Esame di MQAT / Statistica del 2-2-2011

Nome / Mail : ....

1) Si calcoli il tasso medio di laureati a Venezia in base alla tabella di dati

Localita'	% Laureati	Popolazione
Rialto	0.09	70000
Mestre	0.04	150000
Lido	0.06	30000

2) Data la distribuzione Normale $x \sim \mathrm{N}(\mu = 3\,, \sigma^2 = 4)$ ed un campione di dimensione $n = 9$, calcolare la probabilita' condizionata $\mathrm{P}(\bar{x}_n < 2 \,|\, \bar{x}_n < 3.5)$

3) Data la distribuzione Bernoulliana $x \sim \mathrm{B}(\pi = 1/3)$ ed un campione di $n = 3$ osservazioni, calcolare la distorsione e l'efficienza della statistica $\tilde{x} = (\frac{1}{6}\,x_1 + \frac{2}{6}\,x_2 - \frac{3}{6}\,x_3)$ come stimatore per π. E' $\tilde{x}$ corretto e consistente ?

4) Si consideri la tabella di frequenze dell'Esercizio 6.15. Supponendo che l'indice di associazione di Pearson abbia fornito il valore $W = 8.21$, si chiede di verificare la sua significativita' con un test χ^2 al 90% e valutare l'indice di Cramer.

5) Due campioni di dimensioni $n_x = 10$, $n_y = 15$ sul valore di immobili di tipo X e Y ha fornito le somme $\sum_i x_i = 20$, $\sum_i y_i = 45$ e $\sum_i x_i^2 = 50$, $\sum_i y_i^2 = 165$. Calcolare le medie e la varianza "pooled", costruire un intervallo di confidenza al 90 % per la differenza delle medie e verificare l'ipotesi $\delta = \mu_y - \mu_x = 2$.

6) Si consideri la serie y_t in Figura 5.2.2. Le regressioni sui primi 6 dati sono

$$\text{trend lineare} \quad y_t = 576 - 0.29\,x_t + e_t, \quad S_e = 0.79 \tag{1}$$

$$\text{auto regress.} \quad y_t = .97\,y_{t-1} + u_t, \quad S_u = 1.20 \tag{2}$$

dove $x_t = 2001 \ldots 2006$ e $t = 1 \ldots 6$. Quale e' il modello migliore ?

Dalla Figura 5.2.2 gli ultimi valori sono $y_6 = 3.8$ e $y_7 = 4.8$. Si chiede di calcolare le previsoni per l'anno $x_7 = 2007$ ed il relativo intervallo di confidenza al 95%. Quale e' il modello migliore ? Le due previsoni sono compatibili ?

Soluzioni

1) la variabile X e' la % dei laureati, e bisogna calcolarne la media ponderata coi pesi f_j della popolazione: $\bar{X}$=.09*(7/25)+.04*(15/25)+.06*(3/25)=0.0564

2) la variabile e' $\bar{x}_n \sim$N(3, 4/9), avendo $(x < 2) \cap (x < 3.5) = (x < 2)$, la probabilita' cercata e' P($\bar{x}_n$ <2)/P($\bar{x}_n$ < 3.5) e standardizzando P(z < -1.5)/P(z <0.75)

3) per la Distorsione bisogna calcolare E($\tilde{x}$) $= \mu/6 + \mu/3 - \mu/2$=0, poiche' $\mu = \pi$=1/3, la distorsione e' quindi D($\tilde{x}$)=-1/3. Per l'Efficenza bisogna calcolare la varianza V($\tilde{x}$) $= \sigma^2/6^2 + \sigma^2/3^2 + \sigma^2/2^2$=0.086, essendo $\sigma^2 = \pi(1-\pi) = .22$, che e' maggiore di V($\bar{x}_3$)=.22/3. Quindi lo stimatore $\tilde{x}$ e' distorto ed inefficiente.

4) la tabella ha dimensione 3x4 e n=69, quindi la statistica di Pearson W_n ha distribuzione $\chi^2(2*3)$. Il valore critico al 90% e' $\chi^2_{.10}(6)$=10.6 > 8.21=W_n e quindi ci accetta l'ipotesi di indipendenza tra i caratteri. L'indice di Cramer $V = \sqrt{8.21/(69*2)}$=.24 e' basso.

5) dalle somme si calcolano $\bar{x}$=2, $\bar{y}$=3, S^2_x=50/10-4=1, S^2_y=2. La varianza *pooled* e' la media ponderata (per le dimensioni campionarie) delle due S^2_p=1*10/25 +2*15/25=1.6. L'intervallo di confidenza esatto al 90% per la differenza delle medie (supponendo x, y normali e $\sigma_x = \sigma_y$) usera' la studente $t_{.05}$(10+15-2)=1.7, da cui

$$\hat{\delta}_{1,2} = (3-2) \pm 1.7\sqrt{1.6/10 + 1.6/15} = [0.12, 1.88]$$

si rifiutano le ipotesi $\mu_x = \mu_y$ ma anche δ=2. Al 95% invece si accettano

6) il modello migliore e' il primo perche' ha varianza dei residui S_e piu' bassa; le due previsoni saranno $\hat{y}_7$=576-.29*2007=-6, e $\hat{y}_7$=.97*3.8=3.7. Quella che piu' si avvicina al valore reale y_7=4.8 e' la seconda. Gli intervalli di confidenza approssimati saranno $\hat{y}_7 \pm 2 * S$, quindi [-7.6, -4.4], [1.3, 6.1] che sono incompatibili non avendo punti in comune e solo il secondo include il vero y_7.