Probabilidades e Estatística

LEAN, LEGI, LEGM, LMAC, MEAer, MEAmbi, MEC

1º semestre – 2018/2019
10/01/2019 – **11:00**

Duração: 90 minutos **2º teste B**

Justifique convenientemente todas as respostas

Grupo I	10 valores

1. Considere-se que a fração de água disponível diariamente num reservatório (às 00:00) é uma variável aleatória X com função de densidade de probabilidade

$$f_X(x) = \begin{cases} 2\theta\, x\left(1-x^2\right)^{\theta-1}, & 0<x<1 \\ 0, & \text{caso contrário,} \end{cases}$$

onde θ é um parâmetro desconhecido positivo. Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de X.

(a) Mostre que o estimador de máxima verosimilhança do parâmetro θ, com base na amostra aleatória referida acima, é dado por $-\frac{n}{\sum_{i=1}^n \ln(1-X_i^2)}$. (3.0)

- **V.a. de interesse**
  X = fração de água disponível diariamente num reservatório às 00:00 horas
- **F.d.p. de** X
  $$f_X(x) = \begin{cases} 2\theta\, x\left(1-x^2\right)^{\theta-1}, & 0<x<1 \\ 0, & \text{caso contrário,} \end{cases}$$
- **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

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

  Passo 2 — Função de log-verosimilhança

  $$\ln L(\theta \mid \underline{x}) = n\ln(2) + n\ln(\theta) + \sum_{i=1}^n \ln(x_i) + (\theta-1)\sum_{i=1}^n \ln\left(1-x_i^2\right)$$

  Passo 3 — Maximização

  A estimativa de MV de θ passa a ser representada por $\hat\theta$ e

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

$$\hat\theta \quad : \quad \begin{cases} \frac{n}{\hat\theta} + \sum_{i=1}^n \ln\left(1 - x_i^2\right) = 0 \\ -\frac{n}{\hat\theta^2} < 0 \qquad \text{(prop. verdadeira pois } n > 0) \end{cases}$$

$$\begin{cases} \hat\theta = -\frac{n}{\sum_{i=1}^n \ln\left(1-x_i^2\right)} \\ [-\frac{1}{n}\left[\sum_{i=1}^n \ln\left(1 - x_i^2\right)\right]^2 < 0]. \end{cases}$$

Passo 4 — Estimador de MV de θ

$$EMV(\theta) = -\frac{n}{\sum_{i=1}^n \ln\left(1 - X_i^2\right)}.$$

(b) A amostra $(x_1, x_2, \ldots, x_{20})$ conduziu a $\sum_{i=1}^{20} \ln\left(1 - x_i^2\right) = \ln(0.009)$. Calcule a estimativa de máxima verosimilhança da moda de X dada por $\frac{1}{\sqrt{2\theta-1}}$. (1.5)

- **Estimativa de MV de θ**

$$\begin{aligned} \hat\theta &= -\frac{n}{\sum_{i=1}^n \ln\left(1 - x_i^2\right)} \\ &= -\frac{20}{\ln(0.009)} \\ &= 4.245806 \end{aligned}$$

- **Outro parâmetro desconhecido**
  $h(\theta) = mo(X) = \frac{1}{\sqrt{2\theta-1}}$

- **Estimativa de MV de $h(\theta)$**
  Pela propriedade de invariância dos estimadores de máxima verosimilhança, concluímos que a estimativa de MV de $h(\theta)$ é igual a

$$\begin{aligned} \widehat{h(\theta)} &= h(\hat\theta) \\ &= \frac{1}{\sqrt{2\hat\theta - 1}} \\ &\simeq \frac{1}{\sqrt{2 \times 4.245806 - 1}} \\ &\simeq 0.365353. \end{aligned}$$

2. Em determinada região afetada por um surto epidémico, recolheu-se uma amostra casual de 1500 indivíduos, tendo-se encontrado 723 indivíduos contaminados.

(a) Determine um intervalo de confiança a aproximadamente 90% para a verdadeira proporção, p, de indivíduos contaminados na região afetada pelo surto epidémico. (2.5)

- **V.a. de interesse**

$$X = \begin{cases} 1, & \text{se indivíduo está contaminado} \\ 0, & \text{c.c.} \end{cases}$$

- **Situação**
  $X \sim \text{Bernoulli}(p)$
  $p = P(\text{indíviduo contaminado})$ DESCONHECIDA
  $n = 1500 >> 30$ (suficientemente grande).

- **Obtenção de IC aproximado para p**
  Passo 1 — Selecção da v.a. fulcral para p
  [Uma vez que nos foi solicitada a determinação de um IC aproximado para uma probabilidade e a dimensão da amostra é suficientemente grande para justificar o recurso à seguinte v.a. fulcral para p com distribuição aproximada]

$$Z = \frac{\bar{X} - p}{\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}} \overset{a}{\sim} \text{normal}(0,1)$$

Passo 2 — Obtenção dos quantis de probabilidade
Os quantis a utilizar são

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

[Estes enquadram a v.a. fulcral para p com probabilidade aproximadamente igual a $(1-\alpha) = 0.90$.]

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

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

$$P\left[a_\alpha \leq \frac{\bar{X}-p}{\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}} \leq b_\alpha\right] \simeq 1-\alpha$$

$$P\left[\bar{X} - b_\alpha \times \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \leq p \leq \bar{X} - a_\alpha \times \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right] \simeq 1-\alpha$$

$$P\left[\bar{X} - \Phi^{-1}(1-\alpha/2) \times \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \leq p \leq \bar{X} + \Phi^{-1}(1-\alpha/2) \times \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right] \simeq 1-\alpha.$$

Passo 4 — Concretização
Ao ter-se em conta que

- $n = 1500$
- $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i = \frac{723}{1500} = 0.482$ [$\equiv$ proporção observada de indivíduos contaminados]
- $\Phi^{-1}(1-\alpha/2) = 1.6449$,

conclui-se que o intervalo de confiança a aproximadamente 90% para p é dado por

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

$$= \left[0.482 - 1.6449 \times \sqrt{\frac{0.482 \times (1-0.482)}{1500}}, \quad 0.482 + 1.6449 \times \sqrt{\frac{0.482 \times (1-0.482)}{1500}}\right]$$

$$= [0.460778, 0.503222].$$

(b) Com base na amostra referida, confronte as hipóteses $H_0 : p = 0.5$ e $H_1 : p \neq 0.5$. Decida com base no valor-p. (3.0)

- **Hipóteses**
  $H_0 : p = p_0 = 0.5$
  $H_1 : p \neq p_0$

- **Estatística de teste**
  [Sabe-se que o estimador de MV de p é $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, onde $X_i \sim_{i.i.d.} X$. Para além disso, $E(\bar{X}) = E(X) = p$ e $V(\bar{X}) = \frac{1}{n}V(X) = \frac{p(1-p)}{n} < +\infty$. Então pelo TLC pode afirmar-se que $\frac{\bar{X}-E(\bar{X})}{\sqrt{V(\bar{X})}} = \frac{\bar{X}-p}{\sqrt{\frac{p(1-p)}{n}}} \overset{a}{\sim} \text{normal}(0,1)$, pelo que a estatística de teste é]

$$T = \frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{a}{\sim}_{H_0} \text{normal}(0,1).$$

- **Região de rejeição de H_0** (para valores de T)
  Tratando-se de um teste bilateral ($H_1 : p \neq p_0$), a região de rejeição de H_0, escrita para valores da estatística de teste, é do tipo $W = (-\infty, -c) \cup (c, +\infty)$.

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

$$\begin{aligned} t &= \frac{\bar{x}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \\ &= \frac{0.482-0.5}{\sqrt{\frac{0.5\,(1-0.5)}{1500}}} \\ &\simeq -1.39. \end{aligned}$$

  Uma vez que a região de rejeição deste teste é a reunião de dois intervalos simétricos, temos:

$$\begin{aligned} valor-p &= 2\times P(T>|t| \mid H_0) \\ &= 2\times[1-P(T\leq |t| \mid H_0)] \\ &\simeq 2\times[1-\Phi(|t|)] \\ &\simeq 2\times[1-\Phi(1.39)] \\ &\stackrel{calc/tabela}{=} 2\times(1-0.9177) \\ &= 0.1646. \end{aligned}$$

  Consequentemente, é suposto:
  - não rejeitar H_0 a qualquer n.s. $\alpha_0 \leq 16.46\%$, por exemplo, a qualquer dos n.u.s. (1%, 5% e 10%);
  - rejeitar H_0 a qualquer n.s. $\alpha_0 > 16.46\%$.

Grupo II	10 valores

1. Num estudo urbanístico, uma investigadora está interessada na variável aleatória X que representa a proporção de casas por rua de Lisboa que os proprietários pretendem explorar em regime de *Alojamento Local* de entre as casas que se encontram devolutas na mesma rua. A investigadora defende a conjetura H_0 de que X possui função de distribuição dada por

$$P(X\leq x)=1-\left(1-x^2\right)^2, \quad 0\leq x\leq 1.$$

Para avaliar esta conjetura ela selecionou casualmente 50 ruas de Lisboa e registou o valor observado de X para cada uma delas, tendo-se obtido a seguinte tabela de frequências:

Classe	$[0,0.325]$	$]0.325,0.475]$	$]0.475,0.606]$	$]0.606,0.743]$	$]0.743,1]$
Frequência absoluta observada	8	12	9	13	8
Freq. abs. esperada sob H_0	10.00	10.01	9.96	E_4	E_5

(a) Obtenha os valores de E_4 e E_5 (aproximando-os às centésimas). (1.0)

- **V.a. de interesse**
  X = proporção de casas por rua de Lisboa que os proprietários pretendem explorar em regime de A.L. de entre as casas que se encontram devolutas na mesma rua

- **F.d. conjecturada**
  $F(x)=P(X\leq x)=1-\left(1-x^2\right)^2, \quad 0\leq x\leq 1$

- **Frequências absolutas esperadas omissas**
  Atendendo à dimensão da amostra $n=50$ e à f.d. conjecturada, temos

$$\begin{aligned} E_4 &= 50\times P(0.606<X\leq 0.743 \mid H_0) \\ &= 50\times[F(0.743)-F(0.606)] \\ &\simeq 9.99 \\ E_5 &= n-\sum_{i=1}^{4} E_i \\ &\simeq 50-(10.00+10.01+9.96+9.99) \\ &= 10.04. \end{aligned}$$

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

- **Hipóteses**
  $H_0 : X$ possui f.d. $P(X \le x) = 1 - \left(1 - x^2\right)^2, \quad 0 \le x \le 1$
  $H_1 : \neg H_0$

- **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 = 5

  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.

- **Estimativas das frequências absolutas esperadas sob H_0**
  De acordo com a tabela facultada e a alínea (a), as frequências absolutas esperadas sob H_0 (aproximadas às centésimas) são: $E_1 \simeq 10.00$; $E_2 \simeq 10.01$; $E_3 \simeq 9.96$; $E_4 \simeq 9.99$; $E_5 \simeq 10.04$.
  [Não é necessário fazer qualquer agrupamento de classes uma vez que em pelo menos 80% das classes se verifica $E_i \ge 5$ e que $E_i \ge 1$ 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)
  Tratando-se de um teste de ajustamento, a região de rejeição de H_0 é o intervalo à direita $W = (c, +\infty)$, onde

$$\begin{aligned} c &= F^{-1}_{\chi^2_{(k-\beta-1)}}(1-\alpha_0) \\ &= F^{-1}_{\chi^2_{(5-0-1)}}(1-0.05) \\ &\overset{tabela/calc.}{=} 9.488. \end{aligned}$$

- **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	$[0, 0.325]$	8	10.00	$\frac{(8-10.00)^2}{10.00} = 0.4$
2	$]0.325, 0.475]$	12	10.01	$\frac{(12-10.01)^2}{10.01} \simeq 0.396$
3	$]0.475, 0.606]$	9	9.96	0.093
4	$]0.606, 0.743]$	13	9.99	0.907
5	$]0.743, 1]$	8	10.04	0.415
		$\sum_{i=1}^{k} o_i = n = 50$	$\sum_{i=1}^{k} E_i = n = 50$	$t = \sum_{i=1}^{k} \frac{(o_i-E_i)^2}{E_i} \simeq 2.211$

  Dado que $t \simeq 2.211 \not\in W = (9.488, +\infty)$, não devemos rejeitar H_0 ao n.s. de $\alpha_0 = 5\%$ [nem a qualquer outro n.s. inferior a α_0].

2. Medições da percentagem de gordura corporal (x) e do índice de massa corporal (Y) de 52 pacientes de certa clínica conduziram aos seguintes resultados:

$$\sum_{i=1}^{52} x_i = 930.1, \quad \sum_{i=1}^{52} x_i^2 = 21544.71, \quad \sum_{i=1}^{52} y_i = 1332.2, \quad \sum_{i=1}^{52} y_i^2 = 34801.32, \quad \sum_{i=1}^{52} x_i y_i = 25279.19,$$

onde $\left[\min_{i=1,\ldots,52} x_i, \max_{i=1,\ldots,52} x_i\right] = [9, 27]$.

(a) Considere o modelo de regressão linear simples de Y em x e obtenha a estimativa de mínimos (2.0) quadrados do valor esperado do índice de massa corporal de um paciente com percentagem de gordura corporal igual a 10.

- **Estimativa de MQ de** $E(Y \mid x) = \beta_0 + \beta_1 x$ **com** $x = 10$
  Dado que

  $n = 52$

  $\sum_{i=1}^n x_i = 930.1$
  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{930.1}{52} \simeq 17.886538$
  $\sum_{i=1}^n x_i^2 = 21\,544.71$
  $\sum_{i=1}^n x_i^2 - n(\bar{x})^2 \simeq 21\,544.71 - 52 \times 17.886538^2 \simeq 4\,908.441435$

  $\sum_{i=1}^n y_i = 1\,332.2$
  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1\,332.2}{52} \simeq 25.619231$
  $\sum_{i=1}^n y_i^2 = 34\,801.3$
  $\sum_{i=1}^n y_i^2 - n(\bar{y})^2 \simeq 34\,801.32 - 52 \times 25.619231^2 \simeq 671.380154$

  $\sum_{i=1}^n x_i y_i = 25\,279.19$
  $\sum_{i=1}^n x_i y_i - n\,\bar{x}\,\bar{y} \simeq 25\,279.19 - 52 \times 17.886538 \times 25.619231 \simeq 1450.743862,$

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

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - n\,\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n(\bar{x})^2} \\
&\simeq \frac{1\,450.743862}{4\,908.441435} \\
&\simeq 0.295561 \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \times \bar{x} \\
&\simeq 25.619231 - 0.295561 \times 17.886538 \\
&\simeq 20.332668 \\
\hat{\beta}_0 + \hat{\beta}_1 x &\simeq 20.332668 + 0.295561 \times 10 \\
&\simeq 23.288278.
\end{aligned}$$

(b) Após ter enunciado as hipóteses de trabalho que entender convenientes, teste a significância do modelo de regressão linear simples ajustado ao nível de 10%. (3.0)

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

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

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

- **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 ($H_1 : \beta_1 \neq \beta_{1,0}$), 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.,

$$
\begin{aligned}
c &= F^{-1}_{t_{(n-2)}}(1-\alpha_0/2) \\
&= F^{-1}_{t_{(52-2)}}(1-0.1/2) \\
&\overset{tabela/calc.}{\simeq} 1.676.
\end{aligned}
$$

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

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-2}\left[\left(\sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2\right) - (\hat{\beta}_1)^2\left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2\right)\right] \\
&\simeq \frac{1}{52-2}\left(671.380154 - 0.353975^2 \times 4\,908.441435\right) \\
&\simeq 4.851937,
\end{aligned}
$$

  o valor observado da estatística de teste é igual a

$$
\begin{aligned}
t &= \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}}} \\
&\simeq \frac{0.295561 - 0}{\sqrt{\frac{4.851937}{4\,908.441435}}} \\
&\simeq 9.400725.
\end{aligned}
$$

  Como $t \simeq 9.400725 \in W = (-\infty, -1.676) \cup (1.676, +\infty)$ devemos rejeitar H_0 ao n.s. de $\alpha_0 = 10\%$ [assim como a qualquer n.s. superior a 10%].

(c) Calcule e interprete o coeficiente de determinação do modelo de regressão linear simples ajustado. (1.0)

- **Cálculo do coeficiente de determinação**

$$
\begin{aligned}
r^2 &= \frac{\left(\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}\right)^2}{\left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2\right) \times \left(\sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2\right)} \\
&= \frac{1\,450.743862^2}{4\,908.441435 \times 671.380154} \\
&\simeq 0.638659.
\end{aligned}
$$

- **Interpretação coeficiente de determinação**
  Cerca de 63.9% 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 bem ao conjunto de dados [e deverá conduzir a resultados com interesse prático].