Departamento de Matemática

Probabilidades e Estatística

MEFT + MEBM + MEQ + LEMat

Segundo teste - Época de recurso — 2º semestre – 2010/2011

Duração: 90 minutos — 25/06/2011 – 11:00

▼ Justifique convenientemente todas as respostas ▼

Grupo I — 10 valores

1. Considere uma amostra aleatória de dimensão 10, $(X_1, X_2, \cdots, X_{10})$, de uma população X com distribuição normal de valor esperado μ e desvio padrão igual a 2.

(a) Determine o estimador de máxima verosimilhança do parâmetro μ. (3.0)

Função de verosimilhança:
$\mathcal{L}(\lambda; x_1, \ldots, x_{10}) \overset{i.i.d.}{=} \prod_{i=1}^{10} f_X(x_i) = \prod_{i=1}^{10} \frac{1}{\sqrt{2\,\pi \times 4}}\ e^{-\frac{1}{2\times 4}(x_i-\mu)^2} = (8\,\pi)^{-5}\ e^{-\frac{1}{8}\sum_{i=1}^{10}(x_i-\mu)^2}$

$\frac{d\log\mathcal{L}}{d\mu} = 0 \quad\Leftrightarrow\quad \frac{1}{4}\sum_{i=1}^{10}(x_i-\mu) = 0 \quad\Leftrightarrow\quad \mu = \bar{x}$ e $\frac{d^2\log\mathcal{L}}{d\mu^2} = -\frac{5}{2} < 0,\ \forall \mu \in I\!R.$

$\therefore$ Estimador de máxima verosimilhança (MV): $\hat{\mu} = \bar{X}$, cuja estimativa é $\bar{x}$.

(b) Considere os estimadores $T_1 = \bar{X}$ e $T_2 = \dfrac{X_1 + 3X_{10}}{4}$ do parâmetro μ. Qual destes estimadores é mais eficiente para estimar μ? Justifique. (2.0)

$E[T_1] = E[\bar{X}] = \frac{\sum_{i=1}^{10} E[X_i]}{10} \overset{i.d}{=} \mu$ e $E[T_2] = E[\frac{X_1+3X_{10}}{4}] = \frac{E[X_1]+3E[X_{10}]}{4} \overset{i.d}{=} \mu$

$\therefore T_1$ e T_2 são centrados para μ.

$Var[T_1] = Var[\bar{X}] \overset{i.}{=} \frac{\sum_{i=1}^{10} Var[X_i]}{10^2} \overset{i.d}{=} \frac{4}{10}$ e $Var[T_2] = Var[\frac{X_1+3X_{10}}{4}] \overset{i.}{=} \frac{Var[X_1]+9Var[X_{10}]}{16} \overset{i.d}{=} \frac{10}{4}$.

$e(T_1, T_2) = \frac{EQM[T_1]}{EQM[T_2]} = \frac{Var[T_1]+(E[T_1]-\mu)^2}{Var[T_2]+(E[T_2]-\mu)^2} = \frac{\frac{4}{10}}{\frac{10}{4}} = \frac{16}{100} < 1.$

$\therefore T_1$ é um estimador de μ mais eficiente que T_2.

2. O dono de uma ervanária comercializa um chá indicado para curar dores de cabeça. Num inquérito efectuado a 250 clientes dessa ervanária, escolhidos ao acaso entre os clientes que usaram o chá, 195 concordaram que o chá lhes cura as dores de cabeça.

(a) Construa um intervalo de confiança, a aproximadamente 95%, para a proporção de clientes da ervanária que consideram que o chá lhes cura as dores de cabeça. (3.5)

Seja $X_i = 1$, se o cliente i concorda com a cura da dor de cabeça devido ao chá, e 0, no caso contrário, $i = 1, \ldots, n$. Suponha que $(X_1, \ldots, X_n)$ é uma amostra aleatória de $X \sim$ Bernoulli(p), onde $p = P(X = 1)$. Pelo T.L.C. (grande amostras, $n = 250$), tem-se que $Z = \frac{\bar{X}-E(\bar{X})}{\sqrt{Var(\bar{X})}} = \frac{\bar{X}-p}{\sqrt{p(1-p)/n}} \overset{a}{\sim} N(0,1)$. Sabendo que $\bar{X}$ é um estimador consistente de p (i.e., $\bar{X} \to p$ quando $n \to \infty$), a variável fulcral apropriada é $Z^* = \frac{\bar{X}-p}{\sqrt{\bar{X}(1-\bar{X})/n}} \overset{a}{\sim} N(0,1)$

$P(-1.96 < Z^* < 1.96) \approx 0.95 \Leftrightarrow P\left(\bar{X} - 1.96\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} < p < \bar{X} + 1.96\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right) \approx 0.96.$

$\therefore IAC(p; 0.95) = \left(\bar{X} \pm 1.96\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right)$, cuja concretização é dada por

$IC(p; 0.95) = (\frac{195}{250} \pm 1.96\sqrt{\frac{0.78(1-0.78)}{250}}) = (0.78 \pm 0.0513) = (0.7287, 0.8313).$

(b) Que dimensão deveria ter a amostra de modo a que a amplitude do intervalo de confiança derivado na alínea anterior tivesse amplitude de aproximadamente 4%? (1.5)

Amplitude do intervalo de confiança em a): $(0.78 + 1.96\sqrt{\frac{\bar{x}(1-\bar{x})}{n}} - (0.78 - 1.96\sqrt{\frac{\bar{x}(1-\bar{x})}{n}})).$

$2 \times 1.96\sqrt{\frac{\bar{x}(1-\bar{x})}{n}} = 0.04 \ \Leftrightarrow\ n = (\frac{2\times 1.96}{0.04})^2 \times 0.78(1-0.78) \approx 1649.$

Grupo II	10 valores

1. No atendimento telefónico aos clientes de uma empresa, anotaram-se os tempos de espera até início de atendimento (em minutos) de 100 clientes, $(x_1, x_2, \cdots, x_{100})$, cujo agrupamento em 3 classes se apresenta na tabela seguinte:

classes (tempos de espera)	]0,2]	]2,4]	]4,+∞[
nº de clientes	40	35	25

O que pode concluir sobre a seguinte afirmação do responsável por esse serviço da empresa: "O tempo de espera numa chamada telefónica até um cliente ser atendido (X, em minutos) segue uma distribuição exponencial de parâmetro λ"? Argumente com base no valor-p, assumindo que a estimativa de máxima verosimilhança de λ é igual a 1/3. (5.0)

Considerem-se as classes $C_1 =]0,2]$, $C_2 =]2,4]$ e $C_3 =]4,+\infty[$ e seja $p_i = P(X \in C_i)$, $i = 1,2,3$. Pretende-se testar a hipótese $H_0 : X \sim \text{Exponencial}(\lambda)$ contra $H_1 : X \nsim \text{Exponencial}(\lambda)$, recorrendo à estatística de Pearson

$$Q = \sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i} \underset{H_0}{\overset{a}{\sim}} \chi^2_{(1)},$$

tendo em conta que o parâmetro λ será estimado pelo seu e.m.v. cuja estimativa é 1/3 e, por conseguinte, os graus de liberdade da distribuição do χ^2 são $k - m - 1 = 3 - 1 - 1 = 1$.

Classes	O_i	$E_i = np_i^0$	$(O_i - E_i)^2/E_i$
C_1	40	$100 \times F_X(2) = 100 \times 0.4866 = 48.66$	1.541
C_2	36	$100 \times (F_X(4) - F_X(2)) = 100 \times (0.7364 - 04866) = 24.98$	4.862
C_3	24	$100 \times (1 - F_X(4)) = 100 \times (1 - 0.7364) = 26.36$	0.211
	$n = 100$	100	$q = 6.614$

Valor-$p = P(Q \geq 6.614 | H_0) = 1 - F_{\chi^2_{(1)}}(6.614) \approx 1 - 0.9899 = 0.0101$.
A hipótese H_0 deve ser rejeitada para $\alpha > 0.0101$ e não deve ser rejeitada no caso contrário. Como o valor-p se encontra no intervalo de valores usuais para o nível de significância de um teste de hipóteses, a decisão vai naturalmente depender do nível de significância que for fixado.

2. A tabela seguinte apresenta resultados, obtidos no ano de 1980, para uma amostra de 9 países da América do Norte e Central. Para cada país, foi obtida a taxa de natalidade (número de nascimentos nesse ano, por 1 000 pessoas) e a taxa de urbanização (percentagem de população residente em cidades com pelo menos 100 000 habitantes).

	Canadá	Costa Rica	Cuba	EUA	El Salvador	Guatemala	Haiti	Honduras	Jamaica
i	1	2	3	4	5	6	7	8	9
Taxa de natal. (x_i)	16.2	30.5	16.9	16.0	40.2	38.4	41.3	43.9	28.3
Taxa de urban. (y_i)	55.0	27.3	33.3	56.5	11.5	14.2	13.9	19.0	33.1

$\sum_{i=1}^{9} x_i = 271.7$; $\sum_{i=1}^{9} x_i^2 = 9258.69$; $\sum_{i=1}^{9} y_i = 263.8$; $\sum_{i=1}^{9} y_i^2 = 10055.14$; $\sum_{i=1}^{9} x_i y_i = 6542.9$.

Com o objectivo de averiguar se existe uma relação linear entre a taxa de natalidade e de urbanização nos países da América do Norte e Central, ajustou-se um modelo de regressão linear, $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, 2, \ldots, 9$, com as suposições habituais.

(a) Obtenha a recta de regressão linear estimada. (2.0)

Sabe-se que as estimativas de minímos quadrados de β_0 e β_1 são, respectivamente,

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{6542.9 - 9 \times 30.1889 \times 29.3111}{9258.69 - 9 \times 30.1889^2} = \frac{-1420.929}{1056.363} = -1.345$ e

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 29.3111 - (-1.345) \times 30.1889 = 69.919$.

Assim, a equação de regressão linear estimada é $\hat{E}(Y|x) = 69.919 - 1.345\,x$, com $16 \leq x \leq 43.9$ (estimativas apropriadas).

(b) Teste a significância da regressão a um nível de significância de 10%. (3.0)

Pretende-se testar as hipóteses $H_0 : \beta_1 = 0$ contra $H_1 : \beta_1 \neq 0$ com base na estatística do teste $T = \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - n\bar{x}^2}}} \overset{H_0}{\sim} t_{(7)}$, cujo valor observado é $t = \frac{-1.345}{\sqrt{\frac{58.7947}{1056.363}}} \approx -5.702$. Note-se que $\hat{\sigma}^2 = \frac{1}{7}[10055.14 - 9 \times 29.3111^2 - (-1.345)^2(9258.69 - 9 \times 30.1889^2)] \approx 58.7947$. Para um nível de significância de 10%, a região crítica é $RC_{10\%} = (-\infty, -1.895) \cup (1.895, \infty)$ pois $F^{-1}_{t_{(7)}}(0.95) = 1.895$. Como $t \in RC_{1\%}$, rejeita-se a hipótese H_0 ao nível de significância de 10%, ou seja, a recta de regressão é significativa i.e. a taxa de natalidade influencia a taxa de urbanização nos países da América do Norte e Central, a esse nível de significância.