LADE

LABORATÓRIO DE ANÁLISES DE DADOS EDUCACIONAIS E ESTATÍSTICA APLICADA

IFCE - CAMPUS FORTALEZA

Estatística Descritiva com Python

Escore Z, Quartis, percentis e decis, Medidas de Curtose , Assimetria

DEFINIÇÃO

O **escore padronizado**, ou **escore** z, é o número de desvios-padrão pelo qual um valor x dista da média (para mais ou para menos). Obtém-se como segue:

Amostra		População	
$z = \frac{x - \bar{x}}{s}$	ou	$z = \frac{x - \mu}{\sigma}$	

(Arredondar z para duas decimais.)

EXEMPLO As alturas da população de homens adultos têm média $\mu = 69{,}0$ in., desvio-padrão $\sigma = 2{,}8$ in. e distribuição em forma de sino. O jogador de basquete Michael Jordan ganhou reputação de gigante por suas proezas no jogo, mas com 78 in., ele pode ser considerado excepcionalmente alto, comparado com a população geral de homens adultos? Determine o escore z para a altura de 78 in.

SOLUÇÃO Como estamos lidando com parâmetros populacionais, o escore z se calcula como segue:

$$z = \frac{x - \mu}{\sigma} = \frac{78 - 69{,}0}{2{,}8} = 3{,}21$$

Podemos interpretar este resultado dizendo que a altura de Michael Jordan, de 78 in., está 3,21 desvios-padrão acima da média.

A importância dos escores z na estatística reside no fato de que eles permitem distinguir entre valores usuais e valores raros, ou incomuns. Consideramos usuais os valores cujos escores padronizados estão entre $-2{,}00$ e 2,00, e incomuns os valores com escore z inferior a $-2{,}00$ ou superior a 2,00. (Veja Figura 2-11.) A altura de Michael Jordan corresponde a um escore z de 3,21, que consideramos incomum, por ser superior a 2,00. Em comparação com a população geral, Jordan é excepcionalmente alto.

Nosso critério para classificar um escore z como incomum decorre da regra empírica e do teorema de Tchebichev. Recorde que, pela regra empírica, para dados com distribuição em forma de sino, cerca de 95% dos valores estão a menos de 2 desvios-padrão da média. (Veja Figura 2-10 da seção precedente.) Por outro lado, o teorema de Tchebichev afirma que, para qualquer conjunto de dados, ao menos 75% dos valores estão dentro de 2 desvios-padrão a contar da média.

Já vimos que os escores z são úteis para comparar escores de diferentes populações com médias distintas e desvios-padrão diferentes. O exemplo que segue ilustra essa aplicação dos escores z.

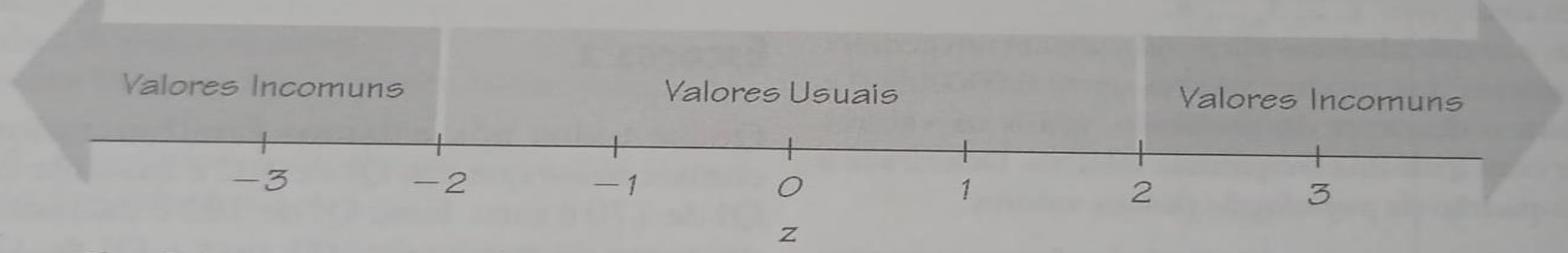


Fig. 2-11 Interpretação do escore z.
Valores com escores z inferiores a $z = 22{,}00$ ou superiores a $z = 2{,}00$ são considerados incomuns.

EXEMPLO Uma professora de estatística aplica dois testes diferentes a duas turmas do seu curso. Os resultados foram

Turma 1: $\bar{x} = 75$ e $s = 14$
Turma 2: $\bar{x} = 40$ e $s = 8$

Que nota é relativamente melhor: 82 no teste da Turma 1, ou 46 no da Turma 2?

SOLUÇÃO Não podemos comparar diretamente as notas 82 e 46 porque provêm de escalas diferentes. Transformamo-las, portanto, em escores z. Para o valor 82 da Turma 1, obtemos o escore z 0,50, porque

$$z = \frac{x - \bar{x}}{s} = \frac{82 - 75}{14} = 0{,}50$$

Para a nota 46 da Turma 2, o escore z correspondente é 0,75, porque

$$z = \frac{x - \bar{x}}{s} = \frac{46 - 40}{8} = 0{,}75$$

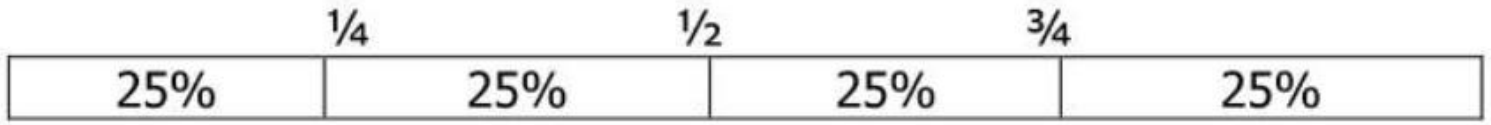
Isso significa que a nota 82 do teste da Turma 1 está 0,5 desvio-padrão acima da média, enquanto a nota 46 do teste da Turma 2 está 0,75 desvio-padrão acima da média. Isso implica que o resultado 46 do teste da Turma 2 é melhor, relativamente. Embora inferior a 82, a nota 46 tem melhor posição relativa no contexto dos outros resultados do teste. Mais adiante vamos utilizar amplamente os escores z.

Medidas Separatrizes

- Medidas que dividem a distribuição em partes iguais
- Servem para descrever posições numa distribuição de dados

Quartil

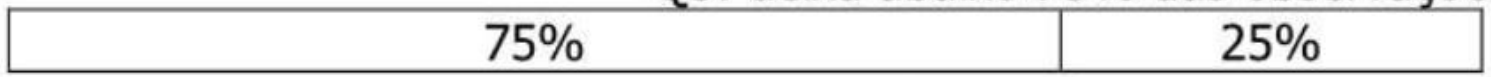
Valores da variável que dividem a distribuição em quatro partes iguais.



Q1: deixa abaixo 25% das observações

Q2: deixa abaixo 50% das observações

Q3: deixa abaixo 75% das observações

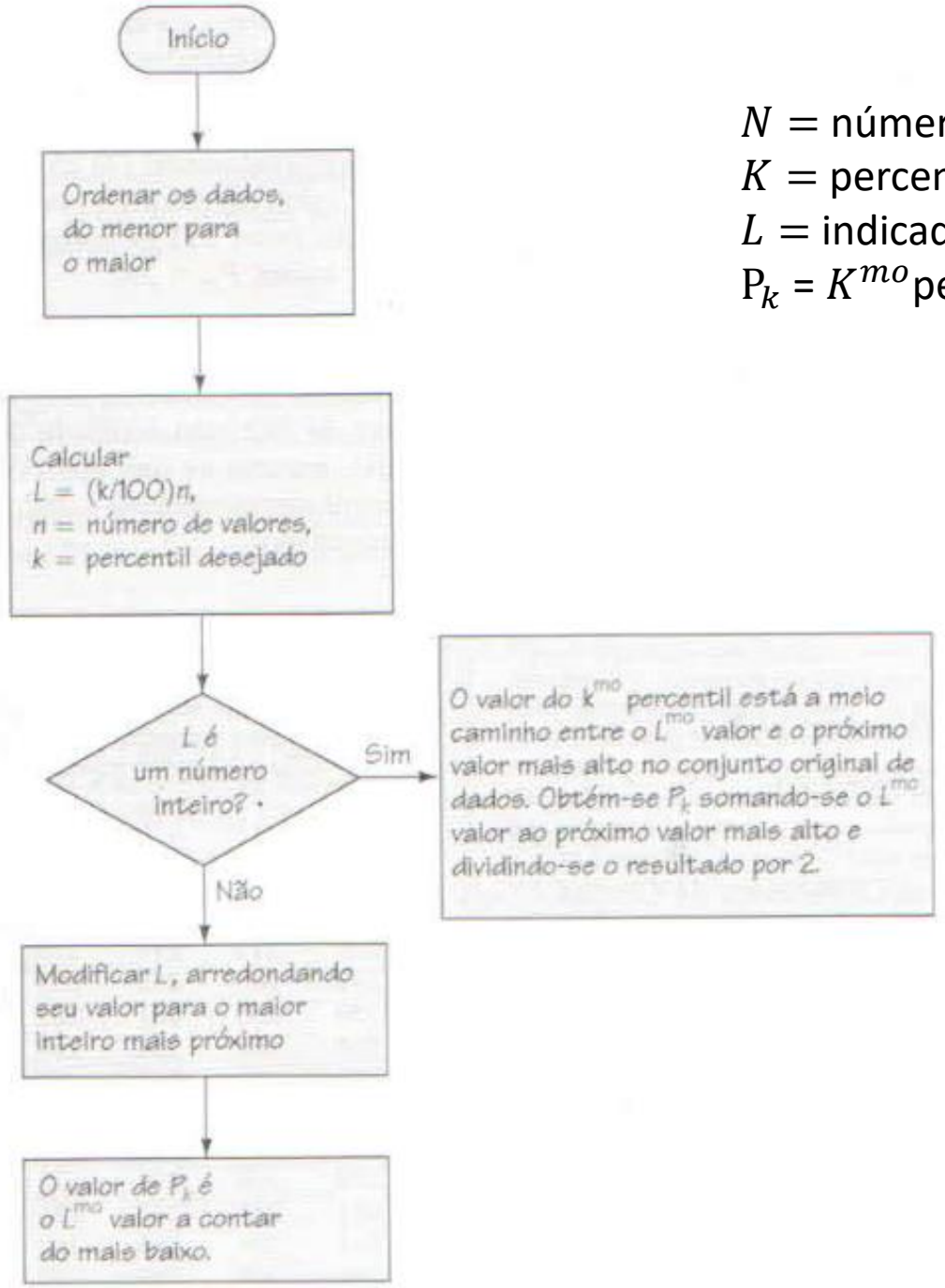


Valor Correspondentes determinado percentil

$$L = (\frac{k}{100}) \times n$$

Percentil do valor

$$x = \frac{Número\ de\ valores\ Inferiores\ a\ x}{Número\ total\ de\ valores} \times 100$$



N = números de escores, ou valor, no conjunto de dados
K = percentil a ser utilizado
L = indicador que dá a posição de um escore
P_k = K^{mo} percentil

Valor Correspondentes determinado percentil

$$L = (\frac{k}{100}) \times n$$

Percentil do valor

$$x = \frac{Número\ de\ valores\ Inferiores\ a\ x}{Número\ total\ de\ valores} \times 100$$

EXEMPLO A Tabela 2-9 relaciona as 175 cargas axiais das latas de alumínio, ordenadas da mais baixa até a mais elevada. Determine o percentil correspondente a 241.

SOLUÇÃO Pela Tabela 2-9, vemos que há 21 valores inferiores a 241, de forma que

$$\text{percentil de } 241 = \frac{21}{175} \cdot 100 = 12$$

A carga axial de 241 é o 12.º percentil.

Valor Correspondentes determinado percentil

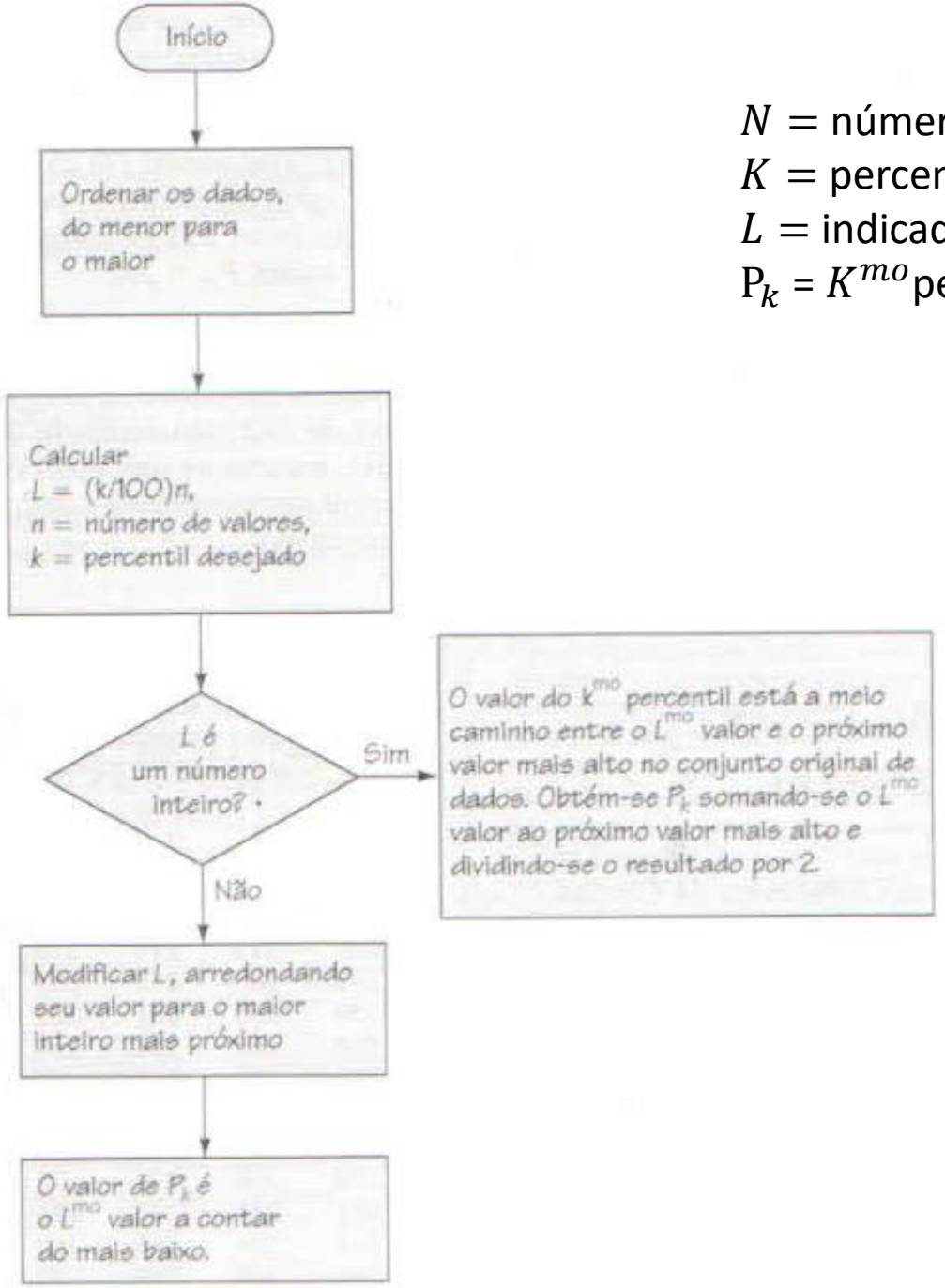
$$L = \left(\frac{k}{100}\right) \times n$$

Percentil do valor

$$x = \frac{\text{Número de valores Inferiores a } x}{\text{Número total de valores}} \times 100$$

TABELA 2-9 Valores Ordenados de Cargas Axiais de Latas de Alumínio

200	201	204	204	206	206	208	208	209	215	217	218	220	223	223
225	228	230	230	234	236	241	242	242	248	250	251	251	252	252
254	256	256	256	257	257	258	259	259	260	261	262	262	262	262
262	263	263	263	263	263	264	265	265	265	266	267	267	268	268
268	268	268	268	268	268	268	269	269	269	269	270	270	270	270
270	270	270	270	271	271	272	272	272	272	272	273	273	273	273
273	273	274	274	274	274	275	275	275	275	276	276	276	276	276
277	277	277	277	277	277	277	277	278	278	278	278	278	278	278
279	279	279	280	280	280	281	281	281	281	282	282	282	282	282
282	283	283	283	283	283	283	284	284	284	284	285	285	285	286
286	286	286	287	287	288	289	289	289	289	289	290	290	290	291
291	292	292	292	293	293	294	295	295	297					



N = números de escores, ou valor, no conjunto de dados
K = percentil a ser utilizado
L = indicador que dá a posição de um escore
P_k = K^{mo} percentil

Valor Correspondentes determinado percentil

$$L = (\frac{k}{100}) \times n$$

Percentil do valor

$$x = \frac{Número\ de\ valores\ Inferiores\ a\ x}{Número\ total\ de\ valores} \times 100$$

EXEMPLO Para as 175 cargas axiais de latas de alumínio da Tabela 2-9, determine o escore correspondente ao 25.° percentil; ou seja, determine o valor de P_{25}.

SOLUÇÃO Recorremos à Figura 2-12 e observamos que os dados já estão ordenados, do menor para o maior. Calculamos a seguir o indicador L como segue:

$$L = \left(\frac{k}{100}\right)n = \left(\frac{25}{100}\right) \cdot 175 = 43{,}75$$

Respondemos *não* à pergunta na Figura 2-12, se 43,75 é um número inteiro, e somos orientados a arredondar *L para cima*, ou seja, arredondar para 44. (Nesse processo em particular arredondamos L para o inteiro superior mais próximo, mas na maior parte das situações neste livro seguimos o processo geral de arredondamento.) O 25.° percentil, denotado por P_{25}, é o 44.° valor, ou escore, a contar do menor. Partindo, pois, do menor valor, 200, percorremos a lista até o 44.° valor, que é 262; assim, $P_{25} = 262$.

Valor Correspondentes determinado percentil

$$L = (\frac{k}{100}) \times n$$

Percentil do valor

$$x = \frac{Número\ de\ valores\ Inferiores\ a\ x}{Número\ total\ de\ valores} \times 100$$

LADE

Suponha agora que queiramos achar o percentil correspondente a um escore de 262. Verificamos que há 41 valores abaixo de 262, não deixando de considerar cada valor individual, mesmo os que aparecem repetidos. Calculando o percentil correspondente a 262, obtemos $(41/175) \cdot 100 = 23$ (arredondado).

Há aqui uma pequena discrepância: no exemplo precedente encontramos 262 para o 25.º percentil, mas no processo inverso, 262 corresponde ao 23.º percentil. À medida que aumenta o número de dados, tais discrepâncias diminuem. Poderíamos eliminá-las utilizando um processo mais complicado, que aplica a interpolação em lugar do arredondamento.

Em razão do tamanho da amostra no exemplo precedente, o indicador L calculado foi inicialmente 43,75, valor que foi arredondado para 44, porque o valor original de L não era inteiro. No próximo exemplo ilustramos um caso em que o valor original de L é um número inteiro. Essa condição nos levará para o ramo direito no fluxograma da Figura 2-12.

Valor Correspondentes determinado percentil

$$L = (\frac{k}{100}) \times n$$

Percentil do valor

$$x = \frac{Número\ de\ valores\ Inferiores\ a\ x}{Número\ total\ de\ valores} \times 100$$

EXEMPLO Determine o 40.° percentil P_{40} das cargas axiais da Tabela 2-9.

SOLUÇÃO Seguindo o processo delineado na Figura 2-12 e notando que os dados já estão ordenados do menor para o maior, calculamos

$$L = \left(\frac{k}{100}\right)n = \left(\frac{40}{100}\right) \cdot 175 = 70 \quad \text{(exatamente)}$$

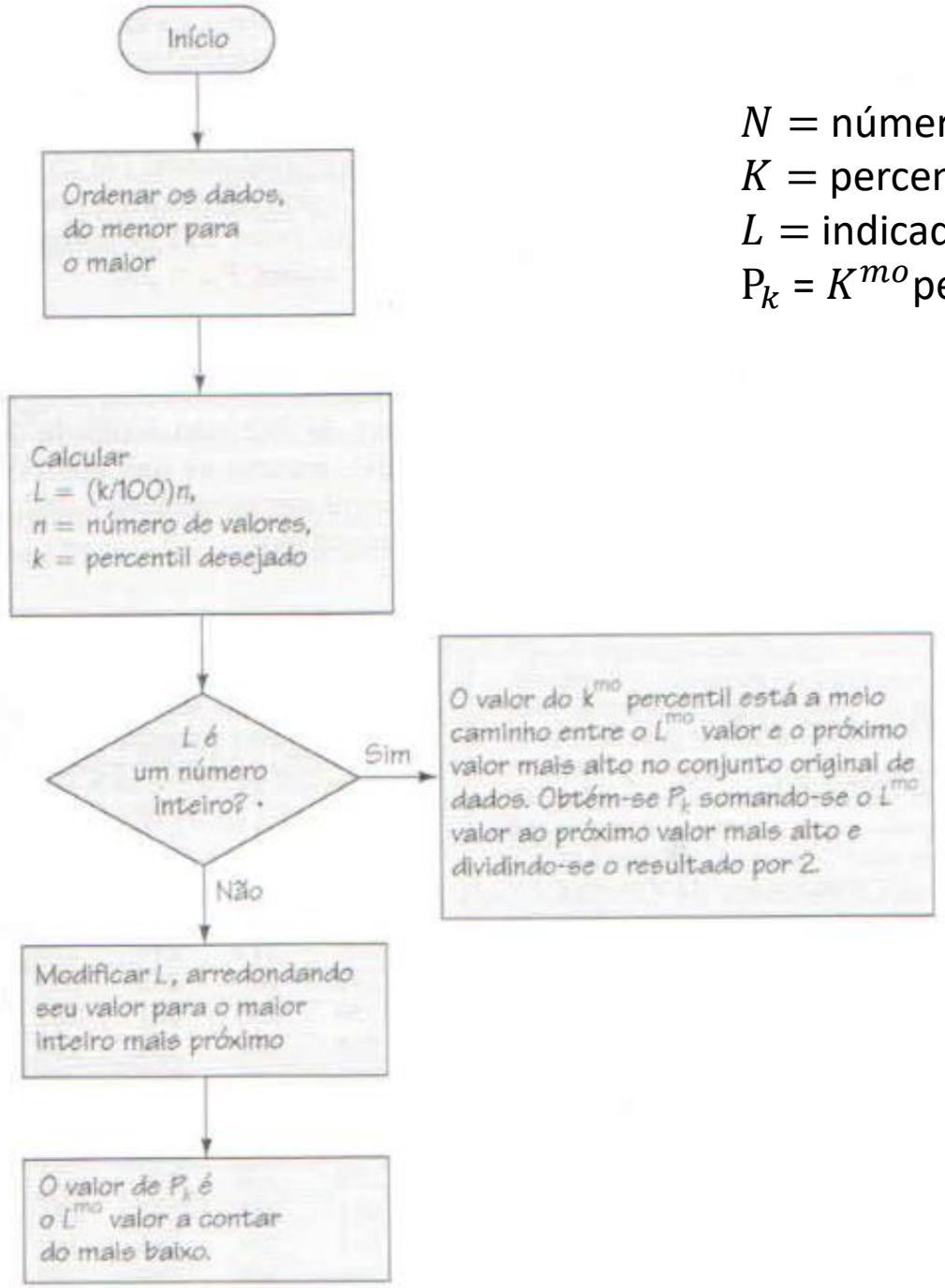
70 é um número inteiro, e a Figura 2-12 indica que P_{40} está a meio caminho entre os 70.° e 71.° valores. E como esses valores são ambos 269, concluímos que o 40.° percentil é 269.

Valor Correspondentes determinado percentil

$$L = (\frac{k}{100}) \times n$$

Percentil do valor

$$x = \frac{Número\ de\ valores\ Inferiores\ a\ x}{Número\ total\ de\ valores} \times 100$$



N = números de escores, ou valor, no conjunto de dados
K = percentil a ser utilizado
L = indicador que dá a posição de um escore
P_k = K^{mo} percentil

Valor Correspondentes determinado percentil

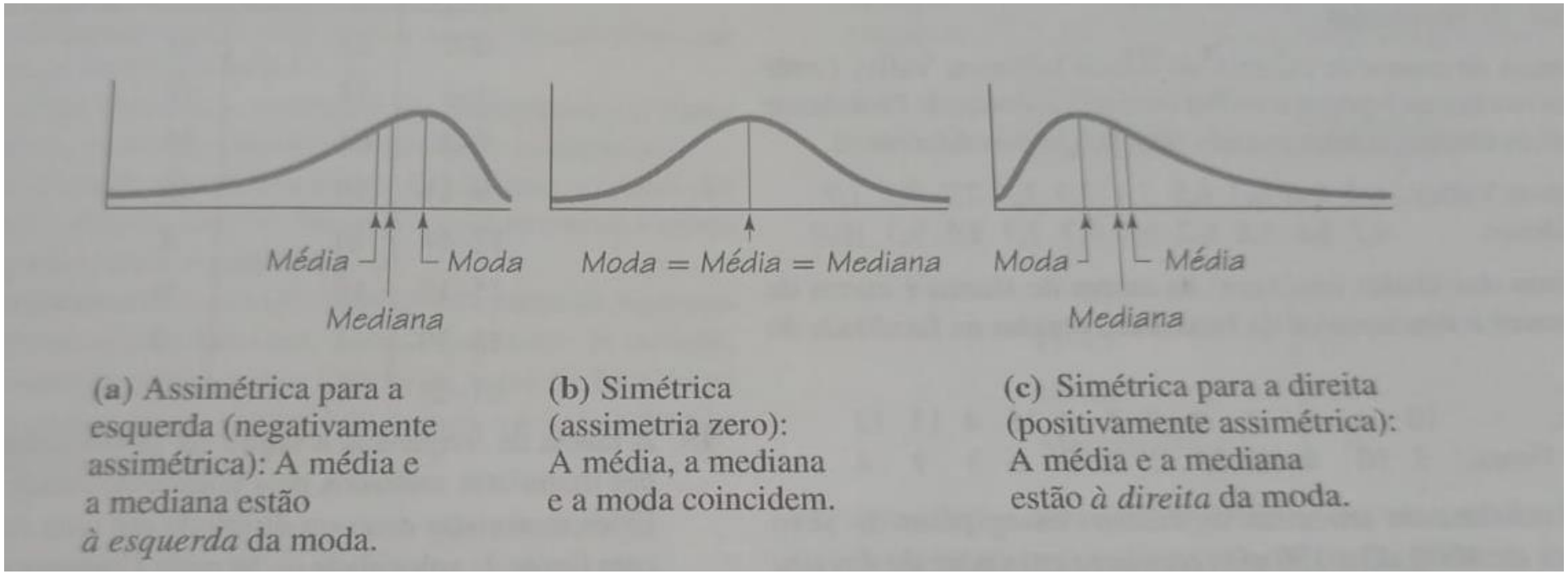
$$L = (\frac{k}{100}) \times n$$

Percentil do valor

$$x = \frac{Número\ de\ valores\ Inferiores\ a\ x}{Número\ total\ de\ valores} \times 100$$

Medidas Assimetria

Possibilitam analisar uma distribuição em relação a sua moda, mediana e média.



Primeiro Coeficiente de Pearson:

$$AS = \frac{Média - Moda}{Desvio\ Padrão}$$

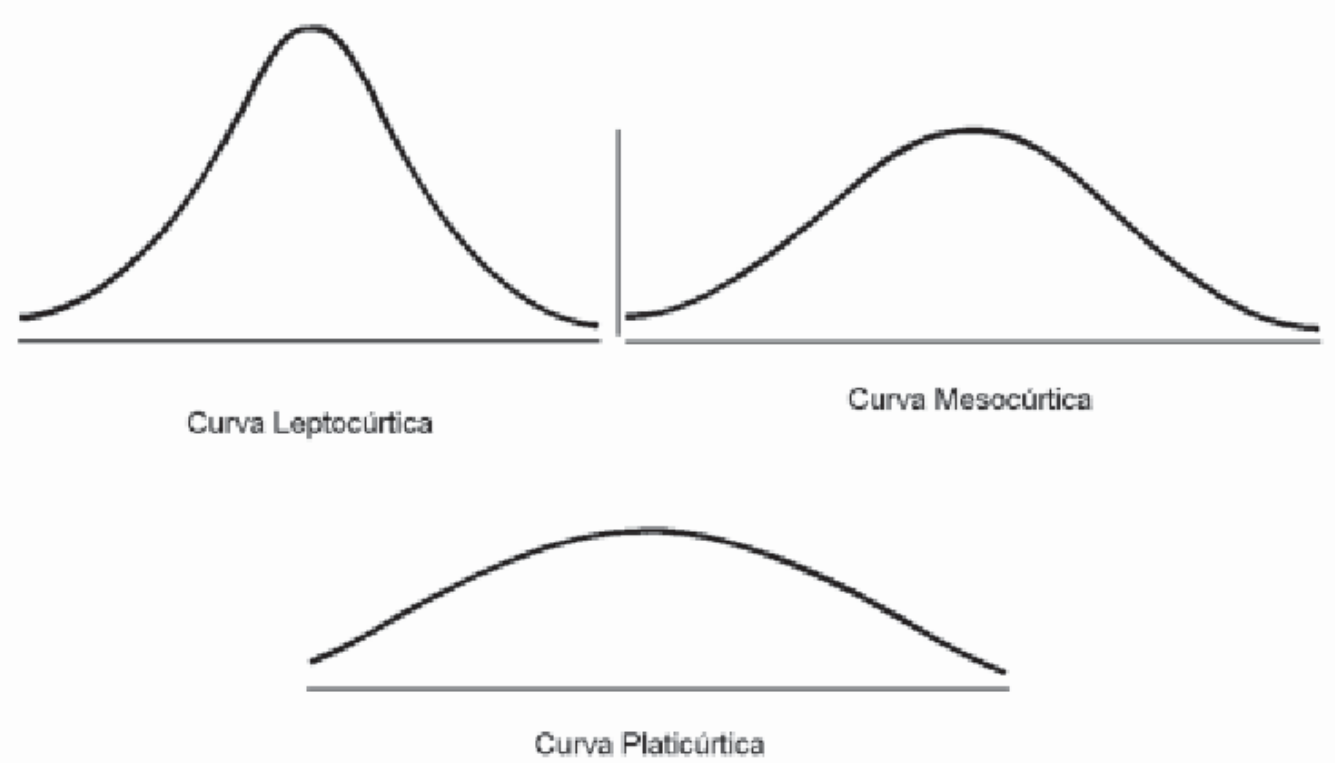
AS= 0 , Média=Moda, Distribuição Simétrica

AS<0 , Média<Moda, Distribuição assimétrica à esquerda ou negativa

AS>0, Média>Moda, Distribuição assimétrica à direita ou positiva

Coeficiente de Curtose

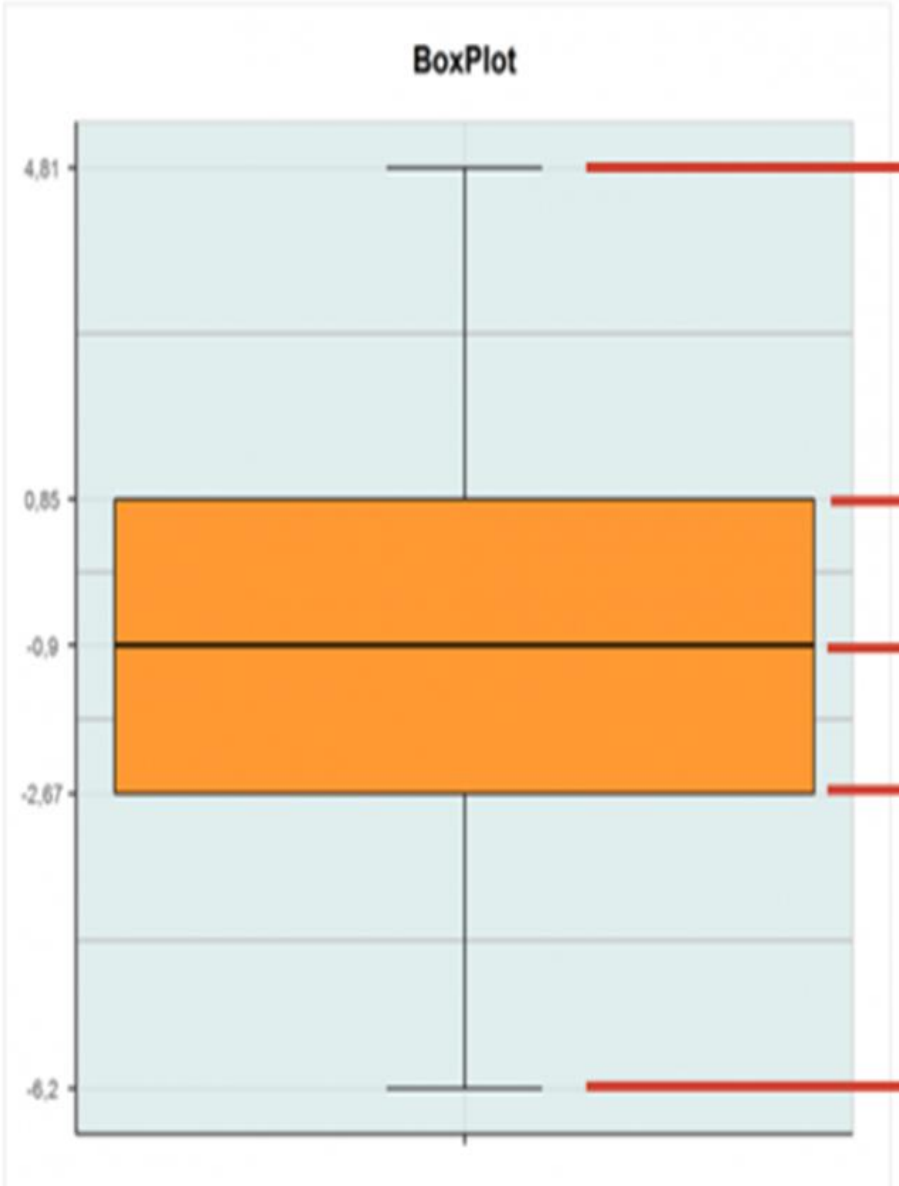
Mesocúrtica - quando a distribuição é normal.
Leptocúrtica - quando a distribuição é mais pontiaguda que a normal
Platicúrtica - quando a distribuição é mais achatada que a normal.



$$C = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}$$

Se C = 0,263 a distribuição é Mesocúrtica
Se C < 0,263 a distribuição é Leptocúrtica
Se C > 0,263 a distribuição é Platicúrtica

Box Plot



Limite Superior

Terceiro Quartil Q_3

Segundo Quartil ou Mediana Q_2

Primeiro Quartil Q_1

Limite Inferior

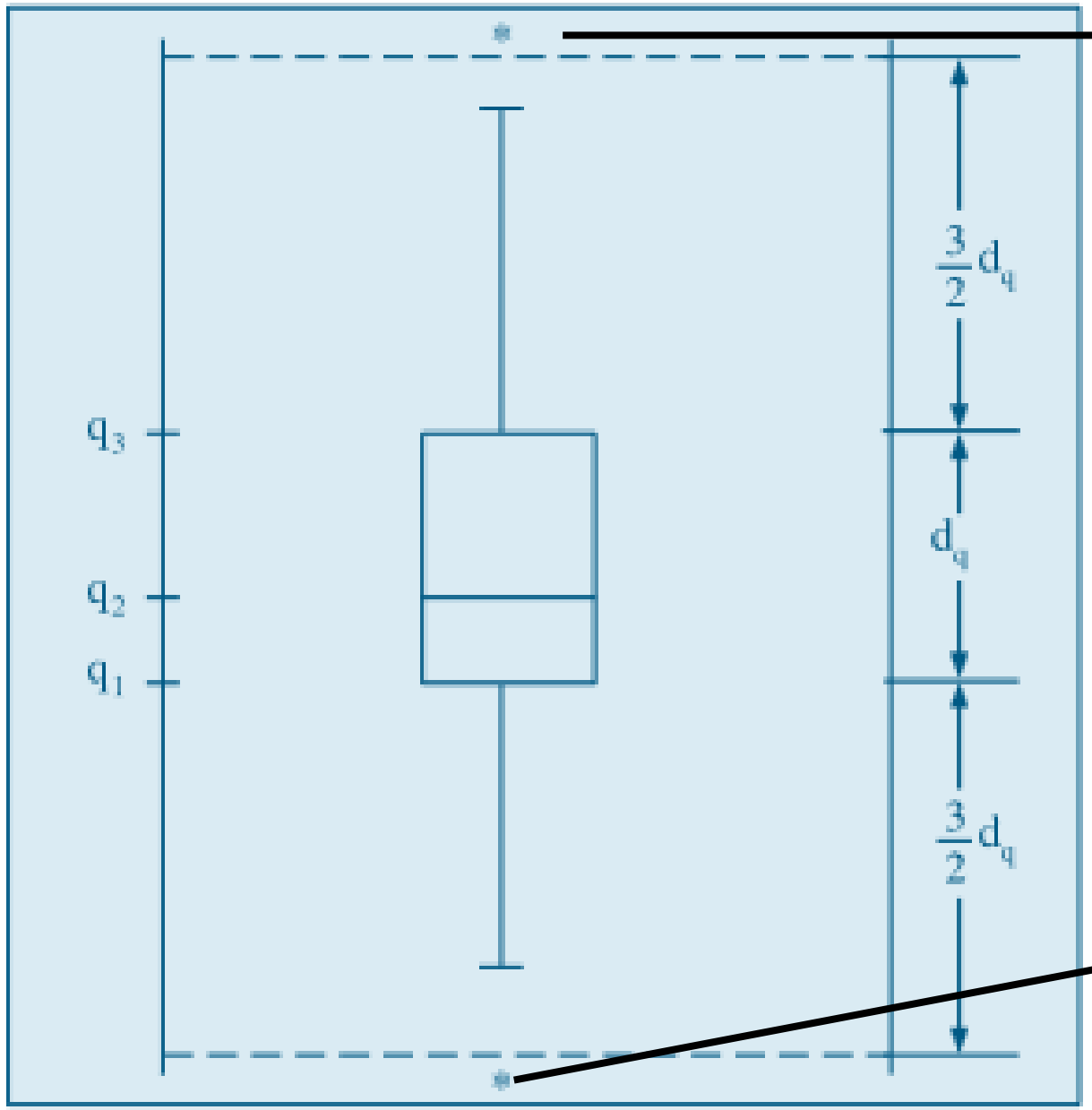
Limite Inferior = $q_1 - 1,5d_q$

Limite Superior = $q_3 + 1,5d_q$

Distância Interquartil

$$d_q = q_3 - q_1$$

LADE



Pontos exteriores

Outliers ou valores atípicos

Pontos exteriores

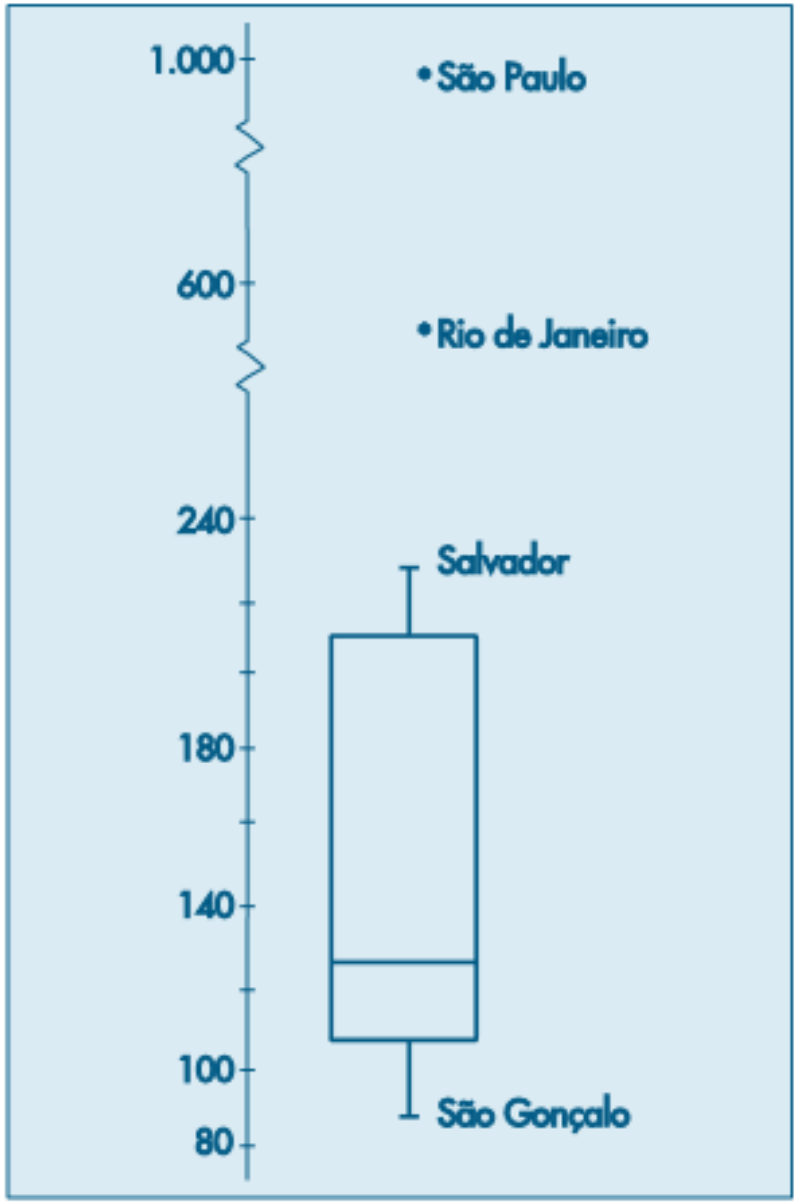
O box plot dá uma idéia da posição, dispersão, assimetria, caudas e dados discrepantes. A posição central é dada pela mediana e a dispersão por d_q.

As posições relativas de $\mathbf{q}_1$, q_2, q_3 dão uma noção da assimetria da distribuição.

Os comprimentos das caudas são dados pelas linhas que vão do retângulo aos valores remotos e pelos valores atípicos

LADE

Box plot para os quinze maiores municípios do Brasil.



EXERCÍCIO

1- Os homens adultos (nos EUA) têm altura média de 69,0 polegadas, com desvio-padrão de 2, 8 polegadas. Determine os escores **z** correspondentes a:

A- O jogador de basquete Mugsy Bogues que tem 5 pés e 3 in.
b- O jogador de basquete Shaquille O'Neal , que tem 7 p é s e 1 polegada.
c- O autor, que é um jogador de golfe e tênis com 69.72 in.

2- Os carros dos estudantes na faculdade do autor têm idade média de 7,90 anos, com desvio-padrão de 3,67 anos. Determine o escores **z** para os carros com as seguintes idades:

a- um Corvette de 12 ano
b- U m a Ferrari de 2 anos
c- Um Porscbe novo

3- Qual dos dois escores abaixo acusa melhor posição relativa?
a- Um escore de 60 em um teste com x = 50 e s = 5.
b- Um escore de 250 em um reste com x = 200 e s = 20.

4- Utilizar as 175 cargas axiais ordenadas. Ache o percentil correspondente ao valor dado.

a- 254 **b-** 265 **c-** 277 **d-** 288

LADE

Usando o Python

Box

```
# Escore Z, Quartis, percentis e decis,  Medidas de Curtose , Assimetria e Bos Plot
EscoreZRenda=(PesquisaNacional['Renda']-PesquisaNacional['Renda'].mean())/PesquisaNacional['Renda'].std()
```

```
EscoreZRenda
```

```
0       -0.361193
1       -0.255878
2       -0.337121
3        0.451231
4       -0.556776
           ...
76835   -0.357582
76836   -0.150564
76837   -0.210744
76838   -0.150564
76839   -0.331103
Name: Renda, Length: 76840, dtype: float64
```

```
PesquisaNacional['RendaPadronizada']=EscoreZRenda
```

```
z=(800-2000.38)/3323.38
z
```

```
-0.3611925208673098
```

```
PesquisaNacional
# z=(800-2000.38)/3323.38
```

	UF	Sexo	Idade	Cor	Anos de Estudo	Renda	Altura	Intervalo de Classe	RendaPadronizada
0	Rondônia	Masculino	23	Parda	11 anos	800	1.603808	(22, 32]	-0.361193
1	Rondônia	Feminino	23	Branca	11 anos	1150	1.739790	(22, 32]	-0.255878
2	Rondônia	Feminino	35	Parda	14 anos	880	1.760444	(32, 42]	-0.337121
3	Rondônia	Masculino	46	Branca	5 anos	3500	1.783158	(42, 52]	0.451231
4	Rondônia	Feminino	47	Parda	8 anos	150	1.690631	(42, 52]	-0.556776
...	...	...	...	...	...	...	...	...	...
76835	Distrito Federal	Feminino	46	Branca	10 anos	812	1.687030	(42, 52]	-0.357582
76836	Distrito Federal	Masculino	30	Preta	6 anos	1500	1.792934	(22, 32]	-0.150564
76837	Distrito Federal	Masculino	32	Parda	11 anos	1300	1.830587	(22, 32]	-0.210744
76838	Distrito Federal	Masculino	57	Parda	3 anos	1500	1.726344	(52, 62]	-0.150564
76839	Distrito Federal	Masculino	38	Parda	3 anos	900	1.658305	(32, 42]	-0.331103

LADE

```python
# Quartis
Q1=np.quantile(PesquisaNacional['Renda'],0.25)
Q2=np.quantile(PesquisaNacional['Renda'],0.50)
Q3=np.quantile(PesquisaNacional['Renda'],0.75)
print(Q1,Q2,Q3)
```

```
788.0 1200.0 2000.0
```

```python
PesquisaNacional['Renda'].median()
```

```
1200.0
```

Primeiro Coeficiente de Pearson:

$$AS = \frac{Média - Moda}{Desvio\ Padrão}$$

```
#Medidas de Curtose , Assimetria e Box Plot
AS=(PesquisaNacional['Renda'].mean()-PesquisaNacional['Renda'].mode())/PesquisaNacional['Renda'].std()
AS
```

```
0    0.364803
```

$$C = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}$$

```
Q1=np.quantile(PesquisaNacional['Renda'],0.25)
Q3=np.quantile(PesquisaNacional['Renda'],0.75)
P10=np.quantile(PesquisaNacional['Renda'],0.10)
P90=np.quantile(PesquisaNacional['Renda'],0.90)
C=(Q3-Q1)/(P90-P10)
C
```

```
0.3320547945205479
```

```
PesquisaNacional['Renda'].describe()
```

```
count      76840.000000
mean        2000.383199
std         3323.387730
min            0.000000
25%          788.000000
50%         1200.000000
75%         2000.000000
max       200000.000000
Name: Renda, dtype: float64
```

```
# Quartis
Q1=np.quantile(PesquisaNacional['Renda'],0.25)
Q2=np.quantile(PesquisaNacional['Renda'],0.50)
Q3=np.quantile(PesquisaNacional['Renda'],0.75)
print(Q1,Q2,Q3)
```

```
788.0 1200.0 2000.0
```

LADE

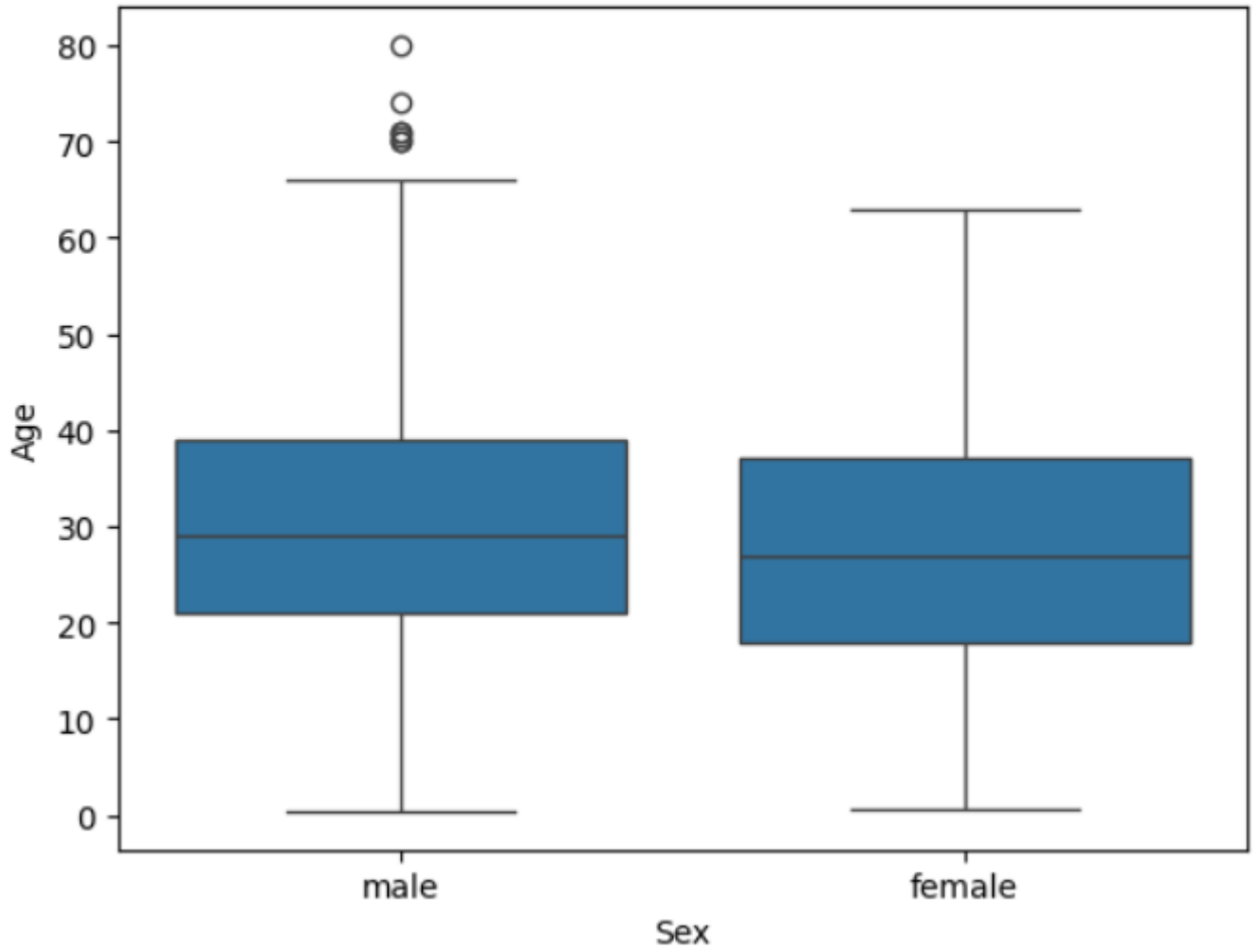
```
Titanic
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

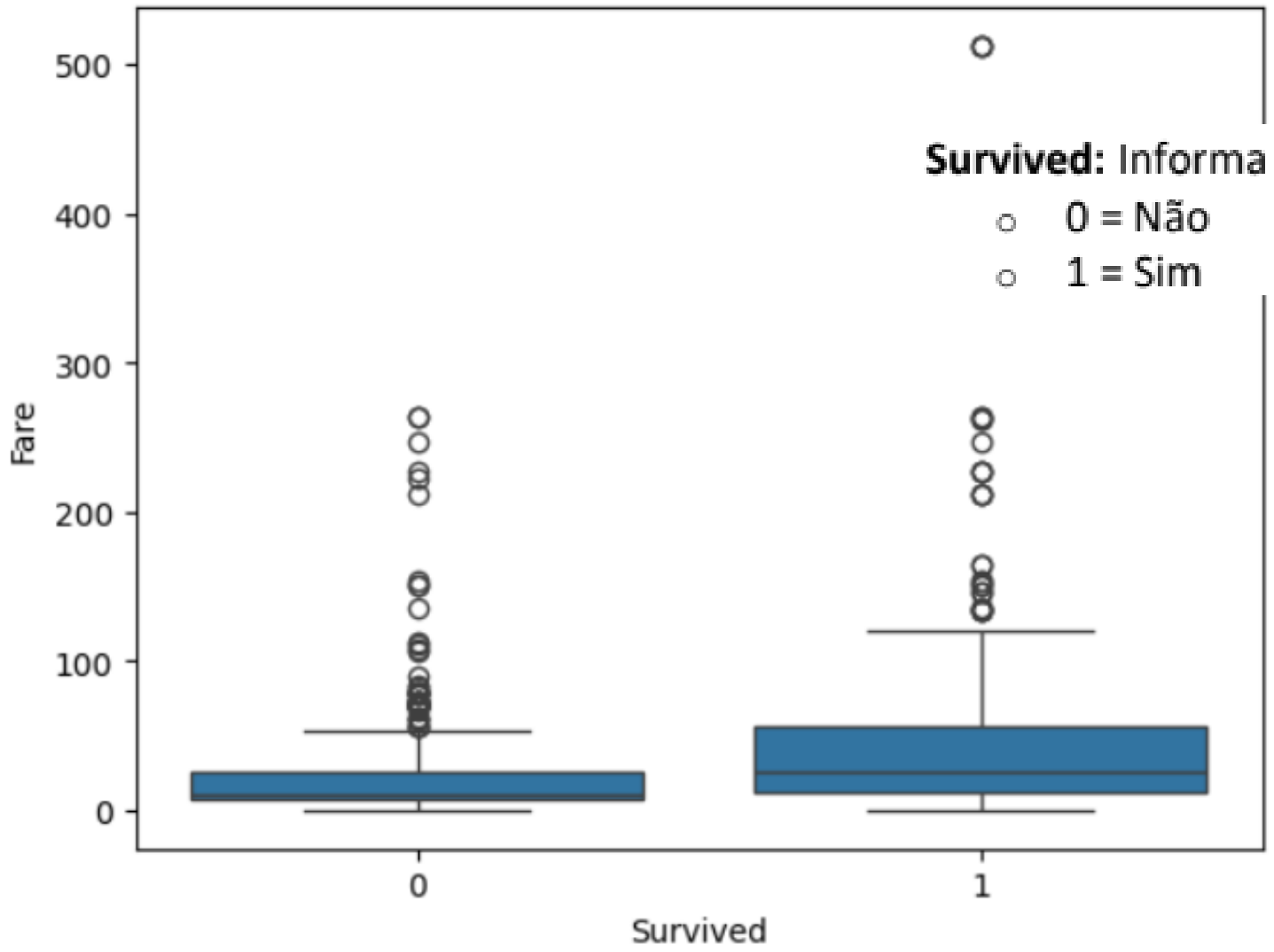
```
sns.boxplot(data=Titanic, x="Sex", y="Age")
```

```
<Axes: xlabel='Sex', ylabel='Age'>
```



```
sns.boxplot(data=Titanic, x="Survived", y="Fare")
```

```
<Axes: xlabel='Survived', ylabel='Fare'>
```



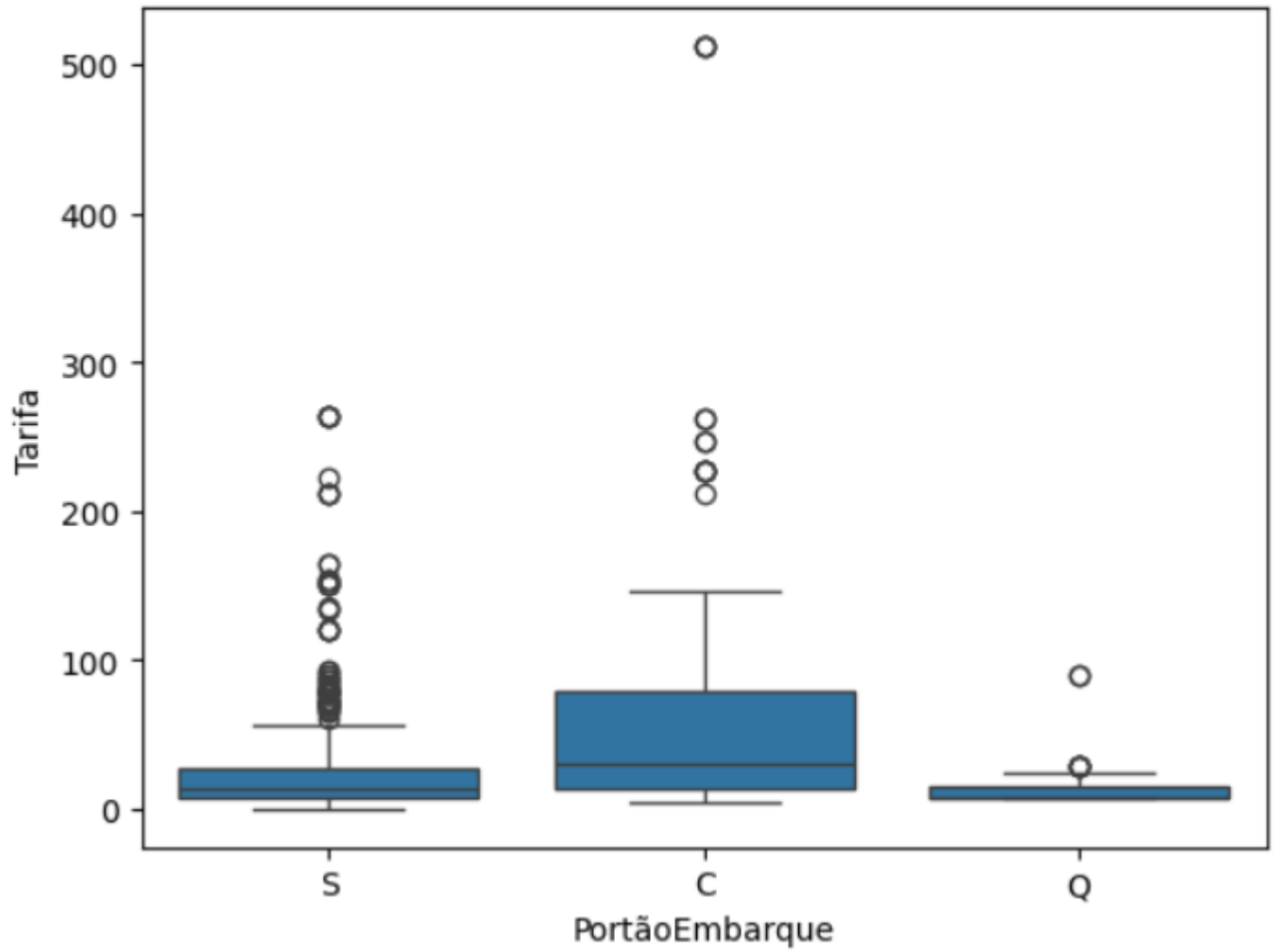
Survived: Informa se o passageiro sobreviveu ao desastre

- 0 = Não
- 1 = Sim

LADE

```
import seaborn as sns
sns.boxplot(data=data1, x="PortãoEmbarque",y="Tarifa")
```

```
<Axes: xlabel='PortãoEmbarque', ylabel='Tarifa'>
```



EXERCÍCIO

1. Em "Ages of Oscar - Winning Best Actors and Actress es " na revisar : Mathematics Teacher , por Richard Brown e Gretchen Davis , utilizam-se diagramas em caixas , ou box plots , para comparar idades dos atores e das atrizes na ocasião em que receberam o Oscar . Relacionam-se adiante os 34 vencedores recentes de cada categoria. Compare s dois conjuntos de dados com auxilio de um diagrama em caixas .

Atores : 32 37 36 32 51 53 33 61 35 45 55 39 76 37 42 40 32 60 38 56 48 48 40 43 62 43 42 44 41 56 39 46 31 47

Atrizes : 50 44 35 80 26 28 41 21 61 38 49 33 74 30 33 41 31 35 41 42 37 26 34 34 35 26 61 60 34 24 30 37 31 27

2- Utilizando a base de dados "Companhia_MB.xlsx", construir um box plot comparativo para a variável salário, utilizando as categorias do estado civil.

LADE

LABORATÓRIO DE ANÁLISES DE DADOS EDUCACIONAIS E ESTATÍSTICA APLICADA

IFCE - CAMPUS FORTALEZA

OBRIGADO!!!