FIRMA DELLO STUDENTE

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

Cognome ______________________ *Nome* ______________________

Numero di matricola ______________________ *Codice corso* ______________________

COMPITO B1

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) La quantità di acqua contenuta nelle bottiglie da un litro e mezzo di una certa marca si distribuisce secondo una normale con valore atteso pari a 1.5 (litri) e varianza pari a 0.04 (litri^2).
   a) Calcolare la probabilità che, scelta a caso una bottiglia, essa contenga una quantità di acqua compresa fra 1.4 e 1.6 litri. **(1 punto)**

X = quantità d'acqua in una bottiglia X ~N(1.5, 0.04)

$$\Pr(1.4 < X < 1.6) = \Pr(X < 1.6) - \Pr(X \le 1.4) = \Pr\left(Z < \frac{0.1}{0.2}\right) - \Pr\left(Z \le -0.5\right) =$$

$$= 0.6915 - (1-0.6915) = \mathbf{0.383}$$

   b) Scelte a caso 150 bottiglie, si calcoli la probabilità che la quantità totale di acqua contenuta nelle 150 bottiglie sia maggiore di 222 litri? **(2 punti)**

$X_1 + \ldots + X_{150} \sim N(150 \cdot 1.5;\ 150 \cdot 0.04) \sim N(225;\ 6)$

$$\Pr(X_1 + \ldots + X_{150} > 222) = \Pr\left(Z > \frac{222-225}{\sqrt{6}}\right) = \Pr(Z > -1.22) = \Pr(Z = 1.22) = \mathbf{0.8888}$$

2) Siano X_1 e X_2 due variabili aleatorie *iid*, con $X_i \sim Be(0.2)$ ($i = 1,2$).
   a) Determinare la distribuzione del vettore aleatorio (X_1, X_2) **(1 punto)**

$$\Pr(X_i = x) = \begin{cases} 0.2 & \text{per } x = 1 \\ 0.8 & \text{per } x = 0 \end{cases}$$. Quindi avremo

X_1 \ X_2	**0**	**1**
0	0.64	0.16
1	0.16	0.04

   b) Fornire la funzione di probabilità di $Y = |X_1 - X_2|$ **(1 punto)**

$$\mathbf{Pr\ (Y_i = y) =} \begin{cases} 0.68 & \text{per } y = 0 \\ 0.32 & \text{per } y = 1 \end{cases}$$

3) Una società che si occupa di sondaggi vuole stimare la proporzione di italiani che intende votare SI ad un referendum. A tal fine vengono intervistati 300 italiani con diritto di voto, 200 dei quali dichiarano di essere intenzionati a votare SI.

a) Fornire una stima della proporzione incognita oggetto di indagine **(1 punto)**

Stimatore: $\overline{X}$ **Stima**: $\bar{x} = \dfrac{200}{300} = 0.6667$

b) Fornire uno stimatore e la corrispondente stima della varianza dello stimatore utilizzato al punto precedente **(2 punti)**

Stimatore: $\overline{X}$ Varianza di stima= $\mathrm{Var}(\overline{X}) = \dfrac{Var(X)}{n} = \dfrac{p(1-p)}{n}$

Stima della varianza di stima: $V\hat{a}r(\overline{X}) = \dfrac{\hat{p}(1-\hat{p})}{n} = \dfrac{0.6667 \cdot 0.3333}{300} = 0.0007$

4) Da una popolazione X ~ Uniforme(θ,θ+2) viene estratto un campione casuale semplice di ampiezza *n*. Al fine di stimare il parametro θ, vengono proposti i due stimatori:

$$T_1 = \overline{X}$$

$$T_2 = \frac{X_1 + X_n - 2}{2}$$

a) Dimostrare che T_1 è uno stimatore distorto rispetto a θ e proporre, a partire da T_1, uno stimatore non distorto per θ (si chiami questo nuovo stimatore T_1^*) **(2 punti)**

E(T_1) = E(X) = $\dfrac{J + (J+2)}{2} = J + 1$. Quindi uno stimatore non distorto può essere T_1^* = T_1-1, poiché E$\left(T_1^*\right)$ = E(T_1 -1) = $J + 1 - 1 = J$

b) Dimostrare che T_2 è uno stimatore non distorto per θ e calcolarne la varianza **(2 punti)**

$$\mathrm{E}\left(\frac{X_1 + X_n - 2}{2}\right) = \frac{E(X_1) + E(X_n) - 2}{2} = \frac{J + 1 + J + 1 - 2}{2} = J$$

$$\mathrm{Var}\left(\frac{X_1 + X_n - 2}{2}\right) = \frac{Var(X_1) + Var(X_n)}{4} = \frac{(J - J - 2)^2}{12} \cdot \frac{2}{4} = 0{,}1667$$

c) Tra T_1 e T_2 quale scegliereste per stimare θ? Giustificare la risposta **(2 punti)**

Per scegliere tra T_1 e T_2 devo confrontare i loro errori quadratici medi, valutati rispetto a J, che troviamo come somma di varianza e distorsione al quadrato.

$EQM_{T_1}(J) = \dfrac{1}{3n} + (1)^2 = 1 + \dfrac{1}{3n}$, mentre $EQM_{T_2}(J) = \dfrac{1}{6} + (0)^2 = \dfrac{1}{6}$. Quindi, per qualsiasi *n* **preferiremo T_2**, perché il suo errore quadratico medio è più basso (qualsiasi sia *n*).

[Nota. $Var(T_1) = Var(\overline{X}) = \dfrac{Var(X)}{n} = \dfrac{(J - J - 2)^2}{12} \cdot \dfrac{1}{n} = \dfrac{1}{3n}$]

5) Si vuole studiare la velocità media (in chilometri orari) μ delle automobili che percorrono una certa autostrada. A tal fine vengono controllate casualmente 200 automobili, ottenendo:

$$\sum_{i=1}^{200} x_i = 24000 \qquad \sum_{i=1}^{200} x_i^2 = 3000000$$

dove x_i rappresenta la velocità della i-esima automobile del campione.

a) Si fornisca uno stimatore non distorto rispetto a μ e se ne calcoli la stima corrispondente in relazione alla realizzazione campionaria osservata **(2 punti)**

Stimatore: $\overline{X}$, media campionaria Stima: $\overline{x} = \dfrac{24000}{200} = 120$

b) Si dia l'espressione dell'intervallo di confidenza asintotico di livello 0.9 per μ **(2 punti)**

I.C. $_{0.90}$ (μ) = $\left(\overline{X} - z_{0.95} \dfrac{S_c}{\sqrt{n}}, \overline{X} + z_{0.95} \dfrac{S_c}{\sqrt{n}} \right)$ Intervallo di confidenza aleatorio di livello 0.9 per μ.

i.c. $_{0.90}$ (μ) = $\left(\overline{x} - z_{0.95} \dfrac{s_c}{\sqrt{n}}, \overline{x} + z_{0.95} \dfrac{s_c}{\sqrt{n}} \right)$ Intervallo di confidenza di livello 0.9 per μ.

c) Si determinino, specificando il valore assunto da tutte le quantità coinvolte, gli estremi dell'intervallo sopra proposto in corrispondenza della realizzazione campionaria osservata **(2 punti)**

$$s_c^2 = \frac{200}{199}\left(\frac{3000000}{200} - (120)^2 \right) = \frac{200}{199}(15000 - 14400) = 603.0151 \qquad s_c = 24.5564$$

$\overline{x} = 120$ $\qquad n = 200$ $\qquad z_{0.95} = 1.645$

$$\text{i.c. }_{0.9}(\mu) = \left(120 - 1.645 \frac{24.5564}{\sqrt{200}}, 120 + 2.8564 \right) = (117.1436, 122.8564)$$

6) Si suppone che la distribuzione del prezzo di un certo prodotto nei supermercati di Milano segua una legge normale, con varianza pari a 1 euro^2. A partire da un campione casuale di 9 supermercati, si vuole verificare il seguente sistema di ipotesi:

H_0: $m=4$ (euro)

H_1: $m=3$ (euro)

dove μ rappresenta il prezzo medio incognito del prodotto in tutti i supermercati milanesi.

a) Si scriva la regione di rifiuto del test più potente al livello $\alpha = 0.1$ per verificare H_0 contro H_1 **(2 punti)**

$$R := \left\{ (x_1, ..., x_9) : \frac{\overline{x} - 4}{\sqrt{\frac{1}{9}}} < -z_{0.9} \right\} = \left\{ (x_1, ..., x_9) : \overline{x} < 4 - \frac{1}{3} \cdot 1.282 \right\} = \{ (x_1, ..., x_9) : \overline{x} < 3.5727 \}$$

b) Si prenda una decisione sulla base del test proposto, supponendo che la somma dei prezzi del prodotto registrati nei 9 supermercati del campione sia pari a 32.4 euro **(2 punti)**

$\overline{x} = 3.6$. Questo significa che $\overline{x} \notin R$ e quindi **accetto** l'ipotesi nulla.

c) Si calcoli la potenza del test utilizzato, specificando il procedimento adottato **(2 punti)**

Per valutare la potenza del test, valuteremo 1- b .

$$b = \Pr\{\overline{X} > 3.5727 | m = 3\} = \Pr\left\{Z > \frac{3.5727 - 3}{\frac{1}{3}}\right\} = \Pr\{Z > 1.7181\} = 1 - 0.9573 = 0.0427$$

Quindi la potenza del test, sarà pari a **0.9573**.

7) Esiste una relazione tra la temperatura e la profondità dell'acqua in un lago? Nella seguente tabella è riportato l'output della regressione lineare della variabile Y = "temperatura dell'acqua" (in gradi centigradi) sulla variabile X = "profondità" (in metri) (la regressione è effettuata in base ad una serie di dati sulle due variabili raccolti in diversi punti di uno stesso lago):

R	*R al quadrato*
0.922	0.850

	Coefficienti	*Errore standard*	*t*	*p-value*
intercetta	23.384	1.592	14.691	.000
PROFONDITA'	-1.435	.213	-6.726	.000

a) Si scriva l'equazione della retta di regressione lineare di Y rispetto a X **(1 punto)**

$$\hat{y} = 23.384 - 1.435x$$

b) Sulla base delle informazioni contenute nelle tabelle, è possibile rifiutare l'ipotesi secondo cui il coefficiente angolare della retta sarebbe uguale a zero? Perchè? **(2 punti)**

Secondo i valori presentati nella tabella, l'ipotesi nulla H_0: $b_1 = 0$ deve essere rifiutata, poiché il *p-value* è pari a 0.000, che è sicuramente più piccolo di un qualsiasi a prefissato.

c) Si valuti, giustificando la risposta, la bontà del modello di regressione lineare proposto per spiegare la variabile Y e si utilizzi il modello per prevedere la temperatura dell'acqua del lago in un punto di profondità pari a 15 metri **(1 punto)**

Un R^2=0.85 (ovvero un valore vicino a 1) indica un'alta capacità della variabile esplicativa X di spiegare il comportamento della variabile dipendente Y.
Per prevedere la temperatura del lago a una profondità di 15 metri basta sostituire a x il valore 15, ottenendo $\hat{y} = 23.384 - 1.435 \cdot 15 = 1.859$.