FACOLTA' DI SOCIOLOGIA – A.A. 2005-2006
ESAME DI ANALISI MULTIVARIATA
QUARTO RECUPERO
Appello del 20.12.05

Avvertenza: Fornire le formule utilizzate e tutti i passaggi dei calcoli eseguiti. Utilizzare almeno 2 cifre decimali.

Esercizio 1:

Il comune di Lodi vuole promuovere una campagna di informazione e di educazione alla mobilità ciclabile. I responsabili della giunta comunale sostengono che la metà dei residenti si dichiara non soddisfatta della sicurezza delle piste ciclabili cittadine e suggerisce alcuni provvedimenti per migliorare la mobilità ciclabile in città.

a) Scelti casualmente 6 cittadini, determinare la probabilità che fra essi ve ne siano *almeno* 2 che si dichiarano non soddisfatti della sicurezza delle piste ciclabili.

b) Il tempo (espresso in ore) che ciascun cittadino trascorre al mese in bicicletta per spostamenti di lavoro/studio è normalmente distribuito con media μ=30 e scarto quadratico medio σ=5. Scegliendo casualmente un cittadino, determinare la probabilità che trascorra in bicicletta al mese un numero di ore compreso tra 20 e 40.

c) Dalla popolazione dei residenti di Lodi si estrae un campione casuale di ampiezza n=3, al fine di stimare la media μ ignota del numero di ore trascorse in bicicletta nel tempo libero nell'ultimo anno (fenomeno X). A tale scopo si considerano i due seguenti stimatori:

$$T_1 = \bar{X} \qquad T_2 = \frac{4}{3}X_1 - X_2 + \frac{2}{3}X_3$$

(*i*) Verificare se gli stimatori T_1 e T_2 sono non distorti.
(*ii*) Decidere quale tra i due stimatori si dovrebbe scegliere, motivando la risposta.

Esercizio 2:

La giunta comunale decide di far condurre un'indagine sui cittadini per individuare i principali provvedimenti da adottare al fine di migliorare la sicurezza delle piste ciclabili. Per questo dalla lista dei residenti all'anagrafe ne vengono casualmente estratti 800 e si chiede loro, tra l'altro, se sono favorevoli ad un ampliamento/miglioramento delle piste (ex. interventi sulla pavimentazione). Tra di essi 200 si dichiarano non favorevoli all'iniziativa.

a) Costruire un intervallo di confidenza al 90% per la percentuale di cittadini che si dichiarano *favorevoli* all'iniziativa.

b) Stabilire qual è il numero minimo di osservazioni campionarie n per ottenere un intervallo con ampiezza al più pari a 0,051, fissando un livello di confidenza pari al 95%.

c) Commentare il risultato ottenuto al punto b) e discutere in linea teorica come varia l'ampiezza dell'intervallo di confidenza al variare:(*i*) del livello di confidenza $(1-\alpha)$; (*ii*) della numerosità campionaria n.

Esercizio 3:

a) Ad altri 100 cittadini di Lodi, estratti casualmente dalle liste dell'anagrafe, viene poi chiesto se sono residenti o meno nel centro città (variabile X) e se essi sono favorevoli o meno alla chiusura totale del centro città al traffico d'auto (variabile Y). I risultati ottenuti sono riassunti nella seguente tabella:

Y / *X*	*FAVOREVOLE ALLA CHIUSURA DEL CENTRO ALLE AUTO*	*NON FAVOREVOLE ALLA CHIUSURA DEL CENTRO ALLE AUTO*	
RESIDENTE IN CENTRO CITTA'	15	5	20
NON RESIDENTE IN CENTRO CITTA'	50	30	80
	75	25	*100*

Verificare se esiste *indipendenza* tra le variabili X e Y utilizzando un opportuno test statistico, considerando $\alpha = 0{,}05$ e commentando il risultato ottenuto.

b) In 15 interviste aggiuntive a cittadini di Lodi che svolgono la propria attività lavorativa all'interno del territorio comunale, viene chiesto quanto tempo (espresso in minuti) essi impiegano/impiegherebbero a raggiungere il posto di lavoro da casa utilizzando la bicicletta; dalle interviste è emerso che mediamente il tempo impiegato è pari a 25 minuti. Sapendo che il tempo T si distribuisce come una Normale con media e varianza ignote e che $\bar{s}^2 = 16$, verificare l'ipotesi (con $\alpha = 0{,}10$) secondo cui il tempo medio impiegato da casa al lavoro in bici è superiore a 30 minuti e commentare il risultato ottenuto.

c) Esporre il concetto e l'utilità di *p-value* e darne una metodologia di interpretazione per l'esecuzione di test statistici a 1 o a 2 code.

FACOLTA' DI SOCIOLOGIA – A.A. 2005-2006
ESAME DI ANALISI MULTIVARIATA
SOLUZIONI APPELLO DEL 20.12.05[1]

QUARTO RECUPERO

Esercizio 1:

a) La probabilità richiesta si ottiene applicando la formula:
$P(X \geq 2) = 1 - [P(X = 0) + P(X = 1)]$, dove

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

p=0,50
$(1-p)$=0,50
n=6
$x \geq 2$

$$P(X \geq 2) = 1 - P(X < 2) = 1 - [P(X=0) + P(X=1)] = 1 - \binom{6}{0}0{,}50^0 0{,}50^6 - \binom{6}{1}0{,}50^1 0{,}50^5 = .$$

$= 1 - 0{,}015625 - 0{,}09375 = 0{,}890625$

b) Viene richiesto di calcolare la seguente probabilità:
(in ore)
$P(20 < X < 40)$

Si standardizza:

$$P\left(\frac{20-30}{5} < \frac{X-30}{5} < \frac{40-30}{5}\right) = P(-2 < Z < 2) =$$
$$= P(Z < 2) - P(Z < -2) =$$
$$= P(Z < 2) - P(Z > 2) =$$
$$= P(Z < 2) - [1 - P(Z < 2)] =$$
$$= 0{,}97724994 - 1 + 0{,}97724994 = 0{,}95449988.$$

c) (*i*) In generale, lo stimatore T per il parametro θ si dice non distorto se $E(T) = \theta$.

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_i X_i\right) = \frac{1}{n}\sum_i E(X_i) = \frac{1}{n}\sum_i E(X) = \frac{1}{n} n\mu = \mu.$$

T_1 (media campionaria) è stimatore non distorto per μ (risultato noto).

$$E(T_2) = E\left[\frac{4}{3}X_1 - X_2 + \frac{2}{3}X_3\right] = \frac{4}{3}E(X_1) - E(X_2) + \frac{2}{3}E(X_3) \overset{i.i.d.}{=} \frac{4\mu}{3} - \mu + \frac{2}{3}\mu = \frac{3}{3}\mu = \mu$$

$E(T_2) = \mu$
T_2 è stimatore non distorto per μ.

(*ii*) Per scegliere tra due stimatori non distorti, si calcolano le loro varianze e si considera lo stimatore con varianza più piccola (ossia il più efficiente).

$$Var(T_1) = Var(\bar{X}) = Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2}\sum_i Var(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} = \frac{1}{3}\sigma^2.$$

(risultato noto)

[1] A cura di C. Colleoni

$$MSE(T_2) = Var(T_2) = Var\left(\frac{4X_1}{3} - X_2 + \frac{2}{3}X_3\right) = Var\left(\frac{4X_1}{3}\right) + Var(-X_2) + Var(\frac{2}{3}X_3) =$$

$$= \frac{16}{9}\sigma^2 + \sigma^2 + \frac{4}{9}\sigma^2 = \frac{29}{9}\sigma^2.$$

Come ci si aspettava, si sceglie lo stimatore T_1 (media campionaria): $MSE(T_1) < MSE(T_2)$

Esercizio 2:

a) Interessa conoscere la proporzione di cittadini che si dichiarano *favorevoli* all'ampliamento/miglioramento delle piste ciclabili ($P=A/N$)

a=600
n=800
p=600/800=0,75 proporzione campionaria
$\alpha = 0{,}10$

Essendo l'ampiezza campionaria, n=800, sufficientemente elevata, è possibile costruire l'IC asintotico di livello ($1-\alpha$) per la percentuale di successi P nella popolazione, dove per successo si intende la situazione in cui il cittadino intervistato è *favorevole* ai provvedimenti proposti per le piste ciclabili.

Tale intervallo coincide con:

$$P\left(p - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \le P \le p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) = 0{,}90$$

$$z_{1-\alpha/2} = z_{0{,}95} = 1{,}65$$

$$\left(0{,}75 - 1{,}65\sqrt{\frac{0{,}75\times 0{,}25}{800}}; 0{,}75 + 1{,}65\sqrt{\frac{0{,}75\times 0{,}25}{800}}\right)$$

(0,72474;0,77526)

b) L'ampiezza dell'intervallo di confidenza appena calcolato è pari a:
$0{,}77526 - 0{,}72474 = 0{,}050521$

$\alpha = 0{,}05$
$1-\alpha = 0{,}95$
$\alpha/2 = 0{,}025$
$1-\alpha/2 = 0{,}975$
$z_{0{,}975} = 1{,}96$

Ci si chiede ora quanto deve essere la numerosità campionaria n se si vuole un intervallo di ampiezza al più pari a 0,051 ed ad un livello di confidenza del 95%.

$$2z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \le 0{,}051$$

$$2^2 \times 1{,}96^2 \frac{0{,}75\times 0{,}25}{n} \le 0{,}051^2$$

$$\frac{2{,}8812}{n} \le 0{,}002552$$

$$n \ge \frac{2{,}8812}{0{,}002552} = 1128{,}845 \cong 1129$$

Si conclude che se si aumenta il livello di confidenza dal 90% al 95%, per avere un intervallo di ampiezza al più pari a 0,051, occorre passare da 800 a 1129 osservazioni campionarie.

c) Si vedano appunti e/o libri di testo.

Esercizio 3:

a) Test di indipendenza: si rifiuta H_0 se $\chi^2 > \chi^2_{1-\alpha,(h-1)(k-1)}$

$$\begin{cases} H_0 : \chi^2 = 0 \\ H_1 : \chi^2 > 0 \end{cases}$$

$$\chi^2 = 100 \times \left(\frac{15^2}{75 \times 20} + \ldots + \frac{30^2}{25 \times 80} - 1 \right) = 100 \times 0{,}066666667 = 6{,}7$$

$\alpha = 0{,}05$

$\chi^2_{1-\alpha,(h-1)(k-1)} = \chi^2_{0{,}95;1} = 3{,}84$

Poiché $6{,}7 > 3{,}84$, si RIFIUTA l'ipotesi nulla di indipendenza tra X e Y *a* livello di significatività 95%.

Nota bene: si sottopone a test il valore χ^2 e non $\tilde{\chi}^2$.

b)

n=15

$\bar{x} = 25$

$s^2 = 16$

$\mu_0 = 30$

$\alpha = 0{,}10$

La *statistica test* da utilizzare è la seguente:

$$\frac{\bar{X} - \mu_0}{s/\sqrt{n}} \approx T_{14}$$

dato che σ è *ignoto*.

Il valore osservato di T (valore sperimentale) è pari a:

$$t = \frac{25 - 30}{4/\sqrt{15}} = -4{,}841 .$$

$H_0 : \mu > 30$

Zona di rifiuto di H_0 : $t < t_\alpha$

$\alpha = 0{,}10$

$t_{\alpha,n-1} = t_{0{,}10;14} = 1{,}345$

Poichè $-4{,}841 < 1{,}345$, si RIFIUTA l'ipotesi nulla.

c) Si vedano appunti e/o libri di testo.