PROVA SCRITTA DI STATISTICA
(COD. 5047 - COD. 4038 - 371-377)
7 luglio 2005

APPROSSIMARE TUTTI I CALCOLI ALLA QUARTA CIFRA DECIMALE

SOLUZIONI MODALITÀ A

Esercizio 1 (9 punti) Supponiamo di aver osservato la seguente distribuzione doppia (frequenze assolute congiunte) relativa ai caratteri X= *voto di diploma* e Y= *voto medio esami sostenuti*, su un collettivo di studenti.

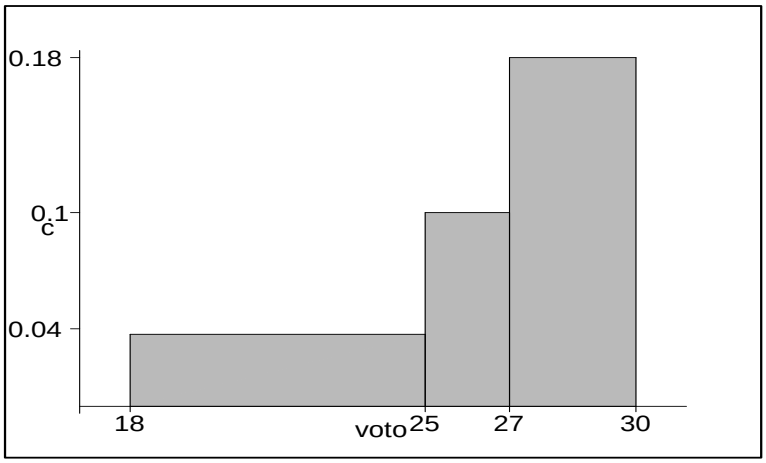
Y / **X**	$[18, 25)$	$[25, 27)$	$[27, 30]$	**tot**
[70, 80)	10	3	17	30
[80, 90)	13	10	27	50
[90, 100]	3	7	10	20
tot	26	20	54	100

a) (1 punto) Ricavare la distribuzione marginale di Y = *voto medio*.

voto	n_i	p_i	δ_i	c_i
[18, 25)	26	0.26	7	0.03714
[25, 27)	20	0.2	2	0.1
[27, 30]	54	0.54	3	0.18

b) (1 punto) Rappresentarla attraverso l'opportuna rappresentazione grafica.

La corretta rappresentazione è data dal seguente istogramma



c) (1 punto) Calcolare la media di Y.

$\mu_y = 0.26(21.5) + 0.2(26) + 0.54(28.5) = 26.18$

d) (2 punti) I due caratteri X e Y sono indipendenti?

No, non lo sono, infatti esiste almeno una frequenza congiunta relativa diversa dal prodotto delle corrispondenti frequenze marginali.
e) (1 punto + 2 punti) È possibile calcolare un indice di correlazione per i caratteri in questione? Se sì, calcolarlo e commentare il valore ottenuto.

Sì, è possibile. Calcoliamo tutte le quantità coinvolte in : $\rho = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)}$

$EX = 84$, $EY = 26.18$
$EX^2 = 7105$, $EY^2 = 694$
$Var(X) = 49$, $\sigma(X) = 7$
$Var(Y) = 8.6076$, $\sigma(Y) = 2.9339$
$EXEY = 2199.12$, $E(XY) = 2200.7$, $Cov(XY) = 1.58$

$$\rho = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)} = \frac{1.58}{7(2.9339)} = 0.0769$$

L'indice mostra la quasi assenza di legami di tipo lineare.

f) (1 punto) È utile a questo punto, calcolare l'indice di connessione relativo $\tilde{\varphi}^2$ per tali dati? Sì, no, perchè?

Sì, poichè, dato il basso valore di $|\rho|$, l'indice di connessione potrebbe indicare la presenza di un legame di tipo non lineare.

Esercizio 2 (3 punti) Si consideri il seguente esperimento aleatorio: si lanciano 2 monete regolari, e si osservano le facce risultanti. Se compare lo stesso simbolo su entrambe le facce, si lancia una terza moneta regolare, altrimenti l'esperimento è concluso.

a) (2 punti) Costruire lo spazio Ω dei possibili risultati dell'esperimento:
assumendo per semplicità che le monete siano lanciate in successione si ha

$$\Omega = \{TTC, TTT, CCT, CCC, TC, CT\};$$

calcolare la probabilità di ciascun evento elementare:

$$P(TTT) = P(TTC) = P(CCT) = P(CCC) = \frac{1}{8}; \qquad P(TC) = P(CT) = \frac{1}{4}.$$

b) (1 punto) Costruire la distribuzione di probabilità della variabile aleatoria X = Numero di teste osservate al termine dell'esperimento.

$$P(X=x) = \begin{cases} \frac{1}{8} & x=0 \\ \frac{5}{8} & x=1 \\ \frac{1}{8} & x=2 \\ \frac{1}{8} & x=3 \end{cases}$$

Esercizio 3 (3 punti) Assumiamo che il punteggio di un test d'ammissione ad un corso di dottorato di ricerca in una prestigiosa università italiana si distribuisca come una v.a. X Uniforme continua nell'intervallo $[36, 60]$.

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

Usando le formule relative ad una distribuzione continua uniforme su $[a, b]$,

$E(X) = \frac{a+b}{2} = \frac{96}{2} = 48$

b) (1 punto) Qual è la probabilità che X risulti maggiore di 57?

$P(X > 57) = 1 - \frac{57-a}{b-a} = 1 - \frac{21}{24} = \frac{1}{8}$.

c) (1 punto) Uno studente viene ammesso soltanto nel caso raggiunga un punteggio superiore a 57. Si presenta alla selezione un campione casuale semplice di 10 studenti. Qual è la probabilità che almeno due dei dieci studenti vengano ammessi?

Sia Y la v.a. che rappresenta il numero degli ammessi su 10 presentati. Allora $Y \sim$ Bin(n, p) con $n = 10$ e $p = 1/8$.

$$P(Y \geq 2) = 1 - P(Y=0) - P(Y=1) = 1 - \binom{10}{0}(\frac{7}{8})^{10} - \binom{10}{1}(\frac{1}{8})(\frac{7}{8})^{9} = 0.3611.$$

Esercizio 4 (2 punti) Siano $(x_1, x_2, \ldots, x_{64})$ i risultati riportati da un campione casuale semplice di 64 aspiranti presentatori ad un test di selezione per una trasmissione televisiva. Il risultato del test, per ogni partecipante alla selezione, è una v.a. X_i con distribuzione incognita di media $\mu = 20$ e varianza $\sigma^2 = 256$.

a) Determinare il valore atteso e la varianza della v.a. $\bar{X}_{64}$ = "punteggio medio".

Per le proprietà della media campionaria: $E(\bar{X}_{64}) = 20$, $Var(\bar{X}_{64}) = \frac{256}{64} = 4$

b) Qual è la probabilità (approssimata) che $\bar{X}_{64}$ risulti inferiore a 22.6?

$$P(\bar{X}_{64} < 22.6) = P\left(\frac{\bar{X}_{64}-\mu}{\frac{\sigma}{\sqrt{n}}} < \frac{22.6-20}{\sqrt{\frac{256}{64}}}\right) = P(Z < \frac{2.6}{\frac{16}{8}}) = \Phi(1.3) = 0.9032.$$

Esercizio 5 (4 punti) Estratto un campione casuale semplice di ampiezza $n = 25$ da una popolazione Normale, di media e varianza entrambe incognite, si ottiene $\sum_{i=1}^{25} x_i = 480$, $\sum_{i=1}^{25} x_i^2 = 10400$. Si vuole testare l'ipotesi nulla $H_0 : \mu \leq 15$ contro l'ipotesi alternativa $H_1 : \mu > 15$.

a) (1 punto) Per un generico livello α dell'errore di prima specie, riportare l'espressione analitica della regione di rifiuto del test.

$$R = \{(x_1, \ldots, x_{25}) : \bar{x}_{25} > 15 + t^{24}_{1-\alpha}\frac{s_c}{\sqrt{25}}\}$$

b) (3 punti) Qual è il *p-value* relativo alla realizzazione campionaria assegnata? Se α fosse fissato a 0.01, rifiutereste o no l'ipotesi nulla?

$$s_c^2 = \frac{n}{n-1}\left[\frac{10400}{25} - (\frac{480}{25})^2\right] = 1.0417(47.36) = 49.3349$$

$$p-value = P(\bar{X}_{25} > \bar{x}_{25}|H_0) = P(T_{24} > \frac{\bar{x}_{25}-15}{s_c/\sqrt{25}}) = P(T_{24} > \frac{19.2-15}{\frac{7.0239}{5}}) = P(T_{24} > 2.9898).$$

Si, rifiuteremmo, perchè dalle tavole della distribuzione T di Student con 24 gradi di libertà sappiamo che $P(T_{24} > 2.492) = 0.01$; quindi $P(T_{24} > 2.9898)$ sarà minore di $\alpha = 0.01$

Esercizio 6 (2 punti) Siano T_n e U_n due stimatori per un parametro incognito θ di una popolazione. Assumiamo che $E(T_n) = \theta$ ovvero che T_n sia non distorto per θ, e che $Var_\theta(U_n) \leq Var_\theta(T_n)$, per ogni θ.

a) (1 punto) È possibile stabilire quale sia lo stimatore più efficiente? Se sì, fornire la risposta.

Per determinare lo stimatore più efficiente è necessario confrontare gli errori quadratici medi dei due stimatori. È noto che $EQM(T_n) = D^2(T_n) + Var(T_n)$. Poichè $Var_\theta(U_n) \leq Var_\theta(T_n)$, ma lo stimatore T_n è non distorto, non possiamo determinare lo stimatore più efficiente.

b) (1 punto) Se fosse anche $E(U_n) = \theta$ la risposta precedente cambierebbe? Se sì , come?

In questo caso U_n risulterebbe lo stimatore più efficiente poichè il confronto degli errori quadratici medi si ridurrebbe al confronto delle varianze.

Esercizio 7 (2 punti) Sia $(X_1, \ldots, X_n)$ un campione casuale semplice estratto da una popolazione Normale $N(\mu, 9)$. Assumiamo che per il campione osservato risulti, $\sum_{i=1}^{n} x_i = 4680$.

a) Determinare l'ampiezza campionaria minima necessaria affinché l'intervallo di confidenza per μ a livello $(1-\alpha) = 0.90$ risulti di lunghezza inferiore a 0.5.

$l = 2(z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) < 0.5$, quindi $2(1.645\frac{3}{\sqrt{n}}) < 0.5$, ovvero $n \geq 390$.

Il minimo valore di n è dato da 390.
b) Calcolare esplicitamente l'intervallo di confidenza in questione utilizzando il valore n trovato al punto a). (Se non si è risolto il punto a) assumere $n = 36$).

Per n=390

$$IC(\mu) = \bar{x}_n \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = [12 \pm 1.645(\frac{3}{19.7484})] = [12 \pm 0.2499] = (11.7501,\ \ 12.2499).$$

Per n=36

$$IC(\mu) = \bar{x}_n \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = [130 \pm 1.645(\frac{3}{6})] = [130 \pm 0.8225] = (129.1775,\ \ 130.8225).$$

Esercizio 8 (2 punti) Riportare l'espressione analitica dell'indice per la valutazione della capacità di adattamento del modello lineare, noto come *coefficiente di determinazione*, esplicitando tutte le quantità coinvolte.

$$R^2 = \frac{SQM}{SQT} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

dove le y_i sono i valori osservati, $\bar{y}$ è la media dei valori osservati, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, sono i valori stimati dal modello.