Frequenze, Indici di Posizione, di Variabilità e di Forma

Test per le organizzazioni

Ottavia M. Epifania
ottavia.epifania@unipd.it
Margherita Calderan
margherita.calderan@unipd.it

Università di Padova

03/2026

1 Statistica descrittiva e inferenziale

2 Dati

3 Frequenze

4 Indici di tendenza centrale

5 Indici di variabilità

6 I percentili

7 Indici di forma

Statistica descrittiva e inferenziale

Statistica descrittiva:

- Riassume, descrive, esplora i dati osservati
- Prima fase nella valutazione delle proprietà psicometriche di uno strumento

Statistica inferenziale:

- Usa i dati di un **campione** per fare inferenze sulla **popolazione**
- Confronti tra gruppi, valutazione di interventi, validazione di test

1 Statistica descrittiva e inferenziale

2 Dati

3 Frequenze

4 Indici di tendenza centrale

5 Indici di variabilità

6 I percentili

7 Indici di forma

Dati

Il dataset sintetico contiene dati individuali relativi a dipendenti di un'azienda e viene utilizzato per modellare il rischio di burnout e analizzare i fattori associati al benessere lavorativo.

Se non l'hai già scaricato, clicca **qui** per scaricare il dataset.

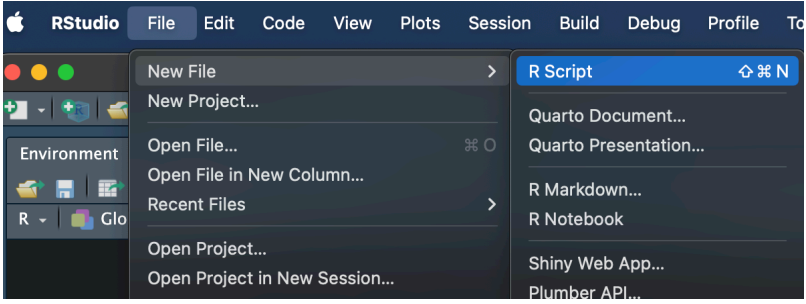
Fonte: **https://www.kaggle.com/datasets/ankam6010/synthetic-hr-burnout-dataset**

Ambiente di lavoro (ideale)

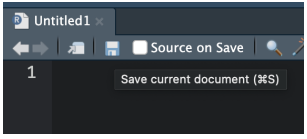
```
Utente
  |
  |--- NomeUtente
         |
         |--- Desktop
         |      |
                |--- test-organizzazioni
                |       |
                        |-- data
                        |     |---- burnout.csv
                        |
                        |- script
                        |     |---
```

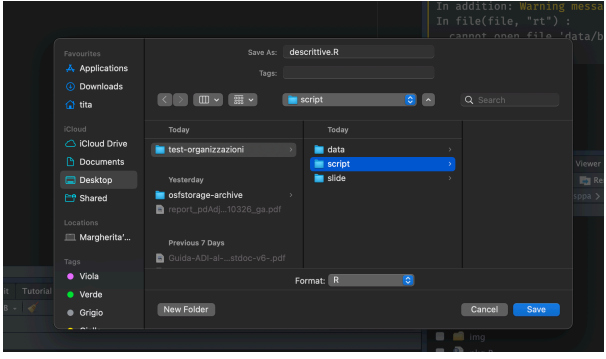
Apriamo

Creiamo uno script:



Salviamo:


Untitled1
Source on Save
1
Save current document (⌘S)


Save As: descrittive.R
Tags:
script
Search
Favourites
Applications
Downloads
tita
iCloud
iCloud Drive
Documents
Desktop
Shared
Locations
Margherita'...
Tags
Viola
Verde
Grigio
Today
test-organizzazioni
Today
data
script
slide
Yesterday
osfstorage-archive
report_pdAdj...10326_ga.pdf
Previous 7 Days
Guida-ADI-al-...stdoc-v6-.pdf
Format: R
New Folder
Cancel
Save

```r
# Pulisco lo spazio di lavoro
rm(list = ls())

# Imposto working directory
# setwd('/Users/tita/Desktop/test-organizzazioni')

# Carico dataset
burnout = read.csv("data/burnout.csv")
```

Tip

In RStudio puoi anche usare **Session → Set Working Directory → Choose Directory** per impostare la cartella senza scrivere il percorso a mano.

Visualizzo la struttura:

```
str(burnout)
```

```
'data.frame':   2000 obs. of  10 variables:
 $ Name              : chr  "Max Ivanov" "Max Wang" "Nina Petrov"
 $ Age               : int  32 40 33 35 59 45 31 39 24 22 ...
 $ Gender            : chr  "Male" "Female" "Female" "Female" ...
 $ JobRole           : chr  "Analyst" "Engineer" "Engineer" "Manag
 $ Experience        : int  3 9 2 6 8 11 5 17 0 1 ...
 $ WorkHoursPerWeek  : int  60 47 44 44 38 44 70 31 63 30 ...
 $ RemoteRatio       : int  21 67 20 70 46 90 30 53 35 5 ...
 $ SatisfactionLevel: num  4.4 2.09 2.58 3.23 4.41 4.31 2.18 3.95
 $ StressLevel       : int  1 2 3 8 1 7 3 4 5 5 ...
 $ Burnout           : int  0 0 0 0 0 0 0 0 0 0 ...
```

Come prima cosa vogliamo esplorare la distribuzione dei livelli di stress (variabile `StressLevel`).

```
str(burnout$StressLevel)
```

```
 int [1:2000] 1 2 3 8 1 7 3 4 5 5 ...
```

```
# definisco la natura della variabile
burnout$StressLevel = as.ordered(burnout$StressLevel)
str(burnout$StressLevel)
```

```
 Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 2 3 8 1 7 3 4 5 5
```

La variabile `StressLevel` è stata raccolta attraverso una scala likert a 10 punti. E' quindi una variabile di tipo ordinale.

1 Statistica descrittiva e inferenziale

2 Dati

3 Frequenze

- Le frequenze assolute semplici
- Le frequenze assolute cumulate
- Le frequenze relative semplici
- Le frequenze relative cumulate
- Visualizzazione

4 Indici di tendenza centrale

5 Indici di variabilità

Notazioni fondamentali

- Sia X la variabile `StressLevel`
- Sia X_i la categoria/livello i-esima di X, dove $i = 1 \dots 10$
- Sia n il totale delle unità statistiche ($n = 2000$)

Le frequenze assolute semplici

La frequenza assoluta semplice di una categoria è il **numero naturale di unità statistiche** che presentano tale categoria.

Se vogliamo calcolare le frequenze assolute semplici per ogni categoria/livello, possiamo utilizzare la funzione `table()`

```
# nome_dataframe$nome_variabile
fi = table(burnout$StressLevel); fi
```

```
  1   2   3   4   5   6   7   8   9  10
220 199 199 200 196 197 191 208 214 176
```

Le frequenze assolute cumulate

La frequenza assoluta cumulata di una categoria/livello è la **somma** delle **frequenze assolute semplici** delle categorie precedenti alla categoria data più la frequenza assoluta semplice della categoria data.

Attraverso la funzione `cumsum()` possiamo calcolare la somma cumulata:

```
# nome_dataframe$nome_variabile
Fi = cumsum(table(burnout$StressLevel)); Fi
```

```
   1    2    3    4    5    6    7    8    9   10
 220  419  618  818 1014 1211 1402 1610 1824 2000
```

Le frequenze relative semplici

La frequenza relativa semplice è data dal rapporto tra la **frequenza assoluta semplice di tale categoria** e **il numero totale** di unità statistiche osservate.

Qual è il numero totale di unità statistiche?

```
nrow(burnout)                # numero righe dataframe
```

```
[1] 2000
```

```
length(burnout$StressLevel)  # lunghezza della variabile
```

```
[1] 2000
```

```
n = nrow(burnout)            # oggetto n = numero unità statistiche
```

Qual è la frequenza relativa semplice associata alle categorie/livelli?

```
pi = table(burnout$StressLevel) / n ; pi
```

```
    1     2     3     4     5     6     7     8     9    10
0.110 0.100 0.100 0.100 0.098 0.098 0.096 0.104 0.107 0.088
```

Una frequenza relativa semplice varia sempre tra 0 e 1.

Le frequenze relative cumulate

La frequenza relativa cumulata di una categoria è la **somma delle frequenze relative semplici** delle categorie precedenti alla categoria data più la frequenza relativa semplice della categoria data.

Come prima, attraverso la funzione `cumsum()`, possiamo calcolare le frequenze relative **cumulate**:

```
Pi = cumsum(table(burnout$StressLevel) / n) ; Pi
```

```
    1     2     3     4     5     6     7     8     9    10
0.110 0.210 0.309 0.409 0.507 0.606 0.701 0.805 0.912 1.000
```

Una frequenza relativa cumulata varia sempre tra 0 e 1.

Creo una tabella con le frequenze:

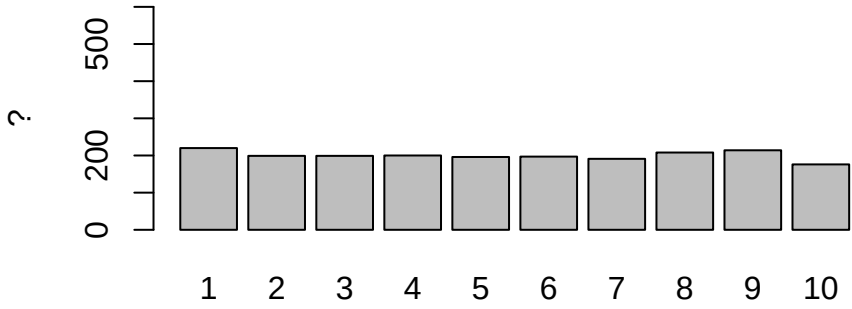
```
#accedo ai livelli della variabile StressLevel
tabella = data.frame(Categoria = levels(burnout$StressLevel),
                     fi = as.numeric(fi), # mi assicuro che sia numerico
                     Fi = as.numeric(Fi),
                     pi = as.numeric(pi),
                     Pi = as.numeric(Pi))
```

```r
# la funzione print ci permette di stampare l'output sulla console
# row.names = FALSE evita che le righe del dataframe siano numerate
print(tabella, row.names = FALSE)
```

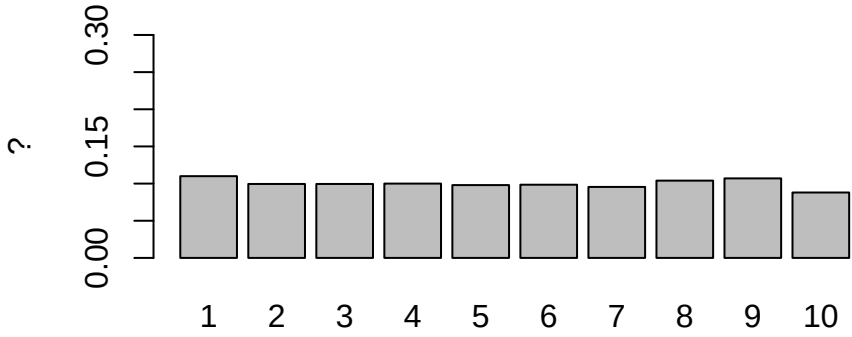
```
 Categoria  fi   Fi     pi     Pi
         1 220  220 0.1100 0.1100
         2 199  419 0.0995 0.2095
         3 199  618 0.0995 0.3090
         4 200  818 0.1000 0.4090
         5 196 1014 0.0980 0.5070
         6 197 1211 0.0985 0.6055
         7 191 1402 0.0955 0.7010
         8 208 1610 0.1040 0.8050
         9 214 1824 0.1070 0.9120
        10 176 2000 0.0880 1.0000
```

Visualizzazione

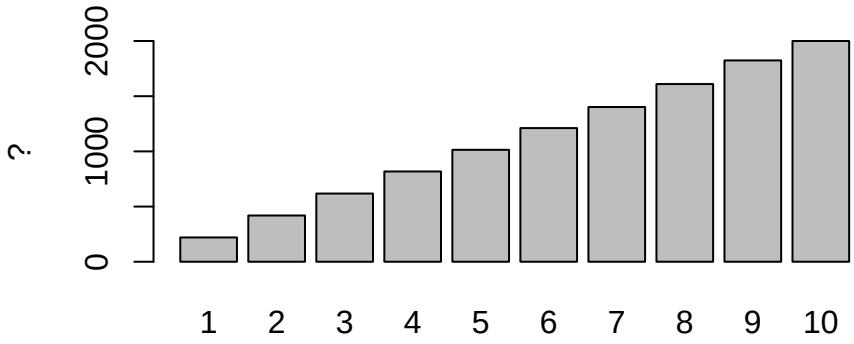
```r
# cosa rappresenta l'asse y?
barplot(table(burnout$StressLevel),
        ylab = "?",
        ylim = c(0, 600))
```



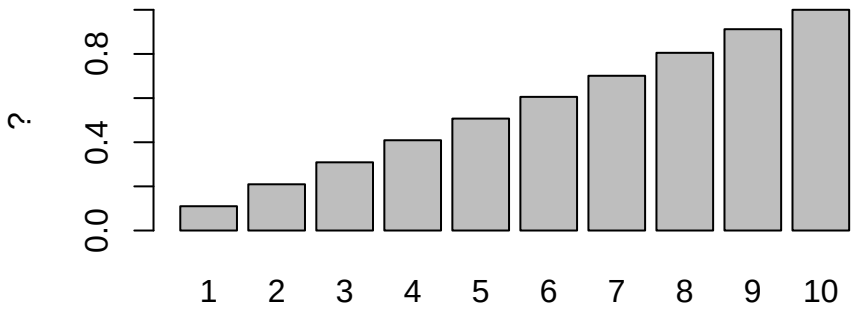
```r
# cosa rappresenta l'asse y?
barplot(table(burnout$StressLevel) / n,
        ylab = "?",
        ylim = c(0, .3))
```



```r
# cosa rappresenta l'asse y?
barplot(cumsum(table(burnout$StressLevel)),
        ylab = "?")
```



```r
# cosa rappresenta l'asse y?
barplot(cumsum(table(burnout$StressLevel)) / n,
        ylab = "?")
```



1 Statistica descrittiva e inferenziale

2 Dati

3 Frequenze

4 Indici di tendenza centrale
- La moda
- La mediana
- La media aritmetica

5 Indici di variabilità

6 I percentili

Indici di tendenza centrale

Un indice di tendenza centrale è un valore che descrive e riassume il centro di una distribuzione di dati.

Indice	Variabile nominale	Variabile ordinale	Variabile quantitativa
Moda	SI	SI	SI
Mediana	NO	SI	SI
Media	NO	NO	SI

La moda

La moda di una distribuzione di dati rilevati sulla variabile X è la categoria che si presenta con la massima frequenza.

Ad esempio, rispetto ai dati relativi ai livelli di stress, la moda è ….

```
table(burnout$StressLevel)
```

```
  1   2   3   4   5   6   7   8   9  10
220 199 199 200 196 197 191 208 214 176
```

La mediana

- La mediana di una distribuzione di dati ordinati rilevati sulla variabile X è il dato che occupa la posizione centrale rispetto alla distribuzione dei dati.

Calcolo della mediana

Si individua la prima frequenza cumulata maggiore o uguale alla posizione cercata:

```
# Qual è la posizione centrale?
i = (n + 1) / 2; i
```

```
[1] 1000.5
```

```
# Calcolo frequenza cumulata
cumsum(table(burnout$StressLevel))
```

```
   1    2    3    4    5    6    7    8    9   10
 220  419  618  818 1014 1211 1402 1610 1824 2000
```

C'è anche la funzione `median()`

```
median(as.numeric(burnout$StressLevel))
```

```
[1] 5
```

La media aritmetica

La media aritmetica di una distribuzione di dati rilevati sulla variabile X è data dalla somma dei dati divisa per il numero di unità statistiche.

In questo caso dobbiamo cambiare variabile perché questo indice richiede una variabile quantitativa:

```
str(burnout$Age)
```

```
 int [1:2000] 32 40 33 35 59 45 31 39 24 22 ...
```

Calcolo della media:

```r
# Numero di unità
n = length(burnout$Age)

# Sommo tutti gli elementi e divido per n
sum(burnout$Age) / n
```

```
[1] 40.6945
```

```r
# Funzione mean
mean(burnout$Age)
```

```
[1] 40.6945
```

1 Statistica descrittiva e inferenziale

2 Dati

3 Frequenze

4 Indici di tendenza centrale

5 Indici di variabilità

- Campo di variazione
- La varianza
- La deviazione standard
- Il coefficiente di variazione
- I quantili e la differenza interquartilica

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 a un indice di tendenza centrale (come la media) comporta una forte semplificazione, e da solo non fornisce informazioni esaurienti sulla distribuzione.
- È 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).

Esempio: numero di ore lavorate

Consideriamo le ore di lavoro settimanali di tre diversi team:

```
team_A = c(38, 42, 44, 36, 39, 42, 38, 41)
mean(team_A)
```

```
[1] 40
```

```
team_B = c(30, 30, 32, 30, 50, 48, 50, 50)
mean(team_B)
```

```
[1] 40
```

```
team_C = c(40, 40, 40, 40, 40, 40, 40, 40)
mean(team_C)
```

```
[1] 40
```

Indici di variabilità o dispersione

- Dobbiamo quindi considerare 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.

Campo di variazione

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

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

```
team_A
```

```
[1] 38 42 44 36 39 42 38 41
```

```
max(team_A) - min(team_A)
```

```
[1] 8
```

```
range(team_A)          # restituisce direttamente [min, max]
```

```
[1] 36 44
```

Confrontiamo il Team B con un ipotetico Team D:

```
team_D = c(30, 30, 32, 31, 31, 30, 30, 50)
team_B
```

```
[1] 30 30 32 30 50 48 50 50
```

```
# Gamma Team B
max(team_B) - min(team_B)
```

```
[1] 20
```

```
# Gamma Team D
max(team_D) - min(team_D)
```

```
[1] 20
```

Commenti?

La varianza

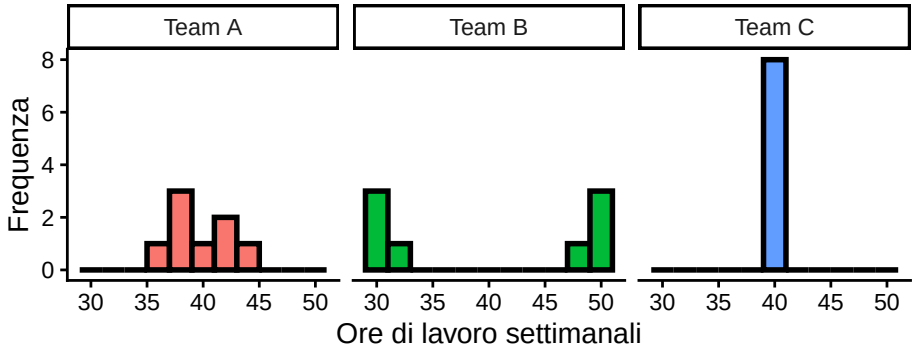
La varianza σ^2 di un insieme di dati è la media degli scarti al quadrato tra i dati e la media dei dati stessi:

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

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 tra i nostri tre team? E quant'è la varianza del team C?



Calcolo della varianza

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

Prendiamo in esame il team A:

```
team_A
```

```
[1] 38 42 44 36 39 42 38 41
```

Il divisore è il numero di osservazioni (membri del team): $n = 8$

```
divisore = length(team_A)
```

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

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

```
media = mean(team_A)
media
```

```
[1] 40
```

$$\sigma^2 = \frac{\sum_i^8 (X_i - 40)^2}{8}$$

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

```
# scarti dalla media
scarti = team_A - media
scarti
```

```
[1] -2  2  4 -4 -1  2 -2  1
```

```
# quadrati
scarti_quadrati = scarti^2
scarti_quadrati
```

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

```
# somma dei quadrati degli scarti
dividendo = sum(scarti_quadrati)
```

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 (divide per n)
varianza
```

```
[1] 6.25
```

Important

La funzione `var()` di R calcola la **varianza campionaria**, che divide per $n-1$ anziché n. Per dati descrittivi (popolazione osservata) si usa n; per stimare la varianza della popolazione da un campione si usa $n-1$.

```
var(team_A)          # divide per n-1 (varianza campionaria)
```

```
[1] 7.142857
```

```
varianza             # divide per n   (varianza descrittiva)
```

```
[1] 6.25
```

La deviazione standard

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

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

Riporta l'indice di variabilità sulla scala della variabile.

Es. Se in un campione di dipendenti è stata rilevata la variabile `WorkHoursPerWeek` con una media pari a 40 ore e una deviazione standard pari a 2.5, si potrà affermare che i dipendenti differiscono mediamente di 2.5 ore dal carico di lavoro medio di 40 ore.

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 **variabilità relativa** che tiene conto, oltre che della **deviazione standard**, anche della **media**. Per questo motivo è fondamentale per eseguire dei **confronti** in termini di variabilità tra fenomeni misurati su **scale diverse**.

Esempio sul dataset Burnout

```
# eseguire questo codice se il dataset non è già stato caricato
burnout = read.csv("data/burnout.csv")
```

Vogliamo capire se c'è più variabilità negli anni di esperienza (`Experience`) dei dipendenti o nelle loro ore lavorative settimanali (`WorkHoursPerWeek`).

Estraiamo media e deviazione standard direttamente dal dataset:

```
# Statistiche per il livello di esperienza
media_esp = abs(mean(burnout$Experience))
# > [1] 10.0745
sd_esp    = sd(burnout$Experience)
# > [1] 9.148267

# Statistiche per le ore lavorative
media_ore = abs(mean(burnout$WorkHoursPerWeek))
# > [1] 49.588
sd_ore    = sd(burnout$WorkHoursPerWeek)
# > [1] 11.83242
```

Naturalmente confrontare le deviazioni standard non è di grande aiuto. Esse dipendono fortemente dalla media dei dati su cui sono state calcolate.

Per poter operare un confronto sulla variabilità è opportuno calcolare i rispettivi coefficienti di variazione:

```
cv_esp = sd_esp / media_esp
cv_ore = sd_ore / media_ore

# CV Esperienza:
round(cv_esp, 2)
```

```
[1] 0.91
```

```
# CV Ore Lavorative:
round(cv_ore, 2)
```

```
[1] 0.24
```

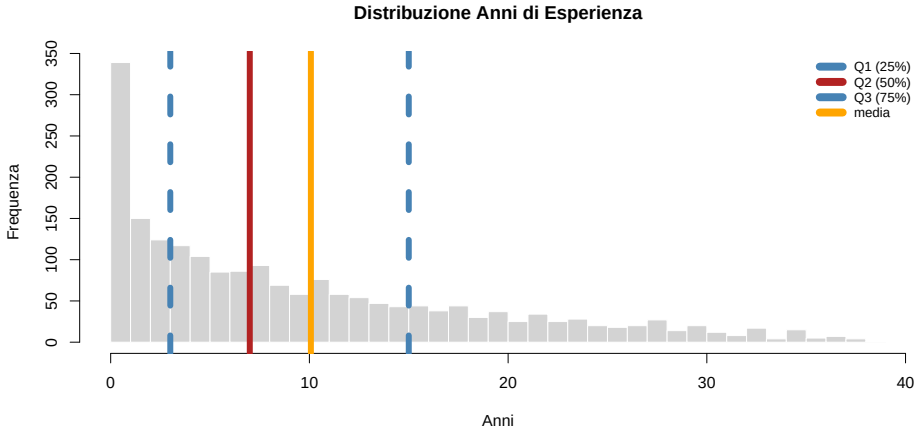
I quantili e la differenza interquartilica

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:

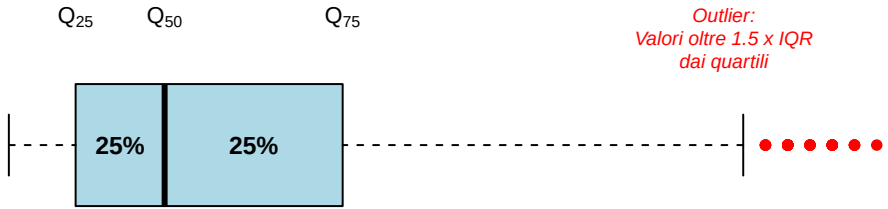
$$IQR = Q_{75} - Q_{25}$$

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.

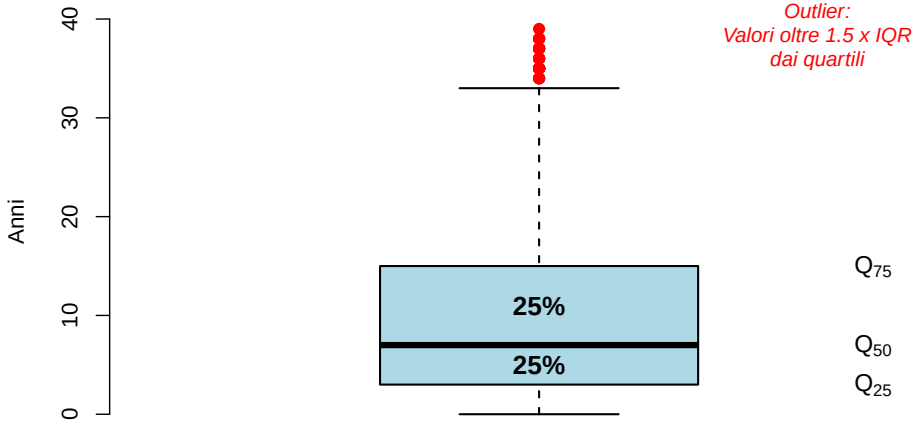
Analizziamo la distribuzione degli anni di esperienza nel dataset burnout:



Boxplot



Il 50% dei dati sta tra il primo e il terzo quartile


Anni
0
10
20
30
40
25%
25%
Outlier:
Valori oltre 1.5 x IQR
dai quartili
Q75
Q50
Q25

Codice boxplot semplificato:

```
# variabile Experience del dataset burnout
bp = boxplot(burnout$Experience,
             horizontal = TRUE,
             # Visualizzo il boxplot in orizzontale
             main = "Boxplot", #titolo
             xlab = "Anni", # nome della variabil
             # (asse x perchè orizzontale)
             col = "lightblue", #colore
             border = "black", #bordo
             outcol = "red", # colore outlier
             outpch = 19, # forma outlier (i.e, pallini)
             lwd = 1.5) # spessore linee
```

La differenza interquartilica

Traduciamo in codice R la formula teorica:

$$[Q_1 - 1.5 \times IQR \quad ; \quad Q_3 + 1.5 \times IQR]$$

```
# 1. Calcoliamo i quartili (attraverso la funzione quantile)
Q1 = quantile(burnout$Experience, prob = 0.25); Q1
```

```
25%
  3
```

```
Q3 = quantile(burnout$Experience, prob = 0.75); Q3
```

```
75%
 15
```

```
# 2. Calcoliamo la differenza
differenza_iqr = Q3 - Q1

# 3. Definiamo i limiti oltre i quali un valore è un outlier
limite_inferiore = Q1 - (1.5 * differenza_iqr)
limite_superiore = Q3 + (1.5 * differenza_iqr)
```

Quali dipendenti sono outlier?

Dal grafico sappiamo che gli outlier in questo caso sono solo i valori maggiori del limite superiore (i.e., 33).

```
subset(burnout, #dataset
       select = c(Name, Experience), # colonne d'interesse
       subset = Experience>limite_superiore) # valuto per riga
```

```
              Name Experience
59         Sam Lee         38
102   Dina Garcia         34
157    Sam Petrov         35
300      Kate Kim         37
395  Ivan Johnson         35
407   Lily Petrov         36
418    Max Ivanov         39
445   Alex Petrov         37
```

Equivalente con dataframe[valuto per riga, colonne d'interesse]:

```
# dataset[condizione da valutare per riga, colonne d'interesse]
burnout[burnout$Experience > limite_superiore, c("Name", "Experience")]
```

	Name	Experience
59	Sam Lee	38
102	Dina Garcia	34
157	Sam Petrov	35
300	Kate Kim	37
395	Ivan Johnson	35
407	Lily Petrov	36
418	Max Ivanov	39
445	Alex Petrov	37
478	Alex Garcia	37
501	Sam Wang	35
510	Kate Brown	35

Outlier

Gli outlier identificati per la variabile `Experience` potrebbero non coincidere con quelli rilevati per altre variabili, o emergere solo quando si considerano più variabili simultaneamente. È buona prassi, quindi, indagarne le possibili cause e valutarli in un'ottica multivariata, tenendo conto del profilo complessivo di ciascuna osservazione su tutte le variabili di interesse.

1 Statistica descrittiva e inferenziale

2 Dati

3 Frequenze

4 Indici di tendenza centrale

5 Indici di variabilità

6 I percentili

- Esempio: livelli di soddisfazione
- Rango percentile

I percentili

Il **percentile** x_p **di ordine** p è quella categoria/valore che è:

- preceduta da almeno $p\%$ dei casi
- superata da almeno $(1-p)\%$ dei casi

Quartili $\rightarrow$ percentili di ordine 25 - 50 - 75

Decili $\rightarrow$ percentili di ordine 10 - 20 - ... - 90

Percentili $\rightarrow$ percentili di ordine 1 - 2 - ... - 98 - 99

Esempio: livelli di soddisfazione

Trovare il valore sotto cui cade circa il 25% dei punteggi di soddisfazione.

La funzione `quantile()` che permette di calcolare i quantili. L'argomento `prob = ...` definisce che percetile vogliamo ottenere.

```
quantile(burnout$SatisfactionLevel, probs = c(0.05, 0.25, 0.50, 0.75, 0.95))
```

```
    5%     25%     50%     75%     95%
1.2100 2.0000 3.0250 4.0000 4.7905
```

Esempio

Il livello di soddisfazione di 6 lavoratori è riportato nella tabella seguente:

Codice partecipante	1	2	3	4	5	6
Punteggio	3.65	4.90	1.06	2.93	3.10	4.97

Valutare i livelli di soddisfazione dei 6 dipendenti alla luce dei dati del campoione

```
quantile(burnout$SatisfactionLevel, probs = c(0.05, 0.25, 0.50, 0.75, 0.95))
```

```
    5%     25%     50%     75%     95%
1.2100 2.0000 3.0250 4.0000 4.7905
```

Rango percentile

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

Per i dipendenti del reparto `Sales`, a 3.05 di soddisfazione è associato il rango percentile di 50 ($Rp_{3.05} = 50$).

Il **50%** dei dipendenti del reparto `Sales` riporta dei livelli di soddisfazione inferiori o uguali a 3.05.

Supponiamo di aver intervistato un dipendente e di aver verificato che il suo livello di soddisfazione equivale a un rango percentile pari a 98.

Come interpretare il suo livello di soddisfazione?

Calcolo dei ranghi percentili

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

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

In R è possibile utilizzare direttamente la funzione `rank()`: *"Qual è il **rango** (la posizione relativa) di ciascun dipendente rispetto al resto?"*

```
rango = rank(burnout$SatisfactionLevel)
head(rango) # ordine crescente (primi 6)
```

```
[1] 1704.0  548.5  786.5 1103.5 1711.5 1664.5
```

```
head(burnout$SatisfactionLevel) # primi 6 valori originali
```

```
[1] 4.40 2.09 2.58 3.23 4.41 4.31
```

```
N = nrow(burnout) # numero totale di osservazioni
```

```r
# Calcolo del rango
Rp_SatisfactionLevel = (rango / N) * 100

# Aggiungo la variabile al dataset
burnout$Rp_SatisfactionLevel = Rp_SatisfactionLevel

# Visualizzo nome, ore e rango percentile
# [tutte le righe, colonne d'interesse]
burnout[, c("Name", "SatisfactionLevel", "Rp_SatisfactionLevel")]
```

	Name	SatisfactionLevel	Rp_SatisfactionLevel
1	Max Ivanov	4.40	85.200
2	Max Wang	2.09	27.425
3	Nina Petrov	2.58	39.325
4	John Ivanov	3.23	55.175
5	John Wang	4.41	85.575
6	Lily Smith	4.31	83.225
7	Max Garcia	2.18	29.550

1 Statistica descrittiva e inferenziale

2 Dati

3 Frequenze

4 Indici di tendenza centrale

5 Indici di variabilità

6 I percentili

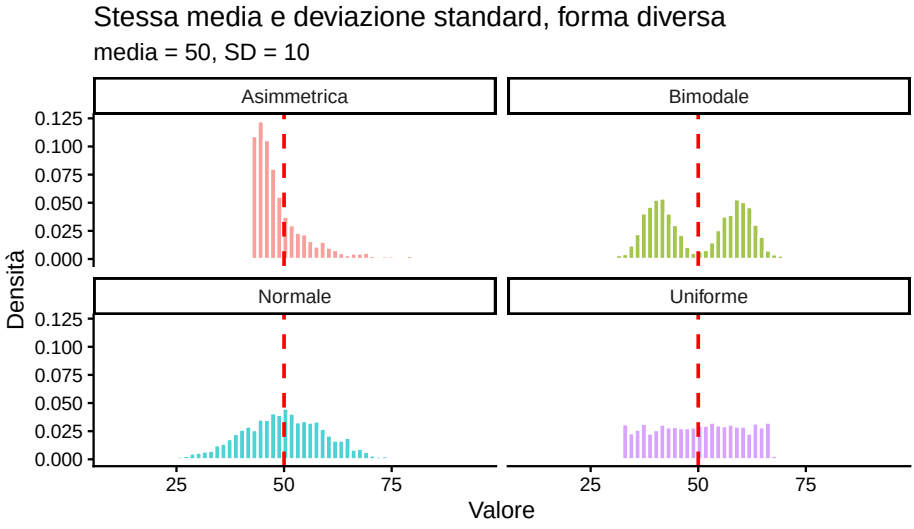
7 Indici di forma

- Asimmetria (Skewness)
- Curtosi (Kurtosis)

Indici di forma

Gli indici di forma descrivono **come** è distribuita una variabile, oltre alla sua posizione centrale e dispersione. Due distribuzioni possono avere stessa media e deviazione standard ma **forma diversa**.

- La distribuzione è **simmetrica** o **sbilanciata** (asimmetrica)?
- Come sono le code (le estremità) della distribuzione?


Stessa media e deviazione standard, forma diversa
media = 50, SD = 10
Asimmetrica
Bimodale
Normale
Uniforme
Densità
0.125
0.100
0.075
0.050
0.025
0.000
25
50
75
Valore

Asimmetria (Skewness)

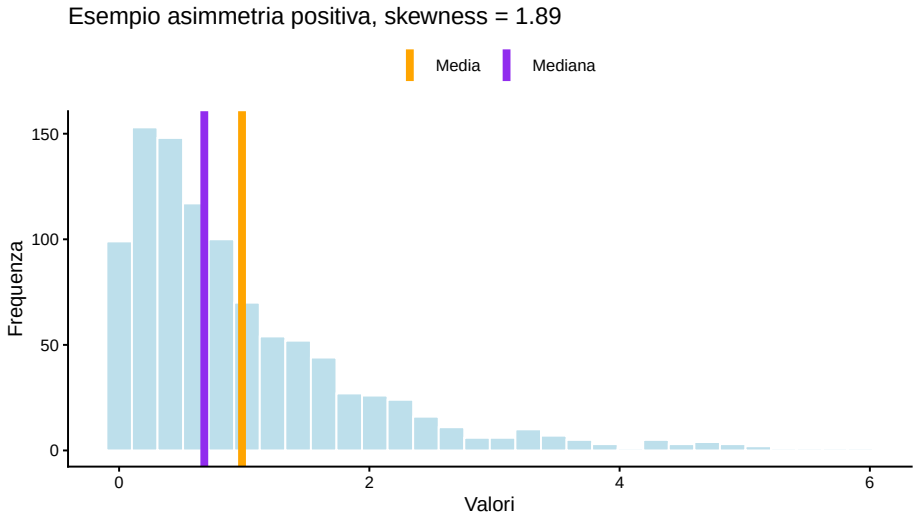
L'**asimmetria** misura quanto una distribuzione si discosta dalla simmetria rispetto al centro.

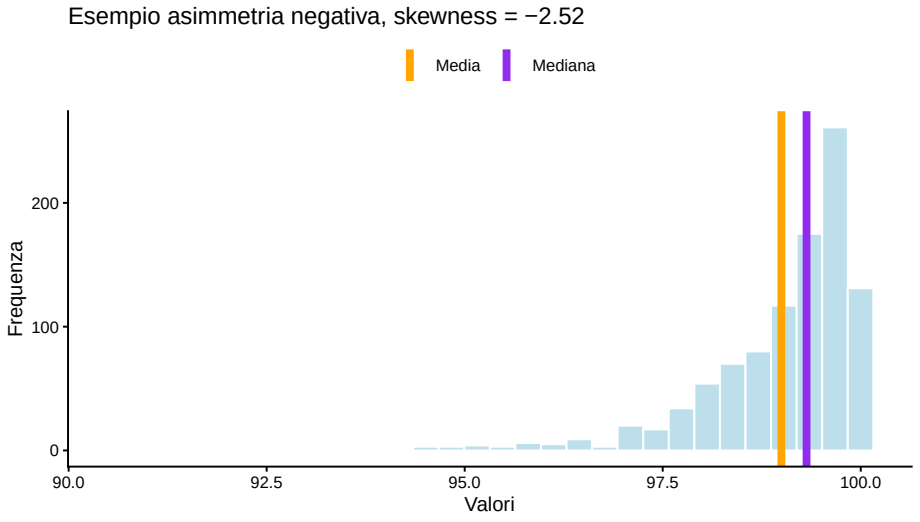
Distribuzione simmetrica: media, mediana e moda coincidono

Distribuzione asimmetrica: i valori si "addensano" da una parte e la coda si allunga dall'altra

Tipi di asimmetria

Tipo	Valore indice	Caratteristiche
Simmetrica	≈ 0	Media $=$ Mediana $=$ Moda
Asimmetria positiva	> 0	Moda $<$ Mediana $<$ Media
Asimmetria negativa	< 0	Media $<$ Mediana $<$ Moda


Esempio asimmetria positiva, skewness = 1.89
Media
Mediana
150
100
50
0
Frequenza
0
2
4
6
Valori


Esempio asimmetria negativa, skewness = −2.52
Media
Mediana
Frequenza
200
100
0
90.0
92.5
95.0
97.5
100.0
Valori

```r
# Installiamo il pacchetto psych
install.packages("psych")

library(psych)

# Calcoliamo l'asimmetria per WorkHoursPerWeek
skew(burnout$WorkHoursPerWeek)
```

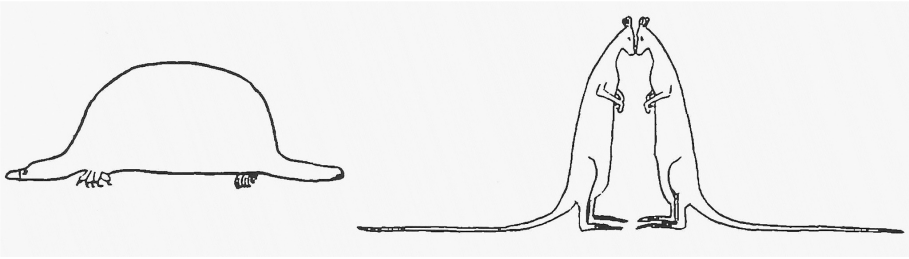
```
[1] 0.05606748
```

Interpretazione: valore vicino a 0 indica distribuzione approssimativamente simmetrica.

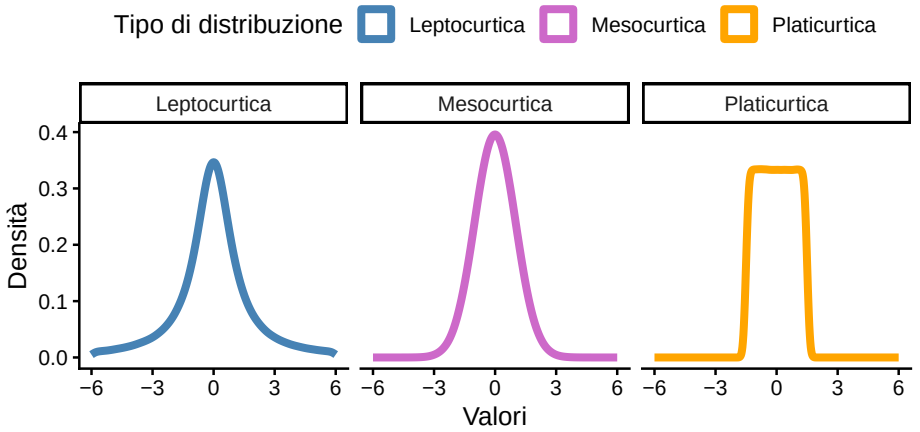
Curtosi (Kurtosis)

La curtosi misura quanto i dati si **concentrano nelle code** della distribuzione. Molti software riportano automaticamente l'**eccesso di curtosi**, che esprime questo valore in relazione alla curtosi della distribuzione normale (il cui valore di curosi è 3).

Tipo	Valore indice	Forma
Mesocurtica	≈ 0	Forma normale
Leptocurtica	> 0	code pesanti
Platicurtica	< 0	code leggere



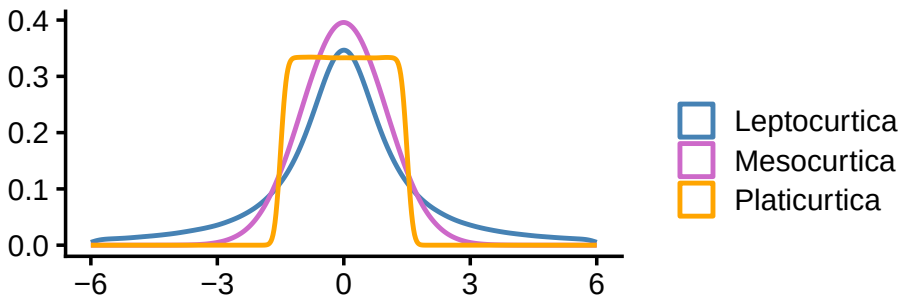
Student, 1927. Errors of routine analysis. Biometrika, 19, 151-164. p. 160


Tipo di distribuzione
Leptocurtica
Mesocurtica
Platicurtica
Leptocurtica
Mesocurtica
Platicurtica
Densità
0.4
0.3
0.2
0.1
0.0
−6
−3
0
3
6
Valori

```r
# Calcoliamo l'eccesso di curtosi per la variabile WorkHoursPerWeek
kurtosi(burnout$WorkHoursPerWeek)
```

```
[1] -1.186166
```

Che tipo di distribuzione mi aspetto che abbia questa variabile?



Confrontiamo asimmetria e curtosi di diverse variabili del dataset:

```
variabili = c("Age", "Experience",
              "WorkHoursPerWeek", "SatisfactionLevel")

risultati = data.frame(
  Variabile = variabili,
  #applico la funzione a tutte le variabili/colonne
  # via sapply(a cosa, funzione)
  Asimmetria = sapply(burnout[, variabili], skew),
  Curtosi = sapply(burnout[, variabili], kurtosi)
)
```

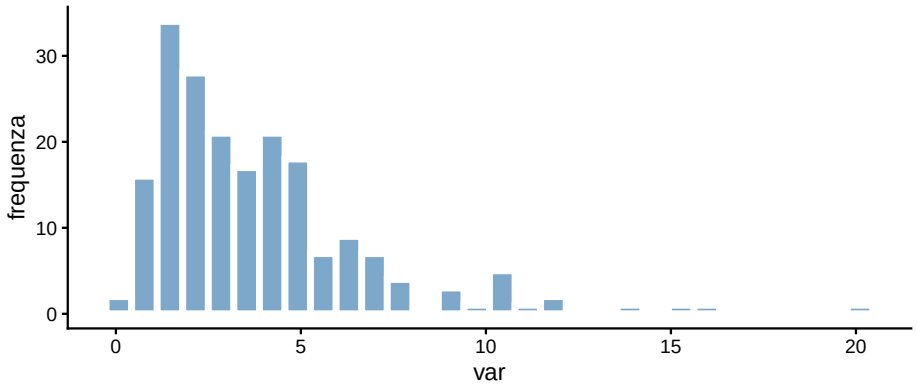
Interpretazione risultati

```
print(risultati, row.names = FALSE, digits = 2)
```

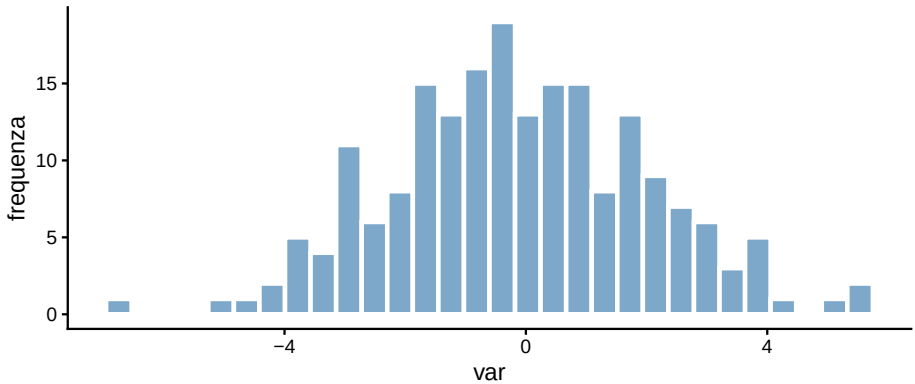
```
          Variabile Asimmetria Curtosi
                Age      0.0354   -1.20
         Experience      0.9966    0.17
  WorkHoursPerWeek      0.0561   -1.19
 SatisfactionLevel     -0.0083   -1.20
```

Domande esempio

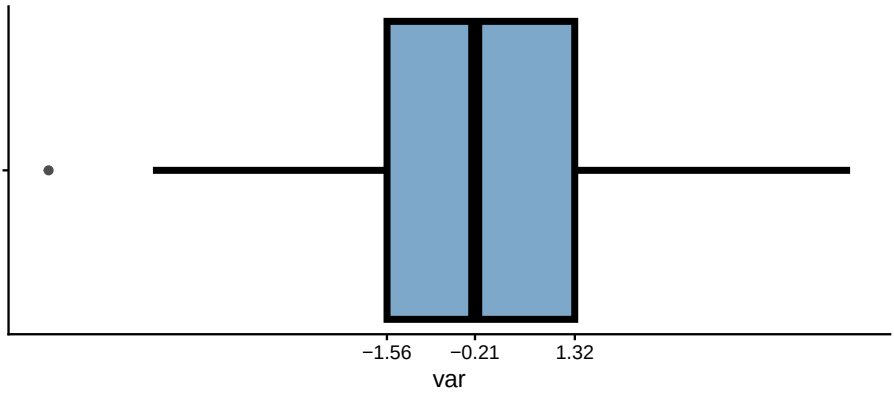
Descrivere il grafico:



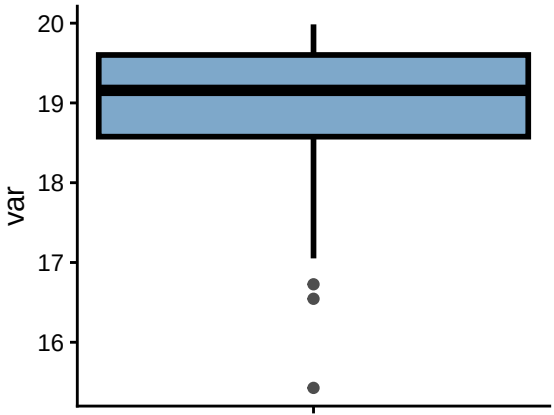
Descrivere il grafico:

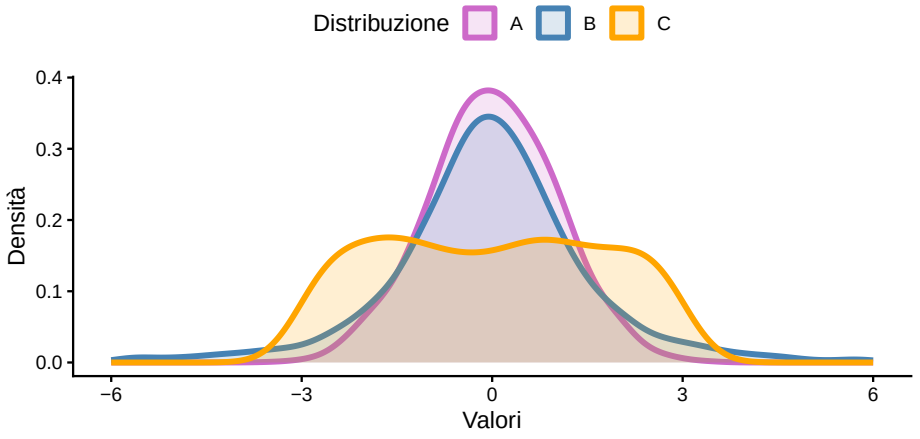


Descrivere il grafico:



Descrivere il grafico:




Distribuzione
A
B
C
Densità
0.4
0.3
0.2
0.1
0.0
−6
−3
0
3
6
Valori

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

Fonte dati