Probabilidades e Estatística

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

1º semestre – 2018/2019
10/01/2019 – **09:00**

Duração: 90 minutos **2º teste A**

Justifique convenientemente todas as respostas

Grupo I	10 valores

1. A variável aleatória X representa o número diário de análises químicas efectuadas por um pequeno laboratório e possui função de probabilidade satisfazendo:

$$P(X=x) = \frac{(x+1)\,(x+2)}{2}\,(1-p)^x\,p^3, \quad x=0,1,2,\ldots,$$

onde $p \in\,]0,1[$ é uma probabilidade desconhecida. Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de X.

(a) Mostre que o estimador de máxima verosimilhança do parâmetro p, com base na amostra aleatória referida acima, é dado por $3/(\bar{X}+3)$. (3.0)

- **V.a. de interesse**
  X = número diário de análises químicas a efectuar por um pequeno laboratório
- **F.p. de** X
  $P(X=x) = \frac{(x+1)\,(x+2)}{2}\,(1-p)^x\,p^3, \quad x=0,1,2,\ldots$
- **Parâmetro desconhecido**
  $p, \quad 0<p<1$
- **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

$$\begin{aligned}
L(p \mid \underline{x}) &= P(\underline{X} = \underline{x}) \\
&\overset{X_i\ indep}{=} \prod_{i=1}^{n} P(X_i = x_i) \\
&\overset{X_i \sim X}{=} \prod_{i=1}^{n} P(X = x_i) \\
&= \prod_{i=1}^{n} \left[\frac{(x_i+1)\,(x_i+2)}{2}\,(1-p)^{x_i}\,p^3\right] \\
&= 2^{-n}\left[\prod_{i=1}^{n}(x_i+1)\,(x_i+2)\right](1-p)^{\sum_{i=1}^n x_i}\,p^{3n}, \quad 0<p<1
\end{aligned}$$

  Passo 2 — Função de log-verosimilhança (caso $\bar{x}>0$)

$$\ln L(p \mid \underline{x}) = -n\ln(2) + \sum_{i=1}^{n} \ln[(x_i+1)\,(x_i+2)] + \ln(1-p)\sum_{i=1}^{n} x_i + 3n\ln(p)$$

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

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

$$\begin{cases} -\frac{\sum_{i=1}^n x_i}{1-\hat{p}} + \frac{3n}{\hat{p}} = 0 \\ -\frac{\sum_{i=1}^n x_i}{(1-\hat{p})^2} - \frac{3n}{\hat{p}^2} < 0 \quad \text{(prop. verdadeira porque } \bar{x}\geq 0 \text{ e } n>0) \end{cases}$$

$$\hat{p} \quad : \quad \begin{cases} -\hat{p}\,n\bar{x} + 3n - 3n\hat{p} = 0 \quad \Leftrightarrow \quad \hat{p} = \frac{3}{\bar{x}+3} \\ \left[-\frac{n\bar{x}}{\left(1-\frac{3}{\bar{x}+3}\right)^2} - \frac{3n}{\left(\frac{3}{\bar{x}+3}\right)^2} = -\frac{n\,(\bar{x}+3)^3}{3\,\bar{x}} < 0 \right]. \end{cases}$$

Passo 4 — Estimador de MV de λ

$$EMV(p) = \frac{3}{\bar{X}+3}.$$

(b) Obtenha a estimativa de máxima verosimilhança de $E(X) = 3\left(\frac{1}{p} - 1\right)$ baseada na concretização $(x_1, x_2, x_3, x_4, x_5) = (2, 3, 1, 4, 0)$. (1.5)

- **Estimativa de MV de p**

$$\begin{aligned} \hat{p} &= \frac{3}{\bar{x}+3} \\ &= \frac{3}{\frac{2+3+1+4+0}{5}+3} \\ &= \frac{3}{2+3} \\ &= 0.6 \end{aligned}$$

- **Outro parâmetro desconhecido**
$h(p) = 3\left(\frac{1}{p} - 1\right)$

- **Estimativa de MV de $h(p)$**
Ao invocar a propriedade de invariância dos estimadores de máxima verosimilhança, obtemos a estimativa de MV de $h(p)$:

$$\begin{aligned} \widehat{h(p)} &= h(\hat{p}) \\ &= 3\left(\frac{1}{\hat{p}} - 1\right) \\ &= 3\left(\frac{1}{0.6} - 1\right) \\ &= 2 \qquad [\equiv \bar{x} \text{ pois } h(\hat{p}) = 3\left(\frac{1}{\hat{p}} - 1\right) = 3\left(\frac{1}{\frac{3}{\bar{x}+3}} - 1\right) = \bar{x}]. \end{aligned}$$

2. De forma a comparar dois grupos de estudantes universitários, considerou-se a variável aleatória X_1 (respetivamente X_2) que representa a altura, em cm, de estudantes com actividade física acima da média (respetivamente abaixo da média). Ao selecionarem-se casualmente 20 estudantes de cada um dos grupos, obtiveram-se $\bar{x}_1 = 173$ e $\bar{x}_2 = 171$. Admita que X_1 e X_2 possuem distribuições normais com valores esperados μ_1 e μ_2 desconhecidos e desvios padrão conhecidos e iguais a $\sigma_1 = 6.4$ e $\sigma_2 = 7.2$.

(a) Obtenha um intervalo de confiança a 95% para $\mu_1 - \mu_2$. (2.5)

- **V.a. de interesse**
X_1 = altura (em cm) de estudante com actividade física acima da média
X_2 = altura (em cm) de estudante com actividade física abaixo da média

- **Situação**
$X_i \overset{indep.}{\sim} \text{normal}(\mu_i, \sigma_i^2), \quad i = 1, 2$
$(\mu_1 - \mu_2)$ DESCONHECIDO
σ_1^2 e σ_2^2 conhecidas
$n_1 = n_2 = 20$

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

  $$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \text{normal}(0,1)$$

  [dado que se pretende determinar um IC para a diferença de valores esperados de duas populações independentes com distribuições normais e com variâncias conhecidas.]

  Passo 2 — Obtenção dos quantis de probabilidade
  Ao ter-se em consideração que $(1-\alpha) \times 100\% = 95\%$, far-se-á uso dos quantis

  $$\begin{cases} a_\alpha = \Phi^{-1}(\alpha/2) = \Phi^{-1}(0.025) = -\Phi^{-1}(1-0.025) \stackrel{tabela/calc.}{=} -1.9600 \\ b_\alpha = \Phi^{-1}(1-\alpha/2) = \Phi^{-1}(0.975) \stackrel{tabela/calc.}{=} 1.9600. \end{cases}$$

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

  $$P(a_\alpha \leq Z \leq b_\alpha) = 1 - \alpha$$
  $$P\left[a_\alpha \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq b_\alpha\right] = 1 - \alpha$$
  $$P\left[(\bar{X}_1 - \bar{X}_2) - b_\alpha \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) - a_\alpha \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right] = 1 - \alpha$$

  Passo 4 — Concretização
  Tendo em conta que

  $$IC_{(1-\alpha)\times 100\%}(\mu_1 - \mu_2) = \left[(\bar{x}_1 - \bar{x}_2) \pm \Phi^{-1}(1-\alpha/2) \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

  e atendendo aos valores dos quantis acima e de n_i, $\bar{x}_i$, σ_i^2 $(i = 1,2)$, segue-se

  $$\begin{aligned} IC_{95\%}(\mu_1 - \mu_2) &= \left[(173 - 171) \pm 1.9600 \times \sqrt{\frac{6.4^2}{20} + \frac{7.2^2}{20}}\right] \\ &\simeq [2 \pm 1.9600 \times 2.15407] \\ &\simeq [-2.22197,\ 6.22197]. \end{aligned}$$

(b) Confronte as hipóteses $H_0 : \mu_1 - \mu_2 = 1$ e $H_1 : \mu_1 - \mu_2 > 1$, calculando para o efeito o valor-p. (3.0)

- **V.a. de interesse e situação**
  Ver alínea (a).

- **Hipóteses**
  $H_0 : \mu_1 - \mu_2 = \mu_0 = 1$
  $H_1 : \mu_1 - \mu_2 > \mu_0$

- **Estatística de teste**

  $$T = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim_{H_0} \text{normal}(0,1)$$

  [uma vez que se pretende efectuar um teste sobre a diferença de valores esperados de duas populações independentes com distribuições normais e com variâncias conhecidas.]

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

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

$$\begin{aligned} t &= \frac{(\bar{x}_1-\bar{x}_2)-\mu_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \\ &= \frac{(173-171)-1}{\sqrt{\frac{6.4^2}{20}+\frac{7.2^2}{20}}} \\ &\simeq \frac{1}{2.154066} \\ &\simeq 0.464238 \end{aligned}$$

  e a região de rejeição deste teste é um intervalo à direita. Donde

$$\begin{aligned} valor-p &= P(T>t\mid H_0) \\ &= 1-\Phi(t) \\ &\simeq 1-\Phi(0.46) \\ &\overset{tabela/calc.}{=} 1-0.6772 \\ &= 0.3228. \end{aligned}$$

  Logo é suposto:
  - não rejeitar H_0 a qualquer n.s. $\alpha_0 \leq 32.28\%$, por exemplo, a qualquer dos n.u.s. (1%, 5% e 10%);
  - rejeitar H_0 a qualquer n.s. $\alpha_0 > 32.28\%$.

Grupo II	10 valores

1. Em determinado processo industrial é crucial caraterizar a variável aleatória X que representa o número de peças mecânicas inspecionadas até se encontrar uma peça defeituosa. Uma engenheira mecânica defende a hipótese H_0 de que X possui distribuição geométrica com parâmetro desconhecido p. Uma amostra aleatória de 200 contagens do número de peças mecânicas inspecionadas até se encontrar uma peça defeituosa conduziu à seguinte tabela de frequências:

Classe	1 a 4	5 a 8	9 a 12	13 a 16	17 a 20	21 ou mais
Frequência absoluta observada	7	7	10	5	8	163
Estimativa da freq. abs. esperada sob H_0	e_1	7.57	7.27	6.99	6.71	e_6

(a) Sabendo que a estimativa de máxima verosimilhança de p é $\hat{p} \simeq 0.01$, obtenha os valores das estimativas e_1 e e_6 (aproximando-as às centésimas). (1.0)

- **V.a. de interesse**
  $X =$ número de peças mecânicas inspecionadas até se encontrar uma peça defeituosa
- **F.p. conjecturada**
  $P(X=x) = (1-p)^{x-1}\,p, \quad x = 1,2,\ldots,$ onde p é uma probabilidade desconhecida.
- **Estimativas das frequências absolutas esperadas omissas**
  Atendendo à dimensão da amostra $n = 200$, à f.p. conjecturada, à estimativa de MV facultada $\hat{p} \simeq 0.01$ e à propriedade de invariância dos EMV, temos:

$$\begin{aligned} e_1 &= n \times P(X\leq 4\mid p=\hat{p}) \\ &= 200 \times \sum_{i=1}^{4}(1-\hat{p})^{x-1}\,\hat{p} \\ &= 200\times\hat{p}\times\frac{1-(1-\hat{p})^4}{1-(1-\hat{p})} \\ &\simeq 200\times(1-0.99^4) \\ &\simeq 7.88; \end{aligned}$$

$$
\begin{aligned}
e_5 &= n \times \hat{P}(X \geq 21) \\
&= n - \sum_{i=1}^{5} e_i \\
&\simeq 200 - (7.88 + 7.57 + 7.27 + 6.99 + 6.71) \\
&= 163.58.
\end{aligned}
$$

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

- **Hipóteses**
  $H_0 : X \sim \text{geométrica}(p)$ (p desconhecido)
  $H_1 : X \not\sim \text{geométrica}(p)$

- **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 = ESTIMADOR da frequência absoluta esperada, sob H_0, da classe i

  β = No. de parâmetros a estimar = 1 [dado que p é uma probabilidade desconhecida.]

- **Estimativas das frequências absolutas esperadas sob H_0**
  De acordo com a tabela facultada e a alínea (a), as estimativas [de MV] das frequências absolutas esperadas sob H_0 aproximadas às centésimas são: $e_1 \simeq 7.88$; $e_2 \simeq 7.57$; $e_3 \simeq 7.27$; $e_4 \simeq 6.99$; $e_5 \simeq 6.71$; $e_6 \simeq 163.58$.
  [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)
  Tratando-se de um teste de ajustamento, a região de rejeição de H_0 é o intervalo à direita $W = (c, +\infty)$, onde

$$
\begin{aligned}
c &= F^{-1}_{\chi^2_{(k-\beta-1)}}(1-\alpha_0) \\
&= F^{-1}_{\chi^2_{(6-1-1)}}(1-0.05) \\
&\overset{tabela/calc.}{=} 9.488.
\end{aligned}
$$

- **Decisão**

i	Classe i	Freq. abs. obs. o_i	Estim. freq. abs. esp. sob H_0 e_i	Parcelas valor obs. estat. teste $\frac{(o_i-e_i)^2}{e_i}$
1	$\{1,2,3,4\}$	7	7.88	$\frac{(7-7.88)^2}{7.88} \simeq 0.098$
2	$\{5,6,7,8\}$	7	7.57	$\frac{(7-7.57)^2}{7.57} \simeq 0.043$
3	$\{9,10,11,12\}$	10	7.27	1.025
4	$\{13,14,15,16\}$	5	6.99	0.567
5	$\{17,18,19,20\}$	8	6.71	0.248
6	$\{21,22,\ldots\}$	163	163.58	0.002
		$\sum_{i=1}^{k} o_i = n = 200$	$\sum_{i=1}^{k} e_i = n = 200$	$t = \sum_{i=1}^{k} \frac{(o_i-e_i)^2}{e_i} \simeq 1.983$

  Uma vez que $t \simeq 1.983 \not\in W = (9.488, +\infty)$, não devemos rejeitar H_0 ao n.s. de $\alpha_0 = 5\%$ [nem a qualquer outro n.s. inferior a α_0].

2. Um conjunto de 20 medições independentes conduziu aos seguintes resultados referentes ao nível de pressão sonora de uma fonte de ruído (x, em dB) e ao aumento da tensão arterial de um indivíduo causado pela exposição a esta fonte de ruído (Y, em mm Hg):

$$\sum_{i=1}^{20} x_i = 1646, \quad \sum_{i=1}^{20} x_i^2 = 138476, \quad \sum_{i=1}^{20} y_i = 86, \quad \sum_{i=1}^{20} y_i^2 = 494, \quad \sum_{i=1}^{20} x_i y_i = 7594,$$

onde $\left[\min_{i=1,\ldots,20} x_i, \max_{i=1,\ldots,20} x_i\right] = [60, 100]$.

(a) Considere o modelo de regressão linear simples de Y em x e determine a estimativa de mínimos quadrados do valor esperado do aumento da tensão arterial de um indivíduo exposto a um ruído com nível de pressão sonora de 110 dB. (2.0)

- **Estimativa de MQ de** $E(Y \mid x) = \beta_0 + \beta_1 x$ **com** $x = 110$
  Dado que

$$n = 20$$

$$\sum_{i=1}^n x_i = 1646$$
$$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1646}{20} = 82.3$$
$$\sum_{i=1}^n x_i^2 = 138476$$
$$\sum_{i=1}^n x_i^2 - n(\bar{x})^2 = 138476 - 20 \times 82.3^2 = 3010.2$$

$$\sum_{i=1}^n y_i = 86$$
$$\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i = \frac{86}{20} = 4.3$$
$$\sum_{i=1}^n y_i^2 = 494$$
$$\sum_{i=1}^n y_i^2 - n(\bar{y})^2 = 494 - 20 \times 4.3^2 = 124.2$$

$$\sum_{i=1}^n x_i y_i = 7594$$
$$\sum_{i=1}^n x_i y_i - n\,\bar{x}\,\bar{y} = 7594 - 20 \times 82.3 \times 4.3 = 516.2,$$

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

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^n x_i^2 - n(\bar{x})^2} \\
&= \frac{516.2}{3010.2} \\
&\simeq 0.171484 \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \times \bar{x} \\
&\simeq 4.3 - 0.171484 \times 82.3 \\
&\simeq -9.813133 \\
\hat{\beta}_0 + \hat{\beta}_1 x &\simeq -9.813133 + 0.171484 \times 110 \\
&\simeq 9.050107.
\end{aligned}$$

- **[Comentário**
  Estamos a cometer um erro de extrapolação ao estimar pontualmente $E(Y \mid x) = \beta_0 + \beta_1 \times x$ uma vez que $x = 110 \not\in [\min_{i=1,\ldots,n} x_i, \max_{i=1,\ldots,n} x_i] = [60, 100]$.]

(b) Após ter enunciado as hipóteses de trabalho que entender convenientes, confronte $H_0 : \beta_1 = 0.2$ e $H_1 : \beta_1 \neq 0.2$, ao nível de significância de 10%. (3.0)

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

- **Hipóteses**
  $H_0 : \beta_1 = \beta_{1,0} = 0.2$
  $H_1 : \beta_1 \neq \beta_{1,0}$

- **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 ($H_1 : \beta_1 \neq 0.2$), 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.,

$$\begin{aligned} c &= F^{-1}_{t_{(n-2)}}(1-\alpha_0/2) \\ &= F^{-1}_{t_{(20-2)}}(1-0.10/2) \\ &= F^{-1}_{t_{(18)}}(0.95) \\ &\overset{tabela/calc.}{=} 1.734. \end{aligned}$$

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

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-2}\left[\left(\sum_{i=1}^n y_i^2 - n\,\bar{y}^2\right) - \left(\hat{\beta}_1\right)^2\left(\sum_{i=1}^n x_i^2 - n\,\bar{x}^2\right)\right] \\ &\simeq \frac{1}{20-2}\left(124.2 - 0.171484^2 \times 3\,010.2\right) \\ &\simeq 1.982209, \end{aligned}$$

  o valor observado da estatística de teste é igual a

$$\begin{aligned} t &= \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n x_i^2 - n\,\bar{x}^2}}} \\ &\simeq \frac{0.171484 - 0.2}{\sqrt{\frac{1.982209}{3\,010.2}}} \\ &= -1.111250. \end{aligned}$$

  Como $t \simeq -1.111250 \not\in W = (-\infty, -1.734) \cup (1.734, +\infty)$ não devemos rejeitar H_0 ao n.s. de 10% [nem a qualquer n.s. inferior a 10%].

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

- **Cálculo do coeficiente de determinação**

$$\begin{aligned} r^2 &= \frac{\left(\sum_{i=1}^n x_i y_i - n\,\bar{x}\,\bar{y}\right)^2}{\left(\sum_{i=1}^n x_i^2 - n\,\bar{x}^2\right) \times \left(\sum_{i=1}^n y_i^2 - n\,\bar{y}^2\right)} \\ &= \frac{516.2^2}{3\,010.2 \times 124.2} \\ &\simeq 0.712720. \end{aligned}$$

- **Interpretação coeficiente de determinação**
  Cerca de 71.3% 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.