A estatística pode ser dividida em duas grandes áreas:

Estatística descritiva

Inferência estatística

Estatística descritiva: é um conjunto de métodos cujo objetivo é sintetizar e representar de forma compreensível a informação contida nos dados. Usam-se frequentemente medidas, tabelas e gráficos para sintetizar e representar a informação.

Inferência estatística: é um conjunto de métodos que permitem fazer estimativas e tirar conclusões sobre a população a partir da informação contida na amostra

Cofinanciado por:

UNIÃO EUROPEIA
Fundo Social Europeu

Ana Cristina Matos

Conceitos Básicos

- **População** ou **Universo Estatístico**: conjunto de elementos sobre o qual incide o estudo estatístico;
- **Característica Estatística** ou **Atributo**: a característica que se observa nos elementos da população;
- **Modalidades** (incompatíveis e exaustivas): as diversas formas em que se apresenta a característica estatística;
- **Amostra**: subconjunto finito da população (razões para a recolha de uma amostra: dimensão excessiva da população, estudo de natureza destrutiva, economia e tempo)

Cofinanciado por:

PORTUGAL 2020

UNIÃO EUROPEIA
Fundo Social Europeu

Exemplos:

- O Gestor de produção de uma fábrica pretende ter uma ideia da percentagem de peças defeituosas que a fábrica produziu em determinado período de tempo.

*A **população** em estudo é constituída por todas as peças produzidas pela fábrica durante aquele período de tempo. A **característica estatística** tem apenas duas modalidades: peça defeituosa e peça não defeituosa.*

- Num estudo de mercado para construção de um centro comercial, interessa estudar o rendimento familiar mensal dos habitantes de uma determinada cidade.

Cofinanciado por:

PORTUGAL 2020

UNIÃO EUROPEIA
Fundo Social Europeu

*A **população** é constituída pelas famílias daquela cidade e **a característica estatística** é o rendimento familiar mensal. As **modalidades** do rendimento familiar mensal não se podem enumerar; são todos os valores desde, por exemplo, 50 contos até 1000 contos.*

- Uma determinada empresa pretende realizar um inquérito aos seus trabalhadores, onde lhes é pedido para classificarem a qualidade do serviço do bar/refeitório segundo a seguinte escala: fraco, razoável, bom ou muito bom.

*Os trabalhadores da fábrica constituem a **população** em estudo, e a **característica estatística** é a opinião acerca da qualidade do serviço do bar/refeitório. Neste estudo o atributo pode manifestar-se nas seguintes **modalidades**: fraco, razoável, bom ou muito bom.*

Cofinanciado por:

POCH PORTUGAL 2020 UNIÃO EUROPEIA Fundo Social Europeu

Tipos de Dados estatísticos

- Quantitativos (e.g., o número diário de nascimentos no hospital de Viseu ou a altura dos alunas da ESTV):
  - **Discretos** (número finito ou infinito numerável de modalidades; e.g., o número diário de nascimentos no hospital de Viseu)
  - **Contínuos** (pode assumir qualquer valor num intervalo de números reais; a distinção entre os dois é por vezes arbitrária; e.g., altura de um aluno da ESTV)

- Qualitativos (e.g., cor dos cabelos, estado civil)

Cofinanciado por:

PORTUGAL 2020

UNIÃO EUROPEIA
Fundo Social Europeu

Escalas de medida de dados estatísticos

Escala Nominal (dados qualitativos): apresentam-se em diferentes categorias ou classes, não ordenáveis. Por exemplo:

- Estado civil dos empregados de uma empresa;
- Religião; cor de cabelos;
- Os números das camisolas dos futebolistas;
- Sexo dos indivíduos de uma população (**característica dicotómica** ou **binária);**
- Numa sondagem de opinião, a resposta à pergunta "É a favor da despenalização do aborto?" (**característica dicotómica** ou **binária)**.

Cofinanciado por:

UNIÃO EUROPEIA
Fundo Social Europeu

Para lidar com este tipo de dados é frequente atribuir um código numérico a cada categoria da característica em estudo, mas cuidado! não faz qualquer sentido usar operações aritméticas e calcular médias, desvios padrões, etc..

Escala Ordinal (dados qualitativos): as diversas categorias possuem uma ordem intrínseca (os códigos numéricos devem ter em conta essa ordem). Por exemplo:

- O sistema de graduação militar: Soldado, Cabo, Sargento, ...
- Num inquérito de opinião pede-se às pessoas que classifiquem um determinado produto como sendo, muito fraco, fraco, razoável, bom ou muito bom. (escala de Likert).

Cofinanciado por:

PORTUGAL 2020

UNIÃO EUROPEIA
Fundo Social Europeu

- Classificação dos clientes de um banco, segundo o volume de capital que movimentam mensalmente: pouco importantes, importantes ou muito importantes.
- Classificação dos alunos de uma escola segundo a sua altura: baixos (menos de 155 cm), médios (entre 155 e 170 cm) ou altos (mais de 170 cm).

Representação de Dados

População ou amostra de **n** indivíduos.

Atributo **A** com **p** modalidades: $\mathbf{A_1}, \mathbf{A_2}, \ldots, \mathbf{A_p}$.

Frequência absoluta ou **efetivo** da modalidade $A_i \rightarrow \mathbf{n_i}$, é o nº de indivíduos que apresentam a modalidade A_i.

Cofinanciado por:

UNIÃO EUROPEIA
Fundo Social Europeu

Frequência relativa da modalidade $A_i \rightarrow$ **f_i**, é a proporção de indivíduos que apresentam a modalidade A_i, $f_i = \frac{n_i}{n}$.

$$\sum_{i=1}^{p} n_i = n \qquad \text{e} \qquad \sum_{i=1}^{p} f_i = 1.$$

Cofinanciado por:

POCH PORTUGAL 2020 UNIÃO EUROPEIA Fundo Social Europeu

• *Representação Tabular – Quadros de Frequências*

Modalidades	Frequências absolutas	Frequências relativas	Frequências absolutas acumuladas	Frequências relativas acumuladas
A_1	n_1	$f_1=n_1/n$	n_1	f_1
A_2	n_2	$f_2=n_2/n$	n_1+n_2	f_1+f_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_P	n_p	$f_p=n_p/n$	$n_1+n_2+\ldots+n_p=n$	$f_1+f_2+\ldots+f_p=1$
Total	n	1	-	-

Cofinanciado por:

PORTUGAL 2020

UNIÃO EUROPEIA
Fundo Social Europeu

Exemplo 1: os dados que se seguem são relativos ao número de telefonemas de pedido de auxilio a um corpo de bombeiros de 30 dias selecionados dentro de um período específico.

120	130	80	100	110	100	90	70	140	120
140	110	100	100	110	70	90	90	130	150
160	80	70	120	100	110	110	80	100	120

Tabela de frequências - dados não agrupados

x_i	Freq. absolutas n_i	Freq. relativas f_i	Freq. absolutas acumuladas	Freq. relativas acumuladas
70	3	3/30	3	3/30
80	3	3/30	6	6/30
90	3	3/30	9	9/30
100	6	6/30	15	15/30
110	5	5/30	20	20/30
120	4	4/30	24	24/30
130	2	2/30	26	26/30
140	2	2/30	28	28/30
150	1	1/30	29	29/30
160	1	1/30	30	1
Total	30	1	-	-

Tabela de frequências com dados agrupados.

Classes de valores	Freq. absolutas n_i	Freq. relativas f_i	Freq. absolutas acum.	Freq. relativas acum.
[60, 80[	3	3/30	3	3/30
[80, 100[	6	6/30	9	9/30
[100, 120[	11	11/30	20	20/30
[120, 140[	6	6/30	26	26/30
[140, 160[	3	3/30	29	29/30
[160, 180[	1	1/30	30	30/30
Total	30	1	-	-

Cofinanciado por:

PORTUGAL 2020

UNIÃO EUROPEIA
Fundo Social Europeu

– Os intervalos de classe podem ter a mesma amplitude ou amplitudes diferentes dependendo da natureza dos fenómenos a estudar.

– Agrupar os dados implica perda de informação.

– Regras práticas para a determinação do nº de classes:

Regra de Sturges – nº de classes $\cong 1+\log_{10}(n)/\log_{10}(2)$

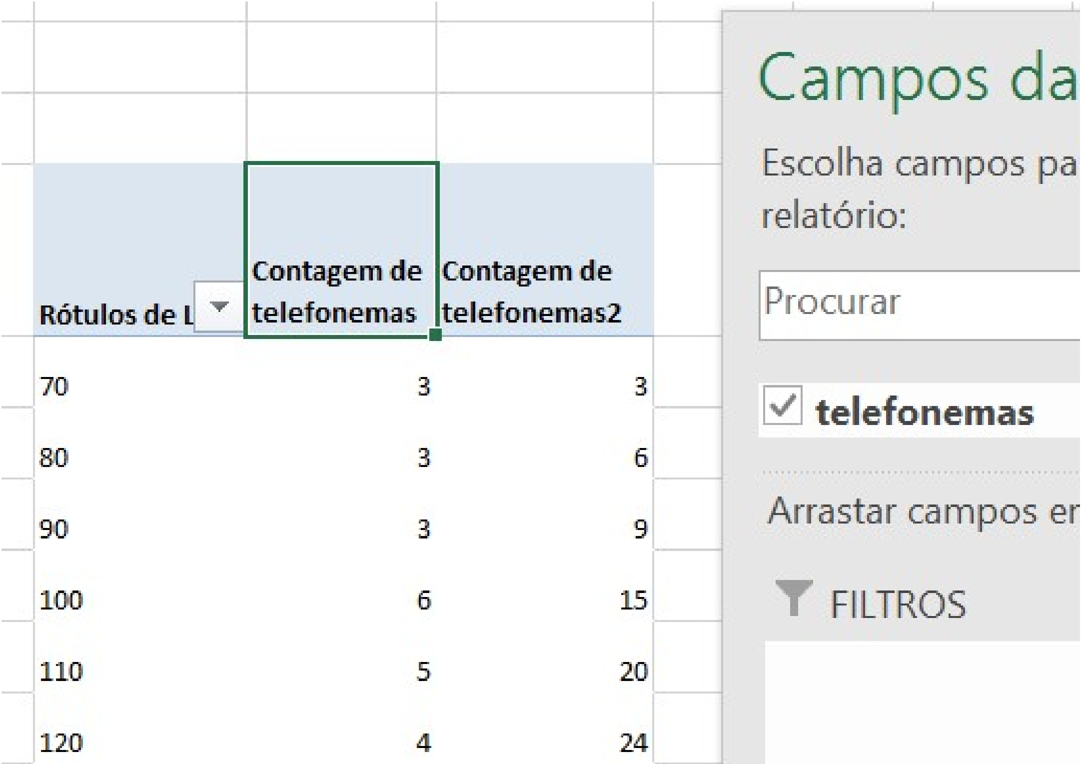
Outra – nº de classes $\cong \sqrt{n}$ (usualmente empregue quando n>25).

Recorrendo ao excel estas tabelas são facilmente elaboradas através das tabelas dinâmicas:

Cofinanciado por:

PORTUGAL 2020

UNIÃO EUROPEIA
Fundo Social Europeu



Rótulos de L	Contagem de telefonemas	Contagem de telefonemas2
70	3	3
80	3	6
90	3	9
100	6	15
110	5	20
120	4	24

Com o botão do lado direito do rato em cima de “Contagem de telefonemas 2”, selecionando `mostar valores como>total corrente em` obtemos as frequências absolutas acumuladas, assim como também existe a opção para as frequências relativas.

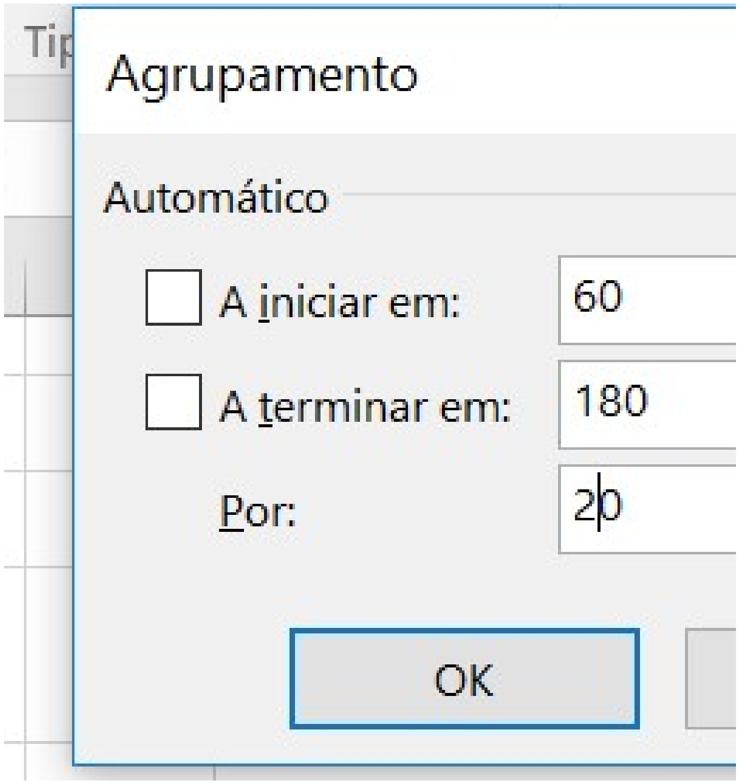
Para se trabalhar com tabelas agrupando a informação basta:

Selecionar uma célula da coluna dos “rótulos de linha” e definir o agrupamento pretendido:

Cofinanciado por:

PORTUGAL 2020

UNIÃO EUROPEIA
Fundo Social Europeu



Rótulos de Linha	Contagem	frequ absc acum
60-79	3	
80-99	6	
100-119	11	
120-139	6	

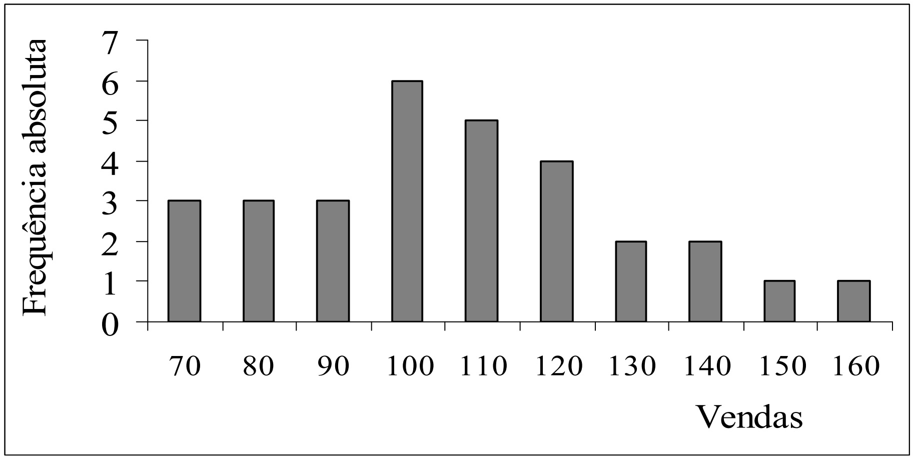
Cofinanciado por:

POCH PORTUGAL 2020 UNIÃO EUROPEIA Fundo Social Europeu

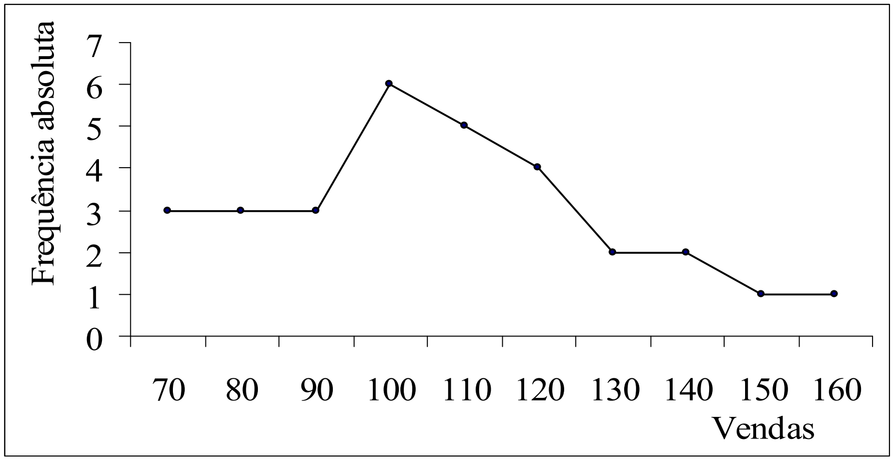
Ana Cristina Matos

• *Representação gráfica:* *DADOS NÃO AGRUPADOS*

Diagrama de barras

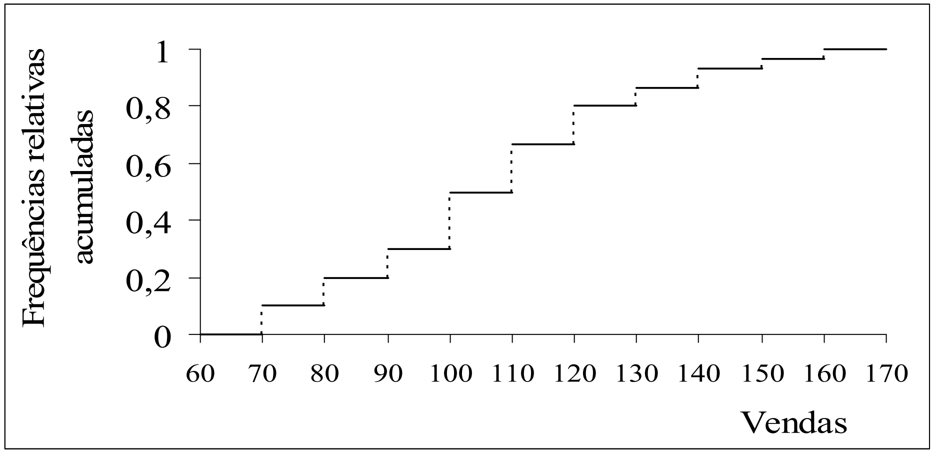


Polígono de frequências



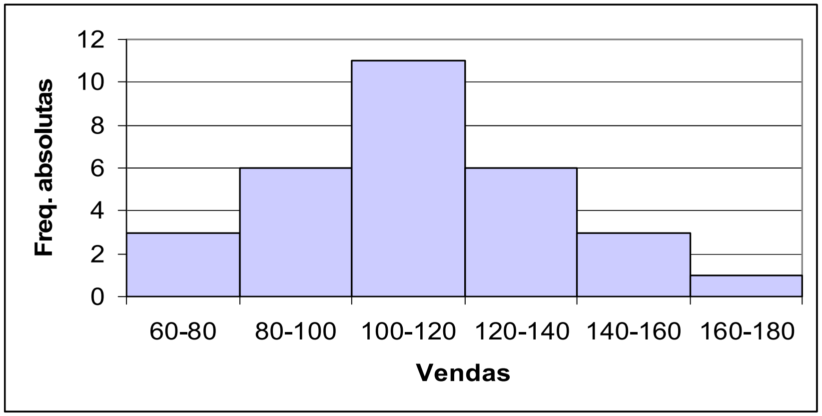
Cofinanciado por:

UNIÃO EUROPEIA
Fundo Social Europeu

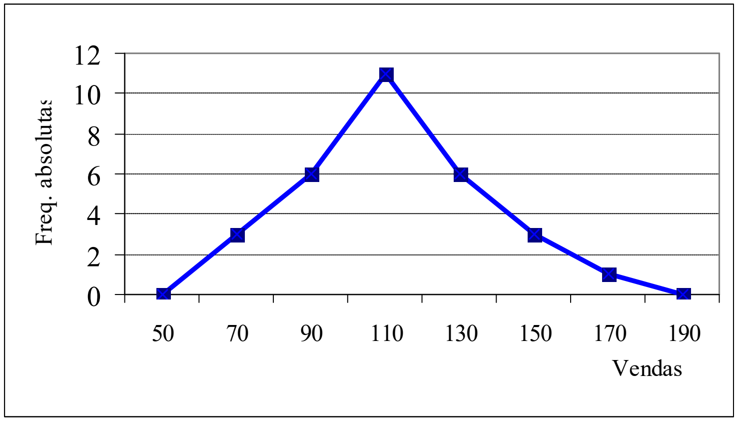


Dados Agrupados

. Histograma



Polígono de frequências

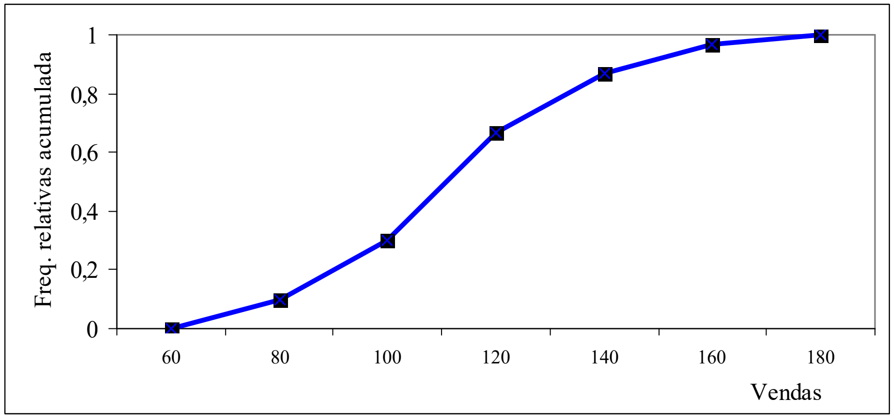


Cofinanciado por:

PORTUGAL 2020

UNIÃO EUROPEIA
Fundo Social Europeu

Polígono de frequências acumuladas

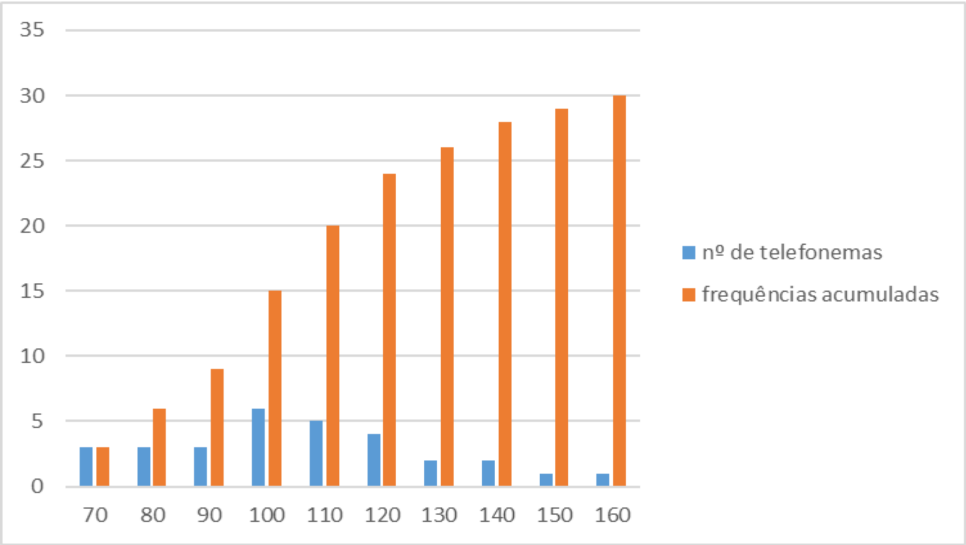


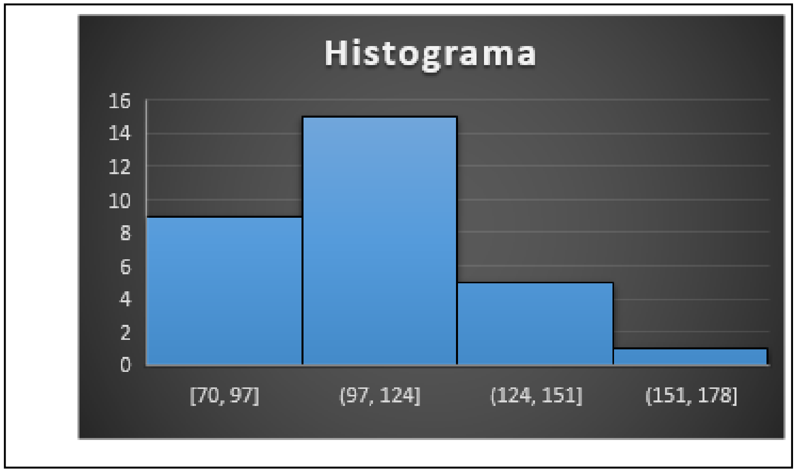
Também no Excel a representação gráfica é muito acessível. Basta selecionar uma célula da tabela dinâmica e no menu inserir, na opção gráficos ou gráficos recomendados aparecem várias hipóteses de representação gráfica. A título de exemplo temos:

Cofinanciado por:

PORTUGAL 2020

UNIÃO EUROPEIA
Fundo Social Europeu





Cofinanciado por:

UNIÃO EUROPEIA
Fundo Social Europeu

Medidas Descritivas

Medidas de Localização ou de Tendência Central (média, moda e mediana): Estas medidas dão-nos uma ideia do "centro" ou "localização" da distribuição dos dados.

Média aritmética

Sejam $x_1, x_2, \ldots, x_p$ os valores distintos de um conjunto de **n** dados, cada um deles com frequência absoluta n_i e frequência relativa f_i. Então a média aritmética representa-se por $\bar{x}$ e é dada por:

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

Cofinanciado por:

PORTUGAL 2020 UNIÃO EUROPEIA Fundo Social Europeu

Para **dados agrupados em classes** toma-se para x_i o ponto médio da *i-ésima* classe; n_i e f_i serão, naturalmente, a frequência absoluta e relativa da *i-ésima* classe, respetivamente.

Exemplo 2: A tabela de frequências que se segue é relativa ao número de pneus produzidos por dia na fábrica MAVOR, para uma amostra de 30 dias.

x_i	Freq. absoluta n_i	Freq. relativa f_i	Freq. abso. acum.	Freq. relat. acum.s	$n_i x_i$
18	2	0.06667	2	0.06667	36
20	3	0.1	5	0.16667	60
21	5	0.16667	10	0.33334	105
24	7	0.23333	17	0.56667	168
25	6	0.2	23	0.76667	150
28	4	0.13333	27	0.9	112
29	3	0.1	30	1	87
Total	30	1	-	-	718

Cofinanciado por:

UNIÃO EUROPEIA
Fundo Social Europeu

A média de pneus produzidos diariamente, para os 30 dias considerados é:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} n_i x_i = \frac{718}{30} = 23.9333.$$

Mediana

Trata-se do valor que divide o conjunto de dados, ordenados por ordem crescente, em duas partes iguais. Isto é, a mediana, como o próprio nome indica, é o ponto mediano de um conjunto de dados ordenados em ordem crescente.

Sejam $x_1, x_2, \ldots, x_n$, **n** observações ordenadas por ordem crescente dos seus valores, e que constituem o conjunto de dados em análise.

$$Me = \begin{cases} x_{(n+1)/2} & se \quad n \quad é \quad impar \\ \dfrac{x_{n/2} + x_{n/2+1}}{2} & se \quad n \quad é \quad par \end{cases}$$

Cofinanciado por:

PORTUGAL 2020

UNIÃO EUROPEIA
Fundo Social Europeu

Exemplo 2, como *n* é par: $Me = \frac{x_{30/2} + x_{30/2+1}}{2} = \frac{x_{15} + x_{16}}{2} = \frac{24+24}{2} = 24$.

Para **dados agrupados em classes**, procuramos a classe mediana, sendo esta tal que a sua frequência absoluta (resp. relativa) acumulada é ≥ $n/2$ (resp. 1/2) e a frequência absoluta (resp. relativa) acumulada da classe anterior é < $n/2$ (resp. 1/2).

Depois de encontrada a classe mediana, $[a_j, a_{j+1}[$, encontra-se a mediana por interpolação linear:

$$Me = a_j + \frac{n/2 - \sum_{i=1}^{j-1} n_i}{n_j}\left(a_{j+1} - a_j\right)$$

Cofinanciado por:

UNIÃO EUROPEIA
Fundo Social Europeu

Moda : É o valor mais frequente num conjunto de dados.

– {2, 3, 4, 4, 5} → Mo=4 (distribuição unimodal);

– {2, 2, 3, 4, 4, 5} → Mo=2 e 4 (distribuição bimodal);

– Havendo mais de 2 valores modais, a distribuição diz-se multimodal.

Exemplo 2 → Mo=24.

Quando os **dados estão agrupados em classes**, a classe modal é aquela que tem maior frequência por unidade de amplitude.

Medidas de Localização não Central – Quantis: Q_p

Cofinanciado por:

POCH PORTUGAL 2020 UNIÃO EUROPEIA Fundo Social Europeu

A mediana divide o conjunto de dados em duas partes iguais. Quando o conjunto de dados ordenados é dividido em 4 partes iguais, os pontos de divisão são chamados os ***quartis***:

- $\mathbf{Q_{1/4}}$, 1º quartil – valor que tem cerca de 25% dos dados abaixo dele;
- $\mathbf{Q_{2/4}}$, 2º quartil – valor que tem cerca de 50% dos dados abaixo dele – trata-se da Mediana;
- $\mathbf{Q_{3/4}}$, 3º quartil – valor que tem cerca de 75% dos dados abaixo dele.

Podemos ainda calcular os ***quintis***, ***decis***, ***percentis***,…

Cálculo do quantil de ordem p, Q_p: *Dados <u>não agrupados</u> em classes*

Sejam $x_1, x_2, \ldots, x_n$, n observações ordenadas por ordem crescente dos seus valores.

Se np não é um inteiro, então $Q_p=x_k$, onde k é o inteiro imediatamente seguinte a np. Caso contrário, sendo np um inteiro, então $Q_p=(x_{np}+x_{np+1})/2$.

Cofinanciado por:

POCH PORTUGAL 2020 UNIÃO EUROPEIA Fundo Social Europeu

Cálculo do quantil de ordem p, Q_p: *Dados agrupados em classes*

Seja **[*a_j*, *a_{j+1}*[** a classe que contém Q_p, i.e., que contém o valor ao qual corresponde a frequência absoluta (resp. relativa) acumulada de *np* (resp. *p*). Por interpolação linear obtém-se Q_p:

$$Q_p = a_j + \frac{np - \sum_{i=1}^{j-1} n_i}{n_j}\left(a_{j+1} - a_j\right)$$

Posição relativa da média, mediana e moda

As distribuições de frequências podem ser simétricas ou não. Considerando apenas distribuições unimodais, temos:

- Distribuições simétricas $\rightarrow \ \overline{x} = Me = Mo$
- Distribuições assimétricas positivas ou enviesada à direita $\rightarrow \ Mo < Me < \overline{x}$

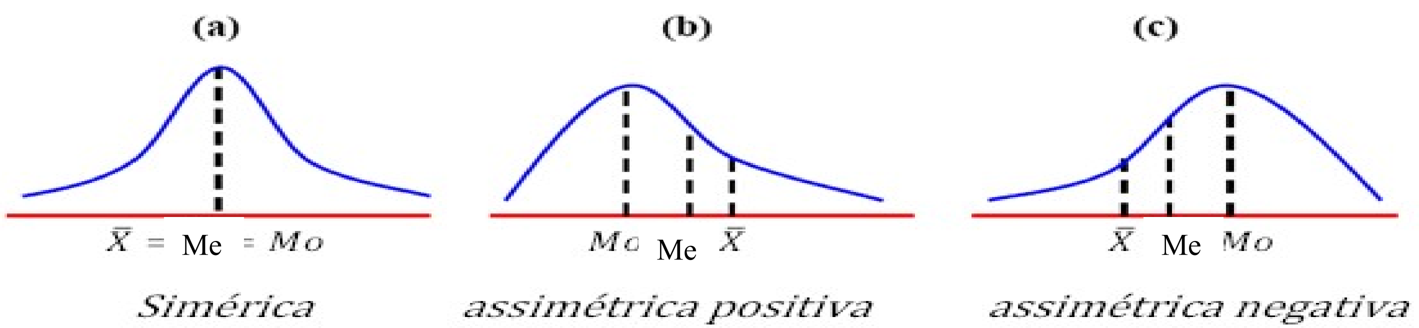
Cofinanciado por:

UNIÃO EUROPEIA
Fundo Social Europeu

A cauda direita é mais longa e menos abrupta do que a esquerda.

- Distribuições assimétricas negativas ou enviesada à esquerda $\rightarrow \; \overline{x} < Me < Mo$

A cauda esquerda é mais longa e menos abrupta do que a direita



Nas distribuições assimétricas os valores extremos da cauda mais longa puxam a média para o lado direito. A mediana, como divide a área em duas partes iguais, para compensar a redução de área no lado abrupto, afasta-se também da moda, mas menos do que a média.

Cofinanciado por:
UNIÃO EUROPEIA
Fundo Social Europeu

Há vários coeficientes para medir a assimetria; estes, de um modo geral, interpretam-se da seguinte maneira:

Distribuição **simétrica** $\rightarrow$ ***coeficiente*** $\cong$ **0**

Distribuição **assimétrica positiva** $\rightarrow$ ***coeficiente >0***

Distribuição **assimétrica negativa** $\rightarrow$ ***coeficiente <0***

Medidas de Dispersão

Exemplo: Duas empresas concorrentes com sede em Viseu, obtiveram os seguintes lucros nos 5 últimos anos:

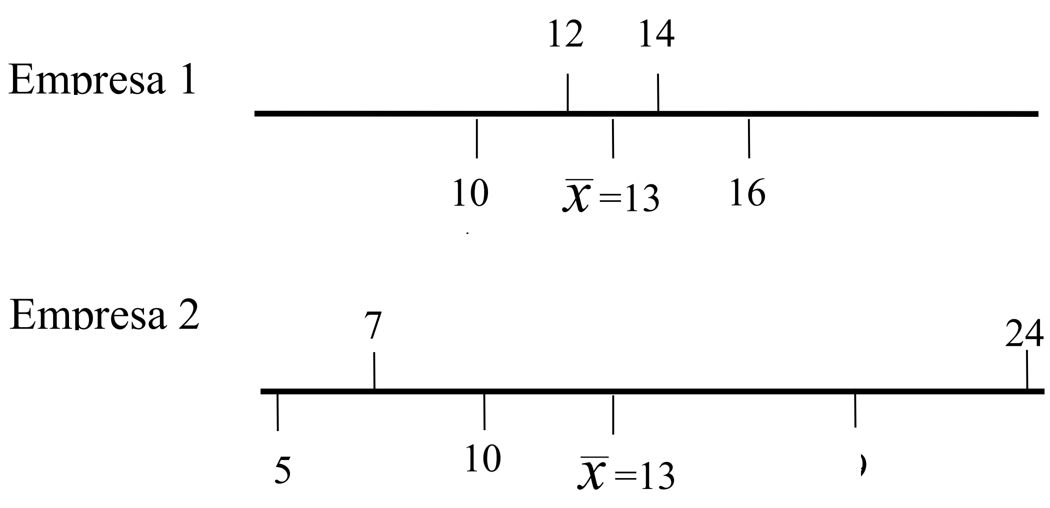
	Lucros em unidades monetárias (u. m.)				
Empresa 1	10	13	12	14	16
Empresa 2	7	5	10	19	24

Cofinanciado por:

PORTUGAL 2020

UNIÃO EUROPEIA
Fundo Social Europeu

O lucro médio das duas empresas nos últimos 5 anos é o mesmo, 13 u.m., no entanto a Empresa 2 apresenta uma maior variabilidade nos lucros do que a Empresa 1.



- O **intervalo interquartis, [$Q_{1/4}$, $Q_{3/4}$]** contém 50% das observações. A amplitude deste intervalo, **amplitude interquartis**, é uma medida de dispersão.

Cofinanciado por:

As medidas de dispersão mais utilizadas são o **desvio padrão** e a **variância** que definimos a seguir:

Sejam x_1, x_2, ..., x_p os valores distintos de um conjunto de ***n*** dados, cada um deles com frequência absoluta n_i e frequência relativa f_i.

Se estes dados constituem observações feitas sobre toda a população, a **variância** denota-se por σ^2 e é calculada da seguinte maneira:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{p} n_i\,(x_i - \overline{x})^2 = \sum_{i=1}^{p} f_i\,(x_i - \overline{x})^2 ,$$ ou equivalentemente,

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{p} n_i\, x_i^2 - \overline{x}^2 = \sum_{i=1}^{p} f_i\, x_i^2 - \overline{x}^2 .$$

Cofinanciado por:

UNIÃO EUROPEIA
Fundo Social Europeu

Escala de medida	Estatísticas descritivas	
	Medidas de tendência central	Medidas de dispersão
Nominal (sem qualquer relação de ordem). ● espécie ou género	**Moda**: valor mais frequente da amostra	Não há....
Ordinal (ordenável, mas sem quantificar). ● nível hierárquico ● ordem de chegada	Moda; **Quartis**: Valor abaixo do qual estão 25%($Q_{1/4}$), 50%($Q_{2/4}$) ou mediana) ou 75%($Q_{3/4}$) doa valores da amostra	Amplitude interquartilica AIQ= $Q_{3/4}$- $Q_{1/4}$ Intervalo interquartis=[$Q_{1/4}$, $Q_{3/4}$]
Quantitativa (ordenável sendo possível quantificar as diferenças). ● nº de células ● altura ● temperatura	Moda Quartis Média	AIQ Desvio padrão $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{p} n_i (x_i - \overline{x})^2}$ Erro padrão da média $s/\sqrt{n}$

Cofinanciado por:

PORTUGAL 2020

UNIÃO EUROPEIA
Fundo Social Europeu

Coeficiente de dispersão e de variação

Medidas de dispersão absolutas: expressas na mesma unidade dos dados a que se referem;

Medidas de dispersão relativas: independentes da unidade de medida dos dados a que se referem

A variância e o desvio padrão são medidas de dispersão absolutas.

Se pretendermos comparar a dispersão de dois conjuntos de dados que não estejam expressos na mesma unidade de medida ou escala teremos de adotar uma medida de dispersão relativa, por exemplo:

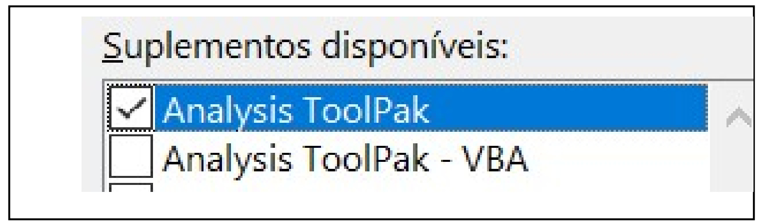
Cofinanciado por:

UNIÃO EUROPEIA
Fundo Social Europeu

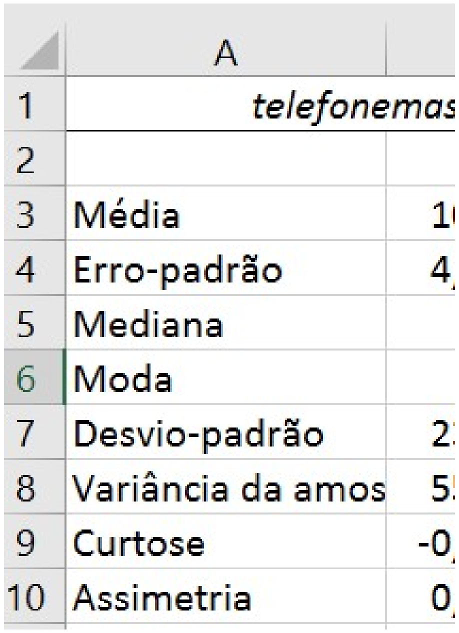
Coeficiente de dispersão: ***cd***= $\frac{s}{|\bar{x}|}$ ou $\frac{\sigma}{|\bar{x}|}$

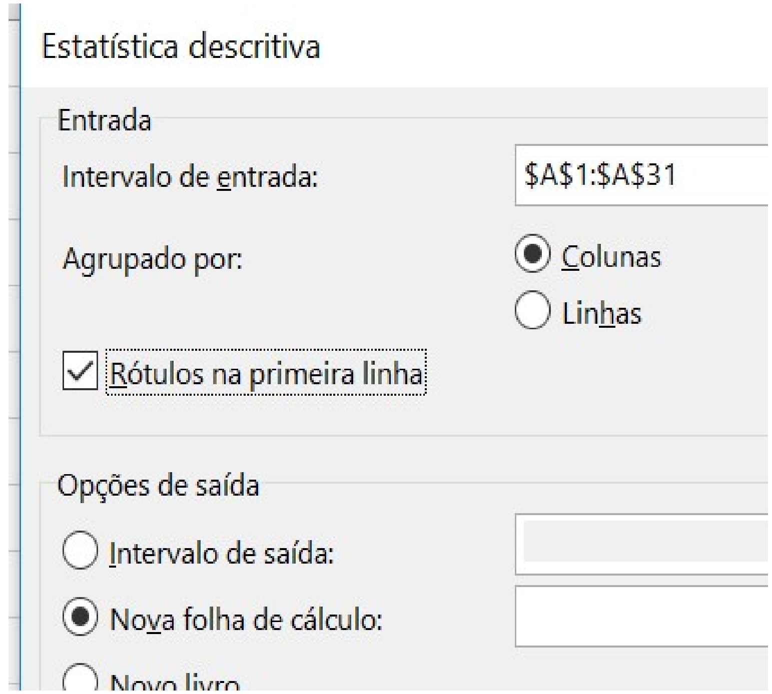
Coeficiente de variação: $cv = cd \times 100\%$

No Excel, há a necessidade de instalados o suplemento Analysis ToolPak: `Menu Ficheiro>opções>suplementos> ir>` clicar em ok. Aparecerá no canto superior direito a opção Análise de Dados



Selecionado a opção estatística descritiva:





Cofinanciado por:

UNIÃO EUROPEIA
Fundo Social Europeu