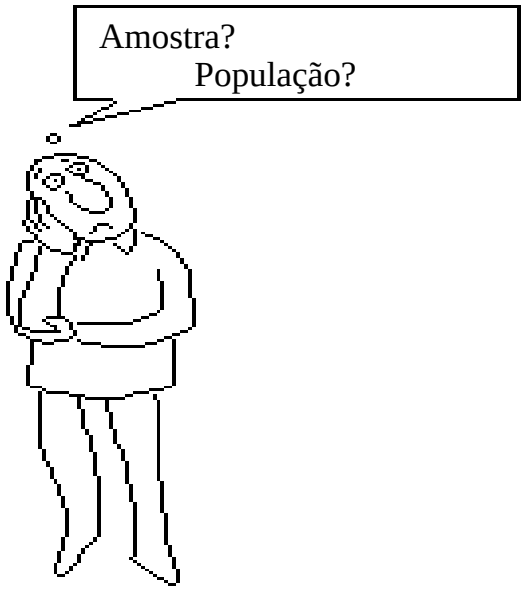
# Introdução
# à Probabilidade e à Estatística

Com complementos de Excel




**Maria Eugénia Graça Martins**

Departamento de Estatística e Investigação Operacional da FCUL
Sociedade Portuguesa de Estatística
Junho 2005

# Introdução à Probabilidade e à Estatística

Com complementos de Excel

**Maria Eugénia Graça Martins**

Departamento de Estatística e Investigação Operacional da FCUL
Sociedade Portuguesa de Estatística
Junho 2005

FICHA TÉCNICA

Título – INTRODUÇÃO À PROBABILIDADE E À ESTATÍSTICA
Com complementos de Excel

Autor – Maria Eugénia Graça Martins



ISBN – 972–8890–03-6
Depósito Legal nº 228501/05

Junho 2005

## Nota prévia

Estas folhas têm como objectivo transmitir os conhecimentos básicos de uma disciplina na área de Probabilidades e Estatística, dando uma visão geral do que é que se pretende com a Estatística e qual a sua utilidade, e ainda porque é que é necessário saber Probabilidade, para fazer Estatística.

Nesta edição, revista a partir de uma edição de 2000, introduzimos alguns complementos de Excel. Embora esta folha de cálculo não seja um software de Estatística, já nos permite fazer muita da Estatística necessária, a nível elementar, e tem a grande vantagem de estar acessível em todos os computadores, que não é o caso de qualquer software de Estatística. Achamos também que o saber trabalhar com esta folha de cálculo, permitirá minimizar os erros e as falhas cometidas ao utilizá-la para fazer Estatística.

Não pretendemos apresentar estas folhas como um produto acabado, mas sim como um novo passo para um trabalho que possa ser continuamente melhorado com as críticas e sugestões, que desde já agradeço, da parte dos alunos a que se destinam e da parte dos colegas.

A autora

*Aos meus alunos*

# Índice

**Capítulo 1**
**O que é a Estatística** 1
1.1 – Introdução 1
1.2 – Aquisição de dados: sondagens e experimentações. População e amostra 2
1.2.1 – Sondagens 2
Técnicas de amostragem aleatória 6
Amostra aleatória simples 6
Como obter uma tabela de números aleatórios 8
Utilização do Excel na selecção de uma amostra aleatória simples 8
Amostra aleatória sistemática 12
Amostragem estratificada 12
Amostragem por "Clusters" ou grupos 13
Amostragem multi-etapas 13
Utilização do Excel na selecção de uma amostra aleatória sistemática 13
Qual a dimensão que se deve considerar para a amostra? 14
Pode-se aumentar a precisão estratificando? 17
1.2.2 – Experimentações 19
1.3 – Exploração de dados 21
1.4 – Inferência Estatística 22
1.5 – Estatística Descritiva e Inferência Estatística 23
Interpretação do intervalo de confiança 25
1.6 – Exemplos de aplicação da Estatística 26
Exercícios 28

**Capítulo 2**
**Análise, representação e redução de dados** 29
2.1 – Introdução 29
2.2 – Tipos de dados 30
2.2.1 – Dados qualitativos 30
Variáveis nominais 30
Variáveis ordinais 30
2.2.2 – Dados quantitativos 33
Variáveis intervalares 33
Variáveis percentuais 33
Outras classificações 33
Como organizar os dados 34
2.2.2.1 – Organização de dados discretos 34
2.2.2.2 – Organização de dados contínuos 35
Utilização do Excel na obtenção de tabelas de frequência 37
2.3 – Representação gráfica de dados 43
2.3.1 – Variáveis discretas. Diagrama de barras 43
Utilização do Excel na construção de diagramas de barras 44
2.3.2 – Variáveis contínuas. Histograma 46
Utilização do Excel na construção de histogramas 48
2.3.3 – Outras representações gráficas 50
2.3.3.1 – Diagrama circular 50
2.3.3.2 – Caule-e-folhas 51

Utilização do caule-e-folhas para comparar duas amostras 55
Utilização do Excel na construção de um caule-e-folhas 56
2.3.3.3 – Função distribuição empírica 58
2.3.3.4 – Box-plot ou Box-and-whisker plot (caixa-com-bigodes) 61
Utilização do Excel na construção de uma Box-plot 66
Exercícios 69
2.4 – Dados bivariados 73
Utilização do Excel na construção de uma tabela de contingência 77
Exercícios 79

**Capítulo 3**
**Características amostrais** 81
3.1 – Introdução 81
3.2 – Medidas de localização 82
3.2.1 – Média 82
3.2.2 – Mediana 86
3.2.3 – Quantis. Quartis e quartos 88
3.2.4 – Médias aparadas e trimédia 89
3.2.5 – Moda 90
Exercícios 91
3.3 – Medidas de dispersão 94
3.3.1 – Variância 94
3.3.2 – Desvio padrão 95
3.3.3 – Amplitude inter-quartil 98
3.3.4 – Dispersão relativa 99
Exercícios 100
Utilização do Excel na obtenção das estatísticas descritivas 101
3.4 – Associação de variáveis 102
3.4.1 – Coeficiente de correlação 102
Utilização do Excel na construção do diagrama de pontos e no cálculo da correlação 108
Exercícios 109
3.4.2 – Associação de variáveis qualitativas 109
Paradoxo de Simpson 112
Exercício 115

**Capítulo 4**
**Regressão** 117
4.1 – Introdução 117
4.2 – Recta dos mínimos quadrados 118
Utilização do Excel na construção da recta de regressão 123
Exercícios 123

**Capítulo 5**
**Probabilidade** 125
5.1 – Introdução 125
5.2 – Experiência aleatória. Espaço de resultados. Acontecimentos 130
5.2.1 – Operações com acontecimentos 136
5.3 – Probabilidade de um acontecimento 138
5.3.1 – Probabilidade frequencista 139

Utilização do Excel na simulação de experiências aleatórias 142
5.3.2 –Probabilidade Laplaciana (ou definição clássica) 145
5.3.3 –Probabilidade subjectivista ou Bayesiana 147
5.3.4 – Definição axiomática de Probabilidade 148
Propriedades da Probabilidade 150
5.4 – Probabilidade condicional. Acontecimentos independentes 152
5.4.1 – Probabilidade condicional 152
Árvore de probabilidades 157
5.4.2 – Probabilidade da Intersecção de acontecimentos ou probabilidade conjunta dos acontecimentos A e B ou regra do produto 159
5.4.3 – Acontecimentos independentes 160
5.5 – Teorema de Bayes 163
Teorema da Probabilidade Total 165
Exercícios 166

**Capítulo 6**
**Variáveis aleatórias** 173
6.1 – Introdução 173
6.2 – Variável aleatória 173
6.2.1 – Variável aleatória discreta 175
Função massa de probabilidade 177
Utilização do Excel na simulação da experiência do lançamento de três dados 180
6.2.2 – Variável aleatória contínua 181
6.3 – Função distribuição 182
6.4 – Função densidade de probabilidade 186
Exercícios 189
6.5 – Pares de variáveis aleatórias 191
6.5.1 – Introdução 191
6.5.2 – Distribuição de probabilidade conjunta 191
6.5.3 – Variáveis aleatórias independentes 193
Exercícios 193

**Capítulo 7**
**Características populacionais** 195
7.1 – Introdução 195
7.2 – Valor médio 196
Lei dos grandes números 197
7.2.1 – Propriedades do valor médio 199
7.3 – Quantil de probabilidade p 200
Mediana 201
7.4 – Variância (populacional) 202
7.4.1 – Desvio padrão (populacional) 202
7.5 – Covariância 204
7.5.1 – Coeficiente de correlação 205
7.6 – Regressão de Y em X 206
Coeficiente de determinação 210
Exercícios 211

**Capítulo 8**
**Alguns modelos de probabilidade** 213
8.1 – Introdução 213
8.2 – Modelos discretos 214
8.2.1 – Modelo Uniforme 214
8.2.2 – Modelo Binomial 214
Amostragem com reposição 219
Amostragem sem reposição em populações infinitas 219
8.2.3 – Modelo Binomial Negativa 220
8.2.4 - Modelo de Poisson 223
Aproximação da distribuição Binomial pela distribuição de Poisson 224
8.2.5 – Modelo Hipergéométrico 228
Utilização do Excel para calcular probabilidades dos Modelos Discretos 231
8.3 – Modelos contínuos 233
8.3.1 – Modelo Normal 233
8.3.2 – Modelo Uniforme 238
Transformação uniformizante 240
8.3.3 – Modelo Exponencial 240
Utilização do Excel para calcular probabilidades dos Modelos Contínuos. 241
8.4 – Compreender a simulação 242
Utilização do Excel para gerar números pseudo-aleatórios com determinadas distribuições 244
Exercícios 244

**Capítulo 9**
**Distribuições de amostragem** 249
9.1 – Introdução 249
9.2 – Distribuição de amostragem da média 251
9.2.1 – Valor médio e desvio padrão da média 251
9.2.2 – Distribuição da média para populações Normais 252
9.2.2.1 – Desvio padrão $\sigma$ conhecido 252
9.2.2.2 – Desvio padrão $\sigma$ desconhecido 252
9.2.3 – Distribuição da média para populações não Normais. Teorema do Limite Central 253
Aplicações do Teorema Limite Central 258
Aproximação da Distribuição Binomial, pela Normal 258
Aproximação da Distribuição de Poisson, pela Normal 259
9.3 – Distribuição de amostragem da proporção 262
9.3.1 – Valor médio e variância do estimador da proporção populacional 265
9.3.2 – Distribuição de amostragem do estimador da proporção 265
Exercícios 266

**Capítulo 10**
**Introdução à Estimação** 269
10.1 – Noções prelimunares sobre estimação. Estimadores pontuais e intervalares 269
10.2 – Estimação da proporção. Intervalo de confiança 270
Confiança e precisão 273
10.3 – Estimação do valor médio. Intervalo de confiança para o valor médio 275

10.3.1 – Intervalo de confiança para o valor médio – σ conhecido 275
10.3.2 - Intervalo de confiança para o valor médio – σ desconhecido 278
Utilização do Excel para obter quantis da Normal e da t-Student 281
Exercícios 281

**Capítulo 11**
**Introdução aos testes de hipóteses** 285
11.1 – Introdução 285
11.2 – Outros exemplos 287
11.3 – Hipótese nula e Hipótese alternativa; erros de tipo 1 e tipo 2; estatística de teste; região de rejeição 288
11.4 – Testes de hipóteses sobre a proporção p 290
11.4.1 – Determinação dos pontos críticos para grandes amostras 293
11.4.2 – P-value 293
11.5 – Vamos conversar acerca de testes 294
11.6 - Testes de hipóteses sobre o valor médio 297
11.6.1 – P-value 300
Exercícios 300

**Capítulo 12**
**Introdução aos testes de ajustamento** 305
12.1 – Introdução 305
12.2 – Generalização do modelo Binomial: o modelo Multinomial 305
12.3 – Teste de ajustamento do Qui-quadrado para variáveis qualitativas 307
12.4 - Teste de ajustamento do Qui-quadrado para variáveis quantitativas discretas 311
12.5 - Teste de ajustamento do Qui-quadrado para variáveis quantitativas contínuas 313
Exercícios 319

**Bibliografia** 321

**Capítulo 1**

# O que é a Estatística?

## 1.1 - Introdução

Não é uma tarefa simples definir o que é a Estatística. Por vezes define-se como sendo um conjunto de técnicas de tratamento de dados, mas é muito mais do que isso! A Estatística é uma "**arte**" e uma **ciência** que permite tirar conclusões e de uma maneira geral fazer inferências a partir de conjuntos de dados.

Até 1900, a Estatística resumia-se ao que hoje em dia se chama Estatística Descritiva. Apesar de tudo, deu contribuições muito positivas em várias áreas científicas.

A necessidade de uma maior formalização nos métodos utilizados, fez com que, nos anos seguintes, a Estatística se desenvolvesse numa outra direcção, nomeadamente no que diz respeito ao desenvolvimento de métodos e técnicas de Inferência Estatística. Assim, por volta de 1960 os textos de Estatística debruçam-se especialmente sobre métodos de estimação e de testes de hipóteses, assumindo determinadas famílias de modelos, descurando os aspectos práticos da análise dos dados.

Porém, na última década, em grande parte devido às facilidades computacionais postas à sua disposição, os Estatísticos têm-se vindo a preocupar cada vez mais, com a necessidade de desenvolver métodos de análise e exploração dos dados, que dêem uma maior importância aos dados e que se traduz na seguinte frase: "**Devemos deixar os dados falar por si**".

O significado dos termos *Estatística Descritiva* e *Inferência Estatística* será precisado, um pouco mais à frente.

Além do significado considerado anteriormente, o termo *estatísticas*, de um modo geral no plural, também é utilizado para indicar números, como por exemplo as estatísticas fornecidas pelo governo: estatísticas da saúde – nº de doentes assistidos em cada centro de saúde; estatísticas da educação – percentagem de alunos candidatos ao ensino superior que tiveram lugar nas instituições públicas; estatísticas da energia – consumo médio de electricidade per capita, etc.

## 1.2 - Aquisição de dados: sondagens e experimentações. População e amostra

O mundo que nos rodeia será mais facilmente compreendido se puder ser quantificado. Em todas as áreas do conhecimento é necessário saber “o que medir” e “como medir”. A Estatística é a ciência que ensina a recolher **dados** válidos, assim como a interpretá-los.

Perante um conjunto de dados podem-se distinguir duas metodologias de aproximação:

- por vezes o estatístico é confrontado com conjuntos de dados sem ter qualquer ideia preconcebida sobre o que é que vai encontrar e então procede a uma **análise exploratória de dados**, quase sempre utilizando processos gráficos, análise esta que revelará aspectos do comportamento dos dados; neste caso não se fala em amostras, mas sim conjuntos de dados (Murteira, 1993) e de uma maneira geral a análise exploratória é suficiente para os fins que se têm em vista;
- em outros casos procede à análise de dados com propósitos bem definidos no sentido de responder a questões específicas. Neste caso os dados têm que ser produzidos por meio de técnicas adequadas de forma a que resultem dados válidos (amostras representativas). Estas técnicas, em que é fundamental a intervenção do **acaso**, revolucionaram e fizeram progredir a maior parte dos campos da ciência aplicada. Pode-se dizer que hoje em dia não existe área do conhecimento para cujo progresso não tenha contribuído a Estatística. Abordaremos de seguida algumas dessas técnicas de produção de dados, em que se distinguem as

**Sondagens** e **Experimentações (aleatorizadas)**

Não é demais realçar a importância desta fase, a que chamamos de Produção ou Aquisição de Dados. Como é referido em Tannenbaum (1998), pag 426: "Behind every statistical statement there is a story, and like a story it has a beginning, a middle, an end, and a moral. In this first statistics chapter we begin with the beginning, which in statistics typically means the process of gathering or collecting data. Data are the raw material of which statistical information is made, and in order to get good statistical information one needs good data".

Antes de começar a recolha de dados é fundamental, face a determinado problema, identificar correctamente a População sobre a qual se pretende recolher informação.

### 1.2.1 - Sondagens

O objectivo de uma **sondagem** é o de recolher informação acerca de uma população, seleccionando e observando um conjunto de elementos dessa população.

**Sondagem** – Estudo estatístico de uma população, feito através de uma amostra, destinado a estudar uma ou mais características tal como elas se apresentam nessa população.

Por exemplo, numa fábrica de parafusos o departamento de controlo de qualidade pretende saber qual a percentagem de parafusos defeituosos. Tempo, custos e outros inconvenientes impedem a inspecção de todos os parafusos. Assim a informação pretendida será obtida à custa de uma parte do conjunto - **amostra**, mas com o objectivo de tirar conclusões para o conjunto todo - **população**. Se se observarem todos os elementos da população tem-se um **recenseamento**. Por

vezes confunde-se sondagem com amostragem. No entanto a amostragem diz respeito ao procedimento da recolha da amostra qualquer que seja o estudo estatístico que se pretenda fazer. Assim, a amostragem é uma das fases das sondagens, já que estas devem incluir ainda o estudo dos dados recolhidos, assim como a elaboração do relatório final.

**População, unidade, amostra**

**População** é o conjunto de objectos, indivíduos ou resultados experimentais acerca do qual se pretende estudar alguma característica comum. Aos elementos da população chamamos **unidades estatísticas**.

**Amostra** é uma parte da população que é observada com o objectivo de obter informação para estudar a característica pretendida.

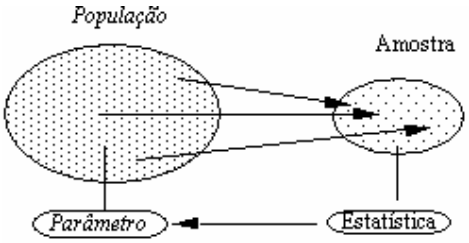
Geralmente, há algumas quantidades numéricas acerca da população que se pretendem conhecer. A essas quantidades chamamos **parâmetros**.

Ao estudar a população constituída por todos os potenciais eleitores para as legislativas, dois parâmetros que podem ter interesse são:

- **idade média** dos potenciais eleitores que estão decididos a votar;
- **percentagem** de eleitores que estão decididos a votar.

Para conhecer aqueles parâmetros, teria de se perguntar a cada eleitor a sua idade, assim como a sua intenção no que diz respeito a votar ou não. Esta tarefa seria impraticável, nomeadamente por questões de tempo e de dinheiro.

Os **parâmetros** são estimados por **estatísticas**, que são números calculados a partir da amostra. No caso do exemplo anterior, à característica populacional "percentagem de eleitores que estão decididos a votar" corresponde a característica amostral "percentagem dos 1000 eleitores (supõe-se que entretanto se recolheu uma amostra de dimensão 1000), que interrogados disseram estar decididos a votar". Estas quantidades são conceptualmente distintas, pois enquanto a característica populacional pode ser considerada um valor exacto, embora desconhecido, a característica amostral é conhecida, embora contendo um certo erro, mas que todavia pode ser considerada uma estimativa útil da característica populacional respectiva.



No entanto para se poder utilizar as estatísticas para estimar parâmetros é necessário que as amostras sejam representativas das populações de onde foram retiradas. Uma amostra que não seja representativa da População diz-se **enviesada** e a sua utilização pode dar origem a interpretações erradas, como se sugere nos seguintes exemplos:

- utilizar uma amostra constituída por 10 benfiquistas, para prever o vencedor do próximo Benfica-Sporting!
- utilizar uma amostra constituída por leitores de determinada revista especializada, para tirar conclusões sobre a população em geral.

Surge assim, a necessidade de fazer um **planeamento da amostragem**, onde se decide quais e como devem ser recolhidos os dados. De um modo geral, o trabalho do Estatístico deve começar antes de os dados serem recolhidos. Deve planear o modo de os recolher, de forma a que, posteriormente, se possa extrair o máximo de informação relevante para o problema em estudo, ou seja para a população de onde os dados foram recolhidos e de modo a que os resultados obtidos possam ser considerados válidos. Vem a propósito referir a seguinte frase de Fisher: "*Ao pedir a um Estatístico que diagnostique dados já recolhidos, muitas vezes só se obtém uma autópsia"*.

O planeamento de um estudo estatístico, que começa com a forma de seleccionar a amostra, deve ser feito de forma a evitar **amostras enviesadas**. Alguns processos que provocam quase sempre amostras enviesadas são, por exemplo, a **amostragem por conveniência** e a obtenção de uma amostra por **resposta voluntária.** Este último processo foi usado, com muita frequência, por uma estação de televisão, com resultados por vezes contraditórios com os que se obtêm quando se utiliza um processo correcto de seleccionar a amostra.

A utilização de uma amostragem por conveniência também se realiza frequentemente, quando se selecciona a amostra a partir de uma listagem dos elementos de determinado clube ou grupo, como por exemplo a Ordem dos Engenheiros. A seguir apresentamos exemplos de más amostras ou amostras enviesadas e resultado da sua aplicação:

<u>Amostra 1</u> - A SIC pretende saber qual a percentagem de pessoas que é a favor da despenalização do aborto. Para isso indicou dois números de telefone, um dos quais para as respostas SIM e o outro para a resposta NÃO.
Resultado - A utilização da percentagem de respostas positivas como indicação da percentagem da população portuguesa que é a favor da despenalização do aborto é enganadora. Efectivamente só uma pequena percentagem da população responde a estas questões e de um modo geral tendem a ser pessoas com a mesma opinião.
<u>Amostra 2</u> - Uma estação de televisão preparou um debate sobre o aumento de criminalidade, onde enfatizou o facto de terem aumentado os crimes violentos. Ao mesmo tempo decorria uma sondagem de opinião sobre se as pessoas eram a favor da implementação da pena de morte.

Esta recolha de opiniões era feita no molde descrito no exemplo anterior, isto é, por resposta voluntária.

Resultado - A utilização da percentagem de SIM´s, que naturalmente se espera elevada, dá uma indicação errada sobre a opinião da população em geral. As pessoas influenciadas pelo debate e pelo medo da criminalidade serão levadas a telefonar dando indicação de estarem a favor da pena de morte.

Amostra 3 - Opiniões de alguns leitores de determinada revista técnica, para representar as opiniões dos portugueses em geral.

Resultado - Diferentes tipos de pessoas lêem diferentes tipos de revistas, pelo que a amostra não é representativa da população. Basta pensar que, de um modo geral, a população feminina ainda não adere às revistas técnicas como a população masculina. A amostra daria unicamente indicações sobre a população constituída pelos leitores da tal revista.

Amostra 4 - Utilizar alguns alunos de uma turma, para tirar conclusões sobre o aproveitamento de todos os alunos da escola.

Resultado - Poderíamos concluir que o aproveitamento dos alunos é pior ou melhor do que na realidade é. As turmas de uma escola não são todas homogéneas, pelo que a amostra não é representativa dos alunos da escola. Poderia servir para tirar conclusões sobre a população constituída pelos alunos da turma.

Amostra 5 - Utilizar os jogadores de uma equipa de basquete de uma determinada escola para estudar as alturas dos alunos dessa escola.

Resultado - O estudo concluiria que os estudantes são mais altos do que na realidade são.

Os exemplos que apresentámos anteriormente são exemplos de amostras não aleatórias porque tiveram a intervenção do factor humano. Estas amostras são quase sempre enviesadas. Com o objectivo de minimizar o enviesamento, no planeamento da escolha da amostra deve ter-se presente o princípio da aleatoriedade de forma a obter uma amostra aleatória.

**Amostra aleatória e amostra não aleatória** – Dada uma população, uma amostra aleatória é uma amostra tal que qualquer elemento da população tem alguma probabilidade de ser seleccionado para a amostra. Numa amostra não aleatória, alguns elementos da população podem não poder ser seleccionados para a amostra.

Apresentamos a seguir algumas técnicas para obter amostras aleatórias.

**Técnicas de amostragem aleatória**

Seguidamente apresentaremos alguns dos planeamentos mais utilizados para seleccionar amostras aleatórias. Dos vários tipos de planeamento utilizados, destacam-se os que conduzem a amostras aleatórias simples, amostras aleatórias com reposição, amostras sistemáticas e amostras estratificadas.

**Amostragem aleatória simples (sem reposição) e amostragem aleatória com reposição**

O plano de amostragem aleatória mais básico é o que permite obter a amostra aleatória simples:

**Amostra aleatória simples** - Dada uma população, uma amostra aleatória simples de dimensão n é um conjunto de n unidades da população, tal que qualquer outro conjunto dos $\binom{N}{n}$ conjuntos diferentes de n unidades, teria igual probabilidade de ser seleccionado.

Se uma população tem dimensão N e se pretende uma amostra aleatória simples de dimensão n, esta amostra é recolhida aleatoriamente de entre todas as $\binom{N}{n} = \frac{N!}{n!\,(N-n)!} =$ $= \frac{N(N-1)(N-2)...(N-n+1)}{n(n-1)(n-2)...1}$ amostras distintas que se podem recolher da população. Isto implica que cada amostra tenha a mesma probabilidade $\binom{N}{n}^{-1}$ de ser seleccionada. Uma amostra destas pode ser escolhida sequencialmente da população, escolhendo um elemento de cada vez, sem reposição, pelo que em cada selecção cada elemento tem a mesma probabilidade de ser seleccionado. Um esquema de amostragem aleatória simples, conduz a que cada elemento da População tenha a mesma probabilidade de ser seleccionado para a amostra. No entanto existem outros esquemas de amostragem em que cada elemento tem igual probabilidade de ser seleccionado, sem que cada conjunto de n elementos tenha a mesma probabilidade de ser seleccionado. É o que se passa com a amostragem aleatória sistemática, de que falaremos adiante.

**Amostragem com reposição**

Na amostragem com reposição, sempre que um elemento é seleccionado, ele é reposto na população, antes de seleccionar o seguinte, ao contrário do que acontece na amostragem sem reposição. Intuitivamente conseguimos apercebermo-nos de que se a dimensão da população for "grande", quando comparada com a dimensão da amostra, estes dois tipos de amostragem podem ser considerados de certo modo equivalentes, já que a probabilidade de seleccionar o mesmo elemento duas vezes é "muito pequena".

Dada uma população de dimensão N, referir-nos-emos a uma **amostra aleatória** de dimensão n, **com reposição**, como um conjunto de n unidades da população, tal que qualquer outro conjunto dos $N^n$ conjuntos diferentes de n unidades, teria igual probabilidade de ser seleccionado.

A probabilidade de cada uma das amostras ser seleccionada é igual a $1/N^n$.

Exemplificamos a seguir um processo de obter uma amostra aleatória simples.

**Exemplo 1.3.1.1** - Consideremos a população constituída pelos 18 alunos de uma turma do 10º ano de uma determinada Escola Secundária, em que a característica de interesse a estudar é a altura média desses alunos. Uma maneira possível de recolher desta população uma amostra aleatória, seria escrever cada um dos indicadores (nº do aluno, nome, …) dos elementos da população num quadrado de papel, inserir todos esses bocados de papel numa caixa e depois seleccionar tantos quantos a dimensão da amostra desejada.
A recolha tem de ser feita **sem reposição** pois quando se retira um papel (elemento da população), ele não é reposto enquanto a amostra não estiver completa (com a dimensão desejada). Qualquer conjunto de números recolhidos desta forma dará origem a uma amostra

aleatória simples, constituída pelas alturas dos alunos seleccionados (desde que se tenha o cuidado de cortar os bocadinhos de papel todos do mesmo tamanho, para ficarem semelhantes, e de os baralhar convenientemente). A partir de cada amostra, pode-se calcular o valor da estatística média, que será uma estimativa do parâmetro a estudar - valor médio da altura dos alunos da turma. Obter-se-ão tantas estimativas, quantas as amostras retiradas.

Chama-se a atenção para o facto de nesta altura não se poder dizer qual das estimativas é "melhor", isto é, qual delas é uma melhor aproximação do parâmetro a estimar, já que esse parâmetro é desconhecido (obviamente que nesta população tão pequena seria possível estudar exaustivamente todos os seus elementos, não sendo necessário recolher nenhuma amostra - este exemplo só serve para ilustrar uma situação)!

O processo que acabámos de descrever não é prático se a população a estudar tiver dimensão elevada. Neste caso, um processo de seleccionar uma amostra aleatória simples consiste em utilizar uma tabela de números aleatórios.

**Dígitos aleatórios** - Uma tabela de dígitos aleatórios é uma listagem dos dígitos 0, 1, 2, 3, 4, 5, 6, 7, 8 ou 9 tal que:

- qualquer um dos dígitos considerados tem igual possibilidade de figurar em qualquer posição da lista;

- a posição em que figura cada dígito é independente das posições dos outros dígitos.

Apresenta-se a seguir um extracto de uma tabela de números aleatórios (Moore, 1997). O facto de os dígitos se apresentarem agrupados 5 a 5 é só para facilidade de leitura.

| **Linha** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **101** | 19223 | 95034 | 05756 | 28713 | 96409 | 12531 | 42544 | 82853 |
| **102** | 73676 | 47150 | 99400 | 01927 | 27754 | 42648 | 82425 | 36290 |
| **103** | 45467 | 71709 | 77558 | 00095 | 32863 | 29485 | 82226 | 90056 |
| **104** | 52711 | 38889 | 93074 | 60227 | 40011 | 85848 | 48767 | 52573 |
| **105** | 95592 | 94007 | 69971 | 91481 | 60779 | 53791 | 17297 | 59335 |
| **106** | 68417 | 35013 | 15529 | 72765 | 85089 | 57067 | 50211 | 47487 |
| **107** | 82739 | 57890 | 20807 | 47511 | 81676 | 55300 | 94383 | 14893 |
| **108** | 60940 | 72024 | 17868 | 24943 | 61790 | 90656 | 87964 | 18883 |
| **109** | 36009 | 19365 | 15412 | 39638 | 85453 | 46816 | 83485 | 41979 |

A partir da tabela de dígitos aleatórios podem-se obter números aleatórios de 2 dígitos - qualquer par dos 100 pares possíveis 00, 01, …98, 99, tem igual probabilidade de ser seleccionado, de 3 dígitos - qualquer triplo dos 1000 triplos possíveis 000, 001, …998, 999, tem igual probabilidade de ser seleccionado, etc, tomando os dígitos da tabela 2 a 2, 3 a 3, etc, a partir de uma linha qualquer e percorrendo-a da esquerda para a direita.

Para seleccionar uma amostra de uma população utilizando a tabela procede-se em duas etapas:
- atribui-se um número a cada elemento da população. Esta atribuição terá de ser feita com as devidas precauções, de forma a que cada número tenha o mesmo número de dígitos, para ter igual probabilidade de ser seleccionado;
- a partir da tabela escolhe-se uma linha ao acaso e começa-se a percorrê-la da esquerda para a direita, tomando de cada vez os dígitos necessários.

**Exemplo 1 (cont) -** Considerando a população do exemplo anterior, constituída por 18 elementos, vamos numerá-los com os números 01, 02, 03, …, 17, 18 (podia ser utilizado qualquer outro conjunto de 18 números de 2 dígitos). Para seleccionar uma amostra de dimensão 4 fixamo-nos numa linha qualquer da tabela, por exemplo a linha 107 e começamos a seleccionar os números de dois dígitos, tendo-se obtido:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 82 | 73 | 95 | 78 | 90 | 20 | 80 | 74 | 75 | 11 | 81 |
| 67 | 65 | 53 | 00 | 94 | 38 | 31 | 48 | 93 | 60 | 94 |
| 07 | 20 | 24 | 17 | 86 | 82 | 49 | 43 | 61 | 79 | 09 |

Tivemos de ler 33 números, dos quais só aproveitámos 4, pois os outros não correspondiam a elementos da população.

**Como obter uma tabela de números aleatórios?**

Um processo poderá consistir em meter numa caixa 10 bolas numeradas de 0 a 9 e fazer várias extracções de uma bola, tantas quantas os dígitos que se pretendem para constituir a tabela. De cada vez que se faz uma extracção, lê-se o número da bola, aponta-se e repõe-se a bola na caixa - extracção *com reposição.* Com este processo qualquer dígito tem igual probabilidade de ser seleccionado. Além disso a saída de qualquer um dos dígitos em qualquer momento, é independente dos dígitos que já saíram anteriormente.

Além das tabelas de números aleatórios também existe a possibilidade de utilizar o computador para os gerar ou uma simples máquina de calcular. Este é o processo mais utilizado hoje em dia, mas convém ter presente que os números que se obtêm são *pseudo-aleatórios*, já que é um mecanismo determinista que lhes dá origem, embora se comportem como números aleatórios (passam numa bateria de testes destinados a confirmar a sua aleatoriedade).

**Utilização do Excel na selecção de uma amostra aleatória simples e de uma amostra aleatória com reposição**

No exemplo seguinte, apresentamos uma forma simples de utilizar o Excel para seleccionar uma amostra aleatória simples e uma amostra aleatória, com reposição, de uma População finita, de que se tenha uma listagem dos elementos.

**Exemplo** – Considere a seguinte a lista de Escolas Secundárias de Portugal continental. Utilizando o Excel, extraia uma amostra aleatória simples, de 10 escolas.

| Nome | Distrito | Local |
|---|---|---|
| Escola secundária Alves Martins | Viseu | Viseu |
| Escola secundária Amélia Rey Colaço | Lisboa | Oeiras |
| Escola secundária com 3º ciclo do ensino básico José Afonso | Lisboa | Loures |
| Escola secundária D. Afonso Henriques | Porto | Santo Tirso |
| Escola secundária da Cidade Universitária | Lisboa | Lisboa |
| Escola secundária da Lourinhã | Lisboa | Lourinhã |
| Escola secundária da Moita | Setúbal | Moita |
| Escola secundária da Sertã | Castelo Branco | Sertã |
| Escola secundária David Mourão Ferreira | Lisboa | Lisboa |
| Escola secundária de Albufeira | Faro | Albufeira |
| Escola secundária de Alves Redol | Lisboa | Vila Franca de Xira |
| Escola secundária de Arganil | Coimbra | Arganil |
| Escola secundária de Avelar Brotero | Coimbra | Coimbra |
| Escola secundária de Benavente | Santarém | Benavente |

| | | |
|---|---|---|
| Escola secundária de Cantanhede | Coimbra | Cantanhede |
| Escola secundária de Cascais | Lisboa | Cascais |
| Escola secundária de Coelho e Castro - Fiães | Aveiro | Santa Maria da Feira |
| Escola secundária de D. Duarte | Coimbra | Coimbra |
| Escola secundária de D. Luís de Castro | Braga | Braga |
| Escola secundária de D. Pedro V | Lisboa | Lisboa |
| Escola secundária de D. Sancho II | Portalegre | Elvas |
| Escola secundária de D.Manuel I | Beja | Beja |
| Escola secundária de Dom Manuel Martins | Setúbal | Setúbal |
| Escola secundária de Domingos Sequeira | Leiria | Leiria |
| Escola secundária de Francisco Rodrigues Lobo | Leiria | Leiria |
| Escola secundária de Gabriel Pereira | Évora | Évora |
| Escola secundária de Gago Coutinho | Lisboa | Vila Franca de Xira |
| Escola secundária de Gil Eanes | Faro | Lagos |
| Escola secundária de Homem Cristo | Aveiro | Aveiro |
| Escola secundária de Jaime Cortesão | Coimbra | Coimbra |
| Escola secundária de João de Deus | Faro | Faro |
| Escola secundária de José Falcão | Coimbra | Coimbra |
| Escola secundária de Júlio Dantas | Faro | Lagos |
| Escola secundária de Loulé | Faro | Loulé |
| Escola secundária de Manuel Teixeira Gomes | Faro | Portimão |
| Escola secundária de Maria Amália Vaz de Carvalho | Lisboa | Lisboa |
| Escola secundária de Montemor-o-Velho | Coimbra | Montemor-o-Velho |
| Escola secundária de Odivelas | Lisboa | Odivelas |
| Escola secundária de Oliveira do Bairro | Aveiro | Oliveira do Bairro |
| Escola secundária de Pombal | Leiria | Pombal |
| Escola secundária de S. João da Talha | Lisboa | Loures |
| Escola secundária de S. João do Estoril | Lisboa | Cascais |
| Escola secundária de Santa Maria - Sintra | Lisboa | Sintra |
| Escola secundária de Seia | Guarda | Seia |
| Escola secundária de Silves | Faro | Silves |
| Escola secundária de Stº André | Setúbal | Barreiro |
| Escola secundária de Tavira | Faro | Tavira |
| Escola secundária de Tomás Cabreira | Faro | Faro |
| Escola secundária de Vendas Novas | Évora | Vendas Novas |
| Escola secundária de Vitorino Nemésio | Lisboa | Lisboa |
| Escola secundária Diogo de Gouveia | Beja | Beja |
| Escola secundária do Dr. Francisco Fernandes Lopes | Faro | Olhão |
| Escola secundária do Eng. Acácio Calazans Duarte | Leiria | Marinha Grande |
| Escola secundária do Forte da Casa | Lisboa | Vila Franca de Xira |
| Escola secundária do Infante D. Pedro | Lisboa | Vila Franca de Xira |
| Escola secundária do Prof. Reynaldo dos Santos | Lisboa | Vila Franca de Xira |
| Escola secundária do Professor Herculano de Carvalho | Lisboa | Lisboa |
| Escola secundária Dr. Bernardino Machado | Coimbra | Figueira da Foz |
| Escola secundária Dr. Manuel Candeias Gonçalves -Odemira | Beja | Odemira |
| Escola secundária Emídio Navarro | Viseu | Viseu |
| Escola secundária Infanta D. Maria | Coimbra | Coimbra |
| Escola secundária Jacôme Ratton | Santarém | Tomar |
| Escola secundária José Belchior Viegas - São Brás de Alportel | Faro | São Brás de Alportel |
| Escola secundária José Saramago | Lisboa | Mafra |
| Escola secundária Marques de Castilho | Aveiro | Águeda |
| Escola secundária Martinho Árias | Coimbra | Soure |
| Escola secundária Monserrate | Viana do Castelo | Viana do Castelo |
| Escola secundária Poeta António Aleixo | Faro | Portimão |
| Escola secundária Santa Maria Maior | Viana do Castelo | Viana do Castelo |
| Escola secundária Sebastião e Silva | Lisboa | Oeiras |

Começámos por criar um ficheiro, em Excel, com os dados das escolas, considerando ainda uma coluna onde inserimos o número da escola, segundo a ordem pela qual as escolas são apresentadas (esta ordem, que no caso presente, é a ordem cescente, não é importante para o que se segue), a que chamámos EscolasSec.xls e do qual apresentamos um pequeno pedaço:

EscolasSec.xls

| | A | B | C | D |
|---|---|---|---|---|
| 1 | | **Nome** | **Distrito** | **Local** |
| 2 | 1 | Escola secundária Alves Martins | Viseu | Viseu |
| 3 | 2 | Escola secundária Amélia Rey Colaço | Lisboa | Oeiras |
| 4 | 3 | Escola secundária com 3º ciclo do ensino básico José Afonso | Lisboa | Loures |
| 5 | 4 | Escola secundária D. Afonso Henriques | Porto | Santo Tirso |
| 6 | 5 | Escola secundária da Cidade Universitária | Lisboa | Lisboa |
| 7 | 6 | Escola secundária da Lourinhã | Lisboa | Lourinhã |
| 8 | 7 | Escola secundária da Moita | Setúbal | Moita |
| 9 | 8 | Escola secundária da Sertã | Castelo Bran | Sertã |
| 10 | 9 | Escola secundária David Mourão Ferreira | Lisboa | Lisboa |

Sheet1 **Sheet2** Sheet3

## Amostra aleatória simples

1º passo - Utilizando a função *RAND()*, atribuir um número aleatório, entre 0 e 1, a cada escola. Para isso basta inserir a função na célula E2 e replicá-la tantas vezes, quantos as escolas (ou seja, 70 vezes):

| | A | B | E |
|---|---|---|---|
| 1 | | Nome | |
| 2 | 1 | Escola secundária Alves Martins | =RAND() |
| 3 | 2 | Escola secundária Amélia Rey Colaço | =RAND() |
| 4 | 3 | Escola secundária com 3º ciclo do ensino básico José Afonso | =RAND() |
| 5 | 4 | Escola secundária D. Afonso Henriques | =RAND() |
| 6 | 5 | Escola secundária da Cidade Universitária | =RAND() |
| 7 | 6 | Escola secundária da Lourinhã | =RAND() |
| 8 | 7 | Escola secundária da Moita | =RAND() |
| 9 | 8 | Escola secundária da Sertã | =RAND() |
| 10 | 9 | Escola secundária David Mourão Ferreira | =RAND() |
| 11 | 10 | Escola secundária de Albufeira | =RAND() |
| 12 | 11 | Escola secundária de Alves Redol | =RAND() |
| 13 | 12 | Escola secundária de Arganil | =RAND() |

**Sheet1** Sheet2 Sheet3

Para visualizar as fórmulas na folha de Excel, bastou seleccionar: *Tools → Options → View → Formulas →Ok*:

Uma vez que a função *RAND()* é uma função volátil, isto é, muda quando se recalcula a folha, no caso de pretendermos ficar com os valores gerados convém ir ao *Edit* e fazer um *Paste Special - Values*, como se indica a seguir:

| | A | B | E |
|---|---|---|---|
| 1 | | Nome | |
| 2 | 1 | Escola secundária Alves Martir | 0,618064 |
| 3 | 2 | Escola secundária Amélia Rey ( | 0,052667 |
| 4 | 3 | Escola secundária com 3º ciclo | 0,119493 |
| 5 | 4 | Escola secundária D. Afonso H | 0,41474 |
| 6 | 5 | Escola secundária da Cidade U | 0,33154 |
| 7 | 6 | Escola secundária da Lourinhã | 0,065462 |
| 8 | 7 | Escola secundária da Moita | 0,946472 |
| 9 | 8 | Escola secundária da Sertã | 0,319074 |
| 10 | 9 | Escola secundária David Mourã | 0,333699 |
| 11 | 10 | Escola secundária de Albufeira | 0,739753 |
| 12 | 11 | Escola secundária de Alves Re | 0,959635 |

**Sheet1** Sheet:

| | A | B | E | F |
|---|---|---|---|---|
| 1 | | Nome | | |
| 2 | 1 | Escola secundária Alves Martir | 0,676871 | 0,6180643 |
| 3 | 2 | Escola secundária Amélia Rey ( | 0,834374 | 0,0526671 |
| 4 | 3 | Escola secundária com 3º ciclo | 0,282121 | 0,1194932 |
| 5 | 4 | Escola secundária D. Afonso H | 0,862547 | 0,4147402 |
| 6 | 5 | Escola secundária da Cidade U | 0,191305 | 0,3315404 |
| 7 | 6 | Escola secundária da Lourinhã | 0,799682 | 0,0654624 |
| 8 | 7 | Escola secundária da Moita | 0,307812 | 0,9464718 |
| 9 | 8 | Escola secundária da Sertã | 0,03148 | 0,3190742 |
| 10 | 9 | Escola secundária David Mourã | 0,883317 | 0,3336986 |
| 11 | 10 | Escola secundária de Albufeira | 0,944583 | 0,7397531 |
| 12 | 11 | Escola secundária de Alves Re | 0,871099 | 0,9596352 |

**Sheet1** Sheet2 She

Colámos os valores na coluna F e fizémos o *Save*. Repare-se que os valores que estavam inicialmente na coluna E foram alterados, dando origem a novos valores (devido ao facto da função *RAND()* ser volátil, como referimos anteriormente);
2º passo – Ordenar o ficheiro, utilizando como critério a coluna F;
3º passo – Como pretendemos uma amostra de dimensão 10, seleccionar as primeiras 10 escolas do ficheiro ordenado:

| | A | B | F |
|---|---|---|---|
| 1 | | Nome | |
| 2 | 24 | Escola secundária de Domingos | 0,0121949 |
| 3 | 34 | Escola secundária de Loulé | 0,0368761 |
| 4 | 37 | Escola secundária de Montemo | 0,0461299 |
| 5 | 2 | Escola secundária Amélia Rey | 0,0526671 |
| 6 | 6 | Escola secundária da Lourinhã | 0,0654624 |
| 7 | 28 | Escola secundária de Gil Eanes | 0,0696173 |
| 8 | 33 | Escola secundária de Júlio Dan | 0,0740176 |
| 9 | 62 | Escola secundária Jacôme Ratt | 0,085772 |
| 10 | 30 | Escola secundária de Jaime Co | 0,0995526 |
| 11 | 13 | Escola secundária de Avelar Br | 0,1023393 |

Sheet1 Sheet2

As escolas seleccionadas foram as números 24, 34, 37, 2, 6, 28, 33, 62, 30 e 13.

Nota: Embora os números anteriores sejam referidos como aleatórios, convém ter presente que os números que se obtêm são *pseudo-aleatórios*, já que é um mecanismo determinista que lhes dá origem. No entanto comportam-se como números aleatórios (passam uma bateria de testes destinados a confirmar a sua aleatoriedade) e daí a sua utilização como tal.

## *Amostra aleatória com reposição*

Vamos apresentar dois processos para seleccionar uma amostra aleatória com reposição, utilizando a função *Randbetween* ou a função *Sampling*:

**1. Função *RANDBETWEEN***

Para seleccionar aleatoriamente uma escola da lista anterior utilizamos a função ***RANDBETWEEN(m;n)***, que devolve um número pseudo-aleatório entre os números **m** e **n** especificados (m<n). Como o nosso ficheiro é constituído por 70 escolas, e pretendemos obter uma amostra de dimensão 10, escolhemos as células E2:E11 para replicar a função ***RANDBETWEEN(1;70)***:

EscolasSec.xls

| | E |
|---|---|
| 2 | =RANDBETWEEN(1;70) |
| 3 | =RANDBETWEEN(1;70) |
| 4 | =RANDBETWEEN(1;70) |
| 5 | =RANDBETWEEN(1;70) |
| 6 | =RANDBETWEEN(1;70) |
| 7 | =RANDBETWEEN(1;70) |
| 8 | =RANDBETWEEN(1;70) |
| 9 | =RANDBETWEEN(1;70) |
| 10 | =RANDBETWEEN(1;70) |
| 11 | =RANDBETWEEN(1;70) |

Esc...

| | E |
|---|---|
| 2 | 45 |
| 3 | 10 |
| 4 | 62 |
| 5 | 37 |
| 6 | 47 |
| 7 | 23 |
| 8 | 32 |
| 9 | 17 |
| 10 | 50 |
| 11 | 11 |

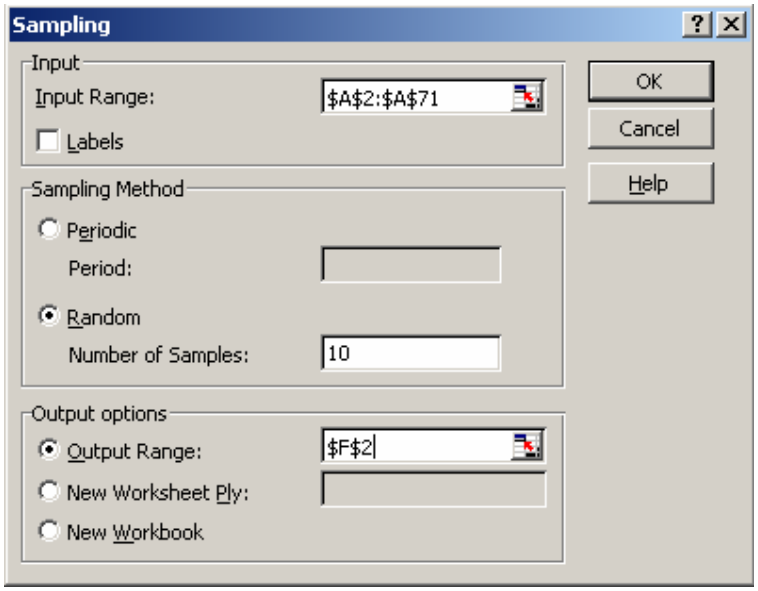
Escolas...

| | E | F |
|---|---|---|
| 2 | 15 | 45 |
| 3 | 56 | 10 |
| 4 | 63 | 62 |
| 5 | 10 | 37 |
| 6 | 23 | 47 |
| 7 | 62 | 23 |
| 8 | 28 | 32 |
| 9 | 43 | 17 |
| 10 | 54 | 50 |
| 11 | 38 | 11 |

A amostra obtida é constituída pelas escolas com os números obtidos nas células E2:E11. Uma vez que a função *RANDBETWEEN* é uma função volátil, isto é, muda quando se recalcula a folha, como pretendíamos ficar com os valores gerados fizémos um *Paste Special* dos valores obtidos para as células F2:F11 e fizémos o *Save*. Repare-se que os valores que estavam inicialmente nas células E2:E11 foram alterados, dando origem a uma nova amostra. Esta amostra alterar-se-á sempre que procedermos a alguma operação na folha de cálculo.

**2. Função *Sampling***

No Excel existe uma função, que permite seleccionar, aleatoriamente, um subconjunto de números, de um conjunto mais vasto de números. Acede-se a esta função seleccionando *Tools→ Data Analysis→Sampling* (se o comando *Data Analysis* não constar do menu, seleccione *Tools* e na opção *Add-Ins*, seleccione *Analysis ToolPack*) e procedendo como se indica a seguir:

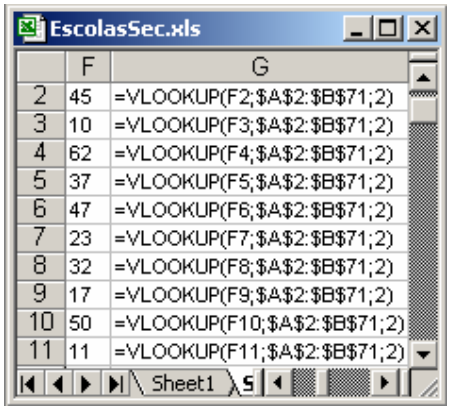



Na caixa de diálogo proceder do seguinte modo:

- Em *Input Range*: colocar o endereço da população de onde pretendemos seleccionar a amostra;
- Em *Sampling Method*: Seleccionar *Random* e em *Number of Samples*, a dimensão da amostra (nós escolhemos a dimensão 10, como no caso anterior);
- Em *Outut options*: Seleccionar a localização para onde pretendemos colocar a amostra (nós optámos por seleccionar a célula $F$2, para que a amostra ficasse colocada nas células $F$2:$F$11, como no caso da selecção através da função Randbetween).

**Função *VLOOKUP***

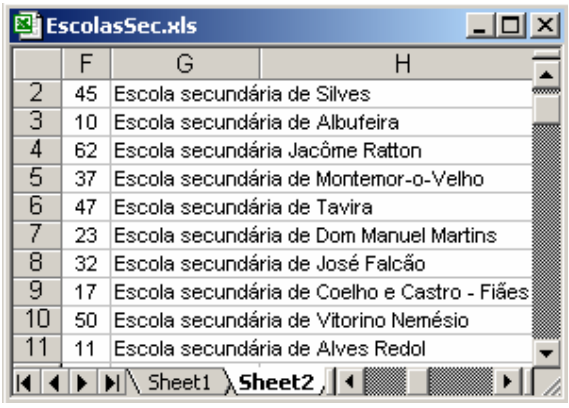
Para seleccionar o nome das escolas correspondentes aos elementos da amostra obtida – células F2:F11, vamos utilizar a função *VLOOKUP(a;b;c)*, que vai à tabela das escolas, constituída pelas 2 colunas com o número e nome das escolas- células A2:B71 (argumento b), seleccionar os nomes das escolas – 2ª coluna da tabela seleccionada (argumento c), que correspondem aos números das células F2:F11 (argumento a):



EscolasSec.xls

| | F | G |
|---|---|---|
| 2 | 45 | =VLOOKUP(F2;$A$2:$B$71;2) |
| 3 | 10 | =VLOOKUP(F3;$A$2:$B$71;2) |
| 4 | 62 | =VLOOKUP(F4;$A$2:$B$71;2) |
| 5 | 37 | =VLOOKUP(F5;$A$2:$B$71;2) |
| 6 | 47 | =VLOOKUP(F6;$A$2:$B$71;2) |
| 7 | 23 | =VLOOKUP(F7;$A$2:$B$71;2) |
| 8 | 32 | =VLOOKUP(F8;$A$2:$B$71;2) |
| 9 | 17 | =VLOOKUP(F9;$A$2:$B$71;2) |
| 10 | 50 | =VLOOKUP(F10;$A$2:$B$71;2) |
| 11 | 11 | =VLOOKUP(F11;$A$2:$B$71;2) |

Sheet1 S



EscolasSec.xls

| | F | G | H |
|---|---|---|---|
| 2 | 45 | Escola secundária de Silves | |
| 3 | 10 | Escola secundária de Albufeira | |
| 4 | 62 | Escola secundária Jacôme Ratton | |
| 5 | 37 | Escola secundária de Montemor-o-Velho | |
| 6 | 47 | Escola secundária de Tavira | |
| 7 | 23 | Escola secundária de Dom Manuel Martins | |
| 8 | 32 | Escola secundária de José Falcão | |
| 9 | 17 | Escola secundária de Coelho e Castro - Fiães | |
| 10 | 50 | Escola secundária de Vitorino Nemésio | |
| 11 | 11 | Escola secundária de Alves Redol | |

Sheet1 Sheet2

Se pretendêssemos seleccionar além do nome da escola, o seu distrito, então teríamos de considerar como argumento b da função *VLOOKUP* a tabela A2:C71 e como argumento c, o valor 3, uma vez que nos estamos a referir à 3ª coluna da tabela considerada.

## Amostra aleatória sistemática

Na prática o processo de seleccionar uma *amostra aleatória simples* de uma população com grande dimensão, não é tão simples como o descrito anteriormente. Se a dimensão da população for grande o processo torna-se muito trabalhoso. Então uma alternativa é considerar uma amostra

aleatória sistemática. Por exemplo, se pretendermos seleccionar uma amostra de 150 alunos de uma Universidade com 6000 alunos, considera-se um ficheiro com o nome dos 6000 alunos ordenados por ordem alfabética. Considera-se o quociente 6000/150=40 e dos primeiros 40 elementos da lista, selecciona-se um aleatoriamente. A partir deste elemento seleccionamos sistematicamente todos os elementos distanciados de 40 unidades. Assim, se o elemento seleccionado aleatoriamente de entre os primeiros 40, foi o 27, os outros elementos a serem seleccionados são 67, 107, 147, etc. Obviamente que o quociente entre a dimensão da população e a da amostra não é necessariamente inteiro, como anteriormente, mas não há problema pois considera-se a parte inteira desse quociente.

**Amostra aleatória sistemática** – Dada uma população de dimensão N, ordenada por algum critério, se se pretende uma amostra de dimensão n, escolhe-se aleatoriamente um elemento de entre os k primeiros, onde k é a parte inteira do quociente N/n. A partir desse elemento escolhido, escolhem-se todos os k-ésimos elementos da população para pertencerem à amostra.

A amostra aleatória sistemática não é uma amostra aleatória simples, já que os elementos da população não têm a mesma probabilidade de pertencerem à amostra. Basta pensar que dois elementos adjacentes não podem ser seleccionados.

**Amostragem estratificada**

Pode acontecer que a população possa ser subdividida em várias subpopulações, mais ou menos homogéneas relativamente à característica a estudar. Por exemplo, se se pretende estudar o salário médio auferido pelas famílias lisboetas, é possível dividir a região de Lisboa segundo zonas mais ou menos homogéneas, *estratos*, quanto à característica em estudo – salário médio, e posteriormente extrair de cada um destes estratos uma percentagem de elementos que irão constituir a amostra, sendo esta percentagem, de um modo geral, proporcional à dimensão dos estratos.

**Amostra estratificada** – Divide-se a população em várias subpopulações – estratos, e de cada uma destes estratos extrai-se aleatoriamente uma amostra. O conjunto de todas estas amostras constitui a amostra pretendida.

As técnicas anteriores podem não ser ainda satisfatórias para resolver determinadas situações.

**Amostragem por "clusters" ou grupos**

Por exemplo, suponha que se pretende estudar o nível de satisfação dos trabalhadores têxteis, das empresas do Norte do País. Não dispondo de uma lista com todos os trabalhadores, considera-se uma lista de todas as empresas têxteis – "clusters", admitindo-se que o conjunto de trabalhadores de cada empresa caracteriza convenientemente a população que se pretende

estudar. A partir dessa lista seleccionam-se aleatoriamente algumas empresas e considera-se a amostra constituída por todos os trabalhadores das empresas seleccionadas.

**Amostra por clusters** – A população é dividida em *clusters*, onde cada *cluster* é representativo da população. Selecciona-se aleatoriamente um conjunto de *clusters* e a amostra é constituída por todos os elementos dos *clusters* seleccionados.

Um outro tipo de amostragem também muito utilizado, é semelhante ao anterior no que diz respeito à primeira fase da selecção dos clusters como se exemplifica a seguir.

**Amostragem multi-etapas**

Suponha que em vésperas de eleições presidenciais se pretende obter uma estimativa das percentagens de cada candidato e não se dispõe de uma lista com todos os eleitores, que são milhões. Mesmo se se dispusesse dessa lista não seria tarefa simples seleccionar aleatoriamente alguns elementos. Então considera-se o País dividido em algumas regiões geográficas, por exemplo Norte, Centro e Sul. Dentro de cada região procede-se ao agrupamento dos centros populacionais com dimensão semelhante. Depois de cada agrupamento são seleccionadas aleatoriamente algumas cidades. As cidades por sua vez ainda estão divididas em Juntas de Freguesia. Algumas destas Juntas de Freguesia são seleccionadas aleatoriamente das cidades seleccionadas no passo anterior. Finalmente de cada freguesia seleccionada, ainda se escolhem aleatoriamente alguns lares para inqurir, por exemplo, o adulto mais jovem.

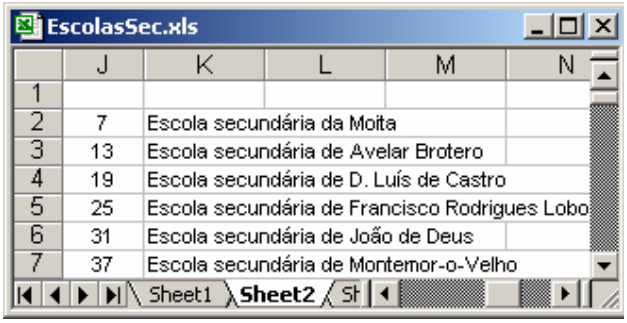
**Amostragem multi-etapas** – Considera-se a população dividida em vários grupos, seleccionando-se aleatoriamente alguns dos grupos. Por sua vez estes grupos ainda estão divididos em grupos, dos quais se seleccionam alguns aleatoriamente. Este processo pode repetir-se até ser possível constituir grupos.

## Utilização do Excel na selecção de uma amostra aleatória sistemática

Vamos considerar ainda o ficheiro EscolasSec.xls para exemplificar a utilização do Excel na selecção de uma amostra aleatória sistemática.

**Exemplo** – Considerando ainda a população finita constituída pelas escolas do ficheiro EscolasSec.xls, seleccione uma amostra aleatória sistemática de dimensão 6.

Temos uma população de dimensão 70, da qual se pretende seleccionar uma amostra de dimensão 6. Vamos utilizar a seguinte metodologia:
Passo 1 – Dividir 70 por 6 e reter a parte inteira que é 11;
Passo 2 – Dos primeiros 11 elementos da lista ordenada das escolas, seleccionar um elemento ao acaso, utilizando a função RANDBETWEEN(1;11), que inserimos na célula I2; copiámos o valor obtido, através de um Paste Special, para a célula J2;
Passo 3 – Coloque o cursor na célula J3 e escreva =J2+6; replique a fórmula da célula J3 pelas células J4:J7;
Passo 4 – Escreva na célula K2 a função VLOOKUP(J2;$A$2:$B$71;2) e replique-a através das células K3:K7:



EscolasSec.xls

| | J | K | L | M | N |
|---|---|---|---|---|---|
| 1 | | | | | |
| 2 | 7 | Escola secundária da Moita | | | |
| 3 | 13 | Escola secundária de Avelar Brotero | | | |
| 4 | 19 | Escola secundária de D. Luís de Castro | | | |
| 5 | 25 | Escola secundária de Francisco Rodrigues Lobo | | | |
| 6 | 31 | Escola secundária de João de Deus | | | |
| 7 | 37 | Escola secundária de Montemor-o-Velho | | | |

Sheet1 Sheet2 Sh

**Qual a dimensão que se deve considerar para a amostra?**

Outro problema que se levanta com a recolha da amostra é o de saber qual a **dimensão** desejada para a amostra a recolher. Este é um problema para o qual, nesta fase, não é possível avançar nenhuma teoria, mas sobre o qual se podem tecer algumas considerações gerais. Pode-se começar por dizer que, para se obter uma amostra que permita calcular estimativas suficientemente precisas dos parâmetros a estudar, a sua dimensão depende muito da variabilidade da população subjacente. Por exemplo, se relativamente à população constituída pelos alunos do 10º ano de uma escola secundária, estivermos interessados em estudar a sua idade média, a dimensão da amostra a recolher não necessita de ser muito grande já que a variável idade apresenta valores muito semelhantes, numa classe etária muito restrita. No entanto se a característica a estudar for o tempo médio que os alunos levam a chegar de casa à escola, já a amostra terá de ter uma dimensão maior, uma vez que a variabilidade da população é muito maior. Cada aluno pode apresentar um valor diferente para esse tempo. Num caso extremo, se numa população a variável a estudar tiver o mesmo valor para todos os elementos, então bastaria recolher uma amostra de dimensão 1 para se ter informação completa sobre a população; se, no entanto, a variável assumir valores diferentes para todos os elementos, para se ter o mesmo tipo de informação seria necessário investigar todos os elementos.

Chama-se a atenção para a existência de técnicas que permitem obter valores mínimos para as dimensões das amostras a recolher e que garantem estimativas com uma determinada **precisão** exigida à partida. Uma vez garantida essa precisão, a opção por escolher uma amostra de maior dimensão, é uma questão a ponderar entre os custos envolvidos e o ganho com o acréscimo de precisão. Vem a propósito a seguinte frase (Mendenhall, 1974,pag. 226):

*"Se a dimensão da amostra é demasiado grande, desperdiça-se tempo e talento; se a dimensão da amostra é demasiado pequena, desperdiça-se tempo e talento".*

Convém ainda observar que a dimensão da amostra a recolher não é directamente proporcional à dimensão da população a estudar, isto é, se por exemplo para uma população de dimensão 1000 uma amostra de dimensão 100 for suficiente para o estudo de determinada característica, não se

exige necessariamente uma amostra de dimensão 200 para estudar a mesma característica de uma população análoga, mas de dimensão 2000, quando se pretende obter a mesma precisão. Como explicava George Gallup, um dos pais da consulta da opinião pública (Tannenbaum, 1998): "Whether you poll the United States or New York State or Baton Rouge (Louisiana) … you need … the same number of interviews or samples. It's no mystery really – if a cook has two pots of soup on the stove, one far larger than the other, and thoroughly stirs them both, he doesn't have to take more spoonfuls from one than the other to sample the taste accurately".

Há no entanto (Vicente, 1996) uma situação de excepção relativamente ao que foi dito, isto é, existe uma situação em que a dimensão da população interfere directamente na dimensão da amostra: quando a amostra é recolhida sem reposição não há independência entre os elementos, facto que terá impacto na fórmula do cálculo da variância do estimador a utilizar.

Finalmente chama-se a atenção para o facto de que se o processo de amostragem originar uma amostra enviesada, aumentar a dimensão não resolve nada, antes pelo contrário!

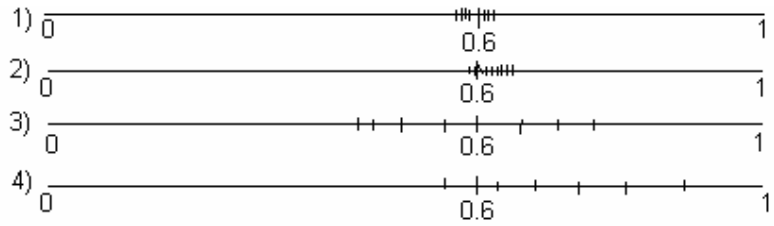
Além do enviesamento, um outro problema que não pode deixar de ser referido é o da precisão.

**Precisão** - Ao utilizar o valor de uma estatística para estimar um parâmetro, vimos que cada amostra fornece um valor para a estatística que se utiliza como estimativa desse parâmetro. Estas estimativas não são iguais devido à variabilidade presente na amostra. Se, no entanto, os diferentes valores obtidos para a estatística forem próximos, podemos ter confiança de que o valor calculado a partir da amostra recolhida (na prática recolhe-se uma única amostra) está próximo do valor do parâmetro (desconhecido).

A **falta de precisão** juntamente com o problema do **enviesamento** são dois tipos de erro com que nos defrontamos num processo de amostragem. Não se devem, contudo, confundir. Enquanto o enviesamento se manifesta por um desvio nos valores da estatística, relativamente ao valor do parâmetro a estimar, sempre no mesmo sentido, a falta de precisão manifesta-se por uma grande variabilidade nos valores da estatística, uns relativamente aos outros. Por outro lado, enquanto o enviesamento se reduz com o recurso a amostras aleatórias, a precisão aumenta-se aumentando a dimensão da amostra.

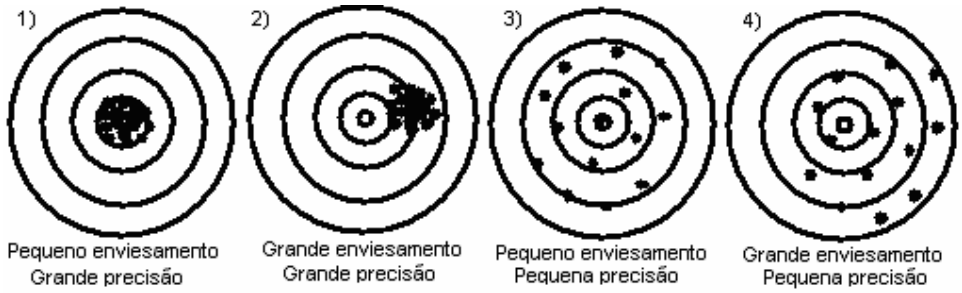
**Exemplo 2 -** Suponhamos que ao pretender estudar a percentagem de eleitores que votariam favoravelmente num candidato à Câmara de determinada cidade, se recolhia uma amostra de 300 eleitores, dos quais 175 responderam que sim. Então uma estimativa para a proporção pretendida seria 0.58. Se considerássemos outra amostra de 300 eleitores, suponhamos que o valor obtido para o número de sim's tinha sido 181. Então o valor obtido para a estatística seria 0.60. A repetição deste processo 15 vezes permitiria obter 15 valores para a estatística, que seriam outras tantas estimativas do parâmetro a estimar - percentagem de eleitores da cidade, potenciais

apoiantes do tal candidato. Representando num eixo os valores obtidos, poderíamos deparar-nos com várias situações:




Se admitirmos que o valor do parâmetro era 0.60, então a situação 1) reflecte um *pequeno* ou ausência de *enviesamento*, pois os valores para a estatística (proporções obtidas a partir das amostras) situam-se para um e outro lado do valor do parâmetro, e a existência de uma pequena variabilidade entre os resultados obtidos para as várias amostras, que se traduz em *grande precisão*. No caso 2) embora se mantenha a *precisão*, existe um *grande enviesamento*, pois os valores da estatística situam-se sistematicamente para a direita do valor do parâmetro. No caso 3) voltamos a ter uma situação de *pequeno enviesamento*, mas de *pequena precisão* devido à grande variabilidade apresentada pelos valores da estatística. Finalmente no caso 4) a *falta de precisão* da situação 3) é acompanhada de um *grande enviesamento*.

Fazendo analogia com o que se passa com um atirador que aponta várias setas a um alvo, em que procurava atingir o centro do alvo, teríamos




**Outros tipos de erros num processo de aquisição de dados**

Além dos erros apontados anteriormente existem ainda outras fontes de erros que não estão relacionadas com o método da recolha da amostra nem com a dimensão da amostra, que são os chamados *erros de não amostragem*. Se, por exemplo, seleccionarmos uma amostra aleatória simples a partir de uma listagem de elementos que não contenha todos os elementos da população, poderemos obter uma amostra enviesada. Efectivamente, muitas vezes a recolha da amostra faz-se de uma população que não é a população que se pretende estudar – *população objectivo*, mas sim de outra população que se pensa representar a primeira – *população inquirida*. Por exemplo, se se pretende estudar uma determinada característica dos residentes em Lisboa, é comum recolher uma amostra seleccionando aleatoriamente alguns números de telefones da lista telefónica de Lisboa, para representar a população lisboeta. Este processo introduz algum

enviesamento, pois existem zonas de Lisboa onde a percentagem de pessoas com telefone é pequena. Além disso, pode acontecer com alguma frequência telefonarem para casa das pessoas quando elas estão ausentes, no trabalho, pelo que a amostra subestimará a percentagem dos lisboetas que trabalham fora de casa. O exemplo que acabámos de descrever refere-se a um **erro de selecção**.

Na recolha da informação também se pode ainda verificar que a informação dada **não seja verdadeira**. Ao responder a um inquérito o inquirido pode sentir-se condicionado pelo inquiridor, face a determinadas perguntas. Isso poderá levá-lo a mentir. Por exemplo ao perguntarem a um indivíduo se ele é racista, ele pode dizer que não, quando na verdade o é.

Finalmente, pode-se ter feito um planeamento adequado da amostra a recolher, mas ao recolher a informação de entre os elementos da amostra a pessoa encarregada dessa recolha pode ver-se defrontada com a **não resposta**. Este problema acontece com frequência quando a amostra é constituída por pessoas, das quais algumas das seleccionadas não são encontradas para darem a informação sobre a variável em estudo, ou então se recusam a responder.

Outro problema que pode surgir é devido a **erros de processamento** que não têm nada a ver com o processo de recolha da amostra, mas que podem influenciar o resultado da estatística, já que esta é calculada com base na informação recolhida. Estes erros surgem com alguma frequência, sendo muitas vezes detectados por serem *outliers*. Efectivamente, se ao digitar um conjunto de valores correspondentes a pesos de pessoas adultas aparecer 566 quilogramas, ao fazer uma representação gráfica aparecerá este valor como *outlier* e imediatamente se concluirá que se trata de um problema de processamento: eventualmente ao carregar a tecla do 6 o tempo de apoio foi um pouco maior e apareceram dois 6.

## Pode-se aumentar a precisão estratificando?

A selecção de uma amostra estratificada, utilizando o Excel, não apresenta qualquer dificuldade, pois não é mais do que a selecção de amostras aleatórias simples das subpopulações que constituem os estratos. Vamos, no entanto apresentar um exemplo sugerido por Hodgson (1998), não pela sua importância em termos da sua resolução com o Excel, mas pela sua relevância na exemplificação da técnica de estratificação.

**Exemplo** – Consideremos uma população constituída por 40 cartões – 20 vermelhos e 20 pretos, numerados, de acordo com a seguinte tabela:

| Nº | 6 | 7 | 8 | 9 | 10 | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|---|---|---|---|
| Freq. | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| Cor | Ver | Ver | Ver | Ver | Ver | Preto | Preto | Preto | Preto | Preto |

A média dos números inscritos nesta população de 40 cartões é de 18 – valor médio da característica populacional em estudo. Admitindo que o valor médio anterior era desconhecido e que se pretendia obter uma estimativa, foram-se seleccionar algumas amostras de dimensão 4 e calcular as médias das amostras obtidas. Para isso construimos um ficheiro, em Excel, com o valor dos cartões e a cor:

Estratificar

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Cartão nº | Valor | Cor | Cartão nº | Valor | Cor | Cartão nº | Valor | Cor |
| 2 | 1 | 6 | Ver | 15 | 9 | Ver | 29 | 28 | Preto |
| 3 | 2 | 6 | Ver | 16 | 9 | Ver | 30 | 28 | Preto |
| 4 | 3 | 6 | Ver | 17 | 10 | Ver | 31 | 28 | Preto |
| 5 | 4 | 6 | Ver | 18 | 10 | Ver | 32 | 28 | Preto |
| 6 | 5 | 7 | Ver | 19 | 10 | Ver | 33 | 29 | Preto |
| 7 | 6 | 7 | Ver | 20 | 10 | Ver | 34 | 29 | Preto |
| 8 | 7 | 7 | Ver | 21 | 26 | Preto | 35 | 29 | Preto |
| 9 | 8 | 7 | Ver | 22 | 26 | Preto | 36 | 29 | Preto |
| 10 | 9 | 8 | Ver | 23 | 26 | Preto | 37 | 30 | Preto |
| 11 | 10 | 8 | Ver | 24 | 26 | Preto | 38 | 30 | Preto |
| 12 | 11 | 8 | Ver | 25 | 27 | Preto | 39 | 30 | Preto |
| 13 | 12 | 8 | Ver | 26 | 27 | Preto | 40 | 30 | Preto |
| 14 | 13 | 9 | Ver | 27 | 27 | Preto | | | |
| 15 | 14 | 9 | Ver | 28 | 27 | Preto | | | |

Sheet1 / Sheet2 / Sheet3

Da população anterior seleccionámos 36 amostras de dimensão 4, calculando ainda as médias dos valores escritos nos cartões seleccionados. Posteriormente considerámos a população constituída por dois estratos – a subpopulação dos cartões vermelhos e a dos cartões pretos e de cada uma destas subpopulações extraímos 2 cartões, calculando ainda a média dos 4 cartões seleccionados. Os resultados obtidos nos dois processos de amostragem encontram-se na figura seguinte, respectivamente na tabela do lado esquerdo e do lado direito, para a amostragem sem estratificação e com estratificação:

| X | Y | Z | AA | AB | AC |
|---|---|---|---|---|---|
| Amostra nº | | | | | média |
| 1 | 10 | 26 | 8 | 30 | 18,50 |
| 2 | 6 | 26 | 27 | 27 | 21,50 |
| 3 | 26 | 29 | 9 | 29 | 23,25 |
| 4 | 29 | 26 | 30 | 28 | 28,25 |
| 5 | 28 | 7 | 6 | 30 | 17,75 |
| 6 | 7 | 7 | 9 | 27 | 12,50 |
| 7 | 26 | 26 | 26 | 29 | 26,75 |
| 8 | 27 | 9 | 9 | 30 | 18,75 |
| 9 | 29 | 10 | 8 | 7 | 13,50 |
| 10 | 6 | 27 | 27 | 6 | 16,50 |
| 11 | 9 | 10 | 28 | 27 | 18,50 |
| 12 | 9 | 7 | 6 | 10 | 8,00 |
| 13 | 30 | 7 | 6 | 29 | 18,00 |
| 14 | 6 | 7 | 10 | 27 | 12,50 |
| 15 | 6 | 6 | 9 | 7 | 7,00 |
| 16 | 10 | 9 | 30 | 6 | 13,75 |
| 17 | 7 | 27 | 30 | 27 | 22,75 |
| 18 | 30 | 7 | 27 | 26 | 22,50 |
| 19 | 27 | 10 | 6 | 27 | 17,50 |
| 20 | 27 | 10 | 7 | 9 | 13,25 |
| 21 | 29 | 30 | 6 | 6 | 17,75 |
| 22 | 9 | 8 | 28 | 27 | 18,00 |
| 23 | 7 | 9 | 8 | 26 | 12,50 |
| 24 | 8 | 9 | 7 | 10 | 8,50 |
| 25 | 27 | 8 | 7 | 9 | 12,75 |
| 26 | 9 | 9 | 8 | 29 | 13,75 |
| 27 | 28 | 9 | 30 | 6 | 18,25 |
| 28 | 9 | 7 | 28 | 26 | 17,50 |
| 29 | 7 | 9 | 30 | 28 | 18,50 |
| 30 | 9 | 7 | 29 | 8 | 13,25 |
| 31 | 10 | 6 | 7 | 6 | 7,25 |
| 32 | 26 | 27 | 7 | 29 | 22,25 |
| 33 | 29 | 26 | 8 | 27 | 22,50 |
| 34 | 30 | 8 | 8 | 10 | 14,00 |
| 35 | 26 | 26 | 27 | 30 | 27,25 |
| 36 | 8 | 8 | 27 | 7 | 12,50 |

Sheet1 / Shee

| L | M | N | O | P | Q |
|---|---|---|---|---|---|
| Amostra nº | | | | | média |
| 1 | 30 | 27 | 9 | 10 | 19,00 |
| 2 | 26 | 30 | 7 | 8 | 17,75 |
| 3 | 30 | 30 | 7 | 10 | 19,25 |
| 4 | 30 | 26 | 9 | 6 | 17,75 |
| 5 | 26 | 28 | 9 | 7 | 17,50 |
| 6 | 26 | 30 | 10 | 7 | 18,25 |
| 7 | 26 | 27 | 10 | 6 | 17,25 |
| 8 | 28 | 26 | 6 | 9 | 17,25 |
| 9 | 28 | 26 | 9 | 10 | 18,25 |
| 10 | 26 | 29 | 9 | 8 | 18,00 |
| 11 | 27 | 30 | 7 | 6 | 17,50 |
| 12 | 27 | 30 | 10 | 8 | 18,75 |
| 13 | 30 | 28 | 9 | 6 | 18,25 |
| 14 | 30 | 26 | 10 | 8 | 18,50 |
| 15 | 28 | 26 | 7 | 6 | 16,75 |
| 16 | 26 | 30 | 10 | 7 | 18,25 |
| 17 | 26 | 28 | 8 | 7 | 17,25 |
| 18 | 27 | 30 | 10 | 6 | 18,25 |
| 19 | 30 | 30 | 9 | 8 | 19,25 |
| 20 | 30 | 30 | 6 | 8 | 18,50 |
| 21 | 6 | 10 | 26 | 26 | 17,00 |
| 22 | 6 | 8 | 27 | 28 | 17,25 |
| 23 | 9 | 8 | 26 | 30 | 18,25 |
| 24 | 7 | 7 | 26 | 26 | 16,50 |
| 25 | 6 | 6 | 27 | 28 | 16,75 |
| 26 | 8 | 7 | 30 | 26 | 17,75 |
| 27 | 10 | 7 | 28 | 29 | 18,50 |
| 28 | 9 | 8 | 29 | 26 | 18,00 |
| 29 | 6 | 8 | 29 | 30 | 18,25 |
| 30 | 8 | 7 | 28 | 28 | 17,75 |
| 31 | 10 | 6 | 29 | 26 | 17,75 |
| 32 | 6 | 6 | 29 | 27 | 17,00 |
| 33 | 10 | 10 | 29 | 29 | 19,50 |
| 34 | 10 | 7 | 27 | 27 | 17,75 |
| 35 | 8 | 6 | 29 | 30 | 18,25 |
| 36 | 10 | 9 | 29 | 26 | 18,50 |

Sheet1 / Sh

Representando num eixo, os valores obtidos para as diferentes estimativas, temos:

Sem estratificação →


Com estratificação →


De acordo com as considerações feitas anteriormente sobre a precisão, são óbvias as vantagens da estratificação (Não esqueçamos que o valor do parâmetro a estimar era 18).

### 1.2.2 - Experimentações

A recolha de dados através de sondagens não é suficiente quando se pretende estudar o efeito ou resposta de um conjunto de indivíduos a determinado estímulo ou tratamento (termo utilizado em estatística). Somos assim conduzidos a um outro processo de aquisição de dados a que chamamos experimentação. Enquanto que o objectivo de uma sondagem é o de recolher informação acerca de uma população seleccionando e observando uma amostra da população tal qual ela se apresenta, pelo contrário, uma experimentação impõe um **tratamento** às unidades experimentais com o fim de observar a **resposta**. O princípio base de uma experimentação é o **método da comparação**, em que se comparam os resultados obtidos na variável resposta de um **grupo de tratamento** com um **grupo de controlo**.

**Exemplo 3** (Moore, 1997) - Será que a aspirina reduz o perigo de um ataque cardíaco? O estudo conhecido por Physicians' Health Study, foi uma experimentação médica levada a cabo com o objectivo de responder a esta questão específica. Metade de um grupo de 22000 médicos (homens) foram escolhidos aleatoriamente para tomar uma aspirina todos os dias. A outra metade dos médicos tomou um **placebo**, que tinha o mesmo aspecto e sabor da aspirina. Depois de vários anos 239 médicos do grupo que tomou placebo, contra 139 do grupo que tomou aspirina, tiveram ataques cardíacos. Esta diferença é suficientemente grande para evidenciar o efeito da aspirina na prevenção dos ataques cardíacos.

**Unidades experimentais, tratamento, variável resposta, variáveis explanatórias**

**Unidades experimentais** são os objectos sobre os quais incide a experimentação e a quem é aplicado uma condição experimental específica, a que chamamos **tratamento**. **Variável resposta** é a variável cujo comportamento pretendemos estudar. **As variáveis explanatórias** são as variáveis que explicam ou causam mudanças na variável resposta.

No estudo considerado anteriormente temos:

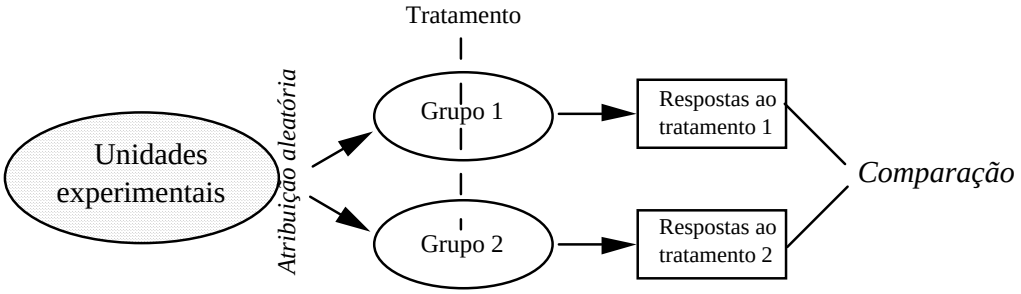
- Unidades experimentais - 22000 médicos
- Tratamentos - aspirina ou placebo
- Variável explanatória - se o indivíduo tomou aspirina ou placebo
- Variável resposta - se o indivíduo teve ou não ataque cardíaco.

Sem a comparação de tratamentos os resultados de experimentações em medicina e em ciências do comportamento, duas áreas onde estes métodos são largamente utilizados, poderiam ser muito influenciados pela selecção dos indivíduos, o efeito do placebo, etc. O resultado poderia vir **enviesado**. Um estudo não controlado de uma nova terapia médica é quase sempre enviesado no sentido de dar ao tratamento um maior sucesso do que ele tem na realidade (efeito placebo).

**Exemplo 4** (Moore, 1997) - Um tratamento utilizado durante vários anos para tratar úlceras do estômago consistia em pôr o doente a aspirar, durante uma hora, uma solução refrigerada que era bombeada para dentro de um balão. Segundo o Journal of the American Medical Association, uma experimentação levada a efeito com este tratamento permitiu concluir que o arrefecimento gástrico reduzia a secreção de ácido, diminuindo a propensão para as úlceras. No entanto, veio-se a verificar mais tarde com um planeamento adequado, que a resposta dos doentes ao tratamento foi influenciada pelo efeito placebo – efeito *confounding*. O que acontece é que há doentes que respondem favoravelmente a qualquer tratamento, mesmo que seja um placebo, possivelmente pela confiança que depositam no médico e pelas expectativas de cura que depositam no tratamento. Num planeamento adequado feito anos mais tarde, um grupo de doentes com úlcera foi dividido em dois grupos, tratando-se um com a solução refrigerada e o outro grupo com um placebo, constituído por uma solução à temperatura ambiente. Os resultados desta experimentação permitiram concluir que dos 82 doentes sujeitos à solução refrigerada - grupo de tratamento, 34% apresentaram melhoras, enquanto que dos 78 doentes que receberam o placebo - grupo de controlo, 38% apresentaram melhoras.

Num planeamento experimental, uma vez identificadas as variáveis e estabelecido o protocolo dos tratamentos, segue-se uma segunda fase que consiste na atribuição de cada unidade experimental a um tratamento.

Esta segunda fase deve ser regida pelo **princípio da aleatoriedade**. Este princípio tem como objectivo fazer com que os grupos que vão ser comparados, tenham à partida constituição semelhante, de forma que as diferenças observadas na variável resposta possam ser atribuídas aos efeitos dos tratamentos. Assim, a atribuição de cada indivíduo ao grupo de tratamento ou de controlo é feita de forma aleatória. Combinando a comparação com a aleatoriedade, podemos esquematizar da seguinte forma o tipo de planeamento mais simples:



Ao comparar os resultados temos de ter presente que haverá sempre alguma diferença que se tem de atribuir ao facto de os grupos não serem perfeitamente idênticos e algumas diferenças que se atribuem ao acaso. O que se pretende é averiguar se as diferenças encontradas não serão "demasiado grandes" para que se possam atribuir a essas causas, ou seja, verificar se não tendo em linha de conta a diferença entre os tratamentos, a probabilidade de obter as diferenças observadas não seria extremamente pequena. Se efectivamente esta probabilidade for inferior a um determinado valor (de que falaremos mais tarde) dizemos que a diferença **é estatisticamente significativa**, sendo de admitir que foi provocada pelos diferentes tratamentos.

Convém ainda observar que numa experimentação os indivíduos seleccionados para cada grupo não devem saber qual o tipo de tratamento a que estão a ser sujeitos, nem o investigador que está a conduzir a experimentação e a medir os resultados deve saber qual o tipo de tratamento que cada indivíduo seguiu. Temos o que se chama uma experimentação *duplamente cega*. Esta precaução é uma forma de evitar o enviesamento, quer nas respostas, quer nas medições (um médico ao observar o efeito de um tratamento que provoque, por exemplo, uma mancha vermelha na pele, pode estar condicionado na interpretação da gravidade dessa mancha se souber qual o tratamento a que o doente foi sujeito).

Em muitas situações os investigadores têm de se cingir aos estudos observacionais, já que não é possível conduzir uma experimentação controlada. Por exemplo, para estudar o efeito do tabaco no cancro do pulmão, o investigador limita-se a observar grupos de indivíduos que fumam ou não, não podendo ser ele próprio a seleccionar um conjunto de indivíduos e depois pô-los aleatoriamente a fumar tabaco ou um placebo. O exemplo que acabámos de abordar sugere a existência de algumas questões éticas associadas às experimentações, que impedem o investigador de prosseguir a recolha de informação da forma que inicialmente teria planeado.

Nesta secção procurámos abordar alguns problemas relacionados com a fase de recolha da amostra e motivar os leitores para a sua importância. O estudo conveniente do planeamento das experiências, assim como da definição da amostra adequada para o estudo em vista, saem fora do âmbito da disciplina a que estas folhas se destinam, pois contêm por si só matéria suficiente para ser objecto de várias disciplinas num curso de Estatística, nomeadamente as disciplinas de Planeamento de Experiências e Amostragem.

### 1.3 - Exploração dos dados - Estatística Descritiva

Uma vez os dados recolhidos, sob a forma de uma amostra, faz-se a redução desses dados, utilizando tabelas, diferentes tipos de gráficos e medidas a que chamamos estatísticas, sendo um dos principais objectivos desta fase, a identificação da estrutura subjacente aos dados, deixando de lado a aleatoriedade presente.

Nesta fase de análise dos dados, além da descrição dos mesmos, em que se procura pôr em evidência as características principais e as propriedades, pretende-se **formular um modelo**. De um modo geral a situação em estudo é bastante complexa, ou nem todos os aspectos da situação têm interesse para o estudo em vista, de modo que se formula um modelo, que nos dá uma visão simplificada da situação real.

O estatístico George Box afirmava que: *Todos os modelos são maus; alguns modelos são úteis*.

O objectivo na escolha de um modelo é o de encontrar um que consiga apreender os aspectos importantes do fenómeno a estudar, mas que seja suficientemente simples para se conseguir trabalhar!

Por exemplo:

1) "Em média, cada cigarro que se fuma por dia, reduz o tempo de vida de uma certa quantidade de tempo, a qual será estimada com base num grande conjunto de dados"
Este modelo ignora muitos factores, tais como a idade, sexo, tipo de vida, etc. No entanto, pode dar uma boa ideia do efeito de fumar na saúde.

2) " O tempo para amanhã depende do tempo de hoje, se tivermos em consideração a pressão atmosférica, a humidade, formações de nevoeiro, e velocidades de vento"
Pode não ser um mau modelo para a previsão do tempo. No entanto, não passa de um modelo e todos sabemos que as previsões nem sempre saem certas!

Esta fase inicial da análise dos dados, a que damos o nome de **Estatística Descritiva** - por vezes é chamada de **Análise Preliminar de Dados**, embora alguns autores (Chatfield, 1985), contestem esta terminologia, pois afirmam que por vezes a análise inicial de dados é suficiente, não havendo necessidade de proceder a qualquer tipo de inferência e daí ser abusivo o termo *preliminar*.

### 1.4 - Inferência Estatística

Seguidamente, o objectivo de um estudo estatístico é, de uma maneira geral, o de **estimar** uma quantidade ou **testar uma hipótese**, utilizando-se técnicas estatísticas convenientes, as quais realçam toda a potencialidade da Estatística, na medida em que vão permitir tirar conclusões acerca de uma População, baseando-se numa pequena amostra, dando-nos ainda *uma medida do erro cometido*. A esta fase chamamos **Inferência Estatística**.

Quando dizemos que de um modo geral é esse o objectivo, significa que por vezes não chegamos a esta fase, de fazer inferências (ver observação do último parágrafo da secção anterior). Podem, por exemplo, os resultados da análise dos dados, ter permitido tirar algumas conclusões, tais como a de os dados serem demasiado pobres para fazer inferência. Por outro lado, os resultados da análise dos dados, podem ser suficientes, para os fins que se têm em vista.

**Exemplo 5** - Numa experiência para comparar resultados de métodos de ensino para ensinar a aritmética, 45 estudantes foram seleccionados aleatoriamente e divididos em 5 grupos de tamanho igual. A dois dos grupos, A e B, aplicou-se o método tradicional (grupos de controlo),

enquanto que aos outros grupos, C, D e E, se aplicaram 3 métodos novos. No fim da experiência todos os estudantes realizaram um teste, cujos resultados se apresentam a seguir:

| | | | | | | | | | | Média | Ampl. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Grupo A | 17 | 14 | 24 | 20 | 24 | 23 | 16 | 15 | 24 | 19.7 | 10 |
| Grupo B | 21 | 23 | 13 | 19 | 13 | 19 | 20 | 21 | 16 | 18.3 | 10 |
| Grupo C | 28 | 30 | 29 | 24 | 27 | 30 | 28 | 28 | 23 | 27.4 | 7 |
| Grupo D | 19 | 28 | 26 | 26 | 19 | 24 | 24 | 23 | 22 | 23.4 | 9 |
| Grupo E | 21 | 14 | 13 | 19 | 15 | 15 | 10 | 18 | 20 | 16.1 | 11 |

Como todos os grupos têm igual dimensão, calculámos a amplitude como medida de dispersão. Ao compararmos as médias são evidentes as divergências entre os grupos. Estas divergências tornam-se mais evidentes ao construirmos as representações em *Box-plot* (a ver posteriormente), que mostram que efectivamente os 5 métodos não são equivalentes.

**1.5 – Estatística Descritiva e Inferência Estatística**

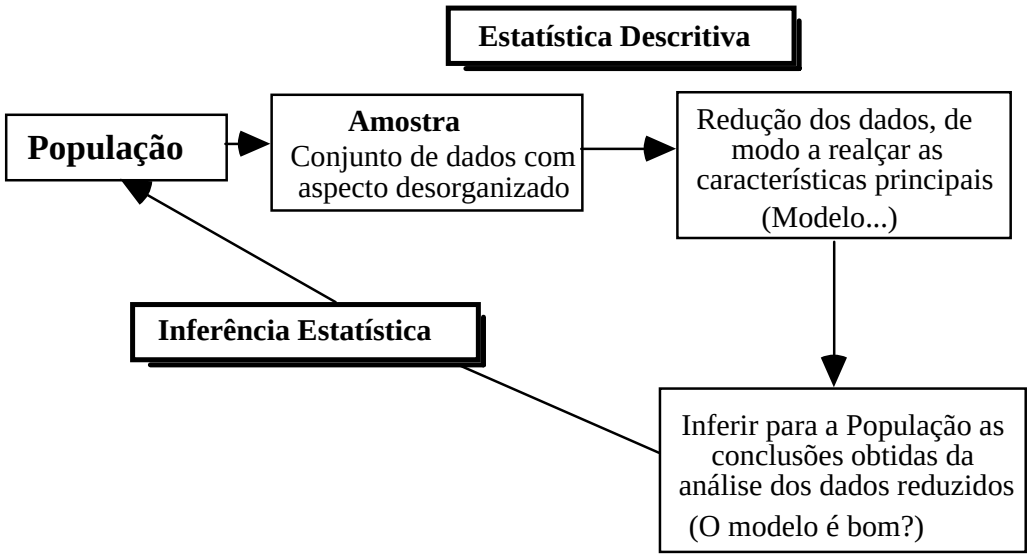
Resumindo, podemos dizer que uma análise estatística envolve, geralmente, duas fases fundamentais, com objectivos distintos:

**Estatística Descritiva** - Procura-se descrever a amostra, pondo em evidência as características principais e as propriedades. Procura-se ainda formular um modelo que traduza, de uma forma simplificada, a situação em estudo.

**Inferência Estatística** - Conhecidas certas propriedades (obtidas a partir de uma análise descritiva da amostra), expressas por meio de proposições, imaginam-se proposições mais gerais, que exprimam a existência de leis (na População). No entanto, ao contrário das proposições deduzidas, não podemos dizer que são falsas ou verdadeiras, já que foram verificadas sobre um conjunto restrito de indivíduos, e portanto não são falsas, mas não foram verificadas para todos os indivíduos da População, pelo que também não podemos afirmar que são verdadeiras! Existe assim um certo grau de incerteza (percentagem de erro) que é medido em termos de PROBABILIDADE.

Nesta fase procuramos estudar a adaptabilidade do modelo sugerido na fase anterior.

Esquematicamente, temos:




Porque é que é necessário o conceito de Probabilidade para se poder fazer Estatística?

De acordo com o que dissemos anteriormente sobre a Inferência Estatística, precisamos aqui da noção de Probabilidade, para medir o grau de incerteza que existe quando tiramos uma conclusão

para a População, a partir da observação da amostra. Seguidamente vamos tentar exemplificar este processo.

Vimos anteriormente que ao fazer uma análise de dados, em que se calculam estatísticas, a que chamamos **estimadores**, temos como objectivo tomar algumas decisões acerca de **parâmetros** desconhecidos, que descrevem as populações de onde foram feitas as observações. Este processo baseia-se **na distribuição de amostragem** da estatística utilizada para estimar o parâmetro em estudo. A distribuição de amostragem descreve a forma como se comporta uma estatística quando varia a amostra que se utilizou para a calcular. Vamos exemplificar de seguida um processo de fazer inferência estatística, nomeadamente num processo de estimação.

**Exemplo 6** - Suponhamos que se pretendia estimar qual a **percentagem p** de estudantes da Universidade de Lisboa que vivem em casa dos pais, no ano lectivo 1997/98. Feito um inquérito a 150 estudantes, seleccionados aleatoriamente das diferentes faculdades, em que se pedia para responderem SIM ou NÃO, caso vivessem ou não em casa dos pais, obtiveram-se 89 SIM, donde uma estimativa para a percentagem pretendida é $\hat{p} = \frac{89}{150} = 0.59$.

Será que podemos dizer que a percentagem pretendida **p** é 0.59? Não, já que se retirarmos outra amostra da mesma dimensão, o valor obtido para a estatística não será necessariamente o mesmo. Por exemplo, poderíamos ter recolhido mais 10 amostras de dimensão 150, e o nº de SIM's obtidos ser 87, 89, 85, 90, 87, 79, 89, 88, 86 e 90 para cada uma das amostras consideradas. Quer dizer que o valor da estatística varia de amostra para amostra.

Então como proceder? Para estudar esta variabilidade apresentada pela estatística, vai-se obter a sua distribuição de amostragem.

**Distribuição de amostragem** - Distribuição de amostragem de uma estatística é a distribuição dos valores que a estatística assume para todas as possíveis amostras, da mesma dimensão, da população.

Então para conhecer a distribuição de amostragem da estatística $\hat{\mathbf{p}}$, utilizada para **estimar** o **parâmetro p,** teríamos de ir considerar todas as amostras possíveis de dimensão 150 da população constituída pelos estudantes da Universidade de Lisboa. Para cada uma dessas amostras constituída por 150 estudantes investigaríamos qual a percentagem de SIM's, para em seguida com todos os valores obtidos para essas percentagens obtermos a distribuição de amostragem da estatística **percentagem**. Felizmente que não é necessário percorrer todo este caminho, pois então teria sido mais simples investigar todos os estudantes sobre a característica em estudo!

Como veremos mais tarde, a **teoria das probabilidades** permite-nos afirmar que se a dimensão n da amostra seleccionada for "suficientemente grande" então a distribuição de amostragem da estatística $\hat{p}$ é conhecida, mais precisamente é a distribuição Normal (com valor médio p e variância igual a $\frac{p(1-p)}{n}$), como veremos num capítulo posterior. Então vai ser possível construir um intervalo aleatório (veremos mais tarde o modo de o fazer),

$$[\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$$

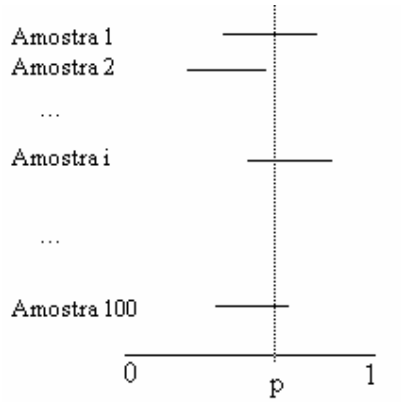
a que se dá o nome de intervalo de confiança para **p**, com uma confiança de 95%, em que

$$\text{Probabilidade}[\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le \mathbf{p} \le \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}] = .95$$

Aquele intervalo é aleatório na medida em que para cada amostra se obtém um valor para $\hat{\mathbf{p}}$ e correspondentemente, limites para o intervalo respectivo.

**Interpretação do intervalo de confiança**

Ao interpretar o intervalo de confiança deve-se ter em atenção que o que é aleatório é o intervalo e não a percentagem p (desconhecida, mas fixa) - a variabilidade existe no processo de amostragem e não no parâmetro. Quando se recolhem várias amostras, o valor de $\hat{\mathbf{p}}$ é diferente de amostra para amostra, pelo que os limites do intervalo variam.
Uma vez seleccionada uma amostra e obtido um valor para $\hat{\mathbf{p}}$, ao calcular um intervalo com 95% de confiança, não significa que a probabilidade do intervalo conter o parâmetro é .95, já que o intervalo contém ou não contém o parâmetro. Como deve ser interpretado o intervalo de confiança é da seguinte forma: ao recolher 100 amostras da mesma dimensão e ao calcular os intervalos correspondentes, aproximadamente 95 destes intervalos contêm o parâmetro p, enquanto que 5 não o contêm:




O processo que acabámos de descrever e que será desenvolvido num capítulo posterior é um exemplo de estimação intervalar.

**Exemplo 6** (cont) - Considerando finalmente o exemplo em estudo e tendo em conta o valor de 0.59 obtido para $\hat{\mathbf{p}}$, tem-se o intervalo [.51, .67] que com uma confiança de 95% contém o valor da percentagem de estudantes da Universidade de Lisboa que vivem em casa dos pais. Ao obtermos uma resposta para a nossa questão – qual a percentagem de estudantes que vivem em casa dos pais no ano lectivo 1997/98, sob a forma de um intervalo, obtivemos também a quantificação do erro cometido ao assumir essa resposta!

### 1.6 - Exemplos de aplicação da Estatística

Os campos de aplicação da Estatística são muitos e os mais variados. Por exemplo:
**Estudos de mercado** - O gerente de uma fábrica de detergentes pretende lançar um novo produto para lavar a loiça, pelo que encarrega uma empresa especialista em estudos de mercado, após realizar uma sondagem, de estimar a percentagem de potenciais compradores desse produto.

*População* - conjunto de todos os agregados familiares do país
*Amostra* - conjunto de alguns agregados familiares inquiridos pela empresa
*Problema* - pretende-se, a partir da percentagem de respostas afirmativas, de entre os inquiridos, sobre a compra do novo produto, obter uma estimativa do número de compradores, de entre todos os agregados familiares do país ( População).

**Medicina -** Pretende-se estudar o efeito de um novo medicamento, para curar determinada doença. É seleccionado um grupo de 20 doentes, administrando-se o novo medicamento a 10 desses doentes, escolhidos ao acaso, e o medicamento habitual aos restantes.

*População* - conjunto de todos os doentes com a doença que o medicamento a estudar pretende tratar

*Amostra* - conjunto de 20 doentes seleccionados

*Problema* - pretende-se a partir dos resultados obtidos, realizar um teste de hipóteses, para tomar uma decisão sobre qual dos medicamentos é melhor.

**Controlo de qualidade** - O administrador de uma fábrica de parafusos pretende assegurar-se de que a percentagem de peças defeituosas, não excede um determinado valor, a partir do qual uma encomenda poderia ser rejeitada (sondagem).

*População* - conjunto de todos os parafusos fabricados ou a fabricar pela fábrica, utilizando o mesmo processo

*Amostra* - conjunto de parafusos escolhidos ao acaso, de entre o lote de produzidos

*Problema* - pretende-se, a partir da percentagem de parafusos defeituosos na amostra, estimar a percentagem de defeituosos em toda a produção.

**Política de ensino** - O Ministério da Educação pretende saber se a prova de aferição em Matemática está bem construída, isto é, se seleccionou efectivamente os melhores alunos (sondagem).

*População* - conjunto de todos os alunos candidatos ao Ensino Superior, e respectivas notas em Matemática no 12º ano e na prova de aferição

*Amostra* - conjunto de alunos seleccionados aleatoriamente em todo o país, de entre a População considerada anteriormente

*Problema* - pretende-se determinar um coeficiente de associação, que indique se existe uma associação forte ou fraca, entre os dois conjuntos de notas, consideradas anteriormente.

**Pedagogia** - Um conjunto de padagogos, desenvolveu uma técnica nova para a aprendizagem da leitura, na escola primária, a qual, segundo dizem, encurta o tempo de aprendizagem, relativamente ao método habitual (pretende-se fazer uma experimentação).

*População* - conjunto de todos os alunos que entram para a escola primária sem saber ler

*Amostra* - conjunto de alunos de algumas escolas, seleccionadas aleatoriamente para este estudo. Os alunos foram separados por dois grupos para se aplicarem as duas técnicas em confronto

*Problema* - do estudo da amostra, decidir qual a melhor técnica.

**Exercícios de revisão**

**1** - Considere a seguinte situação: Um político, candidato a Presidente da República, pretende ter uma ideia de qual a sua representatividade, junto do eleitorado português, pelo que encarrega uma empresa de fazer o estudo conveniente. Identifique: População e Amostra.

**2** - Diga porque é que as seguintes situações representam más amostras:

- Para saber qual o candidato mais votado, para a Câmara de determinada cidade, auscultou-se a opinião dos clientes de determinado supermercado.

- Para conhecer a situação financeira das empresas têxteis portuguesas, verificou-se a situação das empresas que tiveram maior volume de exportações, no último ano.

**3** - Em 1985 verificaram-se, nos Estados Unidos, 19893 assassínios, enquanto que em 1970 se tinham verificado 16848 - um aumento de cerca de 20%. Estes números significam que os Estados Unidos se tornou um país violento no período 1970-1985?

**4** - Num determinado distrito de Portugal, foi levada a cabo uma experiência para verificar o efeito da distribuição de leite às crianças em idade escolar. Assim, foram escolhidas algumas crianças em cada escola para pertencerem ao grupo de tratamento, a quem foi dado leite e outras a quem não foi dado leite, constituindo o grupo de controlo. Para tornar os grupos equivalentes em termos de nível familiar e de saúde, a atribuição de cada criança a cada grupo foi feita aleatoriamente. Contudo, verificou-se que, apesar da atribuição aleatória, havia ainda pequenas diferenças entre os grupos. Permitiu-se então que os professores fizessem a selecção das crianças, com o objectivo de tornar os grupos comparáveis. Terá sido este um procedimento correcto?

**5** - De acordo com um estudo observacional, feito na Califórnia, verificou-se que a taxa de cancro cervical era maior entre as utilizadoras de contraceptivos orais, do que entre as que não os utilizavam, mesmo tendo em consideração os factores idade, educação, estado civil, religião e o facto de ser fumadora ou não. Os investigadores concluíram que a pílula causava o cancro cervical. O que acha desta conclusão?

**6** – A revista "Filhos e Pais" pediu a uma empresa de sondagens que elaborasse um estudo sobre a opinião dos Pais relativamente à utilidade, sob o ponto de vista educacional, de bater nos filhos. Foram postas as seguintes questões aos pais que faziam parte de uma amostra aleatória: i) Acredita que se deve bater nos filhos? ii) Bateu nos seus filhos? iii) Se a resposta à questão anterior foi sim, com que frequência?

No estudo anterior poderá estar envolvido algum tipo de erro de não amostragem?

**7** – Um investigador pretendendo fazer um estudo sobre a relação entre a quantidade de ovos consumidos semanalmente e o nível do colesterol, pediu a colaboração de voluntários para entrarem neste estudo. Apresentaram-se 2589 voluntários. O investigador colheu informação sobre a quantidade de ovos consumida e o nível de colesterol de cada uma das pessoas apresentadas, tendo concluído que existia um a forte associação entre as duas variáveis.

- Estamos perante um estudo observacional ou uma experimentação controlada?
- Baseado neste estudo pode o investigador concluir que o consumo de ovos aumenta o nível de colesterol? Explique.

**8** - Quais os objectivos da Estatística Descritiva e da Inferência Estatística?

**9** - As inferências estatísticas são sempre correctas?

**Capítulo 2**

# Análise, representação e redução de dados

## 2.1 - Introdução

Vimos no capítulo 1, que o objectivo da Estatística é o estudo de Populações, isto é, conjuntos de indivíduos (não necessariamente pessoas) com características comuns, que se pretendem estudar. A uma característica comum, que assume valores diferentes de indivíduo para indivíduo, chamamos **variável**. Sendo então o nosso objectivo o estudo de uma (ou mais) característica da População, vamos identificar População com a variável (característica) que se está a estudar, dizendo que a População é constituída por todos os valores que a variável pode assumir. Por exemplo, relativamente à população portuguesa, se o objectivo do nosso estudo for a característica altura, diremos que a população é constituída por todos os valores possíveis para a variável altura.

Vimos também que um dos conceitos fundamentais em Estatística é o de amostra. Quando falamos em amostras, entendemos conjuntos de dados, que representem convenientemente as Populações subjacentes. Observe-se que estamos, portanto, a identificar amostra com o resultado das observações feitas sobre os elementos da população a que chamámos amostra.

Neste momento vamos admitir que dispomos de um desses conjuntos de dados, sem nos preocuparmos como foram obtidos, e pretendemos desenvolver processos de análise que nos permitam responder a algumas questões, tais como:

- Serão os dados quase todos iguais?
- Serão muito diferentes, uns dos outros?
- De que modo é que são diferentes?
- Existe alguma estrutura subjacente ou alguma tendência?
- Existem alguns agrupamentos especiais?
- Existem alguns dados muito diferentes da maior parte?

Estas questões, de um modo geral, não podem ser respondidas rapidamente, olhando unicamente para o conjunto dos dados! No entanto, se estiverem organizados sob a forma de tabelas ou gráficos, já a resposta às questões anteriores se torna mais simples.

Seguidamente começaremos por dar uma possível classificação para os dados e os processos adequados para a sua representação. Estes processos de redução dos dados permitem realçar

as características principais e a estrutura subjacente, à custa de alguma informação que se perde, mas que não é relevante para o estudo em vista.

## 2.2 - Tipos de dados

As variáveis podem ser de dois tipos: **qualitativas e quantitativas**. Para os dados também se usa a mesma terminologia, conforme resultam da observação de variáveis qualitativas ou quantitativas.

Na determinação da análise estatística apropriada para um conjunto de dados, é importante classificar as variáveis quanto ao tipo. Depois de várias tentativas terem sido feitas, existe um sistema de classificação normalmente aceite e proposto por Stevens - Handbook of experimental psychology (1951), que apresentamos de seguida.

### 2.2.1 - Dados qualitativos

**Dados qualitativos** - Representam a informação que identifica alguma qualidade, categoria ou característica, não susceptível de medida, mas de classificação, assumindo várias modalidades.

Por exemplo, o estado civil de um indivíduo é um dado qualitativo, assumindo as categorias: solteiro, casado, divorciado e viúvo.

Os dados de tipo qualitativo ainda se podem exprimir na *escala nominal* ou *na escala ordinal*:

**a) Variáveis nominais**

Uma variável é nominal se cada observação pertence a uma de várias categorias distintas. Estas categorias não são necessariamente numéricas, embora se possa utilizar números para as representar. Por exemplo, a variável sexo é nominal, já que um indivíduo é do sexo masculino ou feminino. Podemos utilizar os símbolos M e F, mas também podemos utilizar os números 1 e 2 para as categorias masculino e feminino. Uma variável nominal pode apresentar duas ou mais categorias; alguns exemplos de variáveis nominais apresentando mais de duas categorias, são a religião, raça, etc.

A estrutura da escala nominal não é destruída por uma substituição biunívoca. Para estes dados não tem sentido falar em média ou mediana. A única medida de localização que tem sentido é a moda - categoria com maior número de elementos.

**b) Variáveis ordinais**

Para as variáveis ordinais também se utilizam as categorias mas, no entanto, existe uma relação de ordem entre elas. Por exemplo uma tabela que classifica os minerais e rochas, quanto à dureza, tem as categorias ordenadas segundo 10 níveis de dureza, do mais duro - Diamante, ao menos duro - Talco. A estrutura desta escala não é distorcida por uma substituição que preserve

a ordem. No exemplo da classificação dos minerais, em vez dos números de 1 a 10, podem-se utilizar as letras de A a J.

Como é evidente, e do mesmo modo que para as variáveis nominais, continua a não ter sentido o cálculo da média. Pode-se calcular a moda e, já que existe uma ordenação, pode-se calcular a mediana.

Estes dados são organizados na forma de uma **tabela de frequências**, que apresenta o número de elementos - **frequência absoluta** (ou só frequência) de cada uma das categorias ou **classes**.

Numa tabela de frequências, além das frequências absolutas, também se apresentam as **frequências relativas**, onde

$$\textbf{frequência relativa} = \frac{\text{frequência absoluta}}{\text{dimensão da amostra}}$$

entendendo-se por **dimensão** da amostra o número de elementos da amostra.

**Exemplo 1** (De Veaux et al, 2004) – O que aconteceu ao Titanic, na noite de 14 de Abril de 1912, é bem conhecido. Apresentamos de seguida alguns dados relativos aos passageiros e tripulação, nomeadamente no que diz respeito se sim ou não, a pessoa *Sobreviveu* (Morta ou Viva), a sua *Idade* (Adulto ou Criança), *Sexo* (feminino ou Masculino), e a *Classe* em que viajava (Primeira, Segunda, Terceira ou Tripulação):

| Sobreviveu | Idade | Sexo | Classe |
|---|---|---|---|
| Morta | Adulto | Masculino | Terceira |
| Morta | Adulto | Masculino | Tripulação |
| Morta | Adulto | Masculino | Terceira |
| Morta | Adulto | Masculino | Tripulação |
| Morta | Adulto | Masculino | Tripulação |
| Morta | Adulto | Masculino | Tripulação |
| Viva | Adulto | Feminino | Primeira |
| Morta | Adulto | Masculino | Terceira |
| Morta | Adulto | Masculino | Tripulação |

Parte de uma tabela em que se mostra para 9 passageiros, as categorias referentes a 4 variáveis

O que fazer com dados como estes? Um princípio básico de uma análise de dados de qualquer tipo é proceder à sua representação gráfica. Para isso é necessário proceder ao seu agrupamento, através de uma tabela de frequências. Organizando os dados referentes à variável Classe, obteve-se a seguinte distribuição para os 2201 passageiros:

| Classe | Freq. Absoluta | Freq. relativa |
|---|---|---|
| Primeira | 325 | 0.148 |
| Segunda | 285 | 0.129 |
| Terceira | 706 | 0.321 |
| Tripulação | 885 | 0.402 |
| Total | 2201 | 1.000 |

Da tabela anterior concluímos imediatamente que a maior percentagem de passageiros eram tripulantes e que dos que tinham comprado bilhete, eram mais frequentes os que viajavam em 3ª classe, seguidos da 1ª classe e finalmente os menos frequentes eram os que viajavam em 2ª

classe. Estas conclusões não eram evidentes a partir dos dados inicialmente considerados. Ao fazer a redução, sob a forma de uma tabela de frequências, a única informação que se perdeu foi a ordenação inicial dos dados, que neste caso não era relevante.

**Exemplo 2** - A seguinte tabela apresenta a distribuição do pessoal docente (freq. absolutas), segundo os ramos de ensino, em Portugal Continental, durante os anos de 1985-1986, 1986-1987, 1987-1988, 1988-1989, 1989-1990, 1990-1991 (Fonte : Anuário Estatístico de Portugal - 1992):

| | **Básico** | | | **Secund** | | | | |
|---|---|---|---|---|---|---|---|---|
| | Pré-escolar | Primário | Preparat. | Sec. Unific | Sec. comp. | 12ºano | Liceal | Técnico |
| **1985-86** | 5991 | 41534 | 29189 | 28675 | 14187 | 3584 | 3069 | 2216 |
| **1986-87** | 4583(a) | 41553 | 31742 | 28751 | 15171 | 4136 | 3454 | 2656 |
| **1987-88** | 4430 | x | 29486 | 32272 | 18140 | 4738 | 4192 | 2893 |
| **1988-89** | 6368 | x | 35420 | 38881 | 21825 | 7337 | 6740 | 4809 |
| **1989-90** | 7041 | x | 32731 | 40065 | 21628 | 6079 | 5471 | 2857 |
| **1990-91** | 9317 | x | 33015 | 42229 | 23868 | 7003 | 6205 | 3178 |

(cont)

| | **Cursos Profis.** | **Artístico** | **Médio** | | **Superior** | Total |
|---|---|---|---|---|---|---|
| | | | Mag.Infantil | Mag.Primário | | |
| **1985-86** | 1281 | 629 | 535 | 571 | 9620 | 141081 |
| **1986-87** | 969 | 602 | 414 | 485 | 9234(b) | 143750 |
| **1987-88** | 1389 | 736 | x | x | 10769 | x |
| **1988-89** | 1105 | 678 | x | x | 12113 | x |
| **1989-90** | 351 | 716 | x | x | 10405 | x |
| **1990-91** | 372 | 723 | x | x | 14223 | x |

Obs: (a) Não funcionaram 190 estabelecimentos por diversos motivos
(b) A Univ do Porto apenas enviou os elementos relativos ao pessoal docente das Fac de Economia e Arquitectura
x - informação não disponível

Das tabelas anteriores pode-se retirar bastante informação, nomeadamente no que diz respeito à evolução do nº de docentes nas diferentes categorias, desde 1985 até 1991.

Para os únicos anos onde existe informação completa, os anos lectivos 1985-1986 e 1986-1987, considerámos a tabela das frequências relativas, que apresentamos de seguida:

| | | **Básico** | | **Secund** | | | | |
|---|---|---|---|---|---|---|---|---|
| | Pré-escolar | Primário | Preparat. | Sec. Unific | Sec. comp. | 12ºano | Liceal | Técnico |
| **1985-86** | 0.042 | 0.294 | 0.207 | 0.203 | 0.101 | 0.025 | 0.022 | 0.016 |
| **1986-87** | 0.032 | 0.289 | 0.221 | 0.200 | 0.106 | 0.029 | 0.024 | 0.018 |

(cont)

| | **Cursos Profis.** | **Artístico** | **Médio** | | **Superior** | Total |
|---|---|---|---|---|---|---|
| | | | Mag.Infantil | Mag.Primário | | |
| **1985-86** | 0.009 | 0.004 | 0.004 | 0.004 | 0.068 | 1 |
| **1986-87** | 0.007 | 0.004 | 0.003 | 0.003 | 0.064 | 1 |

Da tabela das frequências relativas, rapidamente se conclui que a classe predominante é a dos Professores Primários enquanto que os Professores do ensino Médio e Artístico, são só uma pequena percentagem do total de docentes.

Quando se constrói uma tabela de frequências, a partir de uma amostra, um processo de fácil verificação de que as frequências estão bem calculadas consiste em somá-las para todas as classes consideradas, pois:

**A soma das frequências absolutas é igual à dimensão da amostra**
e
**A soma das frequências relativas é igual a 1**

Como consequência da observação anterior, a utilização das frequências relativas é preferível, relativamente às frequências absolutas, pois assim é possível fazer a comparação de amostras de dimensões diferentes. É o que se passa no caso do exemplo 2, em que as dimensões das amostras relativamente a 1985-1986 e 1986-1987 são respectivamente 141081 e 143750.

### 2.2.2 - Dados quantitativos

**Dados quantitativos** - Representam a informação resultante de características susceptíveis de serem medidas, apresentando-se com diferentes intensidades.

Os dados de tipo quantitativo ainda se podem exprimir na *escala intervalar* e *percentual*:

**a) Variáveis intervalares**

Uma variável intervalar é uma espécie de variável ordinal, mas em que as diferenças entre valores sucessivos são sempre iguais. Por exemplo a temperatura medida em graus Fahrenheit (F) é intervalar, já que a diferença entre 12° e 13° é a mesma que a diferença entre 14° e 15°. Define-se assim uma unidade de medida e um zero arbitrário. Na realidade, a variável temperatura também pode ser medida em graus centígrados (C), correspondendo os 0° C aos 32° F. Assim, a transformação linear $C = 5/9\ (F-32)$
transforma a temperatura F, em graus Fahrenheit, na temperatura C, em graus centígrados. A estrutura desta escala não é destruída por uma substituição que preserve a igualdade dos intervalos.
Para este tipo de variáveis já tem sentido o cálculo da média.

**b) Variáveis percentuais (ou absolutas)**

As variáveis percentuais são as variáveis intervalares para as quais existe um zero absoluto, que representa a origem das medidas. Por exemplo, a variável altura é percentual - podemos dizer que uma altura de 164 cm é o dobro de uma altura de 82 cm. Se mudarmos a unidade de medida para o metro, continuamos ainda a dizer que a 1ª altura é o dobro da 2ª. A estrutura da escala percentual não vem distorcida quando se fazem transformações da forma $x'=kx$.

**Outras classificações**

Além da classificação referida anteriormente, as variáveis também podem ser classificadas em *discretas* e *contínuas*. Uma variável é contínua se pode assumir qualquer valor de um intervalo contido no domínio da variável. Caso contrário será discreta.

Todas as variáveis nominais e ordinais são discretas. As variáveis intervalares e percentuais podem ser discretas ou contínuas.

A classificação de uma variável em discreta ou contínua, é por vezes susceptível de algumas dúvidas. Por exemplo a variável idade, ao contrário do que possa parecer à primeira vista, já que só utilizamos números inteiros para a representar, é uma variável contínua, pois a diferença de idade entre dois indivíduos pode ser tão pequena quanto se queira - um ano, um mês, uma hora, um minuto, …..Como o nome indica, dados dois valores observados de uma variável contínua, passamos de um valor a outro de forma contínua - qualquer valor intermédio ainda é um valor da variável. No caso da variável ser discreta, passamos de um valor a outro por saltos!

**Como organizar os dados?**

Os dados são organizados na forma de uma **tabela de frequências**, do mesmo modo que os dados qualitativos. No entanto convém fazer distinção entre os dados discretos e contínuos, já que a construção da tabela de frequências se processa, de um modo geral, de forma diferente.

### 2.2.2.1 – Organização de dados discretos

No caso de dados discretos, a construção da tabela de frequências é análoga à que foi feita para os dados qualitativos, mas em vez das categorias consideram-se os valores distintos que surgem na amostra, os quais vão constituir as **classes**.

**Exemplo 3** - Numa turma do 10º ano da Escola Secundária Professor Herculano de Carvalho, os alunos registaram o nº de irmãos, tendo-se obtido a seguinte amostra:

1 2 2 1 3 0 0 1 1 2 1 1 1 0 0 3 4 3 1 2

A tabela de frequências correspondente à amostra anterior é a seguinte:

Tabela de frequências

| Classes | Freq. abs. | Freq. rel. |
|---|---|---|
| 0 | 4 | .20 |
| 1 | 8 | .40 |
| 2 | 4 | .20 |
| 3 | 3 | .15 |
| 4 | 1 | .05 |
| Total | 20 | 1 |

Podemos no entanto dispor de uma amostra de dados discretos, mas estes assumirem muitos valores distintos, que torne pouco prático a construção de uma tabela de frequências, onde se consideram todos esses valores. Neste caso procede-se a um agrupamento conveniente para os dados, como se exemplifica a seguir:

**Exemplo 4** - No Distrito Sanitário de Chicago, a escolha dos técnicos é feita mediante um exame. Em 1966, havia 233 candidatos para 15 lugares. O exame teve lugar no dia 12 de Março e os resultados dos testes apresentam-se a seguir (Freedman and al., 1991 *Statistics*, pag.51):

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 26 | 27 | 27 | 27 | 27 | 29 | 30 | 30 | 30 | 30 | 31 | 31 | 31 | 32 | 32 |
| 33 | 33 | 33 | 33 | 33 | 34 | 34 | 34 | 35 | 35 | 36 | 36 | 36 | 37 | 37 |
| 37 | 37 | 37 | 37 | 37 | 39 | 39 | 39 | 39 | 39 | 39 | 39 | 40 | 41 | 42 |
| 42 | 42 | 42 | 42 | 43 | 43 | 43 | 43 | 43 | 43 | 43 | 43 | 44 | 44 | 44 |
| 44 | 44 | 44 | 45 | 45 | 45 | 45 | 45 | 45 | 45 | 46 | 46 | 46 | 46 | 46 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 46 | 47 | 47 | 47 | 47 | 47 | 47 | 48 | 48 | 48 | 48 | 48 | 48 | 48 | 48 |
| 49 | 49 | 49 | 49 | 50 | 50 | 51 | 51 | 51 | 51 | 51 | 52 | 52 | 52 | 52 |
| 52 | 53 | 53 | 53 | 53 | 53 | 54 | 54 | 54 | 54 | 54 | 55 | 55 | 55 | 56 |
| 56 | 56 | 56 | 56 | 57 | 57 | 57 | 57 | 58 | 58 | 58 | 58 | 58 | 58 | 58 |
| 58 | 59 | 59 | 59 | 59 | 60 | 60 | 60 | 60 | 60 | 60 | 61 | 61 | 61 | 61 |
| 61 | 61 | 62 | 62 | 62 | 63 | 63 | 64 | 65 | 66 | 66 | 66 | 67 | 67 | 67 |
| 67 | 68 | 68 | 68 | 69 | 69 | 69 | 69 | 69 | 69 | 69 | 71 | 71 | 72 | 73 |
| 74 | 74 | 74 | 75 | 75 | 76 | 76 | 78 | 80 | 80 | 80 | 80 | 81 | 81 | 81 |
| 82 | 82 | 83 | 83 | 83 | 83 | 84 | 84 | 84 | 84 | 84 | 84 | 84 | 90 | 90 |
| 90 | 91 | 91 | 91 | 92 | 92 | 92 | 93 | 93 | 93 | 93 | 95 | 95 | | |

Neste caso a construção da tabela de frequências poderia processar-se do mesmo modo que no exemplo anterior; resultaria, no entanto, uma tabela com demasiadas classes. Assim, resolvemos tomar como classes uma partição natural, para os dados considerados, que é a seguinte: considerar como classes os intervalos [20, 30[, [30,40[, [40,50[, [50,60[, [60,70[, [70,80[, [80,90[, [90,100[. A forma do intervalo [ , [ significa que o limite inferior do intervalo pertence à classe, enquanto que o limite superior não pertence. Assim, um elemento da amostra igual a 30 pertencerá à 2ª classe e não à 1ª.

Tabela de frequências

| Classes | Freq. abs. | freq. rel. |
|---|---|---|
| [20,30[ | 6 | .027 |
| [30,40[ | 36 | .161 |
| [40,50[ | 52 | .233 |
| [50,60[ | 46 | .206 |
| [60,70[ | 36 | .161 |
| [70,80[ | 12 | .054 |
| [80,90[ | 20 | .090 |
| [90,100[ | 15 | .067 |
| Total | 223 | 1 |

Enquanto que no caso dos dados discretos a construção da tabela de frequências é de um modo geral muito simples, no caso de variáveis contínuas o processo de resumir a informação constituída pelos resultados das suas observações, é um pouco mais elaborado, já que a definição das classes não é tão imediata. Efectivamente não tem sentido considerar, para classes, os diferentes valores que surgem na amostra, pois eventualmente eles são todos diferentes.

### 2.2.2.2 – Organização de dados contínuos

Para a organização e redução de dados contínuos, podem-se considerar as seguintes etapas:

**1- Definição das classes**

a) Determinar a **amplitude** da amostra, isto é, a diferença entre o valor máximo e o valor mínimo;

b) Dividir essa amplitude pelo número **k**[(1)], de classes que se desejam considerar; tomar para **amplitude de classe h**, um valor aproximado <u>por excesso</u>, do valor anteriormente obtido;

c) Construir as classes de modo que tenham todas a mesma amplitude e cuja união contenha todos os elementos da amostra.

Uma metodologia a seguir para a construção das classes $C_i = [c_i, c_{i+1}[$, poderá ser a seguinte: a primeira classe $C_1$ será $C_1$ = [mínimo da amostra, mínimo da amostra + h[. As outras classes serão

(1) A definir posteriormente.

$C_i$= [mínimo da amostra + (i-1) h, mínimo da amostra + i h[ com i=2, …, k.

Nem sempre se consegue aplicar a metodologia anterior de considerar todas as classes com a mesma amplitude. No entanto uma regra a ter presente é que estas classes devem ser todas disjuntas duas a duas e a sua união deve conter todos os elementos da amostra.

**2 - Contagem** do número de elementos de cada classe.

Conta-se o número de elementos da amostra, que pertencem a cada classe. Analogamente ao que foi considerado no caso dos dados discretos, esses valores serão as frequências absolutas das classes.

Quantas classes se devem considerar para fazer a redução de um conjunto de dados? Qual o valor de **k**?

Existe uma regra empírica, chamada regra de Sturges, que nos dá um valor aproximado para o número de classes que se devem considerar e que é a seguinte:

**Regra de Sturges** - Para uma amostra de dimensão n, o nº de classes é dado pelo menor inteiro k tal que $2^k > n$.

**Exemplo 5** - Consideremos a amostra constituída pelas notas obtidas num ponto de Matemática, de uma determinada turma:

12.1; 8.9; 16.2; 8.2; 9.8; 15.1; 14.5; 13.4; 14.7; 7.5; 8.8; 12.4; 16.1; 15.2; 13.5; 13.8; 14.6; 15.5; 7.8; 12.5;13.2; 11.0; 10.5

De acordo com a metodologia apresentada anteriormente, temos:

Amplitude da amostra: 16.2 - 7.5 = 8.7
Número de classes: k = 5
Amplitude de classe: 8.7/5 = 1.74 -> h = 1.8
Classes: [7.5, 9.3[, [9.3, 11.1[, [11.1, 12.9[,[12.9, 14.7[, [14.7, 16.5[

Tabela de frequências

| Classes | Freq. abs. | Freq. rel. |
|---|---|---|
| [7.5, 9.3[ | 5 | .2174 |
| [9.3, 11.1[ | 3 | .1304 |
| [11.1, 12.9[ | 3 | .1304 |
| [12.9, 14.7[ | 6 | .2609 |
| [14.7, 16.5[ | 6 | .2609 |
| Total | 23 | 1 |

Obs: Não esquecer que a amplitude de classe h, é um valor aproximado **por excesso** do quociente $\frac{\text{amplitude da amostra}}{\text{nº de classes}}$. Se, por exemplo, o quociente anterior for igual a 2.15, pode-se considerar 2.2; se for igual a 2, pode-se utilizar o 2.1 (Porque é que, neste caso, não se deve utilizar para amplitude de classe 2?). Uma regra simples poderá ser a de considerar para amplitude de classe um valor com mais uma casa decimal do que o número de casas decimais com que se apresentam os dados. Neste caso, uma boa escolha para a amplitude de classe seria o valor 1.75.

**Nota 1**: A regra enunciada anteriormente para o número de classes, é uma de várias regras existentes e que se verificou dar bons resultados quando se procede à representação gráfica, sob a forma de histograma, a partir dos dados agrupados. Existem outras regras como por exemplo a sugerida por Velleman em 1976, que considera para k o maior inteiro contido em $2\sqrt{n}$ e a

considerada num trabalho de Dixon e Kronmal em 1965, que utilizam para k o maior inteiro contido em $10 \times \log_{10} n$.

**Nota 2**: A metodologia utilizada para a construção das classes não é única. Pode-se, por exemplo, decidir construir as classes fechadas à direita e abertas à esquerda, metodologia seguida pela folha de cálculo Excel, quando se utiliza a função *Frequency*, ou seguir a metodologia indicada, utilizada, por exemplo, no software de Estatística Statview, ou ainda utilizar outras abordagens diferentes.

## Utilização do Excel na obtenção de tabelas de frequências

**1 - Dados de tipo qualitativo ou quantitativo discreto**

Vamos exemplificar a utilização do Excel na construção de tabelas de frequência de dados qualitativos ou quantitativos discretos.

O procedimento para a construção de tabelas de frequência é idêntico, quer tenhamos um conjunto de dados qualitativos ou quantitativos discretos, já que as classes que se consideram para a tabela de frequência são, de um modo geral, como vimos anteriormente, as diferentes categorias ou valores que surgem, respectivamente, no conjunto de dados.

**1.1 - Função *COUNTIF***

**Exemplo** – Considerando ainda o ficheiro EscolasSec.xls, vamos agrupar os dados segundo a variável qualitativa Distrito. Para ver quais as diferentes categorias que a variável assume, um processo simples é proceder à ordenação dos dados segundo aquela variável. Fizémos essa ordenação e escrevemos as diferentes categorias nas células F3:F18. A seguir utilizámos a função ***COUNTIF(a;b)***, que devolve o número de células consideradas no argumento **a**, que são iguais à categoria considerada no argumento **b**:

EscolasSec.xls

| | F | G |
|---|---|---|
| 1 | | |
| 2 | **Categorias** | **Freq. Absoluta** |
| 3 | Aveiro | =COUNTIF($C$2:$C$71;F3) |
| 4 | Beja | =COUNTIF($C$2:$C$71;F4) |
| 5 | Braga | =COUNTIF($C$2:$C$71;F5) |
| 6 | Castelo Branco | =COUNTIF($C$2:$C$71;F6) |
| 7 | Coimbra | =COUNTIF($C$2:$C$71;F7) |
| 8 | Évora | =COUNTIF($C$2:$C$71;F8) |
| 9 | Faro | =COUNTIF($C$2:$C$71;F9) |
| 10 | Guarda | =COUNTIF($C$2:$C$71;F10) |
| 11 | Leiria | =COUNTIF($C$2:$C$71;F11) |
| 12 | Lisboa | =COUNTIF($C$2:$C$71;F12) |
| 13 | Portalegre | =COUNTIF($C$2:$C$71;F13) |
| 14 | Porto | =COUNTIF($C$2:$C$71;F14) |
| 15 | Santarém | =COUNTIF($C$2:$C$71;F15) |
| 16 | Setúbal | =COUNTIF($C$2:$C$71;F16) |
| 17 | Viana do Castelo | =COUNTIF($C$2:$C$71;F17) |
| 18 | Viseu | =COUNTIF($C$2:$C$71;F18) |
| 19 | | =SUM(G3:G18) |

Sheet1 / Sheet

EscolasSec.xls

| | F | G |
|---|---|---|
| 1 | | |
| 2 | **Categorias** | **Freq. Absoluta** |
| 3 | Aveiro | 4 |
| 4 | Beja | 3 |
| 5 | Braga | 1 |
| 6 | Castelo Branc | 1 |
| 7 | Coimbra | 10 |
| 8 | Évora | 2 |
| 9 | Faro | 12 |
| 10 | Guarda | 1 |
| 11 | Leiria | 4 |
| 12 | Lisboa | 21 |
| 13 | Portalegre | 1 |
| 14 | Porto | 1 |
| 15 | Santarém | 2 |
| 16 | Setúbal | 3 |
| 17 | Viana do Cast | 2 |
| 18 | Viseu | 2 |
| 19 | | 70 |

Sheet1

Recomenda-se que ao construir a tabela de frequências, se proceda à soma das frequências absolutas, para confirmar que a soma é igual ao número de elmentos do conjunto de dados que se está a agrupar.

**1.2 - *PivotTable***

Outro processo que pode ser utilizado para construir uma tabela de frequências de dados qualitativos ou quantitativos discretos, é usando a PivotTable, como se mostra a seguir.

**Exemplo** – Utilizando a PivotTable, proceda ao agrupamento dos dados da variável Distrito, do ficheiro EscolasSec.xls.

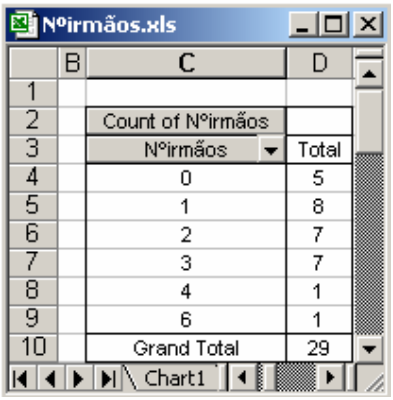
- No menu Data, clicar em *PivotTable and PivotChart Report*;
- No passo 1 da *PivotTable and PivotTable Wizard*, seguir as instruções, e clicar *PivotTable* à pergunta *What kind of report do you want to create*?;
- No passo 2 seguir as instruções, seleccionando os dados que se pretende usar (não esquecer de seleccionar os títulos). Neste caso seleccionar as células C1 a C71;
- No passo 3 seleccionar o lugar onde pretende criar a tabela. Nós optámos por seleccionar a célula E2;
- Arrastar o botão Distrito da barra *PivotTable*, e colocá-lo (drop it) no campo *Row*; Arrastar ainda o botão Distrito e colocá-lo (drop it) no campo *Data*:

| | E | F | G | H | I |
|---|---|---|---|---|---|
| 1 | | | | | |
| 2 | Count of Distrito | | | **Tabela de frequências** | |
| 3 | Distrito | Total | | Classes | Freq. Abs. |
| 4 | Aveiro | 4 | | Aveiro | 4 |
| 5 | Beja | 3 | | Beja | 3 |
| 6 | Braga | 1 | | Braga | 1 |
| 7 | Castelo Branco | 1 | | Castelo Branco | 1 |
| 8 | Coimbra | 10 | | Coimbra | 10 |
| 9 | Évora | 2 | | Évora | 2 |
| 10 | Faro | 12 | | Faro | 12 |
| 11 | Guarda | 1 | | Guarda | 1 |
| 12 | Leiria | 4 | | Leiria | 4 |
| 13 | Lisboa | 21 | | Lisboa | 21 |
| 14 | Portalegre | 1 | | Portalegre | 1 |
| 15 | Porto | 1 | | Porto | 1 |
| 16 | Santarém | 2 | | Santarém | 2 |
| 17 | Setúbal | 3 | | Setúbal | 3 |
| 18 | Viana do Castelo | 2 | | Viana do Castelo | 2 |
| 19 | Viseu | 2 | | Viseu | 2 |
| 20 | Grand Total | 70 | | Total | 70 |

O procedimento anterior conduziu-nos à tabela do lado esquerdo da figura anterior, cujo conteúdo foi copiado para construir a tabela do lado direito, com uma apresentação mais sugestiva. Pode obter as frequências relativas, em vez das absolutas, clicando duas vezes em *Count of Distrito* e seleccionando sucessivamnete *Options>>* → *Show data as:* → *% of total*.

**Exemplo** – A um conjunto de 25 alunos de uma escola, perguntou-se quantos irmãos tinham, tendo-se obtido os seguintes valores: 1, 2, 1, 0, 3, 3, 2, 1, 0, 1, 2, 2, 3, 1, 1, 0, 2, 3, 1, 2, 0, 3, 4, 1, 6. Proceda a um agrupamento conveniente dos dados.

Começámos por inserir os dados numa folha de Excel e utilizando o procedimento anterior, obtivémos a seguinte tabela:



| | B | C | D |
|---|---|---|---|
| 1 | | | |
| 2 | | Count of Nºirmãos | |
| 3 | | Nºirmãos | Total |
| 4 | | 0 | 5 |
| 5 | | 1 | 8 |
| 6 | | 2 | 7 |
| 7 | | 3 | 7 |
| 8 | | 4 | 1 |
| 9 | | 6 | 1 |
| 10 | | Grand Total | 29 |

Observação: Se ao construir uma tabela de frequências de dados quantitativos discretos, faltar algum valor entre o mínimo e o máximo, deve-se considerá-lo na tabela, com frequência nula, se a seguir se pretende construir um diagrama de barras.

## 2 - Dados de tipo contínuo

### 2.1 – Utilização da função COUNTIF

**Exemplo** – Considere o seguinte ficheiro – IdadeTrabalhadores.xls, constituído pelas idades de 180 indivíduos escolhidos aleatoriamente de entre os trabalhadores de uma grande empresa têxtil. Proceda ao agrupamento dos dados, da forma que achar conveniente.

IdadeTrabalhadores.xls

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 18 | 26 | 33 | 40 | 44 | 47 | 51 | 53 | 57 | 21 | 31 | 36 | 42 | 45 | 48 | 52 | 55 | 61 |
| 2 | 19 | 27 | 33 | 40 | 44 | 47 | 51 | 53 | 57 | 22 | 31 | 38 | 42 | 45 | 49 | 52 | 55 | 61 |
| 3 | 19 | 27 | 33 | 40 | 44 | 48 | 51 | 53 | 57 | 22 | 31 | 38 | 42 | 45 | 49 | 52 | 55 | 61 |
| 4 | 19 | 28 | 34 | 40 | 44 | 48 | 51 | 54 | 57 | 22 | 31 | 39 | 42 | 46 | 49 | 52 | 56 | 62 |
| 5 | 20 | 28 | 34 | 41 | 44 | 48 | 51 | 54 | 57 | 23 | 31 | 39 | 42 | 46 | 49 | 52 | 56 | 62 |
| 6 | 20 | 29 | 34 | 41 | 45 | 48 | 51 | 54 | 57 | 23 | 32 | 39 | 43 | 46 | 50 | 52 | 56 | 63 |
| 7 | 20 | 30 | 35 | 41 | 45 | 48 | 51 | 54 | 58 | 24 | 32 | 39 | 43 | 47 | 50 | 52 | 56 | 63 |
| 8 | 21 | 30 | 35 | 41 | 45 | 48 | 51 | 54 | 58 | 25 | 32 | 39 | 43 | 47 | 50 | 52 | 56 | 63 |
| 9 | 21 | 30 | 35 | 42 | 45 | 48 | 52 | 55 | 58 | 25 | 33 | 39 | 43 | 47 | 50 | 52 | 56 | 63 |
| 10 | 21 | 31 | 36 | 42 | 45 | 48 | 52 | 55 | 59 | 25 | 33 | 39 | 43 | 47 | 51 | 53 | 56 | 64 |

Sheet1 / Sheet2 / Sheet3

Vamos utilizar a metodologia indicada na secção 2.2.2.2, para a definição das classes.

Definição das classes

1. Determinar a amplitude da amostra, subtraindo o mínimo do máximo;
2. Calcular a amplitude de classe h, dividindo a amplitude da amostra pelo número K de classes pretendido e tomando para h um valor aproximado por excesso do quociente anteriormente obtido. Existe uma regra empírica que nos dá um valor aproximado para o número K de classes e que consiste no seguinte: para uma amostra de dimensão n, considerar para K o menor inteiro tal que $2^K \geq n$. Uma expressão equivalente para obter K, consiste em considerar K=*INT*(LOG(n;2))+1 ou K=*ROUNDUP*(LOG(n;2);0), em que a função *ROUNDUP*(x;m), devolve um valor de x, arredondado por excesso, com m casas decimais;
3. Construir as classes $C_1$, $C_2$, ..., $C_k$. Vamos considerar como classes os intervalos [mínimo, mínimo + h[, [mínimo + h, mínimo + 2h[, ..., [mínimo + (k-1)h, mínimo + kh[. Uma alternativa a este procedimento seria considerar as classes abertas à esquerda e fechadas à direita, da seguinte forma: ]max – Kh, max – (K-1)h], ]max – (K-1)h, max – (K-2)h], ]max – h, max].

O resultado destes passos são representados na figura seguinte, para a amostra das idades:

IdadeTrabalhadores.xls

| | T | U | V | W | X | Y |
|---|---|---|---|---|---|---|
| 1 | | | | | Classes | |
| 2 | | | | | Lim. Inf. | Lim. Sup. |
| 3 | Mínimo | =MIN(A1:R10) | | $C_1$ | =U3 | =X3+$U$9 |
| 4 | Máximo | =MAX(A1:R10) | | $C_2$ | =Y3 | =X4+$U$9 |
| 5 | Amplitude | =U4-U3 | | $C_3$ | =Y4 | =X5+$U$9 |
| 6 | n | =COUNT(A1:R10) | | $C_4$ | =Y5 | =X6+$U$9 |
| 7 | k | =INT(LOG(U6;2))+1 | | $C_5$ | =Y6 | =X7+$U$9 |
| 8 | Amplitude/k | =U5/U7 | | $C_6$ | =Y7 | =X8+$U$9 |
| 9 | h | =ROUNDUP(U8;1) | | $C_7$ | =Y8 | =X9+$U$9 |
| 10 | | | | $C_8$ | =Y9 | =X10+$U$9 |

Sheet1 / Sheet2 / Shee

IdadeTrabalhadores.xls

| | T | U | W | X | Y |
|---|---|---|---|---|---|
| 1 | | | | Classes | |
| 2 | | | | Lim. Inf | Lim. Sup |
| 3 | Mínimo | 18 | $C_1$ | 18 | 23,8 |
| 4 | Máximo | 64 | $C_2$ | 23,8 | 29,6 |
| 5 | Amplitude | 46 | $C_3$ | 29,6 | 35,4 |
| 6 | n | 180 | $C_4$ | 35,4 | 41,2 |
| 7 | k | 8 | $C_5$ | 41,2 | 47 |
| 8 | Amplitude/k | 5,75 | $C_6$ | 47 | 52,8 |
| 9 | h | 5,8 | $C_7$ | 52,8 | 58,6 |
| 10 | | | $C_8$ | 58,6 | 64,4 |

Sheet1

Para calcular as frequências absolutas das classes obtidas anteriormente utilizámos a função COUNTIF, como se exemplifica a seguir para a classe $C_5$:

=COUNTIF($A$1:$R$10;"<47")-COUNTIF($A$1:$R$10;"<41,2")

IdadeTrabalhadores.xls

| | S | T | U | V | W | X | Y | Z | AA |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | Classes | | | |
| 2 | | | | | | Limite inferior | Limite superior | Freq.abs. | Freq.rel. |
| 3 | | Mínimo | 18 | | $C_1$ | 18 | 23,8 | 16 | 0,089 |
| 4 | | Máximo | 64 | | $C_2$ | 23,8 | 29,6 | 10 | 0,056 |
| 5 | | Amplitude | 46 | | $C_3$ | 29,6 | 35,4 | 23 | 0,128 |
| 6 | | n | 180 | | $C_4$ | 35,4 | 41,2 | 19 | 0,106 |
| 7 | | k | 8 | | $C_5$ | 41,2 | 47 | 28 | 0,156 |
| 8 | | Amplitude/k | 5,75 | | $C_6$ | 47 | 52,8 | 43 | 0,239 |
| 9 | | h | 5,8 | | $C_7$ | 52,8 | 58,6 | 30 | 0,167 |
| 10 | | | | | $C_8$ | 58,6 | 64,4 | 11 | 0,061 |
| 11 | | | | | | | | 180 | 1,000 |

Sheet1 / Sheet2 / Sheet3

Nota: Para obter o valor de h, utilizámos a função ROUNDUP(x;m), referida anteriormente. Neste caso utilizámos a função ROUNDUP(U8;1). Chamamos, no entanto, a atenção, para que nem sempre este arredondamento produz o resultado desejado, já que o valor arredondado por excesso pode vir igual ao valor que se pretende arredondar. Por exemplo, se no caso presente o resultado da célula U8 fosse 5,8, então a função ROUNDUP(U8;1) devolveria o valor 5,8. Assim, é necessário estar atento para eventualmente se proceder a um arredondamento manual.

### 2.2 –Utilização da função *Frequency*

O Excel tem uma função, que é a função *Frequency(Data_array;Bins_array)*, que calcula o número de elementos da variável - cujos valores se encontram na **Data_array**, existentes nas classes - cujos limites se encontram em **Bins_array**. Este vector **Bins_array** é constituído por um conjunto de k valores $b_1$, $b_2$, ..., $b_k$, formando (k+1) classes, tais que:

- A 1ª classe é dada por $(-\infty, b_1]$, isto é, conterá todos os elementos $\leq b_1$;
- A 2ª classe é dada por$]b_1, b_2]$;
- A 3ª classe é dada por $]b_2, b_3]$;
- A késima classe é dada por $]b_{k-1}, b_k]$;
- A (k+1)ésima classe é dada por $]b_k, +\infty)$;

Vamos exemplificar construindo uma tabela de frequências para a variável idade, assumindo como separadores (*bins*) os valores 23,8; 29,6; 35;4; 41,2; 47; 52,8 e 58,6 considerados em 2.1:

Para utilizar a função *Frequency(Data_array;Bins_array)*, procede-se do seguinte modo:

- Definir a coluna de separadores ou limites das classes, que constituirá o **Bins_array**; no nosso caso será {23,8, 29,6, 35,4, 41,2, 47, 52,8, 58,6}
- Seleccionar tantas células em coluna, quantas as classes consideradas para a tabela de frequências (não esquecer que o número de classes é superior em uma unidade ao número de separadores, pelo que o número de células seleccionadas deverá ser, neste caso, de 8);
- Introduzir a função *Frequency*, considerando como primeiro argumento o conjunto de células onde se encontram os dados a agrupar, chamado de **Data_array**, e como segundo argumento as células que constituem o **Bins_array**;
- Carregar CTRL+SHIFT+ENTER

IdadeTrabalhadores.xls

| | A | E | F | G | H |
|---|---|---|---|---|---|
| 12 | | | Classes | | |
| 13 | | | Lim. Inf. | Lim. Sup | Freq.abs. |
| 14 | 23,8 | | 18 | 23,8 | =FREQUENCY(A1:R10;A14:A20) |
| 15 | 29,6 | | 23,8 | 29,6 | =FREQUENCY(A1:R10;A14:A20) |
| 16 | 35,4 | | 29,6 | 35,4 | =FREQUENCY(A1:R10;A14:A20) |
| 17 | 41,2 | | 35,4 | 41,2 | =FREQUENCY(A1:R10;A14:A20) |
| 18 | 47 | | 41,2 | 47 | =FREQUENCY(A1:R10;A14:A20) |
| 19 | 52,8 | | 47 | 52,8 | =FREQUENCY(A1:R10;A14:A20) |
| 20 | 58,6 | | 52,8 | 58,6 | =FREQUENCY(A1:R10;A14:A20) |
| 21 | | | 58,6 | 64,4 | =FREQUENCY(A1:R10;A14:A20) |

Sheet1 / Sheet2 / She

IdadeTrabalhadores.xls

| | A | D | E | F | G | H |
|---|---|---|---|---|---|---|
| 12 | | | | Classes | | |
| 13 | | | | Lim. Inf. | Lim. Sup | Freq.abs. |
| 14 | 23,8 | | $C_1$ | 18 | 23,8 | 16 |
| 15 | 29,6 | | $C_2$ | 23,8 | 29,6 | 10 |
| 16 | 35,4 | | $C_3$ | 29,6 | 35,4 | 23 |
| 17 | 41,2 | | $C_4$ | 35,4 | 41,2 | 19 |
| 18 | 47,0 | | $C_5$ | 41,2 | 47 | 34 |
| 19 | 52,8 | | $C_6$ | 47 | 52,8 | 37 |
| 20 | 58,6 | | $C_7$ | 52,8 | 58,6 | 30 |
| 21 | | | $C_8$ | 58,6 | 64,4 | 11 |

Sheet1 / She

Repare-se que a frequência absoluta da classe $C_5$ vem diferente da obtida em 2.1, pois as classes agora são fechadas à direita e abertas à esquerda, pelo que os elementos iguais a 47 pertencem a esta classe, ao contrário do que acontecia em 2.1, em que pertenciam à classe $C_6$. Assim, se se pretende utilizar a função Frequency, a metodologia para formar as classes deve ser a utilizada em 2.1, com a alternativa considerada no passo 4, isto é:

4. Considerar as classes abertas à esquerda e fechadas à direita, da seguinte forma: ]max – Kh, max – (K-1)h], ]max – (K-1)h, max – (K-2)h], ]max – h, max].

**2.3 - Utilização da *PivotTable***

Outro processo que pode ser utilizado para fazer o agrupamento de uma variável de tipo contínuo é utilizando a PivotTable. Vamos distinguir algumas situações, pois o método de construção das tabelas de frequência sofre algumas alterações, para as quais devemos estar alertas, conforme o tipo de dados a tratar. O processo que vamos utilizar foi sugerido, em parte, por um artigo de Neville Hunt, na revista Teaching Statistics (Volume 25, Number 2, Summer 2003).

**2.3.1 – Dados em formato de inteiro e amplitude de classe também de tipo inteiro**

**Exemplo** – Considere ainda o ficheiro IdadeTrabalhadores.xls utilizado anteriormente e proceda ao agrupamento dos dados utilizando a PivotTable.

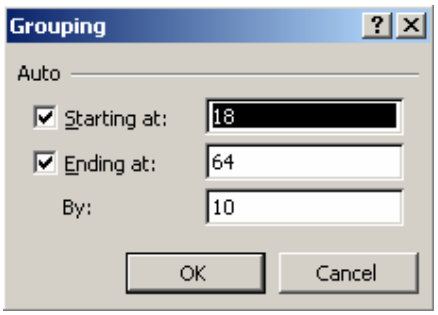
Antes de processarmos os passos associados à construção da tabela, é necessário dispormos os dados numa única coluna, em que na 1ª linha dessa coluna deve estar o nome da variável. Optámos por inserir os dados com o título Idade, nas células A1:A181 da Sheet2 do nosso ficheiro.

Procedimento a seguir:

1. No menu Data, clique em *PivotTable and PivotChart Report*;
2. No passo 1 da *PivotTable and PivotTable Wizard*, siga as instruções, e clique *PivotTable* à pergunta *What kind of report do you want to create*?;
3. No passo 2 siga as instruções, seleccionando os dados que pretende usar. Neste caso seleccione as células A1:A181 (embora os dados estejam nas células A2:A181, o título está na célula A1);
4. No passo 3 seleccione o lugar onde pretende criar a tabela. Nós optámos por seleccionar a célula C2;
5. Arraste o botão Idade da barra *PivotTable*, e coloque-o (drop it) no campo *Row*; Arraste o mesmo botão e coloque-o (drop it) no campo *Data*;
6. Clique duas vezes no botão *Sum of Idade*, da tabela, e seleccione *Count*;

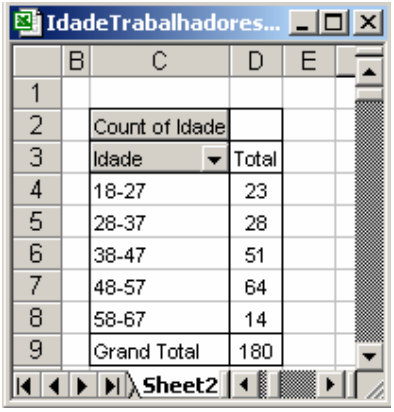
A tabela que aparece depois destas operações, mostra a frequência de cada valor individual (como estamos com dados contínuos, embora inteiros, corremos o risco de termos uma tabela com tantas classes, quantos os dados, todos com frequência igual a 1!). Assim, é necessário proceder a mais algumas operações, para agrupar os dados:

7. Clique em algum dos dados da variável Idade e seleccione *Data → Group and Outline → Group (*ou então clique em algum dos dados com o botão direito do rato e seleccione *Group and Outline → Group*) , fazendo surgir o seguinte diálogo:




Por defeito, no diálogo anterior é considerado como "*Starting at*" e "*Ending at*" respectivamente, o mínimo e o máximo do conjunto de dados a agrupar. Para "*By*" é considerado, também por defeito, um valor que dependerá do número de dados e da grandeza desses dados.

8. Clicando em OK, é produzida a seguinte tabela de frequências:

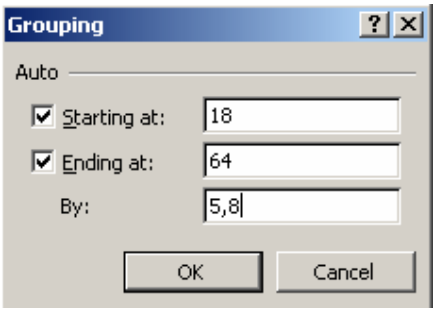


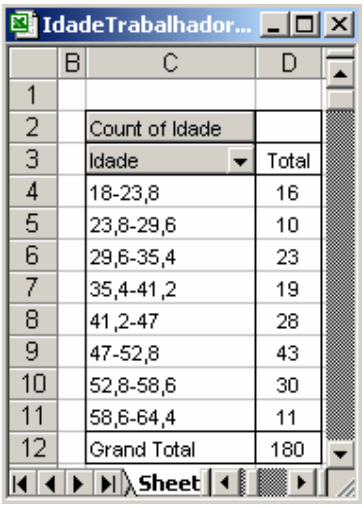
| | B | C | D | E |
|---|---|---|---|---|
| 1 | | | | |
| 2 | | Count of Idade | | |
| 3 | | Idade | Total | |
| 4 | | 18-27 | 23 | |
| 5 | | 28-37 | 28 | |
| 6 | | 38-47 | 51 | |
| 7 | | 48-57 | 64 | |
| 8 | | 58-67 | 14 | |
| 9 | | Grand Total | 180 | |

Observação: Repare-se que na construção automática desta tabela, as classes estão construídas de tal modo que são equivalentes às classes [18, 28[, [28, 38[, [38, 48[, [48, 58[, [58, 68[ (não esqueçamos que estamos a trabalhar com números inteiros e, neste caso, a amplitude de classe também é um número inteiro).

**2.2.2** – **Dados em formato de inteiro e amplitude de classe de tipo não inteiro**

Suponhamos agora que no passo 7, escolhíamos para amplitude de classe o valor 5.8, sugerido em 2.1:



| | B | C | D |
|---|---|---|---|
| 1 | | | |
| 2 | | Count of Idade | |
| 3 | | Idade | Total |
| 4 | | 18-23,8 | 16 |
| 5 | | 23,8-29,6 | 10 |
| 6 | | 29,6-35,4 | 23 |
| 7 | | 35,4-41,2 | 19 |
| 8 | | 41,2-47 | 28 |
| 9 | | 47-52,8 | 43 |
| 10 | | 52,8-58,6 | 30 |
| 11 | | 58,6-64,4 | 11 |
| 12 | | Grand Total | 180 |

Como se verifica, ao contrário do que acontecia quando a amplitude de classe era um inteiro, o limite superior de um intervalo é igual ao limite inferior do intervalo seguinte, ficando a dúvida de saber em que classe inserir um elemento igual a um desses limites (esta situação não de põe neste caso, uma vez que os dados são inteiros). Na verdade estes intervalos funcionam como se fossem fechados à esquerda e abertos à direita (excepto a última classe que também é fechada à direita).

**2.2.3** – **Dados em formato decimal**

Quando os dados são apresentados com casas decimais, a situação á idêntica à anterior. A aparente ambiguidade, de à primeira vista, não se saber a que classe pertence um valor igual a um limite de classe, pode ser resolvida, considerando para amplitude de classe um valor decimal, com uma casa decimal a mais dos que os dados.

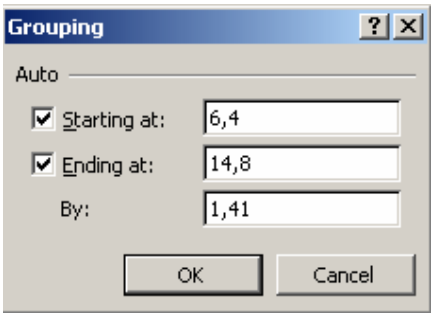
**Exemplo** – Na publicação do INE, *Anuário Estatístico da Região de Lisboa e Vale do Tejo 2002*, verifica-se que a taxa de natalidade para os diferentes concelhos desta região é a apresentada no seguinte ficheiro:

Taxanatalidade.xls

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **Portugal** | **10,9** | | | | | | | | |
| 2 | **Lisboa e Vale Tejo** | **11,4** | | | | | | | | |
| 3 | **Oeste** | **10,7** | **G. Lisboa** | **11,9** | **P. Setúbal** | **11,9** | **Médio Tejo** | **9,3** | **Lezíria Tejo** | **10,0** |
| 4 | Alcobaça | 9,4 | Amadora | 11,6 | Alcochete | 11,5 | Abrantes | 8,4 | Almeirim | 11,4 |
| 5 | Alenquer | 11,7 | Cascais | 12,8 | Almada | 12,0 | Alcanena | 8,5 | Alpiarça | 9,1 |
| 6 | Arruda dos Vinhos | 9,3 | Lisboa | 9,9 | Barreiro | 9,4 | Constância | 9,4 | Azambuja | 8,3 |
| 7 | Bombarral | 9,2 | Loures | 11,7 | Moita | 11,6 | Entroncamento | 13,1 | Benavente | 13,6 |
| 8 | Cadaval | 10,1 | Odivelas | 11,2 | Montijo | 11,3 | Fer. do Zêzere | 6,4 | Cartaxo | 10,8 |
| 9 | Caldas da Rainha | 10,2 | Oeiras | 11,6 | Palmela | 11,6 | Ourém | 11,1 | Chamusca | 6,4 |
| 10 | Lourinhã | 10,1 | Sintra | 14,8 | Seixal | 13,0 | Sardoal | 6,4 | Coruche | 7,6 |
| 11 | Mafra | 12,3 | V. F. Xira | 13,0 | Sesimbra | 12,5 | Tomar | 8,1 | Golegã | 9,3 |
| 12 | Nazaré | 11,7 | | | Setúbal | 12,1 | Torres Novas | 8,9 | Rio Maior | 10,6 |
| 13 | Óbidos | 11,0 | | | | | V. N. Barquinha | 9,5 | Salvat. Magos | 9,1 |
| 14 | Peniche | 10,1 | | | | | | | Santarém | 10,2 |
| 15 | Sobral Monte Agraç | 12,6 | | | | | | | | |
| 16 | Torres Vedras | 10,8 | | | | | | | | |
| 17 | **Fontes**: Informação calculada com base em: INE, Estatísticas Demográficas; INE, Estimativas Provisórias de População Residente, aferidas dos resultados provisórios dos Censos 2001, ajustados com as taxas de cobertura. | | | | | | | | | |

Sheet1 / Chart1 / Sheet2 / St

Procedemos a um agrupamento conveniente para os dados, utilizando a metodologia apresentada anteriormente, tendo obtido a tabela de frequências (absolutas) que se apresenta a seguir:




Taxanatalidade.xls

| | C | D | E | F | G | H |
|---|---|---|---|---|---|---|
| 1 | | | | | | |
| 2 | | Mínimo | 6,4 | | Count of Taxa | |
| 3 | | Máximo | 14,8 | | Taxa | Total |
| 4 | | Amplitude | 8,4 | | 6,4-7,81 | 4 |
| 5 | | n | 51,0 | | 7,81-9,22 | 8 |
| 6 | | k | 6 | | 9,22-10,63 | 13 |
| 7 | | Amplitude/k | 1,4 | | 10,63-12,04 | 16 |
| 8 | | h | 1,41 | | 12,04-13,45 | 8 |
| 9 | | | | | 13,45-14,86 | 2 |
| 10 | | | | | Grand Total | 51 |

Sheet1 / Sheet2

## 2.3 - Representação gráfica de dados

### 2.3.1 - Variáveis discretas. Diagrama de barras

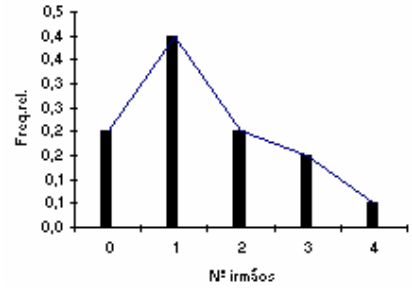
Vimos que, no caso de dados discretos, a construção da tabela de frequências se resume, de um modo geral, a considerar como classes os diferentes valores que surgem na amostra. Uma representação gráfica adequada para estes dados, é o diagrama de barras.

**Diagrama de barras** - Representação gráfica que consiste em marcar num sistema de eixos coordenados, no eixo dos xx, o valor das classes e nesses pontos barras verticais de altura igual à frequência absoluta ou à frequência relativa.

Algumas considerações sobre a metodologia a seguir na construção do diagrama de barras:

- Ordenar a amostra e considerar para classes os diferentes valores aí considerados. Marcar essas classes no eixo dos xx, num sistema de eixos coordenados.
- Nos pontos onde se consideraram as classes, marcar barras de altura igual à frequência absoluta ou relativa, da respectiva classe. De preferência utilizar as frequências relativas, pois se pretendermos comparar diagramas de barras de amostras diferentes, temos a garantia de que a soma das barras em qualquer dos diagramas é igual a 1.

**Exemplo 3** (cont) - O diagrama de barras que representa a distribuição das frequências do nº de irmãos dos alunos da turma considerada, tem o seguinte aspecto:



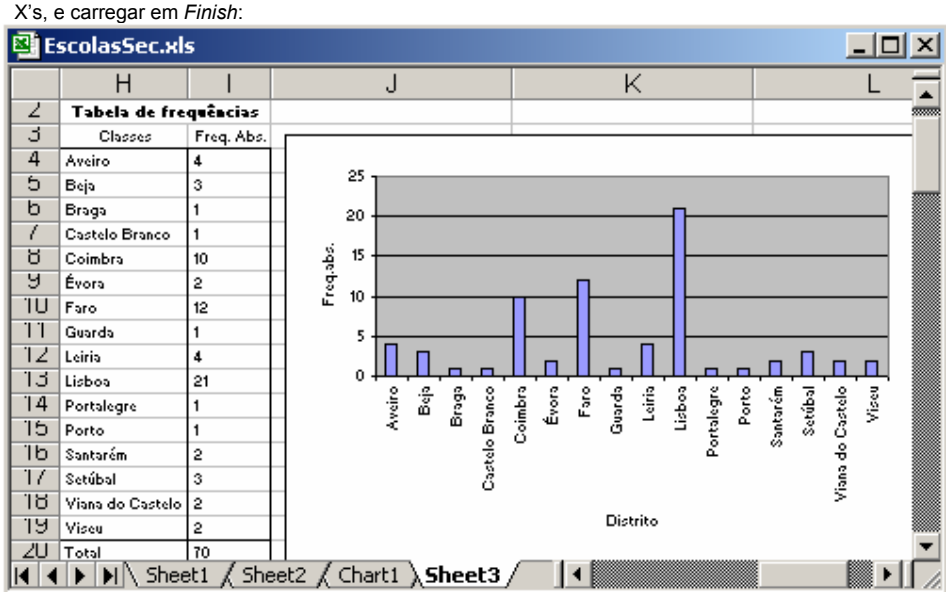

A linha poligonal que une os extremos das barras, chama-se **polígono de frequências**.

## Utilização do Excel na construção de diagramas de barras

**Variável de tipo qualitativo**

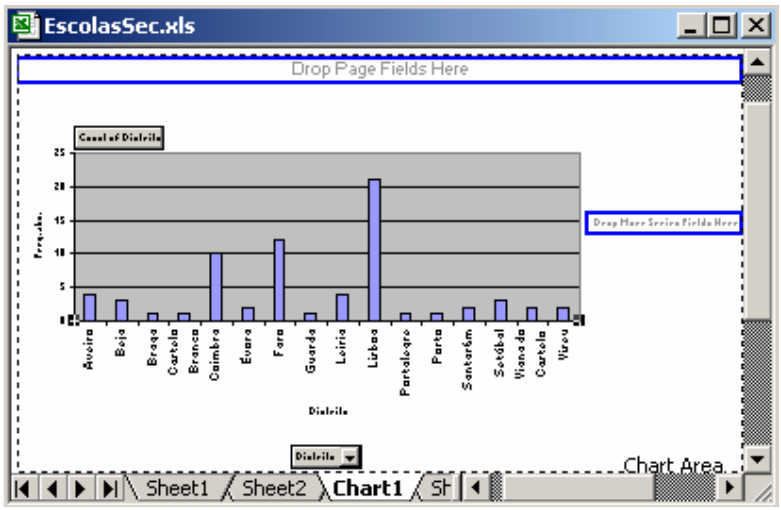
Considere a tabela de frequências obtida, na secção anterior, para os dados do ficheiro EscolasSec.xls e variável Distrito, e construa o diagrama de barras associado.A metodologia seguida para construir o diagrama de barras, consiste em, na folha Excel, que contém a tabela:

- Seleccionar as células que contêm as classes e as frequências absolutas (por exemplo);
- Seleccionar, no menu, o ícone Chart;
- Na caixa de diálogo que aparece, seleccionar a opção *Column*;
- Clicar no botão *Next*, duas vezes, para passar dois passos, até aparecer uma caixa de diálogo, que apresenta várias opções: Em *Legend*, desactivar a legenda e em *Titles*, acrescentar o título no eixo dos Y's e no eixo dos X's, e carregar em *Finish*:

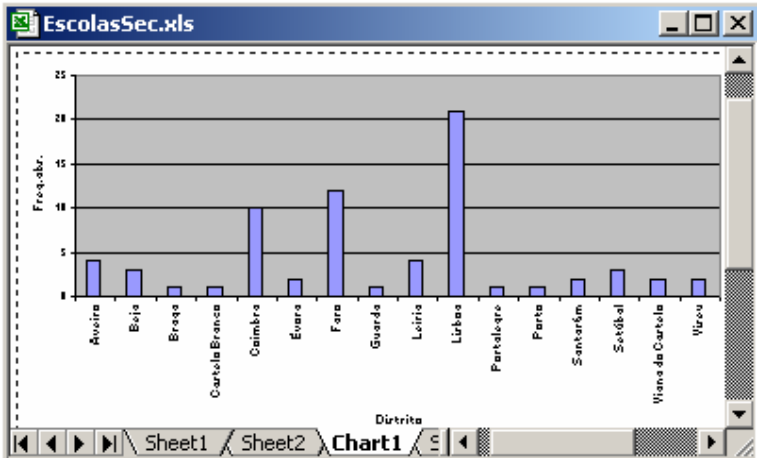



No entanto, se a tabela de frequências tiver sido construída utilizando a metodologia das PivotTables, o procedimento a seguir é o seguinte, como exemplificamos com a tabela obtida ainda para o mesmo ficheiro:

- Clicar em alguma parte da tabela e na barra da *PivotTable* clicar no ícone :

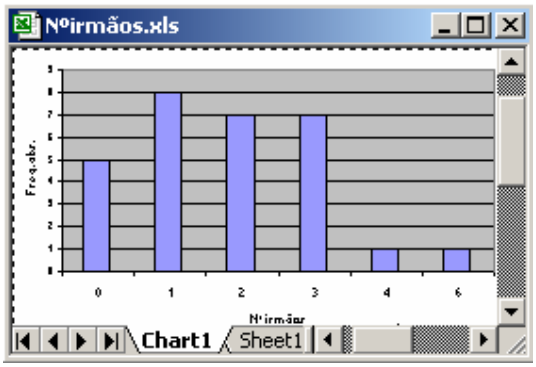



- Surge-nos um gráfico idêntico ao anterior. Do mesmo modo que anteriormente acrescentámos títulos, pelo que só falta esconder os botões, o que se faz clicando com o lado direito do rato num deles e seleccionando *Hide PivotChart Field Buttons*:



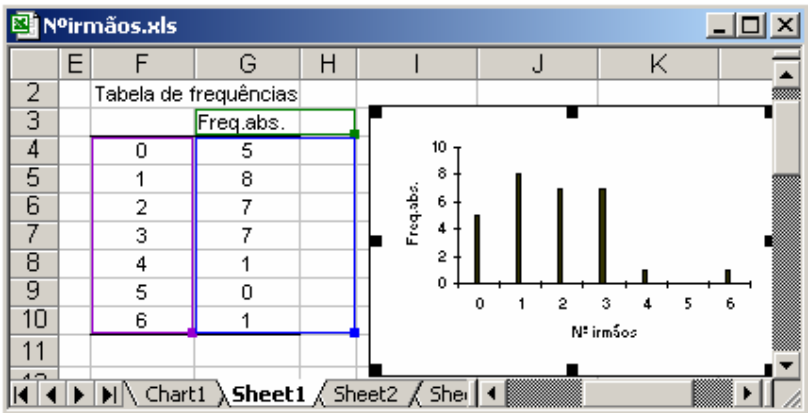

**Variável de tipo quantitativo discreto**

**Exemplo** – Consideremos de novo o exemplo da página 38, em que temos os dados do número de irmãos de 25 alunos. Repare que na amostra seleccionada não existe nenhum aluno com 5 irmãos, pelo que a tabela de frequências não inclui a classe 5. Se utilizar o procedimento anterior para obter o diagrama de barras, obtém-se a seguinte representação gráfica:



Repare-se que a variável Nº irmãos está a ser considerada como qualitativa, pois para termos uma representação gráfica correcta, deveria aparecer o valor 5, embora com uma frequência nula.

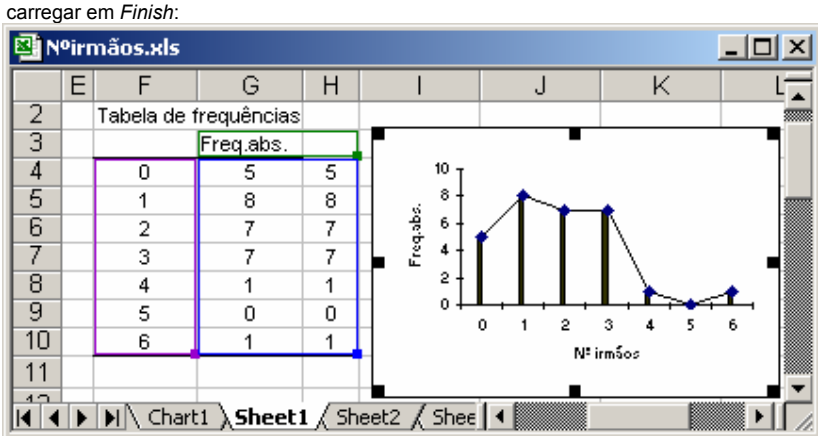
Tendo em consideração a observação feita no seguimento desse exemplo, incluímos na tabela a classe 5 com frequência nula (para isso foi necessário copiar os valores da tabela para outras células, já que não se podem inserir células em tabelas obtidas pelo processo das PivotTables) e procedemos do seguinte modo para obter o diagrama de barras:
Seleccionar as células F3:G10, depois de ter apagado a palavra Classes da célula F3 e proceder como no caso dos dados qualitativos, obtendo-se:




**Gráficos combinados**

Se pretender visiualizar juntamente com o diagrama de barras, o polígono de frequências, basta juntar uma nova coluna com as frequências e proceder do seguinte modo:

- Seleccionar as células que contêm as classes e as frequências que se pretendem representar no gráfico combinado ;
- Seleccionar, no menu, o ícone Chart ;
- Seleccionar Custom Types → Line-Column → Next → Next → Aparece uma caixa de diálogo, que apresenta várias opções: Em *Legend*, desactivar a legenda e em *Titles*, acrescentar o título no eixo dos Y's e no eixo dos X's, e carregar em *Finish*:




## 2.3.2 - Variáveis contínuas. Histograma.

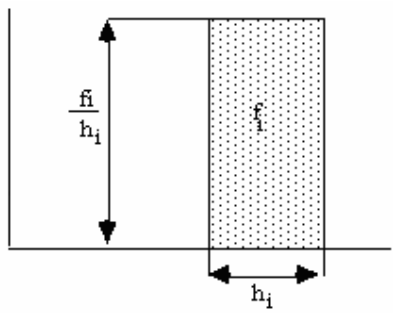
Já vimos anteriormente as etapas que, de um modo geral, se seguem para obter a tabela de frequências de uma amostra de dados contínuos. Ao contrário do caso anterior, agora as classes já não são pontos isolados, mas intervalos. Assim, a representação gráfica já não pode ser o

diagrama de barras, pois não existem pontos isolados, onde elas seriam colocadas. Vejamos então como construir a representação gráfica adequada a que damos o nome de **histograma**.

**Histograma** - Para a representação gráfica de dados contínuos usa-se um *diagrama de áreas* ou histograma, formado por uma sucessão de rectângulos adjacentes, tendo cada um por base um intervalo de classe e por área a frequência relativa (ou a frequência absoluta). Deste modo, a área total coberta pelo histograma é igual a 1 (respectivamente igual a n, a dimensão da amostra).

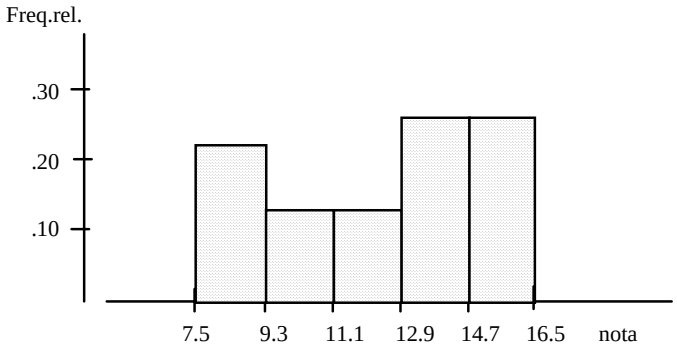
Para construir o histograma, quais as alturas que se devem considerar para os rectângulos?

Se se pretende que a área do rectângulo correspondente à classe $C_i$, seja $f_i$ ou $n_i$, respectivamente a frequência relativa ou absoluta, então a altura desse rectângulo deverá ser $f_i/h_i$ ou $n_i/h_i$, onde $h_i$ representa a amplitude da classe $C_i$.




Se todas as classes tiverem a mesma amplitude, então $h_i = h$. Neste caso facilita-se a construção do histograma, considerando para alturas dos rectângulos as frequências relativas, não esquecendo que a área total ocupada pelo histograma será igual a h e não igual a 1! Efectivamente a área de cada rectângulo é proporcional, e não igual, à frequência relativa da respectiva classe, sendo a constante de proporcionalidade h.

**Exemplo 5** (cont) - No caso da amostra de notas considerada no exemplo 5, o histograma tem o seguinte aspecto

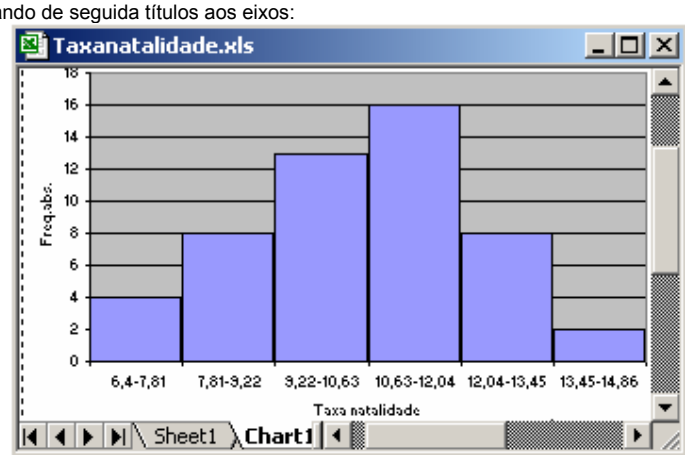


## Utilização do Excel na construção de histogramas

**1. Tabela de frequências obtida a partir da PivotTable**

Voltemos à tabela obtida na página 43, sobre os dados da taxa de natalidade dos concelhos da região de Lisboa e Vale do Tejo 2002. Tendo esta tabela sido obtida pela metodologia das PivotTables, para construior o histograma associado:

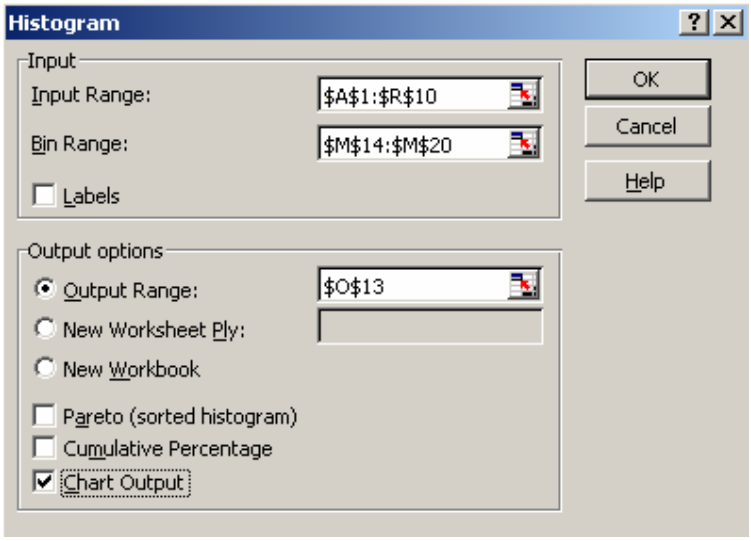
- Clicar em alguma parte da tabela e na barra da *PivotTable* clicar no ícone ;
- Clicar com o lado direito do rato numa das colunas do diagrama de barras que se obtém no passo anterior, e seleccionar *Format data Series →Options → Gap width:0;*
- Esconder os botões clicando com o lado direito do rato num deles e seleccionando *Hide PivotChart Field Buttons* e acrescentando de seguida títulos aos eixos:



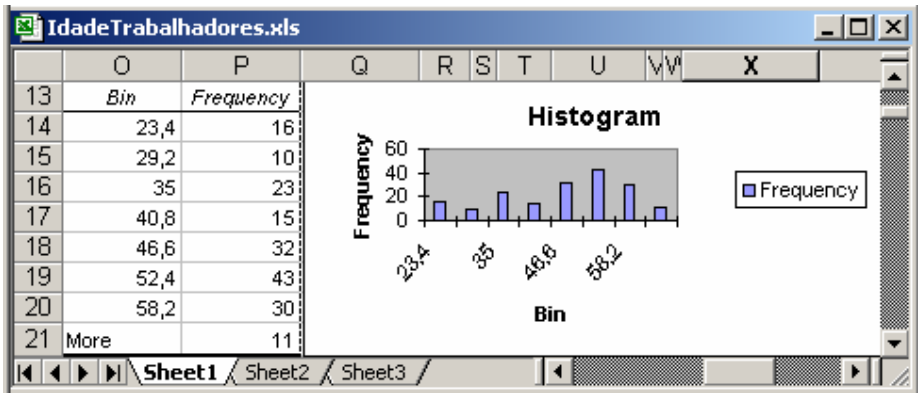

**2. Função Histogram**

No Excel existe uma função, idêntica à função *Frequency*, a que se acede seleccionando *Tools→ Data Analysis→Histogram →OK* (se o comando *Data Analysis* não constar do menu, seleccione *Tools* e na opção *Add-Ins*, seleccione *Analysis ToolPack*). Vamos, para os dados do ficheiro IdadeTrabalhadores.xls, exemplificar a sua utilização.

- Definir a coluna de separadores ou limites das classes, que constituirá o **Bin Range**: no nosso caso construímos as classes subtraindo a amplitude de classe de 5,8, sucessivamente ao máximo, obtendo os valores {23,4, 29,2, 35,0, 40,8, 46,6, 52,4, 58,2}, que colocámos nas células M14:M20;
- Seleccionar *Tools→ Data Analysis→Histogram →OK*:

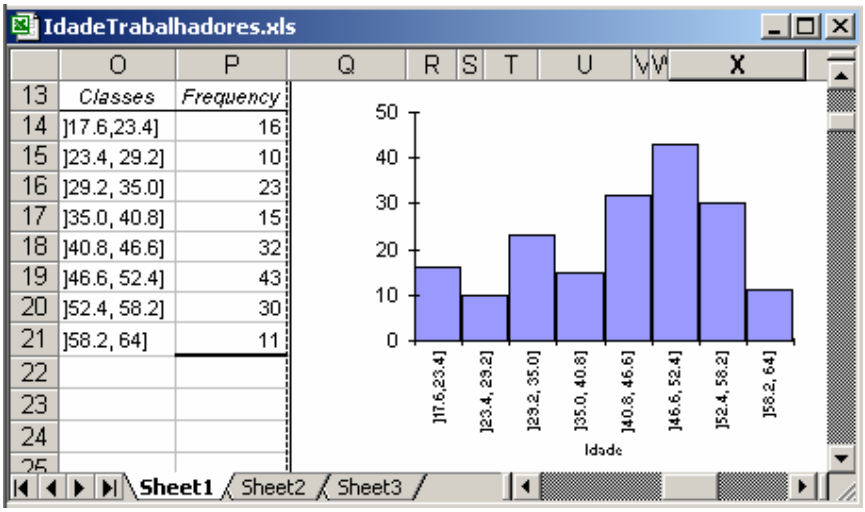


- Em Input Range, indicámos o local dos dados e seleccionámos ainda a opção Chart Output e clicámos OK. Como resultado obtivémos o seguinte:




| | Bin | Frequency |
|---|---|---|
| 14 | 23,4 | 16 |
| 15 | 29,2 | 10 |
| 16 | 35 | 23 |
| 17 | 40,8 | 15 |
| 18 | 46,6 | 32 |
| 19 | 52,4 | 43 |
| 20 | 58,2 | 30 |
| 21 | More | 11 |

- Sustituímos os limites das classes pelos intervalos das classes e arranjámos convenientemente o gráfico, já que a representação que se obtém, ao contrário do que é indicado no título, não é um histograma:



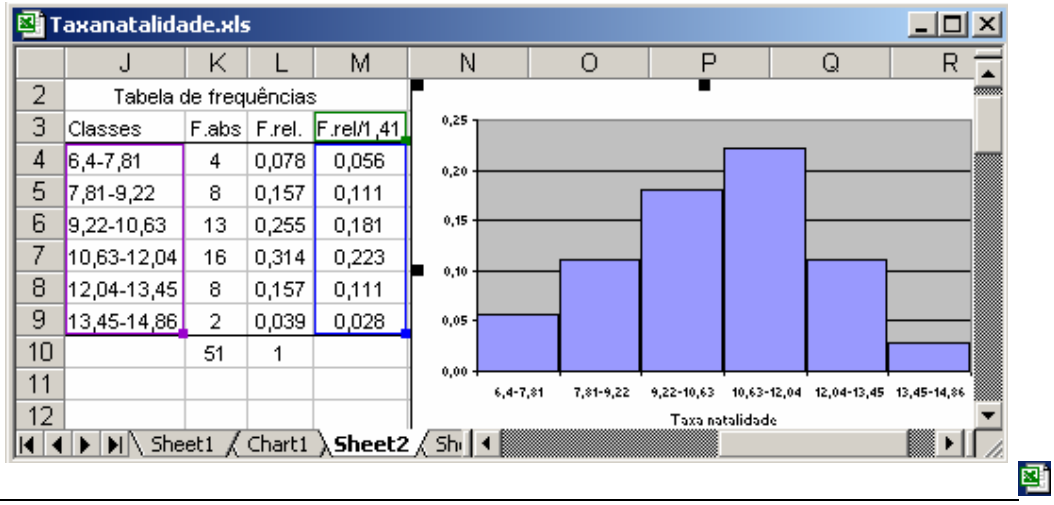

| | Classes | Frequency |
|---|---|---|
| 14 | ]17.6,23.4] | 16 |
| 15 | ]23.4, 29.2] | 10 |
| 16 | ]29.2, 35.0] | 23 |
| 17 | ]35.0, 40.8] | 15 |
| 18 | ]40.8, 46.6] | 32 |
| 19 | ]46.6, 52.4] | 43 |
| 20 | ]52.4, 58.2] | 30 |
| 21 | ]58.2, 64] | 11 |

**3. Tabela de frequências, obtida por um processo qualquer**

O histograma obtido pelo processo anterior tem como área total (amplitude de classe x dimensão da amostra), já que cada rectângulo tem por altura a frequência absoluta. Para construir um histograma cuja área total seja igual a 1, procedemos do seguinte modo:

Na tabela de frequências acrescentar (caso ainda não tenha) uma coluna com as frequências relativas e uma outra com as frequências relativas a dividir pela amplitude de classe e proceder do seguinte modo:

- Seleccionar as células J4:J9 e M4:M9 (para seleccionar células não adjacentes, basta seleccionar as células da primeira coluna e se a coluna seguinte não for adjacente, começar por carregar a tecla *CTRL* e com ela pressionada seleccionar, então, as células pretendidas;
- Proceder como foi indicado em 2.3.1, para a construção de um diagrama de barras;
- Clicar com o lado direito do rato numa das colunas do diagrama de barras que se obtém no passo anterior, e seleccionar *Format data Series →Options → Gap width:0:*




### 2.3.3 - Outras representações gráficas

Além das representações gráficas anteriormente consideradas, isto é, o diagrama de barras e o histograma, especialmente adequadas, respectivamente para dados discretos ou contínuos (embora o histograma também se possa utilizar para dados discretos), há outras representações, que passamos a descrever.

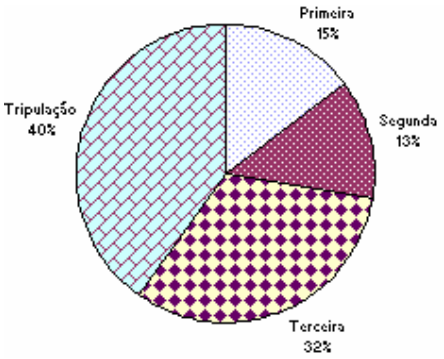
#### 2.3.3.1 - Diagrama circular

Esta representação, utilizada essencialmente para dados qualitativos, é constituída por um círculo, em que se apresentam vários sectores circulares, tantos quantas as classes consideradas na tabela de frequências da amostra em estudo. Os ângulos dos sectores são proporcionais às frequências das classes. Por exemplo uma classe com uma frequência relativa igual a .20, terá no diagrama circular um sector com um ângulo igual a $360 \times .20 = 72$ graus.

**Exemplo 6** (De Veaux et al, 2004)– Represente num diagrama circular os dados da tabela de frequências seguinte em que se apresenta a distribuição dos 2201 passageiros do Titanic segundo a variável Classe:

Tabela de frequências

| Classes | Freq.abs. | Freq.rel. |
|---|---|---|
| Primeira | 325 | 0,148 |
| Segunda | 285 | 0,129 |
| Terceira | 706 | 0,321 |
| Tripulação | 885 | 0,402 |
| | 2201 | 1 |



### 2.3.3.2 - Caule-e-folhas

É um tipo de representação que se pode considerar entre a tabela e o gráfico, uma vez que de um modo geral são apresentados os verdadeiros valores da amostra, mas numa apresentação sugestiva, que faz lembrar um histograma. Quando comparada com o histograma, é uma representação mais simples de construir quando se trabalha com lápis e papel e tem uma vantagem imediata, que é a de facilitar a ordenação dos dados, quando não se dispõe de um computador. Por outro lado, como na maior parte das vezes preserva os dígitos dos dados, ao contrário do histograma que os agrupa, permite a reconstituição da amostra.

A base da construção de uma representação em caule-e-folhas está na escolha de um par de dígitos adjacentes nos dados que vão permitir dividir cada valor do conjunto de dados em duas partes: o *caule* e a *folha*, que se dispõem para um e outro lado de um traço vertical, como exemplificamos a seguir:

**Exemplo 7** - Num determinado teste realizado a 48 estudantes, obtiveram-se as seguintes pontuações:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 75 | 98 | 42 | 75 | 84 | 87 | 65 | 59 | 63 | 86 | 78 | 37 |
| 99 | 66 | 90 | 79 | 80 | 89 | 68 | 57 | 95 | 55 | 79 | 88 |
| 76 | 60 | 77 | 49 | 92 | 83 | 71 | 78 | 53 | 81 | 77 | 58 |
| 93 | 85 | 70 | 62 | 80 | 74 | 69 | 90 | 62 | 84 | 64 | 73 |

Para fazer a representação em caule-e-folhas, consideramos o algarismo das dezenas como caule, enquanto que o algarismo das unidades será a folha. Começa-se então por traçar uma linha vertical e do lado esquerdo os caules, por ordem crescente:

1º passo

```
3 |
4 |
5 |
6 |
7 |
8 |
9 |
```

2º passo

```
3 |
4 |
5 |
6 |
7 |   5
8 |
9 |
```

3ºpasso

```
3 |  7
4 |  2 9
5 |  9 7 5 3 8
6 |  5 3 6 8 0 2 9 2 4
7 |  5 5 8 9 9 6 7  1 8 7 0 4 3
8 |  4 7 6 0 9 8 3 1 5 0 4
9 |  8 9 0 5 2 3 0
```

No 1º passo limitamo-nos a colocar os caules. Agora teremos de pendurar em cada caule as folhas respectivas. O 1º número da amostra é o 75, pelo que vamos pendurar o 5 no caule 7 (2º passo). O processo repete-se até termos esgotado todas as observações (passo 3). Finalmente é usual apresentar as folhas de cada caule ordenadas:

```
3 |  7
4 |  2 9
5 |  3 5 7 8 9
6 |  0 2 2 3 4 5 6 8 9
7 |  0 1 2 4 5 5 6 7 7 8 8 9 9
8 |  0 0 1 3 4 4 5 6 7 8 9
9 |  0 0 2 3 5 8 9
```

**Exemplo 8** - Dado o seguinte conjunto de dados, represente-os na forma de um gráfico de caule-e-folhas:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 21.3 | 27.5 | 21.4 | 28.2 | 23.5 | 28.3 | 23.8 | 23.6 | 28.4 | 28.9 |
| 24.3 | 29.1 | 24.6 | 24.8 | 29.4 | 30.0 | 24.9 | 24.9 | 31.2 | 28.9 | 24.1 |

Começamos por tomar para par de dígitos adjacentes o algarismo das unidades e o das décimas, ficando o caule constituído por dois algarismos. Considerando todos os caules possíveis, ordenam-se e dispõem-se do lado esquerdo dum traço vertical, e a partir daí começam-se a pendurar as folhas respectivas. Depois de ordenar as folhas das linhas correspondentes aos caules considerados, obtém-se a representação seguinte:

| Prof. | | n=21 (unidade=0.1) |
|---|---|---|
| 2 | 21 | 3 4 |
| | 22 | |
| 5 | 23 | 5 6 8 |
| (6) | 24 | 1 3 6 8 9 9 |
| | 25 | |
| | 26 | |
| 10 | 27 | 5 |
| 9 | 28 | 2 3 4 9 9 |
| 4 | 29 | 1 4 |
| 2 | 30 | 0 |
| 1 | 31 | 2 |

Na representação considerámos uma observação sobre as unidades com que se apresentam os dados, que no caso considerado é 0.1. Assim, ao lermos o primeiro valor no caule-e-folhas, nomeadamente o valor 213, teremos de multiplicar pela unidade, para obter o valor original. Juntámos também uma coluna com a profundidade dos dados, sendo esta noção definida a seguir.

**Profundidade de uma observação**

Dado um conjunto de dados ordenados, a qualquer uma das observações podemos associar duas ordens, contando a posição da observação a partir de cada uma das extremidades dos dados ordenados. A *profundidade* da observação é a menor daquelas ordens. Assim, juntamente com a representação de caule-e-folhas, considera-se um conjunto de profundidades em que, exceptuando a linha central, o número apresentado na coluna das profundidades é a profundidade máxima associada com os valores da linha. Na linha que contém a mediana (a definir posteriormente) é indicado, entre parêntesis, o número de folhas da linha. Voltaremos a este assunto mais à frente, quando tratarmos de um outro tipo de representação gráfica, nomeadamente a *Box-plot.*

**Qual o número de linhas (ou caules) adequado para a construção dum caule-e-folhas?**

A escolha do número de linhas, tal como acontece com o número de classes do histograma, depende em grande parte da experiência e da habilidade do estatístico. Os problemas que se levantam são análogos aos já abordados quando da construção do histograma. No entanto, dado o facto de se utilizar a notação decimal, é necessário considerar uma outra metodologia para o

comprimento do intervalo correspondente a cada linha. Assim, utiliza-se normalmente o seguinte procedimento:

Considera-se para número de linhas L um valor que não exceda

$$L = [\, 10 \ \log_{10} n \,]$$

onde n é o número de observações e [x] representa o maior inteiro que não excede x.

Na amostra considerada

$$L = [10 \log_{10} 21], \text{ ou seja } L = 13$$

Esta regra costuma fornecer valores de L convenientes para as dimensões das amostras usuais num tratamento estatístico. É evidente que, se n for muito grande, esta representação torna-se muito pesada e pouco maleável.

Usando L como limite para o número de linhas, levanta-se o problema da determinação dos comprimentos dos intervalos correspondentes a cada linha. O processo mais simples é usar uma potência de 10 como comprimento do intervalo. Assim, dividimos R, a amplitude da amostra, por L e arredondamos por excesso (se necessário) o quociente obtido, até à potência de 10 mais próxima.

Na amostra considerada

$$R = 31.4 - 21.3 = 10.1, \ \ L = 13, \ \frac{R}{L} = \frac{10.1}{13} = 0.78$$

pelo que o arredondamento à potência de 10 mais próxima dá 1.

Pode acontecer que a técnica descrita anteriormente para a construção da representação de caule-e-folhas apresente demasiadas folhas por linha. Então, o processo de resolver este problema é considerar mais linhas, repetindo os seus valores no caule. Assim, uma representação de três linhas, em que os dígitos dominantes fossem 0, 1 e 2, com demasiadas folhas por caule, transformar-se-ia em

| | |
|---|---|
| 0 | 0* |
| 1 | 0. |
| 2 | 1* |
| | 1. |
| | 2* |
| | 2. |

Enquanto que nas linhas marcadas com "*" se colocam as folhas de 0 até 4, nas linhas marcadas com "." registam-se as folhas de 5 até 9. Nesta representação o comprimento do intervalo será 5 vezes uma potência de 10 ($5 \times 10^{-1}$).

Pode acontecer que, mesmo considerando 2 linhas por caule, a representação ainda continue muito pesada, mas que se arredondássemos para a potência de 10 imediatamente abaixo do valor obtido para R/L, também ficasse muito esparsa. Então, resolve-se o problema considerando 5 linhas por caule e indicando-as da maneira que segue:

| | | |
|---|---|---|
| 0* | folhas 0 e 1 | |
| t | folhas 2 e 3 | ("two" e "three") |
| f | folhas 4 e 5 | ("four" e "five") |
| s | folhas 6 e 7 | ("six" e "seven") |
| 0. | folas 8 e 9 | |

Neste caso o comprimento do intervalo é 2 vezes uma potência de 10.

**Exemplo 9** (Hoaglin and al.1983) - Apresentamos de seguida os tempos (meses) até ao início da remissão, em doentes sujeitos a cirurgia, ao cancro do estômago. Alguns dos dados são censurados (indicados com o símbolo +) ( morte ou "perdido de vista")

1+, 1+, 1+, 2+, 2+, 3+, 4+, 4+, 5+, 8+, 9, 9, 9, 9+, 11+, 12, 14, 14+, 14+, 16, 16+, 17, 18, 19, 21+, 22, 26, 27, 28+, 29, 29, 56, 67, 68, 71

$$L = [10 \log_{10} 35] = 15$$

$$R = 70/15 = 4.67$$

pelo que o comprimento do intervalo será 10.

| | |
|---|---|
| 0 | 1 1 1 2 2 3 4 4 5 8 9 9 9 9 |
| 1 | 1 2 4 4 4 6 6 7 8 9 |
| 2 | 1 2 6 7 8 9 9 |
| 3 | |
| 4 | |
| 5 | 6 |
| 6 | 7 8 |
| 7 | 1 |

Considerando para comprimento do intervalo 5 vezes uma potência de 10, vem

| Prof | | n=35 (unidade=mês) |
|---|---|---|
| 9 | 0* | 1 1 1 2.2 3 4 4 4 |
| 15 | 0. | 5 8 9 9 9 9 |
| (4) | 1* | 1 2 4 4 |
| 16 | 1. | 6 6 7 8 9 |
| 11 | 2* | 1 2 |
| 9 | 2. | 6 7 8 9 9 |
| | 3* | |
| | 3. | |
| | 4* | |
| | 4. | |
| | 5* | |
| 4 | 5. | 6 |
| | 6* | |
| 3 | 6. | 7 8 |
| 1 | 7* | 1 |
| | 7. | |

O último exemplo sugere-nos a existência, nos dados, de alguns valores que sobressaem de entre os restantes, por serem demasiado grandes. Este é outro aspecto em que a representação de caule-e-folhas, nos ajuda a detectar esses valores (perturbadores), que chamamos de **outliers**. Veremos posteriormente uma técnica mais elaborada para detectar os outliers.

No quadro seguinte apresenta-se o número de concelhos de cada um dos distritos de Portugal Continental e das Regiões Autónomas de Açores e Madeira (Anuário Estatístico de Portugal, 1992), apresentando-se de seguida uma representação possível em caule-e-folhas:

| Região | Nº concelhos | Região | Nº concelhos |
|---|---|---|---|
| Aveiro | 19 | Lisboa | 15 |
| Beja | 14 | Portalegre | 15 |
| Bragança | 13 | Porto | 17 |
| Braga | 12 | Santarém | 21 |
| Cast.Branco | 11 | Setúbal | 13 |
| Coimbra | 17 | Viana Cast. | 10 |
| Évora | 14 | Vila Real | 14 |
| Faro | 16 | Viseu | 24 |
| Guarda | 14 | Açores | 19 |
| Leiria | 16 | Madeira | 11 |

| | |
|---|---|
| 1* | 0 1 1 |
| t | 2 3 3 |
| f | 4 4 4 4 5 5 |
| s | 6 6 7 7 |
| 1. | 9 9 |
| 2* | 1 |
| t | |
| f | 4 |

Nesta representação utilizámos 5 caules para o número 1, pendurando o 0 e o 1 no primeiro caule, o 2 e o 3 no segundo caule, etc. Procedeu-se de modo análogo com o 2. Utilizou-se esta metodologia, porque se se considerassem unicamente dois caules, obtinha-se uma representação muito pouco elucidativa.

**Utilização do caule-e-folhas para comparar duas amostras**

A representação em caule-e-folhas é muito sugestiva para comparar duas amostras, como se apresenta no exemplo seguinte:

**Exemplo 10** - A seguir apresentam-se os tempos de sono, medidos durante 30 noites seguidas, de dois jovens. Compare-os.

| Pedro | | | David | | |
|---|---|---|---|---|---|
| 8.7 | 9.3 | 8.7 | 7.1 | 9.5 | 7.1 |
| 9.4 | 5.3 | 7.4 | 8.3 | 7.1 | 7.4 |
| 6.6 | 7.3 | 6.3 | 7.1 | 7.5 | 7.4 |
| 6.0 | 6.7 | 5.9 | 7.9 | 7.9 | 7.8 |
| 6.9 | 5.8 | 10.0 | 7.5 | 6.4 | 6.2 |
| 9.9 | 4.7 | 6.5 | 6.2 | 6.2 | 8.6 |
| 6.3 | 5.6 | 8.6 | 8.2 | 7.5 | 8.4 |
| 8.9 | 5.9 | 7.7 | 8.7 | 7.7 | 6.6 |
| 10.1 | 9.4 | 9.0 | 8.5 | 7.6 | 8.1 |
| 9.6 | 7.6 | 7.9 | 7.6 | 8.8 | 7.1 |

Para representar os caule-e-folhas paralelos, determinamos os caules (comuns) a partir da amostra de maior amplitude, ou seja, neste caso, dos dados correspondentes ao David.

| Prof. | n=30 | | | | | | | | | | | | | | | n=30 | Prof. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | | | (unidade=0.1 hora) | |
| 1 | | | | | 7 | 4. | | | | | | | | | | | |
| 2 | | | | | 3 | 5* | | | | | | | | | | | |
| 6 | | 9 | 9 | 8 | 6 | 5. | | | | | | | | | | | |
| 9 | | | 3 | 3 | 0 | 6* | 2 | 2 | 2 | 4 | | | | | | | 4 |
| 13 | | 9 | 7 | 6 | 5 | 6. | 6 | | | | | | | | | | 5 |
| 15 | | | | 4 | 3 | 7* | 1 | 1 | 1 | 1 | 1 | 4 | 4 | | | | 12 |
| 15 | | | 9 | 7 | 6 | 7. | 5 | 5 | 5 | 6 | 6 | 7 | 8 | 9 | 9 | | (9) |
| | | | | | | 8* | 1 | 2 | 3 | 4 | | | | | | | 9 |
| 12 | | 9 | 7 | 7 | 6 | 8. | 5 | 6 | 7 | 8 | | | | | | | 5 |
| 8 | | 4 | 4 | 3 | 0 | 9* | | | | | | | | | | | |
| 4 | | | | 9 | 6 | 9. | 5 | | | | | | | | | | 1 |
| 2 | | | | 1 | 0 | 10* | | | | | | | | | | | |

Os dados relativamente ao Pedro encontram-se para o lado esquerdo, enquanto que os referentes ao David estão para o lado direito. A representação anterior permite realçar a maior dispersão do sono do Pedro, enquanto que o David é mais regular, com uma duração de sono de um modo geral entre as 7 e as 8 horas.

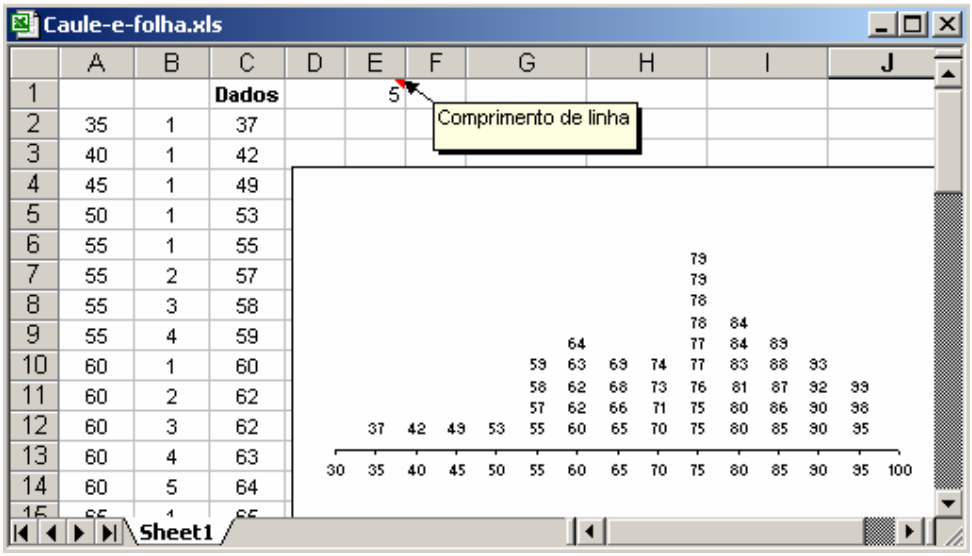
## Utilização do Excel na construção de um caule-e-folhas

Não existe no Excel uma representação imediata para a construção de um caule-e-folhas, pelo que vamos utilizar um processo desenvolvido por Neville Hunt (Hunt, 2001).

**Exemplo**: Construa uma representação em caule-e-folhas, utilizando o Excel, para os dados do exemplo 7.

Construção de um caule-e-folhas utilizando uma folha de Excel:

- 1º passo – Insira os dados na coluna C, começando na célula C2; se não estiverem ordenados, ordene-os por ordem crescente;
- 2º passo – Insira na célula E1 o valor que deseja para o comprimento de linha: uma potência de 10, de 5 ou de 2, isto é $10^m$, $5x10^m$ ou $2\ x10^m$, com m inteiro;
- 3º passo – Na célula A2 escreva a seguinte fórmula = *INT*(C2/E$1)*E$1 e replique-a tantas vezes quantos os dados inseridos no 1º passo, na coluna C;
- 4º passo – Na célula B2 escreva o valor 1. Na célula B3 escreva a fórmula = *IF* (A3=A2; B2+1; 1) e replique a fórmula, tantas vezes quantos os dados inseridos no 1º passo, na coluna C;
- 5º passo – Seleccione as células das colunas A, B e C com os resultados obtidos nos passos anteriores e no módulo *Chart Wizard* (Assistente de Gráficos) escolha *Bubble*;
- 6º passo – Faça um duplo clique numa das bolas representadas e na janela *Format data Series* (ou clique com o botão direito do rato e seleccione *Format data Series*) seleccione:
  *Patterns*
  *Border:* None
  A*rea:* None
  *Data Labels:* Show bubbles sizes
  OK;
- 7º passo – Faça um duplo clique numa das "*Data labels*" (ou clique com o botão direito do rato e seleccione *Format Data Labels*), e na janela *Format Data Labels*, em Alignment:
  *Label Position*: Centre
  OK;
- 8º passo – Clique numa das linhas horizontais que atravessam o gráfico e apague-as com a tecla Delete. Faça o mesmo ao fundo cinzento, seleccionando-o e carregando na tecla Delete. Apague também a legenda.
- 9º passo – Formate convenientemente os eixos.




Se pretender mudar o comprimento de linha para 10, basta substituir na célula E1 o 5 por 10.

**Que característica é que se pretende realçar, quando se representa um conjunto de dados sob a forma de um histograma ou de uma representação em caule-e-folhas?**

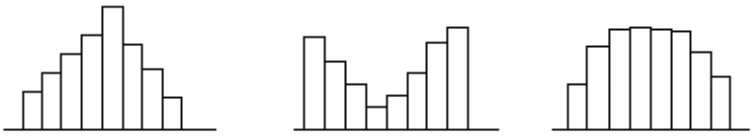
Dada uma amostra, o aspecto do histograma reflecte a forma da distribuição da População subjacente aos dados observados! Este é um dos aspectos da redução dos dados em que se perde alguma informação contida nesses dados, mas em contrapartida obtemos a estrutura da População, que eles pretendem representar.

**Quais os aspectos mais frequentes apresentados por um histograma?**

Alguns histogramas apresentam formas que, pela frequência com que surgem, merecem referência especial. Assim, as distribuições mais comuns apresentadas pelos dados são:

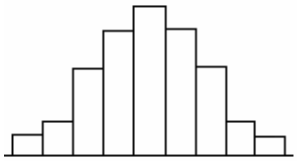
**Distribuições simétricas**

A distribuição das frequências faz-se de forma aproximadamente simétrica, relativamente a uma classe média:



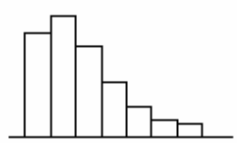
*Caso especial de uma distribuição simétrica*

Um caso especial de uma distribuição simétrica é aquele que sugere a forma de um "sino" e que é apresentado por amostras provenientes de Populações *Normais*. O significado deste termo será explicado mais tarde, no âmbito das Probabilidades.

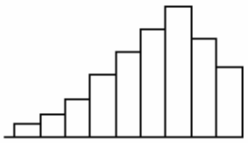


## Distribuições enviesadas

A distribuição das frequências faz-se de forma acentuadamente assimétrica, apresentando valores substancialmente mais pequenos num dos lados, relativamente ao outro:



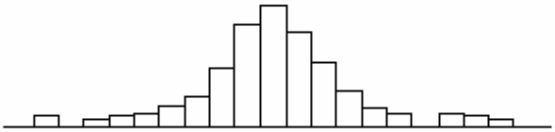
Enviesada para a direita



Enviesada para a esquerda

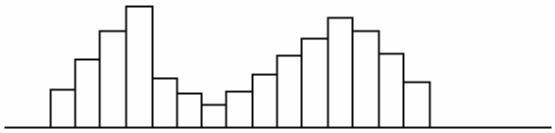
### Distribuições com caudas longas

A distribuição das frequências faz-se de tal forma que existe um grande número de classes nos extremos, cujas frequências são pequenas, relativamente às classes centrais:



### Distribuições com vários "picos" ou modas

A distribuição das frequências apresenta 2 ou mais "picos" a que chamamos modas, sugerindo que os dados são constituídos por vários grupos distintos:



### 2.3.3.3 - Função distribuição empírica

Embora de representações gráficas como um histograma ou um caule-e-folhas, se possa extrair informação relevante para a caracterização dos dados, na medida em que nos mostra a forma como se encontram concentrados, essa representação pode não ser suficiente quando se pretende outro tipo de informação, como seja a de saber qual a percentagem de valores da amostra inferiores ou superiores a um determinado valor.

Assim, quando se pretende este tipo de informação, existe uma representação gráfica conveniente que é a **função distribuição empírica**.

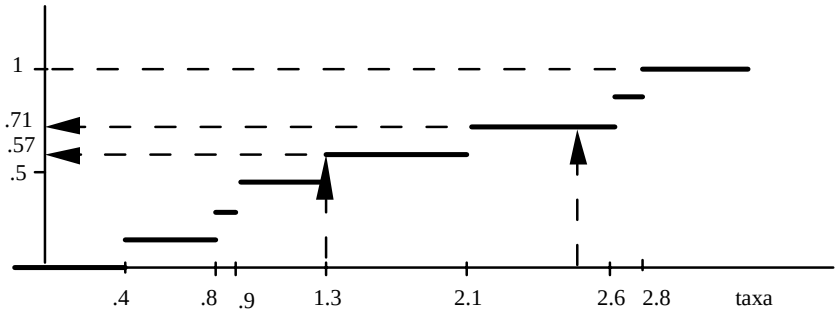
*O que é?* É uma função definida para todo o número x real e que para cada x dá a proporção de elementos da amostra menores ou iguais a x.

*Como se constrói?* Para a sua construção seguem-se as seguintes etapas:

1) Ordenar os n elementos da amostra, por ordem crescente.
2) Considerar um sistema de eixos coordenados e marcar no eixo do xx os valores da amostra.
3) Começar a desenhar a função da esquerda para a direita, atribuindo o valor 0 à esquerda do mínimo, o valor 1/n entre o mínimo e o 2º mínimo, o valor 2/n entre o 2º e o 3º mínimo, e assim sucessivamente até esgotarmos todos os valores da amostra. Para um valor igual ou superior ao máximo, a função toma o valor 1. Se na amostra um *valor se repete d vezes*, então o salto da função nesse ponto será *d/n*, em vez de 1/n.

**Exemplo 11** - Construa o gráfico de uma função distribuição empírica para os seguintes valores, que representam a taxa de crescimento populacional, nas seguintes regiões:

| | | | | | |
|---|---|---|---|---|---|
| África | 2.8 | América Latina | 2.6 | Oceânia | 1.3 |
| Ásia | 2.1 | URSS | .9 | | |
| Amér. do Norte | .8 | Europa | .4 | | |




Suponhamos que se pretendem as seguintes informações:

Qual a percentagem de taxas inferiores ou iguais a 1.3? .57

Qual a percentagem de taxas inferiores ou iguais a 2.5? .71

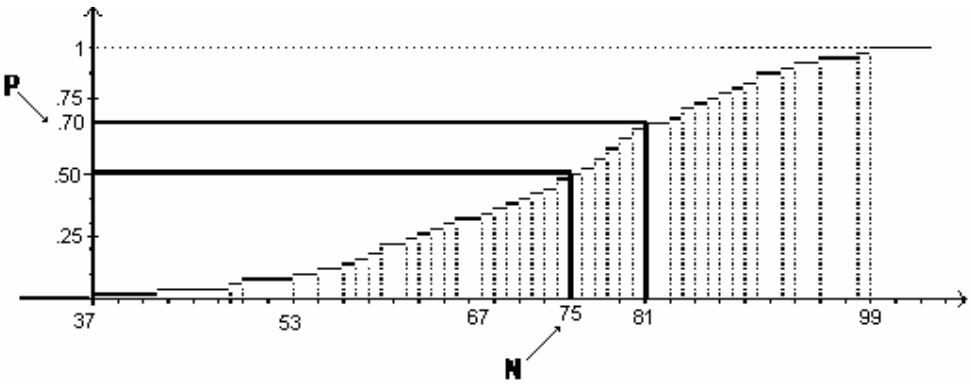
**Exemplo 12** - Num determinado teste realizado a 50 estudantes, obtiveram-se as seguintes pontuações

75 98 42 75 84 87 65 59 63 86 78 37 99 66 90 79 80 89 68 57 95 55
79 88 76 60 77 49 92 83 71 78 53 81 77 58 93 85 70 62 80 74 69 90
62 84 64 73 48 72

Depois de ordenada a amostra construa a função distribuição empírica e determine :

a) A nota N, tal que 50% dos alunos tenham nota menor ou igual a N;

b) A percentagem P de alunos com nota menor ou igual a 81.




a) A nota N é 75

b) A percentagem pedida é 70%

**Função distribuição empírica e percentis. O que são percentis ou quantis?**

Como vimos, a função distribuição empírica permite obter a percentagem, ou proporção, de elementos da amostra que são inferiores ou iguais (maiores ou iguais) a um valor qualquer. Por outro lado, dado um valor **p** qualquer, entre 0 e 1, permite determinar um valor $\mathbf{Q_p}$,tal que a amostra fica dividida em duas partes:

> **100p%** dos elementos da amostra são menores ou iguais a $\mathbf{Q_p}$ e os restantes **100(1-p)%** elementos são maiores ou iguais a $\mathbf{Q_p}$. Ao valor $\mathbf{Q_p}$ dá-se o nome de **percentil** ou **quantil** de ordem p ou percentagem 100p%.

Existem alguns quantis que, pela sua importância, merecem uma referência especial:

**Mediana**- É o percentil correspondente à percentagem de 50%, o que significa que divide a amostra em duas partes com o mesmo número de elementos. Costuma-se representar por **m**.

**Quartis** - O **1º quartil** (ou quartil inferior) é o percentil, correspondente à percentagem de 25%, o que significa que 25% dos elementos da amostra são menores ou iguais a ele e os restantes são maiores ou iguais. O **3º quartil** (ou quartil superior) é o percentil correspondente à percentagem de 75%.

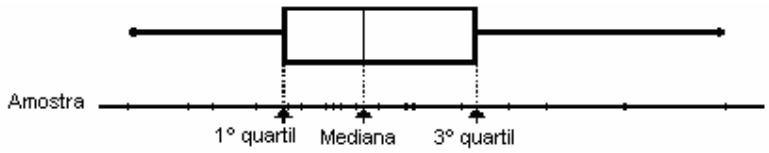
*Para calcular a mediana e os quantis, é sempre necessário construir a função distribuição empírica?*

Veremos que não! Na secção Características Amostrais, veremos um processo de calcular as características mediana e quartis sem fazer intervir a função distribuição empírica.

A seguir apresentamos um processo gráfico de representação dos dados, em que aquelas características têm papel importante.

### 2.3.3.4 – "Box-plot" ou "Box-and-whisker plot" (caixa-com-bigodes)

É um tipo de representação gráfica, em que se realçam algumas características da amostra. O conjunto dos valores da amostra compreendidos entre o 1º e o 3º *quartis*, $Q_{.25}$ e $Q_{.75}$ é representado por um rectângulo (caixa) com a *mediana* indicada por uma barra. Consideram-se seguidamente duas linhas que unem os meios dos lados dos rectângulos com o menor e maior elementos da amostra que estão dentro das **barreiras**, definidas a seguir.




**O que são barreiras?**

Define-se **barreira inferior**, como sendo o valor

$$Q_{.25} - 1.5 \times (Q_{.75} - Q_{.25})$$

Define-se **barreira superior**, como sendo o valor

$$Q_{.75} + 1.5 \times (Q_{.75} - Q_{.25})$$

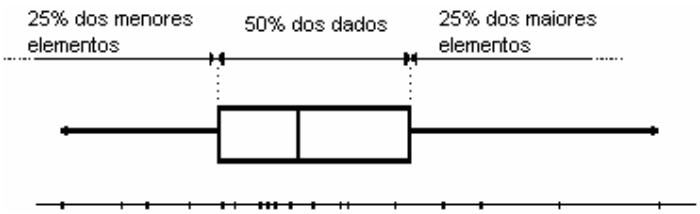
Por vezes surgem na amostra valores, que se distinguem dos restantes por serem muitos grandes ou muito pequenos. A esses valores chamamos **outliers**.

*Quando é que consideramos um valor como outlier?*

Dizemos que um valor é outlier, quando não está compreendido no intervalo [barreira inferior, barreira superior]. Numa representação em *box-plot* os outliers assinalam-se com o símbolo "*".

**Qual a importância da representação em *box-plot*?**

Realça informação importante sobre os dados, nomeadamente sobre o centro da amostra (mediana), variabilidade, simetria e a existência de outliers (valores que se distinguem dos restantes, dando a ideia de não pertencerem ao mesmo conjunto de dados). Repare-se que da forma como o diagrama se constrói, se pode retirar imediatamente a seguinte informação:




**Conjunto de letras-resumo. Outra versão da *box-plot* utilizando o resumo de 5 números**

O primeiro passo na determinação do conjunto de letras-resumo consiste na ordenação da amostra, que vamos representar por

$$x_{1:n}\ ,\ x_{2:n},\ x_{3:n},\ \ldots,\ x_{n:n}$$

onde $x_{1:n} \leq x_{2:n} \leq x_{3:n} \leq \ldots \leq x_{n:n}$ . A $x_{1:n}$ , $x_{2:n}$, $x_{3:n}$, …, $x_{n:n}$ chamamos **estatísticas ordinais** associadas à amostra e $x_{i:n}$ é a **i-ésima estatística ordinal**. Uma vez a amostra ordenada, definimos **ordem ascendente** de uma observação, como sendo a sua posição contada a partir do valor mais pequeno da amostra; **ordem descendente** será a sua posição contada a partir do maior valor. Assim, a observação correspondente à estatística ordinal $x_{i:n}$ tem ordem ascendente i e descendente n-i+1. Para qualquer observação verifica-se que

ordem ascendente + ordem descendente = n+1

Os conceitos de ordem permitem-nos definir **profundidade** (já abordado quando definimos o caule-e-folhas) de uma observação da amostra, como sendo a menor das suas duas ordens ascendente e descendente. À custa da noção de profundidade definiremos algumas estatísticas, as mais simples das quais são os **extremos**: observações cuja profundidade é 1.

Analogamente, utilizando a profundidade, se define a **mediana**, que é a estatística a que corresponde a **profundidade** $\frac{\mathbf{n+1}}{\mathbf{2}}$. Quando **n é ímpar**, corresponde à observação de profundidade $\frac{\mathbf{n+1}}{\mathbf{2}}$. Quando n é par é a **semi-soma** das observações de **profundidade** $\frac{\mathbf{n}}{\mathbf{2}}$ (pois a profundidade $\frac{n+1}{2}$ envolve, neste caso, a fracção $\frac{1}{2}$ ).

Além dos extremos e da mediana, definem-se outro par de estatísticas, as charneiras ou **quartos**, onde

$$\textbf{profundidade do quarto=}\ \frac{[\text{profundidade da mediana}]+1}{2}$$

Sempre que a profundidade do quarto envolver a fracção 1/2, procede-se a uma interpolação, como se fez para a mediana. Chamamos a atenção para o facto de os quartos não coincidirem necessariamente com os quartis (do mesmo modo que nem todos os processos para obter os quartis conduzem aos mesmos valores). Aliás, pode-se mostrar que, para n par os quartos e quartis[(2)] coincidem, enquanto que para n ímpar, só não coincidem se n for múltiplo de 4+1.

O conjunto da mediana, quartos e extremos, constituem o chamado **resumo de 5 números**.

Por vezes, e em particular quando a dimensão da amostra é elevada, é útil utilizar mais alguns números para resumir os dados, fornecendo assim mais detalhe. Então definem-se os **oitavos**, onde

$$\text{profundidade do oitavo=}\ \frac{[\text{profundidade do quarto}]+1}{2}$$

Esta metodologia pode ser continuada, de modo que se definem novas estatísticas à custa das anteriores, calculando a respectiva profundidade através da fórmula

---

[(2)] Os quartis são os quantis de ordem .25 (1º quartil) e .75 ( 3º quartil) e são casos particulares de quantis. Define-se quantil de ordem p ( 0<p<1) e representa-se por $Q_p$, como sendo o valor tal que 100p% dos valores da amostra são $\leq Q_p$ e os restantes 100(1-p)% são $\geq Q_p$. Para a determinação do quantil Qp, utiliza-se a regra seguinte:
$Q_p = x_{([np]+1)}$ se np não é inteiro e $Q_p = (x_{(np)}+x_{(np+1)})/2$ se np é inteiro. Na expressão da determinação dos quantis, a ordem i da observação $x_{(i)}$ é a ordem ascendente.

$$\frac{[\text{profundidade anterior}]+1}{2}$$

Ao conjunto dos números assim determinados para resumir a amostra, podem-se associar letras, chamadas etiquetas, sendo habitual fazê-lo da forma seguinte:

| | |
|---|---|
| M | Mediana |
| F | Quarto ("Fourth") |
| E | Oitavo ("Eight") |
| D | 16-avos |
| … | … |
| A | 128-avos |
| Z | 256-avos |
| Y | 512-avos |
| … | … |
| 1 | Extremos (profundidade) |

Esta associação entre as letras e os valores-resumo faz com que a esses valores se chamem **letras-resumo**.

Um modo de representar um conjunto de letras-resumo, de forma a termos a informação de uma forma sugestiva, é a seguinte:

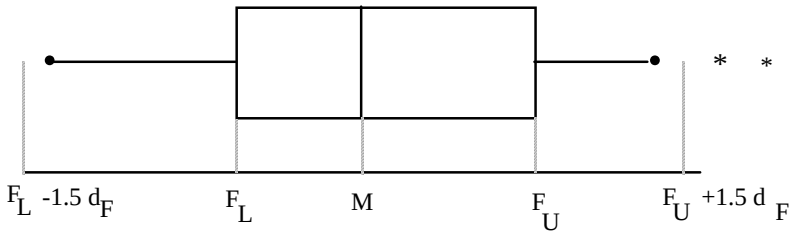
| | | | | |
|---|---|---|---|---|
| # | n | | | |
| M | profundidade da mediana | | mediana | |
| F | profundidade do quarto | quarto inferior | | quarto superior |
| | 1 | extremo inferior | | extremo superior |

**Exemplo 13** - Dada a amostra 12, 14, 16, 24, 26, 27, 32, 34, 45, 46, 46, 47, 57, 58, 59 a representação do resumo de 5 números, é a seguinte:

| | | | | |
|---|---|---|---|---|
| # | 15 | | | |
| M | 8 | | 34 | |
| F | 4.5 | 25 | | 46.5 |
| | 1 | 12 | | 59 |

Se quisermos adicionar mais informação, com outras letras-resumo, basta adicionar as linhas necessárias.

Dada uma amostra, para se proceder à construção da *box-plot*, começa-se por obter o resumo de 5 números, a partir do qual se constrói a *dispersão quartal* – diferença entre os quartos, e as barreiras de *outliers*. A construção da *box-plot* é análoga à já descrita anteriormente, isto é, desenha-se um rectângulo com os lados nos quartos e com uma barra na mediana; em seguida traçamos uma linha que vai do meio do lado do rectângulo até ao valor da amostra mais afastado do rectângulo, que não seja um *outlier*.



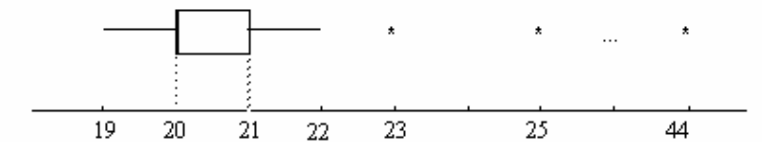
Os *outliers* são representados com um x ou um * e situam-se, portanto, para além das barreiras.

**Exemplo 14** - Dada a amostra seguinte constituída pelas idades dos alunos de uma turma do 2º ano de Estatística, faça a sua representação em *box-plot*

19, 20, 20, 21, 21, 20, 19, 20, 22, 23, 21, 21, 20, 25, 20, 19, 21, 44, 20

| | | | | | |
|---|---|---|---|---|---|
| # | 19 | | | | |
| M | 10 | | 20 | | $d_F$ |
| F | 5.5 | 20 | | 21 | 1 |
| | 1 | 19 | | 44 | |

1.5 dF = 1.5
Barreiras de outliers (18.5, 22.5)
outliers: 23, 25, 44



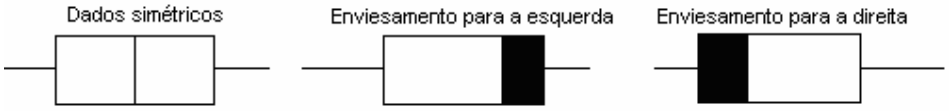

Desta representação, ressalta imediatamente a **forte assimetria** existente entre os dados, já que a **mediana coincide com o quarto inferior**; e a assimetria é positiva, isto é, os dados têm um enviesamento para a direita. Verifica-se também a existência de outliers.

**Como é que se pode reconhecer a simetria ou o enviesamento dos dados, com a *box-plot*?**

Existem fundamentalmente duas características da *box-plot*, que nos dão ideia da simetria ou enviesamento dos dados:

- distância entre a linha indicadora da mediana e os lados do rectângulo
- comprimento das linhas que saem dos lados dos rectângulos

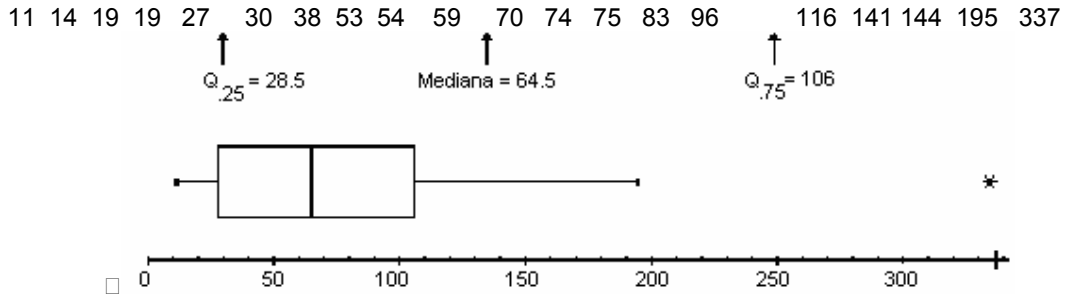
Apresentamos de seguida 3 exemplos de *box-plot*'s, correspondentes a tipos diferentes de distribuição dos dados.




**Exemplo 15** - Num inquérito à comunidade científica sobre a utilização de meios informáticos, realizado pela Fundação para o Desenvolvimento dos meios Nacionais de Cálculo Científico, obtiveram-se os seguintes resultados, quanto ao tipo de problemas tratados:

| | | | |
|---|---|---|---|
| Ajustamento de dados | 337 | Eq. Diferenc. Ordinárias | 54 |
| Análise de Fourier | 195 | Gráfica Computacional | 53 |
| Anál. Estatíst. de Dados | 144 | Integração Numérica | 38 |
| Desenv. de Software | 116 | Inteligência Artificial | 30 |
| Diferenças Finitas | 96 | Interpolação | 27 |
| Diferenciação Numérica | 83 | Método Monte Carlo | 19 |
| Elementos de Fronteira | 75 | Métodos Numéricos | 19 |
| Elementos Finitos | 74 | Simulação | 14 |
| Eq. Algébricas Lineares | 70 | Valores e Vect. Próprios | 11 |
| Eq. Algéb. não Lineares | 59 | Outros | 141 |

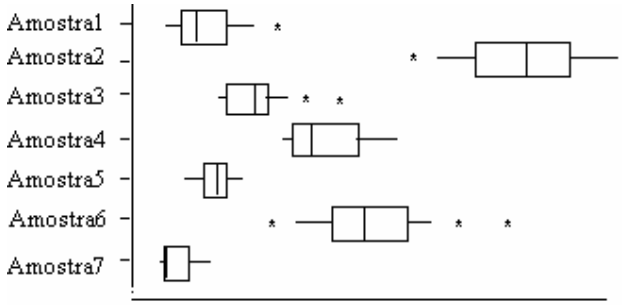
Uma representação em box-plot para estes dados, tem o seguinte aspecto:




Da análise da representação anterior, verifica-se que os dados são um pouco enviesados para a direita e existe um outlier correspondente ao valor 337, que diz respeito à utilização dos meios informáticos para o ajustamento de dados.

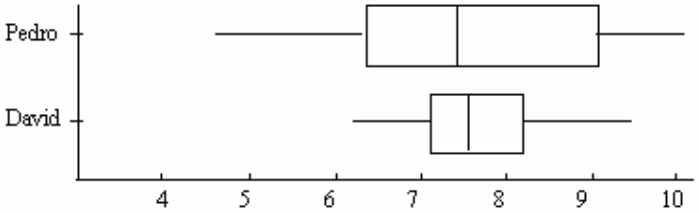
**Utilização da representação em *box-plot* para comparação de amostras**

A representação em *box-plot* é particularmente útil quando se pretendem comparar várias amostras. Para isso consideramos para as diferentes amostras as suas representações *box-plot* dispostas em paralelo. Esta disposição permite comparar as amostras quanto à simetria, comprimento das caudas e *outliers*.



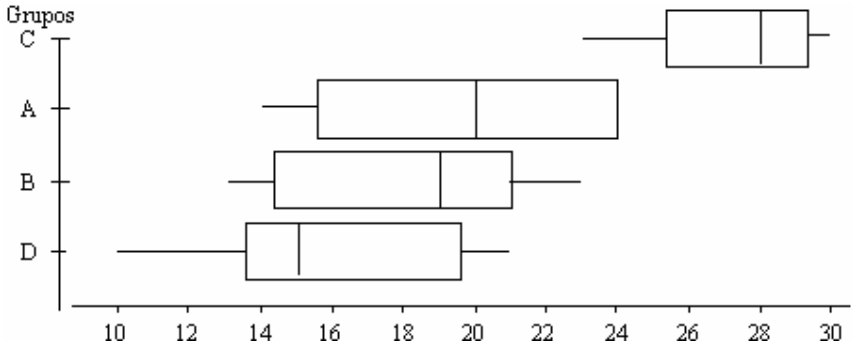

Em vez de dispormos as amostras de forma aleatória, podemos ordená-las de acordo com o valor da mediana. Uma disposição deste género permite verificar um fenómeno que surge frequentemente nos dados, e que é a tendência para o aumento da dispersão, à medida que o nível (localização indicada pela mediana, média,…) aumenta. Este facto não é compatível com a hipótese de igual variabilidade nas diferentes amostras, a qual é muitas vezes necessária para se poderem aplicar determinadas metodologias estatísticas.

**Exemplo 16** – Representado em *box-plot* paralelas os dados apresentados no exemplo 10, relativos à duração do sono do Pedro e do David, obtém-se



evidenciando as características para as quais já se havia chamado a atenção quando se fez a representação em caule-e-folhas.

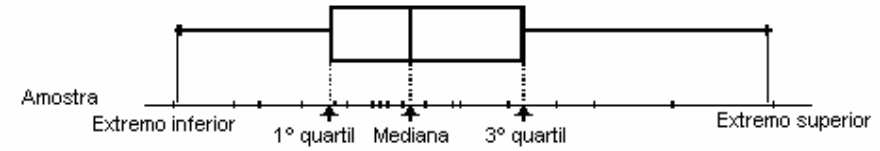
**Exemplo 17** – Considerando os dados do exemplo 5 do capítulo 1, a representação em *box-plot* paralelas realça as diferenças entre os 4 conjuntos de dados:




***Nota***: A construção da Box-plot pode-se fazer quer com os quartos, quer com os quartis, pois a representação gráfica obtida quando não é a mesma, é muito semelhante.

**Diagrama de extremos e quartis**

Uma versão simplificada da representação Box-plot, é o diagrama de extremos e quartis. Para obter esta representação, começa por se recolher da amostra informação sobre 5 números, que são: os extremos, a mediana e os quartis. A representação do diagrama de extremos e quartis tem o seguinte aspecto:



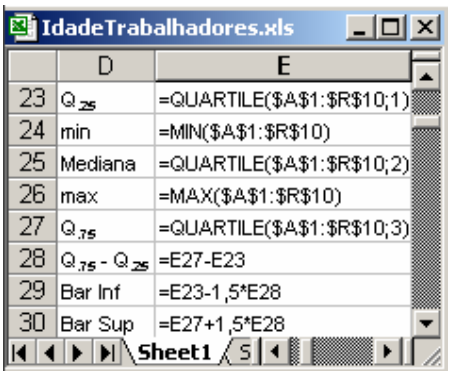

O extremo inferior é o mínimo da amostra, enquanto que o extremo superior é o máximo.

**Utilização do Excel na construção de uma representação em Box-plot**

Mais uma vez estamos perante uma representação gráfica cuja construção, por meio do Excel, necessita de alguns "truques". Assim, o primeiro passo para uma dessas construções, consiste em representar, adequadamente, numa folha de Excel, as estatísticas Mínimo, Máximo. 1º e 3º quartis e mediana.

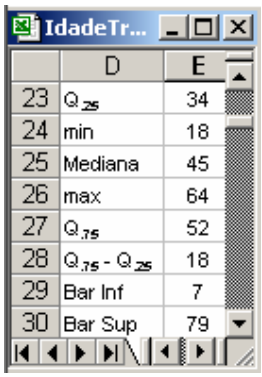
**Exemplo –** Para os dados do ficheiro IdadeTrabalhadores.xls, considerado na demostração com o Excel, da secção 2.2.2.2, construa uma representação em Box-plot, para a variável Idade.

Utilizando o Excel, recomenda-se que se comece por calcular as estatísticas necessárias, que se apresentam a seguir, verificando se existem Outliers:



IdadeTrabalhadores.xls

| | D | E |
|---|---|---|
| 23 | $Q_{.25}$ | =QUARTILE($A$1:$R$10;1) |
| 24 | min | =MIN($A$1:$R$10) |
| 25 | Mediana | =QUARTILE($A$1:$R$10;2) |
| 26 | max | =MAX($A$1:$R$10) |
| 27 | $Q_{.75}$ | =QUARTILE($A$1:$R$10;3) |
| 28 | $Q_{.75}$ - $Q_{.25}$ | =E27-E23 |
| 29 | Bar Inf | =E23-1,5*E28 |
| 30 | Bar Sup | =E27+1,5*E28 |

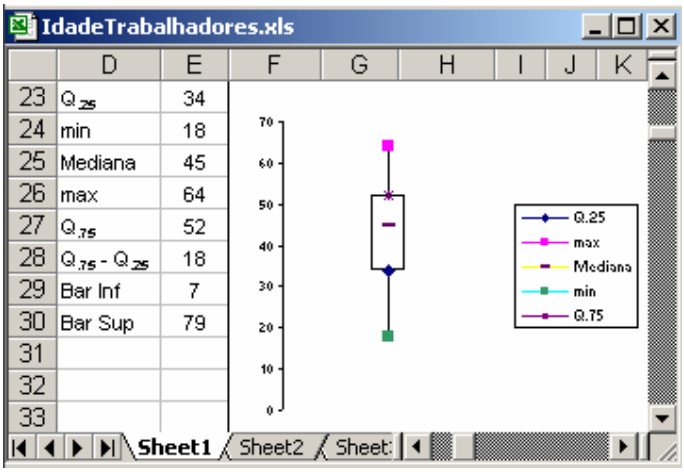


IdadeTr...

| | D | E |
|---|---|---|
| 23 | $Q_{.25}$ | 34 |
| 24 | min | 18 |
| 25 | Mediana | 45 |
| 26 | max | 64 |
| 27 | $Q_{.75}$ | 52 |
| 28 | $Q_{.75}$ - $Q_{.25}$ | 18 |
| 29 | Bar Inf | 7 |
| 30 | Bar Sup | 79 |

Como não existem outliers, a Box-plot resume-se a um diagrama de extremos e quartis, cuja construção segue os seguintes passos:

- Seleccionar as células que contêm as estatísticas 1º quartil, mínimo, mediana, máximo e 3º quartil, assim como as suas etiquetas (é importante que o 1º quartil e o 3º quartil sejam a primeira e a última estatísticas a serem apresentadas na tabela);
- No módulo Chart Wizard (Assistente de Gráficos) seleccionar:
  *Line*
  Seleccionar *Line with markers displayed at each data value*
  Clicar *Next*
  Seleccionar *Series in Rows*
  Clicar *Finish*
- Clicar com o botão direito do rato num dos pontos. Seleccionar:
  *Format Data Series*
  Seleccionar *Options*
  Escolher *High-low lines* e *Up-down bars*; Ajuste à sua escolha *Gap width*;
  *OK*
- Arranjar "esteticamente" o gráfico:



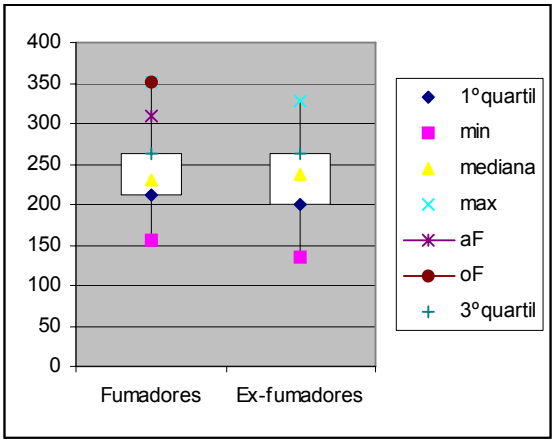

**Exemplo** (De Veaux et al, 2004)– Considere os seguintes dados, que representam o resultado de um estudo para comparar os riscos de doenças cardíacas devidas ao tabaco, utilizando os níveis de colesterol, como termo de comparação. O colestelol foi medido em pessoas que fumam, há pelo menos 25 anos e por ex-fumadores que fumaram no máximo 5 anos:

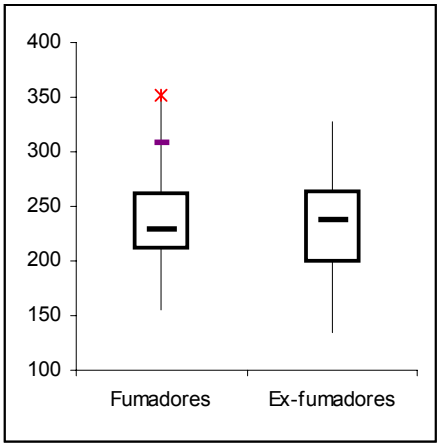
Colesterol.xls

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Fumadores | | | | Ex-fumadores | | | | | Fum | Ex-fum | | | Fumadore | Ex-fumad |
| 2 | 155 | 213 | 232 | 271 | 134 | 217 | 250 | | 1°quartil | 212 | 200 | | 1°quartil | 212 | 200 |
| 3 | 183 | 216 | 232 | 280 | 160 | 217 | 257 | | min | 155 | 134 | | min | 155 | 134 |
| 4 | 196 | 216 | 237 | 280 | 163 | 218 | 263 | | mediana | 230 | 238 | | mediana | 230 | 238 |
| 5 | 200 | 216 | 243 | 280 | 174 | 227 | 267 | | max | 351 | 328 | | max | 351 | 328 |
| 6 | 200 | 216 | 243 | 284 | 175 | 228 | 267 | | 3°quartil | 262,5 | 263 | | aF | 309 | |
| 7 | 200 | 217 | 246 | 287 | 183 | 238 | 271 | | amp | 50,5 | 63 | | oF | 351 | |
| 8 | 200 | 217 | 246 | 288 | 188 | 242 | 292 | | Bar Inf | 136,3 | 105,5 | | 3°quartil | 262,5 | 263 |
| 9 | 209 | 225 | 250 | 305 | 192 | 242 | 300 | | Bar Sup | 338,3 | 357,5 | | | | |
| 10 | 209 | 225 | 256 | 309 | 200 | 243 | 310 | | Outliers | 351 | | | | | |
| 11 | 209 | 225 | 258 | 351 | 213 | 249 | 321 | | | | | | | | |
| 12 | 211 | 230 | 267 | | 213 | 249 | 328 | | | | | | | | |

Sheet1 / Sheet2 / Sheet3

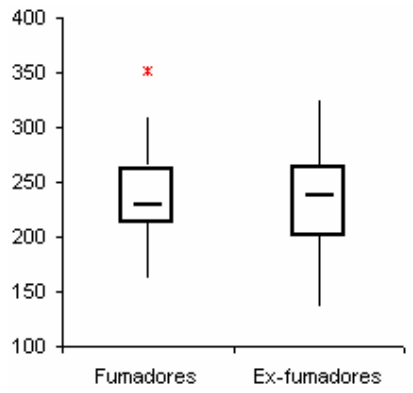
- Os dados são apresentados na tabela do lado esquerdo da figura anterior. Na tabela do meio calculámos as estatísticas necessárias para a construção da Box-plot e verificámos a existência de um outlier nos dados referentes aos Fumadores. Assim, teremos que ligar a barra que sai da caixa no 3º quartil, com o maior valor da amostra que está dentro da barreira e que é o 309. Representámos este valor por aF, que inserimos na tabela do lado direito, assim como inserimos o outlier 351, que representámos por oF (não esquecer que os quartis devem enquadrar a tabela que vai ser utilizada para obter a representação gráfica desejada);
- Seleccionar as células M1:O8, e proceder como no caso anterior:




Obtivémos a representação do lado esquerdo, da figura anterior, que depois de arranjada esteticamente, deu lugar à representação que se encontra no lado direito. Chamamos a atenção para que esta é uma solução, de entre muitas possíveis.

Nota: Numa representação correcta da box-plot, é pressuposto que, para cada amostra, as linhas que saem da caixa estejam ligadas ao menor e maior elementos da amostra que não sejam outliers, e a partir daí marcam-se os outliers com *s. Trabalhando na representação gráfica anterior, obtivémos:




**Exercícios**

**1** - Em 1960 e novamente em 1980 foi feito um inquérito às mulheres americanas sobre o nº de filhos. Os resultados obtidos foram os seguintes:

| *Número de filhos* | *% mulheres 1960* | *% mulheres 1980* |
|---|---|---|
| 0 | 22 | 29 |
| 1 | 17 | 16 |
| 2 | 21 | 22 |
| 3 | 16 | 15 |
| 4 | 10 | 8 |
| 5 | 5 | 4 |
| 6 | 3 | 2 |
| 7 | 2 | 1 |
| 8 | 2 | 1 |
| ≥9 | 3 | 1 |

Construa uma representação gráfica adequada para os dados anteriores e tire conclusões.

**2** - A tabela seguinte mostra a distribuição das frequências relativas do último dígito das idades dos indivíduos adultos. Esta informação foi recolhida relativamente a dois censos diferentes: o Censo de 1880 e o de 1970.

| Dígito | 1880 | 1970 |
|---|---|---|
| 0 | 16.8 | 10.6 |
| 1 | 6.7 | 9.9 |
| 2 | 9.4 | 10.0 |
| 3 | 8.6 | 9.6 |
| 4 | 8.8 | 9.8 |
| 5 | 13.4 | 10.0 |
| 6 | 9.4 | 9.9 |
| 7 | 8.5 | 10.2 |
| 8 | 10.2 | 10.0 |
| 9 | 8.2 | 10.1 |

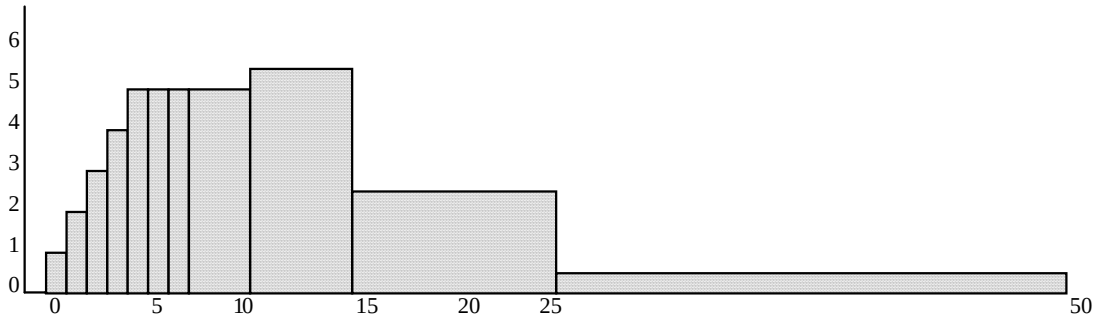
a) Da consulta da tabela verifica a existência de algumas anomalias?

b) Construa diagramas de barras relativamente aos dois censos.

c) Em 1880 havia uma nítida preferência pelos dígitos 0 e 5. Tem alguma explicação para este facto?

d) Em 1970 essa preferência é muito mais fraca. Como explica esse facto?

**3** - O histograma seguinte representa o rendimento familiar, em milhares de dólares de famílias americanas.




Cerca de 1% das famílias têm rendimentos entre 0 e 1000 USD. Estime a percentagem de famílias com rendimentos:

i) a) Entre 1000 USD e 2000 USD
b) Entre 2000 USD e 3000 USD
c) Entre 3000 USD e 4000 USD
d) Entre 4000 USD e 5000 USD
e) Entre 4000 USD e 7000 USD
f) Entre 7000 USD e 10000 USD

ii) a) Haverá mais famílias com rendimentos entre 6000 USD e 7000 USD ou entre 7000USD e 8000 USD? Ou será aproximadamente o mesmo?

b) Haverá mais famílias com rendimentos entre 10000 USD e 11000 USD ou entre 15000USD e 16000 USD? Ou será aproximadamente o mesmo?

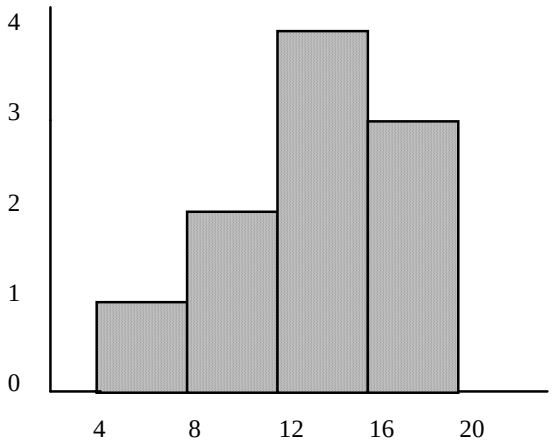
c) Haverá mais famílias com rendimentos entre 10000USD e 12000USD ou entre 15000USD e 20000USD?

R: i) a) 2% b) 3% c) 4% d) 5% e) 15% f) 15%

ii) a) O mesmo b) Mais entre 10000 USD e 11000 USD

c) Mais entre 15000USD e 20000USD

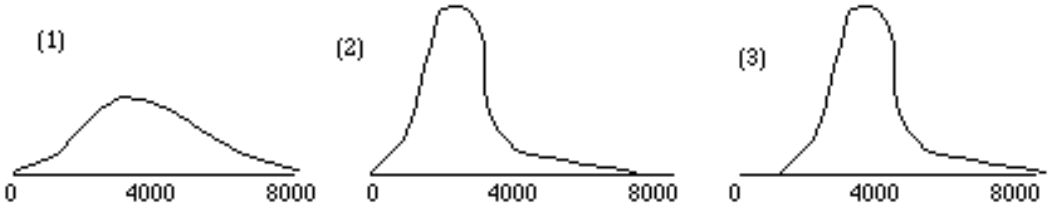
**4** - O histograma seguinte mostra a distribuição das notas finais de Matemática de uma determinada turma.




a) Algum aluno teve nota inferior a 4?
b) 10% dos alunos da turma tiveram nota entre 4 e 8. Qual a percentagem de alunos com nota entre 8 e 12?
c) Qual a percentagem de alunos com nota superior a 12?

R: a) Não b) 20% c) 70%

**5** - Recolheram-se os preços, por hora, de 3 tipos de trabalhadores. Os trabalhadores do grupo B ganham cerca de duas vezes mais do que os trabalhadores do grupo A; os trabalhadores do grupo C ganham mais 1500$ por hora do que os do grupo A. Qual das manchas seguintes, de histogramas, pertence a cada um dos grupos?




R: (1) - B (2) – A (3) - C

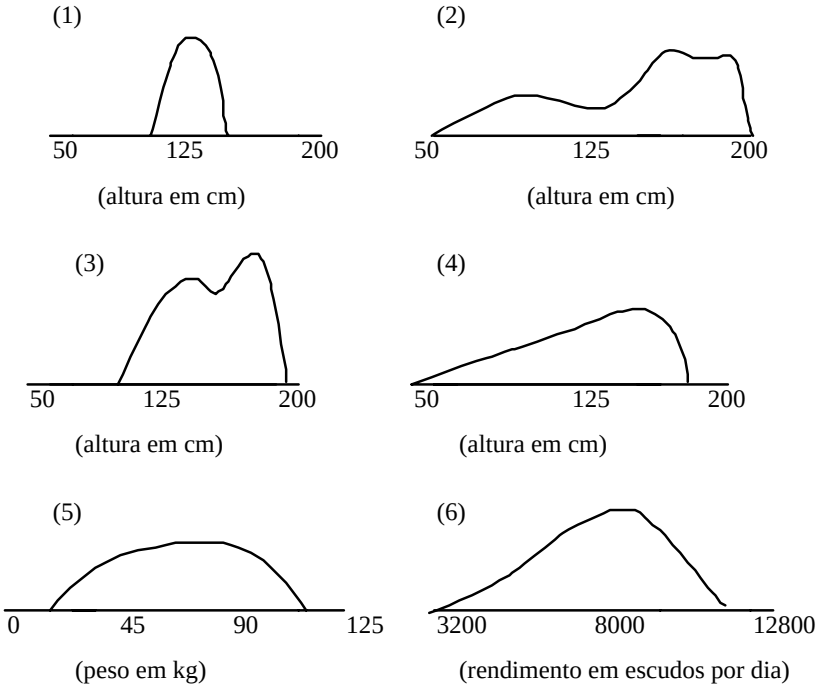
**6** - Seguidamente apresentam-se 6 "manchas" de histogramas, 4 dos quais apresentam os resultados do estudo, numa pequena cidade, das 4 características seguintes :

a) Alturas de todos os elementos das famílias, em que os pais tenham idade inferior a 24 anos.

b) Alturas dos casais (marido e mulher).

c) Alturas de todos os indivíduos da cidade.

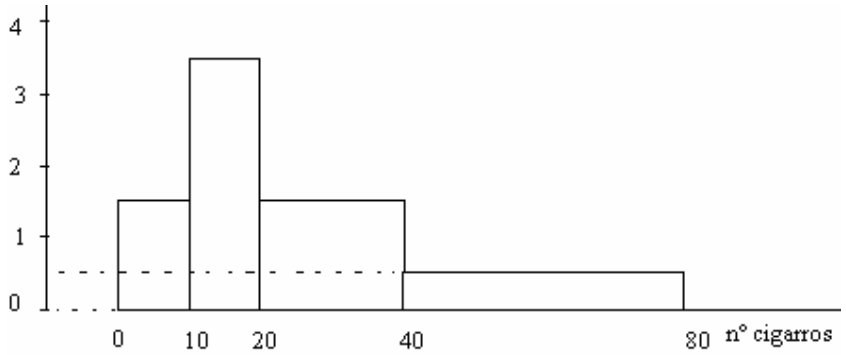
d) Alturas de todos os automóveis.




Quais dos histogramas podem representar cada uma das variáveis anteriores? Explique porquê.

R:a) - (2) b) - (3) c) - (4) d) - (1)

**7** - Um serviço de saúde registou o nº médio de cigarros fumados por dia por cada doente (homem) assistido nesse serviço. Os dados recolhidos permitiram construir o seguinte histograma:




Considerando que a percentagem de fumadores que fuma menos de 10 cigarros por dia é aproximadamente 15%:

a) A percentagem de fumadores que fuma um maço ou mais por dia, mas menos de 2 maços é aproximadamente

1.5% 15% 30% 50%

b) A percentagem de fumadores que fuma um maço ou mais por dia, é aproximadamente

1.5% 15% 35% 50%

c) A percentagem de fumadores que fuma três maços ou mais por dia, é aproximadamente

.25% .5% 10%

d) A percentagem de fumadores que fuma 15 cigarros por dia, é aproximadamente

.30% .5% 1.5% 3.5% 10%

R: a) 30% b) 50% c) 10% d) 3.5%

**8 –** Foi feito um estudo sobre os efeitos secundários da pílula, nomeadamente sobre a tensão arterial. Esse estudo envolveu um pouco mais de 14000 mulheres e os resultados obtidos encontram-se na seguinte tabela:

| Tensão (mm) | Idade n. util. % | 17-24 Utili. % | Idade n. util. % | 25-34 Util. % | Idade n. util. % | 35-44 Util. % | Idade n. util. % | 45-58 Util. % |
|---|---|---|---|---|---|---|---|---|
| <90 | - | 1 | 1 | - | 1 | 1 | 1 | - |
| [90,95[ | 1 | - | 1 | - | 2 | 1 | 1 | 1 |
| [95,100[ | 3 | 1 | 5 | 4 | 5 | 4 | 4 | 2 |
| [100, 105[ | 10 | 6 | 11 | 5 | 9 | 5 | 6 | 4 |
| [105, 110[ | 11 | 9 | 11 | 10 | 11 | 7 | 7 | 7 |
| [110, 115[ | 15 | 12 | 17 | 15 | 15 | 12 | 11 | 10 |
| [115, 120[ | 20 | 16 | 18 | 17 | 16 | 14 | 12 | 9 |
| [120, 125[ | 13 | 14 | 11 | 13 | 9 | 11 | 9 | 8 |
| [125,130[ | 10 | 14 | 9 | 12 | 10 | 11 | 11 | 11 |
| [130,135[ | 8 | 12 | 7 | 10 | 8 | 10 | 10 | 9 |
| [135,140[ | 4 | 6 | 4 | 5 | 5 | 7 | 8 | 8 |
| [140,145[ | 3 | 4 | 2 | 4 | 4 | 6 | 7 | 9 |
| [145,150[ | 2 | 2 | 2 | 2 | 2 | 5 | 6 | 9 |
| [150,155[ | - | 1 | 1 | 1 | 1 | 3 | 2 | 4 |
| [155,160[ | - | - | - | 1 | 1 | 1 | 1 | 3 |
| ≥160 | - | - | - | - | 1 | 2 | 2 | 5 |
| Total | 100 | 98 | 100 | 99 | 100 | 100 | 99 | 99 |
| Nº elementos | 1206 | 1024 | 3040 | 1747 | 3494 | 1028 | 2172 | 437 |

a) Construa os histogramas correspondentes às mulheres com idades compreendidas entre 25 e 34 anos. Tire conclusões sobre a tensão arterial nas utilizadoras e não utilizadoras da pílula.

b) Construa histogramas para as tensões arteriais das não utilizadoras da pílula, com idades compreendidas entre 17-24 e 25-34. O que é que conclui?

9 - A seguinte tabela apresenta os índices gerais de produção industrial, nos diferentes países da comunidade e noutros países ( Fonte : Anuário Estatístico de Portugal - 1992):

| Eur12 | 1984 | Out. países | 1984 | Eur12 | 1990 | Out. países | 1990 |
|---|---|---|---|---|---|---|---|
| Alemanha | 95.3 | Áustria | 95.4 | Alemanha | 117.9 | Áustria | 121.2 |
| Bélgica | 97.6 | Canadá | 95.0 | Bélgica | 118.4 | Canadá | 107.0 |
| Dinamarca | 95.9 | EUA | 98.3 | Dinamarca | 107.8 | EUA | 115.7 |
| Espanha | 98.0 | Finlândia | 96.6 | Espanha | 116.1 | Finlândia | 114.0 |
| França | 99.8 | Japão | 96.5 | França | 113.6 | Japão | 125.4 |
| Grécia | 96.7 | Noruega | 98.0 | Grécia | 103.3 | Noruega | 141.1 |
| Holanda | 96.1 | Suécia | 97.3 | Holanda | 109.1 | Suécia | 105.2 |
| Irlanda | 96.7 | Suiça | 94.2 | Irlanda | 143.8 | Suiça | 118.0 |
| Itália | 98.6 | Turquia | 99.0 | Itália | 117.8 | Turquia | 138.8 |
| Luxemb. | 93.6 | URSS | 95.8 | Luxemb. | 118.0 | URSS | x |
| Portugal | 90.2 | | | Portugal | 135.2 | | |
| Reino Uni. | 94.8 | | | Reino Uni. | 109.3 | | |

Obs: Considerou-se como índice 100 o ano de 1985.
x - Informação não disponível

Faça uma representação gráfica adequada para os dados.

## 2.4 – Dados bivariados. Diagrama de dispersão. Tabela de contingência

Por vezes a População que se pretende estudar aparece sobre a forma de pares de valores, isto é cada indivíduo ou resultado experimental contribui com um conjunto de dois valores. É o que acontece, por exemplo quando se considera para cada aluno candidato ao Ensino Superior, a nota da PGA e a nota da Prova Específica.

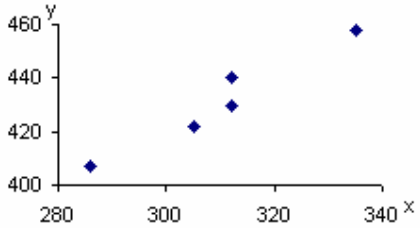

Como representar e organizar este tipo de informação? A representação gráfica utilizada é o diagrama de pontos ou de dispersão.

**Diagrama de dispersão** - É uma representação gráfica para os dados bivariados quantitativos, em que cada par de dados (x,y) é representado por um ponto de coordenadas (x,y), num sistema de eixos coordenados.

Este tipo de representação é muito útil, pois permite realçar algumas propriedades entre os dados, nomeadamente no que diz respeito ao tipo de associação entre os x's e os y's. Seguidamente apresentamos alguns exemplos, para ilustrar o que acabamos de dizer.

**Exemplo 18** - Considere os seguintes dados que representam as medidas em mm, de ossos do braço e da perna, de fósseis do período Neanderthal. Construa o diagrama de dispersão e comente-o.

| Espécie | Braço(Úmero ) x | Perna(Fémur) y |
|---|---|---|
| A | 312 | 430 |
| B | 335 | 458 |
| C | 286 | 407 |
| D | 312 | 440 |
| E | 305 | 422 |



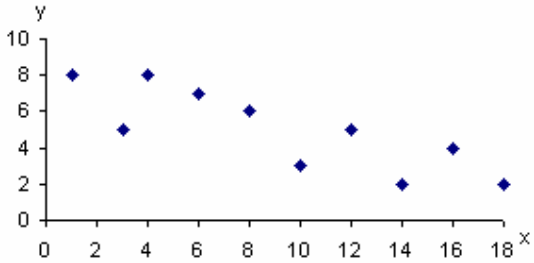

Comentário:

Verifica-se uma associação linear entre as medidas dos ossos do braço e da perna, isto é, aos maiores valores de x correspondem os maiores valores de y. Esta conclusão seria de esperar, pois de um modo geral se as pessoas são grandes, são-no de braços e pernas!

**Exemplo 19** - Considere os seguintes dados, que representam o número de faltas não autorizadas por ano e a distância (em km) a que os empregados de determinado armazém estão de casa.

Construa o diagrama de dispersão e comente-o.

| Distância x | Nº faltas y |
|---|---|
| 1 | 8 |
| 3 | 5 |
| 4 | 8 |
| 6 | 7 |
| 8 | 6 |
| 10 | 3 |
| 12 | 5 |
| 14 | 2 |
| 16 | 4 |
| 18 | 2 |



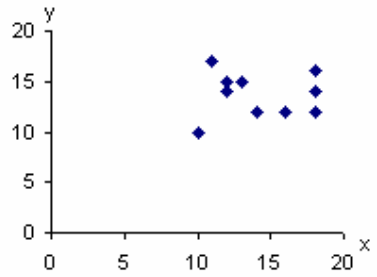

Comentário:

O gráfico mostra uma associação, de sentido contrário, entre o nº de faltas e a distância. Assim, quanto maior é a distância, menor é a tendência para faltar!

**Exemplo 20** - Considere os seguintes dados, que representam as notas obtidas por 10 alunos nas disciplinas de Matemática e Educação Física.

Construa o diagrama de dispersão e comente-o.

| Matemática x | Ed. Física y |
|---|---|
| 12 | 14 |
| 13 | 15 |
| 10 | 10 |
| 11 | 17 |
| 18 | 16 |
| 16 | 12 |
| 12 | 15 |
| 14 | 12 |
| 18 | 14 |
| 18 | 12 |




Comentário:

Aparentemente não existe nenhuma associação linear entre as notas obtidas nas duas disciplinas, uma vez que os pontos se encontram dispersos de forma "aleatória".

Um outro processo de organizar a informação correspondente a dados bivariados, normalmente de tipo qualitativo, é utilizando uma tabela de frequências, a que damos o nome de **tabela de contingência**.

De uma maneira geral, uma tabela de contingência é uma representação dos dados, quer de tipo qualitativo, quer de tipo quantitativo, especialmente quando são de tipo bivariado, isto é, podem ser classificados segundo dois critérios. O aspecto de uma tabela de contingência é o de uma tabela com linhas, correspondentes a um dos critérios, e colunas correspondente ao outro critério. Seguidamente apresentamos um exemplo, para ilustrar o que acabamos de dizer.

**Exemplo 21** – Considerando novamente o exemplo dos passageiros do Titanic (Exemplo 1), classificando os dados relativamente às duas variáveis Classe e Tipo de Sobrevivência, foi possível construir a seguinte tabela (os dados originais não estão disponíveis):

| | | **Classe** | | | | |
|---|---|---|---|---|---|---|
| | | **Primeira** | **Segunda** | **Terceira** | **Tripulação** | **Total** |
| **Sobrev.** | **Vivos** | 202 | 118 | 178 | 212 | **710** |
| | **Mortos** | 123 | 167 | 528 | 673 | **1491** |
| | **Total** | **325** | **285** | **706** | **885** | **2201** |

As células da tabela apresentam as frequências absolutas para cada combinação das modalidades das duas variáveis em estudo. Às distribuições das margens da tabela, chamamos **distribuições marginais**. A coluna da direita representa a distribuição marginal da variável Tipo de Sobrevivência, enquanto que a linha de baixo representa a distribuição marginal da variável Classe.

Normalmente tem mais interesse utilizar as frequências relativas ou percentagens. No entanto, aqui temos vários processos de as calcular: relativamente ao total de passageiros, ou relativamente a cada uma das modalidades, de cada uma das variáveis. Foi isso que fizémos na tabela seguinte:

| | | | Classe | | | | |
|---|---|---|---|---|---|---|---|
| | | | Primeira | Segunda | Terceira | Tripulação | Total |
| Sobrev | Vivos | **Freq.abs.** | **202** | **118** | **178** | **212** | **710** |
| | | %Total | 9,2% | 5,4% | 8,1% | 9,6% | 32,3% |
| | | %Coluna | 62,2% | 41,4% | 25,2% | 24,0% | 32,3% |
| | | %Linha | 28,5% | 16,6% | 25,1% | 29,9% | 100,0% |
| | Mortos | **Freq.abs.** | **123** | **167** | **528** | **673** | **1491** |
| | | %Total | 5,6% | 7,6% | 24,0% | 30,6% | 67,7% |
| | | %Coluna | 37,8% | 58,6% | 74,8% | 76,0% | 67,7% |
| | | %Linha | 8,2% | 11,2% | 35,4% | 45,1% | 100,0% |
| | Total | **Freq.abs.** | **325** | **285** | **706** | **885** | **2201** |
| | | %Total | 14,8% | 12,9% | 32,1% | 40,2% | 100,0% |
| | | %Coluna | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% |
| | | %Linha | 14,8% | 12,9% | 32,1% | 40,2% | 100,0% |

Da tabela anterior podemos tirar várias conclusões, como por exemplo:

- 9,2% (=202/2201) do total de passageiros viajavam em 1ª classe e sobreviveram;
- 62,2% (=202/325) dos passageiros que viajavam em 1ª classe, sobreviveram;
- 28,5% (=202/710) dos passageiros que sobreviveram, viajavam em 1ª classe.

Seria interessante verificar se a distribuição dos passageiros que sobreviveram ou não, teria a ver com a classe em que viajavam. Da tabela anterior, vamos reter as duas tabelas seguintes:

| | Classe | | | | |
|---|---|---|---|---|---|
| | **Primeira** | **Segunda** | **Terceira** | **Tripulação** | **Total** |
| **Vivos.** | 202 | 118 | 178 | 212 | 710 |
| | 28,5% | 16,6% | 25,1% | 29,9% | 100,0% |

e

| | Classe | | | | |
|---|---|---|---|---|---|
| | **Primeira** | **Segunda** | **Terceira** | **Tripulação** | **Total** |
| **Mortos.** | 123 | 167 | 528 | 673 | 1491 |
| | 8,2% | 11,2% | 35,4% | 45,1% | 100,0% |

Como se depreende da tabela anterior, parece não haver independência entre a classe e o tipo de sobrevivência, uma vez que de entre os mortos, predominaram os passageiros que viajavam em terceira classe e os tripulantes. No capítulo 3 voltaremos a estudar a associação entre variáveis de tipo qualitativo.

## Utilização do Excel na construção de uma tabela de contingência

Vamos exemplificar a construção de uma tabela de contingência utilizando a metodologia das PivotTables do Excel.

**Exemplo** – Admita que a seguinte tabela apresenta os dados referentes a 24 passageiros do Titanic:

Titanic.xls

| | A | B | C | D |
|---|---|---|---|---|
| 1 | **Classe** | **Idade** | **Sexo** | **Sobrev** |
| 2 | Tripulante | Adulto | Feminino | Morto |
| 3 | Segunda | Adulto | Masculino | Vivo |
| 4 | Terceira | Adulto | Feminino | Morto |
| 5 | Tripulante | Adulto | Feminino | Morto |
| 6 | Primeira | Criança | Masculino | Morto |
| 7 | Primeira | Adulto | Masculino | Morto |
| 8 | Primeira | Adulto | Masculino | Vivo |
| 9 | Tripulante | Adulto | Masculino | Vivo |
| 10 | Tripulante | Adulto | Feminino | Morto |
| 11 | Tripulante | Adulto | Feminino | Vivo |
| 12 | Terceira | Adulto | Masculino | Morto |
| 13 | Terceira | Criança | Masculino | Vivo |
| 14 | Terceira | Adulto | Masculino | Morto |
| 15 | Tripulante | Adulto | Masculino | Vivo |
| 16 | Segunda | Criança | Feminino | Morto |
| 17 | Tripulante | Adulto | Feminino | Morto |
| 18 | Segunda | Adulto | Masculino | Morto |
| 19 | Segunda | Criança | Masculino | Vivo |
| 20 | Primeira | Adulto | Masculino | Morto |
| 21 | Terceira | Adulto | Feminino | Morto |
| 22 | Tripulante | Adulto | Masculino | Morto |
| 23 | Primeira | Criança | Feminino | Vivo |
| 24 | Tripulante | Adulto | Feminino | Morto |
| 25 | Terceira | Adulto | Masculino | Morto |

Sheet1 Sheet4

Para construir uma tabela de contingência idêntica à apresentada no exemplo 21, em que associa a informação relativa às variáveis Classe e Tipo de sobrevivência, proceda do seguinte modo:

- No menu Data, clique em *PivotTable and PivotChart Report*;
- No passo 1 da *PivotTable and PivotTable Wizard*, siga as instruções, e clique *PivotTable* à pergunta *What kind of report do you want to create*?;
- No passo 2 siga as instruções, seleccionando os dados que pretende usar. Neste caso seleccione as células A1:D25. Se antes de ir ao menu Data, colocar o cursor em alguma célula da tabela a partir da qual quer construir a PivotTable, na janela apresentada neste passo da construção da tabela, as células da tabela aparecem seleccionadas por defeito;
- No passo 3 seleccione o lugar onde pretende criar a tabela;
- Arraste o botão Sobrev da barra *PivotTable*, e coloque-o (drop it) no campo *Row*; Arraste o botão Classe da barra *PivotTable*, e coloque-o (drop it) no campo *Column.* Arraste um dos botões e coloque-o (drop it) no campo *Data* (nós seleccionámos o botão Classe):

Titanic.xls

| | F | G | H | I | J | K |
|---|---|---|---|---|---|---|
| 1 | | | | | | |
| 2 | Count of Classe | Classe | | | | |
| 3 | Sobrev | Primeira | Segunda | Terceira | Tripulante | Grand Total |
| 4 | Morto | 4 | 3 | 4 | 5 | 16 |
| 5 | Vivo | 3 | 2 | 1 | 2 | 8 |
| 6 | Grand Total | 7 | 5 | 5 | 7 | 24 |

Sheet1 Sheet4 Sheet2 She

Do mesmo modo que no exemplo 21, vamos também considerar frequências relativas (Nós optámos por considerar as percentagens de cada célula da tabela, relativas ao total de elementos). Para isso proceda da seguinte forma:

- Faça um duplo clique em *Count of Classe*;
- Na janela que aparece seleccione *Options* e em *Show Data as*, seleccione *% of total*:

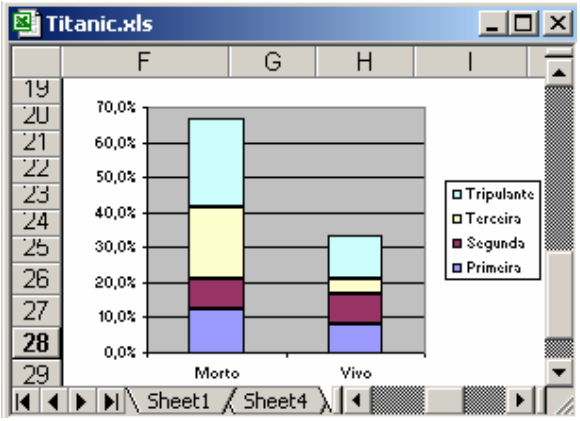
Titanic.xls

| | F | G | H | I | J | K |
|---|---|---|---|---|---|---|
| 1 | | | | | | |
| 2 | Count of Classe | Classe | | | | |
| 3 | Sobrev | Primeira | Segunda | Terceira | Tripulante | Grand Total |
| 4 | Morto | 12,5% | 8,3% | 20,8% | 25,0% | 66,7% |
| 5 | Vivo | 8,3% | 8,3% | 4,2% | 12,5% | 33,3% |
| 6 | Grand Total | 20,8% | 16,7% | 25,0% | 37,5% | 100,0% |

Sheet1 / Sheet4 / Sheet2 / S

Para obter uma representação gráfica associada à tabela anterior, proceda da seguinte forma:

- Clicar em alguma parte da tabela e na barra da *PivotTable* clicar no ícone , criando-se um gráfico numa folha chamada Chart1. No menu Chart seleccione Location e na janela que aparece, em *As object in*, seleccione a folha onde tem a tabela;
- Esconder os botões clicando com o lado direito do rato num deles e seleccionando *Hide PivotChart Field Buttons:*




Pode ainda a partir da tabela dos dados iniciais fazer outros agrupamentos, nomeadamente entrando com mais do que 2 variáveis, como por exemplo:

Titanic.xls

| | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|
| 9 | | | | | | | | |
| 10 | Count of Sexo | Idade | Sexo | | | | | |
| 11 | | Adulto | | Adulto Total | Criança | | Criança Total | Grand Total |
| 12 | Classe | Feminino | Masculino | | Feminino | Masculino | | |
| 13 | Primeira | | 3 | 3 | 1 | 1 | 2 | 5 |
| 14 | Segunda | | 2 | 2 | 1 | 1 | 2 | 4 |
| 15 | Terceira | 2 | 3 | 5 | | 1 | 1 | 6 |
| 16 | Tripulante | 6 | 3 | 9 | | | | 9 |
| 17 | Grand Total | 8 | 11 | 19 | 2 | 3 | 5 | 24 |

Sheet1 / Sheet4 / Sheet2 / Sheet3

**Exercícios**

**1**. Num leilão de computadores em segunda mão verificou-se que, para 10 marcas de computadores, se obtiveram os seguintes preços médios (em escudos) (adaptado de Mendenhall, 1994):

| Tipo computador | Preço médio de venda (novo) | Preço médio proposto no leilão | Preço médio venda no leilão |
|---|---|---|---|
| 20MB PC XT | 120000 | 60000 | 90000 |
| 20MB PC AT | 210000 | 120000 | 172500 |
| IBM XT 089 | 135000 | 60000 | 97500 |
| IBM AT 339 | 210000 | 105000 | 180000 |
| 20MB IBM PS/2 30 | 285000 | 150000 | 217500 |
| 20MB IBM PS/2 50 | 315000 | 210000 | 262500 |
| 60MB IBM PS/2 70 | 600000 | 480000 | 517500 |
| 20MB Compaq SLT | 360000 | 210000 | 262500 |
| Toshiba 1600 | 300000 | 210000 | 270000 |
| Toshiba 1200HB | 345000 | 240000 | 292500 |

a) Construa um diagrama de dispersão que relacione os preços médios propostos, com os preços médios com que são vendidos os computadores no leilão.
b) Construa um diagrama de dispersão que relacione os preços dos computadores novos, com os preços médios com que são vendidos os computadores no leilão
c) Compare os dois gráficos. Qual a relação que parece existir entre as três variáveis?

**2**. *Será que o vinho é bom para o coração*? Há a convicção de que o consumo moderado de vinho ajuda a prevenir ataques cardíacos. Na tabela seguinte apresentamos, para 19 países desenvolvidos, alguns valores respeitantes ao consumo anual de vinho (litros de álcool obtidos a partir do consumo de vinho, por pessoa) e a taxa de mortes anuais por doenças cardíacas (mortes por 100000 pessoas):

| *País* | *Álcool* | *Taxa mortes* | *País* | *Álcool* | *Taxa mortes* |
|---|---|---|---|---|---|
| Austrália | 2.5 | 211 | Holanda | 1.8 | 167 |
| Áustria | 3.9 | 167 | N.Zelândia | 1.9 | 266 |
| Bélgica | 2.9 | 131 | Noruega | 0.8 | 227 |
| Canadá | 2.4 | 191 | Espanha | 6.5 | 86 |
| Dinamarca | 2.9 | 220 | Suécia | 1.6 | 207 |
| Finlândia | 0.8 | 297 | Suiça | 5.8 | 115 |
| França | 9.1 | 71 | R. Unido | 1.3 | 285 |
| Islândia | 0.8 | 211 | EUA | 1.2 | 199 |
| Irlanda | 0.7 | 300 | Alemanha | 2.7 | 172 |
| Itália | 7.9 | 107 | | | |

A partir dos dados anteriores, qual a resposta que daria à questão em estudo?

3. A tabela seguinte compara a previsão do tempo e o tempo que se verificou na realidade, durante o período de 1 ano (De Veaux, 2004):

| | | Tempo verificado | |
|---|---|---|---|
| | | Choveu | Não choveu |
| Previsão | Chove | 27 | 63 |
| | Não chove | 7 | 268 |

a) Qual a percentagem de dias em que choveu?

b) Qual a percentagem de dias em que estava prevista chuva?

c) Qual a percentagem de vezes em que as previsões estiveram correctas?

d) Acha que, de um modo geral, as previsões acertaram? Explique porquê.4

4. Uma empresa fez o rastreio da tensão arterial aos seus colaboradores, tendo obtido os seguintes resultados (De Veaux et al, 2004):

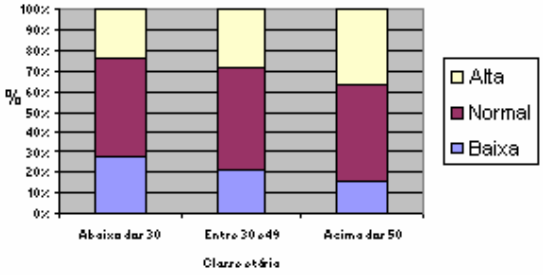
| | | | Idade | |
|---|---|---|---|---|
| | | Abaixo dos 30 | Entre 30 e 49 | Acima dos 50 |
| Tensão arterial | Baixa | 27 | 37 | 31 |
| | Normal | 48 | 91 | 93 |
| | Alta | 23 | 51 | 73 |

a) Determine a distribuição marginal da tensão arterial;

b) Determine a distribuição marginal da tensão arterial, dentro de cada classe etária;

c) Compare graficamente essas distribuições;

Sugestão: Com o Excel obtenha uma representação do tipo:




d) Comente a associação entre a tensão arterial e a idade;

**Capítulo 3**

# Características amostrais

## 3.1 - Introdução

Vimos no capítulo anterior, alguns processos de resumir a informação contida nos dados, utilizando tabelas e gráficos. Veremos neste capítulo, um outro processo de resumir essa informação utilizando determinadas **medidas**, calculadas a partir dos dados, que se chamam **estatísticas**.

Das medidas ou estatísticas que iremos definir, para caracterizar os dados, destacam-se as **medidas de localização**, nomeadamente as que localizam o centro da amostra, e as **medidas de dispersão**, que medem a variabilidade dos dados.

Observemos que ao resumir a informação contida nos dados na forma de alguns números, estamos a proceder a uma redução "drástica" desses dados. Assim, aquelas medidas devem ser convenientemente escolhidas, de modo a representarem o melhor possível o conjunto de dados que pretendem sumariar. Como veremos, definiremos várias medidas possíveis, mas não poderemos dizer, de uma forma geral, que uma é melhor do que outra, já que a sua utilização depende do contexto e da situação em que necessitam de ser calculadas e como vão ser utilizadas.

Será mesmo necessário utilizar os dois tipos de medidas, isto é de localização e de dispersão, para caracterizar um conjunto de dados? O exemplo seguinte procura responder a esta questão.

**Exemplo 1** - Dois alunos do 7º ano obtiveram as seguintes notas no 3º período:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Pedro | 4 | 3 | 3 | 3 | 3 | 3 | 4 | 3 | 4 | 3 |
| João | 5 | 2 | 2 | 3 | 4 | 3 | 5 | 3 | 3 | 3 |

O Pedro e o João tiveram a mesma média de 3.3, mas o João não transitou de ano, pois teve duas negativas. Quer dizer que utilizámos uma medida de redução dos dados, a média, que não é suficiente para caracterizar e diferenciar os dois conjuntos de dados. Efectivamente, se representarmos num diagrama de caule-e-folhas os dois conjuntos, obtemos duas representações com aspecto diferente, já que na segunda representação se verifica uma maior variabilidade, isto é, os dados estão mais dispersos:

| | |
|---|---|
| 3 | 3 3 3 3 3 3 3 |
| 4 | 4 4 4 |

| | |
|---|---|
| 2 | 2 2 |
| 3 | 3 3 3 3 3 |
| 4 | 4 |
| 5 | 5 5 |

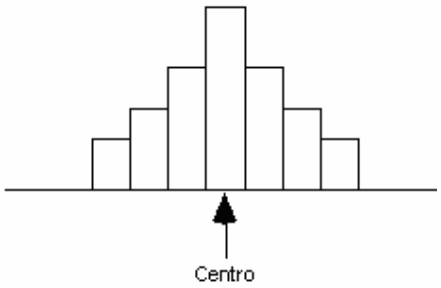
Para definir as medidas que vão ser utilizadas para resumir a informação contida nos dados, e lembramos mais uma vez que estamos na fase da análise estatística conhecida por Estatística Descritiva, utilizamos a seguinte notação para representar a amostra

$$x_1, x_2, x_3, \ldots, x_n$$

onde $x_1$, $x_2$,...., $x_n$, representam, respectivamente, a 1ª observação, a 2ª observação, a n-ésima observação, a serem recolhidas para constituir uma amostra de dimensão n. Esta notação não pressupõe uma ordenação.

## 3.2 - Medidas de localização

De entre as medidas de localização, merecem destaque especial as que localizam o *centro de uma amostra*. Vimos no capítulo anterior, que uma representação gráfica adequada para um conjunto de dados contínuos era, por exemplo, o histograma. Vimos também que um histograma pode ter vários aspectos, nomeadamente pode apresentar uma forma simétrica ou enviesada. No caso particular do histograma ser perfeitamente simétrico, não haveria dúvida em dizer qual o centro dessa distribuição:




No entanto, a situação anterior, a existir, é muito rara, pois devido à aleatoriedade presente nos dados, os histogramas não apresentam aquele aspecto. Por outro lado, quando o histograma é enviesado, a situação ainda se torna mais complicada, pois é difícil dizer o que é o centro. Existem então, vários processos para definir o centro, cujas medidas não dão necessariamente o mesmo resultado. Destas medidas destacamos a média e a mediana, a definir seguidamente.

### 3.2.1 - Média

A média amostral ou simplesmente *média*, é a medida de localização do centro da amostra, mais vulgarmente utilizada. Representa-se por $\overline{x}$ e calcula-se utilizando o seguinte processo:

- Somam-se todos os elementos da amostra;
- divide-se o resultado da soma, pelo número de elementos da amostra.

Utilizando a notação introduzida anteriormente para representar a amostra, a média obtém-se a partir da expressão:

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

**E se os dados se encontram agrupados?**

Neste caso podem-se verificar duas situações:

- Os dados são discretos e as diferentes classes são os diferentes valores que surgem na amostra. Então ainda se pode calcular a média a partir da expressão

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i y_i}{n}$$

onde: k é o número de classes do agrupamento
$n_i$ é a frequência absoluta da classe i
$y_i$ é o ponto correspondente à classe i

- Os dados são discretos ou contínuos e as classes são intervalos. Então já não temos um valor exacto para a média, mas sim um valor aproximado, o qual é dado pela expressão

$$\bar{x} \approx \frac{\sum_{i=1}^{k} n_i y_i}{n}$$

onde: k é o número de classes do agrupamento
$n_i$ é a frequência absoluta da classe i
$y_i$ é o ponto médio da classe i, o qual é considerado como elemento representativo da classe.

**A média será sempre uma medida representativa dos dados?**

Ao determinar a média dos seguintes dados

12.4 13.5 13.6 11.2 15.1 10.6 12.4 14.3 113.5

obteve-se o valor $\bar{x}$ = 24.1.

Embora todos os dados, menos um, estejam no intervalo [10.6, 15.1], o valor obtido para a média está "bem afastado" daquele intervalo! Uma medida que se pretendia representativa dos dados, não está a conseguir esses objectivos, pois se nos disserem que um conjunto de dados tem média 24.1, imediatamente pensamos em valores que não se afastem muito deste valor.

O que acontece é que *a média é muito sensível a valores muito grandes ou muito pequenos*, dizendo-se que é uma medida pouco *resistente*.

No caso do exemplo foi o valor 113.5 que inflacionou a média. Além disso temos alguma razão para pensar que pode ter havido um erro ao digitar o valor 113.5, digitando um 1 a mais! E se em

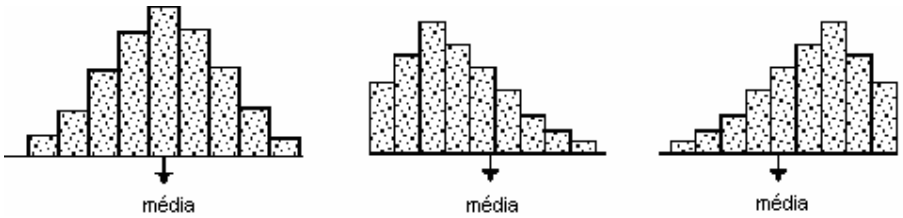
vez de 113.5 o valor correcto fosse 13.5, qual o valor da média? Neste caso para a média dos seguintes dados

12.4 13.5 13.6 11.2 15.1 10.6 12.4 14.3 13.5

obteve-se o valor $\overline{x}$ = 13.0, significativamente diferente do obtido no caso anterior!

**Sendo a média uma medida tão sensível aos dados, é preciso ter cuidado com a sua utilização, pois pode dar uma imagem distorcida dos dados que pretende representar!**

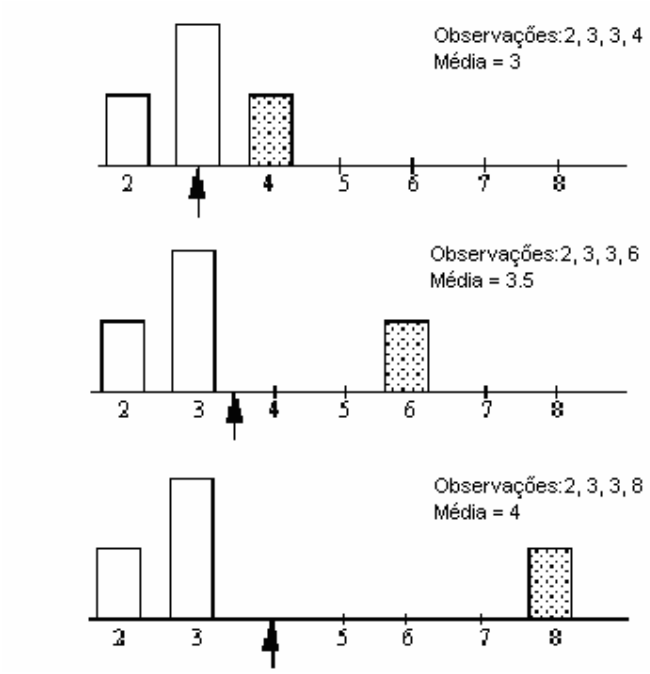
A média dá uma boa indicação do centro da amostra quando a distribuição dos dados for aproximadamente simétrica. Aliás, a sua "popularidade está associada ao facto de quando a distribuição dos dados é "normal" (o histograma correspondente tem a forma aproximada de um sino), então a melhor medida de localização do centro é a média. Ora sendo a Distribuição Normal (como se verá posteriormente, no módulo das distribuições) uma das distribuições mais importantes e que surge com mais frequência nas aplicações, esse facto justifica a grande utilização da média. Esquematicamente podemos posicionar a média da forma que se segue, tendo em conta a representação gráfica na forma de histograma:



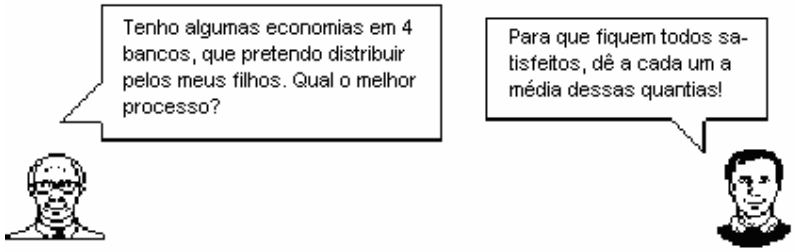

No histograma do lado esquerdo temos uma figura aproximadamente simétrica, pelo que o centro está bem definido. No histograma do centro o enviesamento para a direita provoca uma deslocação da média para a direita; finalmente no histograma da direita o enviesamento provoca uma deslocação da média para a esquerda.

**Exemplo 2** - Considerando os valores 2, 3, 3 e 4, fomos construir um diagrama de barras e posicionar a média e posteriormente alterámos um desses valores para estudar o comportamento da média.

É interessante verificar que um diagrama de barras (ou histograma) se comporta como um balancé, em que o ponto de apoio é a média. Ao contrário da mediana, como se verá adiante, a percentagem de elementos para um e outro lado da média não é necessariamente igual a 50%.




A média tem uma outra característica, que torna a sua utilização vantajosa em certas aplicações: quando o que se pretende representar é a *quantidade total expressa pelos dados*, utiliza-se a média. Na realidade, ao multiplicar a média pelo número total de elementos, obtemos a quantidade pretendida!




*Pode-se sempre calcular a média?*

Chamamos a atenção para que com dados de tipo *qualitativo* não tem sentido calcular a média, mesmo que os dados sejam números. Se, por exemplo, temos um conjunto de “1’s” e “2’s” para representar as classes da variável sexo, em que se utilizou o 1 para representar o sexo masculino e o 2 para o sexo feminino (variável codificada), não tem qualquer significado calcular a média daquele conjunto de dados.

Vamos ver de seguida uma outra medida de localização do centro da amostra, alternativa à média, e que é a mediana.

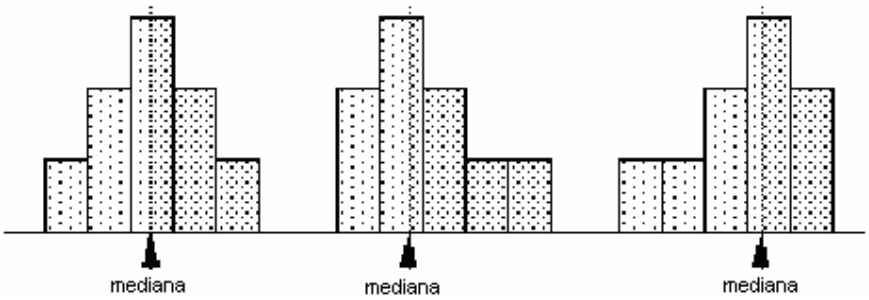
### 3.2.2 - Mediana

A mediana é uma medida de localização do centro da distribuição dos dados, definida do seguinte modo: ordenados os elementos da amostra, a mediana é o valor ( pertencente ou não à amostra), que a divide ao meio, isto é, 50% dos elementos da amostra são menores ou iguais à mediana e os outros 50% são maiores ou iguais à mediana.

Para a determinação da mediana utiliza-se a seguinte regra, depois de ordenada a amostra de n elementos:

- Se **n** é **ímpar**, a mediana é o elemento central;
- se **n** é **par**, a mediana é a semi-soma dos dois elementos centrais.

Dado um histograma é fácil obter a posição da mediana, pois esta está na posição em que passando uma linha vertical por esse ponto o histograma fica dividido em duas partes com áreas iguais.




**Exemplo 3** - Considere o seguinte conjunto de notas de um aluno de Química da FCL:

10 10 10 11 11 11 11 12

A média e a mediana deste conjunto de dados são, respectivamente,

$\overline{x} = 10.75$ e $m = 11$

Admitamos que uma das notas de 10 foi substituída por uma de 18. Então neste caso a mediana continuaria a ser 11, enquanto que a média subiria para 11.75!

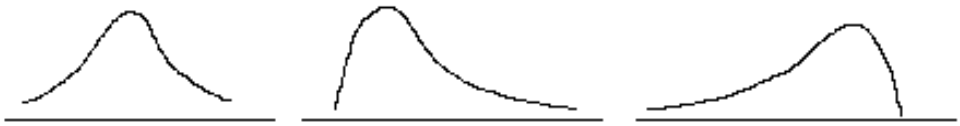
Como medida de localização, a <u>mediana é mais resistente do que a média</u>, pois não é tão sensível aos dados!

**Então qual destas medidas é preferível? Média ou mediana?**

- Quando a distribuição é simétrica, a média e a mediana coincidem.
- A mediana não é tão sensível, como a média, às observações que são muito maiores ou muito menores do que as restantes (*outliers*). Por outro lado, a média reflecte o valor de todas as observações.

<u>Assim, não se pode dizer, em termos absolutos, qual destas medidas é preferível, dependendo do contexto em que estão a ser utilizadas.</u>

Resumindo, como a média é influenciada quer por valores muito grandes, quer por valores muito pequenos, se a distribuição dos dados for enviesada para a direita (alguns valores grandes como *outliers*), a média tende a ser maior que a mediana; se for aproximadamente simétrica, a média aproxima-se da mediana e se for enviesada para a esquerda (alguns valores pequenos como *outliers*), a média tende a ser inferior à mediana. Representando as distribuições dos dados (esta observação é válida para as representações gráficas na forma de diagrama de barras ou de histograma) na forma de uma mancha, temos, de um modo geral:



média ≈ mediana  média > mediana  média < mediana

Observe-se que o simples cálculo da média e da mediana nos pode dar informação sobre a forma da distribuição dos dados.

Observação: O cálculo da mediana pode ser feito à custa da noção de profundidade, como exemplificámos no capítulo anterior.

*Pode-se sempre calcular a mediana?*

Para dados de tipo *qualitativo* pode-se calcular a mediana <u>desde que</u> esteja subjacente uma hierarquia nas diferentes classes ou modalidades que a variável pode assumir.

**Exemplo 4** - Num posto médico há 10 funcionários sendo 4 auxiliares de enfermagem (AE), 3 enfermeiros de 2ª classe (E2), 2 enfermeiros de 1ª classe (E1) e uma enfermeira chefe (EC). A mediana deste conjunto de observações é "enfermeiro de 2ª classe", pois podemos estabelecer uma hierarquia entre as categorias obtendo a amostra ordenada

AE AE AE AE E2 E2 E2 E1 E1 EC
↑
mediana

**Exemplo 5 -** Os salários dos 160 empregados de uma determinada empresa, distribuem-se de acordo com a seguinte tabela de frequências:

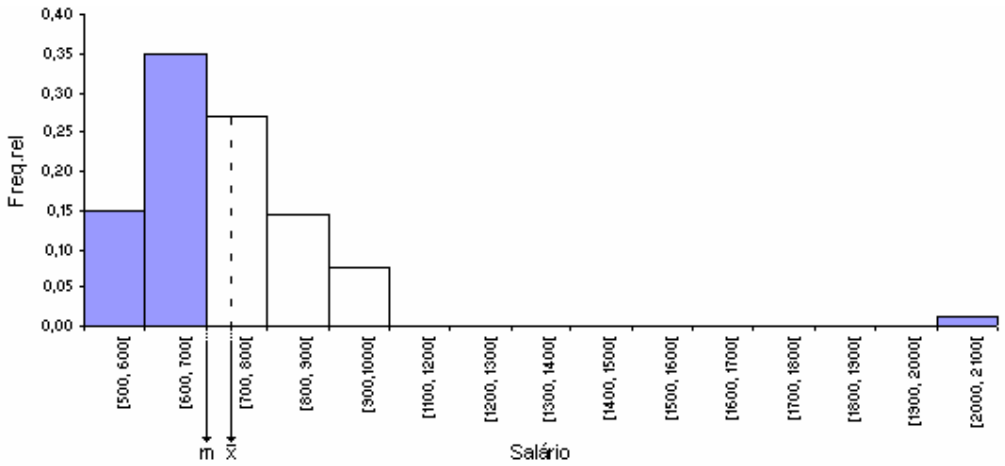
| Salário (euros) | [500, 600[ | [600, 700[ | [700, 800[ | [800, 900[ | [900,1000[ | [2000, 2100[ |
|---|---|---|---|---|---|---|
| Nº empregados | 24 | 56 | 43 | 23 | 12 | 2 |

Calcule a média e a mediana e comente os resultados obtidos.
Cálculo da média:

$\bar{x} \approx (550 \times 24 + 650 \times 56 + \ldots + 950 \times 12 + 2050 \times 2)/ 160$
$\approx 730{,}6$

Cálculo da mediana:

Calculando as frequências relativas e somando, verificará que a soma das duas primeiras classes é 0,5, pelo que podemos considerar como valor aproximado para a mediana
$m \approx 700$




A média é superior à mediana, pois 2 dos valores do conjunto de dados são muito grandes, quando comparados com os restantes, tendo assim inflacionado a média. A mediana dá-nos uma ideia mais correcta do nível dos salários.

A mediana divide a área do histograma em duas partes iguais. A média indica o "ponto de balanço" do histograma, isto é, tem em linha de conta tanto a área das barras, como as suas distâncias ao centro.

No exemplo anterior fomos calcular as áreas dos rectângulos ou partes, para cada lado da linha vertical que passa pela média e multiplicámos pelas distâncias à média (considerámos como distância de um rectângulo à média, a distância entre o ponto médio da sua base e a média) e somámos de forma conveniente os resultados, como se apresenta a seguir:

| | [500, 600[ | [600, 700[ | [700, 730,6[ | [730,6, 800[ | [800, 900[ | [900,1000[ | [2000,2100[ |
|---|---|---|---|---|---|---|---|
| Área rect ou parte | 15 | 35 | 8,22375 | 18,65125 | 14,375 | 7,5 | 1,25 |
| Distância | 180,6 | 80,6 | 15,6 | 34,7 | 119,4 | 219,4 | 1319,4 |
| Áreaxdistância | 2709,0 | 2821,0 | 128,3 | 647,2 | 1716,4 | 1645,5 | 1649,3 |
| | | **5658** | | | **5658** | | |

Assim se compreende que num histograma, quanto mais uma das suas barras estiver afastada das outras, mais influenciará que o centro se desloque na sua direcção.

### 3.2.3 – Quantis. Quartis e quartos

Generalizando a noção de mediana **m**, que como vimos anteriormente é a medida de localização tal que 50% dos elementos da amostra são menores ou iguais a **m**, e os restantes elementos são maiores ou iguais a **m**, temos a noção de *quantil* de ordem p, com 0<p<1, como sendo o valor $\mathbf{Q_p}$ tal que 100p% dos elementos da amostra são menores ou iguais a $\mathbf{Q_p}$ e os restantes 100(1-p)% elementos da amostra são maiores ou iguais a $\mathbf{Q_p}$.

Tal como a mediana, o quantil de ordem p, $\mathbf{Q_p}$, é uma medida que se calcula a partir da amostra ordenada. Para facilitar a sua obtenção vamos considerar a seguinte notação, já utilizada anteriormente, para a amostra ordenada

$$(\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \ldots, \mathbf{x_n}) \xrightarrow{\text{ordenar}} (\mathbf{x}_{1:n}, \mathbf{x}_{2:n}, \mathbf{x}_{3:n}, \ldots, \mathbf{x}_{n:n})$$

Com esta notação, a obtenção do quantil de ordem p faz-se da seguinte forma:

$$Q_p = \begin{cases} x_{[np]+1:n} & \text{se np não é inteiro} \\ \frac{1}{2}(x_{np:n} + x_{np+1:n}) & \text{se np inteiro} \end{cases}$$

onde representamos por [a] a parte inteira de a.

Aos quantis de ordem 1/4 e 3/4 damos respectivamente o nome de 1º quartil e 3º quartil, como já vimos no capítulo anterior.

**Exemplo 6** - Tendo-se decidido registar os pesos dos alunos de uma determinada turma de Matemática do 12º ano, obtiveram-se os seguintes valores (em kg):

52 56 62 54 52 51 60 61 56 55 56 54 57 67 61 49

Um aluno com o peso de 62kg pode ser considerado "normal", isto é, nem demasiado magro, nem demasiado gordo?

Ordenando a amostra anterior, cuja dimensão é 16, temos

49 51 52 52 54 54 55 56 56 56 57 60 61 61 62 67

Para a obtenção dos quartis consideramos:

$16 \times 1/4 = 4$ de onde $Q_{1/4}=(x_{4:16}+x_{5:16})/2 = 53$

$16 \times 3/4 = 12$ de onde $Q_{3/4}=(x_{12:16}+x_{13:16})/2 = 60.5$

Um aluno com o peso de 62 Kg é um bocado “forte”, pois só 25% dos alunos é que têm um peso superior ou igual a 60.5 Kg.

Outras medidas de localização, já consideradas no capítulo anterior, são as letras-resumo, das quais se destacam os ***quartos***, que dão informação em tudo semelhante aos quartis.

### 3.2.4 - Médias aparadas e trimédia

Vimos nas secções anteriores duas medidas de localização do centro da amostra, nomeadamente a média e a mediana. Dissemos que a média é muito sensível a valores muito grandes ou muito pequenos, sendo portanto uma medida *pouco resistente.* Ao contrário a mediana é uma medida *resistente*, pois não é sensível aos *outliers*. No entanto a mediana não representa, tão bem como a média, a totalidade dos dados. Seria então desejável arranjar uma solução de compromisso entre estas duas medidas. Surge assim o conceito de *média aparada*.

Obtém-se uma *média aparada*, eliminando igual número de observações em ambos os extremos da amostra ordenada, onde, se existirem, se situam os *outliers*, e calculando a média com os restantes elementos.

*Quantos elementos é que se devem eliminar de cada um dos extremos?*

Não existe uma regra fixa, pois depende nomeadamente do número de *outliers* existentes. Uma escolha que se costuma fazer é eliminar 10% dos elementos da amostra em cada extremo, resultando num total de 20% de elementos eliminados. Quando a percentagem não der um valor inteiro, considera-se o maior inteiro contido no valor obtido.

**Exemplo 7** - Para os conjuntos A e B calcule a média, a mediana e a média aparada.

| A | B |
|---|---|
| 198 | 198 |
| 175 | 175 |
| 184 | 184 |
| 196 | 196 |
| 168 | 168 |
| 161 | 161 |
| 185 | 185 |
| 184 | 184 |
| 235 | 235 |
| 289 | 96 |

Ordenação

| A | B |
|---|---|
| 161 | 96 |
| 168 | 161 |
| 175 | 168 |
| 184 | 175 |
| 184 | 184 |
| 185 | 184 |
| 196 | 185 |
| 198 | 196 |
| 235 | 198 |
| 289 | 235 |

| | Média | Mediana | Média aparada |
|---|---|---|---|
| Conjunto A: | 197.5 | 184.5 | 190.6 |
| Conjunto B: | 178.2 | 184 | 181.4 |

Outra medida resistente de localização, além da mediana e da média aparada, é a ***trimédia***, definida por

$$\text{Trimédia} = \frac{1}{4}\ (\text{quarto inferior}) + \frac{1}{2}\ (\text{mediana}) + \frac{1}{4}\ (\text{quarto superior})$$

**3.2.5 -Moda**

Para um conjunto de dados, define-se *moda* como sendo o valor que surge com mais frequência, se os dados são discretos, ou o intervalo de classe com maior frequência, se os dados são contínuos.

Assim, das representações gráficas adequadas, para cada um destes tipos de dados, obtém-se imediatamente o valor que representa a moda ou a classe modal.

Esta medida é especialmente útil para reduzir a informação de conjuntos de dados qualitativos, portanto apresentados sob a forma de nomes ou categorias, para os quais não se pode calcular a média e por vezes a mediana ( se não forem susceptíveis de ordenação).

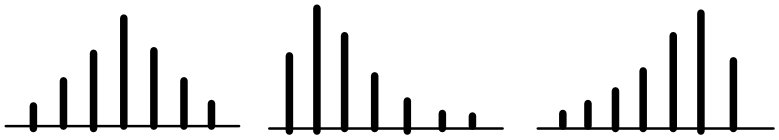
**Exercícios**

**1** - Considere os seguintes conjuntos de números:

1 2 3 4 5 2 3 4 5 6 3 5 7 9 11

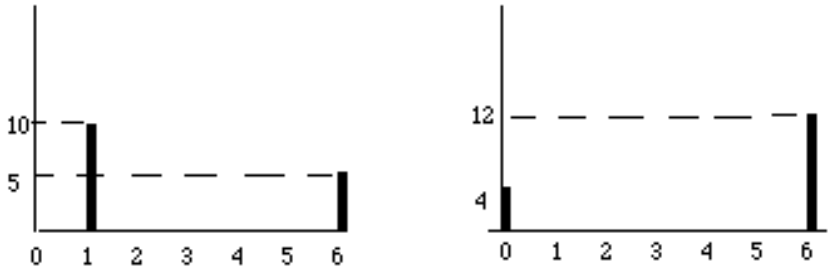
Para cada um destes conjuntos calcule a média. Identifique qual a relação existente entre os conjuntos e diga como poderia obter a média dos dois últimos conjuntos, a partir da média do primeiro conjunto.

**2** - Considere os seguintes diagramas de barras:



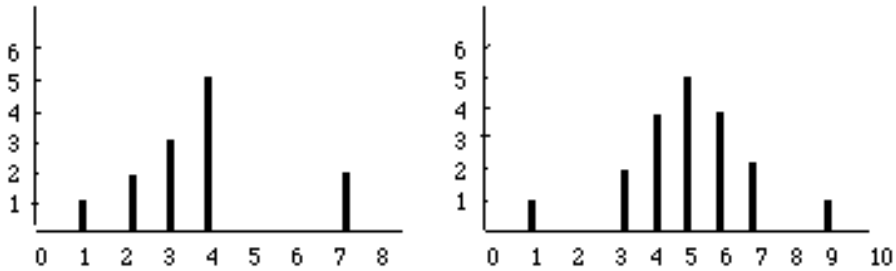
Para cada um deles assinale a posição da média.

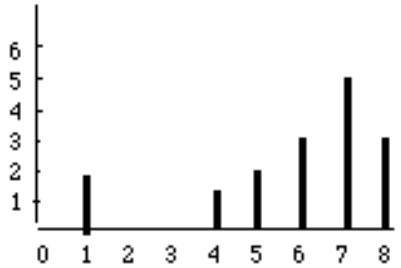
**3** - Faça o mesmo que no exercício anterior para os seguintes diagramas de barras:




Se as barras representarem crianças, em que as frequências absolutas são os respectivos pesos e o eixo horizontal a tábua de um balancé, o que representa o ponto onde marcou a média, relativamente ao balancé, se este estiver em equilíbrio?

**4** - Considere os seguintes diagramas de barras:






Para cada um deles assinale a posição da média e da mediana. O que conclui?

**5** - Numa sala de aulas de 21 alunos, 20 desses alunos têm em média a altura de 145 cm. Se o outro aluno, que no dia em que se fez as medições das alturas tinha faltado, tiver de altura 150, qual é a altura média da turma?

**6** - Numa sala de aulas de 21 alunos, 20 desses alunos têm em média a altura de 145 cm. Qual deve ser a altura do outro aluno, que no dia em que se fez as medições das alturas tinha faltado, para que a altura média da turma aumente de 1 cm?

**7** - Num ponto de Matemática com 5 questões, cada uma valendo 4 valores, verificaram-se os seguintes resultados:

| | | | |
|---|---|---|---|
| 5% | dos alunos | tiveram | 0 |
| 10% | " " | " | 4 |
| 25% | " " | " | 8 |
| 40% | " | " | 12 |
| 15% | " | " | 16 |
| 5% | " | " | 20 |

a) Se o teste foi realizado por 10 alunos, qual a pontuação média obtida?

b) Se o teste foi realizado por 20 alunos, qual a pontuação média obtida?

c) Será que pode calcular a média sem saber o número de alunos? Deduza uma expressão para o cálculo da média, quando os dados estão agrupados e se tem a frequência relativa de cada valor.

d) Qual o valor da mediana?

**8** - Considere os seguintes diagramas de caule-e-folhas:

```
        7
        6
    5 2 4 5
0 1 2 3 4 5 6 7 8 9
```

```
            7
            6
      2 4 5     6
0 1 2 3 4 5 6 7 8 9
```

```
            7
            6
            2   4 5 4
  0 1 2 3 4 5 6 7 8 9
```

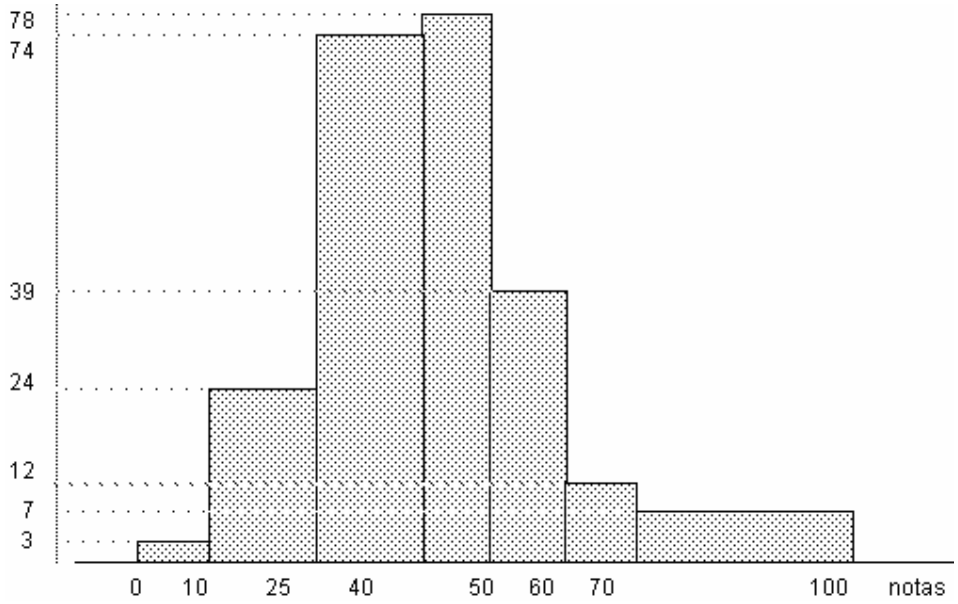
```
            7
            6
            2 4 5 6       4
  0 1 2 3 4 5 6 7 8 9     90
```

Para cada um dos conjuntos de números representados anteriormente, calcule a média e a mediana.

Obs: 1) Nas representações anteriores desenharam-se os traços que separam os caules das folhas horizontalmente, o que torna a representação em caule-e-folhas semelhante ao histograma.

2) Na última representação de caule-e-folhas, utilizou-se uma notação diferente da habitual, pois um dos valores do correspondente conjunto de dados é muito maior do que os outros, optando-se por interromper o traço que separa os caules das pétalas.

**9** - O histograma seguinte representa as notas da prova específica de uma amostra de alunos que entraram para a Faculdade de Ciências, no ano lectivo de 92/93. Cerca de 1% dos alunos tiveram nota inferior a 10.




Relativamente à amostra considerada:

a) Qual a percentagem de alunos com notas da prova específica

i) Entre 10 e 25?

ii) Entre 25 e 50?

iii) Superior a 50?

b) Haverá mais alunos com nota entre 40 e 50, ou entre 25 e 40? Justifique.

c) Indique um valor aproximado para a média.

d) Indique um valor aproximado para a mediana.

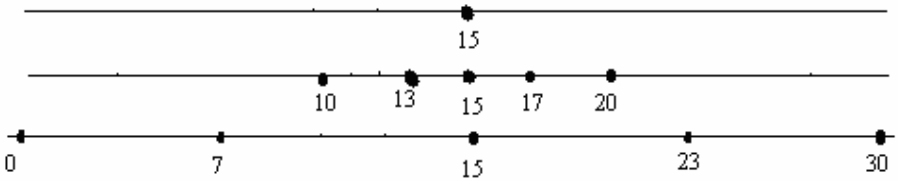
## 3.3 - Medidas de dispersão

Um aspecto importante no estudo descritivo de um conjunto de dados, é o da determinação da variabilidade ou dispersão desses dados, relativamente à medida de localização do centro da amostra. Efectivamente as medidas de localização que estudámos, não são suficientes para caracterizar completamente um conjunto de dados.

Considerem-se os três conjuntos de dados:

| | | | | | |
|---|---|---|---|---|---|
| Conjunto 1 | 15 | 15 | 15 | 15 | 15 |
| Conjunto 2 | 10 | 13 | 15 | 17 | 20 |
| Conjunto 3 | 0 | 7 | 15 | 23 | 30 |

Embora tenham a mesma média, mediana e média aparada, têm um aspecto bem diferente no que diz respeito à variabilidade.




Como a medida de localização mais utilizada é a média, será relativamente a ela que se define a principal medida de dispersão - o desvio padrão, apresentado a seguir. Começamos, no entanto, por definir variância, que serve de base à definição de desvio padrão.

### 3.3.1 - Variância

Define-se a **variância** e representa-se por $s^2$, como sendo a medida que se obtém somando os quadrados dos desvios das observações relativamente à média, e dividindo pelo número de observações menos uma:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

*Quais as razões que nos levam a considerar aquela definição para a variância?*

- <u>Se afinal pretendemos medir a dispersão relativamente à média, porque é que não somamos simplesmente os desvios, em vez de os quadrar</u>?

O que acontece é que a soma dos desvios é igual a zero, pois os desvios positivos estão a cancelar com os desvios negativos,

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x}) + .... +(x_n - \bar{x}) = 0\ !$$

Poderíamos ter utilizado módulos, para evitar que os desvios positivos cancelassem com os desvios negativos, mas é mais fácil trabalhar com os quadrados, do que com os módulos!

- E então porque é que em vez de dividirmos por n, que é o número dos desvios, dividimos por (n-1)?

Na realidade, só aparentemente é que temos n desvios independentes, isto é, se calcular (n-1) desvios, o restante fica automaticamente calculado, uma vez que a sua soma é igual a zero, como vimos no parágrafo anterior. Costuma-se referir este facto, dizendo que se "perdeu" um "grau de liberdade".

Uma vez que a variância envolve a soma de quadrados, a unidade em que se exprime não é a mesma que a dos dados. Assim, para obter uma medida da variabilidade ou dispersão com as mesmas unidades que os dados, tomamos a raiz quadrada da variância e obtemos o *desvio padrão*.

### 3.3.2 - Desvio padrão

Pelas razões apontadas anteriormente, a medida de dispersão que se costuma utilizar é o **desvio padrão**, que se representa por **s** e é a raiz quadrada da variância:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

O desvio padrão é uma medida que **só pode assumir valores não negativos** e quanto **maior for**, **maior** será **a dispersão** dos dados.

Relativamente aos três conjuntos de dados apresentados no início do estudo das medidas de dispersão, verificamos que:

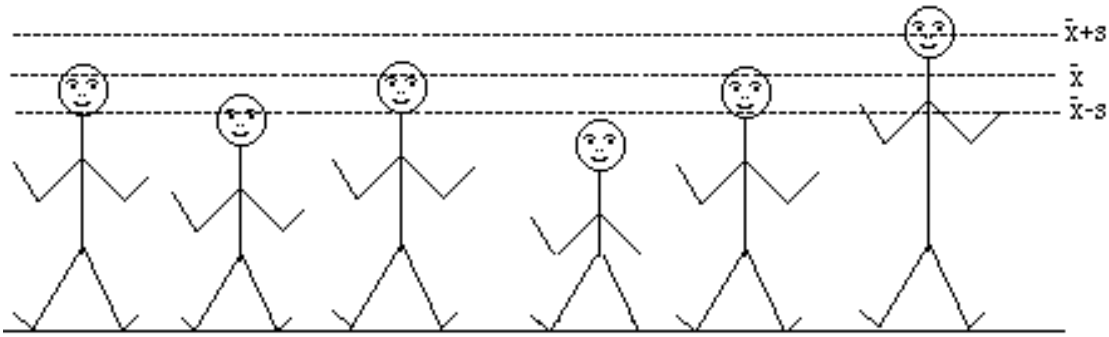
- o conjunto 1 apresenta um desvio padrão igual a zero, como seria de esperar, pois se os valores são todos iguais, a dispersão é nula;
- os conjuntos 2 e 3 apresentam um desvio padrão igual, respectivamente, a 3.8 e 12.0.

O desvio padrão, da mesma forma que a média, é *muito sensível à presença de outliers*, sendo portanto uma medida de dispersão pouco resistente. Assim, um valor grande para o desvio padrão, pode ser devido a uma grande variabilidade nos dados, ou então a uma pequena variabilidade, mas à existência de um ou mais *outliers*.

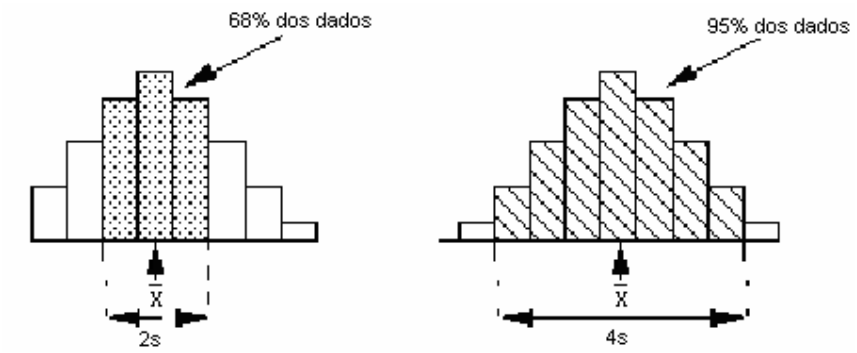
Propriedade para dados com distribuição aproximadamente normal:

Uma propriedade que se verifica se os dados se distribuem de forma aproximadamente *normal*, ou seja, quando o histograma apresenta uma forma característica com uma classe média predominante e as outras classes distribuindo-se à volta desta de forma aproximadamente simétrica e com frequências a decrescer à medida que se afastam da classe média, é a seguinte:

- Aproximadamente 68% dos dados estão no intervalo [$\bar{x}$- s , $\bar{x}$+ s]




- Aproximadamente 95% dos dados estão no intervalo [$\bar{x}$- 2s , $\bar{x}$+ 2s]




- Aproximadamente 100% dos dados estão no intervalo [$\bar{x}$- 3s , $\bar{x}$+ 3s];

Como se depreende do que atrás foi dito, se os dados se distribuem de forma aproximadamente normal, então estão praticamente todos concentrados num intervalo de amplitude 6 vezes o desvio padrão: quanto menor for o desvio padrão, mais concentrada é a distribuição dos dados.

**Exemplo 8** - Perguntou-se o preço da "bica" em 5 cafés, tendo-se obtido os seguintes valores:

50 55 55 55 57.5 60

Calculando a média e o desvio padrão daqueles valores, obtém-se:

$$\bar{x} = \frac{50+55+55+55+57.5+60}{6} = 55.4$$

$$s^2 = \frac{(50-55.4)^2 + 36(55-55.4)^2 + (57.5-55.4)^2 + (60-55.4)^2}{5}$$

= 10.16

de onde s = 3.19

Estes valores significam que, mais de dois terços das vezes, o preço da bica está no intervalo (52.2, 57.6). Obtivemos o intervalo anterior subtraindo e adicionando o valor do desvio padrão à média.

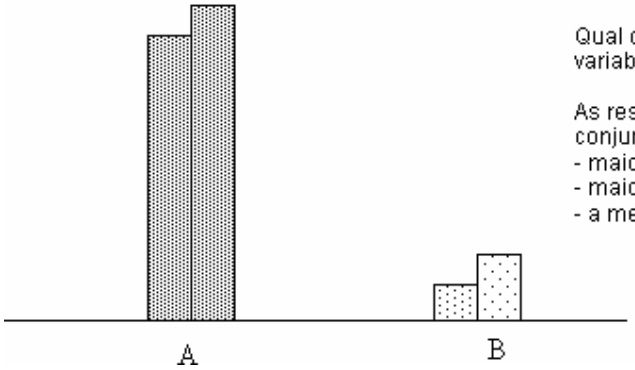
**Exemplo 9 -** O que mede o desvio padrão? Que tipo de variabilidade? (The standard deviation: some drawbacks of an intuitive approach - *Teaching Statistics*, vol 7, n.3, 1985)

A variabilidade apresentada por um conjunto de observações pode-se interpretar como:

- uma medida da diferença entre as obervações, umas relativamente às outras;
- uma medida da diferença entre as observações relativamente a uma medida padrão.

A seguinte experiência dá conta de que nem sempre o desvio padrão é entendido pelos alunos como uma medida da variabilidade relativamente à média.

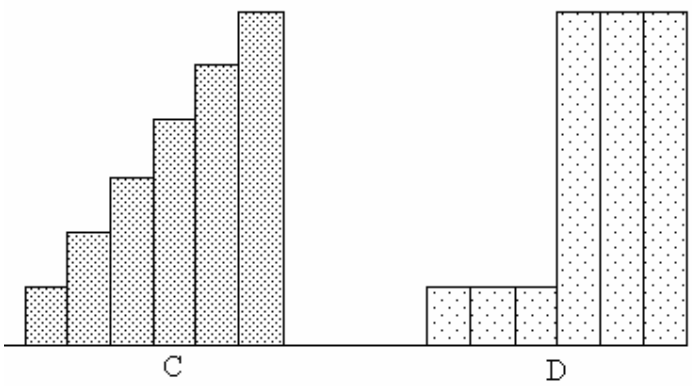
Consideremos dois conjuntos formados cada um por dois blocos: no 1º conjunto os blocos têm altura 45 e 50 cm. No 2º conjunto as alturas dos blocos são 5 e 10 cm:




Qual dos conjuntos apresenta maior variabilidade?

As respostas apresentadas por um conjunto de 154 estudantes, foram:
- maior variab. o A - 11%
- maior variab. o B - 69%
- a mesma variab. - 20%

Apresentou-se seguidamente aos mesmos alunos outros dois conjuntos C e D. No conjunto C os blocos têm alturas 10, 20, 30, 40, 50 e 60 cm; no conjunto D há 3 blocos de altura 10 cm e outros 3 blocos de altura 60 cm:

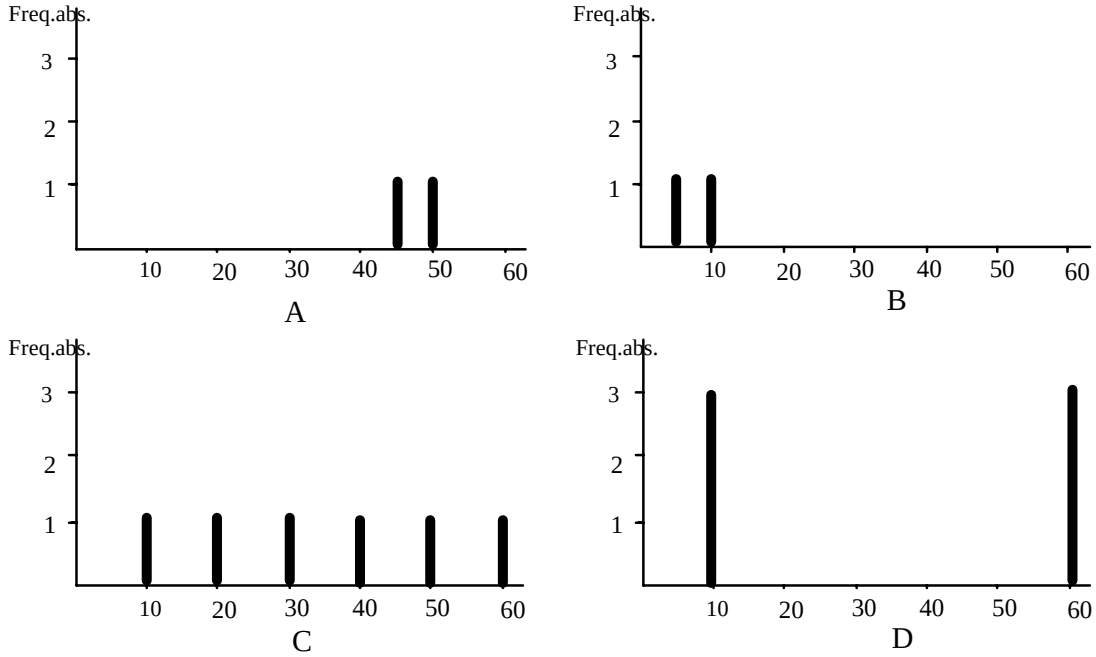



Qual dos conjuntos apresenta maior variabilidade?

As respostas apresentadas pelo conjunto de 154 estudantes, foram:
- maior variab. o C - 50%
- maior variab. o D - 36%
- a mesma variab. - 14%

**Comentário**: o resultado da experiência mostra que intuitivamente os estudantes entendem, de um modo garal, a variabiliadde em termos de "mais ou menos iguais uns relativamente aos outros", independentemente de considerarem um ponto padrão como referência, nomeadamente a média.

Assim para visualizar convenientemente o conceito de variabilidade medida pelo desvio padrão, apresentam-se diagramas de barras. A partir destes gráficos os estudantes podem ver que a variabilidade das alturas pode ser expressa em termos dos desvios relativamente à média:




Pedindo para calcular o desvio padrão das alturas de cada um dos conjuntos os estudantes facilmente verificam que:

desvio padrão de A = desvio padrão de B

desvio padrão de C < desvio padrão de D

Confrontados com os resultados intuitivos, os estudantes concluem que o desvio padrão é uma medida muito específica da variabilidade.

**Expressão alternativa para o cálculo da variância:**

A partir da expressão que define a variância, pode-se deduzir sem dificuldade uma expressão mais simples para o seu cálculo, assim como o do desvio padrão, e que é a seguinte:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2}{n-1} - \frac{n}{n-1}\bar{x}^2$$

Observação: Por vezes, devido a erros de arredondamento, a fórmula anterior dá um valor negativo para a variância, pelo que é necessário ter cuidado!

### 3.3.3 - Amplitude inter-quartil

A medida mais simples para medir a variabilidade é a amplitude, que se representa por um R (range) e se define como a diferença entre o máximo da amostra e o mínimo:

R = máximo - mínimo

A medida anterior tem a grande desvantagem de ser muito sensível à existência, na amostra, de uma observação muito grande ou muito pequena. Assim, define-se uma outra medida, a *amplitude inter-quartil*, que é, de certo modo, uma "solução de compromisso", pois não é afectada, de um modo geral, pela existência de um número pequeno de observações demasiado grandes ou demasiado pequenas. Esta medida é definida como sendo a diferença entre os 1º e 3º quartis:

amplitude inter-quartil= 3º quartil - 1º quartil

ou, utilizando a notação já introduzida,

amplitude inter-quartil = $Q_{3/4} - Q_{1/4}$

Do modo como se define a *amplitude inter-quartil*, concluimos que 50% dos elementos no centro da amostra, estão contidos num intervalo com aquela amplitude. Esta medida já foi, aliás, utilizada na construção da *box-plot*. Esta medida é *não negativa* e *será tanto maior quanto maior for a variabilidade* nos dados.

Mas, ao contrário do que acontece com o desvio padrão, uma amplitude inter-quartil nula, não significa necessariamente, que os dados não apresentem variabilidade. Por exemplo, o seguinte conjunto de dados

10 20 30 30 30 30 30 30 40 50

tem desvio padrão igual a 10.5 e amplitude inter-quartil igual a zero.

*Qual das medidas de dispersão utilizar? Desvio padrão ou amplitude inter-quartil?*

Do mesmo modo que a questão foi posta relativamente às duas medidas de localização mais utilizadas - média e mediana, também aqui se pode por o problema de comparar aquelas duas medidas de dispersão.

**1** - A *amplitude inter-quartil* é mais resistente, relativamente à presença de outliers, do que o *desvio padrão*, que é mais sensível aos dados. Por outro lado, a amplitude inter-quartil não reflecte o conjunto de todos os dados, como o desvio padrão.

**2** - Para uma distribuição dos dados aproximadamente normal, verifica-se a seguinte relação

amplitude inter-quartil $\approx 1.3 \times$ desvio padrão

**3** - Se a distribuição é enviesada, já não se pode estabelecer uma relação análoga à anterior, mas pode acontecer que o desvio padrão seja muito superior à amplitude inter-quartil, sobretudo se se verificar a existência de "outliers".

### 3.3.4 - Dispersão relativa

De um modo geral verifica-se que a variabilidade presente num conjunto de dados aumenta com a localização. Por exemplo se pretendemos comparar vários conjuntos de dados, uma maneira possível é utilizar as *box-plot* paralelas. Quando falámos nesta representação aconselhámos a

dispor as amostras por ordem crescente da mediana (localização), verificando-se normalmente que os comprimentos das caixas também cresciam com a mediana. Assim, para compararmos conjuntos de dados diferentes, convém utilizar uma medida, que dê uma ideia da variabilidade relativamente à localização. Uma medida que se costuma utilizar e que dá a dispersão relativa é o chamado *coeficiente de dispersão ou coeficiente de variação*:

$$\text{coeficiente de dispersão} = \frac{s}{\bar{x}}$$

Ao coeficiente anterior, quando expresso em percentagem, dá-se o nome de *coeficiente de variação*.

**Exercícios**

**1** - Suponha que adicionou 100, a cada um dos valores de uma amostra. O que acontece:

a) ao desvio padrão?
b) à amplitude inter-quartil?
c) à amplitude?
d) à média?
e) à mediana?

E se em vez de adicionar 100, multiplicar por 100? Generalize as conclusões anteriores para uma constante k qualquer.

**2** - Suponha que obteve o valor -40.5 para a variância. O que conclui?

**3** - Suponha que a amplitude de uma amostra é 105.4, e que ao calcular o desvio padrão obteve o valor 160.6. O que conclui?

**4** - Suponha que os resultados de um teste de Matemática, em duas turmas, uma de rapazes e outra de raparigas, se distribuem aproximadamente segundo uma normal. Nesse teste, enquanto que as raparigas tiveram em média 55 pontos, os rapazes tiveram 50. Para ambas as distribuições o desvio padrão foi de 10.

**a)** Qual o valor aproximado para a percentagem de raparigas com nota superior a 75 pontos?

**b)** Qual o valor aproximado para a percentagem de rapazes com nota inferior a 50 pontos? E a 40 pontos?

## Utilização do Excel na obtenção das estatísticas descritivas

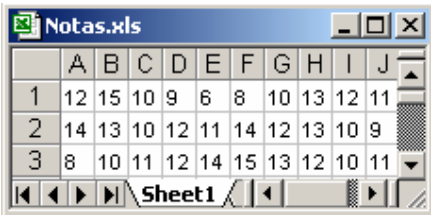
A utilização do Excel no cálculo das estatísticas descritivas não apresenta qualquer dificuldade, como exemplificamos a seguir.

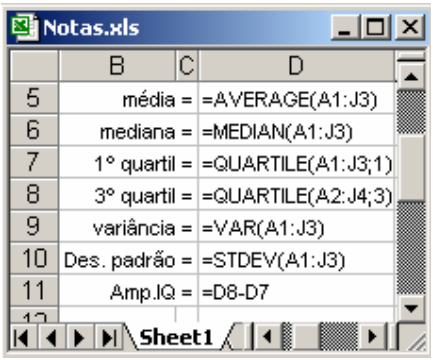
**Exemplo** – Considere os seguintes dados que representam as notas obtidas por 30 estudantes num teste de Estatística:

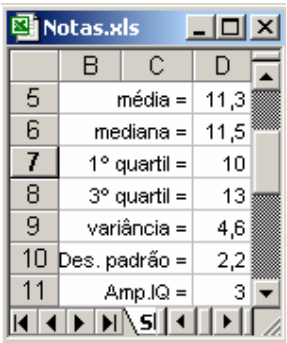
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 12 | 15 | 10 | 9 | 6 | 8 | 10 | 13 | 12 | 11 |
| 14 | 13 | 10 | 12 | 11 | 14 | 12 | 13 | 10 | 9 |
| 8 | 10 | 11 | 12 | 14 | 15 | 13 | 12 | 10 | 11 |

Utilizando o Excel, calcule a média, variância, desvio padrão e quartis e amplitude inter-quartil.

Inserimos os dados nula folha de Excel e depois utilizámos as funções apropriadas:

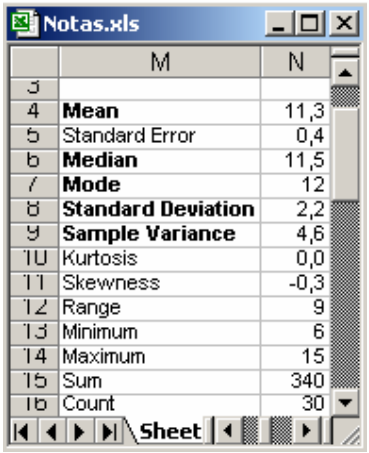



Uma alternativa para o cálculo da mediana, é através da função Quartili(A1:J3;2).

O Excel dispõe ainda de uma função que se obtém seleccionando *Tools* → *data Analysis* → *Descriptive Statistics*, onde se acede a uma janela em que inserimos os endereços das células com os dados, que devem estar numa única coluna, e onde se selecciona *Summary Statistics*, obtendo-se a seguinte tabela de estatísticas:




Na tabela anterior apresentam-se algumas estatísticas, como a *kurtosis* e a *skewness*, que têm a ver com a forma da distribuição dos dados, mas que não definiremos, assim como não definiremos Standard Error. As outras medidas são a Amplitude, Mínimo, Máximo, Soma dos dados e Número de dados da amostra.

## 3.4 – Associação de variáveis

### 3.4.1 - Coeficiente de correlação

Já vimos no capítulo das representações gráficas, que quando dispomos de amostras de dados bivariados, que vamos passar a representar por $(x_i,y_i)$, i=1, …, n, a sua representação num diagrama de dispersão pode mostrar a existência de uma certa *associação linear* entre os factores x e y, que compõem os pares. No que se segue admitimos que as variáveis são de tipo quantitativo.

A medida que se utiliza com mais frequência para medir o grau desta associação linear, é o *coeficiente de correlação*, que se representa por **r,** e se calcula a partir da expressão:

$$\mathbf{r} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \qquad \text{onde} \qquad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Na expressão anterior $\bar{x}$ e $\bar{y}$, representam, respectivamente, as médias dos $x_i$'s e dos $y_i$'s.

Na definição do coeficiente de correlação de pares de variáveis, está implícita a definição de uma medida que dá uma ideia da variabilidade conjunta existente entre as variáveis e que é a *covariância amostral*:

$$\text{Covariância} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Esta medida tem o inconveniente de depender drasticamente das unidades com que se apresentam os elementos da amostra e daí o facto de normalmente não ser utilizada, passando-se imediatamente à definição do coeficiente de correlação (independente das unidades utilizadas), que como facilmente se verifica da expressão anteriormente considerada, vem:

$$\text{Correlação} = \frac{\text{covariância}}{\sqrt{\text{variância}(x)}\sqrt{\text{variância}(y)}}$$

<u>Propriedades do coeficiente de correlação:</u>

**1** – O valor de **r** está no intervalo [-1,1]

**2** – Quanto *maior* for o *módulo* de **r**, *maior* será a *relação linear* existente entre os $x_i$. e os $y_i$.

**3** – O facto de **r** ser *positivo*, significa que a relação entre os x's e os y's é do *mesmo sentido*, isto é, a valores grandes de x, correspondem valores grandes de y e vice-versa. Quando **r** é *negativo*, a relação entre os x's e os y's é de *sentido contrário*, o que significa que a valores grandes de x, correspondem valores pequenos de y e vice-versa.

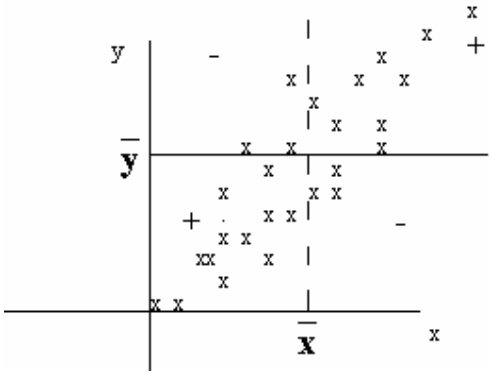
Interpretação geométrica:

**1** – Se aos maiores valores de x estão associados os maiores valores de y, então r>0.

Efectivamente, quando pensamos num valor grande de x, será um valor acima da média. Por outro lado, um valor pequeno de x é um valor abaixo da média. Então se, de um modo geral, aos valores grandes de x estão associados os valores grandes de y, e aos valores pequenos de x estão associados os valores pequenos de y, os produtos

$$(x_i - \overline{x})(y_i - \overline{y})$$

são de um modo geral positivos, já que ambos os factores são positivos ou negativos. Como o denominador da expressão do coeficiente de correlação, não depende da forma como os x's se associam com os y's, então o facto de no numerador somarmos grande número de parcelas positivas, faz com que o valor do coeficiente de correlação seja positivo e tanto maior quantas mais parcelas positivas houver.

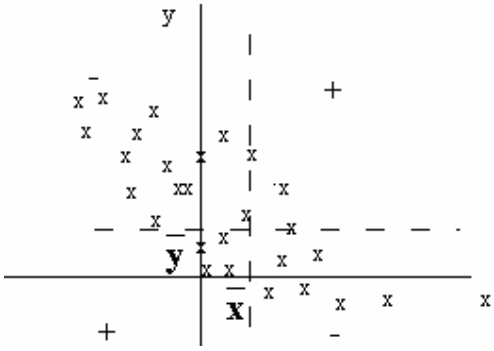



**2** – Se aos maiores valores de x estão associados os menores valores de y, então r<0.

Fazendo o raciocínio como no ponto anterior, verificamos que agora as parcelas são maioritariamente negativas, já que quando x é grande (superior à média dos x's), então y é pequeno (inferior à média dos y's). Assim, os produtos

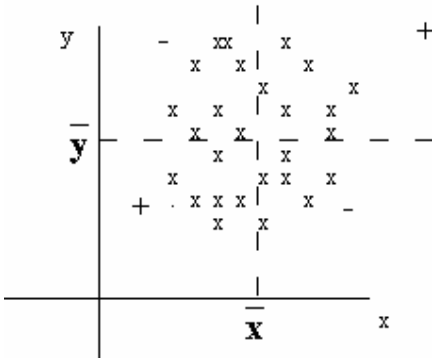
$$(x_i - \overline{x})(y_i - \overline{y})$$

são, de um modo geral, negativos.



**3** – Se não existe qualquer tipo de associação linear entre os x's e os y's, então r=0.

Neste caso tanto podem surgir produtos negativos, como positivos, distribuindo-se de forma mais ou menos equitativa. Então o valor de r vem próximo de zero.

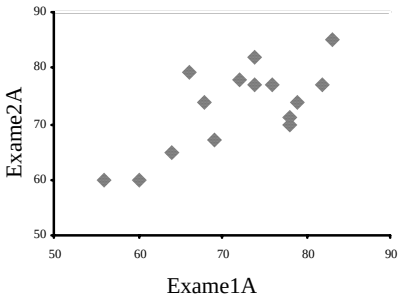


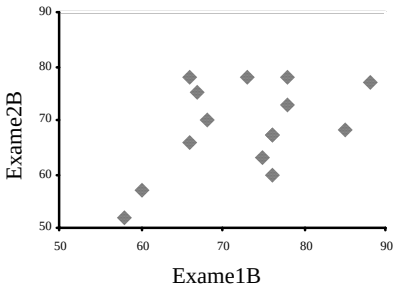

**Observação:** Dada a amostra ($x_1$, $x_2$, $x_3$, … , $x_n$), obtém-se a amostra "estandardizada" ou padronizada ($x_1^*$, $x_2^*$, $x_3^*$, … , $x_n^*$), subtraindo a cada elemento a média, isto é, *centrando* a amostra na origem, e dividindo pelo desvio padrão, ou seja, *reduzindo* os dados de forma a que o desvio padrão dos dados transformados venha igual a 1:

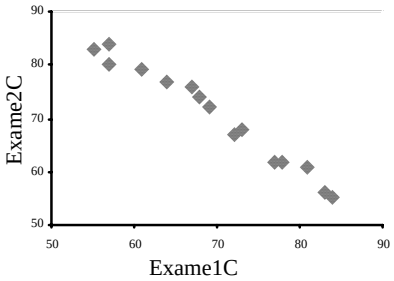
$$x_i^* = \frac{x_i - \bar{x}}{s_x}$$

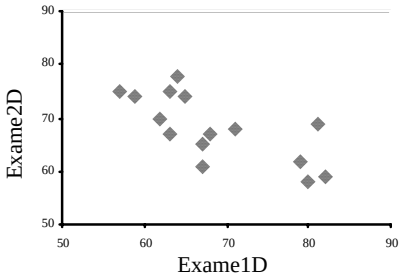
Exercício: Verifique que o coeficiente de correlação da amostra bivariada ($x_i$, $y_i$), i=1,…n, é a covariância da amostra padronizada correspondente.

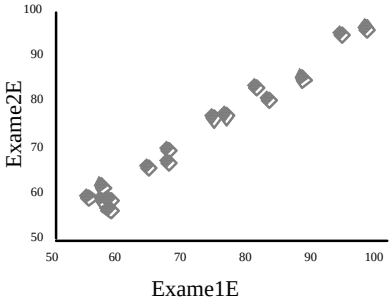
**Exemplo 10** (Rossman, 1996) **-** Considere os seguintes diagramas de dispersão correspondentes aos resultados de 2 exames de 6 classes (A-F).

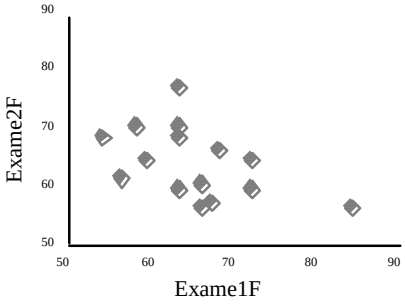




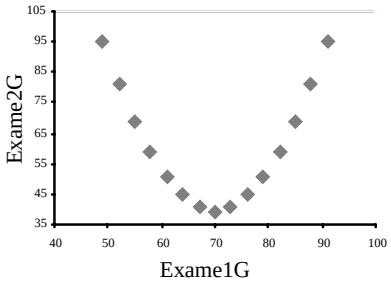

A visualização dos gráficos anteriores leva-nos a supor que entre os dois exames se possa admitir o seguinte tipo de associação:

| | Forte | Moderada | Fraca |
|---|---|---|---|
| Positiva | E | A | B |
| Negativa | C | D | F |

O cálculo do coeficiente de correlação, que se apresenta na tabela seguinte completa a informação da tabela anterior:

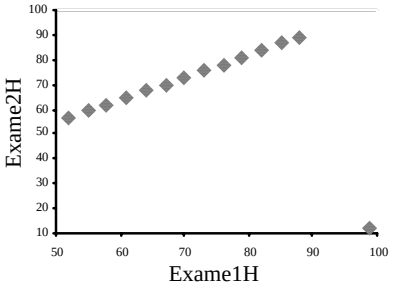
| Classe | Correlação |
|---|---|
| A | 0.71 |
| B | 0.47 |
| C | -0.99 |
| D | -0.72 |
| E | 0.99 |
| F | -0.47 |

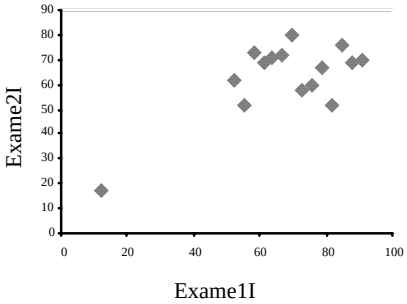
Considere agora a seguinte representação correspondente aos dados de uma classe G:




Como se verifica, existe uma forte associação entre os valores do exame 1 e os valores do exame 2. *Surpreendentemente* ao calcular o coeficiente de correlação obtemos o valor 0! Mas será assim tão surpreendente? Não, se nos lembrarmos que o que o coeficiente de correlação mede é o grau de associação linear e não outro tipo de associação, como a associação curvilínea, presente nos dados da representação anterior.

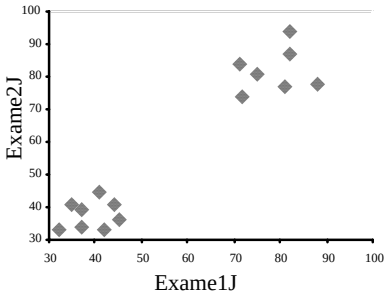
Considere agora as duas representações correspondentes às notas obtidas pelas classes H e I:




O valor para o coeficiente de correlação é respectivamente 0.04 e 0.70 para as classes H e I, o que continua a ser surpreendente! Repare-se que relativamente à classe H todos os pares menos 1 seguem um padrão linear, tendo-se obtido para o coeficiente de correlação um valor próximo de zero, enquanto que para a classe I, em que os valores se apresentam mais ou menos dispersos, obtivemos um valor relativamente alto. No entanto, se retirarmos a cada um dos conjuntos de dados anteriores o "outlier", já o valor do coeficiente de correlação passa para 0.9997 e 0.13, respectivamente para as classes H e I. O exemplo que acabámos de dar mostra que o coeficiente de correlação não é uma medida *resistente*, já que é muito influenciado pelos "outliers". Este facto não é de estranhar, já que no cálculo do coeficiente de correlação entramos com a média, que já vimos ser uma medida não resistente.

Finalmente consideremos o seguinte diagrama de dispersão correspondente à classe J:




Da análise da representação anterior verificamos existirem dois grupos distintos de alunos: uns muito bons e outros muito maus. Embora para cada um dos grupos se verifique uma ligeira tendência para uma associação positiva, o facto é que o valor do coeficiente de correlação é 0.95, bem superior ao valor que seria de esperar.

Os exemplos que acabámos de ver, elucidam-nos sobre as limitações do coeficiente de correlação como medida de associação entre duas variáveis.

Antes de calcular e tentar interpretar o coeficiente de correlação entre duas variáveis, construa um diagrama de pontos. Não esqueça que o coeficiente de correlação só mede a intensidade com que duas variáveis se associam linearmente, pelo que se a representação gráfica não mostrar evidência de associação linear, não tem sentido calculá-lo.

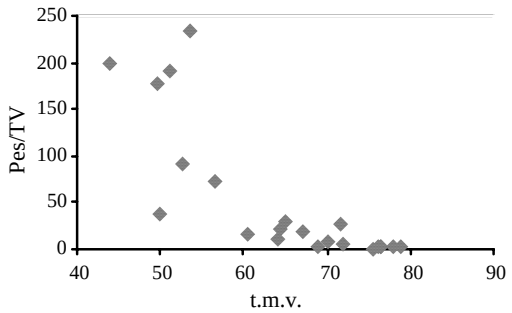
Um outro aspecto que não pode deixar de ser referido quando estamos perante uma correlação forte entre duas variáveis, é que isso não significa necessariamente uma relação de causa-efeito.

Não confundir correlação com relação causa-efeito. Um diagrama de pontos e uma correlação não provam a existência de uma relação causa-efeito. Podem existir outras variáveis, que não são estudadas, mas influenciam as que estão a ser estudadas e que são conhecidas como "lurking variables" (temos dificuldade em arranjar uma tradução adequada, pelo que vamos utilizar o termo "variáveis perturbadoras").

**Exemplo 11** (Rossman, 1996) - A seguinte tabela apresenta para um conjunto de 22 países, o tempo médio de vida e o número de pessoas por aparelho de televisão:

| País | t.m.v. | Pes/TV | País | t.m.v. | Pes/TV |
|---|---|---|---|---|---|
| Angola | 44 | 200 | México | 72 | 6.6 |
| Austrália | 76.5 | 2 | Marrocos | 64.5 | 21 |
| Cambodja | 49.5 | 177 | Paquistão | 56.5 | 73 |
| Canadá | 76.5 | 1.7 | Russia | 69 | 3.2 |
| China | 70 | 8 | África Sul | 64 | 11 |
| Egipto | 60.5 | 15 | Sri Lanka | 71.5 | 28 |
| França | 78 | 2.6 | Uganda | 51 | 191 |
| Haiti | 53.5 | 234 | ReinoUnido | 76 | 3 |
| Iraque | 67 | 18 | EUA | 75.5 | 1.3 |
| Japão | 79 | 1.8 | Vietnam | 65 | 29 |
| Madagáscar | 52.5 | 92 | Yemen | 50 | 38 |

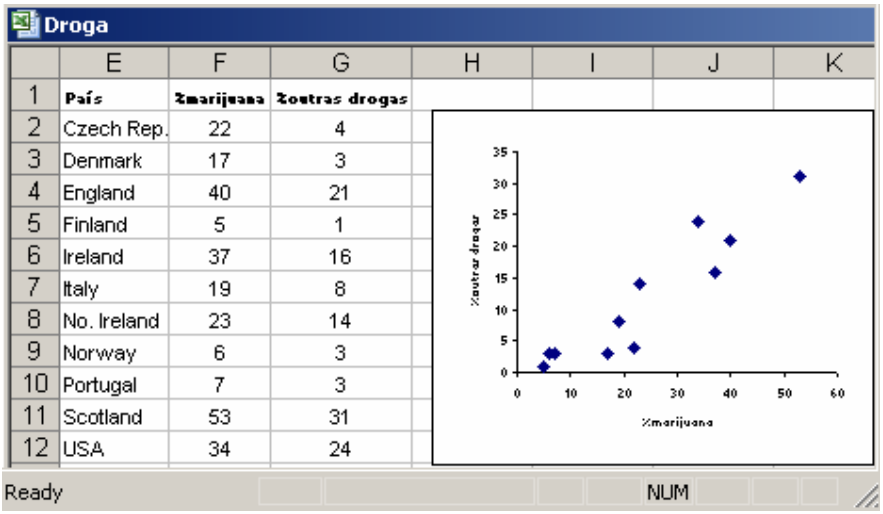
O valor do coeficiente de correlação entre as variáveis t.m.v e Pes/TV é igual a -0.80, o que significa uma forte correlação negativa entre o tempo médio de vida e o número de pessoas por aparelho de TV, ou seja, quanto maior for o número de pessoas por aparelho de TV, menor é o tempo médio de vida. Será que então se pode aumentar o tempo médio de vida da população de um país, aumentando o número de aparelhos de TV? Seria ridículo pensar desta maneira, pois este é um exemplo em que sobressai que não se pode admitir uma relação de causa-efeito. Obviamente existem outras variáveis não observadas -*variáveis perturbadoras* - relacionadas com o nível de vida na população, que provocam alterações nas duas variáveis que estamos a estudar e que explicam a forte correlação verificada. O diagrama de dispersão das variáveis estudadas tem o seguinte aspecto:



## Utilização do Excel na construção do diagrama de pontos e no cálculo da correlação

**Exemplo** (De Veaux et al, 2004) – Foi feito um inquérito nos Estados Unidos e em 10 países europeus, para determinar a percentagem de jovens que usaram marijuana e outras drogas, cujos resultados se apresentam na seguinte tabela.

a) Construa um diagrama de pontos dos dados
b) Calcule o coeficiente de correlação entre as percentagems de jovens que usaram marijuana e outras drogas
c) Será que os resultados confirmam que a marijuana é "uma porta de entrada para a droga", isto é, o uso da marijuana conduz ao uso de outras drogas? Explique.



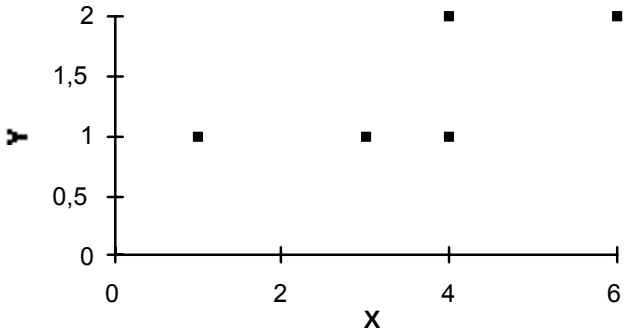

a) Para construir o diagrama de pontos anterior, selccionar as duas colunas com os dados, F2:G12, e de seguida :
   - Seleccionar, no menu, o ícone Chart;
   - Na caixa de diálogo que aparece, seleccionar a opção *XY (Scatter)* e o primeiro *sub-tipo*;
   - Clicar no botão *Next*, duas vezes, para passar dois passos, até aparecer uma caixa de diálogo, que apresenta várias opções: Em *Legend*, desactivar a legenda e em *Titles*, acrescentar o título no eixo dos Y's e no eixo dos X's, e carregar em *Finish*.

b) Como a representação gráfica mostra a existência de associação linear entre as variáveis % de marijuana e % de outras drogas, fomos calcular o coeficiente de correlação. Para isso utilizámos a função CORREL do Excel, que nos devolveu o valor 0.9341. Podemos dizer que existe uma forte associação positiva entre as variáveis em estudo.

c) Não podemos confundir correlação com uma relação de causa-efeito. Neste caso existirão, possivelmente, outras variáveis que predispõem os jovens ao consumo quer da marijuana, quer das outras drogas.

**Exercícios**

**1** - Considere o seguinte diagrama de dispersão:




Responda às seguintes questões:

a) A média dos x's está próxima de 1, 1.5 ou 3?

b) A média dos y's está próxima de 1, 1.5 ou 3?

c) Qual das variáveis apresenta maior variabilidade?

d) Calcule o coeficiente de correlação.

**2** – Numa Conservatória de Registo Civil recolheu-se informação sobre as idades do homem e da mulher de uma amostra de 20 casais. Os resultados foram os seguintes:

| Par | H | M | Par | H | M | Par | H | M | Par | H | M |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 20 | 19 | 6 | 38 | 29 | 11 | 26 | 27 | 16 | 36 | 32 |
| 2 | 25 | 25 | 7 | 35 | 36 | 12 | 32 | 31 | 17 | 19 | 19 |
| 3 | 26 | 24 | 8 | 27 | 26 | 13 | 54 | 56 | 18 | 29 | 20 |
| 4 | 22 | 23 | 9 | 42 | 29 | 14 | 45 | 42 | 19 | 32 | 32 |
| 5 | 28 | 24 | 10 | 25 | 25 | 15 | 28 | 29 | 20 | 45 | 43 |

Calcule o coeficiente de correlação entre as idades do homem e da mulher e interprete-o.

**3** – Durante vários anos consecutivos, e para uma determinada região, registou-se o consumo de gelados, em quilos, e o número de fogos, tendo-se verificado uma forte correlação entre estas duas variáveis. Será que o consumo de gelados provoca incêndios?

**4** – Mostram as estatísticas que existe uma correlação negativa entre o número de horas gastas a ver televisão e a desenvoltura na leitura. Será que ver televisão diminui a capacidade para a leitura?

### 3.4.2 – Associação de variáveis qualitativas

Quando anteriormente estudámos a associação de variáveis, utilizando nomeadamente o diagrama de dispersão e o coeficiente de correlação, assumimos que as variáveis eram de tipo *quantitativo*. Pode, no entanto, acontecer que estejamos interessados em estudar associação de variáveis de tipo *qualitativo* como, por exemplo, sexo e religião, ou então apesar de as variáveis serem de tipo quantitativo, procedemos a agrupamentos de forma que obtemos classes ou

categorias. Como vimos no capítulo 2, uma forma de apresentar os dados é utilizando tabelas de contingência. Vejamos, com um exemplo, uma forma de extrair informação a partir das tabelas de contingência:

**Exemplo 12** – Suponha que uma universidade decidiu estudar o seu corpo docente quanto ao estado civil e categoria profissional, tendo obtido os seguintes resultados:

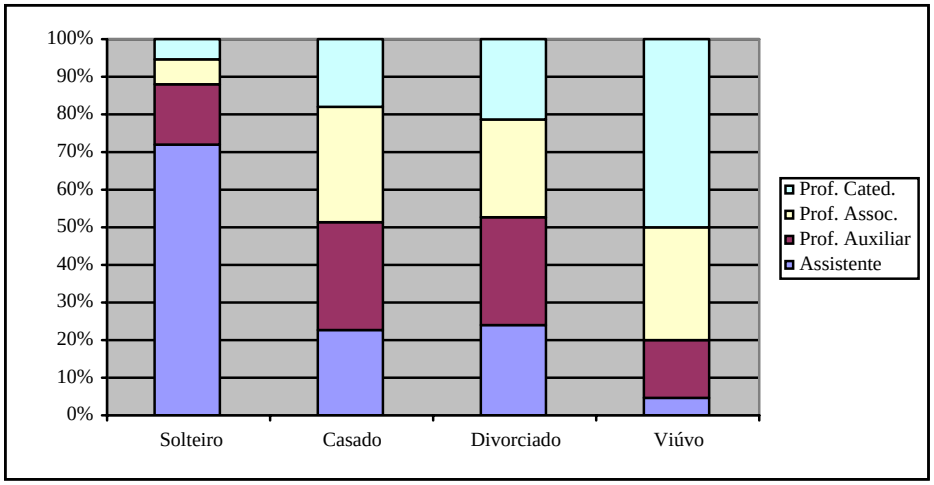
| Estado civil<br>Categoria | Solteiro | Casado | Divorciado | Viúvo | Total |
|---|---|---|---|---|---|
| Assistente | 111 | 43 | 10 | 1 | 165 |
| Prof. Auxiliar | 25 | 54 | 12 | 3 | 94 |
| Prof. Associado | 10 | 58 | 11 | 6 | 85 |
| Prof. Catedrático | 8 | 34 | 9 | 10 | 61 |
| Total | 154 | 189 | 42 | 20 | 405 |

Na última coluna do lado direito apresentamos os totais de linha, que corresponde à *distribuição* da variável "categoria profissional". Analogamente, na última linha estão apresentados os totais de coluna, que correspondem à *distribuição* da variável "estado civil". A estas distribuições chamamos *distribuições marginais* (precisamente por se apresentarem nas margens da tabela!). Estas distribuições apresentadas separadamente não nos dão informação sobre a associação entre as variáveis em estudo. Tão pouco essa informação pode ser dada pelo diagrama de dispersão ou pela correlação.

Uma forma de descrever a relação entre variáveis qualitativas é através do cálculo de percentagens convenientes. Consideremos a tabela seguinte, obtida a partir da tabela anterior, dividindo o valor de cada célula pelo total de coluna correspondente:

| Estado civil<br>Categoria | Solteiro | Casado | Divorciado | Viúvo | |
|---|---|---|---|---|---|
| Assistente | 0.721 | 0.228 | 0.238 | 0.050 | 0.407 |
| Prof. Auxiliar | 0.162 | 0.285 | 0.286 | 0.150 | 0.232 |
| Prof. Associado | 0.065 | 0.307 | 0.262 | 0.300 | 0.210 |
| Prof. Catedrático | 0.052 | 0.180 | 0.214 | 0.500 | 0.151 |
| Total | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

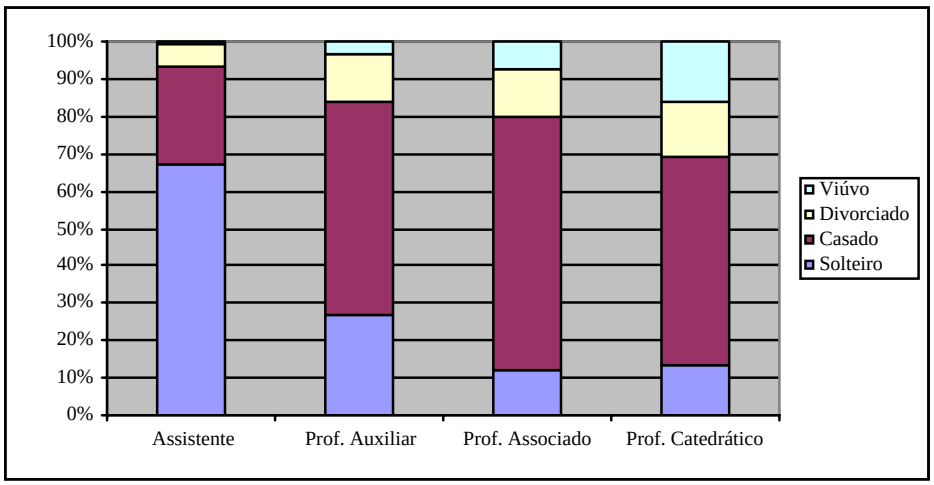
Nesta tabela apresentamos as *distribuições condicionais* da variável categoria profissional, relativamente às classes da outra variável estado civil. Temos assim que, por exemplo, nos solteiros a percentagem de assistentes é de aproximadamente 72%, enquanto que nos casados é de aproximadamente 23%. Estas distribuições condicionais podem ser visualizadas graficamente num diagrama de barras por segmentos, como se apresenta a seguir:




Se estivéssemos interessados nas distribuições condicionais da variável estado civil, condicional à variável categoria profissional, então a tabela a construir seria:

| Estado civil<br>Categoria | Solteiro | Casado | Divorciado | Viúvo | Total |
|---|---|---|---|---|---|
| Assistente | 0.673 | 0.261 | 0.061 | 0.006 | 1.001 |
| Prof. Auxiliar | 0.266 | 0.574 | 0.128 | 0.032 | 1.000 |
| Prof. Associado | 0.118 | 0.682 | 0.129 | 0.071 | 1.000 |
| Prof. Catedrát. | 0.131 | 0.557 | 0.148 | 0.164 | 1.000 |
| | 0.380 | 0.467 | 0.104 | 0.049 | 1.000 |

A leitura que se deve fazer desta tabela é semelhante à que se fez da tabela anterior, mas tendo em atenção que agora a variável que está a condicionar é a categoria profissional. Por exemplo pode obter-se a informação de que aproximadamente 67% dos assistentes são solteiros, enquanto que casados são cerca de 26%. O diagrama de barras por segmentos correspondente a estas distribuições marginais tem o seguinte aspecto:



Podemos finalmente estar interessados na *distribuição conjunta* das duas variáveis, e então em vez de recolher a informação a partir da primeira tabela constrói-se uma outra em que a frequência absoluta de cada célula é substituída pela frequência relativa, relativamente ao total de docentes, pois as frequências relativas são mais fáceis de comparar:

| Estado civil<br>Categoria | Solteiro | Casado | Divorciado | Viúvo | Total |
|---|---|---|---|---|---|
| Assistente | 0.274 | 0.106 | 0.025 | 0.002 | 0.407 |
| Prof. Auxiliar | 0.062 | 0.133 | 0.030 | 0.007 | 0.232 |
| Prof. Associado | 0.025 | 0.143 | 0.027 | 0.015 | 0.210 |
| Prof. Catedrático | 0.020 | 0.084 | 0.022 | 0.025 | 0.151 |
| Total | 0.380 | 0.467 | 0.104 | 0.049 | 1.000 |

Desta tabela imediatamente se conclui que, do pessoal docente, 3% são Professores Auxiliares e casados, enquanto que Assistentes e solteiros são mais de 27%.

**Paradoxo de Simpson**

Vimos na secção anterior que, por vezes, a interpretação do coeficiente de correlação não é imediata, nomeadamente devido ao facto de ser influenciado por variáveis perturbadoras, que podem ocasionar que, por exemplo, entre duas variáveis se obtenha uma forte correlação difícil de explicar, já que o que se esperaria seria uma correlação fraca, ou até de sentido diferente! Ora, o mesmo se passa na leitura das percentagens de uma tabela de contingência, que podem ocasionar interpretações menos correctas. Vejamos o seguinte exemplo:

**Exemplo 13** (Statistics, 1991) – Foi realizado estudo sobre admissão de candidatos na Universidade da Califórnia, tendo-se verificado que durante o período envolvido no estudo se candidataram 8442 homens e 4321 mulheres, tendo sido admitidos cerca de 44% dos homens e 35% das mulheres. Haverá discriminação sexual contra as mulheres? Admitindo que à partida não há razão para diferenciar profissionalmente os candidatos quanto ao sexo, os resultados obtidos mostram uma preferência dos supervisores, encarregados da selecção, pelo sexo masculino. Será verdade? Embora na admissão do pessoal estivessem envolvidos mais de 100 supervisores, vamos ver em particular o que se passou com os 6 maiores que seleccionaram cerca de um terço dos candidatos:

| | Homens | | Mulheres | |
|---|---|---|---|---|
| Supervisor | Número Candidatos | % admitidos | Número Candidatos | % admitidos |
| A | 825 | 62 | 108 | 82 |
| B | 560 | 63 | 25 | 68 |
| C | 325 | 37 | 593 | 34 |
| D | 417 | 33 | 375 | 35 |
| E | 191 | 28 | 393 | 24 |
| F | 373 | 6 | 341 | 7 |

Para cada supervisor, a percentagem de mulheres admitidas é sensivelmente igual à percentagem de homens admitidos, excepto para o supervisor A, que parece ter uma discriminação contra os homens! A maior diferença a favor dos homens verifica-se unicamente para o supervisor E e é unicamente de 4 pontos percentuais. Contudo, quando se considera na globalidade o conjunto de admitidos verifica-se que para os homens a percentagem é de cerca de 44% enquanto que para as mulheres é de cerca de 30%. Como explicar esta diferença de 14%? Esta situação é paradoxal, mas tem uma explicação:

1. Os dois primeiros supervisores eram mais permissivos e tiveram a candidatura de cerca de 50% dos homens.
2. Os outros quatro supervisores eram mais rígidos e tiveram a candidatura de cerca de 90% das mulheres.

Assim, os homens candidataram-se aos supervisores onde era mais fácil de entrar, enquanto que as mulheres fizeram o contrário. Existe aqui o efeito devido à escolha do supervisor que provoca uma interpretação enganadora quanto à variável sexo. Esta situação é conhecida como o paradoxo de Simpson. O paradoxo de Simpson diz respeito à inversão na direcção da associação quando os dados referentes a vários grupos são combinados para formarem um único grupo.

**Exemplo 14 - Um problema de saúde pública** (Tradução livre de um exemplo retirado do endereço **www.cawtech.freeserve.co.uk/Simpsons.2.html**) - O responsável pelo Departamento de Saúde (DS) de determinada região está a braços com um grave problema, que diz respeito a uma doença, conhecida como doença de Grott, frequentemente fatal, mas para a qual não existia, até à data, tratamento. Acontece que chegou a informação que teria sido descoberto um tratamento para a dita doença, havendo até pessoas que já o tomavam, acreditando na sua eficácia. O responsável do DS decidiu encomendar um estudo, cujos resultados foram os seguintes:

| | Não tratamento | Tratamento |
|---|---|---|
| Vivas | 108 | 153 |
| Moras | 123 | 120 |

Afinal o tratamento é útil, concluíu a comissão encarregada do estudo. Os dados indicam que a percentagem de pessoas vivas que fizeram o tratamento é de 56% (=153/273), superior à das pessoas vivas que não fizeram o tratamento, que é só de 46,7% (=108/231).

Conclusão: *Embora não sejam uns resultados espectaculares, vale a pena investir, apesar do tratamento ser bastante caro*, pensou o responsável pelo Departamento de Saúde.

Qual não foi a surpresa deste senhor, quando recebeu uma comissão de mulheres, colocando reticências ao investimento em causa, já que alegávam que o tratamento só beneficiaria os homens, uma vez que tinham verificado o que se passava com os dados referentes às mulheres e

estes indicávam até uma diminuição ligeira na percentagem de mulheres vivas, de entre as que tinham feito o tratamento:

| | Mulheres | |
|---|---|---|
| | Não tratamento | Tratamento |
| Vivas | 57 | 32 |
| Mortas | 100 | 57 |

Efectivamente a percentagem de mulheres vivas de entre as que não fizeram o tratamento é de 36,3%, enquanto que para as que fizeram o tratamento é de 36%! Esperáva-se assim que os homens fossem largamente beneficiádos, tendo em conta os dados da primeira tabela apresentada. Qual não foi o espanto, quando verificáram que, afinal, o tratamento também não beneficiáva os homens:

| | Homens | |
|---|---|---|
| | Não tratamento | Tratamento |
| Vivos | 51 | 121 |
| Mortos | 23 | 63 |

É mesmo verdade que o tratamento não é benéfico para o sexo masculino, já que a percentagem de homens vivos sem tratamento é de 69% (=51/74), contra os 66% (= 121/184) dos que fizeram tratamento.

Conclusão: *O tratamento é prejudicial tanto para os homens, como para as mulheres, embora seja benéfico para o pessoal em geral*!

Estava ainda o responsável do DS atarantado com estas conclusões, a reflectir sobre o que fazer, quando recebe a informção de que o marido da sua secretária tinha morrido com a doença de Grott. Não havia nada a fazer, era uma pessoa com a tensão arterial muito alta. Como se pode comprovar pelos dados seguintes, o tratamento em estudo tem um interesse limitado para os indivíduos de tensão alta, pois não consegue sequer uma percentagem de 50% de cura:

| | Homens tensão alta | |
|---|---|---|
| | Não tratamento | Tratamento |
| Vivos | 4 | 51 |
| Mortos | 6 | 57 |

Repare que a percentagem de vivos é de 40% (=4/10) para os que não seguiram o tratamento, contra 47% (=51/108) para os que seguiram o tratamento.

Já agora, o que se passará com os de tensão normal ou baixa? Vejamos os dados:

| | Homens tensão normal ou baixa | |
|---|---|---|
| | Não tratamento | Tratamento |
| Vivos | 47 | 70 |
| Mortos | 17 | 6 |

Também para estes é benéfico pois a percentagem de vivos é de 92% (=70/76).

Conclusão: *O tratamento é prejudicial aos homens, mas é benéfico para os que têm a tensão alta, e para os que têm a tensão normal ou baixa, é uma autêntica salvação!*

Ainda podemos aumentar a perplexidade do responsável do Departamento de Saúde se considerarmos as mulheres divididas em dois grupos, as jovens e as menos jovens:

| | Mulheres jovens | | Mulheres menos jovens | |
|---|---|---|---|---|
| | Não tratamento | Tratamento | Não tratamento | Tratamento |
| Vivos | 49 | 25 | 8 | 7 |
| Mortos | 19 | 8 | 81 | 49 |

Conclusão: Vimos anteriormente que *o tratamento não era benéfico para as mulheres, mas agora concluímos que é benéfico para as mulheres jovens, pois 76% (=25/33) das que receberam tratamento estão vivas, contra 72% (=49/68) das que não receberam tratamento*.

Depois disto o responsável pelo Departamento de Saúde meteu atestado médico.

Atenção – Quando se calculam proporções ou percentagens entre diferentes grupos, é necessário certificarmo-nos de que os grupos são comparáveis. Este problema do paradoxo de Simpson, foi assim denominado depois que o estatístico Simpson, num seminário em 1951, apresentou algumas fracções com propriedades surpreendentes e que são contrárias à intuição. Quando estamos a comparar duas variáveis, para as quais é possível estar associada uma terceira variável, a comparação deve ser feita para cada nível ou modalidade desta terceira variável, pois quando se comparam os dados para todos os níveis em conjunto, a direcção da associação pode vir invertida.

Os leitores interessados neste tema, encontram referências pesquisando na Internet o assunto "Simpson's Paradox". Além dos exemplos apresentados, encontram outros exemplos interessantes.

**Exercício**

1. Na sua cidade há duas clínicas A e B. O ministério da Saúde pretende tomar uma decisão de escolher uma destas clínicas para fazer parte do plano de saúde pública, pelo que fez um estudo sobre o sucesso em 5 tipos de operações realizadas nestas clínicas:

| | Clínica A | | | Clínica B | | |
|---|---|---|---|---|---|---|
| Tipo operação | Nº operações | Nº bem suc. | % sucesso | Nº operações | Nº bem suc. | % sucesso |
| A | 359 | 292 | .81 | 88 | 70 | .80 |
| B | 1836 | 1449 | .79 | 514 | 391 | .76 |
| C | 299 | 178 | .60 | 222 | 113 | .51 |
| D | 2086 | 434 | .21 | 86 | 12 | .14 |
| E | 149 | 13 | .09 | 45 | 2 | .04 |
| | 4729 | 2366 | .50 | 955 | 588 | .62 |

Repare que em todos os tipos de operações a clínica A tem maior sucesso, ainda que na globalidade a clínica B tenha uma maior percentagem de sucesso. Qual das clínicas escolheria?

**Capítulo 4**

# Regressão

## 4.1 - Introdução

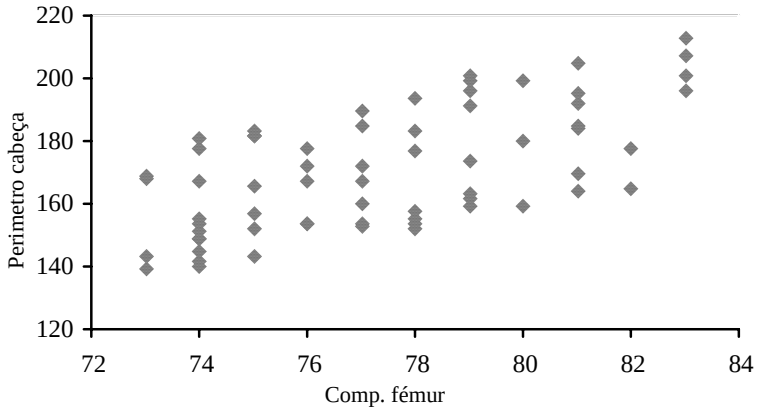
Como vimos no capítulo anterior a correlação mede o grau e o tipo – positivo ou negativo, da associação linear existente entre duas variáveis quantitativas. Quando o diagrama de dispersão realça a existência desta associação linear, então é possível resumir através de uma recta a forma como uma *variável resposta* y é influenciada por uma *variável explanatória* x – a essa recta damos o nome de *recta de regressão*.

Um modelo de *regressão* é um modelo matemático – equação, que descreve a relação entre duas ou mais variáveis. Se o estudo só incluir duas variáveis – a variável explanatória x e a variável resposta y, temos uma *regressão simples*. Se o modelo matemático utilizado for a equação de uma recta, então diz-se *regressão linear simples*

**Exemplo 1** – Os dados seguintes representam o comprimento do fémur, avaliado através de ecografia, de fetos humanos na 30ª semana de gestação (colunas encimadas com X) e o correspondente perímetro da cabeça à nascença (colunas encimadas com Y).

| X | Y | X | Y | X | Y |
|---|---|---|---|---|---|
| 73 | 168 | 76 | 172 | 79 | 163 |
| 73 | 169 | 76 | 154 | 79 | 159 |
| 73 | 143 | 76 | 167 | 79 | 174 |
| 73 | 139 | 76 | 154 | 79 | 191 |
| 74 | 140 | 76 | 178 | 79 | 196 |
| 74 | 178 | 77 | 160 | 80 | 180 |
| 74 | 149 | 77 | 185 | 80 | 199 |
| 74 | 167 | 77 | 172 | 80 | 159 |
| 74 | 155 | 77 | 167 | 81 | 170 |
| 74 | 149 | 77 | 153 | 81 | 184 |
| 74 | 154 | 77 | 154 | 81 | 192 |
| 74 | 145 | 77 | 190 | 81 | 185 |
| 74 | 151 | 78 | 152 | 81 | 164 |
| 74 | 181 | 78 | 158 | 81 | 195 |
| 74 | 142 | 78 | 154 | 81 | 205 |
| 75 | 166 | 78 | 194 | 82 | 165 |
| 75 | 182 | 78 | 183 | 82 | 178 |
| 75 | 143 | 78 | 155 | 83 | 201 |
| 75 | 183 | 78 | 177 | 83 | 207 |
| 75 | 157 | 79 | 201 | 83 | 196 |
| 75 | 152 | 79 | 162 | 83 | 213 |
| 75 | 182 | 79 | 199 | | |

A representação num diagrama de dispersão dos valores observados para o par de variáveis (comprimento do fémur, perímetro da cabeça) tem o seguinte aspecto:




Na representação anterior verifica-se uma certa tendência (linear) para que à medida que o comprimento do fémur aumente, também aumente o perímetro da cabeça.

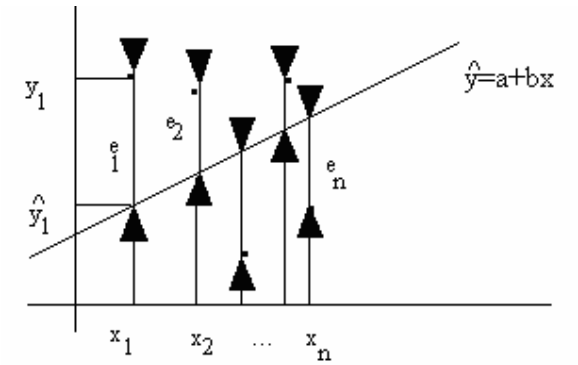
## 4.2 – Recta dos mínimos quadrados

Quando estamos numa situação análoga à anterior em que temos um conjunto de dados $(x_i,y_i)$, i=1,…,n, que seguem um padrão linear, pode ter interesse **ajustar** uma recta da forma

$$\mathbf{y = a + bx}$$

que dê a informação de como se reflectem em y, as mudanças processadas em x. Quando os dados não se dispõem segundo uma linha recta, então há transformações adequadas de forma a linearizá-los.

Um dos métodos mais conhecidos de ajustar uma recta a um conjunto de dados, é o *método dos mínimos quadrados*, que consiste em determinar a recta que minimiza a soma dos quadrados dos desvios (ou erros) entre os verdadeiros valores das ordenadas e os obtidos a partir da recta, que se pretende ajustar.




Esta técnica, embora muito simples, é pouco resistente, já que é muito sensível a dados "estranhos" - valores que se afastam da estrutura da maioria. Efectivamente, quando se pretende minimizar

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

facilmente se obtêm os estimadores do declive e da ordenada na origem, que são respectivamente:

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad a = \bar{y} - b\bar{x}$$

O facto de dependerem, de forma muito estrita, de todos os pontos (além de dependerem da média, que como vimos é uma medida não resistente), torna a recta muito vulnerável aos tais valores "estranhos", pelo que é necessário proceder a uma análise prévia do diagrama de dispersão para ver se não existem alguns desses elementos – "outliers". A expressão que dá o declive da *recta dos mínimos quadrados*, ou também chamada *recta de regressão*, pode ser apresentada com outro aspecto, mais útil para efeitos de cálculo:

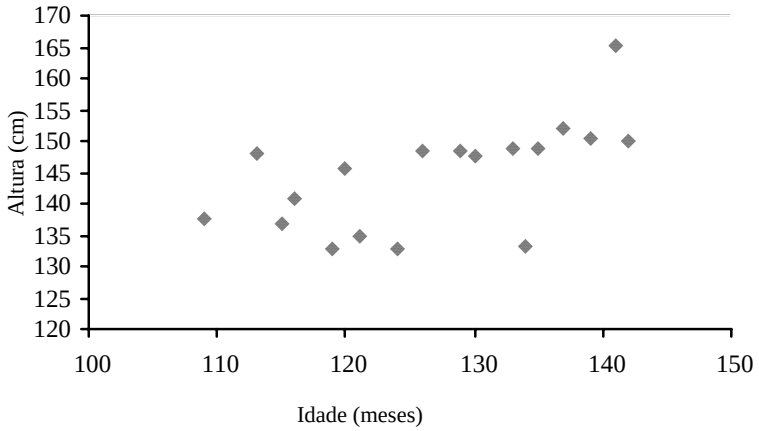
$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

Para exemplificar o cálculo dos coeficientes da recta de regressão consideremos o exemplo seguinte:

**Exemplo 2** - Os dados da tabela seguinte representam a idade e a altura das crianças de uma escola privada.

| Criança | Idade(meses) | Altura(cm) |
|---|---|---|
| 1 | 109 | 137.6 |
| 2 | 113 | 147.8 |
| 3 | 115 | 136.8 |
| 4 | 116 | 140.7 |
| 5 | 119 | 132.7 |
| 6 | 120 | 145.4 |
| 7 | 121 | 135.0 |
| 8 | 124 | 133.0 |
| 9 | 126 | 148.5 |
| 10 | 129 | 148.3 |
| 11 | 130 | 147.5 |
| 12 | 133 | 148.8 |
| 13 | 134 | 133.2 |
| 14 | 135 | 148.7 |
| 15 | 137 | 152.0 |
| 16 | 139 | 150.6 |
| 17 | 141 | 165.3 |
| 18 | 142 | 149.9 |

Construindo o diagrama de dispersão




verifica-se a existência de uma certa associação linear entre a idade e a altura, pelo que vamos construir a recta de regressão da altura na idade. Exemplificamos a seguir a forma de fazer os cálculos:

| Criança | x | $x^2$ | y | xy |
|---|---|---|---|---|
| 1 | 109 | 11881 | 137.6 | 14998.4 |
| 2 | 113 | 12769 | 147.8 | 16701.4 |
| 3 | 115 | 13225 | 136.8 | 15732.0 |
| 4 | 116 | 13456 | 140.7 | 16321.2 |
| 5 | 119 | 14161 | 132.7 | 15791.3 |
| 6 | 120 | 14400 | 145.4 | 17448.0 |
| 7 | 121 | 14641 | 135.0 | 16335.0 |
| 8 | 124 | 15376 | 133.0 | 16492.0 |
| 9 | 126 | 15876 | 148.5 | 18711.0 |
| 10 | 129 | 16641 | 148.3 | 19130.7 |
| 11 | 130 | 16900 | 147.5 | 19175.0 |
| 12 | 133 | 17689 | 148.8 | 19790.4 |
| 13 | 134 | 17956 | 133.2 | 17848.8 |
| 14 | 135 | 18225 | 148.7 | 20074.5 |
| 15 | 137 | 18769 | 152.0 | 20824.0 |
| 16 | 139 | 19321 | 150.6 | 20933.4 |
| 17 | 141 | 19881 | 165.3 | 23307.3 |
| 18 | 142 | 20164 | 149.9 | 21285.8 |
| | $\sum 2283$ | $\sum 291331$ | $\sum 2601.8$ | $\sum 330900.2$ |

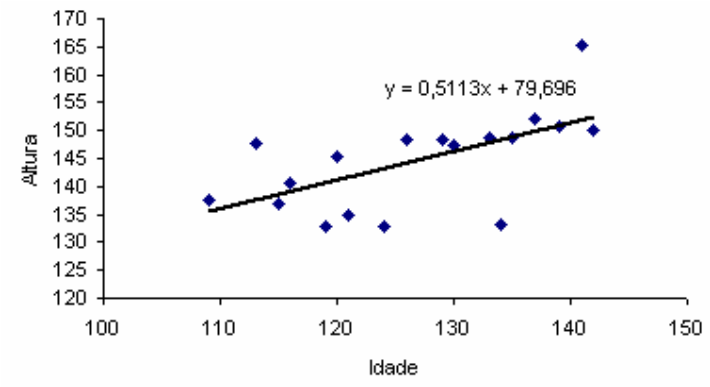
Utilizando as expressões anteriores para o cálculo dos coeficientes da recta, temos

$$b = \frac{18 \times 330900.2 - 2283 \times 2601.8}{18 \times 291331 - 2283^2} = 0.51$$

$$a = \frac{2601.8}{18} - 0.51 \times \frac{2283}{18} = 79.7$$

pelo que a recta de regressão é

$$\hat{y} = 79.7 + 0.51\,x$$




Uma utilização muito frequente da recta de regressão é na obtenção de *predições*. Por exemplo, se estivéssemos interessados em obter o valor para a altura de uma criança com 150 meses, bastaria substituir na equação da recta o valor de x por 150, obtendo-se um valor aproximado de 156 cm para a altura. E se pretendêssemos a altura de um jovem de 240 meses? Comente o resultado obtido.

**Resíduos**

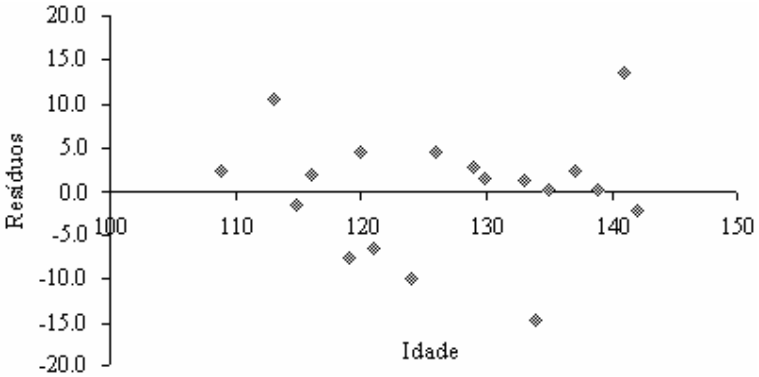
Uma forma de verificar se o modelo ajustado é bom é através dos resíduos, isto é, das diferenças entre os valores observados y e os valores ajustados $\hat{y}$:

resíduos = dados observados– valores ajustados

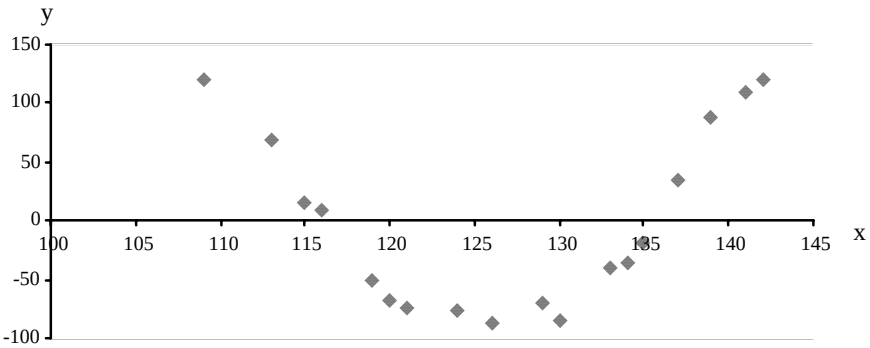
pois se estes não se apresentarem muito grandes, nem com nenhum padrão bem determinado, é sintoma de que o modelo que estamos a ajustar é bom. Chama-se a atenção para o facto de os resíduos gozarem da propriedade (esta propriedade é consequência imediata da forma como se obtêm as expressões para os estimadores a e b da recta de regressão) de a sua soma ser nula

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$$

pelo que uma forma elucidativa de os representar é considerar num diagrama de dispersão os pontos ($x_i$, $e_i$), visualizando-se os desvios positivos e negativos para cima e para baixo do eixo dos x's. No caso do exemplo tem-se



O facto de os desvios se apresentarem aleatoriamente para um e outro lado do eixo dos x´s é sintoma de que o modelo utilizado está correcto. Se por exemplo se tivesse obtido uma representação para os resíduos com o seguinte aspecto (depois de ajustado um modelo linear),




seríamos levados a concluir que o modelo que se deveria ajustar seria o não linear.

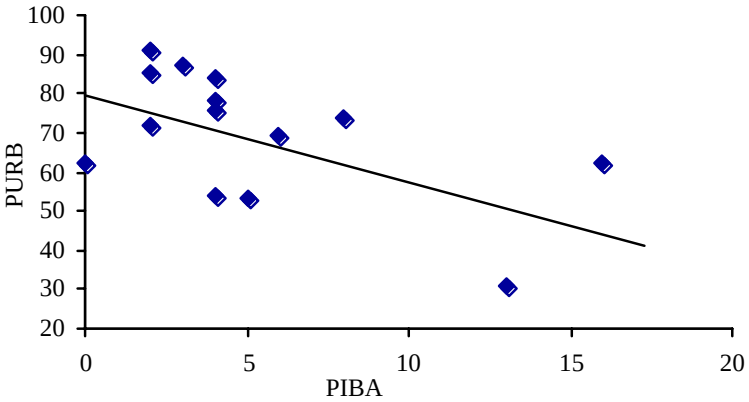
No contexto da regressão, *outliers* são valores com grandes resíduos. Se uma observação não conduzir a um grande resíduo, mas tiver grande influência na recta dos mínimos quadrados diz-se que é uma *observação influente*. Assim, um aspecto sobre a recta de regressão que convém não descurar, e já falado quando iniciámos o seu estudo, prende-se com o facto de ser *não resistente*, pois é muito influenciada por valores perturbadores O seguinte exemplo ilustra este facto:

**Exemplo 3** - Para alguns países da Europa, considerámos alguns indicadores económicos, nomeadamente o PIBA (produto interno bruto, originado pela agricultura) e o PURB (percentagem de população urbana):

| País | PIBA | PURB | País | PIBA | PURB |
|---|---|---|---|---|---|
| Alemanha | 2 | 85 | Grécia | 16 | 62 |
| Áustria | 4 | 54 | Holanda | 4 | 76 |
| Bélgica | 2 | 72 | Itália | 6 | 69 |
| Dinamarca | 4 | 84 | Noruega | 5 | 53 |
| Espanha | 8 | 74 | Portugal | 13 | 31 |
| Finlândia | 0 | 62 | Reino Unido | 2 | 91 |
| França | 4 | 78 | Suécia | 3 | 87 |

A recta dos mínimos quadrados entre as variáveis consideradas é

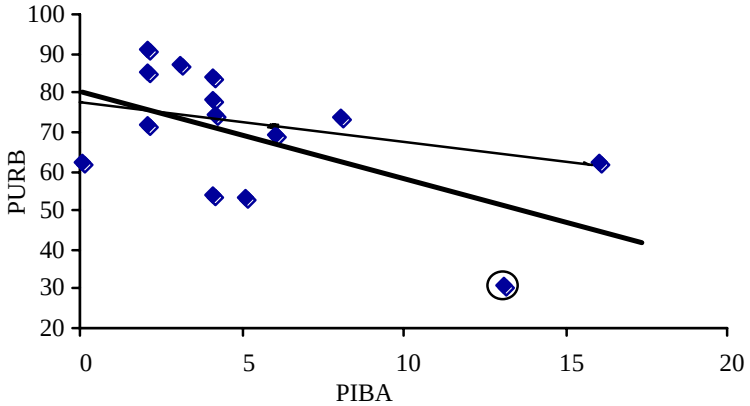
$$\hat{y} = 80.283 - 1.999\,x$$



Se retirarmos aos dados o ponto correspondente a Portugal, que juntamente com a Grécia sobressaem de entre os restantes no que diz respeito ao PIBA, obtém-se a seguinte equação para a recta

$$\hat{y} = 77.308 - 0.967\,x$$

Representando as duas rectas no mesmo gráfico, verifica-se a influência provocada por um único ponto:




Imediatamente se verifica que a inclinação da recta agora considerada é bastante mais pequena do que a que se obtém quando se consideram todos os pontos. A observação correspondente a Portugal diz-se *influente*.

Para obviar a este problema, utiliza-se a técnica da *recta resistente*, que recorre às medianas, que já vimos serem medidas resistentes. É um processo que consiste, basicamente, em dividir o conjunto dos n pontos $(x_i,y_i)$ , i=1,…,n, em três grupos, usar a mediana de cada grupo como ponto representativo do grupo e obter a recta ajustada, a partir dos três pontos (Hoaglin et al. 1983).

### Utilização do Excel na construção da recta de regressão

Para construir uma recta de regressão, deve-se começar por construir o diagrama de pontos. Caso haja evidência de haver associação linear, então vai-se ajustar a recta de regressão. Para isso seleccione o Diagrama de pontos e no menu, em *Chart*, seleccione *Add Trendline* e a opção *Linear*. Ainda na janela de *Add Trendline*, seleccione *Options e Display equation on chart*. A recta de regressão do exemplo 2 foi obtida por este processo.

### Exercícios

**1**.Suponha que um economista está interessado em estudar a relação entre as despesas mensais com a alimentação e os rendimentos mensais das famílias portuguesas. Obviamente que as despesas mensais com a alimentação dependem de vários factores tais como a dimensão do agregado familiar, os gostos dos elementos do agregado, além do rendimento. Como estamos interessado num modelo de regressão simples vamos considerar unicamente como variável explanatória o rendimento. Recolheu-se informação sobre 15 famílias, tendo-se obtido os resultados seguintes:

| Rendimento | Despesas |
|---|---|
| 495 | 110 |
| 340 | 85 |

| | |
|---|---|
| 260 | 80 |
| 450 | 100 |
| 540 | 120 |
| 356 | 90 |
| 250 | 85 |
| 290 | 80 |
| 420 | 110 |
| 560 | 120 |
| 380 | 110 |
| 270 | 90 |
| 330 | 85 |
| 420 | 120 |
| 360 | 115 |

a) Represente as observações num diagrama de dispersão.
b) A representação anterior sugere a existência de alguma relação linear entre as variáveis em estudo?
c) Se na alínea anterior a sua resposta foi afirmativa, obtenha a expressão que traduz essa relação. Interprete os coeficientes da recta obtida.
d)Obtenha uma estimativa para os gastos mensais com a alimentação de uma família cujos rendimentos são de 300 contos mensais.

**2**. A seguinte tabela apresenta, para um conjunto de animais, o tempo médio de vida (em anos) e período de gestação (em dias) (Rossman, 1996):

| Animal | Gestação | Longevidade | Animal | Gestação | Longevidade |
|---|---|---|---|---|---|
| Burro | 365 | 12 | Porco da Guiné | 68 | 4 |
| Baboon | 187 | 20 | Hipopótamo | 238 | 25 |
| Urso preto | 219 | 18 | Cavalo | 330 | 20 |
| Urso cinzento | 225 | 25 | Canguru | 42 | 7 |
| Urso polar | 240 | 20 | Leopardo | 98 | 12 |
| Castor | 122 | 5 | Leão | 100 | 15 |
| Búfalo | 278 | 15 | Macaco | 164 | 15 |
| Camelo | 406 | 12 | Veado | 240 | 12 |
| Gato | 63 | 12 | Rato | 21 | 3 |
| Chimpanzé | 231 | 20 | Opossum | 15 | 1 |
| Esquilo chipmuk | 31 | 6 | Porco | 112 | 10 |
| Vaca | 284 | 15 | Puma | 90 | 12 |
| Gamo | 201 | 8 | Coelho | 31 | 5 |
| Cão | 61 | 12 | Rinoceronte | 450 | 15 |
| Elefante | 645 | 40 | Leão marinho | 350 | 12 |
| Alce | 250 | 15 | Carneiro | 154 | 12 |
| Raposa | 52 | 7 | Esquilo | 4 | 10 |
| Girafa | 425 | 10 | Tigre | 105 | 16 |
| Cabra | 151 | 8 | Lobo | 63 | 5 |
| Gorila | 257 | 20 | Zebra | 365 | 15 |

a) Obtenha a recta dos mínimos quadrados que lhe permita estimar a longevidade a partir do tempo de gestação.
b) Interprete os coeficientes da recta dos mínimos quadrados
c) Represente graficamente os resíduos
d) Algum dos animais é claramente um outlier tanto em longevidade, como em tempo de gestação?
e) Relativamente ao animal considerado na alínea anterior, calcule o seu resíduo. Verifique se é substancialmente maior que os resíduos dos outros animais.
f) Qual dos animais tem o maior resíduo em valor absoluto? O seu período de gestação é maior ou menor do que se esperaria para um animal com a sua longevidade?
g) Retire a girafa dos seus dados e recalcule a recta dos mínimos quadrados. Compare as duas.
h) Faça o mesmo que na alínea anterior, mas agora com o elefante. Conclua das duas alíneas anteriores se, no contexto da regressão, alguma das observações consideradas é *influente* ou *outlier*.

**Capítulo 5**

# Probabilidade

## 5.1 – Introdução

Todos os dias somos confrontados com situações, que nos conduzem a utilizar, intuitivamente, a noção de **Probabilidade**. Nos mais variados aspectos da nossa vida, está presente a incerteza:

- dizemos que existe uma pequena probabilidade de ganhar o totoloto;
- dizemos que existe uma grande probabilidade de chover num dia carregado de nuvens;
- o político interroga-se sobre qual a probabilidade de ganhar as próximas eleições;
- o aluno interroga-se sobre qual a probabilidade de obter positiva num teste de perguntas com resposta múltipla, para o qual não estudou e responde sistematicamente ao acaso;
- o médico pretende saber se um medicamento novo tem maior probabilidade de cura que o medicamento habitual, para tratar determinada doença;
- o comerciante pretende saber se deve rejeitar um determinado carregamento de material, pois ao verificar um certo número de peças, encontrou uma determinada percentagem de defeituosas;
- o fabricante desejaria saber se um produto que pretende lançar no mercado, terá uma boa probabilidade de aceitação;
- o corretor da bolsa interroga-se sobre se será provável que umas acções que tem em vista, aumentem de cotação.

Embora não saibamos, para já, atribuir um valor numérico às probabilidades de realização dos acontecimentos envolvidos nos exemplos anteriores, há situações em que não temos dúvidas nessa atribuição. Por exemplo, ninguém hesita em afirmar que a probabilidade de um bebé nascer com dentes é igual a zero, assim como também não terá dúvida em dizer que é igual a 1 a probabilidade de num dia em que está a chover, haver nuvens! Por outro lado, quando se pretende tomar uma decisão ao acaso, para a qual existem duas opções, e não se sabe qual escolher, também é usual tomar a decisão mediante o resultado da saída de cara ou coroa, no lançamento de uma moeda ao ar, pois existe a convicção que a probabilidade de sair cara ou coroa são iguais a 1/2.

No dia a dia é comum atribuirmos probabilidades a determinados acontecimentos. Ao fazer isto, não estamos mais que a exprimir o nosso grau de convicção na realização desses acontecimentos. Podíamos então ser tentados a definir probabilidade de um determinado acontecimento como uma medida da convicção que temos na realização desse acontecimento. Mas claro, não nos podemos ficar por aqui. Este conceito tão simples só por si é demasiado precário para ser útil à Ciência (Graça Martins et al, 1997). Há necessidade de ir muito mais longe, já que não havendo mais do que meras conjecturas e convicções, diferentes com certeza de indivíduo para indivíduo, e quantas vezes incoerentes, não é possível fazer teoria. Há assim

necessidade de saber como quantificar aquela "medida de convicção" relativamente a qualquer acontecimento. Se em certas situações (como a relacionada com o lançamento de uma moeda) não temos dificuldade, há outras em que isso já se não nos afigura simples, ou por falta de informação, ou por mera incapacidade devido, por exemplo, à própria complexidade de que o acontecimento se reveste. Sabemos, se não por convicção, pelo menos pela própria experiência, que a probabilidade de nos sair o totoloto na próxima vez que jogarmos é extremamente pequena. Mas, quantas pessoas que não tenham estudado cálculo das probabilidades são capazes de atribuir um número a essa probabilidade? Já em face de um dado equilibrado, somos levados a dizer que a probabilidade de sair um 5 num lançamento é 1/6. Porque é que fazemos tal afirmação? Somos, no entanto, capazes de ficar perplexos quando alguém nos afirma que estudos estatísticos indicam que a probabilidade de contrair cancro de pulmão, se se fumar mais de 20 cigarros por dia, é de 7%. Com que base é que se pode fazer uma afirmação desta natureza?
Digamos que, com os dois exemplos apresentados, quantificámos a probabilidade de um acontecimento por dois processos distintos. No segundo caso, a quantificação da probabilidade de contrair cancro de pulmão se se fumar mais de 20 cigarros, foi feita recorrendo à experiência, identificando empiricamente a probabilidade de um acontecimento com a frequência relativa com que esse acontecimento se observa numa amostra representativa da população em estudo. Em termos estatísticos "estimámos" a probabilidade (desconhecida) da realização de um acontecimento pela frequência relativa com que esse acontecimento se verifica. No primeiro caso, o do dado equilibrado, o raciocínio é feito com base no facto de haver uma possibilidade em 6 de, ao lançar o dado uma vez, se observar a face 5. Não precisámos da experiência para quantificar a probabilidade, já que estamos a admitir o *pressuposto da simetria ou de equilíbrio* (este pressuposto da simetria é a base para a definição de probabilidade segundo o "conceito clássico" ou de Laplace, de que falaremos posteriormente), isto é, estamos a admitir que devido à simetria física do dado, não temos razão para atribuir probabilidade diferente à saída de cada face.

Imaginemos, no entanto, que estávamos a jogar um determinado jogo que obrigava ao lançamento de um dado e que a saída da face 5 implicava um bónus. Depois de jogarmos um grande número de vezes descobríamos que a face 5 quase nunca saía. O nosso senso comum levava-nos a supor que "algo estava errado com o dado". Como poderíamos averiguar isso? Lançando o dado um grande número de vezes, digamos n, e calculando a frequência relativa da realização do acontecimento de interesse, isto é, "saída de um 5". Estimávamos assim a probabilidade de no lançamento daquele dado sair a face 5. A intuição diz-nos que se não houver nada de errado com o dado, este valor deve flutuar à volta de 0.166(6).

A palavra probabilidade está presente sempre que estivermos perante *um fenómeno aleatório*, isto é, um fenómeno para o qual não sabemos de antemão o que vai acontecer, na próxima repetição, mas para o qual se admite uma *certa regularidade a longo termo*, ou seja, para um grande número de repetições do fenómeno. Esta regularidade estatística é utilizada para definir a probabilidade

segundo o "conceito frequencista", de que falaremos a seguir. Como veremos, é uma aproximação conceptual da probabilidade, muito utilizada, mas limitativa, na medida em que só permite definir a probabilidade de acontecimentos que se possam repetir um grande número de vezes nas mesmas condições.

Fenómenos aleatórios – são fenómenos cujos resultados individuais são incertos, mas para os quais se admite uma regularidade a longo termo, possibilitando a obtenção de um padrão genérico de comportamento.

Associados às seguintes experiências ou situações temos os seguintes exemplos de fenómenos, considerados aleatórios:

- Chave do totoloto em cada semana;
- Resposta de uma doença a um tratamento feito com determinado medicamento;
- Estado do tempo no dia seguinte;
- Comportamento dos eleitores nas próximas eleições legislativas;
- Comportamento de um aluno no exame de resposta múltipla, para o qual não estudou;
- Comportamento do mercado perante um produto novo para lavar a roupa;
- Etc.

É importante apercebermo-nos do que é que significa a regularidade a longo termo de que falámos anteriormente.

*Será que o acaso pode ser governado*? Então não estamos a admitir que a longo termo é possível obter um padrão genérico de comportamento do fenómeno aleatório?

Efectivamente, quando observamos o fenómeno em estudo um número suficientemente grande de vezes verifica-se um comportamento que pode ser modelado, isto é podemos arrranjar um modelo para exprimir a aleatoriedade. Mas atenção! Esta regularidade não existe a não ser a longo termo!

Na situação comum do lançamento de uma moeda ou de um dado, não podemos dizer qual a face que sai no próximo lançamento. No entanto se lançarmos a moeda ou o dado um número razoável de vezes, esperamos que aproximadamente metade das vezes saia cara e aproximadamnete um sexto das vezes saia a face 1 do dado. Suponha agora que lança a moeda 8 vezes e que obteve a seguinte sequência (representamos a cara por F e a coroa por C):

C, F, C, C, F, F, F, F

Se lançar novamente a moeda, o que é que espera que saia? Embora lhe apetecesse dizer que no próximo lançamento é mais provável que saia coroa (C), para equilibrar o número de caras com o número de coroas, na verdade no próximo lançamento tanto pode sair cara como coroa, já que os sucessivos lançamentos da moeda são independentes uns dos outros (a moeda não tem memória…).

**Exemplo 1** (adaptado de Moore, 1997) – A regularidade a longo termo se não for bem compreendida, pode acarretar alguns dissabores! Foi o que aconteceu com aquele casal que tinha

planeado ter 4 filhos. Depois de nascerem 4 raparigas, e na expectativa de terem um rapazinho, ainda tentaram mais 3 vezes e ficaram com uma linda equipa de 7 raparigas! Depois destas 7 raparigas o médico assegurou-lhes que era praticamente certo que o bebé seguinte fosse rapaz. Infelizmente para este casal, os fenómenos aleatórios que consistem em ter mais uma criança ou lançar mais uma vez a moeda, são idênticos. Efectivamente 8 raparigas de seguida, é muito improvável, mas uma vez nascidas 7 raparigas, não é de todo improvável que o próximo bebé fosse rapariga – e era!

O objectivo da Teoria da Probabilidade é o estudo dos fenómenos aleatórios, através de *modelos matemáticos*, a que chamamos modelos probabilísticos.

**Será possível fazer Estatística sem utilizar a Probabilidade?**

De um modo geral não! A maior parte das vezes em que é necessário utilizar técnicas estatísticas, estamos perante situações em que é necessário fazer inferência estatística, isto é, pretendemos tirar conclusões para um grande conjunto de indivíduos (População), a partir do estudo de um número restrito desses indivíduos (Amostra). Assim, quando a partir do estudo de uma amostra pretendemos inferir para a população de onde a amostra foi recolhida, existe sempre um grau de incerteza, associado à aleatoriedade da escolha da amostra, que é medido em termos de Probabilidade. Alguns exemplos ajudar-nos-ão a desenvolver esta ideia.

**Exemplo 2** – Admita que tem uma moeda equilibrada. Mas o que é uma moeda equilibrada? É aquela em que estamos a admitir, à partida, que existe igual possibilidade de sair cara ou coroa no próximo lançamento que façamos com ela – estamos a admitir o *princípio da simetria*, de que falámos anteriormente. Estamos, assim, a admitir, na nossa cabeça, um *modelo matemático* em que assumimos que em qualquer lançamento da moeda, a probabilidade de sair cara é igual à de sair coroa e igual a 1/2:

Modelo para o resultado do lançamento da moeda equilibrada

| Resultado | Cara (F) | Coroa (C) |
|---|---|---|
| Probabilidade | 1/2 | 1/2 |

Não nos estamos a preocupar, por exemplo, com a força ou direcção com que atiramos a moeda, nem tão pouco com o desgaste acusado pela moeda após sucessivos lançamentos! Também não estamos a encarar a hipótese da moeda cair de pé! Se nos estivéssemos a preocupar em arranjar um modelo que traduzisse mais fielmente a realidade, estaríamos a arranjar um modelo matemático tão complicado que seria impossível de tratar e não nos serviria para nada. O estatístico George Box dizia:

*Todos os modelos são maus, alguns modelos são úteis.*

Assumindo então o modelo anterior, um pouco simplista, para o lançamento da moeda, se lançarmos a moeda repetidas vezes, esperamos que o número de caras seja aproximadamente metade do número de lançamentos. Se, por outro lado, recolhermos uma amostra de dimensão 1, isto é, fizermos um único lançamento, não sabemos qual o resultado que se vai verificar, se será cara ou coroa, mas dizemos que a probabilidade de sair cara é 1/2.

Suponha agora que não podíamos invocar o princípio da simetria, isto é, não sabíamos se a moeda era equilibrada. Neste caso a População que estamos a estudar não é completamente conhecida, pois conhecemos os resultados possíveis em cada lançamento, mas não conhecemos as suas probabilidades - o modelo não está completamente especificado. Como obter alguma informação, para especificar um modelo para o lançamento da moeda? Um modo possível de obter mais alguma informação sobre o modelo probabilístico é proceder a um certo número de lançamentos e calcular a frequência relativa da saída de cara, nos lançamentos efectuados. Este valor vai-nos servir para *estimar* a probabilidade da saída de cara. Por exemplo, se em 1000 lançamentos se obtiveram 324 caras, dizemos que um valor aproximado para a probabilidade de se verificar cara é 0.324 (ao fim de 1000 lançamentos verificou-se uma certa estabilidade à volta deste valor) e o valor aproximado para a probabilidade de sair coroa será 0.676.

O comportamento de grandes grupos de indivíduos, pode ser também considerado aleatório e o processo utilizado para definir um modelo, é o de verificar o que é que se passa com um grande conjunto de indivíduos.

**Exemplo 3** (Moore, 1997) – Se nos perguntassem qual a probabilidade de uma determinada pessoa morrer no próximo ano, obviamente que não saberíamos dizer. No entanto, se observarmos milhões de pessoas, poderemos obter um padrão para o comportamento das mortes. É assim que poderemos dizer que a proporção de homens, com idades compreeendidas entre 25 e 34 anos, que morrerão no próximo ano, anda à volta de 0.0021. Esta proporção, verificada para um conjunto grande de indivíduos, será entendida como a *probabilidade* de que um homem jovem morra no próximo ano. Para as mulheres com aquela idade, a probabilidade de morrer será cerca de 0.0007. Estamos, a partir da observação de resultados verificados numa amostra, a inferir para toda a população constituída pelos indivíduos da classe etária considerada. Estes modelos têm muito interesse para as companhias de seguros, quando se trata nomeadamente de seguros de vida, já que lhes vai permitir definir uma política de preços para as apólices, sendo até natural que cobrem mais por um seguro de vida a um homem, do que a uma mulher.

Com os exemplos anteriores tentámos exprimir o papel relativo da Probabilidade e da Estatística, que resumimos a seguir:

Enquanto que ao assumirmos um determinado modelo de probabilidade – População conhecida, o que foi feito ao admitir que a moeda era *equilibrada*, estamos aptos a raciocinar do geral para o particular, isto é, da População para a Amostra, quando a População não é conhecida utilizamos a Estatística para fazer raciocínios no sentido inverso, isto é, inferir para a População resultados observados na Amostra.

Para formalizarmos um pouco o conceito de Probabilidade, vamos introduzir alguma terminologia própria.

## 5.2 - Experiência aleatória. Espaço de resultados. Acontecimentos.

Dissemos anteriormente que o objectivo da Teoria da Probabilidade é o de estudar fenómenos aleatórios, construindo modelos matemáticos, a que chamamos modelos de probabilidade, que os possam descrever convenientemente. A noção mais básica a de experiência aleatória.

**Experiência aleatória** – é o processo de observar um resultado de um fenómeno aleatório. Numa experiência aleatória obtém-se um **resultado**, de entre um conjunto de resultados conhecidos de antemão, mas **não se tem conhecimento suficiente** de qual o resultado que sai em cada realização da experiência. Admite-se ainda que a experiência se pode **repetir** e que as repetições são realizadas nas mesmas circunstâncias e são independentes.

**Observação**: Esta definição de experiência aleatória, segundo a qual a experiência se pode repetir o número de vezes que se quiser, independentemente umas das outras e sempre nas mesmas circunstâncias, apresentando uma *regularidade estatística*, prepara-nos para a definição de *probabilidade*, segundo *a teoria frequencista*, como veremos mais à frente.

São exemplos de experiências aleatórias:

- contar o número de carros estacionados, na rua, ao sairmos de manhã de casa;
- perguntar a uma pessoa ao acaso, quantas são as pessoas do seu agregado familiar;
- lançar uma moeda ao ar e ver o resultado que sai;
- lançar uma moeda ao ar 20 vezes e ver quantas caras saem;
- medir o tempo que de manhã levamos a chegar ao emprego;
- contar o número de desastres que encontramos, em cada dia, na ida para o emprego.

As situações anteriores são exemplos de experiências aleatórias, pois além de envolverem aleatoriedade, o resultado da experiência está bem especificado. O mesmo não se passa com a seguinte situação: *ao acordar, de manhã, ir à janela*. Efectivamente, na situação anterior não se especificou qual o resultado possível, de modo a termos uma experiência aleatória. No entanto, associado à situação anterior são experiências aleatórias:

- *ao acordar, de manhã, ir à janela* e ver se chove;

- *ao acordar, de manhã, ir à janela* e contar o número de carros encarnados, que passam num período de 5 minutos.

**Espaço de resultados S** - é o conjunto de todos os resultados possíveis, associados à realização de uma experiência aleatória.

Relativamente à experiência aleatória que consiste em observar o resultado do lançamento de uma moeda ao ar, temos:

$$S = \{ \text{cara, coroa} \}$$

Relativamente à experiência aleatória que consiste em observar o número de caras saídas em 20 lançamentos de uma moeda, temos:

$$S = \{0,1,2,...,19,20\}$$

Relativamente à experiência aleatória que consiste em observar de manhã o tempo que se leva a chegar ao emprego, temos

$$S = [0, +\infty [$$

Relativamente à experiência aleatória que consiste em observar o resultado do lançamento de dois dados, temos:

$$S = \{(i,j): i=1,2,...,6; j=1,2,...6\}$$

**Acontecimento** - Define-se acontecimento, como sendo um subconjunto do espaço de resultados S.

Considerando a experiência aleatória que consiste em perguntar a uma pessoa, escolhida ao acaso, quantas pessoas constituem o seu agregado familiar, o espaço de resultados é constituído por todos os números inteiros não negativos (excluindo o zero). Alguns acontecimentos são:

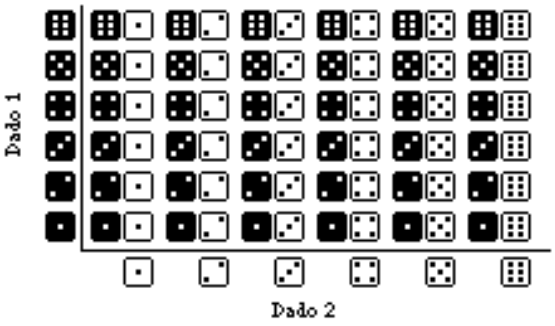
- 3 pessoas, que podemos representar por {3}
- entre 2 e 4 pessoas (inclusive)" " {2, 3, 4)
- mais de 3 pessoas " " {4, 5, 6, …}
- menos de 10 pessoas " " {1, 2, …, 9}

De um modo geral os acontecimentos identificam-se com letras maiúsculas A, B, etc. Diz-se que se realizou o acontecimento A, quando o resultado da experiência pertence a A.

Alguns dos acontecimentos são constituídos por um único resultado: chamam-se **acontecimentos elementares**.

**Exemplo 4** - Considere a experiência aleatória que consiste no lançamento de dois dados. Identifique o espaço de resultados e os acontecimentos “o número de pintas é igual nos dois dados” e “a soma das pintas é 7”.

Para descrever o espaço de resultados vamos considerar dois dados, um preto e um branco, para os distinguir. O espaço de resultados é constituído por todos os pares de dados considerados na figura a seguir. O número de elementos do espaço de resultados é 36 = 6X6.

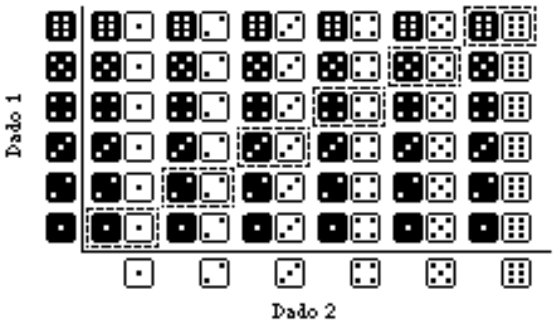



O espaço anterior pode ser descrito de forma mais sintética considerando os pares ordenados (i,j), onde representamos por i o número de pintas do dado 1, ou seja do dado preto, e por j o número de pintas do dado 2, ou seja do dado branco:

$$S = \{(i,j): i=1,2,...,6;\ j=1,2,...6\}$$

Chamamos a atenção que, por exemplo, o par (1,3) não é o mesmo que o par (3,1). No par ordenado, o primeiro elemento refere-se a um dos dados (neste caso o dado preto) e o segundo elemento refere-se ao outro dado (o dado branco).

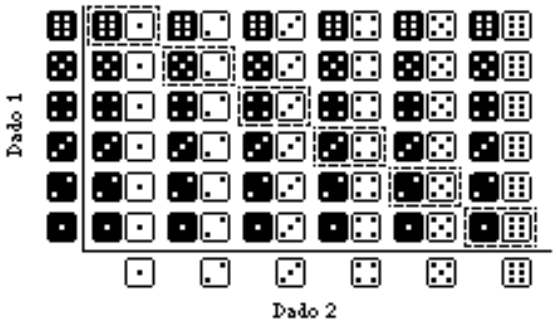
O acontecimento "o número de pintas é igual nos dois dados" é constituído pelos pares assinalados na figura seguinte




ou em notação em termos dos pares ordenados

$$A = \{ (1,1), (2,2), (3,3), (4,4), (5,5), (6,6)\}$$

Finalmente o acontecimento "a soma das pintas é 7" é constituído pelos pares assinalados na figura seguinte




ou em notação em termos dos pares ordenados

B = {(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)}

Observação: Qual a diferença entre o espaço de resultados associado à experiência aleatória do lançamento de dois dados e a experiência que consiste no lançamento do mesmo dado duas vezes? O espaço de resultados é idêntico nas duas experiências. Considerámos dados de cores distintas para justificar a nossa opção para descrever S como um conjunto de pares ordenados, mas é óbvio que este mesmo espaço serve para modelar o lançamento de dois dados idênticosou dois lançamentos de um mesmo dado.

Nota – Associado à experiência que acabámos de descrever no exemplo anterior, poderíamos ter considerado o seguinte espaço de resultados:

S = { saírem dois 1's, sair um 1 e um 2, sair um 1 e um 3, sair um 1 e um 4, sair um 1 e um 5, sair um 1 e um 6, saírem dois 2's, sair um 2 e um 3, sair um 2 e um 4, sair um 2 e um 5, sair um 2 e um 6, saírem dois 3's, sair um 3 e um 4, sair um 3 e um 5, sair um 3 e um 6, saírem dois 4's, sair um 4 e um 5, sair um 4 e um 6, saírem dois 5's, sair um 5 e um 6, saírem dois 6's}

Qual a desvantagem em considerar este espaço de resultados? Como veremos mais à frente, se o espaço de resultados for constituído por resultados igualmente possíveis, o que não acontece nesta situação, podemos utilizar a regra de Laplace, para atribuir probabilidades a acontecimentos associados ao fenómeno em estudo.

**Exemplo 5** - Se lançar 3 dados como é constituído o espaço de resultados? Utilizando uma generalização da notação do exemplo anterior, o espaço de resultados será constituído por todos os triplos (i, j, k), em que o i, j e k, podem assumir os valores de 1 a 6. O i refere-se a um dos dados, por exemplo o 1º a ser lançado, ou se os quisermos distinguir a um dado preto, o j refere-se ao 2º dado a ser lançado, ou a um dado branco e finalmente o k refere-se ao 3º dado a ser lançado, ou a um dado vermelho. O número de elementos do espaço de resultados, ou seja, o número de resultados possíveis é 216 = 6 x 6 x 6.

**Nota histórica** (Statistics, 1991) - No século XVII, os jogadores italianos costumavam fazer apostas sobre o número total de pintas obtidas no lançamento de 3 dados. Acreditavam que a possibilidade de obter um total de 9 era igual à possibilidade de obter um total de 10. Por exemplo, diziam que uma combinação possível para dar um total de 9 seria

1 pinta num dos dados, 2 pintas num outro dado, 6 pintas no terceiro dado

Abreviando o resultado anterior para “1 2 6”, todas as combinações para dar o 9 são:

1 2 6  1 3 5  1 4 4  2 3 4  2 2 5  3 3 3

Analogamente, obtinham 6 combinações para o 10:

1 4 5  1 3 6  2 2 6  2 3 5  2 4 4  3 3 4

Assim, os jogadores argumentavam que o 9 e o 10 deveriam ter a mesma possibilidade de se verificarem. Contudo, a experiência mostrava que o 10 aparecia com uma frequência um pouco superior ao 9. Pediram a Galileu que os ajudasse nesta contradição, tendo este realizado o seguinte raciocínio: Pinte-se um dos dados de branco, o outro de cinzento e o outro de preto. De quantas maneiras se podem apresentar os três dados depois de lançados? O dado branco pode apresentar 6 possibilidades diferentes. Para cada uma destas possibilidades o dado cinzento pode apresentar 6 possibilidades, obtendo-se 6 X 6 possibilidades para os dois dados. Correspondendo a cada uma destas possibilidades, o dado preto pode apresentar 6 possibilidades obtendo-se no total 6 X 6 X 6 = 216 possibilidades. Galileu listou todas as 216 maneiras de 3 dados se apresentarem depois de lançados. Depois percorreu a lista e verificou que havia 25 maneiras de obter um total de 9 e 27 maneiras de obter um total de 10.

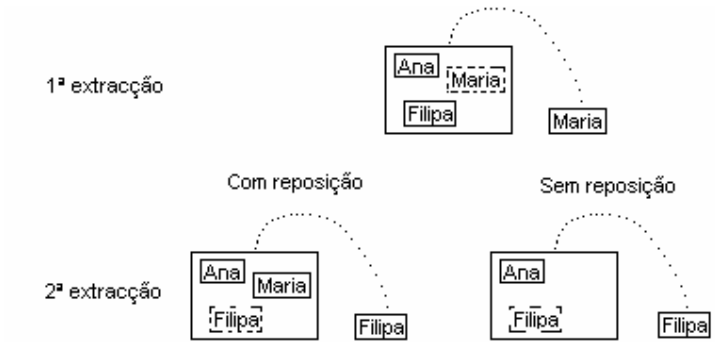
O raciocínio dos jogadores não entrava em linha de conta com as diferentes maneiras como os dados se podiam apresentar. Por exemplo o triplo 3 3 3, que dá o 9, corresponde unicamente a uma forma de os dados se apresentarem, mas o triplo 3 3 4 que dá o 10, corresponde a 3 maneiras diferentes:

pelo que o raciocínio dos jogadores deve ser corrigido de acordo com a tabela seguinte:

| Triplos para o 9 | | | Nº de maneiras de obter o triplo | Triplos para o 10 | | | Nº de maneiras de obter o triplo |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 6 | 6 | 1 | 4 | 5 | 6 |
| 1 | 3 | 5 | 6 | 1 | 3 | 6 | 6 |
| 1 | 4 | 4 | 3 | 2 | 2 | 6 | 3 |
| 2 | 3 | 4 | 6 | 2 | 3 | 5 | 6 |
| 2 | 2 | 5 | 3 | 2 | 4 | 4 | 3 |
| 3 | 3 | 3 | 1 | 3 | 3 | 4 | 3 |
| Total | | | 25 | Total | | | 27 |

## Extracções com reposição e sem reposição

Colocaram-se (Graça Martins. M.E. et al, 1999) numa caixa 3 papéis com o nome de 3 meninas: Ana, Maria e Filipa. Considere a experiência aleatória que consiste em retirar da caixa 2 papéis e verificar os nomes que saíram. Qual o espaço de resultados? Para responder a esta questão é necessário saber se a extracção se faz *com reposição*, isto é, se uma vez retirado um papel e verificado o nome se volta a colocar o papel na caixa, antes de proceder à extracção seguinte, ou se a extracção é feita *sem reposição*, isto é, uma vez retirado um papel, ele não é reposto antes de se proceder à próxima extracção. No esquema seguinte procuramos representar as duas situações.




Admitimos que na 1ª extracção saiu o papel com o nome da Maria. Na 2ª extracção, saiu o nome da Filipa nos dois casos, mas *na extracção com reposição* havia uma possibilidade em três de ele sair, tal como na 1ª extracção, enquanto que na *extracção sem reposição* havia uma possibilidade em duas de ele sair. Quer dizer que neste caso havia uma maior probabilidade de sair o nome da Filipa. Os espaços de resultados $S_c$ e $S_s$ correspondentes às duas situações com reposição e sem reposição, são respectivamente:

$S_c$ = {(Ana, Ana), (Ana, Maria), (Ana, Filipa), (Maria, Ana), (Maria, Maria), (Maria, Filipa), (Filipa, Ana); (Filipa, Maria), (Filipa, Filipa)}

$S_s$ = {(Ana, Maria), (Ana, Filipa), (Maria, Ana), (Maria, Filipa), (Filipa, Ana), (Filipa, Maria)}.

O acontecimento "saiu o nome da Maria" é constituído pelos seguintes resultados, considerando a extracção com reposição e sem reposição, respectivamente:

$A_c$= {(Ana, Maria), (Maria, Ana), (Maria, Maria), (Maria, Filipa), (Filipa, Maria)}

e $A_s$ = {(Ana, Maria), (Maria, Ana), (Maria, Filipa), (Filipa, Maria)}.

**Exemplo 6** - Considere a experiência aleatória que consiste em extrair 2 berlindes, de um saco com 3 berlindes vermelhos e 2 azuis. Qual é o espaço de resultados?

Para já é necessário saber se a extracção se faz com reposição ou sem reposição. Vamos considerar as duas situações. Para identificar o espaço de resultados será mais fácil numerar os berlindes, pelo que vamos numerar os berlindes vermelhos com 1, 2 e 3 e os azuis com 4 e 5.

**Com reposição** - Quando se retira um berlinde verifica-se a cor e torna-se a repor o berlinde no saco antes de extrair o próximo. O espaço de resultados é constituído por todos os resultados, em número de 25, do esquema seguinte:

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 11 | 12 | 13 | 14 | 15 |
| 2 | 21 | 22 | 23 | 24 | 25 |
| 3 | 31 | 32 | 33 | 34 | 35 |
| 4 | 41 | 42 | 43 | 44 | 45 |
| 5 | 51 | 52 | 53 | 54 | 55 |

**Sem reposição** - Neste caso o espaço de resultados é constituído por todos os resultados do espaço do esquema anterior, exceptuando os pares constituídos pelo mesmo berlinde:

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | | 12 | 13 | 14 | 15 |
| 2 | 21 | | 23 | 24 | 25 |
| 3 | 31 | 32 | | 34 | 35 |
| 4 | 41 | 42 | 43 | | 45 |
| 5 | 51 | 52 | 53 | 54 | |

O acontecimento “tirar 2 berlindes de cor diferente” é constituído pelos resultados {(1,4), (1,5), (2,4), (2,5), (3,4), (3,5), (4,1), (4,2), (4,3), (5,1), (5,2), (5,3)} tanto no esquema com reposição, como sem reposição.

### 5.2.1 – Operações com acontecimentos

Uma técnica utilizada para visualizar acontecimentos consiste em utilizar um rectângulo para representar o espaço de resultados e círculos para representar os acontecimentos. A essas representações chamamos *diagramas de Venn*. Vamos utilizar esses diagramas para apresentar a terminologia utilizada quando falamos de acontecimentos. Assim, representando os acontecimentos por A, B, C, ..., temos:

- Acontecimento **Complementar** ou **contrário** do acontecimento **A**:

O acontecimento **complementar** ou **contrário** do acontecimento A, representa-se por $\overline{A}$ ou $\mathbf{A}^{\mathbf{c}}$ e é o acontecimento constituído por todos os resultados de S, que não estão em A.


$\overline{A}$

- Acontecimento **A implica B**

O acontecimento A **implica** a realização do acontecimento B, quando todo o resultado de A é um resultado de B; indica-se este facto escrevendo $\mathbf{A} \subset \mathbf{B}$.


$A \subset B$

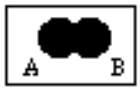
- Acontecimento **Intersecção**

**Intersecção** dos acontecimentos A e B, **A∩B** , ou (A **e** B) é o acontecimento que se realiza sse A e B se realizam simultaneamente.


$A \cap B$

- Acontecimento **União**

**União** dos acontecimentos A e B, **A∪B** , ou (A **ou** B) é o acontecimento que se realiza sse A ou B se realizam.




$A \cup B$

- Acontecimentos **Disjuntos**

Acontecimentos **disjuntos** ou acontecimentos **mutuamente exclusivos**, são acontecimentos em que a realização de um deles implica a não realização do outro.


- Acontecimento **Diferença**

Acontecimento diferença entre A e B, **A-B**, é o acontecimento que se realiza sse A se realiza, sem que B se realize.


- Acontecimento **Impossível**

Acontecimento **impossível** é o acontecimento que resulta da intersecção de acontecimentos mutuamente exclusivos. Analogamente ao que se passa na teoria dos conjuntos, representa-se por $\phi$ ( símbolo do conjunto vazio, mas que aqui se lê acontecimento impossível e não acontecimento vazio). Então, com esta notação introduzida para o acontecimento impossível, temos:

**Se dois acontecimentos são disjuntos, então A∩B =** $\phi$.

**Exemplo 7** - Relativamente à experiência aleatória que consiste no lançamento de um dado, represente com a notação que achar conveniente:

a) O espaço de resultados

b) O acontecimento A, que consiste em sair uma face par (número de pintas par)

c) O acontecimento B que consiste em sair face ímpar

d) O acontecimento C que consiste em sair uma face menor que 3

e) O acontecimento intersecção de A com B. O que conclui acerca dos acontecimentos A e B?

f) O acontecimento união de A com B. O que conclui acerca dessa união?

g) O acontecimento intersecção de B e C.

Resolução:

a) S = { 1, 2, 3, 4, 5, 6}; b) A = { 2, 4, 6}; c) B = { 1, 3, 5};

d) C = { 1, 2}; e) $A \cap B = \phi$, pelo que os acontecimentos A e B são disjuntos

f) $A \cup B = S$, pelo que a união de A e B é o espaço de resultados. Das alíneas e) e f) concluímos que os acontecimentos A e B são complementares.

g) $B \cap C = \{ 1\}$

**Exemplo 8** - Considere a experiência aleatória que consiste em verificar os resultados de um desafio de futebol Benfica-Sporting.

a) Descreva o espaço dos resultados.

b) Represente os acontecimentos : A - empate; B - Benfica ganhou; C - Sporting ganhou.

Resolução:

a) S = { (i,j): i = 0, 1, 2, 3, …; j= 0, 1, 2, 3, …}, isto é, o espaço de resultados é constituído por todos os pares possíveis de números naturais, incluindo o zero

b) A = { (i,i): i = 0, 1, 2, 3, …}; B = { (i,j): i = 1, 2, 3, …; j= 0, 1, 2, 3, … e **i>j**}

C = { (i,j): i = 0, 1, 2, 3, …; j= 1, 2, 3, … e **i<j**}

**Exemplo 9** - Uma empresa que faz a prospecção de petróleo, quando faz um furo pode encontrar petróleo ou gás, ou não encontrar nada. A empresa fez dois furos.

a) Descreva o espaço de resultados associado à experiência aleatória anterior.

b) Represente o acontecimento : a empresa obteve petróleo ou gás.

Resolução:

a) S = {(petróleo ou gás, nada), (petróleo ou gás, petróleo ou gás), (nada, petróleo ou gás), (nada, nada)}

b) A = {(petróleo ou gás, nada), (petróleo ou gás, petróleo ou gás), (nada, petróleo ou gás)}

## 5.3 - Probabilidade de um acontecimento

Dissemos anteriormente que o nosso objectivo é definir modelos de probabilidade para fenómenos aleatórios, que nos interessem estudar. Em espaços finitos, esta definição implica:

- A identificação de um espaço de resultados;
- Uma forma de atribuir probabilidades a cada um dos resultados, isto é, aos acontecimentos elementares.

O processo de atribuir probabilidades deve ser tal, que algumas regras básicas devam ser satisfeitas para todos os modelos. Vamos então considerar as seguintes regras, que são intuitivas:

Regra 1 – Uma probabilidade deve ser um número entre 0 e 1;

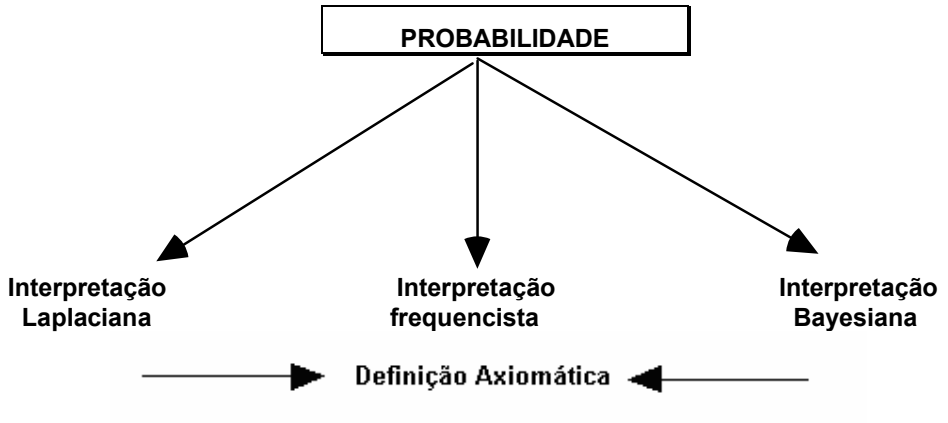
Regra 2 – O conjunto de todos os resultados possíveis tem probabilidade igual a 1;

Admitamos, para já, que tínhamos um processo de definir um modelo de probabilidade. Uma vez definido esse modelo de probabilidade, como obter a probabilidade de acontecimentos?

Uma vez que um acontecimento é um conjunto de resultados, vamos definir a probabilidade do acontecimento A, que representamos por P(A), à custa das probabilidades dos resultados qur compõem A:

Em espaços finitos, a probabilidade de um acontecimento A é a soma das probabilidades dos acontecimentos elementares que compõem A.

A probabilidade é uma medida do grau de incerteza atribuído à realização de um acontecimento. A sua quantificação é susceptível de várias interpretações, que apresentamos a seguir. Assim vamos abordar o conceito de Probabilidade de um acontecimento, fazendo referência à interpretação frequencista, clássica ou Laplaciana, subjectivista ou Bayesiana e finalmente introduzimos a definição axiomática de Probabilidade




### 5.3.1 – Probabilidade frequencista

Retomemos a definição de experiência aleatória. Desta definição, vimos que uma das suas características consistia no facto de se poder *repetir*, nas mesmas circunstâncias, apresentando uma *regularidade estatística*. Vamos então repetir a experiência um grande número de vezes e registar a *frequência relativa* - proporção de vezes, com que um determinado resultado (acontecimento elementar) ocorreu.

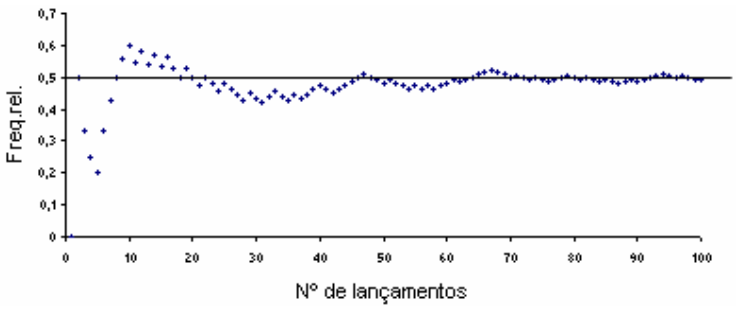
À medida que o número de repetições da experiência aleatória aumenta, a frequência relativa do acontecimento elementar tende para um valor entre 0 e 1. Este limite, é interpretado como sendo a **Probabilidade** desse acontecimento elementar.

Suponhamos, por exemplo, a experiência aleatória que consiste no lançamento de uma moeda ao ar e observar a face que fica virada para cima. Realizaram-se 100 lançamentos, tendo-se obtido os seguintes resultados:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | cara | 21 | cara | 41 | cara | 61 | coroa | 81 | cara |
| 2 | coroa | 22 | coroa | 42 | cara | 62 | cara | 82 | coroa |
| 3 | cara | 23 | cara | 43 | coroa | 63 | coroa | 83 | cara |
| 4 | cara | 24 | cara | 44 | coroa | 64 | coroa | 84 | cara |
| 5 | cara | 25 | coroa | 45 | coroa | 65 | coroa | 85 | coroa |
| 6 | coroa | 26 | cara | 46 | coroa | 66 | coroa | 86 | cara |
| 7 | coroa | 27 | cara | 47 | coroa | 67 | coroa | 87 | cara |
| 8 | coroa | 28 | cara | 48 | cara | 68 | cara | 88 | coroa |
| 9 | coroa | 29 | coroa | 49 | cara | 69 | cara | 89 | coroa |
| 10 | coroa | 30 | cara | 50 | cara | 70 | cara | 90 | cara |
| 11 | cara | 31 | cara | 51 | coroa | 71 | coroa | 91 | coroa |
| 12 | coroa | 32 | coroa | 52 | cara | 72 | cara | 92 | coroa |
| 13 | cara | 33 | coroa | 53 | cara | 73 | cara | 93 | coroa |
| 14 | coroa | 34 | cara | 54 | cara | 74 | coroa | 94 | coroa |
| 15 | cara | 35 | cara | 55 | coroa | 75 | cara | 95 | cara |
| 16 | coroa | 36 | coroa | 56 | cara | 76 | cara | 96 | cara |
| 17 | cara | 37 | cara | 57 | coroa | 77 | coroa | 97 | coroa |
| 18 | cara | 38 | coroa | 58 | cara | 78 | coroa | 98 | cara |
| 19 | coroa | 39 | coroa | 59 | coroa | 79 | coroa | 99 | cara |
| 20 | cara | 40 | coroa | 60 | coroa | 80 | cara | 100 | cara |

Se ao fim dos 100 lançamentos se verificaram 49 coroas, então a frequência relativa com que se verificou coroa foi de 0.49. <u>O limite para que tende a frequência relativa da saída de coroa, ao fim de um grande número de lançamentos, é interpretado como a probabilidade de saída de coroa</u> .

O gráfico obtido para a frequência relativa após cada lançamento, tem o seguinte aspecto:




A frequência relativa, à medida que o número de provas aumenta, tem tendência a estabilizar à volta do valor 0.5. Assim, dizemos que a probabilidade de sair coroa é 0.5.

**Observação**: Chamamos a atenção, ainda relativamente a este exemplo, para o seguinte: não é correcto dizer que à medida que o número de lançamentos aumenta, o número de coroas se aproxima de metade do número de lançamentos. A *regularidade a longo termo* significa que a *proporção* de vezes que saiu coroa tende a estabilizar. Neste caso, ao fim de 100 lançamentos o número de coroas foi de 49; se continuássemos a fazer lançamentos poderia acontecer que ao fim de 500, 1000, 2000 e 3000 lançamentos, o número de coroas obtidas fosse respectivamente de 253, 495, 993 e 1510 como se apresenta na seguinte tabela:

| Nº lançamentos | Nº coroas obtidas x | Metade dos lanç. y | \|y - x\| | Freq. relativa |
|---|---|---|---|---|
| 100 | 49 | 50 | 1 | 0.49 |
| 500 | 253 | 250 | 3 | 0.51 |
| 1000 | 495 | 500 | 5 | 0.50 |
| 2000 | 993 | 1000 | 7 | 0.50 |
| 3000 | 1510 | 1500 | 10 | 0.50 |

Como se verifica, pode acontecer que o número de coroas obtidas se afaste de metade do número de lançamentos, não impedindo que a frequência relativa tenha tendência a estabilizar à volta do valor 0.50.

**Definição frequencista de probabilidade** - Define-se *probabilidade* de um acontecimento A e representa-se por P(A) como sendo o valor para que tende a frequência relativa da realização de A, num grande número de repetições da experiência aleatória

$$P(A) = \text{limite da frequência relativa } \frac{n_A}{n} \text{ com que se realiza o acontecimento A}$$

($n_A$ representa o nº de realizações de A em n repetições da experiência)

**Exemplo 10** - Suponha que lança um dado 1000 vezes e verifica a face que ficou voltada para cima, tendo obtido os seguintes resultados:

| Face | Freq. abs. | Freq. rel.(%) |
|---|---|---|
| 1 | 159 | 15.9% |
| 2 | 163 | 16.3% |
| 3 | 160 | 16.0% |
| 4 | 161 | 16.1% |
| 5 | 86 | 8.6% |
| 6 | 271 | 27.1% |

Perante os resultados anteriores somos levados a sugerir o seguinte modelo de probabilidade para o fenómeno aleatório que consiste em verificar qual a face que sai no lançamento de um dado:

| Face | Probabilidade |
|---|---|
| 1 | 16% |
| 2 | 16% |
| 3 | 16% |
| 4 | 16% |
| 5 | 9% |
| 6 | 27% |

Os resultados anteriores levam-nos ainda a concluir que estamos perante um dado “viciado”, pois as faces não têm todas a mesma probabilidade de saírem, como seria de esperar se o dado fosse “equilibrado”.

**Exemplo 11 -** Qual a probabilidade de ao retirar uma carta ao acaso de um baralho de 52 cartas, ela ser um Ás? Suponha que tem um baralho de cartas e pede a alguém para retirar uma carta; verifica se é Ás e repõe a carta novamente no baralho. Repete esta experiência 1000 vezes, tendo o cuidado de entre duas extracções sucessivas, embaralhar as cartas. Os resultados obtidos foram os seguintes:

| Nº repetições | Freq. abs. Ás | Freq. rel. Ás |
|---|---|---|
| 1000 | 78 | 0.078 |

Perante os resultados anteriores sugere-se a probabilidade de 8% para a saída de Ás.

*Será sempre possível definir a probabilidade de um acontecimento, utilizando a definição anterior?*

Este processo de submeter a atribuição da probabilidade de um acontecimento, à realização do acontecimento um grande número de vezes, é susceptível de crítica, na medida em que nem sempre se pode repetir a experiência as vezes necessárias, de modo a obter a convergência pretendida.

## Utilização do Excel na simulação de experiências aleatórias

Os algoritmos de geração de números pseudo-aleatórios no intervalo [0,1], estão concebidos de tal modo, que ao considerar uma qualquer sequência de números gerados se obtenha aproximadamente a mesma proporção de observações em subintervalos de igual amplitude do intervalo [0,1]. Assim, por exemplo, se se fizer correr um desses algoritmos 100 vezes, é de esperar que caiam 25 dos números gerados em cada quarto do intervalo [0,1] (Loura, L. e Graça Martins, M. E., 2001).

De um modo geral quando falamos em números aleatórios, estamos a referir-nos à obtenção de qualquer real do intervalo [0, 1], de tal forma que a probabilidade de obter um valor de um subintervalo [a, b] de [0, 1], é igual à amplitude desse subintervalo, ou seja (b-a). No Excel, obtemos estes números com a função *RAND*. A função *RANDBETWEEN(m;n)*, já utilizada em capítulos anteriores, gera números pseudo-aleatórios inteiros, no intervalo [m,n].

**Exemplo –** Simule 10 lançamentos de uma moeda equilibrada, utilizando a função RAND() do Excel.

Como admitimos que a moeda é equilibrada, vamos utilizar a função RAND() da seguinte forma: se o resultado for menor que 0,5, simulamos a saída de coroa; caso contrário simulamos a saída de cara. Um procedimento possível para a simulação em causa é o seguinte:

NúmerosAleatórios

| | A | B | C |
|---|---|---|---|
| 1 | =RAND() | 0,947664443788201 | =IF(B1<0,5;"coroa";"cara") |
| 2 | =RAND() | 0,291064707013349 | =IF(B2<0,5;"coroa";"cara") |
| 3 | =RAND() | 0,327702896521203 | =IF(B3<0,5;"coroa";"cara") |
| 4 | =RAND() | 0,857492455318246 | =IF(B4<0,5;"coroa";"cara") |
| 5 | =RAND() | 0,871382593518266 | =IF(B5<0,5;"coroa";"cara") |
| 6 | =RAND() | 0,822363769573869 | =IF(B6<0,5;"coroa";"cara") |
| 7 | =RAND() | 0,102469601168385 | =IF(B7<0,5;"coroa";"cara") |
| 8 | =RAND() | 0,61056813317687 | =IF(B8<0,5;"coroa";"cara") |
| 9 | =RAND() | 0,221923780776478 | =IF(B9<0,5;"coroa";"cara") |
| 10 | =RAND() | 0,590014480461622 | =IF(B10<0,5;"coroa";"cara") |

NUM

Número...

| | B | |
|---|---|---|
| 1 | 0,947664 | cara |
| 2 | 0,291065 | coroa |
| 3 | 0,327703 | coroa |
| 4 | 0,857492 | cara |
| 5 | 0,871383 | cara |
| 6 | 0,822364 | cara |
| 7 | 0,102470 | coroa |
| 8 | 0,610568 | cara |
| 9 | 0,221924 | coroa |
| 10 | 0,590014 | cara |

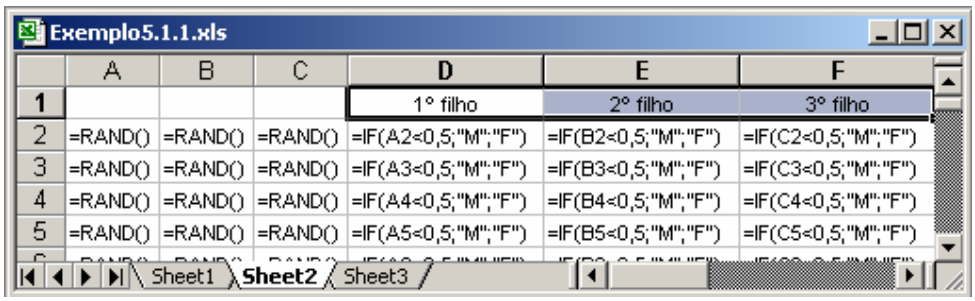
Para obter a tabela do lado esquerdo, inserimos a função *RAND()* na célula A1 e replicámos *(Fill Down*) até à célula A10. Seguidamente copiámos os valores obtidos – utilizando o *Paste Special*, para as células B1:B10. Este procedimento é necessário (se nos quisermos fixar numa determinada amostra), pois a função *RAND* é volátil, pelo que em cada cálculo da folha, gera um novo número. Com a função IF, simulámos o resultado cara ou coroa, que se apresenta na tabela do lado direito, da figura anterior.

**Exemplo (**Exemplo 5.1.1 de Loura, L. e Graça Martins, M. E., 2005) – Suponha um casal que pretende ter um "casal" de filhos, não desejando mais do que 3 filhos e só tentando o 3º filho se anteriormente tiver tido ou dois rapazes ou duas raparigas. Qual a probabilidade de ter efectivamente o casalinho?

Admitindo que a probabilidade de nascer rapaz é igual à de nascer rapariga, vamos utilizar a função RAND, para simular um qualquer destes nascimentos, da seguinte forma: Se o resultado da função RAND for inferior a 0,5, simulamos o nascimento de um rapaz – M. Caso contrário simulamos o nascimento de uma rapariga. Numa folha de Excel vamos simular várias repetições da experiência "nascimento de 3 filhos". Poderíamos ter optádo por começar por simular o nascimento de dois filhos e só simular o 3º filho se não houvesse os dois sexos nos dois primeiros filhos. No entanto, este condicionamento da simulação do 3º filho faz com que cada repetição da experiência dependa do que se obtém anteriormente, o que torna mais demorado o processo da simulação. Assim, simulámos sempre 3 filhos e basta nos dois primeiros haver os dois sexos, para termos como resultado da experiência um sucesso. Assinalamos o sucesso (dois sexos diferentes logo nos dois primeiros filhos ou sexos diferentes nos três filhos) com um 1 – esta notação facilita-nos o cálculo da frequência relativa do nº de sucessos, à medida que repetimos a experiência.

Um procedimemto possível para a simulação em causa, pode ser o seguinte:

- Inserir a função RAND() nas células A2, B2 e C2 e nas células D2, E2 e F2 a função IF(), como se exemplifica na figura seguinte:



Exemplo5.1.1.xls

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | | | | 1º filho | 2º filho | 3º filho |
| 2 | =RAND() | =RAND() | =RAND() | =IF(A2<0,5;"M";"F") | =IF(B2<0,5;"M";"F") | =IF(C2<0,5;"M";"F") |
| 3 | =RAND() | =RAND() | =RAND() | =IF(A3<0,5;"M";"F") | =IF(B3<0,5;"M";"F") | =IF(C3<0,5;"M";"F") |
| 4 | =RAND() | =RAND() | =RAND() | =IF(A4<0,5;"M";"F") | =IF(B4<0,5;"M";"F") | =IF(C4<0,5;"M";"F") |
| 5 | =RAND() | =RAND() | =RAND() | =IF(A5<0,5;"M";"F") | =IF(B5<0,5;"M";"F") | =IF(C5<0,5;"M";"F") |

Sheet1 Sheet2 Sheet3

- Replicar (*Fill down*) as células A2:F2, tantas vezes quantas as vezes que se pretende simular a realização da experiência. Nós replicámos 400 vezes, colocando os resultados nas células A2:F401;
- Copiar (*Paste special*) os valores das células D2:F401, para as células H2:J401 (Este passo tem como objectivo guardar os valores gerados anteriormente, pois a função RAND() é volátil, como já referimos anteriormente;
- Em cada uma das células da coluna K inserir 1 se o resultado da experiência tiver sido sucesso;
- Na coluna L contabilizar o nº de sucessos acumulados;
- Na coluna M contabilizar o nº da experiência;
- Na coluna N calcular a frequência relativa de sucesso, à medida que se vão realizando experiências.

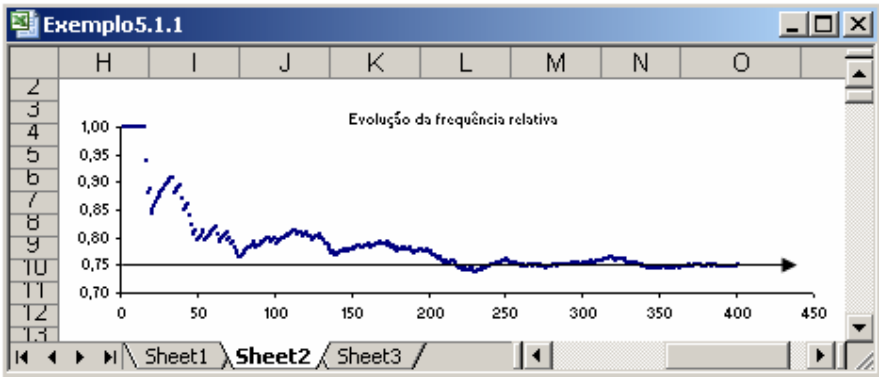
O processo anterior é apresentado na figura seguinte. Por uma questão de espaço só apresentamos a parte inicial e a parte final da tabela:

Exemplo5.1.1.xls

| | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|
| 1 | 1º filho | 2º filho | 3º filho | Sucesso | Nºsuc | Nºexp | fre.rel |
| 2 | F | F | M | 1 | 1 | 1 | 1,000 |
| 3 | F | F | M | 1 | 2 | 2 | 1,000 |
| 4 | M | F | M | 1 | 3 | 3 | 1,000 |
| 5 | F | M | M | 1 | 4 | 4 | 1,000 |
| 6 | F | M | F | 1 | 5 | 5 | 1,000 |
| 7 | M | F | F | 1 | 6 | 6 | 1,000 |
| 8 | M | F | F | 1 | 7 | 7 | 1,000 |
| 9 | M | F | M | 1 | 8 | 8 | 1,000 |
| 10 | F | M | M | 1 | 9 | 9 | 1,000 |
| 11 | M | M | F | 1 | 10 | 10 | 1,000 |
| 12 | M | F | M | 1 | 11 | 11 | 1,000 |
| 13 | F | F | M | 1 | 12 | 12 | 1,000 |
| 14 | F | F | M | 1 | 13 | 13 | 1,000 |
| 15 | F | M | F | 1 | 14 | 14 | 1,000 |
| 16 | M | M | F | 1 | 15 | 15 | 1,000 |
| 17 | F | F | F | | 15 | 16 | 0,938 |
| 18 | F | F | F | | 15 | 17 | 0,882 |
| 19 | M | F | F | 1 | 16 | 18 | 0,889 |
| 20 | M | M | M | | 16 | 19 | 0,842 |
| 21 | F | M | M | 1 | 17 | 20 | 0,850 |
| 22 | M | F | F | 1 | 18 | 21 | 0,857 |
| 23 | M | M | F | 1 | 19 | 22 | 0,864 |
| 388 | M | M | F | 1 | 290 | 387 | 0,749 |
| 389 | M | M | M | | 290 | 388 | 0,747 |
| 390 | M | M | F | 1 | 291 | 389 | 0,748 |
| 391 | F | M | F | 1 | 292 | 390 | 0,749 |
| 392 | F | F | F | | 292 | 391 | 0,747 |
| 393 | M | M | F | 1 | 293 | 392 | 0,747 |
| 394 | M | F | F | 1 | 294 | 393 | 0,748 |
| 395 | M | M | M | | 294 | 394 | 0,746 |
| 396 | F | F | M | 1 | 295 | 395 | 0,747 |
| 397 | F | M | M | 1 | 296 | 396 | 0,747 |
| 398 | M | F | F | 1 | 297 | 397 | 0,748 |
| 399 | M | M | F | 1 | 298 | 398 | 0,749 |
| 400 | M | F | F | 1 | 299 | 399 | 0,749 |
| 401 | F | F | M | 1 | 300 | 400 | 0,750 |

Sheet1 Sheet2 Sheet3

Como se verifica, a frequência relativa estabiliza à volta do valor 0,75, pelo que dizemos que 0,75 é uma estimativa para a probabilidade pretendida (O valor calculado, teoricamente, para esta probabilidade é de 0,75). A título de curiosidade acrescentamos que o resultado da simulação ao fim de 100, 200 e 300 repetições, foi respectivamente 0,790, 0,775 e 0,753. Apresentamos a evolução da frequência relativa na seguinte representação gráfica:



### 5.3.2 –Probabilidade Laplaciana (ou definição clássica)

Considerando ainda a experiência que consiste no lançamento de um dado *equilibrado*, em que podemos à partida considerar que cada resultado (saída de uma face) é igualmente possível, qual a *probabilidade* de sair a face 4?

Como temos 6 faces e existe uma face com o número 4, então o número de possibilidades é de 1 em 6. Assim a probabilidade de sair a face 4 é 1/6. O mesmo se passa com qualquer das outras faces.

Se dado um baralho de cartas, pretendermos saber qual a probabilidade de sair o ás de paus, como temos uma carta favorável para a nossa pretensão (ás de paus) de entre 52 possíveis, então a probabilidade pretendida é 1/52.

Mais geralmente, se o espaço de resultados S é constituído por um número finito **n** de elementos - acontecimentos elementares, todos eles *igualmente possíveis*, a **probabilidade** de cada resultado ou acontecimento elementar é **1/n** (princípio da simetria ou da razão insuficiente).

Considerando de novo a experiência do lançamento do dado, qual a probabilidade de se realizar o acontecimento "sair uma face par"?

Neste momento temos 3 faces favoráveis, de entre 6 possíveis, pelo que a probabilidade pretendida é de 3/6 ou 1/6 + 1/6 +1/6, que é a soma das probabilidades dos resultados que nos interessam.

Definida intuitivamente a probabilidade de um acontecimento elementar, define-se **Probabilidade de um acontecimento A** e representa-se por **P(A)**, como sendo a soma das probabilidades dos resultados que compõem A.

**Definição clássica de Probabilidade** – Define-se probabilidade do acontecimento A como sendo a razão entre o número de resultados **favoráveis** a A (resultados que compõem A) - $\mathbf{n_A}$ e o número de resultados **possíveis** (resultados que constituem S, admitindo-se o princípio da simetria) - **n**:

$$P(A) = \frac{n_A}{n}$$

**Exemplo 12 -** Considere a seguinte experiência aleatória, que consiste em seleccionar dois macacos ao acaso, de entre a seguinte lista, aos quais será administrado um certo medicamento.

| Macaco | Tipo | Idade |
|---|---|---|
| 1 | Baboon | 6 |
| 2 | Baboon | 8 |
| 3 | Spider | 6 |
| 4 | Spider | 6 |

Determine a probabilidade dos seguintes acontecimentos:

A : Os macacos escolhidos são do mesmo tipo

B : Os macacos escolhidos são da mesma idade

Resolução: O espaço dos resultados tem a seguinte forma

$$S = \{ (1,2), (1,3), (1,4), (2,3), (2,4), (3,4)\}$$

Os acontecimentos A e B são

$$A = \{(1,2), (3,4)\} \quad e \quad B = \{(1,3), (1,4), (3,4)\}$$

Então, de acordo com a definição clássica de probabilidade, temos:

$$P(A) = \frac{2}{6} \text{ e } \quad P(B) = \frac{3}{6}$$

**Exemplo 13** - Considere a experiência que consiste em registar o dia de anos de cada um dos 10 alunos que foram seleccionados ao acaso, numa determinada turma. Qual a probabilidade de não haver dois alunos que façam anos no mesmo dia? ( Considere o ano com 365 dias)

Resolução:

O número de possibilidades para os dias de anos dos 10 alunos é $365^{10}$. O número de resultados favoráveis será: 365 x 364 x 363 x 362 x 361 x 360 x 359 x 358 x 357 x 356. Então a probabilidade pretendida será

$$\frac{365 \times 364 \times 363 \times 362 \times 361 \times 360 \times 359 \times 358 \times 357 \times 356}{365^{10}}$$

**Exemplo 14** - Numa empresa de limpezas com 20 trabalhadores, estes têm de ser distribuídos pelos 4 serviços existentes. No serviço 1, dos 6 trabalhadores necessários, 4 pertencem a um determinado grupo étnico. Os outros serviços necessitam respectivamente de 4, 5 e 5 empregados. Qual a probabilidade de, numa distribuição aleatória, os 4 membros da dita etnia terem sido colocados no serviço 1?

Resolução:

O número de maneiras possíveis pelas quais os 20 empregados se podem distribuir pelos 4 serviços é

$$\frac{20!}{6!4!5!5!}$$

Tendo em consideração que os 4 elementos foram colocados no serviço 1, sobram 16 trabalhadores e o número de maneiras possíveis pelas quais se podem distribuir é

$$\frac{16!}{2!4!5!5!}$$

Então a probabilidade pretendida é $\dfrac{\frac{16!}{2!4!5!5!}}{\frac{20!}{6!4!5!5!}} = .0031$

Quando a hipótese de que os acontecimentos elementares são igualmente possíveis não se puder aplicar, e é o que acontece, por exemplo no caso de um dado em que se cortou um canto, e o

dado deixa de ser equilibrado, como é que poderemos atribuir probabilidade a um determinado resultado?

*Será sempre possível definir a probabilidade de um acontecimento, utilizando a definição anterior?*

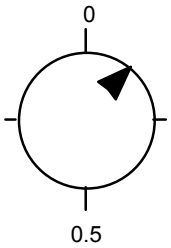
Nem sempre é possível construir um espaço de resultados, em que os resultados sejam igualmente possíveis. Por outro lado, esta definição de probabilidade é grandemente criticável, sob diversos pontos de vista, nomeadamente pelo facto de ser uma definição que tem por base a noção primitiva de igualmente possíveis, que é sinónimo de igualmente prováveis. Aparece assim o conceito de provável para definir probabilidade!

### 5.3.3 - Probabilidade subjectivista ou Bayesiana

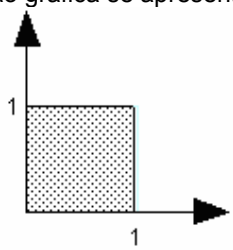
A maior parte das vezes não se pode repetir a experiência as vezes que se quer, nem tão pouco assumir que os resultados da experiência são igualmente possíveis. Por exemplo, qual a probabilidade de um aluno obter uma nota superior a 14 na disciplina de IPE, onde se encontra matriculado no 1º semestre? Nem é desejável que a experiência se repita, nem devemos atribuir igual possibilidade aos acontecimentos nota superior a 14 e nota menor ou igual que 14. No entanto, se formos ver o currículo do aluno poderemos atribuir uma probabilidade elevada (ou baixa) ao acontecimento em causa. A probabilidade diz-se, neste caso, subjectiva, pois foi baseada em informação anterior e num julgamento subjectivo. Uma vez que existe algo de arbitrário na atribuição de probabilidades a acontecimentos seguindo esta teoria, é de difícil aplicação, embora recentemente esteja a ter grande sucesso.

**Definição Bayesiana de Probabilidade** – atribui-se a um acontecimento uma probabilidade com base em experiência e informação anteriores.

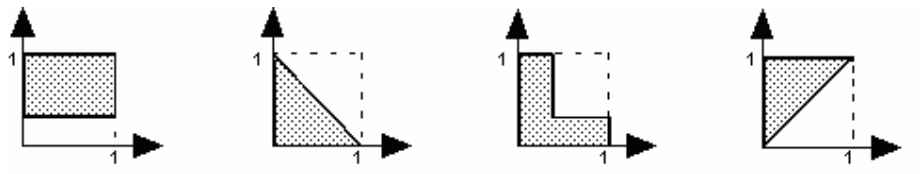
**Exemplo 15** – (Alpuim, T., 1997) – Suponha que vamos rodar uma roleta calibrada de 0 a 1, duas vezes consecutivas:




Se designarmos por $x_1$ o resultado da 1ª vez e por $x_2$ o resultado da 2ª vez, o espaço de resultados será $S=\{(x_1, x_2)\varepsilon[0,1)x[0,1)\}$, cuja repre-sentação gráfica se apresenta a seguir:




Alguns acontecimentos associados a este espaço de resultados são exemplificados a seguir:




A = $\{x_2>1/3\}$ B = $\{x_2+x_1<1\}$ C = $\{\min(x_1,x_2)<1/3\}$ D = $\{x_2 > x_1\}$

Como calcular a probabilidade destes acontecimentos e de outros do espaço de resultados S? Teríamos de calcular a frequência relativa para um número suficientemente grande de repetições da experiência, para todos os acontecimentos de S, que tem um número infinito, não numerável, de subconjuntos, o que tornaria a tarefa impraticável. Neste caso seria intuitivo pensar que a probabilidade associada a um acontecimento A $\varepsilon$ [0,1)x[0,1) é proporcional à sua área, ou seja,

P(A) = Área de A/Área de S = Área de A

No entanto, o que pretendemos não é uma forma de atribuir probabilidades que sirva para um determinado espaço de resultados, mas sim uma forma mais geral, que possa ser aplicada a todos os espaços amostrais, quer sejam finitos ou infinitos. Somos assim conduzidos à definição axiomática de Probabilidade.

### 5.3.4 - Definição axiomática de Probabilidade

Uma definição mais rigorosa e consequentemente mais consistente de Probabilidade, embora menos intuitiva, pode ser dada introduzindo um conjunto de regras ou **axiomas**, nomeadamente a Axiomática de Kolmogorov, a que deve obedecer uma função P, quando aplicada a subconjuntos de um espaço de resultados S.

Dado um espaço de resultados S, finito, representemos por **W** uma família de subconjuntos de S (acontecimentos), tais que :

i) Se o acontecimento **A** $\varepsilon$ **W** , então $\overline{A}$ $\varepsilon$ **W**

ii) Se os acontecimentos **A e B** $\varepsilon$ **W** , então **A**$\cup$**B** $\varepsilon$ **W**

iii) S está em **W**

Vamos ver seguidamente o processo de atribuir probabilidades a todos os acontecimentos de **W**, construindo uma teoria, à custa de um conjunto de três axiomas.

**Axiomática de Kolmogorov**

Dado o par **(S**,**W**) a cada elemento **A** $\varepsilon$ **W**, associa-se um número que se chama **Probabilidade** e se representa por **P(A)**. As probabilidades associadas aos acontecimentos de uma mesma família de acontecimentos satisfazem as seguintes condições ou axiomas :

***1º axioma*** - Qualquer que seja o acontecimento A, **P(A)≥0**

***2º axioma*** - A probabilidade do espaço de resultados é **1**

***3º axioma*** - Se os acontecimentos A e B são disjuntos, isto é,

$$\mathbf{A \cap B = \phi}, \text{ então } \mathbf{P(A \cup B) = P(A) + P(B)}$$

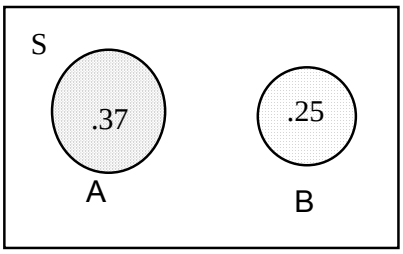
Este axioma é chamado de axioma da aditividade finita e não pode ser generalizado para uniões infinitas. Se admitirmos que o espaço de resultados é infinito numerável (Um conjunto diz-se numerável se pudermos estabelecer uma aplicação bijectiva entre ele e os naturais), $S=\{s_1, s_2, \ldots\}$, então seria desejável que para qualquer subconjunto A de S, finito ou não, a sua probabilidade fosse a soma das probabilidades dos acontecimentos elementares que o compõem. Neste caso, resolve-se o problema substituindo o axioma 3, pelo seguinte axioma:

Axioma 3* - $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ se $A_i \cap A_j = \varnothing$ para todo o $i \neq j$

**Exercício**: Verifique que a definição frequencista e a definição clássica de Probabilidade, conduzem a Probabilidades segundo a axiomática das Probabilidades, isto é, verifica os axiomas anteriores.

**Exemplo 16 -** Num estudo sobre o comportamento da gaivota, considere os seguintes acontecimentos: "a gaivota andava a voar" e "a gaivota estava na água". Admita que os acontecimentos anteriores têm, respectivamente, as probabilidades 0.37 e 0.25. Admitindo ainda que estes acontecimentos representam o comportamento de uma determinada gaivota, num determinado instante:

a) Represente a situação anterior utilizando diagramas de Venn. Será que os acontecimentos são disjuntos?




A - Voar

B - Estar na água

Os acontecimentos são disjuntos.

b) Qual a probabilidade de que a gaivota esteja a voar **ou** na água?

$$\begin{aligned} P(A \textbf{ ou } B) &= P(A \cup B) = P(A) + P(B) \\ &= 0.37 + 0.25 \\ &= 0.62 \end{aligned}$$

c) Qual a probabilidade de que a gaivota esteja a voar **e** na água? (Ver propriedade 1.a seguir)

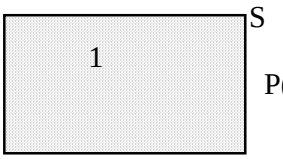
$$P(A \textbf{ e } B) = P(A \cap B) = P(\phi) = 0$$

**Observação:** Em linguagem de conjuntos o "ou" é traduzido pela união, enquanto que o "e" é a intersecção.

**Propriedades da Probabilidade**

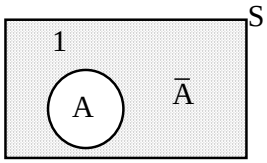
Com a ajuda de diagramas de Venn, e tendo em consideração os axiomas das Probabilidades, facilmente se mostram as seguintes propriedades para a Probabilidade:

**1** - $P(\phi) = 0$



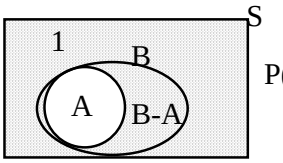

$$P(S \cup \phi) = P(S) + P(\phi) = 1$$

**2** - $P(\overline{A}) = 1 - P(A)$




$$P(A \cup \overline{A}) = P(A) + P(\overline{A}) = 1$$

**3** - Se $A \subset B$ então $P(A) \leq P(B)$



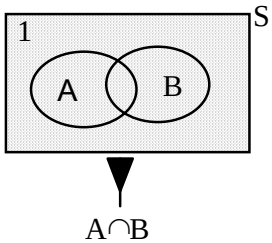

$$P(B) = P(A) + \underbrace{P(B-A)}_{0}$$

**4** - Qualquer que seja o acontecimento A, $0 \leq P(A) \leq 1$

Corolário do resultado anterior.

**5** - Quaisquer que sejam os acontecimentos A e B,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$




$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

**Exemplo 17 -** Num restaurante registaram-se, durante bastante tempo, os pedidos dos clientes, tendo-se chegado à conclusão que, para terminar a refeição, 20% dos clientes pedem só sobremesa, 40% pedem só café e 30% pedem sobremesa e café.

a) Construa um diagrama de Venn para ilustrar a situação anterior.

b) Determine a probabilidade do acontecimento “pedir café”.

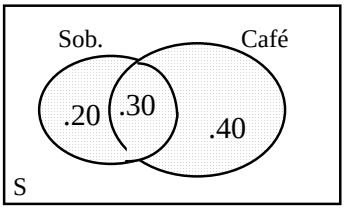
c) Determine a probabilidade do acontecimento “não pedir sobremesa”.

d) Determine a probabilidade do acontecimento “nem pede café nem sobremesa”.

e) Determine a probabilidade do acontecimento “pedir café ou sobremesa”.

f) Os acontecimentos “pedir café” e “pedir sobremesa” são disjuntos?

Resolução: a)




Sob - "Pedir sobremesa"

Café - "Pedir café"

b) P(Café) = .30 + .40 = .70

c) $P(\overline{Sob}) = 1 - P(Sob)$

$= 1 - .50 = .50$

d) $P(\overline{Café\ ou\ Sob}) = 1 - P(Café\ ou\ Sob)$

$= 1 - .90 = .10$

e) P(Café ou Sob) = .90

f) Os acontecimentos não são disjuntos

**Nota histórica** (Adaptado de Freedman, 1991) - **O paradoxo do Cavaleiro De Méré**

No século XVII, os jogadores Franceses costumavam fazer apostas sobre os seguintes acontecimentos: 1º jogo: lançar 4 dados e sair pelo menos um ás (chama-se ás à face com 1 pinta); 2º jogo: lançar 24 vezes um par de dados e sair pelo menos um duplo-ás (um par de dados com as faces 1). Um nobre Francês, o Cavaleiro De Méré, pensava que estes dois acontecimentos tinham igual probabilidade. O seu raciocínio era o seguinte, relativamente ao primeiro jogo:

- No lançamento de um dado, tenho uma probabilidade 1/6 de obter um ás;
- Assim, em 4 dados tenho uma probabilidade 4x1/6 de obter pelo menos um ás:

O seu raciocínio relativamente ao segundo jogo era análogo:

- No lançamento de um par de dados tenho uma probabilidade 1/36 de obter um duplo-ás.
- Assim, em 24 lançamentos, terei uma probabilidade 24x1/36 de obter pelo menos um duplo-ás.

Com este argumento, ambos os acontecimentos tinham a mesma probabilidade, igual a 2/3. Mas a experiência mostrava que o primeiro acontecimento se observava mais vezes que o segundo! Esta contradição ficou conhecida como o paradoxo do Chevalier de Méré.

De Méré questionou o filósofo Blaise Pascal sobre este problema, e Pascal resolveu-o com a ajuda do seu amigo Pierre de Fermat. Fermat era um juíz e membro do parlamento, que é conhecido hoje pelas investigações matemáticas que fazia nas horas vagas. Fermat mostrou que De Méré utilizava a regra da adição (axioma 3) para acontecimentos que não eram mutuamente exclusivos ou disjuntos. Efectivamente é possível obter um às tanto no 1º como no 2º lançamento de um dado. Além do mais, levando o argumento de De Méré um pouco mais longe, concluiríamos que a probabilidade de obter um ás em 6 lançamentos de um dado seria 6/6, ou seja 1. Alguma coisa teria que estar mal.

A questão que se punha agora, era como calcular correctamente estas probabilidades. Pascal e Fermat resolveram o problema, com um tipo de raciocínio matemático, indirecto – o que normalmente deixa os não matemáticos com o sentimento de que estão a ser enganados. Efectivamente, numa resolução directa como a proposta por Galileu (ver secção 5.1) afundar-nos-íamos completamente: com 4 lançamentos de um dado há $6^4$ = 1 296 resultados possíveis; com 24 lançamentos de um par de dados há $36^{24} \approx 2.2 \times 10^{37}$ resultados possíveis. Infelizmente a conversa entre Pascal e Fermat perdeu-se para a história, mas apresentamos seguidamente uma reconstrução.

*Pascal*. Olhemos então em primeiro lugar para o primeiro jogo.

*Fermat*. Vamos a isso. A probabilidade de ganhar é difícil de calcular, pelo que vamos tentar calcular a probabilidade do acontecimento complementar: a de perder. Então

Probabildade de ganhar = 1 – probabilidade de perder

*Pascal*. De acordo. O jogador perde quando nenhum dos 4 dados mostrar um ás. Mas como é que calcula a probabilidade?

*Fermat*. Parece complicado. Vamos começar com um dado. Qual a probabilidade que o primeiro dado não mostre um ás?

*Pascal*. Tem que mostrar entre o 2 e o 6, pelo que essa probabilidade será 5/6.

*Fermat*. É isso. Agora, qual a probabilidade que os primeiros dois lançamentos não mostrem ases?

*Pascal*. A probabilidade que o primeiro lançlamento do dado não mostre um ás é 5/6 = 0.83(3), ou seja, podemos dizer que se espera que em 83,(3)% das vezes que se faz o primeiro lançamento não saia ás. Para que não haja ases nos dois lançamentos, esperamos que em 83,(3)% dessas vezes também não haja ás no segundo lançamento. Como 83.(3)% de 83.(3)% é 83.(3)%x83.(3)%=69,(4)%, deveremos esperar que em 69,(4)% das vezes não haja ases nos dois lançamentos. Repare-se que 69.(4)% não é mais do que $5/6x5/6=(5/6)^2$, ou seja, o produto da probabilidade de não sair ás no primeiro lançamento pela probabilidade de não sair ás no segundo lançamento.

*Fermat*. Então e com 3 lançamentos?

*Pascal*. Será $5/6x5/6x5/6 = (5/6)^3$

*Fermat*. Sim. E agora com 4 lançamentos?

*Pascal*. Deve ser $(5/6)^4$

*Fermat*. Está bem. Significa que se tem uma probabilidade de cerca de 48.2% de perder. Agora

Probabildade de ganhar = 100% – 48.2% = 51.8%

*Fermat*. Então a probabiliddae de ganhar o primeiro jogo é um pouco superior a 50%. E no que diz respeito ao segundo jogo?

*Pascal*. Bem, no lançamento de um par de dados, há uma possibilidade em 36 de obter um duplo-ás, e 35 possibilidades em 36 de não o obter. Pelo mesmo argumento utilizado para o primeiro jogo, em 24 lançamentos de um par de dados, a probabilidade de não obter um duplo-ás é $(35/36)^{24}$.

*Fermat*. Que é cerca de 50.9%. Então como esta é a probabilidade de perder, a

Probabildade de ganhar = 100% – 50.9% = 49.1%

*Pascal*. Exactamente, o que dá uma probabilidade um pouco inferior a 50%. Cá está a razão pela qual se ganhava o segundo jogo com menos frequência que o primeiro. Mas teria de lançar o dado um grande número de vezes para se aperceber da diferença.

## 5.4 – Probabilidade condicional e independência.

### 5.4.1 – Probabilidade condicional

Num exemplo do início do capítulo referimos que a moeda não tem memória…. Efectivamente os sucessivos lançamentos que se fazem com uma moeda são *independentes*, o que significa que não possamos prever o que se vai verificar no próximo lançamento, com base no que se passou em lançamentos anteriores.

Suponhamos agora o seguinte exemplo: Considera-se um baralho de cartas e extrai duas cartas. Ganha 100 euros se a segunda carta for um rei de copas. Qual a probabilidade de ganhar os 100 euros?

Admita que joga este jogo segundo 2 cenários:

1º cenário – Não lhe permitem que veja a 1ª carta;

2º cenário – Quando retira a 1ª carta olha e vê que é o 7 de espadas.

Para obter aquela probabilidade podemos fazer o seguinte raciocínio:

1º cenário - se o baralho está embaralhado, como é pressuposto, a probabilidade do rei de copas estar na 2ª posição é 1/52, já que há 52 posições possíveis, todas igualmente possíveis, das quais só uma é favorável. Assim, P(Rei de copas) = 1/52.

2º cenário – Neste caso temos 51 cartas por uma ordem aleatória, e estamos interessados numa delas que é o rei de copas. Então P(Rei de copas) = 1/51.

Embora o acontecimento de que pretendíamos calcular a probabilidade fosse o mesmo nos dois casos, os contextos eram diferentes. No 1º caso estávamos à procura da probabilidade de na 2ª carta estar o Rei de copas, independentemente do que estivesse na 1ª carta, enquanto que no 2º caso estávamos à procura da probabilidade de na segunda carta estar o Rei de copas, *condicional* a que na 1ª carta estivesse o 7 de espadas. A esta probabilidade chamamos *probabilidade condicional*.

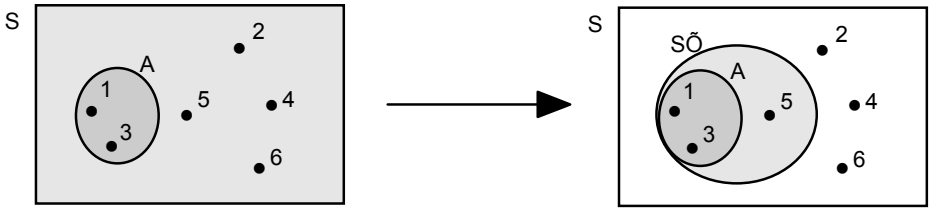
O conceito de probabilidade condicional é um dos conceitos mais importantes da Teoria da Probabilidade e está relacionado com o facto de em muitas situações em que se pretende calcular a probabilidade de um acontecimento, já se dispor de alguma informação sobre o resultado da experiência, a qual permite actualizar a atribuição de probabilidade a esse acontecimento. É uma noção, em geral, intuitiva, quando é aplicada no cálculo de probabilidades de cadeias de acontecimentos (ao retirar bolas de uma urna sucessivamente, sem reposição, a composição da urna altera-se e a probabilidade de se retirar certo tipo de bola depende dos tipos que saíram nas extracções anteriores).

Outro tipo de exemplos que conduzem facilmente à noção de probabilidade condicional são os que envolvem a "extracção" (ou escolha) ao acaso de um indivíduo de uma população cujos indivíduos estão classificados segundo os níveis de duas (ou mais) categorias (escolha ao acaso de um aluno de uma turma onde há rapazes, raparigas, filhos únicos e não filhos únicos).

Notar ainda que em situações de escolha aleatória de um indivíduo de uma população, a probabilidade de ocorrência de A condicional à ocorrência de B não é mais do que a probabilidade de ocorrência de A quando se escolhe ao acaso um indivíduo da subpopulação constituída unicamente pelos indivíduos que verificam a característica determinada pelo acontecimento B.

Consideremos (Graça Martins, M. E. et al, 1999), por exemplo, a experiência aleatória que consiste em lançar um dado e verificar o número de pintas que sai. A probabilidade do acontecimento A, sair "1 ou 3 pintas" é 2/6, já que o nosso espaço de resultados S, é constituído por 6 casos igualmente possíveis, dos quais 2 são favoráveis à realização de A. Se, no entanto, pretendermos a probabilidade desse mesmo acontecimento, sabendo de antemão que saiu um número de pintas ímpar, neste momento já o espaço de resultados S', é constituído por 3 resultados, igualmente possíveis, dos quais 2 são favoráveis, pelo que a probabilidade pretendida é 2/3, o dobro da obtida anteriormente, quando não tínhamos nenhuma informação.

Exemplificando com um diagrama de Venn




Vejamos ainda uma outra situação. Suponhamos, por exemplo, a experiência aleatória que consiste em retirar 2 bolas sem reposição, de uma caixa contendo 4 bolas brancas B1, B2, B3 e B4 e 3 bolas pretas P1, P2, P3. Os **N** diferentes resultados obtidos na realização da experiência são:

| | | | | | |
|---|---|---|---|---|---|
| B1B2 | B1B3 | B1B4 | B1P1 | B1P2 | B1P3 |
| B2B1 | B2B3 | B2B4 | B2P1 | B2P2 | B2P3 |
| B3B1 | B3B2 | B3B4 | B3P1 | B3P2 | B3P3 |
| B4B1 | B4B2 | B4B3 | B4P1 | B4P2 | B4P3 |
| P1B1 | P1B2 | P1B3 | P1B4 | P1P2 | P1P3 |
| P2B1 | P2B2 | P2B3 | P2B4 | P2P1 | P2P3 |
| P3B1 | P3B2 | P3B3 | P3B4 | P3P1 | P3P2 |

Representando por n(Branca1) e n(Branca2), respectivamente, o número de vezes em que se verificou o acontecimento Branca1 – "saiu bola branca na 1ª extracção" e o número de vezes que se realizou o acontecimento Branca2 – "saiu bola branca na 2ª extracção", e por n(Branca1□Branca2) o número de vezes que se realizou o acontecimento Branca1□Branca2 – "saiu branca na 1ª e 2ª extracções", temos:

P(Branca1) = 24/42, P(Branca2) = 24/42, P(Branca1□Branca2) = 12/42

Suponhamos, no entanto, que sabíamos que tinha saído branca na 1ª extracção, isto é, que se tinha verificado o acontecimento Branca1. Qual a probabilidade de sair branca na 2ª extracção, isto é de se verificar o acontecimento Branca2, tendo em conta esta informação adicional? Neste momento o espaço de resultados foi substancialmente reduzido, pois o número de resultados possíveis é 24 (ter saído branca na 1ª extracção),

| | | | | | |
|---|---|---|---|---|---|
| B1B2 | B1B3 | B1B4 | B1P1 | B1P2 | B1P3 |
| B2B1 | B2B3 | B2B4 | B2P1 | B2P2 | B2P3 |
| B3B1 | B3B2 | B3B4 | B3P1 | B3P2 | B3P3 |
| B4B1 | B4B2 | B4B3 | B4P1 | B4P2 | B4P3 |

dos quais só 12 é que são favoráveis, pelo que

P(Branca2 sabendo que Branca1) = 12/24

À probabilidade anterior chamamos *probabilidade condicional* do acontecimento Branca2, sabendo que (ou dado que) se realizou o acontecimento Branca1, e representamos por P(Branca2|Branca1).

Repare-se que

$$P(\text{Branca2}|\text{Branca1}) = \frac{n(\text{Branca1} \cap \text{Branca2})}{n(\text{Branca1})}$$

$$= \frac{\dfrac{n(\text{Branca1} \cap \text{Branca2})}{N}}{\dfrac{n(\text{Branca1})}{N}}$$

$$= \frac{P(\text{Branca1} \cap \text{Branca2})}{P(\text{Branca1})}$$

ou seja $$P(\text{Branca2}|\text{Branca1}) = \frac{P(\text{Branca1} \cap \text{Branca2})}{P(\text{Branca1})}$$

Assim, a *probabilidade condicional* de se realizar o acontecimento Branca2, sabendo que se realizou Branca1, é o quociente entre a probabilidade da realização de Branca1 e Branca2, e a probabilidade da realização de Branca1. Esta probabilidade condicional só tem sentido se P(Branca1) for superior a zero.

Seja S um espaço de resultados e P uma probabilidade nesse espaço. Dados dois acontecimentos A e B, com P(B)>0, define-se probabilidade condicional de A se B (ou probabilidade de A condicional à ocorrência de B) como sendo

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

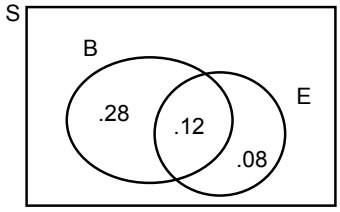
**Exemplo 18** (Parzen, 1960) – Consideremos uma família com dois filhos e suponhamos que existe igual probabilidade de cada filho ser rapaz ou rapariga. Qual a probabilidade de que ambos os filhos sejam rapazes dado que: (i) o filho mais velho é um rapaz, (ii) pelo menos um dos filhos é rapaz.

O espaço de resultados associado ao fenómeno em estudo, isto é, uma família ter dois filhos é S = {MM, MF, FM, FF}. Todos estes resultados são igualmente possíveis tendo em consideração o facto de ser igualmente provável um filho ser rapaz (M) ou rapariga (F). Pretende-se a probabilidade de ambos serem rapazes, sabendo que (i) o filho mais velho é rapaz – este condicionamento provoca que o espaço de resultados se reduza a S' = {MM, MF}, donde P(MM) = 1/2. Condicionando agora no acontecimento (ii) pelo menos um dos filhos é rapaz, já o espaço de resultados é S'' = {MM, MF, FM} pelo que a probabilidade pretendida é P(MM) = 1/3.

Nota: Repare-se que a probabilidade de que "ambos os filhos sejam rapazes" é diferente consoante nada se saiba sobre o sexo dos filhos ou haja conhecimento parcial sobre o sexo de um dos filhos. No primeiro caso a probabilidade é 1/4.

**Exemplo 19** (Siegel et al, 1988) -. Consideremos a experiência aleatória que consiste em observar, numa dada multinacional, a impressão causada (boa ou má) na entrevista dos candidatos a um emprego, assim como se conseguem ou não o emprego. Pensemos nos

acontecimentos B – “o candidato causa boa impressão” e E – “o candidato consegue o emprego”. Suponhamos que os acontecimentos anteriores estão representados num diagrama de Venn e que se conhecem as probabilidades assinaladas:




No diagrama de Venn os números indicados representam:

P(B–E) = 0.28

P(E–B) = 0.08

P(B□E) = 0.12

A partir do diagrama anterior sabemos que

$$P(\text{“Conseguir emprego”}) = 0.12 + 0.08 = 0.20$$

o que significa que 20% dos candidatos, que vão à entrevista, conseguem o emprego. Será que o facto de causar boa impressão, aumenta as possibilidades de ser bem sucedido, na obtenção do emprego? Isto é, será que a informação adicional de que "um candidato causou boa impressão" tem efeito na probabilidade de obter o emprego? Para responder a esta questão, temos de nos cingir unicamente aos candidatos que causam boa impressão, em vez de considerarmos todos os candidatos. A dimensão deste grupo é 40% de todos os candidatos, já que

$$P(\text{"Causar boa impressão"}) = 0.28 + 0.12 = 0.40$$

Para este totao de 40%, qual o contributo dos que conseguem o emprego? A resposta obtém-se restringindo este grupo aos que conseguem o emprego

$$P(\text{"Causar boa impressão } \mathbf{e} \text{ Conseguir o emprego"}) = 0.12$$

Finalmente podemos calcular a probabilidade de uma pessoa que causou boa impressão, conseguir o emprego. Esta probabilidade é dada pela resposta à seguinte questão " 0.12 que percentagem é de 0.40"? , resposta esta que se obtém dividindo 0.12 por 0.40, como aliás se deduz da definição anteriormente dada de probabilidade condicional:

$$P(\text{"Conseguir o emprego"} \mid \text{"Causou boa impressão"}) = \frac{0.12}{0.40} = 0.30$$

Vemos que a probabilidade de conseguir o emprego aumentou de 20% para 30%, com a informação adicional disponível. Isto significa que 30% dos candidatos que causam boa impressão, conseguem o emprego, comparados com unicamente 20% dos candidatos em geral (causando ou não boa impressão). Intuitivamente esperávamos que o facto de um candidato causar boa impressão, aumentasse as suas possibilidades de sucesso, e o que acabamos de medir foi precisamente quão grande é esse efeito.

**Exemplo 20** (Pestana, D. et al, 2002) - Numa caixa estão 5 moedas, duas delas com face (F) em ambos os lados, duas com coroa (C) em ambos os lados, e uma com F num dos lados e C no outro. Escolhe-se uma moeda ao acaso, observando-se no lado que fica virado para cima F. Qual a probabilidade do outro lado ser C?

Estão em jogo 5 faces favoráveis num total de 10 lados, pelo que

$$P(C_2 \cap F_1) = \frac{1}{10} \text{ pelo que } P(C_2 \mid F_1) = \frac{\frac{1}{10}}{\frac{5}{10}} = \frac{1}{5}$$

**Árvore de probabilidades**

Uma árvore de probabilidades é uma representação esquemática, especialmente pensada para apresentar todos os casos possíveis e respectivas probabilidades, em situações que envolvam uma sequência de experiências aleatórias cujos espaços de resultados sejam de dimensão reduzida. Do nó inicial da árvore partem tantos ramos, quantos os acontecimentos elementares do espaço de resultados correspondente à primeira experiência aleatória. No extremos de cada ramo indica-se o acontecimento respectivo e por cima do ramo a sua probabilidade. Passando ao nível seguinte, o extremo de cada ramo será um nó para a segunda experiência aleatória. A informação é registada de forma idêntica à descrita para o primeiro nó, mas tendo agora em conta que as probabilidades são as condicionais ao acontecimento que figura no novo nó. O processo repete-se até atingir a última experiência aleatória (Graça Martins, M. E. e Loura, L. C. C, 2003).

**Exemplo 21** - Duas equipas de baseball, muito equilibradas, disputam um torneio de 4 jogos. Regista-se o resultado de cada jogo (não está previsto o empate).

a) Descreva o espaço de resultados.

b) Seja A o acontecimento: A equipa 1 ganha exactamente 3 jogos. Quais os acontecimentos elementares que compõem A?

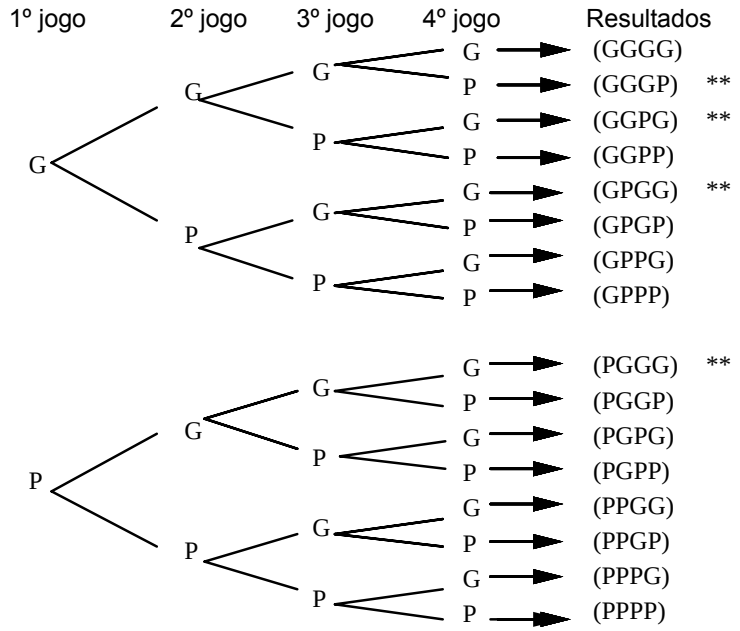
c) Determine a probabilidade do acontecimento A.

Resolução:

a) O espaço de resultados é constituído por todos os conjuntos de 4 elementos da figura seguinte, onde representamos por G e P respectivamente a equipa 1 ganha ou perde .

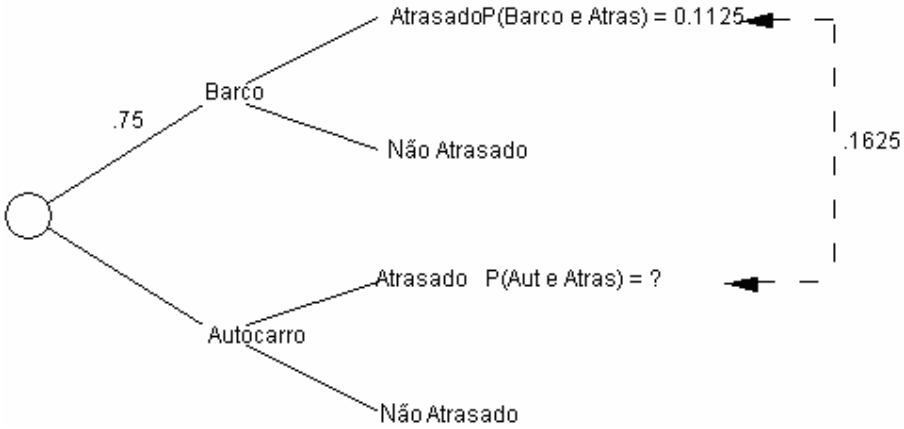
b) Os acontecimentos elementares que compõem A encontram-se assinalados com **.

c) Como os resultados favoráveis são 4 e os possíveis são 16, todos igualmente prováveis, obtemos que P(A) = 4/16 = 1/4




**Exemplo 22 -** Um indivíduo que trabalha em Lisboa, mas reside na margem Sul do Tejo, tem diariamente duas possibilidades para se dirigir ao trabalho: o barco ou o autocarro. Ele gosta muito de ir de barco, pelo que escolhe o barco 75% das vezes. A probabilidade de chegar atrasado ao trabalho é 16.25%. sabe-se ainda que a probabilidade de ir de barco e chegar atrasado é 11.25%. Qual a probabilidade de chegar atrasado, sabendo que veio de barco?

Vamos tentar construir uma árvore de probabilidades onde entre a informação anterior




A informação dada está representada no diagrama anterior. Contudo, à custa dessa informação podemos ir um pouco mais longe, calculando a probabilidade dos acontecimentos complementares.

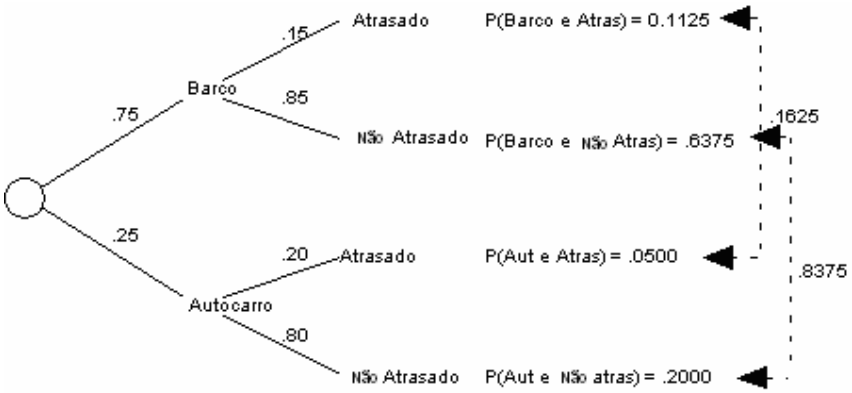
Qual a probabilidade de chegar atrasado **dado** que veio de barco?

$$P(\text{"chegar atrasado"}/\text{"veio de barco"}) = \frac{P(\text{" vir de barco e chegar atrasado"})}{P(\text{" vir de barco"})}$$

$$= \frac{.1125}{.75}$$

$$= .15$$

Esta probabilidade condicional coloca-se ao longo do traço superior, como se indica na figura seguinte, onde também preenchemos as bolas do lado direito com as respectivas probabilidades, o que nos permitiu chegar à seguinte árvore:




Considerando a árvore anterior, vemos que:

P("chegar atrasado **dado** que não veio de barco") = $\frac{.05}{.25}$ = .20

P("chegar atrasado **ou** vir de barco") = .1625 + .6375 = .80
ou
P("chegar atrasado **ou** vir de barco")= P("chegar atrasado")+P("vir de barco") - P("chegar atrasado **e** vir de barco")= .1625 + .75 - .1125 = .80

P("vir de barco" **dado que** "chegou atrasado") = $\frac{.1125}{.1625}$ = .69

P("não chegar atrasado **e** não vir de barco")=1-P("chegar atrasado **ou** vir de barco") = 1 - .80 = .20

**Exercício:** Seja $P_B(A)$ uma função definida da seguinte forma: dado o acontecimento B, com P(B)>0, então para qualquer acontecimento A tem-se $P_B(A)$= P(A|B). Mostre que $P_B(A)$ é uma Probabilidade, isto é, satisfaz a axiomática de Kolmogorov.

### 5.4.2 - Probabilidade da intersecção de acontecimentos ou probabilidade conjunta dos acontecimentos A e B ou regra do produto

Atendendo a que

$$P(A/B) = \frac{P(A \textbf{ e } B)}{P(B)}$$

vem

$$\boxed{P(A \textbf{ e } B) = P(B)\,P(A/B) \quad \text{ou} \quad P(A \textbf{ e } B) = P(A)\,P(B/A)}$$

ou com a notação de intersecção

$$\boxed{P(A \cap B) = P(B)\,P(A/B) \quad \text{ou} \quad P(A \cap B) = P(A)\,P(B/A)}$$

### 5.4.3 - Acontecimentos independentes

O conceito de probabilidade condicional permite-nos definir acontecimentos independentes, como sendo aqueles em que a informação acerca de um não ajuda a determinar a probabilidade de ocorrência do outro. De forma mais rigorosa, dados os acontecimentos A e B, com P(A)>0 e P(B)>0,

O acontecimento **A é independente do acontecimento B**, se a probabilidade de A se verificar, é igual à probabilidade condicional de A se verificar, dado que B se verificou

$$P(A) = P(A/B)$$

Se A é independente de B, então B é independente de A?

Efectivamente assim é! Repare-se que

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{P(B)P(A/B)}{P(A)} = \frac{P(B)P(A)}{P(A)} = P(B)$$

**Outra definição de independência de acontecimentos**

Dois acontecimentos A e B, são **independentes** se a probabilidade conjunta é igual ao produto das probabilidades de cada um deles

$$\boxed{P(A \cap B) = P(A)\,P(B)}$$

Esta definição de independência, embora não seja tão intuitiva, é utilizada com muita frequência, pois não é necessário impor restrições aos valores de P(A) e P(B). Verifique que as duas definições são equivalentes desde que P(A)>0 e P(B)>0.

**Exercício** ( Teaching Statistics, vol16, nº 2) - Tendo dois dados de 12 faces, em que cada um tem 7 faces vermelhas e 5 brancas, perguntou-se a 40 estudantes qual dos acontecimentos era mais provável, no lançamento dos dois dados:
i) Sair 2 faces vermelhas, ou
ii) Sair1 face vermelha e 1 branca.
Trinta e seis estudantes responderam que era mais provável sair 2 faces vermelhas. Está de acordo? Justifique.

Aos mesmos estudantes, mostraram-se 3 dados de 4 faces, cada um com 3 faces vermelhas e uma branca. No lançamento dos 3 dados, qual o acontecimento mais provável:
i) Sair 3 faces vermelhas, ou

ii) Sair 2 faces vermelhas e 1 branca?

Todos os estudantes responderam que o acontecimento i) era o mais provável. Está de acordo? Justifique.

**Exemplo 17** (cont.) - Considere de novo os dados do exemplo 18. Calcule as probabilidades dos seguintes acontecimentos:

a) Pedir café dado que pediu sobremesa R: .6

b) Não pedir café dado que pediu sobremesa R: .4

c) Pedir sobremesa dado que pediu café R: 3/7

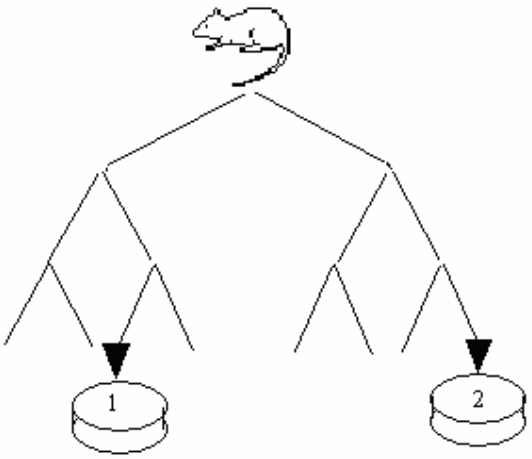
d) Pedir sobremesa dado que não pediu café R: 2/3

e) Será que o café e a sobremesa "ligam" bem? Ou seja, é mais provável um cliente pedir café se pediu sobremesa ou se não pediu sobremesa?

R: É mais provável pedir café se não pediu sobremesa (.8)

f) Os acontecimentos pedir café e pedir sobremesa são independentes? R: Não são independentes.

**Exemplo 23** - Um rato apresenta-se na entrada de um caminho com várias bifurcações, como se apresenta a seguir:




Sempre que se apresenta uma bifurcação o rato tem de optar por virar à esquerda ou à direita, nunca podendo voltar para trás. Em duas das saídas encontram-se dois belos queijos. Qual a probabilidade de o rato chegar a qualquer um dos queijos:

a) Se a probabilidade de virar á esquerda for igual à de virar à direita para todos os cruzamentos.

b) Se a probabilidade de virar à esquerda for 0.3 e a de virar à direita for 0.7.

Resolução: a) Como o rato tem sempre igual probabilidade de virar à esquerda ou à direita, as 8 saídas são todas igualmente possíveis. Como existem duas favoráveis, a probabilidade pretendida será 2/8 = 1/4.

b) Para chegar ao queijo 1 o rato tem de fazer o percurso ($D_1$**e**$E_2$**e**$D_3$), enquanto que para chegar ao queijo 2 terá de fazer ($E_1$**e**$E_2$**e**$E_3$), onde representamos por $D_1$ virar à direita no primeiro cruzamento, $E_2$ virar à esquerda no 2º cruzamento, etc. Então

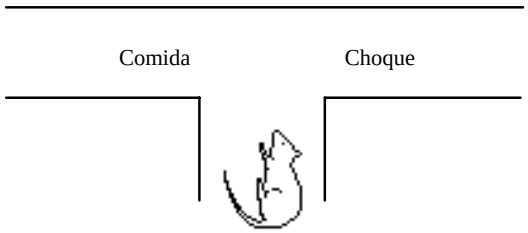
$P(D_1 \mathbf{e} E_2 \mathbf{e} D_3) = P(D_1)\, P(E_2)\, P(D_3)$
$= 0.7 \times 0.3 \times 0.7$
$= 0.147$
$P(E_1 \mathbf{e} E_2 \mathbf{e} E_3) = P(E_1)\, P(E_2)\, P(E_3)$
$= 0.3 \times 0.3 \times 0.3$
$= 0.027$

porque os acontecimentos são independentes, já que a probabilidade de o rato virar à esquerda ou á direita num determinado cruzamento não depende do que é que ele fez antes. Assim, a probabilidade pretendida é

$$0.147 + 0.027 = 0.174$$

Chamamos a atenção para o facto de neste caso não ser possível utilizar a definição clássica de probabilidade, pois as chegadas não são todas igualmente possíveis.

**Exemplo 24 -** Imagine a seguinte experiência laboratorial:




À primeira vez que se apresenta o cruzamento, o rato tem igual probabilidade de virar à esquerda ou à direita. À segunda vez, o rato se recebeu comida à primeira vez, vira à esquerda com probabilidade 0.6 e se recebeu um choque à primeira vez vira à direita com probabilidade 0.2. Calcule a probabilidade do rato virar à direita à segunda vez.

Resolução: Pretende-se a probabilidade do acontecimento "virar à direita à 2ª vez" que vamos representar por $D_2$. Repare-se que para que o rato tenha virado à direita (ou à esquerda) à 2º vez, é necessário que tenha feito uma de 2 coisas à 1ª vez: ou virar á direita ou à esquerda. Assim

$$D_2 \equiv \{(D_1 \mathbf{e} D_2)\ \mathbf{ou}\ (E_1 \mathbf{e} D_2)\}$$

Os acontecimentos $(D_1 \mathbf{e} D_2)$ e $(E_1 \mathbf{e} D_2)$ são disjuntos, pois não se podiam ter verificado simultaneamente, pelo que a probabilidade da sua união é igual à soma das probabilidades. Então

$$P(D_2) = P(D_1 \mathbf{e} D_2) + P(E_1 \mathbf{e} D_2)$$

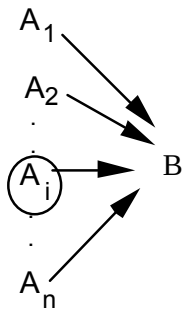
Vejamos agora a que é igual cada uma das parcelas da soma anterior:

$P(D_1 \mathbf{e} D_2) = P(D_1)\, P(D_2|D_1)$
$= 0.5 \times 0.2$
$= 0.10$
$P(E_1 \mathbf{e} D_2) = P(E_1)\, P(D_2|E_1)$
$= 0.5 \times (1 - 0.6) = 0.20$ donde $P(D_2) = 0.30$

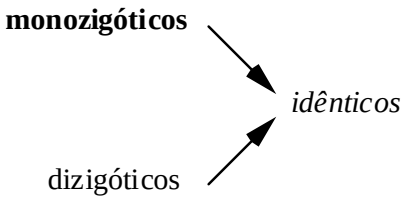
## 5.5 - Teorema de Bayes

O interesse do teorema de Bayes para problemas biológicos e em particular de genética, reside no facto seguinte:

Observa-se um acontecimento B, que nós sabemos ser susceptível de ter sido ocasionado por um qualquer dos acontecimentos $A_1$, $A_2$, … , $A_n$, mutuamente exclusivos. Pretende-se saber qual a probabilidade de ter sido o acontecimento $A_i$ e não outro qualquer a ocasionar B,




**Exemplo 25 -** Um homem de grupo sanguíneo AB e uma mulher de grupo sanguíneo O têm dois rapazes gémeos, de grupo sanguíneo A. Sendo os gémeos idênticos[1] (no nosso exemplo vamos considerar que idênticos significa terem o mesmo sexo e mesmo grupo sanguíneo) pretende-se calcular a probabilidade de serem monozigóticos.




De acordo com a notação introduzida para a probabilidade condicional, pretendemos calcular

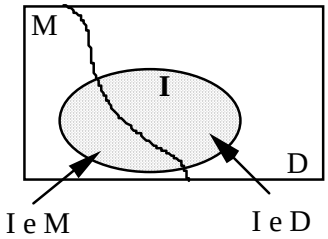
**P(monozigótico / idêntico)** = ?

Para facilitar a escrita, vamos introduzir a seguinte simplificação na notação:

monozigótico - M; dizigótico - D; idêntico - I

Repare-se agora no seguinte esquema:

[1] Dois gémeos são idênticos para um certo número de caracteres, cuja determinação genética está bem estudada. Para cada um destes caracteres esta identidade pode resultar de uma monozigotomia ou de uma coincidência entre irmãos dizigóticos, podendo-se medir esta coincidência em termos probabilísticos.




I = ( I e M ) ou ( I e D )

Por outro lado, existe a seguinte informação disponível, fornecida pelos biólogos:

P(M) = .30 e P(D) = .70

Como consideramos o acontecimento "idêntico" equivalente a ter o mesmo sexo e o mesmo grupo sanguíneo, obtemos as seguintes probabilidades

P(I / M) = P(mesmo sexo **e** mesmo grupo sang./M)=1[1]

P(I / D) = P(mesmo sexo **e** mesmo grupo sang./D)= $\frac{1}{2} \times \frac{1}{2}$

Para calcular esta última probabilidade deve ter-se em atenção o seguinte: a probabilidade de que o 2º gémeo seja um rapaz, assim como o 1º, é 1/2, e a probabilidade de que o grupo sanguíneo seja A é também 1/2, porque o grupo AB dá em média uma vez A e outra vez B e o grupo O é recessivo. Como estes acontecimentos são independentes, porque fazem intervir cromossomas diferentes, vem que a probabilidade conjunta pretendida é o produto das probabilidades dos acontecimentos envolvidos.

O nosso objectivo é o cálculo da probabilidade P(M / I ), a qual, de acordo com a expressão da probabilidade condicional, é dada por

(1) $$\mathbf{P(M / I)} = \frac{P(M \text{ e } I)}{P(I)}$$

No entanto, temos

P( M e I ) = P( M ) P ( I / M )

= $.30 \times 1$ = .30

P( I ) = P[( I e M ) ou ( I e D )]

= P( I e M ) + P ( I e D ) porque os acontecimentos I e M e I e D são disjuntos

= P( M ) P( I / M ) + P ( D ) P( I / D )

= $.30 \times 1 + .70 \times .25$

= .475

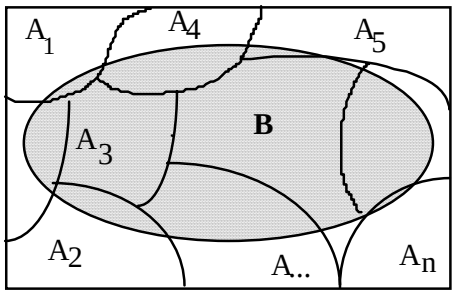
[1]Dois gémeos monozigóticos (gémeos verdadeiros) têm necessariamente o mesmo sexo e grupo sanguíneo.

Substituindo estes valores na fórmula (1), vem

$$\mathbf{P(\ M\ /\ I\ )} = \frac{.30}{.475} = .63$$

O resultado anterior é uma aplicação de um teorema atribuído a um padre presbiteriano inglês, Thomas Bayes, cuja generalização para n causas $A_1, A_2, \ldots, A_n$, é a seguinte:

**Teorema de Bayes**: Se $\{A_1, A_2, \ldots, A_n\}$ constituem uma partição do espaço de resultados, isto é, $A_i$ e $A_j$ são disjuntos dois a dois e a união dos acontecimentos $A_i$, é igual ao espaço, com $P(A_i)>0$, $i=1,2,\ldots,n$, então dado qualquer acontecimento B, com $P(B)>0$, tem-se




$$P(A_i / B) = \frac{P(A_i)\ P(B / A_i)}{\sum P(A_i)\ P(B / A_i)}$$

Observação: o teorema de Bayes permite-nos rever as probabilidades, mediante informação entretanto disponível. Assim, enquanto que às probabilidades $P(A_i)$ chamamos probabilidades **à priori**, às probabilidades $P(A_i/B)$, calculadas após a realização do acontecimento B, chamamos probabilidades **à posteriori**. Estas probabilidades são a base da **teoria subjectivista das Probabilidades**, já referida anteriormente.

No denominador da expressão que dá a probabilidade condicional, aparece uma expressão que só por si merece relevo especial, dada a sua importância, e que é a base da demostração do teorema de Bayes:

**Teorema da Probabilidade Total**

Se $\{A_1, A_2, \ldots, A_n\}$ constituem uma partição do espaço de resultados, isto é, $A_i$ e $A_j$ são disjuntos dois a dois e a união dos acontecimentos $A_i$, é igual ao espaço, com $P(A_i)>0$, $i=1,2,\ldots,n$, então dado qualquer acontecimento B, com $P(B)>0$, tem-se

$$P(B) = \sum_{i=1}^{n} P(B/A_i)\ P(A_i)$$

Dem: P(B) = P(B∩S) = $P(B \cap \bigcup_{i=1}^{n} A_i) = P(\bigcup_{i=1}^{n} (B \cap A_i)) = \sum_{i=1}^{n} P(B \cap A_i)$ porque se os $A_i$ constituem uma partição, o mesmo acontece com $B \cap A_i$.

**Exemplo 26** - Num centro de cálculo existem três impressoras A, B e C, que imprimem a velocidades diferentes. Os ficheiros são enviados para a primeira impressora que estiver disponível. A probabilidade de um ficheiro ser enviado para as impressoras A, B ou C é respectivamente .6, .3 e .1. Ocasionalmente a impressora avaria e destrói a impressão. As impressoras A, B e C avariam com probabilidades .01, .05 e .04. A impressão do seu ficheiro foi destruída! Qual a probabilidade de ter sido enviada para a impressora A?

Resolução: Vamos utilizar a seguinte notação para referenciar os acontecimentos:

A - enviar para a impressora A

B - enviar para a impressora B

C - enviar para a impressora C

D - impressão destruída

Dados: P(A) = .6 P(B) = .3 P(C) = .1

P(D/A) = .01 P(D/B) = .05 P(D/C) = .04

Pretende-se P(A/D) = ?

De acordo com a expressão da probabilidade condicional, temos

$$P(A/D) = \frac{P(AeD)}{P(D)}$$

Mas P(A **e** D) = P(A) P(D/A)

P(A **e** D) = .6 x .01 = .006 e

P(D) = P(A **e** D) + P(B **e** D) + P(C **e** D)

P(D) = P(A) P(D/A) + P(B) P(D/B) + P(C) P(D/C)

P(D) = $.6 \times .01 + .3 \times .05 + .1 \times .04$

P(D) = .025 P(A/D) = $\frac{.006}{.025}$ = .24

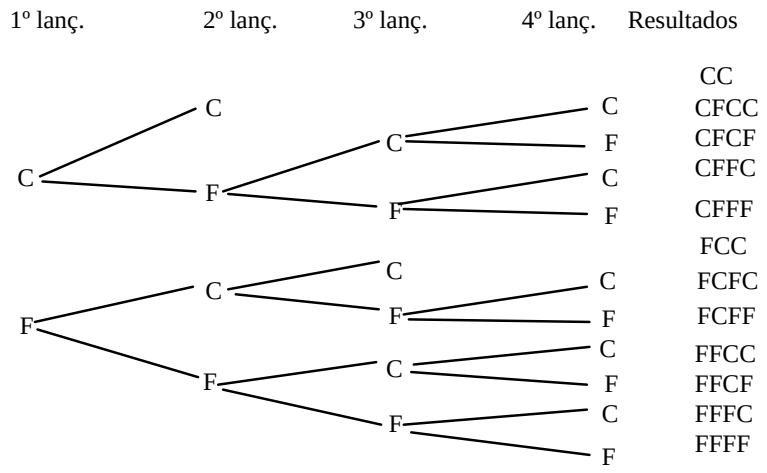
**Exercícios**

**1**. Considere a experiência aleatória que consiste em lançar uma moeda ao ar até sair coroa duas vezes consecutivas ou até se realizarem 4 lançamentos. Qual o espaço de resultados associado a este acontecimento?

Resolução:

Consideremos os dois resultados possíveis do lançamento da moeda: Coroa - **C** e Cara - **F**

Vamos construir um diagrama para exemplificar os sucessivos lançamentos:




S = { CC, CFCC, CFCF, … , FFFC, FFFF}

**2** . Sejam A, B e C três acontecimentos associados a um espaço de resultados S. Exprima com notação conveniente:

a) Pelo menos um dos acontecimentos ocorre

b) Quando muito um dos acontecimentos ocorre

c) Um e um só dos acontecimentos ocorre

d) Pelo menos dois dos acontecimentos ocorrem

e) Exactamente dois dos acontecimentos ocorrem

Resolução:

a) A **ou** B **ou** C

b) $(A \text{ e } \overline{B} \text{ e } \overline{C})$ **ou** $(\overline{A} \text{ e } B \text{ e } \overline{C})$ **ou** $(\overline{A} \text{ e } \overline{B} \text{ e } C)$ **ou** $(\overline{A} \text{ e } \overline{B} \text{ e } \overline{C})$

c) $(A \text{ e } \overline{B} \text{ e } \overline{C})$ **ou** $(\overline{A} \text{ e } B \text{ e } \overline{C})$ **ou** $(\overline{A} \text{ e } \overline{B} \text{ e } C)$

d) ( A **e** B) **ou** (B **e** C) **ou** (A **e** C)

e) $(A \text{ e } B \text{ e } \overline{C})$ **ou** $(\overline{A} \text{ e } B \text{ e } C)$ **ou** $(A \text{ e } \overline{B} \text{ e } C)$

**3** . Considere a experiência aleatória que consiste no lançamento de dois dados. Calcule a probabilidade do acontecimento "soma das pintas igual a 5".

Resolução:

S= {(i,j) : i, j = 1,…, 6}

"soma das pintas igual a 5" = A = {(1,4), (2,3), (3,2), (4,1)}

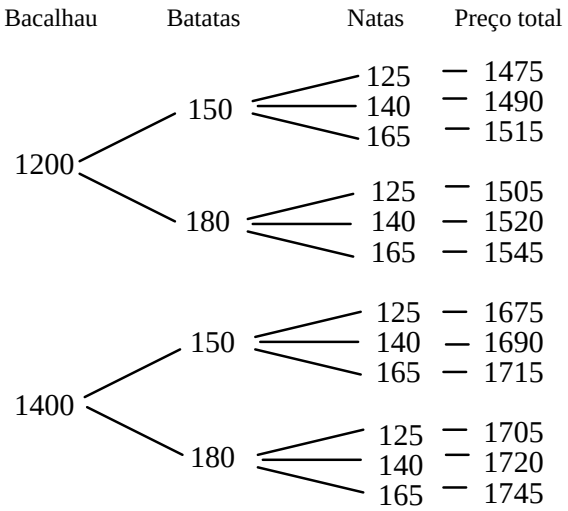
Utilizando a definição clássica de probabilidade, já que todos os resultados do espaço de resultados são igualmente possíveis, temos P(A) = $\frac{4}{36} = \frac{1}{9}$

**4**. Para confeccionar um prato de bacalhau com natas, pode-se optar por bacalhau médio ou bacalhau grande, cujo preço é respectivamente 1200$00 e 1400$00. As batatas podem custar 150$00 ou 180$00 e por outro lado as natas variam entre 125$00, 140$00 ou 165$00. Existe igual probabilidade de escolher qualquer um destes ingredientes. Considerando desprezável o preço

dos outros produtos que entram na confecção do prato, qual a probabilidade do preço da ementa ser superior a 1700$00?

Resolução:

Vamos identificar cada uma das possibilidades pelo respectivo preço. Temos de considerar todas as combinações possíveis e uma maneira simples de o fazer é considerando um diagrama em árvore:




Do diagrama anterior verificamos que das 12 combinações possíveis só 4 é que têm um preço superior a 1700$00. Assim a probabilidade pedida é 4/12 ou seja 1/3.

**5**. Numa determinada Universidade, verificou-se que de entre os alunos do 1º ano:

| | | | | |
|---|---|---|---|---|
| 51 | % | estão inscritos em Análise | | |
| 62 | " | " | " | Álgebra |
| 40 | " | " | " | Probabilidades |
| 28 | " | " | " | simultaneamente em Análise e Álgebra |
| 21 | " | " | " | simultaneamente em Análise e Probabilidades |
| 24 | " | " | " | simultaneamente em Álgebra e Probabilidades |
| 10 | " | " | " | simultaneamente em Análise, Álgebra e Prob. |

a) Represente num diagrama de Venn, os acontecimentos anteriores.

b) Calcule a probabilidade de um aluno escolhido ao acaso:

1) Estar inscrito em Análise ou Álgebra

2) Estar inscrito só em Análise e Álgebra

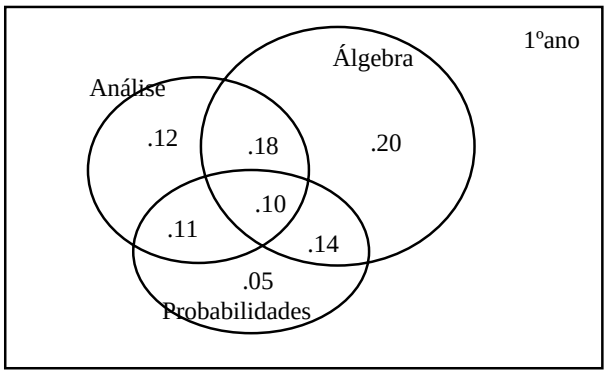
3) Estar inscrito em pelo menos uma das cadeiras

4) Estar inscrito só em Probabilidades

5) Não estar inscrito em nenhuma das cadeiras consideradas

Resolução:

**a**)




Para construir o diagrama anterior, começamos por preencher com a probabilidade .10, o espaço correspondente à intersecção dos 3 acontecimentos. Seguidamente a partir do conhecimento das probabilidades das intersecções dois a dois, preenchemos os espaços correspondentes às probabilidades .18, .14 e .11. Finalmente preencheram-se os espaços resultantes, a partir do conhecimento das probabilidades de cada um dos acontecimentos "estar inscrito em Análise", "estar inscrito em Álgebra" e "estar inscrito em Probabilidades".

**b**) **1**) P( Anál. **ou** Álg.) = P(Anál.) + P(Álg.) - P(Anál. **e** Álg.)
= .51 + .62 - .28 = .85

ou, a partir do diagrama de Venn

P( Anál. **ou** Álg.) = .12 + .11 + .18 + .10 + . 20 + .14 = .85

**2**) P(Anál. **e** Álg. **e** $\overline{\text{Prob}}$ ) = P(Anál. **e** Álg.) - P(Anál. **e** Álg. **e** Prob.)
= .28 - .10 =.18

ou, a partir do diagrama de Venn
P(Anál. **e** Álg. **e** $\overline{\text{Prob}}$ ) = .18

**3**) P(Anál. **ou** Álg. **ou** Prob.) = P(Anál.) + P(Álg.) + P(Prob.) - P(Anal. **e** Álg) - P(Anál. **e** Prob.) - P(Álg. **e** Prob.) + P(Anál. **e** Álg. **e** Prob.)
= .51 + .62 + .40 - .28 - .21 - -24 + .10
= .90

ou, a partir do diagrama de Venn

P(Anál. **ou** Álg. **ou** Prob.) = .12 + .18 + .11 + .10 + .20 + .14 + .05 = .90

**4**) P(Prob. **e** $\overline{\text{Anál.}}$ **e** $\overline{\text{Álg.}}$ )=P(Prob)-P(Prob. **e** Anál.)-P(Prob. **e** Álg.)+P(Prob. **e** Anál. **e** Álg.)
= .40 - .21 - .24 + .10 = .05

ou, a partir do diagrama de Venn
P(Prob.e $\overline{\text{Anál.}}$ **e** $\overline{\text{Álg.}}$ ) = .05

**5**) P( $\overline{\text{Anál. ou Álg. ou Prob.}}$ ) = 1 - P(Anál. **ou** Álg. **ou** Prob.)
= 1 - .90 = .10

**6** . De um lote de 20 rifas, em que 8 têm prémio e 12 não têm, retiraram-se 7. Qual a probabilidade de nas rifas retiradas, haver 3 premiadas e 4 não premiadas?

Resolução:

Vamos utilizar a definição clássica de probabilidade, para calcular a probabilidade pretendida.

Assim, do conjunto de 20 rifas de quantas maneiras possíveis se podem retirar 7 rifas? Será as combinações de 20, 7 a 7 ou seja $\binom{20}{7}$ Destas, nem todas são favoráveis, pois só o serão as que têm 3 premiadas e 4 não premiadas. O nº de possibilidades de tirar 3 premiadas é $\binom{8}{3}$ e não premiadas é $\binom{12}{4}$. Então, como cada uma das possibilidades das 3 premiadas se conjuga com todas as possibilidades das não premiadas, o nº de maneiras possíveis de 3 premiadas e 4 não premiadas é $\binom{8}{3}$x$\binom{12}{4}$ donde a probabilidade pretendida é

$$\binom{8}{3}\text{x}\binom{12}{4}/\binom{20}{7}$$

**7** . Suponha que uma andorinha entrou inadvertidamente numa sala com 4 janelas, em que uma estava aberta e as outras fechadas. A andorinha não se apercebia de qual a janela aberta, de forma que ao tentar sair da sala dirigia-se aleatoriamente para qualquer uma das 4 janelas. Por outro lado, como era uma andorinha esperta, se ao fazer uma tentativa não acertasse com a janela certa, já não tornava a essa janela na tentativa seguinte. Qual a probabilidade de conseguir sair:

a) À primeira tentativa?; b) À segunda tentativa?;c) À terceira tentativa?; d) À quarta tentativa?

Resolução:

a) P(1ª tentativa) = P(escolher a janela aberta) = 1/4

b) P(2ª tentativa) = P(escolher uma fechada à 1ª tentativa e a aberta à 2ª tentativa) = P(escolher uma fechada à 1ª tentativa) P(escolher a aberta à 2ª tentativa| escolheu uma fechada à 1ª tentativa) = $\frac{3}{4}\times\frac{1}{3}$ = 1/4

c) P(3ª tentativa) = P(escolher uma fechada à 1ª tentativa e uma fechada à 2ª tentativa e a aberta à 3ª tentativa) = P(escolher uma fechada à 1ª tentativa) P(escolher uma fechada à 2ª tentativa| escolheu uma fechada à 1ª tentativa) P(escolher a aberta à 3ª tentativa| escolheu uma fechada à 1ª tentativa e uma fechada à 2ª tentativa) = $\frac{3}{4}\times\frac{2}{3}\times\frac{1}{2}=\frac{1}{4}$

d) P(4ª tentativa) = … = $\frac{3}{4}\times\frac{2}{3}\times\frac{1}{2}\times 1$ ( por um raciocínio análogo ao da alínea anterior)

**8** . O José está indeciso em ir passar o fim de semana fora e telefonou para o serviço meteorológico para saber qual a previsão do tempo. Disseram-lhe que havia 20% de possibilidades de chover. Se chover o José tem uma probabilidade de .25 de ir para o Algarve. Se não chover esta probabilidade aumenta para .85.

a) Qual a probabilidade do José ir para o Algarve?

b) O José foi passar o fim de semana ao Algarve. Qual a probabilidade de ter chovido?

Resolução:

a) P(ir Algarve) = P(chover **e** ir Algarve) + P(não chover **e** ir Algarve)

= P(chover) P(ir Algarve / choveu) + P(não chover) P(ir Algarve / não choveu)

$= .20 \times .25 + .80 \times .85$

= .73

b) P(ter chovido / foi Algarve) = $\dfrac{\text{P(chover e ir Algarve)}}{\text{P(ir Algarve)}}$

$= \dfrac{.05}{.73}$

= .07

**Exercícios propostos**

**1.** Numa cervejaria trabalham 3 empregados: o António, o Bernardo e o Constantino. O António serve 40% dos clientes e os outros dois dividem entre si a restante clientela. Ao pedir uma cerveja, o acompanhamento desta por tremoços é deixada ao critério do empregado. O António é sócio da cervejaria, pelo que apenas traz tremoços em 10% das vezes. O Bernardo oferece tremoços em 40% dos casos, enquanto que o Constantino apenas oferece os tremoços a 20% dos clientes.

a) Ao pedir uma cerveja, calcule a probabilidade de que esta venha acompanhada de tremoços.

b) Se ao chegar à mesa de um amigo verificar que ele está a beber cerveja, acompanhada de tremoços, calcule a probabilidade de ele ter sido servido pelo Constantino.

**2**. a) Numa determinada cidade, existem em média 2 daltónicos em cada 1000 indivíduos. Num cruzamento de ruas dessa cidade, o trânsito é regulado por semáforos. Quando algum condutor passa o sinal vermelho, ou é por ser daltónico ou por ser atrevido. Supondo que a probabilidade de um condutor passar o sinal vermelho se for daltónico é .5 e se não for daltónico é .1, determine a probabilidade de um indivíduo ser daltónico, se passou o sinal vermelho.

b) calcule a probabilidade de em 100 indivíduos que passaram o sinal vermelho, não haver nenhum daltónico.

**3**. Estão a decorrer as filmagens do novo filme de um famoso cineasta português, célebre pelo ritmo estonteante que impõe às cenas. Numa cena particular intervêm dois actores, Arnesto e Bicente, tendo cada um de dizer apenas uma frase.

O primeiro a falar é Arnesto que se engana com probabilidade .3. Se Arnesto falha, Bicente que fala de seguida, falha também com probabilidade .9. Porém, se Arnesto acertar na sua frase, Bicente pode falhar a sua com probabilidade .05. A cena é repetida até que ambos acertem as

suas frases. Admita que as diversas filmagens são independentes e que as probabilidades indicadas se mantêm para cada filmagem.

a) Calcule a probabilidade de

i) Arnesto ter falhado a sua frase, sabendo que Bicente falhou a sua.

ii) Arnesto ter acertado a sua frase, sabendo que Bicente acertou a sua.

b) Calcule a probabilidade da cena ter de ser filmada 1, 2, 3, … vezes.

**Capítulo 6**

# Variáveis Aleatórias

## 6.1 - Introdução

Já definimos experiências aleatórias, espaços de resultados e acontecimentos. Também já vimos alguns processos de atribuir probabilidades a acontecimentos. Assim, a uma experiência aleatória, pode-se associar um **modelo de probabilidade**, que pressupõe a construção de um espaço de resultados e a atribuição de uma probabilidade a cada um dos resultados (acontecimentos elementares).

Os resultados de uma experiência aleatória, seriam analisados mais facilmente quando associados a números, mas nem todos os resultados de uma experiência são resultados numéricos! Basta pensar na experiência aleatória, que consiste no lançamento de uma moeda ao ar e verificar qual a face que fica voltada para cima.

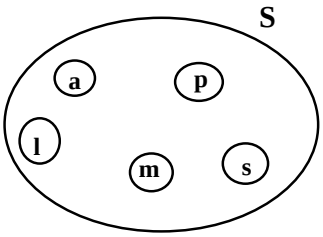
Veremos, no entanto, um processo de associar valores numéricos aos resultados de uma experiência aleatória, entrando com o conceito de variável aleatória, introduzido a seguir.

## 6.2 - Variável aleatória

Consideremos uma experiência aleatória, com o espaço de resultados S, associado.

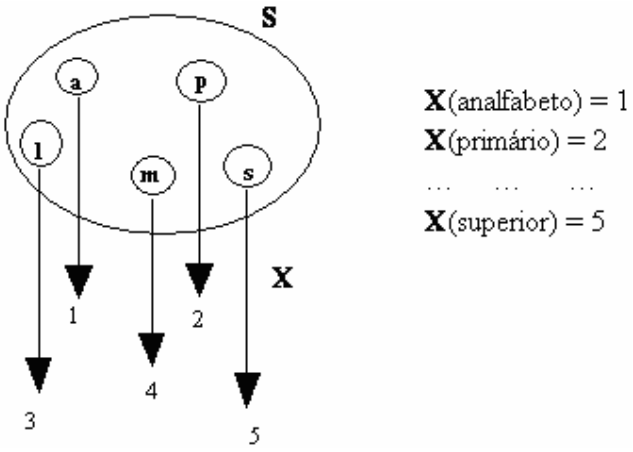
Uma **variável aleatória** (v.a.) **X** é uma função que associa a cada ponto do espaço de resultados S, um número.

**Exemplo 1** - Consideremos a experiência aleatória que consiste em perguntar a uma pessoa, ao acaso, quais as suas habilitações literárias. As respostas possíveis são: analfabeto, instrução primária, liceal, curso médio ou curso superior, que representamos no seguinte esquema




a - analfabeto
p - instrução primária
l - liceal
m - médio
s - superior

Podemos associar aos acontecimentos anteriores valores numéricos da seguinte forma:




A função **X** é uma **variável aleatória**, que assume os valores 1, 2, …, 5.

Seguidamente apresentamos outros exemplos de variáveis aleatórias.

**Exemplo 2** - Considere a experiência aleatória que consiste em lançar um dado ao ar e observar a face que fica voltada para cima. Associada a esta experiência podemos definir a variável aleatória. X, que a cada face associa o número de pintas; então os valores que X pode assumir são

X - 1, 2, …, 6

**Exemplo 3** - Considere a experiência aleatória que consiste em lançar ao ar uma moeda 50 vezes. Associada a esta experiência, podemos definir a variável aleatória Y, que representa o número de vezes que saiu cara, nos 50 lançamentos; então os valores que Y pode assumir são

Y - 0, 1, … , 50

**Exemplo 4** - Considere a experiência aleatória que consiste em lançar uma moeda ao ar até sair cara. Associada a esta experiência, podemos definir a variável aleatória Z, que representa o número de lançamentos necessários para sair cara; então os valores que a variável aleatória Z pode assumir são

Z - 1, 2, 3, …

**Exemplo 5** - Considere a experiência aleatória que consiste em observar a chuva que cai, num dia ao acaso. Associada a esta experiência, podemos definir a variável aleatória U, que representa a quantidade de chuva (em mm); então U pode assumir qualquer valor real, não negativo.

**Exemplo 6** – Considere a variável aleatória que consiste em observar o resultado de um desafio de futebol. Aos 3 resultados possíveis – perde a equipa visitante, empatam ou ganha a equipa visitante, associamos os valores –1, 0 e 1 através da variável aleatória V da seguinte forma:

V(perde a equipa visitante) = -1
V(empate) = 0
V(ganha a equipa visitante) = 1

pelo que a variável aleatória V assume os valores –1, 0 ou 1.

*Observação*: Ao definir variável aleatória, deve-se ter o cuidado de não a confundir com o valor observado, que ela pode assumir. Assim, no exemplo 2, considerado anteriormente, a variável aleatória X, antes de se lançar o dado, pode assumir qualquer valor do conjunto {1, 2,… , 6}. Depois de se ter realizado a experiência, se se obteve o valor 4, por exemplo, diz-se que 4 é um **valor observado** da variável aleatória. Geralmente representa-se um valor observado de uma variável aleatória pela mesma letra com que se representa a variável, mas minúscula. Então se Y for uma variável aleatória, representamos por y um valor observado dessa variável aleatória.

**População e variável aleatória?**

No início do nosso curso dissemos que o objectivo da Estatística é o estudo de **Populações**. Mas então qual a **entidade que representa a População**? A partir deste momento já estamos aptos a responder a esta questão, pois o que acontece é que identificamos População com a variável aleatória associada. Vamos tentar explicitar um pouco melhor esta associação, com o seguinte exemplo:

Suponhamos que estávamos interessados em estudar a **População** constituída pelas alturas dos Portugueses. Podemos considerar a **experiência aleatória** que consiste em perguntar a um português, escolhido ao acaso, qual a sua altura. Os resultados desta experiência constituem a *População* que se pretende estudar, que é o conjunto de todas as alturas possíveis (ver secção 1, capítulo 2). Então, associada a esta experiência podemos considerar a *variável aleatória* **X**, que representa a altura de um português, escolhido ao acaso. Esta variável aleatória pode assumir qualquer valor positivo.

Suponhamos ainda, que estávamos interessados em estudar a População constituída pelo número de chamadas telefónicas, que chegam a uma central, num determinado intervalo de tempo. Então podemos representar essa População pela variável aleatória. Y, que dá o número de chamadas telefónicas nesse intervalo de tempo e que assume os valores 0, 1, 2, ….

### 6.2.1 - Variável aleatória discreta

Uma variável aleatória diz-se **discreta**, se só assume valores de um conjunto, para o qual se possa estabelecer uma correspondência biunívoca com um subconjunto dos números inteiros, isto é, só assume um *número finito* ou *infinito numerável* de valores distintos.

As variáveis aleatórias consideradas nos exemplos 2, 3, 4 e 6 são exemplos de variável aleatória discretas.

Terá sentido falar na **probabilidade de uma** variável aleatória **assumir um determinado valor**? Vamos ver que sim!

Efectivamente, já que aos acontecimentos atribuímos probabilidades, é natural definir probabilidade de uma variável aleatória assumir um determinado valor, como sendo a probabilidade do acontecimento, que fez com que a variável aleatória tivesse esse valor! Por exemplo, quando consideramos a experiência aleatória do lançamento da moeda, os acontecimentos elementares (resultados) são "cara" e "coroa". Se a esta experiência associarmos a variável aleatória X, tal que

X(cara) = 0 e X(coroa) = 1

então dizemos que

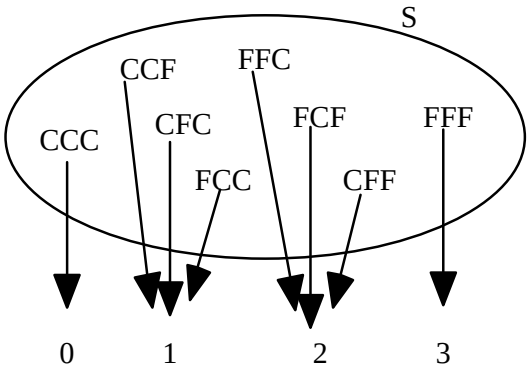
**P( X = 0) = 1/2** porque P("cara") = 1/2

e

**P( X = 1) = 1/2** porque P("coroa") = 1/2

Observação: Através da Probabilidade introduzida para os acontecimentos, estamos a *induzir* uma Probabilidade para a variável aleatória.

**Exemplo 7 –** Considere a experiência aleatória que consiste em verificar o número de caras que saem no lançamento de 3 moedas. Associada a esta experiência consideremos a variável aleatória X que assume os valores 0, 1, 2 ou 3, conforme for 0, 1, 2 ou 3 o número de caras obtidas no lançamento das 3 moedas. Qual a probabilidade de a variável aleatória assumir aqueles valores? Representando por F – cara e C – coroa, o espaço de resultados S é constituído pelos seguintes resultados




P(X=0) = P(CCC) = $\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{8}$ (no cálculo desta probabilidade entrámos com o facto de as moedas serem equilibradas e os lançamentos serem independentes uns dos outros)

P(X=1) = P$(CCF \cup CFC \cup FCC)$ = P(CCF)+P(CFC)+P(FCC) (Porquê?)

$$= \frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}+\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}+\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{3}{8}$$

P(X=2) = P(FFC$\cup$FCF$\cup$CFF) = P(FFC)+P(FCF)+P(CFF)

$$= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{3}{8}$$

$$P(X=3) = P(FFF) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$$

**Função massa de probabilidade**

Atendendo a que a variável aleatória discreta associa números aos resultados de uma experiência, em vez de falarmos nas probabilidades dos acontecimentos elementares (resultados), podemos falar nas probabilidades dos valores que a variável aleatória assume.

À função que dá a probabilidade associada a cada valor numérico, que a variável aleatória assume, chamamos **função massa de probabilidade**.

Uma variável aleatória. **X** fica perfeitamente identificada pela sua f.m.p., isto é, pelos valores $\mathbf{x_i}$ que assume e pelas probabilidades de assumir esses valores

$$\mathbf{p_i = P(X = x_i)}$$

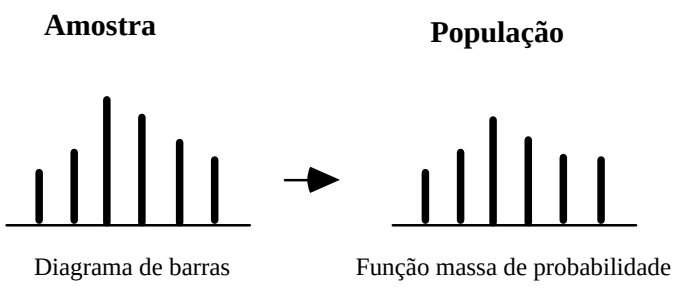
Atendendo à definição de probabilidade é imediato que:

a) $p_i \geq 0$

b) $\sum_i p_i = 1$

**Função massa de Probabilidade**
**versus**
**diagrama de barras**

Recordando a definição de diagrama de barras, construído com as frequências relativas, e tendo em consideração a teoria frequencista da Probabilidade, podemos concluir que o *diagrama de barras* é a imagem estatística da *função massa de probabilidade*! Efectivamente, se a amostra que se recolhe para estudar uma População ou variável aleatória, tem dimensão suficientemente grande, podemos interpretar as frequências relativas dos valores observados na amostra, como as probabilidades dos valores que a variável aleatória pode assumir.



Vem a propósito recordar o que dissemos, no início do curso sobre o que é fazer inferência estatística: consiste em, a partir das propriedades verificadas na amostra, tentar transportar essas propriedades para a população. Então, se ao estudar uma determinada amostra obtivermos um diagrama de barras com um determinado aspecto, esperamos que a função massa de probabilidade da **População** – representada pela variável aleatória associada, de onde foi extraída a amostra, tenha um aspecto semelhante. Vejamos o seguinte exemplo:

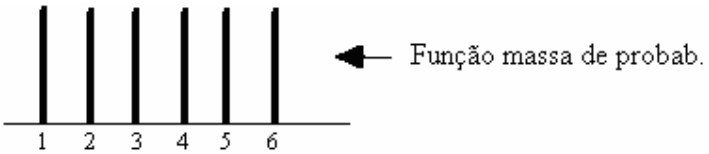
**Exemplo 8** - Consideremos a experiência aleatória que consiste em lançar um dado e verificar a face que fica voltada para cima. Associada a esta experiência, pensemos na variável aleatória que representa o número de pintas dessa face. Precisamos de arranjar um modelo de probabilidade para esta variável aleatória!

Suponhamos então que lançamos o dado 1000 vezes e registamos o número de vezes que se observou cada face, tendo-se verificado os seguintes resultados:

| **Face** | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $f_i$ | .163 | .160 | .167 | .168 | .162 | .170 |


Tendo em atenção os resultados anteriores, em que os valores para as frequências relativas são muito , será natural considerar como modelo de probabilidade para **X**, o seguinte:

| **X** | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $p_i$ | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |




Uma vez feita esta hipótese, de que o dado é equilibrado, existem métodos estatísticos (testes), que nos permitem <u>quantificar o erro</u> que se comete ao admiti-la. Não esqueçamos que estamos a admitir que a população goza de uma determinada propriedade (todos os valores da variável aleatória são igualmente prováveis), a qual foi sugerida por uma propriedade verificada na amostra. Devido à aleatoriedade presente na amostra, existe a possibilidade de estarmos a cometer um erro, ao transportar a propriedade para a população.

Vamos apresentar seguidamente alguns exemplos onde se consideram variáveis aleatórias discretas.

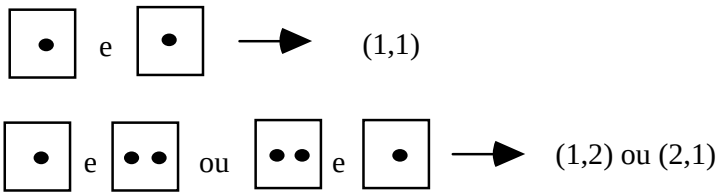
**Exemplo 9 -** Considere a variável aleatória X, que representa a soma das pintas das faces que ficam voltadas para cima, quando se lançam dois dados. Defina completamente essa variável.

Resolução:

Seja X - v.a. que representa a soma das pintas de dois dados. Podemos representar os valores possíveis para X, assim como as respectivas probabilidades, na seguinte tabela:

| **X** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **$p_i$** | 1/36 | 2/36 | 3/36 | 4/36 | 5/36 | 6/36 | 5/36 | 4/36 | 3/36 | 2/36 | 1/36 |

O processo de obter cada uma das probabilidades anteriores foi o seguinte: representando as faces que ficam voltadas para cima, nos dois dados, pelo par ordenado (i,j), com i,j=1,…6, temos, por exemplo,




P(X=2) = P(1,1) = P(1) × P(1) porque os acontecimentos "saída de 1" num dado e "saída de 1" no outro dado são independentes.

$$= \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

P(X=3) = P((1,2) ou (2,1)) = P(1,2) + P(2,1) Porque os acontecimentos (1,2) e (2,1) são disjuntos.

$$= \frac{1}{36} + \frac{1}{36} = \frac{2}{36}$$

P(X=4) = P((1,3) ou (2,2) ou (3,1)) = $\frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{3}{36}$

Analogamente se calculam as outras probabilidades.

**Exemplo 10** – Considere a v.a. Z que representa a soma das faces que ficam voltadas para cima, quando lança 3 dados. Calcule a probabilidade de Z ser maior que 13.

Este problema é idêntico ao anterior, só que agora temos mais casos possíveis. Assim, no lançamento dos 3 dados temos $6^3$ = 216 possibilidades, todas igualmente possíveis, das quais só nos interessam aquelas cuja soma seja superior a 13. Vamos ver então qual a probabilidade da variável aleatória Z assumir os valores 14, 15, 16, 17 ou 18:

O cálculo destas probabilidades reduz-se à contagem do número de possibilidades de obter cada um daqueles valores, como se apresenta a seguir.

| Resultado | $Z=z_i$ | $P(Z= z_i)$ |
|---|---|---|
| (2,6,6), (6,2,6), (6,6,2), (3,5,6), (3,6,5), (5,3,6), (5,6,3), (6,3,5), (6,5,3), (4,4,6), (4,6,4), (6,4,4), (4,5,5), (5,4,5), (5,5,4) | 14 | 15/216 |
| (3,6,6), (6,3,6), (6,6,3), (4,5,6), (4,6,5), (5,4,6), (5,6,4), (6,4,5), (6,5,4), (5,5,5) | 15 | 10/216 |
| (4,6,6), (6,4,6), (6,6,4), (5, 5,6), (5,6,5), (6,5,5) | 16 | 6/216 |
| (5,6,6), (6,5,6), (6,6,5) | 17 | 3/216 |
| (6,6,6) | 18 | 1 |

Assim, a probabilidade de Z>13 será igual a 35/216 = 0,162.

## Utilização do Excel na simulação da experiência do lançamento de três dados

Vamos utilizar o Excel para simular a experiência que consiste em lançar três dados e estimar a probabilidade da soma das pintas das faces que ficam voltadas para cima, ser superior a 13.

Nesta simulação vamos utilizar a função *LOOKUP((lookup value; lookup vector; result vector)*. Esta função pesquisa no vector *lookup vector* o maior valor que não seja superior a *lookup value* e de seguida devolve o valor que está na posição correspondente em *result vector*. Assim, na folha de Excel começámos por considerar estes dois vectores e depois utilizámos a função RAND para simular o lançamento dos dados. Apresentamos de seguida parte da tabela onse se visualiza a simulação correspondente ao 1º dado. Para os outros dados é idêntico:

Simulação

| | A | B | C | D |
|---|---|---|---|---|
| 4 | X | P(X=i) | Prob.acum | |
| 5 | 1 | 0,166666667 | =B5 | 0 |
| 6 | 2 | 0,166666667 | =B6+C5 | 0,166666667 |
| 7 | 3 | 0,166666667 | =B7+C6 | 0,333333334 |
| 8 | 4 | 0,166666667 | =B8+C7 | 0,500000001 |
| 9 | 5 | 0,166666667 | =B9+C8 | 0,666666668 |
| 10 | 6 | 0,166666667 | =B10+C9 | 0,833333335 |
| 11 | | =SUM(B5:B10) | | 1,000000002 |
| 12 | | | | |
| 13 | | 1º dado | | 2ºdado |
| 14 | =RAND() | =LOOKUP(A14;$D$5:$D$11;$A$5:$A$10) | =RAND() | =LOOKUP(C14;$D$5:$D |
| 15 | =RAND() | =LOOKUP(A15;$D$5:$D$11;$A$5:$A$10) | =RAND() | =LOOKUP(C15;$D$5:$D |
| 16 | =RAND() | =LOOKUP(A16;$D$5:$D$11;$A$5:$A$10) | =RAND() | =LOOKUP(C16;$D$5:$D |

Sheet1 Sheet2 Sheet3

Depois de simularmos os três dados, procedemos aos seguintes cálculos:

- Na coluna G calculamos a soma das faces;
- N coluna H testamos se temos sucesso;
- Na coluna I inserimos o número da experiência;
- Na coluna J calculamos a frequência absoluta de sucesso;
- Na coluna K calculamos a frequência relativa de sucesso.

Simulação

| | G | H | I | J | K |
|---|---|---|---|---|---|
| 13 | Soma das faces | Sucesso? | Nºexp. | Freq.abs.sucesso | Freq.rel.sucesso |
| 14 | =B14+D14+F14 | =IF(G14>13;1;0) | 1 | =H14 | =J14/I14 |
| 15 | =B15+D15+F15 | =IF(G15>13;1;0) | 2 | =J14+H15 | =J15/I15 |
| 16 | =B16+D16+F16 | =IF(G16>13;1;0) | 3 | =J15+H16 | =J16/I16 |
| 17 | =B17+D17+F17 | =IF(G17>13;1;0) | 4 | =J16+H17 | =J17/I17 |
| 18 | =B18+D18+F18 | =IF(G18>13;1;0) | 5 | =J17+H18 | =J18/I18 |
| 19 | =B19+D19+F19 | =IF(G19>13;1;0) | 6 | =J18+H19 | =J19/I19 |
| 20 | =B20+D20+F20 | =IF(G20>13;1;0) | 7 | =J19+H20 | =J20/I20 |
| 21 | =B21+D21+F21 | =IF(G21>13;1;0) | 8 | =J20+H21 | =J21/I21 |
| 22 | =B22+D22+F22 | =IF(G22>13;1;0) | 9 | =J21+H22 | =J22/I22 |
| 23 | =B23+D23+F23 | =IF(G23>13;1;0) | 10 | =J22+H23 | =J23/I23 |

Sheet1 Sheet2 Sheet3

Ao fim de 1000 simulações obtivémos o seguinte resultado:

Simulação

| | G | H | I | J | K |
|---|---|---|---|---|---|
| 13 | Soma das faces | Sucesso? | Nºexp. | Freq.abs.sucesso | Freq.rel.sucesso |
| 14 | 10 | 0 | 1 | 0 | 0 |
| 15 | 6 | 0 | 2 | 0 | 0 |
| 16 | 10 | 0 | 3 | 0 | 0 |
| 17 | 6 | 0 | 4 | 0 | 0 |
| 18 | 13 | 0 | 5 | 0 | 0 |
| 19 | 5 | 0 | 6 | 0 | 0 |
| 20 | 10 | 0 | 7 | 0 | 0 |
| 1010 | 12 | 0 | 997 | 162 | 0,162487 |
| 1011 | 13 | 0 | 998 | 162 | 0,162325 |
| 1012 | 7 | 0 | 999 | 162 | 0,162162 |
| 1013 | 7 | 0 | 1000 | 162 | 0,162 |

Sheet1 Sheet

Pelo que consideramos a frequência relativa de 0,162, como uma estimativa para a probabilidade pretendida.

Nota – Sempre que recalcular a folha de Excel, obtém um valor diferente para a estimativa da probabilidade, já que, como dissemos várias vezes, a função RAND é volátil.

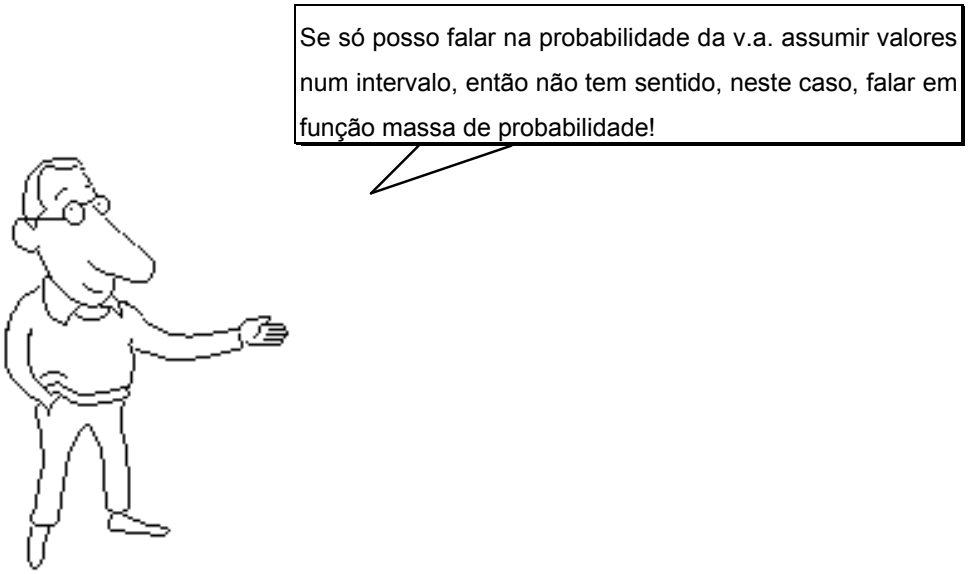
### 6.2.2 - Variável aleatória contínua

As variáveis aleatórias que possam assumir todos os valores de um intervalo, sendo nula a probabilidade de assumirem valores isolados, dizem-se variáveis aleatórias **contínuas**. Enquanto que uma variável aleatória discreta se refere a qualquer tipo de contagem, uma v. a. contínua refere-se a uma medida, como por exemplo o peso, a altura, o tempo, etc.

Exemplos:

- tempo que um cliente espera numa "bicha" dum supermercado
- peso de um bebé de 6 meses

- tempo entre chegadas telefónicas consecutivas




De acordo com a definição de variável aleatória contínua, esta não assume valores em pontos isolados, com probabilidade diferente de zero, ao contrário do que se passa com as variáveis aleatórias discretas. Assim, não tem sentido falar na probabilidade de uma variável aleatória X, contínua, assumir determinado valor x, uma vez que esta probabilidade é sempre nula.

**Então não podemos definir função massa de probabilidade de uma variável aleatória contínua!.**

Existe, no entanto, uma função - a **função densidade de probabilidade** (f.d.p.), que definiremos mais à frente, e que vai assumir, para as variáveis aleatórias contínuas, o papel da função massa de probabilidade no caso das variáveis aleatórias discretas.

## 6.3 - Função distribuição

Outro processo (além da f.m.p. para as v.a. discretas e da f.d.p. para as v.a. contínuas) de exprimir as probabilidades associadas à variável aleatória **X**, é utilizando a **Função distribuição** $F_X(x)$, ou simplesmente F(x), função que para cada valor $\mathbf{x} \in \mathbf{R}$, acumula as probabilidades de todos os valores menores ou iguais a x. Assim

**Função distribuição** de uma variável aleatória X (discreta ou contínua), é a função **F(x)** tal que

$$\forall\ \mathbf{x} \in \mathbf{R},\ \mathbf{F(x) = P(X \leq x)}$$

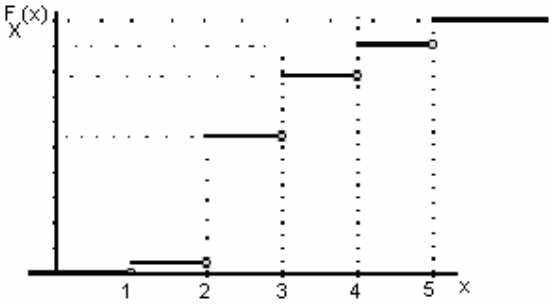
Para poder calcular $P(X \leq x)$, significa que a $X \leq x$ deve estar associado um acontecimento!

Na realidade, dado um ponto qualquer x, X≤x refere-se ao acontecimento constituído pelos resultados de S, tais que os valores associados pela v.a. X são menores ou iguais a x. Vamos ver, seguidamente, um exemplo de uma função de distribuição, para uma variável aleatória discreta.

**Exemplo 11** - Construa a função distribuição da v.a. definida pela seguinte função massa de probabilidade

| X | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $p_i$ | .04 | .50 | .24 | .12 | .10 |

FX(x) = 0 para x<1
0.04 para 1≤x<2
0.54 para 2≤x<3
0.78 para 3≤x<4
0.90 para 4≤x<5
1 para x≥5

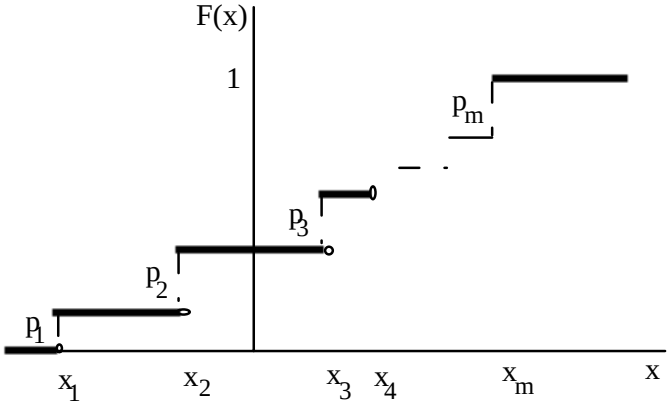



Como se viu no exemplo anterior, a função de distribuição de uma v.a. X discreta, é uma função em escada, com saltos nos pontos $x_i$ onde a v.a. assume valores com probabilidade diferente de zero. Os saltos têm amplitude $p_i$, onde $p_i = P(X = x_i)$

De uma forma genérica, consideremos a v.a. **X**, discreta, com função massa de probabilidade

| **X** | $x_1$ | $x_2$ | $x_3$ | … | $x_m$ |
|---|---|---|---|---|---|
| $P(X=x_i)$ | $p_1$ | $p_2$ | $p_3$ | | $p_m$ |

A função distribuição da v.a. anterior, tem o seguinte aspecto:

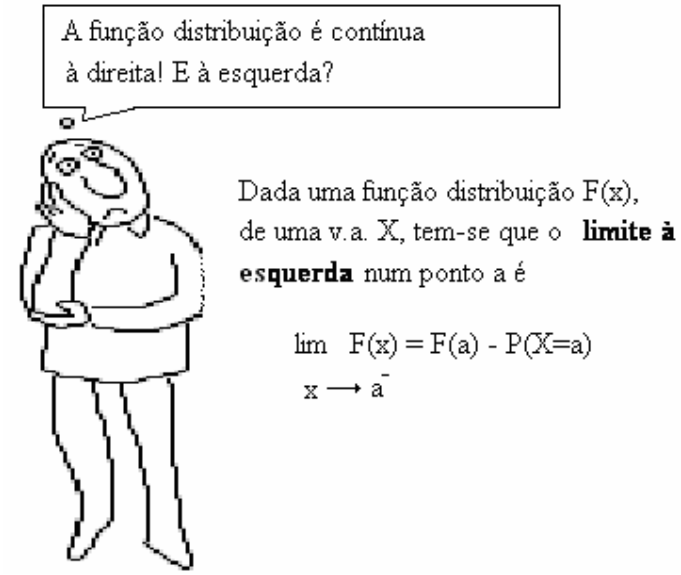



**Exercício:** Dada a função distribuição de uma v.a. discreta, verifique como é que pode obter os valores da v.a. associada.

**Propriedades da função distribuição de uma variável aleatória X ( discreta ou contínua)**

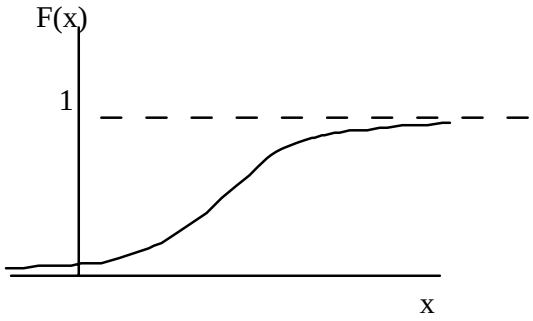
1. $F(-\infty) = 0$ (limite de F(x) quando $x \to -\infty$) porque $P(X \leq -\infty) = 0$

   $F(+\infty) = 1$ (limite de F(x) quando $x \to +\infty$) porque $P(X \leq +\infty) = 1$
2. **F(x)** é uma função **não decrescente**
3. **F(x)** é **contínua à direita** (decorre da forma como foi definida)




Então:

- Se a v.a. **X** é **discreta**, a função distribuição é **descontínua** (só é contínua à direita) - é <u>uma função em escada</u>, com saltos nos pontos onde a v.a. assume valores com probabilidade diferente de zero.
- Se a v.a. **X** é **contínua**, a função distribuição é **contínua**, porque, qualquer que seja o ponto **a**, tem-se P(X=a)=0, pelo que a <u>função também é contínua à esquerda</u>. Uma função distribuição contínua tem o seguinte aspecto

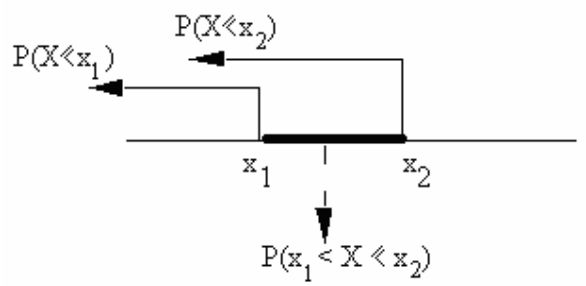



**Qual a utilidade da função distribuição?**

Será que o conhecimento da função distribuição, permite o cálculo da probabilidade de uma v.a. assumir valores num determinado intervalo?

Consideremos uma v.a. com função distribuição F(x). Então, dados dois pontos quaisquer $x_1$ e $x_2$, tem-se

$$P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$$




O conhecimento da função distribuição, permite o cálculo da probabilidade da v.a. X assumir valores num intervalo da forma $]x_1 , x_2]$

E se o intervalo não for dessa forma, isto é, aberto à esquerda e fechado à direita? Então temos de distinguir os casos em que a v.a. é discreta e contínua:

**Variável aleatória X**

| **Discreta** | **Contínua** |
|---|---|
| $P(x_1<X<x_2) = F(x_2) - F(x_1) - P(X=x_2)$ | $P(x_1<X<x_2) = F(x_2) - F(x_1)$ |
| $P(x_1\leq X<x_2) = F(x_2) - F(x_1) - P(X=x_2) + P(X=x_1)$ | $P(x_1\leq X<x_2) = F(x_2) - F(x_1)$ |
| $P(x_1\leq X\leq x_2) = F(x_2) - F(x_1) + P(X=x_1)$ | $P(x_1\leq X\leq x_2) = F(x_2) - F(x_1)$ |

Se a variável aleatória for contínua, não há diferença para o cálculo da probabilidade de um intervalo, se esse intervalo é aberto ou fechado, já que a probabilidade de um ponto é igual a 0.

**Função distribuição**
**versus**
**função distribuição empírica**

Ao fazer o estudo descritivo dos dados de uma amostra, uma das representações utilizadas foi a da **Função distribuição empírica**, que como vimos, dá para cada x, a *proporção* de elementos da amostra menores ou iguais a x. É uma <u>função não decrescente, contínua à direita, que assume valores entre 0 e 1</u>.

Repare-se na analogia entre esta função e a **Função distribuição**, a que também chamamos *Função distribuição populacional*, por dizer respeito à população, ou seja à variável aleatória **X**. Efectivamente a função distribuição dá para cada x, a *probabilidade* dos valores da variável aleatória serem menores ou iguais a x.

Assim, mais uma vez invocando a teoria frequencista da Probabilidade, podemos dizer que a Função distribuição empírica é uma imagem estatística da Função distribuição (populacional), já que, se a amostra com que se construiu a f.d.e. for suficientemente grande, interpretamos as proporções como probabilidades.

## 6.4 - Função densidade de probabilidade (para v. aleatórias contínuas)

Do mesmo modo que a função distribuição empírica é a imagem estatística da função distribuição, podemos dizer que o **histograma** - representação utilizada para dados de tipo contínuo, é a imagem estatística de uma função definida para variáveis aleatórias contínuas, a que damos o nome de **função densidade de probabilidade**. De uma forma mais correcta:

Define-se **função densidade de probabilidade** da v.a. X contínua, e representa-se por **f(x)**, como sendo a derivada, se existir, da função distribuição **F(x)**:

$$\mathbf{f(x) = F'(x)}$$

**Atenção**: A função densidade só está definida para v.a. contínuas! Para as v.a. discretas, temos uma função que desempenha papel análogo, a função massa de probabilidade! (Em alguma bibliografia, à função massa de probabilidade também chamam função densidade)

A partir da definição de função densidade, e das propriedades da função distribuição, facilmente se demonstra que:

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

(Esta é a notação utilizada para a primitiva de f(x), que se anula em -□). Então, do mesmo modo que a partir da função massa de probabilidade das v.a. discretas, se pode obter a função distribuição, também no caso das v.a. contínuas, o conhecimento da função densidade, permite a obtenção da função distribuição. Temos assim:

$$F(x) \xrightarrow{\text{Derivação}} f(x)$$

$$f(x) \xrightarrow{\text{Primitivação}} F(x)$$

A operação de primitivação, ao contrário da derivação, é definida a menos de uma constante, pelo que utilizamos a notação anteriormente considerada, para representar a primitiva especial, que se anula em -∞, pois F(-∞) = 0.

**Propriedades da função densidade**

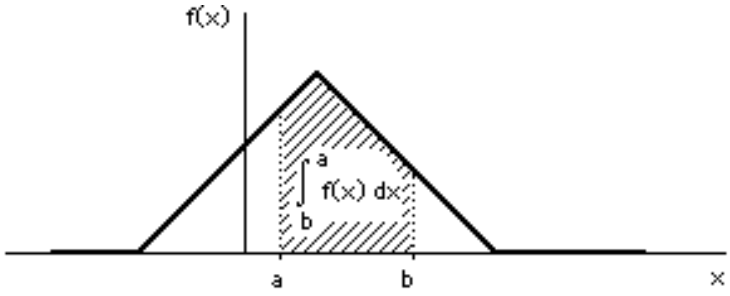
1 . **f(x) ≥ 0** (porque é a derivada de uma função não decrescente

2. $\int_{-\infty}^{+\infty} \mathbf{f(x)\,dx = 1}$ (porque F(+∞) = 1)

3. $\int_{a}^{b} \mathbf{f(x)\,dx = F(b) - F(a)}$ ( propriedade do integral)

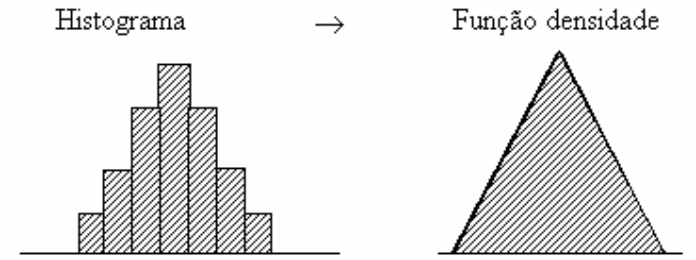
**Propriedade:**

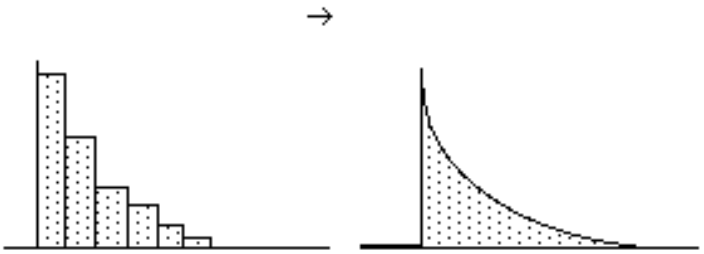
Também da definição de integral, pode-se mostrar que o cálculo de 3. se resume a calcular a **área** compreendida entre o eixo dos xx, o gráfico da função densidade f(x) e as rectas x=a e x=b, como se ilustra na figura seguinte:



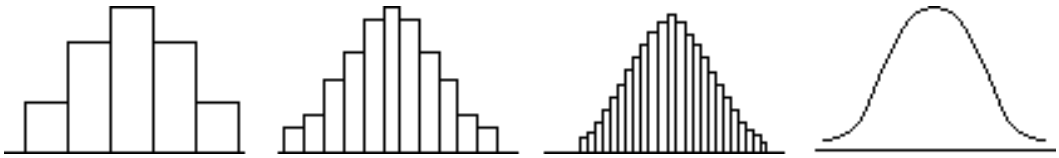

**Função densidade**
**versus**
**histograma**

Ao construirmos o histograma, chamámos a atenção para que os rectângulos, que o compõem, deviam ter áreas iguais às frequências relativas das respectivas classes. Assim, a área total ocupada pelo histograma é igual a 1. Ora o histograma é a imagem estatística da <u>função densidade, que é uma função tal que a área total compreendida entre o seu gráfico e o eixo dos xx, é igual a 1</u> (veja-se a propriedade enunciada anteriormente sobre áreas e a propriedade 2. das funções densidades). Assim a imagem do *histograma* sugere a da *função densidade*, da população subjacente à amostra, com a qual se construiu o histograma:





Se os histogramas anteriores foram obtidos para determinadas dimensões das amostras, ao aumentar substancialmente as dimensões dessas amostras, iríamos considerar um maior nº de classes, o que conduziria a que a amplitude de classe fosse diminuindo. Quanto menor for a amplitude das classes, melhor será a imagem que o histograma dá da função densidade, que pretende ilustrar:



**Exemplo 12 -** Consideremos a v.a. X, que representa o tempo que uma pessoa leva para ir de carro de Lisboa a Coimbra. Admitamos que esse tempo se distribui uniformemente no intervalo [2h , 2h 20m].

a) Qual a probabilidade de que a viagem dure entre 2h e 2h10m?

b) Qual a probabilidade de que a viagem dure entre 2h5m e 2h10m?

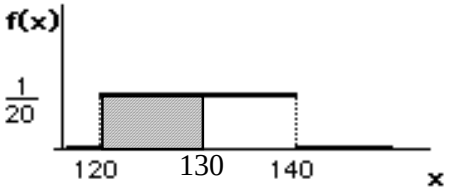
Resolução: A v.a. X é contínua, pois pode assumir qualquer valor do intervalo considerado. Além disso tem uma distribuição uniforme ( este modelo será estudado mais à frente, no capítulo 8 - Algumas distribuições importantes) , pelo que a função densidade é constante nesse intervalo e tem o seguinte aspecto:


Observe-se que a área compreendida entre o gráfico da função e o eixo dos xx, é efectivamente igual a 1. Como a função é não negativa, estão satisfeitas as propriedades da função densidade.

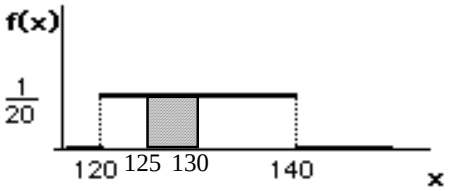
A expressão da função densidade é

$$f(x) = \begin{cases} \frac{1}{20} & 120 \le x \le 140 \\ 0 & \text{caso contrário} \end{cases}$$

a) Para calcular a P(120≤X≤130), basta calcular a área da parte a tracejado, pelo que a probabilidade pretendida é 1/2.




b) Analogamente o cálculo da P(125≤X≤130) se reduz ao cálculo da área a tracejado




obtendo-se para a probabilidade 1/4.

**Exercícios**

**1**. Apresentam-se a seguir exemplos de experiências aleatórias e variáveis aleatórias associadas. Para cada um dos casos identifique quais os valores que a v.a. pode assumir e diga se é discreta ou contínua.

| | |
|---|---|
| a) Realizar um exame de 20 questões | Nº de questões respondidas correctamente |
| b) Observar os carros que chegam a uma portagem durante uma hora | Nº de carros que chegam à portagem |
| c) Observar a chuva que cai num dia de inverno | Quantidade de água, medida em mm, num certo observatório meteorológico |
| d) Lançar um dado até sair a face 6 | Nº de lançam. necess. para sair a face 6 |

**2** . Quais das seguintes funções , são funções massa de probabilidade

| $X=x_i$ | $p_i$ | $Y=y_i$ | $p_i$ | $Z=z_i$ | $p_i$ |
|---|---|---|---|---|---|
| 0 | .20 | -2 | .25 | -1 | .20 |
| 1 | .30 | 2 | .05 | 0 | .50 |
| 2 | .25 | 4 | .10 | 1 | -.10 |
| 3 | .35 | 6 | .60 | 2 | .40 |

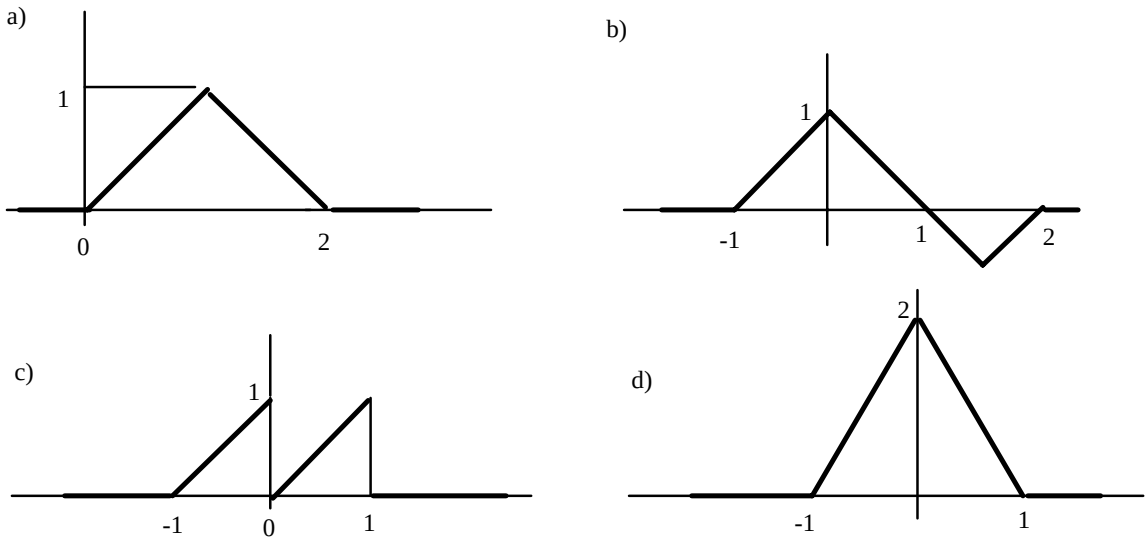
**3**. Diga se as seguintes funções podem ser funções de distribuição (f.d.) de alguma variável aleatória. Se alguma for f.d. diga qual a v.a. associada.

| | |
|---|---|
| **F(x)** = 0 | x< -2 |
| 1/4 | -2 ≤ x < 2 |
| 1/5 | 2 ≤ x < 3 |
| 1/2 | 3 ≤ x < 5 |
| 1 | 5 ≤ x |

| | |
|---|---|
| **F(x)** = 0 | x ≤ 2 |
| 1/4 | 2 < x ≤ 3 |
| 1/3 | 3 < x ≤ 4 |
| 1/2 | 4 < x ≤ 5 |
| 1 | 5 < x |

| | |
|---|---|
| **F(x)** = 0 | x < -2 |
| 1/4 | -2 ≤ x < 2 |
| 1/3 | 2 ≤ x < 3 |
| 1/2 | 3 ≤ x < 4 |
| 1 | 4 ≤ x |

**4**. Diga quais dos seguintes gráficos podem ser representações de funções densidade de v.a. contínuas:




**5** . Relativamente à função densidade da alínea c) do exercício anterior, calcule, para a v.a. X associada

$$P(X\leq 0);\ P(X<0);\ P(X>0)$$

**6** . A probabilidade de em cada instante, se conseguir fazer o "login" num determinado computador, a partir de um terminal remoto, é .70. Seja X a v.a. que representa o nº de tentativas necessárias, até se obter a ligação.

a) Determine os 4 primeiros termos para a f.m.p. de X

b) Determine uma expressão genérica para a f.m.p. de X

c) Determine P(X=6)

d) Determine a expressão de F(x)

e) A partir de F, determine a probabilidade de que sejam necessárias, no máximo, 4 tentativas para conseguir ligar o computador

f) A partir de F, determine a probabilidade de que sejam necessárias, no mínimo, 5 tentativas para conseguir ligar o computador.

**7** . A quantidade de bacalhau (expressa em Kg) vendida diariamente no supermercado do Sr. Manuel é uma variável aleatória X com a seguinte função densidade de probabilidade:

$$f(x)=\begin{cases} kx & 0 \leq x < 5 \\ k(10-x) & 5 \leq x < 10 \\ 0 & x < 0 \text{ ou } x \geq 10 \end{cases}$$

a) Determine k de forma a f(x) poder ser considerada função densidade da v.a. X.

b) Calcule a função distribuição de X.

c) Calcule a mediana da variável aleatória X ( Resolva esta alínea depois de ter estudado o capítulo seguinte).

d) O Sr. Manuel vende o bacalhau ao preço de 1500$00/Kg. Sabendo que ao fim da manhã tinham sido vendidos 4.5 Kg, calcule a probabilidade de até ao fim do dia o Sr. Manuel fazer no máximo 12 000$00.

**8** . Considere a seguinte função:

$$f(x)=\begin{cases} k(x+2)/2 & -2 \le x < 0 \\ k & 0 \le x < 2 \\ k(3-x) & 2 \le x < 3 \\ 0 & x < -2 \text{ ou } x \ge 3 \end{cases}$$

a) Determine k de forma a f(x) poder ser considerada função densidade de uma v.a. X.

b) Determine a função distribuição de X.

c) Determine a mediana de X ( Resolva esta alínea depois de ter estudado o capítulo seguinte).

## 6.5 - Pares de variáveis aleatórias

### 6.5.1 - Introdução

Apesar das distribuições de probabilidade estudadas até aqui envolverem uma única variável, pode acontecer que tenhamos de analisar duas ou mais variáveis em conjunto. Nestas situações, a distribuição de probabilidade resultante é referida como distribuição de probabilidade conjunta.

### 6.5.2 - Distribuição de probabilidade conjunta

Vamos considerar unicamente o caso de termos um **par** de variáveis aleatórias **discretas**, que representaremos por (X,Y). Admitindo que X assume os valores $x_i$ e Y os valores $y_j$, definimos **função massa de probabilidade conjunta**, que representamos por $p_{ij}$, como sendo

$$p_{ij}= P(X = x_i, Y = y_j)$$

As probabilidades anteriores costumam-se representar numa tabela com o seguinte aspecto:

| X \ Y | $y_1$ | $y_2$ | … | $y_j$ | … | $y_k$ | |
|---|---|---|---|---|---|---|---|
| $x_1$ | $p_{11}$ | $p_{12}$ | … | $p_{1j}$ | … | $p_{1k}$ | |
| $x_2$ | $p_{21}$ | $p_{22}$ | … | $P_{2j}$ | … | py | |
| … | | | … | | … | | |
| $x_i$ | $p_{i1}$ | $P_{i2}$ | … | $p_{ij}$ | … | $p_{ik}$ | |
| … | | | … | | … | | |
| $x_m$ | $p_{m1}$ | $P_{m2}$ | … | $p_{mj}$ | … | $p_{mk}$ | |
| | | | … | | … | | 1 |

Suponhamos que pretendíamos a **$P(X = x_1)$**. Como calculá-la a partir da tabela anterior?

Repare-se que o acontecimento traduzido por $X=x_1$, é equivalente a

$(X = x_1) \equiv [(X = x_1 \text{ e } Y = y_1) \text{ ou } (X = x_1 \text{ e } Y = y_2) \text{ ou } \ldots (X = x_1 \text{ e } Y = y_j) \text{ ou } \ldots \text{ ou } (X = x_1 \text{ e } Y = y_k)]$

$$\text{Então, } P(X = x_1) = p_{11} + p_{12} + \ldots + p_{1j} + \ldots + p_{1k}$$

Representando esta probabilidade por $p_{1.}$, e utilizando a mesma metodologia para calcular a probabilidade de X assumir outro valor qualquer ou Y assumir qualquer dos seus valores, temos a tabela com as margens preenchidas:

| **X** \ **Y** | $y_1$ | $y_2$ | … | $y_j$ | … | $y_k$ | |
|---|---|---|---|---|---|---|---|
| $x_1$ | $p_{11}$ | $p_{12}$ | … | $p_{1j}$ | … | $p_{1k}$ | $p_{1.}$ |
| $x_2$ | $p_{21}$ | $p_{22}$ | … | $P_{2j}$ | … | py | $p_{2.}$ |
| … | | | … | | … | | |
| $x_i$ | $p_{i1}$ | $P_{i2}$ | … | $p_{ij}$ | … | $p_{ik}$ | $p_{i.}$ |
| … | | | … | | … | | |
| $x_m$ | $p_{m1}$ | $P_{m2}$ | … | $p_{mj}$ | … | $p_{mk}$ | $p_{m.}$ |
| | $p_{.1}$ | $p_{.2}$ | … | $p_{.j}$ | … | $p_{.k}$ | 1 |

Às funções $\mathbf{p_{i.}} = \sum_j \mathbf{p_{ij}}$, i= 1, 2, …, m

e $\mathbf{p_{.j}} = \sum_i \mathbf{p_{ij}}$, j= 1, 2, …, k

chamamos **funções massa de probabilidade marginais** de X e Y, respectivamente.

**Exemplo 13 -** Suponha que se escolhem 3 pilhas, de um conjunto constituído por 3 pilhas novas, 4 usadas, mas a trabalhar e 5 estragadas. Representando por

X - v.a. que dá o nº de pilhas novas no lote das 3 pilhas retiradas

Y - " " usadas " "

determine a função massa de probabilidade conjunta de (X,Y) e as f.m.p. marginais de X e Y.

Resolução: Os valores que a v.a. X pode assumir são: 0, 1, 2 ou 3, o mesmo acontecendo com a v.a. Y.

$$P(X= 0, Y= 0) = \frac{\binom{5}{3}}{\binom{12}{3}} = \frac{10}{220} \qquad P(X= 0, Y= 1) = \frac{\binom{4}{1}\binom{5}{2}}{\binom{12}{3}} = \frac{40}{220}, \ \ldots$$

$$P(X = 3, Y= 0) = \frac{\binom{3}{3}}{\binom{12}{3}} = \frac{1}{220}$$

De um modo geral, temos a seguinte expressão para calcular a f.m.p. conjunta

$$P(X = i, Y = j) = \frac{\binom{3}{i}\binom{4}{j}\binom{5}{3-i-j}}{\binom{12}{3}} \qquad \text{com } i, j = 0, 1, 2, 3 \text{ e } i+j \leq 3$$

Depois de calculadas, as probabilidades apresentam-se no quadro seguinte, onde se incluem as f.m.p. marginais de X e Y:

| X \ Y | 0 | 1 | 2 | 3 | $p_{i.}$ |
|---|---|---|---|---|---|
| 0 | 10/220 | 40/220 | 30/220 | 4/220 | 84/220 |
| 1 | 30/220 | 60/220 | 18/220 | 0 | 108/220 |
| 2 | 15/220 | 12/220 | 0 | 0 | 27/220 |
| 3 | 1/220 | 0 | 0 | 0 | 1/220 |
| $p_{.j}$ | 56/220 | 112/220 | 48/220 | 4/220 | 1 |

### 6.5.3 - Variáveis aleatórias independentes

Limitando-nos ainda às variáveis aleatórias discretas, dizemos que as variáveis aleatórias **X** e **Y** são **independentes**, se e só se , para todo o par $(x_i,y_j)$ em que (X,Y) está definido, se tem

$$\mathbf{P(X = x_i, Y = y_j) = P( X = x_i) P( Y = y_j)}$$

ou seja, utilizando a notação introduzida anteriormente,

$$\forall\ (i,j) \quad \mathbf{p_{ij} = p_{i.}\, p_{.j}}$$

Relativamente ao exemplo anterior, imediatamente se verifica que as variáveis X e Y não são independentes. Efectivamente, basta existir um zero no interior da tabela para que não se possa verificar a independência.

**Exercícios**

**1** . Sejam X e Y duas v.a. tais que:

X - assume os valores 0 ou 1, conforme seja a máquina A ou B, que produz determinado artigo

Y - assume os valores 0, 1, 2 ou 3 e representa o nº de defeitos de um artigo produzido pelas máquinas A ou B

A seguinte tabela, apresenta a distribuição de probabilidade conjunta das v.a. X e Y:

| X \ Y | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 0 | .1250 | .0625 | .1875 | .1250 |
| 1 | .0625 | .0625 | .1250 | .2500 |

a) Verifica-se que um artigo não tem defeitos. Qual a probabilidade de ter sido produzido pela máquina A?

b) Sabe-se que um artigo foi produzido pela máquina A. Qual a probabilidade de não ter defeitos?

c) Sabe-se que um artigo tem dois ou mais defeitos. Qual a probabilidade de ter sido produzido pela máquina A?

d) O nº de defeitos de um artigo, é influenciado pela máquina que o produz?

**2 .** Considere a seguinte tabela que representa a função massa de probabilidade conjunta do par aleatório (X,Y):

| Y \ X | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 1 | 0.1 | 0.15 | 0.2 | p |
| 2 | P | 0.15 | 0.15 | 0.05 |
| 3 | 0.05 | 0 | P | 0 |

a) Encontre o valor de p e obtenha as funções massa de probabilidade marginais de X e Y.

b) Verifique se X e Y são variáveis aleatórias independentes.

c) Defina a variável aleatória Z = X+Y e calcule o seu valor médio e variância.

d) Calcule $P(X+Y\leq 3|Y \text{ é ímpar})$

**3**. Considere dois acontecimentos A e B tais que P(A)=1/4, P(B|A)=1/2 e P(A|B)=1/4. Considere as variáveis aleatórias definidas do seguinte modo:

X=1 se A ocorre X=0 se A não ocorre

Y=1 se B ocorre Y=0 se B não ocorre

a) Determine a função massa de probabilidade conjunta do par (X,Y).

b) Determine as funções massa de probabilidade marginais de X e de Y.

c) Verifique se X e Y são variáveis aleatórias independentes.

d) Defina a variável aleatória Z=X+Y

**4**. Determine, designando por:

$X_1$ um número escolhido ao acaso do conjunto $\{i\varepsilon N: 1\leq i\leq 4\}$

$X_2$ um segundo número, escolhido ao acaso do conjunto $\{i\varepsilon N: 1\leq i\leq X_1\}$

a) i ) $P(X_2)=1$

ii) $P(X_1=2|X_2=1)$

b) Diga, justificando, se $X_1$ e $X_2$ são independentes.

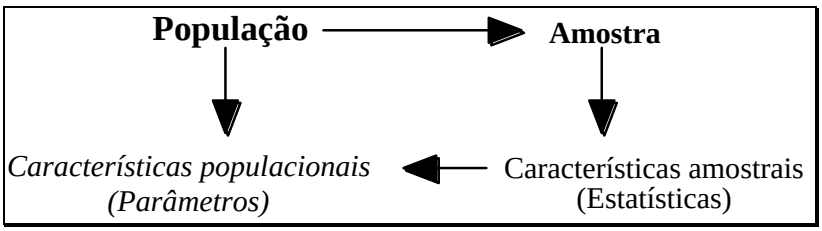
**Capítulo 7**

# Características populacionais

## 7.1 - Introdução

Quando pretendemos estudar uma população, que representamos pela variável aleatória **X**, já vimos que o processo que, de uma maneira geral, se segue, é recolher uma amostra da referida população e calcular as suas características amostrais, além das representações gráficas adequadas.

O objectivo do estudo da amostra é tentar "**inferir**" para a população, de onde a amostra foi recolhida, algumas propriedades. Assim, veremos que existem para a população **X**, medidas análogas às definidas para as amostras.




**Exemplo 1** - Consideremos a população, ou v.a. que representa o nº de pintas que se obtém no lançamento de um dado. Para estudar esta população, que pode assumir os valores 1, 2, 3, 4, 5 ou 6, fomos recolher uma amostra de dimensão 20, constituída pelo nº de pintas em 20 lançamentos. Suponhamos que os resultados obtidos foram os seguintes:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 4 | 2 | 1 | 5 | 2 | 3 | 6 | 2 | 1 |
| 5 | 6 | 4 | 5 | 5 | 3 | 4 | 2 | 3 | 3 |

Para calcular a média, podemos começar por agrupar os dados, pelo que obtemos:

$$\bar{x} = 1\times\frac{3}{20} + 2\times\frac{4}{20} + 3\times\frac{4}{20} + 4\times\frac{3}{20} + 5\times\frac{4}{20} + 6\times\frac{2}{20}$$

$$= 1\times .15 + 2\times .20 + 3\times .20 + 4\times .15 + 5\times .20 + 6\times .10$$

$$= 3.35$$

Para o cálculo da média utilizámos a fórmula

$$\bar{x} = \sum_i x_i f_i$$

pois somámos os produtos dos diferentes valores que surgem na amostra pelas frequências relativas, respectivas. Mas se o nº de provas fosse suficientemente grande, as frequências relativas utilizadas anteriormente para calcular a média, poderiam ser interpretadas como as probabilidades de uma v.a. assumir os valores de 1 a 6 (teoria frequencista da probabilidade).

Então, utilizando uma expressão análoga, vamos multiplicar os valores que a v.a. assume, pelas respectivas probabilidades, tendo em conta o modelo utilizado para modelar a população em estudo:

$$1\times\frac{1}{6}+2\times\frac{1}{6}+3\times\frac{1}{6}+4\times\frac{1}{6}+5\times\frac{1}{6}+6\times\frac{1}{6}=3.5$$

Mas agora não deveríamos continuar a chamar média a este valor, já que utilizámos as **probabilidades** e não as frequências. Neste momento deixámos de ter uma característica amostral, para termos uma *característica populacional*, equivalente à *característica amostral* média. Vamos ver na secção seguinte que a esta característica populacional, chamamos *valor médio*.

## 7.2 - Valor médio

Consideremos uma população representada pala v.a. **X**, discreta, que assume os valores $x_1$, $x_2$, $x_3$, … , com probabilidades $p_1$, $p_2$, $p_3$, …. Então define-se **valor médio** e representa-se por **E(X)** ou $\boldsymbol{\mu_X}$, como sendo a característica que se obtém a partir da seguinte expressão:

$$\boxed{E(X) = \sum_i x_i p_i}$$

Chamamos a atenção para que o valor médio é um **parâmetro**, isto é, uma quantidade numérica fixa, embora por vezes seja desconhecida, que descreve uma característica populacional. É um **parâmetro de localização**, que pretende localizar o centro da distribuição de probabilidades, do mesmo modo que a média é uma medida de localização do centro da amostra.

Ao contrário do *valor médio* que é um número *fixo*, a *média* é uma *variável aleatória* - efectivamente, conforme a amostra recolhida (para uma determinada dimensão), assim obtemos um valor diferente para a média.

Podemos ainda acrescentar o seguinte: a média é uma **estatística**, pois é uma variável aleatória que só depende dos valores da amostra e não depende de parâmetros desconhecidos. Assim, utilizando a terminologia já referida para as variáveis aleatórias, representaremos por $\overline{X}$ a v.a. média e por $\overline{x}$ um valor observado da variável aleatória média.

Então, voltando ao exemplo 1, dizemos que 3.35 é um valor observado da v.a. $\overline{X}$. Se a amostra recolhida tivesse sido a seguinte

1 4 1 1 3 2 3 5 2 1

5 6 4 5 5 3 6 2 3 3

obteríamos um outro valor observado para a v.a. média: o valor 3.25.

**Qual a importância da média, para o estudo da população?**

Quando recolhemos uma amostra, o nosso objectivo, como já temos referido várias vezes, é retirar conclusões para a população subjacente à amostra. É precisamente a média, que nos fornece informação sobre o valor médio! Assim, ao recolhermos a amostra anterior e ao obter o valor 3.25 para a *média*, dizemos que este valor é uma **estimativa** do *valor médio* da v.a. **X**, caracterizada por assumir os valores 1, 2, …,6, com probabilidades 1/6. Mas a amostra inicialmente recolhida tinha dado o valor 3.35 como estimativa para o valor médio! Aliás se continuássemos a recolher amostras diferentes, embora com a mesma dimensão, continuaríamos a obter valores ligeiramente diferentes para as respectivas médias, que seriam outras tantas estimativas para o valor médio. Quer dizer que a v.a. $\overline{X}$ é uma função que fornece *estimativas* para o valor médio - diz-se que é um **estimador** do valor médio.

$\overline{X}$ **é um estimador do**
$\bar{x}$ **é uma estimativa do** → **valor médio**

*Será um bom estimador? Isto é, as estimativas serão boas? Darão valores aproximados do parâmetro que pretende estimar?*

Se é um bom estimador ou não depende das suas propriedades, nomeadamente da variabilidade apresentada. Voltaremos a este assunto numa secção posterior dedicada à média, mas acrescentamos desde já, que efectivamente a média é, de um modo geral, um bom estimador para o valor médio, que é aliás traduzido pela seguinte versão da chamada Lei dos grandes números:

**Lei dos grandes números**

Se uma experiência aleatória se repetir muitas e muitas vezes, a média dos resultados obtidos aproxima-se do valor médio da variável aleatória associada.

Observação: Sobre este ponto gostaríamos ainda de observar que podemos à partida dispor de um bom estimador e obter más estimativas, quando as amostras que serviram para obter essas estimativas não forem representativas da população (relembrar o que foi dito no Capítulo 1 sobre o problema da amostragem).

**Exemplo 2** (adaptado de Moore, 1997) – Uma companhia de seguros instituiu um seguro de vida com a duração de 5 anos, para indivíduos de 21 anos, do sexo masculino, segundo a seguinte modalidade: a companhia paga uma indemnização de 20 mil contos se o segurado morrer nos

próximos 5 anos, sendo o prémio anual de 50 contos. Pretende-se saber qual o lucro esperado para a companhia de seguros, tendo em conta as seguintes probabilidades:

| Idade morte | 21 | 22 | 23 | 24 | 25 | ≥26 |
|---|---|---|---|---|---|---|
| Probabilidade | .0018 | .0019 | .0019 | .0019 | .0019 | .9906 |

Resolução:

Seja X a v.a. que representa o lucro auferido pela companhia de seguros ao longo dos anos em que o seguro é válido:

| Idade morte | 21 | 22 | 23 | 24 | 25 | ≥26 |
|---|---|---|---|---|---|---|
| X | -19950 | -19900 | -19850 | -19800 | -19750 | 250 |
| Probabilidade | .0018 | .0019 | .0019 | .0019 | .0019 | .9906 |

Então tendo em conta a expressão para o cálculo do valor médio temos que o lucro esperado é de aproximadamente 61 contos.

Tendo em conta o resultado anterior estaria disposto a assumir perante um amigo a responsabilidade que a companhia de seguros assume perante os seus segurados?

**Valor médio de uma função da v.a. X**

Dada a v.a. **X**, discreta, e a v.a. **Y**, função de X por intermédio da função g, isto é, **Y = g(X)**, tem-se

$$\mathbf{E(Y) = E[g(X)] = \sum_i g(x_i)\, p_i}$$

**Exemplo 3 –** Na produção de determinado tipo de vidro é necessário que a temperatura a que se aquece o forno atinja uma temperatura C rondando os 550° centígrados. No entanto verificam-se algumas flutuações em torno desta temperatura de acordo com a seguinte distribuição de probabilidades

| Temperatura C | 540° | 545° | 550° | 555° | 560° |
|---|---|---|---|---|---|
| Probabilidade | .10 | .15 | .50 | .20 | .05 |

- Calcule o valor médio de C
- Calcule o valor médio das flutuações verificadas
- Calcule o valor médio da temperatura medida em graus Fahrenheit

Resolução:

a) $E(C) = 540 \times .10 + 545 \times .15 + 550 \times .50 + 555 \times .20 + 560 \times .05$

$= 549.75°$

b) Consideremos uma nova variável aleatória Y = 550° - C, cuja f.m.p. é

| Y = 550 - C | 10° | 5° | 0° | -5° | -10° |
|---|---|---|---|---|---|
| Probabilidade | .10 | .15 | .50 | .20 | .05 |

donde $E(Y) = 10 \times .10 + 5 \times .15 + 0 \times .50 + (-5) \times .20 + (-10) \times .05$

$= .25°$

c) A temperatura medida em graus – F - obtém-se da temperatura medida em graus centígrados – C - a partir da seguinte expressão

$$F = \frac{9}{5}C + 32$$

Pelo que a f.m.p. da variável aleatória F é a seguinte:

| F | 1004° | 1013° | 1022° | 1031° | 1040° |
|---|---|---|---|---|---|
| Probabilidade | .10 | .15 | .50 | .20 | .05 |

Donde $E(F) = 1004 \times .10 + 1013 \times .15 + 1022 \times .50 + 1031 \times .20 + 1040 \times .05$

$= 1021.55°$

**Observação**: A definição de valor médio para populações contínuas, é uma generalização da definição de valor médio para populações discretas, em que agora utilizamos o integral em vez do somatório: assim, dada a v.a. **X**, contínua, definida em R, com função densidade f(x), tem-se

$$E(X) = \int_{-\infty}^{+\infty} x\, f(x)\, dx$$

**7.2.1 - Propriedades do valor médio**

1. Dadas duas v.a. **X** e **Y**, com valores médios respectivamente E(X) e E(Y), então

$$\mathbf{E(X \pm Y) = E(X) \pm E(Y)}$$

2. Dada a v.a. **X** e as constantes **a** e **b**, tem-se

$$\mathbf{E(aX + b) = a\,E(X) + b}$$

3. Se as v.a. X e Y são independentes, então

$$\mathbf{E(XY) = E(X)\,E(Y)}$$

*Atenção*: O valor médio do produto só é igual ao produto dos valores médios, se as v.a. forem independentes.

**Exemplo 3 (cont) –** As alíneas b) e c) deste exemplo poderiam ser imediatamente calculadas a partir da alínea a), tendo em conta a propriedade 2 do valor médio:

$$E(Y) = E(550 - C)$$
$$= 550 - E(C)$$
$$= .25°$$

$$E(F) = E(\frac{9}{5}C + 32)$$
$$= \frac{9}{5}E(C) + 32$$
$$= 1021.55°$$

**Exemplo 4** - O gerente de um restaurante verificou que o nº de pessoas, que compõem os grupos que pretendem mesa segue o seguinte modelo de probabilidade:

| nº pessoas/grupo | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| probabilidade | .10 | .30 | .10 | .20 | .08 | .11 | .03 | .08 |

Determine o tamanho médio dos grupos.

Resolução:

O que se pretende é o valor médio da v.a. X que representa o tamanho do grupo, donde

$E(X) = 1\times.10+2\times.30+3\times.10+4\times.20+5\times.08+6\times.11+7\times.03+8\times.08$

$= 3.71$

Se considerarmos a v.a. $Y=5X^2$, tem-se

$E(Y) = 5(1\times.10+4\times.30+9\times.10+16\times.20+25\times.08+36\times.11+49\times.03+64\times.08)$

$= 89.75$

## 7.3 - Quantil de probabilidade p

Continuando a estabelecer o paralelismo entre características amostrais e características populacionais, vamos definir uma outra *medida de localização*, além do valor médio, e que é o **quantil de probabilidade p** ou quantil de ordem p, onde $0 \le p \le 1$.

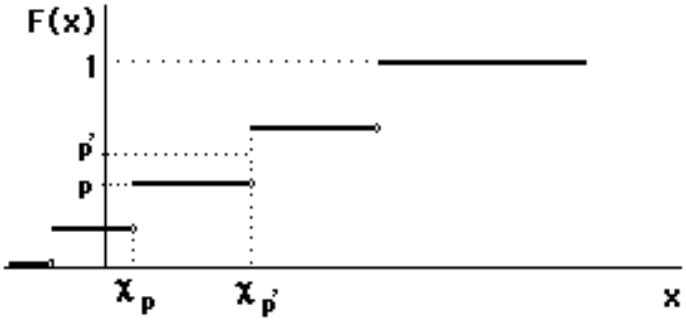
Assim, define-se **quantil de probabilidade p** da v.a. **X** e representa-se por $\chi_p$, como sendo o menor valor da v.a. **X** tal que

$$p \le F(\chi_p) \le p + P(X = \chi_p)$$

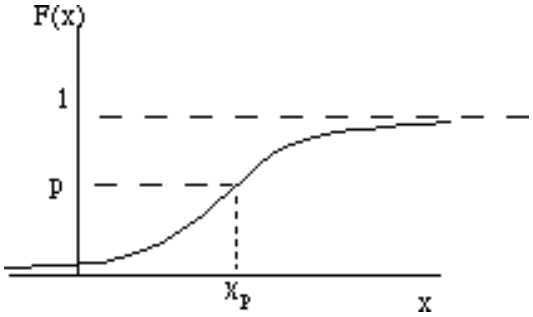
onde representamos por F(x) a função distribuição da v.a. X.

A notação agora utilizada para o quantil populacional, é diferente da utilizada para o quantil amostral, o qual se representou por $Q_p$.

Da definição de quantil, apresentada anteriormente, verifica-se que para as probabilidades p e p' os quantis são os valores representados na figura seguinte, representados respectivamente por $\chi_p$ e $\chi_{p'}$:



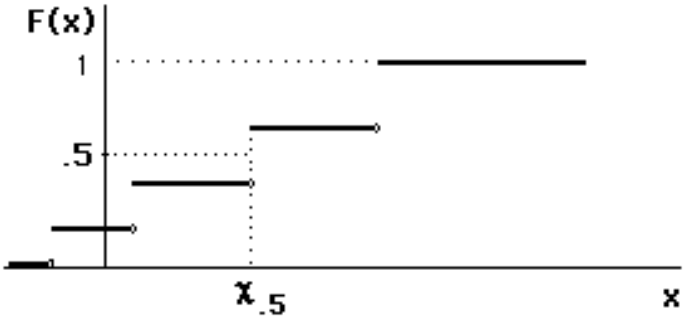
Observação: Quando a v.a. X é contínua, podemos dizer que o quantil de probabilidade p, $\chi_p$, é o valor tal que $F(\chi_p) = p$. Efectivamente, se a v.a. é contínua, a probabilidade de assumir valores em pontos isolados é igual a zero.




Como se vê pela figura, dado qualquer valor de p, no intervalo (0,1), o quantil fica univocamente determinado.

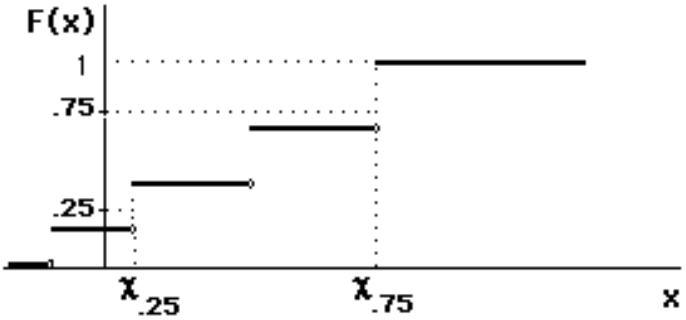
**Mediana**

A mediana é o quantil de probabilidade .5 e representa-se por $\chi_5$




**Quartis**

Quartis são os quantis de probabilidade .25 - 1º quartil e .75 - 3º quartil




**Exemplo 3** (cont) - Determine a mediana e os primeiro e terceiro quartis.

$P(X \leq 1) = .10$

$P(X \leq 2) = .40$ donde $\chi_{.25} = 2$

$P(X \leq 3) = .50$

$P(X \leq 4) = .70$

$P(X \leq 5) = .78$ donde $\chi_{.75} = 5$

## 7.4 - Variância (populacional)

Por oposição à variância amostral, podemos definir também um parâmetro populacional equivalente, a que chamamos variância e representamos por Var(X) ou $\sigma^2_X$. Define-se variância de X como sendo o valor médio do quadrado da diferença entre X e o seu valor médio

$$\mathrm{Var}(X) = E\{ [X - E(X)]^2\}$$

Limitando-nos ao caso de populações discretas, e utilizando a notação introduzida na definição do valor médio, define-se variância da v.a. X como sendo:

$$\mathbf{Var(X) = \sum_i [x_i - E(X)]^2 p_i}$$

Observação: Repare-se na analogia entre a definição da variância populacional e a variância amostral.

**Propriedades da variância**

1**.** Dada a v.a. **X** e as constantes **a** e **b**, tem-se

$$\mathbf{Var(aX+b)= a^2Var(X)}$$

2**.** Dadas as v.a. **X** e **Y independentes**, tem-se

$$\mathbf{Var(X \pm Y) = Var(X) + Var(Y)}$$

### 7.4.1 - Desvio padrão (populacional)

Do mesmo modo que fizemos para a amostra, também se define o **desvio padrão populacional**, ou unicamente desvio padrão, quando não houver dúvidas a qual nos estamos a referir, que se representa por $\sigma_X$, como sendo a raiz quadrada da variância

$$\sigma_X = \sqrt{,E\{ [X - E(X)]^2\}}$$

Observação: Enquanto que a medida de localização do centro da amostra se chama média e a do centro da população se chama valor médio, no caso da variância ou desvio padrão, não existem nomes diferentes, conforme estejamos na amostra ou na população. Assim, quando houver o perigo de confusão, falaremos em desvio padrão amostral ou empírico e em desvio padrão populacional.

*O desvio padrão populacional é uma medida da variabilidade da população, relativamente à medida de localização - valor médio*. Assim, quanto maior for o desvio padrão, maior será a dispersão apresentada pela variável aleatória.

O desvio padrão amostral, da mesma forma que a média, também é uma variável aleatória, que representamos por **S**. Quando se observa uma determinada amostra, então obtemos um valor observado para a v.a. S, que representamos por **s**.

Do mesmo modo que a média se utiliza como estimador do valor médio, também o **desvio padrão amostral** se costuma utilizar como **estimador** do parâmetro **desvio padrão populacional**

**S** é um estimador de → **desvio padrão populacional**

**s** é uma estimativa de → **desvio padrão populacional**

Assim, quando pretendemos estudar uma população **X**, recolhemos uma amostra dessa população, e calculamos a *média* e a *variância amostral*. Estas medidas dão-nos informação sobre os parâmetros populacionais *valor médio* e *variância populacional*, respectivamente.

**Outra expressão para o cálculo da variância:**

A partir da definição de variância, pode-se deduzir uma expressão mais simples para efeitos de cálculo, e que é a seguinte:

$$\mathbf{var(X) = E(X^2) - E^2(X)}$$

**Exemplo 5** - Suponha que lhe propõem o seguinte jogo:

- Receber mil contos sem ter que fazer nada ou
- Receber dois mil contos, se sair cara no lançamento de uma moeda ( se sair coroa não recebe nada).

Qual das situações prefere? Porquê?

Resolução:

As duas situações podem ser caracterizadas pelas v.a. X e Y, respectivamente em que

| **X** | 1000 |
|---|---|
| p | 1 |

| **Y** | 2000 | 0 |
|---|---|---|
| $p_i$ | 1/2 | 1/2 |

| | |
|---|---|
| $E(X) = 1000$ | $E(Y) = 1000$ |
| $Var(X) = 0$ | $Var(Y) = 1000^2$ |
| $\sigma_X = 0$ | $\sigma_Y = 1000$ |

As duas v.a. são caracterizadas por terem o mesmo valor médio, o que significa que, ao fim de várias jogadas, em média, o jogador ganharia o mesmo. No entanto o risco que corre ao aceitar a primeira situação é nulo, enquanto que o que corre ao aceitar a segunda situação é bastante grande. Assim, a primeira situação é preferível à segunda ( a não ser que o jogador goste de correr riscos!).

**Exercícios**

**1** - O João apostou com o seu amigo Pedro que no próximo jogo Benfica - Sporting, o Benfica ganharia. O João recebe 300$ se ganhar a aposta e paga 200$ de perder. Para quem é que é favorável a aposta:

a) Se a probabilidade do Benfica ganhar ao Sporting for de .5?

b) Se a probabilidade anterior for de .3?

Se os montantes implicados na aposta forem respectivamente 200$ e 100$, e tendo em conta a alínea a), o risco corrido pelo João é maior ou menor, do que com os montantes iniciais?

**2** - Um jornal de desporto publica anúncios nas suas páginas, verificando-se que cada página ou não contém anúncios, ou tem 1/3, 2/3 ou a página inteira preenchida com publicidade. O modelo de probabilidade que descreve a proporção da página ocupada com publicidade é dado pela seguinte tabela:

| Prop. da pág. | 0 | 1/3 | 2/3 | 1 |
|---|---|---|---|---|
| p | .408 | .017 | .025 | .550 |

a) Determine a proporção média de cada página ocupada por anúncios.

b) Determine o desvio padrão do modelo de probabilidade definido anteriormente

c) Determine a função de distribuição e represente-a graficamente. Obtenha a mediana.

**3** - A percentagem de peças defeituosas produzidas por uma máquina é de 10%. Se se escolherem aleatoriamente 2 peças dessa máquina, qual o nº médio de peças defeituosas? E a variância do nº de peças defeituosas?

**4** – Calcule o lucro médio esperado para um jogador que joga na raspadinha.

## 7.5 - Covariância

Dadas duas variáveis aleatórias X e Y, existe uma medida adequada para medir a maior ou menor intensidade, com que as v. a. se associam (linearmente) ou acompanham, que se chama covariância entre X e Y e se representa por Cov(X,Y):

$$\boxed{\mathbf{Cov(X,Y) = E\{[X - E(X)][Y - E(Y)]\}}}$$

Tendo em consideração as propriedades do valor médio, tem-se:

$$\mathbf{Cov(X,Y) = E(XY) - E(X)\,E(Y)}$$

**Propriedades da covariância**

**1** - Se as v.a. **X** e **Y** são **independentes**, então

$$\mathbf{Cov(X,Y) = 0}$$

Obs: A propriedade inversa não é necessariamente verdadeira. As variáveis podem ter covariância nula, sem que sejam independentes.

**2** - Dadas as v.a. **X** e **Y**, tem-se

$Var(X \pm Y) = Var(X) + Var(Y) \pm 2\,Cov(X,Y)$

Das propriedades anteriores deduz-se imediatamente que

Se as v.a. **X** e **Y** são **independentes**, então

$$Var(X \pm Y) = Var(X) + Var(Y)$$

**7.5.1 - Coeficiente de correlação**

A covariância depende das unidades com que se exprimem as variáveis aleatórias X e Y. Sendo assim, é conveniente introduzir uma nova medida, chamada **coeficiente de correlação** entre **X** e **Y**, que se representa por $\rho$, e se obtém dividindo a covariância pelo produto dos desvios padrões de X e Y

$$\rho = \frac{\mathbf{Cov(X,Y)}}{\sqrt{\mathbf{var(X)\,var(Y)}}}$$

Como é evidente, $\rho$ **é** a característica populacional correspondente à medida empírica ou amostral **r**.

Do mesmo modo que o coeficiente de correlação empírico r, também o coeficiente de correlação $\rho$ assume valores do intervalo **[-1,1]**.

Assim:

- um valor de $\rho$ <u>próximo de 1</u>, significa uma <u>forte associação linear</u>, <u>positiva</u>, entre as variáveis X e Y;
- um valor de $\rho$ <u>próximo de -1</u>, significa uma <u>forte associação linear</u>, <u>negativa</u>, entre as variáveis X e Y;

- um valor de $\rho$ próximo de 0, significa que essa associação linear não existe ou é muito pequena. Chamamos a atenção para o facto de que, neste caso, as v.a. X e Y, podem estar correlacionadas não linearmente.

No caso em que o coeficiente de correlação é igual a 1 ou a -1, temos que Y é uma função linear de X, respectivamente crescente ou decrescente, pelo que o conhecimento de uma das variáveis permite conhecer a outra das variáveis.

**Exemplo 6** - Dado o par de v.a. (X,Y) pela seguinte tabela

| X \ Y | 1 | 2 |
|---|---|---|
| 0 | 1/21 | 2/21 |
| 1 | 3/21 | 4/21 |
| 2 | 5/21 | 6/21 |

verifique se são correlacionadas linearmente.

Resolução: Começamos por calcular as f.m.p. marginais e a seguir os valores médios, desvios padrões e covariância.

| X \ Y | 1 | 2 | f.m.p.X |
|---|---|---|---|
| 0 | 1/21 | 2/21 | 3/21 |
| 1 | 3/21 | 4/21 | 7/21 |
| 2 | 5/21 | 6/21 | 11/21 |
| f.m.p.Y | 9/21 | 12/21 | 1 |

E(X) = 29/21 E(Y) = 33/21

$\sigma^2_X$ = .5215 $\sigma^2_Y$ = .2449

$$E(XY) = (0\times 1)\times\frac{1}{21}+(1\times 1)\times\frac{3}{21}+(2\times 1)\times\frac{5}{21}+\ldots+(2\times 2)\times\frac{6}{21}=\frac{45}{21}$$

$$Cov(X,Y) = \frac{45}{21}-\frac{29}{21}\times\frac{33}{21}=-.0272$$

$$\rho = \frac{-.0272}{\sqrt{.5215\times .2449}}=-.0761$$

$\Rightarrow$ Existe correlação linear de tipo inverso, mas muito fraca, entre X e Y.

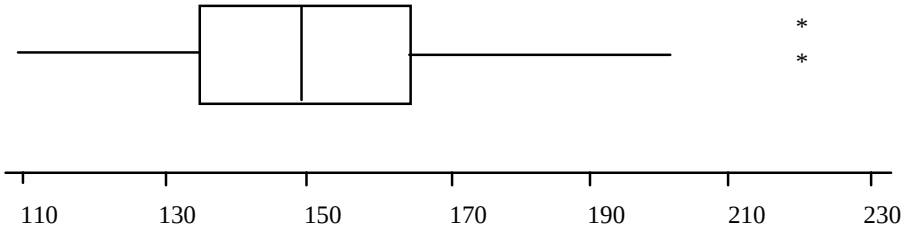
## 7.6 – Regressão de Y em X

Vimos no capítulo 4 que a *recta dos mínimos quadrados* nos dá uma descrição linear da relação (linear) existente entre uma variável explanatória X e uma variável resposta Y. Neste capítulo vimos que para cada característica amostral, existe a característica populacional correspondente. Será que poderemos continuar a estabelecer esse paralelismo, agora que estamos no domínio dos pares aleatórios? Será que existe algo de semelhante, para a População, à recta dos mínimos quadrados definida para as amostras? Efectivamente assim é, pois podemos dizer que a recta dos mínimos quadrados é a imagem estatística da regressão populacional de Y em X. No entanto para definirmos este modelo é necessário admitir determinadas hipóteses que neste momento saem for a do contexto destas folhas:

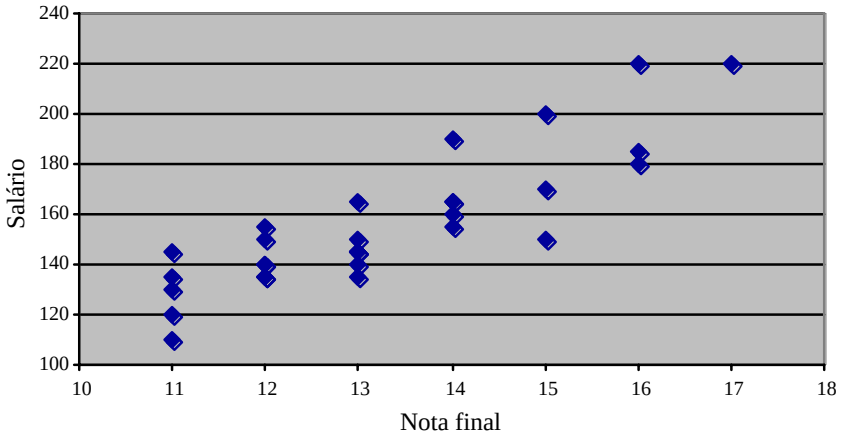
**Exemplo 7** – Uma determinada Universidade que lecciona um curso de gestão fez durante alguns anos um estudo sobre a integração dos seus alunos no campo de trabalho, nomeadamente recolhendo a informação sobre quanto tempo (em meses) tinha decorrido desde que tinham terminado a licenciatura e arranjado emprego e qual o salário auferido no início do trabalho. Obviamente que uma informação relevante para este estudo seria a nota final de curso, na posse dos serviços académicos da Universidade. Apresenta-se a seguir uma amostra dos resultados obtidos:

| Estudante | Média | Tempo proc. Emp. | Salário inicial |
|---|---|---|---|
| 1 | 15 | 1 | 200 |
| 2 | 13 | 2 | 150 |
| 3 | 12 | 3 | 135 |
| 4 | 16 | 0 | 220 |
| 5 | 12 | 4 | 140 |
| 6 | 11 | 5 | 145 |
| 7 | 14 | 3 | 155 |
| 8 | 17 | 0 | 220 |
| 9 | 11 | 5 | 135 |
| 10 | 15 | 2 | 170 |
| 11 | 13 | 3 | 145 |
| 12 | 14 | 2 | 160 |
| 13 | 14 | 1 | 190 |
| 14 | 12 | 4 | 135 |
| 15 | 13 | 3 | 140 |
| 16 | 13 | 4 | 165 |
| 17 | 15 | 1 | 150 |
| 18 | 16 | 1 | 180 |
| 19 | 13 | 4 | 135 |
| 20 | 11 | 5 | 110 |
| 21 | 12 | 4 | 140 |
| 22 | 14 | 2 | 165 |
| 23 | 16 | 1 | 185 |
| 24 | 11 | 6 | 130 |
| 25 | 12 | 5 | 155 |
| 26 | 13 | 3 | 145 |
| 27 | 12 | 5 | 150 |
| 28 | 13 | 4 | 145 |
| 29 | 14 | 2 | 165 |
| 30 | 11 | 4 | 120 |

Os valores anteriores podem ser considerados observações de variáveis aleatórias que representam o comportamento dos estudantes do curso de gestão da dita Universidade, no passado e no futuro (próximo). Vamo-nos ocupar particularmente das variáveis aleatórias X e Y que representam respectivamente a nota final de curso e o salário inicial. A representação em *box-plot* da amostra correspondente aos salários mostra que se distribuem de forma aproximadamente simétrica e não apresentam uma grande variabilidade, já que a amplitude *inter-quartil* é de 30 contos




A média é igual a 156 contos, ligeiramente superior à mediana que é igual a 150 contos, inflacionada devido à existência dos dois valores "outliers". Os alunos que terminem a licenciatura este ano lectivo podem recorrer a esta informação para ter uma estimativa de qual irá ser o seu salário inicial. Poderão nomeadamente servir-se da informação de que 50% dos alunos que terminam a licenciatura ganham entre 135 contos e 165 contos. Um destes alunos pensou que dispunha de mais alguma informação que lhe poderia ser útil para ter uma ideia de qual o salário que iria auferir, pois só lhe faltavam 3 disciplinas que não iriam afectar muito a nota final de curso. O aluno tinha construído um diagrama de dispersão dos pares (X,Y) e tinha verificado a existência de uma certa associação linear entre os pares representados:

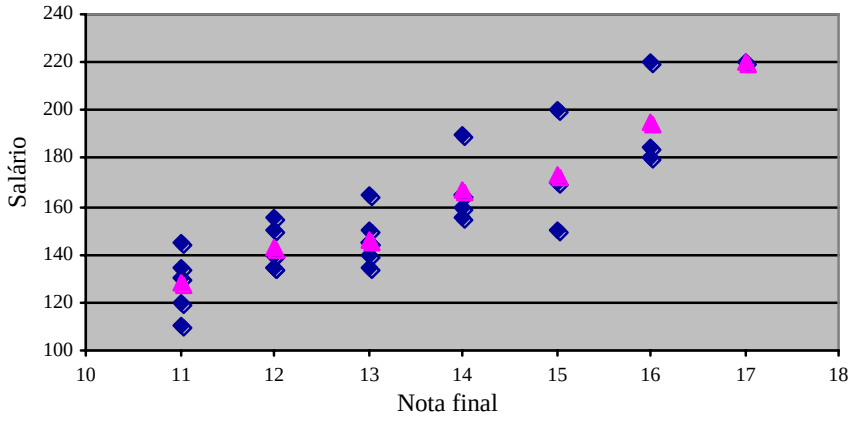



<u>Será que o facto de conhecer a média final de curso, pode tornar um pouco mais precisa a informação sobre o seu salário inicial</u>? Para já, vejamos o que se passa com os salários respeitantes a cada uma das notas:

| 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|
| 120 | 135 | 150 | 190 | 200 | 220 | 220 |
| 130 | 140 | 145 | 160 | 170 | 185 | |
| 135 | 135 | 140 | 155 | 150 | 180 | |
| 110 | 140 | 165 | 165 | | | |
| 145 | 155 | 135 | 165 | | | |
| | 150 | 145 | | | | |
| | | 145 | | | | |
| $\overline{y}_{11}$ =128 | $\overline{y}_{12}$ =142.5 | $\overline{y}_{13}$ =146.4 | $\overline{y}_{14}$ =167 | $\overline{y}_{15}$ =173.3 | $\overline{y}_{16}$ =195 | $\overline{y}_{17}$ =220 |

Efectivamente à medida que a nota final de curso cresce, cresce a média dos salários associados com cada uma das notas, isto é, em média os salários estão a crescer com a nota final. Repare-se

que as médias calculadas anteriormente são médias condicionais ao conhecimento do valor da variável X. A representação gráfica destas médias dá-nos ideia da forma como se processa o crescimento médio da variável Y, em função de X:




As considerações anteriores levam-nos à definição de valor médio condicional da variável aleatória Y, assim como à definição de regressão de Y em X (analogamente se definiria regressão de X em Y).

**Valor médio condicional**

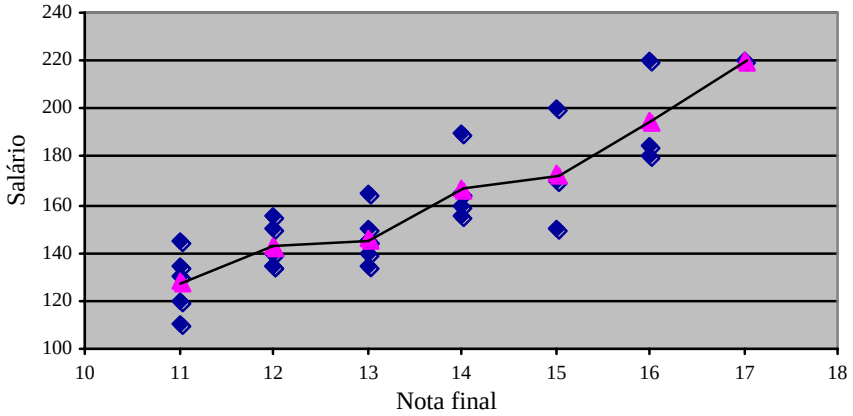
Dado o par de variáveis aleatórias (X, Y), define-se **valor médio condicional de Y dado X=x** e representa-se por $E_{Y|X}(Y)$ como sendo o valor médio de todos os valores de Y que correspondem ao mesmo valor de x.

**Regressão de Y em X**

Uma vez que para cada valor da variável aleatória X se pode definir o valor médio condicional $E_{Y|X}(Y)$, considere-se a função definida pelos pontos (x, $E_{Y|X}(Y)$). A esta função chamamos **regressão** (populacional) de Y em X. Quando esta função é linear dizemos que temos a **regressão linear** e o modelo utilizado é

$$E_{Y|X}(Y) = \alpha + \beta x$$

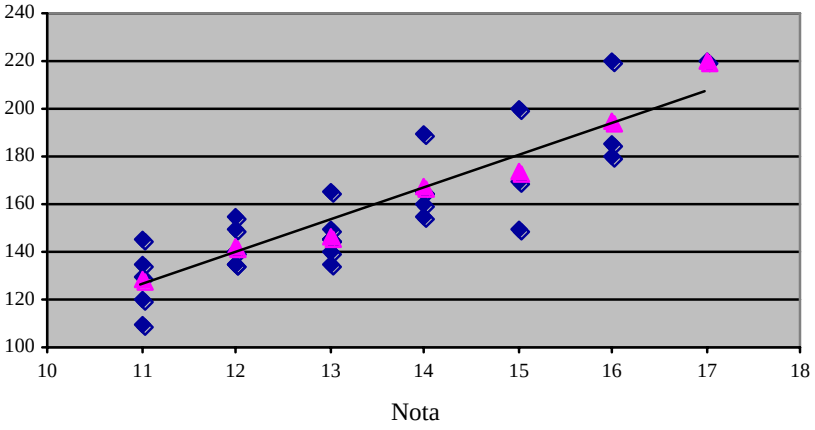
Para estimar a função de regressão a partir de uma amostra de dados bivariados, podemos utilizar vários processos, nomeadamente considerar a curva constituída pelos pontos (x, $\bar{y}_x$), como apresentamos a seguir




Para utilizar este processo de estimar a curva de regressão é necessário dispormos de um número suficientemente grande de obervações, em que os valores de x se repetem muito. Quando pudermos assumir a regressão linear, isto é, quando a representação dos pontos num diagrama de dispersão sugerir que estes podem ser aproximados por uma linha recta, então o processo utilizado para estimar os coeficientes da recta de regressão é o método dos mínimos quadrados, já estudado no capítulo 6. A utilização deste método conduz-nos à seguinte equação para a recta de regressão

$$\hat{y} = -22.277 + 13.498x$$

que se apresenta no diagrama de dispersão




**Coeficiente de determinação**

Embora o assunto da regressão não seja, no âmbito deste curso, mais aprofundado, refira-se que o quadrado do coeficiente de correlação, $r^2$, chamado **coeficiente de determinação**, dá-nos a proporção da variabilidade existente em Y que é explicada pela recta de regressão. No caso do exemplo apresentado $r^2 = 74.3$, pelo que podemos dizer que a recta de regressão ajustada aos pontos explica cerca de 74% da variabilidade existente em Y.

**Como interpretar os coeficientes $\alpha$ e $\beta$ da recta de regressão?**

Ainda referindo-nos ao exemplo anterior, suponhamos que um aluno que tinha terminado a licenciatura com nota final de 0 (obviamente que esta situação seria impossível no contexto em que estamos, isto é, não teria sentido considerar para a variável X o valor 0). Então de acordo com a equação da recta de regressão esperar-se-ia que o salário inicial fosse de aproximadamente –22 contos. Ao considerarmos anteriormente para a variável X o valor 0 para preveremos o valor para o salário inicial, estamos a cometer dois erros: em primeiro lugar não tem sentido no estudo em causa atribuir a X o valor 0; em segundo lugar, quando se pretende prever um valor para Y, utilizando a recta de regressão, não se deve considerar para X um valor que saia fora do intervalo que se considerou para construir a recta de regressão. No caso em estudo os valores para a variável X devem estar incluídos no intervalo [11, 17].

Vejamos agora o que acontece quando aumentamos na equação da recta de regressão o valor de X de uma unidade:

$$\hat{y}(x+1) - \hat{y}(x) = -22.277 + 13.498(x+1) - (-22.277 + 13.498x)$$
$$= 13.498$$

isto é, o acréscimo de uma unidade no valor de X, provoca em Y um acréscimo igual ao valor estimado para $\beta$. Podemos então dizer que um acréscimo, em média, na nota final, provoca um acréscimo, em média, de aproximadamente 13.5 contos no salário inicial. Não esqueçamos que o que a equação da recta de regressão nos dá é a variação em média de Y para um determinado valor de X. Assim, para um aluno particular que aumente de uma unidade a sua nota, não podemos garantir que o seu salário tenha um aumento de 13.5 contos. O que podemos dizer é que relativamente a todos os alunos que tenham aumentado a nota de uma unidade, em média o salário aumentará de 13.5 contos.

Este modelo será objecto de um estudo posterior, pois neste momento sai fora do âmbito deste curso um estudo mais desenvolvido. Um dos problemas que será abordado nessa altura é o da explicitação das hipóteses subjacentes, que conduzem à sua aplicação.

**Exercícios**

1. Hoje em dia, uma das preocupações das companhias de seguros é estudar a desvalorização dos carros, com a idade. Assim, com o objectivo de estudar esse fenómeno recolheu, para um determinado modelo, uma amostra de 10 carros, tendo obtido a seguinte informação sobre a idade (em anos) e o preço (em milhares de escudos):

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Idade | 6 | 4 | 3 | 4 | 5 | 3 | 8 | 4 | 9 | 3 |
| Preço | 650 | 800 | 890 | 750 | 700 | 850 | 500 | 790 | 300 | 930 |

a) Represente graficamente os dados num diagrama de dispersão

- Obtenha a recta de regressão, considerando a variável preço como variável dependente
- Interprete os valores obtidos para os coeficientes da recta de regressão
- Qual o preço previsto para um carro de 7 anos?
- Estime o preço de um carro de 15 anos. Interprete o valor obtido.

**2**. Um professor de ginástica que treina alunos de uma Universidade, pretende investigar o efeito do treino na redução do tempo que leva a correr a maratona. Assim, pôs 9 alunos num plano de treino de 3, 5 ou 7 semanas, tendo obtido os seguintes resultados:

| | | | |
|---|---|---|---|
| Redução do tempo(minutos) | 1.6, 0.8, 1.1 | 2.0, 1.7, 2.6 | 3.6, 2.8, 3.2 |
| Duração do treino (semanas) | 3 | 5 | 7 |

Analise os dados e retire conclusões.

**3**. Uma agência de aluguer de automóveis tem o seguinte plano de aluguer, por um dia, de um determinado modelo de carro: paga-se uma quantia fixa de 10 mil escudos e por cada quilómetro percorrido paga-se 75 escudos. Ao fim do dia a quantia, y, paga por um cliente será função do número de kms percorridos de acordo com a seguinte equação

$$y = 10 + .075\,x$$

a) Qual a quantia paga por um cliente que percorra 100 kms?

b) Suponha que 25 pessoas alugam o carro por um dia e percorrem exactamente 100 kms. Será que cada uma delas vai pagar exactamente a mesma quantia pelo aluguer? Explique.

**Capítulo 8**

# Alguns modelos de probabilidade

## 8.1 - Introdução

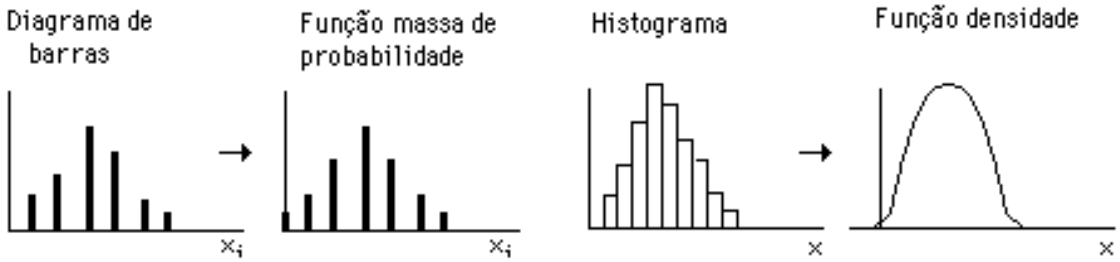
Nos capítulos anteriores, nomeadamente nos que dizem respeito às características amostrais e populacionais, realçámos o facto de que o estudo das características amostrais tem como objectivo principal, a obtenção de informação sobre as correspondentes características populacionais - é esta parte da análise estatística, que nos referimos como sendo a Inferência Estatística.

É nesta perspectiva que, por exemplo, a média e a variância amostral podem ser usadas para fazer inferência sobre os parâmetros populacionais desconhecidos, respectivamente valor médio e variância da População, de onde foi recolhida a amostra (da v.a. X que representa a População).

Pretendendo então estudar uma População **X** (que representamos, portanto, pela v.a. X), o nosso objectivo final será obter o **modelo de probabilidade** para **X**. Recordemos que, no âmbito da estatística descritiva, se estudaram vários processos de resumir a informação contida nos dados da amostra, que se recolheu da População em estudo. Alguns desses processos foram as representações gráficas.

Precisamente a representação gráfica dos dados pode dar uma informação importante sobre a distribuição da População, já que para **dados discretos** se pode interpretar o *diagrama de barras* como a imagem estatística da *função massa de probabilidade*, enquanto que para **dados contínuos** o *histograma* é a imagem estatística da *função densidade de probabilidade*. Por outro lado a *função distribuição empírica* é a imagem estatística da *função distribuição.*

Por exemplo as seguintes representações do diagrama de barras e do histograma sugerem que a f.m.p. ou a função densidade das v.a. subjacente às amostras, sejam simétricas:



Embora possa haver uma grande variedade de formas para as distribuições de probabilidade (quando falamos em distribuições de probabilidade, estamos a referir-nos indiferentemente à função massa de probabilidade ou função densidade, conforme as v.a. sejam discretas ou contínuas, ou à função distribuição), existem alguns modelos que, pela frequência com que surgem nas aplicações, merecem destaque especial.

Desses modelos realçamos três, nomeadamente o **Binomial** e o **Poisson**, para *populações discretas* e o **Normal** para *populações contínuas*.

## 8.2 - Modelo Discretos

### 8.2.1 – Modelo Uniforme

Este é um dos modelos mais simples é é caracterizado por ter uma função massa de probabilidade em que a probabilidade é constante, para um conjunto finito de pontos.

Diz-se que a variável aleatória X tem uma distribuição **uniforme** em n pontos, se assumir os valores $x_1$, $x_2$, ..., $x_n$, com probabilidade $P(X=x_i) = \frac{1}{n}$.

O exemplo mais conhecido é o modelo que descreve o lançamento de um dado equilibrado.

No caso em que $x_i=i$, i=1, 2, ..., n, de que o modelo referido anteriormente é um caso particular com n=6, tem-se $E(X) = \frac{n+1}{2}$ e $Var(X) = \frac{n^2-1}{2}$.

### 8.2.2 – Modelo Binomial

Para introduzirmos o modelo Binomial, vamos considerar a seguinte situação:

Um gerente de um centro comercial, mandou fazer publicidade do seu centro, na televisão, durante uma semana. Passados 15 dias, sobre a apresentação do anúncio, os clientes eram abordados para responderem se a sua visita se devia, ou não, ao anúncio.

Admitindo que o número de clientes a quem foi feita a pergunta é **n**, que as respostas que cada um dá são **independentes** umas das outras, e que cada cliente tem **igual probabilidade** de responder afirmativamente, a experiência anterior tem as seguintes características:

- a experiência é constituída por **n** provas, entendendo-se por **prova** uma repetição em condições idênticas
- as provas são **independentes**
- em cada uma das provas pode-se verificar **um de dois** resultados a que chamamos **sucesso** e **insucesso**, sendo constante a probabilidade de sucesso em cada prova; esta probabilidade representa-se por **p**.

A provas com estas características, chamamos **provas de Bernoulli**.

Seja **X** a v.a. que representa o número de **sucessos** em **n** provas de **Bernoulli**, em que a probabilidade de sucesso é **p**.

Relativamente ao exemplo anterior, **X** é a v.a. que representa o número de clientes, em n, que responderam afirmativamente, isto é, que tinham sido influenciados pelo anúncio.

É evidente que **X** é uma v.a. discreta que assume os valores

$$0, 1, 2, ..., n-1, n$$

**Exemplo 1** - Suponhamos, para simplificar, que foram 4 os clientes a quem foi feita a pergunta, isto é, n=4. Suponhamos ainda que o anúncio influenciou 25% dos potenciais clientes do Centro Comercial. Então a probabilidade de um cliente dizer que foi influenciado é de .25 (probabilidade do sucesso), enquanto que a probabilidade do cliente responder que não foi influenciado é de .75 (probabilidade do insucesso). Se representarmos por X, a v.a. que dá, de entre os 4 clientes, o número de clientes que responderam afirmativamente, temos que os valores possíveis para X são:

X - 0 1 2 3 4

Vejamos como obter a função massa de probabilidade de X: Representando por S - influenciado e por N - não influenciado, temos

$P(X=0)=P(NNNN)=.75^4$

$P(X=1)=P[(SNNN) \cup (NSNN) \cup (NNSN) \cup (NNNS)] = 4 \times .25 \times .75^3$

$P(X=2)=P[(SSNN) \cup (SNSN) \cup (SNNS) \cup (NSSN) \cup (NSNS) \cup (NNSS)]$

$=6 \times .25^2 \times .75^2$

$P(X=3)=P[(SSSN) \cup (SSNS) \cup (SNSS) \cup (NSSS)] = 4 \times .25^3 \times .75$

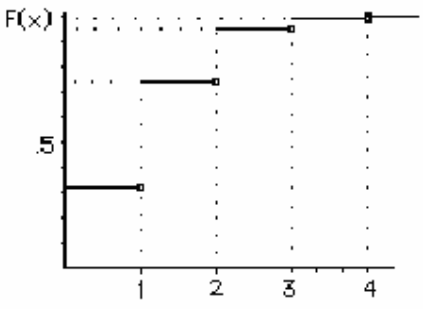
$P(X=4)=P(SSSS) = .25^4$

A f.m.p. encontra-se na tabela seguinte

| $X=x_i$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $p_i$ | .316 | .422 | .211 | .047 | .004 |

Apresenta-se a seguir a função de distribuição da v.a. X

$$\begin{aligned} F(x) &= 0 && \text{se} \quad x<0 \\ &= .316 && \text{se} \quad 0 \le x<1 \\ &= .738 && \text{se} \quad 1 \le x<2 \\ &= .949 && \text{se} \quad 2 \le x<3 \\ &= .996 && \text{se} \quad 3 \le x<4 \\ &= 1 && \text{se} \quad 4 \le x \end{aligned}$$




No caso geral, em que o número de provas é n, então

Seja **X** uma v.a. tal que **P(X=k) =** $\binom{n}{k}$ $\mathbf{p^k(1-p)^{n-k}}$ , k= 0, 1, 2, …, n

A uma v.a. **X** com esta função massa de probabilidade chamamos v.a. **Binomial** de parâmetros **n** e **p** e representamos este facto por

$$\mathbf{X \cap B(n,p)}$$

À sua distribuição chamamos **distribuição Binomial**.

*Será que as probabilidades anteriormente consideradas constituem efectivamente uma função massa de probabilidade?*

Para responder a esta questão é necessário verificar que

$$\sum_{k=0}^{n} P(X = k) = 1$$

Na verdade

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = [p + (1-p)]^n = 1$$

**Aplicação do modelo Binomial**

O modelo Binomial aplica-se sempre que estejamos perante uma situação de <u>n provas repetidas e independentes, em que em cada prova se possa verificar um de dois resultados, geralmente chamados de *sucesso* e *insucesso*, e em que se mantenha constante a probabilidade de *sucesso*</u>.

A *variável aleatória* de interesse é o *número de sucessos nas n provas*.

Situações destas surgem frequentemente em problemas de:

- prospecção de mercado
- controlo de qualidade
- etc

A partir da definição de valor médio e de variância , obtém-se

$$\mathbf{E(X) = np} \qquad \text{e} \qquad \mathbf{Var(X) = np(1-p)}$$

Relativamente ao exemplo considerado em que n=4, temos

$$E(X) = 4 \times .25 = 1$$

o que está de acordo com a intuição, pois se a 4 pessoas se fizer uma pergunta, para a qual existe uma probabilidade de 25% de dizer "sim", esperamos obter, em média, 1 resposta "sim"!

**Exercício**: Verifique que E(X)=np e Var(X)=np(1-p)

**Tabelas com as probabilidades da Binomial**

No caso do exemplo considerado anteriormente, o valor de n=4, é suficientemente pequeno, para que o cálculo das probabilidades não seja muito trabalhoso, o que não aconteceria para valores grandes de n. Assim, existem tabelas que, para alguns valores de **n** e de **p**, nos dão imediatamente os valores das probabilidades, assim como as probabilidades acumuladas, para a construção da função distribuição. Como alternativa às tabelas, temos, por exemplo, o Excel, como veremos mais à frente.

**Exemplo 2** - Outra situação que surge com frequência e em que se aplica o modelo Binomial, é no lançamento de uma moeda. Mais propriamente, o que se passa é o seguinte: lança-se uma moeda ao ar um certo número de vezes e pretende-se estudar a v.a. **X**, que representa o número de "caras" saídas nesses lançamentos. Suponhamos então que se lançou ao ar 20 vezes, uma moeda "equilibrada". Pretende-se estudar a v.a. **X**. que representa o número de caras saídas nos 20 lançamentos.

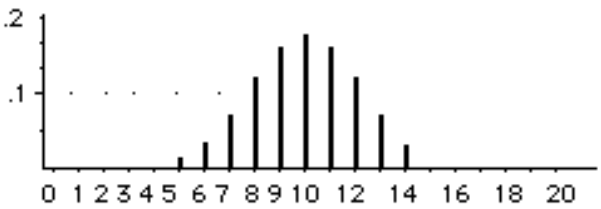
Resolução: A v.a. X assume os valores 0,1, 2,..., 20 e ficará perfeitamente definida depois de calcularmos as probabilidades de assumir esses valores. Esquematicamente, podemos escrever

$$X\begin{cases} k & k = 0,\ 1,\ 2,\ \ldots,\ n \\ p_k = P(X = k) = \binom{20}{k}.5^k.5^{20-k} \end{cases}$$

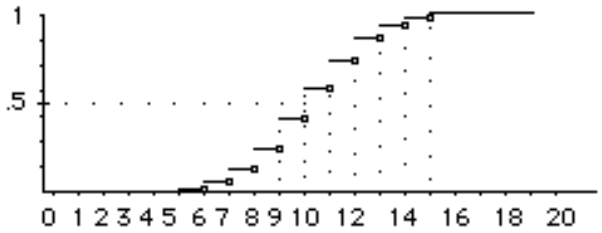
Consultando as tabelas da Binomial com n=20 e p=.5, obtemos:

| | | | |
|---|---|---|---|
| P(X=0)=.0000 | P(X=4)=.0046 | P(X=8)=.1201 | P(X=12)= .1201 |
| P(X=1)=.0000 | P(X=5)=.0148 | P(X=9)=.1602 | ..... |
| P(X=2)=.0002 | P(X=6)=.0370 | P(X=10)=.1762 | |
| P(X=3)=.0011 | P(X=7)=.0739 | P(X=11)=.1602 | |

Neste caso, em que a moeda é equilibrada, tem-se que p=1-p= .5, pelo que imediatamente se conclui que P(X=k) = P(X=20-k) com k=0, 1, …, 9. A função massa de probabilidade tem o seguinte aspecto


No que diz respeito à função distribuição, temos




**Exemplo 3** - Um estudante que não teve tempo para se preparar para um exame, em que cada questão tinha 6 respostas possíveis, em que 1 única é a correcta, decide responder ao acaso. Se o exame for constituído por 18 questões:

a) Qual a probabilidade de responder certo a uma questão?

b) Qual o número esperado de respostas certas que espera obter?

c) Qual a probabilidade de responder certo, a pelo menos11 das questões?

d) Qual a probab. de responder certo a um número de questões entre 2 e 5?

a)


Porque é que o estudante decidiu lançar um dado ao ar, para ver qual a questão a que devia responder?

C-correcto E-errado P(C)=1/6

b) Seja X a v.a. que representa o nº de respostas correctas, nas 18 questões. Então X tem uma distribuição Binomial de parâmetros 18 e 1/6, e o que se pretende é $E(X)=18 \times 1/6=3$.

c) $P(X\geq 11)=1 - P(X\leq 10) \approx 1 - .9998=.0002$

d) $P(2<X<5)=P(X=3)+P(X=4)\approx .2297+.2154=.4441$

Obs: As probabilidades anteriores foram obtidas a partir de tabelas com p=.2, em vez de p=.17, pelo que os valores obtidos são aproximados.

**E se o parâmetro p da Binomial for desconhecido?**

Existem muitas situações em que se pode aplicar o modelo Binomial, mas o parâmetro **p** é desconhecido, ao contrário do que se passa com o valor de **n**, que normalmente é conhecido, pois é possível contar o nº de provas realizadas. Então uma maneira de rodear o problema, é **estimar** o valor de p, isto é tentar obter um valor aproximado para p.

Um estimador que se costuma utilizar para estimar p e que se representa por $\hat{\mathbf{p}}$ é $\frac{\mathbf{X}}{\mathbf{n}}$, onde **X** representa o nº de sucessos em **n** provas. Estamos assim a estimar p pela frequência relativa de sucesso. Quando n for suficientemente grande, temos uma boa aproximação da probabilidade (é altura de recordar o que aprendeu sobre a teoria frequencista da probabilidade!)

**Exercício**: Verifique que efectivamente $\hat{\mathbf{p}} = \frac{\mathbf{X}}{\mathbf{n}}$, onde **X** representa o nº de sucessos em **n** provas, em que cada prova tem probabilidade de sucesso **p**, é um bom estimador da probabilidade **p**. Por outras palavras, pretende-se provar que a frequência relativa, se aproxima da probabilidade, quando o número de provas for suficientemente grande.
Sugestão: Calcule E($\hat{\mathbf{p}}$) e Var($\hat{\mathbf{p}}$).

**Amostragem com reposição**

No processo de amostragem que consiste em retirar aleatoriamente uma amostra de uma população, *com reposição*, em que para cada indivíduo recolhido se verifica se sim ou não tem determinada propriedade, repondo o elemento recolhido antes de proceder a nova extracção, estamos em condições de aplicar o modelo binomial, quando se pretende estudar a variável aleatória que representa o número de indivíduos da amostra, com a dita propriedade.

**Exemplo 4** – O gerente de uma casa que vende material informático fez uma encomenda de 20 impressoras de determinada marca, que será aceite mediante a inspecção de 3 das impressoras, para ver se funcionam ou estão avariados. Quando a encomenda chega o gerente analisa as 3 primeiras impressoras a serem descarregadas. Embora o gerente não saiba, 2 das impressoras têm avarias. Será que estamos perante uma experiência binomial? Resolução: Estamos perante uma experiência constituída por 3 provas, em que em cada prova se pode verificar o sucesso (impressora avariada) ou insucesso (impressora boa). A probabilidade de seleccionar uma impressora defeituosa é 2/20, admitindo que qualquer uma das impressoras poderia ter sido colocada no meio de transporte, em melhores condições de ser a primeira a ser descarregada. No entanto as provas não são independentes, já que a probabilidade de obter uma impressora defeituosa na 2ª prova ou na 3ª prova depende do que aconteceu nas provas anteriores, pelo que a probabilidade de sucesso não se mantém constante ao longo das provas. Assim, não estamos perante uma experiência binomial.

**Amostragem sem reposição em populações "infinitas"**

Se na experiência anterior a dimensão **N** da população, de onde foi recolhida a amostra, fosse suficientemente grande, relativamente à dimensão **n** da amostra recolhida, então a probabilidade de sucesso não sofreria alterações significativas de prova para prova. Nestas condições

poderíamos ainda utilizar o modelo Binomial. Como indicação, para as aplicações, o modelo Binomial não deve ser aplicado se n/N≥.05 (Mendenhall, 1994) (Há autores que consideram que ainda se pode aplicar o modelo Binomial se a dimensão da amostra for inferior a 10% da dimensão da população).

**Somas de variáveis aleatórias independentes com distribuição Binomial**

**Propriedade**: Dadas as v.a. $X_i$ independentes, com distribuição Binomial, $\mathbf{X_i \cap B(n_i,p)}$, i=1, 2, …, n, então a soma $\mathbf{S_n}= X_1 + X_2 + \ldots + X_n$, também tem distribuição Binomial

$$\mathbf{S_n \cap B(\sum_{i=1}^{n} n_i, p)}$$

Dem: Para demonstrar o resultado anterior, basta fazer a demonstração para n=2 (Porquê?).

Consideremos então as v.a. $X_1$ e $X_2$, com $X_i \cap B(n_i,p)$, i=1, 2 e $S_2=X_1 + X_2$ .

$$P(S_2=k) = \sum_{i=0}^{k} P(X_1=i \text{ e } X_2=k-i) = \sum_{i=0}^{k} \binom{n_1}{i} p^i (1-p)^{n_1-i} \binom{n_2}{k-i} p^{k-i} (1-p)^{n_2-k+i}$$

$$= \sum_{i=0}^{k} \binom{n_1}{i}\binom{n_2}{k-i} p^k (1-p)^{n_1+n_2-k}$$

$$= \binom{n_1+n_2}{k} p^k (1-p)^{n_1+n_2-k} \qquad \text{em que } k= 0,1, \ldots, n_1 + n_2 .$$

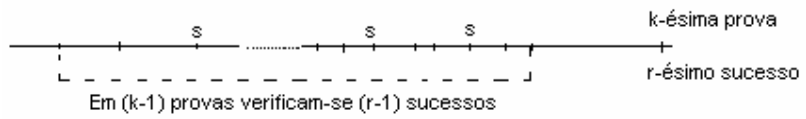
Como consequência da propriedade anterior, temos que uma v.a. **X** com distribuição Binomial de parâmetros **n** e **p**, pode ser considerada a soma de **n** variáveis aleatórias, independentes, cada uma com distribuição Binomial de parâmetros 1 e p ( variáveis aleatórias de Bernoulli).

**Exercício:** Tendo em consideração o que foi dito anteriormente, determine o valor médio e a variância de X.

### 8.2.3 - Modelo Binomial Negativa

Consideremos ainda uma sucessão de provas de Bernoulli, isto é, provas independentes, em que em cada prova a probabilidade de sucesso é constante e igual a **p** (sendo a de insucesso igual a q=1-p). Suponhamos que estamos interessados na variável aleatória X que representa o número de provas necessárias para se obter **r** sucessos. Repare na analogia com o modelo Binomial: enquanto que neste o número de provas é fixo e o número de sucessos é aleatório, na Binomial negativa o que é fixo é o número de sucessos, enquanto que o número de provas é aleatório. Vejamos quais os valores que a variável aleatória X assume e com que probabilidades:

X pode assumir os valores k = r, r+1, r+2, .... Por outro lado, para que na k-ésima prova se verifique o r-ésimo sucesso, é necessário que nas (k-1) provas anteriores se verifiquem (r-1) sucessos:




Assim, a probabilidade de serem necessárias k provas, para se verificarem r sucessos, é

P(X = K) = P(em (K-1) provas verificarem-se (r-1) sucessos e na K-ésima prova verificar-se sucesso)

= P(em (K-1) provas verificarem-se (r-1) sucessos) ×P(sucesso na k-ésima prova)

$$= \binom{k-1}{r-1} p^{r-1}(1-p)^{k-r} \times p$$

$$= \binom{k-1}{r-1} p^{r}(1-p)^{k-r}$$

Pode-se mostrar que E(X) = r/p e Var(X) = r(1-p)/$p^2$.

Uma variável aleatória **X** com função massa de probabilidade

$$P(X=k) = \binom{k-1}{r-1} p^{r}(1-p)^{k-r}, \quad k= r, r+1, r+2, \quad , 0<p<1,$$

diz-se que tem distribuição **Binomial Negativa** e representa-se simbolicamente por X∩BinNeg(r,p).

Caso particular – No caso em que r=1, diz-se que se tem o modelo **Geométrico**, que representa, portanto, o número de provas necessárias para se verificar sucesso (pela 1ª vez), representando-se por Geom(p). A função massa de probabilidade é dada por

$$P(X=k) = (1-p)^{k-1}p, \text{ com } k = 1, 2, 3, ...$$

**Exemplo 5** – Um indivíduo faz anos em Junho. Resolve, na rua, perguntar às pessoas que encontra, qual o mês em que fazem anos, até encontrar duas que façam anos no mesmo mês. Qual a probabilidade de ter de importunar 10 pessoas? E se pretender encontrar 1 pessoa a fazer anos em Junho, em vez de 2? Em média a quantas pessoas tem de fazer a pergunta, até encontrar uma a fazer anos no mesmo mês?

Seja X a v.a. que representa o número de pessoas a quem tem de importunar, para encontrar 2 a fazerem anos em Junho. Então X∩BinNeg(2,1/12), se admitirmos que existe igual probabilidade de fazer-se nos em qualquer um dos 12 meses, donde

$$P(X=10) = \binom{10-1}{2-1}\left(\frac{1}{12}\right)^2\left(1-\frac{1}{12}\right)^{10-2}$$

$$= 9 \times (1/12)^2 \times (11/12)^8$$

= 0.031

Quando r=1, temos Y∩BinNeg(1,1/12) e P(Y = 10) = $(11/12)^9$ x1/12 = 0.038. Neste caso E(Y) = 12

**Exemplo 6** (adaptado de De Veaux et al, 2004) – Os indivíduos com sangue de tipo O, RH-, são chamados de dadores universais. Só 6% da população tem este tipo de sangue. a) Quantos dadores espera observar, na unidade móvel que costuma estacionar em Entrecampos, Lisboa, até obter alguém que seja dador universal? Qual a probabilidade de que o primeiro dador universal se encontre entre os 4 primeiros dadores? b) Suponha que chegam 20 dadores à unidade móvel. Quantos dadores universais espera encontrar? Qual a probabilidade de encontrar 2 ou 3 dadores universais?

a) Se representarmos por X o número de dadores até obter 1 que seja dador universal, podemos considerar que X tem uma distribuição geométrica de parâmetro 0.06.

Então E(X) = 1/0.06 = 16.7, pelo que se espera examinar em média 16.7% de pessoas até encontrar um dador universal.

A probabilidade de que se encontre um dador de tipo O-, nos 4 primeiros é dada por

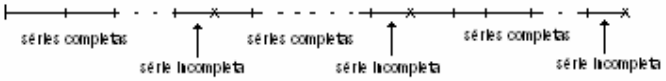
P(X=1)+P(X=2)+P(X=3)+P(X=4) = 0.2193, pelo que cerca de 22% das vezes encontra-se um dador universal nos primeiros 4 dadores que se apresentam.

b) Neste caso temos uma variável aleatória Y com distribuição Binomial de parâmetros 20 e 0.06. Então E(Y) = 20×0.06 = 1.2 e P(Y = 2 ou 3) = P(Y = 2) + P(Y = 3) = 0.3106.

**Exemplo 7** (adaptado de Murteira et al, 2002) – A probabilidade de que determinada máquina se avarie quando efectua uma série de fabrico é p=0.1; quando a máquina se avaria, a série em curso considera-se perdida. As avarias são reparadas por substituição de uma peça, de que a unidade fabril tem duas em reserva. Supõe-se que as avarias são independentes do que se passou nas séries anteriores e que a máquina está presentemente em boas condições de funcionamento. Representando por Y o número de séries completas produzidas até a máquina parar por não haver mais peças de reserva, determine:

- o número esperado de séries completas;
- a probabilidade de se completarem mais de 30 séries;
- o número de peças de reserva R que assegura a produção de pelo menos 50 séries com probabilidade 0.95.

Se Y é a v.a. que representa o número de séries completas até a máquina parar, podemos considerar X = Y + 3, o número de séries completas e incompletas, em que 3 são as peças que foram substituídas – as duas suplentes e a que a máquina tinha e X ∩BinNeg(3,0.1):




Na linguagem até aqui utilizada, podemos dizer que se X é uma variável aleatória que representa o número de provas até se obter r sucessos, então Y = X-r, é a v.a. que representa o número de

insucessos, até se verificarem r sucessos. Facilmente se mostra, a partir da distibuição de X∩BinNeg(r,p), que P(Y = k) = $\binom{r+k-1}{k} p^r (1-p)^k$ , com k=0, 1, 2, ..., com E(Y) = r(1-p)/p e Var(Y) = r(1-p)/$p^2$.

Então E(Y) = 27 e P(Y>30) = 1 – P(Y≤30) = 1 - $\sum_{k=0}^{30} \binom{3+k-1}{k} 0.1^3 \times 0.9^k$ = 0.345

Se R representa o número de peças de reserva, então pretende-se calcular P(X≥50+R+1) com X∩BinNeg(R,0.1) e vai-se procurar o menor R tal que

$\sum_{k=51+R}^{\infty} \binom{k-1}{R-1} 0.1^R \times 0.9^{k-R}$ ≥0.95. Substituindo R por vários valores obtemos o seguinte qadro

| R | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|
| P(X≥50+R+1) | 0.66 | 0.78 | 0.87 | 0.93 | 0.96 |

donde concluímos que o valor de o número de peças de reserva eve ser 11.

### 8.2.4 - Modelo de Poisson

Vamos introduzir seguidamente um outro modelo de probabilidades, também *discreto* e que se aplica em situações em que se está interessado em estudar o número de ocorrências de um acontecimento, num determinado intervalo de tempo ou espaço.

Suponhamos que se verificam as seguintes hipóteses:

- A probabilidade de uma ocorrência do acontecimento, é a mesma para quaisquer dois intervalos de igual amplitude.
- A ocorrência ou não ocorrência do acontecimento num determinado intervalo, é independente da ocorrência ou não ocorrência do acontecimento num outro qualquer intervalo.

Representando por **X** a v.a. que dá o número de ocorrências na unidade de tempo, então **X** tem uma **distribuição de Poisson**, com f.m.p. dada por

$$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}, \qquad \lambda>0, \qquad k=0, 1, 2, \ldots$$

Na expressão da função massa de probabilidade aparece a constante **λ**, que é o único *parâmetro* da distribuição e a que normalmente se dá o nome de intensidade da distribuição.

Uma v.a. com distribuição de Poisson, chama-se v.a. de **Poisson** e representa-se este facto com a seguinte notação

$$\mathbf{X \cap P(\lambda)}$$

Dada uma v.a. X com distribuição de Poisson de parâmetro λ, pode-se mostrar que

$$\mathbf{E(X) = \lambda} \qquad \text{e} \qquad \mathbf{Var(X) = \lambda}$$

**Exercício**: Verifique que $E(X)=\lambda$ e $Var(X)=\lambda$

Repare-se na particularidade do valor médio e da variância serem iguais.

**Aproximação da distribuição Binomial pela Distribuição de Poisson**

**Propriedade**: A distribuição Binomial B(n,p) converge para a distribuição de Poisson $P(\lambda)$, quando $n\to\infty$ (o número de provas aumenta), $p \to 0$ (a probabilidade de sucesso tende para zero) e o produto np se mantém aproximadamente constante, $np = \lambda > 0$ ( o nº médio de sucessos mantém-se aproximadamente constante ao longo das provas).

Dem: Fazendo $p= \lambda/n$, na expressão da P(X=k) da Binomial

$$\binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} (\lambda/n)^k (1 - \lambda/n)^{n-k}$$

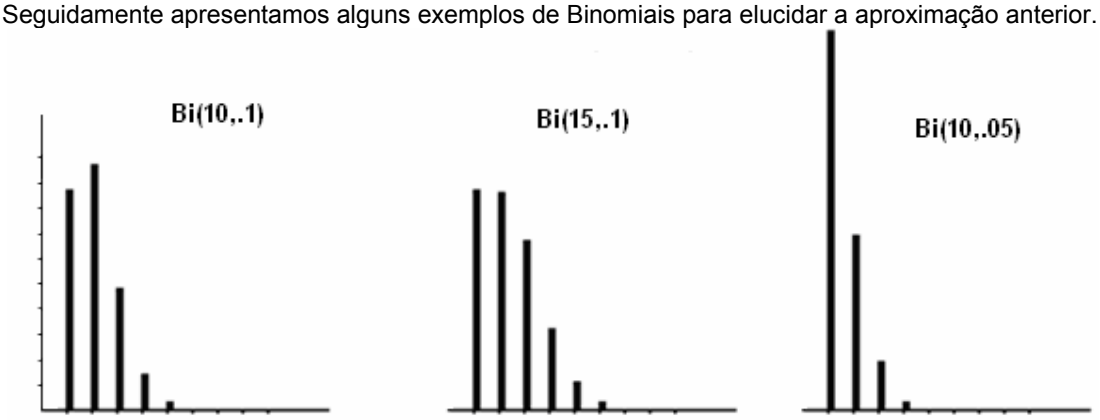
$$= \frac{n(n-1)\ \ldots\ (n-k+1)}{n^k} (1 - \lambda/n)^{-k} (1 - \lambda/n)^n \lambda^k/k!$$

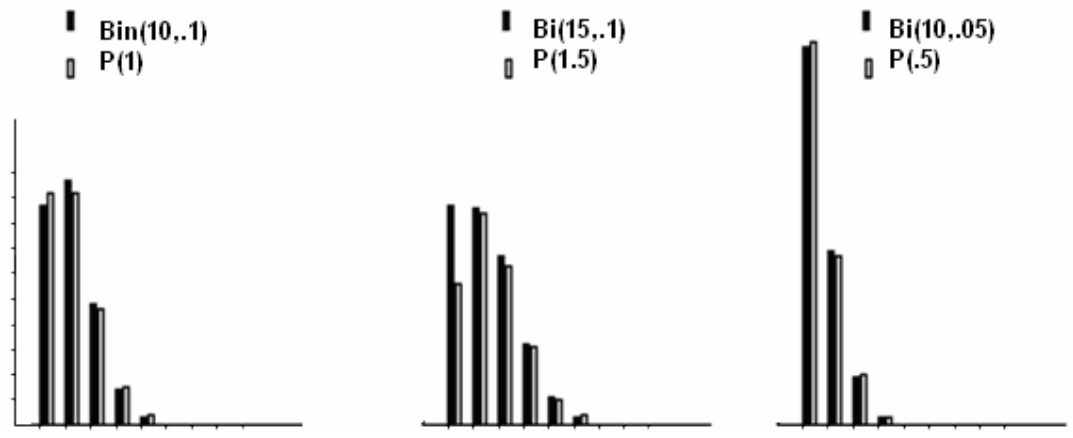
e calculando o limite da expressão anterior quando $n\to\infty$ e $p\to 0$, obtemos a fórmula que nos dá a função massa de probabilidade da Poisson.

O resultado anterior dá-nos as condições em que o modelo Poisson aparece como limite do modelo Binomial, razão pela qual é conhecido como **lei dos acontecimentos raros**. Temos:

- uma situação de muitas provas de Bernoulli ($n\to\infty$);
- com pequena probabilidade de sucesso ($p\to 0$);
- e em que o número esperado de sucessos se mantém constante ($np = \lambda$).

Seguidamente apresentamos alguns exemplos de Binomiais para elucidar a aproximação anterior.






**Aplicação do modelo de Poisson**

O modelo de Poisson aplica-se em situações em que estamos interessados em estudar o número de ocorrências de determinado acontecimento num certo intervalo de tempo ou num certo espaço. Exemplos concretos em que se utiliza este modelo são nomeadamente no estudo:

- do número de partículas radioactivas recebidas por um contador Geiger, num determinado intervalo de tempo
- do número de clientes chegados a um serviço, num determinado intervalo de tempo
- do número de chamadas telefónicas chegadas a uma central, num determinado intervalo de tempo
- do número de bactérias num certo reticulado.

**Exemplo 8** - O número de pedidos de ambulâncias que chegam, por dia, a determinado posto de socorros, é em média de 2. Calcule a probabilidade de que:

a) Num dia, haja pelo menos um pedido.

b) Num dia haja pelo menos um pedido, sabendo que no dia anterior não se registou nenhum.

c) Num dia haja dois pedidos e no dia seguinte também se verifiquem dois pedidos.

Resolução: Seja **X** a v.a. que representa o número de pedidos de ambulâncias por dia. Podemos considerar que X tem distribuição de Poisson de parâmetro $\lambda=2$.

a) $P(X\geq 1)=1 - P(X<1)= 1 - P(X=0)=e^{-2}\, 2/0! = .865$

b) Esta probabilidade é igual à anterior! (Basta ter em atenção a forma como foi introduzido o modelo de Poisson)

c) Vamos representar por

$X_1$ - nº de pedidos num dia; $X_2$ - nº de pedidos no dia seguinte

então, pela mesma razão invocada na alínea anterior, temos

$$P(X_1 =2 \text{ e } X_2 =2)=P(X_1 =2)\, P(X_2 =2) = .271 \times .271 = .073$$

**Exemplo 9** - Em 1945 os alemães bombardearam Londres com as bombas V2. A região londrina está dividida em 576 distritos de superfícies semelhantes, pelo que admitimos que cada distrito tem probabilidade idêntica de ser bombardeado. Calcula-se que o número de bombas recebidas por Londres foi de 535. Calcule as probabilidades de cada distrito receber 0, 1, 2, ..., bombas.

Resolução: Seja X a v.a. que representa o nº de bombas recebidas por cada distrito. Então podemos assumir que $X \cap Bi(535,1/576)$ donde:

P(X=0)=.3947

P(X=1)=.3672

P(X=2)=.1705

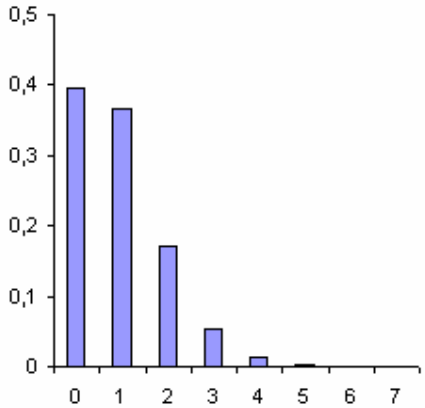
P(X=3)=.0527

P(X=4)=.0121

P(X=5)=.0023

P(X=6)=.0003

$\sum_{k=7}^{535} P(X=k)$=.0002




Como n é grande e p é pequeno, vamos aproximar a Binomial por uma Poisson com parâmetro $\lambda$=535/576. Os resultados obtidos considerando $X \cap P(535/576)$ são os seguintes:

P(X=0)=.3950

P(X=1)=.3669

P(X=2)=.1704

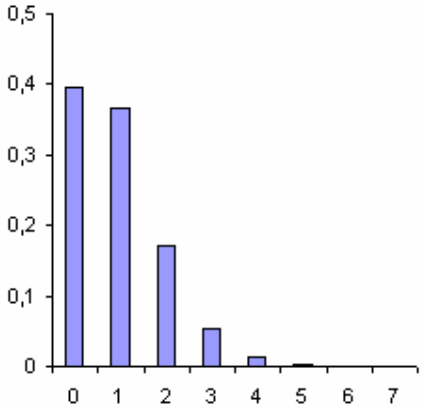
P(X=3)=.0528

P(X=4)=.0122

P(X=5)=.0023

P(X=6)=.0003

$\sum_{k\geq 7} P(X=k)$=.0001




Comparando estes resultados, com os obtidos anteriormente, verificamos que são muito semelhantes.

**Exemplo 10** – Qual a probabilidade de que numa empresa com 550 empregados, exactamente k façam anos no dia de Natal?

Resolução: Podemos considerar um esquema de provas de Bernoulli de 550 provas e em que se considera a probabilidade de sucesso p=1/365. Nestas condições, como n é grande e p é pequeno, podemos utilizar a aproximação da Binomial pela Poisson com $\lambda$=550/365=1.5 e vem para as probabilidades

| k | 0 | 1 | 2 | 3 | 4 | 5 | ≥6 |
|---|---|---|---|---|---|---|---|
| Prob. | .223 | .335 | .251 | .125 | .047 | .015 | .003 |

**E se o valor do parâmetro $\lambda$ for desconhecido?**

Por vezes acontece que estamos em situação de aplicar o modelo de Poisson, mas desconhecemos o valor do parâmetro. Nestas circunstâncias o que se faz, é **estimar** o parâmetro desconhecido. Tendo em atenção que o parâmetro $\lambda$ a estimar é o valor médio da distribuição, e a analogia existente entre características amostrais e populacionais, uma estimativa natural para o parâmetro $\lambda$ é a média.

**Exemplo 11 -** Apresentamos seguidamente os resultados das experiências de Rutherford e Geiger. Estes cientistas observaram o número de partículas $\alpha$ emitidas por uma substância radioactiva, durante 2608 períodos de 7.5 segundos, obtendo os resultados apresentados na tabela seguinte:

| i | $n_i$ | i | $n_i$ |
|---|---|---|---|
| 0 | 57 | 5 | 408 |
| 1 | 203 | 6 | 273 |
| 2 | 383 | 7 | 139 |
| 3 | 525 | 8 | 45 |
| 4 | 532 | 9 | 27 |
| | | 10 | 16 |

Estamos a representar por $n_i$ o número de períodos em que foram emitidas i partículas. Representando por X a v.a. que dá o número de partículas radioactivas emitidas em cada período, podemos considerar que esta v.a. tem distribuição de Poisson. A partir da tabela anterior é possível calcular quantas partículas foram emitidas em média, valor esse que será considerado para estimativa do parâmetro da distribuição considerada. Representando o parâmetro estimado por $\hat{\lambda}$, temos

$$\hat{\lambda} = \sum \frac{n_i . i}{n} = 3.87$$

Então $p_i = P(X=i) = \dfrac{3.87^i e^{-3.87}}{i!}$ para i= 0, 1, 2, …

**Soma de variáveis aleatórias independentes com distribuição de Poisson**

**Propriedade**: Dadas as v.a. $X_i$, independentes, com distribuição de Poisson, $\mathbf{X_i \cap P(\lambda_i)}$, i=1, 2, …, n, então a soma $\mathbf{S_n} = X_1 + X_2 + \ldots + X_n$, também tem distribuição de Poisson

$$S_n \cap P(\sum_{i=1}^{n} \lambda_i)$$

Dem: A demonstração é análoga à que foi feita para o caso da Binomial.

Como consequência da propriedade anterior, temos que uma v.a. **X** com distribuição de Poisson de parâmetro $\lambda$, pode ser considerada a soma de **n** variáveis aleatórias, independentes, cada uma com distribuição de Poisson de parâmetro $\lambda/\mathbf{n}$.

### 8.2.5 – Modelo hipergeométrico

Vimos, no estudo do modelo Binomial, que o modelo Binomial podia ser aplicado para estudar a v.a. que representa o número de elementos de uma amostra que possuem determinada característica, quando a amostra é extraída com reposição ou no caso de ser sem reposição, se a dimensão N da população for suficientemente grande, quando comparada com a dimensão n da amostra. Esta situação advinha do facto de nestas condições as provas (extracções sucessivas) poderem ser consideradas independentes mantendo-se constante a probabilidade de sucesso (o elemento recolhido possuir a característica). Efectivamente a probabilidade de sucesso que é dada pela proporção p de elementos da população possuindo a característica, não se altera substancialmente de prova para prova.

Então o que acontece quando se procede a uma extracção sem reposição, numa população finita (que não é suficientemente grande)? Consideremos o seguinte exemplo:

**Exemplo 12** – Uma caixa contém 12 garrafas de vinho, das quais 3 são de vinho branco e as restantes de vinho tinto. Retira 4 garrafas da caixa. Qual a probabilidade de 2 serem de vinho branco?

Resolução: O número de *maneiras possíveis* de retirar 4 garrafas da caixa se 12 é dado pelas combinações de 12, 4 a 4. Destas maneiras possíveis só são *favoráveis* as que tiverem 2 garrafas de vinho branco e 2 de vinho tinto, cujo número se obtém multiplicando as combinações de 3, 2 a 2 pelas combinações de 9, 2 a 2. De seguida basta usar a definição clássica de probabilidade. Formalizando o raciocínio anterior, vem:

Seja X a v.a. que representa o número de garrafas de vinho branco existentes na amostra de 4 garrafas retiradas de uma população constituída por 12 garrafas – 3 de vinho branco e 9 de vinho tinto

$$P(X=2) = \frac{\binom{3}{2}\binom{9}{2}}{\binom{12}{4}} = .218$$

A v..a. X pode assumir os valores 0, 1, 2 ou 3, pois embora a amostra tenha dimensão 4, o número de garrafas de vinho branco são só 3. Analogamente se calculavam as probabilidades da v.a. X assumir os outros valores 0, 1 ou 3.

Consideremos uma população de **N** elementos dos quais $N_1$ possuem determinada característica – sucessos, enquanto que os restantes $N_2 = N - N_1$ elementos não a possuem. Seja **n** a dimensão de uma amostra retirada da população e **X** a v.a. que representa o número de sucessos na amostra. Então a v.a. X tem uma distribuição **Hipergeométrica** cuja função massa de probabilidade é

$$P(X=k) = \frac{\binom{N_1}{k}\binom{N-N_1}{n-k}}{\binom{N}{n}} \qquad k = \max(0, N_1+n-N), \ldots, \min(N_1,n)$$

**Exercício:** Verificar que as probabilidades anteriores definem efectivamente uma função massa de probabilidade.

Dada uma v.a. X com distribuição Hipergeométrica de parâmetros N, n e $p = \frac{N_1}{N}$ (probabilidade de sucesso num elemento extraído ao acaso), pode-se mostrar que

$$\mathbf{E(X)} = np \text{ e } \mathbf{Var(X)} = np(1-p)\,\frac{N-n}{N-1}.$$

**Exercício**: Mostrar que **E(X)** = np e **Var(X)** = $np(1-p)\,\frac{N-n}{N-1}$.

Resolução: Embora a demonstração das propriedades anteriores possa ser feita por cálculo directo, vamos utilizar o seguinte raciocínio: Seja uma população constituída por $N_1$ elementos possuindo determinada propriedade (sucessos) e $N_2$ elementos sem essa propriedade (insucessos). Retiremos, sem substituição, uma amostra de dimensão n e seja $S_n$ o número de sucessos obtidos. Seja $X_k$ uma v.a. que assume os valores 1 ou 0, conforme o k-ésimo elemento da amostra for sucesso ou insucesso. A probabilidade de $X_k$ ser igual a 1 é $N_1/(N_1+N_2)$, donde

$$E(X_k) = \frac{N_1}{N_1+N_2} \quad \text{e} \quad Var(X_k) = \frac{N_1N_2}{(N_1+N_2)^2}$$

Por outro lado, se $j \neq k$, então $X_jX_k=1$ se os j-ésimo e k-ésimo elementos da amostra forem 1, e isto verifica-se com probabilidade $N_1(N_1-1)/(N_1+N_2)(N_1+N_2-1)$ donde

$$E(X_jX_k) = \frac{N_1(N_1-1)}{(N_1+N_2)(N_1+N_2-1)} \quad \text{e} \quad Cov(X_jX_k) = \frac{-N_1N_2}{(N_1+N_2)^2(N_1+N_2-1)}$$

donde

$$E(S_n) = E(X_1+X_2+\ldots+X_n) = n\frac{N_1}{N_1+N_2} \quad \text{e} \quad Var(S_n) = \frac{nN_1N_2}{(N_1+N_2)^2}\left\{1-\frac{n-1}{N_1+N_2-1}\right\}$$

tendo em consideração que Var($\sum_{K=1}^{n} X_k$) $= \sum_{K=1}^{n} \mathrm{Var}(X_k) + 2\sum_{j,k} \mathrm{Cov}(X_j, X_k)$ com este somatório estendido aos $\binom{n}{2}$ pares ($X_jX_k$) com j<k.

Ao estudar o modelo Binomial dissemos que numa situação de amostragem sem reposição em populações "infinitas", esse modelo ainda poderia ser aplicado, quando efectivamente o modelo correcto é o Hipergeométrico, como foi agora estudado. Verifique que as expressões para o valor médio e variância anteriormente consideradas justificam essa aplicação.

**Exemplo 13** – Uma loja que vende componentes electrónicas, recebe-as em lotes de 12. Algumas das componentes vêm avariadas pelo que o gerente da loja, com o objectivo de minimizar o tempo dispensado a verificar se todas funcionam, estabeleceu o seguinte plano de amostragem: retira 4 e aceita o lote se não encontrar nenhuma defeituosa. Qual a probabilidade de não rejeitar o lote, sabendo que existem 3 componentes defeituosas?

Resolução: O lote será aceite se na amostra recolhida não se verificar nenhuma componente defeituosa. Seja X a v.a. que representa o nº de componentes defeituosas (sucessos) em 4 componentes retiradas de uma população constituída por 9 componentes boas e 3 defeituosas. Pretende-se calcular P(X=0)

$$P(X=0) = \frac{\binom{9}{4}\binom{3}{0}}{\binom{12}{4}} \approx .25$$

A probabilidade do lote não ser rejeitado é aproximadamente de 25%.

**Exemplo 14** (Feller, 1968) – Num país com 50 estados, cada estado tem dois senadores. Num grupo de 50 senadores, qual a probabilidade que um determinado estado esteja representado?

Resolução: Temos N=100 senadores dos quais $N_1$=2 representam o tal estado. Representando por X a v.a. que dá o nº de senadores do tal estado, numa amostra de 50, pretende-se P(X=1)+P(X=2) ou 1-P(X=0), donde

$$P(X=0) = \frac{\binom{2}{0}\binom{98}{50}}{\binom{100}{50}} = .247$$

pelo que a probabilidade pretendida é .753.

E qual a probabilidade de que todos os estados estejam representados?

Neste momento já não estamos em condições de aplicar o modelo Hipergeométrico, mas o problema é suficientemente simples para poder ser resolvido utilizando a definição clássica de probabilidade. Então o <u>número de casos favoráveis</u> que podemos considerar é $2^{50}$, uma vez que se pretende construir uma amostra de dimensão 50, em que cada elemento seja proveniente de

um conjunto de 2. Como o número de casos possíveis é dado pela combinações de 100, 50 a 50, vem para a probabilidade pretendida o valor

$$\frac{2^{50}}{\binom{100}{50}} \approx 4.126.\ 10^{-14}.$$

A distribuição hipergeométrica tem sido sido aplicada com sucesso na estimação da dimensão de populações animais, utilizando métodos de captura e recaptura (Feller, 1968).

## Utilização do Excel para calcular probabilidades dos modelos discretos

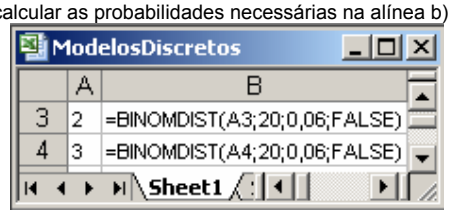
O Excel disipõe de funções que dão as probabilidades dos modelos discretos considerados anteriormente. Assim, temos:

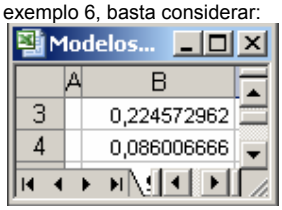
**Modelo Binomial**

Função ***BINOMDIST****(number_s; trials; probability_s; cumulative)*, onde:

- *Number_s* é o número de sucessos nas provas;
- *Trials* é o número de provas independentes;
- *Probability_s* é a probabilidade de sucesso
- *Cumulative* é um valor lógico: para obter a função distribuição, usar TRUE; para obter a função massa de probabilidade, usar FALSE.

Exemplo – Para calcular as probabilidades necessárias na alínea b) do exemplo 6, basta considerar:



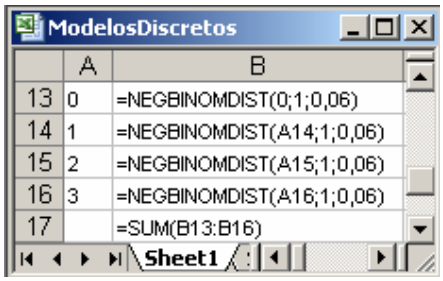



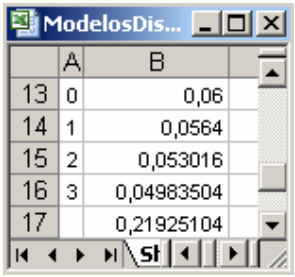

**Modelo Binomial negativa**

Função ***NEGBINOMDIST****(number_f; number_s; probability_s)*, onde:

- Number_f é o número de falhas ou insucessos;
- Number_s é o número se sucessos;
- Probability_s é a probabilidade de sucesso.

Exemplo – Para calcular as probabilidades necessárias na alínea a) do exemplo 6, basta considerar




Nota – Repare-se que Number_f é o número de falhas ou insucessos antes de um sucesso, enquanto que nós definimos a Binomial negativa como o número de provas até se obter sucesso, isto, é contamos os insucessos e a prova em que se deu sucesso.

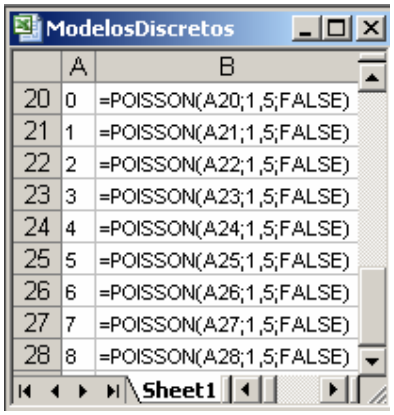
**Modelo Poisson**

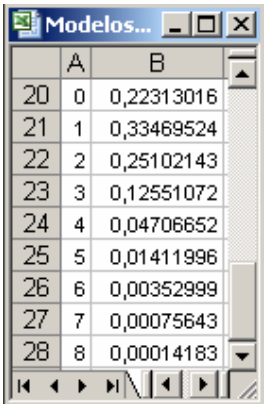
Função ***POISSON****(x; mean; cumulative)*, onde:

- *x* é o número de acontecimentos;
- *Mean* é o valor médio

- *Cumulative* é um valor lógico: para obter a função distribuição, usar TRUE; para obter a função massa de probabilidade, usar FALSE.

Exemplo – Para calcular as probabilidades do exemplo 10, basta considerar:



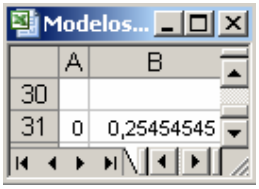




**Modelo Hipergeométrico**

Função ***Hypgeomdist****(sample_s; number_sample; population_s; number_population)*, onde

- Sample_s é o número de sucessos na amostra;
- Number_sample é a dimensão da amostra;
- Population_s é o número de sucessos na população;
- Number_population é a dimensão da população.

Exemplo – Considerando o exemplo 13, a probabilidade pretendida pode ser calculada da seguinte forma:

## 8.3 – Modelos Contínuos

### 8.3.1 - Modelo Normal

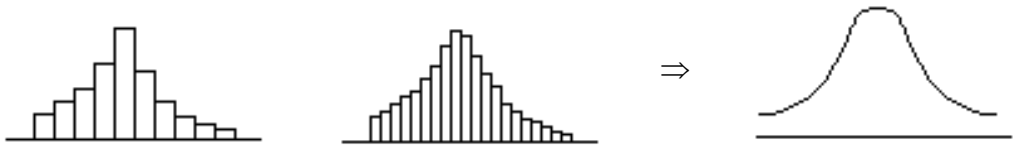
As distribuições consideradas anteriormente referem-se a v.a. discretas, isto é, v.a. que só podem tomar um número finito (modelo Binomial ou hipergeométrico) ou infinito numerável (modelo de Poisson, modelo Binomial negativa) de valores distintos.

Vamos seguidamente estudar uma v.a. de tipo *contínuo*, isto é, uma v.a. que pode assumir qualquer valor de um intervalo e que é identificada pela sua função densidade de probabilidade.

Antes de prosseguirmos, convém recordar alguns aspectos do tratamento das v.a. contínuas, que se distinguem do das v.a. discretas. Assim:

- Não falaremos da probabilidade da v.a. tomar um determinado valor, já que esta probabilidade é, para as v.a. contínuas, nula. Falar-se-á, no entanto, da probabilidade da v.a. assumir valores de um intervalo.
- Por outro lado, o cálculo da probabilidade da v.a. assumir qualquer valor de um intervalo [a,b], será dado pela área compreendida pelo gráfico da função densidade de probabilidade, o eixo das abcissas e as rectas x=a e x=b.

A **distribuição Normal**, das distribuições contínuas, a mais conhecida, foi obtida matematicamente por Gauss, como a distribuição dos erros de medidas, tendo-lhe dado o nome sugestivo de "lei normal dos erros". A partir daí, astrónomos, físicos e mais tarde, cientistas de outros campos, que manipulavam dados, verificaram que muitos dos histogramas que construíam apresentavam a característica seguinte: começavam a crescer gradualmente, até atingirem um ponto máximo, a partir do qual decresciam de forma simétrica.

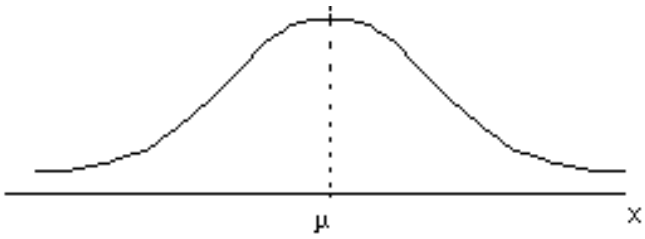


Este aspecto apresentado pelos histogramas, foi o suficiente para desencadear um entusiasmo pela distribuição (População) Normal, com função densidade em forma de sino, a qual se admitia como subjacente aos dados. Chegou-se ao ponto de duvidar de dados, cujos histogramas não tinham aquele comportamento!

Desfeito o mito da distribuição normal, podemos dizer que ela tem ainda hoje um papel importante em estatística, já que muitos dos processos de inferência estatística clássica, têm por base, precisamente a distribuição **Normal**.

Ao falarmos na distribuição **Normal**, estamos na realidade a referir-nos a uma família de distribuições, indexadas pelos parâmetros $\mu$ e $\sigma$. Assim, para cada par de valores destes

parâmetros temos uma distribuição normal, cuja função densidade de probabilidade tem o seguinte aspecto:




Uma v.a. **X** com distribuição **Normal** de parâmetros $\mu$ e $\sigma$ representa-se por

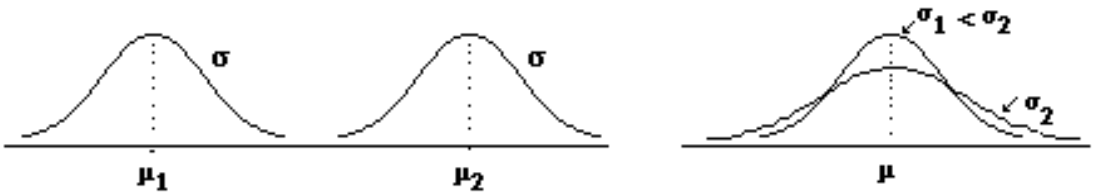
$$X \cap N(\mu,\sigma)$$

Pode-se mostrar que:

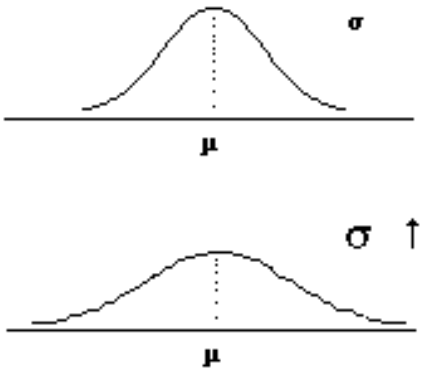
$$E(X) = \mu \qquad e \qquad Var(X) = \sigma^2$$

Vejamos algumas propriedades, relativamente à representação gráfica, da função densidade normal, que se deduzem da sua expressão analítica $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$, $x \varepsilon R$:

- é simétrica relativamente ao seu valor médio $\mu$, de modo que duas curvas correspondentes a duas distribuições com o mesmo desvio padrão têm a mesma forma, diferindo unicamente na localização.
- é tanto mais achatada, quanto maior for o valor de $\sigma$, de modo que duas curvas correspondentes a duas distribuições com o mesmo valor médio, são simétricas, relativamente ao mesmo ponto, diferindo no grau de achatamento.

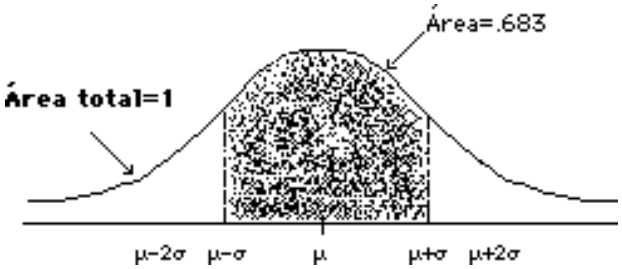


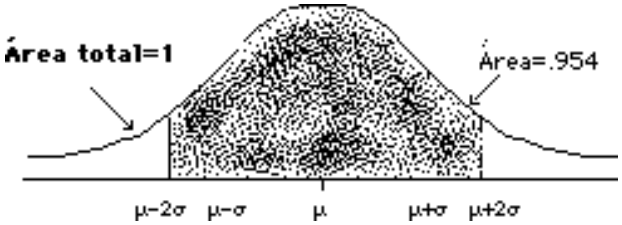

Se deixasse cair um peso em cima da curva da função densidade, ela ficaria mais achatada, o que implicaria um maior desvio padrão!



Para dar uma ideia da **concentração** da distribuição normal, em torno do seu valor médio, apresentamos seguidamente algumas probabilidades:

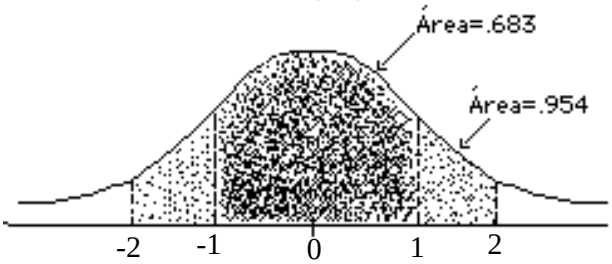
$$P(\mu - \sigma \leq X \leq \mu + \sigma) = .683$$
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = .954$$
$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = .997$$




À distribuição normal que tem valor médio 0 e desvio padrão 1 chamamos distribuição "*standard* " ou *reduzida*, e representamos por

**$Z \cap N(0,1)$**

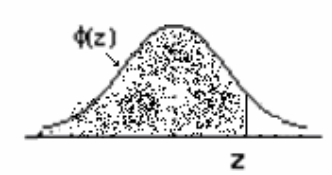



Se a v.a. **X** tiver valor médio $\mu$ e desvio padrão $\sigma$, então a v.a. **Z =** $\frac{X-\mu}{\sigma}$, tem valor médio 0 e desvio padrão 1. Assim

$$X \cap N(\mu,\sigma) \quad \Leftrightarrow \quad Z = \frac{X-\mu}{\sigma} \cap N(0,1)$$

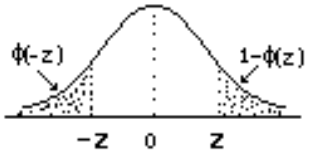
A função distribuição da normal reduzida, tem uma notação especial. Assim, se Z for uma v.a. normal reduzida, representamos

**P( Z ≤ z) = Φ(z)**




**Propriedade**:

Da simetria da curva normal, deduz-se imediatamente a seguinte propriedade:

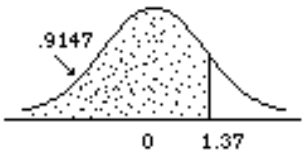



$$\Phi(-z) = 1 - \Phi(z)$$

Existem tabelas extensivas da função distribuição da normal standard, pelo que o cálculo de quaisquer probabilidades, referentes à v.a. Z é imediato (veremos mais adiante a utilização do computador para o cálculo das probabilidades da Normal). A propriedade enunciada anteriormente também permite concluir, que basta haver tabelas para os valores de z≥0 ou de z≤0.

**Exemplo 15** - P(Z≤1.37)

P(Z≤1.37) = Φ(1.37)
= .9147

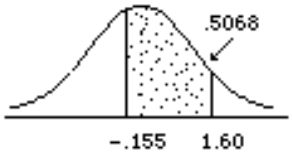



**Exemplo 16** - P(Z>1.37)

P(Z>1.37) = 1 - P(Z≤1.37)
= 1 - .9147
= .0853


**Exemplo 17** - P(-.155 <Z<1.60)

P(-.155 <Z<1.60) = Φ(1.60) - Φ(-.155)
= Φ(1.60)-1+ Φ(.155) (a tabela disponível só tinha os valores positivos(
== .9452 -1 + .5616
== .5068



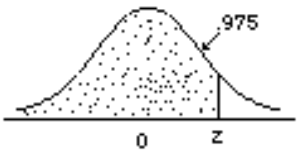
**Exemplo 18**- Determinar o valor de z, tal que P(Z≤z) = .975

Neste caso a consulta da tabela terá de ser feita de maneira inversa.

Temos

$\Phi(z) = .975 \quad \Rightarrow \quad z = \Phi^{-1}(.975)$

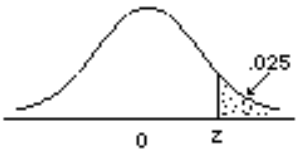
$= 1.96$




**Exemplo 19** - Determinar o valor de z tal que P(Z>z) = .025

$1 - \Phi(z) = .025 \quad \Rightarrow \quad z = \Phi^{-1}(.975)$

$= 1.96$




**Mas se a Normal não tiver valor médio nulo e desvio padrão 1, já não temos tabelas! Como é que vamos calcular as probabilidades?**

Para o cálculo das probabilidades correspondentes a uma distribuição normal de parâmetros $\mu$ e $\sigma$, vamo-nos servir das tabelas da normal reduzida, tendo em atenção a seguinte relação, já apresentada anteriormente:

$$\mathbf{X \cap N(\mu,\sigma)} \quad \Leftrightarrow \quad \mathbf{Z = \frac{X-\mu}{\sigma} \cap N(0,1)}$$

donde:

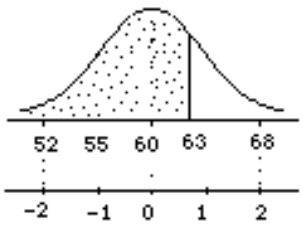
$$P(a \le X \le b) = P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right) \Leftrightarrow P(a \le X \le b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

**Exemplo 20** - Se X∩N(60,4) calcular P(X≤63)

$P(X \le 63) = \Phi\left(\frac{63-60}{4}\right)$

$= \Phi(.75)$

$= .7734$



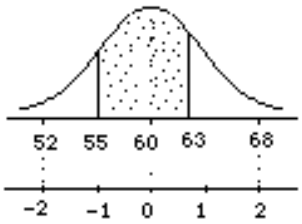

**Exemplo 21** - Se X∩N(60,4) calcular P(55≤X≤63)

$P(55 \le X \le 63) = P\left(\frac{55-60}{4} \le \frac{X-60}{4} \le \frac{63-60}{4}\right)$

$= \Phi(.75). - \Phi(-1.25)$

$= ..7734 - .1056$

$= .6678$



**Exemplo 22 -** Na pastelaria "Gulosa" a quantidade de farinha F utilizada semanalmente, é uma variável aleatória com distribuição normal de valor médio 600kg e desvio padrão 40kg. Havendo no início de determinada semana, um armazenamento de 634kg e não sendo possível receber mais farinha durante a semana:

a) Determine a probabilidade de ruptura do stock de farinha.

b) Qual deveria ser o stock, de modo que a probabilidade de ruptura fosse de .01?

Resolução:

a) Pretende-se calcular a probabilidade de ruptura do stock, isto é, P(F>634), com $F \cap N(600,40)$

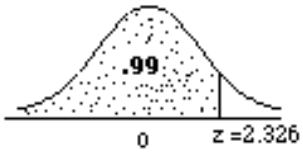
$$P(F > 634)=1 - P(F \leq 634)=1 - P(Z \leq \frac{634-600}{40}) = 1 - \Phi(.85)$$

$$= 1- .8023 =.1977$$

b) $P(F > s)= .01 \Rightarrow 1- \Phi(\frac{s-600}{40})=.01$

$\Phi(\frac{s-600}{40}) = .99 \Rightarrow \frac{s-600}{40} = 2.326$

s = 693kg




Enunciamos seguidamente uma outra propriedade da distribuição normal:

**Propriedade**: A soma de variáveis aleatórias independentes, com distribuição normal, ainda tem distribuição normal:

$$\mathbf{X_i \cap N(\mu_i,\sigma_i)}, \quad i= 1, 2, \ldots , n, \textbf{ independentes}$$

$$\sum_{i=1}^{n} X_i \cap N(\sum_{i=1}^{n} \mu_i, \sqrt{\sum_{i=1}^{n} \sigma_i^2})$$

Obs: A propriedade anterior é um caso particular de uma propriedade mais geral que se pode enunciar da seguinte forma:

Qualquer **combinação linear** de variáveis aleatórias **independentes**, com distribuição **Normal**, ainda tem distribuição **Normal**.

### 8.3.2 - Modelo Uniforme

Uma v.a. **X** diz-se que tem distribuição **Uniforme** no intervalo [a,b], se a sua função densidade de probabilidade for dada por:

$$f(x)=\begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{fora} \end{cases}$$

Da definição de função distribuição a partir de função densidade, obtemos

$$F(x)=\begin{cases} 0 & x<a \\ \dfrac{x-a}{b-a} & a\le x<b \\ 1 & x\ge b \end{cases}$$

Pode-se mostrar que **E(X)** = $\dfrac{(a+b)}{2}$ e **Var(X)**= $\dfrac{(b-a)^2}{12}$.

**Caso particular** - Se U for uma variável aleatória uniforme no intervalo (0, 1), então a sua função distribuição será F(u) = u, com uε(0, 1) (não esquecer que é indiferente o intervalo ser aberto ou fechado, já que para as variáveis contínuas, a probabilidade num ponto é igual a 0), pelo que a probabilidade da v.a.U assumir valores num subintervalo (a, b) de (0, 1) é igual à amplitude desse subintervalo:

$$P(a<U<b) = F(b) - F(a) = b-a$$

**Exemplo 23** – Um gerador de números aleatórios, gera números com distribuição uniforme no intervalo (0, 1). Calcule a probabilidade de um número gerado:

a) Ser menor que 0,5
b) Estar no intervalo (0.15, 0.86)
c) Ser maior que 0.55

Res: a) Seja U∩Uniforme(0,1). Então P(U<0.5) = F(0.5) = 0.5, onde representámos por F, a função distribuição de U.

b) P(0.15<U<0.86) = 0.86 – 0.15 = 0.71

c) P(U>0.55) = 1 – P(U≤0.55) = 1 – 0.55 = 0.45

**Exemplo 24 -** Os automóveis de determinada marca têm averbado no livrete o peso bruto de 1200 kg. Tendo um dos possuidores destes automóveis mandado proceder a algumas modificações, o peso actual varia uniformemente entre 1150kg e 1250kg.

a) Qual a expressão da função densidade da v.a. que representa o peso?

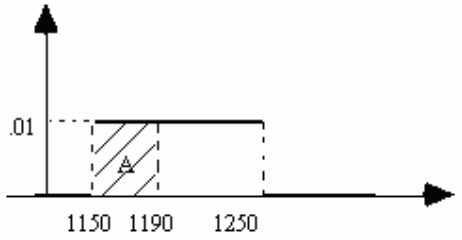
b) Qual a probabilidade de que o carro pese menos de 1190kg?

c)Qual o peso médio dos carros que sofreram as mesmas alterações?

Res:a)

$$f(x)=\begin{cases} .01 & 1150\le x\le 1250 \\ 0 & \text{fora} \end{cases}$$

b)




P(X<1190) = A = 40 × .01 = .4

c) E(X)= 1200 Kg.

**Propriedade – Tranformação uniformizante -** Dada a variável aleatória X, contínua, com função distibuição F, então a variável aleatória U, que se obtém transformando X através de U, U = F(X), tem distribuição uniforme, no intervalo (0, 1). De facto, se U = F(X), então a função distribuição de U, que vamos representar por G, será:

$G(u) = P(U \leq u) = P(F(X) \leq u) = P(X \leq F^{-1}(u))$ (onde representamos por $F^{-1}$, a função inversa de F), donde $G(u) = F(F^{-1}(u)) = u$.

Esta propriedade, conhecida como **transformação uniformizante**, já que transforma qualquer variável aleatória contínua X, cuja função distribuição tenha inversa, numa variável aleatória com distribuição uniforme no intervalo (0, 1), é muito importante, pois permite simular variáveis aleatórias com distribuição F, a partir de uma uniforme, como exemplificaremos mais à frente.

### 8.3.3 - Modelo Exponencial

Diz-se que uma variável aleatória **X** tem distribuição **Exponencial**, com parâmetro $\theta$, se e só se a sua função densidade tiver a forma

$$f(x)=\begin{cases} \frac{1}{\theta}\ e^{-x/\theta}, \theta > 0; & x \geq 0 \\ 0 & x < 0 \end{cases}$$

A função distribuição correspondente tem a forma

$$F(x)=\begin{cases} 0 & x < 0 \\ 1 - e^{-x/\theta} & x \geq 0 \end{cases}$$

Uma v.a. **X** com distribuição Exponencial de parâmetro $\theta$, representa-se por

$$\mathbf{X \cap E(\theta)}$$

Pode-se mostrar que

$$\mathbf{E(X) = \theta \qquad e \qquad Var(X) = \theta^2}$$

O modelo exponencial aplica-se frequentemente quando se pretende estudar o tempo até à falha de componentes electrónicas, em que se admite que o tempo que a componente ainda vai durar, não depende do tempo que já durou. Uma componente com tempo de vida com distribuição exponencial é tão boa nova como velha (Diz-se que não tem memória)!

**Propriedade** – Mostre que $P(X \geq t+h \mid X \geq t) = P(X \geq h)$

**Exemplo 25 -** O tempo de vida, em horas, de certo tipo de componentes electrónicas tem a seguinte função densidade:

$$f(x)=\begin{cases} \dfrac{1}{100} e^{-x/100} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Um aparelho tem três destas componentes, que operam independentemente e o aparelho falha se pelo menos duas das componentes falham. Qual a probabilidade de que o aparelho esteja a funcionar, sem falhas, pelo menos 200 horas?

Resolução:

Representemos por $\underline{\text{Componente}}$ - componente a funcionar

$\overline{\text{componente}}$ - componente a não funcionar

P(aparelho funcionar) = 1 - P(aparelho falhar)

P(aparelho falhar) = 3 P(componente, $\overline{\text{componente}}$, $\overline{\text{componente}}$) +

+ P($\overline{\text{componente}}$, $\overline{\text{componente}}$, $\overline{\text{componente}}$)

P($\overline{\text{componente}}$) = $1-e^{-200/100} = 1- e^{-2} = .865$

P(aparelho falhar) = $3 \times (1-.865) \times .865^2 + .865^3 = .95$

P(aparelho funcionar) = 1 - .95 = .05

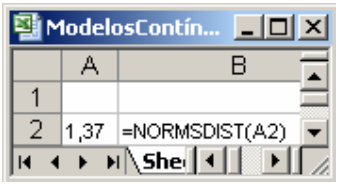
## Utilização do Excel para calcular probabilidades dos modelos contínuos

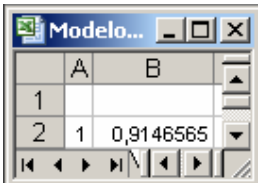
O Excel disipõe de funções que dão as probabilidades dos modelos contínuos considerados anteriormente. Assim, temos:

**Modelo Normal**

Função ***NORMSDIST****(z)*, calcula o valor da função distribuição de uma normal reduzida, no ponto z.

Exemplo – Para calcular a probabilidade pretendida no exemplo 15, basta fazer:

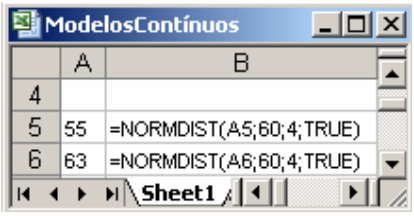


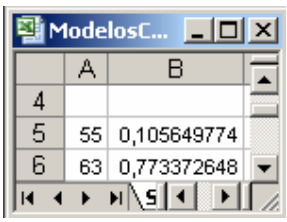




Função ***NORMDIST****(x; mean; standard_dev; cumulative)*, onde:

- x é o valor para o qual pretendemos calcular a distribuição;
- Mean é o valor médio da distribuição;
- Standard_dev é o desvio padrão da distribuição;
- *Cumulative* é um valor lógico: para obter a função distribuição, usar TRUE; para obter a função massa de probabilidade, usar FALSE.

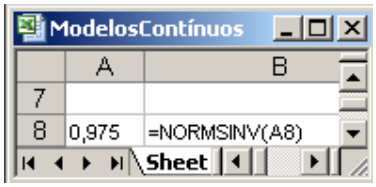
Exemplo – Para calcular as probabilidades pretendidas no exemplo 21, basta fazer:

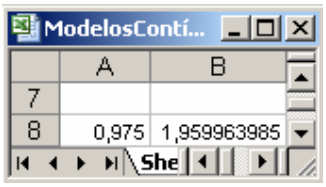



Função ***NORMSINV****(Probability)*, calcula o valor da inversa da distribuição Normal reduzida, para a probabilidade Probability, como se exemplifica a seguir:

Exemplo – Para calcular o valor de z, tal que a função distribuição nesse ponto é 0.975, como se pretende no exemplo 18, basta fazer:

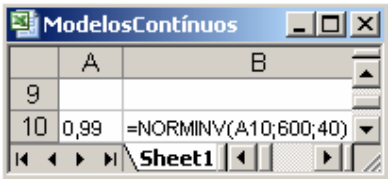


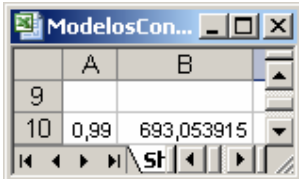




Função ***NORMINV****(probability; mean; standars_dev)*, calcula o valor da inversa da distribuição Normal, onde:

- Probability é o valor da probabilidade;
- Mean é o valor médio da distribuição Normal;
- Standard_dev é o valor do desvio padrão.

Exemplo – No exemplo 22, para determinar o valor de s tal que P(F>s) = 0.01, tem que se formalizar o problema em termos da função distribuição, pelo que a igualdade anterior é equivalente a considerar P(F≤s) = .99. Para calcular o valor de s, basta fazer:



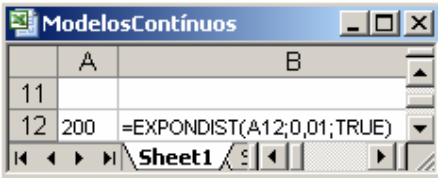



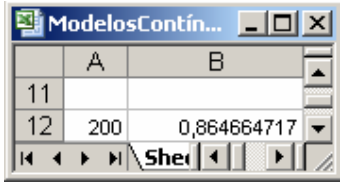

**Modelo Exponencial**

Função EXPONDIST(x; lambda; cumulative), calcula o valor da função distribuição, onde:

- x é o valor onde se pretende calcular a distribuição;
- lambda é o valor do parâmetro (Chamamos a atenção para que no Excel a função distribuição exponencial de parâmetro lambda, apresenta a seguinte expressão: F(x) = 1 – exp(lambda x), pelo que o parâmetro é, no Excel, igual ao inverso do parâmetro da definição que foi dada da distribuição Exponencial);
- *Cumulative* é um valor lógico: para obter a função distribuição, usar TRUE; para obter a função massa de probabilidade, usar FALSE.

Exemplo – No exemplo 25, para calcular o valor da probabilidade da variável aleatória, com distribuição Exponencial de parâmetro 100 ou 0.01 no Excel, ser inferior a 200, basta fazer:




## 8.4 – Compreender a simulação

No capítulo 5, utilizámos o Excel para simular experiências aleatórias. Dissemos na altura, e repetimos agora que, de um modo geral, quando falamos em gerar números aleatórios, estamos a referir-nos à obtenção de qualquer real do intervalo [0, 1], de tal forma que a probabilidade de obter um valor de um subintervalo [a, b] de [0, 1], é igual à amplitude desse subintervalo, ou seja (b-a). No Excel, podemos obter estes números com a função *RAND*. Agora já sabemos que estes números aleatórios não são mais que números com a distribuição **uniforme**, no intervalo [0, 1] (recorde que é indiferente se o intervalo é fechado ou aberto). Além da função *RAND*, podemos

utilizar no Excel, uma componente do *Analysis ToolPaK*, para gerar números pseudo-aleatórios com distribuição uniforme (ou outros tipos de distribuição, de que falaremos mais à frente). Estes números são gerados por algorirmos (deterministas) específicos, que geram cada número a partir do anterior e que começam com um valor a que se chama "semente". Assim, se se utilizar a mesma semente pode-se gerar a mesma sucessão de números. Se os algoritmos que geram os números forem bons, estes comportam-se como se fossem aleatórios, com a distribuição desejada. Assim, temos números que se comportam como se fossem aleatórios, mas que são obtidos por mecanismos deterministas, e daí o chamarem-se **pseudo-aleatórios**.

Suponhamos agora, que se pretendia simular a chegada de chamadas telefónicas a uma central, durante um período de tempo especificado. Estudos anteriores permitem-nos afirmar que a variável aleatória que representa o tempo entre as chamadas sucessivas, pode ser bem modelada por uma distribuição exponencial, com parâmetro igual ao inverso do número médio de chamadas recebidas no período referido, sendo, por sua vez, o número de chamadas recebidas, bem modelado por uma Poisson de parâmetro 2. Admita que, em média, são recebidas 2 chamadas por período. Como é que podemos simular os tempos entre chegadas sucessivas de chamadas à tal central telefónica? Vamos admitir que dispomos de uma série de numeros aleatórios, $u_1$, $u_2$, ..., $u_n$ (com distribuição uniforme no intervalo (0,1)), obtidos através de uma tabela ou gerados em computador. Será que nestas condições conseguimos simular os valores da distribuição exponencial? A resposta é sim, e para o fazer basta lembrar a propriedade que apresentámos como **transformação uniformizante**. De facto:

Se $P(X \leq x) = F(x)$, então $U = F(X)$ tem distribuição uniforme $\Rightarrow$ a tranformada de U por intermédio da inversa de F, tem distribuição F, isto é se $Y = F^{-1}(U)$, $P(Y \leq y) = P(F^{-1}(U) \leq y) = P(X \leq y) = F(y)$. Assim, dado o conjunto de valores aleatórios uniformes, $u_1$, $u_2$, ..., $u_n$, os transformados $x_1=F^{-1}(u_1)$, $x_2=F^{-1}(u_2)$, ..., $x_n=F^{-1}(u_n)$, têm distribuição F. Vejamos o caso da exponencial:

1) Se $X \cap Exp(1/2)$ então $F(x) = 1 - e^{-2x}$, donde $x = -\frac{\ln(1-F(x))}{2}$;

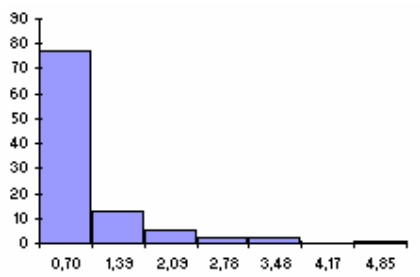
2) Substituindo na expressão anterior F(x) por números aleatórios (uniformes), obtemos números aleatórios exponenciais;

3) Na tabela seguinte apresentamos uma série de números aleatórios com distribuição uniforme e os correspondentes números exponenciais:

| Unif. | Exp(1/2) | Unif. | Exp(1/2) | Unif. | Exp(1/2) | Unif. | Exp(1/2) | Unif. | Exp(1/2) |
|---|---|---|---|---|---|---|---|---|---|
| 0,2349 | 0,1339 | 0,3456 | 0,2120 | 0,4563 | 0,3047 | 0,5670 | 0,4185 | 0,6778 | 0,5662 |
| 0,4315 | 0,2824 | 0,8566 | 0,9710 | 0,2816 | 0,1654 | 0,7067 | 0,6132 | 0,1317 | 0,0706 |
| 0,0747 | 0,0388 | 0,7600 | 0,7135 | 0,4452 | 0,2946 | 0,1305 | 0,0699 | 0,8158 | 0,8458 |
| 0,1346 | 0,0723 | 0,0020 | 0,0010 | 0,8694 | 1,0176 | 0,7367 | 0,6673 | 0,6041 | 0,4633 |
| 0,4196 | 0,2720 | 0,1355 | 0,0728 | 0,8515 | 0,9535 | 0,5674 | 0,4190 | 0,2834 | 0,1666 |
| 0,3714 | 0,2321 | 0,5073 | 0,3539 | 0,6432 | 0,5153 | 0,7791 | 0,7550 | 0,9150 | 1,2326 |
| 0,9939 | 2,5494 | 0,5216 | 0,3687 | 0,0493 | 0,0253 | 0,5771 | 0,4303 | 0,1048 | 0,0554 |
| 0,4986 | 0,3452 | 0,2492 | 0,1433 | 0,9999 | 4,8520 | 0,7506 | 0,6943 | 0,5012 | 0,3478 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0,9498 | 1,4955 | 0,7757 | 0,7474 | 0,6017 | 0,4602 | 0,4276 | 0,2790 | 0,2535 | 0,1462 |
| 0,9556 | 1,5576 | 0,5052 | 0,3518 | 0,0548 | 0,0282 | 0,6044 | 0,4636 | 0,1539 | 0,0836 |
| 0,5939 | 0,4506 | 0,5791 | 0,4327 | 0,5644 | 0,4155 | 0,5496 | 0,3989 | 0,5349 | 0,3827 |
| 0,0097 | 0,0049 | 0,5031 | 0,3497 | 0,9966 | 2,8349 | 0,4899 | 0,3366 | 0,9833 | 2,0473 |
| 0,3752 | 0,2352 | 0,9944 | 2,5966 | 0,6137 | 0,4755 | 0,2329 | 0,1325 | 0,8521 | 0,9556 |
| 0,1252 | 0,0669 | 0,1289 | 0,0690 | 0,1326 | 0,0711 | 0,1362 | 0,0732 | 0,1399 | 0,0754 |
| 0,7805 | 0,7582 | 0,2064 | 0,1156 | 0,6323 | 0,5003 | 0,0582 | 0,0300 | 0,4841 | 0,3310 |
| 0,2642 | 0,1534 | 0,1126 | 0,0598 | 0,9611 | 1,6240 | 0,8096 | 0,8294 | 0,6581 | 0,5366 |
| 0,4683 | 0,3159 | 0,7332 | 0,6606 | 0,9980 | 3,1191 | 0,2629 | 0,1525 | 0,5277 | 0,3751 |
| 0,3797 | 0,2388 | 0,0875 | 0,0458 | 0,7953 | 0,7930 | 0,5030 | 0,3496 | 0,2107 | 0,1183 |
| 0,1045 | 0,0552 | 0,1751 | 0,0963 | 0,2457 | 0,1410 | 0,3163 | 0,1901 | 0,3869 | 0,2446 |
| 0,9609 | 1,6213 | 0,5656 | 0,4169 | 0,1702 | 0,0933 | 0,7748 | 0,7455 | 0,3795 | 0,2386 |

Construindo um histograma dos números exponenciais, obtemos um gráfico com o seguinte aspecto,




que sugere o modelo Exponencial, como esperávamos. Não esqueça que o histograma é a imagem estatística da função densidade de probabilidade. O passo seguinte seria testar a adequabilidade do modelo proposto, o que seria feito com instrumentos disponíveis na inferência estatística, mas que saem fora do âmbito deste curso.

## Utilização do Excel para gerar números pseudo-aleatórios com determinadas distribuições

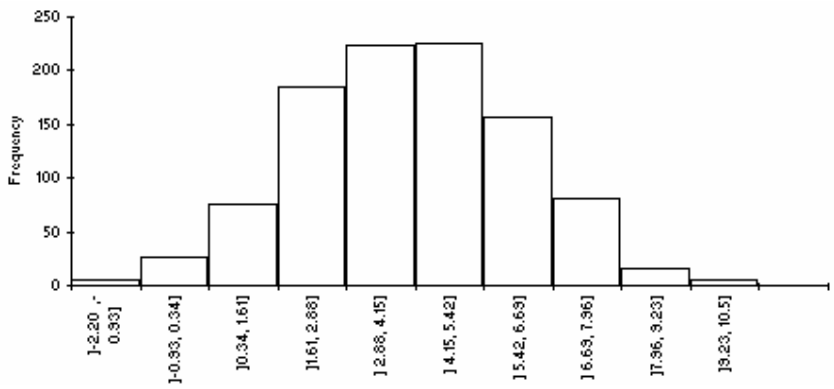
O Excel dispõe de uma componente no *Analysis ToolPaK* que permite obter números pseudo-aleatórios, que se comportam como números aleatórios, com determinadas distribuições. Estão disponíveis os modelos Uniforme, Normal, Bernoulli, Binomial, Poisson ou modelos discretos com determinada função massa de probabilidade. Para obter o gerador destes números seleccione:

*Tools →Data Analysis→Random Number Generation→OK.*

Obtém uma janela, onde deve seleccionar o número de amostras -*Number of Variables* e a dimensão dessas amostra – *Number of Random Numbers*. Tem ainda possibilidade de escolher um valor para a semente – *Random Seed*, se pretende reproduzir o conjunto de números a gerar. Selecciona ainda a Distribuição a gerar, em *Distribution* e de acordo com a distribuição seleccionada, assim terá de introduzir alguns valores para os parâmetros, em *Parameters*.

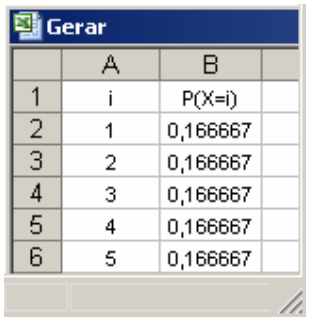
**Exemplo** – Utilize o gerador de números aleatórios do Excel para gerar um conjunto de números com a distribuição Normal (4, 2).

Em *Distribution* seleccionámos o modelo Normal e em *Parameters* escolhemos 4 para *Mean* e 2 para *Standard Deviation*. Gerámos uma única amostra (Escrevendo 1 em *Number of Variables*) de dimensão 1000 (Escrevendo 1000 em *Number of Random Numbers*) e apresentamos a seguir o histograma correspondente



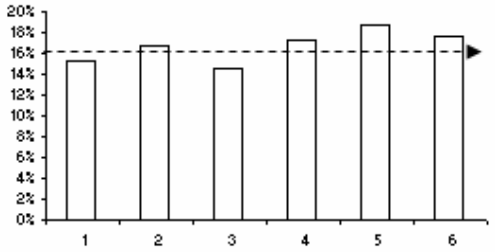

**Exemplo** – Utilize o gerador de números aleatórios do Excel para simular o lançamento de um dado equilibrado.

Para simular esta experiência construímos uma função massa de probabilidade de uma distribuição uniforme discreta em 6 pontos, que foi utilizada pelo *Random Number Generation* para simular a experiência pretendida:



Gerar

| | A | B |
|---|---|---|
| 1 | i | P(X=i) |
| 2 | 1 | 0,166667 |
| 3 | 2 | 0,166667 |
| 4 | 3 | 0,166667 |
| 5 | 4 | 0,166667 |
| 6 | 5 | 0,166667 |



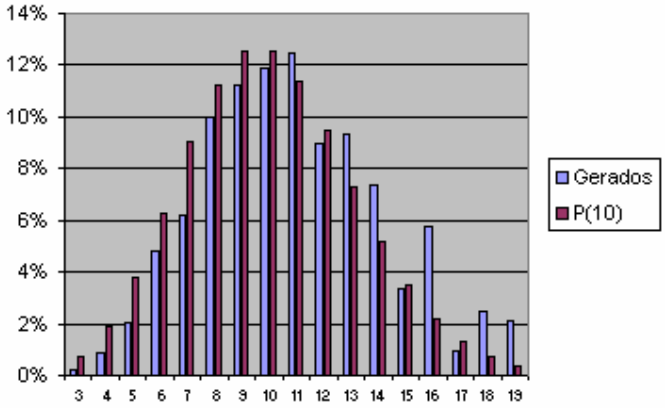

O processo é idêntico ao do exemplo anterior, com a diferença que, em Distribution, seleccionámos *Discrete*.Gerámos 500 números e construimos o diagrama de barras da amostra obtida. Não obtivémos uma distribuição perfeitamente uniforme, como se vê pela figura.

**Nota** – Alguns testes têm revelado algumas deficiências no gerador de números aleatórios do Excel. Assim, para trabalhos de responsabilidade, recomenda-se a utilização de outros processos.

**Exemplo** – Utilize o gerador de números aleatórios do Excel para simular a experiência que consiste em verificar o número de ambulâncias que chegam a determinado hospital, durante um ano, sabendo que em média, por dia, chegam 10 ambulâncias.

Vamos modelar a chegada das ambulâncias, por dia, por um modelo de Poisson, de parâmetro 10. Assim, não temos mais que gerar uma amostra de dimensão 365, de uma Poisson, com aquele valor para o parâmetro:

| Valor | Total |
|---|---|
| 3 | 0,25% |
| 4 | 0,88% |
| 5 | 2,07% |
| 6 | 4,81% |
| 7 | 6,19% |
| 8 | 9,96% |
| 9 | 11,20% |
| 10 | 11,89% |
| 11 | 12,47% |
| 12 | 8,96% |
| 13 | 9,35% |
| 14 | 7,36% |
| 15 | 3,32% |
| 16 | 5,75% |
| 17 | 0,94% |
| 18 | 2,49% |
| 19 | 2,10% |
| Total | 100,00% |



Obtivémos uma amostra cuja tabela de frequências e diagrama de barras se apresenta em cima. Para comparação considerámos também a função massa de probabilidade de uma Poisson, de parâmetro 10, que representámos por P(10).

**Exercícios**

**1**. Num grupo de 40 cães, 20 ladram, 14 não ladram e mordem e 26 mordem.

a) Calcule a probabilidade de ser verdadeira a seguinte frase: "Cão que ladra não morde".

b) Suponha que passa diariamente junto da matilha anterior e selecciona um dos cães aleatoriamente, para fazer festas. Ao fim de uma semana, qual a probabilidade de nunca ser mordido.

c) Num dia em que passam 50 pessoas, em que cada uma selecciona aleatoriamente um dos cães para fazer festas, qual a probabilidade de no máximo serem mordidas 10 pessoas.

**2** - Uma fonte radioactiva é observada durante 4 intervalos de tempo disjuntos, de 6 segundos cada um, tendo-se registado o nº de partículas emitidas em cada intervalo. Admitindo-se que o nº de partículas emitidas segue uma lei de Poisson, em que o nº médio (taxa) de partículas emitidas por segundo é .5, determine a probabilidade de:

a) Em cada um dos 4 intervalos de tempo sejam emitidas 3 ou mais partículas.

b) Em pelo menos um dos 4 intervalos de tempo, sejam emitidas 3 ou mais partículas.

**3 -** O Glorioso tem no seu avançado Marquinhos o seu maior trunfo. É tal a influência deste jogador, que o número de golos marcados pelo Glorioso num jogo em que ele alinha é uma Poisson de valor médio 3, sendo uma Poisson de valor médio 2, quando ele não alinha. Marquinhos é intempestivo, estando várias vezes sujeito a castigos, para além das naturais lesões, pelo que joga apenas 60% dos jogos de uma temporada. Admita que os jogos são independentes.

a) O Glorioso marcou três golos num jogo. Calcule a probabilidade de Marquinhos ter jogado.

b) No final da época, Marquinhos foi vendido. Calcule o número esperado de golos, a marcar pelo Glorioso na próxima época, sabendo que esta é constituída por 30 jogos.

c) Calcule ainda a probabilidade, de na mesma época considerada na alínea anterior, o Glorioso marcar pelo menos 80 golos.

**4** - Pretende-se estudar a incidência de doença pulmonar, numa população em que existem três vezes mais indivíduos não fumadores do que fumadores. Sabe-se que a percentagem de doentes entre os fumadores e os não fumadores é respectivamente de 60% e 20%.

a) Determine a probabilidade de um indivíduo ter doença pulmonar.

b) Determine a probabilidade de um doente pulmonar ser fumador.

c) Qual a probabilidade de numa amostra de 10 doentes, pelo menos 3 serem fumadores?

**5**. Na mercearia da D. Ana vendem-se maçãs vermelhas e amarelas. 70% dessas maçãs são vermelhas. Como as maçãs não são tratadas quimicamente, 20% das maçãs amarelas estão bichosas, enquanto que das vermelhas só 5% é que têm bicho.

a) Determine a probabilidade de uma maçã escolhida ao acaso ser sã.

b) A D. Ana resolveu comer uma maçã. Trincou e … tinha bicho. Qual a probabilidade de se tratar de uma maçã vermelha?

c) Um cliente comprou uma dúzia de maçãs amarelas. Qual a probabilidade de encontrar, no máximo, 4 maçãs bichosas?

**6** - O I.N.I.P. está a proceder a um estudo sobre a sensibilidade ao envenenamento por mercúrio de duas variedades de lagostim. Para o efeito foram recolhidas amostras de dimensão variável, mas contendo um nº apreciável de espécimes das variedades A e B. Sejam $X_1$ e $X_2$ as v.a. que representam o nº de lagostins das variedades A e B, respectivamente, em cada amostra recolhida. Admite-se que $X_1$ e $X_2$ têm distribuição de Poisson de parâmetros $\lambda_1$=8 e $\lambda_2$=12.

Numa das amostras foi observado um total de 15 lagostins das variedades A e B. Qual a probabilidade de que 10 desses lagostins sejam da variedade A?

**7** - A produção diária de determinado artigo, segue uma distribuição Normal com valor médio igual a 185 unidades e desvio padrão igual a 4.5 unidades.

a) Determine a probabilidade da produção diária ser inferior a 190 unidades.

b) Determine a probabilidade da produção diária estar compreendida entre 160 e 190 unidades.

c) O fabricante afirma que 80% das vezes a produção diária é superior a P. Qual é o valor de P?

**8** - As quantidades de margarina (medidas em nº de pacotes de 500g), vendidas por semana em 3 supermercados, Pão Doce, Pinga Pouco e Paga Açúcar, podem ser consideradas v.a. independentes e com distribuições N(551,33), N(250,28) e N(831,42) respectivamente. Determine as probabilidades dos seguintes acontecimentos:

a) O Pão Doce vende numa semana entre 250 e 570.

b) O nº total de vendas numa semana nos 3 supermercados, excede 1800.

c) Numa semana, o nº total de vendas do Pão Doce e do Pinga Pouco excede o nº de vendas do Paga Açúcar.

**9** - Verificou-se que o tempo médio entre acidentes de avião tem uma distribuição exponencial com valor médio de 44 dias. Se ocorreu um acidente no dia 1 de Julho, qual a probabilidade de nesse mês se verificar novo acidente?

**10** - Estima-se em 5% a percentagem de pessoas com mais de 60 anos que sofrem da doença de Paget. Sabe-se que uma medição efectuada por análise ao sangue tem distribuição normal, tanto nos doentes como nos indivíduos que não sofrem desta doença, apresentando valores anormalmente elevados entre os doentes. Um valor superior a 14 (resultado positivo da análise) é motivo para que o indivíduo seja posteriormente submetido a exames, de modo a que seja possível um diagnóstico rigoroso. Apenas são sujeitos à análise os indivíduos com mais de 60 anos.

Sejam $X_1$ e $X_2$ as variáveis aleatórias que representam o valor da referida medição num indivíduo doente e num indivíduo saudável, respectivamente, sendo os valores dos parâmetros os seguintes: $X_1$ $X_2$

| | | |
|---|---|---|
| Valor Médio | 16.4 | 7.2 |
| Desvio Padrão | 6.0 | 3.0 |

a) Qual a probabilidade de um indivíduo saudável ser submetido a exames por ter apresentado um resultado positivo na análise?

b) Calcule a probabilidade de um indivíduo doente não ser diagnosticado.

c) Numa população em que todas as pessoas com mais de 60 anos foram sujeitas a análise ao sangue, que percentagem apresenta resultados positivos?

d) Determine a probabilidade de uma dessas pessoas, que teve resultado positivo na análise, e que será portanto submetida a exames sofra realmente da doença de Paget ?

**11**. Dadas as v.a. $X_1$ e $X_2$ independentes com distribuição de Poisson de parâmetros $\lambda_1$ e $\lambda_2$, respectivamente, mostre que $X_1$ dado $X_1+X_2$ é Binomial, isto é

$P(X_1=k|X_1+X_2=n)=B(n, \frac{\lambda_1}{\lambda_1+\lambda_2})$

**12**. Mostre que se X e Y forem variáveis aleatórias independentes com distribuição Binomial de parâmetros (m,p) e (n, p) respectivamente, então a distribuição condicional de X, dado que X+Y=k, é

$$P(X=i|X+Y=k) = \frac{\binom{n}{i}\binom{m}{k-i}}{\binom{n+m}{k}}$$

**13**. ***Preservação da Poisson, perante uma selecção aleatória***. Suponha que numa situação adequadamente descrita pelo modelo de Poisson, nem todos os acontecimentos são contabilizados, isto é, cada acontecimento pode ser ou não contabilizado e a probabilidade de o ser é **p**. Será que o número de acontecimentos contabilizados ainda segue um modelo de Poisson? A questão anterior pode ser equacionada da seguinte maneira: Se X tem distribuição de Poisson de parâmetro $\lambda$ e a distribuição condicional de Y dado X=n, é Binomial de parâmetros n e p, então Y tem distribuição de Poisson de parâmetro $\lambda p$.

Demonstração: $P(Y=k) = \sum_{n=k}^{\infty} P(Y=k \mid X=n)P(X=n)$

$$= \sum_{n=k}^{\infty}\binom{n}{k}p^k(1-p)^{n-k}e^{-\lambda}\frac{\lambda^n}{n!}$$

$$= e^{-\lambda}\frac{(\lambda p)^k}{k!}\sum_{n=k}^{\infty}\frac{1}{k!(n-k)!}\left[\lambda(1-p)\right]^{n-k}$$

$$= e^{-\lambda p}\frac{(\lambda p)^k}{k!}$$

**Capítulo 9**

# Distribuições de amostragem

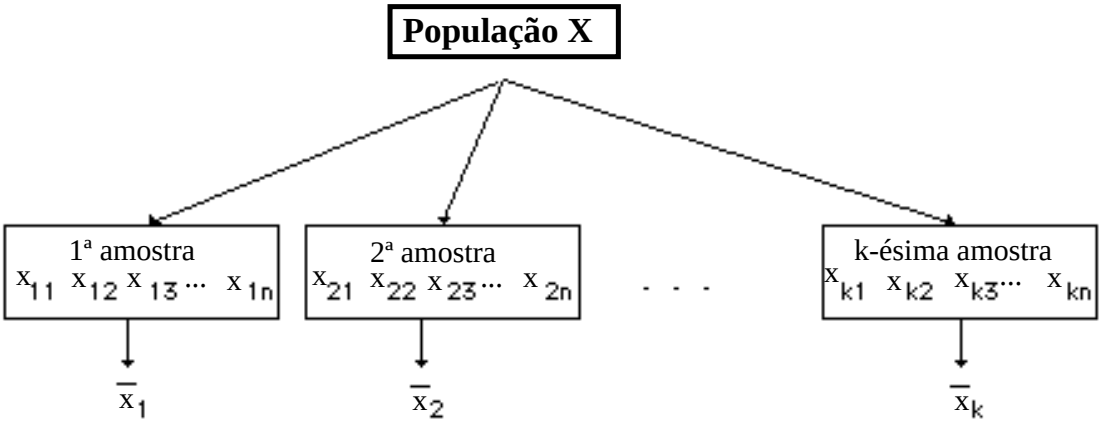
## 9.1 - Introdução

Já vimos no módulo características amostrais, que podemos resumir a informação contida na amostra, utilizando as estatísticas, isto é, funções unicamente dos dados amostrais.

Mas, do mesmo modo que duas amostras extraídas da mesma população, apresentam uma certa variabilidade, também as estatísticas calculadas com amostras diferentes, apresentam variabilidade.

Por exemplo, dada a população **X**, constituída pelas alturas dos alunos inscritos na cadeira de IPE, no ano lectivo de 98/99, se recolhermos uma amostra das alturas de 10 estudantes, a estatística **média** apresentará um certo valor, por exemplo, 1.64m. No entanto, se recolhermos outra amostra, da mesma população, é natural esperar que a média para esta nova amostra seja diferente daquele valor, embora não se afaste muito!

Generalizando o exemplo anterior, podemos considerar o seguinte esquema, se tivermos k amostras de dimensão n, recolhidas da População X:




Relativamente às amostras anteriores, podemos considerar o seguinte:

$x_{11}, x_{21}, \ldots, x_{k1}$ são os valores observados de uma v.a. com distribuição idêntica à de **X**, mas que representamos por $\mathbf{X_1}$, para significar que foi o 1º elemento recolhido nas diferentes amostras;

$x_{12}, x_{22}, \ldots, x_{k2}$ são os valores observados de uma v.a. com distribuição idêntica à de **X**, independente de $\mathbf{X_1}$ (numa amostra aleatória, os valores não podem depender uns dos outros), mas que representamos por $\mathbf{X_2}$, para significar que corresponde ao 2º elemento recolhido;

$x_{1n}, x_{2n}, \ldots, x_{kn}$ são os valores observados de uma v.a. com distribuição idêntica à de **X**, independente de $X_1, X_2, \ldots$ que representamos por $\mathbf{X_n}$, para significar que foi o enésimo elemento a ser recolhido.

Com esta notação, as amostras $(x_{11}, x_{12}, x_{13}, \ldots, x_{1n})$, $(x_{21}, x_{22}, x_{23}, \ldots, x_{2n})$ ..., $(x_{k1}, x_{k2}, x_{k3}, \ldots, x_{kn})$ são **amostras observadas** da **amostra aleatória**

$$(\mathbf{X_1, X_2, \ldots, X_n})$$

Admitindo que a população **X,** que estávamos a estudar, constituída pelas alturas (em cm) dos alunos inscritos na cadeira de IPE, no ano lectivo de 89/90, era tal que $\mathbf{X \cap N(165,10)}$, podemos obter várias amostras observadas, de dimensão 10:

(158, 163, 171, 150, 149, 167, 158, 172, 149, 150)

(167, 149, 168, 153, 162, 160, 170, 161, 160, 149)

… … … … … … … …

(170, 160, 158, 168, 165, 159, 163, 159, 172, 150)

da amostra aleatória $(\mathbf{X_1, X_2, \ldots, X_{10}})$, em que todas as v.a. $\mathbf{X_i}$, i=1,…,10, são independentes e têm distribuição Normal de valor médio 165 e desvio padrão 10.

Tendo em consideração o que foi dito anteriormente, podemos afirmar que $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$, são valores observados da variável aleatória

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

onde $\mathbf{X_1, X_2, \ldots, X_n}$ são variáveis aleatórias independentes e com a mesma distribuição que uma variável aleatória **X** ( população em estudo). Sendo a estatística uma variável aleatória tem uma distribuição de probabilidades, a que damos o nome de distribuição de amostragem.

**Distribuição de amostragem** - Distribuição de amostragem de uma estatística é a distribuição dos valores que a estatística assume para todas as possíveis amostras, da mesma dimensão, da população.

São as distribuições de amostragem das *estatísticas* que nos vão permitir fazer inferências sobre os *parâmetros* populacionais correspondentes. Ao aleatoriezarmos o processo de selecção das amostras, faz com que se possa utilizar a distribuição de amostragem de uma estatística para descrever o comportamento dessa estatística, quando se utiliza para estimar um determinado parâmetro. Por outro lado, para podermos utilizar os resultados da Teoria das Probabilidades, o processo de amostragem que se considera é o de amostragem com reposição. Esta observação é relevante, sobretudo para populações de dimensão pequena, em que a composição da população, relativamente à característica de interesse, se altera quando se retiram alguns elementos; esta situação não se verifica com populações de grande dimensão, que é normalmente a situação de interesse em Estatística.

Assim, se uma população tiver N elementos, para obter as distribuições de amostragem de estatísticas, a partir de amostras de dimensão n, teríamos de seleccionar $N^n$ amostras distintas. Então, para calcular a distribuição de amostragem da média, será necessário considerar todas as amostras possíveis e calcular as respectivas médias? Felizmente não é necessário estar com tanto trabalho, graças a um dos resultados mais importantes da Teoria das Probabilidades, conhecido como Teorema Limite Central, que nos fornece um modelo matemático para a distribuição de amostragem da média, como veremos a seguir.

## 9.2 - Distribuição de amostragem da média

Vamos começar por estudar a distribuição de amostragem da variável aleatória média, considerada anteriormente. Veremos como esta distribuição de amostragem nos vai permitir fazer inferências sobre o valor médio da população de onde foi retirada a amostra que serviu para calcular a média.

Algumas questões que se podem levantar acerca da distribuição de amostragem da estatística média, são as seguintes:

- *A distribuição da média, depende da distribuição da população X, subjacente às amostras?*
- *Será sempre possível conhecer essa distribuição?*

No que se segue procuraremos responder a estas questões, adiantando desde já que, na verdade, a <u>distribuição de amostragem da média depende da distribuição da população subjacente</u> às amostras. Veremos também, que <u>nem sempre é possível obter a distribuição exacta da média</u>, mas sim uma distribuição aproximada.

### 9.2.1 - Valor médio e desvio padrão da média

Dada uma população **X** de valor médio $\mu$ e desvio padrão $\sigma$, então, tendo em consideração as propriedades do valor médio e da variância, pode-se mostrar facilmente que

$$E(\overline{X}) = \mu \qquad \text{e} \qquad Var(\overline{X}) = \frac{\sigma^2}{n}$$

Obs: Não esquecer que $\overline{X}$ é uma combinação linear de variáveis aleatórias independentes e com a mesma distribuição.

Chamamos a atenção para o facto do valor médio da v.a *estimador* $\overline{X}$, coincidir com o *parâmetro* que estamos a estimar, o valor médio, **μ**, da população. Dizemos que o estimador é centrado ou **não enviesado**. Além disso, a variância do estimador decresce com a dimensão da amostra, o que permite concluir que, à medida que aumentamos a dimensão da amostra a variabilidade do estimador, em torno do parâmetro, diminui. Diz-se então que o estimador é **consistente**. Estas propriedades de não enviesamento e de consistência fazem com que a média seja um "bom" estimador do valor médio.

### 9.2.2 - Distribuição da média, para populações Normais

Para calcular a distribuição de $\overline{X}$, vamos distinguir o caso de a população **X** ser Normal e não Normal, distinguindo ainda se o desvio padrão **σ** é conhecido ou não.

#### 9.2.2.1 – Desvio padrão σ conhecido

Já dissemos quando estudamos a distribuição Normal, que qualquer combinação linear de variáveis aleatórias independentes, com distribuição Normal, ainda tem distribuição Normal. Como a média é uma combinação linear de variáveis aleatórias $X_i$, independentes, com distribuição idêntica à de **X**, que por hipótese é **Normal(μ,σ),** vem imediatamente que $\overline{X}$ tem distribuição Normal, com valor médio μ e desvio padrão $\frac{\sigma}{\sqrt{n}}$, pelo que procedendo à standardização, se obtém o seguinte resultado

**Populações Normais, σ conhecido**

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \cap N(0,1)$$

#### 9.2.2.2 – Desvio padrão σ desconhecido

Quando o parâmetro σ é desconhecido, situação que ocorre com frequência, já o resultado anterior não é válido. Assim, estima-se o desvio padrão desconhecido pelo desvio padrão empírico, S, em que

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

e tem-se o seguinte resultado

**Populações Normais, σ desconhecido**

$$T = \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \cap t(n-1)$$

o que significa que se conhece a distribuição exacta da variável aleatória T, que é a chamada distribuição **t-Student** (t de Student), com (n-1) graus de liberdade e que se representa por **t(n-1).**

Este modelo tem uma função densidade semelhante à da Normal, mas com as caudas mais altas, isto é, não é tão concentrada. No entanto, à medida que o número de graus de liberdade aumenta (isto é, à medida que n aumenta), a t-Student confunde-se com a Normal. Do mesmo modo que a Normal, também a distribuição t-Student se encontra tabelada.

### 9.2.3 - Distribuição da média, para populações não normais. Teorema Limite Central

Quando a distribuição da população **X** já não é Normal, a distribuição de amostragem da média dependerá da distribuição de X, não sendo em geral conhecida. No entanto, um dos teoremas fundamentais das probabilidades, dá-nos uma indicação do comportamento da distribuição da média de um número suficientemente grande de variáveis aleatórias independentes e identicamente distribuídas:

**Teorema limite central**

Se $\mathbf{X_1}$, $\mathbf{X_2}$, …, $\mathbf{X_n}$ são variáveis aleatórias independentes e identicamente distribuídas a uma variável aleatória X com valor médio **μ** e variância $\sigma^2$, finita, então a distribuição da soma $\mathbf{S_n} = X_1 + X_2 + \ldots + X_n$, ou da média $\overline{X} = \frac{S_n}{n}$ tende a aproximar-se da distribuição Normal, para **n suficientemente grande**

$$P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le z\right) \approx \Phi(z) \quad \text{e} \qquad P\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \le z\right) \approx \Phi(z)$$

O teorema limite central, dá-nos uma justificação teórica para a grande utilização da distribuição Normal, como modelo de fenómenos aleatórios. Quantidades tais como alturas e pesos de uma população relativamente homogénea, podem ser consideradas como somas de um grande número de causas genéticas e efeitos devido ao meio ambiente, mais ou menos independentes entre si, cada um contribuindo com uma pequena quantidade para a soma.

**O que é que se entende por um valor de n suficientemente grande?**

Uma questão que se pode pôr é a seguinte: quando queremos aplicar o teorema do limite central: *qual o valor de n, para que se possa considerar que temos uma boa aproximação para a distribuição Normal*?

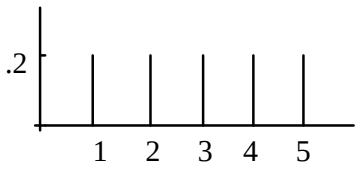
Este valor de n, depende da distribuição subjacente à amostra e será tanto maior quanto mais enviesada for a distribuição da população (o termo enviesado aplica-se como contrário a simétrico). No entanto, de uma maneira geral a aproximação é "rápida", como sugere o exemplo que se segue.

**Exemplificação do Teorema Limite Central:**

Cinco equipas de futebol, resolveram organizar um torneio entre si. As equipas eram todas muito equilibradas, de modo que a classificação final pelo 1º, 2º, ..., 5º lugares pode ser considerada perfeitamente aleatória. O torneio correu tão bem, que resolveram repeti-lo, continuando as equipas equilibradas entre si.
Pretende-se estudar a distribuição da **média** dos pontos obtidos, nos dois torneios, por uma qualquer das equipas, escolhida ao acaso. Considera-se que uma equipa que ficou em 1º lugar tem 1 ponto, em 2º lugar 2 pontos, etc.

Seja **X** a v.a. que representa a pontuação obtida por uma equipa, escolhida ao acaso. Como as equipas são equilibradas, a probabilidade de cada uma se classificar em qualquer dos lugares é igual a 1/5, pelo que a f.m.p. da v.a. X é

| $X$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $p_i$ | .2 | .2 | .2 | .2 | .2 |




Considerando os dois torneios, o espaço dos resultados possíveis é constituído por todos os pares (i,j), com i,j=1, 2,..., 5, em que a pontuação i se refere ao 1º torneio, enquanto a pontuação j se refere ao 2º.

Em termos de variáveis aleatórias podemos considerar duas v.a. $X_1$ e $X_2$, para identificar respectivamente o 1º e o 2º elemento do par (i,j), sendo as v.a. $X_i$, i=1,2, independentes e identicamente distribuídas a X. Com esta notação

$$\overline{\mathbf{X}} = \frac{\mathbf{X}_1 + \mathbf{X}_2}{\mathbf{2}}$$

Vejamos quais os valores que a v.a. $\overline{\mathbf{X}}$ assume:

| $(X_1, X_2)$ | (1,1) | (1,2) | (1,3) | (1,4) | (1,5) | (2,5) | (3,5) | (4,5) | (5,5) |
|---|---|---|---|---|---|---|---|---|---|
| | | (2,1) | (2,2) | (2,3) | (2,3) | (3,4) | (4,4) | (5,4) | |
| | | | (3,1) | (3,2) | (3,3) | (4,3) | (5,3) | | |
| | | | | (4,1) | (4,2) | (5,2) | | | |
| | | | | | (5,1) | | | | |
| $\overline{\mathbf{X}} \rightarrow$ | **1** | **1.5** | **2** | **2.5** | **3** | **3.5** | **4** | **4.5** | **5** |

Observação: Os pares considerados anteriormente correspondem a todas as amostras possíveis de dimensão 2, com reposição, extraídas de uma população que pode assumir os valores 1, 2, 3, 4 ou 5 com igual probabilidade. Estamos numa situação simples em que é fácil considerar todas as amostras possíveis, porque são em número de $5^2$=25.

Tendo em atenção os resultados anteriores, vem imediatamente para a função massa de probabilidade de $\overline{\mathbf{X}}$ a seguinte função:

| $\overline{\mathbf{X}}$ | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 | 4.5 | 5 |
|---|---|---|---|---|---|---|---|---|---|
| $p_i$ | 1/25 | 2/25 | 3/25 | 4/25 | 5/25 | 4/25 | 3/25 | 2/25 | 1/25 |

Obs: Para calcular a probabilidade de $\overline{\mathbf{X}}$ ser igual a 1, temos em consideração que dos 25 pares possíveis, só um dos pares é que conduz a que a média seja igual a 1. Estamos assim a aplicar a definição clássica de probabilidade!

A f.m.p. tem a seguinte representação gráfica:


Se em vez de considerarmos dois torneios, isto é n=2, considerarmos três torneios, portanto agora n=3, teremos de considerar uma nova v.a. $X_3$, independente das anteriores, mas com distribuição idêntica, a qual vai representar a pontuação obtida no 3º torneio.

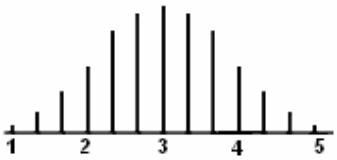
Então

$$\overline{\mathbf{X}} = \frac{\mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3}{3}$$

O processo seguido para calcular a distribuição da média, agora com n=3, é análogo ao que foi feito com n=2 ( embora um pouco mais trabalhoso, pois temos de considerar todas as possibilidades para o terno (i,j,k) com i, j, k=1,…,5), em número de 125, obtendo-se a f.m.p.:

| $\overline{\mathbf{X}}$ | 1 | 1.33 | 1.67 | 2 | 2.33 | 2.67 | 3 | 3.33 | 3.67 | 4 | 4.33 | 4.67 | 5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p_i$ | .008 | .024 | .048 | .08 | .12 | .144 | .152 | .144 | .12 | .08 | .048 | .024 | .008 |

com a seguinte representação gráfica.




Considerando finalmente o caso em que n=5, isto é, em que se consideram 5 torneios, obtemos para a distribuição de amostragem da v.a. $\overline{\mathbf{X}}$

$$\overline{\mathbf{X}} = \frac{\mathbf{X}_1 + \mathbf{X}_2 + \ldots + \mathbf{X}_5}{5}$$

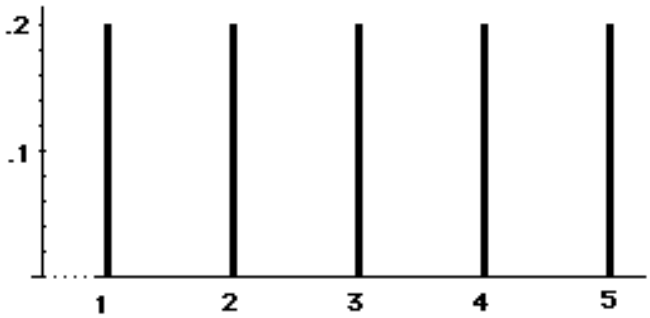
a seguinte f.m.p.:

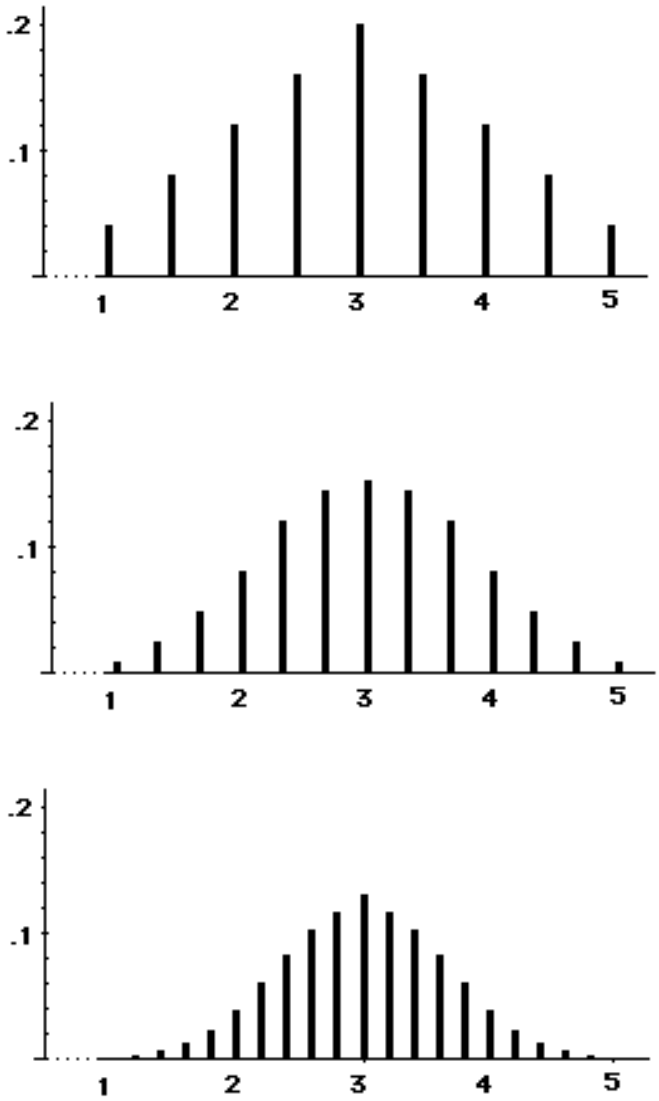
| $\overline{\mathbf{X}}$ | $p_i$ | $\overline{\mathbf{X}}$ | $p_i$ |
|---|---|---|---|
| 1 | 0 | 3.2 | .1168 |
| 1.2 | .0020 | 3.4 | .1024 |
| 1.4 | .0048 | 3.6 | .0813 |
| 1.6 | .0112 | 3.8 | .0592 |
| 1.8 | .0224 | 4 | .0387 |
| 2 | .0387 | 4.2 | .0224 |
| 2.2 | .0592 | 4.4 | .0112 |
| 2.4 | .0813 | 4.6 | .0048 |
| 2.6 | .1024 | 4.8 | .0020 |
| 2.8 | .1168 | 5 | 0 |
| 3 | .1226 | | |

cuja representação gráfica se apresenta a seguir


Observação: O número de amostras de dimensão 5 que considerámos para obter a distribuição anterior foi de $5^5 = 3125$, o que já começa a ser complicado!

Podemos visualizar melhor o processo anterior considerando as f.m.p. todas seguidas:






O processo descrito anteriormente serve para chamar a atenção de que a aproximação da distribuição da média pela distribuição Normal se faz, mesmo que o nº de parcelas não seja muito grande. É evidente que não basta somar 5 parcelas, mas com mais algumas teríamos já uma aproximação razoável (como tudo leva a indicar!).

Repetimos o que já dissémos no início, nomeadamente que o número de parcelas necessárias para se obter uma aproximação razoável depende da forma da distribuição subjacente à população. Não é indiferente, por exemplo, se temos uma população com uma distribuição simétrica ou bastante enviesada. Esta aproximação será um pouco mais desenvolvida a seguir, nas aplicações do Teorema Limite Central.

## Aplicações do Teorema Limite Central

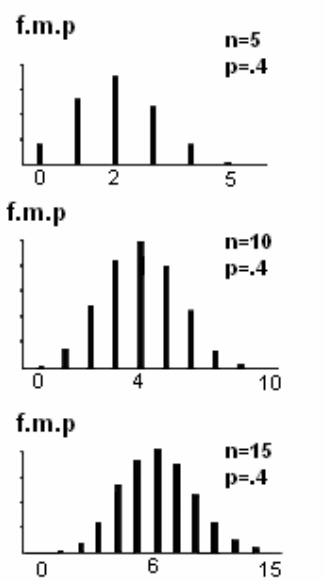
### Aproximação da distribuição Binomial pela distribuição Normal

Já vimos, quando do estudo da distribuição Binomial, que uma v.a. **X** com distribuição Binomial de parâmetros n e p, pode ser considerada a soma de n variáveis aleatórias, independentes, cada uma com distribuição Binomial de parâmetros 1 e p (variáveis aleatórias de Bernoulli). Então, invocando o Teorema do Limite Central, temos o seguinte resultado

Se **X∩ B(n,p)**, então para **n suficientemente grande**

$$P\left(\frac{X-np}{\sqrt{np(1-p)}} \le z\right) \approx \mathbf{\Phi(z)}$$

Seguem-se alguns exemplos de f.m.p. para alguns valores dos parâmetros n e p:




**Regra prática**: Considera-se que se tem uma aproximação razoável da distribuição Binomial pela Normal, quando np>10 e nq>10. Porquê?

Vimos quando estudámos o modelo Normal, que a probabilidade de uma v.a. Normal, com valor médio $\mu$ e desvio padrão $\sigma$, assumir valores menores que $\mu$-3$\sigma$ e maiores que $\mu$+3$\sigma$, é desprezável. Como o suporte da Binomial é constituído pelos inteiros entre 0 e n, inclusivé, vamos exigir que $\mu$-3$\sigma$ > 0 ou seja $\mu$ >3$\sigma$.

No caso da Binomial esta desigualdade implica que *np*>3$\sqrt{np(1-p)}$ , ou seja, $n^2p^2$>9np(1-p), de onde np>9-9p. Como 0≤p≤1, exigimos np>9. Para simplificar, exige-se np>10 e n(1-p)>10, para a outra cauda, para fazer a aproximação desejada.

**Exemplo 1** - Numa determinada cidade, a taxa de desemprego é de 7,9%. Tendo-se recolhido uma amostra de 300 pessoas, aptas para o trabalho, calcule um valor aproximado para a probabilidade de:

a) Haver menos de 18 desempregados na amostra recolhida.

b) Mais de 30 desempregados na referida amostra

Resolução:

a) Representando por X a v.a. que dá o nº de desempregados em 300 pessoas, temos

$$X \cap B(300,.079)$$

$$P(X < 18) = P(X \leq 17) \approx \Phi\left(\frac{17 - 300 \times .079}{\sqrt{300 \times .079 \times .921}}\right) \approx .076$$

b) $$P(X > 30) = 1 - P(X \leq 30) \approx 1 - \Phi\left(\frac{30 - 23.7}{4.67}\right) \approx 1 - .901 = .089$$

**Exemplo 2** – A polícia estima que 85% dos condutores usam cinto de segurança. Decidem fazer uma operação stop para controlar a sua urilização.

a) Quantos carros esperam fazer parar, até encontrarem um condutor sem cinto?
b) Qual a probabilidade de o primeiro condutor a prevaricar no que diz respeito à utilização do cinto, seja o sétimo que mandam parar?
c) Qual a probabilidade de que os primeiros 12 condutores que mandam parar utilizem todos cinto?
d) Se na primeira hora mandarem parar 30 condutores, quantos condutores esperam apanhar sem cinto?
e) Se mandarem parar 150 condutores durante a operação stop, qual a probabilidade de encontrarem pelo menos 25 condutores sem cinto?

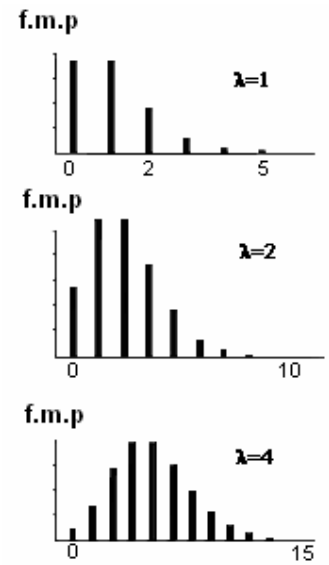
**Aproximação da distribuição de Poisson, pela distribuição Normal**

Ao estudarmos a distribuição de Poisson, vimos que uma v.a. **X** com distribuição de Poisson de parâmetro $\lambda$, pode ser considerada a soma de n variáveis aleatórias, independentes, cada uma com distribuição de Poisson de parâmetro $\lambda/n$. Então, invocando o Teorema do Limite Central, temos o seguinte resultado

Se $\mathbf{X \cap P(\lambda)}$, então para $\lambda$ **suficientemente grande**

$$P\left(\frac{X - \lambda}{\sqrt{\lambda}} \leq z\right) \approx \mathbf{\Phi(z)}$$

Seguem-se alguns exemplos de f.m.p. para alguns valores do parâmetro $\lambda$:




**Regra prática**: Considera-se que se tem uma aproximação razoável da distribuição de Poisson pela Normal, quando $\lambda>20$.

**Exemplo 3** - Durante um dia (de 10 horas), registou-se o número de doentes que chegam a um serviço de urgência, por períodos sucessivos de 15 minutos. Representando por Y o nº de doentes que chegam em intervalos de 15 minutos, os resultados obtidos foram os seguintes:

| Y | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| Freq. | 5 | 11 | 11 | 7 | 4 | 1 | 0 | 1 |

a) Admitindo que o nº de chegadas em cada unidade de tempo(período de 15 minutos), segue uma distribuição de Poisson, obtenha um estimador para o parâmetro da distribuição.

b) Determine um valor aproximado para a probabilidade de o número de doentes que chegam ao fim de 3 horas, ser superior a 30.

Resolução:

a)

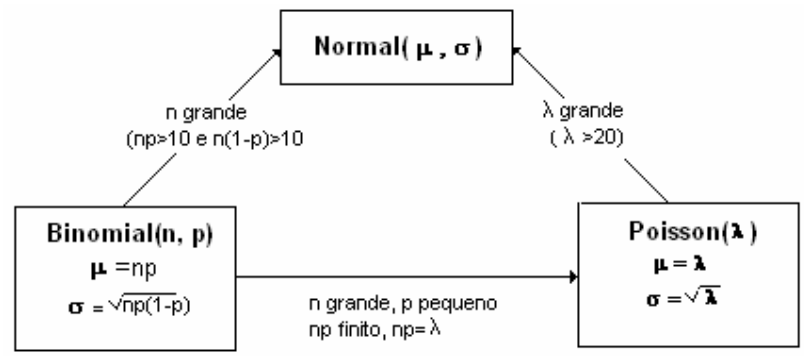
$$\hat{\lambda} = \frac{0\times 5+1\times 11+2\times 11+\ldots+6\times 0+7\times 1}{5+11+11+\ldots+0+1} = 2.05$$

b) Seja X a v.a. que representa o nº de doentes que chegam ao fim de 3 horas. Admitindo que existe independência entre as chegadas em períodos sucessivos, e atendendo a que em 3 horas temos 12 períodos de 15 minutos, vem que

$$X \cap P(12\times 2.05)$$

$$P(X>30)=1 - P(X\leq 30)\approx 1 - \Phi\left(\frac{30-24.6}{\sqrt{24.6}}\right) \approx 1 - .862 = .138$$

No seguinte esquema, resumimos os resultados obtidos, no que diz respeito a aproximações.




**Exemplo 4** - A altura dos homens pertencentes à classe etária [30,35], segue um modelo Normal de valor médio 165 cm e desvio padrão 30 cm. Recolhida uma amostra de dimensão 50, daquela população, qual a distribuição da média das alturas? Essa distribuição é exacta ou aproximada? Calcule a probabilidade da média ser inferior a 160 cm.

Resolução:

Seja X a v.a. que representa a altura dos homens pertencentes à classe etária [30,35]

$$X \cap N(165,30)$$

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_{50}}{50} \quad \text{em que} \quad X \cap N(165,30), \quad i=1,2,\ldots,50$$

Então $\overline{X} \cap N(165, 30/\sqrt{50})$

$$P(\overline{X} \leq 160) = \Phi\left(\frac{160-165}{30/\sqrt{50}}\right) \approx \Phi(-1.18) \approx .12$$

**Exemplo 5** - O gerente da fábrica "Confie", afirma que a percentagem de artigos defeituosos produzidos é de 8%. Um cliente que pretende comprar um lote de 100 peças, diz que não pagará o lote e o devolve, no caso de verificar que existem mais de 10 artigos defeituosos. Qual a probabilidade de o lote ser devolvido?

Resolução: Seja N a v.a. que representa o número de peças defeituosas no lote de 100. Então

$$N \cap Bi(100,.08)$$

$$P(N > 10) = 1 - P(N \leq 10) \approx 1 - \Phi\left(\frac{10 - 100 \times .08}{\sqrt{100 \times .08 \times .92}}\right)$$

$$\approx 1 - \Phi(.74) \approx .23$$

**Exemplo 6** - O número médio de aviões que chegam a determinado aeroporto é de 1 em cada 40 segundos. Qual a probabilidade aproximada de que numa hora, seleccionada ao acaso, ocorram:

a) Pelo menos 75 chegadas.

b) Menos de 100 chegadas.

Resolução: Seja Y a v.a. que representa o número de aviões que chegam ao aeroporto, num período de 40 segundos. Então esta v.a. pode ser bem modelada por uma distribuição de Poisson

$$Y \cap P(1)$$

Seja S a v.a. que representa o número de aviões que chegam durante 1 hora (90 períodos de 40 segundos). Então

$$S \cap P(90)$$

a) $P( S \geq 75 ) = 1 - P( S \leq 74) \approx 1 - \Phi(\frac{74-90}{\sqrt{90}}) \approx 1 - \Phi(-1.69) \approx .9545$

b) $P( S < 100) = P( S \leq 99) \approx \Phi(\frac{99-90}{\sqrt{90}})$

$\approx .8289$

## 9.3 – Distribuição de amostragem da proporção

Suponhamos que temos uma população constituída por indivíduos que pertencem a uma de duas categorias, que representamos por A e $A^C$. Representemos por **p** a proporção (desconhecida) de indivíduos que pertencem à categoria A. Um exemplo desta situação é o que se passa quando se considera a a população de uma determinada cidade e a proporção p de eleitores dessa cidade que estão dispostos a vota num determinado candidato a presidente da Câmara, nas próximas eleições autárquicas.

Pretendemos fazer inferência sobre o parâmetro **p**, pelo que se recolhe da população uma amostra de dimensão **n**. Seja **X** a v.a. que representa o nº de indivíduos da amostra que pertencem à categoria A. Um estimador natural para o parâmetro **p**, é a frequência relativa $\frac{X}{n}$, que representamos por $\hat{\mathbf{p}}$. Do mesmo modo que a média $\overline{\mathbf{X}}$ é uma variável aleatória, também $\hat{\mathbf{p}}$ é uma v.a. cujo valor depende amostra recolhida, por intermédio da v.a. X. Vejamos então como obter a sua distribuição de amostragem.

**Exemplo 7** – No Departamento de Estatística de uma determinada Faculdade, há 5 docentes que são professores associados, dos quais 3 são mulheres – Maria, Ana, Rita e 2 são homens – Pedro e Tiago. Se representarmos por p a percentagem de homens que são professores associados, temos que p=2/5. Suponhamos que pretendíamos estimar esta proporção utilizando amostras de dimensão 2, pelo que vamos construir todas as amostras desta dimensão para obter a distribuição de amostragem da estatística utilizada:

| Amostra | $\hat{\mathbf{p}}$ | Amostra | $\hat{\mathbf{p}}$ |
|---|---|---|---|
| Maria, Maria | 0 | Rita, Pedro | 1/2 |
| Maria, Ana | 0 | Rita, Tiago | 1/2 |
| Maria, Rita | 0 | Pedro, Maria | 1/2 |
| Maria, Pedro | 1/2 | Pedro, Ana | 1/2 |
| Maria, Tiago | 1/2 | Pedro, Rita | 1/2 |

| | | | |
|---|---|---|---|
| Ana, Maria | 0 | Pedro, Pedro | 2/2 |
| Ana, Ana | 0 | Pedro, Tiago | 2/2 |
| Ana, Rita | 0 | Tiago, Maria | 1/2 |
| Ana, Pedro | 1/2 | Tiago, Ana | 1/2 |
| Ana, Tiago | 1/2 | Tiago, Rita | 1/2 |
| Rita, Maria | 0 | Tiago, Pedro | 2/2 |
| Rita, Ana | 0 | Tiago, Tiago | 2/2 |
| Rita, Rita | 0 | | |

A partir da tabela anterior é possível obter a distribuição de amostragem da estatística $\hat{\mathbf{p}}$:

| $\hat{\mathbf{p}}$ | 0 | .5 | 1 |
|---|---|---|---|
| Probabilidade | 9/25 | 12/25 | 4/25 |

$$E(\hat{\mathbf{p}}) = 2/5 \text{ e } Var(\hat{\mathbf{p}}) = 3/25$$

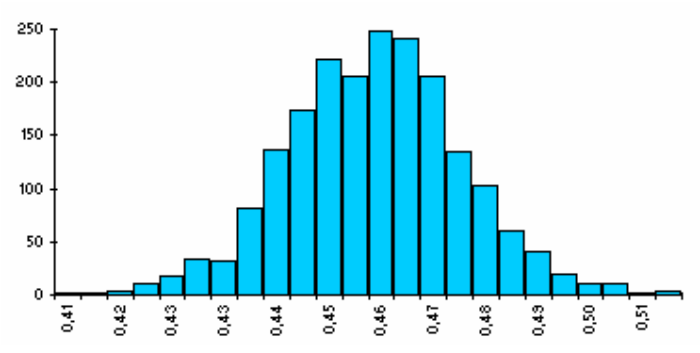
Repare-se que o valor médio da estatística $\hat{\mathbf{p}}$ coincide com o valor do parâmetro **p** que se está a estimar.

No exemplo anterior foi fácil de obter a distribuição de amostragem da proporção, pois a dimensão da população era pequena, o que não é o caso na maior parte das aplicações. Vejamos outro exemplo.

**Exemplo 8** (Adaptado de De Veaux et al, 2004) – No dia 27 de Outubro de 2000, a menos de 2 semanas para as eleições presidenciais, uma sondagem da NBC, em que foram inquiridos 1000 eleitores seleccionados aleatoriamente, dava um resultado de 46% a favor de Al Gore, contra 43% a favor de Bush. Ao mesmo tempo, uma sondagem da CNN, dava 46% para Bush, contra 42% para Al Gore. Será que alguma das sondagens estava errada? Será possível obter estes resultados quando as amostras são bem recolhidas e a população é a mesma? Qual a variabilidade que esperamos numa sondagem? Como é que varia a proporção amostral? Como é que sondagens feitas ao mesmo tempo, pela mesma organização, sobre as mesmas questões, podem dar resultados diferentes? A resposta a esta questão está no âmago da Estatística – *a compreensão da variabilidade, para melhor compreender o mundo*. Efectivamente cada sondagem é baseada numa amostra de 1000 pessoas, mas as pessoas são diferentes e por isso as proporções também são diferentes. A Estatística vai-nos permitir estudar, compreender e prever esta diferença!

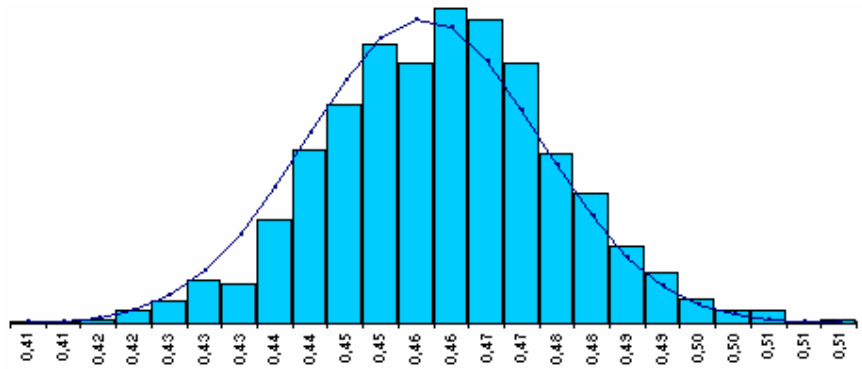
Vamos imaginar todas as amostras possíveis, de dimensão 1000, que poderiam ser recolhidas da população constituída pelos eleitores (no exemplo anterior a população era de dimensão pequena, pelo que foi possível considerar todas as amostras possíveis, de dimensão 2). Como será o aspecto do histograma construído para as proporções de eleitores que pensam votar Bush, em todas as amostras possíveis de dimensão 1000? Em vez de imaginar, vamos simular uma quantidade razoável dessas amostras, considerando como probabilidade de sucesso o valor de p=0.46. Simulámos no Excel 2000 amostras e calculámos a percentagem de sucessos em cada

uma dessas amostras. Obtivémos uma amostra de dimensão 2000 em que obtivémos para a média o valor 0.46, para o desvio padrão 0.016 e para a qual construímos o seguinte histograma:




(Observação: Os valores que aparecem no eixo dos xx, debaixo de alguns intervalos, são limites superiores desses intervalos)

Obtivémos um histograma aproximadamente simétrico, centrado no verdadeiro valor do parâmetro p, cuja forma imediatamente nos faz lembrar o modelo Normal. Efectivamente o modelo Normal é o modelo certo para a distribuição de amostragem da proporção. Para utilizar o modelo Normal, é necessário especificar o seu valor médio e o seu desvio padrão. Como o centro do histograma é p, vamos escolher para $\mu$ o valor de p. E no que diz respeito ao desvio padrão? De um modo geral o conhecimento do valor médio não nos dá qualquer informação sobre o desvio padrão. No entanto, no caso de termos uma **proporção**, a situação é diferente. Como veremos na secção seguinte, o conhecimento de p implica o conhecimento do desvio padrão para a proporção, que é igual a $\sqrt{\frac{p(1-p)}{n}}$. Então um bom modelo para a distribuição de amostragem da proporção $\hat{p}$, como estimador de p, é dado pela Normal(p, $\sqrt{\frac{p(1-p)}{n}}$), como se apresenta a seguir (com p=0.46):




Quando seleccionamos várias amostras aleatórias simples de n indivíduos, a proporção de indivíduos com a característica em estudo, varia de amostra para amostra, de acordo com o modelo da Normal considerado anteriormente.

No caso da eleição presidencial, foi conhecido o valor da verdadeira proporção dos eleitores que votaram Bush, que foi de 47.9%. No dia 27 de Outubro esta própria proporção poderia ser

diferente. Nunca saberemos o verdadeiro valor desta proporção, enquanto decorriam as sondagens. Conhecíamos sim um intervalo que a continha, como veremos mais à frente.

### 9.3.1 – Valor médio e variância do estimador $\hat{\mathbf{p}}$ da proporção p

O estimador $\hat{\mathbf{p}}$ é a frequência relativa com que se verifica na amostra de dimensão n, a característica em estudo, ou seja, é dado por $\frac{X}{n}$, onde X – nº de elementos na amostra com a característica, é uma v.a. com distribuição Binomial de parâmetros n e p. O facto de da v.a. X ter distribuição Binomial, de parâmetros n e p, resulta de termos admitido a hipótese de a selecção da amostra ser feita com reposição, o que implica que a composição da população relativamente à característica de interesse não se altera. Então

$$E(\hat{\mathbf{p}}) = p \text{ e } Var(\hat{\mathbf{p}}) = \frac{p(1-p)}{n}$$

Repare-se que o valor médio do estimador $\hat{\mathbf{p}}$ coincide com o parâmetro a estimar. Esta particularidade já havia sido observada com o estimador do valor médio, ou seja a média, pelo que mais uma vez temos um estimador *não enviesado*. Além de não enviesado, $\hat{\mathbf{p}}$ também é *consistente*, como já havia sido observado com a média $\overline{\mathbf{X}}$. Repare-se que, à medida que a dimensão da amostra aumenta, a variabilifdade de $\hat{\mathbf{p}}$ em torno de p, tende para 0. Mais uma vez estamos a utilizar um "bom" estimador para estimar um parâmetro desconhecido, neste caso a proporção.

### 9.3.2 – Distribuição de amostragem de $\hat{\mathbf{p}}$

Vimos também ao estudar a v.a. Binomial X, que para n suficientemente grande a sua distribuição pode ser aproximada pela distribuição Normal,

$$P\left(\frac{X-np}{\sqrt{np(1-p)}} \le z\right) \approx \Phi(z)$$

de onde

$$P\left(\frac{\frac{X}{n}-p}{\sqrt{\frac{p(1-p)}{n}}} \le z\right) \approx \Phi(z)$$

Então a forma da distribuição de amostragem da proporção é aproximadamente Normal, como consequência do Teorema Limite Central.

Para n suficientemente grande a distribuição de amostragem da proporção $\hat{\mathbf{p}}$ pode ser aproximada pela distribuição Normal, com valor médio p e variância p(1-p)/n

$$\mathbf{P}\left(\frac{\hat{\mathbf{p}}-\mathbf{p}}{\sqrt{\frac{\mathbf{p(1-p)}}{\mathbf{n}}}} \le \mathbf{z}\right) \approx \Phi(\mathbf{z})$$

Como já dissemos neste texto "todos os modelos são maus, alguns modelos são úteis". Vem de novo esta citação de Box, a propósito do seguinte: se a amostragem não tiver sido feita com reposição, já o modelo Binomial utilizado para obter a distribuição de amostragem da proporção, não deveria ser aplicado, uma vez que a probabilidade de sucesso se altera à medida que seleccionamos os elementos para a amostra, uma vez que a composição da população se altera. Assim, para que a probabilidade de sucesso se mantenha aproximadamente constante, é necessário que a amostra seja suficientemente pequena, quando comparada com a população. Mas, por outro lado, para se poder aplicar a aproximação da distribuição Binomial pela Normal, é necessário que a dimensão da amostra seja suficientemente grande. Chegamos assim a uma contradição! De um modo geral esta contradição não causa problemas, pois a maior parte das vezes a dimensão da população é mais do que 10 vezes superior à dimensão da amostra. No que diz respeito à dimensão da amostra exigida, para podermos inferir para a população, propriedades verificadas na amostra, veremos mais à frente que essa dimensão terá de ser tanto maior, quanto mais próximo de 0.5 for o valor de p.

**Exemplo 9** – De acordo com o censo de 91 a percentagem da população portuguesa (residente em Portugal) feminina é de 51.74%. Numa amostra de dimensão 240, escolhida aleatoriamente de entre a população portuguesa, qual a probabilidade da percentagem de mulheres ser superior a 56%?

Resolução: Seja $\hat{\mathbf{p}}$ a percentagem de mulheres na amostra de dimensão 240. Então

$$P(\hat{\mathbf{p}} > .52) = P\left(\frac{\hat{\mathbf{p}} - .5174}{\sqrt{\frac{.5174(1-.5174)}{240}}} > \frac{.56 - .5174}{\sqrt{\frac{.5174(1-.5174)}{240}}}\right) = 1 - \Phi(1.32) = .0934$$

**Exercícios**

**1**. Pretende-se adicionar números num computados. O computador ao receber os números arredonda-os segundo as regras habituais. Admitindo que os erros de arredondamento são independentes de número para número e têm distribuição uniforme no intervalo [-.4, .6], determine:

a) A probabilidade de que o erro total seja maior do que 7, se se adicionarem 75 números (admita que o erro da soma é igual à soma dos erros das parcelas).

b) Quantos números poderão ser adicionados de modo que o erro total seja menor do que 6, com probabilidade .5478?

**2**. Três espécies diferentes de determinada planta são difíceis de distinguir uma semana após a germinação, altura em que devem ser transplantadas. Metade das plantas são de tipo A, 3/8 são de tipo B e 1/8 são de tipo C. Uma semana depois da germinação, a altura (em cm) das plantas de cada tipo segue uma distribuição Normal com os seguintes parâmetros:

| | Valor médio | Variância |
|---|---|---|
| Tipo A | 6.2 | 1.00 |

| | | |
|---|---|---|
| Tipo B | 4.9 | 0.36 |
| Tipo C | 3.3 | 0.25 |

a) Uma semana após a germinação, que proporção de plantas de tipo B têm altura inferior a 4 cm?

b) O terreno para o qual as plantas vão ser transplantadas encontra-se dividido em três lotes, cujo solo foi especialmente preparado para cada um dos três tipos de plantas. As plantas de altura inferior a 4 cm são colocadas no solo preparado para o tipo C, as que têm altura entre 4 cm e 5.2 cm são colocadas no solo preparado para o tipo B e as restantes plantas são colocadas no solo preparado para o tipo A. Determine a proporção de plantas de tipo A que foram transplantadas para o lote de terreno apropriado.

c) Numa amostra de 70 plantas com altura inferior a 4 cm, qual a probabilidade de que pelo menos 12 dessas plantas sejam de tipo B?

**3.** Supõe-se que numa população existem três vezes mais indivíduos fumadores do que não fumadores. Sabe-se que a percentagem de doentes com determinada doença pulmonar , entre os fumadores e não fumadores é respectivamente de 60% e 20%.

a) Determine a probabilidade de um indivíduo ter doença pulmonar.

b) Determine a probabilidade de um doente pulmonar ser fumador.

c) Qual a probabilidade de numa amostra de 10 doentes, pelo menos três serem fumadores?

d) Qual a probabilidade de numa amostra de 225 doentes, mais de metade serem fumadores?

**4**. O número de nascimentos verificados por dia , numa certa maternidade, é uma variável aleatória com distribuição de Poisson.

a) Sabendo que a probabilidade de não haver nascimentos num dia é .368, determine a probabilidade de ocorrerem pelo menos 3 nascimentos por dia.

b) Determine um valor aproximado da probabilidade de se registarem entre 28 e 32 nascimentos (inclusivé) no mês de Abril, sabendo que o número de nascimentos é independente de dia para dia.

c) Sabe-se que com probabilidade igual a .95, o número de nascimentos no mês de Abril não excederá determinado valor. Determine esse valor.

**5**. Num prédio habitam 60 pessoas: 40 adultos e 20 crianças. Admita que os pesos dos adultos e das crianças são variáveis aleatórias N(75,10) e N(35,10), respectivamente.

a) Calcule a probabilidade de um indivíduo do prédio ter peso inferior a 55 Kg.

b) Dado um indivíduo desse prédio com peso superior a 55 Kg, qual a probabilidade de ser criança?

c) O elevador do prédio só funciona com carga inferior a 300 Kg. Duas crianças já o ocupam, quando três adultos pretendem entrar. Qual a probabilidade de poderem seguir juntos?

**6** . A quantidade diária de potássio necessária para o organismo varia entre 2000 a 6000 mg, sendo necessário maiores quantidades nos dias de verão, com o tempo quente. A quantidade de potássio existente nos alimentos varia de alimento para alimento, sendo por exemplo em média de 7 mg numa coca cola, 46 mg numa cerveja, 630 mg numa banana, 300 mg numa cenoura, etc.

Admitindo que o potássio se distribui normalmente nas bananas e nas cenouras, com desvio padrão respectivamente igual a 40 mg e 15 mg, determine a probabilidade de que a quantidade mínima necessária seja excedida se comer 3 bananas e 1 cenoura. Se num dia só comesse bananas, qual o número mínimo de bananas que teria de comer para que com uma probabilidade de .95, excedesse a quantidade máxima de potássio necessária?

**7** . Suponha que o tempo de estudo semanal dos alunos de determinado colégio tem uma distribuição enviesada para a direita com valor médio 9 horas e desvio padrão 3 horas. Determine a probabilidade de que em média o tempo gasto a estudar por 40 estudantes

a) esteja entre 8.5 e 9 horas

b) seja inferior a 8 horas

**Capítulo 10**

# Introdução à estimação

## 10.1 - Noções preliminares sobre estimação. Estimadores pontuais e intervalares.

Dada uma amostra, vimos que é possível fazer a sua redução, através do cálculo de certas estatísticas. No entanto, a importância destas características amostrais não se fica por aqui, pois o nosso objectivo vai ser utilizá-las para inferir algo sobre a população subjacente à amostra. Foi nesta perspectiva que falámos em utilizar:

**i)** a média $\bar{\mathbf{x}}$ como estimativa do valor médio **μ**;
**ii)** a proporção $\hat{\mathbf{p}}$=x/n, onde x representa o nº de sucessos obtidos numa certa amostra de dimensão n, como estimativa da probabilidade **p** de sucesso, na distribuição Binomial, etc.

Quer dizer que as estatísticas referidas permitem-nos obter determinados valores que servem como estimativas dos parâmetros (desconhecidos) ou características das distribuições populacionais - a estes valores chamamos **estimativas pontuais**. Por vezes interessa-nos obter, não um valor que estime o parâmetro em causa, mas um intervalo que contenha, com determinada probabilidade, esse parâmetro - neste caso pretendemos uma **estimativa intervalar** ou um **intervalo de confiança**.

Um **estimador** é uma variável aleatória, função da amostra aleatória, que para valores observados da amostra fornece estimativas pontuais ou estimativas intervalares do parâmetro populacional desconhecido. Então, a v.a. $\overline{\mathbf{X}}$ é um estimador do valor médio, assim como $\hat{\mathbf{p}} = \frac{X}{n}$ é um estimador da probabilidade p. De um modo geral quando nos referimos ao estimador utilizamos letra maiúscula, enquanto que a estimativa se representa com letra minúscula. Esta metodologia por vezes não é seguida, como é por exemplo, no caso anteriormente considerado da proporção.

O que é um **"bom"** estimador?

O facto de termos escolhido a média como estimador do valor médio, não se deve unicamente à analogia existente, entre parâmetros populacionais e parâmetros amostrais. Existem alguns critérios que definem à partida, se um estimador é "bom" ou "mau". Assim, o critério mais utilizado exige que o estimador seja *não enviesado* ou centrado, isto é, que o seu valor médio coincida com o parâmetro populacional a estimar, e de entre os que satisfazem esta condição deve ter *variância mínima*. Estas duas propriedades são, de certo modo intuitivas, pois ao considerar um estimador

esperamos que as estimativas que ele fornece coincidam, em média, com o parâmetro a estimar, e além disso a variabilidade dessas estimativas, em torno do parâmetro, deve ser pequena. Por exemplo, no caso concreto de populações simétricas, podem existir vários estimadores centrados para o valor médio, nomeadamente a média e a mediana. No entanto, escolhe-se o que tem variância mínima, que é a média.

No que diz respeito à variância populacional $\sigma^2$, alguns estimadores possíveis são

$$S^2 = \frac{\Sigma(X_i - \overline{X})^2}{n-1} \qquad \text{ou} \qquad S'^2 = \frac{\Sigma(X_i - \overline{X})^2}{n}$$

Ao considerar estas duas estatísticas, costuma-se dizer que por razões que se prendem com a inferência estatística, a estatística mais utilizada é $S^2$. Neste momento já podemos dar a razão que nos leva a escolher $\frac{\Sigma(X_i - \overline{X})^2}{n-1}$ - é o facto de este estimador, ao contrário de $\frac{\Sigma(X_i - \overline{X})^2}{n}$, ser centrado, pois pode-se mostrar que $E[S^2]=\sigma^2$ com $X_i$, i=1,...,n variáveis aleatórias independentes e identicamente distribuídas a X, enquanto que $E[S'^2]= \frac{n-1}{n}\sigma^2$. Assim, quando a dimensão da amostra é suficientemente grande, $S'^2$ é assintóticamente centrado, pois $\frac{n-1}{n} \to 1$, sendo indiferente utilizar um ou outro estimador.

## 10.2 - Estimação da proporção. Intervalo de confiança para a proporção

Já vimos no capítulo 9 que se tivermos uma população constituída por indivíduos que pertencem a uma de duas categorias, que representamos por A e $A^C$ em que **p** é a proporção (desconhecida) de indivíduos que pertencem à categoria A, um estimador desta proporção é $\hat{\mathbf{p}}$. Vimos que $\hat{\mathbf{p}}$ é um estimador centrado ou não enviesado e tem uma variabilidade que tende para 0, à medida que a dimensão da amostra recolhida aumenta. Podemos dizer que temos um bom estimador, pelo menos relativamente ao critério considerado anteriormente!

Então, quando pretendemos fazer inferência sobre p, recolhemos uma amostra de dimensão n e calculamos $\hat{\mathbf{p}}$. O valor obtido é uma **estimativa pontual** de p. Se recolhermos várias amostras da mesma dimensão e calcularmos outras tantas estimativas para p, não temos possibilidade de saber qual o erro associado com cada uma dessas estimativas, pelo que não temos possibilidade de saber qual a que devemos utilizar. Por exemplo, se dois jornais distintos apresentarem, no mesmo dia, as percentagens de 45% e 52% de pessoas que votarão "Sim" à Constituição Europeia, não sabemos qual a que nos merece mais confiança. Assim, por vezes é preferível utilizar uma estimativa intervalar, ou seja um intervalo aleatório que, como veremos a seguir, nos dá uma ideia do erro cometido, ao ser utilizado para estimar o parâmetro.

Já que, como vimos na secção anterior, a distribuição de amostragem de $\hat{\mathbf{p}}$ pode ser aproximada pela distribuição Normal, quando a dimensão n da amostra utilizada for suficientemente grande, então é possível, dada uma probabilidade **P**, por exemplo **.95**, obter o valor de **z** tal que se tenha

$$P\left(\frac{|\hat{p}-p|}{\sqrt{\frac{p(1-p)}{n}}} \le z\right) = .95.$$

Se $P\left(\frac{|\hat{p}-p|}{\sqrt{\frac{p(1-p)}{n}}} \le z\right) = .95$ então z=1.96, ou seja $P\left(\frac{|\hat{p}-p|}{\sqrt{\frac{p(1-p)}{n}}} \le 1.96\right) = .95$.

Trabalhando a expressão anterior obtemos

$$P\left(\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} \le p \le \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}\right) = .95$$

Se n é suficientemente grande $\hat{p}$ está suficientemente próximo de p, pelo que na expressão anterior vamos substituir p por $\hat{p}$ em $\sqrt{\frac{p(1-p)}{n}}$, obtendo-se

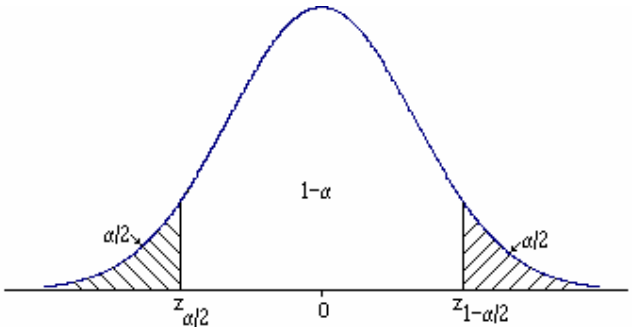
$$P\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx .95$$

Dizemos que o intervalo $\left[\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$ é um intervalo aproximado de confiança para p, com uma confiança de 95%.

De um modo geral se considerarmos uma **confiança de 100(1-α)%** (representamos por $\alpha$ uma probabilidade pequena, que associamos à desconfiança ou ao erro cometido na obtenção do intervalo de confiança), o **intervalo de confiança** para **p** assume o aspecto

$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

onde representamos por $z_{1-\alpha/2}$ o quantil de probabilidade $(1-\alpha/2)$ da N(0,1) e $\hat{p}$ é a proporção de elementos da amostra pertencentes à categoria em estudo.




**Exemplo 1** (Adaptado de De Veaux and al, 2004) – Os corais estão em declínio, em todo o mundo, possívelmente devido à poluição ou mudança da temperatura da água do mar. A morte dos recifes de corais pode ser um aviso das mudanças climáticas e poderá ter um impacto

económico ainda não calculado. Uma espécie muito bonita de coral, conhecida como Leque do Mar, é particularmente afectada pela poluição e pela doença *aspergillosis*. Em Junho de 2000, uma equipa de investigadores recolheu uma amostra de corais desta espécie, a uma profundidade de 40 pés, em Las Redes Reef, Akumal, México. Verificaram que 54 dos 104 corais que recolheram, estavam infectados com aquela doença. O que é que se pode dizer sobre a prevalência desta doença, sobre aquele tipo de corais? Para já, temos uma proporção de corais doentes de 51.9%, mas ninguém nos garante que os investigadores obteriam a mesma proporção se recolhessem outra amostra de 104 corais.O que é que podemos dizer efectivamente sobre a proporção **p** de corais infectados? Apresentamos a seguir uma lista de coisas que poderíamos dizer, ou que por vezes se dizem, e a razão pela qual não são correctas a maior parte delas:

**1.** "*51.9% de todos os corais da espécie Leque do Mar, em Las Redes Ref, estão infectados*" – Não temos informação suficiente para fazer esta afirmação. Só poderíamos fazer esta afirmação se tivéssemos investigado o que se passava com todos os corais. Assim, se recolhessemos outra amostra, obteríamos outra percentagem.

**2.** *"Provavelmente é verdade que 51.9% de todos os corais da espécie Leque do Mar, em Las Redes Ref, estejam infectados"* – Não podemos fazer esta afirmação. Podemos ter quase a certeza de que, qualquer que seja a verdadeira proporção de corais infertados, ela não será exactamente igual a 51.900%.

**3.** *"Não sabemos exactamente qual a proporção de corais infectados, da espécie Leque do Mar, em Las Redes Ref, mas* ***sabemos*** *que essa proporção está no intervalo 51.9% ±1.96* $\sqrt{\frac{0.519\times(1-0.519)}{104}}$*, ou seja 51.9%±9.6%, ou seja ainda entre 42.3% e 61.5%".* Ainda não podemos fazer esta afirmação, pois não podemos ter a certeza que a verdadeira proporção esteja neste intervalo, ou noutro qualquer.

**4.** *"Não sabemos exactamente qual a proporção de corais infectados da espécie Leque do Mar, em Las Redes Ref, mas o intervalo de 42.3% a 61.5%* ***provavelmente*** *contém a verdadeira proporção".* Agora sim, podemos fazer esta afirmação. Começámos por dar o intervalo e em seguida admitir que pensamos que esse intervalo provavelmente contém o verdadeiro valor da proporção.

Esta última afirmação está correcta, mas podemos quantificar o que é que entendemos por *provavelmente*. Podemos dizer que 95% das vezes que construirmos intervalos do tipo considerado anteriormente, conseguimos cobrir o valor de p, pelo que podemos estar 95% confiantes de que aquele intervalo seja um dos que contém p.

**5.** *Temos uma confiança de 95% de que o intervalo entre 42.3% e 61.5% contenha a percentagem de corais infectados, da espécie Leque do Mar, em Las Redes Reef*. A este intervalo chamamos um **intervalo de confiança**.

**Confiança e precisão**

*Qual a dimensão da amostra necessária para obter um intervalo de 100(1-$\alpha$)% de confiança, cuja amplitude não exceda d?* Repare-se que a amplitude do intervalo nos dá a *precisão* – quanto menor for a amplitude, maior será a precisão. Efectivamente não estamos interessados em obter um intervalo com uma grande amplitude, pois numa situação extrema dizemos que o intervalo [0, 1] contém a probabilidade p, que pretendemos estimar, com uma confiança de 100%!

Da forma do intervalo de confiança para p, verificamos que existem duas maneiras de diminuir a sua amplitude, que é igual a $2\, z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ . Assim:

i) Ou diminuímos a confiança, o que implica obter um valor mais pequeno para o quantil $z_{1-\alpha/2}$, ou

ii) aumentamos a dimensão da amostra.

A solução apresentada em i) não é aconselhável - num caso extremo obteríamos um intervalo de amplitude nula (estimativa pontual!), mas com uma confiança de 0%!

Então vejamos como proceder adoptando a solução preconizada em ii). Pretendemos que

$$2\, z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq d$$

de onde

$$n \geq \left(\frac{2z_{1-\alpha/2}}{d}\right)^2 \hat{p}(1-\hat{p})$$

Se não tivermos uma estimativa de p, então devemos considerar o valor máximo do 2º membro da desigualdade anterior, que se obtém quando $\hat{p}=\frac{1}{2}$ donde um limite superior para n será

$$\mathbf{n \approx \left(\frac{z_{1-\alpha/2}}{d}\right)^2}$$

Chamamos a atenção para que este valor de n, de um modo geral, peca por excesso, já que foi obtido para a pior situação do valor do parâmetro a estimar estar próximo de 0.5. Assim, é aconselhável proceder a um estudo prévio, ou recolher informação eventualmente existente, para ter uma ideia do valor do parâmetro, se os custos com a recolha da amostra forem elevados.

Chama-se **margem de erro**, a metade da amplitude do intervalo de confiança. Representando a margem de erro por ME, temos na expressão anterior que dá o valor adequado para a dimensão da amostra:

$$\mathbf{n \approx \left(\frac{z_{1-\alpha/2}}{2ME}\right)^2}$$

Repare que, fixando a dimensão da amostra, quanto maior for a confiança, maior será a margem de erro. Podemos aumentar a confiança até 100%. Mas, na verdade, qual a utilidade de um intervalo, com essa confiança?

**Exemplo 2 -** Numa cidade com 25000 habitantes considerou-se uma amostra de 1600 pessoas para estimar a percentagem de eleitores que votavam num certo candidato. Dos 1600 inquiridos, 917 declararam ter a intenção de votar no dito candidato. Determine um intervalo de 95% de confiança para a proporção de eleitores da cidade que pensam votar no candidato.

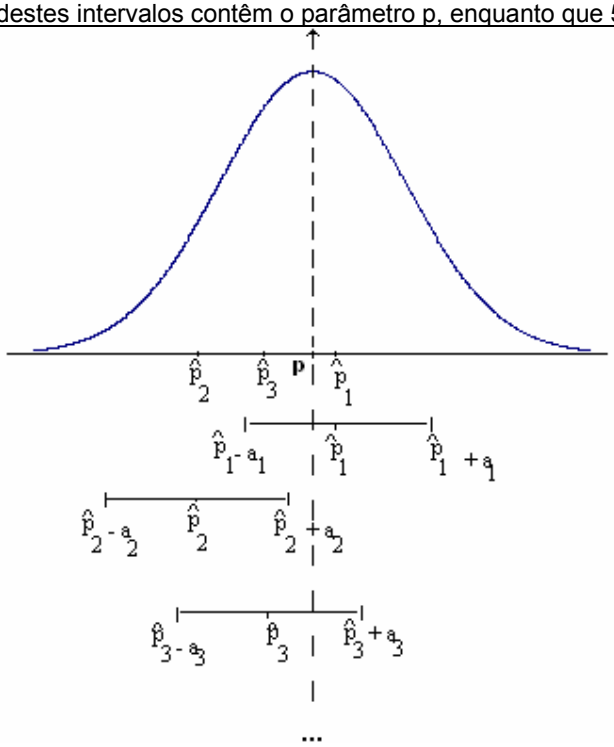
Resolução: Uma estimativa de p é $\hat{p}$=917/1600= .57. O intervalo de confiança pretendido será $[.57 - 1.96\sqrt{\frac{.57\times.43}{1600}}\ ,\ .57 + 1.96\sqrt{\frac{.57\times.43}{1600}}\ ]$, ou seja [.55, .59], pelo que o intervalo tem amplitude .04.

Qual a dimensão da amostra necessária para reduzir a metade a amplitude anterior? Como consideramos o mesmo nível de significância vem $n \approx \frac{1.96^2}{.02^2} \approx 9604$. Repare-se que neste caso já não estão satisfeitas as condições para podermos aplicar o modelo Binomial, pois a dimensão da amostra é muito grande, quando comparada com a dimensão da população.

**Interpretação do intervalo de confiança**

Ao interpretar o intervalo de confiança deve-se ter em atenção que o que é aleatório é o intervalo e não a percentagem p desconhecida - a variabilidade existe no processo de amostragem e não no parâmetro. Quando se recolhem várias amostras, o valor de $\hat{p}$ é diferente de amostra para amostra, pelo que os limites do intervalo variam.

Ao calcular um intervalo com 95% de confiança (a partir de uma amostra entretanto recolhida), não significa que a probabilidade do intervalo conter o parâmetro é .95, já que o intervalo contém ou não contém o parâmetro. Como deve ser interpretado o intervalo de confiança é da seguinte forma: <u>ao recolher 100 amostras da mesma dimensão e ao calcular os intervalos correspondentes, aproximadamente 95 destes intervalos contêm o parâmetro p, enquanto que 5 não o contêm</u>,



com $a_i = \sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n}}$. Na figura anterior representámos 3 intervalos, dos quais 2 contêm p, enquanto um terceiro não contém o valor de p. Chamamos a atenção para que quando calculamos um intervalo de confiança para a proporção, nunca sabemos se ele contém ou não o verdadeiro valor da proporção. Estamos confiantes que sim, já que em 95% das vezes que calculamos esses intervalos, eles contêm o valor de p. Já seria muito azar, o nosso intervalo ser um dos 5% de intervalos que não contêm o valor de p!

## 10.3 - Estimação do valor médio. Intervalo de confiança para o valor médio

Dada uma população **X**, com valor médio **μ**, desconhecido, e desvio padrão **σ**, suponhamos que se pretende estimar o parâmetro **μ**. Já vimos que um bom estimador para o *valor médio* é a *média*, pelo que a maneira de proceder é a seguinte: recolhe-se uma amostra de dimensão n da população a estudar, $x_1, x_2, \ldots, x_n$, e calcula-se a média $\overline{x} = \sum x_i/n$. Este valor é considerado como *estimativa pontual* de μ.

No entanto, se tivesse sido outra a amostra recolhida, nomeadamente $x'_1, x'_2, \ldots, x'_n$, seria natural que a estimativa obtida para μ através desta amostra, diferisse da inicialmente obtida. Qual a confiança que devemos atribuir a uma ou a outra? Surge assim, intuitivamente, a necessidade de um outro processo, que não só nos forneça o método de estimar, mas permita simultaneamente saber qual a confiança que devemos atribuir ao resultado obtido, tal como no caso da proporção.

### 10.3.1 - Intervalo de confiança para o valor médio - σ conhecido

Consideremos a população **X** com distribuição **Normal** de parâmetros **μ** e **σ**, em que o parâmetro **σ é conhecido**. Então, como vimos no capítulo 9, para a distribuição da média, tem-se,

$$\overline{X} \cap N(\mu, \sigma/\sqrt{n})$$

pelo que é possível obter o valor de z tal que

$$P\left[\frac{|\overline{X}-\mu|}{\sigma/\sqrt{n}} \leq z\right] = .95 \qquad \Rightarrow \qquad z=1.96$$

A probabilidade anterior pode-se escrever

$$P[\overline{X} - 1.96\,\sigma/\sqrt{n} \leq \mu \leq \overline{X} + 1.96\,\sigma/\sqrt{n}] = .95$$

ou seja, $[\overline{X} - 1.96\,\sigma/\sqrt{n},\ \overline{X} + 1.96\,\sigma/\sqrt{n}]$ é um intervalo aleatório, que contém o valor médio μ, com uma probabilidade ou confiança igual a .95, ou por outras palavras, se recolhermos um grande número de amostras (de igual dimensão), esperamos que cerca de 95% dos intervalos $[\overline{x} - 1.96\sigma/\sqrt{n},\ \overline{x} + 1.96\sigma/\sqrt{n}]$ obtidos, contenham μ, enquanto 5% dos intervalos não o conterão.

Para considerar um exemplo concreto, admitamos por exemplo, que o peso dos indivíduos do sexo masculino, de 1.65 m de altura, tem distribuição normal com valor médio 60 e desvio padrão

4. Nas cidades de Lisboa, Porto, Coimbra, Braga e Évora recolheram-se amostras de pesos de 10 indivíduos escolhidos ao acaso (com 1.65 m de altura), tendo-se obtido os seguintes resultados:

| | | | | | | | | | | | **Média** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| L | 55.9 | 56.3 | 56.8 | 57.2 | 61.2 | 61.9 | 62.5 | 63.8 | 64.4 | 68.2 | 60.82 |
| P | 55.7 | 55.8 | 57.0 | 57.4 | 59.0 | 59.5 | 59.9 | 60.4 | 64.2 | 67.7 | 59.66 |
| C | 53.0 | 54.6 | 54.7 | 54.8 | 57.6 | 58.6 | 62.4 | 63.5 | 65.5 | 66.6 | 59.13 |
| B | 57.3 | 58.1 | 58.6 | 58.7 | 59.0 | 61.9 | 62.6 | 64.4 | 64.9 | 66.7 | 61.22 |
| E | 49.5 | 50.4 | 52.8 | 54.3 | 55.3 | 57.0 | 61.2 | 62.6 | 63.2 | 64.1 | 57.04 |

Qualquer uma das médias obtidas pode ser considerada como estimativa pontual do valor médio 60. Pensemos ainda na mediana amostral, como estimador de μ (nas distribuições simétricas o valor médio coincide com a mediana). As estimativas obtidas para as diferentes amostras seriam:

L 61.55

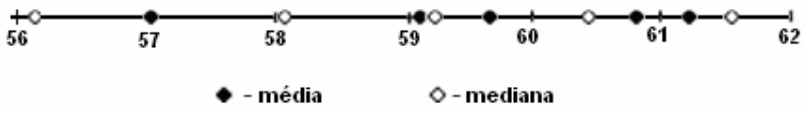
P 59.25

C 58.10

B 60.45

E 56.15

Dispondo os valores obtidos para as médias e as medianas, num segmento de recta, verificamos que a mediana apresenta maior variabilidade do que a média, em relação ao valor médio (embora uma amostra de dimensão 5 não seja significativa!).




Vejamos agora o que se passa com a estimação intervalar. Considerando o intervalo aleatório $[\overline{X} - 1.96 \times \frac{4}{\sqrt{10}}, \overline{X} + 1.96 \times \frac{4}{\sqrt{10}}]$, com confiança de 95%, para as amostras consideradas anteriormente, chegámos aos seguintes resultados:

| Cidade | $\overline{x}$ | $[\overline{x} - 2.48$ , $\overline{x} + 2.48]$ |
|---|---|---|
| L | 60.82 | [58.34 , 63.30] |
| P | 59.66 | [57.18 , 62.14] |
| C | 59.13 | [56.65 , 61.61] |
| B | 61.22 | [58.74 , 63.70] |
| E | 57.04 | [54.56 , 59.52] *** |

Dos intervalos obtidos, concluímos que 4 contêm o valor médio enquanto que um não o contém (assinalado com ***).

Uma questão que se levanta neste momento é a seguinte: o que acontece se exigirmos um intervalo de confiança com uma probabilidade de 99% em vez de 95%? Facilmente se conclui, que quanto maior for o nível de confiança exigido, maior será a amplitude do intervalo obtido. Para um nível de confiança de 99% o intervalo de confiança será $[\overline{X}-2.58\,\sigma/\sqrt{n}\,,\ \overline{X}+2.58\,\sigma/\sqrt{n}\,]$ e na realidade a amplitude pode ser tão grande que deixe de ter significado o cálculo do intervalo. No limite temos um intervalo de amplitude infinita, mais precisamente R, com uma confiança de 100%!

De um modo geral, dada uma população **N(μ,σ)**, um **intervalo de confiança** para o **valor médio**, com um **nível de confiança** de **100(1-α)%**, obtém-se considerando

$$P\left[-z_{1-\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z_{1-\alpha/2}\right] = 1-\alpha$$

onde representamos por $z_{1-\alpha/2}$ o quantil de probabilidade $1-\alpha/2$, da normal (0,1).

A partir da probabilidade anterior conclui-se imediatamente, que o **intervalo de confiança** para o valor médio tem a forma

$$\left[\overline{X} - z_{1-\alpha/2}\,\sigma/\sqrt{n}\,,\ \overline{X} + z_{1-\alpha/2}\,\sigma/\sqrt{n}\right]$$

para uma confiança de $100(1-\alpha)\%$, e qualquer que seja a dimensão da amostra considerada.

Admitamos agora, que a distribuição da população de que se pretende estimar o valor médio já **não é normal**. Neste caso, as conclusões anteriormente obtidas continuam a ser válidas, mas exige-se que a **dimensão da amostra** seja **suficientemente grande** (n>30), para ser possível aplicar o teorema do limite central - os resultados agora não serão exactos, mas sim aproximados.

Resumindo

Dada uma população **N(μ,σ)** e uma amostra de dimensão qualquer, ou uma amostra de dimensão suficientemente grande (n>30), no caso de a população já não ser normal, **σ conhecido**, um **intervalo de confiança** para o **valor médio**, com um **nível de confiança** de **100(1-α) %**, tem a forma

$$\left[\overline{X} - z_{1-\alpha/2}\,\sigma/\sqrt{n}\,,\ \overline{X} + z_{1-\alpha/2}\,\sigma/\sqrt{n}\right]$$

onde representamos por $z_{1-\alpha/2}$ o quantil de probabilidade $1-\alpha/2$, da N(0,1).

Suponhamos ainda que a população X tem distribuição **normal** de valor médio μ desconhecido e desvio padrão σ conhecido, ou que a dimensão da amostra é suficientemente grande. Pretende-se determinar para o valor médio, um intervalo de confiança com um nível de confiança de $100(1-\alpha)\%$ e cuja amplitude não exceda **d**. Qual a dimensão exigida para a amostra? Tendo em conta a forma para o intervalo de confiança, concluímos ainda, que um processo para diminuir a amplitude do intervalo de confiança, será aumentar a dimensão da amostra e essa dimensão terá de ser tal que:

$$2\, z_{1-\alpha/2}\ \sigma/\sqrt{n} \leq d$$

ou

$$\mathbf{n \geq (\, 2\, z_{1-\alpha/2}\ \sigma/d)^2}$$

Repare-se que da expressão anterior podemos concluir, para já, que a dimensão da amostra que deve ser recolhida, depende da variabilidade existente na população.

**10.3.2 - Intervalo de confiança para o valor médio - $\sigma$ desconhecido.**

Em todas as conclusões obtidas até aqui, no que respeita à estimação do valor médio, admitimos que o parâmetro $\sigma$ era conhecido. No entanto na situação mais vulgar, tanto $\mu$ como $\sigma$ são desconhecidos. Para resolver o problema, vamos distinguir dois casos:

**a)** Se a dimensão da amostra for suficientemente grande (n>30), utiliza-se a estatística **S** como estimador de $\sigma$ e o intervalo de confiança, para um nível de confiança de 100(1-$\alpha$)% tem a forma

$$\mathbf{[\,\overline{X} - z_{1-\alpha/2}\, S/\sqrt{n}\ ,\ \overline{X} + z_{1-\alpha/2}\, S/\sqrt{n}\,]}$$

onde representamos por $z_{1-\alpha/2}$ o quantil de probabilidade 1-$\alpha$/2, da normal (0,1), pois para n grande, $\mathbf{\sqrt{n}(\overline{X}-\mu)/S}$ continua a ter distribuição aproximadamente normal.

**b)** Se a dimensão da amostra for pequena, mas a população tem **distribuição normal,** então $\mathbf{\sqrt{n}(\overline{X}-\mu)/S}$ já não tem distribuição normal, mas sim a chamada distribuição **t de Student** com (n-1) graus de liberdade, como já vimos no capítulo 9, no estudo da distribuição de amostragem da média. Nestas condições o intervalo de confiança para a média, para um nível de confiança de 100(1-$\alpha$)% é

$$\mathbf{[\,\overline{X} - t_{1-\alpha/2}(n-1)\, S/\sqrt{n}\ ,\ \overline{X} + t_{1-\alpha/2}(n-1)\, S/\sqrt{n}\,]}$$

onde representamos por $\mathbf{t_{1-\alpha/2}(n-1)}$ o quantil de probabilidade 1-$\alpha$/2, da distribuição t de Student, com n-1 graus de liberdade. Esta distribuição, assim como a normal, encontra-se tabelada.

Convém ainda observar que a distribuição t-Student se aproxima da distribuição normal reduzida, à medida que o número de graus de liberdade aumenta. Assim, tem toda a propriedade utilizar a aproximação feita em a), para grandes amostras.

Observação – Para usar o modelo de Student, é necessário que a população seja Normal. Na prática, é suficiente que os dados sejam provenientes de uma população unimodal e simétrica, se a dimensão da amostra for superior a 15 (De Veaux and al, 2004).

**Qual a dimensão da amostra necessária para que o intervalo de confiança tenha alguma utilidade?**

Já anteriormente definimos margem de erro (ME), como sendo metade da amplitude do intervalo de confiança e dissemos que quanto menor for a margem de erro, maior será a precisão, mas menor será a confiança, para uma mesma dimensão da amostra. Qualquer intervalo de confiança

é uma solução de compromisso entre confiança e precisão. Então o que se faz é fixar a confiança em determinados valores, tais como 90%, 95% ou 99% e recolher uma amostra de dimensão tal que mantenha a margem de erro dentro de certo limite. Considerando então determinados valores para a margem de erro e para a confiança, vejamos qual a dimensão da amostra necessária:

$$ME = t_{1-\alpha/2}(n-1)\ s/\sqrt{n}$$

de onde

$$n = (t_{1-\alpha/2}(n-1)\ s/ME)^2$$

Na expressão anterior podemos fixar um determinado valor para a margem de erro ME, mas estamos perante algumas situações problemáticas. Não conhecemos s, antes de termos recolhido a amostra e precisamente queríamos conhecer n para recolher a amostra! Normalmente o que se faz nestes casos é fazer um estudo piloto que nos dá uma ideia do valor de s. Aliás esta situação è idêntica à que já nos deparámos quando do estudo do intervalo de confiança para a proporção ou probabilidade **p**, em que era necessário conhecer a estimativa de **p**, $\hat{p}$. E no que diz respeito ao valor de $t_{1-\alpha/2}(n-1)$? Novamente precisamos de conhecer n para calcular o valor do quantil de probabilidade $(1-\alpha)$ de uma t-Student com (n-1) graus de liberdade! Neste caso o que se pode fazer é substituir o quantil da t-Student pelo quantil $z_{1-\alpha/2}$ da N(0,1) e ver qual o valor que vem para n. Se este valor for suficientemente grande, podemos utilizá-lo como dimensão da amostra a recolher, já que os quantis da t-Student e da Normal(0,1) são idênticos. Caso contrário, utilizamo-lo para obter o quantil da t-Student e posteriormente recalcular o valor (de n) a partir da fórmula respectiva.

**Exemplo 3 -** Uma máquina está afinada para produzir peças de um certo comprimento. Todavia, observa-se uma certa variação de comprimento de uma peça para outra, podendo tal comprimento ser considerado uma variável aleatória normal.

a) Suponha que foi extraída uma amostra de 16 peças, tendo sido medido o comprimento de cada uma. Os resultados obtidos foram os seguintes:

$$\sum x_i = 80 \text{ cm} \qquad \sum x_i^2 = 535 \text{ cm}^2$$

Determine um intervalo de 95% de confiança para o valor médio do comprimento das peças.

b) Admita que o verdadeiro valor da variância é igual à estimativa obtida naquela amostra. Determine novo intervalo de confiança, com esta informação adicional. Que conclusões tira?

c) Repita a alínea b) admitindo que a amostra recolhida tinha dimensão 25.

Resolução:

n=16 $\qquad \bar{x} = \frac{80}{16} = 5 \qquad s^2 = \frac{535}{15} - \frac{16 \times 25}{15} = 9$

$t_{.975}(15) = 2.131$

a) Intervalo de confiança $\qquad [5 - 2.131 \times \frac{3}{4}, 5 + 2.131 \times \frac{3}{4}] = [3.40, 6.60]$

b) Intervalo de confiança $[5 - 1.96\times \frac{3}{4}, 5 + 1.96\times \frac{3}{4}]$ = [3.53,6.47]

O intervalo de confiança agora calculado tem uma amplitude inferior à do calculado na alínea a), o que seria de esperar pois dispomos de mais informação.

c) Intervalo de confiança $[5 - 1.96\times \frac{3}{5}, 5 + 1.96\times \frac{3}{5}]$ = [ 3.82, 6.18]

A amplitude do intervalo é inferior à do intervalo calculado na alínea b) pois considerámos ainda mais informação ao dispormos de uma amostra de maior dimensão.

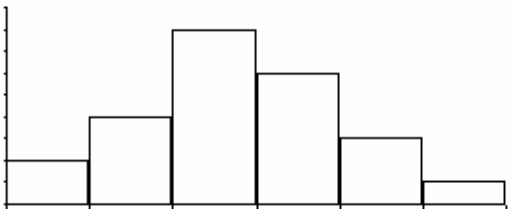
**Exemplo 4** – Numa rua que passa à frente de uma escola, chamada Rua Nova, existe uma passadeira para os peões e um sinal a limitar a velocidade a 50 km por hora. No entanto, a maior parte das vezes, os carros nem sequer abrandam! A polícia, frequentemente, coloca um radar para controlar a velocidade e motivar ao cumprimento daquela regra de trânsito. Os pais das crianças é que não acreditam que esta medida seja suficiente e pretendem que seja colocado um semáforo, que passa a encarnado com velocidade superior aos 50 Km/h. Para poderem ter argumentos perante as instâncias camarárias, resolvem fazer um controlo de velocidades e num certo dia útil, pensam recolher as velocidades médias de alguns dos carros que passarem. Quantos carros devem observar, para obterem um intervalo de confiança de 95%, cuja margem de erro não ultrapasse 2 Km?

Resolução:

Para determinar a dimensão da amostra a recolher, é necessário ter uma ideia de como é que se distribuem as velocidades, nomeadamente se a distribuição dos dados é unimodal e simétrica. Além disso é necessário ter um valor aproximado para a variabilidade. Suponhamos então que se recolheu uma amostra piloto, para recolher a informação necessária:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 50 | 48 | 57 | 57 | 43 | 50 | 48 | 63 | 52 | 42 | 57 | 53 |
| 52 | 45 | 62 | 48 | 43 | 40 | 57 | 60 | 52 | 57 | 60 | 35 |

O histograma que fizémos dos dados mostra que a distibuição é unimodal e aproximadamente simétrica. Não temos razões que nos levem a duvidar da independência dos dados (estamos a admitir que a recolha dos dados não se fez em hora de ponta...).



Para a variância amostral obtivémos o valor de s=7.33. Consideranto o quantil de probabilidade 0.975 da normal, que é igual a 1.96, temos

$$n= \left(\frac{1.96\times 7.33}{2}\right)^2 = 51.5$$

donde necessitamos de uma amostra de dimensão 52. Refazendo os cálculos para a determinação da dimensão da amostra, considerando agora o quantil da t-Student com 51 graus de liberdade, que é igual a 2.008, obtivémos para **n** o valor de 54.

Facilmente se verifica que a margem de erro do intervalo de 95% de confiança, construído com os dados recolhidos para a amostra piloto, é de 3 Km.

### Utilização do Excel para obter quantis da Normal e da t-Student

O Excel não é de grande ajuda na obtenção dos intervalos de confiança. Pode-nos servir unicamente para obter os quantis, nomeadamente da Normal e da t-Student. Assim, para obter:

- $z_{1-\alpha/2}$, faça: *Insert* → *Function* → *NORMSINV* e em *Probability* escreva o valor de $(1-\alpha/2)$;

  Ex: O valor de $z_{.975}$, obtém-se escrevendo em *Probability*, 0,975.

- $t_{1-\alpha/2}(n-1)$, faça: *Insert* → *Function* → *TINV* e em *Probability* escreva o valor de $\alpha$ e em *Deg_freedom* escreva o número de graus de liberbade, ou seja o valor de (n-1).

  Ex: O valor de $t_{.975}$ (51), obtém-se escrevendo em *Probability* 0,05 e em *Deg_freedom*, 51.

Nota: Repare-se na falta de coerência no modo como se obtêm os quantis dos dois modelos, para o mesmo valor da probabilidade. Assim, ao utilizar uma função do Excel, recomenda-se uma leitura atenta das indicações para a utilização da referida função, nomeadamente no que diz respeito ao dignificado dos seus parâmetros.

## Exercícios

**1.** Uma fábrica produz peças, havendo uma certa percentagem de defeituosas. O departamento de controlo de qualidade recolheu uma amostra de 30 peças, encontrando 4 defeituosas. Determine um intervalo de 95% de confiança para a percentagem de peças defeituosas produzidas pela dita máquina.

Qual a dimensão da amostra necessária para obter um intervalo com 95% de confiança, cuja amplitude não exceda .1?

**2.** Perguntou-se a cada um dos 80 estudantes de um determinado curso, qual o seu grau de satisfação relativamente ao curso que frequenta. Obtiveram-se os seguintes resultados:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| NS | MB | B | S | NS | NS | SP | SP |
| NS | B | NS | NS | SP | B | B | MB |
| SP | NS | NS | MB | SP | B | NS | B |
| SP | S | SP | SP | NS | NS | SP | S |
| MB | S | B | MB | NS | S | S | S |
| SP | S | B | NS | S | S | SP | B |
| B | B | MB | NS | B | S | NS | NS |
| B | S | MB | S | MB | NS | MB | SP |
| S | S | NS | B | MB | NS | MB | NS |
| B | MB | SP | MB | S | SP | SP | MB |

NS-"Não Satisfaz"; SP-"Satisfaz Pouco; S-"Satisfaz"; B- "Bom"; MB- "Muito Bom".

a) Faça uma representação gráfica adequada para os dados e indique uma característica amostral.

b) Admitindo que as opiniões destes estudantes são representativas das opiniões dos estudantes dos outros cursos, construa um intervalo de 95% de confiança para a probabilidade de um estudante, escolhido ao acaso, ter uma opinião positiva (Satisfaz, Bom ou Muito Bom) sobre o curso em que está inscrito.

**3 .** Um inquérito realizado a 100 potenciais compradores de um carro novo para o próximo ano, revelou que estão dispostos a pagar em média 14750 euros, com um desvio padrão de 4250 euros.

a) Calcule um intervalo de 95% de confiança para a quantia média que os compradores estão dispostos a pagar.

b) Foi posto à venda um novo tipo de carro, ao preço de 22500 euros. Será que este valor excede significativamente o que os compradores pretendem gastar em média?

**4**. Ao Instituto para a defesa do consumidor têm sido apresentadas queixas, dizendo que as embalagens de determinado produto congelado têm menos peso do que o indicado nas embalagens. Uma recolha preliminar de 40 destas embalagens indicou um peso médio de 975 gramas, com um desvio padrão de 85 gramas. Quantas embalagens devem ser examinadas, de forma a obter uma estimativa do peso médio com erro inferior a 25 gramas, com uma confiança de 95%?

**5**. Os seguintes dados representam o tempo de reacção (em segundos), de 42 indivíduos, a um estímulo luminoso :

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 13.8 | 19.1 | 20.4 | 21.8 | 22.3 | 24.0 | 24.6 | 25.2 | 26.1 | 26.5 | 26.6 | 28.7 |
| 28.8 | 30.2 | 31.2 | 31.7 | 31.7 | 33.6 | 34.6 | 34.8 | 35.4 | 36.0 | 36.3 | 36.8 |
| 37.1 | 38.1 | 40.3 | 40.4 | 41.8 | 42.2 | 42.4 | 43.7 | 43.8 | 44.0 | 44.4 | 44.6 |
| 46.5 | 48.1 | 49.9 | 50.0 | 50.2 | 56.4 | | | | | | |

**a)** Determine as seguintes características amostrais : média, variância, mediana, $Q_{3/5}$ e $Q_{5/14}$.

**b)** Escolha uma amplitude conveniente para o intervalo de classe e construa o histograma correspondente aos dados.

**c)** Construa um intervalo de 99% de confiança para o tempo médio de reacção.

**6**. Os seguintes dados representam o tempo de CPU (em segundos), gastos por um programa que utiliza um determinado software de estatística :

| | | | | | | |
|---|---|---|---|---|---|---|
| 6.2 | 5.8 | 4.6 | 4.9 | 7.1 | 5.2 | 4.4 |
| 8.1 | 3.2 | 3.4 | 4.4 | 8.0 | 7.9 | 3.1 |
| 6.1 | 5.6 | 5.5 | 3.1 | 6.8 | 4.6 | 7.8 |
| 3.8 | 2.6 | 4.5 | 4.6 | 7.7 | 3.8 | 2.9 |
| 4.1 | 6.1 | 4.1 | 4.4 | 5.2 | 1.5 | 5.6 |

**a)** Determine as seguintes características amostrais : média, variância, mediana , $Q_{2/5}$ e $Q_{3/4}$.

**b)** Escolha uma amplitude conveniente para o intervalo de classe e construa o histograma correspondente aos dados.

**c)** Admitindo a normalidade dos dados, construa um intervalo de 95% de confiança para o valor médio dos tempos de CPU gastos pelo programa.

**7** . Recolheu-se uma amostra de 40 alunos a frequentarem o tronco comum de Matemática Aplicada no ano lectivo de 98/99, tendo-se verificado que 10 destes alunos frequentam o curso em 1ª opção.

a) Com base nos resultados determine um intervalo de 95% de confiança para a verdadeira percentagem de estudantes do 1º ano que efectivamente escolheram o curso em 1º opção.

b) Se pretendesse reduzir a metade a amplitude do intervalo obtido anteriormente, com uma amostra da mesma dimensão, qual o maior nível de confiança com que devia trabalhar?

c) Se recolhesse 200 amostras de dimensão 40, a partir das quais construísse outros tantos intervalos de confiança, quantos destes intervalos esperaria que contivessem o verdadeiro valor da percentagem de estudantes que frequentam o curso em 1ª opção?

**8**. Verifique que o intervalo de 90% de confiança para os dados do exemplo 4 é [47.9km; 54.1km]. Explique, porque é que não é correcto dizer o seguinte (Adaptado de De Veaux and al, 2004):

a) *90% de todos os veículos que passam na Rua Nova*, vão a uma velocidade entre 47.9km e 54.1km. (Res: O intervalo de confiança diz respeito à velocidade *média* dos veículos e não à velocidade de cada um dos veículos).

b) Temos uma confiança de 90% de que um *veículo seleccionado aleatoriamente*, vá a uma velocidade entre 47.9km e 54.1km. (Res: Como no caso anterior, estamos a referir-nos a um único veículo, quando, na verdade, estamos 90% confiantes que o intervalo [47.9km; 54.1km] contenha a velocidade *média* de todos os veículos que passam na Rua Nova).

c) A velocidade média dos veículos, é 51km, *90% do tempo*. (Res: esta afirmação dá a ideia que a verdadeira velocidade média varia, quando o que varia é o intervalo, que será diferente, sempre que recolhermos uma amostra diferente).

d) *90% de todas as amostras* têm velocidades médias entre 47.9km e 54.1km. (Res: Esta afirmação dá a ideia de que este intervalo goza de algum privilégio, relativamente a outros. De facto, este intervalo é tão bom ou tão mau, como qualquer dos outos. O que deveremos dizer é que 90% de todas as possíveis amostras permitem construir intervalos que contêm a velocidade média. Nunca saberemos se o nosso intervalo é um dos que contêm ou não).

**Capítulo 11**

# Introdução aos testes de hipóteses

## 11.1 - Introdução

Já vimos um processo de fazer inferência estatística - a estimação, em que utilizámos o modelo Binomial e o modelo Normal. Vamos ainda utilizar o modelo Binomial num outro tipo de inferência estatística a que chamamos **testes de hipóteses**.

O objectivo dos testes de hipóteses, é determinar se uma dada conjectura ou hipótese que fazemos acerca de uma população, é plausível, isto é, tem razão de ser. Precisamente esta plausibilidade é calculada com base na informação obtida a partir de uma amostra da população.

**Exemplo 1** (Teaching Statistics, vol 15, nº1, 1993) - Um professor chega um dia à aula e resolve pôr a seguinte questão: - Há aqui algum aluno que consiga distinguir, pelo sabor, a Coca-Cola da Pepsi-Cola?
Um estudante diz que sim, que consegue distinguir, embora o professor pense que ele efectivamente não o consegue, e se acertar, é por acaso. Depois de alguma discussão em que o aluno afirma que consegue distinguir e o professor diz que ele está a fazer "bluf", resolvem fazer uma aposta, em que apostam uma certa quantia.

Algumas questões que se levantam, relativamente a este problema, são:

**1** - Com que probabilidade consegue o estudante distinguir entre a Coca e a Pepsi?
**2** - Qual o critério que se utiliza para ver quem é o vencedor?
**3** - Usando o critério, a definir em 2:

**a)** *Qual a probabilidade do estudante perder, mesmo que tenha razão*? ( o estudante pode ter acordado mal disposto, estar nervoso, pouco concentrado, …)

**b)** *Qual a probabilidade do estudante ganhar a aposta*, se de facto adivinhou, mas efectivamente *não consegue distinguir* entre a Coca e a Pepsi e responde ao acaso (foi uma questão de sorte…)?

**4** - Quão pequenas devem ser as probabilidades em 3., para que cada um dos apostadores não esteja a correr um risco muito grande?

A perspectiva de levar a cabo a experiência na turma, em frente de toda a gente, é deveras intimidante, pelo que não é de esperar que o voluntário consiga distinguir as duas bebidas 100% das vezes. De modo geral o estudante estabelecerá essa probabilidade entre .7 e .8 como realística. Para o prosseguimento da nossa experiência, vamos admitir que é de **.7**.

Temos agora de delinear a experiência e determinar o **critério** de sucesso para o estudante.

Depois de algumas discussões na aula, o voluntário concorda em provar 15 copos de bebida e dizer se cada uma é Coca ou Pepsi. De acordo com a probabilidade estabelecida anteriormente, para cada prova ele terá uma probabilidade de 70% de dar a resposta correcta.

Qual o *critério* justo, que se deve considerar, para admitir que o estudante tem razão? Ou antes, qual o *critério* justo, que se deve considerar, para que as *duas pessoas que apostaram não estejam a correr um risco demasiado grande*?

Idealmente, gostaríamos que o risco que correm os dois apostadores fossem aproximadamente iguais, isto é, as probabilidades consideradas em 3. deveriam ser aproximadamente iguais.

Com o objectivo de estabelecer um **critério**, o voluntário sugere que **pelo menos 10 respostas certas** significam que tem razão.

Então, P(estudante ganhar a aposta, sabendo distinguir as bebidas) =

= P(nº respostas certas em 15 ser ≥ 10, sabendo que a probabilidade de sucesso é **.7**) =

$$= \sum_{i=10}^{15} \binom{15}{1}.7^{i}.3^{15-i} = .722$$

de onde

P(estudante perder a aposta, sabendo distinguir as bebidas) = 1 - .722 = **.278**

Esta probabilidade de **.278** é o **risco que o estudante** corre.

Qual o risco que o professor corre?

O professor está interessado em calcular a probabilidade de perder o seu dinheiro, se o estudante se limitou a adivinhar e efectivamente não consegue distinguir a Coca da Pepsi. Esta probabilidade é:

P(professor perder a aposta, se o estudante não sabe distinguir as bebidas) =

P(nº respostas certas em 15 ser ≥ 10, sabendo que a probabilidade de sucesso é **.5**) =

$$= \sum_{i=10}^{15} \binom{15}{1}.5^{i}.5^{15-i} = .151$$

Esta probabilidade de **.151** é o **risco que o professor** corre.

<u>Nestas circunstâncias estarão eles dispostos a apostar? E se se aumentar o nº de respostas correctas como critério de ganho ou perca</u>?

**1** - Se se aumentar o nº de respostas correctas necessárias, <u>a probabilidade do estudante perder</u>, embora estando convencido que consegue distinguir, <u>aumenta</u>.

Obs: Se o nº de respostas correctas necessárias fosse k (>10), então o risco que o estudante corria seria

$$P(\text{nº de respostas certas} < k) > P(\text{nº de respostas certas} < 10)(= .278)$$

**2** - Por outro lado, aumentando o nº de respostas correctas necessárias, a probabilidade do professor perder, se o estudante se limita a adivinhar, diminui.

Obs: Se o nº de respostas correctas necessárias fosse k>10, então o risco que o professor corria seria

$$P(\text{nº de respostas certas} \geq k) < P(\text{nº de respostas certas} \geq 10)(= .151)$$

Assim, modificando o critério, estamos a aumentar a probabilidade de um dos tipos de erro e a diminuir a probabilidade do outro tipo de erro.

**3** - Sob a hipótese de que a capacidade de decisão (gustativa…) do estudante continua em forma, aumentando a dimensão da amostra, talvez se consigam diminuir estas probabilidades dos dois tipos de erros.

Por exemplo, se o nº de provas for 20 e o critério para ganhar for de 12 respostas correctas pelo menos, recalculando as probabilidades de cometer os dois tipos de erros, ou sejam, de correr os dois tipos de risco são .113 e .252, respectivamente, para o estudante e o professor.

Se o professor for um "bom desportista" este critério é razoável, se a quantidade de dinheiro posta em jogo não for grande.

Este exemplo servirá para introduzir os conceitos formais de testes de hipóteses, erros de tipo 1 e tipo 2 e as notações associadas com os procedimentos estatísticos. No entanto vamos antes disso, dar outros exemplos de aplicação de testes de hipóteses.

## 11.2 - Outros exemplos

**Exemplo 2** - Numa fábrica de determinadas peças, um lote destas peças é considerado aceitável se tem menos de 8% de peças defeituosas. Já que os lotes têm um grande número de peças, sairia muito caro inspeccionar todas essas peças. A decisão a favor de não rejeitar o lote será tomada no caso de uma amostra a retirar do lote, dar indicação nesse sentido.

**Exemplo 3** - Supõe-se que os estudantes são a favor da avaliação contínua, isto é, mais de 50% dos estudantes preferem a avaliação contínua. Para verificar se existem indícios de que esta hipótese não seja verdadeira, recolhe-se uma amostra de estudantes, registando-se o nº de respostas a favor.

**Exemplo 4** - Um fabricante afirma na garantia que acompanha as lâmpadas que fabrica, que o tempo médio de vida é superior a 450 horas. Ultimamente alguns clientes têm-se queixado das

referidas lâmpadas. Para testar se os clientes têm razão, recolheu-se uma amostra de algumas lâmpadas, registando-se o tempo de vida ( utilizando os chamados testes de vida acelerados, que provocam a falha mais rapidamente).

Todos estes exemplos que acabamos de referir, têm algumas características comuns:

1) Consideram-se **duas hipóteses** complementares acerca de uma **quantidade desconhecida da população**.
2) a informação disponível é dada pela **amostra** que se recolheu da população em estudo.
3) pretende-se verificar se uma das hipóteses a que damos mais importância, é **sustentada** ou **rejeitada** pela informação recolhida da amostra.

No caso 2, por exemplo, as hipóteses a testar são de que o lote é aceitável - $p \leq .08$, ou não - $p > .08$. O que se pretende é verificar que não temos razões para rejeitar a hipótese de que $p \leq .08$.

No caso 3, temos as hipóteses $p \geq .5$ e $p < .5$. O que se pretende testar é se há alguma razão para rejeitar $p \geq .5$.

## 11.3 - Hipótese nula e Hipótese alternativa; erros de tipo 1 e tipo 2; estatística de teste; região de rejeição

Num teste estatístico temos duas hipóteses em alternativa, a que chamamos **hipótese nula ($H_0$)** e **hipótese alternativa ($H_1$)**, sobre um *parâmetro desconhecido* da população. A hipótese nula é a hipótese que reflecte a situação em que não há mudança, sendo pois uma hipótese conservadora e é aquela em que temos mais confiança (resultado de uma experiência passada).

O objectivo de um teste de hipóteses é o de **tomar uma decisão**, no sentido de verificar se existem razões para rejeitar ou não a hipótese nula. Esta decisão é baseada na *informação disponível, obtida a partir de uma amostra, que se recolhe da população*.

No caso em estudo vamos considerar as hipóteses

**$H_0$:** O estudante consegue distinguir Coca da Pepsi

contra (versus)

**$H_1$**: O estudante não consegue distinguir

Estas hipóteses podem-se exprimir em termos da probabilidade de o estudante dar uma resposta correcta

$$\mathbf{H_0}: p = .7 \qquad \text{contra} \qquad \mathbf{H_1}: p = .5$$

A amostra recolhida tem dimensão 15 e vamos utilizar como informação relevante, o nº de respostas correctas, nas 15 provas. Seja **X** a variável aleatória que representa esse número. É esta v.a. que vai permitir tomar uma decisão, recebendo o nome de **estatística de teste**.

Ao tomar uma decisão podemos cometer dois tipos de erros:

- Decidir que o estudante não consegue distinguir, quando efectivamente ele consegue, isto é, **Rejeitar $H_0$, quando $H_0$ é verdadeiro**;
- Decidir que o estudante consegue distinguir, quando efectivamente ele não consegue e responde ao acaso, isto é, **Não rejeitar $H_0$, quando $H_1$ é verdadeiro** .

Ao primeiro erro chamamos erro de tipo 1 e ao segundo, erro de tipo 2. Estes erros são contabilizados em termos de probabilidade.

O **nível de significância** do teste representa-se por $\alpha$ e é o valor máximo para a probabilidade de cometer o erro de tipo 1:

$$\mathbf{P(Rejeitar\ H_0 \mid H_0\ é\ verdadeiro) \leq \alpha}$$

A probabilidade de cometer o erro de tipo 2 representa-se por $\beta$

$$\mathbf{P(Não\ rejeitar\ H_0 \mid H_1\ é\ verdadeiro) = \beta}$$

O risco que o estudante corre é $\alpha$, enquanto que o risco que o professor corre é $\beta$.

O seguinte quadro reflecte a situação verificada quando se realiza um teste de hipóteses:

| | | Situação verdadeira | |
|---|---|---|---|
| | | $H_0$ verdadeira | $H_1$ verdadeira |
| Decisão | Não rejeitar $H_0$ | Decisão correcta | Erro tipo 2<br>P(Erro tipo 2)=$\beta$ |
| | Rejeitar $H_0$ | Erro tipo 1<br>P(Erro tipo 1)=$\alpha$ | Decisão correcta |

Na escolha do teste, o nosso objectivo é controlar o erro de tipo 1, ou seja $\alpha$.

No caso do exemplo o nosso critério de decisão baseou-se na seguinte regra de decisão:

$$\mathbf{Rejeitar\ H_0\ se\ X<10}$$

Considerando a regra anterior vimos que $\alpha$ = **.2784**.

Poderíamos considerar o problema de outra forma, isto é, partir de um determinado nível de significância, e a seguir determinar os valores de **X** que levavam à rejeição de $\mathbf{H_0}$.

Suponhamos que pretendíamos realizar o teste para o nível de significância de **10%**. Os valores possíveis para **X – v.a.** que representa o **nº de respostas correctas**, são todos os inteiros entre 0 e 15. Destes, pretendemos saber quais os que levam a rejeitar $\mathbf{H_0}$, de modo que o que pretendemos é saber qual o valor de c, tal que

$$P(X \leq c \mid X \cap B(15, .7)) \leq .10$$

Obs: Atendendo às hipóteses consideradas somos levados a rejeitar a hipótese nula quando o nº de respostas correctas do estudante for pequeno ($X \leq c$).

Consultando uma tabela da Binomial com parâmetros 15 e .7, verificamos que

$P(X \leq 8) = .1311$
$P(X \leq 7) = .0500$

Então rejeitamos $H_0$ quando o nº de respostas correctas for ≤ 7. Mas para esta região de rejeição a probabilidade de cometer o erro de tipo 2 é

$$P(X \geq 8 \mid X \cap B(15, .5)) = .5, \text{ o qual é muito grande!}$$

## 11.4 - Testes de hipóteses para a proporção p

O exemplo apresentado anteriormente é um caso particular de testes de hipóteses para a proporção **p**, que vamos formalizar seguidamente.

Suponhamos que temos uma população constituída por indivíduos que pertencem a uma de duas categorias, que representamos por A e $A^C$. Representemos por **p** a proporção (desconhecida) de indivíduos que pertencem à categoria A. Pretendemos fazer inferência sobre o parâmetro **p**, pelo que se recolhe da população uma amostra de dimensão n. A estatística de teste que vamos utilizar, para tomar uma decisão, é X - v.a. que representa o nº de indivíduos da amostra que pertencem à categoria A. Na formalização dos testes representamos por $p_0$ o valor da proporção, que se pretende testar.

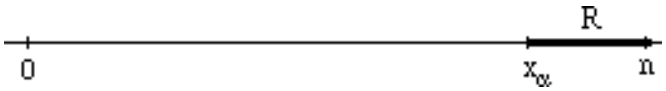
Os testes que vamos realizar são os seguintes:

**1**.

| | | |
|---|---|---|
| $\mathbf{H_0: p = p_0}$ | contra | $\mathbf{H_1: p > p_0}$ |

Rejeitamos $H_0$ quando for elevado o nº de indivíduos da amostra pertencentes à categoria A, ou seja quando $X \geq x_\alpha$. ( Se $H_1$ verdadeiro, ou seja, $p > p_0$, caso em que devemos rejeitar $H_0$, então esperamos encontrar na amostra "muitos" indivíduos pertencentes à categoria A. Entendemos por

"muitos", um número de indivíduos à volta de np, que é superior aos que esperaríamos encontrar caso fosse $H_0$ verdadeiro, ou seja $np_0$).




A determinação do ponto crítico $x_\alpha$ deve fazer-se tendo em atenção o nível de significância $\alpha$, ou seja, vamos calcular o menor inteiro $\mathbf{x_\alpha}$ tal que

$$\mathbf{P[\, X \geq x_\alpha |\, X \cap B(n, p_0)] \leq \alpha}$$

isto é, a região de rejeição R é constituída pelos pontos:

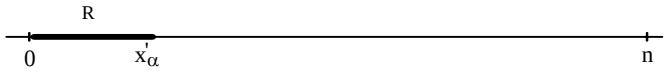
$$R = \{ x \geq x_\alpha |\ [\ \sum_{i=x_\alpha}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i}\ ] \leq \alpha \ \text{ e } \ [\ \sum_{i=x_\alpha-1}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i}\ ] > \alpha \}$$

Obs: A hipótese nula pode-se exprimir na forma $p \leq p_0$ , já que se obtém um teste equivalente.

**2.**

| | | |
|---|---|---|
| $\mathbf{H_0: p = p_0}$ | contra | $\mathbf{H_1: p < p_0}$ |

Rejeitamos $\mathbf{H_0}$ quando for pequeno o nº de indivíduos da amostra pertencentes à categoria A, ou seja quando $X \leq x'_\alpha$.




A determinação do ponto crítico $x'_\alpha$ deve fazer-se tendo em atenção o nível de significância $\alpha$, ou seja, vamos calcular o maior inteiro $\mathbf{x'_\alpha}$ tal que

$$\mathbf{P[\, X \leq x'_\alpha |\, X \cap B(n, p_0)] \leq \alpha}$$

isto é, a região de rejeição R é constituída pelos pontos:

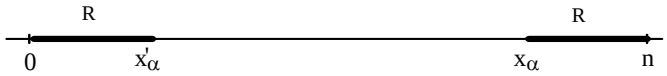
$$R = \{ x \leq x'_\alpha |\ [\ \sum_{i=0}^{x'_\alpha} \binom{n}{i} p_0^i (1-p_0)^{n-i}\ ] \leq \alpha \ \text{ e } \ [\ \sum_{i=0}^{x'_\alpha+1} \binom{n}{i} p_0^i (1-p_0)^{n-i}\ ] > \alpha \}$$

Obs: A hipótese nula pode-se exprimir na forma $p \geq p_0$ , já que se obtém um teste equivalente.

**3.**

| | | |
|---|---|---|
| $\mathbf{H_0: p = p_0}$ | contra | $\mathbf{H_1: p \neq p_0}$ |

Rejeitamos $\mathbf{H_0}$ quando for pequeno ou elevado o nº de indivíduos da amostra pertencentes à categoria A, ou seja quando $X \leq x'_\alpha$ ou $X \geq x_\alpha$.




A determinação dos pontos críticos $x'_\alpha$ e $x_\alpha$ deve fazer-se tendo em atenção o nível de significância $\alpha$. Além disso vamos considerar o chamado teste equilibrado, isto é, atribuir a cada uma das partes da região de rejeição, uma probabilidade igual a metade do nível de significância:

$$\mathbf{P[\, X \le x'_\alpha |\, X \cap B(n, p_0)] \le \alpha/2}$$

e

$$\mathbf{P[\, X \ge x_\alpha |\, X \cap B(n, p_0)] \le \alpha/2}$$

isto é, a região de rejeição R é constituída pelos pontos:

$$R = \{\, x \le x'_\alpha \text{ ou } x \ge x_\alpha \mid [\sum_{i=0}^{x'_\alpha} \binom{n}{i} p_0^i (1-p_0)^{n-i}] \le \alpha/2 \text{ e } [\sum_{i=x_\alpha}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i}] \le \alpha/2 \}$$

Dos três tipos de testes considerados anteriormente, os dois primeiros dizem-se **unilaterais**, enquanto que o último se chama **bilateral**.

**Exemplo 5** - Uma fábrica produz determinado tipo de peças, e sabe-se que a percentagem de defeituosas é de 20%. O director da linha de montagem procedeu a algumas alterações no equipamento, com o objectivo de melhorar a produção, diminuindo nomeadamente a percentagem de peças defeituosas.

Tendo-se recolhido uma amostra de 20 peças, verificou-se que 2 eram defeituosas. Será que há evidência de mudança na percentagem de peças defeituosas?

Resolução:

$$H_0: p \ge .20 \qquad \text{contra} \qquad H_1: p < .20$$

Pretende-se determinar o valor de $x'_\alpha$ tal que

$$P[\, X \le x'_\alpha |\, X \cap B(20, .20)] \le \alpha$$

Consultando uma tabela da Binomial, verificamos que

$$P(X \le 0) = .0115$$
$$P(X \le 1) = .0692$$
$$P(X \le 2) = .2061$$

donde concluímos que:

Se $\alpha = 5\%$ $R=\{0\}$

Se $\alpha = 10\%$ $R=\{0,1\}$

Decisão: Para os níveis usuais de significância, não se deve rejeitar $H_0$, isto é não há evidência de ter havido alteração (para melhor) no processo de fabrico.

### 11.4.1 - Determinação dos pontos críticos $x'_\alpha$ e $x_\alpha$ para grandes amostras

A determinação dos pontos críticos $x'_\alpha$ e $x_\alpha$ dos testes anteriores, pode fazer-se consultando as tabelas com a distribuição Binomial. Pode no entanto acontecer que o valor de n seja demasiado grande, e já não conste nessas tabelas. Então faz-se uma aproximação à Normal, como se descreve a seguir.

Tendo em consideração o teorema do limite central, sabe-se que a distribuição Binomial pode ser aproximada pela distribuição Normal, isto é, se $X \cap B(n,p)$, então

$$P(X \le x) \approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right)$$

Considera-se a estatística de teste

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$$

cuja distribuição pode ser aproximada por uma Normal(0,1) e a determinação dos pontos críticos, para os três tipos de testes considerados anteriormente, faz-se da seguinte forma:

**1**. $P[X \ge x_\alpha | X \cap B(n, p_0)] \le \alpha$ sendo $x_\alpha$ o menor inteiro tal que

$$x_\alpha \ge 1 + np_0 + z_{1-\alpha}\sqrt{np_0(1-p_0)})$$

**2**. $P[X \le x'_\alpha | X \cap B(n, p_0)] \le \alpha$ sendo $x'_\alpha$ o maior inteiro tal que

$$x'_\alpha \le np_0 + z_\alpha\sqrt{np_0(1-p_0)} \quad \text{ou} \quad x'_\alpha \le np_0 - z_{1-\alpha}\sqrt{np_0(1-p_0)})$$

**3**. $P[X \le x'_\alpha | X \cap B(n, p_0)] \le \alpha/2$ e $P[X \ge x_\alpha | X \cap B(n, p_0)] \le \alpha/2$

$$x'_\alpha \le np_0 - z_{1-\alpha/2}\sqrt{np_0(1-p_0)} \text{ e } x_\alpha \ge 1 + np_0 + z_{1-\alpha/2}\sqrt{np_0(1-p_0)})$$

($x'_\alpha$ maior inteiro e $x_\alpha$ menor inteiro satisfazendo respectivamente cada uma das desigualdades anteriores).

Observação – Uma alternativa, equivalente, à estatística de teste X, com distribuição aproximadamente Normal($np_0, \sqrt{np_0(1-p_0)}$), sob $H_0$, é a estatística $\hat{p} = \frac{X}{n}$, com distribuição aproximadamente N($p_0$, $\sqrt{\frac{p_0(1-p_0)}{n}}$).

### 11.4.2 - P-value

Ao realizar um teste de hipóteses, podemos à partida não ter especificado um nível de significância. Então, um processo alternativo para a realização dos testes anteriores é, face ao valor observado $\mathbf{x}_0$ da estatística de teste **X**, calcular a seguinte probabilidade:

Caso 1 : $\mathbf{P = P[X \geq x_0 \,|X \cap B(n, p_0)]}$

Caso 2 **:** $\mathbf{P = P[X \leq x_0 \,|X \cap B(n, p_0)]}$

Caso 3 : $\mathbf{P = 2 \min \{P[X \geq x_0 \,|X \cap B(n, p_0)], P[X \leq x_0 \,|X \cap B(n, p_0)]\}}$

Esta probabilidade **P** é o menor valor para o nível de significância que levaria à rejeição da hipótese nula, para a amostra recolhida. A P chamamos *P-value*. Assim, <u>para tomarmos uma decisão, calcula-se o P-value e para um dado nível de significância $\alpha$, rejeita-se a hipótese nula se</u>

$$\mathbf{P \leq \alpha}$$

A metodologia seguida neste caso é diferente da seguida anteriormente, em que para tomarmos uma decisão era necessário especificar à partida o nível de significância com que pretendíamos realizar o teste, de forma a calcular a região de rejeição. Se o valor observado da estatística de teste pertencesse a essa região então rejeitaríamos a hipótese nula. Agora calculamos o P-value e não é necessário calcular explicitamente a região de rejeição, pois se para um determinado nível de significância se verifica que $\mathbf{P \leq \alpha}$, isto significa necessariamente que o valor observado da estatística de teste pertence à região de rejeição.

## 11.5 - Vamos conversar acerca de testes

Ao longo deste texto já temos referido várias vezes que é objectivo da Estatística arranjar modelos probabilísticos que sirvam para modelar situações do mundo real. Ao formular uma *hipótese* (hipótese nula), como as que formulámos anteriormente, não estamos mais que a propor um modelo para uma situação real. Uma vez o modelo proposto, vamos recolher informação - os *dados*, para averiguar da consistência do modelo. Então, defrontamo-nos com duas situações:

- ou os dados são consistentes com o modelo, e nesse caso *não vemos razão para o rejeitar*,
- ou os dados contradizem fortemente o modelo, e neste caso *pensamos que há evidência para o rejeitar*.

Repare-se que na primeira situação, não dissemos que os dados mostravam que a hipótese é verdadeira! Só dissémos que não víamos razão para a rejeitar. Esta situação é análoga à que se passa nos tribunais – tem que se começar por admitir a presunção de inocência e cabe ao juiz, mostrar que os factos contradizem esta presunção, para admitir a culpabilidade. Na segunda situação, dissémos que pensamos que há evidência para rejeitar o modelo. Mas fica-nos sempre a dúvida se deveremos tomar essa opção, já que rejeitar o modelo proposto, se ele fosse efectivamente verdadeiro, pode acarretar grandes prejuízos. Então precisamos de quantificar essa decisão e essa quantificação é feita probabilisticamente. Assim, calculamos a probabilidade de obter dados como os recolhidos, baseando-nos em que o modelo é verdadeiro. Se esta probabilidade for muito pequena, pensamos que não foi só o acaso, isto é a aleatoriedade

presente na recolha da informação, que nos levou a obter aqueles dados, mas naturalmente é o modelo que não é o correcto, pois "essa probabilidade é demasiado pequena, para ser verdade", e rejeitamos esse modelo. Esta tal probabilidade – *p-value*, dá-nos uma medida do erro que cometemos ao rejeitar o modelo proposto, e quanto menor for, maior será a evidência contra o modelo.

Assim, quando não rejeitamos a hipótese nula, ficamos sempre na dúvida, sobre se terá sido o teste que não teve capacidade para a rejeitar, mesmo sendo ela falsa. Justifica-se, assim, que se procure calcular a probabilidade de se rejeitar a hipótese nula, quando ela é falsa, isto é, P(Rejeitar$H_0$|$H_1$verdadeira). A esta probabilidade chama-se **potência** do teste. Repare-se que para um determinado valor do parâmetro especificado na hipótese alternativa:

Potência do teste = 1 – P(erro de tipo 2)

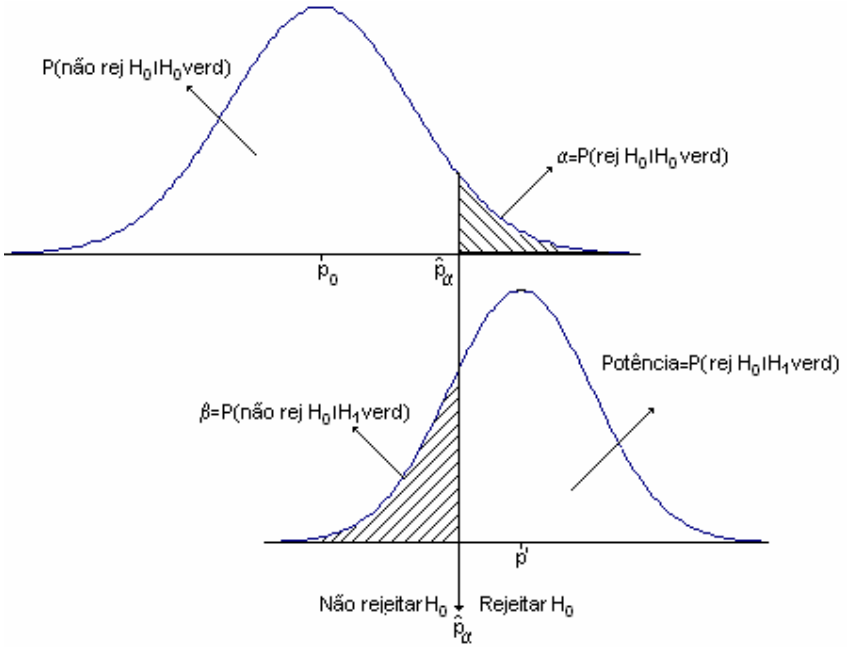
Então, de um modo geral, podemos dizer que pretendemos um teste com nível de significância pequeno e potência grande.

Para visualizar melhor a relação entre estes conceitos, vamos colocar-nos na situação de estarmos a realizar o seguinte teste:

H0: $p=p_0$ contra H1: $p>p_0$

Rejeitamos $H_0$ para valores grandes de X ou de forma equivalente, para valores grandes de $\hat{p}=\frac{X}{n}$, nomeadamente para valores de $\hat{p} \geq \hat{p}_\alpha$, onde $\hat{p}_\alpha = \frac{x_\alpha}{n}$, utilizando notação já introduzida anteriormente.

Se n for grande, a distribuição da estatística de teste é aproximada pela Normal, pelo que temos:



Na figura anterior apresentamos a distribuição da estatística de teste, para o caso de $H_0$ ser verdadeira (Normal superior) e para um valor específico do parâmetro (p'), no caso de ser $H_1$ verdadeira (Normal inferior). A região de rejeição é dada pelo intervalo $[\hat{p}_\alpha, 1]$. Algumas conclusões são evidentes da figura anterior:

- Quanto mais $\hat{p}_\alpha$ estiver para a direita, isto é, menor for o nível de significância do teste, ou a probabilidade de cometer o erro de tipo 1, maior será a probabilidade de cometer o erro de tipo 2. Assim, não é possível minimizar os dois erros ao mesmo tempo, a não ser aumentando a dimensão da amostra. Efectivamente, se se aumentar a dimensão da amostra recolhida, as normais ficam mais "magras", já que a variância diminui;
- Quanto menor for o erro de tipo 2, maior será a potência do teste;
- No caso de $H_0$ ser falsa, a potência do teste será tanto maior, quanto mais afastado de $p_0$, estiver o verdadeiro valor da proporção p (a Normal de baixo afasta-se para a direita).

Formalizando um pouco o que dissémos anteriormente, para realizar um teste de hipóteses, em que as hipótese são quase sempre sobre parâmetros de modelos, é necessário:

- Formular uma hipótese nula $H_0$, que é aquela que reflecte a situação em que não há mudança e em que assumimos um valor para o parâmetro no modelo proposto, e uma hipótese alternativa $H_1$, que reflecte a situação que pensamos ser verdadeira, no caso de não o ser a hipótese nula;
- Arranjar uma estatística de teste, que sirva para medir a discrepância entre o que se observa nos dados e o que se espera quando se considera a hipótese nula (isto é, uma estatística cuja distribuição de amostragem seja conhecida no caso da hipótese nula ser verdadeira, pois a discrepância é medida em termos de probabilidade);
- Face à amostra que entretanto se recolheu, calcular o p-value;
- Tomar uma decisão, que se exprimirá na seguinte forma:

  Rejeitar $H_0$, para o nível de significância $\alpha$ ou

  Não rejeitar $H_0$ para o nível de significância $\alpha$.
- Se tivermos possibilidade de escolher entre vários testes, então para o mesmo nível de significância, deve-se escolher o de potência máxima;
- Se tivermos possibilidade de recolher amostras de dimensão maior, melhor será, pois reduzimos as probabilidades de cometer erros, ao tomar uma decisão, aumentando também a potência do teste.

## 11.6 - Testes de hipóteses sobre o valor médio

Da mesma forma que realizámos testes de hipóteses sobre o parâmetro **p**, também se podem realizar sobre o valor médio **μ**, desconhecido, de uma população. A metodologia a seguir é a mesma, mas agora temos de considerar outra estatística de teste, sendo natural considerar a média ou uma função da média para fazer inferência estatística sobre o valor médio.

Consideremos, por exemplo, um industrial de componentes electrónicas, que afirma que o tempo médio de vida das componentes que fabrica é de 560 horas. Um cliente acha este tempo exagerado, pois tem tido mau resultado com este tipo de material. Então o industrial está interessado em testar que o valor médio da distribuição do tempo de vida das componentes é igual a 560 horas, ou seja de que tem razão. Temos assim uma conjectura ou hipótese sobre a população e que em testes de hipóteses se refere como **Hipótese nula** e se representa por $\mathbf{H_0}$. No entanto a hipótese anterior vai ser testada *contra* uma **Hipótese alternativa** que se representa por $\mathbf{H_1}$, que reflicta a situação que será verdadeira, no caso de não o ser a hipótese nula. Concretamente, no exemplo anteriormente considerado temos as seguintes hipóteses a serem testadas (representando por μ o valor médio da população):

$$\mathbf{H_0}: \mu=560 \text{ horas} \qquad \text{contra} \qquad \mathbf{H_1}: \mu<560 \text{ horas}$$

Escolhemos a hipótese alternativa anterior, pois ela reflecte a situação real, no caso de não se provar que $H_0$ é verdadeira(estamos a pensar nas queixas dos clientes).

Vamos exemplificar a realização de um teste de hipóteses sobre o valor médio através do exemplo dos pesos, referido quando abordámos o problema da estimação.

Suponhamos que estamos interessados em realizar um teste sobre o peso médio da população, constituída pelos indivíduos de 1.65 m de altura, tendo sido levantadas algumas dúvidas sobre se o peso seria de 60 kg. Então

$$\mathbf{H_0}: \mu = 60 \text{ kg} \qquad \text{contra} \qquad \mathbf{H_1}: \mu \neq 60 \text{ kg}$$

Formulamos a hipótese alternativa deste modo já que à partida não tínhamos qualquer informação que nos levasse a considerar quer um valor médio superior, quer um valor médio inferior a 60 kg.

Pensemos na seguinte estatística de teste

$$T = \frac{\mathbf{X}-\mu}{\sigma/\sqrt{\mathbf{n}}}$$

cuja distribuição é conhecida se $H_0$ verdadeira. É fundamental conhecer a distribuição de T, no caso de $H_0$ ser verdadeira, pois esse facto vai-nos permitir determinar a **região de rejeição** $\mathcal{R}$ a partir do nível de significância $\alpha$, definido à priori,

$$P[T \varepsilon \mathcal{R} \mid H_0 \text{ verdadeira}] = \alpha \quad \text{ou seja}$$

$$P[|T_0| > z_{1-\alpha/2}] = \alpha$$

onde $T_0$ se obtém de T substituindo μ por $\mu_0$, sendo $\mu_0$ o valor de μ considerado na hipótese nula (no caso do exemplo $\mu_0$=60).

Então a zona de rejeição é dada pelo seguinte intervalo

$$\mathcal{R} = ]-\infty , -z_{1-\alpha/2}[ \cup ] z_{1-\alpha/2} , +\infty[$$

pelo que se rejeita a hipótese $H_0$, sempre que $t_0 \varepsilon \mathcal{R}$, sendo $t_0$ o valor observado da estatística de teste. Considerando, no exemplo, a amostra correspondente à cidade de Lisboa, temos:

$$t_0 = \sqrt{10}(60.82 - 60)/4 = .65$$

pelo que trabalhando com um nível de significância de 5%, se tem a seguinte região de rejeição:

$$\mathcal{R} = ]-\infty , -1.96[ \cup ] 1.96 , +\infty[$$

Como .65 não pertence à região de rejeição, não vemos razão para rejeitar a hipótese nula.

Ainda no exemplo que temos vindo a considerar, suponhamos que tínhamos começado por recolher a amostra referente à cidade de Évora. Mediante o resultado obtido, uma pessoa mais céptica teria razões para suspeitar que o peso médio seria inferior a 60 kg. Nestas circunstâncias deveríamos proceder ao seguinte teste:

$$\mathbf{H_0}: \mu = 60 \qquad \text{contra} \qquad \mathbf{H_1}: \mu < 60$$

Agora a hipótese alternativa especifica que o valor médio é inferior a 60 e se a hipótese $H_1$ fosse verdadeira esperaríamos obter amostras que levassem a um valor negativo para $t_0$ (porquê?). Quer dizer que vamos rejeitar a hipótese nula se $t_0 < z_\alpha$ , pois

$$P [ T < z_\alpha \mid H_0 \text{ verdadeira}] = \alpha$$

Para o nível de significância de 5% o quantil $z_{.05}$ = -1.6449, pelo que a região de rejeição é

$$\mathcal{R} = ]-\infty , -1.6449[$$

Considerando então a amostra observada em Évora obtemos

$$t_0 = \sqrt{10}(57.04 - 60)/4 = -2.34$$

valor que pertence à região de rejeição, donde concluímos que devemos rejeitar a hipótese nula.

Pensemos agora na cidade de Braga, em que temos razões para suspeitar que os pesos são mais altos (come-se muito bem no Norte..., o que não quer dizer que se coma mal em Évora...), pelo que consideramos o seguinte teste:

$$\mathbf{H_0} : \mu = 60 \qquad \text{contra} \qquad \mathbf{H_1} : \mu > 60$$

Neste momento a hipótese alternativa indica-nos que a zona de rejeição corresponderá a valores de $t_0$ demasiado grandes, ou seja $t_0 > z_{1-\alpha}$. Como

$$t_0 = \sqrt{10}(61.22 - 60)/4 = .96 \qquad \text{e} \qquad z_{.95} = 1.6449$$

não vemos razão para rejeitar a hipótese nula.

Repare-se que dos 3 testes considerados anteriormente, o 1º é de índole diferente dos outros dois, no que diz respeito à hipótese alternativa - no 1º caso estamos perante um teste **bilateral** enquanto que os outros 2 se referem a testes **unilaterais**.

O exemplo anterior pode-se inserir num processo mais geral de testar hipóteses sobre o valor médio, que podemos resumir do modo seguinte:

**1º caso**

**Dados**: É dada uma amostra $(x_1,x_2\ldots,x_n)$, valor observado da amostra aleatória $(X_1,X_2,\ldots,X_n)$ em que $n\geq 30$.

Se a população tem distribuição aproximadamente normal e variância conhecida, então a dimensão da amostra pode ser inferior a 30.

**Hipótese nula: $H_0$**: $\mu = \mu_0$

(Esta hipótese nula é equivalente a $\mu\leq\mu_0$ ou $\mu\geq\mu_0$ conforme se utilizem as hipóteses alternativas b) e c) especificadas a seguir)

**Estatística de teste**: $T_0= \sqrt{n}(\overline{X}-\mu_0)/\sigma$

Obs. Para $n\geq 30$, quando $\sigma$ é desconhecido, pode ser estimado por s.

| **Hipótese alternativa** | *Decisão a tomar para um nível de significância $\alpha$* |
|---|---|
| **a)** $H_1$: $\mu \neq \mu_0$ | Rejeita-se $H_0$ se $\lvert t_0\rvert > z_{1-\alpha/2}$ |
| **b)** $H_1$: $\mu > \mu_0$ | Rejeita-se $H_0$ se $t_0 > z_{1-\alpha}$ |
| **c)** $H_1$: $\mu < \mu_0$ | Rejeita-se $H_0$ se $t_0 < z_\alpha$ |

**2º caso**

**Dados**: É dada uma amostra $(x_1,x_2\ldots,x_n)$, valor observado da amostra aleatória $(X_1,X_2,\ldots,X_n)$ de uma população com distribuição Normal e parâmetro $\sigma$ desconhecido.

**Hipótese nula** - $H_0$: $\mu = \mu_0$

(Esta hipótese nula é equivalente a $\mu\leq\mu_0$ ou $\mu\geq\mu_0$ conforme se utilizem as hipóteses alternativas b) e c) especificadas a seguir)

**Estatística de teste** : $T_1 = \sqrt{n}(\overline{X}-\mu_0)/S$

| **Hipótese alternativa** | *Decisão a tomar para um nível de significância $\alpha$* |
|---|---|
| **a)** $H_1$: $\mu \neq \mu_0$ | Rejeita-se $H_0$ se $\lvert t_1\rvert > t_{1-\alpha/2}(n-1)$ |
| **b)** $H_1$: $\mu > \mu_0$ | Rejeita-se $H_0$ se $t_1 > t_{1-\alpha}(n-1)$ |
| **c)** $H_1$: $\mu < \mu_0$ | Rejeita-se $H_0$ se $t_1 < t_\alpha(n-1)$ |

onde representamos por $t_\alpha(n-1)$ o quantil de probabilidade $\alpha$ da distribuição t de Student com (n-1) graus de liberdade.

**Obs**. Quando a dimensão da amostra for suficientemente grande, a distribuição da estatística $T_1$ é aproximadamente normal, pelo que podemos tratar o segundo caso de modo análogo ao 1º caso. Na realidade, à medida que a dimensão da amostra aumenta e consequentemente o número de graus de liberdade, a distribuição t de Student aproxima-se da distribuição Normal.

### 11.6.1 - P-value

Um processo alternativo de realizar os testes de hipóteses anteriores é calcular o P-value. Mais concretamente, para cada uma das situações consideradas anteriormente, face ao valor observado $t_0$ ou $t_1$ das estatísticas de teste $T_0$ ou $T_1$, calcula-se:

1º caso:

a) $P = 2 \min \{P[T_0 \leq t_0], P[T_0 \geq t_0]\}$

b) $P = P[T_0 \geq t_0]$

c) $P = P[T_0 \leq t_0]$

2º caso:

a) $P = 2 \min \{P[T_1 \leq t_1], P[T_1 \geq t_1]\}$

b) $P = P[T_1 \geq t_1]$

c) $P = P[T_1 \leq t_1]$

Decisão: Para um determinado nível de significância $\alpha$, rejeita-se a hipótese nula quando $P \leq \alpha$.

### Exercícios

**1.** Admita que a mediana da nota da PE, dos alunos que entraram no ano lectivo 91/92 foi de 35. Com base na amostra anterior, verifique se existem razões para suspeitar de que os alunos que entraram no ano lectivo de 92/93, têm tendência para terem notas mais fracas.

Obs. Considere que a população a estudar está dividida em duas categorias: a dos alunos com nota superior a 35 e dos alunos com nota inferior ou igual a 35.

**2.** Supõe-se que numa população existem três vezes mais indivíduos não fumadores do que fumadores.

**a)** Tendo-se recolhido uma amostra de 20 indivíduos, verificou-se que 7 eram fumadores. Teste, ao nível de significância de 5% se a suposição tem razão de ser.

**b)** Na população anterior pretende-se estudar a incidência de doença pulmonar. Sabe-se que a percentagem de doentes entre os fumadores e não fumadores é respectivamente de 60% e 20%.

**(i)** Determine a probabilidade de um indivíduo ter doença pulmonar.

**(ii)** Determine a probabilidade de um doente pulmonar ser fumador.

**(iii)** Qual a probabilidade de numa amostra de 10 doentes, pelo menos três serem fumadores?

**(iv)** Qual a probabilidade de numa amostra de 225 doentes, mais de metade serem fumadores?

**3.** O sr. X não consegue chegar a horas ao emprego. Todos os dias marca o ponto depois da hora estipulada para a sua entrada. No final do mês, juntamente com uma repreensão escrita, recebeu uma folha com um registo dos seus atrasos (em minutos):

| | | | | | | |
|---|---|---|---|---|---|---|
| 0.01 | 2.66 | 3.30 | 3.77 | 4.47 | 5.13 | 7.56 |
| 8.79 | 10.26 | 14.36 | 15.29 | 19.64 | 21.45 | 28.41 |

**a)** Investigue a existência de possíveis outliers na amostra.

**b)** O sr. X acha injusta a repreensão, já que segundo diz, desde que trabalha naquela empresa, mais de 50% das vezes o atraso é inferior a 5 minutos. Com base nos dados anteriores verifique se existe evidência suficiente para dar razão ao sr. X.

**4.** Suponha que uma amostra recolhida de rendimentos de famílias de determinada cidade revelou que 55% dos rendimentos da população se situam entre os 60 e os 120 contos. O presidente da câmara considera-a **"ideal"** !

Desconfia-se que o bairro X não segue a distribuição "ideal" da cidade. Recolheu-se uma amostra de valores de rendimentos familiares nesse bairro, tendo-se obtido os seguintes resultados:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 15 | 24 | 36 | 55 | 58 | 62 | 65 | 67 | 70 | 71 |
| 73 | 76 | 89 | 90 | 92 | 97 | 105 | 112 | 118 | 160 |

Verifique se esta suspeita tem razão de ser.

**5.** Um grupo de 20 indivíduos hipertensos, foi submetido durante 30 dias a um regime de dieta sem sal. Apresentam-se a seguir os valores da pressão sistólica para esses indivíduos:

| sexo | Antes da dieta | Depois da dieta |
|---|---|---|
| M | 17.0 | 15.6 |
| M | 17.7 | 16.6 |
| M | 17.9 | 16.9 |
| F | 18.1 | 15.6 |
| F | 18.1 | 16.0 |
| M | 18.2 | 15.5 |
| F | 18.3 | 16.5 |
| M | 18.4 | 17.2 |
| M | 18.4 | 15.0 |
| F | 18.5 | 17.5 |
| F | 18.5 | 15.9 |
| F | 18.6 | 16.2 |
| M | 18.7 | 17.5 |
| M | 18.8 | 15.8 |
| F | 18.9 | 17.2 |
| M | 19.2 | 17.3 |
| M | 19.3 | 17.8 |
| F | 19.5 | 16.0 |
| F | 19.8 | 16.9 |
| F | 20.1 | 17.5 |

**a**) Um especialista afirma que após um mês em regime de dieta sem sal, pelo menos 80% dos indivíduos apresenta uma diminuição da pressão sistólica superior a 10%. Averigúe se existem razões para duvidar da afirmação do especialista.

**6.** Recolheu-se a opinião de 20 executivos acerca de máquinas fotocopiadoras, verificando-se que 15 preferiam a marca Kodac relativamente à marca Xerox. Pensa-se, no entanto, que na

realidade não existem diferenças significativas entre as máquinas, pelo que a probabilidade de cada uma ser escolhida é de 50%. Poderíamos assim considerar as seguintes hipóteses a testar:

$$H_0: p=.5 \qquad \text{contra} \qquad H_1: p\neq .5$$

em que representamos por p a probabilidade de ser escolhida a máquina Kodac. Se para 20 executivos consultados, representar por X o número dos que preferem Kodac, considere a seguinte regra de decisão:

rejeito $H_0$ se $X<6$ ou se $X>14$

a) Qual o nível de significância associado ao teste anterior?

b) Qual a decisão a tomar relativamente á amostra considerada?

c) Para as hipóteses $H_0$ e $H_1$ especificadas, qual a regra de teste se efectivamente o número de executivos que constituem a amostra fosse de 50, considerando o nível de significância de 5%?

**7.** Admite-se que a quantidade de nicotina (medida em mg.) existente numa dada marca de cigarros, tem distribuição normal. Observaram-se 5 cigarros da referida marca tendo-se obtido:

16 16.5 19 15.4 15.6

O fabricante afirma que a quantidade média de nicotina , por cigarro, é de 13.5 mg.

**a)** Teste, ao nível de significância $\alpha$= 0.10 a hipótese:

$$H_0 : \mu = 13.5 \qquad \text{contra} \qquad H_1 : \mu > 13.5$$

**b)** Determine um intervalo de 95% de confiança para a quantidade média de nicotina existente em cada cigarro.

**8.** O departamento de controlo de qualidade de uma fábrica de conservas, está na disposição de mandar reajustar todo o equipamento, caso se verifique que o peso médio de cada lata é inferior ao especificado na embalagem. Nomeadamente no caso das latas de sardinha, especifica-se que este peso seja de 150 gramas. Com o objectivo de tomar uma decisão, procedeu-se à recolha de algumas latas de sardinha, que se pesaram, usando-se a média $\bar{X}$, como estatística de teste.

a) Formule as hipóteses nula e alternativa, em termos do valor especificado para o peso médio.

b) Tendo em consideração as consequências que advêm de cometer um erro de tipo I, deverá escolher um nível de significância grande ou pequeno? Justifique convenientemente a sua resposta. Qual ou quais os valores que escolheria?

c) Admitindo que o peso das latas de sardinha se distribui de acordo com uma Normal e que os valores observados para os pesos de uma amostra de 10 latas foram ( em gramas):

147 152 145 130 155 148 150 149 146 149

qual a decisão que o gerente da fábrica deve tomar, no que diz respeito ao reajustamento do equipamento?

**9**. Um médico receita aos seus doentes um medicamento para diminuir o número de pulsações por minuto. Recolheu o nº de pulsações a doentes medicados, que já tomam o medicamento há um mês, tendo obtido o seguinte output, obtido através de um software de Estatística, em que seleccionou como opção, utilizar a distribuição t-Student:

Com 95% de confiança: 70.887604 < $\mu$ <74.497011

a) Quais as hipóteses que o investigador teve de admitir para tomar a opção de seleccionar a distribuição t-Student?
b) Explique o que significa o output anterior.
c) Qual a margem de erro do intervalo?
d) Se o intervalo fosse calculado com uma confiança de 99%, a margem de erro aumentaria ou diminuiria?

**10**. Durante um cateterismo para detectar a a extenção da doença cardíaca, verificando o estado das artérias, é introduzido um pequeno tubo, o catéter, através de uma artéria da perna. É importante que catéter tenha um diâmetro de 2.00mm, em média, com um desvio padrão muito pequeno. O processo de fabrico dos cateteres é submetido a um rigoroso controlo de qualidade, de modo que todos os dias são recolhidas algumas medidas, para testar a hipótese nula $H_0$: m = 2.00mm, contra a hipótese alternativa m$\neq$2.00mm, com um nível de significância de 5%, para parar o processo de fabrico, no caso de haver alterações.

a) Estamos perante um teste unilateral ou bilateral? Porque é que isto é importante, no contexto do problema?

b) Explicar o que é que acontece se o departamento de controlo de qualidade cometer um erro de tipo 1?

c) E se cometer um erro de tipo 2?

**11**. Uma fábrica de bolachas com pedacinhos de chocolate, ao anunciar as suas bolachas diz que cada pacote de meio quilo contém, pelo menos, 1000 pedacinhos de chocolate. Os estudantes do Departamento de Estatística de determinada Universidade, decidiram comprar alguns destes pacotes e contar o número de pedacinhos de chocolate, tendo obtido os seguintes resultados:

1219 1214 1087 1200 1419 1121 1325 1345 1244 1258 1356 1132
1191 1270 1295 1135

a) Verifique se estão cumpridas as condições para poder fazer infereência.

b) Obtenha um intervalo de 95% de confiança para o número médio de pedacinhos de chocolate, em cada pacote.

c) O que é que pode concluir sobre o que diz a empresa que vende as bolachas? Utilize o intervalo anterior para testar uma hipótese apropriada para tirar conclusões.

Sugestão: Verifique que o intervalo de confiança é (1187.9, 1288.4) e de seguida calcule P(X<1000) tendo em consideração que a v.a.X, que representa o nº de pedacinhos de chocolate em cada pacote, tem distribuição Normal N(1188, 94) ou N(1288, 94).

**Capítulo 12**

# Introdução aos testes de ajustamento

## 12.1 - Introdução

Será que existe alguma razão para dizer que os nascimentos são influenciados pelas fases da Lua? Será que o signo influencia o futuro, mais ou menos brilhante, de cada indivíduo? Será que é verdade o que a empresa das drageias M&M afirma, sobre as percentagens de cores das drageias em cada embalagem?

A revista Fortune (De Veaux and al, 2004) recolheu os signos de 256 presidentes de 400 das maiores empresas, tendo obtido a seguinte informação:

| | | | |
|---|---|---|---|
| Carneiro | 23 | Balança | 18 |
| Touro | 20 | Escorpião | 21 |
| Gémeos | 18 | Sagitário | 19 |
| Caranguejo | 23 | Capricórnio | 22 |
| Leão | 20 | Aquário | 24 |
| Virgem | 19 | Peixes | 29 |

Na tabela anterior verifica-se que o signo dos Peixes sobressai com maior número de nascimentos, mas será esta diferença suficiente para dizer que os indivíduos que nascem sob este signo têm maior probabilidade de sucesso? Se os nascimentos se distribuissem uniformemente, esperaríamos aproximadamente 21.3 (256/12) nascimentos em cada signo. De que modo é que os valores observados se "ajustam" à **hipótese** (nula) de que os nascimentos se distribuem uniformemente ao longo do ano? Neste caso já não temos, como no capítulo anterior, um teste sobre uma proporção, mas sim sobre 12 proporções, uma para cada signo, pelo que precisamos de arranjar um teste que nos dê uma ideia global sobre *se as proporções observadas diferem muito das conjecturadas* (consideradas na hipótese nula).

## 12.2 – Generalização do modelo Binomial: o modelo Multinomial

Vimos que no caso dos testes sobre a proporção tínhamos como base o modelo Binomial, em que o parâmetro p, era a probabilidade sobre a qual se pretendia fazer inferência estatística. Neste momento já não temos em estudo uma característica da população com probabilidade p, mas admitimos que a população pode ser dividida em $k \geq 2$ categorias disjuntas $A_1, A_2, ..., A_k$, sendo $p_i$, com i=1, ..., k, a proporção de indivíduos pertencentes à classe $A_i$, e $p_1 + p_2 + ... + p_k = 1$. Fazer inferência estatística acerca desta população, resume-se a estudar os parâmetros $p_i$.

A generalização do modelo Binomial é o chamado modelo Multinomial, que consiste no seguinte: Consideram-se n provas idênticas:

- O resultado de cada prova pode pertencer a uma de k classes possíveis $A_1, A_2, ..., A_k$;
- A probabilidade de que o resultado pertença à classe $A_i$ é $p_i$ e é sempre a mesma de prova para prova;
- As provas são independentes;
- As variáveis de interesse são $O_1, O_2, ..., O_k$, em que $O_i$ é o número de vezes que o resultado pertence à classe $A_i$ e $O_1+O_2+...+O_k=n$.

Diz-se que a variável aleatória $(O_1, O_2, ..., O_k)$ é uma v.a. Multinomial de parâmetros n e pi, i=1,...,n, e tem-se

$$P(O_1=o_1, O_2=o_2, ..., O_k=o_k) = \frac{n!}{o_1!o_2!...o_k!} p_1^{o_1} p_2^{o_2} ... p_k^{o_k} \quad \text{com} \quad oi = 0, 1, ..., n; i=1, 2, ..., k.$$

Observação: Repare-se que se k=2, estamos no caso Binomial.

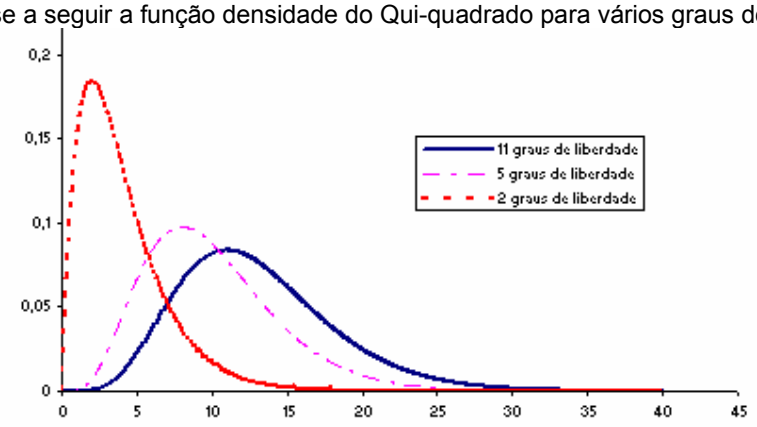
A seguinte propriedade da v.a. Multinomial, tem especial importância para a obtenção de um teste, para testar o ajustamento pretendido na secção anterior.

**Propriedade** – Se $(O_1, O_2, ..., O_k)$ é uma v.a. Multinomial de parâmetros n e $p_i$, i=1,...,n, então a função distribuição da v.a.

$$U = \sum_{i=1}^{k} \frac{(O_i - np_i)^2}{np_i}$$

aproxima-se da função distribuição dum $\chi^2$ com (k-1) graus de liberdade, quando $n \to \infty$.

Observação – O modelo do $\chi^2$ tem uma função densidade com suporte positivo e tem enviesamento para a direita, dependendo a sua forma do número de graus de liberdade. Apresenta-se a seguir a função densidade do Qui-quadrado para vários graus de liberdade:




A distribuição aproximada para U, pode ser obtida de forma intuitiva, do seguinte modo:

Numa experiência multinomial, em que cada resultado pode ser um de k possíveis, o número médio de resultados, em n, que pertencem à classe $A_i$, é $np_i$. Então, $O_i$ tem uma distribuição Binomial(n, $p_i$), pelo que se n for suficientemente grande e $p_i$ pequeno, a distribuição de $O_i$ pode

ser aproximada por uma Poisson de valor médio $np_i$, e a distribuição de $\frac{O_i - np_i}{\sqrt{np_i}}$ pode ser aproximada por uma N(0,1). Finalmente $\left(\frac{O_i - np_i}{\sqrt{np_i}}\right)^2$ tem uma distribuição dum $\chi^2$ com 1 grau de liberdade e $\sum_{i=1}^{k} \frac{(O_i - np_i)^2}{np_i}$ tem uma distribuição dum $\chi^2$ com (k-1) graus de liberdade.

## 12.3 – Teste de ajustamento do Qui-quadrado para variáveis qualitativas

A propriedade anterior vai-nos servir para testar a *hipótese* de que $(O_1, O_2, ..., O_k)$ é uma v. a. Multinomial com parâmetros n, $p_i$, i=1,...,k. Basta para isso calcular, para um conjunto de valores observados $(o_1, o_2, ..., o_k)$, o valor de

$$u = \sum_{i=1}^{k} \frac{(o_i - np_i)^2}{np_i}$$

e rejeitar a *hipótese* se o valor de u for muito grande – situação em que os valores observados $o_i$ se afastam muito dos esperados $np_i$, nomeadamente $u \geq \chi^2_{1-\alpha}$ (k-1), onde $\chi^2_{1-\alpha}$ (k-1) é o quantil de probabilidade (1-$\alpha$) de um $\chi^2$ com (k-1) graus de liberdade.

**Teste de ajustamento do Qui-quadrado**

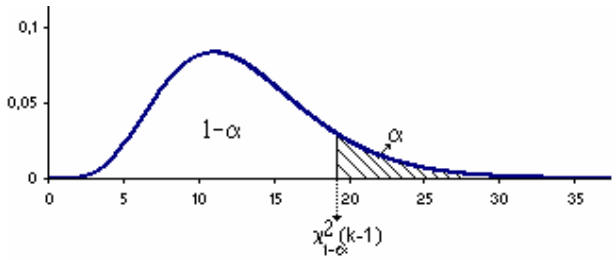
Considerando então a população em estudo, pretendemos realizar testes de hipóteses sobre os parâmetros $p_i$, i=1, ..., k, sendo as **hipóteses** a testar:

**$H_0$**: $p_1 = p_1^0$ , $p_2 = p_2^0$ ,..., $p_k = p_k^0$ contra **$H_1$**: $p_i \neq p_i^0$ , para algum 1=1, ..., k.

**Estatística de teste**: $X^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i}$

onde $O_i$ é a v.a. que representa o número de indivíduos observados da amostra, de dimensão n, que pertencem à classe $A_i$ e $e_i = n\,p_i^0$ , i=1, ..., k, são os valores esperados, isto é, os valores que esperamos obter, no caso de $H_0$ ser verdadeira. Esta estatística, sob a hipótese de $H_0$ ser verdadeira, tem uma distribuição de amostragem aproximada a um $\chi^2$ com (k-1) graus de liberdade.

**Regra de decisão**: Para o nível de significância $\alpha$, rejeita-se a hipótese nula $H_0$, quando $X^2 \geq \chi^2_{1-\alpha}$ (k-1), ou seja, a região de rejeição é constituída pelo intervalo $[\chi^2_{1-\alpha}$ (k-1), +∞[, como se pode ver pela figura seguinte



De forma alternativa, face ao valor observado da estatística de teste, $x_0^2$, calcula-se o P-value $P=P(X^2 \geq x_0^2)$ e rejeita-se $H_0$, quando $P \leq \alpha$.

Observação – Para se utilizar este teste deve-se ter em consideração que os valores esperados $e_i$ não devem ser muito pequenos. Normalmente exige-se que sejam ≥5. Quando isso não acontece, procede-se ao agrupamento de classes.

**Exemplo 1** – Utilize os dados apresentados no início deste capítulo, para verificar se existe evidência de que existam alguns signos mais propícios a que os seus nativos sejam homens de sucesso.

$H_0$: P(Carneiro) = P(Touro) = P(Gémeos) = ... = P(Peixes) = 1/12

contra

$H_1$: Alguma das probabilidades anteriores é diferente de 1/12

Sabemos que, sob $H_0$, a estatística de teste tem uma distribuição aproximada dum $\chi^2$ com 11 graus de liberdade, uma vez que k=12, isto é, temos 12 classes.

Para calcular o valor observado da estatística de teste, vamos considerar a seguinte tabela:

Quiquadrado

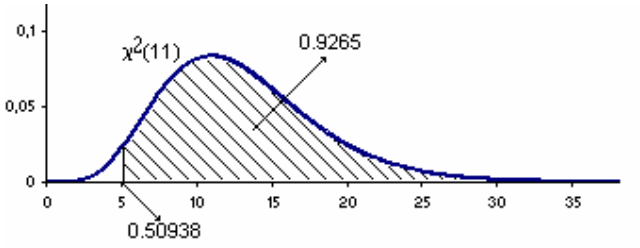
| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | **Classe** | $o_i$ | $e_i$=256×1/12 | $(o_i-e_i)^2$ | $(o_i-e_i)^2/e_i$ |
| 2 | Carneiro | 23 | 21,3333 | 2,7778 | 0,1302 |
| 3 | Touro | 20 | 21,3333 | 1,7778 | 0,0833 |
| 4 | Gémeos | 18 | 21,3333 | 11,1111 | 0,5208 |
| 5 | Caranguejo | 23 | 21,3333 | 2,7778 | 0,1302 |
| 6 | Leão | 20 | 21,3333 | 1,7778 | 0,0833 |
| 7 | Virgem | 19 | 21,3333 | 5,4444 | 0,2552 |
| 8 | Balança | 18 | 21,3333 | 11,1111 | 0,5208 |
| 9 | Escorpião | 21 | 21,3333 | 0,1111 | 0,0052 |
| 10 | Sagitário | 19 | 21,3333 | 5,4444 | 0,2552 |
| 11 | Capricórnio | 22 | 21,3333 | 0,4444 | 0,0208 |
| 12 | Aquário | 24 | 21,3333 | 7,1111 | 0,3333 |
| 13 | Peixes | 29 | 21,3333 | 58,7778 | 2,7552 |
| 14 | | | | x2 = | **5,0938** |

Sheet1 Sheet2 Sheet3

Obtivémos o valor de 5.0938 para a estatística de teste. Será que é um valor grande? Será que é um valor na cauda direita da função densidade? Será que $P(X^2 \geq 5.0938)$ é um valor pequeno, quando a distribuição de $X^2$ é um $\chi^2$ com 11 graus de liberdade? Estas três questões, são outras tantas formas de fazer a mesma pergunta, que é: Há evidência para rejeitar a hipótese nula $H_0$?

Repare-se que neste caso, não temos dificuldade em dizer que não há evidência para rejeitar $H_0$, pois basta ver na figura da função densidade do $\chi^2$ com 11 graus de liberdade, que o valor 5.0938 é relativamente pequeno. De qualquer modo calculámos o P-value associado a este teste, utilizando a função *CHIDIST*(*x;deg_freedom*) do Excel, que devolve o valor de P(X>*x*), onde X é

uma variável aleatória com uma distribuição do Qui-quadrado com *deg_freedom* graus de liberdade. O valor obtido é 0.9265, que se apresenta na figura seguinte:




Decisão: Não há evidência para rejeitar a hipósese de que os nascimentos se distribuem uniformemente pelos signos.

**Não rejeitar a hipótese nula significa que o modelo proposto é o correcto?**

Não! Na verdade o facto de os dados não nos levarem a rejeitar o modelo proposto na hipótese nula, não significa que ele seja verdadeiro. O teste serviu unicamente para mostrar que os dados são consistentes com a teoria (o modelo proposto), mas não para provar que ela é verdadeira.

**Porque é que não podemos provar a hipótese nula?** (De Veaux and al, 2004) – Um biologista pretende mostrar que a sua teoria, sobre a mosca da fruta, é válida. Segundo ele, 10% das moscas são de tipo 1, 70% de tipo 2 e 20% de tipo 3. Fez um teste de ajustamento a partir dos dados que os seus alunos recolheram, sobre 100 moscas, tendo obtido um P-value de 7%. Celebrou este facto, pois sustentava a sua hipótese, até que os seus alunos recolheram informação sobre mais 100 moscas. Com 200 moscas o P-value desceu para 2%. Apesar de já estar a adivinhar que a resposta seria não, ainda perguntou ao estatístico, na esperança de poder deitar fora metade dos dados e ficar com os 100 primeiros! Ora bem, se isto fosse possível, conseguiríamos sempre "provar a hipótese nula" não recolhendo muitos dados. Efectivamente, quanto menos informação tivermos, mais os nossos dados serão consistentes com o que quer que seja, e também nunca rejeitaremos o que quer que seja! Então um teste assim não serve para nada. Como já vimos na secção 11.5, diz-se que um teste destes tem pouca potência, medindo-se a potência de um teste como a probabilidade de rejeitar $H_0$, quando $H_0$ é falsa. Assim, quantos mais dados, melhor, já que nunca poderemos "provar" a hipótese nula.

**Exemplo 2** – Suponha que uma marca conhecida de carros pretende averiguar se existe evidência para afirmar que os compradores mudaram, nos últimos tempos, as suas preferências pelas 4 cores mais vendidas, nomeadamente o cinza prateado, o preto, o branco e o vermelho, em que estas cores eram preferidas por, respectivamente 56.25%, 18.75%, 18.75% e 6.25% dos compradores, segundo informação de alguns anos atrás. Assim, recolheu informação sobre 100 clientes, tendo obtido os seguintes resultados:

| Preto | Cinza prateado | Vermelho | Branco |
|---|---|---|---|
| 59 | 20 | 11 | 10 |

Retire conclusões, para o nível de significância de 5%.

Hipóteses:

$H_0$: P(Cinza prateado) = 0.5625; P(Preto) = 0.1875; P(Branco) = 0.1875; P(Vermelho) = 0.0625

contra

$H_1$: P(Cinza prateado)≠0.5625 **ou** P(Preto)≠0.1875 **ou** P(Branco)≠0.1875 **ou** P(Vermelho)≠0.0625

Estatística de teste: $X^2 = \sum_{i=1}^{4} \frac{(O_i - e_i)^2}{e_i}$ , que sob $H_0$, tem distribuição aproximada dum $\chi^2(3)$.

Quiquadrado

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 16 | **Classe** | $o_i$ | $e_i$=100×pi | $(o_i-e_i)^2$ | $(o_i-e_i)^2/e_i$ |
| 17 | Cinza prat | 59 | 56,25 | 7,5625 | 0,1344444 |
| 18 | Preto | 20 | 18,75 | 1,5625 | 0,0833333 |
| 19 | Branco | 11 | 18,75 | 60,0625 | 3,2033333 |
| 20 | Vermelho | 10 | 6,25 | 14,0625 | 2,25 |
| 21 | | | | | **5,6711111** |

Sheet1 Sheet2 Sheet3

Valor observado da estatística de teste: $x^2$ = 5.671

P-value: $P(X^2 \geq 5.671) = 1-P(X^2 \leq 5.671)$

Para calcular a probabilidade anterior, utilizando o Excel, utiliza-se o facto do $\chi^2$, com k graus de liberdade, ser uma Gamma de parâmetros k/2 e 2, segundo a notação do Excel. Assim, $P(X^2 \leq 5.671)$ = *GAMADIST(5.671; 1.5; 2; TRUE)*, que devolve o valor 0.871245. Finalmente temos que P-value = 0.12855

Decisão: Não rejeitar H0, para os níveis usuais de significância, nomeadamente para o nível de significância de 5%. Só rejeitaríamos $H_0$, para $\alpha \geq 12.855\%$.

Suponhamos, agora, que tinha sido recolhido uma amostra de dimensão 200, tendo obtido o dobro dos valores observados, em cada uma das categorias. Qual a conclusão que se tiraria?

Refazendo os cálculos anteriores, temos:

Quiquadrado

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 16 | **Classe** | $o_i$ | $e_i$=200×pi | $(o_i-e_i)^2$ | $(o_i-e_i)^2/e_i$ |
| 17 | Cinza prat | 118 | 112,5 | 30,25 | 0,2688889 |
| 18 | Preto | 40 | 37,5 | 6,25 | 0,1666667 |
| 19 | Branco | 22 | 37,5 | 240,25 | 6,4066667 |
| 20 | Vermelho | 20 | 12,5 | 56,25 | 4,5 |
| 21 | | | | | **11,342222** |

Sheet1 Sheet2 Sheet3

Valor observado da estatística de teste: $x^2$ = 11.342

P-value: $P(X^2 \geq 11.342) = 1-P(X^2 \leq 11.342)$

= 1- *GAMADIST(11.342; 1.5; 2; TRUE)*=1-0.989988=0.0100

Decisão: Para o nível de significância de 5%, rejeitar $H_0$, isto é, existe evidência de que os compradores mudaram de atitude, quanto ao gosto das cores.

Esta conclusão não é de estranhar, embora seja diferente da retirada anteriormente, pois agora temos mais dados, isto é, mais informação, e podemos dizer que as discrepâncias existentes entre os valores observados e os valores esperados, mostram "maior evidência" contra a hipótese nula.

## 12.4 – Teste de ajustamento do Qui-quadrado para variáveis quantitativas discretas

Suponhamos que pretendemos inferir algo sobre uma característica populacional X, quantitativa discreta e vamos começar por admitir que na hipótese nula especificamos completamente o modelo. Já que o modelo é discreto, esta especificação pode ser feita através da função massa de probabilidade ou da função distribuição:

$H_0$: $P(X=a_i) = p_i$, onde $a_i \varepsilon D$, domínio de variação da v.a. X, ou $X \cap F$
contra
H1: X não tem a distribuição admitida na hipótese nula.

Considera-se então uma partição de D, eventualmente constituída pelos pontos $a_i$, alguns dos quais podem ser agrupados. Representando por $A_1$, $A_2$, ..., $A_k$ essa partição, consideram-se as frequências observadas $O_i$, i=1,...,k, do número de elementos de uma amostra aleatória que pertencem às classes $A_i$, i=1,...,k, e estamos num caso idêntico ao considerado anteriormente, de análise de observações qualitativas, pertencentes a uma de k categorias.

Se o modelo não estiver completamente especificado, terão de se estimar alguns parâmetros, através de estimativas (da máxima verosimilhança) e estamos também na situação descrita anteriormente, da análise de observações pertencentes a uma de k categorias, mas em que a distribuição da estatística de teste não será a mesma, pois agora o número de graus de liberdade do $\chi^2$, diminui de tantas unidades, quantos os parâmetros que tiverem de ser estimados a partir dos dados. Resumindo, temos:

**Hipóteses**: $H_0$: $X \cap F$ contra $H_1$: X não tem a distribuição F

**Estatística de teste**: $X^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i}$

Distribuição da estatística de teste, sob a validade de $H_0$:

a) Se o modelo está completamente especificado, $X^2$ tem uma distribuição assintótica dum $\chi^2(k-1)$.

b) Se o modelo está especificado a menos de m parâmetros desconhecidos, que terão de ser estimados a partir dos dados, $X^2$ tem uma distribuição assintótica dum $\chi^2(k-m-1)$.

**Decisão**: Fixando o nível de significância $\alpha$:

a) Rejeita-se $H_0$ se $X^2 \geq \chi^2_{1-\alpha}$ (k-1) ou alternativamente, face ao valor observado $x^2$ da estatística de teste $X^2$, calcula-se $P = P(X^2 \geq x^2)$ e se $P \leq \alpha$, rejeita-se $H_0$.

b) Análogo à alínea a), mas com a distribuição do Qui-quadrado com (k-m-1) graus de liberdade.

Observação – Convém que o número esperado de elementos em cada classe seja ≥5.

**Exemplo 3** – A procura diária de um determinado produto, foi, em 60 dias escolhidos ao acaso, a seguinte:

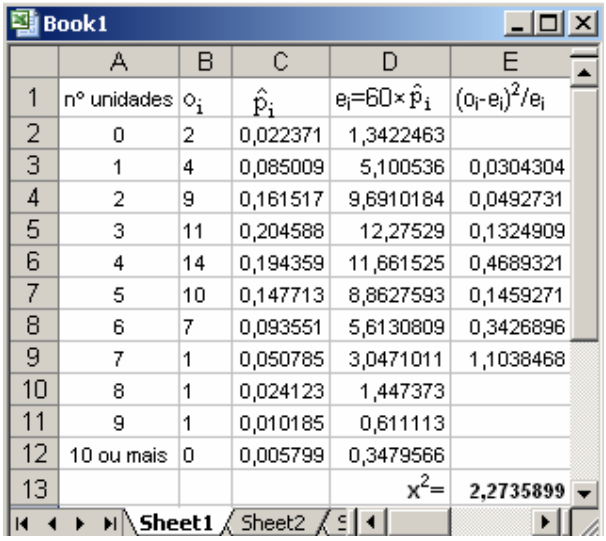
| Nº unidades procuradas | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Nº dias | 2 | 4 | 9 | 11 | 14 | 10 | 5 | 3 | 1 | 1 |

Haverá evidência para duvidar que tal procura se faça segundo um modelo de Poisson?

Resolução: Seja X a v.a. que representa o nº de unidades procuradas, por dia. Então:

$H_0$: $X \cap P(\lambda)$ contra $H_1$: X não tem uma distribuição $P(\lambda)$

Representando o estimador de $\lambda$ por $\hat{\lambda}$, temos que $\hat{\lambda} = \overline{X}$ (não esquecer que no modelo de Poisson, o parâmetro é o valor médio da variável aleatória), pelo que uma estimativa para $\lambda$, é a média dos dados $\overline{x}$=3.8, e as estimativas para as probabilidades $p_i$, obter-se-ão a partir da expressão $P(X=k) = e^{-3.8}\frac{3.8^k}{k!}$. Estas probabilidades foram obtidas no Excel através da função *Poisson(x; mean; cumulative)*, em que *x* é o valor que a v.a. X assume, *mean* é o valor médio e *cumulative* é um valor lógico: para a função distribuição, usar *TRUE*; para a função massa de probabilidade usar *FALSE*. Por exemplo, para obter o valor 0.085009, colocámos o cursor na célula C3 e inserimos a função *=POISSON(B3;3,8;FALSE).*

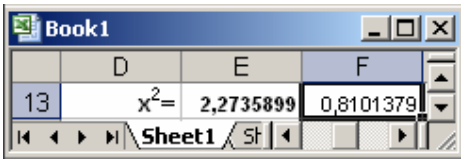


| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | nº unidades | $o_i$ | $\hat{p}_i$ | $e_i=60\times\hat{p}_i$ | $(o_i-e_i)^2/e_i$ |
| 2 | 0 | 2 | 0,022371 | 1,3422463 | |
| 3 | 1 | 4 | 0,085009 | 5,100536 | 0,0304304 |
| 4 | 2 | 9 | 0,161517 | 9,6910184 | 0,0492731 |
| 5 | 3 | 11 | 0,204588 | 12,27529 | 0,1324909 |
| 6 | 4 | 14 | 0,194359 | 11,661525 | 0,4689321 |
| 7 | 5 | 10 | 0,147713 | 8,8627593 | 0,1459271 |
| 8 | 6 | 7 | 0,093551 | 5,6130809 | 0,3426896 |
| 9 | 7 | 1 | 0,050785 | 3,0471011 | 1,1038468 |
| 10 | 8 | 1 | 0,024123 | 1,447373 | |
| 11 | 9 | 1 | 0,010185 | 0,611113 | |
| 12 | 10 ou mais | 0 | 0,005799 | 0,3479566 | |
| 13 | | | | $x^2$= | 2,2735899 |

Chamamos a atenção para o facto de as classes $A_i$ deverem constituir uma partição do domínio da v.a. X. Assim, como o domínio da Poisson é constituído pelos valores inteiros positivos (incluindo o 0) introduzimos a classe 10 ou mais, cuja probabilidade foi calculada fazendo (1-P(X≤9)) (não esquecer que $\sum P(A_i) = 1$). Por outro lado, tendo em conta a observação feita sobre o valor dos

$e_i$, que não devem ser inferiores a 5, agrupámos as classes 0 e 1, numa classe, e as classes 7, 8, 9 e 10 ou mais, noutra classe, tendo ficado assim 7 classes.

Se $H_0$ for verdadeiro, a estatística de teste $X^2 = \sum_{i=1}^{7} \frac{(O_i - e_i)^2}{e_i}$ tem uma distribuição assintótica dum $\chi^2(7\text{-}1\text{-}1)$, ou seja dum Qui-quadrado com 5 graus de liberdade. Segundo a tabela anterior, obtivémos, para a estatística de teste, o valor observado de 2.2736. Para tomar uma decisão, vamos calcular o P-value:$P(X^2 \geq 2.2736) = 0.81$. Este valor foi obtido, inserindo na célula F13, a função = *CHIDIST(E13;5):*




Decisão: Não há evidência para dizer que a distribuição do número de unidades procuradas por dia, não segue uma distribuição de Poisson.

## 12.5 – Teste de ajustamento do Qui-quadrado para variáveis quantitativas contínuas

Este teste para observações contínuas é análogo ao realizado para observações discretas.

Observação: Convém referir, no entanto, o seguinte: agora a escolha das classes $A_i$, que constituem uma partição do domínio da variável aleatória X, já não é feita de uma forma tão óbvia, como no caso dos dados discretos. Assim, de forma a reduzir a arbitrariedade na escolha da partição $A_i$, $1 \leq i \leq k$, é usual escolher os $A_i$, tais que

$P(X \varepsilon A_i | H_0) = 1/k$ ou seja $p_i = 1/k$, $1 \leq i \leq k$.

Como escolher o k?

A escolha de k é feita de modo a garantir que o número esperado $e_i = np_i$, de elementos em cada classe seja ≥5. Assim, deve ter-se $n/k \geq 5$, o que implica que $k \leq n/5$. Considera-se geralmente para k o maior inteiro contido em n/5 (a não ser que este valor seja demasiado grande, como veremos no exemplo a seguir, em que se escolhe um valor inferior), e as classes $A_i$, são assim construídas:

$A_1 = (-\infty, a_1[$, $\quad P(X \varepsilon A_1 | H_0) = 1/k \rightarrow P(X \leq a_1) = F(a_1) = 1/k \rightarrow a_1 = F^{-1}(1/k)$

$A_2 = [a_1, a_2[$, $\quad P(X \varepsilon A_2 | H_0) = 1/k \rightarrow P(a_1 < X \leq a_2) = F(a_2) - F(a_1) = 1/k \rightarrow a_2 = F^{-1}(2/k)$

...

$A_k = [a_{k-1}, \infty[$, $\quad P(X \varepsilon A_k | H_0) = 1/k \rightarrow P(X > a_{k-1}) = 1 - F(a_{k-1}) = 1/k \rightarrow a_{k-1} = F^{-1}((k-1)/k)$

A estatística de teste obtém-se da mesma maneira, assim como a distribuição de amostragem.

**Exemplo 4** – O Sr. Silva, industrial têxtil, decidiu começar a fabricar camisas de homem, destinadas a serem vendidas em Portugal. Para ter alguma informação sobre os moldes que deve considerar, nomeadamente no que diz respeito ao comprimento das mangas, resolveu pedir a uma empresa de Consultoria de Estatística que o ajudasse, dando-lhe algumas indicações sobre a população a que se destinam as camisas.

Vamos delinear o processo utilizado pela tal empresa, para ajudar o Sr. Silva.

1º passo – Recolha de uma amostra

A empresa de Consultoria encarregou o Departamento de Sondagens de recolher uma amostra de dimensão 250, tendo esta fornecido os seguintes dados, relativos ao comprimento do braço direito de 250 homens:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 51.5 | 56.0 | 55.0 | 58.3 | 58.4 | 55.3 | 56.3 | 52.2 | 55.2 | 57.3 |
| 55.4 | 52.9 | 54.0 | 59.7 | 55.4 | 53.0 | 52.6 | 55.5 | 53.1 | 52.4 |
| 57.9 | 57.7 | 55.3 | 53.5 | 55.8 | 57.9 | 54.7 | 55.7 | 54.0 | 52.1 |
| 57.6 | 52.9 | 54.2 | 52.9 | 56.2 | 54.9 | 58.2 | 53.2 | 54.1 | 53.1 |
| 53.9 | 54.9 | 56.7 | 52.1 | 57.7 | 55.4 | 54.9 | 54.9 | 55.5 | 56.6 |
| 56.6 | 54.7 | 55.6 | 53.2 | 54.7 | 53.0 | 57.5 | 55.6 | 56.9 | 57.4 |
| 49.9 | 54.7 | 53.8 | 58.4 | 55.7 | 55.4 | 54.3 | 49.1 | 56.7 | 55.4 |
| 53.0 | 55.3 | 55.7 | 52.1 | 51.0 | 53.1 | 55.3 | 52.1 | 54.3 | 54.9 |
| 55.3 | 56.7 | 57.1 | 54.4 | 53.7 | 58.9 | 53.8 | 54.8 | 55.7 | 55.4 |
| 56.6 | 56.8 | 53.4 | 53.4 | 56.0 | 56.5 | 56.7 | 54.0 | 51.6 | 52.6 |
| 56.4 | 56.8 | 57.4 | 54.7 | 55.5 | 53.2 | 54.7 | 54.7 | 58.4 | 56.3 |
| 58.1 | 53.4 | 56.7 | 58.1 | 54.9 | 54.2 | 56.5 | 53.2 | 51.3 | 56.6 |
| 56.6 | 58.8 | 57.7 | 52.5 | 56.2 | 54.4 | 56.8 | 51.8 | 53.9 | 58.4 |
| 58.7 | 55.2 | 53.0 | 58.0 | 58.6 | 52.3 | 59.2 | 56.5 | 57.1 | 54.2 |
| 55.3 | 55.5 | 56.1 | 52.1 | 53.9 | 53.2 | 52.9 | 58.8 | 55.0 | 54.2 |
| 54.8 | 53.4 | 56.8 | 51.9 | 55.0 | 51.6 | 58.2 | 55.5 | 56.2 | 53.7 |
| 54.6 | 51.7 | 55.5 | 52.8 | 54.4 | 55.7 | 54.0 | 56.8 | 53.3 | 56.8 |
| 54.2 | 50.5 | 54.3 | 54.6 | 53.2 | 52.2 | 55.2 | 55.4 | 55.8 | 55.6 |
| 60.2 | 57.0 | 54.6 | 55.0 | 56.6 | 55.1 | 58.0 | 57.3 | 56.0 | 51.7 |
| 55.1 | 54.5 | 53.8 | 55.1 | 55.7 | 57.1 | 53.2 | 52.4 | 55.5 | 57.2 |
| 56.1 | 55.1 | 55.2 | 56.3 | 57.1 | 55.5 | 53.2 | 54.8 | 55.6 | 56.0 |
| 60.7 | 58.3 | 59.4 | 52.8 | 55.8 | 56.8 | 56.3 | 55.7 | 53.0 | 53.0 |
| 51.9 | 55.7 | 53.4 | 53.8 | 52.1 | 57.5 | 59.8 | 55.3 | 55.0 | 55.0 |
| 54.2 | 57.6 | 55.1 | 56.5 | 58.3 | 53.1 | 55.2 | 53.7 | 48.4 | 54.7 |
| 55.0 | 56.5 | 56.9 | 57.0 | 58.2 | 56.7 | 54.4 | 50.2 | 54.4 | 56.5 |

2º passo – Estudo descritivo

Procedeu-se ao estudo descritivo dos dados anteriores, calculando algumas características amostrais e procedendo à redução dos dados através de uma tabela de frequências e à construção do histograma correspondente. Apresentam-se a seguir os resultados obtidos:

Exemplo4.12

| | A | B | C |
|---|---|---|---|
| 1 | x | | |
| 2 | 51,5 | *Column1* | |
| 3 | 55,4 | | |
| 4 | 57,9 | Mean | 55,1376 |
| 5 | 57,6 | Median | 55,2 |
| 6 | 53,9 | Mode | 54,7 |
| 7 | 56,6 | Standard Deviation | 2,0874136 |
| 8 | 49,9 | Sample Variance | 4,3572954 |
| 9 | 53,0 | Range | 12,3 |
| 10 | 55,3 | Minimum | 48,4 |
| 11 | 56,6 | Maximum | 60,7 |
| 12 | 56,4 | Count | 250 |
| 13 | 58,1 | | |

Sheet1 / Sheet

Decidimos construir uma tabela de frequências com 8 classes, valor sugerido pela regra empírica enunciada quando da construção do histograma, e considerar como amplitude de classe o valor 1.54 (valor aproximado, por excesso, de (max-min)/8).Construímos uma tabela de frequências e o histograma associado, utilizando a metodologia das *PivotTables*.:

Exemplo4.12

| Count of x | |
|---|---|
| x | Total |
| 48,4-49,94 | 3 |
| 49,94-51,48 | 4 |
| 51,48-53,02 | 34 |
| 53,02-54,56 | 49 |
| 54,56-56,1 | 78 |
| 56,1-57,64 | 52 |
| 57,64-59,18 | 24 |
| 59,18-60,72 | 6 |
| Grand Total | 250 |

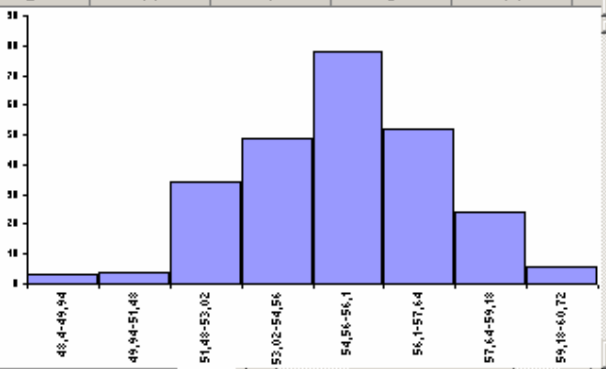



Chart1 / Chart2 / Sheet4 / Chart3 / Shee

O histograma sugere-nos um modelo Normal, pelo que, o passo seguinte será testar se efectivamente tem sentido ajustar um modelo Normal aos dados. Uma questão que se levanta neste momento é a seguinte: terá sentido estar a ajustar aos nossos dados um modelo com suporte R, isto é, que pode assumir qualquer valor real, quando nós sabemos que isso não se passa com o comprimento do braço? Mas se estamos renitentes em ajustar um modelo com suporte em R, talvez pensassemos que seria mais razoável um cujo suporte fosse $R^+$, pois se temos a garantia que o comprimento não pode ser negativo, não sabemos qual o valor máximo que devemos escolher. Ou poderíamos inventar um valor ao acaso como limite superior, por exemplo 150 cm, mas com que legitimidade é que escolhemos este e não outro valor? Também não devemos considerar o valor 60.7 como valor máximo, embora tenha sido o maior valor da amostra que se recolheu. Ninguém nos garante que na população não haja homens com o

comprimento do braço superior a 60.7! Nesta altura, de reflexão sobre qual o modelo a adoptar, recordemos o que se disse sobre a escolha de um modelo para traduzir um fenómeno aleatório – *todos os modelos são maus, alguns são úteis*. No entanto, além do histograma nos sugerir o modelo Normal, devido à semelhança com a função densidade da Normal, também dispomos de alguma informação científica sobre este modelo; e são esses estudos que nos dizem que ele se aplica em situações de fenómenos que possam ser considerados provenientes de uma contribuição aditiva de várias variáveis, como é, por exemplo, o caso da variável em estudo. Então, em posse da informação sobre a proveniência dos dados e dos resultados do estudo descritivo dos mesmos, estamos em condições de propor o modelo Normal.

3º passo – Teste de ajustamento do modelo sugerido no passo anterior

Representando por X, a v.a. que representa o comprimento do braço, consideremos as seguintes hipóteses:

$$H_0: X \cap N(\mu,\sigma) \quad \text{contra} \quad H_1: X \not\cap N(\mu,\sigma)$$

Para utilizarmos o teste de ajustamento do Qui-qudrado, as classes $A_i$ têm que constituir uma partição do suporte da v.a. X. Neste momento podemos seguir dois processos, nomeadamente: utilizar a tabela de frequência anterior, procedendo às modificações adequadas nas classes, de forma a termos uma partição, ou utilizar o processo enunciado anteriormente, para a formação das classes. Vamos exemplificar os dois processos:

Processo 1 – Modificação da tabela de frequências, de forma a termos uma partição de R

Para obter uma partição, basta proceder a uma alteração conveniente na primeira e na última classe, como se apresenta a seguir:

Exemplo4.12

| | D | E | F | G | H | I |
|---|---|---|---|---|---|---|
| 17 | | Classes | $o_i$ | $\hat{p}_i$ | $e_i=250\times\hat{p}_i$ | $(o_i-e_i)^2/e_i$ |
| 18 | | (-∞, 49.94] | 3 | 0,006358 | 1,5895485 | 1,2515336 |
| 19 | | ] 49.94, 51.48] | 4 | 0,033382 | 8,3454037 | 2,2626267 |
| 20 | | ] 51.48, 53.02] | 34 | 0,11512 | 28,779903 | 0,946821 |
| 21 | | ] 53.02, 54.56] | 49 | 0,235681 | 58,920242 | 1,6702443 |
| 22 | | ] 54.56, 56.10] | 78 | 0,286698 | 71,674516 | 0,5582423 |
| 23 | | ] 56.10, 57.64] | 52 | 0,207282 | 51,82049 | 0,0006218 |
| 24 | | ] 57.64, 59.18] | 24 | 0,089033 | 22,258265 | 0,1362928 |
| 25 | | ] 59.18, +∞) | 6 | 0,026447 | 6,6116323 | 0,0565812 |
| 26 | | | | | $x^2=$ | 6,8829638 |

Chart1 / Chart2 / Sheet4

Para calcular estimativas das probabiliaddes $p_i$, utilizámos o modelo Normal(55.14, 2.087), no Excel. Por exemplo, para calcular a probabilidade do intervalo ]49.94, 51.48], colocámos o cursor na célula G19 e escrevemos =*NORMDIST(51,48;55,14;2,087;TRUE)-NORMDIST(49,94;55, 14;2,087;TRUE).*

Como estimámos dois parâmetros a partir dos dados, a estatística de teste $X^2$, tem uma distribuição assintótica dum $\chi^2$(8-2-1), ou seja dum Qui-quadrado com 5 graus de liberdade.

Para tomar uma decisão calculámos o P-value, bastando colocar o cursor na célula J26 e escrever =*CHIDIST(I26;5):*

Exemplo4.12

| | H | I | J |
|---|---|---|---|
| 26 | $x^2$= | 6,8829638 | 0,229491 |

Decisão: Não existe evidência para rejeitar a hipótese do modelo Normal.

Processo 2 – Admitindo que não tinha havido uma fase anterior, em que tinha sido necessário proceder a um agrupamento dos dados, como no caso do exemplo que estamos a tratar, vamos exemplificar o processo sugerido na secção anterior.

Temos n=250, donde k≤250/5. Vamos considerar k=10, isto é, 10 classes. Então os limites de classe $a_1$, $a_2$, ..., $a_9$, com a notação introduzida na secção referida, podem ser obtidos no Excel, da seguinte forma:

Exemplo4.12

| | H | I | J | K |
|---|---|---|---|---|
| 3 | i | i/10 | ai | |
| 4 | 1 | =H4/10 | a1 | =NORMINV(I4;55,14;2,087) |
| 5 | 2 | =H5/10 | a2 | =NORMINV(I5;55,14;2,087) |
| 6 | 3 | =H6/10 | a3 | =NORMINV(I6;55,14;2,087) |
| 7 | 4 | =H7/10 | a4 | =NORMINV(I7;55,14;2,087) |
| 8 | 5 | =H8/10 | a5 | =NORMINV(I8;55,14;2,087) |
| 9 | 6 | =H9/10 | a6 | =NORMINV(I9;55,14;2,087) |
| 10 | 7 | =H10/10 | a7 | =NORMINV(I10;55,14;2,087) |
| 11 | 8 | =H11/10 | a8 | =NORMINV(I11;55,14;2,087) |
| 12 | 9 | =H12/10 | a9 | =NORMINV(I12;55,14;2,087) |

Exemplo4.12

| | H | I | J | K |
|---|---|---|---|---|
| 3 | i | i/10 | ai | |
| 4 | 1 | 0,1 | a1 | 52,47 |
| 5 | 2 | 0,2 | a2 | 53,38 |
| 6 | 3 | 0,3 | a3 | 54,05 |
| 7 | 4 | 0,4 | a4 | 54,61 |
| 8 | 5 | 0,5 | a5 | 55,14 |
| 9 | 6 | 0,6 | a6 | 55,67 |
| 10 | 7 | 0,7 | a7 | 56,23 |
| 11 | 8 | 0,8 | a8 | 56,90 |
| 12 | 9 | 0,9 | a9 | 57,81 |

Uma vez as classes construídas, teremos de contar quais os valores observados. Utilizámos a seguinte tabela feita no Excel, para determinar esses valores, assim como o valor observado da estatística de teste:

Exemplo4.12

| | K | L | M | N | O |
|---|---|---|---|---|---|
| 17 | | $o_i$ | $e_i$ =250*1/10 | $(o_i - e_i)^2/e_i$ | |
| 18 | 52,47 | =COUNTIF($A$2:$A$251;"<52,47") | 25 | =(L18-M18)^2/M18 | |
| 19 | 53,38 | =COUNTIF($A$2:$A$251;"<53,38")-COUNTIF($A$2:$A$251;"<52,47") | 25 | =(L19-M19)^2/M19 | |
| 20 | 54,05 | =COUNTIF($A$2:$A$251;"<54,05")-COUNTIF($A$2:$A$251;"<53,38") | 25 | =(L20-M20)^2/M20 | |
| 21 | 54,61 | =COUNTIF($A$2:$A$251;"<54,61")-COUNTIF($A$2:$A$251;"<54,05") | 25 | =(L21-M21)^2/M21 | |
| 22 | 55,14 | =COUNTIF($A$2:$A$251;"<55,14")-COUNTIF($A$2:$A$251;"<54,61") | 25 | =(L22-M22)^2/M22 | |
| 23 | 55,67 | =COUNTIF($A$2:$A$251;"<55,67")-COUNTIF($A$2:$A$251;"<55,14") | 25 | =(L23-M23)^2/M23 | |
| 24 | 56,23 | =COUNTIF($A$2:$A$251;"<56,23")-COUNTIF($A$2:$A$251;"<55,67") | 25 | =(L24-M24)^2/M24 | |
| 25 | 56,9 | =COUNTIF($A$2:$A$251;"<56,9")-COUNTIF($A$2:$A$251;"<56,23") | 25 | =(L25-M25)^2/M25 | |
| 26 | 57,81 | =COUNTIF($A$2:$A$251;"<57,81")-COUNTIF($A$2:$A$251;"<56,9") | 25 | =(L26-M26)^2/M26 | |
| 27 | | =250-COUNTIF($A$2:$A$251;"<57,81") | 25 | =(L27-M27)^2/M27 | |
| 28 | | =SUM(L18:L27) | | $x^2$= =SUM(N18:N27) | =CHIDIST(N28;7) |

A estatística de teste é a mesma, mas agora tem uma distribuição de amostragem dum Qui-quadrado com 7=(10-2-1) graus de liberdade, uma vez que considerámos 10 classes e estimámos 2 parâmetros:

Exemplo4.12

| | K | L | M | N | O |
|---|---|---|---|---|---|
| 17 | | $o_i$ | $e_i$ =250*1/10 | $(o_i - e_i)^2/e_i$ | |
| 18 | 52,47 | 26 | 25 | 0,04 | |
| 19 | 53,38 | 28 | 25 | 0,36 | |
| 20 | 54,05 | 20 | 25 | 1,00 | |
| 21 | 54,61 | 19 | 25 | 1,44 | |
| 22 | 55,14 | 29 | 25 | 0,64 | |
| 23 | 55,67 | 31 | 25 | 1,44 | |
| 24 | 56,23 | 20 | 25 | 1,00 | |
| 25 | 56,90 | 30 | 25 | 1,00 | |
| 26 | 57,81 | 20 | 25 | 1,00 | |
| 27 | | 27 | 25 | 0,16 | |
| 28 | | 250 | $\chi^2$= | **8.08** | 0.3256 |

Sheet1 Sheet2

Decisão: Uma vez que o P-value é igual a 32.56%, não existe evidência para rejeitar a hipótese de que os dados sejam provenientes de um modelo Normal.

4º passo – Transmissão dos resultados ao industrial têxtil

Agora, nesta fase, justificava-se uma conversa com o Sr. Silva, para a apresentação dos resultados. Pode-se, no entanto, ir adiantando alguma informação, em termos de percentagens dos futuros compradores das camisas. Assim, temos os seguintes números:

- Aproximadamente 68% dos homens têm o comprimento dos braços no intervalo [53, 57]
  $P(55.14-2.087 \leq X \leq 55.14+2.087)=\phi(1)-\phi(-1)= 2\phi(1)-1 \approx 0.68$
- Aproximadamente 95% dos homens têm o comprimento dos braços no intervalo [51, 59]
  $P(55.14-2\times 2.087 \leq X \leq 55.14+2\times 2.087)=\phi(2)-\phi(-2)= 2\phi(2)-1 \approx 0.95$
- Aproximadamente 100% dos homens têm o comprimento dos braços no intervalo [49, 61]
  $P(55.14-3\times 2.087 \leq X \leq 55.14+3\times 2.087)=\phi(2)-\phi(-2)= 2\phi(3)-1 \approx 0.997$

Utilizando ainda o modelo Normal(55.14, 2.087), podemos ser um pouco mais precisos, informando o Sr. Silva sobre os valores do 1º e 3º quartis, que são respectivamente 53.7 cm e 56.5 cm:

Exemplo4.12

| | Q | R |
|---|---|---|
| 18 | 0,25 | =NORMINV(Q18;55,14;2,087) |
| 19 | 0,75 | =NORMINV(Q19;55,14;2,087) |

Sheet1

Exemplo4.12

| | Q | R |
|---|---|---|
| 18 | 0,25 | 53,73233989 |
| 19 | 0,75 | 56,54766011 |

Sh

Assim, o industrial sabe que, por exemplo, só 25% dos homens é que têm o comprimento dos braços inferior a 53.7 cm e que 50% dos homens têm o comprimento dos braços no intervalo [53.7, 56.5]. Esta informação é importante, pois permite fazer uma programação adequada da percentagem de camisas que devem ser fabricadas, para cada tamanho.

**Exercícios**

1. É convicção popular que os nascimentos ocorrem "depois da 9ª lua". Seguidamente apresenta-se uma tabela onde se regista o número de nascimentos nos 7 dias seguintes a cada fase da lua, de crianças seleccionadas ao acaso entre as nascidas numa determinada maternidade, em 1995:

| Lua nova | Crescente | Lua cheia | minguante |
|---|---|---|---|
| 72 | 61 | 68 | 61 |

Com base naqueles dados, teste ao nível de significância de 5%, se a convicção popular tem algum fundamento, não sendo os nascimentos distribuídos, de forma regular, ao longo das diversas fases da lua.

2. Numa amostra aleatória de 200 famílias, cada uma com 4 filhos, registou-se o número de raparigas em cada família, tendo-se obtido os seguintes resultados:

| Nº de filhas | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Nº de famílias | 5 | 32 | 65 | 75 | 23 |

Teste o ajustamento de uma distribuição Binomial a estes dados.

3. Com o objectivo de testar a hipótese de que uma moeda é equilibrada, essa moeda é lançada até se obter cara pela 1ª vez. Repetiu-se a experiência 150 vezes, tendo-se obtido os seguintes resultados:

| Nº lançamentos necessários até obter cara pela 1ª vez (inclusivé) | 1 | 2 | 3 | 4 | 5 ou mais |
|---|---|---|---|---|---|
| Frequência | 60 | 48 | 22 | 11 | 9 |

Que pode concluir?

4. O médico responsável pelo gabinete médico de uma fábrica registou o nº de acidentes, por mês, verificados nessa fábrica, durante os últimos 10 anos:

| nº acidentes/mês | nº meses |
|---|---|
| 0 | 2 |
| 1 | 10 |
| 2 | 15 |
| 3 | 30 |
| 4 | 28 |
| 5 | 15 |
| 6 | 10 |
| 7 | 6 |
| $\geq 8$ | 4 |

Relativamente aos dados anteriores:
a) Determine valores aproximados para a média e para a variância amostral.
b) Faça o resumo de 5 números e investigue a existência de possíveis outliers.
c) Faça uma representação gráfica conveniente.
d) Pensa-se que o nº de acidentes por mês nessa fábrica, é uma v.a. com distribuição de Poisson de valor médio 4. Verifique se existem razões que nos levem a duvidar desta suposição.
e) Encontre um intervalo de 95% de confiança para o nº médio de acidentes por mês.
f) Verifique se existem razões para afirmar que em mais de 15% dos meses, se verificam 6 ou mais acidentes.

5. O sr. Nobre dedica-se à criação de leitões, que vende quando atingem os dois meses de idade e pesam mais de 9kg. Pretendendo fazer um estudo sobre o crescimento dos leitões, resolveu pesar 64 leitões com dois meses de idade, tendo obtido os seguintes valores:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 4.1 | 5.8 | 5.8 | 6.1 | 6.7 | 7.0 | 7.5 | 7.5 | 7.5 | 7.5 | 7.7 | 8.2 |
| 8.3 | 8.5 | 8.7 | 8.8 | 9.0 | 9.0 | 9.1 | 9.1 | 9.1 | 9.2 | 9.2 | 9.2 |
| 9.2 | 9.4 | 9.4 | 9.4 | 9.5 | 9.5 | 9.7 | 9.8 | 10.0 | 10.0 | 10.0 | 10.2 |
| 10.2 | 10.2 | 10.3 | 10.6 | 10.6 | 10.8 | 10.9 | 10.9 | 11.0 | 11.1 | 11.1 | 11.6 |
| 11.7 | 11.8 | 11.8 | 11.8 | 12.0 | 12.2 | 12.2 | 12.3 | 12.5 | 12.6 | 12.7 | 14.0 |
| 14.1 | 14.2 | 15.0 | 16.0 | | | | | | | | |

a) Represente graficamente os dados e calcule a média, a mediana, a variância e o quantil 2/3. Sugira uma distribuição para a população da qual a amostra foi recolhida e averigúe se a sua suposição tem razão de ser.
b) O sr. Nobre afirma que pelo menos 40% dos leitões que vende, têm peso superior a 11.8 kg. Ultimamente os restaurantes que compram os leitões ao sr. Nobre, têm-se queixado, afirmando que os leitões têm menos peso do que o estipulado pelo comerciante. Verifique se existe evidência para afirmar que os donos dos restaurantes têm razão.
c) Encontre um intervalo de 99% de confiança para o peso médio dos leitões vendidos pelo sr. Nobre.

6. Considere os seguintes dados, que dizem respeito ao peso de 37 crianças de uma determinada classe etária:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 18.2 | 17.4 | 17.6 | 16.7 | 17.1 | 20.1 | 17.9 | 16.8 | 19.6 | 18.4 | 17.7 |
| 19.3 | 20.4 | | | | | | | | | |
| 18.4 | 18.6 | 17.8 | 16.9 | 20.6 | 19.8 | 18.7 | 17.5 | 17.8 | 18.3 | 18.9 |
| 19.6 | 19.6 | | | | | | | | | |
| 20.6 | 18.7 | 18.3 | 18.8 | 19.6 | 18.6 | 19.9 | 20.7 | 19.6 | 18.9 | 20.8 |

a) Sugira uma distribuição de probabilidade que lhe pareça ajustar-se à população subjacente aos dados. Justifique a sua escolha.
b) Utilizando um teste de ajustamento adequado, teste a adaptabilidade do modelo sugerido em a).

7. Os dados da tabela seguinte, representam o comprimento, arredondado aos centímetros, de uma amostra de peças de determinado tipo:

| comprimento | [49-52[ | [52-54[ | [55-58[ | [58-61[ | [61-64[ | [64-67[ | [67-70[ |
|---|---|---|---|---|---|---|---|
| frequência | 2 | 10 | 48 | 64 | 56 | 16 | 4 |

a) Represente graficamente os dados e calcule a média, a mediana e a variância. Teste a hipótese de que o comprimento das peças tem uma distribuição Normal com variância 12.
b) Suponha que já sabe que a distribuição é Normal. Justifique que ($\bar{x}$-0.5; $\bar{x}$+0.5) é um intervalo de confiança para o comprimento médio das peças. Determine o grau de confiança.
c) Verifique se existe evidência para afirmar que o comprimento médio das peças é menor que 60 cm.

**Bibliografia**

Na preparação destas folhas seguiu-se essencialmente a seguinte bibliografia:

Alpuim, T. – *Introdução às Probabilidades*, Associação dos Estudantes da FCUL, 1997

De Veaux, R. e Velleman, P. – *Intro Stats*, Pearson Education, Inc, 2004

Feller, W. – *An Introduction to Probability Theory and its Applications*, John Wiley & Sons, 1968

Freedman, D., Pisani, R., Purves, R., Adhikari, A.. - *Statistics*. W. W. Norton & Company, 1991.

Graça Martins, M. E., Cerveira, A. – *Introdução às Probabilidades e à Estatística,* Universidade Aberta, 1999

Graça Martins, M. E., Monteiro, C., Viana, J. P., Turkman, M. A. A. – *Estatística*, Ministério da Educação, Departamento do Ensino Secundário, 1997

Graça Martins, M. E., Monteiro, C., Viana, J. P., Turkman, M. A. A. – *Probabilidades e Combinatória*, Ministério da Educação, Departamento do Ensino Secundário, 1999

Graça Martins, M. E., Loura, L. – *Matemática para as Ciências Sociais*, Anexo para apoio à interpretação do programa, 2001.

Graça Martins, M. E., Loura, L. – *Introdução à Probabilidade*, Projecto Reanimat, Departamento de Estatística e Investigação Operacional, 2003.

Graça Martins, M. E., Loura, L. – *Estatística Computacional*, Anexo para apoio à interpretação do programa do Módulo B2 para os Cursos Profissionais, Departamento de Estatística e Investigação Operacional, 2005.

Hoaglin, D. and al. - *Análise Exploratória de dados. Técnicas robustas*. Edições Salamandra, 1993.

Iman, R. e Conover, W. - *A Modern Approach to Statistics*. John Wiley & Sons, 1983.

Mann, P. – *Introductory Statistics*. John Wiley & Sons, 1995.

Mendenhall. W., Beaver, R. – *Introduction to Probability and Statistics*. Duxbury Press, 1994

Moore, D. – *Statistics* – Concepts and Controversies. Freeman, 1997

Moore, D. – *The Basic Practice of Statistics,* Freeman, 1996

Moore, D., McCabe, G. – *Introduction to the Practice of Statistics,* Freeman, 1996

Murteira, B. - *Análise Exploratória de Dados. Estatística descritiva*. McGraw-Hill, 1993.

Murteira, B. And al. – *Introdução à Estatística*, McGraw-Hill, 2002

Parzen, E. – *Modern Probability and its Applications*, John Wiley & Sons, 1960

Pestana, D. and al. – *Introdução à Probabilidade e à Estatística*, Fund. Calouste Gulbenkian, 2002

Rossman, A. – *Workshop Statistics , Discovery with data*. Springer-Verlag New York, 1996

Siegel, A. - *Statistics and data analysis*. John Wiley & Sons, 1988.

Tannenbaum, P. and al. - *Excursions in modern Mathematics*, Prentice Hall, 1998.

Vicente, P., Reis, E., Ferrão, F. – *Sondagens*, Edições Sílabo, Lda, 1996

**Artigos da revista TEACHING STATISTICS**

Hodgson, T. and al. – Why Statify? Vol 20, 1, 68-71

Neville, H. – Handling Continuous Data in Excel, Vol 25, 2, 42-45

Neville, H. – Charts in Excel, Vol 26, 2, 49-53

**Páginas na Internet**

ALEA - http://www.alea.pt

Instituto Nacional de Estatística - www.ine.pt

(Tem informação sobre Portugal, ao nível da freguesia)

Eurostat – europa.eu.int/comm/eurostat/

(Tem informação relativa aos diversos países da Europa)

World Health Organization – http://www.who.int/research/en/

(Tem informação sobre temas ligados à saúde, para todos os países do mundo)

World in figures – http://.stat.fi/tup/maanum/index_en.html

(Tem informação das mais diversas areas, tais como população e estatísticas vitais, cultura, religiões, emprego, consumo, etc., relativa a todos os países do mundo)