

Duração: 90 minutos

2º teste A

**Justifique convenientemente todas as respostas!**

**Grupo I**

10 valores

1. Seja  $X \sim \text{Geométrica}(p)$ ,  $0 < p < 1$ , e  $(X_1, X_2, \dots, X_{10})$  uma amostra aleatória de dimensão 10 proveniente da população  $X$ . Deduza o estimador de máxima verosimilhança de  $p$  e, sabendo que  $\sum_{i=1}^{10} x_i = 50$ , obtenha as estimativas de máxima verosimilhança de  $p$  e de  $P(X \leq 2)$ . (3.5)

Seja  $(x_1, \dots, x_n) \in \mathbb{N}^n$  uma concretização de  $(X_1, \dots, X_n)$ ,  $n = 10$ .

Função de verosimilhança:

$$\begin{aligned} L(p|x_1, \dots, x_n) &= f_{X_1, \dots, X_n}(x_1, \dots, x_n|p) \stackrel{X_i \text{ ind.}}{=} \prod_{i=1}^n f_{X_i}(x_i|p) \stackrel{X_i \text{ i.d.}}{=} \prod_{i=1}^n f_X(x_i|p) \\ &= \prod_{i=1}^n p(1-p)^{x_i-1} = p^n (1-p)^{\sum_{i=1}^n x_i - n}, \quad 0 < p < 1. \end{aligned}$$

Função de log-verosimilhança:

$$\ell(p|x_1, \dots, x_n) \equiv \log L(p|x_1, \dots, x_n) = n \log p + \left( \sum_{i=1}^n x_i - n \right) \log(1-p)$$

diferenciável em ordem a  $p$ . Maximização de  $\ell(p|x_1, \dots, x_n)$ :

$$\frac{d}{dp} \ell(p|x_1, \dots, x_n) = 0 \Leftrightarrow \frac{n}{p} - \frac{\sum_{i=1}^n x_i - n}{1-p} = 0 \Leftrightarrow n - np - p \sum_{i=1}^n x_i + np = 0 \Leftrightarrow p = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{x}}.$$

Seja  $\hat{p} = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{x}}$ , então  $\hat{p}$  é maximizante de  $\ell(p|x_1, \dots, x_n)$  se e somente se (onde  $\bar{x} \neq 0$ )  $\frac{d^2}{dp^2} \ell(p|x_1, \dots, x_n) \Big|_{p=\hat{p}} < 0$  e

$$\frac{d^2}{dp^2} \ell(p|x_1, \dots, x_n) \Big|_{p=\hat{p}} = \frac{d}{dp} \left( \frac{n}{p} - \frac{\sum_{i=1}^n x_i - n}{1-p} \right) \Big|_{p=\hat{p}} = - \left( \frac{n}{p^2} + \frac{\sum_{i=1}^n x_i - n}{(1-p)^2} \right) \Big|_{p=\frac{1}{\bar{x}}} < 0, \text{ pois } \sum_{i=1}^n x_i \geq n.$$

Logo, o estimador de máxima verosimilhança de  $p$  é  $\frac{1}{\bar{x}}$  e a sua estimativa é  $\frac{1}{\bar{x}} = \frac{10}{50} = 0.2$ .

$P(X \leq 2) = \sum_{x=1}^2 p(1-p)^{x-1} = p + p(1-p) = p(2-p)$ . Pela propriedade de invariância dos estimadores de máxima verosimilhança, o estimador de máxima verosimilhança de  $P(X \leq 2)$  é  $\frac{1}{\bar{x}}(2 - \frac{1}{\bar{x}})$  e a sua estimativa é  $0.2(2 - 0.2) = 0.36$ .

2. Numa loja foram seleccionadas, ao acaso e de forma independente, duas amostras de facturas: uma de 31 facturas com indicação de NIF e outra de 31 facturas sem NIF. Concluiu-se que a soma dos montantes das facturas foi de 630 euros na primeira amostra e 682 na segunda amostra. Os valores dos desvios padrões (corrigidos) nessas amostras foram, respectivamente, 6 e 7 euros. Supondo que, em qualquer dos dois casos, o montante por factura tem distribuição normal:

- (a) Calcule um intervalo de confiança a 90% para a variância do montante de uma factura com indicação de NIF emitida nessa loja. (3.0)

Sejam  $X_1 =$  valor de uma factura com NIF (em euros) e  $X_2 =$  valor de uma factura sem NIF (em euros).

$$n_1 = n_2 = 31, \sum_{i=1}^{n_1} x_{1i} = 630, \sum_{i=1}^{n_2} x_{2i} = 682, s_1 = 6, s_2 = 7, X_i \sim N(\mu_i, \sigma_i^2), i = 1, 2.$$

Variável fulcral:  $T = \frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi_{(n_1-1)}^2 = \chi_{(30)}^2$ .

Quantis:  $1 - \alpha = 0.9 = P(a < T < b)$ ,  $a : F_{\chi_{(30)}^2}(a) = 0.10/2 = 0.05 \Rightarrow a = 18.493$  e  $b : F_{\chi_{(30)}^2}(b) = 1 - 0.10/2 = 0.95 \Rightarrow b = 43.773$ .

Intervalo aleatório de confiança:

$$a \leq \frac{(n_1-1)S_1^2}{\sigma_1^2} \leq b \Leftrightarrow \frac{(n_1-1)S_1^2}{b} \leq \sigma_1^2 \leq \frac{(n_1-1)S_1^2}{a} \Rightarrow IAC_{0.90}(\sigma^2) = \left[ \frac{(n_1-1)S_1^2}{b}, \frac{(n_1-1)S_1^2}{a} \right]$$

Intervalo de confiança:

$$IAC_{0.90}(\sigma^2) = \left[ \frac{(n_1-1)s_1^2}{b}, \frac{(n_1-1)s_1^2}{a} \right] = \left[ \frac{30 \times 6^2}{43.773}, \frac{30 \times 6^2}{18.493} \right] = [24.673, 58.402].$$

- (b) Teste a hipótese de que nessa loja os valores esperados dos montantes dos dois tipos de facturas (com e sem NIF) são iguais, assumindo que as variâncias dos montantes são iguais nos dois casos. Decida através do valor-p do teste, tendo em conta os níveis de significância usuais. (3.5)

Admitindo  $\sigma_1^2 = \sigma_2^2$ ,

Hipóteses:

$H_0 : \mu_1 = \mu_2$  versus  $H_1 : \mu_1 \neq \mu_2$ .

Estatística do teste:

$$T_0 = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \stackrel{H_0}{\sim} t_{(n_1+n_2-2)} = t_{(60)},$$

cujo valor observado é  $t_0 = \frac{20.3225 - 22}{\sqrt{\frac{30 \times 6^2 + 30 \times 7^2}{60} \left( \frac{1}{31} + \frac{1}{31} \right)}} = -1.013$ .

Valor-p:

$p = P(|T_0| \geq |t_0| | H_0 \text{ verdadeiro}) = 2(1 - F_{t_{(60)}}(1.013)) = 2(1 - 0.842) = 0.3151$ .

Decisão: Se  $\alpha \geq 0.3151 \Rightarrow$  Rejeita-se  $H_0$  e se  $\alpha < 0.3151 \Rightarrow$  Não se rejeita  $H_0$ . Logo, aos níveis de significância usuais (0.01, 0.05, 0.10), não se rejeita a hipótese de os valores esperados dos montantes dos dois tipos de facturas (com e sem NIF) serem iguais.

## Grupo II

10 valores

1. Seja  $X$  a variável aleatória que indica o tempo, em segundos, entre chegadas consecutivas de carros a uma portagem. Uma concretização de uma amostra aleatória de dimensão 70 da variável  $X$  conduziu ao agrupamento em classes apresentado na tabela seguinte:

Classe	$\leq 5$	$]5,10]$	$]10,20]$	$>20$
Frequência absoluta	33	18	15	4

Teste ao nível de significância de 10% a hipótese de  $X$  seguir uma distribuição exponencial com valor esperado igual a 8 segundos. (4.5)

Hipóteses:

$H_0 : X \sim \text{Exponencial}(\lambda = 1/E(X) = 1/8)$  versus  $H_1 : X \not\sim \text{Exponencial}(\lambda = 1/8)$ .

Estatística do teste:

$$T_0 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \underset{H_0}{\overset{a}{\sim}} \chi_{(k-\beta-1)}^2 = \chi_{(3)}^2,$$

cujo valor observado é dado por

classe	$o_i$	$p_i$	$e_i = np_i$	$\frac{(o_i - e_i)^2}{e_i}$
$\leq 5$	33	0.4647	32.53 ( $\geq 5$ )	0.007
]5,10]	18	0.2488	17.41 ( $\geq 5$ )	0.020
]10,20]	15	0.2044	14.31 ( $\geq 5$ )	0.034
$>20$	4	0.0821	5.75 ( $\geq 5$ )	0.532
	$n = 70$	1	70	$t_0 = 0.593$

Se  $X \sim \text{Exponencial}(\lambda = 1/8)$ ,  $F_X(x) = \int_{-\infty}^x f_X(y) dy = \int_0^x \lambda e^{-\lambda y} dy = 1 - e^{-\lambda x}$ ,  $x > 0$ . Logo,

$$p_1 = P(X \leq 5 | H_0) = 1 - e^{-5/8} = 0.4647,$$

$$p_2 = P(5 < X \leq 10 | H_0) = P(X \leq 10 | H_0) - P(X \leq 5 | H_0) = e^{-5/8} - e^{-10/8} = 0.2488,$$

$$p_3 = P(10 < X \leq 20 | H_0) = P(X \leq 20 | H_0) - P(X \leq 10 | H_0) = e^{-10/8} - e^{-20/8} = 0.2044,$$

$$p_4 = P(X > 20 | H_0) = 1 - p_1 - p_2 - p_3 = 0.0821.$$

Note-se que  $k = 4$  ( $e_i \geq 5, \forall i$ ) e  $\beta = 0$  (não foram estimados parâmetros).

Região crítica:

$$\alpha = 0.10 = P(\text{Rejeitar } H_0 | H_0 \text{ verdadeiro}) \Leftrightarrow 0.10 = P(T_0 > b | H_0) \Leftrightarrow F_{\chi_{(3)}^2}(b) = 0.90 \Leftrightarrow b = 6.251.$$

$$RC = ]6.251, +\infty[.$$

Decisão: Se  $t_0 > 6.251 \Rightarrow$  Rejeita-se  $H_0$  e se  $t_0 \leq 6.251 \Rightarrow$  Não se rejeita  $H_0$ . Logo, como  $t_0 = 0.593 < 6.251$ , não se rejeita a hipótese de  $X$  seguir uma distribuição exponencial com valor esperado igual a 8 segundos ao nível de significância de 10%.

2. Para estudar a relação entre a altura das ondas ( $X$ , em metros) e o montante  $Y$  (em milhares de euros) dos estragos causados na orla costeira em dias de forte agitação marítima, foram obtidas observações relativas a 20 dias com forte agitação marítima, que conduziram a:

$$\sum_{i=1}^{20} x_i = 142 \quad \sum_{i=1}^{20} x_i^2 = 1334 \quad \sum_{i=1}^{20} x_i y_i = 4940 \quad \sum_{i=1}^{20} y_i = 420 \quad \sum_{i=1}^{20} y_i^2 = 23820$$

Considerando um modelo de regressão linear simples de  $Y$  sobre  $x$ :

- (a) Indicando hipóteses de trabalho convenientes, teste a significância do modelo de regressão linear ao nível de significância de 1%. (4.0)

Hipóteses de trabalho:  $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ , onde  $\epsilon_i \sim N(0, \sigma^2)$ ,  $Cov(\epsilon_i, \epsilon_j) = 0$ ,  $i \neq j$ ,  $i = 1, \dots, n$ ,  $n = 20$ .

Hipóteses:

$$H_0 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0.$$

Estatística do teste:

$$T_0 = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}} \underset{H_0}{\sim} t_{(n-2)} = t_{(18)},$$

cujo valor observado é  $t_0 = 6.0 / \sqrt{\frac{179.5983}{1334 - 20 \times 7.1^2}} = 8.094$ , onde  $\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = 6.01$  e  $\hat{\sigma}^2 = \frac{1}{n-2} [(\sum_{i=1}^n y_i^2 - n\bar{y}^2) - (\hat{\beta}_1)^2 (\sum_{i=1}^n x_i^2 - n\bar{x}^2)] = 179.5983$ .

Região crítica:

$$\alpha = 0.01 = P(\text{Rejeitar } H_0 | H_0 \text{ verdadeiro}) \Leftrightarrow 0.01 = P(|T_0| > b | H_0) \Leftrightarrow F_{t_{(18)}}(b) = 0.995 \Leftrightarrow b = 2.8784.$$

$$RC = ]-\infty, -2.8784[ \cup ]2.8784, +\infty[.$$

Decisão: Se  $|t_0| > 2.8784 \Rightarrow$  Rejeita-se  $H_0$  e se  $|t_0| \leq 2.8784 \Rightarrow$  Não se rejeita  $H_0$ . Logo, como  $t_0 = 8.094 > 2.8784$ , rejeita-se a hipótese de haver significância do modelo de regressão linear ao nível de significância de 1%.

- (b) Calcule o coeficiente de determinação do modelo e comente o valor obtido, tendo também em consideração o resultado da alínea anterior. (1.5)

$$r^2 = \frac{(\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y})^2}{(\sum_{i=1}^n x_i^2 - n\bar{x}^2)(\sum_{i=1}^n y_i^2 - n\bar{y}^2)} = \frac{(4940 - 20 \times 7.1 \times 21)^2}{(1334 - 20 \times 7.1^2)(23820 - 20 \times 21^2)} = 0.784$$

Pode-se afirmar-se que 78.4% da variabilidade observada do montante dos estragos causados foi explicada pela altura das ondas.

Este valor é suficientemente elevado para se dizer que o modelo se ajusta bem aos dados, o que é coerente com o facto de se ter rejeitado  $H_0 : \beta_1 = 0$ , ao nível de significância de 1%.