Probabilidades e Estatística

TODOS OS CURSOS

2º semestre – 2017/2018
04/07/2018 – **15:00**

Duração: 90 minutos **2º Teste C**

Justifique convenientemente todas as respostas

Grupo I	10 valores

1. Admita que os tempos (em centena de hora) entre avarias consecutivas de um sistema são variáveis aleatórias independentes e identicamente distribuídas à variável aleatória X com função de densidade de probabilidade dada por

$$f_X(x) = \begin{cases} \lambda^2\, x\, e^{-\lambda x}, & x > 0 \\ 0, & \text{caso contrário,} \end{cases}$$

onde λ é um parâmetro desconhecido positivo.

(a) Deduza o estimador de máxima verosimilhança de λ com base em $(X_1, \ldots, X_n)$, uma amostra aleatória de dimensão n de X. (3.0)

- **V.a. de interesse**
  X = tempo (em centenas de hora) entre avarias consecutivas de um sistema
- **F.d.p. de** X
  $$f(x) = \begin{cases} \lambda^2\, x\, e^{-\lambda x}, & x > 0 \\ 0, & \text{caso contrário,} \end{cases}$$
- **Parâmetro desconhecido**
  $\lambda, \quad \lambda > 0$
- **Amostra**
  $\underline{x} = (x_1, \ldots, x_n)$ amostra de dimensão n proveniente da população X.
- **Obtenção do estimador de MV de** λ

Passo 1 — Função de verosimilhança

$$\begin{aligned} L(\lambda|\underline{x}) &= f_{\underline{X}}(\underline{x}) \\ &\stackrel{X_i\, indep}{=} \prod_{i=1}^{n} f_{X_i}(x_i) \\ &\stackrel{X_i \sim X}{=} \prod_{i=1}^{n} f_X(x_i) \\ &= \prod_{i=1}^{n} \left[\lambda^2\, x_i\, e^{-\lambda x_i}\right] \\ &= \lambda^{2n} \left(\prod_{i=1}^{n} x_i\right) e^{-\lambda \sum_{i=1}^{n} x_i}, \quad \lambda > 0 \end{aligned}$$

Passo 2 — Função de log-verosimilhança

$$\ln L(\lambda|\underline{x}) = 2n \ln(\lambda) + \sum_{i=1}^{n} \ln(x_i) - \lambda \sum_{i=1}^{n} x_i$$

Passo 3 — Maximização

A estimativa de MV de λ é doravante representada por $\hat{\lambda}$ e

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

$$\begin{cases} \frac{2n}{\hat{\lambda}} - \sum_{i=1}^{n} x_i = 0 \\ -\frac{2n}{\hat{\lambda}^2} < 0 \end{cases}$$

$$\hat{\lambda} \quad : \quad \begin{cases} \hat{\lambda} = \frac{2n}{\sum_{i=1}^{n} x_i} \\ -\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{2n} < 0 \qquad \text{(proposição verdadeira já que } \sum_{i=1}^{n} x_i > 0\text{).} \end{cases}$$

Passo 4 — Estimador de MV de λ

$$EMV(\lambda) = \frac{2n}{\sum_{i=1}^{n} X_i} \qquad \left[\;= \frac{2}{\bar{X}}\; \right].$$

(b) A concretização de uma amostra aleatória de dimensão 20 de X conduziu a $\sum_{i=1}^{20} x_i = 203.8$. Obtenha a estimativa de máxima verosimilhança do valor esperado do tempo entre avarias consecutivas do sistema, $E(X) = \frac{2}{\lambda}$. (1.5)

- **Estimativa de MV de λ**

$$\begin{aligned} \hat{\lambda} &= \frac{2n}{\sum_{i=1}^{n} x_i} \qquad [= 2/\bar{x}] \\ &= \frac{2 \times 20}{203.8} \\ &\simeq 0.196271 \end{aligned}$$

- **Outro parâmetro desconhecido**

$$h(\lambda) = \frac{2}{\lambda}$$

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

$$\begin{aligned} \widehat{h(\lambda)} &= h(\hat{\lambda}) \\ &= \frac{2}{\hat{\lambda}} \\ &\simeq \frac{2}{0.196271} \\ &\simeq 10.19 \qquad \left[= \widehat{h(\lambda)} = h(\hat{\lambda}) = \frac{2}{\hat{\lambda}} = \frac{2}{2/\bar{x}} = \bar{x} = 203.8/20 = 10.19\right]. \end{aligned}$$

2. Uma empresa de construção civil produz provetes de betão. A tensão de compressão máxima a que estas provetes resistem (em mPa) é descrita pela variável aleatória X, possuindo distribuição normal com valor esperado (μ) e variância (σ^2) desconhecidos. Antes de executar uma laje, a empresa avalia a tensão de compressão máxima de 12 provetes de betão, tendo obtido $\sum_{i=1}^{12} x_i = 308.9$ e $\sum_{i=1}^{12} x_i^2 = 8024.67$. Com base nestes valores:

(a) Construa um intervalo de confiança a 95% para σ^2. (2.5)

- **V.a. de interesse**
  X = tensão de compressão máxima

- **Situação**
  $X \sim \text{normal}(\mu, \sigma^2)$
  μ desconhecido
  σ^2 DESCONHECIDO

- **Obtenção de IC para σ**

  Passo 1 — Selecção da v.a. fulcral para σ

$$Z = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

  [uma vez que é suposto determinar um IC para a variância de uma população normal, com valor esperado desconhecido.]

Passo 2 — Obtenção dos quantis de probabilidade
Dado que $n = 12$ e $(1-\alpha) \times 100\% = 95\%$, usaremos os quantis

$$(a_\alpha, b_\alpha) \quad : \quad \begin{cases} P(a_\alpha \leq Z \leq b_\alpha) = 1-\alpha \\ P(Z < a_\alpha) = P(Z > b_\alpha) = \alpha/2. \end{cases}$$

$$\begin{cases} a_\alpha = F^{-1}_{\chi^2_{(n-1)}}(\alpha/2) = F^{-1}_{\chi^2_{(11)}}(0.025) \stackrel{tabela/calc.}{=} 3.816 \\ b_\alpha = F^{-1}_{\chi^2_{(n-1)}}(1-\alpha/2) = F^{-1}_{\chi^2_{(11)}}(0.975) \stackrel{tabela/calc.}{=} 21.92. \end{cases}$$

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

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

$$P\left[a_\alpha \leq \tfrac{(n-1)S^2}{\sigma^2} \leq b_\alpha\right] = 1-\alpha$$

$$P\left[\tfrac{1}{b_\alpha} \leq \tfrac{\sigma^2}{(n-1)S^2} \leq \tfrac{1}{a_\alpha}\right] = 1-\alpha$$

$$P\left[\tfrac{(n-1)S^2}{b_\alpha} \leq \sigma^2 \leq \tfrac{(n-1)S^2}{a_\alpha}\right] = 1-\alpha$$

Passo 4 — Concretização
Tendo em conta que

$$IC_{(1-\alpha)\times 100\%}(\sigma^2) = \left[\frac{(n-1)\,s^2}{F^{-1}_{\chi^2_{(n-1)}}(1-\alpha/2)}, \frac{(n-1)\,s^2}{F^{-1}_{\chi^2_{(n-1)}}(\alpha/2)}\right],$$

o par de quantis acima e o facto de

$$\begin{aligned} s^2 &= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\,(\bar{x})^2\right] \\ &= \frac{1}{12-1}\left\{8024.67 - 12 \times [25.741(6)]^2\right\} \\ &= 6.642652 \end{aligned}$$

temos:

$$\begin{aligned} IC_{95\%}(\sigma^2) &\simeq \left[\frac{(12-1)\times 6.642652}{21.92}, \frac{(12-1)\times 6.642652}{3.816}\right] \\ &\simeq [3.333447, 19.148104]. \end{aligned}$$

(b) Confronte as hipóteses $H_0 : \sigma^2 = 21.5$ e $H_1 : \sigma^2 < 21.5$. Decida com base no valor-p. (3.0)

- **Hipóteses**
  $H_0 : \sigma^2 = \sigma_0^2 = 21.5$
  $H_1 : \sigma^2 < \sigma_0^2$
- **Estatística de teste**

  $$T = \frac{(n-1)S^2}{\sigma_0^2} \sim_{H_0} \chi^2_{(n-1)}$$

  [pois pretendemos efectuar teste sobre a variância de pop. normal, com valor esperado desc.]
- **Região de rejeição de H_0** (para valores de T)
  Tratando-se de um teste unilateral inferior ($H_1 : \sigma^2 < \sigma_0^2$), a região de rejeição de H_0 é do tipo $W = (0, c)$.
- **Decisão (com base no valor-p)**
  O valor observado da estatística de teste é dado por

  $$t = \frac{(n-1)s^2}{\sigma_0^2} \simeq \frac{(12-1)6.642652}{21.5} \simeq 3.39857.$$

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

$$\begin{aligned} valor-p &= P(T < t \mid H_0) \\ &= F_{\chi^2_{(n-1)}}(t) \\ &\simeq F_{\chi^2_{(12-1)}}(3.39857) \\ &\stackrel{calc.}{=} 0.015658. \end{aligned}$$

Consequentemente, é suposto:

- não rejeitar H_0 a qualquer n.s. $\alpha_0 \le 1.5658\%$, nomeadamente ao n.u.s. de 1%;
- rejeitar H_0 a qualquer n.s. $\alpha_0 > 1.5658\%$, por exemplo aos n.u.s. de 5% e 10%.

[Decisão (com base em intervalo para o valor-p)
Recorrendo às tabelas de quantis da distribuição do qui-quadrado podemos adiantar um intervalo para o valor-p:

$$\begin{array}{rcccl} F^{-1}_{\chi^2_{(11)}}(0.01) = 3.053 & < & 3.39857 & < & 3.816 = F^{-1}_{\chi^2_{(11)}}(0.025) \\ 0.01 & < & valor-p = F_{\chi^2_{(11)}}(3.39857) & < & 0.025. \end{array}$$

Consequentemente:

- não devemos rejeitar H_0 a qualquer n.s. $\alpha_0 \le 1\%$, nomeadamente ao n.u.s. de 1%;
- devemos rejeitar H_0 a qualquer n.s. $\alpha_0 \ge 5\%$, por exemplo aos n.u.s. de 5% e 10%.]

Grupo II	10 valores

1. Seja X a variável aleatória que descreve a tensão de rotura (em mPa) de uma resina fenólica. Uma engenheira de materiais defende a hipótese H_0 de que X possui função de distribuição dada por

$$F(x) = \begin{cases} 1 - e^{-\left(\frac{x}{75}\right)^2}, & x > 0 \\ 0, & \text{caso contrário.} \end{cases}$$

Um ensaio de flexão, envolvendo 100 corpos de prova selecionados casualmente, conduziu à seguinte tabela de frequências:

Classe	$]0,40]$	$]40,60]$	$]60,80]$	$]80,+\infty[$
Frequência absoluta observada	12	33	30	25
Frequência absoluta esperada sob H_0	24.76	22.51	E_3	E_4

(a) Obtenha os valores de E_3 e E_4 (aproximando-os às centésimas). (1.0)

- **V.a. de interesse**
  $X =$ tensão de rotura (em mPa) de uma resina fenólica

- **F.d. conjecturada**
  $$F(x) = \begin{cases} 1 - e^{-\left(\frac{x}{75}\right)^2}, & x > 0 \\ 0, & \text{caso contrário.} \end{cases}$$

- **Frequências absolutas esperadas**
  Atendendo à dimensão da amostra $n = 100$ e à f.d. conjecturada, segue-se, para $i = 1, \ldots, 4$:

$$\begin{aligned} E_3 &= n \times [F(80) - F(60)] \\ &= 100 \times \left\{ \left[1 - e^{-\left(\frac{80}{75}\right)^2}\right] - \left[1 - e^{-\left(\frac{60}{75}\right)^2}\right] \right\} \\ &= 20.68; \\ E_4 &= n - \sum_{i=1}^{3} E_i \\ &= 100 - (24.76 + 22.51 + 20.68) \\ &= 32.05. \end{aligned}$$

(b) Teste H_0, ao nível de significância de 10%. (3.0)

- **Hipóteses**
  $H_0 : X$ com f.d. $F_X(x) = F(x),\ x \in \mathbb{R}$
  $H_1 : X$ com f.d. $F_X(x) \neq F(x),$ para algum $x \in \mathbb{R}$

- **Nível de significância**
  $\alpha_0 = 10\%$

- **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 = 4

  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**
  De acordo com (a), os valores das freq. absolutas esperadas sob H_0 são: $E_1 = 24.76$, $E_2 = 22.51$, $E_3 = 20.68$ e $E_4 = 32.05$.
  [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

$$c \;=\; F^{-1}_{\chi^2_{(k-\beta-1)}}(1-\alpha_0) = F^{-1}_{\chi^2_{(4-0-1)}}(1-0.1) = F^{-1}_{\chi^2_{(3)}}(0.9) \overset{tabela/calc.}{=} 6.251.$$

- **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	$]0,40]$	12	24.76	$\frac{(12-24.76)^2}{24.76} \simeq 6.576$
2	$]40,60]$	33	22.51	4.888
3	$]60,80]$	30	20.68	4.200
4	$]80,+\infty[$	25	32.05	1.551
		$\sum_{i=1}^{k} o_i = n$ $= 100$	$\sum_{i=1}^{k} E_i = n$ $= 100$	$t = \sum_{i=1}^{k} \frac{(o_i-E_i)^2}{E_i}$ $\simeq 17.215$

  Como $t \simeq 17.215 \in W = (6.251, +\infty)$, devemos rejeitar H_0 ao n.s. de $\alpha_0 = 10\%$. [ou a qualquer outro n.s. superior a 10%].

2. Medições efetuadas em 12 indivíduos conduziram aos seguintes resultados referentes ao nível de colesterol Y (em mg/100ml) e à idade x (em anos) de um indivíduo:

$$\sum_{i=1}^{12} x_i = 619,\quad \sum_{i=1}^{12} x_i^2 = 34\,059,\quad \sum_{i=1}^{12} y_i = 2\,436,\quad \sum_{i=1}^{12} y_i^2 = 538\,156,\quad \sum_{i=1}^{12} x_i y_i = 132\,520,$$

onde $\left[\min_{i=1,\dots,12} x_i,\ \max_{i=1,\dots,12} x_i\right] = [33\,,76]$.

(a) Considere o modelo de regressão linear simples de Y em x e determine a estimativa de mínimos quadrados do valor esperado do nível de colesterol para um indivíduo com 50 anos. (2.0)

- **Estimativa de MQ de $E(Y \mid x) = \beta_0 + \beta_1\, x$ com $x = 50$**
  Uma vez que

  $n = 12$

  $\sum_{i=1}^n x_i = 619$
  $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i = \frac{619}{12} = 51.58(3)$
  $\sum_{i=1}^n x_i^2 = 34\,059$
  $\sum_{i=1}^n x_i^2 - n\,(\bar{x})^2 = 34\,059 - 12 \times [51.58(3))^2 = 2\,128.91(6)$

  $\sum_{i=1}^n y_i = 2\,436$
  $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i = \frac{2436}{12} = 203$
  $\sum_{i=1}^n y_i^2 = 538\,156$
  $\sum_{i=1}^n y_i^2 - n\,(\bar{y})^2 = 538\,156 - 12 \times 203^2 = 43\,648$

  $\sum_{i=1}^n x_i\,y_i = 132\,520$
  $\sum_{i=1}^n x_i\,y_i - n\,\bar{x}\,\bar{y} = 132\,520 - 12 \times 51.58(3) \times 203 = 6\,863,$

  as estimativas de MQ de β_1, β_0 e $\beta_0 + \beta_1\, x$ são, para este modelo de RLS, iguais a:

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i\,y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^n x_i^2 - n\,(\bar{x})^2} \\
&= \frac{6\,863}{2128.91(6)} \\
&\simeq 3.223705 \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \times \bar{x} \\
&\simeq 203 - 3.223705 \times 51.58(3) \\
&\simeq 36.710550 \\
\hat{\beta}_0 + \hat{\beta}_1\, x &\simeq 36.710550 + 3.223705 \times 50 \\
&\simeq 197.895800.
\end{aligned}
$$

(b) Após ter enunciado as hipóteses de trabalho que entender convenientes, teste a hipótese $H_0 : \beta_1 = 0$, ao nível de significância de 1%. (3.0)

- **Hipóteses de trabalho**
  [$Y_i = \beta_0 + \beta_1\, x_i + \epsilon_i,\ i = 1, 2, \ldots, n,$ com]
  $\epsilon_i \overset{i.i.d.}{\sim}$ Normal$(0, \sigma^2)$, $i = 1, \ldots, n$

- **Hipóteses**
  $H_0 : \beta_1 = \beta_{1,0} = 0$
  $H_1 : \beta_1 \neq 0$

- **Nível de significância**
  $\alpha_0 = 1\%$

- **Estatística de teste**

$$
T = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n x_i^2 - n\,\bar{x}^2}}} \sim_{H_0} t_{(n-2)}
$$

- **Região de rejeição de H_0** (para valores da estatística de teste)
  Estamos a lidar com um teste bilateral ($H_1 : \beta_1 \neq 0$), pelo que a região de rejeição de H_0 é uma reunião de intervalos do tipo $W = (-\infty, -c) \cup (c, +\infty)$, onde $c : P(\text{Rejeitar } H_0 \mid H_0) = \alpha_0$, i.e.,

$$\begin{aligned} c &= F^{-1}_{t_{(n-2)}}(1-\alpha_0/2) \\ &= F^{-1}_{t_{(12-2)}}(1-0.01/2) \\ &= F^{-1}_{t_{(10)}}(0.995) \\ &\overset{tabela/calc.}{=} 3.169. \end{aligned}$$

- **Decisão**

  Tendo em conta os valores obtidos em (a), bem como o de

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-2}\left[\left(\sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2\right) - \left(\hat{\beta}_1\right)^2 \left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2\right)\right] \\ &\simeq \frac{1}{12-2}\left(43\,648 - 3.223705^2 \times 2\,128.91(6)\right) \\ &\simeq 2\,152.371483, \end{aligned}$$

  o valor observado da estatística de teste é igual a

$$\begin{aligned} t &= \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}}} \\ &= \frac{3.223705 - 0}{\sqrt{\frac{2\,152.371483}{2\,128.91(6)}}} \\ &= 3.206092. \end{aligned}$$

  Como $t = 3.206092 \in W = (-\infty, -3.169) \cup (3.169, +\infty)$ devemos rejeitar H_0 ao n.s. de 1% [bem como a qualquer n.s. superior a 1%. Com efeito, podemos concluir que devemos rejeitar a hipótese de o valor esperado da variável aleatória Y não ser função linear da variável explicativa x.]

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

- **Cálculo do coeficiente de determinação**

$$\begin{aligned} r^2 &= \frac{\left(\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}\right)^2}{\left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2\right) \times \left(\sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2\right)} \\ &= \frac{6\,863^2}{2128.91(6) \times 43\,648} \\ &\simeq 0.506880. \end{aligned}$$

- **Interpretação coeficiente de determinação**

  Cerca de 50.7% 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 ajustar-se razoavelmente ao conjunto de dados.