DM
DEPARTAMENTO DE MATEMÁTICA
TÉCNICO LISBOA

Probabilidades e Estatística

LEGM, LEIC-A, LEIC-T, LMAC, MA, MEAer, MEBiol, MEBiom, MEFT

2º semestre – 2020/2021
18/06/2021 – **11:00**

Duração: 60+15 minutos **Teste 2B**

Justifique convenientemente todas as respostas

1. Admita que o número de viaturas que chega por hora a um centro *drive-thru* para testagem de covid-19 é uma variável aleatória X com distribuição de Poisson de parâmetro desconhecido λ ($\lambda > 0$). (4.0)

Deduza a estimativa de máxima verosimilhança do coeficiente de variação de X, $CV(X) = \frac{\sqrt{V(X)}}{E(X)}$, baseada na amostra $(17, 10, 11, 23, 18)$ proveniente da população X.

- **V.a. de interesse**

  $X =$ número de viaturas que chega por hora a um centro *drive-thru* para testagem de covid-19

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

- **F.p. de** X

  $P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$

- **Parâmetro desconhecido**

  $\lambda, \quad \lambda > 0$

- **Amostra**

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

- **Obtenção da estimativa de MV de** p

  Passo 1 — Função de verosimilhança

$$
\begin{aligned}
L(\lambda \mid \underline{x}) &= P(\underline{X} = \underline{x}) \\
&\overset{X_i \overset{i.i.d}{\sim} X}{=} \prod_{i=1}^{n} P(X = x_i) \\
&= \prod_{i=1}^{n} \left(e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}\right]) \\
&= e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}, \quad \lambda > 0
\end{aligned}
$$

  Passo 2 — Função de log-verosimilhança

$$
\ln L(\lambda \mid \underline{x}) = -n\lambda + \ln(\lambda)\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \ln(x_i!), \quad \lambda > 0
$$

  Passo 3 — Maximização

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

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

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

$$\hat{\lambda} \quad : \quad \begin{cases} \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \\ -\frac{n}{\bar{x}} < 0 \qquad \text{(prop. verdadeira já que } \bar{x} \geq 0 \text{ e } n \in \mathbb{N}) \end{cases}$$

Passo 4 — Estimativa de MV de p

$$\hat{\lambda} \quad = \quad \bar{x} = \frac{17+10+11+23+18}{5} = \frac{79}{5} = 15.8$$

- **Outro parâmetro desconhecido**

$$h(\lambda) \quad = \quad CV(X) = \frac{\sqrt{V(X)}}{E(X)} \overset{form.}{=} \frac{\sqrt{\lambda}}{\lambda} = \frac{1}{\sqrt{\lambda}}$$

- **Estimativa de MV de $h(p)$**

  Pela propriedade de invariância dos estimadores de MV, concluímos que a estimativa de MV de $h(\lambda)$ é dada por

$$\begin{aligned} \widehat{h(\lambda)} &= h(\hat{\lambda}) \\ &= \frac{1}{\sqrt{\hat{\lambda}}} \\ &= \frac{1}{\sqrt{15.8}} \\ &\simeq 0.251577. \end{aligned}$$

2. Uma das medidas de desempenho de um teste de diagnóstico é a sua sensibilidade, a probabilidade do teste resultar positivo quando aplicado a um indivíduo com a doença em estudo. (4.0)

Para avaliar a sensibilidade, p, de um teste de diagnóstico ao covid-19 foram testados, de forma independente, um conjunto de $n = 259$ pacientes infectados com covid-19. Destes testes de diagnóstico 143 deram positivo.

Determine um intervalo aproximado de confiança a 97% para a sensibilidade desconhecida do teste, p.

- **V.a. de interesse**

$$X = \begin{cases} 1, & \text{se teste deu positivo} \\ 0, & \text{caso contrário} \end{cases}$$

- **Situação**

  $X \sim \text{Bernoulli}(p)$

  p DESCONHECIDO

- **Obtenção do IC para p**

Passo 1 — Seleção da v.a. fulcral

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

Passo 2 — Obtenção dos quantis de probabilidade

Como $(1-\alpha)\times 100\% = 93\%$, lidaremos com os quantis seguintes:

$$\begin{cases} a_\alpha = \Phi^{-1}(\alpha/2) = -\Phi^{-1}(1-\alpha/2) = -\Phi^{-1}(1-0.03/2) = -\Phi^{-1}(0.985) \overset{tabelas,calc.}{=} -2.1701 \\ b_\alpha = \Phi^{-1}(1-\alpha/2) = \Phi^{-1}(0.985) = 2.1701. \end{cases}$$

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

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

$$P\left[a_\alpha \le \frac{\bar{X}-p}{\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}} \le b_\alpha\right] \simeq 1-\alpha$$

$$P\left[\bar{X} - b_\alpha \times \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \le p \le \bar{X} - a_\alpha \times \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right] \simeq 1-\alpha$$

Passo 4 — Concretização

A expressão geral do IC aproximado para p é

$$IC_{(1-\alpha)\times 100\%}(p) \quad \simeq \quad \left[\bar{x} - \Phi^{-1}(1-\alpha/2) \times \sqrt{\frac{\bar{x}(1-\bar{x})}{n}}\right].$$

Atendendo aos quantis acima e ao facto de $n = 259$ e $\sum_{i=1}^n x_i = 143$, o IC pretendido é

$$\begin{aligned} IC_{93\%}(p) \quad &\simeq \quad \left[\frac{143}{259} \pm 2.1701 \times \sqrt{\frac{\frac{143}{259}\left(1-\frac{143}{259}\right)}{259}}\right] \\ &= \quad [0.5521 \pm 2.1701 \times 0.0309] \\ &= \quad [0.4850, 0.6192]. \end{aligned}$$

3. Uma equipa de infectologistas suspeita que as vítimas de covid-19 apresentem uma perda percentual esperada de massa corporal igual a $\mu_0 = 9$, fruto do gasto metabólico para combater a doença. Para avaliar esta hipótese, a equipa registou a percentagem de perda de massa corporal, X, de cada indivíduo de um grupo de $n = 51$ vítimas de covid-19, escolhidas ao acaso, contabilizando-se um total de $\sum_{i=1}^n x_i = 460.8$ e um desvio padrão amostral de $s = 1.27$. (4.0)

Assumindo que X tem distribuição normal, confronte as hipóteses $H_0 : \mu = \mu_0$ e $H_1 : \mu \neq \mu_0$, calculando para o efeito o valor-p.

- **V.a. de interesse**

  X = percentagem de perda de massa corporal de paciente vítima de covid-19

- **Situação**

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

  $\mu = E(X)$ DESCONHECIDO

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

- **Hipóteses**

  $H_0 : \mu = \mu_0$

  $H_1 : \mu \neq \mu_0$

- **Estatística de teste**

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

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

  O teste é bilateral ($H_1 : \mu \neq \mu_0$), logo a região de rejeição de H_0 é do tipo $W = (-\infty, -c) \cup (c, +\infty)$.

- **Decisão (com base no valor-p)**

  Atendendo a que o valor observado da estatística de teste é igual a

$$t \quad = \quad \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{\frac{460.8}{51} - 9}{\frac{1.27}{\sqrt{51}}} \simeq 0.198465$$

  e

$$valor-p \quad = \quad P(|T| > |t| \mid H_0) = 2 \times \left[1 - F_{t_{(n-1)}}(|t|)\right] = 2 \times \left[1 - F_{t_{(50)}}(0.198465)\right)] \overset{calc.}{\simeq} 0.8435$$

  devemos

  – não rejeitar H_0 a qualquer n.s. $\alpha_0 \leq 84.35\%$, designadamente aos n.u.s. $(1\%, 5\%, 10\%)$;
  – rejeitar H_0 a qualquer n.s. $\alpha_0 > 84.35\%$.

  Alternativamente e recorrendo às tabelas de quantis da distribuição de t-student, podemos obter um intervalo para o valor-p deste teste:

$$\begin{aligned}
F^{-1}_{t_{(50)}}(0.5) = 0 \quad &< \quad |t| = 0.198465 < 0.255 = F^{-1}_{t_{(50)}}(0.6) \\
0.5 \quad &< \quad F_{t_{(50)}}(0.198465) < 0.6 \\
0.8 = 2 \times (1 - 0.6) \quad &< \quad valor-p < 1 = 2 \times (1 - 0.5).
\end{aligned}$$

  Assim, podemos adiantar somente que:

  – não devemos rejeitar H_0 a qualquer n.s. $\alpha_0 \leq 80\%$, designadamente aos n.u.s. $(1\%, 5\%, 10\%)$.

4. Um investigador conjeturou a hipótese H_0 de que a proporção de resultados positivos de um teste de diagnóstico aplicados a doentes com covid-19 é uma variável aleatória X com a seguinte função de distribuição: $F_0(x) = x^4$, para $0 \leq x \leq 1$. (4.0)

Em $n = 200$ grupos de pacientes com covid-19, escolhidos ao caso, registaram-se as correspondentes proporções de pacientes com teste positivo:

Proporção de testes positivos	$[0, 0.40]$	$]0.40, 0.55]$	$]0.55, 0.70]$	$]0.70, 0.85]$	$]0.85, 1.00]$
Frequência absoluta observada	8	18	38	52	84
Frequência esperada sob H_0	5.1	13.2	29.7	E_4	E_5

Após ter calculado as frequências absolutas esperadas sob H_0 omissas, E_4 e E_5 (aproximando-as às décimas), avalie a hipótese de X possuir função de distribuição definida acima. Decida com base no valor-p aproximado.

- **V.a. de interesse**

  X = proporção de testes positivos em cada grupo

- **Hipóteses**

  $H_0 : F_X(x) = F_0(x), \quad x \in \mathbb{R}$

  $H_1 : \neg H_0$

- **Estatística de teste**

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

  onde:

  - k = no. de classes = 5;
  - O_i = freq. abs. observável da classe i;
  - E_i = freq. abs. esperada sob H_0 da classe i;
  - $\beta = 0$.

- **Frequência absolutas esperadas sob H_0 omissas**

$$E_4 = n \times F_{X|H_0}(0.85) - F_{X|H_0}(0.70) = 200 \times \left(0.85^4 - 0.70^4\right) \simeq 56.4$$

$$E_5 = n - \sum_{i=1}^{4} E_i \simeq 200 - (5.1 + 13.2 + 29.7 + 56.4) = 95.6.$$

- **Região de rejeição de H_0** (para valores de T)

  Tratando-se de um teste de ajustamento do qui-quadrado, a região de rejeição de H_0 escrita para valores observados de T é o intervalo à direita $W = (c, +\infty)$.

- **Decisão (com base no valor-p)**

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	$[0, 0.40]$	8	5.1	$\frac{(8-5.1)^2}{5.1} = 1.649$
2	$]0.40, 0.55]$	18	13.2	1.745
3	$]0.55, 0.70]$	38	29.7	2.320
4	$]0.70, 0.85]$	52	56.4	0.343
5	$]0.85, 1.00]$	84	95.6	1.408
		$\sum_{i=1}^{k} o_i = n$ $= 200$	$\sum_{i=1}^{k} e_i = n$ $= 200$	$t = \sum_{i=1}^{k} \frac{(o_i-e_i)^2}{e_i}$ $= 7.465$

Dado que o valor observado da estatística de teste é $t = 7.465$ e $W = (c, +\infty)$, obtemos

$$\begin{aligned} valor-p &= P(T > t \mid H_0) \\ &\simeq 1 - F_{\chi^2_{(k-1)}}(t) \\ &\simeq 1 - F_{\chi^2_{(4)}}(7.465) \\ &\overset{calc.}{\simeq} 0.1133 \end{aligned}$$

e devemos:

- não rejeitar de H_0 a qualquer n.s. $\alpha_0 \le 11.33\%$, nomeadamente aos n.u.s. $(1\%, 5\%, 10\%)$;
- rejeitar de H_0 a qualquer n.s. $\alpha_0 > 11.33\%$.

Alternativamente e recorrendo às tabelas de quantis da distribuição do qui-quadrado, podemos obter um intervalo para o valor-p aproximado deste teste:

$$\begin{array}{rcl} F^{-1}_{\chi^2_{(4)}}(0.85) = 6.745 & < & t = 7.465 < 7.779 = F^{-1}_{\chi^2_{(4)}}(0.90) \\ 0.85 & < & F_{\chi^2_{(4)}}(7.465) < 0.90 \\ 0.10 = 1 - 0.9 & < & valor-p \simeq 1 - F_{\chi^2_{(4)}}(7.465) < 1 - 0.85 = 0.15. \end{array}$$

Assim, podemos adiantar que:

- não devemos rejeitar H_0 a qualquer n.s. $\alpha_0 \leq 10\%$, por exemplo aos n.u.s. $(1\%, 5\%, 10\%)$;
- devemos rejeitar H_0 a qualquer n.s. $\alpha_0 \geq 15\%$.

5. Uma amostra de $n = 7$ robalos foi capturada por uma equipa de biólogos que registou o comprimento x (4.0)
(em milímetros) e a massa Y (em gramas) de cada robalo capturado. Os dados recolhidos conduziram aos seguintes resultados respeitantes a x e a Y:

$$\sum_{i=1}^n x_i = 1811, \quad \sum_{i=1}^n x_i^2 = 489665, \quad \sum_{i=1}^n y_i = 1177, \quad \sum_{i=1}^n y_i^2 = 205687, \quad \sum_{i=1}^n x_i y_i = 317301,$$

onde $\left[\min_{i=1,\dots,n} x_i, \max_{i=1,\dots,n} x_i\right] = [169, 329]$.

Admita que as variáveis x e Y estão relacionadas de acordo com o modelo de regressão linear simples: $Y = \beta_0 + \beta_1 x + \varepsilon$.

Após ter enunciado as hipóteses de trabalho que entender convenientes, obtenha o intervalo de confiança a 99% para o valor esperado de Y quando $x = 329$ e a amplitude deste intervalo.

- **Modelo de RLS**

  $Y =$ massa (v.a. resposta)

  $x =$ comprimento (variável explicativa)

  $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n$

- **Hipóteses de trabalho**

  $\varepsilon_i \overset{i.i.d.}{\sim} \text{normal}(0, \sigma^2), \quad i = 1, \dots, n$

- **Estimativas de MV de β_0 e β_1; estimativa de σ^2**
  - $n = 7$
  - $\sum_{i=1}^n x_i = 1811$

    $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1811}{7} \simeq 258.714286$

    $\sum_{i=1}^n x_i^2 = 489665$

    $\sum_{i=1}^n x_i^2 - n\,\bar{x}^2 \simeq 489665 - 7 \times 258.714286^2 \simeq 21133.428571$
  - $\sum_{i=1}^n y_i = 1177$

    $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i = \frac{1177}{7} \simeq 168.142857$

    $\sum_{i=1}^n y_i^2 = 205687$

    $\sum_{i=1}^n y_i^2 - n\,\bar{y}^2 \simeq 205687 - 7 \times 168.142857^2 \simeq 7782.857143$
  - $\sum_{i=1}^n x_i y_i = 317301$

    $\sum_{i=1}^n x_i y_i - n\,\bar{x}\,\bar{y} \simeq 317301 - 7 \times 258.714286 \times 168.142857 \simeq 12794.285637.$

Deste modo

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i\,y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^n x_i^2 - n\,\bar{x}^2} \simeq \frac{12794.285637}{21133.428571} \simeq 0.605405 \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\,\bar{x} \simeq 168.142857 - 0.605405 \times 258.714286 \simeq 11.515935 \\
\hat{\sigma}^2 &= \frac{1}{n-2}\left[\left(\sum_{i=1}^n y_i^2 - n\,\bar{y}^2\right) - \left(\hat{\beta}_1\right)^2\left(\sum_{i=1}^n x_i^2 - n\,\bar{x}^2\right)\right] \\
&\simeq \frac{1}{7-2}\left(7782.857143 - 0.605405^2 \times 21133.428571\right) \\
&\simeq 7.426809
\end{aligned}$$

- **Obtenção do IC para** $E(Y \mid x = x_0) = \beta_0 + \beta_1\,x_0$

Passo 1 — Seleção da v.a. fulcral

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

Passo 2 — Obtenção dos quantis de probabilidade

$$\begin{cases}
a_\alpha = F_{t_{(n-2)}}(\alpha/2) = -F_{t_{(7-2)}}(1-0.01/2) = -F_{t_{(5)}}(0.995) \stackrel{tabelas,calc.}{=} -4.032 \\
b_\alpha = F_{t_{(n-2)}}(1-\alpha/2) = F_{t_{(5)}}(0.995) \stackrel{tabelas,calc.}{=} 4.032
\end{cases}$$

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

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

$$P\left\{a_\alpha \leq \frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{S} \leq b_\alpha\right\} = 1-\alpha$$

$$P\left\{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - F^{-1}_{t_{(n-2)}}(1-\alpha/2) \times S \leq \beta_0 + \beta_1 x_0 \leq (\hat{\beta}_0 + \hat{\beta}_1 x_0) - F^{-1}_{t_{(n-2)}}(1-\alpha/2) \times S\right\} = 1-\alpha$$

Passo 4 — Concretização

Tendo em conta a expressão geral do IC para $\beta_0 + \beta 1\,x_0$,

$$IC_{(1-\alpha)\times 100\%}(\beta_0 + \beta_1 x_0) = \left[(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm F^{-1}_{t_{(n-2)}}(1-\alpha/2) \times \sqrt{\hat{\sigma}^2 \times \left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n x_i^2 - n\,\bar{x}^2}\right]}\right],$$

o IC pretendido e a sua amplitude são iguais a

$$\begin{aligned}
IC_{99\%}(\beta_0 + \beta_1 x_0) &\simeq \left[(11.515935 + 0.605405 \times 329) \pm 4.032 \times \sqrt{7.426809 \times \left(\frac{1}{7} + \frac{(329-258.714286)^2}{21133.428571}\right)}\right] \\
&\simeq [210.694180 \pm 4.032 \times 1.672435] \\
&\simeq [203.950682, 217.437678]
\end{aligned}$$

$$217.437678 - 203.950682 = 13.486996.$$