Variabilità, eterogeneità, asimmetria

1

La variabilità (di un carattere quantitativo)

- La variabilità esprime la tendenza di una variabile ad assumere valori diversi.
- Sinonimo: dispersione.
- La media (o altra misura di sintesi) non sempre è sufficiente per descrivere un fenomeno.
- Uguali misure di sintesi possono essere associate a indici di variabilità molto differenti.

2

- Esempio:
- Studenti A e B con 3 esami e voti

  A: 18, 24, 30

  B: 23, 24, 25
- Entrambi gli studenti hanno voto medio uguale 24, ma… la variabilità è molto diversa.

3

Gli indici di variabilità

- Per misurare la variabilità si usano appositi indici.
- Un indice di variabilità (I) deve possedere due requisiti:

1. Deve assumere il suo valore minimo se tutti i valori di una variabile sono tra loro uguali

$$x_1 = x_2 = \cdots = x_n \Longrightarrow \min(I)$$

2. Deve aumentare quando aumenta la diversità dei valori.

4

Gli indici di variabilità principali sono:

1. Varianza
2. Deviazione standard
3. Coefficiente di variazione

5

La varianza

- È il più noto indice di variabilità.
- Per *n* dati semplici, la varianza è data da

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

- La varianza è la media degli scostamenti (scarti, differenze) al quadrato dei valori osservati dalla media aritmetica.

6

- Studente A: $x_1 = 18, x_2 = 24, x_3 = 30$

$$\sigma^2 = \frac{(18-24)^2+(24-24)^2+(30-24)^2}{3}$$
$$= \frac{36+0+36}{3}$$
$$= \frac{72}{3} = 24$$

7

- Studente B: $x_1 = 23, x_2 = 24, x_3 = 25$

$$\sigma^2 = \frac{(23-24)^2+(24-24)^2+(25-24)^2}{3}$$
$$= \frac{1+0+1}{3}$$
$$= \frac{2}{3} = 0{,}667$$

8

- Per una distribuzione di frequenze, la varianza è data da

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

9

x_i	n_i	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_i$
1	2	2,7225	5,445
2	6	0,4225	2,535
3	9	0,1225	1,1025
4	3	1,8225	5,4675
Totale	20		14,55

$$\sigma^2 = \frac{14{,}55}{20} = 0{,}727$$

10

- Per una distribuzione di frequenze in classi, la varianza è calcolata come

$$\sigma^2 \approx \frac{\sum_{i=1}^{k}(c_i - \bar{x})^2 n_i}{n}$$

11

Formula indiretta della varianza

- σ^2 = media dei quadrati – quadrato della media
- Per *n* dati semplici,

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

12

Varianza di una trasformazione lineare

- Data una variabile *X* e considerata una trasformazione lineare

$$Y = a + bX$$

allora la varianza della variabile *Y* è pari a

$$\sigma_Y^2 = b^2 \sigma_X^2$$

13

La deviazione standard

- Indice noto anche come scarto quadratico medio
- È la radice quadrata della varianza

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

- È importante perché considera gli scarti nella stessa unità di misura del fenomeno.

14

Il coefficiente di variazione

- Si calcola quando bisogna confrontare la variabilità di due o più distribuzioni

$$CV = \frac{\sigma}{\bar{x}} 100$$

- Si utilizza se le distribuzioni

1. Presentano diverse unità di misura (es. metri e cm)
2. Presentano la stessa unità di misura ma hanno valori medi molto diversi (es. costo appartamenti in diverse città)

15

- La deviazione standard dipende dall'unità di misura e dall'ordine di grandezza del fenomeno.
- Il CV non dipende.

16

Esempio 1

- Valori in metri (*X*): 10, 15, 20 $\bar{x} = 15$
- Valori in cm (*Y*): 1000, 1500, 2000 $\bar{y} = 1500$

$$\sigma_X = \sqrt{\frac{(10-15)^2+(15-15)^2+(20-15)^2}{3}}$$

$$\sigma_X = \sqrt{\frac{25+0+25}{3}} = 4{,}082 \qquad CV_X = \frac{4{,}082}{15}100 = 27{,}2$$

$$\sigma_Y = \sqrt{\frac{(1000-1500)^2+(1500-1500)^2+(2000-1500)^2}{3}}$$

$$\sigma_Y = \sqrt{\frac{250000+0+250000}{3}} = 408{,}2 \qquad CV_Y = \frac{408{,}2}{1500}100 = 27{,}2$$

17

Esempio 2

- Costo casa Salerno (*X*): 30, 35, 40 $\bar{x} = 35$
- Costo casa Milano (*Y*): 60, 70, 80 $\bar{y} = 70$

$$\sigma_X = \sqrt{\frac{(30-35)^2+(35-35)^2+(40-35)^2}{3}}$$

$$\sigma_X = \sqrt{\frac{25+0+25}{3}} = 4{,}082 \qquad CV_X = \frac{4{,}082}{35}100 = 11{,}66$$

$$\sigma_Y = \sqrt{\frac{(60-70)^2+(70-70)^2+(80-70)^2}{3}}$$

$$\sigma_Y = \sqrt{\frac{100+0+100}{3}} = 8{,}165 \qquad CV_Y = \frac{8{,}165}{70}100 = 11{,}66$$

18

Altri indici di variabilità

- Il range è la differenza tra il valore massimo e il valore minimo.

$$R = x_{MAX} - x_{MIN}$$

- Dati anomali influenzano R.

19

- La differenza interquartile è la differenza tra il terzo e il primo quartile

$$W = Q_3 - Q_1$$

- Può essere interpretato come il range del 50% dei valori centrali.
- Dati anomali non influenzano W.

20

Il box-plot

- Il box plot è un grafico che descrive una distribuzione utilizzando cinque valori:
- 1. valore minimo
- 2. Q_1
- 3. Me
- 4. Q_3
- 5. valore massimo

21

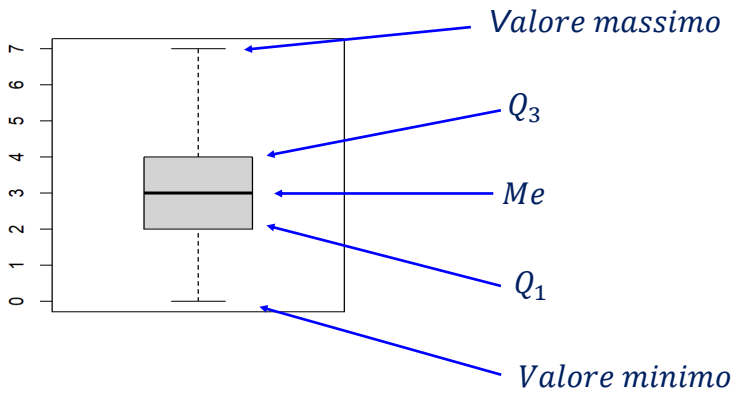
Il box-plot

- Il box-plot è caratterizzato da 3 elementi:

1. Un rettangolo (box) che ha come altezza la differenza interquartile
2. Un segmento che taglia il box in corrispondenza della mediana
3. Due segmenti che partono dal box e che indicano il valore minimo e il valore massimo della distribuzione

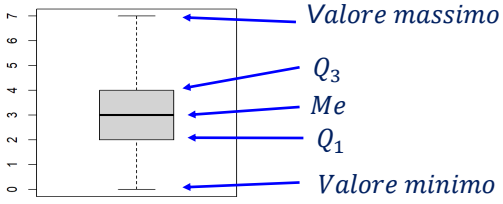
22

Il box-plot



23

Il box-plot



- Il 50% dei valori centrali è compreso tra 2 e 4.
- La differenza interquartile è $W = Q_3 - Q_1 = 4 - 2 = 2$.
- Il valore minimo è 0.
- Il valore massimo è 7.
- Il range è $R = 7$.

24

L'eterogeneità (di un carattere qualitativo)

- Per caratteri qualitativi la variabilità non è misurabile.
- Ciò che possiamo quantificare è il grado di eterogeneità.

25

- Assenza di eterogeneità (massima omogeneità): tutte le unità presentano la stessa modalità.
- Es. 1: colore degli occhi (*n*=6)

  c c c c c c

Colore occhi	*Frequenze*
castani	6
Totale	*6*

26

- Massima eterogeneità: tutte le modalità osservate hanno la stessa frequenza.
- Es. 2: colore degli occhi (*n*=6)

  c c v v a a

Colore occhi	*Frequenze*
castani	2
verdi	2
azzurri	2
Totale	*6*

27

- *3* modalità (*k* = 3)

$$n_1 = n_2 = n_3 = 2$$

- *k* modalità

$$n_1 = n_2 = \cdots = n_k = \frac{n}{k}$$

28

- Eterogeneità: più modalità che non hanno la stessa frequenza.
- Es. 3: colore degli occhi (n=6)

  c c c v v a

Colore occhi	*Frequenze*
castani	3
verdi	2
azzurri	1
Totale	*6*

29

L'indice di eterogeneità di Gini

- L'indice di eterogeneità di Gini è dato da

$$E_1 = 1 - \sum_{i=1}^{k} f_i^2$$

- Versione normalizzata

$$e_1 = \frac{k-1}{k} E_1 = \frac{k}{k-1}\left(1 - \sum_{i=1}^{k} f_i^2\right)$$

- $0 \leq e_1 \leq 1$

30

Interpretazione

- $e_1 = 0$ ➔ Assenza di eterogeneità
- $0 < e_1 < 1$ ➔ Eterogeneità
- $e_1 = 1$ ➔ Massima eterogeneità

31

- Es. 1: colore degli occhi:

c c c c c c

Colore occhi	n_i
castani	6
Totale	*6*

$e_1 = 0$

32

- Es. 2: colore degli occhi:

  c c v v a a

Colore occhi	n_i	f_i	f_i^2
castani	2	1/3	1/9
verdi	2	1/3	1/9
azzurri	2	1/3	1/9
Totale	*6*	*1*	*1/3*

$$e_1 = \frac{k}{k-1}\left(1 - \sum_{i=1}^{k} f_i^2\right) = \frac{3}{2} \cdot \left(1 - \frac{1}{3}\right) = \frac{3}{2} \cdot \frac{2}{3} = 1$$

33

- Es. 3: colore degli occhi:

  c c c v v a

Colore occhi	n_i	f_i	f_i^2
castani	3	3/6	9/36
verdi	2	2/6	4/36
azzurri	1	1/6	1/36
Totale	*6*	*1*	*14/36*

$$e_1 = \frac{k}{k-1}\left(1 - \sum_{i=1}^{k} f_i^2\right) = \frac{3}{2} \cdot \left(1 - \frac{14}{36}\right) = \frac{3}{2} \cdot \frac{22}{36} = 0{,}917$$

34

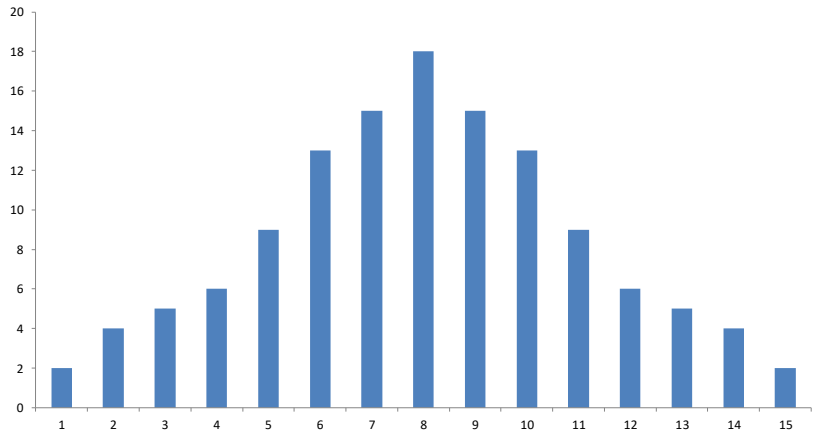
Simmetria e asimmetria

- Una distribuzione è detta simmetrica se è possibile individuare un asse verticale che suddivida la distribuzione in due parti uguali.
- Per una distribuzione simmetrica con *k* modalità:

$$n_1 = n_k, n_2 = n_{k-1}, \ldots$$

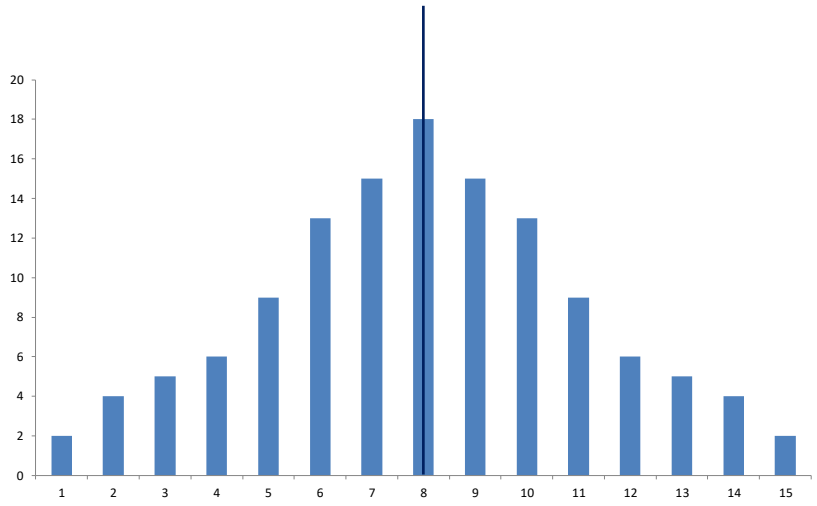
35

Distribuzione simmetrica



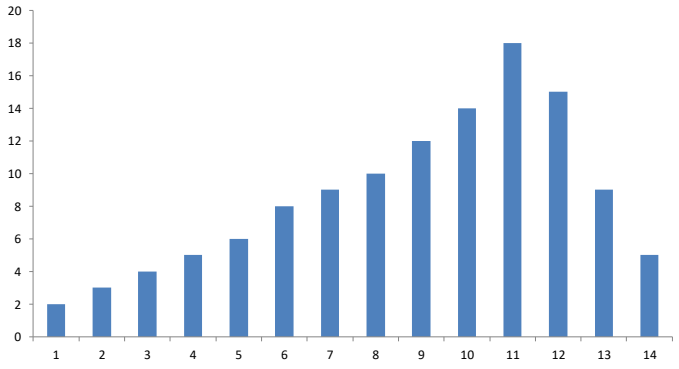
36

Distribuzione simmetrica



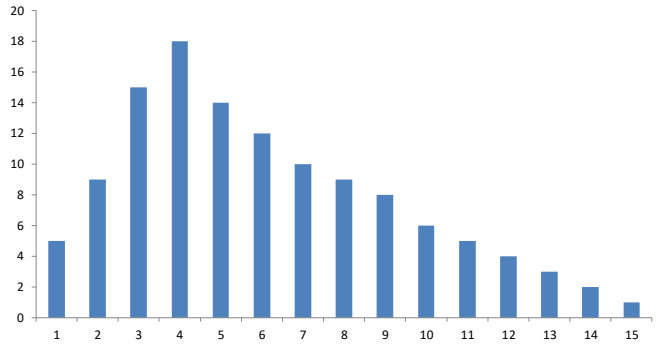
37

- Una distribuzione ha <u>asimmetria negativa</u> se si allunga lungo la coda sinistra (verso valori negativi)



38

- Una distribuzione ha asimmetria positiva se si allunga lungo la coda destra (verso valori positivi).



39

Media, mediana, moda

- Per una distribuzione simmetrica

$$\bar{x} = Me = Mo$$

- Per una distribuzione con asimmetria negativa

$$\bar{x} < Me < Mo$$

- Per una distribuzione con asimmetria positiva

$$\bar{x} > Me > Mo$$

40

Indici di asimmetria

- Il più importante è l'indice β di Fisher

$$\beta = \frac{M_3}{\sigma^3}$$

- M_3 è il momento centrale di ordine 3
- σ^3 è la deviazione standard al cubo

41

- Il momento centrale di ordine 3 è

$$M_3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n}$$

42

- $\beta = 0$ simmetria
- $\beta < 0$ asimmetria negativa
- $\beta > 0$ asimmetria positiva

43

- In generale si definisce il momento centrale di ordine *r*

$$M_r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^r}{n}$$

- Da notare che:

$$M_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n} = 0$$

$$M_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \sigma^2$$

44

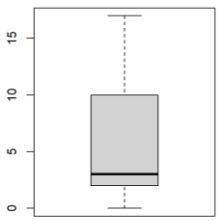
- Per distribuzioni di frequenze, nel calcolo del momento centrale di ordine 3 bisogna considerare anche le frequenze assolute

$$M_3 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^3 n_i}{n}$$

45

Asimmetria e box-plot

- L'asimmetria è anche rilevabile con il box-plot.



- In questo esempio la distribuzione si allunga verso la coda destra: asimmetria positiva. Il primo 50% dei valori è compreso tra 0 e 3, il rimanente 50% è compreso tra 3 e 17.

46