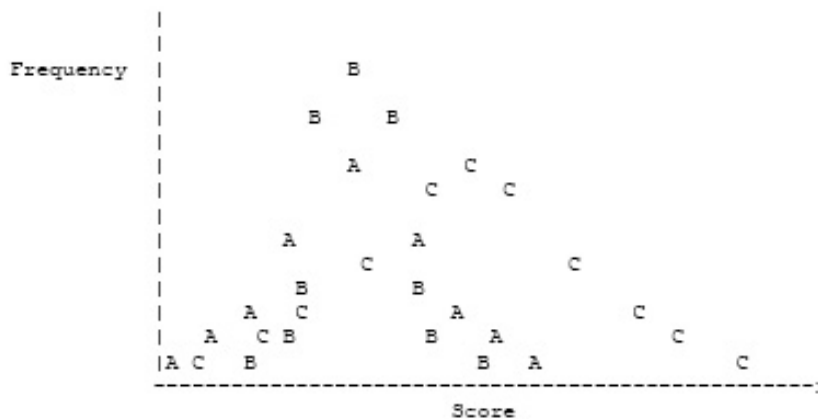


Statistica. Esercizi su Indici di posizione e variabilità

20 novembre 2023

Domande

Domanda 1 Le domande da (a) a (e) si riferiscono alla seguente figura



NOTA: Si uniscano i punti contrassegnati con A con una linea continua per disegnare la distribuzione A, i punti contrassegnati con B per disegnare la distribuzione B, ed infine i C per ottenere la distribuzione C.

- a la media di quale distribuzione differisce dalle altre? → distribuzione c (perchè traslata a destra rispetto alle altre)
- b quale distribuzione ha la più piccola deviazione standard? → distribuzione b (perchè i valori sono più concentrati intorno alla media)
- c quale distribuzione ha la più piccola varianza? → distribuzione b (la varianza è il quadrato della deviazione standard)
- d è corretto affermare che la media della distribuzione A è approssimativamente uguale alla moda della distribuzione B? → Sì
- e Quale di queste distribuzioni è evidentemente asimmetrica? Di che tipo di asimmetria si tratta? Spiegare brevemente cosa vuol dire osservare tale asimmetria per una certa variabile → Nessuna

Domanda 2 Per una distribuzione perfettamente simmetrica la media e la mediana sono (Scegliere una delle seguenti opzioni)

- a Uguali
- b Sempre diverse
- c Dipende, a volte uguali, a volte diverse
- d Non abbiamo sufficienti informazioni per decidere

Soluzione: risposta esatta a

Domanda 3 Una distribuzione di 6 valori ha mediana uguale a 21. Se il valore osservato più alto (cioè il massimo) aumenta di 3, la mediana sarà uguale a:

- a 24
- b 21.5
- c 21
- d Dobbiamo rifare i calcoli

Soluzione: risposta esatta c (non cambia perchè aumenta solo il massimo, non cambia la numerosità del collettivo. Abbiamo sempre 6 valori)

Domanda 4 Se vi viene detto che in un certo collettivo una data variabile ha media pari a 25 e varianza 0, cosa potete concludere? (Scegliere una delle seguenti opzioni)

- a Qualcuno ha commesso un errore di calcolo
- b Che c'è solo un individuo in questo collettivo ($N=1$)
- c Che la numerosità del collettivo è 0 ($N=0$)
- d Che per tutti gli individui del collettivo la variabile assume valore 25
- e Nessuna delle risposte è corretta

Soluzione: risposta esatta d (vi è assenza di variabilità)

Domanda 5 Abbiamo a disposizione solo la media dei punteggi ad un test su 12 studenti. Il punteggio massimo tuttavia non era stato riportato correttamente. Ci è stato segnalato che quello corretto è più alto di 36 punti rispetto a quanto riportato. Quale effetto ha questa segnalazione sulla media? (Scegliere una delle seguenti opzioni)

-
- a che la nuova media sarà più alta di 12 punti
 - b che la media non cambia
 - c che la nuova media sarà più alta di 3 punti
 - d che la nuova media sarà più alta di 36 punti
 - e non c'è modo di sapere quanto varrà la nuova media senza avere a disposizione i dati grezzi

Soluzione: risposta esatta c

Domanda 6 La quantità $\sum (x - \bar{x})$ non viene usata per il calcolo di nessun indice di variabilità perchè (Scegliere una delle seguenti opzioni):

- a è sempre uguale a 0
- b è sempre negativa
- c è troppo difficile da calcolare
- d è sempre positiva

Soluzione: risposta esatta a (proprietà di baricentro della media)

Domanda 7 Data la seguente distribuzione di frequenza assoluta cumulata di una variabile X qualitativa ordinale con quattro modalità,

X	N_i
a	25
b	75
c	95
d	100

- a Quali indici di posizione possiamo calcolare?
- b Quali indici di variabilità?
- c Si calcolino gli indici individuati ai punti [a] e [b].

Soluzione: [a] Per una variabile qualitativa sconnessa l'unico indice di posizione individuabile è la moda

[b] Non conosciamo alcun indice di variabilità per una variabili qualitative sconnesse

[c] Calcoliamo la moda, facendo attenzione che la tabella riporta le frequenze assolute cumulate. Trasformiamole in frequenze assolute semplici:

X	N_i	
a	25	25
b	75	50
c	95	20
d	100	5

La moda, essendo il valore (o modalità) della variabile più frequente è $Moda = b$

Esercizi

Esercizio 1 In riferimento ai dati già utilizzati della durata (in ore) di 40 lampadine prodotte dall'azienda A e dall'azienda B

Azienda A					Azienda B				
684	697	720	723	821	819	836	888	897	903
831	835	848	852	852	907	912	918	942	943
859	860	868	870	876	952	959	962	986	992
893	899	905	909	911	994	1004	1005	1007	1015
922	924	926	926	938	1016	1018	1020	1022	1034
939	943	946	954	971	1038	1072	1077	1077	1082
972	977	984	1005	1014	1096	1100	1113	1113	1116
1016	1041	1052	1080	1093	1153	1154	1174	1188	1230

è stata costruita la seguente distribuzione di frequenza relativa in classi:

Durata (in ore)	f_i Azienda A	f_i Azienda B
[650, 750)	0.1	0
[750, 850)	0.1	0.05
[850, 950)	0.5	0.2
[950, 1050)	0.225	0.4
[1050, 1150)	0.075	0.225
[1150, 1250)	0	0.125

- Calcolare media, mediana, primo e terzo quartile della distribuzione della durata delle lampadine per entrambe le aziende a partire dalla distribuzione disaggregata della variabile (dati grezzi). Quali sono le differenze tra le due distribuzioni?
- Calcolare media, mediana, primo e terzo quartile della distribuzione della durata delle lampadine per entrambe le aziende a partire dalla distribuzione in classi. Ci sono delle differenze tra i valori coalcati al punto [a]? Se sì, perché?
- Che tipo di indici sono quelli calcolati per le aziende A e B? marginali o condizionati? Spiegare.
- Calcolare la media marginale della durata delle lampadine a partire dalla distribuzione disaggregata dei dati

e. Calcolare la media marginale della durata delle lampadine a partire dalla distribuzione in classi

Soluzione: A partire dalla distribuzione dei dati grezzi la media si ottiene con la formula standard (come nell'esercizio precedente) e, per le due distribuzioni essa è:

$$\bar{x}_A = 908.4$$

$$\bar{x}_B = 1018.3$$

(l'azienda B sembra sempre essere migliore, anche in media!)

La mediana la si ottiene dalla distribuzione disaggregata ordinata (lo è già!), individuando il valore centrale tra i 40 valori delle due distribuzioni.

Essendo $n = 40$ pari (per entrambe le distribuzioni), abbiamo che la mediana si colloca tra la 20^a e la 21^a osservazione (nei dati ordinati) sia per la distribuzione dell'azienda A che dell'azienda B.

$$Me_A = \frac{911+922}{2} = 916.5$$

$$Me_B = \frac{1015+1016}{2} = 1015.5$$

(anche in mediana l'azienda B sembra produrre lampadine migliori dell'azienda A!)

Il primo quartile lo si individua sempre nella distribuzione (disaggregata) ordinata ma prendendo i 20 valori a sinistra della mediana esclusa (per forza in quanto non è un valore osservato). Pertanto esso si trova tra la 10^a e la 11^a osservazione della prima metà dei dati:

$$Q1_A = \frac{852+859}{2} = 855.5$$

$$Q1_B = \frac{943+952}{2} = 947.5$$

Il terzo quartile lo si individua sempre nella distribuzione (disaggregata) ordinata ma prendendo i 20 valori a destra della mediana (esclusa). Pertanto esso si trova tra la 10^a e la 11^a osservazione della seconda metà dei dati:

$$Q3_A = \frac{971+972}{2} = 971.5$$

$$Q3_B = \frac{1082+1096}{2} = 1089$$

I risultati sono:

	$\bar{x}$	Q_1	Mediana	Q_3
Azienda A	908.4	855.5	916.5	971.5
Azienda B	1018.3	947.5	1015.5	1089.0

[b]

Per calcolare la **media a partire da una distribuzione in classi** di una variabile quantitativa, basta approssimare i valori di ogni classe con i corrispondenti valori centrali.

Se la distribuzione in classi è costruita con le frequenze assolute allora ogni valore centrale va moltiplicato per la frequenza della classe corrispondente e la somma di tali prodotti diviso la somma delle frequenze (ossia $\sum n_i = n$):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n c_i n_i$$

Se la distribuzione in classi è costruita con le frequenze relative la formula è simile eccetto per il fatto che non si divide più per la somma delle frequenze (in quanto la divisione per n è già inclusa nel calcolo delle frequenze relative):

$$\bar{x} = \sum_{i=1}^n c_i f_i$$

Pertanto con riferimento alle nostre due distribuzioni di frequenza relativa in classi, le medie saranno:

$$\bar{x}_A = 700 \times 0.1 + 800 \times 0.1 + 900 \times 0.5 + 1000 \times 0.225 + 1100 \times 0.075 = 907.5$$

Analogamente per l'azienda B: $\bar{x}_B = 1017.5$

Per la **mediana a partire da una distribuzione di frequenza** occorre ricorrere alle frequenze cumulate (per semplicità quelle relative). E si ricordi che la mediana e il primo e terzo quartile sono i valori quei valori della variabile che lasciano alla propria sinistra le seguenti frazioni di dati: 0.5, 0.25 e 0.75.

Se, come nel nostro caso, la distribuzione è in classi, dopo aver individuato la classe che contiene il quartile ricercato bisogna trovare l'esatto valore dell'indice.

Primo passo: costruire la distribuzione delle frequenze (relative) cumulate:

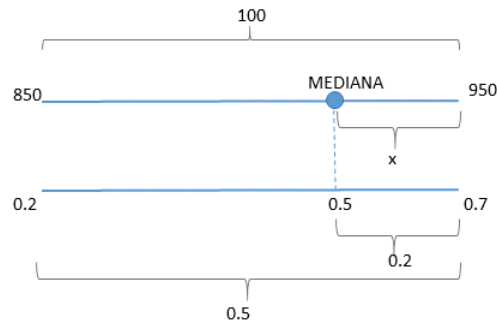
Durata (in ore)	F_i Azienda A	F_i Azienda B
[650, 750)	0.1	0
[750, 850)	0.2	0.05
[850, 950)	0.7	0.25
[950, 1050)	0.925	0.65
[1050, 1150)	1	0.875
[1150, 1250)	1	1

Consideriamo come esempio solo la ricerca della mediana della distribuzione della durata delle lampadine per l'azienda A.

Dato che la mediana è il quantile $Q_{0.5}$, la classe che lo contiene è la classe [850, 950).

Pertanto la ricerca dell'effettivo valore della mediana deve essere effettuata attraverso un metodo di *interpolazione lineare* dato dalla seguente proporzione (riportata visivamente in figura):

$$x : 0.2 = 100 : 0.5 \Rightarrow x = 0.2 \times \frac{100}{0.5} = 40$$



dunque la mediana delle durate per l'azienda A è: $Me_A = 950 - 40 = 910$

E così via anche per gli altri quartili.

I valori trovati quando consideriamo la distribuzione in classi sono:

	$\bar{x}$	Q_1	Mediana	Q_3
Azienda A	907.5	860	910	972.2
Azienda B	1017.5	950	1012.5	1094.4

Se confrontiamo questi valori con quelli ricavati dai dati grezzi, ovviamente vi è qualche differenza dovuta alle approssimazioni che sono necessarie per calcolare valori esatti a partire da intervalli (le classi).

[c.]

Tutti gli indici che abbiamo calcolato fino ad ora sono indici di posizione condizionati alle due aziende. Se vogliamo gli indici “marginali” basta non considerare i gruppi formati dalle aziende.

[d.]

Dalla distribuzione disaggregata della variabile basta sommare tutti i valori delle durate senza guardare all'appartenenza di quelle lampadine ad una delle due aziende. Oppure sfruttare il calcolo che abbiamo già fatto al punto a, ricordandoci che la media marginale date le medie nei gruppi e le numerosità nei gruppi è ottenuta come media ponderata:

$$\bar{x} = \frac{\bar{x}_A n_A + \bar{x}_B n_B}{n_A + n_B} = \frac{908.4 \times 40 + 1018.3 \times 40}{40 + 40} = 963.35$$

[e.]

Allo stesso modo del precedente punto calcoliamo la media marginale a partire dalla distribuzione in classi. Possiamo sfruttare il calcolo del punto b, oppure ricostruire la distribuzione di frequenza marginale (sommando le frequenze **assolute** delle classi nei due gruppi):

Durata (in ore)	i Azienda A	f_i Azienda B	n_i Azienda A e B
[650, 750) [650, 750)	0.1	0	4
[750, 850) [750, 850)	0.1	0.05	4+2=6
[850, 950) [850, 950)	0.5	0.2	20+8=28
[950, 1050) [950, 1050)	0.225	0.4	9+16=25
[1050, 1150) [1050, 1150)	0.075	0.225	3+9=12
[1150, 1250) [1150, 1250)	0	0.125	5

$$\bar{x} = \frac{700 \times 4 + 800 \times 6 + 900 \times 28 + 1000 \times 25 + 1100 \times 12 + 1200 \times 5}{80} = 962.5$$

Esercizio 2 Un questionario è stato somministrato ad un collettivo di 21 studenti. tra le varie domande vi era anche una relativa agli affitti mensili da essi sostenuto (in Euro). Sono stati osservati i seguenti dati (distribuzione disgregata):

(210, 170, 90, 190, 80, 200, 220, 140, 120, 130, 110, 320, 160, 260, 150, 650, 110, 310, 160, 190, 280).

Si costruisca un boxplot per rappresentare graficamente la distribuzione degli affitti rilevata sui 21 studenti.

Soluzione: Per costruire un boxplot abbiamo bisogno della cosiddetta **sintesi a 5** ovvero i valori:

Minimo, Q_1 , Me, Q_3 , Massimo

Avendo a disposizione la distribuzione disaggregata degli affitti, ordiniamola :

80, 90, 110, 110, 120, 130, 140, 150, 160, 160, 170, 190, 190, 200, 210, 220 260, 280, 310, 320, 650

e cerchiamo i quartili desiderati (e poi verificheremo se il minimo e il massimo debbano essere inclusi nel grafico):

$$Q_1 = 130, Me = 170, Q_3 = 220$$

Questi rappresenteranno la scatola del boxplot.

Calcoliamo ora i valori teorici che ci indicheranno se il minimo (o valori estremamente piccoli) e il massimo (o valori estremamente grandi) dovranno essere i valori che chiuderanno i baffi del boxplot oppure no:

- Calcoliamo il limite inferiore $LI = Q_1 - 1.5 \cdot IQR = 130 - 1.5 \cdot (220 - 130) = 130 - 135 = -5$
- e non essendoci valori osservati inferiori ad esso possiamo chiudere il baffo a 80;
- Calcoliamo il limite superiore $LS = Q_3 + 1.5 \cdot IQR = 220 + 1.5 \cdot (220 - 130) = 220 + 135 = 355$
- in questo caso abbiamo un solo valore che eccede questo limite (650), quello immediatamente inferiore ad esso è 320 (il quale verifica il criterio: inferiore a 355), pertanto possiamo chiudere il baffo proprio a 320 e segnalare nel grafico 650 come outlier (mediante un punto o un asterisco)

Esercizio 3 Supponiamo di osservare il voto di diploma (in centesimi) degli studenti in 2 corsi universitari (corso A e corso B). I voti di diploma dei frequentanti il corso A sono i seguenti: 61, 68, 83, 86, 73; i voti di diploma dei frequentanti il corso B sono invece: 68, 74, 82, 83.

-
- a. Calcolare le medie condizionate e a partire da queste calcolare anche la media marginale
- b. Calcolare le varianze nei due gruppi e quella marginale

Soluzione: [a.]

Le medie condizionate sono le medie nelle due classi:

$$\bar{x}_A = 371/5 = 74.2, \bar{x}_B = 307/4 = 76.75$$

A partire da queste è semplice ricavare la media marginale (o totale) mediante la relazione:

$$\bar{x} = \frac{1}{n_A+n_B}(\bar{x}_A \times n_A + \bar{x}_B \times n_B) = \frac{1}{9}(5\bar{x}_A + 4\bar{x}_B) = \frac{74.2 \times 5 + 76.75 \times 4}{5+4} = 75.33$$

[b.]

per le varianze non conosciamo la relazione (che pure esiste) tra varianze condizionate e varianze marginali pertanto usiamo i dati grezzi.

Per la prima varianza condizionata all'appartenenza alla classe A, consideriamo solo i 5 valori in quella classe:

$$\sigma_A^2 = 86.2$$

Per la classe B solo i 4 valori in quella classe:

$$\sigma_B^2 = 37.7$$

Per la marginale tutti e 9 i valori, indipendentemente dalla classe di appartenenza (non guardando l'appartenenza ai gruppi!)

$$\sigma^2 = 66.2$$