Probabilidades e Estatística

TODOS OS CURSOS

2º semestre – 2018/2019
05/07/2019 – **15:00**

Duração: 90 minutos **2º Teste C**

Justifique convenientemente todas as respostas

Grupo I	10 valores

1. O número de pessoas inquiridas até se encontrar o segundo indivíduo que tenha visto o último episódio da série GoT é uma variável aleatória X com função de probabilidade

$$P(X=x)=\begin{cases}(x-1)\,(1-p)^{x-2}\,p^2, & x=2,3,\ldots\\ 0, & \text{outros valores de } x,\end{cases}$$

onde p representa a probabilidade desconhecida de um indivíduo selecionado ao acaso ter visto tal episódio. Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de X.

(a) Deduza o estimador de máxima verosimilhança do parâmetro p, com base na amostra aleatória referida acima. (3.0)

- **V.a. de interesse**
  X = no. de pessoas inquiridas até se encontrar o segundo indivíduo que...
- **F.p. de** X
  $P(X=x)=(x-1)\,(1-p)^{x-2}\,p^2, \quad x=2,3,\ldots$
- **Parâmetro desconhecido**
  $p, \quad 0 \le p \le 1$
- **Amostra**
  $\underline{x}=(x_1,\ldots,x_n)$ amostra de dimensão n proveniente da população X
- **Obtenção do estimador de MV de** p

  Passo 1 — Função de verosimilhança

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

  Passo 2 — Função de log-verosimilhança

$$\ln L(p\mid\underline{x})=\sum_{i=1}^{n}\ln(x_i-1)+\ln(1-p)\sum_{i=1}^{n}(x_i-2)+2n\ln(p), \quad 0<p<1$$

  Passo 3 — Maximização
  A estimativa de MV de p é doravante representada por $\hat{p}$ e

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

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

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

Passo 4 — Estimador de MV de p

$$EMV(p) = \frac{2}{\bar{X}}.$$

(b) Determine a estimativa de máxima verosimilhança de $P(X = 2)$ com base na amostra $(x_1, x_2, \ldots, x_{100})$ tal que $\sum_{i=1}^{100} x_i = 1599$. (1.5)

- **Estimativa de MV de p**

$$\begin{aligned} \hat{p} &= \frac{2}{\bar{x}} \\ &= \frac{2}{\frac{1599}{100}} \\ &\simeq 0.125078 \end{aligned}$$

- **Outro parâmetro desconhecido**
$h(p) = P(X = 2) = p^2$

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

$$\begin{aligned} \widehat{h(p)} &= h(\hat{p}) \\ &= \hat{p}^2 \\ &\simeq 0.125078^2 \\ &\simeq 0.015645. \end{aligned}$$

2. Admita que o tempo (em segundo) entre emissões consecutivas de partículas α por uma fonte radioativa é uma variável aleatória X com distribuição exponencial com valor esperado μ desconhecido. Suponha que uma amostra casual de X com dimensão $n = 35$ conduziu a uma média de tempos entre emissões consecutivas igual a 0.509.

(a) Obtenha um intervalo de confiança a aproximadamente 90% para μ. Considere a variável aleatória fulcral $Z = \sqrt{35}\left(\frac{\bar{X}}{\mu} - 1\right)$, cuja distribuição é aproximadamente normal$(0,1)$. (2.5)

- **V.a. de interesse**
X = tempo (em segundos) entre emissões consecutivas de partículas α

- **Situação**
$X \sim$ Exponencial$(1/\mu)$
$E(X) = \sqrt{V(X)} = \mu > 0$ DESCONHECIDO
$n = 35 > 30$ (suficientemente grande).

- **Obtenção de IC aproximado para $E(X) = \mu$**
Passo 1 — Selecção da v.a. fulcral para μ

$$Z = \sqrt{35}\left(\frac{\bar{X}}{\mu} - 1\right) \overset{a}{\sim} \text{normal}(0,1)$$

[Uma vez que nos foi solicitada a determinação de um IC aproximado para o valor esperado de X e a dimensão da amostra é suficientemente grande para invocar o TLC, faremos uso da seguinte v.a. fulcral para μ:

$$Z = \frac{\bar{X}-E(\bar{X})}{\sqrt{V(\bar{X})}} = \frac{\bar{X}-E(X)}{\sqrt{\frac{V(X)}{n}}} = \frac{\bar{X}-\mu}{\sqrt{\frac{\mu^2}{n}}} = \sqrt{n}\left(\frac{\bar{X}}{\mu} - 1\right) \overset{a}{\sim} \text{normal}(0,1).]$$

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

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

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

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

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

$$P\left[a_\alpha \leq \sqrt{35}\left(\frac{\bar{X}}{\mu}-1\right) \leq b_\alpha\right] \simeq 1-\alpha$$

$$P\left(1+\frac{a_\alpha}{\sqrt{35}} \leq \frac{\bar{X}}{\mu} \leq 1+\frac{b_\alpha}{\sqrt{35}}\right) \simeq 1-\alpha$$

$$P\left(\frac{\bar{X}}{1+\frac{b_\alpha}{\sqrt{35}}} \leq \mu \leq \frac{\bar{X}}{1+\frac{a_\alpha}{\sqrt{35}}}\right) \simeq 1-\alpha.$$

Passo 4 — Concretização
Ao termos em conta que

- $n = 35$
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 0.509$
- $\Phi^{-1}(1-\alpha/2) = 1.6449$,

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

$$\begin{aligned} \left[\frac{\bar{x}}{1+\frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{35}}}, \quad \frac{\bar{x}}{1-\frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{35}}}\right] &= \left[\frac{0.509}{1+\frac{1.6449}{\sqrt{35}}}, \quad \frac{0.509}{1-\frac{1.6449}{\sqrt{35}}}\right] \\ &\simeq [0.398266, 0.705024]. \end{aligned}$$

(b) Teste as hipóteses $H_0: \mu = 0.5$ e $H_1: \mu > 0.5$. Decida com base no valor-p. (3.0)

- **Hipóteses**
  $H_0: \mu = \mu_0 = 0.5$
  $H_1: \mu > \mu_0 = 0.5$

- **Estatística de teste**
  [Pode tirar-se partido da v.a. fulcral utilizada em (a) para obter a seguinte estatística de teste:]

$$T = \sqrt{35}\left(\frac{\bar{X}}{\mu_0}-1\right) \stackrel{a}{\sim}_{H_0} \text{normal}(0,1).$$

- **Região de rejeição de H_0** (para valores de T)
  Tratando-se de um teste unilateral superior ($H_1: \mu = E(X) > \mu_0$) e havendo tendência para os valores tomados por T crescerem à medida que $\bar{X}$ aumenta, a região de rejeição de H_0, escrita para valores da estatística de teste, é do tipo $W = (c, +\infty)$.

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

$$\begin{aligned} t &= \sqrt{35}\left(\frac{\bar{x}}{\mu_0}-1\right) \\ &= \sqrt{35}\left(\frac{0.509}{0.5}-1\right) \\ &\simeq 0.11. \end{aligned}$$

Uma vez que a região de rejeição deste teste é um intervalo à direita, temos:

$$valor-p = P(T > t \mid H_0)$$

$$
\begin{aligned}
valor-p &\simeq 1-\Phi(t) \\
&\simeq 1-\Phi(0.11) \\
&\overset{calc/tabela}{=} 1-0.5438 \\
&= 0.4562.
\end{aligned}
$$

Logo é suposto:

- não rejeitar H_0 a qualquer n.s. $\alpha_0 \leq 45.62\%$, nomeadamente a qualquer dos n.u.s. (1%, 5% e 10%);
- rejeitar H_0 a qualquer n.s. $\alpha_0 > 45.62\%$.

Grupo II 10 valores

1. Um engenheiro biomédico defende a hipótese H_0 de que a variável aleatória X, que representa o tempo de vida (em meses) de um paciente com certo tipo de leucemia, possui função de distribuição dada por $F_0(x) = 1 - e^{-\left(\frac{x}{14.9}\right)^2}$, para $x \geq 0$.

A observação de 200 destes pacientes conduziu aos resultados sumariados na tabela abaixo.

Classe	$[0, 7.4]$	$]7.4, 14.9]$	$]14.9, 22.3]$	$]22.3, 29.7]$	$]29.7, +\infty[$
Frequência absoluta observada	44	85	47	20	4
Frequência absoluta esperada sob H_0	E_1	82.71	52.28	17.53	E_5

(a) Obtenha os valores das frequências E_1 e E_5 (aproximando-as às centésimas). (1.0)

- **V.a. de interesse**
  X = tempo de vida (em meses) de paciente com certo tipo de leucemia
- **F.d. conjecturada**
  $F_0(x) = 1 - e^{-\left(\frac{x}{14.9}\right)^2}$, $x \geq 0$
- **Frequências absolutas esperadas omissas**
  Atendendo à dimensão da amostra $n = 200$ e à f.d. conjecturada, temos:

$$
\begin{aligned}
E_1 &= n \times P(X \leq 7.4) \\
&= n \times F_0(7.4) \\
&= 200 \times \left(1 - e^{-\left(\frac{7.4}{14.9}\right)^2}\right) \\
&\simeq 200 \times 0.218590 \\
&\simeq 43.72;
\end{aligned}
$$

$$
\begin{aligned}
E_5 &= n \times P(X > 29.7) \\
&= n - \sum_{i=1}^{4} E_i \\
&\simeq 200 - (43.72 + 82.71 + 52.28 + 17.53) \\
&= 3.76.
\end{aligned}
$$

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

- **Hipóteses**
  $H_0 : F_X(x) = F_0(x),\ x \in \mathbb{R}$
  $H_1 : \neg H_0$
- **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.

- **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: $E_1 \simeq 43.72$; $E_2 \simeq 82.71$; $E_3 \simeq 52.28$; $E_4 \simeq 17.53$; $E_5 \simeq 3.76$.
  [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)
  Lidamos com um teste de ajustamento, logo 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**
  [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. esp. sob H_0 E_i	Parcelas valor obs. estat. teste $\frac{(o_i-E_i)^2}{E_i}$
1	$[0, 7.4]$	44	43.72	$\frac{(44-43.72)^2}{43.72} \simeq 0.002$
2	$]7.4, 14.9]$	85	82.71	$\frac{(85-82.71)^2}{82.71} \simeq 0.063$
3	$]14.9, 22.3]$	47	52.28	0.533
4	$]22.3, 29.7]$	20	17.53	0.348
5	$]29.7, +\infty[$	4	3.76	0.015
		$\sum_{i=1}^k o_i = n = 200$	$\sum_{i=1}^k E_i = n = 200$	$t = \sum_{i=1}^k \frac{(o_i-E_i)^2}{E_i} \simeq 0.961$

  Uma vez que $t \simeq 0.961 \not\in W = (9.488, +\infty)$, 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 à idade em que uma criança profere a primeira palavra (x, em meses) e a pontuação obtida por ela num teste de aptidão (Y):

$$\sum_{i=1}^{20} x_i = 292, \quad \sum_{i=1}^{20} x_i^2 = 5506, \quad \sum_{i=1}^{20} y_i = 1867, \quad \sum_{i=1}^{20} y_i^2 = 178155, \quad \sum_{i=1}^{20} x_i y_i = 25864,$$

onde $[\min_{i=1,\dots,20} x_i, \max_{i=1,\dots,20} x_i] = [7, 42]$.

(a) Considere um modelo de regressão linear simples de Y em x e estime a reta de regressão de mínimos quadrados. (1.5)

- **Estimativas de MQ de** β_0, β_1 **e da reta de regressão**
  Dado que $n = 20$

$\sum_{i=1}^n x_i = 292$

$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i = \frac{292}{20} = 14.6$

$\sum_{i=1}^n x_i^2 = 5506$

$\sum_{i=1}^n x_i^2 - n(\bar{x})^2 = 5506 - 20 \times 14.6^2 = 1242.8$

$\sum_{i=1}^{n} y_i = 1867$
$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1867}{20} = 93.35$
$\sum_{i=1}^{n} y_i^2 = 178155$
$\sum_{i=1}^{n} y_i^2 - n\,(\bar{y})^2 = 178155 - 20 \times 93.35^2 = 3870.55$

$\sum_{i=1}^{n} x_i y_i = 25864$
$\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y} = 25864 - 20 \times 14.6 \times 93.35 = -1394.2,$

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

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,(\bar{x})^2} \\
&= \frac{-1394.2}{1242.8} \\
&\simeq -1.121822 \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \times \bar{x} \\
&\simeq 93.35 - (-1.121822) \times 14.6 \\
&\simeq 109.728601.
\end{aligned}$$

Consequentemente, a reta de regressão estimada é

$$\hat{E}(Y \mid x) = \hat{\beta}_0 + \hat{\beta}_1 \times x = 109.728601 - 1.121822\,x,$$

para $x \in [\min_{i=1,\ldots,20} x_i, \max_{i=1,\ldots,20} x_i] = [7, 42]$.

(b) Após ter enunciado as hipóteses de trabalho que entender convenientes, obtenha um intervalo de confiança a 90% para o valor esperado do resultado do teste de aptidão efetuado a uma criança que tenha proferido a primeira palavra aos 15 meses. (3.5)

- **Hipóteses de trabalho**
  $\epsilon_i \overset{i.i.d.}{\sim} \text{Normal}(0, \sigma^2)$, $i = 1, \ldots, n$
- **Obtenção do IC para** $E(Y \mid x = x_0) = \beta_0 + \beta_1 x_0$, com $x_0 = 15$
  Passo 1 — V.a. fulcral para $E(Y \mid x = x_0) = \beta_0 + \beta_1 x_0$

$$Z = \frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{\sqrt{\hat{\sigma}^2 \times \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right]}} \sim t_{(n-2)}$$

  Passo 2 — Quantis de probabilidade
  Como $(1-\alpha) \times 100\% = 90\%$, temos $\alpha = 0.1$ e lidaremos com os quantis

$$\begin{cases} a_\alpha = F^{-1}_{t_{(n-2)}}(\alpha/2) = -F^{-1}_{t_{(20-2)}}(1 - 0.1/2) = -F^{-1}_{t_{(18)}}(0.95) \overset{tabela/calc.}{=} -1.734 \\ b_\alpha = F^{-1}_{t_{(n-2)}}(1 - 0.1/2) = F^{-1}_{t_{(18)}}(0.95) \overset{tabela/calc.}{=} 1.734. \end{cases}$$

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

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

$$P\left[a_\alpha \leq \frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{\sqrt{\hat{\sigma}^2 \times \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right]}} \leq b_\alpha\right] = 1 - \alpha$$

$$P\left[(\hat\beta_0+\hat\beta_1 x_0) - b_\alpha \times \sqrt{\hat\sigma^2 \times \left[\frac{1}{n}+\frac{(x_0-\bar x)^2}{\sum_{i=1}^n x_i^2 - n\,\bar x^2}\right]} \le \beta_0+\beta_1 x_0 \right.$$
$$\left. \le (\hat\beta_0+\hat\beta_1 x_0) - a_\alpha \times \sqrt{\hat\sigma^2 \times \left[\frac{1}{n}+\frac{(x_0-\bar x)^2}{\sum_{i=1}^n x_i^2 - n\,\bar x^2}\right]}\right] = 1-\alpha$$

Passo 4 — Concretização

Uma vez que a estimativa de σ^2 é igual a

$$\begin{aligned}
\hat\sigma^2 &= \frac{1}{n-2}\left[\left(\sum_{i=1}^n y_i^2 - n\,\bar y^2\right) - (\hat\beta_1)^2\left(\sum_{i=1}^n x_i^2 - n\,\bar x^2\right)\right] \\
&= \frac{1}{20-2}\left[3\,870.55 - (-1.121822)^2 \times 1\,242.8\right] \\
&\simeq 128.139186
\end{aligned}$$

e a expressão geral do IC pretendido é

$$\begin{aligned}
& IC_{(1-\alpha)\times 100\%}(\beta_0+\beta_1 x_0) \\
&= \left[(\hat\beta_0+\hat\beta_1 x_0) \pm F^{-1}_{t_{(n-2)}}(1-\alpha/2) \times \sqrt{\hat\sigma^2 \times \left[\frac{1}{n}+\frac{(x_0-\bar x)^2}{\sum_{i=1}^n x_i^2 - n\,\bar x^2}\right]}\right],
\end{aligned}$$

temos

$$\begin{aligned}
& IC_{90\%}(\beta_0+\beta_1 \times 15) \\
&\simeq \left[(109.728601 - 1.121822\times 15) \pm 1.734 \times \sqrt{128.139186 \times \left[\frac{1}{20}+\frac{(15-14.6)^2}{1\,242.8}\right]}\right] \\
&\simeq [92.901271 \pm 1.734 \times 2.534454] \\
&= [92.901271 \pm 4.394743] \\
&= [88.506528, 97.296014].
\end{aligned}$$

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

- **Cálculo do coeficiente de determinação**

  O coeficiente de determinação pedido é igual a

$$\begin{aligned}
r^2 &= \frac{\left(\sum_{i=1}^n x_i y_i - n\,\bar x\,\bar y\right)^2}{\left(\sum_{i=1}^n x_i^2 - n\,\bar x^2\right)\times\left(\sum_{i=1}^n y_i^2 - n\,\bar y^2\right)} \\
&\overset{(a)}{=} \frac{(-1394.2)^2}{1\,242.8 \times 3\,870.55} \\
&\simeq 0.404088.
\end{aligned}$$

- **Interpretação do coeficiente de determinação**

  Cerca de 40.4% da variação total da variável resposta Y é explicada pela variável x através do modelo de regressão linear simples ajustado, donde podemos afirmar que a recta estimada parece não se ajustar bem ao conjunto de dados.