

Statistica e valutazione della Formazione

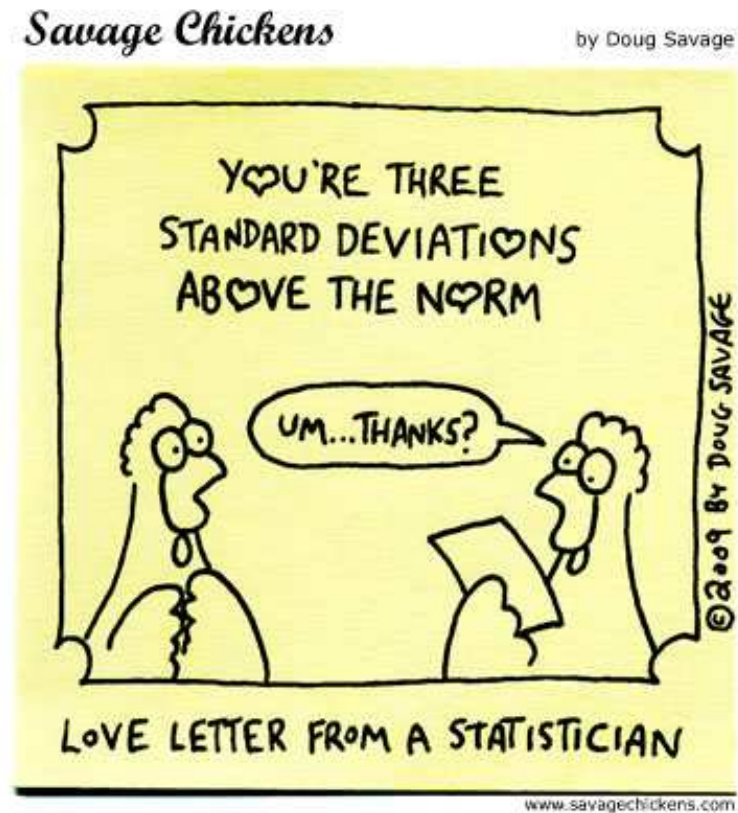
Slides 4

A.A. 2025-2026

Docente: ANNA LINA SARRA

Modulo 1: elementi di statistica descrittiva

- Le misure di variabilità



Obiettivi della lezione

Al termine di questa lezione sarai in grado di...

① Definire la variabilità statistica

Comprendere cosa misura la variabilità e perchè è importante non affidarsi solo alla media

② Calcolare gli indici di eterogeneità

Misurare la variabilità per caratteri qualitativi e quantitativi con l'indice di eterogeneità assoluto e relativo

③ Applicare gli indici di dispersione

Calcolare scostamento semplice medio, deviazione standard e varianza per distribuzioni disaggregate e di frequenza

④ Utilizzare il box-plot

Costruire e interpretare il diagramma a scatola, individuare valori anomali ed estremi e confrontare distribuzioni diverse

⑤ Confrontare la variabilità tra distribuzioni

Usare il coefficiente di variazione e gli indici relativi per confrontare distribuzioni con diverse unità di misura o intensità media

⑥ Applicare le conoscenze a casi reali

Risolvere esercizi su voti, redditi e dati reali per consolidare i concetti appresi

Variabilità

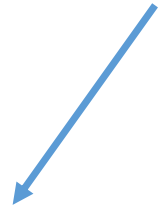
Per variabilità si intende l'attitudine dei fenomeni, naturali e sociali, a manifestarsi in modi differenti.

- La variabilità è l'attitudine di un carattere a presentare modalità differenti nel collettivo in esame.
- La distribuzione di un carattere presenta variabilità nulla se su tutte le unità statistiche si rileva la stessa modalità. In tal caso tutti gli indici di variabilità assumono valore zero.

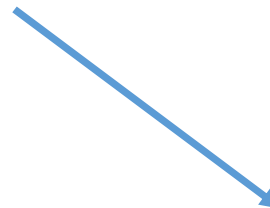
Come la valutiamo/calcoliamo?

Principalmente attraverso **indici di variabilità**.

Stesso concetto degli indici che abbiamo visto fino ad ora per la tendenza centrale: cerchiamo di sintetizzare un aspetto dei dati con un solo numero (o pochi numeri). Esistono molti indici di variabilità, ma presentano tutti due caratteristiche in comune:



1) Sono sempre maggiori o uguali a zero (non ha senso la variabilità negativa)



2) Valgono zero se e solo se la variabile presenta sempre una sola modalità per tutte le unità statistiche

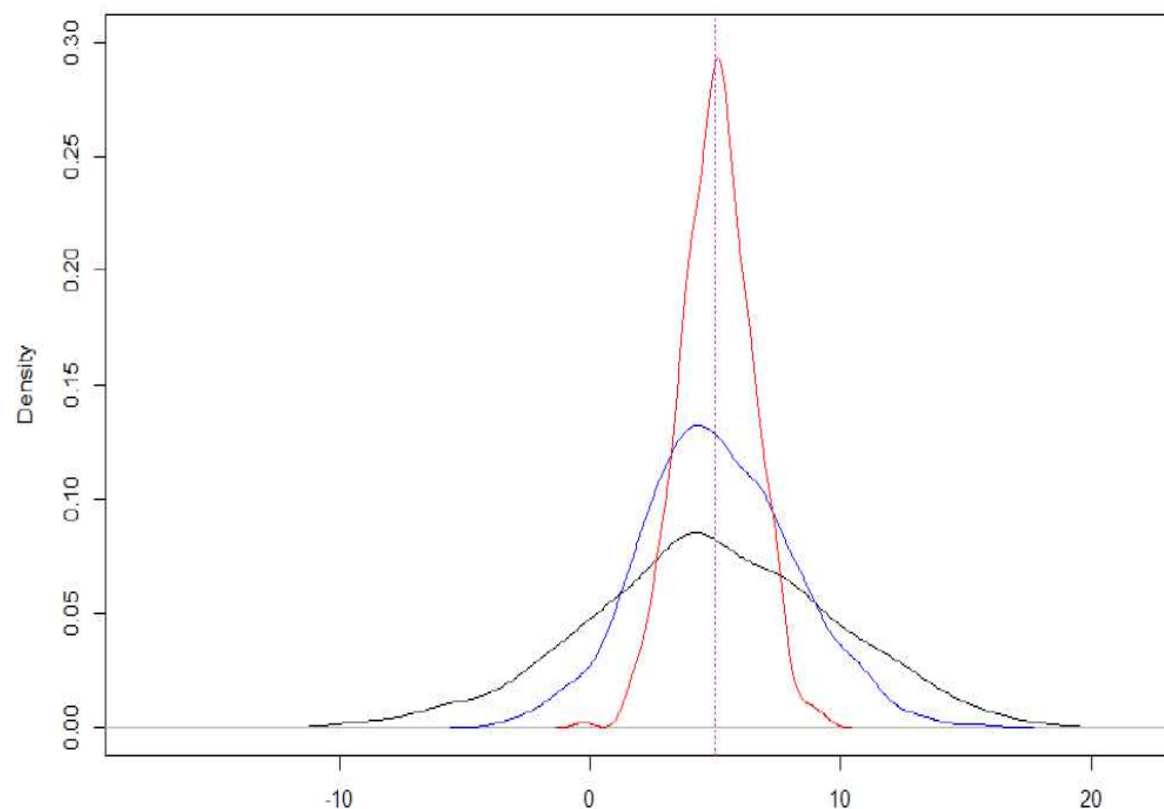
Ma anche alcune rappresentazioni grafiche dicono qualcosa sulla variabilità

Quale presenta maggiore variabilità?

Sono curve di densità
(simili a istogrammi)

3 gruppi di persone

Centrate sul valore 5 ma con
Varianze piuttosto diverse



Cosa aggiunge la variabilità?

Poniamo di avere tre studenti tutti con la stessa media, uguale a 24: Giorgio, Alfredo e Laura.

Tutti e 3 hanno fatto 4 esami. Voi riterreste le loro «carriere universitarie» equivalenti giusto?

Medie uguali...ma moda?

Quale dei 3 studenti mostra minore variabilità?

Quale invece la maggiore variabilità?

Interpretando solo la media, li avremmo considerati **UGUALI**,
MA IN REALTA' SONO STUDENTI CON PERCORSI DIVERSI

Questi sono i voti da loro presi:

<u>Giorgio</u>	<u>Alfredo</u>	<u>Laura</u>
24	22	18
24	23	18
24	25	30
24	26	30

Le misure variabilità sono indici che

1. Segnalano quanto sono tra loro diversi i valori della variabile
2. Evidenziano il grado di dispersione di ciascun valore rispetto ad un punto di riferimento.
3. Misurano la diversità tra due termini della distribuzione o tra due quantili.

Variabilità per caratteri di natura qualsiasi

Eterogeneità

- Misura la variabilità nelle distribuzioni secondo **caratteri qualitativi** o quantitativi
 - **minima eterogeneità (massima omogeneità):** tutte le unità del collettivo hanno la stessa modalità del carattere
 - **massima eterogeneità (minima omogeneità):** le modalità presentano tutte la stessa frequenza.

Indice di eterogeneità

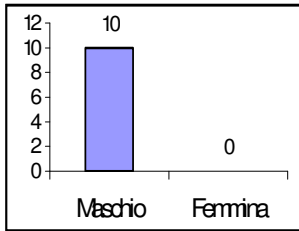
$$E = 1 - \sum_{i=1}^k f_i^2$$

valore minimo 0 (*eterogeneità nulla*)

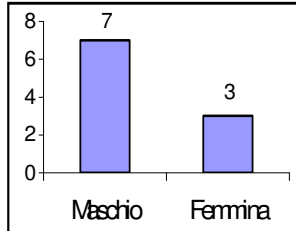
valore massimo $\left(\frac{k-1}{k}\right)$ (*eterogeneità massima*)

Eterogeneità: esempi

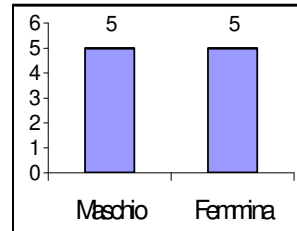
x_i	n_i
Maschio	10
Femmina	0
Totale	10



x_i	n_i
Maschio	7
Femmina	3
Totale	10



x_i	n_i
Maschio	5
Femmina	5
Totale	10



**Eterogeneità
nulla**

**Eterogeneità
massima**

x_i	n_i	f_i	f_i^2
Maschio	10	1	1
Femmina	0	0	0
Totale	10	1	1

$$E = 1 - \sum_{i=1}^k f_i^2 = 1 - 1 = 0$$

x_i	n_i	f_i	f_i^2
Maschio	7	0,7	0,49
Femmina	3	0,3	0,09
Totale	10	1	0,58

$$E = 1 - \sum_{i=1}^k f_i^2 = 1 - 0,58 = 0,42$$

x_i	n_i	f_i	f_i^2
Maschio	5	0,5	0,25
Femmina	5	0,5	0,25
Totale	10	1	0,5

$$E = 1 - \sum_{i=1}^k f_i^2 = 1 - 0,5 = 0,5$$

Indice di eterogeneità: esempio

Ateneo di iscrizione alla Laurea specialistica	n_i	f_i	f_i^2
Stesso Ateneo	1808	0.028	0.0008
Altro Ateneo stessa area	7513	0.118	0.0138
Nord	14756	0.231	0.0534
Centro	35626	0.558	0.3112
Sud-Isole	4163	0.065	0.0042
Totale	63866	1	0.3834

$$E = 1 - \sum_{i=1}^k f_i^2$$

$$E = 1 - 0.3834 = 0.6166$$

Indice di eterogeneità relativo esempio

$$INDICE\ RELATIVO = \frac{INDICE\ ASSOLUTO - MIN(INDICE\ ASSOLUTO)}{MAX(INDICE\ ASSOLUTO) - MIN(INDICE\ ASSOLUTO)}$$

Ateneo di iscrizione alla Laurea specialistica	n_i	f_i	f_i^2
Stesso Ateneo	1808	0.028	0.0008
Altro Ateneo stessa area	7513	0.118	0.0138
Nord	14756	0.231	0.0534
Centro	35626	0.558	0.3112
Sud-Isole	4163	0.065	0.0042
Totale	63866	1	0.3834

$$E = 1 - 0.3834 = 0.6166$$

$$\max(E) = \frac{k-1}{k} = \frac{5-1}{5} = 0.80$$

L'indice di eterogeneità relativo si calcola dividendo l'indice di eterogeneità per il suo massimo

$$e = \frac{E}{\max E} = \frac{0.6166}{0.80} = 0.77$$

Variabilità per caratteri quantitativi: Dispersione rispetto alla media aritmetica

- La dispersione rappresenta la variabilità delle modalità presentata del carattere rispetto ad un valore di sintesi scelto, ad esempio la media aritmetica della distribuzione
- Gli indici che danno una misura della variabilità dei valori della distribuzione rispetto alla media aritmetica sono lo **scostamento semplice medio** e lo **scostamento quadratico medio**.

Misura della variabilità: scostamento semplice medio

- Lo **scostamento semplice medio** è la media aritmetica degli scarti dalla media presi in valore assoluto

*Distribuzione
disaggregata*

$$S_{\mu} = \frac{|x_1 - \mu| + |x_2 - \mu| + \dots + |x_N - \mu|}{N}$$

$$= \frac{1}{N} \sum_{i=1}^N |x_i - \mu|.$$

*Distribuzione di
frequenza*

$$S_{\mu} = \frac{|x_1 - \mu| \cdot n_1 + |x_2 - \mu| \cdot n_2 + \dots + |x_k - \mu| \cdot n_k}{N}$$

$$= \frac{1}{N} \sum_{i=1}^k |x_i - \mu| \cdot n_i = \sum_{i=1}^k |x_i - \mu| \cdot f_i$$

Scostamento semplice medio per la distribuzione disaggregata: calcolo

$$S_{\mu} = \frac{1}{N} \sum_{i=1}^N |x_i - \mu|$$

x_i	$ x_i - \mu $
0	7.14
5	2.14
6	1.14
8	0.86
9	1.86
10	2.86
12	4.86
Totale	20.86

Numero di ore trascorse su Internet in una settimana di 7 ragazzi

0, 5, 6, 8, 9, 10, 12

□ *Scostamento semplice medio:*

$$S_{\mu} = \frac{|0 - 7.14| + |5 - 7.14| + \dots + |12 - 7.14|}{7}$$

$$\mu = 7.14 \quad \text{Numero di ore trascorse su Internet}$$

$$= \frac{20.86}{7} = 2.98 \quad \text{Numero di ore trascorse su Internet}$$

Scostamento semplice medio per una distribuzione di frequenze : calcolo

$$S_{\mu} = \frac{1}{N} \sum_{i=1}^k |x_i - \mu| \cdot n_i$$

N. ore trascorse su Internet

xi	ni	xi*ni	xi-μ	xi-μ *ni
1	1	1	4.42	4.42
2	2	4	3.42	6.84
3	3	9	2.42	7.26
4	3	12	1.42	4.26
5	1	5	0.42	0.42
6	2	12	0.58	1.16
7	1	7	1.58	1.58
8	2	16	2.58	5.16
9	3	27	3.58	10.74
10	1	10	4.58	4.58
	19	103		46.42

□ Media aritmetica

$$\mu = \frac{103}{19} = 5.42$$

□ Scostamento semplice medio

$$S_{\mu} = \frac{|1-5.42| \cdot 1 + \dots + |10-5.42| \cdot 1}{19}$$

$$= \frac{46.42}{19} = 2.42$$

Misura della variabilità: scostamento quadratico medio

Lo **scostamento quadratico medio** o **deviazione standard** è la **media quadratica degli scarti dalla media**

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_N - \mu)^2}{N}} = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}} \quad \text{Distribuzione disaggregata}$$

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 n_1 + (x_2 - \mu)^2 n_2 + \dots + (x_k - \mu)^2 n_k}{N}} = \sqrt{\frac{\sum_{i=1}^k (x_i - \mu)^2 n_i}{N}} \quad \text{Distribuzione di frequenza}$$

Formula
alternativa

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2 - \mu^2}$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^k x_i^2 n_i - \mu^2}$$

Scostamento quadratico medio per la distribuzione disaggregata: calcolo

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

x_i	$(x_i - \mu)^2$
0	51.02
5	4.59
6	1.31
8	0.73
9	3.45
10	8.16
12	23.59
Totale	92.86

Numero di ore trascorse su Internet di 7 ragazzi

0, 5, 6, 8, 9, 10, 12

□ **Scostamento quadratico medio:**

$$\sigma = \sqrt{\frac{(0 - 7.14)^2 + (5 - 7.14)^2 + \dots + (12 - 7.14)^2}{7}}$$

$$= \sqrt{\frac{92.86}{7}} = 3.64$$

Numero di ore trascorse su Internet di 7 ragazzi

$\mu = 7.14$ **Numero di ore trascorse su Internet di 7 ragazzi**

σ per una distribuzione di frequenze a modalità singole: calcolo

Numero di ore trascorse su Internet di 7 ragazzi

$$\sigma = \sqrt{\frac{\sum_{i=1}^k (x_i - \mu)^2 n_i}{N}}$$

xi	ni	xi*ni	(xi- μ)^2	(xi- μ)^2*ni
1	1	1	19.55	19.55
2	2	4	11.70	23.41
3	3	9	5.86	17.58
4	3	12	2.02	6.06
5	1	5	0.18	0.18
6	2	12	0.34	0.67
7	1	7	2.49	2.49
8	2	16	6.65	13.30
9	3	27	12.81	38.43
10	1	10	20.97	20.97
	19	103		142.63

□ Media aritmetica

$$\mu = \frac{103}{19} = 5.42$$

□ Scostamento quadratico medio

$$\sigma = \sqrt{\frac{(1-5.42)^2 \cdot 1 + \dots + (1-5.42)^2 \cdot 10}{19}} =$$
$$= \sqrt{\frac{142.63}{19}} = 7.50$$

Proprietà degli indici S_μ e σ

- Assumono il valore 0 nel caso di assenza di variabilità (*tutte le unità statistiche presentano la stessa modalità del carattere*).
- **Sono espressi nella stessa unità di misura del carattere considerato.**
- Non cambiano se a ciascun termine della distribuzione si aggiunge una quantità costante positiva o negativa.
- La moltiplicazione di ciascun termine della distribuzione per una costante, positiva o negativa, ha come conseguenza la moltiplicazione degli indici per il valore assoluto della costante.

Varianza

Il quadrato della deviazione standard si chiama **varianza**

(è espressa in unità di misura del carattere al quadrato).

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_N - \mu)^2}{N} = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \quad \text{Distribuzione disaggregata}$$

$$\sigma^2 = \frac{(x_1 - \mu)^2 n_1 + (x_2 - \mu)^2 n_2 + \dots + (x_k - \mu)^2 n_k}{N} = \frac{\sum_{i=1}^k (x_i - \mu)^2 n_i}{N} \quad \text{Distribuzione di frequenza}$$

Formula
alternativa

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N x_i^2 - \mu^2$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^k x_i^2 n_i - \mu^2$$

Devianza

La somma dei quadrati degli scarti dalla media (numeratore della varianza) si chiama **devianza**

(è espressa in unità di misura del carattere al quadrato)

$$D = (x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_N - \mu)^2 = \sum_{i=1}^N (x_i - \mu)^2 \quad \text{Distribuzione disaggregata}$$

$$D = (x_1 - \mu)^2 n_1 + (x_2 - \mu)^2 n_2 + \dots + (x_k - \mu)^2 n_k = \sum_{i=1}^k (x_i - \mu)^2 n_i \quad \text{Distribuzione e di frequenza}$$

Varianza e devianza per la distribuzione disaggregata: calcolo

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

x_i	$(x_i - \mu)^2$
0	51.02
5	4.59
6	1.31
8	0.73
9	3.45
10	8.16
12	23.59
Totale	92.86

Numero di ore trascorse su Internet di 7 ragazzi

0, 5, 6, 8, 9, 10, 12

$$\sigma = \sqrt{\frac{92.86}{7}} = 3.64 \quad \text{N. di ore trascorse su Internet}$$

$$\sigma^2 = \frac{92.86}{7} = 13.27 \quad \text{N. di ore trascorse su Internet}^2$$

$$\text{devianza} = 92.86 \quad \text{N. di ore trascorse su Internet}^2$$

$$\mu = 7.14 \quad \text{N. di ore trascorse su Internet}$$

Diversità tra due termini della distribuzione

La distanza tra il valore più piccolo e il valore più grande è indicata come **Campo di variazione (Range)** .



Se utilizziamo la mediana come indice del centro della distribuzione, dividendo la distribuzione in due parti, possiamo usare la stessa idea per misurare la dispersione .

La distanza tra Q_1 e Q_3 è una misura di variabilità detta **INTERVALLO INTERQUARTILE** .



Campo di variazione e differenza interquartile

Sia $x_1, x_2, \dots, x_N$ una distribuzione disaggregata. Sia $y_1, y_2, \dots, y_N$ la stessa distribuzione con i termini disposti in ordine non decrescente.

**campo
di variazione**

$$\Delta_c = y_N - y_1$$

Differenza tra
massimo e minimo

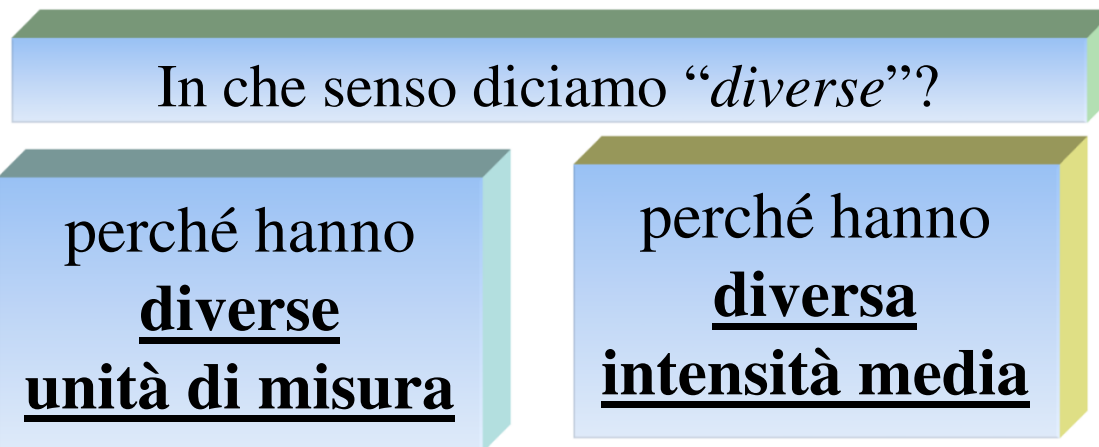
**Intervallo
interquartile**

$$\Delta_q = q_3 - q_1$$

Differenza tra terzo
e primo quartile

Indici di variabilità relativa

Il problema si pone nel confrontare gli indici di variabilità di 2 o più distribuzioni *diverse*.



Indice di variabilità relativa:
numero puro a-dimensionale che elimina l’influenza dell’unità di misura e dell’intensità media.

Indici di variabilità percentuali

- Si chiama **indice di variabilità percentuale** il rapporto, moltiplicato per 100, tra un indice di variabilità assoluto e la media aritmetica.
- Particolare rilievo per le applicazioni ha **coefficiente di variazione**

$$CV = \frac{\sigma}{\mu} \cdot 100$$

Coefficiente di variazione: esempio

È maggiormente variabile il reddito medio annuo familiare o il numero di componenti della famiglia?

Reddito
(in migliaia)

Media : 33364 euro

Deviazione standard : 24636 euro

Numero di
componenti

Media : 2.77 componenti

Deviazione standard : 1.31 componenti

$$CV(\text{reddito}) = \frac{24636 \text{ euro}}{33364 \text{ euro}} 100 = 74\%$$

$$CV(\text{componenti}) = \frac{1.31 \text{ componenti}}{2.77 \text{ componenti}} 100 = 47\%$$

Coefficiente di variazione: esempio

Il reddito medio annuo è maggiormente variabile nell'insieme delle famiglie con 2 o con 4 componenti?

Reddito
(2 componenti)

Media : 24451 euro

Deviazione standard : 21218 euro

Reddito
(4 componenti)

Media : 49260 euro

Deviazione standard : 26050 euro

$$CV(\text{reddito} - 2\text{componenti}) = \frac{21218 \text{ euro}}{24451 \text{ euro}} 100 = 86.4\%$$

$$CV(\text{reddito} - 4\text{componenti}) = \frac{26050 \text{ euro}}{49260 \text{ euro}} 100 = 52.9\%$$

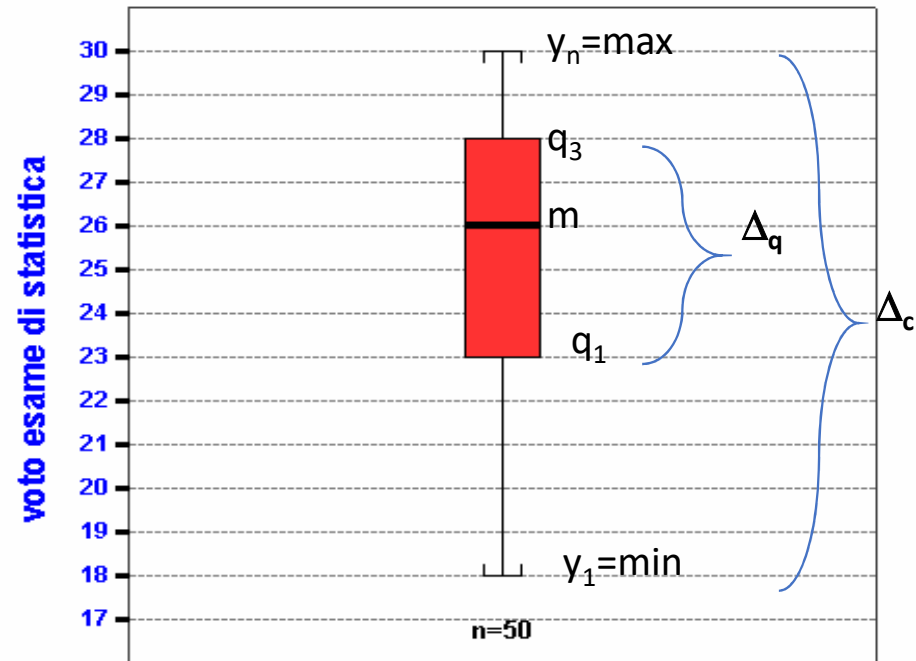
IL BOX-PLOT

- Una descrizione sintetica e abbastanza completa di una distribuzione di frequenze secondo un carattere quantitativo **è data dal box-plot**; questo è un riassunto a cinque numeri.
- I numeri sono i seguenti:
 - - la mediana (che dà informazioni sulla tendenza centrale)
 - - il primo e terzo quartile (la cui differenza dà informazioni sulla variabilità)
 - - i due estremi (la modalità più grande e la modalità più piccola)
- Questi numeri forniscono una descrizione sintetica di un insieme di dati anche quando il numero di unità osservate è elevato.

Box plot o Diagramma a scatola

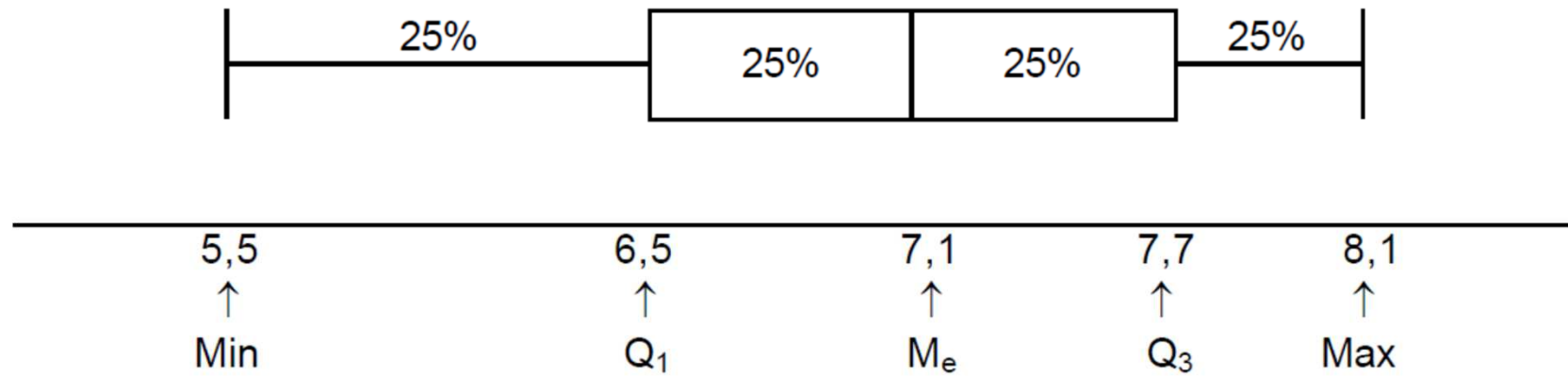
Il box plot di una distribuzione è un grafico caratterizzato da tre elementi principali:

- una linea che indica la posizione della **mediana** della distribuzione
- un rettangolo (box) i cui estremi sono determinati in base ai **quartili Q1 e Q3** della distribuzione e la cui altezza indica la variabilità dei valori prossimi alla mediana
- due segmenti che partono dal rettangolo i cui estremi sono determinati in base ai **valori minimo e massimo** della distribuzione



	y_1	q_1	m	q_3	y_n
voto esame di statistica	18	23	26	28	30

INTERPRETAZIONE DEL BOXPLOT



INTERPRETAZIONE DEL BOXPLOT

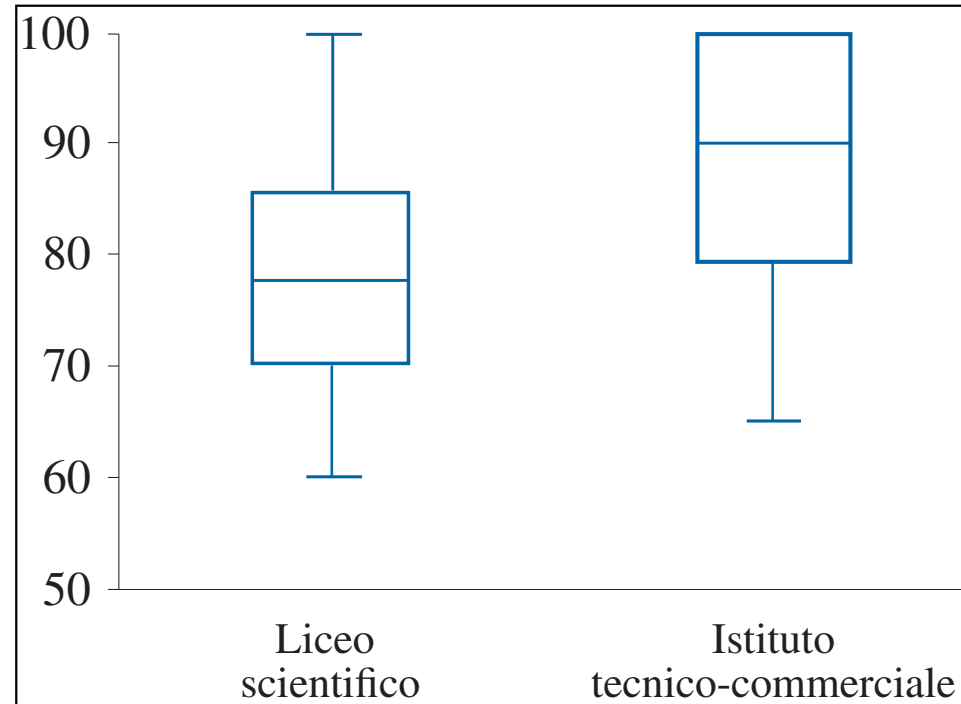
Il box-plot è utile perché riassume mediante pochi numeri molte informazioni su una distribuzione di frequenze.

- **La mediana** riassume la tendenza centrale della distribuzione.
- I **quartili** danno un'indicazione sulla variabilità, perché con essi si calcola lo scarto interquartile (misura più robusta del campo di variazione).
- La **posizione della mediana** rispetto ai quartili fornisce altre utili informazioni (in particolare **sulla asimmetria della distribuzione**).
- Gli estremi forniscono indicazioni non solo sul valore massimo e valore minimo ma soprattutto sull'eventuale **presenza di dati con caratteristiche anomale**.

Diagramma a scatola: esempio

Aspetti salienti delle due distribuzioni

- ❑ **Intensità media** (*mediana*)
- ❑ **La misura della variabilità** (*campo di variazione e differenza interquartile*)
- ❑ La presenza di **asimmetria** positiva o negativa che si desume dal raffronto delle distanze del primo e del terzo quartile dalla mediana



Voti riportati all'esame di Stato in un campione di studenti del corso di laurea in Economia aziendale distinti per scuola frequentata

Valori anomali ed estremi

Un dato è **anomalo** se:

- è maggiore del valore $Q3 + 1.5\Delta_q$
- è minore del valore $Q1 - 1.5\Delta_q$

Un dato è **estremo** (estremamente anomalo) se

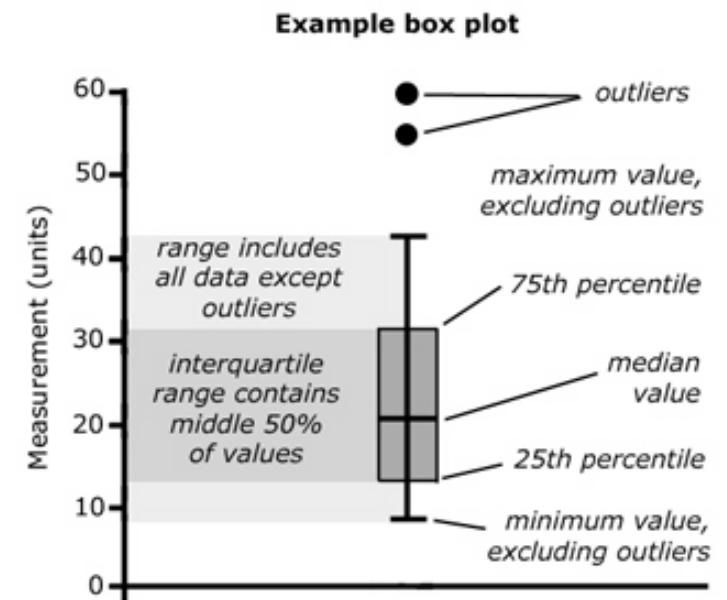
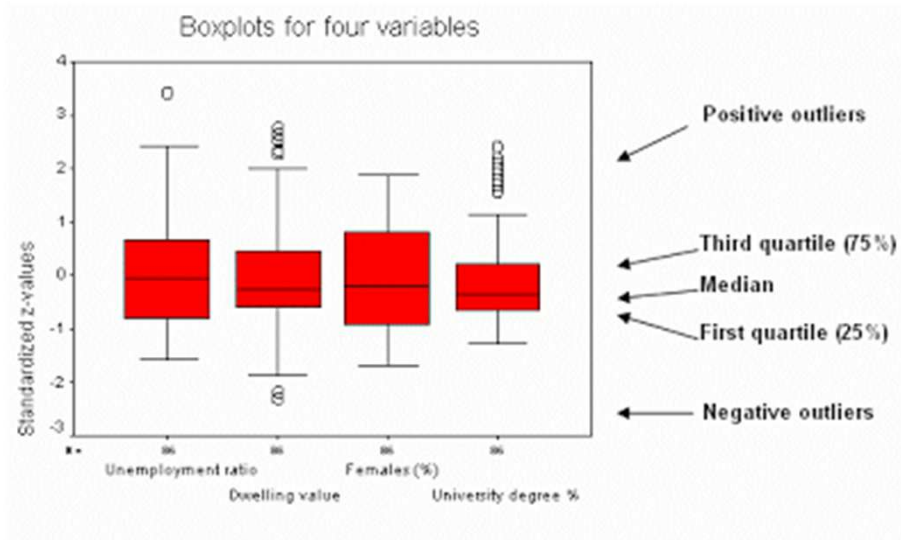
- è maggiore del valore $Q3 + 3\Delta_q$
- è minore del valore $Q1 - 3\Delta_q$

Boxplot e valori anomali ed estremi

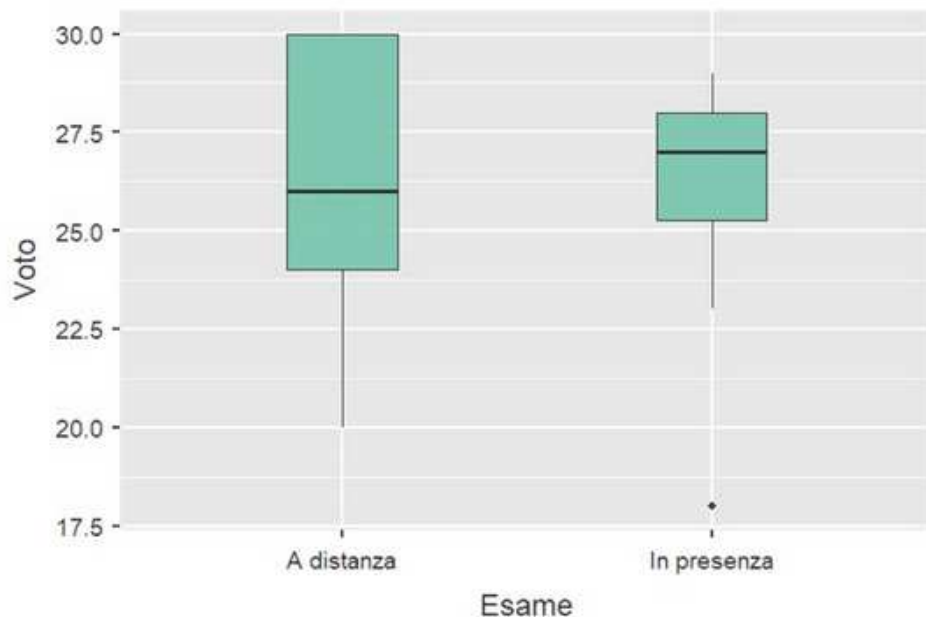
I valori anomali ed estremi possono essere rappresentati nel boxplot.

In questo caso i «baffi» (segmenti che partono dal rettangolo) hanno come estremi

- il valore più grande fra il minimo e $Q1 - 1.5\Delta_q$
- Il valore più piccolo fra il massimo e $Q3 + 1.5\Delta_q$

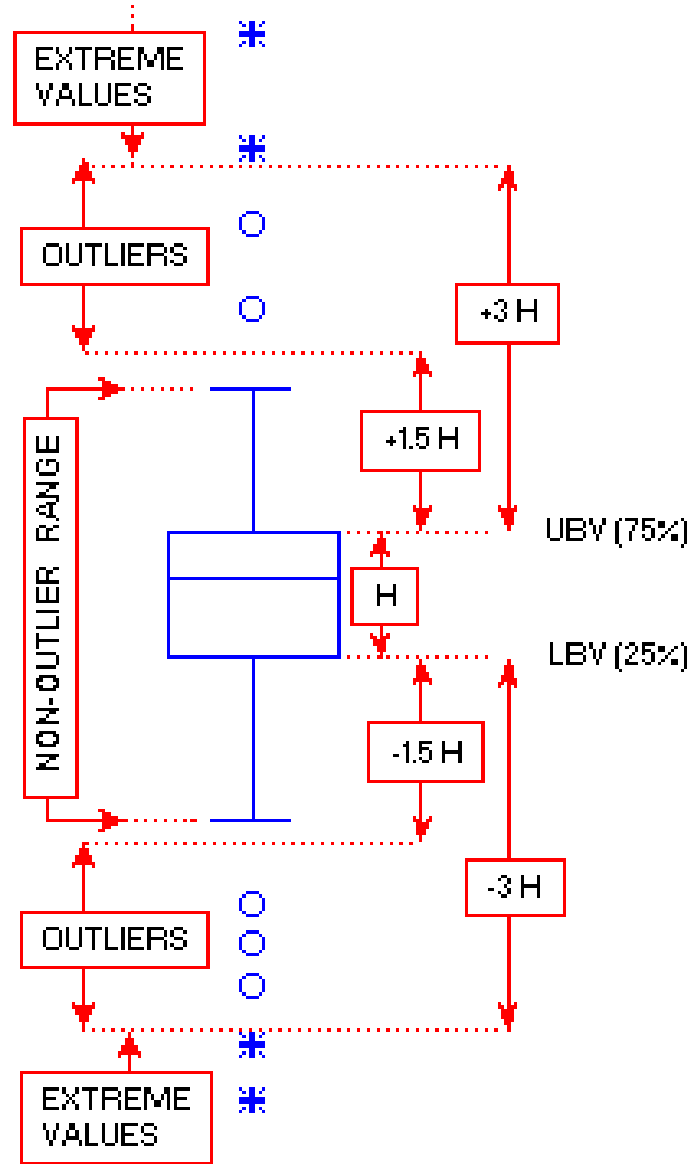


Boxplot- Esempio per il confronto tra due distribuzioni

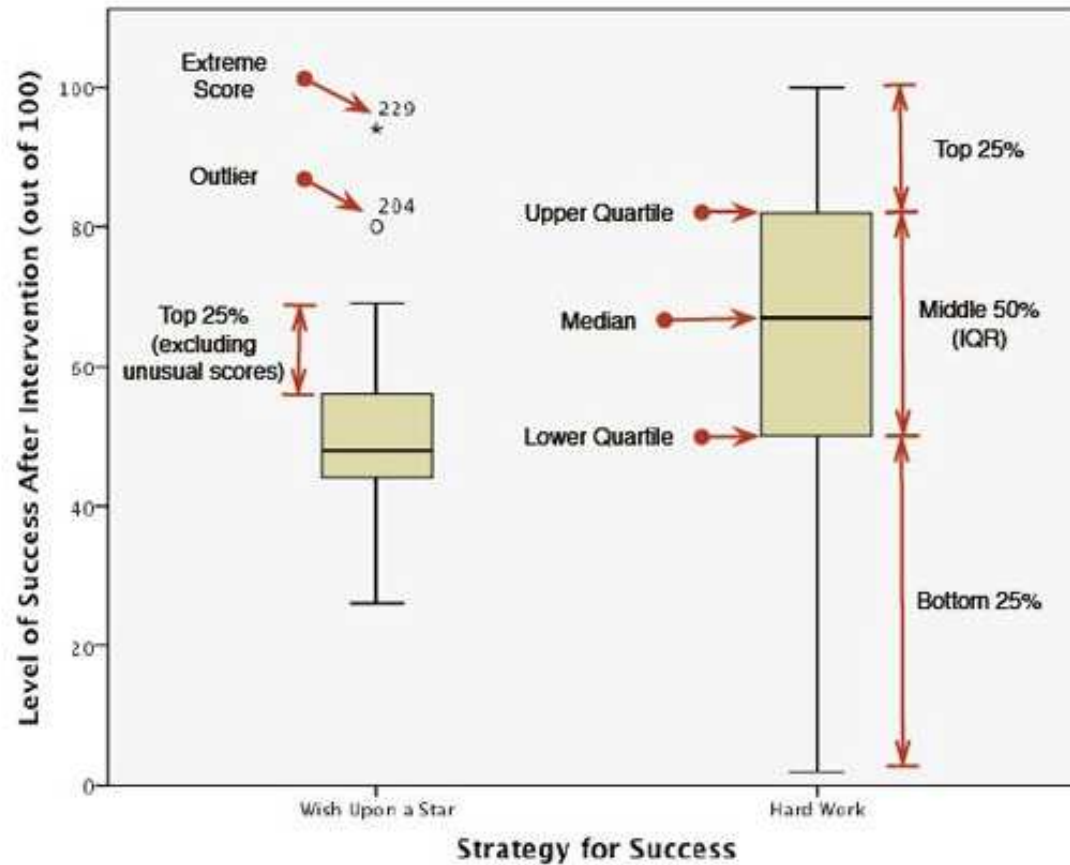


- ✓ Chi ha svolto l'esame in presenza ha conseguito un voto mediano più alto rispetto a chi ha sostenuto l'esame a distanza.
- ✓ Maggiore variabilità tra chi ha sostenuto l'esame a distanza rispetto che in presenza.
- ✓ Infine, relativamente al boxplot a destra, si può osservare che è presente un "outlier" (corrispondente ad uno studente che ha preso solo 18 all'esame) che nel grafico è indicato con un puntino.

Boxplot e valori anomali ed estremi



Boxplot of success scores 5 years after implementing a strategy of working hard or wishing upon a star



Domande (1)

- Se hai la varianza come calcoli la deviazione standard?
 - 1. dividi per la numerosità del collettivo
 - 2. elevi la varianza al quadrato
 - 3. estrai la radice quadrata della varianza
 - 4. sottrai la media
 - 5. calcoli il reciproco della varianza

E se hai la devianza, quali delle operazioni elencate sopra effettui per calcolare la deviazione standard?

Se si vuole calcolare di quanto in media i valori sono distanti dalla media aritmetica, perché non si calcola semplicemente la media aritmetica delle differenze dei valori osservati rispetto alla media?

Domande (2)

Perché si ha bisogno di una misura della variabilità della distribuzione?

Cosa indica una deviazione standard uguale a zero?

I voti riportati da uno studente in 6 esami sono 18, 25, 30, 26, 27, 30. Un altro studente agli esami ha preso 26, 25, 24, 26, 26, 29.

Quale dei due studenti ha un rendimento maggiormente costante?

Confrontare le medie e le deviazioni standard

Trova il range ed il range interquartilico dei seguenti valori:

31, 16, 39, 15, 18, 41, 24, 26, 21, 33, 9, 12

Domande (3)

Disegna il boxplot del tasso di abbandono, evidenziando anche i valori anomali ed estremi

REGIONI	%
Piemonte	6,9
Valle d'Aosta/Vallée d'Aoste	10,8
Lombardia	6,6
Trentino-Alto Adige/Südtirol	3,2
Veneto	4,3
Friuli-Venezia Giulia	4,6
Liguria	7,5
Emilia-Romagna	6,6
Toscana	7,4
Umbria	4,7
Marche	4,8
Lazio	5,8
Abruzzo	6,1
Molise	5,7
Campania	9,3
Puglia	5,3
Basilicata	5,3
Calabria	5,3
Sicilia	9,2
Sardegna	10,4
Bolzano/Bozen	2,9
Trento	3,6

* Valori %, anno di riferimento 2012

Domande (4)

Abbandoni scolastici nella scuola secondaria, 2006–12								
Numero								
	2006	2007	2008	2009	2010	2011	2012	totale
Germany	64,865	64,093	55,905	50,965	49,622	50,791	47,667	383,908
France	5,792	5,797	6,128	6,007	5,869	5,928	4,821	40,342
Italy	32,306	34,439	34,082	34,101	32,761	34,034	33,852	235,575
<i>Totale</i>	102,963	104,329	96,115	91,073	88,252	90,753	86,340	659,825

- Confronta la variabilità degli abbandoni scolastici nella scuola secondaria nei tre Stati