Statistics - Practice

Domenico De Stefano

1 novembre 2023

1 Statistica - Esercizi 1 (Soluzioni)

Vero o Falso?

a. Falso. L'affermazione così proposta è falsa perchè non viene indicata la natura casuale del campione che stiamo prendendo in esame, né il modo in cui è stato racccolto. Non si può quindi dire se il campione riflette le caratteristiche della popolazione o se sia stato eseguito secondo le leggi della probabilità. La numerosità non è indicativa della rappresentatività del campione.

b. Falso. Perchè non ha le stesse caratteristiche di tutta la popolazione degli studenti universitari in Italia e di conseguenza non è rappresentativo.

c. Vero. Rappresenta la restrizione di x ad una certa modalità.

d. Vero. La formula della densitá di una classe frequenza assoluta di Y sull'intervallo/lunghezza dell'intervallo

e. Vero. Rappresenta il punto in cui la frequenza è maggiore, e il picco del grafico.

f. Falso. Dipende dalla mia distribuzione. Se ad esempio esistono nelle modalità valori negativi questi potrebbero portare ad una media negativa.

g. Falso. Essi coincidono se una distribuzione è costante o sei i valori centrali della mia distribuzione sono tutti uguali.

h. Falso. Il valore 0.5 rappresenta la posizione della mediana nelle frequenze relative cumulate.

i. Falso. Ad esempio non può essere calcolata nelle variabili qualitative.

Esercizio 1

a. La popolazione è rappresentata dall'insieme delle famiglie (di un comune, di una regione, ecc.) e le telefonate sono un esempio di variabile quantitativa discreta.

b. La popolazione è composta da tutti i cittadini del comune di Trieste (possibilmente proprietari di un immobile). La lista dei cittadini è reperibile all'anagrafe ad esempio. L'ampiezza della casa di poprietà è un esempio di variabile quantitativa continua (perchè frutto di misurazione con precisione arbitrariamente crescente).

c. La popolazione è rappresentata dai clienti di un centro commerciale. La spesa giornaliera è una variabile quantitiva discreta.

d. La popolazione è composta da tutte le aule dell'Università di Trieste e la presenza di connessione Internet è un variabile qualitativa sconnessa (e dicotomica, dato che può assumere solo due modalità: presenza/assenza)

e. La popolazione è costituita da tutti gli aventi diritto al voto in Italia, pertanto cittadini maggiorenni residenti in Italia. La variabile intenzioni di voto è una variabile qualitativa sconnessa (in quanto si può chiedere alle unità statistiche di esprimere una preferenza per un candidato, un partito o una coalizione).

f. La popolazione è composta da tutti gli iscritti all'Università di Trieste. La variabile è qualitativa ordinale.

g. La popolazione è composta dai dipendenti della regione Friuli Venezia Giulia. Variabile qualitativa ordinale.

Esercizio 2

a. La popolazione è l'insieme delle famiglie Italiane che possiedono almeno un animale domestico

b. Per la variabile al punto 1 le possibili modalità potrebbero essere: supermercato, negozi specializzati, negozi online, ecc.; le variabili ai punti 2, 3 e 4 si possono lasciare aperte (cioè senza predefinire alcuna modalità);

c. L'unica variabile qualitativa è la 1 (qualitativa sconnessa)

d. Le domande 2, 3 e 4 sono quantitative. L'unica tra queste che è continua è la 4 in quanto è una spesa media settimanale (frutto di un calcolo)

e. In generale non sarebbe a priori sbagliato dire che per le variabili 2, 3 e 4 possiamo costruire delle distribuzioni di frequenza in classi in quanto tutte quantitative. Tuttavia dato che la 2 e la 3 sono discrete (e potrebbero assumere valori piuttosto contenuti visto il particolare oggetto di analisi), nello specifico dovremo di sicuro costruire delle classi per la variabile 4 (e valutare dopo la raccolta dati se dobbiamo costruirle anche per la 2 e la 3).

Esercizio 3 [a] Si ricordi che se scegliamo a priori il numero di classi e vogliamo che siano equiampie, una semplice formula per stabilire l'ampiezza delle classi è la seguente:

$$Ampiezza = \frac{CV}{Numeroclassi}$$

dove CV è il campo di variazione (o Range) della variabile che vogliamo ripartire in classi, ossia $CV = x_{max} - x_{min}$.

Per cui $CV = 98.6 - 11.2 = 87.4$, e dunque:

$$Ampiezza = \frac{87.4}{9} = 9.711 \approx 10$$

Un ampiezza esatta di 9.711 è approssimabile tranquillamente a 10!

Inoltre non è necessario partire dal minimo osservato (cioè da $x_{min} = 11.2$), ma è solo necessario che il minimo sia incluso nella prima classe! Scegliamo pertanto un più "leggibile" estremo inferiore per la prima classe, ad esempio 10. Pertanto la prima classe sarè l'intervallo $(10, 20]$ (11.2 è incluso!), la seconda classe è l'intervallo $(20, 30]$, e così via fino all'ultima classe $(90, 100]$. Analogamente alla prima l'unica altra cosa da controllare è se il massimo osservato cade nell'ultima classe. Se è così la ripartizione in classi funziona.

Regola: La costruzione in classi è arbitraria per numero di classi, ampiezza delle classi (potete approssimare come volete) e scelta dell'estremo inferiore della prima classe. L'unica cosa da controllare è se il minimo e il massimo osservati ricadano rispettivamente nella prima e nell'ultima classe.

[b]

Abbiamo scelto di approssimare l'ampiezza esatta 9.711 a 10.

Esercizio 4 Per alcuni dei quesiti di questo esercizio dovranno essere usate le frequenze cumulate che servono a rispondere a questioni del tipo: quante unità statistiche hanno valori della variabile inferiore a, superiore a, ecc.

a. Sono coloro che si sono collocati nella prima classe, 2 candidati (il $2/50 \times 100 = 4\%$ dei candidati)

b. 18 candidati (il $18/50 \times 100 = 36\%$ dei candidati)

c. Qui usiamo le frequenze cumulate, ovvero sommiamo le frequenze delle classi 450-499 e 500-549: quindi 2+18=20 candidati (ossia $20/50 \times 100 = 40\%$ dei candidati)

d. Anche qui usiamo le cumulate ma in pratica essendo 750 l'estremo superiore dell'ultima classe cumuliamo le frequenze di tutte le classi e quindi la risposta è tutti e 50 i candidati (100% dei candidati).

Esercizio 5 [a] Per completare la tabella serve l'ultima frequenza assoluta e per ricavarcela ci serve la numerosità del nostro collettivo, N. L'unico elemento che ci consente di ricavare il valore di N è l'unica frequenza relativa nota. Pertanto ricordando che la frequenza relativa della i-ma modalità di una certa variabile (o di un classe) è:

$$f_i = \frac{n_i}{N}$$

(dove n_i è la corrispondente freequenza assoluta) possiamo facilmente risolvere per N:

$$N = \frac{n_i}{f_i}$$

Nel nostro caso:

$$N = \frac{n_i}{f_i}$$
$$N = 11/0.22 = 50$$

Ora possiamo ricavare la frequenza assoluta della modalità C (che possiamo indicare ad esempio con n_C):

$$n_C = N - \sum_{i \in (A,B)} n_i$$
$$n_C = 50 - (11 + 30) = 9$$

Calcoliamo le frequenze relative delle modalità B e C, con la formula $f_i = \frac{n_i}{N}$

$$f_B = \frac{n_B}{N} = \frac{30}{50} = 0.6$$
$$f_C = \frac{n_C}{N} = \frac{9}{50} = 0.18$$

La percentuale della i-ma modalità corrisponde alla i-ma frequenza relativa moltiplicata per 100:

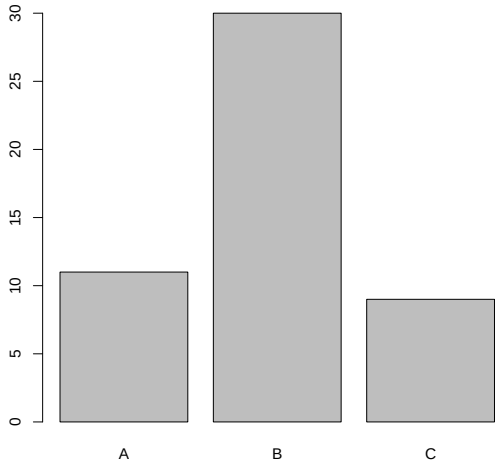
$$\%f_A = (\frac{n_A}{N}) \times 100 = (\frac{11}{50}) \times 100 = 22\%$$
$$\%f_B = (\frac{n_B}{N}) \times 100 = (\frac{30}{50}) \times 100 = 60\%$$
$$\%f_C = (\frac{n_C}{N}) \times 100 = (\frac{9}{50}) \times 100 = 18\%$$

La tabella completa

Modalità	n_i	f_i	%
A	11	0.22	22
B	30	0.6	60
C	9	0.18	18

[b]

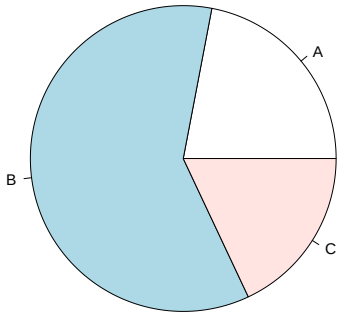
Per costruire un diagramma a barre per una variabile qualitativa sconnessa si riporta semplicemente la frequenza assoluta (o relativa o la percentuale) sull'asse delle Y e le modalita' della variabile sull'asse delle X. Si tracciano poi delle barre di altezza proporzionale alla frequenza scelta



[c]

Per costruire il grafico a torta riportiamo le frequenze assolute (meglio le relative!) su scala dei gradi sessagesimali (cioè in 360^o), in modo da calcolare l'ampiezza dell'angolo poporzionale alla frequenza da riprodurre nel grafico:

$$Angolo_A = f_A \times 360 = 79.2$$
$$Angolo_B = f_B \times 360 = 216$$
$$Angolo_C = f_C \times 360 = 64.8$$



Exercise 6 [a.]

Le unità statistiche sono le 3756 imprese sulle quali è stato rilevato il costo del lavoro sostenuto.

[b.]

Per calcolare le frequenze assolute occorre ricordarsi la definizione di densità e fare attenzione che queste, nel caso specifico, sono state calcolate dalle frequenze relative.

Pertanto la formula da considerare per l'i-ma classe è:

$$\text{densità}_i = \frac{f_i}{ampiezza_i}$$

per cui è facile ricavare tute le frequenze relative e per differenza la frequenza mancante (e da quella la densità mancante). Nota: le densità non sommano a 1, le frequenze relative si!

$$f_i = \text{densità}_i \times ampiezza$$

e dunque per le prime tre classi:

$$f_1 = 0.0125 \times (12 - 0) = 0.15$$
$$f_2 = 0.02 \times (30 - 12) = 0.02 \times 18 = 0.36$$
$$f_3 = 0.02 \times (50 - 30) = 0.02 \times 20 = 0.4$$

La frequenza relativa mancante per la quarta classe è dunque: $1 - (0.15 + 0.36 + 0.4) = 0.09$

E la corrispondente densità:

$$\text{densità}_4 = \frac{0.09}{100-50} = 0.002$$

Pertanto la tabella completa sarà:

Costo lav.	**Densità di frequenza relativa**	**frequenza relativa**	**frequenza assoluta**
(0, 12]	0.0125	0.15	$0.15 \times 3756 = 564$
(12, 30]	0.02	0.36	$0.36 \times 3756 = 1352$
(30, 50]	0.02	0.4	$0.4 \times 3756 = 1502$
(50, 100]	0.0018	0.09	$0.09 \times 3756 = 338$

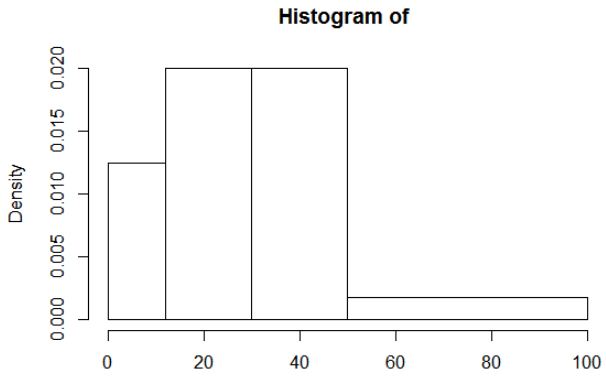
[c]

I valori centrali sono:

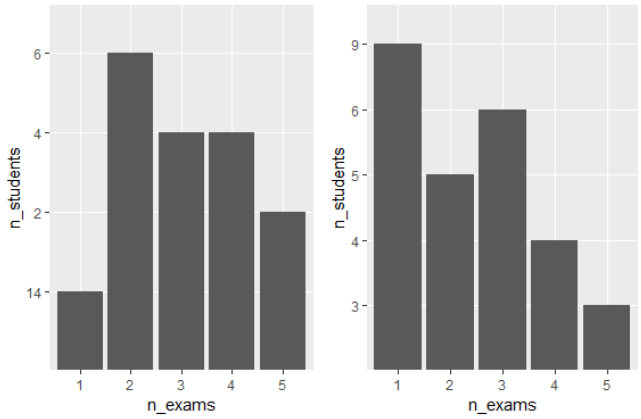
$$x_1 = \frac{0+12}{2} = 6$$
$$x_2 = \frac{12+30}{2} = 21$$
$$x_3 = \frac{30+50}{2} = 40$$
$$x_4 = \frac{50+100}{2} = 75$$

[d]

Il grafico corretto è l'istogramma (con ampiezza diversa per le varie classi):



Esercizio 7 [a] Il grafico piú opportuno é quello a barre:



[b]

	SP	EC	n_i	N_i
1	14	9	23	23
2	6	5	11	34
3	4	6	10	44
4	4	4	8	52
5	2	3	5	57

$M_e = \frac{N+1}{2} = \frac{58}{2} = X_{29} = 2$

$Q_1 = \frac{N+1}{4} \cdot 1 = \frac{57+1}{4} \cdot 1 = X_{14} = 1$

$Q_3 = \frac{N+1}{4} \cdot 3 = \frac{57+1}{4} \cdot 3 = X_{43} = 3$

[c]

	SP	EC	n_i	N_i	F_i
1	14	9	23	23	$23/57 = 0,4$
2	6	5	11	34	$34/57 = 0,6$
3	4	6	10	44	$44/57 = 0,8$
4	4	4	8	52	$52/57 = 0,9$
5	2	3	5	57	$57/57 = 1$

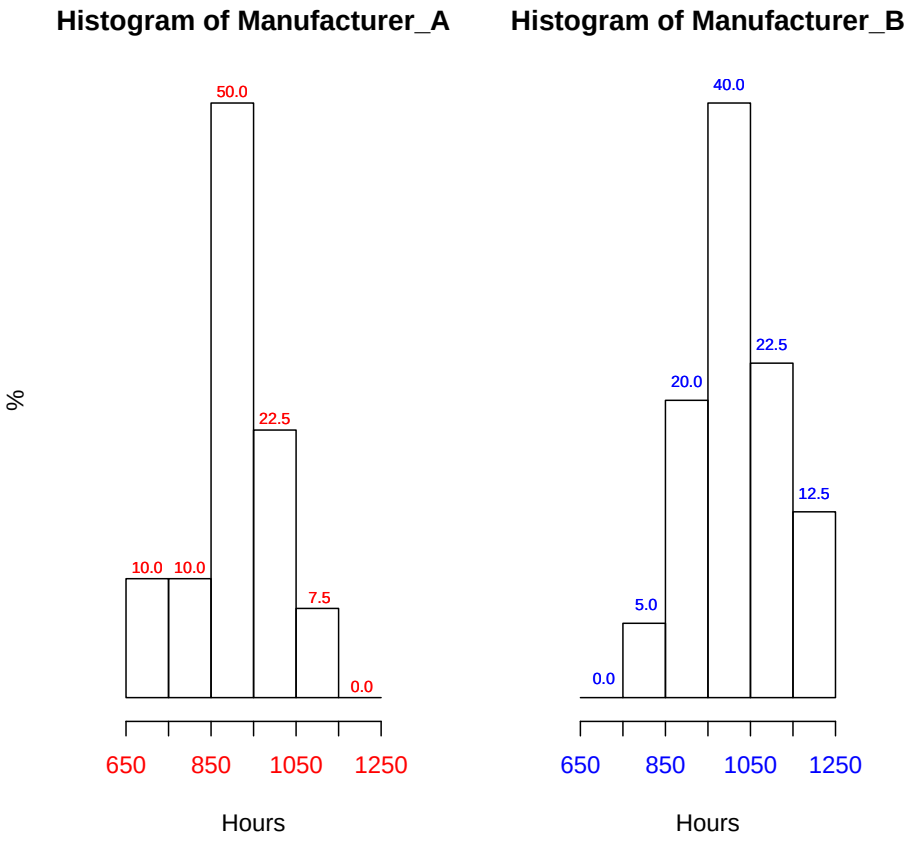
Esercizio 8 [a]

Scegliamo lo stesso numero di classi e la stessa ampiezza per entrambe le distribuzioni condizionate all'impresa. In particolare scegliamo un'ampiezza di 100 (ore).

Qui mostriamo solo la distribuzione percentuale (equivalenti alle frequenze relative).

Durata (in ore)	% Azienda A	% Azienda B
$[650, 750)$	10.0	0.0
$[750, 850)$	10.0	5.0
$[850, 950)$	50.0	20.0
$[950, 1050)$	22.5	40.0
$[1050, 1150)$	7.5	22.5
$[1150, 1250)$	0.0	12.5

[b]



[c]

La distribuzione della durata delle lampadine dell'azienda A sembra più concentrata verso durate inferiori rispetto a quella dell'azienda B (ad esempio l'azienda B non ha nessuna lampadina che dura tra le 650 e le 750 ore, a fronte di un 10% di lampadine in quella classe per l'azienda A)

[d]

Il confronto è ancor più evidente usando le frequenze cumulate

Durata (in ore)	% Azienda A	% Azienda B
$[650, 750)$	10.0	0.0
$[750, 850)$	20.0	5.0
$[850, 950)$	70.0	25.0
$[950, 1050)$	92.5	65.0
$[1050, 1150)$	100.0	87.5
$[1150, 1250)$	100.0	100.0

... che possono anche essere rappresentate graficamente (volendo)

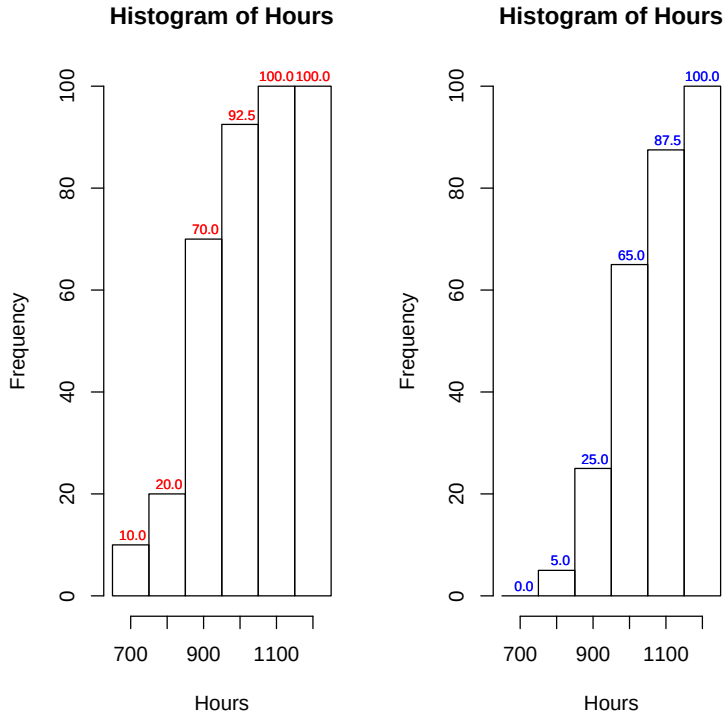


Figura 1: Istogrammi delle distribuzione percentuali cumulata della durata delle lampadine dell'azienda A (a sinistra) e dell'azienda B (a destra).

Dall'analisi delle frequenze (percentuali) cumulate, sembra che l'azienda B produca lampadine migliori dell'azienda A. Infatti la percentuale cumulata per l'azienda B mostra infatti che il 65% delle sue lampadine durano meno di 1050 ore, mentre il 70% di lampadine dell'azienda A, durano meno di 950 ore. Inoltre nessuna lampadina dell'azienda A dura pi ù di 1150 ore, invece il 12.5% di quelle dell'azienda B durano tra le 1150 e le 1250 ore (provate anche a calcolare qualche indice per verificare se danno la stessa informazione)