Probabilidades e Estatística

LEE, LEIC-A, LEIC-T, LMAC, LEMat, LERC, MEAer, MEBiol, MEBiom, MEEC, MEFT, MEMec, MEQ

1º semestre – 2019/2020
09/01/2020 – **9:00**

Duração: 90 minutos **2º Teste A**

Justifique convenientemente todas as respostas

Grupo I 10 valores

1. A variável aleatória X representa o tempo de reparação de um regulador de tensão eléctrica de um automóvel e possui função de densidade de probabilidade dada por

$$f_X(x) = \begin{cases} \frac{0.5\theta}{\sqrt{x}}\, e^{-\theta\sqrt{x}}, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

onde θ ($\theta > 0$) é um parâmetro desconhecido. Seja $(X_1, \ldots, X_n)$ uma amostra aleatória de X.

(a) Deduza o estimador de máxima verosimilhança de θ, com base na amostra aleatória acima. (3.0)

- **V.a. de interesse**
  X = tempo de reparação de um regulador de tensão eléctrica de um automóvel
- **F.d.p. de** X
  $$f_X(x) = \begin{cases} \frac{0.5\theta}{\sqrt{x}}\, e^{-\theta\sqrt{x}}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$
- **Parâmetro desconhecido**
  $\theta, \quad \theta > 0$
- **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
  $$\begin{aligned} L(\theta|\underline{x}) &= f_{\underline{X}}(\underline{x}) \\ &\overset{X_i\ indep}{=} \prod_{i=1}^{n} f_{X_i}(x_i) \\ &\overset{X_i \sim X}{=} \prod_{i=1}^{n} f_X(x_i) \\ &= \prod_{i=1}^{n} \left(\frac{0.5\theta}{\sqrt{x_i}}\, e^{-\theta\sqrt{x_i}} \right) \\ &= 0.5^n \times \theta^n \times \left(\prod_{i=1}^{n} x_i \right)^{-0.5} \times \exp\left(-\theta \sum_{i=1}^{n} \sqrt{x_i} \right), \quad \theta > 0 \end{aligned}$$

  Passo 2 — Função de log-verosimilhança
  $$\ln L(\theta|\underline{x}) = n\ln(0.5) + n\ln(\theta) - 0.5 \sum_{i=1}^{n} \ln(x_i) - \theta \sum_{i=1}^{n} \sqrt{x_i}$$

  Passo 3 — Maximização
  A estimativa de MV de θ é doravante representada por $\hat{\theta}$ e
  $$\hat{\theta} \ : \ \begin{cases} \left.\frac{d\ln L(\theta|\underline{x})}{d\theta}\right|_{\theta=\hat{\theta}} = 0 & \text{(ponto de estacionaridade)} \\ \left.\frac{d^2\ln L(\theta|\underline{x})}{d\theta^2}\right|_{\theta=\hat{\theta}} < 0 & \text{(ponto de máximo)} \end{cases}$$

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

$$\begin{cases} \hat{\theta} = \frac{n}{\sum_{i=1}^n \sqrt{x_i}} \\ -\frac{(\sum_{i=1}^n \sqrt{x_i})^2}{n} < 0 \qquad \text{(proposição verdadeira).} \end{cases}$$

Passo 4 — Estimador de MV de θ

$$EMV(\theta) = \frac{n}{\sum_{i=1}^n \sqrt{X_i}}$$

(b) Tendo-se recolhido uma concretização $(x_1, \ldots, x_{10})$ para a qual $\sum_{i=1}^{10} \sqrt{x_i} \simeq 9.615$, obtenha a estimativa de máxima verosimilhança da probabilidade de tal tempo de reparação exceder 4. (2.0)

Nota: A função de distribuição de X é dada por $F_X(x) = 1 - e^{-\theta\sqrt{x}}$, para $x \geq 0$.

- **Estimativa de MV de θ**

$$\begin{aligned} \hat{\theta} &= \frac{n}{\sum_{i=1}^n \sqrt{x_i}} \\ &\simeq \frac{10}{9.615} \\ &\simeq 1.040042 \end{aligned}$$

- **Outro parâmetro desconhecido**

$$\begin{aligned} h(\theta) &= P(X > 4) \\ &= 1 - F_X(4) \\ &= 1 - \left(1 - e^{-\theta\sqrt{4}}\right) \\ &= e^{-\theta\sqrt{4}} \end{aligned}$$

- **Estimativa de MV de $h(\theta)$**

  Invocando a propriedade de invariância dos estimadores de máxima verosimilhança, concluímos que a estimativa de MV de $h(\theta)$ é

$$\begin{aligned} \widehat{h(\theta)} &= h(\hat{\theta}) \\ &= e^{-\hat{\theta}\sqrt{4}} \\ &\simeq e^{-1.040042\times\sqrt{4}} \\ &\simeq 0.124920. \end{aligned}$$

2. Considere que a variável aleatória X_1 (respetivamente X_2) representa a concentração de PCB (policloretos de bifenilo, em partes por milhão) em peixes provenientes do rio que recebe descargas da fábrica 1 (respetivamente da fábrica 2). Ao selecionarem-se casualmente 40 peixes de cada um dos rios, obtiveram-se os seguintes resultados: $\bar{x}_1 = 6.450$, $s_1^2 = 1.451795$, $\bar{x}_2 = 7.515$, $s_2^2 = 1.641308$. Admitindo que X_1 e X_2 são variáveis aleatórias independentes:

(a) determine um intervalo de confiança aproximado a 90% para $E(X_1) - E(X_2) = \mu_1 - \mu_2$; (2.0)

- **V.a. de interesse**

  X_i = concentração de PCB em peixes prov. do rio que recebe descargas da fábrica 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]
$n_1 = n_2 = 40 > 30$ [i.e., ambas as amostras são suficientemente grandes].

- **Obtenção do IC para $\mu_1 - \mu_2$**

  Passo 1 — Selecção da v.a. fulcral para $\mu_1 - \mu_2$

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

  [dado que pretendemos determinar um IC para a diferença de valores esperados de duas populações com distribuições arbitrárias independentes e com variâncias desconhecidas, dispondo de duas amostras suficientemente grandes.]

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

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

  [que enquadram a v.a. fulcral Z com probabilidade aproximadamente igual a 90%.]

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

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

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

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

  Passo 4 — Concretização
  Tendo em conta que

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

  e aos valores dos quantis acima e de n_i, $\bar{x}_i$, s_i^2 $(i = 1,2)$, temos

$$\begin{aligned} IC_{90\%}(\mu_1 - \mu_2) &= \left[(6.450 - 7.515) \pm 1.6449 \times \sqrt{\frac{1.451795}{40} + \frac{1.641308}{40}}\right] \\ &\simeq [-1.065 \pm 1.6449 \times 0.278078] \\ &\simeq [-1.522411,\ -0.607589]. \end{aligned}$$

(b) confronte as hipóteses $H_0: \mu_1 = \mu_2$ e $H_1: \mu_1 < \mu_2$, calculando para o efeito o valor-p aproximado. (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 suficientemente grandes.]

- **Região de rejeição de H_0** (para valores da estatística de teste)
  Estamos a lidar com um teste unilateral inferior ($H_1 : \mu_1 - \mu_2 < \mu_0$), logo a região de rejeição de H_0 (para valores da estatística de teste) é do tipo $W = (-\infty, c)$.

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

$$
\begin{aligned}
t &= \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\
&= \frac{(6.450 - 7.515) - 0}{\sqrt{\frac{1.451795}{40} + \frac{1.641308}{40}}} \\
&\simeq -3.829856.
\end{aligned}
$$

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

$$
\begin{aligned}
valor - p &= P(T < t \mid H_0) \\
&\simeq \Phi(t) \\
&\simeq \Phi(-3.83) \\
&\simeq 1 - \Phi(3.83) \\
&\overset{calc/tabela}{=} 1 - 0.999936 \\
&= 0.000064.
\end{aligned}
$$

  Logo é suposto:
  - não rejeitar H_0 a qualquer n.s. $\alpha_0 \leq 0.0064\%$;
  - rejeitar H_0 a qualquer n.s. $\alpha_0 > 0.0064\%$, nomeadamente a qualquer dos n.u.s. $(1\%, 5\%, 10\%)$.

Grupo II	10 valores

1. Uma engenheira defende a hipótese H_0 de que o primeiro algarismo (X) da massa atómica de elementos químicos com que habitualmente lida possui função de probabilidade

$$P(X = i) = \log_{10}\left(1 + \frac{1}{i}\right), \quad \text{para} \quad i = 1, 2, \ldots, 9.$$

Uma amostra casual de 100 desses elementos químicos conduziu ao seguinte quadro de frequências:

Primeiro algarismo da massa atómica	1	2	3	4	5	$\{6, \ldots 9\}$
Frequência absoluta observada	47	19	6	4	7	17
Frequência absoluta esperada sob H_0	30.1	E_2	12.5	9.7	7.9	E_6

(a) Calcule os valores das frequências absolutas esperadas E_2 e E_6 (aproximando-os às décimas). (1.0)

- **V.a. de interesse**
  X = primeiro algarismo da massa atómica de elementos químicos com que a engenheira lida

- **F.p. conjecturada**
  $\log_{10}\left(1 + \frac{1}{i}\right)$, $i = 1, 2, \ldots, 9$

- **Frequências absolutas esperadas omissas**
  Atendendo à dimensão da amostra $n = 100$ e à f.p. conjecturada, segue-se:

$$
\begin{aligned}
E_2 &= 100 \times \log_{10}\left(1 + \frac{1}{2}\right) \\
&\simeq 17.6;
\end{aligned}
$$

$$\begin{aligned} E_6 &= n - \sum_{i=1}^{5} E_i \\ &\simeq 100 - (30.1 + 17.6 + 12.5 + 9.7 + 7.9) \\ &= 22.2. \end{aligned}$$

(b) Teste a hipótese conjeturada pela engenheira, ao nível de significância de 5%. (3.0)

- **Hipóteses**
  $H_0 : P(X = i) = \log_{10}\left(1 + \frac{1}{i}\right), \quad$ para $\quad i = 1, 2, \ldots, 9$
  $H_1 : \neg H_0$

- **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 = 6
  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 a distribuição conjecturada em H_0 está completamente especificada, i.e., H_0 é uma hipótese simples.]

- **Frequências absolutas esperadas sob H_0**
  De acordo com a tabela facultada e a alínea (a), os valores das frequências absolutas esperadas sob H_0 aproximados às décimas são: $E_1 \simeq 30.1$; $E_2 \simeq 17.6$; $E_3 \simeq 12.5$; $E_4 \simeq 9.7$; $E_5 \simeq 7.9$; $E_6 \simeq 22.2$;
  [De notar que 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 teriam que ser recalculados...]

- **Região de rejeição de H_0** (para valores de T)
  Tratando-se de um teste de ajustamento, a região de rejeição de H_0 escrita para valores de T é o intervalo à direita $W = (c, +\infty)$, onde
  $$c = F^{-1}_{\chi^2_{(k-\beta-1)}}(1 - \alpha_0) = F^{-1}_{\chi^2_{(6-0-1)}}(1 - 0.05) = F^{-1}_{\chi^2_{(5)}}(0.95) \overset{tabela/calc}{=} 11.07.$$

- **Decisão**
  No cálculo do valor obs. da estatística de teste convém recorrer à seguinte tabela auxiliar:

i	Classe i	Freq. abs. obs. o_i	Freq. abs. esper. sob H_0 E_i	Parcelas valor obs. estat. teste $\frac{(o_i-E_i)^2}{E_i}$
1	$\{1\}$	47	30.1	$\frac{(47-30.1)^2}{30.1} \simeq 9.489$
2	$\{2\}$	19	17.6	0.111
3	$\{3\}$	6	12.5	3.380
4	$\{4\}$	4	9.7	3.349
5	$\{5\}$	7	7.9	0.103
6	$\{6,\ldots,9\}$	17	22.2	1.218
		$\sum_{i=1}^{k} o_i = n = 100$	$\sum_{i=1}^{k} E_i = n = 100$	$t = \sum_{i=1}^{k} \frac{(o_i-E_i)^2}{E_i} \simeq 17.650$

  Uma vez que $t \simeq 17.65 \in W = (11.07, +\infty)$, devemos rejeitar H_0 ao n.s. de $\alpha_0 = 5\%$ [ou qualquer outro n.s. superior a α_0].

2. Um conjunto de 18 medições independentes conduziu aos seguintes resultados referentes ao índice refrativo (x) e à densidade (Y) de pedaços de vidro encontrados em locais onde ocorreram crimes:

$$\bar{x} = 1.517889\,, \quad \sum_{i=1}^{18} x_i^2 = 41.471946\,, \quad \bar{y} = 2.491111\,, \quad \sum_{i=1}^{18} y_i^2 = 111.704964\,, \quad \sum_{i=1}^{18} x_i\, y_i = 68.062907\,,$$

onde $\left[\min_{i=1,\dots,18} x_i,\, \max_{i=1,\dots,18} x_i\right] = [1.514\,, 1.525]$.

(a) Obtenha a estimativa de mínimos quadrados do valor esperado da densidade do vidro com índice refrativo igual a 1.52. (2.0)

- **[Hipóteses de trabalho**
  $E(\epsilon_i) = 0,\ V(\epsilon_i) = \sigma^2,\ i = 1,\dots,n$
  $cov(\epsilon_i, \epsilon_j) = 0,\ i \neq j,\ i,j = 1,\dots,n]$

- **Estimativa de MQ de** $E(Y \mid x) = \beta_0 + \beta_1\, x$ **com** $x = 1.52$
  Dado que

  $n = 18$

  $\bar{x} = 1.517889$
  $\sum_{i=1}^n x_i^2 = 41.471946$
  $\sum_{i=1}^n x_i^2 - n\,(\bar{x})^2 \simeq 41.471946 - 18 \times 1.517889^2 \simeq 0.000180$

  $\bar{y} = 2.491111$
  $\sum_{i=1}^n y_i^2 = 111.704964$
  $\sum_{i=1}^n y_i^2 - n\,(\bar{y})^2 \simeq 111.704964 - 18 \times 2.491111^2 \simeq 0.003552$

  $\sum_{i=1}^n x_i\, y_i = 68.062907$
  $\sum_{i=1}^n x_i\, y_i - n\,\bar{x}\,\bar{y} \simeq 68.062907 - 18 \times 1.517889 \times 2.491111 \simeq 0.000767,$

  as estimativas de MQ de β_1, β_0 e $E(Y \mid x) = \beta_0 + \beta_1\, x$ são, para este modelo de RLS, iguais a:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i\, y_i - n\,\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\,(\bar{x})^2} \\
&\simeq \frac{0.000767}{0.000180} \\
&\simeq 4.261111 \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \times \bar{x} \\
&\simeq 2.491111 - 4.261111 \times 1.517889 \\
&\simeq -3.976783 \\
\hat{E}(Y \mid x) &= \hat{\beta}_0 + \hat{\beta}_1\, x \\
&\simeq -3.976783 + 4.261111 \times 1.52 \\
&\simeq 2.500106.
\end{aligned}$$

(b) Admitindo a validade das hipóteses de trabalho habituais para o modelo de regressão linear simples de Y em x, teste a significância do modelo de regressão linear simples ajustado, ao nível de 1%. (3.0)

- **[Hipóteses de trabalho**
  $\epsilon_i \overset{i.i.d.}{\sim} \text{normal}(0, \sigma^2),\ i = 1,\dots,n]$

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

- **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 de T)
Estamos a lidar com um teste bilateral ($H_1 : \beta_1 \neq 0$), 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.,

$$\begin{aligned} c &= F^{-1}_{t_{(n-2)}}(1 - \alpha_0/2) \\ &= F^{-1}_{t_{(18-2)}}(1 - 0.01/2) \\ &\overset{tabela/calc.}{=} 2.921. \end{aligned}$$

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

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-2}\left[\left(\sum_{i=1}^n y_i^2 - n\,\bar{y}^2\right) - (\hat{\beta}_1)^2\left(\sum_{i=1}^n x_i^2 - n\,\bar{x}^2\right)\right] \\ &\simeq \frac{1}{18-2}\left(0.003552 - 4.261111^2 \times 0.000180\right) \\ &\simeq 0.000018, \end{aligned}$$

o valor observado da estatística de teste é igual a

$$\begin{aligned} t &= \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n x_i^2 - n\,\bar{x}^2}}} \\ &\simeq \frac{4.261111 - 0}{\sqrt{\frac{0.000018}{0.000180}}} \\ &= \frac{4.261111}{\sqrt{0.1}} \\ &\simeq 13.474816. \end{aligned}$$

Como $t \simeq 13.474816 \in W = (-\infty, -2.921) \cup (2.921, +\infty)$ devemos rejeitar H_0 ao n.s. de $\alpha_0 = 1\%$ [bem como a qualquer n.s. superior a α_0].

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

- **Cálculo do coeficiente de determinação**

$$\begin{aligned} r^2 &= \frac{\left(\sum_{i=1}^n x_i y_i - n\,\bar{x}\,\bar{y}\right)^2}{\left(\sum_{i=1}^n x_i^2 - n\,\bar{x}^2\right) \times \left(\sum_{i=1}^n y_i^2 - n\,\bar{y}^2\right)} \\ &\simeq \frac{0.000767^2}{0.000180 \times 0.003552} \\ &\simeq 0.920122. \end{aligned}$$

- **Interpretação coeficiente de determinação**
Cerca de 92.0% 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 muito bem ao conjunto de dados.