Probabilidades e Estatística

TODOS OS CURSOS

1º semestre – 2018/2019
30/01/2019 – **15:00**

Duração: 90 minutos **2º Teste C**

Justifique convenientemente todas as respostas

Grupo I	10 valores

1. Seja $\underline{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória proveniente da população X com função de densidade de probabilidade

$$f_X(x) = e^{-(x-\theta)} \exp\left[-e^{-(x-\theta)}\right], \quad x \in \mathbb{R},$$

onde θ é um parâmetro real desconhecido.

(a) Mostre que o estimador de máxima verosimilhança de θ é $\hat{\theta} = \ln\left(\frac{n}{\sum_{i=1}^n e^{-X_i}}\right)$. (3.0)

- **V.a. de interesse**
  X
- **F.d.p. de** X
  $f_X(x) = e^{-(x-\theta)} \exp\left[-e^{-(x-\theta)}\right], \quad x \in \mathbb{R}$
- **Parâmetro desconhecido**
  $\theta \in \mathbb{R}$
- **Amostra**
  $\underline{x} = (x_1, \ldots, x_n)$ amostra de dimensão n proveniente da população X
- **Obtenção do estimador de MV de** θ

  Passo 1 — Função de verosimilhança

$$\begin{aligned} L(\theta|\underline{x}) &= f_{\underline{X}}(\underline{x}) \\ &\overset{X_i\, indep}{=} \prod_{i=1}^n f_{X_i}(x_i) \\ &\overset{X_i \sim X}{=} \prod_{i=1}^n f_X(x_i) \\ &= \prod_{i=1}^n \left\{ e^{-(x_i-\theta)} \exp\left[-e^{-(x_i-\theta)}\right]\right\} \\ &= e^{-\sum_{i=1}^n (x_i-\theta)} \exp\left[-\sum_{i=1}^n e^{-(x_i-\theta)}\right]. \end{aligned}$$

  Passo 2 — Função de log-verosimilhança

$$\begin{aligned} \ln L(\theta|\underline{x}) &= -\sum_{i=1}^n (x_i - \theta) - \sum_{i=1}^n e^{-(x_i-\theta)} \\ &= -\sum_{i=1}^n x_i + n\theta - e^{\theta} \sum_{i=1}^n e^{-x_i}, \end{aligned}$$

  Passo 3 — Maximização
  A estimativa de MV de θ passa a ser representada por $\hat{\theta}$ e

$$\hat{\theta} \; : \; \begin{cases} \left.\frac{d \ln L(\theta|\underline{x})}{d\theta}\right|_{\theta=\hat{\theta}} = 0 & \text{(ponto de estacionaridade)} \\ \left.\frac{d^2 \ln L(\theta|\underline{x})}{d\theta^2}\right|_{\theta=\hat{\theta}} < 0 & \text{(ponto de máximo)} \end{cases}$$

$$\begin{cases} n - e^{\hat{\theta}} \sum_{i=1}^n e^{-x_i} = 0 \\ -e^{\hat{\theta}} \sum_{i=1}^n e^{-x_i} = -n < 0 \end{cases}$$

$$\hat{\theta} \quad : \quad \begin{cases} \hat{\theta} = \ln\left(\frac{n}{\sum_{i=1}^{n} e^{-x_i}}\right) \\ -n < 0 \qquad\qquad \text{(proposição verdadeira).} \end{cases}$$

Passo 4 — Estimador de MV de θ

$$EMV(\theta) = \ln\left(\frac{n}{\sum_{i=1}^{n} e^{-X_i}}\right).$$

(b) Calcule a estimativa de máxima verosimilhança de $h(\theta) = 2\theta + 1$ com base na amostra $\underline{x} = (x_1, \ldots, x_5) = (1.34, 2.56, -1.78, 4.67, 5.78)$ tal que $\sum_{i=1}^{5} e^{-x_i} \simeq 6.281$. (1.5)

- **Estimativa de MV de θ**

$$\begin{aligned} \hat{\theta} &= \ln\left(\frac{n}{\sum_{i=1}^{n} e^{-x_i}}\right) \\ &\simeq \ln\left(\frac{5}{6.281}\right) \\ &\simeq \ln(0.796052) \\ &\simeq -0.228091 \end{aligned}$$

- **Outro parâmetro desconhecido**
  $h(\theta) = 2\theta + 1$

- **Estimativa de MV de $h(\theta)$**
  Invocando a propriedade de invariância dos estimadores de máxima verosimilhança, pode concluir-se que a estimativa de MV de $h(\theta)$ é dada por

$$\begin{aligned} \widehat{h(\theta)} &= h(\hat{\theta}) \\ &= 2\hat{\theta} + 1 \\ &\simeq 2 \times (-0.228091) + 1 \\ &\simeq 0.543818. \end{aligned}$$

2. Uma instituição financeira admite que o tempo (em meses) até ao pagamento integral de crédito atribuído a pequenas e médias empresas é uma variável aleatória X com distribuição exponencial de valor esperado igual a δ. Da sua carteira de créditos a pequenas e médias empresas foi retirada uma amostra casual de dimensão $n = 120$ que conduziu a $\sum_{i=1}^{120} x_i = 2031.6$.

(a) Obtenha um intervalo de confiança a aproximadamente 96% para δ. Considere a variável aleatória fulcral $Z = \sqrt{120}\left(\frac{\bar{X}}{\delta} - 1\right)$, cuja distribuição é aproximadamente normal$(0,1)$. (2.5)

- **V.a. de interesse**
  X = tempo (em meses) até ao pagamento integral de crédito...

- **Situação**
  $X \sim$ Exponencial$(1/\delta)$
  $E(X) = \sqrt{V(X)} = \delta > 0$ DESCONHECIDO
  $n = 120 > 30$ (suficientemente grande).

- **Obtenção de IC aproximado para $E(X) = \delta$**
  Passo 1 — Selecção da v.a. fulcral para δ

$$Z = \sqrt{120}\left(\frac{\bar{X}}{\delta} - 1\right) \stackrel{a}{\sim} \text{normal}(0,1)$$

  [Uma vez que nos foi solicitada a determinação de um IC aproximado para o valor de X e a dimensão da amostra é suficientemente grande para invocar o TLC, faremos uso da seguinte v.a. fulcral para δ:

$$Z = \frac{\bar{X}-E(\bar{X})}{\sqrt{V(\bar{X})}} = \frac{\bar{X}-E(X)}{\sqrt{\frac{V(X)}{n}}} = \frac{\bar{X}-\delta}{\sqrt{\frac{\delta^2}{n}}} = \sqrt{n}\left(\frac{\bar{X}}{\delta}-1\right) \overset{a}{\sim} \text{normal}(0,1).]$$

Passo 2 — Obtenção dos quantis de probabilidade

Os quantis a utilizar são

$$\begin{cases} a_\alpha = \Phi^{-1}(\alpha/2) = -\Phi^{-1}(1-\alpha/2) = -\Phi^{-1}(0.98) \overset{tabela/calc.}{=} -2.0537 \\ b_\alpha = \Phi^{-1}(1-\alpha/2) = \Phi^{-1}(0.98) = 2.0537. \end{cases}$$

[Estes quantis enquadram a v.a. fulcral para δ com probabilidade aproximadamente igual a $(1-\alpha) = 0.96$.]

Passo 3 — Inversão da desigualdade $a_\alpha \le Z \le b_\alpha$

$$P(a_\alpha \le Z \le b_\alpha) \simeq 1-\alpha$$

$$P\left[a_\alpha \le \sqrt{120}\left(\frac{\bar{X}}{\delta}-1\right) \le b_\alpha\right] \simeq 1-\alpha$$

$$P\left(1+\frac{a_\alpha}{\sqrt{120}} \le \frac{\bar{X}}{\delta} \le 1+\frac{b_\alpha}{\sqrt{120}}\right) \simeq 1-\alpha$$

$$P\left(\frac{\bar{X}}{1+\frac{b_\alpha}{\sqrt{120}}} \le \delta \le \frac{\bar{X}}{1+\frac{a_\alpha}{\sqrt{120}}}\right) \simeq 1-\alpha.$$

Passo 4 — Concretização

Ao termos em conta que

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{2031.6}{120} = 16.93$
- $\Phi^{-1}(1-\alpha/2) = 2.0537$,

conclui-se que o intervalo de confiança a aproximadamente 96% para δ é dado por

$$\left[\frac{\bar{x}}{1+\frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{120}}}, \quad \frac{\bar{x}}{1-\frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{120}}}\right] = \left[\frac{16.93}{1+\frac{2.0537}{\sqrt{120}}}, \quad \frac{16.93}{1-\frac{2.0537}{\sqrt{120}}}\right]$$

$$\simeq \quad [14.257127, 20.836315].$$

(b) Confronte as hipóteses $H_0: \delta = 24$ e $H_1: \delta < 24$, calculando para o efeito o valor-p. (3.0)

- **Hipóteses**
  $H_0: \delta = \delta_0 = 24$
  $H_1: \delta < \delta_0 = 24$

- **Estatística de teste**
  [Pode tirar-se da variável fulcral utilizada em (a) para obter a seguinte estatística de teste:]

$$T = \sqrt{120}\left(\frac{\bar{X}}{\delta_0}-1\right) \overset{a}{\sim}_{H_0} \text{normal}(0,1).$$

- **Região de rejeição de H_0** (para valores de T)
  Tratando-se de um teste unilateral inferior ($H_1: \delta = E(X) < \delta_0$) e haver tendência para os valores tomados por $\bar{X}$ decrescerem à medida que δ diminui, a região de rejeição de H_0, escrita para valores da estatística de teste, é do tipo $W = (-\infty, c)$.

- **Decisão (com base no valor-p)**
  O valor observado da estatística de teste é

$$\begin{aligned} t &= \sqrt{120}\left(\frac{\bar{x}}{\delta_0}-1\right) \\ &= \sqrt{120}\left(\frac{16.93}{24}-1\right) \\ &\simeq -3.23. \end{aligned}$$

Uma vez que a região de rejeição deste teste é um intervalo à esquerda, temos:

$$\begin{aligned} valor-p &= P(T < t \mid H_0) \\ &\simeq \Phi(t) \\ &\simeq \Phi(-3.23) \\ &= 1-\Phi(3.23) \\ &\overset{calc/tabela}{=} 1-0.999381 \\ &= 0.000619. \end{aligned}$$

Logo é suposto:

- não rejeitar H_0 a qualquer n.s. $\alpha_0 \le 0.0619\%$;
- rejeitar H_0 a qualquer n.s. $\alpha_0 > 0.0619\%$, nomeadamente a qualquer dos n.u.s. (1%, 5% e 10%).

Grupo II — 10 valores

1. Os tempos de espera (em segundo) para atendimento de 200 chamadas, escolhidas ao acaso de entre as recebidas nos Centros de Orientação de Doentes Urgentes do INEM, estão sumariados na tabela abaixo. (4.0)

Classe	[0,5]	]5,15]	]15,+∞[
Frequência absoluta observada	76	54	70

Teste, ao nível de significância de 5%, a hipótese de o tempo de espera para atendimento seguir uma distribuição exponencial de parâmetro igual a 0.1.

- **V.a. de interesse**
  X = tempo de espera para atendimento

- **Hipóteses**
  $H_0 : X \sim \text{Exponencial}(\lambda),\ \lambda = 0.1$
  $H_1 : X \not\sim \text{Exponencial}(0.1)$

- **Nível de significância**
  $\alpha_0 = 5\%$

- **Estatística de Teste**

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \overset{a}{\sim}_{H_0} \chi^2_{(k-\beta-1)},$$

  onde:

  k = No. de classes = 3
  O_i = Frequência absoluta observável da classe i
  E_i = Frequência absoluta esperada, sob H_0, da classe i
  β = No. de parâmetros a estimar = 0 [dado que em H_0 se conjectura uma f.d. específica.]

- **Frequências absolutas esperadas sob H_0**
  Atendendo à dimensão da amostra $n = 200$ e à f.d. conjecturada dada por

$$F_{X|H_0}(x) = F_{Exponencial(0.1)}(x) = 1 - e^{-0.1\,x}, \quad x \ge 0,$$

  segue-se, para $i = 1,2,3$:

$$\begin{aligned} E_1 &= n \times [F_{X|H_0}(5) - F_{X|H_0}(0)] \\ &= 200 \times \left[\left(1 - e^{-0.1\times 5}\right) - 0\right] \\ &= 78.69; \end{aligned}$$

$$
\begin{aligned}
E_2 &= n \times [F_{X|H_0}(15) - F_{X|H_0}(5)] \\
&= 200 \times \left[\left(1 - e^{-0.1\times 15}\right) - \left(1 - e^{-0.1\times 5}\right)\right] \\
&= 76.68; \\
E_3 &= n - \sum_{i=1}^{2} E_i \\
&= 200 - (78.69 + 76.68) \\
&= 44.63.
\end{aligned}
$$

[Não é necessário fazer qualquer agrupamento de classes uma vez que em pelo menos 80% das classes se verifica $E_i \geq 5$ e que $E_i \geq 1$ para todo o i. Caso fosse preciso efectuar agrupamento de classes, os valores de k e $c = F^{-1}_{\chi^2_{(k-\beta-1)}}(1-\alpha_0)$ teriam que ser recalculados...]

- **Região de rejeição de** H_0 (para valores de T)

  Trata-se de um teste de ajustamento, logo a região de rejeição de H_0 escrita para valores de T é o intervalo à direita $W = (c, +\infty)$, onde

$$
c = F^{-1}_{\chi^2_{(k-\beta-1)}}(1-\alpha_0) = F^{-1}_{\chi^2_{(3-0-1)}}(1-0.05) = F^{-1}_{\chi^2_{(2)}}(0.95) \overset{tabela/calc.}{=} 5.991.
$$

- **Decisão**

i	Classe i	Freq. abs. obs. o_i	Freq. abs. esp. sob H_0 E_i	Parcelas valor obs. estat. teste $\frac{(o_i-E_i)^2}{E_i}$
1	$[0,5]$	76	78.69	$\frac{(76-78.69)^2}{78.69} \simeq 0.092$
2	$]5,15]$	54	76.68	6.708
3	$]15,+\infty[$	70	44.63	14.422
		$\sum_{i=1}^k o_i = n$ $= 200$	$\sum_{i=1}^k E_i = n$ $= 200$	$t = \sum_{i=1}^k \frac{(o_i-E_i)^2}{E_i}$ $\simeq 21.222$

Como $t \simeq 21.222 \in W = (5.991, +\infty)$, devemos rejeitar H_0 ao n.s. de $\alpha_0 = 5\%$. [ou a qualquer outro n.s. superior a 5%].

2. Num conjunto de dados encontra-se registado o número de anos de escolaridade (x) e o valor da remuneração mensal (em milhares de euros) (Y) para uma amostra casual de 102 trabalhadores. Para analisar o efeito do número de anos de escolaridade na remuneração mensal, usou-se um modelo de regressão linear simples de Y em x e foram obtidos os seguintes resultados:

$$
\sum_{i=1}^{102} x_i = 1094.7, \quad \sum_{i=1}^{102} x_i^2 = 12502.01, \quad \sum_{i=1}^{102} y_i = 151.4, \quad \sum_{i=1}^{102} y_i^2 = 311.42, \quad \hat{\beta}_1 \simeq 0.195422,
$$

onde $[\min_{i=1,\ldots,102} x_i, \max_{i=1,\ldots,102} x_i] = [6.4, 16.0]$.

(a) Confirme que $\sum_{i=1}^n x_i y_i \simeq 1772.09$ e obtenha a estimativa de mínimos quadrados do valor esperado da remuneração mensal de um trabalhador com 12 anos de escolaridade. (2.0)

- **Valor de** $\sum_{i=1}^n x_i y_i$ **e estimativa de MQ de** $E(Y \mid x) = \beta_0 + \beta_1 x$ **com** $x = 12$

  Como

  $n = 102$

  $\sum_{i=1}^n x_i = 1094.7$

  $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1094.7}{102} \simeq 10.732353$

  $\sum_{i=1}^n x_i^2 = 12502.01$

  $\sum_{i=1}^n x_i^2 - n(\bar{x})^2 \simeq 12502.01 - 102 \times 10.732353^2 \simeq 753.303235$

$\sum_{i=1}^{n} y_i = 151.4$
$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{151.4}{102} = 1.484314$
$\sum_{i=1}^{n} y_i^2 = 311.42$
$\sum_{i=1}^{n} y_i^2 - n\,(\bar{y})^2 \simeq 311.42 - 102 \times 1.484314^2 = 86.694902$

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,(\bar{x})^2} \simeq 0.195422$

temos

$$\begin{aligned}
\sum_{i=1}^{n} x_i y_i &= \hat{\beta}_1 \times \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) + n\bar{x}\bar{y} \\
&\simeq 0.195422 \times 753.303235 + 102 \times 10.732353 \times 1.484314 \\
&\simeq 1772.09.
\end{aligned}$$

Mais, as estimativas de MQ de β_0 e $\beta_0 + 12\,\beta_1$ são, para este modelo de RLS, iguais a:

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \times \bar{x} \\
&\simeq 1.484314 - 0.195422 \times 10.732353 \\
&\simeq -0.613024 \\
\hat{\beta}_0 + 12\,\hat{\beta}_1 &\simeq -0.613024 + 12 \times 0.195422 \\
&\simeq 1.732040.
\end{aligned}$$

(b) Após ter enunciado as hipóteses de trabalho que entender convenientes, obtenha e interprete um intervalo de confiança a 95% para β_1. (3.0)

- **Hipóteses de trabalho**
  $\epsilon_i \stackrel{i.i.d.}{\sim} \text{Normal}(0, \sigma^2)$, $i = 1, \ldots, n$

- **Obtenção do IC para β_1**

Passo 1 — V.a. fulcral para β_1

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}}} \sim t_{(n-2)}$$

Passo 2 — Quantis de probabilidade
Como $n = 102$ e $(1-\alpha) \times 100\% = 95\%$, usaremos os quantis de probabilidade simétricos $a_\alpha = -b_\alpha$ dados por:

$$\begin{cases}
a_\alpha = F^{-1}_{t_{(n-2)}}(\alpha/2) = -F^{-1}_{t_{(102-2)}}(1-0.05/2) = -F^{-1}_{t_{(100)}}(0.975) \stackrel{tabela/calc.}{=} -1.984 \\
b_\alpha = F^{-1}_{t_{(102-2)}}(1-0.05/2) = F^{-1}_{t_{(100)}}(0.975) \stackrel{tabela/calc.}{=} 1.984.
\end{cases}$$

Passo 3 — Inversão da desigualdade $a_\alpha \leq Z \leq b_\alpha$

$$P(a_\alpha \leq Z \leq b_\alpha) = 1 - \alpha$$

$$P\left[a_\alpha \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}}} \leq b_\alpha\right] = 1 - \alpha$$

$$P\left[\hat{\beta}_1 - F^{-1}_{t_{(n-2)}}(1-\alpha/2) \times \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}} \leq \beta_1 \leq \hat{\beta}_1 + F^{-1}_{t_{(n-2)}}(1-\alpha/2) \times \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}}\right] = 1 - \alpha$$

Passo 4 — Concretização
Uma vez que a estimativa de σ^2 é igual a

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-2}\left[\left(\sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2\right) - (\hat{\beta}_1)^2\left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2\right)\right] \\
&\simeq \frac{1}{102-2}\left(86.694902 - 0.195422^2 \times 753.303235\right) \\
&\simeq 0.579264
\end{aligned}
$$

e a expressão geral do IC pretendido é

$$
IC_{(1-\alpha)\times 100\%}(\beta_1) = \left[\hat{\beta}_1 \pm F^{-1}_{t_{(n-2)}}(1-\alpha/2) \times \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}}\right],
$$

temos

$$
\begin{aligned}
IC_{95\%}(\beta_1) &\simeq \left[0.195422 \pm 1.984 \times \sqrt{\frac{0.579264}{753.303235}}\right] \\
&\simeq [0.195422 \pm 0.055017] \\
&\simeq [0.140405, 0.250439].
\end{aligned}
$$

- **Comentário**
  Com base nos dados recolhidos e considerando um nível de confiança de 95%, estima-se que a um aumento de um ano na escolaridade corresponda um aumento entre 0.140405 e 0.250439 (milhares de euros) no valor esperado da remuneração mensal.

(c) Calcule o valor do coeficiente de determinação do modelo ajustado e interprete o valor obtido. (1.0)

- **Cálculo do coeficiente de determinação**
  O coeficiente de determinação pedido é igual a

$$
\begin{aligned}
r^2 &= \frac{\left(\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}\right)^2}{\left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2\right) \times \left(\sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2\right)} \\
&\simeq \frac{(1772.09 - 102 \times 10.732353 \times 1.484314)^2}{753.303235 \times 86.694902} \\
&\simeq \frac{147.211765^2}{753.303235 \times 86.694902} \\
&\simeq 0.331835
\end{aligned}
$$

- **Interpretação coeficiente de determinação**
  Cerca de 33.2% da variação total da variável resposta Y é explicada pela variável x através do modelo de regressão linear simples ajustado, donde possamos afirmar que a recta estimada parece não se ajustar bem ao conjunto de dados.