IPS Instituto Politécnico de Setúbal
Escola Superior de Tecnologia de Setúbal

MÉTODOS ESTATÍSTICOS

Estatística Descritiva

Licenciatura em Engenharia Informática

Departamento de Matemática
Escola Superior de Tecnologia de Setúbal
Instituto Politécnico de Setúbal
2020-2021

Estatística Descritiva

- Conceitos básicos.
- Dados qualitativos e quantitativos.
- Organização e interpretação de dados.
- Medidas de localização e dispersão.

Conceitos Básicos

O que é a Estatística?

De uma forma simples, é a ciência que se ocupa da recolha e tratamento de informação.
Também pode ser definida como a ciência que se ocupa da obtenção de **amostras**, da sua descrição e interpretação e, com apoio da teoria das probabilidades, permite efetuar inferências para a **população** e previsões do fenómeno em estudo.

E o que é a Estatística Descritiva?

Corresponde à parte da Estatística que se ocupa da descrição e interpretação de um conjunto de dados.

Conceitos Básicos

População (ou universo)

Conjunto de objetos (pessoas, resultados experimentais, ...) com uma ou mais características comuns, que se pretendem estudar. A população poderá ser finita ou infinita. Aos elementos da população chamam-se **Unidades Estatísticas**.

Amostra

Subconjunto de dados que pertencem à população. Parte da população que é observada com o objectivo de obter informação para estudar a característica pretendida. Estudam-se amostras para tirar conclusões para a população.

Conceitos Básicos

Por que não estudamos sempre toda a População?

À exceção dos casos em que a população tem dimensão "modesta" , raramente é possível analisar todos os elementos da população devido a

- Economia de tempo
- Redução de custos
- Características da população:
  - ser infinita,
  - possibilidade de destruição do objeto de experimentação

Em geral o estudo tem de ser feito sobre um subconjunto da População: a **amostra**. No entanto, não esquecer que uma amostra deverá ser representativa da população, ampla e imparcial. Nem sempre se recolhem amostras que verifiquem estas características, e consequentemente, pode-se tirar conclusões erradas.

Conceitos Básicos

População (ou universo)

Conjunto de objetos (pessoas, resultados experimentais, ...) com uma ou mais características comuns, que se pretendem estudar. A população poderá ser finita ou infinita. Aos elementos da população chamam-se **Unidades Estatísticas**.

Amostra

Subconjunto de dados que pertencem à população. Parte da população que é observada com o objetivo de obter informação para estudar a característica pretendida. Estudam-se amostras para tirar conclusões para a população.

Variável Estatística

Propriedade ou característica que se pretende estudar numa população.

Dado Estatístico

É cada um dos valores observados da variável em estudo.

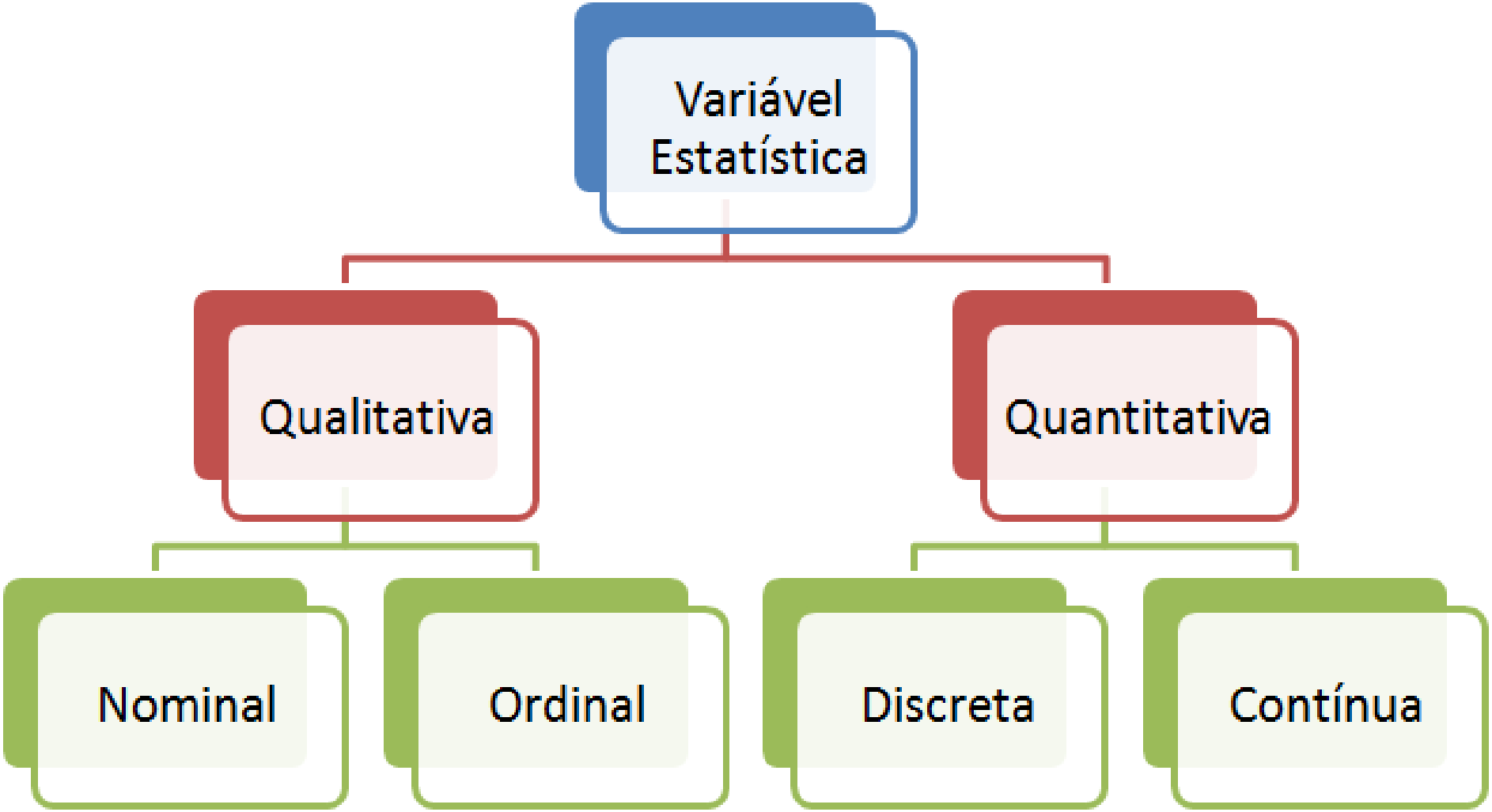
Conceitos Básicos

Exemplos

- **População:** O conjunto dos alunos da ESTSetúbal
- **Unidade Estatística associadas à População:** Alunos da ESTSetúbal
- **Possível Variável de interesse na Unidade Estatísticas:** Número de aprovações dos alunos da ESTSetúbal

- **População:** O conjunto de todas as empresas portuguesas
- **Unidade Estatística associadas à População:** Empresas portuguesas
- **Possível Variável de interesse na Unidade Estatísticas:** Lucro das empresas portuguesas

- **População:** O conjunto de todas as pessoas com uma dada doença
- **Unidade Estatística associadas à População:** Pessoas com uma dada doença
- **Possível Variável de interesse na Unidade Estatísticas:** Idade da pessoa com uma dada doença



Como os dados estatísticos correspondem aos valores observados da variável em estudo, então a sua classificação é idêntica à das variáveis: **Dados Qualitativos e Quantitativos**

Dados Qualitativos e Quantitativos

Dados Qualitativos

Representam a informação que identifica alguma qualidade, categoria ou característica, não suscetível de medida, mas de classificação. Registam-se numa escala:

- **Nominal** - a ordem das categorias não tem significado
  - **Exemplos:**
    - Sexo: feminino, masculino
    - Cor dos olhos: pretos, castanhos, azuis,...
    - Grupo sanguíneo: O-, O+, A-, A+, B-, B+, AB-, AB+
- **Ordinal** - há uma ordem natural das categorias
  - **Exemplos:**
    - Nível de escolaridade: 1º ciclo, 2º ciclo, 3º ciclo,...
    - Classe social: baixa, média, alta
    - Fases de uma doença: inicial, intermédio, terminal

Dados Qualitativos e Quantitativos

Observações:

- Muitas vezes os dados qualitativos (nominais ou ordinais) podem ser representados numericamente, isto é, são associados valores numéricos às diferentes categorias. No entanto esses valores numéricos não têm qualquer significado.

- Por exemplo, é possível associar os valores 1 e 2 às categorias masculino e feminino da variável sexo. Ou os valores 1, 2 e 3 às categorias baixo, médio e alto da variável classe social. Mas estes números são apenas símbolos para representar as categorias.

- Quando os dados são qualitativos do tipo ordinal, a numeração é feita de forma a respeitar a ordem.

Dados Qualitativos e Quantitativos

Dados Quantitativos

Representam a informação resultante de características suscetíveis de serem medidas, apresentando-se com diferentes intensidades. Registam-se numa escala:

- **Discreta** - os valores podem ordenar-se, mas entre dois valores consecutivos não pode existir um valor intermédio (ou seja, o domínio da variável é um conjunto finito ou infinito numerável) - associado a contagens
  - **Exemplos:**
    - número de letras no nome,
    - número de irmãos,
    - número de cigarros fumados por dia.

- **Contínua** - pode tomar qualquer valor num certo intervalo (ou seja, o seu domínio é um conjunto de números reais) - associado a medições.
  - **Exemplos:**
    - tempos efetuados por um atleta para correr os 100 metros,
    - altura das pessoas,
    - peso de objetos,
    - pressão arterial dos indivíduos.

Dados Qualitativos e Quantitativos

Observações:

- Os dados quantitativos são valores numéricos e estes números têm significado.
- Os dados originalmente podem ser quantitativos, mas podem ser recolhidos de forma qualitativa.

  Por exemplo, a variável idade, medida em anos é quantitativa (contínua), mas, se for obtida apenas a faixa etária (0 a 5 anos, 6 a 10 anos, ...), é qualitativa (ordinal).

Exemplos

1. Numa escola com 1200 alunos selecionou-se ao acaso um grupo de 300 alunos para responderem a um inquérito sobre o grau de satisfação em relação ao refeitório dessa escola.
   - Indique:
     - a população e a sua dimensão;
     - a dimensão da amostra;
     - a variável em estudo, classificando-a.

2. Em relação às seguintes frases indique as variáveis estatísticas em estudo e classifique-as em qualitativas (nominal ou ordinal) ou quantitativas (discreta ou contínua).
   - Número de filhos dos casais residentes em Setúbal.
   - Cor do cabelo e idade dos alunos de uma escola.
   - Sexo, nível de escolaridade e classe social dos residentes em Setúbal.
   - Número de ações negociadas na bolsa de valores de Lisboa.
   - Salários dos funcionários do Instituto Politécnico de Setúbal.

Organização e interpretação de dados

Estatística Descritiva

Tem como objetivo sumariar e descrever os aspetos relevantes num conjunto de dados. Para atingir este objetivo recorre-se:

- a tabelas de modo a condensar os dados;
- a representações gráficas dos dados;
- ao cálculo de indicadores numéricos de localização e dispersão.

Organização e interpretação de dados

Estatística Descritiva

Tem como objetivo sumariar e descrever os aspetos relevantes num conjunto de dados. Para atingir este objetivo recorre-se:

- a tabelas de modo a condensar os dados;
- a representações gráficas dos dados;
- ao cálculo de indicadores numéricos de localização e dispersão.

Tabela de frequências

Tabelas de Frequências

- Numa tabela de frequências a informação é organizada, de um modo geral, em 3 colunas:
  1. Coluna dos valores ou modalidades que as variáveis podem assumir, caso sejam variáreis quantitativas ou qualitativas, respetivamente;
  2. Coluna das frequências absolutas;
  3. Coluna das frequências relativas.

- Podem, ainda, ser acrescentadas mais duas colunas, com as frequências acumuladas:
  4. Frequência absoluta acumulada;
  5. Frequência relativa acumulada.

Tabelas de Frequências

- **Frequência absoluta** de um valor x_i da variável é o número de vezes que esse valor foi observado. Representa-se habitualmente por n_i.
  - A soma das frequências absolutas é igual à dimensão da amostra (ou à dimensão da população, caso tenham sido recolhidos dados relativos a todos os indivíduos da população).

- **Frequência relativa** de um valor da variável é o quociente entre a frequência absoluta desse valor e o número n de elementos da população (ou da amostra). Representa-se habitualmente por f_i.
  - É sempre um número entre 0 e 1.
  - Pode ser expressa em percentagem desde que se multiplique o número obtido por 100.
  - A soma das frequências relativas é igual a 1.

Tabelas de Frequências

As Tabelas de Frequências constroem-se de maneira diferente, consoante o tipo de variável.

Assim temos Tabelas de Frequências para

- Variáveis Qualitativas ou Variáveis Quantitativas Discretas (com número pequeno de valores distintos)
- Variáveis Quantitativas Contínuas ou Variáveis Quantitativas Discretas (com número elevado de valores distintos) - Neste caso há a necessidade de agrupar os dados em classes

Tabela de Frequências

Variáveis qualitativas ou quantitativas discretas

Valor da variável	Frequências Absolutas	Frequências Relativas	Frequências Absolutas Acumuladas	Frequências Relativas Acumuladas
x_i	n_i	f_i	N_i	F_i
x_1	n_1	$f_1 = \frac{n_1}{n}$	$N_1 = n_1$	$F_1 = f_1$
x_2	n_2	$f_2 = \frac{n_2}{n}$	$N_2 = n_1 + n_2$	$F_2 = f_1 + f_2$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_k	n_k	$f_k = \frac{n_k}{n}$	$N_k = \sum_{i=1}^{k} n_i = n$	$F_k = \sum_{i=1}^{k} f_i = 1$
Total	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} f_i = 1$		

- Frequência absoluta (n_i) — número de observações iguais a x_i
- Frequência relativa (f_i) — fração do número total de observações iguais a x_i
- Frequência absoluta acumulada (N_i) — número de observações menores ou iguais a x_i
- Frequência relativa acumulada (F_i) — fração do número total de observações menores ou iguais a x_i

Tabela de Frequências

Exemplo

Foi observada uma amostra de 10 atletas do sexo feminino com idades compreendidas entre os 15 e os 20 anos, nas quais tinha sido diagnosticada anemia. Relativamente a cada uma das pacientes, durante a permanência numa unidade hospitalar, foi registada a seguinte informação:

Número de ordem	Dieta equilibrada	Intensidade dos treinos	número de suplementos alimentares (por semana)	Nível de ferro (mg)
1	Sim	Moderada	3	14, 3
2	Sim	Elevada	2	7, 8
3	Não	Baixa	5	27, 0
4	Sim	Moderada	6	11, 0
5	Sim	Elevada	6	9, 9
6	Não	Baixa	3	14, 5
7	Sim	Baixa	4	15, 4
8	Não	Baixa	4	20, 8
9	Não	Elevada	7	10, 5
10	Sim	Baixa	3	15, 9

1. Identifique as variáveis e classifique-as.
2. Construa as tabelas de frequências das variáveis qualitativas e quantitativas discretas.

Tabela de Frequências

Variáveis quantitativas contínuas

(ou variáveis quantitativas discretas com um número elevado de valores distintos)

- Neste caso há a necessidade de agrupar os dados em **classes**.
- Para construir classes de igual amplitude pode ser usada uma regra para determinar o número de classes: **a regra de Sturges**.

Regra de Sturges

Para organizar uma amostra, de dados contínuos, de dimensão n, pode considerar-se para número de classes o valor k, onde k é o menor inteiro tal que $2^k > n$.
Da inequação anterior pode deduzir-se o seguinte resultado:

$$k = \lfloor 1 + \log_2 n \rfloor = \left\lfloor 1 + \frac{\ln n}{\ln 2} \right\rfloor$$

onde n é o número de dados e $\lfloor a \rfloor$ representa a parte inteira de a.

Tabela de Frequências

Variáveis quantitativas contínuas

(ou variáveis quantitativas discretas com um número elevado de valores distintos)

- Para a formação das **classes**, uma estratégia possível consiste em:
  - determinar o máximo $(max(x_i))$ e o mínimo $(min(x_i))$ dos dados
  - a amplitude de cada classe é $h = \frac{max(x_i) - min(x_i)}{k}$, desta forma todas as classes têm a mesma amplitude.
  - Formar as classes como intervalos semiabertos, abertos à esquerda e fechados à direita (ou vice-versa), sendo o extremo esquerdo do primeiro intervalo o mínimo dos dados.

Observação:

- Quando os intervalos são abertos à esquerda e fechados à direita, o primeiro intervalo da tabela de frequências é fechado à esquerda e à direita.
- Quando os intervalos são fechados à esquerda e abertos à direita, o último intervalo da tabela de frequências pode ser fechado à direita caso o seu valor corresponda ao máximo dos dados.

Tabela de Frequências

Variáveis quantitativas contínuas

(ou variáveis quantitativas discretas com um número elevado de valores distintos)

- Uma vez escolhidas as classes, a construção da tabela de frequências é idêntica à considerada para dados discretos.
  1. Coluna das classes onde se indicam todas as classes definidas.
  2. Coluna das frequências absolutas.
  3. Coluna das frequências relativas.

- Podem, ainda, existir mais três colunas:
  4. Coluna do representante da classe (também se pode designar por marca da classe) e corresponde ao ponto médio de cada intervalo.
  5. Coluna das frequências absolutas acumuladas.
  6. Coluna das frequências relativas acumuladas.

Tabela de Frequências

Variáveis quantitativas contínuas

(ou variáveis quantitativas discretas com um número elevado de valores distintos)

Classe $\mathbf{c_i}$	Marca da classe $\mathbf{x'_i}$	Frequências Absolutas $\mathbf{n_i}$	Frequências Relativas $\mathbf{f_i}$	Frequências Absolutas Acumuladas $\mathbf{N_i}$	Frequências Relativas Acumuladas $\mathbf{F_i}$
$c_1 = [min(x_i); b_1]$ (com $b_1 = min(x_i) + h$)	$x'_1 = \frac{x_{min}+b_1}{2}$	n_1	$f_1 = \frac{n_1}{n}$	$N_1 = n_1$	$F_1 = f_1$
$c_2 =]b_1; b_2]$ (com $b_2 = b_1 + h$)	$x'_2 = \frac{b_1+b_2}{2}$	n_2	$f_2 = \frac{n_2}{n}$	$N_2 = n_1 + n_2$	$F_2 = f_1 + f_2$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$c_k =]b_{k-1}; b_k]$ (com $b_k = b_{k-1} + h$)	$x'_k = \frac{b_{k-1}+b_k}{2}$	n_k	$f_k = \frac{n_k}{n}$	$N_k = \sum_{i=1}^{k} n_i = n$	$F_k = \sum_{i=1}^{k} f_i = 1$
Total		$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} f_i = 1$		

- Frequência absoluta (n_i) — número de observações que pertencem à classe c_i
- Frequência relativa (f_i) — fração do número total de observações que pertencem à classe c_i
- Frequência absoluta acumulada (N_i) — número de observações menores que o extremo superior da classe c_i (menores ou iguais, se for a última classe e o intervalo estiver fechado)
- Frequência relativa acumulada (F_i) — fração do número total de observações menores que o extremo superior da classe c_i (menores ou iguais, se for a última classe e o intervalo estiver fechado)

Tabela de Frequências

Exemplo

Foi observada uma amostra de 10 atletas do sexo feminino com idades compreendidas entre os 15 e os 20 anos, nas quais tinha sido diagnosticada anemia. Relativamente a cada uma das pacientes, durante a permanência numa unidade hospitalar, foi registada a seguinte informação:

Número de ordem	Dieta equilibrada	Intensidade dos treinos	número de suplementos alimentares (por semana)	Nível de ferro (mg)
1	Sim	Moderada	3	$14,3$
2	Sim	Elevada	2	$7,8$
3	Não	Baixa	5	$27,0$
4	Sim	Moderada	6	$11,0$
5	Sim	Elevada	6	$9,9$
6	Não	Baixa	3	$14,5$
7	Sim	Baixa	4	$15,4$
8	Não	Baixa	4	$20,8$
9	Não	Elevada	7	$10,5$
10	Sim	Baixa	3	$15,9$

1. Construa a tabela de frequências da variável quantitativa.

Estatística Descritiva

Tem como objetivo sumariar e descrever os aspetos relevantes num conjunto de dados. Para atingir este objetivo recorre-se:

- a tabelas de modo a condensar os dados;
- a representações gráficas dos dados;
- ao cálculo de indicadores numéricos de localização e dispersão.

Estatística Descritiva

Tem como objetivo sumariar e descrever os aspetos relevantes num conjunto de dados. Para atingir este objetivo recorre-se:

- a tabelas de modo a condensar os dados;
- a representações gráficas dos dados;
- ao cálculo de indicadores numéricos de localização e dispersão.

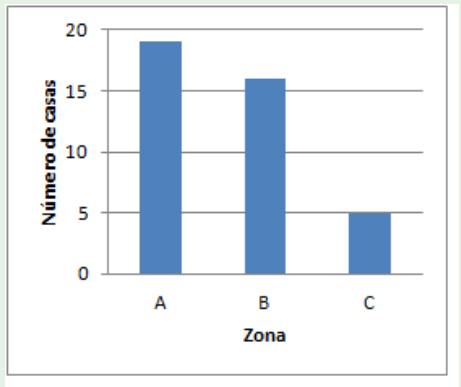
Gráficos

A principal vantagem dos gráficos, relativamente às tabelas, está na rapidez de leitura, pois permitem ter uma perceção imediata de quais as categorias de maior e menor frequência, assim como a ordem de grandeza de cada categoria relativamente às restantes.

Gráficos de Barras

- São muito usados para representar graficamente **dados qualitativos** ou **quantitativos discretos** (não agrupados).
- No eixo horizontal colocam-se as modalidades ou categorias da variável em estudo e no eixo vertical colocam-se as frequências absolutas ou relativas.
- Constrói-se uma barra para cada categoria (no caso da variável ser qualitativa) ou para cada valor assumido pela variável (no caso da variável ser discreta), sendo a altura de cada barra igual (ou proporcional) à respetiva frequência absoluta ou relativa.

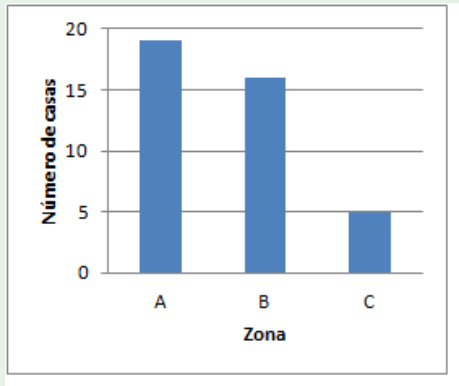
Zonas (x_i)	Número de casas (n_i)
A	19
B	16
C	5



Gráficos de Barras

- Ao contrário das alturas das barras, a largura das barras não transmite qualquer informação.
- As barras devem ter todas a mesma largura (pois barras mais largas podem chamar mais a atenção, induzindo em erro) e a distância entre as barras deve ser a mesma.
- A metodologia apresentada refere-se a gráficos de barras verticais. Se trocar os eixos, então tem-se um gráfico de barras horizontal.

Zonas (x_i)	Número de casas (n_i)
A	19
B	16
C	5



Gráficos de Barras

Exemplo 1

Os seguintes dados correspondem a respostas dadas por 30 pessoas sobre o desporto que praticam com mais frequência nos tempos livres:

Desporto (x_i)	Frequências absolutas (n_i)	Frequências relativas (f_i)
Andebol	2	0.067
Atletismo	6	0.200
Basquetebol	1	0.033
Futebol	9	0.300
Ténis	5	0.167
Voleibol	7	0.233

- Construa um gráfico de barras considerando as frequências absolutas.

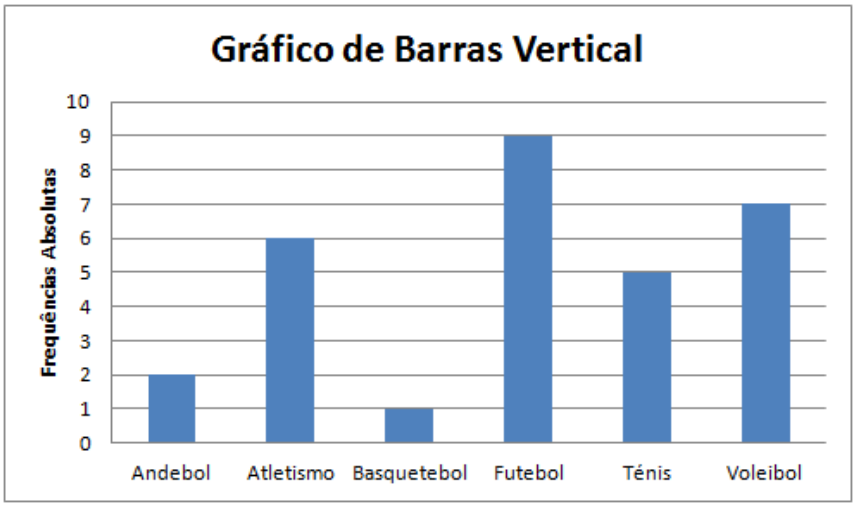
Gráficos de Barras

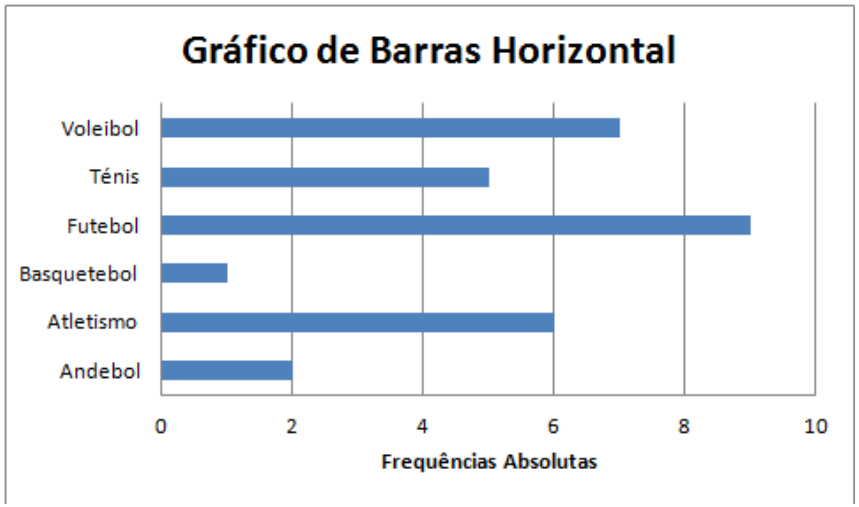
Exemplo 1

Os seguintes dados correspondem a respostas dadas por 30 pessoas sobre o desporto que praticam com mais frequência nos tempos livres:

Desporto (x_i)	Frequências absolutas (n_i)	Frequências relativas (f_i)
Andebol	2	0.067
Atletismo	6	0.200
Basquetebol	1	0.033
Futebol	9	0.300
Ténis	5	0.167
Voleibol	7	0.233

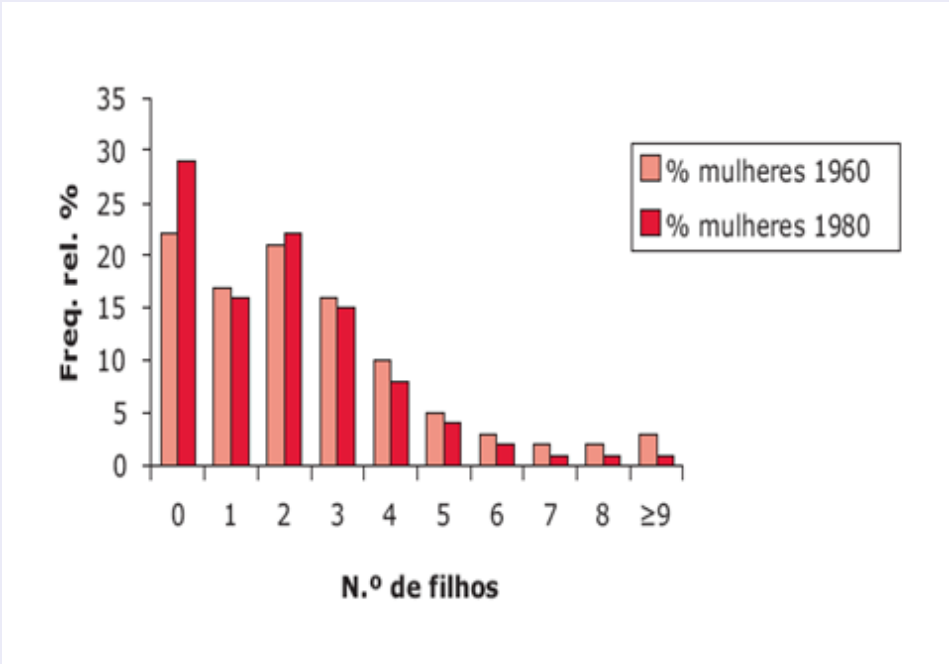
- Construa um gráfico de barras considerando as frequências absolutas.





Gráficos de Barras

- Para comparar diferentes conjuntos de dados, habitualmente, consideram-se gráficos de barras agrupadas.



- Se os conjuntos de dados têm dimensão diferente, devem usar-se gráficos de barras em que sejam representadas as frequências relativas (e não as frequências absolutas).

Gráficos de Barras

Exemplo 2

Os seguintes dados correspondem a respostas dadas por 30 pessoas de Lisboa e 50 pessoas do Porto sobre o desporto que praticam com mais frequência nos tempos livres:

Lisboa		
Desporto (x_i)	Freq. absolutas (n_i)	Freq. relativas (f_i)
Andebol	2	0.067
Atletismo	6	0.200
Basquetebol	1	0.033
Futebol	9	0.300
Ténis	5	0.167
Voleibol	7	0.233

Porto		
Desporto (x_i)	Freq. absolutas (n_i)	Freq. relativas (f_i)
Andebol	5	0.10
Atletismo	10	0.20
Basquetebol	9	0.18
Futebol	14	0.28
Ténis	5	0.10
Voleibol	7	0.14

Construa um gráfico de barras que permita comparar os dois conjuntos de dados.

Gráficos de Barras

Exemplo 2

Os seguintes dados correspondem a respostas dadas por 30 pessoas de Lisboa e 50 pessoas do Porto sobre o desporto que praticam com mais frequência nos tempos livres:

Lisboa		
Desporto (x_i)	Freq. absolutas (n_i)	Freq. relativas (f_i)
Andebol	2	0.067
Atletismo	6	0.200
Basquetebol	1	0.033
Futebol	9	0.300
Ténis	5	0.167
Voleibol	7	0.233

Porto		
Desporto (x_i)	Freq. absolutas (n_i)	Freq. relativas (f_i)
Andebol	5	0.10
Atletismo	10	0.20
Basquetebol	9	0.18
Futebol	14	0.28
Ténis	5	0.10
Voleibol	7	0.14

Construa um gráfico de barras que permita comparar os dois conjuntos de dados.

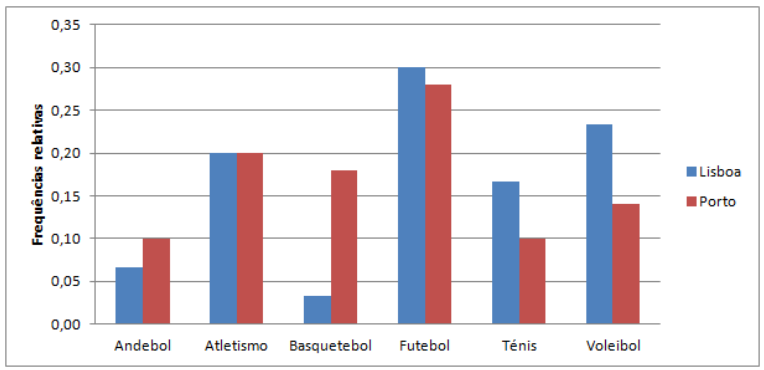


Diagrama Circular

- São mais usados para representar graficamente **dados qualitativos**.
- Esta representação é constituída por um círculo dividido em sectores.
- Tem tantos sectores circulares quantas as categorias ou classes consideradas na tabela de frequências.
- Podem mostrar as frequências absolutas, mas, em geral, apresentam as frequências relativas sob a forma de percentagens.
- O ângulo de cada sector circular é proporcional à frequência observada na modalidade que lhe corresponde, isto é, o ângulo do sector i é $f_i \times 360^o$.

Zonas (x_i)	Número de casas (n_i)	Frequências relativas (f_i)
A	19	0.475
B	16	0.400
C	5	0.125

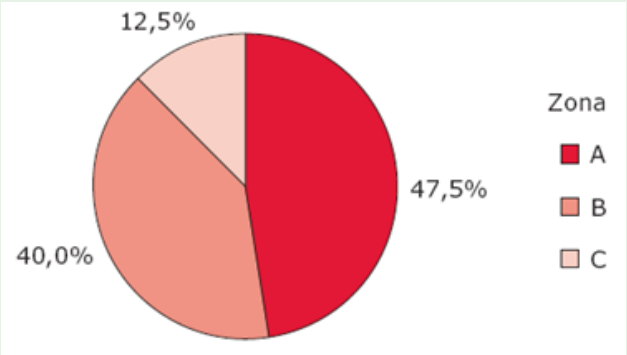


Diagrama Circular

Exemplo

Os seguintes dados correspondem ao número de vitórias, empates e derrotas de uma equipa desportiva durante um campeonato. Represente os dados recorrendo a um diagrama circular.

Resultados (x_i)	Frequências absolutas (n_i)	Frequências relativas (f_i)
vitória	10	0.40
empate	7	0.28
derrota	8	0.32
Total	25	1

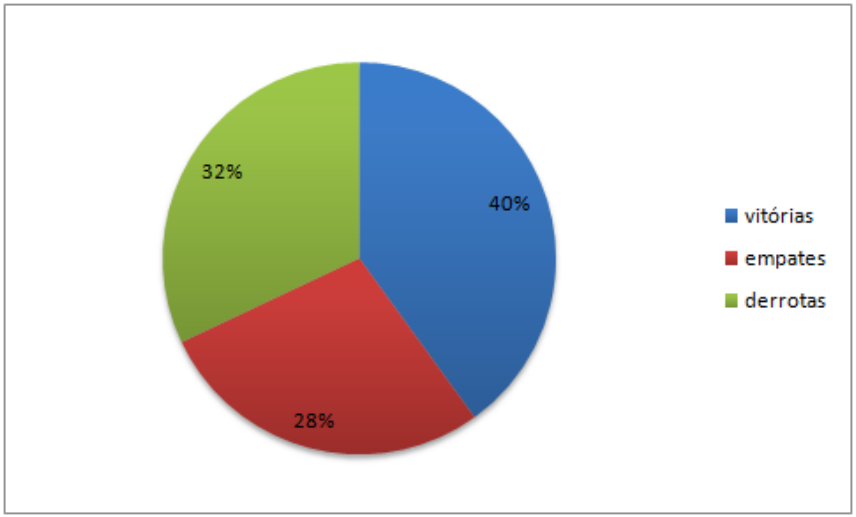
Diagrama Circular

Exemplo

Os seguintes dados correspondem ao número de vitórias, empates e derrotas de uma equipa desportiva durante um campeonato. Represente os dados recorrendo a um diagrama circular.

Resultados (x_i)	Frequências absolutas (n_i)	Frequências relativas (f_i)
vitória	10	0.40
empate	7	0.28
derrota	8	0.32
Total	25	1

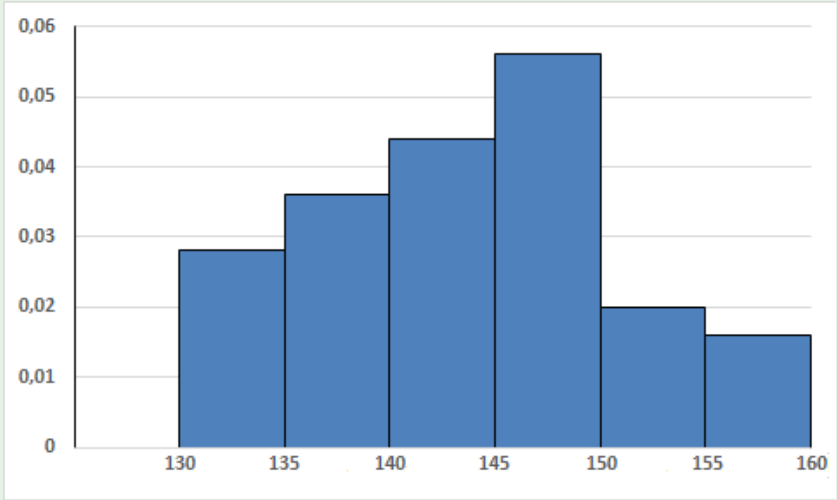
Resultados (x_i)	Amplitude do ângulo $(f_i \times 360^o)$
vitória	144
empate	100.8
derrota	115.2
Total	360



Histogramas

- São usados para representar graficamente **dados quantitativos contínuos** (ou dados discretos quando o número de valores distintos é muito elevado e são agrupados em classes).
- É um diagrama formado por uma sucessão de retângulos adjacentes:
  - a base de cada retângulo representa uma classe,
  - a altura de cada retângulo representa a frequência (relativa ou absoluta) com que os valores dessa classe ocorreram no conjunto de dados.
- Nesta construção dos histogramas estamos a supor que todas as classes têm a mesma amplitude.

Classe (c_i)	F. Relativa (f_i)
$[130, 135]$	0.14
$]135, 140]$	0.18
$]140, 145]$	0.22
$]145, 150]$	0.28
$]150, 155]$	0.10
$]155, 160]$	0.08



Histogramas

Exemplo 1

Uma empresa decidiu fazer um estudo sobre a idade (em anos) dos seus empregados. Para tal recolheu uma amostra de dimensão 70, tendo construído a seguinte tabela de frequências:

Classe	Frequência Absoluta (n_i)	Frequência Relativa (f_i)
$[25, 35]$	30	0.43
$]35, 45]$	20	0.29
$]45, 55]$	11	0.16
$]55, 65]$	9	0.13
Total	70	1

Construa um histograma usando as frequências relativas.

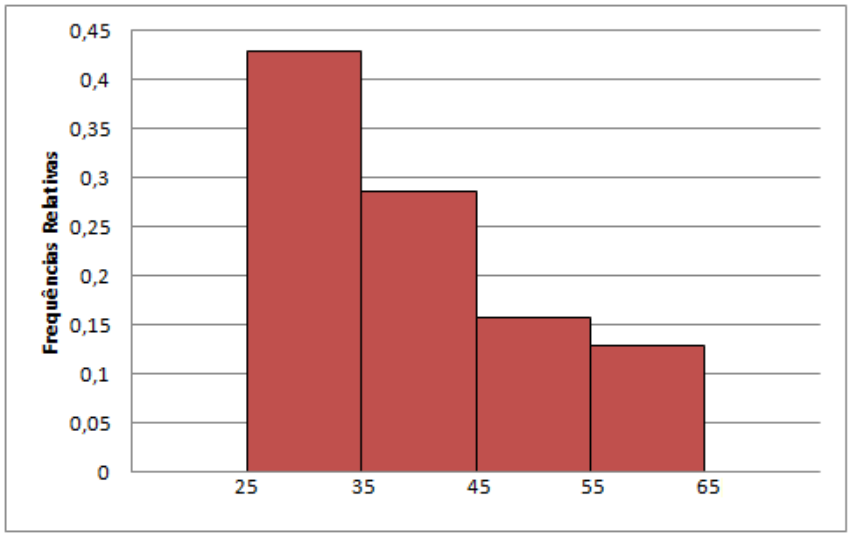
Histogramas

Exemplo 1

Uma empresa decidiu fazer um estudo sobre a idade (em anos) dos seus empregados. Para tal recolheu uma amostra de dimensão 70, tendo construído a seguinte tabela de frequências:

Classe	Frequência Absoluta (n_i)	Frequência Relativa (f_i)
$[25, 35]$	30	0.43
$]35, 45]$	20	0.29
$]45, 55]$	11	0.16
$]55, 65]$	9	0.13
Total	70	1

Construa um histograma usando as frequências relativas.



Histogramas

- Se se pretender comparar vários conjuntos de dados através de histogramas, deve-se ter o cuidado de os construir de modo a que a **área** total seja unitária, para ser possível a comparação.

Exemplo 2

Uma empresa decidiu fazer um estudo sobre a idade (em anos) dos seus empregados. Para tal recolheu uma amostra de dimensão 70 da empresa A e uma amostra de dimensão 50 da empresa B, tendo construído as seguintes tabelas de frequências:

Empresa A	
Classe	Frequência Absoluta (n_i)
$[25, 35]$	30
$]35, 45]$	20
$]45, 55]$	11
$]55, 65]$	9
Total	70

Empresa B	
Classe	Frequência Absoluta (n_i)
$[25, 35]$	25
$]35, 45]$	10
$]45, 55]$	11
$]55, 65]$	4
Total	50

Construa dois histogramas de modo a comparar os resultados.

Histogramas

Exemplo 2

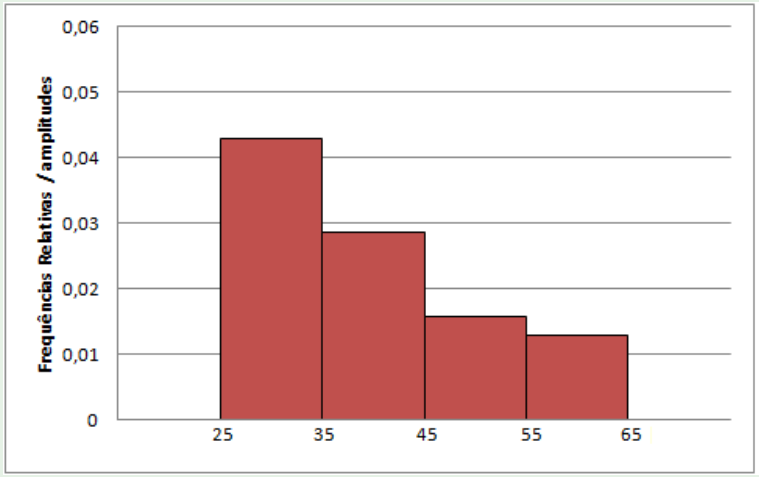
- É necessário calcular as frequências relativas.
- Cada retângulo deve ter **área** igual à frequência relativa (deste modo a área total será unitária). Ou seja:
  - a base de cada retângulo representa uma classe,
  - a altura de cada retângulo representa o quociente entre a frequência relativa e a amplitude da classe.

Empresa A				
Classe	F. Absoluta (n_i)	F. Relativa (f_i)	Base do retângulo (amplitude do intervalo $= a_i$)	Altura do retângulo $\left(\frac{f_i}{a_i}\right)$
$[25, 35]$	30	0.43	$35 - 25 = 10$	0.043
$]35, 45]$	20	0.29	$45 - 35 = 10$	0.029
$]45, 55]$	11	0.16	$55 - 45 = 10$	0.016
$]55, 65]$	9	0.13	$65 - 55 = 10$	0.013
Total	70	1		

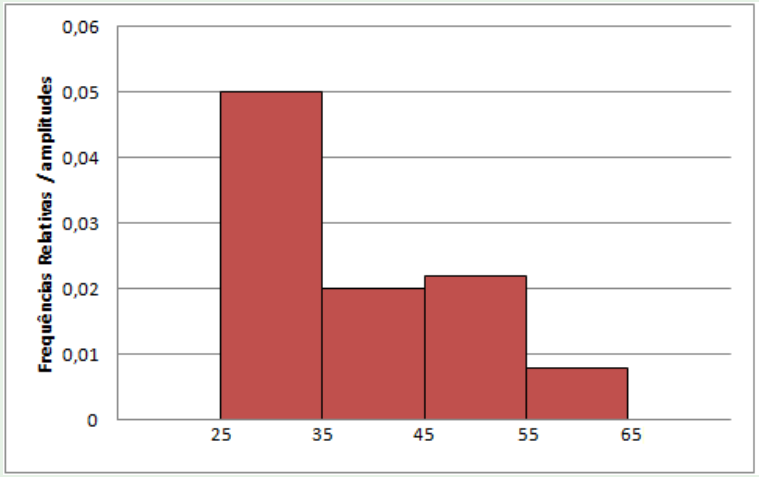
Empresa B				
Classe	F. Absoluta (n_i)	F. Relativa (f_i)	Base do retângulo (amplitude do intervalo $= a_i$)	Altura do retângulo $\left(\frac{f_i}{a_i}\right)$
$[25, 35]$	25	0.50	$35 - 25 = 10$	0.050
$]35, 45]$	10	0.20	$45 - 35 = 10$	0.020
$]45, 55]$	11	0.22	$55 - 45 = 10$	0.022
$]55, 65]$	4	0.08	$65 - 55 = 10$	0.008
Total	50	1		

Histogramas

Exemplo 2



Empresa A



Empresa B

Observação

- Neste exemplo, como as classes das tabelas de frequências são iguais (com a mesma amplitude, o mesmo número de classes e os mesmos limites nas classes), os histogramas podiam ter sido construídos recorrendo à frequência relativa e não era necessário impor que a área fosse 1.
- Quando as classes não têm todas a mesma amplitude, então os histogramas devem ser construídos de modo a que a área total seja 1 (mesmo que o objetivo não seja o de comparação).

Estatística Descritiva

Tem como objetivo sumariar e descrever os aspetos relevantes num conjunto de dados. Para atingir este objetivo recorre-se:

- a tabelas de modo a condensar os dados;
- a representações gráficas dos dados;
- ao cálculo de indicadores numéricos de localização e dispersão.

Estatística Descritiva

Tem como objetivo sumariar e descrever os aspetos relevantes num conjunto de dados. Para atingir este objetivo recorre-se:

- a tabelas de modo a condensar os dados;
- a representações gráficas dos dados;
- ao cálculo de indicadores numéricos de localização e dispersão.

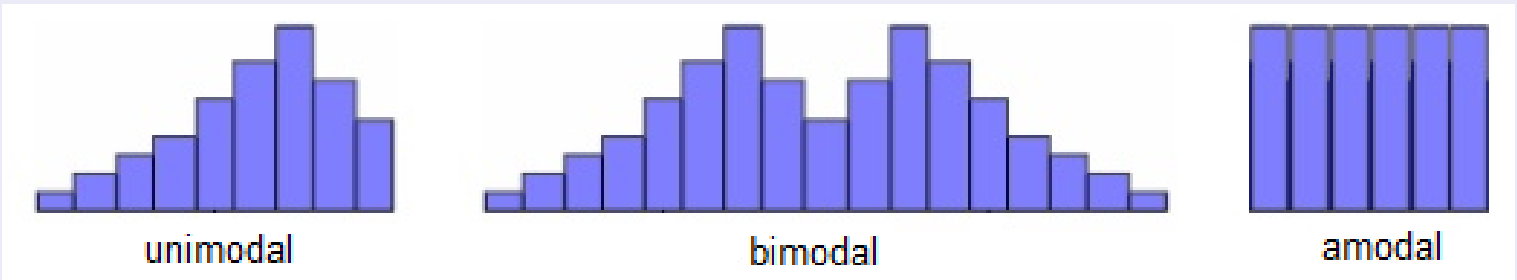
Medidas de Localização

Permitem resumir os dados calculando algumas características numéricas de modo a ter informação sobre a sua localização:

- Medidas de **localização central**:
  - moda,
  - média
  - mediana.
- Medidas de **localização não central**:
  - quantis.

Moda

- Habitualmente representa-se por mo.
- Para dados não agrupados, a moda define-se como o valor mais frequente.
- Para dados agrupados em classes (todas as classes com a mesma amplitude), a classe com maior frequência diz-se a **classe modal**.
- Um conjunto de dados pode não ter moda e diz-se **amodal**.
- Um conjunto de dados pode ter mais que uma moda. Isto acontece quando há dois ou mais valores que têm a maior frequência e diz-se
  - **bimodal** se tem duas modas;
  - **multimodal** ou **plurimodal** se tem mais do que duas modas.



Moda: Exemplos

Determine a moda.

1.

Naturalidade (x_i)	**Número de pessoas** (n_i)
Lisboa	72
Setúbal	42
Coimbra	31
Porto	50

Moda: Exemplos

Determine a moda.

1.

Naturalidade (x_i)	**Número de pessoas** (n_i)
Lisboa	72
Setúbal	42
Coimbra	31
Porto	50

A moda é Lisboa.

2.

14	14	12	13	15	15	16	11	10	9

Moda: Exemplos

Determine a moda.

1.

Naturalidade (x_i)	**Número de pessoas** (n_i)
Lisboa	72
Setúbal	42
Coimbra	31
Porto	50

A moda é Lisboa.

2.

14	14	12	13	15	15	16	11	10	9

Valores observados (x_i)	Frequência absoluta (n_i)
9	1
10	1
11	1
12	1
13	1
14	2
15	2
16	1

A moda é o 14 e o 15 $\rightarrow$ é bimodal

Moda: Exemplos

3. | 14 | 12 | 13 | 15 | 16 | 11 | 10 | 9 |

Moda: Exemplos

3. | 14 | 12 | 13 | 15 | 16 | 11 | 10 | 9 |

Valores observados (x_i)	Frequência absoluta (n_i)
9	1
10	1
11	1
12	1
13	1
14	1
15	1
16	1

Não tem moda.
Não há nenhum valor que seja mais frequente.
↓
é amodal

4. | 14 | 14 | 12 | 13 | 15 | 15 | 16 | 11 | 10 | 13 |

Moda: Exemplos

3.

14	12	13	15	16	11	10	9

Valores observados (x_i)	Frequência absoluta (n_i)
9	1
10	1
11	1
12	1
13	1
14	1
15	1
16	1

Não tem moda.
Não há nenhum valor que seja mais frequente.
↓
é amodal

4.

14	14	12	13	15	15	16	11	10	13

Valores observados (x_i)	Frequência absoluta (n_i)
10	1
11	1
12	1
13	2
14	2
15	2
16	1

A moda é o 13, o 14 e o 15
↓
é multimodal

Moda

Dados agrupados em classes (todas as classes com a mesma amplitude)

- A classe com maior frequência diz-se a **classe modal**.
- Ao valor central da classe modal chama-se **valor modal**. O **valor modal** é um possível valor para a moda (a chamada moda bruta).
- Outra forma de calcular um valor aproximado para a moda é a chamada **moda pelo método de King**:

$$\text{moda} \approx x_i^{\min} + (x_i^{\max} - x_i^{\min}) \times \frac{f_{i+1}}{f_{i-1} + f_{i+1}}$$

onde $x_i^{\min}$ é o limite inferior da classe modal, $x_i^{\max}$ é o limite superior da classe modal, f_{i-1} e f_{i+1} são as frequências relativas, respetivamente, da classe anterior e da classe posterior à classe modal.

Moda: Exemplos

5. Calcule valores aproximados para a moda.

Classe	Frequência Absoluta (n_i)	Frequência Relativa (f_i)
$[130, 135]$	7	0.14
$]135, 140]$	9	0.18
$]140, 145]$	11	0.22
$]145, 150]$	14	0.28
$]150, 155]$	5	0.10
$]155, 160]$	4	0.08
Total	50	1

Moda: Exemplos

5. Calcule valores aproximados para a moda.

Classe	Frequência Absoluta (n_i)	Frequência Relativa (f_i)
$[130, 135]$	7	0.14
$]135, 140]$	9	0.18
$]140, 145]$	11	0.22
$]145, 150]$	14	0.28
$]150, 155]$	5	0.10
$]155, 160]$	4	0.08
Total	50	1

- A **classe modal** é $]145, 150]$.
- Um valor aproximado da moda é o **valor modal** (moda bruta):

$$\text{moda} \approx \frac{150 + 145}{2} = 147.5$$

- Um outro valor aproximado da moda é a **moda pelo método de King**:

$$\text{moda} \approx 145 + (150 - 145) \times \frac{0.10}{0.22 + 0.10} = 146.56$$

Média

- Representa-se por $\overline{x}$ (quando os dados correspondem a uma amostra) ou por μ (quando os dados correspondem à população).
- A média é a medida de localização central mais utilizada, sendo muitas vezes usada como valor "representativo" de um conjunto de dados.
- A **média** define-se como o quociente entre a soma de todos os valores observados e o número de elementos da amostra.

  Isto é, seja $\{x_1, x_2, \ldots, x_n\}$ um conjunto de dados com n observações, define-se média aritmética, ou simplesmente **média**, como

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Não existe média quando a variável é qualitativa.

Média

Dados em Tabelas de Frequências

Não Estão Agrupados em Classes

$$\overline{x} = \begin{cases} \frac{n_1 \times x_1 + n_2 \times x_2 + \ldots + n_k \times x_k}{n} = \frac{\sum_{i=1}^{k} n_i x_i}{n} & \text{, frequências absolutas} \\ f_1 \times x_1 + f_2 \times x_2 + \ldots + f_k \times x_k = \sum_{i=1}^{k} f_i x_i & \text{, frequências relativas} \end{cases}$$

Dados em Tabelas de Frequências

Estão Agrupados em Classes

$$\overline{x} \approx \begin{cases} \frac{n_1 \times x'_1 + n_2 \times x'_2 + \ldots + n_k \times x'_k}{n} = \frac{\sum_{i=1}^{k} n_i x'_i}{n} & \text{, frequências absolutas} \\ f_1 \times x'_1 + f_2 \times x'_2 + \ldots + f_k \times x'_k = \sum_{i=1}^{k} f_i x'_i & \text{, frequências relativas} \end{cases}$$

Média

- Quando os dados estão agrupados em classes, o valor obtido para a média é um valor aproximado, uma vez que esta é calculada com base no representante da classe (x_i').
- A média de uma amostra apenas dá uma ideia da ordem de grandeza dos elementos da população, pois apenas é calculada com base nos elementos que foram incluídos na amostra.
- A média é muito sensível a valores extremos (muito grandes ou muito pequenos) dizendo-se por isso que é uma medida pouco resistente. Em alguns casos, a média pode não ser "representativa" de um conjunto de dados.

Média: Exemplos

Determine a média.

1. | 14 | 14 | 12 | 13 | 15 | 15 | 16 | 11 | 10 | 9 |

Média: Exemplos

Determine a média.

1. | 14 | 14 | 12 | 13 | 15 | 15 | 16 | 11 | 10 | 9 |

$$\overline{x} = \frac{14 + 14 + 12 + 13 + 15 + 15 + 16 + 11 + 10 + 9}{10} = 12.9$$

2.

Valores observados (x_i)	Frequência absoluta (n_i)	Frequência relativa (f_i)
45	3	0.091
47	10	0.303
50	7	0.212
53	10	0.303
54	3	0.091
Total	33	1

Média: Exemplos

Determine a média.

1. | 14 | 14 | 12 | 13 | 15 | 15 | 16 | 11 | 10 | 9 |

$$\overline{x} = \frac{14 + 14 + 12 + 13 + 15 + 15 + 16 + 11 + 10 + 9}{10} = 12.9$$

2.

Valores observados (x_i)	Frequência absoluta (n_i)	Frequência relativa (f_i)
45	3	0.091
47	10	0.303
50	7	0.212
53	10	0.303
54	3	0.091
Total	33	1

$$\overline{x} = \frac{3 \times 45 + 10 \times 47 + 7 \times 50 + 10 \times 53 + 3 \times 54}{33} = 49.909$$

ou

$$\overline{x} = 0.091 \times 45 + 0.303 \times 47 + 0.212 \times 50 + 0.303 \times 53 + 0.091 \times 54 = 49.909$$

Média: Exemplos

3

Classe	Frequência Absoluta (n_i)	Frequência Relativa (f_i)
$[130, 135]$	7	0.14
$]135, 140]$	9	0.18
$]140, 145]$	11	0.22
$]145, 150]$	14	0.28
$]150, 155]$	5	0.10
$]155, 160]$	4	0.08
Total	50	1

Média: Exemplos

3

Classe	Representante da classe (x'_i)	Frequência Absoluta (n_i)	Frequência Relativa (f_i)
$[130, 135]$	132.5	7	0.14
$]135, 140]$	137.5	9	0.18
$]140, 145]$	142.5	11	0.22
$]145, 150]$	147.5	14	0.28
$]150, 155]$	152.5	5	0.10
$]155, 160]$	157.5	4	0.08
Total		50	1

$$\overline{x} \approx \frac{7 \times 132.5 + 9 \times 137.5 + 11 \times 142.5 + 14 \times 147.5 + 5 \times 152.5 + 4 \times 157.5}{50} = 143.8$$

ou

$$\overline{x} \approx 0.14 \times 132.5 + 0.18 \times 137.5 + 0.22 \times 142.5 + 0.28 \times 147.5 + 0.10 \times 152.5 + 0.08 \times 157.5 = 143.8$$

Quantis

- Representa-se por Q_p.
- Dado um número $0 \leq p \leq 1$, define-se quantil de ordem p, Q_p, como o valor contido no intervalo de variação das observações tal que, pelo menos $p \times 100\%$ das observações são inferiores ou iguais a esse valor e pelo menos $(1-p) \times 100\%$ das observações são maiores ou iguais a esse valor.

Seja $\{x_1, x_2, \ldots, x_n\}$ um conjunto de dados com n observações e seja $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$ o mesmo **conjunto ordenado**.

$$Q_p = \begin{cases} \frac{x_{(np)} + x_{(np+1)}}{2} & , \text{ se } np \text{ é inteiro} \\ x_{(\lfloor np \rfloor + 1)} & , \text{ se } np \text{ não é inteiro} \end{cases}$$

onde $\lfloor a \rfloor$ representa a parte inteira de a.

Quantis

- Alguns quantis são muito usados e têm nomes específicos:
  - **Quartis** - dividem a amostra em 4 partes iguais
    - 1^oquartil $= Q_1 = Q_{0.25}$
    - 2^oquartil $= Q_2 = Q_{0.50} =$ mediana
    - 3^oquartil $= Q_3 = Q_{0.75}$
  - **Decis** - dividem a amostra em 10 partes iguais
    - 1^odecil $= D_1 = Q_{0.10}$
    - 2^odecil $= D_2 = Q_{0.20}$
    - $\ldots$
    - 8^odecil $= D_8 = Q_{0.80}$
    - 9^odecil $= D_9 = Q_{0.90}$
  - **Percentis** - dividem a amostra em 100 partes iguais
    - 1^opercentil $= P_1 = Q_{0.01}$
    - 2^opercentil $= P_2 = Q_{0.02}$
    - $\ldots$
    - 98^opercentil $= P_{98} = Q_{0.98}$
    - 99^opercentil $= P_{99} = Q_{0.99}$

Quantis

Mediana

- Um dos quantis mais importantes e mais utilizado em estatística é

$$2^o\text{quartil} = \text{Q}_2 = Q_{0.50} = \text{mediana}$$

- Habitualmente representa-se por $\widetilde{x}$ ou me.
- **A mediana** é o valor que ocupa a posição central quando se ordenam os dados estatísticos. Isto é, a mediana é o valor que separa as 50% das observações inferiores das 50% superiores. Por este motivo a mediana é considerada uma **medida de localização central**.
- A mediana é determinada pelo número de observações e não pelos seus valores, não sendo afetada por valores extremos. Diz-se, por isso, que é mais resistente do que a média.

Quando nos referimos aos quantis no geral, diz-se que são medidas de **localização não central** (a única exceção é a mediana que é uma medida de localização central).

Quantis

Exemplo 1

Calcule o 1^o Quartil, a mediana e o 3^o Quartil:

14	14	12	13	15	15	16	11	10	9	9	16

Quantis

Exemplo 1

Calcule o 1^o Quartil, a mediana e o 3^o Quartil:

14	14	12	13	15	15	16	11	10	9	9	16

- $n = 12$
- Ordenar os dados:

9	9	10	11	12	13	14	14	15	15	16	16

- 1^oquartil $= \mathrm{Q}_1 = Q_{0.25}$
  - $np = 12 \times 0.25 = 3$ é inteiro
  - $\mathrm{Q}_1 = Q_{0.25} = \frac{x_{(3)}+x_{(4)}}{2} = \frac{10+11}{2} = 10.5$
- mediana $= 2^o$quartil $= \mathrm{Q}_2 = Q_{0.50}$
  - $np = 12 \times 0.50 = 6$ é inteiro
  - $\widetilde{x} = \mathrm{Q}_2 = Q_{0.50} = \frac{x_{(6)}+x_{(7)}}{2} = \frac{13+14}{2} = 13.5$
- 3^oquartil $= \mathrm{Q}_3 = Q_{0.75}$
  - $np = 12 \times 0.75 = 9$ é inteiro
  - $\mathrm{Q}_3 = Q_{0.75} = \frac{x_{(9)}+x_{(10)}}{2} = \frac{15+15}{2} = 15$

Quantis

Exemplo 2

Calcule o 1^o Quartil, a mediana e o 3^o Quartil:

219	226	222	229	224	226	221	228	223	230	225

Quantis

Exemplo 2

Calcule o 1^o Quartil, a mediana e o 3^o Quartil:

219	226	222	229	224	226	221	228	223	230	225

- $n = 11$
- Ordenar os dados:

219	221	222	223	224	225	226	226	228	229	230

- 1^oquartil $= \mathrm{Q}_1 = Q_{0.25}$
  - $np = 11 \times 0.25 = 2.75$ não é inteiro
  - $\mathrm{Q}_1 = Q_{0.25} = x_{(\lfloor 2.75 \rfloor + 1)} = x_{(2+1)} = x_{(3)} = 222$
- mediana $= 2^o$quartil $= \mathrm{Q}_2 = Q_{0.50}$
  - $np = 11 \times 0.50 = 5.5$ não é inteiro
  - $\widetilde{x} = \mathrm{Q}_2 = Q_{0.50} = x_{(\lfloor 5.5 \rfloor + 1)} = x_{(5+1)} = x_{(6)} = 225$
- 3^oquartil $= \mathrm{Q}_3 = Q_{0.75}$
  - $np = 11 \times 0.75 = 8.25$ não é inteiro
  - $\mathrm{Q}_3 = Q_{0.75} = x_{(\lfloor 8.25 \rfloor + 1)} = x_{(8+1)} = x_{(9)} = 228$

Quantis

- Para determinar os quantis é necessário ordenar por ordem crescente as observações, pelo que não existem quantis quando a variável é qualitativa. No entanto há quem considere que é possível calcular quantis no caso da variável ser qualitativa ordinal.

- Os quantis são determinados pelo número de observações e não pelos seus valores, não sendo afetados por valores extremos.

Quantis

Dados em Tabelas de Frequências

Não Estão Agrupados em Classes

No caso dos **dados organizados numa tabela**, os quantis podem ser determinados a partir dos valores da frequência relativa acumulada.

- Se existir um valor com frequência relativa acumulada igual a p, o quantil é a média aritmética entre esse valor e o seguinte.

$$Q_p = \begin{cases} \frac{x_i + x_{i+1}}{2} & , \text{ para o valor } i, \text{ tal que } F_i = p \\ \end{cases}$$

Quantis

Dados em Tabelas de Frequências

Não Estão Agrupados em Classes

No caso dos **dados organizados numa tabela**, os quantis podem ser determinados a partir dos valores da frequência relativa acumulada.

- Se existir um valor com frequência relativa acumulada igual a p, o quantil é a média aritmética entre esse valor e o seguinte.
- Se não existir nenhum valor com frequência relativa acumulada igual a p, o quantil é o primeiro valor cuja frequência relativa acumulada ultrapassa p.

$$Q_p = \begin{cases} \frac{x_i + x_{i+1}}{2} & , \text{ para o valor } i, \text{ tal que } F_i = p \\ x_i & , \text{ para o menor valor } i, \text{ tal que } F_i > p \end{cases}$$

Quantis

Mediana: Exemplos

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.70
14	1

$$\widetilde{x} = \mathrm{Q}_2 = Q_{0.50} =$$

Quantis

Mediana: Exemplos

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.70
14	1

$$\widetilde{x} = \mathrm{Q}_2 = Q_{0.50} = 13$$

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.50
13	0.70
14	1

$$\widetilde{x} = \mathrm{Q}_2 = Q_{0.50} =$$

Quantis

Mediana: Exemplos

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.70
14	1

$$\widetilde{x} = \mathrm{Q}_2 = Q_{0.50} = 13$$

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.50
13	0.70
14	1

$$\widetilde{x} = \mathrm{Q}_2 = Q_{0.50} = \frac{12+13}{2} = 12.5$$

Quantis

1^o Quartil e 3^o Quartil: Exemplos

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.80
14	1

$$\mathrm{Q}_1 = Q_{0.25} =$$

Quantis

1^o Quartil e 3^o Quartil: Exemplos

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.80
14	1

$$\mathrm{Q}_1 = Q_{0.25} = 12$$

x_i	Frequências relativas acumuladas (F_i)
11	0.25
12	0.75
13	0.90
14	1

$$\mathrm{Q}_1 = Q_{0.25} =$$

Quantis

1^o Quartil e 3^o Quartil: Exemplos

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.80
14	1

$$\mathrm{Q}_1 = Q_{0.25} = 12$$

x_i	Frequências relativas acumuladas (F_i)
11	0.25
12	0.75
13	0.90
14	1

$$\mathrm{Q}_1 = Q_{0.25} = \frac{11+12}{2} = 11.5$$

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.80
14	1

$$\mathrm{Q}_3 = Q_{0.75} =$$

Quantis

1^o Quartil e 3^o Quartil: Exemplos

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.80
14	1

$$\mathrm{Q}_1 = Q_{0.25} = 12$$

x_i	Frequências relativas acumuladas (F_i)
11	0.25
12	0.75
13	0.90
14	1

$$\mathrm{Q}_1 = Q_{0.25} = \frac{11+12}{2} = 11.5$$

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.80
14	1

$$\mathrm{Q}_3 = Q_{0.75} = 13$$

x_i	Frequências relativas acumuladas (F_i)
11	0.25
12	0.75
13	0.90
14	1

$$\mathrm{Q}_3 = Q_{0.75} =$$

Quantis

1^o Quartil e 3^o Quartil: Exemplos

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.80
14	1

$$\mathrm{Q}_1 = Q_{0.25} = 12$$

x_i	Frequências relativas acumuladas (F_i)
11	0.25
12	0.75
13	0.90
14	1

$$\mathrm{Q}_1 = Q_{0.25} = \frac{11+12}{2} = 11.5$$

x_i	Frequências relativas acumuladas (F_i)
11	0.20
12	0.35
13	0.80
14	1

$$\mathrm{Q}_3 = Q_{0.75} = 13$$

x_i	Frequências relativas acumuladas (F_i)
11	0.25
12	0.75
13	0.90
14	1

$$\mathrm{Q}_3 = Q_{0.75} = \frac{12+13}{2} = 12.5$$

Quantis

Dados em Tabelas de Frequências

Estão Agrupados em Classes

- A primeira classe cuja a frequência relativa acumulada seja maior ou igual a p diz-se a **classe do quantil de ordem** p.
- Um valor aproximado dos quantis pode ser obtido como anteriormente, mas através dos representantes das classes:

$$Q_p \approx \begin{cases} \frac{x'_i + x'_{i+1}}{2} & , \text{ para a classe } i, \text{ tal que } F_i = p \\ x'_i & , \text{ para a menor classe } i, \text{ tal que } F_i > p \end{cases}$$

onde x'_i é o representante da classe do quantil de ordem p e x'_{i+1} é o representante da classe seguinte à classe do quantil de ordem p.

- **Outra possibilidade** para calcular um valor aproximado para os quantis é:

$$Qp \approx x_i^{\min} + (x_i^{\max} - x_i^{\min}) \times \frac{p - F_{i-1}}{f_i}$$

onde $x_i^{\min}$ é o limite inferior da classe do quantil de ordem p, $x_i^{\max}$ é o limite superior da classe do quantil de ordem p, F_{i-1} é a frequência relativa acumulada da classe anterior à classe do quantil de ordem p e f_i é a frequência relativa da classe do quantil de ordem p. (Esta fórmula só se aplica quando todas as classes têm a mesma amplitude.)

Exemplo 1

Calcule o 1^o Quartil, a mediana e o 3^o Quartil.

Classe	Frequência Absoluta (n_i)	Frequência Relativa (f_i)	Frequência Relativa Acumulada (F_i)
$[130, 135]$	7	0.14	0.14
$]135, 140]$	9	0.18	0.32
$]140, 145]$	11	0.22	0.54
$]145, 150]$	14	0.28	0.82
$]150, 155]$	5	0.10	0.92
$]155, 160]$	4	0.08	1
Total	50	1	

Quantis

Exemplo 1 (1^o quartil)

Classe	Frequência Absoluta (n_i)	Frequência Relativa (f_i)	Frequência Relativa Acumulada (F_i)	Representante da classe (x_i')
$[130, 135]$	7	0.14	0.14	$\frac{130+135}{2} = 132.5$
$]135, 140]$	9	0.18	0.32	$\frac{135+140}{2} = 137.5$
$]140, 145]$	11	0.22	0.54	$\frac{140+145}{2} = 142.5$
$]145, 150]$	14	0.28	0.82	$\frac{145+150}{2} = 147.5$
$]150, 155]$	5	0.10	0.92	$\frac{150+155}{2} = 152.5$
$]155, 160]$	4	0.08	1	$\frac{155+160}{2} = 157.5$
Total	50	1		

- $\mathrm{Q}_1 = Q_{0.25}$
- A **classe do 1º Quartil** é $]135, 140]$.
- Como a frequência relativa acumulada da classe do quantil de ordem $p = 0.25$ é $0.32 > 0.25$, um valor aproximado para o 1^o quartil é o representante da classe

$$\mathrm{Q}_1 = Q_{0.25} \approx 137.5$$

- Um outro valor aproximado do 1^o quartil pode ser calculada por

$$\mathrm{Q}_1 = Q_{0.25} \approx 135 + (140 - 135) \times \frac{0.25 - 0.14}{0.18} = 138.06$$

Quantis

Exemplo 1 (mediana)

Classe	Frequência Absoluta (n_i)	Frequência Relativa (f_i)	Frequência Relativa Acumulada (F_i)	Representante da classe $\left(x_i'\right)$
$[130, 135]$	7	0.14	0.14	$\frac{130+135}{2} = 132.5$
$]135, 140]$	9	0.18	0.32	$\frac{135+140}{2} = 137.5$
$]140, 145]$	11	0.22	0.54	$\frac{140+145}{2} = 142.5$
$]145, 150]$	14	0.28	0.82	$\frac{145+150}{2} = 147.5$
$]150, 155]$	5	0.10	0.92	$\frac{150+155}{2} = 152.5$
$]155, 160]$	4	0.08	1	$\frac{155+160}{2} = 157.5$
Total	50	1		

- $\widetilde{x} = \mathrm{Q}_2 = Q_{0.50}$
- A **classe da mediana** é $]140, 145]$.
- Como a frequência relativa acumulada da classe da mediana é $0.54 > 0.50$, um valor aproximado da mediana é o representante da classe

$$\widetilde{x} \approx 142.5$$

- Um outro valor aproximado da mediana pode ser calculado por

$$\widetilde{x} \approx 140 + 5 \times \frac{0.50 - 0.32}{0.22} = 144.09$$

Quantis

Exemplo 1 (3^o quartil)

Classe	Frequência Absoluta (n_i)	Frequência Relativa (f_i)	Frequência Relativa Acumulada (F_i)	Representante da classe (x_i')
$[130, 135]$	7	0.14	0.14	$\frac{130+135}{2} = 132.5$
$]135, 140]$	9	0.18	0.32	$\frac{135+140}{2} = 137.5$
$]140, 145]$	11	0.22	0.54	$\frac{140+145}{2} = 142.5$
$]145, 150]$	14	0.28	0.82	$\frac{145+150}{2} = 147.5$
$]150, 155]$	5	0.10	0.92	$\frac{150+155}{2} = 152.5$
$]155, 160]$	4	0.08	1	$\frac{155+160}{2} = 157.5$
Total	50	1		

- $\mathrm{Q}_3 = Q_{0.75}$
- A **classe do 3º Quartil** é $]145, 150]$.
- Como a frequência relativa acumulada da classe do quantil de ordem $p = 0.75$ é $0.82 > 0.75$, um valor aproximado para o 3^o quartil é o representante da classe

$$\mathrm{Q}_3 = Q_{0.75} \approx 147.5$$

- Um outro valor aproximado do 3^o quartil pode ser calculada por

$$\mathrm{Q}_3 = Q_{0.75} \approx 145 + (150 - 145) \times \frac{0.75 - 0.54}{0.28} = 148.75$$

Quantis

Observação

- Para dados agrupados em classes, se existir uma classe com frequência relativa acumulada igual à ordem p, os valores aproximados para os quantis (recorrendo às fórmulas anteriores) correspondem ao limite superior da classe do quantil de ordem p.

Exemplo 2

Classe	Frequências relativas acumuladas (F_i)
$[130, 135]$	0.25
$]135, 140]$	0.60
$]140, 145]$	0.80
$]145, 150]$	1

$$\mathrm{Q}_1 = Q_{0.25} \approx$$

Quantis

Observação

- Para dados agrupados em classes, se existir uma classe com frequência relativa acumulada igual à ordem p, os valores aproximados para os quantis (recorrendo às fórmulas anteriores) correspondem ao limite superior da classe do quantil de ordem p.

Exemplo 2

Classe	Frequências relativas acumuladas (F_i)
$[130, 135]$	0.25
$]135, 140]$	0.60
$]140, 145]$	0.80
$]145, 150]$	1

$$\mathrm{Q}_1 = Q_{0.25} \approx 135$$

Classe	Frequências relativas acumuladas (F_i)
$[130, 135]$	0.20
$]135, 140]$	0.50
$]140, 145]$	0.70
$]145, 150]$	1

$$\widetilde{x} = \mathrm{Q}_2 = Q_{0.50} \approx$$

Quantis

Observação

- Para dados agrupados em classes, se existir uma classe com frequência relativa acumulada igual à ordem p, os valores aproximados para os quantis (recorrendo às fórmulas anteriores) correspondem ao limite superior da classe do quantil de ordem p.

Exemplo 2

Classe	Frequências relativas acumuladas (F_i)
$[130, 135]$	0.25
$]135, 140]$	0.60
$]140, 145]$	0.80
$]145, 150]$	1

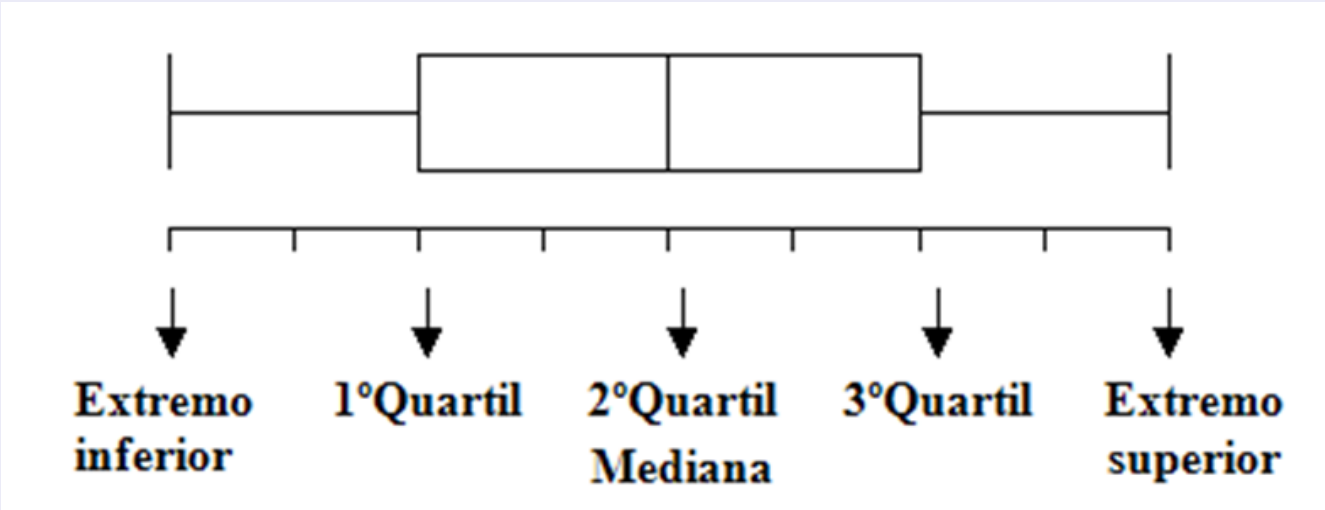
$$\mathrm{Q}_1 = Q_{0.25} \approx 135$$

Classe	Frequências relativas acumuladas (F_i)
$[130, 135]$	0.20
$]135, 140]$	0.50
$]140, 145]$	0.70
$]145, 150]$	1

$$\widetilde{x} = \mathrm{Q}_2 = Q_{0.50} \approx 140$$

Diagrama de extremos e quartis

- O diagrama de extremos e quartis (BoxPlot) é uma forma esquemática de representar uma distribuição por cinco dos seus valores estatísticos: extremo inferior (mínimo), 1^o quartil, mediana ou 2^o quartil, 3^o quartil e extremo superior (máximo).



- Ficam definidas quatro zonas: duas centrais representadas por retângulos e duas caudas. Em cada uma destas zonas está 25% dos dados.
- Quanto mais estreita for uma zona, maior é a concentração de dados aí existente. Por isso, este diagrama dá algumas indicações gerais sobre o tipo de distribuição.

Diagrama de extremos e quartis: Exemplo

Construa o diagrama de extremos e quartis:

40	53	60	72	65	54	60	92	48	87

Diagrama de extremos e quartis: Exemplo

Construa o diagrama de extremos e quartis:

40	53	60	72	65	54	60	92	48	87

- Ordenar os dados:

40	48	53	54	60	60	65	72	87	92

- extremo inferior = mínimo dos dados $= 40$
- extremo superior = máximo dos dados $= 92$
- $n = 10$
  - como $np = 10 \times 0.25 = 2.5$, então $Q_1 = Q_{0.25} = x_{(3)} = 53$
  - como $np = 10 \times 0.50 = 5$, então $Q_2 = Q_{0.50} = \widetilde{x} = \frac{x_{(5)} + x_{(6)}}{2} = \frac{60+60}{2} = 60$
  - como $np = 10 \times 0.75 = 7.5$, então $Q_3 = Q_{0.75} = x_{(8)} = 72$

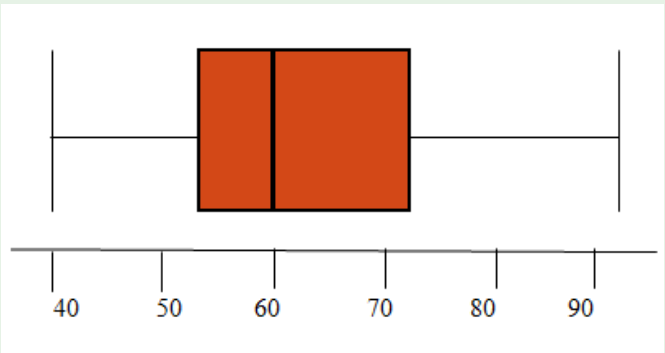
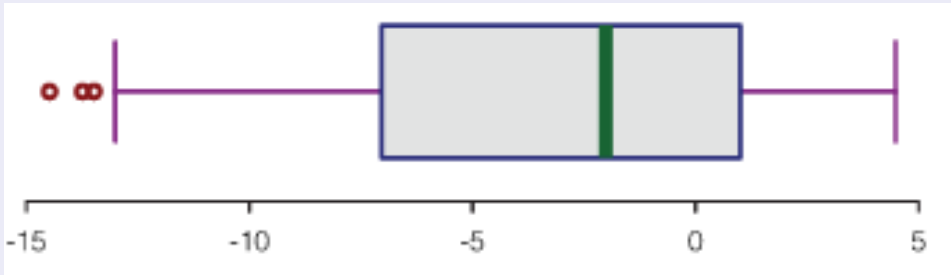


Diagrama de extremos e quartis

- No diagrama de extremos e quartis podemos identificar as observações que se afastam do padrão geral dos dados, os chamados "outliers" (observações discordantes) e representam-se por $*$ ou $\circ$.



- Existem vários critérios para classificar uma observação como um "outlier" :
  - Um valor x_i é um candidato a **"outlier" moderado** (habitualmente representa-se por $\circ$) se estiver
    - entre $Q_1 - 1.5 \times (Q_3 - Q_1)$ e $Q_1 - 3 \times (Q_3 - Q_1)$
      ou
    - entre $Q_3 + 1.5 \times (Q_3 - Q_1)$ e $Q_3 + 3 \times (Q_3 - Q_1)$.
  - Um valor x_i é um candidato a **"outlier" severo** (habitualmente representa-se por $*$) se
    - for maior que $Q_3 + 3 \times (Q_3 - Q_1)$
      ou
    - menor que $Q_1 - 3 \times (Q_3 - Q_1)$.

Diagrama de extremos e quartis: Exemplo

Construa o diagrama de extremos e quartis representando os "outliers" (caso existam).

22	14	23	6	20	21	55	22	25

Diagrama de extremos e quartis: Exemplo

Construa o diagrama de extremos e quartis representando os "outliers" (caso existam).

22	14	23	6	20	21	55	22	25

- Ordenar os dados:

6	14	20	21	22	22	23	25	55

- $n = 9$
  - como $np = 9 \times 0.25 = 2.25$, então $Q_1 = Q_{0.25} = x_{(3)} = 20$
  - como $np = 9 \times 0.50 = 4.5$, então $Q_2 = Q_{0.50} = \widetilde{x} = x_{(5)} = 22$
  - como $np = 9 \times 0.75 = 6.75$, então $Q_3 = Q_{0.75} = x_{(7)} = 23$

Diagrama de extremos e quartis: Exemplo

Dados ordenados:

6	14	20	21	22	22	23	25	55

- $Q_1 = 20$; $\widetilde{x} = 22$; $Q_3 = 23$
- limites dos "outliers" moderados:
  - $Q_1 - 1.5 \times (Q_3 - Q_1) = 20 - 1.5 \times (23 - 20) = 15.5$
  - $Q_3 + 1.5 \times (Q_3 - Q_1) = 23 + 1.5 \times (23 - 20) = 27.5$
- limites dos "outliers" severos:
  - $Q_1 - 3 \times (Q_3 - Q_1) = 20 - 3 \times (23 - 20) = 11$
  - $Q_3 + 3 \times (Q_3 - Q_1) = 23 + 3 \times (23 - 20) = 32$
- "outliers" moderados: 14
- "outliers" severos: 6 e 55
- a caixa só é construída com os dados: 20, 21, 22, 22, 23, 25
  - extremo inferior $= 20$
  - extremo superior $= 25$

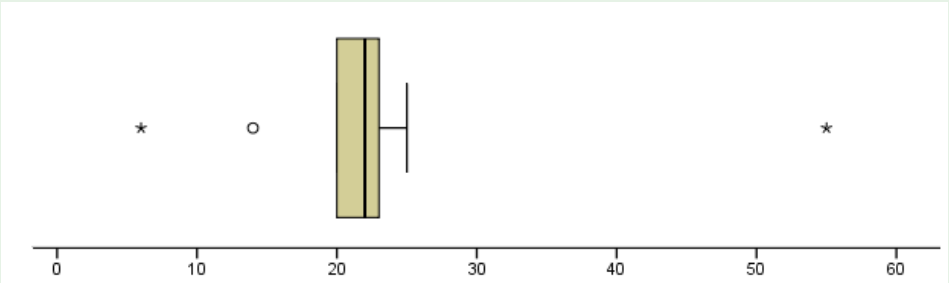
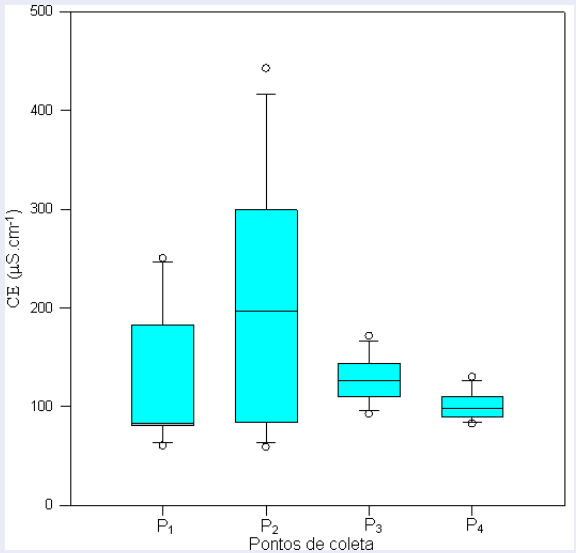
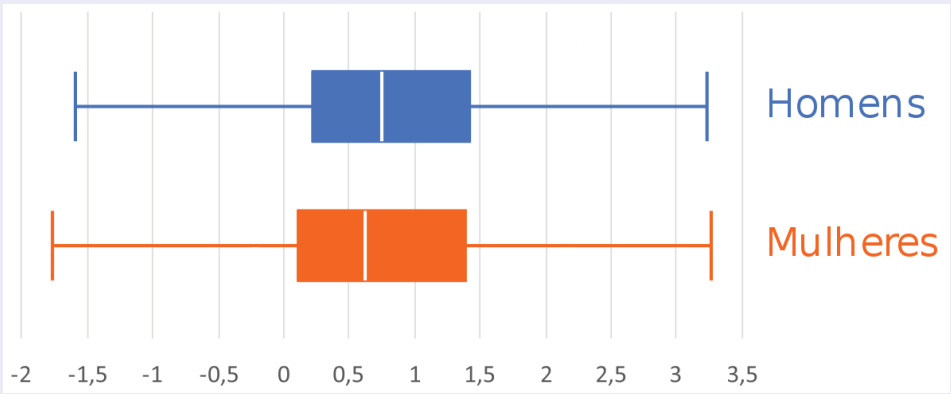


Diagrama de extremos e quartis

- Quando se pretende comparar várias amostras, o recurso a este tipo de diagramas, dispostos paralelamente, é uma ferramenta que permite, de forma fácil, obter uma primeira interpretação e comparação dos conjuntos de dados.



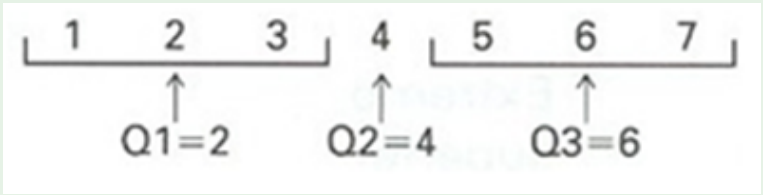


Quartis

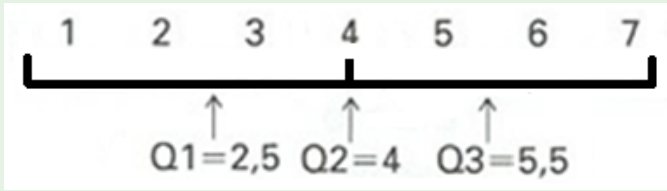
- Há vários métodos para calcular os quartis, nem todos conducentes aos mesmos valores, mas a valores próximos.
- Um dos métodos alternativos é dividir os dados de modo a calcular apenas a mediana:
  - Considerar os dados todos e calcular a mediana, $\widetilde{x} = \mathrm{Q}_2$, pois como vimos esta divide o conjunto de dados em duas partes iguais;
  - depois calcula-se a mediana dos dados que ficam à esquerda de $\widetilde{x} = \mathrm{Q}_2$ e obtém-se o 1^o quartil;
  - de seguida calcula-se a mediana dos dados que ficam à direita de $\widetilde{x} = \mathrm{Q}_2$ e obtém-se o 3^o quartil.
  - No caso da dimensão da amostra ser ímpar, a mediana coincide com um dos elementos da amostra. Este método permite duas opções:
    - considerar que a mediana fica incluída nas duas metades em que fica dividida a amostra
      ou
    - considerar que a mediana não fica incluída em nenhuma das metades em que fica dividida a amostra.

Quartis: Exemplos

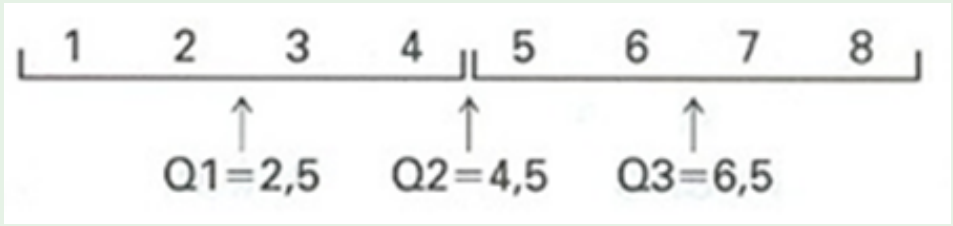
$n = 7$



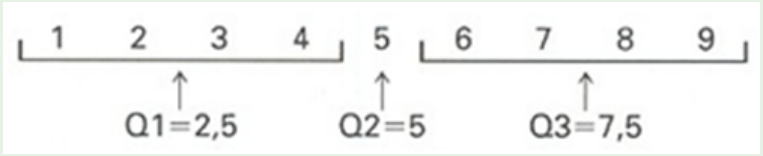
ou



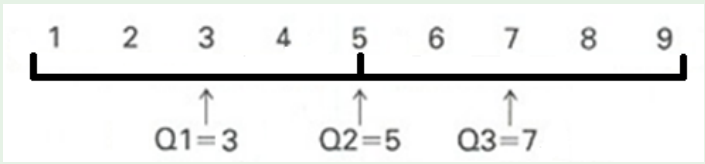
$n = 8$



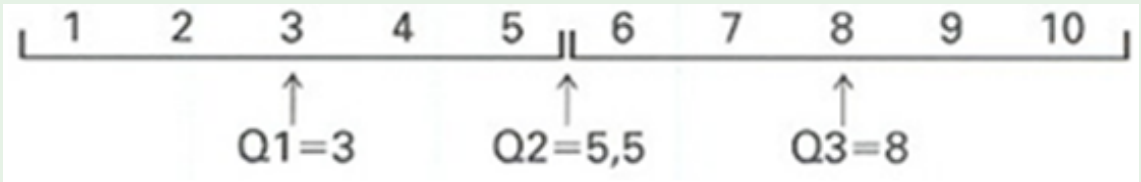
$n = 9$



ou

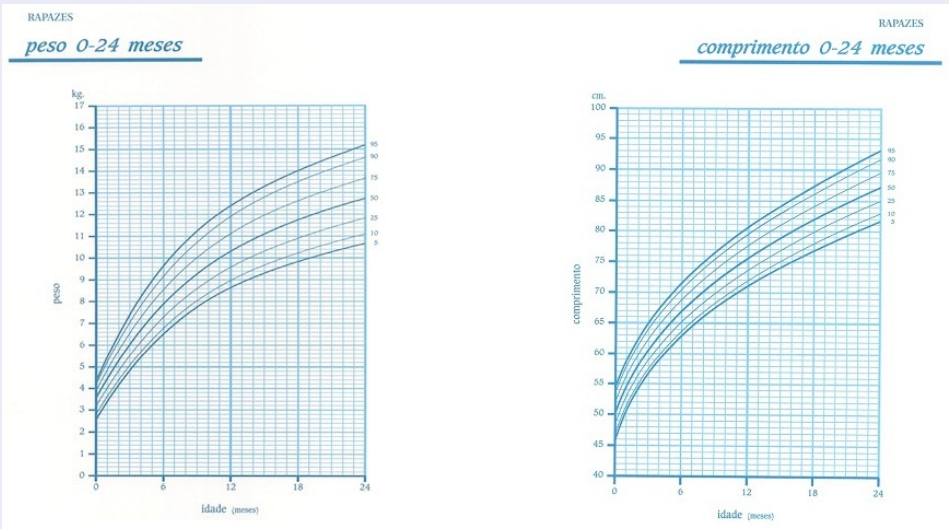


$n = 10$



Percentis

- Como já foi referido, os **Percentis** são quantis onde se divide a amostra em 100 partes iguais:
  - percentil $1 = P_1 = Q_{0.01}$
  - percentil $2 = P_2 = Q_{0.02}$
  - $\ldots$
  - percentil $99 = P_{99} = Q_{0.99}$
- Os percentis são muito usados pelos médicos quando se referem ao peso ou altura de um bebé.
- Existem gráficos e tabelas que apresentam, para cada idade, os valores dos percentis para as variáveis peso e altura:



Percentis

- Suponha a seguinte situação: "Os pais vão, com o seu bebé de 6 meses, à consulta de rotina, do pediatra. Este, depois de pesar e medir a criança, consulta uns gráficos e diz aos pais o seguinte:

  "O vosso filho, no que diz respeito ao peso, está no percentil 95. Vamos ter que ter algum cuidado!"
- Mas o que significa o percentil 95?
- Significa que 95% das crianças com 6 meses têm um peso menor ou igual ao do bebé e só 5% têm um peso maior ou igual, o que significa que um peso no percentil 95 é efetivamente exagerado.
- No entanto um bebé com o peso no percentil 95 e um comprimento no mesmo percentil ou num percentil próximo não terá excesso de peso. Se o comprimento estiver num percentil inferior, aí haverá já peso a mais para aquela estatura, que será tanto maior quanto menor o valor do percentil do comprimento.

Estatística Descritiva

Tem como objetivo sumariar e descrever os aspetos relevantes num conjunto de dados. Para atingir este objetivo recorre-se:

- a tabelas de modo a condensar os dados;
- a representações gráficas dos dados;
- ao cálculo de indicadores numéricos de localização e dispersão.

Estatística Descritiva

Tem como objetivo sumariar e descrever os aspetos relevantes num conjunto de dados. Para atingir este objetivo recorre-se:

- a tabelas de modo a condensar os dados;
- a representações gráficas dos dados;
- ao cálculo de indicadores numéricos de localização e dispersão.

Medidas de Dispersão

Permitem resumir os dados calculando algumas características numéricas de modo a ter informação sobre a sua a variabilidade ou dispersão:

- Medidas de **dispersão absoluta** (depende da unidade em que se exprime a variável):
  - amplitude: amplitude total e amplitude interquartis,
  - variância e desvio padrão.
- Medidas de **dispersão relativa** (não depende da unidade em que se exprime a variável):
  - coeficiente de variação.

Amplitude Total

- Habitualmente representa-se por **A**.
- A **Amplitude Total** é a medida mais simples para medir a variabilidade dos dados.

- Para dados não agrupados, a amplitude total define-se como a diferença entre o maior e o menor valor do conjunto de dados (diferença entre os extremos). Isto é, seja $\{x_1, x_2, \ldots, x_n\}$ um conjunto de dados com n observações,

$$\mathrm{A} = max(x_i) - min(x_i).$$

- Para dados agrupados em classes, a amplitude total é a diferença entre o limite superior da última classe e o limite inferior da primeira classe.

- É uma medida não negativa e será tanto maior quanto maior for a variabilidade dos dados.

Amplitude Total

Exemplos

Determine a amplitude.

1. | 14 | 14 | 12 | 13 | 15 | 15 | 16 | 11 | 10 | 9 |

Amplitude Total

Exemplos

Determine a amplitude.

1. | 14 | 14 | 12 | 13 | 15 | 15 | 16 | 11 | 10 | 9 |

   - Ordenar os dados:

     | 9 | 10 | 11 | 12 | 13 | 14 | 14 | 15 | 15 | 16 |

   - mínimo $= 9$
     máximo $= 16$
   - A $= 16 - 9 = 7$

2.

Classe	Frequência Absoluta (n_i)	Frequência Relativa (f_i)
$[130, 135]$	7	0.14
$]135, 140]$	9	0.18
$]140, 145]$	11	0.22
$]145, 150]$	14	0.28
$]150, 155]$	5	0.10
$]155, 160]$	4	0.08
Total	50	1

Amplitude Total

Exemplos

Determine a amplitude.

1. | 14 | 14 | 12 | 13 | 15 | 15 | 16 | 11 | 10 | 9 |

   - Ordenar os dados:

     | 9 | 10 | 11 | 12 | 13 | 14 | 14 | 15 | 15 | 16 |

   - mínimo $= 9$
     máximo $= 16$
   - A $= 16 - 9 = 7$

2.

Classe	Frequência Absoluta (n_i)	Frequência Relativa (f_i)
$[130, 135]$	7	0.14
$]135, 140]$	9	0.18
$]140, 145]$	11	0.22
$]145, 150]$	14	0.28
$]150, 155]$	5	0.10
$]155, 160]$	4	0.08
Total	50	1

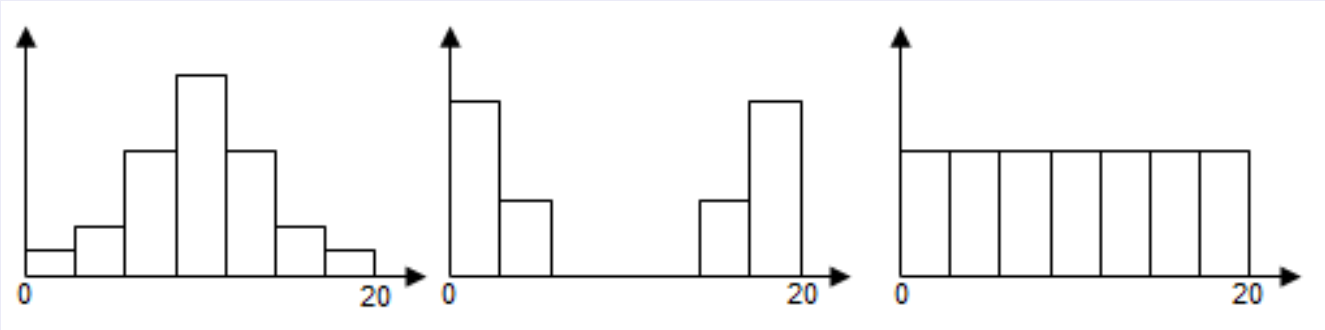
   - limite inferior da primeira classe $= 130$
     limite superior da última classe $= 160$
   - A $= 160 - 130 = 30$

Amplitude Total

Observações

- A amplitude é uma fraca medida de dispersão.

  Exemplo de distribuições com a mesma amplitude, mas com uma dispersão muito diferente:



- Desvantagem da amplitude enquanto medida de dispersão:
  - É insensível às alterações dos valores intermédios (nela só intervêm os extremos).
  - Não diz nada sobre o que se passa no intervalo entre os extremos. Em certas distribuições os valores extremos correspondem a casos excecionais e portanto pouco significativos.

Amplitude interquartis

- Habitualmente representa-se por AIQ.
- A amplitude interquartis define-se como a diferença entre o 3^o quartil e o 1^o quartil:

$$AIQ = \mathrm{Q}_3 - \mathrm{Q}_1 = Q_{0.75} - Q_{0.25}$$

- É uma medida não negativa e será tanto maior quanto maior for a variabilidade dos dados.
- Amplitude interquartis indica a amplitude do intervalo onde se situa a metade central dos dados, sendo pouco sensível aos valores extremos.
- Uma Amplitude Interquartis nula não significa que os dados não apresentem variabilidade.
- Desvantagem desta medida de dispersão:
  - É insensível às alterações dos valores que se encontram antes do 1^o quartil e depois do 3^o quartil.

Amplitude interquartis

Exemplo

Determine a amplitude interquartis.

1. | 14 | 14 | 12 | 13 | 15 | 15 | 16 | 11 | 10 | 9 |
|---|---|---|---|---|---|---|---|---|---|

Amplitude interquartis

Exemplo

Determine a amplitude interquartis.

1. | 14 | 14 | 12 | 13 | 15 | 15 | 16 | 11 | 10 | 9 |

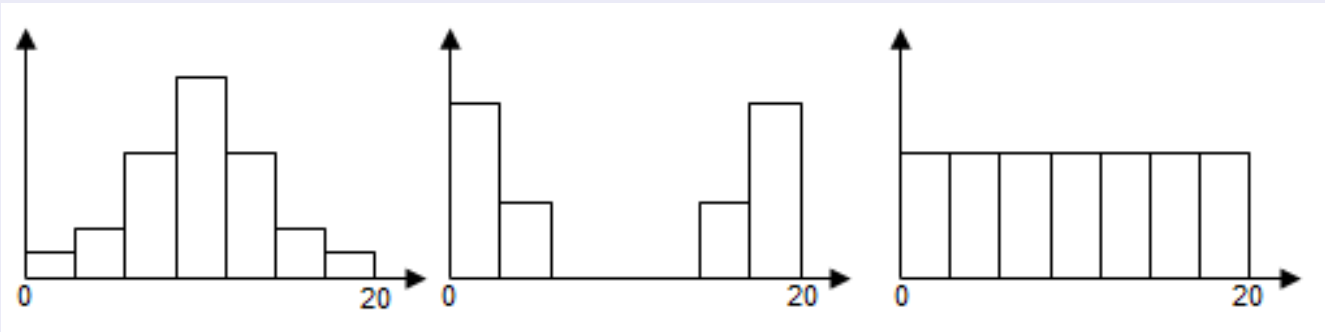
   - Ordenar os dados:

     | 9 | 10 | 11 | 12 | 13 | 14 | 14 | 15 | 15 | 16 |

   - $n = 10$
     - como $np = 10 \times 0.25 = 2.5$, então $Q_1 = Q_{0.25} = x_{(3)} = 11$
     - como $np = 10 \times 0.75 = 7.5$, então $Q_3 = Q_{0.75} = x_{(8)} = 15$
   - $AIQ = 15 - 11 = 4$

Medidas de Dispersão

- Como vimos, embora a amplitude (total ou interquartil) seja uma possibilidade importante para analisar a variabilidade dos dados, tem limitações.
- Outra possibilidade para analisar a variabilidade dos dados consiste em comparar os dados com uma medida de localização central: a média.
- A dispersão dos dados em torno da sua média permite caracterizar um conjunto de dados, pois dados com a mesma média podem ter uma dispersão muito diferente:



- No entanto não é possível caracterizar a variabilidade somando os desvios em relação à média. A soma dos desvios é sempre zero.
- Deve-se considerar uma medida que não leve em conta o sinal dos desvios (o que importa é a magnitude do desvio). Assim, se considerarmos valor absoluto (módulo) dos desvios temos o **Desvio absoluto médio**, mas se considerarmos o quadrado dos desvios temos a **Variância**.

Variância

- Representa-se por s^2 (quando os dados correspondem a uma amostra) ou por σ^2 (quando os dados correspondem à população).
- A variância mede o afastamento dos dados em relação à média.
- A variância é a média dos quadrados dos desvios relativamente à média. Isto é, seja $\{x_1, x_2, \ldots, x_n\}$ um conjunto de dados com n observações, define-se variância como

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1}$$

Variância

Dados em Tabelas de Frequências

Não Estão Agrupados em Classes

- com as frequências absolutas:

$$s^2 = \frac{n_1 (x_1 - \overline{x})^2 + n_2 (x_2 - \overline{x})^2 + \cdots + n_k (x_k - \overline{x})^2}{n-1} =$$

$$= \frac{\sum_{i=1}^{k} n_i (x_i - \overline{x})^2}{n-1} = \frac{\sum_{i=1}^{k} n_i x_i^2 - n\overline{x}^2}{n-1}$$

- com as frequências relativas:

$$s^2 = \frac{n \times \left(f_1 (x_1 - \overline{x})^2 + f_2 (x_2 - \overline{x})^2 + \cdots + f_k (x_k - \overline{x})^2 \right)}{n-1} =$$

$$= \frac{n \times \left(\sum_{i=1}^{k} f_i (x_i - \overline{x})^2 \right)}{n-1} = \frac{n \times \left(\sum_{i=1}^{k} f_i x_i^2 - \overline{x}^2 \right)}{n-1} =$$

Variância

Dados em Tabelas de Frequências

Estão Agrupados em Classes

- com as frequências absolutas:

$$s^2 \approx \frac{n_1 \left(x'_1 - \overline{x}\right)^2 + n_2 \left(x'_2 - \overline{x}\right)^2 + \cdots + n_k \left(x'_k - \overline{x}\right)^2}{n-1} =$$

$$= \frac{\sum_{i=1}^{k} n_i \left(x'_i - \overline{x}\right)^2}{n-1} = \frac{\sum_{i=1}^{k} n_i x_i'^2 - n\overline{x}^2}{n-1}$$

- com as frequências relativas:

$$s^2 \approx \frac{n \times \left(f_1 \left(x'_1 - \overline{x}\right)^2 + f_2 \left(x'_2 - \overline{x}\right)^2 + \cdots + f_k \left(x'_k - \overline{x}\right)^2\right)}{n-1} =$$

$$= \frac{n \times \left(\sum_{i=1}^{k} f_i \left(x'_i - \overline{x}\right)^2\right)}{n-1} = \frac{n \times \left(\sum_{i=1}^{k} f_i x_i'^2 - \overline{x}^2\right)}{n-1}$$

Quando os dados estão agrupados em classes, obtemos um valor aproximado da variância através do representante da classe.

Variância: Exemplos

Determine a variância.

(1)

14	14	12	13	15

Variância: Exemplos

Determine a variância.

(1)

14	14	12	13	15

- calcular a média:

$$\overline{x} = \frac{14 + 14 + 12 + 13 + 15}{5} = 13.6$$

- a variância é

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} = \frac{(14 - 13.6)^2 + (14 - 13.6)^2 + (12 - 13.6)^2 + (13 - 13.6)^2 + (15 - 13.6)^2}{5 - 1} = 1.3$$

ou

- a variância é

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1} = \frac{\left(14^2 + 14^2 + 12^2 + 13^2 + 15^2\right) - 5 \times 13.6^2}{5 - 1} = 1.3$$

Variância: Exemplos

Determine a variância.

(2)

Valores observados (x_i)	Frequência absoluta (n_i)	Frequência relativa (f_i)
45	3	0.091
47	10	0.303
50	7	0.212
53	10	0.303
54	3	0.091
Total	33	1

Variância: Exemplos

Determine a variância.

(2)

Valores observados (x_i)	Frequência absoluta (n_i)	Frequência relativa (f_i)
45	3	0.091
47	10	0.303
50	7	0.212
53	10	0.303
54	3	0.091
Total	33	1

Calcular a variância considerando as frequências absolutas.

- calcular a média:

$$\overline{x} = \frac{45 \times 3 + 47 \times 10 + 50 \times 7 + 53 \times 10 + 54 \times 3}{33} = \frac{1647}{33} = 49.91$$

- a variância é

$$s^2 = \frac{(45 - 49.91)^2 \times 3 + (47 - 49.91)^2 \times 10 + (50 - 49.91)^2 \times 7 + (53 - 49.91)^2 \times 10 + (54 - 49.91)^2 \times 3}{33 - 1} =$$

$$= \frac{302.79}{32} = 9.46$$

Variância: Exemplos

- outra possibilidade de cálculo:

  Calcular a variância considerando as frequências relativas.

  - calcular a média:

Variância: Exemplos

- outra possibilidade de cálculo:

  Calcular a variância considerando as frequências relativas.

  - calcular a média:

    $$\overline{x} = 45 \times 0.091 + 47 \times 0.303 + 50 \times 0.212 + 53 \times 0.303 + 54 \times 0.091 = 49.91$$

  - a variância é

$$s^2 = \frac{33 \times \left[(45-49.91)^2 \times 0.091 + (47-49.91)^2 \times 0.303 + (50-49.91)^2 \times 0.212 + (53-49.91)^2 \times 0.303 + (54-49.91)^2 \times 0.091\right]}{33-1} =$$

$$= \frac{33 \times 9.178}{32} = 9.46$$

Variância: Exemplos

Determine a variância.

(3)

Classe	Representante da classe $\left(x_i'\right)$	Frequência Absoluta (n_i)	Frequência Relativa (f_i)
$[130, 135]$	132.5	7	0.14
$]135, 140]$	137.5	9	0.18
$]140, 145]$	142.5	11	0.22
$]145, 150]$	147.5	14	0.28
$]150, 155]$	152.5	5	0.10
$]155, 160]$	157.5	4	0.08
Total		50	1

Variância: Exemplos

Determine a variância.

(3)

Classe	Representante da classe (x_i')	Frequência Absoluta (n_i)	Frequência Relativa (f_i)
$[130, 135]$	132.5	7	0.14
$]135, 140]$	137.5	9	0.18
$]140, 145]$	142.5	11	0.22
$]145, 150]$	147.5	14	0.28
$]150, 155]$	152.5	5	0.10
$]155, 160]$	157.5	4	0.08
Total		50	1

Calcular a variância considerando as frequências absolutas.

- calcular a média:

$$\overline{x} \approx \frac{132.5 \times 7 + 137.5 \times 9 + 142.5 \times 11 + 147.5 \times 14 + 152.5 \times 5 + 157.5 \times 4}{50} = \frac{7190}{50} = 143.8$$

- a variância é

$$s^2 \approx \frac{(132.5-143.8)^2 \times 7 + (137.5-143.8)^2 \times 9 + (142.5-143.8)^2 \times 11 + (147.5-143.8)^2 \times 14 + (152.5-143.8)^2 \times 5 + (157.5-143.8)^2 \times 4}{50-1} =$$

$$= \frac{2590.32}{49} = 52.86$$

Variância: Exemplos

- outra possibilidade de cálculo:

  Calcular a variância considerando as frequências relativas.

  - calcular a média:

Variância: Exemplos

- outra possibilidade de cálculo:

  Calcular a variância considerando as frequências relativas.

  - calcular a média:

$$\overline{x} \approx 132.5 \times 0.14 + 137.5 \times 0.18 + 142.5 \times 0.22 + 147.5 \times 0.28 + 152.5 \times 0.10 + 157.5 \times 0.08 = 143.8$$

  - a variância é

$$s^2 \approx \frac{50 \times \left[(132.5-143.8)^2 \times 0.14 + (137.5-143.8)^2 \times 0.18 + (142.5-143.8)^2 \times 0.22 + (147.5-143.8)^2 \times 0.28 + (152.5-143.8)^2 \times 0.10 + (157.5-143.8)^2 \times 0.08\right]}{50-1} =$$

$$= \frac{50 \times 51.807}{49} = 52.86$$

Variância

Observações

- Note-se que a **Variância** envolve a soma de quadrados, e por isso a unidade medida em que se exprime não é a mesma que a dos dados, a **unidade de medida** fica **ao quadrado**.

- Vantagem da variância como medida de dispersão:
  - no seu cálculo entram todas as observações.

- Desvantagem da variância como medida de dispersão:
  - não é fácil de interpretar, uma vez que é expressa em unidades da variável ao quadrado;
  - facilmente assume valores muito elevados;
  - é uma medida pouco resistente a valores extremos (muito grandes ou muito pequenos).

Variância

Observações

- Em todas as definições anteriores de variância assumiu-se sempre que se estava a trabalhar com amostras (e não com todos os dados da população). Por isso representou-se a variância por s^2.
- Quando se calcula a variância de dados que correspondem a toda a população (e não a uma amostra), tem-se

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N} = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

onde N é a dimensão da população e μ a média da população.

(A média da população calcula-se do mesmo modo que vimos anteriormente, apenas se usa uma letra diferente para a representar.)

Desvio Padrão

- Representa-se por s (quando os dados correspondem a uma amostra) ou por σ (quando os dados correspondem à população).
- O desvio padrão é a raiz quadrada da variância

$$s = \sqrt{s^2}$$

- O desvio padrão é sempre maior ou igual a zero.
- É a medida de dispersão mais utilizada uma vez que vem expressa na mesma unidade em que estão expressos os dados da amostra.
- O desvio padrão informa sobre o afastamento dos dados em relação à média. Quanto maior for o desvio padrão, maior é o afastamento dos dados em relação à média.
- O Desvio Padrão, assim como a média, é muito sensível a valores extremos, portanto é uma medida pouco resistente.

Desvio Padrão

Exemplo

Determine o desvio padrão (suponha que a unidade de medida dos dados é metros).

14	14	12	13	15

Desvio Padrão

Exemplo

Determine o desvio padrão (suponha que a unidade de medida dos dados é metros).

14	14	12	13	15

média:

$$\overline{x} = \frac{14 + 14 + 12 + 13 + 15}{5} = 13.6 \text{ metros}$$

variância:

$$s^2 = \frac{(14 - 13.6)^2 + (14 - 13.6)^2 + (12 - 13.6)^2 + (13 - 13.6)^2 + (15 - 13.6)^2}{5 - 1} = 1.3 \text{ metros}^2$$

desvio padrão:

$$s = \sqrt{s^2} = \sqrt{1.3} = 1.14 \text{ metros}$$

Coeficiente de Variação

- O desvio padrão por si só não traz muita informação. Ou seja, um desvio padrão de 2 unidades pode ser considerado pequeno para um conjunto de valores cuja média é 200, mas já pode ser considerado grande se a média for de 20.

- Como o desvio padrão vem na mesma unidade de medida dos dados, não se deve usar esta medida de dispersão para comparar conjuntos de dados com **unidades de medida diferentes** ou que **diferem consideravelmente em grandeza**. Neste caso deve-se recorrer ao **Coeficiente de Variação**.

Coeficiente de Variação

- O coeficiente de variação representa-se por CV.
- O coeficiente de variação é uma medida de dispersão relativa e corresponde ao quociente entre o desvio padrão (medida de dispersão) e a média (medida de localização):
  - quando os dados correspondem a uma **amostra**:

$$CV = \frac{s}{\overline{x}} \times 100\%$$

  - quando os dados correspondem à **população**:

$$CV = \frac{\sigma}{\mu} \times 100\%$$

Coeficiente de Variação

- O coeficiente de variação pode ser interpretado como a fração da dispersão pela qual a localização é responsável. Isto é, o coeficiente de variação indica a magnitude relativa do desvio padrão quando comparado com a média do conjunto de valores.
- Quanto maior for o coeficiente de variação, maior é a dispersão dos dados.
- O coeficiente de variação é independente da unidade de medida utilizada, sendo útil para comparar conjuntos de dados.
- Esta medida só deve ser usada quando a variável toma valores de um só sinal, isto é, todos os dados são positivos ou todos os dados são negativos.

Coeficiente de Variação: Exemplo 1

Na tabela seguinte são apresentados os resultados da altura e peso de um grupo de indivíduos:

	Altura	Peso
média $(\overline{x})$:	175 cm	68 kg
desvio padrão (s):	5 cm	2 kg

Qual dos conjuntos de dados apresenta maior dispersão, a altura ou o peso dos indivíduos?

Coeficiente de Variação: Exemplo 1

Na tabela seguinte são apresentados os resultados da altura e peso de um grupo de indivíduos:

	Altura	Peso
média $(\overline{x})$:	175 cm	68 kg
desvio padrão (s):	5 cm	2 kg

Qual dos conjuntos de dados apresenta maior dispersão, a altura ou o peso dos indivíduos?

As **unidades de medidas são diferentes**: a altura está em **centímetros** e o peso está em **quilos**. É necessário calcular o **coeficiente de variação**:

altura

$$CV = \frac{5}{175} \times 100\% = 2,86\%$$

peso

$$CV = \frac{2}{68} \times 100\% = 2,94\%$$

Conclui-se que neste grupo de indivíduos, os pesos apresentam maior grau de dispersão que as alturas.

Coeficiente de Variação: Exemplo 2

Considere os seguintes conjuntos de dados referentes aos preços (em euros) de frigoríficos e batedeiras em 7 lojas distintas:

Frigoríficos	750	800	790	810	820	760	780
Batedeiras	50	45	55	43	52	45	54

$\overline{x} = 787,14$	$s = 25,63$
$\overline{x} = 49.14$	$s = 4.81$

Qual dos produtos tem uma maior variabilidade de preços?

Coeficiente de Variação: Exemplo 2

Considere os seguintes conjuntos de dados referentes aos preços (em euros) de frigoríficos e batedeiras em 7 lojas distintas:

Frigoríficos	750	800	790	810	820	760	780
Batedeiras	50	45	55	43	52	45	54

$\overline{x} = 787,14$	$s = 25,63$
$\overline{x} = 49.14$	$s = 4.81$

Qual dos produtos tem uma maior variabilidade de preços?
As unidades de medidas são iguais mas **diferem consideravelmente em grandeza**.
É necessário calcular o **coeficiente de variação**:

Frigoríficos

$$CV = \frac{25.63}{787.14} \times 100\% = 3.3\%$$

Batedeiras

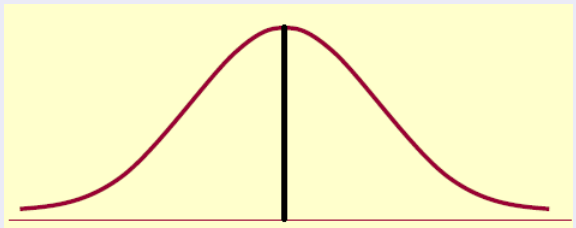
$$CV = \frac{4.81}{49.14} \times 100\% = 9.8\%$$

Conclui-se que neste conjunto de dados, os preços das batedeiras têm uma maior variabilidade do que os preços dos frigoríficos.

Caracterização da Distribuição de Frequências

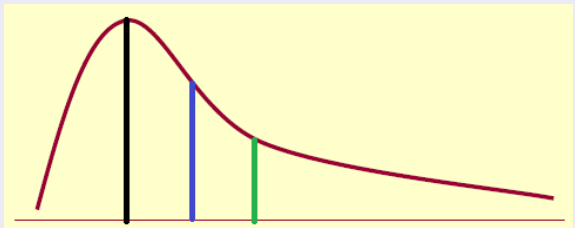
A posição relativa das medidas de localização média, mediana e moda possibilitam classificar as distribuições dos dados como: **Simétricas** ou **Assimétricas**.

- Se a distribuição dos dados for aproximadamente simétrica, a **média** aproxima-se da **mediana** e da **moda**.



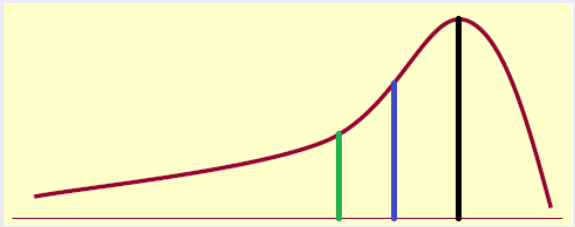
média = mediana = moda

- Se a distribuição dos dados for assimétrica positiva (ou enviesada para a direita), a média tende a ser maior que a mediana e que a **moda**.



moda < mediana < média

- Se a distribuição dos dados for assimétrica negativa (ou enviesada para a esquerda), a média tende a ser inferior à mediana e à **moda**.

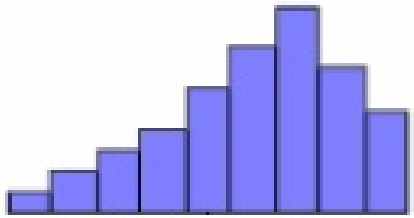


média < mediana < moda

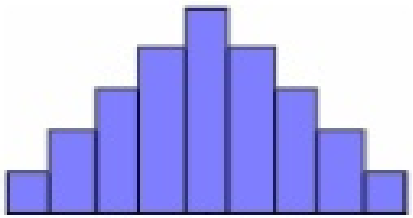
Caracterização da Distribuição de Frequências

Esta caracterização da distribuição de frequências em **Simétrica** ou **Assimétrica** também pode ser observada graficamente:

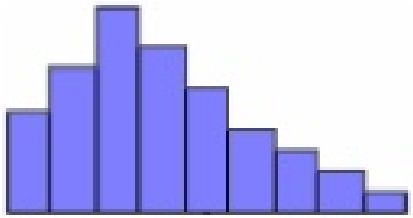
- através do Histograma:



Assimétrica negativa

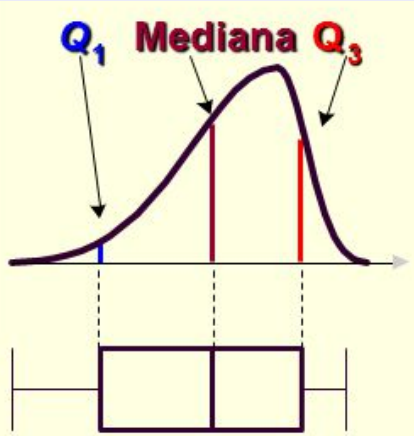


Simétrica

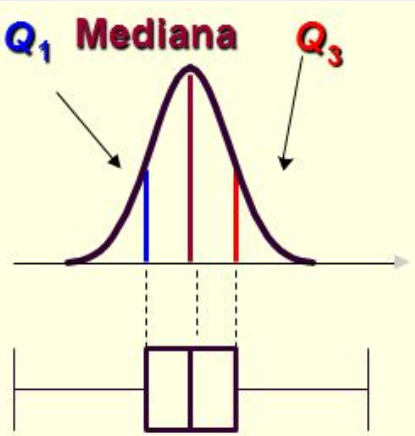


Assimétrica positiva

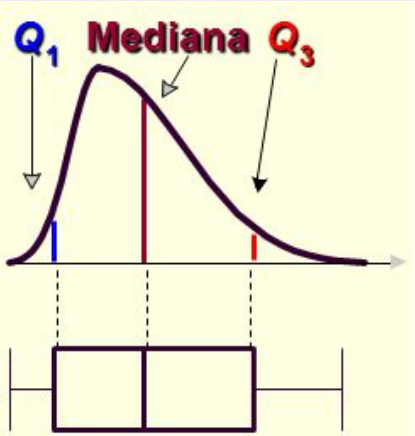
- através do Diagrama de extremos e quartis:



Assimétrico negativo



Simétrico



Assimétrico positivo

Caracterização da Distribuição de Frequências

- As medidas de localização e dispersão, embora forneçam informação importante, são insuficientes para uma boa caracterização da distribuição de frequências.

- Para caracterizar adequadamente a distribuição de frequências é preciso estudar a sua forma, analisando o seu grau de assimetria e de achatamento, recorrendo às seguintes medidas:
  - Medidas de **Assimetria**
  - Medidas de **Curtose** (ou achatamento)

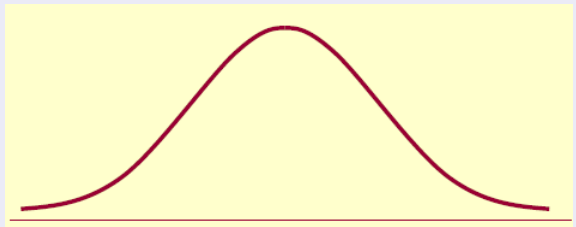
Medidas de Assimetria

Existem diversas medidas de assimetria, o coeficiente b_1 é um dos mais utilizados para avaliar a assimetria:

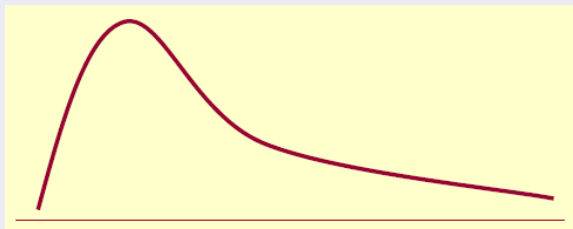
$$b_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^3$$

onde

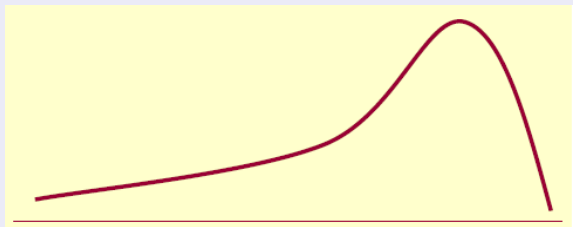
- $b_1 = 0 \rightarrow$ Simétrica



- $b_1 > 0 \rightarrow$ Assimétrica positiva (ou enviesada para a direita)



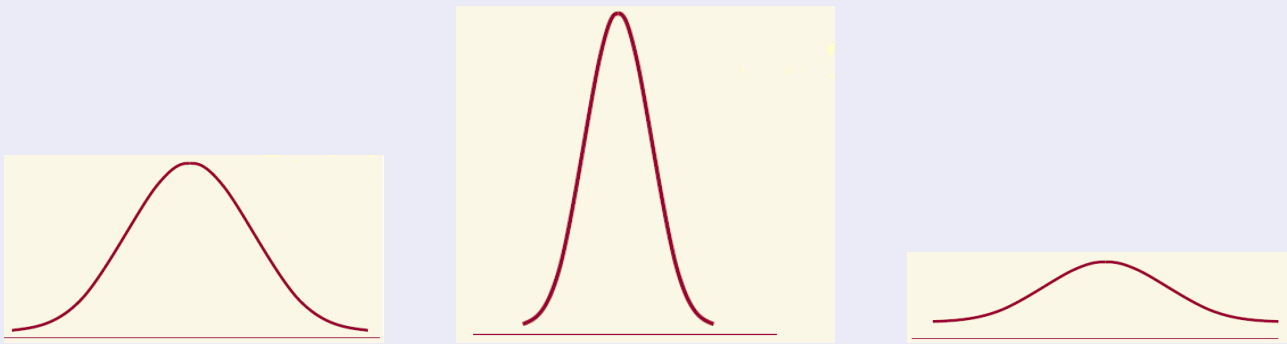
- $b_1 < 0 \rightarrow$ Assimétrica negativa (ou enviesadas para a esquerda)



Medidas de Curtose

- O Achatamento ou Curtose ou "peso das caudas" indica até que ponto a curva de frequências de uma distribuição é mais alongada ou mais achatada do que uma curva padrão, habitualmente denominada de curva normal.

- De acordo com o grau de curtose, podemos ter três tipos de curvas:
  - curva **mesocúrtica**, nem achatada nem alongada, apresenta um grau de achatamento equivalente ao da curva normal
  - curva **leptocúrtica**, com "caudas leves", alongada, apresenta um alto grau de alongamento superior ao da normal
  - curva **platicúrtica**, com "caudas pesadas", achatada, apresenta um alto grau de achatamento superior ao da normal



mesocúrtica leptocúrtica platicúrtica

Medidas de Curtose

Existem diversas medidas de curtose, como por exemplo o coeficiente b_2 que é um dos mais utilizados para avaliar a curtose:

$$b_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^4 - 3$$

- $b_2 = 0 \rightarrow$ curva mesocúrtica, nem achatada nem alongada
- $b_2 > 0 \rightarrow$ curva leptocúrtica, alongada
  - significa que existe uma forte concentração de valores em torno da média e, consequentemente, a variação é pouco elevada.
- $b_2 < 0 \rightarrow$ curva platicúrtica, achatada
  - significa que os valores estão pouco concentrados em torno da média e, consequentemente, existe uma variação elevada

Análise uma base de dados

Antes de efetuar qualquer análise estatística é necessário fazer uma "limpeza" inicial aos dados, de modo a garantir que as conclusões são consistentes. Essa "limpeza" consiste em:

- verificar se existem valores absurdos que só podem ser erros e eliminar esses dados (ou corrigir, caso seja possível);
- verificar se existem dados omissos (em geral representados por NA), registar esse facto e, caso não seja adequado trabalhar com essa falta de informação, retirar os indivíduos nessa situação.

Atenção: Observações discordantes (também chamadas de "outliers") podem não ser erros, mas apenas valores que são possíveis de observar em situações raras.