Testing Psicologico

Anno Accademico 2024/2025

Corrado Caudek

2024-11-11

Indice

Benvenuti

Questo sito web è dedicato al materiale didattico dell'insegnamento di Testing Psicologico (A.A. 2024/2025), rivolto agli studenti del primo anno del Corso di Laurea Magistrale Psicologia Clinica e della Salute e Neuropsicologia dell'Università degli Studi di Firenze.

L'insegnamento si propone quale stimolo e guida per l'apprendimento delle basi dell'assessment psicologico.

Informazioni sull'insegnamento

- **Codice**: B033288 - Testing Psicologico
- **Modulo**: B033288 - Testing Psicologico (Cognomi L-Z)
- **Corso di laurea**: Laurea Magistrale: Psicologia Clinica e della Salute e Neuropsicologia
- **Anno Accademico**: 2024-2025
- **Calendario**: Il corso si terrà dal 4 marzo al 31 maggio 2025.
- **Orario delle lezioni**: Le lezioni si svolgeranno il martedì dalle 10:30 alle 13:30 e il giovedì dalle 8:30 alle 11:30.
- **Luogo**: Le lezioni si terranno presso il Plesso didattico La Torretta.
- **Modalità di svolgimento della didattica**: Le lezioni ed esercitazioni saranno svolte in modalità frontale.

> **Nota**
>
> Questo sito web è la fonte ufficiale per tutte le informazioni relative al programma dell'insegnamento *B033288 - Testing Psicologico* (Cognomi A-K) per l'A.A. 2024-2025 e le modalità d'esame.

Syllabus

Il Syllabus può essere scaricato utilizzando questo link.

Prefazione

Gli obiettivi di questo insegnamento sono:

- presentare i principi metodologici su cui i test psicologici sono fondati;
- mettere gli studenti in condizione di discriminare le diverse tipologie di test e gli obiettivi per cui essi vengono utilizzati;
- introdurre le tematiche dell'assessment psicologico;
- presentare la teoria classica dei test, il metodo dell'analisi fattoriale, i modelli di equazioni strutturali e i modelli IRT.

Viene presentata qui una panoramica degli argomenti che verranno trattati.

Definizione di misurazione

La misurazione psicologica è un pilastro fondamentale nella comprensione e nell'analisi del comportamento umano, fornendo un mezzo quantitativo per esplorare le dinamiche della mente e della personalità. La definizione di misurazione proposta da Stevens (1951), uno dei pionieri della teoria della misurazione, stabilisce che essa consiste nell'assegnare numeri a oggetti o eventi secondo regole definite. Tuttavia, è ormai ampiamente accettato che questa visione sia troppo semplicistica e che la misurazione richieda un approccio più sofisticato. Si concorda comunemente sul fatto che la misurazione debba essere considerata come un processo di creazione di modelli che rappresentano i fenomeni di interesse, principalmente in forma quantitativa.

Di conseguenza, la misurazione si basa su regole che attribuiscono scale o valori alle entità che rappresentano i costrutti di interesse. Come avviene per tutti i modelli, quelli di misurazione, come i test, le scale o le variabili, devono semplificare la realtà per risultare utili. Pertanto, è fondamentale specificare chiaramente i modelli di misurazione per poterli valutare, confutare e migliorare.

Inoltre, anziché chiedersi se un modello sia vero o corretto, è più utile sviluppare diversi modelli alternativi plausibili e porre domande del tipo: quale modello è meno inaccurato? Questo approccio al confronto dei modelli rappresenta la strategia migliore per valutare e perfezionare le procedure di misurazione, consentendo un'analisi più approfondita e accurata delle variabili coinvolte.

Per illustrare l'approccio alla misurazione come descritto, prendiamo in considerazione un esempio concreto: la valutazione dell'intelligenza attraverso il test del quoziente intellettivo (QI).

Iniziamo definendo il concetto di interesse, ovvero l'intelligenza, che può essere concepita come la capacità di apprendere, comprendere e applicare conoscenze, risolvere problemi e adattarsi a nuove situazioni. Tuttavia, trattandosi di un concetto astratto, è necessario operazionalizzarlo in modo misurabile.

Per misurare l'intelligenza, si crea un test di QI che comprende una serie di compiti e domande progettati per valutare diverse dimensioni della capacità cognitiva, quali la memoria, il ragionamento logico e la comprensione verbale.

Ciascun compito nel test di QI è associato a un punteggio. I risultati individuali vengono quindi calcolati e confrontati con una norma statistica per attribuire un punteggio di QI.

Successivamente, il test di QI viene sottoposto a diverse analisi per verificare la sua validità (ovvero se misura effettivamente l'intelligenza) e affidabilità (se fornisce risultati consistenti nel tempo).

Tuttavia, esistono diverse teorie dell'intelligenza, come ad esempio quella delle intelligenze multiple di Gardner, che suggeriscono modelli alternativi di misurazione. Confrontando il modello del QI con questi approcci alternativi, gli psicologi possono valutare quale modello è meno distorto o più adatto per specifici scopi.

In risposta alle critiche, alle nuove scoperte e ai cambiamenti culturali e sociali, il modello del QI viene regolarmente rivisto e adattato per assicurare che continui a essere uno strumento utile di misurazione.

Questo esempio mostra come la misurazione in psicologia non sia semplicemente un atto di assegnare numeri a un costrutto, ma piuttosto un processo complesso che implica la creazione, la valutazione e il continuo perfezionamento di modelli teorici.

Temi Centrali nell'Approccio Psicometrico

1. **Affidabilità**: Questo concetto si riferisce alla capacità di un test di produrre risultati consistenti nel tempo e in contesti diversi, costituendo una base fondamentale per la misurazione psicologica.
2. **Validazione del Costrutto e Test dei Modelli**: L'evoluzione della psicometria ha portato a una sempre maggiore enfasi sulla validazione dei costrutti e sull'importanza dei test di modelli, utilizzando tecniche come i modelli a equazioni strutturali (SEM) per verificare la coerenza e la validità dei costrutti psicologici.
3. **Dimensionalità e Validità Strutturale**: La dimensionalità viene considerata un elemento fondamentale nella valutazione della validità strutturale, poiché permette di esplorare come i diversi aspetti di un costrutto si manifestano e interagiscono all'interno del modello di misurazione.
4. **Costruzione dei Questionari**: La progettazione e la formulazione degli item dei questionari rivestono un ruolo cruciale, in quanto influenzano direttamente l'affidabilità e la validità dei risultati ottenuti. La scelta degli item, il loro ordine e la chiarezza della formulazione sono tutti aspetti che contribuiscono alla qualità e all'efficacia della misurazione psicologica.

Attraverso questi approcci, la misurazione psicologica si adatta alle sfide uniche poste dalla natura astratta e complessa dei costrutti psicologici, cercando di fornire strumenti validi e affidabili per la loro esplorazione e comprensione.

Affidabilità e Generalizzabilità nelle Misure Psicologiche

Nel contesto della misurazione psicologica, così come in altre discipline, è cruciale considerare le variabili che possono influenzare la precisione delle misure. L'affidabilità di uno strumento di misurazione psicologica si riferisce alla sua consistenza nel produrre risultati replicabili nel tempo e in contesti diversi. Gli indici di affidabilità sono utilizzati per quantificare il grado di riproducibilità e l'assenza di errori casuali nelle misurazioni.

Teoria Classica dei Test

L'approccio più ampiamente utilizzato nello studio dell'affidabilità delle misure psicologiche è rappresentato dalla teoria classica dei test, come descritto da Lord e Novick (1968). Secondo questa teoria, ogni misurazione (X) è composta da due componenti distintive: un punteggio "vero" (T) e un errore di misurazione (e). Il concetto di misurazione accurata, o "vera", può essere rappresentato come $X-e$, evidenziando il fatto che ogni misurazione può essere decomposta in tali elementi distinti.

La teoria classica dei test enfatizza l'importanza di condurre misurazioni ripetute per valutare l'affidabilità. Un concetto fondamentale è quello dei test paralleli, che consistono in due test con medie, varianze e distribuzioni identiche, e che mostrano una correlazione simile con variabili esterne. In questa prospettiva, il punteggio vero e l'errore di misurazione sono considerati indipendenti. Di conseguenza, la varianza dei punteggi osservati (Varianza X) è la somma della varianza dei punteggi veri (Varianza T) e della varianza dell'errore di misurazione (Varianza e).

L'affidabilità è quindi definita come il rapporto tra la varianza del punteggio vero e la varianza del punteggio osservato:

$$\text{Affidabilità} = \frac{\text{Varianza}(T)}{\text{Varianza}(X)}.$$

In termini pratici, un'affidabilità di 1 indicherebbe l'assenza di errori, mentre un'affidabilità di 0 implicherebbe che i punteggi derivano esclusivamente dall'errore. La correlazione tra il punteggio osservato e il punteggio vero è la radice quadrata dell'affidabilità, fornendo una stima della precisione della misurazione.

Questo framework fornisce una solida base per comprendere e quantificare l'affidabilità nelle misure psicologiche, sottolineando l'importanza di considerare sia i punteggi veri sia gli errori di misurazione per ottenere misurazioni precise e affidabili.

Evidenze Multiple di Affidabilità

Nonostante la teoria classica dei test fornisca una definizione matematica dei test paralleli, non fornisce dettagliate linee guida sulle procedure specifiche per costruirli. Tuttavia, a partire dagli anni '50, sono stati sviluppati diversi metodi che consentono di valutare empiricamente l'affidabilità delle misurazioni:

1. **Test-Retest**: Questo approccio implica la somministrazione dello stesso test ai partecipanti in due momenti diversi. L'obiettivo è valutare la stabilità dei punteggi nel tempo. Una correlazione elevata tra i punteggi ottenuti nei due momenti indica una buona affidabilità del test-retest.
2. **Equivalenza di Forme Parallele**: Questo metodo prevede l'utilizzo di due versioni diverse del test, ma che coprono lo stesso contenuto, somministrate simultaneamente ai partecipanti. Una forte correlazione tra i punteggi ottenuti dalle due versioni suggerisce che entrambe misurano il medesimo costrutto in modo affidabile.
3. **Split-Half e Coerenza Interna**:
    - **Split-Half**: I partecipanti completano una sola versione del test, la quale è divisa in due parti equivalenti. Si calcola poi la correlazione tra i punteggi delle due metà. Questo metodo valuta la coerenza interna del test.
    - **Coerenza Interna (ad esempio, Omega di McDonals)**: Valuta la correlazione tra tutti gli elementi del test. Un alto valore di coerenza interna indica che tutti gli elementi del test misurano aspetti simili del costrutto.
4. **Valutazione da Giudici Multipli**: In questo caso, i partecipanti sono valutati da più giudici in un'unica occasione. Un alto grado di accordo tra i giudici fornisce un'indicazione dell'affidabilità delle valutazioni.

Ciascuno di questi approcci fornisce indicazioni sull'affidabilità di un test, ma è fondamentale considerare che alcuni potrebbero essere più appropriati di altri in base alla natura del test e del costrutto misurato. L'affidabilità è pertanto un concetto multidimensionale che richiede l'impiego di diversi approcci per una valutazione completa delle misurazioni psicologiche.

Il Ruolo del Coefficiente Alpha nella Misurazione Psicologica

Il coefficiente alpha, introdotto da Cronbach nel 1951, è diventato un importante indicatore di coerenza interna nella letteratura psicologica, principalmente grazie alla sua facilità di calcolo. A differenza dell'affidabilità test-retest, che richiede dati raccolti in due momenti diversi, o dell'affidabilità delle forme parallele, che richiede la costruzione di due versioni alternative di un test, il coefficiente alpha può essere calcolato utilizzando un unico set di dati, rendendolo estremamente pratico come indice di affidabilità.

Tuttavia, è importante correggere un comune malinteso riguardo al coefficiente alpha: esso non misura direttamente l'omogeneità delle intercorrelazioni tra gli elementi o conferma la unidimensionalità di una scala. In realtà, il coefficiente alpha non fornisce informazioni dirette su questi aspetti strutturali della scala.

Per affrontare la questione della unidimensionalità, è necessario ricorrere a approcci più sofisticati come l'analisi fattoriale confermativa e i modelli di equazioni strutturali (SEM). Questi metodi consentono di testare quanto bene la struttura di correlazione degli elementi si adatti a un modello con un singolo

fattore rispetto a modelli multifattoriali, valutando se le correlazioni tra gli elementi possono essere meglio spiegate da un singolo costrutto sottostante.

Nel contesto delle analisi SEM, le saturazioni degli item indicano quanto della varianza di un item sia condivisa con gli altri (e quindi generalizzabile), mentre la varianza residua dell'item cattura l'errore unico associato a quell'item. La presenza di multidimensionalità emerge dalla capacità di un modello multifattoriale di adattarsi meglio ai dati rispetto a un modello a singolo fattore.

Quando un test è considerato multidimensionale, è ancora appropriato utilizzare il coefficiente alpha come indice di affidabilità? La risposta è negativa. In presenza di multidimensionalità, il coefficiente alpha tende a sottostimare l'affidabilità. Pertanto, è consigliabile, in tali casi, utilizzare altri metodi per valutare l'affidabilità, anziché basarsi esclusivamente sul coefficiente alpha.

Il Fenomeno dell'Attenuazione in Relazione all'Affidabilità

All'interno del contesto della teoria classica dei test, come delineato da Lord e Novick (1968), l'affidabilità svolge un ruolo cruciale poiché influisce sulla forza della correlazione che una misura può mostrare con altre variabili, come un criterio esterno. Secondo questa teoria, se l'errore nelle misurazioni è genuinamente casuale, il massimo teorico della correlazione tra una misura e un'altra variabile non è 1.0, ma piuttosto la radice quadrata dell'affidabilità di quella misura.

Ciò implica che, in presenza di un'affidabilità meno che ottimale, la correlazione effettiva tra una misura e qualsiasi altra variabile viene sistematicamente sottostimata, fenomeno noto come attenuazione. Questa attenuazione è direttamente proporzionale all'inadeguatezza dell'affidabilità: più bassa è l'affidabilità di una misura, maggiore sarà la sottostima della sua correlazione con altre variabili. Pertanto, per ottenere stime accurate delle correlazioni e comprendere veramente le relazioni tra diverse variabili, è fondamentale garantire che le misure utilizzate siano il più affidabili possibile. Questa considerazione enfatizza l'importanza dell'accuratezza e della precisione nelle procedure di misurazione psicologica.

La Teoria della Generalizzabilità

La Teoria della Generalizzabilità propone un approccio più completo e flessibile per comprendere l'affidabilità delle misure psicologiche rispetto alla classificazione tradizionale delle tipologie di affidabilità. Invece di limitarsi a categorizzare le misure in base a criteri specifici come test-retest, affidabilità interna o inter-valutatori, la Teoria della Generalizzabilità considera una serie di dimensioni che possono influenzare l'affidabilità in contesti diversi.

Una delle principali criticità della teoria classica dei test è la sua presunzione di uniformità e parallelismo delle misurazioni e degli errori casuali. La Teoria della Generalizzabilità, al contrario, riconosce che l'affidabilità dipende dalla specifica dimensione di generalizzazione considerata. Ad esempio, un test potrebbe essere affidabile per misurare una certa caratteristica in un contesto, ma non altrettanto affidabile in un contesto diverso o per una caratteristica correlata ma non identica.

Per superare le limitazioni della teoria classica dei test, l'American Psychological Association ha proposto l'adozione della Teoria della Generalizzabilità. Tuttavia, nonostante questa proposta, la pratica nei campi di ricerca non si è adeguatamente evoluta e la teoria della generalizzabilità non ha ancora completamente sostituito le nozioni più semplicistiche popolari in psicologia.

La Teoria della Generalizzabilità esamina diverse dimensioni che influenzano l'affidabilità, tra cui la dimensione temporale, delle forme, degli item e dei giudici o osservatori. Questa teoria enfatizza l'importanza di estendere le osservazioni a un'ampia varietà di situazioni e identificare l'impatto specifico delle fonti di varianza nei punteggi dei test in contesti particolari.

Invece dei tradizionali coefficienti di affidabilità come il coefficiente di stabilità o il coefficiente alfa, la Teoria della Generalizzabilità suggerisce l'uso di misure più ampie di affidabilità, come il coefficiente di correlazione intraclasse, per esaminare specifici aspetti dell'affidabilità. Questo approccio è particolarmente utile in ricerche con dati strutturati in maniera nidificata e dove diverse dimensioni possono influenzare l'affidabilità, come nei metodi di valutazione ecologica momentanea.### La Teoria della Risposta agli Item

La Teoria della Risposta agli Item (IRT) rappresenta un avanzamento rispetto alla teoria classica dei test, offrendo un approccio più sofisticato per analizzare le risposte degli individui agli item e la loro relazione con un costrutto latente. Questa teoria stabilisce un collegamento tra le risposte degli individui a un particolare item e il costrutto latente utilizzando una funzione chiamata "curva caratteristica dell'item".

La curva caratteristica dell'item mostra la probabilità che individui con differenti livelli del costrutto latente rispondano correttamente all'item, fornendo inoltre informazioni sulla capacità dell'item di distinguere tra individui con livelli elevati e bassi del tratto latente, oltre a misurare la sua difficoltà. Queste informazioni sono cruciali per identificare eventuali distorsioni negli item, noto come bias. Secondo la IRT, un item è privo di bias nel misurare un costrutto se individui con lo stesso livello del tratto ottengono punteggi attesi simili sull'item, indipendentemente da caratteristiche non rilevanti come genere, etnia o background culturale.

La Teoria della Risposta agli Item offre diversi vantaggi nel processo di creazione e valutazione di scale psicometriche:

1. **Selezione degli Item**: Permette di selezionare gli item in base alla loro difficoltà e alla capacità di discriminazione, superando così la limitazione della teoria classica che si basa esclusivamente sulle correlazioni tra gli item e il punteggio totale.
2. **Testing Adattivo Computerizzato**: La IRT facilita la valutazione della posizione di un individuo su un costrutto latente senza la necessità di somministrare l'intero test, grazie a tecniche come il testing adattivo computerizzato.

In conclusione, la Teoria della Risposta agli Item fornisce strumenti quantitativi per esaminare approfonditamente la relazione tra un item specifico e il costrutto latente, attraverso parametri di difficoltà e discriminazione.

Evoluzione e Comprensione della Validità nelle Misure Psicologiche

La nostra comprensione della validità nelle misure psicologiche ha subito un notevole sviluppo nel corso del tempo, passando da una visione iniziale più frammentata a un approccio più olistico e dinamico. Inizialmente, la validità veniva suddivisa in diversi tipi, tra cui la validità di contenuto, di facciata, orientata al criterio e di costrutto.

La validità di contenuto si riferisce alla rappresentatività degli item di un test rispetto al costrutto che si intende misurare, mentre la validità di facciata valuta se superficialmente gli item sembrano idonei a misurare il costrutto, sebbene questa non sia considerata un indice rigoroso di validità. La validità orientata al criterio si divide ulteriormente in predittiva e concorrente, che valutano la capacità del test di prevedere comportamenti futuri o di correlare con criteri esterni contemporaneamente misurati. Infine, la validità di costrutto indaga se il test misura effettivamente il costrutto in questione, richiedendo una comprensione approfondita sia del costrutto sia della metodologia del test.

Tuttavia, queste distinzioni sono state gradualmente considerate limitate e frammentarie. Un punto di svolta è stato rappresentato dall'approccio olistico di Samuel Messick, che ha enfatizzato che la validità va oltre la misura stessa, coinvolgendo l'interpretazione e l'uso dei punteggi del test. Messick ha sottolineato l'importanza di considerare le evidenze di validità da molteplici fonti e di assicurare la coerenza delle interpretazioni dei punteggi del test con le teorie psicologiche sottostanti.

Un'importante correzione concettuale è stata l'idea che la validità non sia un attributo statico dei test, ma piuttosto un processo continuo di accumulo di evidenze e giustificazioni teoriche. Questo processo di validazione riflette l'evoluzione delle teorie psicologiche e delle metodologie di misurazione, sottolineando che la validità è dinamica e contestuale.

In sintesi, l'evoluzione della concezione di validità nelle misure psicologiche sottolinea l'importanza di un approccio comprensivo, teoricamente informato e basato sull'evidenza per valutare, interpretare e utilizzare i punteggi dei test. Questo approccio moderno incoraggia i ricercatori e i praticanti a considerare la validità come un concetto ampio che incorpora molteplici aspetti della progettazione, dell'implementazione e dell'interpretazione dei test psicologici.

Approfondimento su Tecniche di Validazione di Costrutto e Costruzione di Scale

La discussione sulla evoluzione della validità nelle misure psicologiche può proseguire con l'esame delle tecniche che vengono usate per la validazione di costrutto e per la costruzione di scale. In particolare, gli strumenti maggiormente usati dagli psicometristi sono l'Analisi Fattoriale Confermativa (CFA) e i Modelli di Equazioni Strutturali (SEM).

L'**Analisi Fattoriale Confermativa (CFA)** rappresenta un approccio metodologico rigoroso, basato sull'ipotesi che un insieme di osservazioni possa essere spiegato da pochi costrutti latenti. A differenza dell'Analisi Fattoriale Esplorativa, che non prevede ipotesi a priori sui fattori, la CFA richiede che i ricercatori definiscano anticipatamente un modello teorico. Questo specifica le relazioni tra le variabili osservabili e i costrutti latenti, permettendo di testare l'adeguatezza del modello ai dati. La capacità della CFA di confrontare diversi modelli offre un mezzo potente per identificare la struttura che meglio rappresenta i dati.

Nel contesto della **valutazione della coerenza interna di una scala**, l'utilizzo della CFA supera i limiti dei metodi basati sulla teoria classica dei test, fornendo una valutazione più dettagliata e strutturata delle relazioni tra item e costrutti latenti.

I **Modelli di Equazioni Strutturali (SEM)** estendono le possibilità offerte dalla CFA, abilitando l'analisi delle relazioni di regressione non solo tra variabili manifeste e latenti, ma anche tra i costrutti latenti stessi. Questa caratteristica rende i SEM strumenti eccezionalmente potenti per esplorare le interazioni complesse tra variabili in uno studio psicometrico.

L'**esame della dimensionalità di un costrutto** attraverso la CFA e i SEM consente di testare con precisione le ipotesi sulla struttura dimensionale dei costrutti, verificando se l'organizzazione teorizzata degli item in fattori latenti corrisponde ai dati. Questi strumenti sono quindi fondamentali per confermare la struttura di un costrutto come ipotizzato dalla teoria sottostante.

In aggiunta, l'approccio **Multitrait-Multimethod (MTMM)** per esaminare la validità esterna, incorporando la validità convergente e discriminante, arricchisce ulteriormente la comprensione della misura. L'uso del disegno MTMM permette di distinguere efficacemente tra costrutti correlati ma distinti, assicurando che le misure non solo riflettano accuratamente il costrutto target, ma siano anche discriminanti rispetto ad altri costrutti.

In sintesi, l'integrazione di CFA e SEM nel processo di validazione di costrutti e nella costruzione di scale psicometriche rappresenta un avanzamento metodologico significativo. Questi approcci non solo migliorano la precisione e la comprensione delle relazioni tra variabili osservabili e latenti, ma contribuiscono anche a elevare la qualità e l'affidabilità delle misure psicologiche. Attraverso un uso attento e informato di queste tecniche, i ricercatori possono arricchire la validità e l'utilità delle scale psicometriche. Chi volesse approfondire ulteriormente questi argomenti, può fare riferimento al testo di John e Benet-Martinez (2014).

PART I

Programmazione

Chapter 1

Calendario delle lezioni

Il calendario didattico prevede 14 incontri di 3 ore ciascuno, con una verifica tramite Quiz Moodle e le presentazioni finali degli studenti negli ultimi due incontri.

Incontro	**Data**	**Argomento**	**Orario**
1	4 marzo 2025	Presentazione del corso, introduzione a R	10:30-13:30
2	6 marzo 2025	Concetti di base: Test psicologici e misure; distribuzioni di probabilità; modello lineare	8:30-11:30
3	11 marzo 2025	Teoria Classica dei	10:30-13:30
4	13 marzo 2025	Modello di regressione logistica. Mokken Scale Analysis (MSE)	8:30-11:30
5	18 marzo 2025	Item Response Theory (IRT). Validità.	10:30-13:30
6	20 marzo 2025	Path Analysis. Tutorial di Clement e Bradley-Garcia (2022)	8:30-11:30
7	25 marzo 2025	Elementi di algebra lineare, analisi delle componenti principali	10:30-13:30
8	27 marzo 2025	Analisi fattoriale esplorativa. Il modello statistico delll'analisi fattoriale	8:30-11:30
9	1 aprile 2025	Estrazione dei fattori, rotazione	10:30-13:30
10	3 aprile 2025	Analisi fattoriale confermativa	8:30-11:30
11	8 aprile 2025	Introduzione ai modelli di equazioni strutturali (SEM)	10:30-13:30
12	10 aprile 2025	Modelli multilivello; attendibilità dei giudici. Modelli di crescita latente	8:30-11:30
13	15 aprile 2025	Verifica tramite Quiz Moodle	10:30-13:30
14	17 aprile 2025	Presentazioni finali degli studenti	8:30-11:30

PART II

Punteggi e scale

Chapter 2

Punteggi e scale

Prerequisiti

- Leggere i capitoli 1, *Scores and Scales*, e 2, *Constructs*, del testo *Principles of psychological assessment* di Petersen (2024).
- Si consiglia di ripassare i concetti fondamentali della teoria delle probabilità, in particolare le distribuzioni di massa e di densità di probabilità. Per approfondire, si rimanda al materiale didattico dell'insegnamento di Psicometria disponibile al link [https://ccaudek.github.io/psicometria/].

Concetti e Competenze Chiave

Preparazione del Notebook

```
# Carica il file _common.R per impostazioni di pacchetti e opzioni
here::here("code", "_common.R") |> source()

# Carica pacchetti aggiuntivi
pacman::p_load(MASS, nortest)
```

2.1 Introduzione

Questo capitolo si propone di introdurre l'utilizzo del software "R", ponendo l'attenzione sulla differenza tra valutazioni normative e criteriali.

2.2 Tipologie di Dati

Si possono identificare quattro principali categorie di dati: nominali, ordinali, di intervallo e di rapporto. È opportuno notare che, in funzione dell'utilizzo della variabile, i dati possono rientrare in più di una categoria. La tipologia del dato influisce significativamente sulle modalità di analisi applicabili. A titolo esemplificativo, l'analisi statistica parametrica (come la regressione lineare) presuppone che i dati siano di intervallo o di rapporto.

2.2.1 Dati Nominali

I dati nominali si configurano come categorie distinte, caratterizzate da natura categorica e prive di ordinamento. Tali dati non esprimono affermazioni di natura quantitativa, bensì rappresentano entità nominabili (ad esempio, "felino" e "canino"). Sebbene possano essere rappresentati numericamente, come nel caso dei codici postali o dei codici identificativi di genere, etnia o razza dei partecipanti, è fondamentale sottolineare che valori numerici più elevati non riflettono livelli superiori (o inferiori) del costrutto, in quanto i numeri rappresentano meramente categorie prive di ordine intrinseco.

2.2.2 Dati Ordinali

I dati ordinali si distinguono per essere categorie ordinate: possiedono una denominazione e un ordine. Non forniscono informazioni sulla distanza concettuale tra i ranghi, ma indicano esclusivamente che valori più elevati rappresentano livelli superiori (o inferiori) del costrutto. Un esempio paradigmatico è costituito dalle posizioni in classifica successive a una competizione: il concorrente classificato al primo posto ha concluso la gara prima del secondo classificato, il quale a sua volta ha preceduto il terzo ($1 > 2 > 3 > 4$). È cruciale evidenziare che la distanza concettuale tra numeri adiacenti non è necessariamente equivalente.

2.2.3 Dati di Intervallo

I dati di intervallo sono caratterizzati da un ordine e da distanze significative (ovvero, intervalli equidistanti). Questi dati consentono operazioni di somma (ad esempio, 2 dista 2 unità da 4), ma non di moltiplicazione ($2 \times 2 \neq 4$). Esempi emblematici sono le temperature espresse in gradi Fahrenheit o Celsius: 100 gradi Fahrenheit non equivalgono al doppio di 50 gradi Fahrenheit. È importante sottolineare che, sebbene in psicologia molti dati presentino la medesima distanza matematica tra gli intervalli, è probabile che tali intervalli non rappresentino la medesima distanza concettuale.

2.2.4 Dati di Rapporto

I dati di rapporto si distinguono per essere ordinati, caratterizzati da distanze significative e da uno zero assoluto che rappresenta l'assenza del costrutto. In questa tipologia di dati, le relazioni moltiplicative risultano valide. Un esempio paradigmatico è la temperatura espressa in gradi Kelvin: 100 gradi Kelvin corrispondono effettivamente al doppio di 50 gradi Kelvin. Nel campo della psicologia, l'aspirazione a disporre di scale di rapporto persiste, nonostante la difficoltà di definire uno zero assoluto per i costrutti psicologici: come si potrebbe, infatti, concettualizzare l'assenza totale di depressione?

2.3 Punteggi Grezzi e Trasformati

Nell'ambito dei test psicometrici, il **punteggio grezzo** costituisce la valutazione più immediata e si basa sulla somma delle risposte categorizzate, come quelle corrette o errate, o vero o falso. Nonostante la sua immediatezza, il punteggio grezzo presenta limitazioni interpretative, poiché non considera fattori contestuali quali il numero totale di domande o il livello di difficoltà di queste.

Per mitigare queste limitazioni, i punteggi grezzi vengono spesso convertiti in formati che permettono un'interpretazione più contestualizzata, quali i punteggi standardizzati o scalati. Queste trasformazioni facilitano l'interpretazione dei risultati ottenuti.

L'interpretazione dei risultati dei test necessita di un riferimento comparativo. A seconda del contesto, può essere utile confrontare le prestazioni con una norma di riferimento o con criteri specifici.

Le **interpretazioni basate sulla norma** confrontano la performance di un individuo con quella di un gruppo di riferimento o normativo, offrendo una valutazione relativa alla prestazione tipica o "normale". Un esempio è rappresentato dai test di intelligenza. Al contrario, le **interpretazioni basate sul criterio** valutano le prestazioni rispetto a un livello di competenza specifico, indipendentemente dalla performance altrui.

Un altro approccio interpretativo è offerto dalla **Teoria della Risposta agli Item** (IRT), che fornisce un'analisi avanzata delle prestazioni nei test, permettendo un'esplorazione dettagliata delle risposte individuali.

2.4 Interpretazioni Basate sulla Norma (Norm-Referenced)

Per valutare la performance in un test psicologico, può essere utile confrontarla con quella di un gruppo predefinito. I punteggi grezzi acquisiscono significato quando messi a confronto con le prestazioni di un gruppo normativo. In questo contesto, i punteggi grezzi vengono trasformati in punteggi derivati basati sulle performance di un gruppo normativo specifico.

Un aspetto cruciale in queste interpretazioni è la pertinenza del gruppo di riferimento. È fondamentale che questo gruppo sia rappresentativo degli individui ai quali il test è destinato o con cui il partecipante viene confrontato.

La selezione del campione normativo, chiamato anche campione di standardizzazione, segue il principio del campionamento casuale stratificato proporzionale, assicurando che il campione rifletta proporzionalmente le caratteristiche demografiche nazionali. Tale rappresentatività è vitale per l'interpretazione basata sulla norma, rendendo necessaria l'accurata selezione e descrizione del campione da parte degli sviluppatori del test.

Quando si utilizzano questi test, è cruciale valutare se il campione di standardizzazione è rappresentativo per l'uso previsto e se le caratteristiche demografiche del campione corrispondono a quelle dei soggetti testati. La pertinenza e l'attualità del campione, insieme alla sua dimensione, sono fattori chiave per garantire interpretazioni valide e affidabili.

Una considerazione finale riguardante le interpretazioni basate sulla norma è l'importanza della standardizzazione nella somministrazione. È fondamentale che il campione di riferimento venga sottoposto al test nelle stesse condizioni e secondo le stesse procedure amministrative che saranno utilizzate nella pratica effettiva. Di conseguenza, quando il test viene somministrato in contesti clinici, è cruciale che l'utente del test segua attentamente le procedure amministrative prescritte. Ad esempio, nel caso di test standardizzati, è essenziale leggere le istruzioni testuali esattamente come sono fornite e rispettare rigorosamente i limiti di tempo. Sarebbe irragionevole confrontare la performance dell'esaminando in un test a tempo con quella di un campione di standardizzazione che ha avuto più o meno tempo per completare gli item. Questa necessità di seguire procedure standardizzate si applica a tutti i test standardizzati, sia quelli con interpretazioni basate sulla norma che quelli basati sul criterio.

2.4.1 Punteggi Derivati

In ambito psicometrico, i punteggi derivati da test possono assumere diverse forme, ciascuna con implicazioni specifiche per l'interpretazione dei dati. Esploreremo le tipologie più comuni:

1. **Punteggi Standardizzati:**
   - Questi punteggi trasformano i punteggi grezzi (ad esempio, il numero di risposte corrette) in misure standardizzate. Ciò permette di ottenere valori invarianti rispetto a variabili come l'età dell'individuo.
   - Si calcolano stabilendo una media e una deviazione standard specifiche a priori.
   - Esempi:
     - **z-scores**: Misurano la distanza di un punteggio dalla media, espressa in deviazioni standard. Hanno una media di 0 e una deviazione standard di 1.
     - **T-scores**: Trasformano i punteggi in valori positivi, con una media di 50 e una deviazione standard di 10.
     - **Punteggi di QI**: Tipici delle scale di intelligenza, hanno una media di 100 e una deviazione standard di 15.
2. **Punteggi Standardizzati Normalizzati:**
   - Quando i punteggi originali non seguono una distribuzione normale, si utilizzano trasformazioni non lineari per normalizzarli.
   - Esempi:
     - **Stanine**: Suddividono i punteggi in 9 categorie (da 1 a 9).
     - **Punteggi scalati di Wechsler**: Utilizzati nei test di intelligenza di Wechsler.
     - **Equivalenti della Curva Normale (NCE)**: Esprimono la posizione di un punteggio rispetto alla distribuzione normale.
3. **Ranghi Percentili:**
   - Vanno da 1 a 99 e indicano la posizione relativa di un soggetto rispetto alla popolazione.
   - Ad esempio, un punteggio al 75° percentile significa che il soggetto ha ottenuto un risultato migliore del 75% della popolazione.

2.5 Interpretazioni Basate su Criteri

L'approccio delle valutazioni basate su criteri specifici è diventato sempre più rilevante nel mondo dell'educazione e della psicometria a partire dagli anni Sessanta. Questo approccio, noto anche come

valutazione basata su contenuti, dominio o obiettivi, si concentra sulla misurazione delle competenze individuali rispetto a standard definiti, piuttosto che sul confronto con le prestazioni di un gruppo di riferimento.

Ecco alcune metodologie e applicazioni comuni:

1. **Percentuale di Risposte Corrette**:
   - Questo metodo fornisce un'indicazione diretta delle competenze di uno studente.
   - Ad esempio, se uno studente risponde correttamente all'85% delle domande di matematica, l'insegnante può valutare le sue abilità in modo specifico.
2. **Test di Padronanza**:
   - Questi test determinano se uno studente ha acquisito una competenza specifica.
   - Ad esempio, gli esami per la patente di guida valutano se lo studente ha raggiunto il livello di padronanza richiesto.
3. **Valutazioni Basate su Standard**:
   - Queste valutazioni classificano i risultati in categorie di prestazione (ad esempio, base, competente, avanzato).
   - Spesso, i punteggi vengono correlati a voti letterali basati su una percentuale di correttezza.

I punti di forza delle valutazioni basate su criteri includono:

- **Comparazione con Standard Predefiniti**:
  - Valutano il raggiungimento di competenze o obiettivi specifici, indipendentemente dalle prestazioni altrui.
  - Questo approccio evita il bias derivante dal confronto con altri studenti.
- **Focalizzazione su Competenze Specifiche**:
  - Questi test richiedono una definizione precisa dell'area di conoscenza o abilità valutata.
  - Sono ideali per valutare aree di contenuto specifiche.

2.5.0.1 Benefici

- **Valutazione Mirata delle Competenze**: Fornisce una verifica concreta del conseguimento delle conoscenze e abilità delineate dal programma di studi.
- **Personalizzazione dell'Insegnamento**: Identifica le aree di debolezza, consentendo un approccio didattico più focalizzato e personalizzato.

In conclusione, le valutazioni basate su criteri rappresentano un'alternativa preziosa ai metodi di valutazione tradizionali, specialmente in contesti in cui è fondamentale misurare le competenze individuali. Questo approccio è in crescente adozione in ambiti educativi e formativi, enfatizzando l'importanza dell'acquisizione di conoscenze e abilità mirate.

2.6 Analisi Comparativa tra Valutazioni Normative e Basate su Criteri

La distinzione tra valutazioni **normative** (norm-referenced) e **basate su criteri** (criterion-referenced) è fondamentale per interpretare le prestazioni individuali nei test. Sebbene un test possa teoricamente adottare entrambi gli approcci interpretativi, di solito si orienta verso uno dei due, a seconda dell'obiettivo specifico.

Ecco una panoramica delle differenze:

1. **Valutazioni Normative**:
   - **Versatilità**: Si applicano a test che valutano una vasta gamma di dimensioni, come attitudini, risultati scolastici, interessi, atteggiamenti e comportamenti.
   - **Ampio Quadro**: Ideali per esplorare costrutti generali come l'attitudine generale o l'intelligenza.
   - **Selezione delle Domande**: Preferiscono domande di difficoltà intermedia, evitando quelle troppo semplici o complesse.
2. **Valutazioni Basate su Criteri**:
   - **Specificità**: Associate principalmente a test che mirano a valutare conoscenze o competenze specifiche.
   - **Focalizzazione**: Concentrate su abilità e competenze ben definite.

- **Calibrazione delle Domande**: La difficoltà delle domande è tarata in base alle conoscenze o abilità specifiche da valutare.

È importante notare che queste interpretazioni non sono mutuamente esclusive. Alcuni test offrono sia valutazioni normative che basate su criteri, fornendo una visione completa delle prestazioni relative rispetto a un gruppo di riferimento e del livello di competenza in un ambito specifico. Questa dualità interpretativa è preziosa in vari contesti.

2.7 Analisi dei Punteggi secondo la Teoria della Risposta agli Item

La **Teoria della Risposta agli Item (IRT)** rappresenta un notevole avanzamento nel campo della **psicometria**, fornendo strumenti essenziali per valutare con precisione le capacità e i tratti latenti degli individui.

Fondamenti e Principi dell'IRT: L'IRT si basa sull'assunto che ogni persona possieda un livello di un tratto latente, come l'intelligenza, che è indipendente dalle specifiche domande del test o dal metodo di valutazione utilizzato. Attraverso l'applicazione di modelli matematici complessi, l'IRT consente di posizionare ogni individuo su un continuum di tratto latente, offrendo una misurazione delle capacità più precisa rispetto ai tradizionali punteggi grezzi.

Vantaggi dei Punteggi basati sull'IRT: I punteggi derivati dall'IRT presentano significativi vantaggi. Essi sono trattati come punteggi a intervalli costanti, consentendo comparazioni valide tra le performance di soggetti o gruppi diversi. Inoltre, questi punteggi mantengono una deviazione standard uniforme attraverso diverse fasce d'età, rendendoli particolarmente adatti per monitorare l'evoluzione o il progresso delle abilità nel tempo.

Applicazioni Pratiche e Prospettive Future dell'IRT: Una delle applicazioni più innovative dell'IRT è lo sviluppo dei **test adattivi computerizzati (CAT)**, in cui le domande vengono selezionate dinamicamente in base alle risposte precedenti del candidato. Questo metodo consente valutazioni precise ed efficienti delle abilità in tempo reale. Ad esempio, i punteggi IRT, come i *W-scores* nel *Woodcock-Johnson IV*, vengono utilizzati per analizzare variazioni nelle capacità cognitive legate ai processi di apprendimento o ai declini cognitivi.

2.8 Quali tipi di punteggi usare?

2.9 La Selezione del Punteggio Appropriato per la Valutazione

Determinare il tipo di punteggio più adeguato per un test è essenziale per ottenere informazioni specifiche e pertinenti dalla valutazione. Le diverse categorie di punteggi forniscono risposte a domande distinte riguardo alle prestazioni degli esaminandi:

1. **Punteggi Grezzi:**
   - Rappresentano la quantità totale di risposte corrette accumulate da un individuo.
   - Offrono una visione immediata del livello di prestazione e permettono di stabilire un ordine tra i partecipanti.
   - Sono utili per identificare rapidamente il posizionamento relativo di un individuo all'interno di un gruppo.
2. **Punteggi Norm-Referenced Standard:**
   - Forniscono un confronto diretto tra le prestazioni di un individuo e quelle di un gruppo normativo.
   - Consentono di interpretare la prestazione su una scala relativa, facilitando la comprensione del rendimento in termini di posizione all'interno di una popolazione di riferimento.
3. **Punteggi Criterion-Referenced:**
   - Indicano se un individuo ha raggiunto un determinato standard di competenza.
   - Sono particolarmente indicati per valutare il conseguimento di obiettivi specifici o competenze chiave.
4. **Punteggi Basati sull'IRT (Inclusi i Punteggi Rasch):**
   - Offrono una misurazione su scala a intervalli costanti, riflettendo la posizione di un individuo su un continuum di un tratto latente.

- Sono ideali per tracciare il progresso nel tempo o confrontare le prestazioni attraverso diverse valutazioni di un medesimo tratto.

Ad esempio, nel caso di Giovanni, che ha beneficiato di un programma di supporto alla lettura: - **Punteggi Norm-Referenced:** Fornirebbero insight su come le capacità di lettura di Giovanni si confrontano con quelle dei suoi coetanei dopo l'intervento. - **Punteggi Rasch o IRT:** Consentirebbero di valutare l'evoluzione precisa delle competenze di lettura di Giovanni, misurando il progresso a partire dal suo livello iniziale. - **Punteggi Grezzi:** Darebbero indicazioni sul miglioramento assoluto, sebbene privi della capacità di riflettere le variazioni in termini di difficoltà degli item o di altri fattori. - **Punteggi Criterion-Referenced:** Stabilirebbero se Giovanni ha raggiunto specifici obiettivi di competenza in lettura definiti a priori.

In contesti educativi, l'uso di punteggi norm-referenced standardizzati per età può essere preferibile per determinare se uno studente sta progredendo adeguatamente rispetto ai suoi pari. In contesti clinici, come nella gestione della depressione, i punteggi criterion-referenced possono offrire una valutazione mirata del raggiungimento di soglie di miglioramento clinico significativo.

In conclusione, la scelta del tipo di punteggio da utilizzare è guidata dal contesto di valutazione e dall'obiettivo specifico della misurazione. Diverse tipologie di punteggi illuminano aspetti distinti delle prestazioni, rendendoli più o meno adatti a seconda delle esigenze informative della valutazione.

2.10 Significato e Applicazione delle Norme e dei Punteggi Standardizzati

Per chiarire questi concetti, esaminiamo i dati della Tabella 2.1 di {cite:t}`bandalos2018measurement`. Con degli esempi numerici, analizzeremo vari tipi di punteggi normativi, tra cui:

- **Punteggi Percentili**: Che indicano la posizione relativa di un individuo all'interno del gruppo normativo.
- **Punteggi Standardizzati e Normalizzati**: Che trasformano i punteggi grezzi in una scala standard per facilitare il confronto tra diversi individui o gruppi.
- **Stanini**: Un metodo di punteggio che divide i punteggi in intervalli standardizzati.
- **Equivalenti alla Curva Normale**: Che adattano i punteggi a una distribuzione normale.

Nei capitoli successivi esamineremo come calcolare i punteggi basati sulla teoria IRT.

Iniziamo a leggere i dati.

```
raw_score <- c(
    26, 25, 33, 31, 26, 34, 29, 36, 25, 29, 28, 32, 25,
    30, 27, 31, 30, 30, 35, 30, 27, 26, 34, 32, 26, 34,
    30, 28, 28, 31, 30, 27, 26, 29, 29, 33, 27, 35, 26,
    27, 28, 29, 28, 27, 34, 36, 26, 26, 34, 30, 34, 27
)
```

2.10.1 Distribuzione di frequenze

```
freq <- table(raw_score) # frequency
cumfreq <- cumsum(freq) # cumulative frequency
perc <- prop.table(freq) * 100 # percentage
cumperc <- cumsum(perc) # cumulative percentage
pr <- (cumperc - 0.5 * perc) # percentile rank
cbind(freq, cumfreq, perc, cumperc, pr)
```

A matrix: 12 x 5 of type dbl

	freq	cumfreq	perc	cumperc	pr
25	3	3	5.769231	5.769231	2.884615
26	8	11	15.384615	21.153846	13.461538
27	7	18	13.461538	34.615385	27.884615
28	5	23	9.615385	44.230769	39.423077

		freq	cumfreq	perc	cumperc	pr
29	5	28	9.615385	53.846154	49.038462	
30	7	35	13.461538	67.307692	60.576923	
31	3	38	5.769231	73.076923	70.192308	
32	2	40	3.846154	76.923077	75.000000	
33	2	42	3.846154	80.769231	78.846154	
34	6	48	11.538462	92.307692	86.538462	
35	2	50	3.846154	96.153846	94.230769	
36	2	52	3.846154	100.000000	98.076923	

2.10.2 Punteggi Percentili

I punteggi percentili sono un modo efficace per interpretare e confrontare i punteggi di un individuo con quelli di un campione normativo. Un punteggio percentile indica la posizione relativa di un individuo all'interno di un gruppo normativo. Più specificamente, un punteggio percentile mostra la percentuale di persone nel campione normativo che ha ottenuto un punteggio uguale o inferiore a quello dell'individuo in questione.

Per esemplificare il concetto, consideriamo il calcolo di un quantile di ordine 0.74. Questo significa che stiamo cercando il valore al di sotto del quale si trova il 74% dei punteggi nel campione normativo. In altre parole, un individuo con un punteggio corrispondente a questo quantile ha superato il 74% delle persone nel gruppo normativo.

Il calcolo dei punteggi percentili può essere effettuato attraverso l'analisi statistica dei dati di un campione rappresentativo. Questi dati vengono ordinati in modo crescente, e si identifica il punteggio che corrisponde al percentile desiderato. Nel caso del quantile 0.74, si cerca il punteggio che si trova alla posizione che corrisponde al 74% della lunghezza totale dell'elenco ordinato dei punteggi.

```
# P74
quantile(raw_score, .74)
```

74%: 31.74

```
# Use a different type (see https://en.wikipedia.org/wiki/Quantile#Estimating_quantiles_from_a_sample
quantile(raw_score, .74, type = 6)
```

74%: 32

I punteggi percentili sono particolarmente utili perché offrono una comprensione intuitiva della posizione di un individuo rispetto agli altri. Tuttavia, è importante notare che essi rappresentano una scala ordinale e, pertanto, le differenze tra i punteggi percentili non sono necessariamente uniformi o proporzionali attraverso l'intera gamma di punteggi.

In conclusione, i punteggi percentili sono uno strumento fondamentale nella valutazione psicologica e educativa, poiché forniscono un modo diretto e facilmente interpretabile per valutare le prestazioni di un individuo in confronto a un campione normativo.

2.10.3 Punteggi Standardizzati

I punteggi standardizzati rappresentano una trasformazione essenziale nel campo della psicometria, che consente di convertire i punteggi grezzi ottenuti in un test in una scala unificata. Questa trasformazione permette di confrontare i risultati di individui o gruppi in maniera equa e coerente, superando le variazioni di scala o di difficoltà tra diversi test.

2.10.3.1 Principi Fondamentali dei Punteggi Standardizzati

- **Media e Deviazione Standard Predefinite**: I punteggi standardizzati sono calcolati in modo tale da avere una media e una deviazione standard specifiche, stabilite in anticipo. Per esempio, spesso si utilizza una media di 100 e una deviazione standard di 15 (come nei test di intelligenza) o una media di 0 e una deviazione standard di 1 (come negli z-score).

- **Risultati Confrontabili**: Attraverso questa standardizzazione, i punteggi diventano direttamente confrontabili. Un punteggio standardizzato rispetto a una media di 100 e una deviazione standard di 15, ad esempio, permette di valutare rapidamente se un punteggio è al di sopra, al di sotto o vicino alla media del campione normativo.

2.10.3.2 Come Funziona la Trasformazione

Il processo di standardizzazione implica la sottrazione della media del campione normativo dal punteggio grezzo di un individuo, seguita dalla divisione del risultato per la deviazione standard del campione normativo. In termini matematici, se $ X $ è un punteggio grezzo, $ $ è la media del campione normativo e $ $ è la deviazione standard del campione normativo, allora il punteggio standardizzato $ Z $ è calcolato come:

$$Z = \frac{X - \mu}{\sigma}.$$

2.10.3.3 Utilità dei Punteggi Standardizzati

- **Comparabilità**: Rendono i punteggi ottenuti da test diversi o da campioni diversi direttamente comparabili.
- **Interpretazione Facilitata**: Forniscono un modo semplice per interpretare i punteggi individuali in termini di posizione relativa rispetto alla media del campione normativo.
- **Adattabilità**: Sono utili in una varietà di contesti, da test educativi a valutazioni cliniche.

In conclusione, i punteggi standardizzati sono uno strumento cruciale nella psicometria e nella valutazione educativa. Trasformando i punteggi grezzi in una scala comune con media e deviazione standard specifiche, facilitano il confronto e l'interpretazione dei risultati dei test, rendendo più accessibile l'analisi e la valutazione delle prestazioni individuali e di gruppo.

Nel caso dell'esempio, i calcoli si svolgono in R nel modo seguente:

```
z_score <- (raw_score - mean(raw_score)) / sd(raw_score)
c(mean = mean(z_score), sd = sd(z_score))
```

mean -5.61516645146954e-16sd
1

2.10.3.4 Punteggi T

I punteggi T sono una forma specifica di punteggi standardizzati, utilizzati frequentemente nella psicometria per rendere più accessibili e interpretabili i risultati dei test. A differenza dei punteggi z, che tipicamente hanno una media di 0 e una deviazione standard di 1, i punteggi T sono trasformati in modo da avere una media fissata a 50 e una deviazione standard di 10.

2.10.3.5 Caratteristiche Principali dei Punteggi T

- **Media e Deviazione Standard**: La media fissata a 50 e la deviazione standard di 10 sono scelte per offrire una scala più intuitiva e di facile lettura rispetto agli z-score. Questa trasformazione sposta la scala degli z-score in una gamma numericamente più familiare e più semplice da interpretare per la maggior parte delle persone.
- **Calcolo dei Punteggi T**: Il calcolo dei punteggi T avviene trasformando prima i punteggi grezzi in z-score e poi convertendo questi z-score nella scala dei punteggi T. Matematicamente, se $ Z $ è lo z-score, il punteggio T corrispondente $ T $ è calcolato come:

  $$T = 50 + 10 \times Z.$$

  Questa formula adatta lo z-score in una scala che inizia da 50 e si allarga in entrambe le direzioni con incrementi standard di 10 per ogni deviazione standard.

2.10.3.6 Utilizzo dei Punteggi T

- **Facilità di Interpretazione**: I punteggi T sono particolarmente utili quando si desidera presentare i risultati dei test in un formato che sia immediatamente comprensibile, senza la necessità di ulteriori calcoli o trasformazioni.
- **Comparabilità**: Consentono di confrontare i risultati di test diversi in modo più diretto, grazie alla loro scala standardizzata.
- **Ampio Utilizzo**: Sono ampiamente usati in vari ambiti della valutazione psicologica, inclusi l'educazione, la ricerca e la pratica clinica.

In sintesi, i punteggi T offrono un modo efficace e standardizzato per interpretare i risultati dei test, rendendo i dati più accessibili e immediatamente comprensibili. La loro trasformazione da z-score a una scala con media 50 e deviazione standard 10 facilita la comprensione e la comparazione dei punteggi tra diversi test e diversi individui.

Svolgendo i calcoli in R otteniamo

```
T_score <- z_score * 10 + 50
c(mean = mean(T_score), sd = sd(T_score))
```

mean 50sd
10

2.10.4 Punteggi Stanini

I punteggi Stanini (dall'inglese "standard nine") rappresentano un metodo standardizzato per categorizzare i risultati dei test in psicometria, dividendoli in nove intervalli. Questa scala, progettata per semplificare l'interpretazione dei dati, permette di valutare la posizione relativa di un individuo all'interno di un gruppo di riferimento.

Come funzionano? Ogni intervallo Stanine corrisponde a un range di punteggi grezzi, con un'ampiezza che può variare leggermente a seconda della distribuzione dei dati. Un punteggio Stanine di 5 indica una prestazione media, mentre valori più alti o più bassi indicano prestazioni rispettivamente superiori o inferiori alla media. È importante notare che i punteggi Stanini sono principalmente utilizzati per confronti relativi all'interno di un gruppo, piuttosto che per misurazioni assolute.

Calcolo dei Punteggi Stanini. Per calcolare i punteggi Stanini, è necessario seguire alcuni passaggi:

1. **Determinare Media e Deviazione Standard**: Inizialmente, si calcolano la media e la deviazione standard dei dati del campione normativo.
2. **Applicare la Formula dei Punteggi Stanini**: Per ogni punteggio grezzo, si applica la seguente formula per calcolare il punteggio Stanine corrispondente:

$$\text{Stanine} = \left(\frac{\text{Punteggio Grezzo} - \text{Media}}{\text{Deviazione Standard}} \right) \times 2 + 5.$$

   Questa formula trasforma il punteggio grezzo in un valore sulla scala dei punteggi Stanini.
3. **Arrotondare al Numero Intero Più Vicino**: Infine, si arrotonda il risultato al numero intero più vicino per ottenere il punteggio Stanini finale.

I punteggi Stanini offrono diversi vantaggi:

- Semplicità: La scala a nove punti è facile da comprendere e memorizzare.
- Rapidità: Permettono una valutazione rapida della performance.
- Standardizzazione: Consentono di confrontare i risultati ottenuti in test diversi o da gruppi diversi.

Limitazioni:

Sebbene i punteggi Stanini siano uno strumento utile, è importante considerarne anche i limiti: essi assumono una distribuzione normale dei dati e quindi non sono adatti a tutti i tipi di test.

Per l'esempio presente abbiamo: