Esercitazione 1: Frequenze, indici di posizione e amici

Angela Andreella

20/10/2020

Variabile Qualitativa

Si rileva il tipo di diploma di scuola secondaria di 12 studenti del vostro dipartimento. Identifichiamo con C = classico, S = scientifico, T = tecnico, A = altro.

$$X = (C, T, C, C, T, A, A, T, S, S, T, A)$$

1. Identificare la popolazione, l'unita statistica, variabile rilevata, tipo di variabile rilevata e modalitá della variabile.

Popolazione = tutti gli studenti del vostro dipartimento

Unitá statistica = lo studente

Variabile Rilevata = Diploma di scuola secondaria

Tipo di variabile = Qualitativa nominale, ovvero non ordinabile

Modalità = C, S, T, A

2. Calcolare la tabella di frequenza assoluta e relativa

Richiamiamo un po' di formule:

La nostra variabile osservata è X, per calcolare le frequenze assolute che indichiamo con f_i basta semplicemente contare quante volte si presenta ciascuna modalità.

Per le frequenze relative invece utilizziamo la seguente formula:

$$p_i = \frac{f_i}{n}$$

dove n è il numero totale di osservazioni, nel nostro caso 12.

	f_i	p_i
A	3	$3/12 = 0.250$
C	3	$3/12 = 0.250$
S	2	$2/12 = 0.167$
T	4	$1/12 = 0.333$
Totale	12	1

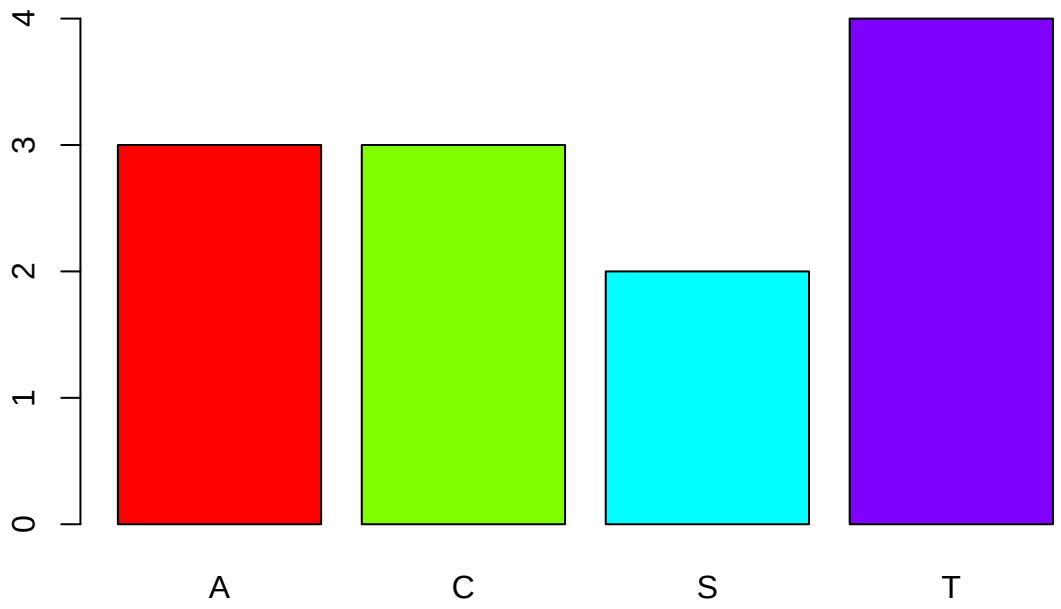
NB: Controllate sempre che

- La somma delle frequenze assolute sia uguale a n, ovvero: $\sum_{i=1}^{k} f_i = n$ nel nostro caso $\sum_{i=1}^{4} f_i = 12$;
- La somma delle frequenze relative sia uguale a 1, ovvero: $\sum_{i=1}^{k} p_i = 1$ nel nostro caso $\sum_{i=1}^{4} p_i = 1$;

Con k (nelle formule dei due punti precedenti) indichiamo le modalità, in questo caso il totale delle modalità è pari a 4.

3. Creare il grafico a barre

Semplicemente dobbiamo riportare sull'asse delle x le nostre modalità e sull'asse delle y le nostre frequenze assolute f_i o relative p_i:



4. Calcolare la Moda

La moda, che indichiamo con Mo, è la modalità dove troviamo la frequenza assoluta o relativa maggiore. Si può vedere chiaramente dal grafico a barre del punto 3. o utilizzando la tabella creata nel punto 2. La moda in questo caso è T, avendo frequenza assoluta pari a 4.

Variabile Quantitativa

Si rileva la durata in secondi di 10 brani musicali di musica heavy metal:

$$X = (180; 300; 250; 60; 250; 60; 60; 270; 300; 270)$$

1. Identificare la popolazione, l'unita statistica, variabile rilevata, tipo di variabile rilevata e modalitá della variabile.

Popolazione = tutti i brani heavy metal

Unitá statistica = il brano musicale

Variabile Rilevata = durata del brano in secondi

Tipo di variabile = Quantitativa a scala a rapporti equivalenti, poichè se il brano ha durata 0 significa che non vi è un brano

Modalità = 60, 180, 250, 270 e 300. In questo caso l'esercizio è parecchio semplificato avendo solo 10 osservazioni, nel caso avessimo più osservazioni sarebbe carino raggruppare le osservazioni in classi per esempio considerando i minuti, ovvero da 0 a 60, da 60 a 120 etc.

2. Calcolare la tabella di frequenza assoluta e relativa e rispettive cumulate

Per le frequenze assolute e relative riguardate le formule dell'esercizio precedente :)

Le frequenze cumulate assolute della modalità i si calcolano sommando le frequenze assolute riferite a i e prima di i, ovvero:

$$F_i = \sum_{j=1}^{i} f_j$$

uguale per le frequenze cumulate relative, dobbiamo sommare le frequenze relative fino a i:

$$P_i = \sum_{j=1}^{i} p_j$$

	f_i	p_i	F_i	P_i
60	3	0.3	3	0.3
180	1	0.1	4	0.4
250	2	0.2	6	0.6
270	2	0.2	8	0.8
300	2	0.2	10	1
Sum	10	1		

NB: Controllate sempre che

- F_i dell'ultima modalità sia uguale a n, nel nostro caso $F_5 = 10$;
- P_i dell'ultima modalità sia uguale a 1, nel nostro caso $P_5 = 1$;

Qualche esempio per capire gli indici e la formula delle frequenze cumulate:

- $F_1 = \sum_{j=1}^{1} f_j = f_1 = 3$;
- $F_2 = \sum_{j=1}^{2} f_j = f_1 + f_2 = 3 + 1 = 4$;
- $F_3 = \sum_{j=1}^{3} f_j = f_1 + f_2 + f_3 = 3 + 1 + 2 = 6$;
- $F_4 = \sum_{j=1}^{4} f_j = f_1 + f_2 + f_3 + f_4 = 3 + 1 + 2 + 2 = 8$;

- $F_5 = \sum_{j=1}^{5} f_j = f_1 + f_2 + f_3 + f_4 + f_5 = 3 + 1 + 2 + 2 + 2 = 10$;

uguale per le frequenze cumulate relative ma considerando p_i invece di f_i.

4. Calcolare la Media

Ok, qui abbiamo due possibilità:

1. Utilizzare i dati grezzi: $X = (180; 300; 250; 60; 250; 60; 60; 270; 300; 270)$, la media sarà uguale a:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{\sum_{i=1}^{10} X_i}{10} = \frac{180 + 300 + 250 + 60 + 250 + 60 + 60 + 270 + 300 + 270}{10} = 200$$

2. Utilizzare la tabella di frequenze che abbiamo appena calcolato. Se utilizziamo le frequenze assolute, la formula della media sarà uguale a:

$$\bar{X} = \frac{\sum_{i=1}^{k} X_i f_i}{n} = \frac{\sum_{i=1}^{5} X_i f_i}{10} = \frac{(60 \cdot 3) + (180 \cdot 1) + (250 \cdot 2) + (270 \cdot 2) + (300 \cdot 2)}{10} = 200$$

se usiamo le frequenze relative, la formula della media sarà uguale a:

$$\bar{X} = \sum_{i=1}^{k} X_i p_i = \sum_{i=1}^{5} X_i p_i = (60 \cdot 0.3) + (180 \cdot 0.1) + (250 \cdot 0.2) + (270 \cdot 0.2) + (300 \cdot 0.2) = 200$$

Deve dare sempre lo stesso risultato ovviamente :) le formule sono equivalenti, la prima se abbiamo molte osservazioni non solo 10 è più antipatica.

5. Calcolare la mediana e i quartili

Anche qui ci sono diversi metodi, usate quello che più vi piace, ancora una volta comunque il risultato deve essere sempre lo stesso! Quindi potete provare con diversi metodi per poter fare un check sui vostri calcoli :)

1. Utilizzando i dati grezzi: $X = (180; 300; 250; 60; 250; 60; 60; 270; 300; 270)$. Per prima cosa dobbiamo ordinarli, poichè la mediana si riferisce a un dato ordinato. 60 60 60 180 250 250 270 270 300 300

   Dobbiamo trovare la modalità che sta in "mezzo". Avendo $n = 10$ osservazioni, l'indice che sta in mezzo è tra 5 e 6. Che fortuna! le modalità in posizione 5 e 6 è la stessa, ovvero 250. Se invece avessimo avuto due diverse modalità per esempio 250 e 270 potevamo prendere come mediana il valore intermedio tra questi due valori (interpolazione lineare), ovvero $(250 + 270)/2$.

2. Utilizzando le frequenze cumulate assolute: sappiamo da prima che la modalità mediana è tra la posizione 5 e 6. Possiamo dunque analizzare le frequenze cumulate assolute e vedere dove si trovano il 5 e il 6, nel nostro caso sono inglobati in $F_3 = 6$, ovvero la mediana è magicamente ancora 250!;

3. Utilizzando le frequenze relative: sappiamo che la mediana è ci indica la modalità che spacca a metà la nostra distribuzione di frequenza, ovvero che raccoglie il 50% delle osservazioni. Possiamo dunque utilizzare le frequenze cumulate relative, vedendo dove si trova lo 0.5, ci riferiamo dunque a $P_3 = 0.6$, ovvero la mediana è ancora 250.

La mediana è semplicemente il secondo quartile, ovvero $Me = Q_2$, dunque possiamo usare i metodi precedenti per calcolare il primo quartile che chiamiamo Q_1, e il terzo quartile che chiamiamo Q_3. Il primo quartile raccoglie il primo 25% della nostra distribuzione, mentre il terzo il 75%.

Anche qui abbiamo tre metodi:

1. Utilizzando i dati grezzi ordinati come per il punto 1 della mediana. Sappiamo che il quartile racchiude il 25% della nostra distribuzione, dunque se moltiplichiamo 0.25 per il numero totale di osservazioni $n = 10$ abbiamo la posizione del primo quartile, ovvero 2.5, cioè tra 2 e 3, che nel nostro caso è 60!. Se avessimo avuto due valori diversi, conviene arrotondare 2.5 al numero intero superiore, ovvero 3 e prendere l'osservazione che sta alla terza posizione;

2. Utilizzando le frequenze cumulate assolute: Come prima sappiamo che il nostro Q_1 è in posizione 2.5. Analizzando la colonna F_i vediamo che 2.5 è contenuto nella prima riga! Quindi ancora una volta il primo quartile è pari a 60;

3. Utilizzando le frequenze cumulate relative: Sappiamo che Q_1 tiene il 25% della distribuzione. Dunque, dove è il 25%? Se guardiamo la la colonna P_i vediamo che 0.25 è contenuto ancora nella prima riga! Quindi ancora una volta il primo quartile è pari a 60 (e per fortuna!);

Facciamo lo stesso procedimento per il terzo quartile, ovvero considerando 0.75 invece di 0.25, se avete problemi scrivetemi :) (Soluzione: $Q_3 = 270$)

6. Calcolare il campo di variazione, lo scarto interquartile

Il campo di variazione è semplicemente la differenza tra massima modalità e minima modalità, ci indica dunque appunto in quale campo prende valori la nostra variabile. In questo caso:

$$CV = X_{max} - X_{min} = 300 - 60 = 240$$

e lo scarto interquartile la differenza tra terzo e primo quartile:

$$Q = Q_3 - Q_1 = 270 - 60 = 210$$

7. Calcolo della varianza, deviazione standard e coefficiente di variazione

Anche qui, come per la media, abbiamo vari metodi (sempre tutti equivalenti) per calcolare la varianza. Ricordiamo che la media $\bar{X}$ è pari a 200.

1. Utilizzando i dati grezzi: $X = (180; 300; 250; 60; 250; 60; 60; 270; 300; 270)$, dunque:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} = \frac{\sum_{i=1}^{10}(X_i - 200)^2}{10} =$$
$$= \frac{(180-200)^2 + (300-200)^2 + (250-200)^2 + (60-200)^2 + (250-200)^2 +}{}$$
$$\frac{(60-200)^2 + (60-200)^2 + (270-200)^2 + (300-200)^2 + (270-200)^2}{10} = 9400$$

2. Utilizzando le frequenze assolute:

$$S^2 = \frac{\sum_{i=1}^{k}(X_i - \bar{X})^2 f_i}{n} = \frac{\sum_{i=1}^{5}(X_i - \bar{X})^2 f_i}{10}$$
$$= \frac{((60-200)^2 \cdot 3) + ((180-200)^2 \cdot 1) + ((250-200)^2 \cdot 2) + ((270-200)^2 \cdot 2) + ((300-200)^2 \cdot 2)}{10} = 9400$$

3. Utilizzando le frequenze relative

$$S^2 = \sum_{i=1}^{k}(X_i - \bar{X})^2 p_i = \sum_{i=1}^{5}(X_i - \bar{X})^2 p_i$$
$$= ((60-200)^2 \cdot 0.3) + ((180-200)^2 \cdot 0.1) + ((250-200)^2 \cdot 0.2) + ((270-200)^2 \cdot 0.2) + ((300-200)^2 \cdot 0.2) = 9400$$

La deviazione standard non è altro che la radice quadrata della varianza, ovvero $S = \sqrt{S^2} = 96.9536$. Il coefficiente di variazione invece è pari a $CV = \dfrac{S}{|\bar{X}|} = \dfrac{96.9536}{200} = 0.484768$.

Esercizi sparsi su frequenze e media

1. Avendo tre osservazioni ma conosciamo solo il valore di due, pari a 80 e 70, e il valore della media pari a 70. Quanto sarà la terza osservazione?

$$\bar{X} = \frac{X_1 + X_2 + X_3}{n} \qquad 70 = \frac{80 + 70 + X_3}{3} \rightarrow X_3 = (70 \cdot 3) - (80 - 70) = 60$$

2. Avendo 3 osservazioni e conoscendo le seguenti frequenze relative: $p_1 = 0.3$, $p_2 = 0.4$, quanto sarà la terza?

$$\sum_{i=1}^{3} p_i = 1 \qquad 0.3 + 0.4 + p_3 = 1 \rightarrow p_3 = 1 - 0.3 - 0.4 = 0.3$$

3. Avendo $n = 10$ e le seguenti frequenze assolute: $f_1 = 3$, $f_2 = 1$, $f_3 = 4$, quanto sarà f_4?

$$\sum_{i=1}^{4} f_i = 10 \qquad 3 + 1 + 4 + f_4 = 10 \rightarrow f_4 = 10 - (3 + 1 + 4) = 2$$

4. Avendo $X_1 = 2$, $f_1 = 1$, $X_2 = 5$, $f_2 = 2$, $X_3 = 10$, e la media uguale a 7, quanto è f_3?

$$\bar{X} = \frac{\sum_{i=1}^{3} X_i f_i}{n} \quad 7 = \frac{(2 \cdot 1) + (5 \cdot 2) + (10 \cdot f_3)}{1 + 2 + f_3} \rightarrow f_3 = 3$$

5. Si consideri una rilevazione di 10 osservazione su una variabile che assume i seguenti valori

X	p_i
1	0.4
3	0.3
4	0.1
5	0.2

Calcolare F_2, media aritmetica, mediana e varianza.

F_2 si riferisce alla frequenza assoluta della seconda modalità, dunque per prima cosa dobbiamo calcolare la tabella delle frequenze assolute. Sapendo che

$$p_i = f_i/n$$

che nel nostro caso significa $p_i = f_i/10$, ovvero $f_i = p_i \cdot 10$. La tabella delle frequenze assolute e frequenze assolute cumulate è pari a:

X	p_i	f_i	F_i
1	0.4	4	4
3	0.3	3	7
4	0.1	1	8
5	0.2	2	10

Dunque F_2 è uguale a 7.

La media aritmetica è uguale a:

$$\bar{X} = \sum_{i=1}^{4} X_i p_i = (1 \cdot 0.4 + 3 \cdot 0.3 + 4 \cdot 0.1 + 5 \cdot 0.2) = 2.7$$

o potete usare le formule spiegate sopra con le frequenze assolute.

La mediana avendo 10 osservazioni è tra le posizioni 5 e 6, ovvero la modalità 3. Provate a usare gli altri metodi spiegati sopra :)

La varianza ${‘e} pari a

$$S_X^2 = \sum_{i=1}^{4} (X_i - 2.7)^2 p_i = (1 - 2.7)^2 \cdot 0.4 + (3 - 2.7)^2 \cdot 0.3 + (4 - 2.7)^2 \cdot 0.1 + (5 - 2.7)^2 \cdot 0.2 = 2.975$$

6. Avendo la variabile X rilevata che assume i valori $2, 3, 4, 5, 6, 8$ e 16 con $F_1 = 5$, $F_2 = 9$, $F_3 = 12$, $F_4 = 17$, $F_5 = 20$, $F_6 = 22$, $F_7 = 25$. Dire qual è la frequenza assoluta della modalità 8.

Sappiamo che F_i indica la frequeza cumulata assoluta della modalità i, dunque dobbiamo cercare f_6. Sapendo che:

$$F_6 = f_1 + f_2 + f_3 + f_4 + f_5 + f_6$$

e

$$F_5 = f_1 + f_2 + f_3 + f_4 + f_5$$

possiamo semplicemente trovare f_6 sottraendo F_5 da F_6:

$$f_6 = F_6 - F_5 = f_1 + f_2 + f_3 + f_4 + f_5 + f_6 - f_1 + f_2 + f_3 + f_4 + f_5 = 2$$

7. Avendo $X = \{1, 2, 2, 2, 2, 3, 4, 5, 6, 6, 7, 7, 7, 8, 8, 9, 9, 9, 10, 11, 11, 15\}$, calcolare i quartili.

Anche qui come prima possiamo fare la tabella delle frequenze assolute/relativa o usare i dati grezzi. In questo caso useremo i dati grezzi ma voi provate tutte le varie alternative :)

Il primo quartile è in posizione $n \cdot 0.25 = 22 \cdot 0.25 = 5.5$, arrotondando al numero intero superiore, cerchiamo la modalità in posizione 6, ovvero 3

Il secondo quartile, ovvero la mediana, è tra la posizione $n/2 = 11$ e $(n)/2+1 = 12$, avendo la stessa modalità nelle posizioni 11 e 12, la mediana è 7.

Il terzo quartile è in posizione $n \cdot 0.75 = 22 \cdot 0.75 = 16.5$, arrotondando al numero intero superiore, cerchiamo la modalità in posizione 17, ovvero 9.

8. Avendo $X = \{1, 50, 55, 60, 65\}$ calcolare la media e la mediana, e commentare.

La media è semplicemente

$$\bar{X} = \frac{1 + 50 + 55 + 60 + 65}{5} = 46.2$$

La mediana invece è in posizione $(n + 1)/2 = 6/2 = 3$, ovvero è pari a 55.

Notiamo che la mediana è un indice robusto verso possibili outliers, difatti la media è influenzata parecchio da quell'1, mentre la mediana no.