Statistica

Indici di posizione, variabilità, ecc.

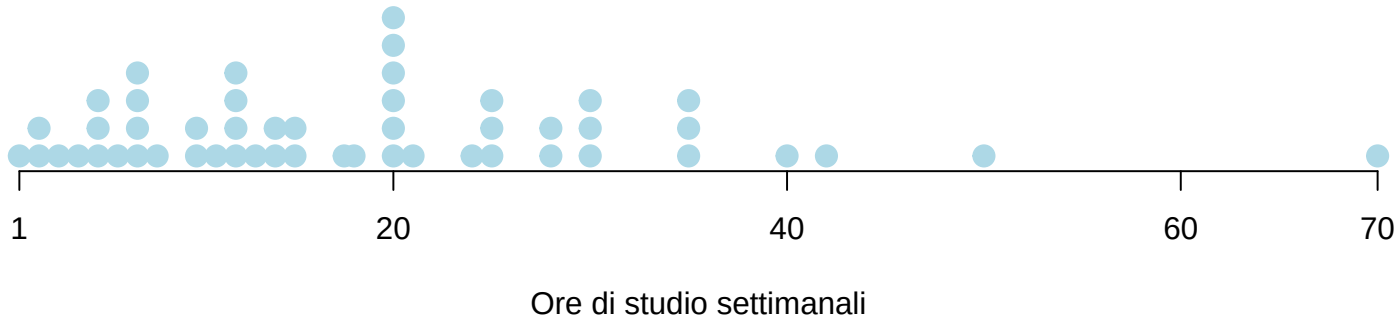
Domenico De Stefano

a.a. 2024/2025

Indice

Diagramma a barre (*o a punti*)

Esempio:Ore di studio per settimana.

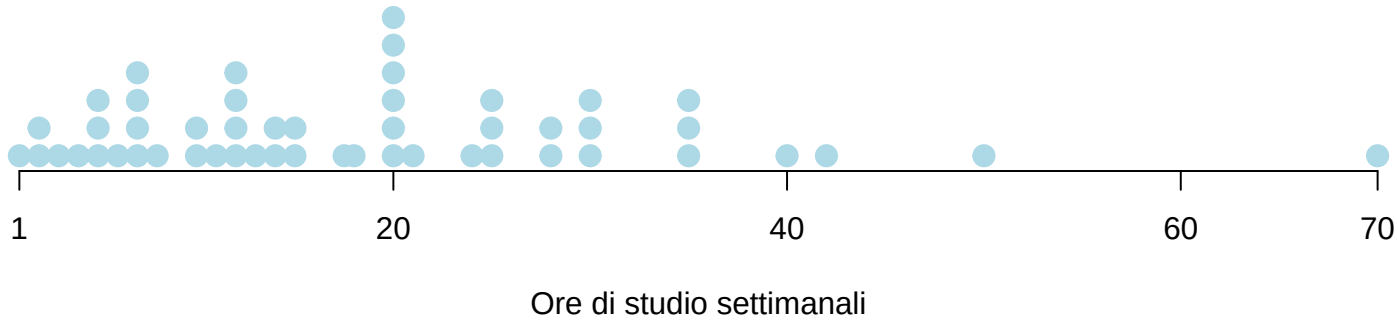


Sapendo che ogni pallino rappresenta una unità statistica... Come descrivereste questa distribuzione? Qual è il valore più frequente? Intorno a quale valore possiamo dire che è posizionata la distribuzione? In altre parole, dove è il centro della distribuzione?

"Posizione" della distribuzione

La domanda precedente ci chiede di sintetizzare la distribuzione in un unico numero che, in un 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.

Misure di posizione: la moda

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.
- Non è detto che sia unica. Una distribuzione è detta *unimodale* se presenta un unico massimo locale; altrimenti è detta *multimodale*.
- È definita anche per variabili qualitative (lo ricorderemo a tempo debito).

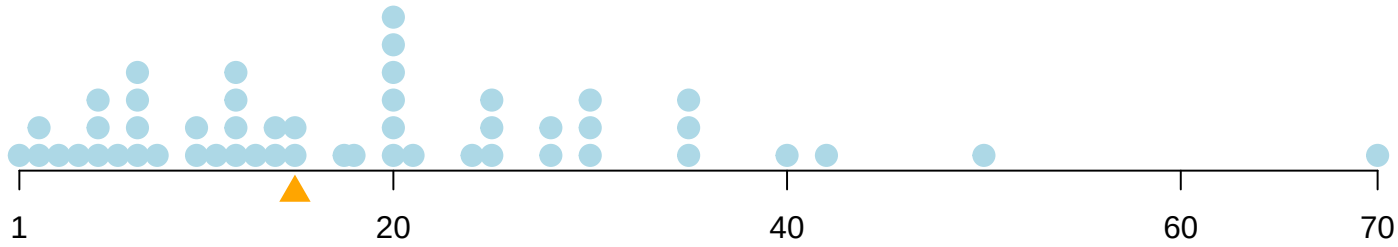
Misure di posizione (cont)

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.

1 2 2 3 4 5 5 5 6 7
7 7 7 8 10 10 11 12 12 12
12 13 14 14 15 15 17.5 18 20 20
20 20 20 20 21 24 25 25 25 28
28 30 30 30 35 35 35 40 42 50
70

Misure di posizione (cont)

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.



Ore di studio settimanali

Misure di posizione: la mediana

Sia $x_1, x_2, \cdots, x_N$ una distribuzione statistica disaggregata.
Sia $x_{(1)}, x_{(2)}, \cdots, x_{(N)}$ la corrispondente distribuzione dei valori ordinati:

- $x_{(1)} = \min(x_1, \ldots, x_N), \quad x_{(N)} = \max(x_1, \ldots, x_N)$;
- $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(N)}$.

La *mediana*, indicata con m, è calcolata come:

$$m = \begin{cases} x_{(N+1)/2} & \text{se } N \text{ dispari} \\ \frac{x_{(N/2)} + x_{(N/2+1)}}{2} & \text{se } N \text{ pari} \end{cases}$$

La mediana è un particolare *quantile*.

Quantili

- Il *quantile* di livello α, indicato con q_α, definito per $0 \leq \alpha \leq 1$, è quel valore che lascia alla sua sinistra una frazione $\alpha\%$ dei dati q_α e una frazione $(1-\alpha)\%$ alla sua destra.

Quantili

- Il *quantile* di livello α, indicato con q_α, definito per $0 \leq \alpha \leq 1$, è quel valore che lascia alla sua sinistra una frazione $\alpha\%$ dei dati q_α e una frazione $(1-\alpha)\%$ alla sua destra.
- La mediana, quindi, è il quantile di livello 0.5, cioè $m = q_{0.5}$.

Quantili

- Il *quantile* di livello α, indicato con q_α, definito per $0 \leq \alpha \leq 1$, è quel valore che lascia alla sua sinistra una frazione $\alpha\%$ dei dati q_α e una frazione $(1-\alpha)\%$ alla sua destra.
- La mediana, quindi, è il quantile di livello 0.5, cioè $m = q_{0.5}$.
- Tra i quantili diversi dalla mediana, $q_{0.25}$ e $q_{0.75}$ sono i più usati, perché basati su una divisione in quarti del collettivo. Sono chiamati *primo quartile* e *terzo quartile*, rispettivamente (la mediana è, di fatto, il *secondo quartile*).
- I quartili sono dunque particolari quantili e saranno quelli che utilizzeremo più spesso per descrivere i nostri dati
- ... ma esistono diversi tipi di quantili come per esempio i percentili. Considerando i percentili la mediana sarà il cinquantesimo percentile, il primo quartile il 25-mo e il terzo il 75-mo percentile

Esempio: altezza

Si calcolino $q_{0.25}$, m e $q_{0.75}$ per la variabile altezza.

Partiamo dai dati grezzi.

```
180 173 170 168 172 185 175 170 176 183
176 181 185 188 180 173 170 187 165 190
187 182 175 183 166 186 190 181 185 170
178 180 160 174 180 184 183 180 175 182
175 160 180 176 178 164 177 170 184 173
173 164 NA 176
```

Esempio: altezza

Ordinando i valori in senso crescente, abbiamo

160 160 164 164 165 166 168 170 170 170
170 170 172 173 173 173 173 174 175 175
175 175 176 176 176 176 177 178 178 180
180 180 180 180 180 181 181 182 182 183
183 183 184 184 185 185 185 186 187 187
188 190 190

Esempio: altezza

Ordinando i valori in senso crescente, abbiamo

160 160 164 164 165 166 168 170 170 170
170 170 172 173 173 173 173 174 175 175
175 175 176 176 176 176 177 178 178 180
180 180 180 180 180 181 181 182 182 183
183 183 184 184 185 185 185 186 187 187
188 190 190

Abbiamo $N = 53$. Quindi $m = x_{(27)} = 177$.

Esempio: altezza

Ordinando i valori in senso crescente, abbiamo

160 160 164 164 165 166 168 170 170 170
170 170 172 173 173 173 173 174 175 175
175 175 176 176 176 176 177 178 178 180
180 180 180 180 180 181 181 182 182 183
183 183 184 184 185 185 185 186 187 187
188 190 190

Abbiamo $N = 53$. Quindi $m = x_{(27)} = 177$.

$q_{0.25}$ è di fatto la mediana di $x_{(1)}, x_{(2)}, \cdots, x_{(27)}$ cioè è $x_{(14)} = 173$.

Misure di posizione: la media aritmetica

- 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.

Esempio: altezze

Si calcoli la media delle altezze (ossia l'altezza media).

Partiamo dai dati grezzi (cioè la distribuzione disaggregata della variabile altezza).

```
180 173 170 168 172 185 175 170 176 183
176 181 185 188 180 173 170 187 165 190
187 182 175 183 166 186 190 181 185 170
178 180 160 174 180 184 183 180 175 182
175 160 180 176 178 164 177 170 184 173
173 164 NA 176
```

Esempio: altezze

Si calcoli la media delle altezze (ossia l'altezza media).

Partiamo dai dati grezzi (cioè la distribuzione disaggregata della variabile altezza).

```
180 173 170 168 172 185 175 170 176 183
176 181 185 188 180 173 170 187 165 190
187 182 175 183 166 186 190 181 185 170
178 180 160 174 180 184 183 180 175 182
175 160 180 176 178 164 177 170 184 173
173 164 NA 176
```

Abbiamo $N = 53$. Quindi:

$$\frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}\sum_{i=1}^{N} x_{(i)} = \frac{9378}{53} = 177.$$

Riassumendo

- 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.

NB: misure marginali e condizionate

Le misure di posizione per variabili condizionate vengono, per semplicità, etichettate come misure di posizione *condizionate*, per distinguerle dalle misure di posizione calcolate sulla variabile non condizionata, ovvero *marginale*.
Possiamo calcolare misure di posizione della durata della gravidanza condizionata al fumo oppure l'altezza condizionata al genere (ecc. ecc. dipende da quale variabile quantitativa e quale variabile gruppo abbiamo) e misure marginali

Esempio: altezze
Sia X l'altezza e Y il genere (con valori M e F).

- Mediana di $X|Y = M \longrightarrow$ 180.5 (mediana condizionata)
- Media di $X|Y = M \longrightarrow$ 180.1 (media condizionata)
- Mediana di $X|Y = F \longrightarrow$ 172 (mediana condizionata)
- Media di $X|Y = F \longrightarrow$ 171.2 (media condizionata)
- Mediana di $X \longrightarrow$ 177 (mediana marginale)
- Media di $X \longrightarrow$ 176.9 (media marginale)

Qualche formula

Fino ad ora, abbiamo introdotto alcune formule per il calcolo delle misure di posizione, immaginando di avere a disposizione i dati grezzi (ovvero la distribuzione statistica disaggregata).

Qualche formula

Fino ad ora, abbiamo introdotto alcune formule per il calcolo delle misure di posizione, immaginando di avere a disposizione i dati grezzi (ovvero la distribuzione statistica disaggregata).

A volte, anche partendo dai dati grezzi, possono esserci delle ambiguità nel calcolo delle misure (o indicatori). Più in generale, i dati possono essere forniti in forma aggregata (per es. sotto forma di tabelle di frequenza!).

Qualche formula

Fino ad ora, abbiamo introdotto alcune formule per il calcolo delle misure di posizione, immaginando di avere a disposizione i dati grezzi (ovvero la distribuzione statistica disaggregata).

A volte, anche partendo dai dati grezzi, possono esserci delle ambiguità nel calcolo delle misure (o indicatori). Più in generale, i dati possono essere forniti in forma aggregata (per es. sotto forma di tabelle di frequenza!).

Ora vedremo cosa fare in questi casi.

Mediana: distribuzione di frequenza per classi

Supponiamo di avere la seguente distribuzione di frequenza:

	$(0,1]$	$(1,2]$	$(2,3]$	$(3,4]$	$(4,5]$
frequenze assolute	1	4	4	2	1

I dati sono 12. La mediana dovrebbe essere scelta tra la 6^a e la 7^a osservazione dal basso

Mediana: distribuzione di frequenza per classi

Supponiamo di avere la seguente distribuzione di frequenza:

	$(0,1]$	$(1,2]$	$(2,3]$	$(3,4]$	$(4,5]$
frequenze assolute	1	4	4	2	1

I dati sono 12. La mediana dovrebbe essere scelta tra la 6^a e la 7^a osservazione dal basso

- Supponiamo (arbitrariamente) che i quattro dati appartenenti al terzo intervallo siano equidistribuiti. Sotto questa assunzione, la mediana è la media dei valori attribuiti alla 6° e alla 7° osservazione dal basso.

Mediana: distribuzione di frequenza per classi

Supponiamo di avere la seguente distribuzione di frequenza:

	$(0,1]$	$(1,2]$	$(2,3]$	$(3,4]$	$(4,5]$
frequenze assolute	1	4	4	2	1

I dati sono 12. La mediana dovrebbe essere scelta tra la 6^a e la 7^a osservazione dal basso

- Supponiamo (arbitrariamente) che i quattro dati appartenenti al terzo intervallo siano equidistribuiti. Sotto questa assunzione, la mediana è la media dei valori attribuiti alla $6°$ e alla $7°$ osservazione dal basso.
  - useremo le frequenze cumulate per identificare la classe mediana (cioè quella classe che contiene la mediana)

Mediana: distribuzione di frequenza per classi

Supponiamo di avere la seguente distribuzione di frequenza:

	$(0,1]$	$(1,2]$	$(2,3]$	$(3,4]$	$(4,5]$
frequenze assolute	1	4	4	2	1

I dati sono 12. La mediana dovrebbe essere scelta tra la 6^a e la 7^a osservazione dal basso

- Supponiamo (arbitrariamente) che i quattro dati appartenenti al terzo intervallo siano equidistribuiti. Sotto questa assunzione, la mediana è la media dei valori attribuiti alla $6°$ e alla $7°$ osservazione dal basso.
  - useremo le frequenze cumulate per identificare la classe mediana (cioè quella classe che contiene la mediana)

	$(0,1]$	$(1,2]$	$(2,3]$	$(3,4]$	$(4,5]$
frequenze assolute cumulate	1	5	9	11	12

  - $m \in (2,3]$ (si legge "la mediana m appartiene alla classe $(2,3]$", ossia che la classe mediana è $(2,3]$)
  - per identificare il valore esatto della mediana (e degli altri quantili) invece useremo una semplice proporzione (vedremo durante le esercitazioni)

Media: distribuzione di frequenza per classi

Supponiamo di avere a disposizione una distribuzione di frequenza per classi del tipo

intervalli	$(c_0, c_1]$	$(c_1, c_2]$	$\cdots$	$(c_{k-1}, c_k]$
frequenze assolute	n_1	n_2	$\cdots$	n_k

dove k indica il numero delle classi. La media non può essere calcolata esattamente.

Media: distribuzione di frequenza per classi

Supponiamo di avere a disposizione una distribuzione di frequenza per classi del tipo

intervalli	$(c_0, c_1]$	$(c_1, c_2]$	$\cdots$	$(c_{k-1}, c_k]$
frequenze assolute	n_1	n_2	$\cdots$	n_k

dove k indica il numero delle classi. La media non può essere calcolata esattamente.

Una approssimazione spesso usata in questi casi è

$$\frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i} = \frac{1}{N} \sum_{i=1}^{k} x_i n_i$$

dove x_i è il valore centrale della classe i-sima, ovvero la seguente semisomma

$$x_i = \frac{c_{i-1} + c_i}{2}$$

Esempio: dataset babies

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

$$\bar{x} = \frac{2500 * 5 + 2700 * 5 + 2900 * 5 + 3100 * 6 + 3300 * 5 + 3500 * 6}{32} = 3018,75.$$

La media calcolata a partire dai dati grezzi è invece $\bar{x} = 3019,875$. (ovviamente questa è più attendibile visto che non dobbiamo approssimare alcun valore come invece accade quando usiamo i valori centrali delle classi)

Importante: media aritmetica ponderata

La media aritmetica calcolata per dati raggruppati è un esempio di *media aritmetica ponderata*

$$\bar{x}_w = \frac{\sum_{i=1}^{k} x_i w_i}{\sum_{i=1}^{k} w_i}$$

dove ad ogni modalità x_i assegnamo un peso non negativo w_i. I pesi w_i possono essere di natura qualsiasi.

Media marginale e medie condizionate

Possiamo calcolare una media marginale a partire dalle medie condizionate.

Media marginale e medie condizionate

Possiamo calcolare una media marginale a partire dalle medie condizionate.

Supponiamo di avere N unità statistiche suddivise in L gruppi, secondo le modalità $y_1, \ldots y_L$ di una variabile qualitativa Y. Siano N_j, $j = 1, \ldots, L$, il numero di osservazioni per ogni gruppo. Ovviamente,

$$N = \sum_{j=1}^{L} N_j.$$

Media marginale e medie condizionate

Possiamo calcolare una media marginale a partire dalle medie condizionate.

Supponiamo di avere N unità statistiche suddivise in L gruppi, secondo le modalità $y_1, \ldots y_L$ di una variabile qualitativa Y. Siano N_j, $j = 1, \ldots, L$, il numero di osservazioni per ogni gruppo. Ovviamente,

$$N = \sum_{j=1}^{L} N_j.$$

Indichiamo poi con $x_{i,j}$ l'osservazione i-sima appartenente al gruppo j, $i = 1, \ldots, N_j$, $j = 1, \ldots, L$.

Esempio: dataset cholesterol

$X \longrightarrow$ livello di fosfato inorganico (mg/dl) nel plasma
$Y \longrightarrow$ tipo di paziente, con modalità y_1 =OI, y_3 =ON, y_3 =C

$X\|Y = y_1$	$X\|Y = y_2$	$X\|Y = y_3$
2.3	3.0	3.0
4.1	4.1	2.6
4.2	3.9	3.1
4.0	3.1	2.2
4.6	3.3	2.1
4.6	2.9	2.4
3.8	3.3	2.8
5.2	3.9	3.4
3.1		2.9
3.7		2.6
3.8		3.1
		3.2
$N_1 = 11$	$N_2 = 8$	$N_3 = 12$

Esempio: dataset cholesterol

$X \longrightarrow$ livello di fosfato inorganico (mg/dl) nel plasma
$Y \longrightarrow$ tipo di paziente, con modalità y_1 =OI, y_3 =ON, y_3 =C

$X\|Y = y_1$	$X\|Y = y_2$	$X\|Y = y_3$
2.3	3.0	3.0
4.1	4.1	2.6
4.2	3.9	3.1
4.0	3.1	2.2
4.6	3.3	2.1
4.6	2.9	2.4
3.8	3.3	2.8
5.2	3.9	3.4
3.1		2.9
3.7		2.6
3.8		3.1
		3.2
$N_1 = 11$	$N_2 = 8$	$N_3 = 12$

Abbiamo $L = 3$ e $N = 31$.

Esempio: dataset `cholesterol` (cont)

$X\|Y=y_1$	$X\|Y=y_2$	$X\|Y=y_3$
2.3	3.0	3.0
4.1	4.1	2.6
4.2	3.9	3.1
4.0	3.1	2.2
4.6	3.3	2.1
4.6	2.9	2.4
3.8	3.3	2.8
5.2	3.9	3.4
3.1		2.9
3.7		2.6
3.8		3.1
		3.2

$X\|Y=y_1$	$X\|Y=y_2$	$X\|Y=y_3$
$x_{1,1}$	$x_{1,2}$	$x_{1,3}$
$x_{2,1}$	$x_{2,2}$	$x_{2,3}$
$x_{3,1}$	$x_{3,2}$	$x_{3,3}$
$x_{4,1}$	$x_{4,2}$	$x_{4,3}$
$x_{5,1}$	$x_{5,2}$	$x_{5,3}$
$x_{6,1}$	$x_{6,2}$	$x_{6,3}$
$x_{7,1}$	$x_{7,2}$	$x_{7,3}$
$x_{8,1}$	$x_{8,2}$	$x_{8,3}$
$x_{9,1}$		$x_{9,3}$
$x_{10,1}$		$x_{10,3}$
$x_{11,1}$		$x_{11,3}$
		$x_{12,3}$

Media marginale e medie condizionate (cont)

Per ogni gruppo j possiamo calcolare la media condizionata

$$\overline{x}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} x_{i,j}.$$

Esempio: dataset `cholesterol`

$X\|Y = y_1$	$X\|Y = y_2$	$X\|Y = y_3$
3,94	3,44	2,78

$X\|Y = y_1$	$X\|Y = y_2$	$X\|Y = y_3$
$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_3$

NB. Si noti che:

$$N_j \overline{x}_j = \sum_{i=1}^{N_j} x_{i,j}.$$

questo risultato ci sarà utile per dimostrare il risultato presentato nella prossima slide...

Media marginale a partire dalle medie condizionate

La media marginale, ossia la media di tutte le osservazioni (senza riferimento al gruppo di appartenenza) è

$$\overline{x} = \frac{1}{N} \sum_{j=1}^{L} \sum_{i=1}^{N_j} x_{i,j}$$

È immediato dimostrare che la media marginale è la media delle medie condizionate, pesata con la numerosità dei gruppi. Infatti:

$$\overline{x} = \frac{1}{N} \sum_{j=1}^{L} \sum_{i=1}^{N_j} x_{i,j} = \frac{1}{N} \sum_{j=1}^{L} \left(\sum_{i=1}^{N_j} x_{i,j} \right) = \frac{1}{N} \sum_{j=1}^{L} N_j \overline{x}_j$$

Guardando oltre al centro della distribuzione

Ci interessa avere anche un'idea di quanto diversi siano i valori assunti dalla variabile, ossia ci interessa avere un'idea della variabilità di un carattere

Guardando oltre al centro della distribuzione

Ci interessa avere anche un'idea di quanto diversi siano i valori assunti dalla variabile, ossia ci interessa avere un'idea della variabilità di un carattere

Per farlo, possiamo vedere come si muovono le osservazioni intorno al centro della distribuzione.

Guardando oltre al centro della distribuzione

Ci interessa avere anche un'idea di quanto diversi siano i valori assunti dalla variabile, ossia ci interessa avere un'idea della variabilità di un carattere

Per farlo, possiamo vedere come si muovono le osservazioni intorno al centro della distribuzione.

E per fare ciò, possiamo usare l'idea di "distanza".

Esempio: assenza di variabilità

Se non c'è variabilità, tutte le unità statistiche mostrano la stessa modalità del carattere.

$$1{,}1 \;\; 1{,}1 \;\; 1{,}1 \;\; 1{,}1 \;\; 1{,}1 \;\; 1{,}1 \;\; 1{,}1 \;\; 1{,}1 \;\; 1{,}1 \;\; 1{,}1 \;\; 1{,}1 \;\; 1{,}1$$

Abbiamo

- $x_{(1)} = 1{,}1 \;\; x_{(N)} = 1{,}1$
- $q_{0.25} = 1{,}1 \;\; m = 1{,}1 \;\; q_{0.75} = 1{,}1$

Misurando distanze dal centro della distribuzione, possiamo costruire indicatori che valgono 0 in assenza di variabilità.

- $|x_i - m| = |x_i - x_j| = 0, \;\; i, j = 1, \ldots, N,$
- $x_{(N)} - m = m - x_{(1)} = 0,$
- $q_{0.75} - m = m - q_{0.25} = 0.$

Indici elementari di variabilità

- $x_{(N)} - x_{(1)}$ è il *campo di variazione* (range).
- $q_{0.75} - q_{0.25}$ è la *distanza interquartilica* (IQR).

Indici elementari di variabilità

- $x_{(N)} - x_{(1)}$ è il *campo di variazione* (range).
- $q_{0.75} - q_{0.25}$ è la *distanza interquartilica* (IQR).

Ovviamente, in presenza di variabilità, sia il campo di variazione che la distanza interquartilica assumono un valore maggiore di zero.

E, in presenza di variabilità, possiamo cercare di rappresentare come variano le modalità.

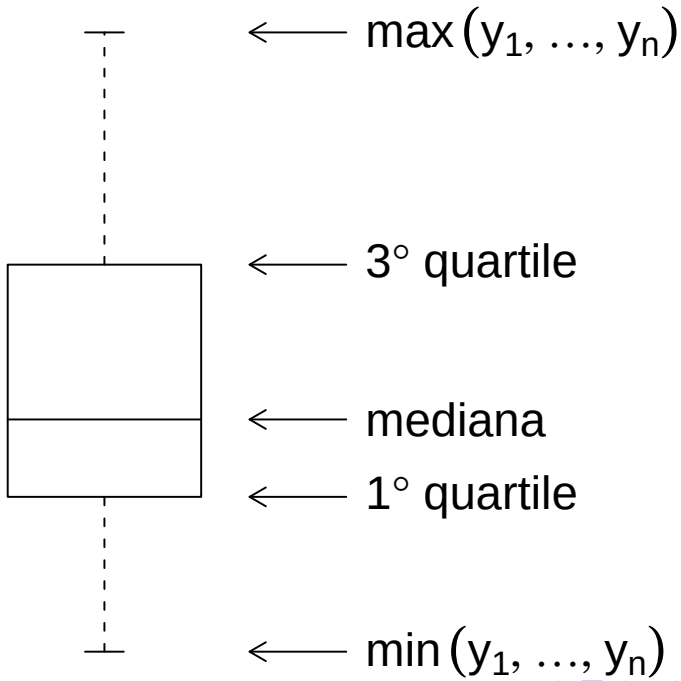
Diagramma a scatola con baffi (*box and whiskers plot o boxplot*)

Il boxplot è un grafico molto utilizzato in statistica. Esso fornisce un'idea schematica di un insieme di dati (di una distribuzione) basata sui quartili.

Diagramma a scatola con baffi (*box and whiskers plot o boxplot*)

Il boxplot è un grafico molto utilizzato in statistica. Esso fornisce un'idea schematica di un insieme di dati (di una distribuzione) basata sui quartili.

Sono costituiti, come dice il nome, da una *scatola* e da due *baffi* costruiti in accordo al disegno sottostante.

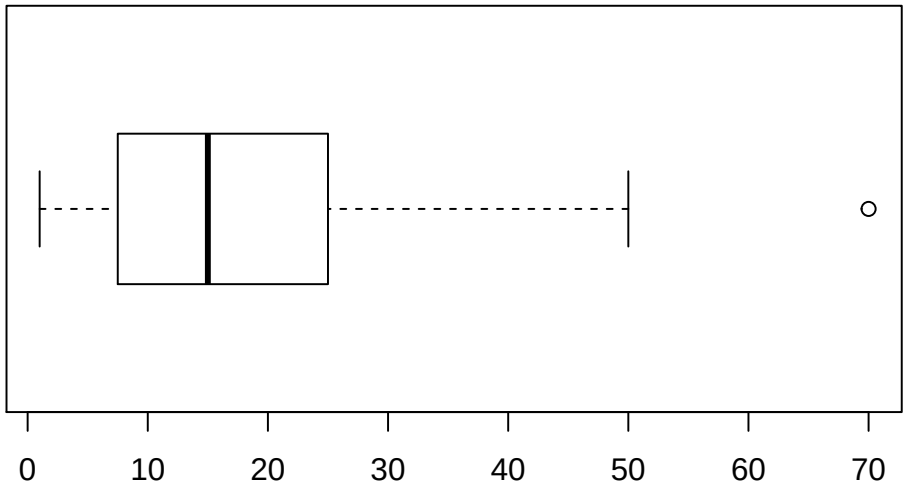


Boxplot (cont)

Una variante del diagramma usata frequentemente può essere costruita come segue:

1. la scatola è costruita come descritto precedentemente a partire dai tre quartili.
2. i baffi si estendono fino ai dati più lontani che siano però non più distanti di cost $\times$ (scarto interquartile) dalla scatola (non accettiamo baffi esageratamente lunghi).
3. cost è una costante arbitraria, tipicamente scelta uguale a 1,5.
4. Le osservazioni che sono oltre i baffi sono disegnate opportunamente sul grafico (ad. esempio utilizzando un pallino o un asterisco). $\Rightarrow$ Queste osservazioni sono dette valori anomali (o outliers), cioè valori particolarmente distanti dal centro della distribuzione (tali da poter essere addirittura considerati errori di rilevazione del dato)

Diagrammi a scatola con baffi: ore di studio settimanali



Esercizio: costruzione di un boxplot

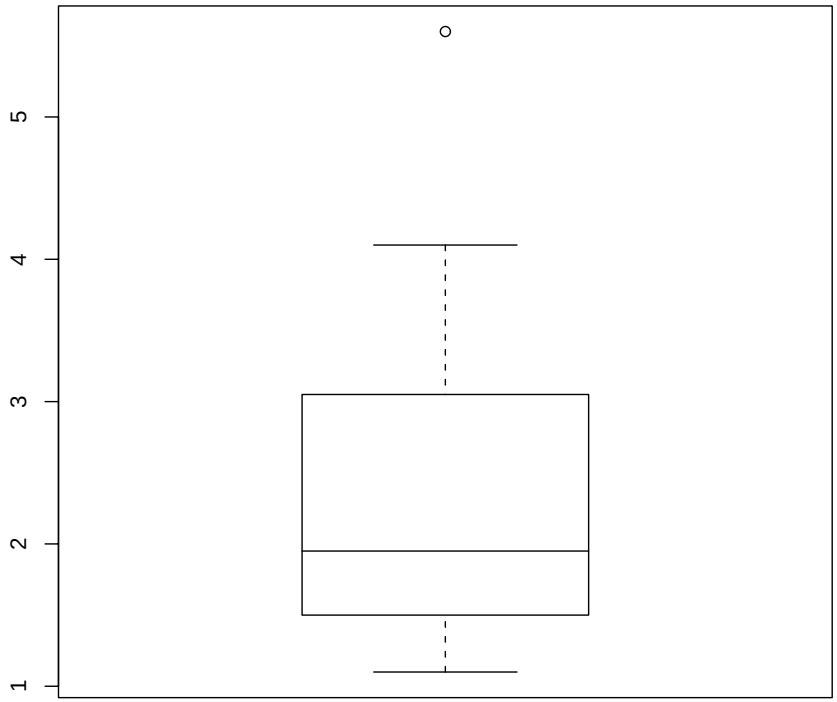
Dati (già ordinati):

$$1{,}1 \quad 1{,}3 \quad 1{,}4 \quad 1{,}6 \quad 1{,}8 \quad 1{,}9 \quad 2{,}0 \quad 2{,}5 \quad 2{,}9 \quad 3{,}2 \quad 4{,}1 \quad 5{,}6$$

Perciò $q_{0.25} = 1{,}5$, $m = 1{,}95$, $q_{0.75} = 3{,}05$,
$1{,}5 \times (q_{0.75} - q_{0.25}) = 1{,}5 \times 1{,}55 = 2{,}325$.

1. *scatola*: da 1,5 a 3,05 con la mediana indicata da una linea a 1,95;
2. *baffo inferiore*: fino all'osservazione più bassa tra quelle maggiori di $q_{0.25} - 2{,}325 = -0{,}825$, ovvero fino a 1,1;
3. *baffo superiore*: fino all'osservazione più alta tra quelle minori di $q_{0.75} + 2{,}325 = 5{,}375$, ovvero fino a 4,1;
4. sono da disegnare esplicitamente nel diagramma le osservazioni più piccole di 1,1 o più grandi di 5,375; in questo caso solamente l'osservazione risultata uguale a 5,6.

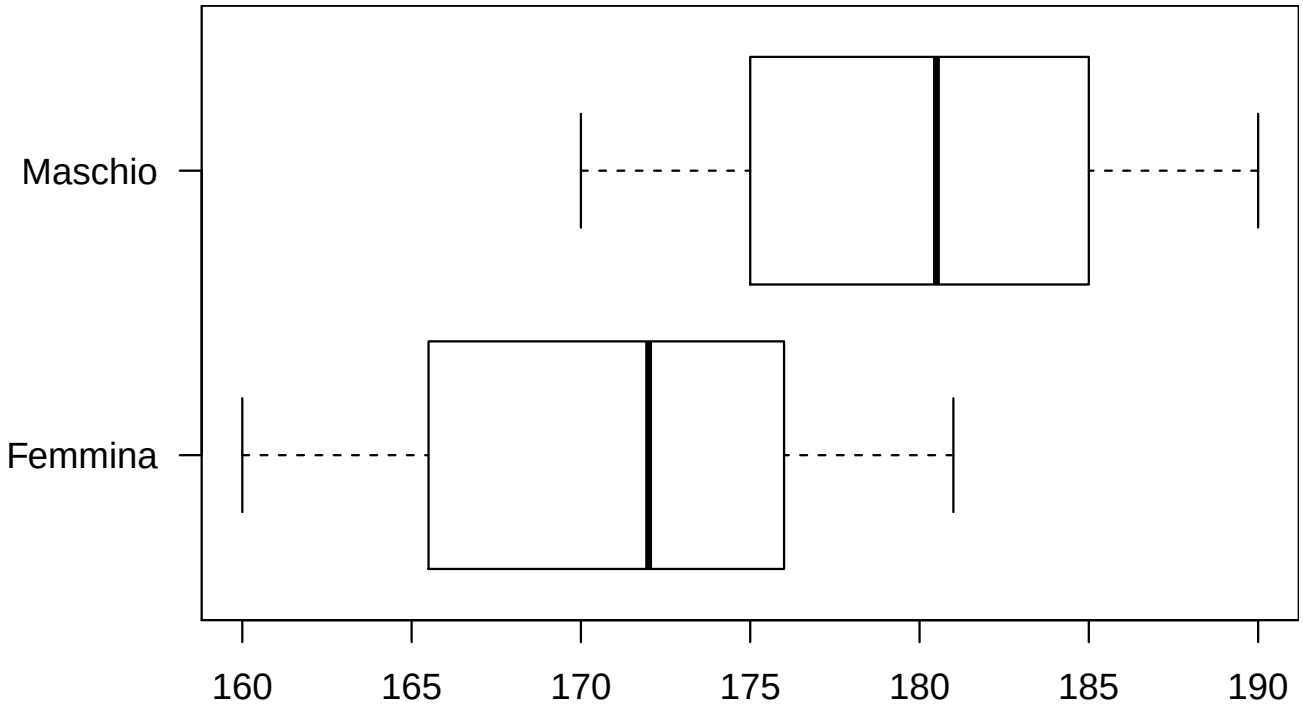
Diagramma a scatola con baffi (esempio precedente)



Box plot: anche per variabili condizionate

Esempio: altezze.

Altezze di maschi e femmine



Come descrivereste queste distribuzioni condizionate?

Outliers

Perché è importante cercare gli outliers?

- Dare una spiegazione a marcate asimmetrie.
- Identificare errori nell'imputazione dei dati.
- Scoprire cose nuove.

Ancora sulla variabilità

Abbiamo detto che per misurare la variabilità, possiamo utilizzare la "distanza" delle osservazioni dal centro della distribuzione.

Ancora sulla variabilità

Abbiamo detto che per misurare la variabilità, possiamo utilizzare la "distanza" delle osservazioni dal centro della distribuzione.

Proviamo a utilizzare la media per caratterizzare il centro della distribuzione.

Ancora sulla variabilità

Abbiamo detto che per misurare la variabilità, possiamo utilizzare la "distanza" delle osservazioni dal centro della distribuzione.

Proviamo a utilizzare la media per caratterizzare il centro della distribuzione.

Siano $x = (x_1, \ldots, x_N)$ i dati osservati, N il loro numero e $\overline{x}$ la loro media aritmetica, ovvero $\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$.

La distanza di ogni osservazione x_i dalla media $\overline{x}$, il cosidetto *scarto dalla media*, può essere misurata così:

$$|x_i - \overline{x}|.$$

Ancora sulla variabilità

Abbiamo detto che per misurare la variabilità, possiamo utilizzare la "distanza" delle osservazioni dal centro della distribuzione.

Proviamo a utilizzare la media per caratterizzare il centro della distribuzione.

Siano $x = (x_1, \ldots, x_N)$ i dati osservati, N il loro numero e $\overline{x}$ la loro media aritmetica, ovvero $\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$.

La distanza di ogni osservazione x_i dalla media $\overline{x}$, il cosidetto *scarto dalla media*, può essere misurata così:

$$|x_i - \overline{x}|.$$

Perché abbiamo bisogno del valore assoluto?

Ancora sulla variabilità (cont)

È ancora meglio se consideriamo lo scarto al quadrato:

$$(x_i - \overline{x})^2.$$

Perché il quadrato?

Ancora sulla variabilità (cont)

È ancora meglio se consideriamo lo scarto al quadrato:

$$(x_i - \overline{x})^2.$$

Perché il quadrato?

Perché il quadrato "amplifica" le distanze grandi e "attenua" quelle piccole.

Esempio: $10^2 = 100$, $0.1^2 = 0.01$.

Ancora sulla variabilità (cont)

È ancora meglio se consideriamo lo scarto al quadrato:

$$(x_i - \overline{x})^2.$$

Perché il quadrato?

Perché il quadrato "amplifica" le distanze grandi e "attenua" quelle piccole.

Esempio: $10^2 = 100$, $0.1^2 = 0.01$.

Quindi, per costruire un indice di variabilità, possiamo costruire queste N quantità (per $i = 1, \ldots, N$) e farne una media.

Varianza

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

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

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:

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

Deviazione standard

La *deviazione standard* è la radice quadrata della varianza ed è espressa nella stessa unità di misura del carattere.

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

La deviazione standard per le ore di studio/settimana degli studenti si calcola come:

$$\sigma = \sqrt{183.12} = 13.53$$

Devianza

La deviazione standard non deve essere confusa con la *devianza*, che è la quantità al numeratore della varianza.

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

La devianza rappresenta quindi la somma dei quadrati degli scarti delle osservazioni dalla propria media.

Varianza campionaria corretta

Quando si lavora con un campione (quindi nella stragrande maggioranza dei casi...), si utilizza spesso la *varianza campionaria corretta*, che differisce dalla varianza campionaria solo per il denominatore (che anzichè N è uguale a $N - 1$):

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

La ragione della modifica del denominatore è legata a proprietà teoriche di s^2 che la rendono una misura di variabilità più comoda quando farete inferenza.

Varianza: una formula operativa

Si osservi che

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

$$= \frac{1}{N}\sum_{i=1}^{N} x_i^2 + \frac{1}{N}\sum_{i=1}^{N}\overline{x}^2 - \frac{1}{N}\sum_{i=1}^{N} 2\overline{x}x_i =$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i^2 + \frac{N\overline{x}^2}{N} - \frac{2\overline{x}}{N}\sum_{i=1}^{N} x_i =$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i^2 + \overline{x}^2 - 2\overline{x}^2$$

Varianza: una formula operativa (cont)

Quindi, possiamo scrivere

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

ovvero

$$(\text{varianza}) = \left(\begin{array}{c} \text{media dei} \\ \text{quadrati} \end{array} \right) - \left(\begin{array}{c} \text{quadrato della} \\ \text{media} \end{array} \right).$$

Formula operativa: esempio di utilizzo

dati: $1, 3, 2, 5$.

media: $\frac{1+3+2+5}{4} = 2.75$.

media dei quadrati: $\frac{1^2+3^2+2^2+5^2}{4} = 9.75$.

varianza: $9.75 - 2.75^2 = 2.19$.

Importante: media potenziata di ordine s

La media dei quadrati è collegata alla media potenziata di ordine 2.
Si dice *media potenziata di ordine s* (con $s \neq 0$) il valore

$$\mu_s = \left(\frac{1}{N} \sum_{i=1}^{N} x_i^s \right)^{\frac{1}{s}}.$$

Osservazioni:

- $s = -1$: media armonica.
- $s = 1$: media aritmetica.
- $s = 2$: media quadratica.
- $s \to 0$: media geometrica (se la variabile ha valori positivi).

Varianza: distribuzione di frequenza per classi

Supponiamo di avere a disposizione una distribuzione di frequenza per classi del tipo

intervalli	$[c_0, c_1)$	$[c_1, c_2)$	$\cdots$	$[c_{k-1}, c_k)$
frequenze assolute	n_1	n_2	$\cdots$	n_k

dove k indica il numero delle classi.

Varianza: distribuzione di frequenza per classi

Supponiamo di avere a disposizione una distribuzione di frequenza per classi del tipo

intervalli	$[c_0, c_1)$	$[c_1, c_2)$	$\cdots$	$[c_{k-1}, c_k)$
frequenze assolute	n_1	n_2	$\cdots$	n_k

dove k indica il numero delle classi.

Per il calcolo della varianza, possiamo fare ricorso alla formula operativa utilizzando la stessa strategia adottata per il calcolo della media da distribuzioni di frequenza, ovvero utilizzando il punto centrale di ogni classe per rappresentare i valori della classe stessa.

Esempio: altezze

Immaginiamo di disporre solo della distribuzione in classi della variabile altezza e che questa sia fatta nel seguente modo:

altezza	frequenza assoluta
(160,170]	10
(170,175]	10
(175,180]	13
(180,190]	18

- $\bar{x} = (10 \times 165 + 10 \times 172.5 + 13 \times 177.5 + 18 \times 185)/51 = 176.72$
- $\frac{1}{N}\sum_{i=1}^{N} x_i^2 = (10 \times 165^2 + 10 \times 172.5^2 + 13 \times 177.5^2 + 18 \times 185^2)/51 = 31283.21$
- $\sigma^2 = 31283.21 - 176.72^2 = 53.25$
- $\sigma = 7.30$

Varianza marginale e varianze condizionate

Riprendiamo le nostre N unità statistiche suddivise in L gruppi, secondo le L modalità di una variabile X (v. dataset `cholesterol`).

La varianza marginale, ossia la varianza di tutte le osservazioni (senza riferimento al gruppo di appartenenza) è

$$\sigma^2 = \frac{1}{N} \sum_{j=1}^{L} \sum_{i=1}^{N_j} (x_{i,j} - \overline{x})^2$$

Per ogni gruppo definito dalle L modalità di una variabile qualitativa Y, possiamo calcolare la varianza condizionata

$$\sigma_j^2 = \frac{1}{N_j} \sum_{i=1}^{N_j} (x_{i,j} - \overline{x}_j)^2.$$

Varianza marginale e varianze condizionate (cont)

Si dimostra che

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{L} N_j \sigma_j^2 + \frac{1}{N}\sum_{j=1}^{L} N_j (\overline{x}_j - \overline{x})^2$$

Il primo addendo sul lato destro della formula è la media delle varianze condizionate σ_j^2 pesate con N_j, detta *varianza entro i gruppi*.

Il secondo addendo è la varianza delle medie condizionate, anche queste pesate con N_j, detta *varianza tra i gruppi*.

Varianza marginale e varianze condizionate (cont)

Si dimostra che

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{L} N_j \sigma_j^2 + \frac{1}{N}\sum_{j=1}^{L} N_j (\overline{x}_j - \overline{x})^2$$

Il primo addendo sul lato destro della formula è la media delle varianze condizionate σ_j^2 pesate con N_j, detta *varianza entro i gruppi*.

Il secondo addendo è la varianza delle medie condizionate, anche queste pesate con N_j, detta *varianza tra i gruppi*.

Questa è chiamata *scomposizione della varianza*.

Varianza marginale e varianze condizionate (cont)

- La scomposizione mostra come la varianza totale, σ^2, sia scomponibile in due parti:
  (i) la prima, il 1° addendo, dovuta alla variabilità *entro* i gruppi e
  (ii) la seconda, il 2° addendo, legata alle differenze *tra* le medie dei gruppi.

  Per questo motivo, i due addendi sono spesso indicati come *varianza entro i gruppi* e *varianza tra i gruppi*.

Scomposizione della varianza: dimostrazione

$$
\begin{aligned}
\sigma^2 &= \frac{1}{N}\sum_{j=1}^{L}\sum_{i=1}^{N_j}(x_{i,j}-\overline{x})^2 = \\
&= \frac{1}{N}\sum_{j=1}^{L}\sum_{i=1}^{N_j}[(x_{i,j}-\overline{x}_j)+(\overline{x}_j-\overline{x})]^2 = \\
&= \frac{1}{N}\sum_{j=1}^{L}\sum_{i=1}^{N_j}[(x_{i,j}-\overline{x}_j)^2+(\overline{x}_j-\overline{x})^2+2(x_{i,j}-\overline{x}_j)(\overline{x}_j-\overline{x})] = \\
&= \frac{1}{N}\sum_{j=1}^{L}\sum_{i=1}^{N_j}(x_{i,j}-\overline{x}_j)^2+\frac{1}{N}\sum_{j=1}^{L}N_j(\overline{x}_j-\overline{x})^2+ \\
&\quad +\frac{2}{N}\sum_{j=1}^{L}(\overline{x}_j-\overline{x})\sum_{i=1}^{N_j}(x_{i,j}-\overline{x}_j) = \\
&= \frac{1}{N}\sum_{j=1}^{L}N_j\sigma_j^2+\frac{1}{N}\sum_{j=1}^{L}N_j(\overline{x}_j-\overline{x})^2.
\end{aligned}
$$

Indice

Indice di eterogeneità

- La varianza e la deviazione standard sono due indici di variabilità per variabili quantitative (perché c'è bisogno di calcolare un valor medio).
- Un'interessante proprietà che si può studiare nel caso di variabili (mutabili) qualitative è la mutabilità (o eterogeneità), cioè l'attitudine del carattere a manifestarsi con modalità diverse tra le unità statistiche.
- Sia X una variabile qualitativa con k modalità ciascuna delle quali con frequenza relativa f_i, per $i = 1, \ldots, k$.
- L'indice di eterogeneità di Gini (assoluto) G si calcola con la seguente: $G = 1 - \sum_{i=1}^{k} f_i^2$

Nota: Maggiore è tale indice più i dati saranno distribuiti in maniera eterogenea tra le k modalità ossia le k modalità hanno frequenze simili. Minore il valore di G minore l'eterogeneità (maggiore omogeneità)

Indice di eterogeneità normalizzato

- G varia nell'intervallo
  - $0 \leq G \leq \frac{(k-1)}{k}$
- Pertanto:
  - Se $G = 0$, ovvero nel caso di minima eterogeneità, i dati assumono tutti un'unica modalità che ha quindi frequenza relativa massima (pari a 1);
  - Se $G = \frac{(k-1)}{k}$, cioè nel caso di massima eterogeneità, i dati sono distribuiti equamente su tutte le k modalità, le quali hanno quindi uguale frequenza relativa.
- Quindi una misura normalizzata tra 0 e 1 (0= min. eterogeneità e 1=max eterogeneità) è:
  - $G_N = G / \frac{k-1}{k}$

Indice

La disuguaglianza di Chebyshev

La *disuguaglianza di Chebyshev* ci aiuta a "descrivere" una distribuzione in termini della sua media e della sua deviazione standard.

La disuguaglianza dice che, dato un numero $h \geq 1$ e N osservazioni per un certo carattere, almeno $[1 - 1/h^2]$ delle osservazioni cadrà nell'intervallo avente come estremi la media $\pm h$ volte la deviazione standard.

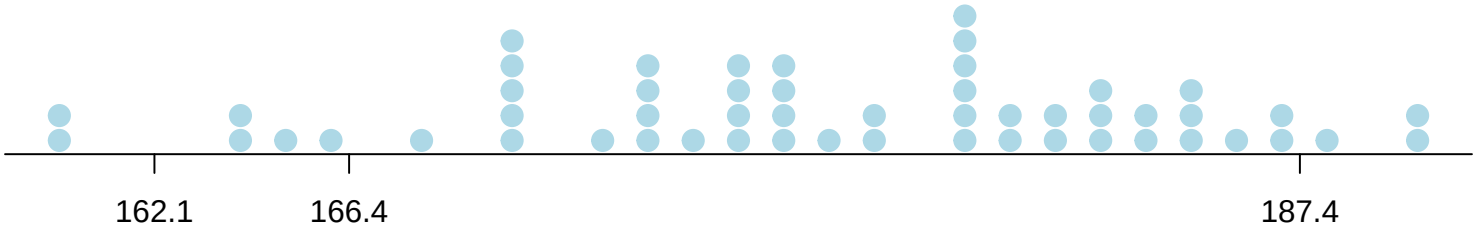
Esempio: $N = 25$, $\bar{y} = 75$, $\sigma = 10$.
Scegliendo $h = 2/\sqrt{3}, \sqrt{2}, 2, 3$, si ha

- almeno il 25% dei dati cade tra 63.5 e 86.5;
- almeno il 50% dei dati cade tra 60.9 e 89.1;
- almeno il 75% dei dati cade tra 55 e 95;
- almeno il 88.9% dei dati cade tra 45 e 105.

La disuguaglianza di Chebyshev: altezze

La disuguaglianza è "conservatrice"

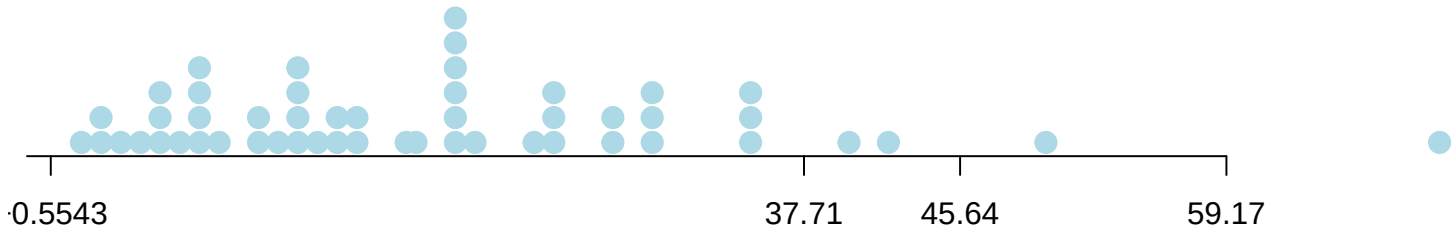
- media=176.9, s.d.=7.398
- per C. il 50% dei dati cade tra 166.4 e 187.4
  - in effetti è il 83%
- per C. il 75% dei dati cade tra 162.1 e 191.7
  - in effetti è il 96.2%
- per C. il 90% dei dati cade tra 154.7 e 199.1
  - in effetti è il 100%



La disuguaglianza di Chebyshev: ore di studio

La disuguaglianza è "conservatrice"

- media=18.58, s.d.=13.53
- per C. il 50% dei dati cade tra -0.5543 e 37.71
  - in effetti è il 92.2%
- per C. il 75% dei dati cade tra -8.48 e 45.64
  - in effetti è il 96.1%
- per C. il 90% dei dati cade tra -22.01 e 59.17
  - in effetti è il 98%



Indice

Forma di una distribuzione

Oltre alla media e alla varianza (e deviazione standard), ci sono altri aspetti da valutare per "descrivere" una distribuzione.

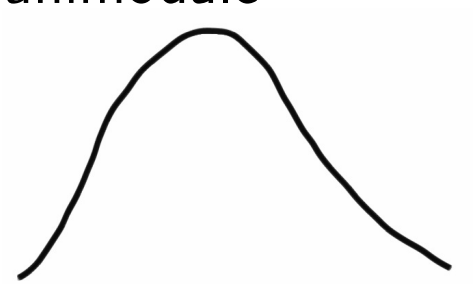
Quanti picchi mostra l'istogramma: uno (distribuzione *unimodale*), molti (distribuzione *bimodale/multimodale*), o nessuno (distribuzione *uniforme*)?

Forma di una distribuzione

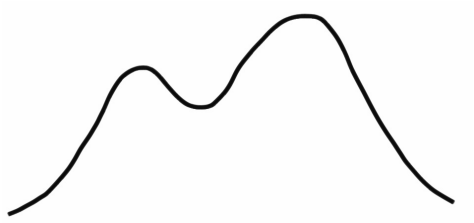
Oltre alla media e alla varianza (e deviazione standard), ci sono altri aspetti da valutare per "descrivere" una distribuzione.

Quanti picchi mostra l'istogramma: uno (distribuzione *unimodale*), molti (distribuzione *bimodale/multimodale*), o nessuno (distribuzione *uniforme*)?

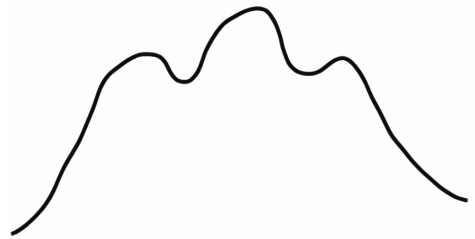
unimodale



bimodale



multimodale



uniforme

Forma della distribuzione: simmetria

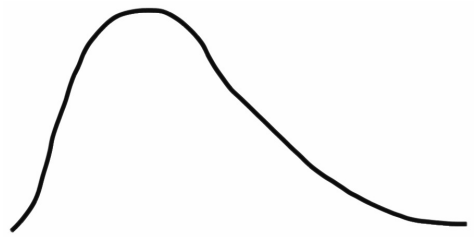
L'istogramma è *asimmetrico a destra*, *asimmetrico a sinistra*, o *simmetrico*?

La direzione (destra/sinistra) della asimmetria è data dalla posizione della coda più lunga.

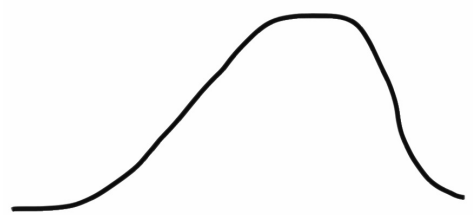
Forma della distribuzione: simmetria

L'istogramma è *asimmetrico a destra*, *asimmetrico a sinistra*, o *simmetrico*?

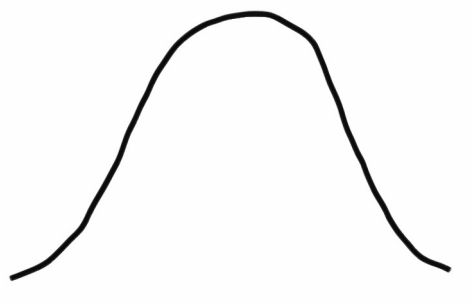
asimmetria a destra



asimmetria a sinistra



simmetria

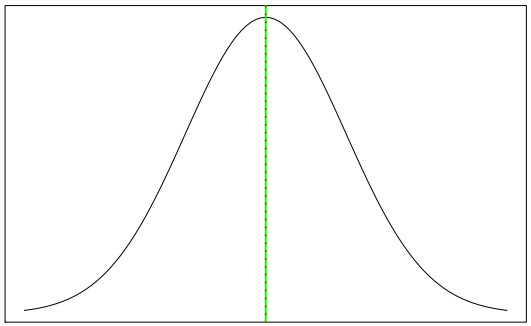


La direzione (destra/sinistra) della asimmetria è data dalla posizione della coda più lunga.

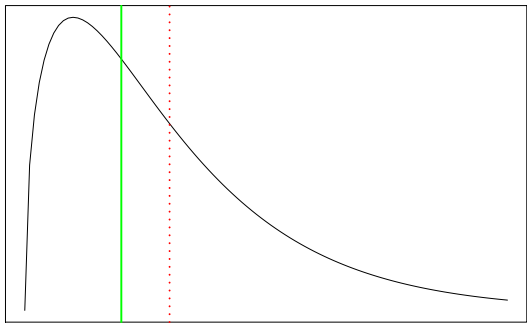
Simmetria: media vs. mediana

Se la distribuzione è simmetrica

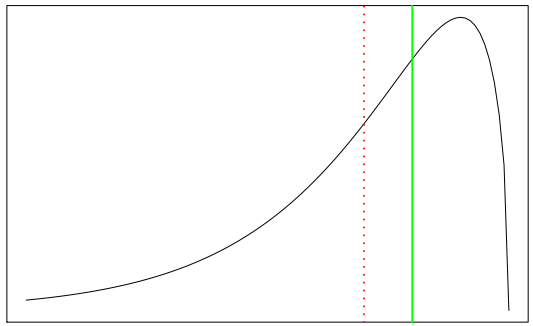
media $\approx$ mediana



Se la distribuzione è asimmetrica a destra (positiva):
media $>$ mediana



Se la distribuzione è asimmetrica a sinistra (negativa):
media $<$ mediana



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 σ

Trasformazioni lineari: esempio notevole

Standardizzazione

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

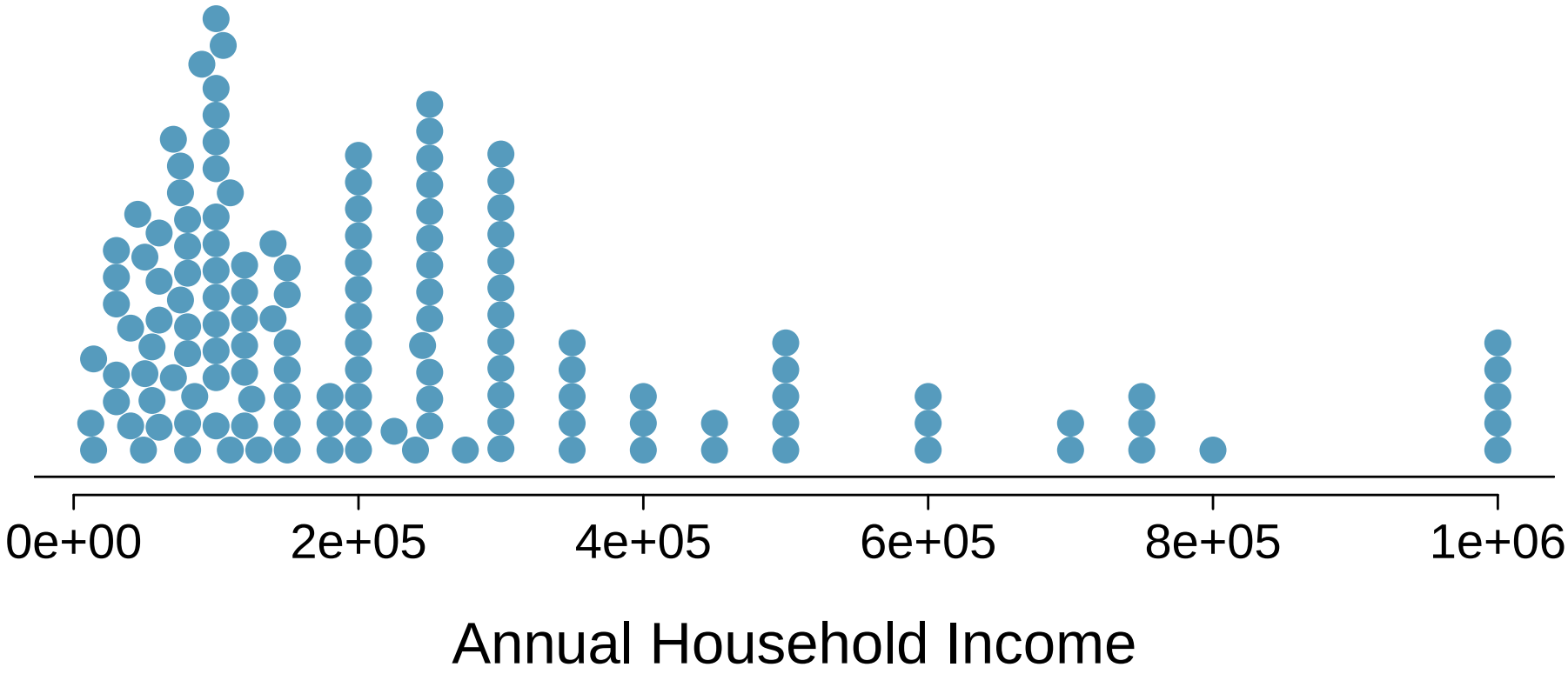
$$z_i = a + bx_i = -\frac{\bar{x}}{\sigma} + \frac{1}{\sigma}x_i.$$

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

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

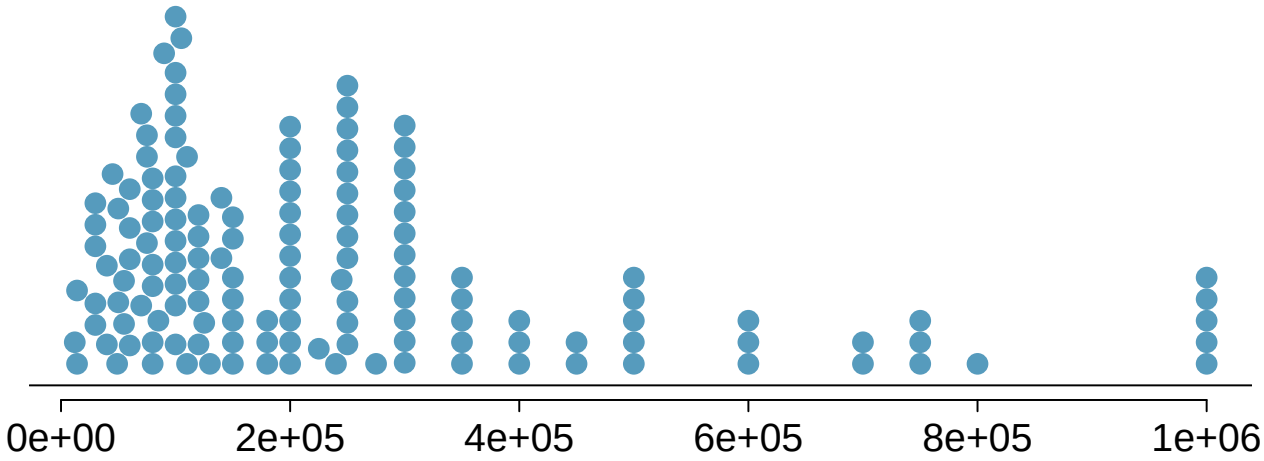
Osservazioni estreme e robustezza

Esempio: reddito annuale di un campione di famiglie.



Come cambierebbero la mediana, la media, la distanza interquartilica e la deviazione standard se l'osservazione più elevata fosse \$10 milioni?

Osservazioni estreme e robustezza



	robusto		**non robusto**	
scenario	m	IQR	$\bar{x}$	σ
dati originali	190K	200K	245K	226K
sposta max a \$10 milioni	190K	200K	309K	853K

Mediana e IQR sono più "stabili" della media e della deviazione standard. Si dice che sono più "robuste". **Ossia in presenza di valori anomali è sempre meglio usare la mediana anzichè la media (e l'IQR anzichè il campo di variazione)**

Proprietà di Cauchy

La media è sempre compresa tra il più piccolo e il più grande dei valori osservati:

$$x_{(1)} \leq \overline{x} \leq x_{(N)}.$$

Infatti, ad esempio, per quanto riguarda la prima disuguglianza

$$x_{(1)} = \frac{\overbrace{x_{(1)} + \cdots + x_{(1)}}^{N \text{ volte}}}{N} \leq \frac{x_1 + x_2 + \cdots + x_N}{N} = \overline{x}$$

Proprietà di baricentro

La somma (e quindi la media) degli scarti dei dati grezzi dalla propria media è sempre zero:

$$\sum_{i=1}^{N}(x_i - \bar{x}) = 0.$$

Infatti, con i dati grezzi, si ha

$$\sum_{i=1}^{N}(x_i - \bar{x}) = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \bar{x} = N\bar{x} - N\bar{x} = 0.$$

Equivarianza rispetto a trasformazioni lineari

Siano $x_1, x_2, \ldots, x_N$ le osservazioni disponibili per il carattere X e sia $\overline{x}$ la loro media.

Equivarianza rispetto a trasformazioni lineari

Siano $x_1, x_2, \ldots, x_N$ le osservazioni disponibili per il carattere X e sia $\overline{x}$ la loro media.

Sia $T = g(X) = a + bX$ una trasformazione lineare e siano $t_1 = g(x_1), t_2 = g(x_2), \ldots, t_N = g(x_N)$ i dati risultanti dalla trasformazione dei dati grezzi $x_1, x_2, \ldots, x_N$

Equivarianza rispetto a trasformazioni lineari

Siano $x_1, x_2, \ldots, x_N$ le osservazioni disponibili per il carattere X e sia $\overline{x}$ la loro media.

Sia $T = g(X) = a + bX$ una trasformazione lineare e siano $t_1 = g(x_1), t_2 = g(x_2), \ldots, t_N = g(x_N)$ i dati risultanti dalla trasformazione dei dati grezzi $x_1, x_2, \ldots, x_N$

La media dei valori $t_1, t_2, \ldots, t_N$, indicata con $\overline{t}$, è la trasformazione tramite $g(\cdot)$ della media dei valori originali, ovvero

$$\overline{t} = g(\overline{x}).$$

Equivarianza rispetto a trasformazioni lineari (cont)

Dimostrazione.

$$
\begin{aligned}
\bar{t} &= \frac{t_1 + t_2 + \cdots + t_N}{N} = \\
&= \frac{(a + bx_1) + (a + bx_2) + \cdots + (a + bx_N)}{N} = \\
&= \frac{\overbrace{a + \cdots + a}^{N \text{ volte}}}{N} + b\frac{x_1 + x_2 + \cdots + x_N}{N} \\
&= a + b\bar{x}.
\end{aligned}
$$