FIRMA DELLO STUDENTE

SECONDA PROVA INTERMEDIA DI STATISTICA
CLEA - CLEFIN - CLAPI (cod. 5047/4038) CLELI (cod. 4038) 371-377
14 Gennaio 2004

Cognome ***Nome***

Numero di matricola ***Codice corso***

COMPITO C1

Ai fini della valutazione si terrà conto solo ed esclusivamente di quanto riportato negli appositi spazi. Al termine della prova, è OBBLIGATORIO consegnare il presente foglio ed il foglio di brutta (DI CUI NON SI TERRÀ CONTO AI FINI DELLA VALUTAZIONE).
APPROSSIMARE TUTTI I CALCOLI ALLA QUARTA CIFRA DECIMALE

1. (2 punti) Sia X una variabile aleatoria Uniforme continua nell'intervallo $[0, 6k]$, $k > 0$.

a) Calcolare il valore atteso di X. **(1 punto)**

b) Calcolare la probabilità che X assuma valori nell'intervallo $\left(\frac{k}{3}, \frac{k}{2}\right]$. **(1 punto)**

a)

$E(X) = 3k$

b)

$$\Pr\left(\frac{k}{3} < X \le \frac{k}{2}\right) = \int_{k/3}^{k/2} \frac{1}{6k}\,dx = \frac{1}{36}$$

2. (3 punti) Date le variabili aleatorie indipendenti X ~ N(3, 16) e Y ~ N(2, 25):

a) Calcolare il valore atteso e la varianza della v.a. W ottenuta dalla seguente combinazione lineare W = 42 + 7X – 3Y. **(2 punti).**

b) Come si distribuisce la v.a. W? (**1 punto**)

a)

$E(W) = E(42 + 7X - 3Y) = 42 + 7E(X) - 3E(Y) = 42 + 21 - 6 = 57$

$Var(W) = Var(42 + 7X - 3Y) = 7^2 Var(X) + 3^2 Var(Y) = 49 \cdot 16 + 9 \cdot 25 = 1009$

b)

$N(57,1009)$

3. (3 punti)) La probabilità di riuscire a salire sulla linea gialla della metropolitana nelle ore di punta è pari a 4/5.

a) Come si distribuisce la v.a. X_i che descrive se la *i*-esima persona riesce a salire sulla metropolitana oppure no? **(1 punto)**

b) Se si trovano sulla banchina di attesa 150 persone, che cercano di salire indipendentemente una dall'altra, come si distribuisce *approssimativamente* la v.a. $\overline{X}$ "frequenza relativa delle persone che riescono a salire all'arrivo del prossimo convoglio"? **(1 punto)**

c) Qual è la probabilità che almeno 114 di queste riescano a salire? **(1 punto)**

a)

$$Ber(4/5) \Rightarrow X_i = \begin{cases} 0 & 1/5 \\ 1 & 4/5 \end{cases}$$

---

b)

$$N\left(\frac{4}{5}, \frac{4}{3750}\right)$$

---

c)

$$\Pr(\overline{X} \geq 114/150) = \Pr\left(Z \geq \frac{\frac{114}{150} - \frac{4}{5}}{\sqrt{\frac{4}{3750}}}\right) = 1 - \Phi(-1.2247) = 0.8897$$

4. (6 punti) Sia ($X_1,\ldots, X_6$) un campione bernoulliano estratto da una popolazione statistica di media $E(X) = 3q$ e $Var(X) = 1$.

Siano $T_1 = \left(\frac{1}{5}X_1 + \frac{2}{15}X_2\right)$ e $T_2 = \frac{X_4 + X_6}{6}$ due stimatori per *q*.

a) Dimostrare che sono entrambi corretti e calcolarne i rispettivi errori quadratici medi. **(4 punti)**

b) Qual è il più efficiente? E perché? **(2 punti)**

a)

$$E(T_1) = \frac{1}{5}3q + \frac{2}{15}3q = q$$

$$E(T_2) = \frac{1}{6}(3q + 3q) = q$$

$$EQM(T_1) = Var(T_1) = \left(\frac{1}{5}\right)^2 + \left(\frac{2}{15}\right)^2 = \frac{1}{25} + \frac{4}{225} = \frac{13}{225}$$

$$EQM(T_2) = Var(T_2) = \left(\frac{1}{6}\right)^2 + \left(\frac{1}{6}\right)^2 = \frac{2}{36} = \frac{1}{18}$$

---

b) T_2 è più efficiente perchè ha un errore quadratico medio minore.

5. (6 punti) Sia ($X_1, \ldots ,X_6$) un campione bernoulliano di osservazioni estratto da una popolazione Normale (μ, σ^2) di media e varianza entrambe incognite.
a) Si definisca con l'opportuna espressione analitica un intervallo di confidenza al livello (1-α) per il parametro μ. **(2 punti).**
b) Se il campione osservato risulta (11, 4, 17, 13, 3, 9), qual è l'intervallo di confidenza per μ a livello 1-α = 0.90? **(2 punti) [** Sugg. $\sum_{i=1}^{n} x_i = 57$, $\sum_{i=1}^{n} x_i^2 = 685$ **]**
c) Se fosse noto che σ = 7, quale sarebbe la dimensione campionaria minima necessaria per ottenere, allo stesso livello di confidenza, un intervallo di lunghezza pari al massimo a 4? (**2 punti)**

a)
$$\left(\overline{X} - t^{n-1}_{1-\frac{a}{2}} \frac{S_c}{\sqrt{n}} ; \overline{X} + t^{n-1}_{1-\frac{a}{2}} \frac{S_c}{\sqrt{n}} \right) \qquad (n=6)$$

---

b)
essendo $\overline{x} = 9.5$, s_c^2 = 28.7 , s_c = 5,3572 e $t^5_{0.95}$=2.015, si ha

(5.0931 ; 13.9069)

---

c)
$$\ell = 2 \cdot z_{0.95} \frac{s}{\sqrt{n}} = 2 \cdot 1.645 \cdot \frac{7}{\sqrt{n}} \le 4 \Rightarrow n \ge \left(\frac{1.645 \cdot 7}{2} \right)^2 = 33.14 \Rightarrow n = 34$$

6. (6 punti) Sia ($X_1, \ldots, X_5$) un campione da una popolazione X ~ N (μ, 144). Si vuole testare l'ipotesi nulla $H_0: \mu = 11$ contro l'ipotesi alternativa $H_1: \mu < 11$.
a) Proponete l'espressione analitica dell'opportuna regione di rifiuto. **(2 punti)**
b) Per α = 0.05, se (7, 4, 12, 18, 5) è il campione osservato, accetto o rifiuto l'ipotesi nulla? **(1 punto)**
c) Se il P-value osservato è 0.02, accetto o rifiuto l'ipotesi nulla? (Giustificare la risposta) **(1 punto)**
d) Se il test fosse caratterizzato dalla regione di rifiuto $R = \{(x ,\ldots, x_n) : \overline{x}_n < 9.5\}$ quale risulterebbe la probabilità dell'errore di prima specie α? **(2 punti)**

a)
$$R = \left\{ (x_1,\ldots,x_n) : \overline{x}_n < m_0 - z_{1-a} \frac{s}{\sqrt{n}} \right\} \qquad m_0 = 11, \quad s = 12, \quad n = 5$$

---

b)
da $z_{1-a} = z_{0.95} = 1,645$ e $\overline{x} = 9.2$ si ha
9.2 > 2.17, quindi accetto l'ipotesi nulla.

c)
Rifiuto l'ipotesi nulla, perchè il valore del P-value è minore del livello di significatività assegnato pari a 0.05.

---

d)

$$\Pr(\overline{X} < 9.5 \mid H_0) = \Pr\left(Z < \frac{9.5 - 11}{12/\sqrt{5}}\right) = \Phi(-0.28) = 0.3897$$

7. (4 punti) Dato un campione di osservazioni $(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)$, assumere che il modello lineare semplice spieghi la dipendenza della variabile aleatoria Y dalla variabile deterministica x significa ipotizzare che $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, per i=1, …, n.
E' noto che come stime puntuali dei parametri incogniti β_0 e β_1 si utilizzano le stime dei minimi quadrati $\hat{b}_0$ e $\hat{b}_1$.

a) Quale ipotesi aggiuntiva sulla distribuzione degli errori ε_i, rende *forti* le ipotesi *deboli* per il modello lineare semplice e consente, ad esempio, di calcolare stime di intervallo per β_0 e β_1? **(2 punti)**
b) La formula sotto riportata è nota sotto il nome di **scomposizione della devianza totale.**

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Quale delle quantità a destra del segno di uguale è la **devianza del modello**? **(1 punto)**

c) Quale rapporto tra le quantità coinvolte nella scomposizione fornisce il coefficiente di determinazione, che misura la bontà di adattamento del modello lineare ? **(1 punto)**

a)

Si assume che gli errori abbiano una *distribuzione normale*.

---

b)

$$\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$$

---

c)

$$\frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$