Matemática Computacional

Capítulo 2

Estatística descritiva

Licenciatura em Engenharia Informática
ISEP
(2024/2025)

Conteúdo

A Estatística descritiva consiste num conjunto de métodos cujo objetivo é organizar, analisar, sintetizar e representar de forma compreensível a informação obtida dos dados.

Caracterização, organização e representação dos dados

- No processo de análise estatística, o objeto de estudo são as variáveis e a informação que estas podem fornecer.
- Variável é uma característica da população que pode tomar vários valores possíveis.
  - Quanto à característica da população que representa, uma variável pode ser:
    - Qualitativa, indicando uma característica não numérica.
    - Quantitativa, indicando uma característica numérica.
  - Quanto ao tipo de valores que pode tomar, uma variável classifica-se em:
    - Discreta, pode tomar apenas um conjunto finito ou infinito numerável de valores.
    - Contínua, toma valores num intervalo real.
  - As variáveis qualitativas são variáveis discretas.
  - As variávis quantitativas podem ser discretas ou contínuas.

- Os valores das variáveis qualitativas pertencem a categorias (ou classes), exaustivas e mutuamente exclusivas. Estas variáveis podem ser medidas numa escala:
  - Nominal: categorias, estados ou ”nomes de coisas”.
    Por exemplo:
    - CorCabelo= {preto, loiro, castanho, cinzento, branco};
    - EstadoCivil, Profissão, Morada.
  - Ordinal: valores sujeitos a uma ordem (*ranking*), segundo uma relação descritível mas não quantificável.
    Por exemplo:
    - Tamanho= {pequeno, médio, grande};
    - Graus académicos, patentes militares;
    - Escalas de *Likert*: 1− muito insatisfeito, 2− insatisfeito, 3− nem insatisfeito nem satisfeito, 4− satisfeito e 5− muito satisfeito.

- A Categoria (ou classe) de uma variável qualitativa é cada um dos valores que a variável pode tomar.

- A Classe de uma variável quantitativa discreta é o valor numérico que a variável pode tomar.

Dados qualitativos e quantitativos discretos

Frequência absoluta da categoria i (n_i)

É o número de observações associadas à categoria i. Verifica-se

$$\sum_{i=1}^{c} n_i = n.$$

sendo n o número total de observações e c o número de categorias.

Frequência relativa da categoria i (f_i)

É o quociente entre a frequência absoluta dessa categoria e o número total de observações efetuadas,

$$f_i = \frac{n_i}{n}.$$

Dados qualitativos e quantitativos discretos

Distribuição de frequência absoluta (ou relativa)

É um arranjo tabular ou uma representação gráfica dos dados que mostra, para cada categoria, a sua frequência absoluta (ou relativa) observada.

Gráfico de barras

É uma representação dos dados em que se usam barras separadas (de igual largura) cuja altura é proporcional à frequência (absoluta e relativa) da categoria correspondente.

Gráfico circular

É uma representação dos dados num círculo dividido em sectores circulares cuja área (e ângulo ao centro correspondente) é proporcional à frequência da categoria/classe que representam.

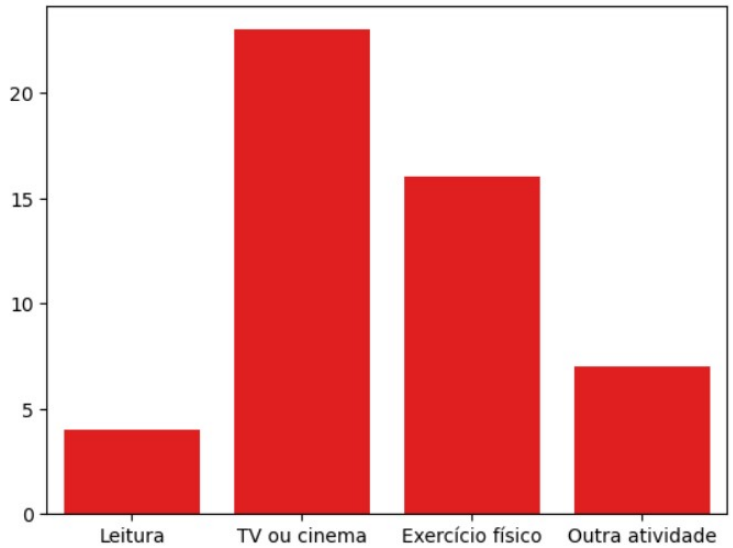
Exemplo 2.1: Realizou-se um inquérito a 50 habitantes de uma cidade para analisar a preferência na ocupação dos tempos livres. Registou-se a preferência (qualitativa) de cada habitante - unidade estatística - e organizaram-se os resultados na tabela seguinte.

Preferência	**Número de habitantes**	**Frequência relativa**
Leitura	4	$4/50 = 8\%$
TV ou cinema	23	$23/50 = 46\%$
Exercício físico	16	$16/50 = 32\%$
Outra atividade	7	$7/50 = 14\%$
Total	50	$50/50 = 100\%$

Dados qualitativos e quantitativos discretos

Exemplo 2.1 (Cont.): Representar os dados num gráfico de barras.

Output:



Comandos Python:

```
import seaborn as sns
x = ['Leitura', 'TV ou cinema',
     'Exercício físico', 'Outra atividade']
y = [4, 23, 16, 7]
sns.barplot(x=x,y=y, color = 'red')
```

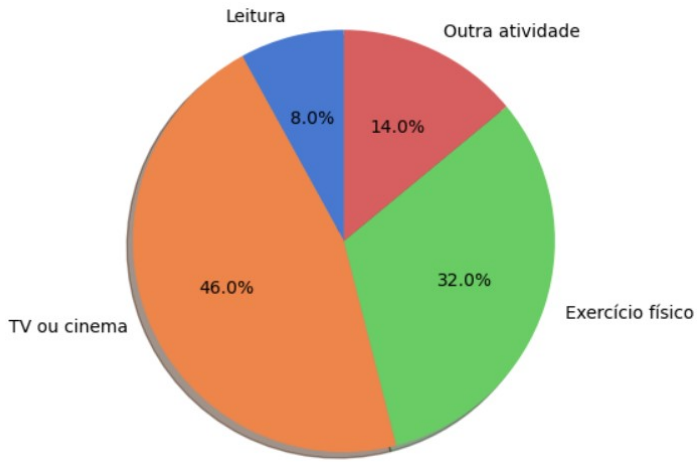
Dados qualitativos e quantitativos discretos

Exemplo 2.1 (Cont.): Representar os dados num gráfico circular.

Comandos Python:

```
import seaborn as sns
import matplotlib.pyplot as plt
y = [4, 23, 16, 7]
txtLabels = 'Leitura', 'TV ou cinema',
      'Exercício físico', 'Outra atividade'
plt.pie(y, labels=txtLabels,
      autopct='%1.1f%%', shadow=True,
      startangle=90,
      colors=sns.color_palette('muted'))
plt.axis('equal')
```

Output:



Dados quantitativos discretos

Frequência acumulada absoluta até à classe i (N_i)

É o número de observações de valor inferior ou igual ao valor característico da classe i:

$$N_i = \sum_{j=1}^{i} n_j.$$

Frequência acumulada relativa até à classe i (F_i)

É o quociente entre a frequência acumulada absoluta até à classe i e o número total n de observações efetuadas (é comum exprimir-se em percentagem):

$$F_i = \frac{N_i}{n} = \sum_{j=1}^{i} f_j = \frac{1}{n} \sum_{j=1}^{i} n_j.$$

Exemplo 2.2: Realizou-se um estudo de uma amostra de 5000 apólices do ramo automóvel em que se observou para cada apólice - unidade estatística - o número de sinistros (quantitativa discreta) ocorridos nos três primeiros anos de seguro.

Núm. de sinistros	Núm. de habitantes	Frequência relativa
0	2913	$2913/5000 = 58\%$
1	1532	$1532/5000 = 31\%$
2	381	$381/5000 = 8\%$
3	102	$102/5000 = 2\%$
4	72	$72/5000 = 1\%$
Total	5000	$5000/5000 = 100\%$

Dados quantitativos contínuos

- A classe de uma variável quantitativa contínua é um intervalo da forma $[a, b[$ ou $]a, b]$ que representa um conjunto de valores que a variável pode tomar.
- A amplitude da classe $[a, b[$ ou $]a, b]$ é a distância entre o limite superior da classe e o seu limite inferior, isto é, amplitude $= b - a$.
- O centro, marca, ponto médio ou valor característico da classe $[a, b[$ ou $]a, b]$ é o ponto médio da classe, isto é, $\frac{a+b}{2}$.
- O número de classes adequado, c, pode ser obtido pela fórmula (Regra de Sturges),

$$c = \text{Int}[1 + 3.3\log_{10}(n)],$$

onde n é o número total de observações e $\text{Int}[x]$ representa a parte inteira de x.

Dados quantitativos contínuos

Histograma

É uma representação gráfica dos dados em que se marcam as classes no eixo horizontal, as frequências no eixo vertical e em que se usam barras de área proporcional à frequência da classe correspondente. As barras contíguas têm uma fronteira comum.

Quando se elabora uma tabela de distribuições de frequência é necessário definir o número de classes, a amplitude de cada classe e o limite inferior da primeira classe.

Exemplo 2.3: Realizou-se um estudo de uma amostra de 104 doentes renais, registando-se, para cada um, o tempo (meses) (quantitativa contínua) de hemodiaise antes da realização do transplante.

Tempo de hemodialise	Freq. absoluta	Freq. relativa
$0 - 15$	9	$9/104 = 8.7\%$
$15 - 30$	35	$35/104 = 33.7\%$
$30 - 45$	20	$20/104 = 19.2\%$
$45 - 60$	20	$20/104 = 19.2\%$
$60 - 75$	7	$7/104 = 6.7\%$
$75 - 90$	4	$4/104 = 3.8\%$
$90 - 105$	5	$5/104 = 4.8\%$
$105 - 120$	1	$1/104 = 1\%$
$120 - 135$	1	$1/104 = 1\%$
$135 - 150$	2	$2/104 = 1.9\%$
Total	104	$104/104 = 100\%$

Dados quantitativos contínuos

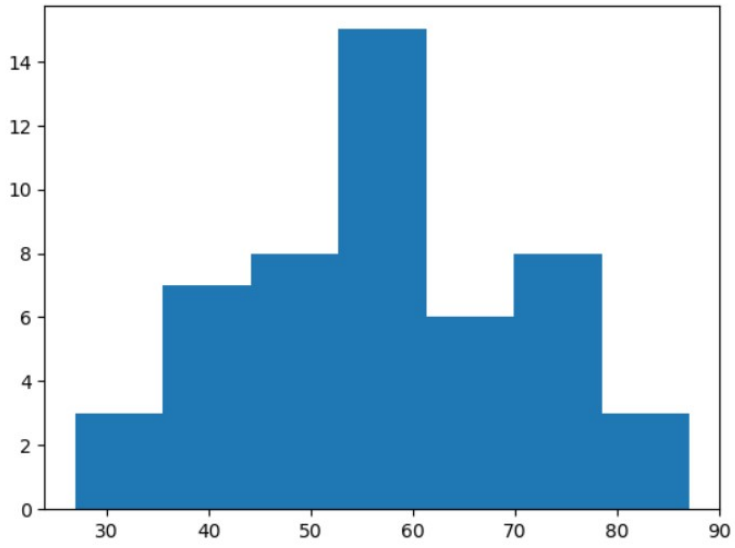
Exemplo 2.4: Registaram-se, numa escala de 0 a 100 pontos percentuais, as classificações dos 50 estudantes de um curso de Estatística e pretende-se mostrar os dados num histograma. As classificações foram as seguintes:

43	52	57	53	41
56	39	44	47	49
57	33	59	43	69
80	79	56	59	58
71	50	45	78	64
66	61	87	65	61
73	74	36	55	52
65	53	69	77	34
74	76	73	55	60
49	51	53	44	27

Dados quantitativos contínuos

Exemplo 2.4 (Cont.):.

Output:



Comandos Python:

```
import seaborn as sns
import matplotlib.pyplot as plt
df = pd.read_csv('notas_est_csv.csv')
plt.hist(df, bins='auto')
```

Frequência acumulada absoluta até à classe i (N_i)

- É o número de observações de valor inferior ou igual ao valor limite superior da classe i, no caso de intervalos do tipo $]a, b]$.
- É o número de observações de valor inferior ao valor limite superior da classe i, no caso de intervalos do tipo $[a, b[$. Ou seja,

$$N_i = \sum_{j=1}^{i} n_j.$$

Frequência acumulada relativa até à classe i (F_i)

É o quociente entre a frequência acumulada absoluta até à classe i e o número total n de observações efetuadas (é comum exprimir-se em percentagem):

$$F_i = \frac{N_i}{n} = \sum_{j=1}^{i} f_j = \frac{1}{n} \sum_{j=1}^{i} n_j.$$

Medidas descritivas

Medidas descritivas

As medidas descritivas são estatísticas amostrais (funções de uma amostra) que sumariam características importantes das amostas e dividem-se em três categorias:

- Medidas de posição ou localização.
  - Medidas de localização central: média, mediana e moda.
  - Medidas de localização não central: quantis (quartis, decis e percentis).
- Medidas de variabilidade: Amplitude interquartil, variância, desvio padrão e coeficiente de variação.
- Medidas de assimetria e de curtose.

Média aritmética (dados não classificados)

A média $\bar{x}$ de um conjunto de n valores observados $x_1, x_2, \ldots, x_n$ é dada por,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Média aritmética (dados classificados)

A média $\bar{x}$ de um conjunto de n valores observados, agrupados em c classes é dada por,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{c} n_i x_i = \sum_{i=1}^{c} f_i x_i,$$

em que n_i é a frequência absoluta da classe i, f_i é a frequência relativa da classe i e x_i é o valor característico da classe i.

Mediana $\tilde{x}$ (dados discretos ou contínuos não classificados)

Dado um conjunto de n valores observados $x_1, x_2, ..., x_n$, seja $x_{(1)}, x_{(2)}, ..., x_{(n)}$ a ordenação dos dados por ordem crescente, ou seja, $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$. A mediana é dada por,

$$\tilde{x} = \begin{cases} \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{se } n \text{ par} \\ x_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ ímpar} \end{cases}.$$

Moda (dados não classificados e dados classificados discretos)

É o valor que ocorre com maior frequência num conjunto de observações.

Mediana $\tilde{x}$ (dados contínuos classificados)

Dado um conjunto de valores de uma variável contínua agrupados em classes, a mediana é dada por,

$$\tilde{x} = \ell_{e-1} + \frac{0.5 - F_{e-1}}{f_e}(\ell_e - \ell_{e-1}),$$

em que e é a classe da mediana (tal que $F_{e-1} < 0.5$ e $F_e \geq 0.5$), f_e é a frequência relativa da classe e, F_{e-1} é a frequência acumulada relativa até à classe $e-1$, e ℓ_{e-1} e ℓ_e são, respetivamente, os limites inferior e superior da classe da mediana.

Moda (dados contínuos classificados)

Numa distribuição de frequência, com intervalos de classe de igual amplitude, a classe modal é a classe com maior frequência.

Exemplo 2.5:

Comandos Python:

```
# Para a média e mediana também se pode usar a biblioteca numpy
import statistics as st
x = [0, 1, 1, 1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 0]
print(f'A média dos valores é st.mean(x):.2f;', end = " ")
print(f'A mediana dos valores é st.median(x):.2f')
print(f'A moda dos valores é st.mode(x)')
```

Output: A média dos valores é 3.79; A mediana dos valores é 3.50
A moda dos valores é 1

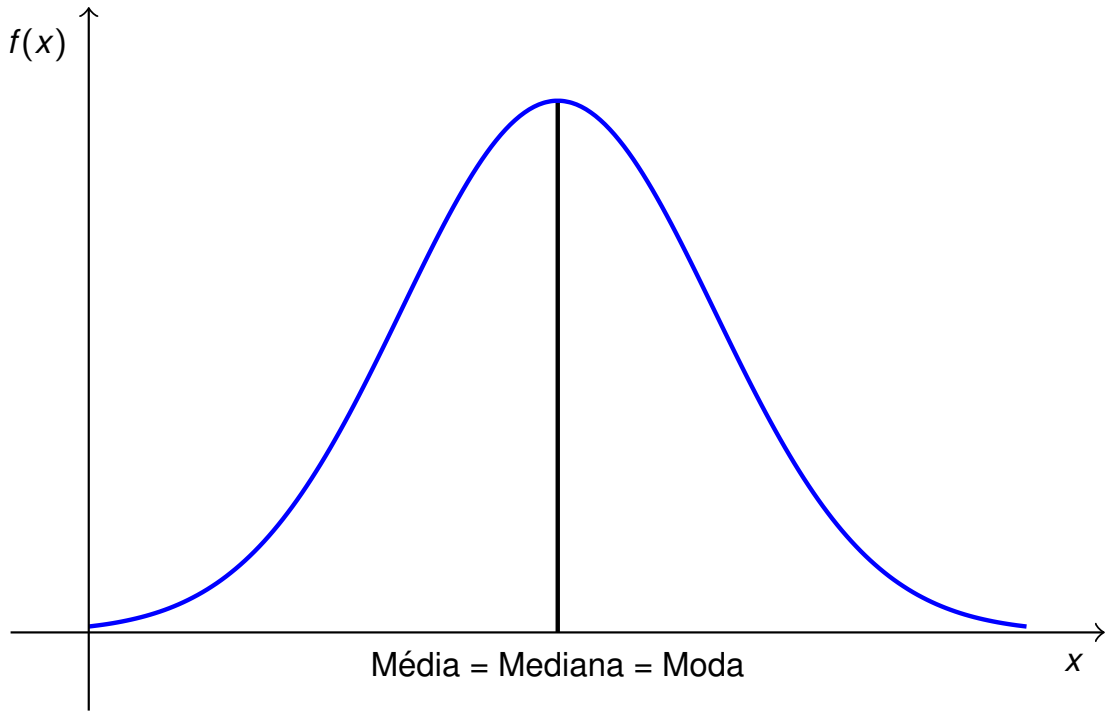
Comandos Python:

```
import pandas as pd
state = pd.read_csv('state.csv')
state['Population'].mean(), state['Population'].median()
```

Output: (6162876.3, 4436369.5)

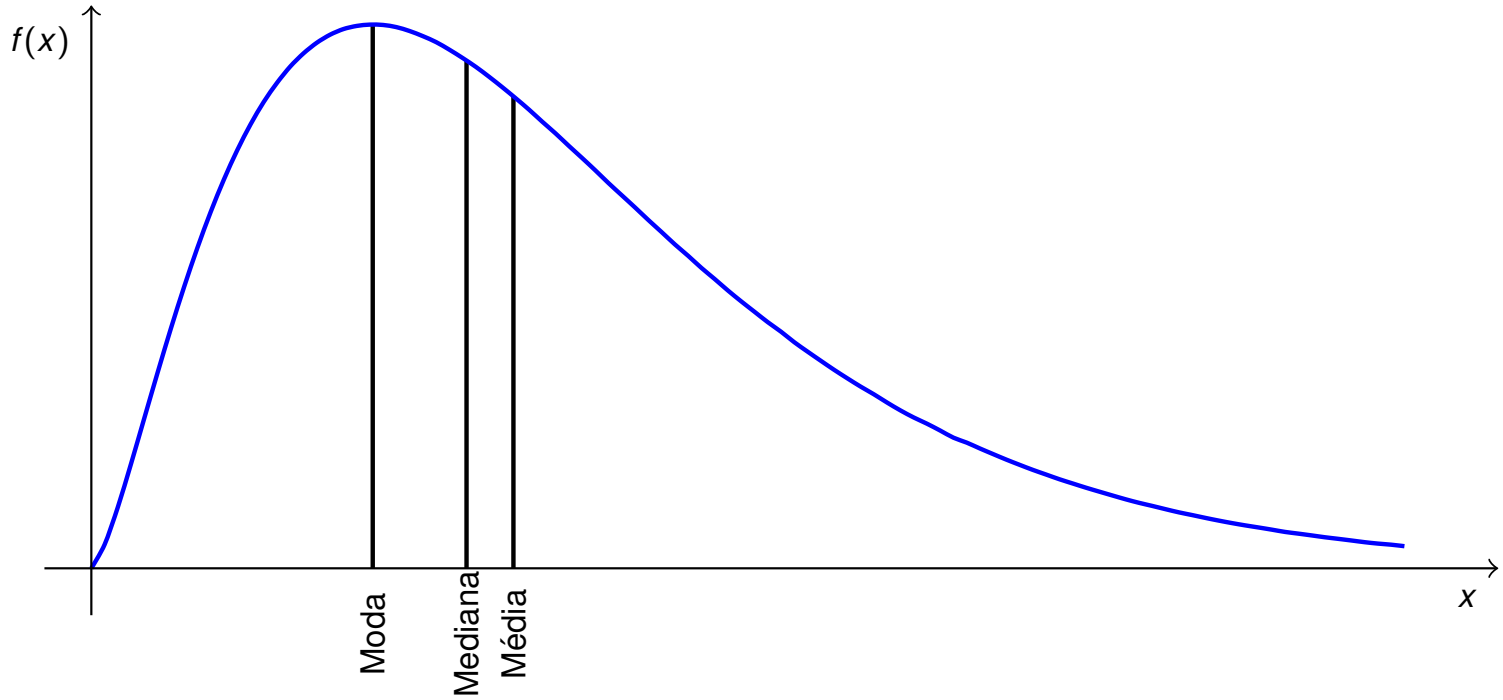
Comparação entre média, mediana e moda

Distribuição simétrica



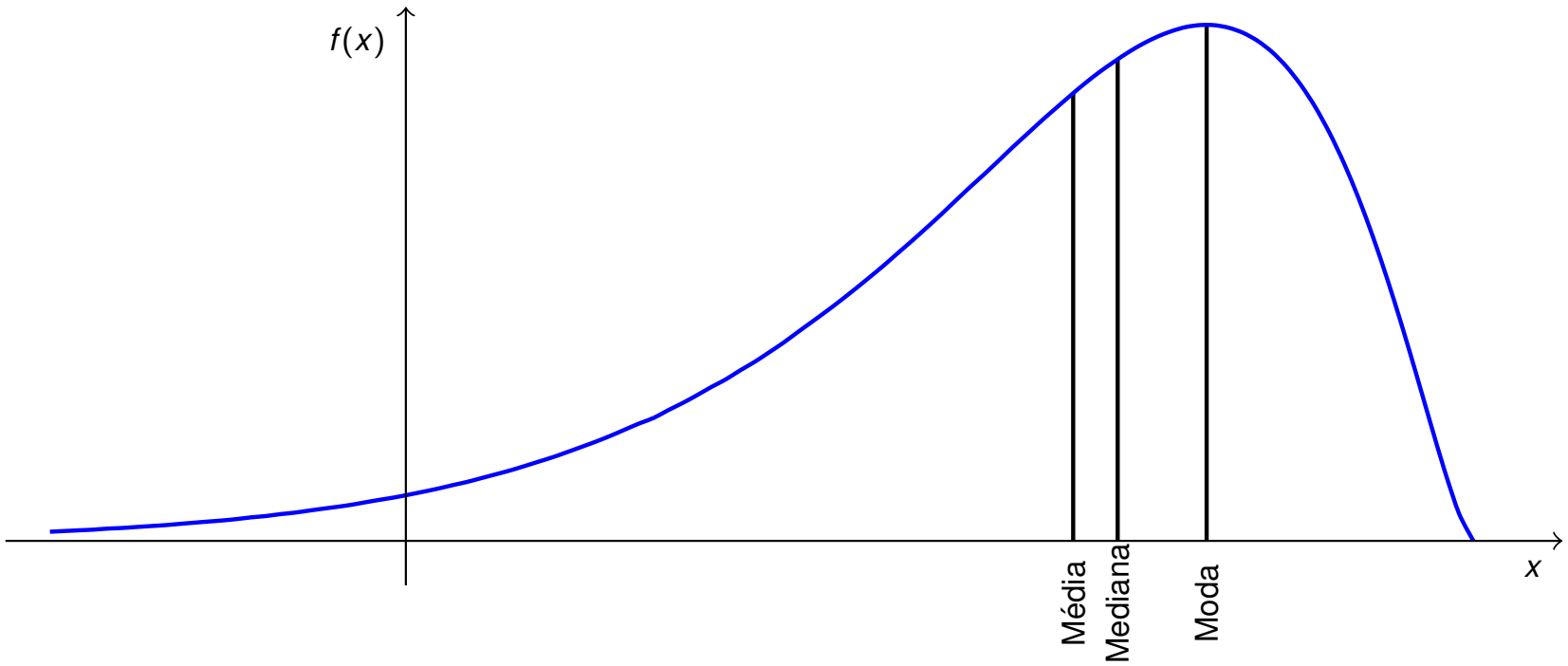
Comparação entre média, mediana e moda

Distribuição assimétrica positiva ou enviesada à direita



Comparação entre média, mediana e moda

Distribuição assimétricas negativa ou enviesada à esquerda



Comparação entre média, mediana e moda

- A média é a medida de localização mais usada, pois é a mais fácil de calcular e tem tratamento matemático simples.
- A média tem a vantagem do seu cálculo incluir todos os valores do conjunto de dados e varia menos com a amostra do que a mediana.
- No entanto, se a distribuição de frequência é assimétrica, a média tende a afastar-se da zona de concentração de observações, devido à influência de valores extremos atípicos ou *outliers*.
- A mediana pode ter mais interesse quando as obervações incluem valores extremos atípicos ou *outliers*.
- A moda não é muito prática na maioria das situações.

Quantis

Percentil

Considere um conjunto de dados ordenado em ordem crescente.

- Os percentis dividem um conjunto de dados em 100 partes iguais.
- O valor do percentil de ordem k ($k = 1, 2, ..., 99$) é denotado por p_k.
- Cerca de k% das observações são menores do que p_k.

Por exemplo, se num conjunto de dados se verifica que $p_{25} = 3.6$, então significa que 25% das observações é menor que 3.6.

Quantis

Decil

Considere um conjunto de dados ordenado em ordem crescente.

- Os decis dividem um conjunto de dados, em 10 partes iguais.
- O valor do decil de ordem k ($k = 1, 2, ..., 9$) é denotado por d_k.
- Cerca de $10k\%$ das observações são menores do que d_k.

Por exemplo, se num conjunto de dados se verifica que $d_5 = 3.6$, então significa que 50% das observações é menor que 3.6.

Quantis

Quartil

Considere um conjunto de dados ordenado em ordem crescente.

- Os quartis dividem um conjunto de dados, em 4 partes iguais.
- O valor do quartil de ordem k ($k = 1, 2, 3$) é denotado por q_k.
- Cerca de $25k\%$ das observações são menores do que q_k.

Por exemplo, se num conjunto de dados se verifica que $q_3 = 3.6$, então significa que 75% das observações é menor que 3.6.

Medidas de localização não central

Comandos Python:

```
import statistics as st
x = [1, 5, 7, 5, 43, 43, 8, 43, 6, 65, 63, 42, 1, 76, 43, 87, 53, 54]
decis = st.quantiles(x, n=10)
quartis = st.quantiles(x, n=4)
print(f'Os 4 quartis são: quartis\n')
percentis = st.quantiles(x, n=100)
print(f'O percentil de ordem 50 é percentis[49]')
print(f'O decil de ordem 5 é decis[4]')
print(f'O quartil de ordem 2 é quartis[1]')
```

Output: Os quartis são: [5.75, 43.0, 56.25]

O percentil de ordem 50 é 43.0
O decil de ordem 5 é 43.0
O quartil de ordem 2 é 43.0

Amplitude total, r

É a diferença entre o maior e o menor dos valores do conjunto de observações:

$$r = \max\{x_1, x_2, ..., x_n\} - \min\{x_1, x_2, ..., x_n\} = x_{(n)} - x_{(1)},$$

se $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ for a ordenação das obervações por ordem crescente.

Amplitude interquartil, r_q

$$r_q = q_3 - q_1,$$

representa a diferença entre o terceiro e o primeiro quartil, ou seja, indica como estão dispersas 50% das observações.

Variância s^2 de uma amostra

Para dados não classificados: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, sendo x_i o valor da observação i.

Para dados classificados: $s^2 = \frac{1}{n-1}\sum_{i=1}^{c} n_i(x_i - \bar{x})^2$, sendo x_i o valor característico da classe i.

Desvio padrão s de uma amostra

É raíz quadrada positiva da variância, $s = \sqrt{s^2}$.

Coeficiente de variação c_V

É dado por, $c_V = \frac{s}{|\bar{x}|}$, para $\bar{x} \neq 0$. Trata-se de uma medida adimensional e representa-se, em geral, na forma de percentagem.

Comandos Python:

```
import pandas as pd
import matplotlib.pyplot as plt
data = pd.DataFrame({
'Gender': ['f', 'f', 'm', 'f', 'm', 'm', 'f', 'm', 'f', 'm', 'm'],
'TV': [3.4, 3.5, 2.6, 4., 4.1, 4.1, 5.1, 3.9, 3.7, 2.1, 4.3]
} )
print(data)
grouped = data.groupby('Gender')
print(grouped.describe())
```

```
   Gender   TV
0       f  3.4
1       f  3.5
2       m  2.6
3       f  4.0
4       m  4.1
5       m  4.1
6       f  5.1
7       m  3.9
8       f  3.7
9       m  2.1
10      m  4.3
```

Output:

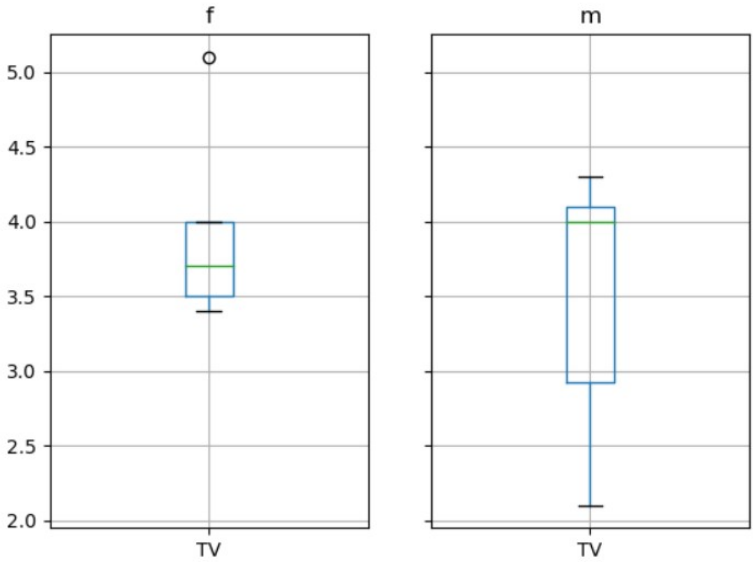
	TV							
	count	mean	std	min	25%	50%	75%	max
Gender								
f	5.0	3.940000	0.687750	3.4	3.500	3.7	4.0	5.1
m	6.0	3.516667	0.926103	2.1	2.925	4.0	4.1	4.3

O diagrama de extremos e quartis (*boxplot*) é uma representação gráfica dos valores: *mínimo*, q_1, *mediana*, q_3, *máximo*.

Verifica-se:

- Os extremos da caixa são os primeiro e terceiro quartis, ou seja, o comprimento da caixa é a amplitude interquartil;
- A mediana é marcada como uma linha na caixa;
- Duas linhas extendem-se até aos valores mínimo e máximo;
- Os *outliers*, que se podem definir como sendo os valores que se afastam da mediana mais do que 1.5 vezes a amplitude interquartil, são representados isoladamente.

```
# plot the data
grouped.boxplot()
plt.show()
```




```
# Get the groups as DataFrames
df_female = grouped.get_group('f')
values_female = df_female.values
print(values_female)
```

```
[['f' 3.4]
 ['f' 3.5]
 ['f' 4.0]
 ['f' 5.1]
 ['f' 3.7]]
```

Momento amostral centrado de ordem r

É a média dos desvios em relação a $\bar{x}$ elevados à potência r (inteiro não negativo).

Para dados não classificados,

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r,$$

sendo x_i o valor da observação i.

Para dados classificados,

$$m_r = \frac{1}{n}\sum_{i=1}^{c} n_i(x_i - \bar{x})^r,$$

sendo x_i o valor característico da observação i.

O coeficiente de assimetria amostral é uma medida de assimetria que indica se uma distribuição de frequências é enviesada ou assimétrica. Seja,

$$a_3 = \frac{m_3}{s^3}$$

- m_3 é o momento centrado de ordem 3;
- s o desvio padrão experimental;
- $a_3 > 0$ para uma distribuição enviesada à direita ou assimétrica positiva;
- $a_3 < 0$ para uma distribuição enviesada à esquerda ou assimétrica negativa;
- $a_3 = 0$ para uma distribuição simétrica.

O coeficiente de curtose amostral dá-nos informação sobre o peso das caudas de uma distribuição e é dado por,

$$a_4 = \frac{m_4}{s^4}$$

- m_4 é o momento centrado de ordem 4;
- s o desvio padrão experimental.
- $a_4 > 3$ para uma distribuição mais esguia, caudas mais pesadas do que a distribuição Normal;
- $a_4 = 3$ para uma distribuição Normal;
- $a_4 < 3$ para uma distribuição mais achatada, caudas menos pesadas do que a distribuição Normal.

O coeficiente de assimetria amostral e o coeficiente de curtose amostral são adimensionais, facilitando a comparação entre distribuições de frequências distintas.

Comandos Python:

```
from scipy import stats

x = [88, 85, 82, 97, 67, 77, 74, 86, 81, 95, 77, 88, 85, 76, 81]

# calcula o coeficiente de assimetria amostral
assimetria = stats.skew(data, bias= False )

# calcula o coeficiente de curtose amostral
curtose = stats.kurtosis(data, fisher = False )

print(f'O valor do coeficiente de assimetria amostral é assimetria:.4f')
print(f'O valor do coeficiente de curtose amostral é curtose:.4f')
```

Output:

O valor do coeficiente de assimetria amostral é 0.0293
O valor do coeficiente de curtose amostral é 2.7073