PROVA SCRITTA DI STATISTICA

CLEA/CLEFIN/CLEMIT (cod. 5047/4038/371/377)

3 Novembre 2004

MOD. A

Esercizio N. 1 (3 punti).

Data la v.s. X avente funzione di densità:

$$f(x) = \begin{cases} 1/20 & 0 \leq x < 1 \\ 9/40 & 1 \leq x < 3 \\ 1/10 & 3 \leq x < 7 \\ 1/20 & 7 \leq x < 9 \\ 0 & \text{altrove} \end{cases}$$

a) Determinare la distribuzione di frequenza.
b) Calcolarne media, moda e mediana

a) (1 punto) La distribuzione di frequenza è:

$$p(x) = \begin{cases} 1/20 & \text{per } 0 \leq x < 1 \\ 9/20 & \text{per } 1 \leq x < 3 \\ 8/20 & \text{per } 3 \leq x < 7 \\ 2/20 & \text{per } 7 \leq x < 9 \\ 0 & \text{altrove} \end{cases}$$

b) (2 punti) La media è = 149/40, la moda è = 2 e la mediana è = 3.

Esercizio N. 2 (2 punti)

Enunciare la disuguaglianza di Chebyshev e verificarla per la v.a.

$$X = \begin{cases} 0 & 4 & 5 & 6 & 10 \\ \\ 0.05 & 0.1 & 0.7 & 0.1 & 0.05 \end{cases}$$

ponendo $\lambda = 2$. Riportare dettagliatamente tutti i passaggi.

Data una variabile statistica X con media μ e varianza σ^2 allora

$Fr(\mu - \lambda\sigma < X < \mu + \lambda\sigma) \geq 1 - 1/\lambda^2$. Nel nostro caso: $\mu = 5$, $\sigma^2 = 2.7$, $\sigma = 1.6432$, quindi
$Fr(\mu - \lambda\sigma < X < \mu + \lambda\sigma) = Fr(5 - 3.2864 < X < 5 + 3.2864) = Fr(1.7136 < X < 8.2864) =$
$0.1 + 0.7 + 0.1 = 0.9 > 1 - ¼ = 0.75$, quindi la disuguaglianza è verificata.

Esercizio N. 3 (3 punti).

E' data una v.s. a due dimensioni (X,Y) con E(X) = 5, E(Y) = 4, Var(X) = 36, Var(Y) = 9, COV(X,Y) = 18.
a) Quanto vale il coefficiente quadratico medio relativo di contingenza $\tilde{\hat{\varphi}}^2$? (Suggerimento: calcolare prima il coefficiente di correlazione lineare ρ)
b) Determinare l'equazione della retta di regressione di Y su X.
c) Prevedere il valore assunto da Y quando X = 8.

a) (1 punto) $\rho = \dfrac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$ = 18 / 18 = 1 Siamo nel caso di perfetta dipendenza lineare quindi anche il coefficiente quadratico medio relativo di contingenza vale 1.

b) (1 punto) Il coefficiente di regressione è b = Cov(X,Y) / Var(X) = 18 / 36 = 0.5; a = E(Y) – b E(X) = 1.5. L'equazione della retta di regressione di Y su X è quindi
y = 0.5 x +1.5

c) (1 punto) Il valore assunto da Y quando X = 8 è y(8) = 5.5

Esercizio N. 4 (4 punti).

Dati i numeri aleatori indipendenti X ed Y distribuiti secondo la legge bernoulliana con parametri rispettivi $p_X = 0.2$ e $p_Y = 0.3$
a) Trovare la distribuzione del numero aleatorio Z = X – Y.
b) Determinare E(Z) e Var(Z) e scrivere le relazioni di E(X) ed E(Y) con E(Z) e di Var(X) e Var(Y) con Var(Z).

a) (2 punti) Nella tabella seguente riportiamo in neretto le differenze Z fra i valori di X e di Y cui sono associate le probabilità congiunte che sono uguali al prodotto delle probabilità marginali, stante l'indipendenza fra X ed Y.

X / Y	**0**	**1**	
0	**0** 0.56	**1** 0.14	0.7
1	**-1** 0.24	**0** 0.06	0.3
	0.8	0.2	1.0

Riclassificando i valori di Z si trova la v.a.

$$Z = \begin{cases} -1 & 0 & 1 \\ 0.24 & 0.62 & 0.14 \end{cases}$$

b) (2 punti) E(Z) = -0.1 = E(X) – E(Y) = 0.2 – 0.3
Var(Z) = 0.37 = Var(X) + Var(Y) = 0.16 + 0.21

Esercizio N. 5 (4 punti).

Un'urna contiene 4 palline bianche e 6 rosse. Si estraggono 3 palline.
a) Verificare se è più probabile estrarre 3 palline bianche con la tecnica di estrazione con reimmissione oppure con quella di estrazione senza reimmissione.
b) Trovare la probabilità che le palline bianche siano più delle palline rosse, nel caso di reimmissione.
c) Trovare la probabilità che fra le palline estratte ce ne sia almeno una rossa, nel caso di reimmissione.

a) (1 punto)
Nel caso di reimmissione la probabilità di estrarre 3 palline bianche è = $(4/10)^3$ = 0.0640.
Nel caso di estrazione senza reimmissione tale probabilità diventa = 4/10*3/9*2/8 = 0.0333 minore della precedente.

b) (2 punti)
La probabilità che le palline bianche siano più delle rosse è uguale alla somma delle seguenti probabilità:
P(3 bianche) = 64/1000
P(2 bianche) = 3*4/10*4/10*6/10 = 288/1000
P(bianche più di rosse) = 64/1000 + 288/1000 = 352/1000 = 0.3520

c) (1 punto)
La probabilità che ci sia almeno una pallina rossa è la probabilità contraria a quella dell'evento "nessuna pallina rossa" (od anche "tre palline bianche) avente probabilità = 0.0640. La probabilità cercata è quindi = 1 – 0.0640 = 0.9360.

Esercizio N. 6 (4 punti)

Estraiamo dalla v.a. X dell'esercizio n. 2 un campione bernoulliano di 100 elementi
a) Trovare media e varianza della variabile media campionaria.
b) Determinare inoltre la probabilità approssimata che la media campionaria cada nell'intervallo (5 , 5.2) .
c) E' possibile che la media campionaria superi il valore 15? Giustificare la risposta.

a) (1 punto) La media campionaria ha media uguale alla media della v.a. di partenza e quindi è = 5.
La varianza della v.a. media campionaria è uguale ad 1/n della varianza della v.a. di partenza e quindi è = 2.7/100 = 0.027

b) (2 punti) Utilizzando il teorema del limite centrale si ha che la media campionaria segue una legge approssimativamente normale con media 5 e varianza = 0.027.
Poniamo z = $\frac{5.2-5}{\sqrt{0.027}}$ = 0.2 / 0.1643 = 1.2172: Dalle tavole della funzione di ripartizione della normale standardizzata si ha Φ(1.2172) = 0.8883. La probabilità approssimata che la media campionaria cada nell'intervallo (5 , 5.2) è = Φ(1.2172) - Φ(0) = 0.8883 – 0.5000 = 0.3883.

c) (1 punto) No, perchè la media deve essere sempre compresa fra il valore minimo (0) ed il valore massimo (10) della v.a. di partenza.

Esercizio N. 7 (2 punti).

Data una v.a. normale con varianza nota pari a 16 si estrae un campione bernoulliano di n elementi. L'intervallo di confidenza per la media, determinato con un livello di confidenza del 95%, ha lunghezza 3.1360. Determinare n.

La lunghezza dell'intervallo di confidenza è:

$L = 2\, z_{1-\alpha/2}\, (\sigma / \sqrt{n}) = 2 * 1.96 * 4 / \sqrt{n} = 15.68 / \sqrt{n} = 3.136$ da cui si ricava n = 25

Esercizio N. 8 (5 punti).

Per testare l'efficacia di un farmaco antinfluenzale questo viene somministrato ad un campione di 100 pazienti. Di questi 80 guariscono entro la prima settimana di cure. Sia p la percentuale di soggetti che potrebbero guarire entro la prima settimana nell'intera popolazione.

a) Definire una regione critica per verificare l'ipotesi $H_0 : p \geq 0.9$ contro l'alternativa $H_1: p < 0.9$. al livello $\alpha = 0.01$

b) Dare la definizione di p-value.

c) Determinare il p-value relativo alla realizzazione campionaria osservata.

d) E' ragionevole accettare l'ipotesi H_0 ? Giustificare la decisione.

a) (2 punti) Per verificare l'ipotesi $H_0 : p \geq 0.9$ contro l'alternativa $H_1: p < 0.9$ si può utilizzare il test asintotico, di dimensione $\alpha = 0.01$, che ha come regione critica:

$R = \{ (x_1, x_2,\ldots,x_n): \bar{x} < 0.9 - z_{0.99} \sqrt{\frac{0.9 * 0.1}{100}} \}$ Sostituendo: $R = \{ (x_1, x_2,\ldots,x_n) : \bar{x} < 0.8302\}$

b) (1 punto) Dato un test con regione critica R dipendente dalla statistica test $T = T(X_1, X_2,\ldots,X_n)$ si supponga di osservare $t = T(x_1, x_2,\ldots,x_n)$. Ciò premesso il p-value è la più alta probabilità, ottenibile sotto l'ipotesi nulla H_0, che T assuma valori più estremi rispetto alla sua realizzazione t.

c) (1 punto) Nel nostro caso il p-value è = $\Pr(\bar{X} < 0.8 \mid p = 0.9)$ dove $\bar{X}$ si distribuisce asintoticamente come una normale con media 0.90, varianza pari a 0.90 * 0.10 / 100 = 0.0009 e deviazione standard $\sqrt{0.0009}$ = 0.03. Si ha quindi:

$\Pr(\bar{X} < 0.8 \mid p = 0.9) = \Pr\{(\bar{X}-0.9)/0.03 < (0.8-0.9)/0.03\} = \Pr\{Z<(0.8-0.9)/0.03\}=\Phi(-3.333)= 0.0004$

d) (1 punto) Dato che la frequenza campionaria osservata 0.80 cade nella regione critica rifiutiamo l'ipotesi H_0.

Il giudizio è confermato dal fatto che il p-value = 0.0004 è minore del livello $\alpha = 0.01$.

PROVA SCRITTA DI STATISTICA

CLEA/CLEFIN/CLEMIT (cod. 5047/4038/371/377)

3 Novembre 2004

MOD. B

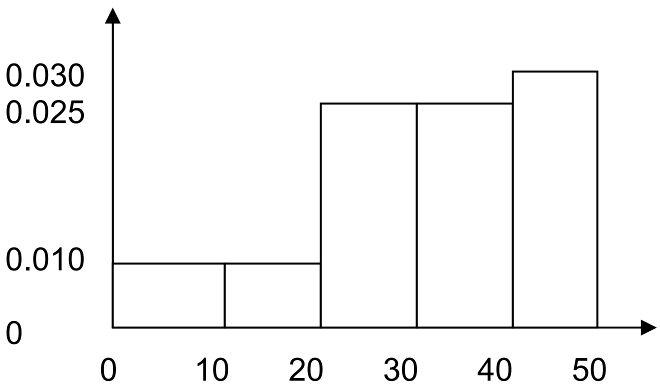
Esercizio N. 1 (6 punti)

Si è fatta un'indagine per stabilire l'efficacia di messaggi pubblicitari su un quotidiano per favorire la vendita di un certo prodotto nei supermercati di una città, per 20 settimane consecutive.
Sia X il numero di messaggi pubblicati in una settimana ed Y l'incasso, in migliaia di euro, per i prodotti venduti nello stesso periodo, riportati nella tabella seguente.

Y \ X	0	1	2	3	4	
[0 , 10]	**1**	**1**				**2**
(10,20]		**1**	**1**			**2**
(20,30]		**1**	**3**	**1**		**5**
(30,40]			**1**	**4**		**5**
(40,50]				**1**	**5**	**6**
	1	**3**	**5**	**6**	**5**	**20**

a) Rappresentare graficamente la distribuzione di frequenze di Y.
b) Calcolare media e moda di Y.
c) Trovare la funzione di regressione di Y su X.
d) Trovare mediana e varianza della distribuzione di Y condizionata ad X = 1.
e) Dare la definizione di indipendenza statistica e di indipendenza regressiva.

a) (1 punto)



b) (2 punti) media = 30.5 moda = 45 (valore centrale della classe modale (40,50])

c) (1 punto) La funzione di regressione di Y su X è data da:

$m_Y(0) = 5$
$m_Y(1) = 15$
$m_Y(2) = 25$
$m_Y(3) = 35$
$m_Y(4) = 45$

d) (1 punto) $(Y|X=1) = \begin{cases} [0\,,\,10] & (10,20] & (20,30] \\ 1/3 & 1/3 & 1/3 \end{cases}$

Mediana = 15
Varianza = $(5 - 15)^2\ 1/3 + 0 + (25 - 15)^2\ 1/3 = 66.6667$

e) (1 punto) Due v.s. X ed Y si dicono statisticamente indipendenti se le frequenze congiunte $p(x_i\,,\,y_k)$ sono uguali al prodotto delle frequenze marginali $p_x(x_i) * p_y(y_k)$, $\forall\,i$ e $\forall\,k$. X si dice regressivamente indipendente da Y quando la funzione di regressione $m_X(y)$ è costante e coincide con M(X); Y si dice regressivamente indipendente da X quando la funzione di regressione $m_y\,(x)$ è costante e coincide con M(Y).

Esercizio N. 2 (3 punti)

Nella seguente tabella a doppia entrata

X Y	0	1	
0			0.6
1			0.4
	0.4	0.6	1.0

si conosce il coefficiente di correlazione lineare ρ = -1.

a) Determinare il valore del coefficiente quadratico medio di contingenza φ^2 .
b) Riportare sul foglio la tabella completata con le frequenze congiunte.

a) (1 punto) Essendoci perfetta dipendenza fra X ed Y il coefficiente quadratico medio di contingenza relativo $\tilde{\varphi}^2$ = 1. Ma in caso di tabelle (2 x 2) $\tilde{\varphi}^2 = \varphi^2$, quindi il coefficiente quadratico medio di contingenza vale anch'esso 1.
b) (2 punti) Per la perfetta dipendenza fra X ed Y deve esserci una sola frequenza diversa da zero per ogni riga ed ogni colonna della tabella delle frequenze congiunte. Tenendo conto che la dipendenza è lineare inversa, l'unica struttura che risponde al problema è:

X Y	0	1	
0	0	0.6	0.6
1	0.4	0	0.4
	0.4	0.6	1.0

Esercizio N. 3 (3 punti)

Un'urna contiene 6 palline bianche, 3 nere ed una pallina rossa.
Si estrae una pallina. Se questa è bianca o nera il gioco termina. Se questa è rossa essa viene rimessa nell'urna e viene estratta un'altra pallina (In ogni caso non si procede ad altra estrazione).

a) Definire lo spazio Ω degli eventi elementari e calcolare la probabilità di ogni evento.
b) Se ogni volta che si estrae una pallina bianca si guadagna 1 euro e per ogni pallina nera si perde un euro, calcolare il guadagno atteso al termine del gioco.

a) (2 punti) $\Omega = \{b,n,rb,rn,rr\}$

Pr(b) = 0.60 Pr(n) = 0.30 Pr(rb) = 0.1*0.6=0.06 Pr(rn)=0.1*0.3=0.03   Pr(rr)=0.1*0.1=0.01

b) (1 punto) G = 1*0.60 – 1*0.30 + 1*0.06 -1*0.03 = 0.33

Esercizio N. 4 (3 punti)

Sia X una v.a. discreta tale che:
Pr(X = 0) = q
Pr(X = 1) = 3q
Pr(X ≠ 0,1) = 0

a) Determinare media e varianza di X.
b) Data $S = X_1 + X_2 + \ldots + X_{50}$ somma di 50 variabili indipendenti tutte uguali ad X e T = 4 S + 2 calcolare la media e la varianza di T.

a) (1 punto) Dobbiamo prima di tutto determinare q con la condizione q + 3 q = 1 da cui q = 0.25; E(X) = 0.75; Var(X) = 0.75 * 0.25 = 0.1875.
Notiamo che si tratta di una bernoulliana con parametro $\theta = 3q$.

b) (2 punti) Per i teoremi sulle variabili somma di variabili indipendenti si ha:
E(S) = 50 E(X) = 37.50
Var(S) = 50 Var(X) = 9.375
E(T) = 4 E(S) + 2 = 150 +2 = 152
Var(T) = 16 Var(S) = 150

Esercizio N. 5 (7 punti)

Sia X una v.a. normale con media μ e s.q.m. = 16. Si estrae da X un campione bernoulliano di n elementi.

a) Determinare il minimo valore di n tale che la probabilità che la media campionaria assuma valori compresi fra $\mu - 5$ e $\mu + 5$ sia pari almeno al 90%.

N.B. Se non si è risolto il punto a) porre n = 20 negli esercizi b) c) d).

b) Avendo trovato che la media del campione definito in precedenza è pari a 48 determinare l'intervallo di confidenza di μ al livello $\alpha = 0.05$.
c) Indicare la regione di rifiuto dell'ipotesi $H_0 : \mu = 50$ contro l'alternativa $H_1 : \mu \neq 50$ fissato $\alpha = 0.05$.

d) Stabilire, sempre sapendo che $\bar{x}$ = 48 e che α =0.05, se possiamo rifiutare l'ipotesi H_0 giustificando la decisione presa.

a) (3 punti) Sappiamo che $\bar{X}$ ~ N(μ,256/n). Deve essere: P(μ - 5 < $\bar{X}$ < μ + 5) ≥ 0.90. Ponendo

$$Z = \frac{\bar{X}-\mu}{16/\sqrt{n}}$$

e ricavando dalle tavole della normale il valore critico $z_{0.95}$ = 1.645 deve valere la condizione (μ + 5 - μ) / $\frac{16}{\sqrt{n}}$ ≥ 1.645 da cui $\frac{16}{\sqrt{n}}$ ≤ 5/1.645 = 3.0395. Si ricava n ≥ $(16/3.0395)^2$ = 27,7097 arrotondato a 28.

b) (2 punti) L'intervallo di confidenza di μ al livello del 5% è:
$ic_{0.95}(\mu) = (\bar{x}_n - z_{0.975}*16/\sqrt{n}\ ,\ \bar{x}_n + z_{0.975}*16/\sqrt{n}) = (48–1.96*16/\sqrt{28}\ ;\ 48 + 1.96*16/\sqrt{28}) =$ (42.0735 , 53.9265).
Se non si è svolto il punto a) $ic_{0.95}(\mu) = (48–1.96*16/\sqrt{20}\ ;48 + 1.96*16/\sqrt{20}) =$
= (40.9877, 55.0123)

c) (1 punto) Per verificare l'ipotesi H_0 **:** μ = 50 contro l'alternativa H_1 **:** μ ≠ 50 si può utilizzare il test di dimensione α = 5% che ha come regione critica:
$R = \{ (x_1,\ldots,x_n) : |\bar{x}_n – 50| > z_{0.975} *16/\sqrt{n} \}$ che nel nostro caso diventa
$R = \{ (x_1,\ldots,x_n) : |\bar{x}_n – 50| > 1.96*16/\sqrt{28} \} = \{ (x_1,\ldots,x_n) : |\bar{x}_n – 50| > 5,9265\}$
Se non si è svolto il punto a)
$R = \{ (x_1,\ldots,x_n) : |\bar{x}_n – 50| > 1.96*16/\sqrt{20} \} = \{ (x_1,\ldots,x_n) : |\bar{x}_n – 50| > 7.0123\}$

d) (1 punto) Con n = 28 accettiamo l'ipotesi H_0 in quanto |48 – 50| < 5.9265
Anche con n = 20 accettiamo l'ipotesi H_0 in quanto |48 – 50| < 7.0123.
La stessa conclusione si raggiunge osservando che in entrambi i casi 50 appartiene all'intervallo di confidenza per μ.

Esercizio N. 6 (5 punti)

E' stata fatta una rilevazione su 10 nuclei famigliari in particolare sui redditi x e sul numero Y di figli, ottenendo i risultati seguenti:

$$\sum_{i=1}^{10} x_i = 250 \qquad \sum_{i=1}^{10} Y_i = 26 \qquad \sum_{i=1}^{10} x_i Y_i = 900 \qquad \sum_{i=1}^{10} x^2{}_i = 9000 \qquad \sum_{i=1}^{10} Y^2{}_i = 80$$

a) Determinare i parametri del modello lineare di Y in x.

b) Conoscendo la somma dei quadrati degli errori $\sum_{i=1}^{10} \hat{\varepsilon}^2 = 5.03$ fornire una stima non distorta della varianza di ε_i .

c) Utilizzando ancora l'informazione data al punto b) verificare l'ipotesi $H_0 : \beta_1 = 0$ contro l'alternativa $H_0 : \beta_1 \neq 0$ fissato α = 0.05 (se non avete svolto il punto b) scrivere la regione di rifiuto dell' l'ipotesi $H_0 : \beta_1 = 0$ contro l'alternativa $H_0 : \beta_1 \neq 0$ fissato α = 0.05).

d) Enunciare le ipotesi del modello lineare distinguendo quelle deboli e quelle forti.

a) (1 punto) Nel modello lineare $Y = \beta_0 + \beta_1 x_i + \varepsilon_i$ (i=1,…,n) lo stimatore del coefficiente β_1 può mettersi nella forma:

$\hat{\beta}_1 = \dfrac{E(xY) - E(x)\,E(Y)}{Var(x)}$ dove

E(x)=250/10=25 E(Y)=26/10=2.6 Var(x)=9000/10 – 25^2=275 E(xY)=900/10=90

Si ha quindi $\hat{\beta}_1$= (90 – 25*2.6) / 275 = 0.0909

$\hat{\beta}_0$ = E(Y) - $\hat{\beta}_1$ E(x) = 2.6 – 0.0909*25 = 0.3275

b) (1 punto) Una stima non distorta della varianza di ε_i è $\hat{\sigma}^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$ / (n - 2). Nel nostro caso $\hat{\sigma}^2$ = 5.03 / 8 = 0.62875

c) (2 punti) La regione di rifiuto dell'ipotesi $H_0: \beta_1 = 0$ contro l'alternativa $H_0: \beta_1 \neq 0$ fissato α = 0.05 è:

$$R = \{ (y_1,x_1),\ldots,(y_n,x_n) : \frac{|\hat{\beta}_1|}{\hat{\sigma}_{\hat{\beta}_1}} > t_{1-\alpha}^{(n-2)} \}$$

Nel nostro caso (se si è svolto il punto b)):

$\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\dfrac{0.62875}{10 * 275}}$ = 0.0151 $t_{1-\alpha/2}^{(n-2)} = t_{0.975}^{(8)}$ = 2.306

Dato che 0.0909 / 0.0151 = 6.0199 > 2.306 rifiutiamo l'ipotesi nulla.

d) (1 punto) Le tre ipotesi deboli alla base del modello lineare sono:
1. $E(\varepsilon_i) = 0$ per ogni i = 1,…,n (il valore atteso di ogni variabile "errore" è nullo)
2. $Var(\varepsilon_i) = \sigma^2$ per ogni i=1,…,n (condizione di omoschedasticità)
3. $Cov(\varepsilon_i, \varepsilon_j) = 0$ per ogni i=1,…,n e j≠i (non correlazione degli "errori")

La quarta ipotesi che trasforma le ipotesi deboli in ipotesi forti è:

4. $\varepsilon_i \sim N(0, \sigma^2)$ per ogni i = 1,…,n (le variabili "errore" seguono la legge Normale con media zero ed hanno tutte la stessa varianza σ^2).