Probabilidades e Estatística

LEGM, LEIC-A, LEIC-T, MA, MEMec

2º semestre – 2017/2018
14/06/2018 – **11:00**

Duração: 90 minutos **2º Teste B**

Justifique convenientemente todas as respostas

Grupo I 10 valores

1. Os indivíduos com idade superior a 110 anos são designados de supercentenários. Um estudo recente considera que a idade (em anos) de um supercentenário é descrita pela variável aleatória X com função de densidade de probabilidade

$$f_X(x) = \begin{cases} \lambda\, e^{-\lambda\,(x-110)}, & x \geq 110 \\ 0, & \text{caso contrário,} \end{cases}$$

onde λ é um parâmetro desconhecido positivo.

(a) Deduza o estimador de máxima verosimilhança de λ com base numa amostra aleatória de dimensão n proveniente desta população. (3.0)

- **V.a. de interesse**
  X = idade (em anos) de um supercentenário
- **F.d.p. de** X
  $$f_X(x) = \begin{cases} \lambda\, e^{-\lambda\,(x-110)}, & x \geq 110 \\ 0, & x < 110 \end{cases}$$
- **Parâmetro desconhecido**
  $\lambda, \quad \lambda > 0$
- **Amostra**
  $\underline{x} = (x_1, \ldots, x_n)$ amostra de dimensão n proveniente da população X [com $x_i > 110, i = 1, \ldots, n$].

- **Obtenção do estimador de MV de** λ

Passo 1 — Função de verosimilhança

$$\begin{aligned} L(\lambda|\underline{x}) &= f_{\underline{X}}(\underline{x}) \\ &\stackrel{X_i\, indep}{=} \prod_{i=1}^{n} f_{X_i}(x_i) \\ &\stackrel{X_i \sim X}{=} \prod_{i=1}^{n} f_X(x_i) \\ &= \prod_{i=1}^{n} \left[\lambda\, e^{-\lambda\,(x_i-110)}\right] \\ &= \lambda^n\, e^{-\lambda \sum_{i=1}^{n}(x_i-110)}, \quad \lambda > 0 \end{aligned}$$

Passo 2 — Função de log-verosimilhança

$$\ln L(\lambda|\underline{x}) = n \ln(\lambda) - \lambda \left[\sum_{i=1}^{n} (x_i - 110)\right]$$

Passo 3 — Maximização

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

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

$$\begin{cases} \frac{n}{\hat{\lambda}} - \sum_{i=1}^{n}(x_i - 110) = 0 \\ -\frac{n}{\hat{\lambda}^2} < 0 \end{cases}$$

$$\hat{\lambda} \quad : \quad \begin{cases} \hat{\lambda} = \frac{n}{\sum_{i=1}^{n}(x_i-110)} \\ -\frac{[\sum_{i=1}^{n}(x_i-110)]^2}{n} < 0 \qquad \text{(proposição verdadeira já que } \sum_{i=1}^{n}(x_i-110) > 0). \end{cases}$$

Passo 4 — Estimador de MV de λ

$$EMV(\lambda) = \frac{n}{\sum_{i=1}^{n}(X_i - 110)} \qquad [= (\bar{X} - 110)^{-1}]$$

(b) Recolheu-se uma amostra $(x_1, \ldots, x_5)$, tendo-se obtido $\sum_{i=1}^{5} x_i = 573.9$. Obtenha a estimativa de máxima verosimilhança da probabilidade de um supercentenário com idade x viver ainda mais um ano, isto é, da probabilidade $P(X > x+1 \mid X > x) = e^{-\lambda}$ com $x > 110$. (1.5)

- **Estimativa de MV de λ**

$$\begin{aligned} \hat{\lambda} &= \frac{n}{\sum_{i=1}^{n}(x_i - 110)} \\ &= \frac{5}{573.9 - 5 \times 110} \\ &\simeq 0.209205 \end{aligned}$$

- **Outro parâmetro desconhecido**

$$\begin{aligned} h(\lambda) &= P(X > x+1 \mid X > x) \\ &= e^{-\lambda} \end{aligned}$$

- **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

$$\begin{aligned} \widehat{h(\lambda)} &= h(\hat{\lambda}) \\ &= e^{-\hat{\lambda}} \\ &\simeq e^{-0.209205} \\ &\simeq 0.811229. \end{aligned}$$

2. De modo a comparar a popularidade de dois sites da internet, considerou-se a variável aleatória X_1 (respetivamente X_2) que representa o número de acessos semanais ao site 1 (respetivamente ao site 2). Ao selecionarem-se casualmente 41 registos semanais de cada um dos dois sites, obtiveram-se os seguintes resultados: $\bar{x}_1 = 2952.8$, $s_1^2 = 3307.53$, $\bar{x}_2 = 3002.4$, $s_2^2 = 3100.20$.

(a) Determine um intervalo de confiança aproximado a 90% para $\mu_1 - \mu_2 = E(X_1) - E(X_2)$. (2.5)

- **V.a. de interesse**
  X_i = número de acessos semanais ao site i, $(i = 1, 2)$

- **Situação**
  X_i v.a. com dist. arbitrária, valor esperado μ_i e variância σ_i^2 $(i = 1, 2)$

  $X_1 \perp\!\!\!\perp X_2$

  $(\mu_1 - \mu_2)$ DESCONHECIDO

  σ_1^2 e σ_2^2 desconhecidas [não necessariamente iguais]

  $n_1 = n_2 = 41 \geq 30$ [i.e., ambas as amostras possuem dimensão suficientemente grande].

- **Obtenção do IC para $\mu_1 - \mu_2$**
  Passo 1 — Selecção da v.a. fulcral para $\mu_1 - \mu_2$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{a}{\sim} \text{normal}(0, 1)$$

[dado que se pretende determinar um IC para a diferença de valores esperados de duas populações com distribuições arbitrárias independentes e com variâncias desconhecidas e dispomos de duas amostras com dimensões suficientemente grandes.]

Passo 2 — Obtenção dos quantis de probabilidade
Ao ter-se em consideração que $(1-\alpha) \times 100\% = 90\%$, far-se-á uso dos quantis

$$\begin{cases} a_\alpha = \Phi^{-1}(\alpha/2) = \Phi^{-1}(0.05) = -\Phi^{-1}(1-0.05) \stackrel{tabela/calc.}{=} -1.6449 \\ b_\alpha = \Phi^{-1}(1-\alpha/2) = \Phi^{-1}(0.95) \stackrel{tabela/calc.}{=} 1.6449. \end{cases}$$

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

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

$$P\left[a_\alpha \le \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \le b_\alpha\right] \simeq 1-\alpha$$

$$P\left[(\bar{X}_1-\bar{X}_2) - b_\alpha \times \sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \le \mu_1-\mu_2 \le (\bar{X}_1-\bar{X}_2) - a_\alpha \times \sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right] \simeq 1-\alpha$$

Passo 4 — Concretização
Tendo em conta que

$$IC_{(1-\alpha)\times 100\%}(\mu_1-\mu_2) = \left[(\bar{x}_1-\bar{x}_2) \pm \Phi^{-1}(1-\alpha/2) \times \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right],$$

bem como os valores dos quantis acima e de n_i, $\bar{x}_i$, s_i^2 $(i=1,2)$, segue-se

$$\begin{aligned} IC_{95\%}(\mu_1-\mu_2) &= \left[(2952.8-3002.4) \pm 1.6449 \times \sqrt{\frac{3307.53}{41}+\frac{3100.2}{41}}\right] \\ &\simeq [-49.6 \pm 1.6449 \times 12.501449] \\ &\simeq [-70.163625,\ -29.036375]. \end{aligned}$$

(b) Confronte as hipóteses $H_0: \mu_1-\mu_2 = 0$ e $H_1: \mu_1-\mu_2 < 0$, calculando para o efeito o valor-p. (3.0)

- **V.a. de interesse e situação**
  Ver alínea (a).
- **Hipóteses**
  $H_0: \mu_1-\mu_2 = \mu_0 = 0$
  $H_1: \mu_1-\mu_2 < \mu_0$
- **Estatística de teste**

$$T = \frac{(\bar{X}_1-\bar{X}_2)-\mu_0}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \stackrel{a}{\sim}_{H_0} \text{normal}(0,1)$$

  [uma vez que se pretende efectuar um teste de igualdade dos valores esperados de duas populações com distribuições arbitrárias independentes e com variâncias desconhecidas, dispondo de duas amostras com dimensões suficientemente grandes.]
- **Região de rejeição de H_0** (para valores de T)
  Estamos a lidar com um teste unilateral inferior $(H_1: \mu_1-\mu_2 < \mu_0)$, logo a região de rejeição de H_0 é do tipo $W = (-\infty, -c)$.
- **Decisão (com base no valor-p)**
  O valor observado da estatística de teste é igual a

$$
\begin{aligned}
t &= \frac{(\bar{x}_1-\bar{x}_2)-\mu_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}\\
&= \frac{(2952.8-3002.4)-0}{\sqrt{\frac{3307.53}{41}+\frac{3100.2}{41}}}\\
&\simeq \frac{-49.6}{12.501449}\\
&\simeq -3.967540.
\end{aligned}
$$

e a região de rejeição deste teste é um intervalo à esquerda. Consequentemente

$$
\begin{aligned}
valor-p &= P(T<t \mid H_0)\\
&\simeq \Phi(t)\\
&\simeq \Phi(-3.97)\\
&= 1-\Phi(3.97)\\
&\overset{tabela/calc.}{=} 1-0.999964\\
&= 0.000036.
\end{aligned}
$$

Logo é suposto:

- não rejeitar H_0 a qualquer nível de significância $\alpha_0 \leq 0.0036\%$;
- rejeitar H_0 a qualquer n.s. $\alpha_0 > 0.0036\%$, nomeadamente a qualquer dos níveis usuais de significância (1%, 5% e 10%).

Grupo II	10 valores

1. Uma estudante recorreu a um *software* estatístico para gerar 1800 números pseudo-aleatórios no intervalo [0, 1], tendo obtido a seguinte tabela de frequências:

Classe	$[0,0.2]$	$]0.2,0.5]$	$]0.5,0.8]$	$]0.8,1]$
Frequência absoluta observada	391	490	580	339

A estudante defende a hipótese H_0 de que o *software* gerou números pseudo-aleatórios que seguem uma distribuição uniforme contínua no intervalo [0, 1].

(a) Calcule os valores das frequências absolutas esperadas sob H_0 de cada uma das classes. (1.0)

- **V.a. de interesse**
  X = número pseudo-aleatório gerado pelo *software* estatístico

- **Distribuição, f.d.p. e f.d. conjecturadas**
  $X \sim$ uniforme contínua$(0,1)$

  $$f_X(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \text{caso contrário} \end{cases}$$

  $$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,dt = \begin{cases} \int_{-\infty}^{x} 0\,dt = 0, & x<0 \\ \int_{-\infty}^{0} 0\,dt + \int_0^x 1\,dt = x, & 0 \leq x \leq 1 \\ \int_{-\infty}^{0} 0\,dt + \int_0^1 1\,dt + \int_1^x 0\,dt = 1, & x>1. \end{cases}$$

- **Frequências absolutas esperadas**
  Atendendo à dimensão da amostra $n = 1800$ e à f.d. conjecturada, segue-se, para $i = 1,\ldots,4$:

  $$
  \begin{aligned}
  E_1 &= n \times [F(0.2)-F(0)]\\
  &= 1800 \times (0.2-0)\\
  &= 360;
  \end{aligned}
  $$

$$
\begin{aligned}
E_2 &= n \times [F(0.5) - F(0.2)] \\
&= 1800 \times (0.5 - 0.2) \\
&= 540; \\
E_3 &= n \times [F(0.8) - F(0.5)] \\
&= 1800 \times (0.8 - 0.5) \\
&= 540; \\
E_4 &= n - \sum_{i=1}^{3} E_i \\
&= 1800 - (360 + 540 + 540) \\
&= 360.
\end{aligned}
$$

(b) Teste H_0, ao nível de significância de 5%. (3.0)

- **Hipóteses**
  $H_0 : X \sim$ uniforme contínua$(0,1)$
  $H_1 : X \not\sim$ uniforme contínua$(0,1)$

- **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 $= 4$

  $O_i =$ Frequência absoluta observável da classe i

  $E_i =$ Frequência absoluta esperada, sob H_0, da classe i

  $\beta =$ No. de parâmetros a estimar $= 0$ [dado que em H_0 se conjectura uma distribuição específica.]

- **Frequências absolutas esperadas sob H_0**
  De acordo com (a), os valores das freq. absolutas esperadas sob H_0 são: $E_1 = E_4 = 360$; $E_2 = E_3 = 540$.
  [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
  $$c = F^{-1}_{\chi^2_{(k-\beta-1)}}(1-\alpha_0) = F^{-1}_{\chi^2_{(4-0-1)}}(1-0.05) = F^{-1}_{\chi^2_{(3)}}(0.95) \overset{tabela/calc.}{=} 7.815.$$

- **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	$[0,0.2]$	391	360	$\frac{(391-360)^2}{360} \simeq 2.669$
2	$]0.2,0.5]$	490	540	4.630
3	$]0.5,0.8]$	580	540	2.963
4	$]0.8,1]$	339	360	1.225
		$\sum_{i=1}^{k} o_i = n$ $= 1800$	$\sum_{i=1}^{k} E_i = n$ $= 1800$	$t = \sum_{i=1}^{k} \frac{(o_i-E_i)^2}{E_i}$ $\simeq 11.487$

Como $t \simeq 11.487 \in W = (7.815, +\infty)$, devemos rejeitar H_0 ao n.s. de $\alpha_0 = 5\%$.

2. A determinação da resistência ao cisalhamento (Y) de soldas por pontos é relativamente difícil, enquanto que a medição do diâmetro da solda (x) é relativamente simples. Um conjunto de 10 medições independentes conduziu aos seguintes resultados respeitantes a x (em polegada) e a Y (em psi):

$$\sum_{i=1}^{10} x_i = 2.325, \quad \sum_{i=1}^{10} x_i^2 = 0.697425, \quad \sum_{i=1}^{10} y_i = 22\,860, \quad \sum_{i=1}^{10} y_i^2 = 67\,719\,400, \quad \sum_{i=1}^{10} x_i y_i = 6\,872.25,$$

onde $\left[\min_{i=1,\ldots,10} x_i, \max_{i=1,\ldots,10} x_i\right] = [0.04\,, 0.4]$.

(a) Considere o modelo de regressão linear simples de Y em x e determine a estimativa de mínimos quadrados do valor esperado da resistência ao cisalhamento de uma solda com diâmetro igual a 0.25 polegada. (2.0)

- **Estimativas de MQ de β_0, β_1 e $E(Y \mid x) = \beta_0 + \beta_1 x$ com $x = 0.25$**
  Dado que

  $n = 10$

  $\sum_{i=1}^n x_i = 2.325$
  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{2.325}{10} = 0.2325$
  $\sum_{i=1}^n x_i^2 = 0.697425$
  $\sum_{i=1}^n x_i^2 - n(\bar{x})^2 = 0.697425 - 10 \times 0.2325^2 = 0.1568625$

  $\sum_{i=1}^n y_i = 22\,860$
  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{22\,860}{10} = 2\,286$
  $\sum_{i=1}^n y_i^2 = 67\,719\,400$
  $\sum_{i=1}^n y_i^2 - n(\bar{y})^2 = 67\,719\,400 - 10 \times 2\,286^2 = 15\,461\,440$

  $\sum_{i=1}^n x_i y_i = 6\,872.25$
  $\sum_{i=1}^n x_i y_i - n\,\bar{x}\,\bar{y} = 6\,872.25 - 10 \times 0.2325 \times 2\,286 = 1\,557.3,$

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

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^n x_i^2 - n(\bar{x})^2} \\
&= \frac{1\,557.3}{0.1568625} \\
&\simeq 9927.803012 \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \times \bar{x} \\
&\simeq 2\,286 - 9927.803012 \times 0.2325 \\
&\simeq -22.2142 \\
\hat{\beta}_0 + \hat{\beta}_1 x &\simeq -22.2142 + 0.25 \times 9927.803012 \\
&\simeq 2459.736553.
\end{aligned}$$

(b) Uma engenheira mecânica defende a conjetura $H_0 : \beta_1 = 10\,000$ ao passo que um engenheiro de materiais é de opinião contrária. Após ter enunciado as hipóteses de trabalho que entender convenientes, teste H_0 ao nível de significância de 5%. (3.0)

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

- **Hipóteses**
  $H_0 : \beta_1 = \beta_{1,0} = 10\,000$
  $H_1 : \beta_1 \neq \beta_{1,0}$

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

- **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 de T)
  Estamos a lidar com um teste bilateral ($H_1 : \beta_1 \neq 10\,000$), logo a região de rejeição de H_0 é do tipo $W = (-\infty, -c) \cup (c, +\infty)$, onde $c : P(\text{Rejeitar } H_0 \mid H_0) = \alpha_0$, i.e.,

$$\begin{aligned} c &= F^{-1}_{t_{(n-2)}}(1-\alpha_0/2) \\ &= F^{-1}_{t_{(10-2)}}(1-0.05/2) \\ &= F^{-1}_{t_{(8)}}(0.975) \\ &\overset{tabela/calc.}{=} 2.306. \end{aligned}$$

- **Decisão**
  Atendendo aos valores obtidos em (a), assim como o de

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-2}\left[\left(\sum_{i=1}^n y_i^2 - n\,\bar{y}^2\right) - (\hat{\beta}_1)^2 \left(\sum_{i=1}^n x_i^2 - n\,\bar{x}^2\right)\right] \\ &\simeq \frac{1}{10-2}\left(15\,461\,440 - 9927.803012^2 \times 0.1568625\right) \\ &\simeq 109.046214, \end{aligned}$$

  o valor observado da estatística de teste é igual a

$$\begin{aligned} t &= \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n x_i^2 - n\,\bar{x}^2}}} \\ &\simeq \frac{9927.803012 - 10\,000}{\sqrt{\frac{109.046214}{0.1568625}}} \\ &= -2.738252. \end{aligned}$$

  Como $t \simeq -2.738252 \in W = (-\infty, -2.306) \cup (2.306, +\infty)$ devemos rejeitar H_0 ao n.s. de 5% [bem como a qualquer n.s. superior que 5%).

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

- **Cálculo do coeficiente de determinação**

$$\begin{aligned} r^2 &= \frac{\left(\sum_{i=1}^n x_i y_i - n\,\bar{x}\,\bar{y}\right)^2}{\left(\sum_{i=1}^n x_i^2 - n\,\bar{x}^2\right) \times \left(\sum_{i=1}^n y_i^2 - n\,\bar{y}^2\right)} \\ &= \frac{1\,557.3^2}{0.156863 \times 15\,461\,440} \\ &\simeq 0.999940. \end{aligned}$$

- **Interpretação coeficiente de determinação**
  Cerca de 99.99% 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 excepcionalmente bem ao conjunto de dados.