

Probabilidades e Estatística

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

2º semestre – 2011/2012

19/07/2012 – 9:00

Exame de Época Especial

Duração: 3 horas

Justifique convenientemente **todas as respostas!**

Grupo I

5 valores

1. De uma caixa, contendo 2 bolas azuis e 3 bolas vermelhas, retira-se ao acaso uma bola e coloca-se numa segunda caixa que já contém 4 bolas azuis e 2 bolas vermelhas. De seguida, extrai-se ao acaso uma bola da segunda caixa.

(a) Qual é a probabilidade de extrair bolas da mesma cor das duas caixas? (1.0)

- **Experiência aleatória**

Seleção ao acaso de uma bola de uma caixa contendo 2 bolas azuis (A) e 3 vermelhas (V), bola essa colocada numa segunda caixa que já contém 4 bolas azuis e 2 vermelhas, seguida de seleção ao acaso de uma bola da segunda caixa.

- **Importante**

Dado que a primeira bola seleccionada é colocada na segunda caixa a proporção de bolas passará de 4 bolas azuis e 2 vermelhas para:

- 5 bolas azuis e 2 vermelhas, caso a primeira bola seleccionada e colocada na segunda caixa seja azul;
- 4 bolas azuis e 3 vermelhas, caso a primeira bola seleccionada e colocada na segunda caixa seja vermelha;

- **Eventos-chave**

A_i = selecção de uma bola azul na i – ésima extracção, $i = 1, 2$

V_i = selecção de uma bola vermelha na i – ésima extracção, $i = 1, 2$

- **Evento**

$\{\text{Seleccionar bolas da mesma cor das duas caixas}\} = \{A_1 \cap A_2, V_1 \cap V_2\}$

- **Prob. pedida**

Tendo em conta que se pretende a probabilidade da reunião de eventos disjuntos e aplicando a lei das probabilidades compostas, obtém-se sucessivamente:

$$\begin{aligned} P(\{A_1 \cap A_2, V_1 \cap V_2\}) &= P(A_1 \cap A_2) + P(V_1 \cap V_2) \\ &= P(A_1)P(A_2 | A_1) + P(V_1)P(V_2 | V_1) \\ &= \frac{2}{2+3} \times \frac{4+1}{4+2+1} + \frac{3}{2+3} \times \frac{2+1}{4+2+1} \\ &= \frac{2}{5} \times \frac{5}{7} + \frac{3}{5} \times \frac{3}{7} \\ &= \frac{10}{35} + \frac{9}{35} \\ &= \frac{19}{35}. \end{aligned}$$

(b) Determine a probabilidade de a bola extraída da segunda caixa ser vermelha. (0.5)

- **Evento**

$\{\text{Seleccionar bola vermelha da 2a. caixa}\} = V_2$

- **Prob. pedida**

$$\begin{aligned}
 P(V_2) &= P(\{A_1 \cap V_2, V_1 \cap V_2\}) \\
 &= P(A_1 \cap V_2) + P(V_1 \cap V_2) \\
 &= P(A_1)P(V_2 | A_1) + P(V_1 \cap V_2) \\
 &\stackrel{(a)}{=} \frac{2}{2+3} \times \frac{2+0}{4+2+1} + \frac{9}{35} \\
 &= \frac{2}{5} \times \frac{2}{7} + \frac{9}{35} \\
 &= \frac{13}{35}.
 \end{aligned}$$

(c) Se a bola extraída da segunda caixa é vermelha, qual é a probabilidade de se ter extraído da primeira caixa uma bola dessa mesma cor? (1.0)

- **Prob. pedida**

Tirando partido dos resultados anteriores, segue-se

$$\begin{aligned}
 P(V_1 | V_2) &= \frac{P(V_1 \cap V_2)}{P(V_2)} \\
 &\stackrel{(a)}{=} \frac{\frac{9}{35}}{\frac{13}{35}} \\
 &= \frac{9}{13}.
 \end{aligned}$$

2. Num determinado cruzamento, o número diário de acidentes rodoviários é uma variável aleatória com distribuição de Poisson de valor esperado 0.001. Considere que, em dias diferentes, os acidentes nesse cruzamento ocorrem de forma independente.

(a) Determine a probabilidade de num ano (365 dias) ocorrerem um ou mais acidentes nesse cruzamento em pelo menos dois dias. (1.5)

- **V.a.**

X = número diário de acidentes por dia no cruzamento

- **Distribuição de X**

$X \sim \text{Poisson}(\lambda)$

- **Parâmetro**

$\lambda : E(X) = 0.001 \Leftrightarrow \lambda = 0.001$

- **F.p. de X**

$P(X = x) = e^{-0.001} \frac{0.001^x}{x!}, x = 0, 1, 2, \dots$

- **Outra v.a.**

Y = número de dias, em 365, em que ocorrem acidentes no cruzamento

- **Distribuição de Y**

Ao admitir-se também que em dias diferentes os acidentes nesse cruzamento ocorrem de forma independente, a v.a. Y corresponde ao número de sucessos em n provas de Bernoulli i.i.d., pelo que $Y \sim \text{Binomial}(n, p)$

- **Parâmetros**

$n = 365$

$p = P(\text{ocorrerem acidentes num dia}) = P(X > 0) = 1 - P(X = 0) = 1 - e^{-0.001} \simeq 0.001$

- **F.p. de Y**

$P(Y = y) = \binom{365}{y} (1 - e^{-0.001})^y (e^{-0.001})^{365-y} y = 0, 1, \dots, 365$

- **Prob. pedida**

$$\begin{aligned}
 P(Y \geq 2) &= 1 - P(Y \leq 1) \\
 &= 1 - \sum_{y=0}^1 \binom{365}{y} (1 - e^{-0.001})^y (e^{-0.001})^{365-y} \\
 &\simeq 1 - (0.999^{365} + 365 \times 0.001 \times 0.999^{364}) \\
 &\simeq 1 - 0.947589 \\
 &= 0.052341.
 \end{aligned}$$

- (b) *Determine o valor esperado e o desvio padrão do número de dias que decorrem até que ocorra pelo menos um acidente nesse cruzamento.* (1.0)

- **Outra v.a.**

Z = número de dias que decorrem até que se verifique um dia com acidentes no cruzamento

- **Distribuição de Z**

Z corresponde ao número provas de Bernoulli i.i.d. até à ocorrência do primeiro sucesso, logo $Z \sim \text{Geométrica}(p)$.

- **Parâmetro**

$$p = P(X > 0) = 1 - P(X = 0) = 1 - e^{-0.001} \simeq 0.001$$

- **Valor esperado de Z**

$$E(Z) \stackrel{\text{form}}{=} \frac{1}{p} \simeq \frac{1}{0.001} = 1\,000 \text{ dias}$$

- **Variância de Z**

$$V(Z) \stackrel{\text{form}}{=} \frac{1-p}{p^2} \simeq \frac{1-0.001}{0.001^2} = 999\,000 \text{ dias}^2$$

- **Desvio-padrão de Z**

$$DP(Z) = \sqrt{V(Z)} \simeq \sqrt{999\,000} \simeq 999.5 \text{ dias.}$$

Grupo II	5 valores
-----------------	-----------

1. *Considere um círculo de raio X (em cm), onde X é uma variável aleatória exponencial de parâmetro igual a 1.*

- (a) *Determine a probabilidade de o diâmetro do círculo não exceder 2 cm. Obtenha a variância do diâmetro do círculo.* (1.0)

- **V.a.**

X = raio do círculo

- **Distribuição de X**

$X \sim \text{Exponencial}(\lambda = 1)$

- **F.d.p. de X**

$$f_X(x) = e^{-x}, x \geq 0$$

- **Outra v.a.**

D = diâmetro do círculo = $2X$

- **Prob. pedida**

$$\begin{aligned}
 P(D \leq 2) &= P(2X \leq 2) \\
 &= P(X \leq 1) \\
 &= \int_{-\infty}^1 f_X(x) dx \\
 &= \int_0^1 e^{-x} dx
 \end{aligned}$$

$$\begin{aligned}
&= (-e^{-x}) \Big|_0^1 \\
&= 1 - e^{-1} \\
&\simeq 0.632121.
\end{aligned}$$

$$V(D) = V(2X) = 4V(X) = 4 \times 1 = 4$$

- (b) Considere agora um novo círculo de raio Y (em cm), onde Y é uma variável aleatória independente de X , também com distribuição exponencial de parâmetro igual a 1. Considerando que os dois círculos estão representados no mesmo plano e que os respectivos centros estão separados por uma distância de 2 cm, determine a probabilidade de os círculos se sobreponem. (1.0)

• **Outra v.a.**

Y = raio de novo círculo

• **Distribuição de Y**

$Y \sim \text{Exponencial}(\lambda = 1)$

$Y \perp\!\!\!\perp X$

• **F.d.p. conjunta de (X, Y)**

Uma vez que X e Y são v.a. independentes, ambas com distribuição Exponencial($\lambda = 1$), tem-se

$$\begin{aligned}
f_{X,Y}(x, y) &= f_X(x) \times f_Y(y) \\
&= e^{-x} \times e^{-y}, \quad x, y \geq 0.
\end{aligned}$$

• **Probab. pedida**

$$\begin{aligned}
P(X + Y > 2) &= 1 - P(X + Y \leq 2) \\
&= 1 - P((X, Y) \in \{(x, y) \in \mathbb{R}^2 : x + y \leq 2\}) \\
&= 1 - \int_{-\infty}^2 \int_{-\infty}^{2-x} f_{X,Y}(x, y) \, dy \, dx \\
&= 1 - \int_0^2 \int_0^{2-x} e^{-x} \times e^{-y} \, dy \, dx \\
&= 1 - \int_0^2 e^{-x} (-e^{-y}) \Big|_0^{2-x} \, dx \\
&= 1 - \int_0^2 e^{-x} [1 - e^{-(2-x)}] \, dx \\
&= 1 - (-e^{-x} - e^{-2}x) \Big|_0^2 \\
&= 1 - (1 - e^{-2} - 2 \times e^{-2}) \\
&= 3 \times e^{-2} \\
&\simeq 0.406006.
\end{aligned}$$

2. O número de computadores de determinado modelo vendidos diariamente numa loja é uma variável aleatória com distribuição uniforme discreta em $\{0, 1, 2, 3, 4\}$. Considera-se que os números de computadores daquele modelo vendidos nessa loja em dias diferentes são variáveis aleatórias independentes.

- (a) Calcule um valor aproximado da probabilidade de em 30 dias serem vendidos na loja mais de 50 computadores daquele modelo. (2.0)

• **V.a.**

X = número de computadores de certo modelo vendidos diariamente

• **Distribuição de X**

$X \sim \text{Uniforme Discreta}(\{0, 1, 2, 3, 4\})$

- **F.p., valor esperado e variância de X_i**

$$P(X = x) = \frac{1}{5}, x = 0, 1, \dots, 4$$

$$E(X) = \mu = \sum_{x=0}^4 x \times P(X = x) = \frac{0+1+2+3+4}{5} = 2$$

$$V(X) = \sigma^2 = E(X^2) - E^2(X) = \sum_{x=0}^4 x^2 \times P(X = x) - 2^2 = \frac{0^2+1^2+2^2+3^2+4^2}{5} - 2^2 = 6 - 4 = 2$$

- **V.a.**

X_i = número de computadores de certo modelo vendidos no i - ésimo dia, $i = 1, \dots, 30$

- **Distribuição, valor esperado e variância de X_i**

$$X_i \stackrel{i.i.d.}{\sim} X, i = 1, \dots, 30$$

$$E(X_i) = E(X) = \mu = 2$$

$$V(X_i) = V(X) = \sigma^2 = 2$$

- **Nova v.a.**

$S = \sum_{i=1}^{30} X_i$ = número de computadores de certo modelo vendidos em 30 dias

- **Valor esperado e variância de S**

$$E(S) = E\left(\sum_{i=1}^{30} X_i\right) = \sum_{i=1}^{30} E(X_i) \stackrel{X_i \sim X}{=} 30 \times E(X) = 30 \times 2 = 60$$

$$V(S) = V\left(\sum_{i=1}^{30} X_i\right) \stackrel{X_i \text{ indep.}}{=} \sum_{i=1}^{30} V(X_i) \stackrel{X_i \sim X}{=} 30 \times V(X) = 30 \times 2 = 60$$

- **Distribuição aproximada de S**

Pelo Teorema do Limite Central (TLC) pode escrever-se

$$\frac{S - E(S)}{\sqrt{V(S)}} = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}} \stackrel{a}{\sim} \text{Normal}(0, 1).$$

- **Prob. pedida — valor aproximado**

$$\begin{aligned} P(S > 50) &= 1 - P\left[\frac{S - E(S)}{\sqrt{V(S)}} \leq \frac{50 - E(S)}{\sqrt{V(S)}}\right] \\ &= 1 - P\left[\frac{S - E(S)}{\sqrt{V(S)}} \leq \frac{50 - 60}{\sqrt{60}}\right] \\ &\stackrel{TLC}{\simeq} 1 - \Phi\left(\frac{50 - 60}{\sqrt{60}}\right) \\ &\simeq 1 - \Phi(-1.29) \\ &= \Phi(1.29) \\ &\stackrel{tabela}{=} 0.9015. \end{aligned}$$

- (b) *A loja só pode fazer encomendas de 30 em 30 dias. Qual o número mínimo de unidades do referido modelo de computador que deve existir em stock, no início de um desses períodos de 30 dias, por forma a que o valor aproximado da probabilidade de haver rotura de stock nesse período seja no máximo de 5%?* (1.0)

- **V.a.**

$S = \sum_{i=1}^{30} X_i$ = número de computadores de certo modelo vendidos em 30 dias

- **Distribuição aproximada de S**

Tivemos ocasião de referir que

$$\frac{S - E(S)}{\sqrt{V(S)}} = \frac{\sum_{i=1}^n X_i - 60}{\sqrt{60}} \stackrel{a}{\sim}_{TLC} \text{Normal}(0, 1).$$

- **Obtenção do stock mínimo**

$$\begin{aligned}
 k &: P(S > k) \leq 0.05 \\
 1 - P(S \leq k) &\leq 0.05 \\
 P(S \leq k) &\geq 0.95 \\
 \Phi\left(\frac{k-60}{\sqrt{60}}\right) &\geq 0.95 \\
 \frac{k-60}{\sqrt{60}} &\geq \Phi^{-1}(0.95) \\
 k &\geq 60 + \sqrt{60} \times \Phi^{-1}(0.95) \\
 k &\geq 60 + \sqrt{60} \times 1.6449 \\
 k &\geq 72.741341.
 \end{aligned}$$

O menor valor de k que satisfaz a condição $P(S > k) \leq 0.05$ é $k^* = 73$.

Grupo III	5 valores
------------------	-----------

1. O tempo (em 10^3 h) até fractura de um tipo de rolamentos de esferas produzidos por um fabricante é uma variável aleatória X com função de densidade de probabilidade

$$f_X(x) = \begin{cases} \frac{2x}{\lambda^2} e^{-\frac{x^2}{\lambda^2}}, & x > 0 \\ 0, & x \leq 0, \end{cases}$$

onde λ é uma constante positiva desconhecida.

- (a) Com base numa amostra aleatória de X , (X_1, \dots, X_n) , deduza o estimador de máxima verosimilhança do parâmetro λ . (1.5)

- **V.a. de interesse**

X = tempo até fractura (em 10^3 h)

- **F.d.p. de X**

$$f_X(x) = \begin{cases} \frac{2x}{\lambda^2} e^{-\frac{x^2}{\lambda^2}}, & x > 0 \\ 0, & x \leq 0, \end{cases}$$

- **Parâmetro DESCONHECIDO**

λ ($\lambda > 0$)

- **Amostra**

$\underline{x} = (x_1, \dots, x_n)$ amostra de dimensão n proveniente da população X

- **Obtenção do estimador de MV de β**

Passo 1 — Função de verosimilhança

$$\begin{aligned}
 L(\lambda|\underline{x}) &\stackrel{X_i \text{ indep}}{=} \prod_{i=1}^n f_{X_i}(x_i) \\
 &\stackrel{X_i \sim X}{=} \prod_{i=1}^n \left(\frac{2x_i}{\lambda^2} e^{-\frac{x_i^2}{\lambda^2}} \right) \\
 &= 2^n \lambda^{-2n} \left(\prod_{i=1}^n x_i \right) e^{-\frac{1}{\lambda^2} \sum_{i=1}^n x_i^2}, \lambda > 0
 \end{aligned}$$

Passo 2 — Função de log-verosimilhança

$$\ln L(\lambda|\underline{x}) = n \ln(2) - 2n \ln(\lambda) + \sum_{i=1}^n \ln(x_i) - \frac{1}{\lambda^2} \sum_{i=1}^n x_i^2$$

Passo 3 — Maximização

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

$$\hat{\lambda} : \begin{cases} \frac{d \ln L(\lambda|\underline{x})}{d\lambda} \Big|_{\lambda=\hat{\lambda}} = 0 & \text{(ponto de estacionaridade)} \\ \frac{d^2 \ln L(\lambda|\underline{x})}{d\lambda^2} \Big|_{\lambda=\hat{\lambda}} < 0 & \text{(ponto de máximo)} \\ -\frac{2n}{\hat{\lambda}} + \frac{2}{\hat{\lambda}^3} \sum_{i=1}^n x_i^2 = 0 \\ \frac{2n}{\hat{\lambda}^2} - \frac{6}{\hat{\lambda}^4} \sum_{i=1}^n x_i^2 < 0 \\ \hat{\lambda} = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2} \\ -\frac{4n^2}{\sum_{i=1}^n x_i^2} < 0, \text{ proposição verdadeira pois } \sum_{i=1}^n x_i^2 > 0 \end{cases}$$

Passo 4 — Estimador de MV de λ

Será representado pela v.a. $\text{EMV}(\lambda) = \sqrt{\frac{1}{n} \sum_{i=1}^n X_i^2}$.

- (b) *Obtenha a estimativa de máxima verosimilhança da probabilidade do tempo até fractura ser superior a 3 (em 10^3 h), sabendo que uma amostra de 20 rolamentos, (x_1, \dots, x_{20}) , conduziu ao valor de 4.5 para a estimativa de máxima verosimilhança do parâmetro λ .* (0.5)

• **Outro parâmetro DESCONHECIDO**

$$\begin{aligned} h(\lambda) &= P(X > 3) \\ &= \int_3^{+\infty} \frac{2x}{\lambda^2} e^{-\frac{x^2}{\lambda^2}} dx \\ &= \left(-e^{-\frac{x^2}{\lambda^2}} \right) \Big|_3^{+\infty} \\ &= e^{-\frac{9}{\lambda^2}} \end{aligned}$$

• **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 $h(\lambda) = P(X > 3) = e^{-\frac{9}{\lambda^2}}$ é

$$\begin{aligned} \widehat{h(\lambda)} &= h(\hat{\lambda}) \\ &= e^{-\frac{9}{\hat{\lambda}^2}} \\ &= e^{-\frac{9}{4.5^2}} \\ &\simeq 0.641180. \end{aligned}$$

2. *Com o objectivo de comparar a autonomia de 2 modelos de telemóveis (A e B), recolheu-se a seguinte informação sobre as durações das cargas das baterias de aparelhos escolhidos ao acaso de cada um dos modelos:*

Modelo	Nº de aparelhos testados	Média amostral	Variância amostral (corrigida)
A	80	6.5 h	1.7^2 h^2
B	100	5.5 h	2.5^2 h^2

- (a) *Construa um intervalo com nível de confiança de aproximadamente 95% para o valor esperado da duração da carga de baterias do modelo A.* (1.5)

• **V.a. de interesse**

X_A = duração da carga de baterias do modelo A

• **Situação**

X_A com distribuição arbitrária

$E(X_A) = \mu_A$ DESCONHECIDO

$V(X_A) = \sigma_A^2$ desconhecida

n_A suficientemente grande (maior que 30).

• **Passo 1 — Selecção da v.a. fulcral para μ_A**

$$Z = \frac{\bar{X}_A - \mu_A}{\frac{S_A}{\sqrt{n_A}}} \stackrel{a}{\sim} \text{normal}(0, 1)$$

uma vez que pretendemos um intervalo aproximado para o valor esperado de uma população com distribuição arbitrária com variância desconhecida e dispomos de amostra com dimensão suficientemente grande.

• **Passo 2 — Obtenção dos quantis de probabilidade**

Como o nível aproximado de confiança é de $(1 - \alpha) \times 100\% = 95\%$ (i.e., $\alpha = 0.05$), recorre-se aos quantis

$$\begin{cases} a_\alpha = -\Phi^{-1}(1 - \alpha/2) = -\Phi^{-1}(0.975) \stackrel{tabela}{=} -1.96 \\ b_\alpha = \Phi^{-1}(1 - \alpha/2) = \Phi^{-1}(0.975) \stackrel{tabela}{=} 1.96, \end{cases}$$

sendo que estes quantis enquadram a v.a. fulcral Z com probabilidade aproximadamente igual a $(1 - \alpha) = 0.95$.

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

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

$$P\left[a_\alpha \leq \frac{\bar{X}_A - \mu_A}{\frac{S_A}{\sqrt{n_A}}} \leq b_\alpha\right] \simeq 1 - \alpha$$

...

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

• **Passo 4 — Concretização**

Neste caso o IC aproximado a $(1 - \alpha) \times 100\%$ é dado por

$$IC(\mu_A) = \left[\bar{x}_A - \Phi^{-1}(1 - \alpha/2) \times \frac{s_A}{\sqrt{n_A}}, \bar{x}_A + \Phi^{-1}(1 - \alpha/2) \times \frac{s_A}{\sqrt{n_A}} \right].$$

Como $n_A = 80$, $\bar{x}_A = 6.5$, $s_A^2 = 1.7^2$, $\alpha = 0.05$ e $\Phi^{-1}(1 - \alpha/2) = 1.96$, tem-se

$$\begin{aligned} IC(\mu_X - \mu_Y) &= \left[6.5 \pm 1.96 \times \frac{1.7}{\sqrt{80}} \right] \\ &\simeq [6.5 \pm 0.372529] \\ &= [6.127471, 6.872529]. \end{aligned}$$

- (b) *Poderá afirmar-se que, ao nível de significância de 2%, há evidência de o valor esperado da duração da carga da bateria do modelo A ser superior ao valor esperado da duração da carga da bateria do modelo B?* (1.5)

• **V.a. de interesse**

X_A = duração da carga da bateria do modelo A

X_B = duração da carga da bateria do modelo B

• **Situação**

X_A e X_B v.a. independentes com distribuições arbitrárias

$E(X_A) - E(X_B) = \mu_A - \mu_B$ DESCONHECIDA

$V(X_A) = \sigma_A^2$ e $V(X_B) = \sigma_B^2$ desconhecidas

n_A, n_B suficientemente grandes (maiores que 30).

As a.a. associadas a X_A e X_B são independentes, pelo que \bar{X}_A e \bar{X}_B são v.a. independentes.

• **Hipóteses**

$H_0 : \mu_A - \mu_B = \mu_0 = 0$

$H_1 : \mu_A - \mu_B > \mu_0 = 0$

- **Nível de significância**

$$\alpha_0 = 2\%$$

- **Estatística de teste**

$$T = \frac{(\bar{X}_A - \bar{X}_B) - \mu_0}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}} \underset{a}{\sim}_{H_0} \text{normal}(0, 1)$$

dado que se pretende efectuar um teste sobre a diferença de valores esperados de duas populações independentes com distribuições arbitrárias e variâncias desconhecidas.

- **Região de rejeição de H_0** (para valores da estatística de teste)

Estamos a lidar com um teste unilateral superior ($H_1 : \mu_A - \mu_B > \mu_0$), logo a região de rejeição de H_0 (para valores da estatística de teste) é um intervalo do tipo $W = (c, +\infty)$, onde $c : P(\text{Rejeitar } H_0 | H_0) = \alpha_0$, i.e.,

$$\begin{aligned} c &= \Phi^{-1}(1 - \alpha_0) \\ &= \Phi^{-1}(0.98) \\ &\stackrel{\text{tabela}}{=} 2.0537. \end{aligned}$$

- **Decisão**

Atendendo a que $n_A = 80$, $\bar{x}_A = 6.5$, $s_A^2 = 1.7^2$, $n_B = 100$, $\bar{x}_B = 5.5$, $s_B^2 = 2.5^2$, o valor observado da estatística de teste é igual a

$$\begin{aligned} t &= \frac{(\bar{x}_A - \bar{x}_B) - \mu_0}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \\ &= \frac{(6.5 - 5.5) - 0}{\sqrt{\frac{1.7^2}{80} + \frac{2.5^2}{100}}} \\ &\simeq 3.184. \end{aligned}$$

Como $t = 3.184245 \in W = (2.0537, +\infty)$, devemos rejeitar H_0 a qualquer n.s. $\alpha \geq 2\%$, [pelo que pode afirmar-se que há evidência para considerar $\mu_A > \mu_B$ a tais n.s.]

Grupo IV	5 valores
-----------------	-----------

1. A tabela seguinte apresenta o registo do número anual de acidentes de trabalho numa empresa de construção civil nos últimos 40 anos:

Nº de acidentes	0	1	2	3 ou mais
Frequência	13	17	6	4

Teste a hipótese de o número anual de acidentes seguir uma distribuição de Poisson com valor esperado unitário. Decida com base no valor- p . (2.0)

- **V.a. de interesse**

X = número anual de acidentes

- **Hipóteses**

$$H_0 : X \sim \text{Poisson}(1)$$

$$H_1 : X \not\sim \text{Poisson}(1)$$

- **Estatística de Teste**

$$T = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \underset{a}{\sim}_{H_0} \chi_{(k-\beta-1)}^2,$$

onde:

k = No. de classes;

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.

- **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 um intervalo à direita $W = (c, +\infty)$.

Frequências absolutas esperadas sob H_0

Para já, note-se que o conjunto de valores possíveis da distribuição Poisson(1) é $\{0, 1, 2, 3, \dots\}$.

Assim, as classes a considerar são $\{0\}$, $\{1\}$, $\{2\}$ e $\{3, 4, \dots\}$ e as frequências absolutas esperadas sob H_0 são iguais a

$$\begin{aligned} E_1 &= n \times p_1^0 \\ &= n \times P[X = 0 \mid X \sim \text{Poisson}(1)] \\ &= 40 \times e^{-1} \\ &= 40 \times F_{\text{Poi}(1)}(0) \\ &\stackrel{\text{tabela}}{=} 40 \times 0.3679 \\ &= 14.716 \\ E_2 &= n \times p_2^0 \\ &= n \times P[X = 1 \mid X \sim \text{Poisson}(1)] \\ &= 40 \times e^{-1} \\ &= 40 \times (F_{\text{Poi}(1)}(1) - F_{\text{Poi}(1)}(0)) \\ &\stackrel{\text{tabela}}{=} 40 \times (0.7358 - 0.3679) \\ &= 40 \times 0.3679 \\ &= 14.716 \\ E_3 &= n \times p_3^0 \\ &= n \times P[X = 2 \mid X \sim \text{Poisson}(1)] \\ &= 40 \times e^{-1} \frac{1}{2!} \\ &= 40 \times (F_{\text{Poi}(1)}(2) - F_{\text{Poi}(1)}(1)) \\ &\stackrel{\text{tabela}}{=} 40 \times (0.9197 - 0.7358) \\ &= 40 \times 0.1839 \\ &= 7.356 \end{aligned}$$

$$\begin{aligned} E_4 &= n \times p_4^0 \\ &= n \times P[X \geq 3 \mid X \sim \text{Poisson}(1)] \\ &= n \times (1 - p_1^0 - p_2^0 - p_3^0) \\ &= 40 \times (1 - 0.3679 - 0.3679 - 0.1839) \\ &= 40 \times 0.0803 \\ &= 3.212. \end{aligned}$$

É sabido que não é necessário qualquer agrupamento de classes se em pelo menos 80% das classes se verifica $E_i \geq 5$ e em todas elas se tem $E_i \geq 1$. Ora, somente as 3 primeiras das 4 classes (i.e., 75%) possuem $E_i \geq 5$, pelo que é preciso agrupar as duas últimas classes.

- **Decisão**

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

i	Classe i	Freq. abs. obs. o_i	Freq. abs. esper. sob H_0 $E_i = n \times p_i^0$	Parcelas valor obs. estat. teste $\frac{(o_i - E_i)^2}{E_i}$
1	{0}	13	14.716	$\frac{(13-14.716)^2}{14.716} = 0.200099$
2	{1}	17	14.716	0.354489
3	$\{2\} \cup \{3, \dots\}$	$6 + 4 = 10$	$7.356 + 3.212 = 10.568$	0.030528
		$\sum_{i=1}^k o_i = n = 40$	$\sum_{i=1}^k E_i = n = 40$	$t = \sum_{i=1}^k \frac{(o_i - E_i)^2}{E_i} = 0.585116$

- **Decisão (com base em intervalo para o valor-p)**

Uma vez que este teste está associado a uma região de rejeição que é um intervalo à direita temos:

$$\begin{aligned}
 \text{valor} - p &= P(T > t \mid H_0) \\
 &= P(T > 0.585116 \mid H_0) \\
 &\simeq 1 - F_{\chi^2_{(3-0-1)}}(0.585116).
 \end{aligned}$$

Recorrendo às tabelas de quantis da distribuição do qui-quadrado podemos adiantar um intervalo para o *valor-p* deste teste. Com efeito, ao enquadrarmos convenientemente $t = 0.585116$, obtemos

$$\begin{aligned}
 F_{\chi^2_{(2)}}^{-1}(0.20) = 0.446 &< 0.585116 < 0.713 = F_{\chi^2_{(2)}}^{-1}(0.30) \\
 0.20 &< F_{\chi^2_{(2)}}(3.2020) < 0.30 \\
 0.70 = 1 - 0.30 &< \text{valor} - p < 1 - 0.20 = 0.80.
 \end{aligned}$$

Logo:

- não devemos rejeitar H_0 a qualquer n.s. $\alpha_0 \leq 70\%$, por exemplo, a qualquer dos níveis usuais de significância de 1%, 5% e 10%;
- devemos rejeitar H_0 a qualquer n.s. $\alpha_0 \geq 80\%$.

- **Alternativa — Decisão (com base no valor-p determinado usando máquina de calcular)**

Dado que este teste está associado a uma região de rejeição que é um intervalo à direita temos:

$$\begin{aligned}
 p - \text{value} &= P(T > t \mid H_0) \\
 &= P(T > 0.585116 \mid H_0) \\
 &\simeq 1 - F_{\chi^2_{(3-0-1)}}(0.585116) \\
 &= 0.746352.
 \end{aligned}$$

Consequentemente:

- não devemos rejeitar H_0 a qualquer n.s. $\alpha_0 \leq 74.6352\%$, por exemplo, a qualquer dos níveis usuais de significância de 1%, 5% e 10%;
- devemos rejeitar H_0 a qualquer n.s. $\alpha_0 > 74.6352\%$.

2. Com o objectivo de construir um modelo que relacione a taxa anual de mortalidade por aterosclerose por 100 000 habitantes em determinado país, Y , com o consumo anual de gordura per capita, x (em Kg), no mesmo país, foram analisados os valores destas variáveis num conjunto de $n = 10$ anos, conduzindo aos seguintes resultados:

$$\bar{x} = 7.82, \quad \sum_{i=1}^{10} x_i^2 - 10 \bar{x}^2 = 5.016, \quad \bar{y} = 2.93, \quad \sum_{i=1}^{10} y_i^2 - 10 \bar{y}^2 = 1.861, \quad \sum_{i=1}^{10} x_i y_i - 10 \bar{x} \bar{y} = 2.504$$

- (a) Indicando as hipóteses de trabalho convenientes, averigüe se os dados permitem concluir a significância do modelo de regressão linear simples de Y em x , considerando um nível de significância de 1%. (2.0)

- [Modelo de RLS

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Y_i = taxa de mortalidade por aterosclerose por 100 000 habitantes no i – ésimo ano

x_i = consumo anual de gordura per capita no i – ésimo ano

ϵ_i = erro aleatório associado à medição de tal taxa no i – ésimo ano]

- Hipóteses de trabalho

$\epsilon_i \sim_{i.i.d.} \text{Normal}(0, \sigma^2)$, $i = 1, \dots, n$ (hipótese de trabalho)

$\beta_0, \beta_1, \sigma^2$ DESCONHECIDOS

- Hipóteses

$$H_0 : \beta_1 = \beta_{1,0} = 0 \quad (\text{ou } \beta_1 = 0)$$

$$H_1 : \beta_1 \neq \beta_{1,0} = 0 \quad (\text{ou } \beta_1 \neq 0)$$

- 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 da estatística de teste)

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

$$c : P(\text{Rejeitar } H_0 | H_0) = \alpha_0$$

$$c = F_{t_{(n-2)}}^{-1}(1 - \alpha_0/2)$$

$$c = F_{t_{(10-2)}}^{-1}(0.995)$$

$$c \stackrel{\text{tabela}}{=} 3.355.$$

- Decisão

Tendo em conta que as estimativas de β_1 e σ^2 são iguais a

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \\ &= \frac{2.504}{5.016} \\ &= 0.499203 \end{aligned}$$

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-2} \left[\left(\sum_{i=1}^n y_i^2 - n \bar{y}^2 \right) - (\hat{\beta}_1)^2 \left(\sum_{i=1}^n x_i^2 - n \bar{x}^2 \right) \right] \\ &= \frac{1}{10-2} (1.861 - 0.499203^2 \times 5.016) \\ &\simeq 0.076374 \end{aligned}$$

(respectivamente), o valor observado da estatística de teste é dado por

$$\begin{aligned} t &= \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}}} \\ &= \frac{0.4992 - 0}{\sqrt{\frac{0.076374}{5.016}}} \\ &= 4.045602. \end{aligned}$$

Como $t = 4.045602 \in W = (-\infty, -3.355) \cup (3.355, +\infty)$, devemos rejeitar $H_0 : \beta_1 = \beta_{1,0} = 0$ a favor de $H_1 : \beta_1 \neq \beta_{1,0}$ a qualquer n.s. $\alpha'_0 \geq \alpha_0 = 1\%$.

Podemos assim concluir que o modelo de regressão considerado é significativo ao nível de significância de 1% [ou a qualquer n.s. superior a 1%].

(b) Calcule o coeficiente de determinação associado ao modelo referido. Comente o valor obtido, (1.0)

nomeadamente relacionando-o com a conclusão da alínea anterior.

Tirando partido do facto de $\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}$, assim como dos valores obtidos anteriormente, segue-se

- **Cálculo do coeficiente de determinação**

$$\begin{aligned} r^2 &= \frac{(\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y})^2}{(\sum_{i=1}^n x_i^2 - n \bar{x}^2) \times (\sum_{i=1}^n y_i^2 - n \bar{y}^2)} \\ &= \frac{2.504^2}{5.016 \times 1.861} \\ &\simeq 0.671684. \end{aligned}$$

- **Comentário ao coeficiente de determinação**

Cerca de 67.2% da variação total da taxa anual de mortalidade por aterosclerose é explicada pelo consumo anual de gordura per capita, através do modelo de regressão considerado.

Ao termos em conta este resultado bem como o da alínea anterior, podemos adiantar que há um ajustamento razoável da recta estimada ao nosso conjunto de dados.