

MAE 5905: Introdução à Ciência de Dados

Lista 1. Primeiro Semestre de 2024. Entregar 02/04/2024.

1. Num conjunto de dados, o primeiro quartil é 10, a mediana é 15 e o terceiro quartil é 20. Indique quais das seguintes afirmativas são verdadeiras, justificando sua resposta:

- a) A distância interquartis é 5.
- b) O valor 32 seria considerado *outlier* segundo o critério utilizado na construção do *boxplot*.
- c) A mediana ficaria alterada de 2 unidades se um ponto com valor acima do terceiro quartil fosse substituído por outro 2 vezes maior.
- d) O valor mínimo é maior do que zero.

2. Num estudo na área de Oncologia, o número de vasos que alimentam o tumor está resumido na seguinte tabela.

Tabela 1: Distribuição de frequências do número de vasos que alimentam o tumor

Número de vasos	Frequência
0 – 5	8 (12%)
5 – 10	23 (35%)
10 – 15	12 (18%)
15 – 20	9 (14%)
20 – 25	8 (12%)
25 – 30	6 (9%)
Total	66 (100%)

Indique a resposta correta.

- a) O primeiro quartil é 25%.
- b) A mediana está entre 10 e 15.
- c) O percentil de ordem 10% é 10.
- d) A distância interquartis é 50.
- e) Nenhuma das respostas anteriores.

3. Em um teste de esforço cardiopulmonar aplicado a 55 mulheres e 104 homens, foram medidas entre outras, as seguintes variáveis:

Tabela 2: VO2MAX

Grupo	n	Média	Mediana	Desvio Padrão
Normais	56	1845	1707	795
Cardiopatas	57	1065	984	434
DPOC	46	889	820	381

Tabela 3: VCO2MAX

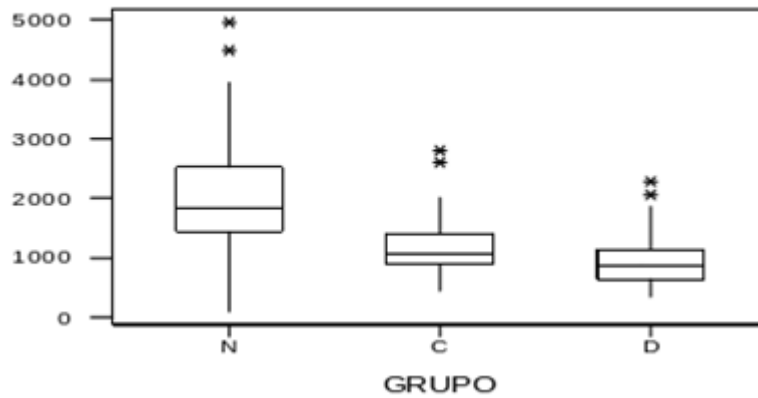
Grupo	n	Média	Mediana	Desvio Padrão
Normais	56	2020	1847	918
Cardiopatas	57	1206	1081	479
DPOC	46	934	860	430

- Grupo: Normais, Cardiopatas ou DPOC (portadores de doença pulmonar obstrutiva crônica).
- VO2MAX: consumo máximo de O₂ (ml/min).
- VCO2MAX: consumo máximo de CO₂ (ml/min).

Algumas medidas descritivas e gráficos são apresentados abaixo nas Tabelas 2 e 3 e Figura 1. Coeficiente de correlação entre VO2MAX e VCO2MAX = 0,92.

- Que grupo tem a maior variabilidade?
 - Compare as médias e as medianas dos 3 grupos.
 - Compare as distâncias interquartis dos 3 grupos para cada variável. Você acha razoável usar a distribuição normal para esse conjunto de dados?
 - O que representam os asteriscos nos *boxplots*?
 - Que tipo de função você ajustaria para modelar a relação entre o consumo máximo de CO₂ e o consumo máximo de O₂? Por quê?
 - Há informações que necessitam verificação quanto a possíveis erros? Quais?
4. O gráfico QQ da Figura 2 corresponde ao ajuste de um modelo de regressão linear múltipla.
- Pode-se afirmar que:
- Há indicações de que a distribuição dos erros é Normal.
 - Há evidências de que a distribuição dos erros é assimétrica.
 - Há evidências de que a distribuição dos erros tem caudas mais leves do que aquelas da distribuição Normal.

Boxplot para VCO2MAX



Boxplot para VO2MAX

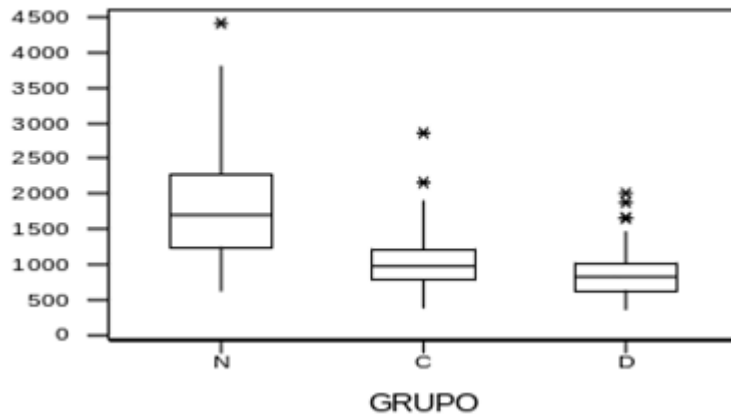


Diagrama de Dispersão VO2MAX x VCO2MAX

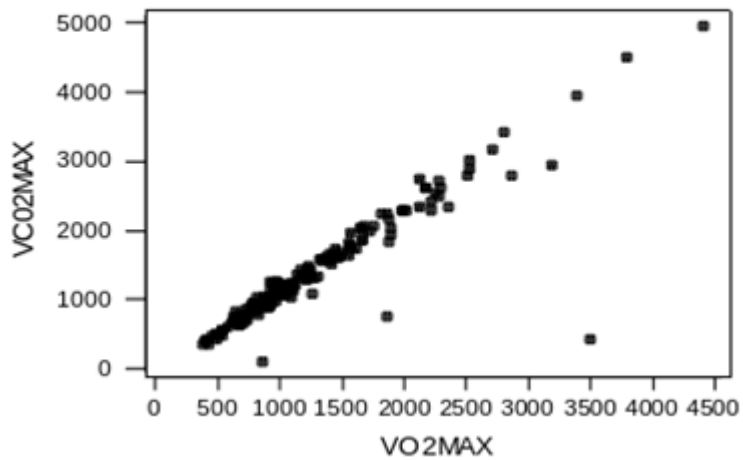


Figura 1: Gráficos para o Exercício 3.

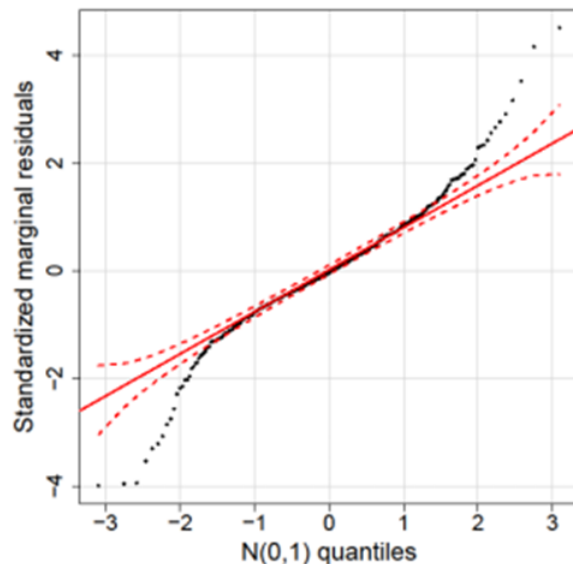


Figura 2: Gráfico QQ correspondente a ajuste de um modelo de regressão linear múltipla.

- d) Há evidências de que a distribuição dos erros tem caudas mais pesadas que aquelas da distribuição Normal.
- e) Nenhuma das anteriores.

5. Para estudar a associação entre gênero (1=Masc, 0=Fem) e idade (anos) e a preferência (1=sim, 0=não) pelo refrigerante Kcola, o seguinte modelo de regressão logística foi ajustado aos dados de 50 crianças escolhidas ao acaso:

$$\log \left\{ \frac{\pi_i(x_i, w_i)}{1 - \pi_i(x_i, w_i)} \right\} = \alpha + \beta x_i + \gamma(w_i - 5),$$

em que x_i (w_i) representa o gênero (idade) da i -ésima criança e $\pi_i(x_i, w_i)$ a probabilidade de uma criança do gênero x_i e idade w_i preferir Kcola. As seguintes estimativas para os parâmetros foram obtidas:

Parâmetro	Estimativa	Erro padrão	Valor p
α	0,69	0,12	< 0,01
β	0,33	0,10	< 0,01
γ	-0,03	0,005	< 0,01

- a) Interprete os parâmetros do modelo por intermédio de chances e razões de chances.
- b) Com as informações acima, estime a razão de chances de preferência por Kcola correspondente à comparação de crianças do mesmo gênero com 10 e 15 anos.
- c) Construa intervalos de confiança (com coeficiente de confiança aproximado de 95%) para $\exp(\beta)$ e $\exp(\gamma)$ e traduza o resultado em linguagem não técnica.

d) Estime a probabilidade de meninos com 15 anos preferirem Kcola.