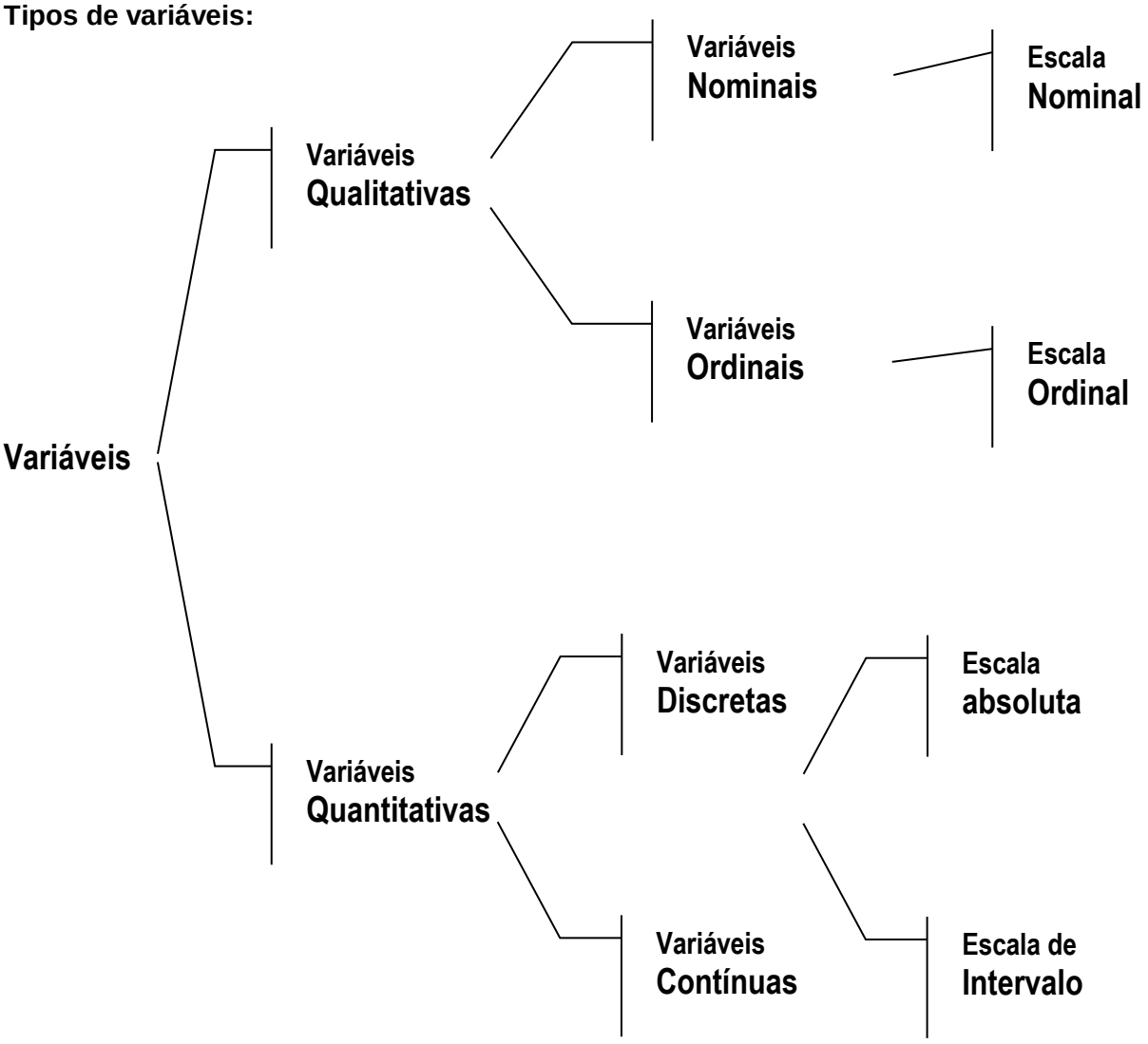
# 1. Variáveis aleatórias, dados estatísticos e escalas

## 1.1 Definições e classificação

**Variável aleatória** é uma entidade que pode assumir os valores de um domínio e que assume esses valores de forma aleatória.

Um conjunto de valores medidos de uma variável aleatória, sobre uma amostra ou uma população, constitui um grupo ou conjunto de **dados estatísticos**.

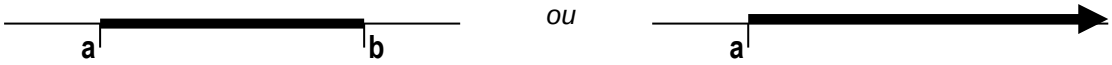
**Tipos de variáveis:**



## Representação de variáveis quantitativas na recta real

*Variáveis Discretas (por exemplo):*


*Variáveis Contínuas (por exemplo):*




*Exemplo 1 -* ***Exemplos de variáveis qualitativas nominais:***

| | |
|---|---|
| *Nome* | *Valores: António , Maria Alice , Felisberto, …* |
| *Fruto* | *Valores: Maçã , Pêra , Noz , Manga , …* |
| *Cor* | *Valores: Amarelo , Branco , Azul, …* |
| *Resposta a …* | *Valores: Sim, Não* |
| *Perigosidade* | *Valores: Tóxico, Inflamável, Provoca queimaduras, …* |

*Exemplo 2 -* ***Exemplos de variáveis qualitativas ordinais:***

| | |
|---|---|
| *Medalha ganha nos J.O.* | *Valores: Ouro , Prata , Bronze, Nenhuma* |
| *Habilitações Académicas* | *Valores: Sem habilitações mínimas, Ensino obrigatório, 12º ano, Bacharelato, Licenciatura, Pós-Graduação, Mestrado, Doutoramento* |
| *Grau de toxicidade* | *Valores: Não tóxico, Pouco tóxico, Tóxico, Muito tóxico* |

*Exemplo 3 -* ***Exemplos de variáveis quantitativas discretas:***

| | |
|---|---|
| *Idade (em anos)* | *Valores: 0, 1, 2, 3, 4, 5, … , 25 , … (escala absoluta)* |
| *Receita da venda do jornal X (em €):* | *Valores (ex.): 0, 1, 2, … 10, … (escala absoluta)* |
| *Impacto de um terramoto* | *Valores: 1 , 2 , 3 … 9 (escala de intervalo)* |
| *N.º de latas infetadas* | *Valores: 0, 1, 2, 3, 4, 5, … , 25 , … (escala absoluta)* |

*Exemplo 4 -* ***Exemplos de variáveis quantitativas contínuas:***

| | | |
|---|---|---|
| *Altura* | *Valores: $\|R^+$* | *(escala absoluta)* |
| *pH* | *Valores: $\|R$* | *(escala de intervalo)* |
| *Concentração* | *Valores: $\|R^+$* | *(escala absoluta)* |
| *Temperatura, em $^0C$* | *Valores: $\|R$* | *(escala de intervalo)* |
| *Potencial eletroquímico* | *Valores: $\|R$* | *(escala de intervalo)* |

### 1.2 Tipo de variável e Excel

É possível definir à partida em cada célula o tipo de variável que aí vai ser introduzida.

Esta atitude evita algumas complicações, sobretudo no que respeita a variáveis qualitativas.

**Procedimento:**

1 – Selecione a(s) célula(s) a formatar.

2 – Selecione o Menu Principal **Base**.

3 – Selecione o campo **Formato do número** (por defeito apresenta selecionado o formato ***Geral***), clicando na seta.

4 – Escolha a categoria mais adequada e, se for caso disso, um número de casas decimais igual àquele com que os seus dados se apresentam:

| Tipo de variável | Categoria a seleccionar |
|---|---|
| Qualitativa Nominal | **Texto ou Geral** |
| Qualitativa Ordinal | **Texto ou Geral** |
| Quantitativa Discreta | **Número ou Geral** |
| Quantitativa Contínua | **Número, Científico ou Geral** |

## 2. Tratamento de dados quantitativos não agrupados (discretos e contínuos)

### 2.1 Características gerais

**Amostra:** $x_1 , x_2 , x_3 , \dots , x_i , \dots , x_n$

**Dimensão da amostra:** **n**

**Valor mínimo:** $x_{mín}$

**Valor máximo:** $x_{máx}$

### 2.2 Cálculos auxiliares

No tratamento dos dados sem utilização de ferramentas informáticas, ajuda realizar previamente os seguintes cálculos:

$$\sum_{i=1}^{n} x_i \quad , \quad \sum_{i=1}^{n} x_i^2 \quad , \quad x_i - \bar{x} \quad , \quad \sum_{i=1}^{n} (x_i - \bar{x})^3 \quad , \quad \sum_{i=1}^{n} (x_i - \bar{x})^4 \qquad [1]$$

*NOTA:* *Estes cálculos são mais comodamente realizados mediante a construção de uma tabela auxiliar de cálculo.*

[1] $\bar{x}$ é a média amostral, cujo cálculo se explica na secção 2.3.

## 2.3 Medidas de localização

**Média amostral:** $\bar{x} = \frac{\sum x_i}{n}$

**Mediana amostral:** **me** A mediana é o valor de x tal que 50% ou mais dos valores amostrais lhe são iguais ou inferiores e 50 % ou mais dos valores amostrais lhe são iguais ou inferiores.

Para determinar a mediana, em primeiro lugar há que ordenar todos os valores da amostra.

*1º processo:* Se n for ímpar, a mediana é igual ao $\frac{n+1^o}{2}$ valor.

Se n for par, a mediana é igual à média aritmética obtida entre o $\frac{n^o}{2}$ e o $\frac{n}{2}+1^o$ valores.

*1º processo:* Realiza-se a conta n / 2 ( ou 0,5 x n ). Se o resultado não for inteiro, arredonda-se ao inteiro seguinte, digamos m. e a mediana é o mº valor na amostra ordenada. Se o resultado for inteiro, digamos m , então a mediana é a média aritmética entre o mº e o (m+1)º valores na amostra ordenada.

**Moda amostral:** **mo** A moda é, à partida, o valor mais frequente na amostra.
ANA verdade, a moda é o valor amostral tal que, quer o valor anterior, quer o valor seguinte, se repetem um número menor de vezes do que ele próprio.
Frequentemente, a moda é o valor que mais vezes se repete na amostra, mas nem sempre isso acontece…

*Algumas situações especiais:*

- Há mais do que um valor repetido igual número (máximo) de vezes e, colocando a amostra ordenada, esses valores são seguidos → A moda é a sua média aritmética.

- Há mais do que um valor repetido igual número (máximo) de vezes e, colocando a amostra ordenada, esses valores não são seguidos → Há duas, três, … modas e diz-se que a amostra é bimodal, trimodal, etc…

- Há mais do que um valor nas condições exigidas (tal que quer o valor anterior quer o valor seguinte se repetem um número menor de vezes do que ele próprio) → Há duas, três, … modas e diz-se que a amostra é bimodal, trimodal, etc, sendo a moda principal o valor que se repete o maior número de vezes.

Qual é a medida de localização mais importante?

Na caracterização da localização de um grupo de dados, a média é quase sempre a melhor medida, dado que é a única sensível a qualquer variação em qualquer valor amostral.
No entanto, em amostras com forte assimetria, a mediana é mais adequada.

## 2.4 Medidas de dispersão

**Amplitude global da amostra:** $\mathbf{A_G} = x_{máx} - x_{mín}$

**Amplitude interquartílica:** $\mathbf{A_{IQ}} = Q_3 - Q_1$ *sendo:*

*1º quartil:* **$Q_1$** – é o valor abaixo do qual se encontram 25% dos valores da amostra (e acima do qual se encontram os restantes 75%).
Calcula-se n/4. Se o resultado não for inteiro, m é o inteiro seguinte e $Q_1$ é o mº valor na amostra ordenada. Se o resultado for inteiro, m é-lhe semelhante e $Q_1$ é a média aritmética entre o mº e o (m+1)º valores da amostra ordenada.

*3º quartil:* **$Q_3$** – é o valor abaixo do qual se encontram 75% dos valores da amostra (e acima do qual se encontram os restantes 25%).
Calcula-se 3xn/4. Se o resultado não for inteiro, m é o inteiro seguinte e $Q_3$ é o mº valor na amostra ordenada. Se o resultado for inteiro, m é-lhe semelhante e $Q_3$ é a média aritmética entre o mº e o (m+1)º valores da amostra ordenada.

*NOTA:* *Existem três quartis que no seu conjunto dividem a amostra em "4 partes iguais", em termos de percentagem de dados, sendo a mediana igual ao 2º quartil, $Q_2$ .*

**Desvio absoluto médio:** $DAM = \dfrac{\sum \left| x_i - \bar{x} \right|}{n}$

**Variância amostral:** $s^2 = \dfrac{\sum x_i^2 - n.\bar{x}^2}{n-1}$ , se n ≤ 30 (amostra pequena)

ou: $s^2 = \dfrac{\sum x_i^2}{n} - \bar{x}^2$ , se n > 30 (amostra grande)

**Desvio padrão:** $s = \sqrt{s^2}$

**Coeficiente de variação:** $cv = \dfrac{s}{\bar{x}}$ ***Nota:*** *Multiplicar por 100, para obter valor percentual.*

Qual é a medida de dispersão mais importante?

Na caracterização da dispersão de uma amostra, o desvio padrão é sem dúvida a medida mais usada.

No entanto, quando pretendemos comparar a dispersão de amostras com diferentes ordens de grandeza, a medida mais indicada é o coeficiente de variação, por ser uma medida adimensional e independente da ordem de grandeza dos valores amostrais.

## 2.5 Medidas de assimetria

**Coeficiente de assimetria amostral:** $g_1 = \frac{n^2}{(n-1).(n-2)} . \frac{m_3}{s^3}$ , se n ≤ 30 (amostra pequena)

ou: $g_1 = \frac{m_3}{s^3}$ , se n > 30 (amostra grande)

*sendo:* $m_3 = \frac{\sum (x_i - \bar{x})^3}{n}$ [2]

O valor do coeficiente de assimetria amostral é:

- nulo, se a amostra for perfeitamente simétrica
- negativo, se a amostra apresentar assimetria negativa ou à esquerda (cauda à esquerda)
- positivo, se a amostra apresentar assimetria positiva ou à direita (cauda à direita)

## 2.6 Medidas de achatamento ou curtose

**Coeficiente de curtose amostral:**

$$g_2 = \frac{n^2 (n+1)}{(n-1)(n-2)(n-3)} . \frac{m_4}{s^4} - 3 . \frac{(n-1)^2}{(n-2)(n-3)}$$ , se n ≤ 30 (amostra pequena)

ou: $g_2 = \frac{m_4}{s^4} - 3$ , se n > 30 (amostra grande)

*sendo:* $m_4 = \frac{\sum (x_i - \bar{x})^4}{n}$ [3]

O valor do coeficiente de curtose amostral é:

- nulo, se a amostra tiver uma curtose semelhante à da distribuição normal – amostra mesocúrtica.
- negativo, se a amostra apresentar uma curtose menos acentuada do que a distribuição normal (mais achatada, com menor concentração de dados ao centro) – amostra platicúrtica.
- positivo, se a amostra apresentar uma curtose mais acentuada do que a distribuição normal (menos achatada, com maior concentração de dados ao centro) – amostra leptocúrtica.

[2] $m_3$ é designado por terceiro momento centrado.

[3] $m_4$ é designado por quarto momento centrado.

## 2.7 Cálculos em Excel

### 2.7.1 Cálculo directo

Utilizando as potencialidades básicas do Excel, é possível realizar todos os cálculos apresentados para Dados Não Agrupados, com comodidade e rigor.

Para efeitos de alguns cálculos, é necessário ordenar valores. Tal também pode ser comodamente realizado em Excel.

**Procedimento para ordenar valores:**

1 – Selecione os valores amostrais.

2 –Selecione o menú **Dados** e depois clique no campo **Ordenar…**

3 – Se não selecionou a linha de cabeçalho (se seleccionou apenas valores), escolha **Não** como resposta à pergunta ***"O intervalo de dados tem linha de cabeçalho?".***

4 – Selecione **OK**

*Exemplo 5 – Veja Exemplo 5, no ficheiro Capítulo 3 Exemplos.xls*

### 2.7.2 Utilização de funções estatísticas pré-definidas

O Excel dispõe de várias funções já definidas, muito úteis no tratamento de dados não agrupados.

Poderá encontrar discrepância de critérios no cálculo de alguns parâmetros, tais como mediana, quartis e, em certos casos, variância, desvio padrão, coeficiente de assimetria e coeficiente de curtose.

**Procedimento para utilizar funções estatísticas:**

1 – Introduza os valores amostrais (um em cada célula), em linha, em coluna ou em matriz rectangular.

2 – Na célula na qual pretende que apareça o resultado, escreva **=**

3 – Prima o ícon ***fx***

4 – Selecione a categoria **Estatística**

5 – Selecione a função pretendida (consulte tabela abaixo)

6 – Selecione **OK**

7 – Premindo o botão esquerdo do rato, Selecione o conjunto dos valores amostrais na página de cálculo

8 – Introduza opções, caso necessário.

9 – Selecione **OK**

| O que pretende calcular? | Função a seleccionar | Opções |
|---|---|---|
| Dimensão amostral, n | CONTAR | |
| Valor mínimo | MÍNIMO / QUARTIL.INC | / 0 |
| Valor máximo | MÁXIMO / QUARTIL.INC | / 4 |
| Média | MÉDIA | |
| Mediana | MED / QUARTIL.INC | / 2 |
| Moda | MODO.MÚLT | |
| Variância | VAR.S | |
| Desvio padrão | DESVPAD.S | |
| 1º Quartil | QUARTIL.INC | 1 |
| 3º Quartil | QUARTIL.INC | 3 |
| Coeficiente de assimetria | DISTORÇÃO | |
| Coeficiente de curtose | CURT | |

*Exemplo 6 – Veja Exemplo 6, no ficheiro Capítulo 3 Exemplos.xls*

### 2.7.3 Utilização da ferramenta Estatística Descritiva

A utilização da ferramenta Estatística Descritiva proporciona rapidamente a obtenção de vários cálculos. Tem porém o inconveniente de ser um método estático (não é um método dinâmico) ; isto é, se quisermos tratar uma nova amostra ou apenas alterar um valor amostral, será necessário voltar a realizar o cálculo.

**Procedimento:**

1 – Introduza os valores amostrais (um em cada célula), em linha ou em coluna.

2 – No Menu **Dados,** clique no campo **Análise de Dados…**

3 – Selecione **Estatística Descritiva.**

4 – Selecione **OK**

5 – Coloque o cursor na caixa ***Intervalo de entrada***. Com o rato Selecione todos os valores amostrais.

6 – Em ***Agrupado por***, caso os dados se encontrem ao longo de uma coluna, seleccionar **Colunas**, caso tenha introduzido os dados ao longo de uma linha, seleccionar **Linhas**.

7 – Em ***Opção de saída***, escolha **Intervalo de saída** e, colocando o cursor na respetiva caixa, Selecione a célula superior esquerda da zona pretendida para saída dos resultados.

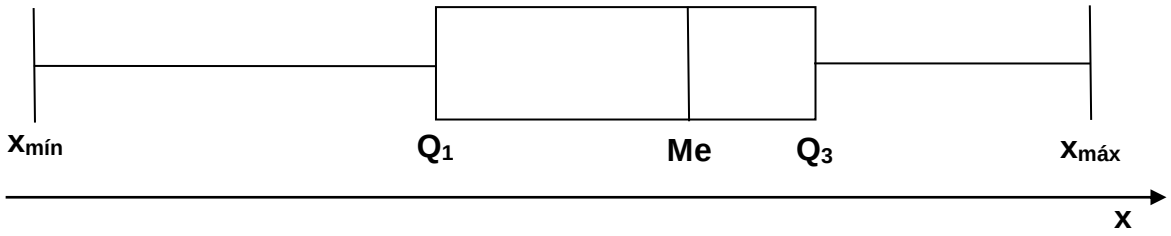
8 – Selecione **Estatísticas de Sumário.**

9 – Selecione **OK.**

*Exemplo 7 – Veja Exemplo 7, no ficheiro Capítulo 3 Exemplos.xls*

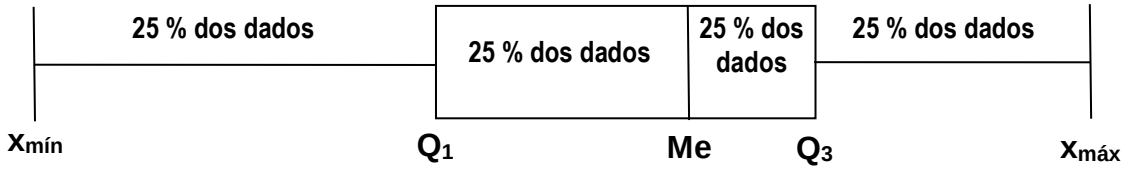
## 2.8 Representação gráfica dos dados

A representação gráfica mais útil para dados quantitativos não agrupados é o diagrama conhecido como **Caixa de Bigodes** (ou **Diagrama de Extremos e Quartis**):




*NOTA: Os diagramas de "Caixa de Bigodes" também podem ser apresentados verticalmente.*

Numa caixa de bigodes, os dados aparecem repartidos em grupos de 25 % de n. Quanto menor for a largura de cada região da caixa de bigodes, maior é a concentração de dados registados no intervalo de valores que lhe corresponde.




Através da observação (visual) da caixa de bigodes podemos tirar conclusões sobre localização, dispersão, assimetria e curtose da amostra, bem como comparar diferentes amostras:

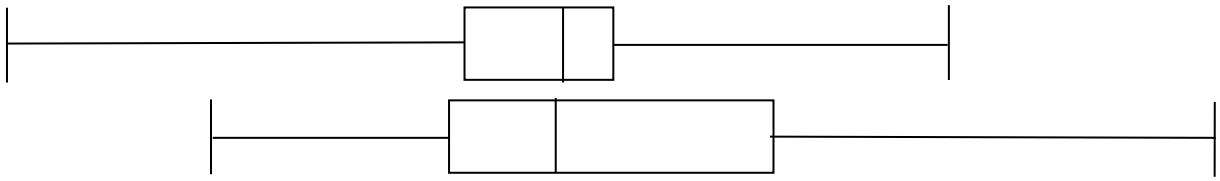
- A localização é principalmente caracterizável pela posição da mediana.

- A dispersão é sobretudo ser avaliável pela largura da caixa, (recorda-se que esta largura corresponde ao intervalo interquartílico, que é uma medida de dispersão).

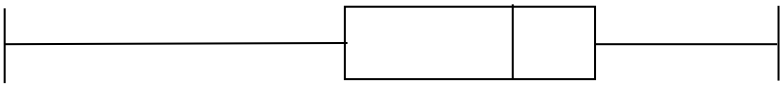
- A assimetria vem dada pela comparação entre os tamanhos relativos de bigode direito e semi-caixa direita com bigode esquerdo e semi-caixa esquerda, respectivamente.

- A curtose pode ser caracterizada por observação da relação entre o tamanho da caixa comparativamente e o tamanho dos bigodes; quanto menor for o tamanho da caixa comparativamente ao tamanho dos bigodes, maior será a curtose amostral.

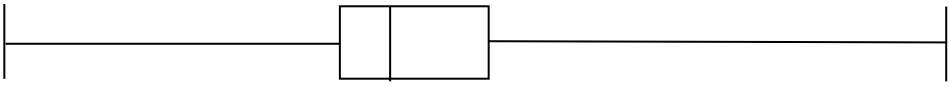
*Exemplo 8 - Estas duas amostras possuem localizações bastante semelhantes ; mas a segunda tem uma dispersão nitidamente superior à primeira.*



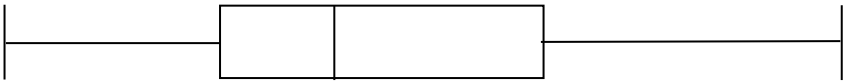
*Exemplo 9 - Esta amostra apresenta algum sintoma de assimetria negativa (cauda à esquerda).*



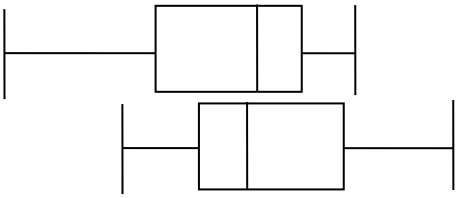
*Exemplo 10 - Esta amostra praticamente não apresenta sintoma de assimetria, mas apresenta sintoma de curtose positiva (grande concentração de dados ao centro).*



*Exemplo 11 – Esta amostra apresenta ligeiros sintomas de assimetria positiva (cauda à direita. Tem uma curtose nula ou negativa (distribuição de dados achatada, não há uma significativa concentração de dados ao centro).*



*Exemplo 8 – Estas duas amostras possuem semelhantes localização e dispersão; mas, enquanto que a primeira apresenta sintomas de assimetria negativa, a segunda apresenta uma assimetria positiva. Em ambos os casos, a curtose é muito baixa.*



Outras representações são possíveis mediante um agrupamento dos dados (ver secções seguintes). No entanto, estas representações poderão mostrar-se pouco elucidativas caso a amostra seja de pequena dimensão.

# 3. Tratamento de dados discretos agrupados

## 3.1 Características gerais

Amostra:

| $x_i$: | $x_1$ | $x_2$ | $x_3$ | … | $x_k$ |
|---|---|---|---|---|---|
| $n_i$: | $n_1$ | $n_2$ | $n_3$ | … | $n_k$ |

sendo $x_i$ o iésimo valor diferente da amostra ordenada e $n_i$ o número de ocorrências que ele ocorre na amostra.

**Dimensão da amostra**: **n** $= \sum_i n_i$

**Número de valores distintos na amostra:** **k** ( $k \leq n$ )

Os valores $x_i$ são ordenados por ordem crescente, pelo que se tem sempre que:

**Valor mínimo:** $x_{mín} = x_1$ e **Valor máximo:** $x_{max} = x_k$

## 3.2 Tabelas de frequências

É muito útil e por si só importante, no tratamento de dados agrupados, a construção de uma tabela de frequências. Nesta devem constar as seguintes frequências:

**Frequência (simples) absoluta**, $\mathbf{n_i}$ – Número de vezes que o valor $x_i$ ocorre na amostra.

**Frequência (simples) relativa, $f_i$ –** Proporção do número de vezes que o valor $x_i$ ocorre na amostra em relação à sua dimensão.
Pode multiplicar-se por 100 para obter um valor percentual – $f\%_i$ .

**Frequência acumulada absoluta, $N_i$–** Número de vezes que o valor $x_i$ ou valores inferiores a $x_i$ ocorrem na amostra.

**Frequência acumulada relativa, $F_i$ –** Proporção do número de vezes que o valor $x_i$ ou valores inferiores a $x_i$ ocorrem na amostra em relação à sua dimensão.
Pode multiplicar-se por 100 para obter um valor percentual – $F\%_i$ .

Fórmulas de relação entre as várias frequências:

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

$$\sum_{i=1}^{k} n_i = n \qquad \sum_{i=1}^{k} f_i = 1 \qquad N_k = n \qquad F_k = 1$$

## 3.3 Cálculos auxiliares

Ajuda no tratamento dos dados realizar previamente os seguintes cálculos:

$$\sum_{i=1}^{k} n_i.x_i \text{ ou } \sum_{i=1}^{k} f_i.x_i$$

$$\sum_{i=1}^{k} n_i.x_i^2 \text{ ou } \sum_{i=1}^{k} f_i.x_i^2$$

$$\sum_{i=1}^{k} n_i.(x_i - \bar{x})^3 \text{ ou } \sum_{i=1}^{k} f_i.(x_i - \bar{x})^3$$

$$\sum_{i=1}^{k} n_i.(x_i - \bar{x})^4 \text{ ou } \sum_{i=1}^{k} f_i.(x_i - \bar{x})^4$$

*NOTA: Estes cálculos são mais comodamente realizados mediante a construção de uma tabela auxiliar de cálculo – pode criar-se, por exemplo, um prolongamento na tabela de frequências.*

## 3.4 Medidas de localização

**Média amostral:** $$\bar{x} = \frac{\sum n_i.x_i}{n} = \sum f_i.x_i$$

**Mediana amostral:** **me** A mediana continua a ser o valor de x tal que 50% ou mais dos valores amostrais lhe são iguais ou inferiores e 50 % ou mais dos valores amostrais lhe são iguais ou inferiores.

Para determinar a mediana, utiliza-se uma frequência acumulada, absoluta ou relativa.

*1º processo:* Em primeiro lugar há que calcular 0,5 x n.
Seguidamente, procura-se o valor obtido entre os valores de $N_i$ .(frequência acumulada absoluta).
Se o valor exacto for encontrado, a mediana é a média aritmética entre o valor de x correspondente e o valor de x seguinte.
Se o valor exacto não for encontrado, a mediana é o valor de x correspondente ao primeiro $N_i$ superior a 0,5 x n.

*2º processo:* Procura-se o valor 0,5 entre os valores de $F_i$ (frequência acumulada relativa).
Se o valor exacto for encontrado, a mediana é a média aritmética entre o valor de x correspondente e o valor de x seguinte.
Se o valor exacto não for encontrado, a mediana é o valor de x correspondente ao primeiro $F_i$ superior a 0,5.

**Moda amostral:** **mo** A moda é o valor de $x_i$ tal que, quer o valor anterior $x_{i-1}$, quer o valor seguinte $x_{i+1}$ apresentam menores frequências simples (absolutas $n_i$ ou relativas $f_i$) do que ele próprio.
Frequentemente, a moda é o valor $x_i$ correspondente ao maior $n_i$ ou $f_i$ ou seja, o valor que mais vezes se repete na amostra, mas nem sempre isso acontece…

*Algumas situações especiais:*

- Há mais do que um valor $x_i$ com as mesmas frequências simples máximas e esses valores são seguidos → A moda é a sua média aritmética.

- Há mais do que um valor $x_i$ com as mesmas frequências simples máximas e esses valores não são seguidos → Há duas, três, … modas e diz-se que a amostra é bimodal, trimodal, etc…

- Há mais do que um valor nas condições exigidas (tal que quer o valor anterior quer o valor seguinte apresentam frequências simples menores do que ele próprio) → Há duas, três, … modas e diz-se que a amostra é bimodal, trimodal, etc, sendo a moda principal o valor ao qual corresponde a maior frequência simples (valor que se repete mais vezes).

## 3.5 Medidas de dispersão

**Amplitude global da amostra:** $A_G = x_{máx} - x_{mín} = x_k - x_1$

**Amplitude interquartílica:** $A_{IQ} = Q_3 - Q_1$

*sendo os 1º e 3º quartiis, respectivamente $Q_1$ e $Q_2$, calculados por:*

*1º processo:* Primeiro, há que calcular 0,25 x n, para $Q_1$, ou 0,75 x n, para $Q_3$.
Seguidamente, procura-se o valor obtido entre os valores de $N_i$ (frequência acumulada absoluta).
Se o valor exacto for encontrado, o quartil é a média aritmética entre o valor de x correspondente e o valor de x seguinte.
Se o valor exacto não for encontrado, o quartil é o valor de x correspondente ao primeiro $N_i$ superior a 0,25 x n ou 0,75 x n, respectivamente.

*2º processo:* Procura-se o valor 0,25 , para $Q_1$, ou 0,75, para $Q_3$ entre os valores de $F_i$ (frequência acumulada relativa).
Se o valor exacto for encontrado, o quartil é a média aritmética entre o valor de x correspondente e o valor de x seguinte.
Se o valor exacto não for encontrado, o quartil é o valor de x correspondente ao primeiro $F_i$ superior a 0,25 ou 0,75, respectivamente.

**Variância amostral:** $s^2 = \frac{\sum n_i.x_i^2 - n.\bar{x}^2}{n-1} = \frac{n}{n-1}.\left(\sum f_i.x_i^2 - \bar{x}^2\right)$ , se n ≤ 30 (am. peq.)

ou: $s^2 = \frac{\sum n_i.x_i^2}{n} - \bar{x}^2 = \sum f_i.x_i^2 - \bar{x}^2$ , se n > 30 (am.grande)

**Desvio padrão:** $s = \sqrt{s^2}$

**Coeficiente de variação:** $cv = \frac{s}{\bar{x}}$ ***Nota:*** *Multiplicar por 100, para obter valor percentual.*

## 3.6 Medidas de assimetria

**Coeficiente de assimetria amostral:**

$$g_1 = \frac{n^2}{(n-1).(n-2)} . \frac{m_3}{s^3} \text{ , se } n \leq 30 \text{ (amostra pequena)}$$

ou: $g_1 = \frac{m_3}{s^3}$ , se $n > 30$ (amostra grande)

*sendo:* $m_3 = \frac{\sum (x_i - \bar{x})^3 . n_i}{n}$ [4]

O valor do coeficiente de assimetria amostral tem o significado apresentado anteriormente.

## 3.7 Medidas de achatamento ou curtose

**Coeficiente de curtose amostral:**

$$g_2 = \frac{n^2 (n+1)}{(n-1)(n-2)(n-3)} . \frac{m_4}{s^4} - 3 . \frac{(n-1)^2}{(n-2)(n-3)} \text{ , se } n \leq 30 \text{ (amostra pequena)}$$

ou: $g_2 = \frac{m_4}{s^4} - 3$ , se $n > 30$ (amostra grande)

*sendo:* $m_4 = \frac{\sum (x_i - \bar{x})^4 . n_i}{n}$ [5]

O valor do coeficiente de curtose amostral tem o significado apresentado anteriormente.

[4] $m_3$ é designado por terceiro momento centrado.

[5] $m_4$ é designado por quarto momento centrado.

## 3.8 Cálculos em Excel

O Excel não possui ferramentas ou funções específicas para o tratamento de dados discretos agrupados. No entanto, as suas potencialidades de base bem aplicadas aos conhecimentos teóricos do analista permitem, de forma expedita, uma análise profunda e completa deste tipo de dados.

*Exemplo 12 – Veja Exemplo 12, no ficheiro Capítulo 3 Exemplos.xls*

## 3.9 Representação gráfica dos dados

A representação gráfica em forma de o **Diagrama de "Caixa de Bigodes"** continua a ser extremamente útil numa análise da amostra. As suas construção e interpretação são em tudo semelhantes às já apresentadas – as únicas variações ocorrem nos cálculos dos valores envolvidos.

Outras representações também muito úteis no caso de dados discretos agrupados são:

- **Gráfico de barras** – representa uma frequência simples (absoluta ou relativa) em função de $x_i$

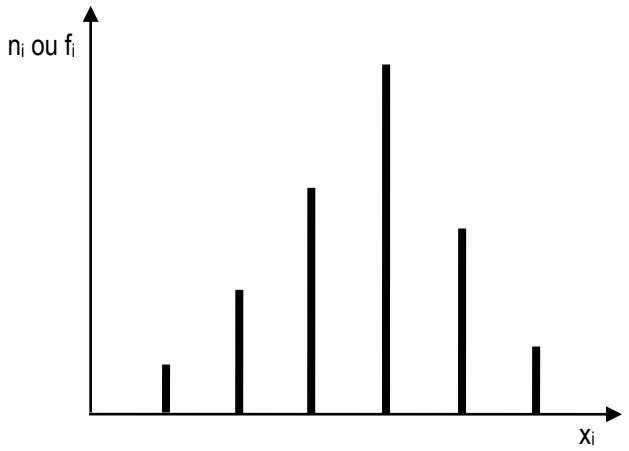
- **Polígono de frequências simples** (absolutas ou relativas) – une os topos das barras do diagrama anterior com rectas. Nos extremos, é prolongado para valores de x imaginários, abaixo do mínimo e acima do máximo, com frequências simples nulas.

- **Gráfico de barras de frequências acumuladas** (absolutas ou relativas) – representa $N_i$ ou $F_i$ em função de $x_i$. Cada coluna i é assente sobre o espaço que une $x_i$ a $x_{i+1}$.

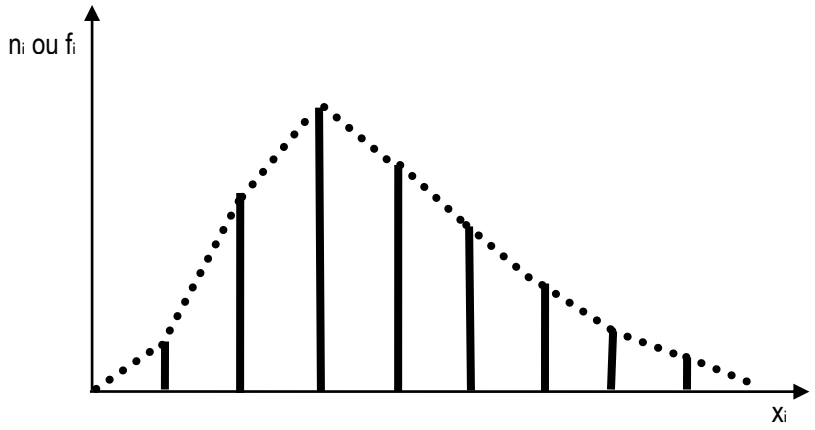
- **Curva de frequências acumuladas** (absolutas ou relativas) – representa $N_i$ ou $F_i$ em função de $x_i$ ; une os pontos ( $x_i$ , $N_i$ ) ou ( $x_i$ , $F_i$ ). Considera-se um ponto imaginário $x_{i-1}$, como sendo o último ponto para o qual a frequência acumulada vale 0. Para valores de x superiores a $x_{máx}$, a frequência acumulada vale o seu máximo ( n ou 1, respectivamente ).

*Exemplo 13 – Veja Exemplo 13, no ficheiro Capítulo 3 Exemplos.xls*

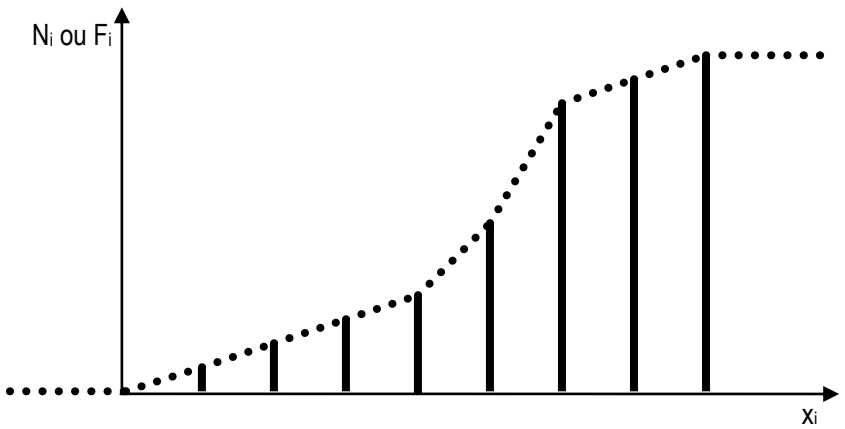
*Exemplo 14 - Esquema de uma representação de uma amostra sob a forma de gráfico de barras. Esta amostra apresenta algum sintoma de assimetria negativa (cauda à esquerda) e uma curtose média..*




*Exemplo 15 - Esquema de uma representação de uma amostra sob a forma de gráfico de barras e também de polígono de frequências.*
*Esta amostra apresenta sintoma de assimetria positiva (cauda à direita) e uma curtose ligeiramente negativa.*




*Exemplo 16 - Gráfico de barras e curva de frequências acumuladas, para uma amostra de dados discretos agrupados, com k = 8.*



# 4. Tratamento de dados contínuos agrupados (por classes)

## 4.1 Características gerais

Amostra:

| i: | 1 | 2 | 3 | | k |
|---|---|---|---|---|---|
| Classe i: | $linf_1$ a $lsup_1$ | $linf_2$ a $lsup_2$ | $linf_3$ a $lsup_3$ | … | $linf_k$ a $lsup_k$ |
| $n_i$: | $n_1$ | $n_2$ | $n_3$ | … | $n_k$ |

sendo $l_{inf\ 1}$ a $l_{sup\ 1}$ a iésima classe diferente da amostra (por exemplo, 25 a 50 ) e $\mathbf{n_i}$ os números de vezes que ocorrem na amostra valores entre cada par de limites, designados por **frequências simples absolutas**.
As classes são sempre mutuamente exclusivas e ordenam-se por ordem crescente dos valores de x.

**Dimensão da amostra**: $\mathbf{n} = \sum_i n_i$

**Número de classes: k** ( $k \leq n$ )

**Valor mínimo:** $x_{mín} = linf_1$ e **Valor máximo:** $x_{max} = lsup_k$

## 4.2 Características de cada classe

Cada classe i está definida entre dois **limites**: $linf_i$ (limite inferior) e $lsup_i$ (limite superior)
Caso o limite superior de cada classe não coincida com o limite inferior da classe seguinte, então convém calcular os **limites reais** (a utilizar nos cálculos):

$$lsup_i = linf_{i+1} = ( lsup_i' + linf_{i+1}' ) / 2$$

NOTA IMPORTANTE: Por analogia a $linf_1$ subtrai-se uma quantidade igual à que foi subtraída nos restantes limites inferiores e a $lsup_k$ adiciona-se uma quantidade igual à que foi adicionada nos restantes limites superiores.

**Intervalo de classe i:** - para as classes 1 a k-1: **[ $\mathbf{linf_i}$ , $\mathbf{lsup_i}$ [**(limites reais)
- para a classe k: **[ $\mathbf{linf_k}$ , $\mathbf{lsup_k}$ ]**

**Amplitude (ou largura ) da classe i:** $a_i = lsup_i - linf_i$

**Ponto médio da classe i:** $\overline{x_i} = \frac{l\inf_i + l\sup_i}{2}$

### 4.3 Como organizar dados não agrupados em dados agrupados por classes?

Logo, à partida, há que definir o número de classes, **k** .

Para tal, há várias referências sugeridas por vários autores. Apresentam-se de seguida duas referências de simples utilização.

Para amostras de dimensão até aproximadamente 100 valores:

$$k \approx \sqrt{n}$$ [6]

Para amostras de dimensão muito grande:

$$\mathbf{k \approx 1 + 3{,}322 \,.\, \log_{10} n}$$ ***(Regra de Sturges)***

devendo obrigatoriamente k ser inteiro.

Quase sempre, pretende-se criar classes com iguais amplitudes. No entanto, em alguns casos particulares, há conveniência em que isso não aconteça (por exemplo, classificação dos indivíduos humanos em faixas etárias, referentes à sua evolução pessoal).

**Para criar classes com igual amplitude, siga os seguintes passos:**

1º Calcular sobre os dados ainda não agrupados: **n**, $\mathbf{x_{mín}}$, $\mathbf{x_{máx}}$ e $\mathbf{A_G}$.

2º Definir / escolher o número de classes a criar, **k**, por uma das regras acima apresentadas.

3º Definir / escolher a amplitude cada classe, tomando como referência que deverá ser próxima, mas nunca inferior a $A_G$ / k, sendo $A_G$ a amplitude global dos dados ainda não agrupados :

$$\mathbf{a_i \geq A_G / k}$$

Para se obter limites de mais fáceis leitura e tratamento, pode-se e deve-se arredondar este valor, porém sempre por excesso,

4º Definir / escolher o limite inferior da primeira classe, que deve ser próximo do valor de $x_{mín}$ da amostra ainda não agrupada, podendo e devendo ser arredondado inferiormente, se necessário:

$$\mathbf{L_{inf1} \leq x_{mín}}$$

5º Antes de criar as classes e proceder à contagem dos valores, convém verificar as opções tomadas para os valores de k, $a_i$ e de $l_{inf1}$, por forma a garantir que todos os valores amostrais pertençam a uma classe. Um modo simples de o fazer é verificar a seguinte condição necessária:

$$\mathbf{l_{inf\,1} + k \,.\, a_i \geq x_{max}}$$ ( $x_{max}$ da amostra ainda não ordenada)

6º Cria-se a primeira classe com:

$$linf_1 \text{ e } lsup_1 = linf_1 + a_1$$

Depois, criam-se as 2ª, 3ª , etc classes, sempre por esta ordem, obedecendo a:

$$linf_i = lsup_{i-1} \text{ e } lsup_i = linf_i + a_i$$

[6] Guimarães, R.C. , Cabral , J.A.S., "Estatística" , McGraw Hill

A amplitude de cada classe, $linf_1$ e os limites de classes devem sempre ser definidos com um número de algarismos significativos igual ou superior ao dos dados amostrais em bruto.

7º Como um valor não pode pertencer em simultâneo a duas classes, há que definir se é o limite inferior que vai ser incluído na classe ou o limite superior:

$[\ l_{inf\ i}\ ;\ l_{sup\ i}\ [$ ou $]\ l_{inf\ i}\ ;\ l_{sup\ i}\ ]$

Ambas as opções são corretas, mas é necessário optar e proceder em coerência.

8º Criadas as k classes, só resta proceder à contagem dos números de ocorrências em cada uma delas, para se obter os valores de $n_i$.

*Exemplo 17 - Sendo n = 200, com $x_{mín}$ = 1,2 e $x_{máx}$ = 17,8 ; como poderemos agrupar estes dados?*

*Primeiro calculamos e escolhemos um número de classes adequado:*
*$k \approx 1 + 3{,}322\ .\ \log_{10} n = 8{,}6 \rightarrow k = 9$ (também seria razoável k = 8.)*

*A amplitude global é: $A = x_{máx} - x_{mín} = 17{,}8 - 1{,}2 = 16{,}59$*

*A amplitude calculada para cada classe é: $a_i = 16{,}59 / 9 = 1{,}843$ . Podemos arredondar este valor, mas nunca inferiormente. Tomemos $a_i = 1{,}9$.*

*O limite inferior da primeira classe deverá ser igual ou ligeiramente inferior ao valor mínimo. Vamos escolher 1,2.*

*Verificação: $1{,}2 + 9 \times 1{,}9 = 18{,}3 \geq 17{,}8$ OK !*

*O limite superior da primeira classe, igual ao limite inferior da segunda classe é: 1,2 + 1,9 = 3,1.*

*O limite superior da segunda classe, igual ao limite inferior da terceira classe é: 3,1 + 1,9 = 5,0.*

*Etc.*

| *i* | *Intervalo de classe* |
|---|---|
| *1* | *[ 1,1 ; 2,3 ]* |
| *2* | *] 2,3 ; 3,5 ]* |
| *3* | *] 3,5 ; 4,7 ]* |
| *4* | *] 4,7 ; 5,9 ]* |
| *5* | *] 5,9 ; 7,1 ]* |
| *6* | *] 7,1 ; 8,3 ]* |
| *7* | *] 8,3 ; 9,5 ]* |
| *8* | *] 9,5 ; 10,7 ]* |
| *9* | *] 10,7 ; 11,9 ]* |

*Pode fechar-se o primeiro intervalo, por forma a incluir este valor.*

*Uma vez criadas as classes, procederíamos de seguida à contagem dos dados por classes, para obtermos os valores de $n_i$.*

### 4.4 Utilização do Excel para agrupar dados

O Excel é sempre útil para agrupar de dados, quanto mais não seja, pela possibilidade de os ordenar, o que facilita imenso a sua contagem.

O Excel dispõe ainda de uma função que faz a contagem automática dos dados – a **função Frequência**. Esta função agrupa sempre os dados em intervalos abertos à esquerda e fechados à direita (e não abertos à esquerda e fechados à direita).

**Procedimento para agrupar dados contínuos em classes:**

1 – Introduza os valores amostrais.

2 – Introduza os limites superiores de cada classe.

3 – Selecione a célula para saída da frequência simples absoluta da primeira classe. Aí, introduza **=,** clique em ***fx*** e depois selecione a função ***Frequência*** (no submenú *Estatística*).

3 – Selecione os dados a agrupar.

4 – Selecione todos os limites superiores de classe (NOTE BEM: apenas os limites superiores).

5 – Selecione **OK**.

6 – Selecione a célula de saída do último resultado e também as células imediatamente abaixo, num total de número de células igual ao número de classes.

7 – Prima ***F2***.

8 – Prima ***Ctrl – Shift – Enter***.

*Exemplo 18 – Veja Exemplo 18, no ficheiro Capítulo 3 Exemplos.xls*

## 4.5 Tabela de frequências

Tal como no caso dos dados discretos agrupados, ajuda muito no tratamento de dados agrupados, e também facilita a sua análise, a construção de uma tabela de frequências. Nesta devem constar as frequências também já anteriormente referidas:

**Frequência (simples) absoluta**, $\mathbf{n_i}$ – Número de vezes que a classe i ocorre na amostra.

**Frequência (simples) relativa, $\mathbf{f_i}$ –** Proporção do número de vezes que a classe i ocorre na amostra em relação à sua dimensão.
Pode multiplicar-se por 100 para obter um valor percentual – $f\%_i$ .

**Frequência acumulada absoluta, $\mathbf{N_i}$–** Número de vezes que a classe i ou classes inferiores a ela ocorrem na amostra.

**Frequência acumulada relativa, $\mathbf{F_i}$ –** Proporção do número de vezes que a classe i ou classes inferiores a ela ocorrem na amostra em relação à sua dimensão.
Pode multiplicar-se por 100 para obter um valor percentual – $F\%_i$ .

Fórmulas de relação entre as várias frequências (semelhantes às anteriores):

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

$$\sum_{i=1}^{k} n_i = n \qquad \sum_{i=1}^{k} f_i = 1 \qquad N_k = n \qquad F_k = 1$$

## 4.6 Cálculos auxiliares

Ajuda no tratamento dos dados realizar previamente os seguintes cálculos:

$$\sum_{i=1}^{k} n_i.\overline{x_i} \text{ ou } \sum_{i=1}^{k} f_i.\overline{x_i}$$

$$\sum_{i=1}^{k} n_i.\overline{x_i}^2 \text{ ou } \sum_{i=1}^{k} f_i.\overline{x_i}^2$$

$$\sum_{i=1}^{k} n_i.(\overline{x_i} - \overline{x})^3 \text{ ou } \sum_{i=1}^{k} f_i.(\overline{x_i} - \overline{x})^3$$

$$\sum_{i=1}^{k} n_i.(\overline{x_i} - \overline{x})^4 \text{ ou } \sum_{i=1}^{k} f_i.(\overline{x_i} - \overline{x})^4$$

*NOTA: Estes cálculos são mais comodamente realizados mediante a construção de uma tabela auxiliar de cálculo – pode criar-se, por exemplo, um prolongamento na tabela de frequências.*

## 4.7 Medidas de localização

**Média amostral:** $$\bar{x} = \frac{\sum n_i . \overline{x_i}}{n} = \sum f_i . \overline{x_i}$$

**Mediana amostral:** **Me** A mediana é uma vez mais o valor de x tal que 50% ou mais dos valores amostrais lhe são iguais ou inferiores e 50 % ou mais dos valores amostrais lhe são iguais ou inferiores.

*1º processo:* Para determinar a mediana, em primeiro lugar há que calcular 0,5 x n .
Seguidamente, procura-se o valor obtido entre os valores de $N_i$ (frequência acumulada absoluta). A primeira classe cujo $N_i$ seja igual ou superior ao valor procurado chama-se **classe mediana** e é por definição a classe que contém a mediana.
Se o valor exacto for encontrado, a mediana é igual ao limite superior real da classe mediana.
Se o valor exacto não for encontrado, então há que calcular a mediana pela seguinte fórmula:

$$Me = l\inf_i + \frac{0{,}5xn - N_{i-1}}{n_i}.a_i$$

*NOTA: i é a classe mediana ; i-1 é a classe anterior à classe mediana.*

*2º processo:* Procura-se o valor 0,5 entre os valores de $F_i$ (frequência acumulada relativa). A primeira classe cujo $F_i$ seja igual ou superior a 0,5 chama-se **classe mediana** e é por definição a classe que contém a mediana.
Se o valor exacto for encontrado, a mediana é igual ao limite superior real da classe mediana.

Se o valor exacto não for encontrado, então há que calcular a mediana pela seguinte fórmula:

$$Me = l\inf_i + \frac{0{,}5 - F_{i-1}}{f_i}.a_i$$

*NOTA: i é a classe mediana ; i-1 é a classe anterior à classe mediana.*

**Moda amostral:** **Mo** A moda é, teoricamente, o valor de x mais frequente, ainda que ele não ocorra na amostra.
Para calcular a moda, há em primeiro lugar que determinar a **classe modal**, que é a classe tal que, quer a classe anterior, quer a classe seguinte apresentam menores frequências simples (absolutas $n_i$ ou relativas $f_i$) do que ela própria.
Frequentemente, a classe modal moda é a classe correspondente ao maior $n_i$ ou $f_i$ ou seja, mas nem sempre isso acontece…

*Algumas situações especiais:*

- Há mais do que uma classe com as mesmas frequências simples máximas e essas classes são seguidas → Agrupam-se as duas, três, … classes numa só, para efeitos de cálculo da moda. A classe modal será depois a classe que a contiver.

- Há mais do que uma classe com as mesmas frequências simples máximas e essas classes não são seguidas → Há duas, três, … classes modais (consequentemente haverá duas, três, … modas) e diz-se que a amostra é bimodal, trimodal, etc.

- Há mais do que uma classe nas condições exigidas (tal que quer a classe anterior quer a classe seguinte apresentam frequências simples menores do que ela própria) → Há duas, três, … classes modais (consequentemente haverá duas, três, … modas) e diz-se que a amostra é bimodal, trimodal, etc, sendo a classe modal e a moda principais as correspondentes à maior frequência simples.

Depois de identificada(s) a(s) classe(s) modal(is), a(s) moda(s) calcula(m)-se pela seguinte fórmula:

$$Mo = l\,\text{inf}_i + \frac{\Delta_1}{\Delta_1 + \Delta_2}.a_i$$

sendo: $\Delta 1 = n_i - n_{i-1}$ e $\Delta_2 = n_i - n_{i+1}$

ou então: $\Delta 1 = f_i - f_{i-1}$ e $\Delta_2 = f_i - f_{i+1}$

*NOTA: i é a classe modal ; i-1 é a classe anterior à classe modal ; i+1 é a classe seguinte à classe modal.*

*NOTA MUITO IMPORTANTE:* Caso as classes não tenham todas igual amplitude, deverá substituir no texto e nas fórmulas anteriores $n_i$ por $n_i / a_i$ e $f_i$ por $f_i / a_i$ .

## 4.8 Medidas de dispersão

**Amplitude global da amostra:** $A_G = x_{máx} - x_{mín} = lsup_k - linf_1$

**Amplitude interquartílica:** $A_{IQ} = Q_3 - Q_1$

*sendo os 1º e 3º quartiis, respectivamente $Q_1$ e $Q_2$, calculados por:*

*1º processo:* Para determinar os quartis, em primeiro lugar há que calcular 0,25 x n ou 0,75 x n, respectivamente consoante se trate do primeiro ou do terceiro quartis .
Seguidamente, procura-se o valor obtido entre os valores de $N_i$ (frequência acumulada absoluta). A primeira classe cujo $N_i$ seja igual ou superior ao valor procurado é a classe que contém o quartil.
Se o valor exacto for encontrado, o quartil é igual ao limite superior real dessa classe.
Se o valor exacto não for encontrado, então há que calcular o quartil por uma das seguintes fórmulas:

$$Q_1 = l\,\text{inf}_i + \frac{0{,}25xn - N_{i-1}}{n_i}.a_i \qquad Q_3 = l\,\text{inf}_i + \frac{0{,}75xn - N_{i-1}}{n_i}.a_i$$

*2º processo:* Procura-se o valor 0,25 ou 0,75, respectivamente consoante se trate do primeiro ou do terceiro quartis, entre os valores de $F_i$ (frequência acumulada relativa). A primeira classe cujo $F_i$ seja igual ou superior a 0,5 é a classe que contém o quartil.
Se o valor exacto for encontrado, o quartil é igual ao limite superior real dessa classe.
Se o valor exacto não for encontrado, então há que calcular o quartil por uma das seguintes fórmulas:

$$Q_1 = l\,\text{inf}_i + \frac{0{,}25 - F_{i-1}}{f_i}.a_i \qquad Q_3 = l\,\text{inf}_i + \frac{0{,}75 - F_{i-1}}{f_i}.a_i$$

**Variância amostral:** $s^2 = \frac{\sum n_i . \overline{x_i}^2 - n.\overline{x}^2}{n-1} = \frac{n}{n-1}.\left(\sum f_i . \overline{x_i}^2 - \overline{x}^2\right)$ , se n ≤ 30

ou: $s^2 = \frac{\sum n_i . \overline{x_i}^2}{n} - \overline{x}^2 = \sum f_i . \overline{x_i}^2 - \overline{x}^2$ , se n > 30

**Desvio padrão:** $s = \sqrt{s^2}$

**Coeficiente de variação:** $cv = \frac{s}{\overline{x}}$ ***Nota:*** *Multiplicar por 100, para obter valor percentual.*

## 4.9 Medidas de assimetria

**Coeficiente de assimetria amostral:**

$$g_1 = \frac{n^2}{(n-1).(n-2)} . \frac{m_3}{s^3}$$ , se n ≤ 30 (amostra pequena)

ou: $g_1 = \frac{m_3}{s^3}$ , se n > 30 (amostra grande)

sendo: $m_3 = \frac{\sum (\overline{x_i} - \overline{x})^3 . n_i}{n}$ [7]

O valor do coeficiente de assimetria amostral tem o significado apresentado anteriormente.

## 4.10 Medidas de achatamento ou curtose

**Coeficiente de curtose amostral:**

$$g_2 = \frac{n^2 (n+1)}{(n-1)(n-2)(n-3)} . \frac{m_4}{s^4} - 3 . \frac{(n-1)^2}{(n-2)(n-3)}$$ , se n ≤ 30 (amostra pequena)

ou: $g_2 = \frac{m_4}{s^4} - 3$ , se n > 30 (amostra grande)

sendo: $m_4 = \frac{\sum (\overline{x_i} - \overline{x})^4 . n_i}{n}$ [8]

O valor do coeficiente de curtose amostral tem o significado apresentado anteriormente.

[7] $m_3$ é designado por terceiro momento centrado.

[8] $m_4$ é designado por quarto momento centrado.

## 4.11 Cálculos em Excel

O Excel não possui ferramentas ou funções específicas para o tratamento de dados discretos agrupados. No entanto, as suas potencialidades de base bem aplicadas aos conhecimentos teóricos do analista permitem, de forma expedita, uma análise profunda e completa deste tipo de dados.

*Exemplo 19 – Veja Exemplo 19, no ficheiro Capítulo 3 Exemplos.xls*

## 4.12 Representação gráfica dos dados

A representação gráfica em forma de o **Diagrama de "Caixa de Bigodes"** continua a ser extremamente útil numa análise da amostra. As suas construção e interpretação são em tudo semelhantes às já apresentadas – as únicas variações ocorrem nos cálculos dos valores envolvidos.

Outras representações também muito úteis no caso de dados contínuos agrupados são:

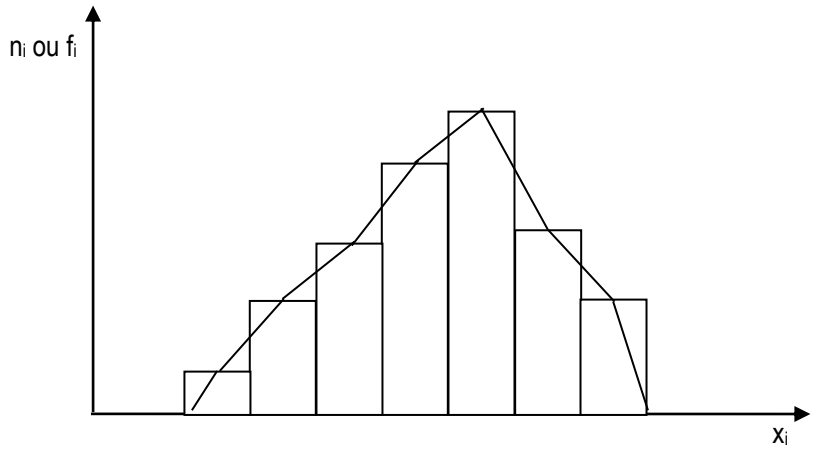
- **Histograma** (de frequências simples) – representa uma frequência simples (absoluta ou relativa em função de $x_i$ ). Obtém-se construindo sobre cada classe uma coluna com a sua largura e uma altura igual à frequência simples em representação.
Caso as classes não possuam todas a mesma amplitude, deve representar-se as frequências simples normalizadas: a representação de $n_i$ ou de $f_i$ é substituída pela representação de $n_i / a_i$ ou $f_i / a_i$ , respectivamente. Este procedimento evita a sobrerrepresentação de classes mais largas e a subrrepresentação de classes mais estreitas.

- **Polígono de frequências simples** (absolutas ou relativas) – une os pontos centrais dos topos das colunas topos do diagrama anterior com rectas. Nos extremos, é une-se o topo da primeira coluna ao ponto ( $linf_1$ , 0 ) e o topo da última coluna ao ponto ( $lsup_k$ , 0 ).

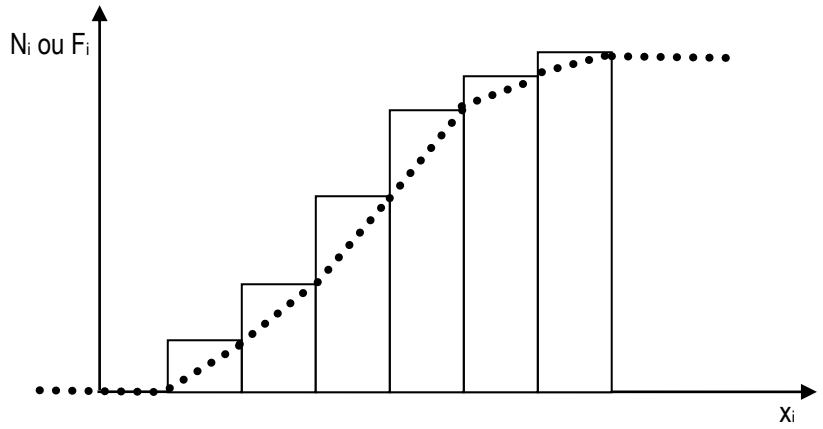
- **Histograma de frequências acumuladas** (absolutas ou relativas) – representa $N_i$ ou $F_i$ em função de $x_i$. Cada coluna i é assente sobre o espaço que une os dois limites de cada classe.
Ainda que as classes não possuam todas a mesma amplitude, não deve nunca fazer-se a normalização das frequências acumuladas.

- **Curva de frequências acumuladas** (absolutas ou relativas) – representa $N_i$ ou $F_i$ em função de $x_i$ ; une os pontos ( $lsup_i$ , $N_i$ ) ou ( $lsup_i$ , $F_i$ ). Para $linf_1 = x_{mín}$, considera-se que a frequência acumulada vale 0, bem como para valores de x a ele inferiores. Para valores de x superiores a $lsup_k = x_{máx}$, a frequência acumulada vale o seu máximo (n ou 1, respectivamente).

*Exemplo 20 - Histograma e polígono de frequências simples.*
*Esta amostra apresenta algum sintoma de assimetria negativa (cauda à esquerda) e uma curtose média..*




*Exemplo 21 - Histograma e curva de frequências acumuladas.*

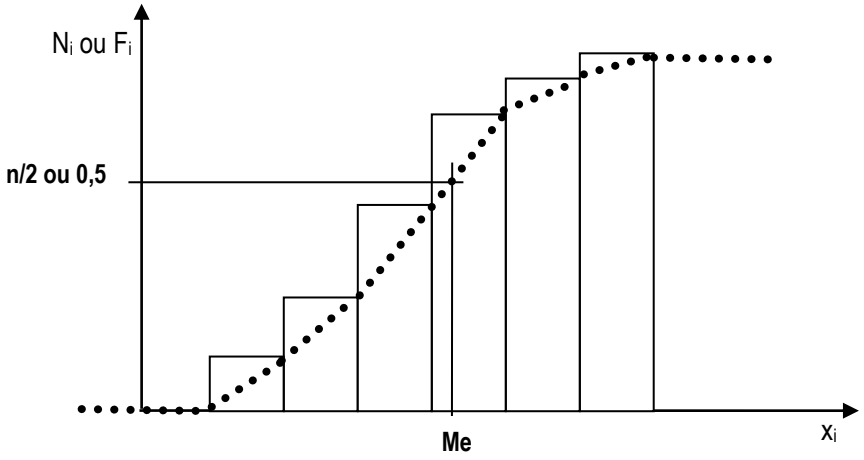



*Exemplo 22 – Veja Exemplo 22, no ficheiro Capítulo 3 Exemplos.xls*

## 4.13 Significados geométricos de moda e mediana amostrais

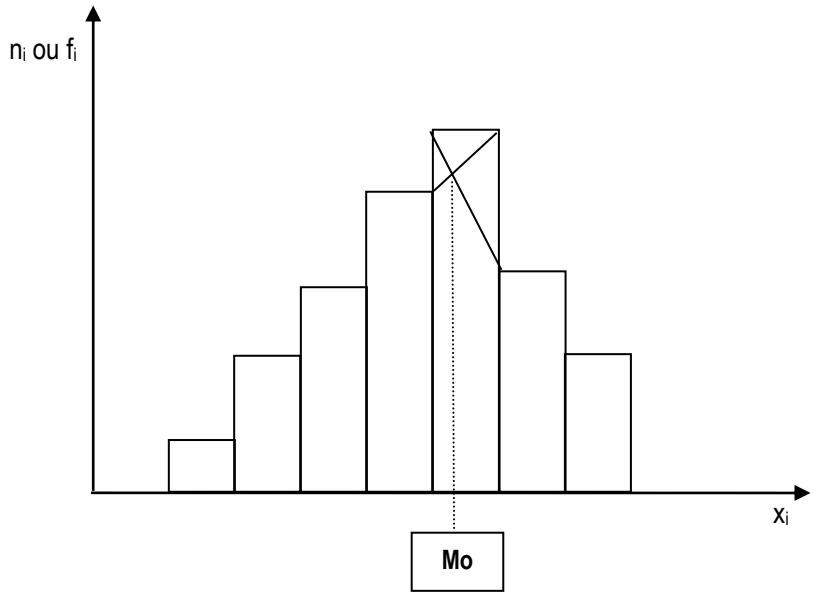
A mediana amostral corresponde ao valor de x para o qual a curva de frequências acumuladas apresenta uma ordenada igual a 0,5 ou a n/2 (consoante se trate de frequências relativas ou absolutas). A mediana é na verdade o resultado de uma interpolação linear entre dois pontos – ver exemplo 23.

*Exemplo 23 - Significado geométrico da mediana amostral.*




O significado de moda amostral pode ser observado num histograma (de frequências simples) e corresponde ao valor de x no cruzamento entre duas linhas – ver exemplo 24.

*Exemplo 24 - Significado geométrico da moda amostral.*



# 5. Tratamento de dados qualitativos

## 5.1 Características gerais

**Dimensão da amostra:** **n**

**Número de categorias:** **k** ( $k \leq n$ )

As classes deverão ser sempre mutuamente exclusivas.

## 5.2 Tabela de frequências

Cada classe é caracterizada por um valor qualitativo e, caso os dados estejam agrupados, por uma frequência simples, $n_i$ ou $f_i$.

**Frequência (simples) absoluta, $n_i$ –** Número de vezes que ocorrem valores da classe i.

**Frequência (simples) relativa, $f_i$ –** Proporção do número de vezes que a classe i ocorre na amostra em relação à sua dimensão.
Pode multiplicar-se por 100 para obter um valor percentual – $f\%_i$ .

As frequências acumuladas apenas têm significado caso os valores sejam medidos numa escala ordinal.

**Frequência acumulada absoluta, $N_i$–** Número de vezes que a classe i ou classes inferiores ocorrem na amostra.

**Frequência acumulada relativa, $F_i$ –** Proporção do número de vezes que a classe i ou classes inferiores ocorrem na amostra em relação à sua dimensão.
Pode multiplicar-se por 100 para obter um valor percentual – $F\%_i$ .

Fórmulas de relação entre as várias frequências (semelhantes às anteriores):

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

$$\sum_{i=1}^{k} n_i = n \qquad \sum_{i=1}^{k} f_i = 1 \qquad N_k = n \qquad F_k = 1$$

## 5.3 Medidas de localização

Média amostral: Não é possível calcular a média amostral de dados qualitativos pois $x_i$ não é um valor numérico.

**Mediana amostral:** **Me** A mediana é o valor de x tal que 50% ou mais dos valores amostrais lhe são iguais ou inferiores e 50 % ou mais dos valores amostrais lhe são iguais ou inferiores.

Para determinar a mediana, em primeiro lugar há que ordenar todos os valores da amostra.

Apenas podemos calcular a mediana de dados qualitativos se a escala for ordinal. Portanto a mediana não é calculável para dados nominais.

*1º processo:* Para determinar a mediana, em primeiro lugar há que calcular 0,5 x n.
Seguidamente, procura-se o valor obtido entre os valores de $N_i$ .(frequência acumulada absoluta).
Se o valor exacto for encontrado, a mediana é a intermédia entre o valor qualitativo de x correspondente e o valor de x seguinte.
Se o valor exacto não for encontrado, a mediana é o valor qualitativo de x correspondente ao primeiro $N_i$ superior a 0,5 x n.

*2º processo:* Procura-se o valor 0,5 entre os valores de $F_i$ (frequência acumulada relativa).
Se o valor exacto for encontrado, a mediana é intermédia entre o valor de x qualitativo correspondente e o valor de x seguinte.
Se o valor exacto não for encontrado, a mediana é o valor qualitativo de x correspondente ao primeiro $F_i$ superior a 0,5.

**Moda amostral:** **Mo** A moda é o valor de x mais frequente.
Podemos sempre calcular a moda de uma amostra de dados qualitativos ; a moda é o valor qualitativo com maior frequência simples (absoluta ou relativa).

*Situação especial:*

- Há mais do que um valor $x_i$ com as mesmas frequências simples máximas quer esses valores sejam ou não sejam seguidos → Há duas, três, … modas e diz-se que a amostra é bimodal, trimodal, etc…

## 5.4 Medidas de dispersão, assimetria e curtose

Não é possível calcular estas medidas para dados qualitativos.
Porém é possível realizar uma análise comparativa por observação de gráficos, quando estamos a lidar com uma escala ordinal (ver secção 5.4.5).

*Nota: Nalguns casos em que a escala seja ordinal poderá eventualmente ser de alguma utilidade o cálculo dos primeiro e terceiro quartis. No entanto, a amplitude inter-quartílica não é calculável.*

## 5.5 Representação gráfica dos dados

A representação gráfica em forma de o diagrama de Caixa de Bigodes quase sempre não é útil pela ausência de uma escala qualitativa. No entanto nalguns casos em que a escala seja ordinal poder-se-á construir um diagrama deste tipo.

Há várias representações muito úteis no caso de dados qualitativos agrupados, entre outras:

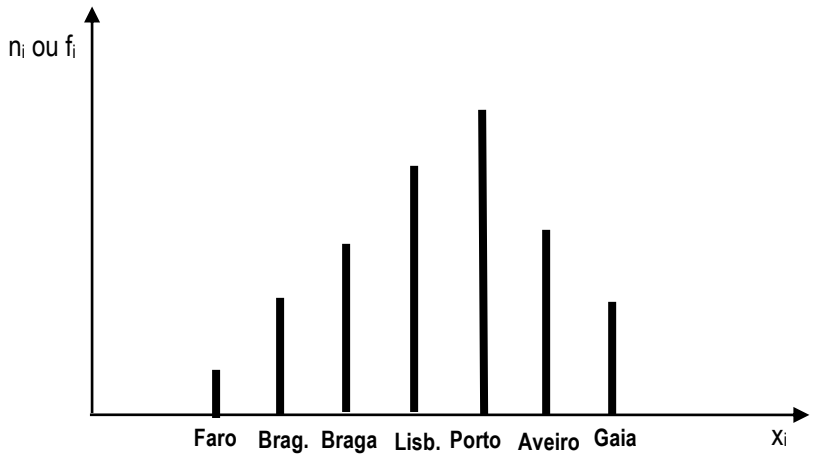
- **Gráfico de barras** e o **Histograma** – representam uma frequência simples em função de $x_i$

**- Diagrama "pie" (tarte)** – diagrama circular.

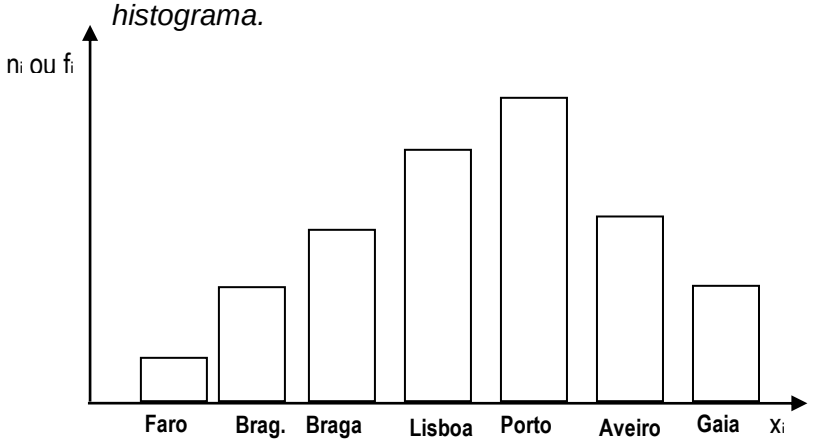
**- Histograma de frequências acumuladas** (absolutas ou relativas) – apenas útil para dados em escala ordinal ; representa $N_i$ ou $F_i$ em função de $x_i$.

**- Curva de frequências acumuladas** (absolutas ou relativas) – também apenas útil para dados ordinais ; representa-se $N_i$ ou $F_i$ em função de $x_i$ ; une os pontos ( $x_i$ , $N_i$ ) ou ( $x_i$ , $F_i$ ). Considera-se um ponto imaginário $x_{i-1}$, como sendo o último ponto para o qual a frequência acumulada vale 0. Para valores de x superiores a $x_{máx}$, a frequência acumulada vale o seu máximo (n ou 1, respectivamente).

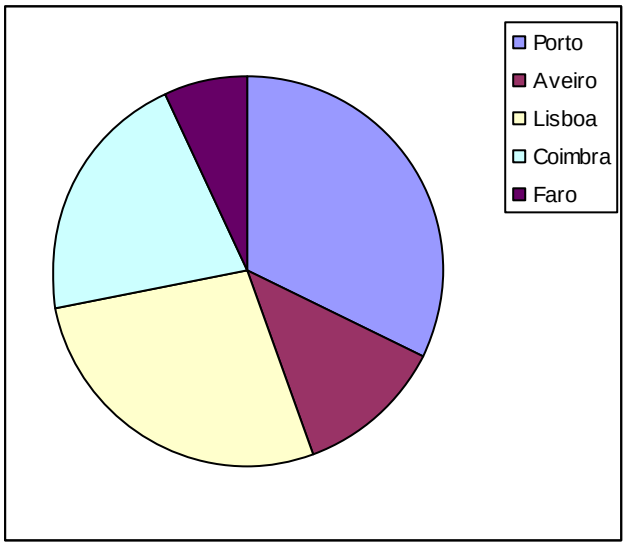
*Exemplo 25 - Esquemas de uma representação de uma amostra de dados qualitativos sob a forma de gráfico de barras.*



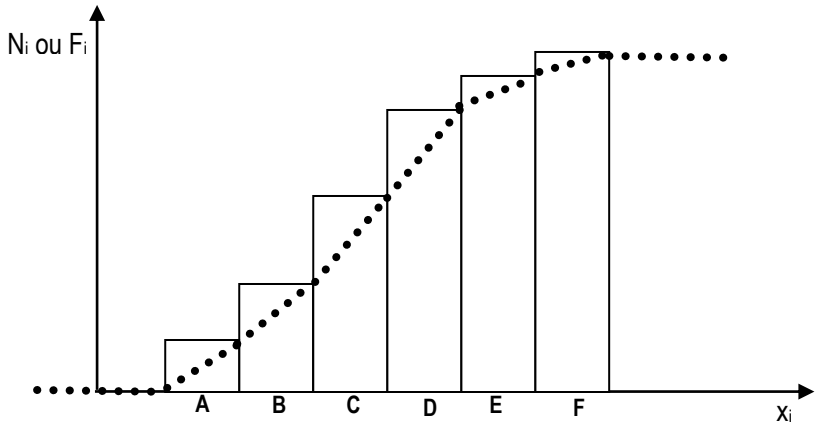

*Exemplo 26 - Representação de uma amostra de dados qualitativos sob a forma de histograma.*



*Exemplo 27 -* *Esquemas de uma representação de uma amostra de dados qualitativos sob a forma de gráfico tipo "pie" (circular).*




*Exemplo 28-* *Esquemas de uma representação de uma amostra de dados qualitativos ordinais sob a forma de histograma de frequências acumuladas e curva de frequências acumuladas.*




*Exemplo 29 – Veja Exemplo 29, no ficheiro Capítulo 3 Exemplos.xls*

*Exemplo 30 – Veja Exemplo 30, no ficheiro Capítulo 3 Exemplos.xls*

- FIM do Capítulo -