# 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 Poisson de parâmetro desconhecido $\lambda$ ($\lambda > 0$). Deduza a estimativa de máxima verosimilhança do coeficiente de variação de $X$, $CV(X) = \frac{\sqrt{V(X)}}{E(X)}$, baseada na amostra $(x_1, \ldots, x_5)$ proveniente da população $X$. (4.0)

- **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 $\lambda$ é doravante representada por $\hat{\lambda}$ e

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

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

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

**Passo 4 — Estimativa de MV de $p$**

$$\begin{aligned} \hat{\lambda} &= \bar{x} \\ &= \frac{x_1 + \ldots x_5}{5} \end{aligned}$$

- **Outro parâmetro desconhecido**

$$\begin{aligned} h(\lambda) &= CV(X) \\ &= \frac{\sqrt{V(X)}}{E(X)} \\ &\overset{form.}{=} \frac{\sqrt{\lambda}}{\lambda} \\ &= \frac{1}{\sqrt{\lambda}} \end{aligned}$$

- **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

$$\begin{aligned} \widehat{h(\lambda)} &= h(\hat{\lambda}) \\ &= \frac{1}{\sqrt{\hat{\lambda}}} \\ &= \frac{1}{\sqrt{\frac{x_1 + \ldots x_5}{5}}} \end{aligned}$$

| **Pergunta 2** | 4 valores |
|---|---|

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.

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

Determine um intervalo aproximado de confiança a $(1-\alpha)\times 100\%$ para 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**

  $$\begin{cases} a_\alpha = \Phi^{-1}(\alpha/2) = -\Phi^{-1}(1-\alpha/2) \\ b_\alpha = \Phi^{-1}(1-\alpha/2) \end{cases}$$

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

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

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

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

  **Passo 4 — Concretização**

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

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

**3.** Uma equipa de infectologistas suspeita que as vítimas de covid-19 apresentem uma perda percentual esperada de massa corporal igual a $\mu_0$, 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 $n$ vítimas de covid-19, escolhidas ao acaso, contabilizando-se um total de $\sum_{i=1}^{n} x_i = a$ e um desvio padrão amostral de $s = b$. (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**

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

- **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

$$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \\ &= \frac{\frac{a}{n} - \mu_0}{\frac{b}{\sqrt{n}}} \end{aligned}$$

  e

$$\begin{aligned} valor-p &= P(|T| > |t| \mid H_0) \\ &= P\left(|T| > \left|\frac{\frac{a}{n} - \mu_0}{\frac{b}{\sqrt{n}}}\right| \mid H_0\right) \overset{calc.}{=} \cdots \end{aligned}$$

  devemos

  - não rejeitar $H_0$ a qualquer n.s. $\alpha_0 \leq valor-p$;
  - rejeitar $H_0$ a qualquer n.s. $\alpha_0 > valor-p$.

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

$$\begin{aligned} F^{-1}_{\chi^2_{(4)}}(p_1) = \quad &< \quad t = \cdots < \quad = F^{-1}_{\chi^2_{(4)}}(p_2) \\ p_1 &< \quad F_{\chi^2_{(4)}}(t) < p_2 \\ 1 - p_2 &< \quad valor-p \simeq 1 - F_{\chi^2_{(4)}}(t) < 1 - p_1. \end{aligned}$$

  Assim, podemos adiantar que

  - não devemos rejeitar $H_0$ a qualquer n.s. $\alpha_0 \leq 2 \times (1 - p_2) \times 100\%$,;
  - devemos rejeitar $H_0$ a qualquer n.s. $\alpha_0 \geq 2 \times (1 - p_1) \times 100\%\%$.

**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 (4.0)

$$F_0(x) = \begin{cases} 0, & x \leq 0 \\ x^{\beta}, & 0 < x < 1 \\ 1, & x \geq 1. \end{cases}$$

Em $n$ 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 | $o_1$ | $o_2$ | $o_3$ | $o_4$ | $o_5$ |
| Frequência esperada sob $H_0$ | $E_1$ | $E_2$ | $E_3$ | $E_4$ | $E_5$ |

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**

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

  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**

$$\begin{aligned} E_4 &= n \times F_{X|H_0}(0.85) - F_{X|H_0}(0.70) \\ &= n \times \left(0.85^{\beta} - 0.70^{\beta}\right) \\ E_5 &= n - \sum_{i=1}^{4} E_i. \end{aligned}$$

- **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)**

  Dado que o valor observado da estatística de teste é

$$t = \sum_{i=1}^{k} \frac{(o_i - E_i)^2}{E_i}$$

  e $W = (c, +\infty)$, obtemos

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

  e devemos

– não rejeitar de $H_0$ a qualquer n.s. $\alpha_0 \le valor - p$;

– rejeitar de $H_0$ a qualquer n.s. $\alpha_0 > valor - p$.

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

$$\begin{array}{rcl} F^{-1}_{\chi^2_{(k-1)}}(p_1) = & < & t = \cdots < \quad = F^{-1}_{\chi^2_{(k-1)}}(p_2) \\ p_1 & < & F_{\chi^2_{(k-1)}}(t) < p_2 \\ 1 - p_2 & < & valor - p \simeq 1 - F_{\chi^2_{(k-1)}}(t) < 1 - p_1. \end{array}$$

Assim, podemos adiantar que

– não devemos rejeitar $H_0$ a qualquer n.s. $\alpha_0 \le (1 - p_2) \times 100\%$;

– devemos rejeitar $H_0$ a qualquer n.s. $\alpha_0 \ge (1 - p_1) \times 100\%\%$.

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

$$\sum_{i=1}^n x_i = \sum_{i=1}^n x_i, \quad \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i^2, \quad \sum_{i=1}^n y_i = \sum_{i=1}^n y_i, \quad \sum_{i=1}^n y_i^2 = \sum_{i=1}^n y_i^2, \quad \sum_{i=1}^n x_i y_i = \sum_{i=1}^n x_i y_i,$$

onde $[\min_{i=1,\dots,n} x_i, \max_{i=1,\dots,n} x_i] = [x_{(1)}, x_{(n)}]$.

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 $(1 - \alpha) \times 100\%$ para o valor esperado de $Y$ quando $x = x_0$ 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**

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

- **Estimativas de MV de $\beta_0$ e $\beta_1$; estimativa de $\sigma^2$**

  ◦ $n = n$

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

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

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

  $\sum_{i=1}^n x_i^2 - n\bar{x}^2 = \sum_{i=1}^n x_i^2 - n\left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2$

  ◦ $\sum_{i=1}^n y_i = \sum_{i=1}^n y_i$

  $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i = \frac{1}{n}\sum_{i=1}^n y_i$

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

  $\sum_{i=1}^n y_i^2 - n\bar{y}^2 = \sum_{i=1}^n y_i^2 - n\left(\frac{1}{n}\sum_{i=1}^n y_i\right)^2$

- $\sum_{i=1}^n x_i y_i = \sum_{i=1}^n x_i y_i$
  $\sum_{i=1}^n x_i y_i - n\,\bar{x}\,\bar{y} = \sum_{i=1}^n x_i y_i - n\left(\frac{1}{n}\sum_{i=1}^n x_i\right)\left(\frac{1}{n}\sum_{i=1}^n y_i\right)$.

Deste modo

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^n x_i^2 - n\,\bar{x}^2} \\
&= \frac{\sum_{i=1}^n x_i y_i - n\left(\frac{1}{n}\sum_{i=1}^n x_i\right)\left(\frac{1}{n}\sum_{i=1}^n y_i\right)}{\sum_{i=1}^n x_i^2 - n\left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} \\
&= \left(\frac{1}{n}\sum_{i=1}^n y_i\right) - \hat{\beta}_1 \times \left(\frac{1}{n}\sum_{i=1}^n x_i\right) \\
\hat{\sigma}^2 &= \frac{1}{n-2}\left[\left(\sum_{i=1}^n y_i^2 - n\,\bar{y}^2\right) - (\hat{\beta}_1)^2\left(\sum_{i=1}^n x_i^2 - n\,\bar{x}^2\right)\right] \\
&= \frac{1}{n-2}\left\{\left[\sum_{i=1}^n y_i^2 - n\left(\frac{1}{n}\sum_{i=1}^n y_i\right)^2\right] - [\hat{\beta}_1]^2\left[\sum_{i=1}^n x_i^2 - n\left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2\right]\right\}
\end{aligned}
$$

- **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**

$$
\begin{cases}
a_\alpha = F_{t_{(n-2)}}(\alpha/2) = -F_{t_{(n-2)}}(1-\alpha/2) = \cdots \\
b_\alpha = F_{t_{(n-2)}}(1-\alpha/2) = \cdots
\end{cases}
$$

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

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

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

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

**Passo 4 — Concretização**

Tendo em conta a expressão geral do IC para $\beta_0 + \beta_1 x_0$, o IC pretendido e a sua amplitude são iguais a

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

$$
2 \times F^{-1}_{t_{(n-2)}}(1-\alpha/2) \times \sqrt{\hat{\sigma}^2 \times \left[\frac{1}{n} + \frac{\left(x_0 - \frac{\sum_{i=1}^n x_i}{n}\right)^2}{\sum_{i=1}^n x_i^2 - n\left(\frac{\sum_{i=1}^n x_i}{n}\right)^2}\right]}.
$$