

Statistica Sociale

Quartili e percentile e differenza interquartile

La mediana è una grandezza che appartiene a un insieme di misure di posizione chiamate **percentili**

Il **p-esimo percentile** è il valore nella distribuzione al di sotto del quale ricade il p% delle osservazioni e al di sopra del quale ricade il (100 - p)% delle osservazioni

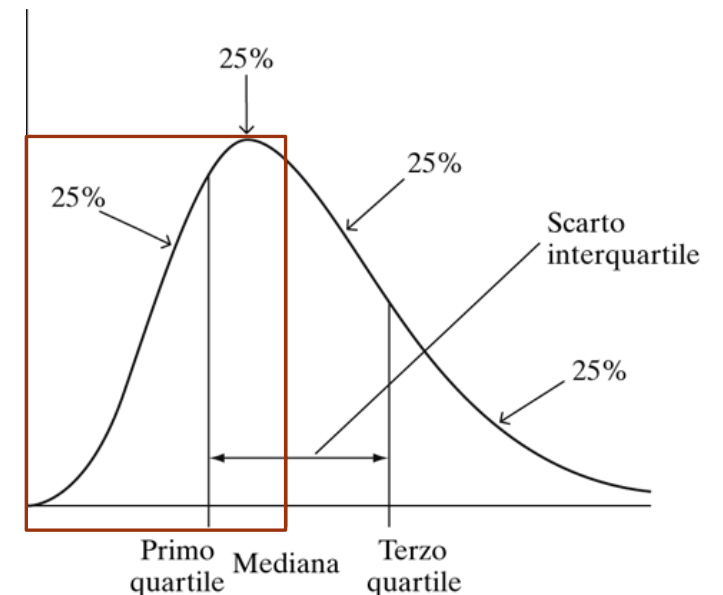
- La **mediana** è il **50-esimo percentile** (o **secondo quartile**)
- Altri due percentili di uso comune sono il **25-esimo percentile** (o **primo quartile**) ed il **75-esimo percentile** (o **terzo quartile**)

I **quartili** (primo quartile, mediana e terzo quartile) dividono la distribuzione in 4 parti ognuna delle quali contiene un quarto delle osservazioni

Differenza(scarto) interquartile è la differenza tra il terzo ed il primo quartile

$$IQR = Q_3 - Q_1$$

- Rappresenta il campo di variazione per il 50% delle unità centrali intorno alla mediana, non sensibile alla presenza di valori anomali



Differenza interquartile e range

Il primo quartile può essere calcolato come la mediana del primo 50% dei dati ed il terzo come la mediana del secondo 50% dei dati

- Q_3 in questo caso è la media tra il 75-esimo ed il 76-esimo valore osservato, mentre Q_1 è la media tra il 25-esimo ed il 26-esimo valore osservato.
- Per il primo collettivo quindi: $Q_3 = 4$ e $Q_1 = 3$
 - $IQR_1 = Q_3 - Q_1 = 4 - 3 = 1$
 - $R_1 = 7 - 1 = 6$
- Per il secondo collettivo invece: $Q_3 = 6$ e $Q_1 = 2$
 - $IQR_2 = Q_3 - Q_1 = 6 - 2 = 4$
 - $R_2 = 7 - 1 = 6$

X	Primo collettivo		Secondo collettivo	
	n_j	N_j	n_j	N_j
1	2	2	10	10
2	5	7	25	35
3	20	27	10	45
4	50	77	10	55
5	15	92	10	65
6	5	97	15	80
7	3	100	20	100
Totale	100		100	

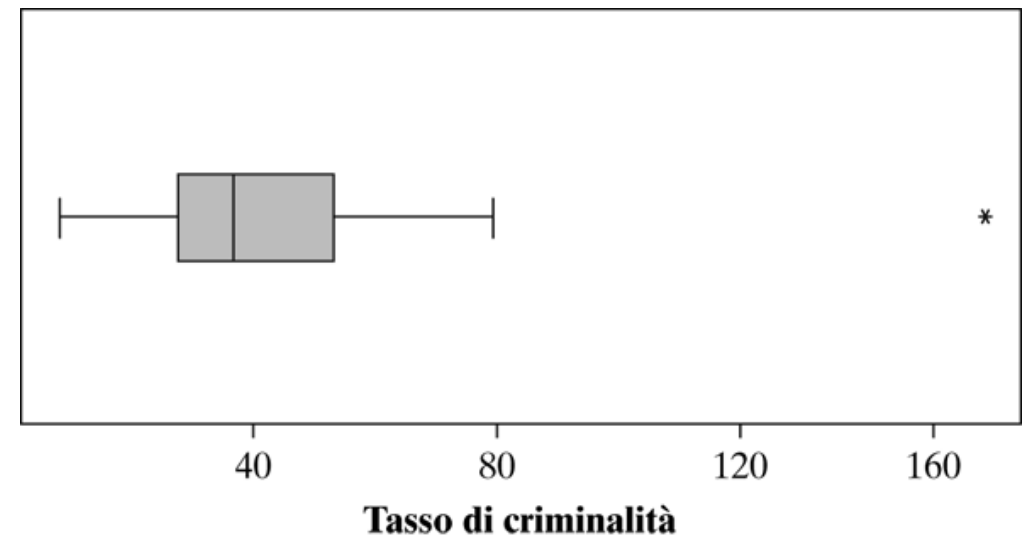
La differenza interquartile non è influenzata dagli outlier

Box plot

- La mediana, i quartili, il massimo e il minimo sono cinque misure di posizione spesso impiegate congiuntamente per descrivere la centralità e la variabilità di una distribuzione
- I cinque numeri forniscono una semplice descrizione dei dati e li chiameremo sintesi-a-cinque-numeri
- Essi sono anche gli elementi base di una rappresentazione grafica chiamata **box plot**
 - Il box (scatola) contiene il 50% centrale della distribuzione, dal primo al terzo quartile
- La mediana è rappresentata da una linea che attraversa il box
- Le linee che si estendono a partire dalla scatola sono chiamate **whiskers** e vanno fino al massimo e fino al minimo a meno che nella distribuzione siano presenti osservazioni **outlier**

- Con un software statistico è stato ottenuto il seguente report riferito alla distribuzione dei tassi di criminalità USA:

100%	Max	79.0
75%	Q3	51.0
50%	Med	36.
25%	Q1	27.0
0%	Min	8.0



Box plot con valori anomali

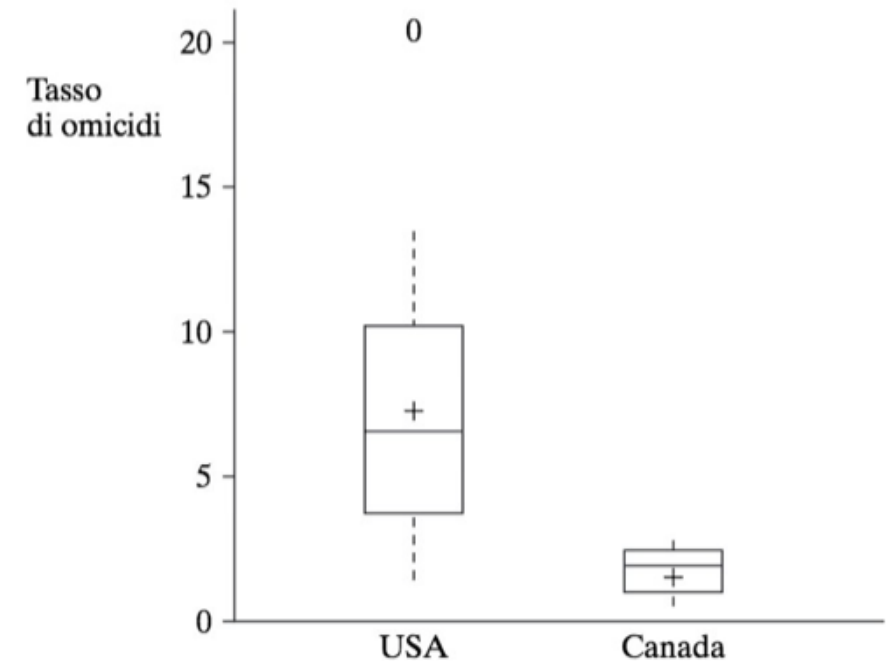
Sono considerati anomali (**outlier**) quei valori più piccoli di

$$Q_1 - 1,5 * IQR$$

o più grandi di

$$Q_3 + 1,5 * IQR$$

- Se nella distribuzione non sono presenti osservazioni outlier, i whisker (*baffi*) del box plot si estendono fino alle osservazioni massima e minima
- Se ci sono outlier, i whisker si estendono fino all'osservazione che è compresa entro $1.5 \times (IQR)$ oltre i quartili; gli outlier sono indicati separatamente nel grafico
- È sempre meglio considerare un'osservazione definita outlier come un potenziale outlier



Età	Numerosità
21	5
23	1
27	1
31	1
32	1
44	1

21,21,21,21,21, 23,27,31,32,44

$$Me = \frac{x_{(5)} + x_{(6)}}{2} = \frac{21 + 23}{2} = 22$$

$$Q_1 = x_{(3)} = 21$$

$$Q_3 = x_{(8)} = 31$$

$$IQR = Q_3 - Q_1 = 31 - 21 = 10$$

$$Baffo\ inf = 21 - 1,5 * 10$$

$$Baffo\ sup = 31 + 1,5 * 10$$

Età	Numerosità
21	3
22	2
24	2
34	1

21,21,21,22, 22,24,24,34

$$Me = \frac{x_{(4)} + x_{(5)}}{2} = \frac{22 + 22}{2} = 22$$

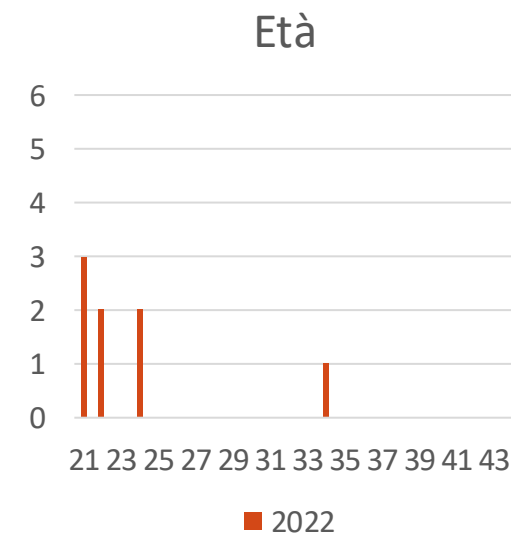
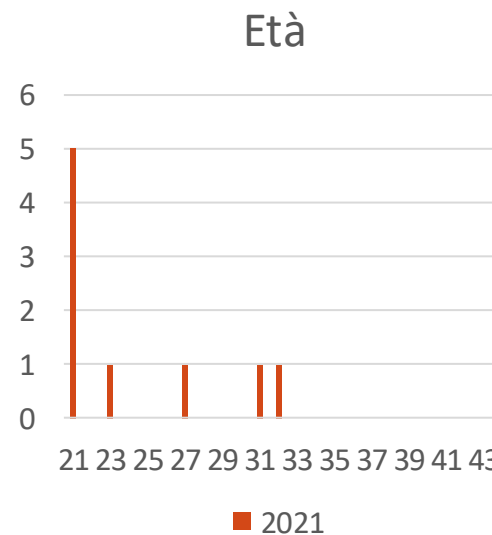
$$Q_1 = \frac{x_{(2)} + x_{(3)}}{2} = \frac{21 + 21}{2} = 21$$

$$Q_3 = \frac{x_{(6)} + x_{(7)}}{2} = \frac{24 + 24}{2} = 24$$

$$IQR = Q_3 - Q_1 = 24 - 21 = 3$$

$$Baffo\ inf = 21 - 1,5 * 3 = 16,5$$

$$Baffo\ sup = 24 + 1,5 * 3 = 28,5$$



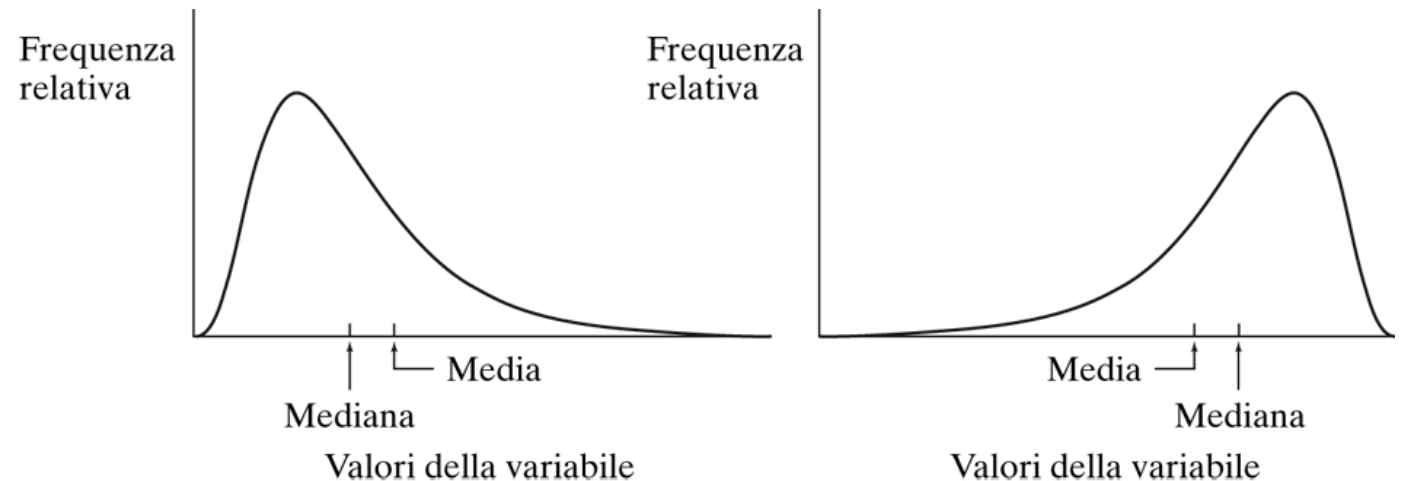
Media vs mediana

Da un'indagine campionaria ISTAT sul reddito delle famiglie italiane nel 2022 è emerso un reddito medio di 2525 euro ed un reddito mediano di 2.197 euro

- Il reddito medio si può interpretare come quel valore che avrebbe dovuto percepire ogni famiglia se l'ammontare totale fosse stato ripartito equamente tra tutte le famiglie
- Il reddito mediano ci dice che il 50% delle famiglie italiane nel 2022 ha percepito un reddito minore o uguale a 2197 euro, mentre l'altro 50% superiore al valore mediano

La mediana è più robusta della media ai valori estremi/code

Quando una distribuzione è simmetrica media e mediana coincidono



Indici sintetici: moda

Può essere calcolata per tutti i tipi di variabili (anche per i qualitativi nominali)

La **moda** è definita come la modalità a cui è associata la frequenza osservata più elevata

Quando il numero di modalità possibili è alto la moda è probabilmente “troppo sintetica”

Se il carattere è suddiviso in classi parliamo di **classe modale**

- Maggiore approssimazione
- Se le classi sono di ampiezza diversa vanno confrontate le densità (vd istogramma) e non le frequenze!

Esistono distribuzioni **unimodali** o **bimodali**

- trova un esempio grafico di distribuzione unimodale e distribuzione bimodale nelle slides delle lezioni precedenti

La moda non è sensibile agli outliers

- Calcolare la moda

del carattere “colore degli occhi” sul campione di 8 unità

Unità	Sesso	Età	Statura	Colore occhi
1	F	24	163	Marrone
2	F	21	165	Azzurri
3	M	34	185	Azzurri
4	F	22	164	Marroni
5	F	21	167	Marroni
6	F	22	175	Verdi
7	M	24	178	Verdi
8	F	21	155	Marroni

Esercizi

Calcola la media, la mediana e la moda per questi valori: 10, 5, 20, 5, 15

- Come cambiano i tre indici se aggiungo 50 (20+50 = 70) al valore più grande osservato?

$$\bar{x} = 11$$

$$\text{Me} = 10$$

$$\text{Mo} = 5$$

$$\bar{x} = 21$$

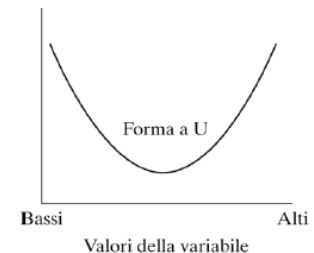
$$\text{Me} = 10$$

$$\text{Mo} = 5$$

Completa la frase:

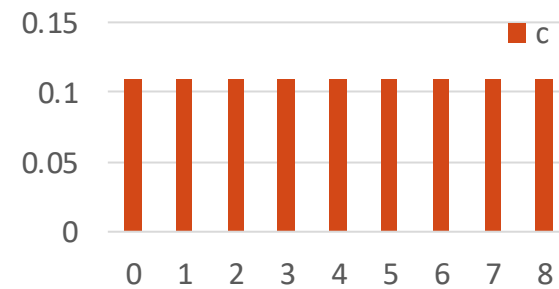
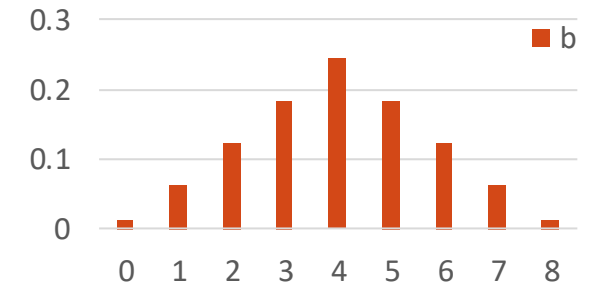
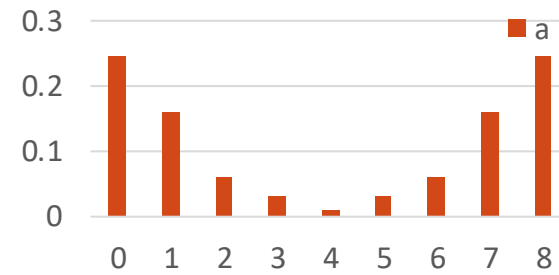
In una distribuzione bimodale...

- A: Media, moda e mediana coincidono se la distribuzione è simmetrica
- B: Media e mediana coincidono se la distribuzione è simmetrica
- C: Media, moda e mediana non coincidono



Rappresentatività media

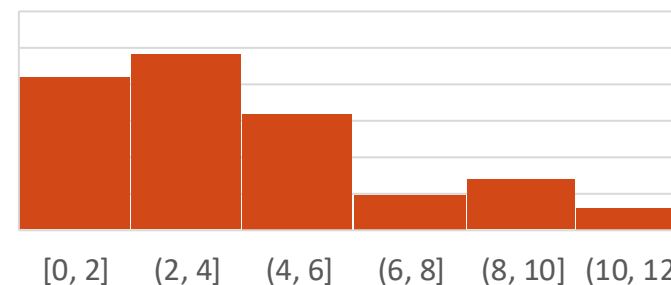
x	$Freq\ rel^a$	$Freq\ rel^b$	$Freq\ rel^c$
0	0,245	0,012	0,11
1	0,16	0,061	0,11
2	0,06	0,122	0,11
3	0,03	0,183	0,11
4	0,01	0,244	0,11
5	0,03	0,183	0,11
6	0,06	0,122	0,11
7	0,16	0,061	0,11
8	0,245	0,012	0,11
1			



$$\bar{x}^a = 4$$

$$\bar{x}^b = 0 * 0,012 + 1 * 0,061 + \dots + 7 * 0,061 + 8 * 0,012 = 4$$

$$\bar{x}^c = \dots = 4$$

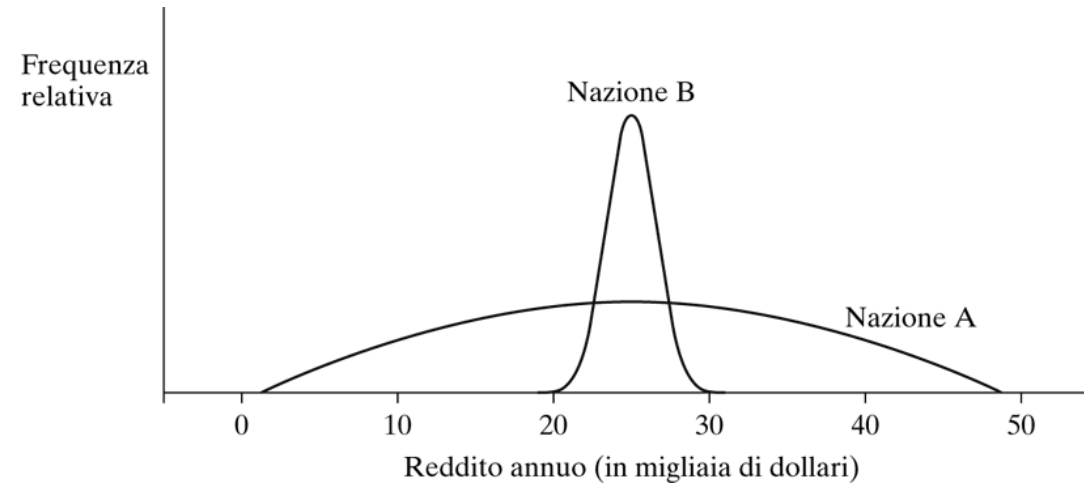


$$\bar{x}^d = 4$$

Gli indici di posizione danno informazioni sulla tendenza centrale della distribuzione ma non sulla **dispersione (variabilità)**

La variabilità di un fenomeno

Gli indici di posizione danno informazioni sulla tendenza centrale della distribuzione ma non sulla **dispersione (variabilità)**



Attraverso una **misura sintetica**, si vuole descrivere la **tendenza delle unità del campione ad assumere modalità diverse**

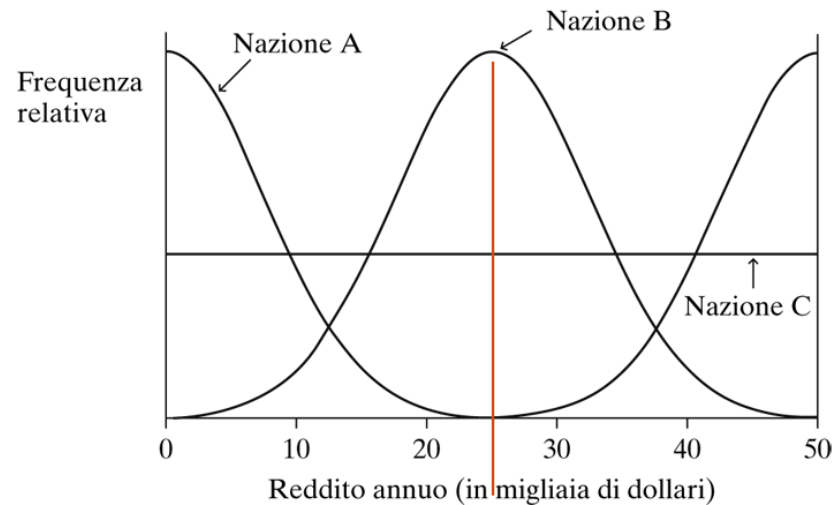
- avrà valore minimo quando tutte le unità della distribuzione presentano uguale modalità
- avrà valore massimo quando le unità assumono modalità diverse tra loro

N.B.: varie misure

1. Campo di variazione (range)

Il **campo di variazione** si calcola come la **differenza** tra il valore **massimo** e **minimo** osservati

Non è sensibile ad altre caratteristiche della variabilità dei dati



In termini di distanza dalla media, quale delle tre distribuzioni ha maggiore variabilità?

Media=\$25.000

Range=\$50.000. Redditi nazione A più distanti dalla loro media: **Nazione A maggiore variabilità**

Variabilità rispetto alla media

Si sceglie un valore caratteristico della distribuzione rispetto a cui calcolare la “diversità” o dispersione delle modalità, uno dei più utilizzati è la media aritmetica

La “diversità” si misura calcolando le differenze tra le unità ed il valore caratteristico scelto. Ad esempio potremmo utilizzare gli **scarti** (o deviazione) **dalla media** ($x_i - \bar{x}$), ma ...

... la media è il baricentro delle osservazioni: la somma degli scarti dalla media delle osservazioni più piccole è uguale alla somma delle distanze di quelle più grandi

La somma degli scarti dalla media è **sempre pari a 0**

$$x_1 = 3, x_2 = 2, x_3 = 2, x_4 = 1, x_5 = 2$$

$$n = 5$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{n} = \frac{3 + 2 + 2 + 1 + 2}{5} = 2$$

Calcoliamo gli scarti dalla media

$$x_1 - \bar{x} = 3 - 2 = 1; x_2 - \bar{x} = 2 - 2 = 0$$

$$x_3 - \bar{x} = 2 - 2 = 0; x_4 - \bar{x} = 1 - 2 = -1$$

$$x_5 - \bar{x} = 2 - 2 = 0$$

$$\sum_{i=1}^5 (x_i - \bar{x}) = (x_1 - \bar{x}) + \dots + (x_5 - \bar{x}) = 1 + 0 + 0 - 1 + 0 = 0$$

Varianza e deviazione standard

Considera la **somma degli scarti al quadrato** dalla media aritmetica e la divide per la numerosità campionaria ($n-1$)

La **varianza** di un insieme di n valori osservati $x_1, x_2, \dots, x_i, \dots, x_n$ di una variabile x con media aritmetica \bar{x} è

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2]$$

È approx una media delle deviazioni al quadrato

Mentre la **deviazione standard** è la radice quadrata della varianza

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

La deviazione standard è espressa nella stessa unità di misura della media

- La ragione per cui si utilizza $(n - 1)$ piuttosto che n nel denominatore di s (e di s^2) riguarda l'inferenza per i parametri della popolazione
- Quando abbiamo dati riferiti ad un'intera popolazione, sostituiamo $(n - 1)$ con l'effettiva ampiezza campionaria: la varianza della popolazione è, allora, esattamente la media delle deviazioni al quadrato

Formula indiretta: $s^2 = M(x^2) -$

Confronto di due campioni

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2]$$

Campione 1: 0, 4, 4, 5, 7, 10

Campo di variazione del campione 1 = 10 - 0 = 10

$$\bar{x}_1 = \frac{0+4+4+5+7+10}{6} = \frac{30}{6} = 5$$

$$s_1^2 = \frac{1}{6-1} [(0-5)^2 + (4-5)^2 + (4-5)^2 + (5-5)^2 + (7-5)^2 + (10-5)^2]$$

$$= \frac{1}{5} [(-5)^2 + (-1)^2 + (-1)^2 + (0)^2 + (2)^2 + (5)^2]$$

$$= \frac{1}{5} [25 + 1 + 1 + 0 + 4 + 25] = \frac{56}{5} = 11,2$$

$$s = \sqrt{s^2} = \sqrt{11,2} = 3,3$$

Campione 2: 0, 0, 1, 9, 10, 10

Campo di variazione del campione 2 = 10-0=10

$$\bar{x}_2 = \frac{0+0+1+9+10+10}{6} = \frac{30}{6} = 5$$

$$s_2^2 = \frac{1}{6-1} [(0-5)^2 + (0-5)^2 + (1-5)^2 + (9-5)^2 + (10-5)^2 + (10-5)^2]$$

$$= \frac{1}{5} [(-5)^2 + (-5)^2 + (-4)^2 + (4)^2 + (5)^2 + (5)^2]$$

$$= \frac{1}{5} [25 + 25 + 16 + 16 + 25 + 25] = \frac{132}{5} = 26,4$$

$$s = \sqrt{s^2} = \sqrt{26,4} = 5,13$$

Attenzione all'ordine delle operazioni: se sommate gli scarti prima di elevarli al quadrato ottenete 0!

Proprietà della varianza/deviazione standard

- la varianza è sempre positiva (somma di differenze al quadrato)
- la varianza è uguale a 0 solo se tutte le osservazioni hanno lo stesso valore

$$x_1 = 10; x_2 = 10; x_3 = 10$$
$$\bar{x} = 10$$

$$s^2 = \frac{1}{2} [(10 - 10)^2 + (10 - 10)^2 + (10 - 10)^2] = 0$$

- più grande è la variabilità intorno alla media maggiore è il valore della varianza
 - le differenze più grandi hanno “più peso” perchè aumentano più che proporzionalmente
- se sui dati viene applicata una trasformazione di scala, anche la deviazione standard viene trasformata
 - Altezza da cm a metri

Interpretazione della deviazione standard

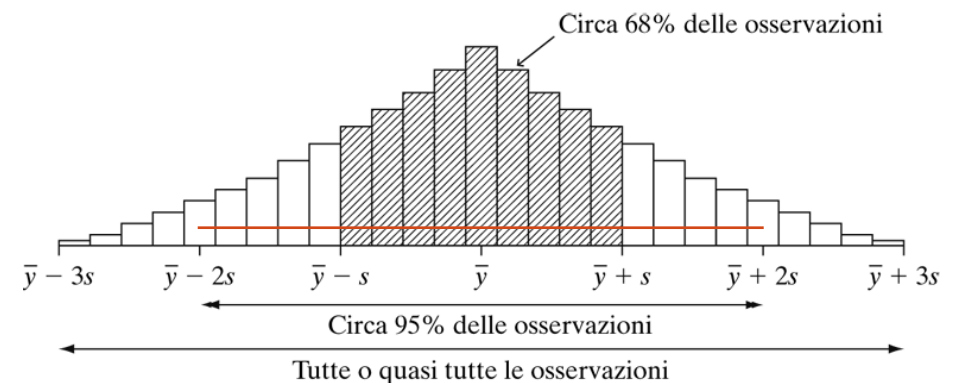
A parità di unità di misura, possiamo confrontare le varianze di due campioni e concludere quale delle due è più variabile

Ma come interpretare il valore della varianza in termini assoluti?

Regola empirica

Se l'istogramma della distribuzione ha una forma approssimativamente campanulare:

1. Circa il 68% delle osservazioni assume valori compresi tra $\bar{x} - s$ e $\bar{x} + s$
2. Circa il 95% delle osservazioni assume valori compresi tra $\bar{x} - 2s$ e $\bar{x} + 2s$
3. Quasi la totalità delle osservazioni assume valori compresi tra $\bar{x} - 3s$ e $\bar{x} + 3s$



Calcolo della varianza

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} [(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2]$$

Unità	Età	$x_i - \bar{x}$	=	$(x_i - \bar{x})^2$
1	22	22-26,3	-4,3	18,49
2	21	21-26,3	-5,3	28,09
3	21	21-26,3	-5,3	28,09
4	21	21-26,3	-5,3	28,09
5	21	21-26,3	-5,3	28,09
6	27	27-26,3	0,7	0,49
7	31	31-26,3	4,7	22,09
8	44	44-26,3	17,7	313,29
9	23	23-26,3	3,3	10,89
10	32	32-26,3	5,7	32,49
				510,1

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{9} [18,49 + 28,09 + \dots + 32,49] = \frac{510,1}{9} = 56,68$$

$$s = \sqrt{s^2} = \sqrt{56,68} = 7,53$$

Unità	Età	$x_i - \bar{x}$	=	$(x_i - \bar{x})^2$
1	24	24-23,625	0,375	0,140625
2	21	21-23,625	-2,625	6,890625
3	34	34-23,625	10,375	107,640625
4	22	22-23,625	-1,625	2,640625
5	21	21-23,625	-2,625	6,890625
6	22	22-23,625	-1,625	2,640625
7	24	24-23,625	0,375	0,140625
8	21	21-23,625	-2,625	6,890625
				133,875

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{7} [0,141 + 6,89 + \dots + 6,89] = \frac{133,875}{7} = 19,125$$

$$s = \sqrt{s^2} = \sqrt{19,125} = 4,37$$

3. Coefficiente di variazione (per confronti)

Utile se si vogliono confrontare variabilità di fenomeni con ordini di grandezza molto diversi (es. altezze di adulti e bambini) o con unità di misura diverse (es peso ed altezza)

Il **coefficiente di variazione** è un indice di variabilità percentuale definito come il rapporto tra la deviazione standard e la media aritmetica (positiva) moltiplicato per 100

$$CV = \frac{S}{\bar{x}} * 100$$

- Confrontiamo le due distribuzioni relative alla quantità di pulviscolo emesso rispetto a due dispositivi anti inquinanti installati sulle ciminiere.

Tipo	Quantità di pulviscolo (g/min)								
A	69	80	44	52	54	54	86	77	66
B	35	62	43	23	30	28	22	40	25

$$\bar{x}_A = 64,67 \text{ g/min} \quad s = 13,65 \text{ g/min}$$

$$\bar{x}_B = 34,22 \text{ g/min} \quad s = 12,02 \text{ g/min}$$

Sembra che tra le due la distribuzione del dispositivo A sia più variabile, ma osservando il coefficiente di variazione...

$$CV_A = \frac{13,65}{64,67} * 100 = 21\%$$

$$CV_B = \frac{12,02}{34,22} * 100 = 35\%$$

Riepilogo

Misure	Definizione	Interpretazione
Centro		
Media	$\bar{y} = \Sigma y_i / n$	Centro di gravità
Mediana	Osservazione centrale del campione ordinato	50-esimo percentile, divide il campione in due parti uguali
Moda	Valore osservato più di frequente	Valore più probabile, valida per tutti i tipi di variabili
Variabilità		
Deviazione standard	$s = \sqrt{\Sigma (y_i - \bar{y})^2 / (n - 1)}$	Regola empirica: se campanulare, 68% entro s da \bar{y} (bar), 95% entro $2s$ da \bar{y}
Campo di variazione	Differenza fra l'osservazione più piccola e l'osservazione più grande	Quanto più è grande maggiore è la variabilità
Scarto interquartile	Differenza fra il terzo quartile (75-esimo percentile) e il primo quartile (25-esimo percentile)	Comprende la metà centrale delle osservazioni

Formula indiretta per calcolo varianza:

$$s^2 = \sum_{i=1}^n x_i^2 / n - [\bar{x}]^2$$