

Statistica

Indici sintetici

Francesco Pauli

A.A. 2019/2020

La parola chiave è riassumere

Tutti gli strumenti che abbiamo visto sinora: tabelle di frequenza e rappresentazioni grafiche; servono a riassumere le osservazioni.



Gli strumenti che vedremo ora sono anch'essi metodi per riassumere le osservazioni.



Come tutti i riassunti, anche questi evidenzieranno certi aspetti trascurandone altri.



In particolare sono tre gli aspetti che vogliamo considerare:

- ▶ posizione, centro della distribuzione;
- ▶ variabilità: “larghezza” della distribuzione;
- ▶ forma: simmetria ma non solo.

Indice

Misure di posizione

Variabilità

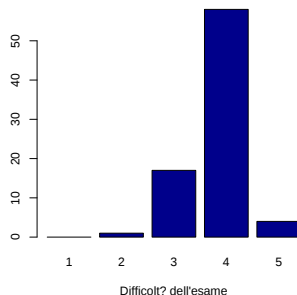
Media e varianza per dati in gruppi

Forma di una distribuzione

Proprietà delle misure di posizione e scala

Esempio: Difficoltà dell'esame di statistica.

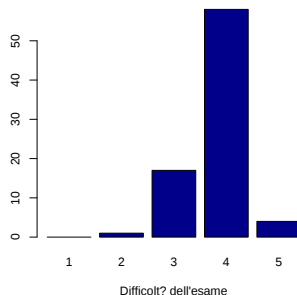
Modalità	Freq
1	0
2	1
3	17
4	58
5	4



Quale potrebbe essere un riassunto efficace per il centro di questa distribuzione?

Esempio: Difficoltà dell'esame di statistica.

Modalità	Freq
1	0
2	1
3	17
4	58
5	4



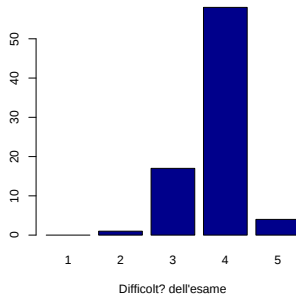
Quale potrebbe essere un riassunto efficace per il centro di questa distribuzione?



Potremmo dire che è 4, il valore che compare più frequentemente.

Esempio: Difficoltà dell'esame di statistica.

Modalità	Freq
1	0
2	1
3	17
4	58
5	4



Quale potrebbe essere un riassunto efficace per il centro di questa distribuzione?



Potremmo dire che è 4, il valore che compare più frequentemente.



Questo valore è chiamato **moda** della distribuzione.

Misure di posizione: la moda

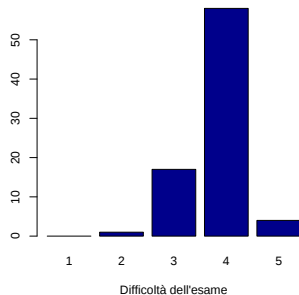
Moda

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

- ▶ In altre parole, la moda esprime la modalità più comune.
- ▶ È definita sia per variabili qualitative che quantitative.
- ▶ N.B.: la moda è la modalità NON la frequenza.

Moda come sintesi

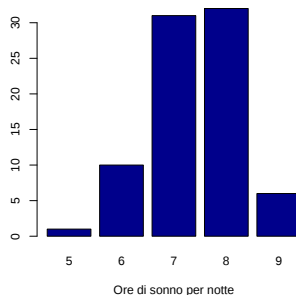
Modalità	Freq
1	0
2	1
3	17
4	58
5	4



La moda è una sintesi efficace (interessante) se c'è una modalità 'molto' più frequente delle altre.

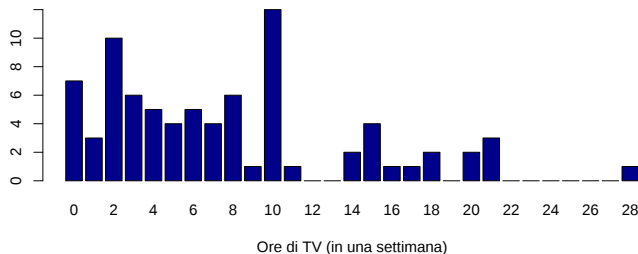
Moda come sintesi

Modalità	Freq
5	0.01
6	0.12
7	0.39
8	0.40
9	0.07



Per le ore di sonno per notte funzionerebbe meno bene.

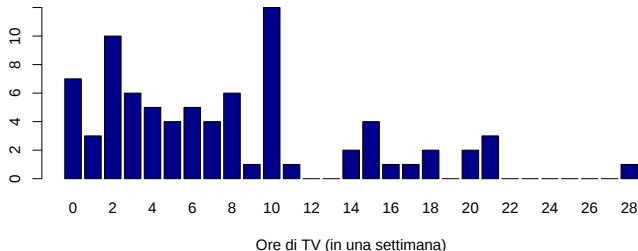
Moda come sintesi



Ma anche per il numero di ore di televisione la moda non appare molto indicativa.

Mediana come misura di posizione

Il centro di una distribuzione potrebbe anche essere pensato come quel valore che “divide a metà” le osservazioni.

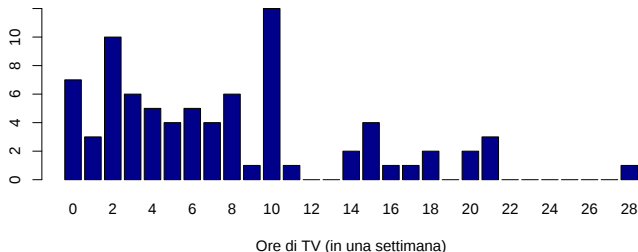


Mettiamo in fila le 80 persone, in ordine di numero di ore guardate.



Mediana come misura di posizione

Il centro di una distribuzione potrebbe anche essere pensato come quel valore che “divide a metà” le osservazioni.

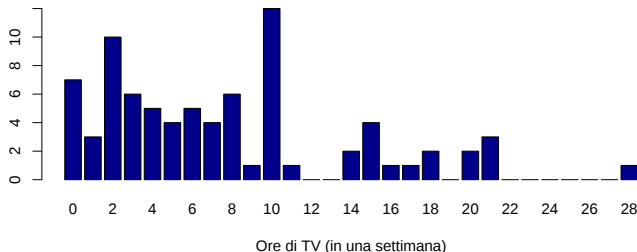


Mettiamo in fila le 80 persone, in ordine di numero di ore guardate.
Dividiamoli a metà.



Mediana come misura di posizione

Il centro di una distribuzione potrebbe anche essere pensato come quel valore che “divide a metà” le osservazioni.



Mettiamo in fila le 80 persone, in ordine di numero di ore guardate.

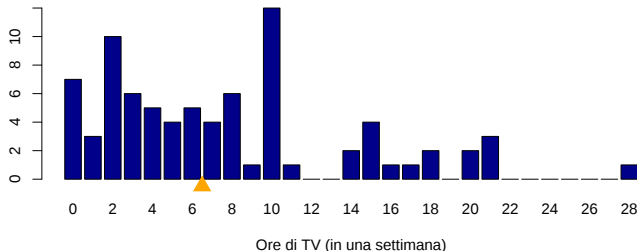
Dividiamoli a metà.

40 guardano meno di 6.5, 40 guardano più di 6.5.



Mediana come misura di posizione

Il centro di una distribuzione potrebbe anche essere pensato come quel valore che “divide a metà” le osservazioni.



Il valore 6.5, che divide a metà le osservazioni, è la **mediana** della distribuzione.



Definizione di mediana

Sia $y_1, y_2, \dots, y_N$ una distribuzione statistica disaggregata.
Indichiamo con $y_{(1)}, y_{(2)}, \dots, y_{(N)}$ le osservazioni ordinate:

$$y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(N)}$$

Se N è pari (quindi $N/2$ è intero)

$$\underbrace{y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(\frac{N}{2})}}_{\frac{N}{2} \text{ osservazioni}} \leq \underbrace{y_{(\frac{N}{2}+1)} \leq \dots \leq y_{(N)}}_{\frac{N}{2} \text{ osservazioni}}$$

e la mediana è

$$m = \frac{y_{(\frac{N}{2})} + y_{(\frac{N}{2}+1)}}{2}$$

Definizione di mediana

Sia $y_1, y_2, \dots, y_N$ una distribuzione statistica disaggregata.
Indichiamo con $y_{(1)}, y_{(2)}, \dots, y_{(N)}$ le osservazioni ordinate:

$$y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(N)}$$

Se N è dispari (quindi $N/2$ non è intero, ma $(N+1)/2$ sì)

$$\underbrace{y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(\frac{N-1}{2})}}_{\frac{N-1}{2} \text{ osservazioni}} \leq y_{(\frac{N+1}{2})} \leq \underbrace{y_{(\frac{N+3}{2})} \leq \dots \leq y_{(N)}}_{\frac{N-1}{2} \text{ osservazioni}}$$

e la mediana è

$$m = y_{(\frac{N+1}{2})}$$

Definizione di mediana

Mediana

La **mediana** delle osservazioni $y_1, \dots, y_N$ è

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

dove $y_{(1)}, y_{(2)}, \dots, y_{(N)}$ sono le osservazioni ordinate.

Quartili

Possiamo 'generalizzare' quello che abbiamo fatto per la mediana dividendo le osservazioni in parti più piccole, ad esempio in quarti



Quartili

Possiamo 'generalizzare' quello che abbiamo fatto per la mediana dividendo le osservazioni in parti più piccole, ad esempio in quarti



I valori in corrispondenza ai quali la distribuzione è divisa in quarti sono detti **quartili**.

Quartili e quantili

Possiamo 'generalizzare' quello che abbiamo fatto per la mediana dividendo le osservazioni in parti più piccole, ad esempio in quarti



I valori in corrispondenza ai quali la distribuzione è divisa in quarti sono detti **quartili**.



Più in generale

Quantile

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

Quantili: esempio

Consideriamo la distribuzione delle ore di TV in un campione più ampio ($N = 590$), cerchiamo il quantile $q_{0.1}$: valore che “lascia a sinistra” il 10% delle osservazioni.

Modalità	Freq	Fr. cum.
0	43	43
1	29	72
2	52	124
3	43	167
4	37	204
5	43	247
6	43	290
7	31	321
8	37	358
9	1	359
10	87	446
11	1	447
12	17	464
13	1	465
14	29	494
15	35	529
16	3	532
17	2	534
18	8	542
20	20	562
21	10	572
22	1	573
25	6	579
28	2	581
30	6	587
41	1	588
42	1	589
50	1	590

Quantili: esempio

Consideriamo la distribuzione delle ore di TV in un campione più ampio ($N = 590$), cerchiamo il quantile $q_{0.1}$: valore che “lascia a sinistra” il 10% delle osservazioni.



Passiamo alle frequenze relative.

Modalità	Freq	Fr. cum.
0	0.073	0.073
1	0.049	0.122
2	0.088	0.210
3	0.073	0.283
4	0.063	0.346
5	0.073	0.419
6	0.073	0.492
7	0.053	0.544
8	0.063	0.607
9	0.002	0.608
10	0.147	0.756
11	0.002	0.758
12	0.029	0.786
13	0.002	0.788
14	0.049	0.837
15	0.059	0.897
16	0.005	0.902
17	0.003	0.905
18	0.014	0.919
20	0.034	0.953
21	0.017	0.969
22	0.002	0.971
25	0.010	0.981
28	0.003	0.985
30	0.010	0.995
41	0.002	0.997
42	0.002	0.998
50	0.002	1.000

Quantili: esempio

Consideriamo la distribuzione delle ore di TV in un campione più ampio ($N = 590$), cerchiamo il quantile $q_{0.1}$: valore che “lascia a sinistra” il 10% delle osservazioni.



Passiamo alle frequenze relative.



La frequenza relativa cumulata in 0 è 0.073, inferiore a 0.1, quella in 1 è 0.122, superiore a 0.1, quindi $q_{0.1} = 1$.

Modalità	Freq	Fr. cum.
0	0.073	0.073
1	0.049	0.122
2	0.088	0.210
3	0.073	0.283
4	0.063	0.346
5	0.073	0.419
6	0.073	0.492
7	0.053	0.544
8	0.063	0.607
9	0.002	0.608
10	0.147	0.756
11	0.002	0.758
12	0.029	0.786
13	0.002	0.788
14	0.049	0.837
15	0.059	0.897
16	0.005	0.902
17	0.003	0.905
18	0.014	0.919
20	0.034	0.953
21	0.017	0.969
22	0.002	0.971
25	0.010	0.981
28	0.003	0.985
30	0.010	0.995
41	0.002	0.997
42	0.002	0.998
50	0.002	1.000

Quantili: esempio

Consideriamo la distribuzione delle ore di TV in un campione più ampio ($N = 590$), cerchiamo il quantile $q_{0.1}$: valore che “lascia a sinistra” il 10% delle osservazioni.



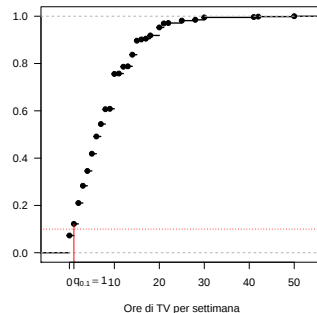
Passiamo alle frequenze relative.



La frequenza relativa cumulata in 0 è 0.073, inferiore a 0.1, quella in 1 è 0.122, superiore a 0.1, quindi $q_{0.1} = 1$.



Possiamo leggere il quantile facilmente sul grafico della funzione di ripartizione empirica.



Mediana come sintesi

Due campioni con la stessa mediana

1, 2, 4, 5, 12, 15



1, 2, 4, 5, 12, 40



La mediana non riflette la presenza dell'osservazione estrema.



Questa caratteristica della mediana non è di per sé uno svantaggio, ma va tenuta presente.

Mediana come sintesi

Due campioni con la stessa mediana

1, 2, 4, 5, 12, 15



1, 2, 4, 5, 12, 40



Una misura di posizione alternativa, sensibile agli estremi è la media aritmetica:

$$1, 2, 4, 5, 12, 15 \rightarrow (1 + 2 + 4 + 5 + 12 + 15)/6 = 6.5$$

$$1, 2, 4, 5, 12, 40 \rightarrow (1 + 2 + 4 + 5 + 12 + 40)/6 = 10.67$$

Media aritmetica

Media aritmetica

La **media aritmetica** delle osservazioni $y_1, \dots, y_N$, indicata con $\bar{y}$, è

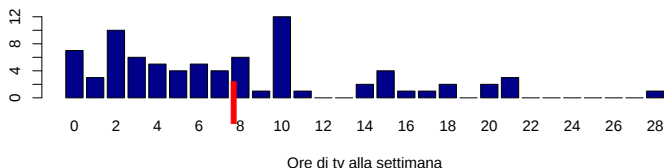
$$\bar{y} = \frac{y_1 + y_2 + \dots + y_N}{N} = \frac{1}{N} \sum_{i=1}^N y_i,$$

Esistono altri tipi di “medie”. Quella aritmetica è senza dubbio quella di utilizzo più comune. Per questo motivo, viene comunemente indicata come “la media” senza nessuna ulteriore aggettivazione.

Esempio: media e mediana per le ore di TV

Abbiamo $N = 80$. Quindi:

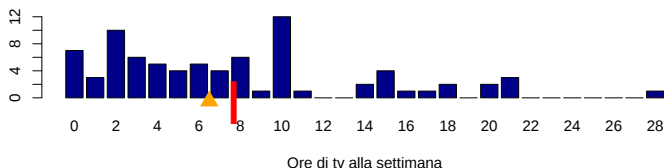
$$\frac{1}{N} \sum_{i=1}^N y_i = \frac{615}{80} = 7.69.$$



Esempio: media e mediana per le ore di TV

Abbiamo $N = 80$. Quindi:

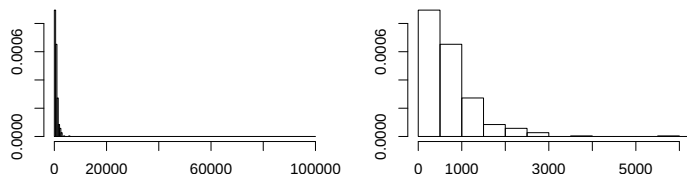
$$\frac{1}{N} \sum_{i=1}^N y_i = \frac{615}{80} = 7.69.$$



La mediana (6.5) è relativamente vicina.

Esempio: media e mediana per il n. di amici su Facebook

Consideriamo delle osservazioni relative al numero di amici su Facebook (il grafico a dx è un ingrandimento che esclude l'osservazione pari a 100000).



Si ha

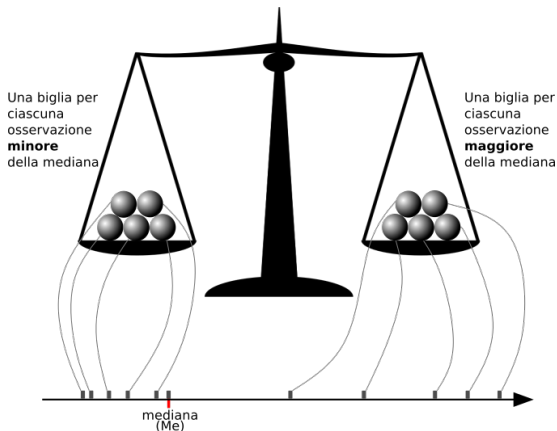
- Mediana: $m = 590$
- Media: $\bar{y} = 944.7$

La grande differenza si deve in parte all'osservazione estrema ($y_{(N)} = 100000$), togliendola si ha

$$m^{(-N)} = 589.5, \quad \bar{y}^{(-N)} = 754.2$$

la differenza rimane relativamente grande perché la distribuzione è asimmetrica (vedi poi).

Media e mediana

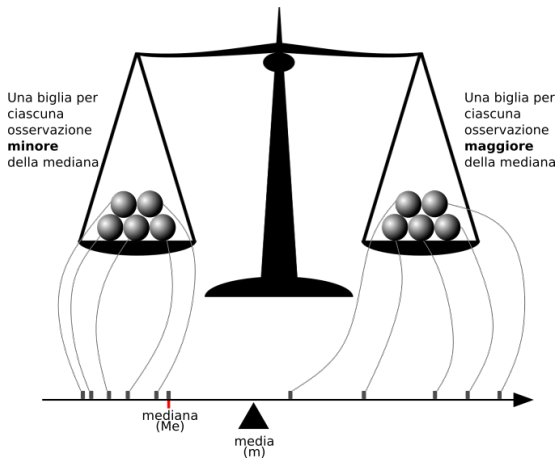


La mediana produce un equilibrio tenendo metà delle osservazioni da una parte, metà dall'altra, **indifferente quanto sono distanti**.



La media è il baricentro, ossia equilibra delle masse (uguali) poste in corrispondenza alle osservazioni.

Media e mediana



La mediana produce un equilibrio tenendo metà delle osservazioni da una parte, metà dall'altra, **indifferente quanto sono distanti.**



La media è il baricentro, ossia equilibra delle masse (uguali) poste in corrispondenza alle osservazioni.

Misure (di posizione) di popolazione e campionarie

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.

Misure di posizione e distribuzioni condizionate

Le misure di posizione sintetizzano (la posizione di) una distribuzione di frequenze.



Il concetto si applica tanto alle distribuzioni marginali quanto alle distribuzioni condizionate.



Ad esempio:

Sia Y l'altezza e X il genere (con valori M e F). Possiamo calcolare misure di posizione dell'altezza condizionata al genere e misure marginali

- ▶ Mediana di $Y|X = M \longrightarrow 180$ (mediana condizionata)
- ▶ Media di $Y|X = M \longrightarrow 180.2$ (media condizionata)
- ▶ Mediana di $Y|X = F \longrightarrow 165$ (mediana condizionata)
- ▶ Media di $Y|X = F \longrightarrow 165.7$ (media condizionata)
- ▶ Mediana di $Y \longrightarrow 174.5$ (mediana marginale)
- ▶ Media di $Y \longrightarrow 173.1$ (media marginale)

Indice

Misure di posizione

Calcolo delle misure di posizione

Variabilità

Media e varianza per dati in gruppi

Forma di una distribuzione

Proprietà delle misure di posizione e scala

Calcolo della media con una distribuzione di frequenza

y_i	n_i	f_i
5	1	0.012
6	10	0.125
7	31	0.388
8	32	0.400
9	6	0.075

Calcolo della media con una distribuzione di frequenza

y_i	n_i	f_i
5	1	0.012
6	10	0.125
7	31	0.388
8	32	0.400
9	6	0.075

$5, \underbrace{6, \dots, 6}_{10 \text{ OSS.}}, \underbrace{7, \dots, 7}_{31 \text{ OSS.}}, \underbrace{8, \dots, 8}_{32 \text{ OSS.}}, \underbrace{9, \dots, 9}_{6 \text{ OSS.}}$

Calcolo della media con una distribuzione di frequenza

y_i	n_i	f_i
5	1	0.012
6	10	0.125
7	31	0.388
8	32	0.400
9	6	0.075

$$5 + \underbrace{6 + \dots + 6}_{10 \text{ volte}} + \underbrace{7 + \dots + 7}_{31 \text{ volte}} + \underbrace{8 + \dots + 8}_{32 \text{ volte}} + \underbrace{9 + \dots + 9}_{6 \text{ volte}}$$

Calcolo della media con una distribuzione di frequenza

y_i	n_i	f_i
5	1	0.012
6	10	0.125
7	31	0.388
8	32	0.400
9	6	0.075

$$5 + \underbrace{6 + \dots + 6}_{=10 \times 6} + \underbrace{7 + \dots + 7}_{=31 \times 7} + \underbrace{8 + \dots + 8}_{=32 \times 8} + \underbrace{9 + \dots + 9}_{=6 \times 9}$$

Calcolo della media con una distribuzione di frequenza

y_i	n_i	f_i
5	1	0.012
6	10	0.125
7	31	0.388
8	32	0.400
9	6	0.075

Si ha, con $y_1, \dots, y_k$ le modalità,

$$\sum_{i=1}^k y_i n_i = 5 \times 1 + 6 \times 10 + 7 \times 31 + 8 \times 32 + 9 \times 6$$

Calcolo della media con una distribuzione di frequenza

y_i	n_i	f_i
5	1	0.012
6	10	0.125
7	31	0.388
8	32	0.400
9	6	0.075

Si ha, con $y_1, \dots, y_k$ le modalità,

$$\bar{y} = \frac{\sum_{i=1}^k y_i n_i}{\sum_{i=1}^k n_i} =$$

Calcolo della media con una distribuzione di frequenza

y_i	n_i	f_i
5	1	0.012
6	10	0.125
7	31	0.388
8	32	0.400
9	6	0.075

Si ha, con $y_1, \dots, y_k$ le modalità,

$$\bar{y} = \frac{\sum_{i=1}^k y_i n_i}{\sum_{i=1}^k n_i} = \sum_{i=1}^k y_i f_i$$

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

Media aritmetica ponderata

La media aritmetica calcolata per dati raggruppati è una **media aritmetica ponderata** delle modalità, con pesi le frequenze.

In generale, una media aritmetica ponderata, con pesi w_i è

$$\bar{y}_w = \frac{\sum_{i=1}^k y_i w_i}{\sum_{i=1}^k w_i}$$

dove ad ogni valore y_i è assegnato un peso non negativo w_i .

Calcolo con una distribuzione di frequenza per classi

Supponiamo di disporre di una distribuzione di frequenze per classi

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

questa non è un'informazione completa (rispetto al conoscere tutte le osservazioni o, che è lo stesso, la distribuzione di frequenze).

Non possiamo, con la sola informazione sulle frequenze delle classi, calcolare media e mediana, possiamo calcolarne delle approssimazioni.

Mediana con una 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. Risposte possibili:

Mediana con una 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. Risposte possibili:

- $m \in (2, 3]$.

Mediana con una 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. Risposte possibili:

- $m \in (2, 3]$.
- Supponiamo (arbitrariamente) che i quattro dati appartenenti al terzo intervallo siano equidistribuiti, cioè che

$$y_{(6)} = 2.25, y_{(7)} = 2.50, y_{(8)} = 2.75, y_{(9)} = 3.00$$

e allora

$$m = \frac{2.25 + 2.50}{2} = 2.375$$

Media con una 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 con una 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.

Un'approssimazione spesso usata in questi casi è

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

dove y_i è il punto centrale della classe i -sima, ovvero

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

Esempio: voti matura

voto (classe) $(c_{i-1}, c_i]$	frequenza assoluta n_i	punto centrale della classe y_i	$y_i n_i$
[60,70]	15	65.0	975.0
(70,80]	35	75.5	2642.5
(80,90]	22	85.5	1881.0
(90,100]	8	95.5	764.0
Totale			6262.5

Da cui la media

$$\bar{y} = \frac{6262.5}{80} = 78.28$$

e la mediana

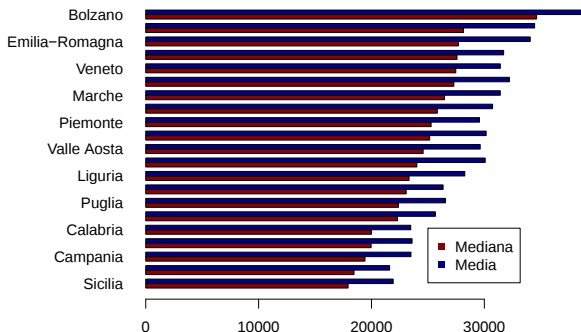
$$m \in (70, 80]$$

La media calcolata con i dati completi è $\bar{y} = 78.38$, la mediana è 77.5.

A cosa mi servono gli indici di posizione?

Ad esempio per fare confronti

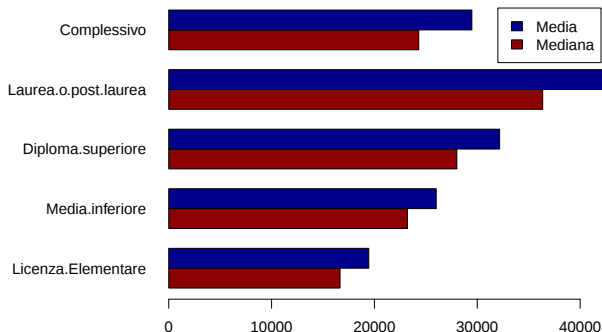
... dei redditi nelle regioni italiane



A cosa mi servono gli indici di posizione?

Ad esempio per fare confronti

... dei redditi per titolo di studio



Indice

Misure di posizione

Variabilità

Media e varianza per dati in gruppi

Forma di una distribuzione

Proprietà delle misure di posizione e scala

La media non basta

Due gruppi di individui, la variabile è l'altezza (*cm*)

150, 151, 156, 146, 157



121, 150, 190, 180, 119

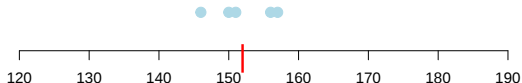


La media è 152cm per entrambi i gruppi.
I gruppi sono diversi!

Misure elementari di variabilità

Due campioni con la stessa media

146, 150, 151, 156, 157



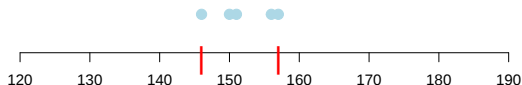
121, 124, 148, 180, 187



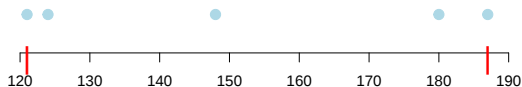
Misure elementari di variabilità: *range*

Due campioni con la stessa media

146, 150, 151, 156, 157



121, 124, 148, 180, 187

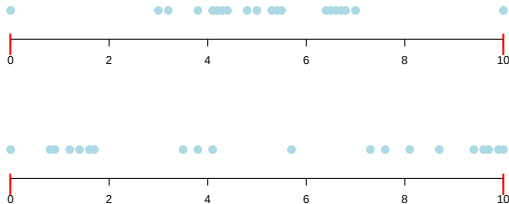


Una misura intuitiva della variabilità di un insieme di dati è la differenza (distanza) tra minimo e massimo, detta **campo di variazione** (range)

$$\text{Range} = y_{(N)} - y_{(1)}$$

Misure elementari di variabilità

Due campioni con lo stesso *range*

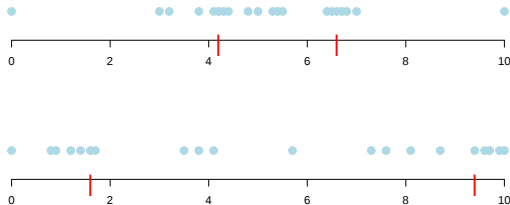


Usare minimo e massimo significa fare riferimento ai casi più estremi.



Misure elementari di variabilità *range*

Due campioni con lo stesso *range*



Usare minimo e massimo significa fare riferimento ai casi più estremi.



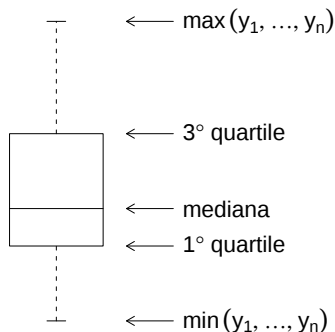
Un'alternativa è considerare la differenza (distanza) tra i quartili, o **distanza interquartilica** (IQR)

$$IQR = q_{0.75} - q_{0.25}$$

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

Fornisce una 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 (variante comune)

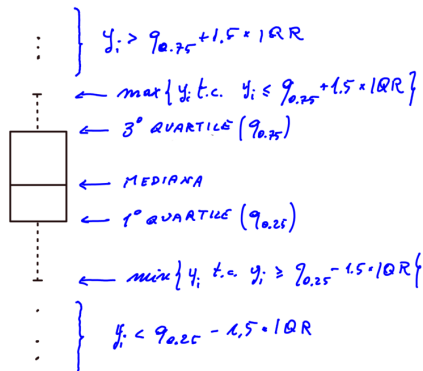
Una variante del diagramma a scatola prevede che i baffi non si estendano sempre fino alle osservazioni più estreme, ed è costruita come segue:

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 $\text{cost} \times (\text{scarto interquartile})$ 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).

La logica è di evidenziare le osservazioni estreme.

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

Schematicamente, la seconda versione del boxplot è come segue



Esempio: costruzione di un boxplot

Dati (già ordinati):

1.1 1.3 1.4 1.6 1.8 1.9 2.0 2.5 2.9 3.2 4.1 5.6

- ▶ $q_{0.25} = 1.5$, $m = 1.95$, $q_{0.75} = 3.05$,
- ▶ $1.5 \times (q_{0.75} - q_{0.25}) = 1.5 \times 1.55 = 2.325$.

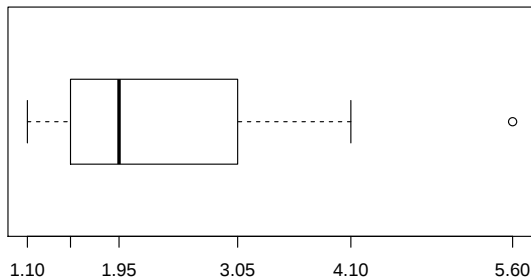
quindi

1. **scatola**: da 1.5 a 3.05 con la mediana indicata da una linea a 1.95;
2. **bafo inferiore**: fino all'osservazione più bassa tra quelle maggiori di $q_{0.25} - 2.325 = -0.825$, ovvero fino a 1.1;
3. **bafo 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 4.375; in questo caso solamente l'osservazione risultata uguale a 5.6.

Esempio: costruzione di un boxplot

Dati (già ordinati):

1.1 1.3 1.4 1.6 1.8 1.9 2.0 2.5 2.9 3.2 4.1 5.6



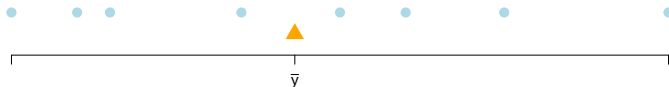
Variabilità e distanza dalla media

La variabilità delle osservazioni può essere misurata in base alla distanza dal centro



Variabilità e distanza dalla media

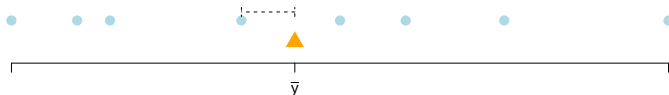
La variabilità delle osservazioni può essere misurata in base alla distanza dal centro



Consideriamo come centro la media aritmetica $\bar{y}$.

Variabilità e distanza dalla media

La variabilità delle osservazioni può essere misurata in base alla distanza dal centro



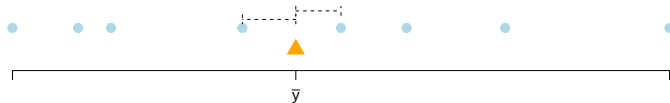
Consideriamo come centro la media aritmetica $\bar{y}$.

Misuriamo, per ciascuna osservazione, quanto dista dal centro come

$$(y_i - \bar{y})^2$$

Variabilità e distanza dalla media

La variabilità delle osservazioni può essere misurata in base alla distanza dal centro

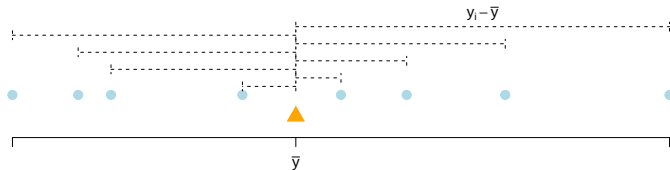


Misuriamo, per ciascuna osservazione, quanto dista dal centro come

$$(y_i - \bar{y})^2$$

Variabilità e distanza dalla media

La variabilità delle osservazioni può essere misurata in base alla distanza dal centro

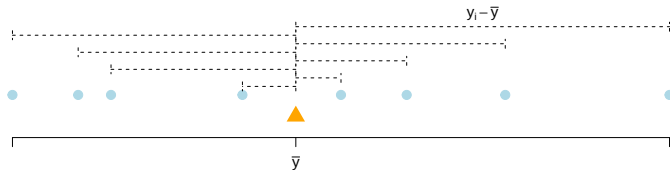


Misuriamo, per ciascuna osservazione, quanto dista dal centro come

$$(y_i - \bar{y})^2$$

Variabilità e distanza dalla media

La variabilità delle osservazioni può essere misurata in base alla distanza dal centro



Misuriamo, per ciascuna osservazione, quanto dista dal centro come

$$(y_i - \bar{y})^2$$

infine, facciamo la media di queste quantità

$$\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2$$

Varianza

Varianza

La **varianza** delle osservazioni $y_1, \dots, y_N$ è la media dei quadrati degli scarti di ogni osservazione dalla media aritmetica.

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

La varianza della variabile Y si indica anche con σ_Y^2 o $V(Y)$.

Varianza: esempio

Esempio: varianza per 5 osservazioni, la media è $\bar{y} = 2.8$

Osservazioni	scarti	(scarti) ²
y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
-1	-3.80	14.44
1	-1.80	3.24
3	0.20	0.04
4	1.20	1.44
7	4.20	17.64
Totale		36.8

La varianza è

$$\sigma^2 = \frac{36.8}{5} = 7.36$$

Varianza con distribuzioni di frequenza

Se la variabile Y ha modalità $y_1, \dots, y_k$ con frequenze assolute $n_1, \dots, n_k$ ($\sum_{i=1}^k n_i = N$) e frequenze relative $f_1, \dots, f_k$ ($f_i = n_i/N$) la **varianza** si calcola con

$$\sigma^2 = \frac{\sum_{i=1}^k n_i (y_i - \bar{y})^2}{N} = \sum_{i=1}^k f_i (y_i - \bar{y})^2$$

Varianza con distribuzioni di frequenza

Se la variabile Y ha modalità $y_1, \dots, y_k$ con frequenze assolute $n_1, \dots, n_k$ ($\sum_{i=1}^k n_i = N$) e frequenze relative $f_1, \dots, f_k$ ($f_i = n_i/N$) la **varianza** si calcola con

$$\sigma^2 = \frac{\sum_{i=1}^k n_i (y_i - \bar{y})^2}{N} = \sum_{i=1}^k f_i (y_i - \bar{y})^2$$

Esempio: ore di sonno per notte, $N = 80$, $\bar{y} = 7.4$

Modalità y_i	Frequenza n_i	scarti $y_i - \bar{y}$	(scarti) ² $(y_i - \bar{y})^2$	scarti pesati $n_i (y_i - \bar{y})^2$
5	1	-2.40	5.7600	5.7600
6	10	-1.40	1.9600	19.6000
7	31	-0.40	0.1600	4.9600
8	32	0.60	0.3600	11.5200
9	6	1.60	2.5600	15.3600
Totale				57.2

La varianza è

$$\sigma^2 = \frac{57.2}{80} = 0.72$$

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 sonno si calcola come:

$$\sigma = \sqrt{0.72} = 0.85$$

Devianza

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

$$\sum_{i=1}^N (y_i - \bar{y})^2$$

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

Varianza campionaria corretta

Quando si lavora con un campione

$$y_1, \dots, y_N$$

si utilizza spesso la **varianza campionaria corretta**, che differisce dalla varianza campionaria solo per il denominatore.

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

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.

Caso limite: assenza di variabilità

È utile tenere presente il caso limite in cui tutte le unità statistiche mostrano la stessa modalità del carattere, ovvero la variabilità è assente, come in

42 42 42 42 42 42 42 42 42 42 42 42

Abbiamo allora

► $y_{(1)} = 42$ $y_{(N)} = 42$ quindi

$$\text{Range: } y_{(N)} - y_{(1)} = 0,$$

► $q_{0.25} = 42$ $m = 42$ $q_{0.75} = 42$ quindi

$$\text{IQR: } q_{0.75} - q_{0.25} = 0,$$

► $\bar{y} = 42 = y_i$, quindi $(y_i - \bar{y}) = 0$, quindi

$$\sigma^2 = 0,$$

Indice

Misure di posizione

Variabilità

Calcolo delle misure di variabilità

Media e varianza per dati in gruppi

Forma di una distribuzione

Proprietà delle misure di posizione e scala

Varianza: una formula operativa

Si osservi che

$$\begin{aligned}
 \sigma^2 &= \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2 = \\
 &= \frac{1}{N} \sum_{i=1}^N y_i^2 + \frac{1}{N} \sum_{i=1}^N \bar{y}^2 - \frac{1}{N} \sum_{i=1}^N 2\bar{y}y_i = \\
 &= \frac{1}{N} \sum_{i=1}^N y_i^2 + \frac{N\bar{y}^2}{N} - \frac{2\bar{y}}{N} \sum_{i=1}^N y_i = \\
 &= \frac{1}{N} \sum_{i=1}^N y_i^2 + \bar{y}^2 - 2\bar{y}^2 \\
 &= \frac{1}{N} \sum_{i=1}^N y_i^2 - \bar{y}^2
 \end{aligned}$$

Varianza: una formula operativa (cont)

Quindi, possiamo scrivere

$$\sigma^2 = \left(\frac{1}{N} \sum_{i=1}^N y_i^2 \right) - \bar{y}^2$$

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

Esempio: varianza per 5 osservazioni, la media è $\bar{y} = 2.8$

Osservazioni	scarti	(scarti) ²	(osservazioni) ²
y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	y_i^2
-1	-3.80	14.44	1
1	-1.80	3.24	1
3	0.20	0.04	9
4	1.20	1.44	16
7	4.20	17.64	49
Totale		36.8	76

La varianza con la “vecchia” formula è

$$\sigma^2 = \frac{36.8}{5} = 7.36$$

La varianza con la “nuova” formula è

$$\sigma^2 = \frac{76}{5} - 2.8^2 = 7.36$$

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

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

Osservazioni:

- ▶ $s = -1$: media armonica.
- ▶ $s = 1$: media aritmetica.
- ▶ $s = 2$: media quadratica.
- ▶ $s \rightarrow 0$: 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	$[c_0, c_1)$	$[c_1, c_2)$	$\dots$	$[c_{k-1}, c_k)$
frequenze assolute	n_1	n_2	$\dots$	n_k

dove k indica il numero degli intervalli.

Varianza: distribuzione di frequenza per classi

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

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

dove k indica il numero degli intervalli.



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.



Si commette un errore che è tanto maggiore quanto più larghe sono le classi.

Esempio: voti matura

voto (classe) $(c_{i-1}, c_i]$	frequenza assoluta n_i	punto centrale della classe y_i	$y_i n_i$	y_i^2	$y_i^2 n_i$
[60,70]	15	65.0	975.0	4225.0	63375.0
(70,80]	35	75.5	2642.5	5700.2	199508.8
(80,90]	22	85.5	1881.0	7310.2	160825.5
(90,100]	8	95.5	764.0	9120.2	72962.0
Totale			6262.5		496671.25

Da cui

$$M(y^2) = \frac{496671.25}{80} = 6208.39$$

e ricordando che $\bar{y} = \frac{6262.5}{80} = 78.28$ si ha la varianza come

$$6208.39 - 78.28^2 = 80.63$$

La varianza calcolata a partire dai dati grezzi è 80.28, naturalmente diversa.

Indice

Misure di posizione

Variabilità

Media e varianza per dati in gruppi

Forma di una distribuzione

Proprietà delle misure di posizione e scala

Dati suddivisi in gruppi

In molte situazioni ci troviamo a voler studiare il comportamento di una variabile (Y) condizionatamente a un'altra (X)



In altre parole, suddividiamo i dati in gruppi in base alle modalità di X e studiamo le distribuzioni di Y nei gruppi.



Ad esempio, sia

► $Y \longrightarrow$ altezza

► $X \longrightarrow$ sesso, con modalità $x_1 = M$, $x_2 = F$



Al fine di studiare Y condizionatamente a X ha senso calcolare misure di sintesi condizionate per le distribuzioni di $Y|X = x_1$ e $Y|X = x_2$.

Esempio: medie condizionate, altezza (Y) e sesso (X)

I gruppi hanno numerosità $N_1 = 20, N_2 = 30$.

Le medie di gruppo sono pari a

$$\bar{y}_1 = \frac{3626}{N_1} = 181.3, \quad \bar{y}_2 = \frac{4939}{N_2} = 164.63$$

La media generale è il totale delle $N = N_1 + N_2$ osservazioni diviso N

$$\bar{y} = \frac{3626 + 4939}{N_1 + N_2} = \frac{8565}{50} = 171.3$$

ma anche

$$\bar{y} = \frac{1}{N}(N_1 \times 181.3 + N_2 \times 164.63)$$

**media pesata delle medie di gruppo con pesi
pari alle numerosità dei gruppi.**

$Y X = M$	$Y X = F$
174	160
170	173
179	168
176	175
170	165
180	160
197	158
181	158
190	157
180	170
187	160
182	163
181	165
187	164
180	174
180	158
185	178
190	169
170	168
187	147
	161
	160
	167
	175
	165
	170
	165
	162
	165
	159
3626	4939

Media marginale e medie condizionate

Supponiamo di avere N unità statistiche suddivise in L gruppi, secondo le modalità $x_1, \dots, x_L$ di una variabile X . Siano $N_j, j = 1, \dots, L$, le numerosità di ciascun gruppo ($N = \sum_{j=1}^L N_j$).

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

$Y X = x_1$	$Y X = x_2$	...	$Y X = x_j$	...	$Y X = x_L$
$y_{1,1}$	$y_{1,2}$	...	$y_{1,j}$	...	$y_{1,L}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$y_{N_1,1}$	$\vdots$		$\vdots$		$\vdots$
	$\vdots$		$y_{N_j,j}$		$\vdots$
	$y_{N_2,2}$				$y_{N_L,L}$

Media marginale e medie condizionate

$Y X = x_1$	$Y X = x_2$	...	$Y X = x_j$	...	$Y X = x_L$
$y_{1,1}$	$y_{1,2}$	...	$y_{1,j}$	...	$y_{1,L}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$y_{N_1,1}$	$\vdots$		$\vdots$		$\vdots$
	$\vdots$		$y_{N_j,j}$		$\vdots$
	$y_{N_2,2}$				$y_{N_L,L}$
$\sum_{i=1}^{N_1} y_{i,1}$	$\sum_{i=1}^{N_2} y_{i,2}$	...	$\sum_{i=1}^{N_j} y_{i,j}$	...	$\sum_{i=1}^{N_L} y_{i,L}$

Per ogni gruppo j si ha la media condizionata

$$\bar{y}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} y_{i,j}.$$

Media marginale e medie condizionate

$Y X = x_1$	$Y X = x_2$	...	$Y X = x_j$	...	$Y X = x_L$
$y_{1,1}$	$y_{1,2}$	...	$y_{1,j}$	...	$y_{1,L}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$y_{N_1,1}$	$\vdots$		$\vdots$		$\vdots$
	$\vdots$		$y_{N_j,j}$		$\vdots$
	$y_{N_2,2}$				$y_{N_L,L}$
$\sum_{i=1}^{N_1} y_{i,1}$	$\sum_{i=1}^{N_2} y_{i,2}$	...	$\sum_{i=1}^{N_j} y_{i,j}$	...	$\sum_{i=1}^{N_L} y_{i,L}$
$\bar{y}_1$	$\bar{y}_2$	...	$\bar{y}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} y_{i,j}$	...	$\bar{y}_L$

Per ogni gruppo j si ha la media condizionata

$$\bar{y}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} y_{i,j}.$$

Media marginale e medie condizionate

$Y X = x_1$	$Y X = x_2$	...	$Y X = x_j$	...	$Y X = x_L$	
$y_{1,1}$	$y_{1,2}$	...	$y_{1,j}$	...	$y_{1,L}$	
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
$y_{N_1,1}$	$\vdots$		$\vdots$		$\vdots$	
	$\vdots$		$y_{N_j,j}$		$\vdots$	
	$y_{N_2,2}$	...			$y_{N_L,L}$	
$\sum_{i=1}^{N_1} y_{i,1}$	$\sum_{i=1}^{N_2} y_{i,2}$	...	$\sum_{i=1}^{N_j} y_{i,j}$	...	$\sum_{i=1}^{N_L} y_{i,L}$	$\sum_{j=1}^L \sum_{i=1}^{N_j} y_{i,j}$
$\bar{y}_1$	$\bar{y}_2$	...	$\bar{y}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} y_{i,j}$	...	$\bar{y}_L$	

La media generale è

$$\bar{y} = \frac{1}{N} \sum_{j=1}^L \sum_{i=1}^{N_j} y_{i,j}$$

Media marginale e medie condizionate

$Y X = x_1$	$Y X = x_2$	...	$Y X = x_j$	...	$Y X = x_L$	
$y_{1,1}$	$y_{1,2}$	...	$y_{1,j}$	...	$y_{1,L}$	
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
$y_{N_1,1}$	$\vdots$		$\vdots$		$\vdots$	
	$\vdots$		$y_{N_j,j}$		$\vdots$	
	$y_{N_2,2}$	...			$y_{N_L,L}$	
$\sum_{i=1}^{N_1} y_{i,1}$	$\sum_{i=1}^{N_2} y_{i,2}$	...	$\sum_{i=1}^{N_j} y_{i,j}$	...	$\sum_{i=1}^{N_L} y_{i,L}$	$\sum_{j=1}^L \sum_{i=1}^{N_j} y_{i,j}$
$\bar{y}_1$	$\bar{y}_2$	...	$\bar{y}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} y_{i,j}$	...	$\bar{y}_L$	$\bar{y} = \frac{1}{N} \sum_{j=1}^L \sum_{i=1}^{N_j} y_{i,j}$

La media generale è

$$\bar{y} = \frac{1}{N} \sum_{j=1}^L \sum_{i=1}^{N_j} y_{i,j}$$

Media marginale e medie condizionate

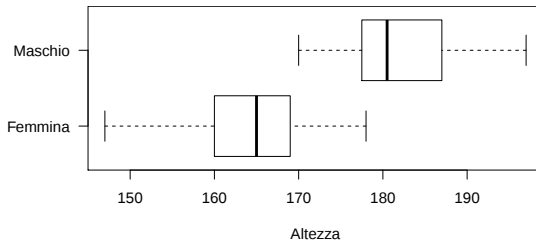
$Y X = x_1$	$Y X = x_2$	...	$Y X = x_j$	...	$Y X = x_L$	
$y_{1,1}$	$y_{1,2}$	...	$y_{1,j}$	...	$y_{1,L}$	
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
$y_{N_1,1}$	$\vdots$		$\vdots$		$\vdots$	
	$\vdots$		$y_{N_j,j}$		$\vdots$	
	$y_{N_2,2}$				$y_{N_L,L}$	
$\sum_{i=1}^{N_1} y_{i,1}$	$\sum_{i=1}^{N_2} y_{i,2}$	...	$\sum_{i=1}^{N_j} y_{i,j}$	...	$\sum_{i=1}^{N_L} y_{i,L}$	$\sum_{j=1}^L \sum_{i=1}^{N_j} y_{i,j}$
$\bar{y}_1$	$\bar{y}_2$	...	$\bar{y}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} y_{i,j}$	...	$\bar{y}_L$	$\bar{y} = \frac{1}{N} \sum_{j=1}^L \sum_{i=1}^{N_j} y_{i,j}$

La media generale è

$$\bar{y} = \frac{1}{N} \sum_{j=1}^L \underbrace{\sum_{i=1}^{N_j} y_{i,j}}_{= N_j \bar{y}_j} = \frac{1}{N} \sum_{j=1}^L N_j \bar{y}_j$$

Boxplot condizionati, sempre con le altezze M/F

Anche mediana e quantili possono essere calcolati per le distribuzioni condizionate, di conseguenza possiamo disegnare dei *boxplot* condizionati



Affiancare due o più *boxplot* è un modo molto efficace di confrontare due distribuzioni.

Esempio: varianze condizionate, altezza (Y) e sesso (X)

Riconsideriamo le osservazioni sulle altezze di $N_1 = 20$ maschi e $N_2 = 30$ femmine, per i quali sappiamo che le medie di gruppo sono $\bar{y}_1 = 181.3$ e $\bar{y}_2 = 164.63$, mentre la media generale è $\bar{y} = 171.3$. Possiamo calcolare le varianze per i due gruppi,

$$\sigma_1^2 = \frac{1}{N_1} \sum_{i=1}^{N_1} (y_{i,1} - \bar{y}_1)^2 = 49.31, \quad \sigma_2^2 = 43.3$$

e la varianza totale (marginale)

$$\sigma^2 = \frac{1}{N_1 + N_2} \left(\sum_{i=1}^{N_1} (y_{i,1} - \bar{y})^2 + \sum_{i=1}^{N_2} (y_{i,2} - \bar{y})^2 \right) = 112.37,$$

Vediamo se c'è un legame anche tra queste.

$Y X = M$	$Y X = F$
174	160
170	173
179	168
176	175
170	165
180	160
197	158
181	158
190	157
180	170
187	160
182	163
181	165
187	164
180	174
180	158
185	178
190	169
170	168
187	147
	161
	160
	167
	175
	165
	170
	165
	162
	165
	159
3626	4939

Varianza marginale e varianze condizionate

$Y X = x_1$	$Y X = x_2$	...	$Y X = x_j$	...	$Y X = x_L$
$y_{1,1}$	$y_{1,2}$	...	$y_{1,j}$	...	$y_{1,L}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$y_{N_1,1}$	$\vdots$		$\vdots$		$\vdots$
	$\vdots$	...	$y_{N_j,j}$		$\vdots$
	$y_{N_2,2}$				$y_{N_L,L}$

La varianza marginale è

$$\sigma^2 = \frac{1}{N} \sum_{j=1}^L \sum_{i=1}^{N_j} (y_{i,j} - \bar{y})^2$$

le varianze condizionate sono

$$\sigma_j^2 = \frac{1}{N_j} \sum_{i=1}^{N_j} (y_{i,j} - \bar{y}_j)^2.$$

Formula di scomposizione della varianza.

Si dimostra che

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

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 (within)**, V_w .

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

Esempio: varianze condizionate, altezza (Y) e sesso (X)

Si ha

$$\sigma_1^2 = 49.31, \quad \sigma_2^2 = 43.3$$

e quindi la **varianza entro i gruppi** è

$$V_w = \frac{20}{50}49.31 + \frac{30}{50}43.3 = 45.7$$

la varianza tra i gruppi è

$$V_b = \frac{20}{50}(181.3 - 171.3)^2 + \frac{30}{50}(164.63 - 171.3)^2 = 66.69$$

la cui somma

$$V_w + V_b = 45.7 + 66.69 = 112.39$$

coincide con la varianza totale ($\sigma^2 = 112.37$), a meno di approssimazioni dovute agli arrotondamenti.

Y X = M	Y X = F
174	160
170	173
179	168
176	175
170	165
180	160
197	158
181	158
190	157
180	170
187	160
182	163
181	165
187	164
180	174
180	158
185	178
190	169
170	168
187	147
	161
	160
	167
	175
	165
	170
	165
	162
	165
	159
3626	4939

Esempio: voto di maturità e sesso

Si ha $\sigma^2 = 72.4$, mentre

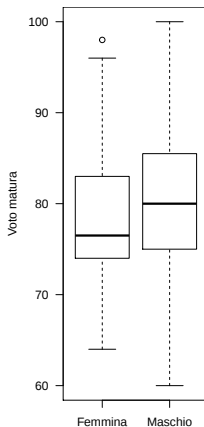
$$V_w = 71.69$$

$$V_b = 0.71$$

In questo caso il contributo di V_b alla varianza totale è minimo, questo perché le medie di gruppo

$$\bar{y}_1 = 79.75, \bar{y}_2 = 78.03$$

sono molto simili.



Indice η^2

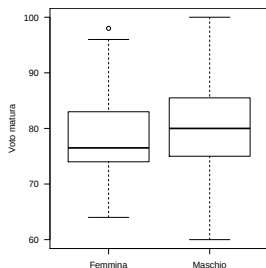
Un indice di quanto diversi sono i gruppi è

$$\begin{aligned}\eta^2 &= \frac{(\text{varianza tra i gruppi})}{(\text{varianza totale})} \\ &= \frac{(\text{varianza tra i gruppi})}{(\text{varianza tra i gruppi}) + (\text{varianza entro i gruppi})} \\ &= \frac{\frac{1}{N} \sum_{j=1}^L N_j (\bar{y}_j - \bar{y})^2}{\sigma^2}\end{aligned}$$

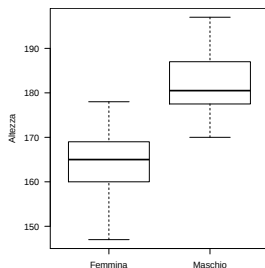
- ▶ è compreso tra 0 e 1
- ▶ è tanto più vicino a 1 quanto più i gruppi sono diversi

Scomposizione della varianza: a cosa serve

È uno strumento per studiare in che misura dei gruppi di unità differiscano **in media** rispetto a una variabile quantitativa.



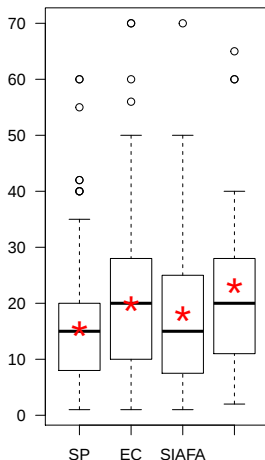
$$\eta^2 = \frac{0.71}{72.4} = 0.01$$



$$\eta^2 = \frac{66.69}{112.37} = 0.59$$

Non c'è differenza in media tra i gruppi quanto a voto di matura, c'è differenza quanto ad altezza.

Scomposizione della varianza: Ore di studio



	N_j	$\bar{y}_j$	σ_j^2
SP	225	15.89	104.92
EC	366	20.45	136.07
SIAFA	51	18.59	183.10
CTF	27	23.67	274.15

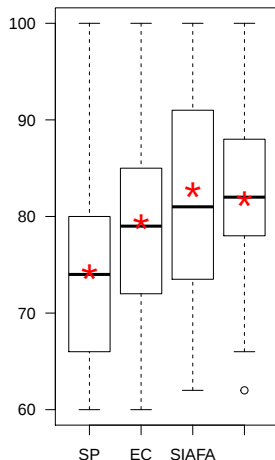
La scomposizione è come segue

$$\begin{aligned} \frac{1}{N} \sum_{j=1}^L N_j \sigma_j^2 &= 134.8 \\ \frac{1}{N} \sum_{j=1}^L N_j (\bar{y}_j - \bar{y})^2 &= 5.294 \\ \hline \sigma^2 &= 140 \end{aligned}$$

Un indice di quanto diversi sono i gruppi è

$$\eta^2 = \frac{5.294}{140} = 0.03781$$

Scomposizione della varianza: Voto di matura



	N_j	$\bar{y}_j$	σ_j^2
SP	230	74.50	108.28
EC	375	79.73	103.73
SIAFA	47	82.98	127.04
CTF	29	82.10	74.51

La scomposizione è come segue

$$\begin{aligned}
 \frac{1}{N} \sum_{j=1}^L N_j \sigma_j^2 &= 105.6 \\
 \frac{1}{N} \sum_{j=1}^L N_j (\bar{y}_j - \bar{y})^2 &= 8.124 \\
 \hline
 \sigma^2 &= 113.8
 \end{aligned}$$

Un indice di quanto diversi sono i gruppi è

$$\eta^2 = \frac{8.124}{113.8} = 0.07139$$

Indice

Misure di posizione

Variabilità

Media e varianza per dati in gruppi

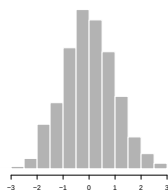
Forma di una distribuzione

Proprietà delle misure di posizione e scala

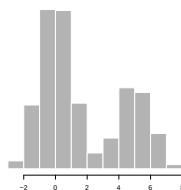
Forma di una distribuzione: numero di mode

Numero di picchi (locali) che mostra l'istogramma: uno (distribuzione **unimodale**), molti (distribuzione **bimodale/multimodale**), o nessuno (distribuzione **uniforme**).

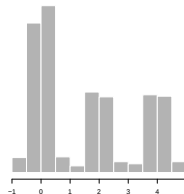
unimodale



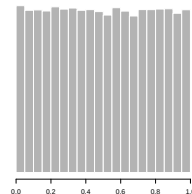
bimodale



multimodale



uniforme



Forma di una distribuzione: numero di mode

Numero di picchi (locali) che mostra l'istogramma: uno (distribuzione **unimodale**), molti (distribuzione **bimodale/multimodale**), o nessuno (distribuzione **uniforme**).

unimodale



bimodale



multimodale



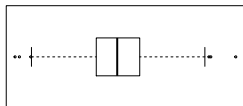
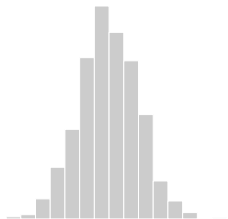
uniforme



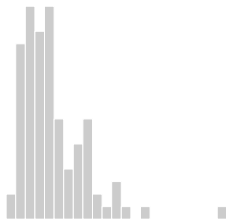
Forma della distribuzione: simmetria

Distinguiamo **simmetria** e **asimmetria** per quest'ultima nelle due possibili direzioni: **a destra** e **a sinistra**.

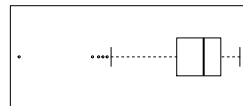
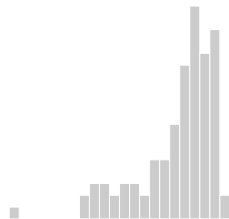
simmetria



asimmetria a destra



asimmetria a sinistra

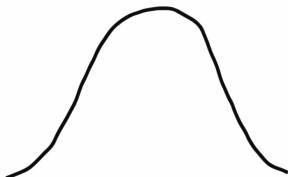


La direzione (destra/sinistra) della asimmetria è data dalla posizione della coda più lunga.

Forma della distribuzione: simmetria

Distinguiamo **simmetria** e **asimmetria** per quest'ultima nelle due possibili direzioni: **a destra** e **a sinistra**.

simmetria



asimmetria a destra



asimmetria a sinistra

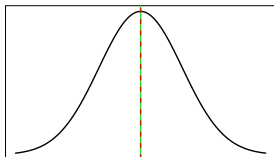


La direzione (destra/sinistra) della asimmetria è data dalla posizione della coda più lunga.

Simmetria: media vs. mediana

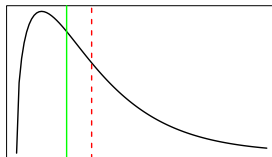
Se la distribuzione è
simmetrica

$\text{media} \approx \text{mediana}$



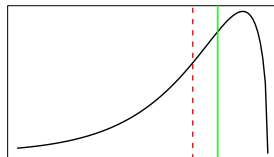
Se la distribuzione è
asimmetrica a destra
(positiva):

$\text{media} > \text{mediana}$



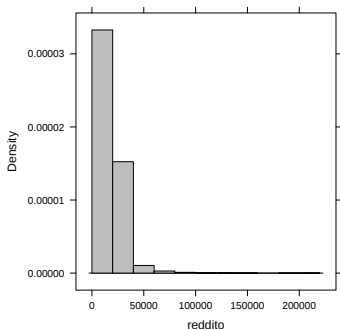
Se la distribuzione è
asimmetrica a sinistra
(negativa):

$\text{media} < \text{mediana}$



Medie e mediane: dati sul reddito

Lavoratori dipendenti



Media 17466

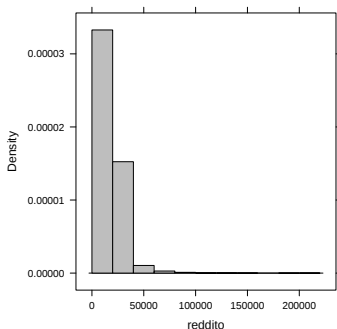
Mediana 16557

Minimo 57

Massimo 214968

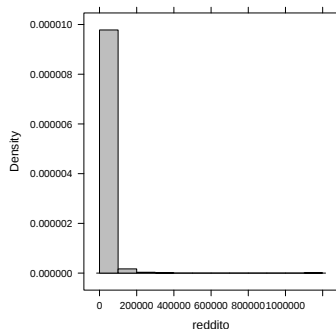
Medie e mediane: dati sul reddito

Lavoratori dipendenti



Media 17466
 Mediana 16557
 Minimo 57
 Massimo 214968

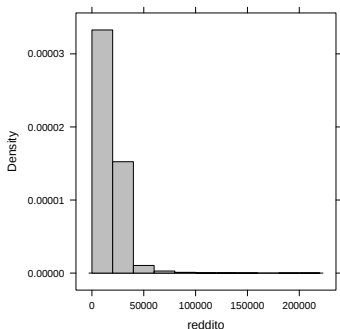
Lavoratori aut con dip



Media 28493
 Mediana 21000
 Minimo 101
 Massimo 1162813

Medie e mediane: dati sul reddito

Lavoratori dipendenti



Media 17466

Mediana 16557

Minimo 57

Massimo 214968

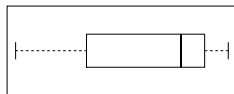
Indice di asimmetria di A.L. Bowley

Una delle prime misure di asimmetria, attribuita a A.L. Bowley e a G.U. Yule, si basa sui quartili

$$B = \frac{(q_{0.75} - m) - (m - q_{0.25})}{q_{0.75} - q_{0.25}} = \frac{q_{0.75} - 2m + q_{0.25}}{q_{0.75} - q_{0.25}}$$

$$-1 \leq B \leq 1$$

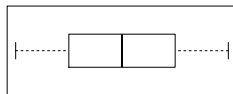
asimmetria a sinistra



$$q_{0.75} - m < m - q_{0.25}$$

$$B < 0$$

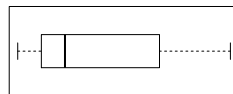
simmetria



$$q_{0.75} - m \approx m - q_{0.25}$$

$$B \approx 0$$

asimmetria a destra



$$q_{0.75} - m > m - q_{0.25}$$

$$B > 0$$

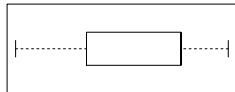
Indice di asimmetria di A.L. Bowley

Una delle prime misure di asimmetria, attribuita a A.L. Bowley e a G.U. Yule, si basa sui quartili

$$B = \frac{(q_{0.75} - m) - (m - q_{0.25})}{q_{0.75} - q_{0.25}} = \frac{q_{0.75} - 2m + q_{0.25}}{q_{0.75} - q_{0.25}}$$

$$-1 \leq B \leq 1$$

asimmetria a sinistra

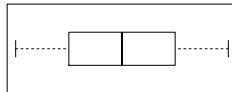


$$q_{0.75} = m$$

$$m - q_{0.25} = q_{0.75} - q_{0.25}$$

$$B = -1$$

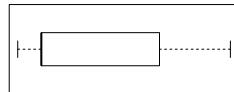
simmetria



$$q_{0.75} - m \approx m - q_{0.25}$$

$$B \approx 0$$

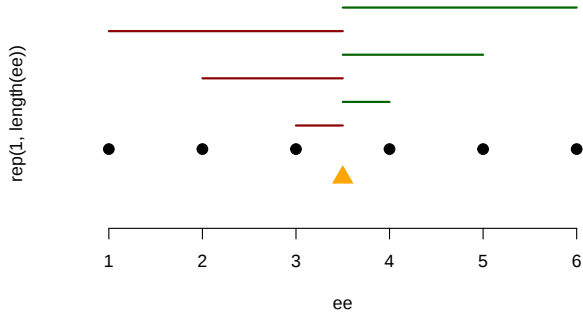
asimmetria a destra



$$q_{0.25} = m$$

$$q_{0.75} - m = q_{0.75} - q_{0.25}$$

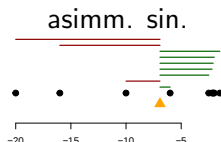
$$B = 1$$



Indice di asimmetria di K. Pearson

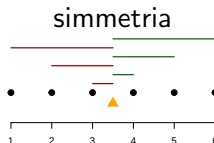
La misura di asimmetria di uso più comune è il cosiddetto **indice di asimmetria standardizzato**, introdotto da K. Pearson e definito come

$$P = \frac{1}{N\sigma^3} \sum_{i=1}^N (y_i - \bar{y})^3 = \frac{\frac{\sum_{i=1}^N (y_i - \bar{y})^3}{N}}{\sigma^3}.$$



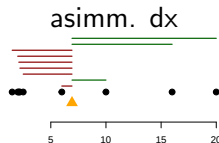
i termini negativi prevalgono
sui positivi

$$P < 0$$



i termini positivi e negativi si
compensano

$$P \approx 0$$

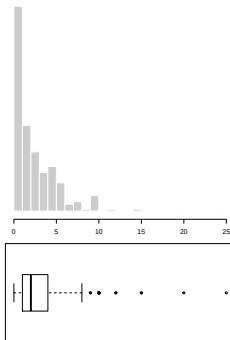


i termini positivi prevalgono
sui negativi

$$P > 0$$

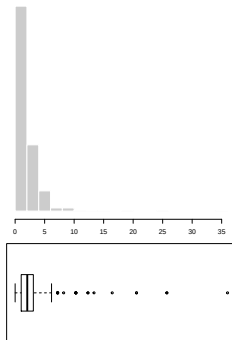
Esempi

Cinema



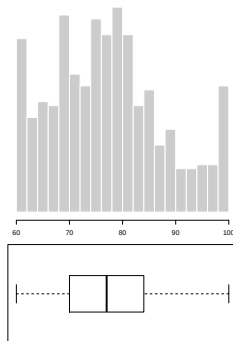
Bowley: 0.33
Pearson: 2.2

Libri



Bowley: 0
Pearson: 5.1

Voto matura

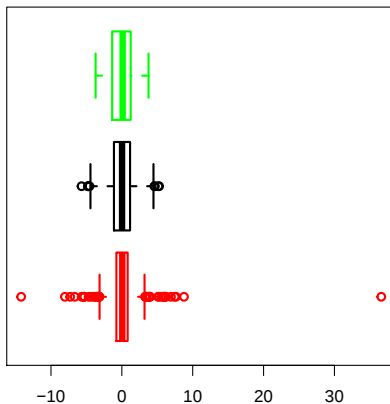
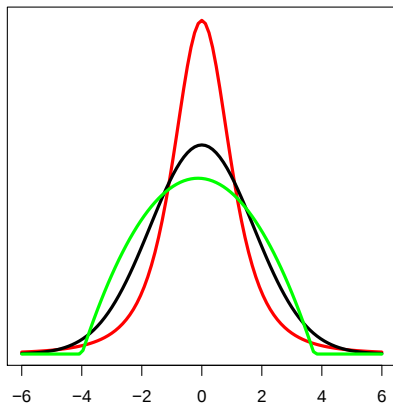


Bowley: 0
Pearson: 0.35

Forma della distribuzione: curtosi

È il “peso” delle code e l’“appuntamento” della distribuzione, si dice

- ▶ platicurtica: code leggere (verde)
- ▶ leptocurtica: code pesanti (rosso)



Forma della distribuzione: curtosi (cont)

Il principale indice usato è l'**indice di curtosi standardizzato**, introdotto da K. Pearson nel 1894 e definito come

$$\frac{1}{N\sigma^4} \sum_{i=1}^N (y_i - \bar{y})^4 = \frac{\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^4}{\sigma^4}.$$

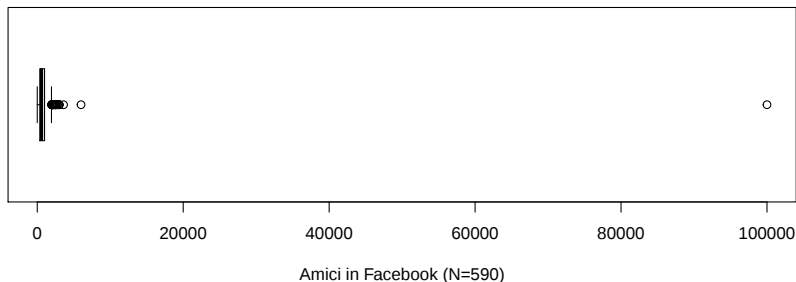


Questo indice può essere visto come un rapporto tra due indici di variabilità. L'indice a numeratore (la media delle potenze quarte degli scarti dalla media aritmetica) è scelto in maniera tale da essere più sensibile alla presenza di code pesanti dell'indice a denominatore (la potenza quarta dello scarto quadratico medio).

Forma della distribuzione: osservazioni anomale (estreme)

Una osservazione potenzialmente anomala (**outlier**) è definita come una osservazione che appare “estrema” rispetto al resto dei dati.

In un boxplot, va oltre i baffi.



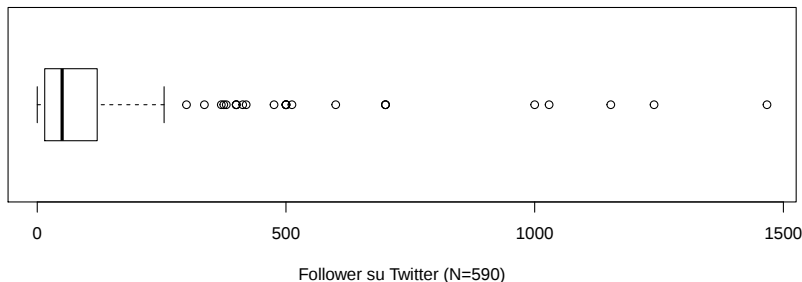
I valori anomali sono rilevanti per

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

Forma della distribuzione: osservazioni anomale (estreme)

Una osservazione potenzialmente anomala (**outlier**) è definita come una osservazione che appare “estrema” rispetto al resto dei dati.

In un boxplot, va oltre i baffi.



I valori anomali sono rilevanti per

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

Osservazioni estreme: robustezza degli indici

Abbiamo già notato che la media è sensibile alla presenza di valori anomali, mentre la mediana è più robusta.

	Tutte oss.	Senza $y_{(N)} = 100000$
Media	944.66	754.17
Mediana	590	589.5

Osservazioni estreme: robustezza degli indici

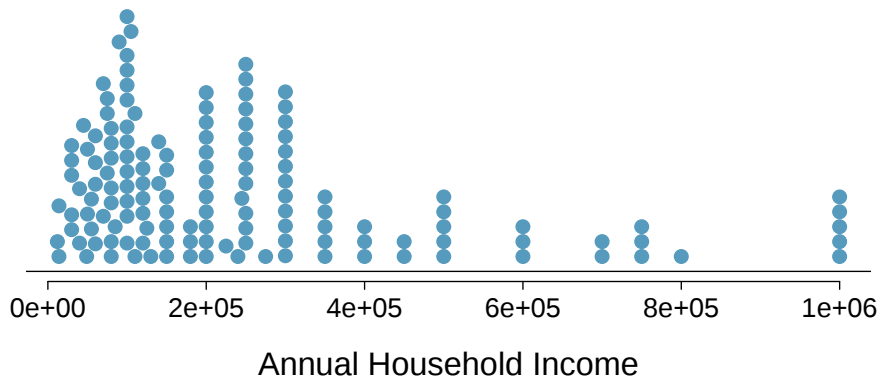
Abbiamo già notato che la media è sensibile alla presenza di valori anomali, mentre la mediana è più robusta.

Un'analoga considerazione vale per varianza e scarto interquartile (il secondo è più robusto).

	Tutte oss.	Senza $y_{(N)} = 100000$
Media	944.66	754.17
Mediana	590	589.5
dev. st.	4392.12	621.29
IQR	650	650

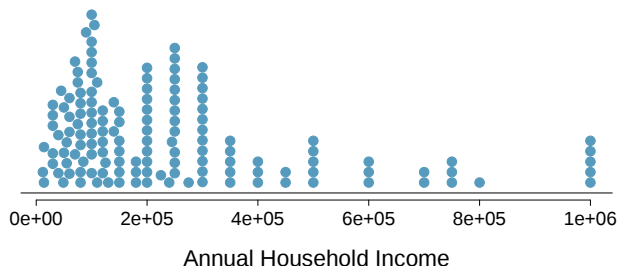
Osservazioni estreme: esempio

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? E come cambierebbe se l'osservazione più piccola fosse spostata a \$10 milioni?

Osservazioni estreme: esempio (cont)



scenario	robusto		non robusto	
	m	IQR	$\bar{y}$	σ
dati originali	190K	200K	245K	226K
sposta max a \$10 milioni	190K	200K	309K	853K
sposta min a \$10 milioni	200K	200K	316K	854K

Mediana e IQR sono più “stabili” della media e della deviazione standard. Si dice che sono più “robuste”.

Indice

Misure di posizione

Variabilità

Media e varianza per dati in gruppi

Forma di una distribuzione

Proprietà delle misure di posizione e scala

Indice

Misure di posizione

Variabilità

Media e varianza per dati in gruppi

Forma di una distribuzione

Proprietà delle misure di posizione e scala

Trasformazioni

Proprietà della mediana

Proprietà della media

Proprietà della varianza

Trasformazioni

Trasformare una variabile Y con osservazioni $y_1, \dots, y_N$ secondo una funzione $g(\cdot)$ significa trasformare ogni modalità secondo tale funzione, producendo una nuova variabile $T = g(Y)$ con osservazioni

$$t_1, \dots, t_i = g(y_i), \dots, t_N$$

Incontriamo trasformazioni ad esempio quando

- ▶ cambiamo unità di misura a una variabile (da gradi Fahrenheit a Celsius, o da metri a centimetri)
- ▶ applichiamo una funzione per migliorare la leggibilità di una distribuzione (trasformazione logaritmica)

Diventa allora rilevante capire cosa succede alle misure di posizione e variabilità di T , ossia cosa possiamo dire di esse in rapporto alle corrispondenti misure di Y .

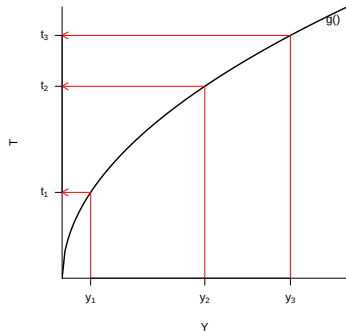
Mediana e trasformazioni monotone

Se $t(\cdot)$ è monotona non decrescente, vale a dire che si mantiene l'ordine

$$y_i \leq y_j \Rightarrow g(y_i) \leq g(y_j)$$

La mediana dei valori $t_1, t_2, \dots, t_N$, indicata con m_T , è la trasformazione tramite $g(\cdot)$ della mediana dei valori originali, ovvero

$$m_T = g(m_Y).$$



Nel disegno, t_2 è la mediana di T e coincide con $g(y_2)$.

Mediana e trasformazioni monotone

Se $t(\cdot)$ è monotona non decrescente, vale a dire che si mantiene l'ordine

$$y_i \leq y_j \Rightarrow g(y_i) \leq g(y_j)$$

La mediana dei valori $t_1, t_2, \dots, t_N$, indicata con m_T , è la trasformazione tramite $g(\cdot)$ della mediana dei valori originali, ovvero

$$m_T = g(m_Y).$$

Se $t(\cdot)$ è monotona non crescente, vale a dire che si inverte l'ordine

$$y_i \leq y_j \Rightarrow g(y_i) \geq g(y_j)$$

vale la stessa conclusione: la mediana m_T è la trasformazione tramite $g(\cdot)$ della mediana dei valori originali, ovvero

$$m_T = g(m_Y).$$

Trasformazioni lineari

Una trasformazione monotona particolarmente importante è la trasformazione lineare, ovvero la trasformazione del tipo

$$g(x) = a + bx.$$

sono esempi

- ▶ Temperatura in gradi Fahrenheit e Celsius: $F = 1.8C + 32$;
- ▶ importi monetari in dollari vengono trasformati in euro (con cambio fisso): $EUR = 0.92\$$.

Una particolare trasformazione lineare è la **standardizzazione**, siano $(y_1, \dots, y_N)$ dati grezzi, con media $\bar{y}$ e deviazione standard σ_Y si definiscono dati standardizzati i trasformati

$$z_i = \frac{y_i - \bar{y}}{\sigma_Y}.$$

Media e trasformazioni lineari

Siano $y_1, y_2, \dots, y_N$ le osservazioni disponibili per Y e $\bar{y}$ la loro media.

Sia $T = g(Y) = a + bY$ e $t_1 = g(y_1), t_2 = g(y_2), \dots, t_N = g(y_N)$ i dati trasformati.

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

$$\bar{t} = g(\bar{y}) = a + b\bar{y}.$$

Media e trasformazioni lineari (DIM)

La media di un trasformato lineare è il trasformato lineare della media.
Dimostrazione.

$$\begin{aligned}
 \bar{t} &= \frac{t_1 + t_2 + \cdots + t_N}{N} = \\
 &= \frac{(a + by_1) + (a + by_2) + \cdots + (a + by_N)}{N} = \\
 &= \frac{\overbrace{a + \cdots + a}^{N \text{ volte}}}{N} + b \frac{y_1 + y_2 + \cdots + y_N}{N} = \\
 &= a + b\bar{y}.
 \end{aligned}$$

Non Equivarianza della media in generale

La media di una trasformazione $T = g(Y)$ non-lineare dei dati non è, in genere, uguale alla stessa trasformazione applicata alla media, cioè, in generale

$$\bar{t} = \frac{1}{N} \sum_{i=1}^N g(y_i) \neq g\left(\frac{1}{N} \sum_{i=1}^N y_i\right) = g(\bar{y})$$

Ad esempio, se $g(Y) = Y^2$, in generale,

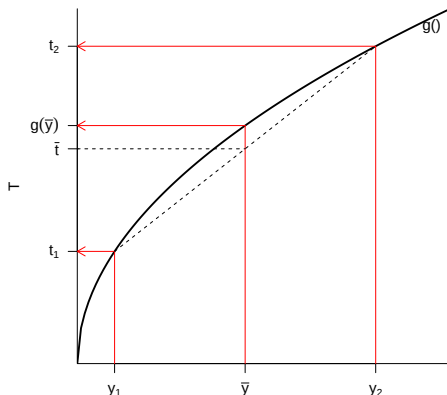
$$\frac{1}{N} \sum_{i=1}^n y_i^2 \neq \left(\frac{1}{N} \sum_{i=1}^N y_i\right)^2$$

ovvero la media dei quadrati dei dati è diversa dal quadrato della media.

Diseguaglianza di Jensen

Se $g(\cdot)$ è concava la retta che congiunge due punti del suo grafico sta sotto il grafico stesso, quindi

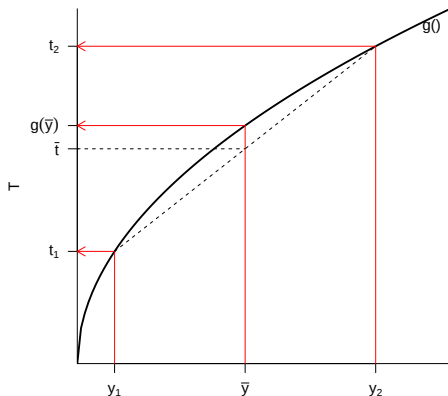
$$g\left(\frac{1}{2}y_1 + \frac{1}{2}y_2\right) > \frac{1}{2}g(y_1) + \frac{1}{2}g(y_2)$$



Diseguaglianza di Jensen

Se $g(\cdot)$ è concava la retta che congiunge due punti del suo grafico sta sotto il grafico stesso, quindi

$$g(\bar{y}) = g\left(\frac{1}{2}y_1 + \frac{1}{2}y_2\right) > \frac{1}{2}g(y_1) + \frac{1}{2}g(y_2) = \bar{t}$$



Diseguaglianza di Jensen

Se $g(\cdot)$ è concava la retta che congiunge due punti del suo grafico sta sotto il grafico stesso, quindi

$$g(\bar{y}) = g\left(\frac{1}{2}y_1 + \frac{1}{2}y_2\right) > \frac{1}{2}g(y_1) + \frac{1}{2}g(y_2) = \bar{t}$$



Se $g(\cdot)$ è convessa la retta che congiunge due punti del suo grafico sta sopra il grafico stesso, quindi

$$g(\bar{y}) = g\left(\frac{1}{2}y_1 + \frac{1}{2}y_2\right) < \frac{1}{2}g(y_1) + \frac{1}{2}g(y_2) = \bar{t}$$

Diseguaglianza di Jensen

Se $g(\cdot)$ è concava la retta che congiunge due punti del suo grafico sta sotto il grafico stesso, quindi

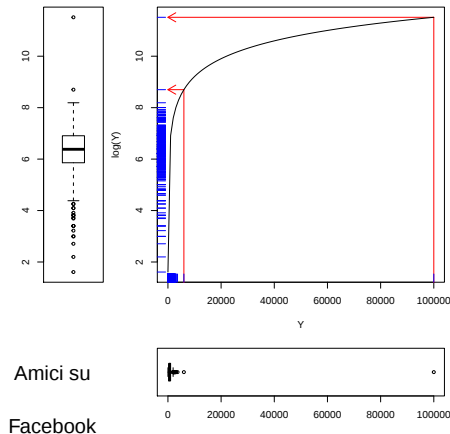
$$g(\bar{y}) = g\left(\frac{1}{2}y_1 + \frac{1}{2}y_2\right) > \frac{1}{2}g(y_1) + \frac{1}{2}g(y_2) = \bar{t}$$



N.B.: le diseguaglianze sono strette se la funzione è strettamente concava/convessa.

Trasformazione logaritmica: Amici di Facebook

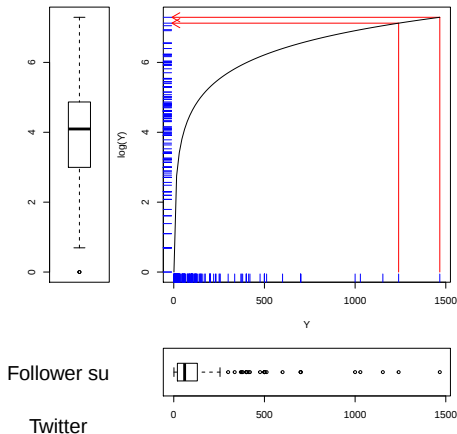
Quando i dati sono fortemente asimmetrici, una trasformazione come il **logaritmo** può essere utile a migliorare la visualizzazione.



Schiaccia le differenze, nel bene e nel male.

Trasformazione logaritmica: Amici di Facebook

Quando i dati sono fortemente asimmetrici, una trasformazione come il **logaritmo** può essere utile a migliorare la visualizzazione.



Schiaccia le differenze, nel bene e nel male.

Pro e contro

- ▶ Outliers tendono ad avvicinarsi:

# amici di Facebook	11	589	3025	...
$\log(\# \text{ Amici di Facebook})$	2.40	6.38	8.01	...

- ▶ 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.

Trasformazione logaritmica media e mediana

Amici di Facebook

$$(y_1, \dots, y_N)$$

si ha

$$\bar{y} = 946.47$$

$$m_Y = 591$$

e si tenga presente che

Log(amici di Facebook)

$$t_i = \log(y_i), \quad i = 1, \dots, N.$$

si ha

$$\bar{t} = 6.32$$

$$m_T = 6.38$$



$$\bar{t} \neq \log(\bar{y}) = 6.85$$

$$m_T = \log(m_Y) = 6.38$$

Varianza e traslazioni

Siano $y_1, y_2, \dots, y_N$ le osservazioni disponibili per Y e σ_Y^2 la loro varianza

Sia $T = g(Y) = a + Y$ e $t_1 = g(y_1), t_2 = g(y_2), \dots, t_N = g(y_N)$ i dati trasformati.

La varianza dei valori $t_1, t_2, \dots, t_N$, indicata con σ_T^2 , è uguale alla varianza dei valori originali.

$$\sigma_T^2 = \sigma_Y^2.$$



La dimostrazione è lasciata per esercizio (chiedetevi che effetto produce una traslazione dei dati sulla distribuzione dei dati).

Omogeneità

Siano $y_1, y_2, \dots, y_N$ le osservazioni disponibili per il carattere Y e sia σ_Y^2 la loro varianza.

Sia $T = g(Y) = bY$, con b costante qualunque e $t_1 = g(y_1), t_2 = g(y_2), \dots, t_N = g(y_N)$ i dati trasformati.

La varianza dei valori $t_1, t_2, \dots, t_N$, indicata con σ_T^2 , è uguale a

$$\sigma_T^2 = b^2 \sigma_Y^2.$$



La dimostrazione è lasciata per esercizio (chiedetevi che effetto produce la moltiplicazione dei dati per una costante sulla distribuzione dei dati).

Una particolare trasformazione: la Standardizzazione

Siano $y_1, y_2, \dots, y_N$ le osservazioni disponibili per il carattere Y , siano $\bar{y}$ la loro media e σ_Y^2 la loro varianza.

Una particolare trasformazione: la Standardizzazione

Siano $y_1, y_2, \dots, y_N$ le osservazioni disponibili per il carattere Y , siano $\bar{y}$ la loro media e σ_Y^2 la loro varianza.

Consideriamo la trasformazione

$$Z = \frac{Y - \bar{y}}{\sigma_Y}$$

e siano $z_1, z_2, \dots, z_N$ i corrispondenti valori standardizzati,

$$z_i = \frac{y_i - \bar{y}}{\sigma_y}$$

Una particolare trasformazione: la Standardizzazione

Siano $y_1, y_2, \dots, y_N$ le osservazioni disponibili per il carattere Y , siano $\bar{y}$ la loro media e σ_Y^2 la loro varianza.

Consideriamo la trasformazione

$$Z = \frac{Y - \bar{y}}{\sigma_Y}$$

e siano $z_1, z_2, \dots, z_N$ i corrispondenti valori standardizzati,

$$z_i = \frac{y_i - \bar{y}}{\sigma_Y}$$

Si ha allora

$$\bar{z} = -\frac{\bar{y}}{\sigma_Y} + \frac{1}{\sigma_Y}\bar{y} = 0$$

e

$$\sigma_Z^2 = \left(\frac{1}{\sigma_Y}\right)^2 \sigma_Y^2 = 1.$$

Quindi, i dati standardizzati hanno media nulla e varianza unitaria.

Indice

Misure di posizione

Variabilità

Media e varianza per dati in gruppi

Forma di una distribuzione

Proprietà delle misure di posizione e scala

Trasformazioni

Proprietà della mediana

Proprietà della media

Proprietà della varianza

Proprietà dei minimi scarti assoluti

La mediana è il valore che minimizza la somma dei valori assoluti degli scarti di un insieme di dati da una costante, ovvero

$$\sum_{i=1}^N |y_i - m| \leq \sum_{i=1}^N |y_i - a|.$$

qualunque sia $a \in \mathbb{R}$.

Indice

Misure di posizione

Variabilità

Media e varianza per dati in gruppi

Forma di una distribuzione

Proprietà delle misure di posizione e scala

Trasformazioni

Proprietà della mediana

Proprietà della media

Proprietà della varianza

Proprietà di Cauchy

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

$$y_{(1)} \leq \bar{y} \leq y_{(N)}.$$



Infatti, ad esempio, per quanto riguarda la prima disuguglianza

$$y_{(1)} = \frac{\overbrace{y_{(1)} + \cdots + y_{(1)}}^{N \text{ volte}}}{N} \leq \frac{y_1 + y_2 + \cdots + y_N}{N} = \bar{y}$$

Proprietà di baricentro

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

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



Infatti, con i dati grezzi, si ha

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

Proprietà dei minimi quadrati

La media aritmetica minimizza la somma (e quindi la media) dei quadrati degli scarti delle singole osservazioni da una costante a .

Sia a un numero qualsiasi. Allora

$$\sum_{i=1}^N (y_i - a)^2 = \sum_{i=1}^N (y_i - \bar{y})^2 + N(\bar{y} - a)^2$$

Dimostrazione.

NB: tutte le sommatorie vanno da 1 a N

$$\begin{aligned} \sum (y_i - a)^2 &= \sum (y_i - a + \bar{y} - \bar{y})^2 = \\ &= \sum [(y_i - \bar{y}) + (\bar{y} - a)]^2 = \\ &= \sum \left[(y_i - \bar{y})^2 + (\bar{y} - a)^2 + 2(\bar{y} - a)(y_i - \bar{y}) \right] = \\ &= \sum (y_i - \bar{y})^2 + \sum (\bar{y} - a)^2 + 2(\bar{y} - a) \sum (y_i - \bar{y}) = \\ &= \sum (y_i - \bar{y})^2 + N(\bar{y} - a)^2 + 2(\bar{y} - a) \times 0. \end{aligned}$$

Quindi, quando $a = \bar{y}$, la quantità $\sum (y_i - a)^2$ è minima.

Indice

Misure di posizione

Variabilità

Media e varianza per dati in gruppi

Forma di una distribuzione

Proprietà delle misure di posizione e scala

Trasformazioni

Proprietà della mediana

Proprietà della media

Proprietà della varianza

Scomposizione della varianza: dimostrazione

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

Connessione con le distanze tra coppie di osservazioni

È interessante notare che la varianza anche essere interpretata come una media delle differenze al quadrato tra tutte le possibili coppie di dati.

Infatti

$$\sigma^2 = \frac{1}{2N^2} \sum_{i=1}^N \sum_{j=1}^N (y_i - y_j)^2$$

Dimostrazione.

$$\begin{aligned} & \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (y_i - y_j)^2 = \\ &= \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N [(y_i - \bar{y}) - (y_j - \bar{y})]^2 = \\ &= \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (y_i - \bar{y})^2 + \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (y_j - \bar{y})^2 - \\ & \quad - \frac{2}{N^2} \sum_{i=1}^N \sum_{j=1}^N (y_i - \bar{y})(y_j - \bar{y}) = \\ &= \frac{2N}{N^2} \sum_{i=1}^N (y_i - \bar{y})^2 - 2 \left[\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y}) \right]^2 = 2\sigma^2. \end{aligned}$$