Universidade Federal da Bahia
Programa de Pós-Graduação em Ciência de Dados e Big Data
Carlos Magno Santos Ribeiro de Brito

Prof Dr. Paulo Canas Rodriques

Respostas da primeira lista de exercício

Questão 01) Um inquérito a 100 pessoas em Salvador sobre as suas idades resultou no quadro seguinte:

Escalão Etário	Frequência absoluta
$[0, 20]$	20
$]20, 40]$	25
$]40, 60]$	40
$]60, 80]$	10
$]80, 100]$	5

a) Obtenha a tabela de frequências completa, incluindo a frequência absoluta acumulada, relativa e relativa acumulada.

A partir dos dados informados na tabela, obtém-se os seguintes resultados:

Escalão Etário	Freq. Abs	Freq. Rel (%)	Freq. Abs. Acum	Freq. Rel. Acum (%)
$[0, 20]$	20	20	20	20
$]20, 40]$	25	25	45	45
$]40, 60]$	40	40	85	85
$]60, 80]$	10	10	95	95
$]80, 100]$	5	5	100	100

b) Calcule, para a idade:

i - A média:

A média pode ser dada pelo valor médio nos intervalos de classe multiplicados pela frequência relativa.

$$\bar{x} = \sum_{i=1}^{n} x_i \cdot \frac{f_i}{100}$$

Onde x_i é a idade média em cada intervalo, f_i é a frequência relativa, e $\bar{x}$ é a média.

$$\bar{x} = 10 \times 0,2 + 30 \times 0,25 + 50 \times 0,4 + 70 \times 0,1 + 90 \times 0,05 = 41 \text{ anos} \tag{1}$$

ii - A mediana:

A mediana pode ser definida somando os dois números do meio e dividindo por dois em valores ordenados. Quando a quantidade de números N que estamos avaliando for par ou escolhendo o número do meio caso N seja ímpar.

Para o caso específico de intervalos de classe, com dados agrupados, podemos fazer uma regra de três com a diferença entre metade do tamanho da amostra $\frac{n}{2} = 50$ e o valor da frequência absoluta acumulada f_{ac} no intervalo anterior $]20, 40] = 45$, juntamente ao valor da frequência absoluta $f_{ab} = 40$ e a amplitude do intervalo $]40, 60] \rightarrow h = 60 - 40 = 20$. Após isso, soma-se o encontrado com o limite inferior da sua classe pertencente ($L_{min} = 40$). Matemáticamente, pode-se representar como:

$$\text{mediana} = L_{min} + \frac{\left(\frac{n}{2} - f_{ac}\right)}{f_{ab}} \cdot h$$

$$\text{mediana} = 40 + \frac{(50 - 45) \cdot 20}{40} = 42,5 \tag{2}$$

iii - A moda:

A moda pode ser definida como a média dos valores mais frequentes. Com essa definição, observa-se inicialmente que o intervalo de classe $[40, 60]$ é o mais frequente. Pela fórmula de Czuber, podemos obter a moda como:

$$\text{moda} = L_{min} + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot h$$

Onde Δ_1 é a diferença entre a frequência da classe modal e a classe anterior e Δ_2 é a diferença entre a frequência da classe modal e a classe seguinte.

$$\Delta_1 = 40 - 25 = 15;$$
$$\Delta_2 = 40 - 10 = 30;$$
$$h = 60 - 40 = 20$$

$$\text{moda} = 40 + \frac{15}{15 + 30} \cdot 20 \approx 42,67 \tag{3}$$

c) Obtenha o boxplot. Interprete o boxplot.

Para montar o boxplot, é necessário primeiro obter os dados dos primeiro e terceiro quartis, além de uma medida de dispersão que pode ser dada como a diferença entre o terceiro e primeiro quartis.

Para o cálculo dos quartis, as equações são similares à mediana (que é o segundo quartil):

$$q_1 = L_{min} + \frac{\left(\frac{n}{4} - f_{ac}\right)}{f_{ab}} \cdot h = 20 + \frac{(25 - 20) \cdot 20}{25} = 24$$
$$q_3 = L_{min} + \frac{\left(3 \cdot \frac{n}{4} - f_{ac}\right)}{f_{ab}} \cdot h = 40 + \frac{(75 - 45) \cdot 20}{40} = 55$$

Com os quartis calculados, obtém-se a dispersão e os límites do boxplot:

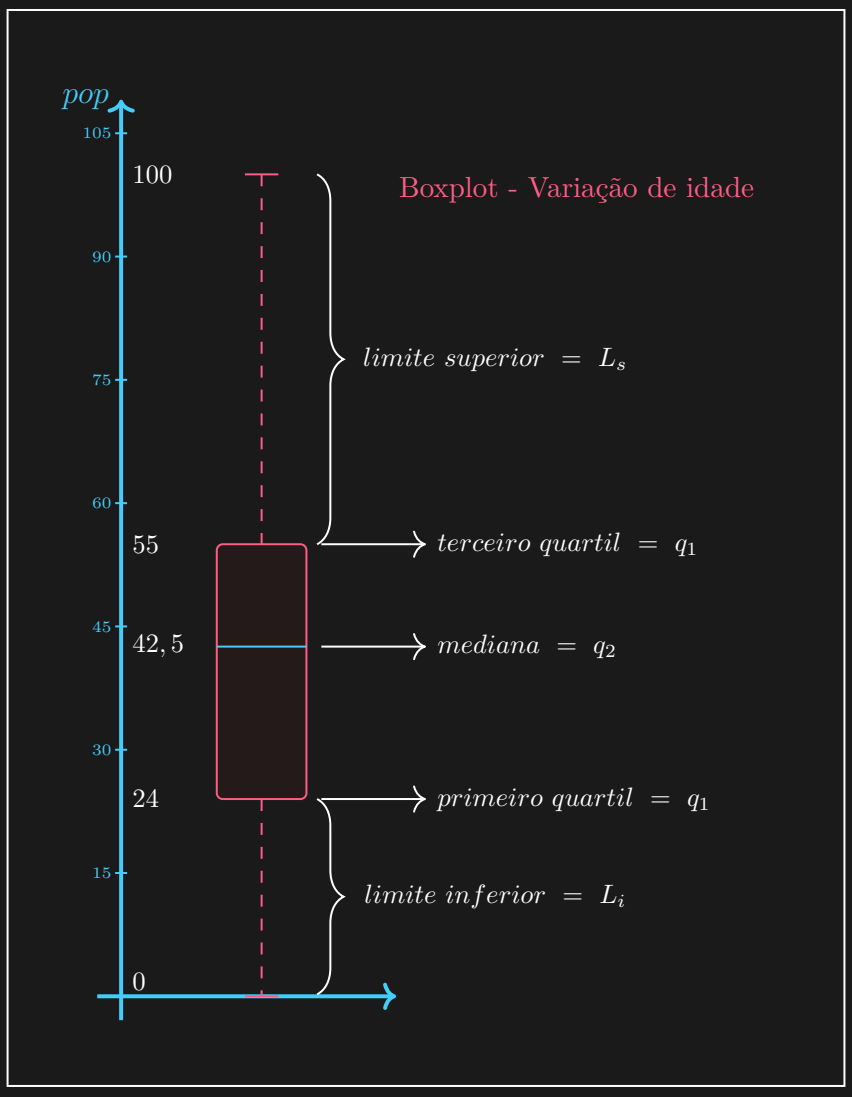
$$d_q = q_3 - q_1 = 55 - 24 = 31$$

$$L_s = q_3 + 1,5 \cdot d_q = 45 + 1,5 \cdot 31 = 101,5$$

$$L_i = q_1 - 1,5 \cdot d_q = 24 - 1,5 \cdot 31 = -22,5$$

Como o limite inferior de idade na amostra é 0, então $L_i = 0$. Assim como o limite superior é 100, então $L_s = 100$.

A partir desses valores, podemos montar o boxplot como:



No boxplot podemos observar uma assimatria que se manifesta na maior concentração de pessoas com idades entre 24 e 55 anos (primeiro e terceiro quartis), com maior prevalencia para idades próximas à mediana. Nesse intervalo de dados, **não há nenhum valor** além do limite superior nem inferior. Quando isso ocorre, esses valores são tidos como outliers. Do ponto de vista estatístico eles podem ser considerados um erro de observação ou arredondamento. Contudo, não necessariamente se pode afirmar tal relação, pois a situação ainda é possível além do limite superior, mesmo sendo atípica.

d) Qual a percentagem de pessoas que têm idade inferior a 50 anos?

Tem-se a soma das frequências relativas dos dois primeiros intervalos de classe, juntamente com metade da frequência do terceiro. Isto é:

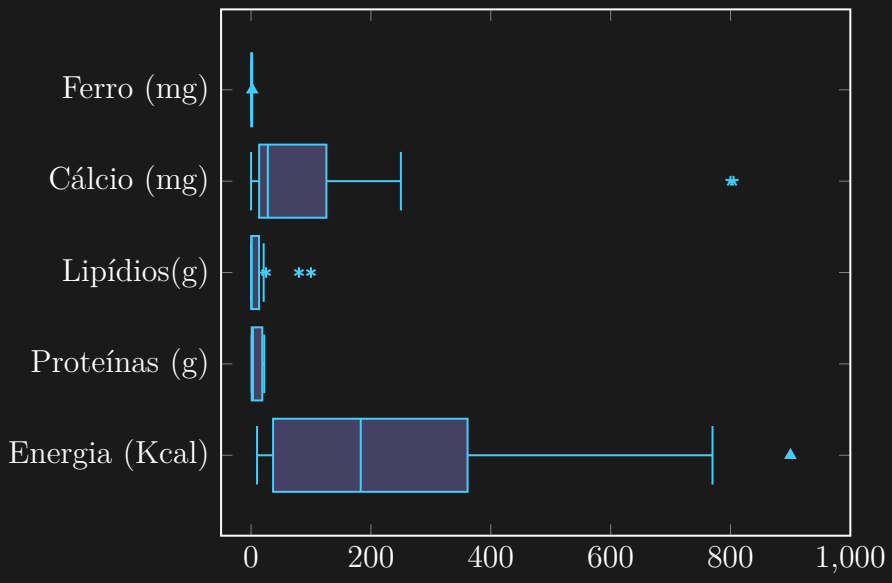
$$f_{rel} = f_{rel}^{[0,20]} + f_{rel}^{]20,40]} + \frac{f_{rel}^{]40,60]}}{2} = 20\% + 25\% + 20\% = 65\%$$

e) Quantas pessoas têm idades entre 40 e 80 anos?

Para essa questão, basta apenas somar os valores da frequência absoluta do terceiro e quarto intervalos:

$$f_{abs} = f_{abs}^{]40,60]} + f_{abs}^{]60,80]} = 40 + 10 = 50 \text{ pessoas}$$

Questão 02) Uma análise às características de 20 alimentos, utilizando o software IBM SPSS 20, resultou nos seguintes resultados:



	Energia (Kcal)	Proteínas (g)	Lipídios (g)	Cálcio (mg)	Ferro (mg)
N - Valid	20	20	20	20	20
N - Missing	0	0	0	0	0
Mean	240,80	8,550	13,775	135,505	1,427
Median	183,00	3,100	0,950	28,00	1,000

Std. Deviation	246,687	9,2488	28,3894	235,9819	1,5640
Percentile (25)	36,75	1,125	0,275	13,500	0,275
Percentile (50)	183,00	3,100	0,950	28,00	1,000
Percentile (75)	361,25	18,750	13,375	125,750	1,750

a) Cite e classifique as variáveis em estudo.

Em termos gerais, tem-se alimentos com valor energético bastante diverso. Isso é visto no *boxplot*, onde temos casos variando de algo próximo de zero (limite inferior) a oitocentos (limite superior). Há um *outlier* (azeite) que tem valor de energia próximo de mil. Ao menos metade dos alimentos apresentam valores energéticos no intervalo de 36 a 360 Kcal, que são os quartis 1 e 3. Além disso, tem-se o Cálcio que também apresenta valores com grande variação e com dois alimentos com elevado potencial (*outliers*). Em relação Ferro, Proteínas e Lipídios, há pouca variação de valores. Isso fica evidente pelo *boxplot* dessas variáveis. Há alimentos atípicos com valores de lipídios fora do limite superior, ainda assim.

b) Qual das variáveis apresenta uma maior dispersão? Qual a medida utilizada para responder a tal pergunta?

Tem-se bem claramente, em primeiro e segunto lugares, que a Energia e o Cálcio são as duas variáveis que apresentam maior dispersão. A medida utilizada para responder a essa pergunta é o Desvio Padrão, que em ambos os casos ultrapassa o valor de 200.

c) É correto dizer que a quantidade média de cálcio é superior à quantidade média de proteínas? Porquê?

Não. Temos médias representadas com unidades de medidas diferentes. Enquanto as proteínas são representadas em gramas, a quantidade de cálcio é representada em miligramas. Para tanto, se o valor de cálcio for convertido em gramas é possível ver que ele é bem menor que a quantidade de proteínas.

d) A partir da tabela e figura acima, escreva um breve parágrafo com a interpretação e comparação dos resultados.

O estudo apresenta resultados com unidades de medidas diferentes (g e mg) e isso pode confundir a priori caso se queira comparar valores de diferentes variáveis entre si. Atentando-se a esse detalhe, observa-se que as proteínas e lipídios são os dois que apresentam maiores valores numéricos. O ferro apresenta valores bastante baixos em todos os alimentos. E o cálcio, por fim, tem quantidades relativamente menores (dados em miligramas) e bem variável nos alimentos, indo de 0 a quase 1,000 miligramas. Em termos energéticos, vemos também

uma grande variabilidade de valores bem evidentes no *boxplot*.

Questão 03) Um inquérito por amostragem sobre o número de filhos em famílias de baixo escalão de rendimento permitiu obter os seguintes números:

$$\{3, 5, 4, 2, 3, 3, 5, 8, 7, 5, 4, 5, 6, 3, 4, 5, 4, 6, 7, 6\}$$

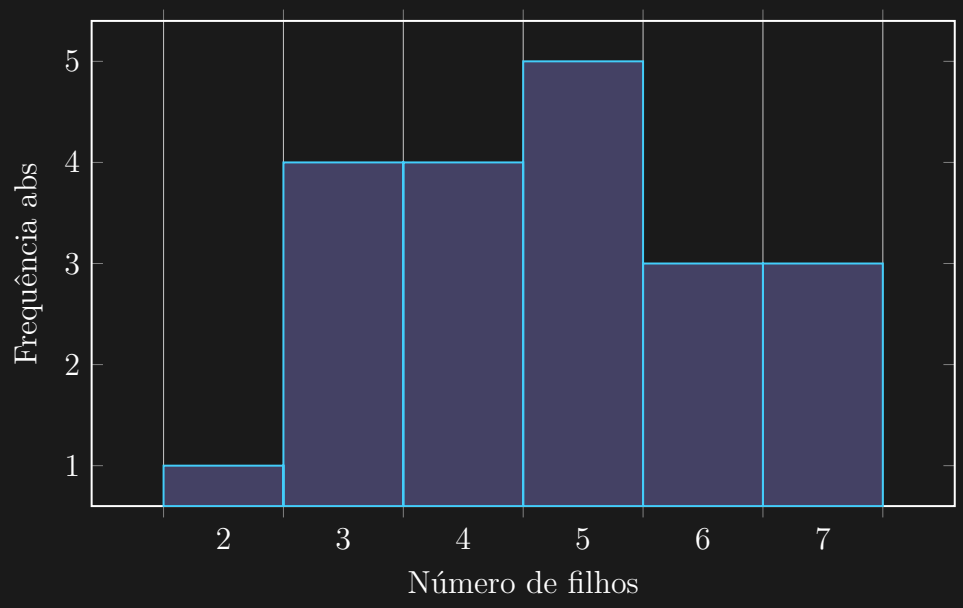
a) Determine a média, a moda e a variância a partir dos dados não classificados.

$$N = 20$$

$$media = \frac{3+5+4+2+3+3+5+8+7+5+4+5+6+3+4+5+4+6+7+6}{20} =$$

$$= \frac{95}{20} = 4,75$$

para a moda, pode-se plotar um histograma e verificar o maior número de ocorrências:



Observa-se que o valor mais frequente é o 5. Logo, a moda é 5.

Para a variância, tem-se:

$$s^2 = \frac{\sum_{n=1}^{20} (x_i - \bar{x})^2}{n-1} \approx 2,51$$

d) Classifique (i.e. agrupe) os dados e verifique os valores calculados em (a). Que conclusões se podem tirar?

Ordenando...

$$\{2,3,3,3,3,4,4,4,4,5,5,5,5,5,6,6,6,7,7,8\}$$

Conforme visto no histograma da questão anterior, as famílias têm 5 filhos com maior frequência e poucas famílias de baixa renda têm poucos filhos (2). Para uma análise mais aprofundada desses dados, as faixas de rendimento também poderiam ser relacionadas.

Questão 04) Considere a seguinte tabela:

Hábito de fumar	Doença no pulmão		Total
	Presente	Ausente	
Sim	25	15	40
Não	10	40	50
Total	35	55	90

a) Verifique se existe associação entre o hábito de fumar e a presença ou ausência de doença do pulmão.

Tem-se que em pessoas que fumam, a percentagem de doenças de pulmão igual a $\frac{25}{40} = 62,5\%$. A ausência equivale a $\frac{15}{40} = 37,5\%$.

Nos não fumantes, as doenças de pulmão são equivalentes a $\frac{10}{50} = 20\%$. A ausência equivale a $\frac{40}{50} = 80\%$.

Nessa amostra, especificamente, foi possível constatar uma relação entre o hábito de fumar e a presença de doença do pulmão, já que entre os fumantes $62,5\%$ deles têm doença no pulmão, enquanto nos não fumantes há uma relação de 20% de pessoas que têm doença do

pulmão. Além disso, os não fumantes apresentam uma quantidade significativa de 80% de pessoas sem doença pulmonar. Dadas essas circunstâncias, é razoável supor que o cigarro é o causador de doenças de pulmão nessa amostra populacional.

b) Qual a percentagem de fumantes?

$$fumantes = \frac{40}{90} \approx 44,4\%$$

c) Quantas pessoas são fumantes e não têm doença do pulmão?

Total de 15 pessoas o que corresponde a $\frac{15}{90} \approx 16,67\%$ da população total.

Questão 05) : Pretende-se estudar a relação entre a idade das pessoas e a quantidade de vezes que comem fruta e vegetais por semana, no estado da Bahia. Descreva detalhadamente como efetuaria esse estudo, desde a definição do problema até à análise e interpretação dos resultados

Inicialmente, faria uma definição geral do problema. Quais as relações principais do estudo, quais as variáveis a serem utilizadas, a população e onde efetuar o estudo. Em seguida, a elaboração de questionário digitais/físicos[1] a serem aplicados com as questões de estudo. Após recolhimento das respostas, no tratamento e interpretação dos resultados, montaria gráficos e tabelas a fim de facilitar a compreensão do estudo. Para montagem desses gráficos e tabelas, utilizaria o software R ou python com algumas bibliotecas padrão das linguagens e com isso, poderia também lançar mão de modelos estatísticos para melhor entender o resultado.

[1] De preferência em formulários online para facilitar na digitalização dos resultados