Corso di Statistica per Ingegneria Fisica
Anno accademico 2020/2021

Esercitazione 1: Statistica Descrittiva

Esercizio 1. I gruppi sanguigni di 12 persone sono

$$B, B, AB, 0, A, 0, A, A, A, B, A, A$$

Si costruisca la tabella di distribuzione di frequenza ed il relativo diagramma a barre.

Esercizio 2. Un certo macchinario produce lotti da 100 pezzi ciascuno. I seguenti dati riportano il numero di pezzi difettosi presenti in 25 lotti ispezionati:

$$1, 5, 3, 1, 3, 2, 2, 1, 2, 5, 3, 0, 1, 4, 3, 7, 1, 3, 1, 7, 2, 1, 2, 4, 8$$

' Si costruiscano

(a) una tabella di distribuzione di frequenza,

(b) il relativo istogramma,

(c) un boxplot.

Si determinino inoltre:

(a) media,

(b) quartili e 80-esimo percentile,

(c) mode,

(d) varianza,

(e) deviazione standard,

(f) differenza interquartile,

(g) range dei dati osservati.

Esercizio 3. I seguenti dati si riferiscono alla temperatura (in $°C$) massima giornaliera raggiunta da un certo apparecchio nell'ultimo mese:

14,7	12,3	18,3	10,2	11,5	15,1	14,2	14,7	13,8	17,3
25,3	26,4	31	19,4	17,5	17,6	16,8	18	13,8	10,7
12,6	14,5	17,8	19,6	17,2	13,1	13,9	14,2	13,7	18,1

Si costruisca una tabella di distribuzione di frequenza e si disegni il relativo istogramma.

Si calcolino media, mediana e deviazione standard per i dati assegnati e per le stesse temperature espresse in $°F$ (*si ricorda che se t ed f rappresentano la medesima temperatura misurata in gradi Celsius e Fahrenheit rispettivamente, allora $f = \frac{9}{5}t + 32$*) *[$\bar{y} = 61.60°F$, $s_y = 8.27°F$, $q_2 = 58.82°F$].*

Esercizio 4. Negli ultimi 5 anni un certo macchinario ha necessitato di 11 interventi di manutenzione. Le cause del malfunzionamento sono state:

- arresto del macchinario per un guasto meccanico (5 volte)
- arresto del macchinario per un guasto del sistema di controllo (4 volte)
- produzione di un numero eccessivo di pezzi difettosi (2 volte)

(a) Costruire la tabella della distribuzione di frequenza

(b) Rappresentare i dati graficamente.

Esercizio 5. In un'azienda il numero di dipendenti maschi e femmine è ripartito per età nel modo seguente:

Età	Femmine	Maschi
$21-30$	220	284
$31-40$	280	427
$41-50$	295	388
$51-55$	104	146
$56-60$	31	125

(a) Quante e quali variabili stiamo considerando in questo set di dati? Che tipo di variabili sono? *[Stiamo considerando due variabili: "Età" e "Sesso". "Età" è una variabile quantitativa e "Sesso" è qualitativa.]*

(b) Vogliamo studiare la distribuzione dell'età dei dipendenti per entrambe le categorie della variabile "Sesso". Costruire la tabella della distribuzione di frequenza dell'età per i dipendenti maschi e femmine (includendo: classi, frequenza assoluta, frequenza relativa, densità e frequenza cumulata).

(c) Costruire l'istogramma delle frequenza relative per le due categorie di dipendenti e commentare il risultato.

Esercizio 6. Un operatore misura il pH di una soluzione per otto volte, utilizzando lo stesso strumento, e ottiene i seguenti dati:

$$7.15 \quad 7.20 \quad 7.18 \quad 7.19 \quad 7.21 \quad 7.20 \quad 7.17 \quad 6.50\,.$$

(a) Calcolare media e varianza campionaria, mediana e IQR. *[$\bar{x} = 7.1$, $s^2 = 0.05914$, $m = 7.185$, $IQR = 0.04$]*

(b) Ripetere il calcolo, escludendo l'ultima misurazione. Commentare i risultati ottenuti.*[$\bar{x}$ = 7.1857, s^2 = 0.00043, m = 7.19, IQR = 0.03. Quantili e IQR sono indici di posizione e dispersione più robusti rispetto a media e varianza campionaria.]*

(c) Disegnare il boxplot sia dei dati completi, sia di quelli in cui avete escluso l'ultima misurazione.

Esercizio 7. Questi dati rappresentano la concentrazione di ferro presente in dieci campioni di sedimento del Lago di Como misurata in $\mu g/l$:

$$5.0,\ 1.5,\ 3.3,\ 5.1,\ 1.8,\ 3.2,\ 3.4,\ 3.4,\ 5.0,\ 2.5$$

Si calcolino: media , varianza, deviazione standard, mediana, quartili e ottantesimo percentile. Se m indica la mediana dei dati, per quali valori di k l'intervallo $[m-k, m+k]$ contiene esattamente il 60% dei dati? *[$\bar{x} = 3.42$, $s^2 = 1.67$, $s = 1.293$, $Q_1 = 2.5$, $Q_2 = 3.35$, $Q_3 = 5$, $q_{0.8} = 5$, $1.55 \leq k < 1.65$]*

Esercizio 8. In una classe di 50 studenti le età dei ragazzi sono così distribuite:

Età	20	21	22	23	24	25	26	27
N.studenti	12	15	6	5	5	3	3	1

Determinare l'età media degli studenti, la deviazione standard dell'età, la mediana e la moda. Si stabilisca inoltre, senza rifare i calcoli, quali dei precedenti indici cambierebbero se avessimo 1 studente di 26 anni e 3 di 27. *[$\bar{x} = 22.040$, $s = 1.958$, $Q_2 = 21$, $moda = 21$. Gli indici che cambierebbero sarebbero media e deviazione standard.]*

Esercizio 9. La seguente tabella sintetizza i dati raccolti nel 1798 dal fisico Henry Cavendish, relativi alla misura della densità della terra espressa come multiplo della densità dell'acqua.

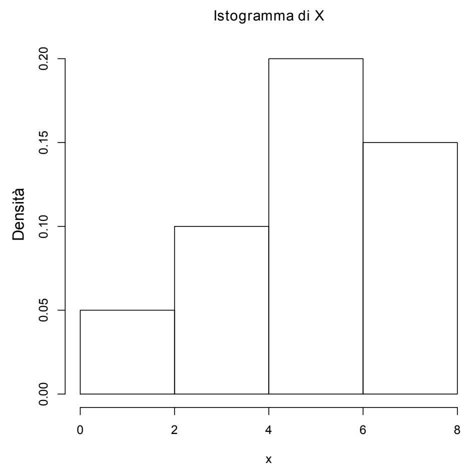
Classi	**Freq. Assoluta**
(4,4.25]	1
(4.25,4.5]	0
(4.5,4.75]	0
(4.75,5]	1
(5,5.25]	1
(5.25,5.5]	14
(5.5,5.75]	9
(5.75,6]	3

(a) Si rappresenti tramite istogramma la distribuzione dei dati raccolti.

(b) Si descriva, sulla base dell'istogramma, la forma della distribuzione dei dati riguardanti la densità della terra. *[Distribuzione unimodale, asimmetria con coda a sinistra (possibile outlier).]*

(c) Si calcoli, in base alla tabella riportata, la media delle misurazioni della densità terrestre. *[*$\overline{x} = 5.435$.*]*

Esercizio 10. L'istogramma seguente rappresenta la distribuzione di frequenza di una variabile X.

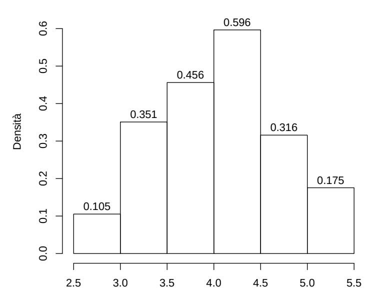


(a) Si costruisca la tabella di distribuzione di frequenza di X.

(b) Si determinino le classi contenenti i quartili di X. *[*$Q_1 \in [2,4)$, $Q_2 \in [4,6)$, $Q_3 \in [6,8)$*]*

(c) Determinare una stima della media, del primo e terzo quartile. *[*$\overline{x} \simeq 4.8$, $Q_1 \simeq 3.5$, $Q_3 \simeq 6.33$.*]*

Esercizio 11. Il seguente istogramma rappresenta la distribuzione di frequenza del volume X della forza di espirazione (FEV: Forced Expiratory Volume), misurata in litri per 57 studenti di medicina. In ordinata è riportata la densità delle classi, ovvero il rapporto fra la loro frequenza relativa e la loro ampiezza.

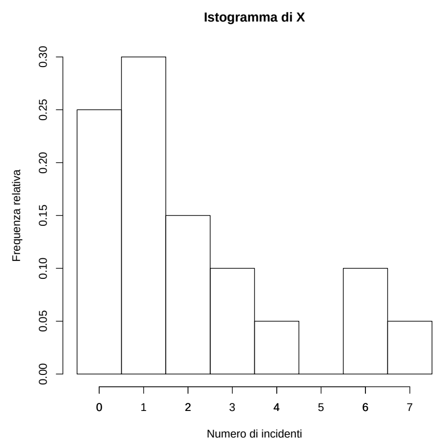


(a) Calcolare la percentuale di studenti la cui FEV è inferiore a 3.5 litri *[22.8%]*.

(b) Stabilire a quale classe appartiene la mediana di X *[*$4 \leq m \leq 4.5$*]*.

(c) Sulla base dell'istogramma, dire quale relazione ci si aspetta fra la media e la mediana di X. *L'istogramma non presenta particolari asimmetrie, quindi mi aspetto che media e varianza siano simili.*

Esercizio 12. Per ognuno dei 180 autisti di autobus di una azienda di trasporti municipale, è stato osservato il numero di incidenti X compiuti durante l'anno 2000. I risultati di questa indagine sono riassunti nel seguente istogramma:



Si calcolino:

(a) la moda, la media e la mediana di X; *[*$moda = 1$, $\overline{x} = 2.05$, $m = 1$*]*

(b) la deviazione standard di X; *[*$s = 2.1148$*]*

(c) il primo e il terzo quartile di X; *[*$Q_1 = 0.5$, $Q_3 = 3$*]*

(d) l'ottantacinquesimo percentile di X. *[*$q_{0.85} = 5$*]*

(e) Quanti autisti hanno fatto un numero di incidenti inferiori alla media? *[126]*

Soluzione 6.

(a) Abbiamo

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{7.15 + 7.20 + \ldots + 6.50}{8} = 7.1$$
$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{8-1}\left(7.15^2 + 7.20^2 + \ldots + 6.50^2 - 8\cdot 7.1^2\right) = 0.05914$$

dove per calcolare la varianza campionaria abbiamo usato la formula alternativa.
Per determinare invece Q1, Q2 e Q3, dobbiamo innanzitutto scrivere i dati ordinati $x_{(1)}, x_{(2)}, \ldots, x_{(8)}$ come segue

$$6.50 \quad 7.15 \quad 7.17 \quad 7.18 \quad 7.19 \quad 7.20 \quad 7.20 \quad 7.21\,.$$

Usando la definizione

$$q_\gamma = \begin{cases} x_{(\lfloor n\gamma\rfloor+1)} & \text{se } n\gamma \notin \mathbb{N} \\ \frac{1}{2}\left(x_{(n\gamma)} + x_{(n\gamma+1)}\right) & \text{se } n\gamma \in \mathbb{N} \end{cases}$$

ricaviamo

$$\mathrm{Q1} = q_{0.25} = \frac{1}{2}\left(x_{(8\cdot 0.25)} + x_{(8\cdot 0.25+1)}\right) = \frac{1}{2}\left(x_{(2)} + x_{(3)}\right) = \frac{1}{2}(7.15 + 7.17) = 7.16$$
$$\mathrm{Q2} = q_{0.50} = \frac{1}{2}\left(x_{(8\cdot 0.50)} + x_{(8\cdot 0.50+1)}\right) = \frac{1}{2}\left(x_{(4)} + x_{(5)}\right) = \frac{1}{2}(7.18 + 7.19) = 7.185$$
$$\mathrm{Q3} = q_{0.75} = \frac{1}{2}\left(x_{(8\cdot 0.75)} + x_{(8\cdot 0.75+1)}\right) = \frac{1}{2}\left(x_{(6)} + x_{(7)}\right) = \frac{1}{2}(7.20 + 7.20) = 7.20$$

perché in tutti e tre casi $n\gamma \in \mathbb{N}$. Di conseguenza, la mediana è $m = \mathrm{Q2} = 7.185$ e il range interquartilico è $\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1} = 7.20 - 7.16 = 0.04$.

(b) Escludendo l'ultima misurazione $x_8 = 6.50$, restano i dati ordinati

$$7.15 \quad 7.17 \quad 7.18 \quad 7.19 \quad 7.20 \quad 7.20 \quad 7.21\,.$$

Gli stessi calcoli di prima per i 7 dati rimasti danno

$$\bar{x} = 7.1857 \qquad s^2 = 0.00043 \qquad \mathrm{Q1} = q_{0.25} = x_{(\lfloor 7\cdot 0.25\rfloor+1)} = x_{(2)} = 7.17$$
$$\mathrm{Q2} = q_{0.50} = x_{(\lfloor 7\cdot 0.50\rfloor+1)} = x_{(4)} = 7.19 \qquad \mathrm{Q3} = q_{0.75} = x_{(\lfloor 7\cdot 0.75\rfloor+1)} = x_{(6)} = 7.20\,.$$

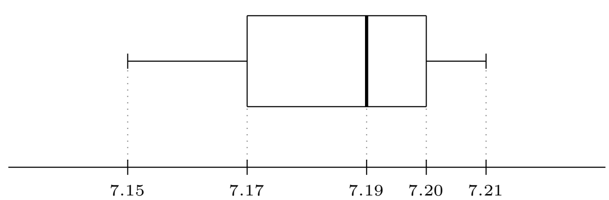
Questa volta per trovare $q_{0.25}$, $q_{0.50}$ e $q_{0.75}$ abbiamo usato la formula per $n\gamma \notin \mathbb{N}$. La mediana è dunque $m = \mathrm{Q2} = 7.19$ e il range interquartilico è $\mathrm{IQR} = 7.20 - 7.17 = 0.03$. Vediamo che, rimuovendo il dato estremo $x_8 = 6.50$, la media e la varianza campionarie si sono modificate sensibilmente (rispettivamente, $7.1 \to 7.1857$ e $0.05914 \to 0.00043$), mentre i quartili e l'IQR sono cambiati di poco (rispettivamente, $7.16 \to 7.17$, $7.185 \to 7.19$, $7.20 \to 7.20$ e $0.04 \to 0.03$). Questo è in accordo col fatto che la media e la varianza campionarie sono molto più sensibili ai dati sulle code di quanto lo siano i quartili. Inoltre, benché entrambi i dataset abbiano una coda a sinistra, evidenziata dal fatto che $\bar{x} < m$ in tutti e due, tale coda è molto più pronunciata nel dataset completo, in cui $\bar{x} - m = -0.0850$ contro $\bar{x} - m = -0.0043$ nel dataset ridotto.

(c) Il boxplot del dataset completo è

in cui:

- il baffo di sinistra si estende da Q1 $= 7.16$ al primo dato compreso tra Q1 $- 1.5 \cdot$ IQR $= 7.16 - 1.5 \cdot 0.04 = 7.10$ e Q1 $= 7.16$, cioè $x_{(2)} = 7.15$;
- il baffo di destra si estende da Q3 $= 7.20$ all'ultimo dato compreso tra Q3 $= 7.20$ e Q3 $+ 1.5 \cdot$ IQR $= 7.20 + 1.5 \cdot 0.04 = 7.26$, cioè $x_{(8)} = 7.21$;
- sono outlier tutti i dati fuori dall'intervallo [Q1 $- 1.5 \cdot$ IQR , Q3 $+ 1.5 \cdot$ IQR] $= [7.10\,,\ 7.26]$, cioè $x_{(1)} = 6.50$.

Il boxplot del dataset senza l'outlier $x_8 = 6.50$ è



in cui:

- il baffo di sinistra si estende da Q1 $= 7.17$ al primo dato compreso tra Q1 $- 1.5 \cdot$ IQR $= 7.17 - 1.5 \cdot 0.03 = 7.125$ e Q1 $= 7.17$, cioè $x_{(1)} = 7.15$;
- il baffo di destra si estende da Q3 $= 7.20$ all'ultimo dato compreso tra Q3 $= 7.20$ e Q3 $+ 1.5 \cdot$ IQR $= 7.20 + 1.5 \cdot 0.03 = 7.245$, cioè $x_{(7)} = 7.21$;
- non ci sono outlier fuori dall'intervallo [Q1 $- 1.5 \cdot$ IQR , Q3 $+ 1.5 \cdot$ IQR] $= [7.125\,,\ 7.245]$.

Si osserva che in entrambi i boxplot (soprattutto in quello con l'outlier) è evidente la coda a sinistra che abbiamo già dedotto dalla disuguaglianza $\bar{x} < m$.

Soluzione 12. Osserviamo innanzitutto che a rigore l'istogramma non sarebbe disegnato correttamente: in ordinata, infatti, non va mai messa la frequenza relativa FR, bensì piuttosto la densità $\frac{\text{FR}}{\text{ampiezza}}$. Tuttavia, questo piccolo abuso è giustificato dal fatto che, essendo i dati discreti, si è scelto ampiezza $= 1$ per ogni classe, e dunque in questo caso la densità coincide con FR. Nell'istogramma leggiamo le frequenze relative di ogni classe e da queste ricaviamo tutte le altre coi semplici passaggi

$$\text{FA}(k) = n \cdot \text{FR}(k)\,, \qquad \text{FC}(k) = \text{FR}(k) + \text{FC}(k-1)\,.$$

Nella tabella seguente, le colonne sono scritte nell'ordine in cui sono state via via ottenute:

Classi	FR	FA	FC
0	0.25	45	0.25
1	0.30	54	0.55
2	0.15	27	0.70
3	0.10	18	0.80
4	0.05	9	0.85
5	0	0	0.85
6	0.10	18	0.95
7	0.05	9	1.00

(a) La moda è il valore nella classe più frequente (eventualmente più valori, se le classi più frequenti sono in numero maggiore di 1). Perciò, vediamo dalla tabella che moda $= 1$. Per dati discreti, la media campionaria si può trovare in modo esatto dalla formula

$$\bar{x} = \sum_{\text{classi } k} k \cdot \text{FR}(k) = 0 \cdot 0.25 + 1 \cdot 0.30 + \ldots + 7 \cdot 0.05 = 2.05\,.$$

Anche la mediana si può trovare in modo esatto dalla tabella:

$$\begin{cases} \text{FC}(0) = 0.25 & \Rightarrow \quad \text{almeno il } (100-25)\% = 75\% \geq 50\% \text{ dei dati è} \geq 1 \\ \text{FC}(1) = 0.55 & \Rightarrow \quad \text{almeno il } 55\% \geq 50\% \text{ dei dati è} \leq 1 \end{cases}$$

$$\Rightarrow \quad m = q_{0.50} = 1\,.$$

(b) Per calcolare la varianza campionaria e da questa la deviazione standard campionaria, usiamo la formula alternativa per dati discreti:

$$\begin{aligned} s^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2\right) = \frac{1}{n-1}\left(\sum_{\text{classi } k} k^2 \cdot \text{FA}(k) - n \cdot \bar{x}^2\right) \\ &= \frac{1}{n-1}\left(0^2 \cdot 45 + 1^2 \cdot 54 + \ldots + 7^2 \cdot 9 - 180 \cdot 2.05^2\right) = 4.4723 \\ s &= \sqrt{s^2} = \sqrt{4.4723} = 2.1148\,. \end{aligned}$$

(c) Il terzo quartile si calcola in modo simile alla mediana:

$$\begin{cases} \text{FC}(2) = 0.70 < 75\% \\ \text{FC}(3) = 0.80 > 75\% \end{cases} \qquad \Rightarrow \quad \text{Q3} = q_{0.75} = 3\,.$$

Per il primo quartile $\mathrm{Q1} = q_{0.25}$, osserviamo invece che $\mathrm{FC}(0) = 0.25$ esattamente, e dunque per calcolarlo occorre fare alcune considerazioni più dettagliate. Infatti, trattandosi sempre di dati discreti, *esattamente* il 25% dei dati è ≤ 0 e il $(100 - 0)\% = 100\%$ è ≥ 0 (cioè tutti). Dunque 0 *potrebbe* essere il quantile $q_{0.25}$ (almeno il 25% dei dati a sinistra e almeno il 75% a destra). Allo stesso modo, però, *esattamente* il 55% dei dati è ≤ 1 e il $(100-25)\% = 75\%$ è ≥ 1. Dunque anche 1 *potrebbe* essere il quantile $q_{0.25}$.
Per capire se $q_{0.25}$ è 0 oppure 1 oppure cade a metà fra questi due valori, occorre risalire ai dati ordinati originari. Siccome abbiamo a che fare con dati discreti, ciò è sempre possibile *senza approssimazioni* a partire dalla tabella delle frequenze.
Poiché $\mathrm{FA}(0) = 45$ e $\mathrm{FA}(1) = 54$, abbiamo

$$x_{(1)} = x_{(2)} = \ldots = x_{(45)} = 0\,, \qquad x_{(46)} = x_{(47)} = \ldots = x_{(45+54)} = x_{(99)} = 1\,.$$

Di conseguenza, usando la definizione

$$q_\gamma = \begin{cases} x_{(\lfloor n\gamma \rfloor + 1)} & \text{se } n\gamma \notin \mathbb{N} \\ \frac{1}{2}\left(x_{(n\gamma)} + x_{(n\gamma+1)}\right) & \text{se } n\gamma \in \mathbb{N} \end{cases},$$

poiché siamo nel caso $n\gamma = 180 \cdot 0.25 = 45 \in \mathbb{N}$, otteniamo

$$\mathrm{Q1} = q_{0.25} = \frac{1}{2}\left(x_{(n\gamma)} + x_{(n\gamma+1)}\right) = \frac{1}{2}\left(x_{(45)} + x_{(46)}\right) = \frac{1}{2}(0+1) = 0.5\,.$$

(d) Vogliamo calcolare $q_{0.85}$, e anche in questo caso vediamo dalla tabella che per ben due classi abbiamo $\mathrm{FC} = 0.85$, e cioè per le classi $\{4\}$ e $\{5\}$. Come nel punto precedente, andiamo allora a vedere in dettaglio i dati ordinati in queste due classi e in quelle immediatamente contigue:

$$\begin{cases} n \cdot \mathrm{FC}(3) = 180 \cdot 0.80 = 144 \\ n \cdot \mathrm{FC}(4) = 180 \cdot 0.85 = 153 \end{cases} \qquad \Rightarrow \qquad x_{(144+1)} = x_{(144+2)} = \ldots = x_{(153)} = 4$$

$$\begin{cases} n \cdot \mathrm{FC}(5) = 180 \cdot 0.85 = 153 \\ n \cdot \mathrm{FC}(6) = 180 \cdot 0.95 = 171 \end{cases} \qquad \Rightarrow \qquad x_{(153+1)} = x_{(153+2)} = \ldots = x_{(171)} = 6$$

(notare che $\mathrm{FR}(5) = 0 \;\Rightarrow$ nessun dato prende il valore 5). Usando sempre la definizione di q_γ come nel punto precedente, osserviamo che anche in questo caso $n\gamma = 180 \cdot 0.85 = 153 \in \mathbb{N}$, e quindi

$$q_{0.85} = \frac{1}{2}\left(x_{(n\gamma)} + x_{(n\gamma+1)}\right) = \frac{1}{2}\left(x_{(153)} + x_{(154)}\right) = \frac{1}{2}(4+6) = 5\,.$$

(e) Il numero di autisti che ha fatto un numero di incidenti inferiore alla media campionaria $\bar{x} = 2.05$ è

$$\begin{aligned} \#\{i \mid x_i < 2.05\} &= \#\{i \mid x_i \leq 2\} && \text{(perché si tratta di dati discreti)} \\ &= n \cdot \mathrm{FC}(2) = 180 \cdot 0.70 = 126\,. \end{aligned}$$