Probabilidades e Estatística

TODOS OS CURSOS

1º semestre – 2017/2018
30/01/2018 – **15:00**

Duração: 90 minutos **2º Teste C**

Justifique convenientemente todas as respostas

Grupo I	10 valores

1. A variável aleatória X representa o número de acessos a um pequeno servidor e possui função de probabilidade

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

onde p é uma probabilidade desconhecida. Sejam $(X_1,\ldots,X_n)$ uma amostra aleatória de X e $\bar{X}=\frac{1}{n}\sum_{i=1}^n X_i$.

(a) Mostre que o estimador de máxima verosimilhança do parâmetro p, com base na amostra aleatória acima, é dado por $2/(\bar{X}+2)$. (2.5)

- **V.a. de interesse**
  X = número de acessos a um pequeno servidor
- **F.p. de** X
  $P(X=x)=(x+1)\,(1-p)^x\,p^2, \quad x=0,1,2,\ldots$
- **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** θ

Passo 1 — Função de verosimilhança

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

Passo 2 — Função de log-verosimilhança

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

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

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

$$\begin{cases} -\frac{\sum_{i=1}^n x_i}{1-\hat{p}}+\frac{2n}{\hat{p}}=0\\ -\frac{\sum_{i=1}^n x_i}{(1-\hat{p})^2}-\frac{2n}{\hat{p}^2}<0\end{cases}$$

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

Passo 4 — Estimador de MV de λ

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

(b) Obtenha a estimativa de máxima verosimilhança de $P(X = 4)$ baseada na concretização $(x_1, x_2, x_3, x_4, x_5) = (3, 9, 8, 18, 8)$. (1.5)

- **Estimativa de MV de p**

$$\begin{aligned} \hat{p} &= \frac{2}{\bar{x}+2} \\ &= \frac{2}{\frac{3+9+8+18+8}{5}+2} \\ &= \frac{10}{3+9+8+18+8+10} \\ &= \frac{5}{28} \\ &\simeq 0.178571 \end{aligned}$$

- **Outro parâmetro desconhecido**

$$\begin{aligned} h(p) &= P(X=4) \\ &= 5\,(1-p)^4\, p^2 \end{aligned}$$

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

$$\begin{aligned} \widehat{h(p)} &= h(\hat{p}) \\ &= 5\,(1-\hat{p})^4\,\hat{p}^2 \\ &\simeq 5 \times (1-0.178571)^4 \times 0.178571^2 \\ &\simeq 0.072589. \end{aligned}$$

(c) Sabendo que $E(X) = \frac{2}{p} - 2$, determine o enviesamento do estimador $\bar{X}$ na estimação de $\frac{1-p}{p}$ e averigúe se $\bar{X}$ é um estimador centrado $\frac{1-p}{p}$. (1.5)

- **Estimador de $\frac{1-p}{p}$**
  $T = \bar{X}$

- **Enviesamento do estimador**

$$\begin{aligned} E(\bar{X}) - \frac{1-p}{p} &= E(X) - \frac{1-p}{p} \\ &= \left(\frac{2}{p} - 2\right) - \frac{1-p}{p} \\ &= \frac{1-p}{p} \end{aligned}$$

- **Conclusão**
  Uma vez que
  - T se diz um estimador centrado de $\frac{1-p}{p}$ caso $E(T) - \frac{1-p}{p} = 0, \forall p$,
  - $E(\bar{X}) - \frac{1-p}{p} \neq 0, \forall p \in (0,1)$,

  podemos concluir que $\bar{X}$ é um estimador não centrado (i.e., enviesado) de $\frac{1-p}{p}$.

2. O custo de produção de um certo artigo (X) possui distribuição normal com valor esperado desconhecido e desvio padrão igual a 5 euro. Ao observarem-se os custos de produção de 10 unidades desse artigo, obteve-se a seguinte estatística: $\sum_{i=1}^{10} x_i = 429$.

(a) Teste a hipótese nula de o valor esperado do custo de produção ser igual a 42 euro contra a alternativa de ser superior a esse valor, ao nível de significância de 5%. (3.0)

- **V.a. de interesse**
  X = custo de produção de um certo artigo

- **Situação**
  $X \sim \text{Normal}(\mu, \sigma^2)$
  μ DESCONHECIDO
  $\sigma = 5$ conhecido

- **Hipóteses**
  $H_0 : \mu = \mu_0 = 42$
  $H_1 : \mu > \mu_0$

- **N.s.**
  $\alpha_0 = 5\%$

- **Estatística de teste**

$$T = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim_{H_0} \text{normal}(0,1)$$

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

- **Região de rejeição de H_0** (para valores da estatística de teste)
  Tratando-se de um teste unilateral superior ($H_1 : \mu > \mu_0$), a região de rejeição de H_0 (para valores da estatística de teste) é do tipo $W = (c, +\infty)$, onde $c : P(\text{Rejeitar } H_0 \mid H_0) = \alpha_0$, i.e.,

$$\begin{aligned} c &: P(T > c \mid H_0) = \alpha_0 \\ & c = \Phi^{-1}(1 - 0.05) \\ & c \overset{tabela/calc.}{=} 1.6449 \end{aligned}$$

- **Decisão**
  O valor observado da estatística de teste é igual a

$$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \\ &= \frac{\frac{429}{10} - 42}{\frac{5}{\sqrt{10}}} \\ &\simeq 0.569210. \end{aligned}$$

  Como $t \simeq 0.56921 \not\in W = (1.6449, +\infty)$, não devemos rejeitar H_0 ao n.s. $\alpha_0 = 5\%$ [ou a qualquer n.s. inferior a $\alpha_0 = 5\%$].

(b) Determine a probabilidade de o procedimento aplicado na alínea (a) conduzir à rejeição de H_0, caso o valor esperado do custo de produção seja igual a 43 euro. (1.5)

- **Prob. pedida**
  Importa notar que $\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \text{normal}(0,1)$. Assim,

$$\begin{aligned} P(\text{Rejeitar } H_0 \mid \mu = 43) &= P(T > c \mid \mu = 43) \\ &= P\left(\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > c \mid \mu = 43\right) \end{aligned}$$

$$\begin{aligned}
P(\text{Rejeitar } H_0 \mid \mu = 43) &= P\left(\frac{\bar{X}-\mu+\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}} > c \mid \mu = 43\right) \\
&= P\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} > c - \frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}} \mid \mu = 43\right) \\
&= 1-\Phi\left(c - \frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right) \\
&= 1-\Phi\left(1.6449 - \frac{43-42}{\frac{5}{\sqrt{10}}}\right) \\
&\simeq 1-\Phi(1.01) \\
&\overset{tabela/calc.}{=} 1-0.8438 \\
&= 0.1562.
\end{aligned}$$

Grupo II	10 valores

1. A contagem do número de fogos florestais de determinada dimensão, numa certa região e num período de 10 semanas, conduziu aos seguintes dados: (4.0)

Dia da semana	segunda	terça	quarta	quinta	sexta	sábado	domingo
Número de fogos florestais	130	150	160	170	180	190	140

Teste a hipótese de os fogos se distribuírem uniformemente pelos 7 dias da semana. Decida com base no valor-p.

- **V.a. de interesse**

  X = dia da semana em que ocorre fogo

  (1 = segunda, ..., 7 = domingo)

- **Hipóteses**

  $H_0 : X \sim$ uniforme discreta$(\{1,\ldots,7\})$

  $H_1 : X \not\sim$ uniforme discreta$(\{1,\ldots,7\})$

- **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 = 7

  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 [dado que em H_0 se conjectura uma distribuição específica.]

- **Frequências absolutas esperadas sob H_0**

  Atendendo à dimensão da amostra $n = 1120$ e ao facto de a f.p. conjecturada ser dada por $P(X = x \mid H_0) = \frac{1}{7}$, $x = 1,\ldots,7$, as frequências absolutas esperadas sob H_0 são, para $i = 1,\ldots,7$, iguais a:

  $$\begin{aligned}
  E_i &= n \times p_i^0 \\
  &= 1120 \times \frac{1}{7} \\
  &= 160.
  \end{aligned}$$

[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 escrita para valores de T é o intervalo à direita $W = (c, +\infty)$.

- **Decisão (com base no valor-p)**

  No cálculo do valor obs. da estat. 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	segunda	130	160	$\frac{(130-160)^2}{160} = 5.625$
2	terça	150	160	0.625
3	quarta	160	160	0
4	quinta	170	160	0.625
5	sexta	180	160	2.5
6	sábado	190	160	5.625
7	domingo	140	160	2.5
		$\sum_{i=1}^k o_i = n = 1120$	$\sum_{i=1}^k e_i = n = 1120$	$t = \sum_{i=1}^k \frac{(o_i-e_i)^2}{e_i} = 17.5$

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

$$\begin{aligned} valor-p &= P(T > t \mid H_0) \\ &\simeq 1 - F_{\chi^2_{(k-\beta-1)}}(t) \\ &= 1 - F_{\chi^2_{(6)}}(17.5) \\ &\overset{calc.}{=} 0.007611. \end{aligned}$$

Logo, é suposto:

- não rejeitar H_0 a qualquer n.s. $\alpha_0 \leq 0.7611\%$;
- rejeitar H_0 a qualquer n.s. $\alpha_0 > 0.7611\%$, nomeadamente a qualquer dos n.u.s. $(1\%, 5\%, 10\%)$.

[Em alternativa, poderíamos recorrer às tabelas de quantis da distribuição do qui-quadrado com 6 graus de liberdade e adiantar um intervalo para o *p-value*:

$$\begin{aligned} F^{-1}_{\chi^2_{(6)}}(0.99) = 16.81 &< t = 17.5 < 18.55 = F^{-1}_{\chi^2_{(6)}}(0.995) \\ 0.99 &< F_{\chi^2_{(6)}}(17.5) < 0.995 \\ 1-0.995 &< 1 - F_{\chi^2_{(6)}}(17.5) < 1-0.99 \\ 0.005 &< valor-p < 0.01. \end{aligned}$$

Assim, é suposto:

- não rejeitar H_0 a qualquer n.s. $\alpha_0 \leq 0.5\%$;
- rejeitar H_0 a qualquer n.s. $\alpha_0 \geq 1\%$, nomeadamente a qualquer dos n.u.s. $(1\%, 5\%, 10\%)$.

2. Um conjunto de dados relativos a 161 países forneceu os seguintes valores relativos ao número de mortes anuais nas estradas por 100000 habitantes (Y) e ao logaritmo (de base e) do número de veículos motorizados por 1000 habitantes (x):

$\sum_{i=1}^{161} x_i = 737.9$, $\sum_{i=1}^{161} x_i^2 = 3792.96$, $\sum_{i=1}^{161} y_i = 2681.4$, $\sum_{i=1}^{161} y_i^2 = 56983.64$, $\sum_{i=1}^{161} x_i y_i = 10787.66$

(a) Considere o modelo de regressão linear simples de Y em x e calcule a estimativa de mínimos (2.0) quadrados de $\beta_0 + \beta_1\, x$.

- **Estimativas de MQ de β_0, β_1 e $\beta_0 + \beta_1\, x$**
  Dado que

$n = 161$

$\sum_{i=1}^n x_i = 737.9$
$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i = \frac{737.9}{161} \simeq 4.583230$
$\sum_{i=1}^n x_i^2 = 3\,792.96$
$\sum_{i=1}^n x_i^2 - n\,(\bar{x})^2 = 3\,792.96 - 161 \times 4.583230^2 = 410.994720$

$\sum_{i=1}^n y_i = 2\,681.4$
$\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i = \frac{2\,681.4}{161} = 16.654658$
$\sum_{i=1}^n y_i^2 = 56\,983.64$
$\sum_{i=1}^n y_i^2 - n\,(\bar{y})^2 = 56\,983.64 - 161 \times 16.654658^2 = 12\,325.839006$

$\sum_{i=1}^n x_i\, y_i = 10\,787.66$
$\sum_{i=1}^n x_i\, y_i - n\,\bar{x}\,\bar{y} = 10\,787.66 - 161 \times 4.583230 \times 16.654658 = -1\,501.812422,$

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

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i\, y_i - n\,\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\,(\bar{x})^2} \\
&= \frac{-1\,501.812422}{410.994720} \\
&\simeq -3.654092 \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \times \bar{x} \\
&\simeq 16.654658 - (-3.654092) \times 4.583230 \\
&\simeq 33.402202 \\
\hat{\beta}_0 + \hat{\beta}_1\, x &\simeq 33.402202 - 3.654092 \times x.
\end{aligned}
$$

(b) Após ter enunciado as hipóteses de trabalho que entender convenientes, teste a significância do modelo de regressão ao nível de 5%. **Nota**: Pode vir a necessitar do quantil $F^{-1}_{t_{(159)}}(0.975) = 1.975$. (3.0)

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

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

- **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 da estatística de teste)
  Estamos a lidar com um teste bilateral ($H_1 : \beta_1 \neq 0$), pelo que a região de rejeição de H_0 é uma reunião de intervalos do tipo $W = (-\infty, -c) \cup (c, +\infty)$, onde $c : P(\text{Rejeitar } H_0 \mid H_0) = \alpha_0$, i.e.,

$$
\begin{aligned}
c &= F^{-1}_{t_{(n-2)}}(1-\alpha_0/2) \\
&= F^{-1}_{t_{(161-2)}}(1-0.05/2) \\
&= F^{-1}_{t_{(159)}}(0.975) \\
&\stackrel{calc.}{=} 1.975
\end{aligned}
$$

- **Decisão**
  Tendo em conta os valores obtidos em (a), bem como o de

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-2}\left[\left(\sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2\right) - \left(\hat{\beta}_1\right)^2 \left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2\right)\right] \\
&\simeq \frac{1}{161-2}\left(12\,325.839006 - (-3.654092)^2 \times 410.994720\right) \\
&\simeq 43.006779,
\end{aligned}
$$

  o valor observado da estatística de teste é igual a

$$
\begin{aligned}
t &= \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}}} \\
&= \frac{-3.654092 - 0}{\sqrt{\frac{43.006779}{410.994720}}} \\
&= -11.296116.
\end{aligned}
$$

  Como $t = -11.296116 \in W = (-\infty, -1.975) \cup (1.975, +\infty)$ devemos rejeitar H_0 ao n.s. de 5% [bem como a qualquer n.s. superior que 5%. Com efeito, concluímos que devemos rejeitar a hipótese de o número de mortes anuais nas estradas por 100000 habitantes (Y) não ser influenciado linearmente pelo logaritmo (de base 10) do número de veículos motorizados por 1000 habitantes (x).]

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

- **Cálculo do coeficiente de determinação**

$$
\begin{aligned}
r^2 &= \frac{\left(\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}\right)^2}{\left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2\right) \times \left(\sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2\right)} \\
&= \frac{(-1\,501.812422)^2}{410.994720 \times 12\,325.839006} \\
&\simeq 0.445224.
\end{aligned}
$$

- **Interpretação coeficiente de determinação**
  Cerca de 44.5% 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 não parece ajustar-se bem ao conjunto de dados.