UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA

Alunos
Gleyce Alves Pereira da Silva
Ivanildo Batista da Silva Júnior
Jaine de Moura Carvalho
Taciana Araújo da Silva

Professor
Dr. Lucian Bogdan Bejan

Resolução da segunda lista de Estatística Aplicada

Recife-PE, 9 de maio de 2021

Sumário

1 Questão 1 . . . . . **1**
1.1 Resolução da questão 1 2
1.1.1 Letra a) 2
1.1.2 Letra b) 4
1.1.3 Letra c) 4
1.1.4 Letra d) 4
1.1.5 Letra e) 4

2 Questão 2 . . . . . **5**
2.1 Resolução da questão 2 5

3 Questão 3 . . . . . **7**
3.1 Resolução da questão 3 8
3.1.1 Letra a) 8
3.1.2 Letra b) 9
3.1.3 Letra c) 9
3.1.4 Letra d) 10
3.1.5 Letra e) 10
3.1.6 Letra f) 11
3.1.7 Letra g) 11

4 Questão 4 . . . . . **12**
4.1 Resolução da questão 4 12

5 Questão 5 . . . . . **13**
5.1 Resolução da questão 5 13

1 Questão 1

Numa pesquisa sobre rotatividade de mão-de-obra, para uma amostra de 40 pessoas foram observadas duas variáveis: número de empregos nos últimos dois anos (X) e salário mais recente, em número de salários mínimos (Y). Os resultados foram:

Indivíduo	X	Y	Indivíduo	X	Y
1	1	6	21	2	4
2	3	2	22	3	2
3	2	4	23	4	1
4	3	1	24	1	5
5	2	4	25	2	4
6	2	1	26	3	2
7	3	3	27	4	1
8	1	5	28	1	5
9	2	2	29	4	4
10	3	2	30	3	3
11	2	5	31	2	2
12	3	2	32	1	1
13	1	6	33	4	1
14	2	6	34	2	6
15	3	2	35	4	2
16	4	2	36	3	1
17	1	5	37	1	4
18	2	5	38	3	2
19	2	1	39	2	3
20	2	1	40	2	5

(a) Usando a mediana, classifique os indivíduos em dois níveis, alto e baixo, para cada uma das variáveis, e construa a distribuição de freqüências conjunta das duas classificações.

(b) Qual a porcentagem das pessoas com baixa rotatividade e ganhando pouco?

(c) Qual a porcentagem das pessoas que ganham pouco?

(d) Entre as pessoas com baixa rotatividade, qual a porcentagem das que ganham pouco?

(e) A informação adicional dada em (d) mudou muito a porcentagem observada em (c)? O que isso significa?

1.1 Resolução da questão 1

1.1.1 Letra a)

Extraindo a base de dados:

```
df1 = df1[0]

df2 = df1[['Indivíduo', 'X', 'Y']]
df3 = df1[['Indivíduo.1', 'X.1', 'Y.1']]
df3 = df3.rename(columns={'Indivíduo.1':'Indivíduo', 'X.1':'X', 'Y.1':'Y'})

df1 = df2.append(df3)

df1=df1.reset_index().drop('index', axis=1)

#primeiras 10 linhas da base de dados
df1.head(10)
```

	Indivíduo	X	Y
0	1	1	6
1	2	3	2
2	3	2	4
3	4	3	1
4	5	2	4
5	6	2	1
6	7	3	3
7	8	1	5
8	9	2	2
9	10	3	2

Definindo as medianas:

```
print('Mediana da variável X: ', df1['X'].median())
print('Mediana da variável Y: ', df1['Y'].median())
```

```
Mediana da variável X:  2.0
Mediana da variável Y:  2.5
```

Classificando as observações de cada coluna:

```
#Classificando a coluna X
lista1=df1['X'].tolist()
lista2=[]
for i in lista1:
  if i >= df1['X'].median():
    i='alto'
  else:
    i='baixo'
  lista2.append(i)

#Classificando a coluna Y
lista3=df1['Y'].tolist()
lista4=[]
for i in lista3:
  if i >= df1['Y'].median():
    i='alto'
  else:
    i='baixo'
  lista4.append(i)

#criando as novas colunas
df1['Class_X']=lista2
df1['Class_Y'] = lista4
```

Criando a tabela de distribuição de frequência conjunta:

```
tabela = pd.crosstab(df1["Class_X"], df1["Class_Y"], margins=True)
tabela
```

Class_Y	alto	baixo	All
Class_X			
alto	13	19	32
baixo	7	1	8
All	20	20	40

Tabela em termos percentuais:

```
tabela/tabela['All'][2]
```

Class_Y	alto	baixo	All
Class_X			
alto	0.325	0.475	0.8
baixo	0.175	0.025	0.2
All	0.500	0.500	1.0

1.1.2 Letra b)

```
print('A porcentagem é de :',(tabela['baixo'][1]/tabela['All'][2])*100, '%')
```

```
A porcentagem é de : 2.5 %
```

1.1.3 Letra c)

```
print('A porcentagem é de :',(tabela['baixo']['All']/tabela['All'][2])*100, '%')
```

```
A porcentagem é de : 50.0 %
```

1.1.4 Letra d)

```
print('A porcentagem é de :',(tabela['baixo'][1]/tabela['All'][1])*100, '%')
```

```
A porcentagem é de : 12.5 %
```

1.1.5 Letra e)

Resposta : Sim, a porcentagem caiu muito (de 50% para 12.5%), isso mostra que as pessoas que ganham pouco possuem uma alta rotatividade.

2 Questão 2

Qual o valor de χ^2 e de C para os dados do Problema 3? E para o Problema 6? Calcule T.

2.1 Resolução da questão 2

Tabela com as informações:

```
tabela
```

Class_Y	alto	baixo	All
Class_X			
alto	13	19	32
baixo	7	1	8
All	20	20	40

Calculando o χ^2:

```
a = tabela[:2][1:2]['alto'][0]
b = tabela[:2][1:2]['baixo'][0]
c = tabela[:2][:1]['alto'][0]
d = tabela[:2][:1]['baixo'][0]
e = tabela[:2][:1]['All'][0]
f = tabela[:2][1:2]['All'][0]
g = tabela[:2][:1]['All'][0]/2
h = tabela[:2][1:2]['All'][0]/2
```

```
chi = ((b-h)**2)/h + ((a-h)**2)/h + ((d-g)**2)/g + ((c-g)**2)/g
chi
```

```
5.625
```

Também pode ser calculado pela função abaixo, da biblioteca *SciPy*:

```
from scipy.stats import chi2_contingency
chi2_contingency(tabela)[0]
```

```
5.625
```

Calculando o valor de C.

```
c = np.sqrt(chi/(chi+tabela['All'][2]))
c
```

```
0.35112344158839 17
```

Calculando o T : No cálculo $r = s = 2$ (categorias alto e baixo).

```
r = 2
s = 2
np.sqrt((chi/tabela['All'][2])/((r-1)*(s-1)))
```

```
0.375
```

3 Questão 3

O departamento de vendas de certa companhia foi formado há um ano com a admissão de 15 vendedores. Nessa época, foram observados para cada um dos vendedores os valores de três variáveis:

T : resultado em um teste apropriado para vendedores;

E : anos de experiência de vendas;

G : conceito do gerente de venda, quanto ao currículo do candidato. O diretor da companhia resolveu agora ampliar o quadro de vendedores e pede sua colaboração para responder a algumas perguntas. Para isso, ele lhe dá informações adicionais sobre duas variáveis:

V : volume médio mensal de vendas em s.m.;

Z : zona da capital para a qual o vendedor foi designado.

O quadro de resultados é o seguinte:

Vendedor	T: teste	E: experiência	G: conceito do gerente	V: vendas	Z: zona
1	8	5	Bom	54	Norte
2	9	2	Bom	50	Sul
3	7	2	Mau	48	Sul
4	8	1	Mau	32	Oeste
5	6	4	Bom	30	Sul
6	8	4	Bom	30	Oeste
7	5	3	Bom	29	Norte
8	5	3	Bom	27	Norte
9	6	1	Mau	24	Oeste
10	7	3	Mau	24	Oeste
11	4	4	Bom	24	Sul
12	7	2	Mau	23	Norte
13	3	3	Mau	21	Sul
14	5	1	Mau	21	Oeste
15	3	2	Bom	16	Norte

Dados: $\sum T = 91$ $\sum T^2 = 601$ $\sum TV = 2.959$

$\sum E = 40$ $\sum E^2 = 128$ $\sum EV = 1.260$

$\sum V = 453$ $\sum V^2 = 15.509$

Mais especificamente, o diretor lhe pede que responda aos sete itens seguintes:

(a) Faça o histograma da variável V em classes de 10, tendo por limite inferior da primeira classe o valor 15.

(b) Encontre a média e a variância da variável V. Suponha que um vendedor seja considerado excepcional se seu volume de vendas é dois desvios padrões superior à média geral. Quantos vendedores excepcionais existem na amostra?

(c) O diretor de vendas anunciou que transferirá para outra praça todos os vendedores cujo volume de vendas for inferior ao 1o quartil da distribuição. Qual o volume mínimo de vendas que um vendedor deve realizar para não ser transferido?

(d) Os vendedores argumentam com o diretor que esse critério não é justo, pois há zonas de venda privilegiadas. A quem você daria razão?

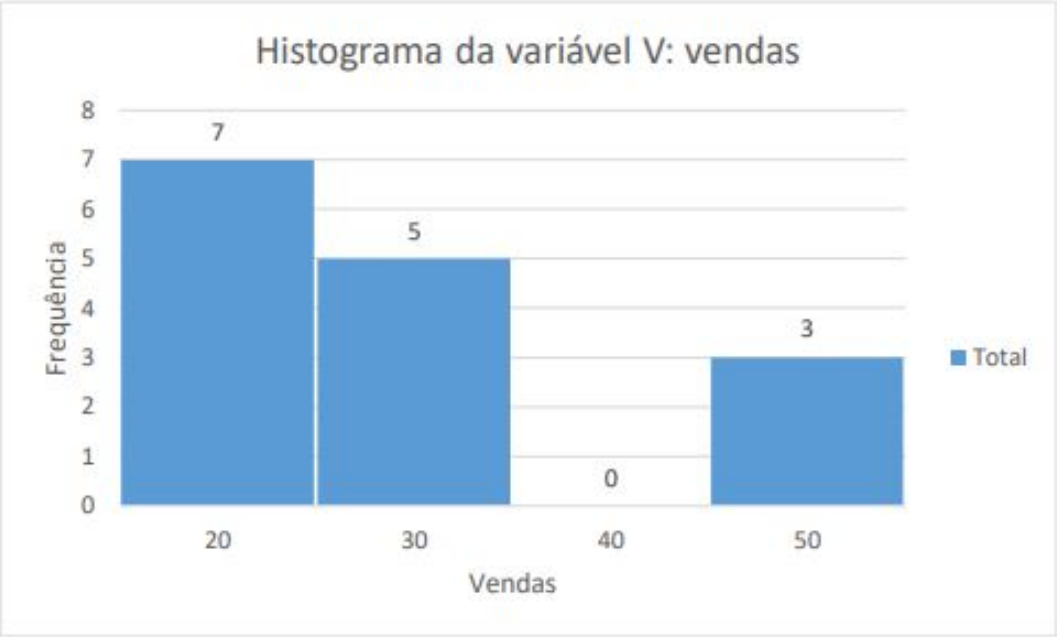
(e) Qual das três variáveis observadas na admissão do pessoal é mais importante para julgar um futuro candidato ao emprego?

(f) Qual o grau de associabilidade entre o conceito do gerente e a zona a que o vendedor foi designado? Você tem explicação para esse resultado?

(g) Qual o grau de associação entre o conceito do gerente e o resultado do teste? E entre zona e vendas?

3.1 Resolução da questão 3

3.1.1 Letra a)



3.1.2 Letra b)

Calculando a média:

$$\overline{X} = \frac{(54 + 50 + 48 + 32 + 30 + 30 + 29 + 27 + 24 + 24 + 24 + 23 + 21 + 21 + 16}{15} = \frac{453}{15} = 30.2$$

Calculando a variância :

$$\begin{aligned} S^2 = \quad & [(54 - 30,2)^2 + (50 - 30,2)^2 + \\ & (48 - 30,2)^2 + (32 - 30,2)^2 + (30 - 30,2)^2 + \\ & (30 - 30,2)^2 + (29 - 30,2)^2 + (27 - 30,2)^2 + \\ & (24 - 30,2)^2 + (24 - 30,2)^2 + (24 - 30,2)^2 + \\ & (23 - 30,2)^2 + (21 - 30,2)^2 + (21 - 30,2)^2 + \\ & (16 - 30,2)^2]/14 \\ & = 1828.4/14 \\ & = 130.6 \end{aligned}$$

Para um vendedor ser considerado excepcional seu volume de vendas deve ser maior que $2S + \overline{X}$. Temos que $S = \sqrt{130.6} \approx 11.43$, logo o valor de S será $(2 \times 11.43) + 30.2 = 53.06$.
Observando na tabela, há apenas um vendedor com vendas acima desse valor.

3.1.3 Letra c)

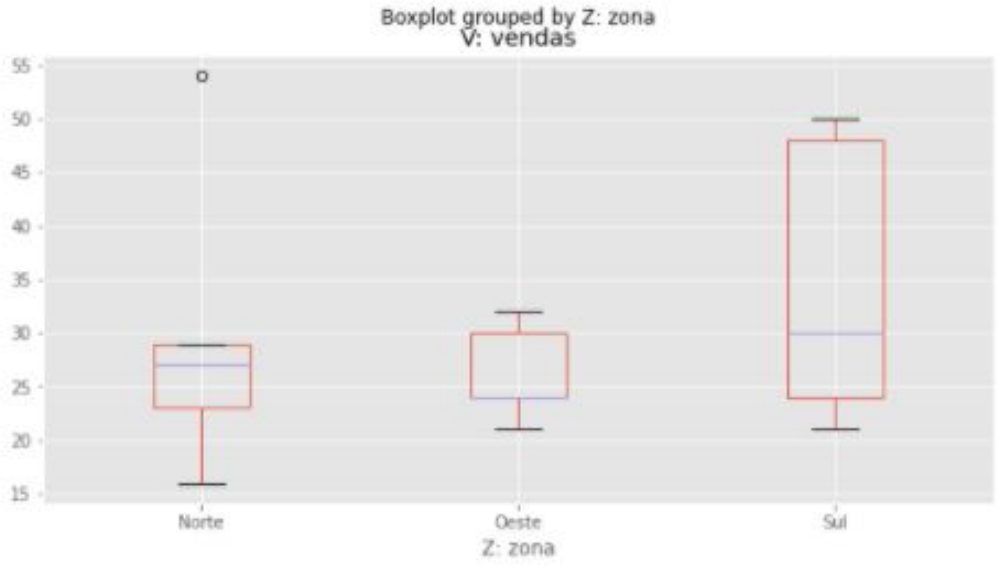
$$Q_1 = \frac{23 + 24}{2} = 23,5$$

Portanto, 23.5 é o volume mínimo de vendas.

3.1.4 Letra d)

Não é justo aplicar a essa forma de avaliação, conforme os *boxplots* abaixo, vemos que os dados de vendas por zonas não possuem um comportamento semelhante.

```
df2.boxplot(column='V: vendas',by='Z: zona', figsize=(10,5));
```



3.1.5 Letra e)

Para avaliarmos qual a variável mais importante, precisaremos analisar a correlação entre as variáveis, mas primeiro irei gerar a tabela de correlação entre essas variáveis:

```
df2[['T: teste','E: experiência','V: vendas']].corr()
```

	T: teste	E: experiência	V: vendas
T: teste	1.000000	0.010317	0.704746
E: experiência	0.010317	1.000000	0.263292
V: vendas	0.704746	0.263292	1.000000

Verificando a variável mais correlacionada com as vendas, vemos que teste é mais importante.

```
df2[['T: teste','E: experiência','V: vendas']].corr()['V: vendas'][:2]
```

```
T: teste          0.704746
E: experiência    0.263292
Name: V: vendas, dtype: float64
```

3.1.6 Letra f)

Tabela 1: valores esperados (e_i) e observados (o_i).

	Zona			
Conceito do gerente	Norte	Sul	Oeste	Total
Bom	4 (2,7)	3 (2,7)	1 (2,7)	8
Mau	1 (2,3)	2 (2,3)	4 (2,3)	7
Total	5 (5,0)	5 (5,0)	5 (5,0)	15

Para calcular o grau de associabilidade entre o conceito do gerente e a zona devemos calcular χ^2.

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i} = \left[\frac{(4-2.7)^2}{2.7} + \frac{(3-2.7)^2}{2.7} + \frac{(1-2.7)^2}{2.7} + \frac{(1-2.3)^2}{2.3} + \frac{(2-2.3)^2}{2.3} + \frac{(4-2.3)^2}{2.3}\right]$$

$$= \left[\frac{1.3^2}{2.7} + \frac{0.3^2}{2.7} + \frac{(-1.7)^2}{2.7} + \frac{(-1.3)^2}{2.3} + \frac{(-0.3)^2}{2.3} + \frac{1.7^2}{2.3}\right] =$$

$$= \left[\frac{1.69}{2.7} + \frac{0.09}{2.7} + \frac{2.89}{2.7} + \frac{1.69}{2.3} + \frac{0.09}{2.3} + \frac{2.89}{2.3}\right] \approx 3.76$$

Então, podemos concluir que existe um baixo grau de associabilidade.

3.1.7 Letra g)

4 Questão 4

Faça o gráfico $q \times q$ para os dois conjuntos de dados em A e B a seguir.

A	65	54	49	60	70	25	87	100	70	102	40	47
B	48	35	45	50	52	20	72	102	46	82	—	—

4.1 Resolução da questão 4

Criando as listas dos dados:

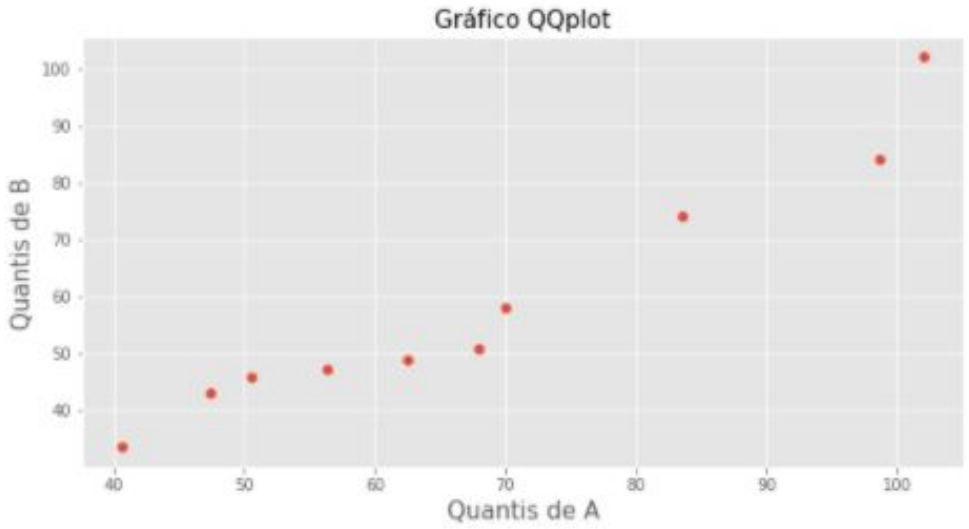
```
lista_a = [65, 54, 49, 60, 70, 25, 87, 100, 70, 102, 40, 47]
lista_b = [48, 35, 45, 50, 52, 20, 72, 102, 46, 82]
```

Criando a lista de quantis de cada lista:

```
x1 = np.quantile(lista_a, [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1])
x2 = np.quantile(lista_b, [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1])
```

Gerando o gráfico *Quantis* x *Quantis* :

```
plt.figure(figsize=(10,5))
plt.scatter(x1,x2)
plt.title('Gráfico QQplot', size=15)
plt.xlabel('Quantis de A', size=15)
plt.ylabel('Quantis de B',size=15)
plt.show()
plt.close()
```



5 Questão 5

A tabela a seguir reporta o número de linhas telefônicas por mil habitantes em cada estado do Brasil, em 2001.

Acre	183,8	Maranhão	86,1	Rio de Janeiro	347,5
Alagoas	125,4	M. Grosso	199,6	R. G. do Norte	150,1
Amapá	193,3	M. G. do Sul	235,3	R. G. do Sul	236,9
Amazonas	162	Minas Gerais	218,6	Rondônia	214,6
Bahia	142,3	Pará	128	Roraima	214,1
Ceará	140,6	Paraíba	125,4	Santa Catarina	257,3
D. Federal	456,8	Paraná	244,2	São Paulo	362,8
E. Santo	228,7	Pernambuco	147,8	Sergipe	140,7
Goiás	231,4	Piauí	118,2	Tocantins	113,8

Determine os seus quartis. Construa o *Box-plot*. Observe se existem alguns pontos discrepantes.

5.1 Resolução da questão 5

Criando uma lista com os dados:

```
estados = ['Acre','Alagoas','Amapá','Amazonas','Bahia','Ceará','D. Federal',
           'E. Santo','Goiás','Maranhão', 'M. Grosso','M. G. do Sul',
           'Minas Gerais','Pará','Paraíba','Paraná','Pernambuco','Piauí',
           'Rio de Janeiro','R. G. do Norte','R. G. do Sul','Rondônia',
           'Roraima','Santa Catarina','São Paulo','Sergipe','Tocantins']

linhas = [183.8,125.4,193.3,162,142.3,140.6,456.8,228.7,231.4,86.1,199.6,235.3,
          218.6,128,125.4,244.2, 147.8,118.2,347.5,150.1,236.9,214.6,214.1,
          257.3,362.8,140.7,113.8]
```

Gerando a base de dados com as listas:

```
df5 = pd.DataFrame(estados, columns=['Estados'])
df5['linhas'] = linhas
df5.head()
```

	Estados	linhas
0	Acre	183.8
1	Alagoas	125.4
2	Amapá	193.3
3	Amazonas	162.0
4	Bahia	142.3

Calculando os quartis 1, 2 e 3.

```
df5.describe().T[['25%','50%','75%']]
```

	25%	50%	75%
linhas	140.65	193.3	233.35

No *Boxplot* dos dados, podemos ver que há **um** valor discrepante (*outlier*) nessa base de dados.

```
plt.figure(figsize=(12,3))
sns.boxplot(x='linhas', data=df5)
plt.xlabel('Número de linhas telefônicas por mil habitantes', size=15)
plt.title('Boxplot da variável de linhas telefônicas', size=15);
```