MAE 5905: Introdução à Ciência de Dados

Lista 1. Primeiro Semestre de 2023

Alunos:

Leonardo de Castro Lima

Leonardo Makoto Kawahara - 7180679

# Questão 1

A. Falsa. A distância interquartis é calculada através da fórmula . Como o primeiro quartil é 10 e o terceiro é 20, .

B. Falsa. De acordo com o critério do boxplot, o limite superior é definido por Como não possuímos , calculamos o limite superior como: = 35. Dado que 35>32, 32 não seria considerado um *outlier*.

C. Falso. A mediana é uma medida resistente, mudando pouco quando um pequeno conjunto de valores é alterado. No caso específico, ela se manterá constante, dado que o ponto no terceiro quartil já está a sua direita e, ao ser multiplicado por 2, continuará a direita.

D. Falso. Não é possível afirmar com exatidão qual o valor mínimo deste conjunto de dados. O valor mínimo do boxplot, considerando a distância interquartis e sem conhecimento do valor mínimo, é -5, o que indica que valores inferiores a zero não seriam sequer considerados outliers para este conjunto de dados.

# Questão 2

A resposta correta é a letra B.

# Questão 3

A. Normais. O desvio padrão é superior aos demais grupos.

B. As médias de todos os grupos superam as medianas, havendo assimetria positiva na distribuição. Além disso, a média (mediana) tanto de VO2MAX quanto de VCO2MAX dos normais é superior à dos cardiopatas, que, por sua vez, é superior ao grupo DPOC. Em termos gerais, conclui-se que, quanto mais saudável o indivíduo, maior o consumo máximo de O2 e CO2.

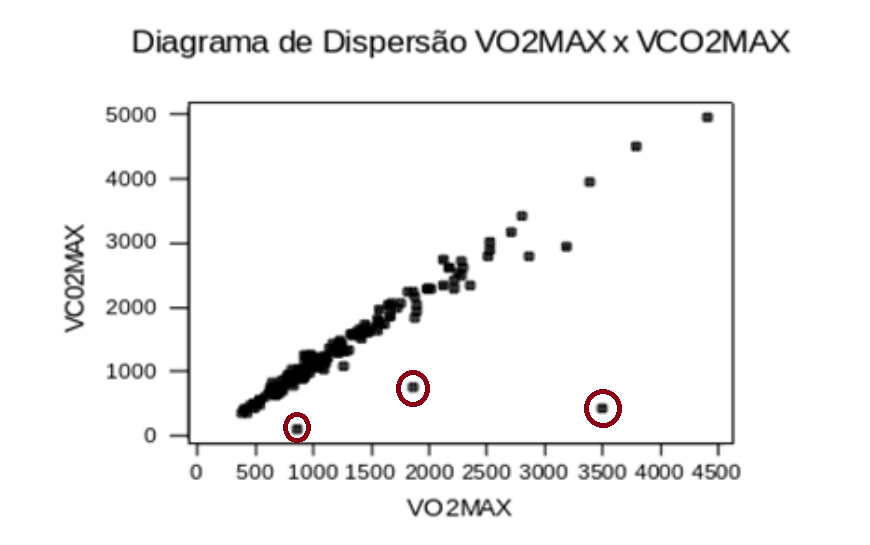
C. Observando a figura, é notável que a distância interquartil dos normais é maior que os demais tanto para o VO2MAX quanto para o VCO2MAX. De forma similar ao caso da média, o grupo de cardiopatas parece ter uma distância interquartis superior ao grupo DPOC.

Não é recomendado utilizar a distribuição normal para este conjunto de dados, dado que há assimetria positiva na distribuição (média superior a mediana) e a distribuição normal não supõe assimetria.

D. Os asteriscos no *boxplot* representam *outliers*.

E. Algum tipo de função linear sem intercepto do tipo . A sugestão advém da observação da relação linear forte entre as duas variáveis, assim como denotado no gráfico e na correlação.

F. Sim. No gráfico de dispersão parece haver uma observação com consumo máximo muito alto de VO2 e baixo de VCO2 que se destaca. Há duas outras observações que levantam dúvidas cuja relação entre VO2MAX e VCO2MAX fica fora do padrão, com o VCO2MAX abaixo dos demais para o mesmo valor de VO2MAX. A figura abaixo traz estas observações circuladas em vermelho.



# Questão 4

A alternativa correta é a letra D: há evidências de que a distribuições dos erros tem caudas mais pesadas que aquelas da distribuição normal. Com base no gráfico QQ, é possível observar que os resíduos dos quantis que estão nas caudas da distribuição se afastam de forma considerável da linha que marca os resíduos de uma normal padrão e do seu respectivo intervalo de confiança (pontos pretos vs linha e pontos em vermelho).

# Questão 5

1. 1. nos dá a chance de uma criança do sexo feminino de 5 anos preferir Kcola. Nesse caso, a chance será de aproximadamente 1,99.

2. é a razão da chance de um indivíduo do sexo masculino preferir Kcola em relação a um indivíduo do sexo feminino. Nesse caso, será de aproximadamente 1,39. Ou seja, a chance de os meninos preferirem Kcola em relação as meninas é 1,39 maior.

3. é a razão da chance de uma criança preferir Kcola em relação a outra criança com um ano a menos. Nesse caso, para cada ano adicional, a chance de uma criança preferir Kcola cai em aproximadamente 0,97

1. A razão das chances de crianças do mesmo gênero de 15 anos em relação a 10 anos poderá ser calculada através de
2. Com relação a beta, temos que os intervalos de confiança a 95% considerando o log da chance são:

Para encontrar o intervalo de confiança da chance, aplicamos exp em relação as bandas do log da chance:

Ou seja, se a amostra das 50 crianças for representativa da população, o aumento da chance de um menino preferir Kcola em relação a uma menina estará entre 1.14 e 1.7 para 95% dos casos.

Para o coeficiente gama, temos:

Aplicando exp:

Ou seja, para cada ano a mais da criança, haverá uma redução da chance de ela preferir Kcola entre 2% e 4% em 95% dos casos.

1. O log da chance de meninos com 15 anos preferirem Kcola pode ser medido por:

Para encontrar a probabilidade dos meninos preferirem Kcola, primeiro exponenciamos ambos os lados, tal que:

Isolando p, temos que:

Ou seja, a probabilidade de um menino de 15 anos preferir Kcola é 0,67.