Indici di Variabilità e di Posizione

Test per le organizzaioni

Ottavia M. Epifania
ottavia.epifania@unipd.it

Margherita Calderan
margherita.calderan@unipd.it

Università di Padova

1 Indici di variabilità

2 Gli indici di posizione

3 Credits

1 Indici di variabilità

- Campo di variazione
- La varianza
- La deviazione standard
- Il coefficiente di variazione
- La differenza interquartilica

2 Gli indici di posizione

3 Credits

Indici di variabilità

- Il concetto di variabilità si riferisce a quanto i punteggi di una distribuzione sono sparsi ovvero quanto siano simili o dissimili tra loro
- Il ricorso ad un indice di tendenza centrale comporta una forte semplificazione, e da solo non fornisce informazioni esaurienti sulla distribuzione.
- E' fondamentale capire quanto i dati siano dispersi intorno all'indice di tendenza centrale.
- La variabilità è una caratteristica fondamentale delle distribuzioni "*Variability is the essence of statistics*" (Cobb, 1992)

Consideriamo i risultati dei compiti di Psicometria ottenuti dagli studenti di tre diversi Professori:

```
prof1 = c(18, 22, 24, 16, 19, 22 , 18, 21)
mean(prof1)
```

```
[1] 20
```

```
prof2 = c(10, 10, 12, 10, 30, 28 , 30, 30)
mean(prof2)
```

```
[1] 20
```

```
prof3 = c(20, 20, 20, 20, 20, 20 , 20, 20)
mean(prof3)
```

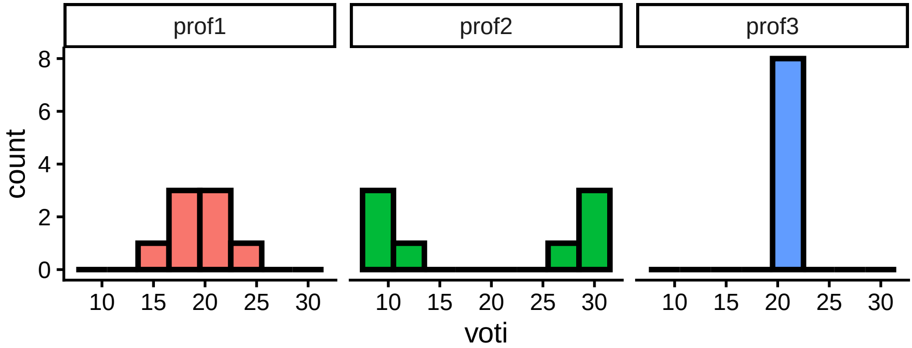
```
[1] 20
```

In ciascun gruppo di studenti la media dei voti è pari a 20, ma è evidente una diversa dispersione intorno a tale valore.

Indici di variabilità o disperisione

- Dobbiamo quindi considerare la la variabilità (o dispersione) di una distribuzione di dati.
- Gli indici di variabilità possono assumere solo valori positivi (non ha senso parlare di dispersione negativa) o nulli (quando i dati osservati hanno tutti lo stesso valore).
- La variabilità minima possibile è 0 e si riferisce a distribuzioni in cui tutti i punteggi sono uguali e dunque non c'è variabilità nei dati.

Quale gruppo è più variabile?



Campo di variazione

Campo di variazione (o gamma) di una distribuzione di dati è la differenza tra il valore massimo e il valore minimo osservato:

$$\text{gamma} = X_{max} - X_{min}$$

```r
prof1
```

```
[1] 18 22 24 16 19 22 18 21
```

```r
max(prof1) - min(prof1)
```

```
[1] 8
```

Esempio 2:

```r
x1 = c(10, 10, 27, 29, 30, 28, 30, 30)
x2 = c(10, 10, 16, 17, 30, 20, 21, 30)

max(x1) - min(x1)
```

```
[1] 20
```

```r
max(x2) - min(x2)
```

```
[1] 20
```

Commenti?

La varianza

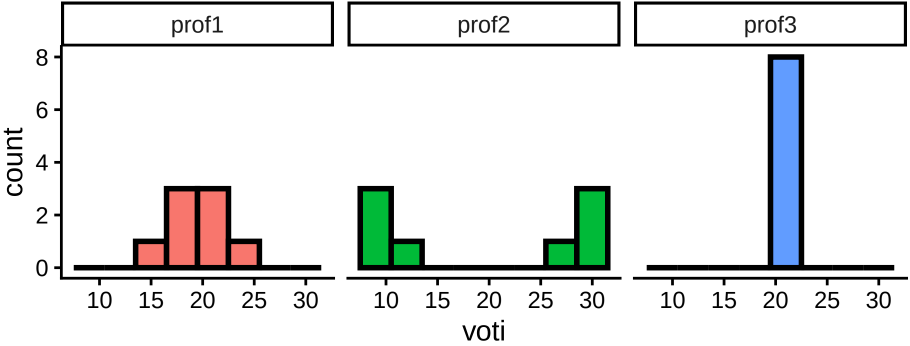
- La varianza σ^2 di un insieme di dati è la media degli scarti al quadrato tra i dati e la media dei dati stessi:

$$\sigma^2 = \frac{\sum_i^n (X_i - \bar{X})^2}{n}$$

La varianza assume valore minimo 0 quando tutti i dati sono uguali tra loro e aumenta all'aumentare della dispersione dei dati rispetto alla media:

$$\sigma^2 \geq 0$$

Qual'è la varianza maggiore? Quant'è la varianza di prof3?



Calcolo della varianza

$$\sigma^2 = \frac{\sum_i^n (X_i - \bar{X})^2}{n}$$

```
prof1
```

```
[1] 18 22 24 16 19 22 18 21
```

Il divisore è il numero di osservazioni: $n = 8$

```
divisore = length(prof1) # numero di osservazioni
```

$$\sigma^2 = \frac{\sum_i^8 (X_i - \bar{X})^2}{8}$$

Calcoliamo la **media** $\bar{X}$

```
media = mean(prof1)
media
```

```
[1] 20
```

$$\sigma^2 = \frac{\sum_i^8 (X_i - 20)^2}{8}$$

Il dividendo è dato dalla **somma dei quadrati degli scarti dalla media** aritmetica:

$$(18-20)^2+(22-20)^2+(24-20)^2+(16-20)^2+(19-20)^2+(22-20)^2+(18-20)^2+(21-20)^2$$

```
prof1
```

```
[1] 18 22 24 16 19 22 18 21
```

```
media
```

```
[1] 20
```

```
# ad ogni elemento di prof1 viene sottrata la media
(prof1-media)^2
```

```
[1]  4  4 16 16  1  4  4  1
```

```
dividendo = sum((prof1-media)^2)
```

La varianza è la **media dei quadrati degli scarti dalla media** aritmetica:

```
dividendo # 4 + 4 + 16 + 16 + 1 + 4 + 4 + 1
```

```
[1] 50
```

```
divisore
```

```
[1] 8
```

```
varianza = dividendo/divisore # varianza descrittiva
```

$$\sigma^2 = \frac{\sum_i^8 (X_i - 20)^2}{8}$$

Varianza descrittiva vs. inferenziale

- **Descrittiva**: descrive la dispersione dei dati osservati (come se fossero la popolazione).
- **Inferenziale**: stima la varianza della popolazione a partire da un campione.

Varianza Descrittiva (Popolazione)

Se i tuoi dati rappresentano *tutta* la popolazione (o vuoi solo descriverli):

$$\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}$$

- Denominatore: **n** (tutti i dati)
- Obiettivo: descrivere la dispersione osservata

Varianza Inferenziale (Campione)

Se hai un campione e vuoi stimare la varianza della popolazione:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

- Denominatore: **n - 1** (correzione di Bessel)
- Obiettivo: stima non distorta della varianza di popolazione

Perché $n-1$?

Per definizione, la somma degli scarti dalla media è sempre zero:

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0$$

Se conosci **n - 1** scarti, l'ultimo è automaticamente determinato (deve far sommare tutto a zero).

Perché n - 1 corregge il bias

Quando usi $\bar{X}$ invece della "vera" media μ:

- $\bar{X}$ è "ottimizzata" per i tuoi dati specifici, è la media del campione non della popolazione (μ)
- Gli scarti $(X_i - \bar{X})$ sono più piccoli di quelli che otterresti con μ
- Dividere per **n** sottostima la vera varianza
- Dividere per **n - 1** (numero più piccolo) aumenta il risultato

```r
x = c(2, 4, 4, 4, 5, 5, 7, 9)
```

Usa σ^2 (n) quando:

- Fai solo descrizione/esplorazione

```r
# Varianza campionaria (n-1) - default
var(x)
```

```
[1] 4.571429
```

Usa s^2 (n - 1) quando:

- Hai un campione
- Vuoi stimare parametri della popolazione

```r
# Varianza descrittiva (n) - manuale
sum((x - mean(x))^2) / length(x)
```

```
[1] 4
```

La deviazione standard

La deviazione standard (o scarto quadratico medio) è la radice della varianza:

$$\sigma = \sqrt{\sigma^2}$$

Riporta l'indice di varibilità sulla scala della variabile.

Es. In campione di 20 soggetti è stata rilevata la variabile peso. In tale campione la media è pari a $70kg$ e la deviazione standard è pari a 10.7. Si potrà affermare che i soggetti **differiscono mediamente** di $10.7kg$ **dal peso medio** di $70kg$.

Il coefficiente di variazione

Il coefficiente di variazione è dato dal rapporto tra la **deviazione standard** e il **valore assoluto della media** dei dati:

$$CV = \frac{\sigma}{|\bar{X}|}$$

Il **CV** è un indice di variabiltà relativa che **tiene conto**, oltre che della **deviazione standard** dei dati, anche della **media**. Per questo motivo è molto utile per eseguire dei confronti in termini di **variabilità tra fenomeni "diversi" tra loro**.

A un gruppo di 15 studenti viene somministrato un test che valuta le conoscenze informatiche generali (20 domande, del tipo *giusto/sbagliato*). I 15 studenti seguono un corso di informatica generale organizzato dal Comune e a fine corso viene somministrato loro un test sulle conoscenze informatiche. Il test prevede 40 domande del tipo *giusto/sbagliato*. I risultati dei due test, in termini di media e deviazione standard di risposte corrette, sono presentati nella seguente tabella:

Test	media	deviazione standard
Pre-test	9	5
Post-test	30	8

C'è più variabilità tra i soggetti al pre-test o al post-test?

- Naturalmente confrontare le deviazioni standard non è di grande aiuto. Esse dipendono fortemente dalle media dei dati su cui sono state calcolate.
- Per poter operare un confronto sulla variabilità dei punteggi ottenuti ai due test è opportuno calcolare i rispettivi coefficienti di variazione:

$$CV_{\text{pre-test}} = \frac{5}{9} = .56$$

$$CV_{\text{post-test}} = \frac{8}{30} = .27$$

Osservando i risultati si può concludere che le conoscenze informatiche dei soggetti sono più omogenee al post-test.

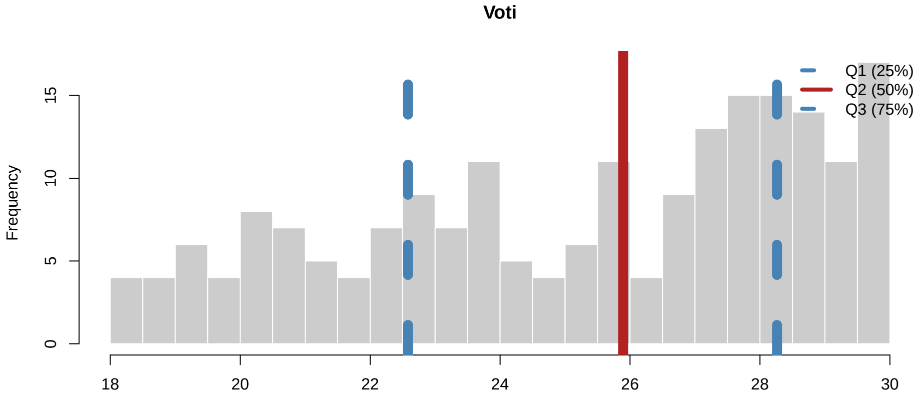
La differenza interquartilica

La differenza interquartilica di una distribuzione è la differenza tra il terzo e il primo quartile (o equivalentemente tra il 75—esimo e il 25—esimo percentile) dei dati:

$$IQR = Q_{75} - Q_{25}$$

Cosa vuol dire?

Immaginiamo di avere i voti di 200 studenti:



La differenza interquartilica

- Q1 (25%) indica una soglia sotto cui cade circa un quarto della classe;
- Q2 (50%) indica una soglia sotto cui cade la metà della classe;
- Q3 (75%) lascia sotto di sé circa tre quarti dei voti.

Cos'è Q2?

1 Indici di variabilità

2 Gli indici di posizione

3 Credits

1 Indici di variabilità

2 Gli indici di posizione
- I quantili
- Rango percentile

3 Credits

I quantili

Data una distribuzione di dati, si definisce come **Quantile di indice p** e si indica con Q_p , il dato al di sotto del quale si situa una percentuale **p** di dati.

Ad esempio, la mediana può essere considerata come il quantile Q_{50}, e cioè il dato al di sotto del quale si situa il **50%** dei dati.

I quantili

- Esistono diverse tipologie di quantili.
- Rispetto all'utilizzo nelle applicazioni in psicologia, i più importanti sono i **Quartili** e i **Percentili**.

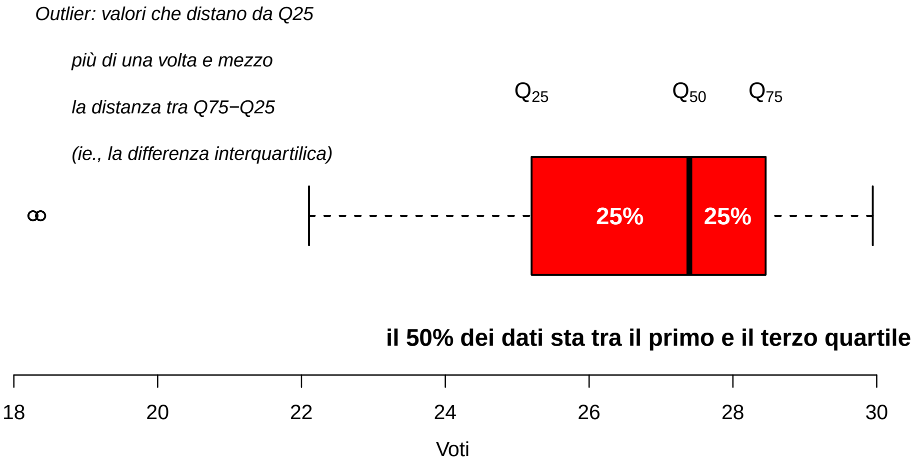
I quartili

I quartili dividono in 4 parti uguali la distribuzione dei dati. Essi sono:

- Il primo quartile Q_{25} (o Q_1) : il dato al di sotto del quale si situa il **25%** dei dati.
- Il secondo quartile (o mediana) Q_{50} (o Q_2): il dato al di sotto del quale si situa il **50%** dei dati.
- Il terzo quartile Q_{75} (o Q_3): il dato al di sotto del quale si situa il **75%** dei dati.

I quartili vengono rappresentati all'interno di un grafico molto utile per descrivere i dati detto diagramma a scatola (boxplot).

Boxplot


Outlier: valori che distano da Q25
più di una volta e mezzo
la distanza tra Q75−Q25
(ie., la differenza interquartilica)
Q25
Q50
Q75
25%
25%
il 50% dei dati sta tra il primo e il terzo quartile
18
20
22
24
26
28
30
Voti

La differenza interquartilica

La differenza interquartilica (IQR) misura la dispersione del 50% centrale dei dati (tra primo e terzo quartile), quindi i valori estremi incidono poco o per nulla sul suo valore. Per questo è chiamato un indice di variabilità robusto.

```
IQR(voti) # funzione per il calcolo automatico
```

```
[1] 3.235585
```

Outilier

- I baffi si estendono fino all'ultimo dato entro 1.5 volte l'IQR dai quartili. Tutto ciò che sta oltre i baffi è un outlier.
- Un outlier è un valore che dista dai quartili (Q_1 o Q_3) più di 1.5 volte la differenza interquartile (IQR).
- È un outlier qualsiasi osservazione che cade fuori dall'intervallo $[Q_1 - 1.5 \times IQR \quad ; \quad Q_3 + 1.5 \times IQR]$.

I percentili

I percentili, spesso indicati con la lettera maiuscola **P**, dividono in **cento parti** la distribuzione dei dati.

Alcuni percentili molto importanti, sia dal punto di vista statistico che rispetto alle applicazioni in psicologia, sono:

$P_5 \quad P_{25} \quad P_{50} \quad P_{75} \quad P_{95}$

I quantili

Per parlare di percentili/quantili dobbiamo prima ordinare i voti dal più basso al più alto:

```
head(voti, n = 6) # primi 6 voti
```

```
[1] 21.33 18.01 24.13 18.17 18.78 29.46
```

```
tail(voti, n = 6) # ultimi 6 voti
```

```
[1] 27.38 29.18 29.76 28.73 27.82 29.28
```

I quantili

Per parlare di percentili/quantili dobbiamo prima ordinare i voti dal più basso al più alto:

```
v_sorted = sort(voti) # riordino ?sort

head(v_sorted, n = 6) # primi 6 voti
```

```
[1] 18.01 18.17 18.17 18.17 18.19 18.31
```

```
tail(v_sorted, n = 6) # ultimi 6 voti
```

```
[1] 29.72 29.76 29.88 29.92 29.96 29.98
```

I quantili

Trovare il valore sotto cui cade circa il 25% dei voti.

```
n = length(voti) # numero di osservazioni

p = 0.25 # voglio il 25° percentile

k = n * p  # quanti voti rappresentano il 25% del totale
#(25% di 200 = 50)

Q_25 = v_sorted[k] # predo il voto in quella posizione
Q_25
```

```
[1] 21.32
```

```
sum(voti <= Q_25)  # quanti sono <= soglia
```

```
[1] 50
```

I quantili

Fortunatamente esiste la funzione `quantile()` che permette di calcolare i quantili:

```
# probs specifica che quantili vogliamo

quantile(voti, probs = c(0.05, 0.25,0.50,0.75, 0.95))
```

```
     5%     25%     50%     75%     95%
18.7780 21.3275 24.8200 27.7975 29.4610
```

Esempio: Workaholism (dipendenza da lavoro)

A 6 lavoratori adulti è stato somministrato un test standardizzato a livello nazionale sul workaholism (punteggio totale). Il punteggio di ciascun partecipante è riportato nella tabella seguente:

Codice partecipante	1	2	3	4	5	6
Punteggio (totale)	40	50	30	80	23	42

Valutare le prestazioni dei 6 partecipanti alla luce dei valori normativi del test:

Percentile	P5	P25	P50	P75	P95
Punteggio	31	42	51	68	78

Rango percentile

Il **Rango percentile** (Rp) indica la **percentuale di casi** che hanno un **valore uguale o inferiore** a un particolare punteggio X_i

Supponiamo di disporre di un test che misura il **ragionamento numerico** in candidati per posizioni amministrative.
Il punteggio di un soggetto al test è ottenuto come la somma di risposte corrette a 20 domande.

Per i candidati con diploma di scuola superiore, al punteggio 15 è associato il rango percentile di 60 ($Rp_{15} = 60$).

Il **60%** dei candidati italiani con diploma di scuola superiore ottiene un punteggio al test inferiore o uguale a 15.

Rango percentile

Supponiamo di aver somministrato il test a un candidato con diploma e di aver verificato sul manuale che il suo punteggio equivale a un rango percentile pari a 5.

Come interpretare la sua performance?

Calcolo dei ranghi percentili

Per calcolare un rango percentile, basta seguire questi semplici step:

1. Determinare quanti casi sono nel gruppo.
2. Determinare quanti casi cadono "sotto" o al punteggio di interesse
3. Dividere il numero di casi ottenuti allo step 2 per il numero totale di casi nel gruppo (step 1)
4. Moltiplicare il risultato dello step 3 per 100

Rango percentile

$$P_r = \frac{B}{N} \times 100 = \text{rango percentile di } X_i$$

- X_i = punteggio di interesse
- N = numero totale di casi
- B = numero di casi $\leq X_i$
- P_r = rango percentile

Il file `data_work.rda` contiene il numero di lavoratori che, su 1.000, subiscono un infortunio sul lavoro.

```r
load("data/data_work.rda")
data_work
```

```
          Paese Infortuni_ogni_1000_lavoratori
1     Singapore                           2.06
2      Giappone                           2.86
3       Francia                          17.11
4        Spagna                           3.86
5     Australia                           4.10
6        Italia                          20.01
7   Stati Uniti                          12.68
8          Cina                           2.26
9       Turchia                           4.03
10      Marocco                           5.12
11      Bolivia                          27.21
12      Etiopia                          32.10
13    Mozambico                          33.44
```

Rango percentile

Supponiamo che il nostro caso di interesse siano gli Stati Uniti. Prima di tutto riordiniamo i dati in ordine crescente.

```
data_sorted=data_work[order(data_work$Infortuni_ogni_1000_lavoratori), ]
data_sorted
```

```
         Paese Infortuni_ogni_1000_lavoratori
1    Singapore                           2.06
8         Cina                           2.26
2     Giappone                           2.86
4       Spagna                           3.86
9      Turchia                           4.03
5    Australia                           4.10
10     Marocco                           5.12
7  Stati Uniti                          12.68
3      Francia                          17.11
6       Italia                          20.01
14 Afghanistan                          25.02
11     Bolivia                          27.21
12     Etiopia                          32.10
```

Rango percentile

Determiniamo il numero di casi con tassi inferiori o uguali al nostro caso di interesse.

Il numero di casi con un tasso di infortuni inferiore a quello degli Stati Uniti (tasso migliore) sono 7 paesi: Marocco, Australia, Turchia, Spagna, Giappone, Cina, Singapore - hanno tassi inferiori a 12.68

```
data_sorted[data_sorted$Infortuni_ogni_1000_lavoratori <= 12.68,]
```

```
          Paese Infortuni_ogni_1000_lavoratori
1     Singapore                           2.06
8          Cina                           2.26
2      Giappone                           2.86
4        Spagna                           3.86
9       Turchia                           4.03
5     Australia                           4.10
10      Marocco                           5.12
7  Stati Uniti                           12.68
```

Rango percentile

Calcoliamo il rango percentile:

$$P_r = \frac{B}{N} \times 100 = \text{rango percentile di } X_i$$

```
# numero di casi minori o uguali al punteggio degli stati uniti
(B = sum(data_sorted$Infortuni_ogni_1000_lavoratori <= 12.68))
```

```
[1] 8
```

```
#numero di casi totali
(N = nrow(data_sorted))
```

```
[1] 14
```

```
(P = (B/N) * 100)
```

```
[1] 57.14286
```

Il 57.14286% per cento di casi ha valori uguali o inferiori agli Stati Uniti.

In R è possibile utilizzare direttamente la funzione `rank()` per ottenere l'inidice di posizione per ciascun valore, e poi dividere per il numero totale per ottenere il rango percentile per ogni valore

```
data_work$Infortuni_ogni_1000_lavoratori
```

```
 [1]  2.06  2.86 17.11  3.86  4.10 20.01 12.68  2.26  4.03  5.12 27.21 32.10
[13] 33.44 25.02
```

```
rango = rank(data_work$Infortuni_ogni_1000_lavoratori)
rango # indica l'ordine crescente
```

```
 [1]  1  3  9  4  6 10  8  2  5  7 12 13 14 11
```

```
(N = nrow(data_work)) # quante osservazioni totali
```

```
[1] 14
```

Rango percentile

```
# aggiungo la variabile Pr al dataset
data_work$Pr = rango/N #divido l'indice di posizione per il valore totale

data_work
```

```
         Paese Infortuni_ogni_1000_lavoratori         Pr
1    Singapore                           2.06 0.07142857
2     Giappone                           2.86 0.21428571
3      Francia                          17.11 0.64285714
4       Spagna                           3.86 0.28571429
5    Australia                           4.10 0.42857143
6       Italia                          20.01 0.71428571
7  Stati Uniti                          12.68 0.57142857
8         Cina                           2.26 0.14285714
9      Turchia                           4.03 0.35714286
10     Marocco                           5.12 0.50000000
11     Bolivia                          27.21 0.85714286
12     Etiopia                          32.10 0.92857143
13   Mozambico                          33.44 1.00000000
14 Afghanistan                          25.02 0.78571429
```

1 Indici di variabilità

2 Gli indici di posizione

3 Credits

1 Indici di variabilità

2 Gli indici di posizione

3 Credits

Credits

Altoè, G. (2022). Corso di Testing Psicologico, Scienze psicologiche dello sviluppo, della personalità e delle relazioni interpersonali, A.A. 2022/23

Marci, M. (2025). Corso di Testing Psicologico, Scienze psicologiche dello sviluppo, della personalità e delle relazioni interpersonali, A.A. 2025/26