# Analisi dei Dati

## Introduzione e concetti di base. Parte 2

Domenico De Stefano

a.a. 2024/2025

https://www.geogebra.org/m/UsoH4eNl

# Indice

# Indice

# Distribuzione statistica disaggregata

Si consideri un collettivo statistico di $n$ unità, dove si sia osservata la variabile $X$. Si chiama distribuzione statistica disaggregata secondo la variabile $X$ l'insieme delle osservazioni (rappresentate da numeri o da espressioni verbali a seconda della natura della variabile) relative alle $n$ unità del collettivo (più semplicemente questi sono i cosiddetti **dati grezzi**).
In simboli, la distribuzione disaggregata sarà indicata come

$$x_1, x_2, \ldots, x_n$$

dove $x_1$ è l'osservazione relativa all'unità identificata dal numero 1, $x_2$ l'osservazione relativa all'unità identificata dal numero 2 e così via
(NB: attenzione il minuscolo non è messo a caso: la variabile in se si indica con la $X$ maiuscola, le sue **modalità** osservate sulle unità statistiche con le $x$ minuscole!)

I dati grezzi non consentono una facile visione d'insieme!

# Distribuzione di frequenza assoluta

Si consideri ancora la variabile $X$. Si chiama distribuzione di frequenza assoluta la lista delle modalità osservate di $X$ accompagnata dal numero di volte in cui queste vengono osservate, ossia accompagnata dalle rispettive frequenze assolute.

È molto facile ottenere distribuzioni di frequenza assoluta per caratteri qualitativi e quantitativi discreti. In presenza di caratteri quantitativi continui (o anche discreti, se assumono tantissime modalità), abbiamo bisogno di qualche operazione preliminare per trattarli.

# Esempio: peso alla nascita neonati

Talvolta, per variabili quantitative, è conveniente definire classi di modalità (o intervalli) contigue ed effettuare il conteggio delle unità che appartengono a ciascuna classe.

| peso | frequenza assoluta |
|---|---|
| $(2400, 2600]$ | 5 |
| $(2600, 2800]$ | 5 |
| $(2800, 3000]$ | 5 |
| $(3000, 3200]$ | 6 |
| $(3200, 3400]$ | 5 |
| $(3400, 3600]$ | 6 |

NB: la scelta delle classi è condizionata dal livello di disaggregazione con cui i dati sono stati rilevati. In altre parole è un'operazione arbitraria (decidete voi numero e ampiezza classi!) sulla base di come sono "disperse" le modalità della variabile in questione

# Classi di differenti lunghezze

Può capitare, o per scelta (si vuole fornire informazioni più dettagliate su parte della distribuzione),
o per necessità (quando i dati sono già stati raggruppati in classi da qualcuno... nel caso ad es. delle classi di età in cui talvolta le classi estreme sono lasciate aperte usando le paroline "...e oltre", es. 20–39; 40–59; 60–79; 80 e oltre),
di costruire delle classi utilizzando intervalli di lunghezza differente.

In questo caso è conveniente definire anche la densità di frequenza.

La densità è definita come:

$$\left(\begin{array}{c}\text{densità}\\ \text{di una classe}\end{array}\right) = \frac{\text{frequenza assoluta di } Y \text{ sull'intervallo}}{\text{lunghezza dell'intervallo}}.$$

Per capire la definizione si pensi alla popolazione. E' la densità della popolazione non il numero totale di abitanti che ci dice quanto gli individui sono *addensati* in una certa regione geografica.

# Esempio classi di diversa ampiezza

| peso | frequenza assoluta | densità |
|---|---|---|
| $(2400, 2600]$ | 5 | $5/200{=}0.025$ |
| $(2600, 2800]$ | 5 | $5/200{=}0.025$ |
| $(2800, 3000]$ | 5 | $5/200{=}0.025$ |
| $(3000, 3200]$ | 6 | $6/200{=}0.030$ |
| $(3200, 3600]$ | 11 | $11/400{=}0.0275$ |

La densità ci dice il numero atteso di unità statistiche per ogni unità di misura della variabile. Nella prima classe, per esempio, ci aspettiamo di osservare 2,5 neonati ogni 100 grammi di peso (ovvero, 2,5 neonati con peso tra 2400 e 2500 e 2,5 neonati con peso tra 2500 e 2600).

# Esempio: distribuzione di frequenza per gruppi

Peso alla nascita da madri non fumatrici e da madri fumatrici.

| Fumo=N | |
|---|---|
| durata | frequenza assoluta |
| $(2400, 2600]$ | 2 |
| $(2600, 2800]$ | 2 |
| $(2800, 3000]$ | 2 |
| $(3000, 3200]$ | 3 |
| $(3200, 3400]$ | 3 |
| $(3400, 3600]$ | 4 |

| Fumo=S | |
|---|---|
| durata | frequenza assoluta |
| $(2400, 2600]$ | 3 |
| $(2600, 2800]$ | 3 |
| $(2800, 3000]$ | 3 |
| $(3000, 3200]$ | 3 |
| $(3200, 3400]$ | 2 |
| $(3400, 3600]$ | 2 |

# Frequenze relative

Dividendo una frequenza assoluta per il numero totale di unità statistiche nel collettivo analizzato ($n$ nel nostro caso) otteniamo le cosiddette frequenze relative (o proporzioni), ovvero

$$\left(\begin{array}{c}\text{frequenze}\\ \text{relative}\end{array}\right) = \frac{\left(\begin{array}{c}\text{frequenze}\\ \text{assolute}\end{array}\right)}{\left(\begin{array}{c}\text{numero totale di}\\ \text{osservazioni}\end{array}\right)}$$

Hanno il vantaggio, rispetto alle frequenze assolute, di permettere di confrontare distribuzioni di frequenza basate su numeri differenti di unità statistiche.

# Esempio: effetti del fumo sul peso dei neonati

| peso | frequenza relativa |
|---|---|
| $(2400, 2600]$ | $5/32 = 0.15625$ |
| $(2600, 2800]$ | $5/32 = 0.15625$ |
| $(2800, 3000]$ | $5/32 = 0.15625$ |
| $(3000, 3200]$ | $6/32 = 0.18750$ |
| $(3200, 3400]$ | $5/32 = 0.15625$ |
| $(3400, 3600]$ | $6/32 = 0.18750$ |

# Esercizio: esiti ammissione a Berkeley, 1973

I seguenti dati rappresentano gli esiti dell'ammissione all'Università di California, Berkeley (USA) nel 1973. È riportato l'esito dell'ammissione ( Admit), il sesso dei candidati ( Gender) e il Dipartimento erogante il corso di studi scelto dai candidati ( Dept).

| Admit | Gender | Dept | Frequenza assoluta |
|---|---|---|---|
| Admitted | Male | A | 512 |
| Rejected | Male | A | 313 |
| Admitted | Female | A | 89 |
| Rejected | Female | A | 19 |
| Admitted | Male | B | 353 |
| Rejected | Male | B | 207 |
| Admitted | Female | B | 17 |
| Rejected | Female | B | 8 |
| Admitted | Male | C | 120 |
| Rejected | Male | C | 205 |
| Admitted | Female | C | 202 |
| Rejected | Female | C | 391 |
| Admitted | Male | D | 138 |
| Rejected | Male | D | 279 |
| Admitted | Female | D | 131 |
| Rejected | Female | D | 244 |
| Admitted | Male | E | 53 |
| Rejected | Male | E | 138 |
| Admitted | Female | E | 94 |
| Rejected | Female | E | 299 |
| Admitted | Male | F | 22 |
| Rejected | Male | F | 351 |
| Admitted | Female | F | 24 |
| Rejected | Female | F | 317 |

È una matrice dei dati? Quante sono le variabili rilevate? Di che tipo sono? Quante sono le unità statistiche?

# Frequenze cumulate

- La frequenza cumulata ha senso se la variabile $X$ è almeno ordinata, quindi

$$x_1 < x_2 < \ldots < x_k$$

- La frequenza assoluta (o anche relativa, perchè no?) cumulata per la modalità/classe $x_i$ è la somma delle frequenze assolute (relative) per le modalità/classi $\leq x_i$

$$F_i = f_1 + \ldots + f_i = \sum_{h=1}^{i} f_h$$

| modalità/classe | frequenze cumulate assolute | frequenze cumulate relative |
|---|---|---|
| $x_1$ | $n_1$ | $F_1 = f_1$ |
| $x_2$ | $n_1 + n_2$ | $F_2 = f_1 + f_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $x_i$ | $n_1 + \ldots + n_i$ | $F_i = f_1 + \ldots + f_i$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $x_k$ | $n$ | ? |

# Esercizio: dataset babies

Si costruisca la distribuzione di frequenze cumulate per la durata della gravidanza nel dataset babies (v. slides precedenti).

Partendo dalla distribuzione di frequenze assolute, abbiamo

| durata | frequenza assoluta | frequenza cumulata |
|---|---|---|
| 34 | 1 | 1 |
| 35 | 3 | 4 |
| 36 | 3 | 7 |
| 37 | 2 | 9 |
| 38 | 5 | 14 |
| 39 | 7 | 21 |
| 40 | 3 | 24 |
| 41 | 3 | 27 |
| 42 | 5 | 32 |

# Indice

# Funzione di ripartizione empirica

La distribuzione di frequenze relative cumulate è collegata ad una importante rappresentazione dell'andamento di una variabile quantitativa, ossia la *funzione di ripartizione empirica*.

$$\left(\begin{array}{c}\text{funzione di}\\ \text{ripartizione empirica}\\ \text{calcolata in } y\end{array}\right) = \frac{\left(\begin{array}{c}\text{numero di}\\ \text{osservazioni minori o}\\ \text{uguali a } y\end{array}\right)}{\left(\begin{array}{c}\text{numero totale di}\\ \text{osservazioni}\end{array}\right)}$$

# Funzione di ripartizione empirica

Formalizzando, detto $Y$ il carattere oggetto di studio, la funzione di ripartizione empirica calcolata a partire dal campione $(y_1, y_2, \ldots, y_N)$ è la funzione

$$F_Y(y) = \frac{\#\{Y \leq y\}}{n}.$$

Il dominio della funzione è dato da $\mathbb{R}$; il codominio è l'intervallo $[0, 1]$.

# Esempio: dataset babies

Si costruisca la funzione di ripartizione empirica per la durata della gravidanza nel dataset babies (v. Lezione 3).

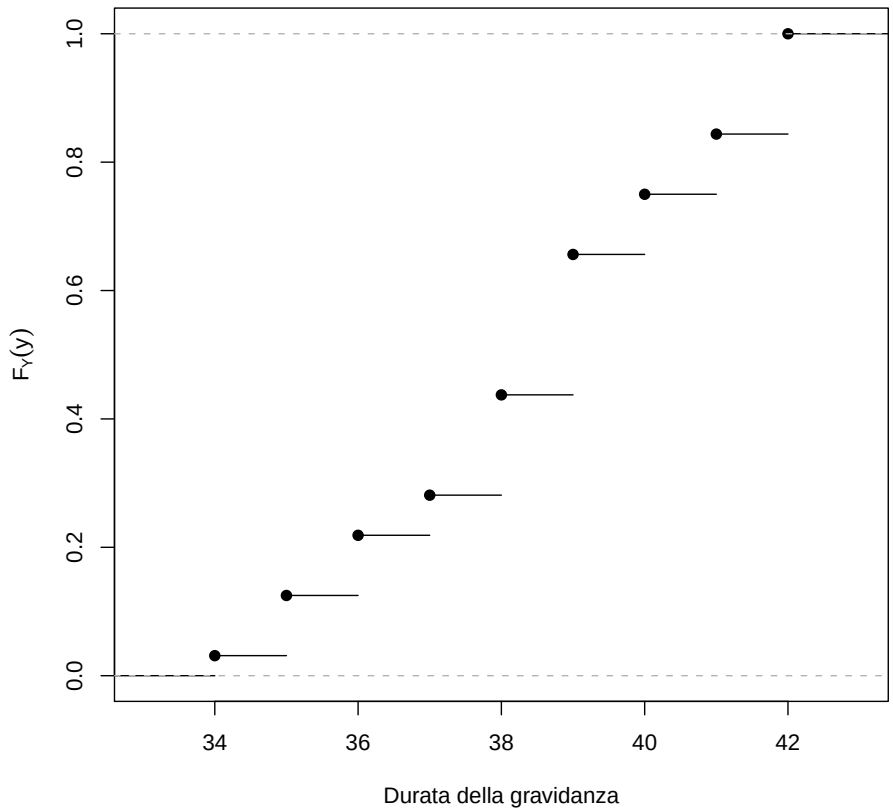
Partendo dalla distribuzione di frequenze relative cumulate, la funzione è così definita.

$$F_Y(y) = \begin{pmatrix} 0 & y < 34 \\ 1/32 & 34 \leq y < 35 \\ 4/32 & 35 \leq y < 36 \\ 7/32 & 36 \leq y < 37 \\ 9/32 & 37 \leq y < 38 \\ 14/32 & 38 \leq y < 39 \\ 21/32 & 39 \leq y < 40 \\ 24/32 & 40 \leq y < 41 \\ 27/32 & 41 \leq y < 42 \\ 1 & y \geq 42 \end{pmatrix}$$

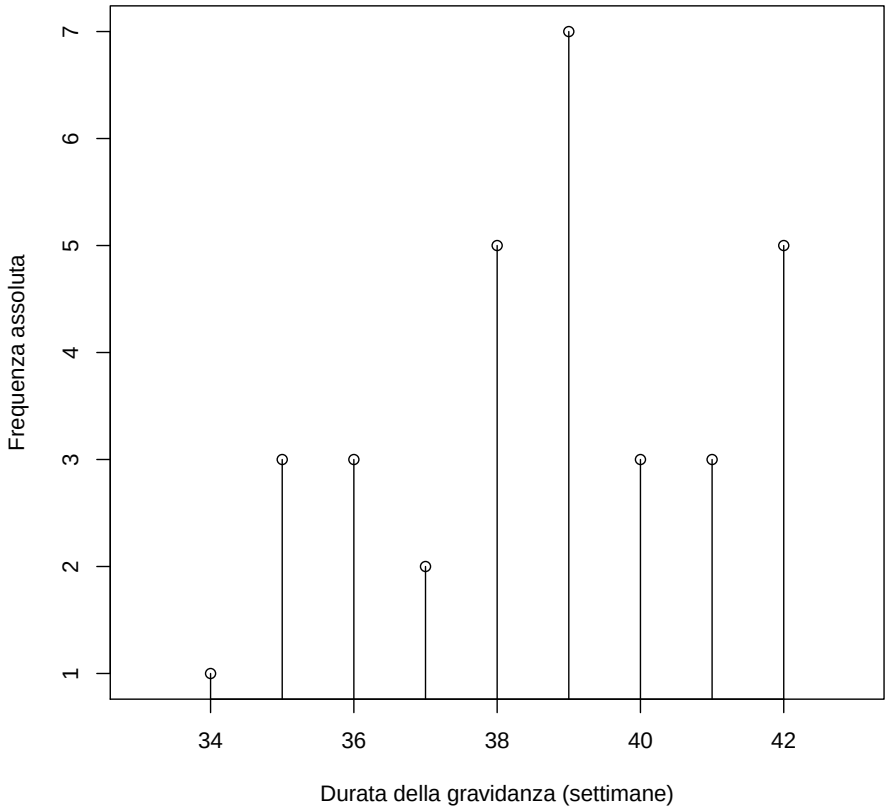
# Funzione di ripartizione empirica: rappresentazione grafica

La funzione di ripartizione empirica può essere rappresentata graficamente.

Esempio: dataset `babies`, durata della gravidanza



# Esempio: dataset `babies`

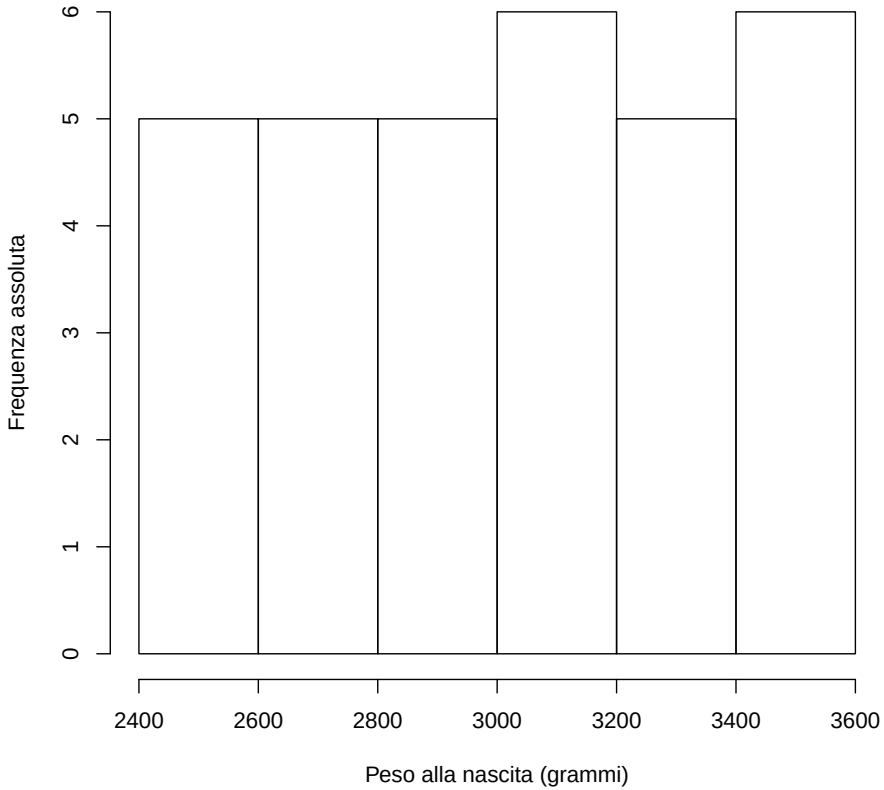



Il grafico è stato costruito ponendo

$$\text{asse}x = \begin{pmatrix}\text{modalità riportate}\\ \text{nella distribuzione}\\ \text{di frequenza}\end{pmatrix}$$

$$(\text{altezza barre}) = (\text{frequenze assolute})$$

# Esempio: dataset babies



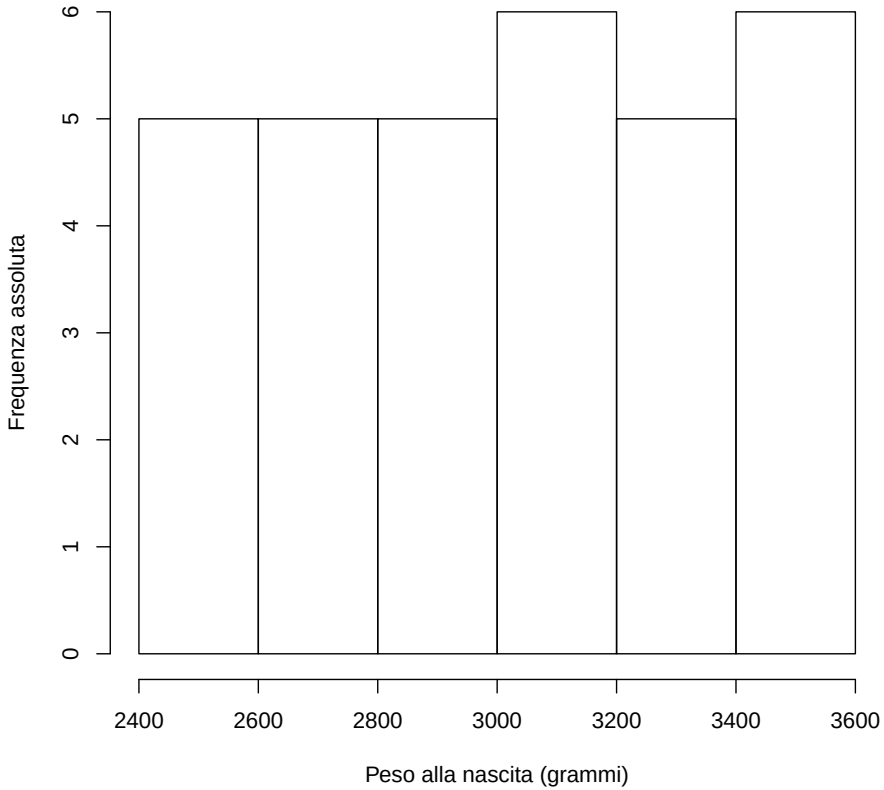

Il grafico è stato costruito ponendo

$$(\text{base rettangoli}) = \begin{pmatrix} \text{intervallini riportati} \\ \text{nella } 1^\circ \text{ colonna} \\ \text{della distribuzione} \\ \text{di frequenza} \end{pmatrix}$$

$$(\text{area rettangoli}) \propto (\text{frequenze assolute})$$

Il simbolo $\propto$ significa "proporzionale a".

# Esempio: dataset babies




Il grafico è stato costruito ponendo

$$(\text{base rettangoli}) = \begin{pmatrix} \text{intervallini riportati} \\ \text{nella } 1^\circ \text{ colonna} \\ \text{della distribuzione} \\ \text{di frequenza} \end{pmatrix}$$

$$(\text{area rettangoli}) \propto (\text{frequenze assolute})$$

Il simbolo $\propto$ significa "proporzionale a".

Essendo l'area dei rettangoli uguale a base$\times$altezza, se le gli intervalli hanno uguale ampiezza, di fatto l'altezza coincide con (o è proporzionale a) la frequenza assoluta:

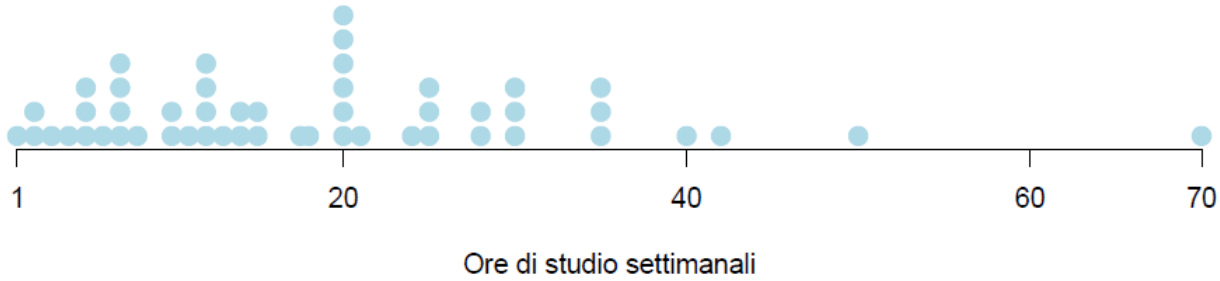
$$(\text{altezza rettangoli}) = (\text{frequenze assolute})$$

# Indice

# Indice

# Indici di posizione

Esempio:Ore di studio per settimana.

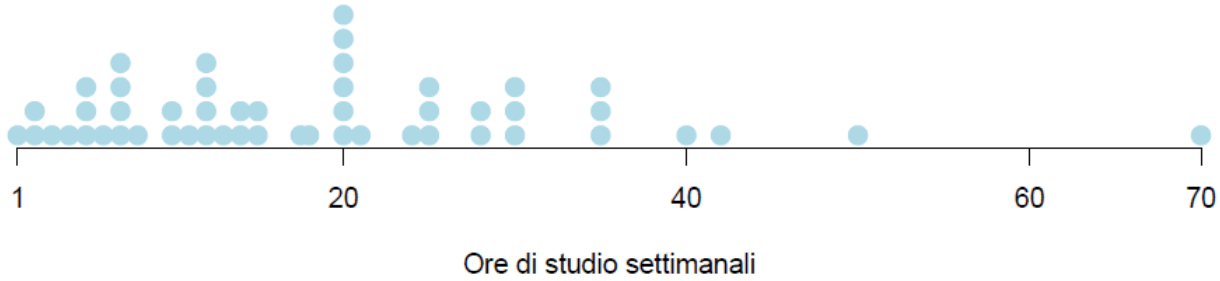



Sapendo che ogni pallino rappresenta una unità statistica... Come descrivereste questa distribuzione? Qual è il valore più frequente? Intorno a quale valore possiamo dire che è posizionata la distribuzione? In altre parole, dove è il centro della distribuzione?

# Indici di posizione

La domanda precedente ci chiede di sintetizzare la distribuzione in un unico numero che, in una qualche senso, indichi dove la distribuzione stessa è "posizionata".

Si potrebbe dire che la distribuzione è posizionata sul valore che compare più frequentemente.




Questo valore è chiamato *moda* della distribuzione.

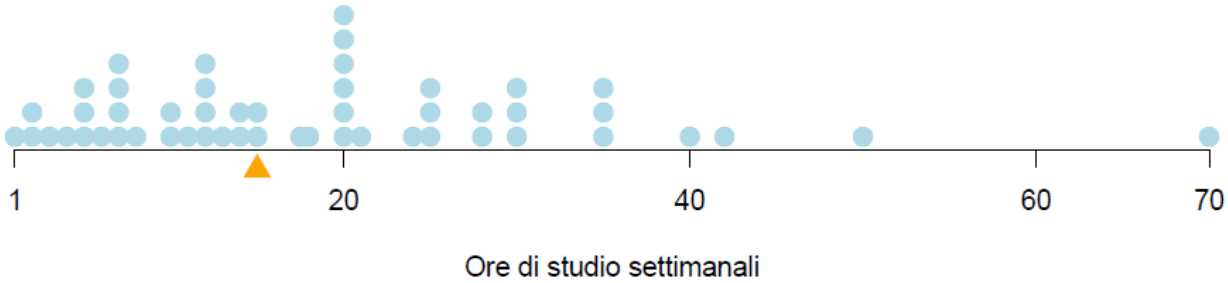
# Indici di posizione

La *moda* di una distribuzione è il valore del supporto cui è associata la più grande frequenza relativa.

- La moda esprime la modalità più comune.
- È definita anche per variabili qualitative (lo ricorderemo a tempo debito).

# Indici di posizione

Ma il centro di una distribuzione potrebbe anche essere pensato come quel valore che lascia alla sua destra ed alla sua sinistra esattamente il 50% delle osservazioni.



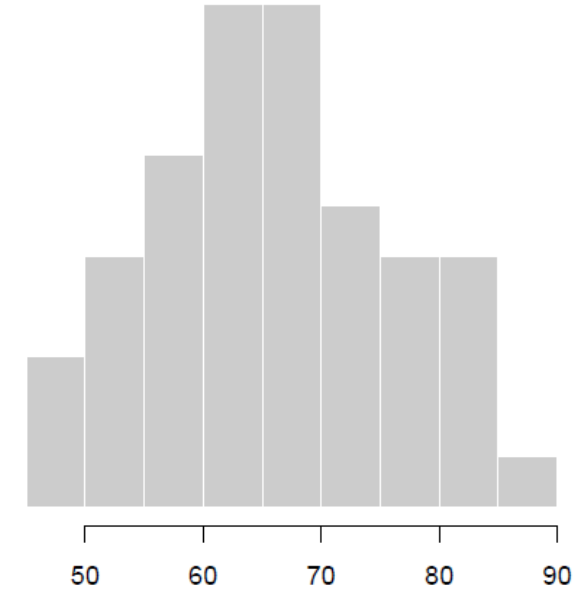
# Indici di posizione

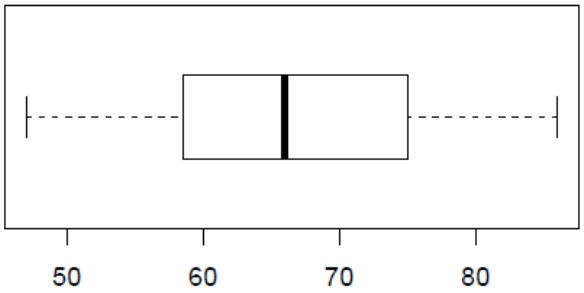
- La *media aritmetica*, indicata con $\bar{x}$, è calcolata come:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N} \sum_{i=1}^{N} x_i,$$

  dove $(x_1, x_2, \cdots, x_N)$ rappresenta la distribuzione disaggregata dei valori osservati per $X$ sulle $N$ unità statistiche del nostro collettivo
- Esistono altri tipi di "medie". Quella aritmetica è senza ogni dubbio quella di utilizzo più comune. Per questo motivo, viene comunemente indicata come "la media" senza nessuna ulteriore aggettivazione.

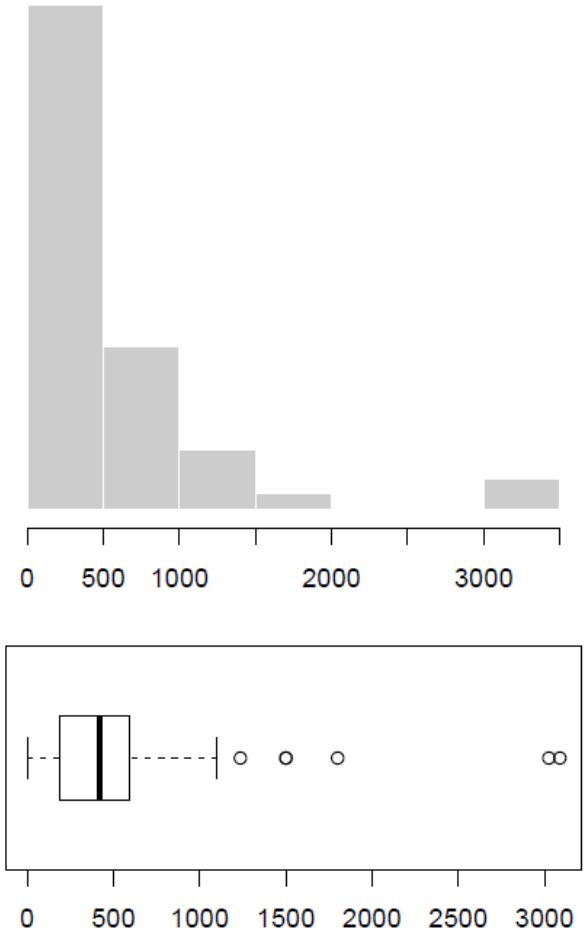
# Indici di posizione/ esempi



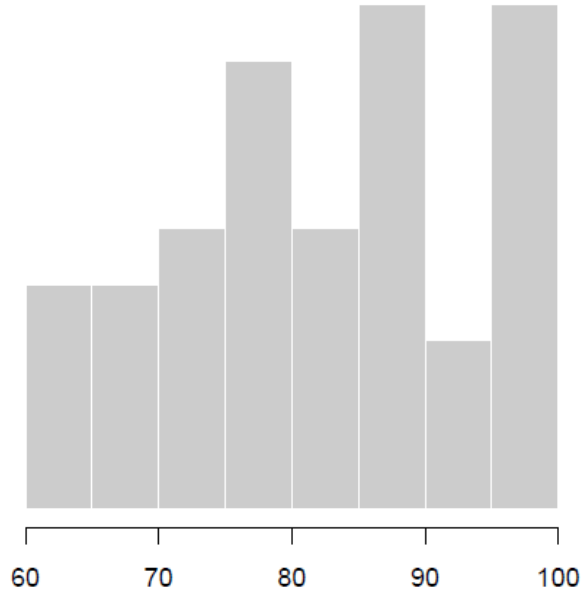

| | |
|---|---|
| Media: | 66.8 |
| Mediana: | 66 |
| Varianza: | 107 |
| Deviaz. Standard: | 10.4 |

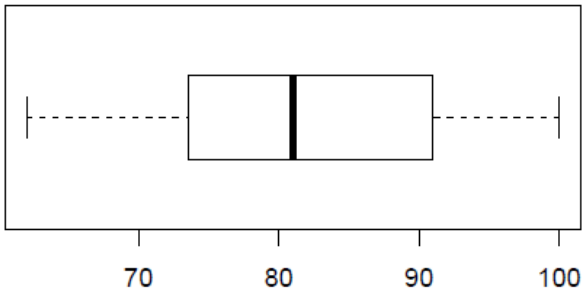
# Indici di posizione/ esempi




| | |
|---|---|
| Media: | 573 |
| Mediana: | 419 |
| Varianza: | $3.95 \times 10^5$ |
| Deviaz. Standard: | 629 |

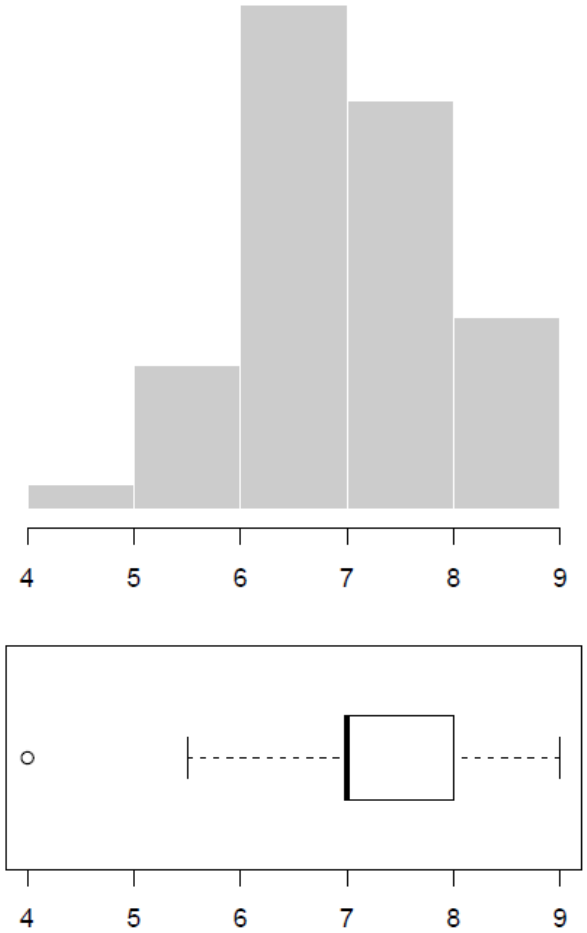
# Indici di posizione/ esempi




| | |
|---|---|
| Media: | 83 |
| Mediana: | 81 |
| Varianza: | 127 |
| Deviaz. Standard: | 11.3 |

# Indici di posizione/ esempi




| | |
|---|---|
| Media: | 7.43 |
| Mediana: | 7 |
| Varianza: | 1.03 |
| Deviaz. Standard: | 1.01 |

# Indici di posizione

- La *moda*, *mediana* e la *media aritmetica* sono tutte misure di posizione.
- Se lavoriamo sull'intera popolazione (abbiamo cioè un censimento), le misure vengono chiamate *di popolazione* (è tradizione indicarle con simboli diversi, spesso lettere greche). Come abbiamo detto, è raro lavorare con l'intera popolazione.
- Se lavoriamo con un campione, come è quasi sempre il caso, le misure vengono dette *campionarie*. Se il campione è rappresentativo, in generale le misure campionarie sono buone "indicazioni" delle misure calcolate sulla intera popolazione.

# Indice

# Trasformazioni di dati: la trasformazione lineare

Spesso per vari motivi (ad es. spesso per cambiare unità di misura oppure a causa di marcate asimmetrie nella distribuzione di una variabile) servirà trasformare i valori originari di una variabile quantitativa $X$ mediante una funzione $g(x)$ opportuna.
Una trasformazione particolarmente importante è la trasformazione lineare, ovvero la trasformazione del tipo: $g(x) = a + bx$

Esempio: Temperatura in gradi Farenheit e Celsius.

$$F^\circ = C^\circ 1,8 + 32$$

# Trasformazioni lineari: esempio notevole

*Standardizzazione*

$(x_1, \ldots, x_N)$ dati grezzi, con media $\bar{x}$ e deviazione standard $\sigma$

# Trasformazioni lineari: esempio notevole

## *Standardizzazione*

$(x_1, \ldots, x_N)$ dati grezzi, con media $\bar{x}$ e deviazione standard $\sigma$
$(z_1, \ldots, z_N)$ dati *standardizzati*, ottenuti come

$$z_i = a + bx_i = -\frac{\bar{x}}{\sigma} + \frac{1}{\sigma}x_i.$$

# Trasformazioni lineari: esempio notevole

*Standardizzazione*

$(x_1, \ldots, x_N)$ dati grezzi, con media $\bar{x}$ e deviazione standard $\sigma$
$(z_1, \ldots, z_N)$ dati *standardizzati*, ottenuti come

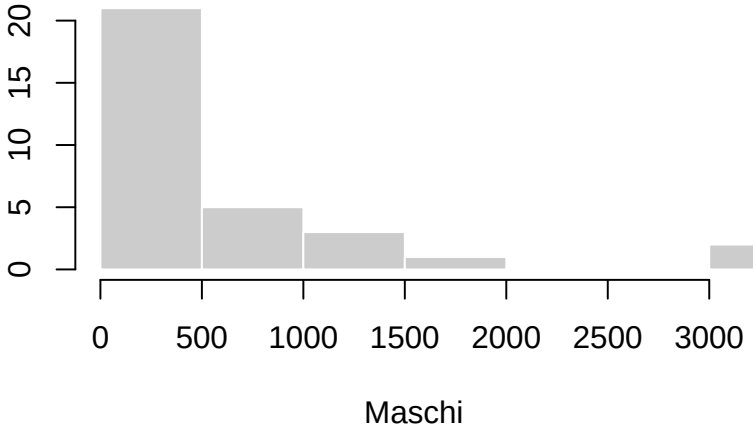
$$z_i = a + bx_i = -\frac{\bar{x}}{\sigma} + \frac{1}{\sigma}x_i.$$

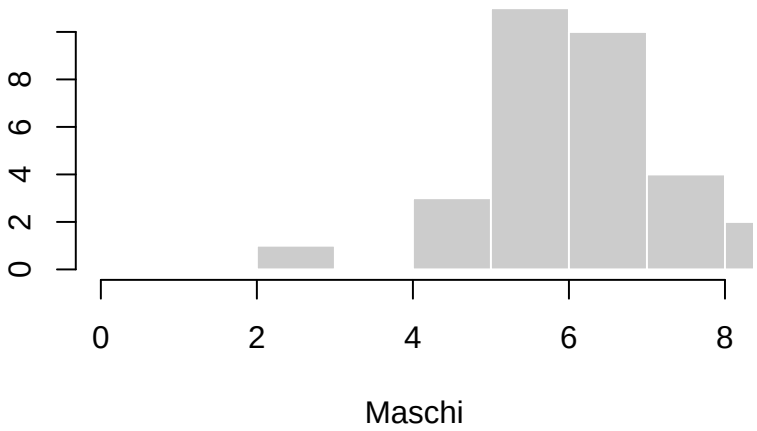
Questa trasformazione è molto usata in statistica (sarà chiaro il perché in seguito).
È facile verificare che la trasformazione può essere scritta anche così:

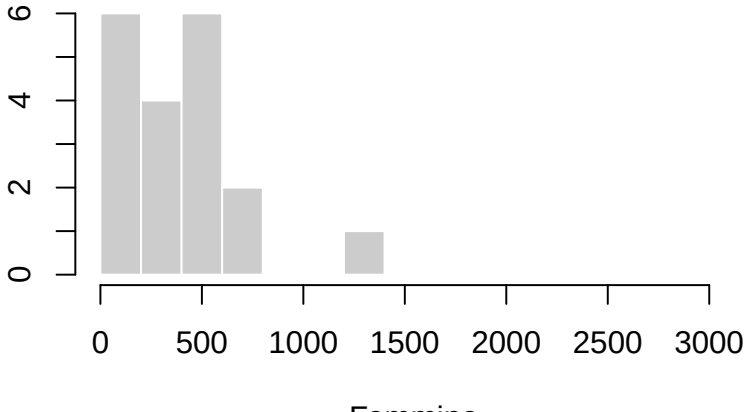
$$z_i = \frac{x_i - \bar{x}}{\sigma}.$$

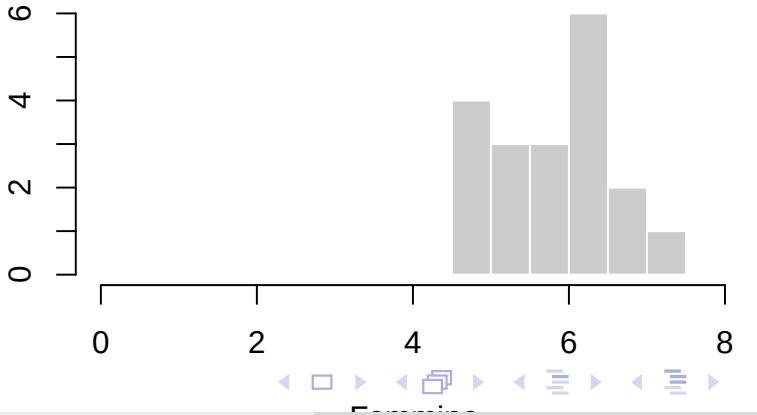
# Trasformazione logaritmica: esempio degli Amici di Facebook

Quando i dati sono distribuiti in modo fortemente asimmetrico, una trasformazione potrebbe consentire di capirne meglio la distribuzione. Una trasformazione comune è il *logaritmo*.



# Pro e contro

- Outliers tendono ad avvicinarsi:

| | | | | |
|---|---|---|---|---|
| # amici di Facebook | 11 | 589 | 3025 | $\cdots$ |
| log(# Amici di Facebook) | 2.40 | 6.38 | 8.01 | $\cdots$ |

- ma i risultati potrebbero essere difficili da interpretare, perché lavoriamo su una scala (quella logaritmica, nello specifico), che non è quella naturale;
- e, in generale, non è facile ricondurre i risultati di sintesi alla scala naturale.

# Pro e contro (cont)

Esempio: Amici di Facebook

$(y_1, \ldots, y_N)$ Amici di Facebook

# Pro e contro (cont)

Esempio: Amici di Facebook
$(y_1, \ldots, y_N)$ Amici di Facebook
$(t_1, \ldots, t_N)$ Logaritmi del numero di amici di Facebook, cioè
$t_i = \log(y_i), \ i = 1, \ldots, N.$

# Pro e contro (cont)

Esempio: Amici di Facebook
$(y_1, \ldots, y_N)$ Amici di Facebook
$(t_1, \ldots, t_N)$ Logaritmi del numero di amici di Facebook, cioè
$t_i = \log(y_i), \ i = 1, \ldots, N.$

Si ha
$\bar{y} = 572.63$

# Pro e contro (cont)

Esempio: Amici di Facebook
$(y_1, \ldots, y_N)$ Amici di Facebook
$(t_1, \ldots, t_N)$ Logaritmi del numero di amici di Facebook, cioè
$t_i = \log(y_i), \ i = 1, \ldots, N.$

Si ha
$\bar{y} = 572.63$
$\bar{t} = 5.82$

# Pro e contro (cont)

Esempio: Amici di Facebook
$(y_1, \ldots, y_N)$ Amici di Facebook
$(t_1, \ldots, t_N)$ Logaritmi del numero di amici di Facebook, cioè
$t_i = \log(y_i), \ i = 1, \ldots, N.$

Si ha
$\bar{y} = 572.63$
$\bar{t} = 5.82$

NB:
$\bar{y} \neq \exp(\bar{t})$

# Pro e contro (cont)

Esempio: Amici di Facebook
$(y_1, \ldots, y_N)$ Amici di Facebook
$(t_1, \ldots, t_N)$ Logaritmi del numero di amici di Facebook, cioè
$t_i = \log(y_i), \ i = 1, \ldots, N.$

Si ha
$\bar{y} = 572.63$
$\bar{t} = 5.82$

NB:
$\bar{y} \neq \exp(\bar{t})$
$\bar{t} \neq \log(\bar{y})$

# Pro e contro (cont)

Esempio: Amici di Facebook
$(y_1, \ldots, y_N)$ Amici di Facebook
$(t_1, \ldots, t_N)$ Logaritmi del numero di amici di Facebook, cioè
$t_i = \log(y_i),\ i = 1, \ldots, N.$

Si ha
$\bar{y} = 572.63$
$\bar{t} = 5.82$

NB:
$\bar{y} \neq \exp(\bar{t})$
$\bar{t} \neq \log(\bar{y})$
Infatti
$\exp(\bar{t}) = 337$

# Pro e contro (cont)

Esempio: Amici di Facebook
$(y_1, \ldots, y_N)$ Amici di Facebook
$(t_1, \ldots, t_N)$ Logaritmi del numero di amici di Facebook, cioè
$t_i = \log(y_i),\ i = 1, \ldots, N.$

Si ha
$\bar{y} = 572.63$
$\bar{t} = 5.82$

NB:
$\bar{y} \neq \exp(\bar{t})$
$\bar{t} \neq \log(\bar{y})$
Infatti
$\exp(\bar{t}) = 337$
$\log(\bar{y}) = 6.35$

# Trasformazioni monotone

La trasformazione logaritmica è una trasformazione strettamente monotona.

Richiamo: una funzione $g(x)$ si dice strettamente monotona in un intervallo aperto $A$, se, comunque si scelgano $x_1$ e $x_2$ in $A$ si ha

- se $x_1 < x_2$ allora $g(x_1) < g(x_2)$ (strettamente crescente)
- se $x_1 < x_2$ allora $g(x_1) > g(x_2)$ (strettamente decrescente)

# Indice

# Varianza

La *varianza* è la media dei quadrati degli scarti di ogni osservazione dalla media aritmetica.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}$$

Esempio: ore di studio per settimana

- La media è $\bar{x} = 18.58$ .
- La varianza è calcolata come:

$$\sigma^2 = \frac{(2 - 18.58)^2 + (30 - 18.58)^2 + \cdots + (42 - 18.58)^2}{51} = 183.12$$