# UFRJ - CCMN - IM - Departamento de Métodos Estatísticos

**Segunda Avaliação de Probabilidade e Estatística** **14-11-2019**
**Atenção: Não serão aceitas respostas sem justificativa.**
**Resolver as questões nos espaços apropriados.**

---

**Q1)** Sejam $X$ e $Y$ variáveis aleatórias cuja função de probabilidade conjunta seja dada como abaixo, onde $p(i,j) = P(X = i, Y = j)$, para $i = 1,2,3$ e $j = 0,1,2$:

$$p(1,0) = 0,1;\ p(2,0) = 0,1;\ p(3,0) = 0,2;$$
$$p(1,1) = 0,1;\ p(2,1) = 0;\ p(3,1) = 0,2;$$
$$p(1,2) = 0,2;\ p(2,2) = 0,1;\ p(3,2) = 0$$

(a) Determine as distribuições marginais de $X$ e $Y$.

(b) $X$ e $Y$ são independentes? Justifique.

(c) Determine a distribuição de $S = X + Y$ e calcule E(S), a esperança de S.

(d) Calcule Var(S), a variância de S.

**Q2)** Espécimes machos de aranhas do gênero *Tidarren* amputam voluntariamente um de seus pedipalpos (direito ou esquerdo) logo antes de atingir a maturidade sexual. Especula-se que a amputação contribui para o aumento da velocidade, o que favoreceria a procura por fêmeas. A velocidade de 9 aranhas (em cm/s) foi medida antes e depois da amputação voluntária.

| Velocidade antes ($x_i$) | | | | | | | | | $\sum x_i$ | $\sum x_i^2$ | $\sum x_i y_i$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1,3 | 2,9 | 2,4 | 3,1 | 3,4 | 3,0 | 2,3 | 2,9 | 3,0 | 24,3 | 68,7 | 92,3 |
| **Velocidade depois ($y_i$)** | | | | | | | | | $\sum y_i$ | $\sum y_i^2$ | |
| 2,1 | 3,3 | 3,6 | 4,3 | 4,2 | 4,0 | 3,3 | 4,5 | 3,7 | 33,0 | 125,2 | |

(a) Calcule os quartis das medidas de velocidade antes e depois da amputação voluntária.

(b) Construa dois *box-plots* na mesma escala para a velocidade antes e depois da amputação voluntária. Comente sobre a diferença observada. Há valores discrepantes?

(c) Construa um diagrama de dispersão para as medidas de velocidade antes e depois da amputação voluntária e calcule o coeficiente de correlação.

**Q3)** Marque verdadeiro ou falso para cada uma das afirmações abaixo, justificando sua resposta.

(a) Considere que se deseja estimar uma proporção populacional $p$ desconhecida com uma precisão especificada. Nesse contexto, denote por $n_1$ o menor tamanho de amostra necessário para garantir tal precisão quando não se tem nenhuma informação sobre $p$ e denote por $n_2$ o menor tamanho de amostra necessário caso tenhamos alguma informação sobre $p$ (ou seja, se soubermos que $p$ pertence a algum intervalo $[a,b] \subset (0,1)$). Podemos afirmar que $n_1$ sempre será no máximo 4 vezes superior a $n_2$, ou seja, que $n_1 \leq 4n_2$.

(b) O erro quadratico médio do estimador média amostral sempre converge para zero quando $n \to \infty$ (considerando que esse estimador está sendo utilizado para estimar o parâmetro média populacional com variância finita e conhecida).

(c) Com intenção de estimar um parâmetro $\theta$ desconhecido, foram propostos inicialmente dois estimadores $\hat{\theta}_1$ e $\hat{\theta}_2$, que satisfazem $E(\hat{\theta}_1) = \theta$ e $E(\hat{\theta}_2) = \frac{(n-1)\theta}{n}$. Pode-se afirmar que um novo estimador $\hat{\theta}_3 = \frac{2}{3}\hat{\theta}_1 + \frac{1}{3}\hat{\theta}_2$ vai ser tendencioso para estimar $\theta$ e que seu viéis será igual a $\frac{\theta}{n}$.

**Q4)** Os dados correspondem ao tempo, em minutos, de 40 funcionários para a realização de uma mesma tarefa:

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 42 | 106 | 44 | 59 | 80 | 48 | 61 | 50 | 63 | 54 | 94 | 99 | 96 | 54 | 62 | 78 | 61 | 93 | 62 | 27 |
| 21 | 88 | 121 | 95 | 49 | 94 | 67 | 68 | 54 | 121 | 62 | 85 | 57 | 83 | 111 | 78 | 165 | 42 | 58 | 103 |

Denotando por $x_i$ as medições acima, temos: $\sum_{i=1}^{40} x_i = 2955$ e $\sum_{i=1}^{40} x_i^2 = 250199$.
Admita que o tamanho amostral é suficientemente grande para que seja válida as aproximações fornecidas pelo TCL.

(a) Construa um Intervalo de 95% de confiança para tempo médio, $\mu$.

(b) Construa um Intervalo de 98% de confiança, não conservativo, para a proporção de tempos maiores que 100 min.

(c) Qual a probabilidade de que o erro absoluto na estimação de $\mu$ (média populacional), com base nessa amostra, seja inferior a 8 minutos?

(d) Qual seria o tamanho de uma nova amostra para que, com 0,95 de probabilidade, o erro absoluto na estimação de $\mu$ seja inferior a 8 minutos?

---

Boa Prova!

# Solução

Q1) (a) $\Omega_X = \{1, 2, 3\}$,
$p_X(1) = P(X = 1) = p(1, 0) + p(1, 1) + p(1, 2) = 0,4$
$p_X(2) = P(X = 2) = p(2, 0) + p(2, 1) + p(2, 2) = 0,2$
$p_X(3) = P(X = 3) = p(3, 0) + p(3, 1) + p(3, 2) = 0,4$

$\Omega_Y = \{0, 1, 2\}$;
$p_Y(0) = P(Y = 0) = p(1, 0) + p(2, 0) + p(3, 0) = 0,4$
$p_Y(1) = P(Y = 1) = p(1, 1) + p(2, 1) + p(3, 1) = 0,3$
$p_Y(2) = P(Y = 2) = p(1, 2) + p(2, 2) + p(3, 2) = 0,3$

(b) Não. $P(X = 2, Y = 1) = 0 \neq P(X = 2)P(Y = 1)$

(c) $P(S = 1) = p(1, 0) = 0,1$; $P(S = 2) = p(2, 0) + p(1, 1) = 0,2$;
$P(S = 3) = p(3, 0) + p(2, 1) + p(1, 2) = 0,4$; $P(S = 4) = p(3, 1) + p(2, 2) = 0,3$
$E(S) = 0,1 + 2 \times 0,2 + 3 \times 0,4 + 4 \times 0,3 = 2,9$

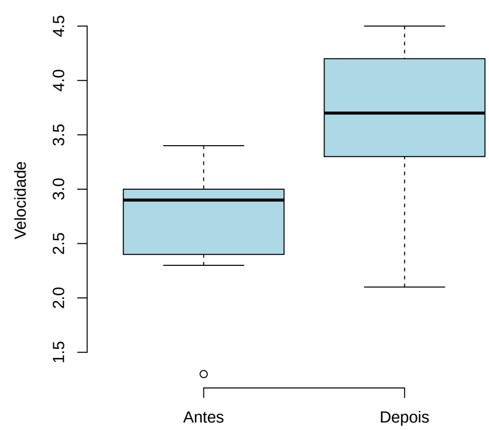
(d) $E(S^2) = 0,1 + 4 \times 0,2 + 9 \times 0,4 + 16 \times 0,3 = 9,3$
$Var(S) = E(S^2) - (E(S))^2 = 9,3 - (2,9)^2 = 9,3 - 8,41 = 0,89$

Q2) (a) Seja $(x_{(1)}, x_{(2)}, \ldots, x_{(9)})$ correspondendo ao conjunto de dados ordenados. Os quartis ocupam as posições

$$\begin{aligned} Q1 &= x_{(\frac{n+3}{4})} = x_{(\frac{9+3}{4})} = x_{(3)}, \\ Q2 &= x_{(\frac{n+1}{2})} = x_{(\frac{9+1}{2})} = x_{(5)}, \\ Q3 &= x_{(\frac{3n+1}{4})} = x_{(\frac{3\times 9+1}{4})} = x_{(7)}. \end{aligned}$$

Assim, os quartis *antes da amputação* são $Q1 = 2,4$, $Q2 = 2,9$ e $Q3 = 3,0$; e *depois da amputação* são $Q1 = 3,3$, $Q2 = 3,7$ e $Q3 = 4,2$.

(b) Apontaremos como valores discrepantes, dados abaixo do limite inferior $LI = Q1 - 1,5(Q3 - Q1)$ e acima do limite superior $LS = Q3 + 1,5(Q3 - Q1)$.
Para os dados *antes da amputação*, temos $LI = 2,4 - 1,5(3,0 - 2,4) = 1,5$ e $LS = 3,0 + 1,5(3,0 - 2,4) = 3,9$ $\Rightarrow$ o valor $1,3$ é discrepante.
Para os dados *depois da amputação*, temos $LI = 3,3 - 1,5(4,2 - 3,3) = 1,95$ e $LS = 4,2 + 1,5(4,2 - 3,3) = 5,55$ $\Rightarrow$ não há valores discrepantes.




(c) O coeficiente de correlação é

$$r_{XY} = \frac{s_{xy}}{s_x s_y},$$

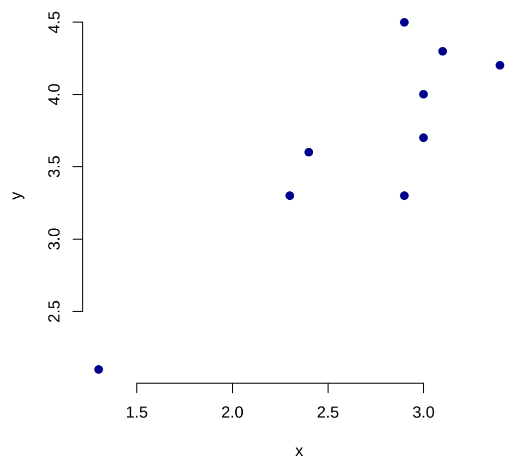
em que

$$s_{xy} = \left(\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}\right)/(n-1) = \left(92,3 - \frac{24,3 \times 33,0}{9}\right)/(9-1) = 0,4,$$

$$s_x = \sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)/(n-1)} = \sqrt{68,7 - \frac{24,3^2}{9}/(9-1)} = 0,6214,$$

$$s_y = \sqrt{\left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right)/(n-1)} = \sqrt{125,2 - \frac{33,0^2}{9}/(9-1)} = 0,7246.$$

Assim, $r_{xy} = 0,4/(0,6214 \times 0,7246) = 0,8884$.




Q3) **(a)** FALSO.

Nesse contexto, estamos buscando o menor valor de $n$ que satisfaz a expressção $P(|\hat{p} - p| < d) \geq \gamma$, para algum $d > 0$ e $\gamma \in (0,1)$. Considerando a distribuição assintótica $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$, temos que:

$$P(|\hat{p}-p| < d) \geq \gamma \ \Rightarrow\ P(-d < \hat{p}-p < d) \geq \gamma \ \Rightarrow\ P(\hat{p}-p < d) \geq \frac{\gamma+1}{2} \ \Rightarrow\ P\left(\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < \frac{d}{\sqrt{\frac{p(1-p)}{n}}}\right) \geq \frac{\gamma+1}{2}.$$

Daí, denotando por $Z_{(\alpha)}$ o quantil $\alpha$ da normal padrão, obtemos que:

$$\frac{d\sqrt{n}}{\sqrt{p(1-p)}} \geq Z_{\left(\frac{\gamma+1}{2}\right)} \quad \Rightarrow \quad \sqrt{n} \geq \frac{Z_{\left(\frac{\gamma+1}{2}\right)}\sqrt{p(1-p)}}{d} \quad \Rightarrow \quad n \geq \frac{\left[Z_{\left(\frac{\gamma+1}{2}\right)}\right]^2 p(1-p)}{d^2} = [C_{\gamma,d}]\ p(1-p),$$

onde $C_{\gamma,d}$ está simplesmente representando tudo que não depende de $p$ na expressão de $n$ e que consequentemente não será relevante na nossa justificativa. Como buscamos o menor tamanho de amostra que satisfaz a inequação acima, iremos tomar como resultado o menor $n$ inteiro que satisfaz a desigualdade acima.

Quando não temos nenhuma informação sobre $p$, devemos substituir $p = 0,5$ na expressão de $n$ (pois esse é o valor de $p$ que maximiza a expressção), obtendo $n_1 = \frac{C_{\gamma,d}}{4}$. Quando sabemos que $p \in [a,b] \subset (0,1)$, iremos substituir $p$ na expressão de $n$ pelo valor mais próximo de $0,5$ dentro do intervalo $[a,b]$. Caso esse valor seja tal que sua utilização na expressão leve a $p(1-p) < \frac{1}{16}$, teremos $n_1 > 4n_2$, refutando a afirmação. Isso ocorreria por exemplo se tivessemos informação de que $p \in (0; 0,05]$.

**(b)** VERDADEIRO.

Temos que EQM$(\bar{X}) \underset{n\to\infty}{\longrightarrow} 0$ que equivale a $E(\bar{X}) \underset{n\to\infty}{\longrightarrow} \mu$ e $Var(\bar{X}) \underset{n\to\infty}{\longrightarrow} 0$. Ambas as afirmações são verdadeiras como podemos ver abaixo:

$$E(\bar{X}) = E\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$$

$$Var(\bar{X}) = Var\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \underset{n\to\infty}{\longrightarrow} 0$$

**(c)** FALSO.

De fato $\hat{\theta}_3$ vai ser tendencioso para estimar $\theta$, mas o viéis desse estimador não será $\frac{\theta}{n}$. Vamos provar que esse viéis está incorreto ao calcular o valor correto:

$$E(\hat{\theta}_3) = E\left(\frac{2}{3}\hat{\theta}_1 + \frac{1}{3}\hat{\theta}_2\right) = \frac{2}{3}E(\hat{\theta}_1) + \frac{1}{3}E(\hat{\theta}_2) = \frac{2}{3}\theta + \frac{1}{3}\left(\frac{n-1}{n}\right)\theta = \left(\frac{3n-1}{3n}\right)\theta$$

Logo, o viéis de $\hat{\theta}_3$ será dado por $|E(\hat{\theta}_3) - \theta| = \frac{\theta}{3n}$.

Q4) (a) $\bar{x} = \frac{2955}{40} = 73,88;$ $\quad s^2 = \frac{1}{39}(250199 - \frac{2955^2}{40}) = 817,91 = 28,60^2.$
Como n=40 é um tamanho de amostra razoavelmente grande poderemos usar a distribuição Normal. Assim,
$d = z_{1-\alpha/2}\frac{s}{\sqrt{n}} = 1,96\frac{28,6}{\sqrt{40}} \approx 8,86$. Então,
$l_{inf} = \bar{x} - d = 73,88 - 8,86 = 65,02$ e
$l_{sup} = \bar{x} + d = 73,88 + 8,86 = 82,74$.

(b) I.C., não conservativo para p, com 98% $\quad \hat{p} = 6/40 = 0,15$
$d = 2,33\sqrt{\frac{0,15\times 0,85}{40}} \approx 0,13$. Então,
$l_{inf} = \hat{p} - d = 0,15 - 0,13 = 0,02$ e
$l_{sup} = \hat{p} + d = 0,15 + 0,13 = 0,28$.

(c) $P(|\bar{x} - \mu| < 8) = P(|Z| < 8\sqrt{n}/s) = 2\Phi(8\sqrt{40}/28,6) - 1 = 2\Phi(1,77) - 1 = 0,9233$.

(d) $n = (\frac{z_{0,975}\times s}{d})^2 = (\frac{1,96\times 28,6}{8})^2 = 49,10 \approx 50$ funcionários.