


Probabilità e Statistica – Corso di Laurea in Informatica
A.A. 2020/2021

ESERCITAZIONE 11

E11.1 . L'esercizio può essere svolto completamente con carta, penna, calcolatrice e molta pazienza oppure con l'ausilio del computer. Un file di testo, chiamato **dati.txt** e contenente i dati elencati sotto, è disponibile su Moodle, nella cartella relativa a questa esercitazione.

Alla prima prova parziale dell'insegnamento di *Probabilità e Statistica* per il Corso di Laurea in Informatica, A.A. 2018/2019, hanno partecipato 104 studenti. I voti ottenuti sono stati:

28	27	30	29	26	5	21	22	15	23	30	29	11
28	19	19	18	27	29	25	28	30	28	18	27	26
13	33	21	33	29	26	22	29	27	28	25	22	22
26	14	24	33	30	16	33	23	23	33	19	21	19
27	27	29	29	26	21	23	19	25	25	13	24	30
18	18	26	23	30	28	21	30	26	21	26	21	30
16	23	33	6	15	15	25	30	28	26	28	25	16
22	29	30	33	33	33	24	30	27	33	14	26	27

- (a) Derivare la distribuzione delle frequenze assolute, relative e relative cumulative per i dati raggruppati in classi di ampiezza 1, 3 e 5.
- (b) Nei tre casi di cui sopra, disegnare gli istogrammi delle frequenze relative (scegliendo le altezze in modo che siano le aree dei rettangoli ad essere pari alle frequenze) e i diagrammi delle frequenze relative cumulative.
- (c) Determinare media campionaria, mediana, moda e varianza campionaria dei dati.
- (d) Determinare la differenza interquartile e costruire il boxplot.
- (e) Determinare il 65° percentile.

Soluzione. Risposta (a). Otteniamo le seguenti distribuzioni delle frequenze.

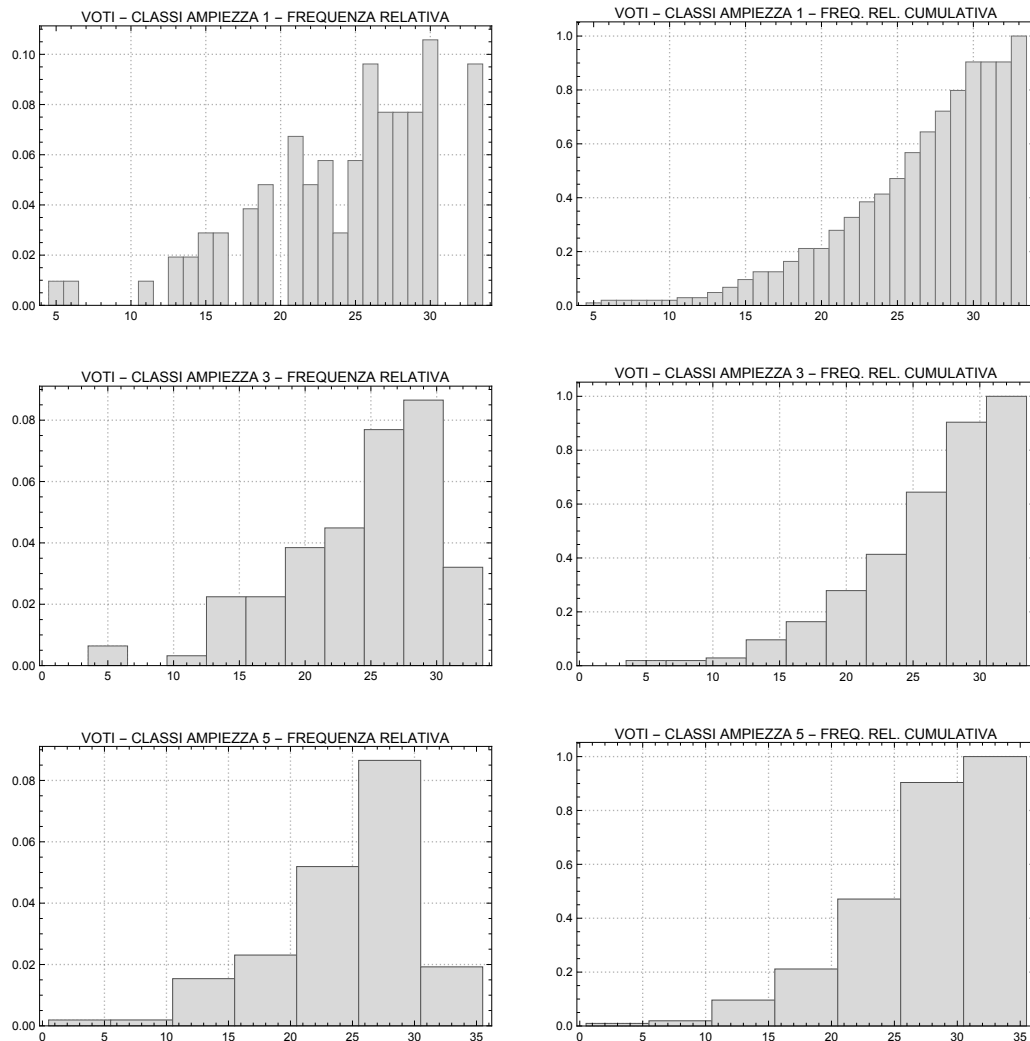
Classi ampiezza 1			
VOTO	f_a	f_r	F_r
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	1	1/104	1/104
6	1	1/104	1/52
7	0	0	1/52
8	0	0	1/52
9	0	0	1/52
10	0	0	1/52
11	1	1/104	3/104
12	0	0	3/104
13	2	1/52	5/104
14	2	1/52	7/104
15	3	3/104	5/52
16	3	3/104	13/104
17	0	0	13/104
18	4	1/26	17/104
19	5	5/104	11/52
20	0	0	11/52
21	7	7/104	29/104
22	5	5/104	17/52
23	6	3/52	5/13
24	3	3/104	43/104
25	6	3/52	49/104
26	10	5/52	59/104
27	8	1/13	67/104
28	8	1/13	75/104
29	8	1/13	83/104
30	11	11/104	94/104
31	0	0	94/104
32	0	0	94/104
33	10	5/52	104/104
Somma	104	1	

Classi ampiezza 3			
CLASSE VOTO	f_a	f_r	F_r
1-3	0	0	0
4-6	2	1/52	1/52
7-9	0	0	1/52
10-12	1	1/104	3/104
13-15	7	7/104	5/52
16-18	7	7/104	17/104
19-21	12	3/26	29/104
22-24	14	7/52	43/104
25-27	24	3/13	67/104
28-30	27	27/104	47/52
31-33	10	5/52	52/52
Somma	104	1	

Classi ampiezza 5			
CLASSE VOTO	f_a	f_r	F_r
1-5	1	1/104	1/104
6-10	1	1/104	1/52
11-15	8	1/13	5/52
16-20	12	3/26	11/52
21-25	27	27/104	49/104
26-30	45	45/104	94/104
31-35	10	5/52	104/104
Somma	104	1	

▲ Nelle tabelle precedenti, le frequenze relative sono state lasciate in forma di frazione perché così scritte sono esatte. Se si scrivono in forma decimale e si eseguono degli arrotondamenti è importante controllare che sommino ad uno.

Risposta (b). Otteniamo i seguenti istogrammi.



Risposta (c). La media campionaria risulta $\bar{x} \approx 24.47$.

La mediana è il secondo quartile, cioè $M = Q_2 = q_{0.5}$. Poiché $(104)(0.5) = 52$ è intero, la mediana sarà la media aritmetica dei dati in posizione 52 e 53. Dalla distribuzione delle frequenze per classi di ampiezza unitaria possiamo dedurre che $x_{52} = x_{53} = 26$. Quindi $M = 26$.

La moda è 30 (!), voto con frequenza massima.

La varianza risulta $s^2 \approx 36.1$.

Risposta (d). La differenza interquartile è data da $IQR = Q_3 - Q_1$. Calcoliamo il primo e terzo quartile.

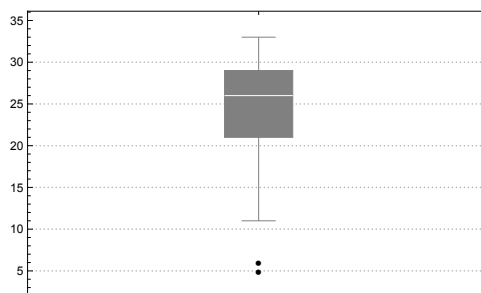
- Il primo quartile è $Q_1 = q_{0.25}$. Poiché $(104)(0.25) = 26$ è intero, il primo quartile sarà la media aritmetica dei dati in posizione 26 e 27. Dalla distribuzione delle frequenze per classi di ampiezza unitaria possiamo dedurre che $x_{26} = x_{27} = 21$. Quindi $Q_1 = 21$.
- Il terzo quartile è $Q_3 = q_{0.75}$. Poiché $(104)(0.75) = 78$ è intero, il terzo quartile sarà la media aritmetica dei dati in posizione 78 e 79. Dalla distribuzione delle frequenze per classi di ampiezza unitaria possiamo dedurre che $x_{78} = x_{79} = 29$. Quindi $Q_3 = 29$.

Pertanto, otteniamo $IQR = 29 - 21 = 8$.

Determiniamo ora i limiti inferiore e superiore per capire quanto estendere i baffi del boxplot ed individuare eventuali outliers. Si ha

$$L = Q_1 - 1.5IQR = 21 - (1.5)8 = 9 \quad \text{e} \quad U = Q_3 + 1.5IQR = 29 + (1.5)8 = 41$$

e quindi il baffo inferiore va esteso scendendo dal primo quartile fino a 11, mentre quello superiore si estende salendo fino a 33. I voti 5 e 6 sono da considerarsi outliers.



Risposta (e). Il 65° percentile è $q_{0.65}$. Poiché $(104)(0.65) = 67.6$ non è intero e $\lfloor 67.6 \rfloor = 67$, il 65° percentile sarà il dato in posizione $67 + 1 = 68$. Dalla distribuzione delle frequenze per classi di ampiezza unitaria possiamo dedurre che $x_{68} = 28$. Quindi $q_{0.65} = 28$.

E11.2. Consideriamo il seguente campione, proveniente da una distribuzione gaussiana di parametri μ e σ^2 (entrambi incogniti):

0.39 0.68 0.82 1.35 1.38 1.62 1.70 1.71 1.85 2.14 2.89 3.69.

- Calcolare media e varianza campionaria.
- Stimare il valore dei parametri μ e σ in base al campione dato.
- Raggruppare i dati grezzi nelle 4 classi: $[0, 1)$, $[1, 2)$, $[2, 3)$ e $[3, 4)$. Calcolare frequenze assolute e relative di queste classi.
- Calcolare la probabilità che una variabile aleatoria gaussiana X con i parametri stimati al punto (b) appartenga a ciascuna delle classi costruite al punto (c) e confrontare con le frequenze relative. C'è un buon adattamento dei dati empirici al modello teorico?

Soluzione.

- (a) Cominciamo col calcolare la media del campione. Abbiamo

$$\bar{x} = \frac{1}{12}(0.39 + 0.68 + 0.82 + 1.35 + 1.38 + 1.62 + 1.70 + 1.71 + 1.85 + 2.14 + 2.89 + 3.69) = 1.685.$$

Ora passiamo alla varianza. Per facilitare le operazioni, conviene costruire la tabella

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
0.39	-1.295	1.6770
0.68	-1.005	1.0100
0.82	-0.865	0.7482
1.35	-0.335	0.1122
1.38	-0.305	0.0930
1.62	-0.065	0.0042
1.70	0.015	0.0002
1.71	0.025	0.0006
1.85	0.165	0.0272
2.14	0.455	0.2070
2.89	1.205	1.4520
3.69	2.005	4.0200
Somma		9.3516

da cui si ricava $s^2 = \frac{1}{11}(9.3516) \approx 0.85$.

- (b) Contesto: stiamo considerando un modello statistico con densità gaussiana, di parametri (ignoti) μ e σ^2 , da cui estraiamo un campione casuale (X_1, \dots, X_{12}) di ampiezza 12. Per stimare il valore dei parametri, usiamo gli stimatori

- $T_1 = T_1(X_1, \dots, X_{12}) = \frac{1}{12} \sum_{i=1}^{12} X_i$ per il valor medio;
- $T_2 = T_2(X_1, \dots, X_{12}) = \frac{1}{11} \sum_{i=1}^{12} (X_i - \bar{X}_{12})^2$ per la varianza.

Otteniamo le stime valutando gli stimatori nel campione dato, quindi

$$\hat{\mu} = T_1(x_1, \dots, x_{12}) = \bar{x} = 1.685 \quad \text{e} \quad \hat{\sigma} = +\sqrt{T_2(x_1, \dots, x_{12})} = +\sqrt{s^2} \approx 0.922.$$

- (c) Otteniamo la seguente distribuzione delle frequenze:

CLASSE	f_a	f_r
[0, 1)	3	0.250
[1, 2)	6	0.500
[2, 3)	2	0.167
[3, 4)	1	0.083
Somma	12	1

- (d) In base all'analisi fatta in precedenza, abbiamo $X \sim N(1.685, 0.85)$. Calcoliamo la probabilità che X appartenga alle 4 classi individuate al punto precedente. Per $k \in \{1, 2, 3, 4\}$, si ha

$$P(k-1 \leq X < k) = P\left(\frac{k-2.685}{0.922} \leq Z < \frac{k-1.685}{0.922}\right) = \Phi\left(\frac{k-1.685}{0.922}\right) - \Phi\left(\frac{k-2.685}{0.922}\right),$$

dove $Z = \frac{X-1.685}{0.922} \sim N(0, 1)$. Ricordando la proprietà $\Phi(-z) = 1 - \Phi(z)$, calcoliamo

- per $k = 1$,

$$P(X \in [0, 1)) = \Phi(-0.74) - \Phi(-1.83) = \Phi(1.83) - \Phi(0.74) = 0.9664 - 0.7704 = 0.194;$$

- per $k = 2$,

$$P(X \in [1, 2)) = \Phi(0.34) - \Phi(-0.74) = \Phi(0.34) + \Phi(0.74) - 1 = 0.6331 + 0.7704 - 1 = 0.404;$$

- per $k = 3$,

$$P(X \in [2, 3)) = \Phi(1.43) - \Phi(0.34) = 0.9236 - 0.6331 = 0.291;$$

- per $k = 4$,

$$P(X \in [3, 4)) = \Phi(2.51) - \Phi(1.43) = 0.9940 - 0.9236 = 0.070.$$

Riassumiamo quanto appena trovato e quanto visto al punto precedente in tabella:

CLASSE	f_r	Probabilità
$[0, 1)$	0.250	0.194
$[1, 2)$	0.500	0.404
$[2, 3)$	0.167	0.291
$[3, 4)$	0.083	0.070
Somma	1	0.959

L'adattamento dei dati empirici al modello teorico è discreto.

E11.3. Stima puntuale della varianza (nel caso di media nota). Sia (X_1, \dots, X_n) un campione casuale estratto da una popolazione di densità f_{θ} , che ammette valor medio μ e varianza σ^2 . Supponiamo di conoscere μ (cosa abbastanza poco frequente!) e di voler stimare σ^2 . Mostrare che la statistica

$$T = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$$

è uno stimatore corretto di σ^2 .

Soluzione. Per mostrare che T è uno stimatore corretto della varianza dobbiamo mostrare che $E_{\theta}(T) = \sigma^2$. Calcoliamo

$$E_{\theta}(T) \stackrel{(\text{linearità})}{=} \frac{1}{n} \sum_{i=1}^n E_{\theta}[(X_i - \mu)^2] \stackrel{(\text{def. varianza})}{=} \frac{1}{n} \sum_{i=1}^n \text{Var}_{\theta}(X_i) \stackrel{(\text{equidistribuzione})}{=} \sigma^2,$$

da cui la conclusione.

▲ T è una statistica solo se il valore $\mu = E(X_i)$ è noto! Se μ è ignoto, T non è una statistica, perché dipende da uno dei parametri da stimare. Ricordiamo che, nel caso di popolazione con media e varianza ignoti, lo stimatore per la varianza è S_n^2 .

E11.4. Si consideri il modello statistico, dipendente dal parametro $\lambda > 0$, definito come

$$f_{\lambda}(x) = \begin{cases} c_{\lambda} \exp(-4\lambda x) & \text{se } x \in [0, +\infty) \\ 0 & \text{altrimenti.} \end{cases}$$

- Determinare il valore c_{λ} tale per cui f_{λ} risulta una densità.
- Calcolare il valor medio.
- Siano X_1, \dots, X_n variabili aleatorie i.i.d. con densità f_{λ} e sia

$$T = \frac{2X_1 + X_3 + 2X_4}{5}$$

uno stimatore del parametro λ . Lo stimatore T è corretto?

Soluzione.

- (a) Imponiamo la condizione di normalizzazione. Otteniamo

$$\int_{\mathbb{R}} f_{\lambda}(x) dx = 1 \Leftrightarrow c_{\lambda} \int_0^{+\infty} e^{-4\lambda x} dx = 1 \Leftrightarrow c_{\lambda} \left[-\frac{e^{-4\lambda x}}{4\lambda} \right]_0^{+\infty} = 1 \Leftrightarrow c_{\lambda} = 4\lambda$$

e quindi la densità risulta

$$f_{\lambda}(x) = \begin{cases} 4\lambda e^{-4\lambda x} & \text{se } x \in [0, +\infty) \\ 0 & \text{altrimenti.} \end{cases}$$

- (b) Poiché la densità f_{λ} è una densità esponenziale di parametro 4λ , il valor medio vale $\frac{1}{4\lambda}$.
 (c) T è uno stimatore corretto di λ se risulta $E_{\lambda}(T) = \lambda$. Poiché si ha

$$E_{\lambda}(T) \stackrel{(\text{linearità})}{=} \frac{2}{5}E_{\lambda}(X_1) + \frac{1}{5}E_{\lambda}(X_3) + \frac{2}{5}E_{\lambda}(X_4) \stackrel{(\text{equidistribuzione})}{=} \frac{1}{4\lambda},$$

la statistica T non è uno stimatore corretto di λ (lo è però del valor medio $\frac{1}{4\lambda}$).

E11.5 (📺 video). Siano X_1, \dots, X_n variabili aleatorie i.i.d. con densità $\text{Exp}(\lambda)$, con $\lambda > 0$. Sia, inoltre, $T_n = \frac{1}{n} \sum_{i=1}^n X_i$ uno stimatore per $\frac{1}{\lambda}$.

- (a) La statistica T_n è uno stimatore corretto di $\frac{1}{\lambda}$?
 (b) Si fissi $n = 1$ e si consideri la statistica $H = T_1(1 - aT_1)$, con $a \in \mathbb{R}$. Si determini il valore di a tale per cui H risulta uno stimatore corretto di $\frac{1}{\lambda}(1 - \frac{1}{\lambda})$.

Soluzione.

- (a) Sì, perché la media campionaria è sempre uno stimatore corretto (e consistente) del valor medio di un modello statistico.
 (b) Osserviamo che $T_1 \equiv X_1$ e calcoliamo il valor medio di H . Si ha

$$\begin{aligned} E_{\lambda}(H) &= E_{\lambda}[X_1(1 - aX_1)] \\ &= E_{\lambda}(X_1) - aE_{\lambda}(X_1^2) && (\text{linearità}) \\ &= E_{\lambda}(X_1) - a[Var_{\lambda}(X_1) + E_{\lambda}(X_1)^2] && (\text{dall'identità per la varianza}) \\ &= \frac{1}{\lambda} - a \left[\frac{1}{\lambda^2} + \left(\frac{1}{\lambda} \right)^2 \right] && (\text{poiché } X_1 \sim \text{Exp}(\lambda)) \\ &= \frac{1}{\lambda} \left(1 - \frac{2a}{\lambda} \right). \end{aligned}$$

La statistica H è uno stimatore corretto di $\frac{1}{\lambda}(1 - \frac{1}{\lambda})$ se risulta $E_{\lambda}(H) = \frac{1}{\lambda}(1 - \frac{1}{\lambda})$. Quindi per determinare a dobbiamo risolvere l'equazione

$$\frac{1}{\lambda} \left(1 - \frac{2a}{\lambda} \right) = \frac{1}{\lambda} \left(1 - \frac{1}{\lambda} \right),$$

da cui si ricava $a = \frac{1}{2}$.

E11.6 (📺 tratto da appello; 📺 video). Sia $0 < \theta \leq 2$. Consideriamo la funzione

$$f_{\theta}(x) = \begin{cases} c_{\theta} x + 1 - \frac{\theta}{2} & \text{se } 0 \leq x \leq 1 \\ 0 & \text{altrimenti.} \end{cases}$$

- (a) Determinare il valore c_θ per cui la funzione f_θ risulta una densità di probabilità.
- (b) Calcolare il valor medio.
- (c) Consideriamo il modello statistico $\{f_\theta | 0 < \theta \leq 2\}$. Sia (X_1, \dots, X_n) un campione casuale di ampiezza n e $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ la relativa media campionaria. Definiamo le statistiche

$$T_1 = \frac{1}{12} - 6\bar{X}_n \quad \text{e} \quad T_2 = 12\left(\bar{X}_n - \frac{1}{2}\right).$$

Dire se le statistiche T_1 e T_2 sono stimatori corretti del parametro θ .

- (d) Modificare gli eventuali stimatori distorti del punto precedente in modo che diventino stimatori corretti del parametro θ .

Soluzione.

- (a) Imponiamo la condizione di normalizzazione. Otteniamo

$$\int_{\mathbb{R}} f_\theta(x) dx = 1 \Leftrightarrow \int_0^1 (c_\theta x + 1 - \frac{\theta}{2}) dx = 1 \Leftrightarrow \left[\frac{c_\theta x^2}{2} + (1 - \frac{\theta}{2})x \right]_0^1 = 1 \Leftrightarrow c_\theta = \theta$$

e quindi la densità risulta

$$f_\theta(x) = \begin{cases} \theta x + 1 - \frac{\theta}{2} & \text{se } 0 \leq x \leq 1 \\ 0 & \text{altrimenti.} \end{cases}$$

- (b) Calcoliamo

$$\int_{\mathbb{R}} x f_\theta(x) dx = \int_0^1 [\theta x^2 + (1 - \frac{\theta}{2})x] dx = \left[\frac{\theta x^3}{3} + (1 - \frac{\theta}{2}) \frac{x^2}{2} \right]_0^1 = \frac{\theta}{12} + \frac{1}{2}.$$

- (c) Una statistica T è uno stimatore corretto del parametro θ se $E_\theta(T) = \theta$. Usando il fatto che $E_\theta(\bar{X}_n) = \frac{\theta}{12} + \frac{1}{2}$, calcoliamo

$$E_\theta(T_1) \stackrel{(\text{linearità})}{=} \frac{1}{12} - 6E_\theta(\bar{X}_n) = \frac{1}{12} - 6\left(\frac{\theta}{12} + \frac{1}{2}\right) = -\frac{35}{12} - \frac{\theta}{2}$$

e

$$E_\theta(T_2) \stackrel{(\text{linearità})}{=} 12[E_\theta(\bar{X}_n) - \frac{1}{2}] = 12\left(\frac{\theta}{12} + \frac{1}{2} - \frac{1}{2}\right) = \theta.$$

Pertanto, la statistica T_2 risulta essere uno stimatore corretto di θ , mentre T_1 no.

- (d) Dobbiamo correggere la statistica T_1 . Sappiamo che $E_\theta(T_1) = -\frac{35}{12} - \frac{\theta}{2}$. Da questa equazione ricaviamo θ , si ha

$$E_\theta(T_1) = -\frac{35}{12} - \frac{\theta}{2} \Leftrightarrow \theta = -2[E_\theta(T_1) + \frac{35}{12}] \stackrel{(\text{linearità})}{=} E_\theta[-2T_1 - \frac{35}{6}].$$

Pertanto, la statistica $\hat{T}_1 = -2T_1 - \frac{35}{6}$ è la correzione di T_1 che stavamo cercando e che risulta uno stimatore corretto del parametro θ .

E11.7 (tratto da appello; video). Consideriamo un modello statistico con densità uniforme sull'intervallo $(\theta - \frac{1}{4}, \theta + \frac{1}{2})$, dove $\theta \in \mathbb{R}$. Sia (X_1, \dots, X_n) un campione casuale di ampiezza n e $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ la relativa media campionaria.

- (a) Dire se la statistica $T = 8\bar{X}_n - 1$ è uno stimatore corretto del parametro θ e, nel caso in cui non lo sia, correggerlo.

- (b) Dire se lo stimatore corretto dato/trovato al punto precedente è consistente.
- (c) Supponiamo si siano registrati i seguenti dati campionari

0.963 1.086 1.111 0.975 0.793 1.332 1.463 0.877 1.436 0.948.

Usando lo stimatore corretto dato/trovato al punto (a), dare una stima del parametro θ .

Soluzione. Innanzitutto osserviamo che, poiché le variabili aleatorie X_1, \dots, X_n sono identicamente distribuite con densità $U(\theta - \frac{1}{4}, \theta + \frac{1}{2})$, risulta $E(X_i) = \theta + \frac{1}{8}$ e $\text{Var}(X_i) = \frac{3}{64}$ per ogni i .

- (a) La statistica T è uno stimatore corretto del parametro θ se $E_\theta(T) = \theta$. Usando il fatto che $E_\theta(\bar{X}_n) = \theta + \frac{1}{8}$, otteniamo

$$E_\theta(T) = E_\theta(8\bar{X}_n - 1) \stackrel{(\text{linearità})}{=} 8E_\theta(\bar{X}_n) - 1 = 8\theta$$

e quindi T non è uno stimatore corretto di θ . Affinché diventi corretto, basta moltiplicarlo per $\frac{1}{8}$. Pertanto, la statistica $\tilde{T} = \frac{T}{8} = \bar{X}_n - \frac{1}{8}$ è uno stimatore corretto di θ , ottenuto come correzione di T .

- (b) Dobbiamo verificare se lo stimatore \tilde{T} (corretto) sia o meno consistente. Uno stimatore è consistente se la sua varianza tende a zero al crescere dell'ampiezza del campione. Calcoliamo

$$\begin{aligned} \text{Var}_\theta(\tilde{T}) &= \text{Var}_\theta\left(\bar{X}_n - \frac{1}{8}\right) \stackrel{(\text{ propr. varianza})}{=} \text{Var}_\theta(\bar{X}_n) \\ &\stackrel{(\text{ propr. varianza } + \text{ indipendenza})}{=} \frac{1}{n^2} \sum_{i=1}^n \text{Var}_\theta(X_i) \stackrel{(\text{ equidistr.})}{=} \frac{3}{64n}, \end{aligned}$$

che va a zero se n tende all'infinito. Pertanto, \tilde{T} risulta essere uno stimatore consistente.

- (c) Per ottenere una stima del parametro θ usiamo lo stimatore \tilde{T} (corretto). La media campionaria dei dati registrati vale

$$\bar{x} = \frac{1}{10}(0.963 + 1.086 + 1.111 + \dots + 1.436 + 0.948) = 1.0984$$

e quindi risulta $\hat{\theta} = \bar{x} - \frac{1}{8} = 1.0984 - 0.125 = 0.9734$ (il parametro reale, che ho usato per generare i dati, era 1).