DM
DEPARTAMENTO DE MATEMÁTICA
TÉCNICO LISBOA

Probabilidades e Estatística

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

2º semestre – 2016/2017
16/06/2017 – **9h:00**

Duração: 90 minutos **2º teste**

Justifique convenientemente todas as respostas

Grupo I	10 valores

1. O tempo de vida (em minuto) de um microrganismo de uma determinada espécie é descrito por uma variável aleatória X, com distribuição exponencial com valor esperado desconhecido $\frac{1}{\lambda}$.

(a) Deduza o estimador de máxima verosimilhança de λ com base numa amostra aleatória $(X_1,\ldots,X_n)$ proveniente da população X. (3.0)

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

  Passo 1 — Função de verosimilhança
  $$\begin{aligned} L(\lambda|\underline{x}) &= f_{\underline{X}}(\underline{x}) \\ &\overset{X_i\, indep}{=} \prod_{i=1}^{n} f_{X_i}(x_i) \\ &\overset{X_i \sim X}{=} \prod_{i=1}^{n} f_X(x_i) \\ &= \prod_{i=1}^{n} \left(\lambda\, e^{-\lambda x_i}\right) \\ &= \lambda^n\, e^{-\lambda \sum_{i=1}^n x_i}, \quad \lambda > 0 \end{aligned}$$

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

  Passo 3 — Maximização
  A estimativa de MV de λ é doravante representada por $\hat{\lambda}$ e
  $$\hat{\lambda} \;:\; \begin{cases} \left.\frac{d \ln L(\lambda|\underline{x})}{d\lambda}\right|_{\lambda=\hat{\lambda}} = 0 & \text{(ponto de estacionaridade)} \\ \left.\frac{d^2 \ln L(\lambda|\underline{x})}{d\lambda^2}\right|_{\lambda=\hat{\lambda}} < 0 & \text{(ponto de máximo)} \end{cases}$$
  $$\begin{cases} \frac{n}{\hat{\lambda}} - \sum_{i=1}^n x_i = 0 \\ -\frac{n}{\hat{\lambda}^2} < 0 \end{cases}$$
  $$\begin{cases} \hat{\lambda} = \frac{n}{\sum_{i=1}^n x_i} \\ -\frac{\left(\sum_{i=1}^n x_i\right)^2}{n} < 0 & \text{(proposição verdadeira já que } \sum_{i=1}^n x_i > 0\text{).} \end{cases}$$

Passo 4 — Estimador de MV de λ

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

(b) Recolheu-se a amostra $(x_1,\ldots,x_8) = (133,24,84,30,108,57,5,45)$, tendo-se observado que $\sum_{i=1}^{8} x_i = 486$. Calcule a estimativa de máxima verosimilhança da probabilidade de o tempo de vida dum microrganismo ser inferior a 90 minutos. (2.0)

- **Estimativa de MV de λ**

$$\begin{aligned}\hat{\lambda} &= \frac{n}{\sum_{i=1}^{n} x_i} \\ &= \frac{8}{486} \\ &\simeq 0.016461\end{aligned}$$

- **Outro parâmetro desconhecido**

$$\begin{aligned}h(\lambda) &= P(X < 90) \\ &= F_X(90) \\ & [= \int_{-\infty}^{90} f_X(x)\,dx = -e^{-\lambda x}\Big|_0^{90}] \\ &= 1 - e^{-90\lambda}\end{aligned}$$

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

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

$$\begin{aligned}\widehat{h(\lambda)} &= h(\hat{\lambda}) \\ &= 1 - e^{-90\hat{\lambda}} \\ &= 1 - e^{-90\times\frac{8}{486}} \\ &\simeq 0.772699.\end{aligned}$$

2. O engenheiro mecânico responsável pela produção de um dado tipo de pneus admite que a sua duração (em milha percorrida) possui distribuição normal. Foram testados 10 pneus desse tipo, tendo-se obtido uma amostra com média e variância iguais a $\bar{x} = 37\,840$ e $s^2 = 6\,556\,000$, respetivamente.

(a) Obtenha intervalos de confiança a 90% para o valor esperado e para a mediana da duração desse tipo de pneu. (2.5)

- **V.a. de interesse**

  X = duração (em milhas percorridas) de novo tipo de pneu

- **Situação**

  $X \sim \text{Normal}(\mu,\sigma^2)$

  μ DESCONHECIDO

  σ^2 desconhecido

- **Obtenção do IC para μ**

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

$$Z = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

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

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

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

$$\begin{cases} a_\alpha = F^{-1}_{t_{(n-1)}}(\alpha/2) = F^{-1}_{t_{(10-1)}}(0.05) = -F^{-1}_{t_{(9)}}(1-0.05) \stackrel{tabela/calc.}{=} -1.833 \\ b_\alpha = F^{-1}_{t_{(n-1)}}(1-\alpha/2) = F^{-1}_{t_{(10-1)}}(0.95) \stackrel{tabela/calc.}{=} 1.833. \end{cases}$$

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

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

$$P\left[a_\alpha \le \frac{\bar{X}-\mu}{S/\sqrt{n}} \le b_\alpha\right] = 1-\alpha$$

$$P\left[\bar{X} - b_\alpha \times \frac{S}{\sqrt{n}} \le \mu \le \bar{X} - a_\alpha \times \frac{S}{\sqrt{n}}\right] = 1-\alpha$$

Passo 4 — Concretização
Atendendo aos quantis acima, às concretizações de $\bar{X}$ e S^2,

$$\begin{aligned} \bar{x} &= 37\,840 \\ s^2 &= 6\,556\,000, \end{aligned}$$

e ao facto de

$$IC_{(1-\alpha)\times 100\%}(\mu) = \left[\bar{x} - F^{-1}_{t_{(n-1)}}(1-\alpha/2) \times \frac{s}{\sqrt{n}}, \quad \bar{x} + F^{-1}_{t_{(n-1)}}(1-\alpha/2) \times \frac{s}{\sqrt{n}}\right],$$

temos

$$\begin{aligned} IC_{90\%}(\mu) &= \left[37\,840 - 1.833 \times \frac{\sqrt{6\,556\,000}}{\sqrt{10}}, \quad 37\,840 + 1.833 \times \frac{\sqrt{6\,556\,000}}{\sqrt{10}}\right] \\ &= [36\,355.8,\ 39\,324.2]. \end{aligned}$$

- **Comentário**
A f.d.p. de X é simétrica em relação a μ logo $me = me(X) \equiv E(X) = \mu$. Por consequência, $IC_{90\%}(me) \equiv IC_{90\%}(\mu)$.

(b) Um engenheiro afirma que o valor esperado da duração desse tipo de pneus é igual a 40 000 milhas, ao passo que uma consultora defende que tal valor esperado é inferior a 40 000 milhas. Confronte a hipótese defendida pelo engenheiro com a defendida pela consultora, tomando para o efeito uma decisão com base no valor-p. (2.5)

- **Hipóteses**
$H_0 : \mu = \mu_0 = 40\,000$
$H_1 : \mu < \mu_0$

- **Estatística de teste**

$$T = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim_{H_0} t_{(n-1)}$$

[pois pretendemos efectuar um teste sobre o valor esperado de uma população normal, com variância desconhecida.]

- **Região de rejeição de H_0** (para valores de T)
Tratando-se de um teste unilateral inferior ($H_1 : \mu < \mu_0$), 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)**
O valor observado da estatística de teste é igual a

$$
\begin{aligned}
t &= \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} \\
&= \frac{37\,840-40\,000}{\sqrt{\frac{6\,556\,000}{10}}} \\
&\simeq -2.66768.
\end{aligned}
$$

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

$$
\begin{aligned}
valor-p &= P(T < t \mid H_0) \\
&= F_{t_{(n-1)}}(t) \\
&= F_{t_{(9)}}(-2.66768) \\
&= 1-F_{t_{(9)}}(2.66768) \\
&\overset{calc.}{=} 1-0.987140 \\
&= 0.012860.
\end{aligned}
$$

Assim sendo, é suposto:

- não rejeitar H_0 a qualquer n.s. $\alpha_0 \leq 1.2860\%$, pelo que $H_0 : \mu = \mu_0 = 40\,000$ é consistente por exemplo ao n.s. de 1%;
- rejeitar H_0 a qualquer n.s. $\alpha_0 > 1.2860\%$, nomeadamente aos n.u.s. de 5% e 10%.

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

$$
\begin{array}{rcccl}
F^{-1}_{t_{(9)}}(0.975) = 2.262 & < & 2.66768 & < & 2.821 = F^{-1}_{t_{(9)}}(0.99) \\
0.01 = 1-0.99 & < & valor-p = 1-F_{t_{(9)}}(2.66768) & < & 1-0.975 = 0.025.
\end{array}
$$

Consequentemente:

- não devemos rejeitar H_0 a qualquer n.s. $\alpha_0 \leq 1\%$;
- devemos rejeitar H_0 a qualquer n.s. $\alpha_0 > 2.5\%$, nomeadamente aos n.u.s. de 5% e 10%.]

Grupo II	10 valores

1. Para averiguar sobre a validade da hipótese de o número de furtos diários numa certa zona ser descrito por uma variável aleatória com distribuição de Poisson, recolheram-se dados relativos a 60 dias de observação: (3.5)

Número de furtos diários	0	1	2	3	4 ou mais
Frequência absoluta observada	8	20	17	5	10

Sabendo que a estimativa de máxima verosimilhança do valor esperado do número diário de furtos, calculada com base na amostra agrupada, é igual a 1.9, teste a hipótese referida, ao nível de significância de 5%.

- **V.a. de interesse e f.p.**
  X = número de furtos diários
- **Hipóteses**
  $H_0 : X \sim \text{Poisson}(\lambda)$, onde λ é um parâmetro DESCONHECIDO
  $H_1 : X \not\sim \text{Poisson}(\lambda)$
- **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 = ESTIMADOR da frequência absoluta esperada, sob H_0, da classe i

β = No. de parâmetros a estimar = 1 [dado que em H_0 se conjectura uma família de distribuições e não uma distribuição em particular.]

- **Estimativas das frequências absolutas esperadas sob** H_0

[Como λ é desconhecido o mesmo acontece com p_i^0 e com a frequência absoluta esperada, sob H_0, da classe i,

$$n \times p_i^0 \quad = \quad n \times \begin{cases} \frac{e^{-\lambda} \times \lambda^{i-1}}{(i-1)!}, & i = 1,2,3,4 \\ 1 - (p_1^0 + \ldots p_4^0), & i = 5.] \end{cases}$$

Uma vez que a estimativa de MV do valor esperado (avaliada com base na amostra agrupada) é igual a $\hat{\lambda} = 1.9$, podemos obter as seguintes estimativas das frequências absolutas esperadas sob H_0:

$$\begin{aligned} e_i &= n \times \hat{p}_i^0 \\ &= n \times \begin{cases} \frac{e^{-\hat{\lambda}} \times \hat{\lambda}^{i-1}}{(i-1)!}, & i = 1,2,3,4 \\ 1 - (\hat{p}_1^0 + \ldots \hat{p}_4^0), & i = 5. \end{cases} \\ &\overset{\hat{\lambda}=1.9}{=} 60 \times \begin{cases} F_{Poisson(1.9)}(0) \overset{tabela/calc.}{=} 0.1496, & i = 1 \\ F_{Poisson(1.9)}(1) - F_{Poisson(1.9)}(0) \overset{tabela/calc.}{=} 0.4337 - 0.1496 = 0.2841, & i = 2 \\ F_{Poisson(1.9)}(2) - F_{Poisson(1.9)}(1) \overset{tabela/calc.}{=} 0.7037 - 0.4337 = 0.2700, & i = 3 \\ F_{Poisson(1.9)}(3) - F_{Poisson(1.9)}(2) \overset{tabela/calc.}{=} 0.8747 - 0.7037 = 0.1710, & i = 4 \\ 1 - (0.1496 + 0.2841 + 0.2700 + 0.1710) = 0.1253, & i = 5 \end{cases} \\ &= \begin{cases} 8.976, & i = 1 \\ 17.046, & i = 2 \\ 16.200, & i = 3 \\ 10.260, & i = 4 \\ 7.518, & i = 5. \end{cases} \end{aligned}$$

[Constata-se 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

$$\begin{aligned} c &= F^{-1}_{\chi^2_{(k-\beta-1)}}(1 - \alpha_0) \\ &= F^{-1}_{\chi^2_{(5-1-1)}}(1 - 0.05) \\ &= F^{-1}_{\chi^2_{(3)}}(0.95) \\ &\overset{tabela/calc.}{=} 7.815. \end{aligned}$$

- **Decisão**

No cálculo do valor observado da estatística de teste convém recorrer à seguinte tabela auxiliar:

i	Classe i	Freq. abs. obs. o_i	Estim. freq. abs. esp. sob H_0 $e_i = n \times \hat{p}_i^0$	Parcelas valor obs. estat. teste $\frac{(o_i-e_i)^2}{e_i}$
1	$\{0\}$	8	8.976	$\frac{(8-8.976)^2}{8.976} \simeq 0.106$
2	$\{1\}$	20	17.046	0.512
3	$\{2\}$	17	16.200	0.040
4	$\{3\}$	5	10.260	2.697
5	$\{4,5,\dots\}$	10	7.518	0.819
		$\sum_{i=1}^k o_i = n$ $= 60$	$\sum_{i=1}^k e_i = n$ $= 60$	$t = \sum_{i=1}^k \frac{(o_i-e_i)^2}{e_i}$ $\simeq 4.174$

Como $t \simeq 4.174 \not\in W = (7.815, +\infty)$, não devemos rejeitar H_0 ao n.s. de $\alpha_0 = 5\%$ [nem a qualquer outro n.s. inferior a α_0].

2. A massa de certa espécie de peixe é descrita pela variável x (em grama) e o respectivo comprimento é representado pela variável aleatória Y (em centímetro). Uma amostra de 30 peixes desta espécie conduziu aos seguintes valores:

$$\sum_{i=1}^{30} x_i = 875.4, \quad \sum_{i=1}^{30} x_i^2 = 25778.00, \quad \sum_{i=1}^{30} y_i = 605.7, \quad \sum_{i=1}^{30} y_i^2 = 12553.10, \quad \sum_{i=1}^{30} x_i\, y_i = 17931.50,$$

onde $\left[\min_{i=1,\dots,30} x_i, \max_{i=1,\dots,30} x_i\right] = [27.9\,,31.1]$.

(a) Calcule as estimativas de mínimos quadrados dos parâmetros da reta de regressão linear simples de Y em x. (1.5)

- **Estimativas de MQ de β_0 e β_1**

  Dado que

$$n = 30$$

$$\sum_{i=1}^n x_i = 875.4$$
$$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i = \frac{875.4}{30} = 29.18$$
$$\sum_{i=1}^n x_i^2 = 25778.00$$
$$\sum_{i=1}^n x_i^2 - n\,(\bar{x})^2 = 25778.00 - 30 \times 29.18^2 = 233.828$$

$$\sum_{i=1}^n y_i = 605.7$$
$$\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i = \frac{605.7}{30} = 20.19$$
$$\sum_{i=1}^n y_i^2 = 12553.1$$
$$\sum_{i=1}^n y_i^2 - n\,(\bar{y})^2 = 12553.1 - 30 \times 20.19^2 = 324.017$$

$$\sum_{i=1}^n x_i y_i = 17931.50$$
$$\sum_{i=1}^n x_i y_i - n\,\bar{x}\,\bar{y} = 17931.50 - 30 \times 29.18 \times 20.19 = 257.174,$$

  as estimativas de MQ de β_1 e β_0 são, para este modelo de RLS, iguais a:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^n x_i^2 - n\,(\bar{x})^2} \\
&= \frac{257.174}{233.828} \\
&\simeq 1.099843
\end{aligned}$$

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \times \bar{x} \\
&\simeq 20.19 - 1.099843 \times 29.18 \\
&\simeq -11.903419.
\end{aligned}$$

(b) Obtenha a estimativa de mínimos quadrados do valor esperado do comprimento de um peixe desta espécie, se a sua massa for igual a 29.1g. (1.0)

- **Estimativa de MQ para** $E(Y \mid x_0) = \beta_0 + \beta_1 x_0$

$$\begin{aligned}\hat{E}(Y \mid x_0) &= \hat{\beta}_0 + \hat{\beta}_1 x_0 \\ &\simeq -11.903419 + 1.099843 \times 29.1 \\ &\simeq 20.102012.\end{aligned}$$

[Não estamos a cometer qualquer erro de extrapolação ao estimar pontualmente $E(Y \mid x = 29.1) = \beta_0 + \beta_1 \times 29.1$ dado que $29.1 \in [\min_{i=1,\ldots,30} x_i, \max_{i=1,\ldots,30} x_i] = [27.9, 31.1]$.]

(c) Teste a significância do modelo de regressão linear simples, ao nível de significância de 5%. Enuncie as hipóteses de trabalho que assumir. (3.0)

- **Hipóteses de trabalho**
  No modelo de RLS, $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, consideraremos $\epsilon_i \overset{i.i.d.}{\sim} \text{Normal}(0, \sigma^2)$, $i = 1, \ldots, n$.
- **[Obs.**
  Pretende confrontar-se
  - $H_0 : \beta_1 = \beta_{1,0} = 0$ (regressão linear não é significativa, i.e., o valor esperado da variável resposta Y, $E(Y) = \beta_0 + \beta_1 x$, não depende linearmente da variável explicativa x)
  - $H_1 : \beta_1 \neq \beta_{1,0}$ (regressão linear é significativa).]
- **Hipóteses**
  $H_0 : \beta_1 = \beta_{1,0} = 0$
  $H_1 : \beta_1 \neq \beta_{1,0}$
- **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 ($H_1 : \beta_1 \neq \beta_{1,0}$), pelo que a região de rejeição de H_0 é $W = (-\infty, -c) \cup (c, +\infty)$, onde

$$\begin{aligned} c \quad : \quad & P(\text{Rejeitar } H_0 | H_0) = \alpha_0 \\ & c = F^{-1}_{t_{(n-2)}}\left(1 - \frac{\alpha_0}{2}\right) \\ & c = F^{-1}_{t_{(28)}}(0.975) \\ & c \overset{tabela/calc}{=} 2.048 \end{aligned}$$

- **Decisão**
  Tendo em conta que

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-2}\left[\left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right) - (\hat{\beta}_1)^2 \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right)\right] \\ &\simeq \frac{1}{30-2}(324.017 - 1.099843^2 \times 233.828) \\ &\simeq 1.470210,\end{aligned}$$

o valor observado da estatística de teste é dado por

$$
\begin{aligned}
t &= \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n x_i^2 - n\,\bar{x}^2}}} \\
&= \frac{1.099843 - 0}{\sqrt{\frac{1.470210}{233.828}}} \\
&\simeq 13.870409.
\end{aligned}
$$

Como $t \simeq 13.870409 \in W = (-\infty, -2.048) \cup (2.048, +\infty)$, devemos rejeitar H_0 [hipótese de inexistência de relação de tipo linear entre o valor esperado da variável resposta Y e a variável explicativa x], ao nível de significância $\alpha_0 = 5\%$ [ou a qualquer outro n.s. superior 5%].

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

- **Cálculo do coeficiente de determinação**

$$
\begin{aligned}
r^2 &= \frac{\left(\sum_{i=1}^n x_i y_i - n\,\bar{x}\,\bar{y}\right)^2}{\left(\sum_{i=1}^n x_i^2 - n\,\bar{x}^2\right) \times \left(\sum_{i=1}^n y_i^2 - n\,\bar{y}^2\right)} \\
&= \frac{257.174^2}{233.828 \times 324.017} \\
&\simeq 0.873.
\end{aligned}
$$

- **Interpretação coeficiente de determinação**

  Cerca de 87.3% da variação total do comprimentos dos peixes é explicada pela respectiva massa, através do modelo de regressão linear simples ajustado. Logo podemos adiantar que a recta estimada parece ajustar-se bem ao conjunto de dados.