Departamento de Matemática

Probabilidades e Estatística

MEFT + MEBM + MEQ + LEMat

Segundo teste - versão 1 — 2º semestre – 2010/2011

Duração: 90 minutos — 09/06/2011 – 16:00

▼ Justifique convenientemente todas as respostas ▼

Grupo I	10 valores

1. Considere que o tempo entre chegadas consecutivas de chamadas telefónicas a uma central de táxis (X, em segundos) possui distribuição exponencial de parâmetro λ. Uma concretização de uma amostra aleatória de dimensão 100 da variável X conduziu a $\sum_{i=1}^{100} x_i = 1\,121$ segundos. Com base na amostra dada:

(a) Determine a estimativa de máxima verosimilhança de λ. (3.0)

Função de verosimilhança: $\mathcal{L}(\lambda; x_1, \dots, x_n) \overset{i.i.d.}{=} \prod_{i=1}^n f_X(x_i) = \prod_{i=1}^n \lambda\, e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$

$\frac{d \log \mathcal{L}}{d\lambda} = 0 \quad \Leftrightarrow \quad \frac{n}{\lambda} - \sum_{i=1}^n x_i = 0 \quad \Leftrightarrow \quad \lambda = \bar{x}^{-1}$ e $\frac{d^2 \log \mathcal{L}}{d\lambda^2} = -\frac{n}{\lambda^2} < 0,\ \forall \lambda > 0, \forall n \in I\!N$.

$\therefore$ Estimador de máxima verosimilhança (MV): $\hat{\lambda} = \bar{X}^{-1}$, cuja estimativa é $(\frac{1121}{100})^{-1} \approx 0.0892$.

(b) Obtenha a estimativa de máxima verosimilhança da probabilidade do tempo entre chegadas consecutivas de chamadas telefónicas à central ser inferior a 5 segundos. (2.0)

Sendo $\phi = P(X < 5) = \int_0^5 \lambda\, e^{-\lambda x} dx = 1 - e^{-5\lambda}$, o seu estimador de MV é $\hat{\phi} = 1 - e^{-5\bar{X}^{-1}}$, pela propriedade de invariância dos estimadores de MV.

Na amostra observada, a estimativa de MV de ϕ é $1 - e^{-500/1121} \approx 0.3598$.

2. Para avaliar a qualidade do ar na proximidade de duas cimenteiras, A (que possui co-incineração) e B (que não possui co-incineração), consideraram-se as variáveis aleatórias Y_1 e Y_2, que indicam a concentração de partículas em suspensão no ar (em microgramas por m^3) na proximidade das cimenteiras A e B, respectivamente. Suponha que Y_1 e Y_2 têm distribuições normais com variâncias desconhecidas mas iguais. A concretização de duas amostras aleatórias independentes, de dimensões 10 e 15, de Y_1 e Y_2, conduziu aos seguintes valores:

$$\sum_{i=1}^{10} y_{1i} = 850; \qquad \sum_{i=1}^{15} y_{2i} = 1\,350; \qquad \sum_{i=1}^{10} y_{1i}^2 = 73\,250; \qquad \sum_{i=1}^{15} y_{2i}^2 = 123\,000.$$

(a) Determine um intervalo de confiança a 98% para a diferença entre os valores esperados das concentrações de partículas em suspensão no ar na proximidade das duas cimenteiras. (3.5)

Se $Y_1 \sim N(\mu_1, \sigma^2)$ e $Y_2 \sim N(\mu_2, \sigma^2)$, independentes, a variável fulcral do intervalo de confiança (IC) de $\mu_1 - \mu_2$ é $W = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{9\,S_1^2 + 14\,S_2^2}{23}\left(\frac{1}{10} + \frac{1}{15}\right)}} \sim t_{(23)}$. Assim, $P(-2.5 < W < 2.5) = 0.98 \Leftrightarrow$

$$P\left(\bar{X}_1 - \bar{X}_2 - 2.5\sqrt{\frac{9S_1^2 + 14S_2^2}{23}\left(\frac{1}{10} + \frac{1}{15}\right)} < \mu_1 - \mu_2 < \bar{X}_1 - \bar{X}_2 + 2.5\sqrt{\frac{9S_1^2 + 14S_2^2}{23}\left(\frac{1}{10} + \frac{1}{15}\right)}\right) = 0.98.$$

$\therefore IAC(\mu_1 - \mu_2; 0.98) = \left(\bar{X}_1 - \bar{X}_2 \pm 2.5\sqrt{\frac{9\,S_1^2 + 14\,S_2^2}{23}\left(\frac{1}{10} + \frac{1}{15}\right)}\right)$, cuja concretização é dada por

$$IC(\mu_1 - \mu_2; 0.98) = 85 - 90 \pm 2.5\sqrt{\frac{9 \times 111.111 + 14 \times 107.143}{23}\left(\frac{1}{10} + \frac{1}{15}\right)} = (-15.64, 5.64).$$

(b) Com base no intervalo de confiança obtido na alínea anterior, teste ao nível de significância de 2% a veracidade da seguinte afirmação feita por um técnico da área ambiental: “Os valores médios das concentrações de partículas em suspensão no ar na proximidade das duas cimenteiras são iguais”. (1.5)

Pretende-se testar $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 \neq \mu_2$, usando $\alpha = 0.02$, com base no $IC(\mu_1 - \mu_2; 0.98)$ em a). Como $0 \in (-15.64, 5.64)$, não há nada contra a veracidade da afirmação do técnico da área ambiental (H_0) para um nível de 2% de significância, visto que com um grau de 98% de confiança, $\mu_1 - \mu_2 = 0$ é um valor plausível para essa diferença de valores esperados.

Grupo II — 10 valores

1. Para um determinado cruzamento rodoviário tem sido registado o número mensal de acidentes lá ocorridos. Uma amostra de 120 meses (10 anos) conduziu a resultados que se encontram agrupados na seguinte tabela: Será que os dados corroboram a hipótese de o número mensal de acidentes nesse

nº de acidentes	0	1	2	3 ou mais
frequência	79	27	12	2

(4.0)

cruzamento seguir a distribuição de Poisson de parâmetro 0.5? Recorra para o efeito ao cálculo do valor–p.

Seja X o número mensal de acidentes nesse cruzamento. Considerem-se inicialmente as classes $C_1 = \{0\}$, $C_2 = \{1\}$, $C_3 = \{2\}$ e $C_4 = \{3, \ldots\}$, fazendo junção de classes quando não houver i) todos E_i maiores ou iguais a 1 e ii) pelo menos 80% deles maiores ou iguais a 5, onde $p_i = P(X \in C_i)$, $i = 1, 2, 3, 4$. Pretende-se testar a hipótese $H_0 : X \sim \text{Poisson}(\lambda\!=\!0.5)$ contra $H_1 : X \nsim \text{Poisson}(\lambda\!=\!0.5)$, recorrendo à estatística de Pearson

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

Classes	O_i	$E_i = np_i^0$	$(O_i - E_i)^2/E_i$
C_1	79	$120 \times \mathrm{e}^{-0.5} = 72.784$	0.531
C_2	27	$120 \times \mathrm{e}^{-0.5} \times 0.5 = 36.392$	2.424
C_3	12 ⌉ 14	$120 \times \mathrm{e}^{-0.5} \times 0.5^2/2 = 9.098$ ⌉ 10.824	0.932
(C_4)	2 ⌋	$120 \times (1 - \sum_{x=0}^{2} \frac{\mathrm{e}^{-0.5} 0.5^x}{x!}) = 1.726$ ⌋	
	$n = 120$		$q = 3.887$

Valor-$p = P(Q \geq q | H_0) = 1 - F_{\chi^2_{(2)}}(3.887) \approx 1 - 0.8569 = 0.1431$.
A hipótese H_0 não deve ser rejeitada para $\alpha > 0.1431$ e deve ser rejeitada no caso contrário. Como o valor-p é superior aos valores usuais para o nível de significância de um teste de hipóteses (1%, 5% e 10%), há evidência a favor da hipótese de o número mensal de acidentes nesse cruzamento seguir a distribuição de Poisson de parâmetro 0.5.

2. Um especialista em motricidade humana pretende verificar se o modelo de regressão linear simples, $Y = \beta_0 + \beta_1 x + \epsilon$, com as hipóteses de trabalho habituais, é adequado para estudar a relação entre a altura (x, em cm) e o número de pulsações por minuto logo após a realização de uma determinada actividade física, Y. Para isso, efectuou medições em 12 pessoas praticantes dessa actividade, tendo obtido os seguintes resultados:

$\sum_{i=1}^{12} x_i = 1\,986$; $\sum_{i=1}^{12} x_i^2 = 336\,752$; $\sum_{i=1}^{12} y_i = 1\,572$; $\sum_{i=1}^{12} y_i^2 = 225\,932$. $\sum_{i=1}^{12} x_i y_i = 265\,831$

A recta estimada de mínimos quadrados é: $\widehat{E(Y|x)} = 14.807 + 0.702\,x$.

(a) Teste ao nível de significância de 1% a significância da recta de regressão. (4.0)

Pretende-se testar as hipóteses $H_0 : \beta_1 = 0$ contra $H_1 : \beta_1 \neq 0$ com base na estatística do teste $T = \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - n\bar{x}^2}}} \overset{H_0}{\sim} t_{(10)}$, cujo valor observado é $t = \frac{0.702}{\sqrt{\frac{1602.36}{8069}}} \approx 1.575$. Note-se que $\hat{\sigma}^2 = \frac{1}{10}[225932 - 12\times 131^2 - 0.702^2(336752 - 12\times 165.5^2)] \approx 1602.366$. Para um nível de significância de 1%, a região crítica é $RC_{1\%} = (-\infty, -3.169) \cup (3.169, \infty)$ pois $F^{-1}_{t_{(10)}}(0.995) = 3.169$. Como $t \notin RC_{1\%}$, não se rejeita a hipótese H_0 ao nível de significância de 1%, ou seja, a recta de regressão não é significativa i.e. a altura não influencia o número de pulsações por minuto logo após a realização de uma determinada actividade física, a esse nível de significância.

(b) Determine o coeficiente de determinação associado ao modelo de regressão considerado. Comente os resultados obtidos, tendo em conta o resultado desta alínea e o da alínea anterior. (2.0)

$R^2 = \frac{\left(\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}\right)^2}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \times \left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right)} = \frac{(5665)^2}{8069 \times 20000} \approx 0.1989.$

Cerca de 20% da variação total do número de pulsações por minuto após a actividade física é explicada pelo modelo de regressão linear simples com a altura como variável independente, indicando assim um mau ajustamento desse modelo.

Esse mau ajustamento já era previsível com a aceitação da hipótese de não significância da recta de regressão (hipótese H_0) na alínea anterior.