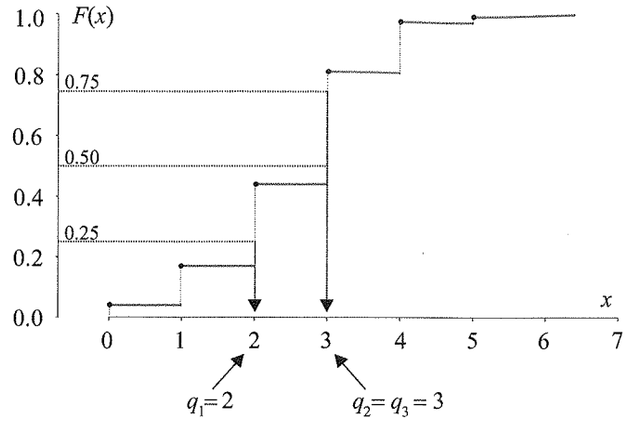


Figura 4.2
Funzione di ripartizione e quantili per una distribuzione di frequenza.

4.11.1 Il caso delle distribuzioni di frequenza in classi

Quando le modalità sono raggruppate in classi, il calcolo della mediana avviene in due fasi: si individua dapprima la classe mediana (cioè la classe in cui cade la mediana); poi, all'interno di questa, si calcola il valore della mediana. Per la prima operazione, si applica la stessa regola stabilita in precedenza per le distribuzioni di frequenza. Individuata la classe mediana, il calcolo della mediana viene generalmente effettuato assumendo l'uniforme distribuzione delle unità all'interno della classe.

PROPOSIZIONE 4.6

Siano $c_{h(inf)}$ e $c_{h(sup)}$ i limiti inferiore e superiore della classe mediana. Assumendo l'uniforme distribuzione delle unità all'interno della classe, la mediana è data da

$$m = c_{h(inf)} + \frac{n/2 - N_{h-1}}{N_h - N_{h-1}}[c_{h(sup)} - c_{h(inf)}], \tag{4.14}$$

dove N_{h-1} e N_h sono le frequenze cumulate in corrispondenza di $c_{h(inf)}$ e $c_{h(sup)}$. La quantità N_{h-1} va posta uguale a zero quando $h = 1$.

La dimostrazione della (4.14) è contenuta nel Paragrafo A.2.2 dell'Appendice A. Si osservi che nell'applicazione della (4.14) vanno utilizzati gli estremi reali di classe.

■

La Proposizione (4.6) può essere estesa al caso di un quantile qualsiasi. Basta sostituire $n/2$ con $n \cdot (l/m)$ nella (4.14) e definire coerentemente i nuovi limiti di classe. Ad esempio, il primo quartile è dato da

$$q_1 = c_{h(inf)} + \frac{n/4 - N_{h-1}}{N_h - N_{h-1}}[c_{h(sup)} - c_{h(inf)}],$$

dove $c_{h(inf)}$ e $c_{h(sup)}$ sono i limiti inferiore e superiore della classe del primo quartile (la classe in cui si trova il primo quartile), mentre N_{h-1} e N_h sono le frequenze cumulate in $c_{h(inf)}$ e $c_{h(sup)}$.

ESEMPIO 4.19

Si consideri la distribuzione nella Tabella 4.11 relativa a un campione di nati vivi di sesso maschile per peso alla nascita. Si determinino i quartili della distribuzione.

Tabella 4.11
Nati vivi maschi secondo il peso.

Peso (g)	Frequenza	Freq. cum.	Peso (g)	Frequenza	Freq. cum.
< 500	39	39	3250-3499	58 093	162 412
500-999	548	587	3500-3749	50 640	213 052
1000-1499	1 149	1 736	3750-3999	30 051	243 103
1500-1999	2 020	3 756	4000-4499	20 257	263 360
2000-2499	7 691	11 447	4500-4999	2 369	265 729
2500-2749	12 772	24 219	5000-5499	192	265 921
2750-2999	26 922	51 141	5500-5999	27	265 948
3000-3249	53 178	104 319	≥ 6000	13	265 961
			Totale	265 961	–

Soluzione

- Primo quartile: poiché $n \cdot (1/4) = 265\,961 \cdot (1/4) = 66\,490.25$, si ha

$$N_7 < 66\,490.25 < N_8,$$

da cui si evince che il primo quartile cade nell'ottava classe. Ne segue che

$$q_1 = 2999.5 + \frac{66\,490.25 - 51\,141}{104\,319 - 51\,141}(3249.5 - 2999.5) = 3071.66.$$

- Secondo quartile: poiché $n \cdot (2/4) = 265\,961 \cdot (2/4) = 132\,980.5$, si ha

$$N_8 < 132\,980.5 < N_9,$$

da cui si evince che il secondo quartile cade nella nona classe. Ne segue che

$$q_2 = 3249.5 + \frac{132\,980.5 - 104\,319}{162\,412 - 104\,319} \cdot (3499.5 - 3249.5) = 3372.84.$$

- Terzo quartile: poiché $n \cdot (3/4) = 265\,961 \cdot (3/4) = 199\,470.75$, si ha

$$N_9 < 199\,470.75 < N_{10},$$

da cui si evince che il terzo quartile cade nella decima classe. Ne segue che

$$q_3 = 3499.5 + \frac{199\,470.75 - 162\,412}{213\,052 - 162\,412} \cdot (3749.5 - 3499.5) = 3682.45.$$

■

Analogamente a quanto fatto con la Figura 4.2, la Figura 4.3 riporta il grafico della funzione di ripartizione della distribuzione di frequenza in classi vista nell'Esempio 3.5. Con essa viene mostrata la procedura per individuare i quantili a partire da $F(x)$. Ad esempio, per trovare q_2 si traccia la linea orizzontale che interseca l'asse delle ordinate in $F(x) = 0.5$ e si individua il valore sull'asse delle ascisse in corrispondenza del quale la linea interseca la curva, 47.8 in questo caso. Procedendo allo stesso modo, si trova che $q_1 = 45.5$ e $q_3 = 49.9$. Chiaramente, questa procedura permette di determinare qualsiasi quantile.

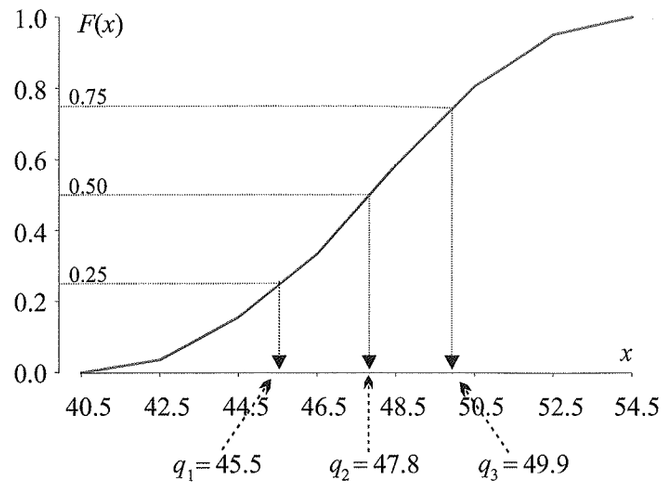


Figura 4.3
Funzione di ripartizione e quantili per una distribuzione di frequenza in classi.

4.12 Valore centrale

Il *valore centrale* è definito come segue.

DEFINIZIONE 4.9

Data la serie di osservazioni $x_1, x_2, \ldots, x_n$, sia $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, dove $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, la corrispondente sequenza non decrescente. Il valore centrale è dato da

$$\mu_c = \frac{x_{(1)} + x_{(n)}}{2}.$$

La definizione può essere generalizzata, con opportuni aggiustamenti, alle distribuzioni di frequenza e alle distribuzioni di frequenza in classi. Il valore centrale coincide con la media aritmetica quando le osservazioni formano una progressione aritmetica (si veda l'Esercizio 4.17*).

ESEMPIO 4.20

Data la seguente distribuzione delle superfici degli appartamenti di un piccolo condominio:

124, 55, 110, 117, 124, 55, 110, 130, 42, 110,

il valore centrale è dato da

$$\mu_c = \frac{42 + 130}{2} = 86.$$

■

4.13 Moda

Diversamente da quanto fin qui fatto, si definirà la moda a partire da una distribuzione di frequenza, per la ragione che sarà subito evidente.

DEFINIZIONE 4.10

La moda di una distribuzione di frequenza è la modalità che presenta la frequenza più alta.

ESEMPIO 4.21

La moda è A nell'Esempio 2.1 e 1 nell'Esempio 2.2. ■

È evidente che in una distribuzione di frequenza a modalità singole si possono avere due o più mode, perché possono esservi due o più modalità con la stessa frequenza massima. In questo caso, si parla di distribuzioni plurimodali (distribuzioni bimodali quando le mode sono due). Talvolta, questa terminologia viene utilizzata in senso estensivo quando vi sono una o più modalità con frequenze che, pur essendo inferiori a quella della moda, sono sensibilmente più alte rispetto a quelle delle altre modalità.

Quando si ha a che fare con una distribuzione di frequenza in classi, si definisce la classe modale come la classe con la più alta densità di frequenza.

ESEMPIO 4.22

La classe modale della distribuzione di frequenza analizzata nell'Esempio 2.7 è 47-48. ■

Quanto detto sopra a proposito delle distribuzioni plurimodali vale anche, *mutatis mutandis*, per il caso delle distribuzioni di frequenza in classi.

4.14 Medie e dati anomali

Come già accennato nel Paragrafo 2.5, un insieme di dati può contenere osservazioni eccezionalmente elevate e/o basse, i cosiddetti dati anomali. Ci si può chiedere come e fino a che punto una determinata media sia sensibile alla presenza di dati anomali. Le medie analitiche, essendo funzioni di tutte le osservazioni, risentono della presenza di questi valori, anche se in misura diversa. Ad esempio, la media quadratica è fortemente influenzata da valori eccezionalmente elevati, mentre la media armonica è fortemente influenzata da valori eccezionalmente piccoli, ma è meno influenzata da valori eccezionalmente grandi. Per la media geometrica, i valori anomali piccoli sono molto più influenti di quelli grandi. Anche la media aritmetica risente di questo fenomeno. Quando una media risente poco della presenza di valori anomali viene qualificata come robusta.

Un metodo pensato per rendere robuste le medie è il troncamento. Consiste nel rimuovere una piccola percentuale dei valori più grandi e più piccoli prima di calcolare la media. Si consideri, ad esempio, la serie di osservazioni $x_1, x_2, \ldots, x_n$ e la corrispondente sequenza non decrescente $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. Se si decide di rimuovere l'osservazione più piccola e quella più grande, la media troncata è data da

$$\text{Media troncata} = \frac{x_{(2)} + x_{(3)} + \cdots + x_{(n-2)} + x_{(n-1)}}{n-2}.$$

Una variante della media troncata, nota come media troncata con sostituzioni, consiste nel sostituire l'osservazione più piccola e quella più grande con le quantità, rispettivamente, più piccola e più grande tra le rimanenti osservazioni. In tal caso, la formula precedente diventa

$$\text{Media troncata con sostituzioni} = \frac{2x_{(2)} + x_{(3)} + \cdots + x_{(n-2)} + 2x_{(n-1)}}{n}.$$

Per quanto riguarda le medie di posizione, la mediana è certamente una media robusta, non essendo influenzata dai valori estremi.

ESEMPIO 4.23

Si considerino le seguenti 11 osservazioni:

$$3,\ 4,\ 4,\ 5,\ 7,\ 8,\ 8,\ 10,\ 11,\ 11,\ 1000.$$

Qui il valore 1000 è chiaramente un valore anomalo. Per le diverse medie analitiche e di posizione si ottengono i valori:

$$\mu = 97.4,\ \mu_a = 6.4,\ \mu_g = 10.2,\ \mu_q = 301.6,$$
$$\mu_c = 501.5,\ m = 8,\ q_1 = 4,\ q_3 = 11.$$

La media troncata e la media troncata con sostituzioni assumono i valori 7.1 e 7.5. Da questi numeri emerge chiaramente l'impatto forte del valore anomalo sulle medie aritmetica e quadratica e sul valore centrale; minore è l'impatto sulla media armonica e su quella geometrica. Viceversa, la mediana, insieme alle altre due medie robuste, non risente della presenza di valori anomali. ■

4.15 La scelta della media

Cominciando con le medie analitiche, è appena il caso di ricordare che la media aritmetica è quella maggiormente utilizzata. Le medie armonica e geometria vengono impiegate in situazioni specifiche, in base alla tipologia dei dati.

La media armonica è appropriata quando i reciproci delle osservazioni possono essere interpretati come le osservazioni di una variabile correlata. Ciò accade, ad esempio, quando la variabile di interesse è il tempo (in ore) utilizzato da un impiegato nell'esecuzione di uno specifico compito. In questo caso, i reciproci delle osservazioni coincidono con le produttività orarie degli impiegati.

La media geometrica è appropriata quando si studia il tasso di variazione di un fenomeno nel tempo, come nell'Esempio 4.3. È consigliabile anche quando la trasformazione logaritmica dei dati (con il logaritmo naturale) e il successivo raggruppamento in classi danno luogo a un istogramma con forma a "campana" (simile a una distribuzione normale; si veda Paragrafo 6.3). Questo è il motivo per cui è frequentemente utilizzata in biologia e nelle analisi che coinvolgono dati ambientali.

La media quadratica è usata molto raramente come sintesi dei dati, ma è impiegata nella definizione di altri indicatori statistici riguardanti altri aspetti dei dati, come si vedrà nei prossimi capitoli (variabilità, dipendenza).

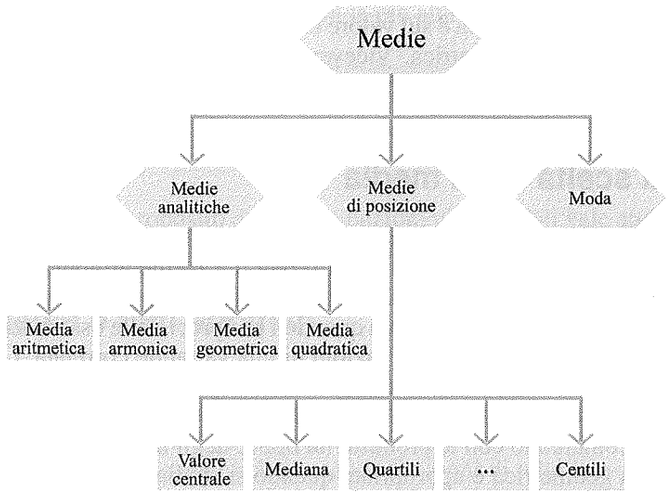
Tra le medie di posizione, la mediana, insieme alle altre due medie robuste introdotte nel paragrafo precedente, è un valido sostituto della media aritmetica quando si sospetta che possano essere presenti valori anomali nei dati (si veda l'Esempio 4.23). La mediana è spesso utilizzata insieme ad altri quantili per descrivere più in dettaglio le distribuzioni di frequenza. Così, ad esempio, quartili, quintili o decili vengono utilizzati dagli uffici statistici nazionali per riassumere la distribuzione del reddito del Paese. Infatti, sapere, ad esempio, che i quartili della distribuzione del reddito familiare (in euro) sono $q_1 = 20\,340$, $q_2 = 35\,480$ e $q_3 = 85\,480$ è molto più informativo che conoscere solo la media aritmetica.

Queste grandezze consentono di farsi un'idea più circostanziata della distribuzione: dicono che il 25% delle famiglie ha un reddito inferiore o uguale a 20 340, che il 25% delle famiglie ha un reddito compreso tra 20 340 e 35 480, e così via. Naturalmente, i quintili o i decili fornirebbero informazioni più affinate sulla distribuzione.

Vale la pena notare che i tre quartili, insieme alle osservazioni estreme, ossia minimo e massimo, costituiscono i cosiddetti cinque valori di sintesi utilizzati per rappresentare un insieme di dati mediante il cosiddetto grafico a scatola (si veda il Paragrafo 7.2.1).

La moda, che è un po' la cenerentola tra le medie, è l'unica calcolabile quando la variabile di interesse è nominale. Riportare il valore modale o la classe modale, insieme ad altre medie, può essere utile soprattutto quando la sua frequenza (o densità di frequenza nel caso delle classi) è molto più alta delle altre frequenze (o densità di frequenza). In tal caso, la moda ha una forte capacità rappresentativa. Si pensi, ad esempio, al settore manifatturiero: i produttori di abbigliamento femminile saranno chiaramente interessati alla classe modale della distribuzione di frequenza secondo la statura o la taglia delle scarpe delle giovani donne.

Figura 4.4
Quadro sinottico sulle medie.



Riassunto

Le medie rappresentano uno degli strumenti principali della statistica descrittiva. Ad eccezione della moda, tutte le medie discusse in questo capitolo possono essere classificate in due gruppi: medie analitiche e medie di posizione. Il primo gruppo comprende le medie: aritmetica, armonica, geometrica e quadratica. Sono dette analitiche perché ciascuna di esse è ottenuta applicando operazioni algebriche a tutte le osservazioni. Il secondo include: la mediana, i quartili, i decili (e più in generale i quantili), nonché il valore centrale. Ciascuna di queste corrisponde a un termine o alla media aritmetica di due termini che occupano particolari posizioni nella graduatoria non decrescente delle osservazioni.

Le medie analitiche possono essere calcolate solo su dati quantitativi. Le medie di posizione possono essere calcolate sia su dati quantitativi sia su dati ordinali. Sui dati nominali è possibile calcolare solo la moda.

In questo capitolo, le medie vengono dapprima definite per le serie di osservazioni; le loro espressioni algebriche o le loro procedure di calcolo vengono poi estese al caso delle distribuzioni di frequenza. Per le medie analitiche vengono discusse anche le versioni ponderate, da utilizzare quando alla singola modalità viene attribuito un peso, una misura di importanza relativa.

Parole chiave

centile, 110
classe mediana, 113
classe modale, 116
decile, 110
distribuzioni plurimodali, 116
linearità della media aritmetica, 94
media aritmetica, 93
media aritmetica ponderata, 105
media armonica, 96
media di potenza, 99
media geometrica, 97
media quadratica, 99
media troncata, 116
media troncata con sostituzioni, 116
mediana, 107
medie analitiche, 93
medie di classe, 102
medie di posizione, 93
medie robuste, 116
misure di tendenza centrale, 93
moda, 115
percentile, 110
quantile, 110
quartile, 109
valore centrale, 115

Commenti e curiosità

1. Le medie secondo Chisini

Sia $x_1, x_2, \ldots, x_n$ una serie di osservazioni della variabile quantitativa X. Sia $f(x_1, x_2, \ldots, x_n)$ un'operazione matematica sui termini della serie, ad esempio la somma, il prodotto ecc. Una costante numerica M è chiamata media secondo questa operazione matematica, se vale la relazione

$$f(M, M, \ldots, M) = f(x_1, x_2, \ldots, x_n),$$

cioè se la sostituzione della costante M a $x_1, x_2, \ldots, x_n$ non altera il risultato dell'operazione matematica.

Una media che soddisfa questa proprietà è chiamata *media secondo Chisini*, in omaggio a Oscar Chisini (matematico italiano, 1889-1967) a cui si deve tale definizione. Sono medie secondo Chisini tutte le medie analitiche introdotte in questo capitolo: per ciascuna di esse infatti la proprietà (b) è un modo equivalente per esprimere la relazione $f(M, M, \ldots, M) = f(x_1, x_2, \ldots, x_n)$, con $f(\cdot)$ che è: la somma per la media aritmetica, la somma dei reciproci per la media armonica, il prodotto per la media geometrica e la somma dei quadrati per la media quadratica.

2. Come "predire" il totale sulla base della media aritmetica

Come si è visto, moltiplicando la media aritmetica per il numero di osservazioni, si ottiene il totale della variabile. Secondo alcuni studiosi, questa proprietà, da sola, giustifica il ruolo centrale della media aritmetica. In pratica, essa è molto utile per calcoli rapidi sull'ordine di grandezza di certi fenomeni. Ad esempio, conoscendo il numero dei nuclei familiari in una certa area residenziale, una quantificazione approssimata della popolazione totale dell'area è data dal prodotto dell'ampiezza media delle famiglie, nota per l'intero comune, per il numero delle famiglie.

3. La proprietà associativa della media aritmetica

Data una serie di osservazioni $x_1, x_2, \ldots, x_n$, la si suddivida in L sottoinsiemi disgiunti aventi numerosità $n^{(1)}, n^{(2)}, \ldots, n^{(L)}$ e medie aritmetiche $\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(L)}$. Allora la media aritmetica della serie iniziale è uguale alla media aritmetica ponderata delle medie aritmetiche dei sottoinsiemi con pesi dati dalle rispettive numerosità:

$$\mu = \frac{\mu^{(1)} \cdot n^{(1)} + \mu^{(2)} \cdot n^{(2)} + \cdots + \mu^{(L)} \cdot n^{(L)}}{n^{(1)} + n^{(2)} + \cdots + n^{(L)}}.$$

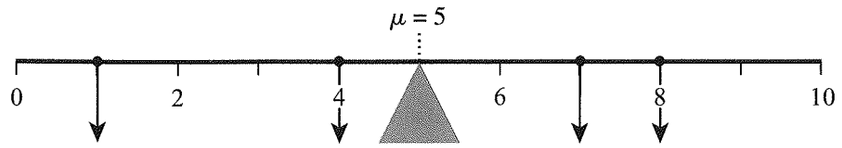
La proprietà, che vale anche per le distribuzioni di frequenza, è dimostrata nel Paragrafo A.2.1 dell'Appendice A.

4. La media aritmetica come "punto di equilibrio"

La Figura 4.5 mostra come la media possa essere interpretata fisicamente come il punto di equilibrio,

il punto di fulcro, in corrispondenza del quale un'asta graduata rimane in equilibrio quando vengono applicati pesi uguali (simboleggiati da frecce verso il basso) in punti che rappresentano un insieme di misurazioni su una determinata scala. Nell'esempio, le osservazioni sono: 1, 4, 7 e 8. Il peso può essere una costante arbitraria (ad esempio 50 g): se i pesi sono uguali, la media aritmetica ponderata è uguale alla media aritmetica semplice $(1 + 4 + 7 + 8)/4 = 5$.

Figura 4.5
Media aritmetica come punto d'equilibrio di un'asta.

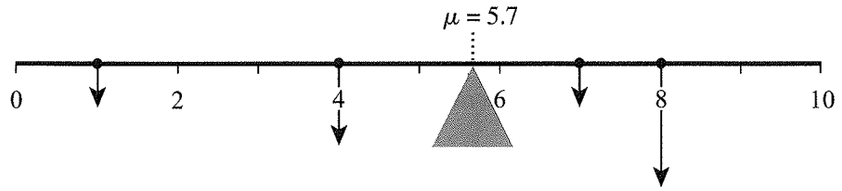


La Figura 4.6 illustra la stessa idea quando i pesi non sono uguali. Le modalità sono le stesse del caso precedente; i pesi sono: 0.5, 1, 0.5, 1.5. La media aritmetica ponderata è data da

$$\frac{1 \cdot 0.5 + 4 \cdot 1 + 7 \cdot 0.5 + 8 \cdot 1.5}{0.5 + 1 + 0.5 + 1.5} = 5.7.$$

Come si vede, il punto di equilibrio si sposta a destra "attratto" dal peso della modalità 8.

Figura 4.6
Media aritmetica ponderata come punto d'equilibrio di un'asta.



5. La percentuale media non è uguale alla media delle percentuali

La percentuale di persone favorevoli alla politica economica del governo è del 31% tra i più giovani (18-30 anni), del 44% nella fascia di età intermedia (31-50 anni) e del 53% tra i più anziani (51 anni o più). Se si volesse conoscere la percentuale media per tutti gli adulti, non si può calcolare la media $(31 + 44 + 55)/3$. Si deve invece utilizzare la media aritmetica ponderata, prendendo come pesi la proporzione di persone nelle tre classi di età. Se queste proporzioni fossero 0.16, 0.33 e 0.51, la percentuale di popolazione adulta favorevole alla politica economica del governo sarebbe data da

$$\frac{31 \cdot 0.16 + 44 \cdot 0.33 + 55 \cdot 0.51}{0.16 + 0.33 + 0.51} = 48.$$

6. Quando la frequenza relativa è una media aritmetica

Si consideri la distribuzione di frequenza di una variabile dicotomica nominale le cui modalità, A e $\bar{A}$ (si legga "non A"), hanno frequenze relative f_1 e $1 - f_1$. Si assegni convenzionalmente il numero 1 ad A (a significare "presenza di A") e il numero 0 a $\bar{A}$ (a significare "assenza di A"). La media aritmetica di questa speciale distribuzione di frequenza è

$$\mu = 0 \cdot (1 - f_1) + 1 \cdot f_1 = f_1.$$

7. Il paradosso di Simpson: quando il confronto di due medie perde di significato

I decessi per Covid-19 dal 20 febbraio 2020 al 13 gennaio 2021, per la popolazione in età di 40 anni e più, nella Provincia autonoma di Trento (nel seguito indicata come situazione A) e nella regione Marche (nel seguito indicata come situazione B) sono stati 1027 e 1696 con un tasso di letalità (rapporto tra decessi e contagi) pari a 6.07% e 5.67% (Fonte: Istituto superiore di sanità). Si potrebbe pensare che nella situazione B si registri una letalità minore: come si vedrà, questa conclusione non regge a una analisi più approfondita. Se si disaggregano i dati per classi di età e si calcolano i tassi a livello di classe, si perviene alla conclusione opposta, come è mostrato nella tabella che segue:

Età	Situazione A			Situazione B		
	Contagi	Decessi	Letalità %	Contagi	Decessi	Letalità %
40-59	7 764	16	0.21	15 259	70	0.46
60-79	4 975	260	5.23	9 399	530	5.64
≥ 80	4 194	751	17.91	5 256	1096	20.85
Totale	16 933	1027	6.07	29 914	1696	5.67

Nelle tre classi di età, è la situazione A che presenta una letalità inferiore. La spiegazione di una simile contraddizione, denominata *paradosso di Simpson*, va ricercata nel fatto che la letalità cresce con l'età delle persone contagiate, cosicché se le situazioni a confronto presentano distribuzioni per età sensibilmente diverse può verificarsi l'inconveniente logico descritto.

A determinare la contraddizione c'è il meccanismo aritmetico sottostante al confronto che può essere chiarito tramite la proprietà associativa della media aritmetica (si vedano le note 3 e 6 di questa sezione). In base ad essa, il tasso di letalità generale nelle due situazioni può essere visto come media aritmetica ponderata dei tassi di letalità nelle classi con pesi pari al numero dei contagi nelle classi.

Nella situazione A il tasso di letalità può essere così espresso

$$6.07 = \frac{0.0021 \cdot 7764 + 0.0523 \cdot 4975 + 0.1791 \cdot 4194}{16\,933} \cdot 100$$
$$= (0.0021 \cdot 0.4585 + 0.0523 \cdot 0.2938 + 0.1791 \cdot 0.2477) \cdot 100.$$

Analogamente, nella situazione B, si ha

$$5.67 = \frac{0.0046 \cdot 15\,259 + 0.0564 \cdot 9399 + 0.2085 \cdot 5256}{29\,914} \cdot 100$$
$$= (0.0046 \cdot 0.5101 + 0.0564 \cdot 0.3142 + 0.2085 \cdot 0.1757) \cdot 100.$$

Confrontando le medie aritmetiche ponderate entro parentesi, si vede come: alla prima frequenza relativa, quella più piccola, è associato un peso minore in A che non in B; alla seconda frequenza, quella intermedia, è associato un peso non molto dissimile nelle due situazioni; alla terza frequenza relativa, quella più grande, è associato un peso maggiore in A che non in B. Il modo di combinarsi dei pesi con le frequenze, che dipende dalla distribuzione di frequenza dei contagi per classe di età, fa sì che avvenga l'inversione della disuguaglianza.

Il paradosso segnala che l'età, come peraltro è largamente noto, incide sul tasso di letalità e che pertanto ha senso confrontare la letalità solo a livello di classe: in questo modo infatti il confronto avviene "a parità di condizioni", cioè a parità del fattore che ha incidenza sulla letalità.

Un infortunio statistico simile a quello descritto, ma ben più grave, ha riguardato il confronto del tasso di letalità per Covid-19 tra vaccinati e non vaccinati e ha innalzato il paradosso di Simpson agli onori dei grandi media nazionali e internazionali, negli ultimi mesi del 2021. Infatti, il confronto, fatto impropriamente a livello delle due popolazioni, portava a una letalità simile per i vaccinati e per i non vaccinati, in ragione della minore anzianità media dei secondi, e induceva gli scettici del vaccino a metterne in dubbio l'efficacia.

In generale, il paradosso di Simpson si manifesta quando si confrontano le medie di una stessa variabile in due popolazioni ignorando uno o più fattori che hanno impatto sulla variabile. Nella letteratura statistica, sono descritti svariati episodi al riguardo.

8. Galton e il peso del bue

Nel 1906, Francis Galton girovagando all'interno di una fiera di bestiame, a Plymouth, s'imbatté in una gara di stima del peso di un bue. I partecipanti alla competizione, circa 800, avevano scritto in un biglietto

il loro pronostico: Galton chiese di poter controllare tutti i biglietti, calcolò la media aritmetica dei pronostici e ottenne 1197 libbre, equivalenti a circa 540 chili. Il peso effettivo del bue era 1198 libbre: l'errore commesso era dunque inferiore a mezzo kg (una libbra equivale a 454 grammi). Commentando il fatto, Galton ebbe a dire che il risultato era più affidabile di quanto ci si potesse aspettare da una giuria democratica.

9. Perché il nome media armonica?

La media armonica fu scoperta dai pitagorici (circa 500 a.C.). Dati due numeri, a e c, la media armonica, b, è il numero definito dalla proporzione $(b-a)/a = (c-b)/c$. Infatti, dalla proporzione si ricava $b = 2ac/(a+c)$ che è la media armonica di a e c. Per l'origine dell'aggettivo "armonica" va forse considerato che i Greci con la parola armonia intendevano l'arte di individuare un numero caratterizzato da una speciale posizione, come quella di essere medio proporzionale tra due numeri assegnati.

Esercizi

4.1 Si consideri la serie storica di cui all'Esempio 2.11.

a) Si calcoli la media annua dei nati vivi nel periodo considerato con la media aritmetica.

b) Si verifichi che la somma degli scarti è zero.

c) Si determini la somma dei quadrati degli scarti.

4.2* Si verifichi che se si aggiunge a una serie di osservazioni un nuovo dato pari alla media della serie stessa, la media aritmetica della serie così modificata resta invariata. (*Suggerimento*: si consideri che la nuova media si può scrivere come $(n\,\mu + \mu)/(n + 1)$.)

4.3 La media della temperatura corporea, in gradi Celsius, di un gruppo di adulti sani è di 36.8 °C. Quale sarebbe la media se le temperature fossero espresse in gradi Fahrenheit? (*Suggerimento*: si deve ricordare che la conversione da Celsius a Fahrenheit è data dalla relazione °F = 32 + 1.8 °C.)

4.4 Si considerino i dati utilizzati nell'Esercizio 3.1 (prima colonna del *file* "Iris"). Si calcoli la somma dei quadrati degli scarti dalla media aritmetica e si verifichi che questa somma è minore della somma dei quadrati degli scarti da un numero a scelta, diverso dalla media aritmetica.

4.5 Si riprenda la distribuzione delle famiglie italiane per numero di componenti di cui all'Esercizio 2.4.

a) Si calcoli la media aritmetica assumendo che il numero medio dei componenti delle famiglie con 6 o più componenti sia 6.2.

b) Si verifichi che la somma degli scarti dalla media aritmetica moltiplicati per le rispettive frequenze relative sia zero.

c) Si calcolino le medie armonica e geometrica utilizzando le informazioni di cui al punto a).

d) Si verifichi che la media armonica è minore della media geometrica e che la media geometrica è minore della media aritmetica.

4.6 Si consideri la distribuzione di frequenza in classi esaminata nell'Esercizio 3.4.

a) Si calcoli la media aritmetica e la si confronti con il valore determinato nell'Esercizio 4.4 a partire dalla serie di osservazioni.

b) A puro titolo di esercizio, si calcolino le medie armonica e geometrica (non ci sono ragioni specifiche per usarle).

4.7 Imprese e addetti nell'industria (dati Istat):

N. di addetti	N. di imprese (migliaia)	Totale addetti (migliaia)
1	186	186
2-9	260	1028
10-19	56	750
20-49	26	772
50-249	11	1023
> 250	2	1176
Totale	541	4935

a) Si calcoli la media aritmetica del numero degli addetti utilizzando l'informazione contenuta nell'ultima colonna.

b) Si ripeta il calcolo della media aritmetica del numero degli addetti usando i valori centrali di classe (per l'ultima classe si utilizzi la media di classe) e si discuta sulla differenza tra i due risultati.

4.8 Distribuzione di frequenza dei comuni in base all'ampiezza demografica (si veda l'Esercizio 1.3):

Ampiezza demografica (migliaia)	N. di comuni	Popolazione
< 5	33	74 370
5-10	11	76 422
10-15	2	26 227
15-20	5	84 417
20-30	3	70 915
30-40	2	72 572
> 40	3	266 898
Totale	59	671 821

a) Si determini l'ampiezza demografica media dei comuni (si utilizzino le informazioni disponibili sull'ampiezza demografica complessiva dei comuni appartenenti a ciascuna classe).

b) Si calcoli la media quadratica.

4.9 La tabella seguente riporta la dimensione media delle famiglie in quattro Paesi europei (Fonte: Eurostat; anno 2017):

Paese	N. medio componenti	N. di famiglie (migliaia)
Francia	2.2	29 314
Italia	2.3	25 865
Portogallo	2.5	4 103
Spagna	2.5	18 513

a) Si calcoli la dimensione media delle famiglie dei quattro Paesi (presi insieme).

b) Si ripeta il calcolo in a) utilizzando la media aritmetica semplice e si illustri perché è maggiore di quella ottenuta al punto precedente.

4.10 Dei 30 studenti che seguono le lezioni di un dato corso universitario, 27 hanno un'età media di 19.6 anni, i restanti tre hanno, rispettivamente, 49, 52 e 53 anni. Si determini l'età media dell'intero gruppo.

4.11* Si verifichi che la media aritmetica ponderata coincide con la media aritmetica semplice se i pesi sono tutti uguali fra loro. (*Suggerimento*: si scriva la media aritmetica ponderata nella forma $\mu = (x_1 \cdot w + x_2 \cdot w + \cdots + x_k \cdot w)/(w + w + \cdots + w)$, essendo w il peso uniforme, e si semplifichi.)

4.12 Con riferimento alla matrice dei dati di cui all'Esercizio 1.6, si calcoli il peso mediano degli individui di età non superiore a 25 anni. Si calcolino anche il primo e il terzo quartile.

4.13 Si calcolino i quartili della distribuzione di cui all'Esercizio 2.4.

4.14 Si calcolino i quartili della distribuzione in classi di cui all'Esercizio 4.6.

4.15 Si calcolino i quartili della distribuzione in classi di cui all'Esercizio 4.8.

4.16 Autovetture a gasolio circolanti in Italia secondo la cilindrata (dati ACI, anno 2017):

Cilindrata	Autovetture (migliaia)
< 800	78
801-1200	92
1201-1600	7 977
1601-2000	6 673
2001-2500	1 412
> 2500	665
Totale	16 897

Si determinino i quartili della distribuzione.

4.17* Si verifichi che il valore centrale coincide con la media aritmetica se i termini della distribuzione sono tutti diversi ed equispaziati. (*Suggerimento*: si esprima il numeratore della media aritmetica come somma di termini in progressione aritmetica.)

4.18 Con riferimento alla distribuzione di cui all'Esercizio 4.16, si individui la classe modale.

Medie con Excel

Con Excel si possono calcolare le medie aritmetica, armonica e geometrica utilizzando le funzioni MEDIA( ), MEDIA.ARMONICA( ) e MEDIA.GEOMETRICA( ). I passi da compiere sono di seguito indicati con riferimento alla media aritmetica.

- Si apre il *file* dati e, in una casella a scelta, si scrive la funzione MEDIA( ) preceduta dal segno di uguale e si seleziona l'area del foglio contenente le osservazioni.
- Si preme "Invio" e si ottiene il risultato nella stessa casella in cui è scritta la funzione.

L'esecuzione delle altre due funzioni segue la stessa procedura.

Quanto alle medie di posizione, Excel prevede le funzioni MEDIANA( ), QUARTILE( ) e PERCENTILE( ), la cui esecuzione è immediata. Per l'impiego della prima, basta specificare all'interno delle parentesi l'area del foglio in cui sono contenuti i dati. Le altre due richiedono, rispettivamente, oltre alla selezione dell'area dei dati, un numero compreso tra 1 e 3 che identifica il quartile, e un numero compreso tra 0 e 1 che indica il valore della funzione di ripartizione corrispondente al percentile desiderato, ad esempio, 0.10 per il decimo percentile, 0.95 per il 95-esimo ecc.

Vale la pena notare che, nel calcolo dei quartili e dei centili, Excel utilizza approssimazioni numeriche che producono risultati leggermente diversi da quelli che si ottengono dall'applicazione delle procedure illustrate in questo volume.

CAPITOLO

5

Variabilità

Sommario del capitolo

5.1 Introduzione
5.2 Scostamenti medi
5.3 Differenze medie
5.4 Campo di variazione e differenza interquartile
5.5 Indici percentuali di variabilità
5.6 Indici di disuguaglianza o di concentrazione
5.7 Eterogeneità
5.8 Quale indice impiegare
Riassunto
Commenti e curiosità
Esercizi

5.1 Introduzione

In termini molto generali, per variabilità si intende l'attitudine dei fenomeni naturali e sociali a manifestarsi in modi differenti. La variabilità costituisce la ragione stessa dell'esistenza della statistica. Tutta la metodologia statistica, dalle medie alle più sofisticate tecniche dell'inferenza statistica, ha a che fare con la variabilità: per "neutralizzarla", con le medie, per "misurarla", con gli indici di variabilità, oggetto di questo capitolo, per "spiegarla", con le tecniche della regressione e in generale con i metodi dell'inferenza statistica.

Proseguendo la discussione del capitolo precedente sulle medie, se si assume il punto di vista di chi utilizza l'informazione statistica, la necessità di misure di variabilità che affianchino le medie è di facile intuizione. Un investitore in borsa è bene che valuti non solo la quotazione media del titolo a cui è interessato, ma anche la variabilità delle quotazioni – la volatilità, per usare il termine tecnico – del titolo nel tempo. Allo stesso modo, l'economista che voglia confrontare il benessere economico di due Paesi deve basarsi non solo sul reddito pro-capite, ma anche su una misura di disuguaglianza dei redditi.

Questo capitolo è dedicato agli indici di variabilità, che possono essere distinti in tre categorie a seconda che esprimano la variabilità dei dati come:

- media delle distanze delle osservazioni dalla loro media aritmetica (scostamento semplice medio, deviazione standard);
- media delle differenze, prese in valore assoluto, tra tutte le coppie di osservazioni (differenze medie);
- differenza tra osservazioni che occupano specifiche posizioni o tra medie di posizione (campo di variazione, differenza interquartile).

Gli indici di variabilità verranno definiti, descrivendone le proprietà, con riferimento a serie di osservazioni. Le loro formule saranno poi adattate al caso delle distribuzioni di frequenza.

Trasformando opportunamente questi indici, si otterranno i cosiddetti indici percentuali di variabilità, che permettono di effettuare corretti confronti della variabilità di due o più insiemi di dati espressi in unità di misura differenti o con ordini di grandezza diversi.

Si affronterà poi il problema della misurazione della variabilità dei dati concernenti le variabili trasferibili, come il reddito e la ricchezza, con l'introduzione degli indici di concentrazione.

Infine, si esaminerà brevemente il problema della misurazione della variabilità delle distribuzioni di frequenza di variabili qualitative, introducendo gli indici di eterogeneità.

5.2 Scostamenti medi

Data una serie di osservazioni, $x_1, x_2, \ldots, x_n$, si considerino le distanze dei termini dalla media μ:

$$|x_1 - \mu|, |x_2 - \mu|, \ldots, |x_n - \mu|,$$

rappresentate in Figura 5.1.

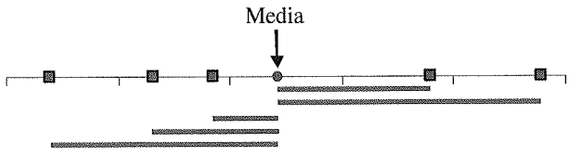


Figura 5.1
Distanze delle osservazioni dalla media come lunghezze dei segmenti che vanno dalle singole osservazioni a μ.

Queste quantità rappresentano le distanze delle osservazioni da μ e sono nulle quando $x_i = \mu$ per ogni i (assenza di variabilità). Facendo un'opportuna sintesi di tali distanze tramite una media, si ottiene una misura di variabilità, che è nulla in assenza di variabilità ed è tanto maggiore quanto maggiore è la "distanza media" delle osservazioni da μ.

Una prima misura di variabilità, il cui significato è immediato, è lo *scostamento semplice medio*.

DEFINIZIONE 5.1

Lo scostamento semplice medio della serie di osservazioni $x_1, x_2, \ldots, x_n$ avente media μ è definito come la media aritmetica degli scarti da μ presi in valore assoluto:

$$S_\mu = \frac{|x_1 - \mu| + |x_2 - \mu| + \cdots + |x_n - \mu|}{n} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \mu|. \qquad (5.1)$$

Definendo come distanza tra due numeri la loro differenza in valore assoluto, la (5.1) può essere interpretata come la distanza media delle osservazioni dalla media aritmetica μ.

ESEMPIO 5.1

Si calcoli lo scostamento semplice medio per la seguente serie di osservazioni relativa alla dimensione (in m^2) di quattro appartamenti:

110, 97, 120, 125.

Soluzione

La superficie media è 113; gli scarti in valore assoluto delle osservazioni da μ sono:

$$|110 - 113| = 3, \ |97 - 113| = 16, \ |120 - 113| = 7, \ |125 - 113| = 12.$$

Lo scostamento semplice medio è allora

$$S_\mu = \frac{3 + 16 + 7 + 12}{4} = 9.5.$$

Ciò significa che mediamente le superfici delle quattro abitazioni differiscono da μ per 9.5 m^2. ■

Una seconda misura di variabilità, di gran lunga più utilizzata, è lo *scostamento quadratico medio* o *deviazione standard*.

DEFINIZIONE 5.2

La deviazione standard della serie di osservazioni $x_1, x_2, \ldots, x_n$ avente media μ è definita come la media quadratica degli scarti da μ:

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}. \qquad (5.2)$$

Assumendo come distanza tra due numeri il quadrato della loro differenza, la (5.2) può essere interpretata come la distanza media delle osservazioni da μ.

La (5.2) può essere espressa nella forma alternativa

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2},$$

denominata formula operativa, perché rende più agevole il calcolo di σ (per la sua dimostrazione, si veda l'Esercizio 5.5).

Lo scostamento semplice medio e la deviazione standard possono essere visti come misure diverse della distanza media tra le osservazioni e la loro media aritmetica.

ESEMPIO 5.2

Si calcoli la deviazione standard della serie di osservazioni dell'Esempio 5.1.

Soluzione

Applicando la (5.2), si ottiene

$$\sigma = \sqrt{\frac{(110 - 113)^2 + (97 - 113)^2 + (120 - 113)^2 + (125 - 113)^2}{4}} = 10.70.$$

Dunque, le superfici dei quattro appartamenti si discostano mediamente da μ per 10.7 m^2. Per gli stessi dati, nell'Esempio 5.1, si è ottenuto uno scostamento semplice medio di 9.5 m^2. Ciò si spiega con il fatto che la (5.1) è la media aritmetica degli scarti in valore assoluto, mentre la (5.2) è la media quadratica dei medesimi scarti. ■

Un utile esercizio per la comprensione dell'argomento in discussione consiste nel comparare la variabilità di più serie di osservazioni con l'esame delle rispettive rappresentazioni grafiche. Nella Figura 5.2 sono rappresentate quattro serie di osservazioni, ciascuna delle quali è formata da 5 elementi.

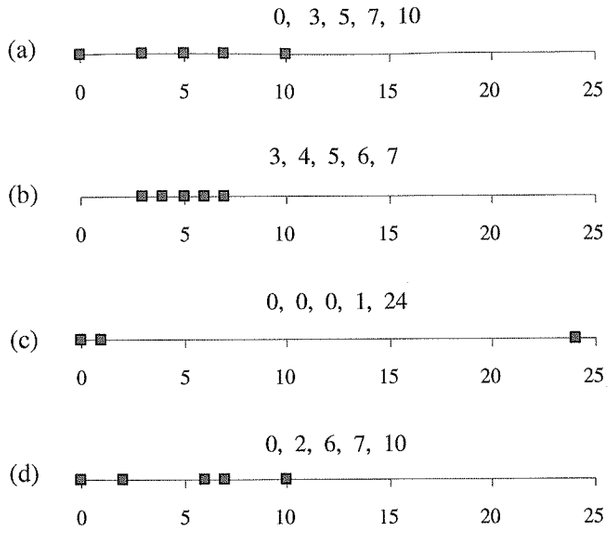


Figura 5.2
Serie di osservazioni con la stessa media e diverso grado di variabilità. La deviazione standard assume i valori (procedendo dall'alto verso il basso): 3.41, 1.41, 9.51 e 3.58.

La figura mostra che la serie con minore variabilità è la (b), dove le osservazioni sono più vicine tra loro e quindi più vicine alla loro media. Al contrario, la serie con variabilità più alta è la (c), dove un punto è molto distante dagli altri quattro. Le serie (a) e (d) presentano una variabilità pressoché uguale.

PROPOSIZIONE 5.1

Lo scostamento semplice medio e la deviazione standard soddisfano le proprietà di seguito indicate:

(a) sono sempre non negativi e sono nulli se e solo se tutte le osservazioni sono uguali;

(b) non variano se si aggiunge la stessa quantità a (sia positiva che negativa) a ciascuna osservazione;

(c) sono moltiplicati per $|b|$ se si moltiplicano tutte le osservazioni per un numero b (sia positivo che negativo).

La prima proprietà è ovvia, le altre due sono dimostrate nell'Esercizio 5.6. ■

Il quadrato della deviazione standard,

$$\sigma^2 = \frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_n-\mu)^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2, \qquad (5.3)$$

è denominato varianza. La (5.3) non è una vera e propria misura di variabilità: a differenza di σ, non è espressa nella stessa unità di misura della variabile, bensì nel quadrato dell'unità di misura. Per esempio, se le osservazioni x_i sono espresse in cm, la varianza è espressa in cm^2.

Un'altra quantità che vale la pena di menzionare è la devianza. Si tratta della somma dei quadrati degli scarti:

$$D = (x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_n-\mu)^2 = \sum_{i=1}^{n}(x_i-\mu)^2.$$

5.2.1 Il caso delle distribuzioni di frequenza

Le (5.1) e (5.2) vanno adattate per poterle utilizzare nel caso di distribuzioni di frequenza. Per brevità, ci si limiterà a presentare il caso della deviazione standard (per S_μ si procede in modo analogo):

$$\sigma = \sqrt{\frac{(x_1-\mu)^2 n_1 + (x_2-\mu)^2 n_2 + \cdots + (x_k-\mu)^2 n_k}{n}}$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{k}(x_i-\mu)^2 n_i}. \qquad (5.4)$$

La (5.4) si deduce immediatamente dalla (5.2) applicando la formula della media quadratica per le distribuzioni di frequenza agli scarti da μ, $|x_1-\mu|, |x_2-\mu|, \ldots, |x_k-\mu|$.

L'indice può essere posto nella forma

$$\sigma = \sqrt{\sum_{i=1}^{k}(x_i-\mu)^2 f_i},$$

dove $f_i = n_i/n$ è la frequenza relativa della modalità x_i; oppure, adattando al caso in esame la formula operativa introdotta nel paragrafo precedente, nella forma

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i - \mu^2}. \qquad (5.5)$$

Le proprietà della deviazione standard descritte nella Proposizione 5.1 continuano a valere anche in questo contesto.

ESEMPIO 5.3

Si calcoli la deviazione standard per la distribuzione di cui all'Esempio 4.5, utilizzando prima la (5.4) e poi la formula operativa.

Soluzione

Si ricorda che la media della distribuzione in questione è 2.55. I calcoli sono illustrati nella Tabella 5.1, dove, nella terza colonna, sono riportati i prodotti dei quadrati degli scarti per le rispettive frequenze. Per esempio, la quantità 91.03 è il risultato dell'operazione $(0 - 2.55)^2 \cdot 14$.

N. di esami (x_i)	Frequenza (n_i)	$(x_i-\mu)^2 \cdot n_i$	$x_i^2 \cdot n_i$
0	14	91.03	0
1	41	98.50	41
2	83	25.11	332
3	116	23.49	1044
4	56	117.74	896
5	5	30.01	125
Totale	315	385.88	2438

Tabella 5.1
Calcolo della deviazione standard per la distribuzione di frequenza dell'Esempio 4.5.

Con la (5.4), si ha

$$\sigma = \sqrt{\frac{385.88}{315}} = 1.11.$$

Con la (5.5), si ha

$$\sigma = \sqrt{\frac{2438}{315} - 2.55^2} = 1.11.$$

■

5.2.2 Il caso delle distribuzioni di frequenza in classi

Quando i dati sono raggruppati in classi, il calcolo della deviazione standard richiede che si determinino i valori centrali di classe, $\bar{x}_i = (c_{i-1} + c_i)/2$, $i = 1, 2, \ldots, k$. La deviazione standard viene poi ricavata sostituendo x_i con $\bar{x}_i$ nella (5.4):

$$\sigma = \sqrt{\frac{(\bar{x}_1 - \mu)^2 n_1 + (\bar{x}_2 - \mu)^2 n_2 + \cdots + (\bar{x}_k - \mu)^2 n_k}{n}}$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{k} (\bar{x}_i - \mu)^2 n_i}$$

o nella formula operativa:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{k} \bar{x}_i^2 n_i - \mu^2}\,.$$

Come già osservato per le medie, il calcolo della deviazione standard con i valori centrali di classe introduce un'approssimazione: prendere il valore centrale è come ipotizzare che in tutte le unità appartenenti alla classe la variabile assuma un'unica modalità pari al valore centrale di classe. È chiaro che, se disponibili, vanno utilizzati preferenzialmente i dati individuali per il calcolo di σ come di ogni altro indice di variabilità.

ESEMPIO 5.4

Si calcoli la deviazione standard della distribuzione di cui all'Esempio 4.19, dove, ai fini del calcolo della media e della deviazione standard, il limite inferiore della prima classe e il limite superiore dell'ultima classe sono stati posti pari a 399 e 6999.

Soluzione

I calcoli sono esposti in modo dettagliato nella Tabella 5.2. Si tenga presente che la media aritmetica è 3362.73. In particolare, nell'ultima colonna della tabella è riportata la quantità $(\bar{x}_i - \mu)^2 f_i$. Per esempio, il primo numero in alto è il risultato dell'operazione

$$(449 - 3362.73)^2(39/265\,961).$$

Peso (g)	Frequenza (n_i)	Valore centrale ($\bar{x}_i$)	$(\bar{x}_i - \mu)^2 \cdot f_i$
399-499	39	449.0	1 244.93
500-999	548	749.5	14 070.77
1000-1499	1 149	1249.5	19 292.82
1500-1999	2 020	1749.5	19 766.33
2000-2499	7 691	2249.5	35 837.25
2500-2749	12 772	2624.5	26 171.24
2750-2999	26 922	2874.5	24 128.94
3000-3249	53 178	3124.5	11 347.68
3250-3499	58 093	3374.5	30.26
3500-3749	50 640	3624.5	13 047.14
3750-3999	30 051	3874.5	29 593.11
4000-4499	20 257	4249.5	59 893.43
4500-4999	2 369	4749.5	17 129.95
5000-5499	192	5249.5	2 569.93
5500-5999	27	5749.5	578.32
6000-6999	13	6499.5	480.94
Totale	265 961	–	275 183.03

Tabella 5.2
Calcolo della deviazione standard per la distribuzione di frequenza di cui all'Esempio 4.19.

Si ha allora $\sigma = \sqrt{275\,183.03} = 524.58$. Questo numero indica che mediamente il grado di diversità dei pesi dei neonati rispetto alla media è 524.58 g. ■

Se per ogni classe è noto il totale della variabile, come nell'Esempio 5.5, è preferibile utilizzare le medie di classe μ_i $(i = 1, 2, \ldots, k)$ al posto dei valori centrali.

ESEMPIO 5.5

Si calcoli la deviazione standard della distribuzione di frequenza dei comuni italiani classificati per superficie (Fonte: Istat).

Soluzione

Si noti che per ogni classe è disponibile la superficie totale dei comuni. Questo permette di calcolare esattamente la media della distribuzione di frequenza come 30 133/8100 = 3.72, e l'esatta media della classe i-esima μ_i $(i = 1, 2, \ldots, 5)$ dividendo la superficie totale della classe per la frequenza di classe. I calcoli sono riportati di seguito:

Superficie (ha)	N. di comuni n_i	Superficie totale (migliaia di ha)	Media di classe μ_i	$(\mu_i - \mu)^2 \cdot n_i$
< 1000	1738	1 071	0.616	16 745.310
1001-2000	2058	3 013	1.464	10 474.265
2001-6000	2979	10 282	3.451	215.563
6001-25 000	1259	13 364	10.615	59 854.150
> 25 000	66	2 403	36.409	70 525.668
Totale	8100	30 133		157 814.956

Tabella 5.3
Calcolo della deviazione standard per la distribuzione di frequenza dei comuni italiani per classi di superfici.

La deviazione standard della distribuzione è data da

$$\sigma = \sqrt{\frac{157\,814.956}{8100}} = 4.414.$$

Dunque, mediamente il grado di diversità delle superfici dei comuni rispetto alla superficie media generale è 4414. ■

Nella Figura 5.3, sono rappresentate quattro distribuzioni aventi la stessa media, ottenute dalla distribuzione esaminata nell'Esempio 5.4 modificando opportunamente le densità di frequenza. Stante la simmetria quasi perfetta delle distribuzioni, bisogna considerare che la media si colloca al centro dell'asse delle ascisse. Pertanto, la variabilità è tanto minore quanto maggiori sono le densità di frequenza nella parte centrale del grafico (si veda il punto 1 di "Commenti e curiosità").

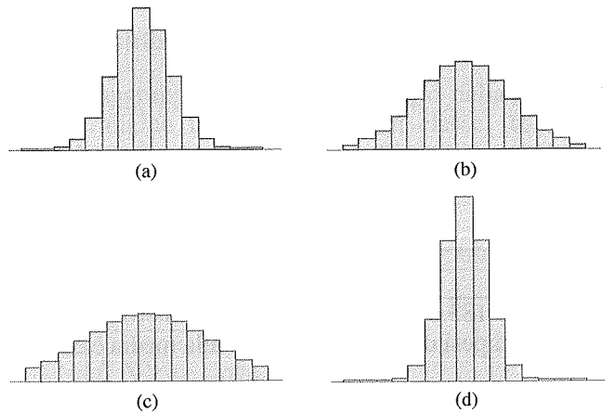


Figura 5.3
Distribuzioni con la stessa media e con deviazione standard pari a: 524 in (a), 853 in (b), 1158 in (c) e 404 in (d).

5.3 Differenze medie

Data una serie di osservazioni, si vuole definire una misura di variabilità confrontando ogni osservazione non con la media, ma con tutte le altre osservazioni. In questo modo, l'osservazione x_1 genera le $n-1$ differenze in valore assoluto:

$$|x_1 - x_2|, |x_1 - x_3|, \ldots, |x_1 - x_n|,$$

l'osservazione x_2 genera le $n-1$ differenze in valore assoluto:

$$|x_2 - x_1|, |x_2 - x_3|, \ldots, |x_2 - x_n|$$

e così via. Si ottengono così $n(n-1)$ differenze $|x_i - x_j|$. Una misura di variabilità può essere allora costruita facendo una opportuna sintesi di queste $n(n-1)$ differenze.

DEFINIZIONE 5.3

La differenza semplice media[1] della serie di osservazioni $x_1, x_2, \ldots, x_n$ è definita come la media aritmetica di tutte le differenze in valore assoluto tra le $n(n-1)$ coppie di osservazioni:

$$\Delta = \frac{1}{n(n-1)} \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} |x_i - x_j|. \qquad (5.6)$$

[1] Altri indici di variabilità si possono costruire sintetizzando le $n(n-1)$ distanze $|x_i - x_j|$ con formule diverse dalla media aritmetica, per esempio con la media quadratica.

ESEMPIO 5.6

Si riprendano i dati dell'Esempio 5.1,

$$110,\ 97,\ 120,\ 125,$$

e si calcoli la differenza semplice media.

Soluzione

Si ottengono le seguenti 12 differenze in valore assoluto tra le osservazioni della serie:

$$|110 - 97|,\ |110 - 120|,\ |110 - 125|,\ |97 - 110|,$$

$$|97 - 120|,\ |97 - 125|,\ |120 - 110|,\ |120 - 97|,$$

$$|120 - 125|,\ |125 - 110|,\ |125 - 97|,\ |125 - 120|.$$

La differenza semplice media è data da

$$\Delta = \frac{13 + 10 + \cdots + 5}{12} = \frac{188}{12} = 15.7,$$

valore che indica di quanto in media le dimensioni dei quattro appartamenti differiscono tra di loro. ■

Le differenze che appaiono nella (5.6) sono a due a due uguali (al riguardo si osservi l'esempio precedente), in quanto, per ogni coppia ij, $|x_i - x_j| = |x_j - x_i|$. Ne segue che la (5.6) può essere posta nella forma

$$\Delta = \frac{2}{n(n-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} |x_i - x_j|. \qquad (5.7)$$

La differenza semplice media presenta le stesse proprietà della deviazione standard (si veda la Proposizione 5.1).

PROPOSIZIONE 5.2

La differenza semplice media soddisfa le proprietà di seguito indicate:

(a) è sempre non negativa ed è nulla se e solo se tutte le osservazioni sono uguali;

(b) non varia se si aggiunge la stessa quantità a (sia positiva che negativa) a ciascuna osservazione;

(c) è moltiplicata per $|b|$ se si moltiplicano tutte le osservazioni per un numero b (sia positivo che negativo).

Queste proprietà possono essere verificate come nel caso della Proposizione 5.1. ■

5.3.1 Il caso delle distribuzioni di frequenza

Il procedimento di calcolo della differenza media va modificato nel caso di una distribuzione di frequenza. La (5.7) diventa

$$\Delta = \frac{2}{n(n-1)} \sum_{i=2}^{k} \sum_{j=1}^{i-1} |x_i - x_j| n_i n_j. \tag{5.8}$$

Invero, poiché la modalità x_i si presenta n_i volte e la modalità x_j n_j volte, la differenza $|x_i - x_j|$ si presenta $n_i \cdot n_j$ volte.

ESEMPIO 5.7

Si calcoli con la (5.8), la differenza media per la distribuzione di cui all'Esempio 5.3.

Soluzione

Si ha

$$\Delta = \frac{2}{315 \cdot 314}(|1-0|41 \cdot 14 + |2-0|83 \cdot 14 + \cdots + |5-4|5 \cdot 56) = 1.21. \quad \blacksquare$$

Distribuzioni di frequenza in classi

Quando i dati sono nella forma di distribuzioni di frequenza in classi, il calcolo di Δ si effettua con la (5.8) sostituendo a x_i e x_j i corrispondenti valori centrali di classe, a meno che non si conoscano le medie di classe, nel qual caso, come è stato detto per la deviazione standard, queste sono da preferire ai valori centrali.

5.4 Campo di variazione e differenza interquartile

Il *campo di variazione* è il più semplice indice di variabilità.

DEFINIZIONE 5.4

Il campo di variazione della serie di osservazioni $x_1, x_2, \ldots, x_n$ è dato dalla differenza

$$\Delta_c = x_{(n)} - x_{(1)},$$

dove $x_{(1)}$ e $x_{(n)}$ sono l'osservazione più piccola e quella più grande.

Quando i dati sono sotto forma di distribuzione di frequenza, la formula del campo di variazione deve essere modificata coerentemente.

L'indice Δ_c ha le proprietà che seguono: (a) è maggiore o uguale a zero, assume il valore zero se le osservazioni sono tutte uguali; (b) resta invariato se si aggiunge a ogni osservazione una quantità costante; (c) se ogni osservazione è moltiplicata per una costante, l'indice risulta moltiplicato per il valore assoluto della costante.

L'indice dipende unicamente dalle osservazioni estreme (minimo e massimo): assume lo stesso valore per tutte le serie che hanno gli stessi estremi, indipendentemente dai valori delle altre $n - 2$ osservazioni, ed è molto sensibile alla presenza di dati anomali.

Per questi motivi, l'indice è di "qualità inferiore" rispetto a quelli discussi nei paragrafi precedenti; la sua applicazione pratica è dunque limitata.

DEFINIZIONE 5.5

La differenza interquartile della serie di osservazioni $x_1, x_2, \ldots, x_n$ è data dalla differenza tra il terzo e il primo quartile della serie:

$$\Delta_q = q_3 - q_1. \tag{5.9}$$

La differenza interquartile soddisfa le seguenti proprietà: (a) è maggiore o uguale a zero, assume il valore zero se le osservazioni sono tutte uguali; (b) resta invariata se si aggiunge a ogni osservazione una quantità costante; (c) se ogni osservazione è moltiplicata per una costante, l'indice viene moltiplicato per il valore assoluto della costante.

Quanto alla proprietà (a), va notato che l'indice può essere nullo anche quando le osservazioni non sono tutte uguali. Ad esempio, si considerino i dati che seguono:

$$1,\ 2,\ 10,\ 10,\ 10,\ 10,\ 10,\ 10,\ 10,\ 10,\ 11,\ 12.$$

Come è facile verificare, si ha $q_1 = q_3 = 10$ e quindi $\Delta_q = 0$ pur non essendo nulla la variabilità.

Un vantaggio della differenza interquartile rispetto al campo di variazione è che è meno sensibile alla presenza dei valori anomali.

ESEMPIO 5.8

Si calcoli la differenza interquartile della distribuzione di frequenza dell'Esempio 4.19.

Soluzione

Il primo e terzo quartile di questa distribuzione, calcolati come soluzione dell'Esempio 4.19, sono $q_1 = 3071.66$ e $q_3 = 3682.45$. L'applicazione della (5.9) dà allora

$$\Delta_q = 3682.45 - 3071.66 = 610.79. \quad \blacksquare$$

5.5 Indici percentuali di variabilità

Gli indici di variabilità discussi finora sono espressi nella stessa unità di misura della variabile in esame. Si consideri, ad esempio, un processo industriale che produce bustine di camomilla del peso medio di 2 g. I pesi di un campione di cinque bustine sono:

$$\text{A}: 1.98,\ 2.03,\ 2.01,\ 1.99,\ 2.00.$$

La media e la deviazione standard di questi dati sono rispettivamente 2.002 g e 0.0172 g. La deviazione standard è interpretabile, in questo caso, come una misura dell'oscillazione (entro certi limiti fisiologica) delle caratteristiche fisiche dei pezzi che escono dal processo produttivo.

Si consideri ora un secondo processo industriale che produce confezioni di pasta alimentare del peso medio di 500 g. I pesi (in g) di un campione di cinque confezioni sono:

$$\text{B}: 495.0, 507.5, 502.5, 497.5, 500.0.$$

La media e la deviazione standard sono 500.5 g e 4.3012 g, rispettivamente. Si noti che le osservazioni del campione B si ottengono moltiplicando per 250 ogni osservazione del campione A. Inoltre, la media e la deviazione standard del campione B sono pari alla media e alla deviazione standard del campione A moltiplicate per 250, in virtù della proprietà (e) della media aritmetica e della proprietà (c) della deviazione standard. Ci si può domandare se il processo B presenti una variabilità maggiore rispetto al processo A. La risposta è negativa, perché le due serie sono strutturalmente identiche e differiscono unicamente per la scala (ossia le osservazioni in B sono proporzionali a quelle in A).

Continuando con l'esempio, si consideri un terzo processo produttivo in cui si osservano 5 manufatti che hanno i seguenti pesi (in g):

$$\text{C}: 501.98, 502.03, 502.01, 501.99, 502.00.$$

Si noti che questi pesi sono derivati dal campione A aggiungendo 500 a ciascun valore. La media e la deviazione standard sono 502.002 g e 0.0172 g (esse si possono ottenere dalla media e dalla deviazione standard calcolate in A applicando la proprietà (e) alla media e la proprietà (b) alla deviazione standard).

Anche in questo caso, la domanda è se si possa o meno affermare, in base ai valori della deviazione standard, che il grado di variabilità sia lo stesso nei processi A e C. La risposta è no, perché la variabilità è assai minore nel processo C che non nel processo A, una volta tenuto conto del fatto che l'ordine di grandezza delle osservazioni è di gran lunga superiore in C che non in A.

Alla luce delle osservazioni precedenti, si può concludere che per un adeguato confronto della variabilità dei tre insiemi di dati, è necessario esprimere la deviazione standard come percentuale della media. Quindi, si ottiene:

$$\text{Processo A}: \frac{0.0172}{2.002} \cdot 100 = 0.86, \quad \text{Processo B}: \frac{4.3012}{500.5} \cdot 100 = 0.86,$$

$$\text{Processo C}: \frac{0.0172}{502.002} \cdot 100 = 0.0034.$$

Se ne deduce che gli insiemi di dati A e B presentano lo stesso grado di variabilità, mentre C presenta una variabilità minore rispetto a quella di A.

L'esempio mostra che gli indici di variabilità fin qui illustrati, denominati indici assoluti di variabilità, non sono adatti per confrontare la variabilità di insiemi di dati che hanno ordini di grandezza molto diversi. La stessa conclusione si applica quando si confrontano due serie di dati con diverse unità di misura, ad esempio centimetri e chilogrammi. In questo caso, gli indici assoluti di variabilità sono espressi in unità di misura differenti, di conseguenza non sono confrontabili.

DEFINIZIONE 5.6

Un indice percentuale di variabilità è definito come il rapporto, moltiplicato per 100, tra uno degli indici assoluti di variabilità introdotti in precedenza e la media aritmetica. Si ha così lo scostamento semplice medio percentuale, la deviazione standard percentuale, la differenza media percentuale, il campo di variazione percentuale e la differenza interquartile percentuale.

Un ruolo importante nelle applicazioni ha la deviazione standard percentuale, nota come coefficiente di variazione (qui si assume che $\mu > 0$):

$$CV = \frac{\sigma}{\mu} \cdot 100. \tag{5.10}$$

Con il calcolo dell'indice percentuale si elimina, in un certo senso, l'influenza che il livello medio della variabile ha sulla misura della variabilità: il rapporto (5.10) è infatti la soluzione in x della proporzione

$$\sigma : \mu = x : 100,$$

da cui si evince che CV è il livello di variabilità che corrisponde a una media posta uguale a 100. Considerazioni identiche valgono per gli altri possibili indici percentuali. Chiaramente, essi sono puri numeri, ossia non dipendono dall'unità di misura della variabile oggetto di studio.

Gli indici percentuali di variabilità, diversamente dagli indici assoluti, sono puri numeri che consentono di mettere a confronto la variabilità di due o più insiemi di dati espressi in unità di misura differenti o aventi medie sensibilmente diverse.

Vale la pena osservare che calcolare un indice percentuale di variabilità per la serie di osservazioni $x_1, x_2, \ldots, x_n$ è equivalente a calcolare l'indice di variabilità assoluto per la serie:

$$\frac{x_1}{\mu} \cdot 100, \ \frac{x_2}{\mu} \cdot 100, \ \ldots, \ \frac{x_n}{\mu} \cdot 100.$$

Ciò è immediatamente verificabile per ogni misura assoluta di variabilità, applicando la proprietà (c) con $b = (1/\mu) \cdot 100$.

ESEMPIO 5.9

La media e la deviazione standard delle quotazioni giornaliere dei titoli azionari A e B, durante l'ultimo anno, sono state:

$$\text{A: } \mu = 9.89, \ \sigma = 1.16,$$
$$\text{B: } \mu = 3.03, \ \sigma = 0.57.$$

Si consideri un risparmiatore che voglia investire i propri risparmi in uno dei due titoli dando la preferenza a quello che presenta una minore "volatilità" (grado di oscillazione delle quotazioni). Quale indice dovrebbe utilizzare?

Soluzione

La deviazione standard non è appropriata per confrontare la "volatilità" dei due titoli, perché tale indice dipende dall'ordine di grandezza delle quotazioni e le quotazioni di A hanno in media un ordine di grandezza superiore al triplo di quello delle quotazioni di B. L'indice appropriato è il coefficiente di variazione che assume i valori di seguito indicati:

$$\text{A: } CV = 11.73\%,$$
$$\text{B: } CV = 18.81\%.$$

Su questa base, il risparmiatore potrebbe orientarsi verso il primo titolo, le cui quotazioni presentano una minore "volatilità". ■

5.6 Indici di disuguaglianza o di concentrazione

Questo paragrafo è dedicato alla variabilità di dati relativi a un particolare tipo di variabili quantitative, le cosiddette variabili trasferibili, quali reddito, ricchezza ecc., per cui sono ipotizzabili trasferimenti di valore da unità meno dotate a

unità più dotate o viceversa. Ad esempio, i redditi delle famiglie di un Paese possono essere modificati attraverso politiche governative di redistribuzione con cui vengono trasferite quote di reddito da famiglie ricche a famiglie povere.

Nella terminologia economica, si dice che una variabile trasferibile X è tanto più concentrata quanto minore è la frazione sul totale della variabile posseduta dalle unità "più povere", cioè dalle unità in cui X è al di sotto di un dato limite. Ad esempio, la ricchezza di un Paese è tanto più concentrata quanto minore è la frazione della ricchezza totale posseduta dalle famiglie meno abbienti. Gli indici di variabilità per queste variabili sono chiamati indici di disuguaglianza o indici di concentrazione.

Siano $x_1, x_2, \ldots, x_n$ le osservazioni, assunte tutte non negative, di una variabile trasferibile X su n unità (ad esempio, il reddito delle 500 famiglie di una piccolo comune) avente media μ. Sia $x_1 + x_2 + \cdots + x_n = n \cdot \mu$ la somma di questi valori. Siano $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ le stesse quantità disposte in ordine non decrescente. Si può pensare a questa serie di osservazioni come a uno specifico elemento dell'insieme delle possibili serie di n osservazioni di X aventi lo stesso totale $n \cdot \mu$. In questo insieme, vi sono due situazioni estreme: la prima, quando X assume lo stesso valore μ su tutte le n unità:

$$x_{(1)} = x_{(2)} = \cdots = x_{(n)} = \mu;$$

la seconda, quando X assume il valore 0 su $n - 1$ unità e il valore $n \cdot \mu$ su una sola unità:

$$x_{(1)} = x_{(2)} = \cdots = x_{(n-1)} = 0 \text{ e } x_{(n)} = n \cdot \mu.$$

Nel primo caso, si parla di equidistribuzione; nel secondo, di massima concentrazione. Questa idea è illustrata nell'Esempio 5.10.

ESEMPIO 5.10

Si considerino i redditi mensili di cinque famiglie (dati ipotetici): 3.5, 6, 7, 9.5, 12, il cui totale è 38. L'insieme delle possibili serie di cinque osservazioni con lo stesso totale è schematizzato nella Tabella 5.4.

Tabella 5.4
Insieme delle serie di 5 osservazioni il cui totale è 38.

1	7.6	7.6	7.6	7.6	7.6
2	5.2	7.6	7.6	7.6	10.0
⋮	⋮	⋮	⋮	⋮	⋮
r	3.5	6	7	9.5	12
⋮	⋮	⋮	⋮	⋮	⋮
∞	0	0	0	0	38

Dunque, l'r-esima serie si colloca idealmente in tale insieme delimitato da due situazioni estreme: la serie i cui termini sono tutti uguali (equidistribuzione) e quella opposta di massima concentrazione. ■

5.6.1 Rapporto di concentrazione di Gini

Data la serie di osservazioni $x_1, x_2, \ldots, x_n$ di una variabile trasferibile X e la corrispondente serie ordinata $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, si consideri la somma

$$A_i = x_{(1)} + x_{(2)} + \cdots + x_{(i)},$$

che rappresenta l'ammontare della variabile X posseduto dalle prime i unità, disposte in ordine non decrescente del loro valore di X. Ovviamente, $A_n = n \cdot \mu$ è la somma delle n osservazioni, denominata ammontare totale della variabile. Si definiscano i rapporti:

$$P_i = \frac{i}{n}, \quad Q_i = \frac{A_i}{A_n}, \quad i = 1, 2, \ldots, n,$$

che rappresentano la frazione di unità in cui $X \le x_{(i)}$ e la frazione dell'ammontare totale della variabile posseduta da tali unità. Si noti che nel caso di equidistribuzione $A_i = i \cdot \mu$ e quindi $Q_i = (i \cdot \mu)/(n \cdot \mu) = P_i$, mentre, nel caso di massima concentrazione, $Q_1 = Q_2 = \cdots = Q_{n-1} = 0$, $Q_n = 1$. Perciò, quanto più le Q_i differiscono dalle P_i, $i = 1, 2, \ldots, n - 1$, tanto maggiore è la disuguaglianza nei dati. Sulla base di queste considerazioni, viene ora data la definizione della misura di disuguaglianza nota come *rapporto di concentrazione di Gini*.

DEFINIZIONE 5.7

Sia X una variabile trasferibile a valori non negativi e sia $x_1, x_2, \ldots, x_n$ una serie di osservazioni della variabile X. Allora una misura di concentrazione di tale serie è data dal rapporto di concentrazione di Gini:

$$G = \frac{\sum_{i=1}^{n-1} (P_i - Q_i)}{\sum_{i=1}^{n-1} P_i}. \tag{5.11}$$

Il rapporto di concentrazione di Gini è l'indice universalmente utilizzato per la misura della disuguaglianza nella distribuzione dei redditi.

Si osservi che vale la disuguaglianza

$$\sum_{i=1}^{n-1} (P_i - Q_i) \le \sum_{i=1}^{n-1} P_i.$$

L'espressione a sinistra del segno di disuguaglianza è essa stessa una misura di concentrazione, che è nulla nel caso di equidistribuzione ed è massima, cioè pari a $\sum_{i=1}^{n-1} P_i$, nel caso di massima concentrazione. Ne segue che il rapporto (5.11) è un indice di concentrazione che varia tra 0 e 1, essendo uguale a 0 nel caso di equidistribuzione e uguale a 1 nel caso di massima concentrazione. Per questa proprietà, l'indice G, diversamente dagli indici di variabilità assoluti, è idoneo per il confronto della concentrazione di due o più serie di osservazioni.

La (5.11) può essere posta nella forma[2]

$$G = \frac{2}{n-1} \sum_{i=1}^{n-1} (P_i - Q_i). \tag{5.12}$$

[2] Ciò perché

$$\sum_{i=1}^{n-1} P_i = \frac{1 + 2 + \cdots + n - 1}{n},$$

da cui, applicando la formula della somma di termini in progressione aritmetica, si ottiene

$$\sum_{i-1}^{n-1} P_i = \frac{1}{n}\left[\frac{1 + (n-1)}{2}(n-1)\right] = \frac{n-1}{2}.$$

ESEMPIO 5.11

Si calcoli l'indice G per l'r-esima serie dell'Esempio 5.10 usando la (5.12).

Soluzione

Il calcolo può essere impostato convenientemente secondo lo schema indicato nella Tabella 5.5.

Tabella 5.5
Calcolo del rapporto di concentrazione G sui dati dell' Esempio 5.10.

i	$x_{(i)}$	A_i	P_i	Q_i	$P_i - Q_i$
1	3.5	3.5	0.2	0.09	0.11
2	6.0	9.5	0.4	0.25	0.15
3	7.0	16.5	0.6	0.43	0.17
4	9.5	26.0	0.8	0.68	0.12
5	12.0	38.0	1.0	1.00	–
Totale	38.0	–	–	–	0.55

Applicando la (5.12), si ha

$$G = \frac{2}{4} \cdot 0.55 = 0.28.$$

■

PROPOSIZIONE 5.3

L'indice di Gini gode delle proprietà di seguito indicate:

(a) assume valori nell'intervallo [0, 1]; è uguale a 0 nel caso di equidistribuzione, è uguale a 1 nel caso di massima concentrazione;

(b) diminuisce se si aggiunge a ogni osservazione della serie una quantità positiva;

(c) non cambia se ogni osservazione della serie è moltiplicata per una costante positiva;

(d) soddisfa il principio di trasferibilità: date due osservazioni, $x_{(i)}$ e $x_{(j)}$, con $x_{(i)} < x_{(j)}$, se si diminuisce $x_{(i)}$ di un ammontare τ e si aumenta $x_{(j)}$ dello stesso ammontare (mantenendo l'ordinamento delle osservazioni), G aumenta.

La dimostrazione di queste proprietà è contenuta nel Paragrafo A.3 dell'Appendice A. ■

Interpretazione geometrica del rapporto di concentrazione

La (5.11) si presta a una interessante interpretazione geometrica tramite la curva di concentrazione, nota anche come curva di Lorenz.

Data una serie di osservazioni $x_1, x_2, \ldots, x_n$, si rappresentino su di un piano cartesiano i punti di coordinate (P_i, Q_i), $i = 1, 2, \ldots, n$. Si chiama curva di concentrazione la spezzata che si ottiene congiungendo con segmenti di retta le coppie consecutive dei punti di coordinate $(0, 0), (P_1, Q_1), (P_2, Q_2), \ldots, (P_n, Q_n)$, come è mostrato nella Figura 5.4.

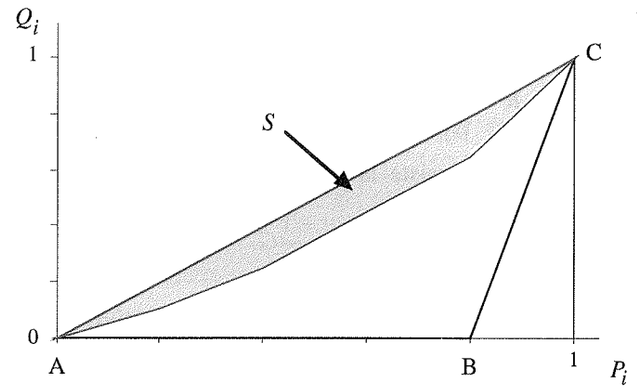


Figura 5.4
Curva di concentrazione.

Nella Proposizione 5.4, la (5.11) viene riesaminata e interpretata in termini geometrici. Nel caso di equidistribuzione, la curva si riduce al segmento di retta AC, detto segmento di equidistribuzione, avente come estremi i punti $(0, 0)$ e $(1, 1)$. Nel caso di massima concentrazione, la curva coincide con la spezzata ABC, che unisce i punti di coordinate:

$$\mathrm{A} = (0, 0), \quad \mathrm{B} = \left(\frac{n-1}{n}, 0\right), \quad \mathrm{C} = (1, 1),$$

spezzata che viene chiamata curva di massima concentrazione.

PROPOSIZIONE 5.4

Sia S la superficie racchiusa tra il segmento di equidistribuzione e la curva di concentrazione. Sia $\max S$ il massimo di S, dato dall'area racchiusa tra il segmento di equidistribuzione e la curva di massima concentrazione. Allora, il rapporto di concentrazione G definito dalla (5.11) può essere espresso nella forma

$$G = \frac{S}{\max S},$$

dove

$$\max S = \frac{1}{2} \cdot \frac{n-1}{n}.$$

La dimostrazione è contenuta nel Paragrafo A.3 dell'Appendice A. ■

La quantità $\max S$ non è altro che l'area del triangolo ABC che ha altezza 1 e base $(n-1)/n$ (lunghezza del segmento AB). L'espressione $G = S/\max S$ aiuta a capire il significato dell'indice: poiché $\max S$ è costante, si può dire che G cresce con S e quindi è tanto più grande quanto più la curva di concentrazione si allontana dal segmento di equidistribuzione.

Quando n è sufficientemente grande, l'area di massima concentrazione è approssimativamente uguale a 1/2, essendo $(n-1)/n \approx 1$. Con questa approssimazione, la (5.12) diviene

$$R = \frac{2}{n} \sum_{i=1}^{n-1} (P_i - Q_i),$$

espressione che (posto $P_0 = Q_0 = 0$), come è facile verificare, può essere scritta nella forma

$$R = \sum_{i=1}^{n} \frac{1}{n}[(P_{i-1} - Q_{i-1}) + (P_i - Q_i)]. \tag{5.13}$$

Tra i due indici, G e R, sussiste la relazione di proporzionalità di immediata verifica

$$R = \frac{n-1}{n} G.$$

5.6.2 Rapporto di concentrazione delle distribuzioni di frequenza

Le formule appena introdotte per il calcolo degli indici G ed R valgono per serie di osservazioni. Diversamente da quanto osservato per le medie e per gli indici di variabilità, non è immediato adattare tali formule al caso delle distribuzioni di frequenza.

PROPOSIZIONE 5.5

Si consideri la distribuzione di frequenza di una variabile trasferibile X. Sia N_i la frequenza cumulata di x_i, e sia

$$A'_i = x_1 \cdot n_1 + x_2 \cdot n_2 + \cdots + x_i \cdot n_i$$

l'ammontare della variabile X posseduto dalle unità in cui $X \le x_i$. Si considerino i due rapporti:

$$P'_i = \frac{N_i}{n}, \quad Q'_i = \frac{A'_i}{A'_k},$$

dove $A'_k = \sum_{j=1}^{k} x_j n_j = n \cdot \mu$ è l'ammontare totale della variabile. Il rapporto di concentrazione R assume la forma

$$R = \sum_{i=1}^{k} [(P'_{i-1} - Q'_{i-1}) + (P'_i - Q'_i)] \frac{n_i}{n}, \tag{5.14}$$

dove $P'_0 = Q'_0 = 0$.

La dimostrazione è omessa per brevità. ■

La (5.14) presenta la stessa struttura della (5.13), da cui si può dedurre per analogia: nel caso di una distribuzione di frequenza infatti la frequenza relativa associata a x_i (che è uguale a $1/n$ nel caso serie di osservazioni) è n_i/n, mentre P'_i e Q'_i sostituiscono le omologhe quantità P_i e Q_i.

ESEMPIO 5.12

Si calcoli l'indice R per la distribuzione di frequenza che segue:

Tabella 5.6
Aziende di credito di una certa regione secondo il numero degli sportelli operanti nella regione.

Numero di sportelli	5	8	15	20	52
Aziende di credito	2	3	2	4	1

Soluzione

I calcoli sono esposti nella Tabella 5.7, dove, nell'ultima colonna, è contenuta la quantità

$$\pi_i = [(P'_{i-1} - Q'_{i-1}) + (P'_i - Q'_i)] \cdot n_i.$$

Tabella 5.7
Calcolo dell'indice R per la distribuzione riportata nella Tabella 5.6.

N. di sportelli x_i	Aziende di credito n_i	$x_i \cdot n_i$	N_i	A'_i	P'_i	Q'_i	$P'_i - Q'_i$	π_i
5	2	10	2	10	0.17	0.05	0.12	0.24
8	3	24	5	34	0.42	0.17	0.25	1.11
15	2	30	7	64	0.58	0.33	0.25	1.00
20	4	80	11	144	0.92	0.73	0.19	1.76
52	1	52	12	196	1.00	1.00	–	0.19
Totale	12	196	–	–	–	–	–	4.30

Si ha allora

$$R = \frac{4.30}{12} = 0.36.$$

La curva di concentrazione è mostrata nella Figura 5.5.

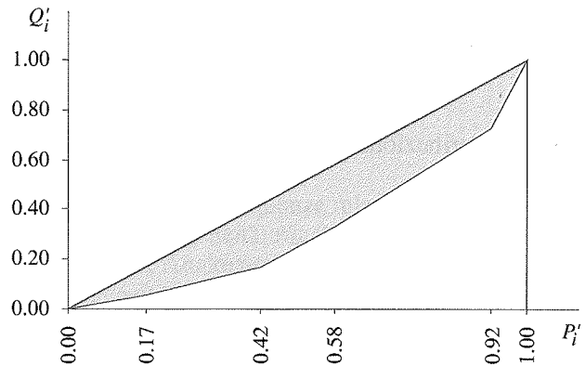


Figura 5.5
Curva di concentrazione della distribuzione di cui alla Tabella 5.6.

5.6.3 Il caso delle distribuzioni di frequenza in classi

La (5.14) si applica anche nel caso delle distribuzioni di frequenza in classi. Occorre distinguere due casi: (a) è nota per ogni classe la somma delle modalità della variabile X per le unità appartenenti alla classe, somma denominata ammontare totale di classe e indicata con t_i, $i = 1, 2, \ldots, k$; (b) non si dispone di questo dato. Nel primo caso, le quantità A'_i, necessarie per il calcolo di Q'_i, sono date da

$$A'_i = t_1 + t_2 + \cdots + t_i, \quad i = 1, 2, \ldots, k.$$

Nel secondo, l'ammontare totale di classe t_i viene approssimato mediante il prodotto $\bar{x}_i \cdot n_i$, dove $\bar{x}_i$ è il valore centrale della classe i-esima. Chiaramente, A'_i è dato da

$$A'_i = \bar{x}_1 \cdot n_1 + \bar{x}_2 \cdot n_2 + \cdots + \bar{x}_i \cdot n_i.$$

L'applicazione della (5.14) è poi immediata in entrambi i casi.

ESEMPIO 5.13

Si calcoli l'indice di concentrazione R per la distribuzione di frequenza dei comuni di una regione italiana secondo la superficie (Fonte: Istat), riportata nelle prime due colonne della Tabella 5.8. La terza colonna contiene in ciascuna riga l'ammontare totale di classe, cioè la superficie complessiva dei comuni appartenenti a quella classe.

Soluzione

I calcoli sono presentati nella tabella che segue:

Tabella 5.8
Calcolo dell'indice R per la distribuzione di frequenza dei comuni di una regione italiana classificati secondo la superficie.

Sup.	Numero comuni.	Sup. Totale t_i	N_i	A'_i	P'_i	Q'_i	$P'_i - Q'_i$	π_i
≤ 1000	8	6 029	8	6 029	0.11	0.02	0.09	0.72
1001-2000	14	21 024	22	27 053	0.30	0.08	0.22	4.34
2001-4000	27	78 445	49	105 498	0.66	0.32	0.34	15.12
4001-6000	9	44 871	58	150 369	0.78	0.46	0.32	5.94
6001-10 000	8	57 034	66	207 403	0.89	0.64	0.25	4.56
10 001-25 000	8	118 944	74	326 347	1.00	1.00	–	2.00
Totale	74	326 347	–	–	–	–	–	32.68

La quantità π_i posta nell'ultima colonna è stata già definita nell'Esempio 5.12. L'indice R si ottiene dal rapporto

$$R = \frac{32.68}{74} = 0.44.$$

Se ne deduce che la distribuzione considerata presenta una concentrazione di livello intermedio. ■

5.7 Eterogeneità

Per le distribuzioni di frequenza secondo una variabile nominale è possibile introdurre indici che misurino il grado di diversità delle osservazioni. È opportuno partire dalla definizione di omogeneità ed eterogeneità.

DEFINIZIONE 5.8

Si consideri una distribuzione di frequenza secondo una variabile nominale avente k modalità. Essa presenta la minima eterogeneità, ovvero la massima omogeneità, quando tutte le unità hanno la stessa modalità. Al contrario, si dice che essa presenta la massima eterogeneità quando le unità sono equiripartite tra le k modalità.

La definizione è del tutto plausibile: quando la variabile assume la stessa modalità su tutte le unità, si può dire che vi è massima omogeneità. A mano a mano che ci si allontana da questa situazione, cresce l'eterogeneità; la situazione di massima eterogeneità si realizza quando le unità sono ripartite tra le categorie in parti uguali. In simboli, le due situazioni estreme sono rappresentate nelle Tabelle 5.9 e 5.10.

Modalità	x_1	x_2	...	x_i	...	x_k	Totale
Frequenza	0	0	...	n	...	0	n

Tabella 5.9
Distribuzione di frequenza nel caso di massima omogeneità.

Modalità	x_1	x_2	...	x_i	...	x_k	Totale
Frequenza	n/k	n/k	...	n/k	...	n/k	n

Tabella 5.10
Distribuzione di frequenza nel caso di massima eterogeneità.

Vengono ora definite due misure di eterogeneità.

DEFINIZIONE 5.9

Sia data una distribuzione di frequenza relativa di una variabile nominale con k modalità

Modalità	x_1	x_2	...	x_k	Totale
Freq. rel.	f_1	f_2	...	f_k	1

due indici di eterogeneità sono espressi dalle formule che seguono:

$$e_1 = 1 - \sum_{i=1}^{k} f_i^2, \tag{5.15}$$

$$e_2 = -\sum_{i=1}^{k} f_i \ln(f_i), \tag{5.16}$$

dove f_i, $i = 1, 2, \ldots, k$, sono le frequenze relative. L'indice (5.15) raggiunge il minimo, 0, nel caso di omogeneità; raggiunge il massimo, $(k-1)/k$, nel caso di massima eterogeneità. Analogamente, l'indice (5.16) raggiunge il minimo, 0, nel caso di omogeneità; raggiunge il massimo, $\ln(k)$, nel caso di massima eterogeneità.

ESEMPIO 5.14

Nella tabella che segue è riportata la distribuzione percentuale della popolazione francese secondo il credo religioso (Fonte: Eurobarometro, 2018):

Religione	Percentuale
Cattolicesimo	50
Nessuna religione	38
Islam	5
Altre religioni	7
Totale	100

Tabella 5.11
Popolazione francese per religione.

Si calcoli l'indice e_1.

Soluzione

Il calcolo fornisce

$$e_1 = 1 - 0.50^2 - 0.38^2 - 0.05^2 - 0.07^2 = 0.60.$$

Se ne deduce che la distribuzione presenta un'eterogeneità piuttosto elevata (il massimo di e_1 è 0.75). ■

5.8 Quale indice impiegare

Analogamente a quanto fatto per le medie, ha senso chiedersi quale indice impiegare per la misura della variabilità nelle situazioni reali. Va subito detto che l'indice che trova più frequente applicazione è la deviazione standard. La ragione va ricercata anche nel ruolo che l'indice e le sue trasformazioni (varianza e devianza) hanno assunto tradizionalmente nell'inferenza statistica.

In pratica, quando un indice di variabilità viene utilizzato nell'ambito della statistica descrittiva, non esiste una regola assoluta per scegliere tra scostamento semplice medio, deviazione standard e differenza media. Il campo di variazione e la differenza interquartile trovano un'applicazione limitata, anche perché i requisiti dell'immediatezza e della facilità di calcolo non sono oggi dirimenti.

Quanto agli indici percentuali, la scelta di uno di essi, ad esempio del coefficiente di variazione, è necessaria quando si confronta la variabilità di due o più insiemi di dati espressi in unità di misura differenti o aventi ordini di grandezza molto diversi.

Gli indici di concentrazione e di eterogeneità sono due misure particolari di variabilità: i primi vengono utilizzati per dati di variabili trasferibili, come il reddito o la ricchezza familiare, i secondi quando si hanno distribuzioni di frequenza di variabili nominali.

Riassunto

Gli indici di variabilità sono uno strumento con cui viene espresso numericamente il grado di diversità presente in un insieme di osservazioni. In un certo senso, la funzione degli indici di variabilità è quella di completare l'informazione fornita dalle medie. Per esempio, se si vuole mettere a confronto il benessere economico di due Stati, è utile disporre, oltre che del reddito medio, del grado di disuguaglianza della distribuzione dei redditi. E ancora, se si deve decidere il vestiario da mettere in valigia in vista di una vacanza in un paese lontano, è importante conoscere non solo il livello medio della temperatura, ma anche le sue oscillazioni (nella giornata, nella settimana ecc.).

Gli indici di variabilità possono essere distinti in tre categorie a seconda che esprimano la variabilità dei dati come:

- media delle distanze delle osservazioni dalla loro media aritmetica (scostamento semplice medio, deviazione standard);
- media di tutte le differenze, prese in valore assoluto, tra le coppie di osservazioni (differenze medie);
- differenza tra osservazioni che occupano posizioni specifiche o tra medie di posizione (campo di variazione, differenza interquartile).

Tutti questi indici sono misure di variabilità inadatte per il confronto della variabilità di due o più insiemi di dati espressi in unità di misura diverse, o aventi ordini di grandezza molto diversi. Per consentire confronti adeguati, sono necessari gli indici percentuali di variabilità, dati dal rapporto (moltiplicato per 100) tra uno degli indici di variabilità sopra indicati (scostamento semplice medio, deviazione standard, differenza media ecc.) e la media.

Altri indici di variabilità particolari sono gli indici di concentrazione, che sono definiti per variabili trasferibili, come il reddito familiare e la ricchezza. In questo contesto, il cosiddetto rapporto di concentrazione di Gini è universalmente utilizzato per descrivere la disuguaglianza nelle distribuzioni del reddito.

Infine, una breve trattazione è dedicata al problema della misurazione della variabilità (eterogeneità) di una distribuzione di frequenza di una variabile nominale. Al riguardo, vengono introdotti due indici basati sulle frequenze relative.

Parole chiave

ammontare totale della variabile, 141
ammontare totale di classe, 145
campo di variazione, 136
coefficiente di variazione, 139
concentrazione, 140
curva di concentrazione, 142
curva di Lorenz, 142
curva di massima concentrazione, 143
devianza, 130
deviazione standard, 129
differenza interquartile, 137
differenza semplice media, 134
equidistribuzione, 140
eterogeneità, 146
formula operativa della deviazione standard, 129
indici assoluti di variabilità, 138
indici di eterogeneità, 147
indici di concentrazione, 140
indici percentuali di variabilità, 138
massima concentrazione, 140
omogeneità, 146
principio di trasferibilità, 142
rapporto di concentrazione di Gini, 141
scostamento quadratico medio, 129
scostamento semplice medio, 128
segmento di equidistribuzione, 143
variabili trasferibili, 139
varianza, 130

Commenti e curiosità

1. Forma della distribuzione e misura della variabilità

È interessante saper "predire" l'ordine di grandezza della deviazione standard a partire dalla struttura del grafico della distribuzione di frequenza.

Le distribuzioni di frequenza in classi rappresentate nella Figura 5.6 hanno stesse classi, stessa media aritmetica, 60, e frequenze di classe diverse; quindi, gli scarti dei valori centrali delle classi dalla media sono −35, −25, −15, −5, 5, 15, 25 e 35 in tutte le distribuzioni. La deviazione standard, essendo una media (quadratica) di queste quantità, oscillerà tra un minimo di 5 e un massimo di 35 al variare delle frequenze di classe. Assumerà valori elevati quando sono relativamente più alti gli istogrammi delle classi più estreme; viceversa, assumerà valori piccoli quando le densità di frequenza sono maggiori per le classi centrali. Infatti, la singola densità di frequenza funge da peso nella determinazione di σ "attraendo" l'indice verso lo scarto dalla media (in valore assoluto) del valore centrale a cui la densità di frequenza si riferisce. Con questo criterio di lettura, è facile comparare la deviazione standard di due distribuzioni e valutarne finanche il valore.

2. Media aritmetica e deviazione standard, una coppia inscindibile

Il binomio media aritmetica-deviazione standard si incontra frequentemente nei rapporti di ricerca e nelle analisi quantitative nei più svariati campi. L'abbinamento tra le due quantità è del tutto naturale se si considera che la media aritmetica è l'indicatore principe deputato a fornire un'idea dell'ordine di grandezza del fenomeno oggetto di studio, come la deviazione standard è l'indice generalmente impiegato per quantificare il grado di diversità tra le manifestazioni del fenomeno. Tuttavia, esso è pienamente informativo solo quando la distribuzione di frequenza si comporta approssimativamente secondo il modello di Gauss (si veda Paragrafo 6.3). Infatti, come si vedrà nel Capitolo 14, si può affermare che: circa il 68% dei valori osservati cade nell'intervallo $(\mu - \sigma, \mu + \sigma)$, circa il 95% nell'intervallo $(\mu - 2\sigma, \mu + 2\sigma)$, circa il 99,7% nell'intervallo $(\mu - 3\sigma, \mu + 3\sigma)$.

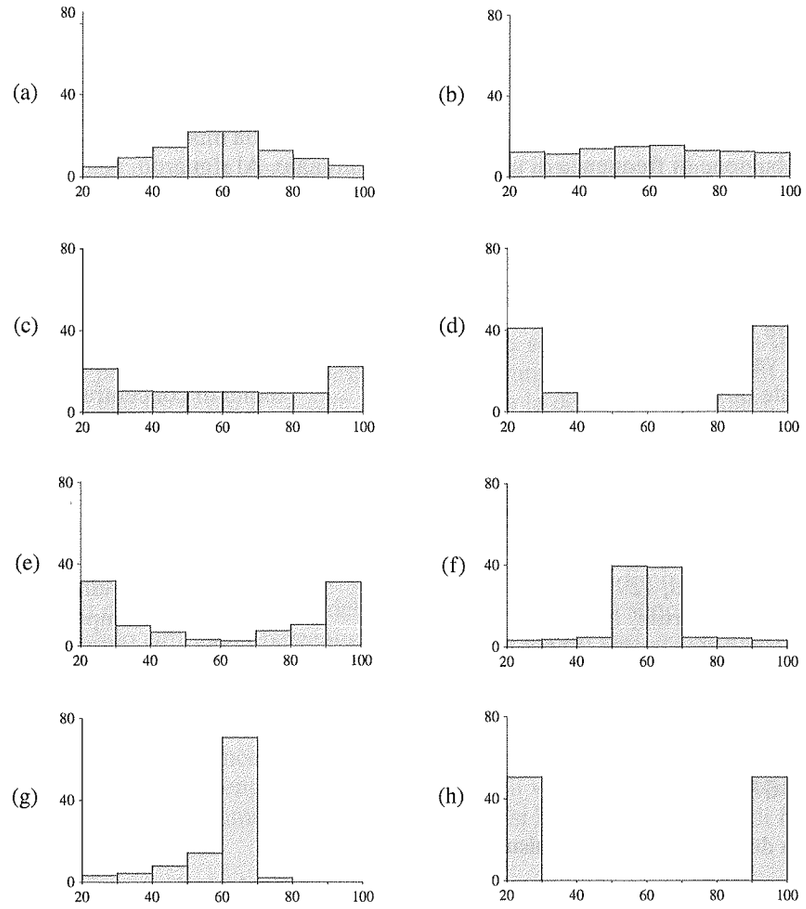


Figura 5.6
Distribuzione con la stessa media ($\mu = 60$) e deviazioni standard differenti: 17.7 in (a), 22 in (b), 26.3 in (c), 33.5 in (d), 30.3 in (e), 12.5 in (f), 10 in (g), 35 in (h).

3. Standardizzazione

È opportuno soffermarsi brevemente sul significato di una operazione che verrà effettuata in seguito, quella di dividere gli scarti, $x_i - \mu$, per la deviazione standard σ. Si consideri dapprima la quantità

$$\frac{|x_i - \mu|}{\sigma}.$$

Questa può essere interpretata come la distanza dell'osservazione dalla media aritmetica rapportata alla distanza media che le n osservazioni presentano rispetto alla medesima media. Così, se

$$\frac{|x_i - \mu|}{\sigma} > 1,$$

si dirà che l'osservazione i-esima è "più deviante" da μ di quanto non lo sia in media l'insieme di tutte le osservazioni. All'opposto, si dirà che l'osservazione i-esima è "meno deviante" da μ se

$$\frac{|x_i - \mu|}{\sigma} < 1.$$

Un'interpretazione simile si ha per la quantità

$$\frac{x_i - \mu}{\sigma}.$$

In questo caso, il segno dà l'informazione ulteriore della posizione dell'osservazione rispetto alla media: se negativo indica che x_i è alla sinistra di μ, se positivo che x_i è alla destra di μ.

Data la serie di osservazioni $2, 3, 3, 4, 5$ avente media 3.4 e deviazione standard 1.02, le osservazioni standardizzate che se ne deducono sono: $-1.37, -0.39, -0.39, 0.59, 1.57$. In generale, alla serie di osservazioni $x_1, x_2, \ldots, x_n$ corrisponde la serie standardizzata

$$\frac{x_1 - \mu}{\sigma}, \frac{x_2 - \mu}{\sigma}, \ldots, \frac{x_n - \mu}{\sigma}.$$

È facile verificare che ogni serie standardizzata ha media aritmetica 0 e deviazione standard 1.

4. La strategia del "sei sigma"

Il controllo della variabilità delle caratteristiche degli oggetti prodotti in un processo industriale è un problema molto serio: da tale controllo dipende la qualità dei manufatti e quindi la capacità di soddisfare i bisogni del cliente. Per ogni caratteristica (peso, resistenza ecc.) dell'oggetto prodotto vengono fissati i limiti di oscillazione, detti "specifiche". Se il singolo pezzo non rientra nelle specifiche, esso viene qualificato come difettoso.

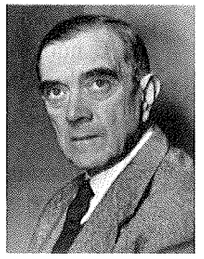
Le fonti della variabilità delle caratteristiche dei pezzi prodotti sono molteplici: il fattore umano, la qualità delle materie prime, le macchine e così via. In queste condizioni, se si considera un numero elevato di pezzi prodotti nelle stesse condizioni e si misura su di essi una data caratteristica, la distribuzione statistica che ne risulta tende a disporsi secondo la curva di Gauss. Ne segue che gli intervalli

$$(\mu - \sigma, \mu + \sigma),\ (\mu - 2\sigma, \mu + 2\sigma),\ (\mu - 3\sigma, \mu + 3\sigma),$$

includono, approssimativamente, il 68%, il 95% e il 99,7% dei casi.

La strategia del "sei sigma" consiste nell'organizzare l'attività produttiva perseguendo l'obiettivo di livelli eccellenti di qualità tali che le difettosità quasi scompaiano: si tratta di controllare la variabilità del processo in modo che il numero di pezzi difettosi sia dell'ordine di 6.8 per ogni milione di pezzi prodotti. Si osservi che questa frazione (0.0000068) è la frequenza relativa con cui si presenta un valore esterno all'intervallo $(\mu - 4.5\sigma, \mu + 4.5\sigma)$ nella curva di Gauss. Ciò significa che si deve diminuire σ in modo tale che gli estremi di questo intervallo coincidano con i limiti di specifica.

5. Corrado Gini e la misura della concentrazione dei redditi



Corrado Gini (1884-1965), conseguita la laurea in giurisprudenza a Bologna, intraprese una carriera universitaria dagli esiti assai brillanti: a soli 26 anni vinse la cattedra di statistica presso l'Università di Cagliari. Insegnò poi presso l'Università di Padova e successivamente nell'Università di Roma. Nel 1926 contribuì alla nascita e all'organizzazione dell'Istituto nazionale di statistica, del cui organo di governo – il Consiglio superiore – fu presidente. Nel 1936 fondò – all'Università di Roma – la Facoltà di Scienze statistiche, demografiche e attuariali, la prima in Italia.

Gini fu un uomo geniale, dal carattere forte, con un innato senso di deontologia scientifico-professionale e con interessi multidisciplinari: i suoi contributi scientifici spaziarono infatti dalla statistica all'economia e alla sociologia. In ambito statistico, diede un impulso eccezionale allo sviluppo del metodo statistico creando una vera e propria scuola italiana di statistica. È considerato il più eminente statistico italiano di ogni tempo.

Nella sua amplissima produzione scientifica, primeggia senza dubbio la teoria della concentrazione e in particolare l'indice di concentrazione che porta il suo nome e che è applicato ovunque nel mondo per la misura della disuguaglianza della distribuzione dei redditi e della ricchezza.

6. Area di concentrazione e misura della concentrazione

Per area di concentrazione si intende l'area compresa tra la curva di Lorenz e il segmento di equidistribuzione. La misura della concentrazione può essere valutata osservando l'area di concentrazione.

Nella Figura 5.7 sono rappresentate quattro curve di concentrazione. Dall'esame delle aree di concentrazione è possibile comparare a due a due le diverse situazioni e finanche predire i valori del rapporto di concentrazione.

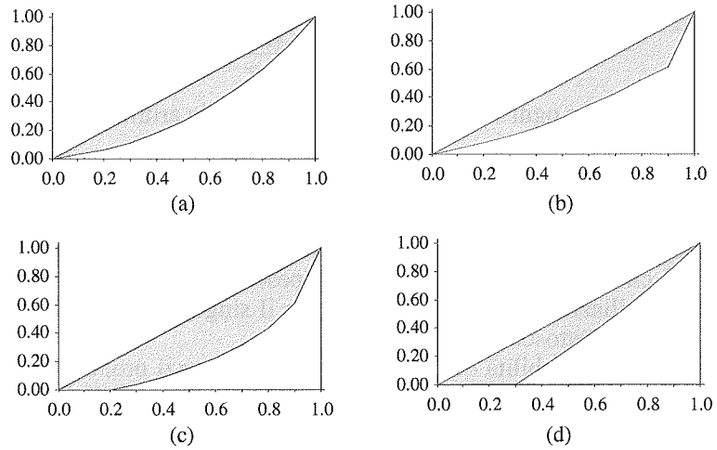


Figura 5.7
Dall'area di concentrazione al valore di *R*.

Esercizi

5.1 I dati che seguono si riferiscono agli incassi giornalieri in euro di un ristorante nei sei giorni di apertura (da martedì a domenica) della prima settimana di giugno:

310, 274, 364, 230, 377, 482.

a) Si determini la deviazione standard e se ne indichi il significato.

b) Si determinino la varianza e la devianza.

c) Si moltiplichi ogni osservazione per 0.4 e si verifichi numericamente che la deviazione standard è pari a quella calcolata al punto a) moltiplicata per 0.4.

5.2 Il nuovo contratto di lavoro dei 50 dipendenti di un'impresa commerciale prevede un aumento dello stipendio mensile nella misura del 5%, uguale per tutti. Con il vecchio contratto la deviazione standard degli stipendi era pari a 263 euro. Si determini la deviazione standard dei nuovi stipendi.

5.3 Utilizzando i dati dell'Esercizio 5.1, si calcoli lo scostamento semplice medio. Si verifichi poi che l'indice non cambia se ad ogni osservazione viene aggiunta la quantità 30.

5.4 Tornando all'Esercizio 2.4, si calcoli la deviazione standard, assumendo che il numero medio dei componenti delle famiglie con 6 o più componenti sia 6.2.

5.5* Si dimostri che vale la seguente identità

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - \mu^2}.$$

(*Suggerimento*: basta sviluppare il quadrato del binomio $(x_i-\mu)^2$, riferire la sommatoria ai singoli elementi dello sviluppo e semplificare.)

5.6* Si dimostri che per gli indici (5.1) e (5.2) valgono le due seguenti proprietà:

a) non cambiano se si aggiunge a ogni osservazione della serie una costante;

b) la moltiplicazione delle osservazioni della serie per una costante positiva o negativa comporta la moltiplicazione degli indici per la costante in valore assoluto.

5.7* Si consideri una serie di osservazioni $x_1, x_2, \ldots, x_n$ avente media μ e deviazione standard σ. Si dimostri che la deviazione standard della serie $x_1, x_2, \ldots, x_n, \mu$ (ottenuta dalla prima aggiungendo una nuova osservazione pari alla media μ) è minore o uguale a σ.

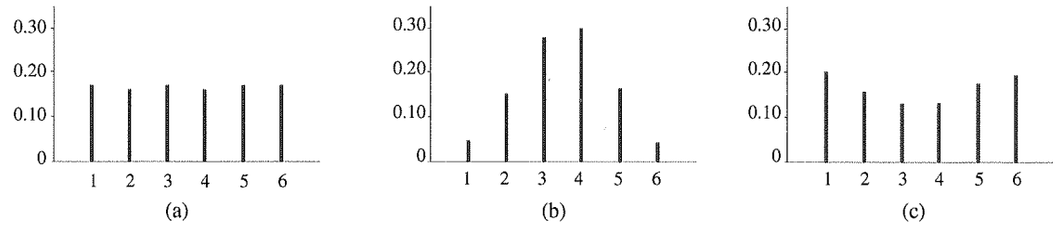
5.8 Le temperature corporee, espresse in gradi Celsius, del gruppo di adulti sani di cui al *file* "Temperature corporee" presentano una deviazione standard pari a 0.34. Quale sarebbe la deviazione standard se le temperature fossero espresse in gradi Fahrenheit? (*Suggerimento*: bisogna tenere presente che dalla scala Celsius si passa alla scala Fahrenheit tramite la relazione °F = 32 + 1.8 °C; inoltre, occorre tenere conto dell'effetto che una trasformazione lineare dei dati ha sulla deviazione standard, si veda l'Esercizio 5.6.)

5.9 Si consideri la distribuzione di frequenza dell'Esercizio 3.3.

a) Si calcolino lo scostamento semplice medio e la deviazione standard.

b) Si spieghi perché il secondo indice è maggiore del primo.

5.10 Nelle tre distribuzioni di seguito rappresentate, aventi la stessa media 3.51, la deviazione standard assume i valori 1.8, 1.2 e 1.7 (indicati in ordine arbitrario).



Si assegni a ciascuna distribuzione l'appropriato valore della deviazione standard.

5.11 Si consideri la distribuzione di frequenza in classi dell'Esercizio 2.3:

Lunghezza Sepalo (cm)	Frequenza
4.25-4.45	4
4.45-4.65	5
4.65-4.85	7
4.85-5.05	12
5.05-5.25	11
5.25-5.45	6
5.45-5.65	2
5.65-5.85	3
Totale	50

Si calcolino lo scostamento semplice medio e la deviazione standard.

5.12 Si consideri la seguente distribuzione di frequenza in classi tratta dalla classificazione secondo la circonferenza toracica degli individui in età fino ai 30 anni (si veda l'Esercizio 1.6):

Circ. toracica	86-90	91-95	96-100	101-105	106-110	111-116	Totale
Frequenza	4	11	7	10	3	3	38

a) Si determini la deviazione standard della distribuzione.

b) Si ripeta il calcolo sulla serie di osservazioni individuali riportata nell'Esercizio 1.6 e si argomenti sulla differenza tra i due risultati.

5.13 Riprendendo l'Esercizio 4.7, si calcoli lo scostamento semplice medio e la deviazione standard. (*Suggerimento*: si usino i dati nell'ultima colonna per calcolare le medie di classe.)

5.14 Si riprenda la distribuzione dei comuni secondo l'ampiezza demografica (Esercizio 4.8).

a) Si determini la deviazione standard della distribuzione utilizzando le medie di classe.

b) Si ripeta il calcolo del punto precedente utilizzando i valori centrali di classe (per l'ultima classe si usi la media di classe) e si confrontino i risultati.

5.15 Riprendendo i dati dell'Esercizio 4.9, si misuri la variabilità della dimensione della famiglia nei quattro Paesi utilizzando la deviazione standard. (*Suggerimento*: si usi il numero di famiglie come peso per calcolare sia la media che la deviazione standard.)

5.16 Si calcoli la differenza semplice media per i dati dell'Esercizio 5.1.

5.17 Si calcoli la differenza semplice media per la distribuzione di frequenza dell'Esercizio 3.3.

5.18 Si calcoli la differenza semplice media per la distribuzione di frequenza dell'Esercizio 5.11.

5.19 Si calcoli la differenza interquartile per la distribuzione di frequenza dell'Esercizio 3.3.

5.20 Si calcoli il coefficiente di variazione per la distribuzione di frequenza dell'Esercizio 5.12.

5.21 Riprendendo i dati dell'Esercizio 1.6, si confronti la variabilità della serie di osservazioni individuali relative ad altezza e peso utilizzando la misura di variabilità più appropriata.

5.22 Comuni italiani per area geografica (Fonte: Istat, anno 2019):

Area geografica	N. di comuni
Nord-Ovest	2996
Nord-Est	1397
Centro	971
Sud	1783
Isole	767
Totale	7914

a) Si disegni la curva di Lorenz.

b) Si calcoli il rapporto di concentrazione G.

5.23 Nella tabella che segue sono riportati i depositi (in milioni di euro) a fine anno presso le banche italiane per tipologia (dati Banca d'Italia).

Tipo di deposito	Importo
Depositi a Risparmio	68 477
Certificati di deposito	31 961
Depositi in conto corrente	467 218
Depositi in valute non euro di residenti in Italia	15 618
Totale	583 274

Si rappresenti la curva di concentrazione e si determini il rapporto di concentrazione G.

5.24 Si riprenda la distribuzione di frequenza dell'Esercizio 4.8.

a) Si rappresenti la curva di concentrazione e si calcoli il rapporto di concentrazione R utilizzando i totali di classe.

b) Si ripeta il calcolo di R utilizzando i valori centrali di classe (per l'ultima classe si prenda la media di classe) e si spieghi perché si ottengono risultati diversi.

5.25 Si calcoli il rapporto di concentrazione R sulla distribuzione di cui all'Esercizio 4.7, utilizzando il totale degli addetti per classe.

5.26 Riprendendo l'Esercizio 5.2, si supponga di applicare l'indice di Gini G alla serie degli stipendi sia del vecchio che del nuovo contratto. Si spieghi perché l'indice rimane invariato. Cosa accadrebbe se il nuovo contratto prevedesse un aumento di €150 per ogni dipendente?

5.27 La seguente distribuzione di frequenza deriva dalla tabella dell'Esercizio 2.12:

Colore degli occhi	Azzurro	Chiaro	Bruno	Scuro	Totale
Frequenza	718	1580	1774	1315	5387

Si calcolino gli indici di eterogeneità e_1 ed e_2.

Indici di variabilità con Excel

Con Excel si possono calcolare lo scostamento semplice medio e la deviazione standard di una serie di osservazioni tramite le funzioni MEDIA.DEV( ) e DEV.ST.POP( ), il cui funzionamento è del tutto simile. Ci si limiterà a illustrare la seconda funzione.

- Si apre il *file* dei dati e si scrive in una casella a scelta la funzione DEV.ST.POP( ) preceduta dall'uguale; all'interno delle parentesi va indicata l'area del foglio in cui sono contenuti i dati.
- Si dà "Invio" e si ottiene il risultato nella stessa casella in cui è scritta la funzione.

Quanto alla differenza media, all'intervallo interquartile e alle misure di concentrazione, non sono disponibili funzioni già predisposte.

CAPITOLO

6

Forma delle distribuzioni di frequenza

Sommario del capitolo

6.1 Introduzione
6.2 Asimmetria
6.3 Curtosi
Riassunto
Commenti e curiosità
Esercizi

6.1 Introduzione

Si pensi alla rappresentazione grafica della distribuzione di frequenza di una variabile quantitativa. Il grafico sarà un diagramma ad aste, come quello in Figura 3.2, se la distribuzione non è in classi, o un istogramma, come quello in Figura 3.8, se la distribuzione è in classi. Un'ispezione del grafico dà la visione immediata della forma della distribuzione.

Due aspetti della forma meritano attenzione: l'asimmetria e la curtosi. L'asimmetria riguarda il maggiore o minore allontanamento della distribuzione osservata dalla forma simmetrica. La curtosi ha a che fare con il grado di allontanamento della distribuzione osservata da uno specifico modello teorico, la distribuzione normale o di Gauss.

Asimmetria e curtosi vanno prima definite e poi misurate con appropriati indici. Ebbene, saranno questi gli argomenti delle pagine che seguono.

6.2 Asimmetria

Si inizia con la definizione di distribuzione di frequenza simmetrica.

DEFINIZIONE 6.1

Sia data una distribuzione di frequenza di una variabile discreta:

Modalità	x_1	x_2	...	x_k	Totale
Frequenza	n_1	n_2	...	n_k	n

Sia m la sua mediana. La distribuzione è detta simmetrica se le modalità di ciascuna delle coppie x_1 e x_k, x_2 e x_{k-1}, x_3 e x_{k-2} ecc. sono equidistanti da m e hanno la stessa frequenza.[1]

[1] In simboli, indicata con δ una quantità che è 0 se k è pari e 1 se k è dispari, si dice che una distribuzione è simmetrica se valgono simultaneamente le seguenti relazioni di identità:

$$(x_i - m) = -(x_{k+1-i} - m),\ n_i = n_{k+1-i},\ i = 1, 2, \ldots, (k - \delta)/2.$$

ESEMPIO 6.1

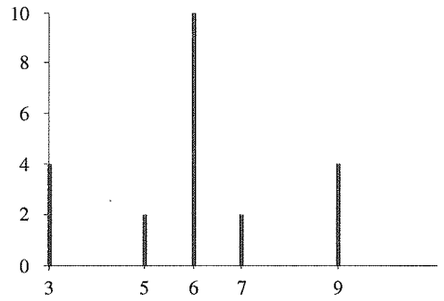
La distribuzione presentata nella tabella che segue è simmetrica:

Tabella 6.1
Distribuzione di un campione di aziende per numero di addetti.

Modalità	3	5	6	7	9	Totale
Frequenza	4	2	10	2	4	22

La mediana è $m = 6$, le modalità 3 e 9 (prima e ultima) hanno la stessa distanza dalla mediana e frequenze uguali; ciò accade anche per le modalità 5 e 7 (seconda e penultima). La distribuzione è rappresentata nella Figura 6.1.

Figura 6.1
Grafico della distribuzione di cui alla Tabella 6.1.

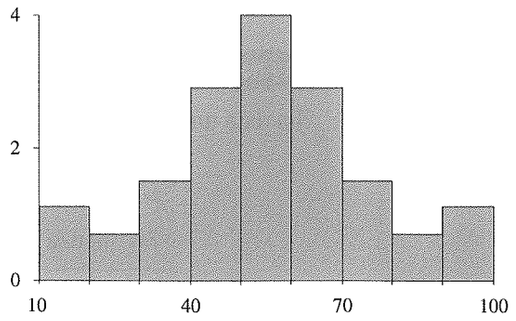


■

Con linguaggio geometrico, si può dire che una distribuzione è simmetrica quando l'asse di simmetria suddivide il suo grafico in due parti tali che l'una è l'immagine speculare dell'altra.

Per le distribuzioni di frequenza in classi, la precedente definizione di simmetria va adattata, sostituendo il termine "modalità" con quello di "classe" e il termine "frequenza" con l'espressione "densità di frequenza". Si veda la Figura 6.2.

Figura 6.2
Distribuzione di frequenza in classi simmetrica.



PROPOSIZIONE 6.1

Una distribuzione di frequenza simmetrica soddisfa le proprietà di seguito indicate:

(a) la media aritmetica coincide con la mediana:

$$\mu = m;$$

(b) la somma degli scarti dalla media aritmetica elevati alla terza potenza è uguale a 0:

$$\sum_{i=1}^{k}(x_i - \mu)^3 n_i = 0;$$

(c) il primo e il terzo quartile hanno la stessa distanza dalla mediana:

$$(m - q_1) = (q_3 - m),$$

da cui

$$(m - q_1) - (q_3 - m) = 0.$$

La dimostrazione della Proposizione 6.1 è contenuta nel Paragrafo A.4 dell'Appendice A.

■

ESEMPIO 6.2

Si verifichi che la distribuzione di cui all'Esempio 6.1 gode delle proprietà indicate nella Proposizione 6.1.

Soluzione

La proprietà (a) è soddisfatta, in quanto media e mediana sono entrambe uguali a 6. Per la verifica della proprietà (b), essendo $\mu = 6$, si ha

$$(3 - 6)^3 \cdot 4 + (5 - 6)^3 \cdot 2 + (6 - 6)^3 \cdot 10 + (7 - 6)^3 \cdot 2 + (9 - 6)^3 \cdot 4$$

$$= -108 - 2 + 0 + 2 + 108 = 0.$$

Quanto alla proprietà (c), si osservi che il primo e terzo quartile sono $q_1 = 5$ e $q_3 = 7$, da cui

$$(m - q_1) = (q_3 - m) = 1.$$

■

Le distribuzioni di frequenza non simmetriche vengono qualificate come asimmetriche. Ai fini dell'introduzione degli indici di asimmetria, è necessario distinguere tra asimmetria positiva e asimmetria negativa.

Un criterio qualitativo per distinguere l'asimmetria positiva da quella negativa si richiama alla rappresentazione grafica della distribuzione. Preso come riferimento il punto dell'asse delle ascisse corrispondente alla mediana, si ha asimmetria positiva se il grafico tende a "prolungarsi" più a destra di tale punto che non a sinistra; si ha asimmetria negativa nel caso opposto.

Nella Figura 6.3 sono riportati i grafici di nove distribuzioni di frequenza in classi che presentano strutture molto diversificate: è utile l'esercizio di desumere il tipo di asimmetria che caratterizza la singola distribuzione dalla forma del rispettivo grafico.

In base al criterio sopra indicato, sono simmetriche le distribuzioni (a), (d) ed (e); presentano asimmetria positiva le distribuzioni (b), (h) e (i); presentano asimmetria negativa le distribuzioni (c), (f) e (g).

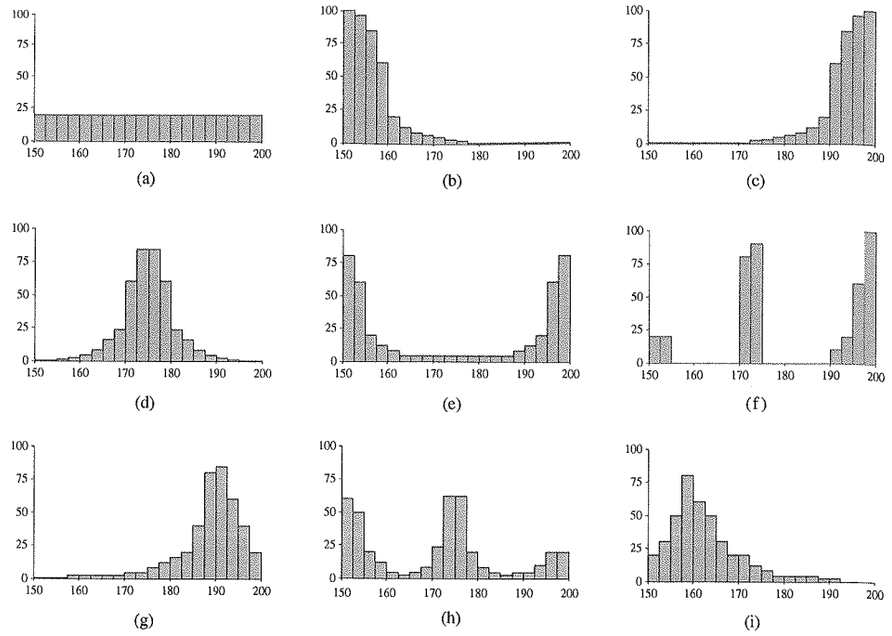


Figura 6.3
Distribuzioni di frequenza in classi con diversa asimmetria.

DEFINIZIONE 6.2

Una prima misura di asimmetria è data dall'indice

$$\alpha_1 = \frac{\mu - m}{\sigma}. \tag{6.1}$$

Esso è nullo nel caso di simmetria; tende ad assumere valori positivi o negativi a seconda che la distribuzione presenti asimmetria positiva o negativa.

L'indice (6.1) si annulla nel caso di distribuzioni simmetriche, per la proprietà (a) della Proposizione 6.1. Il segno dell'indice tende a essere positivo quando la distribuzione presenta asimmetria positiva, perché in questo caso, come è facile intuire, la media aritmetica tende a superare la mediana. Viceversa, l'indice tende ad assumere valori negativi nel caso di asimmetria negativa. Si può dimostrare che vale la disuguaglianza

$$|\mu - m| \leq \sigma,$$

da cui segue che α_1 è un indice normalizzato, assumendo valori nell'intervallo $[-1, 1]$. Inoltre, poiché numeratore e denominatore della (6.1) sono espressi nella stessa unità di misura, α_1 è un numero puro, cioè il suo valore è indipendente dall'unità di misura in cui è espressa la variabile.

ESEMPIO 6.3

Nella distribuzione di cui all'Esempio 5.3, dove $\mu = 2.55$, $m = 3$ e $\sigma = 1.11$, l'indice (6.1) assume il valore

$$\alpha_1 = \frac{2.55 - 3}{1.11} = -0.41,$$

quantità che segnala una non trascurabile asimmetria negativa, in accordo con la forma della distribuzione rappresentata nella Figura 3.2. ■

DEFINIZIONE 6.3

Una seconda misura di asimmetria è data dall'indice

$$\alpha_2 = \frac{1}{\sigma^3}\left[\frac{1}{n}\sum_{i=1}^{k}(x_i - \mu)^3 n_i\right]. \tag{6.2}$$

Esso è nullo nel caso di simmetria; tende ad assumere valori positivi se la distribuzione presenta asimmetria positiva e valori negativi nel caso opposto.

L'indice (6.2) si annulla nel caso di simmetria in forza della proprietà (b) della Proposizione 6.3. Tende a essere positivo nelle distribuzioni positivamente asimmetriche, perché in queste gli scarti positivi molto grandi presentano frequenze più elevate che non nel caso di simmetria. Tende a essere negativo nelle distribuzioni negativamente asimmetriche per le ragioni opposte. È un numero puro molto sensibile ai valori estremi, in quanto basato sugli scarti standardizzati al cubo.

ESEMPIO 6.4

Si calcoli l'indice α_2 per la distribuzione di frequenza considerata nell'Esempio 5.3.

Soluzione

Si ricorda che la media della distribuzione è 2.55. I calcoli sono esposti nella tabella che segue:

N. di esami (x_i)	Frequenza (n_i)	$(x_i - \mu)^2 \cdot n_i$	$(x_i - \mu)^3 \cdot n_i$
0	14	91.04	–232.14
1	41	98.50	–152.68
2	83	25.11	–13.81
3	116	23.49	10.57
4	56	117.74	170.72
5	5	30.01	73.53
Totale	315	385.89	–143.81

Tabella 6.2
Calcolo dell'indice di asimmetria α_1 per la distribuzione di frequenza dell'Esempio 5.3.

L'indice α_2 è uguale a

$$\alpha_2 = \frac{-143.81/315}{(\sqrt{385.89/315})^3} = -0.34.$$

La distribuzione è dunque caratterizzata da asimmetria negativa, come si evince del resto dall'esame della Figura 3.2. ■

ESEMPIO 6.5

Si calcoli α_2 per la distribuzione di frequenza esaminata nell'Esempio 2.4.

Soluzione

Si tenga presente che la media aritmetica della distribuzione è 47.73. I calcoli dell'indice sono illustrati nella tabella che segue:

Tabella 6.3
Calcolo dell'indice di asimmetria α_2 nella distribuzione di frequenza di cui all'Esempio 2.4.

Lunghezza avambraccio (cm)	Frequenza n_i	Valore centrale $\bar{x}_i$	$(\bar{x}_i - \mu)^2 \cdot n_i$	$(\bar{x}_i - \mu)^3 \cdot n_i$
41-42	5	41.5	194.06	–1209.02
43-44	17	43.5	304.18	–1286.68
45-46	25	45.5	124.32	–277.24
47-48	35	47.5	1.85	–0.43
49-50	32	49.5	100.25	177.45
51-52	20	51.5	284.26	1071.65
53-54	6	53.5	199.76	1152.60
Totale	140	–	1208.68	–371.67

L'indice di asimmetria è uguale a

$$\alpha_2 = \frac{-371.67/140}{(\sqrt{1208.68/140})^3} = -0.10.$$

La distribuzione presenta dunque una lieve asimmetria negativa, cosa che trova conferma dall'esame della Figura 3.8.

■

DEFINIZIONE 6.4

Una terza misura di asimmetria è data da

$$\alpha_3 = \frac{(q_3 - m) - (m - q_1)}{(q_3 - m) + (m - q_1)} = \frac{q_1 + q_3 - 2m}{q_3 - q_1}, \qquad (6.3)$$

dove m, q_1 e q_3 sono la mediana, il primo e il terzo quartile. L'indice α_3 è nullo nel caso di simmetria; tende ad assumere valori positivi se la distribuzione presenta asimmetria positiva e valori negativi nel caso opposto.

La (6.3) assume valore 0 nelle distribuzioni simmetriche in virtù della proprietà (c) della Proposizione 6.1. Nel caso di asimmetria positiva, poiché il grafico "si prolunga" maggiormente nella parte a destra della mediana, la distanza tra terzo quartile e mediana tende a essere superiore alla distanza tra primo quartile e mediana; ne segue che l'indice tende ad assumere valori positivi. Il fenomeno opposto si registra nel caso di asimmetria negativa. Il campo di variazione della (6.3) è l'intervallo $[-1, 1]$: il primo valore è assunto quando $m = q_3$, il secondo quando $m = q_1$. Come α_1 e α_2, α_3 è un numero puro. L'indice non è influenzato dai valori estremi, essendo basato sui quartili.

ESEMPIO 6.6

Si calcoli l'indice α_3 per la distribuzione esaminata nell'Esempio 4.19.

Soluzione

I quartili della distribuzione sono stati già calcolati nell'esempio 4.19:

$$q_1 = 3071.66, \quad m = 3372.84, \quad q_3 = 3682.45.$$

Si ha allora

$$\alpha_3 = \frac{(3682.45 - 3372.84) - (3372.84 - 3071.66)}{(3682.45 - 3372.84) + (3372.84 - 3071.66)} = 0.014.$$

L'indice segnala la quasi simmetria della distribuzione.

■

Nell'esempio precedente, gli indici di asimmetria sono stati introdotti per le distribuzioni di frequenza. Chiaramente, possono essere definiti anche per le serie di osservazioni, nel qual caso, le formule di calcolo vanno modificate. Per esempio, l'indice α_2 assume la forma

$$\alpha_2 = \frac{1}{\sigma^3}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^3\right].$$

6.3 Curtosi

Come accennato nell'introduzione, per curtosi s'intende il grado di "diversità" della distribuzione di frequenza osservata rispetto al modello teorico rappresentato dalla distribuzione normale, un modello matematico che può essere impiegato per approssimare una distribuzione reale nel senso di seguito indicato.

Si consideri la distribuzione in classi rappresentata dall'istogramma della Figura 6.4.

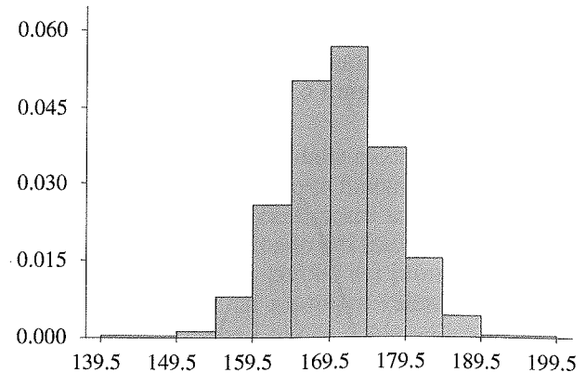


Figura 6.4
Istogramma di una ipotetica distribuzione di frequenza.

Il grafico è stato costruito ponendo sull'asse delle ordinate le densità di frequenza relativa f_i/d_i in modo tale che la superficie complessiva dell'istogramma sia uguale a 1 (si veda il Paragrafo 3.2.2).

Si voglia ora approssimare l'istogramma di frequenza con la curva descritta dall'equazione

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2},$$

detta funzione di densità normale o curva di Gauss, dove la variabile x può assumere tutti i valori dell'asse reale, mentre μ e σ sono due costanti numeriche che sono interpretabili come la media aritmetica e la deviazione standard della distribuzione teorica. La funzione presenta la tipica forma campanulare mostrata nella Figura 6.5.

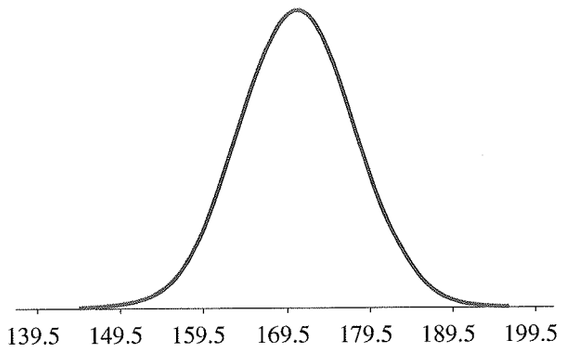


Figura 6.5
Curva normale.

La media μ coincide con il punto di massima densità, mentre σ determina il grado di concentrazione della curva attorno alla media: al diminuire di σ, la curva diventa più concentrata intorno a μ. Si può dimostrare inoltre che l'area sottesa alla curva è pari a 1. La curva normale è un modello appropriato per descrivere un'ampia gamma di fenomeni naturali e anche molti processi industriali.

Si pensi, ad esempio, all'istogramma di frequenza della distribuzione in classi del peso dei pezzi prodotti in un processo industriale. Se il processo è in controllo (cioè se opera in modo stabile e prevedibile), tale istogramma potrà essere approssimato con la curva normale.

Il modo più ragionevole per approssimare la distribuzione osservata mediante la curva gaussiana consiste nell'utilizzare la media e la deviazione standard osservate come media e deviazione standard della distribuzione normale. Con questa procedura, l'istogramma in Figura 6.4 è approssimato dalla curva gaussiana mostrata nella Figura 6.6.

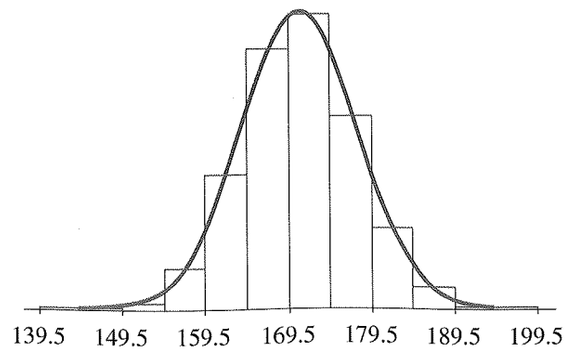


Figura 6.6
Approssimazione di una distribuzione osservata con la normale.

Si distingue tra ipernormalità e iponormalità. La distribuzione ipernormale è caratterizzata dalla eccedenza (ordinate più elevate) rispetto alla normale attorno al valore medio e alle code, mentre la distribuzione iponormale è contraddistinta dal fenomeno opposto. I due tipi di deviazione dalla normalità sono illustrati nella Figura 6.7.

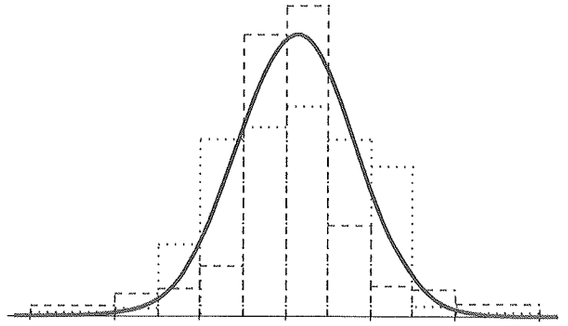


Figura 6.7
Normale (—), Ipernormale (- - -), Iponormale (· · ·).

DEFINIZIONE 6.5

Una misura di curtosi è data da

$$\gamma = \frac{1}{\sigma^4}\left[\frac{1}{n}\sum_{i=1}^{k}(x_i-\mu)^4 n_i\right] - 3. \qquad (6.4)$$

L'indice assume valori vicini allo 0 se la distribuzione è molto prossima alla normale;[2] è positivo nel caso di distribuzioni ipernormali; è negativo nel caso di distribuzioni iponormali.

Il fatto che l'indice sia positivo o negativo a seconda che la distribuzione osservata sia ipernormale o iponormale può essere spiegato come segue. In una distribuzione di frequenza vicina alla normale, $\gamma \approx 0$, da cui

$$\frac{1}{n}\sum_{i=1}^{k}(x_i-\mu)^4 n_i \approx 3\sigma^4 = 3\left[\frac{1}{n}\sum_{i=1}^{k}(x_i-\mu)^2 n_i\right]^2.$$

Si concentri ora l'attenzione sugli scarti dalla media molto grandi in valore assoluto, quelli associati alle modalità di X molto lontane (in una direzione o nell'altra) da μ. In distribuzioni ipernormali, questi scarti sono più frequenti che non nella distribuzione normale e hanno un impatto più forte sul primo che non sul secondo membro della relazione precedente, perché nel primo membro sono elevati alla quarta. Di conseguenza, γ tende ad assumere valori maggiori di 0. Con analogo ragionamento, si intuisce come l'indice tenda ad assumere valori negativi in distribuzioni iponormali.

[2] La (6.4) è vicina allo 0 nel caso di quasi normalità della distribuzione. Si può dimostrare infatti che l'indice, opportunamente riformulato per una distribuzione teorica, è nullo quando la distribuzione è normale.

ESEMPIO 6.7

Si calcoli l'indice di curtosi per la distribuzione esaminata nell'Esempio 6.5.

Soluzione

I calcoli sono mostrati nella tabella che segue:

Tabella 6.4
Calcolo dell'indice di curtosi per la distribuzione di frequenza in classi dell'Esempio 6.5.

Lunghezza avambraccio (cm)	Frequenza n_i	Valore centrale $\bar{x}_i$	$(\bar{x}_i - \mu)^2 \cdot n_i$	$(\bar{x}_i - \mu)^4 \cdot n_i$
41-42	5	41.5	194.06	7 532.21
43-44	17	43.5	304.18	5 442.65
45-46	25	45.5	124.32	618.24
47-48	35	47.5	1.85	0.10
49-50	32	49.5	100.25	314.08
51-52	20	51.5	284.26	4 040.13
53-54	6	53.5	199.76	6 650.50
Totale	140	–	1208.68	24 597.91

L'indice di curtosi assume il valore

$$\gamma = \frac{24\,597.91/140}{(\sqrt{1208.68/140})^4} - 3 = -0.64.$$

La distribuzione è dunque iponormale. ■

Riassunto

Accanto alle medie e agli indici di variabilità, gli indici di forma concorrono a descrivere sinteticamente le caratteristiche di insiemi di dati quantitativi. Specificatamente, essi misurano l'allontanamento della distribuzione di frequenza osservata dalla forma simmetrica (asimmetria) o dal modello teorico rappresentato dalla curva di Gauss (curtosi). In questo capitolo, vengono illustrati tre indici di asimmetria e un indice di curtosi.

Parole chiave

asimmetria negativa, 159
asimmetria positiva, 159
curtosi, 163
curva di Gauss, 164
distribuzioni di frequenze asimmetriche, 159
distribuzioni di frequenze simmetriche, 157
funzione di densità normale, 164
ipernormalità, 165
iponormalità, 165
simmetria, 157

Commenti e curiosità

L'azzeramento degli indici di forma non ha significato univoco

Gli indici di forma introdotti in questo capitolo si annullano anche quando la distribuzione non rispecchia perfettamente il modello di riferimento (simmetria per gli indici α_1, α_2 e α_3, normalità per l'indice γ).

Nella Tabella 6.5 sono date due distribuzioni con le stesse classi e frequenze diverse. Nella prima, rappresentata nella Figura 6.8(a), l'indice α_2 è nullo; nella seconda, rappresentata nella Figura 6.8(b), l'indice γ è nullo.

Tabella 6.5
Distribuzioni di frequenza in cui sono nulli l'indice di asimmetria (a) e l'indice di curtosi (b).

Classe	Frequenza (a)	Frequenza (b)
0-10	1 264	446
10-20	1 038	328
20-30	1 221	524
30-40	937	947
40-50	687	2 736
50-60	800	1 498
60-70	680	1 891
70-80	1 996	669
80-90	679	476
90-100	698	485
Totale	10 000	10 000

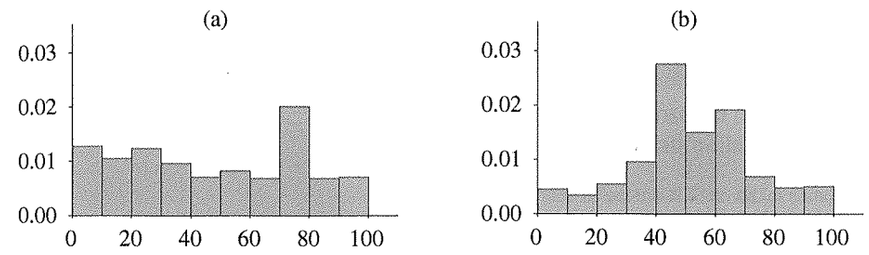


Figura 6.8
Distribuzione non simmetrica (a) e distribuzione non normale (b) in cui α_2 e γ sono nulli.

Esercizi

6.1 Si consideri la serie di osservazioni relativa alla circonferenza toracica degli individui in età inferiore a 25 anni nella matrice dei dati di cui all'Esercizio 1.6. Si determini l'indice α_2.

6.2 Si consideri la distribuzione delle famiglie italiane per numero di componenti data nell'Esercizio 2.4. Supponendo che il numero medio di membri delle famiglie con 6 o più membri sia 6.2, si calcolino α_2 e α_3.

6.3 Si consideri la distribuzione di frequenza dell'Esercizio 5.11.

a) Si determinino gli indici di asimmetria α_1 e α_2.

b) Si calcoli l'indice di curtosi.

6.4 Si riprenda la distribuzione dei comuni secondo l'ampiezza demografica (Esercizio 4.8). Si determinino l'indice di asimmetria α_2 e l'indice di curtosi, sfruttando l'informazione sull'ampiezza demografica complessiva dei comuni appartenenti alla singola classe.

6.5 Si rappresenti con un istogramma di frequenza la seguente distribuzione relativa ai tempi di attesa tra due eruzioni successive di geyser (Hand *et al.*, 1994):

Tempo	Frequenza	Tempo	Frequenza
42.5-47.5	4	77.5-82.5	61
47.5-52.5	34	82.5-87.5	43
52.5-57.5	30	87.5-92.5	21
57.5-62.5	22	92.5-97.5	11
62.5-67.5	11	97.5-102.5	1
67.5-72.5	19	102.5-107.5	0
72.5-77.5	41	107.5-112.5	1

Sulla base del grafico, si esprima un giudizio sulla eventuale asimmetria della distribuzione e se ne verifichi la validità calcolando l'indice α_3.

CAPITOLO

7

Uno sguardo d'insieme alle costanti caratteristiche

Sommario del capitolo

7.1 Introduzione
7.2 Grafici e costanti caratteristiche
7.3 Disuguaglianza di Chebyshev
Riassunto
Commenti e curiosità
Esercizi

7.1 Introduzione

Le costanti caratteristiche introdotte nei tre capitoli precedenti – medie, indici di variabilità e indici di forma – servono a descrivere aspetti particolari dei dati: le medie sono utili perché individuano il centro attorno al quale i dati tendono a collocarsi; gli indici di variabilità danno l'informazione complementare sul grado di dispersione dei dati rispetto al centro; gli indici di forma misurano la maggiore o minore deviazione della struttura dei dati da modelli di riferimento quali la simmetria e la normalità.

ESEMPIO 7.1

Si confrontino tramite le costanti caratteristiche riportate in Tabella 7.2 i due insiemi di dati presentati nella Tabella 7.1.

Tabella 7.1 Voti riportati all'esame di Stato da un campione di studenti del corso di laurea in Economia distinti per scuola superiore frequentata.

Scuola	Voto riportato all'esame di Stato
Liceo scientifico	68, 100, 100, 93, 78, 77, 77, 80, 69, 66, 98, 84, 71, 93, 85, 71, 100, 70, 60, 70, 79, 78, 83, 78, 86, 69.
Istituto tecnico commerciale	100, 88, 90, 74, 98, 85, 100, 100, 65, 100, 65, 100, 90, 87, 68, 79, 68, 92, 90, 100, 80.

Tabella 7.2 Statistiche di sintesi dei due insiemi di dati riportati nella Tabella 7.1.

Statistiche descrittive	Scuola	
	Liceo scientifico	Tecnico commerciale
Media aritmetica	80.12	86.62
Primo quartile	70.00	79.00
Mediana	78.00	90.00
Terzo quartile	86.00	100.00
Deviazione standard	11.30	11.92
Coefficiente di variazione	14.10	13.76
Indice di asimmetria α_1	0.19	−0.28
Indice di asimmetria α_2	0.39	−0.55

Dal confronto delle statistiche di sintesi, si deduce che le medie del secondo insieme di dati sono tutte superiori a quelle del primo. Più precisamente, si può osservare una forte differenza tra le mediane (90 contro 78) e tra i terzi quartili (100 contro 86), segno che nel secondo insieme sono molto più frequenti le votazioni elevate.

Per contro, presentano una variabilità pressoché uguale, come indica il coefficiente di variazione, misura di variabilità qui utilizzata perché le medie dei due gruppi sono sensibilmente diverse.

Infine, gli indici di forma α_1 e α_2 sono entrambi positivi per il primo insieme ed entrambi negativi per il secondo, mostrando in tutti e due i casi una leggera asimmetria.

■

7.2 Grafici e costanti caratteristiche

Una rappresentazione grafica – come un diagramma ad aste o un istogramma – riassume visivamente tutte le caratteristiche dell'insieme di dati considerato. Dall'esame del grafico è possibile "predire" la media e la deviazione standard, oltre che stabilire il segno e il grado di asimmetria. Per esempio, dal grafico (b) della Figura 7.1, si desume che: la media aritmetica è largamente inferiore a 175 (valore centrale della distribuzione), perché la frequenza delle osservazioni al di sotto di questa quantità è di gran lunga superiore a quella delle osservazioni che si trovano al di sopra; la deviazione standard non supera quota 10, perché la maggior parte delle osservazioni si trova nell'intervallo [150, 160]; la distribuzione presenta una forte asimmetria positiva.

L'esame congiunto dei grafici di due distribuzioni di frequenza permette di confrontare le stesse in termini di tendenza centrale, variabilità e forma. Ad esempio, osservando i grafici di Figura 7.1, è facile stabilire che la media aritmetica della distribuzione (b) è notevolmente inferiore a quella della distribuzione (a), e che la media di quest'ultima è molto più piccola di quella della distribuzione (c).

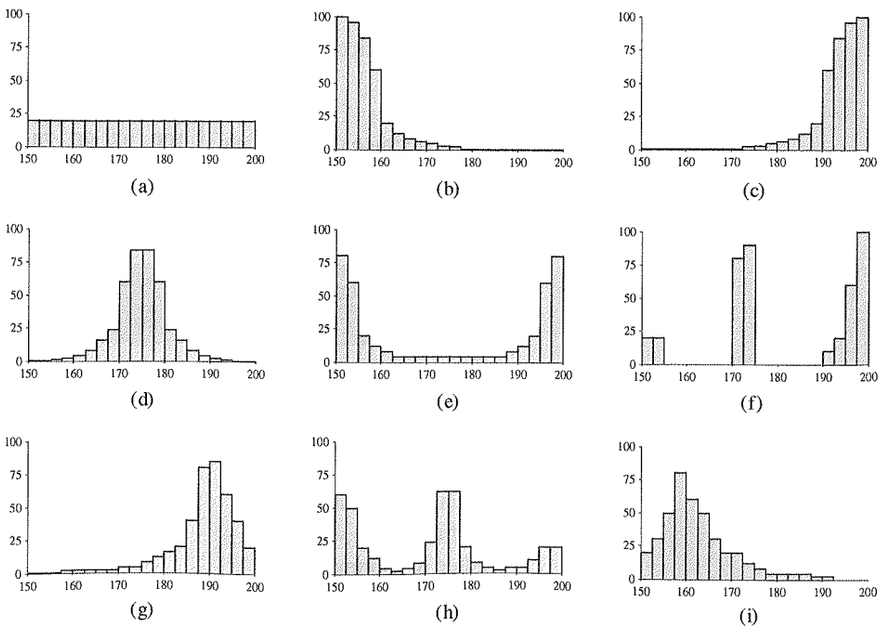


Figura 7.1
Grafici e costanti caratteristiche
μ, σ and α_2.
(a): 175, 14.4, 0;
(b): 156.4, 5.8, 2.6;
(c): 193.6, 5.8, −2.6;
(d): 175, 5.5, 0;
(e): 175, 20.6, 0;
(f): 182.2, 15.3, −0.4;
(g): 188.9, 7.1, −1.7;
(h): 170.2, 14.7, 0.3;
(i): 162.4, 7.6, 1.2.

Ciò perché i valori piccoli (per esempio, i valori compresi tra 150 e 160) hanno frequenza molto più alta in (b) che non in (a), mentre i valori elevati (quelli tra 190 e 200) si presentano più frequentemente in (c). Un criterio per predire la media aritmetica di una distribuzione dal relativo grafico consiste nell'osservare come le frequenze sono associate alle modalità della variabile: la media infatti tende a essere più vicina alle modalità con le frequenze più alte.

Tale esercizio per la deviazione standard porta alle seguenti conclusioni: le distribuzioni (b) e (c) presentano una deviazione standard decisamente minore rispetto a quella di (a); il motivo di ciò è che nei primi due casi la maggior parte dei termini è raggruppata in un intervallo ristretto (tra 150 e 165 in (b) e tra 185 e 200 in (c)). Con analogo ragionamento, si deduce che la deviazione standard della distribuzione (c) è minore di quella della distribuzione (e), perché nella prima la maggior parte delle modalità è concentrata nell'intervallo di estremi 180 e 200, entro cui è compresa anche la media, mentre nella seconda i termini sono molto più dispersi attorno alla media, che è uguale a 175.

7.2.1 Grafico a scatola

Il grafico a scatola (*boxplot*) è una rappresentazione grafica che descrive l'insieme dei dati (una serie di osservazioni o una distribuzione di frequenza) con i cosiddetti cinque valori di sintesi rappresentati da: minimo, massimo, primo quartile q_1, mediana m e terzo quartile q_3.

Questo grafico può essere disposto orizzontalmente, cioè con la scala di misura posta in orizzontale, o verticalmente. Nella descrizione che segue, si adotta il secondo criterio.

Il grafico è composto da: un rettangolo di larghezza arbitraria detto scatola, i cui lati inferiore e superiore corrispondono ai punti (sulla scala di misura) del primo e terzo quartile; un segmento di retta orizzontale all'interno del rettangolo corrispondente al punto della mediana; un segmento di retta perpendicolare che collega il punto centrale del lato superiore della scatola al punto del massimo; un segmento di retta perpendicolare che collega il punto centrale del lato inferiore della scatola al punto del minimo.

Tutto ciò è illustrato nella Figura 7.2.

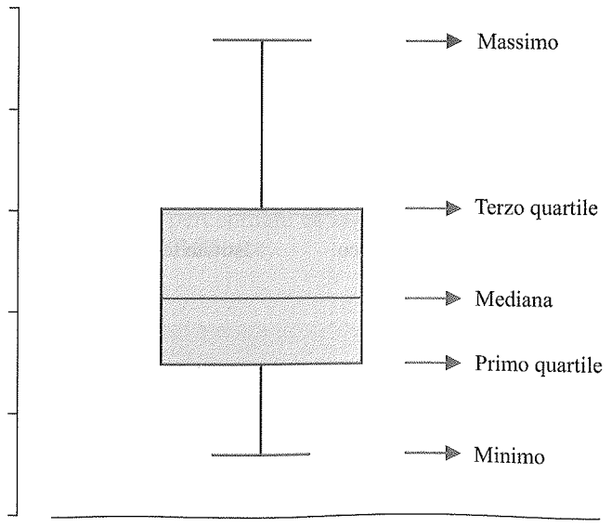


Figura 7.2
Grafico a scatola.

La lunghezza del segmento di retta che va dal minimo al massimo è il campo di variazione, mentre l'altezza della scatola è la differenza interquartile. La posizione della linea della mediana all'interno della scatola consente di stabilire se i dati presentano simmetria o asimmetria e, nel secondo caso, se si tratta di asimmetria positiva o negativa. Precisamente: se la linea della mediana è equidistante dalle linee del primo e terzo quartile, si ha simmetria; se essa è più vicina alla linea del primo quartile, si ha asimmetria positiva; se essa è più vicina alla linea del terzo quartile, si ha asimmetria negativa. Il grado di asimmetria è tanto maggiore quanto più la linea della mediana si discosta dalla posizione di equidistanza. Quanto detto sulla simmetria/asimmetria si basa sulle proprietà dell'indice α_3 (si veda la Definizione 6.4).

In sintesi, il grafico a scatola fornisce le seguenti indicazioni: una misura di tendenza centrale, con la linea della mediana; due misure di variabilità, il campo di variazione e la differenza interquartile; il segno e il grado dell'asimmetria.

ESEMPIO 7.2

Nella Figura 7.3 sono rappresentati, con diagrammi a scatola, i due insiemi di dati presentati nell'Esempio 7.1.

Il grafico permette di confrontare rapidamente le principali caratteristiche dei due insiemi di dati. Si percepisce chiaramente come la seconda scatola sia fortemente spostata verso l'alto, il che significa che tutti e tre i quartili sono molto più grandi di quelli del primo insieme di dati. Per quanto riguarda la variabilità, si vede che la differenza interquartile è più piccola nel primo gruppo che non nel secondo, mentre il campo di variazione è un po' più grande nel primo. Infine, la posizione della linea della mediana mostra una leggera asimmetria positiva nel primo gruppo e una leggera asimmetria negativa nel secondo, come già notato nell'Esempio 7.1.

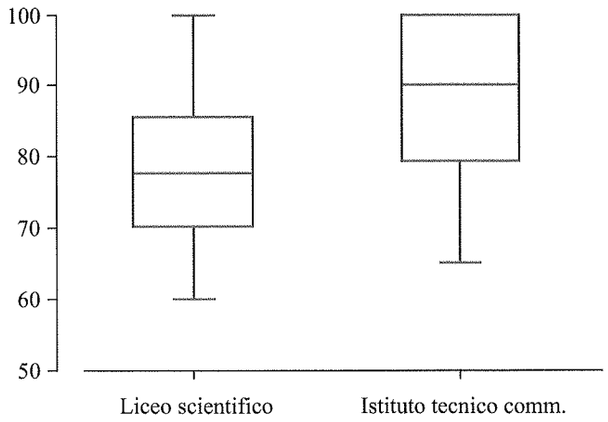


Figura 7.3
Grafici a scatola degli insiemi di dati contenuti nella Tabella 7.1.

■

7.3 Disuguaglianza di Chebyshev

Come si è già detto, gli indici di variabilità integrano l'informazione fornita dalle medie. È ora interessante vedere come l'uso congiunto della media aritmetica e della deviazione standard porga un'informazione aggiuntiva.

PROPOSIZIONE 7.1

Sia $x_1, x_2, \ldots, x_n$ una serie di n osservazioni con media μ e deviazione standard σ. Per qualsiasi numero $k > 0$, si considerino tutti quei termini della serie appartenenti all'intervallo $(\mu - k\sigma, \mu + k\sigma)$. Questi elementi sono identificati dall'insieme di indici

$$I_k = \{j, 1 \leq j \leq n : |x_j - \mu| < k\sigma\}.$$

Sia $n(I_k)$ il numero di elementi in I_k. Allora, per qualsiasi $k \geq 1$, vale la disuguaglianza

$$\frac{n(I_k)}{n} \geq 1 - \frac{1}{k^2}. \tag{7.1}$$

Per la dimostrazione si rinvia al Paragrafo A.5 dell'Appendice A. ■

Questo risultato, noto come disuguaglianza di Chebyshev, è molto generale e vale per qualsiasi insieme di dati, di cui si conosce solo la media e la deviazione standard. Intuitivamente può essere letta nel modo seguente: tra le n osservazioni $x_1, x_2, \ldots, x_n$ la frequenza relativa di quelle che cadono nell'intervallo compreso tra $\mu - k\sigma$ e $\mu + k\sigma$ è non inferiore a $1 - 1/k^2$.

ESEMPIO 7.3

Si consideri un ipotetico insieme di dati avente media 100 e varianza 16. Si applichi disuguaglianza di Chebyshev con $k = 2, 2.5, 3, 5$.

Soluzione

I risultati sono riportati nella Tabella 7.3:

k	Intervallo $\mu - k\sigma, \mu + k\sigma$	Frequenza rel. $1 - 1/k^2$
2.0	(92, 108)	0.75
2.5	(90, 110)	0.84
3.0	(88, 112)	0.89
5.0	(80, 120)	0.96

Tabella 7.1
Applicazione della disuguaglianza di Chebyshev a un insieme generico di dati avente media 100 e varianza 16.

Ciò significa che: la percentuale di osservazioni comprese nell'intervallo (92, 108) è non inferiore al 75%; la percentuale di osservazioni comprese nell'intervallo (90, 110) non è inferiore all'84% ecc. ■

ESEMPIO 7.4

Il peso medio e la deviazione standard dei semilavorati appartenenti a un determinato lotto sono 70 e 2 gr. Tali semilavorati, per essere utilmente impiegati nel processo produttivo, devono avere un peso compreso nell'intervallo (64, 76). Ci si chiede quale sia la percentuale massima di scarti (cioè di semilavorati non utilizzabili).

Soluzione

Dato che

$$k\sigma = \frac{76 - 64}{2} = 6,$$

$k = 6/2 = 3$, cosicché la frequenza relativa dei semilavorati con peso interno all'intervallo (64, 76) è maggiore o uguale a $1 - 1/3^2 = 0.89$. Ne segue che la percentuale massima di scarti è pari all'11%. ■

La disuguaglianza di Chebyshev dà informazioni valide solo se $k > 1$: ciò assicura che il secondo membro della (7.1) assuma un valore positivo. Ovviamente, la disuguaglianza non presenta alcuna utilità quando la distribuzione è nota, nel qual caso, la percentuale dei termini compresi nell'intervallo $(\mu - k\sigma,\ \mu + k\sigma)$ può essere calcolata direttamente, senza approssimazioni.

Il limite inferiore per la frequenza relativa dei casi appartenenti all'intervallo $(\mu - k\sigma, \mu + k\sigma)$ vale universalmente, cioè per tutti gli insiemi di dati aventi la stessa media e la stessa deviazione standard: può accadere allora che, per uno specifico insieme di dati, questo limite sottostimi grossolanamente la frequenza relativa effettiva.

Riassunto

Le medie, gli indici di variabilità e gli indici di forma sono indicatori di sintesi che mettono in luce le caratteristiche dei dati. Uno degli obiettivi di questo breve capitolo è quello di illustrare come tali caratteristiche possano essere "lette" dalle rappresentazioni grafiche.

Una particolare attenzione è dedicata al cosiddetto grafico a scatola che visualizza i cinque valori riassuntivi di un insieme di dati: il minimo, il massimo e i quartili.

Viene introdotta la disuguaglianza di Chebyshev, che, relativamente a un insieme di dati qualsiasi di cui si conoscono solo media μ e deviazione standard σ, fornisce un'approssimazione della frequenza relativa dei casi che cadono nell'intervallo $(\mu - k\sigma,\ \mu + k\sigma)$, con $k > 1$.

Parole chiave

cinque valori di sintesi, 171
disuguaglianza di Chebyshev , 173
grafico a scatola, 171

Commenti e curiosità

Grafico a violino: un arricchimento del grafico a scatola

Come si è visto nel Paragrafo 7.2, il grafico a scatola è la rappresentazione dei cinque valori di sintesi, minimo, massimo e quartili, la cui lettura combinata fa emergere le proprietà principali dei dati (media, variabilità e forma). Il *grafico a violino* costituisce un'evoluzione del grafico a scatola: oltre a rappresentare i cinque valori di sintesi, fornisce un'approssimazione della distribuzione di frequenza dei dati.

Per comprendere come si costruisce, conviene esaminare la Figura 7.4, realizzata con R, dove sono rappresentati i dati relativi alla lunghezza dei sepali di tre specie di iris (si veda il *file* "Iris"). La parte centrale della singola figura riproduce in miniatura il grafico a scatola; ai due lati viene disegnata, simmetricamente, la curva continua che approssima l'istogramma di frequenza dei dati (istogramma lisciato), ruotato di 90 gradi in una direzione e nell'altra.

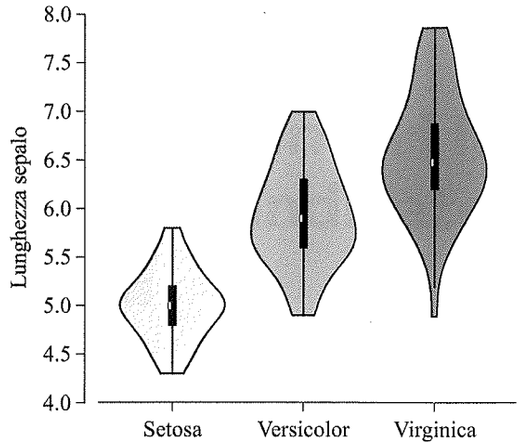


Figura 7.4
Grafici a violino.

Esercizi

7.1 Utilizzando le statistiche appropriate (medie, indici di variabilità e di forma), si descriva la distribuzione dell'Esercizio 2.4, assumendo che il numero medio di componenti della famiglia con 6 o più componenti sia 6.2.

7.2 Si consideri la distribuzione dell'Esercizio 3.2.
a) Se ne faccia una esaustiva descrizione tramite le appropriate costanti caratteristiche.
b) La si rappresenti mediante un grafico a scatola.

7.3 Si descriva con un grafico a scatola la distribuzione di cui all'Esercizio 3.3.

7.4 Si costruiscano l'istogramma e il grafico a scatola della distribuzione di cui all'Esercizio 6.5. Osservando i due grafici, si predicano la media aritmetica e la deviazione standard; si preveda il segno dell'indice α_1. Si confrontino poi le predizioni con i risultati degli appropriati calcoli.

7.5 Si costruisca una distribuzione di 100 unità raggruppate in 10 classi di ampiezza 5 con le seguenti caratteristiche:

$$\text{minimo} = 30,\ \text{massimo} = 80,\ q_1 = 40,\ m = 50,\ q_3 = 75.$$

7.6 Si consideri la serie di osservazioni della variabile altezza nel *file* "Dati antropometrici" (si veda l'Esercizio 1.6), avente media 179.60 e deviazione standard 6.45. Si applichi la disuguaglianza di Chebyshev per approssimare la frequenza relativa dei casi appartenenti agli intervalli:

$$179.61 - 1.5 \cdot 6.45,\ 179.61 + 1.5 \cdot 6.45;$$
$$179.61 - 2.0 \cdot 6.45,\ 179.61 + 2.0 \cdot 6.45;$$
$$179.61 - 2.5 \cdot 6.45,\ 179.61 + 2.5 \cdot 6.45.$$

Si confrontino i risultati con le frequenze effettive.

CAPITOLO

8

Sommario del capitolo

8.1 Introduzione
8.2 Serie storiche di numeri indici
8.3 Numeri indici sintetici
8.4 Indicatori compositi
Riassunto
Commenti e curiosità
Esercizi

Numeri indici e indicatori compositi

8.1 Introduzione

I numeri indici sono particolari rapporti statistici che servono per mettere a confronto i valori assunti da una variabile quantitativa in tempi o luoghi diversi. Sono numeri indici, ad esempio, i rapporti tra: il prezzo di un chilogrammo di pane in Italia in due tempi diversi, il prezzo di un litro di benzina a Parigi e a Roma a una certa data, il fatturato mensile di due aziende industriali di un dato settore merceologico. Si tratta di numeri puri, perché rapporti tra quantità espresse nella stessa unità di misura, molto utili per effettuare comparazioni in modo più efficiente di quanto non si possa fare con semplici differenze.

Particolarmente importanti sono le serie di numeri indici derivabili da serie storiche, che hanno la funzione di facilitare la comprensione di come le variabili di interesse si modificano nel tempo. In questo ambito, verranno definiti anche alcuni concetti basilari, come quelli di variazione relativa in un certo arco temporale e di variazione relativa per unità di tempo.

I numeri indici si possono impiegare anche per la descrizione delle serie territoriali. In questo caso, la comparazione avviene tra i valori che la variabile di interesse assume nelle articolazioni territoriali (ad esempio le regioni italiane) e un valore medio riferito all'intero territorio (Italia, nel caso delle regioni). Nelle pagine che seguono, si farà riferimento solo alle serie storiche.

Quando il confronto nel tempo riguarda non una singola variabile, ma un insieme di variabili omogenee, come ad esempio i prezzi di un insieme di beni e servizi, è necessario sintetizzare i numeri indici delle singole variabili con appropriate formule, denominate numeri indici sintetici, che esprimono la dinamica temporale media dell'insieme delle variabili. Le applicazioni principali in campo economico dei numeri indici sintetici sono la misura della variazione media dei prezzi e delle quantità in un dato arco temporale. In questo capitolo, l'attenzione sarà rivolta sul primo aspetto.

Un argomento che ha punti di contatto concettuali con i numeri indici sintetici è quello dei cosiddetti indicatori compositi. Si tratta di strumenti con cui si confrontano più unità statistiche, tipicamente unità territoriali (comuni, regioni, Stati ecc.) riguardo a un fenomeno socio-economico complesso, che viene quantificato tramite un'appropriata aggregazione di un insieme di indicatori elementari, che descrivono i singoli aspetti del fenomeno. L'esito di queste elaborazioni è quello di poter mettere in graduatoria le unità statistiche in base alle "misure" del fenomeno complesso loro assegnate.

8.2 Serie storiche di numeri indici

Si consideri la serie storica

$$a_1, a_2, \ldots, a_k,$$

dove $a_1, a_2, \ldots, a_k$ sono i valori della variabile considerata ai tempi $1, 2, \ldots, k$, come, ad esempio, il prodotto interno lordo di uno Stato in una successione di anni o i matrimoni celebrati nei diversi mesi di un dato anno.

Un problema preliminare alla costruzione dei numeri indici a partire da una serie storica è quello di individuare la quantità di riferimento, detta base, con cui confrontare le quantità $a_1, a_2, \ldots, a_k$. Se come base si assume il valore della variabile a un tempo prefissato, ad esempio al tempo 1, si parla di *numeri indici a base fissa*.

DEFINIZIONE 8.1

Data la serie storica $a_1, a_2, \ldots, a_k$, i numeri indici a base fissa si ottengono dividendo ciascun termine della serie storica per il termine del tempo assunto come base. Indicato con h il tempo base, la serie dei numeri indici con base h è data da:

$$I_{1|h} = \frac{a_1}{a_h}, \quad I_{2|h} = \frac{a_2}{a_h}, \ldots, I_{k|h} = \frac{a_k}{a_h}. \tag{8.1}$$

I numeri indici percentuali si deducono dalla (8.1) moltiplicando ogni elemento per 100.

ESEMPIO 8.1

Si calcolino i numeri indici percentuali per la serie storica che segue, assumendo come anno base il 2010:

Tabella 8.1
Arrivi nelle strutture ricettive dei paesi EU dal 2010 al 2016 (migliaia).

Anno	2010	2011	2012	2013	2014	2015	2016
Arrivi	227 297	243 458	248 578	262 443	272 539	288 283	296 790

Soluzione

Bisogna dividere ciascuno dei valori della seconda riga della Tabella 8.1 per 227 297 e moltiplicare il risultato per 100. La serie storica che ne risulta è data nella Tabella 8.2:

Tabella 8.2
Numeri indici percentuali con base 2010 per la serie storica riportata nella Tabella 8.1.

Anno	2010	2011	2012	2013	2014	2015	2016
Numero indice	100.0	107.1	109.4	115.5	119.9	126.8	130.6

Questi numeri mostrano la dinamica temporale del fenomeno in modo più efficiente di quanto possano fare i dati di partenza. Dall'esame dei numeri indici, si deduce che, fatto 100 il numero degli arrivi nel 2010, si registrano 107.1 arrivi nel 2011, con un incremento del 7.1%, 109.4 arrivi nel 2012, con un incremento del 9.4% sempre rispetto al 2010 e così via. ■

La Proposizione 8.1 indica il procedimento per il cambiamento di base.

PROPOSIZIONE 8.1

Data una serie di numeri indici con base h, è possibile passare a una nuova base h' dividendo ogni numero indice per $I_{h'|h}$, cioè per il vecchio numero indice della nuova base. Con riferimento al generico tempo t, si ha

$$I_{t|h'} = \frac{I_{t|h}}{I_{h'|h}}. \tag{8.2}$$

La (8.2) segue dalle identità:

$$I_{t|h'} = \frac{a_t}{a_{h'}} = \frac{a_t/a_h}{a_{h'}/a_h} = \frac{I_{t|h}}{I_{h'|h}}.$$

ESEMPIO 8.2

Si calcolino i numeri indici percentuali con base l'anno 2013 a partire dalla serie di numeri indici con base l'anno 2010 (Tabella 8.2).

Soluzione

Il numero indice del 2013 (nuovo anno base) con base 2010 (vecchio anno base) è 1.155. Pertanto, la nuova serie di numeri indici si ottiene come segue:

$$I_{2010|2013} = \frac{1.000}{1.155} = 0.886, \; I_{2011|2013} = \frac{1.071}{1.155} = 0.928, \ldots,$$

$$I_{2016|2013} = \frac{1.306}{1.155} = 1.131.$$

La serie completa è riportata nella Tabella 8.3:

Tabella 8.3
Numeri indici percentuali con base 2013 calcolati in base ai dati della Tabella 8.2.

Anno	2010	2011	2012	2013	2014	2015	2016
Numero indice	86.6	92.8	94.7	100.0	103.8	109.8	113.1

■

In alcune situazioni, è più appropriato o più conveniente mettere a confronto il singolo termine della serie storica non con il termine del tempo base, ma con il termine immediatamente precedente. Procedendo in questo modo, si ottengono i *numeri indici a base mobile*.

DEFINIZIONE 8.2

Data la serie storica $a_1, a_2, \ldots, a_k$, i numeri indici a base mobile si ottengono dividendo il generico termine a_t per il termine precedente a_{t-1}, ($t = 2, 3, \ldots, k$):

$$I_{2|1} = \frac{a_2}{a_1}, \; I_{3|2} = \frac{a_3}{a_2}, \ldots, I_{k|k-1} = \frac{a_k}{a_{k-1}}.$$

ESEMPIO 8.3

Si calcolino i numeri indici a base mobile percentuali per la serie storica riportata nella Tabella 8.1.

Soluzione

Ogni numero indice è calcolato come rapporto, moltiplicato per 100, tra il termine corrispondente all'anno considerato e quello dell'anno precedente. I numeri indici percentuali che ne risultano sono riportati nella tabella che segue:

Tabella 8.4
Numeri indici percentuali a base mobile calcolati in base ai dati della Tabella 8.1.

Anno	2010	2011	2012	2013	2014	2015	2016
N. indice a base mobile	–	107.1	102.1	105.6	103.8	105.8	103.0

■

I numeri indici a base mobile possono essere determinati a partire dai numeri indici a base fissa, come indicato nella Proposizione 8.2.

PROPOSIZIONE 8.2

Data la serie di numeri indici a base fissa $I_{1|h}, I_{2|h}, \ldots, I_{k|h}$, si possono calcolare i numeri indici a base mobile, $I_{2|1}, I_{3|2}, \ldots, I_{k|k-1}$, dividendo ogni numero indice a base fissa per il precedente. Così, per il generico tempo t ($t = 2, 3, \ldots, k$), si ha

$$I_{t|t-1} = \frac{I_{t|h}}{I_{t-1|h}}. \tag{8.3}$$

La (8.3) segue dalle identità:

$$I_{t|t-1} = \frac{a_t}{a_{t-1}} = \frac{a_t/a_h}{a_{t-1}/a_h} = \frac{I_{t|h}}{I_{t-1|h}}.$$

■

In base alla (8.3), è possibile calcolare i numeri indici a base mobile a prescindere dai livelli assoluti della variabile d'interesse, $a_1, a_2, \ldots, a_k$, che potrebbero non essere disponibili.

ESEMPIO 8.4

Si calcolino i numeri indici a base mobile sulla base dei numeri indici a base fissa riportati nella Tabella 8.2.

Soluzione

Applicando la (8.3), si ottiene:

$$I_{2011|2010} = \frac{1.071}{1.000} = 1.071, \; I_{2012|2011} = \frac{1.094}{1.071} = 1.021, \ldots,$$

$$I_{2016|2015} = \frac{1.306}{1.268} = 1.130.$$

Moltiplicando per 100 i numeri indici così calcolati si ottengono i numeri indici percentuali riportati nella Tabella 8.5.

Anno	2010	2011	2012	2013	2014	2015	2016
N. indice a base mobile	–	107.1	102.1	105.6	103.8	105.8	103.0

Tabella 8.5
Numeri indici percentuali a base mobile calcolati in base ai dati della Tabella 8.2.

■

La Proposizione 8.3 mostra come convertire numeri indici a base mobile in numeri indici a base fissa.

PROPOSIZIONE 8.3

Data una serie di numeri indici a base mobile, $I_{2|1}, I_{3|2}, \ldots, I_{k|k-1}$, il numero indice del tempo t con base h può essere ottenuto nel modo seguente:

$$I_{t|h} = \begin{cases} I_{h+1|h} \cdot I_{h+2|h+1} \cdot \cdots \cdot I_{t|t-1}, & \text{se } h < t, \\ (I_{t+1|t} \cdot I_{t+2|t+1} \cdot \cdots \cdot I_{h|h-1})^{-1}, & \text{se } h > t. \end{cases} \tag{8.4}$$

La (8.4) può essere facilmente provata sostituendo a ogni fattore del prodotto la rispettiva espressione e, poi, semplificando. Ad esempio, se $h = 1$ e $t = 4$, si ha

$$I_{4|1} = I_{2|1} \cdot I_{3|2} \cdot I_{4|3} = \frac{a_2}{a_1} \cdot \frac{a_3}{a_2} \cdot \frac{a_4}{a_3} = \frac{a_4}{a_1}.$$

■

In altri termini, il numero indice del tempo t con base h si ottiene come indicato qui di seguito. Se $h < t$, bisogna moltiplicare tra loro i numeri indici a base mobile successivi alla base fino a quello del tempo t. Viceversa, se $h > t$, bisogna moltiplicare tra loro i numeri indici a base mobile, a partire da quello successivo al tempo t fino a quello relativo alla base, e prendere l'inverso del risultato ottenuto.

ESEMPIO 8.5

Si calcolino i numeri indici percentuali con base l'anno 2013 a partire dai numeri indici a base mobile dati nella Tabella 8.4.

Soluzione

Applicando le (8.4) si ottiene:

$$I_{2010|2013} = (1.071 \cdot 1.021 \cdot 1.056)^{-1} = 0.866,$$
$$I_{2011|2013} = (1.021 \cdot 1.056)^{-1} = 0.928,$$
$$I_{2012|2013} = (1.056)^{-1} = 0.947,$$
$$I_{2013|2013} = 1,$$
$$I_{2014|2013} = 1.038,$$
$$I_{2015|2013} = 1.038 \cdot 1.058 = 1.098,$$
$$I_{2016|2013} = 1.038 \cdot 1.058 \cdot 1.030 = 1.131.$$

I numeri indici percentuali corrispondenti sono mostrati nella Tabella 8.6.

Anno	2010	2011	2012	2013	2014	2015	2016
Numero indice	86.6	92.8	94.7	100.0	103.8	109.8	113.1

Tabella 8.6
Numeri indici percentuali con base l'anno 2013, calcolati utilizzando i dati della Tabella 8.4.

■

Variazioni relative

La variazione relativa tra due tempi, h e t, con $h < t$, di una serie storica è così definita

$$C_{t|h} = \frac{a_t - a_h}{a_h}.$$

La variazione relativa percentuale tra h e t si ottiene chiaramente moltiplicando $C_{t|h}$ per 100.

Poiché

$$C_{t|h} = \frac{a_t - a_h}{a_h} = \frac{a_t}{a_h} - 1 = I_{t|h} - 1,$$

la variazione relativa può essere anche ottenuta sottraendo 1 dal numero indice del tempo t rispetto al tempo h. Chiaramente, vale anche la relazione inversa:

$$I_{t|h} = C_{t|h} + 1.$$

ESEMPIO 8.6

Si calcolino le variazioni percentuali annue a partire dagli indici percentuali a base mobile della Tabella 8.4.

Soluzione

Le variazioni percentuali annue si ottengono sottraendo 100 dagli indici percentuali a base mobile riportati nella Tabella 8.4:

Tabella 8.7
Variazioni percentuali annue calcolate utilizzando i numeri indici percentuali a base mobile contenuti nella Tabella 8.4.

Anno	2010	2011	2012	2013	2014	2015	2016
Variazione perc.	–	7.1	2.1	5.6	3.8	5.8	3.0

Ciò significa che: nel 2011 si registra un incremento del 7.1% rispetto al 2010, nel 2012 un incremento del 2.1% rispetto al 2011 e così via. ■

Data una serie storica, la variazione relativa media (per unità di tempo) tra due tempi, h e t, con $h < t$ e $t - h \geq 2$, è data da

$$\bar{C}_{t|h} = \sqrt[t-h]{I_{h+1|h} \cdot I_{h+2|h+1} \cdot \cdots \cdot I_{t|t-1}} - 1. \tag{8.5}$$

Si osservi che la (8.5) è la media geometrica, diminuita di 1, dei numeri indici a base mobile dei tempi che vanno da h a t. In virtù della prima delle (8.4), la (8.5) può essere espressa in modo equivalente come

$$\bar{C}_{t|h} = \sqrt[t-h]{\frac{a_t}{a_h}} - 1 = \sqrt[t-h]{I_{t|h}} - 1, \tag{8.6}$$

dove a_h e a_t sono i termini h e t della serie storica $a_1, a_2, \ldots, a_k$. La (8.6) mostra che la variazione relativa media tra i tempi h e t può essere anche espressa in funzione di a_h e a_t, se questi sono noti, o in funzione del numero indice $I_{t|h}$.

La variazione percentuale media si ottiene moltiplicando per 100 la variazione relativa media.

ESEMPIO 8.7

Si calcoli la variazione percentuale media fra il 2013 e il 2016 per la serie storica riportata nella Tabella 8.1.

Soluzione

Se si vuole impiegare la (8.5), la variazione percentuale media è data da

$$\bar{C}_{2016|2013} = \sqrt[3]{1.038 \cdot 1.058 \cdot 1.030} - 1 = 0.042,$$

dove i fattori del prodotto sotto radice cubica sono i numeri indici a base mobile degli anni 2014, 2015 e 2016. Dunque, la variazione percentuale media tra il 2013 e il 2016 è 4.2%. Lo stesso risultato può essere ottenuto tramite la (8.6) utilizzando i termini a_{2013} e a_{2016} della serie storica riportata nella Tabella 8.1,

$$\bar{C}_{2016|2013} = \sqrt[3]{\frac{296\,790}{262\,443}} - 1 = 0.042,$$

oppure il numero indice $I_{2016|2013}$ contenuto nella Tabella 8.6,

$$\bar{C}_{2016|2013} = \sqrt[3]{1.131} - 1 = 0.042.$$

■

8.3 Numeri indici sintetici

I numeri indici esaminati nel paragrafo precedente si impiegano per descrivere le variazioni di una singola variabile (il prezzo di un certo prodotto, la popolazione di un dato Paese ecc.). Un problema importante in economia è quello di misurare la variazione media di due o più variabili. Per esempio, uno dei compiti principali di ogni Ufficio nazionale di statistica è quello di misurare la variazione media dei prezzi dei beni e dei servizi. A tale scopo, per ogni genere di consumo o servizio considerato, viene costruito un numero indice che misura la variazione relativa del prezzo. Il problema è quello di sintetizzare con un solo numero la variazione relativa dei prezzi di tutti i generi osservati. È intuitivamente ragionevole che una sintesi di tali variazioni relative possa essere ottenuta tramite un'appropriata media.

In generale, il problema è come combinare i numeri indici di un insieme di variabili per ottenere un numero indice di sintesi che esprima la variazione relativa media delle variabili dell'insieme. Il tema è dunque come costruire tali formule di sintesi denominate numeri indici sintetici.

ESEMPIO 8.8

Si voglia misurare l'incremento medio dei prezzi dei generi alimentari in un anno (ad esempio dal gennaio 2018 al gennaio 2019) per una famiglia di tre persone composta da due adulti e un bambino. A questo fine, si considerino, per semplicità, un insieme di 10 generi di consumo che rappresentano la struttura dei consumi alimentari della famiglia. La Tabella 8.8 mostra il numero indice del prezzo di gennaio 2019 con base gennaio 2018 e la spesa (in euro) sostenuta per ciascuno dei 10 prodotti nel gennaio 2018.

Nella prima colonna della Tabella 8.8, A può indicare, ad esempio, "bistecca di vitello", B "latte" ecc. Il problema è quello di calcolare la variazione media dei prezzi dei 10 generi di consumo.

Tabella 8.8
Numeri indici e spese per un insieme di generi di consumo.

Genere di consumo	Numero indice	Spesa a gennaio 2018
A	1.039	58.5
B	1.034	27.0
C	1.045	27.6
D	1.043	75.4
E	1.037	28.3
F	1.051	21.8
G	1.005	35.9
H	1.008	9.3
I	1,021	11.5
L	1.016	7.0
Totale	10.299	302.3

Soluzione

Una prima, fragile, soluzione consiste nel calcolare la media aritmetica dei dieci numeri indici e nel sottrarre 1 al risultato. In questo modo, si ottiene

$$\frac{1.039 + 1.034 + \cdots + 1.016}{10} = 1.030,$$

da cui si deduce un incremento relativo medio dei prezzi pari a 0.030 (3%).

Con questo calcolo si assegna lo stesso peso a ogni genere di consumo, mentre l'intuizione suggerisce che i pesi dovrebbero variare in ragione dell'importanza relativa che il singolo genere ha nel bilancio della famiglia. Per esempio, i prodotti A e L dovrebbero ricevere pesi molto diversi, di modo che, una data variazione relativa del prezzo di A abbia un impatto (sul risultato finale) molto maggiore di una variazione dello stesso livello nel prezzo di L. Ne segue che un procedimento più appropriato consiste nel calcolare il numero indice medio prendendo la media aritmetica ponderata dei 10 numeri indici con pesi pari alle spese sostenute per l'acquisto dei vari generi di consumo. Nell'esempio, tale media è data da

$$\frac{1.039 \cdot 58.5 + 1.034 \cdot 27 + \cdots + 1.016 \cdot 7}{58.5 + 27 + \cdots + 7} = 1.035.$$

Dunque, l'incremento relativo medio diviene 0.035 (3.5%), un valore sensibilmente diverso da quello ottenuto con la media aritmetica non ponderata. Ciò significa che per acquistare i generi A, B, ..., L nel gennaio 2019 negli stessi quantitativi e della stessa qualità di quelli del gennaio 2018, la famiglia deve sostenere una spesa pari a

$$302.3 + \frac{302.3 \cdot 3.5}{100} = 312.9.$$

■

Qui di seguito, verranno illustrati i due principali numeri indici sintetici con riferimento al problema della misura della variazione relativa media dei prezzi.

Siano $p_{1b}, p_{2b}, \ldots, p_{kb}$ i prezzi nel tempo base di k beni o servizi appartenenti al cosiddetto paniere, e siano $p_{1t}, p_{2t}, \ldots, p_{kt}$ i prezzi degli stessi beni e servizi nel tempo corrente. I rapporti

$$\frac{p_{1t}}{p_{1b}}, \frac{p_{2t}}{p_{2b}}, \cdots, \frac{p_{kt}}{p_{kb}} \tag{8.7}$$

sono i numeri indici che misurano le variazioni relative dei prezzi dei singoli beni e servizi del paniere dal tempo base al tempo corrente. Per definire i numeri indici sintetici dei prezzi, si indichino con $q_{1b}, q_{2b}, \ldots, q_{kb}$ e $q_{1t}, q_{2t}, \ldots, q_{kt}$ le quantità di beni e servizi del paniere acquistate nel tempo base e in quello corrente, rispettivamente. I prodotti

$$p_{1b} \cdot q_{1b}, p_{2b} \cdot q_{2b}, \ldots, p_{kb} \cdot q_{kb}$$

sono i valori dei beni e servizi del paniere ai prezzi del tempo base. Analogamente,

$$p_{1t} \cdot q_{1t}, p_{2t} \cdot q_{2t}, \ldots, p_{kt} \cdot q_{kt}$$

sono i valori dei medesimi beni e servizi ai prezzi correnti.

Un primo indice sintetico dei prezzi è dato dalla media aritmetica ponderata dei rapporti (8.7) con pesi uguali ai valori dei beni e servizi del paniere al tempo base:

$$I_L = \frac{\sum_{i=1}^{k} \frac{p_{it}}{p_{ib}} p_{ib} q_{ib}}{\sum_{i=1}^{k} p_{ib} q_{ib}}.$$

Questa espressione può essere posta nella forma

$$I_L = \frac{\sum_{i=1}^{k} p_{it} q_{ib}}{\sum_{i=1}^{k} p_{ib} q_{ib}}. \tag{8.8}$$

La (8.8) è nota come indice di Laspeyres dei prezzi (perché proposta dall'economista tedesco Etienne Laspeyres, 1834-1913). La variazione relativa media dei prezzi dal tempo base al tempo corrente si ottiene sottraendo 1 alla (8.8). Naturalmente, l'indice percentuale si ottiene moltiplicando per 100 il secondo membro della (8.8).

ESEMPIO 8.9

La tabella che segue riporta i prezzi (in euro) e il numero (in milioni) delle azioni circolanti di tre società all'inizio e alla fine del 2019 (Fonte: Bloomberg):

Tabella 8.9
Prezzi e numero di azioni circolanti di tre società.

Società	1 gennaio 2019		31 dicembre 2019	
	Quotazione p_{ib}	N. di azioni q_{ib}	Quotazione p_{it}	N. di azioni q_{it}
Ferrari	87.45	188.954	147.90	185.866
Fiat Chrysler	14.91	1540.090	13.19	1567.519
Leonardo S.p.A.	9.92	578.150	10.45	578.150

Si calcoli la variazione percentuale media del valore delle azioni delle tre società, usando l'indice dei prezzi di Laspeyres.

Soluzione

L'indice di Laspeyres è dato da

$$I_L = \frac{147.90 \cdot 188.954 + 13.19 \cdot 1540.090 + 10.45 \cdot 578.150}{87.45 \cdot 188.954 + 14.91 \cdot 1540.090 + 9.92 \cdot 578.150} = 1.201.$$

La conclusione è che l'incremento percentuale medio annuo del valore dei tre titoli è pari a 20.1%.

■

Una via alternativa per misurare la variazione media dei prezzi consiste nel calcolare la media armonica ponderata dei numeri indici elencati nella (8.7) assumendo come pesi i valori monetari dei beni e servizi del paniere ai prezzi del tempo corrente, valori dati da $p_{1t} \cdot q_{1t}, p_{2t} \cdot q_{2t}, \ldots, p_{kt} \cdot q_{kt}$. Procedendo in questo modo, si perviene alla formula

$$I_P = \frac{\sum_{i=1}^{k} p_{it} q_{it}}{\sum_{i=1}^{k} \frac{p_{it} q_{it}}{p_{it}/p_{jb}}},$$

che, semplificando, si riduce a

$$I_P = \frac{\sum_{i=1}^{k} p_{it} q_{it}}{\sum_{i=1}^{k} p_{ib} q_{it}}. \tag{8.9}$$

La (8.9) è denominata indice dei prezzi di Paasche (dovuto all'economista tedesco Hermann Paasche, 1851-1925). La variazione relativa media dei prezzi si deduce dalla (8.9) sottraendo 1. Come sempre, la variazione percentuale media si ottiene da quella relativa moltiplicando questa per 100.

ESEMPIO 8.10

Si calcoli la variazione percentuale media del valore dei tre titoli azionari considerati nell'Esempio 8.9 impiegando l'indice di Paasche.

Soluzione

Dall'applicazione della (8.9) ai dati in esame, si ottiene

$$I_P = \frac{147.90 \cdot 185.866 + 13.19 \cdot 1567.519 + 10.45 \cdot 578.15}{87.45 \cdot 185.866 + 14.91 \cdot 1567.519 + 9.92 \cdot 578.15} = 1.195.$$

Dunque, nell'anno 2019, la variazione percentuale media delle quotazioni dei tre titoli azionari è 19.5%.

■

In passato, vi è stato dibattito tra gli economisti sull'idoneità degli indici dei prezzi del tipo di quelli appena introdotti a misurare la variazione reale del "costo della vita". A rigore, la misura effettiva della variazione media del costo della vita sarebbe data dal rapporto avente al denominatore la spesa al tempo base per un assegnato paniere di beni e servizi e al numeratore la spesa minima per un paniere che assicuri la stessa soddisfazione (utilità) del tempo base. Questa impostazione è fondata sul cosiddetto "effetto di sostituzione" per cui il consumatore adatta il proprio comportamento di spesa aumentando il consumo dei beni e servizi i cui prezzi sono cresciuti di meno e diminuendo il consumo di quelli con incrementi di prezzo maggiori. Ad esempio, se il prezzo della carne di vitello aumenta, il consumatore potrebbe passare al consumo di una tipologia di carne i cui prezzi crescono di meno.

Ne segue che l'indice di Laspeyres, mantenendo ferme le quantità del tempo base, tende a sovrastimare il costo della vita, mentre, all'opposto, l'indice di Paasche tende a sottostimarlo.[1]

Un numero indice alternativo a I_L e I_P è la formula ideale di Fisher (dal nome dell'economista americano Irving Fisher, 1867-1947), data dalla media geometrica degli indici di Laspeyres e di Paasche:

$$I_F = \sqrt{I_L \cdot I_P}.$$

8.4 Indicatori compositi

In questo paragrafo, verrà affrontato in breve il problema del confronto tra unità statistiche, tipicamente unità territoriali (comuni, regioni, Paesi ecc.), mediante la quantificazione, per ciascuna di esse, di un fenomeno multidimensionale, nel senso che si manifesta con una pluralità di aspetti o dimensioni, che vanno colti simultaneamente.

Si pensi, ad esempio, al "Benessere equo e sostenibile" (BES in breve) oggetto di misurazione da parte dell'Istat. Si tratta di un concetto multidimensionale, che, nella più recente impostazione, viene declinato in 12 dimensioni: salute, istruzione e formazione, lavoro e conciliazione tempi di vita, benessere economico, relazioni sociali, politica e istituzioni, sicurezza, benessere soggettivo, paesaggio e patrimonio culturale, ambiente, ricerca e innovazione, qualità dei servizi. A loro volta, le singole dimensioni si articolano in sottodimensioni a cui vengono associati indicatori elementari. Ad esempio, la dimensione salute presenta 15 sottodimensioni misurate da altrettanti indicatori elementari.[2]

L'esempio del BES consente di dare una definizione generale di *indicatore composito*.

DEFINIZIONE 8.3

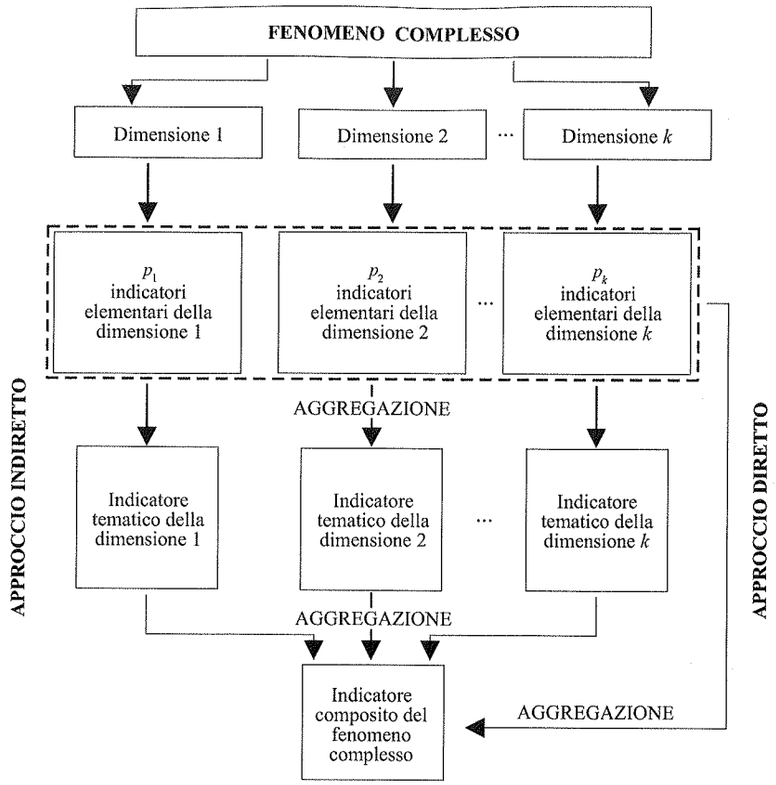
L'indicatore composito di un fenomeno complesso avente k dimensioni, ciascuna delle quali è descritta da un insieme di indicatori elementari, è un indice statistico costruito come segue. Per ciascuna dimensione, si determina un indicatore tematico, aggregando gli indicatori elementari associati alla dimensione. L'indicatore composito è la risultante dell'aggregazione degli indicatori tematici delle k dimensioni (approccio indiretto). Quando gli indicatori elementari non sono raggruppati in dimensioni o quando si ha un numero limitato di dimensioni, è possibile ottenere l'indicatore composito finale come aggregazione diretta di tutti gli indicatori elementari (approccio diretto).

Nella Figura 8.1, viene rappresentato il processo di costruzione dell'indicatore composito per la singola unità statistica. Il primo livello in alto è quello delle k dimensioni del fenomeno complesso. Per la generica dimensione j ($j = 1, 2, \ldots, k$),

[1] Si può dimostrare che l'indice di Laspeyres è maggiore dell'indice di Paasche se tra i rapporti di prezzo p_{it}/p_{ib} e i rapporti di quantità q_{it}/q_{ib} vi è correlazione negativa, se, cioè, all'aumentare del prezzo p_{it}, la quantità q_{it} tende a diminuire.

[2] Mortalità evitabile (0-74 anni), mortalità infantile, mortalità per incidenti stradali (15-34 anni), mortalità per tumore (20-64 anni), mortalità per demenze e malattie del sistema nervoso (65 anni e più), multicronicità e limitazioni gravi (75 anni e più), speranza di vita senza limitazioni nelle attività a 65 anni, eccesso di peso, fumo, alcol, sedentarietà (tassi standardizzati), adeguata alimentazione.

Figura 8.1
Processo di costruzione di un indicatore composito.



si hanno p_j indicatori elementari; da questi si costruisce un indicatore tematico della dimensione j. Si perviene poi all'indicatore composito finale.

Pur non potendo scendere nei dettagli delle tecniche statistiche, anche avanzate, che vengono impiegate nei diversi stadi del processo di costruzione degli indici compositi, è indispensabile accennare alle principali questioni di metodo statistico che si pongono lungo questo percorso.

Trasformazioni

Prima di procedere alla loro aggregazione, è importante verificare se gli indicatori elementari presentano livelli di misurazione confrontabili. In assenza di confrontabilità, è necessario procedere a trasformazioni la cui natura dipende dal tipo di dati.

Con riferimento alla generica dimensione j, sia $I_{1l}, I_{2l}, \ldots, I_{nl}$ la serie dei valori assunti dall'indicatore elementare l nelle n unità oggetto della comparazione (viene omesso dalla simbologia l'indice j per semplicità). Siano $\min_i(I_{il})$, $\max_i(I_{il})$, μ_l, σ_l, il minimo, il massimo, la media e la deviazione standard di questa serie.

Ciò posto, si riportano di seguito alcune delle trasformazioni a cui può essere sottoposto l'indicatore elementare I_{il}:

$$\frac{I_{il}}{\mu_l}, \frac{I_{il}}{\sigma_l}, \frac{I_{il}-\mu_l}{\sigma_l}, \frac{I_{il}}{\max_i(I_{il})}, \frac{I_{il}-\min_i(I_{il})}{\max_i(I_{il})-\min_i(I_{il})}.$$

La pluralità delle tecniche segnala che la trasformazione da applicare può variare da un indicatore elementare all'altro, da una dimensione all'altra.

Una trasformazione speciale è quella che serve a conferire polarità positiva agli indicatori elementari che avessero polarità negativa, intendendo per polarità positiva il fatto che la variazione positiva dell'indicatore segnala il miglioramento dell'aspetto che esso rappresenta. È importante che tutti gli indicatori elementari presentino polarità positiva, perché ciò assicura che all'aumentare di uno di essi e a parità di tutti gli altri aumenterà anche l'indice composito.

Il cambio di polarità, da negativa a positiva, per l'indicatore elementare I_{il} si ottiene tramite una delle seguenti trasformazioni:

$$\max_i(I_{il}) - I_{il}, \quad \frac{1}{I_{il}}.$$

In definitiva, le trasformazioni servono a rendere omogenei, in qualche senso, gli indicatori elementari per poter procedere alla loro ponderazione e aggregazione.

Ponderazione

Per definire i pesi che esprimono la diversa importanza e il diverso contributo degli indicatori elementari nella costruzione degli indicatori compositi si possono seguire tre approcci:

- approccio soggettivo;
- approccio oggettivo;
- approccio ibrido.

Con il primo approccio, quando non si hanno sufficienti informazioni *a priori* sulla loro rilevanza, gli indicatori elementari entrano nell'aggregazione con lo stesso peso. Viceversa, quando ci sono le condizioni per differenziarne l'importanza, il sistema di pesi viene definito da un gruppo di esperti attraverso un processo dialettico (*forum* di discussione).

Con l'approccio oggettivo, i pesi vengono determinati attraverso opportune procedure statistiche solitamente di tipo multivariato (ad esempio, l'analisi in componenti principali).

Con l'approccio ibrido, i pesi vengono determinati in base al giudizio di valore espresso da un campione statistico di individui e utilizzando metodi statistici del tipo di quelli appena evocati.

Aggregazione

Definito il sistema di ponderazione, l'indicatore composito si ottiene mediante l'aggregazione degli indicatori di livello inferiore. I metodi di aggregazione si distinguono in additivi e moltiplicativi. Tra i primi, il più usato è la media aritmetica ponderata dei valori degli indicatori originari (eventualmente trasformati); tra i secondi, il più usato è la media geometrica ponderata dei medesimi indicatori. I metodi additivi ipotizzano la completa sostituibilità tra le componenti dell'indice, nel senso che valori bassi di alcuni indicatori possano essere bilanciati da valori alti di altri indicatori.

Come si è detto all'inizio del paragrafo, gli indicatori compositi servono per la comparazione di unità spaziali (Stati, regioni, comuni ecc.) in termini di misura del fenomeno complesso di interesse. Essi vengono spesso usati per formare graduatorie tra le unità statistiche oggetto della comparazione e per valutare e spiegare le variazioni nel tempo in termini di variazioni negli indicatori elementari.

In considerazione dell'importanza assunta dagli indicatori compositi come strumento di monitoraggio e di orientamento delle politiche pubbliche e dell'importanza della loro standardizzazione per rendere possibili comparazioni a livello sovranazionale, l'OECD, in collaborazione con la Commissione europea, ha predisposto un manuale sul problema della loro costruzione (si veda OECD, 2008).

ESEMPIO 8.11

A titolo di esemplificazione, viene descritto il processo di costruzione dell'indice di digitalizzazione dell'economia e della società (DESI), con cui dal 2014 la Commissione europea monitora, con cadenza annuale, i progressi digitali realizzati dai 27 Paesi dell'UE. Il DESI è un indice composito che consente di valutare lo stato di digitalizzazione di ciascun Paese membro e di identificare le aree che richiedono investimenti e azioni prioritari.

A partire dal 2021, le dimensioni del DESI sono quattro: capitale umano (nel quale rientrano le competenze digitali di base e avanzate), connettività, integrazione delle tecnologie digitali, servizi pubblici digitali. Le quattro dimensioni si suddividono a loro volta in 10 sottodimensioni (due per la prima, quattro per la seconda, tre per la terza e una per la quarta) e ciascuna dimensione è descritta da un insieme di indicatori elementari (sette per la prima, 10 per la seconda, 11 per la terza e cinque per la quarta). Gli indicatori elementari di ciascuna dimensione vengono sottoposti alla trasformazione

$$\frac{I_{il} - \min_i(I_{il})}{\max_i(I_{il}) - \min_i(I_{il})}.$$

L'indicatore composito finale per il Paese membro i ($i = 1, 2, \ldots, 27$) è il risultato dell'aggregazione degli indicatori compositi delle quattro dimensioni tramite la media aritmetica ponderata con pesi uguali.

Nella Figura 8.2 viene mostrata la graduatoria dei 27 Paesi UE e dell'UE in base al DESI del 2021 (Fonte: Eurostat).

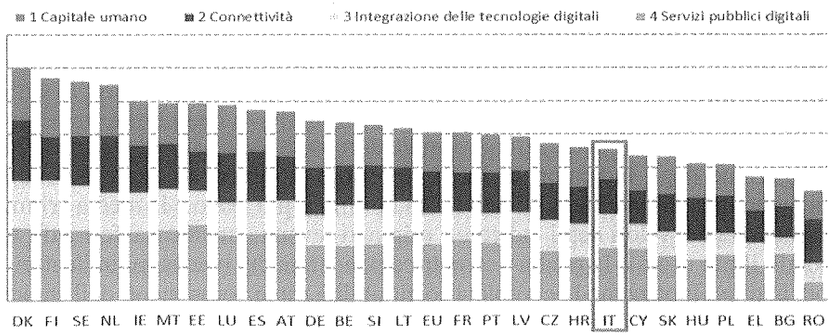


Figura 8.2
Graduatoria dei 27 Paesi UE e dell'UE in base all'indice DESI del 2021.

Legenda: Belgio (BE), Bulgaria (BG), Cechia (CZ), Danimarca (DK), Germania (DE), Estonia (EE), Irlanda (IE), Grecia (EL), Spagna (ES), Francia (FR), Croazia (HR), Italia (IT), Cipro (CY), Lettonia (LV), Lituania (LT), Lussemburgo (LU), Ungheria (HU), Malta (MT), Paesi Bassi (NL), Austria (AT), Polonia (PL), Portogallo (PT), Romania (RO), Slovenia (SI), Slovacchia (SK), Finlandia (FI), Svezia (SE), Unione Europea (EU).

■

Riassunto

I numeri indici sono particolari rapporti statistici che servono per mettere a confronto i valori assunti da una variabile quantitativa in tempi o luoghi diversi. Particolarmente importanti sono i numeri indici derivabili dalle serie storiche, che hanno la funzione di facilitare la comprensione della dinamica temporale della variabile d'interesse.

Quando il confronto nel tempo riguarda non una singola variabile, ma un insieme di variabili omogenee, come ad esempio i prezzi di un insieme di beni e servizi, è necessario sintetizzare i numeri indici delle singole variabili con appropriate formule, denominate numeri indici sintetici, che esprimono la dinamica temporale media dell'insieme delle variabili. In questo ambito, vengono presentate due fondamentali misure, la formula di Laspeyres e quella di Paasche, con cui si misura la variazione media dei prezzi di un paniere di beni e servizi dal tempo base al tempo corrente. Viene anche affrontato l'argomento dei cosiddetti indicatori compositi che ha attinenza con i numeri indici sintetici. Si tratta di strumenti con cui si confrontano più unità statistiche, tipicamente unità territoriali (comuni, regioni, Stati ecc.) riguardo al livello di fenomeno socio-economico complesso, che viene quantificato facendo la sintesi di un insieme di indicatori elementari che riguardano i diversi aspetti del fenomeno.

Parole chiave

dimensioni, 187
formula ideale di Fisher, 187
indicatori compositi, 187
indicatori elementari, 187
indicatori tematici, 187
indice dei prezzi di Laspeyres, 185
indice dei prezzi di Paasche, 186
numeri indici a base fissa, 178
numeri indici a base mobile, 179
numeri indici percentuali, 178
numeri indici sintetici, 183
paniere, 184
polarità positiva, 189
variazione percentuale media, 182
variazione relativa, 182
variazione relativa media, 182

Commenti e curiosità

1. Incrementi e decrementi percentuali

Una grandezza che è espressa da numeri non negativi non può diminuire per più del 100%. Una tale diminuzione infatti implica l'azzeramento del valore iniziale. Capita talvolta di ascoltare affermazioni paradossali che contrastano con questa ovvia proprietà. Per converso, non esistono limiti superiori per gli incrementi percentuali. Se, per esempio, la produzione di un'azienda passa da 3.865.800 unità di un anno a 11.597.400 dell'anno successivo, essa si incrementa del 200%.

2. I numeri indici dei prezzi al consumo elaborati dall'Istat

L'Istat elabora tre diversi numeri indici dei prezzi al consumo:

- indice nazionale dei prezzi al consumo per l'intera collettività (NIC), che è impiegato come misura dell'inflazione a livello dell'intero sistema economico e costituisce un punto di riferimento essenziale per l'impostazione delle politiche economiche nazionali;
- indice dei prezzi al consumo per le famiglie di operai e impiegati (FOI), che fa riferimento solo ai consumi delle famiglie che fanno capo a un lavoratore dipendente in settori extra-agricoli. Si tratta dell'indice al quale la legislazione italiana assegna gli effetti amministrativi più importanti per l'adeguamento periodico di valori espressi in moneta corrente (per esempio, per l'adeguamento periodico degli affitti);
- indice armonizzato dei prezzi al consumo per i paesi dell'Unione europea (IPCA), che è stato sviluppato per misurare la dinamica dell'inflazione nel tempo con criteri che assicurano la comparabilità a livello europeo.

Il calcolo dei numeri indici è basato su un paniere rappresentativo di prodotti. Nel paniere del 2022 utilizzato per gli indici NIC e FOI figurano 1772 prodotti elementari, raggruppati in 1031 prodotti, che a loro volta sono raccolti in 422 aggregati. In quello per il IPCA i prodotti elementari sono invece 1792, raggruppati in 1050 prodotti e 426 aggregati.

Una parte rilevante delle quotazioni di prezzo proviene dalla grande distribuzione organizzata (dati da *scanner*). Un'altra parte proviene dalla raccolta sul territorio ad opera di un campione rappresentativo di Comuni (80 nel 2022).

La formula di base per il calcolo della variazione media dei prezzi è l'indice di Laspeyres con pesi riferiti all'anno precedente a quello corrente. Infatti, a partire dal 1999, l'Istat procede annualmente alla revisione sia del paniere dei prodotti che del sistema di ponderazione. I pesi vengono stimati in base ai dati relativi ai consumi delle famiglie rilevati attraverso la contabilità generale e l'indagine sui bilanci delle famiglie.

3. Come determinare le variazioni percentuali dei prezzi al consumo per l'intera collettività

Nel sito dell'Istat sono disponibili i numeri indici mensili per gli anni che vanno dal 1996 a oggi. Per calcolare la variazione percentuale dei prezzi per un dato livello di aggregazione di prodotti (sottoclasse, classe, ..., intero paniere) tra due mesi successivi alla base corrente (i cui numeri indici presentano la stessa base 2015) basta rapportare il numero indice del mese finale al numero indice del mese iniziale, sottrarre 1 a tale rapporto e moltiplicare il risultato per 100.

In simboli, indicati con m_1 e m_2 il mese iniziale e quello finale, la variazione percentuale dei prezzi da m_1 a m_2 è data da $\left(\frac{I_{m_2|2015}}{I_{m_1|2015}} - 1\right) \cdot 100$. Se invece m_1 è posteriore al 2010 (la base precedente a quella attuale) e anteriore al 2016, mentre m_2 è posteriore al 2016, la variazione percentuale richiede che sia applicato il coefficiente di raccordo, Rac, dato dal numero indice medio dell'anno 2015 (quindi con base 2010) diviso per cento. La formula da applicare è dunque: $\left(\frac{I_{m_2|2015}}{I_{m_1|2010}} \cdot \text{Rac} - 1\right) \cdot 100$. Naturalmente, se tra m_1 e m_2 sono intervenuti due cambiamenti di base, i coefficienti di raccordo da applicare sono due e così via.

Come si è detto in precedenza, a partire da 1999, i pesi impiegati nella formula di Laspeyres per aggregare ai diversi livelli (sottoclasse, classe, ..., intero paniere) i numeri indici risultano dalla revisione anno per anno sia del paniere sia del sistema di ponderazione. Tutto ciò comporta che la base di calcolo dei numeri indici mensili dei dodici mesi di ciascun anno sia il mese di dicembre dell'anno precedente. Tuttavia, per consentire il calcolo delle variazioni percentuali in un arco temporale più ampio, l'Istat provvede a riportare i numeri indici così calcolati a una comune base di riferimento applicando la tecnica dei numeri indici a catena (per questo e per altri dettagli tecnici, si rinvia al documento Istat scaricabile in pdf: Indici dei prezzi al consumo - aspetti generali e metodologia di rilevazione, 2013).

Esercizi

8.1 Nuove immatricolazioni di automobili nei paesi UE (UE-28) nel 2019 (Fonte: European automobile manufacturers association, 2020):

Mese	Automobili (migliaia)	Mese	Automobili (migliaia)
Gen	1194	Lug	1293
Feb	1113	Ago	1041
Mar	1721	Set	1248
Apr	1302	Ott	1176
Mag	1399	Nov	1175
Giu	1444	Dic	1213

a) Si calcolino i numeri indici assumendo gennaio come tempo base.

b) Si calcolino i numeri indici a base mobile.

c) Si calcoli la variazione percentuale media mensile.

8.2 Addetti alla ricerca e sviluppo nel settore privato in Italia e nel Regno Unito dal 2012 al 2017 (Fonte: Eurostat, 2018):

Anno	Italia	Regno Unito
2012	49 950	113 589
2013	52 718	123 684
2014	54 216	133 891
2015	62 170	136 397
2016	72 393	141 873
2017	82 060	155 814

a) Si costruisca la serie dei numeri indici a base mobile per l'Italia e il Regno Unito e si metta a confronto l'andamento del fenomeno nei due paesi.

b) Si calcoli la variazione percentuale media annua per ciascuna serie.

8.3 La tabella che segue contiene la variazione percentuale annua del prodotto interno lordo pro-capite dei paesi europei (UE-27) dal 2010 al 2019 (Fonte: Eurostat, 2020):

Anno	Var. perc. annua	Anno	Var. perc. annua
2010	–	2015	2,13
2011	1.72	2016	1.88
2012	–0.98	2017	2.51
2013	–0.20	2018	1.99
2014	1.44	2019	1.34

a) Si costruisca la serie dei numeri indici percentuali a base mobile.

b) Sulla base della serie ottenuta al punto a), si determini la serie di numeri indici percentuali con base 2010.

c) Sapendo che il prodotto interno lordo pro-capite nel 2011 è 25 328 euro, si calcoli il prodotto interno lordo pro-capite per i restanti anni.

8.4 La tabella che segue riporta le esportazioni trimestrali del Nord-Est d'Italia dal primo trimestre 2018 al quarto trimestre 2019 (Fonte: Istat, 2020):

Anno 2018	Esportazioni (milioni di euro)	Anno 2019	Esportazioni (milioni di euro)
I trimestre	45 878	I trimestre	45 473
II trimestre	45 406	II trimestre	45 925
III trimestre	46 515	III trimestre	46 082
IV trimestre	46 263	IV trimestre	45 270

a) Si costruisca la serie dei numeri indici percentuali a base mobile.

b) Si calcoli la variazione percentuale media dal primo trimestre del 2018 al quarto trimestre del 2019.

8.5 Nella tabella che segue sono indicati i prezzi dell'elettricità e del gas metano insieme ai relativi consumi di una famiglia italiana per gli anni 2015 e 2019:

Anno	Elettricità		Gas metano	
	Prezzo (euro/kWh)	Consumo (kWh)	Prezzo (euro/mc)	Consumo (mc)
2015	18.7	2350	0.82	1620
2019	20.7	2500	0.83	1400

a) Si calcoli la variazione percentuale media dei prezzi dell'elettricità e del gas metano dal 2015 al 2019 per la famiglia considerata, usando l'indice di Laspeyres.

b) Si ripeta il calcolo del punto precedente usando l'indice di Paasche.

8.6 La tabella che segue mostra le quotazioni di chiusura del 31 gennaio 2018 e del 30 dicembre 2019 relative alle azioni delle S.p.A. Campari, Ferrari e Leonardo (Fonte: Bloomberg):

S.p.A.	Quotazione	
	31 gennaio 2018	30 dicembre 2019
Campari	6.41	8.14
Ferrari	118.09	165.30
Leonardo	9.72	10.45

a) Si calcoli l'indice di Laspeyres percentuale per un portafoglio composto, al 31 gennaio 2018, da: 1500 azioni Campari, 2000 azioni Ferrari e 3400 azioni Leonardo.

b) Si calcoli l'indice di Paasche percentuale per un portafoglio composto, al 30 dicembre 2019, da: 2200 azioni Campari, 1500 azioni Ferrari e 2300 azioni Leonardo.

CAPITOLO

9

Analisi della dipendenza nelle tabelle di contingenza

Sommario del capitolo

9.1 Introduzione
9.2 Dipendenza tra le variabili di una tabella di contingenza
9.3 Dipendenza in media
Riassunto
Commenti e curiosità
Esercizi

9.1 Introduzione

Presentare una serie doppia di osservazioni nella forma di distribuzione doppia di frequenza è necessario quando una o entrambe le variabili sono qualitative; è altamente conveniente quando la numerosità della serie è consistente. Nel Capitolo 2, è stato illustrato il procedimento per la costruzione di una tabella di contingenza a partire da una serie doppia di osservazioni. Conviene perciò riprendere lo schema generale di tale tabella già visto in quella sede.

Nella Tabella 9.1, $x_1, x_2, \ldots, x_s$ sono le modalità della variabile X, $y_1, y_2, \ldots, y_t$ quelle della variabile Y e n_{ij} è la frequenza congiunta della coppia di modalità (x_i, y_j). Nella colonna del totale, n_{i0} indica la somma delle frequenze della riga corrispondente alla modalità x_i di X e rappresenta il numero di unità in cui $X = x_i$. Analogamente, nella riga del totale, n_{0j} indica la somma delle frequenze della colonna corrispondente alla modalità y_j di Y e rappresenta il numero di unità in cui $Y = y_j$. La somma delle frequenze nell'ultima colonna è ovviamente uguale alla somma delle frequenze nell'ultima riga ed entrambe sono pari a n, il numero totale di unità.

Tabella 9.1 Distribuzione doppia di frequenza.

Variabile X	Variabile Y						Totale
	y_1	y_2	$\cdots$	y_j	$\cdots$	y_t	
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1t}	n_{10}
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2t}	n_{20}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{it}	n_{i0}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_s	n_{s1}	n_{s2}	$\cdots$	n_{sj}	$\cdots$	n_{st}	n_{s0}
Totale	n_{01}	n_{02}	$\cdots$	n_{0j}	$\cdots$	n_{0t}	n

Le variabili X e Y possono essere entrambe qualitative, una qualitativa e l'altra quantitativa, entrambe quantitative. Le modalità dei caratteri quantitativi (discreti o continui) sono generalmente raggruppate in classi.

Il problema centrale nell'analisi delle distribuzioni doppie di frequenza è lo studio della relazione associativa tra le due variabili: si tratta

di stabilire, ad esempio, se la variabile Y manifesta un legame di dipendenza dalla variabile X e di misurare la forza di tale legame con appropriati indici.

ESEMPIO 9.1

Si consideri la distribuzione doppia di frequenza contenuta nella Tabella 9.2 (Hand *et al.*, 1994):

Tabella 9.2
Distribuzione di 1329 soggetti maschi secondo il livello di colesterolo e i disturbi cardiaci.

Colesterolo	Disturbi cardiaci Sì	No	Totale
< 219	20	553	573
220-259	31	439	470
> 259	41	245	286
Totale	92	1237	1329

Con questi dati ha senso indagare sulla dipendenza dei disturbi cardiaci dal livello di colesterolo. ■

Per illustrare il criterio con cui si stabilisce e si misura il legame associativo tra le variabili di una distribuzione doppia di frequenza, è necessario enucleare dalla tabella a doppia entrata le distribuzioni semplici (ossia di una sola variabile) indicate di seguito.

Distribuzione marginale di Y

Modalità di Y	y_1	y_2	...	y_j	...	y_t	Totale
Frequenza	n_{01}	n_{02}	...	n_{0j}	...	n_{0t}	n

Distribuzione marginale di X

Modalità di X	x_1	x_2	...	x_i	...	x_s	Totale
Frequenza	n_{10}	n_{20}	...	n_{i0}	...	n_{s0}	n

Distribuzione condizionata di Y per $X = x_i$

Modalità di Y	y_1	y_2	...	y_j	...	y_t	Totale
Frequenza	n_{i1}	n_{i2}	...	n_{ij}	...	n_{it}	n_{i0}

Distribuzione condizionata di X per $Y = y_j$

Modalità di X	x_1	x_2	...	x_i	...	x_s	Totale
Frequenza	n_{1j}	n_{2j}	...	n_{ij}	...	n_{sj}	n_{0j}

Le distribuzioni marginali sono le distribuzioni di frequenza della sola variabile Y o della sola variabile X: è come se le n unità fossero classificate in base alle modalità della Y senza tener conto dell'altra variabile o viceversa.

Le distribuzioni condizionate sono distribuzioni di frequenza di una delle due variabili nel sottoinsieme di unità nelle quali l'altra variabile presenta una modalità assegnata. Per esempio, la distribuzione condizionata di Y per $X = x_i$ è la distribuzione di frequenza della variabile Y che riguarda esclusivamente le unità nelle quali $X = x_i$. È evidente che le distribuzioni condizionate di Y sono s e quelle di X sono t.

Le distribuzioni marginali e quelle condizionate sono state "estratte" dalla tabella a doppia entrata allo scopo di metterle in risalto: le si possono leggere direttamente da tale tabella e descrivere verbalmente. Così la distribuzione marginale di una variabile è costituita dalle modalità della variabile con le rispettive frequenze marginali; la distribuzione condizionata di una variabile è costituita dalle modalità della variabile con le frequenze della riga o della colonna corrispondente alla modalità prefissata dell'altra variabile.

ESEMPIO 9.2

Si riprenda la distribuzione doppia di frequenza considerata nell'Esempio 9.1. Si individuino le distribuzioni marginali e si descriva una distribuzione condizionata a scelta per ciascuna variabile.

Soluzione

La distribuzione marginale della variabile disturbi cardiaci è costituita dalle modalità sì e no con le rispettive frequenze 92 e 1237. La distribuzione marginale della variabile livello di colesterolo è costituita dalle classi < 219, 220-259 e > 259 con le rispettive frequenze 573, 470 e 286. La distribuzione condizionata della variabile disturbi cardiaci per un livello di colesterolo appartenente alla classe < 219 è costituita dalle modalità sì e no con le frequenze 20 e 553. La distribuzione condizionata del livello di colesterolo corrispondente alla modalità no della variabile disturbi cardiaci è costituita dalle classi < 219, 220-259 e > 259 con le frequenze 553, 439 e 245. ■

L'idea alla base dell'analisi della dipendenza è quella di stabilire se e in che misura le distribuzioni condizionate di una variabile cambiano al variare delle modalità dell'altra variabile. Ad esempio, lo studio della dipendenza di Y da X implica che si mettano a confronto le distribuzioni condizionate della variabile Y e se ne misuri in qualche modo il grado di diversità. Ma occorre fare attenzione: non si possono confrontare le distribuzioni condizionate così come sono:[1] è necessario renderle confrontabili, passando alle distribuzioni di frequenza relativa corrispondenti. Le distribuzioni di frequenza relativa condizionate della variabile Y si ottengono dividendo le frequenze di ciascuna riga della tabella a doppia entrata per il totale di riga. Analogamente, le distribuzioni di frequenza relativa condizionate della variabile X si determinano dividendo le frequenze di ciascuna colonna della tabella a doppia entrata per il totale di colonna.

[1] Le distribuzioni di frequenza assoluta di una medesima variabile non sono confrontabili, perché la loro comparazione porterebbe al paradosso di considerare diverse due distribuzioni aventi frequenze assolute tra loro proporzionali (come nel caso, ad esempio, sia la singola frequenza dell'una sia il doppio della frequenza corrispondente dell'altra), mentre in realtà esse devono essere considerate uguali, avendo le stesse proprietà statistiche: funzione di ripartizione, medie ecc.

ESEMPIO 9.3

Considerando la distribuzione doppia di frequenza dell'Esempio 9.1, si determinino le distribuzioni di frequenza relativa condizionate delle due variabili.

Soluzione

Nella Tabella 9.3, si possono leggere le distribuzioni di frequenza relativa condizionate della variabile disturbi cardiaci ottenute dalla Tabella 9.2 dividendo la frequenza della singola riga per il totale di riga.

Tabella 9.3
Distribuzioni di frequenza relativa condizionate della variabile disturbi cardiaci.

Colesterolo	Disturbi cardiaci Sì	No	Totale
< 219	0.035	0.965	1.000
220-259	0.066	0.934	1.000
> 259	0.143	0.857	1.000
Totale	0.069	0.931	1.000

Nella Tabella 9.4, si possono leggere le distribuzioni di frequenza relativa condizionate della variabile livello di colesterolo ottenute dalla Tabella 9.2 dividendo la frequenza della singola colonna per il totale di colonna.

Tabella 9.4
Distribuzioni di frequenza relativa condizionate della variabile colesterolo.

Colesterolo	Disturbi cardiaci Sì	No	Totale
< 219	0.217	0.447	0.431
220-259	0.337	0.355	0.354
> 259	0.446	0.198	0.215
Totale	1.000	1.000	1.000

■

9.2 Dipendenza tra le variabili di una tabella di contingenza

Conviene partire dalla nozione di indipendenza tra due variabili.

DEFINIZIONE 9.1

Data una distribuzione doppia di frequenza, si dice che le variabili X e Y sono indipendenti se le distribuzioni di frequenza relativa condizionate di una di esse sono uguali tra loro. Se ciò non accade, si dice che le due variabili presentano un legame associativo.

Dunque, in base alla Definizione 9.1, solo se le distribuzioni di frequenza relativa condizionate di una variabile non sono uguali tra loro si può dire che tra le variabili X e Y intercorre un legame associativo, in ragione del fatto che il variare delle modalità di una variabile, ad esempio X, influisce sulle distribuzioni condizionate della variabile Y.

ESEMPIO 9.4

Tornando alla Tabella 9.2 dell'Esempio 9.1, è possibile notare che le distribuzioni di frequenza relativa condizionate della variabile disturbi cardiaci non sono uguali tra loro. Infatti, all'aumentare del colesterolo aumenta l'incidenza dei disturbi cardiaci (si passa da 0.035 a 0.066 e a 0.143). In altre parole, il livello di colesterolo ha un impatto sulle distribuzioni di frequenza relativa condizionate dei disturbi cardiaci: ne segue che le due variabili presentano un legame associativo. Si può anche dire che i disturbi cardiaci dipendono dal livello di colesterolo. ■

Una tabella a doppia entrata in cui le due variabili sono indipendenti è detta *tabella di indipendenza*. Le frequenze congiunte di questa speciale tabella sono indicate nella definizione che segue.

DEFINIZIONE 9.2

Data una distribuzione doppia di frequenza, si chiama tabella di indipendenza la tabella a doppia entrata che ha le stesse distribuzioni marginali della distribuzione doppia osservata e presenta nella generica casella (i, j) la frequenza congiunta data da

$$\hat{n}_{ij} = \frac{n_{i0} \cdot n_{0j}}{n}, \quad \text{per ogni coppia } (i, j). \tag{9.1}$$

Si può facilmente dimostrare che le seguenti tre proposizioni sono modi equivalenti per affermare l'indipendenza tra X e Y: (a) la frequenza congiunta osservata n_{ij} è pari a $\hat{n}_{ij}$ per ogni coppia (i, j); (b) le distribuzioni di frequenza relativa condizionate di Y sono uguali tra loro; (c) le distribuzioni di frequenza relativa condizionate di X sono uguali tra loro. ■

ESEMPIO 9.5

Riprendendo la Tabella 9.2, si costruisca la tabella di indipendenza e si verifichi numericamente che in essa le distribuzioni di frequenza relativa condizionate di ciascuna variabile sono tra loro uguali.

Soluzione

Sostituendo a ciascuna frequenza congiunta della Tabella 9.2 la quantità ottenuta dall'applicazione della (9.1), si ottiene la tabella di indipendenza che segue:

Tabella 9.5
Tabella di indipendenza corrispondente alla Tabella 9.2.

Colesterolo	Disturbi cardiaci Sì	No	Totale
< 219	39.7	533.3	573.0
220-259	32.5	437.5	470.0
> 259	19.8	266.2	286.0
Totale	92.0	1237.0	1329.0

Da questa tabella, dividendo ciascuna frequenza per il rispettivo totale di riga, si ottiene:

Tabella 9.6
Distribuzioni di frequenza relativa condizionate della variabile disturbi cardiaci nell'ipotesi di indipendenza.

Colesterolo	Disturbi cardiaci Sì	No	Totale
< 219	0.069	0.931	1.000
220-259	0.069	0.931	1.000
> 259	0.069	0.931	1.000
Totale	0.069	0.931	1.000

Dalla stessa Tabella 9.5, dividendo ciascuna frequenza per il rispettivo totale di colonna, si ottiene:

Tabella 9.7
Distribuzioni di frequenza relativa condizionate della variabile colesterolo nell'ipotesi di indipendenza.

Colesterolo	Disturbi cardiaci Sì	No	Totale
< 219	0.432	0.432	0.432
220-259	0.353	0.353	0.353
> 259	0.215	0.215	0.215
Total	1.000	1.000	1.000

Come si vede, le distribuzioni di frequenza relativa condizionate sono tra loro uguali per entrambe le variabili. ■

La situazione opposta a quella di indipendenza è la *dipendenza perfetta*, che si ha quando a ciascuna modalità di una variabile corrisponde una sola modalità dell'altra variabile.

ESEMPIO 9.6

Le tabelle che seguono sono esempi di dipendenza perfetta: nel primo caso, la dipendenza perfetta sussiste in entrambe le direzioni (vi è dipendenza perfetta di Y da X e viceversa); nel secondo, la dipendenza perfetta vale solo in una direzione: vi è dipendenza perfetta di Y da X.

Tabella 9.8
Distribuzione doppia di frequenza ipotetica in cui vi è dipendenza perfetta di una variabile rispetto all'altra.

X	Y y_1	y_2	y_3	Totale
x_1	12	0	0	12
x_2	0	0	30	30
x_3	0	27	0	27
Totale	12	27	30	69

Tabella 9.9
Distribuzione doppia di frequenza ipotetica in cui Y dipende perfettamente da X.

X	Y y_1	y_2	y_3	Totale
x_1	0	0	15	15
x_2	0	50	0	50
x_3	23	0	0	23
x_4	0	11	0	11
Totale	23	61	15	99

■

DEFINIZIONE 9.3

Una tabella di dipendenza perfetta con s righe e t colonne è definita come: (a) la tabella in cui per ogni i vi è un solo j tale che $n_{ij} > 0$ e viceversa; ciò può avvenire solo se $s = t$; (b) la tabella in cui per ogni i vi è un solo j tale che $n_{ij} > 0$; ciò può avvenire solo se $t < s$; (c) la tabella in cui per ogni j vi è un solo i tale che $n_{ij} > 0$; ciò può avvenire solo se $s < t$.

9.2.1 Misura della dipendenza

La via più immediata per costruire una misura della dipendenza tra X e Y consiste nel confrontare le frequenze n_{ij} della tabella di contingenza osservata con le corrispondenti frequenze $\hat{n}_{ij}$ della tabella di indipendenza. A tale scopo, si considerino i rapporti:

$$c_{ij} = \frac{n_{ij} - \hat{n}_{ij}}{\hat{n}_{ij}}, \quad \text{per ogni coppia } (i, j),$$

noti come contingenze, che sono alla base di un primo indice di dipendenza.

DEFINIZIONE 9.4

Data una distribuzione doppia di frequenza, una misura di dipendenza tra X e Y è data da

$$\Phi = \sqrt{\frac{1}{n}\sum_{i=1}^{s}\sum_{j=1}^{t}\frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}} = \sqrt{\sum_{i=1}^{s}\sum_{j=1}^{t}\frac{n_{ij}^2}{n_{i0}\cdot n_{0j}} - 1}, \qquad (9.2)$$

formula interpretabile come media quadratica ponderata delle contingenze c_{ij} con pesi pari a $\hat{n}_{ij}$.[2]

[2] La (9.2) può essere posta nella forma

$$\Phi = \sqrt{\frac{1}{n}\sum_{i=1}^{s}\sum_{j=1}^{t}\left(\frac{n_{ij} - \hat{n}_{ij}}{\hat{n}_{ij}}\right)^2 \hat{n}_{ij}},$$

da cui, essendo $\sum_{i=1}^{s}\sum_{j=1}^{t}\hat{n}_{ij} = n$, si evince che Φ è la media quadratica ponderata delle contingenze con pesi $\hat{n}_{ij}$. Il passaggio dalla prima alla seconda espressione di Φ nella (9.2) è basato sullo sviluppo di $(n_{ij} - \hat{n}_{ij})^2$, sulla sostituzione di $(n_{i0}n_{0j})/n$ a $\hat{n}_{ij}$ e sulle conseguenti semplificazioni.

La formula (9.2), come misura della "distanza" tra la tabella di contingenza osservata e quella di indipendenza, esprime la forza del legame associativo tra X e Y: quanto maggiore è questa distanza tanto più elevato è il valore assunto da Φ.

ESEMPIO 9.7

Si calcoli l'indice Φ per la distribuzione doppia di frequenza dell'Esempio 9.1.

Soluzione

Si considerino le frequenze osservate della Tabella 9.2 e le frequenze teoriche della tabella di indipendenza dell'Esempio 9.5. Applicando la formula (9.2), si ottiene

$$\Phi = \sqrt{\frac{1}{1329}\left[\frac{(20-39.7)^2}{39.7} + \frac{(553-533.3)^2}{533.3} + \cdots + \frac{(245-266.2)^2}{266.2}\right]} = 0.16.$$

■

PROPOSIZIONE 9.1

L'indice Φ varia tra 0 e $\sqrt{\min(s,t)-1}$, dove $\min(s,t)$ è il valore più piccolo tra il numero di righe e il numero di colonne della tabella di contingenza. Assume il valore 0 nel caso di indipendenza, ossia quando $n_{ij} = \hat{n}_{ij}$, $i = 1,2,\ldots,s$, $j = 1,2,\ldots,t$; raggiunge il valore massimo nel caso di dipendenza perfetta.

La dimostrazione che $\Phi \leq \sqrt{\min(s,t)-1}$ è contenuta nel Paragrafo A.6 dell'Appendice A.

■

Talora, come misura di dipendenza viene impiegato, in luogo della (9.2), l'indice

$$\chi^2 = \sum_{i=1}^{s}\sum_{j=1}^{t}\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}} = n\left(\sum_{i=1}^{s}\sum_{j=1}^{t}\frac{n_{ij}^2}{n_{i0}\cdot n_{0j}} - 1\right), \tag{9.3}$$

che è legato a Φ dalla relazione $\chi^2 = n\Phi^2$. L'indice Φ è una misura simmetrica di dipendenza, nel senso che il suo valore non dipende da quale variabile è definita come X o Y.

Dividendo Φ per il suo massimo, si perviene a una misura relativa di dipendenza, nota come indice V di Cramer:

$$V = \frac{\Phi}{\sqrt{\min(s,t)-1}}. \tag{9.4}$$

L'indice (9.4) varia tra 0 e 1. Assume il valore 0 nel caso di indipendenza e il valore 1 nel caso di dipendenza perfetta. L'indice V di Cramer è preferibile all'indice Φ in un contesto di analisi descrittiva, perché è di facile interpretazione: tanto più V è vicino a 1 tanto più forte è la dipendenza.

ESEMPIO 9.8

Si calcoli l'indice V di Cramer per la distribuzione doppia di frequenza dell'Esempio 9.1.

Soluzione

L'indice Φ, già calcolato nell'Esempio 9.6, è pari a 0.16. Il denominatore dell'indice V è 1, essendo

$$\sqrt{\min(s,t)-1} = \sqrt{\min(3,2)-1} = 1.$$

Ne segue che $V = 0.16$.

■

ESEMPIO 9.9

La Tabella 9.10 riporta i dati di un'indagine condotta nel 2011 a Somerville (Massachussetts) in cui fu chiesto agli intervistati di esprimersi sulla bellezza della città e sul loro grado di felicità. Si calcolino gli indici χ^2 e V.

Tabella 9.10
Rispondenti nell'indagine in questione per giudizio sulla bellezza della città e sulla loro felicità.

Felicità	Bellezza della città			Totale
	Scarsa	Accettabile	Buona	
Infelice	90	123	62	275
Così così	555	972	610	2137
Felice	541	1426	1406	3373
Totale	1186	2521	2078	5785

Soluzione

Si determina innanzitutto la tabella di indipendenza utilizzando la (9.1):

Tabella 9.11
Tabella di indipendenza corrispondente alla Tabella 9.10.

Felicità	Bellezza della città			Totale
	Scarsa	Accettabile	Buona	
Infelice	56.4	119.8	98.8	275.0
Così così	438.1	931.3	767.6	2137.0
Felice	691.5	1469.9	1211.6	3373.0
Totale	1186.0	2521.0	2078.0	5785.0

Poi, applicando la (9.3), si ottiene

$$\chi^2 = \frac{(90-56.4)^2}{56.4} + \frac{(123-119.8)^2}{119.8} + \cdots + \frac{(1406-1211.6)^2}{1211.6} = 164.4.$$

V è dato da

$$V = \sqrt{\frac{164.4}{5785\cdot 2}} = 0.12.$$

■

9.3 Dipendenza in media

In sede di definizione della misura di dipendenza tra X e Y, l'attenzione è stata rivolta ad accertare se e in che misura, al variare delle modalità di una variabile, mutano le distribuzioni di frequenza relativa condizionate dell'altra variabile. Si supponga ora che almeno una variabile, per esempio Y, sia quantitativa. In questa situazione, conviene utilizzare un approccio diverso per confrontare le distribuzioni condizionate della variabile Y: anziché comparare le distribuzioni di frequenza condizionate, sfruttando la maggiore informazione disponibile (dovuta al fatto che la variabile è quantitativa), si possono comparare delle sintesi di queste distribuzioni, come quelle fornite dalla media aritmetica.

ESEMPIO 9.10

Si consideri la distribuzione doppia di frequenza presentata nella Tabella 9.12. In una situazione del genere, ha senso chiedersi come varia la spesa media per consumi alimentari (in euro) al variare dell'ampiezza della famiglia.

Tabella 9.12
Distribuzione di un campione di famiglie secondo il numero di componenti e la spesa per consumi alimentari.

Ampiezza (X)	Spesa per consumi alimentari (Y)				Totale
	115-200	200.1-400	400.1-600	> 600	
1	5	4	0	0	9
2	0	11	3	1	15
3	0	4	7	4	15
4	1	4	4	7	16
5	0	0	0	7	7
Totale	6	23	14	19	62

Soluzione

Assumendo 850 come valore centrale dell'ultima classe, si possono calcolare le medie aritmetiche delle cinque distribuzioni condizionate della spesa per consumi alimentari:

$$\mu_Y(1) = (157.5 \cdot 5 + 300.05 \cdot 4 + 500.05 \cdot 0 + 850 \cdot 0)/9 = 220.86,$$

$$\mu_Y(2) = (157.5 \cdot 0 + 300.05 \cdot 11 + 500.05 \cdot 3 + 850 \cdot 1)/15 = 376.71,$$

$$\mu_Y(3) = (157.5 \cdot 0 + 300.05 \cdot 4 + 500.05 \cdot 7 + 850 \cdot 4)/15 = 540.04,$$

$$\mu_Y(4) = (157.5 \cdot 1 + 300.05 \cdot 4 + 500.05 \cdot 4 + 850 \cdot 7)/16 = 581.74,$$

$$\mu_Y(5) = (157.5 \cdot 0 + 300.05 \cdot 0 + 500.05 \cdot 0 + 850 \cdot 7)/7 = 850.00,$$

dove $\mu_Y(x_i)$ è la media aritmetica della distribuzione condizionata di Y corrispondente alla generica modalità di X. Dall'esame delle medie sopra calcolate, risulta chiaro come la spesa media per consumi alimentari vari in funzione dell'ampiezza della famiglia.

■

In generale, considerando una distribuzione doppia di frequenza come quella esposta nella Tabella 9.1, la media aritmetica della generica distribuzione condizionata

di Y è data da

$$\mu_Y(x_i) = \frac{1}{n_{i0}} \sum_{j=1}^{t} y_j n_{ij}, \quad i = 1, 2, \ldots, s,$$

denominata media condizionata di Y per $X = x_i$.

La media della distribuzione marginale di Y è data da

$$\mu_Y = \frac{1}{n} \sum_{j=1}^{t} y_j n_{0j}.$$

È necessario ora definire il concetto di dipendenza in media.

DEFINIZIONE 9.5

Si dice che la variabile Y dipende in media dalla variabile X se le medie condizionate di Y sono diverse tra loro. Ciò implica che siano diverse almeno due delle medie $\mu_Y(x_1), \mu_Y(x_2), \ldots, \mu_Y(x_s)$.

9.3.1 Misura della dipendenza in media

Come indice di dipendenza in media di Y da X è ragionevole prendere un'appropriata misura di diversità delle medie condizionate $\mu_Y(x_i)$. Considerando che la media della distribuzione marginale di Y è la media ponderata delle s medie condizionate,

$$\mu_Y = \frac{1}{n} \sum_{i=1}^{s} \mu_Y(x_i) n_{i0},$$

la devianza delle medie condizionate può essere scritta come

$$D_S = \sum_{i=1}^{s} [\mu_Y(x_i) - \mu_Y]^2 n_{i0}.$$

D_S è una misura di diversità delle medie condizionate e quindi è un indice adatto per misurare la dipendenza in media. Assume il valore 0 nel caso di indipendenza in media, cioè quando le medie condizionate sono tutte uguali; è tanto maggiore quanto più le medie condizionate sono diverse tra loro. È preferibile tuttavia una versione normalizzata dell'indice che assuma valori nell'intervallo [0, 1]. A questo fine, si fa ricorso alla scomposizione della devianza.

PROPOSIZIONE 9.2

La devianza della distribuzione marginale di Y, D_Y, può essere scritta nella forma

$$\sum_{j=1}^{t} (y_j - \mu_Y)^2 n_{0j} = \sum_{i=1}^{s} [\mu_Y(x_i) - \mu_Y]^2 n_{i0} + \sum_{i=1}^{s} \sum_{j=1}^{t} [y_j - \mu_Y(x_i)]^2 n_{ij}, \qquad (9.5)$$

dove il primo termine a destra dell'uguale è la devianza delle medie condizionate D_S, mentre il secondo, indicato con D_R, è la somma delle devianze delle s distribuzioni condizionate di Y.

La dimostrazione della (9.5) è data nel Paragrafo A.6 dell'Appendice A.

■

La relazione (9.5) può essere posta per brevità nella forma

$$D_Y = D_S + D_R,$$

dove D_Y, D_S e D_R sono dette devianza totale, devianza spiegata[3] e devianza residua. Ciò posto, una misura plausibile della dipendenza in media di Y da X è data dal cosiddetto *rapporto di correlazione*.

DEFINIZIONE 9.6

Il rapporto di correlazione si ottiene rapportando la devianza spiegata alla devianza totale:

$$\eta^2_Y = \frac{\sum_{i=1}^{s}[\mu_Y(x_i) - \mu_Y]^2 n_{i0}}{\sum_{j=1}^{t}(y_j - \mu_Y)^2 n_{0j}} = \frac{D_S}{D_Y}. \qquad (9.6)$$

Poiché, per la (9.5), $D_S = D_Y - D_R$, un'espressione alternativa per η^2_Y è data da

$$\eta^2_Y = 1 - \frac{\sum_{i=1}^{s}\sum_{j=1}^{t}[y_j - \mu_Y(x_i)]^2 n_{ij}}{\sum_{j=1}^{t}(y_j - \mu_Y)^2 n_{0j}} = 1 - \frac{D_R}{D_Y}.$$

Considerando che al denominatore della (9.6) è posto il massimo di D_S, il rapporto di correlazione varia tra 0 e 1. Il minimo è assunto quando D_S è 0, ossia nel caso di indipendenza in media. Il massimo è raggiunto quando la devianza D_R è nulla, ossia nel caso di dipendenza perfetta di Y da X (si veda la Definizione 9.3). Inoltre, l'indice aumenta al crescere della variabilità delle medie condizionate, ovvero al diminuire della variabilità delle distribuzioni condizionate.

ESEMPIO 9.11

Si calcoli il rapporto di correlazione per la distribuzione esaminata nell'Esempio 9.10.

Soluzione

Le medie condizionate sono state già calcolate nell'Esempio 9.10. La media della distribuzione marginale di Y è

$$\mu_Y = (157.5 \cdot 6 + 300.05 \cdot 23 + 500.05 \cdot 14 + 850 \cdot 19)/62 = 499.95.$$

Le devianze (spiegata e totale) sono:

$$D_S = (220.86 - 499.95)^2 \cdot 9 + (376.71 - 499.95)^2 \cdot 15 + \cdots + (850 - 499.95)^2 \cdot 7 = 1\,917\,729.32;$$

$$D_Y = (157.5 - 499.95)^2 \cdot 6 + (300.05 - 499.95)^2 \cdot 23 + \cdots + (850 - 499.95)^2 \cdot 19 = 3\,950\,877.43.$$

Ne segue il rapporto di correlazione

$$\eta^2_Y = \frac{1\,917\,729.32}{3\,950\,877.43} = 0.49.$$

■

[3] Alla luce della (9.5), D_S rappresenta la parte di variabilità totale di Y (misurata da D_Y) dovuta alla variabilità delle medie condizionate.

Riassunto

Oggetto di questo capitolo è lo studio della relazione di dipendenza tra le variabili X e Y di una distribuzione doppia di frequenza. Vengono innanzitutto definite le distribuzioni marginali e quelle condizionate, che compongono la distribuzione doppia di frequenza.

Ai fini della misura della dipendenza, una nozione cruciale è quella di indipendenza tra le variabili X e Y, intesa come invarianza delle distribuzioni di frequenza relativa condizionate di una variabile al variare delle modalità dell'altra. In base a ciò, viene introdotta una misura di dipendenza, Φ, che esprime la distanza tra la tabella doppia di frequenza osservata e quella di indipendenza, che è la tabella a doppia entrata che si sarebbe osservata nel caso di indipendenza tra X e Y.

L'ultima parte del capitolo è dedicata allo studio della dipendenza in media, che può essere definita solamente nel caso in cui almeno una delle due variabili è quantitativa. Assumendo che Y sia quantitativa, si tratta di stabilire se e in che misura le medie delle distribuzioni condizionate di Y variano al variare delle modalità di X. Come misura della dipendenza in media, viene introdotto un indice, denominato rapporto di correlazione, che misura il grado di diversità tra le medie condizionate di Y.

Parole chiave

contingenze, 201
devianza residua, 206
devianza spiegata, 206
devianza totale, 206
dipendenza in media, 205
dipendenza perfetta, 201
distribuzioni condizionate, 197
distribuzioni di frequenza relativa condizionate, 197
distribuzioni marginali, 197
frequenze congiunte, 197
medie condizionate, 205
rapporto di correlazione, 206
scomposizione della devianza, 205
tabella di indipendenza, 199
variabili indipendenti, 198

Commenti e curiosità

1. Analisi della dipendenza, un tema problematico

Mentre le medie, gli indici di variabilità e gli indici di forma si configurano come indicatori con funzione di descrizione dei dati di facile intelligibilità, un aspetto più sfuggente all'intuizione entra in campo con il concetto di dipendenza tra due variabili.

Per semplicità, si pensi alla distribuzione degli studenti che hanno sostenuto l'esame di statistica per sesso e votazione conseguita. In una situazione del genere, può essere interessante studiare la dipendenza tra il voto e il sesso – nel senso di verificare se le distribuzioni dei voti dei maschi e delle femmine sono diverse e di quanto – mediante uno degli indici introdotti in questo capitolo. Questi indici assumeranno valori positivi e con ciò segnaleranno, sostanzialmente, che la distribuzione percentuale secondo il voto delle femmine è diversa da quella dei maschi.

Può accadere che le femmine presentino frequenze percentuali più alte rispetto ai maschi nelle classi di voto più elevate e quindi presentino anche un voto medio superiore. Ovviamente, si tratta di un'informazione che può avere qualche interesse se riferita alla specifica situazione, ma che non autorizza ad affermare che le donne riescono meglio degli uomini nello studio della statistica. L'impossibilità di generalizzare il risultato discende dal fatto che non sono stati "controllati" i fattori che insieme al sesso concorrono a determinare il profitto in statistica. Non si sa, per esempio, se i due gruppi (femmine e maschi) si differenziano o meno rispetto al lavoro, alla scuola superiore frequentata, alla classe sociale di appartenenza, ecc.

Ci si può chiedere allora perché l'argomento dell'analisi della dipendenza sia stato inserito in questa parte del testo dedicata alle statistica descrittiva. La risposta è che i concetti e le misure della dipendenza illustrati in questo capitolo hanno comunque un ruolo: quello di "analisi esplorativa", propedeutica, qualora se ne ravvisi l'utilità, a studi più affinati.

2. Il paradosso di Simpson

Quando entrambe le variabili di una distribuzione doppia presentano due sole modalità, la dipendenza si presta a una interpretazione più immediata, specialmente quando le modalità di Y sono ordinate. Si supponga, senza perdita di generalità, che le due modalità di Y si possano interpretare come insuccesso (–) e successo (+). La distribuzione doppia assume allora la struttura illustrata nella Tabella 9.13.

Tabella 9.13
Schema di distribuzione doppia 2 x 2.

Variabile (X)	Variabile (Y)		Totale
	–	+	
x_1	n_{11}	n_{12}	n_{10}
x_2	n_{21}	n_{22}	n_{20}
Totale	n_{01}	n_{02}	n

In questa situazione, esiste dipendenza tra X e Y se e solo se le frequenze relative di successo, $f_{12} = n_{12}/n_{10}$ e $f_{22} = n_{22}/n_{20}$, sono diverse: se la prima è maggiore della seconda, si dirà che la modalità x_1 produce esiti positivi più di quanto avvenga con la modalità x_2; viceversa nel caso opposto. Si illustra ora un fenomeno apparentemente anomalo che può presentarsi quando due o più tabelle a doppia entrata vengono aggregate, fenomeno noto come paradosso di Simpson, già esaminato in un diverso contesto nel Capitolo 4.

In un'ipotetica classe di un certo liceo (classe A), gli alunni sono classificati per sesso e voto in matematica (vedi Tabella 9.14), distinto in due categorie, < 6 (–) e ≥ 6 (+).

Tabella 9.14
Distribuzione per sesso e voto in matematica nella classe A.

Sesso (X)	Voto (Y)		Totale
	–	+	
F	8	9	17
M	6	6	12
Totale	14	15	29

Dall'esame della tabella, emerge un legame associativo tra voto e sesso, perché la frequenza relativa delle femmine con la sufficienza in matematica è 0.53 contro lo 0.50 dei maschi. In un'altra classe dello stesso liceo (classe B), l'analoga distribuzione doppia è:

Tabella 9.15
Distribuzione per sesso e voto in matematica nella classe B.

Sesso (X)	Voto (Y)		Totale
	–	+	
F	2	4	6
M	7	12	19
Totale	9	16	25

Anche in questa distribuzione, la frequenza relativa delle femmine con la sufficienza in matematica è superiore a quella dei maschi: 0.67 contro 0.63.

Ora le due distribuzioni vengono aggregate in un'unica tabella (Tabella 9.16). Rispetto alle due distribuzioni precedenti, la relazione tra le due variabili è ora invertita: la frequenza relativa dei maschi con la sufficienza (0.58) supera quella delle femmine (0.57).

Nel caso esemplificato, la relazione tra sesso e voto si presenta con caratteristiche diverse nelle due classi: nella prima, la percentuale di studenti con sufficienza in matematica è attorno al 50% (53% per le femmine e 50% per i maschi); nella seconda invece tale percentuale supera ampiamente il 60% (67% per

Sesso (X)	Voto (Y)		Totale
	–	+	
F	10	13	23
M	13	18	31
Totale	23	31	54

Tabella 9.16
Distribuzione per sesso e voto in matematica nell'unione delle classi A e B.

le femmine e 63% per i maschi). Questo elemento e il fatto che nella prima classe sono più numerose le femmine, mentre nella seconda prevalgono i maschi, fanno sì che l'aggregazione delle due tabelle produca il fenomeno segnalato.

Si noti che con l'aggregazione di due tabelle, non si tiene conto degli effetti di una terza variabile – nota come variabile confondente – le cui modalità sono associate alle singole tabelle. Nell'esempio appena esposto, la terza variabile è la classe (le cui modalità sono A e B).

Esercizi

9.1* Si dimostri che le frequenze della tabella di indipendenza date nella Definizione 9.1 sono tali che le distribuzioni di frequenza relativa condizionate della variabile Y sono uguali. (*Suggerimento*: si consideri la generica riga della tabella di indipendenza, le cui frequenze sono:

$$\frac{n_{i0}n_{01}}{n}, \frac{n_{i0}n_{02}}{n}, \ldots, \frac{n_{i0}n_{0t}}{n};$$

si divida ciascuna frequenza per il totale di riga n_{i0} e si osservi che le frequenze relative che ne risultano non dipendono da i.)

9.2 Si consideri la distribuzione doppia di frequenza riportata nell'Esercizio 2.12.

a) Si determini la distribuzione di frequenza percentuale condizionata della variabile colore degli occhi associata alla modalità scuro della variabile colore dei capelli.

b) Si costruisca la tabella di indipendenza.

c) Si calcoli l'indice V di Cramer e, sulla base del valore ottenuto, si commenti l'entità della dipendenza tra i due caratteri.

9.3 La seguente tabella di contingenza è stata ottenuta dal *file* "Diamanti":

Colore	Lucentezza					Totale
	IF	VS1	VS2	VVS1	VVS2	
D	1	7	2	2	4	16
E	5	18	7	5	9	44
F	8	25	14	18	17	82
G	11	12	9	12	21	65
H	11	11	14	10	15	61
I	8	8	7	5	12	40
Totale	44	81	53	52	78	308

a) Si costruisca la distribuzione di frequenza percentuale condizionata della variabile colore associata alla modalità IF della variabile lucentezza. Si costruisca poi la distribuzione marginale percentuale della variabile colore. Confrontando le due distribuzioni, si stabilisca se vi è dipendenza tra le variabili colore e lucentezza.

b) Si calcoli l'indice V di Cramer.

9.4 La seguente tabella è stata ottenuta dalla US *General Social Survey* (sda.bekeley.edu/gss) del 2014:

Reddito	Felicità			Totale
	Non troppo felice	Abbastanza felice	Molto felice	
Sotto la media	194	474	174	842
Vicino alla media	101	620	394	1115
Sopra la media	42	294	210	546
Totale	337	1388	778	2503

a) Si costruiscano le distribuzioni di frequenza percentuale condizionate della variabile felicità e in base a quanto ottenuto si esprima un giudizio sulla dipendenza tra le due variabili.

b) Si calcoli l'indice χ^2.

9.5 Si consideri la tabella di contingenza dell'Esercizio 2.13.

a) Si determinino le cinque distribuzioni di frequenza percentuale condizionate della variabile dimensione demografica.

b) Si calcoli l'indice χ^2 e si commenti il risultato.

9.6 Abitazioni in Italia secondo la dimensione e il titolo di godimento:

Dimensione	Titolo di possesso		Totale
	Proprietà	Affitto	
< 30	45	51	96
30-39	189	173	362
40-49	492	404	896
50-59	759	465	1 224
60-79	2 841	1249	4 090
80-99	4 060	1127	5 187
100-119	3 032	502	3 534
120-149	2 070	218	2 288
≥ 150	1 965	139	2 104
Totale	15 453	4328	19 781

Si misuri la dipendenza in media della dimensione dal titolo di godimento dell'abitazione, assumendo che le medie della prima e dell'ultima classe siano 25 e 165.

9.7 Si consideri la distribuzione doppia di frequenza dell'Esercizio 2.14.

a) Si calcoli l'indice V di Cramer e si commenti il risultato.

b) Si determinino la devianza spiegata e la devianza totale della variabile lunghezza del dito medio.

c) Si calcoli il rapporto di correlazione con riferimento ai risultati ottenuti al punto precedente.

9.8 Automobili in Italia per tipo di carburante e cilindrata (Fonte: dati ACI 2017; frequenze espresse in migliaia):

Carburante	Cilindrata					Totale
	≤ 1 200	1 201-1 600	1 601-2 000	2 001-3 000	> 3 000	
Benzina	8841	1 586	1098	163	153	11 841
Benzina o metano/GPL	668	2 272	250	31	15	3 236
Diesel	170	7 977	6673	2029	48	16 897
Totale	9679	11 835	8021	2223	216	31 974

a) Si determinino le medie delle distribuzioni condizionate della cilindrata. In base ai risultati ottenuti, si dica se la dimensione del motore dipende in media dal tipo di carburante. Si utilizzi 800 come limite inferiore della prima classe e 3500 come limite superiore dell'ultima classe.

b) Si determini il rapporto di correlazione e si commenti il risultato.

CAPITOLO

10

Regressione lineare semplice

Sommario del capitolo

10.1 Introduzione
10.2 Regressione lineare semplice
10.3 Adattamento della retta di regressione ai punti osservati
Riassunto
Commenti e curiosità
Esercizi

10.1 Introduzione

In matematica, una variabile si dice funzione di un'altra variabile se a ogni valore della seconda corrisponde, in base a una specifica relazione algebrica, un unico valore della prima. È una funzione, ad esempio, il legame che associa l'area, A, al raggio, r, di un cerchio: $A = \pi r^2$.

Nell'ambito dei fenomeni naturali, economici e sociali – campo di applicazione della statistica – le relazioni tra variabili non sono descrivibili perfettamente con equazioni matematiche. Per esempio, a determinare il livello dei consumi delle famiglie concorrono, accanto al reddito familiare, altri molteplici fattori, quali l'ampiezza della famiglia, l'età dei componenti, lo stile di vita ecc., cosicché non esiste un legame funzionale perfetto tra consumi e redditi: è ragionevole pensare infatti che famiglie con lo stesso reddito abbiano livelli diversi di spesa. Analoghe considerazioni valgono per coppie di variabili come statura e peso delle persone, costo di riscaldamento e superficie delle abitazioni, consumo di carburante e cilindrata delle automobili e così via.

In generale, date due variabili quantitative, X e Y, sia X la variabile indipendente e Y la variabile dipendente o risposta. Si dirà che tra le stesse intercorre una relazione statistica se le variazioni di X aiutano, in qualche misura, a "spiegare" o a predire le variazioni di Y.

Lo studio della relazione statistica tra X e Y richiede che si disponga di una serie doppia di osservazioni delle due variabili,

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

Queste coppie ordinate verranno chiamate per brevità punti osservati, alludendo al relativo grafico di dispersione.

ESEMPIO 10.1

Si consideri la serie doppia di osservazioni riportata nella Tabella 10.1 (tratta dalla matrice dei dati di cui all'Esercizio 1.6) e si costruisca il grafico di dispersione. I dati si riferiscono alle circonferenze del polso e del collo rilevate su 38 individui di sesso maschile.

Soluzione

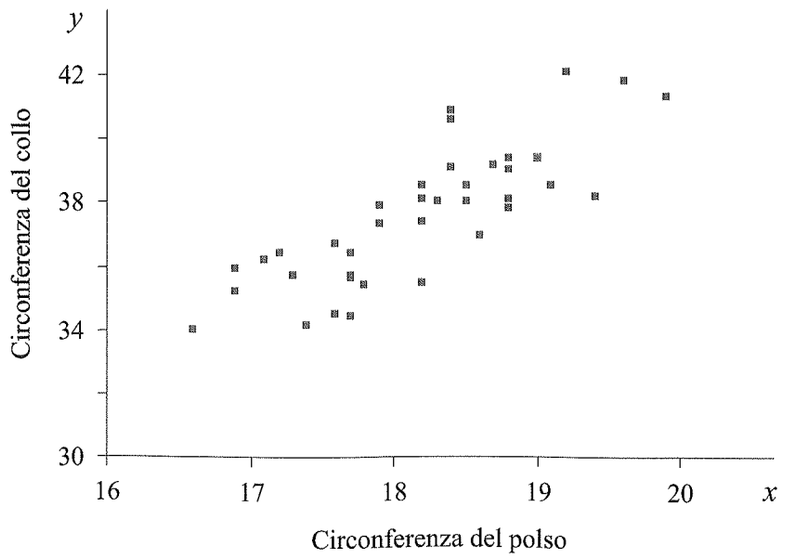
Il grafico di dispersione presentato nella Figura 10.1 mostra come al crescere della circonferenza del polso, la circonferenza del collo tende a crescere. Una tale relazione non può essere tradotta nell'equazione

di una funzione matematica, ma una funzione matematica, come ad esempio la retta, può essere impiegata utilmente per approssimare l'andamento dei dati osservati.

Tabella 10.1
Circonferenza del polso e del collo in 38 maschi adulti.

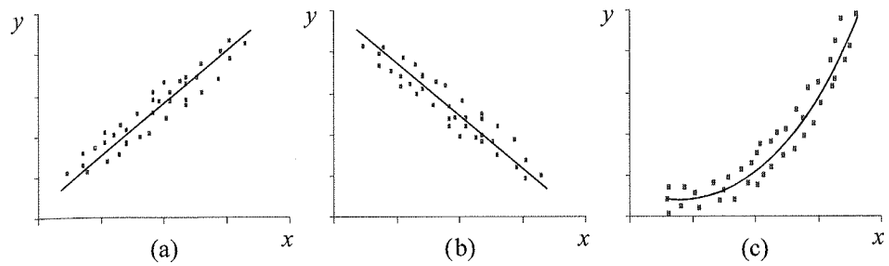
N.	Circonferenza del polso	Circonferenza del collo	N.	Circonferenza del polso	Circonferenza del collo
1	18.2	38.5	20	19.6	41.8
2	16.6	34.0	21	19.0	39.4
3	17.1	36.2	22	17.7	35.7
4	19.2	42.1	23	17.2	36.4
5	18.2	35.5	24	18.8	38.1
6	18.3	38.0	25	17.4	34.1
7	17.7	34.4	26	17.9	37.9
8	18.8	39.0	27	19.4	38.2
9	17.3	35.7	28	18.5	38.0
10	18.7	39.2	29	18.4	39.1
11	18.8	37.8	30	19.9	41.3
12	18.2	38.1	31	17.6	34.5
13	18.4	40.9	32	17.7	35.6
14	16.9	35.2	33	19.1	38.5
15	18.2	37.4	34	18.6	37.0
16	17.7	36.4	35	17.6	36.7
17	18.5	38.5	36	17.9	37.3
18	18.4	40.6	37	18.8	39.4
19	17.8	35.4	38	16.9	35.9

Figura 10.1
Grafico di dispersione dei dati della Tabella 10.1.



L'analisi della relazione statistica tra le variabili X e Y inizia con la scelta della funzione matematica, $y = f(x)$, caratterizzata da pochi parametri, che si reputa adatta ad approssimare l'andamento dei punti osservati. Questa funzione è chiamata modello di regressione o funzione interpolatrice. Potrebbe essere la retta di equazione $y = \beta_0 + \beta_1 x$, la funzione esponenziale $y = \beta_0 x^{\beta_1}$, la parabola $y = \beta_0 + \beta_1 x + \beta_2 x^2$ ecc. I parametri della funzione sono le costanti che appaiono nella relazione algebrica che lega Y a X: ad esempio, β_0 e β_1 nelle prime due funzioni, β_0, β_1 e β_2 nella terza.

La scelta della funzione interpolatrice è basata sull'esame del grafico di dispersione. Nella Figura 10.2, vengono presentati i grafici di sei insiemi di dati artificiali in cui la disposizione dei punti suggerisce specifiche funzioni interpolatrici: la retta in (a) e in (b), la funzione esponenziale in (c) e così via.



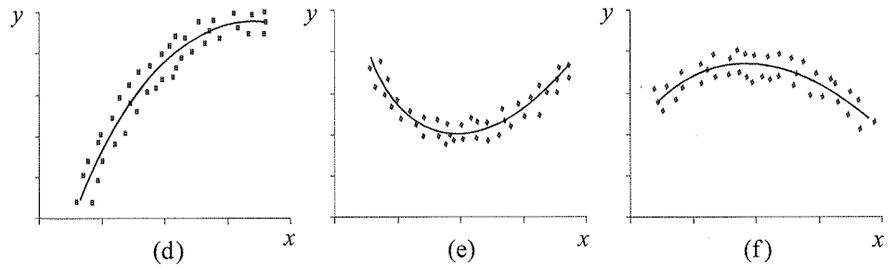


Figura 10.2
Grafici di dispersione e funzioni interpolatrici.

Alla scelta della funzione fanno seguito l'assegnazione dei valori ai parametri e la misura del grado di prossimità della funzione interpolatrice ai dati osservati. In questo capitolo, si farà esclusivo riferimento alla regressione lineare semplice, cioè al caso in cui la relazione intercorrente tra X e Y viene descritta mediante una retta. L'aggettivo semplice viene usato per rimarcare che la variabile dipendente è messa in relazione con una sola variabile indipendente, anziché con due o più, come avviene nella regressione lineare multipla.

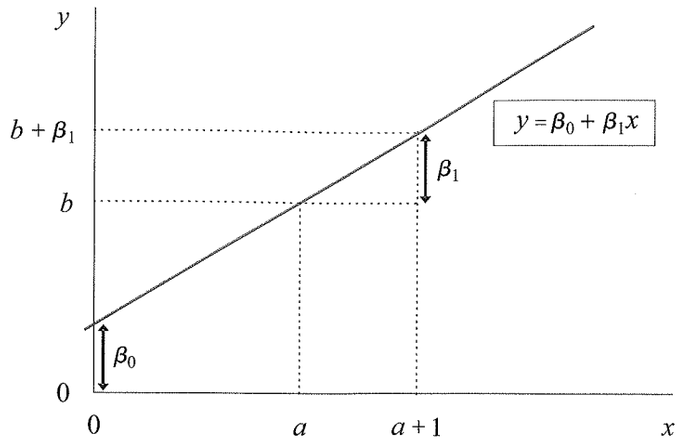
10.2 Regressione lineare semplice

Il modello di regressione lineare semplice è rappresentato dall'equazione

$$y = \beta_0 + \beta_1 x, \tag{10.1}$$

dove β_0 e β_1 sono i parametri. In particolare β_0 è l'intercetta (ordinata all'origine) e β_1 è il coefficiente angolare della retta, come è illustrato nella Figura 10.3.

Figura 10.3
Retta e relativi parametri.



Il problema fondamentale è quello di assegnare ai parametri β_0 e β_1 valori tali che la retta approssimi nel miglior modo possibile i punti osservati. In altri termini, si tratta di individuare, tra le infinite rette del piano, quella che presenta il migliore adattamento a tali punti.

Per avvicinarsi in modo intuitivo alla soluzione del problema, si immagini di assegnare, in qualche modo, a β_0 e β_1 nella (10.1) la coppia di numeri b_0 e b_1. Questa scelta consente di calcolare i cosiddetti valori teorici di Y,

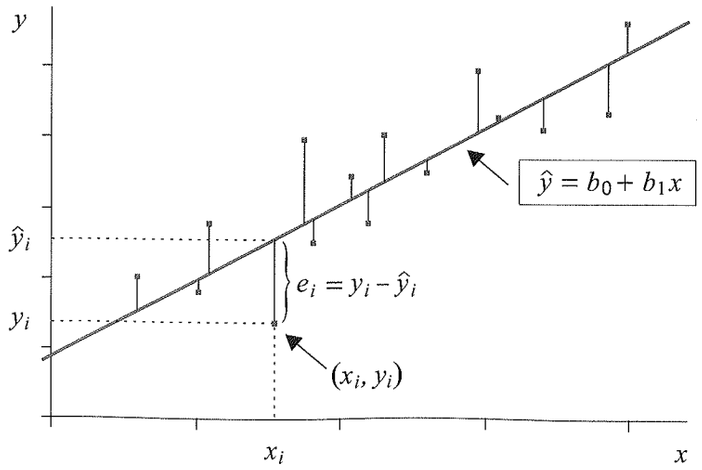
$$\hat{y}_i = b_0 + b_1 x_i, \; i = 1, 2, \ldots, n, \tag{10.2}$$

cioè i valori di Y che si ottengono dalla retta prescelta (avente l'intercetta e il coefficiente angolare assegnati), detta retta interpolatrice. La bontà di questa scelta può essere valutata confrontando valori teorici e valori osservati di Y tramite le differenze:

$$e_1 = y_1 - \hat{y}_1, e_2 = y_2 - \hat{y}_2, \ldots, e_n = y_n - \hat{y}_n.$$

Tali quantità, denominate errori di approssimazione o errori di predizione o residui, indicano le deviazioni verticali dei punti osservati dalla retta interpolatrice (si veda la Figura 10.4).

Figura 10.4
Retta interpolatrice e residui.



I residui dipendono dalla posizione che la retta interpolatrice ha nel piano, posizione determinata da b_0 e b_1; in altri termini, variano in funzione della coppia di numeri b_0 e b_1 assegnati a β_0 e β_1.

Una misura complessiva degli errori di approssimazione, che si compiono approssimando i valori effettivi di Y con quelli teorici, è data dalla somma dei quadrati dei residui:

$$S_q = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2.$$

Ebbene, è ragionevole concludere che la scelta migliore dell'intercetta e del coefficiente angolare è quella che rende minima la quantità S_q definita sopra. Se ne deduce la retta interpolatrice dei minimi quadrati, detta retta di regressione di Y su X.

PROPOSIZIONE 10.1

Data la serie doppia di osservazioni delle variabili X e Y, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, i valori dei parametri del modello (10.1) che rendono minima la somma dei quadrati dei residui,

$$S_q = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2,$$

sono dati da:

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \mu_X)(y_i - \mu_Y)}{\sum_{i=1}^{n} (x_i - \mu_X)^2}, \quad b_0 = \mu_Y - b_1 \mu_X, \tag{10.3}$$

dove μ_X e μ_Y sono le medie aritmetiche di X e di Y.

La dimostrazione delle (10.3) è contenuta nel Paragrafo A.7 dell'Appendice A. ■

Il procedimento con cui si sono determinati i parametri della retta di regressione va sotto il nome di metodo dei minimi quadrati.

Il numeratore della prima delle (10.3) è la codevianza,[1] denotata con C_{XY}, mentre il denominatore è la devianza di X, indicata con D_X. Con queste notazioni,

$$b_1 = \frac{C_{XY}}{D_X} = \frac{\sigma_{XY}}{\sigma_X^2},$$

dove σ_{XY} è la covarianza di X e Y, mentre σ_X^2 è la varianza di X. La seconda espressione di b_1 si deduce dalla prima dividendo per n numeratore e denominatore.

Calcolati b_0 e b_1 con le (10.3), è possibile scrivere l'equazione della retta di regressione di Y su X:

$$\hat{y} = b_0 + b_1 x. \tag{10.4}$$

Con il metodo dei minimi quadrati si assegnano all'intercetta e al coefficiente angolare della retta interpolatrice i valori per cui è minima la somma dei quadrati dei residui.

[1] La codevianza può anche essere scritta nella forma

$$C_{XY} = \sum_{i=1}^{n} x_i y_i - n \mu_X \mu_Y$$

(si veda l'Esercizio 10.2*).

Il coefficiente di regressione b_1 può essere così interpretato. Poiché la retta di regressione "passa tra i punti osservati", b_1 misura la variazione media della variabile dipendente a fronte di un aumento unitario della variabile indipendente.

La retta di regressione ha due proprietà interessanti: (a) passa per il baricentro dei dati, cioè per il punto di coordinate (μ_X, μ_Y); (b) la somma dei residui è nulla:

$$\sum_{i=1}^{n} e_i = 0.$$

La proprietà (a) può essere verificata immediatamente. Infatti, ponendo nella (10.4) $b_0 = \mu_Y - b_1\mu_X$ e $x = \mu_X$, si ricava $\hat{y} = \mu_Y$.

La dimostrazione della proprietà (b) richiede un po' di algebra (si veda il Paragrafo A.7 dell'Appendice A).

ESEMPIO 10.2

Riprendendo i dati dell'Esempio 10.1, si determinino i parametri della retta di regressione della circonferenza del collo sulla circonferenza del polso.

Soluzione

È conveniente organizzare i calcoli come indicato nella Tabella 10.2. Le medie aritmetiche di X e di Y sono: $\mu_X = 18.18$ e $\mu_Y = 37.57$.

Tabella 10.2
Calcolo dei parametri della retta di regressione di Y su X per i dati della Tabella 10.1.

x_i	y_i	$x_i - \mu_X$	$y_i - \mu_Y$	$(x_i - \mu_X)^2$	$(x_i - \mu_X)(y_i - \mu_Y)$
18.2	38.5	0.02	0.93	0.0004	0.0186
16.6	34.0	−1.58	−3.57	2.4964	5.6406
17.1	36.2	−1.08	−1.37	1.1664	1.4796
⋮	⋮	⋮	⋮	⋮	⋮
18.8	39.4	0.62	1.83	0.3844	1.1346
16.9	35.9	−1.28	−1.67	1.6384	2.1376
691.0	1427.8	–	–	22.5512	50.2848

Poiché $C_{XY} = 50.2848$ e $D_X = 22.5512$, l'applicazione delle (10.3) dà:

$$b_1 = \frac{50.2848}{22.5512} = 2.23, \quad b_0 = 37.57 - 2.23 \cdot 18.18 = -2.97.$$

L'equazione della retta di regressione è allora

$$\hat{y} = -2.97 + 2.23x,$$

dove il coefficiente angolare 2.23 indica che la circonferenza del collo aumenta mediamente di 2.23 centimetri a fronte di un aumento della circonferenza del polso di un centimetro o, equivalentemente, che a una differenza di un centimetro nella circonferenza del polso corrisponde mediamente una differenza di 2.23 centimetri nella circonferenza del collo.

La retta di regressione trovata è rappresentata nella Figura 10.5, insieme al grafico di dispersione. Come si vede, presenta una inclinazione positiva, "riproducendo" così l'andamento crescente dei punti osservati.

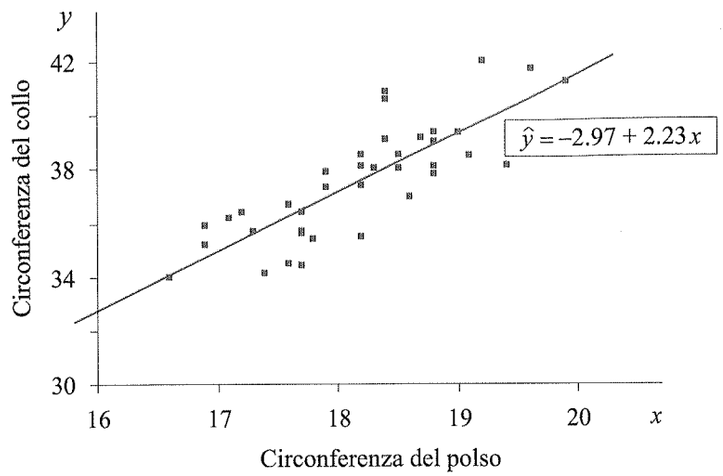


Figura 10.5
Grafico di dispersione dei dati della Tabella 10.1.

Serie storiche

Data una serie storica, si può essere interessati a descrivere in modo approssimato l'andamento temporale della variabile dipendente con una retta. In questa situazione, a fronte dei valori della variabile indipendente, $x_1, x_2, \ldots, x_n$, che sono i tempi (istanti o intervalli temporali espressi da numeri in progressione aritmetica), vi sono i valori corrispondenti della variabile dipendente, $y_1, y_2, \ldots, y_n$.

Per la determinazione di β_0 e β_1, si possono usare le stesse formule illustrate in precedenza. L'unica particolarità è il ricorso che si fa quasi sempre alla traslazione dell'asse delle ascisse assumendo come origine il tempo immediatamente precedente al primo o il tempo intermedio. Tutto ciò allo scopo di semplificare i calcoli.

ESEMPIO 10.3

Si determini la retta interpolatrice dei minimi quadrati per la serie storica riportata nella Tabella 10.3:

Tabella 10.3
Arrivi presso le strutture ricettive turistiche dell'EU-28 (Fonte: Eurostat; valori in migliaia).

Anno	2010	2011	2012	2013	2014	2015	2016
Arrivi	227	243	249	262	273	288	297

Soluzione

I calcoli necessari sono mostrati nella Tabella 10.4.

Tabella 10.4
Calcolo di b_0 e b_1 per la serie storica riportata nella Tabella 10.3.

Anno	x_i	y_i	$x_i - \mu_X$	$y_i - \mu_Y$	$(x_i - \mu_X)^2$	$(x_i - \mu_X)(y_i - \mu_Y)$
2010	1	227	–3	–35.7	9	107.1
2011	2	243	–2	–19.7	4	39.4
2012	3	249	–1	–13.7	1	13.7
2013	4	262	0	–0.7	0	0
2014	5	273	1	10.3	1	10.3
2015	6	288	2	25.3	4	50.6
2016	7	297	3	34.3	9	102.9
Totale	28	1839	–	–	28	324.0

Poiché $\mu_X = 4$ e $\mu_Y = 262.7$, b_1 e b_0 sono uguali a:

$$b_1 = \frac{324}{28} = 11.57, \quad b_0 = 262.7 - 11.57 \cdot 4 = 216.42.$$

La retta di regressione stimata è allora

$$\hat{y} = 216.42 + 11.57x.$$

■

La traslazione dell'asse delle ascisse (come quella effettuata nell'esempio precedente) non altera i risultati dell'analisi.

10.3 Adattamento della retta di regressione ai punti osservati

L'analisi di regressione non termina con la determinazione dei parametri del modello (10.1), ma include come momento importante la verifica dell'idoneità del modello a rappresentare la relazione statistica intercorrente tra le variabili X e Y. A questo fine, è necessario costruire un apposito indice che misuri la bontà dell'adattamento della retta ai punti osservati, la cui costruzione è basata sulla scomposizione della devianza.

PROPOSIZIONE 10.2

Data la serie doppia di osservazioni delle variabili X e Y, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, la devianza di Y può essere così scomposta

$$D_Y = \sum_{i=1}^{n}(y_i - \mu_Y)^2 = \sum_{i=1}^{n}(\hat{y}_i - \mu_Y)^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \qquad (10.5)$$

dove $\hat{y}_i$ sono i valori teorici di Y desunti dalla retta di regressione di Y su X.

La dimostrazione è data nel Paragrafo A.7 dell'Appendice A. ■

Il primo addendo del secondo membro della (10.5) è la devianza spiegata (D_{SL} per brevità). Il secondo è la devianza residua (D_{RL} per brevità).[2] Si noti la stretta analogia tra D_{SL} e D_{RL} e le omologhe quantità D_S e D_R introdotte nel Paragrafo 9.4.1 per la misura della dipendenza in media.

DEFINIZIONE 10.1

Data la serie doppia di osservazioni delle variabili X e Y, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, una misura della bontà dell'adattamento della retta di regressione ai punti osservati, denominata indice di determinazione, è data dal rapporto

[2] Stante la (10.5), D_{SL} rappresenta la parte della variabilità totale dei valori di Y (misurata dalla devianza totale) dovuta alla variabilità dei valori teorici $\hat{y}_i$, mentre D_{RL} è la devianza dei residui $y_i - \hat{y}_i$.

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \mu_Y)^2}{\sum_{i=1}^{n}(y_i - \mu_Y)^2} \qquad (10.6)$$

(in breve, $r^2 = D_{SL}/D_Y$). Questo indice può anche essere scritto nella forma

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \mu_Y)^2} \qquad (10.7)$$

(in breve, $r^2 = 1 - D_{RL}/D_Y$).

Una terza espressione di r^2, equivalente alle due precedenti, è data da (si veda il Paragrafo A.7 dell'Appendice A)

$$r^2 = \frac{[\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)]^2}{\sum_{i=1}^{n}(x_i - \mu_X)^2 \sum_{i=1}^{n}(y_i - \mu_Y)^2}. \qquad (10.8)$$

ESEMPIO 10.4

Si calcoli l'indice di determinazione per i dati dell'Esempio 10.1.

Soluzione

La Tabella 10.5 mostra i calcoli necessari per determinare r^2. Qui, a scopo illustrativo, vengono calcolate entrambe le quantità D_{SL} e D_{RL}. In pratica, solo una di esse è necessaria: la prima se si applica la (10.6), la seconda se si applica la (10.7).

Tabella 10.5
Calcolo di r^2 per la retta di regressione dell'Esempio 10.2.

Unità	x_i	y_i	$\hat{y}_i$	$(\hat{y}_i - \mu_Y)^2$	$(y_i - \hat{y}_i)^2$	$(y_i - \mu_Y)^2$
1	18.2	38.5	37.616	0.0021	0.7815	0.8649
2	16.6	34.0	34.048	12.4045	0.0023	12.7449
3	17.1	36.2	35.163	5.7936	1.0754	1.8769
⋮	⋮	⋮	⋮	⋮	⋮	⋮
37	18.8	39.4	38.954	1.9155	0.1989	3.3489
38	16.9	35.9	34.717	8.1396	1.3995	2.7889
Totale	691.0	1427.8	1428.070	112.1462	56.6295	168.7542

Sulla base dei calcoli esposti, si ottiene

$$r^2 = \frac{112.1462}{168.7542} = 0.66$$

oppure

$$r^2 = 1 - \frac{56.6295}{168.7542} = 0.66.$$

Con la (10.8), si ha

$$r^2 = \frac{50.2848^2}{22.5512 \cdot 168.7542} = 0.66,$$

dove i valori di C_{XY} e D_X sono stati ripresi dall'Esempio 10.2. Il livello dell'indice segnala un buon adattamento della retta ai punti osservati. ■

La (10.8) è particolarmente utile per i calcoli perché, diversamente dalle (10.6) e (10.7), coinvolge le quantità C_{XY}, D_X e D_Y, tra cui le prime due sono già note in quanto richieste per il calcolo del coefficiente angolare della retta di regressione.

La Proposizione 10.3 mostra le principali proprietà dell'indice di determinazione.

PROPOSIZIONE 10.3

L'indice di determinazione r^2 ha le seguenti proprietà:

(a) varia tra 0 e 1;

(b) assume il valore 0 se e solo se $D_{SL} = 0$, cioè se e solo se la retta di regressione è parallela all'asse delle ascisse;

(c) assume il valore 1 se e solo se $D_{RL} = 0$, circostanza che si verifica se e solo se tutti i punti osservati giacciono su una retta.

La prima proprietà deriva dalla (10.5), per la quale $0 \leq D_{SL} \leq D_Y$.

Quanto alla seconda proprietà, dalla (10.6) si desume che r^2 è nullo se e solo se la devianza spiegata è nulla, cioè se e solo se $\hat{y}_i = \mu_Y$, ciò si verifica quando la retta di regressione ha coefficiente angolare nullo (retta parallela all'asse delle ascisse).

La terza proprietà, meno immediata, viene dimostrata nel Paragrafo A.7 dell'Appendice A. ■

L'indice di determinazione rappresenta la proporzione della variabilità della variabile dipendente spiegata dalla retta di regressione. Per esempio, $r^2 = 0,7$ indica che la retta di regressione spiega il 70% della variabilità di Y.

Tali proprietà mostrano che l'indice r^2 assume valori tanto maggiori quanto più i punti osservati sono addensati attorno a una retta, ossia quanto più forte è la relazione di dipendenza lineare di Y da X. Per questo r^2 è un indice che dà la misura dell'adeguatezza della retta a esprimere la relazione statistica che lega la variabile dipendente alla variabile indipendente: valori dell'indice vicini all'unità segnalano un forte adattamento della retta ai punti osservati; valori dell'indice vicini allo zero indicano una totale inidoneità della retta a esprimere la relazione statistica tra i due fenomeni; valori intermedi segnalano ovviamente situazioni intermedie.

Poiché D_Y e D_{SL} misurano la variabilità di Y e quella dei valori teorici $\hat{y}_i$ dedotti dalla retta di regressione stimata, il rapporto $r^2 = D_{SL}/D_Y$ esprime quanta parte della variabilità di Y è "spiegata" dalla retta di regressione. In altre parole, fatta 100 la variabilità di Y, $r^2 \cdot 100$ è la percentuale di questa variabilità attribuibile alla relazione lineare tra Y e X.

Errore medio di predizione

In base alla (10.7), la devianza residua può essere scritta come funzione di r^2:

$$D_{RL} = D_Y(1 - r^2),$$

da cui, dividendo entrambi i membri per n ed estraendo la radice quadrata, si ottiene l'errore medio di predizione:

$$\sigma_{RL} = \sigma_Y \sqrt{1 - r^2}.$$

L'indice dà la misura dell'errore medio che si commette sostituendo ai valori osservati di Y i valori teorici dedotti dalla retta di regressione. Come si vede, l'errore medio di predizione è tanto minore quanto maggiore è r^2. È nullo quando $r^2 = 1$, ossia quando la retta di regressione si adatta perfettamente ai dati osservati. È uguale a σ_Y quando $r^2 = 0$, cioè quando la retta di regressione è parallela all'asse delle ascisse, nel qual caso non è di alcuna utilità ai fini della predizione di Y.

ESEMPIO 10.5

Si calcoli l'errore medio di predizione per i dati dell'Esempio 10.1.

Soluzione

L'errore medio di predizione è dato da

$$\sigma_{RL} = \sqrt{\frac{168.7542}{38}(1 - 0.66)} = 1.23.$$

Questo è l'errore medio che si commette se la circonferenza del collo viene approssimata a partire dalla circonferenza del polso, tramite l'equazione della retta di regressione

$$\hat{y} = -2.97 + 2.23x.$$

■

Riassunto

La regressione è un argomento centrale della metodologia statistica, soprattutto nell'ambito dell'inferenza. In questo capitolo, viene presentato il modello di regressione lineare semplice come strumento di approssimazione della relazione statistica tra due variabili quantitative, X e Y, in cui X è la variabile indipendente e Y la variabile dipendente.

Data la serie doppia di osservazioni di X e Y, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, il problema è quello di determinare i parametri della retta (l'intercetta e il coefficiente angolare) in modo che questa approssimi al meglio i punti osservati. L'obiettivo è raggiunto applicando il metodo dei minimi quadrati, ossia minimizzando la somma dei quadrati delle distanze verticali dei punti osservati dalla retta.

Il secondo passo dell'analisi di regressione è la valutazione del grado di adattamento della retta ai punti osservati. A questo fine, viene introdotto l'indice di determinazione, che esprime quanta parte della variabilità della variabile dipendente è spiegata dalla relazione di dipendenza lineare di Y da X.

Viene infine definito l'errore medio di predizione, l'indice che esprime l'errore medio che si commette nell'approssimare i valori di Y tramite l'equazione della retta.

Parole chiave

codevianza, 217
coefficiente di regressione, 218
covarianza, 217
devianza residua, 220
devianza spiegata, 220
errori di approssimazione, 216
errori di predizione, 216
errore medio di predizione, 222
funzione interpolatrice, 215
indice di determinazione, 220
metodo dei minimi quadrati, 217
modello di regressione, 215
punti osservati, 213
regressione lineare semplice, 215
relazioni statistiche, 213
residui, 216
retta di regressione, 217
retta interpolatrice, 216
retta interpolatrice dei minimi quadrati, 217
risposta, 213
valori teorici, 216
variabile dipendente, 213
variabile indipendente, 213

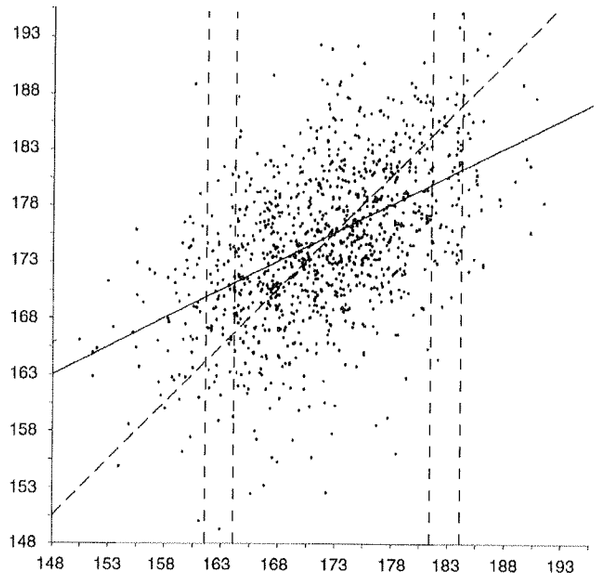
Commenti e curiosità

1. La regressione nella storia

Nella Figura 10.6 è contenuto il grafico di dispersione relativo alle stature di 1078 coppie di padri e figli.[3] Indicando con X la statura dei padri e con Y quella dei figli, le due distribuzioni marginali sono caratterizzate dai seguenti valori di sintesi:

$$\mu_X = 173,\ \sigma_X = 6.75,\ \mu_Y = 175.5,\ \sigma_Y = 6.75.$$

Figura 10.6
Grafico di dispersione per le stature dei padri e dei figli.



Nel grafico, la linea tratteggiata è la retta che passa per il baricentro dei dati, cioè per il punto di coordinate (173, 175.5), e che forma un angolo di 45 gradi con l'asse delle ascisse; mentre la linea continua è la retta di regressione di Y su X definita dall'equazione

$$\hat{y} = 89 + 0.5x.$$

Delle due bande verticali, quella di sinistra comprende le coppie padre-figlio in cui la statura del padre è intorno a 163 centimetri, mentre quella di destra comprende le coppie in cui la statura del padre è intorno a 183 centimetri. Come si vede, le stature dei figli i cui padri sono alti circa 163 cm sono piuttosto disperse e, comunque, in prevalenza al di sopra della linea tratteggiata; la loro media è 170.5. Per contro, le stature dei figli i cui padri sono alti circa 183 cm, anch'esse molto disperse, sono prevalentemente al di sotto della linea tratteggiata, la loro media è 180.5. Questo è il fenomeno che Galton denominò regressione (a cui si è fatto cenno nel Paragrafo 1.2): padri più bassi della media hanno figli più alti; padri più alti della media hanno figli più bassi; le stature dei figli sembrano tendere ad avvicinarsi alla media.

2. Le rette di regressione sono due

Per completezza, vale la pena di segnalare che, oltre alla retta di regressione di Y su X fin qui esaminata, si può definire la retta di regressione di X su Y, scambiando i ruoli delle variabili. Tecnicamente, ossia dal punto di vista formale, ciò è sempre fattibile; tuttavia, in molte situazioni, una variabile si presenta come

[3] Dati usati da K. Pearson in uno studio sui caratteri ereditari.

antecedente logico dell'altra e assume il ruolo naturale di variabile indipendente. Nell'Esercizio 10.9, ad esempio, è senza dubbio l'età la variabile indipendente.

In ogni caso, le formule per la determinazione dei parametri della retta di regressione di X su Y si ottengono immediatamente dalle (10.3). Per esempio, il coefficiente angolare della nuova retta, che qui si denota con b'_1, è dato da

$$b'_1 = \frac{C_{XY}}{D_Y}.$$

La misura della bontà dell'adattamento è la stessa per entrambe le rette di regressione: ciò può essere verificato immediatamente osservando che l'Equazione (10.8) non muta se si scambiano i ruoli delle due variabili.

3. La nullità dell'indice di determinazione non implica assenza di relazione statistica tra le variabili

Se l'indice di determinazione è nullo o molto piccolo, non si deve concludere che non sussiste alcuna relazione tra le variabili studiate: significa soltanto che è da escludere una relazione di tipo lineare (cioè secondo una retta). È sufficiente un controesempio per verificare la validità dell'affermazione. Nella Figura 10.7 è rappresentato un insieme di dati per il quale l'indice di determinazione è nullo, pur essendo molto forte il legame associativo tra X e Y.

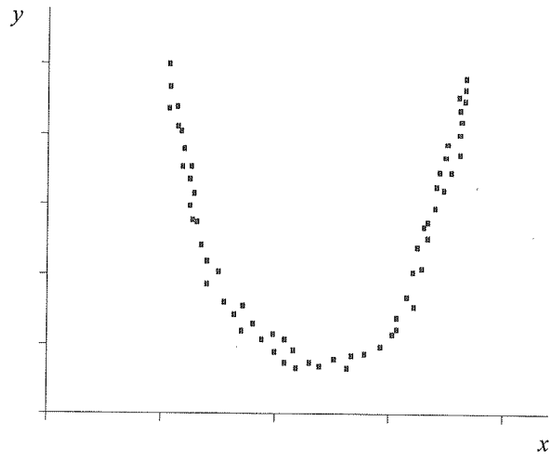


Figura 10.7
Distribuzione doppia in cui $r^2 = 0$ pur essendoci un fortissimo legame associativo tra le due variabili.

4. Un'applicazione della regressione alla misura della volatilità delle quotazioni azionarie: i coefficienti beta

Nella valutazione di un titolo azionario è molto rilevante il parametro della "volatilità", intendendo con questo termine una misura delle variazioni (per esempio settimanali) delle sue quotazioni di borsa. Un modo per misurare la volatilità consiste nel calcolare il coefficiente angolare della retta di regressione delle variazioni percentuali del titolo considerato sulle variazioni percentuali di un indice di mercato, per esempio il Mibtel per il mercato italiano. Il valore assunto dal coefficiente angolare, detto comunemente "coefficiente beta", indica la variazione percentuale che presenta l'azione considerata per una variazione percentuale unitaria dell'indice di mercato. Così, se il coefficiente è pari a 1.6, l'azione presenta una volatilità pari a 1.6 volte quella media di mercato; se, viceversa, il coefficiente assume il valore 0.4, l'azione ha una volatilità pari al 40% di quella media di mercato.

5. Il ruolo dei grafici di dispersione nell'analisi di regressione

La Tabella 10.6 mostra quattro insiemi di dati costruiti artificialmente con le seguenti caratteristiche: danno luogo alla stessa retta di regressione, $\hat{y} = 3 + 0.5x$, e hanno lo stesso indice di determinazione $r^2 = 0.67$ (Anscombe, 1973, pp. 17-21). Sulla base del solo indice di determinazione, si potrebbe trarre la conclusione

Tabella 10.6
Distribuzioni doppie che presentano la stessa retta di regressione.

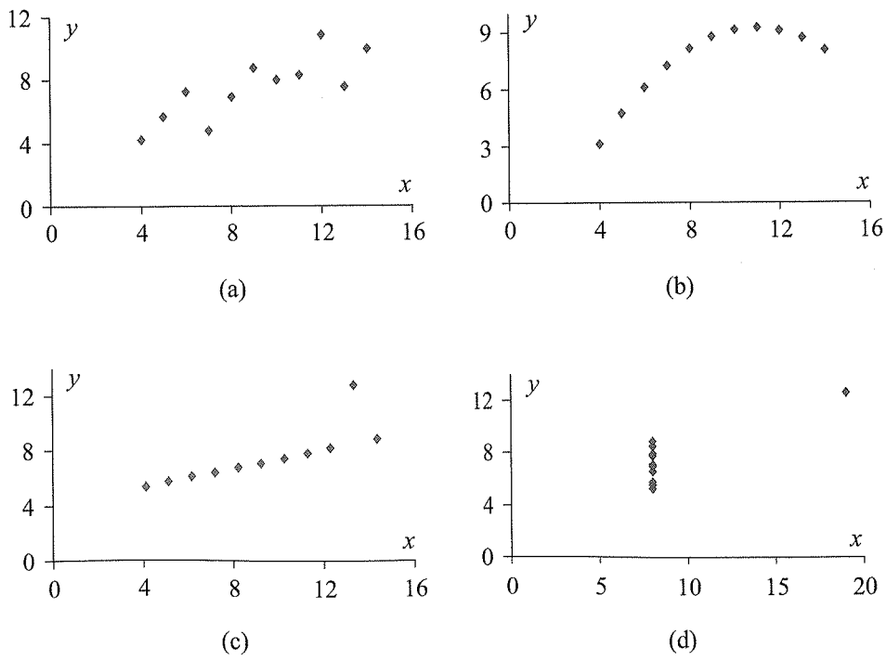
I		II		III		IV	
x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

che in tutti e quattro i casi la retta di regressione è un modello adatto a descrivere la relazione tra le due variabili.

Si considerino ora i grafici di dispersione delle quattro serie doppie mostrate nella Figura 10.8. Dall'esame dei diagrammi, si può constatare che le quattro situazioni sono estremamente diverse e che per tre di esse è addirittura problematico parlare di analisi di regressione. Infatti, il grafico II suggerisce come più appropriato un modello di regressione non lineare; i grafici III e IV mettono in luce invece la presenza di un punto anomalo, che va rimosso nel primo caso prima di procedere all'analisi di regressione, mentre nel secondo la rimozione del punto anomalo preclude la possibilità di determinare la retta di regressione.

L'indicazione da trarre da questo esercizio è che per procedere all'analisi di regressione è necessario appurare se ve ne sono le condizioni osservando il grafico di dispersione dei dati.

Figura 10.8
Grafici di dispersione dei quattro insiemi di dati contenuti nella Tabella 10.6.



Esercizi

10.1 Medici e infermieri professionali (in migliaia) in otto paesi europei nel 2017 (Fonte: Eurostat):

Paese	N. medici X	N. infermieri prof. Y
Austria	45	59
Danimarca	21	56
Francia	209	681
Germania	345	893
Italia	240	338
Polonia	88	197
Portogallo	50	67
Spagna	178	256

a) Si determini la retta di regressione di Y su X.
b) Si calcolino l'indice di determinazione e l'errore medio di predizione.
c) Si verifichi numericamente la validità dell'Equazione (10.5).

10.2* Si verifichi che la codevianza può essere scritta nella forma

$$C_{XY} = \sum_{i=1}^{n} x_i y_i - n\mu_X\mu_Y.$$

(*Suggerimento*: occorre sviluppare il prodotto

$$(x_i - \mu_X)(y_i - \mu_Y) = x_i y_i - x_i\mu_Y - \mu_X y_i + \mu_X\mu_Y,$$

riferire la sommatoria a ogni elemento dello sviluppo e semplificare.)

10.3 I dati della tabella che segue si riferiscono alla distanza stradale (Y) e a quella in linea d'aria (X) tra 20 coppie di luoghi nella zona di Sheffield (Hand *et al.*, 1994):

Distanza in linea d'aria (x_i)	Distanza stradale (y_i)	Distanza in linea d'aria (x_i)	Distanza stradale (y_i)
9.5	10.7	9.8	11.7
5.0	6.5	19.0	25.6
23.0	29.4	14.6	16.3
15.2	17.2	8.3	9.5
11.4	18.4	21.6	28.8
11.8	19.7	26.5	31.2
12.1	16.6	4.8	6.5
22.0	29.0	21.7	25.7
28.2	40.5	18.0	26.5
12.1	14.2	28.0	33.1

a) Si determinino i parametri della retta di regressione della distanza stradale sulla distanza in linea d'aria.

b) Si determini la bontà dell'adattamento della retta ai punti osservati tramite la (10.7).

c) Si verifichi che la retta passa per il baricentro dei dati.

10.4 Si consideri il *file* "Statura e pulsazioni" relativo a un campione di pazienti di un dato ospedale (Hand *et al* ., 1994):

Statura	160	167	162	...	170	175
Pulsazioni	68	80	84	...	84	72

a) Si costruisca il grafico di dispersione e si stabilisca se la retta è idonea a rappresentare la relazione statistica tra le due variabili.

b) Si calcoli l'indice di determinazione tramite la formula (10.8) e si verifichi se il suo valore è in accordo con quanto si percepisce dal grafico.

10.5 Si inseriscano i dati mancanti nella seguente serie doppia di osservazioni in modo che sia 1 l'indice di determinazione r^2 relativo alla regressione di Y su X:

X	5	7	8	10	6	12	15	12
Y		46	49	55	43	61	70	

(*Suggerimento*: si ricordi che l'indice di determinazione è uguale a 1 solo se vi è perfetta relazione lineare tra le due variabili.)

10.6 Si consideri il *file* "Test d'ingresso" concernente il punteggio d'ingresso e la valutazione di fine anno (GPA) di un campione casuale di 114 immatricolati in una piccola università statunitense (dati tratti da Kutner *et al.*, 2005, con modifiche):

Punteggio d'ingresso	21	28	22	...	16	28
Valutazione di fine anno	3.897	3.778	2.540	...	1.86	2.948

a) Si costruisca il grafico di dispersione.

b) Sulla base di tale grafico, si stabilisca se è appropriato procedere all'analisi di regressione su questi dati.

c) Nel caso di risposta affermativa alla domanda del punto precedente, si determini la retta di regressione e si calcoli l'indice di determinazione.

10.7 Con riferimento alla matrice dei dati dell'Esercizio 1.6, si determinino i parametri della retta di regressione del peso sulla statura e si calcoli l'errore medio di predizione.

10.8 Si riprendano i dati relativi all'altezza e al volume di 31 esemplari di cedro nero presentati nell'Esercizio 1.4.

a) Si determini la retta di regressione del volume sull'altezza.

b) Sulla base della retta di regressione trovata nel punto precedente, si calcoli il valore teorico del volume corrispondente a un'altezza di 20.2.

c) Si rappresenti il grafico di dispersione unitamente alla retta di regressione trovata nel punto b).

10.9 Si consideri il *file* "Massa muscolare" che riporta l'età e la massa muscolare di un campione di pazienti di sesso femminile (Kutner *et al.*, 2005):

Età	43	41	47	...	72	76
Massa muscolare	106	106	97	...	70	74

a) Si determini la retta di regressione e la si rappresenti graficamente insieme ai punti osservati.

b) In base alla retta trovata al punto a), si determini la differenza della misura della massa muscolare nelle donne con una differenza di età di due anni.

c) Si trovi, in base alla retta, la massa muscolare per un'età di 65 anni.

d) Si misuri la bontà dell'adattamento della retta ai punti osservati e si commenti il risultato.

10.10 Si considerino le quotazioni giornaliere di borsa delle azioni Ferrari dall'11 Febbraio al 15 Marzo 2019 (contenute nel *file* "Quotazioni"):

Data	11 Feb	12 Feb	13 Feb	...	14 Mar	15 Mar
Quotazioni	109.55	111.1	110.75	...	115.75	116.6

a) Si determinino i parametri della retta interpolatrice.

b) Si calcoli l'indice di determinazione.

Regressione lineare semplice con Excel

Ci sono diversi modi per determinare i parametri della retta di regressione con Excel. Conviene indicare anzitutto le funzioni INTERCETTA( ), PENDENZA( ), RQ( ) e PREVISIONE( ). Le prime tre restituiscono l'intercetta, il coefficiente angolare e l'indice di determinazione. Esse richiedono come argomenti le aree occupate dalle osservazioni di Y e di X. La quarta restituisce il valore teorico corrispondente a un valore assegnato di X e richiede che sia specificato questo valore. Tali funzioni vengono illustrate qui di seguito applicandole ai dati dell'Esercizio 10.3.

E6 =PREVISIONE(40;B2:B21;A2:A21)

	A	B	C	D	E	F
1	**Distanza in linea d'aria**	**Distanza stradale**				
2	9.5	10.7		Intercetta	0.37908	
3	5	6.5		Coefficiente angolare	1.26943	
4	23	29.4		Indice determinazione	0.93913	
5	15.2	17.2				
6	11.4	18.4		Valore teorico x=40	51.1563	
7	11.8	19.7				

Figura 10.9
Funzioni Excel per la regressione lineare.

In alternativa alle funzioni sopra indicate, è disponibile la funzione REGR.LIN( ) che dipende da quattro argomenti: l'area occupata dai valori di Y, l'area occupata dai valori di X, vero o falso a seconda che non si voglia o si voglia che la retta passi per l'origine, vero o falso a seconda che si vogliano o no ottenere alcune statistiche di sintesi.

L'applicazione si svolge in tre passi:

- si seleziona un'area di stampa di 5 righe e 2 colonne;
- nell'apposito spazio (casella bianca), si scrive la funzione specificando gli argomenti (si veda la Figura 10.10);

Figura 10.10
Uso della funzione REGR.LIN( ).

FALSO =REGR.LIN(B2:B21;A2:A21;VERO;VERO)

	A	B	C	D	E	F	G
1	**Distanza in linea d'aria**	**Distanza stradale**					
2	9.5	10.7					
3	5	6.5					
4	23	29.4					
5	15.2	17.2					
6	11.4	18.4					
7	11.8	19.7					
8	12.1	16.6					
9	22	29					

VERO)

- si esegue la funzione premendo simultaneamente i tasti Ctrl, Maiusc e Invio. Si ottengono così i risultati indicati nella Figura 10.11.

Figura 10.11
Risultati dell'applicazione della funzione REGR.LIN( ).

D3 {=REGR.LIN(B2:B21;A2:A21;VERO;VERO)}

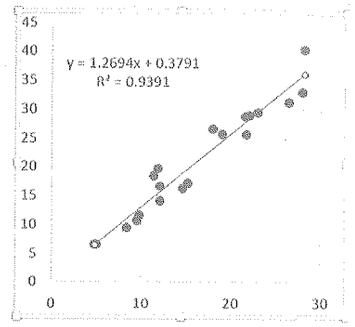
	A	B	C	D	E	F	G
1	**Distanza in linea d'aria**	**Distanza stradale**					
2	9.5	10.7					
3	5	6.5		1.26943	0.37908		
4	23	29.4		0.07617	1.34401		
5	15.2	17.2		0.93913	2.43615		
6	11.4	18.4		277.72838	18.00000		
7	11.8	19.7		1648.26305	106.82645		
8	12.1	16.6					
9	22	29					

I risultati che interessano nella statistica descrittiva sono quelli evidenziati, e sono, dall'alto verso il basso e da sinistra a destra: la pendenza b_1, l'intercetta b_0, il coefficiente di determinazione r^2, la devianza spiegata D_{SL} e la devianza residua D_{RL}.

Excel consente anche di disegnare la retta di regressione con una opzione del grafico di dispersione, nel modo seguente:

- si realizza il grafico di dispersione dal menu Inserisci;
- si preme il tasto destro del mouse sopra uno dei punti del grafico e si seleziona "Aggiungi linea di tendenza";
- dal pannello "Formato linea di tendenza" (si veda la Figura 10.12), si sceglie "lineare" e si selezionano "Visualizza l'equazione sul grafico" e "Visualizza il valore R quadrato sul grafico".

	A	B	C	D	E
1	**Distanza in linea d'aria**	**Distanza stradale**			
2	9.5	10.7			
3	5	6.5			
4	23	29.5			
5	15.2	17.2			
6	11.4	18.4			
7	11.8	19.7			
8	12.1	16.6			
9	22	29			
10	28.2	40.5			
11	12.1	14.2			
12	9.8	11.7			
13	19	25.6			
14	14.6	16.3			
15	8.3	9.5			
16	21.6	28.8			
17	26.5	31.2			
18	4.8	6.5			
19	21.7	25.7			
20	18	26,5			



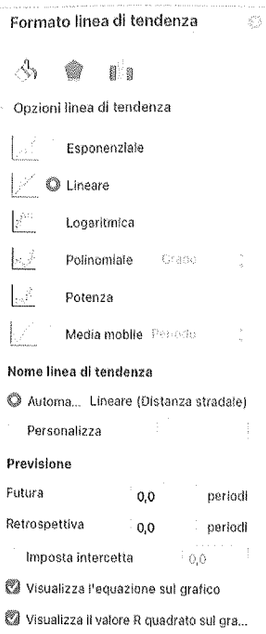


Figura 10.12
Linea di tendenza nel grafico a dispersione.

CAPITOLO

11

Correlazione

Sommario del capitolo

11.1 Introduzione
11.2 Misura della correlazione
Riassunto
Commenti e curiosità
Esercizi

11.1 Introduzione

In statistica, il significato del termine correlazione non differisce da quello che ad esso viene dato nel linguaggio corrente: indica la relazione reciproca tra due fenomeni, precisamente, la relazione tra due variabili quantitative. Ad esempio, si può parlare di correlazione tra statura e peso delle persone, tra prezzo di vendita e costo di produzione di beni e servizi, tra voti riportati in due distinti esami universitari dagli studenti di un corso di laurea.

Lo studio della correlazione tra due variabili ha senso solo se è possibile ipotizzare un legame associativo tra le stesse, alla luce di conoscenze teoriche o di evidenze empiriche.

Come nella regressione, si suppone di disporre di una serie doppia di osservazioni delle variabili quantitative X e Y:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

Conviene partire dalla definizione di *correlazione positiva* e *correlazione negativa*.

DEFINIZIONE 11.1

Data una serie doppia di osservazioni delle variabili quantitative X e Y, le due variabili presentano correlazione positiva o negativa a seconda che al crescere delle modalità dell'una le modalità dell'altra tendano a crescere o a decrescere.

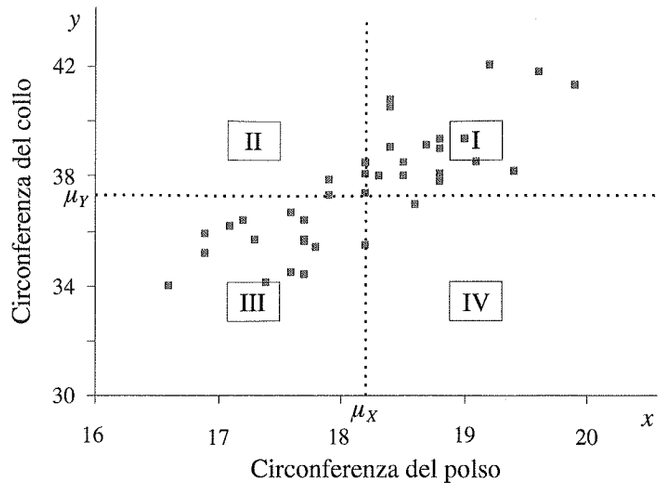
ESEMPIO 11.1

La serie doppia di osservazioni dell'Esempio 10.1 è rappresentata dal grafico di dispersione della Figura 11.1, dove è disegnato un nuovo sistema di assi cartesiani (linee tratteggiate) avente come origine il punto (μ_X, μ_Y), essendo μ_X e μ_Y le medie di X e di Y.

Commento

Come si può vedere, è manifesta la forte tendenza delle due variabili a crescere insieme. Le due variabili sono dunque correlate positivamente e i punti osservati si trovano prevalentemente nel I e nel III quadrante del nuovo piano cartesiano. ■

gura 11.1
afico di dispersione dei
i della Tabella 10.1.



ESEMPIO 11.2

Nella Tabella 11.1 sono riportati i dati di un esperimento volto allo studio della relazione tra livello medio della temperatura esterna e consumo settimanale di gas per il riscaldamento di un'abitazione (Hand *et al.*, 1994, con modificazioni).

bella 11.1
mperatura media esterna
onsumo settimanale di
s per il riscaldamento di
abitazione.

Temp. esterna (°C)	Consumo di gas ($ft^3 \cdot 1000$)	Temp. esterna (°C)	Consumo di gas ($ft^3 \cdot 1000$)
0.0	7.2	6.0	4.4
0.0	6.9	6.2	4.5
0.4	6.4	6.3	4.6
2.5	6.0	6.9	3.7
2.9	5.8	7.0	3.9
3.2	5.8	7.4	4.2
3.6	5.6	7.5	4.0
3.9	4.7	7.5	3.9
4.2	5.8	7.6	3.5
4.3	5.2	8.0	4.0
5.4	4.9	8.5	3.6
6.0	4.9	9.1	3.1
6.0	4.3	10.0	2.6

Questi dati sono rappresentati nel grafico di dispersione della Figura 11.2.

Commento

Il grafico suggerisce una forte correlazione negativa fra le due variabili: quasi tutti i punti osservati si collocano nel II e nel IV quadrante del nuovo piano cartesiano mostrando un andamento decrescente. ■

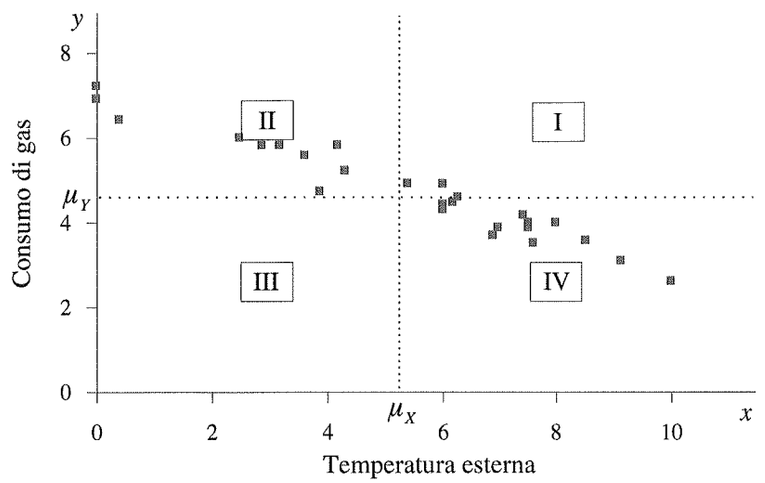


Figura 11.2
Grafico di dispersione dei dati della Tabella 11.1.

11.2 Misura della correlazione

Si consideri la serie doppia di osservazioni delle variabili X e Y, (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n). Siano μ_X e σ_X la media e la deviazione standard della variabile X; siano μ_Y e σ_Y la media e la deviazione standard della variabile Y. Siano

$$z_{x_i} = \frac{x_i - \mu_X}{\sigma_X} \quad \text{e} \quad z_{y_i} = \frac{y_i - \mu_Y}{\sigma_Y}$$

gli scarti standardizzati corrispondenti a x_i e y_i. Il prodotto

$$c_i = z_{x_i} \cdot z_{y_i} \tag{11.1}$$

è positivo o negativo a seconda che i due fattori abbiano lo stesso segno o segni diversi. È chiaro che $c_i > 0$ se x_i e y_i sono entrambi al di sopra o entrambi al di sotto delle rispettive medie. Al contrario, $c_i < 0$ se uno di essi è al di sopra e l'altro è al di sotto della propria media.

Esaminando di nuovo le Figure 11.1 e 11.2, è facile riconoscere che le quantità c_i corrispondenti ai punti osservati che si collocano nei quadranti I e III sono positive, mentre quelle associate ai punti giacenti nei quadranti II e IV sono negative. Ne segue che la media aritmetica delle quantità $c_1, c_2, \ldots, c_n$ sarà positiva se i punti osservati giacenti nei quadranti I e III sono più numerosi e presentano deviazioni più grandi da μ_X e μ_Y che non quelli che si collocano nei quadranti II e IV. Tale media sarà negativa nella situazione opposta. In altre parole, la media aritmetica sopra definita tenderà a essere positiva o negativa a seconda che tra le variabili X e Y vi sia correlazione positiva o negativa. Questo ragionamento conduce alla definizione del *coefficiente di correlazione lineare di Bravais* (dovuto a Auguste Bravais, 1811-1863).

DEFINIZIONE 11.2

Data la serie doppia di osservazioni delle variabili X e Y $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, il coefficiente di correlazione lineare è dato dalla media aritmetica delle quantità $c_1, c_2, \ldots, c_n$ definite dalla (11.1):

$$r = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \mu_X}{\sigma_X} \cdot \frac{y_i - \mu_Y}{\sigma_Y} \right) = \frac{\sum_{i=1}^{n} (x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{n} (x_i - \mu_X)^2 \sum_{i=1}^{n} (y_i - \mu_Y)^2}}. \tag{11.2}$$

Alla seconda espressione di r nella (11.2) si perviene in base all'identità

$$n\sigma_X\sigma_Y = \sqrt{n^2\sigma_X^2\sigma_Y^2} = \sqrt{\sum_{i=1}^{n}(x_i-\mu_X)^2\sum_{i=1}^{n}(y_i-\mu_Y)^2}.$$

La (11.2) corrisponde alla radice quadrata del coefficiente di determinazione (10.8). Questo fatto mette in relazione il coefficiente di correlazione con la retta di regressione e consente di affermare che la correlazione (positiva o negativa) è tanto più forte quanto più forte è la dipendenza lineare di Y da X, ossia, in termini geometrici, quanto più i punti osservati sono addensati attorno a una retta.

La (11.2) può anche essere scritta nella forma (si veda il Paragrafo A.8 dell'Appendice A)

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\mu_X\mu_Y}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\mu_X^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\mu_Y^2\right)}}, \tag{11.3}$$

più adatta per i calcoli perché non fa uso degli scarti $x_i - \mu_X$ e $y_i - \mu_Y$.

PROPOSIZIONE 11.1

Il coefficiente di correlazione lineare ha le proprietà di seguito indicate:

(a) assume valori nell'intervallo $[-1, 1]$; i valori estremi -1 e 1 vengono raggiunti se e solo se tra le variabili X e Y c'è perfetta relazione lineare; in particolare, è uguale a -1 quando i punti osservati si trovano su una retta con coefficiente angolare negativo, mentre è uguale a 1 quando tali punti si trovano su una retta con coefficiente angolare positivo;

(b) è positivo quando la retta di regressione di Y su X ha coefficiente angolare positivo; è negativo nel caso opposto;

(c) non cambia se le modalità della singola variabile vengono moltiplicate per una costante positiva o se vengono aumentate (diminuite) di una costante positiva.

La proprietà (a) si desume dal fatto che r è la radice quadrata dell'indice di determinazione r^2 della retta di regressione di Y su X, indice che assume valori nell'intervallo $[0, 1]$. Dalla stessa circostanza, si deduce che l'indice assume i valori estremi -1 e 1 se e solo se tra X e Y vi è una perfetta relazione lineare (si veda la Proposizione 10.3). La proprietà (b) si trae dalla circostanza che il coefficiente di correlazione e il coefficiente angolare della retta di regressione di Y su X sono espressi da rapporti che hanno in comune lo stesso numeratore, la codevianza, da cui dipende il loro segno. Quanto alla proprietà (c), la semplice verifica è lasciata per esercizio. ■

Alla luce delle proprietà appena indicate, è chiaro che l'indice r è una misura di correlazione lineare: valori dell'indice nell'intervallo $[-1, 0)$ segnalano una correlazione negativa, che è massima nel caso di perfetta relazione lineare negativa tra X e Y. Viceversa, valori dell'indice nell'intervallo $(0, 1]$ segnalano una correlazione positiva, che è massima quando vi è una perfetta relazione lineare positiva tra le due variabili. Infine, $r = 0$ indica l'assenza di correlazione lineare, il che non esclude che le due variabili siano legate da una relazione di tipo non lineare (si veda la Nota 3 di Commenti e curiosità del Capitolo 10).

ESEMPIO 11.3

Si calcoli il coefficiente di correlazione sui dati dell'Esempio 11.1, sapendo che $\mu_X = 18.18$ e $\mu_Y = 37.57$.

Soluzione

Il calcolo di r è illustrato nella Tabella 11.2, limitatamente alle prime tre e alle ultime due osservazioni.

Oss.	x_i	y_i	$x_i-\mu_X$	$y_i-\mu_Y$	$(x_i-\mu_X)^2$	$(y_i-\mu_Y)^2$	$(x_i-\mu_X)(y_i-\mu_Y)$
1	18.2	38.5	0.02	0.93	0.0004	0.8649	0.0186
2	16.6	34.0	−1.58	−3.57	2.4964	12.7449	5.6406
3	17.1	36.2	−1.08	−1.37	1.1664	1.8769	1.4796
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
37	18.8	39.4	0.62	1.83	0.3844	3.3489	1.1346
38	16.9	35.9	−1.28	−1.67	1.6384	2.7889	2.1376
Totale	–	–	–	–	22.5512	168.7542	50.2848

Tabella 11.2
Calcolo del coefficiente di correlazione sui dati dell'Esempio 11.1.

Da cui

$$r = \frac{50.2848}{\sqrt{22.5512 \cdot 168.7542}} = 0.81,$$

valore che segnala una correlazione positiva molto forte. ■

Nella Figura 11.3 sono rappresentati i grafici di dispersione di nove serie di osservazioni composte da 52 termini, in cui le deviazioni standard sono mantenute costanti (pari a 16.98 e 35.19). Un utile esercizio consiste nel predire il segno e il valore che l'indice r assume nella singola serie doppia per poi confrontarlo con il valore effettivo posto al margine della figura.

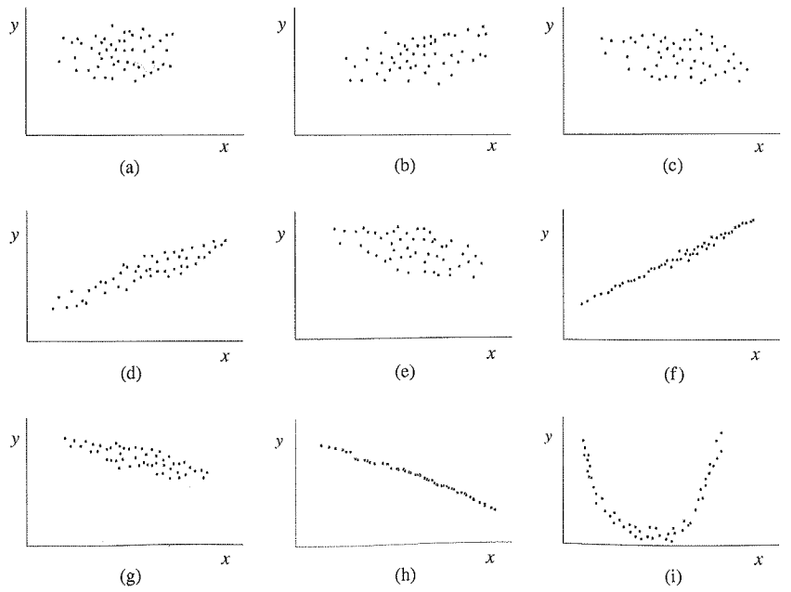


Figura 11.3
Dal grafico di dispersione al coefficiente di correlazione. L'indice r assume i valori: 0.02 in (a); 0.48 in (b); −0.30 in (c); 0.90 in (d); −0.52 in (e); 0.99 in (f); −0.80 in (g); −0.99 in (h); 0 in (i).

11.2.1 Correlazione tra graduatorie

Una variante del coefficiente di correlazione, utile in alcuni contesti di studio, consiste nel calcolare r non già sulla serie doppia $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, ma sulla serie doppia che si ottiene sostituendo a x_i e y_i i rispettivi posti, p_{x_i} e p_{y_i}, nelle graduatorie non decrescenti delle modalità di X e di Y. Ad esempio, alla serie doppia di osservazioni

$$(174, 72), (182, 75), (171, 64), (175, 69),$$

corrisponde la serie doppia dei corrispondenti posti in graduatoria:

$$(2, 3), (4, 4), (1, 1), (3, 2).$$

Se la singola variabile presenta modalità coincidenti, a ciascuna di queste va attribuito uno stesso posto, dato dalla media aritmetica dei posti che ad esse competerebbero nel caso in cui fossero diverse tra loro. Ad esempio, se le osservazioni di X fossero:

$$x_1 = 24,\ x_2 = 27,\ x_3 = 30,\ x_4 = 27,\ x_5 = 29,\ x_6 = 27,\ x_7 = 24,$$

si avrebbe:

$$p_{x_1} = 1.5,\ p_{x_2} = 4,\ p_{x_3} = 7,\ p_{x_4} = 4,\ p_{x_5} = 6,\ p_{x_6} = 4,\ p_{x_7} = 1.5.$$

ESEMPIO 11.4

Si calcoli il coefficiente di correlazione tra graduatorie utilizzando i dati dell'Esempio 11.2.

Soluzione

Si tratta di passare da ogni valore x_i di X al corrispondente posto nella graduatoria non decrescente dei 26 valori della variabile, come indicato all'inizio del paragrafo. La medesima operazione va effettuata sui valori della variabile Y. Il risultato di questa trasformazione è esposto nella Tabella 11.3.

Tabella 11.3
Trasformazione dei valori delle variabili X e Y in posti in graduatoria.

x_i	p_{x_i}	y_i	p_{y_i}	x_i	p_{x_i}	y_i	p_{y_i}
0	1.5	6.9	25	6	13	4.9	16.5
0	1.5	7.2	26	6.2	15	4.5	13
0.4	3	6.4	24	6.3	16	4.6	14
2.5	4	6	23	6.9	17	3.7	5
2.9	5	5.8	21	7	18	3.9	6.5
3.2	6	5.8	21	7.4	19	4.2	10
3.6	7	5.6	19	7.5	20.5	3.9	6.5
3.9	8	4.7	15	7.5	20.5	4	8.5
4.2	9	5.8	21	7.6	22	3.5	3
4.3	10	5.2	18	8	23	4	8.5
5.4	11	4.9	16.5	8.5	24	3.6	4
6	13	4.3	11	9.1	25	3.1	2
6	13	4.4	12	10	26	2.6	1

Il calcolo del coefficiente di correlazione sui posti in graduatoria così determinati dà $r = -0.96$, un numero che suggerisce una correlazione negativa molto forte tra le graduatorie non decrescenti delle modalità di X e di Y. ■

In assenza di modalità coincidenti, si può dimostrare che il coefficiente di correlazione tra le graduatorie può essere posto nella forma

$$r_S = 1 - \frac{6 \sum_{i=1}^{n} (p_{x_i} - p_{y_i})^2}{n(n^2 - 1)}. \tag{11.4}$$

L'indice (11.4) è chiamato indice di correlazione di Spearman, dal nome dello psicologo britannico Charles Edward Spearman (1863-1945), che l'ha introdotto in modo diretto (non come caso particolare del coefficiente di correlazione di Bravais).

Vale la pena di osservare che l'indice r_S, a differenza del coefficiente di correlazione di Bravais, si avvicina a 1 o a -1 se i punti osservati sono vicini a una curva crescente o decrescente, anche se diversa dalla retta.

Riassunto

Due variabili quantitative, X e Y, si dicono correlate positivamente se al crescere dell'una l'altra tende a crescere. Al contrario, si dicono correlate negativamente se al crescere dell'una l'altra tende a decrescere.

Come misura della correlazione, viene introdotto il coefficiente di correlazione lineare di Bravais r, dato dal rapporto tra la codevianza di X e Y e la radice quadrata del prodotto delle devianze delle due variabili. L'indice varia nell'intervallo $[-1, 1]$, assumendo valori negativi nel caso di correlazione negativa e valori positivi nel caso opposto. Viene chiarito come r sia un indice di correlazione lineare, essendo la radice quadrata dell'indice di determinazione r^2, che è una misura di dipendenza lineare tra Y da X.

Il coefficiente di correlazione viene poi esteso al caso delle graduatorie, dove alle coppie di osservazioni (x_i, y_i) vengono sostituiti i rispettivi posti nelle graduatorie non decrescenti delle osservazioni di X e di Y.

Parole chiave

coefficiente di correlazione lineare, 235
correlazione negativa, 233
correlazione positiva, 233
indice di correlazione di Spearman, 239

Commenti e curiosità

1. Correlazione e dati anomali

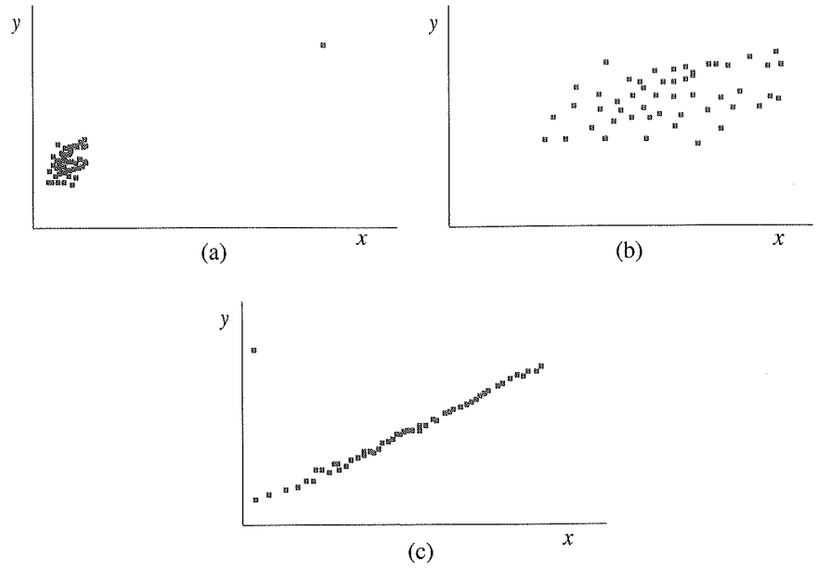
La presenza di dati anomali (punti caratterizzati da una o entrambe le coordinate molto piccole o molto grandi rispetto a quelle degli altri punti) può provocare un forte aumento o una forte diminuzione del coefficiente di correlazione. Nel grafico di dispersione 11.4 (a), caratterizzato dalla presenza di un punto anomalo, r è uguale a 0.84; nel diagramma 11.4 (b), relativo agli stessi dati depurati del punto anomalo, r è uguale a 0.43. Nel diagramma 11.4 (c), la presenza di un punto anomalo fa sì che il coefficiente di correlazione di Bravais sia pari a 0.83, contro un valore molto vicino all'unità che si ottiene eliminando il punto anomalo.

2. Correlazione e relazione di causa ed effetto

Nel linguaggio comune, il termine correlazione viene spesso utilizzato per indicare l'esistenza di un nesso di causalità tra fenomeni. Questo è il senso di affermazioni del tipo "Il cancro al polmone è fortemente correlato al fumo di sigarette". Vale la pena di riflettere sul fatto che una forte correlazione tra due variabili segnalata da un elevato valore assoluto di r non implica necessariamente un tale nesso. Invero, due variabili possono presentare una forte correlazione in assenza di un rapporto di causalità, nelle seguenti circostanze.

- Entrambe le variabili sono collegate da un nesso di causalità con una terza variabile. Ad esempio, la probabile alta correlazione tra spese delle famiglie per consumi di lusso e cilindrata dell'automobile del capofamiglia si spiega con il fatto che entrambe le variabili dipendono dal livello del reddito disponibile.

Figura 11.4
Correlazione e dati anomali.

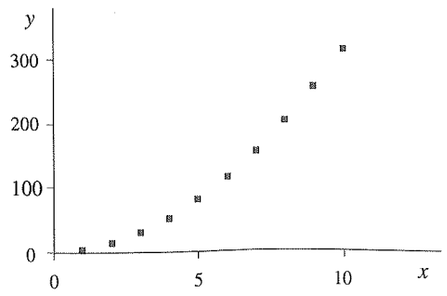


- Entrambe le variabili si modificano nel tempo. Un'alta correlazione tra il prezzo delle sigarette e il prezzo delle arance nell'arco di 36 mesi è spiegabile con la circostanza che i prezzi di tutti i generi tendono a crescere nel tempo.
- La correlazione è dovuta all'esistenza di diversi fattori che operano congiuntamente sulla variabile dipendente, i cui contributi non sono scindibili. Per esempio, la possibile alta correlazione tra il tasso di mortalità per incidenti stradali e i limiti di velocità in un insieme di paesi non implica necessariamente che la diminuzione di tali limiti provochi una diminuzione del tasso di mortalità; ciò in quanto sul tasso di mortalità incidono svariati elementi, quali la densità di traffico nelle strade, le regole di circolazione vigenti nei diversi Paesi, ecc.
- Un'alta correlazione potrebbe essere dovuta all'uso di indicatori non normalizzati. La probabile forte correlazione tra consumo totale di alcool e numero di rapine in città con dimensioni molto diverse è dovuto all'uso di numeri assoluti anziché degli appropriati rapporti normalizzati.
- L'associazione può essere una pura coincidenza.

3. Massima correlazione e valore di r.

Per rimarcare che il coefficiente di correlazione r non assume sempre il valore 1 quando vi è massima correlazione positiva, si osservi il grafico di dispersione 11.5, dove le coordinate dei punti osservati sono tali che $y_i = \pi x_i^2$. Naturalmente, si è in presenza di massima correlazione positiva, ma per essa r non assume il valore 1.

Figura 11.5
Distribuzione con una correlazione positiva massima.



Esercizi

11.1 Si consideri la serie doppia di osservazioni di cui all'Esercizio 10.1. Si misuri correlazione tra le due variabili.

11.2 Si riprendano i dati dell'Esercizio 10.3. Si calcoli il coefficiente di correlazione lineare e si commenti il risultato.

11.3* Si verifichi che moltiplicando per due costanti positive qualsiasi, a e b, i valori di X e Y, il coefficiente di correlazione lineare non cambia. (*Suggerimento*: si prenda la formula (11.3), sostituendo x_i con ax_i e y_i con by_i; si svolgano le conseguenti operazioni algebriche e si semplifichi.)

11.4 Si considerino le tre serie doppie di osservazioni:

x_i	y_i
3	25
5	18
9	32
7	27
10	39
12	42
11	42

x_i	y_i
25	3
18	5
32	9
27	7
39	10
42	12
42	11

x_i	y_i
25	6
18	10
32	18
27	14
39	20
42	24
42	22

Senza effettuare i calcoli, si spieghi perché nelle tre serie il coefficiente di correlazione lineare assume lo stesso valore.

11.5 Si consideri il *file* "Età_coniugi" contenente le età di un campione di 100 coniugi (Hand *et al.*, 1994, con variazioni):

Età del marito	25	40	32	...	64	59
Età della moglie	28	30	27	...	64	56

a) Si rappresenti la serie doppia di osservazioni con un grafico di dispersione e, sulla base del grafico, si stabilisca se la correlazione è debole, media o forte.

b) Si calcoli il coefficiente di correlazione lineare.

11.6 Si prenda il *file* "Quotazioni". Si calcoli l'indice r per le quotazioni giornaliere delle azioni delle società ENI S.p.A. e STMicroelectronics dall'1 gennaio al 28 febbraio 2020 e si commenti il risultato.

11.7 Riprendendo i dati dell'Esercizio 1.4, si calcoli il coefficiente di correlazione lineare tra le variabili altezza e volume.

11.8 Si calcoli l'indice di correlazione tra graduatorie sui dati dell'Esercizio 11.6.

Correlazione con Excel

In Excel è presente la funzione CORRELAZIONE( ) che restituisce il coefficiente di correlazione tra due variabili. Nella Figura 11.6, viene illustrata questa funzione utilizzando la serie di osservazioni discussa nell'Esercizio 10.3.

Figura 11.6
Funzione CORRELAZIONE( ).

D2 =CORRELAZIONE(A2:A21;B2:B21)

	A	B	C	D	E	F	G
1	**Distanza in linea d'aria**	**Distanza stradale**					
2	9.5	10.7		0.97			
3	5	6.5					
4	23	29.5					
5	15.2	17.2					
6	11.4	18.4					
7	11.8	19.7					

PARTE

II

Fondamenti di probabilità

Capitolo 12
Probabilità

Capitolo 13
Variabili casuali

Capitolo 14
Alcune distribuzioni di probabilità parametriche

Capitolo 15
Distribuzioni di probabilità congiunte

Capitolo 16
Legge dei grandi numeri e teorema del limite centrale

Capitolo 17
Campioni casuali e distribuzioni campionarie

CAPITOLO

12

Probabilità

Sommario del capitolo

12.1 Introduzione
12.2 Esperimenti casuali, spazio campionario ed eventi
12.3 Probabilità
12.4 Assegnazione delle probabilità agli eventi
12.5 Probabilità condizionata
12.6 Indipendenza tra eventi
12.7 Formula di Bayes
Riassunto
Commenti e curiosità
Esercizi

12.1 Introduzione

L'idea di probabilità è intuitivamente associata a quella di incertezza, come è del resto testimoniato dal linguaggio di tutti i giorni quando si usano frasi del tipo: "È molto probabile che la prossima settimana il prezzo del petrolio aumenti"; "È improbabile che domani piova"; "È quasi certo che la squadra A batta la squadra B".

Come disciplina matematica, la probabilità è l'insieme dei criteri con cui si misura la possibilità che un evento incerto si manifesti. La nascita ufficiale di questa disciplina si fa risalire alla metà del XVII secolo, precisamente al carteggio tra Pascal e Fermat, in cui i due matematici analizzavano i quesiti su alcuni giochi d'azzardo posti a Pascal dal Cavalier de Méré. Lo sviluppo dei primi fondamenti della probabilità è associato ai nomi di eminenti matematici, tra i quali si ricordano Jacob Bernoulli, Christian Huygen, Pierre-Simon Laplace, Abraham de Moivre, Siméon-Denis Poisson e Thomas Bayes.

Nei prossimi paragrafi, si userà spesso il termine esperimento casuale per indicare un'operazione o un'attività che può essere eseguita più volte e il cui risultato a ogni replica è incerto, come avviene nei giochi di sorte, quali il lancio di una moneta o di un dado, l'estrazione di una pallina da un'urna, ecc. Una caratteristica importante di questi esperimenti è che essi possono essere ripetuti nelle stesse condizioni. Ciò consente di costruire modelli matematici che descrivono il comportamento di eventi casuali associati alla ripetizione dell'esperimento. Ad esempio, è possibile definire un modello che permette di calcolare la probabilità che su n lanci di un dado equilibrato l'evento faccia 1 si verifichi un certo numero di volte.

L'importanza di questi modelli è rappresentata dalla loro applicabilità in molti ambiti (scientifici, tecnologici, sociali), in situazioni in cui le osservazioni empiriche possono essere riguardate come il risultato della replica di esperimenti casuali ripetibili nelle stesse condizioni. Si pensi: alle sperimentazioni cliniche, in cui i pazienti ricevono un trattamento del quale viene studiata l'efficacia; al monitoraggio della qualità dei pezzi prodotti in un'azienda industriale tramite un campione prelevato dalla linea di produzione; all'indagine condotta da un agenzia di ricerca per stimare la percentuale di persone che sostengono una determinata politica di governo e così via. In ciascuna di queste situazioni, si ha un insieme di dati che: (a) sono ottenuti effettivamente replicando più volte un espe-

rimento casuale; (b) sono equiparabili, sotto certe condizioni, a dati del tipo descritto al punto (a).

Sono generati da un meccanismo casuale, ad esempio, i dati provenienti dai sondaggi delle agenzie di ricerca basati su campioni casuali. Un campione casuale di ampiezza 1000 può essere formato, tra gli altri modi, con un procedimento che equivale dell'estrazione con ripetizione di 1000 palline da un'urna che contiene tante palline quante sono le unità della popolazione recanti i numeri progressivi da 1 a N (essendo N il numero delle unità della popolazione), cioè i numeri identificativi delle unità stesse.

Al calcolo delle probabilità sono dedicati interi volumi. Qui e nei prossimi cinque capitoli, verranno esaminati gli aspetti più importanti e quelli che costituiscono la base necessaria per la discussione dei problemi di inferenza statistica che verranno affrontati nei capitoli dal 18 al 24.

Il resto di questo capitolo è organizzato come segue. Verranno introdotte innanzitutto le nozioni di esperimento casuale, spazio campionario ed eventi. A ciò farà seguito la definizione assiomatica della probabilità insieme alle principali concezioni di questa (classica, frequentista e soggettivista). Si esaminerà poi il problema dell'assegnazione delle probabilità agli eventi. Successivamente, l'attenzione verrà concentrata sul concetto di probabilità condizionata, ossia sulla probabilità di un evento assumendo che un altro evento si sia verificato, e sulla nozione cruciale di indipendenza tra eventi. Infine, verrà illustrata la cosiddetta formula di Bayes.

12.2 Esperimenti casuali, spazio campionario ed eventi

In generale, un esperimento è un processo o un'attività che può essere condotta e ripetuta in condizioni controllate per studiare un fenomeno di interesse. Un esperimento è detto deterministico se, eseguito nelle condizioni prescritte, produce un unico prevedibile risultato. Ad esempio, se un bollitore di acqua pura viene portato a una temperatura di 100 °C e 760 mmHg di pressione, l'acqua bolle. Ciascuna replica dell'esperimento, nelle condizioni specificate, dà luogo allo stesso risultato.

Al contrario, un esperimento i cui esiti non possono essere previsti con certezza è chiamato esperimento casuale. Esempi di esperimenti casuali sono: il lancio di una moneta osservandone il risultato, il prelievo di un pacchetto di pasta da una linea di confezionamento per determinarne il peso, l'osservazione della frequenza dei pezzi difettosi in una catena di montaggio, l'osservazione della variazione giornaliera di un indice di borsa. La singola esecuzione di un esperimento casuale è detta prova.

Ci sono esperimenti casuali che possono essere eseguiti ripetutamente in condizioni identiche e altri che possono essere concepiti come tali. Un esempio del primo tipo di esperimento è il lancio di una moneta. Un esperimento in cui un determinato farmaco viene somministrato a cavie può essere considerato ripetibile, sostanzialmente nelle stesse condizioni, se si assume che le cavie usate siano simili nelle loro caratteristiche rilevanti.

Una caratteristica degli esperimenti casuali è che i possibili esiti sono noti e specificabili prima che l'esperimento venga eseguito. Il singolo esito è chiamato evento elementare ed è indicato con ω. L'insieme di tutti gli eventi elementari è chiamato spazio campionario ed è indicato con Ω. A volte, in questo paragrafo, lo spazio campionario sarà costruito con l'ausilio di un diagramma chiamato albero degli eventi.

ESEMPIO 12.1

Si descriva l'esperimento che consiste nel lanciare tre volte una moneta e nell'osservare le facce che si presentano.

Soluzione

In un singolo lancio può uscire testa (T) o croce (C). Tutti i possibili risultati dell'esperimento si possono ottenere con l'aiuto dell'albero degli eventi rappresentato nella Figura 12.1. Lo spazio campionario è dato da

$$\Omega = \{TTT, TTC, TCT, TCC, CTT, CTC, CCT, CCC\}$$

ed è rappresentato dagli otto punti all'interno del rettangolo nella Figura 12.1.

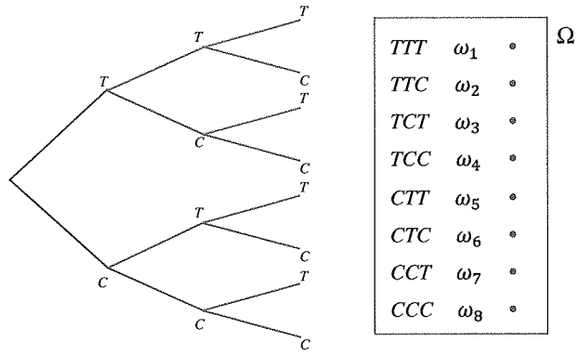


Figura 12.1
Albero degli eventi e spazio campionario nell'esperimento dei tre lanci di moneta.

■

ESEMPIO 12.2

Si costruisca lo spazio campionario dell'esperimento che consiste nel lancio di due dadi e nell'osservazione dei numeri che appaiono sulle facce superiori.

Soluzione

Lo spazio campionario è dato da

$$\Omega = \{(1,1), (1,2), \ldots, (1,6), \ldots, (6,1), (6,2), \ldots, (6,6)\},$$

dove in ogni coppia il primo numero indica il risultato apparso sulla faccia superiore del primo dado (dado 1) e il secondo il risultato relativo all'altro dado (dado 2). Lo spazio campionario è rappresentato nella Figura 12.2.

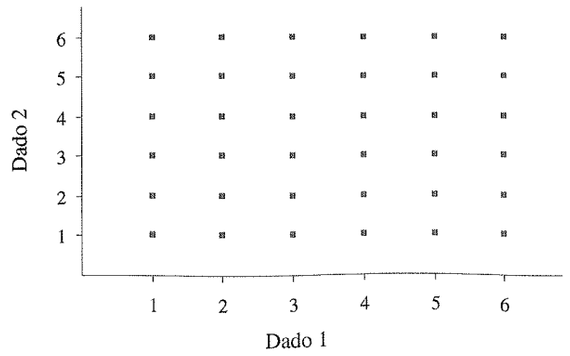


Figura 12.2
Spazio campionario nell'esperimento del lancio di due dadi.

■

ESEMPIO 12.3

Si descriva lo spazio campionario dell'esperimento che consiste nel lanciare una moneta ripetutamente finché non esce testa.

Soluzione

Lo spazio campionario dell'esperimento è costituito da una infinità numerabile di elementi:

$$\Omega = \{T, CT, CCT, CCCT, CCCCT, \ldots\}.$$

■

ESEMPIO 12.4

Si immagini di osservare la durata di una lampadina elettrica prodotta da una certa azienda produttrice di materiale elettrico. Si definisca lo spazio campionario connesso all'esperimento.

Soluzione

I possibili risultati dell'esperimento sono tutti i numeri reali dell'intervallo $[0, T]$, dove T è la durata massima desumibile dalle caratteristiche tecniche del processo produttivo. Si può allora scrivere

$$\Omega = \{\omega : 0 \leq \omega \leq T\}.$$

■

Come si comprende dagli esempi, lo spazio campionario può contenere un numero finito (Esempi 12.1 and 12.2), una infinità numerabile (Esempio 12.3) o una infinità non numerabile di eventi elementari (Esempio 12.4). Spazi campionari costituiti da un numero finito o da una infinità numerabile di elementi si chiamano discreti; quelli costituiti da una infinità non numerabile di elementi si dicono continui. Va considerato che gli spazi campionari non numerabili sono solo una idealizzazione matematica di situazioni reali: l'osservazione dei fenomeni reali conduce sempre, come conseguenza dell'approssimazione delle misure, a spazi campionari discreti. Così nell'Esempio 12.4, la durata della lampadina può essere misurata in ore, minuti o altra unità; in ogni caso, lo spazio campionario conseguente contiene un numero finito, seppur elevato, di elementi.

Uno spazio campionario è sempre riconducibile, in modo naturale (come negli Esempi 12.2 e 12.4) o convenzionale (come nell'Esempio 12.1), a un insieme di punti, dove ogni punto rappresenta un evento elementare.

DEFINIZIONE 12.1

Un evento è un sottoinsieme dello spazio campionario Ω. Un evento è detto semplice se è costituito da un singolo evento elementare, è detto composto se è costituito da due o più eventi elementari. Un evento si realizza se si manifesta uno degli eventi elementari che lo compongono.

ESEMPIO 12.5

Nell'esperimento del lancio di un dado, in cui lo spazio campionario è $\Omega = \{1, 2, 3, 4, 5, 6\}$, eventi semplici sono i sottoinsiemi $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, $\{6\}$. Invece, esempi di eventi composti sono dati dai sottoinsiemi $A = \{1, 3\}$, $B = \{2, 3, 5\}$, $C = \{1, 5, 6\}$. L'evento composto C, tra gli altri, si realizza se nel lancio del dado si osserva il numero 1, il numero 5 oppure il numero 6.

■

ESEMPIO 12.6

Si riprenda lo spazio campionario dell'esperimento dei tre lanci di moneta:

$$\Omega = \{TTT, TTC, TCT, TCC, CTT, CTC, CCT, CCC\}.$$

Eventi semplici sono i sottoinsiemi $TTT, TTC, \ldots, CCC$. Eventi composti sono, per esempio, $A = \{TTT, TTC, TCT, CTT\}$ e $B = \{CTT, CTC, CCT, CCC\}$.

■

Nelle applicazioni pratiche, gli eventi di maggiore interesse sono quelli composti i cui eventi elementari condividono qualche proprietà comune. Eventi di questo tipo sono i sottoinsiemi A e B dell'esempio precedente. Infatti, gli eventi elementari di A hanno la proprietà comune di includere almeno due teste, gli eventi elementari di B condividono la proprietà di presentare una croce al primo lancio.

Data la corrispondenza fra eventi e insiemi di punti, lo studio delle relazioni tra eventi è riconducibile allo studio delle relazioni tra insiemi. Una illustrazione molto utile per descrivere gli insiemi e per mostrarne le relazioni reciproche è il diagramma di Venn (dal nome del logico matematico J. Venn, 1834-1923). Si tratta di un diagramma che rappresenta l'insieme con i punti contenuti in un cerchio, in un rettangolo o in un'altra figura geometrica piana. Generalmente, lo spazio campionario viene rappresentato con un rettangolo e un generico evento con un cerchio o un'ellisse all'interno del rettangolo.

Prima di procedere, conviene definire alcune importanti relazioni tra eventi, facendo ricorso a risultati elementari della teoria degli insiemi.

DEFINIZIONE 12.2

Siano A e B due eventi arbitrari dello spazio campionario Ω. Allora:

- l'evento unione di A e B, indicato con $A \cup B$, è formato da tutti gli eventi elementari che appartengono ad A, a B o a entrambi;
- l'evento intersezione di A e B, indicato con $A \cap B$, è formato da tutti gli eventi elementari che appartengono sia ad A che a B. I due eventi si dicono disgiunti o incompatibili se non hanno nessun evento elementare in comune, cioè se $A \cap B = \varnothing$, essendo $\varnothing$ l'insieme vuoto, chiamato anche evento impossibile;
- l'evento complementare di A, indicato con $\bar{A}$, è formato da tutti gli eventi elementari di Ω che non appartengono ad A.

La Figura 12.3 illustra le relazioni tra eventi sopra definite per mezzo di diagrammi di Venn.

ESEMPIO 12.7

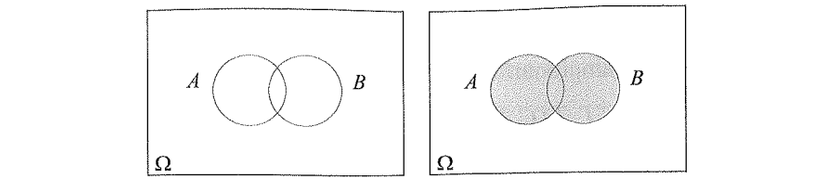
Si consideri l'esperimento descritto nell'Esempio 12.1. Si indichino con A e B gli eventi "Almeno due volte testa" e "Croce al primo lancio". Si descrivano gli eventi $A \cup B$, $A \cap B$ e $\bar{A}$.

Soluzione

Gli eventi sono così formati (si veda anche la Figura 12.4):

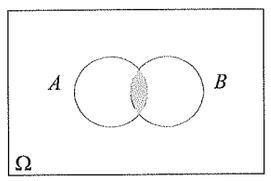
$$A \cup B = \{\omega_1, \omega_2, \omega_3, \omega_5, \omega_6, \omega_7, \omega_8\},\ A \cap B = \{\omega_5\},\ \bar{A} = \{\omega_4, \omega_6, \omega_7, \omega_8\}.$$

■

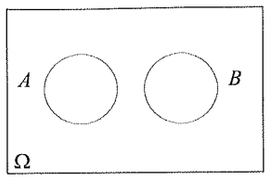


Eventi A e B

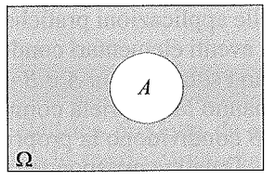
La zona ombreggiata è $A \cup B$



La zona ombreggiata è $A \cap B$



Eventi disgiunti



La zona ombreggiata è $\bar{A}$

Figura 12.3
Diagrammi di Venn di alcune operazioni tra eventi.

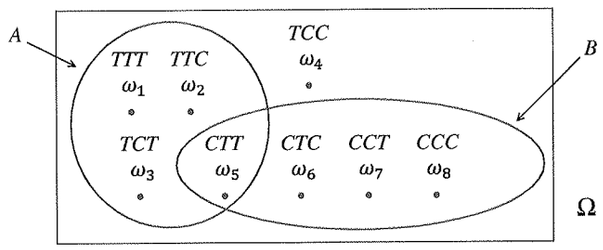


Figura 12.4
Diagramma di Venn degli eventi A e B.

ESEMPIO 12.8

Si consideri un campione di quattro mutui per la casa, ognuno dei quali può essere a tasso fisso (F) o variabile (V). Si considerino i seguenti eventi: "Tre dei mutui selezionati sono a tasso fisso" (evento A), "Al più uno dei quattro mutui è a tasso variabile" (evento B), "Tutti e quattro i mutui sono dello stesso tipo" (evento C). Si descrivano gli eventi: $\bar{A}$, $B \cup C$ e $B \cap C$.

Soluzione

Lo spazio campionario è dato da

$$\Omega = \{VVVV,\ FVVV,\ VFVV,\ VVFV,\ VVVF,\ FFVV,\ FVFV,\ FVVF,\ VFFV,\ VVFF,\ VFVF,\ VFFF,\ FVFF,\ FFVF,\ FFFV,\ FFFF\}.$$

Si ha allora:

$$\bar{A} = \{VVVV,\ FVVV,\ VFVV,\ VVFV,\ VVVF,\ FFVV,\ FVFV,\ FVVF, VFFV,\ VVFF,\ VFVF,\ FFFF\},$$

$$B \cup C = \{VFFF, FVFF, FFVF, FFFV, FFFF, VVVV\},$$

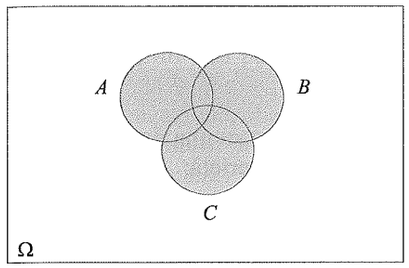
$$B \cap C = FFFF.$$

■

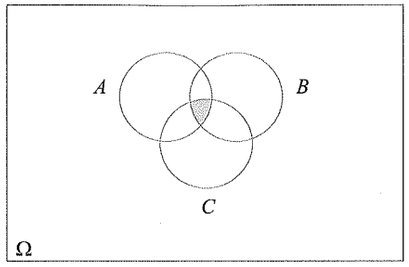
Le operazioni di unione e intersezione possono essere estese a più di due eventi. Siano $A_1, A_2, \ldots, A_k$ k eventi definiti sullo spazio campionario Ω. Allora:

- la loro unione, indicata con $A_1 \cup A_2 \cup \cdots \cup A_k$, è data dall'insieme degli eventi elementari che appartengono ad almeno uno di essi;
- la loro intersezione, indicata con $A_1 \cap A_2 \cap \cdots \cap A_k$, è data dall'insieme degli eventi elementari che appartengono a tutti gli A_i, $i = 1, 2, \ldots, k$.

Queste operazioni sono illustrate nella Figura 12.5 per il caso di tre eventi.



La zona ombreggiata è $A \cup B \cup C$



La zona ombreggiata è $A \cap B \cap C$

Figura 12.5
Diagramma di Venn degli eventi $A \cup B \cup C$ e $A \cap B \cap C$.

DEFINIZIONE 12.3

Gli eventi $A_1, A_2, \ldots, A_k$ costituiscono una partizione dello spazio campionario Ω se sono soddisfatte le tre condizioni:

- $A_i \neq \emptyset$, $i = 1, 2, \ldots, k$;
- $A_i \cap A_j = \emptyset$, $i \neq j = 1, 2, \ldots, k$;
- $A_1 \cup A_2 \cup \cdots \cup A_k = \Omega$.

ESEMPIO 12.9

Nell'esperimento del lancio di un dado, un esempio di partizione dello spazio campionario ($\Omega = \{1, 2, 3, 4, 5, 6\}$) è costituita dagli eventi: $A_1 = \{1, 3\}$, $A_2 = \{2, 4, 5\}$, $A_3 = \{6\}$. È immediato vedere che questi eventi sono a due a due disgiunti e che la loro unione dà Ω.

■

12.3 Probabilità

Si voglia ora definire una funzione $P(\cdot)$ che assegna una probabilità a ogni elemento di una famiglia di sottoinsiemi di Ω. Matematicamente, $P(\cdot)$ è una funzione il cui dominio non è costituito dai punti dell'asse reale, bensì da una collezione di insiemi; per questo viene chiamata funzione d'insieme.

Per avvicinarsi in modo intuitivo alla definizione di $P(\cdot)$, è utile partire dalla nozione di frequenza relativa. Si immagini di aver ripetuto n volte un esperimento, per esempio il lancio di un dado. Indicando con A e B due eventi disgiunti (per esempio "Numero pari" e "Faccia numero 5"), si assuma che le frequenze con cui si sono osservati gli eventi A e B siano n_A e n_B, a cui corrispondono le frequenze relative $f_A = n_A/n$ e $f_B = n_B/n$. L'esperienza insegna che queste frequenze relative sono erratiche per piccoli valori di n e tendono a stabilizzarsi attorno a

dei valori costanti, p_A e p_B (3/6 e 1/6, nell'esempio), al crescere di n. Questo fatto porta a identificare le probabilità degli eventi A e B con le loro frequenze relative, p_A e p_B, con cui ci si aspetta che si verifichino quando l'esperimento è ripetuto un grande numero di volte. Ovviamente, f_A e f_B sono non negative e la frequenza relativa dell'unione di A e B è $f_{A\cup B} = n_A/n + n_B/n$, con $f_{A\cup B}$ vicina a $p_A + p_B$ (3/6 + 1/6) quando n è molto grande. Infine, la frequenza relativa, f_Ω, dell'evento certo ("Uno qualsiasi dei numeri tra 1 e 6") è pari a 1.

La definizione di *misura di probabilità* che viene data qui di seguito è in linea con quanto appena detto intorno alle proprietà della frequenza relativa.

DEFINIZIONE 12.4

Una misura di probabilità $P(\cdot)$ è una funzione definita su una data famiglia di sottoinsiemi (eventi) di Ω che soddisfa le seguenti proprietà:

(a) $P(A) \geq 0$, per ogni evento A;

(b) $P(\Omega) = 1$;

(c) $P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$, per ogni sequenza di eventi disgiunti.

La Proprietà (a) stabilisce che la probabilità è non negativa. La Proprietà (b) assegna probabilità 1 all'evento certo che uno qualsiasi degli eventi elementari si verifichi. La Proprietà (c) asserisce che, dati due o più eventi disgiunti, la probabilità che uno di essi si realizzi è pari alla somma delle loro probabilità.

Una discussione rigorosa intorno alla "famiglia di sottoinsiemi" sulla quale è definita $P(\cdot)$ va oltre gli scopi di questo testo. Nondimeno, si può dire che: 1) lo spazio campionario Ω deve appartenere alla famiglia; 2) se gli eventi $A_1, A_2, \ldots$ appartengono alla famiglia, anche la loro unione deve appartenere alla famiglia; 3) se l'evento A appartiene alla famiglia, anche il suo complemento $\bar{A}$ deve appartenere alla famiglia. Qualche approfondimento su questo punto è dato nel Paragrafo B.1 dell'Appendice B.

I tre assiomi enunciati nella Definizione 12.4 costituiscono i fondamenti della teoria della probabilità e sono conosciuti come assiomi della probabilità di Kolmogorov, dal nome del matematico Andrey N. Kolmogorov (1903-1987).

Dai tre assiomi si possono derivare le seguenti ulteriori proprietà per la funzione $P(\cdot)$:

- $P(\varnothing) = 0$;
- $P(A) \leq 1$, per ogni evento A;
- $P(\bar{A}) = 1 - P(A)$, per ogni evento A;
- $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$, dove A_1 e A_2 sono due eventi arbitrari.

La terza proprietà viene chiamata regola dell'evento complementare; la quarta regola della somma. Le prime tre proprietà si possono verificare facilmente per mezzo dei diagrammi di Venn. La dimostrazione della quarta è contenuta nel Paragrafo B.2 dell'Appendice B.

12.3.1 Interpretazione della probabilità

Il calcolo delle probabilità ha a che fare sia con la matematica che con la filosofia. I fondamenti matematici della teoria sono gli assiomi di Kolmogorov che

stabiliscono le regole che la misura di probabilità $P(\cdot)$ deve soddisfare. L'aspetto filosofico riguarda invece l'interpretazione della probabilità, alla quale sono stati dedicati interi volumi (si veda, ad esempio, Gillies, 2000). Si menzionano qui brevemente le tre principali scuole di pensiero.

Conviene partire dalla cosiddetta interpretazione classica esposta da Pierre-Simon Laplace (1749-1827) nel suo saggio del 1814 (*Essai Philosophique sur les Probabilités*), ma già emersa nel periodo 1650-1800 dalle discussioni tra matematici sui problemi dei giochi di sorte. Secondo questa interpretazione, la probabilità dell'evento A è il rapporto tra il numero di casi favorevoli al verificarsi di A e il numero totale di casi possibili, purché questi siano equiprobabili (ugualmente possibili secondo le parole di Laplace).

Questa concezione, che è stata dominante tra i matematici per un secolo, presenta due chiari limiti: in primo luogo, soffre di circolarità, impiegando l'espressione eventi elementari ugualmente probabili nella definizione di probabilità; in secondo luogo, è applicabile solo quando gli eventi elementari sono equiprobabili.

Una seconda interpretazione della probabilità è nota come interpretazione frequentista. Si consideri il lancio di una moneta bilanciata e si immagini di replicare l'esperimento n volte in modo indipendente. Un possibile risultato potrebbe essere la sequenza

$$CTCCTCTTTCTCCCCTTT\ldots$$

Si consideri la frequenza relativa di testa. L'esperienza suggerisce che all'aumentare di n, questa frequenza relativa fluttua sempre meno e tende a stabilizzarsi intorno al valore limite 0.5, che è la probabilità di T (si veda la Figura 12.6).

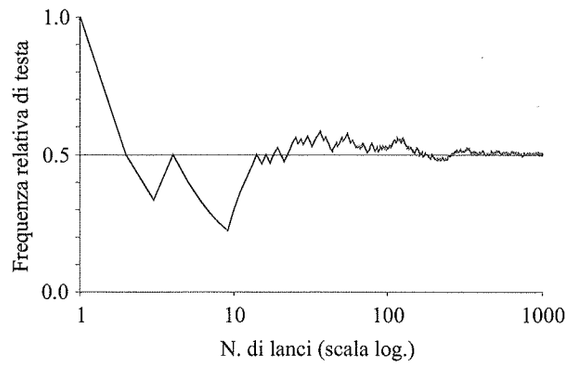


Figura 12.6
Frequenza relativa di testa nel lancio di una moneta regolare all'aumentare del numero delle prove.

Questa idea può essere generalizzata nella cosiddetta legge empirica del caso (si veda Gillies, 2000, p. 92): se si ripete un esperimento un gran numero di volte nelle stesse condizioni, ciascuno dei possibili eventi si verificherà con una frequenza relativa approssimativamente uguale alla sua probabilità, e l'approssimazione ordinariamente aumenterà all'aumentare del numero delle prove. Questa regolarità statistica si osserva non solo nei giochi di sorte, ma anche nel campo dei fenomeni fisici, biologici, economici e sociali.

L'interpretazione frequentista, richiedendo che l'esperimento sia ripetibile, non copre il campo degli eventi non ripetibili la cui incertezza può essere comunque espressa in termini probabilistici. Si pensi, ad esempio al vasto ambito dei concorsi a pronostico, quali lotto, totocalcio ecc.

Per superare queste difficoltà, è stata proposta una terza interpretazione, nota come interpretazione soggettivista della probabilità, secondo la quale la probabilità di un evento è il grado di convinzione che un individuo ha, sulla base delle

informazioni possedute in un dato momento, nel verificarsi dell'evento. Tra i più importanti sostenitori di questa teoria si ricordano: Frank P. Ramsey (1903-1930), Bruno de Finetti (1906-1985) e Leonard J. Savage (1917-1971).

Vale la pena notare che individui diversi con la stessa o diversa quantità di informazioni possono arrivare a probabilità differenti per lo stesso evento. Ad esempio, un operatore di borsa può credere che la probabilità che il prezzo di un'azione aumenti nella settimana successiva sia 0.6, mentre una persona con minori informazioni potrebbe assegnare a tale evento una probabilità anche molto diversa.

12.4 Assegnazione delle probabilità agli eventi

Fin qui si sono esaminate le proprietà della misura di probabilità $P(\cdot)$. Si rivolga ora l'attenzione al problema di come assegnare le probabilità agli eventi.

Va detto che la famiglia di eventi sui quali è definita la funzione $P(\cdot)$ può essere molto ampia anche per uno spazio campionario che contiene un numero moderato di punti. Il processo di assegnazione delle probabilità agli eventi è enormemente semplificato quando lo spazio campionario è discreto. In questo caso, è sufficiente assegnare una probabilità a ogni evento elementare: la probabilità di un qualsiasi evento A si potrà poi ottenere usando l'assioma (c).

Quando lo spazio campionario è continuo, come nell'Esempio 12.4, l'assegnazione della probabilità $P(\cdot)$ va fatta prendendo in considerazione famiglie di intervalli, essendo impossibile assegnare probabilità ai singoli punti.

Si esamini ora il problema con riferimento a spazi campionari composti da un numero finito N di eventi elementari, nel caso di informazione completa, cioè quando le probabilità di tutti gli eventi elementari sono note. In questo caso, una volta assegnate le probabilità a tutti gli eventi elementari, la probabilità di un qualsiasi evento A si ottiene sommando le probabilità degli eventi elementari che appartengono ad A. In simboli, si ha

$$P(A) = \sum_{A} P(\omega_i), \tag{12.1}$$

dove la somma è estesa a tutti gli eventi elementari in A.

Quando gli eventi elementari sono equiprobabili, cioè quando $P(\omega_i) = 1/N$, $i = 1, 2, \ldots, N$, la (12.1) diventa $P(A) = N_A/N$, dove N_A è il numero di eventi elementari contenuti in A. In questo caso, la probabilità dell'evento A è la proporzione di eventi elementari contenuti in A. Dunque, per calcolare $P(A)$ bisogna determinare N_A ed N. Per questa operazione è spesso utile applicare le regole del calcolo combinatorio, i cui fondamenti sono forniti nel Paragrafo B.3 dell'Appendice B.

ESEMPIO 12.10

Si riprenda l'Esempio 12.1 e, assumendo che la moneta sia bilanciata, si determinino le probabilità degli eventi: "Almeno due volte testa", indicato con A; "Croce nel primo lancio", indicato con B; "Croce nel primo lancio o almeno due volte testa", indicato con $A \cup B$.

Soluzione

Lo spazio campionario è composto da otto eventi elementari tutti con la stessa probabilità, $1/8$, perché la moneta è bilanciata. Con la tecnica del conteggio degli eventi elementari contenuti negli insiemi interessati, si ricava:

$$P(A) = \frac{4}{8} = 0.5,\ P(B) = \frac{4}{8} = 0.5,\ P(A \cup B) = \frac{7}{8} = 0.875.$$

L'ultima probabilità può anche essere ottenuta con la regola della somma, considerando che l'intersezione $A \cap B$ contiene solo un evento elementare. Si ha

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{4}{8} + \frac{4}{8} - \frac{1}{8} = \frac{7}{8}.$$

■

ESEMPIO 12.11

Tra i 194 studenti di economia aziendale che frequentano il corso di statistica, 150 studenti hanno fatto l'esame di matematica, 112 hanno fatto quello di microeconomia e 89 li hanno fatti entrambi. Si calcoli la probabilità che uno studente estratto a sorte abbia fatto matematica oppure microeconomia o entrambi.

Soluzione

Sia A l'evento "Lo studente ha fatto matematica" e B l'evento "Lo studente ha fatto microeconomia". Si deve trovare la probabilità di $A \cup B$. I casi favorevoli al verificarsi dell'evento $A \cup B$ sono

$$150 + 112 - 89 = 173.$$

Ciò perché gli 89 studenti che hanno fatto sia matematica che microeconomia sono già inclusi sia tra gli studenti che hanno fatto matematica sia tra quelli che hanno fatto microeconomia. Si ha allora

$$P(A \cup B) = \frac{173}{194} = 0.8918.$$

Allo stesso risultato si arriva applicando la regola della somma:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{150}{194} + \frac{112}{194} - \frac{89}{194} = 0.8918.$$

■

ESEMPIO 12.12

Da un'urna contenente dieci palline rosse e sette bianche si estraggono tre palline. Si determini la probabilità che le palline estratte siano tutte rosse nell'ipotesi che l'estrazione sia con ripetizione. Si calcoli la medesima probabilità nell'ipotesi che l'estrazione sia senza ripetizione.

Soluzione

Sia A l'evento "Estrazione di tre palline rosse". Se le palline sono estratte con ripetizione, è facile vedere che il numero di eventi elementari che compongono lo spazio campionario è[1]

$$N = 17^3 = 4913.$$

Allo stesso modo, il numero di eventi elementari che formano l'evento A è dato da

$$N_A = 10^3 = 1000.$$

[1] Nel linguaggio del calcolo combinatorio, si tratta del numero delle disposizioni con ripetizione di 17 elementi presi a 3 a 3 (si veda la Sezione B.3 dell'Appendice B).

La probabilità dell'evento A è pertanto

$$P(A) = \frac{N_A}{N} = \frac{1000}{4913} = 0.2035.$$

Se le palline sono estratte senza ripetizione, il numero di palline nell'urna è 17 alla prima estrazione, 16 alla seconda e 15 alla terza. Perciò, il numero di eventi elementari che compongono lo spazio campionario è

$$N = 17 \cdot 16 \cdot 15 = 4080,$$

e il numero di eventi elementari contenuti in A è

$$N_A = 10 \cdot 9 \cdot 8 = 720.$$

Ne segue la probabilità

$$P(A) = \frac{N_A}{N} = \frac{720}{4080} = 0.1765.$$

■

ESEMPIO 12.13

Uno studente intende acquistare due *pen drive* da 64 GB. Nel negozio ci sono sei *pen drive* da 64 GB, due della marca H e quattro della marca J. Poiché la qualità e i prezzi dei prodotti disponibili sono identici e poiché lo studente non è interessato a una marca specifica, sceglie i due oggetti a caso. Qual è la probabilità dell'evento A: "Lo studente acquista un esemplare della marca H e uno della marca J"?

Soluzione

Indicando con H_1 e H_2 gli esemplari della marca H e con J_1, J_2, J_3 e J_4 gli esemplari della marca J, lo spazio campionario Ω è composto dai seguenti 15 eventi elementari:

$$\Omega = \{H_1H_2,\ H_1J_1,\ H_1J_2,\ H_1J_3,\ H_1J_4,\ H_2J_1,\ H_2J_2,\ H_2J_3,\ H_2J_4,\ J_1J_2,\ J_1J_3,\ J_1J_4,\ J_2J_3,\ J_2J_4,\ J_3J_4\}.$$

Gli eventi elementari contenuti nell'evento A ("Lo studente acquista un esemplare della marca H e uno della marca J") sono otto. La probabilità di A è data allora da

$$P(A) = \frac{8}{15} = 0.53.$$

Si noti che il numero totale degli eventi elementari dell'esperimento si può calcolare anche tramite le regole del calcolo combinatorio, in particolare, con la formula che dà il numero delle combinazioni di sei elementi presi a due a due (si veda il Paragrafo B.3 dell'Appendice B):

$$\frac{6!}{2!(6-2)!} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(2 \cdot 1) \cdot (4 \cdot 3 \cdot 2 \cdot 1)} = 15.$$

■

ESEMPIO 12.14

Al concorso per l'assegnazione di quattro borse di studio bandite da un istituto bancario partecipano sette uomini e tre donne. Nell'ipotesi che i candidati abbiano titoli equivalenti e che la commissione proceda mediante estrazione a sorte, si calcoli la probabilità che almeno una donna riceva la borsa di studio.

Soluzione

Indicando con A l'evento "Almeno una donna riceve la borsa di studio", conviene calcolare prima la probabilità di $\bar{A}$ (che nessuna donna riceva la borsa di studio) e applicare poi la regola dell'evento complementare. Il numero di casi favorevoli al verificarsi di $\bar{A}$ è dato dal numero delle combinazioni che si possono formare con sette elementi presi a quattro a quattro:

$$\frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(4 \cdot 3 \cdot 2 \cdot 1) \cdot (3 \cdot 2 \cdot 1)} = 35.$$

Il numero di casi possibili è il numero delle combinazioni che si possono formare con 10 elementi presi a quattro a quattro:

$$\frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(4 \cdot 3 \cdot 2 \cdot 1) \cdot (6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)} = 210.$$

La probabilità dell'evento $\bar{A}$ è perciò

$$P(\bar{A}) = \frac{35}{210},$$

da cui

$$P(A) = 1 - P(\bar{A}) = 1 - \frac{35}{210} = \frac{175}{210}.$$

■

12.5 Probabilità condizionata

Dati due eventi A e B di Ω, interessa talvolta la probabilità di B, supponendo che A si sia già verificato. Con riferimento all'esperimento dei tre lanci di una moneta, ci si può chiedere, per esempio, quale sia la probabilità dell'evento "Croce nel primo lancio" (evento B), ammesso che si sia verificato l'evento "Almeno due volte testa" (evento A). Questa probabilità, che verrà indicata con $P(B|A)$, è chiamata *probabilità condizionata*.

DEFINIZIONE 12.5

Se A e B sono due eventi dello spazio campionario Ω e $P(A) > 0$, la probabilità condizionata di B dato A è

$$P(B|A) = \frac{P(A \cap B)}{P(A)}. \tag{12.2}$$

ESEMPIO 12.15

Si riprenda l'Esempio 12.1 e si trovi la probabilità dell'evento "Croce nel primo lancio", ammesso di sapere che si è verificato l'evento "Almeno due volte testa".

Soluzione

Siano A e B gli eventi "Almeno due volte testa" e "Croce nel primo lancio". Si ha allora:

$$P(A \cap B) = \frac{1}{8},\ P(A) = \frac{4}{8}.$$

Ne segue che

$$P(B\,|\,A) = \frac{1/8}{4/8} = 0.25.$$

■

ESEMPIO 12.16

Un'urna contiene 15 palline bianche e 8 nere. Si trovi la probabilità che in due estrazioni senza ripetizione si estragga una pallina bianca nella seconda estrazione (evento B), ammesso che nella prima sia stata estratta una pallina bianca (evento A).

Soluzione

Le probabilità di A e $A \cap B$ sono

$$P(A) = \frac{15}{23} \quad \text{e} \quad P(A \cap B) = \frac{15 \cdot 14}{23 \cdot 22},$$

si ha allora

$$P(B\,|\,A) = \frac{(15 \cdot 14)/(23 \cdot 22)}{15/23} = 0.6364.$$

■

Dalla (12.2) si trova che la probabilità dell'evento $A \cap B$ può essere scritto nella forma

$$P(A \cap B) = P(B\,|\,A)P(A). \tag{12.3}$$

La (12.3) è nota come regola del prodotto. Si noti che, se $P(B) > 0$, scambiando i ruoli di A e B, si può scrivere

$$P(A \cap B) = P(A\,|\,B)P(B).$$

La regola del prodotto può essere generalizzata al caso di k eventi. Per esempio nel caso di tre eventi, A, B e C, con $P(B \cap C) > 0$, si ha

$$\begin{aligned} P(A \cap B \cap C) &= P[A \cap (B \cap C)] \\ &= P(A\,|\,B \cap C) \cdot P(B \cap C) \\ &= P(A\,|\,B \cap C) \cdot P(B\,|\,C) \cdot P(C). \end{aligned}$$

ESEMPIO 12.17

Un'urna contiene 15 palline bianche e 8 nere. Si considerino tre estrazioni senza ripetizione. Si disegni l'albero degli eventi e si individui lo spazio campionario. Si determini poi la probabilità di ogni singolo evento elementare.

Soluzione

La Figura 12.7 mostra l'albero degli eventi e lo spazio campionario Ω. Ad eccezione delle probabilità associate ai due rami che escono dalla radice, tutte le altre probabilità indicate sui rami dell'albero sono probabilità condizionate che alla successiva estrazione esca una pallina del colore indicato alla destra del ramo, dato il risultato delle estrazioni avvenute fino a quel punto. Per esempio, la probabilità condizionata di osservare una pallina bianca alla terza estrazione, dato che nelle prime due estrazioni sono state estratte due palline bianche è

$$P(\text{la terza pallina è bianca}\,|\,\text{le prime due palline sono bianche}) = 13/21.$$

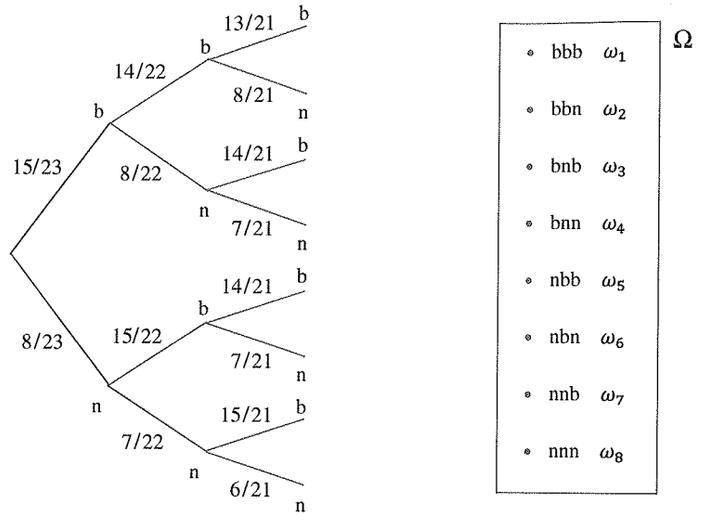


Figura 12.7
Albero degli eventi (con le probabilità condizionate associate a ogni ramo) e spazio campionario.

Allora per trovare le probabilità dei singoli eventi elementari $\omega_1, \omega_2, \ldots, \omega_8$, basta moltiplicare le probabilità sui rami seguendo la regola del prodotto. Per esempio, la probabilità del primo evento elementare è data da

$$\begin{aligned} P(\omega_1) &= P(1^{\text{a}}\text{ bianca},\ 2^{\text{a}}\text{ bianca},\ 3^{\text{a}}\text{ bianca}) \\ &= P(3^{\text{a}}\text{ bianca}\,|\,2^{\text{a}}\text{ bianca}, 1^{\text{a}}\text{ bianca}) \cdot P(2^{\text{a}}\text{ bianca}\,|\,1^{\text{a}}\text{ bianca}) \cdot P(1^{\text{a}}\text{ bianca}) \\ &= (13/21) \cdot (14/22) \cdot (15/23) = 2730/10\,626 = 0.257. \end{aligned}$$

Allo stesso modo, è possibile determinare le probabilità di tutti gli altri eventi elementari ottenendo:

$$P(\omega_2) = 0.158,\ P(\omega_3) = 0.158,\ P(\omega_4) = 0.079,$$
$$P(\omega_5) = 0.158,\ P(\omega_6) = 0.079,\ P(\omega_7) = 0.079,\ P(\omega_8) = 0.032.$$

■

12.6 Indipendenza tra eventi

Un concetto fondamentale nel calcolo delle probabilità è quello di indipendenza tra eventi.

DEFINIZIONE 12.6

Due eventi A e B si dicono indipendenti se

$$P(A \cap B) = P(A)P(B). \tag{12.4}$$

L'indipendenza tra due eventi può essere definita in modo equivalente tramite le probabilità condizionate $P(B\,|\,A)$ e $P(A\,|\,B)$, come indicato nella Proposizione 12.1.

PROPOSIZIONE 12.1

Dati due eventi A e B, con $P(A) > 0$ e $P(B) > 0$, le seguenti tre condizioni sono equivalenti:

(a) $P(A \cap B) = P(A)P(B)$;

(b) $P(B \mid A) = P(B)$;

(c) $P(A \mid B) = P(A)$.

Si dimostra anzitutto che la condizione (a) implica la condizione (b). A questo fine, in forza della regola del prodotto, si ha

$$P(A \cap B) = P(A)P(B \mid A) = P(A)P(B),$$

dove la seconda uguaglianza è vera se e solo se $P(B \mid A) = P(B)$, come si voleva dimostrare. Viceversa, la condizione (b) permette di scrivere

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = P(B),$$

da cui si deduce che la seconda uguaglianza è valida se e solo se $P(A \cap B) = P(A)P(B)$. Dunque, la condizione (b) implica la condizione (a).

Procedendo allo stesso modo, si può dimostrare che la condizione (c) implica la condizione (a) e viceversa. ■

È opportuno mettere in luce il significato della Proposizione 12.1. La condizione (b) stabilisce che, se A e B sono indipendenti, la probabilità di B non viene alterata dall'informazione che l'evento A si è verificato. Reciprocamente, la condizione (c) indica che, se A e B sono indipendenti, la probabilità dell'evento A non viene alterata dall'informazione che l'evento B si è verificato. La Proposizione 12.1 mostra anche come la relazione di indipendenza tra due eventi sia simmetrica.

ESEMPIO 12.18

Si determini la probabilità che in due lanci successivi di un dado regolare appaia due volte la faccia 1.

Soluzione

Siano A l'evento "Faccia 1 nel primo lancio" e B l'evento "Faccia 1 nel secondo lancio". Sulla base di considerazioni fisiche, si può ammettere che i due eventi sono indipendenti. La probabilità di $A \cap B$ è pertanto

$$P(A \cap B) = P(A)P(B) = \frac{1}{6} \cdot \frac{1}{6}.$$

Si arriva allo stesso risultato assegnando la probabilità direttamente all'evento $A \cap B$. Invero, considerando che tale insieme contiene un solo elemento e che lo spazio campionario è costituito da 36 eventi elementari equiprobabili, si ha $P(A \cap B) = 1/36$. ■

ESEMPIO 12.19

Le prime due domande di un test per la valutazione del profitto in statistica sono:

1. Gli eventi "Eleggere il Presidente della Repubblica" e "Eleggere a Presidente della Repubblica un candidato di sesso femminile" sono incompatibili.

   Vero □ Falso □

2. La probabilità di estrarre una pallina rossa da un'urna contenente 10 palline bianche:

   a) non è calcolabile;

   b) è uguale a 0;

   c) è uguale a 1;

   d) nessuna delle precedenti risposte è corretta.

Si voglia calcolare la probabilità che uno studente impreparato, rispondendo a caso, dia la risposta corretta a entrambe le domande.

Soluzione

Gli eventi "Risposta corretta alla prima domanda" (evento A) e "Risposta corretta alla seconda domanda" (evento B) si possono assumere come indipendenti. La probabilità di A è 0.5, la probabilità di B è 0.25. La probabilità che lo studente risponda correttamente a entrambe le domande è, in base alla (12.4),

$$P(A \cap B) = P(A)P(B) = 0.5 \cdot 0.25 = 0.125.$$

Un modo alternativo per risolvere il problema è quello di enumerare tutti gli eventi elementari (equiprobabili) dello spazio campionario. Indicando con v e f le risposte "Vero" e "Falso" alla prima domanda, gli esiti possibili dell'esperimento sono:

$$(v, a), (v, b), (v, c), (v, d), (f, a), (f, b), (f, c), (f, d).$$

Vi è un solo caso favorevole al verificarsi dell'evento $A \cap B$, l'evento elementare (f, b). Si ha pertanto

$$P(A \cap B) = \frac{1}{8} = 0.125.$$

■

ESEMPIO 12.20

Da un lotto di 5000 confezioni di spaghetti da 0.5 Kg ne vengono estratte a sorte due con ripetizione per controllare se il peso effettivo è inferiore a quello dichiarato. Sapendo che il lotto contiene 110 confezioni difettose (con peso inferiore a 0.5 Kg), si determini la probabilità che entrambe le confezioni siano difettose.

Soluzione

Siano A l'evento "Confezione difettosa nella prima estrazione" e B l'evento "Confezione difettosa nella seconda estrazione". I due eventi sono indipendenti in quanto l'estrazione è con ripetizione. Si ha allora

$$P(A \cap B) = P(A)P(B) = \frac{110}{5000} \cdot \frac{110}{5000} = \frac{12\,100}{25\,000\,000} = 0.000484.$$

■

La definizione di indipendenza può essere estesa al caso di tre o più eventi. Qui si dà la definizione esplicita per il caso di tre eventi.

DEFINIZIONE 12.7

I tre eventi A, B e C si dicono indipendenti se sono soddisfatte le condizioni che seguono:

$$P(A \cap B) = P(A)P(B), \quad P(A \cap C) = P(A)\,P(C), \quad P(B \cap C) = P(B)P(C),$$
$$P(A \cap B \cap C) = P(A)P(B)P(C).$$

12.7 Formula di Bayes

Si supponga di avere due urne: la prima, U_1, contenente 4 palline bianche e 6 nere; la seconda, U_2, contenente 3 palline bianche e 5 nere. Si estragga a sorte un'urna e poi una pallina dall'urna selezionata. Ammesso che la pallina estratta sia di colore bianco, ci si può chiedere: qual è la probabilità che essa provenga dall'urna U_1, se le due urne hanno la stessa probabilità di essere selezionate?

Problemi simili si presentano quando un evento A può essere visto come il risultato ("effetto") di uno tra k possibili eventi ("cause"), $C_1, C_2, \ldots, C_k$, che costituiscono una partizione dello spazio campionario, e interessa calcolare la probabilità di C_i, ammesso che l'evento A si sia verificato.

PROPOSIZIONE 12.2

Si consideri la partizione dello spazio campionario Ω formata dai k eventi $C_1, C_2, \ldots, C_k$. Allora, per qualsiasi evento A tale che $P(A) > 0$, la probabilità di C_i, ammesso che A si sia verificato, è data da

$$P(C_i \,|\, A) = \frac{P(C_i)P(A \,|\, C_i)}{\sum_{j=1}^{k} P(C_j)P(A \,|\, C_j)}. \tag{12.5}$$

Per dimostrare la Proposizione 12.2, bisogna considerare che $P(C_i \,|\, A)$ è una probabilità condizionata che, in forza della Definizione 12.5, può essere scritta nella forma

$$P(C_i \,|\, A) = \frac{P(C_i \cap A)}{P(A)}. \tag{12.6}$$

Poi, sapendo che $C_1, C_2, \ldots, C_k$ formano una partizione dello spazio campionario, si può scrivere

$$A = A \cap \Omega = A \cap (C_1 \cup C_2 \cup \cdots \cup C_k) = (A \cap C_1) \cup (A \cap C_2) \cup \cdots \cup (A \cap C_k),$$

dove $A \cap C_1, A \cap C_2, \ldots, A \cap C_k$ sono eventi disgiunti, in quanto sono disgiunti gli eventi $C_1, C_2, \ldots, C_k$ (si osservi la Figura 12.8). Ne segue che

$$\begin{aligned} P(A) &= P[(A \cap C_1) \cup (A \cap C_2) \cup \cdots \cup (A \cap C_k)] \\ &= P(A \cap C_1) + P(A \cap C_2) + \cdots + P(A \cap C_k) \\ &= P(C_1)P(A \,|\, C_1) + P(C_2)P(A \,|\, C_2) + \cdots + P(C_k)P(A \,|\, C_k), \end{aligned}$$

dove l'ultima riga è ottenuta applicando la regola del prodotto a ogni termine della riga precedente. Questo modo di scrivere la probabilità di un evento va sotto il nome di formula delle probabilità totali.

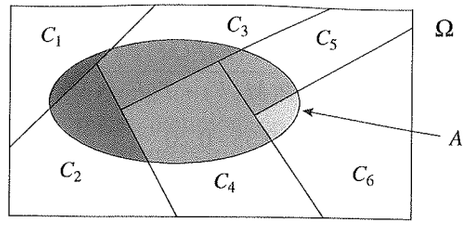


Figura 12.8
Partizione di A per mezzo di $C_1, C_2, \ldots, C_k$.

Poiché $P(C_i \cap A) = P(C_i)\,P(A \,|\, C_i)$, sostituendo a $P(A)$ al denominatore della (12.6) l'espressione finale nelle identità precedenti, si ottiene

$$P(C_i \,|\, A) = \frac{P(C_i)\,P(A \,|\, C_i)}{\sum_{j=1}^{k} P(C_j)\,P(A \,|\, C_j)}.$$

La Proposizione 12.2 è dunque dimostrata. ■

La (12.5) è nota come formula di Bayes (dal nome di Thomas Bayes, 1702-1761, che l'ha introdotta). $P(C_i \,|\, A)$ è conosciuta come probabilità a posteriori, mentre $P(C_i)$ è chiamata probabilità a priori della causa C_i.

ESEMPIO 12.21

Si riprenda l'esempio introduttivo. Si hanno due urne: la prima, U_1, contiene 4 palline bianche e 6 nere, la seconda, U_2, contiene 3 palline bianche e 5 nere. Si estrae una pallina da un'urna selezionata a caso. Qual è la probabilità che l'urna selezionata sia U_1, sapendo che è stata estratta una pallina bianca?

Soluzione

Indicando con A, C_1 e C_2 gli eventi: "pallina bianca", "selezione dell'urna U_1" e "selezione dell'urna U_2", si ha

$$\begin{aligned} P(C_1 \,|\, A) &= \frac{P(C_1)P(A \,|\, C_1)}{P(C_1)P(A \,|\, C_1) + P(C_2)P(A \,|\, C_2)} \\ &= \frac{(1/2)(4/10)}{(1/2)(4/10) + (1/2)(3/8)} = 0.52. \end{aligned}$$

■

ESEMPIO 12.22

È noto che in una data popolazione la percentuale dei fumatori è pari al 35%. Si sa anche che il 20% dei fumatori e il 6% dei non fumatori sono affetti da una malattia respiratoria cronica. Si voglia determinare la probabilità che un individuo affetto dalla malattia sia fumatore.

Soluzione

Definiti gli eventi "Fumatore" (F), "Non fumatore" ($\bar{F}$) e "Malato" (M), le informazioni disponibili consentono di scrivere:

$$P(F) = 0.35, \; P(\bar{F}) = 0.65, \; P(M \,|\, F) = 0.20, \; P(M \,|\, \bar{F}) = 0.06.$$

Allora si ha

$$P(F \mid M) = \frac{P(F)P(M \mid F)}{P(F)P(M \mid F) + P(\bar{F})P(M \mid \bar{F})}$$

$$= \frac{0.35 \cdot 0.20}{0.35 \cdot 0.20 + 0.65 \cdot 0.06} = 0.64.$$

■

ESEMPIO 12.23

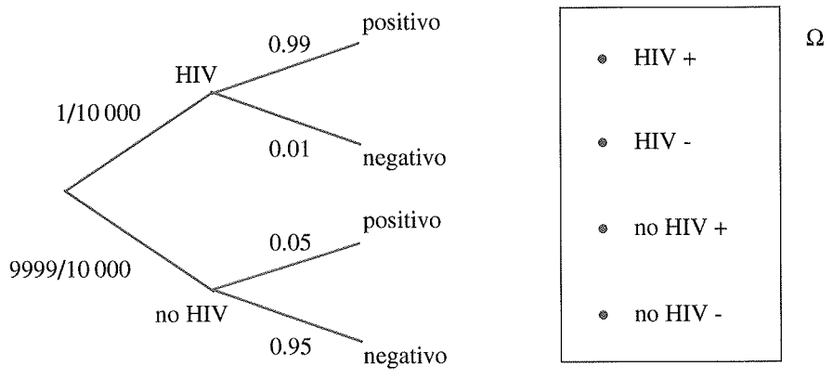
Si supponga che in una popolazione un individuo ogni 10 000 abbia l'HIV. Si estrae a sorte un soggetto da questa popolazione e lo si sottopone a un test per stabilire la presenza del virus. Si sappia che il test è positivo con probabilità 0.99 se l'individuo ha l'HIV e 0.05 se l'individuo è sano. Sapendo che il soggetto è risultato positivo al test, si determini la probabilità che egli abbia effettivamente l'HIV.

Soluzione

È utile organizzare le informazioni disponibili in un albero degli eventi (si veda la Figura 12.9). Indicando con A l'evento "Il soggetto ha l'HIV" e con T l'evento "Il test è positivo", si tratta di applicare la formula di Bayes:

$$P(A \mid T) = \frac{P(A) \cdot P(T \mid A)}{P(A) \cdot P(T \mid A) + P(\bar{A}) \cdot P(T \mid \bar{A})}$$

$$= \frac{(1/10\,000) \cdot 0.99}{(1/10\,000) \cdot 0.99 + (9999/10\,000) \cdot 0.05} = 0.002.$$



ura 12.9
t dell'HIV:
ero degli eventi
pazio campionario.

È utile sottolineare come la probabilità $P(T)$ che il test risulti positivo sia praticamente uguale alla probabilità $P(\bar{A} \cap T)$ di ottenere un "falso positivo". Si ha infatti

$$P(T) = P(A \cap T) + P(\bar{A} \cap T) = P(A) \cdot P(T \mid A) + P(\bar{A}) \cdot P(T \mid \bar{A})$$

$$= 0.000099 + 0.049995 = 0.050094.$$

Ciò spiega il valore molto basso assunto da $P(A \mid T)$.

■

Riassunto

Il capitolo è dedicato all'illustrazione delle regole basilari del calcolo delle probabilità. Si parte dalle idee di esperimento casuale, spazio campionario ed eventi. Vengono poi introdotti gli assiomi della probabilità e discussi le principali interpretazioni di questa (classica, frequentista e soggettivista) e il problema dell'assegnazione delle probabilità agli eventi. La parte centrale del capitolo è dedicata ai concetti di probabilità condizionata (la probabilità di un evento assumendo che un secondo evento si sia già verificato) e di indipendenza tra eventi. Il capitolo si chiude con la presentazione della formula di Bayes.

Parole chiave

albero degli eventi, 246
assiomi della probabilità, 252
esperimento casuale, 246
eventi disgiunti, 249
eventi incompatibili, 249
eventi indipendenti, 259
evento, 248
evento certo, 252
evento complementare, 249
evento composto, 248
evento elementare, 246
evento impossibile, 249
evento intersezione, 249
evento semplice, 248
evento unione, 249
formula delle probabilità totali, 262
formula di Bayes, 263
funzione d'insieme, 251
interpretazione classica della probabilità, 253
interpretazione frequentista della probabilità, 253
interpretazione soggettivista della probabilità, 253
legge empirica del caso, 253
misura di probabilità, 252
partizione dello spazio campionario, 251
probabilità a posteriori, 263
probabilità a priori, 263
probabilità condizionata, 257
prova, 246
regola della somma, 252
regola dell'evento complementare, 252
regola del prodotto, 258
spazio campionario, 246
spazio campionario continuo, 248
spazio campionario discreto, 248

Commenti e curiosità

1. Sopra le scoperte dei dadi di Galileo Galilei

Tra il 1612 e il 1623, Galilei, forse su richiesta del Gran Duca di Toscana, risponde al quesito circa la ragione per cui nel gioco del lancio di tre dadi (con 6 facce numerate da 1 a 6) esca più frequentemente 10 che non 9 come somma dei numeri delle facce superiori, nonostante che tali somme si ottengano con lo stesso numero di tripartizioni (6 in entrambi i casi). Usando le sue stesse parole la questione è la seguente: "Che nel gioco de' dadi alcuni punti sieno più vantaggiosi di altri, vi ha la sua ragione assai manifesta, la quale è il poter quelli più facilmente e più frequentemente scoprirsi che questi".

La spiegazione del fenomeno viene fatta tramite l'enumerazione di tutti i casi favorevoli al verificarsi delle somme da 3 a 10 (per le somme da 11 a 18 si avvale del principio della simmetria, per cui 11 si comporta come il 10, il 12 come il 9 e così via), casi favorevoli che sono elencati nella tabella seguente, dove, accanto a ogni tripartizione, viene indicato il numero delle "maniere" (eventi elementari) in cui gli elementi della tripartizione si possono presentare:

10		9		8		7		6		5		4		3	
6.3.1	6	6.2.1	6	6.1.1	3	5.1.1	3	4.1.1	3	3.1.1	3	2.1.1	3	1.1.1	1
6.2.2	3	5.3.1	6	5.2.1	6	4.2.1	6	3.2.1	6	2.2.1	3				
5.4.1	6	5.2.2	3	4.3.1	6	3.3.1	3	2.2.2	1						
5.3.2	6	4.4.1	3	4.2.2	3	3.2.2	3								
4.4.2	3	4.3.2	6	3.3.2	3										
4.3.3	3	3.3.3	1												
	27		25		21		15		10		6		3		1

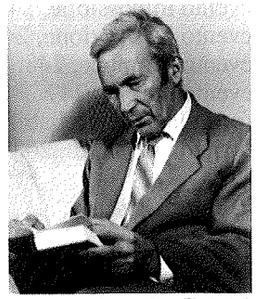
Bayes e la probabilità a posteriori

omas Bayes (1702-1761), figlio di uno dei primi ministri presbiteriani non nformisti ordinati in Inghilterra, svolse i suoi studi di logica e teologia pres- l'Università di Edimburgo, non essendogli stato consentito di studiare né Oxford né a Cambridge. Matematico e ministro presbiteriano anche lui, de- la sua fama ai suoi studi nel campo della matematica e della filosofia. li è soprattutto noto nella statistica per la formula che porta il suo no-, contenuta nel volume (pubblicato dopo la sua morte) dal titolo: *Essay wards Solving a Problem in the Doctrine of Chances*. La formula e il punto di vista che sottende è a base di un approccio all'inferenza statistica noto come statistica Bayesiana.

Legge empirica del caso: gli esperimenti di Buffon e Pearson

legge empirica del caso – per la quale la frequenza relativa con cui un evento si verifica in una serie prove ripetute un gran numero di volte, nelle stesse condizioni, tende ad assumere un valore prossimo a probabilità dell'evento – è stata oggetto di esperimenti nel passato. Si ricordano qui gli esperimenti G. L. Buffon (1707-1788) e di K. Pearson (1857-1936). Il primo in 4040 lanci di una moneta regolare servò 2048 volte testa, con una frequenza relativa pari a 0.5069; il secondo in 24 000 lanci di una moneta golare ottenne testa 12 012 volte, con una frequenza relativa pari a 0.5005.

Kolmogorov e l'assiomatizzazione del calcolo delle probabilità



Sovfoto/Universal Images Group/Shutterstock

drej Nikolaevič Kolmogorov (1903-1987) è stato un matematico russo. Pro-ssore di teoria della probabilità, può essere considerato uno dei più grandi pro-bilisti di tutti i tempi. Fondatore dell'impostazione assiomatica del calcolo del-probabilità, esposta nel volume *Foundations of the Theory of Probability* del 33, diede innumerevoli contributi di altissimo livello in questa disciplina. In tistica, importante è il suo contributo nell'ambito dei test per la verifica del-potesi che un campione casuale provenga da una popolazione di forma nota. lmogorov dedicò gran parte della sua vita all'educazione dei più giovani: egli ndò anche un collegio per studenti matematicamente dotati, per i quali teneva zioni e organizzava gite. Ci sono vari aneddoti che lo rendono una figura affa-nante. Uno dei suoi sport preferiti era nuotare nei fiumi in disgelo.

de Finetti e la probabilità soggettiva

uno de Finetti (1906–1985) è stato uno dei maggiori matematici italiani del No-cento. Dopo la laurea in matematica applicata conseguita presso l'Università di ilano, svolse attività presso l'ufficio matematico del neonato Istat (Istituto Nazio-le di Statistica) e successivamente come attuario e statistico presso un'importante cietà assicurativa italiana. Vincitore della cattedra universitaria di matematica fi-nziaria, fu docente prima presso l'Università di Trieste e poi presso l'Università Roma. Negli anni settanta, come responsabile di una testata giornalistica, a causa lla pubblicazione di articoli a favore dell'obiezione di coscienza, venne arrestato esso la sede dell'*Accademia dei Lincei* all'uscita dalla cerimonia di inaugurazione ll'anno accademico, ma subito dopo gli venne revocato il mandato di arresto.

Gli effetti del suo spirito critico e anticonformista si notarono anche in ambito accademico, dove mo-rò la sua piena ostilità nei confronti dei burocrati. I suoi contributi scientifici riguardano soprattutto la obabilità. Egli è stato uno dei fautori della concezione soggettivista della probabilità, secondo la quale: a probabilità non è nient'altro che il grado di fiducia (speranza, timore) nel fatto che qualcosa di atteso emuto, sperato o indifferente) si verifichi e risulti vero". I suoi studi hanno dato vita a una nuova impo-azione della probabilità, la cosiddetta impostazione soggettivista. La sua produzione scientifica è nota a vello mondiale. La maggior parte dei manoscritti, delle lettere, degli appunti e alcuni volumi della biblio-ca privata di Bruno de Finetti sono stati acquistati dall'Università di Pittsburgh (USA).

6. Fallacia dell'accusatore

Semplificando un po', si può dire che i test del DNA utilizzati attualmente ai fini investigativi non analizzano l'intera sequenza del DNA, ma solo alcune caratteristiche di questa sequenza. La prima conseguenza di ciò è che mentre si può ritenere che il DNA individui univocamente un soggetto, il test del DNA non lo fa. In altre parole, più soggetti possono avere lo stesso risultato al test del DNA.

Per illustrare il problema, si fanno le seguenti assunzioni:

(a) il test del DNA, se applicato ripetutamente allo stesso soggetto, fornisce sempre lo stesso risultato;

(b) la probabilità che un individuo scelto a caso dalla popolazione abbia lo stesso DNA (ossia lo stesso risultato al test del DNA) di un prefissato individuo è pari a circa uno su due milioni;

(c) un individuo ha commesso un crimine e sulla scena del delitto viene trovato materiale organico contenente DNA appartenente al criminale;

(d) per motivi del tutto casuali, un soggetto viene sottoposto al test del DNA e si scopre che ha lo stesso DNA (sempre nel senso del risultato al test) trovato sulla scena del crimine.

In passato, si è spesso giunti alla conclusione che c'è un'altissima probabilità che il soggetto sottoposto al test sia colui che ha effettivamente commesso il crimine. Si vuole mostrare che questa conclusione è sbagliata.

Sia N la dimensione della popolazione. Si considerino le due probabilità condizionate:

$$P(\text{DNA corr.} \mid \text{Colpevole}), \quad P(\text{Colpevole} \mid \text{DNA corr.}),$$

dove DNA corr. indica l'evento corrispondenza del DNA con quello trovato sul luogo del crimine. La prima è pari a 1, perché, se si ammette che l'individuo estratto a sorte è colpevole, il suo DNA corrisponderà con certezza a quello rilevato sulla scena del crimine. La seconda è la probabilità che, in una estrazione casuale dal sottoinsieme della popolazione costituito dagli individui il cui DNA coincide con quello trovato sulla scena del crimine, venga selezionato proprio il colpevole. Per l'assunzione (b), la numerosità di questo sottoinsieme è $1 + (N - 1)/2\,000\,000$). Ne segue che

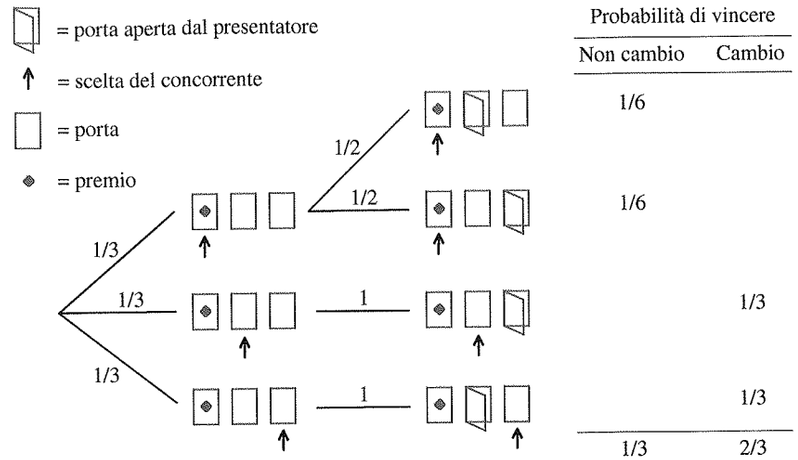
$$P(\text{Colpevole} \mid \text{DNA corr.}) = \frac{1}{1 + (N-1)/2\,000\,000}.$$

Ad esempio, per una popolazione di 60 milioni, tale probabilità è uguale a $1/31 = 0.0323$. Come si vede, si tratta di un numero ben distante dalla "ragionevole certezza" necessaria per una condanna. Dunque, il fatto che il test del DNA di un individuo corrisponda a quello della scena del crimine di per sé non implica che egli sia colpevole con alta probabilità. L'idea che tale corrispondenza sia una prova quasi certa di colpevolezza è basata sul riferimento erroneo alla prima probabilità condizionata, anziché alla seconda. Al riguardo si parla di fallacia dell'accusatore.

7. Dilemma di Monty Hall

Il dilemma di Monty Hall è un problema che ha avuto origine dal gioco televisivo a premi "Let's Make a Deal" andato in onda inizialmente negli Stati Uniti a partire dal 1963 e richiamato più di recente anche nel film "21" del 2008. Il gioco consiste nel mostrare a un concorrente tre porte, dietro una delle quali si nasconde un premio. Dopo che il giocatore ha scelto una delle tre porte, ma non l'ha ancora aperta, il conduttore televisivo (che sa dove si trova il premio) apre una delle altre due porte (stando attento a non aprire la porta che nasconde il premio) e chiede al giocatore se desidera cambiare o meno la propria scelta iniziale, passando all'altra porta ancora chiusa. La soluzione del gioco ha attirato all'inizio molta attenzione in quanto non era ben chiaro quale fosse la strategia migliore da seguire. Anche se ad alcuni potrebbe sembrare controintuitivo, la strategia corretta è quella di cambiare la scelta iniziale. Per capire come questo possa essere vero, si analizzi la Figura 12.10 in cui, per semplificare l'analisi (ma senza nessuna perdita di generalità), si è assunto che il premio sia nascosto dietro la prima porta.

La strategia di non cambiare la porta scelta inizialmente conduce a vincere in due casi, entrambi con probabilità 1/6. D'altra parte, la strategia di cambiare la scelta iniziale conduce a vincere negli altri due casi, aventi entrambi probabilità 1/3. Quindi, la probabilità complessiva di vincere non cambiando la scelta iniziale è pari a 1/3, mentre cambiando è pari a 2/3.



ura 12.10
mma di Monty Hall:
ero degli eventi.

sercizi

12.1 Siano dati i tre eventi: $A = \{x : 0 \le x \le 4\}$, $B = \{x : 3 \le x \le 10\}$, $C = \{x : -1 \le x \le 3\}$. Si determinino gli eventi:

a) $A \cup B \cup C$;

b) $A \cap B \cap C$;

c) $A \cap B \cap \bar{C}$.

12.2 Uno spazio campionario contiene nove eventi elementari, $\omega_1, \omega_2, \ldots, \omega_9$, le cui probabilità sono:

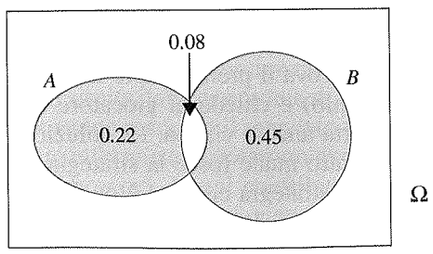
$$P(\omega_1) = P(\omega_2) = 0.08, \quad P(\omega_3) = P(\omega_4) = P(\omega_5) = 0.1,$$
$$P(\omega_6) = P(\omega_7) = 0.2, \quad P(\omega_8) = P(\omega_9) = 0.07.$$

Siano $A = \{\omega_1, \omega_5, \omega_8\}$, $B = \{\omega_2, \omega_5, \omega_8, \omega_9\}$. Si calcolino le probabilità:

a) $P(A)$, $P(B)$, $P(A \cap B)$ e $P(A \cup B)$;

b) $P(\bar{B})$ in due modi: prima in base a $P(B)$, poi direttamente individuando gli eventi elementari che appartengono a $\bar{B}$.

12.3 Si considerino i due eventi A e B rappresentati dai diagrammi di Venn qui sotto, in cui sono indicate le probabilità di alcuni sottoinsiemi di interesse (per esempio, 0.22 è la probabilità dell'intersezione di A e $\bar{B}$):



Si determinino le probabilità degli eventi:

a) A non si verifica;

b) A si verifica e B non si verifica.

12.4 Per raggiungere il pareggio di bilancio un ente può ricorrere a due azioni: aumentare le entrate (evento A), tagliare le spese (evento T). Si esprima a parole il significato dei seguenti eventi:

$$\bar{A}, \quad A \cup T, \quad \bar{T}, \quad A \cup \bar{T}, \quad A \cap T, \quad A \cap \bar{T}.$$

12.5 Per gli eventi A e B si conosce che $P(\bar{A}) = 0.3$, $P(B) = 0.4$, $P(A \cap \bar{B}) = 0.5$. Si trovino le seguenti probabilità:

a) $P(A)$;

b) $P(A \cap B)$;

c) $P(A \cup B)$.

12.6 Dati due eventi incompatibili A e B, tali che $P(A) = 0.35$ e $P(B) = 0.40$, si trovino le seguenti probabilità:

$$P(\bar{A}), \quad P(A \cap B), \quad P(A \cup B), \quad P(\bar{A} \cup \bar{B}), \quad P(\bar{A} \cap \bar{B}).$$

12.7 Le valutazioni di un esperto di borsa circa il rendimento, nel prossimo anno, di una determinata azione del settore industriale sono indicate nella tabella che segue:

Tasso di rendimento %	≤ -5	$(-5, 0]$	$(0, 5]$	$(5, 10]$	> 10
Probabilità	0.02	0.12	0.33	0.47	0.06

Si definiscano i due eventi "Il tasso di rendimento sarà negativo o nullo" (evento A) e "Il tasso di rendimento sarà maggiore di 5" (evento B).

a) Si descriva l'evento complementare di A.

b) Si determini la probabilità di A.

c) Si decriva l'evento intersezione di A e B.

d) Si descriva l'evento unione di A e B.

e) Si determini la probabilità dell'evento unione di A e B.

f) Si determini la probabilità dell'intersezione di A e B.

12.8 Per gli eventi A e B si conosce che: $P(A) = 0.48$, $P(B) = 0.39$, $P(A \cap B) = 0.18$. Si completi la tabella che segue inserendo le appropriate probabilità:

	A	$\bar{A}$	
B			
$\bar{B}$			

Si calcoli la probabilità degli eventi $A \cap \bar{B}$ e $\bar{A} \cap \bar{B}$.

12.9 Un'azienda di elettronica riceve forniture di un certo componente in lotti da 500 pezzi. Sulla base della passata esperienza, le probabilità associate al numero di pezzi difettosi in un lotto sono le seguenti:

N. di pezzi difettosi	0	1	2	3	4 o più
Probabilità	0.32	0.40	0.18	0.07	0.03

Con riferimento a un lotto, si calcoli la probabilità degli eventi di seguito indicati:

a) Il numero dei pezzi difettosi è al massimo 2;

b) Il numero dei pezzi difettosi è 3 o più;

c) Il numero dei pezzi difettosi è 0 o 2;

d) Il numero dei pezzi difettosi è compreso tra 1 e 3, estremi inclusi.

Dati i due eventi "Il numero dei pezzi difettosi è almeno 2" (evento A) e "Il numero dei pezzi difettosi è non più di 3" (evento B), i due eventi sono incompatibili? Si calcoli la probabilità dell'evento intersezione $A \cap B$.

12.10 Delle 80 confezioni di yogurt esposte nel bancone di un supermercato, 10 scadono fra una settimana, 50 fra due settimane e le restanti 20 fra tre settimane. Si calcoli la probabilità che su cinque confezioni scelte a caso 2 scadano fra una settimana, 2 scadano fra due settimane e 1 fra tre settimane.

12.11 Un'urna contiene 100 palline numerate da 1 a 100. Ne vengono estratte 5 con ripetizione. Qual è la probabilità che tra le palline estratte ve ne siano almeno due uguali?

12.12 Da un mazzo di 52 carte da poker se ne estraggono a sorte cinque. Si determini la probabilità che delle cinque carte tre siano assi.

12.13 Si consideri l'esperimento casuale dell'estrazione a sorte consecutivamente di due carte da un mazzo da 40.

a) Si calcoli la probabilità che escano due assi assumendo che la prima carta sia reinserita nel mazzo prima dell'estrazione della seconda.

b) Si ripeta il calcolo del punto precedente assumendo che la prima carta non sia reinserita nel mazzo prima dell'estrazione della seconda.

12.14 Un dado regolare viene lanciato due volte.

a) Qual è la probabilità che l'esito del primo lancio sia cinque se la somma dei punteggi è sette?

b) Qual è la probabilità che l'esito del secondo lancio sia un numero doppio dell'esito del primo lancio?

12.15 I dipendenti di sesso maschile di una casa di cura sono stati classificati secondo il peso corporeo e la pressione sanguigna ottenendo la distribuzione doppia di frequenza relativa che segue:

Pressione sanguigna	Peso corporeo			Totale
	Sovrappeso	Peso normale	Sottopeso	
Iperteso	0.08	0.12	0.03	0.23
Non iperteso	0.14	0.52	0.11	0.77
Totale	0.22	0.64	0.14	1.00

a) Si determini la probabilità che una persona estratta a sorte da questo gruppo abbia la pressione alta.

b) Qual è la probabilità che una persona estratta a sorte da questo gruppo sia sovrappeso e anche ipertesa?

c) Una persona, estratta casualmente dal gruppo, è risultata sovrappeso: qual è la probabilità che essa sia anche non ipertesa?

12.16 Si dimostri che, se gli eventi A e B sono indipendenti, allora anche A e $\bar{B}$ sono indipendenti.

12.17 Un dado regolare viene lanciato cinque volte. Si trovi la probabilità la faccia 1 appaia almeno una volta.

12.18 Siano A e B due eventi tali che $P(A) = 0.7$, $P(A \cup B) = 0.8$. Si trovi $P(B)$ nei seguenti casi:

a) A e B sono disgiunti;

b) A e B sono indipendenti;

c) $P(A \mid B) = 0.6$.

12.19 Si consideri una popolazione in cui la percentuale di sieropositivi è pari allo 0.2%. È noto che il test diagnostico dà un risultato corretto nel 95% dei casi se il soggetto è malato, e nel 90% dei casi se il soggetto è sano. Qual è la probabilità che una persona, risultata positiva al test, sia malata?

12.20 Sulla base della passata esperienza, si sa che il 75% degli studenti che provano un certo esame ha una preparazione sufficiente. Si sa inoltre che uno studente preparato ha una probabilità di 0.95 di superare l'esame, mentre uno studente non preparato supera l'esame con probabilità 0.10. Si determini la probabilità che uno studente che ha superato l'esame abbia una preparazione sufficiente.

CAPITOLO

13

Sommario del capitolo

13.1 Introduzione
13.2 Variabili casuali discrete
13.3 Variabili casuali continue
13.4 Ulteriori proprietà delle variabili casuali
Riassunto
Commenti e curiosità
Esercizi

Variabili casuali

13.1 Introduzione

Una variabile casuale è una quantità il cui valore dipende dall'esito di un esperimento casuale. Locuzioni alternative a quella di variabile casuale sono variabile aleatoria e variabile stocastica.

A titolo di esempio, si consideri l'esperimento del triplo lancio di una moneta (si veda l'Esempio 12.1). Si può creare una variabile casuale assegnando a ogni esito dell'esperimento il numero di volte che appare testa, come illustrato nella Figura 13.1.

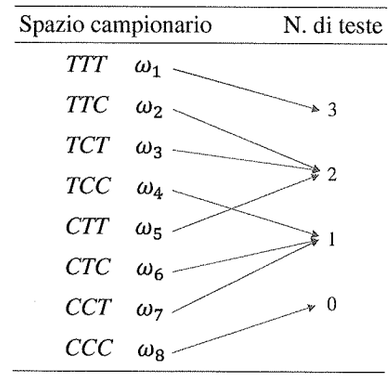


Figura 13.1 Dallo spazio campionario Ω allo spazio numerico nell'esperimento del triplo lancio di una moneta.

Come altro esempio, si consideri l'esperimento del lancio di due dadi (si veda l'Esempio 12.2) e si associ a ogni esito dell'esperimento la somma dei numeri sulle facce superiori, come illustrato nella Figura 13.2.

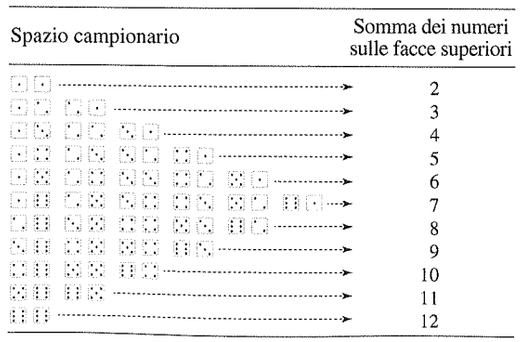


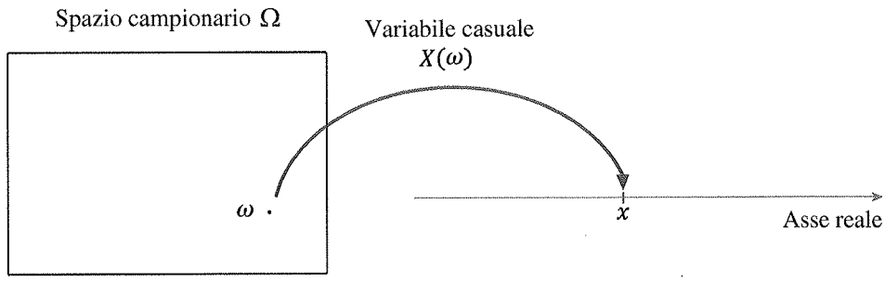
Figura 13.2 Dallo spazio campionario Ω allo spazio numerico nell'esperimento del lancio di due dadi.

L'idea di associare numeri agli eventi elementari o esiti di un esperimento casuale seguendo una regola ben specificata può essere generalizzata come segue.

DEFINIZIONE 13.1

Dato lo spazio campionario Ω di un esperimento casuale, una variabile casuale, $X(\omega)$ – X per brevità – è una funzione definita su Ω che associa un numero reale x a ogni evento elementare ω di Ω.

Con l'introduzione della variabile casuale X si opera la trasformazione degli eventi elementari di Ω in punti dell'asse reale, come illustrato nella Figura 13.3.



ura 13.3
llo spazio campionario Ω
asse reale.

13.2 Variabili casuali discrete

Una variabile casuale X è detta discreta se può assumere un numero finito o un'infinità numerabile di valori. Le variabili casuali rappresentate nelle Figure 13.1 e 13.2 sono chiaramente entrambe discrete.

Indicato con x uno specifico valore della variabile casuale X, la probabilità da assegnare a questo valore è data dalla probabilità che si verifichi uno degli eventi elementari per cui $X = x$.

ESEMPIO 13.1

Si consideri la variabile casuale X rappresentata nella Figura 13.1. Assumendo che la moneta sia bilanciata, si calcoli la probabilità di ogni valore di X.

Soluzione

La variabile casuale assume i valori interi da 0 a 3. Il valore $x = 0$ è associato al solo evento elementare ω_8. Posto $A_0 = \{\omega_8\}$, si ha allora

$$P(X = 0) = P(A_0) = \frac{1}{8}.$$

Il valore $x = 1$ è associato agli eventi elementari ω_4, ω_6 e ω_7. Posto $A_1 = \{\omega_4, \omega_6, \omega_7\}$, si ha allora

$$P(X = 1) = P(A_1) = \frac{3}{8}.$$

Procedendo in modo analogo per gli altri due valori di X, si perviene alla situazione esposta di seguito, in cui ai quattro possibili valori di X sono associate le

rispettive probabilità:

x	0	1	2	3
$P(X = x)$	0.125	0.375	0.375	0.125

■

ESEMPIO 13.2

Assumendo che i dadi siano bilanciati, si associ a ciascun valore della variabile casuale X descritta nella Figura 13.2 l'appropriato livello di probabilità.

Soluzione

La variabile casuale assume i valori che vanno da 2 a 12. Osservando la Figura 13.2, si trova che il valore $x = 2$ è associato a un solo evento elementare, la coppia (1, 1), la cui probabilità è 1/36. Posto $A_2 = \{(1, 1)\}$, si ha allora

$$P(X = 2) = P(A_2) = \frac{1}{36}.$$

Il valore $x = 3$ è associato a due eventi elementari, le coppie (1, 2) e (2, 1). Posto $A_3 = \{(1, 2), (2, 1)\}$, ne segue che

$$P(X = 3) = P(A_3) = \frac{2}{36}.$$

Procedendo allo stesso modo per i rimanenti valori di X, si ottengono le probabilità indicate nella tabella che segue:

x	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

■

In definitiva, la regola con cui si assegna le probabilità al singolo valore x di X può essere così enunciata. Sia A_x l'insieme unione degli eventi elementari ω di Ω per i quali $X(\omega) = x$:

$$A_x = \{\omega : X(\omega) = x\}.$$

La probabilità che $X = x$ è allora data da

$$P(X = x) = P(A_x).$$

L'insieme dei valori che la variabile casuale X può assumere con i rispettivi livelli di probabilità costituisce la distribuzione di probabilità di X. Per una variabile casuale avente k valori, la distribuzione di probabilità può essere presentata in una tabella come quella che segue:

x	x_1	x_2	...	x_k
$P(X = x)$	p_1	p_2	...	p_k

dove $x_1, x_2, \ldots, x_k$ sono i valori che la variabile può assumere e $p_1, p_2, \ldots, p_k$ le rispettive probabilità. La distribuzione di probabilità può essere anche espressa tramite la funzione

$$f(x) = P(X = x), \quad x = x_1, x_2, \ldots, x_k,$$

denominata funzione di probabilità, che gode delle due ovvie proprietà:

$$f(x) \geq 0, \text{ per ogni } x, \quad \sum_x f(x) = 1,$$

dove la sommatoria è estesa a tutti i valori di X.[1]

La distribuzione di probabilità può essere descritta tramite la funzione di ripartizione, $F(x)$, la quale associa a ogni valore x di X la somma delle probabilità corrispondenti a x e a tutti i valori inferiori:

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t).$$

La funzione di ripartizione gode delle proprietà di seguito indicate:

- è definita per ogni x;
- assume valori nell'intervallo $[0, 1]$;
- è una funzione non decrescente in x, cioè, $F(a) \leq F(b)$, per ogni coppia di valori a e b, con $a < b$.[2]

La rappresentazione grafica della funzione di probabilità $f(x)$ si effettua mediante un sistema di assi cartesiani, ponendo sull'asse delle ascisse i valori di X e su quello delle ordinate le corrispondenti probabilità, usualmente raffigurate con aste.

La rappresentazione grafica della $F(x)$ dà luogo invece a un grafico a gradini, analogamente a quanto visto per le distribuzioni di frequenza.

ESEMPIO 13.3

Sia X una variabile casuale discreta descritta dalla funzione di probabilità indicata nella tabella che segue:

x	1	2	3.5	4
$f(x)$	0.1	0.5	0.3	0.1

Si trovi la funzione di ripartizione $F(x)$ di X e si disegnino le funzioni $f(x)$ e $F(x)$.

Soluzione

La funzione di ripartizione $F(x)$ si trova sommando successivamente i valori della funzione di probabilità $f(x)$:

[1] Chiaramente, se X può assumere un'infinità numerabile di valori, $x_1, x_2, \ldots$, la tabella precedente e la definizione di funzione di probabilità vanno modificate di conseguenza.

[2] Ulteriori proprietà matematiche della funzione di ripartizione sono (si veda il Paragrafo B.4 dell'Appendice B): $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$, $\lim_{h \to 0^+} F(x + h) = F(x)$.

$F(x) = 0$, per $x < 1$;

$F(x) = f(1) = 0.1$, per $1 \leq x < 2$;

$F(x) = f(1) + f(2) = 0.1 + 0.5 = 0.6$, per $2 \leq x < 3.5$;

$F(x) = f(1) + f(2) + f(3.5) = 0.1 + 0.5 + 0.3 = 0.9$, per $3.5 \leq x < 4$;

$F(x) = f(1) + f(2) + f(3.5) + f(4) = 0.1 + 0.5 + 0.3 + 0.1 = 1$, per $x \geq 4$.

La Figura 13.4 riporta i grafici di $f(x)$ e $F(x)$ della distribuzione di probabilità in esame.

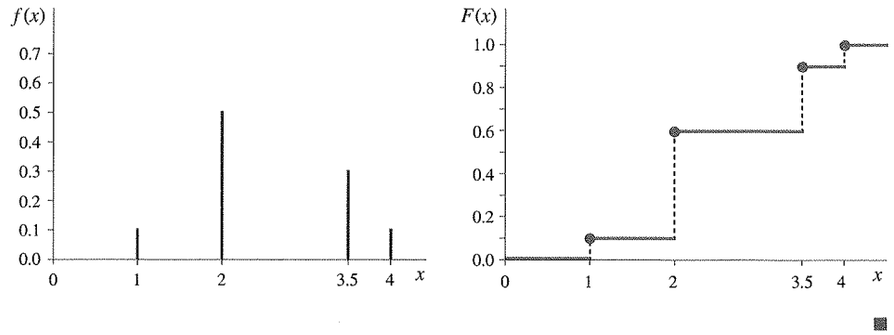


Figura 13.4
Grafici delle funzioni $f(x)$ e $F(x)$ della distribuzione di probabilità dell'Esempio 13.1.

ESEMPIO 13.4

Nella Figura 13.5 è rappresentata la funzione di probabilità e la funzione di ripartizione della distribuzione di probabilità dell'Esempio 13.2.

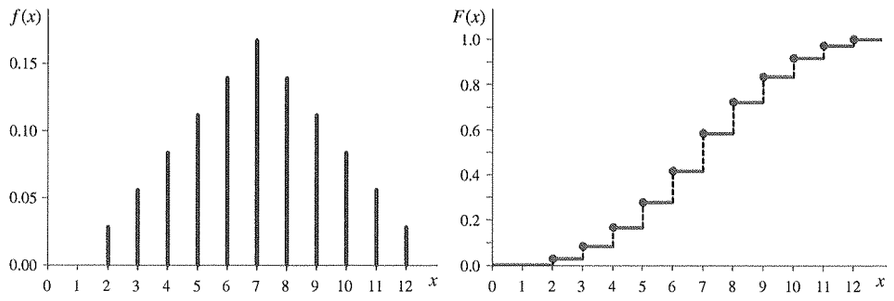


Figura 13.5
Grafici delle funzioni $f(x)$ e $F(x)$ della distribuzione di probabilità dell'Esempio 13.2.

Si noti che la funzione di probabilità $f(x)$ può essere dedotta dalla funzione di ripartizione per differenze successive. Infatti, per una variabile casuale con k valori, $x_1, x_2, \ldots, x_k$, si può scrivere $f(x_1) = F(x_1)$ e $f(x_i) = F(x_i) - F(x_{i-1})$, $i = 2, 3, \ldots, k$. Così nell'Esempio 13.3, si ha:

$f(1) = F(1) = 0.1$;

$f(2) = F(2) - F(1) = 0.6 - 0.1 = 0.5$;

$f(3.5) = F(3.5) - F(2) = 0.9 - 0.6 = 0.3$;

$f(4) = F(4) - F(3.5) = 1 - 0.9 = 0.1$.

13.2.1 Media e varianza

Per una distribuzione di frequenza, la media aritmetica può essere scritta nella forma (4.6) come somma dei prodotti tra modalità e frequenze relative:

$$\mu = x_1 f_1 + x_2 f_2 + \cdots + x_k f_k = \sum_{i=1}^{k} x_i f_i.$$

In modo analogo, la media di una variabile casuale discreta X si ottiene come somma dei prodotti tra i valori di X e le rispettive probabilità.

DEFINIZIONE 13.2

La media o valore atteso di una variabile casuale discreta X, indicata con $\mathrm{E}(X)$ oppure con μ,[3] è il numero dato da

$$\mathrm{E}(X) = \sum_x x f(x), \tag{13.1}$$

dove la somma è estesa a tutti i valori di X.

Per una distribuzione di probabilità, come per una distribuzione di frequenza, la media indica l'ordine di grandezza del fenomeno rappresentato dalla variabile casuale ed è espressa nella stessa unità di misura di questo.

L'espressione "valore atteso" è legata storicamente ai giochi di sorte e allude alla vincita media di un giocatore in un grande numero di giocate. Ad esempio, nella semplice scommessa del lancio di una moneta bilanciata, in cui il giocatore vince un euro se esce testa e zero euro se esce croce, il valore atteso, pari a 0.5, rappresenta il valore medio delle vincite (future) in un grande numero di lanci.

ESEMPIO 13.5

Si calcoli la media della variabile casuale considerata nell'Esempio 13.1.

Soluzione

L'applicazione della (13.1) dà

$$\mathrm{E}(X) = 0 \cdot 0.125 + 1 \cdot 0.375 + 2 \cdot 0.375 + 3 \cdot 0.125 = 1.5.$$

■

ESEMPIO 13.6

In base all'esperienza del recente passato, il ritardo di un autobus del servizio pubblico urbano di una città, a una data fermata, nella fascia oraria dalle 10 alle 12 dei giorni lavorativi, è descritto dalla distribuzione di probabilità:

Ritardo (minuti) x	0	1	2	3	4	5	6	7
Probabilità $f(x)$	0.18	0.23	0.27	0.13	0.08	0.05	0.03	0.03

Si calcoli il valore atteso.

[3] Il simbolo E sta per l'iniziale delle espressioni inglesi *Expectation* oppure *Expected value*.

Soluzione

Applicando la (13.1) alla distribuzione in esame, si ottiene

$$\mu = 0 \cdot 0.18 + 1 \cdot 0.23 + \cdots + 7 \cdot 0.03 = 2.12.$$

■

La media gode della proprietà di linearità: se la variabile casuale X ha media μ_X, la variabile casuale

$$Y = a + bX,$$

dove a e b sono costanti assegnate, ha media

$$\mu_Y = a + b\mu_X. \tag{13.2}$$

La dimostrazione è contenuta nel Paragrafo B.5 dell'Appendice B.

ESEMPIO 13.7

Con riferimento all'Esempio 13.6, posto $a = 3$ e $b = 2$, si determini il valore atteso della variabile casuale $Y = 3 + 2X$.

Soluzione

L'applicazione della (13.2) dà

$$\mu_Y = \mathrm{E}(3 + 2X) = 3 + 2 \cdot 2.12 = 7.24.$$

■

Un'altra costante caratteristica di rilievo è la varianza. Anche in questo caso, è utile il parallelo con le distribuzioni di frequenza (si veda il Capitolo 5): per queste la varianza è definita dalla somma dei quadrati degli scostamenti delle modalità dalla media aritmetica, moltiplicati per le rispettive frequenze relative. Per una variabile casuale discreta, la struttura è la stessa, salvo la sostituzione delle frequenze relative con le probabilità.

DEFINIZIONE 13.3

Sia X una variabile casuale discreta con media μ. La varianza di X, indicata con $\mathrm{Var}(X)$ o σ^2, è data da

$$\mathrm{Var}(X) = \sum_x (x - \mu)^2 f(x), \tag{13.3}$$

dove la sommatoria è estesa a tutti i valori di X.

La (13.3) non è altro che il valore atteso del quadrato della variabile casuale scarto dalla media, $(X - \mu)^2$, e può anche essere posta nella forma

$$\sigma^2 = \sum_x x^2 f(x) - \mu^2,$$

come è dimostrato nel Paragrafo B.5 dell'Appendice B.

La varianza è sempre non negativa ed è nulla se e solo se X assume un unico valore con probabilità 1. Il suo valore è tanto più grande quanto maggiore è la dispersione dei valori di X attorno alla media.

Si può dimostrare (si veda il Paragrafo B.5 dell'Appendice B) che la varianza della variabile casuale

$$Y = a + bX,$$

dove a e b sono costanti assegnate, è data da

$$\sigma_Y^2 = b^2\sigma_X^2, \tag{13.4}$$

essendo σ_X^2 e σ_Y^2 le varianze di X e di Y.

La radice quadrata della varianza va sotto il nome di deviazione standard. Si tratta di una misura di dispersione dei valori della variabile casuale X espressa nella stessa unità di misura di questa.

ESEMPIO 13.8

Si calcoli la varianza della distribuzione di probabilità descritta nell'Esempio 13.1.

Soluzione

Si ricordi che la media di questa distribuzione di probabilità è $\mu = 1.5$. Si ha pertanto

$$\sigma^2 = (0 - 1.5)^2\,0.125 + (1 - 1.5)^2\,0.375 + \cdots + (3 - 1.5)^2\,0.125 = 0.75.$$

La deviazione standard è $\sigma = \sqrt{0.75} = 0.87$. ■

ESEMPIO 13.9

Sia X la variabile casuale di cui all'Esempio 13.1. Si voglia determinare la varianza della variabile casuale

$$Y = -1 + 4X.$$

Soluzione

Sulla base della (13.4), la varianza di Y è data da

$$\sigma_Y^2 = 16 \cdot \sigma_X^2,$$

e, essendo che $\sigma_X^2 = 0.75$ (si veda l'Esempio 13.8), si ricava

$$\sigma_Y^2 = 16 \cdot 0.75 = 12.$$

■

13.3 Variabili casuali continue

Si riprenda l'Esempio 12.4 relativo alla durata di una lampadina elettrica. In questo caso, lo spazio campionario dell'esperimento è $\Omega = \{\omega : 0 \le \omega \le T\}$. Se si considera la variabile casuale definita dalla funzione $X = \omega$, questa può assumere un qualsiasi valore nell'intervallo $[0, T]$ dell'asse reale, dove T è una costante nota.

In prima approssimazione, si può dire che una variabile casuale è continua se può assumere tutti i valori di un determinato intervallo (o insieme di intervalli) di numeri reali. Per la definizione formale, conviene partire dalla funzione di ripartizione che, come nel caso discreto, per ogni assegnato valore x, dà la probabilità che la variabile casuale assuma un valore minore o uguale a x: $F(x) = P(X \le x)$.

DEFINIZIONE 13.4

Una variabile casuale X è detta continua se esiste una funzione $f(x)$ tale che la funzione di ripartizione $F(x) = P(X \le x)$ è data dall'area sottesa a $f(x)$ a sinistra di x. La funzione $f(x)$ è denominata funzione di densità di probabilità o funzione di densità.[4]

La relazione appena stabilita tra funzione di ripartizione e funzione di densità è illustrata nella Figura 13.6.

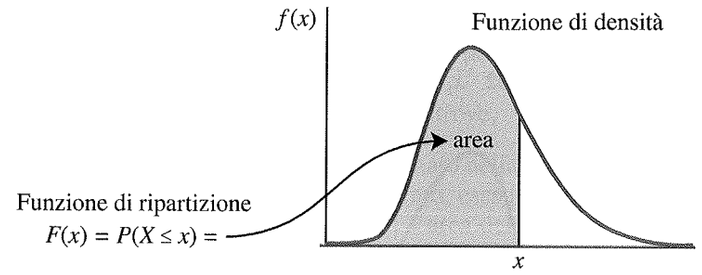


Figura 13.6
Relazione tra funzione di ripartizione e funzione di densità.

La funzione di densità assolve un compito simile a quello che la funzione di probabilità svolge per le variabili casuali discrete.

Per chiarire questo punto, si supponga che la variabile casuale continua X assuma valori tra l_1 e l_2 e si immagini di suddividere l'intervallo (l_1, l_2) in N piccoli sottointervalli

$$(x_1 - \delta/2, x_1 + \delta/2), (x_2 - \delta/2, x_2 + \delta/2), \ldots, (x_N - \delta/2, x_N + \delta/2),$$

di lunghezza

$$\delta = (l_2 - l_1)/N$$

e valori centrali $x_1, x_2, \ldots, x_N$. Siano $p_1, p_2, \ldots, p_N$ le probabilità che X assuma valori nel primo, nel secondo, ..., nell'N-esimo sottointervallo.

Se si rappresenta questa distribuzione di probabilità con un istogramma, si ottiene il grafico della Figura 13.7, dove il generico rettangolo ha area uguale a p_i e altezza uguale a p_i/δ, quantità interpretabile come densità di probabilità per unità di ampiezza. Se, ad esempio, X è il peso di neonati in grammi e δ è uguale a 50 g, l'area del singolo rettangolo è la probabilità che il peso di un neonato sia all'interno dell'intervallo alla base del rettangolo, mentre l'altezza rappresenta la densità di probabilità per unità di ampiezza. In questo modo, viene descritta la maggiore o minore "densità di probabilità" al variare di X in (l_1, l_2).

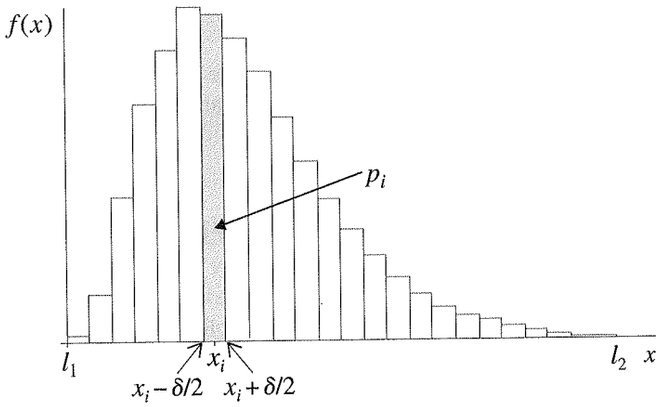
È evidente che l'area complessiva degli N rettangoli dell'istogramma è uguale a 1, perché è la probabilità che X assuma un valore qualsiasi nell'intervallo (l_1, l_2). Se si aumenta N o, equivalentemente, si diminuisce δ, il profilo del grafico tende a una curva come quella indicata nella Figura 13.8, che assume la veste e il significato di funzione di densità.

[4] La relazione matematica tra funzione di ripartizione e funzione di densità è espressa da

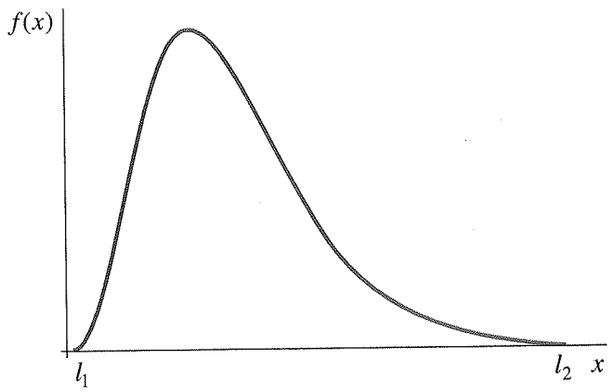
$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t,$$

da cui si deduce che $\dfrac{\mathrm{d}}{\mathrm{d}x}F(x) = f(x)$, per ogni punto in cui $f(x)$ è continua.

ura 13.7
ogramma di probabilità
na variabile casuale
tinua.



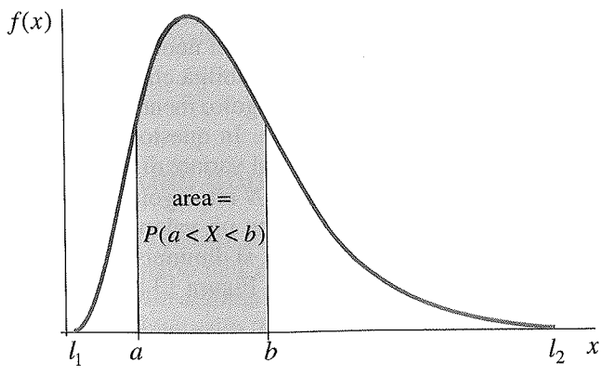
gura 13.8
nzione di densità di una
riabile casuale continua.



Una funzione $f(x)$ può essere considerata una funzione di densità se è non negativa e se l'area al di sotto di essa è uguale a 1.

Si esamini ora la Figura 13.9. In essa la probabilità che la variabile casuale X assuma un valore compreso tra a e b è raffigurata con l'area sotto la curva descritta dalla funzione $f(x)$ nell'intervallo (a, b).[5]

gura 13.9
robabilità dell'intervallo
, b) per una variabile
asuale continua.



[5] In termini matematici, la probabilità in questione è espressa da

$$P(a < X < b) = \int_a^b f(x)\,dx.$$

Questa probabilità può essere espressa come differenza tra i valori assunti dalla funzione di ripartizione in b e in a:

$$P(a < X < b) = P(X < b) - P(X < a) = P(X \le b) - P(X \le a) = F(b) - F(a),$$

dove vengono sfruttate le identità $P(X < a) = P(X \le a)$ e $P(X < b) = P(X \le b)$, in quanto $P(X = a) = P(X = b) = 0$ per la continuità di X.

ESEMPIO 13.10

Si consideri la funzione di densità

$$f(x) = \begin{cases} 2(x - 10)/50, & \text{se } 10 \le x \le 15, \\ 2(20 - x)/50, & \text{se } 15 \le x \le 20, \\ 0, & \text{altrove,} \end{cases}$$

detta densità triangolare, il cui grafico è presentato nella Figura 13.10. È facile verificare che $f(x)$ è una funzione di densità. Infatti, essa assume valori non negativi e l'area che sottende è uguale a 1, essendo 1 l'area del triangolo con i vertici nei punti (10, 0), (20, 0) e (15, 0.2).

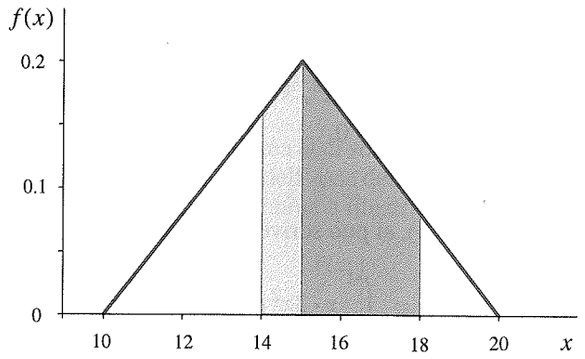


Figura 13.10
Funzione di densità triangolare.

Si voglia calcolare la probabilità $P(14 < X < 18)$, data dall'area sotto la curva nell'intervallo (14, 18), area che può essere calcolata come somma delle aree dei due trapezi indicati nella Figura 13.10: il primo (a sinistra) ha base minore 0.16 (valore assunto dalla funzione di densità per $x = 14$), base maggiore 0.20 e altezza 1; il secondo ha base minore 0.08 (valore assunto dalla funzione di densità per $x = 18$), base maggiore 0.20 e altezza 3. Si ha allora

$$P(14 < X < 18) = \frac{0.16 + 0.20}{2}(15 - 14) + \frac{0.08 + 0.20}{2}(18 - 15) = 0.6.$$

Naturalmente, questa probabilità può essere determinata mediante il calcolo integrale.[6]

■

[6] L'area sotto la curva di densità nell'intervallo (14, 18) è data da

$$P(14 < X < 18) = \int_{14}^{18} f(x)\,dx = \int_{14}^{15} [2(x - 10)/50]\,dx + \int_{15}^{18} [2(20 - x)/50]\,dx$$
$$= \left[(x - 10)^2/50\right]_{14}^{15} - \left[(20 - x)^2/50\right]_{15}^{18} = 0.6.$$

La funzione di densità e la funzione di ripartizione racchiudono, sostanzialmente, la stessa informazione sulla distribuzione di probabilità di X.[7]

13.3.1 Media e varianza

La media e la varianza di una variabile casuale continua sono espresse dagli integrali:

$$\mu = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x, \quad \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x. \tag{13.5}$$

Le (13.5) hanno la stessa struttura delle formule corrispondenti viste per le variabili casuali discrete (13.1 e 13.3). Del resto, l'integrale definito può essere visto come una "somma nel continuo". La radice quadrata della varianza dà la deviazione standard.

Come nel caso discreto, la media indica l'ordine di grandezza del fenomeno aleatorio rappresentato dalla variabile casuale, mentre la deviazione standard ne mette in luce il grado di dispersione attorno alla media.

ESEMPIO 13.11

Si calcolino la media e la varianza della variabile casuale continua avente funzione di densità

$$f(x) = \begin{cases} 3x^2, & 0 \le x \le 1, \\ 0, & \text{altrove.} \end{cases}$$

Soluzione

In base alla prima delle (13.5), si può scrivere

$$\mu = \int_0^1 x \cdot 3x^2\,\mathrm{d}x = 0.75.$$

Per la varianza, usando la seconda delle (13.5), si ha

$$\sigma^2 = \int_0^1 (x-0.75)^2 \cdot 3x^2\,\mathrm{d}x = 0.0375.$$

La verifica dei due risultati è lasciata per esercizio. ■

Come si è già visto per il caso discreto, se una variabile casuale continua X ha media μ_X e varianza σ_X^2, la variabile casuale

$$Y = a + bX,$$

dove a e b sono costanti, ha media

$$\mu_Y = a + b\mu_X,$$

e varianza

$$\sigma_Y^2 = b^2 \sigma_X^2.$$

[7] Si fa notare che a una data funzione di ripartizione corrispondono più funzioni di densità che differiscono, al più, in un'infinità numerabile di punti (si veda il Paragrafo B.4 dell'Appendice B).

La varianza può essere posta anche nella forma

$$\sigma^2 = \mathrm{E}(X^2) - \mu^2.$$

La dimostrazione di queste proprietà è contenuta nel Paragrafo B.5 dell'Appendice B.

13.3.2 Quantili

Nei capitoli relativi all'inferenza, si farà spesso riferimento ai quantili di distribuzioni di probabilità di variabili casuali continue. Se ne dà subito la definizione.

DEFINIZIONE 13.5

Considerata la distribuzione di probabilità di una variabile casuale continua, descritta dalla funzione di ripartizione $F(x)$, fissato un livello di probabilità p, si chiama quantile di livello p la quantità x_p in corrispondenza della quale la funzione di ripartizione assume il valore p:

$$F(x_p) = P(X \le x_p) = p.$$

Indicando con $F^{-1}(\cdot)$ l'inversa della funzione di ripartizione $F(x)$, nel caso in cui l'inversa esista, ossia nel caso in cui $F(x)$ non ha tratti costanti, si può scrivere

$$x_p = F^{-1}(p).$$

Poiché $F(x_p)$ è interpretabile come l'area al di sotto della curva di densità a sinistra di x_p, il quantile di livello p è il punto dell'asse delle ascisse alla cui sinistra l'area sotto la curva di densità è uguale a p, come è illustrato nella Figura 13.11.

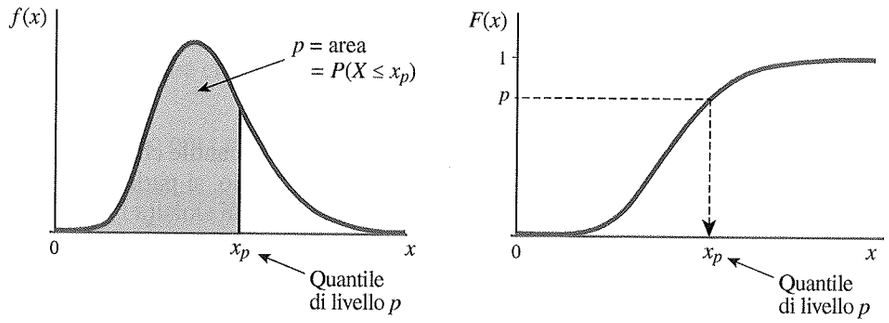


Figura 13.11
Quantile di livello p per una variabile casuale continua.

ESEMPIO 13.12

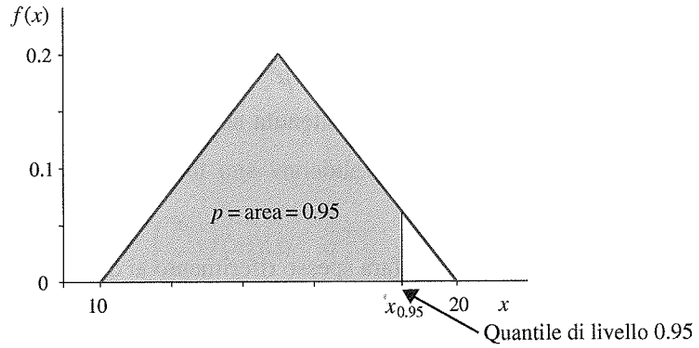
Si trovi il quantile $x_{0.95}$ della distribuzione triangolare dell'Esempio 13.10.

Soluzione

Bisogna individuare sull'asse delle ascisse il valore $x_{0.95}$ che lascia a sinistra un'area sotto la curva di densità uguale a 0.95 (si osservi la Figura 13.12). Poiché quest'area è uguale al complemento a 1 della superficie del triangolo (porzione bianca della Figura 13.12) avente come base $(20 - x_{0.95})$ e come altezza l'ordinata della funzione di densità nel punto $x_{0.95}$, basta risolvere l'equazione

$$1 - (20 - x_{0.95}) \cdot \frac{2(20 - x_{0.95})}{50} \cdot \frac{1}{2} = 0.95,$$

ura 13.12
antile $x_{0.95}$ della
ribuzione di cui
Esempio 13.10.



la cui soluzione è $x_{0.95} = 18.42$. ■

ESEMPIO 13.13

Si trovi il quantile di livello 0.10 della distribuzione di probabilità descritta da

$$f(x) = \begin{cases} 3x^2, & 0 \le x \le 1, \\ 0, & \text{altrove.} \end{cases}$$

Soluzione

La funzione di ripartizione di questa variabile casuale è data da

$$F(x) = \int_0^x 3t^2 \, dt = x^3.$$

Il quantile di livello 0.10 si ricava dalla soluzione dell'equazione

$$x^3 = 0.10,$$

da cui

$$x_{0.10} = \sqrt[3]{0.10} = 0.4642.$$

■

Spesso si usa il termine centile o percentile per indicare il quantile che corrisponde a una probabilità p espressa in centesimi. In questo modo, si parla di: primo, secondo, ..., 99-esimo centile, per intendere l'ascissa alla cui sinistra l'area sotto la curva di densità è, rispettivamente, 0.01, 0.02, ..., 0.99.

13.4 Ulteriori proprietà delle variabili casuali

Questo paragrafo contiene argomenti un po' eterogenei: da una parte la standardizzazione e i momenti di variabili casuali, dall'altra le disuguaglianze di Chebyshev e di Markov. C'è però un elemento che accomuna i quattro temi: essi hanno a che fare con le proprietà di trasformazioni di variabili casuali.

DEFINIZIONE 13.6

Data la variabile casuale X, discreta o continua, avente media μ e varianza σ^2, si chiama variabile casuale standardizzata la quantità

$$Y = \frac{X - \mu}{\sigma}.$$

Ogni variabile casuale così definita gode delle proprietà indicate nella Proposizione 13.1.

PROPOSIZIONE 13.1

Ogni variabile casuale standardizzata ha media 0 e varianza 1.

La dimostrazione è immediata. Basta considerare che $Y = (X - \mu)/\sigma$ è una trasformazione lineare di X del tipo $Y = a + bX$, con $a = -\mu/\sigma$ e $b = 1/\sigma$. Per le proprietà della media e della varianza, si ha:

$$\mu_Y = a + b\,\mathrm{E}(X) = -\frac{\mu}{\sigma} + \frac{1}{\sigma}\mu = 0,$$

$$\sigma_Y^2 = b^2 \mathrm{Var}(X) = \frac{1}{\sigma^2}\sigma^2 = 1.$$

■

ESEMPIO 13.14

Si standardizzi la variabile casuale considerata nell'Esempio 13.1.

Soluzione

Bisogna sottrarre a ogni valore di X la media 1.5 e dividere il risultato per la deviazione standard 0.87. La distribuzione di probabilità che ne risulta è esposta nella tabella che segue:

$(x - \mu)/\sigma$	−1.72	−0.57	0.57	1.72
Probabilità	0.125	0.375	0.375	0.125

È facile accertare che la nuova distribuzione ha media 0 e varianza 1. ■

Si definiscono ora i momenti di ordine r di una distribuzione di probabilità.

DEFINIZIONE 13.7

Sia X una variabile casuale e r un numero intero positivo. Si chiama momento non centrato di ordine r la quantità

$$\mu_r = \mathrm{E}(X^r).$$

Si chiama momento centrato di ordine r la quantità

$$\bar{\mu}_r = \mathrm{E}[(X - \mu)^r],$$

dove $\mu = \mathrm{E}(X)$.

È immediato verificare che il momento non centrato di ordine uno è la media, $\mu_1 = \mathrm{E}(X^1) = \mathrm{E}(X)$, e che il momento centrato di ordine due è la varianza, $\bar{\mu}_2 = \mathrm{E}[(X - \mu)^2] = \mathrm{Var}(X)$.

La disuguaglianza di Chebyshev è una proprietà generale delle distribuzioni di probabilità di variabili casuali sia discrete sia continue, molto utile nella teoria della probabilità.

PROPOSIZIONE 13.2

Sia X una variabile casuale qualsiasi con media μ e varianza σ^2. Sia $\delta > 0$ una quantità prefissata. Vale allora la disuguaglianza

$$P(|X - \mu| \geq \delta) \leq \frac{\sigma^2}{\delta^2}. \tag{13.6}$$

La dimostrazione è data nel Paragrafo B.6 dell'Appendice B. ■

La disuguaglianza (13.6) può essere scritta nella forma equivalente

$$P(|X - \mu| < \delta) = P(\mu - \delta < X < \mu + \delta) \geq 1 - \frac{\sigma^2}{\delta^2}. \tag{13.7}$$

Dunque, in base alla (13.7), per un valore assegnato di $\delta > 0$, la probabilità che la variabile casuale X assuma valori nell'intervallo $(\mu - \delta, \mu + \delta)$ è non inferiore a $1 - \sigma^2/\delta^2$, quantità che cresce al diminuire della varianza σ^2.

ESEMPIO 13.15

Data una variabile casuale con media $\mu = 100$ e varianza $\sigma^2 = 4$, si determini il limite inferiore della probabilità che la variabile casuale assuma un valore nell'intervallo (97, 103).

Soluzione

Il valore di δ è dato dalla differenza $103 - 100 = 3$. In base alla (13.7), si ha

$$P(97 < X < 103) \geq 1 - \frac{4}{9} = 0.56.$$

Si può dunque affermare che la probabilità che la variabile casuale X assuma valori all'interno dell'intervallo (97, 103) è almeno uguale a 0.56. Se la varianza fosse 1.5, tale probabilità sarebbe 0.83. ■

La disuguaglianza di Chebyshev dà un'informazione che dipende non già dalla forma della distribuzione di probabilità di X, ma solo da μ e da σ^2. Il risultato è riferibile a tutte le variabili casuali le cui distribuzioni di probabilità hanno la stessa media e la stessa varianza di quella in esame.

Nella Proposizione 13.3, è enunciata la disuguaglianza di Markov, che limita superiormente la probabilità che una trasformazione positiva di una variabile casuale superi un valore assegnato.

PROPOSIZIONE 13.3

Sia X una variabile casuale qualsiasi e $g(\cdot)$ una funzione non negativa. Vale allora la disuguaglianza

$$P[g(X) \geq \delta] \leq \frac{\mathrm{E}[g(X)]}{\delta}, \quad \text{per ogni } \delta > 0.$$

La dimostrazione è data nel Paragrafo B.7 dell'Appendice B. ■

Riassunto

Per variabile casuale si intende una funzione che associa un numero reale a ogni punto dello spazio campionario. Se lo spazio campionario è costituito da un numero finito o da un'infinità numerabile di eventi elementari, la variabile casuale è discreta. Al contrario, se lo spazio campionario è continuo, la variabile casuale può essere sia discreta che continua; in quest'ultimo caso, potrà assumere tutti i valori possibili all'interno di un dato intervallo o di un insieme di intervalli. Per ogni variabile casuale, la distribuzione di probabilità può essere descritta per mezzo della funzione di ripartizione.

Per una variabile casuale discreta è possibile elencare i valori che questa può assumere e associare a ciascuno di essi un livello di probabilità: in questo modo, si viene a definire una funzione di probabilità.

Quando la variabile casuale è continua, a ogni suo possibile valore si associa non un livello di probabilità, ma una densità di probabilità, in base a una funzione non negativa il cui integrale è pari a uno.

La distribuzione di probabilità di una variabile casuale può essere descritta sinteticamente mediante le costanti caratteristiche. L'attenzione viene posta sulle due principali, la media e la varianza. Per le variabili casuali continue, vengono anche definiti i quantili, che sono dei particolari valori corrispondenti a specifici livelli della funzione di ripartizione.

Nell'ultima parte del capitolo, con riferimento a una variabile casuale discreta o continua, vengono illustrati: la standardizzazione, i momenti, le disuguaglianze di Chebyshev e di Markov.

Parole chiave

centile, 286
deviazione standard, 280
distribuzione di probabilità, 275
disuguaglianza di Chebyshev, 287
disuguaglianza di Markov, 288
expectation, expected value, 278
funzione di densità, 281
funzione di densità triangolare, 283
funzione di probabilità, 276
funzione di ripartizione, 276
linearità della media, 279
media, 278
momento centrato, 287
momento non centrato, 287
percentile, 286
quantile, 285
valore atteso, 278
variabile aleatoria, 273
variabile casuale, 273
variabile casuale continua, 281
variabile casuale discreta, 274
variabile casuale standardizzata, 286
variabile stocastica, 273
varianza, 279

Commenti e curiosità

1. Funzione di densità come approssimazione di un istogramma

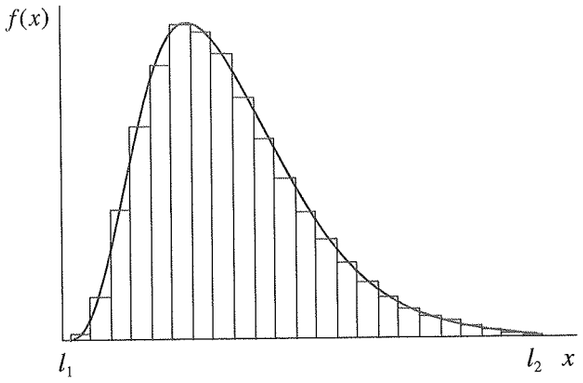
Si può pensare alla funzione di densità come alla curva che approssima un istogramma (si veda la Figura 13.13), anche per "aggirare" un po' di matematica.

Con questa rappresentazione, si comprende facilmente perché le probabilità sono date da aree sottese alla curva e si può fare a meno del calcolo integrale per la soluzione approssimata dei problemi di seguito indicati:

(a) calcolo dell'area sottesa alla curva in un dato intervallo (a, b);
(b) calcolo del valore atteso;
(c) calcolo della varianza.

Per quanto riguarda il primo punto, basta suddividere l'intervallo (a, b) in un numero sufficientemente grande di intervalli più piccoli della stessa ampiezza δ, moltiplicare l'ordinata della funzione di densità nel punto centrale di ciascuno di questi intervalli per δ e sommare i risultati.

Per il calcolo approssimato del valore atteso, occorre suddividere l'intervallo in cui la variabile è definita in un numero sufficientemente grande di intervalli più piccoli di ampiezza δ, dopodiché, per ognuno di questi, si effettua il prodotto di tre fattori: il valore centrale dell'intervallo, l'ordinata della funzione di



ura 13.13
nzione di densità come
va che approssima un
ogramma.

nsità nel punto centrale dell'intervallo e la costante δ; infine, si sommano i risultati delle operazioni ecedenti.

In modo analogo, si calcola il valore approssimato della varianza: il procedimento segue gli stessi ssi visti per la media, con la sola variante che nei prodotti sopra indicati ai valori centrali vanno sostituiti quadrati degli scarti tra i valori centrali stessi e la media.

Valore atteso, giochi equi e paradosso di San Pietroburgo

molte situazioni, il valore atteso può essere interpretato come la vincita media in un gran numero di ommesse. Per esempio, nel caso del lancio di un dado regolare in cui si vincono un euro se esce la ccia uno, due euro se esce la faccia due, e così via, indicando con X l'ammontare della vincita (X è una ariabile casuale discreta che assume i valori $1, 2, \ldots, 6$ con probabilità $1/6$), il valore atteso di X, ossia il alore atteso della vincita, è pari a 3.5 euro. Pur essendo impossibile vincere 3.5 euro nel singolo lancio, uesto valore rappresenta la vincita media su un gran numero di lanci. Se si dovesse stabilire un prezzo da agare per partecipare a questo gioco di sorte, 3.5 euro rappresenterebbe il prezzo equo: un prezzo inferiore arantirebbe al giocatore un vantaggio nel lungo andare, mentre un prezzo superiore lo svantaggia.

Non sempre però tutto funziona in modo così lineare. Si consideri il seguente gioco di sorte. Una oneta bilanciata viene lanciata fino a che non esce testa per la prima volta. Se esce testa al primo lancio vincono due euro, se esce testa al secondo lancio si vincono quattro euro, e così via. Quanto saremmo isposti a pagare per partecipare a questo gioco? Praticamente nessuno è disposto a pagare più di qualche uro (l'esperienza insegna che molti studenti a cui è stato proposto questo gioco si rifiutano di pagare più i due o tre euro). In realtà, il prezzo equo del gioco tende a infinito. Si consideri l'albero degli eventi appresentato nella Figura 13.14 e sia X la variabile casuale che indica il valore della vincita. È facile edere che il valore atteso di X è pari a

$$\mathrm{E}(X) = \sum_x x f(x) = 2 \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} + 8 \cdot \frac{1}{8} + 16 \cdot \frac{1}{16} + \cdots = 1 + 1 + 1 + 1 + \cdots = \infty.$$

igura 13.14
aradosso di San
ietroburgo: spazio
ampionario e relative
ncite.

ω	$P(\omega)$	$X(\omega)$
T	1/2	2
CT	1/4	4
CCT	1/8	8
$CCCT$	1/16	16
$\vdots$	$\vdots$	$\vdots$

Questa divaricazione tra valore atteso e prezzo che si è disposti a pagare per partecipare a questo gioco prende il nome di paradosso di San Pietroburgo. Il paradosso si risolve prendendo in considerazione il concetto di utilità marginale e il fatto che il tempo a disposizione del giocatore è limitato (e non infinito).

Riguardo alla teoria dell'utilità, sono stati condotti numerosi studi per capire i meccanismi in base ai quali la gente effettua le scelte. In uno di questi, a un campione di persone fu sottoposta l'alternativa tra le due opzioni: A) un regalo certo di 240 euro, B) 1000 euro con probabilità di 0.25 e zero euro con probabilità 0.75. La maggior parte delle persone scelse l'opzione A), anche se il valore atteso connesso alla scelta B) è superiore a 240 euro. La decisione riflette probabilmente l'adagio per cui "è meglio un uovo oggi che una gallina domani". E ancora, tra le alternative A) una perdita certa di 740 euro e B) una perdita di 1000 euro con probabilità del 75% e una perdita zero con probabilità del 25%, la maggior parte delle persone scelse B), nonostante la perdita attesa derivante dall'opzione B) fosse superiore alla perdita sicura derivante dalla prima. Il secondo comportamento è dovuto, forse, allo stesso meccanismo psicologico per cui la gente gioca o si assicura contro gli incendi, i furti ecc. Sembra che gli individui annettano alle basse probabilità (di vincere o di perdere) un peso più elevato di quanto non esprimano i loro valori numerici.

Mentre il valore atteso guida, tra le altre, le compagnie di assicurazione nella determinazione dei premi, così che per una polizza vita si chiederanno premi crescenti all'aumentare dell'età dell'assicurato, in quanto con l'età cresce la probabilità di morte, al contrario, nei giochi di sorte (lotto, lotterie ecc.), il giocatore non tiene conto del valore atteso della vincita. Diversamente, non parteciperebbe al gioco, considerato che esso non è mai equo: il valore atteso della vincita è enormemente inferiore all'importo che paga per partecipare al gioco. Si deve allora concludere che il comportamento del giocatore è irrazionale? Non necessariamente, considerando la teoria dell'utilità, ma anche il fatto che la partecipazione al gioco è fonte di altre forme di utilità (il piacere del rischio, l'illusione ecc.).

3. Valore atteso e valutazione di derivati finanziari

Il concetto di valore atteso trova applicazione in molti campi dell'economia, in particolare in finanza. Un importante problema è quello della valutazione di un derivato finanziario. Si consideri un'azione del prezzo di un euro che al termine di un periodo di riferimento può raddoppiare di valore (con probabilità 1/2) o dimezzarsi (con probabilità 1/2). Si esamini quindi il seguente contratto (derivato): al termine del periodo di riferimento, se l'azione aumenta la banca d'affari pagherà un euro all'investitore, se l'azione diminuisce, la banca d'affari non pagherà nulla. Quanto dovrebbe pagare l'investitore per acquistare questo contratto dalla banca d'affari? Considerando il contratto come un gioco di sorte, senza nessun dubbio, il prezzo equo sarebbe pari a 0.5: nel lungo andare, cioè immaginando di ripetere lo stesso investimento (ovvero di comperare lo stesso contratto) molte e molte volte, la "vincita" media sarebbe esattamente pari a 0.5. Tuttavia, si può dimostrare che, se l'investitore pagasse il contratto mezzo euro, la banca d'affari avrebbe un guadagno garantito di 1/6. Per verificare ciò, si consideri lo schema nella Figura 13.15.

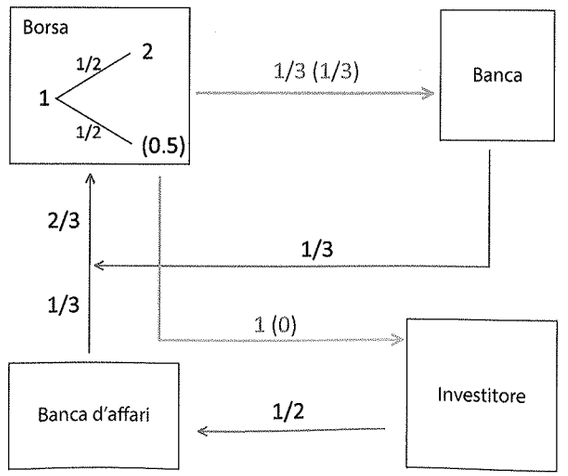


Figure 13.15
Schema per la valutazione di un derivato finanziario.

Si assuma che l'investitore paghi 1/2, ovvero 3/6, per il contratto. La banca d'affari terrà per se 1/6 impiegherà i rimanenti 2/6 = 1/3 per acquistare 2/3 di azione (chiedendo in prestito 1/3 a una banca ommerciale). Al termine del periodo, se l'azione raddoppia, la banca d'affari avrà un ritorno di 4/3 e otrà sia ripagare il prestito (di 1/3) che liquidare l'investitore (con 3/3 = 1 euro). Se invece l'azione si mezza, la banca d'affari avrà un ritorno di 1/3 e ancora potrà sia ripagare il prestito (di 1/3) che liquidare nvestitore, che, in questo caso, si aspetta zero euro. In ogni caso, la banca d'affari avrà tenuto per sé 1/6.

L'idea racchiusa in questo semplice esempio è alla base di tutta la teoria sulla valutazione dei derivati nanziari. L'aspetto più interessante è che il "vero prezzo equo" del contratto si può ancora ricavare come n valore atteso, ma non rispetto alle probabilità "vere" dell'azione (1/2 e 1/2), ma rispetto a delle proba- lità fittizie che fanno sì che l'andamento di borsa sia un gioco equo. Per ricavare queste probabilità basta onsiderare l'equazione

$$1 = 2 \cdot \pi + 0.5 \cdot (1 - \pi),$$

ne risolta in π fornisce $\pi = 1/3$. Quindi, il valore atteso del contratto sarà pari a $1 \cdot (1/3) + 0 \cdot (2/3) = 1/3$, questo, come già visto, rappresenta il prezzo che non garantisce un guadagno certo alla banca d'affari.

Esercizi

13.1 Sei lotti di semilavorati sono pronti per essere consegnati da parte di un certo fornitore. Il numero di pezzi difettosi in ciascun lotto è il seguente:

Lotto	A	B	C	D	E	F
N. di pezzi difettosi	0	2	0	1	2	0

Si supponga che uno di questi lotti venga estratto a sorte e consegnato a una data azienda manifatturiera. Sia X la variabile casuale che esprime il numero di pezzi difettosi presenti nel lotto ricevuto dall'azienda.

a) Si determini la funzione di probabilità.
b) Si determini la funzione di ripartizione.
c) Si calcolino il valore atteso e la varianza.

13.2 Il numero X di contratti di locazione conclusi in una settimana da un'agenzia immobiliare di una grande città è definito dalla seguente distribuzione di probabilità:

N. di contratti x	0	1	2	3	4	5	6
Probabilità $f(x)$	0.16	0.29	0.27	0.15	0.09	0.03	0.01

a) Si rappresenti graficamente la funzione di probabilità.
b) Si determinino le probabilità: $P(X > 4)$, $P(2 \le X < 6)$ e $P(3 < X \le 6)$.
c) Si calcolino il valore atteso e la deviazione standard.
d) Supponendo che l'agenzia guadagni 500 euro per ogni contratto concluso, si determini il guadagno medio per settimana.

13.3 Una scatola contiene sei biglietti. Due di essi valgono 300 000 euro ciascuno, gli altri 200 000 euro ciascuno.

a) Si determini il valore atteso della variabile casuale che registra il valore di un biglietto estratto a sorte dalla scatola.
b) Si determini il valore atteso della variabile casuale che registra la somma del valore di due biglietti estratti a sorte dalla scatola senza reinserimento.

13.4 In una fase del processo di lavorazione di bicchieri viene usato un getto di fiamma la cui temperatura è descritta da una variabile casuale con la seguente distribuzione:

Temperatura	543	550	554	560
Probabilità	0.11	0.35	0.41	0.13

a) Si calcolino la media e la deviazione standard della distribuzione.
b) Si standardizzi la distribuzione.

13.5 Si consideri la funzione $f(x)$ definita da

$$f(x) = \frac{2(b - x)}{b^2}, \quad 0 \le x \le b,$$

e $f(x) = 0$ altrove. Si dimostri che $f(x)$ è una funzione di densità di probabilità.

13.6 Si consideri una variabile casuale continua con funzione di densità $f(x)$ definita da

$$f(x) = \frac{3}{8}x^2, \quad 0 \le x \le 2,$$

e $f(x) = 0$, altrove. Si calcolino il valore atteso e la varianza.

13.7 Si consideri la funzione $f(x)$ definita da

$$f(x) = kx^3, \quad 0 \le x \le 10,$$

e $f(x) = 0$, altrove.

a) Si determini il numero k per cui $f(x)$ è una funzione di densità di probabilità.
b) Utilizzando il valore di k ottenuto nel punto precedente e assumendo che $f(x)$ sia la densità della variabile casuale continua X, si calcoli la probabilità che $5 < X < 8$.

13.8 Il numero di chiamate telefoniche che arrivano a un centralino dalle 9 alle 13 di un giorno feriale è una variabile casuale con media 63 e deviazione standard 5.5. Si applichi la disuguaglianza di Chebyshev con $\delta = 11$ per determinare il limite inferiore della probabilità che la variabile casuale assuma un valore nell'intervallo (52, 74).

13.9 L'incasso nelle giornate festive di un bar del centro di una grande città è assimilabile a una variabile casuale con media 6200 e deviazione standard 570. Quale informazione può dare la disuguaglianza di Chebyshev per $\delta = 1000$?

13.10 Una variabile casuale ha media 175 e varianza 25. Tramite la disuguaglianza di Chebyshev, si determinino i valori a e b per i quali

$$P(a < X < b) \ge 0.90.$$

13.11 Data una variabile casuale con media μ non nota e varianza 20.5, si determini il limite inferiore della probabilità

$$P(\mu - 10 < X < \mu + 10).$$

13.12 Sia X una variabile casuale con $E(X) = 3$ e $Var(X) = 4$. Si determini $E(2 + 3X - 5X^2)$.

CAPITOLO

14

Alcune distribuzioni di probabilità parametriche

Sommario del capitolo

14.1 Introduzione
14.2 Distribuzione uniforme discreta
14.3 Distribuzione di Bernoulli
14.4 Distribuzione binomiale
14.5 Distribuzione di Poisson
14.6 Distribuzione uniforme continua
14.7 Distribuzione esponenziale
14.8 Distribuzione normale
14.9 Distribuzione chi-quadrato
Riassunto
Commenti e curiosità
Esercizi

14.1 Introduzione

Nel capitolo precedente, si è visto come una variabile casuale sia il risultato di un processo che associa numeri reali agli eventi elementari dello spazio campionario di un esperimento casuale, e come la sua distribuzione di probabilità possa essere descritta da una funzione di probabilità (densità nel caso continuo) o da una funzione di ripartizione.

In questo capitolo, verranno illustrate alcune distribuzioni di probabilità parametriche, nel senso che sono definite da funzioni di probabilità (densità) caratterizzate da una o più costanti numeriche dette parametri.

ESEMPIO 14.1

Si consideri la funzione

$$f(x) = \frac{2}{n(n+1)}x, \quad x = 1, 2, \ldots, n,$$

che assegna probabilità linearmente crescenti ai numeri interi consecutivi $1, 2, \ldots, n$. Chiaramente, la distribuzione dipende da n. Ad esempio, se $n = 2$, X assume i valori 1 e 2 con probabilità $1/3$ e $2/3$, se $n = 3$, X assume i valori 1, 2 e 3 con probabilità $1/6$, $1/3$ e $1/2$ e così via. Dunque, questa formula definisce non una, ma una famiglia di distribuzioni. ■

Quando una distribuzione di probabilità è descritta da una funzione matematica che dipende da uno o più parametri, si ha a che fare con una famiglia di distribuzioni di probabilità che differiscono tra loro solo per i valori di questi parametri.

Nelle pagine che seguono verranno esaminate alcune famiglie di distribuzioni di probabilità di uso comune. Alcune di esse, come le distribuzioni binomiale e normale, verranno ripetutamente applicate nei capitoli successivi. Altre, come la distribuzione di Poisson, verranno presentate per la loro importanza nell'applicazione della teoria della probabilità ai problemi reali.

14.2 Distribuzione uniforme discreta

In generale, una variabile casuale ha *distribuzione uniforme discreta* se assume un certo insieme (finito) di valori con la stessa probabilità. La definizione che segue si riferisce al caso particolare in cui tale insieme è costituito dai primi n numeri naturali.

DEFINIZIONE 14.1

Una variabile casuale, X, ha distribuzione uniforme sui primi n numeri naturali se la sua funzione di probabilità è espressa da

$$f(x) = \frac{1}{n}, \quad x = 1, 2, \ldots, n.$$

PROPOSIZIONE 14.1

La media e la varianza della distribuzione uniforme sui primi n numeri naturali sono date da:

$$\mathrm{E}(X) = \frac{n+1}{2}, \quad \mathrm{Var}(X) = \frac{n^2-1}{12}.$$

La dimostrazione è lasciata per esercizio (si veda l'Esercizio 14.1). ■

ESEMPIO 14.2

Si consideri l'esperimento del lancio di un dado bilanciato e sia X la variabile casuale che registra il numero che appare sulla faccia superiore. Si rappresentino graficamente le funzioni di probabilità e di ripartizione. Si calcolino poi media e varianza.

Soluzione

La funzione di probabilità e quella di ripartizione sono rappresentate in Figura 14.1.

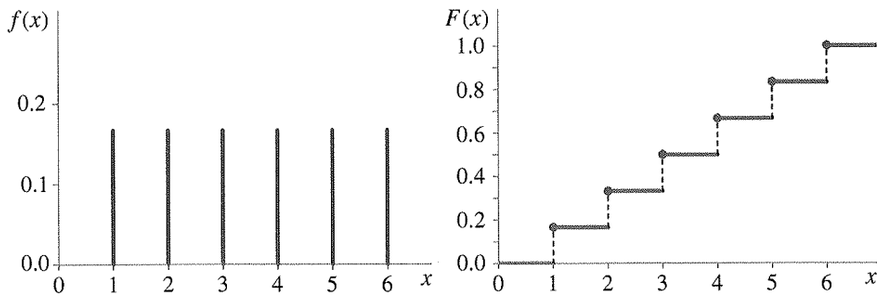


Figura 14.1
Funzioni di probabilità e di ripartizione della distribuzione uniforme discreta sui primi 6 numeri naturali.

La media e la varianza della distribuzione sono:

$$\mathrm{E}(X) = \frac{6+1}{2} = 3.5, \quad \mathrm{Var}(X) = \frac{6^2-1}{12} = 2.917.$$

■

14.3 Distribuzione di Bernoulli

Si consideri un esperimento casuale i cui esiti sono due eventi incompatibili A e $\bar{A}$; si pensi ad esempio al lancio di una moneta, dove A e $\bar{A}$ sono gli eventi testa e croce, oppure all'estrazione casuale con ripetizione di una unità da una popolazione finita, dove A e $\bar{A}$ possono essere maschio e femmina, malato e sano ecc. La singola esecuzione di un tale esperimento è chiamata prova Bernoulliana (dal nome di Jacob Bernoulli, matematico svizzero del XVII secolo). Gli eventi A e $\bar{A}$ sono denominati convenzionalmente successo e insuccesso e codificati come 1 e 0. Indicando con p la probabilità di A, la probabilità di $\bar{A}$ è $1 - p$.

DEFINIZIONE 14.2

Una variabile casuale discreta, X, ha distribuzione di Bernoulli se assume valore 0 e 1 con probabilità $1 - p$ e p, dove $0 < p < 1$. La sua funzione di probabilità è espressa da

$$f(x) = p^x(1-p)^{1-x}, \quad x = 0, 1.$$

PROPOSIZIONE 14.2

La media e la varianza della distribuzione di Bernoulli di parametro p sono date da:

$$\mathrm{E}(X) = p, \quad \mathrm{Var}(X) = p(1-p).$$

La dimostrazione è lasciata per esercizio (si veda l'Esercizio 14.2). ■

ESEMPIO 14.3

Si consideri l'esperimento del lancio di un dado bilanciato e si definisca la variabile casuale X che assume il valore 1 se esce 1 o 2 e il valore 0 se esce uno degli altri quattro numeri. Si trovi la funzione di probabilità della variabile casuale e se ne calcolino media e varianza.

Soluzione

In questo caso, l'evento successo è il numero 1 o 2. La funzione di probabilità della variabile casuale è riportata nella tabella che segue:

x	0	1
$f(x)$	2/3	1/3

La media e la varianza della distribuzione sono:

$$\mathrm{E}(X) = 1/3, \quad \mathrm{Var}(X) = (1/3) \cdot (2/3) = 2/9.$$

■

14.4 Distribuzione binomiale

La distribuzione binomiale è un modello probabilistico utile in molte applicazioni e pertanto merita un'attenzione speciale. Si consideri un processo industriale che produce semilavorati nel settore dell'elettronica. L'esperienza suggerisce che l'1% dei pezzi prodotti è difettoso. Da una partita di pezzi prodotti nell'ultima settimana lavorativa, vengono estratti a sorte con ripetizione 20 pezzi. La selezione casuale

del singolo pezzo può essere pensata come una prova Bernoulliana, dove l'evento successo è definito come "osservazione di un pezzo difettoso". Chiaramente, la probabilità di successo è 0.01 e rimane invariata da un'estrazione all'altra. Il numero di pezzi difettosi nelle 20 estrazioni è una variabile casuale i cui possibili valori sono i numeri interi da 0 a 20.

Come altro esempio, si consideri un'indagine sociale su occupazione e disoccupazione basata su 500 interviste a persone (appartenenti alle forze di lavoro) estratte a sorte (con ripetizione). In ogni intervista, si può osservare uno dei due possibili esiti, occupato o disoccupato. Definendo il primo risultato come successo e assumendo che il 91% della popolazione sia occupato, l'evento successo ha probabilità di 0.91, invariata da una estrazione e all'altra. Il numero di occupati osservabili nelle 500 interviste è una variabile casuale i cui possibili valori sono i numeri interi da 0 a 500.

Le situazioni sopra descritte hanno in comune i seguenti elementi:

- l'esperimento Bernoulliano viene ripetuto un certo numero di volte;
- la probabilità di successo non varia da una prova all'altra;
- le prove sono tra loro indipendenti.

Si fa notare che la probabilità di successo rimane invariata e le prove sono indipendenti per il fatto che le estrazioni vengono eseguite con ripetizione.

Generalizzando questa idea, si considerino n prove Bernoulliane indipendenti, ciascuna con la stessa probabilità p di successo, e sia X la variabile casuale che conta quante volte si verifica l'evento successo. L'obiettivo è quello di determinare la distribuzione di probabilità di questa variabile casuale. L'esempio che segue si riferisce al caso $n = 3$.

ESEMPIO 14.4

Si considerino tre lanci di una moneta non bilanciata con probabilità di testa (successo) pari a p, e sia X la variabile casuale che conta il numero di teste. La tabella che segue riporta: il risultato dei lanci nella prima colonna; la loro probabilità nella seconda; il valore che assume la variabile casuale nella terza.

Si noti che la probabilità del singolo evento elementare ω si ottiene moltiplicando le probabilità degli eventi T o C occorrenti nelle singole estrazioni. Per esempio, la probabilità di TTT è $p \cdot p \cdot p = p^3$, la probabilità di TTC è $p \cdot p \cdot (1-p) = p^2(1-p)$ e così via.

ω	$P(\omega)$	$X(\omega)$
TTT	p^3	3
TTC	$p^2(1-p)$	2
TCT	$p^2(1-p)$	2
TCC	$p(1-p)^2$	1
CTT	$p^2(1-p)$	2
CTC	$p(1-p)^2$	1
CCT	$p(1-p)^2$	1
CCC	$(1-p)^3$	0

Ne segue che la funzione di probabilità di X è:

x	0	1	2	3
$f(x)$	$(1-p)^3$	$3p(1-p)^2$	$3p^2(1-p)$	p^3

■

DEFINIZIONE 14.3

Una variabile casuale discreta, X, ha distribuzione binomiale se la sua funzione di probabilità è espressa da

$$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n, \tag{14.1}$$

dove $0 < p < 1$, n è un numero intero positivo e il simbolo $n!$ è il fattoriale di n.[1]

■

Il procedimento che porta alla funzione di probabilità (14.1) è illustrato nel Paragrafo B.8 dell'Appendice B.

PROPOSIZIONE 14.3

La media e la varianza della distribuzione binomiale di parametri n e p sono date da:

$$\mathrm{E}(X) = np, \ \mathrm{Var}(X) = np(1-p). \tag{14.2}$$

La dimostrazione è data nel Paragrafo B.8 dell'Appendice B. ■

ESEMPIO 14.5

Un processo industriale dà luogo a pezzi difettosi con una probabilità dell'1%. Cinque pezzi sono estratti casualmente e indipendentemente dalla linea di produzione. Si determini la distribuzione di probabilità del numero di pezzi difettosi e se ne calcolino media e varianza.

Soluzione

La variabile casuale può assumere i valori $0, 1, \ldots, 5$. La sua funzione di probabilità, ottenuta applicando la (14.1) con $p = 0.01$ e $n = 5$, è esposta nella tabella che segue:

[1] Le quantità

$$\frac{n!}{0! \cdot n!}, \ \frac{n!}{1! \cdot (n-1)!}, \cdots, \ \frac{n!}{n! \cdot 0!},$$

che appaiono nella (14.1), per $x = 0, 1, \ldots, n$, si chiamano coefficienti binomiali, perché sono i coefficienti moltiplicativi dei termini dello sviluppo del binomio di Newton:

$$[p + (1-p)]^n = \frac{n!}{0! \cdot n!} p^0 (1-p)^n + \frac{n!}{1! \cdot (n-1)!} p^1 (1-p)^{n-1} + \cdots + \frac{n!}{n! \cdot 0!} p^n (1-p)^0.$$

x	0	1	2	3	4	5
$f(x)$	0.951	0.048	0.001	0.000	0.000	0.000

I coefficienti binomiali utilizzati per il calcolo di queste probabilità sono:

$$\frac{5!}{0!\cdot 5!}=1,\quad \frac{5!}{1!\cdot 4!}=5,\quad \frac{5!}{2!\cdot 3!}=10,\quad \frac{5!}{3!\cdot 2!}=10,\quad \frac{5!}{4!\cdot 1!}=5,\quad \frac{5!}{5!\cdot 0!}=1.$$

In base alla (14.2), la media e la varianza della distribuzione sono così determinati:

$$\mathrm{E}(X)=5\cdot 0.01=0.05,\quad \mathrm{Var}(X)=5\cdot 0.01\cdot 0.99=0.0495.$$

■

La distribuzione binomiale dipende dai parametri n e p. La Figura 14.2 mostra come questi incidano sulla forma della distribuzione di probabilità della variabile casuale.

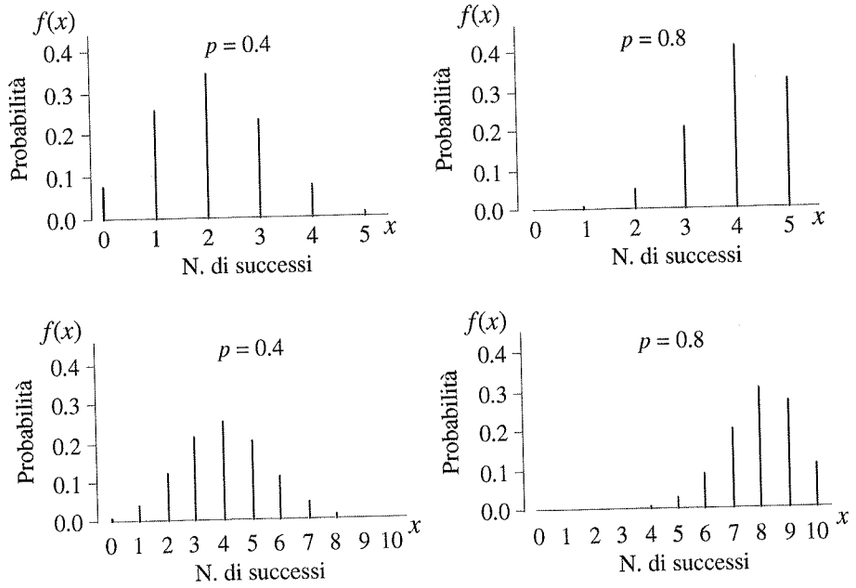


Figura 14.2
Distribuzione binomiale per $p = 0.4, 0.8$ e $n = 5, 10$.

Quando $p = 0.5$ la distribuzione binomiale è simmetrica. Sussistono infatti le identità:

$$f(x)=f(n-x),\quad x=0,1,\ldots,(n-\delta)/2,$$

dove δ assume il valore 0 o 1 a seconda che n sia pari o dispari.

ESEMPIO 14.6

Nella Figura 14.3 viene mostrato il grafico della distribuzione binomiale con parametri $n = 10$ e $p = 0.5$. Come atteso, la distribuzione è simmetrica in quanto: $f(0) = f(10) = 0.0010$, $f(1) = f(9) = 0.0098$, ..., $f(4) = f(6) = 0.2051$, $f(5) = 0.2461$.

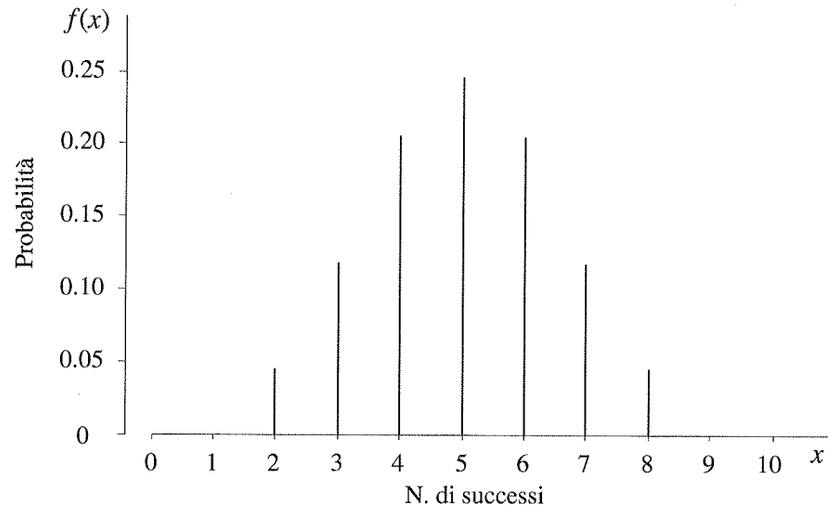


Figura 14.3
Distribuzione binomiale con $p = 0.5$ e $n = 10$.

Si può dimostrare che la distribuzione binomiale è asimmetrica positivamente se $p < 0.5$ e negativamente se $p > 0.5$.

14.5 Distribuzione di Poisson

La distribuzione di Poisson prende il nome da Siméon-Denis Poisson (matematico francese, 1781-1840) ed è così definita.

DEFINIZIONE 14.4

Una variabile casuale discreta, X, ha distribuzione di Poisson se la sua funzione di probabilità è espressa da

$$f(x)=\frac{\lambda^x}{x!}e^{-\lambda},\quad x=0,1,2,\ldots,$$

dove il parametro λ è una costante positiva.

PROPOSIZIONE 14.4

La media e la varianza della distribuzione di Poisson di parametro λ sono date da:

$$\mathrm{E}(X)=\lambda,\quad \mathrm{Var}(X)=\lambda. \tag{14.3}$$

La dimostrazione è contenuta nel Paragrafo B.8 dell'Appendice B.

■

La distribuzione di Poisson trova applicazione quando il fenomeno aleatorio consiste nel contare le occorrenze di un evento casuale in un dato intervallo di tempo o di spazio. Il modello è appropriato nei casi in cui la probabilità dell'evento aleatorio è molto piccola. Per questo motivo è nota anche come legge degli eventi rari.

ESEMPIO 14.7

Per studiare il processo con cui la densità di una sostanza radioattiva si riduce nel tempo, è stato osservato un frammento di radio durante 2608 intervalli di tempo di 7.5 secondi e registrato il numero di particelle emesse in ciascun intervallo, per un totale di 10 094 particelle. Si è potuta così costruire la distribuzione di frequenza del numero di occorrenze (numero di particelle emesse) per unità di tempo, che è presentata nelle prime due colonne della Tabella 14.1. Queste cifre mostrano che sono state osservate 0 particelle nel 2.19% dei casi, 1 particella nel 7.78% dei casi e così via. Si stabilisca se la distribuzione di Poisson è adatta a rappresentare questo fenomeno aleatorio.

bella 14.1
prossimazione di una tribuzione di frequenza pirica con la tribuzione di Poisson.

N. di particelle nel singolo intervallo (x)	Frequenza relativa osservata	Probabilità $f(x)$
0	0.0219	0.0209
1	0.0778	0.0807
2	0.1469	0.1562
3	0.2013	0.2015
4	0.2040	0.1949
5	0.1564	0.1509
6	0.1047	0.0973
7	0.0533	0.0538
8	0.0173	0.0260
9	0.0103	0.0112
≥ 10	0.0061	0.0066

Soluzione

Il numero medio di occorrenze per intervallo di tempo è $10\,094/2608 = 3.87$. Assumendo questo valore come media della distribuzione di Poisson, le probabilità fornite dal modello per i diversi valori di X sono quelle riportate nella terza colonna della Tabella 14.1.

Il confronto delle probabilità con le frequenze relative osservate induce a pensare che la distribuzione di Poisson si presti molto bene a descrivere il fenomeno aleatorio considerato. ■

I grafici della Figura 14.4 mostrano come il parametro λ influenzi la forma della distribuzione. In generale, si può dimostrare che la distribuzione di Poisson è asimmetrica positivamente e che l'asimmetria diminuisce all'aumentare di λ.

Si può anche dimostrare che la distribuzione binomiale si avvicina alla distribuzione di Poisson quando n diventa via via più grande e p via via più piccolo in modo tale che np resti fisso e pari a λ.[2] Questo risultato consente di usare la distribuzione di Poisson per approssimare la distribuzione binomiale quando n è molto

[2] Formalmente, si può affermare che

$$\lim_{n\to\infty} \frac{n!}{x!(n-x)!} p^x(1-p)^{n-x} = \frac{\lambda^x}{x!} e^{-\lambda},$$

con $np = \lambda$.

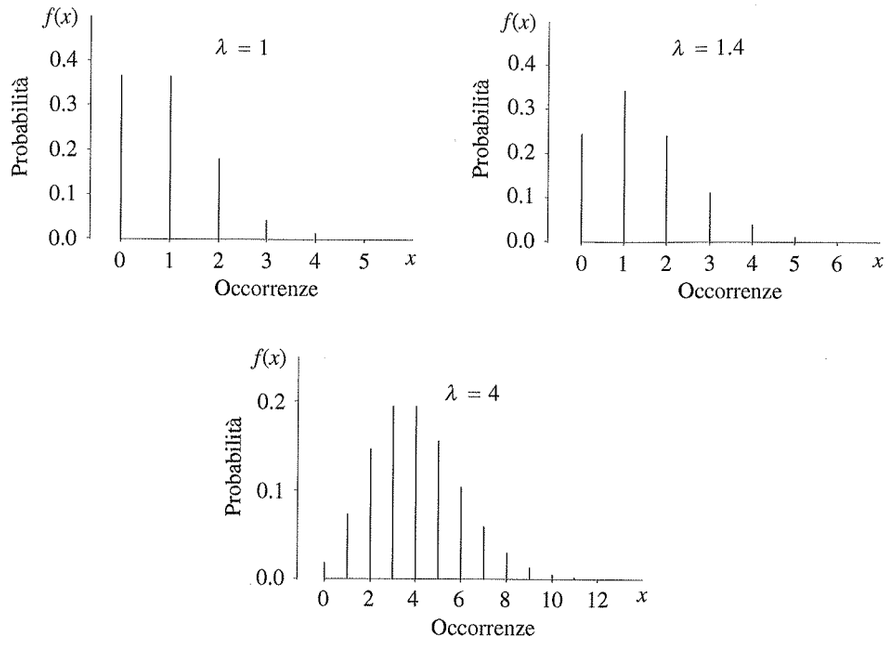


Figura 14.4
Distribuzione di Poisson per alcuni valori di λ.

grande e p è molto piccolo. Un criterio empirico per stabilire se la distribuzione di Poisson approssima adeguatamente la distribuzione binomiale è rappresentato dalla disuguaglianza $n/p > 500$.

ESEMPIO 14.8

Si approssimi la distribuzione binomiale di parametri $p = 0.1$ e $n = 15$ con la distribuzione di Poisson.

Soluzione

La Tabella 14.2 riporta, per valori di X tra 0 e 15, i livelli di probabilità previsti dalla distribuzione binomiale e dalla distribuzione di Poisson di parametro $\lambda = 15 \cdot 0.1 = 1.5$.

N. di successi	Binomiale	Poisson	N. di successi	Binomiale	Poisson
0	0.2059	0.2231	8	0.0000	0.0001
1	0.3432	0.3347	9	0.0000	0.0000
2	0.2669	0.2510	10	0.0000	0.0000
3	0.1285	0.1255	11	0.0000	0.0000
4	0.0428	0.0471	12	0.0000	0.0000
5	0.0105	0.0141	13	0.0000	0.0000
6	0.0019	0.0035	14	0.0000	0.0000
7	0.0003	0.0008	15	0.0000	0.0000

Tabella 14.2
Approssimazione della distribuzione binomiale con la distribuzione di Poisson per $p = 0.01$ e $n = 15$.

Come si può vedere, l'approssimazione è molto buona, sebbene n e p non rispettino la regola empirica sopra richiamata. ■

14.6 Distribuzione uniforme continua

Tra le distribuzioni di variabili casuali continue, quella uniforme, detta anche distribuzione rettangolare, è sicuramente la più semplice.

DEFINIZIONE 14.5

Una variabile casuale, X, ha distribuzione uniforme continua tra a e b se la sua funzione di densità è espressa da

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b, \\ 0, & \text{altrove}, \end{cases}$$

dove a e b sono due costanti.

PROPOSIZIONE 14.5

La media e la varianza della distribuzione uniforme continua tra a e b sono date da:

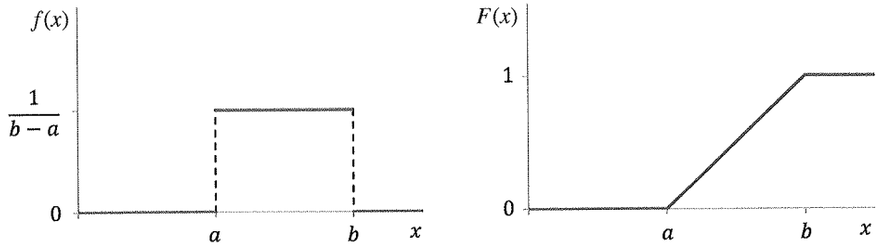
$$\mathrm{E}(X) = \frac{a+b}{2}, \quad \mathrm{Var}(X) = \frac{(a-b)^2}{12}.$$

La dimostrazione è lasciata per esercizio (si veda l'Esercizio 14.11). ■

La funzione di ripartizione, come area sottesa a $f(x)$ tra a e x, è data da (si veda la Figura 14.5)

$$F(x) = \frac{x-a}{b-a},$$

dove $a \le x \le b$.



gura 14.5
nzione di densità e ripartizione della stribuzione uniforme ntinua tra a e b.

Quantili

Il quantile di livello p della distribuzione uniforme continua tra a e b si ricava risolvendo rispetto a x l'equazione

$$F(x) = \frac{x-a}{b-a} = p,$$

da cui

$$x_p = a + p(b-a).$$

Questa operazione è illustrata nella Figura 14.6.

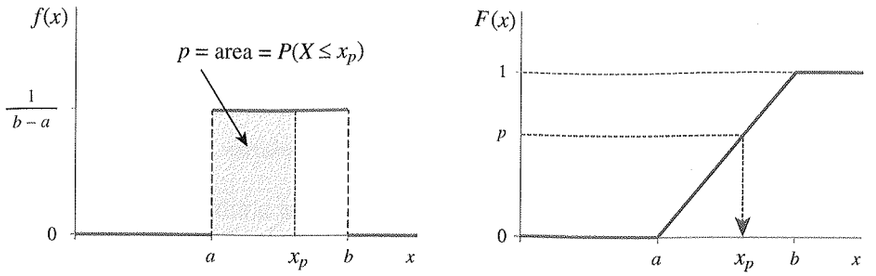


Figura 14.6
Quantile della distribuzione uniforme continua.

14.7 Distribuzione esponenziale

La distribuzione esponenziale è così definita.

DEFINIZIONE 14.6

Una variabile casuale continua, X, ha distribuzione esponenziale se la sua funzione di densità è espressa da

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \ge 0, \\ 0, & x < 0, \end{cases}$$

dove λ è una costante positiva.

PROPOSIZIONE 14.6

La media e la varianza della distribuzione esponenziale di parametro λ sono date da:

$$\mathrm{E}(X) = \frac{1}{\lambda}, \quad \mathrm{Var}(X) = \frac{1}{\lambda^2}.$$

La dimostrazione è lasciata per esercizio (si veda l'Esercizio 14.12). ■

La funzione di ripartizione assume la forma (si veda l'Esercizio 14.12)

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \ge 0. \end{cases}$$

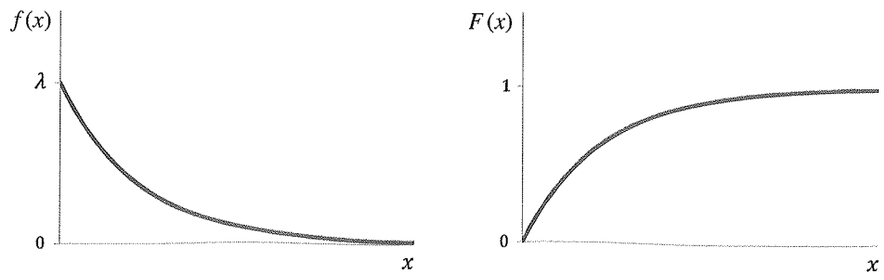


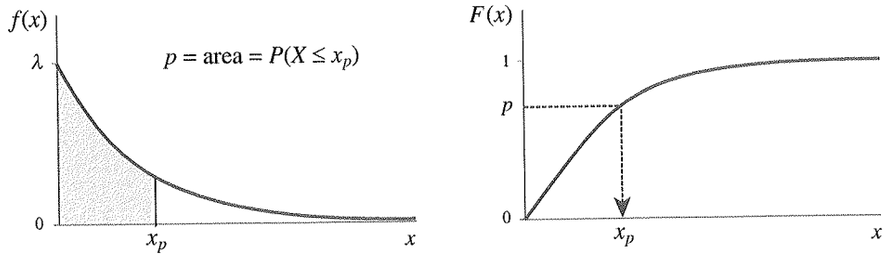
Figura 14.7
Funzione densità e di ripartizione della distribuzione esponenziale.

Quantili

Il quantile di livello p della distribuzione esponenziale si ricava dall'equazione

$$F(x) = 1 - e^{-\lambda x} = p,$$

che, risolta rispetto a x, dà $x_p = -\ln(1-p)/\lambda$.



gura 14.8
uantile di livello p della
stribuzione esponenziale.

14.8 Distribuzione normale

La distribuzione normale o di Gauss è senza dubbio la distribuzione più importante della teoria della probabilità.

DEFINIZIONE 14.7

Una variabile casuale continua, X, ha distribuzione normale se la sua funzione di densità è espressa da

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty, \tag{14.4}$$

dove μ e σ^2 sono due parametri tali che $-\infty < \mu < \infty$ e $0 < \sigma^2 < \infty$.

La funzione di densità (14.4) ha le seguenti proprietà:

- è simmetrica rispetto a $x = \mu$, dato che $f(\mu - \delta) = f(\mu + \delta)$ per ogni δ positivo;
- è crescente nell'intervallo $(-\infty, \mu)$ e decrescente nell'intervallo (μ, ∞); raggiunge il massimo, $1/(\sigma\sqrt{2\pi})$, per $x = \mu$;
- ha l'asse x come asintoto orizzontale, ossia $f(x) \to 0$ per $x \to -\infty$ e per $x \to \infty$;
- ha due punti di flesso in $x = \mu - \sigma$ e $x = \mu + \sigma$; è concava (verso il basso) nell'intervallo $(\mu - \sigma, \mu + \sigma)$ e convessa altrove.

Le prime tre proprietà si desumono direttamente dall'esame della funzione di densità; la quarta richiede lo studio della funzione tramite il calcolo differenziale.

PROPOSIZIONE 14.7

La media e la varianza della distribuzione normale di parametri μ e σ^2 sono date da:

$$\mathrm{E}(X) = \mu, \quad \mathrm{Var}(X) = \sigma^2.$$

La dimostrazione è lasciata per esercizio (si veda l'Esercizio 14.13). ■

La distribuzione normale sarà ampiamente utilizzata nei prossimi capitoli. Per brevità, si adotta la notazione

$$X \sim \mathrm{N}(\mu, \sigma^2)$$

per indicare che la variabile casuale X ha distribuzione normale con media μ e varianza σ^2.

Dunque, la media e la varianza sono i due parametri che definiscono la famiglia delle distribuzioni normali. La media determina la posizione della curva di densità sull'asse delle ascisse. Ad esempio, a valori crescenti di μ corrispondono curve che si spostano verso destra. La varianza determina l'altezza e la larghezza della curva. Una varianza grande produce una curva schiacciata e ampia; una varianza piccola dà luogo a una curva alta e stretta.

La Figura 14.9 illustra come la media e la varianza incidano sulla posizione e sulla forma della normale.

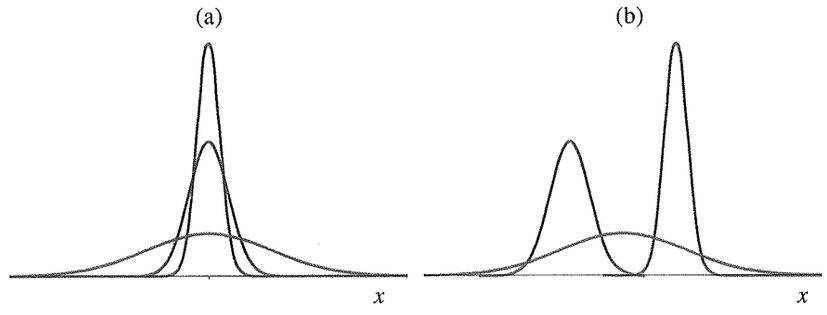


Figura 14.9
Densità normali: (a) con medie uguali e varianze diverse; (b) con medie e varianze diverse.

Come qualsiasi altra funzione di densità, la (14.4) consente di calcolare la probabilità che X assuma valori all'interno di un qualsiasi intervallo (a, b), con a e b costanti arbitrarie. Questa probabilità è data dall'area sottesa alla curva in detto intervallo (si veda la Figura 14.10).

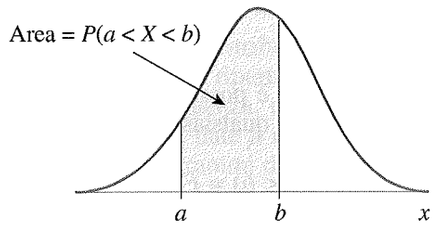
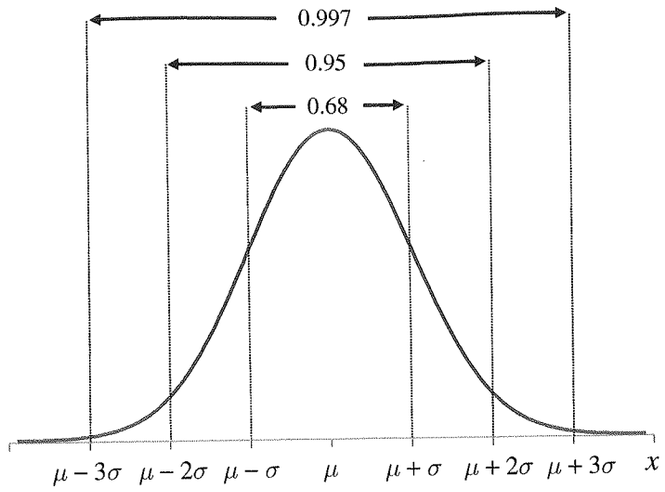


Figura 14.10
Probabilità che $a < X < b$ come area sottesa alla curva tra a e b.

Per farsi un'idea più precisa della struttura della distribuzione normale, si può osservare che le aree sottese alla curva negli intervalli $(\mu-\sigma, \mu+\sigma)$, $(\mu-2\sigma, \mu+2\sigma)$ e $(\mu - 3\sigma, \mu + 3\sigma)$ sono approssimativamente uguali a 0.68, 0.95 e 0.997, rispettivamente (si veda l'Esercizio 14.17). In altri termini: (a) la probabilità di osservare un valore di X entro $1 \cdot \sigma$ al di sotto e al di sopra della media è pari a 0.68; (b) la probabilità di osservare un valore di X entro $2 \cdot \sigma$ al di sotto e al di sopra della media è pari a 0.95; (c) la probabilità di osservare un valore di X entro $3 \cdot \sigma$ al di sotto e al di sopra della media è pari a 0.997. Questa proprietà è denominata regola empirica ed è illustrata nella Figura 14.11.

La distribuzione normale è usata per descrivere un gran numero di fenomeni aleatori in vari campi di applicazione (scienze naturali, scienze sociali, economia ecc.), ma è anche usata per approssimare altre distribuzioni, come quella

ura 14.11
gola empirica.

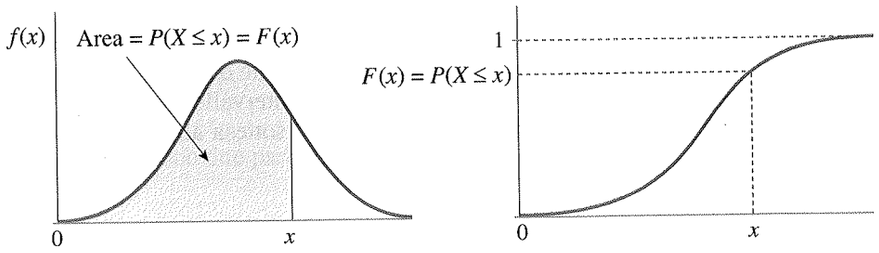


binomiale e quella di Poisson. Questo argomento sarà esaminato in dettaglio nel Capitolo 16.

Funzione di ripartizione

Geometricamente, la funzione di ripartizione della distribuzione normale è l'area sottesa alla curva a sinistra di x,[3] come mostrato nella Figura 14.12.

gura 14.12
unzione di ripartizione
ella distribuzione normale.



Tramite la funzione di ripartizione è possibile determinare la probabilità che X assuma un valore appartenente a un qualsiasi intervallo (a, b):

$$P(a < X < b) = F(b) - F(a). \tag{14.5}$$

Ciò è illustrato nella Figura 14.13: in base alla relazione (14.5), la probabilità che $a < X < b$ è uguale alla differenza tra l'area sottesa alla curva a sinistra di b (area in azzurro scuro + area in azzurro chiaro) e quella sottesa alla curva a sinistra di a (area in azzurro scuro).

Tale probabilità può essere calcolata utilizzando uno dei tanti programmi di calcolo statistico oggi disponibili (Excel, R ecc.), che forniscono i valori della

[3] Con il linguaggio dell'analisi matematica, $F(x)$ è espressa da

$$F(x) = P(X \le x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-\mu)^2/2\sigma^2} \mathrm{d}t.$$

Questo integrale non ha una soluzione analitica, ma è possibile ottenere un'approssimazione numerica molto buona.

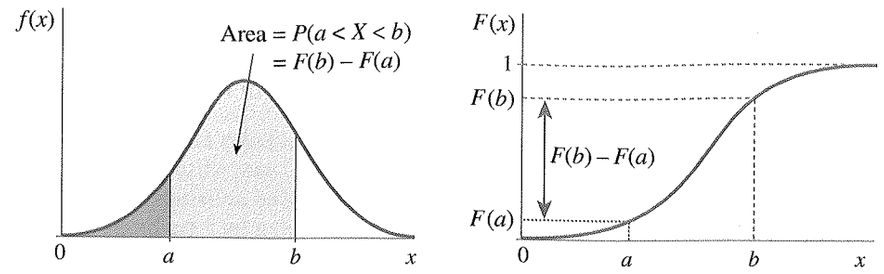


Figura 14.13
Probabilità associata a un intervallo (a, b) come differenza tra i valori di $F(x)$ in b e in a.

funzione di ripartizione di una distribuzione normale $\mathrm{N}(\mu, \sigma^2)$ per valori assegnati di μ e σ^2. In alternativa, si può usare un'apposita tavola che verrà illustrata nel Paragrafo 14.8.2.

Quantili

In base alla definizione generale, il quantile di livello p della distribuzione normale $\mathrm{N}(\mu, \sigma^2)$ è il punto x in cui la funzione di ripartizione $F(x)$ è uguale a p, come mostrato nella Figura 14.14.

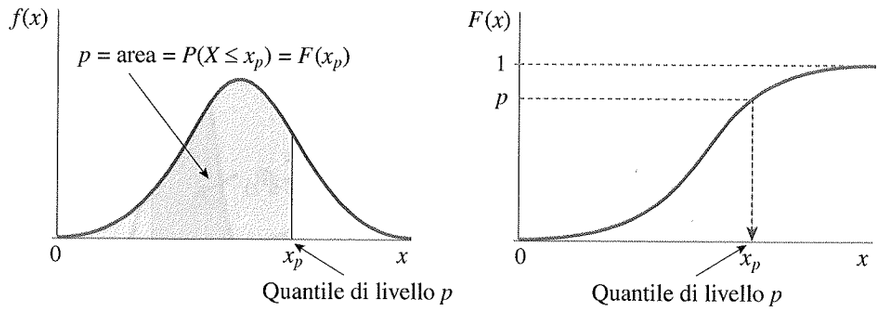


Figura 14.14
Quantile di livello p della distribuzione normale.

Diversamente dai due casi precedenti (distribuzione uniforme continua e distribuzione esponenziale), qui la funzione di ripartizione $F(x)$ non può essere espressa in forma analitica chiusa. Di conseguenza, non esiste una espressione analitica chiusa neppure per x_p, il cui valore, date μ e σ^2, può essere approssimato (con alta precisione) usando uno dei programmi di calcolo statistico sopra richiamati, oppure ricorrendo all'apposita tavola prima evocata.

14.8.1 Distribuzione normale standard

La distribuzione normale standard è una distribuzione normale con media 0 e varianza 1. La sua funzione di densità è

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty.$$

Coerentemente con la simbologia adottata, la distribuzione normale standard viene indicata con $\mathrm{N}(0, 1)$.

La Proposizione 14.8 stabilisce la relazione tra la distribuzione normale $\mathrm{N}(\mu, \sigma^2)$ e la distribuzione normale standard $\mathrm{N}(0, 1)$.

PROPOSIZIONE 14.8

Sia X una variabile casuale con distribuzione normale di media μ e varianza σ^2. Allora la variabile casuale standardizzata

$$Z = \frac{X - \mu}{\sigma}$$

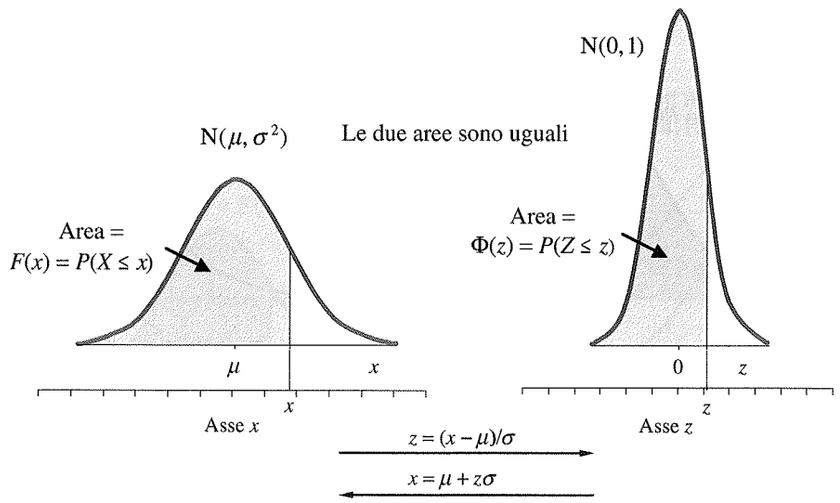
ha distribuzione normale standard. Inoltre, la funzione di ripartizione di X può essere ottenuta come

$$F(x) = P(X \le x) = P\left(Z \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right), \qquad (14.6)$$

dove $\Phi(\cdot)$ è la funzione di ripartizione della distribuzione normale standard.[4]

La (14.6) mostra che l'area sotto la curva normale $N(\mu, \sigma^2)$ a sinistra di un dato punto x è uguale all'area sotto la curva normale standard $N(0, 1)$ a sinistra del punto $(x - \mu)/\sigma$. Questa proprietà è illustrata nella Figura 14.15.

Figura 14.15
Alla distribuzione normale $N(\mu, \sigma^2)$ alla distribuzione $N(0, 1)$ e viceversa.



Sulla base della (14.6), si può scrivere

$$P(a < X < b) = F(b) - F(a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \qquad (14.7)$$

Dunque, la probabilità che una variabile casuale X con distribuzione normale $N(\mu, \sigma^2)$ assuma un valore nell'intervallo (a, b) è uguale alla probabilità che la variabile casuale normale standard $N(0, 1)$ assuma un valore nell'intervallo di estremi $(a - \mu)/\sigma$ e $(b - \mu)/\sigma$.

[4] Per coloro che hanno familiarità con il calcolo integrale, si fa notare che la (14.6) segue da

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-\mu)^2/2\sigma^2}\, dt = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

avendo posto $u = (t - \mu)/\sigma$ e quindi $dt = \sigma du$.

14.8.2 Probabilità e quantili tramite la tavola della distribuzione normale standard

La (14.7) mette in luce l'importanza della funzione di ripartizione $\Phi(z)$ nel calcolo delle probabilità relative alla distribuzione normale $N(\mu, \sigma^2)$. Di seguito viene mostrato come utilizzare i valori di $\Phi(z)$ che sono immancabilmente presentati in ogni testo di statistica nella forma di una tabella a doppia entrata, la tavola della normale standard. Qui è contenuta nell'Appendice C, Tavola C.1 dal titolo "Funzione di ripartizione della distribuzione normale standard". Con essa è possibile calcolare, in modo approssimato, la probabilità che una variabile casuale normale, con media e varianza date, assuma un valore nell'intervallo (a, b), dove a e b sono numeri arbitrari.

La tavola contiene i valori della funzione di ripartizione della distribuzione normale standard per valori non negativi di z compresi tra 0 e 3.49, con un passo di 0.01. I valori di $\Phi(z)$ per i valori negativi di z possono essere ottenuti sfruttando la relazione

$$\Phi(-z) = 1 - \Phi(z), \qquad (14.8)$$

dove z è un numero positivo. La (14.8) è illustrata nella Figura 14.16, in cui si mostra che, data la simmetria della curva normale, per ogni z positivo assegnato, l'area a sinistra di $-z$, $\Phi(-z)$, è uguale all'area a destra di z, $1 - \Phi(z)$. Il motivo della restrizione ai valori di $z \le 3.49$ è che $\Phi(z) \approx 1$ quando $z \ge 3.50$.

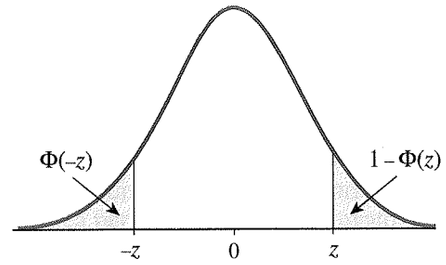


Figura 14.16
Illustrazione grafica della relazione $\Phi(-z) = 1 - \Phi(z)$.

Chiaramente, la variabile casuale normale standard assume valori nell'intervallo $(-\infty, \infty)$. Tuttavia, ai fini applicativi, basta considerare quanto avviene nell'intervallo (−3.49, 3.49). Infatti, la probabilità complessiva di un valore a destra di 3.49 è appena 0.0002, pari alla probabilità complessiva di un valore a sinistra di −3.49. Per questo motivo, i valori di z considerati nella Tabella C.1 non superano 3.49.

Probabilità

Se la variabile casuale ha distribuzione normale standard, la probabilità $P(z_1 < Z < z_2)$ si ottiene immediatamente come

$$P(z_1 < Z < z_2) = \Phi(z_2) - \Phi(z_1),$$

dove $\Phi(z_1)$ e $\Phi(z_2)$ si deducono dalla Tavola C.1.

ESEMPIO 14.9

Si calcoli la probabilità che una variabile casuale con distribuzione normale standard assuma un valore compreso tra −1.25 e 0.58.

Soluzione

Sulla base della Tavola C.1 (nella Figure 14.17 ne è riportato un estratto), si ha

$$\begin{aligned} P(-1.25 < Z < 0.58) &= \Phi(0.58) - [1 - \Phi(1.25)] \\ &= 0.7190 - (1 - 0.8944) \\ &= 0.6134. \end{aligned}$$

■

ura 14.17
colo di $\Phi(z)$ con la tavola
a normale standard.

	Seconda cifra decimale di z									
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
…	…	…	…	…	…	…	…	…	…	…
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
…	…	…	…	…	…	…	…	…	…	…
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
…	…	…	…	…	…	…	…	…	…	…

Se la variabile casuale X ha distribuzione normale $N(\mu, \sigma^2)$, la probabilità che $a < X < b$ si calcola applicando la (14.6) come segue:

- si standardizzano a e b, ottenendo $z_1 = (a - \mu)/\sigma$ e $z_2 = (b - \mu)/\sigma$;
- poi, trovati i valori di $\Phi(z_1)$ e $\Phi(z_2)$ nella Tavola C.1, si applica la (14.7) ottenendo

$$P(a < X < b) = \Phi(z_2) - \Phi(z_1).$$

ESEMPIO 14.10

Il peso delle mele servite in un ristorante ha una distribuzione normale con media 150 g e deviazione standard 12 g. Qual è la probabilità che al prossimo cliente sia servita una mela con peso compreso tra 145 e 165 g?

Soluzione

Standardizzando 145 e 165, si ricava:

$$z_1 = \frac{145 - 150}{12} = -0.42, \; z_2 = \frac{165 - 150}{12} = 1.25.$$

Ne segue che

$$\begin{aligned} P(145 < X < 165) &= \Phi(1.25) - \Phi(-0.42) \\ &= 0.8944 - (1 - 0.6628) = 0.5572. \end{aligned}$$

■

Quantili

Se la distribuzione è normale standard, il problema è trovare il valore di z che soddisfa l'equazione

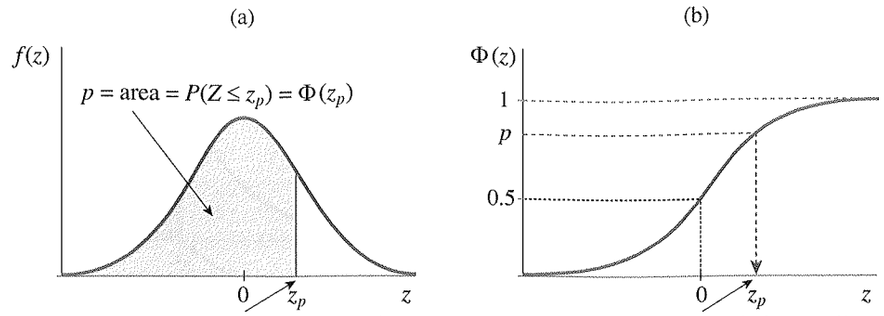
$$\Phi(z) = p,$$

dove p è il livello di probabilità assegnato. La soluzione dell'equazione rispetto a z, z_p, è il quantile cercato (si veda la Figura 14.18 per l'usuale interpretazione geometrica).

Se $p \geq 0.5$, per la soluzione approssimata del problema con la Tavola C.1, è necessario individuare all'interno di questa, la casella contenente la probabilità più vicina a p. Dopodiché, z_p è il numero con due cifre decimali che si ottiene combinando i valori di z della riga e della colonna corrispondenti a quella casella, come illustrato nell'esempio che segue.

ESEMPIO 14.11

Si calcoli il quantile di livello 0.90 (90-esimo centile) della distribuzione normale standard.



Quantile di livello p della N(0, 1) — Quantile di livello p della N(0, 1)

Figura 14.18
Quantile di livello p della distribuzione normale standard.

Soluzione

Si localizzi nella Tavola C.1 il numero più prossimo a 0.90 e si individui il valore di z corrispondente. Come indicato nella Figura 14.19 (estratto della Tavola C.1), il quantile cercato è approssimativamente uguale a 1.28: $z_{0.90} \approx 1.28$.

	Seconda cifra decimale di z									
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
…	…	…	…	…	…	…	…	…	…	…
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
…	…	…	…	…	…	…	…	…	…	…

Figura 14.19
Calcolo del quantile di livello 0.90 con la Tavola C.1.

■

Se $p < 0.5$, il procedimento appena illustrato non si può seguire, perché la Tavola C.1 contiene esclusivamente probabilità ≥ 0.5. Si fa allora ricorso alla relazione (14.8): si individua, all'interno della Tavola C.1, la probabilità più vicina a $1 - p$; si ricava, con lo stesso procedimento descritto sopra, il valore corrispondente di z, z_{1-p}; infine, si moltiplica per -1 il valore così trovato. Il tutto è riassunto nella relazione

$$z_p = -z_{1-p},$$

dove $p < 0.5$.

Come già accennato, i quantili calcolati con la Tavola C.1 sono generalmente approssimati, perché il livello di probabilità p per cui si ricerca il quantile solo raramente coincide esattamente con una delle probabilità contenute nella tavola. Per i valori di p più utilizzati nelle applicazioni (p = 0.90, 0.95, 0.975, 0.99, 0.995, 0.999 ecc.), quantili più precisi (fino alla terza cifra decimale) sono indicati in fondo alla Tavola C.1.

I quantili di una distribuzione normale $N(\mu, \sigma^2)$, con media e varianza qualsiasi, possono essere determinati come segue:

- si calcola, dapprima, il quantile z_p della distribuzione normale standard con il procedimento sopra illustrato;
- si calcola poi il quantile x_p della distribuzione normale $N(\mu, \sigma^2)$ sfruttando la relazione

$$x_p = \mu + z_p\sigma.$$

ESEMPIO 14.12

Nel misurare il livello di glicemia nel sangue, usando un nuovo strumento, l'errore che si commette si distribuisce normalmente con media 1.5 e deviazione standard 3.2. Si determinino il 25-esimo e l'80-esimo centile.

I quantili della distribuzione normale standard sono usati frequentemente nell'inferenza statistica. Il quantile di livello p è il punto sull'asse orizzontale a sinistra del quale l'area sotto la curva normale standard è uguale a p. Vale la pena notare che i quantili realmente utilizzati sono quelli corrispondenti a valori molto piccoli o molto grandi di p (per esempio, $p = 0.01, 0.05, 0.95, 0.99$).

Soluzione

Con il procedimento sopra illustrato, si trova $z_{0.25} \approx -0.67$, da cui

$$x_{0.25} \approx 1.5 - 0.67 \cdot 3.2 = -0.64.$$

In modo analogo, si ricava $x_{0.80} \approx 4.19$. ■

14.9 Distribuzione chi-quadrato

La distribuzione chi-quadrato è una delle più importanti distribuzioni di probabilità per la sua ampia applicazione nell'inferenza statistica. Essa è così definita.

DEFINIZIONE 14.8

Una variabile casuale continua, X, ha distribuzione chi-quadrato se la sua funzione di densità è espressa da

$$f(x) = \begin{cases} ax^{\frac{r}{2}-1}e^{-x/2}, & x \geq 0, \\ 0, & x < 0, \end{cases} \tag{14.9}$$

dove r è un numero intero positivo detto gradi di libertà e a è una costante positiva il cui valore dipende da r.

La funzione di densità (14.9) è asimmetrica positivamente. L'asimmetria si attenua al crescere dei gradi di libertà (si veda la Figura 14.20).

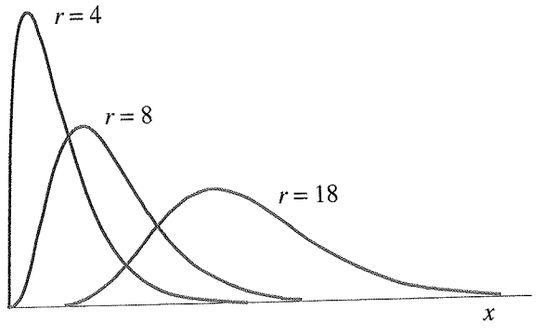


Figura 14.20
Funzione di densità chi-quadrato per $r = 4, 8, 18$.

PROPOSIZIONE 14.9

La media e la varianza della distribuzione chi-quadrato con r gradi di libertà sono date da:

$$\mathrm{E}(X) = r, \quad \mathrm{Var}(X) = 2r.$$

La dimostrazione è contenuta nel Paragrafo B.8 dell'Appendice B. ■

Quantili

La Figura 14.21 mostra il quantile di livello p della distribuzione chi-quadrato.

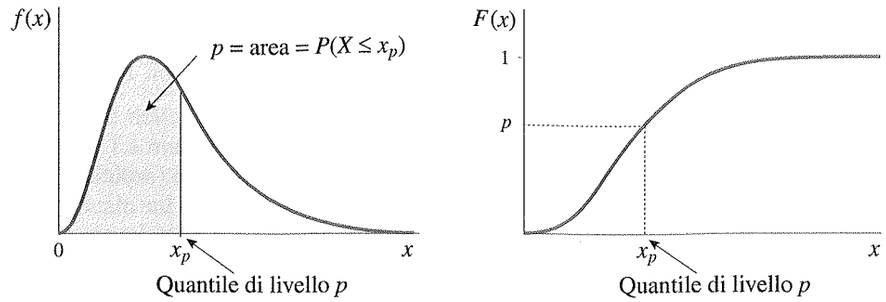


Figura 14.21
Quantile di livello p della distribuzione chi-quadrato.

I quantili della distribuzione chi-quadrato possono essere determinati utilizzando uno dei vari programmi di calcolo statistico oggi disponibili (ad esempio Excel, R ecc.).

In alternativa, la Tavola C.2 dell'Appendice C fornisce i quantili della distribuzione chi-quadrato per combinazioni specifiche di gradi di libertà e livelli di probabilità p. La prima riga di questa tavola contiene 14 livelli di probabilità p in ordine crescente da 0.005 a 0.50 e da 0.75 a 0.999, corrispondenti ai quantili più estremi. La prima colonna della tabella elenca i gradi di libertà da 1 a 40 e poi da 45 a 100 con passo 5. Chiaramente, la singola casella della tabella contiene il quantile della distribuzione chi-quadrato corrispondente al livello di probabilità p e ai gradi di libertà r prescelti.

ESEMPIO 14.13

Si determinino il quantile di livello 0.01 della distribuzione chi-quadrato con 10 gradi di libertà e il quantile di livello 0.99 della distribuzione chi-quadrato con 11 gradi di libertà.

Soluzione

Nella Figura 14.22 è riportato un estratto della Tavola C.2. Il primo quantile da determinare si trova all'intersezione tra la colonna individuata da $p = 0.01$ e la riga corrispondente a 10 gradi di libertà. Si ha allora $x_{0.01} = 2.56$. In modo analogo, si trova $x_{0.99} = 24.73$.

	Area sottesa alla curva a sinistra del quantile													
r	0.005	0.010	0.025	0.050	0.100	0.250	0.500	0.750	0.900	0.950	0.975	0.990	0.995	0.999
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.73	26.76	31.26
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Figura 14.22
Calcolo dei quantili della distribuzione chi-quadrato con la Tavola C.2.

■

Si può dimostrare che, al tendere di r a infinito, la distribuzione chi-quadrato tende alla normale con media r e varianza $2r$. Ne segue che, se r è sufficientemente grande, i quantili della distribuzione chi-quadrato possono essere determinati approssimativamente sfruttando la relazione tra una normale con media r e varianza $2r$ e la normale standard

$$x_p \approx r + z_p\sqrt{2r},$$

dove z_p è il quantile di livello p della normale standard.

ESEMPIO 14.14

Si determini il 95-esimo centile della distribuzione chi-quadrato con 100 gradi di libertà usando l'approssimazione normale.

Soluzione

Si ha

$$x_{0.95} \approx 100 + 1.645\sqrt{200} = 123.26.$$

Il valore esatto è 124.34. ■

iassunto

uesto capitolo si occupa delle distribuzioni da cui non si può prescindere in un volume di statistica di se. Da una parte, le distribuzioni di variabili casuali discrete: uniforme, di Bernoulli, binomiale e di isson. Dall'altra, le distribuzioni di variabili casuali continue: uniforme continua, esponenziale, normale chi-quadrato. Viene dedicata particolare attenzione alle distribuzioni binomiale e normale per il loro portante ruolo nelle tecniche inferenziali oggetto dei prossimi capitoli.

Di ogni distribuzione vengono date la funzione di probabilità (o di densità), la media e la varianza. el caso continuo, vengono definiti anche i quantili e indicate le modalità di calcolo.

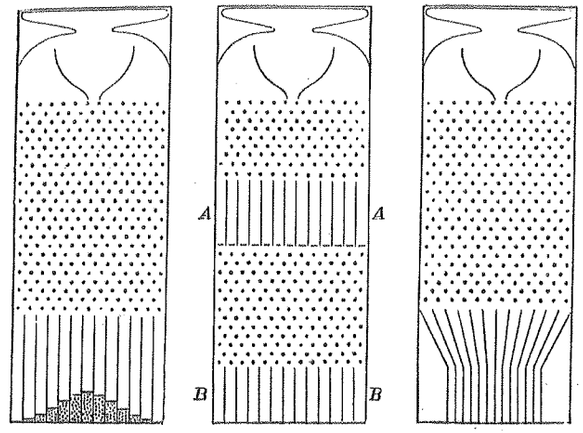
arole chiave

efficiente binomiale, 299
stribuzione binomiale, 299
stribuzione chi-quadrato, 314
stribuzione di Bernoulli, 296
stribuzione di Poisson, 301
stribuzione esponenziale, 305
stribuzione normale, 306
stribuzione normale standard, 309
stribuzione uniforme continua (rettangolare), 304
distribuzione uniforme discreta, 296
famiglia di distribuzioni di probabilità, 295
gradi di libertà, 314
legge degli eventi rari, 301
parametro, 295
prova Bernoulliana, 297
regola empirica, 307
tavola della normale standard, 311

ommenti e curiosità

La macchina di Galton

ancis Galton (1822-1911), cugino di Charles Darwin, svolse un'attività scientifica sta e molteplice: si occupò di meteorologia, antropologia, acustica fisiologica, ma suo nome è soprattutto legato agli studi sull'ereditarietà biologica che lo portarono la creazione della cosiddetta eugenetica, che, secondo Galton, è "lo studio dei fattori ttoposti al controllo sociale che possono migliorare o peggiorare le qualità razziali lle generazioni future, sotto l'aspetto tanto fisico, quanto psichico". In questo cam- di ricerca – mal interpretando il messaggio del cugino Darwin – Galton sviluppò a sua teoria sulla razza umana inaccettabile, che lo portò a conclusioni eticamente erranti che meritano di cadere nell'oblio. Tuttavia, nello sviluppo della sua teoria, li apportò notevoli contributi alla metodologia statistica: effettuò per primo la somministrazione dei que- ionari, sviluppò le tecniche di campionamento, introdusse il metodo dei percentili, per individuare la rma della distribuzione delle variabili quantitative, e il concetto e la misura della correlazione. Nello stu- o sperimentale sull'ereditarietà, introdusse il termine di "inversione"; estendendo questo studio al caso mano, sostituì questo termine con quello molto noto in statistica di "regressione".



Fu anche l'inventore di una particolare macchina, che porta il suo nome, per mostrare meccanicamente, sotto opportune condizioni, l'approssimazione della distribuzione binomiale a quella normale. La macchina di Galton (detta anche scatola di Galton) è un dispositivo formato da un piano verticale sul quale sono piantati perpendicolarmente alcuni pioli o chiodi posizionati secondo la configurazione del quinconce, ossia come la rappresentazione del numero 5 sulla faccia di un comune dado da gioco. Da una apertura, posta in cima a tale piano, viene fatto cadere un certo numero di palline. Sul fondo sono collocati dei raccoglitori cilindrici, dove le palline si depositano l'una sull'altra, formando delle pile. Al termine dell'esperimento, le altezze di queste pile assumono approssimativamente la forma di una curva normale.

2. Distribuzione di Poisson e morti per calcio di cavallo

Nel 1898 Ladislaus Bortkiewicz (1868-1931) pubblicò uno studio sui decessi di soldati dell'esercito prussiano in seguito a calcio di cavallo. Analizzando i verbali di 20 anni di 10 reggimenti dell'esercito prussiano, egli constatò che c'erano stati in totale 122 morti dovuti a questa causa con la distribuzione di frequenza riportata nella tabella che segue:

N. morti	0	1	2	3	4
Frequenza	109	65	22	3	1

Ammettendo l'equivalenza dei reggimenti e la costanza del fenomeno nel tempo, i dati possono essere considerati come il risultato di 200 osservazioni in un anno su un reggimento tipo. Lo studioso notò come la distribuzione di Poisson, la "legge degli eventi rari", fosse adatta a rappresentare il fenomeno. Alcuni storici della statistica osservano che la distribuzione di Poisson dovrebbe essere chiamata "distribuzione di Bortkiewicz".

3. Poisson e Galois

Siméon Denis Poisson (1781-1840), studioso francese, si occupò di matematica, fisica, astronomia e statistica. Nonostante la sua modesta situazione economica, riuscì a entrare all'*École Polytechnique* di Parigi, di cui divenne successivamente docente succedendo a Fourier. Oltre ai sui importanti studi in matematica e fisica, diede notevoli contributi anche nel calcolo delle probabilità e in statistica introducendo la distribuzione che porta il suo nome. Nella sua attività accademica si ricorda l'episodio che lega Poisson al geniale e brillante giovane matematico francese Evariste Galois (1811-1832), padre della teoria che porta il suo nome – la Teoria di Galois, che rappresenta un'importante branca dell'algebra moderna – e fervente repubblicano, che la notte prima di un duello (che lo portò alla morte a soli 20 anni) cercò di sistemare velocemente i suoi brillanti lavori matematici, annotando nelle varie pagine che gli mancava il tempo per un'esposizione più completa e chiara. L'episodio è il seguente: Galois inviò a Poisson un breve riassunto dei suoi studi, chiedendogli di presentare il suo lavoro all'Accademia. Nello stesso anno, mentre era in carcere, in seguito alla sua attività rivoluzionaria, Galois ricevette la risposta di Poisson che rifiutava il lavoro, affermando che l'esposizione non era chiara, e lo invitava a renderla più rigorosa e comprensibile. Probabilmente, negli scritti di Galois era necessaria una maggiore chiarezza espositiva; quello che però si può dire, con il senno del poi, è che sicuramente Poisson perse un'occasione importante per accreditare in ambito accademico e scientifico un giovane studioso che ben presto si rivelerà come uno dei più brillanti matematici della storia moderna.

Distribuzione esponenziale e assenza di memoria

r una variabile casuale X con distribuzione esponenziale si vede facilmente che, se a e b sono due valori merici positivi, $P(X > a + b \mid X > a) = P(X > b)$. Infatti,

$$P(X > a+b \mid X > a) = \frac{P(X > a+b,\, X > a)}{P(X > a)}$$
$$= \frac{P(X > a+b)}{P(X > a)} = \frac{e^{-\lambda(a+b)}}{e^{-\lambda a}} = e^{-\lambda b} = P(X > b).$$

uesta caratteristica prende il nome di proprietà di assenza di memoria: se X rappresenta un tempo di attesa er esempio in minuti), la probabilità condizionata di aspettare altri b minuti, dato che si è già aspettato r a minuti, è uguale alla probabilità (non condizionata) di aspettare b minuti.

Il peana di Youden alla curva normale

funzione di densità normale fu introdotta nel 1733 da Abraham de Moivre (1667-1754) come ap-ossimazione della distribuzione binomiale. Più tardi fu utilizzata da Carl Friedrich Gauss (1777-1855) ll'ambito dello studio degli errori accidentali.

Tra le altre curiosità sulla distribuzione normale, merita menzione l'inno di lode dedicato alla curva llo statistico William John Youden (1900-1971):

ura 14.23
ana di Youden
a curva normale.

THE
NORMAL
LAW OF ERROR
STANDS OUT IN THE
EXPERIENCE OF MANKIND
AS ONE OF THE BROADEST
GENERALIZATIONS OF NATURAL
PHILOSOPHY + IT SERVES AS THE
GUIDING INSTRUMENT IN RESEARCHES
IN THE PHYSICAL AND SOCIAL SCIENCES AND
IN MEDICINE AGRICULTURE AND ENGINEERING +
IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE
INTERPRETATION OF THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT

Gauss e la curva normale

arl Friedrich Gauss (1777-1855) è considerato uno dei maggiori matematici del-storia; egli ricevette il titolo di *princeps mathematicorum* (principe dei mate-atici). Oltre che di matematica, Gauss si occupò di fisica, geodesia e astrono-ia. È quindi considerato uno dei più grandi geni scientifici di tutti i tempi. Sin bambino Gauss mostrò un'eccezionale capacità aritmetica, riuscendo a trovare luzioni rapide a problemi non semplici: egli fu infatti un bambino prodigio. Esi-ono diversi aneddoti riguardo alla sua precocità; per esempio, si racconta che da mbino, il suo maestro Büttner, per tranquillizzare i suoi vivaci scolari, ordinò ro di fare la somma dei numeri interi da 1 a 100. Il bambino Gauss diede in chi secondi la risposta esatta, sorprendendo il maestro. Si narra che il metodo guito da Gauss fu il seguente: mise in una riga i 100 numeri e sommò il primo l'ultimo (1+100) ottenendo 101; poi sommò il secondo e il penultimo (2+99) tenendo sempre 101 e così via fino a sommare infine 50 e 51 ottenendo sempre 101. Avendo ottenuto 101 r 50 volte, ottenne che la somma totale dei numeri da 1 a 100 era pari a $50 \cdot 101 = 5050$. In particolare, auss "scoprì" e immediatamente applicò la formula che dà la somma di un numero finito di termini di na progressione aritmetica. Il duca di Brunswick, impressionato dalle sue capacità, finanziò gli studi di auss al *Collegium Carolinum*. Successivamente, Gauss studiò all'Università di Gottinga, dove formulò na serie di importanti teoremi.

Georgios Kollidas/Alamy Stock Photo

Egli fu un perfezionista e un grande lavoratore. Non fu uno scrittore molto prolifico, rifiutando di pubblicare qualcosa che non fosse assolutamente perfetto. Il suo motto era infatti *pauca sed matura*. I suoi diari personali testimoniano che egli compì molte importanti scoperte matematiche anni o decenni prima che i suoi contemporanei le pubblicassero. Alcuni storici della matematica sostengono che, se Gauss avesse pubblicato per tempo tutte le sue scoperte, avrebbe anticipato i matematici di almeno cinquant'anni.

Fra i principali contributi in matematica, Gauss diede la prima dimostrazione rigorosa del cosiddetto teorema fondamentale dell'algebra. In astronomia, tra l'altro, introdusse metodi poi divenuti fondamentali nella moderna teoria del calcolo delle probabilità e della statistica. A soli 17 anni, ancor prima di entrare all'università, Gauss ideò il metodo dei minimi quadrati che applicò successivamente allo studio delle orbite di piccoli pianeti appena allora scoperti (Cerere, Pallade, Giunone) e di comete apparse in quegli anni, al fine di ridurre al minimo gli errori di misurazione. Gauss sostenne di aver inventato e utilizzato il metodo dei minimi quadrati per primo, anche se la prima pubblicazione ufficiale sul metodo è di Legendre. Sicuramente a Gauss si può attribuire il merito di aver formalizzato il metodo sotto le ipotesi di normalità degli errori. Nella sua *Theoria motus corporum coelestium*, Gauss perfezionò il metodo dei minimi quadrati per la correzione degli errori e utilizzò la distribuzione normale, detta anche distribuzione degli errori accidentali o distribuzione gaussiana. Egli ottenne la distribuzione normale come modello della distribuzione degli errori di misura di grandezze astronomiche. Si noti che, benché porti il nome di Gauss, la distribuzione normale fu "scoperta" da studiosi precedenti, come De Moivre e Laplace; Gauss la riscoprì nella misurazione della posizione degli astri.

Esercizi

14.1* Si determinino la media e la varianza di una variabile casuale con distribuzione uniforme discreta sui primi n numeri naturali. Si noti che per questo è opportuno utilizzare le seguenti due proprietà, relative alla somma dei primi n numeri naturali e alla somma dei quadrati dei primi n numeri naturali:

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}; \quad \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}.$$

14.2* Si dimostri che una variabile casuale discreta con distribuzione di Bernoulli ha media p e varianza $p(1-p)$.

14.3 Data una variabile casuale discreta X con distribuzione binomiale di parametri $n = 6$ e $p = 0.7$, si effettuino le elaborazioni sotto indicate:

a) si costruisca la distribuzione di probabilità di X disponendo in una tabella le quantità x e $f(x)$;

b) si calcolino la media e la varianza della distribuzione con i dati della tabella di cui al punto precedente, applicando le Formule (13.1) e (13.3);

c) si controllino i risultati del punto precedente utilizzando le (14.2).

14.4 La segreteria di uno studio dentistico sa per esperienza che il 10% dei pazienti arriva in ritardo rispetto all'appuntamento stabilito. Si determini la probabilità che su sette pazienti scelti a caso due arrivino in ritardo.

14.5 In una clinica ostetrica l'osservazione delle nascite in un lungo periodo di tempo ha indicato che, in media, su 100 nati vivi 51 sono di sesso maschile. Se in una data settimana sono ammesse 6 donne in attesa di parto, si determini la probabilità degli eventi sotto indicati:

a) nascano tanti maschi quante femmine;

b) il numero dei maschi sia inferiore a quello delle femmine;

c) nascano almeno 4 maschi.

14.6 Una compagnia aerea sostiene che il 95% dei propri voli arriva in orario. Considerando dieci voli scelti a caso, si determinino le probabilità degli eventi sotto indicati:

a) almeno sette voli siano in orario;

b) al più otto voli siano in orario;

c) non più di due voli siano in ritardo.

14.7 Il numero di appartamenti venduti da un'agenzia immobiliare in una settimana è una variabile casuale con distribuzione di Poisson di media $\lambda = 3$; le probabilità che l'agenzia venda $0, 1, 2, 3, \ldots, 10, 11$ appartamenti sono date, rispettivamente, da 0.0498, 0.1494, 0.2240, 0.2240, 0.1680, 0.1008, 0.0504, 0.0216, 0.0081, 0.0027, 0.0008, 0.0002. Si effettuino le elaborazioni sotto indicate:

a) si calcoli la media della distribuzione e si verifichi che (salvo il margine di errore dovuto agli arrotondamenti) essa coincide con λ;

b) si determini la deviazione standard della distribuzione e si verifichi che (fatta salva l'approssimazione dovuta agli arrotondamenti) essa coincide con $\sqrt{\lambda}$, in linea con la (14.3).

14.8 Il numero di telefonate in arrivo in un minuto al centralino di una grande azienda segue la distribuzione di Poisson con media 1.6. Si determini la probabilità degli eventi sotto indicati:

a) in un minuto, non arrivi alcuna telefonata;

b) in un minuto, arrivino almeno due telefonate.

14.9 Il numero di errori di battitura per pagina commessi da una esperta dattilografa segue la distribuzione di Poisson con una media di tre errori per pagina. Si determini la probabilità che, in una pagina scelta a caso, compaiano più di tre errori.

14.10 La probabilità che un'agenzia immobiliare venda un appartamento in una contrattazione è 0.036. Si determini, con la distribuzione di Poisson, la probabilità che in 100 contrattazioni l'agenzia venda almeno un appartamento e si confronti il risultato con quello fornito dalla binomiale.

14.11* Si determinino la media e la varianza di una variabile casuale X con distribuzione uniforme continua tra a e b.

14.12* Si consideri una variabile casuale X avente distribuzione esponenziale.

a) Si dimostri che $F(x) = 1 - e^{-\lambda x}$.

b) Si verifichi che la media e la varianza di X sono, rispettivamente, $1/\lambda$ e $1/\lambda^2$.

14.13* Si dimostri che la distribuzione normale ha media μ e varianza σ^2. (*Suggerimento*: per quanto riguarda il primo punto, l'integrale

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}\, \mathrm{d}x$$

va risolto per sostituzione, ponendo $z = (x - \mu)/\sigma$; per il secondo punto, conviene scrivere la varianza come

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2 = \int_{-\infty}^{\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}\, \mathrm{d}x - \mu^2$$

e poi procedere mediante integrazione per parti.)

14.14 Una variabile casuale X ha distribuzione normale con media 160 e deviazione standard 4.8.

a) Si determini il numero a per cui $P(X \leq a) = 0.56$.

b) Si determini il numero b per cui $P(X \geq b) = 0.12$.

c) Si determini il numero c per cui $P(X \leq c) = 0.27$.

14.15 Una variabile casuale si distribuisce secondo una normale standard. Si determinino le quantità di seguito indicate:

a) gli estremi dell'intervallo con centro 0 entro cui la probabilità è pari a 0.60;

b) il numero che è superato con probabilità 0.15;

c) il numero alla cui sinistra vi è una probabilità pari a 0.32.

14.16 La durata del battistrada degli pneumatici di una data marca è distribuita normalmente con media 42 300 km e deviazione standard 5300. Si effettuino le elaborazioni di seguito indicate.

a) Si calcoli la probabilità che la durata del battistrada sia superiore a 43 000 km.

b) Si determini la probabilità che la durata del battistrada sia compresa tra 37 000 e 45 000 km.

14.17 Si consideri una variabile casuale normale $\mathrm{N}(\mu, \sigma^2)$. Si verifichi che le probabilità associate agli intervalli $(\mu - \sigma, \mu + \sigma)$, $(\mu - 2\sigma, \mu + 2\sigma)$ e $(\mu - 3\sigma, \mu + 3\sigma)$ sono 0.68, 0.95 e 0.997 (fatta salva l'approssimazione dovuta agli arrotondamenti).

14.18 Il tempo impiegato da un meccanico per assemblare un certo tipo di bicicletta può essere considerato una variabile casuale con distribuzione normale con media 32 minuti e deviazione standard 3.5 minuti. Indicando con T il tempo impiegato per assemblare la bicicletta acquistata, si calcolino le probabilità che:

a) T non superi 25 minuti;

b) T sia compreso tra 26 e 33 minuti;

c) T sia compreso tra 25 e 28 minuti.

14.19 Le sfere di acciaio prodotte da una fabbrica hanno un diametro che segue la distribuzione normale. Sapendo che il 20% delle sfere ha un diametro inferiore a 1.2 cm e che il 10% ha un diametro maggiore di 1.4 cm, si effettuino le elaborazioni di seguito indicate.

a) Si determini il valore atteso e la deviazione standard del diametro di una sfera scelta a caso.

b) Sulla base dei risultati ottenuti nel punto precedente, si determini la probabilità che il diametro della sfera sia maggiore di 1.5.

14.20 La durata in mesi di un componente di un'apparecchiatura elettronica segue la distribuzione chi-quadrato con 25 gradi di libertà. Si determinino le quantità sotto indicate:

a) la durata media;

b) la durata mediana;

c) la probabilità che la durata sia compresa tra 20 e 40 mesi (si usi un foglio elettronico o la Tavola C.2, prendendo i numeri che più si avvicinano agli estremi considerati).

Distribuzioni di probabilità con Excel

Qui di seguito sono indicate le funzioni Excel riguardanti le distribuzioni di probabilità binomiale, di Poisson, normale e chi-quadrato:

Binomiale: DISTRIB.BINOM.N( );

Poisson: DISTRIB.POISSON( );

Normale: DISTRIB.NORM.N( );

Chi-quadrato: DISTRIB.CHI.QUAD( ).

onviene illustrare in dettaglio il funzionamento della prima funzione.

Si apre un *file* Excel o un foglio di lavoro vuoto. Si inseriscono in una colonna i valori che la variabile casuale può assumere. Si inserisce in una casella a scelta la funzione DISTRIB.BINOM.N( ) preceduta dall'uguale. Appaiono, come mostrato nella Figura 14.24, gli argomenti da inserire tra parentesi, che sono: il numero dei successi, il numero delle prove, la probabilità di successo e "vero" o "falso" per invocare, rispettivamente, la funzione di ripartizione e la funzione di probabilità. Chiaramente, conviene mettere come numero di successi il riferimento alla prima cella della colonna dei dati.

gura 14.24
nzione
STRIB.BINOM.N( ) per
a distribuzione binomiale
parametri 8 e 0.4.

DISTRIB.BINOM.N =DISTRIB.BINOM.N(B3;8;0.4;FALSO)

	A	B	C	D	E	F	G	H	I	J	K
1		Distribuzione binomiale									
2											
3			0	=DISTRIB.BINOM.N(B3;8;0.4;FALSO)							
4			1								

Inseriti gli argomenti, si preme "Invio" e si ottiene il risultato nella stessa casella in cui è stata scritta la funzione.

Copiando la funzione nelle celle successive, si ottiene l'intera distribuzione di probabilità binomiale (funzione di probabilità o funzione di ripartizione a seconda della scelta effettuata). Il foglio dei risultati viene mostrato nella Figura 14.25 e si riferisce a una distribuzione binomiale con parametri 8 e 0.4.

igura 14.25
oglio dei risultati
ella funzione
ISTRIB.BINOM.N( ).

C3 =DISTRIB.BINOM.N(B3;8;0.4;FALSO)

	A	B	C	D	E	F	G	H	I	J	K
1		Distribuzione binomiale									
2											
3			0	0.016796							
4			1	0.08958							
5			2	0.209019							
6			3	0.278692							
7			4	0.232243							
8			5	0.123863							
9			6	0.041288							
10			7	0.007864							
11			8	0.000655							

La funzione DISTRIB.POISSON( ) richiede come argomenti il valore di X (numero di occorrenze), la media e "vero" o "falso" per invocare, rispettivamente, la funzione di ripartizione o la funzione di probabilità. l procedimento di applicazione è identico a quello descritto nel punto precedente.

La funzione DISTRIB.NORM.N( ) ha come argomenti il valore di X, la media, la deviazione standard e "vero" o "falso" a seconda che si voglia la funzione di ripartizione o la funzione di densità. Con la Figura 14.26 viene illustrata l'applicazione della funzione per il calcolo della funzione di ripartizione, nel punto $X = 172$, di una distribuzione normale con media 175 e deviazione standard 7.5.

Figura 14.26
Funzione
DISTRIB.NORM.N( ) per
una distribuzione normale
di parametri 175 e 7.5^2.

DISTRIB.NORM.N =DISTRIB.NORM.N(B3;175;7.5;VERO)

	A	B	C	D	E	F	G	H	I	J	K
1		Distribuzione normale									
2											
3			172	=DISTRIB.NORM.N(B3;175;7.5;VERO)							
4											

Nel caso della normale standard, il valore della funzione di densità o di ripartizione può essere ottenuto con la funzione DISTRIB.NORM.ST.N( ), che ha come unico argomento il valore di X.

La funzione DISTRIB.CHI.QUAD( ) ha come argomenti il valore di X, i gradi di libertà e "vero" o "falso" a seconda che si voglia la funzione di ripartizione o la funzione di densità.

Per le variabili casuali normale e chi-quadrato sono disponibili le funzioni INV.NORM.N( ) e INV.CHI.QUAD( ) per determinare i quantili di livello p. La prima ha tre argomenti: p, la media e la deviazione standard. La seconda ha due argomenti: p e i gradi di libertà. Nella Figura 14.27, viene illustrato il calcolo del decimo centile di una distribuzione normale con media 175 e deviazione standard 7.5.

DISTRIB.NORM.N =INV.NORM.N(B3;175;7.5)

	A	B	C	D	E	F	G	H	I	J	K
1		Calcolo dei quantili di una normale									
2											
3			0.1	=INV.NORM.N(B3;175;7.5)							

Figura 14.27
Funzione INV.NORM.N( ) per una distribuzione normale di parametri 175 e 7.5^2.

CAPITOLO

15

Distribuzioni di probabilità congiunte

Sommario del capitolo

15.1 Introduzione
15.2 Variabili casuali doppie
15.3 Distribuzione di probabilità congiunta di due variabili casuali discrete
15.4 Distribuzione di probabilità congiunta di due variabili casuali continue
15.5 Distribuzione di probabilità congiunta di *n* variabili casuali
Riassunto
Commenti e curiosità
Esercizi

15.1 Introduzione

Nei Capitoli 13 e 14, sono state studiate le distribuzioni di probabilità di variabili casuali singole. Nelle applicazioni pratiche, è talora necessario esaminare simultaneamente due o più variabili casuali, perché l'esperimento casuale prevede che si associno due o più numeri reali a ogni evento elementare. Per esempio, ai fini dello studio della relazione tra livello di istruzione e remunerazione tra gli addetti di un dato settore economico, si dovranno osservare, su soggetti estratti a sorte, due numeri: gli anni di scuola frequentati e il reddito.

In questo capitolo, si parlerà di distribuzioni di probabilità che descrivono il comportamento congiunto di due o più variabili casuali. Verrà prima introdotta, per semplicità, la funzione di probabilità congiunta, $f(x, y)$, di due variabili casuali discrete X e Y. Saranno poi studiati i seguenti aspetti: la determinazione, a partire da $f(x, y)$, delle distribuzioni di probabilità delle variabili casuali singole X e Y (distribuzioni di probabilità marginali); la definizione della distribuzione di probabilità di una variabile casuale ammettendo che l'altra abbia assunto un determinato valore (distribuzioni di probabilità condizionate); la definizione del concetto di indipendenza tra X e Y; l'illustrazione di due misure di associazione tra X e Y, la covarianza e il coefficiente di correlazione lineare.

Queste nozioni saranno riesaminate sinteticamente per il caso di due variabili casuali continue e poi in parte estese a n variabili casuali (discrete o continue), concentrando l'attenzione su due aspetti rilevanti per le applicazioni: la distribuzione di probabilità di n variabili casuali indipendenti e quella delle combinazioni lineari di n variabili casuali.

15.2 Variabili casuali doppie

Conviene partire da un esempio introduttivo.

ESEMPIO 15.1

Si riprenda l'esperimento di cui all'Esempio 13.1. Si associ a ogni evento elementare, oltre al numero di teste, il numero delle variazioni di sequenza, cioè la frequenza con cui si passa dal simbolo T al simbolo C o viceversa. Così l'evento TTC dà luogo a una variazione, l'evento TCT a due e così via. Sia X il numero di teste e Y il numero delle variazioni di sequenza.

A ciascun evento elementare dello spazio campionario è allora associata una coppia di numeri, come indicato nella Tabella 15.1.

bella 15.1
sociazione di coppie numeri agli eventi ementari dello spazio mpionario di cui Esempio 13.1.

Spazio camp.		X	Y
TTT	ω_1	3	0
TTC	ω_2	2	1
TCT	ω_3	2	2
TCC	ω_4	1	1
CTT	ω_5	2	1
CTC	ω_6	1	2
CCT	ω_7	1	1
CCC	ω_8	0	0

Data la coppia di numeri (x, y), la probabilità da associare all'evento $(X = x) \cap (Y = y)$, indicata con $P(X = x, Y = y)$, è data dalla somma delle probabilità degli eventi elementari nei quali si realizza la doppia uguaglianza: $X = x$, $Y = y$. Così alla coppia $(3, 0)$ è associatala probabilità $P(X = 3, Y = 0) = 1/8$, perché vi è un solo evento elementare (avente probabilità $1/8$) in cui $X = 3$ e $Y = 0$. Alla coppia $(2, 1)$ corrisponde la probabilità $P(X = 2, Y = 1) = 2/8$, perché vi sono due eventi elementari (entrambi aventi probabilità $1/8$) in cui $X = 2$ e $Y = 1$. Procedendo in questo modo per tutte le coppie di numeri possibili, si ottengono le probabilità esposte nella Tabella 15.2.

bella 15.2
obabilità delle coppie numeri (x, y) di cui a Tabella 15.1.

X	Y 0	1	2	Totale
0	0.125	0	0	0.125
1	0	0.250	0.125	0.375
2	0	0.250	0.125	0.375
3	0.125	0	0	0.125
Totale	0.250	0.500	0.250	1

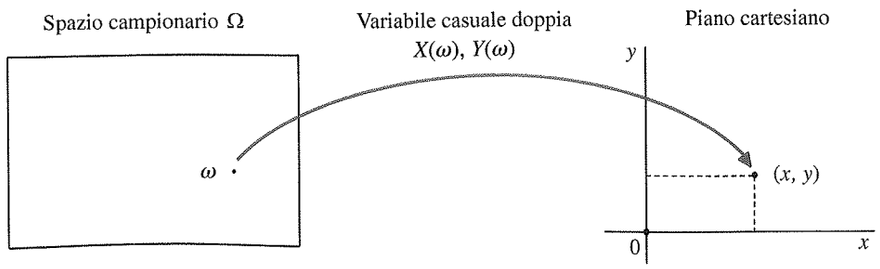
Si dà ora la definizione di *variabile casuale doppia.*

DEFINIZIONE 15.1

Con riferimento a uno spazio campionario Ω, una variabile casuale doppia, (X, Y), è una coppia di funzioni $X(\omega)$ e $Y(\omega)$ che associa una coppia di numeri reali, (x, y), a ogni evento elementare $\omega \in \Omega$.

Questa nozione è rappresentata schematicamente nella Figura 15.1.

gura 15.1
allo spazio campionario Ω piano cartesiano.



15.3 Distribuzione di probabilità congiunta di due variabili casuali discrete

Una variabile casuale doppia (X, Y) si dice discreta se sia X sia Y sono discrete. Analogamente al caso di una sola variabile casuale, le probabilità associate alle singole coppie di valori (x, y) assunti dalla variabile casuale doppia vengono descritte con la funzione di probabilità congiunta così definita

$$f(x, y) = P(X = x, Y = y). \tag{15.1}$$

Nella (15.1), $P(X = x, Y = y)$ è la probabilità dell'evento $\{\omega : X(\omega) = x, Y(\omega) = y\}$, cioè la probabilità dell'insieme di eventi elementari per i quali $X = x$ e $Y = y$.

La funzione di probabilità congiunta gode delle due seguenti proprietà:

$$f(x, y) \geq 0, \quad \sum_x \sum_y f(x, y) = 1,$$

dove la somma è estesa a tutti i possibili valori di X e di Y.

Quando sia X che Y assumono un numero limitato di valori, è utile rappresentare la distribuzione di probabilità congiunta con una tabella a doppia entrata (si veda la Tabella 15.3).

ESEMPIO 15.2

Si consideri l'esperimento del lancio di due dadi (si veda la Figura 13.2) e si associ a ciascun evento elementare una coppia di numeri, il primo dei quali è la differenza in valore assoluto dei numeri che appaiono sulle due facce e il secondo è la somma dei medesimi numeri. Così, per esempio, all'evento elementare $(1, 1)$ è associata la coppia $(0, 2)$, a entrambi gli eventi elementari $(1, 2)$ e $(2, 1)$ è associata la coppia $(1, 3)$ e così via.

Tabella 15.3
Funzione di probabilità congiunta desunta dall'esame della Figura 13.2.

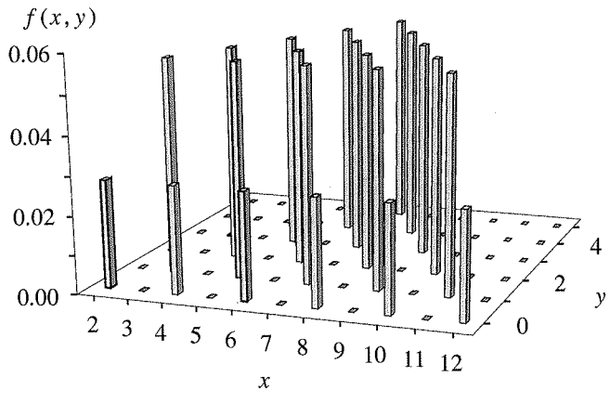
X	Y 2	3	4	5	6	7	8	9	10	11	12	Totale
0	1/36	0	1/36	0	1/36	0	1/36	0	1/36	0	1/36	6/36
1	0	1/18	0	1/18	0	1/18	0	1/18	0	1/18	0	5/18
2	0	0	1/18	0	1/18	0	1/18	0	1/18	0	0	4/18
3	0	0	0	1/18	0	1/18	0	1/18	0	0	0	3/18
4	0	0	0	0	1/18	0	1/18	0	0	0	0	2/18
5	0	0	0	0	0	1/18	0	0	0	0	0	1/18
Totale	1/36	1/18	3/36	2/18	5/36	3/18	5/36	2/18	3/36	1/18	1/36	1

È chiaro che ogni casella della tabella contiene il valore della funzione di probabilità congiunta $f(x, y)$, cioè la probabilità della coppia (x, y) che corrisponde a quella casella. Per esempio:

$$f(0, 2) = \frac{1}{36}, \quad f(2, 8) = \frac{1}{18}, \quad f(3, 9) = \frac{1}{18}.$$

La funzione di probabilità congiunta si rappresenta mediante un sistema di tre assi cartesiani: i punti del piano rappresentano le coppie (x, y), mentre la terza coordinata rappresenta le probabilità associate ai singoli punti del piano. La Figura 15.2 è relativa alla funzione di probabilità congiunta della Tabella 15.3.

ura 15.2
ppresentazione grafica
a funzione di probabilità
giunta della
ella 15.3.



15.3.1 Distribuzioni di probabilità marginali

Data la funzione di probabilità congiunta $f(x, y)$, è possibile ottenere facilmente le funzioni di probabilità delle singole variabili casuali. Si consideri, per esempio, la funzione di probabilità congiunta della Tabella 15.3. La funzione di probabilità di X si ottiene semplicemente sommando le probabilità per riga:

$$\begin{aligned} P(X=0) &= f(0,2)+f(0,3)+\cdots+f(0,12)=6/36,\\ P(X=1) &= f(1,2)+f(1,3)+\cdots+f(1,12)=5/18,\\ &\vdots\\ P(X=5) &= f(5,2)+f(5,3)+\cdots+f(5,12)=1/18. \end{aligned}$$

In modo analogo, si può ottenere la funzione di probabilità di Y.

DEFINIZIONE 15.2

La funzione di probabilità di X, indicata con $f_X(x)$, è data da

$$f_X(x)=\sum_y f(x,y),$$

dove la somma è estesa a tutti i possibili valori di Y. Analogamente, la funzione di probabilità di Y, indicata con $f_Y(y)$, è data da

$$f_Y(y)=\sum_x f(x,y),$$

dove la somma è estesa a tutti i possibili valori di X.

Le funzioni di probabilità $f_X(x)$ e $f_Y(y)$ si chiamano funzioni di probabilità marginali o distribuzioni di probabilità marginali di X e di Y.

ESEMPIO 15.3

Si consideri la distribuzione di probabilità congiunta della Tabella 15.3 e si trovino le distribuzioni di probabilità marginali di X e di Y.

Soluzione

La distribuzione di probabilità marginale di X è costituita dai sei valori di X e dalle rispettive probabilità contenute nell'ultima colonna della Tabella 15.3:

Valore di X	0	1	2	3	4	5
$f_X(x)$	1/6	5/18	2/9	1/6	1/9	1/18

Analogamente, la distribuzione di probabilità marginale di Y è data dagli 11 valori di Y e dalle rispettive probabilità che si trovano nell'ultima riga della Tabella 15.3:

Valore di Y	2	3	4	5	6	7	8	9	10	11	12
$f_Y(y)$	1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36

■

Le distribuzioni di probabilità marginali possono essere rappresentate sinteticamente con le costanti caratteristiche introdotte nel Capitolo 13. Di seguito vengono riscritte le formule della media e della varianza adattandole alle distribuzioni di probabilità marginali.

La media e la varianza della distribuzione di probabilità marginale di X sono date da:

$$\mu_X=\sum_x x f_X(x), \quad \sigma_X^2=\sum_x (x-\mu_X)^2 f_X(x).$$

Analogamente, la media e la varianza della distribuzione di probabilità marginale di Y sono:

$$\mu_Y=\sum_y y f_Y(y), \quad \sigma_Y^2=\sum_y (y-\mu_Y)^2 f_Y(y).$$

ESEMPIO 15.4

Si calcolino la media e la varianza della distribuzione di probabilità marginale della X esaminata nell'Esempio 15.3.

Soluzione

Si ha:

$$\mu_X=0\cdot\frac{1}{6}+1\cdot\frac{5}{18}+\cdots+5\cdot\frac{1}{18}=1.94,$$

$$\sigma_X^2=(0-1.94)^2\frac{1}{6}+(1-1.94)^2\frac{5}{18}+\cdots+(5-1.94)^2\frac{1}{18}=2.05.$$

■

15.3.2 Distribuzioni di probabilità condizionate

Data una variabile casuale doppia discreta (X, Y), è possibile definire la *distribuzione di probabilità condizionata* di una variabile casuale, dato uno specifico valore dell'altra.

DEFINIZIONE 15.3

Considerata una variabile casuale doppia discreta (X, Y), la funzione di probabilità condizionata di Y, dato che X assume il valore x, per il quale $f_X(x) > 0$, è definita da

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)},$$

dove y varia nell'insieme dei possibili valori di Y. Analogamente, la funzione di probabilità condizionata di X, dato che Y assume il valore y, per il quale $f_Y(y) > 0$, è definita da

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)},$$

dove x varia nell'insieme dei possibili valori di X.

Si noti che questa definizione ricalca quella di $P(B|A)$, cioè quella della probabilità condizionata che l'evento B si realizzi, dato che l'evento A si è realizzato (si veda la 12.2). Qui, A è l'evento $X = x$ e B l'evento $Y = y$.

È immediato verificare che $f_{Y|X}(y|x)$ e $f_{X|Y}(x|y)$ sono entrambe non negative e che

$$\sum_y f_{Y|X}(y|x) = \sum_x f_{X|Y}(x|y) = 1.$$

ESEMPIO 15.5

Per comodità di ragionamento, si consideri un esempio molto schematico di distribuzione di probabilità congiunta in cui X è il numero di anni di scolarità (13 per i periti industriali e 18 per i possessori di laurea quinquennale in ingegneria) e Y è il punteggio a un test di selezione per l'assunzione in un'azienda industriale:

X	Y 70	85	100	$f_X(x)$
13	0.2	0.2	0.0	0.4
18	0.0	0.1	0.5	0.6
$f_Y(y)$	0.2	0.3	0.5	1.0

Si calcolino $f_{Y|X}(y|13)$ e $f_{Y|X}(y|18)$ per $y = 70, 85, 100$.

Soluzione

Si ha:

$$f_{Y|X}(70|13) = \frac{f(13,70)}{f_X(13)} = \frac{0.2}{0.4} = 0.5, \quad f_{Y|X}(85|13) = \frac{f(13,85)}{f_X(13)} = \frac{0.2}{0.4} = 0.5,$$

$$f_{Y|X}(100|13) = \frac{f(13,100)}{f_X(13)} = \frac{0.0}{0.4} = 0.0,$$

$$f_{Y|X}(70|18) = \frac{f(18,70)}{f_X(18)} = \frac{0.0}{0.6} = 0.0, \quad f_{Y|X}(85|18) = \frac{f(18,85)}{f_X(18)} = \frac{0.1}{0.6} = 0.17,$$

$$f_{Y|X}(100|18) = \frac{f(18,100)}{f_X(18)} = \frac{0.5}{0.6} = 0.83.$$

È conveniente sistemare i dati ottenuti in una tabella a doppia entrata:

$f_{Y|X}(y|x)$	Y 70	85	100	Totale
$f_{Y|X}(y|13)$	0.50	0.50	0.00	1.00
$f_{Y|X}(y|18)$	0.00	0.17	0.83	1.00

Osservando la tabella, si può constatare che le due distribuzioni condizionate di Y sono sensibilmente diverse. Dunque, il livello di scolarizzazione incide sulla distribuzione di probabilità del punteggio al test e quindi è utile per predire i risultati del test.

■

15.3.3 Indipendenza tra due variabili casuali

La nozione di indipendenza tra variabili casuali riveste un ruolo cruciale nell'applicazione della teoria della probabilità all'inferenza statistica.

DEFINIZIONE 15.4

Due variabili casuali discrete X e Y si dicono indipendenti se, per ogni coppia (x, y), la funzione di probabilità congiunta $f(x, y)$ è uguale al prodotto delle funzioni di probabilità marginali $f_X(x)$ e $f_Y(y)$:

$$f(x,y) = f_X(x) \cdot f_Y(y). \tag{15.2}$$

L'indipendenza tra X e Y può essere espressa equivalentemente mediante le funzioni di probabilità condizionate, come indicato nella proposizione che segue.

PROPOSIZIONE 15.1

Date le variabili casuali discrete X e Y, per ogni coppia (x, y), tale che $f_X(x) > 0$ e $f_Y(y) > 0$, le seguenti tre condizioni sono equivalenti:

(a) $f_{Y|X}(y|x) = f_Y(y)$;

(b) $f(x,y) = f_X(x) \cdot f_Y(y)$;

(c) $f_{X|Y}(x|y) = f_X(x)$.

La dimostrazione è immediata. Assumendo che sia vera la condizione (a), si può scrivere

$$f(x,y) = f_{Y|X}(y|x) \cdot f_X(x) = f_Y(y) \cdot f_X(x),$$

da cui si deduce che anche la condizione (b) è vera. Viceversa, se la condizione (b) è vera, allora

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{f_X(x) \cdot f_Y(y)}{f_X(x)} = f_Y(y),$$

e quindi anche la (a) è vera. In modo analogo, si può dimostrare che la condizione (c) implica la (b) e viceversa.

■

La Proposizione 15.1 consente di capire più a fondo il concetto di indipendenza tra due variabili casuali. Sulla base della relazione (a), si può dire che le variabili casuali X e Y sono indipendenti se per ogni y la probabilità condizionata $f_{Y|X}(y|x)$ non dipende da x, essendo uguale alla probabilità marginale $f_Y(y)$. Ciò equivale a dire che la conoscenza del valore di X non è di nessun aiuto nel predire il valore di Y. Allo stesso modo, sulla base della relazione (c), si può dire che la conoscenza del valore di Y non è di nessun aiuto nel predire il valore di X.

ESEMPIO 15.6

Allo scopo di illustrare la Proposizione 15.1, si assuma che le variabili casuali X e Y dell'Esempio 15.5 siano indipendenti e che la loro funzione di probabilità congiunta sia la seguente:

X	Y			$f_X(x)$
	70	85	100	
13	0.08	0.12	0.20	0.40
18	0.12	0.18	0.30	0.60
$f_Y(y)$	0.20	0.30	0.50	1.00

Si noti che le probabilità congiunte all'interno della tabella soddisfano la relazione (15.2). Ora, se da queste si derivano le probabilità condizionate $f_{Y|X}(y|13)$ e $f_{Y|X}(y|18)$, per $y = 70, 85, 100$, si ottiene la tabella che segue:

$f_{Y\|X}(y\|x)$	Y			Totale
	70	85	100	
$f_{Y\|X}(y\|13)$	0.20	0.30	0.50	1.00
$f_{Y\|X}(y\|18)$	0.20	0.30	0.50	1.00

Analogamente, se si calcolano le probabilità condizionate $f_{X|Y}(x|70)$, $f_{X|Y}(x|85)$ e $f_{X|Y}(x|100)$, per $x = 13, 18$, si ottiene:

X	$f_{X\|Y}(x\|y)$		
	$f_{X\|Y}(x\|70)$	$f_{X\|Y}(x\|85)$	$f_{X\|Y}(x\|100)$
13	0.40	0.40	0.40
18	0.60	0.60	0.60
Totale	1.00	1.00	1.00

Questo esempio mostra che, quando X e Y sono indipendenti, le funzioni di probabilità condizionate $f_{Y|X}(y|x)$ e $f_{X|Y}(x|y)$ sono uguali alle funzioni di probabilità marginali $f_Y(y)$ e $f_X(x)$. ■

ESEMPIO 15.7

Si mostri che le variabili casuali della Tabella 15.2 non sono indipendenti.

Soluzione

È del tutto evidente che le variabili casuali non sono indipendenti: basta controllare che almeno una delle probabilità all'interno della Tabella 15.2 non è uguale

al prodotto delle corrispondenti probabilità marginali. Per esempio, per la prima casella (in alto a sinistra) si ha

$$f(0,0) = 0.125 \neq f_X(0) \cdot f_Y(0) = 0.125 \cdot 0.250.$$

Ne segue che X e Y non sono indipendenti. ■

15.3.4 Covarianza e correlazione

Viene ora data la definizione di *covarianza*, che è una misura della relazione associativa tra le variabili casuali X e Y.

DEFINIZIONE 15.5

Data una variabile casuale doppia discreta (X, Y) con funzione di probabilità congiunta $f(x, y)$, si chiama covarianza la quantità[1]

$$\sigma_{XY} = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)] = \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y),$$

dove μ_X e μ_Y sono le medie di X e di Y, e la somma è estesa a tutti i possibili valori di X e di Y.

La covarianza è una misura del legame associativo (lineare) tra le variabili casuali X e Y: un valore positivo indica che esse tendono a variare nella stessa direzione, mentre un valore negativo indica che esse tendono a variare in direzioni opposte. Un valore pari a zero indica assenza di legame lineare, ma non esclude la presenza di altri tipi di legame associativo.

ESEMPIO 15.8

Si calcoli la covarianza della distribuzione di probabilità congiunta introdotta nell'Esempio 15.1.

Soluzione

Innanzitutto, le medie di X e di Y sono date da:

$$\mu_X = 0 \cdot 0.125 + 1 \cdot 0.375 + 2 \cdot 0.375 + 3 \cdot 0.125 = 1.5,$$
$$\mu_Y = 0 \cdot 0.250 + 1 \cdot 0.500 + 2 \cdot 0.250 = 1.$$

La covarianza è allora

$$\sigma_{XY} = (0 - 1.5)(0 - 1) \cdot 0.125 + (1 - 1.5)(1 - 1) \cdot 0.25 + \cdots + (3 - 1.5)(0 - 1) \cdot 0.125 = 0.$$

■

Sebbene la covarianza fornisca un'indicazione della direzione del legame associativo (lineare) tra X e Y, è difficile capire se un dato valore di σ_{XY} corrisponda a un legame debole o forte. Infatti, σ_{XY} è una quantità che dipende fortemente

[1] Si tratta del valore atteso di una funzione di due variabili casuali, la cui definizione è fornita nel Paragrafo B.9 dell'Appendice B.

dall'unità di misura usata per X e Y. Una misura più appropriata della forza del legame lineare tra due variabili casuali è fornita dal *coefficiente di correlazione lineare*.

DEFINIZIONE 15.6

Siano X e Y due variabili casuali discrete aventi, rispettivamente, medie μ_X e μ_Y e varianze σ_X^2 e σ_Y^2. Il coefficiente di correlazione lineare di X e Y è dato da

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \tag{15.3}$$

Si può dimostrare che ρ assume valori nell'intervallo $[-1, 1]$. L'indice ha lo stesso segno di σ_{XY}, in quanto il denominatore della (15.3) è positivo. Assume valori positivi quando le due variabili tendono a crescere o a decrescere insieme (correlazione positiva). Assume valori negativi quando le variabili tendono a variare in direzioni opposte (correlazione negativa). I valori estremi, -1 e 1, si hanno nel caso di una relazione lineare perfetta tra X e Y, cioè quando $Y = a + bX$, dove $b \neq 0$, con probabilità 1. Un valore nullo indica che tra X e Y non intercorre una relazione lineare. In questo caso, le due variabili si dicono incorrelate.

PROPOSIZIONE 15.2

Se le variabili casuali X e Y sono indipendenti, allora σ_{XY} e ρ sono entrambi nulli.

La dimostrazione è lasciata per esercizio (si veda l'Esercizio 15.10). ■

Il viceversa della Proposizione 15.2 non è vero, perché due variabili casuali possono essere incorrelate, ma non indipendenti (si veda l'Esercizio 15.11).

15.4 Distribuzione di probabilità congiunta di due variabili casuali continue

Una discussione approfondita di questo argomento richiederebbe strumenti matematici che esulano dal bagaglio di conoscenze richiesto per la lettura di questo testo. Qui vengono fatte delle estensioni intuitive al caso continuo delle principali nozioni introdotte nei paragrafi precedenti.

La distribuzione di probabilità di una variabile casuale doppia continua è descritta da una funzione non negativa $f(x, y)$, cioè da una superficie tridimensionale definita nel piano xy, tale che il volume ad essa sotteso sia pari a 1 (analogamente al caso di una sola variabile, in cui è 1 l'area sottesa alla curva di densità). Questa superficie si chiama funzione di densità congiunta.

ESEMPIO 15.9

Si consideri la funzione di densità congiunta

$$f(x, y) = \begin{cases} 4xy, & 0 \leq x \leq 1,\ 0 \leq y \leq 1, \\ 0, & \text{altrove.} \end{cases}$$

La sua rappresentazione grafica è mostrata nella Figura 15.3.

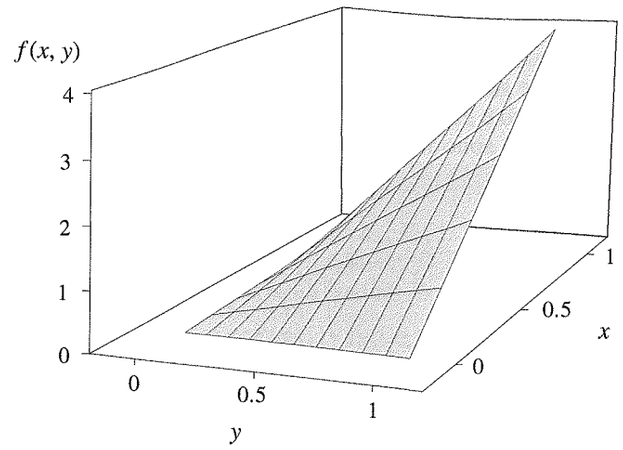


Figura 15.3
Grafico della funzione di densità congiunta della variabile casuale doppia continua in esame.

Dalla funzione di densità congiunta è possibile ottenere le funzioni di densità marginali[2] $f_X(x)$ e $f_Y(y)$, il cui significato e le cui proprietà ricalcano quanto illustrato per il caso di una sola variabile casuale (si veda il Paragrafo 13.3).

A sua volta, la conoscenza delle funzioni di densità marginali consente di calcolare le medie e le varianze di X e di Y (si veda il Paragrafo 13.3.1 per la media e la varianza di una variabile casuale continua).

Analogamente a quanto visto per le variabili casuali discrete, due variabili casuali continue si dicono indipendenti se per ogni coppia (x, y) la funzione di densità congiunta è uguale al prodotto delle funzioni di densità marginali:

$$f(x, y) = f_X(x) \cdot f_Y(y).$$

La definizione di covarianza è formalmente la stessa del caso discreto:[3]

$$\sigma_{XY} = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)].$$

È formalmente la stessa del caso discreto anche la definizione di coefficiente di correlazione lineare (si veda il Paragrafo 15.3.4).

ESEMPIO 15.10

Si consideri la funzione di densità congiunta

$$f(x, y) = \frac{1}{2\pi\sqrt{\sigma_X^2 \sigma_Y^2 (1 - \rho^2)}} \exp\left\{ -\frac{1}{2(1 - \rho^2)} \left[\left(\frac{x - \mu_X}{\sigma_X} \right)^2 + \left(\frac{y - \mu_Y}{\sigma_Y} \right)^2 - 2\rho \left(\frac{x - \mu_X}{\sigma_X} \right) \left(\frac{y - \mu_Y}{\sigma_Y} \right) \right] \right\}, \quad -\infty < x, y < \infty,$$

[2] Se X e Y hanno funzione di densità congiunta $f(x, y)$, le funzioni di densità marginali di X e di Y sono date da:

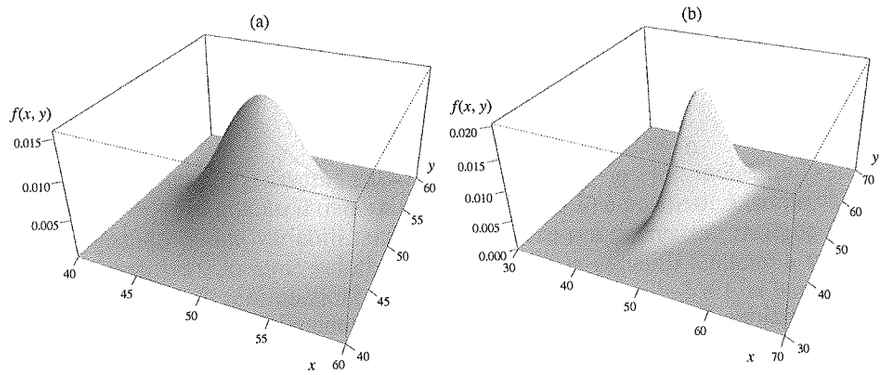
$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, \mathrm{d}y, \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, \mathrm{d}x.$$

[3] Se X e Y hanno funzione di densità congiunta $f(x, y)$, la covarianza σ_{XY} è espressa da

$$\sigma_{XY} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y)\, \mathrm{d}x\, \mathrm{d}y.$$

dove exp{·} sta per e elevato all'espressione tra parentesi. Si tratta della funzione di densità della distribuzione normale doppia. Essa dipende da cinque parametri: la media e la deviazione standard di X e di Y e il coefficiente di correlazione lineare tra X e Y. Il suo grafico è una superficie a campana.

Nella Figura 15.4, sono rappresentate due distribuzioni normali doppie con parametri diversi.



ura 15.4
nzioni di densità di due tribuzioni normali doppie n parametri diversi. In rticolare, ρ è nullo nel afico (a) e pari a 0.75 l grafico (b).

Le distribuzioni di probabilità marginali, che si ottengono dalla funzione di densità congiunta della distribuzione normale doppia, assumono la forma:

$$f_X(x) = \frac{1}{\sigma_X\sqrt{2\pi}} e^{-(x-\mu_X)^2/2\sigma_X^2},$$

$$f_Y(y) = \frac{1}{\sigma_Y\sqrt{2\pi}} e^{-(y-\mu_Y)^2/2\sigma_Y^2}.$$

Si riconosce che la prima è la funzione di densità di una distribuzione normale con media μ_X e varianza σ_X^2, mentre la seconda è la funzione di densità di una distribuzione normale con media μ_Y e varianza σ_Y^2. Ne segue che, quando $\rho = 0$, la funzione di densità della distribuzione normale doppia si riduce al prodotto delle suddette funzioni di densità marginali. Ciò significa che per la distribuzione normale doppia $\rho = 0$ implica l'indipendenza tra X e Y. Questa è un'eccezione alla regola generale per cui l'incorrelazione non implica l'indipendenza.

■

15.5 Distribuzione di probabilità congiunta di *n* variabili casuali

La nozione di distribuzione di probabilità congiunta di due variabili casuali può essere generalizzata al caso di n variabili casuali, che traggono origine dall'associazione di n numeri reali a ogni evento elementare dell'esperimento casuale.

La distribuzione di probabilità congiunta di n variabili casuali $X_1, X_2, \ldots, X_n$ può essere descritta da una funzione, $f(x_1, x_2, \ldots, x_n)$, di probabilità congiunta o di densità congiunta a seconda che le variabili casuali siano discrete o continue.

Nei prossimi capitoli, in particolare nel Capitolo 17, si avrà a che fare con i campioni casuali, definiti come sequenze di variabili casuali indipendenti. È quindi opportuno introdurre fin d'ora la nozione di distribuzione di probabilità congiunta di n variabili casuali indipendenti. Questo aspetto è affrontato nella Definizione 15.7. Oltre a ciò, è opportuno studiare la distribuzione di probabilità della media aritmetica, $(X_1 + X_2 + \cdots + X_n)/n$, e di altre semplici funzioni di n variabili casuali.

DEFINIZIONE 15.7

Siano date n variabili casuali indipendenti $X_1, X_2, \ldots, X_n$ con funzioni di probabilità (densità) marginali $f_{X_1}(x_1), f_{X_2}(x_2), \ldots, f_{X_n}(x_n)$. La funzione di probabilità (densità) congiunta delle variabili casuali $X_1, X_2, \ldots, X_n$ è data da

$$f(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdot \cdots \cdot f_{X_n}(x_n).$$

15.5.1 Combinazioni lineari di *n* variabili casuali

In molte applicazioni, le tecniche di inferenza coinvolgono funzioni lineari di due o più variabili casuali. Esempi di queste funzioni sono la differenza di due variabili casuali, $X_1 - X_2$, la media aritmetica di n variabili casuali, $\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n$, e così via. In generale, si considerano *combinazioni lineari di n variabili casuali* definite da

$$W = a_1X_1 + a_2X_2 + \cdots + a_nX_n, \tag{15.4}$$

dove $a_1, a_2, \ldots, a_n$ sono costanti assegnate.

PROPOSIZIONE 15.3

Siano date n variabili casuali $X_1, X_2, \ldots, X_n$, discrete o continue. La media e la varianza della combinazione lineare

$$W = a_1X_1 + a_2X_2 + \cdots + a_nX_n,$$

dove $a_1, a_2, \ldots, a_n$ sono costanti assegnate, sono date da:

$$\mu_W = a_1\mu_{X_1} + a_2\mu_{X_2} + \cdots + a_n\mu_{X_n}, \tag{15.5}$$

$$\sigma_W^2 = a_1^2\sigma_{X_1}^2 + a_2^2\sigma_{X_2}^2 + \cdots + a_n^2\sigma_{X_n}^2 + \sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n} a_ia_j\sigma_{X_iX_j}, \tag{15.6}$$

dove μ_{X_i} e $\sigma_{X_i}^2$ sono la media e la varianza della i-esima variabile casuale X_i $(i = 1, 2, \ldots, n)$, mentre $\sigma_{X_iX_j}$ è la covarianza tra X_i e X_j.

La dimostrazione è riportata nel Paragrafo B.9 dell'Appendice B.

■

Se le variabili casuali $X_1, X_2, \ldots, X_n$ sono indipendenti, la varianza di W si semplifica e diviene

$$\sigma_W^2 = a_1^2\sigma_{X_1}^2 + a_2^2\sigma_{X_2}^2 + \cdots + a_n^2\sigma_{X_n}^2. \tag{15.7}$$

Ciò perché l'assunzione di indipendenza delle $X_1, X_2, \ldots, X_n$ implica l'indipendenza (a due a due) tra X_i e X_j, per ogni coppia (X_i, X_j), con $i \neq j$, il che, a sua volta, implica $\sigma_{X_iX_j} = 0$.

Per $n = 2$, le (15.5) e (15.6) diventano:

$$\mu_W = a_1\mu_{X_1} + a_2\mu_{X_2}, \tag{15.8}$$

$$\sigma_W^2 = a_1^2\sigma_{X_1}^2 + a_2^2\sigma_{X_2}^2 + 2a_1a_2\sigma_{X_1X_2}. \tag{15.9}$$

ESEMPIO 15.11

Si determinino la media e la varianza della variabile casuale $W = 2X_1 - X_2$, sapendo che $\mu_{X_1} = 10$, $\sigma_{X_1}^2 = 21$, $\mu_{X_2} = 16$, $\sigma_{X_2}^2 = 9$ e $\sigma_{X_1X_2} = -4.5$.

Soluzione

Applicando le (15.8) e (15.9), si ottiene:

$$\mu_W = 2 \cdot 10 - 16 = 6,$$

$$\sigma_W^2 = 4 \cdot 21 + 9 + 2 \cdot 2(-1)(-4.5) = 111.$$

■

PROPOSIZIONE 15.4

Siano date n variabili casuali indipendenti $X_1, X_2, \ldots, X_n$, ognuna con distribuzione normale di media μ_{X_i} e varianza $\sigma_{X_i}^2$. Allora la combinazione lineare

$$W = a_1X_1 + a_2X_2 + \cdots + a_nX_n,$$

dove $a_1, a_2, \ldots, a_n$ sono costanti assegnate, ha distribuzione normale di media

$$\mu_W = a_1\mu_{X_1} + a_2\mu_{X_2} + \cdots + a_n\mu_{X_n}$$

e varianza

$$\sigma_W^2 = a_1^2\sigma_{X_1}^2 + a_2^2\sigma_{X_2}^2 + \cdots + a_n^2\sigma_{X_n}^2.$$

Per la dimostrazione si rinvia a Hogg *et al.* (2013), p. 174. ■

Se nella (15.4) si pone $a_1 = a_2 = \cdots = a_n = 1/n$, la combinazione lineare W diventa la media aritmetica delle variabili casuali $X_1, X_2, \ldots, X_n$, che verrà indicata con $\bar{X}$. Questa variabile casuale trova ampio impiego nei problemi di inferenza sulla media della popolazione, quindi merita un'attenzione particolare.

PROPOSIZIONE 15.5

Siano date n variabili casuali $X_1, X_2, \ldots, X_n$, indipendenti e tutte con la stessa media μ e la stessa varianza σ^2. Allora la media aritmetica di queste variabili casuali,

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

ha valore atteso e varianza date da:

$$\mu_{\bar{X}} = \mu, \quad \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}. \tag{15.10}$$

Il primo risultato si deduce immediatamente dalla (15.5) sostituendo a_i con $1/n$ e μ_{X_i} con μ. Il secondo si ottiene, con dei semplici passaggi algebrici, dalla (15.7) sostituendo a_i con $1/n$ e $\sigma_{X_i}^2$ con σ^2. ■

Riassunto

Oggetto di questo capitolo sono le distribuzioni di probabilità che descrivono il comportamento congiunto di due o più variabili casuali. L'attenzione è rivolta in particolare alla distribuzione di probabilità congiunta di due variabili casuali discrete, per le quali vengono anche definiti: le distribuzioni di probabilità marginali; le distribuzioni di probabilità condizionate; il concetto di indipendenza; due misure di associazione, la covarianza e il coefficiente di correlazione lineare. Queste nozioni sono poi estese, con un approccio intuitivo, al caso di due variabili casuali continue. Successivamente, viene definita la distribuzione di probabilità congiunta di n variabili casuali, in generale e nel caso di indipendenza. Viene infine esaminato sinteticamente il problema della distribuzione di probabilità di una combinazione lineare di n variabili casuali.

Parole chiave

coefficiente di correlazione lineare, 334
combinazioni lineari di n variabili casuali, 337
covarianza, 333
distribuzione di probabilità marginale, 328
distribuzione normale doppia, 336
funzione di densità congiunta, 334
funzione di densità marginale, 335
funzione di probabilità condizionata, 330
funzione di probabilità congiunta, 327
funzione di probabilità marginale, 328
variabile casuale doppia, 326
variabili casuali incorrelate, 334
variabili casuali indipendenti, 331

Commenti e curiosità

Combinazioni lineari di variabili casuali e investimenti in borsa

Le formule (15.8) e (15.9) si prestano a essere applicate nelle valutazioni finanziarie sui portafogli titoli. Si supponga, per semplicità, che il portafoglio di un investitore in titoli azionari sia composto di N_1 azioni di tipo A e di N_2 azioni di tipo B. Le quotazioni azionarie dei due titoli possono essere rappresentate dalle variabili casuali X_1 e X_2. Il valore complessivo del portafoglio è pertanto dato da

$$W = N_1X_1 + N_2X_2,$$

combinazione lineare delle variabili casuali X_1 e X_2. Se l'investitore vuole valutare la volatilità del suo portafoglio, deve ricorrere al calcolo della varianza di W tramite la (15.9), ottenendo

$$\sigma_W^2 = N_1^2\sigma_{X_1}^2 + N_2^2\sigma_{X_2}^2 + 2(N_1 \cdot N_2)\sigma_{X_1X_2},$$

dove $\sigma_{X_1}^2$ e $\sigma_{X_2}^2$ sono le varianze delle quotazioni dei due titoli, che ne misurano le rispettive volatilità, e $\sigma_{X_1X_2}$ è la covarianza tra le quotazioni degli stessi titoli. Ammesso che le varianze e la covarianza siano note (spesso si utilizzano a questo fine i dati storici), l'investitore, che potesse scegliere tra coppie alternative di titoli aventi le medesime volatilità ($\sigma_{X_1}^2$ e $\sigma_{X_2}^2$), dovrebbe suddividere la somma disponibile tra i due titoli (ossia scegliere N_1 e N_2), in primo luogo, in base alla loro redditività attesa e, eventualmente, in base al segno della correlazione tra le rispettive quotazioni: se volesse coprirsi contro il rischio di una eccessiva volatilità, potrebbe scegliere titoli con quotazioni correlate negativamente, perché in tal caso il terzo addendo di σ_W^2 sarebbe di segno negativo (essendo $\sigma_{X_1X_2}$ negativo) e quindi la volatilità del portafoglio sarebbe inferiore rispetto al caso di correlazione positiva o nulla.

Il discorso può essere esteso al caso di più di due titoli, utilizzando la (15.5) e la (15.6) per calcolare la redditività media e la volatilità di una determinata combinazione di titoli.

sercizi

15.1 Una moneta bilanciata viene lanciata quattro volte. Si associno a ogni evento elementare i numeri x e y, tali che x è la frequenza dell'evento testa e y è la differenza tra la frequenza dell'evento testa e quella dell'evento croce, che assume, convenzionalmente, i valori -1, 0 e 1 a seconda che la differenza sia negativa, nulla o positiva.

a) Si trovi la funzione di probabilità congiunta della variabile casuale doppia sopra definita.

b) Si determinino le due funzioni di probabilità marginali.

15.2 Un'azienda di trasporto urbano riceve occasionalmente reclami dalla clientela per i ritardi e per il comfort nel viaggio. Sulla base della passata esperienza, la distribuzione di probabilità congiunta del numero di reclami per i ritardi e per il comfort, in una settimana, è la seguente:

N. di reclami per i ritardi (X)	N. di reclami per il comfort (Y) 0	1	2
1	0.42	0.07	0.09
2	0.15	0.12	0.07
3	0.04	0.03	0.01

a) Qual è la funzione di probabilità marginale di X?

b) Si determini la funzione di probabilità condizionata di Y per $X = 2$.

c) Le due variabili casuali sono indipendenti?

d) Si determini la covarianza.

15.3 Sia data la distribuzione doppia di probabilità:

X	Y 1	2	4
2	0.1	0.0	0.1
3	0.0	0.2	0.0
5	0.2	0.1	0.1
6	0.0	0.0	0.2

a) Si stabilisca se le variabili casuali X e Y sono indipendenti.

b) Si determinino la media e la varianza della variabile casuale $W = X + Y$.

15.4 Si consideri la distribuzione di probabilità congiunta:

X	Y 1	2	3
1	1/36	2/36	3/36
2	2/36	4/36	6/36
3	3/36	6/36	9/36

a) Si determinino le distribuzioni marginali e si stabilisca se le variabili casuali X e Y sono indipendenti.

b) Si determinino le tre distribuzioni condizionate di Y.

c) Si verifichi che le distribuzioni condizionate di X sono uguali alla distribuzione marginale della stessa variabile casuale.

d) Si calcoli la covarianza.

15.5 Relativamente a un gruppo di utenti di due siti internet, le tabelle che seguono riportano le distribuzioni condizionate del numero di pagine visitate nel secondo sito (Y), dato il numero di pagine visitate nel primo sito (X):

Valore di Y	0	1	2
$f_{Y\|X}(y\|0)$	1/8	3/8	4/8

Valore di Y	0	1	2
$f_{Y\|X}(y\|1)$	2/9	2/9	5/9

Invece, la distribuzione condizionata del numero di pagine visitate nel primo sito, dato che il numero di pagine visitate nel secondo sito è pari a 0, è data da

Valore di X	0	1	2
$f_{X\|Y}(x\|0)$	1/6	2/6	3/6

Inoltre, $P(Y = 1) = 7/27$ e $P(Y = 2) = 14/27$.

a) Si trovino la funzione di probabilità congiunta e le funzioni di probabilità marginali di X e di Y.

b) Si accerti se le variabili casuali X e Y sono indipendenti.

15.6 Si considerino le tre seguenti distribuzioni condizionate di Y, dato che $X = x$, per $x = 0, 1, 2$:

$f_{Y\|X}(y\|x)$	Valore di Y 1	2	4
$f_{Y\|X}(y\|0)$	1/5	2/5	2/5
$f_{Y\|X}(y\|1)$	1/4	1/4	2/4
$f_{Y\|X}(y\|2)$	2/6	1/6	3/6

e si supponga che $P(X = 0) = 1/7$, $P(X = 1) = 4/7$ e $P(X = 2) = 2/7$.

a) Si calcoli il valore atteso di X.

b) Si ricostruisca la distribuzione di probabilità congiunta di X e Y.

c) Si calcoli il valore atteso di Y.

d) Si calcoli il coefficiente di correlazione lineare.

15.7 Siano X e Y due variabili casuali discrete con distribuzione di probabilità congiunta:

X	Y 1	2	4
1	0.1	0.2	0.3
2	0.1	0.2	0.1

a) Si stabilisca se le due variabili sono indipendenti.

b) Si calcoli la probabilità $P(Y > 1 \mid X = 2)$.

c) Si calcoli la covarianza.

d) Si determini il valore atteso della variabile casuale $Z = X + Y$.

e) Si calcoli il momento secondo, $E(W^2)$, della variabile casuale $W = XY$.

15.8 Sia X una variabile casuale con distribuzione Binomiale di parametri n e p_1 e sia Y una variabile casuale con distribuzione Bernoulliana di parametro p_2. Assumendo che X e Y siano indipendenti, si determini $E(XY)$.

15.9 Un esperto di borsa formula le seguenti valutazioni probabilistiche sulle quotazioni di due titoli azionari al termine del prossimo mese finanziario:

Quotazione del titolo A (euro)	Quotazione del titolo B (euro) 14	16	18.3
12.5	0.04	0.13	0.13
13.4	0.12	0.07	0.05
15.7	0.42	0.03	0.01

Si calcolino il valore atteso e la varianza del valore di un portafoglio formato da 150 azioni di tipo A e 235 di tipo B. Vi è correlazione positiva o negativa tra le quotazioni dei due titoli?

5.10* Si dimostri che la covarianza di due variabili casuali discrete indipendenti è nulla. (*Suggerimento*: si scriva l'espressione della covarianza sostituendo a $f(x,y)$ il prodotto $f_X(x)f_Y(y)$; dopodiché si potrà scrivere

$$\sigma_{XY} = \sum_x (x-\mu_X) f_X(x) \sum_y (y-\mu_Y) f_Y(y),$$

da cui segue il risultato perché le due sommatorie si annullano, come è facile verificare.)

15.11 Si consideri la funzione di probabilità congiunta

$$f(x,y) = \frac{1}{3}, \quad \text{per } (x,y) = (0,-1),(1,0),(-1,0).$$

Si verifichi che $\sigma_{XY} = 0$, sebbene tra le due variabili casuali valga la relazione $Y = X^2 - 1$.

15.12 Siano X e Y due variabili casuali discrete con funzione di probabilità congiunta

$$f(x,y) = k\frac{3^{x+y}}{x!y!}, \quad x = 0,1,2,\ldots,\ y = 0,1,2,\ldots$$

a) Si determini la costante k.
b) Si verifichi se le variabili casuali X e Y sono indipendenti.
c) Si determinino le funzioni di probabilità marginali.

5.13* Data la funzione di densità congiunta $f(x,y)$ definita da

$$f(x,y) = \begin{cases} x+y, & \text{se } 0 \le x \le 1,\ 0 \le y \le 1, \\ 0, & \text{altrove,} \end{cases}$$

si determinino le funzioni di densità marginali. (*Suggerimento*: si veda la Nota 2.)

5.14* Sia (X, Y) una variabile casuale doppia con funzione di densità

$$f(x,y) = \begin{cases} xye^{-(x+y)}, & \text{se } x \ge 0,\ y \ge 0, \\ 0, & \text{altrove.} \end{cases}$$

a) Si determini la funzione di densità marginale di X.
b) Si calcoli la probabilità $P(X \ge 1, Y \le 1)$.

(*Suggerimento*: per il primo punto, si veda la Nota 2; per il secondo, si consideri che è possibile scrivere $P(X \ge 1, Y \le 1) = \int_1^\infty xe^{-x}\,dx \int_0^1 ye^{-y}\,dy$.)

15.15 Siano X_1 e X_2 due variabili casuali tali che $E(X_1) = 1$, $E(X_2) = 2$, $Var(X_1) = 1$, $Var(X_2) = 3$. Si consideri la combinazione lineare $W = 2X_1 + X_2$.

a) Si trovino la media e la varianza di W, assumendo che $\sigma_{X_1X_2} = 0.5$.
b) Si trovino il valore atteso e la varianza di W, assumendo che X_1 e X_2 siano indipendenti.

15.16 Un portafoglio azionario è costituito da 60 azioni di tre diversi titoli: 10 azioni del primo titolo, 20 azioni del secondo titolo e 30 azioni del terzo titolo. Si supponga che la quotazione giornaliera dei tre titoli (in euro) possa essere descritta dalle variabili casuali X_1, X_2 e X_3 e che: $E(X_1) = 3$, $E(X_2) = 2.5$, $E(X_3) = 4$, $Var(X_1) = 0.5$, $Var(X_2) = 0.8$, $Var(X_3) = 1$.

a) Supponendo che i tre titoli non si comportino in modo indipendente e che $\sigma_{X_1X_2} = 0.6$, $\sigma_{X_1X_3} = 0.6$ e $\sigma_{X_2X_3} = 0.85$, si calcolino media e varianza del valore giornaliero del portafoglio.
b) Si determinino la media e la varianza del valore giornaliero del portafoglio nell'ipotesi che i tre titoli si comportino in modo indipendente.

15.17 Sia X una variabile casuale avente distribuzione normale con valore atteso $\mu = 2$ e varianza $\sigma^2 = 5$, e sia Y un'altra variabile casuale avente distribuzione normale con valore atteso $\mu = 3$ e varianza $\sigma^2 = 1$. Assumendo che X e Y siano indipendenti, si calcoli la probabilità $P(W \le 9.5)$, dove $W = X + 2Y$.

CAPITOLO

16

Legge dei grandi numeri e teorema del limite centrale

Sommario del capitolo

16.1 Introduzione
16.2 Legge dei grandi numeri
16.3 Teorema del limite centrale
Riassunto
Commenti e curiosità
Esercizi

16.1 Introduzione

Si consideri un esperimento casuale. Sia X una variabile casuale, i cui valori dipendono dai risultati di questo esperimento, distribuita secondo la funzione di probabilità o di densità $f_X(x)$ con media μ e varianza σ^2. A n repliche indipendenti dell'esperimento, effettuate nelle stesse condizioni, corrisponderà una sequenza di n variabili casuali $X_1, X_2, \ldots, X_n$, indipendenti e tutte con la stessa distribuzione di probabilità $f_X(x)$.

Data dunque una sequenza di n variabili casuali ottenute come descritto sopra, se ne prenda la media aritmetica, $\bar{X}_n = (X_1+X_2+\cdots+X_n)/n$. In base alle (15.10), il valore atteso e la varianza di $\bar{X}_n$ sono μ e σ^2/n, essendo μ e σ^2 il valore atteso e la varianza di X. Oggetto di questo capitolo è lo studio del comportamento della distribuzione di probabilità di $\bar{X}_n$ quando n aumenta indefinitamente.

Si assuma, ad esempio, che X abbia distribuzione di Poisson con valore atteso 1.5. In questo caso, la media aritmetica $\bar{X}_n$ delle n variabili casuali, indipendenti e aventi distribuzione di Poisson, ottenute ripetendo n volte l'esperimento, ha valore atteso 1.5 e varianza $1.5/n$. La distribuzione di $\bar{X}_n$ è illustrata nella Figura 16.1, per $n = 2, 5, 10, 20$ (per brevità, si tralasciano i dettagli con cui si arriva al risultato).

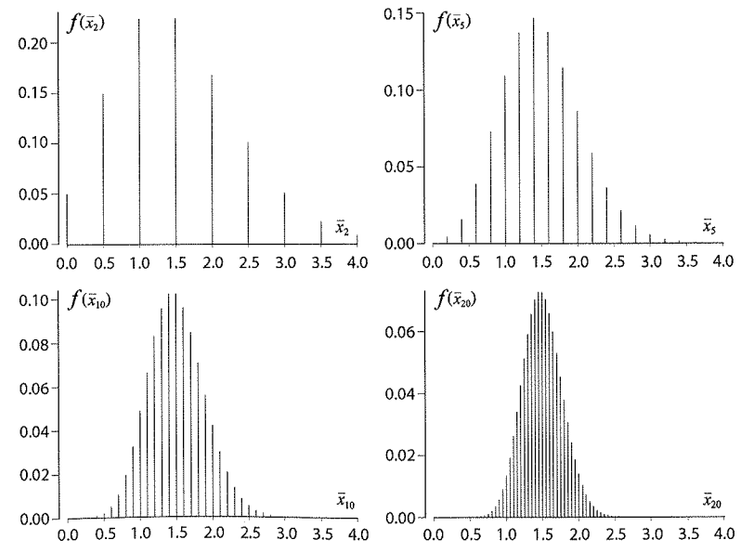


Figure 16.1 Distribuzione di probabilità della media aritmetica $\bar{X}_n$ di n variabili casuali di Poisson con parametro $\lambda = 1.5$, per $n = 2, 5, 10, 20$.

Al riguardo, la domanda è se, all'aumentare di n, la distribuzione di probabilità di $\bar{X}_n$ tenda a concentrarsi sempre più attorno al suo valore atteso, pari a 1.5. Dall'osservazione della Figura 16.1, sembra potersi dare una risposta positiva.

Una seconda domanda è se, all'aumentare di n, la distribuzione di probabilità di $\bar{X}_n$ tenda ad avvicinarsi al modello della distribuzione normale. Anche in questo caso, la Figura 16.1 sembra suggerire una risposta positiva.

In questo capitolo, si studieranno questi due comportamenti per qualsiasi distribuzione $f_X(x)$ di X, illustrando la legge dei grandi numeri, relativamente al primo comportamento, e il teorema del limite centrale, relativamente al secondo.

16.2 Legge dei grandi numeri

La legge dei grandi numeri riguarda il comportamento, al crescere indefinito di n, della distribuzione della media aritmetica $\bar{X}_n$ di una sequenza di n variabili casuali $X_1, X_2, \ldots, X_n$, indipendenti e tutte con la stessa distribuzione, che rappresentano il risultato di n repliche, nelle stesse condizioni, di un esperimento casuale.

PROPOSIZIONE 16.1

Siano date n variabili casuali $X_1, X_2, \ldots, X_n$, indipendenti e tutte con la stessa distribuzione, con valore atteso $\mathrm{E}(X_i) = \mu$ e varianza $\mathrm{Var}(X_i) = \sigma^2$, $i = 1, 2, \ldots, n$. Sia

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

la media aritmetica di queste variabili casuali. Allora, per ogni $\varepsilon > 0$ piccolo a piacere, si ha

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \varepsilon) = 0,$$

oppure, equivalentemente,

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1. \tag{16.1}$$

La dimostrazione è immediata. In base alle (15.10),

$$\mathrm{E}(\bar{X}_n) = \mu \quad \text{e} \quad \mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

L'applicazione della disuguaglianza di Chebyshev (si veda la Proposizione 13.2), per ogni $\varepsilon > 0$, consente di scrivere

$$P\left(|\bar{X}_n - \mu| < \varepsilon\right) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}.$$

Al tendere di n a infinito, il termine di destra della disuguaglianza tende a 1, da cui segue che $\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1$. ∎

È utile far notare che la probabilità dell'evento $|\bar{X}_n - \mu| < \varepsilon$ può anche essere scritta nella forma

$$P(\mu - \varepsilon < \bar{X}_n < \mu + \varepsilon),$$

dalla quale si evince che, per n che tende a infinito, tende a 1 la probabilità che $\bar{X}_n$ assuma un valore nell'intervallo $(\mu - \varepsilon, \mu + \varepsilon)$, cioè un valore molto vicino a μ, essendo ε un numero positivo piccolo a piacere.

La legge dei grandi numeri implica che per n sufficientemente grande ci sia un'altissima probabilità che la variabile casuale $\bar{X}_n$ assuma un valore assai vicino a μ. Così, se un dado bilanciato fosse lanciato molte e molte volte, ad esempio 10 000 volte, sarebbe molto probabile ottenere come media dei numeri usciti un valore molto vicino a 3.5, valore atteso della variabile casuale associata all'esperimento.

Il risultato enunciato nella Proposizione 16.1 è spesso chiamato legge debole dei grandi numeri per distinguerla dalla cosiddetta legge forte dei grandi numeri.[1]

ESEMPIO 16.1

Si riprenda la Figura 16.1 e si delimiti, in ciascun grafico, con due linee verticali tratteggiate, come viene fatto nella Figura 16.2, i valori di $\bar{X}_n$ che appartengono all'intervallo $(1.5-\varepsilon, 1.5+\varepsilon)$, dove 1.5 è il valore atteso di $\bar{X}_n$ e ε, quantità piccola a piacere, è posta uguale a 0.3.

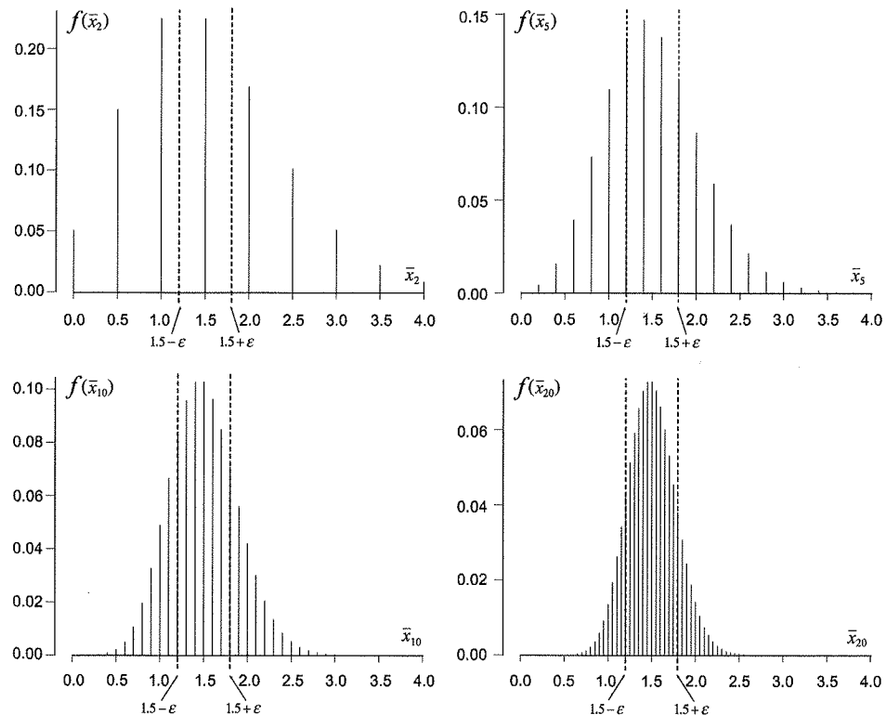


Figura 16.2
Rappresentazione delle stesse distribuzioni di probabilità della Figura 16.1 con enfasi sulla probabilità che $\bar{X}_n$ assuma valori nell'intervallo $(1.5 - \varepsilon, 1.5 + \varepsilon)$ dove $\varepsilon = 0.3$.

La somma delle probabilità corrispondenti ai valori di $\bar{X}_n$ compresi tra le due linee tratteggiate è: 0.22 per $n = 2$, 0.28 per $n = 5$, 0.48 per $n = 10$, 0.68 per $n = 20$. Dunque, la probabilità dell'evento $(1.5 - \varepsilon < \bar{X}_n < 1.5 + \varepsilon)$ aumenta con n.

[1] La legge forte dei grandi numeri afferma che

$$P\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1.$$

La legge debole afferma che, per n abbastanza grande, $\bar{X}_n$ è vicino a μ con un'altissima probabilità. Questo però non esclude, di per sé, che $|\bar{X}_n - \mu| > \varepsilon$, una o più volte, per $n \to \infty$. La legge forte afferma invece che, con probabilità uno, questo non succede: essa afferma che la disuguaglianza $|\bar{X}_n - \mu| < \varepsilon$ è verificata, con probabilità uno, per tutti i valori di n sufficientemente grandi.

Si deve notare che si arriverebbe a conclusioni simili anche prendendo in considerazione altri valori di n e ε. Ciò è dovuto al fatto che, all'aumentare di n, la distribuzione di $\bar{X}_n$ si concentra sempre di più attorno al suo valore atteso. ■

Si riprenda la Proposizione 16.1 assumendo che le variabili casuali $X_1, X_2, \ldots, X_n$ siano indipendenti e tutte distribuite secondo una distribuzione Bernoulliana di parametro p. In questo caso, ogni X_i assume valore 1 (successo) o valore 0 (insuccesso) con probabilità p e $1 - p$, sicché la somma $\sum_{i=1}^{n} X_i$ è la frequenza assoluta e $\bar{X}_n = \sum_{i=1}^{n} X_i/n$ la frequenza relativa dei successi che qui si indica con $\hat{p}_n$. La (16.1) diviene allora

$$\lim_{n\to\infty} P(|\hat{p}_n - p| < \varepsilon) = 1.$$

Questa formulazione è conosciuta come legge dei grandi numeri di Bernoulli. Dunque, considerato un evento di probabilità p associato a un dato esperimento casuale, all'aumentare oltre ogni limite delle repliche indipendenti dell'esperimento, tende a 1 la probabilità che la frequenza relativa con cui l'evento si verifica disti dalla probabilità p per meno di un numero positivo ε piccolo quanto si vuole.

Viene da chiedersi che rapporto ci sia tra la legge dei grandi numeri di Bernoulli e la legge empirica del caso di cui si è fatto cenno nel Capitolo 12 (Paragrafo 12.3). Va chiarito che la prima non è la dimostrazione della seconda. Invero, questa si basa sull'osservazione empirica che la frequenza relativa tende a stabilizzarsi attorno alla probabilità dell'evento; quella stabilisce che all'aumentare di n cresce la probabilità che ciò avvenga.

16.3 Teorema del limite centrale

Del teorema esistono numerose versioni che differiscono tra loro per le assunzioni che vengono fatte sulle variabili casuali coinvolte. Qui se ne dà la versione standard, adatta ai fini delle applicazioni che ne verranno fatte nei prossimi capitoli.

PROPOSIZIONE 16.2

Sia data una sequenza di variabili casuali $X_1, X_2, \ldots, X_n$, indipendenti e tutte con la stessa distribuzione di probabilità, con valore atteso $\mathrm{E}(X_i) = \mu$ e varianza $\mathrm{Var}(X_i) = \sigma^2$, $i = 1, 2, \ldots, n$. Sia $\bar{X}_n = \sum_{i=1}^{n} X_i/n$ la media aritmetica di queste variabili casuali avente, per le (15.10), valore atteso μ e varianza σ^2/n. Allora, posto

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}},$$

per ogni numero reale z, si ha

$$\lim_{n\to\infty} P(Z_n \le z) = \lim_{n\to\infty} P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le z\right) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du = \Phi(z), \qquad (16.2)$$

essendo $\Phi(z)$ la funzione di ripartizione della distribuzione normale standard.

Per la dimostrazione si rinvia a Hogg *et al.* (2013), pp. 307-309. ■

La (16.2) afferma che al tendere di n a infinito, la funzione di ripartizione della variabile casuale Z_n tende alla funzione di ripartizione della normale standard e ciò a prescindere da quale sia la distribuzione di probabilità delle variabili casuali della successione $X_1, X_2, \ldots, X_n$. L'unica condizione è che tali variabili siano indipendenti e abbiano la stessa distribuzione di probabilità.

Il teorema del limite centrale prende il suo nome probabilmente dal titolo di un articolo di G. Polya del 1920, in cui appare l'espressione *On the central limit theorem of calculus of probability* riferita all'importanza, alla centralità, che ha rivestito per decenni il problema della distribuzione della media di n variabili casuali quando n tende a infinito.

ESEMPIO 16.2

Si riprenda l'Esempio 16.1 concernente la media aritmetica di n variabili casuali indipendenti e distribuite secondo una Poisson con parametro $\lambda = 1.5$. Nella Figura 16.3, è rappresentata, con la linea continua, la funzione di ripartizione della media aritmetica standardizzata, $Z_n = (\bar{X}_n - 1.5)/(1.5/\sqrt{n})$, per $n = 2, 5, 10, 20$, e, con la linea tratteggiata, la funzione di ripartizione della distribuzione normale standard.

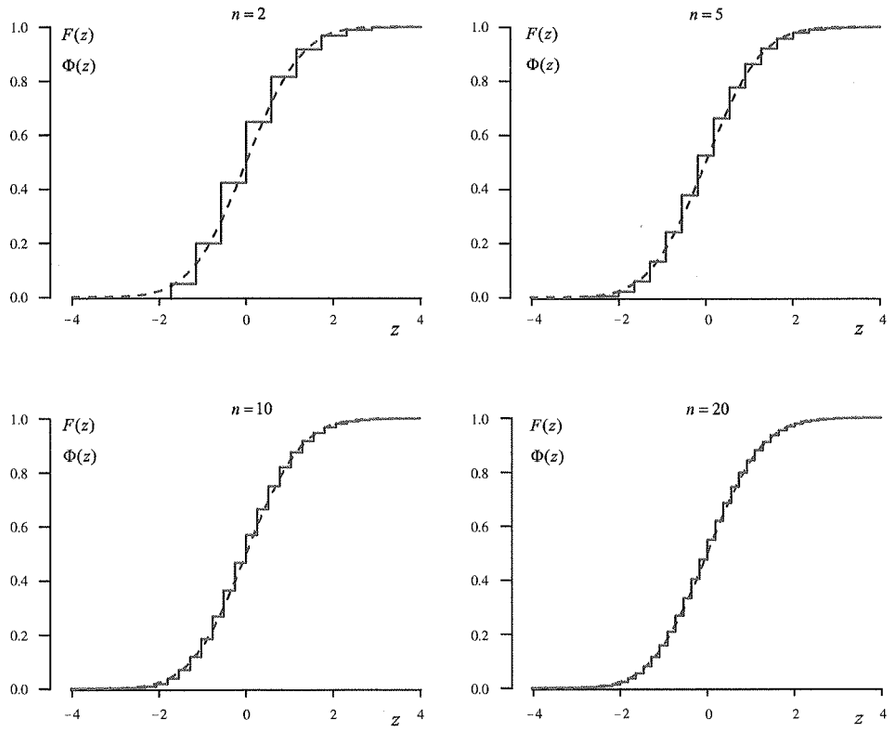


Figura 16.3
Funzione di ripartizione della media aritmetica standardizzata di n ($n = 2, 5, 10, 20$) variabili casuali indipendenti con distribuzione di Poisson di parametro $\lambda = 1.5$ (linea continua) e sua approssimazione con la funzione di ripartizione della distribuzione normale standard (linea tratteggiata).

Come si può osservare, al crescere di n, la funzione di ripartizione effettiva di Z_n tende ad avvicinarsi sempre più alla funzione di ripartizione della distribuzione normale standard. ■

L'importanza pratica di questo teorema è dovuta alla possibilità di approssimare la distribuzione di probabilità di $\bar{X}_n$, quando n è molto grande, con la distribuzione normale. Ciò consente di scrivere, per ogni coppia di numeri a e b (con $a < b$),

$$P(a < \bar{X}_n < b) = P\left(\frac{a-\mu}{\sigma/\sqrt{n}} < \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} < \frac{b-\mu}{\sigma/\sqrt{n}}\right) \approx \Phi\left(\frac{b-\mu}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{a-\mu}{\sigma/\sqrt{n}}\right).$$

Per rafforzare la comprensione della proposizione, è utile osservare la Figura 16.4, dove viene rappresentata la distribuzione di probabilità della media $\bar{X}_n$, per $n = 2, 5, 20$, nel caso delle seguenti assunzioni per le variabili casuali $X_1, X_2, \ldots, X_n$:

(a) distribuzione di Poisson di parametro $\lambda = 0.5$;

(b) distribuzione uniforme continua nell'intervallo (0, 1);

(c) distribuzione esponenziale di parametro $\lambda = 1$.

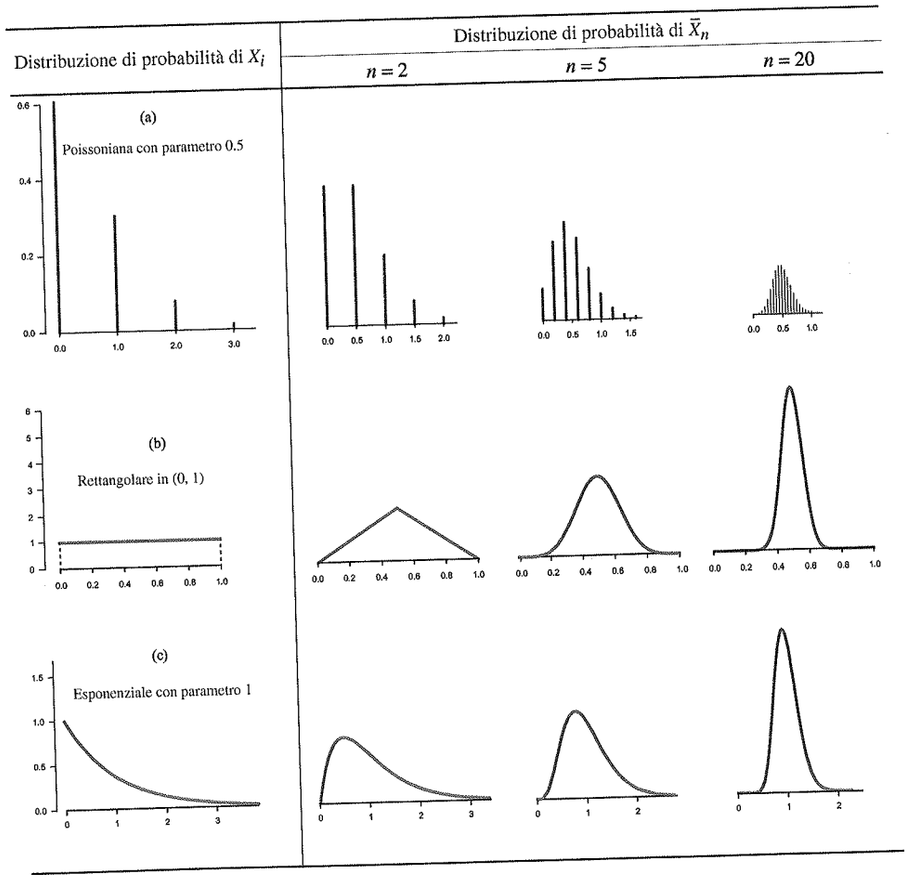


Figura 16.4
Distribuzione di probabilità di $\bar{X}_n$ per tre diverse ipotesi sulla distribuzione delle variabili casuali $X_1, X_2, \ldots, X_n$:
(a) Poissoniana;
(b) Rettangolare;
(c) Esponenziale.

I tre grafici a destra della linea verticale rappresentano per ogni n l'effettiva funzione di probabilità o di densità, che qui non viene riprodotta per brevità. In tutti e tre i casi, passando da $n = 2$ a $n = 5$, a $n = 20$, la funzione di probabilità o di densità si avvicina sempre di più alla forma della normale. La velocità di avvicinamento alla curva normale è più forte nel caso (b), in cui la distribuzione delle variabili casuali è simmetrica.

È facile stabilire che il teorema del limite centrale vale anche per la somma

$$S_n = \sum_{i=1}^{n} X_i,$$

mantenendo le stesse assunzioni di indipendenza e di indentità distributiva delle variabili casuali $X_1, X_2, \ldots, X_n$. Infatti, poiché (come è facile verificare) $\mathrm{E}(S_n) = n\mu$, $\mathrm{Var}(S_n) = n\sigma^2$ e $(S_n - n\mu)/\sqrt{n\sigma^2} = (\bar{X}_n - \mu)/(\sigma/\sqrt{n})$, sussiste la seguente relazione

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \le z\right) = \lim_{n\to\infty} P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le z\right) = \Phi(z).$$

Ciò implica che, sotto le ipotesi fatte, se n è abbastanza grande, la probabilità che S_n assuma valori nell'intervallo (a, b) può essere approssimata con la distribuzione normale:

$$P(a < S_n < b) \approx \Phi\left(\frac{b-n\mu}{\sqrt{n\sigma^2}}\right) - \Phi\left(\frac{a-n\mu}{\sqrt{n\sigma^2}}\right).$$

16.3.1 Approssimazione della distribuzione binomiale con la normale

Come si è visto nel Paragrafo 16.2, quando le variabili casuali della successione $X_1, X_2, \ldots, X_n$ sono indipendenti e tutte distribuite secondo una Bernoulliana con parametro p, la somma

$$S_n = \sum_{i=1}^{n} X_i$$

è una variabile casuale con distribuzione binomiale (che qui viene indicata con X, in linea con la notazione del Capitolo 14), cioè è il numero di successi in n prove Bernoulliane indipendenti tutte con probabilità di successo pari a p. Se n è sufficientemente grande, l'applicazione del teorema del limite centrale a questa somma, che ha valore atteso np e varianza $np(1-p)$, consente di scrivere

$$P(a \le X \le b) \approx \Phi\left(\frac{b-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-np}{\sqrt{np(1-p)}}\right).$$

L'approssimazione migliora se si adotta la correzione per continuità, in base alla quale la probabilità che, in una distribuzione binomiale di parametri n e p, la variabile casuale X assuma valori nell'intervallo $[a, b]$ viene approssimata con l'area sottesa alla curva normale con media np e varianza $np(1-p)$, da $a - 0.5$ a $b + 0.5$:

$$P(a \le X \le b) \approx \Phi\left(\frac{b+0.5-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-0.5-np}{\sqrt{np(1-p)}}\right).$$

Studi empirici hanno dimostrato che l'approssimazione con la distribuzione normale è soddisfacente se i parametri n e p soddisfano le disequazioni:

$$np \ge 5, \quad n(1-p) \ge 5.$$

ESEMPIO 16.3

Si approssimi con la normale la distribuzione binomiale di parametri $n = 90$ e $p = 0.30$. In particolare, si determini la probabilità che il numero dei successi sia compreso nell'intervallo $[20, 37]$.

Soluzione

Si ha

$$P(20 \le X \le 37) \approx \Phi\left(\frac{37.5 - 90 \cdot 0.3}{\sqrt{90 \cdot 0.3 \cdot 0.7}}\right) - \Phi\left(\frac{19.5 - 90 \cdot 0.3}{\sqrt{90 \cdot 0.3 \cdot 0.7}}\right)$$
$$= 0.9922 - (1 - 0.9582) = 0.9504.$$

Si noti che la probabilità esatta, calcolata applicando la distribuzione binomiale, è 0.9517. ■

La qualità dell'approssimazione è molto accurata quando $p = 0.5$ (la distribuzione binomiale è simmetrica), lo è meno quando $p \neq 0.5$ (la distribuzione binomiale è asimmetrica). Ciò emerge anche dall'esame della Figura 16.5, dove sono rappresentate le distribuzioni binomiali esatte,[2] per $p = 0.3,\ 0.5$, e per $n = 5, 10, 50$. Come si può vedere, al crescere di n, l'avvicinamento della distribuzione binomiale alla normale è più veloce quando $p = 0.5$.

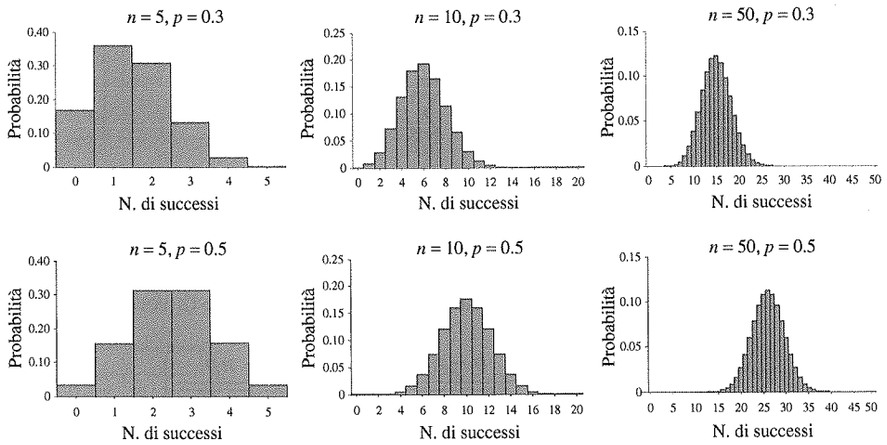


Figura 16.5
Distribuzione binomiale per diversi valori di n e p.

ESEMPIO 16.4

Data una variabile casuale X avente distribuzione binomiale con $n = 50$ e $p = 0.4$, si approssimi con la normale la probabilità che $X = 24$.

Soluzione

Ovviamente, bisogna considerare l'intervallo [23.5, 24.5]. Si ha allora

$$P(X = 24) \approx \Phi\left(\frac{24.5 - 50 \cdot 0.4}{\sqrt{50 \cdot 0.4 \cdot 0.6}}\right) - \Phi\left(\frac{23.5 - 50 \cdot 0.4}{\sqrt{50 \cdot 0.4 \cdot 0.6}}\right)$$
$$= 0.9032 - 0.8438 = 0.0594.$$

La probabilità esatta, cioè quella che si ottiene usando la distribuzione binomiale, è 0.0584. ■

[2] Si noti che la distribuzione binomiale è rappresentata con un istogramma. La probabilità del singolo valore x è pari all'area del rettangolo (che coincide con l'altezza di questo) centrato su x, avente come base il segmento di estremi $x - 0.5$ e $x + 0.5$. Ciò per poter meglio confrontare la distribuzione binomiale con quella normale per mezzo delle aree sottese alla curva.

Riassunto

Questo breve capitolo presenta due risultati fondamentali della teoria della probabilità: la legge dei grandi numeri e il teorema del limite centrale. La legge dei grandi numeri stabilisce che la media aritmetica $\bar{X}_n$ di una sequenza di variabili casuali $X_1, X_2, \ldots, X_n$, indipendenti e tutte con la stessa distribuzione di probabilità, e perciò con lo stesso valore atteso μ e la stessa varianza σ^2, soddisfa la disuguaglianza

$$|\bar{X}_n - \mu| < \varepsilon,$$

dove ε è un numero positivo piccolo a piacere, con una probabilità che tende a 1 al tendere di n a infinito.

Sotto le stesse ipotesi sulle $X_1, X_2, \ldots, X_n$, il teorema del limite centrale stabilisce che la distribuzione di probabilità della variabile casuale $Z_n = (\bar{X}_n - \mu)/(\sigma/\sqrt{n})$ tende alla distribuzione normale standard al tendere di n a infinito. L'importanza del teorema è dovuta alla possibilità di approssimare la distribuzione di $\bar{X}_n$ con la distribuzione normale quando n è sufficientemente grande.

Parole chiave

correzione per continuità, 351
legge dei grandi numeri, 346
legge debole dei grandi numeri, 347
legge dei grandi numeri di Bernoulli, 348
legge forte dei grandi numeri, 347
teorema del limite centrale, 349

Commenti e curiosità

1. Legge dei grandi numeri e monete bilanciate

Come indicato alla fine del Capitolo 12 riguardo alla verifica della legge empirica del caso, si è fatto riferimento all'esperimento di G.L. Buffon (1707-1788), che in 4040 lanci di moneta ottenne 2048 volte testa, con una frequenza relativa pari a 0.5069, e a quello di K. Pearson (1857-1936), che in 24 000 lanci di moneta ottenne 12 012 volte testa, con una frequenza relativa pari a 0.5005. Per quanto sia difficile pensare che Karl Pearson abbia effettivamente lanciato una moneta 24 000 volte, il fatto che figure così autorevoli abbiamo parlato di simili esperimenti è dovuto probabilmente al desiderio di trasmettere l'idea di formidabili regolarità statistiche: su molti lanci, circa metà delle volte esce testa e metà delle volte croce, e, all'aumentare del numero dei lanci, il numero delle teste tende ad avvicinarsi al numero delle croci. Nei due esperimenti, infatti, la differenza tra teste e croci è 56 su 4040 lanci e 24 su 24 000 lanci.

A essere rigorosi, la regolarità statistica non attiene ai numeri assoluti (numero di teste e numero di croci), ma alle frequenze relative. Per rendersene conto, basta simulare il lancio ripetuto di una moneta mediante una passeggiata casuale, in cui si fa un passo a destra (+1) se esce testa e uno a sinistra (−1) se esce croce. La Figura 16.6 mostra una tipica traiettoria di una siffatta passeggiata casuale, mentre la Figura 16.7 mostra questa stessa traiettoria divisa per il numero dei passi.

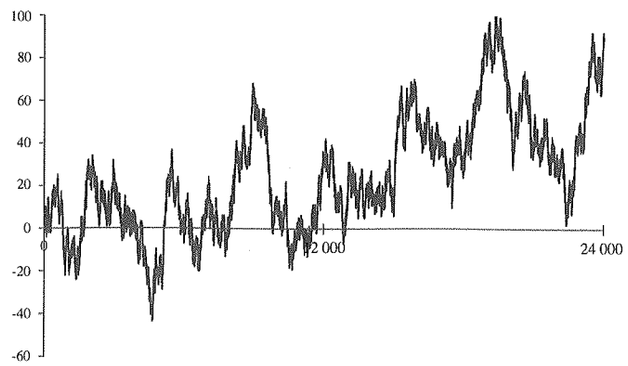
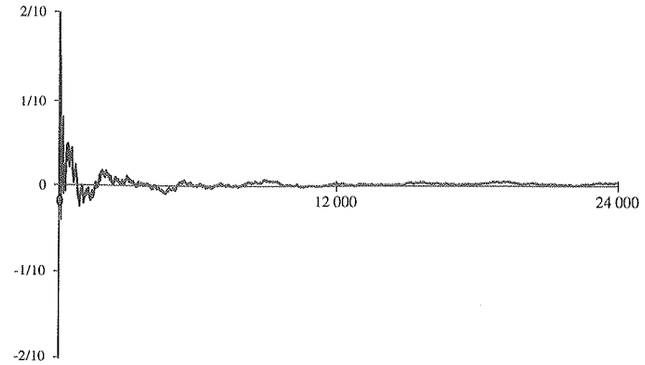


Figura 16.6
Tipica traiettoria di una passeggiata casuale associata ai lanci di una moneta bilanciata.

Figura 16.7
Tipica traiettoria di una passeggiata casuale divisa per il numero di passi.



Come si deduce dalla Figura 16.6, all'aumentare del numero di passi, il numero delle teste tende a essere sempre più lontano dal numero delle croci. La regolarità statistica, implicita nella legge empirica del caso e illuminata dalla legge dei grandi numeri (formulazione di Bernoulli), non riguarda la differenza assoluta tra il numero delle teste e il numero delle croci, bensì le corrispondenti frequenze relative, come si evince dall'osservazione della Figura 16.7.

2. Bernoulli e la prima versione della legge dei grandi numeri

INTERFOTO/Alamy Stock Photo

Jacob Bernoulli (1654-1705), figlio del matematico Nicolaus Bernoulli, svolse i suoi studi universitari in teologia, ma, contro la volontà dei genitori, studiò anche matematica e astronomia. Ottenuta la laurea, si trasferì a Ginevra dove lavorò come istitutore. In seguito, si spostò prima in Francia, dove studiò con gli allievi di Cartesio, e successivamente in Olanda, dove incontrò molti matematici illustri. Tornato a Basilea, Jacob ebbe la sua prima cattedra in matematica nel 1687. In questo periodo, avviò agli studi della matematica infinitesimale suo fratello Johann, con cui poi iniziò una collaborazione scientifica che presto però si trasformò in forte competizione. Gli interessi scientifici di Jacob Bernoulli si concentrarono sulla logica, l'algebra, la geometria e il calcolo delle probabilità. In quest'ultimo campo, i suoi maggiori contributi sono contenuti nel volume dal titolo *Ars Conjectandi*, pubblicato postumo nel 1713. Al suo nome è associata la prima versione della legge dei grandi numeri, quella collegata a sequenze di variabili casuali indipendenti che assumono valore 1 e 0 con probabilità p e $1 - p$.

Esercizi

16.1 Si immagini di estrarre a sorte, con reinserimento, n numeri da un'urna contenente palline numerate da 1 a 10 e si consideri la variabile casuale data dalla media aritmetica dei numeri estratti. Applicando la disuguaglianza di Chebyshev:

a) si calcoli il limite inferiore della probabilità che tale variabile casuale disti (in valore assoluto) dalla media per meno di 0.3 per $n = 150$;

b) si ripeta il calcolo di cui al punto precedente assumendo di effettuare $n = 250$ estrazioni;

c) si dimostri che il limite inferiore della probabilità in questione aumenta all'aumentare di n.

16.2 Dalla passata esperienza si sa che il numero di errori di stampa presenti in una pagina di prime bozze segue la distribuzione di Poisson con parametro 3.5. Si calcoli, usando il teorema del limite centrale, la probabilità che il numero medio di errori per pagina nelle 320 pagine di prime bozze di un volume sia compreso tra 3.2 e 3.7.

16.3 Un componente elettronico per lavastoviglie usato da un'azienda meccanica ha una durata in mesi descritta da una variabile casuale chi-quadrato con 50 gradi di libertà. Si calcoli, usando il teorema del limite centrale, la probabilità che la durata media dei componenti installati nelle 127 lavastoviglie prodotte in un giorno da un reparto dell'azienda sia maggiore di 51.8 mesi.

16.4* Si assuma che il risultato di un certo esperimento sia descritto da una variabile casuale discreta con la seguente distribuzione di probabilità:

x	$\mu-2$	$\mu-1$	μ	$\mu+1$	$\mu+2$
$f(x)$	1/10	2/10	4/10	2/10	1/10

Si può verificare che il valore atteso di questa distribuzione è pari a μ e che la varianza è pari a 1.2. Si assuma inoltre di ripetere questo esperimento n volte nelle stesse condizioni. Indicando con $\bar{X}_n$ la media aritmetica degli n risultati ottenuti, si facciano le seguenti operazioni:

a) si determinino $\mathrm{E}(\bar{X}_n)$ e $\mathrm{Var}(\bar{X}_n)$;

b) assumendo $n = 100$, si determini, tramite la disuguaglianza di Chebyshev, il limite inferiore per la probabilità che $\bar{X}_n$ disti da μ per meno di 0.3;

c) si ripeta il punto precedente usando il teorema del limite centrale;

d) si confrontino i risultati ottenuti ai punti b e c e si commenti;

e) utilizzando il teorema del limite centrale, si stabilisca quale dovrebbe essere il numero n di ripetizioni dell'esperimento affinché $|\bar{X}_n - \mu| < 0.2$ con una probabilità di 0.95.

16.5 Nel sommare 200 numeri reali, ogni numero viene arrotondato allo 0 o all'intero più vicino. Assumendo che l'arrotondamento sia descritto da una variabile casuale uniforme nell'intervallo $(-0.5,\ 0.5)$, si calcoli la probabilità che l'errore nella somma sia minore di 7.07.

16.6 Gli opuscoli pubblicitari stampati da una tipografia sono confezionati in pacchi da 100 esemplari. Si supponga che il peso del singolo opuscolo sia assimilabile a una variabile casuale avente media 25 g e deviazione standard 1.2 g. Qual è la probabilità che un pacco scelto a caso abbia un peso maggiore di 2.52 kg?

16.7 Sia X una variabile casuale discreta con distribuzione binomiale di parametri $n = 28$ e $p = 0.6$. Si calcolino le probabilità degli eventi $X = 19$, $12 \leq X \leq 24$, $15 < X < 22$, usando l'approssimazione normale. Si confrontino poi i risultati con le probabilità fornite dalla binomiale (per i calcoli relativi alla binomiale, si consiglia di utilizzare un foglio elettronico o un altro programma di calcolo).

16.8 Si determini la probabilità di ottenere almeno 25 volte testa in 40 lanci di una moneta bilanciata. Si utilizzi l'approssimazione normale con e senza correzione per continuità e si confrontino i risultati con la probabilità che si ottiene con l'applicazione della binomiale (per i calcoli relativi alla binomiale, si consiglia di utilizzare un foglio elettronico o un altro programma di calcolo statistico).

16.9 Il tasso di disoccupazione in una data città è pari al 7.6%. Vengono estratte a sorte (con ripetizione) 200 unità dalle forze di lavoro (parte della popolazione composta da chi lavora o cerca lavoro). Si utilizzi l'approssimazione normale per calcolare le probabilità indicate di seguito:

a) la probabilità di osservare meno di 15 disoccupati;

b) la probabilità di osservare non meno di 18 disoccupati.

16.10 Il peso delle scatole di pelati dell'azienda alimentare A si distribuisce secondo una legge normale con media 1 kg e deviazione standard 0.009 kg:

a) si calcoli la probabilità che una scatola scelta a caso pesi meno di 990 g;

b) si calcoli la probabilità che su 70 scatole scelte a caso ve ne siano almeno 2 che pesano meno di 990 g, utilizzando l'approssimazione normale.

16.11 In un processo di controllo industriale si assume che il peso di una confezione di pasta sia una variabile casuale con distribuzione normale con valore atteso pari a 500 g e varianza pari a 25. Se

il peso di una confezione è inferiore a 490 g o superiore a 510 g, la confezione viene rifiutata. Considerando un campione di 100 confezioni scelte a caso, si calcolino le probabilità di seguito indicate:

a) la probabilità che al più 10 confezioni siano rifiutate;
b) la probabilità che almeno 80 confezioni siano accettate, approssimando la binomiale con la normale.

CAPITOLO

17

Campioni casuali e distribuzioni campionarie

Sommario del capitolo

17.1 Introduzione
17.2 Campione casuale
17.3 Distribuzione di probabilità del campione casuale
17.4 Statistiche e distribuzioni campionarie
17.5 Distribuzione campionaria della media
17.6 Distribuzione campionaria della varianza
17.7 Distribuzione campionaria della media quando la varianza σ^2 non è nota
17.8 Stima dei parametri e verifica delle ipotesi
Riassunto
Commenti e curiosità
Esercizi

17.1 Introduzione

Nei Capitoli 12-16, sono state presentate le basi della probabilità. Questo capitolo costituisce l'anello di congiunzione tra la teoria della probabilità e l'inferenza statistica, avendo come protagonisti il campione casuale, le statistiche campionarie e le distribuzioni campionarie.

Un campione casuale è un insieme di variabili casuali indipendenti, tutte con la stessa distribuzione di probabilità, che originano dall'estrazione casuale (con ripetizione) di un certo numero di elementi da una popolazione finita o dalla ripetizione di un esperimento casuale un certo numero di volte nelle stesse condizioni. Una statistica campionaria è una variabile casuale definita come funzione del campione casuale, come ad esempio la media e la varianza campionaria. Infine, una distribuzione campionaria non è altro che la distribuzione di probabilità di una statistica campionaria.

All'introduzione di queste nozioni, farà seguito l'illustrazione in dettaglio della distribuzione campionaria della media e della varianza, in considerazione dell'importanza che queste due statistiche hanno nell'ambito dei metodi di inferenza che saranno trattati nei Capitoli 18-24.

17.2 Campione casuale

Per introdurre il fondamentale concetto di campione casuale, si inizia con alcuni semplici esempi.

ESEMPIO 17.1

Presso un'azienda industriale un lotto di semilavorati è sottoposto al controllo di qualità. L'azienda è interessata a osservare il numero di difetti per unità, X. Si assuma che la distribuzione di frequenza relativa di X nel lotto sia quella indicata nella Tabella 17.1.

Tabella 17.1 Semilavorati di un ipotetico lotto per numero di difetti.

N. di difetti x	0	1	2	3	4
Frequenza relativa $f(x)$	$f(0)$	$f(1)$	$f(2)$	$f(3)$	$f(4)$

Si supponga che il controllo di qualità proceda osservando il numero di difetti su 100 pezzi estratti a sorte (con ripetizione) dal lotto. Siano $X_1, X_2, \ldots, X_{100}$ i valori osservabili sull'unità selezionata alla prima, alla seconda, ..., alla centesima estrazione.

È immediato comprendere che X_1 è una variabile casuale che assume i valori 0, 1, 2, 3, 4 con probabilità $f(0)$, $f(1)$, $f(2)$, $f(3)$ e $f(4)$. Le variabili casuali $X_2, X_3, \ldots, X_{100}$ hanno tutte la stessa distribuzione di probabilità di X_1, perché le unità vengono estratte con ripetizione. Per lo stesso motivo, $X_1, X_2, \ldots, X_{100}$ sono indipendenti. In conclusione, $X_1, X_2, \ldots, X_{100}$ sono variabili casuali indipendenti con la stessa distribuzione di probabilità.

L'obiettivo del controllo di qualità è quello di stimare il numero medio di difetti per unità nel lotto e decidere se accettarlo oppure no. ■

ESEMPIO 17.2

In uno studio clinico, 20 pazienti ricevono un dato trattamento. Si supponga che ci sia omogeneità nelle caratteristiche dei pazienti (età, abitudini, presenza di altre malattie ecc.) che insieme al trattamento potrebbero avere un peso nel determinare l'esito della malattia. Sia p la probabilità che un paziente curato guarisca. L'esito del trattamento è una variabile casuale dicotomica, X, che assume i valori 1 (guarigione) e 0 (non guarigione) con probabilità p e $1 - p$.

Ciascuna prova (che consta della somministrazione del trattamento a un paziente e dell'osservazione dell'esito) genera una variabile casuale avente una distribuzione di Bernoulli con parametro p. Ne segue che le 20 prove producono un pari numero di variabili casuali, $X_1, X_2, \ldots, X_{20}$, indipendenti e aventi la stessa distribuzione di Bernoulli.

Il ricercatore è interessato a valutare l'ignota probabilità p a fini di confronto con trattamenti concorrenti. ■

ESEMPIO 17.3

Una linea di produzione di un'azienda alimentare confeziona un certo tipo di pasta in sacchetti il cui peso medio è 500 g con una variabilità pari a $\sigma^2 = 15$. In assenza di fattori di disturbo (cambiamento dei fornitori di materie prime, guasti alle apparecchiature, errori umani ecc.), è ragionevole assumere che il peso della singola confezione, X, si comporti come una variabile casuale normale con media 500 e varianza 15. Si immagini di prelevare 50 confezioni dalla linea di produzione e siano $X_1, X_2, \ldots, X_{50}$ il peso della confezione prelevata per prima, per seconda, ..., per 50-esima.

Chiaramente, $X_1, X_2, \ldots, X_{50}$ sono variabili casuali indipendenti, aventi tutte distribuzione normale con media 500 e varianza 15. Sono indipendenti perché il peso della singola confezione non influisce sul peso delle altre. Hanno la stessa distribuzione di probabilità perché le confezioni sono ottenute nelle stesse condizioni di funzionamento del processo.

In questo caso, si è interessati al peso medio delle confezioni prodotte dalla linea. Nello specifico, il problema è stabilire se tale media è coerente con il valore desiderato (500 g), ossia se il processo è in controllo. ■

L'idea base che accomuna le situazioni illustrate nei tre esempi è la seguente. L'effettuazione di un esperimento casuale (l'estrazione a sorte di un elemento da un lotto di semilavorati, come nell'Esempio 17.1; l'osservazione dell'esito di una malattia in un paziente sottoposto a un trattamento, come nell'Esempio 17.2; l'osservazione del peso di una confezione di pasta, come nell'Esempio 17.3) origina una variabile casuale X.

A n repliche indipendenti dell'esperimento, condotte nelle stesse condizioni,[1] sono associate le variabili casuali $X_1, X_2, \ldots, X_n$ indipendenti e aventi la stessa distribuzione di probabilità di X. Ebbene, queste variabili casuali costituiscono il cosiddetto *campione casuale*.

DEFINIZIONE 17.1

Si chiama campione casuale di ampiezza n una collezione di n variabili casuali, $(X_1, X_2, \ldots, X_n)$, indipendenti e aventi la stessa distribuzione di probabilità.

Il concetto di campione casuale implica quello di popolazione. In linea di principio, il termine popolazione dovrebbe essere utilizzato solo quando un campione è fisicamente estratto da una totalità reale di elementi, cioè da una popolazione finita (si veda il Paragrafo 1.8.1), come nell'Esempio 17.1. Ma è anche comunemente usato quando si tratta di dati generati da esperimenti (si veda il Paragrafo 1.8.2), come negli Esempi 17.2 e 17.3, per indicare i risultati ipotetici di una sequenza infinita di ripetizioni indipendenti dell'esperimento in condizioni identiche. D'ora in poi, per brevità, una tale sequenza infinita viene denominata popolazione infinita.

Nel caso di popolazioni finite, la Definizione 17.1 vale solo se il campione è estratto con ripetizione; non vale quindi se il campione è estratto con una qualsiasi delle tecniche di campionamento per indagini campionarie menzionate nel Paragrafo 1.8.1: campione casuale semplice,[2] campione sistematico, campione stratificato e campione a grappoli.

Sia $f_X(x)$ la funzione di probabilità o di densità che descrive la distribuzione di probabilità di X, di seguito denominata distribuzione di probabilità della popolazione. La funzione $f_X(x)$ può essere nota (o ipotizzata) a meno di una o più costanti numeriche (parametri) indicate con θ. In tal caso, $f_X(x)$ può essere scritta come $f_X(x; \theta)$.

Conviene ora dare la definizione di parametro, l'oggetto dell'inferenza statistica.

DEFINIZIONE 17.2

Si chiama parametro una costante caratteristica della popolazione che può coincidere con la costante θ della distribuzione di probabilità della popolazione $f_X(x; \theta)$.

La Definizione 17.1 può essere riformulata dicendo che un campione casuale è una collezione di n variabili casuali, $(X_1, X_2, \ldots, X_n)$, indipendenti e aventi la stessa distribuzione di probabilità della variabile casuale X oggetto di studio. In simboli, per ogni valore x di X, si può scrivere

$$f_{X_1}(x) = f_{X_2}(x) = \cdots = f_{X_n}(x) = f_X(x), \tag{17.1}$$

dove $f_{X_i}(\cdot)$, $i = 1, 2, \ldots, n$, è la funzione di probabilità o di densità di X_i.

[1] Il significato dell'espressione "nelle stesse condizioni" dipende dal contesto: nell'Esempio 17.1, si fa riferimento all'estrazione con ripetizione; nell'Esempio 17.2, si assume l'omogeneità dei pazienti trattati; nell'Esempio 17.3, si ipotizza che il processo produttivo sia in controllo.

[2] Si ricorda che un campione casuale semplice viene formato estraendo casualmente, senza ripetizione, n unità dalla popolazione.

Nel seguito, verranno usate espressioni come popolazione Bernoulliana, popolazione esponenziale, popolazione normale e così via, per indicare la distribuzione di probabilità della popolazione generatrice, cioè della popolazione da cui proviene il campione.

17.3 Distribuzione di probabilità del campione casuale

Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale proveniente dalla popolazione di interesse. Allora un campione casuale osservabile è una n-upla di numeri, $(x_1, x_2, \ldots, x_n)$, dove x_1 è uno dei possibili valori di X_1, x_2 uno dei possibili valori di X_2 e così via.

A ogni campione casuale osservabile è possibile assegnare l'appropriata probabilità o densità congiunta espressa da

$$f(x_1, x_2, \ldots, x_n) = f_X(x_1) \cdot f_X(x_2) \cdot \cdots \cdot f_X(x_n). \tag{17.2}$$

Questa scrittura è resa possibile dalle due proprietà del campione casuale: l'indipendenza delle variabili casuali $X_1, X_2, \ldots, X_n$ – che consente di esprimere $f(x_1, x_2, \ldots, x_n)$ come prodotto delle probabilità (o densità) marginali – e dal fatto che, in forza della (17.1), la funzione di probabilità o di densità di ogni X_i è identica a quella della variabile casuale X oggetto di studio.

La (17.2) descrive quella che viene chiamata la distribuzione di probabilità del campione casuale. Gli esempi che seguono hanno lo scopo di chiarire questo concetto.

ESEMPIO 17.4

Si consideri la tabella che segue:

Tabella 17.2
Distribuzione di probabilità del campione casuale (X_1, X_2).

Campione osservabile	Probabilità	Campione osservabile	Probabilità
(0, 0)	0.0100	(4, 1)	0.0200
(0, 1)	0.0400	(2, 3)	0.0400
(1, 0)	0.0400	(3, 2)	0.0400
(0, 2)	0.0200	(1, 5)	0.0200
(2, 0)	0.0200	(5, 1)	0.0200
(1, 1)	0.1600	(2, 4)	0.0100
(0, 3)	0.0200	(4, 2)	0.0100
(3, 0)	0.0200	(3, 3)	0.0400
(1, 2)	0.0800	(2, 5)	0.0100
(2, 1)	0.0800	(5, 2)	0.0100
(0, 4)	0.0050	(3, 4)	0.0100
(4, 0)	0.0050	(4, 3)	0.0100
(1, 3)	0.0800	(3, 5)	0.0100
(3, 1)	0.0800	(5, 3)	0.0100
(2, 2)	0.0400	(4, 4)	0.0025
(0, 5)	0.0050	(4, 5)	0.0025
(5, 0)	0.0050	(5, 4)	0.0025
(1, 4)	0.0200	(5, 5)	0.0025

In essa, la prima e terza colonna contengono i campioni di ampiezza 2 estraibili con ripetizione dalla popolazione rappresentata dalla distribuzione delle famiglie di un certo comune per numero di figli:

N. di figli x	0	1	2	3	4	5
Probabilità $f_X(x)$	0.10	0.40	0.20	0.20	0.05	0.05

A ogni campione è associata la probabilità trovata applicando la (17.2). Si ha:

$$\begin{aligned} f(0,0) &= f_X(0) \cdot f_X(0) = 0.10 \cdot 0.10 = 0.01; \\ f(0,1) &= f_X(0) \cdot f_X(1) = 0.10 \cdot 0.40 = 0.04; \\ &\vdots \\ f(5,5) &= f_X(5) \cdot f_X(5) = 0.05 \cdot 0.05 = 0.0025. \end{aligned}$$

La lista dei campioni osservabili con le rispettive probabilità della Tabella 17.2 costituisce la distribuzione di probabilità del campione casuale (X_1, X_2) in esame. ■

ESEMPIO 17.5

La Tabella 17.3 riporta i possibili campioni di ampiezza 4 estraibili dalla popolazione Bernoulliana di parametro p con le rispettive probabilità. Qui la funzione di probabilità della popolazione è

$$f_X(x) = p^x(1-p)^{1-x}, \quad x = 0, 1.$$

La probabilità di un generico campione, (x_1, x_2, x_3, x_4), è data da

$$\begin{aligned} f(x_1, x_2, x_3, x_4) &= p^{x_1}(1-p)^{1-x_1} \cdot p^{x_2}(1-p)^{1-x_2} \cdot p^{x_3}(1-p)^{1-x_3} \cdot p^{x_4}(1-p)^{1-x_4} \\ &= p^{\sum_{i=1}^4 x_i}(1-p)^{4-\sum_{i=1}^4 x_i}. \end{aligned}$$

Ne segue che:

$$\begin{aligned} f(0,0,0,0) &= p^0 \cdot (1-p)^4; \\ f(0,0,0,1) &= p^1 \cdot (1-p)^3; \\ &\vdots \\ f(1,1,1,1) &= p^4 \cdot (1-p)^0. \end{aligned}$$

Campione	Probabilità	Campione	Probabilità
(0, 0, 0, 0)	$(1-p)^4$	(0, 1, 0, 1)	$p^2(1-p)^2$
(0, 0, 0, 1)	$p(1-p)^3$	(1, 0, 1, 0)	$p^2(1-p)^2$
(0, 0, 1, 0)	$p(1-p)^3$	(1, 0, 0, 1)	$p^2(1-p)^2$
(0, 1, 0, 0)	$p(1-p)^3$	(0, 1, 1, 1)	$p^3(1-p)$
(1, 0, 0, 0)	$p(1-p)^3$	(1, 1, 1, 0)	$p^3(1-p)$
(0, 0, 1, 1)	$p^2(1-p)^2$	(1, 1, 0, 1)	$p^3(1-p)$
(0, 1, 1, 0)	$p^2(1-p)^2$	(1, 0, 1, 1)	$p^3(1-p)$
(1, 1, 0, 0)	$p^2(1-p)^2$	(1, 1, 1, 1)	p^4

Tabella 17.3
Campioni di dimensione $n = 4$ da una popolazione Bernoulliana con parametro p e relative probabilità.

■

ESEMPIO 17.6

Si consideri una popolazione normale $N(\mu, \sigma^2)$. I campioni osservabili di dimensione $n = 3$ sono tutte le terne (x_1, x_2, x_3), dove ognuna delle tre componenti può assumere qualsiasi valore tra $-\infty$ e ∞. Per ogni dato campione è possibile scrivere la funzione di densità congiunta sulla base della (17.2), ottenendo

$$\begin{aligned} f(x_1, x_2, x_3) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_1-\mu)^2/2\sigma^2} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_2-\mu)^2/2\sigma^2} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_3-\mu)^2/2\sigma^2} \\ &= \frac{1}{(\sigma\sqrt{2\pi})^3} e^{-[(x_1-\mu)^2+(x_2-\mu)^2+(x_3-\mu)^2]/2\sigma^2}. \end{aligned}$$

■

17.4 Statistiche e distribuzioni campionarie

I procedimenti di stima dei parametri e di verifica delle ipotesi si basano su funzioni dei dati del campione, quali la media e la varianza campionaria.

Sia $(x_1, x_2, \ldots, x_n)$ un campione casuale osservabile. La media aritmetica di questi numeri,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

è una funzione di $x_1, x_2, \ldots, x_n$.

Una seconda quantità d'interesse è rappresentata dalla varianza campionaria, che è data da

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Questa espressione non è altro che la varianza definita nel Capitolo 5 moltiplicata per il fattore $n/(n-1)$ ed è facile stabilire che si tratta di una funzione di $x_1, x_2, \ldots, x_n$.[3]

In generale, sia $t(x_1, x_2, \ldots, x_n)$ una funzione, dove $t(\cdot)$ indica l'operazione da effettuare sui numeri $x_1, x_2, \ldots, x_n$ (somma e divisione per n, per la media campionaria; somma dei quadrati degli scarti dalla media campionaria, e divisione del risultato per $n-1$, per la varianza campionaria ecc.). La funzione $t(\cdot)$, se applicata a una n-upla di numeri dà un numero, se applicata al campione casuale $(X_1, X_2, \ldots, X_n)$ definisce una variabile casuale.

DEFINIZIONE 17.3

Una statistica campionaria o più semplicemente una statistica è una funzione $t(X_1, X_2, \ldots, X_n)$ del campione casuale $(X_1, X_2, \ldots, X_n)$.

Dunque, ogni statistica campionaria è una variabile casuale e come tale è dotata di una propria distribuzione di probabilità, che prende il nome di distribuzione campionaria.

[3] La ragione di questa trasformazione sarà chiarita nel Capitolo 18, quando si parlerà di s^2 come stima della varianza della popolazione.

Per le molteplici applicazioni in cui sono utilizzate, la media e la varianza campionaria meritano un'attenzione speciale. Qui vengono riscritte in lettere maiuscole per rimarcare che sono intese come variabili casuali:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Gli esempi che seguono servono a introdurre in modo intuitivo l'argomento delle distribuzioni campionarie.

ESEMPIO 17.7

Si trovino le distribuzioni campionarie della media e della varianza campionaria per campioni di ampiezza 2 provenienti dalla popolazione dell'Esempio 17.4.

Soluzione

Nella Tabella 17.4 sono riportate, accanto a ciascun campione, la media, la varianza campionaria e la probabilità già calcolata nell'Esempio 17.4.

Campione	$\bar{x}$	s^2	Probabilità	Campione	$\bar{x}$	s^2	Probabilità
(0, 0)	0.00	0.0	0.0100	(4, 1)	2.50	4.5	0.0200
(0, 1)	0.50	0.5	0.0400	(2, 3)	2.50	0.5	0.0400
(1, 0)	0.50	0.5	0.0400	(3, 2)	2.50	2.0	0.0400
(0, 2)	1.00	2.0	0.0200	(1, 5)	3.00	8.0	0.0200
(2, 0)	1.00	2.0	0.0200	(5, 1)	3.00	8.0	0.0200
(1, 1)	1.00	0.0	0.1600	(2, 4)	3.00	2.0	0.0100
(0, 3)	1.50	4.5	0.0200	(4, 2)	3.00	2.0	0.0100
(3, 0)	1.50	4.5	0.0200	(3, 3)	3.00	0.0	0.0400
(1, 2)	1.50	0.5	0.0800	(2, 5)	3.50	4.5	0.0100
(2, 1)	1.50	0.5	0.0800	(5, 2)	3.50	4.5	0.0100
(0, 4)	2.00	8.0	0.0050	(3, 4)	3.50	0.5	0.0100
(4, 0)	2.00	8.0	0.0050	(4, 3)	3.50	0.5	0.0100
(1, 3)	2.00	2.0	0.0800	(3, 5)	4.00	2.0	0.0100
(3, 1)	2.00	2.0	0.0800	(5, 3)	4.00	2.0	0.0100
(2, 2)	2.00	0.0	0.0400	(4, 4)	4.00	0.0	0.0025
(0, 5)	2.50	12.5	0.0050	(4, 5)	4.50	0.5	0.0025
(5, 0)	2.50	12.5	0.0050	(5, 4)	4.50	0.5	0.0025
(1, 4)	2.50	4.5	0.0200	(5, 5)	5.00	0.0	0.0025

Tabella 17.4
Campioni, valori di $\bar{X}$ e S^2, e probabilità relative al campione casuale (X_1, X_2) di cui all'Esempio 17.4.

La distribuzione campionaria della media è contenuta nella Tabella 17.5.

$\bar{x}$	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
$f(\bar{x})$	0.01	0.08	0.20	0.20	0.21	0.13	0.10	0.04	0.0225	0.005	0.0025

Tabella 17.5
Distribuzione campionaria della media per campioni di ampiezza 2 provenienti dalla popolazione dell'Esempio 17.4.

Si noti che la probabilità del singolo valore di $\bar{X}$ è data dalla somma delle probabilità dei campioni in cui la media assume tale valore. Per esempio:

$$f(2) = P(\bar{X} = 2) = 0.005 + 0.005 + 0.080 + 0.080 + 0.040 = 0.21,$$

dove gli addendi sono le probabilità dei campioni (0, 4), (4, 0), (1, 3), (3, 1), (2, 2).

La distribuzione campionaria della varianza è esposta nella Tabella 17.6.

bella 17.6
stribuzione campionaria
la varianza per campioni
ampiezza 2 provenienti
lla popolazione
l'Esempio 17.4.

s^2	0.0	0.5	2.0	4.5	8.0	12.5
$f(s^2)$	0.255	0.345	0.240	0.100	0.050	0.010

Anche qui, la probabilità di ciascun valore di S^2 si ottiene sommando le probabilità dei campioni in cui la varianza campionaria assume tale valore. Per esempio, la probabilità che $S^2 = 0$ è data da

$$f(0) = P(S^2 = 0) = 0.0100+0.1600+0.0400+0.0400+0.0025+0.0025 = 0.255,$$

dove gli addendi sono le probabilità dei campioni $(0,0)$, $(1,1)$, $(2,2)$, $(3,3)$, $(4,4)$, $(5,5)$. ■

Si consideri ora il caso di una popolazione Bernoulliana. Se $(X_1, X_2, \ldots, X_n)$ è un campione casuale da questa popolazione, l'i-esimo elemento X_i assume valore 1 (successo) o 0 (insuccesso) con probabilità p e $1-p$. In accordo con quanto detto nel Capitolo 16, la somma $\sum_{i=1}^{n} X_i$ e la media $\sum_{i=1}^{n} X_i/n$ diventano la frequenza assoluta e relativa dei successi nel campione. Quest'ultima è anche chiamata proporzione campionaria e indicata con $\hat{p}$.

ESEMPIO 17.8

Si determini la distribuzione campionaria della proporzione $\hat{p}$ per campioni di ampiezza 4 da una popolazione Bernoulliana di parametro p.

Soluzione

La Tabella 17.7 è ottenuta dalla Tabella 17.3 aggiungendo la colonna relativa alla proporzione campionaria $\hat{p}$:

abella 17.7
alori di $\hat{p}$ e probabilità
ei campioni di ampiezza 4
a una popolazione
ernoulliana
parametro p.

Campione	$\hat{p}$	Probabilità	Campione	$\hat{p}$	Probabilità
$(0,0,0,0)$	0.00	$(1-p)^4$	$(0,1,0,1)$	0.50	$p^2(1-p)^2$
$(0,0,0,1)$	0.25	$p(1-p)^3$	$(1,0,1,0)$	0.50	$p^2(1-p)^2$
$(0,0,1,0)$	0.25	$p(1-p)^3$	$(1,0,0,1)$	0.50	$p^2(1-p)^2$
$(0,1,0,0)$	0.25	$p(1-p)^3$	$(0,1,1,1)$	0.75	$p^3(1-p)$
$(1,0,0,0)$	0.25	$p(1-p)^3$	$(1,1,1,0)$	0.75	$p^3(1-p)$
$(0,0,1,1)$	0.50	$p^2(1-p)^2$	$(1,1,0,1)$	0.75	$p^3(1-p)$
$(0,1,1,0)$	0.50	$p^2(1-p)^2$	$(1,0,1,1)$	0.75	$p^3(1-p)$
$(1,1,0,0)$	0.50	$p^2(1-p)^2$	$(1,1,1,1)$	1.00	p^4

La distribuzione campionaria di $\hat{p}$ è contenuta nella Tabella 17.8. In essa, al singolo valore di $\hat{p}$ corrisponde la somma delle probabilità dei campioni in cui $\hat{p}$ assume tale valore. È facile riconoscere che la distribuzione di probabilità di $\hat{p}$ è una distribuzione binomiale di parametri 4 e p.

$\hat{p}$	0.00	0.25	0.50	0.75	1.00
$f(\hat{p})$	$(1-p)^4$	$4p(1-p)^3$	$6p^2(1-p)^2$	$4p^3(1-p)$	p^4

Tabella 17.8
Distribuzione campionaria di $\hat{p}$ per campioni di ampiezza 4 tratti da una popolazione Bernoulliana.

■

Gli esempi precedenti suggeriscono che la distribuzione campionaria di una statistica dipende dalla forma della distribuzione di probabilità $f_X(x)$ della popolazione. A parte i casi elementari descritti sopra, determinare la distribuzione campionaria di una statistica T (a partire dalla distribuzione di probabilità di una data popolazione) è un'operazione che richiede un bagaglio di conoscenze matematiche che va oltre quello richiesto per la lettura di questo volume. Qui ci si limita all'affermazione intuitiva che segue. La probabilità o la densità da assegnare a uno specifico valore v di una statistica campionaria si può ottenere dalla somma delle probabilità o dalla "somma nel continuo" delle densità di probabilità di tutti i campioni osservabili nei quali si realizza l'uguaglianza $t(x_1, x_2, \ldots, x_n) = v$.

17.5 Distribuzione campionaria della media

Conviene iniziare con il valore atteso e la varianza della media campionaria.

PROPOSIZIONE 17.1

Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale proveniente da una popolazione qualsiasi con media μ e varianza σ^2. Allora il valore atteso $\mu_{\bar{X}}$ e la varianza $\sigma^2_{\bar{X}}$ di $\bar{X}$ sono date da:

$$\mu_{\bar{X}} = \mu, \quad \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}. \tag{17.3}$$

Indipendentemente dalla forma della popolazione, il valore atteso e la varianza della media campionaria sono uguali, rispettivamente, alla media della popolazione e alla varianza della popolazione divisa per n.

Le (17.3) seguono immediatamente dalla Proposizione 15.5. ■

ESEMPIO 17.9

Si consideri una popolazione con media 100 e varianza 18. Sia $\bar{X}$ la media di un campione casuale di 20 elementi proveniente da tale popolazione. Le (17.3) consentono di affermare che il valore atteso e la varianza di $\bar{X}$ sono:

$$\mu_{\bar{X}} = 100, \quad \sigma^2_{\bar{X}} = \frac{18}{20} = 0.9.$$

■

La deviazione standard della media campionaria, radice quadrata di $\sigma^2_{\bar{X}}$, è una misura della distanza media dei valori di $\bar{X}$ da μ.

Quando la popolazione è Bernoulliana, le (17.3) divengono:

$$\mu_{\hat{p}} = p, \quad \sigma^2_{\hat{p}} = \frac{p(1-p)}{n}. \tag{17.4}$$

La transizione dalle (17.3) alle (17.4) è immediata considerando che: la media campionaria $\bar{X}$ è la proporzione campionaria $\hat{p}$, la media e la varianza di una popolazione Bernoulliana sono p e $p(1-p)$.

ESEMPIO 17.10

Si determinino il valore atteso e la varianza della proporzione campionaria $\hat{p}$ per campioni di ampiezza $n = 65$ provenienti da una popolazione Bernoulliana di parametro $p = 0.4$.

Soluzione

L'applicazione delle (17.4) dà:

$$\mu_{\hat{p}} = \mathrm{E}(\hat{p}) = 0.4, \quad \sigma^2_{\hat{p}} = \frac{0.4 \cdot 0.6}{65} = 0.0037.$$

■

La distribuzione campionaria della media, come quella di ogni altra statistica, dipende dalle caratteristiche della popolazione. Per chiarire questo punto, si considerino due specifici casi: la popolazione Bernoulliana e quella esponenziale. Si può dimostrare che la distribuzione campionaria della media è descritta, nel primo caso, dalla funzione di probabilità

$$f(\hat{p}) = \frac{n!}{(n\hat{p})!(n - n\hat{p})!} p^{n\hat{p}}(1-p)^{n-n\hat{p}}, \quad \hat{p} = 0, 1/n, 2/n, \ldots, 1;$$

nel secondo, dalla funzione di densità

$$f(\bar{x}) = \frac{(n\lambda)^n}{(n-1)!}\bar{x}^{n-1}e^{-n\lambda\bar{x}}, \quad \bar{x} \geq 0.$$

Come dire che a ogni specifica popolazione corrisponde una diversa distribuzione campionaria della media. Tuttavia, il problema è reso meno complicato da due circostanze:

- in molte applicazioni, si può assumere che la popolazione sia normale;
- l'influenza della struttura della popolazione sulla distribuzione campionaria di $\bar{X}$ si attenua all'aumentare della numerosità del campione. In forza del teorema del limite centrale, quando la numerosità campionaria è sufficientemente grande, la distribuzione campionaria della media può essere approssimata con la normale, qualunque sia la forma della popolazione.

Queste questioni saranno trattate nei prossimi due paragrafi.

17.5.1 Popolazione normale

Quando la popolazione è normale, anche la media campionaria ha distribuzione normale.

Il caso in cui la popolazione è normale merita un'attenzione particolare, perché una parte importante dell'inferenza statistica riguarda la stima di parametri o la verifica di ipotesi su popolazioni normali e coinvolge la media campionaria.

PROPOSIZIONE 17.2

Sia $\bar{X}$ la media di un campione casuale di dimensione n proveniente da una popolazione normale con media μ e varianza σ^2. Allora $\bar{X}$ ha distribuzione normale con media μ e varianza σ^2/n. Dunque, la funzione di densità di $\bar{X}$ è data da

$$f(\bar{x}) = \frac{1}{(\sigma/\sqrt{n})\sqrt{2\pi}} e^{-(\bar{x}-\mu)^2/(2\sigma^2/n)}.$$

La Proposizione 17.2 è un caso particolare del risultato più generale enunciato nella Proposizione 15.4, per cui ogni combinazione lineare di variabili casuali normali indipendenti ha distribuzione normale. Basta osservare infatti che la media campionaria è una particolare combinazione lineare delle variabili casuali $X_1, X_2, \ldots, X_n$ con coefficienti numerici tutti uguali a $1/n$. ■

La Proposizione 17.2 è illustrata nella Figura 17.1, dove appaiono la distribuzione di probabilità della popolazione e la distribuzione campionaria di $\bar{X}$.

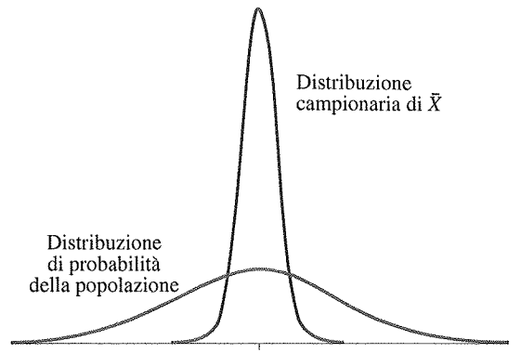


Figura 17.1
Distribuzione di probabilità della popolazione e distribuzione campionaria di $\bar{X}$.

Il risultato descritto nella Proposizione 17.2 dà la possibilità di calcolare la probabilità che $\bar{X}$ assuma valori in un intervallo qualsiasi (a, b) come

$$P(a < \bar{X} < b) = P\left(\frac{a-\mu}{\sigma/\sqrt{n}} < Z_{\bar{X}} < \frac{b-\mu}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{b-\mu}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{a-\mu}{\sigma/\sqrt{n}}\right),$$

dove $Z_{\bar{X}} = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ ha distribuzione normale standard (si veda la Proposizione 14.8).

ESEMPIO 17.11

Si consideri un campione casuale di ampiezza 15 proveniente da una popolazione normale con media 175 e varianza 47. Si determini la probabilità che $\bar{X}$ assuma valori nell'intervallo (173, 177.5).

Soluzione

Sulla base della Proposizione 17.2, si può scrivere

$$\begin{aligned} P(173 < \bar{X} < 177.5) &= P\left(\frac{173-175}{\sqrt{47/15}} < Z_{\bar{X}} < \frac{177.5-175}{\sqrt{47/15}}\right) \\ &= P(-1.13 < Z_{\bar{X}} < 1.41) = \Phi(1.41) - \Phi(-1.13) = 0.7915. \end{aligned}$$

■

17.5.2 Grandi campioni

Qualunque sia la forma della popolazione, la distribuzione campionaria della media tende ad avvicinarsi alla normale all'aumentare dell'ampiezza del campione.

La Proposizione 17.2 si basa sull'ipotesi di normalità. Tuttavia, se n è sufficientemente grande, la distribuzione campionaria di $\bar{X}$ può essere approssimata con la distribuzione normale in virtù del teorema del limite centrale enunciato nella Proposizione 16.2. Come regola empirica, la numerosità del campione si ritiene sufficientemente grande se $n \geq 30$.

Dunque, se l'ampiezza del campione è sufficientemente grande, la probabilità che $\bar{X}$ assuma valori nell'intervallo (a, b) è approssimativamente pari a

$$P(a < \bar{X} < b) \approx P\left(\frac{a-\mu}{\sigma/\sqrt{n}} < Z_{\bar{X}} < \frac{b-\mu}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{b-\mu}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{a-\mu}{\sigma/\sqrt{n}}\right), \quad (17.5)$$

qualunque sia la forma della popolazione da cui proviene il campione.

ESEMPIO 17.12

Si consideri un campione casuale di dimensione 30 estratto da una popolazione esponenziale di parametro $\lambda = 0.05$. Si trovi la probabilità che $\bar{X}$ sia compresa nell'intervallo $(\mu - 2\sigma/\sqrt{n},\ \mu + 2\sigma/\sqrt{n})$.

Soluzione

La media μ e la varianza σ^2 della popolazione esponenziale sono 20 e 400 (si veda la Proposizione 14.6). Sulla base della (17.5), si ha

$$P\left(20 - 2\sqrt{\frac{400}{30}} < \bar{X} < 20 + 2\sqrt{\frac{400}{30}}\right) \approx \Phi(2) - \Phi(-2) = 0.9545.$$

La probabilità esatta (calcolata usando la distribuzione esatta di $\bar{X}$) è 0.9562, con un errore di approssimazione (in valore assoluto) di 0.0017. L'approssimazione, già molto soddisfacente, migliora al crescere di n. Per esempio se $n = 120$, la probabilità effettiva è 0.9549 e l'errore di approssimazione si riduce a 0.0004. ■

ESEMPIO 17.13

Si consideri una popolazione Bernoulliana di parametro $p = 0.30$. Si determini la probabilità che la proporzione campionaria $\hat{p}$ in campioni casuali di ampiezza 60 assuma un valore compreso tra 11/60 e 23/60, estremi inclusi.

Soluzione

Applicando la (17.5) e la correzione per continuità (si veda il Paragrafo 16.3.1), si ha

$$\begin{aligned} P(11/60 \le \hat{p} \le 23/60) &\approx P\left[\frac{10.5/60 - 0.30}{\sqrt{(0.30 \cdot 0.70)/60}} \le Z_{\hat{p}} \le \frac{23.5/60 - 0.30}{\sqrt{(0.30 \cdot 0.70)/60}}\right] \\ &= \Phi(1.55) - \Phi(-2.55) = 0.9220, \end{aligned}$$

dove $Z_{\hat{p}} = (\hat{p} - 0.30)/\sqrt{(0.30 \cdot 0.70)/60}$.

La probabilità esatta (fornita dalla distribuzione binomiale di parametri $n = 60$ e $p = 0.3$) è 0.9229, un valore assai vicino a quello ottenuto con l'approssimazione normale. ■

17.6 Distribuzione campionaria della varianza

Si consideri di nuovo la varianza campionaria

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Come per ogni statistica, la distribuzione di S^2 dipende dalla struttura della popolazione da cui proviene il campione. Tuttavia, le formule del valore atteso e della varianza di questa statistica si possono ricavare in generale per qualsiasi popolazione.

PROPOSIZIONE 17.3

Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale proveniente da una popolazione qualsiasi. Allora il valore atteso e la varianza di S^2 sono dati da:

$$\mu_{S^2} = \sigma^2, \quad \sigma^2_{S^2} = \frac{\sigma^4}{n}\left(\gamma + 2\frac{n}{n-1}\right), \quad (17.6)$$

dove σ^2 e γ sono la varianza e l'indice di curtosi della popolazione, una misura quest'ultima della discrepanza tra la distribuzione di probabilità della popolazione e la distribuzione normale.[4]

Qualunque sia la popolazione, il valore atteso della varianza campionaria è uguale alla varianza della popolazione.

La dimostrazione della prima delle (17.6) è data nel Paragrafo B.10 dell'Appendice B. ■

17.6.1 Popolazione normale

La distribuzione campionaria della statistica S^2 assume una forma ben definita quando la popolazione generatrice è normale.

PROPOSIZIONE 17.4

Sia S^2 la varianza campionaria di un campione casuale di ampiezza n proveniente da una popolazione normale con media μ e varianza σ^2. Allora la variabile casuale

$$V = \frac{(n-1)S^2}{\sigma^2}$$

ha distribuzione di probabilità chi-quadrato con $n - 1$ gradi di libertà.

Per la dimostrazione si può vedere Hogg *et al.* (2013), pp. 194-195. ■

Nella Figura 17.2 è rappresentata la distribuzione campionaria di S^2 per campioni di dimensione 11 da una popolazione normale con $\mu = 175$ e $\sigma^2 = 32$.

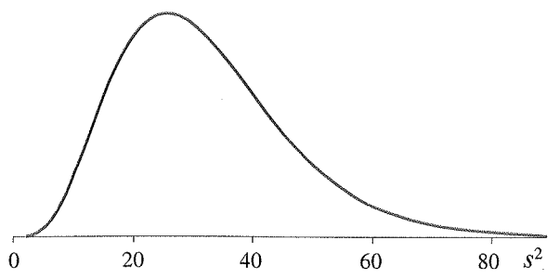


Figura 17.2
Distribuzione campionaria di S^2 per campioni di ampiezza 11 provenienti da una popolazione normale di media 175 e varianza 32.

[4] L'indice di curtosi è dato da

$$\gamma = \frac{\bar{\mu}_4}{\sigma^4} - 3,$$

dove $\bar{\mu}_4$ è il momento quarto centrato della popolazione (si veda il Paragrafo 13.4).

ESEMPIO 17.14

Il peso di una confezione di pasta alimentare ha distribuzione normale con media $\mu = 1$ e varianza $\sigma^2 = 0.0007$. Considerando un campione di 10 confezioni, si determini la probabilità che $0.0005 < S^2 < 0.0008$.

Soluzione

Utilizzando la funzione DISTRIB.CHI.QUAD( ) di Excel, si ottiene

$$\begin{aligned} P(0.0005 < S^2 < 0.0008) &= P\left(\frac{9 \cdot 0.0005}{0.0007} < \frac{9 \cdot S^2}{0.0007} < \frac{9 \cdot 0.0008}{0.0007}\right) \\ &= P(6.43 < V < 10.29) = 0.3687, \end{aligned}$$

in quanto la variabile casuale $V = (9 \cdot S^2)/0.0007$ ha distribuzione chi-quadrato con 9 gradi di libertà. ■

17.7 Distribuzione campionaria della media quando la varianza σ^2 non è nota

Si consideri la variabile casuale

$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}. \tag{17.7}$$

Se il campione proviene da una popolazione normale, questa variabile casuale ha distribuzione normale standard per la Proposizione 17.2. Per campioni casuali provenienti da popolazioni non normali, il rapporto si distribuisce approssimativamente secondo la normale standard, se l'ampiezza del campione è sufficientemente grande, in virtù del teorema del limite centrale.

17.7.1 Popolazione normale

Se si assume che la popolazione di interesse sia normale con varianza non nota, il rapporto che si ottiene dalla (17.7), sostituendo a σ la deviazione standard campionaria S, ha *distribuzione T di Student*, come indicato nella proposizione che segue.

PROPOSIZIONE 17.5

Siano $\bar{X}$ e S^2 la media e la varianza campionarie di un campione casuale di dimensione n proveniente da una popolazione normale con media μ e varianza σ^2 non nota. Allora la variabile casuale

$$T_{\bar{X}} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \tag{17.8}$$

si distribuisce come una T di Student, la cui funzione di densità è data da

$$f(t) = b\left(1 + \frac{t^2}{r}\right)^{-(r+1)/2}, \quad -\infty < t < \infty,$$

dove $r = n - 1$ è un parametro chiamato gradi di libertà e b è un numero positivo che dipende da r.

Per la dimostrazione che $T_{\bar{X}}$ ha distribuzione T di Student con $n - 1$ gradi di libertà, si rinvia a Hogg *et al.* (2013), pp. 189-190. ■

La distribuzione prende questo nome in omaggio allo studioso inglese Student, pseudonimo di William Sealy Gosset (1876-1937).

La distribuzione T di Student ha le seguenti proprietà:

- ha valore atteso 0 e varianza $r/(r-2)$, per $r > 2$;
- è simmetrica;
- tende alla normale standard al tendere di r a infinito.

Nella Figura 17.3 è rappresentata la funzione di densità T di Student per tre diversi valori di r, assieme alla funzione di densità normale standard, individuata con la notazione $r = \infty$.

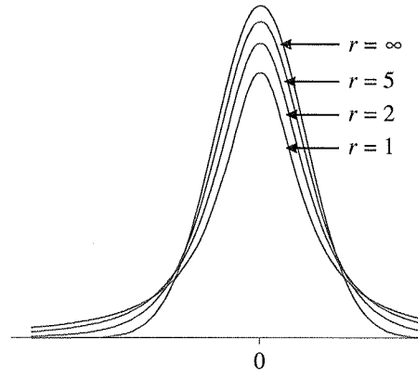


Figura 17.3
Funzione di densità T di Student per alcuni valori di r.

La Figura 17.4 riporta il quantile di livello p, cioè il valore t_p per cui la funzione di ripartizione assume il valore p o, equivalentemente, il punto dell'asse delle ascisse a sinistra del quale l'area sottesa alla densità T di Student è pari a p.

Data la simmetria della distribuzione, vale la relazione

$$t_p = -t_{1-p},$$

che è utile per determinare i quantili (centili) inferiori (primo, secondo, ...) a partire dai quantili superiori corrispondenti (99-esimo, 98-esimo, ...). Nella Tavola C.3 dell'Appendice C, sono riportati i quantili di livello p, per $r = 1, 2, \ldots, 40, 60, 120, \infty$ e $p = 0.60, 0.70, 0.80, 0.90, 0.95, 0.975, 0.99, 0.995, 0.999$.

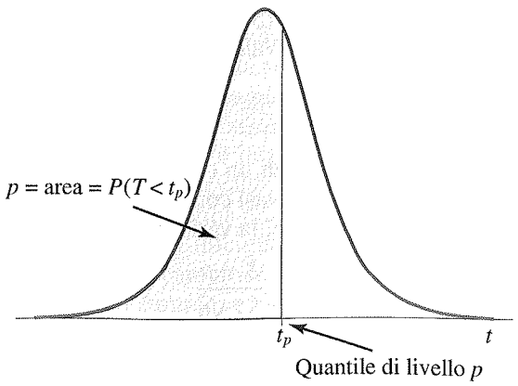


Figura 17.4
Quantile t_p della distribuzione T di Student.

ESEMPIO 17.15

Si consideri la distribuzione T di Student con 12 gradi di libertà. Si determinino il quinto e il 95-esimo centile.

Soluzione

Conviene iniziare con il 95-esimo centile. Come indicato nella Figura 17.5, in corrispondenza di 12 gradi di libertà, si trova $t_{0.95} = 1.782$. Il quinto centile è allora $t_{0.05} = -t_{0.95} = -1.782$.

Figura 17.5
Calcolo di t_p con la Tavola C.3.

	p								
r	0.60	0.70	0.80	0.90	0.95	0.975	0.990	0.995	0.999
...	...	...	...	...	...	...	...	...	...
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012	3.852
...	...	...	...	...	...	...	...	...	...

Si osservi che nella Tavola C.3, per ogni p, il quantile t_p decresce al crescere dei gradi di libertà e che, per la terza proprietà della distribuzione, si avvicina sempre più al quantile di livello p della normale standard, posto nell'ultima riga, quella contrassegnata dal simbolo ∞.

ESEMPIO 17.16

Si consideri un campione casuale di 10 unità proveniente da una popolazione normale con media 45. Si determini la probabilità che la variabile casuale $T_{\bar{X}}$ assuma un valore nell'intervallo $(-2, 2)$.

Soluzione

La probabilità, calcolata usando la funzione DISTRIB.T.N( ) di Excel, è pari a

$$P\left(-2 < \frac{\bar{X} - 45}{S/\sqrt{10}} < 2\right) = 0.9234.$$

17.7.2 Grandi campioni

Si riprenda la variabile casuale $(\bar{X} - \mu)/(S/\sqrt{n})$ e si assuma che la dimensione del campione sia sufficientemente grande. Vale allora la proposizione che segue.

PROPOSIZIONE 17.6

Siano $\bar{X}$ e S^2 la media e la varianza di un campione casuale di dimensione n proveniente da una popolazione qualsiasi con media μ e varianza σ^2 non nota. Allora, se n è sufficientemente grande, la variabile casuale

$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \tag{17.9}$$

si distribuisce approssimativamente secondo la normale standard.

Questa approssimazione può essere intuitivamente spiegata come segue. Se n è sufficientemente grande, la media campionaria ha distribuzione prossima alla normale per il teorema del limite centrale e S è una buona approssimazione di σ. È quindi naturale che le variabili casuali $(\bar{X} - \mu)/(S/\sqrt{n})$ e $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ si comportino in modo simile.

Qualunque sia la forma della popolazione, la distribuzione campionaria di $(\bar{X} - \mu)/(S/\sqrt{n})$ può essere approssimata con la normale standard se n è sufficientemente grande.

Le (17.8) e (17.9), pur essendo formalmente identiche, sono state indicate con $T_{\bar{X}}$ e $Z_{\bar{X}}$ perché sottendono assunzioni molto diverse: la prima presuppone che la popolazione sia normale, la seconda richiede solo che il campione sia sufficientemente grande.

La Proposizione 17.6, valendo per qualsiasi popolazione, può essere impiegata anche quando la popolazione è Bernoulliana. In questo caso, la (17.9) diviene

$$Z_{\hat{p}} = \frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p})/n}}. \tag{17.10}$$

Dunque, se n è sufficientemente grande, la variabile casuale $Z_{\hat{p}}$ ha, approssimativamente, distribuzione normale standard.

Per capire come il denominatore della (17.9) si trasformi nel denominatore della (17.10), si veda l'Esercizio 17.20.

17.8 Stima dei parametri e verifica delle ipotesi

I concetti introdotti nei paragrafi precedenti costituiscono la base dell'inferenza statistica, la cui finalità è quella di fare luce su uno o più parametri della popolazione di interesse (media, proporzione, varianza ecc.), a partire dai dati di un campione casuale.

I metodi dell'inferenza statistica appartengono a due aree distinte:

- la stima dei parametri;
- la verifica delle ipotesi.

L'obiettivo della stima dei parametri è quello di assegnare al parametro incognito un valore o un insieme di valori minimizzando, in qualche modo, il rischio di errore.

Nella verifica delle ipotesi, i dati del campione vengono utilizzati per decidere se rifiutare o non rifiutare un'ipotesi sul parametro di interesse minimizzando, in qualche modo, la probabilità di errore.

Sia la stima dei parametri sia la verifica delle ipotesi, direttamente o indirettamente, utilizzano le statistiche e le loro distribuzioni campionarie.

La determinazione della distribuzione di probabilità delle statistiche campionarie è una questione di teoria della probabilità: quest'ultima consente di trasformare le conoscenze sulla struttura della popolazione in previsioni probabilistiche sui valori che la statistica campionaria potrà assumere in un campione casuale di una certa dimensione. Per esempio, se si sa che la popolazione è normale e se ne conoscono media e varianza, si è in grado di dire che la media campionaria assumerà un valore in un dato intervallo (a, b) con una certa probabilità. Questo processo va sotto il nome di problema diretto. Si parla invece di problema inverso con riferimento alle tecniche dell'inferenza statistica.

Il percorso che porta dal campione alla popolazione può essere così schematizzato: si dispone di un campione casuale proveniente dalla popolazione di interesse; su di esso si calcola il valore di una appropriata statistica campionaria; questo valore viene "inquadrato" all'interno della distribuzione di probabilità

della statistica campionaria; ciò consente di rifiutare o non rifiutare un'ipotesi formulata sul parametro di interesse oppure di assegnare a questo un valore o un insieme di valori plausibili.

Nella Figura 17.6 sono raffigurati il percorso diretto e quello inverso: il primo va dal modello distributivo della popolazione alla distribuzione di probabilità della statistica campionaria, legata anch'essa ai possibili valori del parametro incognito. Il valore che la statistica campionaria assume nel campione è un punto sull'asse delle ascisse della curva relativa alla distribuzione di probabilità della statistica campionaria. Da questo punto prende il via il percorso inverso che va verso il modello probabilistico della popolazione e quindi verso la popolazione.

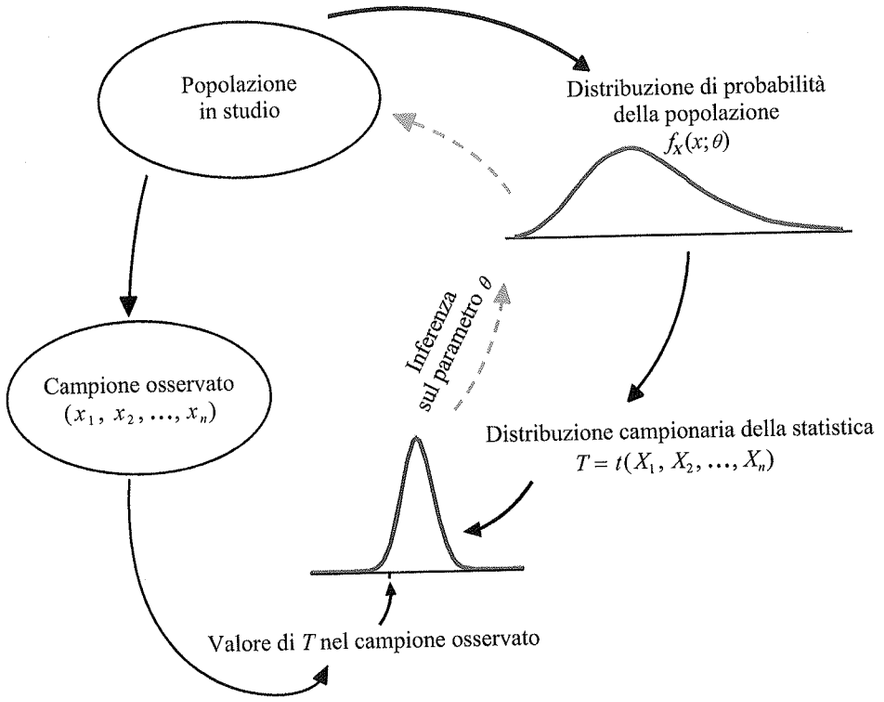


Figura 17.6
Problema diretto e problema inverso.

Riassunto

Questo capitolo costituisce l'anello di congiunzione tra teoria della probabilità e statistica inferenziale. La parte iniziale del capitolo ruota attorno ai concetti di: campione casuale, statistica campionaria, distribuzione di probabilità di una statistica campionaria. Un campione casuale è definito come una collezione di variabili casuali indipendenti, tutte distribuite come la popolazione da cui proviene il campione. Una statistica campionaria è una funzione del campione casuale e come tale è una variabile casuale che ha una distribuzione di probabilità denominata distribuzione campionaria.

La seconda parte del capitolo è dedicata alla distribuzione campionaria di due statistiche particolarmente importanti nell'economia del testo, la media e la varianza campionaria.

Viene anche illustrato che nel caso di normalità della popolazione la variabile casuale

$$T_{\bar{X}} = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

segue la distribuzione T di Student con $n - 1$ gradi di libertà.

Parole chiave

campione casuale, 359
campione casuale osservabile, 360
distribuzione campionaria, 362
distribuzione di probabilità del campione casuale, 360
distribuzione di probabilità della popolazione, 359
distribuzione T di Student, 370
gradi di libertà, 370
parametro, 359
popolazione Bernoulliana, 360
popolazione esponenziale, 360
popolazione generatrice, 360
popolazione infinita, 359
popolazione normale, 360
problema diretto, 373
problema inverso, 373
proporzione campionaria, 364
statistica campionaria, 362

Commenti e curiosità

1. Simulazione delle distribuzioni campionarie con il metodo Monte Carlo

Le statistiche campionarie sono funzioni del campione e sono definite nell'insieme di tutti i possibili campioni di dimensione n. Se è nota la distribuzione della popolazione, è possibile estrarre mediante il calcolatore elettronico un numero molto elevato di campioni, ad esempio 2000, di ampiezza prefissata, e calcolare per ognuno di essi la statistica di interesse. La distribuzione di frequenza relativa dei 2000 valori ottenuti è un'approssimazione della distribuzione campionaria della statistica. Ovviamente quanto più numerosi sono i campioni estratti tanto migliore è l'approssimazione che si ottiene.

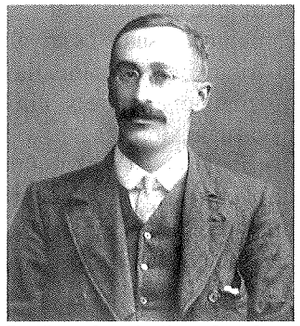
Nella Tabella 17.9 sono riportati i risultati della simulazione della distribuzione della media campionaria basata sull'estrazione, da una popolazione normale con media 175 e varianza 44, di 1000 campioni di ampiezza 15. Come si può osservare, la distribuzione di frequenza relativa ottenuta raggruppando in classi i valori delle 1000 medie campionarie simulate è molto simile alla vera distribuzione campionaria. Ciò è in accordo con le attese basate sulla legge dei grandi numeri, per cui la frequenza relativa con la quale la media campionaria simulata ricade in una certa classe tende a essere vicina alla probabilità di questo evento.

Distribuzione simulata		Distribuzione campionaria esatta	
169-170	0.001	169-170	0.001
170-171	0.010	170-171	0.007
171-172	0.032	171-172	0.028
172-173	0.087	172-173	0.080
173-174	0.171	173-174	0.159
174-175	0.212	174-175	0.225
175-176	0.209	175-176	0.225
176-177	0.161	176-177	0.159
177-178	0.073	177-178	0.080
178-179	0.033	178-179	0.028
179-180	0.011	179-180	0.007
180-181	0.000	180-181	0.001
Totale	1.000	Totale	1.000

Tabella 17.9
Distribuzione di frequenza simulata della media campionaria in 1000 campioni Monte Carlo di ampiezza 15 estratti dalla popolazione normale N(175, 44), e distribuzione campionaria esatta.

Con lo stesso procedimento, si può ottenere la distribuzione campionaria simulata per ogni altra statistica $T = t(X_1, X_2, \ldots, X_n)$, come la mediana campionaria, la varianza campionaria e così via.

William Sealy Gosset, storia di uno pseudonimo


The History Collection/Alamy Stock Photo

William Sealy Gosset (1876–1937), al termine degli studi universitari in chiica e matematica, venne assunto presso la birreria Guinness di Dublino, ove ebbe modo di applicare le sue competenze statistiche alla produzione ella birra. In questa attività di statistico, si rese conto che le condizioni con quali venivano raccolti i dati mutavano di continuo, cosicché la disponibili- di pochi dati rilevati nelle stesse condizioni sperimentali non consentivano applicazione del teorema del limite centrale. Grazie anche a un rapporto di micizia con Karl Pearson, trascorse un anno presso il laboratorio di biomeia dell'University College di Londra, diretto dallo stesso Pearson. Questi udi, unitamente all'esperienza empirica maturata nella sua attività lavorati-a, gli consentirono di pubblicare sulla rivista Biometrika, sotto lo pseudo-mo di Student, l'articolo dal titolo *The probable error of the mean*, in cui trodusse la distribuzione denominata T di Student.

Gosset dovette usare uno pseudonimo perché la fabbrica Guinness vietava la pubblicazione di articoli a parte dei suoi dipendenti per evitare la divulgazione di segreti di produzione. I suoi studi influenzarono i vori di molti statistici, tra cui, per esempio, Egon Pearson (figlio di Karl Pearson) e Jerzy Neyman. Gosset oncluse la sua carriera presso una nuova birreria Guinness a Londra, dove venne nominato direttore e sponsabile del settore scientifico.

Esercizi

17.1 Indicare quale delle seguenti quantità è un parametro e quale una statistica:

a) mediana della popolazione;
b) terzo quartile del campione;
c) media geometrica della popolazione;
d) valore centrale della popolazione;
e) deviazione standard del campione.

17.2 In una università viene condotta un'indagine campionaria per stimare la proporzione di studenti che vivono in appartamento durante il corso. Qual è la popolazione di riferimento e qual è il parametro di interesse?

17.3 Un'urna contiene cinque palline numerate da 1 a 5. Si immagini di estrarre a sorte due palline dall'urna con ripetizione.

a) Si elenchino tutti i possibili campioni.
b) Si determini la distribuzione campionaria della media.

17.4 Si consideri una popolazione in cui la variabile X ha la seguente distribuzione:

x	2	5	7
$f(x)$	0.3	0.6	0.1

a) Si elenchino tutti i campioni di ampiezza 3 che si possono estrarre con ripetizione da tale popolazione e si assegni a ciascun campione la relativa probabilità.
b) Si determini la distribuzione campionaria della media.
c) Si determinino il valore atteso e la varianza della distribuzione di cui al punto precedente e si verifichi che la prima quantità è uguale alla media della popolazione.
d) Si determini la distribuzione campionaria della mediana.

17.5 Si verifichi che, nel caso della popolazione Bernoulliana, il valore atteso e la varianza della media campionaria (proporzione campionaria) sono dati da p e $p(1-p)/n$.

17.6 Si stabilisca come varia la deviazione standard della media campionaria se la dimensione del campione viene:

a) diminuita da 280 a 70;
b) aumentata da 100 a 400;
c) aumentata da 50 a 450.

17.7 Da un'urna con cinque palline numerate 1, 3, 5, 7, 9 si estrae, con ripetizione, un campione di ampiezza 25. Si determinino il valore atteso e la deviazione standard della media campionaria.

17.8 Si consideri la media campionaria per campioni provenienti da una popolazione con media 175 e varianza 42. Applicando la disuguaglianza di Chebyshev, si determini:

a) il limite inferiore della probabilità che la media campionaria sia compresa tra 170 e 180 assumendo un'ampiezza campionaria di 50;
b) la dimensione campionaria tale che sia soddisfatta la disuguaglianza:

$$P(174 < \bar{X} < 176) \geq 0.90.$$

17.9 Si consideri un campione casuale di ampiezza 25 proveniente da una popolazione normale con media 120 e varianza 16. Si calcoli:

a) la probabilità che la media campionaria assuma un valore maggiore di 122;
b) la probabilità che la media campionaria sia compresa tra 112 e 119;
c) il valore superato dal 70% delle medie campionarie.

17.10 L'altezza degli alberi di una foresta ha distribuzione normale con media 11.4 m e deviazione standard 1.3 m. Con riferimento a un campione casuale di 20 unità:

a) si calcoli la probabilità che la media campionaria sia compresa tra 9.7 e 11;
b) si determinino gli estremi dell'intervallo centrato nella media della popolazione entro cui è compresa la media campionaria con probabilità 0.80.

17.11 Il tempo impiegato da un meccanico per assemblare un certo tipo di bicicletta può essere considerato una variabile casuale normale con media 32 minuti e deviazione standard 3.5 minuti. Si trovi la probabilità che il tempo medio impiegato per assemblare 10 biciclette:

a) non superi 33 minuti;
b) sia compreso tra 28.5 e 31.5 minuti.

17.12 Il 35% degli studenti di una data università affitta un appartamento durante l'anno accademico. Si determini la probabilità che in un campione casuale di 200 studenti la percentuale di coloro che affittano un appartamento sia:

a) minore di 37%;
b) compresa tra 33% e 40%.

17.13 Gli iscritti di un'università sono 35 740. Di questi, 4310 hanno più di 26 anni. Si determini, utilizzando l'approssimazione normale, la probabilità che in un campione casuale di 500 studenti la frazione di coloro che hanno più di 26 anni sia compresa tra 0.10 e 0.12.

17.14 Una compagnia aerea sostiene che il 98% dei propri voli arriva in orario. Sia $\hat{p}$ la frazione dei voli in orario su 50 voli scelti a caso. Si determini la probabilità che la frazione di voli in orario sia maggiore o uguale a 48/50 con l'approssimazione normale e con la distribuzione binomiale. Si confrontino poi i risultati.

17.15 Sia $\bar{X}$ la media aritmetica di un campione casuale di 100 unità estratto da una popolazione con media μ incognita e varianza $\sigma^2 = 9$. Si determini il numero a per cui

$$P(|\bar{X} - \mu| < a) = 0.90.$$

(*Suggerimento*: si utilizzi l'approssimazione normale.)

17.16 Sia $(X_1, X_2, \ldots, X_{80})$ un campione casuale proveniente da una popolazione con distribuzione descritta dalla funzione di densità $f(x) = 3x^2$, $0 < x < 1$. Si determini la probabilità che la media campionaria sia minore di 0.8. (*Suggerimento*: si applichi il teorema del limite centrale.)

17.17 Con riferimento all'Esercizio 17.3, si determini la distribuzione campionaria della varianza S^2 e sulla base di essa si verifichi che $\mathrm{E}(S^2) = \sigma^2$.

17.18 Per un campione casuale di ampiezza 20 da una popolazione normale con varianza 14, si trovi (usando la funzione DISTRIB.CHI.QUAD( ) di Excel) la probabilità che la varianza campionaria:

a) assuma un valore maggiore di 14;

b) sia compresa nell'intervallo (18, 21).

17.19 Una variabile casuale X ha distribuzione T di Student con 15 gradi di libertà. Si determinino, usando un qualche *software* statistico:

a) la probabilità che la variabile casuale assuma un valore esterno all'intervallo (0.6, 1.7);

b) il numero c tale che sia soddisfatta la relazione

$$P(-1.4 < X < c) = 0.60.$$

17.20* Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale proveniente da una popolazione Bernoulliana. Si dimostri che sussiste l'identità

$$S^2 = \frac{n}{n-1}\hat{p}(1-\hat{p}).$$

(*Suggerimento*: si scriva la varianza campionaria come

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \frac{n}{n-1}\bar{X}^2.$$

Poi, si sostituisca $\bar{X}$ con $\hat{p}$ e si consideri che $\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i = n\hat{p}$.)

Distribuzione T di Student con Excel

La funzione di Excel relativa alla distribuzione T di Student è DISTRIB.T.N( ). La funzione richiede come argomenti il valore di T, i gradi di libertà e "vero" o "falso" a seconda che si voglia la funzione di ripartizione o quella di densità.

La Figura 17.7 illustra il calcolo della funzione di ripartizione di una variabile casuale T di Student con 20 gradi di libertà nel punto $T = 2.4$.

Figura 17.7
Funzione DISTRIB.T.N( ).

C3 fx =DISTRIB.T.N(B3;20;VERO)

	A	B	C	D	E	F	G
1		Distribuzione T di Student					
2							
3		2.4	0.98687556				
4							

quantili della distribuzione T di Student possono essere calcolati utilizzando la funzione INV.T.2T( ), che presenta come argomenti il livello di probabilità e il numero di gradi di libertà. Per determinare il quantile di livello $(1 - \alpha)$, è necessario inserire come probabilità la quantità 2α, che è la somma delle aree sotto la curva di densità a sinistra di $-t_{1-\alpha}$ e alla destra di $t_{1-\alpha}$.

Nella Figura 17.8, la funzione INV.T.2T( ) viene applicata per trovare il 95-esimo centile della distribuzione T di Student con 20 gradi di libertà.

C3 fx =INV.T.2T(B3;20)

	A	B	C	D	E	F	G
1		Distribuzione T di Student					
2							
3		0.1	1.72471824				
4							

Figura 17.8
Funzione INV.T.2T( ).

PARTE

III

Inferenza statistica

Capitolo 18
Stima puntuale

Capitolo 19
Stima intervallare

Capitolo 20
Verifica delle ipotesi

Capitolo 21
Confronti tra due popolazioni

Capitolo 22
Verifica di ipotesi per dati di frequenza

Capitolo 23
Inferenza sul modello di regressione lineare

Capitolo 24
Inferenza su popolazioni finite

Capitolo 25
Regressione lineare multipla

Capitolo 26
Analisi della varianza

CAPITOLO

18

Stima puntuale

Sommario del capitolo

18.1 Introduzione

18.2 Proprietà degli stimatori

18.3 Ricerca dello stimatore ottimo

18.4 Margine di errore nella stima della media

Riassunto

Commenti e curiosità

Esercizi

18.1 Introduzione

In questo capitolo, verrà affrontato il problema di come assegnare un valore plausibile a un parametro (come la media, la varianza ecc.) della popolazione di interesse, sulla base dei dati di un campione casuale. L'espressione "stima puntuale" allude al fatto che lo scopo è di assegnare al parametro incognito un singolo valore, non già un insieme di valori come nella stima per intervallo che sarà l'oggetto del Capitolo 19.

Si debba, ad esempio, stimare il reddito medio delle famiglie di una grande città con i dati osservati in un campione casuale di 3000 famiglie, $(x_1, x_2, \ldots, x_{3000})$. Si supponga che il campione osservato, $(50.4, 36.2, \ldots, 43.8)$ (dati in migliaia di euro), abbia media $\bar{x} = 42.7$ e mediana $m = 38.3$. Una stima intuitivamente plausibile del reddito medio della popolazione è la media campionaria, 42.7. Una stima alternativa potrebbe essere la mediana campionaria, 38.3. Chiaramente, non sono precluse altre soluzioni.

DEFINIZIONE 18.1

Sia θ il parametro da stimare e sia $(X_1, X_2, \ldots, X_n)$ un campione casuale proveniente dalla popolazione di interesse. Si chiama stimatore del parametro θ la statistica, $T = t(X_1, X_2, \ldots, X_n)$, usata per stimare il parametro. Si chiama stima lo specifico valore, t, assunto dallo stimatore T nel campione osservato $(x_1, x_2, \ldots, x_n)$.

La questione centrale della stima puntuale è quella di definire dei criteri con cui valutare la "qualità" di differenti stimatori. Intuitivamente, un buono stimatore dovrebbe essere in grado di produrre stime "vicine" al valore effettivo di θ. Naturalmente, la vicinanza al valore del parametro può essere intesa ed espressa in modi diversi, a cui corrispondono altrettante proprietà desiderabili per uno stimatore.

Queste proprietà saranno illustrate nel Paragrafo 18.2, mentre nel Paragrafo 18.3 verranno presentati gli elementi principali di teoria della stima che consentono di costruire stimatori con buone proprietà.

Sebbene le formulazioni siano generali, nelle applicazioni e negli esempi, l'attenzione sarà focalizzata sulla stima dei parametri media e varianza. In particolare, il Paragrafo 18.4 sarà dedicato alla quantificazione del rischio di errore connesso all'uso della media campionaria per la stima del parametro media.

18.2 Proprietà degli stimatori

Qui di seguito, vengono definiti i criteri basilari che consentono di comparare più stimatori al fine della scelta della migliore soluzione del problema di stima.

18.2.1 Non distorsione

Conviene partire da un semplice esempio con cui viene descritto il comportamento di due stimatori alternativi della media, nell'insieme di tutti i possibili campioni di ampiezza 3 provenienti da una piccola popolazione artificiale (le piccolissime dimensioni del campione e della popolazione consentono di elencare tutti i campioni possibili).

ESEMPIO 18.1

La Tabella 18.1 contiene tutti i campioni di ampiezza $n = 3$ che possono essere estratti, con ripetizione, dalla popolazione di 4 unità, A, B, C e D, in cui la variabile d'interesse assume i valori 1, 2, 3 e 4. La media della popolazione è $\mu = 2.5$. Nella tabella, accanto a ogni campione, sono riportate: la media campionaria $\bar{x}$, la differenza $\bar{x} - \mu$, la mediana campionaria, m, e la differenza $m - \mu$. Le quantità $\bar{x} - \mu$ e $m - \mu$ sono gli errori di stima che si compiono stimando la media della popolazione con la media campionaria e con la mediana campionaria.

Discussione

Come si può osservare, gli errori di stima sono sia negativi sia positivi. I valori negativi indicano che lo stimatore sottostima μ; i valori positivi indicano il fenomeno opposto. Si prendano ora i valori attesi delle differenze $\bar{X} - \mu$ e $M - \mu$:

$$\mathrm{E}(\bar{X} - \mu) = \frac{1}{64}(-1.50 - 1.17 + \cdots + 1.50) = 0,$$

$$\mathrm{E}(M - \mu) = \frac{1}{64}(-1.50 - 1.50 + \cdots + 1.50) = 0,$$

dove M indica la variabile casuale mediana campionaria. Se ne deduce che, in entrambi i casi, l'errore di stima è mediamente nullo, essendoci un bilanciamento tra errori di stima positivi e negativi. Questo dato è interessante, perché un risultato negativo (positivo) avrebbe implicato una sottostima (sovrastima) in media di μ. Stimatori con questa proprietà sono detti *non distorti*. ■

Si consideri il generico stimatore $T = t(X_1, X_2, \ldots, X_n)$ del parametro θ. Sia

$$T - \theta$$

la variabile casuale che descrive l'errore di stima che si commette stimando θ con il valore fornito dallo stimatore T. Il valore atteso di questa differenza,

$$\mathrm{E}(T - \theta),$$

può essere negativo, positivo o nullo. Nel primo caso, T sottostima in media θ. Nel secondo, T sovrastima in media θ. È allora desiderabile che sia

$$\mathrm{E}(T - \theta) = 0,$$

ossia che valga la relazione

$$\mathrm{E}(T) = \theta.$$

Campione	$\bar{x}$	$\bar{x} - \mu$	m	$m - \mu$	Campione	$\bar{x}$	$\bar{x} - \mu$	m	$m - \mu$
1, 1, 1	1.00	−1.50	1.00	−1.50	1, 3, 4	2.67	0.17	3.00	0.50
1, 1, 2	1.33	−1.17	1.00	−1.50	1, 4, 3	2.67	0.17	3.00	0.50
1, 2, 1	1.33	−1.17	1.00	−1.50	3, 1, 4	2.67	0.17	3.00	0.50
2, 1, 1	1.33	−1.17	1.00	−1.50	3, 4, 1	2.67	0.17	3.00	0.50
1, 2, 2	1.67	−0.83	2.00	−0.50	4, 1, 3	2.67	0.17	3.00	0.50
2, 1, 2	1.67	−0.83	2.00	−0.50	4, 3, 1	2.67	0.17	3.00	0.50
2, 2, 1	1.67	−0.83	2.00	−0.50	2, 2, 4	2.67	0.17	2.00	−0.50
1, 1, 3	1.67	−0.83	1.00	−1.50	2, 4, 2	2.67	0.17	2.00	−0.50
1, 3, 1	1.67	−0.83	1.00	−1.50	4, 2, 2	2.67	0.17	2.00	−0.50
3, 1, 1	1.67	−0.83	1.00	−1.50	2, 3, 3	2.67	0.17	3.00	0.50
1, 2, 3	2.00	−0.50	2.00	−0.50	3, 2, 3	2.67	0.17	3.00	0.50
1, 3, 2	2.00	−0.50	2.00	−0.50	3, 3, 2	2.67	0.17	3.00	0.50
2, 1, 3	2.00	−0.50	2.00	−0.50	2, 3, 4	3.00	0.50	3.00	0.50
2, 3, 1	2.00	−0.50	2.00	−0.50	2, 4, 3	3.00	0.50	3.00	0.50
3, 1, 2	2.00	−0.50	2.00	−0.50	3, 2, 4	3.00	0.50	3.00	0.50
3, 2, 1	2.00	−0.50	2.00	−0.50	3, 4, 2	3.00	0.50	3.00	0.50
1, 1, 4	2.00	−0.50	1.00	−1.50	4, 2, 3	3.00	0.50	3.00	0.50
1, 4, 1	2.00	−0.50	1.00	−1.50	4, 3, 2	3.00	0.50	3.00	0.50
4, 1, 1	2.00	−0.50	1.00	−1.50	1, 4, 4	3.00	0.50	4.00	1.50
2, 2, 2	2.00	−0.50	2.00	−0.50	4, 1, 4	3.00	0.50	4.00	1.50
1, 2, 4	2.33	−0.17	2.00	−0.50	4, 4, 1	3.00	0.50	4.00	1.50
1, 4, 2	2.33	−0.17	2.00	−0.50	3, 3, 3	3.00	0.50	3.00	0.50
4, 1, 2	2.33	−0.17	2.00	−0.50	2, 4, 4	3.33	0.83	4.00	1.50
4, 2, 1	2.33	−0.17	2.00	−0.50	4, 2, 4	3.33	0.83	4.00	1.50
2, 1, 4	2.33	−0.17	2.00	−0.50	4, 4, 2	3.33	0.83	4.00	1.50
2, 4, 1	2.33	−0.17	2.00	−0.50	3, 3, 4	3.33	0.83	3.00	0.50
1, 3, 3	2.33	−0.17	3.00	0.50	3, 4, 3	3.33	0.83	3.00	0.50
3, 3, 1	2.33	−0.17	3.00	0.50	4, 3, 3	3.33	0.83	3.00	0.50
3, 1, 3	2.33	−0.17	3.00	0.50	3, 4, 4	3.67	1.17	4.00	1.50
2, 2, 3	2.33	−0.17	2.00	−0.50	4, 3, 4	3.67	1.17	4.00	1.50
2, 3, 2	2.33	−0.17	2.00	−0.50	4, 4, 3	3.67	1.17	4.00	1.50
3, 2, 2	2.33	−0.17	2.00	−0.50	4, 4, 4	4.00	1.50	4.00	1.50

Tabella 18.1
Campioni di ampiezza 3 estraibili dalla popolazione data e corrispondenti medie, mediane ed errori di stima.

DEFINIZIONE 18.2

Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale proveniente da una popolazione di cui interessa stimare il parametro θ. Uno stimatore $T = t(X_1, X_2, \ldots, X_n)$ di θ è detto non distorto se

$$\mathrm{E}(T) = \theta,$$

per ogni possibile valore di θ.

Sono esempi di stimatori non distorti la media campionaria $\bar{X}$, la proporzione campionaria $\hat{p}$ e la varianza campionaria S^2. Infatti, i loro valori attesi sono, rispettivamente, μ, p e σ^2.

Gli stimatori che non godono della proprietà appena definita si dicono distorti. La differenza $\mathrm{D}(T) = \mathrm{E}(T) - \theta$ è chiamata distorsione e può essere negativa o positiva: nel primo caso, lo stimatore sottostima mediamente il parametro, nel secondo, lo sovrastima mediamente. La distorsione è nulla se lo stimatore è non distorto.

Un esempio di stimatore distorto è dato dalla varianza del campione calcolata con la formula della statistica descrittiva:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

ESEMPIO 18.2

Si verifichi che $\hat{\sigma}^2$ è uno stimatore distorto della varianza della popolazione.

Soluzione

È chiaro che tra $\hat{\sigma}^2$ e S^2 intercorre la relazione

$$\hat{\sigma}^2 = \frac{n-1}{n}S^2.$$

Ne segue che

$$\mathrm{E}(\hat{\sigma}^2) = \mathrm{E}\left(\frac{n-1}{n}S^2\right) = \frac{n-1}{n}\mathrm{E}(S^2) = \frac{n-1}{n}\sigma^2.$$

La distorsione dello stimatore è data da

$$\mathrm{D}(\hat{\sigma}^2) = \mathrm{E}(\hat{\sigma}^2) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

Dunque, lo stimatore presenta una distorsione negativa: le stime di σ^2 che esso produce sono in media al di sotto di σ^2 per la quantità σ^2/n. ■

ESEMPIO 18.3

Dato un campione casuale di ampiezza 5 proveniente da una popolazione di cui interessa stimare la media μ, si verifichi se

$$T = 0.1X_1 + 0.2X_2 + 0.2X_3 + 0.2X_4 + 0.2X_5$$

è uno stimatore non distorto.

Soluzione

Per la (15.5), il valore atteso dello stimatore è

$$\mathrm{E}(T) = 0.1\mu + 0.2\mu + 0.2\mu + 0.2\mu + 0.2\mu = 0.9\mu,$$

da cui segue che lo stimatore è distorto negativamente; la distorsione è pari a

$$\mathrm{D}(T) = \mathrm{E}(T) - \mu = 0.9\mu - \mu = -0.1\mu.$$

■

18.2.2 Efficienza

Si introduce ora la nozione di efficienza con l'ausilio dei dati dell'Esempio 18.1.

ESEMPIO 18.4

I due stimatori considerati nell'Esempio 18.1 sono entrambi non distorti. È naturale chiedersi quale dei due produca stime mediamente più vicine al valore di μ. Per rispondere al quesito, si può ricorrere al valore atteso dell'errore di stima al quadrato:

$$\mathrm{E}(\bar{X} - \mu)^2 = \frac{1}{64}[(-1.50)^2 + (-1.17)^2 + \cdots + 1.50^2] = 0.42,$$

$$\mathrm{E}(M - \mu)^2 = \frac{1}{64}[(-1.50)^2 + (-1.50)^2 + \cdots + 1.50^2] = 0.88.$$

La media campionaria si comporta meglio della mediana campionaria, perché dà stime in media più vicine a μ di quanto non faccia la mediana. ■

In generale, una misura della vicinanza delle stime al valore di θ è espressa da

$$\mathrm{E}|T - \theta|^r,$$

dove i valori più frequentemente assegnati a r sono 1 e 2. Più spesso si pone $r = 2$, come si fa qui, per ragioni di trattabilità analitica, ottenendo l'*errore quadratico medio*.

DEFINIZIONE 18.3

Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale proveniente da una popolazione di cui interessa stimare il parametro θ. L'errore quadratico medio dello stimatore $T = t(X_1, X_2, \ldots, X_n)$ è definito come

$$\mathrm{MSE}(T) = \mathrm{E}(T - \theta)^2.$$

Chiaramente, l'errore quadratico medio è una misura della qualità dello stimatore: quanto più esso è piccolo, tanto migliore è lo stimatore.

L'errore quadratico medio di T può essere interpretato come la media dei quadrati degli errori di stima che si osserverebbero ripetendo la stima di θ con T molte e molte volte su campioni casuali indipendenti della stessa ampiezza estratti dalla popolazione oggetto di studio.

PROPOSIZIONE 18.1

L'errore quadratico medio di uno stimatore T di θ può essere scritto come

$$\mathrm{MSE}(T) = \mathrm{Var}(T) + [\mathrm{D}(T)]^2, \tag{18.1}$$

dove $\mathrm{Var}(T) = \mathrm{E}[T - \mathrm{E}(T)]^2$ è la varianza di T.

La dimostrazione della (18.1) è delineata nell'Esercizio 18.6*. ■

È di immediata evidenza che l'errore quadratico medio di uno stimatore non distorto coincide con la sua varianza:

$$\mathrm{MSE}(T) = \mathrm{Var}(T) \iff \mathrm{E}(T) = \theta.$$

Dalla (18.1) si evince che la qualità di uno stimatore è determinata dalla sua varianza e dall'eventuale sua distorsione: può capitare che uno stimatore distorto abbia varianza più piccola che non uno stimatore non distorto e, nel contempo, un errore quadratico medio superiore.

ESEMPIO 18.5

Si verifichi che lo stimatore considerato nell'Esempio 18.3 ha una varianza inferiore a quella della media campionaria.

Soluzione

In forza della (15.7), la varianza dello stimatore è data da

$$\text{Var}(T) = 0.1^2\sigma^2 + 0.2^2\sigma^2 + 0.2^2\sigma^2 + 0.2^2\sigma^2 + 0.2^2\sigma^2 = 0.17\sigma^2.$$

La varianza della media campionaria è data da

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{5} = 0.2\sigma^2.$$

Come si vede, T ha una varianza inferiore a quella di $\bar{X}$; tuttavia, il suo errore quadratico medio,

$$\text{MSE}(T) = 0.17\sigma^2 + [-0.1\mu]^2,$$

è superiore a quello di $\bar{X}$, pari a $0.2\sigma^2$, se $0.01\mu^2 > 0.03\sigma^2$. ■

La Tabella 18.2 riporta l'errore quadratico medio degli stimatori $\bar{X}$, $\hat{p}$ e S^2 (già esaminati nel Capitolo 17) che verranno largamente usati nei prossimi capitoli per l'inferenza sui parametri μ, p e σ^2.

Tabella 18.2
Errore quadratico medio di $\bar{X}$, $\hat{p}$ e S^2.

Parametro	Stimatore	Errore quadratico medio
μ	$\bar{X}$	$\text{MSE}(\bar{X}) = \text{Var}(\bar{X}) = \dfrac{\sigma^2}{n}$
p	$\hat{p}$	$\text{MSE}(\hat{p}) = \text{Var}(\hat{p}) = \dfrac{p(1-p)}{n}$
σ^2	S^2	$\text{MSE}(S^2) = \text{Var}(S^2) = \dfrac{\sigma^4}{n}\left(\gamma + 2\dfrac{n}{n-1}\right)$

Come indicato in precedenza, questi tre stimatori sono non distorti e quindi per essi vi è coincidenza tra errore quadratico medio e varianza. Va osservato inoltre che l'errore quadratico medio di $\bar{X}$ non dipende da μ, diversamente da $\text{MSE}(\hat{p})$ e $\text{MSE}(S^2)$ che sono funzioni di p e di σ^2, rispettivamente.

ESEMPIO 18.6

In un campione di 150 studenti del secondo anno di un corso di laurea in economia aziendale, il voto medio riportato all'esame di matematica generale è risultato 25.4. Si stimi la media della popolazione e si calcoli la deviazione standard delle stime, sapendo che la deviazione standard della popolazione è $\sigma = 2.6$.

Soluzione

La media della popolazione viene stimata attribuendo a essa il valore assunto dalla media campionaria; la stima è dunque 25.4. La deviazione standard delle stime è data da

$$\sigma_{\bar{X}} = \frac{2.6}{\sqrt{150}} = 0.21.$$

Questo numero dà un'indicazione del grado di precisione dello stimatore nella specifica situazione (cioè con l'ampiezza campionaria e la variabilità della popolazione assegnate): ciò non significa che la stima ottenuta presenta un errore pari a 0.21, ma che sarebbe 0.21 l'errore medio di stima se si ripetesse il campionamento molte e molte volte nelle stesse condizioni (con campioni della stessa ampiezza estratti dalla stessa popolazione). ■

ESEMPIO 18.7

In un sondaggio condotto sulle persone in età 60 anni e più, 195 intervistati di un campione casuale di ampiezza 300 dichiarano di essere in difficoltà nell'eseguire una transazione con il computer. Si stimi la proporzione nella popolazione delle persone che si trovano in questa condizione.

Soluzione

La proporzione p della popolazione è qui stimata con la proporzione campionaria

$$\hat{p} = \frac{195}{300} = 0.65.$$

L'errore medio che si commette usando $\hat{p}$, in questa specifica situazione, è fornito dalla deviazione standard dello stimatore, $\sqrt{\text{Var}(\hat{p})} = \sqrt{p(1-p)/300}$, una quantità che dipende dall'ignota proporzione della popolazione. Poiché $p(1-p)$ raggiunge il suo massimo per $p = 0.5$ (si osservi la Figura 18.1, in cui è rappresentato l'errore quadratico medio di $\hat{p}$ per $n = 300$ al variare di p tra 0 e 1), vale la disuguaglianza $\sqrt{p(1-p)/300} \leq \sqrt{(0.5 \cdot 0.5)/300} = 0.0289$. In definitiva, con un campione di ampiezza 300, l'errore medio che si commette stimando p con $\hat{p}$ non supera 0.0289.

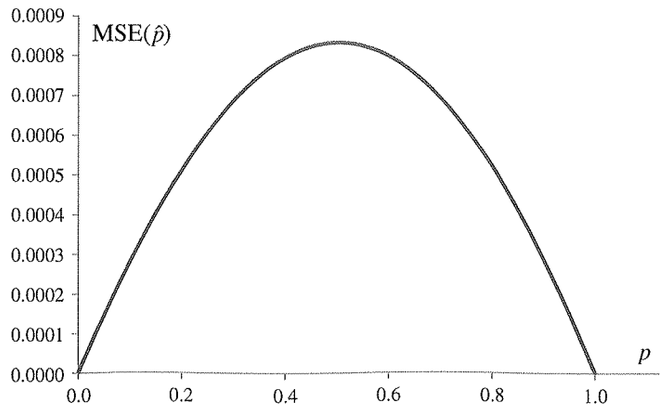


Figura 18.1
Errore quadratico medio di $\hat{p}$ in funzione di p per $n = 300$.

■

ESEMPIO 18.8

Una casa farmaceutica produce compresse con un determinato principio attivo. L'ufficio qualità dell'azienda intende stimare la varianza del principio attivo. A questo fine, viene estratto un campione di 20 compresse in cui si osserva il contenuto di principio attivo (in mg):

2.85, 2.80, 2.94, 3.20, 2.84, 2.85, 2.93, 3.10, 3.16, 3.03,
3.15, 2.85, 2.82, 2.88, 2.82, 3.12, 2.94, 3.03, 2.73, 3.06.

Soluzione

La varianza campionaria è $s^2 = 0.02$. La stima puntuale della varianza σ^2 della popolazione è dunque 0.02. Se si ipotizza che la popolazione sia normale, la precisione dello stimatore S^2 può essere calcolata in modo approssimato sostituendo nella formula della deviazione standard di S^2, data dalla radice quadrata della seconda delle (17.6), $s^2 = 0.02$ a σ^2 e 0 a γ:

$$\hat{\sigma}_{S^2} = \sqrt{\frac{2 \cdot 0.02^2}{20} \cdot \frac{20}{29}} = 0.0065.$$

■

Viene ora introdotta la nozione di *efficienza*.

DEFINIZIONE 18.4

Siano $T_1 = t_1(X_1, X_2, \ldots, X_n)$ e $T_2 = t_2(X_1, X_2, \ldots, X_n)$ due stimatori del parametro θ. Allora T_1 è detto più efficiente di T_2 se

$$\text{MSE}(T_1) \leq \text{MSE}(T_2), \text{ per ogni } \theta,$$

dove la disuguaglianza stretta deve valere per almeno un valore di θ.

La disuguaglianza stretta invocata nella definizione merita una breve digressione. Va ricordato innanzitutto che l'errore quadratico medio può essere funzione di θ. Pertanto, dati gli stimatori T_1 e T_2, convenendo di dire che uno di essi prevale sull'altro, per un assegnato valore di θ, se esso presenta un errore quadratico medio più piccolo, T_1 è considerato più efficiente di T_2 se si verificano le due condizioni: (a) T_2 non prevale mai su T_1, qualunque sia il valore di θ; (b) T_1 prevale su T_2 per almeno un valore di θ.

La Figura 18.2 illustra il caso in cui tra T_1 e T_2 non esiste lo stimatore più efficiente, perché, in un insieme di valori del parametro, T_1 prevale su T_2, mentre nell'insieme complementare si verifica l'opposto.

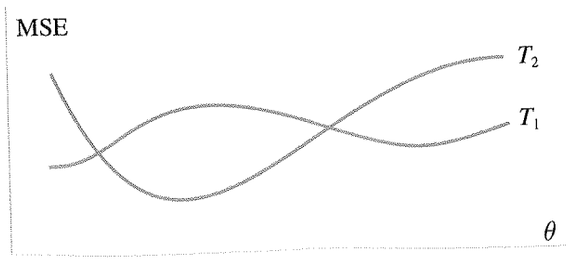


Figura 18.2
Errore quadratico medio di T_1 e T_2 in funzione di θ.

ESEMPIO 18.9

Si può dimostrare che la mediana campionaria è uno stimatore non distorto della media di una popolazione normale. È legittimo chiedersi se, nel caso in esame, la media campionaria sia più efficiente della mediana campionaria.

Soluzione

Se la popolazione è normale, l'errore quadratico medio della mediana campionaria, dato da

$$\text{E}(M - \mu)^2 = \frac{\pi}{2} \cdot \frac{\sigma^2}{n}$$

(si veda Kendall e Stuart, 1973, Vol. 2, p. 7), supera quello della media campionaria (pari a σ^2/n) per qualsiasi valore di μ. Pertanto, nella stima della media di una popolazione normale, la media campionaria è più efficiente della mediana campionaria (si veda la Figura 18.3).

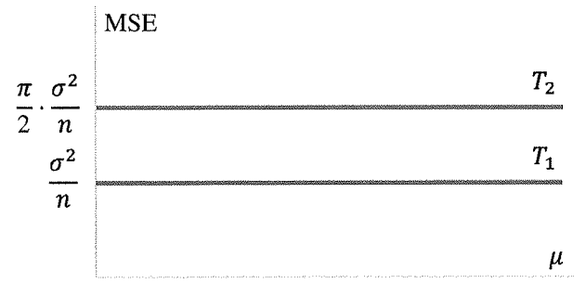


Figura 18.3
Errore quadratico medio della media campionaria, T_1, e della mediana campionaria, T_2, nel caso di una popolazione normale.

■

ESEMPIO 18.10

La Figura 18.4 illustra l'errore quadratico medio di due stimatori della varianza, $T_1 = S^2$ e $T_2 = 50$, per un campione di ampiezza $n = 30$ proveniente da una popolazione normale. Il secondo stimatore è assurdo, perché assegna a σ^2 sempre lo stesso valore, indipendentemente dal campione osservato. Gli errori quadratici medi dei due stimatori sono:

$$\text{MSE}(T_1) = \frac{\sigma^4}{30} \cdot \frac{2 \cdot 30}{29} = 0.069\sigma^4,$$

$$\text{MSE}(T_2) = \text{E}(50 - \sigma^2)^2 = \sigma^4 - 100\sigma^2 + 2500$$

(l'errore quadratico medio di S^2 si deduce dalla seconda delle 17.6, ponendo $n = 30$ e $\gamma = 0$).

Discussione

Gli errori quadratici medi dei due stimatori sono funzioni di σ^2. Come si vede, T_2 supera in efficienza T_1 quando la varianza della popolazione assume valori prossimi a 50.

■

L'argomento di cui sopra può essere applicato a qualsiasi altro stimatore. La conclusione è che non esiste uno stimatore migliore in assoluto, cioè rispetto a tutti i possibili stimatori e per qualsiasi valore di θ: è necessario concentrare l'attenzione su una sottoclasse di tutti i possibili stimatori, ad esempio, su quella degli stimatori non distorti.

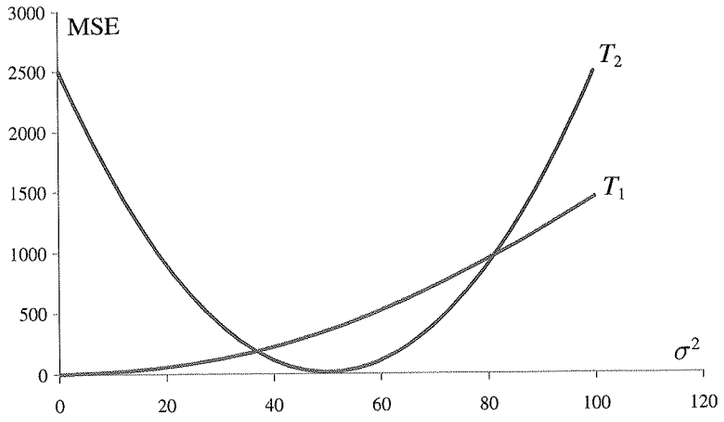


Figura 18.4
Errore quadratico medio di $T_1 = S^2$ e $T_2 = 50$ in funzione di σ^2.

Lo stimatore più efficiente in questa classe è quello che ha una varianza minore di qualsiasi altro stimatore non distorto per tutti i possibili valori del parametro. Un tale stimatore è detto stimatore non distorto con varianza uniformemente minima.

18.2.3 Proprietà asintotiche

Se è legittimo aspettarsi che un campione casuale di grande dimensione rispecchi abbastanza fedelmente le caratteristiche della popolazione da cui è estratto, è altrettanto legittima l'aspettativa che uno stimatore applicato a un tale campione fornisca una stima del parametro molto vicina al suo valore reale. Quando si studia il comportamento degli stimatori in una successione di campioni di dimensioni sempre più grandi, si parla di proprietà asintotiche.

ESEMPIO 18.11

Dato un campione casuale proveniente da una popolazione normale con media non nota μ e varianza $\sigma^2 = 30$, si calcoli la probabilità che il valore assoluto dell'errore di stima $\bar{X} - \mu$ sia inferiore a un numero piccolo, ad esempio $\varepsilon = 0.5$, per campioni di dimensione $n = 10, 100, 300$.

Soluzione

$\bar{X}$ segue una distribuzione normale con media μ e varianza $30/n$ (si veda il Paragrafo 17.5.1). Pertanto, la probabilità cercata è data da

$$P(|\bar{X} - \mu| < 0.5) = P(\mu - 0.5 < \bar{X} < \mu + 0.5) = \Phi\left(\frac{0.5}{\sqrt{30/n}}\right) - \Phi\left(\frac{-0.5}{\sqrt{30/n}}\right)$$

ed è riportata nella Tabella 18.3 per le tre dimensioni campionarie.

Tabella 18.3
Probabilità $P(|\bar{X} - \mu| < 0.5)$ per alcuni valori di n.

Ampiezza del campione	$P(\lvert\bar{X} - \mu\rvert < 0.5)$
10	0.227
100	0.639
300	0.886

La probabilità che il valore assoluto dell'errore di stima sia inferiore a 0.5 aumenta rapidamente con n, come illustrato nella Figura 18.5, che mostra la distribuzione di probabilità di $\bar{X} - \mu$ per ciascuna dimensione del campione. Come si vede, passando da $n = 10$ a $n = 100$ e poi a $n = 300$, la distribuzione di probabilità di $\bar{X} - \mu$ diventa sempre più concentrata attorno allo 0. Di conseguenza, aumenta la probabilità che l'errore di stima appartenga all'intervallo $(-0.5, 0.5)$. Risultati simili si otterrebbero per qualsiasi altro valore positivo di ε.

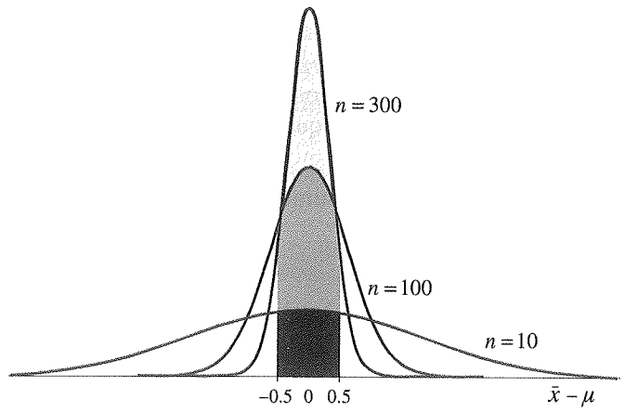


Figura 18.5
Distribuzione di probabilità di $\bar{X} - \mu$ per $n = 10, 100, 300$.

Sia T uno stimatore del parametro θ e sia ε un numero positivo arbitrariamente piccolo. Si consideri la probabilità che T differisca da θ (in valore assoluto) per meno di ε:

$$P(|T - \theta| < \varepsilon) = P(\theta - \varepsilon < T < \theta + \varepsilon).$$

Si noti che $P(|T - \theta| < \varepsilon)$ può essere considerata come una misura della vicinanza di T a θ, nel senso che essa è tanto maggiore quanto più la distribuzione di T è concentrata attorno a θ. La questione è stabilire come tale probabilità varia quando la dimensione del campione tende a infinito.

DEFINIZIONE 18.5

Uno stimatore $T_n = t_n(X_1, X_2, \ldots, X_n)$ del parametro θ è detto consistente[1] se, per ogni ε positivo arbitrariamente piccolo, si ha

$$\lim_{n \to \infty} P(|T_n - \theta| < \varepsilon) = 1.$$

Il deponente n in T sta a ricordare che la distribuzione campionaria dello stimatore T dipende dalla dimensione del campione.

In parole semplici, la proprietà della consistenza può essere così espressa: uno stimatore consistente, se calcolato su un campione molto grande, fornisce con alta probabilità una stima molto vicina al valore vero del parametro.

In parole semplici, la proprietà della consistenza può essere così espressa: uno stimatore consistente, se calcolato su un campione molto grande, fornisce con alta probabilità stime molto vicine al valore vero del parametro.

[1] La consistenza può essere declinata in altri modi. Ad esempio, uno stimatore è detto consistente in media quadratica se il suo errore quadratico medio tende a 0 quando n tende a infinito. Si può dimostrare che uno stimatore consistente in media quadratica è anche consistente nel senso della Definizione 18.5 o, come anche si dice, è consistente in probabilità. La dimostrazione è data nella Sezione B.10 dell'Appendice B.

PROPOSIZIONE 18.2

La media campionaria $\bar{X}$, la proporzione campionaria $\hat{p}$ e la varianza campionaria S^2 sono stimatori consistenti:

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1, \quad \lim_{n\to\infty} P(|\hat{p}_n - p| < \varepsilon) = 1,$$

$$\lim_{n\to\infty} P(|S_n^2 - \sigma^2| < \varepsilon) = 1.$$

La consistenza della media e della proporzione campionaria si deduce dalla legge dei grandi numeri (Paragrafo 16.2). La consistenza della varianza campionaria si deduce dal fatto che $\text{MSE}(S_n^2)$ tende a 0 quando n tende a ∞ (si veda la nota 1). ■

Una seconda proprietà asintotica degli stimatori è la *non distorsione asintotica*.

DEFINIZIONE 18.6

Uno stimatore $T_n = t_n(X_1, X_2, \ldots, X_n)$ del parametro θ è asintoticamente non distorto se la sua distorsione tende a 0 al tendere di n a infinito.

ESEMPIO 18.12

Si verifichi che lo stimatore

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

è asintoticamente non distorto.

Soluzione

La distorsione dello stimatore è $-\sigma^2/n$. Ne segue che

$$\lim_{n\to\infty} \text{D}(\hat{\sigma}_n^2) = \lim_{n\to\infty}\left(-\frac{\sigma^2}{n}\right) = 0.$$

■

18.3 Ricerca dello stimatore ottimo

Nel paragrafo precedente, sono state definite le proprietà principali degli stimatori (non distorsione, efficienza e consistenza), proprietà che consentono di effettuare confronti tra diversi stimatori dello stesso parametro, ad esempio tra la media campionaria e la mediana campionaria per stimare la media della popolazione. Nulla è stato detto su come costruire questi stimatori. In alcune situazioni, l'intuizione può suggerire l'impiego di specifici stimatori, ad esempio la media campionaria e la varianza campionaria per stimare le omologhe quantità della popolazione. Ma è chiaro che occorrono regole oggettive, cioè basate su argomenti razionali, che, alla luce delle informazioni disponibili sulla popolazione, consentano di individuare lo stimatore migliore.

La statistica matematica fornisce molti risultati teorici al riguardo quando è nota la forma della funzione di probabilità o di densità della popolazione. Questi risultati consentono di identificare lo stimatore ottimo, cioè con varianza uniformemente minima nella classe degli stimatori non distorti. L'approfondimento di

tali aspetti va al di là degli scopi di questo volume (il lettore interessato può consultare qualche testo di statistica matematica, ad esempio Rohatgi e Saleh, 2015, pp. 361-368). Qui ci si limiterà a chiarire in che senso si possono considerare ottimi i tre stimatori su cui si è concentrata l'attenzione, la media, la proporzione e la varianza campionarie.

Si può dimostrare che:

- la media campionaria $\bar{X}$ è lo stimatore ottimo della media di una popolazione normale; ha anche varianza uniformemente minima tra tutti i possibili stimatori non distorti espressi da una combinazione lineare di $X_1, X_2, \ldots, X_n$, qualunque sia la popolazione di provenienza del campione;
- la proporzione campionaria $\hat{p}$ è lo stimatore ottimo del parametro p di una popolazione di Bernoulli;
- la varianza campionaria S^2 è lo stimatore ottimo della varianza di una popolazione normale.

18.3.1 Metodo della massima verosimiglianza

Una tecnica fondamentale per la selezione degli stimatori è il metodo della massima verosimiglianza. Per semplicità, viene qui presentato per il caso della popolazione di Bernoulli, la cui funzione di probabilità è

$$f(x; p) = p^x(1-p)^{1-x}, \; x = 0,1.$$

La funzione di probabilità congiunta del campione osservato $(x_1, x_2, \ldots, x_n)$ è data da

$$L(p) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

La stima di massima verosimiglianza di p è il valore di p, $\hat{p}$, per il quale la probabilità congiunta $L(p)$ raggiunge il suo massimo. La soluzione di questo problema di massimo dà

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

(si veda l'Esercizio 18.16*). Estendendo il risultato al campione casuale $(X_1, X_2, \ldots, X_n)$, si ottiene lo stimatore di massima verosimiglianza

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

La procedura appena illustrata può essere generalizzata come segue. Sia $f(x; \theta)$ la funzione di probabilità o di densità della popolazione in esame, allora la stima di massima verosimiglianza di θ si ottiene in due passi. In primo luogo, si scrive la cosiddetta funzione di verosimiglianza, data da

$$L(\theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdot \cdots \cdot f(x_n; \theta).$$

Questa è la probabilità o la densità congiunta del campione osservato $(x_1, x_2, \ldots, x_n)$ considerata in funzione del parametro θ. In secondo luogo, si ricerca il valore di θ che massimizza la funzione di verosimiglianza. Questo valore è la stima di massima verosimiglianza di θ.

Si può dimostrare che gli stimatori di massima verosimiglianza hanno molte proprietà desiderabili, in particolare nel caso di campioni di grandi dimensioni (proprietà asintotiche).

ESEMPIO 18.13

Si determini la stima di massima verosimiglianza della media di una popolazione normale con varianza 1.

Soluzione

La funzione di verosimiglianza è

$$L(\mu) = \frac{1}{\sqrt{2\pi}} e^{-(x_1-\mu)^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-(x_2-\mu)^2/2} \cdot \ldots \cdot \frac{1}{\sqrt{2\pi}} e^{-(x_n-\mu)^2/2}.$$

È immediato verificare che $L(\mu)$ è massima quando

$$\sum_{i=1}^{n} (x_i - \mu)^2$$

è minima. Ciò accade quando μ è uguale alla media delle osservazioni campionarie. Perciò, la stima di massima verosimiglianza di μ è

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

■

18.4 Margine di errore nella stima della media

In questo paragrafo e nei prossimi capitoli, si utilizzeranno spesso i quantili della distribuzione normale standard. Nel Paragrafo 14.8, si è già indicato come determinarli. Vale la pena tuttavia tornare brevemente sull'argomento per sottolineare il ruolo di quelli che possono essere chiamati quantili estremi, ossia quei punti dell'asse z che staccano nelle code della curva normale standard valori di probabilità molto bassi.

Nella Figura 18.6, sono presentati i quantili di livello $\alpha/2$ e $1-\alpha/2$, indicati con i simboli $-z_{1-\alpha/2}$ e $z_{1-\alpha/2}$, dove α è una probabilità molto piccola (che di solito gli statistici pongono uguale a 0.05, 0.01, 0.005, 0.001):

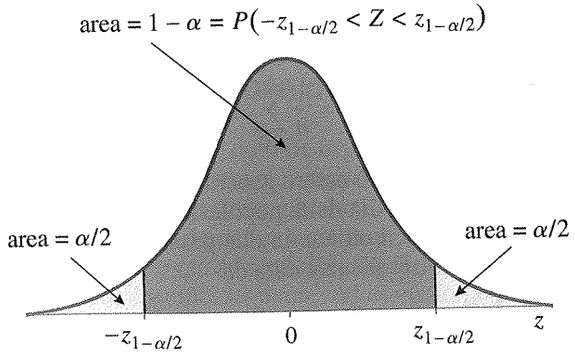


Figura 18.6
Quantili di livello $\alpha/2$ e $1-\alpha/2$ di una distribuzione normale standard.

Il quantile di livello $\alpha/2$ è il punto dell'asse z alla sinistra del quale l'area al di sotto della curva normale è $\alpha/2$. Il quantile di livello $1-\alpha/2$ è il punto dell'asse z alla sinistra del quale l'area sotto la curva normale è $1-\alpha/2$. Formalmente, i due quantili soddisfano le equazioni:

$$\Phi(-z_{1-\alpha/2}) = 1 - \Phi(z_{1-\alpha/2}) = \alpha/2, \quad \Phi(z_{1-\alpha/2}) = 1 - \alpha/2.$$

Per la loro posizione, questi quantili sono anche chiamati valori critici, per ragioni che saranno chiarite nel Capitolo 20. I valori di $1-\alpha/2$ più frequentemente usati e i corrispondenti quantili sono riportati nella tabella che segue:

$1-\alpha/2$	0.950	0.975	0.990	0.995
$z_{1-\alpha/2}$	1.645	1.96	2.326	2.576

Tabella 18.4
Valori critici della distribuzione normale standard.

Se la popolazione è normale o la dimensione del campione è sufficientemente grande, è semplice trovare la probabilità che il valore assoluto dell'errore di stima sia inferiore a un certo limite. Infatti, siccome $\bar{X} - \mu$ ha distribuzione normale o approssimativamente normale con media 0 e varianza σ^2/n, si può scrivere, ad esempio

$$P\left(-1.96\frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < 1.96\frac{\sigma}{\sqrt{n}}\right) = P\left(|\bar{X} - \mu| < 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

Dunque, il valore assoluto dell'errore che si commette nella stima di μ con $\bar{X}$ è minore di $1.96\,\sigma/\sqrt{n}$ con probabilità 0.95.

In generale, per un livello di probabilità $1-\alpha$ qualsiasi, si può scrivere

$$P\left(-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = P\left(|\bar{X} - \mu| < z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha. \quad (18.2)$$

La (18.2) indica che con probabilità $1-\alpha$ l'errore di stima connesso con l'uso di $\bar{X}$ non supera la quantità

$$d = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad (18.3)$$

detta margine di errore.

ESEMPIO 18.14

Si determini il margine di errore al 95% per la media campionaria nel caso di un campione di ampiezza $n = 80$ proveniente da una popolazione normale con varianza $\sigma^2 = 45$.

Soluzione

Il margine di errore assume il valore

$$d = 1.96\sqrt{\frac{45}{80}} = 1.47.$$

Si può dunque concludere che, per il problema in esame, il valore assoluto dell'errore di stima non supera 1.47 con probabilità 0.95. ■

Nel caso particolare di una popolazione Bernoulliana e in presenza di un campione di grande dimensione, considerando che la variabile casuale

$$Z_{\hat{p}} = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

si distribuisce approssimativamente secondo la normale standard, con le stesse argomentazioni con cui si è pervenuti alla (18.3), si trova che il margine di errore

dello stimatore $\hat{p}$ è dato da

$$d = z_{1-\alpha/2}\sqrt{p(1-p)/n}.$$

Si tratta di una quantità che dipende dal parametro p che si intende stimare. Poiché il prodotto $p(1-p)$ raggiunge il massimo quando $p = 0.5$, il margine di errore è al più uguale a

$$d = z_{1-\alpha/2}\frac{0.5}{\sqrt{n}}. \tag{18.4}$$

In definitiva, il margine di errore di $\hat{p}$ è minore o uguale a $0.5 \cdot z_{1-\alpha/2}/\sqrt{n}$ con probabilità $1-\alpha$.

18.4.1 Ampiezza del campione e margine di errore

Poiché d è funzione di n, è possibile determinare la dimensione del campione che garantisce un margine di errore assegnato con probabilità $1-\alpha$: basta risolvere rispetto a n l'Equazione (18.3). La dimensione del campione è allora

$$n = z^2_{1-\alpha/2}\frac{\sigma^2}{d^2}. \tag{18.5}$$

ESEMPIO 18.15

La produzione di olive per pianta in un uliveto segue una distribuzione normale con media non nota e deviazione standard 2.4 kg. Si determini la dimensione del campione necessaria per stimare la media con un margine di errore di 0.5 kg. Si ponga $1-\alpha = 0.95$.

Soluzione

In base alla (18.5), si ha

$$n = \frac{1.96^2 \cdot 2.4^2}{0.5^2} = 88.5.$$

Dunque, nella stima della produzione media con $\bar{X}$, per avere un margine di errore di 0.5 al 95% è necessario un campione di ampiezza 89. ■

Quando la popolazione è Bernoulliana, la ricerca della dimensione del campione, che assicura un dato livello del margine di errore con probabilità $1-\alpha$, è basata sulla (18.4), dalla quale, risolvendo rispetto a n, si ricava

$$n = \frac{0.25 \cdot z^2_{1-\alpha/2}}{d^2}.$$

ESEMPIO 18.16

Qual è la dimensione campionaria che assicura un margine di errore pari a 0.03 al 95% nella stima della proporzione p con $\hat{p}$?

Soluzione

La dimensione campionaria richiesta è

$$n = \frac{0.25 \cdot 1.96^2}{0.03^2} = 1067.$$

■

Riassunto

In questo capitolo, viene affrontato il problema della stima di un parametro incognito della popolazione oggetto di studio calcolando con le osservazioni di un campione casuale una conveniente statistica campionaria, detta stimatore. La qualità di uno stimatore si stabilisce sulla base delle proprietà che possiede tra quelle desiderabili: non distorsione, efficienza e consistenza. Uno stimatore è detto non distorto se il suo valore atteso è uguale al valore del parametro oggetto di stima. Siano T_1 e T_2 due stimatori dello stesso parametro θ. Allora si dice che T_1 è più efficiente di T_2 se le stime fornite da T_1 sono più vicine, in media, al valore reale del parametro che non quelle fornite da T_2. La misura di prossimità generalmente utilizzata è l'errore quadratico medio, cioè il valore atteso dei quadrati degli errori di stima. Uno stimatore è detto consistente se, al tendere di n a infinito, tende a 1 la probabilità che l'errore di stima sia molto piccolo.

Quanto agli stimatori di maggiore rilievo in questa trattazione – la media, la proporzione e la varianza campionarie – essi sono tutti e tre non distorti e consistenti. Inoltre: la media campionaria è lo stimatore più efficiente tra tutti i possibili stimatori non distorti della media di una popolazione normale; la proporzione campionaria è lo stimatore più efficiente tra tutti i possibili stimatori non distorti della proporzione della popolazione; la varianza campionaria S^2 è lo stimatore più efficiente tra tutti gli stimatori non distorti della varianza di una popolazione normale.

Quando la popolazione è descritta da una funzione di probabilità o di densità nota, una procedura importante per costruire stimatori di alta qualità è il metodo di massima verosimiglianza che assegna al parametro incognito il valore per il quale la funzione di probabilità o di densità congiunta del campione osservato è massima.

L'ultimo paragrafo è dedicato alla quantificazione del rischio di errore connesso alla stima della media della popolazione con la media campionaria.

Parole chiave

distorsione, 386
errore quadratico medio, 387
errore di stima, 384
funzione di verosimiglianza, 395
margine di errore, 397
metodo della massima verosimiglianza, 395
proprietà asintotiche, 392
quantili estremi, 396
stima, 383
stima di massima verosimiglianza, 395
stimatore, 383
stimatore asintoticamente non distorto, 394
stimatore consistente, 393
stimatore consistente in media quadratica, 393
stimatore consistente in probabilità, 393
stimatore distorto, 386
stimatore non distorto, 385
stimatore non distorto con varianza uniformemente minima, 392
stimatore ottimo, 394
stimatore più efficiente, 390
valori critici, 397

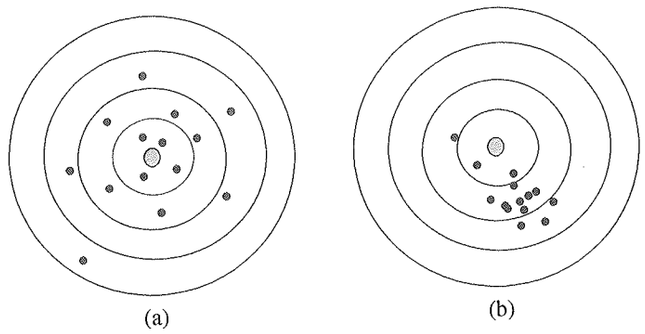
Commenti e curiosità

1. Stime come tiri al bersaglio

Stimare esattamente un parametro è come centrare il bersaglio da parte di un tiratore. Stime ripetute del parametro si possono immaginare come tiri ripetuti. Nella Figura 18.7 (a) vengono rappresentate stime ripetute prodotte da uno stimatore non distorto: le stime sono "disperse" attorno al valore del parametro (centro) senza che si manifestino deviazioni in una particolare direzione. Nella Figura 18.7 (b) invece le stime tendono a concentrarsi in una particolare zona al di sotto del centro, manifestando così la presenza di una distorsione.

L'osservazione della Figura 18.7 (b) suggerisce poi come la varianza non sia la misura appropriata per valutare l'errore medio di stima: infatti, la variabilità delle stime risulta più contenuta rispetto alla situazione (a), ma esse sono decentrate rispetto al valore del parametro. Questa è la conferma intuitiva che, ai fini della comparazione tra due stimatori, occorre considerare insieme sia la variabilità delle stime sia l'eventuale distorsione.

gura 18.7
ffigurazione simbolica
l comportamento
due stimatori, uno
n distorto (a) e uno
storto (b).

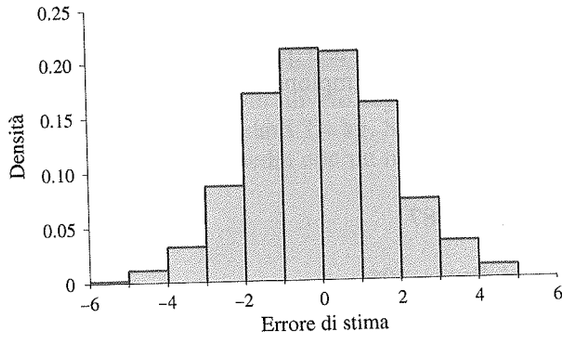


. L'errore di stima della media campionaria nel campionamento ripetuto

Nell'esempio che segue, viene mostrato il comportamento della media campionaria nel campionamento ripetuto da una popolazione normale con media e varianza note.

Da una popolazione normale con media 175 e varianza 44, vengono estratti 1000 campioni casuali di ampiezza 15. Per ogni campione, vengono calcolate la media e l'errore di stima. La distribuzione di frequenza dei mille errori di stima è rappresentata con un istogramma nella Figura 18.8.

Figura 18.8
Distribuzione di frequenza degli errori di stima osservati.



Dall'esame della figura, si trova la conferma che l'errore di stima è in media 0. Gli errori di stima oscillano tra −5 e 5, ma sono fortemente concentrati tra −2 e 2 (la percentuale degli errori compresi in questo intervallo è circa il 75%). La media dei 1000 errori di stima al quadrato è 2.96, un valore coerente con l'errore quadratico medio esatto $\sigma^2_{\bar{X}} = 44/15 = 2.93$.

Per dare maggiore concretezza all'esercizio di simulazione, nella Tabella 18.5, sono elencati alcuni dei 1000 campioni estratti, per la precisione i primi cinque e gli ultimi cinque, con l'indicazione della media e dell'errore di stima.

Tabella 18.5
Medie ed errori di stima per alcuni campioni osservati.

Campione	Media campionaria	Errore di stima
1	173.98	−1.02
2	172.25	−2.75
3	175.66	0.66
4	175.80	0.80
5	173.61	−1.39
⋮	⋮	⋮
996	173.81	−1.19
997	175.73	0.73
998	179.10	4.10
999	175.88	0.88
1000	175.06	0.06

3. Consistenza come smorzamento, all'aumentare di n, delle oscillazioni delle stime attorno al parametro

Si osservi la Figura 18.9, dove è rappresentato l'andamento dei valori assunti dalla media campionaria $\bar{X}$ nei 1000 campioni estratti da una popolazione normale in un esercizio di simulazione condotto con tre diverse ampiezze campionarie: $n = 10, 20, 50$.

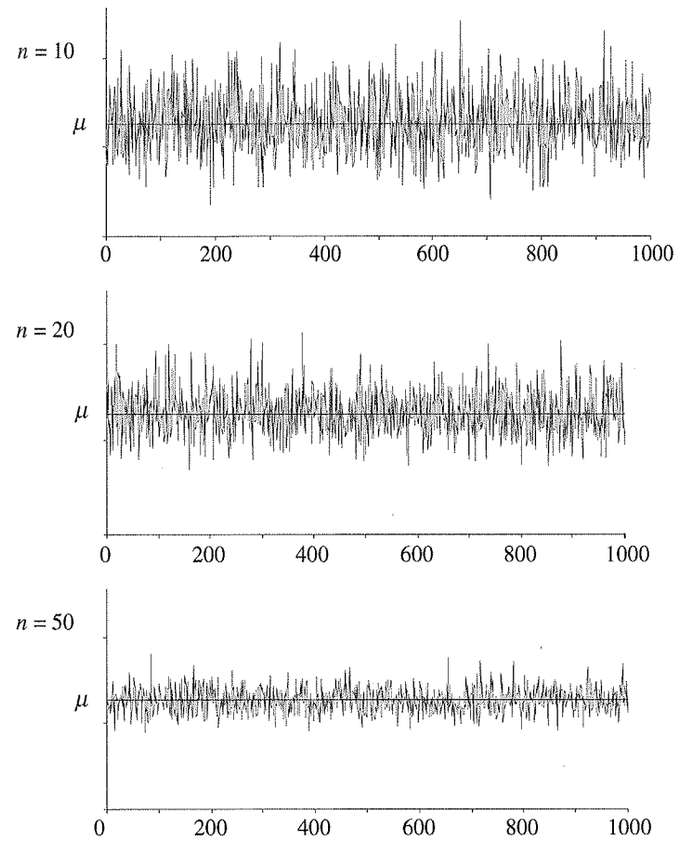


Figura 18.9
Oscillazione della media campionaria attorno alla media della popolazione all'aumentare di n.

Come si può constatare, all'aumentare di n, si attenua l'oscillazione delle stime attorno alla media della popolazione (rappresentata dalla linea orizzontale). Ciò significa che gli errori di stima tendono a divenire più piccoli, in valore assoluto, a mano a mano che cresce n, e che, di conseguenza, cresce la frazione sul totale degli errori di stima (in valore assoluto) che sono al di sotto di un certo limite.

4. Ronald Fisher e il metodo della massima verosimiglianza

History and Art Collection/Alamy Stock Photo

Ronald Aylmer Fisher (1890-1962), matematico, statistico e genetista inglese, compì i suoi studi universitari in matematica presso il Gonville and Caius College, Cambridge. Dopo una serie di lavori come statistico presso la City di Londra e come insegnante di fisica e matematica in scuole private, dal 1919 al 1933 svolse ricerca statistica presso la stazione sperimentale di Rothamsted (Hertfordshire), specializzata in esperimenti in agricoltura. Questa attività contribuì senza dubbio a plasmare il suo profilo di ricercatore. Furono infatti proprio i temi dei piani degli esperimenti e della connessa metodologia di analisi dei dati, la cosiddetta analisi della varianza, che costruirono la sua grande reputazione come biostatistico. Nel 1933 fu chiamato a dirigere il dipartimento di eugenetica presso l'University College London (un dipartimento nato dalla scissione in due del dipartimento di statistica applicata diretto fino ad allora da Karl Pearson; si veda al riguardo la nota 1 di "Commenti e curiosità"

del Capitolo 20), dove svolse un'intensa attività fino al 1943. Nel 1934 divenne direttore degli *Annals of Eugenics* (ora *Annals of Human Genetics*). Nel 1943 si trasferì all'università di Cambridge, dove operò fino alla pensione.

Tra i suoi contributi alla statistica moderna, si ricordano il metodo della massima verosimiglianza e l'analisi della varianza. Il suo libro *The Design of Experiments*, pubblicato nel 1935, è un testo fondamentale per la teoria dei piani degli esperimenti e dell'analisi dei dati che ne conseguono. Fondamentali sono anche i suoi contributi nel campo della verifica delle ipotesi e delle distribuzioni campionarie. Per i suoi altissimi meriti scientifici, Ronald Fisher è stato definito come "un genio che quasi da solo ha creato le fondamenta della scienza statistica moderna".

Fra Ronald Fisher giovane e Karl Pearson già affermatissimo direttore della rivista *Biometrika* (tra i due vi era una differenza di età di oltre 30 anni; si veda la nota biografica di Pearson in "Commenti e curiosità" del Capitolo 22) si manifestarono aspri contrasti per motivi accademici e caratteriali. Una delle dispute tra i due ebbe a oggetto un test statistico introdotto da Pearson nel 1900, basato sulla distribuzione chi-quadrato. Fisher dimostrò che il numero dei gradi di libertà della distribuzione chi-quadrato utilizzata era sbagliato. Famosa è anche la motivazione addotta da Pearson nel rifiutare un articolo presentato da Fisher per la pubblicazione su *Biometrika*: "Nelle condizioni finanziarie attuali, sono costretto mio malgrado a escludere tutto ciò che a mio giudizio è erroneo; ciò in quanto non mi posso consentire controversie di sorta" (Pearson, 1968, p. 453). Altrettanto famosa è la sprezzante definizione che Fisher diede della scuola di Pearson: "Un esercito di industriosi robot obbedienti magicamente ai cenni del maestro". I pessimi rapporti tra i due si riverberarono anche sui rapporti personali e accademici tra Fisher ed Egon Pearson, il figlio di Karl Pearson, che dal 1933 subentrò al padre nella direzione del dipartimento di statistica applicata dell'University College London e nella direzione di *Biometrika*.

Esercizi

18.1 Un'azienda produttrice di oggetti in plastica vuole valutare la resistenza di blocchi di plastica da arredamento. In un campione di 50 blocchi trova i seguenti valori di un indice di solidità:

282, 275, 278, 239, 335, 303, 240, 255, 282, 270, 269, 250, 233, 301, 289, 241, 268,
279, 228, 265, 286, 279, 252, 272, 235, 264, 315, 278, 246, 295, 252, 229, 255, 267,
255, 281, 302, 256, 262, 194, 292, 264, 274, 268, 283, 261, 275, 277, 277, 260.

Si stimi la media della popolazione utilizzando la media campionaria e si determini la deviazione standard dello stimatore, sapendo che la varianza della popolazione è 485.

18.2 L'osservazione della durata (in ore) delle batterie per cellulare di una data marca in 24 esemplari di prodotto ha dato luogo ai seguenti risultati:

58.7, 64.9, 76.9, 67.8, 41.7, 56.7, 64.5, 69.7, 82.1, 82.5, 40.8, 74.9,
71.5, 75.4, 67.3, 73.0, 70.4, 104.0, 82.3, 90.4, 86.8, 72.8, 71.8, 54.5.

Si stimi la media della popolazione e si misuri l'accuratezza delle stime, assumendo che $\sigma^2 = 230$.

18.3 Siano

$$T_1 = (X_1 + X_2 + X_3 + X_4)/4 \quad \text{e} \quad T_2 = (3X_1 + 4X_2 + X_3 + 2X_4)/10$$

due stimatori della media della popolazione basati su un campione di ampiezza 4.

a) Si verifichi che T_2 è non distorto.

b) Si determini l'errore quadratico medio dei due stimatori.

c) In base al risultato ottenuto al punto b, quale dei due stimatori è più efficiente?

18.4 Si consideri lo stimatore della media dato da $T = a + b\bar{X}$, con b maggiore di 0. Se ne determini l'errore quadratico medio e si stabilisca se T è consistente.

18.5 Si consideri lo stimatore $T = 0.4X_1 + 0.2X_2 + 0.4X_3$ per la stima della media della popolazione con un campione $(X_1, X_2, \ldots, X_n)$ di ampiezza n.

a) Lo stimatore è non distorto?

b) Lo stimatore è consistente?

18.6* Si dimostri che l'errore quadratico medio dello stimatore T del parametro θ può essere così scomposto

$$\text{MSE}(T) = \text{Var}(T) + [\text{D}(T)]^2.$$

(*Suggerimento*: si scriva l'espressione esplicita dell'errore quadratico medio $\text{MSE}(T) = \text{E}(T - \theta)^2$ e si sfrutti l'identità $\text{E}(T - \theta)^2 = \text{E}[(T - \mu_T) + (\mu_T - \theta)]^2$, essendo μ_T il valore atteso dello stimatore; dopodiché dallo sviluppo del quadrato del binomio $[(T - \mu_T) + (\mu_T - \theta)]^2$, considerando che $\text{E}(T - \mu_T) = 0$, si perviene, con facili passaggi, al risultato.)

18.7 Siano T_1 e T_2 due stimatori alternativi della media della popolazione di interesse. Sapendo che

$$\text{E}(T_1) = \mu, \quad \text{Var}(T_1) = 27.5, \quad \text{E}(T_2) = \mu + 3, \quad \text{Var}(T_2) = 17,$$

quale dei due stimatori è più efficiente?

18.8 L'ufficio statistico di un Comune ha condotto un sondaggio tra i cittadini maggiorenni per accertare il gradimento di un nuovo sistema di raccolta differenziata dei rifiuti. In un campione casuale di 530 individui, 360 si sono dichiarati favorevoli. Si stimi la proporzione dei favorevoli nella popolazione.

18.9 In una ricerca condotta da un centro studi sull'utilizzo di Internet è stato trovato che, sui 540 intervistati di un campione casuale, 248 ricevono più di 15 messaggi di posta elettronica al giorno. Si determini la stima della proporzione nella popolazione degli utilizzatori di Internet con la caratteristica sopra indicata.

18.10 In una sperimentazione clinica si vuole stimare il tasso di guarigione p connesso all'uso di un nuovo farmaco. Quale deve essere il numero delle unità da includere nell'esperimento affinché l'errore di stima per $\hat{p}$ sia inferiore, in valore assoluto, a 0.05 con probabilità 0.95?

18.11 La deviazione standard della statura degli studenti iscritti in una università del Nord Italia è 5.8. Quanti studenti si devono estrarre a sorte dalla popolazione se si vuole con probabilità del 90% che il valore assoluto dell'errore di stima di $\bar{X}$ non superi 2 cm?

18.12 Un ricercatore vuole stimare la media di una popolazione normale con deviazione standard σ non nota. L'obiettivo è quello di ottenere con probabilità 0.95 una stima che non differisca (in valore assoluto) dalla media della popolazione per più del 5% di σ. Si determini l'ampiezza del campione che assicura un tale risultato.

18.13 Data una popolazione normale, si determini l'errore quadratico medio dello stimatore

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

18.14* Si dimostri che la varianza campionaria S^2 è uno stimatore consistente di σ^2. (*Suggerimento*: si applichi la disuguaglianza di Chebyshev alla variabile casuale S^2 con media e varianza date dalle 17.6 e poi si faccia il limite per $n \to \infty$ del primo e del secondo membro della disuguaglianza.)

18.15 Il direttore di un grande magazzino ha osservato i tempi di servizio alla cassa riservata ai clienti con un numero di articoli non superiore a 7 in situazione di affollamento. In 58 osservazioni ha trovato un tempo medio $\bar{x} = 2.5$ e una deviazione standard $s = 1.6$. Che cosa si può dire, con un livello di probabilità del 99%, circa l'errore massimo in cui si incorre nella stima della media?

18.16* Si dimostri che il punto di massimo della seguente funzione di verosimiglianza

$$L(p) = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

(dove p è la proporzione nella popolazione) è $\hat{p} = \sum_{i=1}^{n} x_i/n$.

18.17 Si consideri una popolazione descritta dalla funzione di densità esponenziale

$$f(x;\lambda) = \lambda e^{-\lambda x},\ x \geq 0.$$

Si stimi il parametro λ con il metodo della massima verosimiglianza, sulla base delle seguenti osservazioni campionarie:

15.4, 26.9, 10.5, 32.5, 38.2, 44.1.

CAPITOLO

19

Stima intervallare

Sommario del capitolo

19.1 Introduzione
19.2 Stima intervallare della media di una popolazione normale con varianza nota
19.3 Stima intervallare della media di una popolazione normale con varianza non nota
19.4 Stima intervallare della media nel caso di grandi campioni
19.5 Stima intervallare della varianza di una popolazione normale
Riassunto
Commenti e curiosità
Esercizi

19.1 Introduzione

Come illustrato nel Capitolo 18, con la stima puntuale si attribuisce un unico valore all'ignoto parametro θ, applicando un'appropriata statistica a un campione casuale osservato. Tale numero non coinciderà quasi certamente con il reale valore del parametro, ma ne costituirà un'approssimazione.

Con la stima intervallare, l'idea di raggiungere una conoscenza approssimata del parametro incognito si concretizza con l'individuazione di un insieme di valori plausibili per θ, sempre sulla base di statistiche applicate ai dati di un campione casuale osservato.

In questo capitolo, l'attenzione sarà concentrata sulla stima intervallare dei parametri media e varianza. Applicazioni in altri ambiti sono contenute nei capitoli 21 e 23.

Si dà ora la definizione di *intervallo di confidenza* con riferimento a un generico parametro θ.

DEFINIZIONE 19.1

Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale estratto da una popolazione della quale interessa stimare il parametro θ. Siano

$$L_1 = g_1(X_1, X_2, \ldots, X_n) \quad \text{e} \quad L_2 = g_2(X_1, X_2, \ldots, X_n)$$

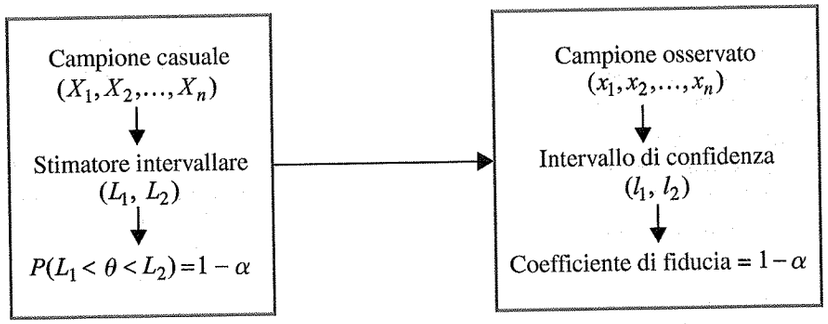
due statistiche campionarie tali che

$$P(L_1 < \theta < L_2) = 1 - \alpha, \tag{19.1}$$

dove α è un numero molto piccolo ($\alpha = 0.10, 0.05, 0.01, \ldots$) che non dipende da θ. L'intervallo (L_1, L_2) si chiama stimatore intervallare del parametro θ. Siano l_1 e l_2 i valori assunti dalle statistiche L_1 e L_2 nel campione casuale osservato $(x_1, x_2, \ldots, x_n)$. Allora (l_1, l_2) è l'intervallo di confidenza per θ con un coefficiente di fiducia pari a $1 - \alpha$.

Dunque, L_1 e L_2 sono variabili casuali e di conseguenza (L_1, L_2) è un intervallo casuale che contiene al suo interno il valore del parametro incognito θ con probabilità $1-\alpha$. L'intervallo di confidenza (l_1, l_2) è invece una realizzazione di (L_1, L_2), nel senso che i suoi estremi sono i valori assunti da L_1 e L_2 nel campione casuale osservato. Come tale, esso non si presta ad alcuna valutazione probabilistica.

In altri termini, è errato asserire che esso contiene il valore del parametro con probabilità $1-\alpha$. Questa quantità, se riferita a (l_1, l_2), cessa di essere una probabilità e diviene un indicatore dell'affidamento che si fa su di esso, per i motivi indicati qui di seguito. Si immagini di estrarre ripetutamente dalla popolazione d'interesse, per molte e molte volte, un campione casuale di ampiezza n e di calcolare per ciascun campione l'intervallo di confidenza (l_1, l_2) con lo stesso livello $1-\alpha$. Per la legge dei grandi numeri (versione di Bernoulli), la frequenza relativa dei campioni validi, ossia di quelli per cui (l_1, l_2) include al suo interno il parametro incognito, sarà anch'essa molto alta e vicina alla probabilità $1-\alpha$ che appare nella (19.1). Per questo, si può confidare che l'attuale intervallo di confidenza, quello ottenuto con il campione osservato, sia valido. In definitiva, la quantità $1-\alpha$ è una probabilità se riferita all'intervallo (L_1, L_2), è una misura del grado di fiducia nella sua validità se riferita all'intervallo osservato (l_1, l_2). Queste idee sono illustrate con la Figura 19.1.



igura 19.1
allo stimatore intervallare
l'intervallo di confidenza.

I due elementi che caratterizzano l'intervallo di confidenza (l_1, l_2) sono: il coefficiente di fiducia $1-\alpha$ e l'ampiezza dell'intervallo, $A = l_2 - l_1$. È desiderabile che il primo sia molto elevato, perché quanto maggiore è $1-\alpha$ tanto maggiore è l'affidamento sul fatto che l'intervallo (l_1, l_2) contenga al suo interno il valore del parametro incognito θ. È ugualmente desiderabile che l'ampiezza A dell'intervallo sia quanto più possibile piccola, giacché, come è ovvio, l'informazione sul parametro che si trae da (l_1, l_2) è tanto più precisa quanto minore è la distanza tra l_1 e l_2.

I livelli di $1-\alpha$ più frequentemente usati sono: 0.90, 0.95 e 0.99. Nel seguito, l'intervallo (l_1, l_2) ottenuto con un certo livello di $1-\alpha$ verrà talvolta indicato come intervallo di confidenza all'$(1-\alpha)\cdot 100\%$.

19.2 Stima intervallare della media di una popolazione normale con varianza nota

Nel caso di una popolazione normale con varianza nota, l'intervallo di confidenza per il parametro μ viene costruito a partire dalla variabile casuale[1]

$$Z_{\bar{X}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}},$$

[1] $Z_{\bar{X}}$ (come le altre funzioni che verranno usate nei prossimi paragrafi) viene chiamata quantità pivot (nel senso di quantità risolutiva), perché ha alcune proprietà che la rendono adatta per la costruzione dell'intervallo di confidenza: è una funzione monotona del parametro incognito e ha distribuzione di probabilità che non dipende da tale parametro.

che ha distribuzione normale standard qualunque sia il valore di μ (si veda il Paragrafo 17.5.1). Ciò è illustrato nella Figura 19.2, dove si ricorda come la variabile casuale $Z_{\bar{X}}$ sia compresa tra i quantili $-z_{1-\alpha/2}$ e $z_{1-\alpha/2}$ con probabilità $1-\alpha$ (rappresentata dall'area sottesa alla curva normale standard tra i detti quantili).

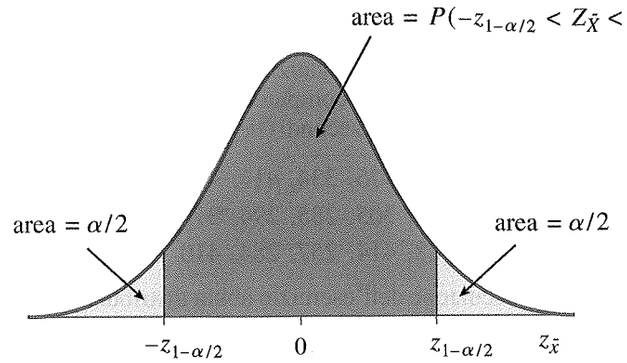


Figura 19.2
Quantili di livello $\alpha/2$ e $1-\alpha/2$ della normale standard e probabilità che $-z_{1-\alpha/2} < Z_{\bar{X}} < z_{1-\alpha/2}$.

PROPOSIZIONE 19.1

Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale estratto da una popolazione normale con media μ incognita e varianza σ^2 nota. Si consideri la variabile casuale

$$Z_{\bar{X}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}},$$

che ha distribuzione normale standard. Lo stimatore intervallare di μ, a un livello di probabilità $1-\alpha$, ha estremi dati da:

$$L_1 = \bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ L_2 = \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},$$

dove $-z_{1-\alpha/2}$ e $z_{1-\alpha/2}$ sono i quantili di livello $\alpha/2$ e $1-\alpha/2$ della normale standard. Ne segue che gli estremi dell'intervallo di confidenza per μ, con coefficiente di fiducia $1-\alpha$, sono:

$$l_1 = \bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ l_2 = \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \qquad (19.2)$$

dove $\bar{x}$ è la media del campione osservato $(x_1, x_2, \ldots, x_n)$.

Ai fini della determinazione di L_1 e L_2, si consideri l'evento aleatorio di probabilità $1-\alpha$

$$-z_{1-\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}.$$

Risolvendo le due disequazioni rispetto a μ, si perviene all'evento aleatorio equivalente

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Si noti che il primo evento implica il secondo e viceversa (in questo senso si dicono equivalenti), pertanto essi hanno la stessa probabilità $1-\alpha$. Si può allora scrivere

$$P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

Ciò vuol dire che μ è contenuto tra le variabili casuali $L_1 = \bar{X} - z_{1-\alpha/2}(\sigma/\sqrt{n})$ e $L_2 = \bar{X} + z_{1-\alpha/2}(\sigma/\sqrt{n})$ con probabilità $1-\alpha$. A partire da L_1 e L_2 è poi immediato trovare gli estremi dell'intervallo di confidenza.[2] ■

ESEMPIO 19.1

Nell'ambito della selezione del personale impiegatizio di una grande azienda, viene richiesto lo svolgimento di un esercizio al computer per il quale viene rilevato il tempo di esecuzione. I tempi osservati (in secondi) dei 24 concorrenti sono:

345, 389, 363, 417, 476, 334, 417, 427,
269, 370, 289, 336, 305, 285, 294, 319,
370, 338, 297, 328, 414, 257, 284, 410.

Ammettendo che il tempo di esecuzione dell'esercizio abbia distribuzione normale con deviazione standard 50, si costruisca l'intervallo di confidenza al 95% per la media della popolazione.

Soluzione

La media del campione è 347.21. Gli estremi dell'intervallo di confidenza sono:

$$l_1 = 347.21 - 1.96\frac{50}{\sqrt{24}} = 327.21 \quad \text{e} \quad l_2 = 347.21 + 1.96\frac{50}{\sqrt{24}} = 367.21.$$

Si può ritenere, con una fiducia del 95%, che la media della popolazione sia racchiusa nell'intervallo (327.21, 367.21). ■

Si ribadisce qui quello che si è detto in generale nel paragrafo precedente: l'intervallo di confidenza per μ non è assoggettabile a una valutazione probabilistica. Non si può dire cioè che la media della popolazione è compresa nell'intervallo (l_1, l_2) con probabilità $1-\alpha$, perché l_1 e l_2 sono due numeri, realizzazioni delle variabili casuali L_1 e L_2. Viene quindi a mancare l'evento casuale a cui poter assegnare una probabilità. Così nell'esempio precedente, sarebbe sbagliato affermare che la media della popolazione è compresa nell'intervallo (327.21, 367.21) con probabilità 0.95. Si può solo esprimere un forte grado di affidamento sul fatto che tale intervallo includa la media della popolazione.

Con la Figura 19.3 viene riproposta per l'intervallo di confidenza per la media l'interpretazione del coefficiente di fiducia formulata in generale nel paragrafo precedente.

[2] Quello descritto non è l'unico modo per determinare un intervallo di confidenza per μ. Vi sono infiniti modi di scegliere una coppia di numeri, z_1 e z_2 (sull'asse delle ascisse del grafico della normale standard), tali che l'evento aleatorio

$$z_1 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_2$$

e quindi l'evento aleatorio equivalente

$$\bar{X} - z_2\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} - z_1\frac{\sigma}{\sqrt{n}}$$

abbiano probabilità $1-\alpha$, da cui è poi immediato ricavare l'intervallo di confidenza. Si può dimostrare tuttavia che l'intervallo di confidenza basato sui quantili $-z_{1-\alpha/2}$ e $z_{1-\alpha/2}$ è quello che ha ampiezza minima.

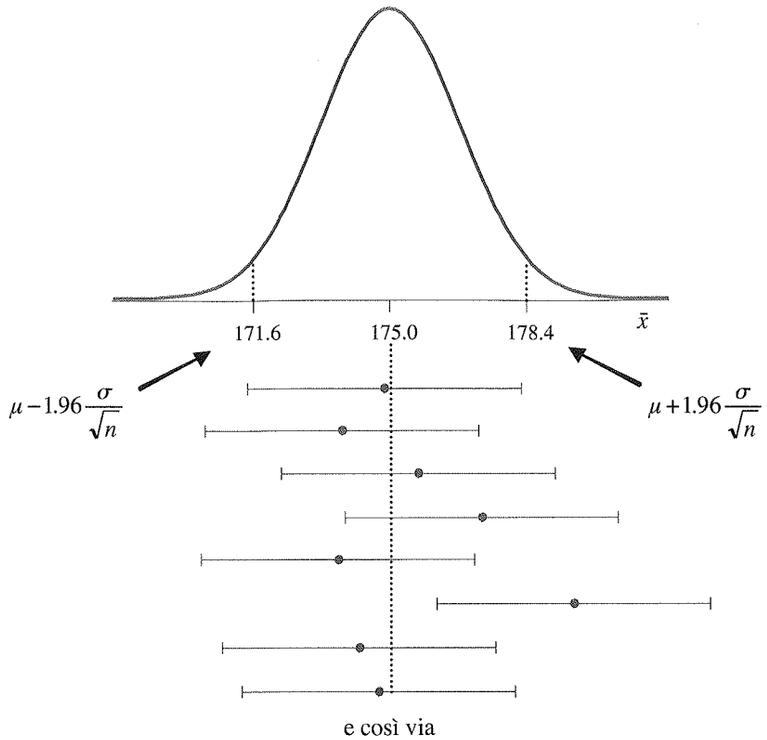


Figura 19.3
Intervalli di confidenza nello schema del campionamento ripetuto.

In essa, il grafico si riferisce alla distribuzione campionaria di $\bar{X}$ per $n = 15$, nel caso di una popolazione normale con media 175 e varianza 44. Sull'asse delle ascisse, oltre alla media μ, sono riportati i valori 171.6 e 178.4 che formano un intervallo centrato su $\mu = 175$ tale che la probabilità di includere al suo interno $\bar{X}$ è pari a 0.95:[3]

$$P(171.6 < \bar{X} < 178.4) = 0.95.$$

Al di sotto del grafico, viene descritto un ipotetico processo di estrazione ripetuta di campioni di ampiezza 15 dalla popolazione menzionata. Dunque, si immagini di estrarre un primo campione e di determinare l'intervallo di confidenza al 95% con la (19.2), cioè sottraendo e aggiungendo $1.96\sqrt{44/15} = 3.4$ dalla media campionaria, rappresentata da un punto in alto a sinistra della linea verticale tratteggiata. Il conseguente intervallo di confidenza è rappresentato da un segmento orizzontale, che, come si vede, include al suo interno la media della popolazione. Ciò avviene perché la media campionaria osservata è compresa nell'intervallo (171.6, 178.4). Si immagini di estrarre un secondo campione (sempre di ampiezza 15) dalla stessa popolazione e di determinare il relativo intervallo di confidenza al 95% con le stesse modalità del caso precedente. Anche questo intervallo copre la media della popolazione per lo stesso motivo indicato nel caso precedente. Se l'estrazione di campioni fosse ripetuta molte e molte altre volte ancora con la determinazione dei conseguenti intervalli di confidenza, la percentuale degli

[3] Tali valori sono così ottenuti:

$$171.6 = 175 - 1.96\sqrt{\frac{44}{15}} = \mu - 1.96\frac{\sigma}{\sqrt{n}}, \quad 178.4 = 175 + 1.96\sqrt{\frac{44}{15}} = \mu + 1.96\frac{\sigma}{\sqrt{n}}.$$

intervalli che includono al loro interno la media della popolazione, cioè la percentuale dei campioni la cui media cade nell'intervallo (171.6, 178.4), sarebbe molto vicina al 95%. Questa previsione è resa possibile dalla legge dei grandi numeri (formulazione di Bernoulli), per cui la frequenza relativa di un evento aleatorio tende a essere molto vicina alla probabilità dell'evento stesso quando il numero delle prove è molto alto. In questo processo di estrazione ripetuta di campioni, l'evento aleatorio è che l'intervallo contenga la media della popolazione.

$$\left(\bar{X} - 1.96\sqrt{\frac{44}{15}} < \mu < \bar{X} + 1.96\sqrt{\frac{44}{15}}\right).$$

La probabilità di questo evento è 0.95 ed è uguale alla probabilità dell'evento equivalente

$$\left(\mu - 1.96\sqrt{\frac{44}{15}} < \bar{X} < \mu + 1.96\sqrt{\frac{44}{15}}\right).$$

È evidente che nelle applicazioni pratiche si è di fronte a un solo intervallo (l_1, l_2), costruito con l'unico campione disponibile, e non è dato sapere se esso contiene o non contiene al suo interno la media della popolazione. Tuttavia, sapendo che un evento ad altissima probabilità si realizza quasi certamente, vi è la forte attitudine a credere (ad avere fiducia) che l'intervallo osservato sia valido.

Per corroborare il ragionamento appena svolto, nella Figura 19.4 vengono mostrati gli intervalli di confidenza al 95% ottenuti in un esercizio di simulazione eseguito prelevando 50 campioni casuali di dimensione 15 dalla stessa popolazione considerata nella Figura 19.3. Questi intervalli sono rappresentati da segmenti verticali il cui punto centrale è la media campionaria. La linea tratteggiata rappresenta invece la media della popolazione. Come si vede, tutti gli intervalli di confidenza tranne due (il 19° e il 42°) contengono la media della popolazione. In questo caso, quindi, il 96% degli intervalli contiene la media reale contro una percentuale teorica del 95%.

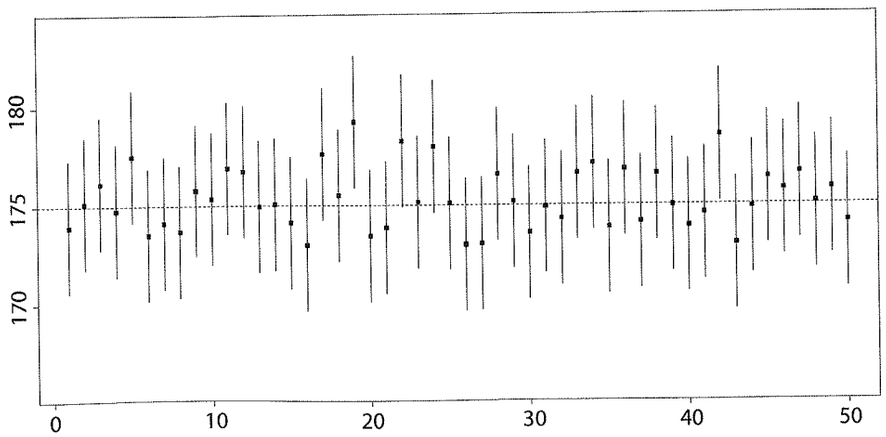


Figura 19.4
Intervalli di confidenza al 95% in 50 campioni di ampiezza 15 estratti da una popolazione normale N(175, 44).

Ampiezza dell'intervallo di confidenza

Come già osservato, minore è l'ampiezza dell'intervallo di confidenza, maggiore è la precisione della stima. L'ampiezza dell'intervallo di confidenza per la media della popolazione è data da

$$A = l_2 - l_1 = 2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}. \tag{19.3}$$

È dunque funzione di α, n e σ e soddisfa le seguenti proprietà:

- cresce al diminuire di α;
- decresce all'aumentare di n;
- è direttamente proporzionale a σ.

Per quanto riguarda la prima proprietà, va osservato che, al diminuire di α, $z_{1-\alpha/2}$ aumenta e con esso aumenta l'ampiezza dell'intervallo. Ciò significa che un aumento del coefficiente di fiducia produce un intervallo di confidenza più ampio e quindi più povero di informazione.

La seconda proprietà si desume dalla (19.3) e afferma che la precisione dell'intervallo di confidenza può essere migliorata utilizzando un campione di dimensione più alta.

La terza proprietà implica che maggiore è la varianza della popolazione oggetto di studio, più ampio è l'intervallo di confidenza. Si noti che, a differenza di α e di n, la varianza σ^2 non è modificabile dal ricercatore.

L'intervallo di confidenza può essere interpretato facendo ricorso all'idea del campionamento ripetuto. Se si immagina di estrarre dalla popolazione un numero molto elevato di campioni della stessa dimensione e di determinare per ciascuno di questi l'intervallo di confidenza con coefficiente di fiducia $(1-\alpha)100\%$, circa l'$(1-\alpha)100\%$ degli intervalli osservati conterrà il valore della media incognita. Ciò è supportato dalla legge dei grandi numeri (versione di Bernoulli).

ESEMPIO 19.2

Un urbanista è interessato alla superficie media delle abitazioni della propria città. Uno studio precedente indica che $\sigma = 8$ m^2. In un campione di 50 appartamenti si osserva $\bar{x} = 120$ m^2. Si determini l'intervallo di confidenza per μ al 95%. Si confronti poi l'ampiezza di tale intervallo con quella dell'intervallo che si avrebbe con un coefficiente di confidenza del 99%.

Soluzione

Gli estremi dell'intervallo di confidenza al 95% sono: $120 \pm 1.96 \cdot (8/\sqrt{50})$, cioè 117.78 e 122.22. L'ampiezza dell'intervallo è

$$A = 2 \cdot 1.96 \cdot \frac{8}{\sqrt{50}} = 4.43.$$

Se il coefficiente di fiducia fosse 0.99, l'ampiezza dell'intervallo di confidenza sarebbe

$$A = 2 \cdot 2.576 \cdot \frac{8}{\sqrt{50}} = 5.83.$$

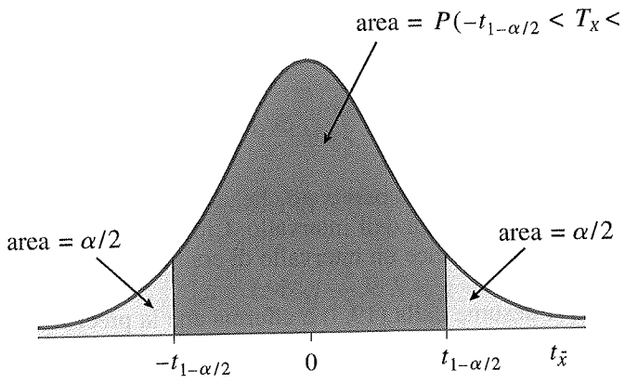
■

19.3 Stima intervallare della media di una popolazione normale con varianza non nota

Se la popolazione è normale, ma la varianza σ^2 non è nota, per determinare l'intervallo di confidenza per μ si utilizza la variabile casuale (introdotta nel Paragrafo 17.7)

$$T_{\bar{X}} = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

che ha distribuzione T di Student con $n-1$ gradi di libertà. Tale distribuzione è illustrata nella Figura 19.5, dove, sull'asse delle ascisse, sono riportati i quantili $-t_{1-\alpha/2}$ e $t_{1-\alpha/2}$, che delimitano un'area al di sotto della funzione di densità pari a $1-\alpha$.



ura 19.5
antili di livello $\alpha/2$ e
$\alpha/2$ della distribuzione
di Student con $n-1$ gradi
ibertà, e probabilità
l'evento
$_{-\alpha/2} < T_{\bar{X}} < t_{1-\alpha/2}$.

PROPOSIZIONE 19.2

Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale estratto da una popolazione normale con media e varianza non note. Si consideri la variabile casuale

$$T_{\bar{X}} = \frac{\bar{X}-\mu}{S/\sqrt{n}},$$

che ha distribuzione T di Student con $n-1$ gradi di libertà. Lo stimatore intervallare di μ, a un livello di probabilità $1-\alpha$, ha estremi dati da:

$$L_1 = \bar{X} - t_{1-\alpha/2}\frac{S}{\sqrt{n}},\ L_2 = \bar{X} + t_{1-\alpha/2}\frac{S}{\sqrt{n}},$$

dove $-t_{1-\alpha/2}$ e $t_{1-\alpha/2}$ sono i quantili di livello $\alpha/2$ e $1-\alpha/2$ della distribuzione T di Student con $n-1$ gradi di libertà. Ne segue che gli estremi dell'intervallo di confidenza per μ, con coefficiente di fiducia $1-\alpha$, sono:

$$l_1 = \bar{x} - t_{1-\alpha/2}\frac{s}{\sqrt{n}},\ l_2 = \bar{x} + t_{1-\alpha/2}\frac{s}{\sqrt{n}},$$

dove $\bar{x}$ e s sono la media e la deviazione standard del campione casuale osservato $(x_1, x_2, \ldots, x_n)$.

Si consideri l'evento aleatorio

$$-t_{1-\alpha/2} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{1-\alpha/2}$$

di probabilità $1-\alpha$. Risolvendo le due disequazioni rispetto a μ, si perviene all'evento aleatorio equivalente

$$\bar{X} - t_{1-\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{1-\alpha/2}\frac{S}{\sqrt{n}},$$

cosicché si può scrivere

$$P\left(\bar{X} - t_{1-\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{1-\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha,$$

da cui si deduce che gli estremi dello stimatore intervallare sono:

$$L_1 = \bar{X} - t_{1-\alpha/2}\frac{S}{\sqrt{n}},\quad L_2 = \bar{X} + t_{1-\alpha/2}\frac{S}{\sqrt{n}}.$$

Da qui seguono poi immediatamente gli estremi dell'intervallo di confidenza per μ. ■

L'ampiezza dell'intervallo di confidenza è data da

$$A = l_2 - l_1 = 2t_{1-\alpha/2}\frac{s}{\sqrt{n}}.$$

Come si vede, l'ampiezza dell'intervallo non è costante, ma dipende dalla deviazione standard campionaria s.

ESEMPIO 19.3

Il produttore di una certa marca di sigarette desidera controllare il quantitativo medio di catrame in esse contenuto. A questo scopo, viene osservato un campione casuale di 25 sigarette, ottenendo i risultati (espressi in mg) di seguito indicati:

11.1, 10.5, 11.4, 10.7, 11.4, 10.8, 11.2, 10.3, 11.6, 10.6, 12.7, 10.6, 11.7
10.6, 10.7, 12.2, 10.7, 11.0, 10.5, 11.0, 10.9, 10.7, 10.1, 11.3, 10.5.

Si determini l'intervallo di confidenza per μ al 99% assumendo che la popolazione sia normale.

Soluzione

La media del campione è $\bar{x} = 11$, la varianza campionaria è $s^2 = 0.357$. Poiché $t_{0.995} = 2.797$ (i gradi di libertà sono 24), gli estremi dell'intervallo di confidenza sono:

$$l_1 = 11 - 2.797\sqrt{\frac{0.357}{25}} = 10.67,\ l_2 = 11 + 2.797\sqrt{\frac{0.357}{25}} = 11.33.$$

In conclusione, si può ritenere, con una fiducia del 99%, che la media della popolazione sia compresa nell'intervallo (10.67, 11.33). ■

19.4 Stima intervallare della media nel caso di grandi campioni

Come è noto, se l'ampiezza del campione è sufficientemente grande, la variabile casuale

$$Z_{\bar{X}} = \frac{\bar{X}-\mu}{S/\sqrt{n}}$$

può essere approssimata con la normale standard, qualunque sia la forma della popolazione da cui il campione è estratto (si veda la Proposizione 17.6).

PROPOSIZIONE 19.3

Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale di dimensione sufficientemente grande proveniente da una popolazione qualsiasi. Si consideri la variabile casuale

$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

la cui distribuzione può essere approssimata con la normale standard. Lo stimatore intervallare di μ, a un livello di probabilità (approssimata) $1 - \alpha$, ha estremi dati da:

$$L_1 = \bar{X} - z_{1-\alpha/2}\frac{S}{\sqrt{n}}, \quad L_2 = \bar{X} + z_{1-\alpha/2}\frac{S}{\sqrt{n}}.$$

Ne segue che gli estremi dell'intervallo di confidenza per μ, con coefficiente di fiducia (approssimato) $1 - \alpha$, sono:

$$l_1 = \bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \quad l_2 = \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}},$$

dove $\bar{x}$ e s sono la media e la deviazione standard del campione casuale osservato $(x_1, x_2, \ldots, x_n)$.

La dimostrazione è analoga a quella della Proposizione 19.1, con i necessari adattamenti. ■

Chiaramente, quest'ultimo intervallo ha una validità più ampia rispetto a quelli presentati nelle Proposizioni 19.1 e 19.2; esso richiede solamente che l'ampiezza del campione sia sufficientemente grande.

ESEMPIO 19.4

Un'indagine condotta su 160 lavoratori di una certa città indica che la distanza media tra casa e sede di lavoro è 12.5 km con una deviazione standard campionaria di 4.6 km. Si determini l'intervallo di confidenza al 99% per la media della popolazione.

Soluzione

Gli estremi dell'intervallo di confidenza sono:

$$l_1 = 12.5 - 2.576\frac{4.6}{\sqrt{160}} = 11.56, \quad l_2 = 12.5 + 2.576\frac{4.6}{\sqrt{160}} = 13.44.$$

In definitiva, si può ritenere, con una fiducia del 99%, che la distanza media tra casa e lavoro nella popolazione sia compresa nell'intervallo (11.56, 13.44). ■

Popolazioni Bernoulliane

Come si è visto nel Paragrafo 17.7.2, se n è sufficientemente grande, la distribuzione campionaria di

$$Z_{\hat{p}} = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

può essere approssimata con una normale standard. Si può allora enunciare la proposizione che segue.

PROPOSIZIONE 19.4

Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale di dimensione sufficientemente grande estratto da una popolazione Bernoulliana. La variabile casuale

$$Z_{\hat{p}} = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

ha, approssimativamente, distribuzione normale standard. Lo stimatore intervallare di p, a un livello di probabilità $1-\alpha$, ha estremi dati da:

$$L_1 = \hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad L_2 = \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

dove $\hat{p} = \sum_{i=1}^{n} X_i/n$ è la frequenza relativa dei successi nel campione. Ne segue che gli estremi dell'intervallo di confidenza per p, con coefficiente di fiducia $1 - \alpha$, sono:[4]

$$l_1 = \hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad l_2 = \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

dove $\hat{p} = \sum_{i=1}^{n} x_i/n$ è la frequenza relativa dei successi nel campione osservato.

Si consideri l'evento aleatorio

$$-z_{1-\alpha/2} < \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} < z_{1-\alpha/2}$$

di probabilità (approssimata) $1 - \alpha$. Risolvendo le due disequazioni rispetto a p, si perviene all'evento aleatorio equivalente

$$\hat{p} - z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p} + z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n},$$

cosicché si può scrivere

$$P\left[\hat{p} - z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < p < z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right] \approx 1 - \alpha.$$

Gli estremi dello stimatore intervallare di p sono allora dati da:

$$L_1 = \hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad L_2 = \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Da questi si ricavano poi facilmente gli estremi dell'intervallo di confidenza. ■

[4] Gli estremi dell'intervallo di confidenza sono formalmente identici a quelli dello stimatore intervallare, in quanto lo stesso simbolo $\hat{p}$ indica sia la variabile casuale frequenza relativa dei successi che la sua realizzazione.

ESEMPIO 19.5

Nel processo di controllo del peso delle confezioni di spaghetti da 500 g, un pastificio considera non conformi agli standard i pesi al di sotto di 485 g o al di sopra di 515 g. Per valutare l'efficienza di una confezionatrice, l'azienda esamina un campione di 830 confezioni e trova che 42 di esse hanno un peso al di fuori di detti limiti. Si determini l'intervallo di confidenza al 95% per p (proporzione nella popolazione o probabilità di osservare una confezione non conforme).

Soluzione

La stima puntuale di p è $\hat{p} = 0.051$; gli estremi dell'intervallo di confidenza sono:

$$l_1 = 0.051 - 1.96\sqrt{\frac{0.051 \cdot 0.949}{830}} = 0.036,$$

$$l_2 = 0.051 + 1.96\sqrt{\frac{0.051 \cdot 0.949}{830}} = 0.066.$$

Si può concludere, con una fiducia del 95%, che il parametro p è compreso nell'intervallo (0.036, 0.066). ■

19.5 Stima intervallare della varianza di una popolazione normale

Si consideri un campione casuale di ampiezza n estratto da una popolazione normale con media μ e varianza σ^2, entrambe incognite. È noto che la variabile casuale

$$V = \frac{(n-1)S^2}{\sigma^2}$$

ha distribuzione chi-quadrato con $n-1$ gradi di libertà (si veda il Paragrafo 17.6). La Figura 19.6 presenta il grafico della distribuzione di V con l'indicazione dei quantili $\chi^2_{\alpha/2}$ e $\chi^2_{1-\alpha/2}$ che delimitano un'area sottesa alla curva pari a $1-\alpha$.

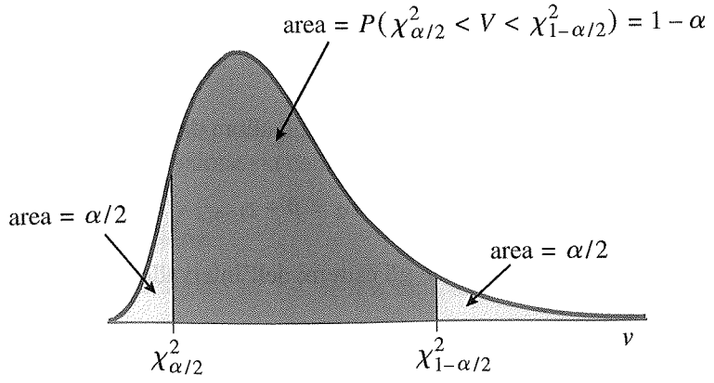


Figura 19.6
Distribuzione campionaria di V e probabilità dell'evento $\chi^2_{\alpha/2} < V < \chi^2_{1-\alpha/2}$.

La Proposizione 19.5 definisce l'intervallo di confidenza per σ^2.

PROPOSIZIONE 19.5

Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale estratto da una popolazione normale con media μ e varianza σ^2, entrambe incognite. Si consideri la variabile casuale

$$V = \frac{(n-1)S^2}{\sigma^2}$$

avente distribuzione chi-quadrato con $n-1$ gradi di libertà. Lo stimatore intervallare di σ^2, a un livello di probabilità di $1-\alpha$, ha estremi dati da:

$$L_1 = \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}, \quad L_2 = \frac{(n-1)S^2}{\chi^2_{\alpha/2}},$$

dove $\chi^2_{\alpha/2}$ e $\chi^2_{1-\alpha/2}$ sono i quantili di livello $\alpha/2$ e $1-\alpha/2$ della distribuzione chi-quadrato con $n-1$ gradi di libertà. Ne segue che gli estremi dell'intervallo di confidenza per σ^2, con coefficiente di fiducia $1-\alpha$, sono:

$$l_1 = \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}, \quad l_2 = \frac{(n-1)s^2}{\chi^2_{\alpha/2}},$$

dove s^2 è la varianza del campione osservato $(x_1, x_2, \ldots, x_n)$.

Si consideri l'evento aleatorio

$$\chi^2_{\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{1-\alpha/2}$$

di probabilità $1-\alpha$. Risolvendo le due disequazioni rispetto a σ^2, si perviene all'evento aleatorio equivalente

$$\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\alpha/2}},$$

cosicché si può scrivere

$$P\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\alpha/2}}\right] = 1-\alpha.$$

Gli estremi dello stimatore intervallare di σ^2 sono allora dati da:

$$L_1 = \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}, \quad L_2 = \frac{(n-1)S^2}{\chi^2_{\alpha/2}}.$$

Da questi si ricavano poi facilmente gli estremi dell'intervallo di confidenza. ■

ESEMPIO 19.6

Un'azienda produttrice di orologi di precisione dichiara che questi sono soggetti a un'oscillazione massima di 1 secondo al mese. Viene osservato l'errore mensile

di 20 orologi e si trova $s^2 = 0.14$. Si determini l'intervallo di confidenza al 95% per σ^2.

Soluzione

I quantili della distribuzione chi-quadrato richiesti sono $\chi^2_{0.025} = 8.91$ e $\chi^2_{0.975} = 32.9$, essendo 19 i gradi di libertà. Gli estremi dell'intervallo di confidenza sono:

$$l_1 = \frac{19 \cdot 0.14}{32.9} = 0.08, \quad l_2 = \frac{19 \cdot 0.14}{8.91} = 0.30.$$

È lecito pertanto affermare, con una fiducia del 95%, che la varianza in questione è compresa nell'intervallo (0.08, 0.30). ■

Riassunto

La costruzione di un intervallo di confidenza per un parametro consiste nell'identificare due statistiche campionarie, L_1 e L_2, con $L_1 < L_2$, tali che vi sia un'altissima probabilità che l'intervallo casuale (L_1, L_2), denominato stimatore intervallare, contenga al suo interno il parametro θ. L'intervallo di confidenza per θ si ottiene poi dallo stimatore intervallare sostituendo L_1 e L_2 con i valori da essi assunti nel campione osservato. Questa procedura generale viene applicata per ottenere l'intervallo di confidenza per: la media di una popolazione normale con varianza nota e non nota; la media di una popolazione qualsiasi nel caso di grandi campioni; la varianza di una popolazione normale.

Parole chiave

ampiezza dell'intervallo di confidenza, 406
coefficiente di fiducia, 405
intervallo di confidenza, 405
quantità pivot, 406
stimatore intervallare, 405

Commenti e curiosità

Intervalli di confidenza per la media e margine di errore

È interessante notare che gli estremi degli intervalli di confidenza concernenti il parametro media, nelle diverse situazioni esaminate, possono essere scritti nella forma $\bar{x} \pm$ *margine di errore*, dove il margine di errore (si veda il Paragrafo 18.4) è un numero noto (quando è nota la varianza della popolazione) oppure è una stima campionaria. La prossima tabella riporta le espressioni del margine di errore riferite alle diverse situazioni trattate in questo capitolo.

Popolazione	Margine di errore
Popolazione normale con varianza nota	$z_{1-\alpha/2}\,(\sigma/\sqrt{n})$
Popolazione normale con varianza non nota	$t_{1-\alpha/2}\,(s/\sqrt{n})$
Popolazione Bernoulliana e grandi campioni	$z_{1-\alpha/2}\,\sqrt{\hat{p}(1-\hat{p})/n}$
Popolazione qualsiasi e grandi campioni	$z_{1-\alpha/2}\,(s/\sqrt{n})$

Il margine di errore equivale alla metà dell'ampiezza dell'intervallo di confidenza, in quanto tale è una misura dell'accuratezza della stima dell'intervallo.

Esercizi

19.1 L'osservazione della durata (in ore) delle batterie per cellulare di una data marca in 24 esemplari di prodotto ha dato luogo ai risultati di seguito indicati:

58.7, 64.9, 76.9, 67.8, 41.7, 156.7, 64.5, 69.7, 82.1, 82.5, 40.8, 74.9,
71.5, 75.4, 67.3, 73.0, 70.4, 104.0, 82.3, 90.4, 86.8, 72.8, 71.8, 54.5.

Si determini l'intervallo di confidenza al 95% per la media della popolazione, supponendo che la popolazione sia normale con varianza $\sigma^2 = 230$.

19.2 I dati che seguono si riferiscono alla durata (in migliaia di km) di un convertitore catalitico in un campione di 15 osservazioni:

115.4, 85.2, 89.1, 118.3, 88.4, 109.3, 104.3, 69.3,
105.5, 106.8, 103.1, 101.6, 102.9, 89.6, 109.3.

Si costruisca l'intervallo di confidenza al 99% per la media, assumendo che la popolazione generatrice sia normale.

19.3 Di seguito sono riportati i chilometri percorsi in un giorno da un campione di taxi operanti in una grande città:

172, 195, 115, 122, 154, 149, 120, 148,
152, 68, 132, 91, 120, 148, 103, 101.

Sapendo che l'intervallo di confidenza per la media della popolazione ottenuto con questi dati è risultato (116.55, 144.70), si individui il coefficiente di fiducia utilizzato.

19.4 Un'azienda produce rotoli di stoffa della lunghezza di 70 m. Tali rotoli possono presentare difetti di tessitura di varia natura. L'azienda è interessata a stimare il numero medio di difetti presenti nei rotoli prodotti. In un campione casuale di 85 rotoli si è trovata la seguente distribuzione di frequenza del numero di difetti:

N. di difetti	0	1	2	3	4	5	6
Frequenza	16	26	22	13	5	2	1

Si determini l'intervallo di confidenza al 99% per la media dei difetti presenti nei rotoli di stoffa prodotti dall'azienda.

19.5 Una grande impresa di trasporti utilizza un certo tipo di pneumatici per camion di cui intende stimare la durata media (in km). A questo fine, osserva un campione di 56 pneumatici. Si costruisca l'intervallo di confidenza al 95% per la durata media degli pneumatici sapendo che nel campione la media e la deviazione standard sono risultate 43 340 e 2749, rispettivamente.

19.6 I pesi in grammi di un campione di 36 bustine di tè di una data marca sono riportati qui di seguito:

5.65, 5.34, 5.57, 5.62, 5.47, 5.32, 5.77, 5.50, 5.61,
5.45, 5.50, 5.42, 5.45, 5.53, 5.46, 5.47, 5.29, 5.42,
5.63, 5.44, 5.54, 5.40, 5.56, 5.40, 5.67, 5.57, 5.32,
5.50, 5.44, 5.57, 5.40, 5.52, 5.54, 5.44, 5.61, 5.49.

Si costruisca l'intervallo di confidenza per la media al 90% e si calcoli la sua ampiezza.

19.7 Il Direttore di un grande magazzino ha osservato i tempi di servizio alla cassa riservata ai clienti con un numero di articoli non superiori a 7 in condizione di affollamento. In 58 osservazioni ha trovato un tempo medio $\bar{x} = 5.4$ e una deviazione standard $s = 2.6$. Si determini l'ampiezza dell'intervallo di confidenza al 99% e si stimi il margine di errore.

19.8 L'ufficio statistico di un comune ha condotto un sondaggio tra i cittadini maggiorenni per accertare il gradimento di un nuovo sistema di raccolta differenziata dei rifiuti. In un campione casuale di 530 individui, l'ufficio statistico ha trovato che il 68% degli intervistati è favorevole. Si costruisca l'intervallo di confidenza al 95% per la percentuale dei favorevoli nella popolazione.

19.9 In una ricerca condotta da un centro studi sull'utilizzo di Internet è stato trovato che, sui 540 intervistati di un campione casuale, 248 ricevono oltre 15 messaggi di posta elettronica al giorno. Si determini l'intervallo di confidenza al 99% per la proporzione nella popolazione degli utilizzatori di Internet con la caratteristica sopra indicata.

19.10 Con riferimento all'Esercizio 19.4, si consideri che un rotolo è ritenuto vendibile se presenta un massimo di tre difetti. Sulla base dello stesso campione di cui all'Esercizio 19.4, si costruisca l'intervallo di confidenza al 95% per la proporzione dei rotoli prodotti considerati vendibili.

19.11 Con i dati del campione dato nell'Esercizio 19.3, si costruisca l'intervallo di confidenza al 99% per la varianza, assumendo che la popolazione sia normale.

19.12 Riprendendo i dati dell'Esercizio 19.6, si determini l'intervallo di confidenza al 95% per la varianza della popolazione, ammesso che questa sia normale.

19.13* Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale proveniente da una popolazione normale con media μ nota e varianza incognita. Si costruisca un intervallo di confidenza per σ^2. (*Suggerimento*: si consideri che la quantità $\sum_{i=1}^{n}(X_i - \mu)^2/\sigma^2$ ha distribuzione chi-quadrato con n gradi di libertà.)

CAPITOLO

20

Verifica delle ipotesi

Sommario del capitolo

20.1 Introduzione
20.2 Verifica delle ipotesi
20.3 Test di ipotesi sulla media di una popolazione normale con varianza nota
20.4 Test di ipotesi sulla media di una popolazione normale con varianza non nota
20.5 Test di ipotesi sulla media di una popolazione nel caso di grandi campioni
20.6 Test di ipotesi sulla varianza di una popolazione normale
20.7 Criteri di ottimizzazione nella verifica delle ipotesi
Riassunto
Commenti e curiosità
Esercizi

20.1 Introduzione

La verifica delle ipotesi è forse la parte dell'inferenza statistica che più rispecchia il canone del metodo scientifico per cui "una teoria è vera finché non si dimostra il contrario". Nell'accezione qui assunta, la "teoria" è rappresentata da una congettura, un'asserzione, su un parametro della popolazione di interesse.

Si consideri, ad esempio, una sperimentazione agraria con cui si intende provare la maggiore efficacia, nella coltivazione di un dato cereale, del nuovo fertilizzante F rispetto a quello tradizionalmente utilizzato. L'esperimento viene realizzato osservando la resa media (in quintali per ettaro) con l'impiego del fertilizzante F in un campione casuale di 50 particelle di terreno omogenee (per esposizione, caratteristiche del terreno ecc.). Ammesso che la resa media assicurata dal fertilizzante tradizionale (su terreni in tutto simili a quelli delle particelle sperimentali suddette) sia nota e pari a $\mu = 85$, la ricerca dovrà mettere a confronto l'ipotesi che $\mu = 85$ con l'ipotesi che $\mu > 85$. La "teoria" da "falsificare" è qui l'affermazione che l'adozione del fertilizzante A non migliora la produzione media per ettaro. In sostanza, le ipotesi a confronto sono le due asserzioni sul parametro media, $\mu = 85$ e $\mu > 85$.

Un secondo esempio può essere tratto dall'ambito aziendale. Si pensi a un processo industriale di confezionamento di pacchetti di riso. Gli standard del processo garantiscono un peso medio per pacchetto pari a 500 g. Il responsabile della qualità ritiene che il processo non sia in controllo, nel senso che produce confezioni di peso medio superiore a 500 g, con conseguente spreco di risorse. Egli procede all'osservazione di un campione casuale di 100 confezioni. Anche in questo caso, l'ipotesi da verificare riguarda il parametro media. Vi sono due asserzioni (ipotesi) riguardo al parametro media: $\mu = 500$ e $\mu > 500$.

Si immagini ora che l'azienda meccanica A utilizzi, per una sua produzione, i semilavorati dell'azienda meccanica B. Il contratto di fornitura prevede che la percentuale massima di pezzi difettosi tollerata sia del 2%. Alla consegna di ogni lotto, l'ufficio qualità dell'azienda A preleva un campione casuale di 300 unità e determina la percentuale di pezzi difettosi nel campione allo scopo di stabilire se l'accordo contrattuale è rispettato. In termini statistici, il parametro è la proporzione p di pezzi difettosi; l'ipotesi da verificare è la congettura che tale percentuale sia inferiore o uguale al 2%. Dunque, le ipotesi a confronto qui sono le due asserzioni riguardo al parametro p: $p \leq 0.02$ e $p > 0.02$.

In generale, formulare un'ipotesi statistica consiste nell'assegnare un valore o un insieme di valori a un parametro della popolazione di interesse, come illustrato negli esempi.

DEFINIZIONE 20.1

Sia θ il parametro di interesse, cioè la caratteristica di interesse della popolazione indagata. Un'ipotesi statistica è un'affermazione su θ.

Un'ipotesi statistica è dunque un'asserzione del tipo: $\theta = \theta_0$, $\theta \geq \theta_0$, $\theta < \theta_0$, essendo θ_0 un determinato valore del parametro. Un'ipotesi che specifica un singolo valore di θ è detta semplice, mentre un'ipotesi che specifica un insieme di valori di θ è detta composta. È semplice, ad esempio, l'ipotesi $\mu = 85$; è invece composta l'ipotesi $\mu > 85$.

L'ipotesi sottoposta a verifica va sotto il nome di ipotesi nulla ed è indicata con H_0; l'ipotesi che contraddice H_0 è chiamata ipotesi alternativa ed è indicata con H_1. Così, nel primo esempio, l'ipotesi nulla è $\mu = 85$ e l'ipotesi alternativa è $\mu > 85$. In molte applicazioni, come l'espressione suggerisce, l'ipotesi nulla esprime la negazione della tesi della ricerca.

Questo capitolo è così organizzato. Saranno anzitutto introdotti i concetti che sono alla base delle procedure di verifica delle ipotesi. Verranno poi illustrate le tecniche di verifica delle ipotesi sui i parametri media, proporzione e varianza. Infine, saranno delineati gli elementi essenziali della teoria della verifica delle ipotesi. In questa discussione, verrà seguita l'impostazione di Neyman-Pearson, ma si farà anche uso del livello di significatività osservato, uno strumento proprio dell'impostazione alternativa dovuta a Fisher (per un approfondimento sui differenti paradigmi filosofici alla base della verifica delle ipotesi, si rinvia a Spanos, 1999, p. 681 e segg.).

20.2 Verifica delle ipotesi

Per introdurre gli elementi chiave della verifica delle ipotesi, si farà riferimento a un solo parametro di una singola popolazione. Inoltre, salvo indicazioni in senso contrario, l'ipotesi nulla avrà la forma $H_0 : \theta = \theta_0$, mentre l'ipotesi alternativa assumerà una delle forme indicate qui di seguito:

- ipotesi alternativa unidirezionale a destra, $\theta > \theta_0$;
- ipotesi alternativa unidirezionale a sinistra, $\theta < \theta_0$;
- ipotesi alternativa bidirezionale, $\theta \neq \theta_0$.

DEFINIZIONE 20.2

Per test statistico o test di ipotesi si intende il procedimento con cui, sulla base di un campione casuale osservato, si decide se rifiutare o non rifiutare l'ipotesi nulla.

L'elemento basilare della verifica delle ipotesi è la statistica test, che generalmente è uno stimatore del parametro θ, qui indicato con

$$W = w(X_1, X_2, \ldots, X_n),$$

dove $(X_1, X_2, \ldots, X_n)$ è un campione casuale. Come stimatore del parametro θ, W è una variabile casuale con una distribuzione di probabilità che dipende da θ. La

decisione se rifiutare o no l'ipotesi nulla è basata sul valore che la statistica test assume nel campione osservato. Se H_0 è vera, la statistica test dovrebbe assumere un valore, w, vicino a θ_0. Infatti, se w è sufficientemente vicino a θ_0, la differenza può essere attribuita al caso e quindi l'ipotesi nulla non viene rifiutata. Al contrario, forti deviazioni da θ_0, rendendo poco plausibile l'ipotesi nulla, portano a rigettarla.

L'insieme dei possibili valori della statistica test è suddiviso in due sottoinsiemi disgiunti chiamati zona di rifiuto e zona di non rifiuto. La prima contiene i valori di W che sono "molto distanti" da θ_0, la seconda tutti gli altri. Il significato dell'espressione "molto distante" dipende dal tipo di ipotesi alternativa:

- per l'ipotesi alternativa $\theta > \theta_0$, un valore di W è considerato molto distante da θ_0 se esso è "molto più grande" di θ_0;
- per l'ipotesi alternativa $\theta < \theta_0$, un valore di W è considerato molto distante da θ_0 se esso è "molto più piccolo" di θ_0;
- per l'ipotesi alternativa $\theta \neq \theta_0$, un valore di W è considerato molto distante da θ_0 se esso è molto più grande o molto più piccolo di θ_0.

Il valore (i valori) che separa (separano) la zona di rifiuto da quella di non rifiuto è noto (sono noti) come valore critico (valori critici). La decisione di rifiutare o non rifiutare l'ipotesi nulla è presa in condizioni di incertezza, in quanto dipende dai dati osservati nel campione casuale. Ne deriva che essa è soggetta a due tipi di errore: può essere rifiutata un'ipotesi vera o non rifiutata un'ipotesi falsa.

DEFINIZIONE 20.3

Si chiama errore di I tipo la decisione di rifiutare H_0 quando è vera. Si chiama errore di II tipo la decisione di non rifiutare H_0 quando è falsa.

Nella tabella che segue, sono descritte le quattro possibili combinazioni tra gli stati di natura di H_0 (vera o falsa) e le decisioni su H_0 (rifiuto o non rifiuto):

Decisione su H_0	Stato di natura	
	H_0 è vera	H_0 è falsa
Rifiuto	Errore di I tipo	Corretto rifiuto
Non rifiuto	Corretto non rifiuto	Errore di II tipo

Nel decidere se rifiutare o non rifiutare l'ipotesi nulla, si possono commettere due possibili errori, l'errore di I tipo, quando si rifiuta un'ipotesi nulla vera, e l'errore di II tipo, quando non si rifiuta un'ipotesi nulla falsa.

La probabilità di commettere un errore di I tipo, indicata con α e chiamata livello di significatività, è definita come la probabilità che la statistica test appartenga alla zona di rifiuto quando il parametro θ è uguale a θ_0, valore specificato dall'ipotesi nulla. In simboli, si ha

$$\alpha = P(W \in R\,;\, \theta = \theta_0),$$

dove R indica la zona di rifiuto.

La probabilità di commettere un errore di II tipo, indicata con $\beta(\theta)$, è definita come la probabilità che la statistica test non appartenga alla zona di rifiuto quando il parametro θ assume un valore tra quelli postulati dall'ipotesi alternativa. In simboli, si ha

$$\beta(\theta) = P(W \notin R\,;\, H_1),$$

dove la scrittura $\beta(\theta)$ indica che la probabilità dell'errore di II tipo è funzione del valore che θ assume all'interno dell'insieme definito dall'ipotesi alternativa (per esempio, $\theta > \theta_0$).

Il complemento a 1 di $\beta(\theta)$, denominato potenza del test e indicato con $\pi(\theta)$, è la probabilità di rifiutare l'ipotesi nulla quando essa è falsa. In simboli, $\pi(\theta)$ è espressa da

$$\pi(\theta) = 1 - \beta(\theta) = P(W \in R\,;\, H_1).$$

ESEMPIO 20.1

Queste idee sono illustrate qui di seguito con riferimento all'esempio del processo di riempimento di pacchetti di riso (si veda il Paragrafo 20.1). Come si è detto, l'ufficio qualità ritiene che il processo tenda a sovra riempire i pacchetti. Il problema è dunque quello di sottoporre a verifica l'ipotesi nulla $\mu = 500$ contro l'alternativa $\mu > 500$. Si supponga che il peso del sacchetto segua una distribuzione normale con varianza 35. Per condurre il test di ipotesi, l'ufficio qualità decide di:

- fissare la dimensione campionaria a 40;
- scegliere come statistica test la media campionaria, $W = \bar{X}$;
- rifiutare l'ipotesi nulla se la media campionaria è maggiore di 501.5 g.

Così, la zona di rifiuto è costituita dai valori della media campionaria che superano il valore critico 501.5. Supponendo che l'ipotesi nulla sia vera (processo in controllo), si commetterà un errore di I tipo se la statistica test supera questa soglia. Ciò potrà accadere con probabilità

$$\alpha = P(\bar{X} > 501.5\,;\, \mu = 500) = P\left(\frac{\bar{X} - 500}{\sqrt{35/40}} > \frac{501.5-500}{\sqrt{35/40}}\,;\, \mu = 500\right) = 0.0548.$$

All'opposto, ammettendo che l'ipotesi nulla sia falsa, cioè che il peso medio dei pacchetti sia effettivamente maggiore di 500, si incorrerà nell'errore di II tipo se la statistica test assume un valore al di sotto del valore critico 501.5. La probabilità di questo evento dipende dal valore che μ assume all'interno dell'insieme specificato dall'ipotesi alternativa ($\mu > 500$). Per esempio, se $\mu = 500.5$, si ha

$$\beta(500.5) = P(\bar{X} < 501.5\,;\, \mu{=}500.5) = P\left(\frac{\bar{X}-500.5}{\sqrt{35/40}} < \frac{501.5-500.5}{\sqrt{35/40}}\,;\, \mu{=}500.5\right)$$
$$= 0.8577.$$

Se $\mu = 501$, si ha

$$\beta(501) = P(\bar{X} < 501.5\,;\, \mu = 501) = P\left(\frac{\bar{X}-501}{\sqrt{35/40}} < \frac{501.5-501}{\sqrt{35/40}}\,;\, \mu = 501\right)$$
$$= 0.7019.$$

Chiaramente, la probabilità $\beta(\mu)$ è tanto minore quanto maggiore è il valore del parametro μ.

La potenza del test, ad esempio per $\mu = 501$, è data da

$$\pi(501) = P(\bar{X} > 501.5\,;\, \mu = 501) = P\left(\frac{\bar{X}-501}{\sqrt{35/40}} > \frac{501.5-501}{\sqrt{35/40}}\,;\, \mu = 501\right)$$
$$= 1 - 0.7019 = 0.2981.$$

■

Interpretando la probabilità come frequenza relativa nel lungo andare, α può essere pensata come la frequenza relativa dei casi in cui H_0 viene rifiutata quando il test di ipotesi viene ripetuto molte e molte volte su campioni indipendenti della stessa dimensione estratti dalla medesima popolazione (dove $\theta = \theta_0$). Analogamente, $\beta(\theta)$ può essere pensata come frequenza relativa nel lungo andare dei casi in cui H_0 non viene rifiutata quando il parametro θ assume un valore appartenente all'insieme specificato da H_1.

La bontà di un test di ipotesi basato sulla statistica W è determinata dai livelli di probabilità α e β o π. Sarebbe desiderabile poter minimizzare simultaneamente sia α sia β. Sfortunatamente, ciò non è possibile.

Siccome l'errore di I tipo è generalmente più rilevante che non quello di II tipo, la soluzione che si adotta è quella di fissare il valore di α e di ricercare la zona di rifiuto che ne deriva. Valori standard di α sono: 0.05, 0.01, 0.005, 0.001.

La zona di rifiuto del test di ipotesi si determina assumendo che l'ipotesi nulla sia vera.

Si ricordi che la regola decisionale del test di ipotesi è di rifiutare l'ipotesi nulla quando la statistica test W assume un valore molto lontano dal valore θ_0 (contemplato da H_0) nella direzione dell'ipotesi alternativa, e che la zona di rifiuto è tale che la probabilità dell'errore di I tipo sia uguale al livello prefissato α. Qui di seguito, viene descritto in generale il procedimento con cui viene costruita la zona di rifiuto.

Ipotesi alternativa unidirezionale a destra

$$H_0: \theta = \theta_0 \text{ contro } H_1: \theta > \theta_0.$$

La zona di rifiuto è costituita dai valori della statistica test W che sono al di sopra della costante c:

$$\{w : w > c\}, \tag{20.1}$$

dove c è il valore critico definito dalla relazione

$$P(W > c\,;\, \theta_0) = \alpha.$$

Ipotesi alternativa unidirezionale a sinistra

$$H_0: \theta = \theta_0 \text{ contro } H_1: \theta < \theta_0.$$

La zona di rifiuto è costituita dai valori della statistica test W al di sotto della costante c:

$$\{w : w < c\}, \tag{20.2}$$

dove c è il valore critico definito dalla relazione

$$P(W < c\,;\, \theta_0) = \alpha.$$

Ipotesi alternativa bidirezionale

$$H_0: \theta = \theta_0 \text{ contro } H_1: \theta \neq \theta_0.$$

La zona di rifiuto è costituita dai valori della statistica test W che sono al di sotto della costante c_1 o al di sopra della costante c_2:

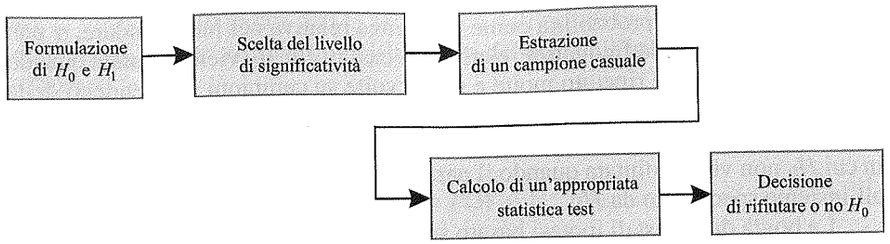
$$\{w : w < c_1 \text{ o } w > c_2\}, \tag{20.3}$$

dove c_1 e c_2 sono i valori critici definiti dalla relazione

$$P(W < c_1\,;\, \theta_0) + P(W > c_2\,;\, \theta_0) = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha.$$

La Figura 20.1 illustra i passi da compiere per effettuare un test di ipotesi.

Figura 20.1
Passi da compiere per condurre il test di ipotesi.



ESEMPIO 20.2

Si riprenda l'Esempio 20.1 e si verifichi l'ipotesi $H_0: \mu = 500$ contro l'alternativa $H_1: \mu > 500$ al livello di significatività $\alpha = 0.01$, sapendo che la media calcolata su un campione di ampiezza 40 è pari a 502.9.

Soluzione

La statistica test è la media campionaria, la cui distribuzione campionaria, sotto l'ipotesi nulla, è normale con media 500 e varianza $\sigma^2 = 35/40$. Il valore critico c può essere ottenuto come segue

$$P(\bar{X} > c\,;\, \mu = 500) = P\left(\frac{\bar{X} - 500}{\sqrt{35/40}} > \frac{c - 500}{\sqrt{35/40}}; \mu = 500\right) = 0.01.$$

Non rifiutare l'ipotesi nulla non equivale ad accettarla: vuol dire solo che nei dati campionari non c'è evidenza sufficiente per rifiutarla.

Poiché, sotto l'ipotesi nulla, $(\bar{X} - 500)/\sqrt{35/40}$ ha distribuzione normale standard, si può porre

$$\frac{c - 500}{\sqrt{35/40}} = 2.326,$$

da cui

$$c = 500 + 2.326\sqrt{35/40} = 502.18.$$

La zona di rifiuto è allora

$$\{\bar{x}: \bar{x} > 502.18\}.$$

La media campionaria osservata, 502.90, cade nella zona di rifiuto. Ne segue il rifiuto dell'ipotesi nulla. ■

Quando la statistica test cade nella zona di rifiuto, si usano spesso espressioni del tipo "i dati osservati contengono sufficiente evidenza contro l'ipotesi nulla".

PROPOSIZIONE 20.1

La zona di rifiuto del test di ipotesi unidirezionale a destra, costruito per verificare l'ipotesi nulla $H_0: \theta = \theta_0$ e dato dalla (20.1), può essere anche utilizzato per verificare l'ipotesi nulla

$$H_0: \theta \le \theta_0.$$

Analogamente, la zona di rifiuto del test di ipotesi unidirezionale a sinistra, costruito per verificare l'ipotesi nulla $H_0: \theta = \theta_0$ e data dalla (20.2), può essere anche applicata per la verifica dell'ipotesi nulla

$$H_0: \theta \ge \theta_0.$$

La logica sottostante alle regole di decisione sopra indicate è la seguente. Nel caso del test di ipotesi unidirezionale a destra, ogni valore osservato w della statistica test W qualificato come molto più grande di θ_0, e come tale appartenente alla zona di rifiuto, è *a fortiori* molto più grande di θ', dove θ' è un qualsiasi valore del parametro minore di θ_0. Un analogo ragionamento vale per il test di ipotesi unidirezionale a sinistra. ■

Vale la pena di osservare che la probabilità dell'errore di I tipo è esattamente uguale ad α quando le zone di rifiuto (20.1) e (20.2) sono utilizzate per verificare l'ipotesi nulla $H_0: \theta = \theta_0$. La probabilità dell'errore di I tipo associata alle anzidette zone di rifiuto è minore o uguale ad α se la prima è impiegata per la verifica dell'ipotesi nulla $H_0: \theta \le \theta_0$ e la seconda per la verifica dell'ipotesi nulla $H_0: \theta \ge \theta_0$. Ciò è illustrato nell'esempio che segue.

ESEMPIO 20.3

Si sottoponga a test l'ipotesi che la spesa mensile media μ per il telefonino dei giovani di età tra i 18 e 30 anni sia minore o uguale a 30 euro contro l'alternativa che sia maggiore, sulla base delle seguenti 15 osservazioni campionarie:

43.5, 32, 26.4, 37.8, 25, 31, 46.8, 19,
28.5, 35, 23.7, 32.5, 37.9, 36, 31.4.

Si assuma che la popolazione sia normale con varianza 50 e si ponga $\alpha = 0.05$.

Soluzione

L'ipotesi nulla è $H_0: \mu \le 30$, mentre l'ipotesi alternativa è $H_1: \mu > 30$. In base alla Proposizione 20.1, la zona di rifiuto si determina con il procedimento illustrato nell'Esempio 20.2, ponendo

$$c = 30 + 1.645\sqrt{50/15} = 33.$$

Cosicché la zona di rifiuto è data da

$$\{\bar{x}: \bar{x} > 33\}.$$

La media campionaria è uguale a 32.4 e non appartiene alla zona di rifiuto. Pertanto, l'ipotesi nulla non viene rifiutata.

Va osservato che il valore critico $c = 33$ è stato determinato assumendo che $\mu = 30$. Ne segue che la probabilità dell'errore di I tipo soddisfa l'equazione

$$P(\bar{X} > 33\,;\, \mu = 30) = 0.05.$$

È facile stabilire che la probabilità dell'errore di I tipo è al di sotto di 0.05 per ogni altro valore di θ appartenente all'insieme specificato da H_0. Ad esempio, se μ fosse 29, la probabilità dell'errore di I tipo sarebbe

$$P(\bar{X} > 33\,;\, \mu = 29) = P\left(\frac{\bar{X} - 29}{\sqrt{35/40}} > \frac{33 - 29}{\sqrt{50/15}}; \mu = 29\right) = 0.014.$$

La conclusione è che la probabilità dell'errore di I tipo del test di ipotesi qui esaminato è minore o uguale a 0.05. ■

Di seguito viene introdotto un altro concetto cruciale nella verifica delle ipotesi, quello di *livello di significatività osservato.*

DEFINIZIONE 20.4

Il livello di significatività osservato – in simboli α_{oss} – è la probabilità che, assunta come vera l'ipotesi nulla, la statistica test W contraddica H_0 con la stessa forza del valore w osservato nel campione o ancora più fortemente.

Nei prossimi paragrafi sarà reso esplicito il procedimento di calcolo del livello di significatività osservato per gli specifici problemi di verifica di ipotesi di volta in volta affrontati.

ESEMPIO 20.4

Si calcoli il livello di significatività osservato per il test di ipotesi considerato nell'Esempio 20.2.

Soluzione

La media campionaria osservata è 502.9. Pertanto, in base alla Definizione 20.4, si ha

$$\alpha_{oss} = P(\bar{X} \geq 502.9; \mu = 500) = P\left(\frac{\bar{X}-500}{\sqrt{35/40}} \geq \frac{502.9-500}{\sqrt{35/40}}; \mu = 500\right)$$
$$= 0.0010.$$

Il risultato si può così commentare. Ammesso che l'ipotesi nulla sia vera, la probabilità che la statistica test $\bar{X}$ assuma un valore distante dall'ipotesi nulla come quello osservato nel campione, $\bar{x} = 502.9$, o ancora più distante è molto piccola. Ciò significa che il valore della statistica test osservato nel campione è estremamente distante dal valore specificato dall'ipotesi nulla. ■

Chiaramente, un piccolo α_{oss} rende l'ipotesi nulla implausibile e conduce al suo rifiuto.

PROPOSIZIONE 20.2

Si può dimostrare che il livello di significatività osservato è minore di α se e solo se la statistica test cade nella zona di rifiuto. Ne segue che un criterio alternativo per decidere se rifiutare l'ipotesi nulla è quello di confrontare α_{oss} con α e di rifiutare l'ipotesi nulla se il primo è minore del secondo. ■

Nelle applicazioni pratiche, il criterio del livello di significatività osservato è molto più usato rispetto a quello basato sulla zona di rifiuto. Va osservato che i principali *software* statistici (SAS, SPSS ecc.) adottano questo criterio. Tale preferenza è forse dovuta al fatto che il livello di significatività osservato esprime in modo più immediato quanto è forte l'evidenza contro l'ipotesi nulla contenuta nel campione.

ESEMPIO 20.5

Il livello di significatività osservato calcolato nell'Esempio 20.4, 0.0010, è minore del livello di significatività, $\alpha = 0.01$. L'ipotesi nulla viene quindi rifiutata. È la stessa decisione presa nell'Esempio 20.2 in base alla zona di rifiuto. ■

20.3 Test di ipotesi sulla media di una popolazione normale con varianza nota

Si consideri una popolazione normale con varianza nota. Sia $H_0 : \mu = \mu_0$ l'ipotesi nulla, dove μ_0 è un valore specifico. L'ipotesi alternativa può assumere una delle seguenti forme: $\mu > \mu_0$ (test di ipotesi unidirezionale a destra); $\mu < \mu_0$ (test di ipotesi unidirezionale a sinistra); $\mu \neq \mu_0$ (test di ipotesi bidirezionale).

La statistica test per la verifica delle ipotesi sulla media della popolazione è la media campionaria $\bar{X}$, che è lo stimatore più efficiente di μ (si veda il Paragrafo 18.3). Se l'ipotesi nulla $H_0 : \mu = \mu_0$ è vera, $\bar{X}$ ha distribuzione normale con media μ_0 e varianza σ^2/n, cosicché $(\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ ha distribuzione normale standard.

Ipotesi alternativa unidirezionale a destra

$$H_0 : \mu = \mu_0 \text{ contro } H_1 : \mu > \mu_0.$$

In questo caso, la zona di rifiuto (20.1) assume la forma

$$\{\bar{x} : \bar{x} > c\},$$

dove c è tale che

$$P(\bar{X} > c\,;\, \mu_0) = \alpha,$$

da cui

$$P(\bar{X} > c\,;\, \mu_0) = P\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > \frac{c-\mu_0}{\sigma/\sqrt{n}}; \mu_0\right) = \alpha.$$

Se $\mu = \mu_0$, $(\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ ha distribuzione normale standard. Ne segue che deve essere

$$\frac{c-\mu_0}{\sigma/\sqrt{n}} = z_{1-\alpha},$$

dove $z_{1-\alpha}$ è il quantile di livello $1-\alpha$ della distribuzione normale standard. Risolvendo questa equazione rispetto a c, si ottiene

$$c = \mu_0 + z_{1-\alpha}\,\sigma/\sqrt{n}.$$

La zona di rifiuto è allora

$$\{\bar{x} : \bar{x} > \mu_0 + z_{1-\alpha}\,\sigma/\sqrt{n}\}.$$

Passando da $\bar{X}$ alla statistica test standardizzata $Z_{\bar{X}} = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$, la zona di rifiuto diviene

$$\{z_{\bar{x}} : z_{\bar{x}} > z_{1-\alpha}\}, \tag{20.4}$$

dove $z_{\bar{x}}$ è il valore osservato di $(\bar{X} - \mu_0)/(\sigma/\sqrt{n})$. Ciò segue dalla relazione

$$\bar{x} > \mu_0 + z_{1-\alpha}\sigma/\sqrt{n} \quad \Leftrightarrow \quad z_{\bar{x}} > z_{1-\alpha}.$$

La zona di rifiuto (20.4) è illustrata nella Figura 20.2.

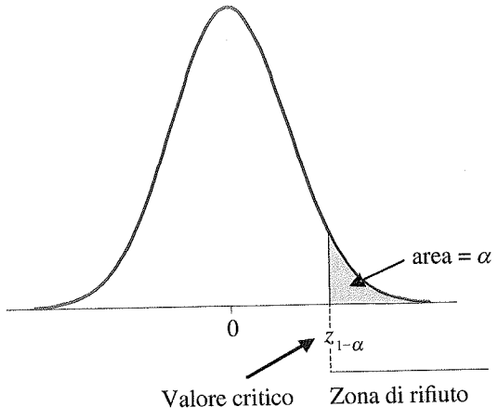


Figura 20.2
Zona di rifiuto della statistica test $Z_{\bar{X}}$; ipotesi alternativa unidirezionale a destra.

ESEMPIO 20.6

Un'azienda produttrice di sigarette sostiene che il contenuto di nicotina della marca M non supera 1.5 mg per sigaretta. Si verifichi l'ipotesi $H_0: \mu \leq 1.5$, sapendo che il contenuto medio di nicotina osservato in un campione casuale di 36 sigarette di quella marca è risultato pari a 1.56. Si assuma che la popolazione sia normale con deviazione standard $\sigma = 0.15$ e si ponga $\alpha = 0.01$.

Soluzione

La zona di rifiuto è

$$\{z_{\bar{x}} : z_{\bar{x}} > 2.326\}.$$

La statistica test assume il valore

$$z_{\bar{x}} = \frac{1.56 - 1.5}{0.15/\sqrt{36}} = 2.4,$$

che appartiene alla zona di rifiuto. Si può dunque concludere che vi è sufficiente evidenza per rifiutare l'ipotesi nulla. ■

Si osservi che il valore critico $z_{1-\alpha}$ è tanto maggiore quanto minore è α. Per esempio, con $\alpha = 0.05$, l'ipotesi nulla viene rifiutata se la statistica test supera 1.645, mentre con $\alpha = 0.01$ questo limite diviene 2.326. Come si vede, nel secondo caso si richiede che lo scarto tra media osservata e media ipotizzata sia sensibilmente superiore che non nel primo. Dunque, l'evidenza empirica contro l'ipotesi nulla richiesta per la decisione di rifiuto deve essere tanto più forte quanto minore è il livello di significatività α.

I valori critici che corrispondono ai livelli di significatività più frequentemente usati, 0.05, 0.01, 0.005, 0.001, sono: 1.645, 2.326, 2.576, 3.090.

Ipotesi alternativa unidirezionale a sinistra

$$H_0: \mu = \mu_0 \text{ contro } H_1: \mu < \mu_0.$$

In questo caso, la zona di rifiuto (20.2) diviene

$$\{\bar{x} : \bar{x} < c\},$$

dove il valore critico c soddisfa l'equazione

$$P(\bar{X} < c\,;\, \mu_0) = P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < \frac{c - \mu_0}{\sigma/\sqrt{n}}\,;\, \mu_0\right) = \alpha.$$

Se l'ipotesi nulla è vera, $Z_{\bar{X}} = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ ha distribuzione normale standard; perciò deve essere

$$\frac{c - \mu_0}{\sigma/\sqrt{n}} = -z_{1-\alpha},$$

da cui si ricava

$$c = \mu_0 - z_{1-\alpha}\,\sigma/\sqrt{n}.$$

La zona di rifiuto è allora data da

$$\{\bar{x} : \bar{x} < \mu_0 - z_{1-\alpha}\,\sigma/\sqrt{n}\}.$$

Con la statistica test standardizzata, la zona di rifiuto diviene

$$\{z_{\bar{x}} : z_{\bar{x}} < -z_{1-\alpha}\}$$

ed è illustrata nella Figura 20.3.

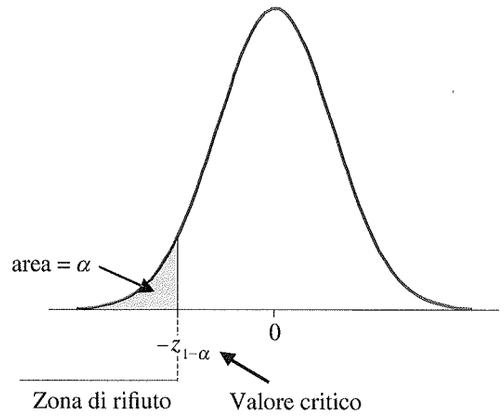


Figura 20.3
Zona di rifiuto della statistica test $Z_{\bar{X}}$; ipotesi alternativa unidirezionale a sinistra.

ESEMPIO 20.7

Il laboratorio di analisi di un'azienda produttrice di farmaci esamina un campione di 32 compresse di un nuovo farmaco per verificare l'ipotesi che la concentrazione media del principio attivo sia 1.25% (come previsto dalle specifiche del processo di produzione) contro l'alternativa che sia minore. La deviazione standard della popolazione è conosciuta ed è pari a $\sigma = 0.0076$. Nel campione la concentrazione media è $\bar{x} = 1.247$. Il livello di significatività è fissato al 5%.

Soluzione

La zona di rifiuto è data da

$$\{z_{\bar{x}} : z_{\bar{x}} < -1.645\}.$$

La statistica test assume il valore

$$z_{\bar{x}} = \frac{1.247 - 1.25}{0.0076/\sqrt{32}} = -2.23.$$

Poiché

$$z_{\bar{x}} = -2.23 < -1.645 = -z_{0.95},$$

l'ipotesi nulla viene rifiutata. ■

Ipotesi alternativa bidirezionale

$$H_0 : \mu = \mu_0 \text{ contro } H_1 : \mu \neq \mu_0.$$

In questo caso, l'ipotesi nulla viene rifiutata per valori di $\bar{X}$ molto più grandi o molto più piccoli di μ_0. In base alla (20.3), vanno determinate le costanti c_1 e c_2 tali che:

$$P(\bar{X} < c_1 ; \mu_0) = \frac{\alpha}{2}, \quad P(\bar{X} > c_2 ; \mu_0) = \frac{\alpha}{2},$$

da cui, con lo stesso procedimento seguito in precedenza, si ottiene:

$$c_1 = \mu_0 - z_{1-\alpha/2}\sigma/\sqrt{n}, \quad c_2 = \mu_0 + z_{1-\alpha/2}\sigma/\sqrt{n}.$$

La zona di rifiuto è allora data da

$$\{\bar{x} : \bar{x} \notin (\mu_0 - z_{1-\alpha/2}\sigma/\sqrt{n},\ \mu_0 + z_{1-\alpha/2}\sigma/\sqrt{n}\}.$$

Con la statistica test standardizzata, si ha

$$\{z_{\bar{x}} : |z_{\bar{x}}| > z_{1-\alpha/2}\}$$

(si veda la Figura 20.4).

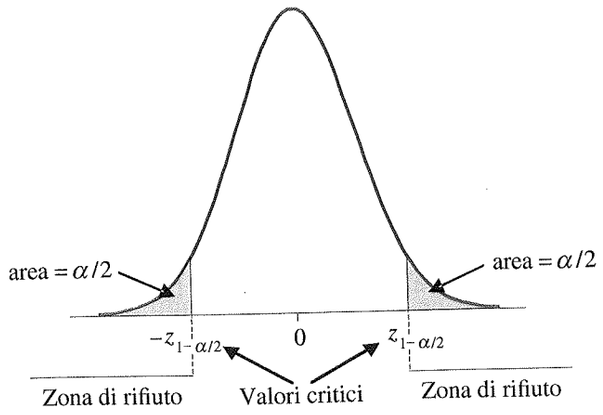


Figura 20.4
Zona di rifiuto della statistica test $Z_{\bar{X}}$; ipotesi alternativa bidirezionale.

ESEMPIO 20.8

Riprendendo i dati dell'Esempio 20.7, si verifichi l'ipotesi $H_0 : \mu = 1.25$ contro l'alternativa $H_1 : \mu \neq 1.25$, al livello di significatività $\alpha = 0.05$.

Soluzione

La zona di rifiuto è data da

$$\{z_{\bar{x}} : |z_{\bar{x}}| > 1.96\}.$$

La statistica test assume il valore

$$z_{\bar{x}} = \frac{1.247 - 1.25}{0.0076/\sqrt{32}} = -2.23.$$

Poiché

$$|z_{\bar{x}}| = 2.23 > 1.96 = z_{0.975},$$

si può concludere che vi è evidenza sufficiente per rifiutare l'ipotesi nulla. ■

Le zone di rifiuto illustrate sopra sono riassunte nella Tabella 20.1.

Ipotesi H_0	Ipotesi H_1	Statistica test	Zona di rifiuto
$\mu = \mu_0$	$\mu > \mu_0$	$Z_{\bar{X}} = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$\{z_{\bar{x}} : z_{\bar{x}} > z_{1-\alpha}\}$
	$\mu < \mu_0$		$\{z_{\bar{x}} : z_{\bar{x}} < -z_{1-\alpha}\}$
	$\mu \neq \mu_0$		$\{z_{\bar{x}} : \lvert z_{\bar{x}}\rvert > z_{1-\alpha/2}\}$

Tabella 20.1
Zone di rifiuto della statistica test $Z_{\bar{X}} = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ per la verifica delle ipotesi sulla media di una popolazione normale con varianza nota.

Per la Proposizione 20.1, i due test unidirezionali illustrati sopra per l'ipotesi nulla semplice $\mu = \mu_0$ sono validi anche nel caso di ipotesi nulla composta ($\mu \leq \mu_0$ o $\mu \geq \mu_0$), come già illustrato con l'Esempio 20.6.

20.3.1 Livello di significatività osservato

Qui di seguito, viene data l'espressione esplicita del livello di significatività osservato per ciascuno dei tre test di ipotesi, applicando la Definizione 20.4 alla statistica test standardizzata $Z_{\bar{X}}$.

Ipotesi alternativa unidirezionale a destra

$$\alpha_{oss} = P(Z_{\bar{X}} \geq z_{\bar{x}};\ \mu = \mu_0) = 1 - \Phi(z_{\bar{x}}),$$

dove $\Phi(\cdot)$ è la funzione di ripartizione della normale standard.

Ipotesi alternativa unidirezionale a sinistra

$$\alpha_{oss} = P(Z_{\bar{X}} \leq z_{\bar{x}};\ \mu = \mu_0) = \Phi(z_{\bar{x}}).$$

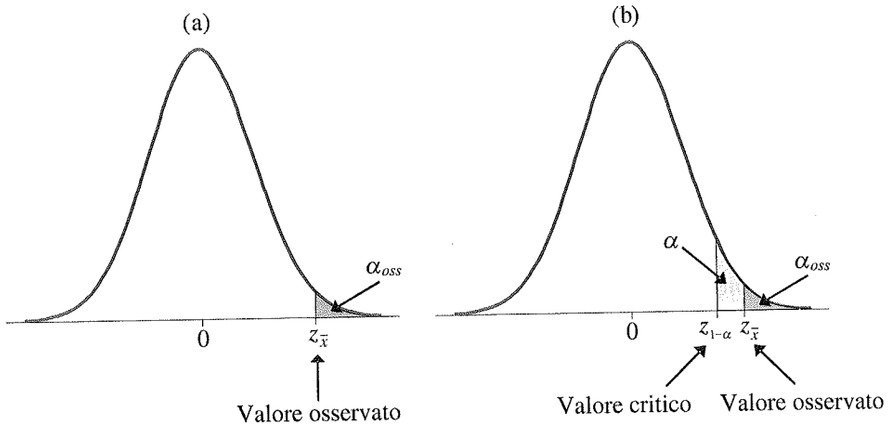
Ipotesi alternativa bidirezionale

$$\alpha_{oss} = P(|Z_{\bar{X}}| \geq |z_{\bar{x}}|;\ \mu = \mu_0) = 2[1 - \Phi(|z_{\bar{x}}|)].$$

Come indicato nella Proposizione 20.2, rifiutare l'ipotesi nulla quando α_{oss} è minore di α equivale a rifiutare l'ipotesi nulla quando la statistica test cade nella zona di rifiuto. Ciò è illustrato nella Figura 20.5 per il test di ipotesi unidirezionale a

destra. In particolare, nella Figura 20.5 (a), viene mostrato che α_{oss} è uguale all'area al di sotto della curva normale standard a destra della statistica test osservata $z_{\bar{x}}$, nella Figura 20.5 (b) invece viene illustrato che α_{oss} è minore di α se e solo se $z_{\bar{x}}$ appartiene alla zona di rifiuto.

Figura 20.5
Livello di significatività osservato, nel caso di ipotesi alternativa unidirezionale a destra, nella verifica delle ipotesi sulla media di una popolazione normale con varianza nota.



ESEMPIO 20.9

Riprendendo l'Esempio 20.6, si proceda alla verifica d'ipotesi mediante il livello di significatività osservato.

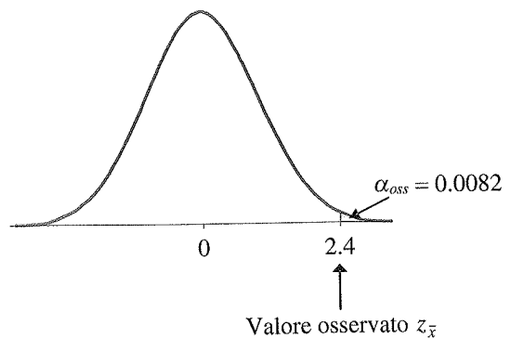
Soluzione

La statistica test è pari a $z_{\bar{x}} = 2.4$. Il livello di significatività osservato è allora

$$\alpha_{oss} = P(Z_{\bar{X}} \geq 2.4; \mu = 1.5) = 1 - \Phi(2.4) = 0.0082,$$

un valore minore del livello di significatività $\alpha = 0.01$ (si veda la Figura 20.6). Ne segue che l'ipotesi nulla viene rifiutata, così come avvenuto nell'Esempio 20.6 mediante la zona di rifiuto.

Figura 20.6
Livello di significatività osservato per il problema di verifica delle ipotesi dell'Esempio 20.6.



■

ESEMPIO 20.10

Si riprenda l'Esempio 20.7 e si ripeta la verifica delle ipotesi con il livello di significatività osservato.

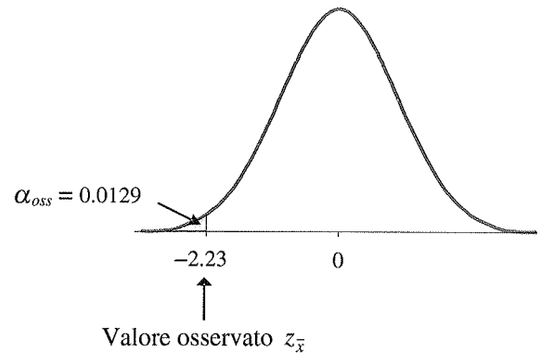
Soluzione

Il livello di significatività osservato è dato da

$$\alpha_{oss} = P(Z_{\bar{X}} \leq -2.23; \mu = 1.25) = \Phi(-2.23) = 0.0129.$$

La conclusione è quella di rifiutare l'ipotesi nulla in linea ovviamente con la decisione assunta nell'Esempio 20.7 (si veda la Figura 20.7).

Figura 20.7
Livello di significatività osservato per il problema dell'Esempio 20.7.



■

ESEMPIO 20.11

Si riprenda l'Esempio 20.8 e si ripeta la verifica delle ipotesi con il livello di significatività osservato.

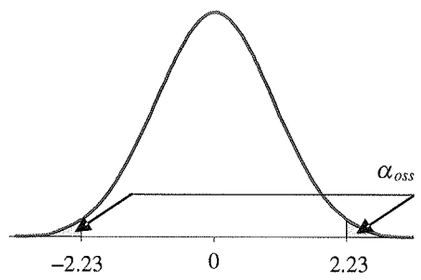
Soluzione

Il livello di significatività osservato è dato da (si veda la Figura 20.8)

$$\alpha_{oss} = P(|Z_{\bar{X}}| \geq 2.23; \mu = 1.25) = 2[1 - \Phi(2.23)] = 0.0258.$$

Poiché $\alpha_{oss} = 0.0258 < 0.05 = \alpha$, l'ipotesi nulla viene rifiutata.

Figura 20.8
Livello di significatività osservato per il problema dell'Esempio 20.8.



■

20.4 Test di ipotesi sulla media di una popolazione normale con varianza non nota

Quando la varianza σ^2 della popolazione normale oggetto di studio non è nota, è necessario sostituire a σ lo stimatore S nella statistica test standardizzata utilizzata nel paragrafo precedente

$$Z_{\bar{X}} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

Con questa sostituzione, si ottiene la statistica test

$$T_{\bar{X}} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$$

Se l'ipotesi nulla, $H_0 : \mu = \mu_0$, è vera, $T_{\bar{X}}$ ha distribuzione T di Student con $n - 1$ gradi di libertà.

Con lo stesso ragionamento con cui si sono costruite le zone di rifiuto per la statistica test $Z_{\bar{X}}$, è facile pervenire alle zone di rifiuto per la statistica test $T_{\bar{X}}$ appena introdotta. In queste, $t_{\bar{x}}$ e t_p indicano il valore osservato della statistica test e il quantile di livello p della distribuzione T di Student con $n - 1$ gradi di libertà.

Ipotesi alternativa unidirezionale a destra

Zona di rifiuto per $H_0 : \mu = \mu_0$ contro $H_1 : \mu > \mu_0$

$$\{t_{\bar{x}} : t_{\bar{x}} > t_{1-\alpha}\}.$$

Ipotesi alternativa unidirezionale a sinistra

Zona di rifiuto per $H_0 : \mu = \mu_0$ contro $H_1 : \mu < \mu_0$

$$\{t_{\bar{x}} : t_{\bar{x}} < -t_{1-\alpha}\}.$$

Ipotesi alternativa bidirezionale

Zona di rifiuto per $H_0 : \mu = \mu_0$ contro $H_1 : \mu \neq \mu_0$

$$\{t_{\bar{x}} : |t_{\bar{x}}| > t_{1-\alpha/2}\}.$$

Le tre zone di rifiuto sono illustrate nella Figura 20.9:

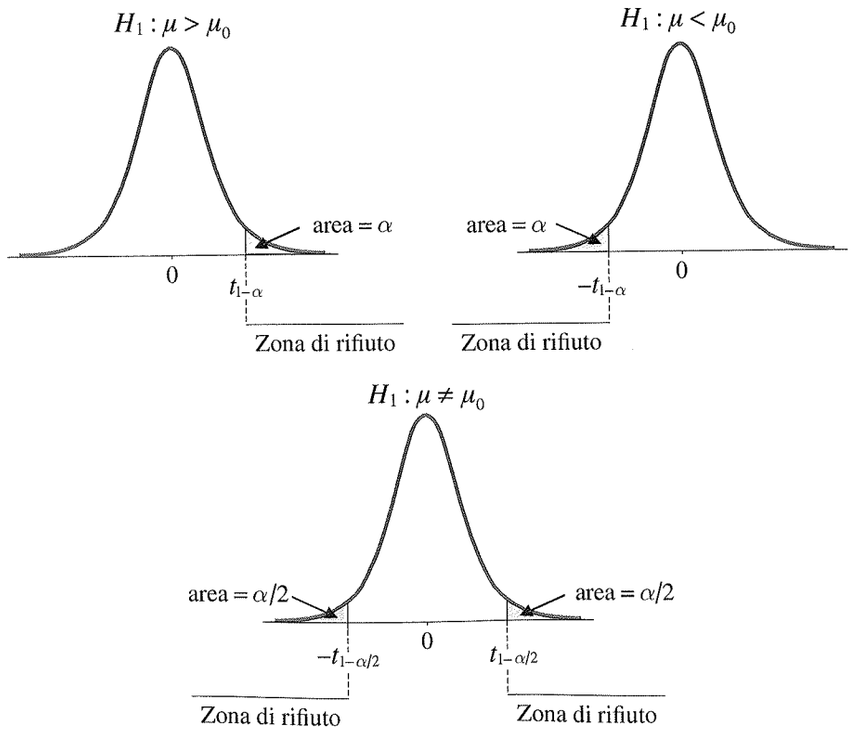


Figura 20.9
Zone di rifiuto della statistica test $T_{\bar{X}} = (\bar{X} - \mu_0)/(S/\sqrt{n})$ per la verifica delle ipotesi sulla media di una popolazione normale con varianza non nota.

Le zone di rifiuto illustrate sopra sono riassunte nella Tabella 20.2.

Ipotesi H_0	Ipotesi H_1	Statistica test	Zona di rifiuto
$\mu = \mu_0$	$\mu > \mu_0$	$T_{\bar{X}} = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$	$\{t_{\bar{x}} : t_{\bar{x}} > t_{1-\alpha}\}$
	$\mu < \mu_0$		$\{t_{\bar{x}} : t_{\bar{x}} < -t_{1-\alpha}\}$
	$\mu \neq \mu_0$		$\{t_{\bar{x}} : \lvert t_{\bar{x}}\rvert > t_{1-\alpha/2}\}$

Tabella 20.2
Zone di rifiuto della statistica test $T_{\bar{X}}$ per la verifica delle ipotesi sulla media di una popolazione normale con varianza incognita.

Come indicato nella Proposizione 20.1, la zona di rifiuto per il test di ipotesi unidirezionale a destra definita sopra è valida anche per l'ipotesi nulla composta $H_0 : \mu \leq \mu_0$. Analogamente, la zona di rifiuto per il test di ipotesi unidirezionale a sinistra è valida anche quando l'ipotesi nulla è composta, $H_0 : \mu \geq \mu_0$.

20.4.1 Livello di significatività osservato

In forza della Definizione 20.4, il livello di significatività osservato per la statistica test $T_{\bar{X}}$ può essere calcolato come segue.

Ipotesi alternativa unidirezionale a destra

$$\alpha_{oss} = P(T_{\bar{X}} \geq t_{\bar{x}};\ \mu = \mu_0).$$

Ipotesi alternativa unidirezionale a sinistra

$$\alpha_{oss} = P(T_{\bar{X}} \leq t_{\bar{x}};\ \mu = \mu_0).$$

Ipotesi alternativa bidirezionale

$$\alpha_{oss} = P(|T_{\bar{X}}| \geq |t_{\bar{x}}|;\ \mu = \mu_0).$$

Va segnalato che la Tavola C.3 (contenente i quantili della distribuzione T di Student) non consente generalmente di determinare il livello di significatività osservato con la dovuta precisione. Per calcoli più precisi è necessario fare ricorso a qualche *software* statistico.

ESEMPIO 20.12

L'osservazione delle pulsazioni cardiache (battiti per minuto) su un campione di 10 studenti maschi del primo anno di un corso di laurea in medicina ha dato i risultati indicati di seguito:

64, 75, 80, 60, 63, 78, 68, 65, 68, 67.

È noto che il valore clinico normale della frequenza media di pulsazioni per i maschi giovani è 72 battiti al minuto. Il quesito a cui si vuole rispondere è se i dati osservati sono in linea con il suddetto valore di riferimento (72 battiti al minuto). Si ponga $\alpha = 0.05$, e si proceda sia con la zona di rifiuto sia con il livello di significatività osservato.

Soluzione

L'ipotesi da verificare è $H_0: \mu = 72$, a fronte dell'alternativa $\mu \neq 72$. Poiché $\alpha/2 = 0.025$, il quantile della distribuzione T di Student con 9 gradi di libertà che delimita la zona di rifiuto è $t_{0.975} = 2.262$. La zona di rifiuto è allora

$$\{t_{\bar{x}}: |t_{\bar{x}}| > 2.262\}.$$

La media e la varianza del campione sono: $\bar{x} = 68.80$ e $s^2 = 44.62$. Il valore della statistica test $T_{\bar{X}}$ è allora

$$t_{\bar{x}} = \frac{68.80 - 72}{\sqrt{44.62/10}} = -1.51.$$

La non appartenenza di tale quantità alla zona di rifiuto porta a non rifiutare l'ipotesi nulla.

Naturalmente, alla stessa decisione si arriva con il livello di significatività osservato, che è uguale a

$$\alpha_{oss} = P(|T_{\bar{X}}| \geq 1.51; \mu = 72) = 0.16,$$

un valore superiore a 0.05. ■

20.5 Test di ipotesi sulla media di una popolazione nel caso di grandi campioni

Quando il campione è sufficientemente grande e la varianza della popolazione non è nota, la statistica test da utilizzare per la verifica dell'ipotesi $H_0: \mu = \mu_0$ è

$$Z_{\bar{X}} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$$

Se l'ipotesi nulla è vera, tale variabile casuale si distribuisce, approssimativamente, secondo una normale standard (si veda la Proposizione 17.6), indipendentemente da come è distribuita la popolazione. Pertanto, il procedimento di verifica di ipotesi coincide, *mutatis mutandis*, con quello descritto nel caso di varianza nota, a meno della sostituzione di s a σ nelle espressioni delle zone di rifiuto (Tabella 20.3).

Tabella 20.3
Zone di rifiuto per la verifica delle ipotesi sulla media di una popolazione qualsiasi nel caso di grandi campioni.

Ipotesi H_0	Ipotesi H_1	Statistica test	Zona di rifiuto
	$\mu > \mu_0$		$\{z_{\bar{x}}: z_{\bar{x}} > z_{1-\alpha}\}$
$\mu = \mu_0$	$\mu < \mu_0$	$Z_{\bar{X}} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$	$\{z_{\bar{x}}: z_{\bar{x}} < -z_{1-\alpha}\}$
	$\mu \neq \mu_0$		$\{z_{\bar{x}}: \|z_{\bar{x}}\| > z_{1-\alpha/2}\}$

Per il calcolo del livello di significatività osservato, si ricorre naturalmente all'approssimazione normale. Per esempio, se l'ipotesi alternativa è $H_1: \mu > \mu_0$, α_{oss} è dato da

$$\alpha_{oss} = P\left(\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \geq z_{\bar{x}}; \mu = \mu_0\right) \approx 1 - \Phi(z_{\bar{x}}),$$

essendo $z_{\bar{x}} = (\bar{x} - \mu_0)/(s/\sqrt{n})$.

ESEMPIO 20.13

La pubblicità di un programma per il dimagrimento afferma: "Perderai 10 kg in 4 mesi". In uno studio di un'associazione di consumatori condotto su un campione casuale di 60 individui, viene osservata una perdita media di peso di 9.1 kg con una deviazione standard $s = 4.2$. Si verifichi l'ipotesi di veridicità della pubblicità a un livello di significatività dell'1%.

Soluzione

L'ipotesi nulla è $H_0: \mu = 10$, l'ipotesi alternativa è $H_1: \mu < 10$. La dimensione del campione consente di utilizzare la statistica test $Z_{\bar{X}} = (\bar{X} - \mu_0)/(S/\sqrt{n})$. La zona di rifiuto è allora

$$\{z_{\bar{x}}: z_{\bar{x}} < -2.326\},$$

dove $z_{\bar{x}}$ è il valore osservato della statistica test. Considerando che

$$z_{\bar{x}} = \frac{9.1 - 10}{4.2/\sqrt{60}} = -1.66$$

non appartiene alla zona di rifiuto, si deve concludere che non vi è evidenza sufficiente per rifiutare l'ipotesi nulla al livello di significatività dell'1%.

Chiaramente, conduce alla stessa conclusione il livello di significatività osservato. Infatti, si ha

$$\alpha_{oss} = P(Z_{\bar{X}} \leq -1.66; \mu = 10) \approx \Phi(-1.66) = 0.048 > 0.01.$$

■

20.5.1 Popolazioni Bernoulliane

Il problema è quello della verifica delle ipotesi sulla media p di una popolazione Bernoulliana. La statistica test naturale è la proporzione campionaria $\hat{p}$, stimatore di p non distorto, efficiente (si veda il Paragrafo 18.3) e con distribuzione approssimativamente normale quando la dimensione del campione è sufficientemente grande.

L'ipotesi nulla è $H_0: p = p_0$. Se H_0 è vera, $\hat{p}$ ha valore atteso p_0 e varianza $p_0(1 - p_0)/n$, cosicché la proporzione standardizzata, data da

$$Z_{\hat{p}} = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}},$$

si distribuisce (approssimativamente) secondo la normale standard.

Considerando che $\hat{p}$ è una media campionaria, le zone di rifiuto di questa statistica test possono essere ottenute con lo stesso procedimento seguito per $\bar{X}$ nel Paragrafo 20.3. La Tabella 20.4 riporta le zone di rifiuto per $Z_{\hat{p}}$.

Tabella 20.4
Zone di rifiuto della statistica test $Z_{\hat{p}}$ per la verifica delle ipotesi sulla media di una popolazione Bernoulliana nel caso di grandi campioni.

Ipotesi H_0	Ipotesi H_1	Statistica test	zona di rifiuto
	$p > p_0$		$\{z_{\hat{p}}: z_{\hat{p}} > z_{1-\alpha}\}$
$p = p_0$	$p < p_0$	$Z_{\hat{p}} = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$	$\{z_{\hat{p}}: z_{\hat{p}} < -z_{1-\alpha}\}$
	$p \neq p_0$		$\{z_{\hat{p}}: \|z_{\hat{p}}\| > z_{1-\alpha/2}\}$

Il livello di significatività osservato si calcola con l'usuale procedimento.

ESEMPIO 20.14

In un sondaggio basato su 300 interviste, 96 intervistati dichiarano l'intenzione di votare per il partito P. Ci si chiede se il consenso del partito sia realmente diminuito, rispetto al 34% ottenuto alle elezioni di due anni prima, o se la differenza sia dovuta a fattori casuali. Si usi un livello di significatività del 5%.

Soluzione

Il problema può essere così schematizzato. Le ipotesi nulla e alternativa sono: $H_0: p=0.34$ e $H_1: p<0.34$. La zona di rifiuto è data da

$$\{z_{\hat{p}}: z_{\hat{p}} < -1.645\}.$$

La proporzione campionaria è

$$\hat{p} = \frac{96}{300} = 0.32,$$

da cui

$$z_{\hat{p}} = \frac{0.32 - 0.34}{\sqrt{(0.34 \cdot 0.66)/300}} = -0.73.$$

Essendo

$$z_{\hat{p}} = -0.73 > -1.645 = -z_{0.95},$$

l'ipotesi nulla non viene rifiutata al livello di significatività del 5%.

Alla stessa conclusione si perviene con il livello di significatività osservato. Infatti, si ha

$$\alpha_{oss} = P(Z_{\hat{p}} \leq -0.73; p = 0.34) \approx \Phi(-0.73) = 0.23 > 0.05. \quad \blacksquare$$

20.6 Test di ipotesi sulla varianza di una popolazione normale

Vi sono situazioni in cui si ha interesse a verificare ipotesi sulla varianza di una popolazione. Per esempio, nel controllo di un processo industriale, ha senso monitorare, oltre che la media, anche la variabilità del processo. Tornando all'esempio del confezionamento dei pacchetti di riso (Paragrafo 20.1), importante è non solo la stabilità del contenuto medio del pacchetto, ma anche la stabilità della variabilità del peso da un pacchetto all'altro. Ne segue che può essere importante effettuare test di ipotesi sulla varianza del peso dei pacchetti.

Si assuma che la popolazione oggetto di studio sia normale con media μ e varianza σ^2 entrambe incognite. L'ipotesi nulla ha la forma $H_0: \sigma^2 = \sigma_0^2$. In questo caso, la statistica test più naturale da assumere è la varianza campionaria S^2, che è lo stimatore non distorto ed efficiente di σ^2 (si veda il Paragrafo 18.3). Se l'ipotesi nulla è vera, il rapporto

$$V = \frac{(n-1)S^2}{\sigma_0^2}$$

ha distribuzione chi-quadrato con $n-1$ gradi di libertà (Proposizione 17.4).

Qui di seguito, viene data la zona di rifiuto per la statistica test S^2 per ciascuna delle tre possibili ipotesi alternative, indicando con χ_p^2 e s^2 il quantile di livello p della distribuzione chi-quadrato con $n-1$ gradi di libertà e il valore osservato di S^2.

Ipotesi alternativa unidirezionale a destra

$$H_0: \sigma^2 = \sigma_0^2 \text{ contro } H_1: \sigma^2 > \sigma_0^2.$$

Se l'ipotesi nulla è vera, il valore atteso di S^2 è uguale a σ_0^2, cosicché valori di S^2 "molto più grandi" di σ_0^2 rendono poco plausibile l'ipotesi nulla. In base alla regola di decisione generale (20.1), è necessario determinare la costante c tale che

$$P(S^2 > c; \sigma_0^2) = \alpha.$$

Questa relazione può essere posta nella forma

$$P\left[\frac{(n-1)S^2}{\sigma_0^2} > \frac{(n-1)c}{\sigma_0^2}; \sigma_0^2\right] = \alpha,$$

da cui, sapendo che $(n-1)S^2/\sigma_0^2$ ha distribuzione chi-quadrato con $n-1$ gradi di libertà, si deduce l'equazione in c

$$\frac{(n-1)c}{\sigma_0^2} = \chi_{1-\alpha}^2,$$

che, risolta, dà

$$c = \frac{\sigma_0^2}{n-1}\chi_{1-\alpha}^2.$$

La zona di rifiuto è allora

$$\left\{s^2: s^2 > \frac{\sigma_0^2}{n-1}\chi_{1-\alpha}^2\right\}.$$

Ipotesi alternativa unidirezionale a sinistra

$$H_0: \sigma^2 = \sigma_0^2 \text{ contro } H_1: \sigma^2 < \sigma_0^2.$$

Dalla (20.2), con le stessa linea di ragionamento, si ottiene

$$\left\{s^2: s^2 < \frac{\sigma_0^2}{n-1}\chi_{\alpha}^2\right\}.$$

Ipotesi alternativa bidirezionale

$$H_0: \sigma^2 = \sigma_0^2 \text{ contro } H_1: \sigma^2 \neq \sigma_0^2.$$

Dalla (20.3), con un procedimento simile a quello illustrato nella derivazione della zona di rifiuto del test bidirezionale per la media (Paragrafo 20.3), si ottiene

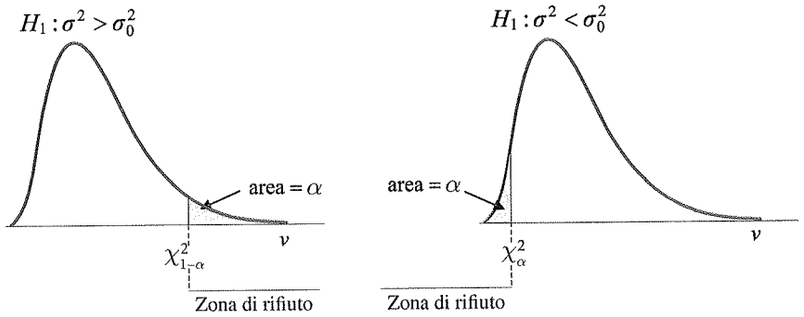
$$\left\{s^2: s^2 \notin \left[\frac{\sigma_0^2}{n-1}\chi_{\alpha/2}^2, \frac{\sigma_0^2}{n-1}\chi_{1-\alpha/2}^2\right]\right\}.$$

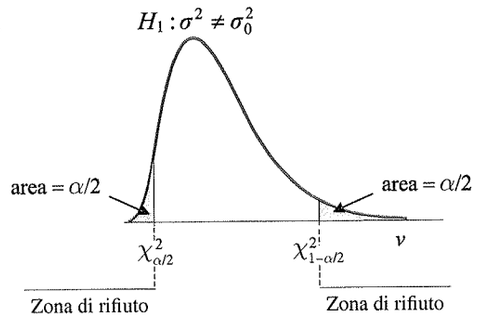
Le zone di rifiuto possono essere determinate, equivalentemente, utilizzando come statistica test V. Queste sono riportate nella Tabella 20.5 e illustrate nella Figura 20.10, dove v è il valore osservato di V.

lla 20.5
di rifiuto della
tica test
$(n-1)S^2/\sigma_0^2$ per
rifica delle ipotesi
varianza di una
lazione normale.

Ipotesi H_0	Ipotesi H_1	Statistica test	Zona di rifiuto
	$\sigma^2 > \sigma_0^2$		$\{v: v > \chi^2_{1-\alpha}\}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$V = \dfrac{(n-1)S^2}{\sigma_0^2}$	$\{v: v < \chi^2_{\alpha}\}$
	$\sigma^2 \neq \sigma_0^2$		$\{v: v \notin (\chi^2_{\alpha/2}, \chi^2_{1-\alpha/2})\}$

ra 20.10
di rifiuto della
tica test
$(n-1)S^2/\sigma_0^2$ per
rifica delle ipotesi
varianza di una
lazione normale.





ESEMPIO 20.15

Il peso (in grammi) delle uova prodotte da un'azienda avicola segue una distribuzione normale. Si verifichi l'ipotesi che la varianza della popolazione sia $\sigma^2 = 36$ contro l'alternativa $H_1: \sigma^2 < 36$, al livello di significatività del 5%, sulla base dei pesi osservati in un campione di 12 uova:

61, 57, 58, 65, 54, 63, 56, 68, 67, 53, 64, 66.

Soluzione

La zona di rifiuto del test è

$$\{v: v < \chi^2_{0.05} = 4.57\},$$

dove 4.57 è il quinto centile della distribuzione chi-quadrato con 11 gradi di libertà. La varianza campionaria è $s^2 = 27.45$, per cui la statistica test è uguale a

$$v = \frac{(n-1)s^2}{\sigma_0^2} = \frac{11 \cdot 27.45}{36} = 8.39,$$

un valore che è al di fuori della zona di rifiuto. La conclusione da trarre è quella di non rifiutare l'ipotesi nulla. ■

20.6.1 Livello di significatività osservato

Sulla base della definizione generale di livello si significatività osservato, si ottengono immediatamente le espressioni di seguito indicate.

Ipotesi alternativa unidirezionale a destra

$$\alpha_{oss} = P(V \geq v;\ \sigma^2 = \sigma_0^2).$$

Ipotesi alternativa unidirezionale a sinistra

$$\alpha_{oss} = P(V \leq v;\ \sigma^2 = \sigma_0^2).$$

Ipotesi alternativa bidirezionale

$$\alpha_{oss} = 2\min[P(V \geq v;\ \sigma^2 = \sigma_0^2),\ P(V \leq v;\ \sigma^2 = \sigma_0^2)].$$

ESEMPIO 20.16

Si ripeta la verifica di ipotesi dell'esempio precedente usando il livello di significatività osservato.

Soluzione

Il livello di significatività osservato è dato da

$$\alpha_{oss} = P(V \leq 8.39; \sigma^2 = 36) = 0.32.$$

Questo risultato (che può essere ottenuto con un *software* statistico) porta, come previsto, alla conferma della decisione presa nell'Esempio 20.15. ■

20.7 Criteri di ottimizzazione nella verifica delle ipotesi

Nel Paragrafo 20.2, è stato detto incidentalmente che un giudizio sulla qualità di un test di ipotesi va basato sulle probabilità α e β (o π). Qui, verranno proposti alcuni approfondimenti al riguardo.

Considerando che ad α viene preliminarmente assegnato un valore, la qualità di un test dipende esclusivamente dal livello di β o di π. La discussione che segue è imperniata sulla seconda quantità.

Si ricordi che la potenza, definita come la probabilità di rifiutare un'ipotesi nulla falsa, è funzione del parametro θ, che può variare all'interno dell'insieme specificato dall'ipotesi alternativa. Si parla perciò di funzione di potenza, che per una generica statistica test W è data da

$$\pi(\theta) = P(W \in R;\ \theta),$$

dove R è la zona di rifiuto.

ESEMPIO 20.17

Si consideri il problema della verifica delle ipotesi sulla media di una popolazione normale con varianza nota. Sia $H_0: \mu = \mu_0$ l'ipotesi nulla e $H_1: \mu = \mu_1$ (con $\mu_1 > \mu_0$) l'ipotesi alternativa. La Figura 20.11 illustra la zona di rifiuto e le probabilità α e β della statistica test $\bar{X}$. La curva (a) descrive la distribuzione campionaria di

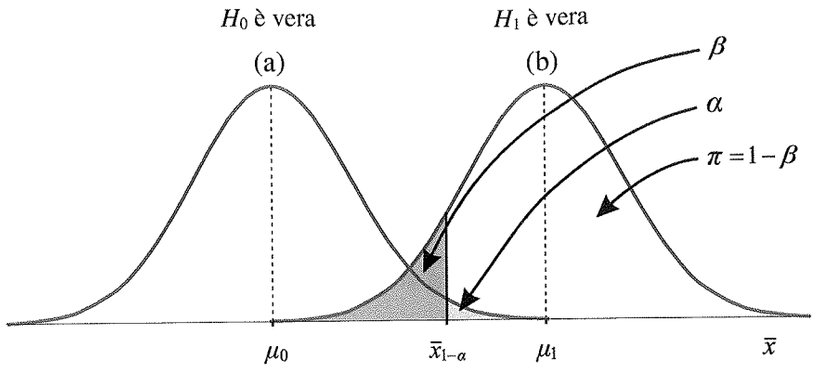


Figura 20.11
Probabilità degli errori I e II tipo.

$\bar{X}$ sotto l'ipotesi nulla, una distribuzione normale con media μ_0 e varianza σ^2/n. La curva (b) descrive la distribuzione campionaria della medesima statistica sotto l'ipotesi alternativa, una distribuzione normale con media μ_1 e varianza σ^2/n. La probabilità α dell'errore di primo I tipo è rappresentata dall'area sotto la curva (a) a destra del valore critico $\bar{x}_{1-\alpha} = \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}$. La probabilità β dell'errore di II tipo è rappresentata dall'area sotto la curva (b) a sinistra di $\bar{x}_{1-\alpha}$. La potenza è rappresentata dall'area sotto la curva (b) alla destra di $\bar{x}_{1-\alpha}$. ■

ESEMPIO 20.18

Per lo stesso problema dell'Esempio 20.17, la Figura 20.12 mostra quattro distribuzioni di probabilità della statistica test $\bar{X}$: la prima (da sinistra) è costruita sotto l'ipotesi nulla $H_0 : \mu = \mu_0$, le altre tre sotto l'ipotesi alternativa, con tre valori di μ, μ_1, μ_2 e μ_3, tutti e tre maggiori di μ_0. La figura mostra come la potenza del test cresca via via che la media si allontana da μ_0. Come si può osservare, passando da μ_1 a μ_2, da μ_2 a μ_3, la curva rappresentativa della distribuzione campionaria di $\bar{X}$ si sposta a destra, cosicché l'area sotto la curva a sinistra di $\bar{x}_{1-\alpha}$, cioè la probabilità $\beta(\mu)$, decresce, mentre l'area a destra di $\bar{x}_{1-\alpha}$ (la potenza) cresce.

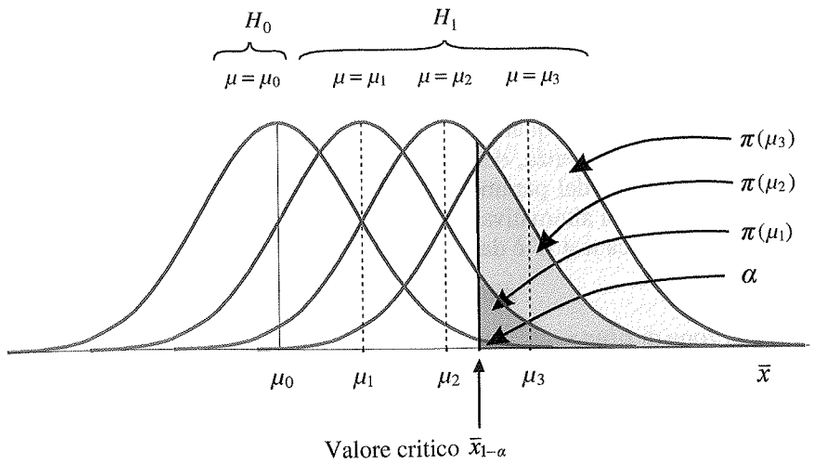


Figura 20.12
Potenza del test $\bar{X}$ per valori crescenti di μ.

Chiaramente, la potenza del test è funzione di μ, che cresce via via che μ si allontana da μ_0, convergendo verso il valore 1. ■

ESEMPIO 20.19

Si consideri il problema della verifica delle ipotesi sulla media di una popolazione normale con varianza nota. Sia $H_0 : \mu = \mu_0$ l'ipotesi nulla e $H_1 : \mu > \mu_0$ l'ipotesi alternativa. La funzione di potenza del test è data da

$$\begin{aligned}\pi(\mu) &= P\left[\bar{X} > \mu_0 + z_{1-\alpha}\sigma/\sqrt{n};\ \mu\right] \\ &= P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > -\frac{\mu-\mu_0}{\sigma/\sqrt{n}} + z_{1-\alpha};\ \mu\right) \\ &= 1 - \Phi\left(-\frac{\mu-\mu_0}{\sigma/\sqrt{n}} + z_{1-\alpha}\right).\end{aligned}$$

Se l'ipotesi alternativa è $H_1 : \mu < \mu_0$, la funzione di potenza è definita da

$$\begin{aligned}\pi(\mu) &= P\left(\bar{X} < \mu_0 - z_{1-\alpha}\sigma/\sqrt{n};\ \mu\right) \\ &= P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < -\frac{\mu-\mu_0}{\sigma/\sqrt{n}} - z_{1-\alpha};\ \mu\right) \\ &= \Phi\left(-\frac{\mu-\mu_0}{\sigma/\sqrt{n}} - z_{1-\alpha}\right).\end{aligned}$$

La questione sarà ripresa nell'Esempio 20.23. ■

ESEMPIO 20.20

Si scriva la funzione di potenza del test per la verifica dell'ipotesi nulla H_0: $\mu = \mu_0$ contro l'alternativa $H_1 : \mu > \mu_0$ per una popolazione normale con varianza non nota.

Soluzione

La funzione di potenza del test per l'ipotesi alternativa H_1: $\mu > \mu_0$ è data da

$$\pi(\mu) = P\left(\frac{\bar{X}-\mu_0}{S/\sqrt{n}} > t_{1-\alpha};\ \mu\right) = P\left[\frac{\bar{X}-\mu+(\mu-\mu_0)}{S/\sqrt{n}} > t_{1-\alpha};\ \mu\right],$$

dove $t_{1-\alpha}$ è il quantile di livello $1-\alpha$ della distribuzione T di Student con $n-1$ gradi di libertà. Dividendo numeratore e denominatore del rapporto all'interno delle parentesi quadre per $\sigma/\sqrt{n}$, si ottiene

$$\pi(\mu) = P\left[\frac{Z+\delta}{\sqrt{V/(n-1)}} > t_{1-\alpha};\ \mu\right],$$

dove $\delta = \sqrt{n}(\mu - \mu_0)/\sigma$, mentre $Z = \sqrt{n}(\bar{X} - \mu)/\sigma$ e $V = (n-1)S^2/\sigma^2$ sono variabili casuali indipendenti distribuite, rispettivamente, secondo una normale standard e una chi-quadrato con $n-1$ gradi di libertà. Il rapporto in questione ha distribuzione T non centrale con $n-1$ gradi di libertà e parametro di non centralità δ. In questo caso, per la quantificazione della funzione di potenza bisogna ricorrere a un *software* statistico (per esempio, a R). Il parametro di non centralità dipende dalla deviazione standard della popolazione. La sostituzione di questa con la stima campionaria introduce ovviamente un'approssimazione nel calcolo di $\pi(\mu)$. ■

ESEMPIO 20.21

Si determini la funzione di potenza del test basato sulla proporzione campionaria $\hat{p}$ per la verifica dell'ipotesi $H_0: p = p_0$ contro l'alternativa $H_1: p > p_0$.

Soluzione

La funzione di potenza è data da

$$\pi(p) = P\left(\hat{p} > p_0 + z_{1-\alpha}\sqrt{p_0(1-p_0)/n}; p\right)$$

$$= P\left[\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} > \frac{p_0 - p}{\sqrt{p(1-p)/n}} + \frac{z_{1-\alpha}\sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}}; p\right]$$

$$\approx 1 - \Phi\left[\frac{p_0 - p}{\sqrt{p(1-p)/n}} + \frac{z_{1-\alpha}\sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}}\right].$$

■

ESEMPIO 20.22

Si consideri il problema della verifica delle ipotesi sulla varianza di una popolazione normale. Siano $H_0: \sigma^2 = \sigma_0^2$ l'ipotesi nulla e $H_1: \sigma^2 < \sigma_0^2$ l'ipotesi alternativa. Si determini la funzione di potenza del test basato sulla varianza campionaria e se ne calcoli il valore per il caso in cui: $H_0: \sigma^2 = 64$, $H_1: \sigma^2 = 52$, $n = 40$ e $\alpha = 0.05$.

Soluzione

Il valore critico del test è $[\sigma_0^2/(n-1)] \cdot \chi_\alpha^2$. Si può allora scrivere

$$\pi(\sigma^2) = P\left(S^2 < \frac{\sigma_0^2}{n-1} \cdot \chi_\alpha^2; \sigma^2\right).$$

Moltiplicando entrambi i lati della disequazione all'interno delle parentesi per $(n-1)/\sigma^2$, si ottiene

$$\pi(\sigma^2) = P\left[\frac{(n-1)S^2}{\sigma^2} < \frac{\sigma_0^2}{\sigma^2} \cdot \chi_\alpha^2; \sigma^2\right].$$

Poiché $(n-1)S^2/\sigma^2$ ha distribuzione chi-quadrato con $n-1$ gradi di libertà, la funzione di potenza di S^2 è la funzione di ripartizione di una variabile casuale chi-quadrato con $n-1$ gradi di libertà nel punto $(\sigma_0^2/\sigma^2) \cdot \chi_\alpha^2$. Come si può osservare, $\pi(\sigma^2)$ ha le seguenti proprietà: assume il valore minimo, pari ad α, quando $\sigma^2 = \sigma_0^2$; è funzione decrescente di σ^2; tende a 1 via via che σ^2 si avvicina allo 0.

Supponendo che $H_0: \sigma^2 = 64$, $H_1: \sigma^2 = 52$, $n = 40$ e $\alpha = 0.05$, $\pi(\sigma^2)$ diviene

$$\pi(52) = P\left(\frac{39 \cdot S^2}{52} < \frac{64}{52} \cdot 25.70; \sigma^2 = 52\right),$$

dove 25.70 è il quinto centile della distribuzione chi-quadrato con 39 gradi di libertà. Dall'applicazione di *software* statistico (R, Excel ecc.), si ottiene $\pi = 0.21$ (il valore approssimato che si desume dalla Tavola C.2 è 0.25, essendo 32.74 il valore più vicino a (64/52)25.70 = 31.63 che si legge nella tavola).

■

Qui di seguito, viene data la definizione di *test uniformemente più potente* per il caso di ipotesi nulla semplice.

DEFINIZIONE 20.5

Si consideri il problema della verifica dell'ipotesi nulla $H_0: \theta = \theta_0$ contro una delle tre possibili ipotesi alternative, $H_1: \theta > \theta_0$, $H_1: \theta < \theta_0$, $H_1: \theta \neq \theta_0$. Un test statistico, avente un livello di significatività α, è detto uniformemente più potente se la sua potenza è maggiore o uguale a quella di ogni altro test statistico, avente lo stesso livello di significatività α, per tutti i valori di θ appartenenti all'insieme specificato dall'ipotesi alternativa.

Chiaramente, un test uniformemente più potente è il migliore test possibile. Sfortunatamente, le condizioni per l'esistenza di tali test sono piuttosto restrittive.[1]

È possibile dimostrare che i test unidirezionali per la verifica delle ipotesi sulla media di una popolazione normale con varianza nota basati su $\bar{X}$ sono uniformemente più potenti, mentre il test bidirezionale basato sulla medesima statistica non lo è (si veda l'Esempio 20.23). Le funzioni di potenza di detti test unidirezionali sono rappresentate nella Figura 20.13 per $\alpha = 0.05$ (curve (a) e (b)).

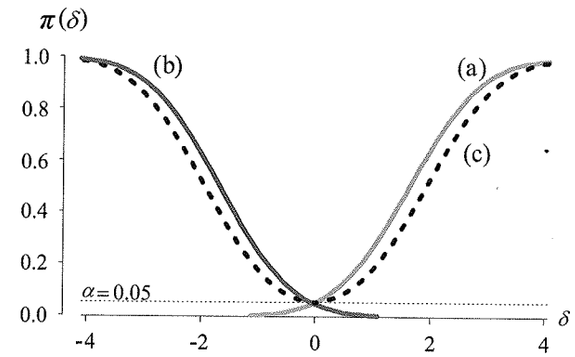


Figura 20.13
Funzione di potenza della statistica test $\bar{X}$ per le ipotesi alternative: $H_1: \mu > \mu_0$, curva (a), $H_1: \mu < \mu_0$, curva (b), $H_1: \mu \neq \mu_0$, curva (c). Sull'asse orizzontale sono riportati i valori di $\delta = (\mu - \mu_0)/(\sigma/\sqrt{n})$.

ESEMPIO 20.23

Si consideri il problema della verifica dell'ipotesi nulla $H_0: \mu = \mu_0$ contro l'alternativa $H_1: \mu \neq \mu_0$ per una popolazione normale con varianza nota. Si può dimostrare che il test bidirezionale basato su $\bar{X}$ non è uniformemente più potente. La funzione di potenza di questo test è data da

$$\pi(\mu) = P\left[\bar{X} \notin \left(\mu_0 - z_{1-\alpha/2}\sigma/\sqrt{n},\ \mu_0 + z_{1-\alpha/2}\sigma/\sqrt{n}\right); \mu\right]$$

$$= P\left[\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \notin \left(-\frac{\mu - \mu_0}{\sigma/\sqrt{n}} - z_{1-\alpha/2},\ -\frac{\mu - \mu_0}{\sigma/\sqrt{n}} + z_{1-\alpha/2}\right); \mu\right]$$

$$= 1 - \Phi\left(-\frac{\mu - \mu_0}{\sigma/\sqrt{n}} + z_{1-\alpha/2}\right) + \Phi\left(-\frac{\mu - \mu_0}{\sigma/\sqrt{n}} - z_{1-\alpha/2}\right).$$

[1] Si può dimostrare che il test uniformemente più potente esiste solo se la funzione di probabilità di densità che descrive la variabile casuale oggetto di studio soddisfa alcune proprietà. Per approfondimenti sull'argomento si rinvia a Abadir et al. (2018), p. 663 e segg.

Questa funzione è illustrata nella Figura 20.13 (per $\alpha = 0.05$) – curva (c) – insieme a quelle dei due test unidirezionali esaminati nell'Esempio 20.19. Come si può vedere, la funzione di potenza del test bidirezionale è dominata da quella del test unidirezionale a destra, quando $\mu > \mu_0$ ($\delta > 0$) – curva (a) –, e da quella del test unidirezionale a sinistra, quando $\mu < \mu_0$ ($\delta < 0$) – curva (b). Da ciò si desume che il test bidirezionale basato su $\bar{X}$ non è uniformemente più potente. ■

DEFINIZIONE 20.6

Si consideri il problema della verifica dell'ipotesi nulla $H_0: \theta = \theta_0$ contro una delle tre possibili alternative, $H_1: \theta > \theta_0$, $H_1: \theta < \theta_0$, $H_1: \theta \neq \theta_0$. Un test con un livello di significatività α è detto non distorto se la sua potenza è maggiore o uguale al livello di significatività α per tutti i valori di θ appartenenti all'insieme specificato dall'ipotesi alternativa.

È questa una proprietà desiderabile per un test statistico, perché, se non valesse, vi sarebbero valori di θ appartenenti all'insieme definito da H_1 per i quali la probabilità di rifiutare un'ipotesi nulla vera è maggiore della probabilità di rifiutare un'ipotesi nulla falsa.

La non distorsione inoltre può essere la chiave con cui risolvere il problema della ricerca del test ottimo. Infatti, in molte situazioni, è possibile individuare il test uniformemente più potente restringendo la ricerca alla classe dei test non distorti. Per esempio, si può dimostrare che il test bidirezionale basato su $\bar{X}$ per la media di una popolazione normale con varianza nota non è, come già detto, uniformemente più potente in assoluto, ma lo è nell'ambito dei test non distorti.

Ciò è illustrato nella Figura 20.13. Come si può vedere, il test bidirezionale, se confrontato con i test unidirezionali, non è uniformemente più potente (perché è dominato dal test unidirezionale a destra quando $\mu > \mu_0$, e dal test unidirezionale a sinistra quando $\mu < \mu_0$). Lo diventa se si restringe l'attenzione ai test non distorti, perché i due test unidirezionali non fanno parte della classe dei test non distorti per il problema di verifica delle ipotesi in questione.

Anche i tre test basati sulla statistica $T_{\bar{X}}$, unidirezionale a destra, unidirezionale a sinistra e bidirezionale, sono non distorti e uniformemente più potenti tra i test non distorti.

Potenza del test e ampiezza campionaria

La potenza di un test dipende anche dalla dimensione del campione: per un dato valore di θ appartenente all'insieme specificato dall'ipotesi alternativa, la potenza cresce all'aumentare della dimensione campionaria, a parità di livello di significatività. Ciò è intuitivamente plausibile se si pensa che un campione più grande, rappresentando meglio la popolazione, rende meno incerta la decisione se rifiutare o non rifiutare l'ipotesi nulla. Ha senso allora il problema della ricerca della dimensione del campione che assicuri un assegnato livello di potenza. Qui di seguito, vengono dati due esempi al riguardo.

ESEMPIO 20.24

Si determini la dimensione campionaria che assicura una potenza pari a π del test (basato su $\bar{X}$) sulla media di una popolazione normale con varianza nota, nel caso di un'ipotesi alternativa unidirezionale a destra.

Soluzione

Sia μ_1 un particolare valore di μ, maggiore di μ_0. La potenza del test è data da (si veda Esempio 20.19)

$$\pi = 1 - \Phi\left(-\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} + z_{1-\alpha}\right),$$

da cui

$$1 - \pi = \Phi\left(-\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} + z_{1-\alpha}\right).$$

Indicato con $z_{1-\pi}$ il punto in cui la funzione di ripartizione della normale standard assume il valore $1 - \pi$, si ha

$$z_{1-\pi} = -\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} + z_{1-\alpha}.$$

Risolvendo l'equazione rispetto a n, si ottiene

$$n = \frac{\sigma^2(z_{1-\alpha} - z_{1-\pi})^2}{(\mu_1 - \mu_0)^2}.$$

■

ESEMPIO 20.25

Si determini la dimensione del campione tale che sia π la potenza del test unidirezionale a sinistra per la varianza di una popolazione normale.

Soluzione

Sia σ_1^2 un particolare valore del parametro σ^2 inferiore a σ_0^2 (si veda l'Esempio 20.22). La potenza del test S^2 è data da

$$\pi = P\left(S^2 < \frac{\sigma_0^2}{n-1} \cdot \chi_\alpha^2;\ \sigma_1^2\right) = P\left[\frac{(n-1)S^2}{\sigma_1^2} < \frac{\sigma_0^2}{\sigma_1^2} \cdot \chi_\alpha^2;\ \sigma_1^2\right].$$

Indicato con χ_π^2 il punto in cui la funzione di ripartizione di una variabile casuale chi-quadrato (con $n - 1$ gradi di libertà) assume il valore π, si può scrivere

$$(\sigma_0^2/\sigma_1^2) \cdot \chi_\alpha^2 = \chi_\pi^2.$$

Sostituendo a χ_α^2 e χ_π^2 i quantili approssimati calcolati con la normale, si ottiene

$$(\sigma_0^2/\sigma_1^2)\left[n - 1 + z_\alpha\sqrt{2(n-1)}\right] = n - 1 + z_\pi\sqrt{2(n-1)}.$$

La soluzione di tale equazione rispetto a n dà

$$n \approx 1 + 2\left[\frac{z_\pi - z_\alpha(\sigma_0^2/\sigma_1^2)}{1 - \sigma_0^2/\sigma_1^2}\right]^2.$$

■

Riassunto

Il test statistico è la procedura con cui si decide se rifiutare o non rifiutare una congettura, la cosiddetta ipotesi nulla, sul valore di un parametro della popolazione oggetto di studio, in base ai dati di un campione casuale osservato. La decisione è basata su un appropriato stimatore del parametro di interesse, denominato statistica test, i cui possibili valori vengono suddivisi in due sottoinsiemi, il primo costituito dai valori per i quali l'ipotesi nulla viene rifiutata (zona di rifiuto) e il secondo dai valori per i quali si prende la decisione opposta (zona di non rifiuto).

La decisione è soggetta a due possibili errori: l'errore di I tipo, che consiste nel rifiutare l'ipotesi nulla uando è vera; l'errore di II tipo, che consiste nel non rifiutare l'ipotesi nulla quando è falsa. La procedura revede che la probabilità dell'errore di primo tipo non superi un limite molto basso, il cosiddetto livello di ignificatività, pari, ad esempio, a 0.05. La scelta della zona di rifiuto poi deve esser tale che la probabilità ell'errore di secondo tipo sia la più bassa possibile.

Intuitivamente, la zona di rifiuto è formata dai valori della statistica test che sono molto lontani dal alore del parametro specificato dall'ipotesi nulla ed è costruita sulla base della distribuzione campionaria ella statistica test determinata assumendo che l'ipotesi nulla sia vera.

I problemi di verifica delle ipotesi trattati in questo capitolo riguardano: la media di una popolazione ormale con varianza nota o non nota, la media di una popolazione qualsiasi per il caso di grandi campioni, a varianza di una popolazione normale.

Nell'ultima parte del capitolo, vengono forniti alcuni elementi di teoria per la scelta ottimale del test tatistico.

Parole chiave

distribuzione T non centrale, 445
errore di I tipo, 423
errore di II tipo, 423
funzione di potenza, 443
ipotesi alternativa, 422
ipotesi nulla, 422
ipotesi statistica, 422
ipotesi statistica composta, 422
ipotesi statistica semplice, 422
livello di significatività, 423
livello di significatività osservato, 428
parametro di non centralità, 445
potenza di un test, 424
statistica test, 422
test di ipotesi, 422
test non distorto, 448
test statistico, 422
test uniformemente più potente, 447
valore critico, 423
zona di non rifiuto, 423
zona di rifiuto, 423

Commenti e curiosità

1. Il Lemma di Neyman-Pearson: un contributo fondamentale alla teoria dei test

Il Lemma di Neyman-Pearson è un risultato basilare nello sviluppo della teoria dei test statistici. Prende il nome dai due illustri statistici, Jerzy Neyman (1894-1981) e Egon S. Pearson (1895-1980), che lo introdussero nel lavoro dal titolo *On the problem of the most efficient tests of statistical hypotheses* (si veda Neyman e Pearson, 1933). Il lemma può essere così enunciato.

Sia $H_0 : \theta = \theta_0$ l'ipotesi nulla semplice e $H_1 : \theta = \theta_1$ l'ipotesi alternativa, anch'essa semplice, essendo θ il parametro della funzione di probabilità o di densità $f(x;\theta)$ della popolazione oggetto di studio. Sia $L(\theta_0; x_1, x_2, \ldots, x_n)$ e $L(\theta_1; x_1, x_2, \ldots, x_n)$ la funzione di verosimiglianza del campione osservabile $(x_1, x_2, \ldots, x_n)$ per $\theta = \theta_0$ e $\theta = \theta_1$. Allora la zona di rifiuto del test più potente a un livello di significatività pari ad α è costituita dall'insieme dei campioni, C, che soddisfano la disuguaglianza

$$\frac{L(\theta_1; x_1, x_2, \ldots, x_n)}{L(\theta_0; x_1, x_2, \ldots, x_n)} \geq k,$$

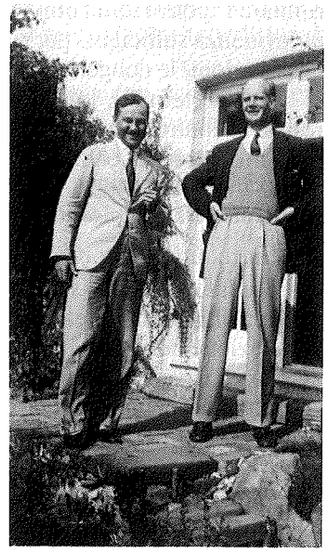
dove k è la costante tale che $P[(X_1, X_2, \ldots, X_n) \in C; \theta_0] = \alpha$.

È opportuno sottolineare che il lemma non fornisce l'espressione esplicita del test più potente, ma ne garantisce l'esistenza e l'unicità e indica la procedura per ottenerlo.

Seguono brevi biografie dei due autori.

Jerzy Neyman (1894-1981) nacque a Bender (Transnistria) da una famiglia di origine polacca. Iniziò i suoi studi universitari in matematica e fisica presso l'Università di Charkiv (Ucraina), che concluse presso l'Università di Varsavia con una tesi sull'applicazione del calcolo delle probabilità agli esperimenti in agricoltura. Nel 1924 svolse attività di ricerca presso l'*University College London* diretto da Karl Pearson, dove ebbe modo di incontrare personaggi chiave della statistica, tra cui Ronald Aylmer Fisher, William Sealy Gosset ed Egon Pearson. Dopo un soggiorno di studi a Parigi nel 1925, tornò in Polonia dove costituì e diresse fino al 1934 il laboratorio di biometria presso l'Istituto Nencki di biologia sperimentale di Varsavia. Nel 1934 fu nominato professore associato dell'*University College London*. Nel 1938 si trasferì all'Università di Berkeley dove concluse la sua carriera accademica.

I contributi scientifici più rilevanti di Neyman riguardano: la teoria dei test statistici, che, insieme a Egon Pearson, contribuì a costruire con una serie di articoli – tra cui quello sopra citato, al quale è legata la sua notorietà internazionale –, la teoria della stima, per l'introduzione dell'intervallo di confidenza, il campionamento da popolazioni finite, di cui si parlerà nel punto 2 di "Commenti e curiosità" del Capitolo 24.



Neyman and Pearson at The Cell
Reprinted with the permission of
the Institute of Mathematical Statistics.

Egon Pearson (1895-1980), figlio di Karl Pearson, compì i suoi studi universitari presso il *Trinity College, Cambridge*. La sua attività di ricerca si orientò inizialmente verso la fisica solare. Nel 1921 ebbero inizio i suoi studi in campo statistico presso il dipartimento di statistica applicata dell'*University College London* diretto da suo padre. Nel 1922 videro la luce i suoi primi importanti contributi scientifici in statistica. Nel 1933, quando suo padre andò in pensione, il dipartimento di statistica applicata fu scisso in due: nel dipartimento di eugenetica, diretto da Ronald Fisher, quale docente di eugenetica, e nel nuovo dipartimento di statistica applicata affidato a Egon Pearson, nella veste di professore associato e poi, dal 1935, di professore ordinario di statistica. Alla morte del padre nel 1936, assunse la direzione della rivista *Biometrika* che tenne fino al 1966.

Fecondi per la sua produzione scientifica furono i contatti e le collaborazioni con importanti statistici dell'epoca, quali W.S. Gosset (Student) con cui si occupò, all'inizio della sua carriera accademica, di distribuzioni di probabilità di statistiche campionarie nel caso di piccoli campioni, H.O. Hartley, co-curatore della revisione delle tavole statistiche di suo padre (*Tables for Statisticians and Biometricians*) e, soprattutto, J. Neyman, coautore di importanti articoli scientifici, basilari per lo sviluppo della teoria dei test statistici, che gli diedero notorietà internazionale. Nella parte finale della sua attività accademica, si occupò di applicazioni della statistica all'industria, con particolare riferimento al controllo di qualità.

2. Ancora sull'espressione "ipotesi nulla"

Nel Paragrafo 20.1, si è accennato al fatto che l'ipotesi nulla è quella che il ricercatore tende a "falsificare", a confutare, sulla base dell'evidenza fornita dai dati del campione. Infatti, quando un ricercatore progetta una ricerca, ha in mente una "ipotesi di studio" (per esempio, che i diabetici sono ipertesi). Sul piano logico, l'ipotesi di studio è provata se viene confutata l'ipotesi contraria (per esempio, che i diabetici non sono ipertesi), da qui l'espressione "ipotesi nulla". Si può anche dire che l'ipotesi nulla rappresenta lo *status quo*, il livello attuale della conoscenza, mentre l'ipotesi alternativa rappresenta l'innovazione, il superamento dello *status quo*.

L'idea sottesa ai test di ipotesi è che l'ipotesi di studio (ipotesi alternativa) debba prevalere sull'ipotesi contraria (ipotesi nulla) solo se è molto forte l'evidenza contro quest'ultima. Si può spiegare così il diverso modo con cui si fronteggiano gli errori di I e di II tipo. Dall'errore di I tipo ci si difende con il livello di significatività, a cui viene assegnato un valore molto piccolo; il controllo dell'errore di II tipo invece potrebbe essere effettuato tramite l'ampiezza del campione, ma questa non è una variabile nella disponibilità del ricercatore, essendo spesso determinata da circostanze obiettive (il budget a disposizione, il numero di pazienti disponibili nelle sperimentazioni cliniche ecc.).

3. Il non rifiuto dell'ipotesi nulla non vuol dire che essa è vera

Nella presentazione dei test di ipotesi e nel commento dei risultati delle applicazioni, si è evitato di usare il termine "accettazione": si è sempre parlato di "rifiuto" e "non rifiuto" dell'ipotesi nulla. In effetti, non

ifiutare l'ipotesi nulla non vuol dire che essa è vera: vuol dire soltanto che i dati osservati non contengono n'evidenza sufficiente per rigettarla. Questa impostazione è in accordo con l'idea Popperiana secondo cui e asserzioni, le congetture, per essere scientifiche, devono poter essere confutabili.

In altri termini, le osservazioni non possono stabilire la verità di una asserzione. Se una asserzione esiste al controllo non la si può definire vera per questo; essa potrebbe crollare al prossimo controllo o alla rossima osservazione.

. Divario tra complessità dei problemi reali e schematizzazione delle procedure di verifica delle ipotesi

Nell'introdurre il problema e i metodi della verifica delle ipotesi su medie e varianze, per semplicità, si proceduto in modo schematico. È necessario ricordare che il presupposto per l'applicazione corretta ei test di ipotesi sulle medie è la casualità del campione osservato, a cui si aggiunge la normalità della popolazione nel caso in cui il campione ha ampiezza non elevata. Questa seconda condizione è necessaria nche per la verifica delle ipotesi sulla varianza.

Quando dalla teoria si passa al mondo reale, sorgono difficoltà e incertezze sulla sussistenza delle condizioni di applicabilità dei metodi in questione. È ben raro infatti che si possano avere certezze al iguardo. Si pensi, ad esempio, a una ricerca di mercato rivolta alla rilevazione del gradimento di un nuovo prodotto cosmetico. Ammesso che l'agenzia di ricerca abbia intervistato 560 donne della fascia di età 24-45, scegliendole in modo che il campione sia "bilanciato" per età, area geografica, livello di istruzione professione, ci si chiede se il campione possa essere considerato casuale. La risposta è negativa se per campione casuale si intende quello formato ricorrendo ai classici criteri di selezione casuale delle unità come l'estrazione di palline numerate da un'urna). La risposta è più aperta se si accerta che la selezione indenne da fattori di distorsione rilevanti: per esempio, il fatto che una quota consistente di intervistate siano state selezionate da una lista di consumatrici abituali di altri prodotti cosmetici della stessa marca. Bisogna quindi sottoporre a indagine critica le osservazioni campionarie disponibili, per poter raggiungere a convinzione che il campione osservato sia "equiparabile" a un campione casuale.

Quanto alla normalità della popolazione, si tratta di studiare la natura del fenomeno aleatorio oggetto di studio per stabilire la plausibilità dell'assunzione. Esistono appropriate tecniche di inferenza statistica che consentono di verificare se il campione osservato può essere considerato come proveniente da una popolazione normale (si veda Paragrafo 22.2).

5. Verifica delle ipotesi e stima intervallare

Tra la verifica delle ipotesi e gli intervalli di confidenza sussiste una relazione molto stretta. Per semplicità, si consideri il sistema di ipotesi, $H_0 : \mu = \mu_0$ e $H_1 : \mu \neq \mu_0$, sulla media di una popolazione normale con varianza nota. La zona di rifiuto del test basato su $\bar{X}$ è data da

$$\left\{\bar{x} : \bar{x} \notin \left(\mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)\right\}.$$

Dunque, l'ipotesi nulla non viene rifiutata se

$$\mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},$$

cioè se

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu_0 < \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Come si vede, con il sistema di ipotesi qui considerato, si rifiuta l'ipotesi nulla se μ_0 è esterna all'intervallo di confidenza per μ con un coefficiente di fiducia pari a $1 - \alpha$.

6. Errori di I e di II tipo: analogie con problemi reali

Una corte d'assise deve giudicare l'accusato di un omicidio. Si assuma che l'ipotesi nulla sia l'affermazione che l'accusato è innocente; l'ipotesi alternativa è l'affermazione opposta, cioè che l'accusato è colpevole. In tale contesto, l'errore di I tipo comporta la condanna di un innocente (rifiuto dell'ipotesi nulla) e l'impunità del vero responsabile del delitto. L'errore di II tipo (non rifiuto dell'ipotesi nulla quando è falsa) comporta invece il proscioglimento del vero colpevole. Nel comune sentire, il primo errore (condanna di un innocente) è più grave del secondo.

Un'analisi clinica per stabilire se un soggetto è sano o malato può condurre a due possibili errori: diagnosi positiva quando il soggetto è sano (falso positivo), diagnosi negativa quando il soggetto è malato (falso negativo). Identificando l'ipotesi nulla con l'affermazione che il soggetto è sano, nella terminologia della verifica delle ipotesi, il falso positivo rappresenta l'errore di primo tipo, mentre il falso negativo costituisce l'errore di II tipo. In questa situazione, tra i due errori, il falso negativo sembra essere il più grave: la malattia non diagnosticata, se grave, può avere esito fatale; a differenza del falso positivo che potrebbe comportare l'avvio di una terapia non necessaria o, più probabilmente, la ripetizione dell'analisi.

7. Inferenza Bayesiana

L'inferenza statistica è un procedimento induttivo con cui, sulla base dei dati campionari osservati, si vuole risalire a caratteristiche non note (il parametro θ) della funzione di probabilità o di densità $f(x;\theta)$ che descrive la popolazione oggetto di studio. Le assunzioni alla base di tale processo induttivo sono state oggetto di dibattito sin dalla elaborazione originaria del problema, che si può far risalire agli studi di Laplace della seconda metà del XVIII secolo.

Con un po' di semplificazione, le principali impostazioni inferenziali sono due: quella *frequentista* o *classica* – adottata anche in questo testo – e quella *Bayesiana*, che prende il nome dal ministro presbiteriano inglese Thomas Bayes (1702-1761), al quale si deve la celebre formula della probabilità condizionata. Se l'impostazione frequentista è, senza dubbio, quella ancora più diffusa nella letteratura e nelle applicazioni, a partire approssimativamente dagli anni '60 del Novecento, l'approccio Bayesiano ha ricevuto nuova attenzione e registra oggi una diffusione molto superiore a quella di pochi decenni fa.

Sia il paradigma frequentista sia quello Bayesiano partono dal presupposto che i dati osservati, $\mathbf{x}_n = (x_1, x_2, \ldots, x_n)$, siano realizzazioni di altrettante variabili casuali, $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$, e che $f(\mathbf{x}_n;\theta)$, la distribuzione di probabilità di $\mathbf{X}_n$, sia nota a meno del parametro θ, il cui vero valore, scalare o vettoriale, appartiene allo spazio Θ. La caratteristica dell'impostazione frequentista è di valutare le procedure inferenziali in base alle loro proprietà distributive.

Gli stimatori puntuali, ad esempio, si valutano attraverso il loro errore quadratico medio; le stime intervallari attraverso i coefficienti di fiducia; le statistiche test in base alle probabilità di errore ad esse associate. In altri termini, le procedure frequentiste vengono valutate e selezionate in funzione del cosiddetto principio del campionamento ripetuto, ossia a seconda del loro comportamento in un'ipotetica successione infinita di repliche dell'esperimento. In questo approccio, il valore di θ si assume incognito, ma non viene trattato come aleatorio.

La caratteristica dell'impostazione Bayesiana consiste invece nel considerare θ non più una costante, ma una variabile casuale e nell'assegnare ad essa una distribuzione di probabilità, $\pi(\theta)$, detta *distribuzione iniziale* o *a priori*, da aggiornare tenendo conto dell'informazione che i dati osservati $\mathbf{x}_n$ forniscono sul parametro. L'inferenza Bayesiana si concretizza quindi nei seguenti passi fondamentali:

(a) scelta di una distribuzione iniziale $\pi(\theta)$, che esprima informazione/incertezza su θ;

(b) osservazione dei dati campionari $\mathbf{x}_n$, determinazione di n variabili casuali aventi distribuzione congiunta condizionata $f(\mathbf{x}_n \mid \theta)$;

(c) determinazione della distribuzione condizionata $\pi(\theta \mid \mathbf{x}_n)$, detta anche *finale* o *a posteriori*, ovvero aggiornamento della distribuzione iniziale $\pi(\theta)$ alla luce dell'informazione contenuta nelle osservazioni $\mathbf{x}_n$.

Il passo (c) della procedura si realizza sfruttando la formula di Bayes che, nel caso in cui Θ sia un sottoinsieme dei numeri reali, assume la forma

$$\pi(\theta \mid \mathbf{x}_n) = \frac{f(\mathbf{x}_n \mid \theta)\,\pi(\theta)}{f(\mathbf{x}_n)},$$

dove $f(\mathbf{x}_n \mid \theta)$ è chiamata *funzione di verosimiglianza* di θ e dove $f(\mathbf{x}_n) = \int_\Theta f(\mathbf{x}_n \mid \theta)\pi(\theta)d\theta$ rappresenta la funzione di probabilità o di densità congiunta di $\mathbf{X}_n$ calcolata nel campione osservato $\mathbf{x}_n$. Dal momento che

$f(\mathbf{x}_n)$ non dipende da θ, le caratteristiche di $\pi(\theta \mid \mathbf{x}_n)$ (ossia la forma, la variabilità, il punto di massimo ecc.) dipendono esclusivamente dal prodotto tra la distribuzione *a priori* $\pi(\theta)$ e la funzione $f(\mathbf{x}_n \mid \theta)$. La distribuzione *a posteriori* $\pi(\theta \mid \mathbf{x}_n)$ sintetizza in termini probabilistici tutta l'informazione disponibile su θ: quella pre-sperimentale rappresentata da $\pi(\theta)$, e quella sperimentale (basata cioè sui dati), incorporata in $f(\mathbf{x}_n \mid \theta)$.

Si consideri il tradizionale esempio del lancio di una moneta. In questo caso, ogni risultato sperimentale può assumere solo i valori $x = 1$ (se esce T) o $x = 0$ (se esce C). Il parametro θ indica la probabilità incognita che in un lancio esca T, ed è quindi un numero non noto dell'intervallo $[0, 1]$. Si supponga che, prima di effettuare l'esperimento e osservare i dati, sia possibile ipotizzare, sulla base di informazioni pregresse, che la moneta sia sbilanciata a favore dell'esito T e che le informazioni disponibili si traducano nella distribuzione iniziale $\pi(\theta)$ rappresentata in Figura 20.14 (grafico a sinistra, curva a puntini), che ha come punto di massimo il valore $\theta = 0.7$ ed è tale da concentrare il 95% della probabilità complessiva sull'intervallo $[0.35, 0.91]$. L'ampiezza di tale intervallo esprime, in termini probabilistici, una sostanziale incertezza iniziale su quale sia il vero valore di θ. Si supponga poi che, effettuati $n = 10$ lanci della moneta, il numero dei successi sia pari a 3. Ciò si traduce nella funzione $f(\mathbf{x}_n \mid \theta)$ (grafico a sinistra, curva a tratti) centrata sul valore di θ uguale a $\bar{x}_n = 0.3$ (*stima di massima verosimiglianza* del parametro). La densità *a posteriori* risultante, $\pi(\theta \mid \mathbf{x}_n)$ (grafico a sinistra, curva continua), è concentrata intorno al valore 0.47 e rappresenta anche graficamente un "compromesso" tra $\pi(\theta)$ e $f(\mathbf{x}_n \mid \theta)$.

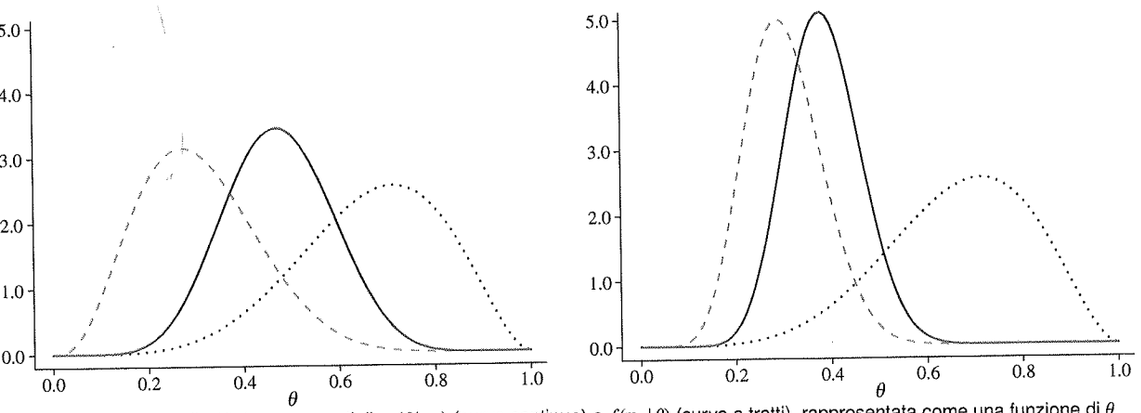


Figura 20.14 Grafici di $\pi(\theta)$ (curva a puntini), $\pi(\theta \mid \mathbf{x}_n)$ (curva continua) e $f(\mathbf{x}_n \mid \theta)$ (curva a tratti), rappresentata come una funzione di θ e per $\bar{x}_n = 0.3$. Grafici a sinistra: $n = 10$; grafici a destra: $n = 30$. ■

Come si può osservare, i grafici a destra nella Figura 20.14 mostrano la funzione di verosimiglianza e la densità *a posteriori* di θ nel caso in cui il numero di osservazioni aumenta ($n = 30$); restano inalterate sia la stima di massima verosimiglianza di θ ($\bar{x}_n = 0.3$) sia la distribuzione iniziale. L'obiettivo dell'impostazione Bayesiana consiste nella derivazione e nell'analisi della distribuzione *a posteriori* che, come detto, ingloba tutta l'informazione disponibile sul parametro incognito θ, sia quella *a priori* sia quella basata sui dati campionari.

Dunque, a parità di informazione iniziale, al crescere della numerosità campionaria, n, il "peso" relativo della funzione di verosimiglianza nella distribuzione *a posteriori* aumenta: quanto più grande è n, tanto più risulterà forte l'attrazione di $\pi(\theta \mid \mathbf{x}_n)$ da parte di $f(\mathbf{x}_n \mid \theta)$ e minore l'impatto di $\pi(\theta)$ nelle conclusioni inferenziali.

La distribuzione *a posteriori* $\pi(\theta \mid \mathbf{x}_n)$ costituisce la base dell'inferenza Bayesiana. Così la stima puntuale si ottiene prendendo una opportuna media *a posteriori* di θ (ad esempio, il valore atteso o la moda). Il corrispettivo dell'intervallo di confidenza dell'impostazione frequentista, detto *intervallo di credibilità*, si ottiene determinando, sempre in base alla distribuzione *a posteriori* $\pi(\theta \mid \mathbf{x}_n)$, due numeri, l_1 e l_2, tali che il parametro θ sia incluso tra l_1 e l_2 con probabilità *a posteriori* molto alta (0.95, 0.99 ecc.), con il vincolo, ad esempio, che le "aree alle code" siano uguali. Nell'esempio precedente, come intervallo di credibilità al 95% (caso $n = 10$) si trova $(0.26, 0.69)$, intervallo che rende molto meno plausibili i valori superiori a 0.8 rispetto a quanto ipotizzato *a priori*.

Quanto alla verifica dell'ipotesi H_0: $\theta \leq \theta_0$ contro l'alternativa H_1: $\theta > \theta_0$, la procedura inferenziale consiste nella determinazione delle probabilità *a posteriori* $P(H_0 \mid \mathbf{x}_n)$ e $P(H_1 \mid \mathbf{x}_n)$ e nel rifiutare l'ipotesi nulla se $P(H_0 \mid \mathbf{x}_n) < P(H_1 \mid \mathbf{x}_n)$. Nell'esempio, ponendo $\theta_0 = 0.5$, la probabilità di H_0 è $P(H_0 \mid \mathbf{x}_n) = 0.14$, mentre la probabilità di H_1 è $P(H_1 \mid \mathbf{x}_n) = 0.86$, il che porta a scegliere H_1.

Esercizi

20.1 Il contenuto medio di nicotina dichiarato di una certa marca di sigarette è 0.25 milligrammi con una deviazione standard di 0.015. Un'associazione di consumatori ritiene che il contenuto di nicotina dichiarato sia al di sotto di quello effettivo. Si sottoponga a verifica l'appropriato sistema di ipotesi, sapendo che in un campione casuale di 30 sigarette si è osservata una media campionaria pari a 0.264. Si ponga $\alpha = 0.01$.

20.2 I dati seguenti si riferiscono alla durata (in migliaia di chilometri) di un convertitore catalitico in un campione di 15 osservazioni:

115.4, 85.2, 89.1, 118.3, 88.4, 109.3, 104.3, 69.3,
105.5, 106.8, 103.1, 101.6, 102.9, 89.6, 109.3.

Si verifichi l'ipotesi che la durata media sia pari a 100 contro l'alternativa che essa sia minore, supponendo che la popolazione sia normale. Si assuma $\alpha = 0.05$.

20.3 Da una formazione geologica vengono estratte 40 piccole porzioni di roccia che vengono successivamente sottoposte ad analisi per verificare il contenuto percentuale di cadmio. Si osserva una percentuale media di 17.4 con una deviazione standard s pari a 4.2. Si supponga che l'estrazione di minerale sia economicamente conveniente se il contenuto medio percentuale di cadmio nella formazione è maggiore di 15.

a) Si specifichi il sistema di ipotesi.

b) Si stabilisca se le osservazioni campionarie suffragano la convenienza economica dello sfruttamento del giacimento (si ponga $\alpha = 0.01$).

c) Si ripeta il punto precedente utilizzando il livello di significatività osservato.

20.4 Di seguito sono riportati i chilometri percorsi in un giorno da un campione di taxi operanti in una grande città:

172, 195, 115, 122, 154, 149, 120, 148,
152, 68, 132, 91, 120, 148, 103, 101.

Assumendo che la popolazione sia normale, si verifichi l'ipotesi che la percorrenza media giornaliera per la totalità dei taxi della città sia 125 contro l'alternativa che sia superiore ($\alpha = 0.01$).

20.5 Il produttore di una certa marca di pillole sostiene che la potenza media delle stesse dopo un'esposizione all'umidità o al calore della durata di una settimana è almeno pari a 69. Il distributore, prima di acquistare una partita di quel farmaco, osserva la potenza in un campione casuale di 120 pillole dopo averle sottoposte all'esposizione sopra specificata, riscontrando una potenza media di 66.8 con una deviazione standard campionaria di 10.6.

a) Si specifichi il sistema di ipotesi.

b) Si sottoponga a verifica l'ipotesi nulla a un livello di significatività del 5%.

20.6 In un campione casuale di 342 famiglie di una certa città è stata rilevata la spesa mensile destinata all'acquisto di carne, ottenendo una media di 240 euro e una deviazione standard campionaria pari a 41.6. Si verifichi l'ipotesi che la spesa mensile media nella popolazione sia pari a 230 contro l'alternativa che sia superiore. Si utilizzi il livello di significatività osservato e si ponga $\alpha = 0.01$.

20.7 Il fornitore di pneumatici di una grande impresa di trasporti sostiene che la durata media di un certo tipo di pneumatici per camion è di 45 000 km. L'impresa sottopone a test l'affermazione del fornitore osservando un campione di 56 pneumatici utilizzati dai propri veicoli. Qual è la conclusione a cui giunge l'impresa se trova una durata media di 43 740 km con una deviazione standard campionaria di 2749 km? Si usi la tecnica del livello di significatività osservato, ponendo $\alpha = 0.01$.

20.8 Un'industria farmaceutica afferma che un proprio farmaco è efficace in almeno l'80% dei casi. L'osservazione su un campione casuale di 180 persone che lo hanno utilizzato ne ha rivelato l'efficacia in 148 casi.

a) Si specifichino le ipotesi nulla e alternativa.

b) Si effettui il test di ipotesi con un livello di significatività del 5%.

c) Si ripeta il punto b usando il livello di significatività osservato.

20.9 Una moneta viene lanciata 80 volte, ottenendo 45 volte l'esito testa. Vi è sufficiente evidenza per ritenere che la moneta sia truccata? Si ponga $\alpha = 0.05$. (*Suggerimento*: si tratta di verificare l'ipotesi nulla che la probabilità dell'evento testa sia 0.5 contro l'alternativa che sia diversa da 0.5.)

20.10 L'ufficio statistico di un comune ha condotto un sondaggio tra i cittadini maggiorenni per accertare il gradimento di un nuovo sistema di raccolta differenziata dei rifiuti. In un campione casuale di 530 individui, l'ufficio statistico trova che il 68.4% degli intervistati si dimostra soddisfatto. Si verifichi l'ipotesi che il 70% dei cittadini sia soddisfatto contro l'alternativa che tale percentuale sia inferiore, assumendo un livello di significatività del 10%.

20.11 Un fornitore di semilavorati assicura all'azienda cliente che i pezzi difettosi nell'ultima fornitura non superano il 2%. L'addetto al controllo di accettazione dell'azienda osserva un campione di 85 unità e trova che la percentuale di pezzi difettosi è 2.15%. Si sottoponga a verifica l'ipotesi $H_0: p = 0.02$ a un livello di significatività del 5%.

20.12 Uno strumento di precisione è garantito per una precisione di lettura dell'ordine di ± 2 unità. L'impiego di 10 esemplari dello strumento per misurare uno stesso oggetto ha prodotto i seguenti risultati:

451, 453, 452, 451, 453, 452, 451, 454, 453, 452.

Si verifichi l'ipotesi che la varianza del grado di precisione dello strumento sia uguale a 0.9 contro l'alternativa che sia superiore. Si prenda $\alpha = 0.05$.

20.13 Il peso dei pacchetti di riso confezionati automaticamente in un'azienda del settore alimentare è assimilabile a una variabile casuale normale con varianza 4.35. Il responsabile del controllo di processo ritiene che una confezionatrice dia luogo a una variabilità dei pesi superiore a 4.35. In un campione casuale di 25 confezioni, tratte dalla confezionatrice sotto osservazione, si è trovata una varianza campionaria pari a $s^2 = 5.11$.

a) Vi è sufficiente evidenza per sostenere che la varianza osservata differisce significativamente da quella ipotizzata per la popolazione? Si ponga $\alpha = 0.05$.

b) Si ripeta il punto precedente utilizzando il livello di significatività osservato.

20.14* Nel problema della verifica delle ipotesi sulla varianza di una popolazione normale siano $H_0: \sigma^2 = \sigma_0^2$ l'ipotesi nulla e $H_1: \sigma^2 > \sigma_0^2$ l'ipotesi alternativa. Nel caso in cui la media della popolazione è nota, si costruisca la zona di rifiuto utilizzando la statistica test $\sum_{i=1}^{n}(X_i - \mu)^2/\sigma_0^2$ che ha distribuzione chi-quadrato con n gradi di libertà.

20.15 Con riferimento all'Esercizio 20.10, si determini la potenza del test per l'ipotesi alternativa che la percentuale dei favorevoli sia 65%.

20.16 Per una generica voce dell'inventario di un'impresa mercantile sia X la differenza tra valore inventariato e valore certificato. Da un campione di 120 voci, un certificatore contabile ha ottenuto $\bar{x} = 25.3$ e $s^2 = 13\,240$.

a) Si sottoponga a verifica l'ipotesi $\mu = 0$ contro l'alternativa $\mu > 0$ (cioè inventario gonfiato), con $\alpha = 0.01$.

b) Si calcoli la potenza del test di cui al punto precedente per l'ipotesi alternativa $H_1: \mu = 30$, supponendo che la varianza della popolazione sia $\sigma^2 = 12\,500$.

20.17 Con riferimento all'Esercizio 20.3, si determini la probabilità dell'errore di II tipo per il sistema di ipotesi $H_0: \mu = 15$, $H_1: \mu = 16$, con $\alpha = 0.01$, assumendo che la deviazione standard della popolazione sia $\sigma = 4$.

20.18 Con riferimento all'Esercizio 20.9, si determini la potenza del test per l'ipotesi alternativa $p = 0.55$.

20.19 Con riferimento all'Esercizio 20.13, si determini la potenza del test per l'ipotesi alternativa $\sigma^2 = 7$.

20.20* Si consideri il problema della verifica delle ipotesi sulla media di una popolazione normale con varianza nota. Siano

$$H_0: \mu = \mu_0, \quad H_1: \mu = \mu_1, \quad \text{con} \quad \mu_1 > \mu_0,$$

l'ipotesi nulla e quella alternativa. Si dimostri che la potenza del test basato su $\bar{X}$ è una funzione crescente della dimensione campionaria n. (*Suggerimento*: si consideri che, nel caso in esame, la potenza del test è data da

$$\pi(\mu_1) = P\left(\bar{X} > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}; \mu_1\right) = P\left(\frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} > -\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} + z_{1-\alpha}; \mu_1\right),$$

dove la quantità

$$-\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}$$

decresce al crescere di n.)

20.21* Si consideri il problema della verifica dell'ipotesi $H_0: \mu = \mu_0$ contro l'alternativa $H_1: \mu \neq \mu_0$ nel caso di una popolazione normale con varianza nota. Si dimostri che la dimensione campionaria che assicura, approssimativamente, il livello di potenza π, per due specifici valori di μ_0 e μ_1, è dato dalla formula ottenuta nell'Esempio 20.24 per il test unidirezionale a destra sostituendo $z_{1-\alpha/2}$ a $z_{1-\alpha}$. (*Suggerimento*: si parta dalla funzione di potenza data nell'Esempio 20.23 e si consideri che

$$\Phi\left(-\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} - z_{1-\alpha/2}\right) \approx 0,$$

se $\mu_1 > \mu_0$, mentre

$$\Phi\left(-\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} + z_{1-\alpha/2}\right) \approx 1,$$

se $\mu_1 < \mu_0$.)

20.22* Si consideri il problema della verifica dell'ipotesi $H_0 : p = p_0$ contro l'alternativa $H_1 : p > p_0$. Si dimostri che la dimensione campionaria che assicura un livello di potenza π quando $p = p_1$, con $p_1 > p_0$, è data da

$$n \approx \frac{\left[z_{1-\pi}\sqrt{p_1(1-p_1)} - z_{1-\alpha}\sqrt{p_0(1-p_0)}\right]^2}{(p_1 - p_0)^2}.$$

CAPITOLO

21

Confronti tra due popolazioni

Sommario del capitolo

21.1 Introduzione
21.2 Inferenza sulle medie di due popolazioni normali
21.3 Inferenza sulle medie di due popolazioni nel caso di grandi campioni
21.4 Verifica dell'ipotesi di uguaglianza delle varianze di due popolazioni normali
Riassunto
Commenti e curiosità
Esercizi

21.1 Introduzione

I Capitoli 19 e 20 hanno avuto per oggetto gli intervalli di confidenza e la verifica di ipotesi sui parametri media e varianza di una singola popolazione. In questo capitolo, si applicheranno queste tecniche inferenziali per confrontare medie e varianze di due popolazioni.

L'esigenza di mettere a confronto due popolazioni emerge, ad esempio, negli studi clinici controllati volti a stabilire l'efficacia di un nuovo trattamento e condotti su due campioni casuali indipendenti costituiti dal gruppo sperimentale che riceve il nuovo trattamento e dal gruppo di controllo a cui viene somministrato il placebo.

Il capitolo è organizzato nel modo seguente. Verranno dapprima illustrate le procedure di inferenza per il confronto tra le medie di due popolazioni normali. Poi l'analisi sarà estesa al caso di due popolazioni qualsiasi, nel presupposto che le dimensioni campionarie siano sufficientemente grandi. In questo ambito, si esaminerà anche il caso di due popolazioni Bernoulliane. Si affronterà infine il problema della verifica dell'ipotesi di uguaglianza delle varianze di due popolazioni normali.

Si indicheranno con:

- $\mathcal{P}_1$ e $\mathcal{P}_2$ le popolazioni oggetto del confronto;
- μ_1 e μ_2 le medie di $\mathcal{P}_1$ e $\mathcal{P}_2$;
- σ_1^2 e σ_2^2 le varianze di $\mathcal{P}_1$ e $\mathcal{P}_2$;
- C_1 e C_2 i due campioni indipendenti provenienti da $\mathcal{P}_1$ e $\mathcal{P}_2$;
- n_1 e n_2 le dimensioni di C_1 e C_2;
- $\bar{X}_1$ e $\bar{X}_2$ le medie di C_1 e C_2;
- S_1^2 e S_2^2 le varianze di C_1 e C_2.

21.2 Inferenza sulle medie di due popolazioni normali

Ai fini dell'inferenza sulle medie di due popolazioni normali, è cruciale conoscere la distribuzione campionaria della differenza tra le medie di due campioni indipendenti provenienti da tali popolazioni. Con la Proposizione 21.1 viene stabilita la normalità della distribuzione della variabile casuale $\bar{X}_1 - \bar{X}_2$ su cui è basato il confronto tra le due popolazioni.

PROPOSIZIONE 21.1

Siano $\bar{X}_1$ e $\bar{X}_2$ le medie di due campioni indipendenti C_1 e C_2 provenienti dalle popolazioni $\mathcal{P}_1$ e $\mathcal{P}_2$ distribuite normalmente. Allora anche la variabile casuale $\bar{X}_1 - \bar{X}_2$ è distribuita normalmente con media

$$\mathrm{E}(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$$

e varianza

$$\sigma^2_{\bar{X}_1-\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Ne segue che la differenza standardizzata

$$Z_{\bar{X}_1-\bar{X}_2} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{X}_1-\bar{X}_2}} \qquad (21.1)$$

ha distribuzione normale standard.

La dimostrazione che $\bar{X}_1 - \bar{X}_2$ è distribuita normalmente segue immediatamente dall'applicazione della Proposizione 15.4 per $n = 2$, $a_1 = 1$ e $a_2 = -1$. ■

Il risultato appena esposto è illustrato nella Figura 21.1.

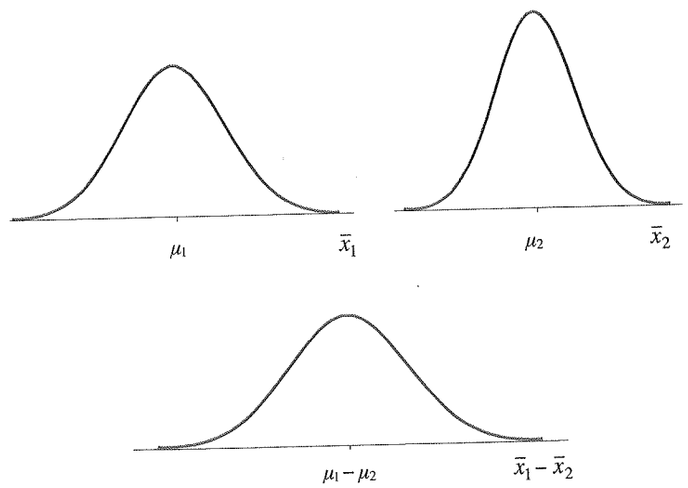


Figura 21.1
Distribuzioni campionarie di $\bar{X}_1$, $\bar{X}_2$ e $\bar{X}_1 - \bar{X}_2$.

Stima intervallare

Sia $\delta = \mu_1 - \mu_2$ la differenza tra le medie delle popolazioni $\mathcal{P}_1$ e $\mathcal{P}_2$, normali con varianze note. Per la costruzione dell'intervallo di confidenza, si può impiegare lo stimatore naturale di δ, $\bar{X}_1 - \bar{X}_2$.

In forza della Proposizione 21.1, si ha

$$P\left[-z_{1-\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{X}_1-\bar{X}_2}} < z_{1-\alpha/2}\right] = 1 - \alpha.$$

Risolvendo le disequazioni rispetto a $\mu_1 - \mu_2$, è facile stabilire che

$$P[(\bar{X}_1 - \bar{X}_2) - z_{1-\alpha/2}\,\sigma_{\bar{X}_1-\bar{X}_2} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + z_{1-\alpha/2}\,\sigma_{\bar{X}_1-\bar{X}_2}] = 1 - \alpha,$$

da cui si ricavano gli estremi dell'intervallo di confidenza per $\mu_1 - \mu_2$ all'$(1 - \alpha)100\%$:

$$l_1 = (\bar{x}_1 - \bar{x}_2) - z_{1-\alpha/2}\,\sigma_{\bar{X}_1-\bar{X}_2},$$

$$l_2 = (\bar{x}_1 - \bar{x}_2) + z_{1-\alpha/2}\,\sigma_{\bar{X}_1-\bar{X}_2}.$$

ESEMPIO 21.1

In una prova sul carico di rottura di due tipi di corda, si dispone di due campioni di ampiezza 26 e 35, le cui medie sono 185.3 kg e 175.2 kg. Supponendo che le deviazioni standard siano note e pari a 14.8 e 10.6, si determini l'intervallo di confidenza al 95% per la differenza delle medie.

Soluzione

Occorre innanzitutto calcolare la deviazione standard della differenza tra le medie campionarie. Si ha

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{14.8^2}{26} + \frac{10.6^2}{35}} = 3.41.$$

Gli estremi dell'intervallo di confidenza sono allora:

$$l_1 = (185.3 - 175.2) - 1.96 \cdot 3.41 = 3.42,$$

$$l_2 = (185.3 - 175.2) + 1.96 \cdot 3.41 = 16.78.$$

Si può dunque concludere che la differenza tra le medie delle due popolazioni è compresa nell'intervallo (3.42, 16.78) con una fiducia del 95%. ■

Verifica di ipotesi

L'ipotesi nulla che si intende sottoporre a verifica è $H_0 : \delta = \mu_1 - \mu_2 = 0$. L'ipotesi alternativa può essere unidirezionale, $H_1 : \delta > 0$ o $H_1 : \delta < 0$, oppure bidirezionale, $H_1 : \delta \neq 0$. Di seguito si considera in dettaglio solamente il primo caso.

La statistica test naturale è la differenza $\bar{X}_1 - \bar{X}_2$. Se l'ipotesi nulla è vera, il valore atteso di questa variabile casuale è zero,

$$\mathrm{E}(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2 = 0,$$

cosicché la differenza standardizzata

$$Z_{\bar{X}_1-\bar{X}_2} = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1-\bar{X}_2}}$$

ha, per la Proposizione 21.1, distribuzione normale standard.

In analogia con quanto illustrato in merito alla verifica di ipotesi sulla media di una sola popolazione (Paragrafo 20.3), si rifiuta l'ipotesi nulla se la statistica test $Z_{\bar{X}_1-\bar{X}_2}$ assume valori molto elevati, cioè maggiori di $z_{1-\alpha}$. La logica sottostante a questa regola di decisione è che valori molto grandi della statistica test non sono verosimilmente attribuibili al caso, ma al fatto che $\mu_1 > \mu_2$, come contemplato dall'ipotesi alternativa.

Le zone di rifiuto per le altre due ipotesi alternative, $H_1 : \delta = \mu_1 - \mu_2 < 0$ e $H_1 : \delta = \mu_1 - \mu_2 \neq 0$, si ottengono in modo analogo (si veda la Tabella 21.1).

abella 21.1
one di rifiuto della
tatistica test
$\bar{X}_1 - \bar{X}_2 = (\bar{X}_1 - \bar{X}_2)/\sigma_{\bar{X}_1 - \bar{X}_2}$
ella verifica di ipotesi
i uguaglianza delle medie
i due popolazioni normali
on varianze note.

Ipotesi H_0	Ipotesi H_1	Statistica test	Zona di rifiuto
	$\delta = \mu_1 - \mu_2 > 0$		$\{z_{\bar{x}_1 - \bar{x}_2} > z_{1-\alpha}\}$
$\mu_1 - \mu_2 = 0$	$\delta = \mu_1 - \mu_2 < 0$	$Z_{\bar{X}_1 - \bar{X}_2} = \dfrac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}}$	$\{z_{\bar{x}_1 - \bar{x}_2} < -z_{1-\alpha}\}$
	$\delta = \mu_1 - \mu_2 \neq 0$		$\{\lvert z_{\bar{x}_1 - \bar{x}_2}\rvert > z_{1-\alpha/2}\}$

Le zone di rifiuto possono anche essere espresse in funzione della statistica $\bar{X}_1 - \bar{X}_2$. Per esempio, se si considera l'ipotesi alternativa $H_1 : \delta = \mu_1 - \mu_2 > 0$, la zona di rifiuto può essere scritta come

$$\{\bar{x}_1 - \bar{x}_2 : \bar{x}_1 - \bar{x}_2 > z_{1-\alpha}\,\sigma_{\bar{X}_1 - \bar{X}_2}\}.$$

Come illustrato nel Capitolo 20, per decidere se rifiutare o meno l'ipotesi nulla, si può fare ricorso al livello di significatività osservato. Ad esempio, per l'ipotesi alternativa $H_1 : \delta > 0$, esso è espresso da

$$\alpha_{oss} = P\left(Z_{\bar{X}_1 - \bar{X}_2} > \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}};\ \delta = 0\right),$$

dove $\bar{x}_1 - \bar{x}_2$ è la differenza tra le medie campionarie osservate. L'ipotesi nulla viene poi rifiutata se $\alpha_{oss} < \alpha$.

ESEMPIO 21.2

Riprendendo i dati dell'Esempio 21.1, si verifichi l'ipotesi che le medie delle due popolazioni siano uguali contro l'alternativa che siano diverse, a un livello di significatività del 5%.

Soluzione

La zona di rifiuto è data da

$$\{z_{\bar{x}_1 - \bar{x}_2} : |z_{\bar{x}_1 - \bar{x}_2}| > 1.96\},$$

dove

$$z_{\bar{x}_1 - \bar{x}_2} = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}}.$$

La statistica test assume il valore

$$z_{\bar{x}_1 - \bar{x}_2} = \frac{185.3 - 175.2}{\sqrt{\dfrac{14.8^2}{26} + \dfrac{10.6^2}{35}}} = 2.96,$$

che appartiene alla zona di rifiuto e porta a rifiutare l'ipotesi nulla. In altri termini, vi è sufficiente evidenza per ritenere che la differenza tra le medie osservate nei campioni non sia attribuibile al solo effetto del caso.

Per il livello di significatività osservato, occorre calcolare la probabilità

$$\alpha_{oss} = P(|Z_{\bar{X}_1 - \bar{X}_2}| > 2.96;\ \delta = 0) = 2[1 - \Phi(2.96)] = 0.003,$$

che, essendo minore di 0.05, porta al rifiuto dell'ipotesi nulla. ■

21.2.1 Varianze non note e piccoli campioni

Quando le varianze σ_1^2 e σ_2^2 non sono note, è necessario stimarle con i dati dei campioni. Se le dimensioni campionarie non sono sufficientemente grandi, le tecniche di inferenza danno origine a variabili casuali distribuite, esattamente o approssimativamente, secondo una T di Student, in analogia con quanto visto nel Paragrafo 20.4 per una singola popolazione. Qui di seguito viene affrontato il caso particolare in cui le varianze incognite si suppongono uguali: $\sigma_1^2 = \sigma_2^2$.

PROPOSIZIONE 21.2

Siano C_1 e C_2 due campioni casuali indipendenti provenienti dalle popolazioni $\mathcal{P}_1$ e $\mathcal{P}_2$ distribuite normalmente con varianze incognite, ma uguali. Sia

$$S_c^2 = \frac{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

la varianza campionaria combinata, data dalla media aritmetica ponderata delle due varianze campionarie con pesi uguali ai gradi di libertà. Allora la variabile casuale

$$T_{\bar{X}_1 - \bar{X}_2} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_c\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \tag{21.2}$$

ha distribuzione T di Student con $n_1 + n_2 - 2$ gradi di libertà.

Per la dimostrazione si rinvia a Casella e Berger (2002), p. 223. ■

Stima intervallare

Sulla base della Proposizione 21.2, si ha

$$P\left[-t_{1-\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_c\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} < t_{1-\alpha/2}\right] = 1 - \alpha,$$

dove $-t_{1-\alpha/2}$ e $t_{1-\alpha/2}$ sono gli appropriati quantili della distribuzione T di Student con $n_1 + n_2 - 2$ gradi di libertà. Con il solito procedimento, dall'equazione di cui sopra, è possibile ottenere gli estremi dell'intervallo di confidenza all'$(1-\alpha)100\%$ per la differenza $\delta = \mu_1 - \mu_2$:

$$l_1 = (\bar{x}_1 - \bar{x}_2) - t_{1-\alpha/2}\,s_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \quad l_2 = (\bar{x}_1 - \bar{x}_2) + t_{1-\alpha/2}\,s_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

ESEMPIO 21.3

In un'azienda addetta all'imballaggio di una certa merce sono in uso due macchinari diversi. Ci si chiede se i tempi di esecuzione dell'imballaggio siano differenti per i due macchinari. A questo scopo, vengono osservati i tempi di esecuzione (in secondi) di 10 operazioni di imballaggio. Per il primo macchinario si ottiene un

tempo medio $\bar{x}_1 = 55$ con una deviazione standard $s_1 = 1.4$, per il secondo un tempo medio $\bar{x}_2 = 53$ con una deviazione standard $s_2 = 1.5$. Si costruisca l'intervallo di confidenza al 95% per $\delta = \mu_1 - \mu_2$ (differenza tra i tempi medi di esecuzione dell'imballaggio con i due macchinari).

Soluzione

La deviazione standard congiunta osservata è

$$s_c = \sqrt{\frac{1.4^2 \cdot 9 + 1.5^2 \cdot 9}{18}} = 1.45.$$

L'appropriato quantile della distribuzione T di Student è $t_{0.975} = 2.101$. Gli estremi dell'intervallo di confidenza sono allora dati da:

$$l_1 = (55-53) - 2.101 \cdot 1.45 \sqrt{\frac{2}{10}} = 0.64, \quad l_2 = (55-53) + 2.101 \cdot 1.45 \sqrt{\frac{2}{10}} = 3.36.$$

La differenza $\mu_1 - \mu_2$ è dunque compresa tra 0.64 e 3.36 con una fiducia del 95%. ■

Verifica di ipotesi

Si consideri il problema della verifica dell'ipotesi nulla $H_0: \delta = \mu_1 - \mu_2 = 0$ contro l'alternativa $H_1: \delta = \mu_1 - \mu_2 > 0$. Una statistica test per questa situazione si ottiene ponendo $\mu_1 - \mu_2 = 0$ nella (21.2):

$$T_{\bar{X}_1 - \bar{X}_2} = \frac{\bar{X}_1 - \bar{X}_2}{S_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Si rifiuta l'ipotesi nulla se $T_{\bar{X}_1 - \bar{X}_2}$ assume un valore positivo molto elevato. Con le usuali argomentazioni, la zona di rifiuto viene definita come

$$\{t_{\bar{x}_1 - \bar{x}_2} : t_{\bar{x}_1 - \bar{x}_2} > t_{1-\alpha}\},$$

dove

$$t_{\bar{x}_1 - \bar{x}_2} = \frac{\bar{x}_1 - \bar{x}_2}{s_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

e $t_{1-\alpha}$ è il quantile di livello $1 - \alpha$ della distribuzione T di Student con $n_1 + n_2 - 2$ gradi di libertà.

Le zone di rifiuto per le altre due ipotesi alternative possono essere determinate in modo analogo (si veda la Tabella 21.2).

Tabella 21.2
Zone di rifiuto della statistica test $T_{\bar{X}_1 - \bar{X}_2}$ nella verifica dell'ipotesi di uguaglianza delle medie di due popolazioni normali con varianze incognite, ma uguali.

Ipotesi H_0	Ipotesi H_1	Statistica test	Zona di rifiuto
$\mu_1 - \mu_2 = 0$	$\delta = \mu_1 - \mu_2 > 0$	$T_{\bar{X}_1 - \bar{X}_2} = \dfrac{\bar{X}_1 - \bar{X}_2}{S_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$\{t_{\bar{x}_1 - \bar{x}_2} > t_{1-\alpha}\}$
	$\delta = \mu_1 - \mu_2 < 0$		$\{t_{\bar{x}_1 - \bar{x}_2} < -t_{1-\alpha}\}$
	$\delta = \mu_1 - \mu_2 \neq 0$		$\{\lvert t_{\bar{x}_1 - \bar{x}_2} \rvert > t_{1-\alpha/2}\}$

ESEMPIO 21.4

Un'industria produttrice di lampadine elettriche intende sottoporre a test la durata di una lampadina di nuova progettazione. Viene allora formulata l'ipotesi nulla che la durata media, μ_1, delle lampadine di nuovo tipo sia uguale a quella, μ_2, delle lampadine di vecchio tipo; come ipotesi alternativa si assume $H_1: \delta = \mu_1 - \mu_2 > 0$. Per verificare l'ipotesi, viene osservato un campione di ampiezza 31 per ciascun tipo di lampadina. Le medie e le varianze dei due campioni sono:

$$\bar{x}_1 = 1195.16, \quad s_1^2 = 118.13, \quad \bar{x}_2 = 1180.05, \quad s_2^2 = 124.34.$$

Si verifichi l'ipotesi a un livello di significatività dell'1%, assumendo che le varianze delle popolazioni siano uguali ($\sigma_1^2 = \sigma_2^2 = \sigma^2$).

Soluzione

La zona di rifiuto è data da

$$\{t_{\bar{x}_1 - \bar{x}_2} : t_{\bar{x}_1 - \bar{x}_2} > t_{0.99}\},$$

dove

$$t_{\bar{x}_1 - \bar{x}_2} = \frac{\bar{x}_1 - \bar{x}_2}{s_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

e $t_{0.99} = 2.390$ (i gradi di libertà sono 60). La statistica test è data da

$$t_{\bar{x}_1 - \bar{x}_2} = \frac{1195.16 - 1180.05}{\sqrt{121.23\left(\frac{1}{31} + \frac{1}{31}\right)}} = 5.40,$$

in cui 121.23 è la varianza campionaria combinata calcolata come segue

$$s_c^2 = \frac{118.13 \cdot 30 + 124.34 \cdot 30}{60} = 121.23.$$

La statistica test osservata fa parte della zona di rifiuto. L'ipotesi nulla viene quindi rifiutata.

■

Si faccia ora cadere l'ipotesi di uguaglianza delle varianze σ_1^2 e σ_2^2. In questo caso, tali varianze vanno stimate singolarmente, venendo meno il presupposto per l'uso dello stimatore S_c^2. La variabile casuale a cui fare riferimento per il confronto tra due medie, in sostituzione della (21.2), è allora

$$T'_{\bar{X}_1 - \bar{X}_2} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

la cui distribuzione campionaria può essere approssimata con una T di Student con

$$\nu = \frac{(a+b)^2}{a^2/(n_1 - 1) + b^2/(n_2 - 1)}$$

gradi di libertà, essendo $a = S_1^2/n_1$ e $b = S_2^2/n_2$.[1]

L'applicazione di $T'_{\bar{X}_1-\bar{X}_2}$ per costruire l'intervallo di confidenza per la differenza $\mu_1 - \mu_2$ o per verificare l'ipotesi che quest'ultima sia 0 segue le stesse linee del caso precedente. Per brevità, si dà solo un esempio dell'intervallo di confidenza per $\mu_1 - \mu_2$.

ESEMPIO 21.5

In una clinica vengono proposti due diversi programmi di dimagrimento. Con il programma A, in un campione di 13 pazienti, è stata registrata una diminuzione media del peso di 9 kg con una deviazione standard pari a 5.3. Con il programma B, in un campione di 20 pazienti, è stata registrata una diminuzione media del peso di 6.2 kg con una deviazione standard di 4.6. Si costruisca l'intervallo di confidenza al 95% per la differenza tra le medie delle due popolazioni, senza assumere che queste abbiano varianze uguali.

Soluzione

Con l'usuale procedimento applicato questa volta alla variabile casuale $T'_{\bar{X}_1-\bar{X}_2}$, si ottengono facilmente gli estremi dell'intervallo di confidenza al 95% :

$$l_1 = (9 - 6.2) - 2.069\sqrt{\frac{5.3^2}{13} + \frac{4.6^2}{20}} = -0.91,$$

$$l_2 = (9 - 6.2) + 2.069\sqrt{\frac{5.3^2}{13} + \frac{4.6^2}{20}} = 6.51,$$

dove 2.069 è il quantile di livello 0.975 della distribuzione T di Student con 23 gradi di libertà. Questo numero è dato da

$$\nu = \frac{\left(\frac{5.3^2}{13} + \frac{4.6^2}{20}\right)^2}{\left(\frac{5.3^2}{13}\right)^2 \frac{1}{12} + \left(\frac{4.6^2}{20}\right)^2 \frac{1}{19}} = 23.$$

■

21.3 Inferenza sulle medie di due popolazioni nel caso di grandi campioni

L'assunzione di normalità è essenziale per le tecniche di inferenza fin qui presentate riguardo alla differenza tra le medie di due popolazioni. Se i due campioni hanno dimensioni sufficientemente grandi, tale assunzione non è più necessaria, e tanto meno è necessario ammettere che le varianze delle due popolazioni siano uguali. In questo caso, la variabile casuale

[1] La ricerca della distribuzione di probabilità della variabile casuale $T'_{\bar{X}_1-\bar{X}_2}$ è nota in letteratura come problema di Behrens-Fisher. Una delle soluzioni proposte approssima la distribuzione di $T'_{\bar{X}_1-\bar{X}_2}$ con una distribuzione T di Student con un numero di gradi di libertà espresso da una specifica funzione delle varianze campionarie s_1^2 e s_2^2 (si veda Rohatgi e Saleh, 2015, p. 486).

$$Z_{\bar{X}_1-\bar{X}_2} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{\bar{X}_1-\bar{X}_2}}, \tag{21.3}$$

dove

$$\hat{\sigma}_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},$$

ha distribuzione prossima alla normale standard, indipendentemente dalla forma e dalle varianze delle popolazioni a confronto. La (21.3) ha una struttura analoga alla (21.1), dalla quale si può derivare sostituendo al denominatore σ_1^2 e σ_2^2 con i rispettivi stimatori S_1^2 e S_2^2.

Stima intervallare

La distribuzione campionaria della variabile casuale $Z_{\bar{X}_1-\bar{X}_2}$ definita dalla (21.3) consente di scrivere

$$P\left[-z_{1-\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{\bar{X}_1-\bar{X}_2}} < z_{1-\alpha/2}\right] \approx 1 - \alpha,$$

da cui è facile determinare gli estremi dell'intervallo di confidenza per la differenza $\mu_1 - \mu_2$:

$$l_1 = (\bar{x}_1 - \bar{x}_2) - z_{1-\alpha/2}\,\hat{\sigma}_{\bar{X}_1-\bar{X}_2},$$

$$l_2 = (\bar{x}_1 - \bar{x}_2) + z_{1-\alpha/2}\,\hat{\sigma}_{\bar{X}_1-\bar{X}_2}.$$

ESEMPIO 21.6

In un campione di 64 famiglie di tre componenti nella città A, il numero medio di automobili possedute è pari a 2.1 con una deviazione standard $s_1 = 0.29$. In un campione di 75 famiglie della stessa dimensione nella città B, il numero medio di automobili possedute è 1.7 con una deviazione standard $s_2 = 0.24$. Si determini l'intervallo di confidenza al 99% per la differenza tra le medie delle due popolazioni.

Soluzione

Si procede al calcolo della stima

$$\hat{\sigma}_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{0.29^2}{64} + \frac{0.24^2}{75}} = 0.047.$$

Gli estremi dell'intervallo di confidenza sono dati da:

$$l_1 = (2.1 - 1.7) - 2.576 \cdot 0.047 = 0.28,$$

$$l_2 = (2.1 - 1.7) + 2.576 \cdot 0.047 = 0.52.$$

Si può allora concludere che la differenza $\mu_1 - \mu_2$ è compresa nell'intervallo (0.28, 0.52) con una fiducia del 99%.

■

Verifica di ipotesi

Se l'ipotesi nulla è vera, dalla (21.3) si trae la statistica test

$$Z_{\bar{X}_1-\bar{X}_2} = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_{\bar{X}_1-\bar{X}_2}},$$

con cui, con il solito procedimento, si individuano le zone di rifiuto riportate nella Tabella 21.3.

Tabella 21.3
Zone di rifiuto della statistica test $Z_{\bar{X}_1-\bar{X}_2} = (\bar{X}_1 - \bar{X}_2)/\hat{\sigma}_{\bar{X}_1-\bar{X}_2}$ nella verifica dell'ipotesi di uguaglianza delle medie di due popolazioni qualsiasi, nel caso di grandi campioni.

Ipotesi H_0	Ipotesi H_1	Statistica test	Zona di rifiuto
$\mu_1 - \mu_2 = 0$	$\delta = \mu_1 - \mu_2 > 0$	$Z_{\bar{X}_1-\bar{X}_2} = \dfrac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_{\bar{X}_1-\bar{X}_2}}$	$\{z_{\bar{x}_1-\bar{x}_2} > z_{1-\alpha}\}$
	$\delta = \mu_1 - \mu_2 < 0$		$\{z_{\bar{x}_1-\bar{x}_2} < -z_{1-\alpha}\}$
	$\delta = \mu_1 - \mu_2 \neq 0$		$\{\lvert z_{\bar{x}_1-\bar{x}_2}\rvert > z_{1-\alpha/2}\}$

ESEMPIO 21.7

Con i dati dell'Esempio 21.6, si verifichi l'ipotesi che il numero medio di automobili possedute sia lo stesso nelle due città, contro l'alternativa che sia superiore nella città A. Si impieghi un livello di significatività dell'1%.

Soluzione

La statistica test assume il valore

$$z_{\bar{x}_1-\bar{x}_2} = \frac{2.1 - 1.7}{\sqrt{\dfrac{0.29^2}{64} + \dfrac{0.24^2}{75}}} = 8.77.$$

La zona di rifiuto è costituita dall'insieme

$$\{z_{\bar{x}_1-\bar{x}_2} : z_{\bar{x}_1-\bar{x}_2} > 2.326\}.$$

La statistica test cade nella zona di rifiuto, cosicché l'ipotesi nulla viene rifiutata. ■

21.3.1 Popolazioni Bernoulliane

L'inferenza sulla differenza tra le medie di due popolazioni Bernoulliane presenta aspetti peculiari che giustificano una trattazione specifica. Si fa riferimento ovviamente al caso di grandi campioni. Si ricorda che le medie e le varianze delle due popolazioni sono: $\mu_1 = p_1$, $\mu_2 = p_2$, $\sigma_1^2 = p_1(1 - p_1)$, $\sigma_2^2 = p_2(1 - p_2)$. Oggetto dell'inferenza è la differenza $\delta = p_1 - p_2$.

Stima intervallare

La (21.3) prende la forma

$$Z_{\hat{p}_1-\hat{p}_2} = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}}$$

e ha distribuzione prossima alla normale standard. Si ha pertanto

$$P\left[-z_{1-\alpha/2} < \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}} < z_{1-\alpha/2}\right] \approx 1 - \alpha,$$

da cui è facile ottenere, con il solito procedimento, gli estremi dell'intervallo di confidenza per $p_1 - p_2$:

$$l_1 = (\hat{p}_1 - \hat{p}_2) - z_{1-\alpha/2}\,\hat{\sigma}_{\hat{p}_1-\hat{p}_2},$$

$$l_2 = (\hat{p}_1 - \hat{p}_2) + z_{1-\alpha/2}\,\hat{\sigma}_{\hat{p}_1-\hat{p}_2},$$

dove

$$\hat{\sigma}_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}.$$

ESEMPIO 21.8

Nei sondaggi pre-elettorali effettuati in due città, si osservano i risultati di seguito indicati. Città A: in un campione di 1000 unità il 29% dichiara che voterà per il partito P. Città B: in un campione di 800 unità il 33% dichiara che voterà per il partito P. Si determini l'intervallo di confidenza al 99% per la differenza delle proporzioni delle intenzioni di voto per il partito P nelle due popolazioni.

Soluzione

Considerato che $z_{0.995} = 2.576$ e che

$$\hat{\sigma}_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{0.29 \cdot 0.71}{1000} + \frac{0.33 \cdot 0.67}{800}} = 0.022,$$

l'intervallo di confidenza cercato è delimitato da:

$$l_1 = (0.29 - 0.33) - 2.576 \cdot 0.022 = -0.10,$$
$$l_2 = (0.29 - 0.33) + 2.576 \cdot 0.022 = 0.02.$$

La differenza $p_1 - p_2$ è allora compresa nell'intervallo $(-0.10, 0.02)$ con una fiducia del 99%. ■

Verifica di ipotesi

L'ipotesi nulla da verificare è $H_0 : \delta = p_1 - p_2 = 0$, a fronte di una delle tre ipotesi alternative $H_1 : \delta = p_1 - p_2 > 0$, $H_1 : \delta = p_1 - p_2 < 0$ e $H_1 : \delta = p_1 - p_2 \neq 0$. Va osservato che l'ipotesi nulla ha impatto sulla varianza della variabile casuale $\hat{p}_1 - \hat{p}_2$. Infatti, essendo

$$\sigma^2_{\hat{p}_1-\hat{p}_2} = \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2},$$

se $p_1 = p_2 = p$,

$$\sigma^2_{\hat{p}_1-\hat{p}_2} = p(1 - p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right). \tag{21.4}$$

Se l'ipotesi nulla è vera, uno stimatore non distorto della (21.4) è dato da

$$\hat{\sigma}^2_{\hat{p}_1-\hat{p}_2} = \hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right),$$

dove

$$\hat{p}_c = \frac{\hat{p}_1 \cdot n_1 + \hat{p}_2 \cdot n_2}{n_1 + n_2}$$

è la proporzione dei successi calcolata aggregando i due campioni. La statistica test da usare è allora

$$Z_{\hat{p}_1-\hat{p}_2} = \frac{\hat{p}_1 - \hat{p}_2}{\hat{\sigma}_{\hat{p}_1-\hat{p}_2}}.$$

Poiché $Z_{\hat{p}_1-\hat{p}_2}$ ha distribuzione prossima alla normale standard, è facile stabilire che le zone di rifiuto del test sono quelle riportate nella Tabella 21.4.

abella 21.4
one di rifiuto della
atistica test
$_{\hat{p}_1-\hat{p}_2} = (\hat{p}_1 - \hat{p}_2)/\hat{\sigma}_{\hat{p}_1-\hat{p}_2}$
ella verifica dell'ipotesi
uguaglianza delle
roporzioni di due
opolazioni Bernoulliane.

Ipotesi H_0	Ipotesi H_1	Statistica test	Zona di rifiuto
	$\delta = p_1 - p_2 > 0$		$\{z_{\hat{p}_1-\hat{p}_2} > z_{1-\alpha}\}$
$p_1 - p_2 = 0$	$\delta = p_1 - p_2 < 0$	$Z_{\hat{p}_1-\hat{p}_2} = \frac{\hat{p}_1 - \hat{p}_2}{\hat{\sigma}_{\hat{p}_1-\hat{p}_2}}$	$\{z_{\hat{p}_1-\hat{p}_2} < -z_{1-\alpha}\}$
	$\delta = p_1 - p_2 \neq 0$		$\{\lvert z_{\hat{p}_1-\hat{p}_2}\rvert > z_{1-\alpha/2}\}$

ESEMPIO 21.9

Un partito politico ha commissionato a un ente di ricerca un'indagine sull'orientamento della popolazione in merito a un prossimo referendum. Al partito interessa sapere se l'opinione dei votanti è la stessa nelle regioni chiave A e B. I dati raccolti sono i seguenti: nella regione A, su 500 intervistati, 300 hanno dichiarato che voteranno sì. Nella regione B, su 600 intervistati, 340 hanno dichiarato che voteranno sì. Si ponga $\alpha = 0.05$.

Soluzione

L'ipotesi da verificare è H_0: $\delta = p_1 - p_2 = 0$, contro l'alternativa H_1: $\delta = p_1 - p_2 \neq 0$. La zona di rifiuto è data dall'insieme

$$\{z_{\hat{p}_1-\hat{p}_2} : |z_{\hat{p}_1-\hat{p}_2}| > 1.96\}.$$

Il valore osservato di $\hat{p}_c$ è

$$\hat{p}_c = \frac{300+340}{500+600} = 0.58,$$

da cui

$$z_{\hat{p}_1-\hat{p}_2} = \frac{300/500 - 340/600}{\sqrt{0.58 \cdot 0.42\left(\frac{1}{500}+\frac{1}{600}\right)}} = 1.12.$$

Il valore osservato della statistica test non appartiene alla zona di rifiuto. L'ipotesi nulla non viene quindi rifiutata. ■

21.4 Verifica dell'ipotesi di uguaglianza delle varianze di due popolazioni normali

Il confronto tra due popolazioni non è limitato alle sole medie; vi sono situazioni in cui vi è interesse a comparare le varianze. Si pensi al controllo degli standard di qualità dei manufatti di un processo industriale. Date certe caratteristiche (peso, lunghezza ecc.), è desiderabile che il manufatto rispecchi la media stabilita dalle cosiddette specifiche del processo di produzione; è anche desiderabile che la variabilità sia la più bassa possibile. È dunque importante confrontare la variabilità – misurata mediante la varianza – di una stessa caratteristica dei manufatti prodotti in differenti processi industriali.

La necessità di sottoporre a verifica l'ipotesi di uguaglianza di due varianze emerge anche nel caso del confronto tra le medie di due popolazioni normali mediante la statistica test $T_{\bar{X}_1-\bar{X}_2} = (\bar{X}_1 - \bar{X}_2)/(S_c\sqrt{1/n_1 + 1/n_2})$. Come si è detto in precedenza, l'uso di questa statistica presuppone che le varianze delle due popolazioni siano uguali, e tale assunzione necessita di essere verificata preliminarmente.

PROPOSIZIONE 21.3

Siano X_1 e X_2 due variabili casuali indipendenti aventi distribuzioni chi-quadrato con r_1 e r_2 gradi di libertà, rispettivamente. Allora la distribuzione di probabilità del rapporto

$$Y = \frac{X_1/r_1}{X_2/r_2}$$

è descritta dalla funzione di densità

$$f(y) = c\,y^{\frac{r_1}{2}-1}\left(1+\frac{r_1}{r_2}y\right)^{-\frac{r_1+r_2}{2}}, \quad 0 \leq y < \infty,$$

dove c è un numero positivo che dipende da r_1 e r_2.

Per la dimostrazione si rinvia a Hogg *et al.* (2013), pp. 191-193. ■

La distribuzione di probabilità di Y è chiamata distribuzione F (da R.A. Fisher). La figura che segue illustra la forma della funzione di densità della distribuzione F, che dipende dai parametri r_1 e r_2.

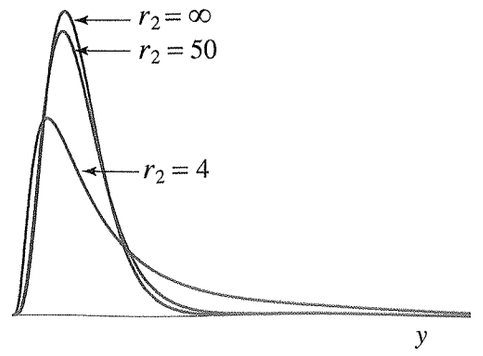


Figura 21.2
Funzione di densità della distribuzione F per $r_1 = 10$ e $r_2 = 4, 50, \infty$.

Pearson

Statistica
Principi e metodi
Quarta edizione

Statistica

Principi e metodi

Quarta edizione

Giuseppe Cicchitelli
Pierpaolo D'Urso
Marco Minozzo

Con la Figura 21.3, viene illustrato il significato del quantile di livello p della distribuzione F.

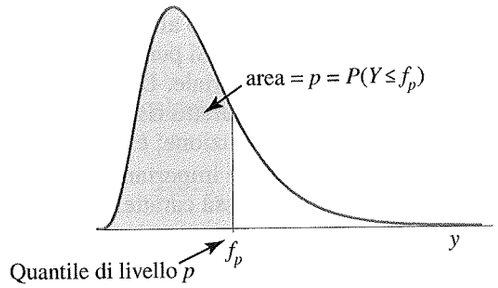


Figura 21.3
Quantile di livello p della distribuzione F.

I quantili superiori della distribuzione F per $p = 0.999, 0.99, 0.95$, sono riportati nella Tabella C.4 dell'Appendice C.

I quantili inferiori della distribuzione F con r_1 e r_2 gradi di libertà possono essere calcolati come reciproci dei corrispondenti quantili superiori della distribuzione F con r_2 e r_1 gradi di libertà. Per esempio, il quantile di livello 0.01 della distribuzione F con $r_1 = 9$ e $r_2 = 12$ è uguale a $1/5.11 = 0.20$, essendo 5.11 il quantile di livello 0.99 della distribuzione F con $r_1 = 12$ e $r_2 = 9$.

PROPOSIZIONE 21.4

Siano S_1^2 e S_2^2 le varianze dei campioni C_1 e C_2 indipendenti, provenienti dalle popolazioni $\mathcal{P}_1$ e $\mathcal{P}_2$ distribuite normalmente e aventi la stessa varianza. Allora la statistica

$$F = \frac{S_1^2}{S_2^2} \tag{21.5}$$

ha distribuzione F con $n_1 - 1$ e $n_2 - 1$ gradi di libertà.

La dimostrazione è lasciata per esercizio (Esercizio 21.15*). ■

In base alla Proposizione 21.4 e usando la (21.5) come statistica test, si perviene facilmente alle zone di rifiuto indicate nella Tabella 21.5 per la verifica dell'ipotesi H_0: $\sigma_1^2 = \sigma_2^2$ contro una delle alternative: $\sigma_1^2 > \sigma_2^2$, $\sigma_1^2 < \sigma_2^2$ e $\sigma_1^2 \neq \sigma_2^2$.

Tabella 21.5
Zone di rifiuto della statistica test $F = S_1^2/S_2^2$ per la verifica dell'ipotesi di uguaglianza delle varianze di due popolazioni normali.

Ipotesi H_0	Ipotesi H_1	Statistica test	Zona di rifiuto
	$\sigma_1^2 > \sigma_2^2$		$\{f > f_{1-\alpha}\}$
$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	$F = \frac{S_1^2}{S_2^2}$	$\{f < f_\alpha\}$
	$\sigma_1^2 \neq \sigma_2^2$		$\{f \notin (f_{\alpha/2},\ f_{1-\alpha/2})\}$

Nella tabella, f_α e $f_{1-\alpha}$ sono i quantili di livello α e $1 - \alpha$ della distribuzione F con $n_1 - 1$ e $n_2 - 1$ gradi di libertà. Significato analogo hanno $f_{\alpha/2}$ e $f_{1-\alpha/2}$.

ESEMPIO 21.10

Riprendendo i dati dell'Esempio 21.5, si sottoponga a verifica l'ipotesi nulla di uguaglianza delle varianze, contro l'ipotesi alternativa che la prima sia maggiore. Si ponga $\alpha = 0.05$.

Soluzione

La zona di rifiuto è

$$\{f: f > 2.31\}.$$

La statistica test assume il valore

$$f = \frac{5.3^2}{4.6^2} = 1.33,$$

che, non appartenendo alla zona di rifiuto, conduce a non rifiutare l'ipotesi nulla. ■

Riassunto

Questo capitolo è dedicato principalmente ai problemi di inferenza riguardanti le medie di due popolazioni. A tal fine, vengono impiegate le tecniche di inferenza introdotte, per una sola popolazione, nei Capitoli 19 e 20. Le situazioni specifiche considerate riguardano:

(a) due popolazioni normali con varianze note;
(b) due popolazioni normali con varianze incognite, ma uguali;
(c) due popolazioni aventi distribuzioni arbitrarie e campioni sufficientemente grandi;
(d) due popolazioni Bernoulliane e campioni sufficientemente grandi.

L'ultimo paragrafo è dedicato alla verifica dell'ipotesi di uguaglianza delle varianze di due popolazioni normali.

Parole chiave

problema di Beherens-Fisher, 466
varianza campionaria combinata, 463
distribuzione F, 471

Commenti e curiosità

Assenza di evidenza non equivale a evidenza dell'assenza

Questa espressione, che può sembrare uno scioglilingua, vuole segnalare che il non rifiuto dell'ipotesi nulla sulla differenza tra due medie (assenza di evidenza) non implica l'accettazione dell'ipotesi nulla di uguaglianza tra le stesse (assenza di differenza). La questione è particolarmente avvertita nelle applicazioni mediche dell'inferenza statistica. In questo settore, spesso gli esperimenti controllati mirano a comparare l'efficacia di due terapie concorrenti. L'ipotesi nulla afferma che le due terapie sono ugualmente efficaci (in termini di percentuali di guarigioni o di medie relative a misure legate al disturbo studiato). L'ipotesi alternativa afferma che la terapia in sperimentazione è più efficace. Ora, se il test non porta al rifiuto dell'ipotesi nulla, si può dire che i dati non contengono un'evidenza così forte contro l'ipotesi nulla da consentirne il rifiuto; ma ciò non implica che le medie delle due popolazioni siano uguali. Una differenza statisticamente non significativa può ritenersi importante dal punto di vista medico. In definitiva, la non significatività della differenza non chiude il problema in modo definitivo: nuovi esperimenti controllati potrebbero essere realizzati, magari con un numero maggiore di osservazioni per diminuire l'incertezza della decisione statistica.

Al riguardo, è opportuno menzionare un esempio tratto dalla letteratura medica (vedi Campbell e Machin, 1999). In un esperimento clinico su 1239 colpiti da infarto non fatale fu osservata una mortalità dell'8% in un gruppo trattato con aspirina e del 10.7% in un gruppo che aveva ricevuto il placebo. L'intervallo di confidenza al 95% per la differenza tra le percentuali di mortalità fu (−0.5, 6), indicando la non significatività statistica del differenziale di mortalità. In uno studio successivo, fatto su 6292 pazienti, venne riscontrata una mortalità del 9.2% tra i soggetti trattati con aspirina e dell'11.5% tra i soggetti curati con placebo. L'intervallo di confidenza al 95% per la differenza risultò però (0.8, 3.8), da cui si deduce la significatività statistica della differenza tra le percentuali. Questa digressione segnala implicitamente che una maggiore ampiezza dei campioni consente di meglio discriminare le popolazioni a confronto in virtù dell'accresciuta potenza del test.

Esercizi

21.1 Di seguito sono riportati i dati sulla durata delle batterie di due tipi di cellulare, osservati su due campioni casuali.

Cellulari di tipo A:	62,	72,	67,	68,	44,	57,	70,	69,	77,	70,	47,	73,	71,
	65,	80,	74,	84,	100,	80,	90,	83,	71,	61,	60,	70.	
Cellulari di tipo B:	78,	105,	72,	96,	79,	91,	79,	65,	82,	85,	82,	71,	62,
	94,	77,	85,	74,	83,	82,	86,	75,	54,	83,	102.		

Si determini l'intervallo di confidenza al 95% per la differenza tra la durata media delle batterie dei due tipi di cellulare, assumendo che le popolazioni siano normali con la stessa varianza.

21.2 Con riferimento all'esercizio precedente, si verifichi l'ipotesi nulla di uguaglianza tra la durata media delle batterie dei due tipi di cellulare, contro l'alternativa che la durata media delle batterie del cellulare di tipo A sia inferiore. Si ponga $\alpha = 0.01$.

21.3 È stato condotto uno studio per comprendere come la somministrazione di un ormone della crescita influisca sul peso dei ratti in stato di gravidanza, osservando l'aumento di peso durante la gestazione su 10 ratti trattati con l'ormone e su altrettanti ratti non trattati. I dati sono esposti nella tabella che segue:

	Ratti trattati	Ratti non trattati
Media	63.4	43.8
Deviazione standard campionaria	15.9	12.2

a) Quali sono le assunzioni necessarie per applicare la statistica $T_{\bar{X}_1-\bar{X}_2}$ per la verifica dell'ipotesi di uguaglianza tra le medie delle due popolazioni?

b) Si stabilisca se l'accrescimento medio del peso dei ratti trattati con l'ormone è significativamente superiore a quello dei ratti non trattati. Si usi il metodo del livello di significatività osservato e si ponga $\alpha = 0.05$.

21.4* Si ripeta l'esercizio precedente per il sistema di ipotesi:

$$H_0: \delta = \mu_1 - \mu_2 = 10, \quad H_1: \delta = \mu_1 - \mu_2 > 10.$$

(*Suggerimento*: la verifica dell'ipotesi nulla $H_0: \delta = \mu_1 - \mu_2 = \delta_0$ contro una delle alternative possibili richiede l'utilizzo della statistica test

$$(\bar{X}_1 - \bar{X}_2 - \delta_0)/\left(S_c\sqrt{1/n_1 + 1/n_2}\right),$$

che ha distribuzione T di Student con $n_1 + n_2 - 2$ gradi di libertà.)

21.5 In seguito a un'indagine sul contenuto di energia metabolizzabile (in kcal per grammo) di due diversi tipi di cibo per gatti, si sono registrati i risultati esposti nella tabella seguente.

	Cibo di tipo A	Cibo di tipo B
Dimensione del campione	28	29
Media	1.19	1.54
Deviazione standard campionaria	0.26	0.35

Si verifichi l'ipotesi che i due tipi di cibo abbiano lo stesso contenuto medio di energia metabolizzabile contro l'alternativa che il cibo B abbia un contenuto medio superiore. Si fissi il livello di significatività al 5% e si assuma che valgano le condizioni standard per l'applicazione della statistica test $T_{\bar{X}_1-\bar{X}_2}$.

21.6 I dati che seguono si riferiscono all'età dei partecipanti a un concorso interno bandito da un ente pubblico irlandese per il passaggio a una qualifica superiore.

Età dei non vincitori:	34,	37,	37,	38,	41,	42,	43,	44,	44,	45,	45,	45,	46,	48,	49,
	53,	53,	54,	54,	55,	56,	57,	60.							
Età dei vincitori:	27,	33,	36,	37,	38,	38,	39,	42,	42,	43,	43,	44,	44,	44,	45,
	45,	45,	45,	46,	46,	47,	47,	48,	48,	49,	49,	51,	51,	52,	54.

Alcuni dei candidati non vincitori fecero ricorso sostenendo che i risultati del concorso erano inficiati da una discriminazione verso i candidati meno giovani. Considerando i due insiemi di dati come campioni casuali provenienti da popolazioni normali, si stabilisca se la differenza tra l'età media nei due campioni è significativamente maggiore di 0. Si usi la statistica test che si ottiene dalla quantità $T'_{\bar{X}_1-\bar{X}_2}$ introdotta nel Paragrafo 21.2.1, con un livello di significatività del 5%. (*Suggerimento*: si ricorra alla funzione INV.T.2T( ) di Excel qualora i gradi di libertà richiesti non siano presenti nella Tavola C.4 dell'Appendice C.)

21.7 In un esperimento sugli effetti dell'alcool furono rilevati gli errori in prove di lettura e di capacità motoria su 21 soggetti a cui era stato fatto assumere alcool e su altrettanti soggetti a cui era stato fatto assumere un placebo. I punteggi medi e le varianze nei due campioni sono presentate nella tabella che segue:

	Gruppo trattato con alcool	Gruppo trattato con placebo
Dimensione campionaria	21	21
Media	4.20	1.71
Varianza	4.84	0.52

I dati sorreggono l'opinione corrente che assumere alcool è pericoloso per chi guida? Per rispondere alla domanda si indichi il sistema di ipotesi, si esplicitino le assunzioni necessarie per l'utilizzo della quantità $T'_{X_1-X_2}$ e si ponga $\alpha = 0.01$.

21.8 Per studiare l'effetto del consumo di pesce sul rischio di malattie cardiache, furono osservati due campioni di individui per un congruo periodo di tempo. Il primo campione era costituito da non consumatori di pesce e il secondo da consumatori di almeno 45 grammi di pesce al giorno. Nella tabella che segue sono riportati, oltre alle dimensioni dei campioni, le medie e le varianze campionarie del livello di colesterolo rilevato al termine del periodo di osservazione:

	Non consumatori di pesce	Consumatori di pesce (>45 grammi al giorno)
Dimensione campionaria	159	79
Livello medio di colesterolo	158	146
Varianza campionaria	66	75

Si determini l'intervallo di confidenza al 99% per la differenza tra i livelli medi di colesterolo nei due gruppi.

21.9 Nell'esercizio precedente, lo studio aveva come obiettivo quello di dimostrare che una dieta a base di pesce abbassa i livelli di colesterolo e quindi il rischio di malattie cardiache. Specificando l'appropriato sistema di ipotesi, si stabilisca se le medie campionarie sono significativamente diverse a un livello di significatività del 5%.

21.10 Per decidere se acquistare un nuovo lettore ottico per il controllo delle scorte, un grande magazzino ha condotto delle prove sul nuovo lettore e su quello in uso. I seguenti dati si riferiscono al numero di codici a barre letti al secondo:

	Nuovo lettore	Vecchio lettore
Dimensione campionaria	54	60
Media	39	31
Varianza campionaria	29.5	21.8

Si determini l'intervallo di confidenza al 99% per la differenza $\mu_1 - \mu_2$.

21.11 Un ricercatore intende confrontare l'efficacia di due trattamenti per una data malattia: il trattamento A, attualmente in uso, e il trattamento B reputato superiore in base ad alcuni studi preliminari. A questo fine, osserva 174 individui. A 85 soggetti scelti a caso somministra il trattamento A, mentre ai restanti soggetti somministra il trattamento B. La media e la deviazione standard della risposta ai trattamenti sono riportate nella tabella che segue:

	Trattamento A	Trattamento B
Media	104	127
Deviazione standard campionaria	37.6	47.2

Si specifichi l'appropriato sistema di ipotesi e si verifichi se le medie campionarie sono significativamente diverse. Si utilizzi il metodo del livello di significatività osservato, ponendo $\alpha = 0.01$.

21.12 In uno studio sulla diffusione del lavoro a tempo parziale tra gli studenti, il Direttore di un dipartimento universitario commissiona un'indagine campionaria. Di seguito sono riportati i risultati distintamente per sesso:

	Maschi	Femmine
Lavoratori	235	127
Non lavoratori	1150	683

Si determini l'intervallo di confidenza al 95% per la differenza delle proporzioni dei lavoratori a tempo parziale nelle due popolazioni.

21.13 Con riferimento all'esercizio precedente, si verifichi l'ipotesi nulla $H_0: p_1 = p_2$, contro l'alternativa $H_1: p_1 \neq p_2$, dove p_1 e p_2 sono le proporzioni (calcolate nelle popolazioni) di lavoratori a tempo parziale tra i maschi e le femmine, rispettivamente. Si ponga $\alpha = 0.01$.

21.14 Si considerino due campioni casuali indipendenti, il primo di 350 persone della classe di età 25-40, il secondo di altrettante persone della classe di età 55-70. Per ogni individuo, viene rilevato il numero di ore dormite per notte, ottenendo quanto riportato nella tabella che segue:

Età	Ore di sonno ≤ 7	Ore di sonno >7	Totale
25-40	156	194	350
55-70	183	167	350
Totale	339	361	700

Vi è evidenza sufficiente per affermare che la proporzione delle persone che dormono non più di 7 ore per notte è maggiore per la classe di età 55-70 che non per la classe 25-40? Si ponga $\alpha = 0.05$.

21.15* Si considerino due campioni casuali indipendenti di ampiezza n_1 e n_2, estratti da due popolazioni normali aventi la stessa varianza. Si dimostri che la variabile casuale

$$F = \frac{S_1}{S_2}$$

ha distribuzione F con $n_1 - 1$ e $n_2 - 1$ gradi di libertà. (*Suggerimento*: si ricordi che, se i campioni provengono da distribuzioni normali con varianze σ_1^2 e σ_2^2, $(n_1 - 1)S_1^2/\sigma_1^2$ e $(n_2 - 1)S_2^2/\sigma_2^2$ sono variabili casuali chi-quadrato con $n_1 - 1$ e $n_2 - 1$ gradi di libertà, rispettivamente.)

21.16 Il responsabile del controllo di processo di un'industria alimentare intende confrontare l'efficienza di due confezionatrici relativamente alle confezioni di pasta da 500 grammi, per quanto riguarda la variabilità dei pesi. Il dubbio è che la confezionatrice A produca confezioni di peso con maggiore variabilità. In un campione casuale di 31 confezioni provenienti dalla confezionatrice A, egli osserva una deviazione standard dei pesi pari a 7.08; in un campione casuale di 28 confezioni provenienti dalla confezionatrice B, osserva una deviazione standard uguale a 4.31. Si indichino le assunzioni necessarie sulle popolazioni e si verifichi se vi sia sufficiente evidenza per rifiutare l'ipotesi che i pesi delle confezioni prodotte dalle due confezionatrici hanno la stessa variabilità. Si ponga $\alpha = 0.01$.

21.17 Con riferimento all'Esercizio 21.7, si verifichi l'ipotesi dell'uguaglianza delle varianze delle due popolazioni contro l'alternativa che la popolazione associata al gruppo dei trattati con alcool abbia maggiore variabilità, assumendo un livello di significatività di 0.05. Sulla base di questo risultato, si chiarisca come, nell'esercizio in questione, sia stato appropriato l'uso della quantità $T'_{\bar{X}_1 - \bar{X}_2}$ per la verifica dell'ipotesi considerata.

21.18 Con riferimento all'Esercizio 21.9, si calcoli la probabilità dell'errore di II tipo per la verifica dell'ipotesi nulla $H_0: \delta = \mu_1 - \mu_2 = 0$ contro l'alternativa $H_1: \delta = \mu_1 - \mu_2 = 5$, ponendo α a 0.01 e assumendo che le varianze delle due popolazioni siano: $\sigma_1^2 = 69$ e $\sigma_2^2 = 71$ (si tenga presente che μ_1 indica il livello medio di colesterolo dei non consumatori di pesce).

CAPITOLO

22

Verifica di ipotesi per dati di frequenza

Sommario del capitolo

22.1 Introduzione

22.2 La statistica test chi-quadrato

Riassunto

Commenti e curiosità

Esercizi

22.1 Introduzione

In questo capitolo, si affronteranno alcuni problemi di verifica di ipotesi in cui i dati sono costituiti da conteggi o frequenze derivanti dalla classificazione delle unità campionarie secondo le modalità di una o due variabili.

Nei capitoli precedenti, è capitato di parlare di popolazione Bernoulliana per indicare la popolazione in cui le unità sono classificate secondo due sole modalità, successo e insuccesso. È ora opportuno introdurre la nozione di popolazione multinomiale. Si tratta di un'estensione della popolazione Bernoulliana: le unità che la compongono sono classificate secondo k modalità o k classi della variabile oggetto di studio, con $k \geq 3$, come indicato nella tabella che segue:

Tabella 22.1 Popolazione multinomiale.

Modalità	A_1	A_2	...	A_k	Totale
Probabilità	p_1	p_2	...	p_k	1

dove $A_1, A_2, \ldots, A_k$ sono le modalità o le classi della variabile considerata e $p_1, p_2, \ldots, p_k$ le rispettive frequenze relative, interpretabili come probabilità se si immagina di estrarre a sorte una unità da tale popolazione. In termini più generali, si può parlare di popolazione multinomiale pensando a un esperimento casuale i cui esiti possibili, $A_1, A_2, \ldots, A_k$, hanno probabilità $p_1, p_2, \ldots, p_k$.

Si immagini ora di ripetere n volte l'esperimento, con prove indipendenti e tali che restino invariate le probabilità $p_1, p_2, \ldots, p_k$ da prova a prova. Il campione casuale che così si ottiene può essere organizzato nella forma di distribuzione di frequenza:

Tabella 22.2 Campione casuale da una popolazione multinomiale.

Modalità	A_1	A_2	...	A_k	Totale
Frequenza	n_1	n_2	...	n_k	n

dove n_1 è la frequenza di A_1, n_2 la frequenza di A_2 e così via. Chiaramente, tali frequenze assumono la veste di variabili casuali.

ESEMPIO 22.1

Per verificare l'ipotesi nulla che i quattro ingressi di un supermercato siano utilizzati con la stessa frequenza, vengono osservati gli ingressi usati da un campione di 116 clienti. La distribuzione di frequenza risultante è riportata nella Tabella 22.3:

Tabella 22.3
Frequenza d'uso dei quattro ingressi del supermercato.

Ingresso	A_1	A_2	A_3	A_4	Totale
Frequenza d'uso	28	37	31	20	116

dove A_1, A_2, A_3 e A_4 sono gli ingressi del supermercato. L'ipotesi nulla equivale a dire che il campione proviene dalla seguente popolazione multinomiale:

Tabella 22.4
Popolazione di provenienza del campione di cui alla Tabella 22.3 sotto l'ipotesi nulla.

Ingresso	A_1	A_2	A_3	A_4	Totale
Probabilità	0.25	0.25	0.25	0.25	1

■

ESEMPIO 22.2

Si consideri la distribuzione di frequenza riportata nella tabella che segue:

Tabella 22.5
Campione casuale di 400 individui per classi di altezza.

Classi di altezza (in cm)	< 165	165-170	170-175	175-180	> 180	Totale
Frequenza	7	51	190	124	28	400

Se l'ipotesi nulla è che il campione casuale provenga da una popolazione normale con media $\mu = 174$ e varianza $\sigma^2 = 16$, la distribuzione di frequenza nella Tabella 22.5 può essere pensata come un campione casuale di dimensione 400 dalla seguente popolazione multinomiale:

Tabella 22.6
Popolazione di provenienza del campione di cui alla Tabella 22.5 sotto l'ipotesi nulla.

Classi altezza (in cm)	< 165	165-170	170-175	175-180	> 180	Totale
Probabilità	0.012	0.147	0.440	0.334	0.067	1

dove le probabilità sono state calcolate usando la distribuzione normale con media $\mu = 174$ e varianza $\sigma^2 = 16$. La probabilità della prima classe è così data da

$$\Phi\left(\frac{165 - 174}{4}\right) = 0.012,$$

dove $\Phi(\cdot)$ è la funzione di ripartizione della normale standard. La probabilità della seconda classe è data da

$$\Phi\left(\frac{170 - 174}{4}\right) - \Phi\left(\frac{165 - 174}{4}\right) = 0.147$$

e così via. ■

Nel prossimo paragrafo, verrà introdotta una procedura di base per la verifica dell'ipotesi che un campione casuale provenga da una data popolazione multinomiale. Il metodo verrà poi generalizzato per coprire situazioni più complesse in cui le probabilità $p_1, p_2, \ldots, p_k$ sono funzione di parametri stimati con i dati del campione.

I risultati enunciati nelle Proposizioni 22.2 e 22.3 sono dati senza dimostrazioni. Per queste si rinvia a Rao (1973), p. 391 e segg.

22.2 La statistica test chi-quadrato

L'ipotesi nulla sui parametri di una popolazione multinomiale consiste nell'assegnare valori specifici a $p_1, p_2, \ldots, p_k$, come illustrato negli Esempi 22.1 e 22.2. In simboli, l'ipotesi nulla è H_0: $p_i = p_{i0},\ i = 1, 2, \ldots, k$, in cui $p_{10}, p_{20}, \ldots, p_{k0}$ sono valori assegnati.

Si riprendano le Tabelle 22.1 e 22.2. Si consideri che, in una prova dell'esperimento casuale, un risultato specifico, ad esempio A_1, si verifica con probabilità p_1, e l'evento complementare con probabilità $1 - p_1$. In n prove indipendenti, in cui rimane ferma la probabilità p_1, il numero di unità che presentano la modalità A_1, n_1, è una variabile casuale binomiale con parametri n e p_1. Ne segue che il valore atteso di n_1 è np_1. Con lo stesso ragionamento, si trova che: il valore atteso di n_2 è $np_2, \ldots$, il valore atteso di n_k è np_k.

PROPOSIZIONE 22.1

Sotto l'ipotesi nulla

$$H_0: p_1 = p_{10},\ p_2 = p_{20}, \ldots, p_k = p_{k0} \tag{22.1}$$

i valori attesi di $n_1, n_2, \ldots, n_k$ – frequenze delle modalità $A_1, A_2, \ldots, A_k$, in un campione casuale di dimensione n proveniente da una popolazione multinomiale – sono dati, rispettivamente, da $np_{10}, np_{20}, \ldots, np_{k0}$.

Associando alle modalità $A_1, A_2, \ldots, A_k$, rispettivamente, i valori attesi delle frequenze $n_1, n_2, \ldots, n_k$ della Tabella 22.2 (intese come variabili casuali) sotto l'ipotesi nulla (22.1), si ottiene la distribuzione di frequenza attesa riportata nella tabella che segue:

Tabella 22.7
Distribuzione di frequenza attesa sotto l'ipotesi nulla (22.1).

Modalità	A_1	A_2	...	A_k	Totale
Frequenza attesa	np_{10}	np_{20}	...	np_{k0}	n

Un criterio ragionevole per verificare l'ipotesi nulla (22.1) consiste nel misurare la distanza tra le frequenze $n_1, n_2, \ldots, n_k$ (Tabella 22.2) e quelle attese sotto l'ipotesi nulla (Tabella 22.7). Una conveniente misura di distanza è data dalla statistica test di Pearson, nota come statistica test chi-quadrato:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}}. \tag{22.2}$$

PROPOSIZIONE 22.2

Sotto l'ipotesi nulla (22.1), la statistica test chi-quadrato ha, approssimativamente, distribuzione chi-quadrato con $k - 1$ gradi di libertà, purché n sia sufficientemente grande.

Dunque, se n è sufficientemente grande, si può usare la distribuzione chi-quadrato per determinare la zona di rifiuto della statistica test (22.2). L'approssimazione è considerata soddisfacente se tutte le frequenze attese nella Tabella 22.7 non sono inferiori a 5.

Poiché la (22.2) è una misura della distanza tra i dati del campione e i dati attesi sotto l'ipotesi nulla, questa verrà rifiutata per valori molto alti della statistica test, ossia per valori maggiori di $\chi^2_{1-\alpha}$, essendo $\chi^2_{1-\alpha}$ il quantile di livello $1-\alpha$ della distribuzione chi-quadrato con $k-1$ gradi di libertà, cioè il valore tale che

$$P(\chi^2 \leq \chi^2_{1-\alpha}\,;\, H_0) = 1-\alpha.$$

In definitiva, la zona di rifiuto è

$$\{\chi^2 : \chi^2 > \chi^2_{1-\alpha}\}.$$

La procedura appena illustrata è nota come test della bontà dell'adattamento.

ESEMPIO 22.3

Si riprenda l'Esempio 22.1 e si verifichi l'ipotesi nulla ivi indicata a un livello di significatività di $\alpha = 0.05$.

Soluzione

La distribuzione di frequenza attesa è esposta nella tabella che segue:

Tabella 22.8
Distribuzione di frequenza attesa per i dati dell'Esempio 22.1.

Ingresso	A_1	A_2	A_3	A_4	Totale
Frequenza d'uso attesa	116/4	116/4	116/4	116/4	116

La statistica test chi-quadrato assume il valore

$$\chi^2 = \frac{(28-116/4)^2}{116/4} + \frac{(37-116/4)^2}{116/4} + \frac{(31-116/4)^2}{116/4} + \frac{(20-116/4)^2}{116/4} = 5.17.$$

I gradi di libertà sono 3 e l'appropriato quantile della distribuzione chi-quadrato è $\chi^2_{0.95} = 7.815$. Poiché $\chi^2 = 5.17 < 7.815 = \chi^2_{0.95}$, non vi è evidenza sufficiente per il rifiuto dell'ipotesi nulla. ■

ESEMPIO 22.4

Si riprenda l'Esempio 22.2 e si verifichi l'ipotesi che la popolazione di provenienza del campione sia normale di media 174 e varianza 16, con $\alpha = 0.01$.

Soluzione

Nella Tabella 22.9 sono riportate le frequenze attese ottenute moltiplicando le probabilità della Tabella 22.6 per 400:

Tabella 22.9
Distribuzione di frequenza attesa per i dati della Tabella 22.5.

Classi di altezza (in cm)	< 165	165-170	170-175	175-180	> 180	Totale
Frequenza attesa	4.8	58.8	176.0	133.6	26.8	400.0

La statistica test chi-quadrato assume il valore

$$\chi^2 = \frac{(7-4.8)^2}{4.8} + \frac{(51-58.8)^2}{58.8} + \cdots + \frac{(28-26.8)^2}{26.8} = 3.9.$$

Poiché la soglia critica è $\chi^2_{0.99} = 13.277$ (i gradi di libertà sono 4), l'ipotesi nulla non viene rifiutata. ■

Quando le probabilità $p_{10}, p_{20}, \ldots, p_{k0}$ dipendono da uno o più parametri incogniti stimati con i dati del campione, la procedura di verifica di ipotesi introdotta sopra va modificata come indicato nella proposizione che segue.

PROPOSIZIONE 22.3

Si supponga che le probabilità $p_{10}, p_{20}, \ldots, p_{k0}$ previste dall'ipotesi nulla (22.1), dipendano da q parametri incogniti, $\theta_1, \theta_2, \ldots, \theta_q$. Siano $\hat{p}_{10}, \hat{p}_{20}, \ldots, \hat{p}_{k0}$ le stime di queste probabilità calcolate usando le stime di $\theta_1, \theta_2, \ldots, \theta_q$ ottenute con i dati del campione. Se l'ipotesi nulla è vera e la dimensione del campione è sufficientemente grande, si può dimostrare che la statistica test

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n\hat{p}_{i0})^2}{n\hat{p}_{i0}} \tag{22.3}$$

si distribuisce approssimativamente secondo una chi-quadrato con $k-q-1$ gradi di libertà.

ESEMPIO 22.5

Da una ricerca sugli incidenti stradali, condotta per 65 settimane in un dato tratto di autostrada, si è ottenuta la distribuzione di frequenza esposta nella tabella che segue:

Tabella 22.10
Occorrenza di incidenti stradali in un tratto di autostrada.

N. di incidenti	0	1	2	≥ 3	Totale
Frequenza	48	15	2	0	65

Si verifichi l'ipotesi nulla che la popolazione di provenienza del campione segua una distribuzione di Poisson, a un livello di significatività $\alpha = 0.01$.

Soluzione

Sotto l'ipotesi nulla, le probabilità $p_{10}, p_{20}, \ldots, p_{k0}$ sono date da $\lambda^x e^{-\lambda}/x!$ $(x = 0, 1, \ldots)$ e dipendono dal parametro incognito λ. La stima di massima verosimiglianza di λ è la media campionaria: $(0\cdot 48 + 1\cdot 15 + 2\cdot 2)/65 = 0.29$. Sostituendo λ con 0.29 nel modello di Poisson, le probabilità $p_{10}, p_{20}, \ldots$ sono stimate come segue: $\hat{p}_{10} = e^{-0.29} = 0.748$, per $x = 0$; $\hat{p}_{20} = 0.29\, e^{-0.29} = 0.217$, per $x = 1$ ecc. La Tabella 22.11 riporta le frequenze attese, ottenute moltiplicando per 65 ciascuna delle probabilità appena stimate (le modalità 2 e ≥ 3 sono state aggregate per migliorare l'approssimazione del chi quadrato):

Tabella 22.11
Distribuzione di frequenza attesa per i dati della Tabella 22.10.

N. di incidenti	0	1	≥ 2	Totale
Frequenza attesa	48.620	14.105	2.275	65

La statistica test (22.3) assume il valore

$$\chi^2 = \frac{(48-48.62)^2}{48.62} + \frac{(15-14.105)^2}{14.105} + \frac{(2-2.275)^2}{2.275} = 0.098.$$

Dato che $\chi^2 = 0.098 < 6.63 = \chi^2_{0.99}$ (i gradi di libertà sono dati da $k-q-1 = 3-1-1 = 1$), l'ipotesi nulla non viene rifiutata. In altre parole, non vi è evidenza sufficiente per escludere che il modello di Poisson sia idoneo a descrivere il fenomeno oggetto di studio. ■

22.2.1 Verifica dell'ipotesi nulla di indipendenza

In questo paragrafo, viene affrontato il problema della verifica dell'ipotesi di indipendenza tra due variabili con la statistica test chi-quadrato. Per fare ciò, si consideri la popolazione multinomiale doppia contenuta nella Tabella 22.12:

Tabella 22.12
Popolazione multinomiale doppia.

A	B						Totale
	B_1	B_2	...	B_j	...	B_t	
A_1	p_{11}	p_{12}	...	p_{1j}	...	p_{1t}	p_{10}
A_2	p_{21}	p_{22}	...	p_{2j}	...	p_{2t}	p_{20}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_i	p_{i1}	p_{i2}	...	p_{ij}	...	p_{it}	p_{i0}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_s	p_{s1}	p_{s2}	...	p_{sj}	...	p_{st}	p_{s0}
Totale	p_{01}	p_{02}	...	p_{0j}	...	p_{0t}	1

dove: $A_1, A_2, \ldots, A_s$ e $B_1, B_2, \ldots, B_t$ sono, rispettivamente, le modalità delle variabili A e B; p_{ij} è la probabilità che, in una prova dell'esperimento casuale, si verifichi l'evento $A_i \cap B_j$; p_{i0} e p_{0j}, date da

$$p_{i0} = p_{i1} + p_{i2} + \cdots + p_{it} \quad \text{e} \quad p_{0j} = p_{1j} + p_{2j} + \cdots + p_{sj},$$

sono le probabilità di A_i e di B_j.

Se le variabili A e B sono indipendenti, si ha

$$p_{ij} = P(A_i \cap B_j) = P(A_i)P(B_j) = p_{i0}p_{0j}, \quad i = 1, 2, \ldots, s,\ j = 1, 2, \ldots, t. \tag{22.4}$$

Analogamente a quanto visto nel Paragrafo 22.1 per la popolazione multinomiale semplice, se si immagina di ripetere l'esperimento n volte, con prove indipendenti e tali che le probabilità p_{ij} non mutino da una prova all'altra, si ottiene un campione casuale di dimensione n. Indicando con n_{ij} la variabile casuale frequenza dell'evento $A_i \cap B_j$, il campione può essere organizzato sotto forma di distribuzione doppia di frequenza, come indicato nella tabella che segue:

Tabella 22.13
Campione casuale di dimensione n da una popolazione multinomiale doppia.

A	B						Totale
	B_1	B_2	...	B_j	...	B_t	
A_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1t}	n_{10}
A_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2t}	n_{20}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{it}	n_{i0}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_s	n_{s1}	n_{s2}	...	n_{sj}	...	n_{st}	n_{s0}
Totale	n_{01}	n_{02}	...	n_{0j}	...	n_{0t}	n

dove n_{i0} e n_{0j} sono le frequenze marginali di A_i e di B_j:

$$n_{i0} = n_{i1} + n_{i2} + \cdots + n_{it}, \quad n_{0j} = n_{1j} + n_{2j} + \cdots + n_{sj}.$$

ESEMPIO 22.6

Si consideri la seguente distribuzione doppia di frequenza di un campione casuale di forze di lavoro.

Tabella 22.14
Forze di lavoro per sesso e stato occupazionale.

Sesso	Stato occupazionale		Totale
	Occupati	In cerca di occupazione	
Femmina	69	11	80
Maschio	141	9	150
Totale	210	20	230

Si voglia verificare l'ipotesi nulla che vi sia indipendenza tra sesso e condizione occupazionale. ■

In base alla (22.4), l'ipotesi nulla di indipendenza tra A e B è definita da

$$H_0: p_{ij} = p_{i0}p_{0j},\ i = 1, 2, \ldots, s,\ j = 1, 2, \ldots, t.$$

Analogamente al caso di una sola variabile, il valore atteso di n_{ij}, sotto l'ipotesi nulla, è $np_{i0}p_{0j}$.

Per verificare l'ipotesi nulla, si deve adeguare la statistica test (22.3) alla nuova situazione. A tal fine, si devono confrontare le frequenze n_{ij} della Tabella 22.13 con quelle attese sotto l'ipotesi nulla, date da $np_{i0}p_{0j}$, $i = 1, 2, \ldots, s,\ j = 1, 2, \ldots, t$. Poiché le probabilità p_{i0} e p_{0j} non sono note, esse vanno stimate utilizzando i dati del campione. Si può dimostrare che le stime di massima verosimiglianza di p_{i0} e p_{0j} sono le frequenze relative di A_i e B_j nel campione:

$$\hat{p}_{i0} = \frac{n_{i0}}{n}, \quad \hat{p}_{0j} = \frac{n_{0j}}{n}.$$

Ne segue che la stima del valore atteso di n_{ij}, sotto l'ipotesi nulla, è

$$n\hat{p}_{i0}\hat{p}_{0j} = \frac{n_{i0}n_{0j}}{n}.$$

Applicando la (22.3) alla situazione in esame, si perviene alla statistica test

$$\chi^2 = \sum_{i=1}^{s}\sum_{j=1}^{t} \frac{[n_{ij} - (n_{i0}n_{0j})/n]^2}{(n_{i0}n_{0j})/n}, \tag{22.5}$$

che, sotto l'ipotesi nulla, ha, approssimativamente, distribuzione chi-quadrato con $(s-1)(t-1)$ gradi di libertà, in forza della Proposizione 22.3, a patto che la dimensione del campione sia sufficientemente grande.[1]

[1] I gradi di libertà si ottengono dalla formula data nella Proposizione 22.3, come segue. Il numero dei possibili esiti dell'esperimento è uguale al numero di celle della Tabella 22.13, che è $s \cdot t$. Il numero di parametri stimati è $s - 1 + t - 1$. Infatti, sono da stimare gli $s - 1$ parametri p_{i0} e i $t - 1$ parametri p_{0j}. Pertanto, i gradi di libertà sono

$$k - q - 1 = s \cdot t - 1 - (s-1) - (t-1) = (s-1)(t-1).$$

La (22.5) è una misura della distanza tra le frequenze osservate, n_{ij} (Tabella 22.13), e le corrispondenti frequenze attese (stimate) sotto l'ipotesi nulla, $(n_{i0}n_{0j})/n$. L'ipotesi nulla viene quindi rifiutata quando la statistica test assume un valore molto elevato, in base alla logica che una grande distanza tra le frequenze osservate e quelle attese è molto improbabile che sia effetto del caso. Dunque, una volta fissato il livello di significatività α, la zona di rifiuto del test è

$$\{\chi^2 : \chi^2 > \chi^2_{1-\alpha}\},$$

dove $\chi^2_{1-\alpha}$ è il quantile di livello $1-\alpha$ della distribuzione chi-quadrato con $(s-1)(t-1)$ gradi di libertà.

ESEMPIO 22.7

Si riprenda l'Esempio 22.6 e si verifichi l'ipotesi di indipendenza al livello di significatività $\alpha = 0.05$.

Soluzione

L'applicazione della (22.5) dà

$$\chi^2 = \frac{[69 - (210 \cdot 80)/230]^2}{(210 \cdot 80)/230} + \frac{[11 - (20 \cdot 80)/230]^2}{(20 \cdot 80)/230} + \frac{[141 - (210 \cdot 150)/230]^2}{(210 \cdot 150)/230} + \frac{[9 - (20 \cdot 150)/230]^2}{(20 \cdot 150)/230} = 3.947.$$

L'appropriato quantile della distribuzione chi-quadrato è $\chi^2_{0.95} = 3.84$ (c'è un solo grado di libertà). Si può allora concludere che l'ipotesi di indipendenza tra sesso e condizione professionale va rifiutata. ■

22.2.2 Verifica dell'ipotesi di omogeneità

Si illustra ora un problema di inferenza concettualmente diverso da quello affrontato nel paragrafo precedente, ma identico dal punto di vista del procedimento di verifica delle ipotesi.

ESEMPIO 22.8

Un'associazione di consumatori, allo scopo di esaminare le qualità di tre diverse marche di batterie per automobili, osserva la durata di 70 batterie della marca A, di 85 della marca B e di 63 della marca C. I risultati dell'indagine sono riportati nella Tabella 22.15.

Tabella 22.15
Distribuzione di frequenza della durata osservata in campioni casuali di tre marche di batteria.

Marca	Durata			Totale
	Bassa	Media	Alta	
A	12	35	23	70
B	11	44	30	85
C	9	28	26	63
Totale	32	107	79	218

L'obiettivo dell'analisi è quello di stabilire se i tre campioni osservati provengano da popolazioni omogenee, cioè se la probabilità di osservare una batteria con una certa durata (bassa, media, alta) è la stessa per le tre marche. ■

Si considerino s popolazioni multinomiali con le stesse modalità, $A_1, A_2, \ldots, A_t$, disposte in una tabella a doppia entrata, come segue:

Tabella 22.16
Un insieme di s popolazioni multinomiali con le stesse modalità.

Popolazione	Modalità						Totale
	A_1	A_2	...	A_j	...	A_t	
1	p_{11}	p_{12}	...	p_{1j}	...	p_{1t}	1
2	p_{21}	p_{22}	...	p_{2j}	...	p_{2t}	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	p_{i1}	p_{i2}	...	p_{ij}	...	p_{it}	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
s	p_{s1}	p_{s2}	...	p_{sj}	...	p_{st}	1

In sostanza, la Tabella 22.16 riporta s popolazioni multinomiali le cui unità sono classificate in base alle modalità $A_1, A_2, \ldots, A_t$ della stessa variabile.

Si è interessati a verificare l'ipotesi nulla che le s popolazioni abbiano la stessa distribuzione di probabilità. Ciò equivale a dire che, in una prova dell'esperimento casuale associato a una qualsiasi delle s popolazioni, la probabilità che si verifichi l'evento A_j è sempre la stessa, pari a p_j. Formalmente, l'ipotesi nulla è espressa da

$$H_0 : p_{i1} = p_1,\ p_{i2} = p_2, \ldots, p_{it} = p_t,\quad i = 1, 2, \ldots, s.$$

Per verificare l'ipotesi è necessario disporre di un campione casuale per ogni popolazione. I dati dei campioni possono essere sistemati in una tabella a doppia entrata:

Tabella 22.17
Un insieme di s campioni casuali provenienti dalle popolazioni della Tabella 22.16.

Campione	Modalità						Totale
	A_1	A_2	...	A_j	...	A_t	
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1t}	n_{10}
2	n_{21}	n_{22}	...	n_{2j}	...	n_{2t}	n_{20}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{it}	n_{i0}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
s	n_{s1}	n_{s2}	...	n_{sj}	...	n_{st}	n_{s0}
Totale	n_{01}	n_{02}	...	n_{0j}	...	n_{0t}	n

Per il campione della popolazione i-esima, i valori attesi di $n_{i1}, n_{i2}, \ldots, n_{it}$, sotto l'ipotesi nulla, sono:

$$n_{i0}p_1, n_{i0}p_2, \ldots, n_{i0}p_t,$$

dove le probabilità $p_1, p_2, \ldots, p_t$ non sono note e vengono stimate con dati del campione. Se H_0 è vera, le stime di massima verosimiglianza di queste probabilità

sono:

$$\frac{n_{01}}{n}, \frac{n_{02}}{n}, \ldots, \frac{n_{0t}}{n}.$$

Ne segue che le frequenze attese stimate delle modalità $A_1, A_2, \ldots, A_t$ nel campione proveniente dalla popolazione i-esima sono:

$$\frac{n_{i0}n_{01}}{n}, \frac{n_{i0}n_{02}}{n}, \ldots, \frac{n_{i0}n_{0t}}{n}.$$

Applicando la solita formula della statistica test chi-quadrato, si ottiene

$$\chi^2 = \sum_{i=1}^{s}\sum_{j=1}^{t} \frac{[n_{ij} - (n_{i0}n_{0j})/n]^2}{(n_{i0}n_{0j})/n}.$$

La statistica è formalmente identica alla (22.5) introdotta per la verifica dell'ipotesi di indipendenza.

Sotto l'ipotesi nulla, la statistica test segue, approssimativamente, una distribuzione chi-quadrato con $(s-1)(t-1)$ gradi di libertà.[2] Come al solito, la zona di rifiuto è

$$\{\chi^2 : \chi^2 > \chi^2_{1-\alpha}\}.$$

ESEMPIO 22.9

Si riprenda l'Esempio 22.8 e si verifichi l'ipotesi di omogeneità sulla base dei campioni riportati nella Tabella 22.15, al livello di significatività $\alpha = 0.05$.

Soluzione

La statistica test chi-quadrato assume il valore

$$\chi^2 = \frac{[12 - (32 \cdot 70)/218]^2}{(32 \cdot 70)/218} + \frac{[35 - (107 \cdot 70)/218]^2}{(107 \cdot 70)/218} + \cdots + \frac{[26 - (79 \cdot 63)/218]^2}{(79 \cdot 63)/218} = 1.566.$$

La soglia critica è $\chi^2_{0.95} = 9.488$ (i gradi di libertà sono 4). Poiché $\chi^2 = 1.566 < 9.49 = \chi^2_{0.95}$, l'ipotesi nulla non viene rifiutata. ■

Riassunto

I problemi di verifica di ipotesi affrontati in questo capitolo si riferiscono a situazioni in cui i dati sono costituiti da conteggi o frequenze derivanti dalla classificazione delle unità del campione secondo le modalità di una o due variabili. Nell'impostazione di base, la verifica di ipotesi viene eseguita come segue. Si considera una popolazione multinomiale:

Modalità	A_1	A_2	...	A_k	Totale
Probabilità	p_1	p_2	...	p_k	1

dove $A_1, A_2, \ldots, A_k$ sono le modalità della variabile oggetto di studio e $p_1, p_2, \ldots, p_k$ le probabilità di osservarle in una estrazione casuale dalla popolazione.

[2] Questo numero si ottiene come segue. Ci sono s vincoli rappresentati dai totali di riga nella Tabella 22.17 (cioè le dimensioni assegnate dei campioni) e $t-1$ stime delle probabilità $p_1, p_2, \ldots, p_t$. Perciò, i gradi di libertà sono: $s \cdot t - s - (t-1) = (s-1)(t-1)$.

Estrarre casualmente una unità dalla popolazione multinomiale è come fare una prova di un esperimento casuale i cui esiti sono $A_1, A_2, \ldots, A_k$ con probabilità $p_1, p_2, \ldots, p_k$, rispettivamente. Effettuare n prove indipendenti dell'esperimento in modo tale che dette probabilità non varino da una prova all'altra equivale a estrarre un campione casuale di ampiezza n dalla popolazione multinomiale. Indicando con n_i la frequenza con cui A_i si verifica, il campione può essere organizzato sotto forma di distribuzione di frequenza:

Modalità	A_1	A_2	...	A_k	Totale
Frequenza	n_1	n_2	...	n_k	n

dove n_1 è la frequenza di A_1, n_2 la frequenza di A_2 e così via.

L'ipotesi nulla sui parametri di una popolazione multinomiale consiste nell'assegnare un dato valore alle probabilità $p_1, p_2, \ldots, p_k$. Formalmente, l'ipotesi nulla è $H_0 : p_i = p_{i0}$, $i = 1, 2, \ldots, k$. Sotto H_0, il valore atteso di $n_1, n_2, \ldots, n_k$ è dato, rispettivamente, da $np_{10}, np_{20}, \ldots, np_{k0}$.

Le frequenze $n_1, n_2, \ldots, n_k$ del campione casuale vengono confrontate con le corrispondenti frequenze attese sotto l'ipotesi nulla, usando la statistica test chi-quadrato

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}},$$

che, se H_0 è vera, ha, approssimativamente, distribuzione chi-quadrato con $k-1$ gradi di libertà, purché la dimensione del campione n sia sufficientemente ampia. L'ipotesi nulla viene poi rifiutata per valori elevati della statistica test.

Se le probabilità $p_1, p_2, \ldots, p_k$ dipendono da q parametri incogniti, stimati utilizzando i dati del campione, la statistica test chi-quadrato che ne risulta si distribuisce approssimativamente come una chi-quadrato con $k-q-1$ gradi di libertà. La statistica test chi-quadrato, opportunamente adattata, può essere utilizzata anche per verificare l'ipotesi di indipendenza tra due variabili e l'ipotesi nulla che s campioni provengano da s popolazioni multinomiali uguali, cioè da popolazioni in cui A_j, $j = 1, 2, \ldots, t$, ha la stessa probabilità p_j in tutte le s popolazioni.

Parole chiave

distribuzione di frequenza attesa, 481
popolazione multinomiale, 479
popolazione multinomiale doppia, 484
statistica test chi-quadrato, 481
test della bontà dell'adattamento, 482

Commenti e curiosità

Karl Pearson e il test chi-quadrato

History and Art Collection/Alamy Stock Photo

Karl Pearson (1857-1936), statistico, antropologo e filosofo della scienza, succedette al suo amico e maestro Galton nella cattedra di eugenetica presso l'University College London. Il suo nome di battesimo era Carl, che, in onore di Karl Marx, cambiò in Karl. Grazie ai suoi studi, la teoria statistica compì importantissimi progressi. Le applicazioni dei suoi risultati teorici trovarono un'ampia diffusione anche per opera dei suoi numerosi allievi.

I contributi di Pearson alla teoria statistica riguardano i seguenti ambiti: curve di densità (costruzione di un sistema di curve di densità i cui parametri vengono stimati con il metodo dei momenti); correlazione (studio del coefficiente di correlazione e delle sue generalizzazioni, anche alla luce dei lavori di Galton); valutazione del grado di adattamento di distribuzioni teoriche ai dati osservati basata sulla statistica test chi-quadrato (statistica utilizzata, tra l'altro, per la verifica dell'ipotesi di indipendenza tra le variabili di una distribuzione doppia); costruzione delle tavole statistiche come strumento di lavoro per statistici e biometristi.

La maggior parte delle sue ricerche di statistica fu pubblicata su Biometrika, rivista fondata insieme a Galton e Weldon, e da lui diretta per i primi 35 anni della sua esistenza. Pearson è dunque, a pieno titolo, uno dei fondatori della statistica moderna.

Caratterialmente forte, egli era portato a rifiutare le imposizioni. Curioso a tale riguardo è il seguente aneddoto. Nel suo periodo di studi a Cambridge, era fortemente contrariato dal fatto che gli studenti dovessero obbligatoriamente frequentare la cappella universitaria e seguire le lezioni di religione. Dopo una dura battaglia sui principi, che vinse, stupì i responsabili accademici con la sua frequentazione sia della cappella sia delle lezioni di religione.

Il suo forte carattere incise negativamente anche sui rapporti personali con un altro famoso statistico inglese, Ronald Fisher; rapporti caratterizzati da insanabili divergenze e aspri scontri, dei quali si è fatto cenno nella nota 4 di "Commenti e curiosità" del Capitolo 18.

Esercizi

22.1 Nella tabella che segue viene presentata la distribuzione delle abitazioni occupate secondo il titolo di godimento (dati Istat, Censimento della popolazione, anno 2011):

Titolo di godimento	Proprietà	Affitto	Altro titolo
Frequenza rel.	0.71	0.20	0.09

Si supponga che in un'indagine campionaria del 2019 condotta su 2146 abitazioni si sia osservata la distribuzione di frequenza contenuta nella tabella che segue:

Titolo di godimento	Proprietà	Affitto	Altro titolo
Frequenza	1627	373	146

Si stabilisca se la distribuzione di frequenza osservata nell'indagine campionaria differisce significativamente da quella rilevata nel censimento del 2011. Si ponga $\alpha = 0.01$. (*Suggerimento*: si usi il test χ^2 per la verifica della bontà dell'adattamento.)

22.2 I dati della tabella che segue si riferiscono alla distribuzione di 100 foglie di albero classificate secondo la lunghezza:

Lunghezza (in mm)	Frequenza	Lunghezza (in mm)	Frequenza
80-89	1	140-149	23
90-99	0	150-159	11
100-109	3	160-169	8
110-119	10	170-179	1
120-129	19	180-189	2
130-139	21	190-199	1

Si verifichi l'ipotesi che il campione provenga da una popolazione normale ($\alpha = 0.05$).

22.3 La tabella che segue si riferisce alla distribuzione di un campione di 749 studenti inglesi classificati secondo il sesso e il profitto in matematica conseguito all'ultimo anno della scuola superiore (Hand *et al.*, 1994).

Sesso	Profitto A	B	C	D	Totale
Femmina	74	23	184	85	366
Maschio	82	13	176	112	383
Totale	156	36	360	197	749

Si verifichi l'ipotesi di indipendenza tra il sesso e il profitto in matematica, ponendo $\alpha = 0.01$.

22.4 La distribuzione che segue si riferisce a un campione di 1329 soggetti classificati secondo la pressione sanguigna e la presenza di disturbi cardiaci (Hand *et al.*, 1994).

Pressione	Disturbi cardiaci Sì	No	Totale
< 127	20	388	408
127-146	28	527	555
147-166	20	204	224
> 166	24	118	142
Totale	92	1237	1329

C'è evidenza sufficiente per escludere che vi sia indipendenza tra disturbi cardiaci e pressione sanguigna? Si ponga $\alpha = 0.05$.

22.5 In uno studio volto a comparare l'efficacia di due terapie, A e B, a un campione casuale di 65 pazienti fu assegnata la terapia A, mentre a un campione di 73 pazienti fu assegnata la terapia B. Al termine del periodo di cura, fu valutato lo stato di salute dei pazienti classificandoli in una delle tre categorie "Eccellente", "Medio", "Non buono". I risultati della valutazione sono esposti nella tabella che segue:

Terapia	Stato di salute Eccellente	Medio	Non buono	Totale
A	34	21	10	65
B	30	26	17	73
Totale	64	47	27	138

Si stabilisca se vi sia evidenza sufficiente per ritenere che le due terapie hanno efficacia diversa, ponendo $\alpha = 0.01$.

CAPITOLO

23

Inferenza sul modello di regressione lineare

Sommario del capitolo

23.1 Introduzione
23.2 Stima di β_0 e β_1
23.3 Stima della varianza del termine di errore
23.4 Stima intervallare e verifica di ipotesi su β_0 e β_1
23.5 Stima intervallare del valore atteso di Y
23.6 Intervallo predittivo per Y
Riassunto
Commenti e curiosità
Esercizi

23.1 Introduzione

Nel Capitolo 10, è stato illustrato come descrivere in modo approssimato la relazione statistica tra due variabili, X (variabile indipendente) e Y (variabile dipendente), mediante una retta, calcolandone intercetta e coefficiente angolare con il metodo dei minimi quadrati. È stato anche introdotto un indice – il coefficiente di determinazione – per misurare il grado di adattamento della retta trovata ai dati osservati. In questo capitolo, il tema verrà ripreso e sviluppato in ottica inferenziale.

Sia x un valore specifico della variabile indipendente X. Si assuma che la variabile dipendente Y sia in relazione con x secondo il modello probabilistico

$$Y = \beta_0 + \beta_1 x + \varepsilon, \tag{23.1}$$

dove β_0 e β_1 sono costanti non note (parametri) e ε è una variabile casuale con valore atteso uguale a 0, chiamata termine di errore.

La relazione (23.1) è detta modello di regressione lineare semplice. Si parla invece di modello di regressione lineare multipla quando la variabile dipendente è posta in relazione lineare con due o più variabili indipendenti. In questo capitolo, la discussione sarà limitata al primo caso.

Il modello (23.1) presenta due componenti: una deterministica, $\beta_0 + \beta_1 x$, che mostra come varia Y al variare della variabile indipendente; l'altra stocastica, il termine di errore ε, che rappresenta il contributo di tutti i fattori diversi dalla variabile indipendente che possono influire su Y e che non vengono considerati esplicitamente nel modello.

ESEMPIO 23.1

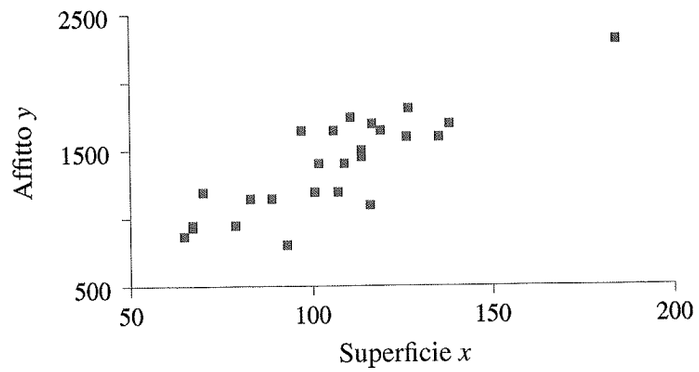
I dati della Tabella 23.1 si riferiscono a 25 appartamenti di un quartiere di una grande città americana, dei quali è data la dimensione (in m^2) e l'importo (in dollari) dell'affitto mensile (Levine *et al.*, 2012, p. 532).

Il grafico di dispersione mostra che l'affitto tende a crescere all'aumentare della superficie dell'abitazione. Questo andamento può essere descritto dal modello (23.1), dove $\beta_0 + \beta_1 x$ ha il ruolo di approssimare l'andamento crescente dei dati, mentre il termine di errore ε dà conto dell'impatto sul livello dell'affitto di altre variabili, diverse dalla superficie, non incluse nel modello (come la qualità delle finiture, il piano in cui si trova l'abitazione, il numero delle stanze, il numero di bagni ecc.).

abella 23.1
bitazioni secondo
superficie e l'affitto
ensile.

Abitazione i	Superficie x_i	Affitto y_i	Abitazione i	Superficie x_i	Affitto y_i
1	79	950	14	127	1800
2	135	1600	15	109	1400
3	101	1200	16	114	1450
4	114	1500	17	116	1100
5	67	950	18	117	1700
6	138	1700	19	107	1200
7	106	1650	20	83	1150
8	67	935	21	126	1600
9	65	875	22	97	1650
10	89	1150	23	70	1200
11	102	1400	24	93	800
12	119	1650	25	111	1750
13	184	2300			

Figura 23.1
Grafico di dispersione
relativo ai dati della
Tabella 23.1.



Il modello (23.1) può essere scritto, in modo equivalente, nella forma

$$\mathrm{E}(Y) = \beta_0 + \beta_1 x. \tag{23.2}$$

L'Equazione (23.2) mostra come il valore atteso della variabile dipendente Y associata al valore x della variabile indipendente sia legato linearmente a x. Geometricamente, al variare di x, il valore atteso $\mathrm{E}(Y)$ si muove lungo una retta. Nella (23.2) β_0 è l'intercetta e β_1 il coefficiente angolare. Chiaramente, β_0 è il valore di $\mathrm{E}(Y)$ quando $x = 0$, mentre β_1 rappresenta la variazione che subisce $\mathrm{E}(Y)$ per un incremento unitario di x.

I modelli (23.1) e (23.2) sono illustrati nella Figura 23.2. Le tre curve rappresentano la distribuzione di probabilità delle variabili casuali Y_1, Y_2 e Y_3 associate ai valori x_1, x_2 e x_3 di X. Ad esempio, la distribuzione di probabilità di Y_1 è descritta dalla funzione di densità $f(y_1)$ definita per tutti i punti della linea tratteggiata che partono da x_1. In base all'Equazione (23.2), i valori attesi $\mathrm{E}(Y_1)$, $\mathrm{E}(Y_2)$ e $\mathrm{E}(Y_3)$ giacciono su una retta (linea a punti).

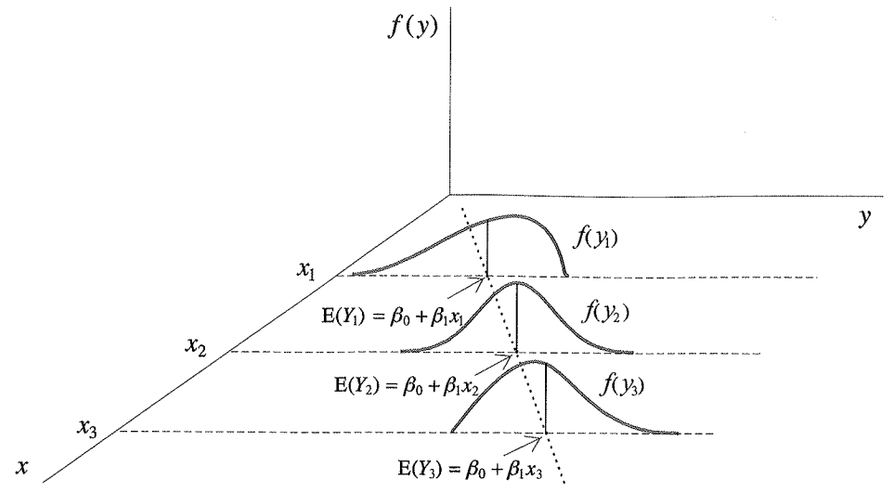


Figura 23.2
Interpretazione geometrica delle Equazioni (23.1) e (23.2).

Si supponga di disporre della serie doppia di osservazioni:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

dove $x_1, x_2, \ldots, x_n$ sono i valori assegnati alla variabile X e $y_1, y_2, \ldots, y_n$ le determinazioni delle variabili casuali $Y_1, Y_2, \ldots, Y_n$ associate a $x_1, x_2, \ldots, x_n$.

Il primo passo da compiere nell'analisi di regressione è quello di rappresentare il campione osservato con un grafico di dispersione, come si è fatto nell'Esempio 23.1. Se il grafico mostra un andamento lineare, l'analisi procede con la stima dell'intercetta e del coefficiente angolare e con l'inferenza sulla retta di regressione.

23.2 Stima di β_0 e β_1

Il metodo dei minimi quadrati può essere utilizzato per ottenere l'equazione della retta che interpola i dati in un contesto descrittivo, come si è fatto nel Capitolo 10, oppure come tecnica di stima puntuale dei parametri del modello (23.2), nel qual caso, è necessario formulare precise ipotesi sul processo generatore dei dati affinché le stime da esso prodotte siano di buona qualità. Tali ipotesi riguardano il termine di errore ε.

Come si è detto, nei dati osservati, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, le quantità $y_1, y_2, \ldots, y_n$ sono pensate come determinazioni delle variabili casuali

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

dove si assume che $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$:

- abbiano valore atteso 0: $\mathrm{E}(\varepsilon_i) = 0$, $i = 1, 2, \ldots, n$;
- abbiano la stessa varianza: $\mathrm{Var}(\varepsilon_i) = \sigma^2$, $i = 1, 2, \ldots, n$;
- siano incorrelati: $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$, per ogni $i \neq j$.

Queste ipotesi sono dette assunzioni di base dei minimi quadrati. In particolare, la seconda è denominata omoschedasticità.

Poiché le variabili casuali ε_i e Y_i differiscono solo per una costante, le ipotesi di omoschedasticità e incorrelazione valgono anche per le variabili casuali Y_i.

Data la serie doppia di osservazioni $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, il problema principale dell'analisi di regressione riguarda la stima dei parametri β_0 e β_1 dell'Equazione (23.2) con il *metodo dei minimi quadrati*, i cui dettagli sono stati descritti nel Capitolo 10 al quale si rinvia. La Proposizione 23.1 ripropone i risultati con qualche cambiamento di notazione.

PROPOSIZIONE 23.1

Le stime dei minimi quadrati dei parametri β_1 e β_0 del modello (23.2) sono date da:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad b_0 = \bar{y} - b_1\bar{x}, \tag{23.3}$$

dove

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{e} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

sono le medie campionarie delle variabili indipendente e dipendente.

Come nel Capitolo 10, si indicano con C_{XY} e D_X il numeratore e il denominatore del rapporto della prima delle (23.3), che sono la codevianza tra X e Y e la devianza di X.

Una volta ottenuti b_0 e b_1, si dispone della retta di regressione stimata:

$$\hat{y} = b_0 + b_1 x. \tag{23.4}$$

Nella Figura 23.3 viene rappresentata la retta di regressione stimata insieme ai punti osservati. Tale retta, tra tutte le possibili rette del piano, è quella che rende minima la somma dei quadrati delle differenze tra i valori osservati $y_1, y_2, \ldots, y_n$ e i valori teorici, $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$, che si ottengono sostituendo a x i valori $x_1, x_2, \ldots, x_n$ nella (23.4). Tali differenze, date da

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n,$$

e rappresentate dalle deviazioni verticali dei punti osservati rispetto alla retta di regressione, sono chiamati residui.

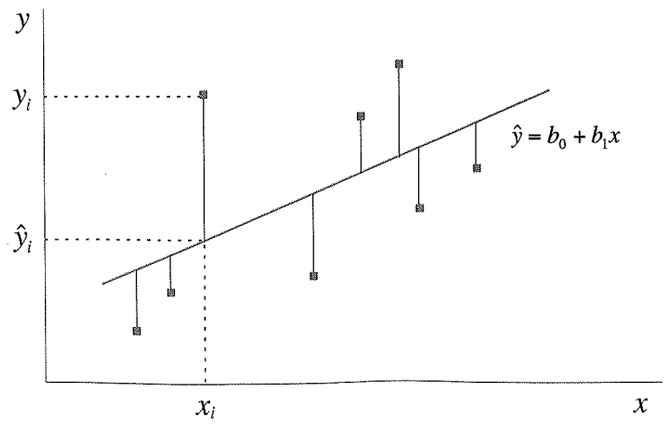


Figura 23.3
Punti osservati, retta di regressione stimata e residui.

ESEMPIO 23.2

Si stimi la retta di regressione per i dati dell'Esempio 23.1 e la si rappresenti nel grafico di dispersione.

Soluzione

La stima di β_0 e β_1 richiede il calcolo delle quantità:

$$\bar{x} = 105.44, \quad \bar{y} = 1386.40,$$

$$D_X = \sum_{i=1}^{25}(x_i - \bar{x})^2 = 17\,192.16, \quad C_{XY} = \sum_{i=1}^{25}(x_i - \bar{x})(y_i - \bar{y}) = 197\,219.60.$$

Ne segue che:

$$b_1 = \frac{197\,219.60}{17\,192.16} = 11.47, \quad b_0 = 1386.40 - 11.47 \cdot 105.44 = 177.$$

La stima $b_1 = 11.47$ indica che a un incremento di 1 m^2 della dimensione dell'abitazione corrisponde mediamente un aumento di 11.47 \$ dell'affitto mensile. La retta di regressione stimata è rappresentata nella Figura 23.4.

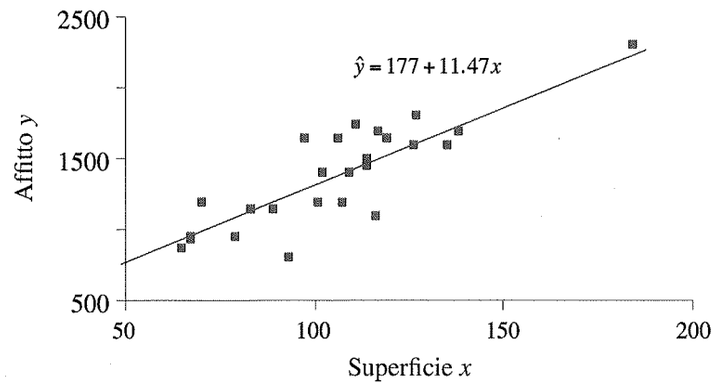


Figura 23.4
Grafico di dispersione e retta di regressione stimata per i dati dell'Esempio 23.1.

■

Proprietà degli stimatori dei minimi quadrati

La prima delle (23.3) può anche essere scritta nella forma

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} w_{1i} y_i,$$

dove

$$w_{1i} = \frac{x_i - \bar{x}}{D_X}.$$

Analogamente, è facile stabilire che

$$b_0 = \bar{y} - b_1\bar{x} = \sum_{i=1}^{n} w_{0i} y_i,$$

dove

$$w_{0i} = \frac{1}{n} - w_{1i}\bar{x}$$

(si veda l'Esercizio 23.1). Dunque, b_0 e b_1 sono particolari combinazioni lineari di $y_1, y_2, \ldots, y_n$. Se, nelle espressioni precedenti, si sostituiscono alle osservazioni $y_1, y_2, \ldots, y_n$ le variabili casuali $Y_1, Y_2, \ldots, Y_n$, si passa dalle stime b_0 e b_1 ai corrispondenti stimatori:

$$B_1 = \sum_{i=1}^{n} w_{1i} Y_i \quad \text{e} \quad B_0 = \sum_{i=1}^{n} w_{0i} Y_i$$

denominati *stimatori dei minimi quadrati* di β_1 e β_0, le cui proprietà sono descritte nella Proposizione 23.2.

PROPOSIZIONE 23.2

Gli stimatori dei minimi quadrati B_0 e B_1 dei parametri β_0 e β_1 sono entrambi non distorti, in quanto:

$$\mathrm{E}(B_0) = \beta_0, \quad \mathrm{E}(B_1) = \beta_1.$$

Le loro varianze sono date da:

$$\sigma^2_{B_0} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{D_X}\right), \quad \sigma^2_{B_1} = \frac{\sigma^2}{D_X}.$$

La dimostrazione è riportata nel Paragrafo B.10 dell'Appendice B. ■

Le deviazioni standard di B_0 e B_1 sono allora:

$$\sigma_{B_0} = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{D_X}}, \quad \sigma_{B_1} = \frac{\sigma}{\sqrt{D_X}}. \tag{23.5}$$

Esse misurano gli errori medi che si commettono stimando β_0 e β_1 con B_0 e B_1. Per esempio, se la deviazione standard di B_1 è pari a 2.27, questo numero esprime l'errore medio che si commette stimando β_1 con B_1.

Osservando le (23.5), si desume che la precisione dei due stimatori:

- aumenta al diminuire di σ;
- aumenta all'aumentare di D_X;
- aumenta all'aumentare di n se, come avviene ordinariamente, D_X aumenta con n.

Si può dimostrare che B_0 e B_1 sono i "migliori stimatori lineari".

PROPOSIZIONE 23.3

Gli stimatori dei minimi quadrati B_0 e B_1 di β_0 e β_1 sono i più efficienti nell'insieme degli stimatori non distorti che sono combinazioni lineari delle variabili casuali $Y_1, Y_2, \ldots, Y_n$, a condizione che le ipotesi di base dei minimi quadrati siano valide.

Per la dimostrazione, si veda Casella e Berger (2002), pp. 545-548. ■

23.3 Stima della varianza del termine di errore

Una stima della varianza del termine di errore è necessaria per stimare le deviazioni standard di B_0 e B_1, che a loro volta sono indispensabili per fare inferenza sulla retta di regressione.

Si consideri la variabile casuale

$$\hat{Y}_i = B_0 + B_1 x_i,$$

per la quale vale la relazione

$$\mathrm{E}(\hat{Y}_i) = \beta_0 + \beta_1 x_i. \tag{23.6}$$

La (23.6) mette in luce come $\hat{Y}_i$ sia uno stimatore non distorto del valore atteso della variabile dipendente Y_i associata al valore x_i della variabile indipendente. Si considerino ora le variabili casuali

$$e_i = Y_i - \hat{Y}_i, \; i = 1, 2, \ldots, n,$$

ossia le differenze tra le variabili casuali Y_i e gli stimatori del loro valore atteso (che sono i residui visti come variabili casuali). Tali differenze rappresentano la base per la stima della varianza del termine di errore.[1]

PROPOSIZIONE 23.4

Uno stimatore non distorto della varianza σ^2 del termine di errore è dato da

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2. \tag{23.7}$$

La dimostrazione è contenuta nel Paragrafo B.10 dell'Appendice B. ■

La somma dei quadrati dei residui nella (23.7) è detta devianza residua (già esaminata nel Capitolo 10 in ambito descrittivo) e può essere espressa, in modo alternativo, come

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 - B_1^2 D_X$$

(si veda il Paragrafo A.7 dell'Appendice A).

Una volta stimata σ^2, le deviazioni standard di B_0 e B_1 si stimano come segue:

$$\hat{\sigma}_{B_0} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{D_X}}, \quad \hat{\sigma}_{B_1} = \frac{\hat{\sigma}}{\sqrt{D_X}}. \tag{23.8}$$

[1] In base alle Equazioni (23.1) e (23.2), si ha

$$\varepsilon_i = Y_i - \mathrm{E}(Y_i).$$

È quindi ragionevole predire ε_i usando $e_i = Y_i - \hat{Y}_i$.

ESEMPIO 23.3

Si stimi σ^2 nella regressione dell'Esempio 23.2.

Soluzione

Nella Tabella 23.2 sono riportati i valori osservati y_i e i valori teorici $\hat{y}_i$, calcolati come indicato in precedenza. Ad esempio, ponendo $x = x_1 = 79$, si ricava $\hat{y}_1 = 177 + 11.47 \cdot 79 = 1083.13$.

abella 23.2
alori osservati e valori orici di Y nella gressione di cui l'Esempio 23.2.

Valore osservato y_i	Valore teorico $\hat{y}_i$	Valore osservato y_i	Valore teorico $\hat{y}_i$
950	1083.13	1800	1633.69
1600	1725.45	1400	1427.23
1200	1335.47	1450	1484.58
1500	1484.58	1100	1507.52
950	945.49	1700	1518.99
1700	1759.86	1200	1404.29
1650	1392.82	1150	1129.01
935	945.49	1600	1622.22
875	922.55	1650	1289.59
1150	1197.83	1200	979.90
1400	1346.94	800	1243.71
1650	1541.93	1750	1450.17
2300	2287.48		

Applicando la (23.7), si ottiene

$$\hat{\sigma}^2 = \frac{(950 - 1083.13)^2 + (1600 - 1725.45)^2 + \cdots + (1750 - 1450.17)^2}{23}$$
$$= 38\,144.54.$$

Infine, in base alle (23.8), si ha:

$$\hat{\sigma}_{B_0} = \sqrt{38\,144.54\left(\frac{1}{25} + \frac{105.44^2}{17\,192.16}\right)} = 161.84, \quad \hat{\sigma}_{B_1} = \sqrt{\frac{38\,144.54}{17\,192.16}} = 1.49.$$

■

23.4 Stima intervallare e verifica di ipotesi su β_0 e β_1

D'ora in poi in questo capitolo, si assumerà che i termini di errore $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ siano indipendenti e distribuiti normalmente con la stessa varianza σ^2. Questa ipotesi, che implica normalità e indipendenza di $Y_1, Y_2, \ldots, Y_n$ alla luce dell'Equazione 23.1, consente di determinare facilmente la distribuzione campionaria di alcune variabili casuali utili per condurre l'inferenza sul modello di regressione.

La Figura 23.5 descrive l'ipotesi di normalità delle variabili $Y_1, Y_2, \ldots, Y_n$ per il caso di $n = 3$.

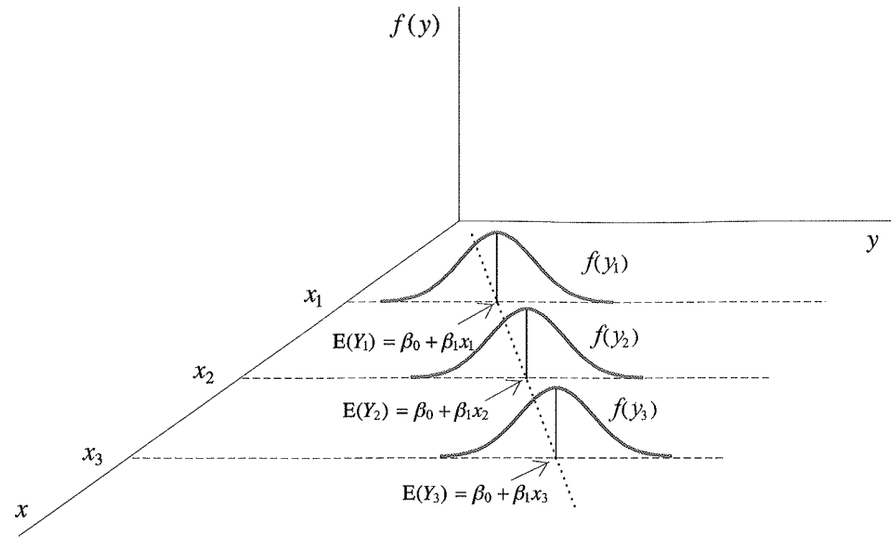


Figura 23.5
Ipotesi di normalità nel modello di regressione.

PROPOSIZIONE 23.5

Se le variabili casuali $Y_1, Y_2, \ldots, Y_n$ sono indipendenti e distribuite normalmente con la stessa varianza, le variabili casuali

$$T_{B_0} = \frac{B_0 - \beta_0}{\hat{\sigma}_{B_0}} \quad \text{e} \quad T_{B_1} = \frac{B_1 - \beta_1}{\hat{\sigma}_{B_1}}$$

hanno entrambe distribuzione T di Student con $n - 2$ gradi di libertà.

Per la dimostrazione, si veda Hogg *et al.* (2013), p. 500. ■

Stime intervallari

Sia $t_{1-\alpha/2}$ il quantile di livello $1 - \alpha/2$ della T di Student con $n-2$ gradi di libertà. Allora, grazie alla Proposizione 23.5, le variabili casuali

$$L_1 = B_0 - t_{1-\alpha/2}\hat{\sigma}_{B_0} \quad \text{e} \quad L_2 = B_0 + t_{1-\alpha/2}\hat{\sigma}_{B_0}$$

includono il parametro β_0 con probabilità $1 - \alpha$. Gli estremi dell'intervallo di confidenza all'$(1 - \alpha)100\%$ per β_0 sono allora:

$$l_1 = b_0 - t_{1-\alpha/2}\hat{\sigma}_{B_0}, \quad l_2 = b_0 + t_{1-\alpha/2}\hat{\sigma}_{B_0}.$$

In modo analogo, si ricavano facilmente gli estremi dell'intervallo di confidenza all'$(1 - \alpha)100\%$ per β_1:

$$l_1 = b_1 - t_{1-\alpha/2}\hat{\sigma}_{B_1}, \quad l_2 = b_1 + t_{1-\alpha/2}\hat{\sigma}_{B_1}.$$

ESEMPIO 23.4

Si costruisca l'intervallo di confidenza al 95% per β_0 e β_1 nella regressione dell'Esempio 23.2.

Soluzione

Le stime di β_0 e β_1 sono:

$$b_0 = 177, \quad b_1 = 11.47;$$

le corrispondenti deviazioni standard stimate sono:

$$\hat{\sigma}_{B_0} = 161.84, \quad \hat{\sigma}_{B_1} = 1.49$$

(si veda l'Esempio 23.3). Gli estremi dell'intervallo di confidenza per β_0 sono:

$$l_1 = 177 - 2.069 \cdot 161.84 = -157.85, \quad l_2 = 177 + 2.069 \cdot 161.84 = 511.85,$$

essendo $t_{0.975} = 2.069$ il quantile di livello 0.975 della distribuzione T di Student con 23 gradi di libertà. In modo analogo, gli estremi dell'intervallo di confidenza per β_1 sono:

$$l_1 = 11.47 - 2.069 \cdot 1.49 = 8.39, \quad l_2 = 11.47 + 2.069 \cdot 1.49 = 14.55.$$

Si può concludere, con una fiducia del 95%, che all'aumentare di 1 m^2 della superficie dell'abitazione l'affitto mensile aumenta mediamente di una misura che va da 8.39 a 14.55 dollari. ■

Verifica di ipotesi

Sia $H_0 : \beta_1 = \bar{\beta}_1$ l'ipotesi nulla, dove $\bar{\beta}_1$ è un particolare valore del coefficiente angolare della retta di regressione. Sia $H_1 : \beta_1 \neq \bar{\beta}_1$ l'ipotesi alternativa. Un'idonea statistica test è data da

$$T_{B_1} = \frac{B_1 - \bar{\beta}_1}{\hat{\sigma}_{B_1}}, \tag{23.9}$$

che, se è vera l'ipotesi nulla, ha distribuzione T di Student con $n-2$ gradi di libertà (si veda la Proposizione 23.5). Per l'ipotesi alternativa bidirezionale, si perviene alla zona di rifiuto

$$R = \{t_{b_1} : |t_{b_1}| > t_{1-\alpha/2}\},$$

dove

$$t_{b_1} = \frac{b_1 - \bar{\beta}_1}{\hat{\sigma}_{B_1}}$$

è il valore osservato della statistica test.

Le versioni unidirezionali della statistica test (23.9) possono essere facilmente derivate e sono pertanto omesse per brevità.

L'ipotesi nulla $H_0 : \beta_1 = 0$ riveste particolare interesse, perché indica l'assenza di legame lineare tra X e Y. La statistica test assume la forma

$$T_{B_1} = \frac{B_1}{\hat{\sigma}_{B_1}}. \tag{23.10}$$

Un modo equivalente per verificare l'ipotesi $H_0 : \beta_1 = 0$ contro l'alternativa bidirezionale è basato sul rapporto

$$F = \frac{D_{SL}}{D_{RL}/(n-2)}, \tag{23.11}$$

dove

$$D_{SL} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \quad \text{e} \quad D_{RL} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

sono la devianza spiegata e quella residua. La (23.11) si ottiene elevando al quadrato la (23.10) e sfruttando la relazione $D_{SL} = B_1^2 D_X$ (si veda il Paragrafo A.7 dell'Appendice A).

Si può dimostrare che il rapporto in questione ha distribuzione F con 1 e $n-2$ gradi di libertà. Di conseguenza, indicando con $f_{1-\alpha}$ il quantile di livello $1-\alpha$ di tale distribuzione, l'ipotesi nulla di assenza di legame lineare viene rifiutata quando il valore osservato di F, f, è maggiore di $f_{1-\alpha}$.

Le quantità necessarie per condurre il test F sono generalmente presentate nella cosiddetta tabella di analisi della varianza, di cui alla Tabella 23.3.

Fonte di variabilità	gl	Dev	Dev/gl	F	α_{oss}
Regressione	1	D_{SL}	D_{SL}	f	$P(F > f\,;\, H_0)$
Errore	$n-2$	D_{RL}	$\dfrac{D_{RL}}{n-2}$		
Totale	$n-1$	D_T			

Tabella 23.3
Tabella di analisi della varianza

dove $D_T = D_{SL} + D_{RL}$ (si veda il Paragrafo A.7 dell'Appendice A).

ESEMPIO 23.5

I dati della Tabella 23.4 si riferiscono alle vendite settimanali di scarpe sportive di due marche diverse, A e B, osservate in 12 settimane successive in un negozio di una grande città.

Marca A (x)	75	62	64	72	66	53	60	64	67	78	72	80
Marca B (y)	164	130	140	158	144	126	132	141	149	175	166	182

Tabella 23.4
Vendite settimanali di scarpe sportive delle marche A e B.

Si determini la retta di regressione che mette in relazione le vendite della marca A (X) con quelle della marca B (Y) e si verifichi l'ipotesi nulla $\beta_1 = 0$ contro l'alternativa $\beta_1 \neq 0$, ponendo $\alpha = 0.05$.

Soluzione

Come è facile verificare, si ha:

$$b_0 = -1.72, \quad b_1 = 2.25, \quad \sum_{i=1}^{12} (y_i - \hat{y}_i)^2 = 210.68, \quad D_X = 686.25.$$

Il valore osservato della statistica test T_{B_1}, dato da

$$t_{b_1} = \frac{2.25 - 0}{\sqrt{\dfrac{210.68}{10}\dfrac{1}{686.25}}} = 12.84,$$

è maggiore del quantile di livello 0.975 della distribuzione T di Student con 10 gradi di libertà, che è pari a $t_{0.975} = 2.228$. L'ipotesi nulla viene pertanto rifiutata.

Si perviene al medesimo risultato utilizzando il test F, i cui dettagli di calcolo sono riportati nella Tabella 23.5.

bella 23.5
bella di analisi della ianza per la verifica l'ipotesi nulla di assenza dipendenza lineare per ti riportati nella bella 23.4.

Fonte di variabilità	gl	Dev	Dev/gl	F	α_{oss}
Regressione	1	3468.2369	3468.2369	164.6226	0.0000
Errore	10	210.6798	21.0678		
Totale	11	3678.9167			

Il valore assunto da α_{oss} è vicino allo 0; ciò porta al rifiuto dell'ipotesi nulla. ■

Per la verifica dell'ipotesi $H_0: \beta_0 = \bar{\beta}_0$, dove $\bar{\beta}_0$ è un valore assegnato, la statistica test da utilizzare è T_{B_0} (si veda la Proposizione 23.5), seguendo una procedura simile a quella descritta sopra per T_{B_1}. Particolare interesse riveste l'ipotesi di intercetta nulla, $H_0: \beta_0 = 0$. Va osservato che il non rifiuto di questa ipotesi equivale ad assumere che la retta di regressione passi per l'origine. In questa situazione, si rende necessaria la determinazione di una nuova stima di β_1, per tener conto del vincolo di passaggio della retta per l'origine (si veda l'Esercizio 23.5).

23.5 Stima intervallare del valore atteso di Y

Nel paragrafo precedente, sono state illustrate le tecniche di inferenza relative ai parametri della retta di regressione. L'obiettivo di questo paragrafo è di costruire l'intervallo di confidenza per il valore atteso della variabile dipendente associato a un valore specifico, x_0, della variabile indipendente. In altri termini, si vuole rispondere alla domanda: se la variabile indipendente assumesse il livello x_0, quale valore assumerebbe in media la variabile dipendente?

In questo caso, il parametro di interesse è il valore atteso

$$\mathrm{E}(Y_0) = \beta_0 + \beta_1 x_0,$$

il cui stimatore naturale è

$$\hat{Y}_0 = B_0 + B_1 x_0.$$

Poiché $B_0 = \bar{Y} - B_1\bar{x}$, la variabile casuale $\hat{Y}_0$ può essere scritta nella forma

$$\hat{Y}_0 = \bar{Y} + (x_0 - \bar{x})B_1. \tag{23.12}$$

Dunque, $\hat{Y}_0$ è una combinazione lineare delle variabili casuali $\bar{Y}$ e B_1 con coefficienti 1 e $(x_0 - \bar{x})$, rispettivamente.[2] Applicando le (15.5) e (15.7) per determinare il valore atteso e la varianza della combinazione lineare appena richiamata, si trova che $\hat{Y}_0$ è uno stimatore non distorto di $\mathrm{E}(Y_0)$ con varianza

$$\sigma^2_{\hat{Y}_0} = \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{D_X}\right] \tag{23.13}$$

(si veda il Paragrafo B.10 dell'Appendice B).

[2] È facile verificare che la quantità $\hat{Y}_0$ può anche essere scritta nella forma di combinazione lineare delle variabili casuali $Y_1, Y_2, \ldots, Y_n$:

$$\hat{Y}_0 = \sum_{i=1}^{n}\left[\frac{1}{n} + \frac{x_i - \bar{x}}{D_X}(x_0 - \bar{x})\right]Y_i.$$

PROPOSIZIONE 23.6

Si assuma che $Y_1, Y_2, \ldots, Y_n$ siano indipendenti e distribuite normalmente con la stessa varianza. Allora il rapporto

$$T_{\hat{Y}_0} = \frac{\hat{Y}_0 - \mathrm{E}(Y_0)}{\hat{\sigma}_{\hat{Y}_0}}$$

ha distribuzione T di Student con $n - 2$ gradi di libertà.

Con questo risultato, è immediato ricavare gli estremi dell'intervallo di confidenza all'$(1 - \alpha)100\%$ per $\mathrm{E}(Y_0)$:

$$l_1 = \hat{y}_0 - t_{1-\alpha/2}\hat{\sigma}_{\hat{Y}_0}, \quad l_2 = \hat{y}_0 + t_{1-\alpha/2}\hat{\sigma}_{\hat{Y}_0}.$$

L'ampiezza dell'intervallo, pari a $2t_{1-\alpha/2}\hat{\sigma}_{\hat{Y}_0}$, è minima per $x_0 = \bar{x}$ e cresce all'aumentare della differenza, in valore assoluto, tra x_0 e $\bar{x}$, come mostra la Figura 23.6, che riporta l'ampiezza dell'intervallo di confidenza per $\mathrm{E}(Y_0)$ in funzione di x_0.

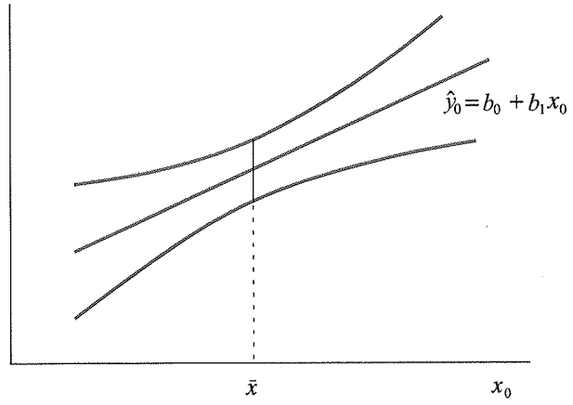


Figura 23.6
Ampiezza dell'intervallo di confidenza per $\mathrm{E}(Y_0)$ in funzione di x_0.

ESEMPIO 23.6

Si riprendano i dati dell'Esempio 23.5 e si determini l'intervallo di confidenza al 95% per il valore atteso della variabile dipendente, dato $x = 65$.

Soluzione

Poiché

$$\sum_{i=1}^{12}(y_i - \hat{y}_i)^2 = 210.68, \; D_X = 686.25, \; \hat{y}_0 = 144.53, \; t_{0.975} = 2.228$$

(le prime due quantità sono state calcolate nell'Esempio 23.5), si ha:

$$l_1 = 144.53 - 2.228\sqrt{\frac{210.68}{10}\left[\frac{1}{12} + \frac{(65 - 67.75)^2}{686.25}\right]} = 141.39,$$

$$l_2 = 144.53 + 2.228\sqrt{\frac{210.68}{10}\left[\frac{1}{12} + \frac{(65 - 67.75)^2}{686.25}\right]} = 147.67.$$

Si può pertanto affermare (con una fiducia del 95%) che per un volume di vendite pari a 65 per la marca A il volume medio di vendite per la marca B è compreso nell'intervallo (141.39, 147.67). ■

23.6 Intervallo predittivo per Y

Un problema sostanzialmente diverso da quello appena affrontato riguarda la previsione del valore di Y associato a un nuovo valore di X. Sia x_{n+1} una nuova osservazione di X. Indicato con Y_{n+1} il valore di Y associato a x_{n+1}, si vuole individuare un intervallo che includa con alta probabilità Y_{n+1}. A questo fine, si considerino le quantità

$$\hat{Y}_{n+1} = B_0 + B_1 x_{n+1} \quad \text{e} \quad W_{n+1} = \hat{Y}_{n+1} - Y_{n+1},$$

in cui $\hat{Y}_{n+1}$ ha la stessa struttura della (23.12), fatta salva la sostituzione di x_{n+1} a x_0, mentre W_{n+1}, nota come errore di predizione, è una combinazione lineare di $\hat{Y}_{n+1}$ e Y_{n+1}. Le varianze di $\hat{Y}_{n+1}$ e Y_{n+1} sono:

$$\sigma^2_{\hat{Y}_{n+1}} = \sigma^2 \left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{D_X} \right], \quad \sigma^2_{Y_{n+1}} = \sigma^2,$$

dove $\sigma^2_{\hat{Y}_{n+1}}$ si deduce dalla (23.13) con la sostituzione di x_{n+1} a x_0. Impiegando l'usuale regola per la determinazione della varianza di una combinazione lineare di variabili casuali indipendenti,[3] si ottiene

$$\sigma^2_{W_{n+1}} = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{D_X} \right].$$

Con questa premessa, è possibile enunciare la proposizione che segue.

PROPOSIZIONE 23.7

Se si assume che $Y_1, Y_2, \ldots, Y_n$ siano indipendenti e distribuite normalmente con la stessa varianza, il rapporto

$$T_{W_{n+1}} = \frac{\hat{Y}_{n+1} - Y_{n+1}}{\hat{\sigma}_{W_{n+1}}},$$

dove $\hat{\sigma}_{W_{n+1}}$ è lo stimatore di $\sigma_{W_{n+1}}$, ha distribuzione T di Student con $n - 2$ gradi di libertà.

Da questo risultato si deducono facilmente gli estremi dell'intervallo cercato:

$$l_1 = \hat{y}_{n+1} - t_{1-\alpha/2}\hat{\sigma}_{W_{n+1}}, \quad l_2 = \hat{y}_{n+1} + t_{1-\alpha/2}\hat{\sigma}_{W_{n+1}},$$

dove $t_{1-\alpha/2}$ è il quantile di livello $(1 - \alpha/2)$ della distribuzione T di Student con $n - 2$ gradi di libertà. Tale intervallo viene chiamato intervallo predittivo per distinguerlo dall'usuale intervallo di confidenza, in quanto l'oggetto dell'inferenza non è un parametro, ma una variabile casuale.

[3] Poiché $\hat{Y}_{n+1}$ è una combinazione lineare delle variabili casuali $Y_1, Y_2, \ldots, Y_n$, anche W_{n+1} è una combinazione lineare delle variabili casuali $Y_1, Y_2, \ldots, Y_{n+1}$.

ESEMPIO 23.7

Con gli stessi dati dell'esempio precedente, si determini l'intervallo predittivo al 95% per il valore della variabile dipendente associato a $x_{13} = 70$.

Soluzione

Si ha: $\hat{y}_{13} = -1.72 + 2.25 \cdot 70 = 155.78$ e $t_{0.975} = 2.228$ (i gradi di libertà sono 10). Di conseguenza, gli estremi dell'intervallo di predizione sono:

$$l_1 = 155.78 - 2.228 \sqrt{\frac{210.68}{10} \left[1 + \frac{1}{12} + \frac{(70 - 67.75)^2}{686.25} \right]} = 145.10,$$

$$l_2 = 155.78 + 2.228 \sqrt{\frac{210.68}{10} \left[1 + \frac{1}{12} + \frac{(70 - 67.75)^2}{686.25} \right]} = 166.46.$$

Dunque, per un volume di vendite di 70 per la marca A, quello per la marca B si colloca verosimilmente tra 145.10 e 166.46. ■

Riassunto

Questo capitolo è dedicato agli aspetti inferenziali del modello di regressione lineare semplice

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

dove ε è una variabile casuale a media nulla. I dati alla base dell'analisi di regressione sono rappresentati dalle n coppie di numeri:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

dove $x_1, x_2, \ldots, x_n$ sono valori assegnati alla variabile indipendente e $y_1, y_2, \ldots, y_n$ determinazioni delle variabili casuali $Y_1, Y_2, \ldots, Y_n$ associate a $x_1, x_2, \ldots, x_n$, rispettivamente, secondo il modello probabilistico sopra ricordato. Sotto l'ipotesi che $Y_1, Y_2, \ldots, Y_n$ abbiano la stessa varianza e siano incorrelate, i parametri β_0 e β_1 vengono stimati impiegando il metodo dei minimi quadrati. Vengono poi introdotti gli intervalli di confidenza e le procedure di verifica delle ipotesi per questi parametri, assumendo che le variabili casuali Y_i siano indipendenti e distribuite normalmente con la stessa varianza. Infine, vengono illustrati l'intervallo di confidenza per il valore atteso di Y e l'intervallo predittivo per il valore di Y associati a uno specifico valore della variabile indipendente.

Parole chiave

assunzioni di base dei minimi quadrati, 495
devianza residua, 499
errori di predizione, 506
intervallo predittivo, 506
modello di regressione lineare semplice, 493
omoschedasticità, 495
residui, 496
stime dei minimi quadrati, 496
stimatori dei minimi quadrati, 498
tabella di analisi della varianza, 503
termine di errore, 493
valori teorici, 496

Commenti e curiosità

1. Minimi quadrati: storia di una disputa

Nel volume *Theoria motus corporum coelestium in sectionibus conicis solem abientium* pubblicato nel 1809, Gauss applicò il metodo dei minimi quadrati nello studio delle orbite dei corpi celesti, utilizzando la curva normale per descrivere la distribuzione di probabilità della componente di errore dell'equazione di

Carl Friedrich Gauss
Georgios Kollidas/Alamy Stock Photo

regressione e proponendo algoritmi per gli aspetti computazionali delle stime (Stigler, 1981). In questo lavoro, Gauss citò regolarmente il volume del matematico francese Adrien-Marie Legendre (1752-1833), *Nouvelle méthodes pour la détermination des orbites des cometés*, pubblicato nel 1805 (Lizana, 2013), dove veniva presentato il metodo dei minimi quadrati (*méthode de moindre carrés*) senza alcun riferimento alla teoria della probabilità, cioè in un'ottica descrittiva. Pur riconoscendo implicitamente a Legendre la paternità del nome "metodo dei minimi quadrati" e l'antecedenza temporale della pubblicazione, Gauss sostenne di aver fatto uso di tale metodo già nel 1795.

a pubblicazione di Gauss suscitò l'immediata reazione di Legendre, che inviò una lettera a Gauss per vendicare la primogenitura del metodo dei minimi quadrati, mostrando un forte risentimento, come se on fosse la prima volta che il matematico tedesco disconoscesse i suoi contributi. L'attacco a Gauss roseguì fino al 1820, quando il matematico francese dette alla luce un supplemento al suo lavoro del 1805 Stigler, 1977), in cui rivendica ancora la paternità del metodo.

A favore della tesi di Gauss milita una nota del 1818 di H. W. Olbers, medico e astronomo tedesco Plackett, 1972), che conferma l'esistenza di annotazioni da parte di Gauss riguardanti il metodo dei mimi quadrati già prima del 1805. Va aggiunto che altri due astronomi, B. A. Lindenau e F. X. von Zach, chiamano il metodo dei minimi quadrati in alcuni lavori del 1806 e 1807: pur non facendo riferimento splicito a Gauss, non è inverosimile che questi abbia comunicato loro le sue annotazioni sul metodo elaorate prima del 1805. La tesi che Gauss avesse fatto uso del metodo dei minimi quadrati già nel 1795 rova sostegno infine in una sua lettera pubblicata nel 1799 in *Allgemeine Geographische Ephemeriden*, ove a proposito dei minimi quadrati parla di "mio metodo" (*meine methode*).

. L'onnipresenza della distribuzione *T* di Student

i sarà notato come, in questo capitolo, la distribuzione T di Student entri in gioco ogni volta che si parla i intervalli di confidenza o di verifica delle ipotesi. Ciò è dovuto all'assunzione di normalità che è alla ase di queste tecniche inferenziali. La variabile casuale T di Student ha la seguente struttura tipica

$$T_{\hat{\theta}} = \frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}},$$

ove $\hat{\theta}$ è lo stimatore non distorto di θ (avente distribuzione normale) e $\hat{\sigma}_{\hat{\theta}}$ è lo stimatore della deviazione tandard di $\hat{\theta}$. La variabile casuale considerata è una T di Student con un numero di gradi di libertà dato all'ampiezza del campione diminuita del numero di stime campionarie utilizzate nel calcolo di $\hat{\sigma}_{\hat{\theta}}$.

. Gli effetti della violazione delle assunzioni di base

'assunzione che $Y_1, Y_2, \ldots, Y_n$ siano indipendenti e distribuite normalmente con la stessa varianza hanno n ruolo fondamentale per la validità delle tecniche di inferenza esposte in questo capitolo. È lecito pertanto hiedersi quali siano gli effetti della violazione di tali assunzioni.

Premesso che l'assunzione di indipendenza non viene generalmente messa in discussione (salvo il aso in cui le osservazioni della variabile dipendente costituiscono una serie storica), va osservato che il enir meno della omoschedasticità non produce effetti sui valori attesi degli stimatori dei minimi quadrati $_0$ e B_1, che continuano a essere non distorti, ma ne lede l'efficienza: tali stimatori non sono più i "migliori timatori lineari", secondo quanto indicato nella Proposizione 23.3. Inoltre, viene alterata la distribuzione i probabilità di $T_{\hat{\theta}}$ ($\hat{\theta} = B_0, B_1, \hat{Y}_0, W_{n+1}$) e messa in dubbio la validità dello stimatore $\hat{\sigma}^2$.

Meno marcati sono gli effetti di moderate deviazioni dalla normalità. Violazioni più gravi producono effetti sul grado di copertura degli intervalli di confidenza e sui rischi di errore connessi ai procedimenti di verifica delle ipotesi. Una dimensione campionaria moderatamente elevata aiuta a superare queste difficoltà.

Esercizi

23.1* Si dimostri che gli stimatori B_0 e B_1 dei parametri β_0 e β_1 del modello 23.2 sono entrambi combinazioni lineari delle variabili casuali $Y_1, Y_2, \ldots, Y_n$.

23.2 Si riprendano i dati dell'Esercizio 10.1 e si svolgano le elaborazioni di seguito indicate:

a) si stimi il modello di regressione lineare del personale ausiliario sul personale medico;

b) si dia l'interpretazione del coefficiente angolare della retta di regressione stimata nel punto a;

c) si verifichi l'ipotesi di assenza di dipendenza lineare fra le due grandezze, assumendo un livello di significatività del 5%.

23.3 È stato condotto uno studio per stabilire se esista una dipendenza lineare fra la resistenza alla rottura, Y, di travi di legno e il peso specifico, X, del legno stesso. A tale scopo sono state estratte casualmente dieci travi della stessa sezione e sono state sottoposte a un test per verificarne la resistenza. I risultati sono esposti nella tabella che segue:

Trave i	Peso specifico x_i	Resistenza y_i	Trave i	Peso specifico x_i	Resistenza y_i
1	0.499	11.14	6	0.528	12.60
2	0.558	12.74	7	0.418	11.13
3	0.604	13.13	8	0.480	11.70
4	0.441	11.51	9	0.406	11.02
5	0.550	12.38	10	0.467	11.41

a) Si trovi l'equazione della retta di regressione di Y su X e la si rappresenti graficamente insieme ai punti osservati.

b) Si determini l'intervallo di confidenza al 95% per β_1 e se ne indichi il significato.

23.4 Si considerino i dati dell'Esercizio 1.4 concernenti il volume, il diametro e l'altezza di 31 esemplari di cedro nero.

a) Si stimino il modello di regressione del volume sul diametro e quello del volume sull'altezza.

b) Si adotti il modello che meglio si adatta ai dati osservati.

c) Sul modello prescelto si verifichi l'ipotesi di assenza di dipendenza lineare contro l'alternativa bidirezionale, calcolando il livello di significatività osservato (si ponga $\alpha = 0.01$).

23.5 Si riprendano i dati dell'Esercizio 10.3 e si svolgano le elaborazioni indicate di seguito:

a) si stimi il modello di regressione lineare della distanza stradale sulla distanza in linea d'aria;

b) si verifichi l'ipotesi $H_0 : \beta_0 = 0$ contro l'alternativa $H_1 : \beta_0 \neq 0$, ponendo $\alpha = 0.05$;

c) sulla base del risultato ottenuto al punto precedente, si stimi β_1;

d) utilizzando il modello prescelto, si determinino i residui.

23.6 I dati che seguono si riferiscono a una serie doppia di osservazioni dell'età (in settimane) del propellente e della potenza (in kg per cm^2):

Oss. i	Età x_i	Potenza y_i	Oss. i	Età x_i	Potenza y_i	Oss. i	Età x_i	Potenza y_i
1	15.50	151.91	8	2.50	181.20	15	22.00	124.22
2	23.75	118.09	9	7.50	165.93	16	18.00	144,51
3	8.00	162.98	10	11.00	158.80	17	6.00	169.90
4	17.00	145.05	11	13.00	152.37	18	12.50	154.85
5	5.50	155.34	12	3.75	168.86	19	2.00	186.78
6	19.00	120.21	13	25.00	125.24	20	21.50	123.41
7	24.00	125.59	14	9.75	164.44			

a) Si stimino i parametri del modello $Y = \beta_0 + \beta_1 x + \varepsilon$.

b) Si stimi σ^2.

c) Si determini l'intervallo di confidenza al 99% per β_1.

d) Si determinino l'intervallo di confidenza per il valore atteso e l'intervallo predittivo al 95% in corrispondenza di $x = 10$. Si metta in evidenza il diverso significato dei due intervalli.

23.7 Nella tabella che segue sono riportati, per 11 autovetture di marche diverse, i dati relativi al numero di km percorribili in città con un litro di carburante e alla cilindrata del motore:

Autovettura	Cilindrata x_i	km per litro y_i	Autovettura	Cilindrata x_i	km per litro y_i
A	1590	13.11	G	1590	10.99
B	2835	6.73	H	1393	12.40
C	1590	10.99	I	2835	6.73
D	2474	8.15	L	1590	10.99
E	5735	6.02	M	1868	11.34
F	1950	9.57			

a) Si stimino i parametri del modello $Y = \beta_0 + \beta_1 x + \varepsilon$.

b) Si determinino gli intervalli di confidenza al 95% per β_0 e β_1.

Inferenza sul modello di regressione con Excel

Nel Capitolo 10 è stato illustrata la procedura di Excel relativa all'analisi di regressione. In quella sede, la presentazione è stata limitata alla: stima dei parametri della retta di regressione; alla misura della bontà dell'adattamento con il coefficiente di determinazione; rappresentazione grafica della retta di regressione stimata insieme ai punti osservati.

Si illustrano ora le funzionalità di Excel per condurre l'inferenza sulla retta di regressione, utilizzando i dati dell'Esempio 23.1.

- Nel menu Analisi dei dati, si seleziona la voce Regressione.
- Nella finestra di dialogo, si selezionano le aree che contengono i dati relativi alla variabile dipendente e a quella indipendente. Inoltre, si selezionano le opzioni di output (si veda la Figura 23.7).
- Premendo OK, si ottengono i risultati descritti nella Figura 23.8.

I risultati vengono presentati in quattro distinte tabelle.

La tabella denominata "Regressione" riporta le seguenti quantità: R multiplo (la radice quadrata del coefficiente di determinazione r^2); R quadro (il coefficiente di determinazione r^2); R quadro aggiustato (una quantità non utilizzata nel modello di regressione semplice); errore standard (una stima di σ).

La seconda tabella, denominata "Analisi della varianza", si riferisce alla verifica dell'ipotesi H_0: $\beta_1 = 0$, contro l'alternativa $H_1 : \beta_1 \neq 0$, con la statistica test F, come illustrato nel Paragrafo 23.4.

La terza tabella contiene, nella prima riga (Intercetta): la stima di β_0; la stima della sua deviazione standard; il valore della statistica test T_{B_0} per la verifica dell'ipotesi nulla $H_0 : \beta_0 = 0$, contro l'alternativa $H_1 : \beta_0 \neq 0$; il livello di significatività osservato della statistica test T_{B_0}; gli estremi dell'intervallo di confidenza al 95% per β_0 (al riguardo, se nella finestra di dialogo si seleziona un livello di confidenza diverso da 95%, vengono dati sia gli estremi dell'intervallo di confidenza per il livello di confidenza selezionato sia quelli corrispondenti al livello di confidenza 95%). Nella seconda riga della tabella, sono riportate le stesse informazioni sopra elencate ma riferite al parametro β_1.

Infine, nell'ultima tabella, denominata "Output residui", vengono riportati i valori teorici $\hat{y}_i$ e i residui.

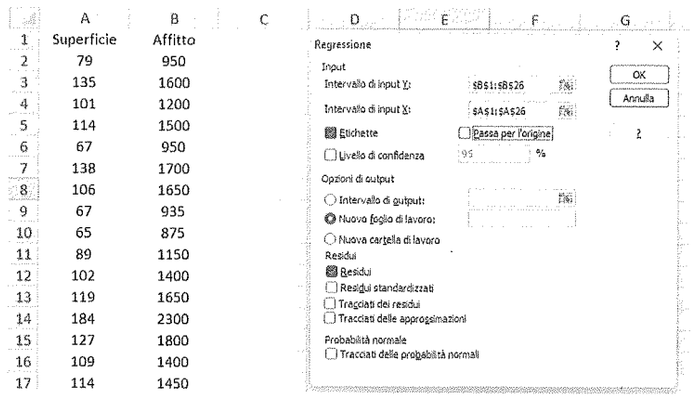


Figura 23.7
Finestra di dialogo della procedura di regressione in Excel.

	A	B
1	Superficie	Affitto
2	79	950
3	135	1600
4	101	1200
5	114	1500
6	67	950
7	138	1700
8	106	1650
9	67	935
10	65	875
11	89	1150
12	102	1400
13	119	1650
14	184	2300
15	127	1800
16	109	1400
17	114	1450

OUTPUT RIEPILOGO

Statistica della regressione	
R multiplo	0.848865683
R al quadrato	0.720572948
R al quadrato co	0.708423946
Errore standard	195.3062679
Osservazioni	25

ANALISI VARIANZA

	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	1	2262401.619	2262401.619	59.31128599	8.18639E-08
Residuo	23	877324.3806	38144.53829		
Totale	24	3139726			

	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Superiore 95%*	*Inferiore 95.0%*	*Superiore 95.0%*
Intercetta	176.8466557	161.8412057	1.092717117	0.28583175	-157.9473861	511.6406975	-157.9473861	511.6406975
Superficie	11.47148468	1.489535846	7.701382083	8.18639E-08	8.390145013	14.55282434	8.390145013	14.55282434

OUTPUT RESIDUI

Osservazione	*Previsto Affitto*	*Residui*
1	1083.093945	-133.0939451
2	1725.497087	-125.497087
3	1335.466608	-135.466608
4	1484.595909	15.40409117
5	945.436129	4.563870974
6	1759.911541	-59.91154107
7	1392.824031	257.1759686
8	945.436129	-10.43612903
9	922.4931597	-47.49315967
10	1197.808792	-47.80879191

Figura 23.8
Risultati dell'analisi di regressione.

CAPITOLO

24

Inferenza su popolazioni finite

Sommario del capitolo

24.1 Introduzione

24.2 Piani di campionamento, parametri e stimatori

24.3 Campionamento casuale semplice

24.4 Campionamento stratificato

24.5 Campionamento a due stadi semplici

24.6 Considerazioni finali

Riassunto

Commenti e curiosità

Esercizi

24.1 Introduzione

Questo capitolo è dedicato all'introduzione delle nozioni basilari del campionamento statistico propriamente detto, cioè quello che si realizza con un'effettiva operazione fisica di estrazione casuale di una parte (generalmente molto piccola) delle unità che compongono la popolazione reale oggetto di studio. La collocazione di questa materia nel capitolo terminale del volume non è il segno della sua marginalità, ma è dovuta al fatto che questa parte della statistica ha una connotazione specialistica, essendo applicabile esclusivamente alle popolazioni finite, cioè a quegli insiemi di unità statistiche costituite da entità (persone, famiglie, aziende, territori ecc.) individuabili e potenzialmente osservabili in modo esaustivo. Peraltro, il campionamento statistico è nato proprio come alternativa alle rilevazioni censuarie, che per definizione hanno come obiettivo l'osservazione completa di tutte le unità della popolazione. Si potrebbe dire che le popolazioni finite sono quelle per le quali è possibile ipotizzare una rilevazione censuaria. Per questo, il campionamento statistico non si applica agli studi sperimentali e osservazionali; l'ambito di applicazione è dunque quello delle indagini statistiche (si veda il Paragrafo 1.8.1).

Campi di applicazione privilegiata del campionamento sono gli ambiti socio-economici: indagini sociali, studi sulle aziende, sull'occupazione, analisi di mercato, sondaggi di opinione ecc. Nelle indagini sociali, la popolazione è costituita da tutte le famiglie o da tutti gli individui della zona geografica studiata (comune, regione, Stato ecc.); negli studi sulle aziende, la popolazione è costituita dalle aziende che operano nel settore o nei settori di interesse; nelle indagini di mercato la popolazione è formata dai potenziali consumatori di un prodotto o di un servizio; nei sondaggi di opinione le popolazioni di riferimento sono tutti gli individui appartenenti a determinate fasce d'età, per esempio i maggiorenni.

È opportuno segnalare subito che dell'apparato tecnico-concettuale dei Capitoli 17-23 verrà utilizzata la parte della stima puntuale e intervallare. Sarà assente la verifica delle ipotesi, per il fatto che questa metodologia è per sua natura congeniale agli studi basati sui dati provenienti da esperimenti.

L'indagine statistica mira all'ottenimento di informazioni sintetiche sulla popolazione. Così, in una indagine sulle aziende di un comprensorio, può interessare il fatturato totale, il totale degli addetti ecc. In una indagine di mercato, l'analisi può riguardare la percentuale delle famiglie che consumano un determinato prodotto, il consumo medio del prodotto

stesso ecc. In definitiva, si tende a conoscere le costanti caratteristiche della popolazione, costituite da totali, medie, percentuali e così via.

24.2 Piani di campionamento, parametri e stimatori

Sia $\mathcal{U}$ una popolazione finita composta da N unità,

$$\mathcal{U} = \{u_1, u_2, \ldots, u_i, \ldots, u_N\} = \{1, 2, \ldots, i, \ldots, N\},$$

dove, per semplicità, l'elemento i di $\mathcal{U}$, u_i, viene identificato con il numero i, chiamato etichetta, termine che allude agli elementi identificativi dell'unità statistica. L'etichetta identifica la singola unità della popolazione nel senso che consente di "localizzarla" e di osservarla, nel caso in cui essa sia inclusa nel campione. Per esempio, in una indagine sulle abitazioni di un dato comune, la numerazione implica che sussista un elenco di tutte le abitazioni oggetto di studio, dove la singola unità è identificata da un numero d'ordine seguito dall'ubicazione dell'abitazione (indirizzo, eventuale interno ecc.). L'elenco di tutte le unità della popolazione è denominato lista o base di campionamento.

Nelle indagini reali, le variabili sono decine, se non centinaia. Pensando al questionario dell'indagine, le variabili sono tante quante le domande che lo compongono. Nelle pagine che seguono, si farà riferimento a una sola variabile, ma è chiaro che il discorso varrà per ciascuna delle variabili rilevate. Sia dunque X la generica variabile d'interesse e siano $x_1, x_2, \ldots, x_N$ i valori non noti di X nelle N unità della popolazione.

Il problema inferenziale che verrà trattato principalmente in questo capitolo è quello della stima del parametro μ, media della popolazione $\mathcal{U}$:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

Altri parametri sono il totale,

$$\tau = N\mu = \sum_{i=1}^{N} x_i,$$

e la frequenza relativa o proporzione di una specifica caratteristica delle unità della popolazione, ad esempio, la proporzione delle famiglie di un Paese con redditi inferiori a un certo limite.

A proposito di costanti caratteristiche della popolazione, è necessario considerare la varianza, data da

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2,$$

che non sarà oggetto di inferenza, ma una sua trasformazione apparirà in alcune formule relative alla stima dei parametri sopra indicati. La trasformazione è così definita

$$\tilde{\sigma}^2 = \frac{N}{N-1}\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2.$$

La nozione di campione che viene qui introdotta è diversa da quella data nel Capitolo 17. Nel caso delle popolazioni finite, il campione, qualificato come probabilistico perché le unità che lo compongono vengono scelte con un meccanismo casuale, è così definito.

DEFINIZIONE 24.1

Un campione probabilistico di una popolazione finita è un sottoinsieme di unità estratte con uno specifico piano di campionamento, che assegna a ogni unità della popolazione una probabilità non nulla di essere selezionata.

Indicati con $\mathbb{C}$ e con $p(c)$, rispettivamente, l'insieme di tutti i campioni che si possono estrarre dalla popolazione con un dato piano di campionamento, denominato spazio dei campioni, e la probabilità associata al campione c, valgono le relazioni:

$$p(c) \geq 0, \text{ per ogni } c \in \mathbb{C}, \quad \sum_{c\in\mathbb{C}} p(c) = 1.$$

Dunque, il piano di campionamento assegna ai possibili campioni probabilità maggiori o uguali a zero, la cui somma è 1.

ESEMPIO 24.1

Data una popolazione di cinque unità, $\mathcal{U} = \{1, 2, 3, 4, 5\}$, i campioni possibili di ampiezza due sono:

$$c_1 = \{1, 2\},\ c_2 = \{1, 3\},\ c_3 = \{1, 4\},\ c_4 = \{1, 5\},\ c_5 = \{2, 3\},$$

$$c_6 = \{2, 4\},\ c_7 = \{2, 5\},\ c_8 = \{3, 4\},\ c_9 = \{3, 5\},\ c_{10} = \{4, 5\}.$$

Un piano di campionamento possibile è quello definito dalle probabilità:

$$p(c_1) = p(c_3) = p(c_5) = p(c_7) = p(c_9) = 0.08,$$

$$p(c_2) = p(c_4) = p(c_6) = p(c_8) = p(c_{10}) = 0.12.$$

Come si vede, sono soddisfatti i requisiti che caratterizzano un piano di campionamento: ai campioni di ampiezza due estraibili dalla popolazione data sono associate probabilità la cui somma è 1.

■

In termini generali, un campione probabilistico, c, di ampiezza n è l'insieme di etichette

$$c = \{i_1, i_2, \ldots, i_n\},$$

dove $i_1, i_2, \ldots, i_n$ sono le n etichette che identificano le unità della popolazione, $u_{i_1}, u_{i_2}, \ldots, u_{i_n}$, che fanno parte del campione. Nel seguito, si userà la lettera maiuscola C per indicare l'insieme casuale delle etichette campionarie $\{i_1, i_2, \ldots, i_n\}$ prima che il campione sia estratto, e, come in precedenza, la lettera minuscola c per il singolo campione, una volta che esso è stato estratto.

Ai fini dell'inferenza su μ, le etichette hanno il compito di veicolare, oltreché gli elementi identificativi delle unità campionarie, le modalità che la variabile X assume nelle unità incluse nel campione. I valori di X nelle unità campionarie sono allora:

$$x_{i_1}, x_{i_2}, \ldots, x_{i_n},$$

dove x_{i_1} è la modalità di X nell'unità i_1, x_{i_2} è la modalità di X nell'unità i_2 e così via.

I piani di campionamento di maggiore rilievo concettuale e pratico sono quelli indicati qui di seguito.

- Campionamento casuale semplice. Le unità del campione sono selezionate con un meccanismo aleatorio tale da assicurare a tutte le unità della popolazione la stessa probabilità di essere incluse nel campione. Costruire un campione casuale semplice di ampiezza n di una popolazione finita equivale idealmente a estrarre a sorte n palline da un'urna contenente N palline recanti i numeri da 1 a N che individuano le unità della popolazione (l'estrazione va intesa senza ripetizione, cioè senza rimettere la pallina nell'urna dopo ogni estrazione).

- Campionamento stratificato. Supponendo di avere informazioni aggiuntive sulla popolazione, in base a queste si suddivide la popolazione in sottopopolazioni denominate strati, ognuno dei quali contiene unità tra loro omogenee in base a qualche criterio; poi, da ciascuno strato, si estrae un campione casuale di opportuna ampiezza.

- Campionamento a grappoli. Supponendo che la popolazione sia suddivisa in un certo numero di sottoinsiemi chiamati grappoli, si prende un campione casuale di tali grappoli e si inseriscono nel campione tutte le unità appartenenti ai grappoli prescelti.

- Campionamento a due stadi. Può essere visto come un campionamento a grappoli in cui, anziché rilevare la totalità delle unità all'interno del singolo grappolo selezionato, se ne prende un campione casuale. In questo modo, si hanno due stadi di campionamento: nel primo stadio si estraggono i grappoli, nel secondo le unità statistiche oggetto di studio.

Nella Figura 24.1 è contenuta una semplice schematizzazione grafica dei campioni: casuale semplice, stratificato, a grappoli e a due stadi.

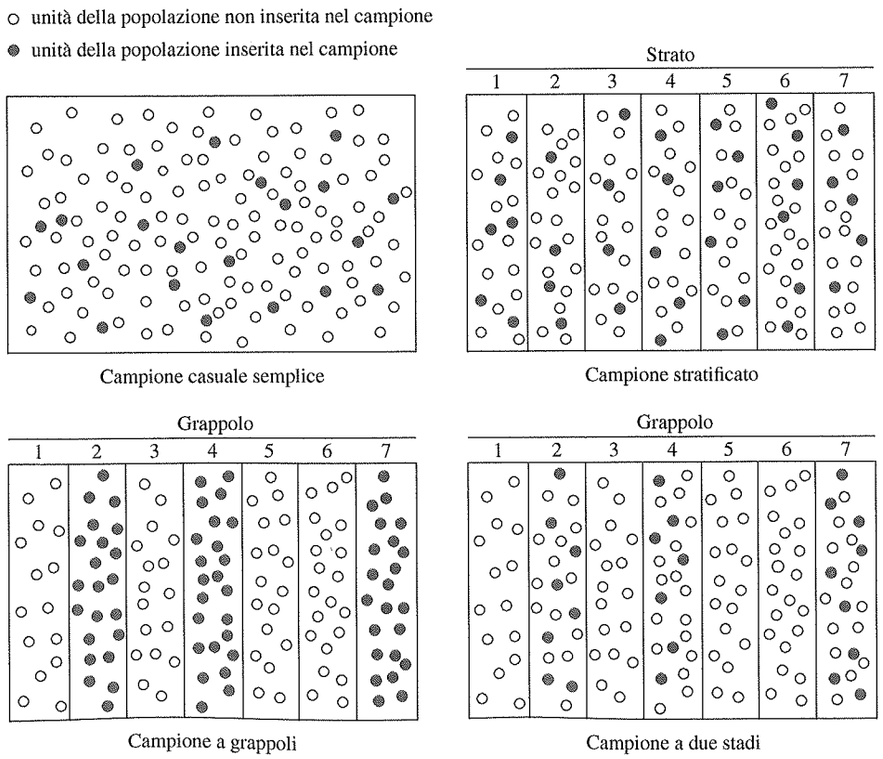


Figura 24.1
Schematizzazione grafica dei campioni: casuale semplice, stratificato, a grappoli e a due stadi.

Nelle pagine che seguono verranno illustrati: il campionamento casuale semplice, il campionamento stratificato e il campionamento a due stadi nel caso particolare in cui sia nel primo sia nel secondo stadio il campionamento è casuale semplice.

Come si vede, il meccanismo di estrazione del campione è più o meno complesso: la complessità è connessa, da una parte, a fattori organizzativi e, dall'altra, alla opportunità di tener conto di informazioni *a priori* sulla popolazione per la costruzione di un campione che la "rappresenti al meglio".

Per fattori organizzativi si intendono le questioni che attengono alla natura della lista disponibile, alle modalità, ai costi, ai tempi e alle tecniche di rilevazione. Per esempio, in un'indagine di mercato sulle preferenze delle famiglie riguardo a un determinato prodotto, non è pensabile di estrarre direttamente un campione dalle liste delle famiglie italiane: una tale lista non esiste, esistono invece liste a livello di singolo comune presso l'ufficio anagrafe; è ragionevole allora formare il campione in due stadi, scegliendo prima un campione di comuni, e poi estraendo da ciascun comune selezionato un campione di famiglie. Anche se teoricamente possibile, un campione di famiglie estratto a livello nazionale sarebbe sconsigliabile per altre ragioni: le famiglie così estratte sarebbero "disperse" su tutto il territorio nazionale (alcune a Torino, altre a Roma, altre a Cagliari ecc.), e ciò, se le interviste fossero faccia a faccia, darebbe luogo a una lievitazione dei costi di rilevazione (spese di viaggio dei rilevatori, ampliamento della rete di rilevazione ecc.).

Quanto alle informazioni *a priori* sulla popolazione, dette anche informazioni ausiliarie, si tratta di conoscenze utili a livello di selezione delle unità o nella fase di stima. Nell'esempio dell'indagine di mercato, potrebbe essere interessante raggruppare i comuni per dimensioni demografiche, distinguendoli in comuni di grandi, medie e piccole dimensioni, per poi estrarre a sorte un certo numero di comuni per ciascuna tipologia. Il vantaggio che ne deriva è chiaro: ciascuna tipologia di comuni sarebbe rappresentata nel campione, cosa che potrebbe non realizzarsi, con conseguenze negative sulla rappresentatività, se si procedesse all'estrazione senza tale accorgimento.

Sia θ un generico parametro della popolazione (ad esempio, la media aritmetica, il totale della variabile, la frequenza relativa delle unità che presentano una data caratteristica). Siano $T(C)$ e $T(c)$, rispettivamente, uno stimatore di θ e la corrispondente stima basata sul campione osservato c. Il valore atteso di $T(C)$ è dato da

$$\mathrm{E}[T(C)] = \sum_{c \in \mathbb{C}} T(c)p(c).$$

Questa scrittura è resa possibile dal fatto che $T(C)$ è una variabile casuale discreta, essendo finito il numero dei campioni in $\mathbb{C}$.

La nozione di non distorsione, già introdotta nel Capitolo 18, resta valida in questo nuovo contesto. Con la simbologia adottata, uno stimatore è non distorto se vale la relazione

$$\mathrm{E}[T(C)] = \theta, \text{ per ogni valore di } \theta.$$

La distorsione è data dalla differenza $\mathrm{E}[T(C)] - \theta$ ed è nulla se lo stimatore è non distorto.

L'errore quadratico medio dello stimatore è dato da

$$\mathrm{MSE}[T(C)] = \mathrm{E}[T(C) - \theta]^2 = \sum_{c \in \mathbb{C}} [T(c) - \theta]^2 p(c)$$

e, come nel caso dei dati provenienti da esperimenti (Capitolo 18), può essere scritto come somma della varianza di $T(C)$ e del quadrato della distorsione:

$$\text{MSE}[T(C)] = \text{Var}[T(C)] + \{\text{E}[T(C)] - \theta\}^2,$$

dove $\text{Var}[T(C)]$ indica la varianza dello stimatore $T(C)$.

L'errore quadratico medio o la varianza, se lo stimatore è non distorto, è una misura della precisione dello stimatore. Tra due stimatori $T_1(C)$ e $T_2(C)$, quello che presenta l'errore quadratico medio più piccolo è detto più efficiente.

Per la stima della media della popolazione, si farà riferimento allo stimatore non distorto per eccellenza, noto come stimatore di Horvitz-Thompson, dato da[1]

$$\hat{\mu} = \frac{1}{N} \sum_{i \in c} \frac{x_i}{\pi_i} \tag{24.1}$$

dove π_i è la cosiddetta probabilità di inclusione, cioè la probabilità che l'unità i della popolazione ha di essere inserita nel campione.

Lo stimatore (24.1) ha un chiaro significato intuitivo. Infatti, potendosi scrivere

$$\sum_{i \in c} \frac{x_i}{\pi_i} = \sum_{i \in c} x_i \cdot \frac{1}{\pi_i},$$

si vede che la singola osservazione campionaria viene moltiplicata per il "coefficiente di espansione" $1/\pi_i$, maggiore di 1, il cui ordine di grandezza medio è N/n, che può essere pensato come il numero di unità della popolazione rappresentate dall'unità i del campione. La somma di questi prodotti è uno stimatore del totale τ; questa somma, divisa N, è quindi lo stimatore della media μ.

Si può dimostrare facilmente che lo stimatore (24.1) è non distorto e che uno stimatore non distorto della sua variabilità di campionamento, $\hat{\sigma}^2_{\hat{\mu}}$, è dato da

$$\hat{\sigma}^2_{\hat{\mu}} = \frac{1}{N^2} \sum_{i \in c} \sum_{j \in c} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{x_i}{\pi_i} \frac{x_j}{\pi_j}, \tag{24.2}$$

dove π_{ij} è la probabilità di inclusione della coppia di unità i e j. Le dimostrazioni sono contenute nel Paragrafo B.11 dell'Appendice B.

Sulla base degli stimatori $\hat{\mu}$ e $\hat{\sigma}^2_{\hat{\mu}}$, è poi possibile costruire, sotto certe condizioni, uno stimatore intervallare per μ per un assegnato coefficiente di fiducia $1 - \alpha$. Gli estremi dell'intervallo sono dati da:

$$l_1 = \hat{\mu} - z_{1-\alpha/2}\hat{\sigma}_{\hat{\mu}}, \quad l_2 = \hat{\mu} + z_{1-\alpha/2}\hat{\sigma}_{\hat{\mu}}.$$

Dopo queste premesse generali, nei prossimi tre paragrafi verrà affrontato il problema della stima della media per i piani di campionamento casuale semplice, stratificato e a due stadi. Per ciascuno di questi piani, verranno presentati gli stimatori $\hat{\mu}$ e $\hat{\sigma}^2_{\hat{\mu}}$. Non verrà invece presentata la variabilità di campionamento $\sigma^2_{\hat{\mu}}$, per ragioni di brevità, ma anche perché non è indispensabile nell'economia di questo capitolo: invero, essa è utile per i confronti di efficienza tra piani di campionamento diversi o tra stimatori alternativi.

24.3 Campionamento casuale semplice

È il piano di campionamento basilare: trova applicazione sia direttamente sia come componente nell'ambito di schemi di campionamento più complessi. Con questo piano di campionamento le n unità vengono estratte a sorte, una alla volta, senza ripetizione e con probabilità uguali. Ciò significa che:

- dopo ogni estrazione, l'unità selezionata viene rimossa dalla popolazione (estrazione senza ripetizione);
- alla estrazione i-esima, alle unità restanti della popolazione vengono attribuite probabilità di estrazione uguali pari a $1/(N - i + 1)$, $i = 1, 2, \ldots, n$ (estrazione con probabilità uguali).

Un tale meccanismo di selezione può essere concepito pensando a un'urna contenente N palline uguali recanti le N etichette che contrassegnano le unità della popolazione da cui vengono estratte a sorte successivamente n palline. In pratica, l'estrazione di un campione siffatto può essere effettuata con il procedimento descritto qui di seguito:

- siano $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ N numeri casuali indipendenti appartenenti all'intervallo (0,1), dove ε_i è associato all'unità i della popolazione. Tali numeri vengono poi ordinati in ordine crescente;
- le n unità statistiche che corrispondono agli n valori più piccoli tra $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ costituiscono un campione casuale semplice.

Quando la posizione delle unità nella lista può essere considerata casuale, una modalità di estrazione, che sotto l'aspetto applicativo dà risultati analoghi a quelli del campionamento casuale semplice, è il campionamento sistematico. Sia p la parte intera del rapporto N/n; si estragga un numero casuale minore o uguale a p; sia r il numero casuale estratto; allora il campione sistematico è dato dall'insieme di unità identificate dalle etichette:

$$r,\ r + p,\ r + 2p,\ \ldots,\ r + (n - 1)p.$$

ESEMPIO 24.2

Si estragga un campione casuale semplice di $n = 20$ unità dalla popolazione dei comuni della Svezia descritta nel *file* MU284 (Särndal *et al.*, 1992).

Soluzione

Il campione casuale semplice ottenuto applicando il procedimento descritto sopra è individuato dal seguente insieme di etichette

{259, 182, 211, 177, 201, 204, 72, 63, 274, 168, 47, 37, 181, 141, 265, 157, 64, 94, 82, 246}.

Dunque, formano il campione le unità della popolazione individuate dai numeri 259, 182, ..., 246. ■

Nella proposizione che segue viene data l'espressione dello stimatore non distorto di μ insieme allo stimatore non distorto della sua variabilità di campionamento.

PROPOSIZIONE 24.1

Si consideri un campione casuale semplice di ampiezza n di una popolazione finita di N unità. Lo stimatore non distorto della media della popolazione è dato da

$$\hat{\mu}_{ccs} = \frac{1}{n} \sum_{i \in c} x_i = \bar{x}. \tag{24.3}$$

[1] L'operatore $\sum_{i \in c} q_i$ indica la somma delle quantità q_i con i che prende i valori indicati nell'insieme c. Per esempio, se $c = \{2, 30, 7, 204\}$, $\sum_{i \in c} q_i = q_2 + q_{30} + q_7 + q_{204}$.

Lo stimatore non distorto della sua variabilità di campionamento è dato da

$$\hat{\sigma}_{\hat{\mu}_{ccs}} = (1-f)\frac{s^2}{n}, \tag{24.4}$$

dove $f = n/N$ e

$$s^2 = \frac{1}{n-1}\sum_{i\in c}(x_i - \bar{x})^2$$

è la varianza campionaria.

Lo stimatore (24.3) si deduce immediatamente dalla (24.1) sapendo che nel campionamento casuale semplice la probabilità di inclusione dell'unità i della popolazione è

$$\pi_i = \frac{n}{N},\ i = 1, 2, \ldots, N$$

(la dimostrazione è contenuta nel Paragrafo B.11 dell'Appendice B). La (24.4) si ottiene dalla (24.2), come illustrato in dettaglio nel Paragrafo B.11 dell'Appendice B. ■

La quantità $1 - f$, che appare nella (24.4), è denominata correzione per popolazione finita; la quantità f è denominata tasso di sondaggio.

Dalla (24.3) è immediato ricavare lo stimatore non distorto $\hat{\tau}$ del totale τ. Basta moltiplicare per N lo stimatore di μ:

$$\hat{\tau}_{ccs} = N\bar{x}.$$

Analogamente, dalla (24.4), moltiplicando per N^2, si deduce lo stimatore non distorto della variabilità di campionamento di $\hat{\tau}_{ccs}$.

L'intervallo di confidenza per μ, per un prefissato coefficiente di fiducia, ha estremi dati da:

$$l_1 = \bar{x} - z_{1-\alpha/2}\sqrt{1-f}\frac{s}{\sqrt{n}},\ l_2 = \bar{x} + z_{1-\alpha/2}\sqrt{1-f}\frac{s}{\sqrt{n}},$$

dove s è la deviazione standard del campione osservato.

ESEMPIO 24.3

Riprendendo il campione casuale semplice dell'Esempio 24.2, si costruisca l'intervallo di confidenza al 95% per la media della variabile SS82 (seggi del partito socialdemocratico nei consigli comunali).

Soluzione

Passando dalle etichette ai valori della variabile SS82, si ricava la serie di osservazioni $15, 18, \ldots, 37$. La media e la varianza campionaria sono pari a 23.6 e 98.7. Gli estremi dell'intervallo di confidenza sono allora:

$$l_1 = 23.6 - 1.96\sqrt{\frac{284-20}{284}}\sqrt{\frac{98.7}{20}},\ l_2 = 23.6 + 1.96\sqrt{\frac{284-20}{284}}\sqrt{\frac{98.7}{20}}.$$

In definitiva, l'intervallo di confidenza è (19.4, 27.8). ■

Dimensione campionaria

Il problema della determinazione della dimensione del campione – già affrontato nel caso di dati provenienti da esperimenti (Paragrafo 18.4) – coinvolge la variabilità di campionamento dello stimatore, che è espressa da

$$\sigma^2_{\hat{\mu}_{ccs}} = (1-f)\frac{\tilde{\sigma}^2}{n}$$

(si veda l'Esercizio 24.2). Indicato con d il limite massimo prefissato per l'errore di stima (margine di errore), assumendo l'approssimata normalità della distribuzione dello stimatore $\hat{\mu}_{ccs}$, si può scrivere

$$P(|\hat{\mu}_{ccs} - \mu| < d) \approx 1 - \alpha,$$

da cui si evince che deve essere

$$d = z_{1-\alpha/2}\sqrt{\frac{N-n}{N}\frac{\tilde{\sigma}^2}{n}}.$$

Risolvendo questa equazione rispetto a n, dopo alcuni passaggi algebrici, si ottiene

$$n = \frac{Nz^2_{1-\alpha/2}\tilde{\sigma}^2}{Nd^2 + z^2_{1-\alpha/2}\tilde{\sigma}^2}. \tag{24.5}$$

L'applicazione della formula presuppone che la varianza σ^2 sia nota o che si disponga di una stima preliminare di essa (stima basata, per esempio, su una precedente indagine campionaria sulla stessa popolazione o su popolazioni simili).

24.3.1 Popolazioni Bernoulliane

Un caso particolarmente importante per le applicazioni è quello in cui la variabile X è qualitativa dicotomica, cioè presenta due sole modalità A e $\bar{A}$, presente e assente. Si pensi, ad esempio, allo stato occupazionale delle persone (occupato, non occupato), alla presenza o assenza nelle unità statistiche di una data caratteristica (consumatore o non consumatore di un prodotto, possessore o non possessore di un dato titolo di studio).

In situazioni come queste, si assegnano a X due valori: 1 per indicare la modalità A o la presenza della caratteristica considerata e 0 per indicare la modalità $\bar{A}$ o l'assenza della caratteristica. In tal modo, la media della popolazione è la frequenza relativa della modalità A (si veda anche il punto 6 di "Commenti e curiosità" del Capitolo 4), quantità che nel seguito verrà indicata con il simbolo p. Con lo stesso ragionamento, si trova che la media del campione casuale semplice è la frequenza relativa della modalità A nel campione, che viene indicata con $\hat{p}$. A sua volta, la varianza campionaria può essere così riscritta

$$s^2 = \frac{n}{n-1}\hat{p}(1-\hat{p})$$

(si veda l'Esercizio 17.22*).

Ne segue che le (24.3) e (24.4) divengono:

$$\hat{\mu}_{ccs} = \hat{p},\ \hat{\sigma}^2_{\hat{\mu}_{ccs}} = \hat{\sigma}^2_{\hat{p}} = (1-f)\frac{\hat{p}(1-\hat{p})}{n-1}.$$

Tutto ciò si riflette naturalmente sulle espressioni degli estremi dell'intervallo di confidenza per p, che assumono la forma

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{(1-f)\frac{\hat{p}(1-\hat{p})}{n-1}}.$$

Per quanto riguarda la dimensione campionaria che assicura un margine di errore al di sotto di un certo limite con alta probabilità, sostituendo a $\tilde{\sigma}^2$ la quantità $Np(1-p)/(N-1)$, la (24.5) assume la forma

$$n = \frac{Nz^2_{1-\alpha/2}p(1-p)}{(N-1)d^2 + z^2_{1-\alpha/2}p(1-p)}. \tag{24.6}$$

Questa espressione dipende dal parametro incognito p. Tuttavia, considerando che il massimo di $p(1-p)$ è raggiunto per $p = 0.5$, per cui il valore di n che si ricava dalla (24.6) ponendo $p = 0.5$, cioè

$$n = \frac{Nz^2_{1-\alpha/2} \cdot 0.25}{(N-1)d^2 + z^2_{1-\alpha/2} \cdot 0.25}, \tag{24.7}$$

è adeguato nel caso in cui p sia esattamente pari a 0.5 e più grande del dovuto per ogni altro valore di p.

ESEMPIO 24.4

Da un lotto di 5100 unità di prodotto semilavorato viene estratto un campione di 200 pezzi per la valutazione del tasso di difettosità. L'ispezione del campione rivela la presenza di 5 pezzi difettosi.

Soluzione

L'intervallo di confidenza al 95% per il tasso di difettosità dell'intero lotto è dato da

$$0.025 \pm 1.96\sqrt{(4900/5100)[(0.025 \cdot 0.975)/199]};$$

l'intervallo di confidenza è dunque $(0.0037, 0.0463)$. ■

ESEMPIO 24.5

Con riferimento al problema precedente, si determini la dimensione del campione tale che l'errore di stima del tasso di difettosità non superi la soglia di 0.02 con probabilità 0.95.

Soluzione

La soluzione si ottiene immediatamente dall'applicazione della (24.7):

$$n = \frac{5100 \cdot 1.96^2 \cdot 0.25}{5099 \cdot 0.02^2 + 1.96^2 \cdot 0.25} = 1633.$$

■

È opportuno precisare che la stima intervallare della media e del totale può essere condotta con il metodo indicato nella Proposizione 19.3, se le unità del campione vengono estratte con ripetizione. In questo caso, vale infatti la definizione di campione casuale, data nel Paragrafo 17.2, come n-upla di variabili casuali indipendenti distribuite come la variabile casuale della popolazione, con tutto quel che ne segue sullo stimatore non distorto di μ, $\bar{X}$, sulla sua varianza, σ^2/n, e sullo stimatore non distorto di quest'ultima. In definitiva, l'intervallo di confidenza ha estremi dati da

$$\bar{x} \pm z_{1-\alpha/2}\frac{s}{\sqrt{n}}.$$

Come si vede, rispetto all'espressione data nella pagina precedente, scompare il fattore $\sqrt{1-f}$. Naturalmente, ciò si riflette anche sull'espressione delle formule della dimensione campionaria che assicura un dato margine di errore: in luogo delle (24.5) e (24.7), vanno applicate la (18.5) e il suo adattamento per il caso della popolazione Bernoulliana.

24.4 Campionamento stratificato

Il campionamento stratificato è caratterizzato dal fatto che la popolazione viene suddivisa preliminarmente in sottopopolazioni omogenee (strati), da ciascuna delle quali si estrae un campione casuale semplice.

L'omogeneità negli strati va intesa in senso diverso a seconda delle finalità che si vogliono perseguire con la stratificazione. Ad esempio, la stratificazione geografica in un'indagine su una certa area può rispondere a esigenze di tipo organizzativo, nel senso che i singoli strati possono rappresentare porzioni dell'area totale affidate per la rilevazione e il controllo a organismi distinti. Il singolo strato poi potrebbe delimitare uno specifico dominio di studio, cioè una sottopopolazione di cui si vogliono stimare i parametri: una stratificazione delle città in un'indagine a carattere socio-economico in città di grandi, medie e piccole dimensioni può tra l'altro essere giustificata dall'opportunità di ottenere risultati distinti per i tre tipi di comune che configurano così differenti domini di studio.

Ma la motivazione che sta alla base della stratificazione è che essa consente un miglioramento della precisione delle stime: in questo caso, l'omogeneità va intesa come ridotta variabilità della variabile in studio nello strato, quale può essere misurata, ad esempio, dalla deviazione standard. Si intuisce infatti che una forte somiglianza tra le unità di ciascuno strato fa sì che a livello di strato il campione rappresenti meglio le caratteristiche della strato, ne consegue che dal campione stratificato, cioè dall'unione di campioni di strato si attendono stime più accurate rispetto a quelle prodotte da un campione non stratificato della stessa dimensione.

Il problema è quello di come formare strati omogenei senza conoscere la variabile rispetto alla quale l'omogeneità si deve realizzare. Evidentemente, occorrerà basarsi su una o più caratteristiche conosciute per tutte le unità della popolazione che presentino una forte correlazione con la variabile X oggetto di studio. Così, per esempio, in una indagine sulla produzione industriale in un determinato settore produttivo, una variabile utile per la stratificazione potrebbe essere la dimensione delle aziende, in termini di addetti, ammesso che l'informazione sul numero di addetti sia disponibile per tutte le aziende della popolazione considerata.

E ancora, in una ricerca di mercato sul gradimento di un nuovo genere di consumo di lusso da parte delle famiglie italiane è ragionevole stratificare le famiglie secondo il livello economico e il settore di attività del capofamiglia in quanto è plausibile che queste caratteristiche influiscano sul comportamento al consumo. Poiché tali informazioni non si possiedono a livello di singola famiglia, una stratificazione del genere potrebbe essere attuata a livello di comune (per i comuni sono disponibili indicatori di benessere, nonché i tassi di attività per settore produttivo rilevati con i censimenti). In tal caso, gli strati sarebbero formati dalle famiglie residenti nei comuni omogenei per le due caratteristiche di stratificazione.

In generale, se la stratificazione è realizzata sulla base di r variabili, il numero di strati in cui è suddivisa la popolazione è dato dal prodotto di r fattori, ciascuno dei quali corrisponde al numero di modalità presenti nella variabile a cui si riferisce. Ad esempio, se le variabili di stratificazione sono due, di cui la prima presenta 4 modalità (o classi) e la seconda 3 modalità (o classi), il numero di strati sarà $4 \cdot 3 = 12$.

Si indichi con:

- N il numero di unità della popolazione;
- L il numero di strati;
- N_h, $h = 1, 2, \ldots, L$, il numero di unità nello strato h;
- n_h, $h = 1, 2, \ldots, L$, la dimensione del campione casuale semplice dello strato h;
- $f_h = n_h/N_h$, $h = 1, 2, \ldots, L$, il tasso di sondaggio dello strato h.

Valgono ovviamente le relazioni:

$$\sum_{h=1}^{L} N_h = N, \quad \sum_{h=1}^{L} n_h = n.$$

Il campionamento stratificato consiste nell'estrazione di un campione casuale semplice da ciascuno strato.

PROPOSIZIONE 24.2

Si consideri un campione stratificato di ampiezza n di una popolazione di ampiezza N suddivisa in L strati di dimensioni $N_1, N_2, \ldots, N_L$. Sia $\bar{x}_h$ la media aritmetica del campione dello strato h. Allora lo stimatore non distorto della media della popolazione è dato da

$$\hat{\mu}_{str} = \sum_{h=1}^{L} \frac{N_h}{N} \bar{x}_h. \tag{24.8}$$

Lo stimatore non distorto della variabilità di campionamento di $\hat{\mu}_{str}$ è dato da

$$\hat{\sigma}^2_{\hat{\mu}_{str}} = \sum_{h=1}^{L} \frac{N_h^2}{N^2}(1 - f_h)\frac{s_h^2}{n_h}, \tag{24.9}$$

dove s_h^2 è la varianza del campione dello strato h.

La (24.8) si deduce facilmente dalla (24.1) considerando che l'unità k della popolazione che appartiene allo strato h ha probabilità di inclusione pari a

$$\pi_k = \frac{n_h}{N_h}.$$

La dimostrazione della (24.9) è contenuta nel Paragrafo B.11 dell'Appendice B. ■

Se anziché la media, si vuole stimare il totale τ della popolazione, le formule pertinenti si ottengono dalla (24.8) moltiplicando per N e dalla (24.9) moltiplicando per N^2.

Stima intervallare

Se la dimensione dei campioni di strato è sufficientemente elevata, la distribuzione di probabilità di $\hat{\mu}_{str}$ può essere approssimata con la normale. L'intervallo di confidenza per μ, con un coefficiente di fiducia pari a $1 - \alpha$, è allora dato da

$$\hat{\mu}_{str} \pm z_{1-\alpha/2}\hat{\sigma}_{\hat{\mu}_{str}}.$$

24.4.1 Allocazione delle unità negli strati

Finora, le dimensioni dei campioni di strato n_h sono state considerate come quantità assegnate. In realtà, chi pianifica il campione deve anche decidere, oltreché la dimensione complessiva n, le dimensioni dei campioni di strato. Questo problema va sotto il nome di allocazione del campione negli strati.

Data la dimensione complessiva n, due semplici soluzioni del problema sono l'allocazione uniforme e l'allocazione proporzionale. Con il primo criterio, la dimensione totale n viene ripartita tra gli strati in parti uguali (o quasi uguali se n non è multiplo di L); con il secondo, la suddivisione di n avviene in proporzione alla dimensione degli strati, ponendo

$$n_h = n\frac{N_h}{N}, \quad h = 1, 2, \ldots, L.$$

Una soluzione più efficiente del problema consiste nell'assegnare a $n_1, n_2, \ldots, n_L$ i valori che minimizzano la variabilità di campionamento $\sigma^2_{\hat{\mu}_{str}}$.

PROPOSIZIONE 24.3

Le dimensioni campionarie di strato che minimizzano la variabilità di campionamento $\hat{\mu}_{str}$ sono espresse da

$$n_h \approx n\frac{N_h\tilde{\sigma}_h}{\sum_{j=1}^{L} N_j\tilde{\sigma}_j}, \quad h = 1, 2, \ldots, L, \tag{24.10}$$

dove $\tilde{\sigma}_j$ è la deviazione standard di X nello strato j.

Per la dimostrazione si veda Conti e Marella (2012), p. 132. ■

Quella descritta dalla formula (24.10) va sotto il nome di allocazione di Neyman. La formula si presta a una chiara interpretazione: la dimensione del campione di strato è proporzionale al livello della variabilità dello strato (misurata dalla deviazione standard $\tilde{\sigma}_h$) moltiplicato per la dimensione dello strato stesso. Il criterio tiene quindi conto, a un tempo, del grado di omogeneità all'interno degli strati e dell'ampiezza degli stessi. L'impiego della (24.10) implica che si conoscano le quantità $\tilde{\sigma}_h$, $h = 1, 2, \ldots, L$, o che si disponga di stime preliminari.

Dimensione globale del campione

Il problema della determinazione del campione coinvolge la variabilità di campionamento dello stimatore $\hat{\mu}_{str}$, che è espressa da

$$\sigma^2_{\hat{\mu}_{str}} = \sum_{h=1}^{L} \frac{N_h^2}{N^2}(1 - f_h)\frac{\tilde{\sigma}_h^2}{n_h},$$

dove $\tilde{\sigma}_h^2$ è la varianza della variabile di interesse nello strato h (si veda l'Esercizio 24.5).

La determinazione della dimensione globale del campione implica che si sia risolto il problema dell'allocazione delle unità campionarie tra gli strati. Qui viene data la formula che fornisce l'ampiezza del campione che garantisce un errore di stima $|\hat{\mu}_{str} - \mu| < d$ a un livello di probabilità $1 - \alpha$ prestabilito. Operando come

nel campionamento casuale semplice, dalla relazione

$$d = z_{1-\alpha/2}\sigma_{\hat{\mu}_{str}} = z_{1-\alpha/2}\left(\sum_{h=1}^{L}\frac{N_h^2}{N^2}\,\frac{N_h - n_h}{N_h}\,\frac{\tilde{\sigma}_h^2}{n_h}\right)^{1/2},$$

fissate le quantità $w_h = n_h/n$, si ha

$$N^2 d^2 = z_{1-\alpha/2}^2 \sum_{h=1}^{L} N_h^2 \frac{N_h - nw_h}{N_h}\,\frac{\tilde{\sigma}_h^2}{nw_h};$$

da cui, risolvendo rispetto a n, si ricava

$$n = \frac{z_{1-\alpha/2}^2 \sum_{h=1}^{L} N_h^2 \tilde{\sigma}_h^2 / w_h}{N^2 d^2 + z_{1-\alpha/2}^2 \sum_{h=1}^{L} N_h \tilde{\sigma}_h^2}.$$

ESEMPIO 24.6

Si stimino μ e la variabilità di campionamento per la variabile SS82 (seggi del partito socialdemocratico nei consigli comunali) della popolazione dei comuni della Svezia (*file* MU284). Si utilizzi un campione stratificato di ampiezza 50, adottando come variabile di stratificazione la regione (variabile Reg) e come criterio di allocazione delle unità campionarie quello proporzionale.

Soluzione

Gli strati hanno le seguenti dimensioni: $N_1 = 25$, $N_2 = 48$, $N_3 = 32$, $N_4 = 38$, $N_5 = 56$, $N_6 = 41$, $N_7 = 15$, $N_8 = 29$. Con l'allocazione proporzionale, le ampiezze dei campioni sono: $n_1 = 4$, $n_2 = 8$, $n_3 = 6$, $n_4 = 7$, $n_5 = 10$, $n_6 = 7$, $n_7 = 3$, $n_8 = 5$. I campioni casuali semplici estratti dagli strati sono risultati: $c_1 = \{17, 5, 24, 6\}$, $c_2 = \{48, 19, 16, 40, 14, 22, 43, 25\}$, $c_3 = \{31, 11, 6, 26, 19, 5\}$, $c_4 = \{35, 5, 30, 14, 31, 20, 37\}$, $c_5 = \{1, 31, 6, 46, 37, 14, 50, 44, 36, 43\}$, $c_6 = \{21, 26, 10, 25, 30, 35, 39\}$, $c_7 = \{13, 5, 1\}$, $c_8 = \{17, 25, 22, 9, 13\}$. Le medie e le varianze di strato sono: $\bar{x}_1 = 23.25$, $\bar{x}_2 = 24$, $\bar{x}_3 = 23.83$, $\bar{x}_4 = 19.43$, $\bar{x}_5 = 19.4$, $\bar{x}_6 = 22.29$, $\bar{x}_7 = 26.33$, $\bar{x}_8 = 23.4$; $s_1^2 = 66.92$, $s_2^2 = 100.57$, $s_3^2 = 86.97$, $s_4^2 = 55.62$, $s_5^2 = 55.60$, $s_6^2 = 40.24$, $s_7^2 = 57.33$, $s_8^2 = 66.3$.

Sulla base di questi dati, l'applicazione della (24.8) dà la stima del numero medio dei seggi del partito socialdemocratico nei consigli comunali della Svezia:

$$\hat{\mu}_{str} = \frac{25}{284}23.25 + \frac{48}{284}24 + \cdots + \frac{29}{284}23.4 = 22.21.$$

La stima della variabilità di campionamento si ottiene applicando la (24.9):

$$\hat{\sigma}^2_{\hat{\mu}_{str}} = \left(\frac{25}{284}\right)^2\left(1 - \frac{4}{25}\right)\frac{66.92}{4} + \left(\frac{48}{284}\right)^2\left(1 - \frac{8}{48}\right)\frac{100.57}{8} + \cdots$$
$$+ \left(\frac{29}{284}\right)^2\left(1 - \frac{5}{29}\right)\frac{66.30}{5} = 1.11.$$

■

24.5 Campionamento a due stadi semplici

Come si è già accennato in precedenza, nel campionamento a due stadi, le unità elementari, ossia quelle su cui viene realizzata la rilevazione dei dati, vengono estratte in due fasi: nella prima si estrae un campione di unità di primo stadio (costituite da grappoli di unità elementari), nella seconda, da ciascuna unità di primo stadio viene estratto un campione di unità elementari. Qui si illustrerà il caso in cui i campioni di primo e secondo stadio sono entrambi casuali semplici.

Il campionamento a due stadi viene in genere adottato non perché fornisce stime più efficienti rispetto ad altri piani di campionamento, ma per ragioni di tipo economico-organizzativo. Invero, le unità elementari all'interno delle unità di primo stadio tendono ad assomigliarsi. Si pensi alle famiglie, ai comuni, ai condomini: le unità elementari cioè le persone (nelle famiglie, nei comuni, nei condomini), sono caratterizzate da una contiguità spaziale o di altro genere che può determinare una somiglianza di condizioni economico-sociali, di visioni e di comportamenti, deleteria per la precisione delle stime, perché può implicare una forte somiglianza anche per la variabile di interesse. Può allora avvenire che il campione di secondo stadio sia costituito da unità così omogenee da ledere la qualità delle informazioni che da esse si traggono. Al limite, se le unità elementari fossero perfettamente omogenee (per quanto riguarda la variabile d'interesse), un campione di 100 unità potrebbe valere quanto un campione unitario (perché, ad esempio, tutte le persone potrebbero dare la stessa risposta a una data domanda di una ipotetica intervista).[2]

Al di là di questa intuizione, si può dimostrare che, a parità di numerosità campionaria, il campionamento a due stadi è meno efficiente del campionamento casuale semplice, e che il divario in efficienza è funzione crescente del grado di somiglianza tra le unità elementari all'interno delle unità di primo stadio.

I motivi economico-organizzativi che rendono più conveniente il ricorso al campionamento a due stadi sono diversi. Basti pensare, ad esempio, che, ove non esista una lista delle unità finali da cui estrarre il campione, bisognerebbe costruirla con costi economici e rischi di non elevata accuratezza. Nel campionamento a due stadi, questa attività potrebbe essere limitata alle unità di primo stadio inserite nel campione. Inoltre, se l'intervista è faccia a faccia, il campione a due stadi consente un risparmio di risorse, perché i rilevatori si devono muovere solo entro i confini dell'unità di primo stadio. Naturalmente, questo vantaggio scompare quando la rilevazione avviene tramite telefono o via *web*. Vi sono situazioni in cui il campionamento a due stadi è più efficiente rispetto ad altri piani, a parità di costi per la realizzazione dell'indagine.

[2] È interessante osservare come la somiglianza delle unità negli strati sia un punto di forza nel campionamento stratificato, e come la somiglianza all'interno delle unità di primo stadio sia un punto di debolezza nel campionamento a due stadi. Questa apparente antinomia può essere così risolta. Nel campionamento stratificato, gli strati (salvo il caso della stratificazione su base geografica) sono dei raggruppamenti artificiali di unità elementari fatti in modo che vi sia una forte omogeneità interna e una forte disomogeneità tra gli strati. Poiché nel campionamento stratificato si prende un campione di unità da ciascuno strato, il campione globale che ne risulta "rappresenta meglio" l'intera popolazione, della quale coglie tutte le specificità. All'opposto, l'omogeneità all'interno delle unità di primo stadio milita contro l'efficienza delle stime. Ciò per due motivi, in primo luogo, le unità di primo stadio sono generalmente dei raggruppamenti naturali che non necessariamente presentano forte disomogeneità tra di loro. In secondo luogo, perché solo una parte delle unità di primo stadio contribuisce alla formazione del campione finale. Il rischio è allora quello di avere un campione di unità elementari composto in parte da unità molto somiglianti che può non rispecchiare alcuni aspetti della popolazione.

Si indichi con:

- M il numero di unità di primo stadio della popolazione;
- m il numero di unità di primo stadio del campione;
- $f_1 = m/M$ il tasso di sondaggio di primo stadio;
- N_i il numero di unità elementari dell'unità di primo stadio i della popolazione;
- n_i l'ampiezza del campione di unità elementari relativo all'unità di primo stadio i della popolazione;
- $f_{2i} = n_i/N_i$ il tasso di sondaggio di secondo stadio per l'unità di primo stadio i;
- $N = \sum_{i=1}^{M} N_i$ il numero di unità elementari della popolazione.

Nella proposizione che segue, vengono date le espressioni dello stimatore della media della popolazione, $\hat{\mu}_{2st}$, e del relativo stimatore della variabilità di campionamento, $\hat{\sigma}^2_{2st}$.

PROPOSIZIONE 24.4

Sia c_1 un campione casuale semplice di m unità di primo stadio. Sia c_{2i} un campione casuale semplice di ampiezza n_i dell'unità di primo stadio i contenente N_i unità elementari. Indicata con $\bar{x}_i$ la media campionaria di c_{2i}, lo stimatore non distorto della media della popolazione è dato da

$$\hat{\mu}_{2st} = \frac{1}{N}\frac{M}{m}\sum_{i\in c_1} N_i \bar{x}_i. \tag{24.11}$$

Lo stimatore non distorto della variabilità di campionamento è dato da

$$\hat{\sigma}^2_{\hat{\mu}_{2st}} = \frac{1}{N^2}\left[M^2\frac{1-f_1}{m}\frac{1}{m-1}\sum_{i\in c_1}(\hat{\tau}_i - \hat{\bar{\tau}})^2 + \frac{M}{m}\sum_{i\in c_1}(1-f_{2i})N_i^2\frac{s^2_{2i}}{n_i}\right], \tag{24.12}$$

dove

$$\hat{\tau}_i = N_i\bar{x}_i,\ i \in c_1,$$

sono le stime dei totali delle m unità di primo stadio nel campione,

$$\hat{\bar{\tau}} = \frac{1}{m}\sum_{i\in c_1}\hat{\tau}_i$$

e s^2_{2i} è la varianza del campione di unità elementari dell'unità di primo stadio i.

La (24.11) si deduce facilmente dalla (24.1) tenendo presente che l'unità elementare k appartenente all'unità di primo stadio i ha una probabilità di inclusione pari a

$$\pi_k = \frac{m}{M}\frac{n_i}{N_i}.$$

Per la dimostrazione della (24.12) si rinvia a Särndal (1992), pp. 138-142. ■

ESEMPIO 24.7

Si stimino la media e la variabilità di campionamento per la variabile SS82 (seggi del partito socialdemocratico nei consigli comunali) della popolazione dei comuni della Svezia (*file* MU284), con un campione a due stadi formato da un campione casuale semplice di primo stadio, di ampiezza $m = 4$, estratto dalle 50 unità di primo stadio in cui è suddivisa la popolazione (si tratta di gruppi di comuni contigui individuati dalla variabile CL), e da un campione casuale semplice di secondo stadio, di ampiezza $n_i = 3$, estratto da ciascuna unità selezionata al primo stadio. Si supponga che il campione di primo stadio sia $\{4, 47, 22, 10\}$ e che le osservazioni dei campioni di secondo stadio siano, rispettivamente: $(41, 23, 27)$, $(17, 14, 20)$, $(15, 32, 19)$, $(11, 20, 38)$.

Soluzione

La stima della media dei seggi del partito socialdemocratico si ricava dall'applicazione della (24.12), in cui compaiono le medie dei campioni di secondo stadio che sono: $30.33, 17, 22, 23$ e le dimensioni delle unità di primo stadio che sono: $5, 5, 6, 5$. Si ha allora

$$\hat{\mu}_{2st} = \frac{1}{284}\frac{50}{4}(5\cdot 30.33 + 5\cdot 17 + 6\cdot 22 + 5\cdot 23) = 21.3.$$

La stima della variabilità di campionamento si ottiene dall'applicazione della (24.12). Nella formula compaiono le stime dei totali delle unità di primo stadio nel campione, pari a: $5\cdot 30.33 = 151.65$, $5\cdot 17 = 85$, $6\cdot 22 = 132$, $5\cdot 23 = 115$, e le varianze dei campioni di secondo stadio, che sono: 89.33, 9, 79, 189. Poiché

$$\sum_{i\in c_1}(\hat{\tau}_i - \hat{\bar{\tau}})^2 = (151.65 - 120.91)^2 + (85 - 120.91)^2 + \\ +(132 - 120.91)^2 + (115 - 120.91)^2 = 2392.39$$

e

$$\sum_{i\in c_1}(1-f_{2i})N_i^2\frac{s^2_{2i}}{n_i} = \left(1-\frac{3}{5}\right)25\frac{89.33}{3} + \left(1-\frac{3}{5}\right)25\frac{9}{3} + \left(1-\frac{3}{6}\right)36\frac{79}{3} + \\ +\left(1-\frac{3}{5}\right)25\frac{189}{3} = 1431.77,$$

si ha in definitiva

$$\hat{\sigma}^2_{\hat{\mu}_{2st}} = \frac{1}{284^2}\left[\frac{50^2}{4}\left(1-\frac{4}{50}\right)\frac{2392.39}{3} + \frac{50}{4}1431.77\right] = 5.91.$$

■

24.6 Considerazioni finali

Come già accennato in precedenza, l'intendimento di questo capitolo è quello di fornire alcune basi concettuali del vastissimo tema dell'inferenza su popolazioni finite, la cui trattazione esaustiva richiederebbe un intero volume. Sono state date le formule esplicite degli stimatori di Horvitz-Thompson della media, con i relativi stimatori della variabilità di campionamento, per i tre piani di campionamento più rilevanti, casuale semplice, stratificato e a due stadi semplici, cercando di spiegare le ragioni logiche che sono alla base della loro adozione. È opportuno ora delineare brevemente come e in quali direzioni questo nucleo conoscitivo iniziale può essere arricchito e sviluppato. Va subito chiarito che l'espansione del quadro conoscitivo può avvenire sia dal lato degli stimatori sia dal lato dei piani di campionamento.

Ad esempio, per la stima della media in un campione casuale semplice, anziché lo stimatore $\hat{\mu}_{ccs}$ qui illustrato, potrebbe essere usato il cosiddetto stimatore per quoziente dato dal prodotto di $\hat{\mu}_{ccs}$ per un fattore correttivo costituito dal rapporto tra la media nota di una variabile ausiliaria e lo stimatore di Horvitz-Thompson della media di tale variabile. Se la correlazione tra la variabile ausiliaria e la variabile oggetto di studio è elevata, questo stimatore è più efficiente perché il coefficiente correttivo tende a essere maggiore di 1 se $\hat{\mu}_{ccs}$ sottostima μ e minore di 1 nel caso opposto.

Un'altra via per ottenere uno stimatore più efficiente è quella di "rettificare" lo stimatore $\hat{\mu}_{ccs}$ aggiungendo ad esso la differenza tra la media nota di una variabile ausiliaria e la media stimata (con lo stimatore di Horvitz-Thompson) della stessa variabile moltiplicata per il coefficiente di regressione tra la variabile X e la variabile ausiliaria. Lo stimatore che ne risulta è il cosiddetto stimatore per regressione.

Questi due stimatori sono approssimativamente non distorti e più efficienti di $\hat{\mu}_{ccs}$, sotto condizioni abbastanza generali. Inoltre, essi possono essere abbinati anche al campionamento stratificato, nel senso che è possibile definire uno stimatore per quoziente o per regressione all'interno di un piano di campionamento stratificato, beneficiando in tal modo di un doppio miglioramento della precisione delle stime: quello indotto dalla stratificazione e quello ascrivibile allo stimatore.

Dal lato dei piani di campionamento, direttrici di sviluppo riguardano in particolare il campionamento a due stadi, che può diventare a tre o più stadi. Ad esempio, in un'indagine sugli studenti della scuola superiore, è possibile configurare un campionamento a tre stadi, in cui le unità di primo stadio sono le scuole, le unità di secondo stadio le classi, e le unità di terzo stadio, cioè elementari, gli studenti nelle classi. In queste situazioni, occorrerà definire gli appropriati stimatori della media o di altri parametri con i rispettivi stimatori della variabilità campionaria.

Nella pratica delle indagini campionarie, il campionamento a due o più stadi può giovarsi di tecniche di miglioramento della precisione delle stime legate all'uso di probabilità di inclusione variabili. Nel campionamento a due stadi in cui le unità di primo stadio sono i comuni, la dimensione demografica dei comuni è una variabile ausiliaria particolarmente importante, perché un notevole miglioramento della precisione delle stime può essere tratto dall'assegnazione ai comuni di una probabilità di inclusione proporzionale alla loro dimensione demografica. Ciò viene fatto, ad esempio, nell'indagine sulle forze di lavoro: l'Istat usa un piano di campionamento a due stadi, in cui le unità di primo stadio sono i comuni e le unità di secondo stadio le famiglie. I comuni vengono distinti in due categorie, quelli che hanno una popolazione superiore a un dato limite, detti "auto-rappresentativi" e tutti gli altri. I primi vengono inseriti nel campione con probabilità 1 e ciascuno forma uno strato a sé stante. I secondi vengono stratificati all'interno della singola provincia in base alla dimensione campionaria e da ciascuno strato viene estratto un comune con probabilità proporzionale alla dimensione.

Come notazione generale, va osservato che la struttura del piano di campionamento è determinata dalla natura e dagli obbiettivi della singola indagine reale. Quindi non vi possono essere soluzioni standard stabilite *a priori*. Ad esempio, nelle grandi indagini campionarie, tipicamente condotte dagli uffici nazionali di statistica, in cui le variabili rilevate sono molto numerose, i piani di campionamento sono quasi sempre complessi, nel senso che la loro architettura è la risultante di una molteplicità di procedure campionarie. Inoltre, il campione è strutturato in modo che si possano ottenere stime valide a livello territoriale sub-nazionale, ad esempio a livello di regione. Molto spesso vi sono due o più stadi di campionamento; quasi sempre le unità di primo stadio vengono stratificate, le unità elementari vengono talora scelte con campionamento sistematico se vi sono le condizioni per cui questo criterio possa essere equiparato al campionamento casuale semplice. E anche i metodi di stima possono variare con la natura delle variabili.

La precisione delle stime campionarie a cui si è fatto riferimento in questo capitolo è intesa come misura della variabilità campionaria del singolo stimatore intorno al valore vero del parametro: quanto minore è la variabilità campionaria dello stimatore tanto maggiore è la sua precisione. Tale misura di variabilità può essere vista come errore medio che si commette assegnando al parametro incognito il valore assunto dallo stimatore nel campione. Si tratta dell'errore campionario, così chiamato perché causato dal fatto che la stima del parametro viene fatta con una sola parte (il campione) delle unità della popolazione. Ne segue che queste misure non danno conto delle altre possibili fonti di variabilità delle stime campionarie, come l'imperfezione della lista, le mancate risposte, gli errori di misura e così via, fonti che producono i cosiddetti errori non campionari di cui si è parlato in generale nel Capitolo 2, a cui si rinvia.

Il protagonista di questo capitolo è il campione probabilistico, che, nella teoria classica del campionamento da popolazioni finite, è il presupposto necessario per fare inferenza sui parametri della popolazione. Come già detto nel Paragrafo 1.8.1, non sempre le indagini campionarie reali sono condotte in base a campioni probabilistici. Molte agenzie private che operano nel settore delle indagini di mercato e demoscopiche si avvalgono di campioni a scelta ragionata, come il campionamento per quote. Per questo e per altri tipi di campione a scelta ragionata, si rinvia al Paragrafo 1.8.1.

Riassunto

In questo capitolo, viene trattato in modo sintetico un argomento di natura specialistica, l'inferenza su popolazioni finite, intendendo con questa espressione quegli insiemi di unità statistiche costituite da entità (persone, famiglie, aziende, territori ecc.) individuabili e potenzialmente osservabili in modo esaustivo. La peculiarità di questo tema sta nel fatto che nel caso delle popolazioni finite il campione di osservazioni non è la risultante della "ripetizione di un esperimento nelle stesse condizioni", ma si ottiene osservando le unità di un campione estratto fisicamente dalla popolazione con meccanismi casuali. È pertanto diversa la fonte da cui trae origine la distribuzione campionaria delle statistiche.

Di questo tema amplissimo il capitolo tratta alcuni aspetti essenziali. Vengono innanzitutto date le definizioni di campione probabilistico, di piano di campionamento, di stimatore e di errore quadratico medio, mettendo in luce le affinità e le difformità rispetto agli analoghi aspetti esaminati nei capitoli precedenti. Viene definito lo stimatore non distorto per eccellenza, lo stimatore di Horvitz-Thompson, del parametro media. A ciò fa seguito la presentazione di tre specifici piani di campionamento, ritenuti essenziali, il campionamento casuale semplice, il campionamento stratificato e quello a due stadi, mettendo in rilevo le ragioni che ne guidano l'adozione. Per ciascuno di essi vengono poi dati lo stimatore non distorto della media e il relativo stimatore della precisione delle stime.

Parole chiave

allocazione del campione negli strati, 525
allocazione di Neyman, 525
allocazione proporzionale, 525
allocazione uniforme, 525
campionamento a due stadi, 516
campionamento a grappoli, 516
campionamento casuale semplice, 515
campionamento sistematico, 519
campionamento stratificato, 516
campione probabilistico, 515

correzione per popolazione finita, 520
dominio di studio, 523
etichetta, 514
informazioni ausiliarie, 517
piano di campionamento, 515
popolazione finita, 513
probabilità di inclusione, 518
probabilità uguali, 519
spazio dei campioni, 515
stimatore di Horvitz-Thompson, 518
stimatore per quoziente, 530
stimatore per regressione, 530
strati, 516
tasso di sondaggio, 520
tasso di sondaggio di primo stadio, 528
tasso di sondaggio di secondo stadio, 528
unità di primo stadio, 527
unità elementari, 527
variabilità di campionamento, 518

Commenti e curiosità

1. Il metodo rappresentativo di Kiaer

La questione della validità delle indagini campionarie fu affrontata per la prima volta in una sede ufficiale a Berna nel 1895, nella V riunione dell'Istituto internazionale di statistica (International Statistical Institute, ISI in acronimo), un'associazione scientifica internazionale tra accademici del settore statistico, i responsabili degli Uffici nazionali di statistica e gli studiosi ed esperti di indagini statistiche. In quell'occasione, Anders Nicolai Kiaer (1838-1919), fondatore e primo Direttore dell'Ufficio nazionale di statistica della Norvegia, presentò una relazione in cui sosteneva come non fosse necessario procedere a una rilevazione totale, a un censimento, per avere utili informazioni sull'intera popolazione. Alla luce di una pluriennale esperienza di "indagini parziali" condotte dal suo Ufficio – tra cui quella sull'opinione dei cittadini su una proposta governativa di istituzione della pensione di vecchiaia – argomentò come per ottenere tali informazioni fosse sufficiente procedere a un'indagine parziale basata su un sottoinsieme di unità della popolazione convenientemente scelte in modo che riproducessero in miniatura l'intera popolazione. Il processo di selezione delle unità di questo sottoinsieme fu da lui denominato "metodo rappresentativo". Questa idea era stata avanzata e poi abbandonata da uno dei più importanti statistici del XIX secolo, Adolphe Quetelet (si veda Lusinchi, 2022, p. 176).

La rappresentatività del campione veniva quindi intesa come somiglianza tra campione e popolazione rispetto ad alcune variabili osservate per l'intera popolazione nel più recente censimento, come la residenza (città o campagna), il sesso, l'età ecc., intendendo per somiglianza, il fatto che il tasso di urbanizzazione, la proporzione dei maschi, quella delle femmine ecc. nel campione fossero uguali o molto vicini a quelli dell'intera popolazione.

Degli otto studiosi che intervennero nella discussione della riunione di Berna solo tre espressero qualche interesse per la proposta di Kiaer, sostenendo l'opportunità che la discussione venisse ripresa nelle riunioni successive. Prevalsero le critiche dai toni talora molto aspri. L'argomentazione principale fu che le rilevazioni parziali avessero valore solo con riferimento all'ambito ristretto in cui venivano realizzate e che non fosse possibile "sostituire con calcoli" l'effettiva osservazione dei fenomeni. Le indagini sembrarono in effetti ai più come un pericoloso arretramento rispetto alle rilevazioni totali che costituivano un paradigma a quel tempo e che venivano via via sempre meglio organizzate per lo sviluppo degli Istituti nazionali di statistica. Al termine della discussione, a stretta maggioranza, passò la proposta di porre la questione all'ordine del giorno delle riunioni a venire.

Kiaer continuò a sostenere la sua idea nelle successive riunioni scientifiche, in particolare in quelle di Budapest del 1901 e di Berlino del 1903, in cui il metodo rappresentativo fu di nuovo oggetto di dibattito esplicito. Nella prima, si registrò un affievolimento dell'opposizione: vi furono solo due interventi critici, anche se in uno di essi si sostenne che il metodo rappresentativo fosse una metodologia obsoleta. Peraltro, i responsabili di alcuni Uffici nazionali di statistica (del Belgio, della Danimarca e della Svizzera) avevano nel frattempo applicato il metodo rappresentativo, sia pure con risultati alterni.

Fu a Berlino nel 1903 che la vicenda del metodo rappresentativo ebbe alfine il suo felice compimento. Una commissione precedentemente costituita propose la seguente risoluzione: "La Commissione, considerando che la corretta applicazione del metodo rappresentativo in un certo numero di casi può fornire esatte e dettagliate informazioni i cui risultati possono essere generalizzati entro certi limiti, ne raccomanda l'uso a patto che nella pubblicazione dei risultati siano completamente specificate le condizioni sotto le quali è avvenuta la selezione delle unità osservate".

Dunque, occorse quasi un decennio di dibattiti e studi perché il metodo campionario potesse essere accettato come strumento di studio dei fenomeni economico-sociali. Dovranno passare altri tre decenni perché il metodo campionario potesse assumere una fisionomia somigliante a quella attuale, acquisendo i crismi della scientificità con l'uso del calcolo delle probabilità.

2. Verso una teoria dei campioni

Se Kiaer ebbe il merito di imporre all'attenzione degli studiosi il tema delle rilevazioni parziali, si deve a Arthur Lyon Bowley (1869-1957) la formulazione di una prima "teoria dei campioni". Considerando il campionamento casuale semplice e servendosi di una particolare versione del teorema del limite centrale, fornì un metodo per valutare la precisione delle stime ottenute con campioni di dimensione elevata estratti da ampie popolazioni.

Ma molti problemi rimanevano ancora aperti: ad esempio, lo schema di Bowley non era applicabile a molteplici situazioni concrete, quali quelle che richiedono più stadi di selezione (per esempio, estrazione prima dei comuni campione e poi, all'interno di questi, delle unità elementari), ovvero la selezione di grappoli di unità elementari. Nel frattempo, in mancanza di una teoria organica, le rilevazioni parziali seguivano vie e soluzioni empiriche. La situazione di stallo durò per circa un trentennio: il punto di svolta si ebbe nel 1934 con il contributo di Jerzy Neyman, di cui si parlerà tra breve.

Nel 1924, l'ISI nominò una commissione con il compito di studiare le applicazioni del metodo rappresentativo. Nella relazione che ne seguì venne delineata una completa rassegna dei metodi usati e della teoria statistica a essi sottesa. In relazione al meccanismo di scelta utilizzato, venne fatta la distinzione fondamentale tra campionamento probabilistico e campionamento a scelta ragionata: il primo inteso nel senso che a tutte le unità della popolazione viene assegnata la stessa probabilità di selezione, il secondo fondato su un criterio di scelta delle unità che, muovendo da informazioni su alcune caratteristiche della popolazione indagata, tende a riprodurre in miniatura la popolazione stessa.

Va notato che: la Commissione pose sullo stesso piano i due metodi di campionamento; fu prevista per entrambi i metodi la possibilità di valutare con i dati del campione la precisione delle stime campionarie; non fu previsto alcunché circa la selezione casuale di grappoli di unità elementari.

Il contributo di Neyman allo sviluppo della teoria dei campioni è contenuto in un articolo presentato alla *Royal Statistical Society* nel 1934 (si veda Neyman, 1934). I punti salienti dello schema di Neyman si possono così compendiare:

- i campioni casuali di unità elementari o di grappoli di unità elementari sono "misurabili", nel senso che consentono di stimare il grado di precisione delle stime;
- il numero delle unità campionarie deve essere elevato;
- è possibile utilizzare informazioni ausiliarie sulla popolazione nella costruzione del piano di campionamento, senza far venir meno la natura casuale del campione (l'autore si riferisce sostanzialmente alla possibilità della stratificazione);
- un metodo di campionamento di generale validità è il campionamento stratificato;
- vi sono situazioni specifiche in cui il campionamento a scelta ragionata può essere usato con successo.

3. I grandi campioni non sono necessariamente garanzia di buone stime: un caso di scuola

La qualità dell'informazione che si trae da un'indagine campionaria non è determinata dall'ampiezza del campione. Certo, per un fissato piano di campionamento e a parità di altre condizioni (tecniche di intervista, qualità della rilevazione ecc.), la precisione delle stime tende a crescere con l'ampiezza del campione. È chiaro però che un grande campione può dare luogo a stime meno precise rispetto a un campione di dimensione molto più piccola, ma caratterizzato da un piano di campionamento e da metodi di rilevazione più accurati. Al riguardo, va ricordato il caso classico dell'indagine condotta dalla rivista americana *Literary Digest* per la previsione del risultato delle elezioni presidenziali degli Stati Uniti nel 1936.

In quella tornata elettorale, i candidati alla presidenza erano Franklin D. Roosevelt e Alfred M. Landon. La rivista pronosticò per Landon il 57% dei voti, mentre in realtà ottenne solo il 36%. Fu un grande

smacco per una rivista molto popolare, grazie anche alla reputazione che le proveniva dall'aver previsto correttamente i risultati di tre precedenti elezioni presidenziali. Per questo sondaggio, la rivista aveva inviato un questionario postale a 10 milioni di persone, i cui nominativi erano stati presi dagli elenchi telefonici, dai registri automobilistici e da elenchi di iscritti ad associazioni varie. Risposero circa 2.4 milioni di persone.

La domanda è come possa essere spiegato un errore di previsione così clamoroso con un campione di tale dimensione. L'opinione prevalente ha ricondotto l'evento a due concause: la scarsa rappresentatività degli iscritti al voto della lista delle persone utilizzata per l'invio del questionario e, soprattutto, l'alta percentuale di non risposte (76%). Il primo fattore può aver determinato la sovra-rappresentazione nel campione delle persone più abbienti (abbonati al telefono, possessori di automobili ecc.), dato che a quell'epoca era prevalentemente la classe più abbiente ad avere il telefono o a possedere un'automobile, cioè la classe tendenzialmente più favorevole al partito repubblicano. Quanto al secondo fattore, è ragionevole pensare che a rispondere al questionario siano state prevalentemente le persone di orientamento repubblicano, più motivate a schierarsi contro la riconferma di un presidente democratico, che, per di più, aveva governato in anni di grave depressione. Il divario tra ampiezza del campione e povertà del risultato si delineò con maggiore nettezza anche perché per le stesse elezioni un sondaggio dell'Agenzia Gallup, basato su un campione di rispondenti (anch'esso non probabilistico, ma per quote) di circa 50 000 iscritti al voto, prodisse correttamente la vittoria di Roosevelt.

Esercizi

24.1 Si consideri la popolazione degli studenti universitari riportata nel *file* "Studenti universitari". Si determini la stima del voto medio finale con un campione casuale semplice di dimensione 20, supponendo di aver estratto il seguente campione di etichette:

$$\{98, 128, 39, 131, 88, 99, 7, 89, 106, 15, 46, 12, 101, 71, 80, 57, 62, 65, 105, 102\}.$$

Si determini anche l'intervallo di confidenza al 95%.

24.2* Si dimostri che la variabilità di campionamento dello stimatore $\hat{\mu}_{ccs}$ è data da

$$\sigma^2_{\hat{\mu}_{ccs}} = (1 - f)\frac{\tilde{\sigma}^2}{n}.$$

24.3 In un campione casuale semplice di 30 famiglie estratte da una popolazione di 15 000 unità, sono stati rilevati il numero di componenti della famiglia (X_1) e il numero delle persone che hanno consultato il medico nel mese precedente a quello della rilevazione (X_2). I dati osservati sono riportati nella tabella che segue:

N.	X_1	X_2	N.	X_1	X_2	N.	X_1	X_2
1	5	5	11	3	1	21	4	2
2	6	0	12	2	1	22	3	0
3	3	2	13	7	0	23	3	0
4	3	3	14	4	4	24	1	0
5	2	0	15	3	1	25	2	2
6	3	0	16	5	2	26	4	2
7	3	0	17	4	0	27	3	0
8	3	0	18	4	0	28	4	2
9	4	0	19	3	1	29	2	0
10	4	0	20	3	3	30	4	1

a) Si stimi il numero delle famiglie con almeno 4 componenti.

b) Si stimino il numero medio di persone che hanno consultato il medico e la relativa variabilità di campionamento.

24.4 Con riferimento alla popolazione di cui all'Esercizio 24.1, si estragga un campione stratificato di ampiezza $n = 25$.

a) Si stratifichi in base al sesso con l'allocazione di Neyman, sapendo che la popolazione è composta da 72 maschi e 61 femmine e che la deviazione standard ($\tilde{\sigma}$) della variabile di interesse è 3.27 per i maschi e 2.94 per le femmine.

b) Si stratifichi congiuntamente in base al sesso e al voto ottenuto nella prova preliminare, suddividendo tale variabile nelle tre classi: [18 – 22], [23 – 26] e [27 – 30]; si adotti poi l'allocazione proporzionale.

c) In entrambi i casi, si proceda alla stima della media del voto finale e della variabilità di campionamento.

24.5* Si dimostri che la variabilità di campionamento dello stimatore $\hat{\mu}_{str}$ è data da

$$\sigma^2_{\hat{\mu}_{str}} = \sum_{h=1}^{L} \frac{N_h^2}{N^2}(1 - f_h)\frac{\tilde{\sigma}_h^2}{n_h},$$

dove $\tilde{\sigma}_h^2$ è la varianza della variabile di interesse nello strato h.

24.6 Si consideri la popolazione dei comuni della provincia di Roma (*file* "Comuni provincia di Roma"). Escludendo il comune di Roma (etichetta 97), si estragga un campione stratificato di ampiezza 20, suddividendo i comuni in tre strati in base alla popolazione residente come segue. Primo strato: comuni con popolazione fino a 1000 residenti; secondo strato: comuni con un numero di abitanti compreso tra 1001 e 10 000; terzo strato: comuni con popolazione superiore a 10 000. Adottando l'allocazione proporzionale, si stimino il totale delle abitazioni e la relativa variabilità di campionamento.

24.7 Nello scaffale di uno studente universitario vi sono 390 libri. Per ottenere la stima del totale delle parole in essi contenute, si usi un campione a due stadi, in cui le unità di primo stadio sono i libri e le unità elementari sono le singole pagine. Si prendano un campione casuale semplice di $m = 5$ libri e un campione di unità elementari di 3 pagine per libro. Si proceda alla stima del totale e della variabilità di campionamento. Si assuma che il campione selezionato abbia dato i risultati indicati nella tabella che segue:

Libro selezionato	Numero di pagine	Pagine selezionate	Numero di parole
81	178	36	429
		82	444
		173	350
186	128	15	294
		94	360
		117	378
195	278	48	305
		61	350
		212	338
288	243	62	343
		99	320
		111	432
387	191	8	295
		71	340
		178	374

CAPITOLO

25

Regressione lineare multipla

ommario del capitolo

5.1 Introduzione

5.2 Il modello di regressione lineare multipla

5.3 Stima dei parametri del modello con il metodo dei minimi quadrati

5.4 Stima dei parametri del modello con il metodo della massima verosimiglianza

5.5 Inferenza sul modello di regressione multipla

5.6 Controllo di validità delle ipotesi di base tramite l'analisi dei residui

Il Capitolo 25 è disponibile sulla piattaforma Pearson MyLab.

CAPITOLO

26

Analisi della varianza

Sommario del capitolo

26.1 Introduzione

26.2 Studi sperimentali e studi di osservazione a un solo fattore

26.3 Studi sperimentali e studi di osservazione a due fattori

Il Capitolo 26 è disponibile sulla piattaforma Pearson MyLab.

APPENDICE

A

Statistica descrittiva: dimostrazioni

Sommario dell'appendice

A.1 L'operatore di sommatoria
A.2 Medie
A.3 Concentrazione
A.4 Forma delle distribuzioni di frequenza
A.5 Disuguaglianza di Chebyshev
A.6 Analisi della dipendenza nelle tabelle di contingenza
A.7 Regressione lineare semplice
A.8 Correlazione

A.1 L'operatore di sommatoria

Date le quantità $x_1, x_2, \ldots, x_n$, la loro somma viene sinteticamente così indicata

$$x_1 + x_2 + \cdots + x_n = \sum_{i=1}^{n} x_i.$$

L'espressione a destra dell'uguale si legge nel modo seguente: "sommatoria di x_i per i che varia da 1 a n". L'indice i della sommatoria può naturalmente essere sostituito con qualsiasi altra lettera:

$$\sum_{i=1}^{n} x_i = \sum_{h=1}^{n} x_h = \sum_{j=1}^{n} x_j = \cdots$$

Gli estremi del suo campo di variazione possono essere omessi quando risultano chiari dal contesto di applicazione.

Seguono degli esempi:

$$\sum_{i=1}^{5} (x_i - 2)^2 = (x_1 - 2)^2 + (x_2 - 2)^2 + (x_3 - 2)^2 + (x_4 - 2)^2 + (x_5 - 2)^2,$$

$$\sum_{i=1}^{4} x_i n_i = x_1 n_1 + x_2 n_2 + x_3 n_3 + x_4 n_4,$$

$$\sum_{i=0}^{3} x^i = x^0 + x^1 + x^2 + x^3,$$

$$\sum_{i=1}^{2} i^2 = 1^2 + 2^2.$$

Di seguito vengono indicate alcune proprietà elementari:

$$\sum_{i=1}^{n} c = \underbrace{c + c + \cdots + c}_{n \text{ volte}} = nc,$$

$$\sum_{i=1}^{n} cx_i = cx_1 + cx_2 + \cdots + cx_n = c \sum_{i=1}^{n} x_i,$$

$$\sum_{i=1}^{n} (x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + \cdots + (x_n + y_n) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.$$

Quando gli elementi da sommare dipendono da due indici, si utilizza l'operatore di sommatoria doppia:

$$x_{11} + x_{12} + \cdots + x_{1n} + x_{21} + x_{22} + \cdots + x_{2n} + \cdots + x_{mn} = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}.$$

A.2 Medie

A.2.1 Media aritmetica

Proprietà (a) della Proposizione 4.1

Siano $x_{(1)}$ e $x_{(n)}$ il valore più piccolo e quello più grande della serie di osservazioni $x_1, x_2, \ldots, x_n$. È evidente che $x_{(1)} \leq x_i \leq x_{(n)}$, da cui seguono le relazioni di equivalenza:

$$\sum_{i=1}^{n} x_{(1)} \leq \sum_{i=1}^{n} x_i \leq \sum_{i=1}^{n} x_{(n)} \Leftrightarrow nx_{(1)} \leq \sum_{i=1}^{n} x_i \leq nx_{(n)}$$

$$\Leftrightarrow x_{(1)} \leq \frac{1}{n} \sum_{i=1}^{n} x_i \leq x_{(n)} \Leftrightarrow x_{(1)} \leq \mu \leq x_{(n)}.$$

La proprietà è così dimostrata.

Proprietà (b) della Proposizione 4.1

Questa proprietà può essere dedotta direttamente dalla definizione di media aritmetica:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i \Leftrightarrow n \cdot \mu = \sum_{i=1}^{n} x_i.$$

Proprietà (c) della Proposizione 4.1

Si può scrivere

$$\sum_{i=1}^{n} (x_i - \mu) = \sum_{i=1}^{n} x_i - n\mu = 0,$$

dove è stata utilizzata la proprietà (b) di μ, per la quale $\sum_{i=1}^{n} x_i = n\mu$. La proprietà è così dimostrata.

Proprietà (d) della Proposizione 4.1

Si vuole provare la validità della disuguaglianza

$$\sum_{i=1}^{n} (x_i - \mu)^2 \leq \sum_{i=1}^{n} (x_i - c)^2,$$

dove c è una costante qualsiasi. Conviene partire dall'espressione

$$\sum_{i=1}^{n} (x_i - c)^2 = \sum_{i=1}^{n} [(x_i - \mu) - (c - \mu)]^2,$$

da cui, sviluppando il quadrato del binomio dentro le parentesi quadre, si ottiene

$$\begin{aligned} \sum_{i=1}^{n} (x_i - c)^2 &= \sum_{i=1}^{n} [(x_i - \mu)^2 + (c - \mu)^2 - 2(x_i - \mu)(c - \mu)] \\ &= \sum_{i=1}^{n} (x_i - \mu)^2 + \sum_{i=1}^{n} (c - \mu)^2 - 2 \sum_{i=1}^{n} (x_i - \mu)(c - \mu). \end{aligned}$$

Poiché

$$\sum_{i=1}^{n} (c - \mu)^2 = n(c - \mu)^2$$

e

$$\sum_{i=1}^{n} (x_i - \mu)(c - \mu) = (c - \mu) \sum_{i=1}^{n} (x_i - \mu) = 0,$$

si può scrivere

$$\sum_{i=1}^{n} (x_i - c)^2 = \sum_{i=1}^{n} (x_i - \mu)^2 + n(c - \mu)^2 \geq \sum_{i=1}^{n} (x_i - \mu)^2,$$

essendo $n(c - \mu)^2 \geq 0$. La Proprietà è così dimostrata.

Proprietà (e) della Proposizione 4.1

Si indichi con μ_X la media aritmetica della serie $x_1, x_2, \ldots, x_n$, e con μ_Y la media aritmetica della serie

$$y_1 = a + bx_1, y_2 = a + bx_2, \ldots, y_n = a + bx_n,$$

ottenuta dalla serie iniziale mediante la trasformazione $y_i = a + bx_i$, $i = 1, 2, \ldots, n$, essendo a e b due costanti arbitrarie. Si ha

$$\mu_Y = \frac{1}{n} \sum_{i=1}^{n} (a + bx_i) = \frac{1}{n} \sum_{i=1}^{n} a + b \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} na + b\mu_X = a + b\mu_X.$$

La proprietà è così dimostrata.

Proprietà associativa

Siano $n^{(1)}, n^{(2)}, \ldots, n^{(L)}$ e $\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(L)}$ le numerosità e le medie aritmetiche di L sottoinsiemi disgiunti della serie $x_1, x_2, \ldots, x_n$. Il totale del generico sottoinsieme può essere così espresso

$$n^{(i)}\mu^{(i)}, \; i = 1, 2, \ldots, L.$$

Il totale dell'intera serie, cioè $\sum_{i=1}^{n} x_i$, è allora dato da

$$\sum_{i=1}^{n} x_i = n^{(1)}\mu^{(1)} + n^{(2)}\mu^{(2)} + \cdots + n^{(L)}\mu^{(L)},$$

da cui

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{n^{(1)}\mu^{(1)} + n^{(2)}\mu^{(2)} + \cdots + n^{(L)}\mu^{(L)}}{n^{(1)} + n^{(2)} + \cdots + n^{(L)}}.$$

Si tratta della media aritmetica ponderata delle medie aritmetiche dei sottoinsiemi in cui è suddivisa la serie $x_1, x_2, \ldots, x_n$, con pesi pari alle loro numerosità.

A.2.2 Mediana

Proprietà (b) della Proposizione 4.5

Si consideri la serie di osservazioni $x_1, x_2, \ldots, x_n$. Sia $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ la corrispondente serie ordinata con $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$. Supponendo che n sia pari, si formino le seguenti coppie di valori: $x_{(1)}$ e $x_{(n)}$, $x_{(2)}$ e $x_{(n-1)}, \ldots, x_{(n/2)}$ e $x_{(n/2+1)}$. La somma degli scarti in valore assoluto dei termini della serie da un punto qualsiasi x può essere scritta nella forma

$$\sum_{i=1}^{n} |x_i - x| = \sum_{i=1}^{n/2} (|x_{(i)} - x| + |x_{(n-i+1)} - x|).$$

È facile verificare che, per ogni i, la somma dei due addendi tra parentesi nella sommatoria a destra dell'uguale è minima e pari a $x_{(n-i+1)} - x_{(i)}$ quando $x_{(i)} \le x \le x_{(n-i+1)}$. Poiché la mediana, per definizione, è tale che

$$x_{(i)} \le m \le x_{(n-i+1)}, \quad \text{per } i = 1, 2, \ldots, n/2,$$

la somma $\sum_{i=1}^{n} |x_i - x|$ è minima quando $x = m$.

Se n è dispari, l'argomentazione precedente va applicata alla somma

$$\sum_{i=1}^{n} |x_i - x| = \sum_{i=1}^{(n-1)/2} (|x_{(i)} - x| + |x_{(n-i+1)} - x|) + |x_{(n+1)/2} - x|,$$

che è minima quando $x = m = x_{(n+1)/2}$. La proprietà (b) della Proposizione 4.5 è così dimostrata.

Mediana nel caso delle distribuzioni di frequenza in classi

Sia $c_{h(inf)}$-$c_{h(sup)}$ la classe mediana. Nell'ipotesi di uniforme distribuzione delle unità nella classi, la funzione di ripartizione assume la forma (si veda il Paragrafo 3.2.2)

$$F(x) = F_{h-1} + \frac{1}{n} \cdot \frac{n_h}{d_h} \cdot [x - c_{h(inf)}].$$

Nel caso continuo, la mediana è il valore per cui la funzione di ripartizione è uguale a $1/2$. Per trovare la mediana basta quindi risolvere rispetto a m l'equazione

$$F_{h-1} + \frac{1}{n} \cdot \frac{n_h}{d_h}[m - c_{h(inf)}] = \frac{1}{2},$$

da cui, con semplici passaggi algebrici, si ottiene

$$m = c_{h(inf)} + \frac{n/2 - N_{h-1}}{N_h - N_{h-1}}[c_{h(sup)} - c_{h(inf)}].$$

A.3 Concentrazione

Proprietà (a) della Proposizione 5.3

La proprietà segue dalla definizione stessa dell'indice.

Proprietà (b) della Proposizione 5.3

Si considerino le due serie di osservazioni ordinate

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)} \quad \text{e} \quad x_{(1)} + c, x_{(2)} + c, \ldots, x_{(n)} + c,$$

dove ogni termine della seconda serie è dato dal corrispondente termine della prima aumentato della quantità positiva c. Si vuole dimostrare che, passando dalla prima alla seconda serie, l'indice di concentrazione di Gini diminuisce. Indicando con Q_i e Q_i^+ la i-esima frazione dell'ammontare totale della variabile nella prima e nella seconda serie, si consideri la differenza

$$Q_i^+ - Q_i = \frac{A_i + ic}{A_n + nc} - \frac{A_i}{A_n},$$

dove $A_i = x_{(1)} + x_{(2)} + \cdots + x_{(i)}$. Questa differenza è positiva per ogni $i < n$, salvo che $x_{(1)} = x_{(2)} = \cdots = x_{(n)}$. Si ha infatti

$$\begin{aligned} Q_i^+ - Q_i &= \frac{A_i + ic}{A_n + nc} - \frac{A_i}{A_n} \\ &= \frac{A_i A_n + icA_n - A_i A_n - ncA_i}{A_n(A_n + nc)} \\ &= \frac{icn\left(\frac{A_n}{n} - \frac{A_i}{i}\right)}{A_n(A_n + nc)} > 0, \ i = 1, 2, \ldots, n-1, \end{aligned}$$

essendo

$$\frac{A_n}{n} - \frac{A_i}{i} > 0,$$

perché la media aritmetica degli n elementi della serie è non inferiore alla media dei primi i elementi ordinati.

Stanti le disuguaglianze $Q_i^+ > Q_i$, $i = 1, 2, \ldots, n-1$, e considerando la (5.12), la proprietà è dimostrata.

Proprietà (c) della Proposizione 5.3

Si considerino le seguenti due serie di osservazioni ordinate

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)} \quad \text{e} \quad cx_{(1)}, cx_{(2)}, \ldots, cx_{(n)},$$

dove il singolo elemento della seconda serie è pari al corrispondente elemento della prima moltiplicato per la costante positiva c. Si vuole dimostrare che, nelle due serie, l'indice di concentrazione di Gini assume lo stesso valore.

A questo fine, indicando con Q_i e Q_i^* l'i-esima frazione dell'ammontare totale della variabile nella prima e nella seconda serie, è sufficiente mostrare che $Q_i = Q_i^*$, $i = 1, 2, \ldots, n-1$. Si ha

$$Q_i = \frac{x_{(1)} + x_{(2)} + \cdots + x_{(i)}}{x_{(1)} + x_{(2)} + \cdots + x_{(n)}} = \frac{cx_{(1)} + cx_{(2)} + \cdots + cx_{(i)}}{cx_{(1)} + cx_{(2)} + \cdots + cx_{(n)}} = Q_i^*.$$

La proprietà è così dimostrata.

Proprietà (d) della Proposizione 5.3

Siano $x_{(i)}$ e $x_{(j)}$ due termini di una serie di osservazioni con $x_{(i)} < x_{(j)}$. Si vuole dimostrare che, diminuendo $x_{(i)}$ di una quantità τ e aumentando $x_{(j)}$ della stessa quantità (mantenendo invariato l'ordine delle posizioni), l'indice di concentrazione G aumenta.

Con tale trasferimento, non mutano le quantità

$$Q_1,\ Q_2,\ \ldots,\ Q_{i-1} \quad \text{e} \quad Q_j,\ Q_{j+1},\ \ldots,\ Q_n,$$

mentre diminuiscono le quantità $Q_i, Q_{i+1}, \ldots, Q_{j-1}$. È evidente, per la (5.12), che ciò comporta un aumento dell'indice di concentrazione G. La Proprietà è dunque dimostrata.

Proposizione 5.4

L'area di concentrazione S è data dalla somma $S = \sum_{i=1}^{n} s_i$, essendo s_i la superficie ombreggiata nella Figura A.1, data dalla differenza

$$s_i = \frac{(P_i - P_{i-1})(P_{i-1} + P_i)}{2} - \frac{(P_i - P_{i-1})(Q_{i-1} + Q_i)}{2},$$

dove il primo termine a destra dell'uguale è l'area del trapezio DEHI, mentre il secondo è l'area del trapezio FGHI. Per la validità dell'espressione precedente si deve porre $P_0 = Q_0 = 0$. Si può anche scrivere:

$$s_i = \frac{1}{2}(P_i - P_{i-1})(P_{i-1} - Q_{i-1} + P_i - Q_i),$$

$$S = \frac{1}{2}\sum_{i=1}^{n}(P_i - P_{i-1})[(P_{i-1} - Q_{i-1}) + (P_i - Q_i)]. \tag{A.1}$$

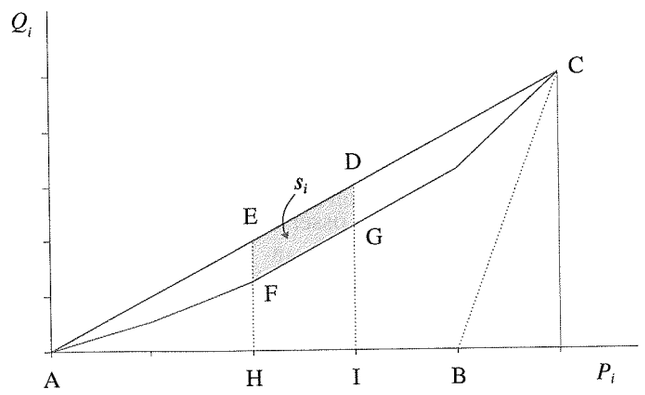


Figura A.1
Area di concentrazione.

Si consideri ora che

$$P_i - P_{i-1} = \frac{1}{n}, \quad i = 1, 2, \ldots, n,$$

e che

$$\sum_{i=1}^{n}(P_{i-1} - Q_{i-1}) = \sum_{i=1}^{n}(P_i - Q_i) = \sum_{i=1}^{n-1}(P_i - Q_i).$$

Si ha perciò

$$S = \frac{2}{2n}\sum_{i=1}^{n-1}(P_i - Q_i).$$

Poiché $\max S$ è pari all'area del triangolo di vertici ABC, con base $(n-1)/n$ e altezza 1, si può scrivere

$$\max S = \frac{n-1}{n}\frac{1}{2}.$$

In definitiva, si ha

$$\frac{S}{\max S} = \frac{2}{n-1}\sum_{i=1}^{n-1}(P_i - Q_i) = G.$$

La relazione tra l'indice di concentrazione G e l'area di concentrazione è quindi dimostrata.

A.4 Forma delle distribuzioni di frequenza

Proprietà (a) della Proposizione 6.1

Per una distribuzione simmetrica valgono le seguenti uguaglianze (si veda la Nota 1 del Capitolo 6):

$$x_i - m = -(x_{k+1-i} - m), \quad n_i = n_{k+1-i}, \quad i = 1, 2, \ldots, (k-\delta)/2,$$

dove δ è 0 o 1 a seconda che k sia pari o dispari. Da queste uguaglianze segue che:

$$\begin{aligned}(x_1 - m)n_1 &= -(x_k - m)n_k,\\ (x_2 - m)n_2 &= -(x_{k-1} - m)n_{k-1},\\ &\vdots\\ [x_{(k-\delta)/2} - m]n_{(k-\delta)/2} &= -[x_{(k-\delta)/2+1} - m]n_{(k-\delta)/2+1}.\end{aligned}$$

Si ha allora

$$\sum_{i=1}^{(k-\delta)/2}[(x_i - m)n_i + (x_{k-i+1} - m)n_{k-i+1}] + \delta[x_{(k+\delta)/2} - m]n_{(k+\delta)/2} = 0,$$

essendo $\delta[x_{(k+\delta)/2} - m]n_{(k+\delta)/2} = 0$ se k è pari e $x_{(k+\delta)/2} = x_{(k+1)/2} = m$ se k è dispari. Dalla relazione precedente si deduce che $m = \mu$, in quanto la media aritmetica è l'unico valore per cui la somma degli scarti è 0. Ciò dimostra la Proprietà.

Proprietà (b) della Proposizione 6.1

Poiché nelle distribuzioni di frequenza simmetriche $m = \mu$, dalle relazioni di uguaglianza del punto precedente, discendono le seguenti:

$$\begin{aligned}(x_1 - \mu)^3 n_1 &= -(x_k - \mu)^3 n_k,\\ (x_2 - \mu)^3 n_2 &= -(x_{k-1} - \mu)^3 n_{k-1},\\ &\vdots\\ [x_{(k-\delta)/2} - \mu]^3 n_{(k-\delta)/2} &= -[x_{(k-\delta)/2+1} - \mu]^3 n_{(k-\delta)/2+1},\end{aligned}$$

da cui

$$\sum_{i=1}^{k}(x_i-\mu)^3 n_i = \sum_{i=1}^{(k-\delta)/2}[(x_i-\mu)^3 n_i+(x_{k-i+1}-\mu)^3 n_{k-i+1}]+\delta[x_{(k+\delta)/2}-\mu]^3 n_{(k+\delta)/2} = 0.$$

Ciò perché la sommatoria a destra dell'uguale è pari a zero, essendo nullo ogni suo addendo, ed è nulla anche la quantità $\delta[x_{(k+\delta)/2} - \mu]^3 n_{(k+\delta)/2}$, in quanto, se k è pari, $\delta = 0$, mentre, se k è dispari, $x_{(k+\delta)/2} = x_{(k+1)/2} = \mu$. La proprietà è perciò dimostrata.

Proprietà (c) della Proposizione 6.1

Si assuma, per semplicità, che k sia pari e n non sia un multiplo di 4. Nelle distribuzioni di frequenza simmetriche di questo tipo, la frequenza cumulata ha la seguente proprietà

$$N_h = n - N_{k-h},\ h = 1, 2, \ldots, k/2.$$

In base alla (4.13), il primo quartile è la più piccola modalità di X, x_h, per cui la frequenza cumulata N_h è maggiore di $n/4$. Si consideri ora che un modo per determinate il terzo quartile consiste nell'individuare la più grande modalità di X, $x_{h'}$, per cui la frequenza retrocumulata $n-N_{k-h'}$ è maggiore di $n/4$. Dalla relazione tra frequenza cumulata e frequenza retrocumulata mostrata sopra, si deduce che, se $q_1 = x_h$ è il primo quartile, il terzo quartile deve essere $q_3 = x_{k-h}$. Quindi, i due quartili sono tali che $|q_1 - x_1| = |q_3 - x_k|$. Ciò implica che q_1 e q_3 siano equidistanti dalla mediana.

A.5 Disuguaglianza di Chebyshev

Sia $x_1, x_2, \ldots, x_n$ una serie di osservazioni con media μ e varianza σ^2. Sia inoltre

$$I_k = \{i, 1 \le i \le n : |x_i - \mu| < k\sigma\}$$

l'insieme degli indici i che identificano gli elementi della serie per cui lo scarto dalla media (in valore assoluto) è minore di $k\sigma$. Si indichi con $N(I_k)$ il numero di elementi appartenenti all'insieme I_k. Si può scrivere

$$\begin{aligned} n\sigma^2 &= \sum_{i=1}^{n}(x_i - \mu)^2 \\ &= \sum_{i\in I_k}(x_i - \mu)^2 + \sum_{i\notin I_k}(x_i - \mu)^2 \\ &\ge \sum_{i\notin I_k}(x_i - \mu)^2 \\ &\ge \sum_{i\notin I_k} k^2\sigma^2, \end{aligned}$$

dove la prima disuguaglianza sussiste perché la somma dei quadrati degli scarti dalla media viene limitata solo agli elementi che non appartengono a I_k, mentre la seconda segue dal fatto che $(x_i - \mu)^2 \ge k^2\sigma^2$. Ne segue che

$$\sum_{i\notin I_k} k^2\sigma^2 \le n\sigma^2,$$

da cui, dividendo entrambi i membri per $nk^2\sigma^2$, si ottiene

$$\frac{1}{n}\sum_{i\notin I_k}(1) \le \frac{1}{k^2} \Leftrightarrow \frac{n - n(I_k)}{n} \le \frac{1}{k^2},$$

da cui

$$\frac{n(I_k)}{n} \ge 1 - \frac{1}{k^2}.$$

La Proposizione 7.1 è così dimostrata.

A.6 Analisi della dipendenza nelle tabelle di contingenza

Massimo del coefficiente Φ

Si vuole dimostrare che in una tabella di contingenza con s righe e t colonne

$$\Phi \le \sqrt{\min(s,t) - 1},$$

dove $\min(s,t)$ è il valore più piccolo tra il numero delle righe e quello delle colonne. A tal fine, si scriva

$$\Phi^2 = \sum_{i=1}^{s}\sum_{j=1}^{t}\frac{n_{ij}}{n_{i0}}\cdot\frac{n_{ij}}{n_{0j}} - 1.$$

Ora, per ogni coppia (i, j), vale la relazione

$$\frac{n_{ij}}{n_{i0}}\cdot\frac{n_{ij}}{n_{0j}} \le \frac{n_{ij}}{n_{0j}},$$

perché $n_{ij}/n_{i0} \le 1$, cosicché si può scrivere

$$\Phi^2 = \sum_{i=1}^{s}\sum_{j=1}^{t}\frac{n_{ij}}{n_{i0}}\cdot\frac{n_{ij}}{n_{0j}} - 1 \le \sum_{i=1}^{s}\sum_{j=1}^{t}\frac{n_{ij}}{n_{0j}} - 1 = \sum_{j=1}^{t}\frac{1}{n_{0j}}\sum_{i=1}^{s}n_{ij} - 1 = \sum_{j=1}^{t}\frac{1}{n_{0j}}n_{0j} - 1,$$

da cui

$$\Phi^2 \le t - 1,$$

dove vale l'uguaglianza se nella Tabella 9.1, per ogni i, vi è un solo valore di j per cui $n_{ij} > 0$, ossia se la variabile Y dipende perfettamente dalla variabile X.

Con procedimento analogo, si ha

$$\Phi^2 = \sum_{i=1}^{s}\sum_{j=1}^{t}\frac{n_{ij}}{n_{i0}}\cdot\frac{n_{ij}}{n_{0j}} - 1 \le \sum_{i=1}^{s}\sum_{j=1}^{t}\frac{n_{ij}}{n_{i0}} - 1 = \sum_{i=1}^{s}\frac{1}{n_{i0}}\sum_{j=1}^{t}n_{ij} - 1 = \sum_{i=1}^{s}\frac{1}{n_{i0}}n_{i0} - 1,$$

da cui si deduce che

$$\Phi^2 \le s - 1,$$

dove vale l'uguaglianza se nella Tabella 9.1, per ogni j, vi è un solo valore di i per cui $n_{ij} > 0$, cioè se X dipende perfettamente da Y.

Dalle due relazioni di disuguaglianza, $\Phi^2 \le s - 1$ e $\Phi^2 \le t - 1$, si deduce infine che

$$\Phi^2 \le \min(s,t) - 1 \Leftrightarrow \Phi \le \sqrt{\min(s,t) - 1}.$$

La disuguaglianza $\Phi \le \sqrt{\min(s,t) - 1}$ è dunque dimostrata.

Scomposizione della devianza

Si consideri una distribuzione doppia di frequenza in cui la variabile Y è quantitativa. La devianza della distribuzione marginale di Y può essere scritta nella forma

$$\sum_{j=1}^{t}(y_j-\mu_Y)^2 n_{0j} = \sum_{j=1}^{t}(y_j-\mu_Y)^2 \sum_{i=1}^{s} n_{ij} = \sum_{i=1}^{s}\sum_{j=1}^{t}(y_j-\mu_Y)^2 n_{ij}.$$

Si ha anche

$$\sum_{i=1}^{s}\sum_{j=1}^{t}(y_j-\mu_Y)^2 n_{ij} = \sum_{i=1}^{s}\sum_{j=1}^{t}\{[y_j-\mu_Y(x_i)]+[\mu_Y(x_i)-\mu_Y]\}^2 n_{ij},$$

dove $\mu_Y(x_i)$, $i = 1,2,\ldots,s$, sono le medie condizionate della variabile Y. Lo sviluppo del quadrato del binomio entro le parentesi graffe al secondo membro consente di scrivere

$$\begin{aligned}\sum_{i=1}^{s}\sum_{j=1}^{t}(y_j-\mu_Y)^2 n_{ij} = {} & \sum_{i=1}^{s}\sum_{j=1}^{t}[y_j-\mu_Y(x_i)]^2 n_{ij} + \sum_{i=1}^{s}\sum_{j=1}^{t}[\mu_Y(x_i)-\mu_Y]^2 n_{ij} \\ & +2\sum_{i=1}^{s}\sum_{j=1}^{t}[y_j-\mu_Y(x_i)][\mu_Y(x_i)-\mu_Y]n_{ij},\end{aligned}$$

dove il terzo addendo a destra dell'uguale è nullo, potendosi scrivere

$$\sum_{i=1}^{s}\sum_{j=1}^{t}[y_j-\mu_Y(x_i)][\mu_Y(x_i)-\mu_Y]n_{ij} = \sum_{i=1}^{s}[\mu_Y(x_i)-\mu_Y]\sum_{j=1}^{t}[y_j-\mu_Y(x_i)]n_{ij} = 0,$$

dato che, per ogni i, $\sum_{j=1}^{t}[y_j-\mu_Y(x_i)]n_{ij} = 0$. Infine, considerando che

$$\sum_{i=1}^{s}\sum_{j=1}^{t}[\mu_Y(x_i)-\mu_Y]^2 n_{ij} = \sum_{i=1}^{s}[\mu_Y(x_i)-\mu_Y]^2 n_{i0},$$

si può scrivere

$$\sum_{j=1}^{t}(y_j-\mu_Y)^2 n_{0j} = \sum_{i=1}^{s}[\mu_Y(x_i)-\mu_Y]^2 n_{i0} + \sum_{i=1}^{s}\sum_{j=1}^{t}[y_j-\mu_Y(x_i)]^2 n_{ij}.$$

La Proposizione 9.2 è così dimostrata.

A.7 Regressione lineare semplice

Determinazione di b_0 e b_1

Si vogliono determinare le quantità b_0 e b_1 che rendono minima la somma

$$S_q = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2. \qquad \text{(A.2)}$$

Chiaramente, per ogni valore prefissato b_1, S_q è funzione della sola b_0. Pertanto, assumendo come noto b_1, si procede determinando il valore b_0 che minimizza S_q. A tale proposito, conviene riscrivere S_q nella forma

$$S_q = \sum_{i=1}^{n}(w_i - b_0)^2,$$

dove le quantità $w_i = y_i - b_1 x_i$ $(i = 1,2,\ldots,n)$ vanno pensate come numeri noti. È facile riconoscere che il valore di b_0 che rende minima S_q è la media aritmetica delle quantità $w_1, w_2, \ldots, w_n$ (per la Proprietà (d) della media aritmetica). Si ha pertanto

$$b_0 = \frac{1}{n}\sum_{i=1}^{n} w_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - b_1 x_i) = \mu_Y - b_1\mu_X.$$

Per determinare b_1, basta sostituire l'espressione di b_0 nella (A.2) e minimizzare S_q rispetto a b_1. Si ha

$$S_q = \sum_{i=1}^{n}(y_i - \mu_Y + b_1\mu_X - b_1 x_i)^2 = \sum_{i=1}^{n}[(y_i-\mu_Y) - b_1(x_i-\mu_X)]^2.$$

Svolgendo il quadrato del binomio dentro le parentesi quadre si ottiene

$$S_q = b_1^2\sum_{i=1}^{n}(x_i-\mu_X)^2 - 2b_1\sum_{i=1}^{n}(x_i-\mu_X)(y_i-\mu_Y) + \sum_{i=1}^{n}(y_i-\mu_Y)^2.$$

In questa espressione, tutte le quantità diverse da b_1 sono numeri noti. Dunque, S_q è l'equazione di una parabola del tipo $S_q = k_1 b_1^2 + k_2 b_1 + k_3$, con la concavità rivolta verso l'alto, essendo $k_1 = \sum_{i=1}^{n}(x_i-\mu_X)^2 > 0$. Ne segue che il valore di b_1 che minimizza S_q è l'ascissa del vertice della parabola, ossia la quantità

$$b_1 = -\frac{k_2}{2k_1} = \frac{2\sum_{i=1}^{n}(x_i-\mu_X)(y_i-\mu_Y)}{2\sum_{i=1}^{n}(x_i-\mu_X)^2}.$$

Si può concludere che i valori di b_1 e b_0 che minimizzano la (A.2) sono dati da:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i-\mu_X)(y_i-\mu_Y)}{\sum_{i=1}^{n}(x_i-\mu_X)^2},\quad b_0 = \mu_Y - b_1\mu_X.$$

La Proposizione 10.1 è dunque dimostrata.

Nullità della somma dei residui

Il risultato segue dalla serie di uguaglianze:

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\hat{y}_i) &= \sum_{i=1}^{n}[y_i-(b_0+b_1x_i)] \\ &= \sum_{i=1}^{n}[y_i-(\mu_Y-b_1\mu_X+b_1x_i)] \\ &= \sum_{i=1}^{n}y_i - n\mu_Y + nb_1\mu_X - b_1\sum_{i=1}^{n}x_i = 0,\end{aligned}$$

essendo $\sum_{i=1}^{n} y_i = n\mu_Y$ e $\sum_{i=1}^{n} x_i = n\mu_X$ per la Proprietà (b) della media aritmetica.

Scomposizione della devianza

Si vuole dimostrare la (10.5) della Proposizione 10.2, che qui si ripropone

$$\sum_{i=1}^{n}(y_i - \mu_Y)^2 = \sum_{i=1}^{n}(\hat{y}_i - \mu_Y)^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

dove $\hat{y}_i = b_0 + b_1 x_i$ sono i valori teorici forniti dalla retta di regressione. Si consideri la catena di uguaglianze:

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 &= \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 \\
&= \sum_{i=1}^{n}(y_i - \mu_Y + b_1\mu_X - b_1 x_i)^2 \\
&= \sum_{i=1}^{n}[(y_i - \mu_Y) - b_1(x_i - \mu_X)]^2 \\
&= \sum_{i=1}^{n}(y_i - \mu_Y)^2 + b_1^2\sum_{i=1}^{n}(x_i - \mu_X)^2 - 2b_1\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y) \\
&= \sum_{i=1}^{n}(y_i - \mu_Y)^2 + b_1^2\sum_{i=1}^{n}(x_i - \mu_X)^2 - 2b_1 \cdot b_1\sum_{i=1}^{n}(x_i - \mu_X)^2 \\
&= \sum_{i=1}^{n}(y_i - \mu_Y)^2 - b_1^2\sum_{i=1}^{n}(x_i - \mu_X)^2,
\end{aligned}$$

dove, nel penultimo passaggio, è stata utilizzata la relazione

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{\sum_{i=1}^{n}(x_i - \mu_X)^2} \Rightarrow \sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y) = b_1\sum_{i=1}^{n}(x_i - \mu_X)^2.$$

In definitiva, la devianza residua può essere scritta nellaa forma

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \mu_Y)^2 - b_1^2\sum_{i=1}^{n}(x_i - \mu_X)^2, \tag{A.3}$$

dove il secondo termine a destra dell'uguale può essere posto nella forma

$$\begin{aligned}
b_1^2\sum_{i=1}^{n}(x_i - \mu_X)^2 &= \sum_{i=1}^{n}(b_1 x_i - b_1\mu_X)^2 \\
&= \sum_{i=1}^{n}(\mu_Y - b_1\mu_X + b_1 x_i - \mu_Y)^2 \\
&= \sum_{i=1}^{n}(\hat{y}_i - \mu_Y)^2.
\end{aligned} \tag{A.4}$$

Inserendo quest'ultimo nella (A.3) e riordinando i termini, si ottiene il risultato desiderato.

Indice di determinazione in funzione di C_{XY}, D_X e D_Y

Si vuole dimostrare che

$$r^2 = \frac{C_{XY}^2}{D_X D_Y},$$

dove D_X, D_Y e C_{XY} sono la devianza di X, la devianza di Y e la codevianza tra X e Y.

Sulla base delle (10.6) e (A.4), si può scrivere

$$r^2 = \frac{1}{D_Y}\sum_{i=1}^{n}(\hat{y}_i - \mu_Y)^2 = \frac{1}{D_Y}b_1^2 D_X,$$

da cui, sostituendo a b_1 la sua espressione e semplificando, si ottiene il risultato desiderato.

Proprietà (c) della Proposizione 10.3

Si vuole dimostrare che

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0$$

se e solo se

$$y_i = c + dx_i,\ i = 1, 2, \ldots, n,$$

cioè, se e solo se la variabile Y è legata alla variabile X da una relazione lineare perfetta. A questo fine, si suppone che sussista la precedente relazione. In tal caso, si ha

$$\mu_Y = \frac{1}{n}\sum_{i=1}^{n}y_i = \frac{1}{n}\sum_{i=1}^{n}(c + dx_i) = c + d\mu_X.$$

Inoltre,

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(c + dx_i - c - d\mu_X)}{\sum_{i=1}^{n}(x_i - \mu_X)^2} = \frac{d\sum_{i=1}^{n}(x_i - \mu_X)^2}{\sum_{i=1}^{n}(x_i - \mu_X)^2} = d,$$

e

$$b_0 = \mu_Y - b_1\mu_X = c + d\mu_X - d\mu_X = c.$$

La conclusione è che, se i punti osservati sono su una retta, questa coincide con la retta di regressione e $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0$. Viceversa, $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0$ implica che i punti osservati si trovino su una retta, dovendo essere $\hat{y}_i = b_0 + b_1 x_i = y_i$, $i = 1, 2, \ldots, n$. La Proprietà (c) della Proposizione 10.3 è quindi dimostrata.

A.8 Correlazione

Formula operativa per l'indice r

Come è noto, la codevianza può essere posta nella forma (si veda Esercizio 10.2*)

$$C_{XY} = \sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y) = \sum_{i=1}^{n}x_i y_i - n\mu_X\mu_Y.$$

Si sa anche che

$$\sigma_X^2 = \frac{\sum_{i=1}^n x_i^2}{n} - \mu_X^2,$$
$$\sigma_Y^2 = \frac{\sum_{i=1}^n y_i^2}{n} - \mu_Y^2$$

(si veda il Paragrafo 5.2, formula operativa della deviazione standard), da cui

$$\sum_{i=1}^n (x_i - \mu_X)^2 = n\sigma_X^2 = \sum_{i=1}^n x_i^2 - n\mu_X^2,$$
$$\sum_{i=1}^n (y_i - \mu_Y)^2 = n\sigma_Y^2 = \sum_{i=1}^n y_i^2 - n\mu_Y^2.$$

Ne segue la formula operativa per il calcolo di r:

$$r = \frac{\sum_{i=1}^n (x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_{i=1}^n (x_i - \mu_X)^2 \sum_{i=1}^n (y_i - \mu_Y)^2}} = \frac{\sum_{i=1}^n x_i y_i - n\mu_X\mu_Y}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n\mu_X^2\right)\left(\sum_{i=1}^n y_i^2 - n\mu_Y^2\right)}}.$$

APPENDICE

B

Probabilità e inferenza: dimostrazioni

Sommario dell'appendice

B.1 Famiglie di eventi e misure di probabilità
B.2 Regola della somma
B.3 Elementi di calcolo combinatorio
B.4 Funzione di ripartizione e funzione di densità
B.5 Valore atteso di trasformazioni di variabili casuali
B.6 Disuguaglianza di Chebyshev
B.7 Disuguaglianza di Markov
B.8 Distribuzioni di probabilità
B.9 Valore atteso di combinazioni lineari di funzioni di due o più variabili casuali
B.10 Argomenti di stima puntuale
B.11 Inferenza su popolazioni finite

B.1 Famiglie di eventi e misure di probabilità

Una misura di probabilità $P(\cdot)$ non è necessariamente definita per tutti i sottoinsiemi dello spazio campionario Ω, ma solo per una appropriata famiglia di questi, chiamata algebra di eventi, che soddisfa le condizioni:

- lo spazio campionario appartiene alla famiglia;
- se gli eventi $A_1, A_2, \ldots$ appartengono alla famiglia, allora anche la loro unione appartiene alla famiglia;
- se l'evento A appartiene alla famiglia, allora anche il suo complemento $\bar{A}$ appartiene alla famiglia.

Ci sono motivazioni sia pratiche sia matematiche che portano a non definire la misura di probabilità $P(\cdot)$ su tutti i sottoinsiemi di Ω.

ESEMPIO B.1

Si consideri il lancio di un dado (che potrebbe essere bilanciato oppure no) con 6 facce per il quale si conosce solamente che $P(1) = 1/6$ e $P(6) = 1/6$. Si determini la famiglia di eventi per questa situazione.

Soluzione

Lo spazio campionario dell'esperimento è dato da $\Omega = \{1, 2, 3, 4, 5, 6\}$. Con le informazioni disponibili, è possibile assegnare una probabilità solo agli eventi:

$$\emptyset,\ \{1\},\ \{6\},\ \{2,3,4,5\},\ \{1,6\},\ \{1,2,3,4,5\},\ \{2,3,4,5,6\},\ \Omega.$$

È facile verificare che questa collezione di eventi soddisfa le tre condizioni sopra indicate e costituisce perciò un'algebra di eventi rispetto alla quale può essere definita una misura di probabilità $P(\cdot)$. ■

La definizione di una famiglia di eventi si pone in tutte le situazioni in cui è conveniente o necessario, per qualche ragione, restringere la misura di probabilità $P(\cdot)$ a una specifica collezione di sottoinsiemi dello spazio campionario.

Se lo spazio campionario è continuo, come nell'Esempio 12.4, $P(\cdot)$ è ristretta a domini costituiti da intervalli aperti. Ciò è reso necessario dalla impossibilità di assegnare una probabilità a tutti i sottoinsiemi dello spazio campionario. Una trattazione più approfondita di questi aspetti va oltre gli scopi di questo volume (al riguardo si veda Spanos, 1999).

B.2 Regola della somma

Si riscriva l'evento $A_1 \cup A_2$ come unione dei tre eventi disgiunti $A_1 \cap \bar{A}_2$, $A_2 \cap \bar{A}_1$ e $A_1 \cap A_2$:

$$A_1 \cup A_2 = (A_1 \cap \bar{A}_2) \cup (A_2 \cap \bar{A}_1) \cup (A_1 \cap A_2).$$

Ne deriva, per il terzo assioma della probabilità, che

$$P(A_1 \cup A_2) = P(A_1 \cap \bar{A}_2) + P(A_2 \cap \bar{A}_1) + P(A_1 \cap A_2). \tag{B.1}$$

Dalle ovvie relazioni:

$$P(A_1) = P(A_1 \cap \bar{A}_2) + P(A_1 \cap A_2),$$
$$P(A_2) = P(A_2 \cap \bar{A}_1) + P(A_1 \cap A_2),$$

è possibile ricavare $P(A_1 \cap \bar{A}_2)$ e $P(A_2 \cap \bar{A}_1)$:

$$P(A_1 \cap \bar{A}_2) = P(A_1) - P(A_1 \cap A_2),$$
$$P(A_2 \cap \bar{A}_1) = P(A_2) - P(A_1 \cap A_2).$$

Sostituendo queste espressioni in (B.1), si ottiene

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2).$$

La regola della somma è così dimostrata.

B.3 Elementi di calcolo combinatorio

Il calcolo combinatorio è una branca della matematica che si occupa delle proprietà di certi raggruppamenti che si possono formare con un numero finito di oggetti.

Permutazioni

Le permutazioni di un insieme di n oggetti distinti sono i gruppi che possono essere formati ordinando gli elementi dell'insieme in tutti i modi possibili. Per esempio, le permutazioni delle tre lettere a, b e c sono:

$$abc, acb, bac, bca, cab, cba.$$

PROPOSIZIONE B.1

Il numero di permutazioni di un insieme di n oggetti distinti è dato da

$$P_n = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1. \tag{B.2}$$

Questo numero si indica con $n!$ e si legge "n fattoriale". Per convenzione, si pone $0! = 1$.

La (B.2) può essere ottenuta come segue. Il primo elemento della sequenza degli n oggetti può essere scelto in n modi diversi, il secondo in $n-1$ modi diversi e così via, l'ultimo elemento potrà essere scelto in un solo modo, essendo rimasto un unico elemento da collocare. Dunque, il numero di sequenze che si possono formare con n oggetti è $n!$. ■

Disposizioni

Le disposizioni, anche dette permutazioni di dimensione m, sono i gruppi che si possono formare orinando in tutti i modi possibili m oggetti presi da un insieme di n oggetti distinti.

PROPOSIZIONE B.2

Il numero di disposizioni, ovvero di permutazioni di dimensione m, di un insieme di n oggetti distinti, dove m è minore o uguale a n, è dato da

$$P_{n,m} = n \cdot (n-1) \cdots (n-m+1). \tag{B.3}$$

La dimostrazione si basa sullo stesso ragionamento seguito per per la dimostrazione della Proposizione B.1. ■

Dati quattro oggetti, a, b, c, d, il numero di permutazioni di dimensione 2 che si possono generare da questo insieme è

$$P_{4,2} = 4 \cdot 3 = 12.$$

Esse sono: ab, ba, ac, ca, ad, da, bc, cb, bd, db, cd, dc.

Combinazioni

Le combinazioni di dimensione m di un insieme di n oggetti sono i gruppi di m elementi che si possono formare prendendo m elementi alla volta da questo insieme. L'ordine degli elementi nei gruppi non ha importanza: due gruppi sono diversi se differiscono per almeno un elemento. Per esempio, date le quattro lettere a, b, c e d, le combinazioni di dimensione 2 sono:

$$ab,\ ac,\ ad,\ bc,\ bd,\ cd.$$

PROPOSIZIONE B.3

Il numero di combinazioni di dimensione m di un insieme di n oggetti distinti, dove m è minore o uguale a n, è dato da

$$C_{n,m} = \binom{n}{m} = \frac{n!}{m!(n-m)!}, \tag{B.4}$$

dove il simbolo $\binom{n}{m}$ si legge "n su m."

Si consideri che le $P_{n,m}$ disposizioni si possono ottenere formando prima le $C_{n,m}$ combinazioni, e poi permutando gli elementi di ognuna di queste combinazioni. Si può allora scrivere

$$P_{n,m} = C_{n,m} \cdot m!,$$

da cui

$$C_{n,m} = \frac{P_{n,m}}{m!} = \frac{n \cdot (n-1) \cdots (n-m+1)}{m!} = \frac{n!}{m!(n-m)!}.$$

Per passare dalla terza alla quarta espressione delle relazioni precedenti basta moltiplicare il numeratore e il denominatore della terza espressione per $(n-m)!$. ■

Come esempio, si calcoli il numero di combinazioni di 4 oggetti, a, b, c e d, presi a 2 a 2. Si ha

$$\binom{4}{2} = \frac{4!}{2!(4-2)!} = 6.$$

Le (B.3) e (B.4) possono essere generalizzate al caso in cui sono permesse le ripetizioni. In altre parole, si vuole trovare in quanti modi si possono scegliere m elementi da un insieme di n oggetti, nel caso in cui ognuno di questi oggetti può essere selezionato più di una volta.

Disposizioni con ripetizione

Conviene partire da un esempio. Le disposizioni con ripetizione di dimensione 3 di 2 oggetti, a e b, sono: *aaa*, *abb*, *bab*, *bba*, *aba*, *baa*, *aab*, *bbb*.

PROPOSIZIONE B.4

Il numero di disposizioni con ripetizione di dimensione m di un insieme di n oggetti distinti è dato da

$$P_{n,m}^{W} = n^m. \tag{B.5}$$

La (B.5) può essere dimostrata seguendo lo stesso ragionamento usato per la (B.2). Il primo elemento della sequenza di lunghezza m può essere scelto in n modi diversi, il secondo ancora in n modi diversi e così via fino ad arrivare all'm-esima posizione. ■

Applicando la (B.5) all'esempio iniziale, si trova che il numero di disposizioni con ripetizione di dimensione 3 di 2 oggetti, a e b, è effettivamente 8:

$$P_{2,3}^{W} = 2^3 = 8.$$

Combinazioni con ripetizione

Si vogliano costruire le combinazioni con ripetizione di dimensione 4 dell'insieme di due oggetti, a e b. Si ha

aaaa, *aaab*, *aabb*, *abbb*, *bbbb*.

PROPOSIZIONE B.5

Il numero di combinazioni con ripetizione di dimensione m di un insieme di n oggetti distinti è dato da

$$C_{n,m}^{W} = \binom{m+n-1}{m}. \tag{B.6}$$

Per dimostrare ciò, si osservi che ogni combinazione con ripetizione di dimensione m di n oggetti, si può rappresentare con una sequenza di $n-1$ barre e m asterischi, in cui le $n-1$ barre delimitano n diverse celle. Per esempio, le combinazioni con ripetizione di dimensione due viste sopra si possono rappresentare così:

$$****|,\quad ***|*,\quad **|**,\quad *|***,\quad |****,$$

dove ogni sequenza, contenente $m+n-1 = 4+2-1 = 5$ simboli, rappresenta una combinazione con ripetizione. Ora, è facile notare che il numero di combinazioni con ripetizione di dimensione m corrisponde al numero di modi di scegliere le $n-1$ posizioni in cui collocare le barre tra le $m+n-1$ posizioni contenenti m asterischi e $n-1$ barre. Quindi, il numero di combinazioni con ripetizione di dimensione m è pari al numero di combinazioni (senza ripetizione) di $m+n-1$ oggetti presi a m a m. ■

Applicando la (B.6) all'esempio iniziale, si ottiene

$$C_{2,4}^{W} = \binom{m+n-1}{m} = \binom{4+2-1}{4} = \binom{5}{4} = 5.$$

Dunque, il numero di combinazioni con ripetizione di dimensione 4 che si possono ottenere con due oggetti distinti è 5, a conferma di quanto trovato con la costruzione effettiva dei raggruppamenti.

Permutazioni con oggetti indistinguibili

Quando gli n oggetti da permutare non sono tutti distinguibili, il numero di permutazioni è minore di $n!$. Per esempio, le sequenze ordinate che si possono formare con le quattro lettere della parola inglese TREE non sono 24, ma 12:

TREE, RTEE, RETE, REET, ETRE, ETER, ERTE,
ERET, EERT, EETR, TEER, TERE.

PROPOSIZIONE B.6

Il numero di permutazioni di n oggetti, dei quali n_1 di tipo 1, n_2 di tipo 2, ..., n_k di tipo k, è dato da

$$P_{n;n_1,n_2,\ldots,n_k} = \frac{n!}{n_1! \cdot n_2! \cdots n_k!}. \tag{B.7}$$

Per ottenere la (B.7), si può così argomentare. Gli n_1 oggetti di tipo 1 si possono sistemare nelle n posizioni in C_{n,n_1} modi, lasciando $n-n_1$ posizioni libere. Poi, gli n_2 oggetti di tipo 2 si possono sistemare nelle $n-n_1$ posizioni rimanenti in C_{n-n_1,n_2} modi, lasciando $n-n_1-n_2$ posizioni libere. Procedendo così fino a che gli n_k oggetti di tipo k sono sistemati nelle $n-n_1-n_2-\cdots-n_{k-1}$ posizioni in $C_{n-n_1-n_2-\cdots-n_{k-1},n_k}$ modi, cioè in un solo modo, si ottiene

$$\begin{aligned} P_{n;n_1,n_2,\ldots,n_k} &= \frac{n!}{n_1!(n-n_1)!} \cdot \frac{(n-n_1)!}{n_2!(n-n_1-n_2)!} \cdots \frac{(n-n_1-n_2-\cdots-n_{k-1})!}{n_k!0!} \\ &= \frac{n!}{n_1! \cdot n_2! \cdots n_k!}, \end{aligned}$$

come volevasi dimostrare. ■

Riprendendo l'esempio delle lettere della parola inglese TREE, si vede che il numero delle permutazioni si ottiene applicando la (B.7) come segue

$$P_{4;2,1,1} = \frac{4!}{2!\,1!\,1!} = 12.$$

B.4 Funzione di ripartizione e funzione di densità

Come si è visto, la funzione di ripartizione è così definita

$$F(x) = P(X \le x) = \begin{cases} \sum_{t \le x} f(t), & \text{se } X \text{ è discreta,} \\ \int_{-\infty}^{x} f(t)\,\mathrm{d}t, & \text{se } X \text{ è continua.} \end{cases}$$

Nel caso discreto, $F(x)$ è una funzione costante a tratti; nel caso continuo, $F(x)$ è una funzione continua.

La funzione di ripartizione di una qualsiasi variabile casuale ha le seguenti proprietà:

- tende a 0 per $x \to -\infty$, ossia $\lim_{x\to-\infty} F(x) = 0$;
- tende a 1 per $x \to \infty$, ossia $\lim_{x\to\infty} F(x) = 1$;
- è non decrescente, cioè $F(a) \le F(b)$ per ogni coppia di numeri reali a e b tali che $a < b$;
- è continua a destra, cioè $\lim_{h\to0^+} F(x+h) = F(x)$.

La funzione di ripartizione permette di calcolare la probabilità che la variabile casuale X assuma valori nell'intervallo, aperto a sinistra e chiuso a destra, $(a, b]$, per ogni a e b, come segue

$$P(a < X \le b) = P(X \le b) - P(X \le a) = F(b) - F(a).$$

Per una variabile casuale continua, la funzione di ripartizione è derivabile quasi ovunque, a meno che in una infinità numerabile di punti, e tra la funzione di densità e la funzione di ripartizione sussiste la relazione

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x},$$

per ogni punto di continuità di $f(x)$.

Nell'esempio che segue si mostra come ricavare, per una variabile casuale continua, la funzione di ripartizione dalla funzione di densità e viceversa.

ESEMPIO B.2

Sia data una variabile casuale continua X con funzione di densità

$$f(x) = \begin{cases} x, & 0 \le x \le 1, \\ 1/2, & 1.5 \le x \le 2.5, \\ 0, & \text{altrove,} \end{cases}$$

il cui grafico è dato nella Figura B.1.

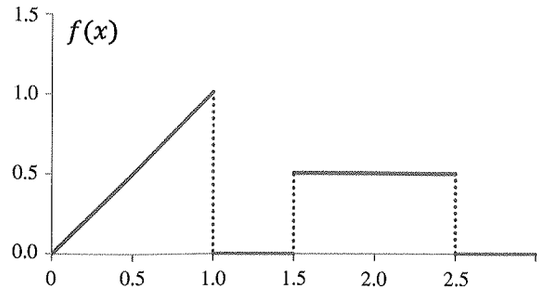


Figura B.1
Grafico della funzione di densità in esame.

La funzione di ripartizione della variabile casuale può essere determinata nel modo seguente. Per $x < 0$, $F(x) = 0$; per $x > 2.5$, $F(x) = 1$; per $1 < x < 1.5$, $F(x) = 0.5$; per $0 \le x \le 1$, si ha

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \int_{-\infty}^{0} 0\,\mathrm{d}t + \int_{0}^{x} t\,\mathrm{d}t = 0 + \left[\frac{t^2}{2}\right]_0^x = \frac{x^2}{2};$$

per $1.5 \le x \le 2.5$, si ha

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \int_{-\infty}^{0} 0\,\mathrm{d}t + \int_{0}^{1} t\,\mathrm{d}t + \int_{1}^{1.5} 0\,\mathrm{d}t + \int_{1.5}^{x} 0.5\,\mathrm{d}t$$

$$= 0 + 0.5 + 0 + \left[0.5\,t\right]_{1.5}^{x} = 0.5 + (0.5 \cdot x - 0.5 \cdot 1.5) = 0.5x - 0.25.$$

Riassumendo, la funzione di ripartizione è data da

$$F(x) = \begin{cases} 0, & x < 0, \\ x^2/2, & 0 \le x \le 1, \\ 0.5, & 1 < x < 1.5, \\ 0.5x - 0.25, & 1.5 \le x \le 2.5, \\ 1, & x > 2.5. \end{cases}$$

Il grafico di $F(x)$ è dato nella Figura B.2.

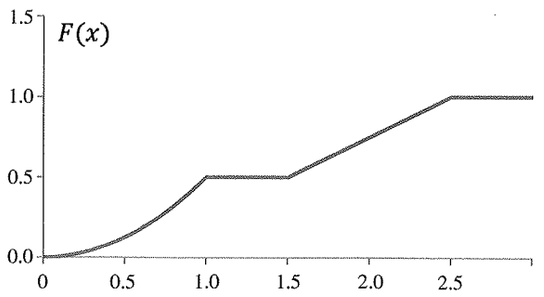


Figura B.2
Grafico della funzione di ripartizione in esame.

Si consideri ora il problema inverso di ricavare la funzione di densità a partire dalla funzione di ripartizione. Si trova che:

$$f(x) = 0$$

negli intervalli in cui $F(x)$ è costante;

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x}\frac{x^2}{2} = x,$$

per $0 < x < 1$;

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x}(0.5x - 0.25) = 0.5,$$

per $1.5 < x < 2.5$.

Si osservi che la funzione di densità risulta indeterminata per $x = 1, 1.5, 2.5$, in quanto in questi punti non esiste la derivata di $F(x)$. Nondimeno, nei punti $x = 1, 1.5, 2.5$ si può definire $f(x)$ a piacere. Una possibilità, per esempio, è di porre $f(1) = 1$ e $f(1.5) = f(2.5) = 0.5$. ■

I valori arbitrari che si possono assegnare alla funzione di densità nei punti (in numero finito o infinito numerabile) in cui $F(x)$ non è differenziabile non alterano la distribuzione di probabilità di X. Infatti, indipendentemente dai valori che si assegnano alla funzione di densità $f(x)$ in questi punti, resta inalterata la probabilità $P(a < X \leq b)$ per ogni coppia di valori a e b. Questo fatto mette in luce una proprietà generale: a una data funzione di ripartizione corrisponde una pluralità di funzioni di densità che differiscono per i valori assunti in un numero finito o in un'infinità numerabile di punti.

B.5 Valore atteso di trasformazioni di variabili casuali

In molte situazioni, si è interessati al valore atteso di una funzione, chiamata anche trasformazione, $Y = g(X)$, di una variabile casuale. Si assuma che la funzione di probabilità o di densità di Y sia $h(y)$. Allora, sulla base della definizione data nel Capitolo 13 (si veda il Paragrafo 13.2.1 per il caso discreto e il Paragrafo 13.3.1 per il caso continuo), il valore atteso di Y è dato da

$$\mathrm{E}(Y) = \sum_y y\,h(y) \quad \text{oppure} \quad \int_{-\infty}^{\infty} y\,h(y)\,\mathrm{d}y,$$

secondo che Y sia discreta o continua. C'è però un modo alternativo ed equivalente di calcolare $\mathrm{E}(Y)$ che non richiede la funzione di probabilità o di densità, $h(y)$, come enunciato nella Proposizione B.7.

PROPOSIZIONE B.7

Sia X una variabile casuale con funzione di probabilità o di densità $f(x)$ e sia $g(X)$ una funzione a valori reali di X. Allora si può mostrare che il valore atteso di $Y = g(X)$ è dato da

$$\mathrm{E}(Y) = \sum_x g(x)f(x), \quad \text{se } X \text{ è discreta,}$$

$$\mathrm{E}(Y) = \int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{d}x, \quad \text{se } X \text{ è continua.}$$

Proprietà del valore atteso

Le proprietà fondamentali del valore atteso sono enunciate nella Proposizione B.8.

PROPOSIZIONE B.8

Il valore atteso soddisfa le seguenti proprietà:

(a) se k è una costante, allora $\mathrm{E}(k) = k$;

(b) se k è una costante e $g(\cdot)$ è una funzione, allora $\mathrm{E}[kg(X)] = k\mathrm{E}[(g(X)]$;

(c) se k_1 e k_2 sono costanti e $g_1(\cdot)$ e $g_2(\cdot)$ sono funzioni, allora

$$\mathrm{E}[k_1g_1(X) + k_2g_2(X)] = k_1\mathrm{E}[(g_1(X)] + k_2\mathrm{E}[(g_2(X)].$$

Le Proprietà (a) e (b) sono immediate. Per quanto riguarda la Proprietà (c), nel caso discreto, si ha

$$\begin{aligned}\mathrm{E}[k_1g_1(X) + k_2g_2(X)] &= \sum_x [k_1g_1(x) + k_2g_2(x)]\,f(x)\\ &= k_1\sum_x g_1(x)f(x) + k_2\sum_x g_2(x)f(x)\\ &= k_1\mathrm{E}[g_1(X)] + k_2\mathrm{E}[g_2(X)].\end{aligned}$$

Analogamente, nel caso continuo, si può scrivere

$$\begin{aligned}\mathrm{E}[k_1g_1(X) + k_2g_2(X)] &= \int_{-\infty}^{\infty} [k_1g_1(x) + k_2g_2(x)]\,f(x)\mathrm{d}x\\ &= k_1\int_{-\infty}^{\infty} g_1(x)f(x)\mathrm{d}x + k_2\int_{-\infty}^{\infty} g_2(x)f(x)\mathrm{d}x\\ &= k_1\mathrm{E}[(g_1(X)] + k_2\mathrm{E}[(g_2(X)].\end{aligned}$$

■

La Proprietà (c) può essere estesa a più di due funzioni. Infatti, se $k_1, k_2, \ldots, k_m$ sono costanti e $g_1(\cdot), g_2(\cdot), \ldots, g_m(\cdot)$ sono funzioni, si ha

$$\begin{aligned}\mathrm{E}[k_1g_1(X) + k_2g_2(X) + \cdots + k_mg_m(X)] &= k_1\mathrm{E}[(g_1(X)] + k_2\mathrm{E}[(g_2(X)] + \cdots\\ &\quad + k_m\mathrm{E}[(g_m(X)].\end{aligned}$$

Valore atteso e varianza di $Y = a + bX$

Le proprietà enunciate nella Proposizione B.8 si possono usare per trovare la media e la varianza della variabile casuale $Y = a + bX$, dove a e b sono costanti. Si vuole dimostrare che

$$\mathrm{E}(a + bX) = a + b\mathrm{E}(X) \quad \text{e} \quad \mathrm{Var}(a + bX) = b^2\mathrm{Var}(X).$$

La prima relazione segue direttamente dalla proprietà (c) della Proposizione B.8 ponendo $k_1 = a$, $k_2 = b$, $g_1(X) = 1$ e $g_2(X) = X$.

Per quanto riguarda la seconda, si ha

$$\begin{aligned}\mathrm{Var}(a + bX) &= \mathrm{E}\{[a + bX - \mathrm{E}(a + bX)]^2\} = \mathrm{E}\{[a + bX - a - b\mathrm{E}(X)]^2\}\\ &= \mathrm{E}\{b^2[X - \mathrm{E}(X)]^2\} = b^2\mathrm{E}\{[X - \mathrm{E}(X)]^2\} = b^2\mathrm{Var}(X).\end{aligned}$$

Formula alternativa per Var(X)

Indicando con μ la media della variabile casuale X, si può mostrare che la varianza di X può essere scritta nella forma

$$\text{Var}(X) = \text{E}(X^2) - \mu^2.$$

Infatti, la varianza è il valore atteso della funzione $g(X) = (X-\mu)^2 = X^2-2\mu X+\mu^2$. Conseguentemente,

$$\text{Var}(X) = \text{E}[(X-\mu)^2] = \text{E}(X^2) - 2\mu\text{E}(X) + \mu^2 = \text{E}(X^2) - 2\mu\text{E}(X) + \mu^2,$$

da cui segue il risultato.

B.6 Disuguaglianza di Chebyshev

La dimostrazione è limitata al caso discreto. Si consideri la Figura B.3 in cui è rappresentato l'insieme, A, di valori di X per cui vale la disuguaglianza $|X-\mu| < \delta$.

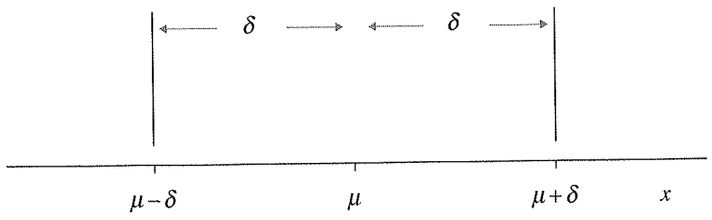


Figura B.3
Valori della variabile casuale X tali che $|X-\mu| < \delta$.

Indicando con $\bar{A}$ il complemento dell'insieme A, contenente i valori di X per i quali vale la disuguaglianza $|X-\mu| \geq \delta$, la varianza di X può essere scritta come

$$\sigma^2 = \sum_x (x-\mu)^2 f(x) = \sum_{x\in A} (x-\mu)^2 f(x) + \sum_{x\in\bar{A}} (x-\mu)^2 f(x).$$

Poiché le due somme nell'ultimo nell'ultima espressione di σ^2 sono entrambe non negative, è possibile scrivere la disuguaglianza

$$\sigma^2 \geq \sum_{x\in\bar{A}} (x-\mu)^2 f(x).$$

Inoltre, poiché i valori di X in $\bar{A}$ soddisfano la disuguaglianza $(x-\mu)^2 \geq \delta^2$, si ha

$$\sigma^2 \geq \sum_{x\in\bar{A}} (x-\mu)^2 f(x) \geq \sum_{x\in\bar{A}} \delta^2 f(x) = \delta^2 \sum_{x\in\bar{A}} f(x),$$

dalla quale si ottiene

$$\sum_{x\in\bar{A}} f(x) = P(|X-\mu| \geq \delta) \leq \frac{\sigma^2}{\delta^2},$$

come volevasi dimostrare.

B.7 Disuguaglianza di Markov

Per dimostrare la disuguaglianza di Markov per una variabile casuale discreta X con funzione di probabilità $f(x)$, si fa osservare che, per ogni $\delta > 0$ e per ogni funzione $g(\cdot)$ non negativa, il valore atteso di $g(X)$ può essere scritto nella forma

$$\text{E}[g(X)] = \sum_x g(x)f(x) = \sum_{x:\, g(x)<\delta} g(x)f(x) + \sum_{x:\, g(x)\geq\delta} g(x)f(x).$$

Poiché le due somme dell'ultima espressione di E[g(X)] sono entrambe non negative, si può scrivere

$$\text{E}[g(X)] \geq \sum_{x:\, g(x)\geq\delta} g(x)f(x) \geq \delta \sum_{x:\, g(x)\geq\delta} f(x) = \delta P[g(X) \geq \delta].$$

da cui

$$\text{E}[g(X)] \geq \delta P[g(X) \geq \delta]$$

e quindi

$$P[g(X) \geq \delta] \leq \frac{\text{E}[g(X)]}{\delta},$$

come volevasi dimostrare.

B.8 Distribuzioni di probabilità

Distribuzione binomiale

Per determinare la funzione di probabilità binomiale, si considerino n prove indipendenti di un esperimento Bernoulliano, in cui l'evento successo A ha probabilità p. I risultati possibili delle n prove sono esposti nella Tabella B.1.

Evento elementare	N. di successi x	Probabilità
$\bar{A}\cap\bar{A}\cap\bar{A}\cap\cdots\cap\bar{A}\cap\bar{A}$	0	$p^0(1-p)^n$
$A\cap\bar{A}\cap\bar{A}\cap\cdots\cap\bar{A}\cap\bar{A}$	1	$p^1(1-p)^{n-1}$
$\bar{A}\cap A\cap\bar{A}\cap\cdots\cap\bar{A}\cap\bar{A}$	1	$p^1(1-p)^{n-1}$
$\bar{A}\cap\bar{A}\cap A\cap\cdots\cap\bar{A}\cap\bar{A}$	1	$p^1(1-p)^{n-1}$
$\vdots$	$\vdots$	$\vdots$
$\bar{A}\cap\bar{A}\cap\bar{A}\cap\cdots\cap\bar{A}\cap A$	1	$p^1(1-p)^{n-1}$
$A\cap A\cap\bar{A}\cap\cdots\cap\bar{A}\cap\bar{A}$	2	$p^2(1-p)^{n-2}$
$\bar{A}\cap A\cap A\cap\cdots\cap\bar{A}\cap\bar{A}$	2	$p^2(1-p)^{n-2}$
$A\cap\bar{A}\cap A\cap\cdots\cap\bar{A}\cap\bar{A}$	2	$p^2(1-p)^{n-2}$
$\vdots$	$\vdots$	$\vdots$
$\bar{A}\cap\bar{A}\cap\bar{A}\cap\cdots\cap A\cap A$	2	$p^2(1-p)^{n-2}$
$\vdots$	$\vdots$	$\vdots$
$A\cap A\cap A\cap\cdots\cap A\cap A$	n	$p^n(1-p)^0$

Tabella B.1
Risultati e rispettive probabilità di n prove bernoulliane.

Se si concentra l'attenzione sul numero x di successi, si nota che: l'evento $x = 0$ ha una sola modalità di presentarsi (si deve verificare $\bar{A}$ in tutte le n prove); l'evento $x = 1$ si può presentare in n modi diversi (i modi di collocare A tra le n posizioni disponibili: A si può verificare alla prima, alla seconda, ..., o alla n-esima prova); l'evento $x = 2$ può realizzarsi in $n!/[2!(n-2)!]$ modi diversi (i modi in cui si possono collocare due A tra le n posizioni disponibili: A può presentarsi alla prima e alla seconda prova, o alla prima e alla terza, ..., o alla $(n-1)$-esima e all'n-esima). Procedendo in modo analogo, si trova infine che l'evento $x = n$ si può presentare in un unico modo (si deve verificare A in tutte le n prove).

In generale, il numero di modi in cui si può avere x volte A in n prove è

$$\binom{n}{x} = \frac{n!}{x!(n-x)!},$$

che non è altro che il numero di combinazioni di n elementi presi a x a x. Poiché ogni evento elementare con x successi (cioè con x volte A) e $n-x$ insuccessi (cioè con $n-x$ volte $\bar{A}$) ha probabilità

$$p^x(1-p)^{n-x},$$

la probabilità complessiva di x successi in n prove bernoulliane indipendenti è

$$P(X = x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}.$$

La (14.1) è quindi dimostrata.

Valore atteso e varianza della distribuzione binomiale

La media della distribuzione binomiale si ottiene come segue

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{x=0}^{n} x\,\frac{n!}{x!(n-x)!}\,p^x(1-p)^{n-x} \\
&= \sum_{x=1}^{n} x\,\frac{n(n-1)!}{x(x-1)![n-1-(x-1)]!}\,p\cdot p^{x-1}(1-p)^{[n-1-(x-1)]} \\
&= np\sum_{y=0}^{n-1}\frac{(n-1)!}{y!(n-1-y)!}\,p^y(1-p)^{n-1-y} \\
&= np[p+(1-p)]^{n-1} = np,
\end{aligned}$$

dove, nel passaggio dalla seconda alla terza riga, si è posto $y = x-1$, e, nel passaggio dalla terza alla quarta, è stata applicata la regola dello sviluppo del binomio di Newton. La prima delle (14.2) è così dimostrata.

Per quanto riguarda la varianza, si ha

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{x=0}^{n} x^2\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} - n^2p^2 \\
&= \sum_{x=1}^{n}[x(x-1)+x]\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} - n^2p^2 \\
&= \sum_{x=1}^{n} x(x-1)\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} + \sum_{x=1}^{n} x\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} - n^2p^2 \\
&= \sum_{x=2}^{n} x(x-1)\frac{n(n-1)(n-2)!}{x(x-1)(x-2)![n-2-(x-2)]!}\,p^2\cdot p^{x-2}(1-p)^{[n-2-(x-2)]} \\
&\quad + np - n^2p^2.
\end{aligned}$$

Portando $n(n-1)p^2$ fuori dalla sommatoria e ponendo $y = x-2$, si ottiene

$$\mathrm{Var}(X) = n(n-1)p^2\sum_{y=0}^{n-2}\frac{(n-2)!}{y!(n-2-y)!}p^y(1-p)^{n-2-y} + np - n^2p^2,$$

dove, per la regola dello sviluppo del binomio di Newton,

$$\sum_{y=0}^{n-2}\frac{(n-2)!}{y!(n-2-y)!}p^y(1-p)^{n-2-y} = [p+(1-p)]^{n-2} = 1.$$

Sostituendo questo risultato nell'ultima espressione di $\mathrm{Var}(X)$, si ricava

$$\mathrm{Var}(X) = np(1-p).$$

La seconda delle (14.2) è così dimostrata.

Valore atteso e varianza della distribuzione di Poisson

Per il valore atteso della distribuzione di Poisson si può scrivere

$$\mathrm{E}(X) = \sum_{x=0}^{\infty} x\frac{\lambda^x}{x!}e^{-\lambda} = e^{-\lambda}\sum_{x=1}^{\infty} x\frac{\lambda\cdot\lambda^{x-1}}{x(x-1)!} = \lambda e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!},$$

da cui, posto $y = x-1$, si ottiene

$$\mathrm{E}(X) = \lambda e^{-\lambda}\sum_{y=0}^{\infty}\frac{\lambda^y}{y!} = \lambda\lambda e^{-\lambda}\lambda e^{\lambda} = \lambda,$$

dove è stata sfruttata la relazione

$$\sum_{y=0}^{\infty}\frac{\lambda^y}{y!} = e^{\lambda}.$$

La prima delle (14.3) è quindi dimostrata.

Quanto alla varianza, si può scrivere

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{x=0}^{\infty} x^2\frac{\lambda^x}{x!}e^{-\lambda} - \lambda^2 \\
&= \sum_{x=0}^{\infty}[x(x-1)+x]\frac{\lambda^x}{x!}e^{-\lambda} - \lambda^2 \\
&= e^{-\lambda}\sum_{x=2}^{\infty} x(x-1)\frac{\lambda^2\cdot\lambda^{x-2}}{x(x-1)(x-2)!} + \sum_{x=0}^{\infty} x\frac{\lambda^x}{x!} - \lambda^2 \\
&= \lambda^2e^{-\lambda}\sum_{x=2}^{\infty}\frac{\lambda^{x-2}}{(x-2)!} + \lambda - \lambda^2,
\end{aligned}$$

da cui, posto $y = x-2$, si ottiene

$$\text{Var}(X) = \lambda^2 e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} + \lambda - \lambda^2 = \lambda^2 e^{-\lambda} e^{\lambda} + \lambda - \lambda^2 = \lambda,$$

dove è stata sfruttata la relazione

$$\sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = e^{\lambda}.$$

La seconda delle (14.3) è così dimostrata.

Valore atteso e varianza della distribuzione chi-quadrato

Per trovare la media e la varianza della distribuzione chi-quadrato, bisogna specificare la costante di normalizzazione a che appare nella (14.9). Si ha allora

$$f(x) = \begin{cases} \dfrac{1}{2^{r/2}\Gamma\left(\frac{r}{2}\right)} x^{\frac{r}{2}-1} e^{-x/2}, & x \geq 0, \\ 0, & \text{altrove}, \end{cases}$$

dove $\Gamma(\cdot)$ è la funzione gamma.

La media della distribuzione chi-quadrato è data da

$$\text{E}(X) = \int_0^{\infty} x \frac{1}{2^{r/2}\Gamma\left(\frac{r}{2}\right)} x^{\frac{r}{2}-1} e^{-x/2}\, \text{d}x = r \int_0^{\infty} \frac{1}{2^{(r+2)/2}\Gamma\left(\frac{r+2}{2}\right)} x^{\frac{(r+2)}{2}-1} e^{-x/2}\, \text{d}x = r,$$

in quanto $\Gamma((r+2)/2) = (r/2)\Gamma(r/2)$, e l'ultimo integrale è pari a 1, essendo l'area sottesa alla funzione di densità chi-quadrato con $r+2$ gradi di libertà.

Analogamente, si ha

$$\begin{aligned} \text{E}(X^2) &= \int_0^{\infty} x^2 \frac{1}{2^{r/2}\Gamma\left(\frac{r}{2}\right)} x^{\frac{r}{2}-1} e^{-x/2}\, \text{d}x \\ &= 4\left(\frac{r}{2}\right)\left(\frac{r+2}{2}\right) \int_0^{\infty} \frac{1}{2^{(r+4)/2}\Gamma\left(\frac{r+4}{2}\right)} x^{\frac{(r+4)}{2}-1} e^{-x/2}\, \text{d}x = r^2 + 2r, \end{aligned}$$

in quanto $\Gamma((r+4)/2) = [(r+2)/2](r/2)\Gamma(r/2)$, e l'ultimo integrale è pari a 1, essendo l'area sottesa alla funzione di densità chi-quadrato con $r+4$ gradi di libertà. Se ne deduce che

$$\text{Var}(X) = \text{E}(X^2) - [\text{E}(X)]^2 = r^2 + 2r - r^2 = 2r,$$

come volevasi dimostrare.

B.9 Valore atteso di combinazioni lineari di funzioni di due o più variabili casuali

PROPOSIZIONE B.9

Siano date due variabili casuali X_1 e X_2 con funzione di probabilità o di densità congiunta $f(x_1, x_2)$. Siano date inoltre m costanti, $k_1, k_2, \ldots, k_m$, e m funzioni, $g_1(X_1, X_2), g_2(X_1, X_2), \ldots, g_m(X_1, X_2)$ delle variabili X_1 e X_2. Allora

$$\text{E}\left[\sum_{i=1}^{m} k_i g_i(X_1, X_2)\right] = \sum_{i=1}^{m} k_i \text{E}[g_i(X_1, X_2)]. \qquad \text{(B.8)}$$

Si limiterà la dimostrazione al caso particolare in cui le variabili casuali sono discrete e $m = 2$. Si ha

$$\begin{aligned} \text{E}[k_1 g_1(X_1, X_2) + k_2 g_2(X_1, X_2)] &= \sum_{x_1} \sum_{x_2} [k_1 g_1(x_1, x_2) + k_2 g_2(x_1, x_2)] f(x_1, x_2) \\ &= k_1 \sum_{x_1} \sum_{x_2} g_1(x_1, x_2) f(x_1, x_2) \\ &\quad + k_2 \sum_{x_1} \sum_{x_2} g_2(x_1, x_2) f(x_1, x_2), \end{aligned}$$

da cui

$$\text{E}[k_1 g_1(X_1, X_2) + k_2 g_2(X_1, X_2)] = k_1 \text{E}[g_1(x_1, x_2)] + k_2 \text{E}[g_2(x_1, x_2)].$$

La Proposizione B.9 può esssere estesa a più di due variabili casuali. Pertanto, se $g_1(\cdot), g_2(\cdot), \ldots, g_m(\cdot)$ sono funzioni di n variabili casuali (discrete o continue), si ha

$$\text{E}\left[\sum_{i=1}^{m} k_i g_i(X_1, X_2, \ldots, X_n)\right] = \sum_{i=1}^{m} k_i \text{E}[g_i(X_1, X_2, \ldots, X_n)]. \qquad \text{(B.9)}$$

Valore atteso e varianza di combinazioni lineari di variabili casuali

Siano date due variabili casuali X_1 e X_2 con valore atteso, rispettivamente, μ_{X_1} e μ_{X_2}. Si consideri la combinazione lineare di X_1 e X_2

$$W = a_1 X_1 + a_2 X_2,$$

dove a_1 e a_2 sono costanti arbitrarie. Utilizzando la Proposizione B.9, si possono facilmente trovare il valore atteso e la varianza di W.

Per quanto riguarda il valore atteso, ponendo $g_1(X_1, X_2) = X_1$ e $g_2(X_1, X_2) = X_2$, si ha

$$\text{E}(W) = \text{E}(a_1 X_1 + a_2 X_2) = a_1 \text{E}(X_1) + a_2 \text{E}(X_2) = a_1 \mu_{X_1} + a_2 \mu_{X_2}.$$

Per quanto riguarda la varianza, si può scrivere

$$\begin{aligned} \text{Var}(W) &= \text{E}(a_1 X_1 + a_2 X_2 - a_1 \mu_{X_1} - a_2 \mu_{X_2})^2 \\ &= \text{E}[a_1 (X_1 - \mu_{X_1}) + a_2 (X_2 - \mu_{X_2})]^2 \\ &= \text{E}[a_1^2 (X_1 - \mu_{X_1})^2 + a_2^2 (X_2 - \mu_{X_2})^2 \\ &\quad + 2 a_1 a_2 (X_1 - \mu_{X_1})(X_2 - \mu_{X_2})], \end{aligned}$$

da cui, applicando la (B.8) con $m = 3$, $k_1 = a_1^2$, $k_2 = a_2^2$, $k_3 = 2a_1 a_2$, $g_1(X_1, X_2) = (X_1 - \mu_{X_1})^2$, $g_2(X_1, X_2) = (X_2 - \mu_{X_2})^2$ e $g_3(X_1, X_2) = (X_1 - \mu_{X_1})(X_2 - \mu_{X_2})$, si ottiene

$$\begin{aligned} \text{Var}(W) &= a_1^2 \text{E}[(X_1 - \mu_{X_1})^2 + a_2^2 \text{E}(X_2 - \mu_{X_2})^2 \\ &\quad + 2 a_1 a_2 \text{E}[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})], \end{aligned}$$

dove i tre addendi a destra dell'uguale sono: la varianza di X_1 moltiplicata per a_1^2, la varianza di X_2 moltiplicata per a_2^2, la covarianza tra X_1 e X_2 moltiplicata per $2a_1 a_2$. In definitiva, indicando con $\sigma_{X_1}^2$ la varianza di X_1, con $\sigma_{X_2}^2$ la varianza di

X_2 e con $\sigma_{X_1X_2}$ la covarianza tra X_1 e X_2, si può scrivere

$$\text{Var}(W) = a_1^2\sigma_{X_1}^2 + a_2^2\sigma_{X_2}^2 + 2a_1a_2\sigma_{X_1X_2}.$$

Le (15.8) e (15.9) sono perciò dimostrate.

Utilizzando la (B.9), con lo stesso procedimento descritto sopra, si possono determinare il valore atteso e la varianza di una combinazione lineare di n variabili casuali,

$$W = a_1X_1 + a_2X_2 + \cdots + a_nX_n,$$

ottenendo:

$$\text{E}(W) = \sum_{i=1}^{n} a_i\text{E}(X_i),$$

$$\text{Var}(W) = \sum_{i=1}^{n} a_i^2\sigma_{X_i}^2 + \sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n} a_ia_j\sigma_{X_iX_j},$$

dove $\sigma_{X_i}^2$ è la varianza di X_i e $\sigma_{X_iX_j}$ è la covarianza tra X_i e X_j.

B.10 Argomenti di stima puntuale

Non distorsione della varianza campionaria S^2

Si osservi innanzitutto che la varianza campionaria può essere scritta come:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \frac{n}{n-1}\bar{X}^2.$$

Utilizzando le proprietà del valore atteso (si veda il Paragrafo B.9), si ha

$$\text{E}(S^2) = \text{E}\left(\frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \frac{n}{n-1}\bar{X}^2\right) = \frac{1}{n-1}\left[\sum_{i=1}^{n}\text{E}(X_i^2) - n\text{E}(\bar{X}^2)\right]. \quad \text{(B.10)}$$

Poiché per una qualsiasi variabile casuale X vale la relazione

$$\text{E}(X^2) = \text{Var}(X) + [\text{E}(X)]^2,$$

i valori attesi di X_i^2 e di $\bar{X}^2$ possono essere scritti come segue:

$$\text{E}(X_i^2) = \text{Var}(X_i) + [\text{E}(X_i)]^2 = \sigma^2 + \mu^2,$$

$$\text{E}(\bar{X}^2) = \text{Var}(\bar{X}) + [\text{E}(\bar{X})]^2 = \frac{\sigma^2}{n} + \mu^2.$$

Sostituendo questi risultati nell'ultimo termine della relazione (B.10), con qualche passaggio, si ottiene

$$\text{E}(S^2) = \sigma^2.$$

La prima delle (17.6) è quindi dimostrata.

Consistenza in media quadratica e consistenza in probabilità

Si vuole ora dimostrare che, se uno stimatore è consistente in media quadratica, allora esso è anche consistente in probabilità.

Sia $T_n = t(X_1, X_2, \ldots, X_n)$ uno stimatore consistente in media quadratica del parametro θ. Poiché

$$P(|T_n - \theta| < \varepsilon) = P(|T_n - \theta|^2 < \varepsilon^2),$$

applicando la disuguaglianza di Markov alla funzione $g(T_n) = (T_n - \theta)^2$, ponendo $\delta = \varepsilon^2$, si può scrivere

$$P(|T_n - \theta| < \varepsilon) \geq 1 - \frac{\text{E}(T_n - \theta)^2}{\varepsilon^2} = 1 - \frac{\text{MSE}(T_n)}{\varepsilon^2},$$

da cui, prendendo il limite per $n \to \infty$, si ottiene

$$\lim_{n\to\infty} P(|T_n - \theta| < \varepsilon) \geq \lim_{n\to\infty}\left(1 - \frac{\text{MSE}(T_n)}{\varepsilon^2}\right)$$

e quindi

$$\lim_{n\to\infty} P(|T_n - \theta| < \varepsilon) = 1,$$

quando $\lim_{n\to\infty}\text{MSE}(T_n) = 0$, cioè quando T_n è consistente in media quadratica, come volevasi dimostrare.

Valore atteso e varianza degli stimatori dei minimi quadrati B_0 e B_1

Si inizi con il considerare il valore atteso di B_1. Applicando la formula del valore atteso di una combinazione lineare di variabili casuali, si può scrivere

$$\text{E}(B_1) = \text{E}\left[\frac{1}{D_X}\sum_{i=1}^{n}(x_i - \bar{x})Y_i\right] = \frac{1}{D_X}\sum_{i=1}^{n}(x_i - \bar{x})\text{E}(Y_i),$$

dove, in base all'Equazione (23.2), $\text{E}(Y_i) = \beta_0 + \beta_1x_i$. Ne segue che

$$\begin{aligned}\text{E}(B_1) &= \frac{1}{D_X}\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1x_i)\\ &= \frac{\beta_0}{D_X}\sum_{i=1}^{n}(x_i - \bar{x}) + \frac{\beta_1}{D_X}\sum_{i=1}^{n}(x_i - \bar{x})x_i,\end{aligned}$$

dove

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \text{ e } \sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = D_X.$$

Si ha allora

$$\text{E}(B_1) = \beta_1.$$

Ciò dimostra che B_1 è uno stimatore non distorto di β_1.

Quanto al valore atteso di B_0, si ha

$$\text{E}(B_0) = \text{E}(\bar{Y} - B_1\bar{x}) = \text{E}(\bar{Y}) - \bar{x}\text{E}(B_1).$$

Poiché, come è facile verificare,

$$\text{E}(\bar{Y}) = \beta_0 + \beta_1\bar{x},$$

si ha

$$\text{E}(B_0) = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0.$$

Si è così dimostrato che B_0 è uno stimatore non distorto di β_0.

Per la determinazione della varianza degli stimatori B_0 e B_1, si ricorda che questi sono particolari combinazioni lineari delle variabili casuali $Y_1, Y_2, \ldots, Y_n$, che sono, per ipotesi, indipendenti e omoschedastiche (cioè hanno tutte la stessa varianza σ^2). Lo stimatore B_0 può essere posto nella forma

$$B_0 = \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{D_X} \right] Y_i,$$

la varianza la sua varianza è data da

$$\mathrm{Var}(B_0) = \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{D_X} \right]^2 \sigma^2_{Y_i},$$

da cui, ponendo $\sigma^2_{Y_i} = \sigma^2$, $i = 1, 2, \ldots, n$, si ottiene

$$\mathrm{Var}(B_0) = \sigma^2 \sum_{i=1}^{n} \left[\frac{1}{n^2} + \frac{(x_i - \bar{x})^2 \bar{x}^2}{D_X^2} - \frac{2}{n} \frac{(x_i - \bar{x})\bar{x}}{D_X} \right],$$

che, dopo qualche passaggio algebrico, diventa

$$\mathrm{Var}(B_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{D_X} \right),$$

come volevasi dimostrare.

Quanto alla varianza di B_1, potendosi scrivere

$$\mathrm{Var}(B_1) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{D_X^2} \sigma^2_{Y_i},$$

e quindi ponendo $\sigma^2_{Y_i} = \sigma^2$, $i = 1, 2, \ldots, n$, si ottiene immediatamente

$$\mathrm{Var}(B_1) = \frac{\sigma^2}{D_X},$$

come volevasi dimostrare.

Non distorsione di $\sum_{i=1}^{n} e_i^2/(n-2)$ nella regressione lineare semplice

Conviene riscrivere la somma dei quadrati dei residui nella (23.7) in modo diverso:

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 - B_1^2 D_X.$$

Ne segue che

$$\begin{aligned} \mathrm{E}\left[\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \right] &= \sum_{i=1}^{n} \mathrm{E}(Y_i^2) - n\mathrm{E}(\bar{Y}^2) - D_X \mathrm{E}(B_1^2) \\ &= \sum_{i=1}^{n} \{\mathrm{Var}(Y_i) + [\mathrm{E}(Y_i)]^2\} - n\{\mathrm{Var}(\bar{Y}) + [\mathrm{E}(\bar{Y})]^2\} \\ &\quad - D_X\{\mathrm{Var}(B_1) + [\mathrm{E}(B_1)]^2\}, \end{aligned}$$

dove, nella seconda riga, i valori attesi dei quadrati delle variabili casuali Y_i, $\bar{Y}$ e B_1 sono stati espressi in termini delle loro varianze. Infine, sostituendo i valori attesi e le varianze, che compaiono nell'ultimo termine della relazione precedente, con le loro espressioni esplicite, che sono:

$$\mathrm{Var}(Y_i) = \sigma^2, \quad \mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i, \quad \mathrm{Var}(\bar{Y}) = \sigma^2/n, \quad \mathrm{E}(\bar{Y}) = \beta_0 + \beta_1 \bar{x},$$
$$\mathrm{Var}(B_1) = \sigma^2/D_X, \quad \mathrm{E}(B_1) = \beta_1,$$

dopo qualche passaggio algebrico, si ottiene

$$\mathrm{E}\left[\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \right] = (n-2)\sigma^2.$$

Ovviamente, ciò implica che $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2/(n-2)$ è uno stimatore non distorto di σ^2, come volevasi dimostrare.

Varianza dello stimatore $\hat{Y}_0$ nella regressione lineare semplice

Si ricordi che $\hat{Y}_0$ può essere scritto come combinazione lineare delle variabili casuali $Y_1, Y_2, \ldots, Y_n$ (si veda la Nota 2 del Capitolo 23), che, per ipotesi, sono incorrelate e con la stessa varianza:

$$\hat{Y}_0 = \sum_{i=1}^{n} \left[\frac{1}{n} + \frac{(x_i - \bar{x})}{D_X} (x_0 - \bar{x}) \right] Y_i.$$

L'applicazione dell'usuale regola per la determinazione della varianza di una combinazione lineare di variabili casuali dà

$$\mathrm{Var}(\hat{Y}_0) = \sum_{i=1}^{n} \left[\frac{1}{n} + \frac{(x_i - \bar{x})}{D_X} (x_0 - \bar{x}) \right]^2 \sigma^2_{Y_i},$$

da cui, ponendo $\sigma^2_{Y_i} = \sigma^2$, si ricava

$$\mathrm{Var}(\hat{Y}_0) = \sigma^2 \sum_{i=1}^{n} \left[\frac{1}{n^2} + \frac{(x_i - \bar{x})^2}{D_X^2} (x_0 - \bar{x})^2 + \frac{2}{n} \frac{(x_i - \bar{x})}{D_X} (x_0 - \bar{x}) \right].$$

Da qui, con facili passaggi algebrici, si ottiene

$$\mathrm{Var}(\hat{Y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{D_X} \right],$$

che è il risultato desiderato.

B.11 Inferenza su popolazioni finite

Probabilità di inclusione

Per un dato piano di campionamento, la probabilità π_i che l'unità i della popolazione sia inclusa nel campione – in breve la probabilità di inclusione – è uguale alla somma delle probabilità dei campioni che la contengono; la somma è ovviamente estesa all'insieme $\mathbb{C}$ di tutti i campioni possibili:

$$\pi_i = \sum_{c \in \mathbb{C}} \delta_i p(c) = \mathrm{E}(\delta_i),$$

dove δ_i è la variabile casuale indicatrice che assume il valore 1 se l'unità i della popolazione è presente nel campione e il valore 0 nel caso contrario.

Analogamente, per un dato piano di campionamento, la probabilità π_{ij} che la coppia di unità i e j della popolazione sia inclusa nel campione – in breve la probabilità di inclusione della coppia di unità i e j – è uguale alla somma delle probabilità dei campioni che contengono entrambe le unità:

$$\pi_{ij} = \sum_{c \in \mathbb{C}} \delta_i \delta_j p(c) = \mathrm{E}(\delta_i \delta_j),$$

con ovvio significato delle variabili indicatrici δ_i e δ_j.

Non distorsione dello stimatore $\hat{\mu}$

Si dimostra innanzitutto la non distorsione dello stimatore $\hat{\mu}$,

$$\hat{\mu} = \frac{1}{N} \sum_{i \in c} \frac{x_i}{\pi_i}. \tag{B.11}$$

A questo fine, si riscrive la formula precedente nella forma

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \frac{x_i}{\pi_i} \delta_i, \tag{B.12}$$

dove δ_i è la variabile casuale indicatrice definita sopra. Scritta in questa forma, $\hat{\mu}$ è una combinazione lineare delle variabili casuali $\delta_1, \delta_2, \ldots, \delta_N$ con coefficienti $x_1/\pi_1, x_2/\pi_2, \ldots, x_N/\pi_N$, il cui valore atteso è uguale alla combinazione lineare dei valori attesi di $\delta_1, \delta_2, \ldots, \delta_N$ (si veda la Proposizione 15.3):

$$\mathrm{E}(\hat{\mu}) = \frac{1}{N} \sum_{i=1}^{N} \frac{x_i}{\pi_i} \mathrm{E}(\delta_i) = \sum_{i=1}^{N} \frac{x_i}{\pi_i} \pi_i,$$

essendo $\mathrm{E}(\delta_i) = \pi_i$ per definizione. In definitiva, si ha

$$\mathrm{E}(\hat{\mu}) = \frac{1}{N} \sum_{i=1}^{N} x_i = \mu,$$

il che dimostra che lo stimatore $\hat{\mu}$ gode della proprietà della non distorsione.

Variabilità di campionamento di $\hat{\mu}$

Si vuole dimostrare che la variabilità di campionamento dello stimatore di $\hat{\mu}$ è data da

$$\sigma^2_{\hat{\mu}} = \frac{1}{N^2} \sum_{i=1}^{N} \frac{1 - \pi_i}{\pi_i} x_i^2 + \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) x_i x_j. \tag{B.13}$$

La dimostrazione della (B.13) è immediata se si applica la formula (15.6) (varianza di una combinazione lineare di variabili casuali) alla (B.12):

$$\sigma^2_{\hat{\mu}} = \frac{1}{N^2} \sum_{i=1}^{N} \frac{x_i^2}{\pi_i^2} \sigma^2_{\delta_i} + \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \frac{x_i}{\pi_i} \frac{x_j}{\pi_j} \sigma_{\delta_i \delta_j}.$$

Ora $\sigma^2_{\delta_i} = \pi_i(1-\pi_i)$, perché δ_i è una variabile casuale Bernoulliana che assume il valore 1 con probabilità π_i e il valore 0 con probabilità $1 - \pi_i$, e la varianza di questa variabile casuale è $\pi_i(1 - \pi_i)$. Si ha inoltre

$$\sigma_{\delta_i \delta_j} = \mathrm{E}[(\delta_i - \pi_i)(\delta_j - \pi_j)] = \mathrm{E}(\delta_i \delta_j) - \pi_i \pi_j = \pi_{ij} - \pi_i \pi_j.$$

La sostituzione di questi due risultati nell'ultima espressione di $\sigma^2_{\hat{\mu}}$ dà subito la (B.13).

Stimatore non distorto della variabilità di campionamento di $\hat{\mu}$

Si vuole dimostrare che lo stimatore non distorto della variabilità di campionamento di $\hat{\mu}$ è data da

$$\hat{\sigma}^2_{\hat{\mu}} = \frac{1}{N^2} \sum_{i \in c} \frac{1 - \pi_i}{\pi_i^2} x_i^2 + \frac{1}{N^2} \sum_{\substack{i \in c \\ i \neq j}} \sum_{j \in c} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} \right) x_i x_j. \tag{B.14}$$

A questo fine, si riscriva $\hat{\sigma}^2_{\hat{\mu}}$ come segue

$$\hat{\sigma}^2_{\hat{\mu}} = \frac{1}{N^2} \sum_{i=1}^{N} \frac{1 - \pi_i}{\pi_i^2} x_i^2 \delta_i + \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} \right) x_i x_j \delta_i \delta_j$$

e se ne prenda il valore atteso. Si ha

$$\mathrm{E}(\hat{\sigma}^2_{\hat{\mu}}) = \frac{1}{N^2} \sum_{i=1}^{N} \frac{1 - \pi_i}{\pi_i^2} x_i^2 \mathrm{E}(\delta_i) + \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} \right) x_i x_j \mathrm{E}(\delta_i \delta_j),$$

da cui, essendo $\mathrm{E}(\delta_i) = \pi_i$ e $\mathrm{E}(\delta_i \delta_j) = \pi_{ij}$, si perviene immediatamente alla (B.14). Si osservi che la (B.14) è un modo equivalente di scrittura della (24.2): in essa vengono scritti con una sommatoria distinta gli addendi in cui $i = j$, nel qual caso $\pi_i = \pi_j$ e $\pi_{ij} = \pi_i$.

Probabilità di inclusione nel campionamento casuale semplice

Nel campionamento casuale semplice, i campioni dello spazio dei campioni sono in numero pari alle combinazioni che si possono formare con le N oggetti prendendone n alla volta:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}.$$

Ne segue che

$$\pi_i = \sum_{c \in \mathbb{C}} \delta_i p(c)$$

La probabilità del campione $c = \{i_1, i_2, \ldots, i_n\}$ è data da

$$p(c) = \frac{n}{N} \cdot \frac{n-1}{N-1} \cdot \ldots \cdot \frac{1}{N-n+1} = \frac{1}{\binom{N}{n}}.$$

Infatti, la probabilità che nella prima estrazione venga selezionata una qualsiasi delle n unità del campione è n/N, la probabilità che nella seconda estrazione venga

selezionata una qualsiasi delle $n-1$ rimanenti unità del campione è $(n-1)/(N-1)$ e così via. Dunque, tutti i campioni hanno la medesima probabilità.

La probabilità di inclusione di una specifica unità, i, della popolazione è

$$\pi_i = \sum_{c \in \mathbb{C}} \delta_i p(c) = \frac{1}{\binom{N}{n}} \sum_{c \in \mathbb{C}} \delta_i.$$

Poiché la somma $\sum_{c \in \mathbb{C}} \delta_i$ è il numero dei campioni che contengono l'unità i, si tratta di determinare questo numero. Il risultato si raggiunge agevolmente ragionando nel modo seguente. Se l'unità i viene esclusa dalla popolazione, il numero dei campioni casuali semplici di ampiezza $n-1$ formati con le rimanenti $N-1$ unità è pari al numero delle combinazioni di $N-1$ oggetti presi a $n-1$ a $n-1$; ne segue che i campioni di ampiezza n che contengono l'unità i si ottengono dai campioni precedenti aggiungendo questa unità. Il numero di campioni di ampiezza n che contengono l'unità i è allora dato da

$$\binom{N-1}{n-1}.$$

Si ha pertanto

$$\pi_i = \frac{1}{\binom{N}{n}} \cdot \binom{N-1}{n-1} = \frac{N-1)!}{(n-1)!(N-n)!} \cdot \frac{n!(N-n)!}{N!},$$

da cui

$$\pi_i = \frac{n}{N}.$$

Con un analogo ragionamento, si può determinare la probabilità di inclusione della coppia di unità i e j della popolazione. Invero, è facile stabilire che il numero dei campioni che contengono la coppia di unità i e j è pari al numero di combinazioni che si possono formare con $N-2$ elementi presi a $n-2$ a $n-2$. Si può allora scrivere

$$\pi_{ij} = \frac{1}{\binom{N}{n}} \sum_{c \in \mathbb{C}} \delta_i \delta_i = \frac{1}{\binom{N}{n}} \cdot \binom{N-2}{n-2},$$

da cui, con facili passaggi algebrici, si ottiene

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)}.$$

Stimatore non distorto della varianza di campionamento di $\hat{\mu}_{ccs}$

Ponendo nella (B.14) $\pi_i = n/N$ e $\pi_{ij} = n(n-1)/N(N-1)$, si ottiene

$$\hat{\sigma}^2_{\hat{\mu}_{ccs}} = \frac{1}{N^2} \sum_{i \in c} \frac{1-n/N}{n^2/N^2} x_i^2 + \frac{1}{N^2} \sum_{\substack{i \in c \\ i \neq j}} \sum_{j \in c} \left[\frac{n(n-1)/N(N-1) - n^2/N^2}{(n^2/N^2)n(n-1)/N(N-1)} \right] x_i x_j.$$

È facile stabilire che valgono le identità:

$$\frac{1}{N^2} \frac{1-n/N}{n^2/N^2} = \frac{1-n/N}{n^2}, \quad \frac{1}{N^2} \frac{n(n-1)/N(N-1) - n^2/N^2}{(n^2/N^2)n(n-1)/N(N-1)} = -\frac{1}{n^2} \frac{1-n/N}{n-1},$$

$$\sum_{\substack{i \in c \\ i \neq j}} \sum_{j \in c} x_i x_j = \left(\sum_{i \in c} x_i \right)^2 - \sum_{i \in c} x_i^2 = n^2 \bar{x}^2 - \sum_{i \in c} x_i^2.$$

Tenendo conto di queste identità, si può scrivere

$$\hat{\sigma}^2_{\hat{\mu}_{ccs}} = \frac{1}{n^2}(1-n/N) \sum_{i \in c} x_i^2 - \frac{1}{n^2} \frac{1-n/N}{n-1} \left(n^2 \bar{x}^2 - \sum_{i \in c} x_i^2 \right),$$

da cui, dopo qualche passaggio algebrico, si ricava

$$\hat{\sigma}^2_{\hat{\mu}_{ccs}} = (1-f)\frac{s^2}{n}. \tag{B.15}$$

Stimatore non distorto della variabilità di campionamento di $\hat{\mu}_{str}$

Per ottenere la (24.9), è necessario adattare la (24.2) al caso stratificato. Si può allora scrivere

$$\hat{\sigma}^2_{\hat{\mu}_{str}} = \frac{1}{N^2} \sum_{h=1}^{L} \left(\sum_{k \in c_h} \sum_{l \in c_h} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{x_k}{\pi_k} \frac{x_l}{\pi_l} \right),$$

dove c_h è il campione casuale semplice dello strato h. Poiché l'espressione che appare all'interno delle parentesi tonde è lo stimatore non distorto della variabilità di campionamento di $\bar{x}_h$ (media del campione dello strato h) moltiplicato per N_h^2, sulla base della (B.15), si può scrivere

$$\hat{\sigma}^2_{\hat{\mu}_{str}} = \frac{1}{N^2} \sum_{h=1}^{L} \left(N_h^2 (1-f_h) \frac{s_h^2}{n_h} \right),$$

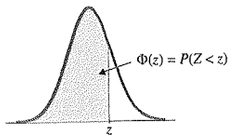
come volevasi dimostrare.

APPENDICE

C

Tavole statistiche

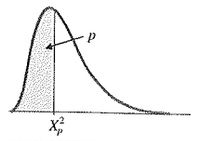
Tavola C.1 Funzione di ripartizione della normale standard.



	Seconda cifra decimale di z									
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

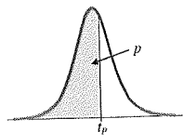
z	1.282	1.645	1.960	2.326	2.576	3.090	3.291	3.891	4.417
$\Phi(z)$	0.90	0.95	0.975	0.99	0.995	0.999	0.9995	0.99995	0.999995

Tavola C.2 Quantili della distribuzione chi-quadrato.



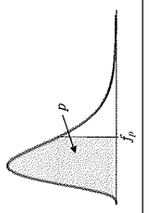
	p													
r	0.005	0.010	0.025	0.050	0.100	0.250	0.500	0.750	0.900	0.950	0.975	0.990	0.995	0.999
1	0.00	0.00	0.00	0.00	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.96	26.12
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.73	26.76	31.26
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	4.60	5.23	6.26	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00	45.31
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67	59.70
31	14.46	15.66	17.54	19.28	21.43	25.39	30.34	35.89	41.42	44.99	48.23	52.19	55.00	61.10
32	15.13	16.36	18.29	20.07	22.27	26.30	31.34	36.97	42.58	46.19	49.48	53.49	56.33	62.49
33	15.82	17.07	19.05	20.87	23.11	27.22	32.34	38.06	43.75	47.40	50.73	54.78	57.65	63.87
34	16.50	17.79	19.81	21.66	23.95	28.14	33.34	39.14	44.90	48.60	51.97	56.06	58.96	65.25
35	17.19	18.51	20.57	22.47	24.80	29.05	34.34	40.22	46.06	49.80	53.20	57.34	60.27	66.62
36	17.89	19.23	21.34	23.27	25.64	29.97	35.34	41.30	47.21	51.00	54.44	58.62	61.58	67.99
37	18.59	19.96	22.11	24.07	26.49	30.89	36.34	42.38	48.36	52.19	55.67	59.89	62.88	69.35
38	19.29	20.69	22.88	24.88	27.34	31.81	37.34	43.46	49.51	53.38	56.90	61.16	64.18	70.70
39	20.00	21.43	23.65	25.70	28.20	32.74	38.34	44.54	50.66	54.57	58.12	62.43	65.48	72.05
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.81	55.76	59.34	63.69	66.77	73.40
45	24.31	25.90	28.37	30.61	33.35	38.29	44.34	50.98	57.51	61.66	65.41	69.96	73.17	80.08
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49	86.66
55	31.73	33.57	36.40	38.96	42.06	47.61	54.33	61.67	68.80	73.31	77.38	82.29	85.75	93.17
60	35.53	37.48	40.48	43.19	46.46	52.29	59.33	66.98	74.40	79.08	83.30	88.38	91.95	99.61
65	39.38	41.44	44.60	47.45	50.88	56.99	64.33	72.28	79.97	84.82	89.18	94.42	98.11	105.99
70	43.28	45.44	48.76	51.74	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.43	104.21	112.32
75	47.21	49.48	52.94	56.05	59.79	66.42	74.33	82.86	91.06	96.22	100.84	106.39	110.29	118.60
80	51.17	53.54	57.15	60.39	64.28	71.14	79.33	88.13	96.58	101.88	106.63	112.33	116.32	124.84
85	55.17	57.63	61.39	64.75	68.78	75.88	84.33	93.39	102.08	107.52	112.39	118.24	122.32	131.04
90	59.20	61.75	65.65	69.13	73.29	80.62	89.33	98.65	107.57	113.15	118.14	124.12	128.30	137.21
95	63.25	65.90	69.92	73.52	77.82	85.38	94.33	103.90	113.04	118.75	123.86	129.97	134.25	143.34
100	67.33	70.06	74.22	77.93	82.36	90.13	99.33	109.14	118.50	124.34	129.56	135.81	140.17	149.45

Tavola C.3 Quantili della distribuzione T di Student.



	p								
r	0.60	0.70	0.80	0.90	0.95	0.975	0.990	0.995	0.999
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.657	318.309
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169	4.144
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750	3.385
31	0.256	0.530	0.853	1.309	1.696	2.040	2.453	2.744	3.375
32	0.255	0.530	0.853	1.309	1.694	2.037	2.449	2.738	3.365
33	0.255	0.530	0.853	1.308	1.692	2.035	2.445	2.733	3.356
34	0.255	0.529	0.852	1.307	1.691	2.032	2.441	2.728	3.348
35	0.255	0.529	0.852	1.306	1.690	2.030	2.438	2.724	3.340
36	0.255	0.529	0.852	1.306	1.688	2.028	2.434	2.719	3.333
37	0.255	0.529	0.851	1.305	1.687	2.026	2.431	2.715	3.326
38	0.255	0.529	0.851	1.304	1.686	2.024	2.429	2.712	3.319
39	0.255	0.529	0.851	1.304	1.685	2.023	2.426	2.708	3.313
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704	3.307
60	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660	3.232
120	0.254	0.526	0.845	1.289	1.658	1.980	2.358	2.617	3.160
∞	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576	3.090

Tavola C.4 Quantili della distribuzione F di Fischer.



		r_1													
p	r_2	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120	∞
0.950	1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.90	245.95	248.02	250.10	252.20	253.25	254.32
0.990		4052.2	4999.3	5403.5	5624.3	5764.0	5859.0	5928.3	5981.0	6022.4	6055.9	6106.7	6157.0	6208.7	6260.4	6313.0	6339.5	6365.6
0.950	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.46	19.48	19.49	19.50
0.990		98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.36	99.39	99.40	99.42	99.43	99.45	99.47	99.48	99.49	99.50
0.950	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.62	8.57	8.55	8.53
0.990		34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.05	26.87	26.69	26.50	26.32	26.22	26.13
0.999		167.06	148.49	141.10	137.08	134.58	132.83	131.61	130.62	129.86	129.22	128.32	127.36	126.43	125.44	124.45	123.98	123.46
0.950	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.75	5.69	5.66	5.63
0.990		21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.84	13.65	13.56	13.46
0.999		74.13	61.25	56.17	53.43	51.72	50.52	49.65	49.00	48.47	48.05	47.41	46.76	46.10	45.43	44.75	44.40	44.05
0.950	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.50	4.43	4.40	4.37
0.990		16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.38	9.20	9.11	9.02
0.999		47.18	37.12	33.20	31.08	29.75	28.83	28.17	27.65	27.24	26.91	26.42	25.91	25.39	24.87	24.33	24.06	23.79
0.950	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81	3.74	3.70	3.67
0.990		13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.23	7.06	6.97	6.88
0.999		35.51	27.00	23.71	21.92	20.80	20.03	19.46	19.03	18.69	18.41	17.99	17.56	17.12	16.67	16.21	15.98	15.75
0.950	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.38	3.30	3.27	3.23
0.990		12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	5.99	5.82	5.74	5.65
0.999		29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33	14.08	13.71	13.32	12.93	12.53	12.12	11.91	11.70
0.950	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08	3.01	2.97	2.93
0.990		11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.20	5.03	4.95	4.86
0.999		25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77	11.54	11.19	10.84	10.48	10.11	9.73	9.53	9.33

(segue)

Tavola C.4 Quantili della distribuzione F di Fischer. (*continua*)

		r_1																
p	r_2	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120	∞
0.950	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86	2.79	2.75	2.71
0.990		10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.65	4.48	4.40	4.31
0.999		22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89	9.57	9.24	8.90	8.55	8.19	8.00	7.81
0.950	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.70	2.62	2.58	2.54
0.990		10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.25	4.08	4.00	3.91
0.999		21.04	14.90	12.55	11.28	10.48	9.93	9.52	9.20	8.96	8.75	8.45	8.13	7.80	7.47	7.12	6.94	6.76
0.950	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.57	2.49	2.45	2.40
0.990		9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	3.94	3.78	3.69	3.60
0.999		19.69	13.81	11.56	10.35	9.58	9.05	8.65	8.35	8.12	7.92	7.63	7.32	7.01	6.68	6.35	6.18	6.00
0.950	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.47	2.38	2.34	2.30
0.990		9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.70	3.54	3.45	3.36
0.999		18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	7.29	7.00	6.71	6.40	6.09	5.76	5.59	5.42
0.950	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.38	2.30	2.25	2.21
0.990		9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.51	3.34	3.25	3.17
0.999		17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98	6.80	6.52	6.23	5.93	5.63	5.30	5.14	4.97
0.950	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.31	2.22	2.18	2.13
0.990		8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.35	3.18	3.09	3.00
0.999		17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58	6.40	6.13	5.85	5.56	5.25	4.94	4.77	4.60
0.950	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.25	2.16	2.11	2.07
0.990		8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.21	3.05	2.96	2.87
0.999		16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08	5.81	5.54	5.25	4.95	4.64	4.48	4.31
0.950	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.19	2.11	2.06	2.01
0.990		8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.10	2.93	2.84	2.75
0.999		16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.20	5.98	5.81	5.55	5.27	4.99	4.70	4.39	4.23	4.06

(*segue*)

Tavola C.4 Quantili della distribuzione F di Fischer. (*continua*)

		r_1																
p	r_2	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120	∞
0.950	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.15	2.06	2.01	1.96
0.990		8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.00	2.83	2.75	2.65
0.999		15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75	5.58	5.32	5.05	4.78	4.48	4.18	4.02	3.85
0.950	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.11	2.02	1.97	1.92
0.990		8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	2.92	2.75	2.66	2.57
0.999		15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56	5.39	5.13	4.87	4.59	4.30	4.00	3.84	3.67
0.950	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.07	1.98	1.93	1.88
0.990		8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.84	2.67	2.58	2.49
0.999		15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39	5.22	4.97	4.70	4.43	4.14	3.84	3.68	3.51
0.950	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.04	1.95	1.90	1.84
0.990		8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.78	2.61	2.52	2.42
0.999		14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08	4.82	4.56	4.29	4.00	3.70	3.54	3.38
0.950	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.01	1.92	1.87	1.81
0.990		8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.72	2.55	2.46	2.36
0.999		14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11	4.95	4.70	4.44	4.17	3.88	3.58	3.42	3.26
0.950	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	1.98	1.89	1.84	1.78
0.990		7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.67	2.50	2.40	2.31
0.999		14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99	4.83	4.58	4.33	4.06	3.78	3.48	3.32	3.15
0.050	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	1.96	1.86	1.81	1.76
0.010		7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.62	2.45	2.35	2.26
0.001		14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89	4.73	4.48	4.23	3.96	3.68	3.38	3.22	3.05
0.950	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.94	1.84	1.79	1.73
0.990		7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.58	2.40	2.31	2.21
0.999		14.03	9.34	7.55	6.59	5.98	5.55	5.24	4.99	4.80	4.64	4.39	4.14	3.87	3.59	3.29	3.14	2.97

(*segue*)

Tavola C.4 Quantili della distribuzione F di Fischer. *(continua)*

		r_1																
p	r_2	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120	∞
0.950	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.92	1.82	1.77	1.71
0.990		7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.54	2.36	2.27	2.17
0.999		13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71	4.56	4.31	4.06	3.79	3.52	3.22	3.06	2.89
0.950	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.90	1.80	1.75	1.69
0.990		7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.50	2.33	2.23	2.13
0.999		13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64	4.48	4.24	3.99	3.72	3.44	3.15	2.99	2.82
0.950	27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.88	1.79	1.73	1.67
0.990		7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.47	2.29	2.20	2.10
0.999		13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57	4.41	4.17	3.92	3.66	3.38	3.08	2.92	2.75
0.050	28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.87	1.77	1.71	1.65
0.990		7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.44	2.26	2.17	2.06
0.999		13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50	4.35	4.11	3.86	3.60	3.32	3.02	2.86	2.69
0.950	29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.85	1.75	1.70	1.64
0.990		7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.41	2.23	2.14	2.03
0.999		13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45	4.29	4.05	3.80	3.54	3.27	2.97	2.81	2.64
0.050	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.84	1.74	1.68	1.62
0.990		7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.39	2.21	2.11	2.01
0.999		13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24	4.00	3.75	3.49	3.22	2.92	2.76	2.59
0.950	35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.04	1.96	1.88	1.79	1.68	1.62	1.56
0.990		7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88	2.74	2.60	2.44	2.28	2.10	2.00	1.89
0.999		12.90	8.47	6.79	5.88	5.30	4.89	4.59	4.36	4.18	4.03	3.79	3.55	3.29	3.02	2.72	2.56	2.38
0.950	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.74	1.64	1.58	1.51
0.990		7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.20	2.02	1.92	1.80
0.999		12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02	3.87	3.64	3.40	3.15	2.87	2.57	2.41	2.23

(segue)

Tavola C.4 Quantili della distribuzione F di Fischer. *(continua)*

		r_1																
p	r_2	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120	∞
0.950	50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.87	1.78	1.69	1.58	1.51	1.44
0.990		7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.56	2.42	2.27	2.10	1.91	1.80	1.68
0.999		12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.82	3.67	3.44	3.20	2.95	2.68	2.38	2.21	2.03
0.950	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.65	1.53	1.47	1.39
0.990		7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.03	1.84	1.73	1.60
0.999		11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69	3.54	3.32	3.08	2.83	2.55	2.25	2.08	1.89
0.950	70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.89	1.81	1.72	1.62	1.50	1.44	1.35
0.990		7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59	2.45	2.31	2.15	1.98	1.78	1.67	1.54
0.999		11.80	7.64	6.06	5.20	4.66	4.28	3.99	3.77	3.60	3.45	3.23	2.99	2.74	2.47	2.16	1.99	1.79
0.950	80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.88	1.79	1.70	1.60	1.48	1.41	1.32
0.990		6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.42	2.27	2.12	1.94	1.75	1.63	1.49
0.999		11.67	7.54	5.97	5.12	4.58	4.20	3.92	3.70	3.53	3.39	3.16	2.93	2.68	2.41	2.10	1.92	1.72
0.950	90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.86	1.78	1.69	1.59	1.46	1.39	1.30
0.990		6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52	2.39	2.24	2.09	1.92	1.72	1.60	1.46
0.999		11.57	7.47	5.91	5.06	4.53	4.15	3.87	3.65	3.48	3.34	3.11	2.88	2.63	2.36	2.05	1.87	1.66
0.950	100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.85	1.77	1.68	1.57	1.45	1.38	1.28
0.990		6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.37	2.22	2.07	1.89	1.69	1.57	1.43
0.999		11.50	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44	3.30	3.07	2.84	2.59	2.32	2.01	1.83	1.62
0.950	∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.46	1.32	1.22	1.00
0.990		6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.70	1.47	1.32	1.00
0.999		10.83	6.91	5.42	4.62	4.10	3.74	3.47	3.27	3.10	2.96	2.74	2.51	2.27	1.99	1.66	1.45	1.00

Bibliografia

Abadir K., Heijmans R., Magnus J. R. (2018), *Statistics*, Cambridge University Press, Cambridge.

Anscombe F. J. (1973), Graphs in Statistical Analysis, *The American Statistician*, Vol. 27, pp. 17-21.

Bernstein S., Bernstein R. (1999), *Elements of Statistics*, Vol. 1, McGraw-Hill, New York.

Campbell M. J., Machin D. (1999), *Medical Statistics – A Commonsense Approach*, Wiley, New York.

Casella G., Berger R. L. (2002), *Statistical Inference*, Duxbury Press, Pacific Grove, California.

Conti P. L., Marella D. (2012), *Campionamento da popolazioni finite. Il disegno campionario*, Springer-Verlag Italia, Milano.

Fracassi R. (1961), *Dal censimento dell'Unità ai censimenti del centenario. Un secolo di vita della statistica italiana 1861-1961*, Istat, Roma.

Freedman D., Pisani R., Purves R. (1998), *Statistica*, McGraw-Hill, Milano.

Gillies D. (2000), *Philosophical Theories of Probability*, Routledge, London.

Hand D. J., Daly F., Lunn A. D., McConway K. J., Ostrowski E. (1994), *Handbook of Small Data Sets*, Chapman & Hall, London.

Hogg R. V., McKean J. W., Craig A. T. (2013), *Introduction to Mathematical Statistics*, Pearson, New York.

Kendall M. G., Stuart A. (1973), *The Advanced Theory of Statistics*, Vol. 2, Griffin, London.

Kutner M. H., Nachtsheim C. J., Neter J., Li W. (2005), *Applied Linear Statistical Models*, McGraw-Hill, London.

Levine D., Krehbiel T. C., Berenson M. L. (2012), *Basic Business Statistics: A First Course*, Prentice Hall, New Jersey.

Lizana A. R. (2013), *Gauss. Una rivoluzione nella teoria dei numeri*, RBA Italia.

Lusinchi D. (2022), Kiær and the rebirth of the representative method: A case-study in controversy management at the International Statistical Institute (1895-1903), *Journal of the History of the Behavioral Sciences*, 58, pp. 163-182.

Montgomery D. C. (2005), *Design and Analysis of Experiments*, Wiley, New York.

Neyman J. (1934), On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection, *Journal of the Royal Statistical Society*, Vol. 97, pp. 557-625.

Neyman J., Pearson E. S. (1933), On the problem of the most efficient tests of statistical hypotheses, *Philosophical Transactions of the Royal Society*, A, Vol. 231, pp. 289-337.

OECD (2008), *Handbook on Constructing Composite Indicators: Methodology and User Guide*, OECD publishing.

Pace L., Salvan A. (1996), *Introduzione alla statistica – Statistica descrittiva*, Cedam, Padova.

Patriarca S. (2011), La disciplina della democrazia: sapere statistico e Risorgimento, in *Atti del workshop: Il percorso storico della statistica nell'Italia unita*, Istat, Roma.

Pearson E. S. (1968), Some early correspondence between W. S. Gosset, R. A. Fisher and Karl Pearson, with notes and comments, *Biometrika*, 55, pp. 445-447.

Plackett R. L. (1972), The discovery of the method of least squares, *Biometrika*, Vol. 59, pp. 239-251.

Rao C. R. (1973), *Linear Statistical Inference and its Applications*, Wiley, New York.

Rohatgi V. K., Saleh A. K. Md. Ehsanes (2015), *An Introduction to Probability and Statistics*, Wiley, New York.

Särndal C. E., Swensson B., Wretman J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Spanos A. (1999), *Probability Theory and Statistical Inference*, Cambridge University Press, Cambridge.

Stigler S. M. (1977), An attack on Gauss published by Legendre in 1920, *Historia Mathematica*, Vol. 4, pp. 31-35.

Stigler S. M. (1981), Gauss and the invention of least squares, *The Annals of Statistics*, Vol. 9, pp. 465-474.

Stigler S. M. (1986), *The History of Statistics*, Harvard University Press, Cambridge, Massachusetts.

Indice analitico

A

albero degli eventi, 246
allocazione
  del campione negli strati, 525
  di Neyman, 525
  proporzionale, 525
  uniforme, 525
ammontare totale
  della variabile, 141
  di classe, 145
ampiezza
  dell'intervallo di confidenza, 406
  di classe, 51
analisi statistica
  bivariata, 42
  univariata, 42
asimmetria
  negativa, 159
  positiva, 159
assiomi della probabilità, 252
assunzioni di base dei minimi quadrati, 495

B

base di campionamento, 16, 514

C

campionamento
  a due stadi, 516
  a grappoli, 516
  casuale semplice, 515
  sistematico, 519
  stratificato, 516
campione
  a due stadi, 17
  a grappoli, 17
  a valanga, 18
  casuale, 359
    osservabile, 360
    semplice, 17
  per quote, 18
  per testimoni privilegiati, 19
  probabilistico, 515, 531
  sistematico, 17
  stratificato, 17
campo di variazione, 136
cartogramma, 82
centile, 110, 286
chiusura delle classi, 48
cinque valori di sintesi, 171
classe
  mediana, 113
  modale, 116
  reale, 48
codevianza, 217
coefficiente
  di correlazione lineare, 235, 334
  di fiducia, 405
  di regressione, 218
  di variazione, 139
combinazioni lineari di n variabili casuali, 337
concentrazione, indici di, 140
contingenze, 201
correlazione
  negativa, 233
  positiva, 233
correzione
  per continuità, 351
  per popolazione finita, 520
costanti caratteristiche, 169
covarianza, 217, 333
curtosi, 163
curva
  di concentrazione, 142
  di Gauss, 164
  di Lorenz, 142
  di massima concentrazione, 143

D

dati
  anomali, 58, 116
  mancanti, 57
decile, 110
densità di frequenza, 51
devianza, 130
  residua, 206, 220, 499
  spiegata, 206, 220
  totale, 206
deviazione standard, 129, 280
differenza
  interquartile, 137
  percentuale, 26
  relativa, 26
  semplice media, 134
dilemma di Monty Hall, 267
dimensioni, 187
dipendenza
  analisi della, 195
  in media, 205
  perfetta, 201
distorsione, 386
distribuzione(i)
  bimodali, 116
  binomiale, 299
  campionaria, 362
  chi-quadrato, 314
  condizionate, 197
  di frequenza, 44
    asimmetriche, 159
    attesa, 481
    cumulata, 45
    in classi, 49
    percentuale, 45
    relativa, 45
    relativa condizionate, 197
  di Poisson, 301
  di probabilità, 275
    condizionate, 329
    congiunta, 327, 334
    del campione casuale, 360
    della popolazione, 359
    marginali, 328
  doppie di frequenza, 53
  esponenziale, 305
  F, 471
  marginali, 197
  normale, 306
    doppia, 336
    standard, 309
  plurimodali, 116
  T di Student, 370
  T non centrale, 445
  uniforme, 52
    continua, 304
    discreta, 296
disuguaglianza
  di Chebyshev, 173, 287
  di Markov, 288
dominio di studio, 523

E

equidistribuzione, 140
errore(i)
  campionario, 59, 531
  di approssimazione, 216
  di I tipo, 423
  di II tipo, 423
  di misura, 58
  di predizione, 506
  di stima, 384
  di predizione, 216, 222
  quadratico medio, 387
  non campionari, 59
esperimento, 15
  casuale, 246
eterogeneità, 146
etichetta, 514
Eurostat, 9
evento(i)
  certo, 252
  complementare, 249
  composto, 248
  disgiunti, 249
  elementare, 246
  impossibile, 249
  incompatibili, 249
  indipendenti, 259
  intersezione, 249
  semplice, 248
  unione, 249

F

fallacia dell'accusatore, 267
famiglia di distribuzioni
    di probabilità, 295
fattori, 21
  sub-sperimentali, 22
fonti
  primarie, 15
  secondarie, 15
formula
  delle probabilità totali, 262
  di Bayes, 263
  ideale di Fisher, 187
  operativa della deviazione standard, 129
frequenze, 44
  congiunte, 195
  cumulate, 45
    percentuali, 46
  percentuali, 44
  relative, 44
    cumulate, 46
funzione(i)
  d'insieme, 251
  di densità, 281
    congiunta, 334
    marginali, 335
    triangolare, 283
    normale, 164
  di frequenza, 68
  di probabilità, 276
    condizionate, 330
    congiunta, 327
    marginali, 328
  di potenza, 443
  di ripartizione, 69, 276
  di verosimiglianza, 395
  interpolatrice, 215

G

gradi di libertà, 314, 370
grafico
  ad aste, 68
  a bolle, 79
  a linee, 79
  a nastri, 69
    multiplo, 77
    sovrapposti, 77
  a punti, 67
  a scatola, 171
  a torta, 76
  a violino, 174
  di dispersione, 78
gruppo
  di controllo, 22
  sperimentale, 22

I

indagini
  campionarie, 16
  censuarie, 16
  statistiche, 15
indicatori
  compositi, 187
  elementari, 187
  tematici, 187
indice(i)
  assoluti di variabilità, 138
  dei prezzi di Paasche, 186
  di concentrazione, 140
  di correlazione di Spearman, 239
  di determinazione, 220
  di dipendenza degli anziani, 27
  di disuguaglianza, 140
  di eterogeneità, 147
  di Laspeyres dei prezzi, 185
  di vecchiaia, 27
  percentuali di variabilità, 138
inferenza statistica, 25, 373
  Bayesiana, 453
informazioni ausiliarie, 517
interpretazione
  classica della probabilità, 253
  frequentista della probabilità, 253
  soggettivista della probabilità, 253
intervallo
  di confidenza, 405
  di credibilità, 454
  di tolleranza, 15, 48
  predittivo, 506
intervista
  diretta, 20
  postale, 20
  telefonica, 20
  via Internet, 20
ipernormalità, 165
iponormalità, 165
ipotesi
  alternativa, 422
  nulla, 422, 451
  statistica, 422
    composta, 422
    semplice, 422
Istat, 8
istogramma di frequenza, 72

L

legge
  debole dei grandi numeri, 347
  degli eventi rari, 301
  dei grandi numeri, 346
    di Bernoulli, 348
  empirica del caso, 253
  forte dei grandi numeri, 347
linearità della media, 94, 279
lista, 16, 514
livello di significatività, 423
  osservato, 428

M

mancate risposte
  parziali, 57
  totali, 57
mappe tematiche, 82
margine di errore, 397
massima concentrazione, 140
matrice dei dati, 24, 56
media(e), 278
  analitiche, 93
  aritmetica, 93
    ponderata, 105
  armonica, 96
  campionaria, 362
  condizionate, 205
  di posizione, 93
  di potenza, 99
  geometrica, 97
  quadratica, 99
  robuste, 116
  secondo Chisini, 119
  troncata, 116
    con sostituzioni, 116
mediana, 107
metodo
  dei minimi quadrati, 217
  della massima verosimiglianza, 395
misure
  di probabilità, 252
  di tendenza centrale, 93
moda, 115
modalità, 12
modello
  di regressione, 215
    lineare semplice, 493
momento
  centrato, 287
  non centrato, 287

N

numeri indici
  a base fissa, 178
  a base mobile, 179
  percentuali, 178
  sintetici, 183

O

omogeneità, 146
omoschedasticità, 495

P

paniere, 184
paradosso
  di San Pietroburgo, 290
  di Simpson, 120, 208
parametro, 295, 359
  di non centralità, 445
partizione dello spazio campionario, 251
passo di campionamento, 17
percentile, 110, 286
perdita di informazioni, 49
piano
  dell'indagine, 19
  di campionamento, 19, 515
polarità positiva, 11, 189
popolazione, 359
  Bernoulliana, 360
  esponenziale, 360
  finita, 16, 513
  generatrice, 360
  infinita, 359
  multinomiale, 479
    doppia, 484
  normale, 360
potenza del test, 424
principio di trasferibilità, 142
probabilità
  a posteriori, 263
  a priori, 263
  condizionata, 257
  di inclusione, 518
  uguali, 519
problema
  di Behrens-Fisher, 466
  diretto, 3, 373
  inverso, 3, 373
proporzioni, 44
  campionarie, 364
proprietà
  asintotiche, 392
  associativa della media aritmetica, 119
  di assenza di memoria, 318
protocollo sperimentale, 23
prova, 246
  Bernoulliana, 297
punti osservati, 213

Q

quantili, 110, 285
  estremi, 396
quantità pivot, 406
quartili, 109
questionario, 19

R

rapporti
  di coesistenza, 27
  di composizione, 27
  di densità, 28
rapporto
  di concentrazione di Gini, 141
  di correlazione, 206
  di mascolinità, 27
regola
  del prodotto, 258
  dell'evento complementare, 252
  della somma, 252
  empirica, 307
regressione lineare semplice, 215, 493
relazione statistica, 213
residui, 216, 496
retta
  di regressione, 217
  interpolatrice, 216
    dei minimi quadrati, 217
risposta, 213

S

scala Likert, 19
scomposizione della devianza, 205, 220
scostamento
  quadratico medio, 129
  semplice medio, 128
segmento di equidistribuzione, 143
serie
  di osservazioni, 42
  doppie di osservazioni, 42
  multiple di osservazioni, 42
  semplici di osservazioni, 42
  storiche, 55
  territoriali, 56
  triple di osservazioni, 42
simmetria, 157
spazio
  campionario, 246
    continuo, 248
    discreto, 248
  dei campioni, 515
statistica
  campionaria, 16, 362
  descrittiva, 25
  test, 422
  test chi-quadrato, 481
stima(e), 383
  dei minimi quadrati, 496
  dei parametri, 25, 373
  di massima verosimiglianza, 395
stimatore(i), 383
  asintoticamente non distorto, 394
  consistente, 393
    in media quadratica, 393
    in probabilità, 393
  dei minimi quadrati, 498
  di Horvitz-Thompson, 518
  distorti, 386
  intervallare, 405
  non distorto, 385
    con varianza uniformemente minima, 392
  ottimo, 394
  per quoziente, 530
  per regressione, 530
  più efficiente, 390
strati, 17, 516, 523
studi
  osservazionali, 15
  sul campo, 15

T

tabella
  di analisi della varianza, 503
  di contingenza, 54
  di frequenza, 44
  di indipendenza, 199
  di spoglio, 43
tasso, 28
  di natalità, 28
  di sondaggio, 520
    di primo stadio, 528
    di secondo stadio, 528
tavola della normale standard, 311
teorema del limite centrale, 349
termine di errore, 493
test
  della bontà dell'adattamento, 482
  di ipotesi, 422
  non distorto, 448
  statistico, 422
  uniformemente più potente, 447
trattamento, 21

U

unità statistica, 12
  di primo stadio, 527
  elementare, 527
uomo medio, 4

V

valore atteso, 278
valore(i)
  centrali di classe, 51
  critico(i), 397, 423
  teorici, 216, 496
variabile(i), 12
  aleatoria, 273
  casuale(i), 273
    continua, 281
    discreta, 274
    doppia, 326
    incorrelate, 334
    indipendenti, 331
    standardizzata, 286
  di flusso, 81
  di stato, 81

dipendente, 213
esplicative, 21
indipendenti, 198, 213
qualitative
nominali, 12
ordinali, 12
quantitative
continue, 13
discrete, 13
non trasferibili, 13
trasferibili, 13, 139
stocastica, 273
variabilità, 127
di campionamento, 518
varianza, 130, 279
campionaria, 362
combinata, 463
variazione
percentuale media, 182
relativa, 182
media, 182
verifica delle ipotesi, 25, 373

Z

zona
di non rifiuto, 423
di rifiuto, 423

Soluzione degli esercizi

Le soluzioni degli esercizi mostrate nelle pagine che seguono sono state ottenute con l'impiego del foglio di lavoro Excel e sono in gran parte illustrate anche nel *file* Excel dal titolo "Soluzioni degli esercizi con Excel". Questa sorta di duplicazione è giustificata dal fatto che il foglio elettronico rende trasparente il procedimento di calcolo che conduce ai risultati.

Come è noto, nei calcoli di Excel vengono usate molte cifre decimali. Ne segue che lo svolgimento degli esercizi con una calcolatrice manuale, con annotazione su un foglio dei risultati parziali, richiede che si utilizzi un congruo numero di cifre decimali (almeno quattro), affinché i risultati siano comparabili con quelli qui esposti.

Salvo avviso in senso contrario, le soluzioni che implicano l'uso delle distribuzioni normale e chi-quadrato sono ottenute adoperando le Tavole C.1 e C.2.

Capitolo 1

1.1 a) La popolazione è composta da tutte le imprese italiane. L'unità statistica è la singola impresa.

b) La popolazione è costituita dagli individui dei paesi dell'UE appartenenti alla forza lavoro (individui di età pari o superiore a 15 anni che sono occupati, disoccupati o in cerca di occupazione). L'unità statistica è il singolo individuo.

c) La popolazione è composta da tutti i cittadini italiani maggiorenni. L'unità statistica è il singolo cittadino.

d) La popolazione è composta dalle famiglie italiane; l'unità statistica è il singolo nucleo familiare.

1.2 La popolazione è composta da tutti gli studenti universitari che frequentano il corso presso l'università considerata. L'unità statistica è il singolo studente. Nella tabella seguente, sono elencate alcune possibili variabili di interesse e le loro modalità.

Variabile	Modalità
Mezzo di trasporto	Auto, scooter treno, autobus, bicicletta, nessuno.
Tempo di percorrenza	Numeri reali
Costo	Numeri reali
Sesso	Femmina, maschio.
Residenza	Comune sede dell'università, altro comune della provincia, altro comune della regione, altra regione.
Scuola frequentata	Liceo, Istituo tecnico comm. Istituto tecnico ind., altra scuola.

1.3 La popolazione è composta da tutti i comuni della provincia considerata. L'unità statistica è il singolo comune. La variabile è quantitativa discreta.

1.4 I 31 alberi di cedro nero sono un campione di osservazioni di una popolazione non specificata. L'unità statistica è il singolo albero. Le variabili sono tutte quantitative continue.

1.5

Variabile	Tipo
Sesso	Qualitativa nominale
Esami	Quantitativa discreta
Altezza	Quantitativa continua
Peso	Quantitativa continua
Colore dei capelli	Qualitativa nominale
Voto	Qualitativa ordinale

1.6

Variabile	Ampiezza intervallo di tolleranza
Età	1 anno
Peso	1 kg
Altezza	1 cm
Circ. collo	1 mm
Circ. torace	1 cm
Circ. addome	1 cm
Circ. coscia	1 cm
Lunghezza avamb.	1 mm
Circ. polso	1 mm

1.7

Variabile	Tipo
Dimensione	Quantitativa continua
Colore	Qualitativa nominale
Lucentezza	Qualitativa ordinale
Tipo di certificazione	Qualitativa nominale
Prezzo	Quantitativa discreta

1.8 a) b)

Nazione	Densità popolazione (ab. per km^2)	Tasso di natalità (per 1000 ab.)
Austria	105.6	9.66
Danimarca	135.3	10.59
Francia	102.7	11.04
Germania	232.4	11.50
Ungheria	105.1	9.18
Italia	199.3	7.29
Polonia	122.8	10.11
Portogallo	111.5	8.47
Spagna	92.2	7.92
Regno Unito	233.5	11.01

c) 19 661 979.

Capitolo 2

2.1 a)

Voto	A	B	C	D	F
Frequenza	17	26	13	5	3

b)

Voto	A	B	C	D	F
Frequenza perc.	26.6	40.6	20.3	7.8	4.7

2.2 a)

N. di tabelle	0	1	2	3	4	5
Frequenza	6	6	3	1	3	1

b)

N. di tabelle	0	1	2	3	4	5
Frequenza cumulata	6	12	15	16	19	20
Frequenza relativa cumulata	0.30	0.60	0.75	0.80	0.95	1.00

2.3

a)

Lung. sepalo	Frequenza
4.25-4.45	4
4.45-4.65	5
4.65-4.85	7
4.85-5.05	12
5.05-5.25	11
5.25-5.45	6
5.45-5.65	2
5.65-5.85	3
Totale	50

b)

Lung. sepalo	Frequenza cumulata	Freq. relativa cumulata
4.25-4.45	4	0.08
4.45-4.65	9	0.18
4.65-4.85	16	0.32
4.85-5.05	28	0.56
5.05-5.25	39	0.78
5.25-5.45	45	0.90
5.45-5.65	47	0.94
5.65-5.85	50	1.00

2.4 a)

N. di membri	1	2	3	4	5	≥ 6
Freq. relativa cumulata	0.319	0.594	0.790	0.947	0.988	1.000

b)

N. di membri	1	2	3	4	5	≥ 6
Frequenza	8251	7113	5070	4061	1060	310

2.5 a)

Peso	fino a 70	71-80	81-90	91 e oltre
Frequenza	8	10	10	10

b)

Peso	fino a 70	71-80	81-90	91 e oltre
Freq. relativa cumulata	0.211	0.474	0.737	1.000

2.6 a) 0-3; 3-6; 6-11; 11-16; 16-21; 21 e oltre.

b) Numero di coppie, con una durata del matrimonio nell'intervallo $[3, 6)$, che hanno chiesto l'adozione.

c)

Durata del matrimonio	0-3	3-6	6-11	11-16	16-21	21 ed oltre
Frequenza perc.	5.2	24.0	38.0	21.0	7.6	4.2

d)

Durata del matrimonio	0-3	3-6	6-11	11-16	16-21	21-27
Densità di frequenza	128.7	600.0	568.8	314.6	113.8	52.5

e) 4075.8.

2.7 a) 75.8%. b) 40.5%. c) 67.1%.

2.8

Rendimento	(–3, –2]	(–2, –1]	(–1, 0]	(0, 1]	(1, 2]	(2, 3]
Frequenza	5	10	29	31	12	0

2.9

Esame	Voto					Totale
	A	B	C	D	F	
45-49	0	0	0	0	1	1
50-54	0	0	0	0	0	0
55-59	0	0	0	2	2	4
60-64	0	0	1	2	0	3
65-69	0	0	4	1	0	5
70-74	0	0	3	0	0	3
75-79	0	2	4	0	0	6
80-84	0	10	1	0	0	11
85-89	1	8	0	0	0	9
90-94	11	6	0	0	0	17
95-99	5	0	0	0	0	5
Totale	17	26	13	5	3	64

2.10 38.5.

2.11

Circonferenza torace	Circonferenza addome				Totale
	fino a 80	81-90	91-100	oltre 100	
fino a 90	7	1	0	0	8
91-100	6	15	5	0	26
101-110	0	6	14	4	24
oltre 110	0	0	1	4	5
Totale	13	22	20	8	63

2.12 a)

Colore degli occhi	Colore dei capelli				
	Chiaro	Rosso	Bruno	Scuro	Nero
Azzurro	45.4	5.3	33.6	15.3	0.4
Scuro	7.5	3.7	30.6	51.8	6.5

b) 7.34%. c) 46.0%. d) 13.3%.

2.13 a)

Area geografica	Ampiezza demografica ≤ 2 000	2 001-5 000	5 001-20 000	20 001-50 000	50 001-100 000	≥ 100 001	Totale
Nord-Ovest	54.12	23.92	18.03	3.09	0.61	0.24	100
Nord-Est	29.25	29.91	34.76	4.34	0.58	1.16	100
Centro	35.78	25.54	27.61	7.96	2.28	0.83	100
Sud	43.52	24.90	23.44	5.55	2.13	0.45	100
Isole	39.90	28.03	22.95	6.65	1.69	0.78	100

b)

Area geografica	Ampiezza demografica ≤ 2 000	2,001-5 000	5,001-20 000	20,001-50 000	50,001-100 000	≥ 100 001	Totale
Totale	43.72	25.79	23.85	4.82	1.26	0.57	100

c) Nord-Ovest.

2.14 a) Sono le frequenze della distribuzione marginale della lunghezza del dito medio. Sono le frequenze della distribuzione marginale della lunghezza del piede.

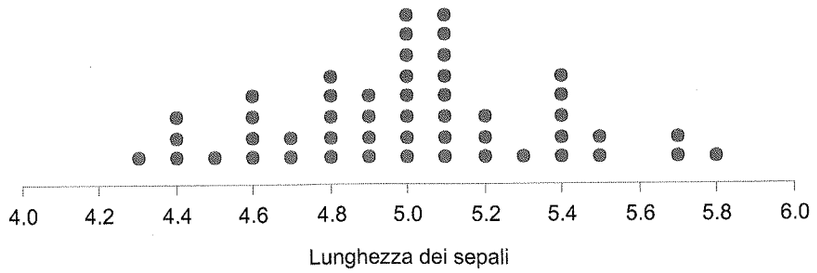
b) La posizione delle frequenze all'interno della tavola suggerisce che la lunghezza del piede tende ad aumentare all'aumentare della lunghezza del dito medio.

c) 10.8%.

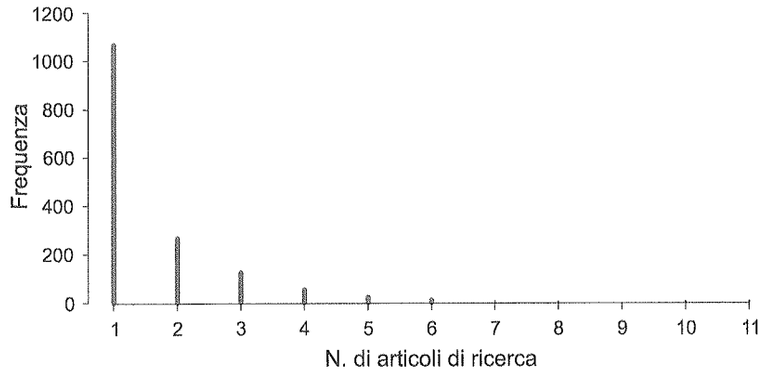
d) È possibile determinare approssimativamente questo numero assumendo l'uniforme distribuzione delle osservazioni nella classe 105-109.

Capitolo 3

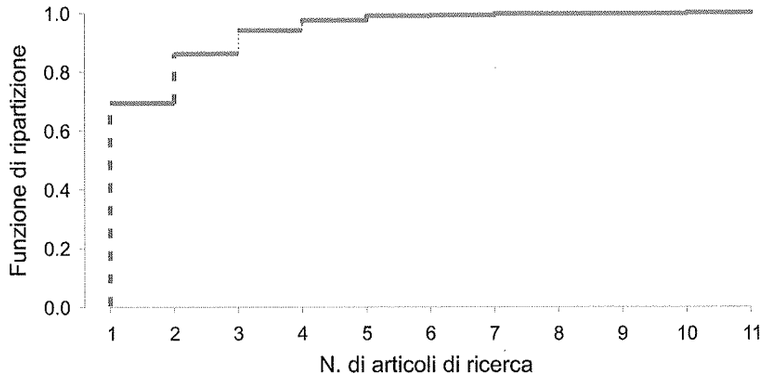
3.1



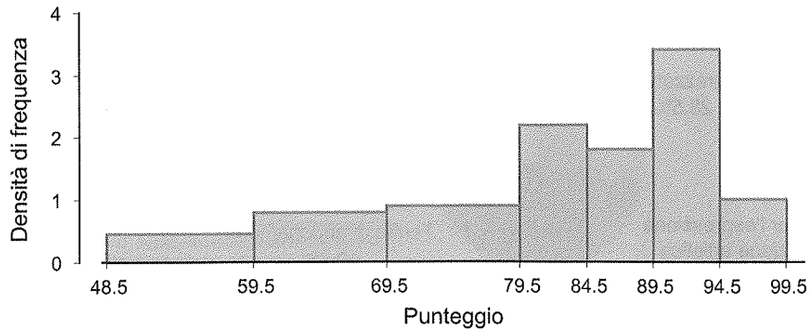
3.2 a)



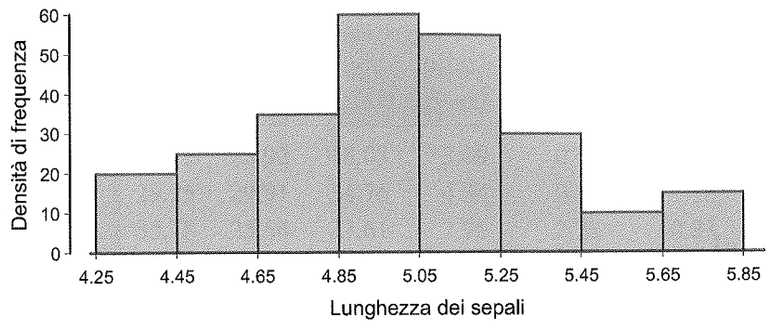
b)



3.3

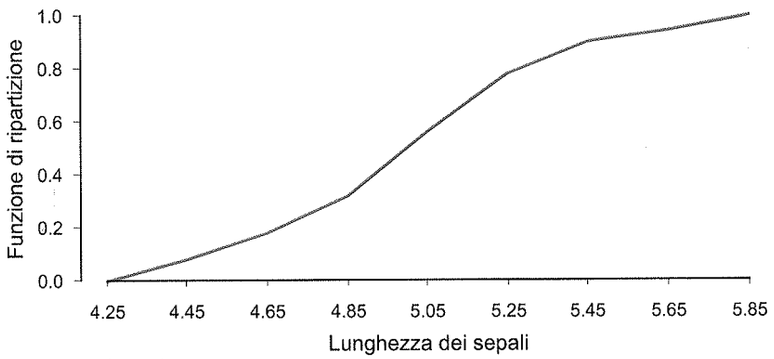


3.4 a)



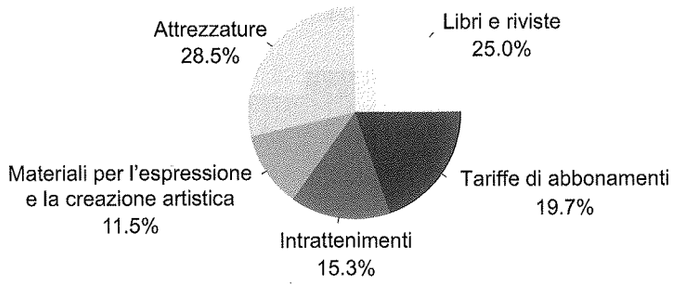
b) Per costruire l'istogramma possono essere usate validamente le frequenze, poiché le classi hanno la stessa larghezza.

c)

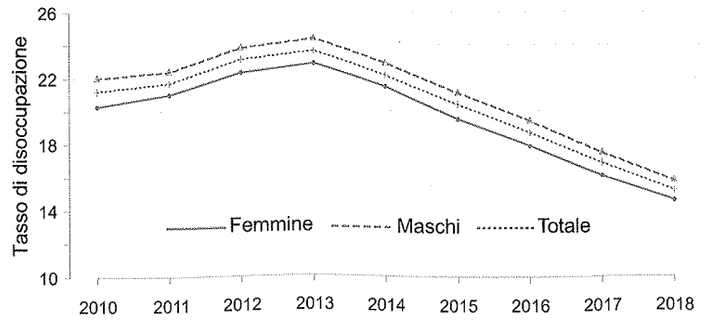


$F(5.2) = 0.725$

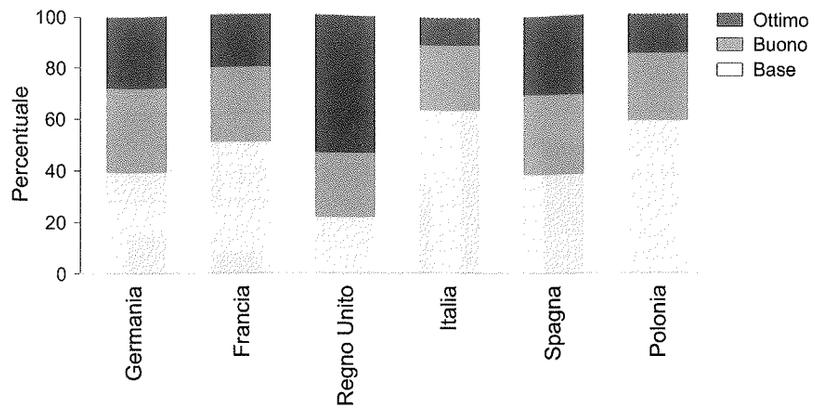
3.5



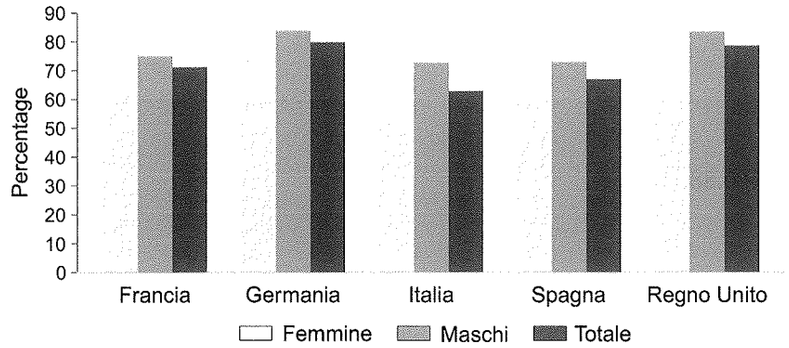
3.6



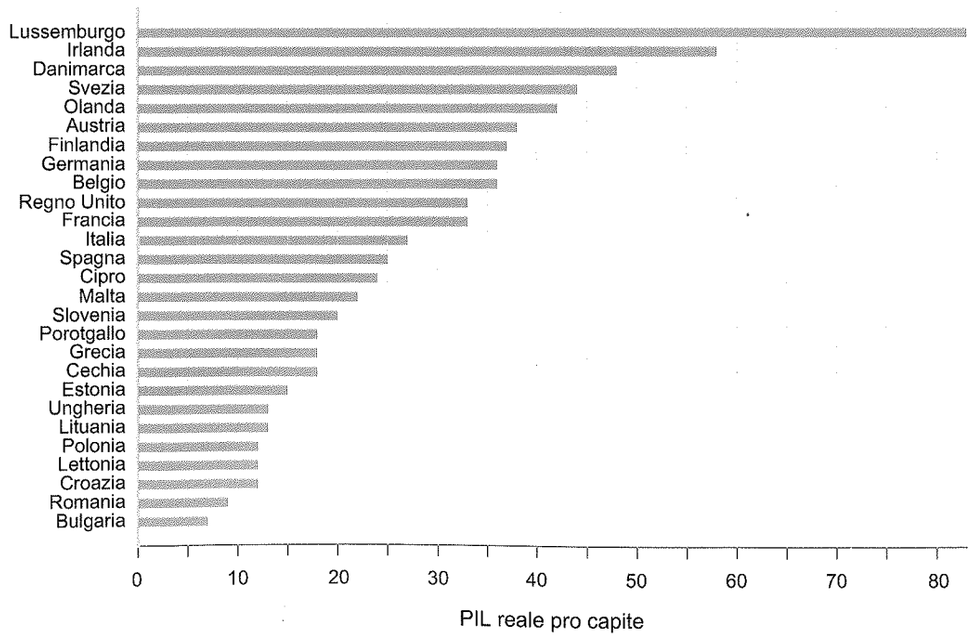
3.7



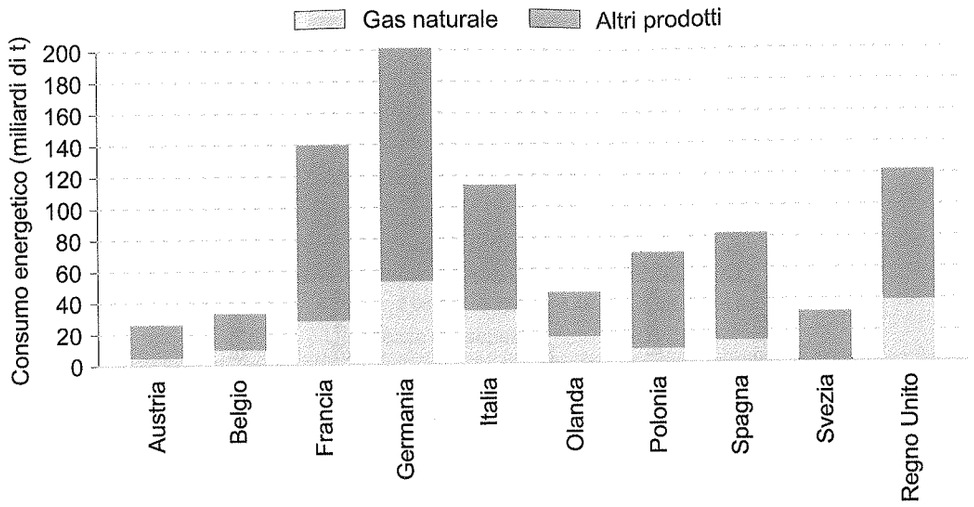
3.8



3.9



3.10



Capitolo 4

4.1 a) 487.13. c) 6640.88.

4.3 98.24.

4.4 6.088.

4.5 a) $\mu = 2.36$. c) $\mu_a = 1.75$; $\mu_g = 2.04$.

4.6 a) $\mu = 4.998$. b) $\mu_a = 4.972$; $\mu_g = 4.985$.

4.7 a) 9.12. b) 11.36.

4.8 a) $\mu = 11\,386.8$. b) $\mu_q = 22\,744.8$.

4.9 a) 2.32. b) 2.38.

4.10 22.77.

4.12 $m = 82$; $q_1 = 71$; $q_3 = 90$.

4.13 $q_1 = 1$; $m = 2$; $q_3 = 3$.

4.14 $q_1 = 4.75$; $m = 5.00$; $q_3 = 5.22$.

4.15 $q_1 = 2.23$; $m = 4.47$; $q_3 = 10.63$.

4.16 $q_1 = 1403.8$; $m = 1618.6$; $q_3 = 1871.8$.

4.18 1201-1600.

Capitolo 5

5.1 a) 81.10. b) 6 577.25; 39 463.50. c) 32.44.

5.2 276.15.

5.3 68.17.

5.4 1.261.

5.8 0.612.

5.9 a) 10.24; 12.42.

5.10

Distribuzione	a	b	c
Deviazione standard	1.7	1.2	1.8

5.11 0.286; 0.359.

5.12 a) 7.11. b) 6.60.

5.13 10.55; 37.94.

5.14 a) 19 689.25. b) 19 607.67.

5.15 0.122.

5.16 108.2.

5.17 13.823.

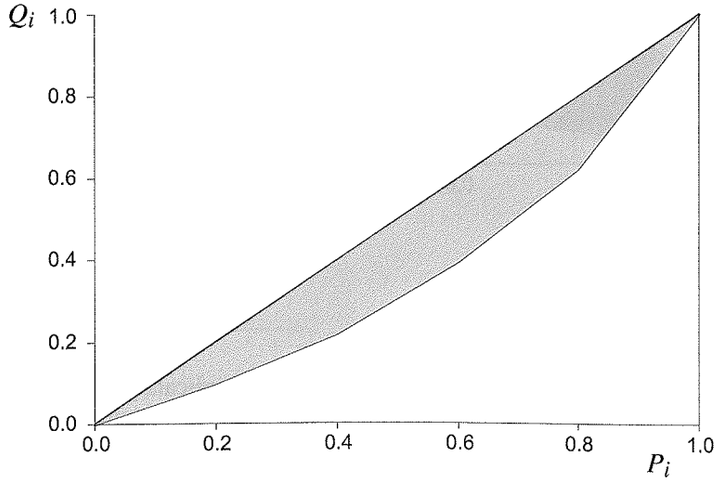
5.18 0.409.

5.19 18.43.

5.20 7.19%.

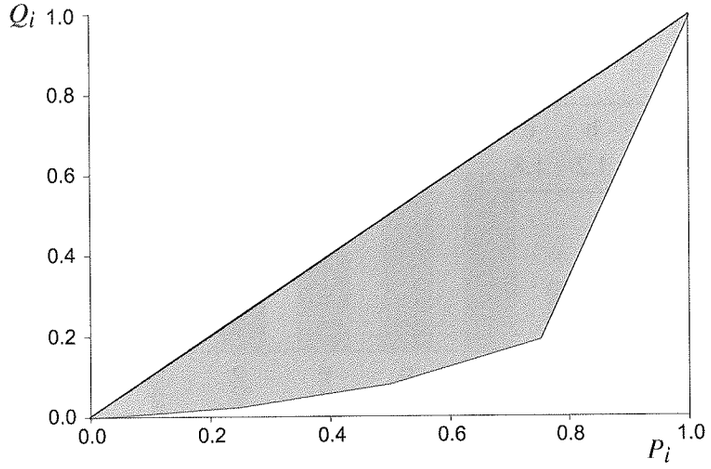
5.21 Altezza: $CV = 3.59\%$; Peso: $CV = 16.11\%$.

5.22 a)



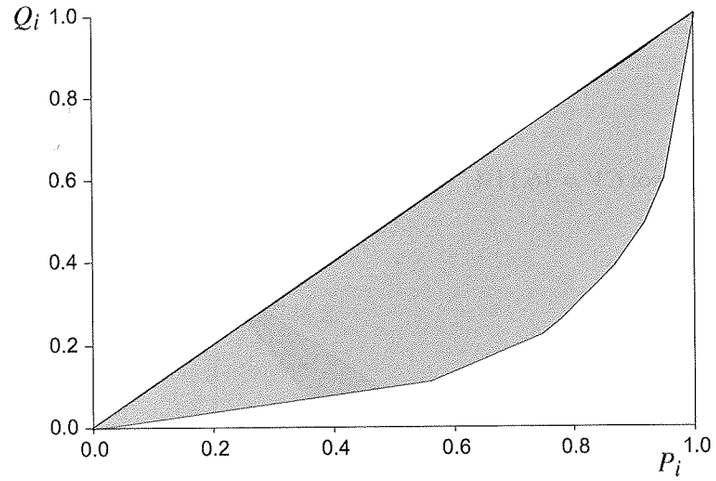
b) 0.333.

5.23



$G = 0.80$.

5.24 a)



$R = 0.64$.

b) 0.62.

5.25 0.70.

5.26 Il coefficiente di Gini non cambia se ogni osservazione viene moltiplicata per un numero positivo. Diminuisce se a ciascuna osservazione viene aggiunto lo stesso numero positivo.

5.27 0.728; 1.338.

Capitolo 6

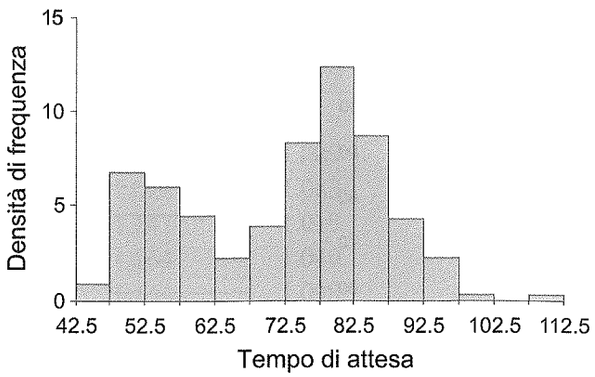
6.1 0.65.

6.2 a) 0.67; 0.

6.3 a) –0.01; 0.12. b) –0.38.

6.4 3.14; 9.38.

6.5



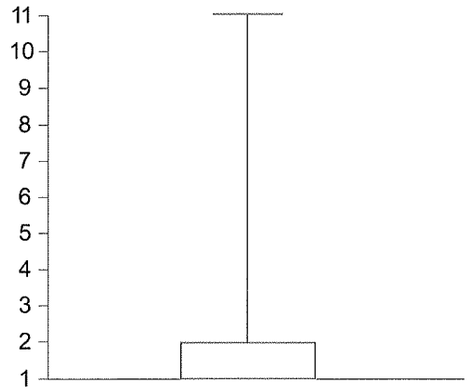
–0.44

Capitolo 7

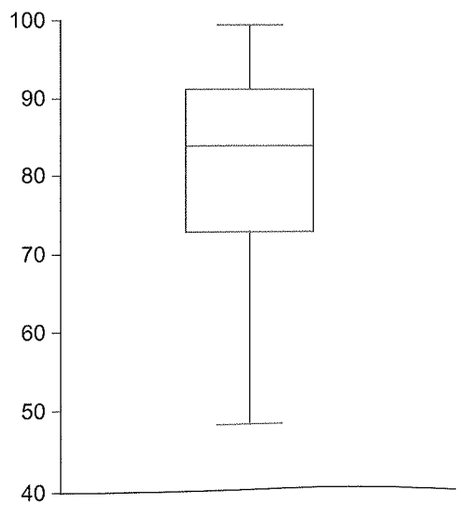
7.1 $\mu = 2.36$; $q_1 = 1$; $m = 2$; $q_3 = 3$; $\sigma = 1.26$; $\alpha_2 = 0.67$.

7.2 a) $\mu = 1.55$; $q_1 = 1$; $m = 1$; $q_3 = 2$; $\sigma = 1.07$; $\alpha_2 = 2.90$.

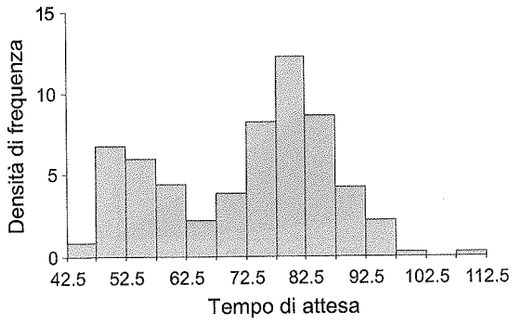
b)

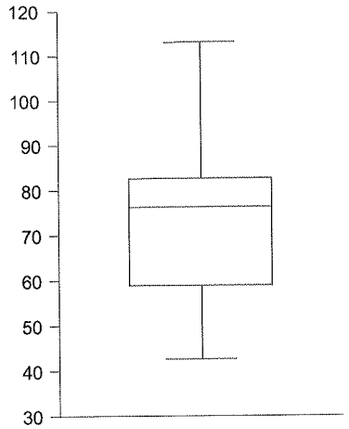


7.3



7.4





$\mu = 72.41$, $\sigma = 13.95$, $\alpha_1 = -0.265$.

7.6 Per $k = 1.5$, la disuguaglianza di Chebyshev dà 0.556 contro una proporzione effettiva di 0.913. Per $k = 2$, la disuguaglianza di Chebyshev dà 0.750 contro una proporzione effettiva di 0.957. Per $k = 2.5$, la disuguaglianza di Chebyshev dà 0.840 contro una proporzione effettiva di 0.989.

Capitolo 8

8.1 a)

Mese	Indice	Mese	Indice
Gennaio	1.000	Luglio	1.083
Febbraio	0.932	Agosto	0.872
Marzo	1.441	Settembre	1.045
Aprile	1.090	Ottobre	0.985
Maggio	1.172	Novembre	0.984
Giugno	1.209	Dicembre	1.016

b)

Mese	Indice	Mese	Indice
Gennaio	–	Luglio	0.895
Febbraio	0.932	Agosto	0.805
Marzo	1.546	Settembre	1.199
Aprile	0.757	Ottobre	0.942
Maggio	1.075	Novembre	0.999
Giugno	1.032	Dicembre	1.032

c) 0.14%.

8.2 a)

Anno	Italia	Regno Unito
2012	–	–
2013	1.055	1.089
2014	1.028	1.083
2015	1.147	1.019
2016	1.164	1.040
2017	1.134	1.098

b) Italia 10.44%; Regno Unito 6.53%.

8.3 a)

Anno	Indice	Anno	Indice
2010	–	2015	102.13
2011	101.72	2016	101.88
2012	99.02	2017	102.51
2013	99.80	2018	101.99
2014	101.44	2019	101.34

b)

Anno	Indice	Anno	Indice
2010	100.00	2015	104.14
2011	101.72	2016	106.10
2012	100.72	2017	108.76
2013	100.52	2018	110.93
2014	101.97	2019	112.41

c)

Anno	PIL	Anno	PIL
2010	24899.7	2015	25930.9
2011	25328.0	2016	26418.4
2012	25079.8	2017	27081.5
2013	25029.6	2018	27620.4
2014	25390.1	2019	27990.5

8.4 a)

Anno 2018	Indice	Anno 2019	Indice
I trimestre	–	I trimestre	98.29
II trimestre	98.97	II trimestre	100.99
III trimestre	102.44	III trimestre	100.34
IV trimestre	99.46	IV trimestre	98.24

b) –0.19%.

8.5 a) 110.42%. b) 110.47%.

8.6 a) 135.68%. b) 135.72%.

Capitolo 9

9.2 a)

Colore degli occhi				Totale
Azzurro	Chiaro	Bruno	Scuro	
7.9	13.5	29.6	49.0	100.0

b)

Colore degli occhi	Colore dei capelli				
	Chiaro	Rosso	Bruno	Scuro	Nero
Azzurro	193.9	38.1	284.8	185.4	15.7
Chiaro	426.7	83.9	626.8	408.0	34.6
Bruno	479.1	94.2	703.7	458.1	38.9
Scuro	355.2	69.8	521.7	339.6	28.8

c) $V = 0.277$.

9.3 a)

Colore (Lucentezza = IF)						Totale
D	E	F	G	H	I	
2.27	11.36	18.18	25.00	25.00	18.18	100.00

Colore						Totale
D	E	F	G	H	I	
5.19	14.29	26.62	21.10	19.81	12.99	100.00

b) 0.131.

9.4 a)

Reddito	Felicità			Totale
	Non troppo felice	Abbastanza felice	Molto felice	
Sotto la media	23.0	56.3	20.7	100.0
Vicino alla media	9.1	55.6	35.3	100.0
Sopra la media	7.7	53.8	38.5	100.0

b) 132.76.

9.5 a)

Area geografica	Ampiezza demografica						Totale
	≤ 2 000	2 001-5 000	5 001-20 000	20 001-50 000	50 001-100 000	≥ 100 001	
Nord-Ovest	54.1	23.9	18.0	3.1	0.6	0.2	100.0
Nord-Est	29.3	29.9	34.8	4.3	0.6	1.2	100.0
Centro	35.8	25.5	27.6	8.0	2.3	0.8	100.0
Sud	43.5	24.9	23.4	5.6	2.1	0.4	100.0
Isole	39.9	28.0	22.9	6.6	1.7	0.8	100.0

b) 386.24.

9.6 0.07.

9.7 a) 0.42.
b) 122 869.79; 67 446.04.
c) 0.55.

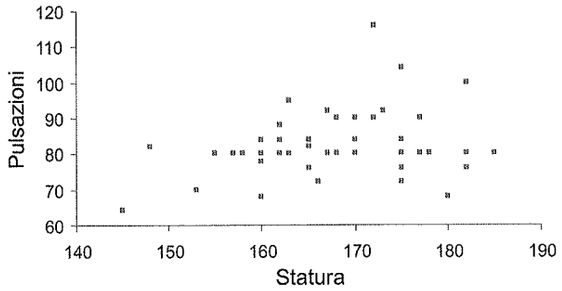
9.8 a) 1031.50; 1327.71; 1690.48
b) 0.36.

Capitolo 10

10.1 a) $\hat{y} = -45.35 + 2.47x$.
b) $r^2 = 0.827$; $\sigma_{RL} = 121.36$.

10.3 a) $b_1 = 1.27$; $b_0 = 0.38$.
b) $r^2 = 0.94$.

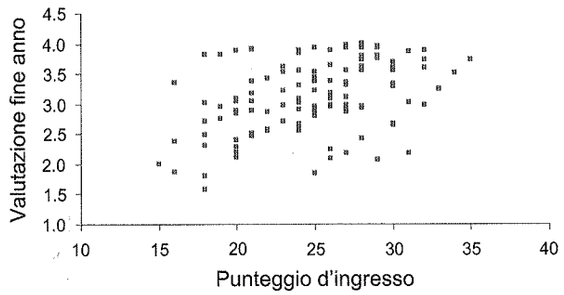
10.4 a)



b) $r^2 = 0.05$.

10.5 40; 61.

10.6 a)



b) Sì.

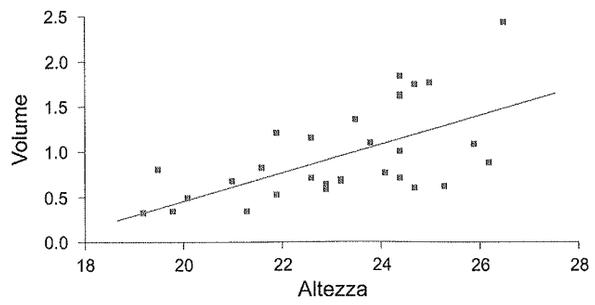
c) $b_0 = 1.671$; $b_1 = 0.057$; $r^2 = 0.18$.

10.7 $b_1 = 1.15$; $b_0 = -126.57$; $\sigma_{RL} = 10.56$.

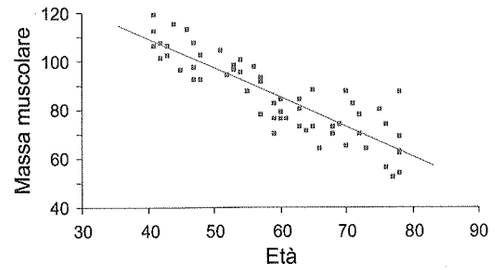
10.8 a) $b_1 = 0.143$; $b_0 = -2.467$.

b) 0.42.

c)



10.9 a) $\hat{y} = 156.35 - 1.19x$.



b) −2.38.

c) 79.

d) $r^2 = 0.75$.

10.10 a) $\hat{y} = 109.6 + 0.272x$.

b) $r^2 = 0.86$.

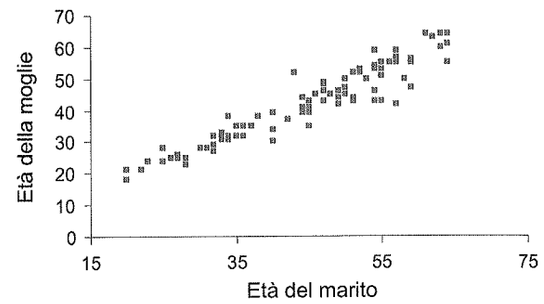
Capitolo 11

11.1 0.91.

11.2 0.97.

11.4 La seconda serie si ottiene dalla prima scambiando i valori di X e Y. Questa operazione non ha effetto sul valore di r. La terza serie si ottiene dalla seconda moltiplicando i valori di Y per 2. Sulla base della proprietà (c) della Proposizione 11.1, questa operazione non cambia il valore di r.

11.5 a)



b) $r = 0.95$.

11.6 −0.38.

11.7 $r = 0.60$.

11.8 a) $r = -0.50$.

Capitolo 12

12.1 a) $\{x : -1 \leq x \leq 10\}$. b) $\{x : x = 3\}$. c) $\{x : 3 < x \leq 4\}$.

12.2 a) $P(A) = 0.25$; $P(B) = 0.32$; $P(A \cap B) = 0.17$; $P(A \cup B) = 0.4$. b) 0.68.

12.3 a) 0.7. b) 0.22.

12.4 "Non aumentare le entrate". "Aumentare le entrate o tagliare le spese". "Non tagliare le spese". "Aumentare le entrate o non tagliare le spese". "Aumentare le entrate e tagliare le spese". "Aumentare le entrate e non tagliare le spese".

12.5 a) 0.7. b) 0.2. c) 0.9.

12.6 $P(\bar{A}) = 0.65$; $P(A \cap B) = 0$; $P(A \cup B) = 0.75$; $P(\bar{A} \cup \bar{B}) = 1$; $P(\bar{A} \cap \bar{B}) = 0.25$.

12.7 a) "Il tasso di rendimento sarà positivo." b) 0.14. c) "Il tasso di rendimento sarà negativo o nullo e, contemporaneamente, maggiore di 5". d) "Il tasso di rendimento sarà negativo, nullo o maggiore di 5." e) 0.67. f) 0.

12.8

	A	$\bar{A}$	
B	0.18	0.21	0.39
$\bar{B}$	0.30	0.31	0.61
	0.48	0.52	1

$P(A \cap \bar{B}) = 0.30$; $P(\bar{A} \cap \bar{B}) = 0.31$.

12.9 a) 0.90. b) 0.10. c) 0.50. d) 0.65. No; 0.25.

12.10 0.0459.

12.11 0.0965.

12.12 0.0018.

12.13 a) 0.01. b) 0.0077.

12.14 a) 1/6. b) 1/12.

12.15 a) 0.23. b) 0.08. c) 0.6364.

12.17 0.5981.

12.18 a) 0.1. b) 1/3. c) 1/4.

12.19 0.0187.

12.20 0.9661.

Capitolo 13

13.1 a)

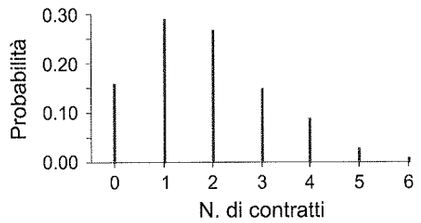
x	0	1	2
$f(x)$	3/6	1/6	2/6

b) Nella tabella che segue sono riportati i valori della funzione di ripartizione della variabile casuale per $x = 0, 1, 2$. Naturalmente: $F(x) = 0$, se $x < 0$; $F(x) = 0.5$, se $0 \leq x < 1$; $F(x) = 0.6667$, se $1 \leq x < 2$; $F(x) = 1$, se $x \geq 2$.

x	0	1	2
$F(x)$	3/6	4/6	6/6

c) 0.8333; 0.8056.

13.2 a)



b) 0.04; 0.54; 0.13. c) 1.85; 1.3592. d) 925.

13.3 a) 233 333.33. b) 466 666.67.

13.4 a) 552.17; 4.50.

b)

z	−2.0382	−0.4823	0.4068	1.7404
$f(z)$	0.11	0.35	0.41	0.13

13.5 La funzione è una funzione di densità in quanto è non negativa e il suo integrale tra 0 e b è uguale a 1.

13.6 1.5; 0.15.

13.7 a) 0.0004. b) 0.3471.

13.8 0.75.

13.9 L'informazione che l'incasso del bar nelle giornate festive è compreso nell'intervallo (5200, 7200) con probabilità maggiore o uguale a 0.6751.

13.10 159.1886; 190.8114.

13.11 0.795.

13.12 −54.

Capitolo 14

14.3 a)

x	0	1	2	3	4	5	6
$f(x)$	0.0007	0.0102	0.0595	0.1852	0.3241	0.3025	0.1176

b) 4.2; 1.26.

14.4 0.1240.

14.5 a) 0.3121. b) 0.3252. c) 0.3627.

14.6 a) 0.9990. b) 0.0861. c) 0.9885.

14.8 a) 0.2019. b) 0.4751.

14.9 0.3528.

14.10 0.9727; 0.9744.

14.14 a) 160.72. b) 165.62. c) 157.07.

14.15 a) (−0.84, 0.84). b) 1.04. c) −0.47.

14.16 a) 0.4483. b) 0.5363.

14.18 a) 0.0228. b) 0.5705. c) 0.1043.

14.19 a) 1.2793; 0.0943. b) 0.0096.

14.20 a) 25. b) 24.34. c) 0.7250.

Capitolo 15

15.1 a)

X	Y		
	−1	0	1
0	1/16	0	0
1	4/16	0	0
2	0	6/16	0
3	0	0	4/16
4	0	0	1/16

b)

x	0	1	2	3	4
$f_X(x)$	1/16	4/16	6/16	4/16	1/16

y	−1	0	1
$f_Y(y)$	5/16	6/16	5/16

15.2 a)

x	1	2	3
$f_X(x)$	0.58	0.34	0.08

b)

y	0	1	2
$f_{Y\|X}(y\|2)$	0.44	0.35	0.21

c) No, perché la distribuzione condizionata di cui al punto b) è diversa dalla distribuzione marginale di Y.

d) 0.08.

15.3 a) Le variabili casuali X e Y non sono indipendenti. b) 6.7; 4.81.

15.4 a)

x	1	2	3
$f_X(x)$	1/6	2/6	3/6

y	1	2	3
$f_Y(y)$	1/6	2/6	3/6

Le variabili casuali X e Y sono indipendenti, in quanto $f(x, y) = f_X(x) \cdot f_Y(y)$, per ogni coppia (x, y), $x = 1, 2, 3$ e $y = 1, 2, 3$.

b) Le distribuzioni condizionate di Y, dato $X = x$, per $x = 1, 2, 3$, sono:

y	1	2	3
$f_{Y\|X}(y\|1)$	1/6	2/6	3/6

y	1	2	3
$f_{Y\|X}(y\|2)$	1/6	2/6	3/6

y	1	2	3
$f_{Y\|X}(y\|3)$	1/6	2/6	3/6

Queste distribuzioni condizionate sono tutte uguali tra loro e uguali alla distribuzione marginale di Y.

c) Le distribuzioni condizionate di X, dato $Y = y$, per $y = 1, 2, 3$, sono:

x	1	2	3
$f_{X\|Y}(x\|1)$	1/6	2/6	3/6

x	1	2	3
$f_{X\|Y}(x\|2)$	1/6	2/6	3/6

x	1	2	3
$f_{X\|Y}(x\|3)$	1/6	2/6	3/6

Queste distribuzioni condizionate sono tutte uguali tra loro e uguali alla distribuzione marginale di X.

d) Siccome X e Y sono indipendenti, per la Proposizione 15.2, $\sigma_{XY} = 0$.

15.5 a) La distribuzione di probabilità congiunta e le distribuzioni di probabilità marginali di X e Y sono contenute nela tabella che segue:

X	Y 0	1	2	Totale
0	1/27	3/27	4/27	8/27
1	2/27	2/27	5/27	9/27
2	3/27	2/27	5/27	10/27
Totale	6/27	7/27	14/27	1

b) Le variabili casuali X e Y non sono indipendenti.

15.6 a) 8/7.

b)

X	Y 1	2	4
0	1/35	2/35	2/35
1	1/7	1/7	2/7
2	4/42	2/42	6/42

c) 2.7048.

d) 0.0048.

15.7 a) Le variabili casuali X e Y non sono indipendenti.

b) 0.75.

c) −0.14.

d) 4.

e) 15.7.

15.8 np_1p_2.

15.9 5 718.30; 94 803.04; vi è correlazione negativa.

15.11 Si verifica immediatamente che:
$\mathrm{E}(X) = 0, \mathrm{E}(Y) = -1/3, \mathrm{E}(XY) = 0$ e $\sigma_{XY} = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = 0$.

15.12 a) e^{-6}

b) Le variabili casuali sono indipendenti.

c) $f_X(x) = e^{-3} \cdot \dfrac{3^x}{x!}, x = 0, 1, 2, \ldots;\ f_Y(y) = e^{-3} \cdot \dfrac{3^y}{y!}, y = 0, 1, 2, \ldots$

15.15 a) 4; 9. b) 4; 7.

15.16 a) 200; 2890. b) 200; 1270.

15.17 0.6915.

Capitolo 16

16.1 a) 0.3889. b) 0.6333. c) La quantità $1 - 8.25/(n \cdot 0.09)$ è funzione crescente di n e tende a 1 al tendere di n a infinito.

16.2 0.9698.

16.3 0.0212.

16.5 0.9164.

16.6 0.0475.

16.7 Approssimazione normale: 0.1054; 0.9778; 0.6564; binomiale: 0.1103; 0.9778; 0.6637.

16.8 Approssimazione normale con correzione per continuità: 0.0778; approssimazione normale senza correzione per continuità: 0.1038; binomiale: 0.0769.

16.9 L'estrazione del campione casuale in questione va intesa con ripetizione; in questo modo, a ogni estrazione è associata una variabile casuale di Bernoulli con parametro $p = 0.076$. Il numero dei disoccupati nel campione è quindi assimilabile alla somma di 200 variabili casuali di Bernoulli indipendenti e aventi la stessa media p.

a) 0.4247. b) 0.2709.

I calcoli sono stati effettuati applicando la correzione per continuità.

16.10 a) 0.1335. b) 0.9971.

16.11 a) 0.9978. b) 1 (calcolato con Excel).

Capitolo 17

17.1 a) Parametro. b) Statistica. c) Parametro. d) Parametro. e) Statistica.

17.2 La popolazione di riferimento è data da tutti gli studenti universitari frequentanti il corso. Il parametro di interesse è la probabilità che uno studente viva in appartamento.

17.3 a)

1, 1	1, 2	1, 3	1, 4	1, 5
2, 1	2, 2	2, 3	2, 4	2, 5
3, 1	3, 2	3, 3	3, 4	3, 5
4, 1	4, 2	4, 3	4, 4	4, 5
5, 1	5, 2	5, 3	5, 4	5, 5

b)

1	1.5	2	2.5	3	3.5	4	4.5	5
1/25	2/25	3/25	4/25	1/5	4/25	3/25	2/25	1/25

17.4 a)

2, 2, 2	0.027	5, 2, 2	0.054	7, 2, 2	0.009
2, 2, 5	0.054	5, 2, 5	0.108	7, 2, 5	0.018
2, 2, 7	0.009	5, 2, 7	0.018	7, 2, 7	0.003
2, 5, 2	0.054	5, 5, 2	0.108	7, 5, 2	0.018
2, 5, 5	0.108	5, 5, 5	0.216	7, 5, 5	0.036
2, 5, 7	0.018	5, 5, 7	0.036	7, 5, 7	0.006
2, 7, 2	0.009	5, 7, 2	0.018	7, 7, 2	0.003
2, 7, 5	0.018	5, 7, 5	0.036	7, 7, 5	0.006
2, 7, 7	0.003	5, 7, 7	0.006	7, 7, 7	0.001

b)

2	3	3.67	4	4.67	5	5.33	5.67	6.33	7
0.027	0.162	0.027	0.324	0.108	0.216	0.009	0.108	0.018	0.001

c) 4.3; 0.87.

d)

2	5	7
0.216	0.756	0.028

17.6 a) Diventa il doppio. b) Si dimezza. c) Si riduce a 1/3.

17.7 5; 0.566.

17.8 a) 0.9664. b) 420.

17.9 a) 0.0062. b) 0.1056. c) 119.584.

17.10 a) 0.0838. b) (11.03, 11.77).

17.11 a) 0.8159. b) 0.3256.

17.12 a) 0.7224. b) 0.6530.

17.13 0.4047.

17.14 0.9357; 0.9216.

17.15 0.4935.

17.16 0.9896.

17.17

0	0.5	2	4.5	8
5/25	8/25	6/25	4/25	2/25

17.18 a) 0.4568. b) 0.1060.

17.19 a) 0.7762. b) 0.51.

Capitolo 18

18.1 267.76; 3.11.

18.2 70.89; 3.10.

18.3 a) T_2 è non distorto. b) $\sigma^2/4$; $3\sigma^2/10$. c) T_1.

18.4 $\text{MSE}(T) = b^2\text{Var}(\bar{X}) + [a + (b-1)\mu]^2$. T non è consistente.

18.5 a) Lo stimatore è non distorto. b) Lo stimatore non è consistente.

18.7 T_2 è più efficiente di T_1 in quanto $\text{MSE}(T_1) = 27.5$, mentre $\text{MSE}(T_2) = 26$.

18.8 0.68.

18.9 0.46.

18.10 385.

18.11 23.

18.12 1537.

18.13 $2\frac{n-1}{n^2}\sigma^4$.

18.15 Una stima del margine di errore è data da $2.576 \cdot \frac{1.6}{\sqrt{58}} = 0.54$.

18.17 0.0358.

Capitolo 19

19.1 (68.99, 81.13).

19.2 (89.78, 109.97).

19.3 0.90.

19.4 (1.33, 2.08).

19.5 (42 620, 44 060).

19.6 (5.47, 5.53); 0.06.

19.7 1.76; 0.88.

19.8 (0.64, 0.72).

19.9 (0.40, 0.51).

19.10 (0.84, 0.97).

19.11 (471.9, 3365.2).

19.12 (0.008, 0.019).

Capitolo 20

20.1 Ipotesi nulla: $\mu = 0.25$; ipotesi alternativa: $\mu > 0.25$. Valore della statistica test: $z_{\bar{x}} = 5.112$. Valore critico: $z_{0.99} = 2.326$. Decisione: rifiuto.

20.2 Ipotesi nulla: a) $\mu = 100$; ipotesi alternativa: $\mu < 100$. Valore della statistica test: $t_{\bar{x}} = -0.038$. Valore critico: $-t_{0.95} = -1.761$. Decisione: non rifiuto.

20.3 a) Ipotesi nulla: $\mu \leq 15$; ipotesi alternativa: $\mu > 15$.

b) Valore della statistica test: $t_{\bar{x}} = 3.614$. Valore critico: $t_{0.99} = 2.426$. Decisione: rifiuto.

c) Livello di significatività osservato: 0.0004. Decisione: rifiuto.

20.4 Ipotesi nulla: $\mu = 125$; ipotesi alternativa: $\mu > 125$. Valore della statistica test: $t_{\bar{x}} = 0.700$. Valore critico: $t_{0.99} = 2.602$. Decisione: non rifiuto.

20.5 a) Ipotesi nulla: $\mu \geq 69$; ipotesi alternativa: $\mu < 69$.

b) Valore della statistica test: $z_{\bar{x}} = -2.274$. Valore critico: $-z_{0.95} = -1.645$. Decisione: rifiuto.

20.6 Ipotesi nulla: $\mu = 230$; ipotesi alternativa: $\mu > 230$. Valore della statistica test: $z_{\bar{x}} = 4.45$. Livello di significatività osservato (si usi Excel): 0.0000043. Decisione: rifiuto.

20.7 Ipotesi nulla: $\mu = 45\,000$; ipotesi alternativa: $\mu < 45\,000$. Valore della statistica test: $z_{\bar{x}} = -3.43$. Livello di significatività osservato: 0.0003. Decisione: rifiuto.

20.8 a) Ipotesi nulla: $p \geq 0.8$; ipotesi alternativa: $p < 0.8$.

b) Valore della statistica test: $z_{\hat{p}} = 0.75$. Valore critico: $-z_{0.95} = -1.645$. Decisione: non rifiuto.

c) Livello di significatività osservato: 0.7734. Decisione: non rifiuto.

20.9 Ipotesi nulla: $p = 0.5$; ipotesi alternativa: $p \neq 0.5$. Valore della statistica test: $z_{\hat{p}} = 1.118$. Valore critico: $z_{0.975} = 1.96$. Decisione: non rifiuto.

20.10 Ipotesi nulla: $p = 0.7$; ipotesi alternativa: $p < 0.7$. Valore della statistica test: $z_{\hat{p}} = -0.804$. Valore critico: $-z_{0.90} = -1.282$. Decisione: non rifiuto.

20.11 Ipotesi nulla: $p \leq 0.02$; ipotesi alternativa: $p > 0.02$. Valore della statistica test: $z_{\hat{p}} = 0.099$. Valore critico: $z_{0.95} = 1.645$. Decisione: non rifiuto.

20.12 Ipotesi nulla: $\sigma^2 = 0.9$; ipotesi alternativa: $\sigma^2 > 0.9$. Valore della statistica test: $v = 10.670$. Valore critico: $\chi^2_{0.95} = 16.92$. Decisione: non rifiuto.

20.13 a) Ipotesi nulla: $\sigma^2 = 4.35$; ipotesi alternativa: $\sigma^2 > 4.35$. Valore della statistica test: $v = 28.193$. Valore critico: $\chi^2_{0.95} = 36.42$. Decisione: non rifiuto.

b) Livello di significatività osservato (da calcolare con Excel, oppure, in modo approssimato, con la Tavola C.2): 0.25. Decisione: non rifiuto.

20.15 Ipotesi nulla: $p = 0.7$; ipotesi alternativa: $p = 0.65$. Livello di significatività: $\alpha = 0.10$. Statistica test: $Z_{\hat{p}}$. Potenza:

$$P\left(\hat{p} < 0.7 - 1.282\sqrt{\frac{0.7 \cdot 0.3}{530}} = 0.6745\,;\ p = 0.65\right),$$

da cui, standardizzando, si ricava

$$P\left[(\hat{p} - 0.65)\Big/\sqrt{\frac{0.65 \cdot 0.35}{530}} < 1.18;\ p = 0.65\right] = 0.8810.$$

20.16 a) Ipotesi nulla: $\mu = 0$; ipotesi alternativa: $\mu > 0$. Valore della statistica test: $z_{\bar{x}} = 2.409$. Valore critico: $z_{0.99} = 2.326$. Decisione: rifiuto.

b) Ipotesi nulla: $\mu = 0$; ipotesi alternativa: $\mu = 30$. Statistica test: $Z_{\bar{X}}$. Potenza:

$$P\left(\bar{X} > 2.326\sqrt{\frac{12\,500}{120}} = 23.7396\,;\ \mu = 30\right),$$

da cui, standardizzando, si ricava

$$P\left[(\bar{X} - 30)\Big/\sqrt{\frac{12\,500}{120}} > -0.61;\ \mu = 30\right] = 0.7291.$$

20.17 La probabilità richiesta è data da

$$P\left(\bar{X} < 15 + 2.326\frac{4}{\sqrt{40}} = 16.4711\,;\ \mu = 16\right),$$

da cui, standardizzando, si ricava

$$P\left[(\bar{X} - 16)\Big/\frac{4}{\sqrt{40}} < 0.74;\ \mu = 16\right] = 0.7704.$$

20.18 Ipotesi nulla: $p = 0.5$; ipotesi alternativa: $p = 0.55$. Livello di significatività: $\alpha = 0.05$. Statistica test: $Z_{\hat{p}}$ bidirezionale. Potenza:

$$P\left(\hat{p} < 0.5 - 1.96\sqrt{\frac{0.5 \cdot 0.5}{80}} = 0.3904;\ p = 0.55\right)$$
$$+ P\left(\hat{p} > 0.5 + 1.96\sqrt{\frac{0.5 \cdot 0.5}{80}} = 0.6096;\ p = 0.55\right),$$

da cui, standardizzando, si ottiene

$$P\left[(\hat{p} - 0.55)\Big/\sqrt{\frac{0.55 \cdot 0.45}{80}} < -2.87;\ p = 0.55\right]$$
$$+ P\left[(\hat{p} - 0.55)\Big/\sqrt{\frac{0.55 \cdot 0.45}{80}} > 1.07;\ p = 0.55\right]$$
$$= 0.1444.$$

20.19 Ipotesi nulla: $\sigma^2 = 4.35$; ipotesi alternativa: $\sigma^2 = 7$. Livello di significatività: $\alpha = 0.05$. Statistica test: V. Potenza:

$$P\left(S^2 > \frac{\sigma_0^2}{n-1}\chi_{0.95};\ \sigma^2 = \sigma_1^2\right) = P\left(S^2 > \frac{4.35}{24}36.42 = 6.6011;\ \sigma^2 = 7\right),$$

da cui, moltiplicando entrambi i membri della disuguaglianza per 24/7, si ottiene

$$P\left(\frac{24 \cdot S^2}{7} > 22.63\ ;\ \sigma^2 = 7\right) = 0.5417.$$

Capitolo 21

21.1 (−17.17, −3.47).

21.2 Ipotesi nulla: $\delta = \mu_1 - \mu_2 = 0$; ipotesi alternativa: $\delta = \mu_1 - \mu_2 < 0$. Valore della statistica test: $t_{\bar{x}_1-\bar{x}_2} = -3.030$. Valore critico: $-t_{0.99} = -2.408$. Decisione: rifiuto.

21.3 a) Le due popolazioni devono essere normali con la stessa varianza.

b) Ipotesi nulla: $\delta = \mu_1 - \mu_2 = 0$; ipotesi alternativa: $\delta = \mu_1 - \mu_2 > 0$. Valore della statistica test: $t_{\bar{x}_1-\bar{x}_2} = 3.093$. Livello di significatività osservato: 0.0031. Decisione: rifiuto.

21.5 Ipotesi nulla: $\delta = \mu_1 - \mu_2 = 0$; ipotesi alternativa: $\delta = \mu_1 - \mu_2 < 0$. Valore della statistica test: $t_{\bar{x}_1-\bar{x}_2} = -4.274$. Valore critico: $-t_{0.95} = -1.673$. Decisione: rifiuto.

21.6 Ipotesi nulla: $\delta = \mu_1 - \mu_2 = 0$; ipotesi alternativa: $\delta = \mu_1 - \mu_2 > 0$. Valore della statistica test: $t'_{\bar{x}_1-\bar{x}_2} = 1.635$. Gradi di libertà: 42. Valore critico: $t_{0.95} = 1.682$. Decisione: non rifiuto.

21.7 Ipotesi nulla: $\delta = \mu_1 - \mu_2 = 0$; ipotesi alternativa: $\delta = \mu_1 - \mu_2 > 0$. Valore della statistica test: $t'_{\bar{x}_1-\bar{x}_2} = 4.929$. Gradi di libertà: 24. Valore critico: $t_{0.99} = 2.492$. Decisione: rifiuto.

21.8 (8.99, 15.01).

21.9 Ipotesi nulla: $\delta = \mu_1 - \mu_2 = 0$; ipotesi alternativa: $\delta = \mu_1 - \mu_2 > 0$. Valore della statistica test: $z_{\bar{x}_1-\bar{x}_2} = 10.273$. Valore critico: $z_{0.95} = 1.645$. Decisione: rifiuto.

21.10 (5.54, 10.46).

21.11 Ipotesi nulla: $\delta = \mu_1 - \mu_2 = 0$; ipotesi alternativa: $\delta = \mu_1 - \mu_2 < 0$. Valore della statistica test: $z_{\bar{x}_1-\bar{x}_2} = -3.563$. Livello di significatività osservato: 0.00018. Decisione: rifiuto.

21.12 (−0.019, 0.045).

21.13 Ipotesi nulla: $\delta = p_1 - p_2 = 0$; ipotesi alternativa: $\delta = p_1 - p_2 \neq 0$. Valore della statistica test: $z_{\hat{p}_1-\hat{p}_2} = 0.786$. Valore critico: $z_{0.995} = 2.576$. Decisione: non rifiuto.

21.14 Ipotesi nulla: $\delta = p_1 - p_2 = 0$; ipotesi alternativa: $\delta = p_1 - p_2 < 0$. Valore della statistica test: $z_{\hat{p}_1-\hat{p}_2} = -2.042$. Valore critico: $-z_{0.95} = -1.645$. Decisione: rifiuto.

21.16 Ipotesi nulla: $\sigma_1^2 = \sigma_2^2$; ipotesi alternativa: $\sigma_1^2 > \sigma_2^2$. Valore della statistica test: $f = 2.698$. Valore critico: $f_{0.99} = 2.47$. Decisione: rifiuto.

21.17 Ipotesi nulla: $\sigma_1^2 = \sigma_2^2$; ipotesi alternativa: $\sigma_1^2 > \sigma_2^2$. Valore della statistica test: $f = 9.308$. Valore critico: $f_{0.95} = 2.12$. Decisione: rifiuto.

21.18 La probabilità β è data da

$$\beta = P\left(\bar{X}_1 - \bar{X}_2 < 2.326\sqrt{\frac{69}{159} + \frac{71}{79}} = 2.6852\,;\ \delta = 5\right),$$

da cui, standardizzando, si ricava

$$\beta = P\left[\left(\bar{X}_1 - \bar{X}_2 - 5\right)\Big/\sqrt{\frac{69}{159} + \frac{71}{79}} < -2.31;\ \delta = 5\right] = 0.0104.$$

Capitolo 22

22.1 Ipotesi nulla: "La distribuzione del 2019 delle abitazioni occupate secondo il titolo di godimento non è mutata rispetto a quella osservata nel censimento del 2011"; ipotesi alternativa: "La distribuzione delle abitazioni occupate è diversa". Valore della statistica test: $\chi^2 = 25.873$. Valore critico: $\chi^2_{0.99} = 9.21$. Decisione: rifiuto.

22.2 Ipotesi nulla: "La popolazione generatrice è normale"; ipotesi alternativa: "La popolazione generatrice non è normale". Valore della statistica test: $\chi^2 = 2.463$ (le prime quattro frequenze attese sono state aggregate in un'unica categoria, lo stesso per le ultime quattro; rispetto alle classi riportate, sono state aggiunte le classi estreme "fino a 79.5" e "oltre 199.5"). Valore critico: $\chi^2_{0.95} = 11.07$. Decisione: non rifiuto.

22.3 Ipotesi nulla: "Il profitto in matematica e il sesso sono indipendenti"; ipotesi alternativa: "Il profitto in matematica e il sesso non sono indipendenti". Valore della statistica test: $\chi^2 = 6.684$. Valore critico (i gradi di libertà sono 3): $\chi^2_{0.99} = 11.34$. Decisione: non rifiuto.

22.4 Ipotesi nulla: "C'è indipendenza tra disturbi cardiaci e pressione sanguigna"; ipotesi alternativa: "Esiste una relazione associativa tra pressione sanguigna e disturbi cardiaci". Valore della statistica test: $\chi^2 = 28.966$. Valore critico (i gradi di libertà sono 3): $\chi^2_{0.95} = 7.81$. Decisione: rifiuto.

22.5 Ipotesi nulla: "Le due terapie sono equivalenti"; ipotesi alternativa: "Le due terapie non sono equivalenti". Valore della statistica test: $\chi^2 = 2.140$. Valore critico (i gradi di libertà sono 2): $\chi^2_{0.99} = 9.21$. Decisione: non rifiuto.

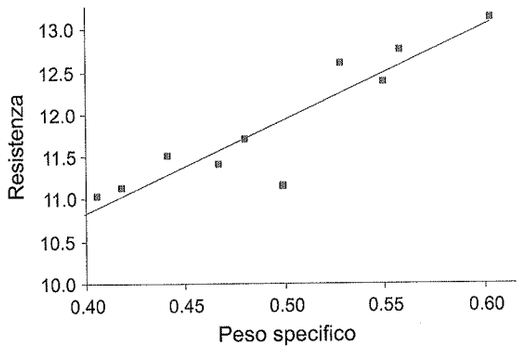
Capitolo 23

23.2 a) $\hat{y} = -45.35 + 2.47x$.

b) All'aumentare di una unità di personale medico, il personale ausiliario aumenta mediamente di 2.47 unità.

c) Ipotesi nulla: $\beta_1 = 0$; ipotesi alternativa: $\beta_1 \neq 0$. Valore della statistica test: $t_{b_1} = 5.351$. Valore critico: $t_{0.975} = 2.447$. Decisione: rifiuto.

23.3 a) $\hat{y} = 6.51 + 10.83x$;



b) (6.89, 14.77); ciò significa che all'aumentare di una unità del peso specifico del legno, la resistenza cresce, molto verosimilmente, in una misura che va da 6.89 a 14.77.

23.4 a) $\hat{y} = -1.046 + 0.056x$; $r^2 = 0.94$; $\hat{y} = -2.467 + 0.143x$; $r^2 = 0.36$.

b) Il modello che meglio si adatta ai dati osservati è il primo.

c) Ipotesi nulla: $\beta_1 = 0$; ipotesi alternativa: $\beta_1 \neq 0$. Valore della statistica test: $t_{b_1} = 20.484$. Livello di significatività osservato: $8.5744 \cdot 10^{-19}$. Decisione: rifiuto.

23.5 a) $\hat{y} = 0.379 + 1.269x$.

b) Ipotesi nulla: $\beta_0 = 0$; ipotesi alternativa: $\beta_0 \neq 0$. Valore della statistica test: $t_{b_0} = 0.282$. Valore critico: $t_{0.975} = 2.101$. Decisione: non rifiuto.

c) 1.289.

d)

x_i	Residuo	x_i	Residuo
9.5	−1.546	9.8	−0.932
5.0	0.055	19.0	1.109
23.0	−0.247	14.6	−2.519
15.2	−2.393	8.3	−1.199
11.4	3.705	21.6	0.958
11.8	4.490	26.5	−2.959
12.1	1.003	4.8	0.313
22.0	0.642	21.7	−2.271
28.2	4.150	18.0	3.298
12.1	−1.397	28.0	−2.992

23.6 a) $\hat{y} = 184.92 - 2.61x$. b) 45.7494.

c) (−3.20, −2.03). d) (155.29, 162.26); (144.14, 173.41).

23.7 a) $\hat{y} = 13.36 - 0.0016x$. b) (11.02, 15.69); (−0.0025, −0.0007).

Capitolo 24

24.2 24.70; (23.62, 25.78)

24.3 a) 6500. b) 1.07; 0.06.

24.4 28 200 848; $3\,381\,567^2$.