Probabilidades e Estatística

LEAN + LEE + LEGI + LEMat + MEAer + MEAmbi + MEBiol + MEBiom + MEEC + MEMec + MEQ

1º semestre – 2013/2014
09/01/2014 – 11:00

Duração: 90 minutos **2º teste B**

Justifique convenientemente todas as respostas!

Grupo I 10 valores

1. O tempo (X, em anos) decorrido entre avarias consecutivas de uma máquina de determinado modelo pode ser modelado por uma distribuição Exponencial(λ), $\lambda > 0$. Considere que $(X_1, X_2, \ldots, X_n)$, com $n > 2$, é uma amostra aleatória dessa distribuição.

(a) Tendo em vista a estimação do valor esperado entre duas avarias consecutivas da máquina, compare os estimadores $T_1 = \bar{X}$ e $T_2 = 0.5(X_1 + X_n)$, em termos de eficiência. (1.5)

$E[T_1] = E[X] = \frac{1}{\lambda}$, $Var[T_1] = \frac{Var[X]}{n} = \frac{1}{n\lambda^2}$
$E[T_2] = E[X] = \frac{1}{\lambda}$, $Var[T_2] = \frac{1}{4}(Var[X_1] + Var[X_2]) = \frac{1}{2\lambda^2}$
$e_{1/\lambda}(T_2, T_1) = \frac{EQM[T_1]}{EQM[T_2]} = \frac{Var[T_1]+(E[T_1]-1/\lambda)^2}{Var[T_2]+(E[T_2]-1/\lambda)^2} = \frac{\frac{1}{n\lambda^2}}{\frac{1}{2\lambda^2}} = \frac{2}{n} < 1, \forall n > 2 \iff T_1$ é mais eficiente que T_2 na estimação de $E[X] = 1/\lambda$.

(b) Determine os estimadores de máxima verosimilhança do parâmetro λ e da probabilidade de o tempo entre duas avarias consecutivas da máquina ser superior a 1 ano. (3.5)

$\mathscr{L}(\lambda; x_1, \ldots, x_n) = \prod_{i=1}^n f_X(x_i; \lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$
$\log \mathscr{L}(\lambda; x_1, \ldots, x_n) = n \log \lambda - \lambda \sum_{i=1}^n x_i$ (diferenciável em ordem a λ em $I\!R^+$)
$\frac{d \log \mathscr{L}(\lambda; x_1,\ldots,x_n)}{d\lambda} = 0 \iff \frac{n}{\lambda} - \sum_{i=1}^n x_i = 0 \iff \lambda = \bar{x}^{-1}$
$\frac{d^2 \log \mathscr{L}(\lambda; x_1,\ldots,x_n)}{d\lambda^2} = -\frac{n}{\lambda^2} < 0, \ \forall \lambda \in I\!R^+$.
$\therefore \hat{\lambda}_{MV} = \bar{X}^{-1}$.
Seja $g(\lambda) = P(X > 1) = \int_1^{+\infty} \lambda e^{-\lambda x}\, dx = e^{-\lambda}$. Pela invariância dos estimadores de máxima verosimilhança tem-se $\hat{h}_{MV}(\lambda) = h(\hat{\lambda}_{MV}) = e^{-\frac{1}{\bar{X}}}$.

2. Uma amostra casual de 236 estudantes universitários foi recentemente inquirida sobre hábitos tabágicos, tendo-se verificado que 47 dos estudantes são fumadores.

(a) Na década passada admitia-se que a proporção de fumadores na população universitária não era inferior a 30%. Averigue se os dados actuais permitem rejeitar essa hipótese a um nível de significância de 6%. (3.0)

A amostra observada é uma concretização de uma amostra aleatória de $X \sim Ber(p)$, em que $p = P$(um estudante universitário ser fumador). Quer-se testar $H_0: p \geq 0.30$ contra $H_1: p < 0.30$.
Uma vez que o tamanho da amostra é suficientemente grande temos, pelo TLC, $Z = \frac{\bar{X}-E[X]}{\sqrt{\frac{Var[X]}{236}}} = \frac{\bar{X}-p}{\sqrt{\frac{p(1-p)}{236}}} \overset{a}{\sim} N(0,1)$. Sob H_0, admitindo $p = 0.30$, obtemos a estatística do teste, $Z_0 = \frac{\bar{X}-0.30}{\sqrt{\frac{0.30\times0.70}{236}}} \overset{a}{\sim} N(0,1)$.
Para $\alpha = 0.06$ deve rejeitar-se H_0 se $Z_0 < \Phi^{-1}(0.06) = -1.5548$. Para a amostra observada temos $\bar{x} = 47/236$ e $z_0 = -3.3807$. Como z_0 pertence à região de rejeição então H_0 é rejeitada para $\alpha = 0.06$.
Alternativa: valor$-p = \Phi(-3.3807) = 3.6 \times 10^{-4} < 0.06$.

(b) Calcule a probabilidade aproximada de o procedimento que aplicou na alínea anterior conduzir a uma decisão errada quando aplicado a uma população de elevada dimensão com 32% de fumadores. (2.0)

Pretende-se $P\left(Z_0 < -1.5548 \,|\, p = 0.32\right)$. Dado que $p = 0.32$, temos agora que $Z^* = \frac{\bar{X}-0.32}{\sqrt{\frac{0.32\times0.68}{236}}} \overset{a}{\sim} N(0,1)$.
$P\left(Z_0 < -1.5548 \,|\, p = 0.32\right) = P\left(Z^* < \frac{-1.5548\sqrt{0.30\times0.70}-0.02\sqrt{236}}{\sqrt{0.32\times0.68}}\right) \approx \Phi(-2.1861) = 0.0144$.

Grupo II	10 valores

1. O tempo de vida, X, em centenas de horas, de um tipo de peças electrónicas foi registado para um conjunto de 1000 peças selecionadas ao acaso. Os dados obtidos encontram-se agrupados em classes na tabela seguinte: (4.5)

Tempo de vida	≤ 2	$]2,4]$	$]4,6]$	>6
Nº de peças	30	470	470	30

Será que os dados corroboram a hipótese de X ter distribuição normal com valor esperado igual a 400 horas e desvio padrão igual a 100 horas? Calcule, justificando, o valor-p do teste e decida com base no valor obtido, tendo em conta os níveis de significância usuais.

Pretende-se testar $H_0 : X \sim N(4,1)$ contra $H_1 : X \not\sim N(4,1)$.

Seja $p_i^0 = P(X \in \text{Classe}_i \mid H_0) = \begin{cases} P(X \leq 2 \mid H_0) = \Phi(-2) = 0.0228, & i=1 \\ P(2 < X \leq 4 \mid H_0) = \Phi(0) - \Phi(-2) = 0.4772, & i=2 \end{cases}.$

Pela simetria da distribuição $N(4,1)$ relativamente a $\mu = 4$ tem-se $p_3^0 = p_2^0$ e $p_4^0 = p_1^0$.

i	Classe_i	o_i	p_i^0	$e_i = np_i^0$
1	$]-\infty,2]$	30	0.0228	22.8
2	$]2,4]$	470	0.4772	477.2
3	$]4,6]$	470	0.4772	477.2
4	$]6,+\infty[$	30	0.0228	22.8
		$n = 1000$		

Como não é necessário agrupar classes ($k=4$) e não há qualquer parâmetro estimado ($\beta = 0$), a estatística de teste é $Q_0 = \sum_{i=1}^4 \frac{(O_i - E_i)^2}{E_i} \underset{H_0}{\overset{a}{\sim}} \chi^2_{(3)}$.

Tem-se $q_0 = 4.7646$ e valor$-p = P(Q_0 > q_0 \mid H_0) = 1 - F_{\chi^2_{(3)}}(4.7646) = 0.1899$. Deve-se rejeitar H_0 para níveis de significância ≥ 0.1899 e não rejeitar no caso contrário. Para os níveis de significância usuais, $\alpha \in [0.01, 0.1]$, não há evidência suficiente para rejeitar H_0.

2. Por vezes é necessário medir o recobrimento de armaduras em estruturas de betão armado já construídas. Para esse efeito existem dois métodos: o método directo (fornecendo o valor Y, em mm), que implica uma perfuração da estrutura, e um método indirecto (fornecendo o valor x, em mm), para o qual não é necessária a perfuração. Para comparar os dois métodos realizaram-se medidas com cada um dos métodos em 15 pontos de uma dada estrutura. Os dados sumariados foram os seguintes:

$$\sum_{i=1}^{15} x_i = 387 \quad \sum_{i=1}^{15} y_i = 468 \quad \sum_{i=1}^{15} x_i^2 = 11327 \quad \sum_{i=1}^{15} y_i^2 = 16040 \quad \sum_{i=1}^{15} x_i y_i = 13418$$

Considerando o modelo de regressão linear simples adequado, $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, 2, \ldots, 15$, com os pressupostos habituais para que este modelo tenha validade estatística:

(a) Determine um intervalo de confiança a 90% para o declive da recta de regressão. Que pode concluir sobre a relação entre os dois métodos de medição? (4.0)

Sejam $T = \frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2-15\bar{x}^2}}} \sim t_{(13)}$ e $a = F^{-1}_{t_{(13)}}(0.95) = 1.7709$

$P(-a \le T \le a) = 0.90 \Leftrightarrow P\left(\hat{\beta}_1 - a\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2-15\bar{x}^2}} \le \beta_1 \le \hat{\beta}_1 + a\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2-15\bar{x}^2}}\right) = 0.90$

$\mathrm{IAC}_{0.90}(\beta_1) = \left[\hat{\beta}_1 - 1.7709\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2-15\bar{x}^2}}, \hat{\beta}_1 + 1.7709\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2-15\bar{x}^2}}\right]$

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = 1.001$

$\hat{\sigma}^2 = \frac{1}{n-2}\left[\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) - \left(\hat{\beta}_1\right)^2\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\right] = 7.200$

$\mathrm{IC}_{0.90}(\beta_1) = [0.871, 1.131]$

A um nível de significância de 0.1 pode-se concluir que há uma relação de tipo linear entre x e Y ($\beta_1 \neq 0$) uma vez que $0 \notin IC_{0.90}(\beta_1)$. Por outro lado $1 \in IC_{0.90}(\beta_1)$ o que leva a não se rejeitar a hipótese $\beta_1 = 1$ ao mesmo nível de significância, o que indica que os 2 métodos conduzem a resultados concordantes a menos de uma possível constante aditiva.

(b) Calcule o coeficiente de determinação associado ao modelo considerado. Comente os resultados obtidos, tendo em conta o resultado desta alínea e o da alínea anterior. (1.5)

$R^2 = \frac{\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)^2}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\times\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)} = 0.935.$

Conclui-se que 93.5% da variabilidade observada nas medições obtidas pelo método directo é explicada pelo MRLS o que evidencia o bom ajustamento desse modelo aos dados. Este resultado é concordante com a rejeição da hipótese $H_0 : \beta_1 = 0$ na alínea anterior.