DM DEPARTAMENTO DE MATEMÁTICA TÉCNICO LISBOA

Probabilidades e Estatística

LEAN, LEE, LEGI, LERC, LMAC, MEAer, MEAmbi, MEBiol, MEBiom, MEEC, MEFT, MEMat, MEQ

2º semestre – 2017/2018
14/06/2018 – **09:00**

Duração: 90 minutos

2º Teste A

Justifique convenientemente todas as respostas

Grupo I — 10 valores

1. Uma engenheira biológica admite que o número de certos parasitas por hospedeiro infetado é uma variável aleatória X com função de probabilidade dada por

$$P(X=x) = \frac{(1-\alpha)^x}{x}\,(-\ln\alpha)^{-1}, \quad \text{onde} \quad x = 1,2,\ldots$$

e α é um parâmetro desconhecido tal que $0<\alpha<1$. Seja $(x_1, x_2, \ldots, x_n)$ a concretização de uma amostra aleatória de X.

(a) Calcule a primeira derivada da função de log-verosimilhança e obtenha uma equação satisfeita pela estimativa de máxima verosimilhança de α, $\hat{\alpha}$. (3.0)

- **V.a. de interesse**
  X = número de parasitas por hospedeiro infetado
- **Função de probabilidade de X**
  $P(X=x) = \frac{(1-\alpha)^x}{x}\,(-\ln\alpha)^{-1}, \qquad x = 1,2,\ldots$
- **Parâmetro desconhecido**
  α, tal que $0<\alpha<1$
- **Amostra**
  $\underline{x} = (x_1,\ldots,x_n)$ amostra de dimensão n proveniente da população X
- **Função de verosimilhança**

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

Função de log-verosimilhança

$$\ln L(\alpha \mid \underline{x}) = \sum_{i=1}^{n} x_i \times \ln(1-\alpha) - n\ln(-\ln\alpha) - \sum_{i=1}^{n} \ln(x_i)$$

- **Equação satisfeita pela estimativa de MV de α**
  A estimativa de MV de α será representada doravante por $\hat{\alpha}$ e satisfaz

$$\left.\frac{d\ln L(\alpha \mid \underline{x})}{d\alpha}\right|_{\alpha=\hat{\alpha}} = 0 \quad \text{(ponto de estacionaridade).}$$

Consequentemente

$$-\frac{\sum_{i=1}^n x_i}{1-\hat{\alpha}} - n\frac{\left(-\frac{1}{\hat{\alpha}}\right)}{(-\ln\hat{\alpha})} = 0$$

$$\frac{n\bar{x}}{1-\hat{\alpha}} + \frac{n}{\hat{\alpha}\ln(\hat{\alpha})} = 0 \qquad [\Leftrightarrow \quad \frac{\hat{\alpha}-1}{\hat{\alpha}\ln(\hat{\alpha})} = \bar{x}].$$

(b) Uma amostra $(x_1, x_2, \ldots, x_{20})$ conduziu a $\hat{\alpha} \simeq 0.527804$. Calcule a estimativa de máxima verosimilhança de $E(X) = \frac{\alpha-1}{\alpha\ln(\alpha)}$. (1.5)

- **Parâmetro desconhecido**

$$\begin{aligned} h(\alpha) &= E(X) \\ &= \frac{\alpha-1}{\alpha\ln(\alpha)} \end{aligned}$$

- **Estimativa de MV de** $h(\alpha) = E(X)$
  De acordo com a propriedade de invariância dos estimadores de máxima verosimilhança, a estimativa de MV de $h(\alpha)$ é dada por:

$$\begin{aligned} \widehat{E(X)} &= \widehat{h(\alpha)} \\ &= h(\hat{\alpha}) \\ &= \frac{\hat{\alpha}-1}{\hat{\alpha}\ln(\hat{\alpha})} \qquad [= \bar{x} = 1.4]. \\ &\simeq \frac{0.527804-1}{0.527804\times\ln(0.527804)} \\ &\simeq 1.4. \end{aligned}$$

2. Considere que a variável aleatória X representa a quantidade de água (em litro) de cada descarga de um novo modelo de autoclismo e que uma concretização de uma amostra aleatória de dimensão 40 da referida população conduziu aos seguintes resultados $\sum_{i=1}^{40} x_i = 450$ e $\sum_{i=1}^{40} x_i^2 = 10\,300$.

(a) Deduza um intervalo de confiança aproximado a 90% para $E(X)$. (2.5)

- **V.a. de interesse**
  X = quantidade de água (em litro) de cada descarga de um novo modelo de autoclismo
- **Situação**
  X com distribuição arbitrária
  $\mu = E(X)$ DESCONHECIDO
  $\sigma^2 = V(X)$ desconhecido
  $n = 40 \geq 30$
- **Obtenção do IC aproximado para** $\mu = E(X)$
  Passo 1 — Selecção da v.a. fulcral para μ

$$Z = \frac{\bar{X}-\mu}{S/\sqrt{n}} \overset{a}{\sim} \text{normal}(0,1)$$

[pois pretende-se IC aproximado para o valor esperado de pop. com distribuição arbitrária com variância desconhecida e estamos a lidar com amostra suficientemente grande.]

Passo 2 — Obtenção dos quantis de probabilidade
[Observe que os quantis simétricos que se seguem dizem respeito à distribuição aproximada da v.a. fulcral para μ e enquadram-na com probabilidade aproximadamente igual a $(1-\alpha) = 0.95$:]

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

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

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

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

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

Passo 4 — Concretização
Tendo em conta que o IC aproximado a $(1-\alpha)\times 100\%$ para μ é dado por

$$IC(\mu) = \left[\bar{x} \pm \Phi^{-1}(1-\alpha/2) \times \frac{s}{\sqrt{n}}\right],$$

onde

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{450}{40} = 11.25 \\ s^2 &= \frac{1}{n-1}\left[\left(\sum_{i=1}^{n} x_i^2\right) - n(\bar{x})^2\right] \\ &= \frac{1}{40-1}(10300 - 40 \times 11.25^2) \\ &= 134.2949. \end{aligned}$$

Assim,

$$\begin{aligned} IC(\mu) &= \left[11.25 \pm 1.6449 \times \sqrt{\frac{134.2949}{40}}\right] \\ &\simeq [8.2360,\ 14.2640]. \end{aligned}$$

(b) Confronte as hipóteses $H_0 : E(X) = 9$ e $H_1 : E(X) > 9$, calculando para o efeito o valor-p. (3.0)

- **V.a. de interesse e situação**
  Ver alínea a).
- **Hipóteses**
  $H_0 : E(X) = \mu = \mu_0 = 9$
  $H_1 : E(X) = \mu > \mu_0 = 9$
- **Estatística de teste**

  $$T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \overset{a}{\sim}_{H_0} \text{normal}(0,1)$$

  [uma vez que pretendemos efectuar um teste para o valor esperado de população com distribuição arbitrária com variância desconhecida e dispomos de uma amostra suficientemente grande.]
- **Região de rejeição de H_0** (para valores de T)
  Estamos a lidar com um teste unilateral superior ($H_1 : \mu > \mu_0$), donde a região de rejeição de H_0 seja do tipo $W = (c, +\infty)$.
- **Decisão (com base no valor-p)**
  O valor observado da estatística de teste é igual a

  $$\begin{aligned} t &= \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \\ &\simeq \frac{11.25-9}{\sqrt{134.2949}/\sqrt{40}} \\ &\simeq 1.23 \end{aligned}$$

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

$$
\begin{aligned}
valor-p &= P(T > t \mid H_0) \\
&\simeq 1-\Phi(t) \\
&\simeq 1-\Phi(1.23) \\
&\overset{tabela/calc.}{=} 1-0.8907 \\
&= 0.1093.
\end{aligned}
$$

Logo é suposto:

- não rejeitar H_0 a qualquer nível de significância $\alpha_0 \leq 10.93\%$, nomeadamente a qualquer dos níveis usuais de significância(1%, 5% e 10%);
- rejeitar H_0 a qualquer n.s. $\alpha_0 > 10.93\%$.

Grupo II 10 valores

1. Seja X a variável aleatória que descreve o número semanal de ataques cibernéticos a um conjunto de servidores. Um engenheiro informático defende a hipótese H_0 de que X possui função de probabilidade $P(X=x) = (x+1)\,0.8^x\,0.2^2, \quad x = 0,1,2,\ldots$

A contagem do número semanal de ataques cibernéticos a este conjunto de servidores, num período de 200 semanas selecionadas casualmente, conduziu à seguinte tabela de frequências:

Número semanal de ataques cibernéticos	0	1	2	3	mais de 3
Frequência absoluta observada	7	17	8	20	148
Frequência absoluta esperada sob H_0	E_1	12.80	15.36	16.38	E_5

(a) Obtenha os valores de E_1 e E_5 (aproximando-os às centésimas). (1.0)

- **V.a. de interesse**
  X = número semanal de ataques cibernéticos a um conjunto de servidores
- **F.p. conjecturada**
  $P(X=x) = (x+1)\,0.8^x\,0.2^2, \qquad x = 0,1,2,\ldots$
- **Frequências absolutas esperadas omissas**
  Atendendo à dimensão da amostra $n = 200$ e à f.p. conjecturada temos:

$$
\begin{aligned}
E_1 &= n \times P(X=0) \\
&= 200 \times (0+1)\,0.8^0\,0.2^2 \\
&= 8.00;
\end{aligned}
$$

$$
\begin{aligned}
E_5 &= n \times P(X \geq 4) \\
&= n - \sum_{i=1}^{4} E_i \\
&\simeq 200 - (8.00 + 12.80 + 15.36 + 16.38) \\
&= 147.46.
\end{aligned}
$$

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

- **Hipóteses**
  $H_0 : P(X=x) = (x+1)\,0.8^x\,0.2^2, \quad x = 0,1,2,\ldots$
  $H_1 : P(X=x) \neq (x+1)\,0.8^x\,0.2^2, \quad$ para algum $x \in \{0,1,2,\ldots\}$
- **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 [dado que em H_0 se conjectura uma distribuição específica.]

- **Frequências absolutas esperadas sob H_0**
  De acordo com a tabela facultada e a alínea a), as frequências absolutas esperadas sob H_0 aproximados às centésimas são: $E_1 = 8.00$; $E_2 = 12.80$; $E_3 = 15.36$; $E_4 = 16.38$; $E_5 = 147.46$.
  [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

$$\begin{aligned} c &= F^{-1}_{\chi^2_{(k-\beta-1)}}(1-\alpha_0) \\ &= F^{-1}_{\chi^2_{(5-0-1)}}(1-0.05) \\ &\overset{tabela/calc.}{=} 9.488. \end{aligned}$$

- **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	8.00	$\frac{(7-8.00)^2}{8.00} = 0.125$
2	$\{1\}$	17	12.80	$\frac{(17-12.80)^2}{12.80} \simeq 1.378$
3	$\{2\}$	8	15.36	3.527
4	$\{3\}$	20	16.38	0.800
5	$\{4,5,\ldots\}$	148	147.46	0.002
		$\sum_{i=1}^k o_i = n = 200$	$\sum_{i=1}^k E_i = n = 200$	$t = \sum_{i=1}^k \frac{(o_i-E_i)^2}{E_i} \simeq 5.832$

  Uma vez que $t \simeq 5.832 \notin W = (9.488, +\infty)$, 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 relativas a um pequeno bairro residencial conduziu aos seguintes resultados respeitantes à temperatura média diária x (em grau Celsius) e ao consumo diário de eletricidade Y (em kWh):

$$\sum_{i=1}^{20} x_i = 633.4, \quad \sum_{i=1}^{20} x_i^2 = 20\,383.92, \quad \sum_{i=1}^{20} y_i = 7\,857.6, \quad \sum_{i=1}^{20} y_i^2 = 3\,126\,303.02, \quad \sum_{i=1}^{20} x_i y_i = 252\,003.88,$$

onde $\left[\min_{i=1,\ldots,20} x_i, \max_{i=1,\ldots,20} x_i\right] = [25.0, 40.5]$.

(a) Considere o modelo de regressão linear simples de Y em x e determine a estimativa de mínimos quadrados do valor esperado do consumo diário de eletricidade num dia com temperatura média igual a 29 graus Celsius. (2.0)

- **Estimativa de MQ de $E(Y \mid x) = \beta_0 + \beta_1 x$ com $x = 29$**
  Uma vez que

  $n = 20$

  $\sum_{i=1}^{n} x_i = 633.4$
  $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{633.4}{20} = 31.67$
  $\sum_{i=1}^{n} x_i^2 = 20383.92$
  $\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2 = 20\,383.92 - 20 \times 31.67^2 = 324.1420$

  $\sum_{i=1}^{n} y_i = 7\,857.6$
  $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{7857.6}{20} = 392.88$
  $\sum_{i=1}^{n} y_i^2 = 3\,126\,303.02$
  $\sum_{i=1}^{n} y_i^2 - n(\bar{y})^2 = 3\,126\,303.02 - 20 \times 392.88^2 = 39\,209.1320$

  $\sum_{i=1}^{n} x_i y_i = 252\,003.88$
  $\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y} = 252\,003.88 - 20 \times 31.67 \times 392.88 = 3\,153.6880,$

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

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2} \\
&= \frac{3\,153.6880}{324.1420} \\
&\simeq 9.729341 \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \times \bar{x} \\
&\simeq 392.88 - 9.729341 \times 31.67 \\
&\simeq 84.751771 \\
\hat{\beta}_0 + \hat{\beta}_1 x &\simeq 84.751771 + 9.729341 \times 29 \\
&\simeq 366.902660.
\end{aligned}$$

(b) Após ter enunciado as hipóteses de trabalho que entender convenientes, teste se há evidência de uma relação de natureza linear entre as variáveis x e Y, ao nível de significância de 5%. (3.0)

- **Hipóteses de trabalho**
  $\epsilon_i \overset{i.i.d.}{\sim} \text{Normal}(0, \sigma^2)$, $i = 1, \ldots, n$
- **Hipóteses**
  $H_0 : \beta_1 = \beta_{1,0} = 0$
  $H_1 : \beta_1 \neq 0$
- **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.,

$$
\begin{aligned}
c &= F^{-1}_{t_{(n-2)}}(1-\alpha_0/2) \\
&= F^{-1}_{t_{(20-2)}}(1-0.05/2) \\
&= F^{-1}_{t_{(18)}}(0.975) \\
&\overset{calc.}{=} 2.101.
\end{aligned}
$$

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

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-2}\left[\left(\sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2\right) - \left(\hat{\beta}_1\right)^2\left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2\right)\right] \\
&\simeq \frac{1}{20-2}\left(39\,209.1320 - 9.729341^2 \times 324.1420\right) \\
&\simeq 473.656975,
\end{aligned}
$$

  o valor observado da estatística de teste é igual a

$$
\begin{aligned}
t &= \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}}} \\
&= \frac{9.729341 - 0}{\sqrt{\frac{473.656975}{324.1420}}} \\
&= 8.048577.
\end{aligned}
$$

  Como $t = 8.048577 \in W = (-\infty, -2.101) \cup (2.101, +\infty)$ devemos rejeitar H_0 ao n.s. de 5% [bem como a qualquer n.s. superior que 5%. De facto podemos concluir que devemos rejeitar a hipótese de a variável aleatória Y não ser influenciada linearmente pela variável explicativa x.]

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

- **Cálculo do coeficiente de determinação**

$$
\begin{aligned}
r^2 &= \frac{\left(\sum_{i=1}^{n} x_i\,y_i - n\,\bar{x}\,\bar{y}\right)^2}{\left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2\right) \times \left(\sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2\right)} \\
&= \frac{3\,153.6880^2}{324.1420 \times 39\,209.1320} \\
&\simeq 0.782555.
\end{aligned}
$$

- **Interpretação coeficiente de determinação**
  Cerca de 78.26% 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.