Statistica e Valutazione della Formazione

Slides

A.A. 2025-2026

Docente: ANNA LINA SARRA

Modulo 2 : I modelli di risposta all’item

Struttura della lezione

1. **Quadro teorico**
   Modello funzionalista e valutazione quantitativa
2. **Fondamenti della misurazione**
   Tratti latenti, indicatori, item analysis
3. **Fondamenti TCT**
   Equazione fondamentale, errori, ipotesi
4. **Item Analysis**
   Difficoltà, discriminatività, punto-biseriale
5. **Attendibilità**
   Split-half, test-retest, alpha di Cronbach
6. **Validità e Guessing**
   Validità del test e correzione per guessing
7. **Caso di Studio**
   Analisi di un questionario reale

La valutazione quantitativa degli apprendimenti

La valutazione quantitativa degli apprendimenti trova il suo fondamento nell'approccio teorico **del modello funzionalista**. Il sistema di riferimento orienta la scelta dei criteri, dei soggetti, degli oggetti, degli strumenti di misura.

Cosa include il modello teorico di riferimento?

Filosofia di pensiero

Un approccio specifico alla valutazione

Substrato teorico

Principi e concetti di riferimento

Strumenti empirici

Metodi per affrontare la valutazione concretamente

Il modello funzionalista

Il modello funzionalista — nato dall'esigenza di abolire azioni didattiche basate sulla soggettività — si caratterizza per un approccio quantitativo alla valutazione. Il focus è centrato sulla verifica della conformità tra obiettivi programmati e risultati conseguiti.

Le tre garanzie della valutazione quantitativa:

ATTENDIBILITÀ	OGGETTIVITÀ	VALIDITÀ
Costanza della rilevazione: il risultato deve rimanere invariato al variare del somministratore, del tempo e del contesto.	Controllo di tutti gli aspetti soggettivi. Il peso di ogni item è stabilito a priori (risposta esatta, errata, omessa).	La misura fornisce informazioni accurate per assumere decisioni efficaci ed efficienti riguardo alle qualità indagate.

L'affidabilità di uno strumento di misura

Definizione:

Quando si ripete più volte una misura e si ottiene lo stesso risultato, la misura è detta **affidabile, attendibile o fedele**. L'affidabilità si riferisce alla costanza della misura di una data prestazione.

Perché è importante?

- Se uno **strumento è attendibile**, le variazioni nei dati non dipendono da imperfezioni dello strumento.
- Le variazioni osservate riflettono il mutare del fenomeno misurato (errore casuale), non difetti dello strumento.
- Nelle ricerche in campo educativo, la ripetizione della stessa rilevazione raramente dà un risultato assolutamente identico.

I tratti latenti (Latent Trait)

LORD et al. (1968)

Costrutti non direttamente osservabili ma desumibili da un insieme di osservazioni (prove attitudinali, questionari, prove fisiche, mentali…) che consentono di effettuare una stima dei fenomeni stessi.

Esempi di tratti latenti:

Intelligenza
Capacità cognitiva generale, non misurabile direttamente

Abilità matematica
Desunta da serie di prove numeriche e logiche

Ansia
Inferita da indicatori comportamentali e fisiologici

Competenza di lettura
Stimata tramite item di comprensione del testo

Operazionalizzazione dei tratti latenti

I tratti latenti sono dimensioni teoriche (costrutti) non osservabili direttamente, ma misurabili mediante l'individuazione di indicatori osservabili. L'indicatore è una variabile osservabile che si ipotizza cogliere il costrutto — o parte di esso.

⚠ *La scelta degli indicatori non è ovvia: dipende dalle teorie dei ricercatori e dagli strumenti adottati.*

INDICATORI RIFLETTIVI

Gli indicatori riflettono il costrutto: ne rappresentano la manifestazione empirica come conseguenza della presenza del costrutto.

Esempio: l'autostima si riflette nell'accezione del sé, nella relazione con l'altro, nell'interazione con il contesto.

INDICATORI FORMATIVI

Gli indicatori formano, contribuiscono o causano il costrutto.

Esempio: la competenza di lettura è determinata dal comprendere parole, individuare informazioni, cogliere relazioni logiche…

Item Analysis

In questo modulo ci occupiamo di tecniche di analisi statistica legate alla **valutazione**, tema centrale nella didattica. Modelli di progettazione che vedono la valutazione strettamente connessa agli obiettivi di apprendimento (Biggs & Tang, 2011; Wiggins & McTighe, 2005) ne danno la giustificazione: l'efficacia di un percorso didattico può essere provata quando, attraverso adeguati strumenti di valutazione, siamo in grado di dire se gli studenti hanno raggiunto i comportamenti attesi espressi come obiettivi educativi.

Valutare permette di analizzare non solo gli apprendimenti degli studenti, ma anche l'efficacia del percorso formativo. Gli strumenti di valutazione andrebbero disegnati in base al tipo di apprendimenti da misurare: ***la misurazione è parte del processo di valutazione.***

Item Analysis — Introduzione

Definizione:

L'item analysis è l'insieme delle tecniche che permettono di ricavare informazioni sull'affidabilità di una prova nel suo complesso e sul funzionamento di ciascuna domanda. Le tecniche si fondano sull'assunto che le singole domande e i soggetti debbano avere un comportamento coerente.

A cosa serve?

Qualità degli item

Misurare quanto le domande siano appropriate e quanto bene misurino le abilità degli intervistati.

Item bank

Riutilizzare gli item in test diversi con conoscenza preliminare delle loro prestazioni statistiche.

Prova pilota

Richiede un tryout su un campione sufficientemente ampio con caratteristiche simili alla popolazione target.

Perché usare l'Item Analysis?

1. Stimare le abilità degli studenti come tratti latenti non direttamente misurabili. Le risposte fornite nelle prove di valutazione sono solo una manifestazione parziale del processo di apprendimento (De Luca & Lucisano, 2011).

2. Esaminare le prove di valutazione per verificare se le domande sono scritte correttamente, se sono troppo semplici o difficili, se aiutano a distinguere gli studenti preparati da quelli meno preparati, se tutte le domande fanno riferimento a uno stesso tratto latente.

3. Trovare un modello predittivo che indichi con quale probabilità gli studenti possano rispondere correttamente a una domanda, in base alle loro abilità e alle caratteristiche delle domande stesse.

Misurazione ed errore casuale

Come altri ambiti scientifici – la fisica, la statistica – ci hanno insegnato nell'ultimo secolo, la **misurazione è sempre affetta da un errore casuale ε**. Ogni grandezza oggetto di misurazione è quindi una variabile casuale, con tutte le conseguenze e le proprietà che ciò comporta.

Gli strumenti quantitativi di valutazione – in particolare i **questionari a risposta chiusa** – sono quelli a cui applichiamo le tecniche di **item analysis**, le cui basi sono da rintracciare nell'analisi psicometrica dei quesiti. Esse sono più veloci da somministrare, vantaggiose su grandi numeri di studenti, permettono di indagare molti argomenti e sono ritenute più oggettive.

Su quali dati si lavora?

Dopo aver prodotto un questionario, è buona regola proporlo a un campione di studenti per verificarne l'efficacia. Nella pratica non sempre si riesce a realizzare una fase di ***tryout***: la prima somministrazione del questionario diventa quindi la prova sperimentale.

In queste fasi iniziali di utilizzo di una scala o di un questionario, l'item analysis può essere molto utile. Più in generale, ogni volta che vengono raccolte risposte a una serie di item, le tecniche di item analysis possono essere usate per:

- stimare l'abilità degli studenti
- verificare le caratteristiche delle domande
- migliorare la qualità complessiva del test

I due approcci fondamentali

Classical Test Theory (CTT)

Modello classico (tradizionale): si identificano alcuni indicatori per valutare gli item.

L'abilità dello studente corrisponde alla **somma delle risposte corrette** date al questionario.

Item Response Theory (IRT)

Modello proposto da **Rasch negli anni Sessanta**. Centrale il calcolo delle probabilità e la ricerca di un modello che stimi l'abilità dello studente tenendo conto anche delle caratteristiche delle singole domande.

Modulo 4.1: La teoria classica dei test (TCT)

$$X = T + E$$

Observed Score = True Score + Error

Nascita e sviluppo della TCT

TIMELINE

Fine '800

Alfred Binet e collaboratori sviluppano i primi test psicologici (1894). Nasce l'interesse per la misurazione dei costrutti psico-sociali.

Inizi '900

Spearman formalizza l'equazione fondamentale della TCT (1904): X = T + E

Anni '30

Impiego su vasta scala. La TCT diventa lo standard per la valutazione psicometrica.

Oggi

Ancora ampiamente usata, affiancata dalla IRT (Item Response Theory) per superarne i limiti.

DUPLICE INTERESSE

La TCT nasce con due obiettivi fondamentali che ancora oggi guidano l'analisi dei test:

Misurazione del costrutto

Quantificare caratteristiche latenti (intelligenza, abilità, competenze) non direttamente osservabili.

Validazione dello strumento

Verificare che il test misuri effettivamente ciò che intende misurare, con precisione e coerenza.

L'equazione fondamentale della TCT

Il punteggio empirico X ottenuto da una persona in un test è formato da due componenti:

$$X = T + E$$

X

Punteggio Osservato

Il punteggio effettivo che la persona ottiene nella somministrazione del test.

T

Punteggio Vero

Il punteggio che otterrebbe in assenza di errori; non direttamente osservabile.

E

Errore (non sistematico)

La componente casuale che non possiamo controllare; la sua media è zero.

Il punteggio vero (True Score)

$T = E(X)$

Il punteggio vero è l'aspettativa matematica (valore atteso) del punteggio empirico. Corrisponde a ciò che si otterrebbe in media se si somministrasse infinite volte lo stesso test.

Proprietà del punteggio vero:

Distribuzione normale

I punteggi osservati ripetuti si distribuiscono normalmente attorno al valore vero.

Media = punteggio vero

La media di infinite somministrazioni dello stesso test coincide con T.

Errori positivi ↔ negativi

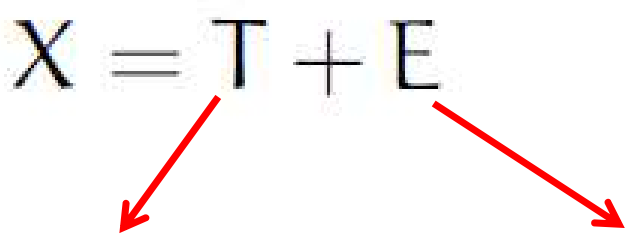
Gli errori positivi compensano quelli negativi: la somma degli errori tende a 0.

Non direttamente osservabile

T è un costrutto teorico: possiamo solo stimarlo tramite i punteggi empirici.

La teoria classica dei test (TCT)

Questo modello consiste nel **sostenere che il punteggio che una persona ottiene nel test, che denominiamo punteggio empirico,** e che solitamente è indicato dalla lettera X, è formato da due componenti:



- **IL PUNTEGGIO VERO**
- **L'ERRORE NON SISTEMATICO**

L'errore può essere dovuto a molte cause che non possiamo controllare. Per questo la TCT **si preoccupa di determinare con precisione l'errore di misurazione.**

Errori di misurazione: casuali e sistematici

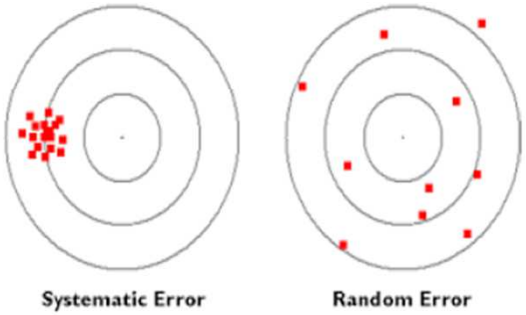
L'errore è una componente integrante e inevitabile di qualsiasi processo di misurazione — anche nelle scienze fisiche.

ERRORI CASUALI

- Variano in maniera casuale tra misurazioni diverse
- Non predicibili in anticipo
- La loro somma tende a 0 su infinite misurazioni
- Esempi: distrazione, benessere fisico del candidato, disturbi ambientali

ERRORI SISTEMATICI (BIAS)

- Si presentano in maniera costante e predicibile
- Influenzano tutte le misure nello stesso modo
- Esempi: errori nella formulazione degli item, ambiguità dei distrattori
- Meno rilevanti se si studiano differenze individuali



L'errore di misurazione casuale e non sistematico

L' errore di misurazione casuale e non sistematico non è una proprietà della caratteristica misurata ma è conseguente alla misurazione effettuata.
Esso è, quindi, inversamente proporzionale all'affidabilità:

MAGGIORE è LA COMPONENTE DI ERRORE

MINORE SARA' L'AFFIDABILITA' DELLA MISURAZIONE

Le tre ipotesi fondamentali della TCT

La TCT si basa su ipotesi molto forti che ne delimitano il campo di applicazione:

IPOTESI 1

Il punteggio vero è il valore atteso

$$T = E(X)$$

Il punteggio vero T è l'aspettativa matematica del punteggio empirico X. Se somministrassimo il test infinite volte allo stesso soggetto, la media convergerebbe a T.

IPOTESI 2

Errore e punteggio vero sono indipendenti

$$Cov(T, E) = 0$$

Il valore del punteggio vero T è indipendente dall'errore di misurazione: non c'è relazione tra l'entità del punteggio vero e la quantità di errore che lo accompagna.

IPOTESI 3

Gli errori di test diversi sono incorrelati

$$r(E_j, E_k) = 0$$

Gli errori di misura in un test concreto non sono legati agli errori di un altro test differente. Errori commessi in un'occasione non influenzano quelli di un'altra.

Le tre ipotesi della TCT

I. **Il punteggio vero (T) è l'aspettativa matematica del punteggio empirico**:
   $T = E(X)$

II. ***Il valore del punteggio vero (T)* è indipendente dall'errore di misurazione.**

II. ***Gli errori* di misura in un test concreto non sono legati agli errori di misura di un altro test diverso.**

Gli errori commessi in un'occasione non influenzerebbero quelli commessi in un'altra occasione.

In sintesi…..

LA TEORIA CLASSICA DEI TEST ASSUME CHE

1. **L'ERRORE È NORMALMENTE DISTRIBUITO**
2. **INCORRELATO CON PUNTEGGIO VERO**
3. **HA MEDIA ZERO**

.

Attendibilità

Attendibilità, affidabilità e fedeltà

sono tre sinonimi utilizzati per riferirsi, in ambito statistico, al grado di accuratezza e precisione di una procedura di misurazione.

•Si dice allora che un test è affidabile, quando si può affermare che i punteggi ottenuti da un gruppo di soggetti allo stesso

"*sono coerenti, stabili nel tempo e costanti dopo molte somministrazioni e in assenza di cambiamenti evidenti quali variazioni psicologiche e fisiche degli individui che si sottopongono al test, o anche all'ambiente in cui questo ha luogo*"

(Pedrabissi, Santinello, 1997)

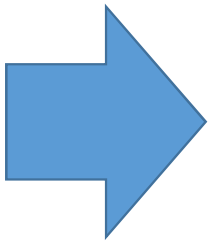
Attendibilità

Abbiamo detto che ogni misurazione è affetta da errori.

Gli errori sistematici posso essere identificati perchè predicibili.

Gli errori casuali non sono predicibili.

.



Nella equazione X = T + E come possiamo separare T da E in modo da capire quanta informazione vera (T) è contenuta in X?

L'attendibilità misura il grado di coerenza e di stabilità di un test o anche il grado di precisione con cui una scala misura un costrutto.

La varianza

La varianza sarà dovuta a componenti vere e a componenti d'errore, cioè dipenderà dal diverso grado di possesso di una certa caratteristica da parte dei soggetti e da fattori occasionali estranei al test e non controllabili.

Possiamo allora affermare che la varianza totale della distribuzione dei punteggi equivale alla somma della varianza delle componenti vere e di quella delle componenti d'errore, ossia

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2$$

σ_X^2
corrisponde alla varianza dei punteggi osservati;

σ_T^2
corrisponde alla varianza del punteggio vero

σ_E^2
corrisponde alla varianza del componente d'errore.

L'attendibilità

L'attendibilità di un test può essere definita allora come la porzione di varianza vera rispetto alla varianza totale di una distribuzione di punteggi, cioè il **rapporto tra varianza dei punteggi veri e varianza dei punteggi osservati** può essere definito come una valutazione del livello di affidabilità,

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2}$$

È IL COEFFICIENTE DI ATTENDIBILITÀ DEL TEST

L'analisi classica degli item

Comprende:

- Frequenze di risposta
- Indice di distrattività
- Indice di difficoltà
- Indice di discriminatività
- Correlazione item-totale
- Alfa di Cronbach

Frequenza di risposta e indice di distrattività

Frequenza di risposta

Numero di occorrenze di ogni modalità di risposta.

Può essere espressa come:
- Frequenza assoluta
- Frequenza relativa (%)
- Frequenza cumulata

Regola: se N < 100 si usano le frequenze assolute anziché le percentuali.

Indice di distrattività

Misura la capacità dei singoli distrattori di far deviare dalla risposta corretta.

Situazione ideale (4 opzioni):

- Tutti i distrattori hanno uguale frequenza di scelta
- La risposta corretta ha frequenza più alta
- Nessun distrattore ha frequenza prossima a zero (altrimenti è palesemente errato)
- Nessun distrattore attrae troppo (potrebbe nascondere ambiguità)

Frequenze di risposta

Quante volte ricorre una risposta in un quesito, per tutte le risposte possibili, per tutti i quesiti.

Possono essere

- ✓ASSOLUTE
- ✓RELATIVE
- ✓CUMULATE
- ✓PERCENTUALI

Quesito 1	Fq assoluta	Fq relativa	Fq cumulata	Fq %
A	6	0,188	6	19%
B	7	0,219	13	22%
C	4	0,125	17	12%
D (corretta)	15	0,469	32	47%
totale	32	1	-	100%

Indice di distrattività

L'indice di distrattività individua la capacità dei singoli distrattori di far deviare dalla risposta corretta.

La distrattività del quesito si misura tenendo conto della distribuzione delle risposte errate sui distrattori.

Nell'ipotesi di quesiti con possibilità di scelta tra quattro opzioni (una risposta esatta e tre distrattori), **la situazione ideale è quella per cui tutti gli item hanno la stessa capacità di attrarre risposte e quindi la stessa frequenza**.
Se uno dei distrattori è palesemente errato e nessuno lo sceglie, le opzioni vere per il rispondente rimangono ridotte a tre e il ruolo della casualità nella scelta esatta sarebbe più forte, con un abbassamento del grado di validità del quesito.

Indice di distrattività

L'indice di distrattività si calcola, sul complesso degli errori, quanti, per ciascun *item*, si riferiscono a ciascun distrattore.

Indice di distrattività = D/Et

dove:

D = numero alunni che hanno scelto i diversi distrattori

E = errori totali commessi

Ad esempio si supponga che un test sia svolto da 80 studenti e il quesito a risposta multipla n. 1 abbia avuto il seguente esito:

- risposta esatta a: n. 50
- distrattore b: n. 20
- distrattore c: n. 8
- distrattore d: n. 2

Totale risposte fornite ai distrattori = n. 30

Db = 20/30 = 0,67 → efficace

Dc= 8/30 = 0,26 → abbastanza efficace

Dd = 2/30 = 0,07 → inefficace

Indice di distrattività

I. ▪Si adopera per valutare il funzionamento dei quesiti a scelta multipla
II. ▪Misura la capacità di ogni singolo distrattore di far deviare dalla risposta corretta a ciascun quesito
III. ▪È la percentuale di studenti che sceglie ciascun distrattore.
IV. ▪..senza includere le risposte lasciate in bianco, le doppie risposte o quelle incomprensibili

▪Se N<100 NON si usa la frequenza percentuale ma direttamente le frequenze assolute

Come si interpreta la distrattività

I. Serve ad individuare quei distrattori poco o per nulla attrattivi
II. Teoricamente, la risposta corretta dovrebbe avere sempre la frequenza assoluta più elevata e i distrattori dovrebbero dividere equamente i restanti rispondenti
III. Se un distrattore non viene scelto vuol dire che non è plausibile (molto spesso per ragioni indipendenti dai contenuti) e rende per questo più facile il quesito.

Dicotomizzare le risposte

Per calcolare gli indici di difficoltà/facilità e di discriminatività è necessario trasformare le risposte in una variabile dicotomica:

Risposta ESATTA

Risposta ERRATA (o omessa)

Attenzione: la dicotomizzazione semplifica il calcolo ma comporta la perdita di informazioni sui singoli distrattori.

Dicotomizzare le risposte

Per calcolare gli' indici di DIFFICOLTÀ/FACILITÀ e di DISCRIMINATIVITÀ occorre DICOTOMIZZARE le risposte.

1=risposta esatta
0= risposta errata

Studente	D1	D2	D3	D4	D5	Studente	D1	D2	D3	D4	D5
A1	A	C	C	SI	V	**A1**	1	1	1	1	0
A2	B	B	C	SI	F	**A2**	0	0	1	1	1
A3	B	C	D	NO	V	**A3**	0	1	0	0	0
A4	B	A	C	NO	F	**A4**	0	0	1	0	1
A5	C	C	C	SI	F	**A5**	0	1	1	1	1
A6	B	C	D	SI	F	**A6**	0	1	0	1	1
A7	D	D	A	SI	V	**A7**	0	0	0	1	0
A8	C	A	B	SI	F	**A8**	0	0	0	1	1
A9	D	C	C	NO	F	**A9**	0	1	1	0	1
A10	A	A	C	NO	V	**A10**	1	0	1	0	0
CHIAVE	**A**	**C**	**C**	**SI**	**F**						

Difficoltà e discriminatività

Indici di base nell'analisi classica dell'andamento dei quesiti

INDICE DI DIFFICOLTA'

$$Df_i = \frac{nE_i}{N}$$

INDICE DI FACILITÀ

$$Fc_i = \frac{nC_i}{N}$$

$$Fc_i = 1 - Df_i$$

Esempio

oscillazione tra 0 e 1

	Quesito 1	Quesito 2	Quesito 3	Quesito 4
Corrette	20	22	24	4
Errate	**5**	**3**	1	**21**
Totale	25	25	25	25
Difficoltà	0,2	0,12	0,04	0,84
Difficoltà %	20%	12%	4%	84%
Item facility	*80%*	*88%*	*96%*	*16%*

Quesito più difficile

Quesito più facile

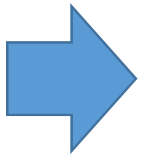
Come si interpreta l'indice di difficoltà?

Non ci sono regole sempre valide, tutto dipende dagli scopi della prova.

Solitamente se

✓se <0.3 è troppo FACILE

✓se>0.7 è troppo DIFFICILE



All'interno della stessa prova ci devono essere sia quesiti facili che quesiti più difficili.

Indice di difficoltà

Secondo la TCT, la scelta che viene usualmente fatta è quella di una dispersione moderata e simmetrica del livello di difficoltà attorno ad un valore leggermente superiore al valore che sta a metà tra il livello 1/n, dove n è il numero delle alternative di risposta all'item, e il punteggio pieno.

Nel caso di un questionario a quattro alternative di risposta, il livello del caso per ogni item è pari a 1/4 =0,25. Il livello ottimale di difficoltà media sarà, quindi, (0,25+1,00)/2=0,62.

Nel caso di item con sole due alternative di risposta (es.Si/ No,Vero/Falso,ecc.) il livello ottimale di difficoltà media è (0,50+1,00)/2=0,75.

Aspetti critici del livello di difficoltà nella TCT

Funzionamento ideale della difficoltà

(Losito pag.107)

	Q1	Q2	Q3	Q4	Q5	%
Studente 1	1	1	1	1	1	100
Studente 2	0	1	1	1	1	80
Studente 3	0	0	1	1	1	60
Studente 4	0	0	0	1	1	40
Studente 5	0	0	0	0	1	20
Df	0,80	0,60	0,40	0,20	0	

Aspetti critici del livello di difficoltà nella TCT

Se **aggiungiamo uno studente con un andamento di risposta anomalo**, la situazione non è più facilmente interpretabile.

Gli studenti 4 e 6 hanno lo stesso punteggio % ma hanno risposto a quesiti di difficoltà diversa.

	Q1	Q2	Q3	Q4	Q5	%
Studente 1	1	1	1	1	1	100
Studente 2	0	1	1	1	1	80
Studente 3	0	0	1	1	1	60
Studente 4	0	0	0	1	1	40
Studente 5	0	0	0	0	1	20
Studente 6	1	1	0	0	0	40
Df	0,66	0,50	0,50	0,33	0,17	

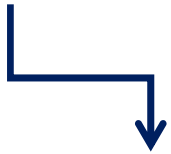
Aspetti critici del livello di difficoltà nella TCT

	Q1	Q2	Q3	Q4	Q5	Q6	%
Studente 1	1	1	1	1	1	0	83
Studente 2	0	1	1	1	1	0	67
Studente 3	0	0	1	1	1	0	50
Studente 4	0	0	0	1	1	0	33
Studente 5	0	0	0	0	1	1	33
Df	0,80	0,60	0,40	0,20	0	0,80	

Indice di discriminatività

Nella composizione di un test o di una prova oggettiva è importante analizzare, tra le altre cose, la capacità che hanno i singoli item di differenziare i soggetti "preparati" da quelli meno preparati.

DOMANDA



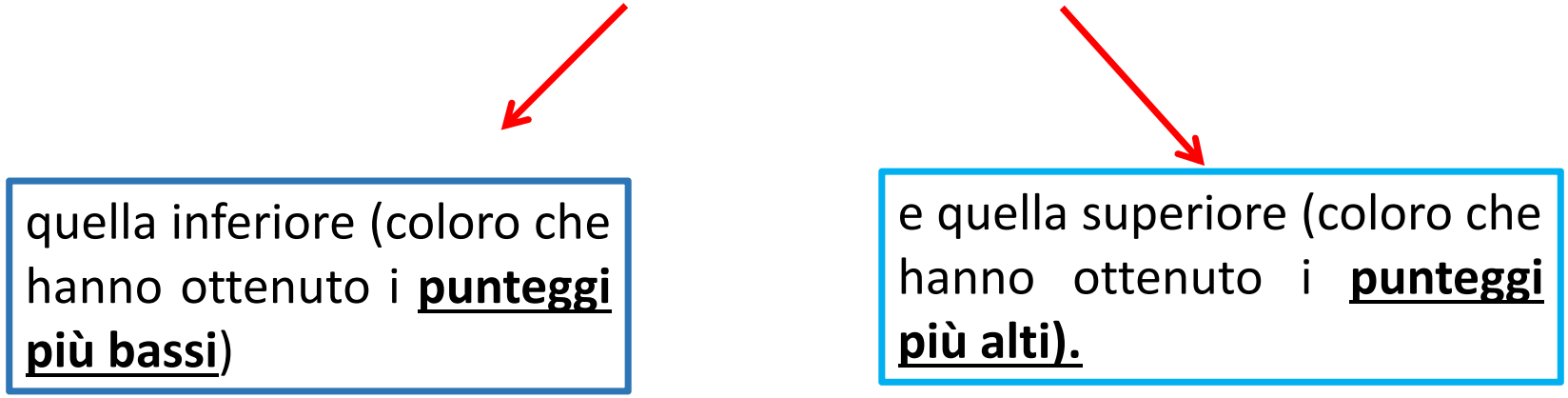
In che misura una domanda posta in un questionario può separare i soggetti che possiedono le abilità che si vogliono verificare con una prova da coloro che non le possiedono?

Uno degli strumenti più utilizzati in tal senso è l'**indice di discriminatività.**

Indice di discriminatività

Il calcolo di questo indice assume che, le risposte giuste date ad una domanda dal gruppo di soggetti più bravi dovrebbero essere più numerose delle risposte esatte date allo stesso quesito dal gruppo che ha conseguito i punteggi più bassi.

È dunque necessario paragonare le due fasce estreme dei risultati



Indice di discriminatività

L'indice di discriminatività/selettività si calcola nel modo seguente

Indice di discriminatività = (Es – Ei)/N

dove:

Es = numero delle risposte esatte registrate nell'estremo superiore (numero di allievi della fascia alta che hanno risposto correttamente all'item)

Ei = numero delle risposte esatte registrate nell'estremo inferiore (numero degli allievi della fascia bassa che hanno risposto correttamente all'item)

n = numero dei soggetti che costituiscono ciascun gruppo estremo

L'indice varia da -1 a +1.
Il valore zero indica che l'item non è discriminativo, quando l'estremo superiore risponde meglio il segno è positivo.

Un buon indice di selettività è compreso tra 0,30 e 0,60:
sotto 0,3 → item non è discriminante
sopra 0,6 → item troppo selettivo

Interpretazione dell'indice di discriminatività

> 0.35 **Item OTTIMO**	Discrimina bene tra soggetti con abilità diverse. Da tenere nel test.
0.20 – 0.34 **Item ACCETTABILE**	Discreto potere discriminante. Valutare il contesto.
0.00 – 0.19 **Item DEBOLE**	Scarso potere discriminante. Rivedere la formulazione.
< 0 **Item PROBLEMATICO**	Discriminatività NEGATIVA: i meno bravi rispondono meglio. Da eliminare o rivedere completamente.

Indice di discriminatività

indica quanto l'item riesce a discriminare tra rispondenti con alte abilità (gruppo 1) rispetto a quelli con basse abilità (gruppo 2).

Nello specifico rappresenta la correlazione tra la probabilità di scegliere una data opzione e l'abilità complessiva del rispondente.

In un test a scelta multipla,
tale legame deve essere negativo per le opzioni di risposta non corrette e positivo solo per quella esatta.
Una domanda è ben formulata se, in media, coloro che rispondono correttamente a quella domanda lo fanno anche a buona parte delle altre che hanno lo stesso livello di difficoltà.

Si considerano funzionanti gli item con un valore > 0,35

Indice di discriminatività

L'indice di discriminatività assume valori compresi tra -1 e+1. L'indice discrimina correttamente se assume un valore compreso tra 0,30 e 0,60.

Se per un item l'indice assume valore +1, significa che al quesito hanno risposto correttamente solo studenti di un livello di apprendimento elevato.

Se per un item l'indice assume valori negativi (discriminatività negativa), occorre riflettere sulla formulazione dell'item, in quanto se l'item è troppo fuorviante puo' indurre a risposte casuali.

Esempio

Ad esempio si supponga che il test sia stato svolto da 25 allievi;
si individueranno 3 fasce, la prima e la terza di 8 allievi, la seconda di 9 allievi: la prima degli alunni migliori, la terza dei peggiori, la seconda dei mediani.
Si supponga che all'*item* n. 1 abbiano risposto correttamente 6 allievi della fascia alta e 2 allievi della fascia bassa
S1 = (6-2)/8 = 0,50
che all'*item* n. 2, rispettivamente 7 e 5 allievi
S2 = (7-5)/8 = 0,25
Nel caso da noi ipotizzato l'item n. 2 non è discriminante, il n.1 sì.

La discriminatività dei distrattori

Di ogni opzione dovrebbe essere calcolata anche la **discriminatività,** che nel caso **dei distrattori dovrebbe essere negativa.**

Dovrebbe cioè accadere che coloro che hanno punteggio basso nel complesso della prova hanno più facilità di scegliere le risposte errate rispetto a coloro che hanno punteggio alto.

Quando un distrattore ha una discriminatività positiva abbiamo una prova che il tranello insito nel distrattore ha danneggiato i più performanti, e che quindi ha reso meno affidabile la misura operata dal quesito.

Quali quesiti rivedere e perché?

	dom. 1	dom. 2	dom. 3	dom. 4	dom. 5	dom. 6	dom. 7
A	13	10	53	7	9	2	4
B	13	13	5	58	27	4	28
C	45	6	17	3	38	62	9
D	6	48	2	0	2	9	34
Missing	0	0	0	9	1	0	2

	dom. 1	dom. 2	dom. 3	dom. 4	dom. 5	dom. 6	dom. 7
corrette	45	48	53	7	38	62	28
errate	32	29	24	70	39	15	49
Difficoltà	42%	38%	31%	91%	51%	19%	64%
Discrimin.	0,5	0,4	0,4	-0,1	0,7	0,7	0,4

Quali quesiti rivedere e perché?

	Dom. 1	Dom. 2	Dom. 3
A*	24	6	16
B	2	32	11
C	14	5	13
D	10	7	10
diff	0,52	0,88	0,68
discr	0,43	-0,21	0,87

Esempio:
prove classi prime matematica

Partecipazione al campionamento: 3 classi prime

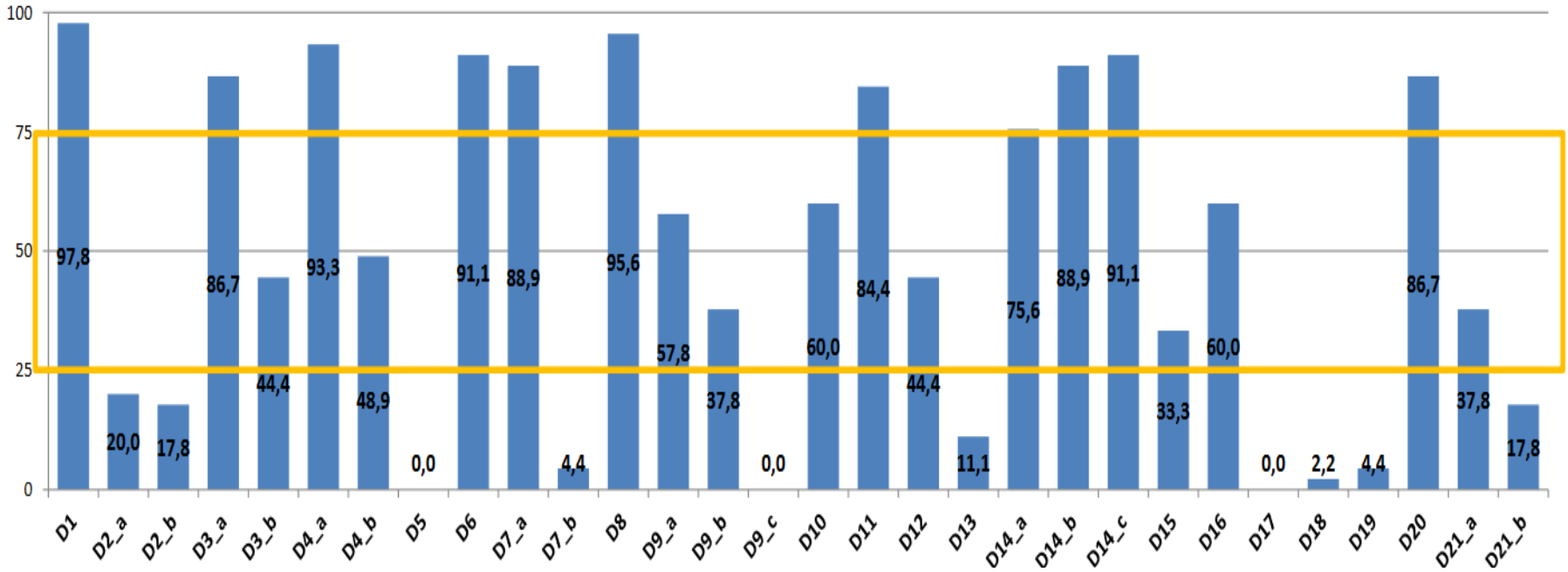
- TOTALE = 45 studenti delle classi prime di 3 classi di 3 scuole diverse.

30 quesiti:

- 15 quesiti a risposta multipla (4 opzioni di risposta A, B, C, D)
- 14 quesiti a risposta libera "breve" → ricodifica omogenea 1/0 dove 1 se corretto e 0 se errato
- 1 quesito complesso D5, composto da 4 item

Statistiche Descrittive		
N studenti partecipanti = 45	# esatte	% esatte
MEDIA	14,8	49,4
MEDIANA	15	50,0
MIN	6	20,0
MAX	18	60,0
RANGE (Max - Min)	12	40,0

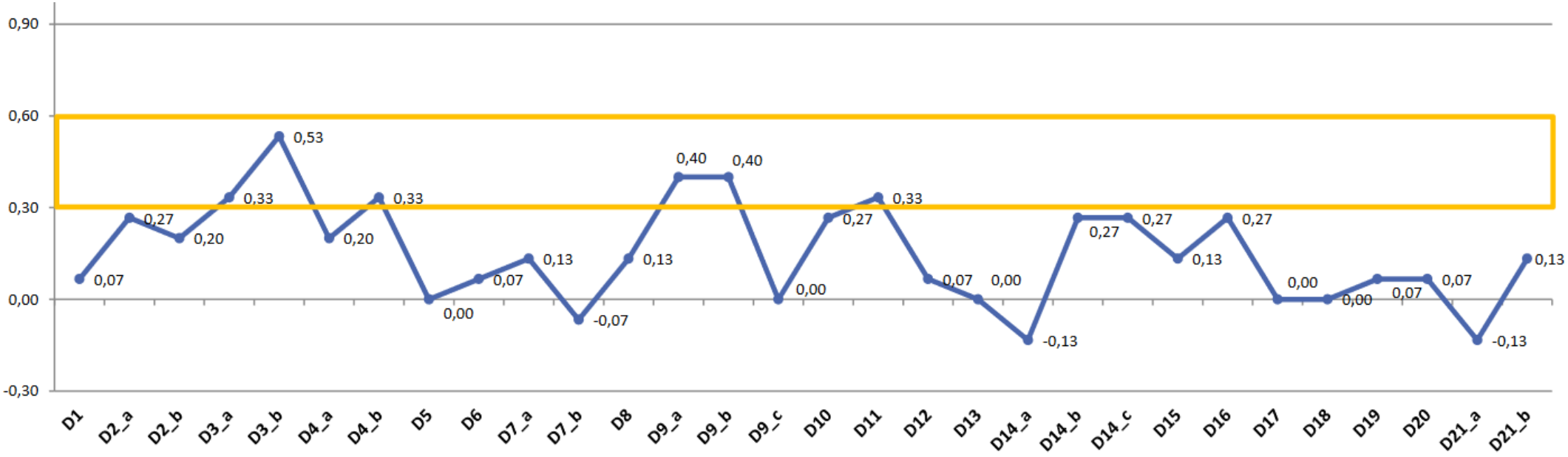
Indice di difficoltà/facilità Studenti classi prime –matematica (N=45)



L'indice di difficoltà /facilità è accettabile se la % di risposte corrette è compresa tra 25% e 75% (oltre 75% l'item è molto facile (FF), sotto il 25% è molto difficile (DD).

- 9 quesiti con indice di difficoltà/facilità accettabile (6 D, 4 F)
- 10 quesiti troppo facili (FF)
- 10 quesiti troppo difficili (DD)

Indice di discriminatività/selettività Studenti classi prime –matematica (N=45)



L'indice di selettività deve essere compreso tra 0,30 e 0,60
sotto 0,3 → item non è discriminante
sopra 0,6 → item molto selettivo

- 6 quesiti con indice di selettività buono
- Gli altri non discriminanti (di questi 5 sono vicini a 0,30)

Il coefficiente di correlazione punto-biseriale

Il coefficiente di correlazione punto-biseriale (rpb) è un coefficiente di correlazione utilizzato quando una variabile è dicotomica. La variabile può essere effettivamente dicotomica, come il genere, o può essere dicotomizzata artificialmente. Il coefficiente di correlazione punto-biseriale è che la correlazione di Pearson tra l'item e il punteggio totale al test. La sua espressione è data nella seguente

$$r_{pb} = \frac{M_1 - M_0}{\sigma_n}\sqrt{\frac{n_1 n_0}{n^2}}$$

rappresenta la deviazione standard dell'intera popolazione

M1 è il **valore medio della variabile continua** X per tutti i dati del **gruppo 1**, rappresenta quindi la media dei punteggi dei soggetti che hanno risposto in maniera corretta all'item;

M0 è il **valore medio delle X per tutti i dati del gruppo 2**; n1 è la numerosità del gruppo 1; n0 è la numerosità del gruppo 2 ed n è la dimensione totale del campione.

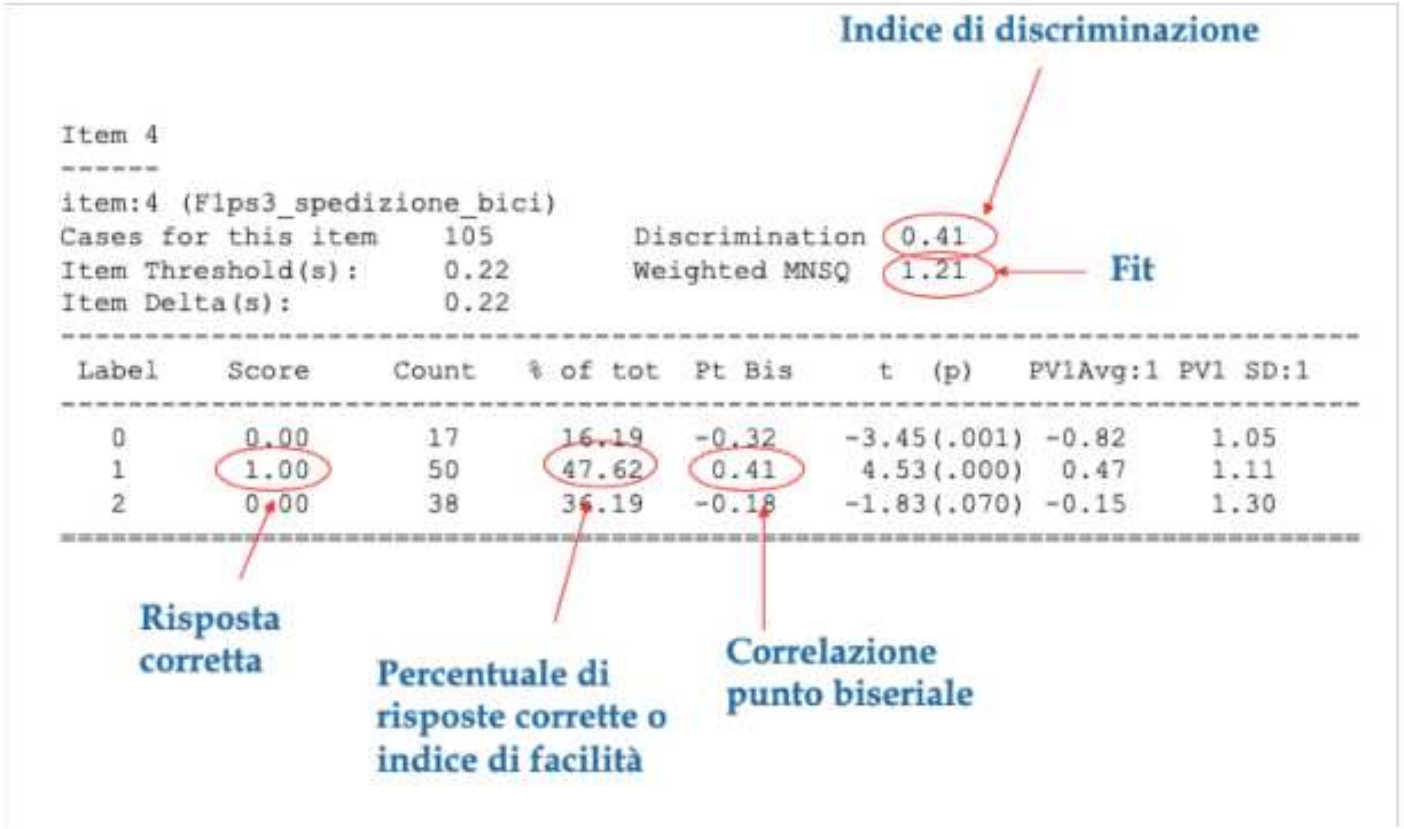
Il coefficiente di correlazione punto-biseriale

Il coefficiente di correlazione punto-biseriale stabilisce una correlazione tra due variabili

una dicotomica che consiste nell'aver risposto in maniera esatta o errata ad uno specifico item

l'altra continua che consiste nel punteggio complessivo ottenuto da coloro che hanno risposto esattamente all'item

Il coefficiente di correlazione punto-biseriale



Il coefficiente di correlazione punto-biseriale

Il coefficiente di correlazione punto-biseriale stabilisce una correlazione tra tutte le risposte date ad un quesito e tutti i punteggi grezzi degli allievi

Differisce dalla discriminatività che è un semplice confronto tra gli estremi

Varia da -1 a +1 e si interpreta

- ✓<0= DA SCARTARE
- ✓<0.19= RIVEDERE IL QUESITO
- ✓0.20-0.29= QUESITI ACCETTABILI
- ✓0.30-0.39= QUESITI BUONI
- ✓>0.40= QUESITI OTTIMI

Il coefficiente di discriminatività e la correlazione punto-biseriale

	Quesito 1	Quesito 2	Quesito 3
Estremo superiore	5	4	6
Estremo inferiore	1	5	0
Numero studenti per fascia	6	6	6
Discriminatività	0,7	-0,2	1
Correlazione punto biseriale	0,6	-0,2	0,7

La correlazione punto-biseriale in Excel

Esempio

	Items									
Students	1	2	3	4	5	6	7	8	9	10
Kid-A	1	1	1	1	1	1	1	1	0	1
Kid-B	1	1	1	1	1	1	1	0	1	0
Kid-C	1	1	1	1	1	1	0	1	0	0
Kid-D	1	1	1	1	1	0	1	0	1	0
Kid-E	1	1	1	1	1	1	0	1	0	0
Kid-F	1	1	1	0	1	0	0	0	0	0
Kid-G	1	1	0	1	0	1	0	0	0	0
Kid-H	1	0	1	0	1	0	0	0	0	0
Kid-I	0	1	1	0	0	0	0	0	0	0

	A	B	C	D	E	F	G	H	I	J	K	L
1	Items / Students	1	2	3	4	5	6	7	8	9	10	Student Total Score
2	Kid-A	1	1	1	1	1	1	1	1	0	1	9
3	Kid-B	1	1	1	1	1	1	1	0	1	0	8
4	Kid-C	1	1	1	1	1	1	0	1	0	0	7
5	Kid-D	1	1	1	1	1	0	1	0	1	0	7
6	Kid-E	1	1	1	1	1	1	0	1	0	0	7
7	Kid-F	1	1	1	0	1	0	0	0	0	0	4
8	Kid-G	1	1	0	1	0	1	0	0	0	0	4
9	Kid-H	1	0	1	0	1	0	0	0	0	0	3
10	Kid-I	0	1	1	0	0	0	0	0	0	0	2
11	Item total	8	8	8	6	7	5	3	3	2	1	

=sum(B2:K2)

=sum(B2:B10)

La correlazione punto-biseriale in Excel

Esempio

=L2-B2

		M	N	O	P	Q	R	S	T	U	V
1	Students	total-item1	total-item2	total-item3	total-item4	total-item5	total-item6	total-item7	total-item8	total-item9	total-item10
2	Kid-A	8	8	8	8	8	8	8	8	9	8
3	Kid-B	7	7	7	7	7	7	7	8	7	8
4	Kid-C	6	6	6	6	6	6	7	6	7	7
5	Kid-D	6	6	6	6	6	7	6	7	6	7
6	Kid-E	6	6	6	6	6	6	7	6	7	7
7	Kid-F	3	3	3	4	3	4	4	4	4	4

=CORREL(B2:B10, M2:M10)

		M	N	O	P	Q	R	S	T	U	V
		total-item1	total-item2	total-item3	total-item4	total-item5	total-item6	total-item7	total-item8	total-item9	total-item10
12	Point-Biserial	0.46	0.29	0.12	0.73	0.49	0.49	0.59	0.46	0.26	0.40

Analisi dell'affidabilità di un test

Lo studio dell'affidabilità o attendibilità di un questionario riguarda l'analisi della coerenza tra i punteggi ottenuti dalla somministrazione di una prova se questa viene somministrata in momenti successivi.

ATTENDIBILITA'  è definibile come il rapporto fra la componente sistematica e la variabilità totale della misura

Affinché una misura sia valida, si richiede che essa misuri sistematicamente qualcosa. L'attendibilità rappresenta la precisione di una misura (ciò che nella misura non è errore).

Analisi dell'affidabilità di un test

La variazione dei punteggi può essere dovuta a componenti vere e a componenti di errore. Si può allora affermare che la varianza totale è pari alla somma della varianza delle componenti vere e di quella delle componenti di errore

$$\sigma^2_X = \sigma^2_T + \sigma^2_E$$

L'attendibilità di un test può essere definita come la proporzione della varianza vera rispetto alla varianza totale di una distribuzione dei punteggi

$$\rho^2_{XT} = \sigma^2_T / \sigma^2_X$$

Analisi dell'affidabilità di un test

L'attendibilità di un test può essere scritta come

$$\rho^2_{XT} = 1 - \frac{\sigma^2_E}{\sigma^2_X}$$

da cui si ricava che

$$\sigma_E = \sqrt{\sigma^2_X(1 - \rho^2_{XT})}$$

Si comprende la relazione inversa che lega l'affidabilità all'errore

$\text{se } \rho^2_{XT} = 1$ ➡ Tutta la variazione dei valori osservati è attribuibile ai punteggi veri

$\text{se } \rho^2_{XT} = 0$ ➡ Tutta la variazione dei valori osservati è attribuibile all'errore

Interpretazione dell'affidabilità

Varia tra:

- **0, il punteggio è composto solo da errore**
  1, il punteggio osservato è vero

da 0 a 1 si esprimono valori intermedi di attendibilità

L'attendibilità è buona se superiore a 0,70 – 0,80; è massima se è 1.

Analisi dell'affidabilità di un test

Poiché la varianza dei punteggi osservati può essere considerata in termini di somma della varianza del punteggio vero e della varianza dell'errore, possiamo esprimere

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2}$$

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

da cui si deduce che il quadrato della correlazione tra i punteggi osservati e i punteggi veri è uguale alla correlazione tra i punteggi osservati di due misurazioni parallele.

Quindi l'affidabilità dei punteggi diventa maggiore se la percentuale di varianza dell'errore è piccola e viceversa.

La radice quadrata dell'indice di affidabilità rappresenterà la correlazione tra i punteggi veri e i punteggi osservati.

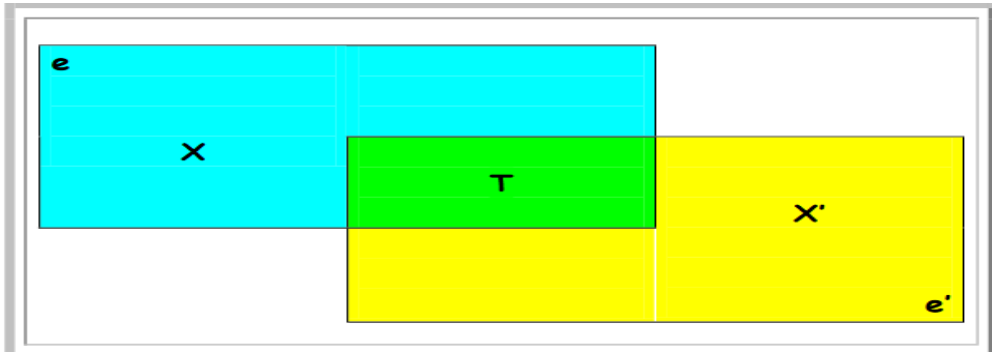
Formalmente si ha….

La correlazione tra le due misure ripetute x e x' consente di stimare l'affidabilità di una misura.

Le due misure ripetute (x e 'x) sono dette parallele se presentano:

1. uguali punteggi veri attesi per ciascun oggetto,
2. uguali varianze dell'errore o, in modo equivalente, errori standard di misurazione, ovvero se risultano vere le seguenti relazioni

$$X = T + E \qquad X' = T' + E' \qquad T = T' \qquad \sigma_E^2 = \sigma_{E'}^2$$



Formalmente si ha….

La correlazione tra misure parallele può essere espressa in funzione dell'errore, del punteggio vero e del punteggio osservato:

$$\rho_{XX'} = \frac{\sigma_{XX'}}{\sigma_X \sigma_{X'}} = \frac{\sigma(T+E)(T+E')}{\sigma_X \sigma_{X'}} = \frac{\sigma_T^2 + \sigma_{TE} + \sigma_{TE'} + \sigma_{EE'}}{\sigma_X \sigma_{X'}}$$

ASSUMIAMO CHE

1. **GLI ERRORI NON SONO CORRELATI NE' TRA LORO, NE' CON I PUNTEGGI VERI**
2. **LE DEVIAZIONI STANDARD DELLE MISURE PARALLELE SONO UGUALI**

La correlazione tra misure parallele è uguale al rapporto tra varianza dei punteggi veri e varianza dei punteggi osservati:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2}$$

Formalmente si ha….

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2}$$

Tale risultato è importante in quanto consente di esprimere la varianza del punteggio vero (inosservabile) in termini di $\rho_{XX'}$ e σ_X^2, entrambi osservabili:

$$\sigma_T^2 = \sigma_X^2 \rho_{XX'}$$

ovvero la varianza del punteggio vero è uguale al prodotto tra la varianza osservata e la correlazione tra misure parallele. Ricordando che l'affidabilità è stata definita come $\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2}$, ne segue che la stima dell'affidabilità non è altro che la correlazione tra misure parallele:

$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_X^2 \rho_{XX'}}{\sigma_X^2} = \rho_{XX'}$. Tale risultato rappresenta un importante risultato nel tentativo di stimare l'affidabilità delle misure empiriche.

L'attendibilità

Il coefficiente ρ^2_{XT}

È il coefficiente di attendibilità del test

RELIABILITY

$\rho_{XX'}$ → È l'indice di affidabilità

Serve a valutarla empiricamente

L'indice di affidabilità tuttavia fornisce indicazioni circa il livello di attendibilità dell'intero questionario, ma non dà alcuna informazione in merito all'affidabilità dei singoli item; questo limite viene superato, come vedremo, dall'IRT.

Analisi dell'affidabilità di un test

La Teoria Classica dei Test studia l'attendibilità di un questionario in modo non univoco.

Il coefficiente di attendibilità è stimato in vari modi:

1) come correlazione tra i punteggi conseguiti da un gruppo di soggetti in due metà dello stesso test (metodo dello Split-Half);
2) come correlazione tra le due distribuzioni di punteggi ottenute applicando due volte uno stesso test ad uno stesso gruppo di soggetti (metodo test-retest);
3) come correlazione tra le due serie di punteggi ottenute somministrando allo stesso gruppo di soggetti due forme parallele dello stesso test (parallel form)
4) come studio della coerenza o omogeneità tra gli item (alpha di Cronbach)

Lo split-half

Si somministra il test in un unico tempo **T1.**

Si divide il test a metà e si considerano le due metà come forme parallele (stessa media e stessa dev. St.)

La suddivisione delle variabili in due gruppi deve essere effettuato in modo che essi risultino omogenee tra loro.

I questionari per la valutazione degli apprendimenti sono costruiti in modo tale che i diversi item abbiano un livello di difficoltà crescente. Di conseguenza una suddivisione a metà in relazione all'ordinamento potrebbe determinare che nella parte più semplice si ottengano risultati migliori rispetto alla parte più complessa. Per ovviare a questa problematica si adotta il metodo "odd-even".

Lo split-half

Prendiamo i seguenti dati di un test di 10 item somministrato a 11 soggetti:

Item	1	2	3	4	5	6	7	8	9	10
Sogg.										
1	1	0	1	1	0	1	0	1	1	1
2	0	0	1	0	1	1	0	1	0	0
3	1	1	0	1	1	1	1	0	1	1
4	0	0	1	0	0	1	0	1	0	0
5	1	0	1	1	0	1	0	0	1	1
6	1	1	0	0	1	0	1	1	0	0
7	0	0	1	0	0	0	1	0	0	0
8	1	1	1	1	0	1	0	0	0	0
9	0	1	0	0	1	1	1	0	0	1
10	1	0	0	1	0	0	1	0	0	1
11	1	1	1	0	1	0	0	1	1	0

Sommiamo gli item pari e dispari

Sogg.	Item Dispari	Item Pari
1	3	4
2	2	2
3	4	4
4	1	2
5	3	3
6	3	2
7	2	0
8	2	3
9	2	3
10	2	2
11	4	2
Media	2,55	2,45
Dev.St.	0,89	1,08

La correlazione tra le due metà=0.40

Interpretazione coefficiente split-half

L'attendibilità dipende molto dalla lunghezza del test.

la correlazione split-half è una sottostima dell'attendibilità.
Infatti la divisione del test a metà ne dimezza la lunghezza.

Ci sono delle formule che permettono di correggere tale sottostima.

Formula di Spearman-Brown

$$\rho^{*}_{XX'} = \frac{n\rho_{xx'}}{1 + (n-1)\rho_{xx'}}$$

Attendibilità dell'intero test

Attendibilità iniziale

n= numero delle volte che il test viene "allungato" o "abbreviato".
Si ottiene come rapporto tra numero degli item iniziali su numero item finali.

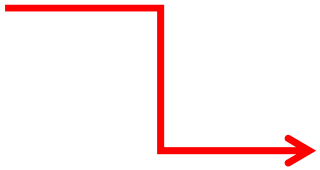
Applicazione al problema precedente

$$\rho_{xx'} = 0.40 \qquad n = \frac{10}{5} = 2$$

$$\rho^{*}_{XX'} = \frac{2x0.40}{1 + (2-1)x0.40} = \frac{0.80}{1.40} = 0.58$$

Il metodo del Test-Retest

Si somministra il **test al tempo T1 ed al tempo T2 e si calcola** la correlazione tra i punteggi.
Questo metodo non necessità di ulteriori specificazioni. Basta saper calcolare la **r di Pearson tra due serie di** punteggi.



Il coefficiente di attendibilità ottenuto esprime, quindi, il grado di stabilità del test e di generalizzabilità dei risultati in caso di somministrazioni diverse.
Quanto più alto risulterà il coefficiente di attendibilità, tanto minore sarà l'influenza delle variabili accidentali sui punteggi.

Il metodo del Test-Retest

	PG T1	PG T2
ss1	11	12
ss2	15	14
ss3	17	14
ss4	20	19
ss5	20	21
ss6	25	27
ss7	22	18
ss8	21	24
ss9	34	31
ss10	38	36
ss11	40	37
Media	23,91	23,00
Dev St.	9,03	8,39

$$Cov\ (PG\ T1, PG\ T2) = 73.45$$

$$Dev.St.(PG\ T1) = 9.03$$

$$Dev.St.(PG\ T2) = 8.39$$

$$\rho_{xx'} = \frac{Cov\ (PG\ T1, PG\ T2)}{Dev.St.(PG\ T1) \times Dev.St.(PG\ T2)}$$

$$73.45 / 9.03 \times 8.39 =$$

$$73.45 / 75.74 = \mathbf{.96}$$

Interpretazione coefficiente test-retest

- Buoni coefficienti test-retest dovrebbero superare .80 (livello piuttosto esigente).

- Il coefficiente test-retest si riduce all'aumentare del tempo trascorso fra le rilevazioni.

- Il coefficiente test-retest è interpretabile se si assume che il concetto misurato non si modifichi nel tempo

Parallel form

Si somministrano due versioni equivalenti del test (**stessa media e stessa dev. St.).** Quindi si calcola la correlazione tra i le due forme come stima dell'attendibilità test-retest.

OSSERVAZIONE:

•Se le due versioni del test sono state somministrate a distanza di tempo, tale coefficiente di correlazione diventa un indice sia della stabilità del tempo, sia della coerenza delle risposte date a diversi campioni di prove che presentano contenuti simili**.**

•Se invece le due forme vengono somministrate consecutivamente, il coefficiente di correlazione trovato esprimerà il grado di attendibilità tra le due versioni, ma non tra i due tempi di applicazione.

Parallel form

Il problema principale è quello di verificare che le due forme siano effettivamente parallele. Ciò significa verificare che le due forme abbiano la stessa media e la stessa varianza.

Prendiamo i seguenti dati:

	T1 Forma A	T2 Forma B
ss1	11	12
ss2	15	14
ss3	17	14
ss4	20	19
ss5	20	21
ss6	25	27
ss7	22	18
ss8	21	24
ss9	34	31
ss10	38	36
ss11	40	37
Media	23,91	23,00
Dev St.	9,03	8,39

t-test sulle due medie

t_{sp}=1.3; *Gdl*=10; t_{cr}=2,23

Le medie sono uguali.

Fare Test F sulle due varianze

F_{sp}=1.15, *gdl* = 10,10
F_{cr}=2.97

Le varianze sono uguali.

Correlazione

$$\rho_{xx'} = 0.96$$

L'alpha di Cronbach

L'attendibilità di un test può essere studiata in funzione della **coerenza od omogeneità degli item all'interno di un test.**

L'attendibilità viene quindi stimata utilizzando l'informazione contenuta negli item. Il coefficiente più utilizzato è l'alpha di Cronbach.

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_X^2}\right)$$

Si utilizza per item dicotomici o politomici

in cui:
k = numero di item
δ^2_i = varianza di ciascun item
δ^2_x = varianza totale del test

Esprime una misura del peso relativo della variabilità associata agli item rispetto alla variabilità associata alla loro somma.

alpha di Cronbach: esempio dati dicotomici

Item Sogg.	1	2	3	4	5	6	7	8	9	10	Totale
1	1	0	1	1	0	1	0	1	1	1	7
2	0	0	1	0	1	1	0	1	0	0	4
3	1	1	0	1	1	1	1	0	1	1	8
4	0	0	1	0	0	1	0	1	0	0	3
5	1	0	1	1	0	1	0	0	1	1	6
6	1	1	0	0	1	0	1	1	0	0	5
7	0	0	1	0	0	0	1	0	0	0	2
8	1	1	1	1	0	1	0	0	0	0	5
9	0	1	0	0	1	1	1	0	0	1	5
10	1	0	0	1	0	0	1	0	0	1	4
11	1	1	1	0	1	0	0	1	1	0	6
Media	0,64	0,45	0,64	0,45	0,45	0,64	0,45	0,45	0,36	0,45	5,00
Dev.St.	0,48	0,50	0,48	0,50	0,50	0,48	0,50	0,50	0,48	0,50	1,65
Varianza	0,23	0,25	0,23	0,25	0,25	0,23	0,25	0,25	0,23	0,25	2,73

Il coefficiente α di Cronbach

$$\alpha = \frac{K}{K-1} \times \left(1 - \frac{\sum \sigma^2_i}{\sigma^2_{totale}}\right) =$$

$$= \frac{10}{9} \times \left(1 - \frac{.23 \times 4 + .25 \times 6}{2.73}\right) =$$

$$= \frac{10}{9} \times \left(1 - \frac{2.42}{2.73}\right) = .1261$$

Alpha di Cronbach: intepretazione

l' Alpha di Cronbach:
dipende dalla media delle intercorrelazioni tra tutti gli item del test, e dalla relazione di ogni item del test con il punteggio totale.

Nella prassi si valuta così:

α di Cronbach	Coerenza interna del test
$\alpha \geq 0.9$	Eccellente
$0.8 \leq \alpha < 0.9$	Buono
$0.7 \leq \alpha < 0.8$	Accettabile
$0.6 \leq \alpha < 0.7$	Discutibile
$0.5 \leq \alpha < 0.6$	Povero
$0.\alpha < 0.5$	Inaccettabile

Considerazioni pratiche

Per meglio valutare l'Alpha di Cronbach, bisogna ricordare che essa **dipende da due fattori:**

1. **intercorrelazioni tra gli item**
2. **lunghezza della scala (numero di item)**

Infatti, a parità di condizioni, all'aumentare del numero degli item, aumenta il valore del coefficiente di attendibilità.

Perciò, per esempio, basterebbe aumentare il numero di item di una scala che ha un coefficiente sufficiente per ottenerne uno buono.

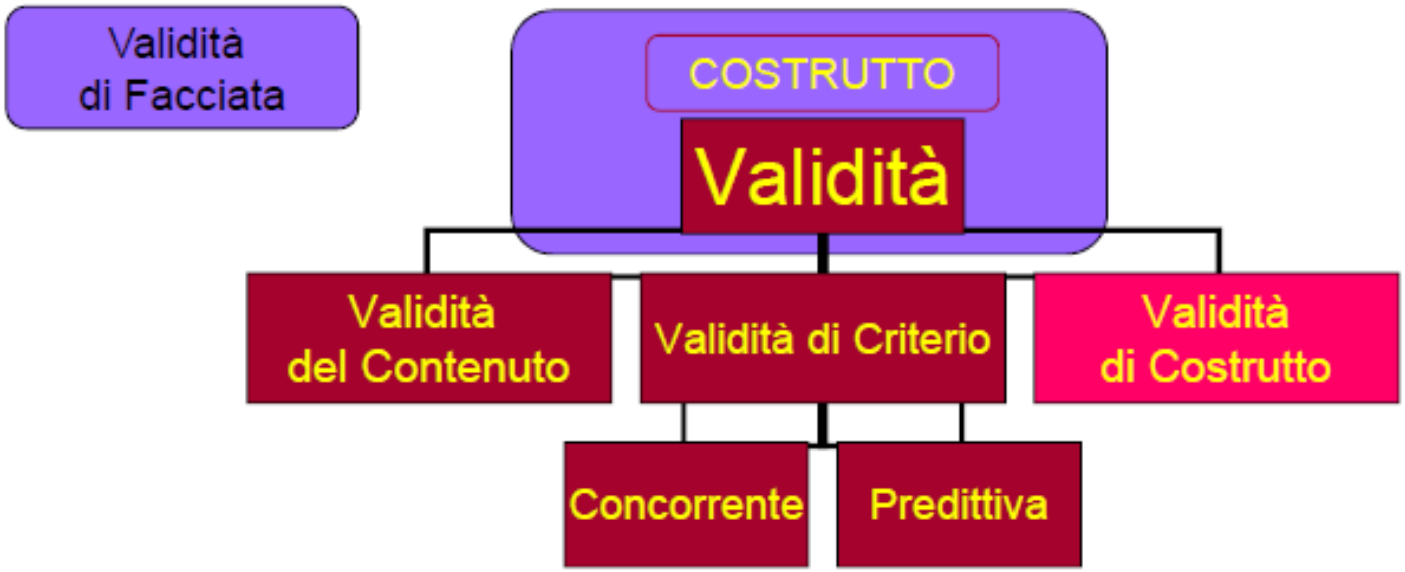
Analisi della validità di un test

- Una misura è valida quando misura ciò che intende misurare.

- Si tratta di "*un giudizio complessivo della misura in cui prove empiriche e principi teorici supportano l'adeguatezza e l'appropriatezza delle conclusioni basate su punteggi al test*" (Messick, 1989)

- La validità si articola in diverse sfumature, e in diverse modalità di giudicare, raggiunte o meno, differenti forme di validità

Tassonomia tradizionale



I diversi aspetti della validità

Validità interna:

gli item sono misure del costrutto sufficientemente correlate tra loro. L'analisi fattoriale consente di verificare tale aspetto.

Validità di criterio

grado di corrispondenza tra una misura ed un criterio di riferimento. Si distingue in:

Validità concorrente

misura e criterio sono rilevati nello stesso momento

Validità predittiva

il criterio è misurato successivamente alla misura.

Validità di criterio

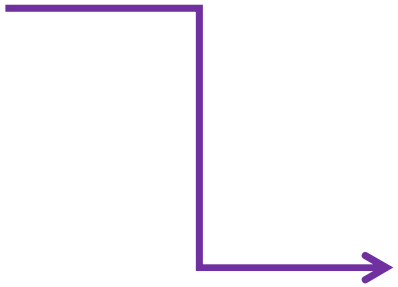
Esempio di validità concorrente:
Correlazione tra quoziente intellettivo misurato tramite un test di intelligenza e la risoluzione di un determinato problema di una certa complessità cognitiva (criterio)

Esempio di validità predittiva:
Correlazione tra il punteggio ottenuto ad un test attitudinale nella fase di selezione del personale per una certa azienda ed il successo lavorativo (ad es., velocità della carriera interna) ottenuto negli anni successivi (criterio).

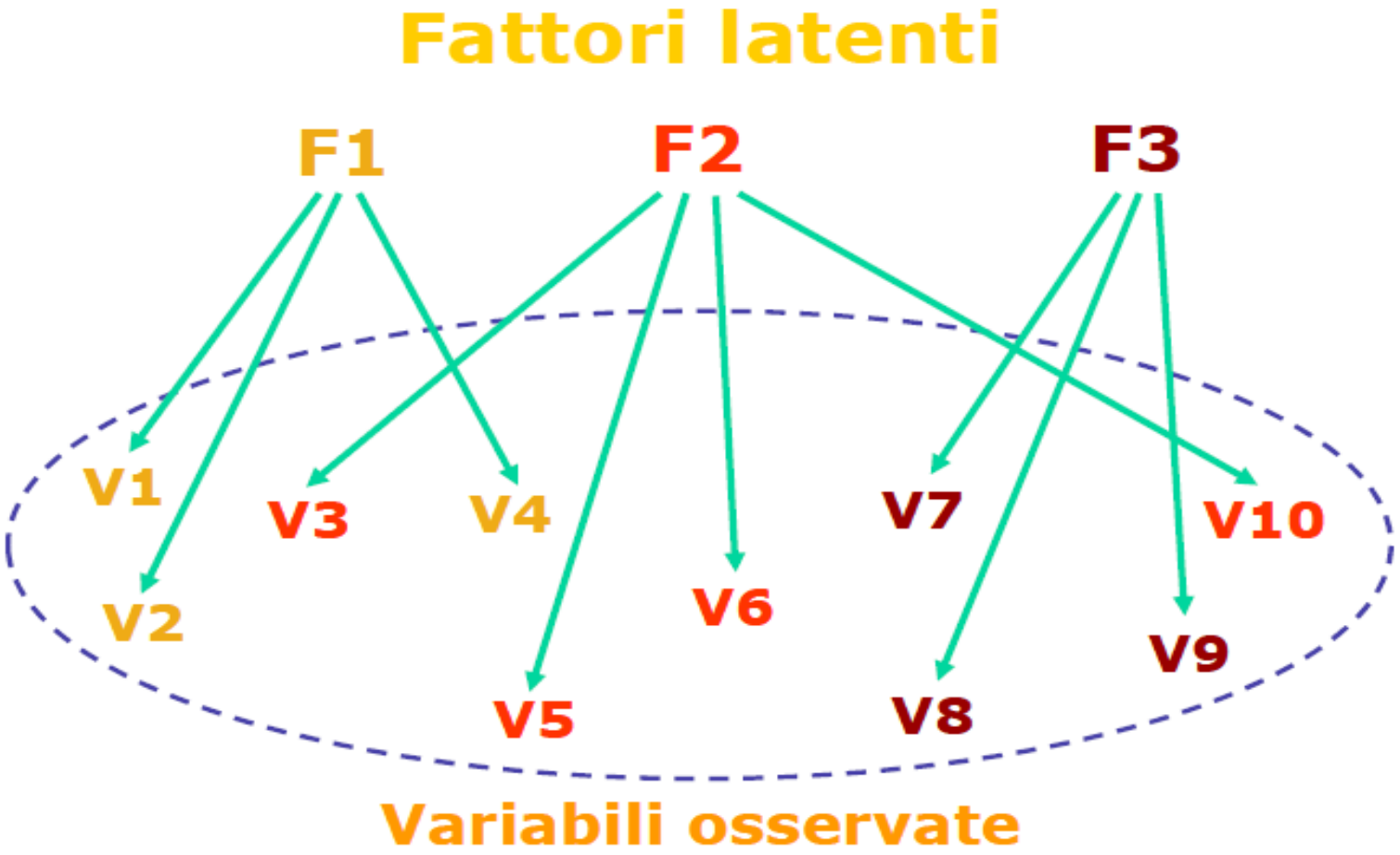
Analisi Fattoriale
la dimensionalita' di un test

Studiare la dimensionalità di un test significa studiare le dimensioni latenti ad un insieme di item e ciò viene effettuato tramite **l'Analisi Fattoriale.**

La dimensione latente può essere considerata un costrutto teorico ipotetico, idealmente corrispondente al costrutto teorico inizialmente ipotizzato.



L'analisi fattoriale viene utilizzata per identificare un numero di fattori latenti (dimensioni, tratti, componenti) che spieghino le correlazioni tra le variabili osservate (indicatori, item) **in modo parsimonioso**.
I fattori perciò sono sempre in **numero molto minore rispetto agli item analizzati.**


Fattori latenti
F1
F2
F3
V1
V2
V3
V4
V5
V6
V7
V8
V9
V10
Variabili osservate

Correzione per guessing

La **correzione per il guessing** scompone Il numero totale di risposte corrette in due componenti:

1. **le risposte corrette dovute alle conoscenze del soggetto**
2. **e quelle che risultano corrette come effetto del caso**.

Individui con lo stesso numero di risposte corrette, ma un diverso numero di errate e/o omesse, non otterranno lo stesso punteggio corretto per "guessing."

$$P_g = C - \frac{E}{(K-1)}$$

Dove:

- P_g è il PUNTEGGIO CORRETTO PER GUESSING
- C indica il numero delle risposte CORRETTE
- E indica il numero delle risposte ERRATE
- K indica le possibili alternative di risposta

Correzione per guessing
Esempio

Consideriamo il caso **di Tonio e Riccardo, che hanno sostenuto la stessa prova d'esame. Il questionario è composto da 40 item a 5 alternative di risposta.**

Tonio risponde correttamente a 32 risposte, ne sbaglia 2 e ne omette 6; Riccardo ottiene 32corrette, 6 errate e 2 omesse, quindi sommando semplicemente le risposte corrette entrambi gli studenti otterrebbero un punteggio grezzo di 32.

I loro punteggi corretti saranno, invece:

$$X_{Tonio} = 32 - \frac{2}{(5-1)} = 31,50 \qquad X_{Riccardo} = 32 - \frac{6}{(5-1)} = 30,50$$

Per effetto della correzione per guessing a parità di risposte esatte, Tonio ottiene un punteggio superiore rispetto a Riccardo nella graduatoria finale dei punteggi grezzi.

Correzione per Guessing
-Esempio

Un questionario può essere costituito da item di diversa tipologia risposte a scelta multipla, risposte dicotomiche, risposte di tipo cloze,ecc..

Per fare un semplice esempio concreto, **supponiamo che un questionario sia composto da 5 domande Vero/Falso (gruppo A), 4 domande a tre alternative di'risposta (gruppo B) e 4 domande di tipo cloze a quattro alternative di completamento (gruppo C)**. Tonio ottiene 3 risposte corrette al gruppo A di domande, 2 al gruppo B e 3 al gruppo C. Il suo punteggio totale sarebbe quindi pari a 8. Il suo punteggio totale corretto, invece, per ogni blocco di domande sarà:

$$X_{Tonio,A} = 3 - \frac{2}{(2-1)} \qquad X_{Tonio,B} = 2 - \frac{2}{(3-1)}$$

$$X_{Tonio,C} = 3 - \frac{1}{(4-1)}$$

da cui un punteggio totale corretto è uguale a:

$$X_{Tot} = X_{Tonio,A} + X_{Tonio,B} + X_{Tonio,C} = 2 + 1 + 2,66 = 5,66$$

ANALISI DI UN TEST SECONDO LA TCT:

Caso di studio

Un questionario per la valutazione dell'abilità di comprensione del testo è stato somministrato a tutti gli studenti delle classi quarte di una Scuola Primaria.

1. Hanno compilato il questionario 84 studenti
2. Prova costituita da 23 item a scelta multipla

Analisi di facilità degli item

Proportions for each level of response:

	0	1	logit
X1t	0.1548	0.8452	1.6977
X2t	0.0595	0.9405	2.7600
X3t	0.0119	0.9881	4.4188
X4t	0.2738	0.7262	0.9754
X5t	0.0476	0.9524	2.9957
X6t	0.3571	0.6429	0.5878
X7t	0.0595	0.9405	2.7600
X8t	0.1786	0.8214	1.5261
X9t	0.0357	0.9643	3.2958
X10t	0.1548	0.8452	1.6977
X11t	0.2738	0.7262	0.9754
X12t	0.4762	0.5238	0.0953
X13t	0.8690	0.1310	-1.8926
X14t	0.4167	0.5833	0.3365
X15t	0.3095	0.6905	0.8023
X16t	0.3929	0.6071	0.4353
X1l	0.4881	0.5119	0.0476
X2l	0.2024	0.7976	1.3715
X3l	0.2619	0.7381	1.0361
X4l	0.5952	0.4048	-0.3857
X5l	0.2976	0.7024	0.8587
X6l	0.2262	0.7738	1.2299
X7l	0.2857	0.7143	0.9163

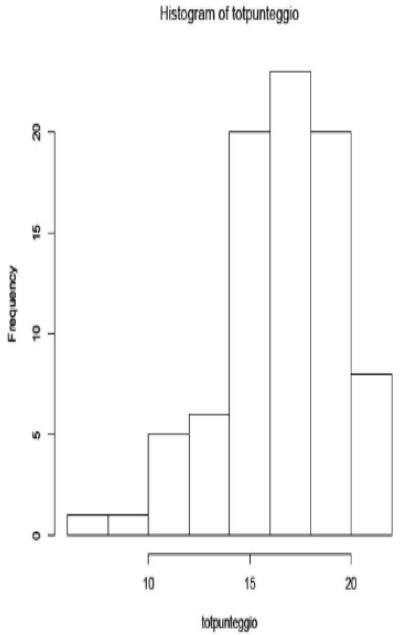
troppo facile

troppo difficile

Analisi della correlazione punto biseriale e dell'alpha di Cronbach

La correlazione punto-biseriale
Varia da -1 a 1

	Point Biserial correlation with Total Score:	Cronbach's alpha all Items: **0,6733**
X1t	0.3883	0.6587
X2t	0.3531	0.6624
X3t	0.1878	0.6716
X4t	0.1241	0.6891
X5t	0.4511	0.6577
X6t	0.4664	0.6513
X7t	0.3531	0.6624
X8t	0.3395	0.6637
X9t	0.4080	0.6613
X10t	0.3277	0.6644
X11t	0.4438	0.6536
X12t	0.2917	0.6748
X13t	0.0294	0.6878
X14t	0.5486	0.6400
X15t	0.4890	0.6481
X16t	0.1561	0.6898
X1l	0.3687	0.6650
X2l	0.5796	0.6379
X3l	0.5285	0.6431
X4l	0.4138	0.6587
X5l	0.3460	0.6657
X6l	0.3919	0.6591
X7l	0.2000	0.6817



Analisi dei distrattori

$`item_ X4t`			
score.level			
resp.	lower	middle	upper
*A	0.703	0.619	0.846
B	0.027	0.143	0.000
C	0.162	0.095	0.154
D	0.108	0.143	0.000

$`item_ X13t`			
score.level			
resp.	lower	middle	upper
A	0.378	0.238	0.269
*B	0.135	0.190	0.077
C	0.351	0.524	0.615
D	0.135	0.048	0.038

$`item_ X16t`			
score.level			
resp.	lower	middle	upper
A	0.054	0.000	0.000
B	0.135	0.048	0.000
C	0.243	0.381	0.308
*D	0.568	0.571	0.692

$`item_ X7l`			
score.level			
resp.	lower	middle	upper
A	0.135	0.190	0.077
*B	0.676	0.667	0.808
C	0.135	0.048	0.077
D	0.054	0.095	0.038

13t. I tempi dei fatti narrati sono:

A. Recenti e precisati.

B. Passati e non precisati.

C. Recenti e non precisati.

D. Passati e precisati.

Limiti della TCT

I limiti della TCT riguardano l'impossibilità

- di separare le caratteristiche delle persone da quelle degli item;

- di determinare, nella pratica, indici per la verifica dell'affidabilità dei test;

- di studiare il comportamento di un singolo individuo nei confronti di un singolo item in quanto si limita a fornire statistiche a livello generale dei test.