Testing psicologico

Modelli e metodi statistici per la misurazione in psicologia

Antonio Calcagnì

Dipartimento di Psicologia dello Sviluppo e della Socializzazione (DPSS)
Università di Padova

A.A. 2019/2020

Introduzione

Nel modulo A abbiamo affrontato la misurazione da una *prospettiva fisica* con enfasi sulla teoria degli errori casuali e la teoria GUM dell'incertezza. Nel modulo B sono stati invece richiamati alcuni fondamenti statistici alla base di questi due paradigmi sulla misurazione.

L'approccio psicometrico alla misurazione maggiormente utilizzato condivide, in parte, alcuni fondamenti alla base della prospettiva fisica (es.: misurandi, indicatori, principio della ripetibilità della misurazione) e le tecniche psicometriche utilizzate richiedono i presupposti statistici richiamati nel modulo B.

Tuttavia affrontare la misurazione da una prospettiva psicometrica è arduo, per ragioni sia di carattere epistemologico che metodologico. La tradizione psicometrica ha negli anni fornito diversi modelli per la misurazione, per i quali non c'è accordo e molte sono le critiche avanzate alla *misurabilità* in psicologia.

Introduzione

In questo corso non affronteremo la questione della *misurabilità* degli oggetti psicologici: il dibattito è abbastanza avanzato e diverse sono le posizioni in merito. Recentemente, tale questione è stata anche affrontata dalla metrologia, proponendo - come richiamato nel modulo A - uno studio approfondito delle c.d. *soft measurements*.

Nella sezione APPROFONDIMENTI sono disponibili tre articoli dello psicometrista australiano *Joe Michell* sulla questione della misurazione in psicometria:

Michell, J. (2013). Constructs, inferences, and mental measurement. *New Ideas in Psychology*, 31(1), 13-21.

Michell, J. (2008). Is psychometrics pathological science?. *Measurement*, 6(1-2), 7-24.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88(3), 355-383.

In essi si possono trovare argomentazioni alle domande: *cosa misuriamo in psicologia? Cosa sono i costrutti latenti? Esiste la possibilità di misurare facoltà umane? La psicometria offre davvero strumenti di misura?* Ecc.

Introduzione

In questo corso invece verranno fornite alcune metodologie e tecniche psicometriche adottando una *prospettiva statistica*, con maggiore enfasi sul lato tecnico ed applicato.

In questo senso, i metodi e le tecniche che verranno presentati nei moduli C-D sono intesi come tecniche di *esplorazione ed analisi dei dati* rilevati utilizzando strumenti tipici delle discipline psicologiche (es.: questionari, test).

Partiremo dal *problema del dato* così come è stato rilevato (qualitativo, quantitativo) e ci chiederemo quali metodi e tecniche statistiche permettono di meglio indagarlo.

Terminologia psicometrica e metrologica

Un **test** è uno strumento di misura usato per quantificare oggetti di interesse psicologico non osservabili (**costrutti latenti**). Tipicamente consiste di un insieme di **items** (o domande/indicatori) che possono essere aggregati insieme per formare una **scala**. Questa è utilizzata per quantificare un costrutto latente ed è un *proxy* di quest'ultimo. Un test è solitamente formato da più scale (spesso correlate tra loro). Il numero di costrutti latenti, individuato dalle scale, è detto **dimensione di un test**.

metrologia	psicometria
misurando	costrutto/tratto latente
misura	scala (se aggregata)
misura	misura osservabile (non aggregata)
misura	item/indicatore
precisione di una misura	attendibilità di una scala
misurazione indiretta	misure composite
strumento di misura	test

Nota: Useremo in maniera interscambiabile i termini metrologici e psicometrici.

Teoria Classica dei Test

FONTI: BN(2.1,2.2)

Questo modello, noto anche come *modello del punteggio vero*, è il modello standard utilizzato per la costruzione e la valutazione dei test. Questo modello è ampiamente basato sulla *teoria degli errori accidentali* presentata nel modulo A ed ha diverse connessioni con l'approccio GUM alla gestione dell'incertezza.

Testo fondamentale per la formalizzazione statistica del modello classico (disponibile in APPROFONDIMENTI):

> Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Welsley Publishing Company [ristampa: 2008, editore IAP]

Diversi sono i modelli psicometrici che integrano o superano il modello classico, ad esempio *Item Response Theory* (IRT), *Generalizability Theory* (G-Theory).

Teoria Classica dei Test

Studio individuale

BN(1.1-1.4)

cap.1 di Lord & Novick (1968) disponibile in APPROFONDIMENTI

Teoria Classica dei Test

Richiami di probabilità: Covarianza e correlazione

https://en.wikipedia.org/wiki/Correlation_and_dependence

Date due v.a. X_1 e X_2 con media finita $\mathbb{E}[X_1] = \mu_{X_1}$ e $\mathbb{E}[X_2] = \mu_{X_2}$, la **covarianza** tra X_1 e X_2 è definita come di seguito:

$$\mathbb{C}\text{ov}[X_1, X_2] = \mathbb{E}[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})]$$

ed esprime l'ammontare di *variazione congiunta* di X_1 e X_2. La covarianza può essere maggiore o minore di 0, ossia $\mathbb{C}\text{ov}[X_1, X_2] \in \mathbb{R}$, ed è indicata anche come $\sigma_{X_1 X_2}$.

La **correlazione** è definita come:

$$\mathbb{C}\text{or}[X_1, X_2] = \frac{\mathbb{C}\text{ov}[X_1, X_2]}{\sqrt{\mathbb{V}\text{ar}[X_1]}\sqrt{\mathbb{V}\text{ar}[X_2]}}$$

ed assume valori in $[-1, 1]$. Spesso è indicata con la lettera greca $\rho_{X_1 X_2}$.

Nota: $\sqrt{\mathbb{V}\text{ar}[X]}$ è lo scarto quadratico medio (o *standard deviation*) di X ed anche indicato con σ_X.

Introduzione

Richiami di probabilità: Covarianza e correlazione

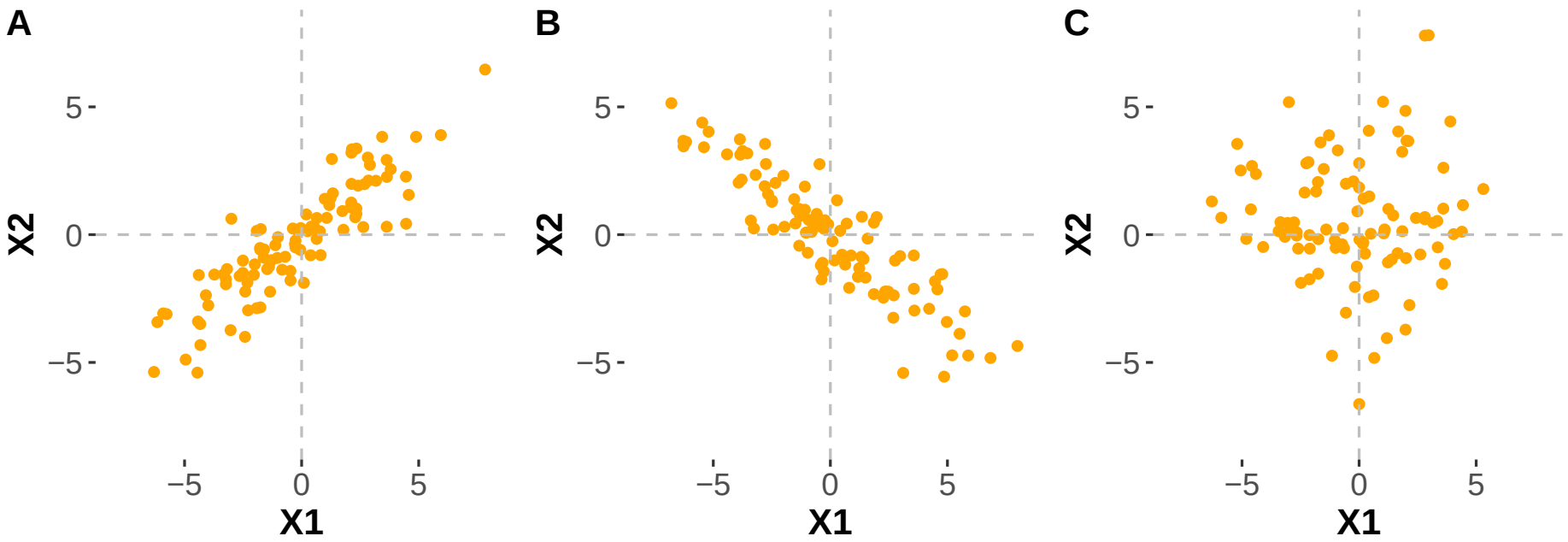


Grafico a dispersione bivariato: Rappresentazione grafica dei valori osservati congiuntamente delle variabili X_1 e X_2.
A: Grafico con associazione lineare positiva $\sigma_{X_1 X_2} = 6.003$.
B: Grafico con associazione lineare negativa $\sigma_{X_1 X_2} = -6.003$.
C: Grafico con nessuna associazione lineare $\sigma_{X_1 X_2} \approx 0$.
In tutti e tre i grafici $\mu_{X_1} = \mu_{X_2} = 0$ e varianze $\sigma^2_{X_1} = 9.140$ e $\sigma^2_{X_2} = 5.165$.

Teoria Classica dei Test

FONTI: BN(2.1,2.2)

Consideriamo un insieme di misure (**items**) indicizzate tramite $j = 1, \ldots, p$ e un insieme di individui (**persone**) $i = 1, \ldots, n$ (estratti casualmente da una popolazione $\mathcal{P}$) su cui le misurazioni sono effettuate usando un determinato test.

Alla j-esima misurazione è associata la realizzazione della variabile aleatoria

$$(X_j, T_j, E_j)$$

i cui esiti sono ottenuti campionando su $\mathcal{P}$.

Fissato j,

- X_j: v.a. che governa la realizzazione della j-esima misura *osservabile*
- T_j: v.a. che governa la realizzazione della j-esima misura *vera*
- E_j: v.a. che governa la realizzazione della j-esima componente di *errore*

Teoria Classica dei Test

FONTI: BN(2.1,2.2)

Della tripla (X_j, T_j, E_j) conosciamo solo X_j mentre le altre componenti sono in relazione a questa attraverso la definizione fondamentale della TCT:

$$X_j = T_j + E_j$$

dove:

(i) $\mathbb{E}[E_j] = 0$ l'errore ha media nulla (errore non sistematico)

(ii) $\mathbb{C}\text{or}[E_j, T_j] = 0$ l'errore e il valore vero della misura sono non correlati

(iii) $\mathbb{C}\text{or}[E_j, E_{j'}] = 0$ due misurazioni differenti j e j' hanno errori non correlati

(iv) $\mathbb{E}[X_j] = \mathbb{E}[T_j + E_j]^{(1)} = \mathbb{E}[T_j] + \cancelto{0}{\mathbb{E}[E_j]}$

(v) $\mathbb{V}\text{ar}[X_j] = \mathbb{V}\text{ar}[T_j + E_j]^{(2)} = \mathbb{V}\text{ar}[T_j] + \mathbb{V}\text{ar}[E_j] + \cancelto{0}{2\mathbb{C}\text{ov}[T_j, E_j]}^{(3)}$

---

(1) linearità dell'operatore atteso: $\mathbb{E}[A + B] = \mathbb{E}[A] + \mathbb{E}[B]$

(2) varianza di una combinazione lineare: $\mathbb{V}\text{ar}[A + B] = \mathbb{V}\text{ar}[A] + \mathbb{V}\text{ar}[B] + 2\mathbb{C}\text{ov}[A, B]$

(3) nota che $\mathbb{C}\text{ov}[A, B] = \mathbb{C}\text{or}[A, B]\sqrt{\mathbb{V}\text{ar}[A]}\sqrt{\mathbb{V}\text{ar}[B]}$

Teoria Classica dei Test

FONTI: BN(2.1-2.4)

Abbiamo detto che della tripla (X_j, T, E_j) conosciamo soltanto X_j e la relazione $X_j = T + E_j$. Le assunzioni (i)-(iii) della TCT ci hanno permesso di derivare le definizioni (iv)-(v), le quali ci confortano sul fatto che:

(iv) possiamo determinare il valore vero del misurando (non osservabile) T attraverso l'osservabile X_j

(v) la varianza dell'osservabile X_j è determinata da due fonti di variabilità, quella di T e quella di E_j

Attualmente non abbiamo ancora un'espressione per le quantità (non osservate) $\mathbb{V}\text{ar}[T]$ e $\mathbb{V}\text{ar}[E_j]$ che tuttavia possono essere ottenute ricorrendo all'ausilio delle c.d. **misure parallele**.

Teoria Classica dei Test

FONTI: BN(2.4)

Consideriamo una misurazione effettuata utilizzando due *distinte* osservabili X_j e $X_{j'}$ per un medesimo misurando[†] (o costrutto) T.

Le osservabili X_j e $X_{j'}$ sono dette **parallele** (p_1) quando:

$$X_j = T + E$$
$$X_{j'} = T + E$$

ossia X_j e $X_{j'}$ misurano la stessa quantità latente T con la stessa quantità di errore E.

In questo modello si ha inoltre che

$$\mathbb{E}[X_j] = \mathbb{E}[X_{j'}] \qquad \text{e} \qquad \mathbb{V}\text{ar}[X_j] = \mathbb{V}\text{ar}[X_{j'}]$$

[†] Usando la definizione (iv) della TCT ricordiamo che $\mathbb{E}\left[X_j\right] = \mathbb{E}\left[X_{j'}\right] = \mathbb{E}[T] = \tau$, il valore vero del misurando è il valore atteso delle misurazioni fatte mediante le osservabili X_j e $X_{j'}$.

Teoria Classica dei Test

FONTI: BN(2.4)

Consideriamo una misurazione effettuata utilizzando due *distinte* osservabili X_j e $X_{j'}$ per un medesimo misurando[†] (o costrutto) T.

Le osservabili X_j e $X_{j'}$ sono dette **τ-equivalenti** (p_2) quando:

$$X_j = T + E_j$$
$$X_{j'} = T + E_{j'}$$

ossia X_j e $X_{j'}$ misurano la stessa quantità latente T con diverse quantità di errore E_j e $E_{j'}$.

In questo modello si ha inoltre che

$$\mathbb{E}[X_j] = \mathbb{E}[X_{j'}] \qquad \text{e} \qquad \mathbb{V}\text{ar}[X_j] \neq \mathbb{V}\text{ar}[X_{j'}]$$

[†] Usando la definizione (iv) della TCT ricordiamo che $\mathbb{E}\left[X_j\right] = \mathbb{E}\left[X_{j'}\right] = \mathbb{E}[T] = \tau$, il valore vero del misurando è il valore atteso delle misurazioni fatte mediante le osservabili X_j e $X_{j'}$.

Teoria Classica dei Test

FONTI: BN(2.4)

Consideriamo una misurazione effettuata utilizzando due *distinte* osservabili X_j e $X_{j'}$ per un medesimo misurando[†] (o costrutto) T.

Le osservabili X_j e $X_{j'}$ sono dette **essenzialmente τ-equivalenti** (p_3) quando:

$$X_j = \alpha + T + E_j$$
$$X_{j'} = \alpha + T + E_{j'}$$

ossia X_j e $X_{j'}$ misurano la stessa quantità latente T con diversa precisione e diverse quantità di errore E_j e $E_{j'}$.

In questo modello si ha inoltre che

$$\mathbb{E}[X_j] = \alpha + \mathbb{E}[X_{j'}] \qquad \text{e} \qquad \mathbb{V}\text{ar}[X_j] \neq \mathbb{V}\text{ar}[X_{j'}]$$

Teoria Classica dei Test

FONTI: BN(2.4)

Esempio di misure **essenzialmente τ-equivalenti**:

> *For example, consider a test designed to measure the latent variable depression in which each item is measured on a 5-point Likert-like scale, from strongly disagree to strongly agree. Responses to the items "I feel sad sometimes" and "I almost always feel sad" are likely to share similar distributions, though perhaps with different modes. This might be due to the fact that, though both questions measure the same latent variable on the same scale, the second question is worded more strongly than the first. As long as the variances of these questions are similar across respondents, they are both measuring depression in the same scale, whereas their precision in measuring depression differs.*
>
> *Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. Educational and psychological measurement, 66(6), 930-944.*

Teoria Classica dei Test

FONTI: BN(2.4)

Consideriamo una misurazione effettuata utilizzando due *distinte* osservabili X_j e $X_{j'}$ per un medesimo misurando[†] (o costrutto) T.

Le osservabili X_j e $X_{j'}$ sono dette **congeneriche** (p_4) quando:

$$X_j = \alpha + T\beta + E_j$$
$$X_{j'} = \alpha + T\beta + E_{j'}$$

ossia X_j e $X_{j'}$ misurano la stessa quantità latente T con diversa precisione, diverse scale e diverse quantità di errore E_j e $E_{j'}$.

In questo modello si ha inoltre che

$$\mathbb{E}[X_j] = \alpha + \mathbb{E}[X_{j'}]\,\beta \qquad \text{e} \qquad \mathbb{V}\text{ar}[X_j] \neq \mathbb{V}\text{ar}[X_{j'}]$$

[†] Usando la definizione (iv) della TCT ricordiamo che $\mathbb{E}\left[X_j\right] = \mathbb{E}\left[X_{j'}\right] = \mathbb{E}[T] = \tau$, il valore vero del misurando è il valore atteso delle misurazioni fatte mediante le osservabili X_j e $X_{j'}$.

Teoria Classica dei Test

FONTI: BN(2.4)

Interpretazione delle proprietà (p_1)-(p_4):

(p_1) le misurazioni X_j e $X_{j'}$ hanno lo stesso punteggio vero per tutte le persone che è misurato allo stesso modo da entrambe le misure.

(p_2) sebbene le misurazioni X_j e $X_{j'}$ hanno lo stesso punteggio vero per tutte le persone, questo non è misurato allo stesso modo dalle due misure (le varianze sono infatti diverse).

(p_3) le misurazioni X_j e $X_{j'}$ hanno lo stesso punteggio vero per tutte le persone a meno di una costante α (traslazione), unitamente a varianze differenti.

(p_4) le misurazioni X_j e $X_{j'}$ sono in relazione lineare ed hanno ancora varianze differenti.

Nota: le proprietà (p_1)-(p_4) sono ordinate dalla più forte (p_1) alla più debole (p_4).

Teoria Classica dei Test

FONTI: BN(2.1-2.4)

Per derivare le quantità ignote $\mathbb{V}\text{ar}[E_j]$ e $\mathbb{V}\text{ar}[T_j]$, consideriamo due misure X_j e $X_{j'}$, parallele nel senso di (p_1), e definiamo la loro correlazione:

$$\begin{aligned}\mathbb{C}\text{or}[X_j, X_{j'}] &= \frac{\mathbb{C}\text{ov}[X_j, X_{j'}]}{\sqrt{\mathbb{V}\text{ar}[E_j]}\sqrt{\mathbb{V}\text{ar}[E_{j'}]} } \\ &= \ldots \text{ sviluppando } \mathbb{C}\text{ov}\left[X_j, X_{j'}\right] = \mathbb{E}\left[X_j\, X_{j'}\right] \text{ ed usando gli assunti (i)-(iii) della TCT} \\ &= \frac{\mathbb{V}\text{ar}[T]}{\mathbb{V}\text{ar}[X_j]}\end{aligned}$$

da cui:

(vi) $\mathbb{V}\text{ar}[T] = \mathbb{C}\text{or}[X_j, X_{j'}]\,\mathbb{V}\text{ar}[X_j]$

La varianza del vero misurando T è uguale al prodotto tra la varianza della misura X_j (oppure $X_{j'}$) e la correlazione tra la coppia di misure parallele X_j e $X_{j'}$.

Nota: La varianza ignota di T è ora espressa in funzione esclusivamente di ciò che osserviamo/misuriamo.

Teoria Classica dei Test

FONTI: BN(2.1-2.4)

Proseguendo, utilizziamo la definizione (vi) nella (v) come segue:

$$\begin{aligned}\mathbb{V}\text{ar}\left[X_j\right] &= \mathbb{V}\text{ar}\left[T_j\right] + \mathbb{V}\text{ar}\left[E_j\right] \\ &= \mathbb{C}\text{or}\left[X_j, X_{j'}\right]\mathbb{V}\text{ar}\left[X_j\right] + \mathbb{V}\text{ar}\left[E_j\right]\end{aligned}$$

ri-arrangiando i termini:

$$\mathbb{V}\text{ar}\left[E_j\right] = \mathbb{V}\text{ar}\left[X_j\right] - \mathbb{C}\text{or}\left[X_j, X_{j'}\right]\mathbb{V}\text{ar}\left[X_j\right]$$

da cui:

(vii) $\mathbb{V}\text{ar}\left[E_j\right] = \mathbb{V}\text{ar}\left[X_j\right](1 - \mathbb{C}\text{or}\left[X_j, X_{j'}\right])$

La varianza dell'errore E_j è uguale al prodotto tra la varianza della misura X_j (oppure $X_{j'}$) e uno meno la correlazione tra la coppia di misure parallele X_j e $X_{j'}$.

Nota: La varianza ignota di E_j è ora espressa in funzione esclusivamente di ciò che osserviamo/misuriamo.

Teoria Classica dei Test

FONTI: BN(2.1-2.4)

TCT in sintesi

$$X_j = T + E_j$$

(i) $\mathbb{E}[E_j] = 0$

(ii) $\mathbb{C}\text{or}[E_j, T] = 0$

(iii) $\mathbb{C}\text{or}[E_j, E_{j'}] = 0$

(iv) $\mathbb{E}[X_j] = \mathbb{E}[T]$

(v) $\mathbb{V}\text{ar}[X_j] = \mathbb{V}\text{ar}[T] + \mathbb{V}\text{ar}[E_j]$

(vi) $\mathbb{V}\text{ar}[T] = \mathbb{C}\text{or}[X_j, X_{j'}]^{\dagger}\, \mathbb{V}\text{ar}[X_j]$

(vii) $\mathbb{V}\text{ar}[E_j] = \mathbb{V}\text{ar}[X_j]\left(1 - \mathbb{C}\text{or}[X_j, X_{j'}]^{\dagger}\right)$

† La quantità $\mathbb{C}\text{or}\left[X_j, X_{j'}\right]$ è ottenuta utilizzando una coppia di misure (osservabili) parallele nel senso di (p_1).

Teoria Classica dei Test

FONTI: BN(6.1.3)

Validità di una misura

Quando il nostro interesse è rivolto a capire *cosa* un determinato test misuri siamo nell'ambito della validità di un test (o misura). Un modo per studiare la validità di una misura, ad esempio X, è quello di confrontare quest'ultima con un'altra misura (o più misure), poniamo Y, detta in questo caso *criterio*.

La quantificazione della *validità secondo il criterio Y* avviene usando un indice di correlazione tra X e Y detto *coefficiente di validità*. In questo caso, si parla di validità lineare in quanto l'operatore $\mathbb{C}\text{or}[X, Y]$ è lineare.

Teoria Classica dei Test

FONTI: BN(6.1.3)

Validità di una misura

«The validity coefficient of a measurement can therefore be stated only in relation to a second measurement; thus it makes no sense to speak of the validity coefficient of a measurement. When speaking of the validity coefficient of a measurement, we shall always make explicit the second measurement unless the identity of this measurernent can be clearly inferred from the context» (Lord & Novick, 1968, p. 59)

Teoria Classica dei Test

FONTI: BN(6.1.3)

Validità di una misura

Consideriamo una misura X e un'altra misura Y (criterio). Il *coefficiente di validità* allora può essere semplicemente espresso dal valore assoluto della correlazione tra le due misure:

$$|\mathbb{C}\text{or}[X, Y]| = \frac{|\mathbb{C}\text{ov}[X, Y]|}{\sqrt{\mathbb{V}\text{ar}[X]}\sqrt{\mathbb{V}\text{ar}[Y]}}$$

L'interpretazione che ne risulta è semplice: $|\mathbb{C}\text{or}[X, Y]| \approx 0$ indica scarsa concordanza tra X e il criterio Y (X non è valida secondo Y) mentre $|\mathbb{C}\text{or}[X, Y]| \approx 1$ indica perfetta concordanza (X è valida secondo Y).

Teoria Classica dei Test

Studio individuale

Altre forme di validità:

BN(6.1.1), BN(6.1.2), BN(6.1.4), BN(6.1.5), BN(6.1.6)

Teoria Classica dei Test

FONTI: BN(6.1.3,2.3)

Dalla validità di una misura all'attendibilità

Nel caso in cui X e Y siano parallele, nel senso di (p_1), il numeratore $\mathbb{C}\text{ov}[X, Y]$ si semplifica come segue:

$$\begin{aligned}\mathbb{C}\text{ov}[X, Y] &= \mathbb{C}\text{ov}[(T + E_X), (T + E_Y)] \\ &= \mathbb{C}\text{ov}[T, T] + \cancelto{0}{\mathbb{C}\text{ov}[T, E_X]} + \cancelto{0}{\mathbb{C}\text{ov}[T, E_Y]} + \cancelto{0}{\mathbb{C}\text{ov}[E_Y, E_X]} \\ &= \mathbb{V}\text{ar}[T]\end{aligned}$$

Analogamente il denominatore

$$\sqrt{\mathbb{V}\text{ar}[X]}\sqrt{\mathbb{V}\text{ar}[Y]} = \mathbb{V}\text{ar}[X]$$

Dalla validità di una misura all'attendibilità

In questo caso, il coefficiente di validità $|\mathbb{C}\text{or}[X, Y]|$ si semplifica esprimendo l'*attendibilità* di un test o misura:

$$\rho^2_{XT} = \frac{\mathbb{V}\text{ar}[T]}{\mathbb{V}\text{ar}[X]}$$

che esprime, in altri termini, la validità di una misura rispetto ad un criterio parallelo Y.

Teoria Classica dei Test

FONTI: BN(2.3)

Attendibilità di una misura

L'attendibilità:

$$\rho^2_{XT} = \frac{\mathbb{V}\text{ar}[T]}{\mathbb{V}\text{ar}[X]} = 1 - \frac{\mathbb{V}\text{ar}[E]}{\mathbb{V}\text{ar}[X]}$$

esprime la variazione del misurando T (non osservato) rispetto al misurabile X (osservato). Se $\mathbb{V}\text{ar}[E] = 0$, $\rho^2_{XT} = 1$, viceversa se $\mathbb{V}\text{ar}[E] = \mathbb{V}\text{ar}[X]$ allora $\rho^2_{XT} = 0$.

L'attendibilità esprime informazione rispetto alla precisione/imprecisione della misura (nella definizione è coinvolta difatti la varianza): in particolare, $\mathbb{V}\text{ar}[E]$ indica l'imprecisione della misura mentre ρ^2_{XT} esprime al contrario la precisione del test. Quando $\mathbb{V}\text{ar}[E] \to 0$, la precisione della misura cresce.

L'*indice di attendibilità* è ottenuto semplicemente come $\sqrt{\rho^2_{XT}} = \frac{\sqrt{\mathbb{V}\text{ar}[T]}}{\sqrt{\mathbb{V}\text{ar}[X]}}$.

Misure composite

Finora abbiamo considerato misurazioni fatte usando più misure (parallele o non parallele) *non gerarchiche*. Ad esempio, X_j e $X_{j'}$ sebbene distinte contribuiscono allo stesso modo a determinare il valore vero T.

Ci sono casi tuttavia - e sono la maggioranza - in cui le diverse misure sono utilizzate per comporre altre misure che sono sintesi delle precedenti (vedi *misure indirette*, modulo A). In questo caso, le misure sono definite in maniera *gerarchica*.

Diremo dunque che X è una *misura composta* quando $X = \phi(Y_1, \ldots, Y_m)$, con $\phi(.)$ funzione di legame nota, e chiameremo *misure componenti* le osservabili $Y_1, \ldots, Y_m$.

Prenderemo in considerazione il caso semplice $X = Y_1 + Y_2$ in cui le misure componenti (o di livello inferiore) compongono additivamente la misura composta (o di livello superiore) X, con $m = 2$.

Teoria Classica dei Test

FONTI: BN(2.6)

Misure composite

Consideriamo due misure (Y_1, T_1, E_1) e (Y_2, T_2, E_2) e definiamo la misura composta (X, T, E) secondo la relazione:

$$X = Y_1 + Y_2$$

Ricordando le assunzioni (i)-(iii) della TCT, abbiamo quando segue:

(viii) $\mathbb{E}[X] = \mathbb{E}[Y_1] + \mathbb{E}[Y_2]$

(ix) $\mathbb{V}\text{ar}[X] = \mathbb{V}\text{ar}[Y_1] + \mathbb{V}\text{ar}[Y_2] + 2\mathbb{C}\text{ov}[Y_1, Y_2]$

(x) $\mathbb{V}\text{ar}[T] = \mathbb{V}\text{ar}[T_1] + \mathbb{V}\text{ar}[T_2] + 2\mathbb{C}\text{ov}[T_1, T_2]$

(xi) $\mathbb{V}\text{ar}[E] = \mathbb{V}\text{ar}[E_1] + \mathbb{V}\text{ar}[E_2] + \cancelto{0}{2\mathbb{C}\text{ov}[E_1, E_2]}$

Nota: Nel caso Y_1 e Y_2 siano misure parallele nel senso di (p_1), (viii)-(xi) si semplificano poiché $\mathbb{E}[Y_1] = \mathbb{E}[Y_2]$ e $\mathbb{V}\text{ar}[Y_1] = \mathbb{V}\text{ar}[Y_2]$.

Teoria Classica dei Test

TCT in sintesi (caso di un test di lunghezza doppia, $m = 2$)

$$\underbrace{X}_{\text{misura composta}} = \underbrace{T_1 + E_1}_{\text{misura componente } Y_1} + \underbrace{T_2 + E_2}_{\text{misura componente } Y_2}$$

(i) $\mathbb{E}[E_1] = 0, \quad \mathbb{E}[E_2] = 0$

(ii) $\mathbb{C}\text{or}[E_1, T_1] = 0, \quad \mathbb{C}\text{or}[E_2, T_2] = 0, \quad \mathbb{C}\text{or}[E_2, T_1] = 0, \quad \mathbb{C}\text{or}[E_1, T_2] = 0$

(iii) $\mathbb{C}\text{or}[E_1, E_2] = 0$

(viii) $\mathbb{E}[X] = \mathbb{E}[Y_1] + \mathbb{E}[Y_2]$

(ix) $\mathbb{V}\text{ar}[X] = \mathbb{V}\text{ar}[Y_1] + \mathbb{V}\text{ar}[Y_2] + 2\mathbb{C}\text{ov}[Y_1, Y_2]$

(x) $\mathbb{V}\text{ar}[T] = \mathbb{V}\text{ar}[T_1] + \mathbb{V}\text{ar}[T_2] + 2\mathbb{C}\text{ov}[T_1, T_2]$

(xi) $\mathbb{V}\text{ar}[E] = \mathbb{V}\text{ar}[E_1] + \mathbb{V}\text{ar}[E_2]$

Teoria Classica dei Test

FONTI: BN(2.1-2.4)

Attendibilità delle misure composite

Consideriamo Y_1, Y_2, Y_3, Y_4 misure componenti *parallele* e

$$X_1 = Y_1 + Y_2 \qquad X_2 = Y_3 + Y_4$$

le rispettive misure composte. Se le misure componenti sono parallele, anche le misure composte X_1 e X_2 sono parallele (dunque entrambe sottendono unico T). L'attendibilità del test a misure composte è pari a:

$$\rho^2_{XT} = \rho^2_{X_1 X_2} = \frac{2\rho^2_{Y_1 Y_2}}{1 + \rho^2_{Y_1 Y_2}}$$

dove $\rho^2_{Y_1 Y_2}$ è l'attendibilità della misura componente Y_1 (rispetto alla misura parallela Y_2).

Nota: La formula può essere scritta anche usando X_3, X_4 e Y_3, Y_4 o più in generale X, X' e Y, Y' ad indicare qualunque coppia di misure parallele (almeno due).

Teoria Classica dei Test

FONTI: BN(2.6)

Attendibilità delle misure composite

In generale, per qualsiasi coppia di misure composte parallele X, X' e misure componenti parallele Y, Y', l'attendibilità è pari a:

$$\rho^2_{XX'} = \frac{2\rho^2_{YY'}}{1 + \rho^2_{YY'}} \qquad \text{con} \quad \rho^2_{XX'} \in [0, 1]$$

che è nota anche come *formula di Sperman-Brown* per un test di lunghezza doppia (poiché abbiamo due misure composte). In particolare, vale la seguente

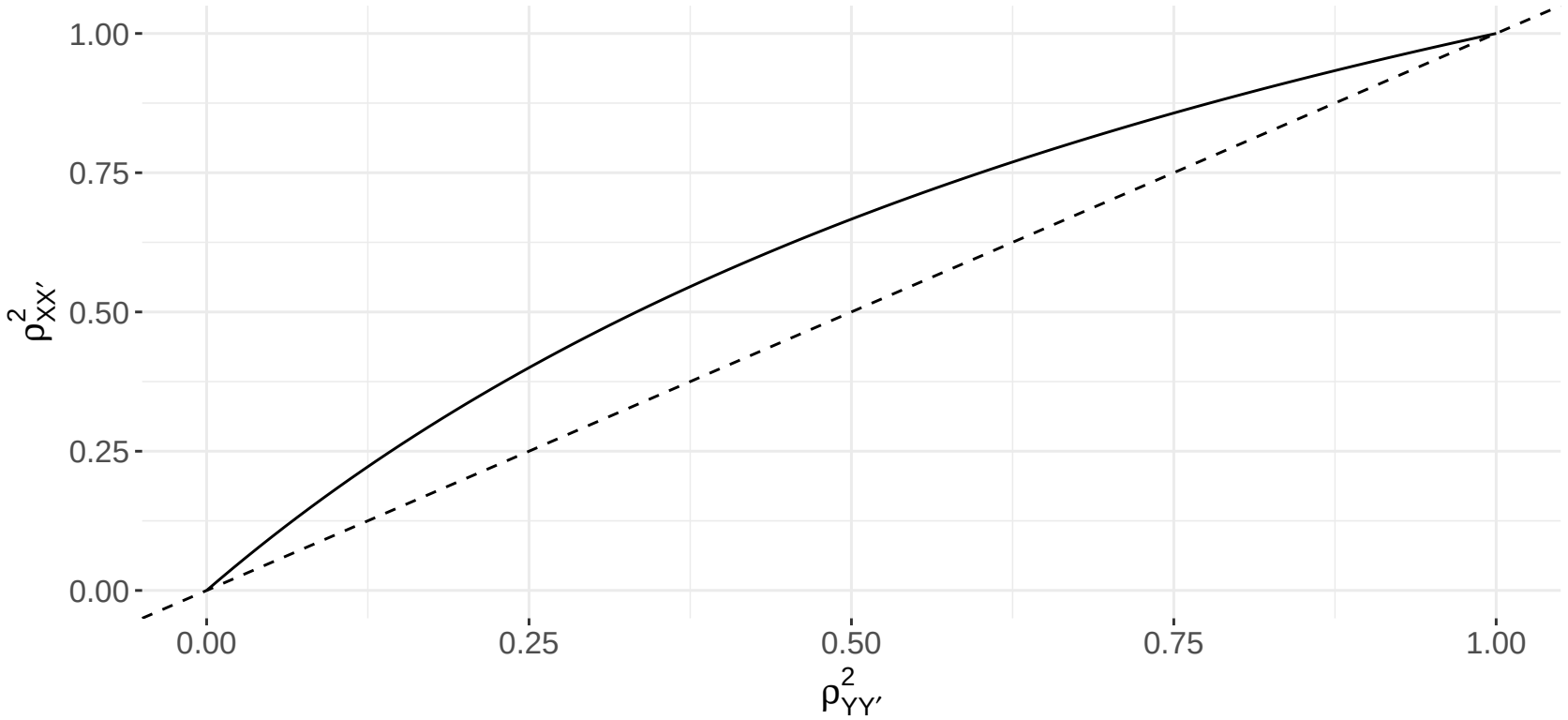
$$\rho^2_{XX'} > \rho^2_{YY'}$$

ad indicare che l'attendibilità (precisione) di un test composito (test di lunghezza doppia in questo caso) è sempre maggiore di quella delle singole componenti (test di lunghezza unitaria).

Esempio: se l'attendibilità del test di lunghezza unitario è pari a $\rho^2_{YY'} = 0.5$, l'attendibilità del test di lunghezza doppia ottenuto per composizione da Y e Y' è pari a $\rho^2_{YY'} = 0.67$.

Teoria Classica dei Test

FONTI: BN(2.6)



Attendibilità di un test composito di lunghezza doppia: Curva di attendibilità calcolata secondo la formula di Sperman-Brown per un test di lunghezza doppia (in ordinata) rispetto ad un test di lunghezza unitaria (in ascissa). L'andamento della curva di attendibilità mostra che un test a misure composte è più preciso rispetto ad un test fatto dalle singole misure componenti. La precisione $\rho^2_{XX'}$ è simile a $\rho^2_{YY'}$ per valori estremi di attendibilità.

Teoria Classica dei Test

FONTI: BN(2.6)

Misure composite e lunghezza del test

I risultati ottenuti per le misure composite di lunghezza due possono essere generalizzate quando $m > 2$.

Ci sono due risultati importanti nel caso $m > 2$ che coinvolgono (i) la varianza del valore vero della misura composta $\mathbb{V}\text{ar}[T_m]$ e (ii) la varianza della componente di errore della misura composta $\mathbb{V}\text{ar}[E_m]$.

Tali risultati supportano l'idea che all'aumentare della lunghezza di un test m, aumenti anche l'attribuzione della varianza delle misurazioni alla componente vera T piuttosto che all'errore E.

In altri termini, all'aumentare della lunghezza di un test riusciamo a separare maggiormente la varianza del costrutto latente T dalla varianza dell'errore E.

Teoria Classica dei Test

FONTI: BN(2.6)

Misure composite e lunghezza del test

Di seguito i due risultati importanti:

$$\mathbb{V}\text{ar}[T_m] = m^2\mathbb{V}\text{ar}[T] \qquad T \text{ indica il valore vero delle misure componenti parallele}$$

$$\mathbb{V}\text{ar}[E_m] = m\mathbb{V}\text{ar}[E] \qquad E \text{ indica l'errore delle misure componenti parallele}$$

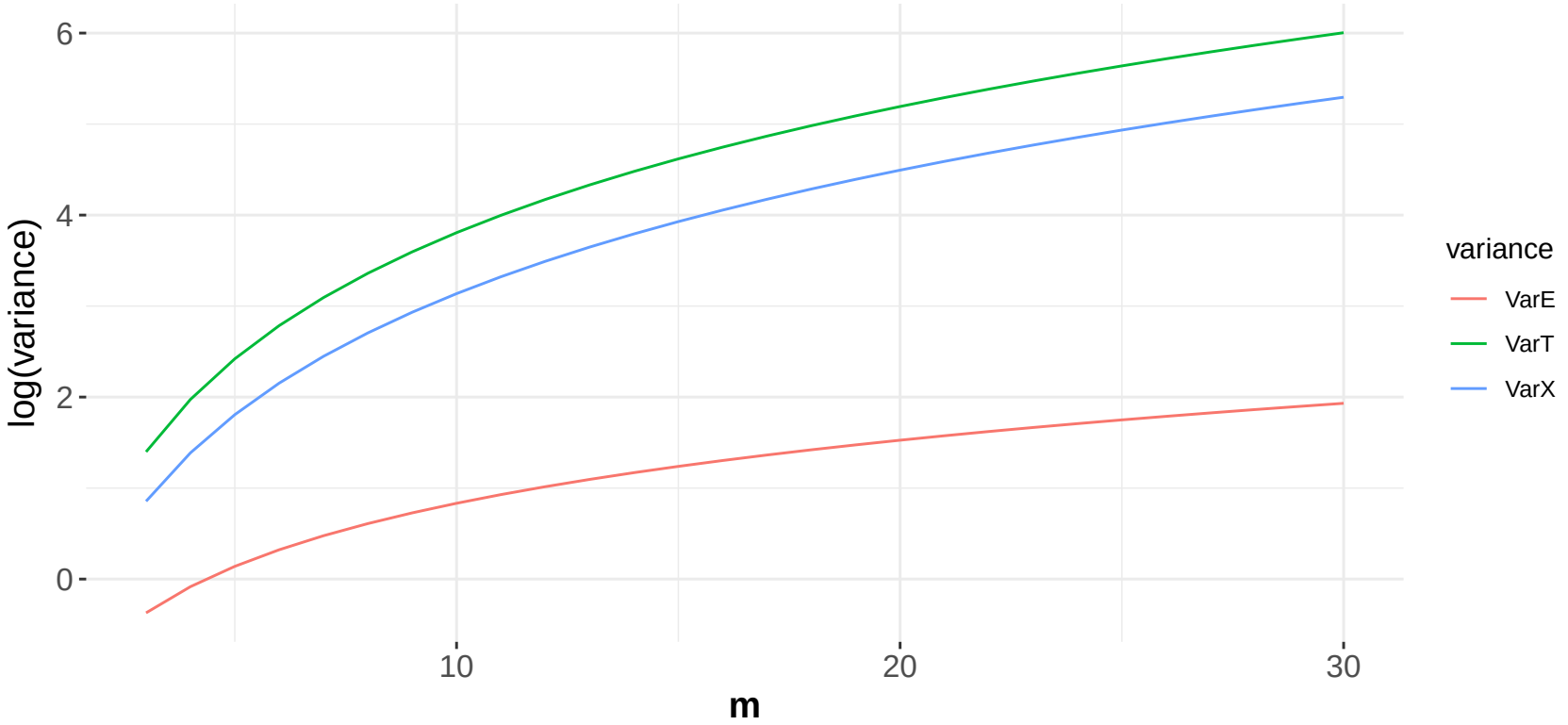
Notiamo che $\mathbb{V}\text{ar}[T_m]$ cresce in maniera quadratica al crescere di m mentre $\mathbb{V}\text{ar}[E_m]$ cresce in maniera lineare al crescere di m.

Questo implica che aumentare la misura composita (o test) con m misure parallele fa aumentare $\mathbb{V}\text{ar}[T_m]$ più velocemente di $\mathbb{V}\text{ar}[E_m]$ con il risultato di ottenere un test più preciso (attendibilità più alta).[†]

[†] Ciò è semplicemente dovuto al fatto che nella definizione di attendibilità $\rho^2_{XT} = \frac{\mathbb{V}\text{ar}[T]}{\mathbb{V}\text{ar}[X]}$ la quantità $\mathbb{V}\text{ar}[T]$ è posta al numeratore.

Teoria Classica dei Test

FONTI: BN(2.6)



Componenti di varianza per una misura composita: varianza di errore $\mathbb{V}\text{ar}\,[E_m]$ (linea rossa), varianza della misurazione $\mathbb{V}\text{ar}\,[X_m]$ (linea blu) e varianza del valore vero $\mathbb{V}\text{ar}\,[T_m]$ (linea verde) per una misura composita (X_m, T_m, E_m) di lunghezza m prefissata. Nota: in ascissa sono riportate le lunghezze m della composizione (da 3 a 30) mentre in ordinata sono riportate le tre componenti di varianza rappresentate su scala logaritmica. Il grafico evidenzia come $\mathbb{V}\text{ar}\,[T_m] > \mathbb{V}\text{ar}\,[E_m]$ per ogni intero m fissato.

Misure composite e lunghezza del test

In generale, data una misura composita iniziale X (o test) avente attendibilità $\rho^2_{XX'}$ possiamo valutare come questa vari in funzione di una data lunghezza m.

Questo può essere utile, ad esempio, quando si vuole valutare quante misure componenti (o items) occorre aggiungere per avere una desiderata attendibilità.

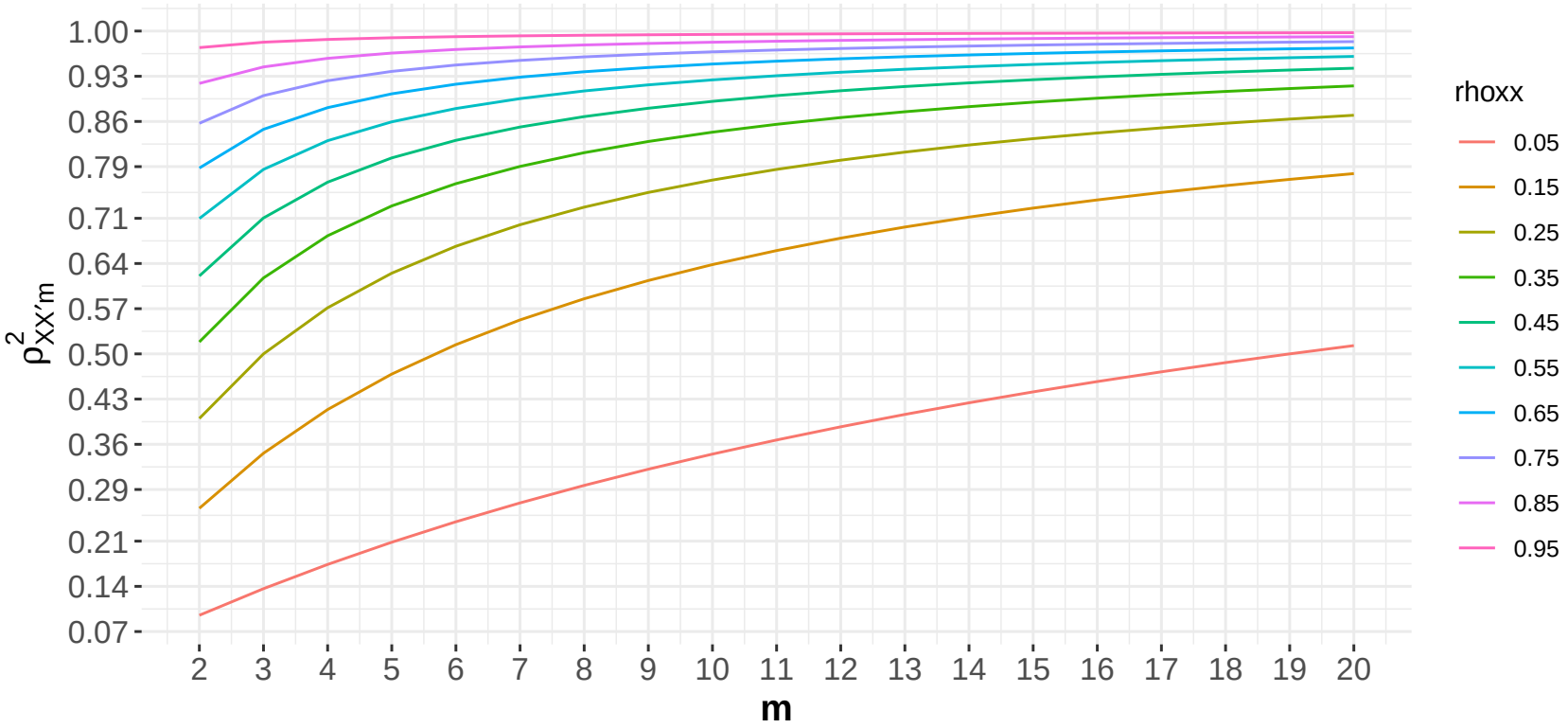
La risposta a tale quesito è la generalizzazione della formula di Sperman-Brown:

$$\rho^2_{XX'_m} = \frac{m\rho^2_{XX'}}{1 + (m-1)\rho^2_{XX'}}$$

dove $\rho^2_{XX'}$ indica l'attendibilità del test iniziale (o test di origine).

Teoria Classica dei Test

FONTI: BN(2.6)



Attendibilità di una misura composita e lunghezza: Curve di attendibilità di una misura composita (o test) quando questa è *allungata* per un intero *m*. Le curve sono in funzione dell'attendibilità iniziale $\rho^2_{XX'}$ della misura composita (colore delle curve) e mostrano come questa, tenendo fisso tale valore, cambia in precisione $\rho^2_{XX'_m}$ (ordinata) quando *m* nuove misure componenti (ascissa) sono aggiunte a formare il test.

Teoria Classica dei Test

FONTI: BN(2.5)

Stime dell'attendibilità di misure composite

Consideriamo m misure componenti $(Y_1, T_1, E_1), \ldots, (Y_m, T_m, E_m)$ che formano la misura composta $X = \sum_{j=1}^{m} Y_j$. Una stima (approssimata) dell'attendibilità della misura composta $\rho^2_{XX'}$ è la seguente:

$$\hat{\rho}^2_{XX'} = \frac{m}{m-1}\left(1 - \frac{\sum_{j=1}^{m} \mathbb{V}\text{ar}[Y_j]}{\mathbb{V}\text{ar}[X]}\right)$$

nota anche come α *di Cronbach*.

Tale formulazione fornisce un'interpretazione dell'attendibilità in termini di *coerenza interna* della misura composita. A differenza di altre tecniche per valutare l'attendibilità di un test (ad esempio, *split-half*), l'α di Cronbach utilizza solo l'informazione derivante dalle misure componenti (o items).

Teoria Classica dei Test

FONTI: BN(2.5)

Stime dell'attendibilità di misure composite

Per apprezzare meglio tale formulazione, utilizziamo il risultato (ix) della TCT generalizzato al caso di m misure componenti:

$$\mathbb{V}\text{ar}\left[X\right] = \sum_{j=1}^{m} \mathbb{V}\text{ar}\left[Y_j\right] + \sum_{j \neq h} \mathbb{C}\text{ov}\left[Y_j, Y_h\right]^{(1)}$$

L'indice α diventa dunque:

$$\hat{\rho}^2_{XX'} = \frac{m}{m-1}\left(1 - \frac{\sum_{j=1}^{m} \mathbb{V}\text{ar}\left[Y_j\right]}{\sum_{j=1}^{m} \mathbb{V}\text{ar}\left[Y_j\right] + \sum_{j \neq h} \mathbb{C}\text{ov}\left[Y_j, Y_h\right]}\right)$$

dove maggiore è la componente di covarianza tra le misure componenti, maggiore sarà l'indice α finale. Quando invece la covarianza tra le misure è prossima allo zero, l'indice assumerà anch'esso valore prossimo allo zero.

(1) La sommatoria è sulle misure componenti che sono diverse. Infatti quando $j = h$ abbiamo $\mathbb{C}\text{ov}\left[Y_j, Y_h\right] = \mathbb{V}\text{ar}\left[Y_j, Y_h\right]$ già calcolata nella sommatoria precedente. Tale formulazione permette di operare sulle covarianze delle coppie di misure realmente differenti tra loro.

Teoria Classica dei Test

FONTI: BN(2.5)

Stime dell'attendibilità di misure composite

$$\hat{\rho}^2_{XX'} = \frac{m}{m-1}\left(1 - \frac{\sum_{j=1}^{m} \mathbb{V}\text{ar}\left[Y_j\right]}{\sum_{j=1}^{m} \mathbb{V}\text{ar}\left[Y_j\right] + \sum_{j \neq h} \mathbb{C}\text{ov}\left[Y_j, Y_h\right]}\right)$$

Nota: La componente di covarianza tra le misure componenti $\mathbb{C}\text{ov}\left[Y_j, Y_h\right]$ fornisce un'indicazione sulla *coerenza* delle misure tra loro. Per tale ragione l'indice α si dice quantifichi la *coerenza interna* di una scala (o misura composita).

Teoria Classica dei Test

FONTI: BN(2.5)

Stime dell'attendibilità di misure composite

Note:

- lo stimatore α è non distorto per $\rho^2_{XX'}$ se le misure componenti sono parallele (p_1), τ-equivalenti (p_2) o essenzialmente τ-equivalenti (p_3)
- interpreta l'attendibilità di una misura composta in termini di coerenza interna rispetto alle misure componenti che la formano
- presuppone che le misure componenti *formino bene* la misura composta, nel senso dato dalla intercorrelazione tra le misure componenti
- Alcuni valori soglia per l'attendibilità:

  $\alpha > 0.9$ *ottima*
  $0.8 \leq \alpha \leq 0.9$ *buona*
  $0.7 \leq \alpha < 0.8$ *discreta*
  $0.6 \leq \alpha < 0.7$ *sufficiente*
  $\alpha < 0.6$ *insufficiente*

Teoria Classica dei Test

Studio individuale

Altre stime dell'attendibilità:

BN(2.5): Indici di Guttman, Kuder-Richardson, Rulon

Approfondimento (articolo presente nella cartella APPROFONDIMENTI):

Pastore M (2017). Tra Alfa e Omega c'è di mezzo la CFA? *Giornale Italiano di Psicologia*, 3, 761-782

Teoria Classica dei Test

FONTI: BN(2.10)

Fattori che influenzano l'attendibilità

- qualità del campione di individui su cui le misurazioni sono effettuate (rappresentatività)
- qualità delle misure componenti (o items)
- campionamento realmente casuale ed indipendenza delle osservazioni
- condizione di somministrazione del test
- lunghezza del test adeguata
- aspetti cognitivi degli individui testati: fatica, ricordo, bassa compliance

Teoria Classica dei Test

Sintesi complessiva e comparativa

Test lunghezza unitaria ($m = 1$)

$$X = T + E$$

(i) $\mathbb{E}\left[E_j\right] = 0$

(ii) $\mathbb{C}\text{or}\left[E_j, T\right] = 0$

(iii) $\mathbb{C}\text{or}\left[E_j, E_{j'}\right] = 0$

(iv) $\mathbb{E}\left[X_j\right] = \mathbb{E}\left[T\right]$

(v) $\mathbb{V}\text{ar}\left[X_j\right] = \mathbb{V}\text{ar}\left[T\right] + \mathbb{V}\text{ar}\left[E_j\right]$

(vi) $\mathbb{V}\text{ar}\left[T\right] = \mathbb{C}\text{or}\left[X_j, X_{j'}\right] \mathbb{V}\text{ar}\left[X_j\right]$

(vii) $\mathbb{V}\text{ar}\left[E_j\right] = \mathbb{V}\text{ar}\left[X_j\right]\left(1 - \mathbb{C}\text{or}\left[X_j, X_{j'}\right]\right)$

$\rho^2_{XT} = 1 - \frac{\mathbb{V}\text{ar}[E]}{\mathbb{V}\text{ar}[X]}$ (attendibilità)

Test lunghezza doppia ($m = 2$)

$$\underbrace{X}_{\text{misura composta}} = \underbrace{(Y_1 + Y_2)}_{\text{misure componenti}} = \underbrace{(T_1 + E_1)}_{Y_1} + \underbrace{(T_2 + E_2)}_{Y_2}$$

(i) $\mathbb{E}\left[E_j\right] = 0$

(ii) $\mathbb{C}\text{or}\left[E_j, T_j\right] = 0$

(iii) $\mathbb{C}\text{or}\left[E_j, E_{j'}\right] = 0$

(viii) $\mathbb{E}\left[X\right] = \mathbb{E}\left[Y_1\right] + \mathbb{E}\left[Y_2\right]$

(ix) $\mathbb{V}\text{ar}\left[X\right] = \mathbb{V}\text{ar}\left[Y_1\right] + \mathbb{V}\text{ar}\left[Y_2\right] + 2\mathbb{C}\text{ov}\left[Y_1, Y_2\right]$

(x) $\mathbb{V}\text{ar}\left[T\right] = \mathbb{V}\text{ar}\left[T_1\right] + \mathbb{V}\text{ar}\left[T_2\right] + 2\mathbb{C}\text{ov}\left[T_1, T_2\right]$

(xi) $\mathbb{V}\text{ar}\left[E\right] = \mathbb{V}\text{ar}\left[E_1\right] + \mathbb{V}\text{ar}\left[E_2\right]$

$\rho^2_{XX} = \frac{2\rho^2_{YY'}}{1+\rho^2_{YY'}}$ (attendibilità)

$\rho^2_{YY'}$ indica l'attendibilità della misura componente

Teoria Classica dei Test

FONTI: BN(2.9)

Esempio di applicazione della TCT
Lord & Novick (1968), pp.156-157

Immaginiamo di aver misurato una certa quantità di interesse su un campione di $n = 10$ individui utilizzando un test con $p = 1$ item (indicatori). Ogni individuo è stato sottoposto a $R = 2$ misurazioni (repliche) *parallele*.

I dati sono riportati come di seguito:

	1	2	3	4	5	6	7	8	9	10
$r = 1$	125	119	109	104	101	98	97	94	90	81
$r = 2$	120	122	107	108	98	106	96	99	93	87

Esempio di applicazione della TCT

Lord & Novick (1968), pp.156-157

Usando i dati osservati, ci interessa determinare le quantità ignote:

$$\mathbb{E}[T] \quad \mathbb{V}\text{ar}[E] \quad \mathbb{V}\text{ar}[T] \quad \rho^2_{XT}$$

Avendo $R = 2$ repliche parallele per ognuno dei $n = 10$ soggetti, possiamo determinare le quantità di interesse anche a livello individuale.

Per ottenere le quantità di interesse possiamo usare le formule presentate finora (quando abbiamo a disposizioni grandi campioni costituiti da misure parallele). In caso contrario, occorre correggere le stime per ottenere stime non distorte delle quantità vere di popolazione.

Teoria Classica dei Test

FONTI: BN(2.9)

Esempio di applicazione della TCT

Lord & Novick (1968), pp.156-157

Per stimare $\mathbb{V}\text{ar}[E]$ procediamo come segue:

$$\hat{\sigma}^2_E = \frac{1}{n}\sum_{i=1}^{n}\hat{\sigma}^2_{E_i}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{R-1}\sum_{r=1}^{R}(x_{ir}-\overline{x}_i)^2\right)$$

dove:

- $\hat{\sigma}^2_E$ è la stima della *varianza di gruppo*
- $\hat{\sigma}^2_{E_i}$ è la stima della *varianza individuale*
- $\overline{x}_i$ è la misurazione media per ciascun individuo

Nell'esempio: $\hat{\sigma}^2_E = 9.9$

Teoria Classica dei Test

FONTI: BN(2.9)

Esempio di applicazione della TCT
Lord & Novick (1968), pp.156-157

Per stimare $\mathbb{V}\text{ar}[T]$ procediamo come segue:

$$\hat{\sigma}^2_T = \frac{1}{nR^2}\left(\sum_{i=1}^{n}(a_i)^2 - \frac{\sum_{i=1}^{n} a_i}{n}\right) - \frac{1}{R}\hat{\sigma}^2_E$$

dove:

- $\hat{\sigma}^2_E$ è la stima della *varianza dell'errore*
- $a_i = \sum_{r=1}^{R} y_{ir}$ è la *somma delle repliche* per ciascun individuo

Nell'esempio: $\hat{\sigma}^2_T = 140.67$

Nota: $\hat{\sigma}^2_T \approx 0$ indica che le differenze osservate tra individui possono essere dovute semplicemente al campionamento casuale (gli individui sono abbastanza omogenei rispetto al misurando, ossia al tratto latente).

Esempio di applicazione della TCT

Lord & Novick (1968), pp.156-157

Per stimare ρ^2_{XT} procediamo utilizzando le quantità precedentemente ottenute:

$$\hat{\rho}^2_{XT} = \frac{\hat{\sigma}^2_T}{\hat{\sigma}^2_T + \hat{\sigma}^2_E}$$

dove l'attendibilità è stimata rapportando la quantità di varianza della componente vera alla varianza complessiva, dovuta cioè alla varianza della componente d'errore e della componente vera stessa (vedi: *coefficiente di correlazione intraclasse*).

Nell'esempio: $\rho^2_{XT} = 0.934$

Teoria Classica dei Test

FONTI: BN(2.9)

Esempio di applicazione della TCT

Lord & Novick (1968), pp.156-157

Infine, per stimare i punteggi veri $\mathbb{E}[T]$ ottenuti dagli $n = 10$ individui utilizziamo lo stimatore lineare:

$$\tau_i = \rho^2_{XT} x_i + (1 - \rho^2_{XT})\mu_x$$

dove $\mu_x = \frac{1}{NR} \sum_{i=1}^{n} \sum_{r=1}^{R} y_{ir}$ è la media complessiva delle misurazioni effettuate (stima della media dei punteggi veri nella popolazione).

Nota: lo stimatore utilizzato per τ potrebbe non essere adeguato quando la relazione tra X e T è *non lineare*. Inoltre, quando la precisione del test ρ^2_{XT} è bassa, si attribuisce maggior peso alla componente di popolazione μ_x.

Nell'esempio:

1	2	3	4	5	6	7	8	9	10
121.20	119.33	107.65	105.78	99.71	102.05	96.91	96.91	92.24	85.23

Teoria Classica dei Test

Stima delle quantità della TCT in sintesi (solo per misure parallele)

$\tau_i = \rho^2_{XT} x_i + (1 - \rho^2_{XT})\mu_x$ stima del punteggio vero $\mathbb{E}[T_i]$

$\hat{\rho}^2_{XT} = \frac{\hat{\sigma}^2_T}{\hat{\sigma}^2_T + \hat{\sigma}^2_E}$ stima dell'attendibilità ρ^2_{XT}

$\hat{\sigma}^2_T = \frac{1}{nR^2}\left(\sum_{i=1}^n (a_i)^2 - \frac{\sum_{i=1}^n a_i}{n}\right) - \frac{1}{R}\hat{\sigma}^2_E$ stima della varianza del punteggio vero $\mathbb{V}\text{ar}[T]$

$\hat{\sigma}^2_E = \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{R-1}\sum_{r=1}^R (x_{ir} - \overline{x}_i)^2\right)$ stima della varianza della componente di errore $\mathbb{V}\text{ar}[E]$

Note:

$\overline{x}_i$ è la misurazione media per ciascun individuo

$a_i = \sum_{r=1}^R y_{ir}$ è la *somma delle repliche* per ciascun individuo

Durante il laboratorio 5 vedremo come ottenere le quantità della TCT utilizzando R.

Teoria Classica dei Test

FONTI: BN(2.11)

Selezione delle misure osservabili

Un test è formato dall'aggregazione di scale che raggruppano a loro volta un insieme *coerente* di misure osservabili (indicatori). Un test formato da buoni items si caratterizza per un alto livello di precisione (attendibilità) nel quantificare il misurando. Al contrario, test con items scarsi presentano bassi livelli di precisione (attendibilità).

$$\text{bassa precisione/attendibilità} \implies \mathbb{V}\text{ar}[E] > \mathbb{V}\text{ar}[T]$$

$$\text{alta precisione/attendibilità} \implies \mathbb{V}\text{ar}[T] > \mathbb{V}\text{ar}[E]$$

Per migliorare la precisione di un test è dunque necessario che le osservabili che lo compongono siano *buone* (la loro inclusione aumenta la precisione test). Il processo di selezione delle osservabili è detto **analisi degli items**.

Considereremo due criteri per la selezione degli items:

- *difficoltà*
- *capacità discriminativa*

Teoria Classica dei Test

FONTI: BN(2.11)

Selezione delle misure osservabili

- *difficoltà*: esprime quanto un item/indicatore risulta di difficile utilizzo da parte dei soggetti sottoposti al test
- *capacità discriminativa*: indica la capacità di un item/indicatore di discriminare tra soggetti che presentano bassi valori del misurando e soggetti che invece presentano alti valori del misurando (potere discriminante dell'item)

Teoria Classica dei Test

FONTI: BN(2.11.1)

Selezione delle misure osservabili dicotomiche

Una misura osservabile X_j è dicotomica quando si esprime con due sole categorie non ordinate, ad esempio $X_j \in \{0, 1\}$. Items di questa categoria hanno modalità di risposta del tipo sì/no, vero/falso, presente/assente.

Si assume che l'item dicotomico segua in distribuzione la legge di Bernoulli, $X_j \sim \text{Bern}(\pi)$, con $\pi \in (0, 1)$ proporzione di casi che rispondono correttamente all'item.

Dalla teoria della probabilità ricordiamo che

$$\mathbb{E}[X_j] = \pi \qquad \mathbb{V}\text{ar}[X_j] = \pi(1 - \pi)$$

dove, per una misurazione con n prove indipendenti, il parametro è stimato come segue:

$$\hat{\pi} = \frac{\sum_{i=1}^{n} X_{ji}}{n}$$

Teoria Classica dei Test

FONTI: BN(2.11.1)

Selezione delle misure osservabili dicotomiche

Difficoltà

$$h_j = \hat{\pi}_j = \frac{\sum_{i=1}^{n} X_{ji}}{n} \qquad \text{difficoltà media dell'item } X_j$$

$$h_{\text{tot}} = \frac{\sum_{j=1}^{p} \hat{\pi}_j}{p} \qquad \text{difficoltà media della scala } X$$

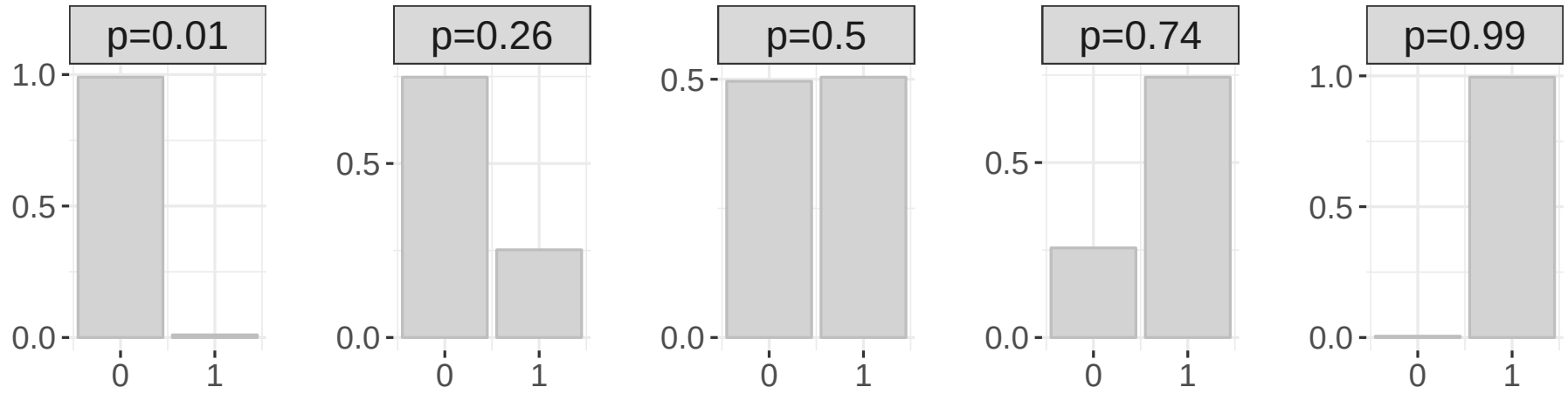
Nel caso di item a risposta dicotomica (vero/falso) con *formato a scelta multipla*, l'indice h_j deve essere corretto per il fattore di **guessing** (indovinare la risposta a caso):

$$h_j = \hat{\pi}_j = \frac{(\sum_{i=1}^{n} X_{ji}) - g_j}{n}$$

dove $g_j = \frac{n - (\sum_{i=1}^{n} X_{ji})}{K-1}$

K indica il numero di alternative di risposta

Teoria Classica dei Test



h	h_cor	g
0.99	0.98	10

h	h_cor	g
0.748	0.496	252

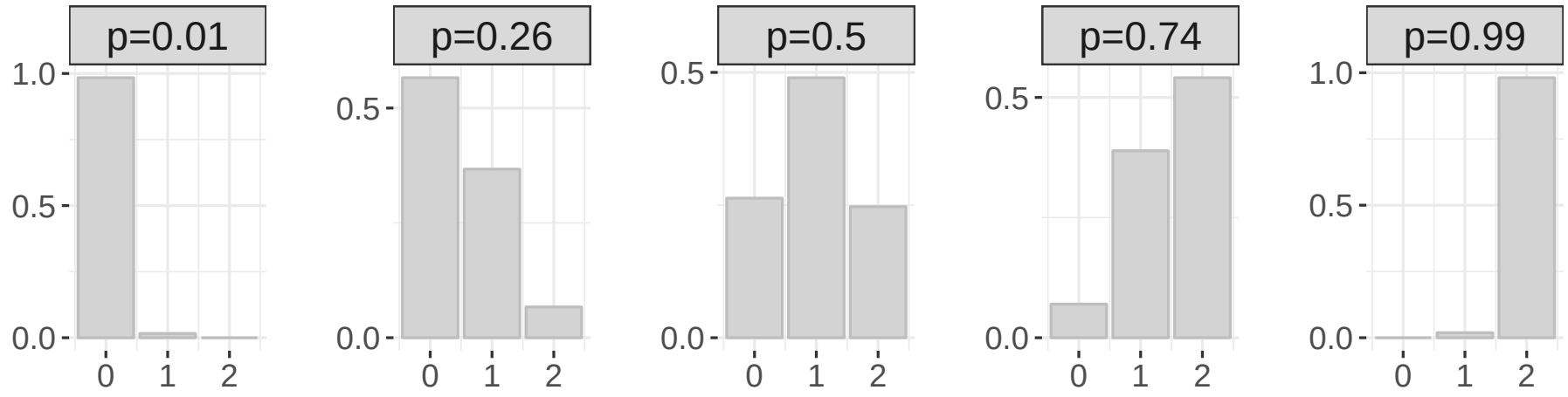
h	h_cor	g
0.496	0	504

h	h_cor	g
0.256	0	744

h	h_cor	g
0.005	0	995

Difficoltà di un item dicotomico: Distribuzioni di massa di probabilità per un item dicotomico e calcolo degli indici h_j e h_j corretto (`h_cor`) per il fattore g. Nell'esempio, la risposta corretta è $X = 0$, $n = 1000$ e $K = 2$. Notiamo come all'aumentare di π (`p`) diminuisce la proporzione di risposte corrette e aumenta la difficoltà (h_j diminuisce progressivamente). Si nota come h_j corretto diminuisce più velocemente di h_j semplice e già per $\pi = 0.5$ indica che l'item è molto difficile (`h_cor=0`).

Teoria Classica dei Test



h	h_cor	g
0.984	0.976	8

h	h_cor	g
0.566	0.349	217

h	h_cor	g
0.263	0	368.5

h	h_cor	g
0.07	0	465

h	h_cor	g
0	0	500

Difficoltà di un item dicotomico con più alternative: Distribuzioni di massa di probabilità per un item dicotomico a scelta multipla e calcolo degli indici h_j e h_j corretto (`h_cor`) per il fattore g. Nell'esempio, la risposta corretta è $X = 0$, $n = 1000$ e $K = 3$. Notiamo come all'aumentare di π (`p`) diminuisce la proporzione di risposte corrette e aumenta la difficoltà (h_j diminuisce progressivamente). Si nota come h_j corretto diminuisce più velocemente di h_j semplice e già per $\pi = 0.5$ indica che l'item è molto difficile (`h_cor=0`).

Nota: all'aumentare del numero di alternative di risposta l'uso della legge Binomiale perde di significato e X_j può essere rappresentato usando la legge Normale. Per tale ragione, l'uso degli indici di difficoltà e capacità discriminativa possono essere scelti tra quelli proposti per gli item politomici.

Teoria Classica dei Test

FONTI: BN(2.11.1)

Selezione delle misure osservabili dicotomiche

Capacità discriminativa

$$d_j = \frac{\sum_{i=1}^{n_a} x_{ji}^a}{n_a} - \frac{\sum_{i=1}^{n_b} x_{ji}^b}{n_b}$$

dove:

$\mathbf{x}_j^a = \{x_{ji} : F(x_{ji}) \leq f_0\}$ e $\mathbf{x}_j^b = \{x_{ji} : F(x_{ji}) \geq f_1\}$

$F(x_{ji})$ indica la funzione cumulata di x_{ji}, f_0 e f_1 indicano dei percentili (di solito $f_0 = 0.25$ e $f_1 = 0.75$) mentre n_a e n_b sono le numerosità corrispondenti a $\mathbf{x}_j^a$ e $\mathbf{x}_j^b$

In pratica, usando i percentili si divide la variabile osservata nei gruppi *a* e *b*, si calcolano le proporzioni di risposte corrette nei due gruppi e si sottraggono le quantità risultanti.

Teoria Classica dei Test

FONTI: BN(2.11.1)

Selezione delle misure osservabili dicotomiche

Capacità discriminativa

$$d_j = \frac{\sum_{i=1}^{n_a} x_{ji}^a}{n_a} - \frac{\sum_{i=1}^{n_b} x_{ji}^b}{n_b}$$

$d_j \in [-1, 1]$, in particolare (Ebel, 1965)

$d_j \geq 0.30$: l'item discrimina bene

$0.20 \leq d_j < 0.3$: l'item discrimina in modo sufficiente, richieste revisioni parziali

$d_j \leq 0.20$: l'item non discrimina per niente, richieste revisioni totali o l'eliminazione

Teoria Classica dei Test

FONTI: BN(2.11.2)

Selezione delle misure osservabili politomiche

Una misura osservabile X_j è politomica quando si esprime con più di due categorie ordinate $X_j \in (0, 1, 2, \ldots, K)$. Esempio notevole di tale tipologia di item è la scala Likert.

In questo contesto l'item segue in distribuzione la legge Multinomiale (generalizzazione della legge Binomiale), $X_j \sim \text{Multinom}(\pi_1, \ldots, \pi_K)$, con $\pi_k \in (0, 1)$ proporzioni di casi che rispondono alle K categorie ordinate sotto il vincolo $\sum_{k=1}^{K} \pi_k = 1$.

Nella pratica dell'analisi degli items, tuttavia, si utilizza l'approssimazione $X_j \sim \text{F}(\theta)$ con F generica distribuzione di probabilità *simmetrica* e *centrata*. Solitamente, $\text{F} = \text{N}(\mu, \sigma^2)$ soprattutto per K grande.

Nota: tale approssimazione è in molti casi non ottimale e può distorcere i risultati dell'analisi.

Teoria Classica dei Test

FONTI: BN(2.11.2)

Selezione delle misure osservabili politomiche

In virtù dell'approssimazione prima richiamata, l'analisi degli items politomici viene sovente effettuata utilizzando rappresentazioni distribuzionali degli items (es.: istogramma), insieme agli indici di tendenza centrale (media, mediana), forma (simmetria, curtosi) e dispersione.

Si considerano dunque buoni items quelli che presentano una distribuzione simmetrica e centrata sul valore teorico della scala politomica, con bassa varianza, media e mediana vicine.

Esistono tuttavia tecniche più sofisticate per analizzare osservabili politomiche e/o continue, ad esempio mediante l'utilizzo dell'analisi fattoriale o della decomposizione della matrice di covarianza delle osservabili. Vedremo tali tecniche nel modulo [D].

Teoria Classica dei Test

FONTI: BN(2.11.2)

Selezione delle misure osservabili politomiche

Difficoltà

$$h_j = |\text{median}(\mathbf{x}_j) - x_j^{\dagger}|$$

dove $x_j^{\dagger}$ è la mediana teorica della scala (esempio, se $K = 7$, $x_j^{\dagger} = 4$)

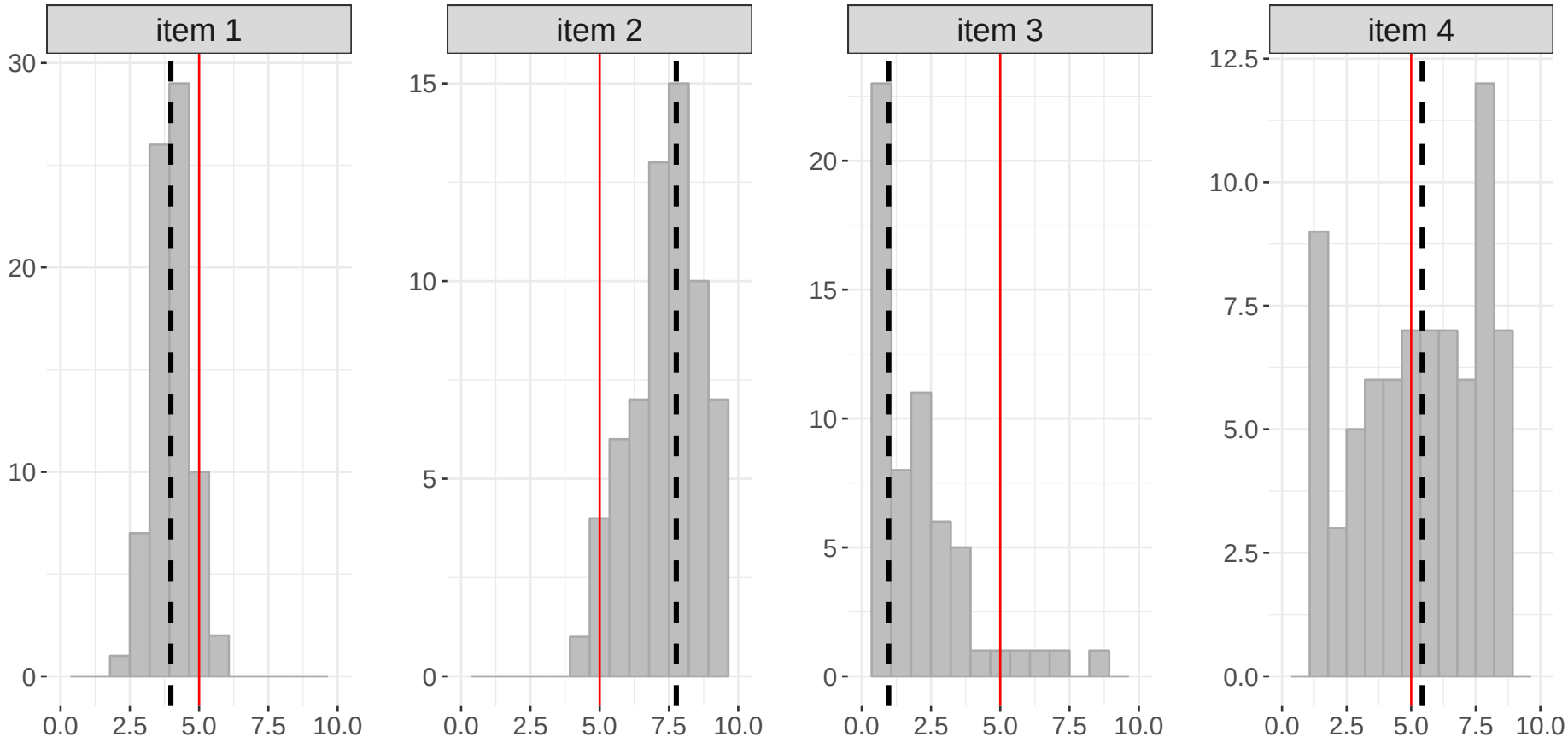
L'item è considerato difficile se $h_j \geq 1.5$.

Capacità discriminativa

Si utilizza una versione dell'indice d_j approssimata che si ottiene applicando una riduzione di scala (da politomica a dicotomica) all'item. In alternativa, si possono utilizzare indici di effect size (ad esempio, d di Cohen) per quantificare la discriminazione tra i due gruppi a e b.

Vedremo tali procedure durante il laboratorio 6.

Teoria Classica dei Test



item	min	sd	median	max	skew	kurtosis
1	2.38	0.65	3.98	5.52	0.05	−0.26
2	4.51	1.65	7.76	11.72	0.25	−0.45
3	0.02	1.75	0.97	8.5	1.62	2.67
4	1.13	2.35	5.43	8.82	−0.24	−1.21

Teoria Classica dei Test

FONTI: BN(2.12.1-2.12.2)

Calcolo dei punteggi di un test

La costruzione dei punteggi di un test (le misurazioni/quantificazioni finali ottenute per ciascun individuo sottoposto al test) avviene utilizzando le misure componenti di una scala. Date $Y_1, \ldots, Y_p$ misure componenti, il *punteggio totale* di una scala (o misura composita) è solitamente calcolato come combinazione lineare delle misure componenti.

Esempi includono,

- somma: $X = \sum_{j=1}^{p} Y_j$
- media: $X = \frac{1}{p} \sum_{j=1}^{p} Y_j$
- media pesata: $X = \frac{1}{p} \sum_{j=1}^{p} Y_j w_j$ (può essere utile richiedere che: $w_j \in (0, 1)$ e $\sum_{j=1}^{p} w_j = 1$)

Teoria Classica dei Test

FONTI: BN(2.12.1-2.12.2)

Calcolo dei punteggi di un test

Quando una scala è somministrata ad un campione di n individui, le misure componenti $\mathbf{y}_{i1}, \ldots, \mathbf{y}_{ip}$ rilevate sul campione sono utilizzate per formare il *punteggio grezzo* $\mathbf{x}$ (ottenuto ad esempio per somma, media, media pesata).

Il processo di calcolo del punteggio grezzo individuale usando gli items di un test è detto *scoring*.

L'interpretazione del composito $\mathbf{x}$ (rilevato sul campione) viene effettuata per confronto con i *valori normativi* del test, ottenuti questi ultimi durante la fase di taratura del test.

I valori normativi esprimono le caratteristiche di X a livello di popolazione e sono quantificate utilizzando i momenti di X (solitamente *media* e *varianza*).

Teoria Classica dei Test

FONTI: BN(2.12.1-2.12.2)

Calcolo dei punteggi di un test

Per valutare/confrontare i punteggi ottenuti dagli individui ad una scala o tra più scale è opportuno trasformare i punteggi grezzi in *punteggi standardizzati*. Questi possono essere impiegati altresì per la costruzione dei *profili individuali*.

Diversi sono i modi per calcolare i punteggi standardizzati, ad esempio mediante:

(a) rango percentile

(b) punteggio z

(c) punteggio T

Teoria Classica dei Test

FONTI: BN(2.12.1-2.12.2)

Calcolo dei punteggi di un test

Rango percentile

$$r_x = \frac{F_i - 0.5n_i}{n} \cdot 100$$

dove:

$F_i = F(X \leq x_i)$ frequenza cumulata

n_i frequenza assoluta di x_i nel campione di n unità

Il valore r_x rappresenta la percentuale di soggetti che hanno ottenuto un punteggio grezzo minore o uguale al punteggio x_i ed esprime dunque un dato aggregato.

Teoria Classica dei Test

FONTI: BN(2.12.1-2.12.2)

Calcolo dei punteggi di un test

Punteggio z

$$z_i = \frac{x_i - \overline{x}}{s_x}$$

dove:

$\overline{x}$ è la media del campione

s_x è lo scarto quadratico medio (standard deviation) del campione

I punteggi z hanno media nulla e varianza unitaria poiché utilizzano la Normale standardizzata come legge di riferimento.

I punteggi z si esprimono su una scala centrata sullo zero e con unità di misura lo scarto quadratico medio: le distanze tra punteggi infatti sono espresse in termini di $\pm 1, 2, \ldots, K$ deviazioni standard.

Teoria Classica dei Test

FONTI: BN(2.12.1-2.12.2)

Calcolo dei punteggi di un test

Punteggio t

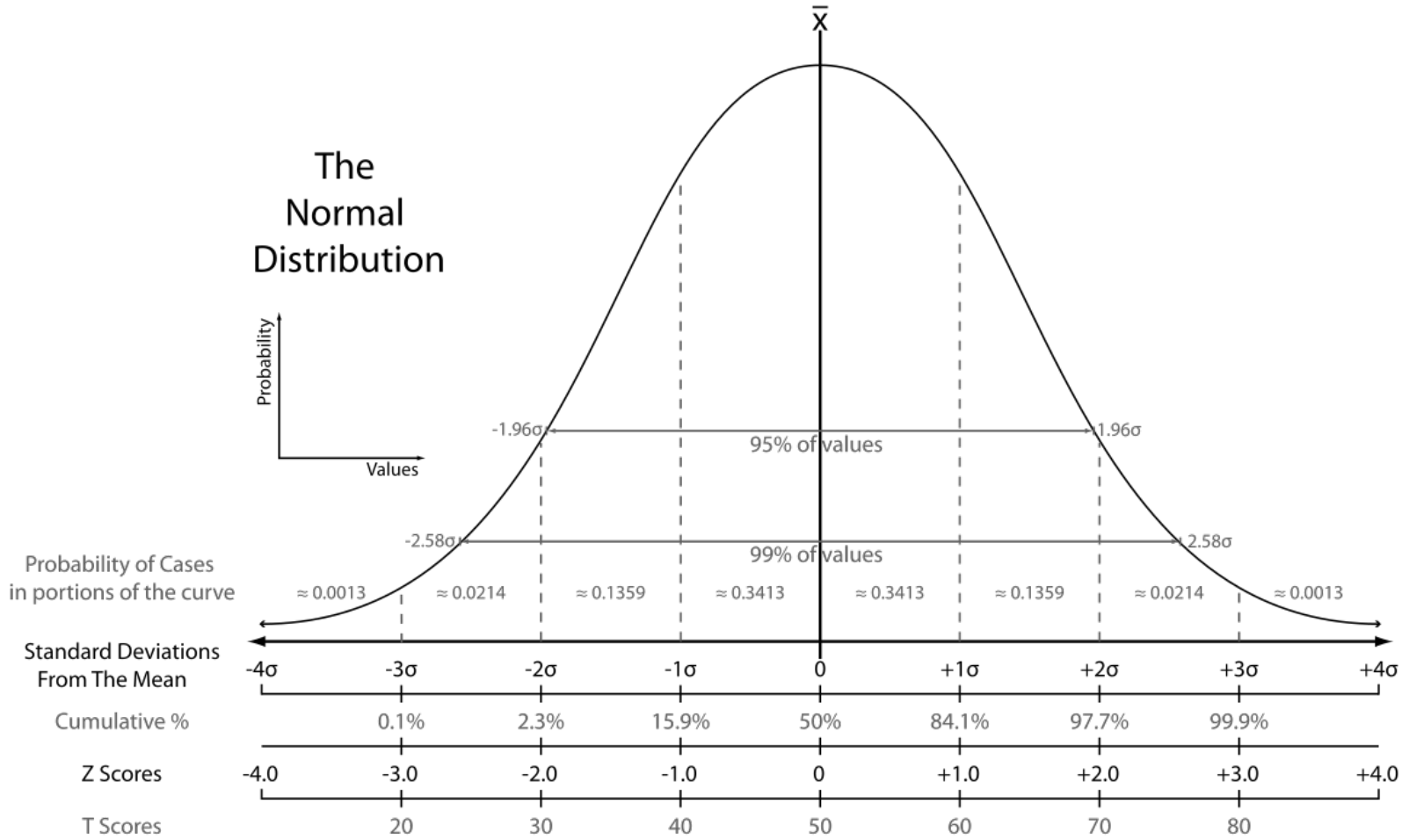
Si ottiene per trasformazione lineare dei punteggi z:

$$t_i = 50 + 10z_i$$

I punteggi t hanno media pari a 50 e deviazione standard pari a 10 e vengono utilizzati preferibilmente per rendere i punteggi finali tutti positivi (i punteggi z infatti possono assumere valori negativi).

Teoria Classica dei Test

FONTI: https://en.wikipedia.org/wiki/Standard_score



Teoria Classica dei Test

Studio individuale

BN(2.12.3): Taratura di un test

Teoria Classica dei Test

Alcune note di sintesi

- La TCT decompone la variabilità della misurazione X in una componente di errore E (accidentale) ed una componente attribuita al misurando T (latente)
- Obiettivo della TCT è di costruire un test di misura che abbia massima precisione/attendibilità, ossia quando $\mathbb{V}\text{ar}[T] > \mathbb{V}\text{ar}[E]$
- L'attendibilità viene costruita utilizzando misure parallele coerenti tra loro
- Le componenti di variabilità individuale (differenze individuali) non vengono modellate se non attraverso il campione/gruppo
- La TCT si focalizza maggiormente sulle caratteristiche del test più che su quelle degli items