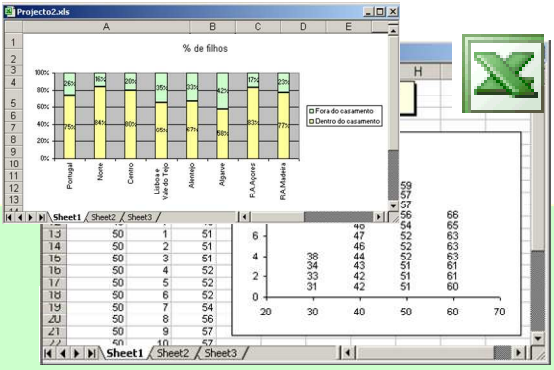


Dossiês Didáticos

XIII a – Estatística Descritiva com Excel - Complementos

LUÍSA CANTO E CASTRO LOURA

MARIA EUGÉNIA GRAÇA MARTINS

Departamento de Estatística e Investigação Operacional da Faculdade de Ciências da Universidade de Lisboa (Janeiro de 2009)

Versão atualizada para Excel 2010

MARIA EUGÉNIA GRAÇA MARTINS

Outubro 2012

Nota

Este dossiê é uma versão integral do dossiê **XIII - Estatística Descritiva com Excel** – **Complementos**, em que se procedeu a uma atualização dos procedimentos realizados em Excel 2003, para Excel 2010.

Aproveitou-se para atualizar o ficheiro que serve para exemplificar os conceitos e procedimentos, que é agora constituído pelos dados dos deputados da XII legislatura, a que demos o nome de *DeputadosXII*.

1. Noções básicas sobre amostragem

1.1- Introdução[1]

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* ou Análise de Dados. 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**".

Do que dissemos anteriormente, podemo-nos aperceber que a Estatística é uma ciência que trata de dados e que num procedimento estatístico estão envolvidas duas fases importantes, nomeadamente a fase que diz respeito à organização de dados - Análise de Dados, e a fase em que se procura retirar conclusões a partir dos dados, dando ainda informação de qual a confiança que devemos atribuir a essas conclusões - Inferência Estatística. Existe, no entanto, uma fase pioneira, que diz respeito à *Produção ou Aquisição de Dados*. Para realçar a importância desta fase consideremos, por analogia, o que se passa quando se pretende realizar um determinado cozinhado. Começa-se por selecionar os ingredientes, que serão depois manipulados de acordo com determinada receita. O resultado do cozinhado pode ser desastroso, embora de aspeto agradável. Efetivamente se os ingredientes não estiverem em condições, resulta um prato de aspeto semelhante ao que se obteria com ingredientes bons, mas de sabor intragável. O mesmo se passa com o procedimento estatístico. Se os dados não forem bons, embora se aplique a técnica correcta, o resultado pode ser desastroso, na medida em que se pode ser levado e retirar conclusões erradas.

[1] Este capítulo segue de perto o texto *Introdução à Probabilidade e à Estatística – Com complementos de Excel*, de Maria Eugénia Graça Martins, edição da Sociedade Portuguesa de Estatística, 2005.

Hoje em dia com a utilização cada vez maior de **dados** nas mais variadas profissões e nas mais diversas situações do dia a dia, torna-se necessário acompanhar este processo de uma cultura estatística que cada vez mais abarque um maior número de pessoas, para que mais facilmente se consiga compreender o mundo que nos rodeia.

Sendo a Estatística a ciência que trata dos dados, gostaríamos desde já de chamar a atenção para que fazer estatística é muito mais do que fazer cálculos e manipular fórmulas. Também não é matemática, embora utilize a matemática. Efectivamente, ao fazer estatística trabalhamos com dados, que são mais do que números! Como diz David Moore (1997) "*Data are numbers, but they are not "just numbers*". ***Data are numbers with a context***. *The number 10.5, for example, carries no information by itself. But if we hear that a friend's new baby weighed 10.5 pounds at birth, we congratulate her on the healthy size of the child. The context engages our background knowledge and allows us to make judgements. We know that a baby weighing 10.5 pounds is quite large, and that it isn't possible for a human baby to weigh 10.5 ounces or 10.5 kilograms. The context makes the number informative*".

Da experiência que temos no dia a dia com os dados já concluímos, com certeza, que estes apresentam **variabilidade**. Por exemplo é comum que um pacote de açúcar que na embalagem tenha escrito um quilograma, não pese exatamente um quilograma. Por outro lado ao pesar duas vezes o mesmo pacote possivelmente não obteremos o mesmo valor. Assim, ao dizermos que o peso do pacote é um determinado valor, não podemos ter a certeza que esse valor seja correto. Esta variabilidade está presente em todas as situações do mundo que nos rodeia, pelo que as conclusões que tiramos a partir dos dados que se nos apresentam, têm inerente um certo grau de incerteza.

A Estatística trata e estuda esta variabilidade apresentada pelos dados. Permite-nos a partir dos dados retirar conclusões, mas também exprimir o grau de confiança que devemos ter nessas conclusões. É precisamente nesta particularidade que se manifesta toda a potencialidade da Estatística.

Podemos então, e tal como refere David Moore em Perspectives on Contemporary Statistics, considerar três grandes áreas nesta ciência dos dados:

1. Aquisição de dados
2. Análise dos dados
3. Inferência a partir dos dados

Neste capítulo vamos abordar o primeiro tema considerado, ou seja o que diz respeito à Aquisição de Dados, numa perspectiva de que pretendemos obter dados, a partir dos quais seria possível responder a determinadas questões, isto é, posteriormente retirar conclusões para as Populações a partir das quais esses dados são adquiridos – contexto em que tem sentido fazer inferência estatística. Vamos assim, preocupar-nos em obter amostras representativas de Populações que se pretendem estudar.

1.2 – Aquisição de dados: sondagens e experimentações. População e amostra. Parâmetro e Estatística.

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". Na Estatística ensina-se a recolher dados válidos, assim como a interpretá-los.

Perante um conjunto de dados podem-se distinguir duas situações:

- Aquela em que 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;
- Uma outra em que procede à análise de dados com propósitos bem definidos no sentido de responder a questões específicas. Neste caso os dados têm que ser produzidos ou adquiridos 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 das técnicas de aquisição de dados, que se enquadram nesta última situação, em que se distinguem as

Sondagens e **Experimentações (aleatoriezadas)**

Gostaríamos desde já de realçar que o objectivo deste texto é o de explorar, de uma forma simples, algumas das técnicas de amostragem, com vista à realização de sondagens, situações que se encontram de um modo geral nas Ciências Sociais, ao contrário das Ciências experimentais, tais como Física ou Química, em que a recolha de dados se faz fundamentalmente recorrendo a experiências. Por exemplo, a população constituída pelos eleitores, a população constituída pela contas sedeadas num banco, etc, que só contêm um número finito de elementos, ao contrário da População conceptual de respostas geradas por um processo químico.

Não é demais realçar a importância desta fase, a que chamamos de Produção ou Aquisição de Dados. Como é referido em Tannenbaum (1998), página 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*".

1.2.1 –Sondagens. População e amostra. Parâmetro e Estatística.

Estas noções, que já foram dadas num módulo anterior, são aqui de novo apresentadas, unicamente com o objectivo de enquadrar o estudo seguinte, ou seja, o de introduzir algumas noções de Amostragem.

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 tais 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, pelo que 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.

Popuação, unidade, amostra

População é o conjunto de objectos, indivíduos ou resultados experimentais acerca do qual se pretende estudar alguma característica comum. As Populações podem ser finitas ou infinitas, existentes ou conceptuais. 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**.

Por exemplo, 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;

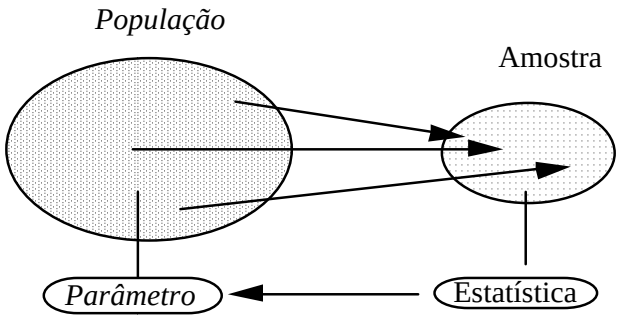
- **percentagem** de eleitores que estão decididos a votar.

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 dos dados que constituem a amostra. No caso do exemplo anterior, se se tivesse recolhido uma amostra de dimensão 1000, à característica populacional "percentagem de eleitores que estão decididos a votar" corresponde a característica amostral "percentagem dos 1000

eleitores, que interrogados disseram estar decididos a votar". Estas quantidades são conceptualmente distintas, pois enquanto a característica populacional (parâmetro) pode ser considerada um valor exacto, embora desconhecido, a característica amostral (estatística) é conhecida, embora difira de amostra para amostra, mas que todavia pode ser considerada uma estimativa útil da característica populacional respectiva.

Um **parâmetro** é uma característica numérica da população, enquanto que a **estatística** é uma característica numérica da amostra.



No entanto, para se poder utilizar as estatísticas, para estimar parâmetros é necessário que as amostras sejam representativas das populações de onde foram retiradas.

Observação – Anteriormente dissémos que uma **estatística** é um número calculado a partir dos dados da amostra, que se utiliza para estimar um parâmetro. Como, de um modo geral, podemos recolher muitas amostras diferentes, embora da mesma dimensão, teremos muitas estatísticas diferentes, como estimativas do parâmetro em estudo. Tantas as amostras diferentes (2 amostras da amesma dimensão serão diferentes se diferirem pelo menos num dos elementos) que se puderem obter da população, tantas as estimativas eventualmente diferentes que se podem calcular para o parâmetro. Então podemos considerar que todas estas estimativas são os valores observados de uma função dos elementos da amostra, a que se dá o nome de **estimador**. A esta função também se dá o nome de estatística, utilizando-se assim, indevidamente, o mesmo termo para a variável e o valor observado da variável.

É oportuno chamar a atenção para o seguinte: por vezes a População que se estuda, ou seja a **População inquirida**, não é a objecto do estudo – **População alvo ou População objectivo**. Por exemplo, se se pretende estudar a População constituída pelos indivíduos adultos de nacionalidade portuguesa - População alvo, a População inquirida pode, no entanto, ser constituída pelos indivíduos adultos de nacionalidade portuguesa e residentes no território português, à data do inquérito.

1.2.1.1 – Amostra enviesada. Amostra aleatória e amostra não aleatória.

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 opinião da população em geral.

Um processo de amostragem diz-se **enviesado** quando tende sistematicamente a seleccionar elementos de alguns segmentos da População, e a não seleccionar sistematicamente elementos de outros segmentos da População.

Surge assim, a necessidade de fazer um **planeamento da amostragem**, onde se decide quais e como devem ser seleccionados os elementos da População, com o fim de serem observados, relativamente à característica de interesse. 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 é usado, com muita frequência, pelas estações de televisão ou jornais, com resultados por vezes contraditórios com os que se obtêm quando se utiliza um processo correcto (aleatório) 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:

Amostra 1 - 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.

Amostra 2 - Uma estação de televisão preparou um debate sobre o aumento de criminalidade, onde enfatizou o facto de ter aumentado o número de crimes violentos. Ao mesmo tempo, e inserida no mesmo programa, 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 – Recolha de 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 – Utilização de 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 - Utilização dos 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 enviesadas porque tiveram a intervenção do factor humano. 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.

Quando se pretende recolher uma amostra de dimensão n, de uma População de dimensão N, podemos recorrer a vários processos de amostragem. Como normalmente o objectivo é, a partir das propriedades estudadas na amostra, *inferir* propriedades para a População, gostaríamos de obter processos de amostragem que dêem origem a "bons" estimadores. Embora a classificação de um estimador como "bom" ou não, saia fora do âmbito deste trabalho, podemos adiantar que essa análise só pode ser efectuada se conseguirmos estabelecer um plano de amostragem que atribua a cada amostra seleccionada uma determinada *probabilidade*, e esta atribuição só pode ser feita com planos de amostragem aleatórios. Assim, é importante termos sempre presente o princípio da aleatoriedade, quando vamos proceder a um estudo em que procuramos alargar para a População as propriedades estudadas na amostra.

Numa secção posterior apresentaremos **técnicas** para obter **amostras aleatórias.**

Exercícios

População e Amostra

Identifique, no que se segue, População e Amostra:

a) Numa determinada empresa, pretende-se saber qual o salário médio dos seus empregados, pelo que se recolheu informação sobre os salários mensais, auferidos pelos empregados dessa empresa;
b) Prendia-se saber a nota média obtida na prova global de Matemática no ano lectivo 2000-2001, dos alunos do 10º ano da Escola Secundária Prof. Herculano de Carvalho, pelo que se recolheu informação sobre as notas obtidas nessa disciplina por todos os alunos da Escola;
c) Pretendia-se averiguar a idade média dos alunos do 10º ano da Escola Secundária Prof. Herculano de Carvalho, pelo que se recolheu informação sobre a idade de 45 alunos do 10º ano dessa Escola;
d) Pretendia-se averiguar a quantidade de vinho produzida no Alentejo, no ano de 1999, pelo que se recolheu informação sobre as quantidades de vinho produzidas por 10 agricultores da região do Alentejo;
e) Pretendia-se estudar o salário médio auferido pelos trabalhadores da indústria têxtil, pelo que se recolheu informação sobre os salários mensais auferidos por 250 desses trabalhadores;
f) Pretendia-se averiguar a quantidade mensal de batata consumida nos lares portugueses, pelo que se recolheu informação sobre as quantidades de batata consumidas mensalmente em 100 lares portugueses;
g) Pretendia-se estudar a eficácia de um medicamento novo para curar determinada doença, pelo que se seleccionaram 20 doentes padecendo dessa doença;
h) Pretendia-se averiguar o nº de carros vendidos num dia por um stand de automóveis, pelo que se investigou junto de por cada um dos 5 empregados desse stand, quantos carros tinha vendido;
i) Pretendia-se averiguar o número de leitores dos jornais diários, pelo que se investigou junto de 6 jornais diários, o número de leitores.
j) Pretendia-se averiguar a percentagem de raparigas que frequentam o tronco comum de Matemática Aplicada da FCUL, pelo que se seleccionaram 50 alunos do dito curso.

Parâmetro e Estatística

1. Diga se são verdadeiras ou falsas as seguintes afirmações:
a) Uma estatística é um número que se calcula a partir da amostra;
b) Os parâmetros utilizam-se para estimar estatísticas;
c) A média populacional é um parâmetro;
d) Um parâmetro é uma característica numérica da variável que se está a estudar na População.

2. Identifique cada uma das quantidades seguintes, a carregado, como parâmetro ou estatística:
a) Nas últimas eleições para a Associação de Estudantes da Escola, **67%** dos estudantes que votaram, fizeram-no na lista vencedora;
b) Para obter uma estimativa do número de irmãos dos alunos que frequentam o 4º ano de uma escola básica, perguntou-se a 30 alunos, escolhidos ao acaso, quantos irmãos tinham. Verificou-se que em média, tinham **1.5** irmãos.
c) Dos 230 deputados que compõem a VIII legislatura, **21.3**% são mulheres.
d) Perguntou-se a 80 deputados qual o partido que representavem, tendo-se concluído que **49**% representavam o PS.
e) Perguntou-se a 10 deputados qual a sua idade, tendo-se concluído que a idade média era de **45** anos.

Amostras enviesadas e amostras aleatórias

1. (Adaptado de Rossman, 2001) Considere a População constituída pelos deputados da VIII legislatura, que se encontra em anexo. Seleccione 5 deputados de que já tenha ouvido falar.

a) Estes deputados constituem uma amostra ou uma população?
b) Quantos deputados, nos 5 seleccionados, pertencem ao círculo eleitoral da sua residência?
c) Suponha que está interessada em estudar o nº médio de anos de serviço dos deputados que constituem a VIII legislatura. Considera o conjunto de deputados seleccionados representativos da população? Porquê?
d) Se calculasse a média dos anos de serviço dos deputados seleccionados esperava obter um valor superior ou inferior ao da média populacional?
e) Se na sua aula ou outros colegas seleccionassem conjuntos de 5 deputados, pelo mesmo processo, isto é, deputados que lhe sejam familiares, espera que a média dos anos de serviço, tenha a mesma tendência, de sistematicamente exibir um enviesamento em determinado sentido? Explique.
f) Se tivesse seleccionado pelo mesmo processo 10 deputados, obteria uma amostra mais representativa do que a constituída pelos 5 deputados? Explique.

*1.2.2 - Experimentações

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 1.2.2.1 (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:

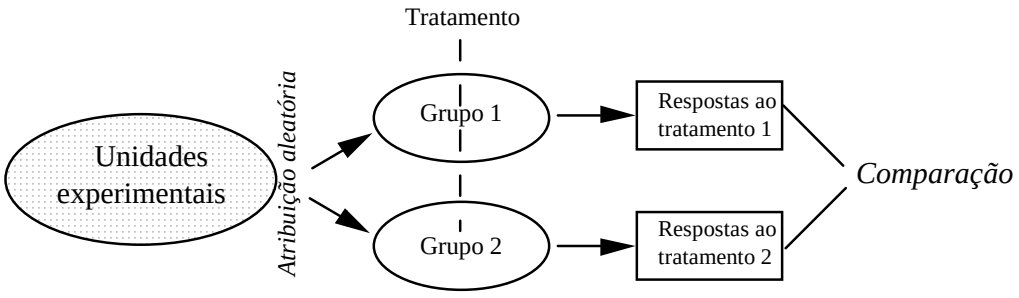
- Unidades experimentais - 22000 médicos
- Tratamentos - aspirina ou placebo

- 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 1.2.2.2 (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.

No capítulo seguinte abordaremos de forma introdutória o estudo de alguns planos de amostragem, já que um estudo conveniente do planeamento das experiências, assim como da definição da amostra adequada para o estudo em vista 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 - 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.

1.3.1 - 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 selecionada. 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 selecionado.

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 (n.º 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)!

1.3.1.1 – Números aleatórios[2]

O processo que acabámos de descrever não é prático se a população a estudar tiver dimensão elevada. Neste caso, um dos processos de seleccionar uma amostra aleatória simples consiste em utilizar números aleatórios. Estes eram fornecidos através de tabelas de dígitos aleatórios, cuja utilização já não se justifica pois, além dessass tabelas 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). No exemplo seguinte vamos utilizar o computador, mais precisamente o programa Excel, para fazer a selecção de uma amostra aleatória simples e de uma amostra aleatória com reposição.

Antes de prosseguirmos, convém ter presente que:

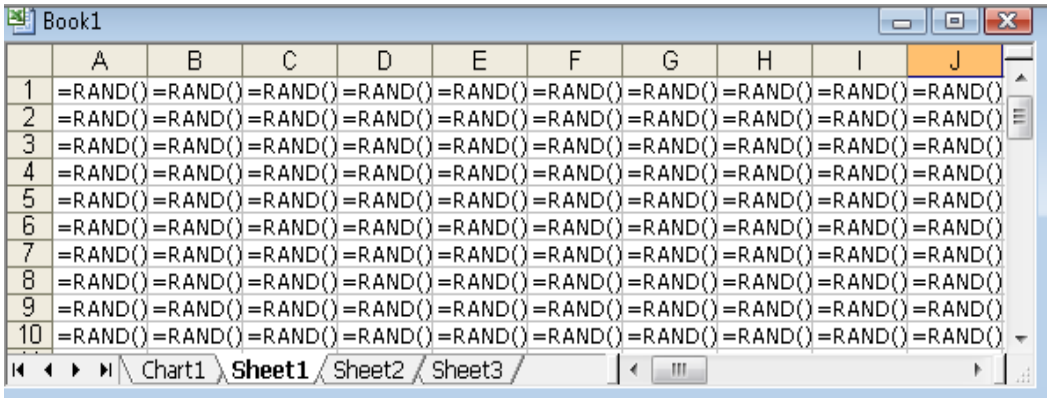
- ✓ De um modo geral, quando falamos em números aleatórios, sem qualquer outra referência, estamos a referir-nos a números reais do intervalo [0, 1].
- ✓ Os algoritmos de geração de números pseudo-aleatórios estão concebidos de modo a que ao considerar uma qualquer sequência de números gerados se obtenha aproximadamente a mesma proporção de observações em sub intervalos de igual amplitude do intervalo [0,1].

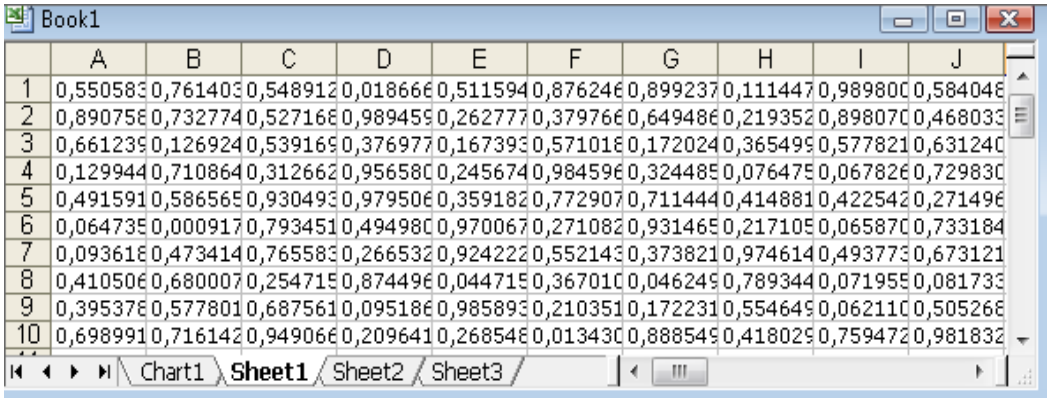
[2] Esta secção foi alterada pois com a facilidade que temos, hoje em dia, de utilizar computadores ou máquinas de calcular, já não justifica a utilização de tabelas de dígitos aleatórios.

- ✓ Assim, por exemplo, se se fizer correr o algoritmo 100 vezes, é de esperar que caiam 25 dos números gerados em cada quarto do intervalo [0,1].

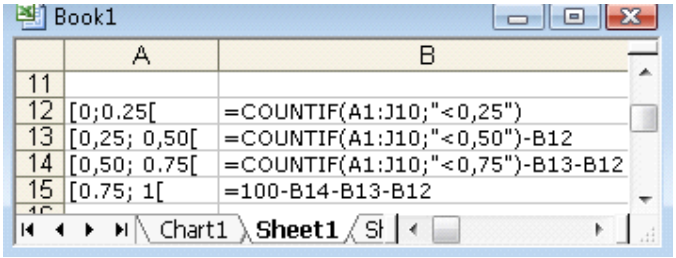
No Excel, os números pseudo-aleatórios (no intervalo [0,1]) são gerados utilizando a função *RAND()*:

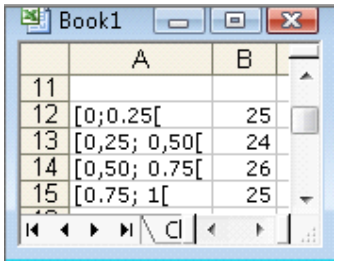
Exemplo de 100 números pseudo-aleatórios obtidos através da função *RAND()* do Excel





Para contar quantos números pertencem a cada intervalo, usamos a função *COUNTIF*, do Excel:



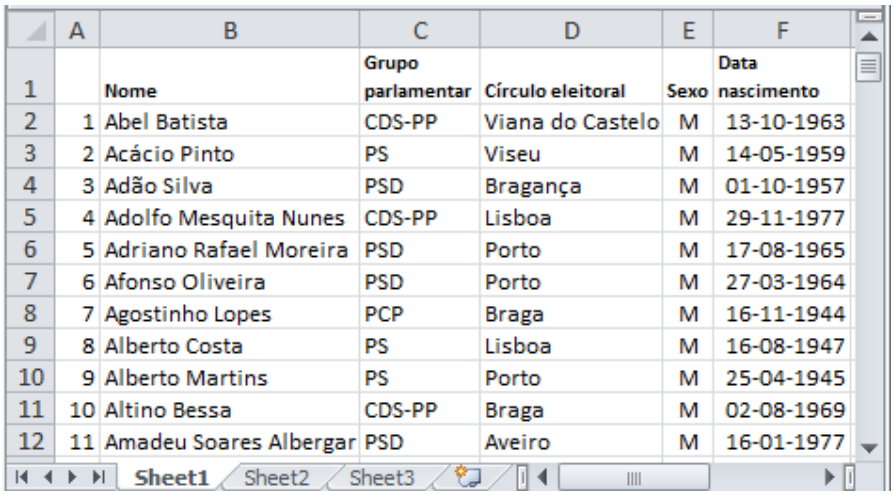


Como se verifica da tabela anterior, 25 números pertencem ao intervalo [0;0,25[, 24 ao intervalo [0,25;0,50[, 26 ao intervalo [0,50;0,75[e 24 ao intervalo [0,75;1[. Quaisquer outros 100 números, dariam outros valores para os intervalos.

1.3.1.2 - Utilização do Excel para recolher uma amostra aleatória simples e 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 1.3.1.2 – Considere a população constituída pelos 230 deputados da actual (XII) legislatura e que se encontra em Anexo. Para obter esta tabela fomos ao "site" da Assembleia da Republica, onde está uma lista ordenada com o nome de todos os deputados (coluna B), o respectivo grupo parlamentar (coluna C) e o círculo eleitoral (coluna D). Consideramos uma coluna com identificação do sexo (coluna E). Este exemplo vai-nos servir para introduzir alguns conceitos importantes, pelo que fomos completar esta lista com a idade dos deputados, acedendo à página de cada um e recolhendo a informação sobre a data de nascimento (coluna F). Nas situações de interesse, que surgem na vida real, não se vai recolher a informação sobre determinada característica, para a população toda, mas unicamente para os elementos seleccionados para a amostra. Apresentamos a seguir uma pequena parcela desse ficheiro, a que chamámos *DeputadosXII.* Este ficheiro tem uma primeira coluna (coluna A), onde é indicado o número do deputado, quando estes estão ordenados por ordem alfabética:



	A	B	C	D	E	F
1		Nome	Grupo parlamentar	Círculo eleitoral	Sexo	Data nascimento
2	1	Abel Batista	CDS-PP	Viana do Castelo	M	13-10-1963
3	2	Acácio Pinto	PS	Viseu	M	14-05-1959
4	3	Adão Silva	PSD	Bragança	M	01-10-1957
5	4	Adolfo Mesquita Nunes	CDS-PP	Lisboa	M	29-11-1977
6	5	Adriano Rafael Moreira	PSD	Porto	M	17-08-1965
7	6	Afonso Oliveira	PSD	Porto	M	27-03-1964
8	7	Agostinho Lopes	PCP	Braga	M	16-11-1944
9	8	Alberto Costa	PS	Lisboa	M	16-08-1947
10	9	Alberto Martins	PS	Porto	M	25-04-1945
11	10	Altino Bessa	CDS-PP	Braga	M	02-08-1969
12	11	Amadeu Soares Albergar	PSD	Aveiro	M	16-01-1977

Sheet1 Sheet2 Sheet3

Como dissemos anteriormente, vamos utilizá-lo para trabalhar alguns conceitos importantes, tais como:

1. **Obtenção de uma amostra aleatória simples e de uma amostra aleatória, com reposição, utilizando o Excel**
2. **Estatística e parâmetro**
3. **Variabilidade amostral**
4. **Precisão**

1. Obtenção de uma amostra aleatória simples e de uma amostra aleatória, com reposição, utilizando o Excel

Seguidamente vamos ver como seleccionar uma amostra de 10 deputados da população constituída pelos 230 deputados que constam no ficheiro *DeputadosXII*. Consideraremos os processos que conduzem a uma amostra aleatória simples sem reposição e a uma amostra aleatória com reposição.

Amostra aleatória simples

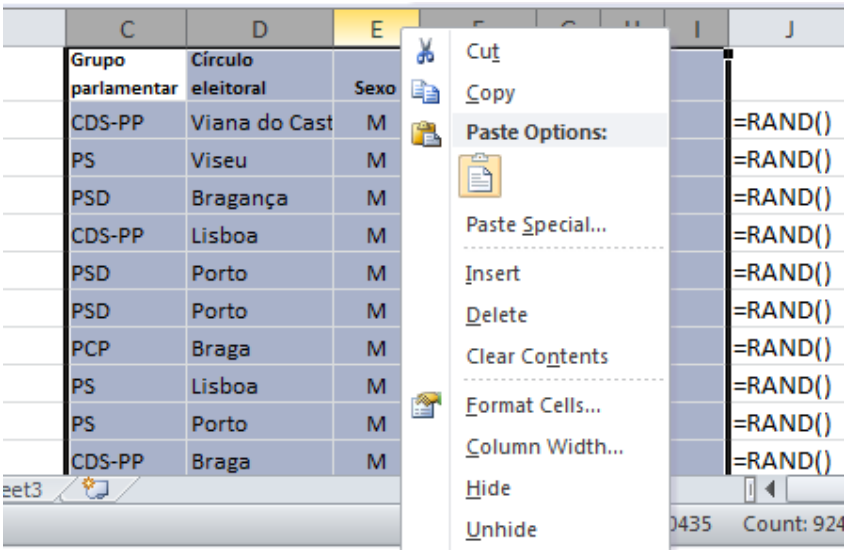
1º passo - Utilizando a função *RAND()*, atribuir um número aleatório, entre 0 e 1, a cada deputado. Para isso basta inserir a função na célula J2 e replicá-la tantas vezes, quantos os deputados (ou seja, 230 vezes):

	A	B	J
1		Nome	
2	1	Abel Batista	=RAND()
3	2	Acácio Pinto	=RAND()
4	3	Adão Silva	=RAND()
5	4	Adolfo Mesquita Nunes	=RAND()
6	5	Adriano Rafael Moreira	=RAND()
7	6	Afonso Oliveira	=RAND()
8	7	Agostinho Lopes	=RAND()
9	8	Alberto Costa	=RAND()
10	9	Alberto Martins	=RAND()
11	10	Altino Bessa	=RAND()

Sheet1 Sheet2

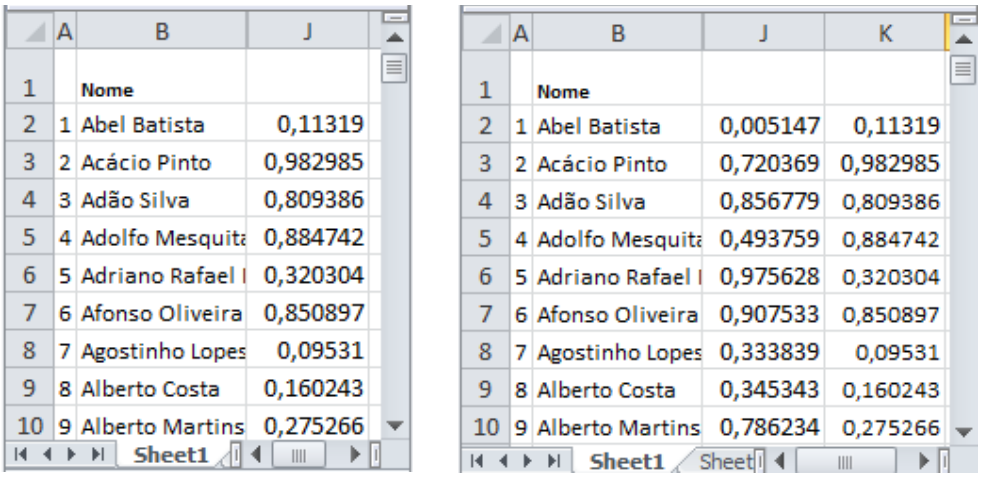
Para visualizar as fórmulas na folha de Excel basta seleccionar na barra
Formulas
Show Formulas
Conforme a imagem seguinte

Para esconder as colunas de C a I, seleccionar essas colunas, carregar com o botão direito do rato em qualquer posição da área seleccionada, fazendo aparecer o menu da figura seguinte e escolher *Hide*.



Na folha de Excel à coluna B segue-se a coluna J, dando indicação de que as colunas intermédias estão escondidas. Para mostrar as colunas escondidas, seleccionar as colunas adjacentes (neste caso a coluna B e a coluna J) e depois carregar com o botão direito do rato em qualquer posição da área seleccionada, fazendo aparecer o menu onde se escolhe *Unhide*.

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 seleccionar o botão *Paste* e fazer um *Paste Special* - *Values*, como se indica a seguir:



Colámos os valores na coluna K e fizémos o *Save*. Repare-se que os valores que estavam inicialmente na coluna J 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 K;

3º passo – Como pretendemos uma amostra de dimensão 10, seleccionar os primeiros 10 deputados do ficheiro ordenado:

	A	B	K
1		Nome	
2	87	Idália Salvador Serrão	0,002291
3	12	Ana Catarina Mendonça	0,007008
4	195	Pedro Delgado Alves	0,014659
5	112	João Soares	0,02874
6	82	Heloísa Apolónia	0,031576
7	148	Maria das Mercês Borges	0,033612
8	181	Nuno Sá	0,035773
9	153	Maria João Ávila	0,041943
10	198	Pedro Filipe Soares	0,044781
11	170	Miguel Tiago	0,045747

Sheet1 Sheet2

Os deputados seleccionados foram os números 87, 12, 195, 112, 82, 148,181, 153, 198 e 170.

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

a) Utilize a função *RANDBETWEEN*(m;n), para obter números pseudo-aleatórios entre m e n. Esta função devolve um número pseudo-aleatório entre os limites especificados nos argumentos. Para simular a extracção de uma amostra aleatória de dimensão 10, da população dos deputados, replicamos na coluna L, nas células L2:L11, a função R*ANDBETWEEN*(1;230):

	A	B	L
1		Nome	
2	1	Abel Batista	=RANDBETWEEN(1;230)
3	2	Acácio Pinto	=RANDBETWEEN(1;230)
4	3	Adão Silva	=RANDBETWEEN(1;230)
5	4	Adolfo Mesquita Nunes	=RANDBETWEEN(1;230)
6	5	Adriano Rafael Moreira	=RANDBETWEEN(1;230)
7	6	Afonso Oliveira	=RANDBETWEEN(1;230)
8	7	Agostinho Lopes	=RANDBETWEEN(1;230)
9	8	Alberto Costa	=RANDBETWEEN(1;230)
10	9	Alberto Martins	=RANDBETWEEN(1;230)
11	10	Altino Bessa	=RANDBETWEEN(1;230)

Sheet1 Sheet2

	A	B	L
1		Nome	
2	1	Abel Batista	3
3	2	Acácio Pinto	226
4	3	Adão Silva	143
5	4	Adolfo Mesquita Nunes	131
6	5	Adriano Rafael Moreira	158
7	6	Afonso Oliveira	131
8	7	Agostinho Lopes	214
9	8	Alberto Costa	44
10	9	Alberto Martins	129
11	10	Altino Bessa	134

Sheet1

Os deputados seleccionados foram os números 3, 226, 143, 131, 158, 131, 214, 44, 129 e 134 (Repare-se que houve um deputado que foi seleccionado 2 vezes, o que é possível uma vez que a selecção é com reposição).

Uma vez que a função *RANDBETWEEN* é uma função volátil, procedemos como anteriormente com a função *RAND*, e colámos os 10 valores obtidos, na coluna M:

	B	L	M
1	**Nome**		
2	Abel Batista	33	3
3	Acácio Pinto	192	226
4	Adão Silva	102	143
5	Adolfo Mesquita Nunes	207	131
6	Adriano Rafael Moreira	35	158
7	Afonso Oliveira	72	131
8	Agostinho Lopes	63	214
9	Alberto Costa	114	44
10	Alberto Martins	138	129
11	Altino Bessa	120	134

Sheet1

b) Da tabela dos deputados, seleccionar o Nome e o Grupo parlamentar dos deputados cujo número seja um dos elementos da amostra obtida anteriormente.

Para seleccionar o nome e o grupo parlamentar dos deputados correspondentes aos 10 números obtidos, vamos utilizar uma função do Excel, a função *VLOOKUP*, do seguinte modo:

	A	B	M	N	O
1		**Nome**			
2	1	Abel Batista	3	=VLOOKUP(M2;$A$2:$C$231;2)	=VLOOKUP(M2;$A$2:$C$231;3)
3	2	Acácio Pinto	226	=VLOOKUP(M3;$A$2:$C$231;2)	=VLOOKUP(M3;$A$2:$C$231;3)
4	3	Adão Silva	143	=VLOOKUP(M4;$A$2:$C$231;2)	=VLOOKUP(M4;$A$2:$C$231;3)
5	4	Adolfo Mesquita Nunes	131	=VLOOKUP(M5;$A$2:$C$231;2)	=VLOOKUP(M5;$A$2:$C$231;3)
6	5	Adriano Rafael Moreira	158	=VLOOKUP(M6;$A$2:$C$231;2)	=VLOOKUP(M6;$A$2:$C$231;3)
7	6	Afonso Oliveira	131	=VLOOKUP(M7;$A$2:$C$231;2)	=VLOOKUP(M7;$A$2:$C$231;3)
8	7	Agostinho Lopes	214	=VLOOKUP(M8;$A$2:$C$231;2)	=VLOOKUP(M8;$A$2:$C$231;3)
9	8	Alberto Costa	44	=VLOOKUP(M9;$A$2:$C$231;2)	=VLOOKUP(M9;$A$2:$C$231;3)
10	9	Alberto Martins	129	=VLOOKUP(M10;$A$2:$C$231;2)	=VLOOKUP(M10;$A$2:$C$231;3)
11	10	Altino Bessa	134	=VLOOKUP(M11;$A$2:$C$231;2)	=VLOOKUP(M11;$A$2:$C$231;3)

Sheet1 Sheet2 Sheet3

Esta função vai à tabela dos deputados, constituída pelas células (A2:C231) seleccionar o nome (**2ª** coluna da tabela seleccionada) e o Grupo Parlamentar (**3ª** coluna da tabela seleccionada) correspondente ao número da coluna A que está na coluna M, obtendo-se a seguinte amostra:

	A	B	M	N	O
1		Nome			
2	1	Abel Batista	3	Adão Silva	PSD
3	2	Acácio Pinto	226	Valter Ribeiro	PSD
4	3	Adão Silva	143	Margarida Almeida	PSD
5	4	Adolfo Mesquita Nunes	131	Luís Fazenda	BE
6	5	Adriano Rafael Moreira	158	Mariana Aiveca	BE
7	6	Afonso Oliveira	131	Luís Fazenda	BE
8	7	Agostinho Lopes	214	Rui Jorge Santos	PS
9	8	Alberto Costa	44	Carlos Zorrinho	PS
10	9	Alberto Martins	129	Lídia Bulcão	PSD
11	10	Altino Bessa	134	Luís Montenegro	PSD

Sheet1 Sheet2 Sheet3

2. Parâmetro e Estatística.

Como sabemos o parâmetro é uma característica numérica da população, enquanto que a estatística é uma característica numérica da amostra. Os parâmetros são normalmente desconhecidos e são estimados pelas estatísticas. No nosso caso, o parâmetro "Percentagem de deputados do PSD" na população, poderia ser estimado pela estatística "Percentagem de deputados do PSD" na amostra, mas não necessitamos de recorrer a uma estimativa, uma vez que podemos calcular o valor do parâmetro (esta situação não é a normalmente verificada em Estatística...).

c) Calcule a percentagem de deputados do PSD na população e na amostra

Vamos começar por utilizar a função *COUNTIF*, que inserimos na célula P2, e que conta o nº de células, de entre um conjunto especificado de células, que satisfazem determinado critério, sendo este critério, no caso presente, o de serem iguais a "PSD":

P2 =COUNTIF(C2:C231;"PSD")

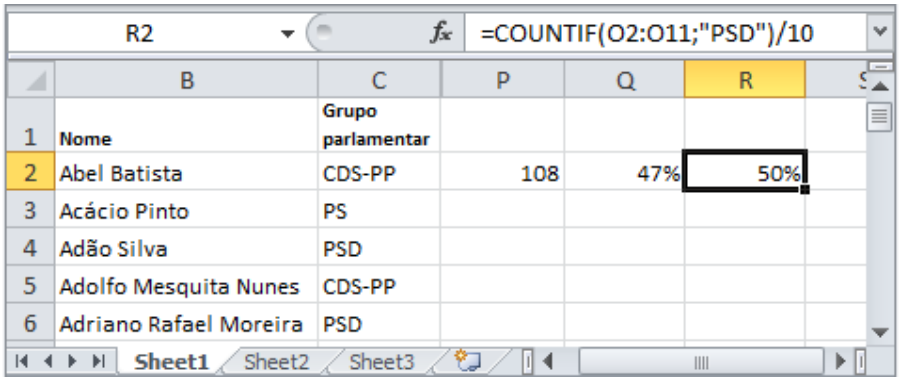
	B	C	P	Q	R
1	Nome	Grupo parlamentar			
2	Abel Batista	CDS-PP	108		
3	Acácio Pinto	PS			
4	Adão Silva	PSD			
5	Adolfo Mesquita Nunes	CDS-PP			
6	Adriano Rafael Moreira	PSD			

Sheet1 Sheet2 Sheet3

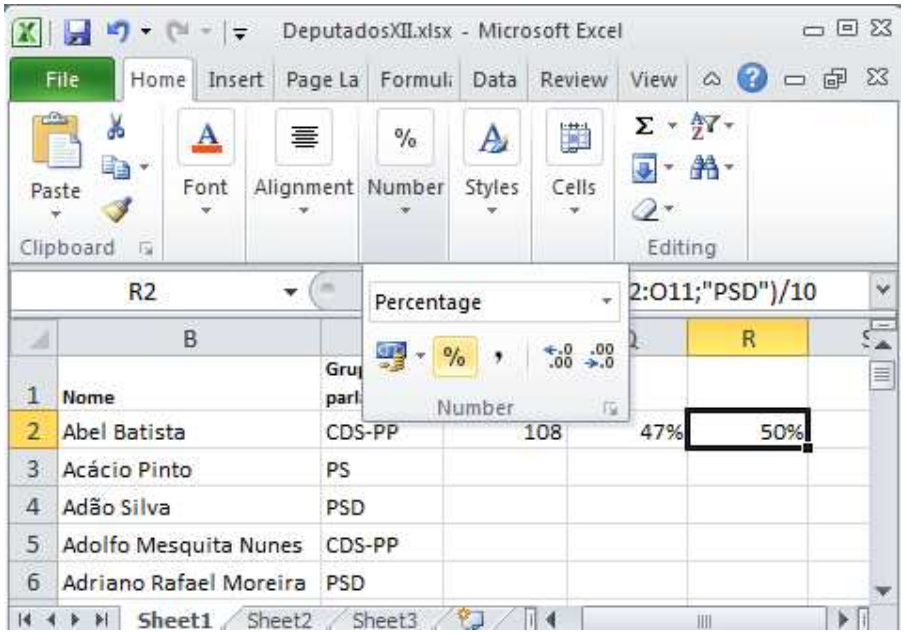
O valor devolvido pela função *COUNTIF*(C2:C231; "PSD") foi 108, pelo que o valor do parâmetro em estudo é 47% (=$\frac{108}{230}$ *100%) (Ver célula Q2 da figura seguinte).

Caso não fosse possível obter o valor do parâmetro, teríamos calculado uma estimativa desse valor a partir da amostra de 10 elementos anteriormente seleccionados. Assim, uma

estimativa seria dada pelo "nº de deputados do PSD na amostra de 10 deputados"/10, ou seja 50%



Para visualizar o valor obtido, em percentagem, basta seleccionar no menu o botão *Number* e aí seleccionar %



3. Variabilidade amostral

d) Repita 10 vezes o processo descrito nas alíneas anteriores, de seleccionar 10 deputados e calcular a percentagem de deputados do PSD, e registe numa tabela os resultados obtidos.

Gerámos 10 amostras e obtivémos os seguintes resultados para a estatística - percentagem de deputados PSD, em cada uma das amostras:

Amostra	1	2	3	4	5	6	7	8	9	10
% PSD	40%	30%	50%	30%	50%	40%	60%	40%	60%	40%

Repare-se na variabilidade apresentada nos resultados obtidos para as diferentes amostras. Os 10 valores obtidos para a percentagem de deputados do PSD existentes em cada uma delas, representam outras tantas estimativas para a verdadeira proporção de deputados

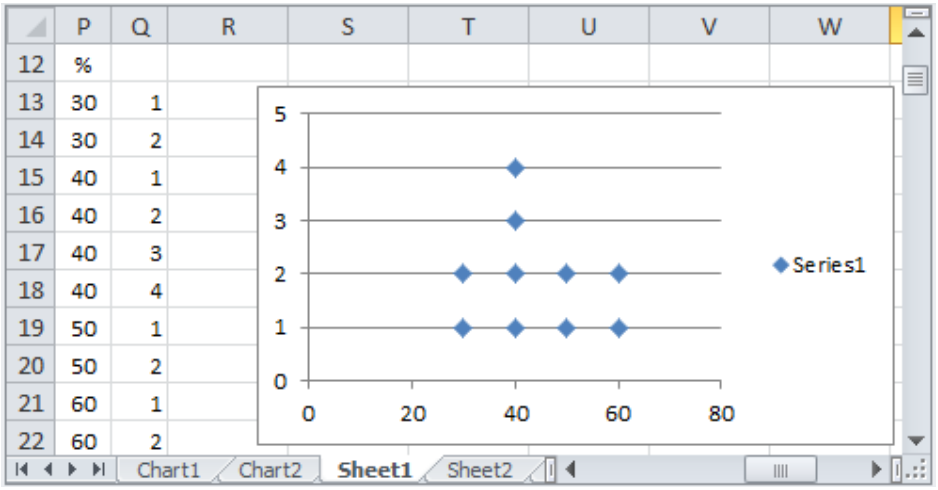
existentes na População. Iremos ilustrar esta variabilidade, representando os valores num diagrama de pontos, utilizando uma opção gráfica do Excel, o *Scatter*. Para obter a representação gráfica pretendida, é necessário começar por construir uma tabela adequada:

	P	Q
12	%	
13	30	1
14	30	2
15	40	1
16	40	2
17	40	3
18	40	4
19	50	1
20	50	2
21	60	1
22	60	2

Para construir esta tabela, pode-se utilizar a seguinte metodologia: consideram-se duas colunas, onde na primeira coluna se representam todos os elementos do conjunto de dados, pela ordem em que aparecem, e na segunda coluna indica-se a frequência absoluta com que cada elemento surge no conjunto de dados, à medida que se vai percorrendo a coluna, de cima para baixo. Por exemplo, ao lado do primeiro elemento que é o 60%, indicamos um 1, mas a segunda vez que aparece o 60%, indicamos um 2, etc. Se, à partida, dispusessemos de uma tabela de frequências, para construir esta nova tabela, bastaria repetir cada elemento da amostra, tantas vezes quantas a sua frequência absoluta.

Na folha do Excel, seleccionam-se as céluas que contêm os dados (P13:Q22) e na barra menu selecionar *Insert* →*Scatter*, 1º subtipo.

Obtém-se o diagrama de pontos com o seguinte aspecto:

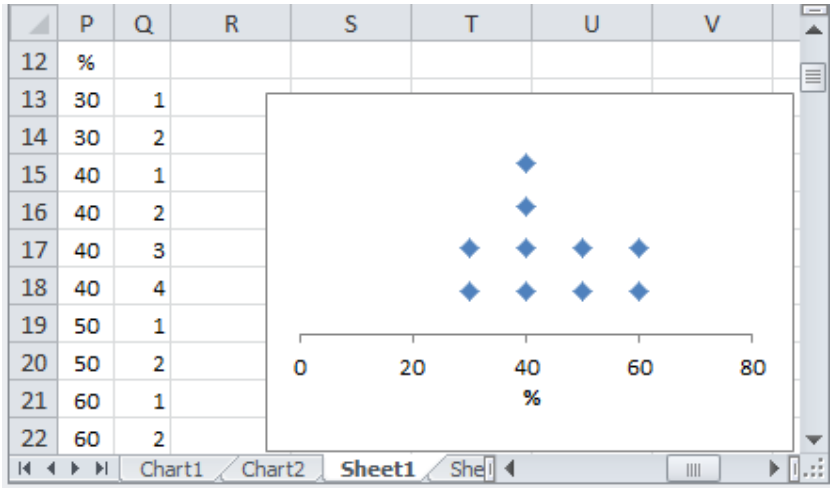


Trabalhámos "esteticamente" esta representação, seguindo os seguintes passos:

Com o gráfico selecionado, selecionar:

Legenda e carregar no botão *Delete*;
As linhas e carregar no botão *Delete*;
Layout →*Axes* →*Primary* →*Vertical Axis* →*None*;
Layout →*Axis Titles* →*Primary horizontal Axis Title* →*Title Below Axis*
Escrever %;

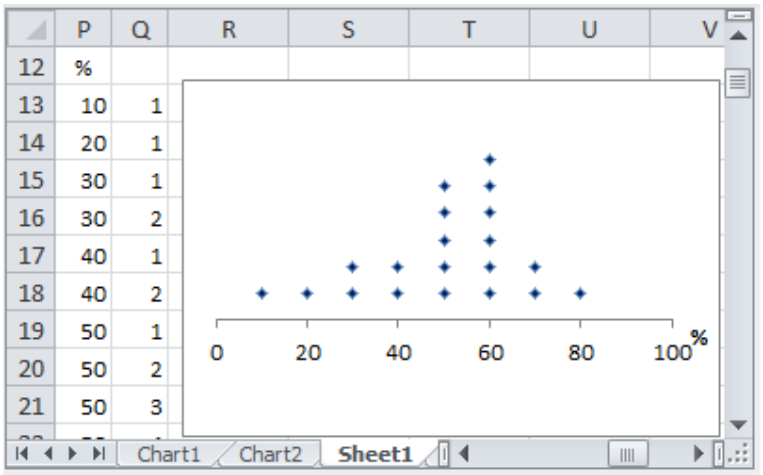
Temos finalmente a seguinte representação:



Da representação gráfica anterior começamos a adivinhar que a distribuição das estimativas apresenta um padrão com uma certa simetria relativamente a um valor compreendido entre 40% e 50%.

e) Considere agora 20 amostras de dimensão 10, calcule para cada uma o valor da estatística em estudo, e construa o diagrama de pontos dos valores obtidos.

Seleccionámos 20 amostras de dimensão 10, caculámos a percentagem de deputados do PSD em cada uma delas e com os resultados obtidos construímos a seguinte representação:

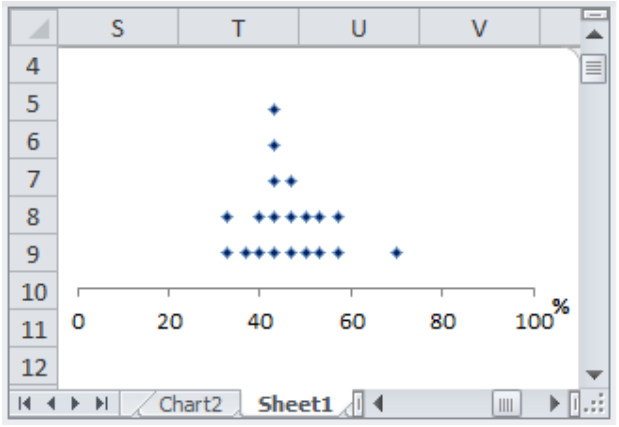


Esta representação é mais elucidativa e reforça a ideia avançada anteriormente, de que o valor do parâmetro em estudo – percentagem de deputados do PSD, se deve situar perto de 50%. Tendo em conta que a verdadeira percentagem de deputados do PSD na população é 47%, apesar de o valor apresentado pela estatística variar de amostra para amostra – **variabilidade amostral**, estes valores apresentam uma distribuição que nos dá informação sobre o parâmetro, já que essa distribuição se localiza ou está **centrada** em torno do parâmetro.

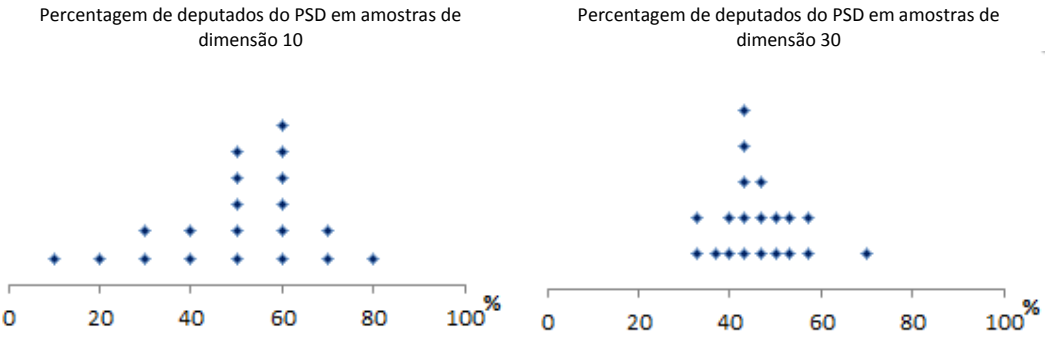
4. Precisão

f) Considere agora 20 amostras de dimensão 30, calcule para cada uma o valor da estatística em estudo, e construa o diagrama de pontos dos valores obtidos. Compare a representação obtida, com a que obteve na alínea e).

Seguimos um processo idêntico ao da alínea e), com a particularidade de as dimensões das amostras terem dimensão 30, em vez de 10. Com as percentagens de deputados do PSD existentes em cada uma delas construímos a seguinte representação gráfica:



Comparando as duas representações obtidas quando se consideram amostras de dimensão 10 ou de dimensão 30, verificamos que a variabilidade apresentada pelos valores da estatística - percentagem de deputados do PSD, no caso das amostras de maior dimensão, é inferior à apresentada pela estatística no caso das amostras de menor dimensão, como se vê na figura seguinte:



A **precisão** de um estimador é caracterizada pela variabilidade apresentada pelas diferentes estimativas, obtidas quando se consideram diferentes amostras. Quanto menor for a variabilidade apresentada pelas diferentes estimativas, maior é a precisão apresentada pelo estimador.

De um modo geral, diz-se que uma estatística é um "bom" estimador de um certo parâmetro, se a distribuição dos valores dessa estatística, calculados para diversas amostras, revelar uma **localização em torno do parâmetro** e apresentar **pequena variabilidade**. Em alguns casos essa análise pode fazer-se do ponto de vista teórico. No entanto, hoje em dia,

cada vez se recorre mais à simulação para decidir se um estimador é ou não, um "bom" estimador do parâmetro de interesse.

Observação: Este exemplo que acabámos de descrever tem como objectivo apresentar alguns conceitos importantes, como o da variabilidade e das propriedades de um estimador. Efectivamente, neste caso, já que temos informação sobre o grupo parlamentar de cada deputado, não teria muito sentido ir recolher uma amostra para obter a percentagem de deputados em cada grupo parlamentar. Repare-se, no entanto, que se o que estivesse em estudo fosse "ter uma ideia" sobre o número médio de filhos dos deputados portugueses e suas idades, já faria sentido recolher uma amostra, pois para obter a informação desejada não seria necessário interrogar todos os deputados e só se interrogariam os seleccionados para a amostra.

1.3.2 - Amostragem 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 – os elementos são escolhidos de uma maneira regular percorrendo a lista.

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 nem todas as amostras possíveis de dimensão n, têm a mesma probabilidade de serem seleccionadas.

1.3.2.1 - Utilização do Excel para recolher uma amostra aleatória sistemática

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

Exemplo 1.3.2.1 – Considere novamente o ficheiro DeputadosXII, que contém o nome, filiação partidária, sexo e data de nascimento dos 230 deputados da actual legislatura e que se encontra em Anexo. Utilizando o processo de amostragem sistemática, obtenha uma amostra de 12 deputados, registando para cada um deles o sexo.

Temos uma população de dimensão 230 e pretendemos obter uma amostra de dimensão 12. Vamos utilizar a seguinte metodologia:

Passo 1 – Dividindo 230 por 12 e retendo a parte inteira, obtém-se o valor 19.

Passo 2 – Dos primeiros 19 elementos da lista ordenada dos deputados, selecciona-se um elemento ao acaso. Vimos na secção anterior que basta utilizar a função Randbetween(1;19), que inserimos na célula T2. A utilização desta função devolveu-nos o deputado número 15.

Passo 3 – A amostra será constituída pelos deputados números 15, 34, 53, 72, 91, 110, 129, 148, 167, 186, 205, 224, que obtivémos adicionando sucessivamente 19, até obtermos 12 elementos (células T2:T13).

Passo 4 - Utilizando a função *VLOOKUP*(T2;$A$2:$E$231;5), replicada pelas 12 células U2:U13, obteve-se finalmente a informação solicitada, constituída pelo sexo dos 12 deputados seleccionados para a amostra:

	T	U
2	15	F
3	34	F
4	53	M
5	72	M
6	91	F
7	110	M
8	129	F
9	148	F
10	167	M
11	186	M
12	205	M
13	224	F

1.3.3 – Amostragem estratificada

Pode acontecer que a população possa ser dividida em várias subpopulações ou estratos, mais ou menos homogéneos, relativamente à característica a estudar. Nesta situação existe uma técnica importante e apropriada, que é a amostragem por estratificação. Apresentamos de seguida um exemplo em que privilegiaremos a exemplificação da técnica, em detrimento da apresentação em Excel, uma vez que o tipo de amostragem utilizado, se resume a uma amostragem aleatória simples, já exemplificada anteriormente.

Exemplo 1.3.3.1 (Ted Hodgson and John Borkowski *in* Getting the Best from Teaching Statistics) – Consideremos uma população constituída por 40 cartões numerados (20 vermelhos e 20 pretos) de acordo com a seguinte tabela:

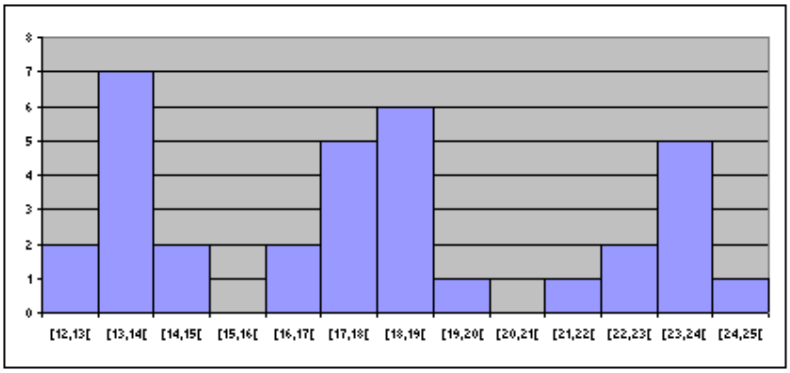
Nº	6	7	8	9	10	26	27	28	29	30
Freq.	4	4	4	4	4	4	4	4	4	4
Cor	Ver	Ver	Ver	Ver	Ver	Preto	Preto	Preto	Preto	Preto

A média dos números inscritos nesta população de 40 cartões é de 18 – valor médio da característica populacional em estudo.

Pretende-se, através de uma amostra, obter alguma indicação sobre a média dos números inscritos nos cartões (a qual neste exemplo fictício é conhecida). Colocam-se os cartões num saco e pede-se a cada aluno da turma que retire uma amostra de 4 cartões – amostra aleatória simples, e que calcule a média dos números dos cartões seleccionados. Numa turma de 34 alunos, obtiveram-se os seguintes resultados:

Amostra nº					Média
1	26	7	10	6	12,25
2	10	26	9	6	12,75
3	29	6	7	10	13
4	6	8	9	29	13
5	6	9	8	30	13,25
6	9	8	7	29	13,25
7	7	7	30	9	13,25
8	9	9	10	26	13,5
9	9	8	8	30	13,75

10	9	10	8	29	14
11	10	9	29	9	14,25
12	6	27	6	26	16,25
13	7	7	26	27	16,75
14	28	8	6	26	17
15	7	6	29	26	17
16	6	29	26	8	17,25
17	9	6	26	29	17,5
18	26	9	8	28	17,75
19	7	10	26	29	18
20	27	6	30	9	18
21	6	29	28	10	18,25
22	8	29	26	10	18,25
23	6	8	30	30	18,5
24	26	9	30	10	18,75
25	8	11	28	30	19,25
26	26	27	6	27	21,5
27	30	26	27	6	22,25
28	8	26	29	28	22,75
29	10	26	26	30	23
30	29	6	30	27	23
31	28	9	30	26	23,25
32	27	26	30	10	23,25
33	30	10	29	26	23,75
34	29	30	7	30	24



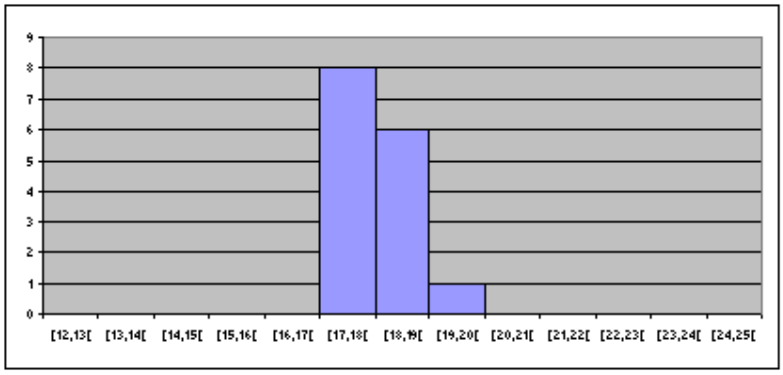
Esta distribuição não nos ajuda muito a dizer qual a estimativa para o valor médio da população (média dos números inscritos). Gostaríamos de ter obtido para a amostra, cujos elementos são as diferentes médias, uma distribuição com pouca variabilidade, para podermos argumentar que a média destes elementos era uma "boa" estimativa para o parâmetro em estudo, ou seja, o valor médio dos números inscritos nos cartões (Ver secção seguinte).

Diz-se então aos alunos que estamos perante duas subpopulações, a de cartões vermelhos e a de cartões pretos, embora não seja esta a característica em estudo e sobre a qual seria importante haver diferença entre os estratos ou subpopulações. De qualquer modo aqueles são informados que poderá haver diferenças relativamente à característica de interesse e que um processo de amostragem adequado levaria em conta essas diferenças.

Procede-se então a uma selecção da amostra, de forma a obter 2 cartões vermelhos e 2 cartões pretos – estes valores devem reflectir a dimensão dos estratos (que no nosso exemplo são iguais). Os resultados obtidos foram os seguintes:

Amostra nº					média
1	6	7	27	28	17
2	8	9	26	27	17,5
3	8	6	28	28	17,5
4	7	8	29	26	17,5
5	9	9	26	26	17,5
6	6	9	29	27	17,75
7	8	10	26	27	17,75
8	10	6	27	28	17,75
9	9	9	28	26	18

10	6	8	28	30	18
11	10	8	27	28	18,25
12	10	7	28	29	18,5
13	9	9	27	29	18,5
14	8	9	29	29	18,75
15	9	10	28	29	19



A partir dos dados obtidos para as amostras, confirma-se que efectivamente temos dois estratos distintos, relativamente à característica de interesse – um estrato com cartões com números mais pequenos e outro estrato com cartões com números maiores.

Estes resultados mostram que as médias das amostras estratificadas estão consistentemente próximas do valor médio da população (o qual só deve ser dito aos alunos depois das simulações serem feitas), podendo-se assim observar que a estratificação conduziu a um aumento da precisão.

*1.3.4 – Estimador centrado e não centrado. Precisão

Uma vez escolhido um plano de amostragem aleatório, ao pretendermos estimar um parâmetro, pode ser possível utilizar várias estatísticas (estimadores) diferentes. Por exemplo, quando pretendemos estudar a variabilidade presente numa População, que pode ser medida pela variância populacional σ^2, sabemos que podemos a partir de uma amostra, obter duas estimativas diferentes para essa variância, a partir das expressões

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad \text{ou} \qquad s'^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Quais as razões que nos podem levar a preferir s^2 em vez de s'^2?

Um critério que costuma ser aplicado é o de escolher um "bom" estimador como sendo aquele que é *centrado* e que tem uma boa *precisão.* Escolhido um plano de amostragem, define-se:

Estimador centrado – Um estimador diz-se *centrado* quando a média das estimativas obtidas para todas as amostras possíveis que se podem extrair da População, segundo o esquema considerado, coincide com o parâmetro a estimar. Quando se tem um estimador *centrado*, também se diz que é *não enviesado*.

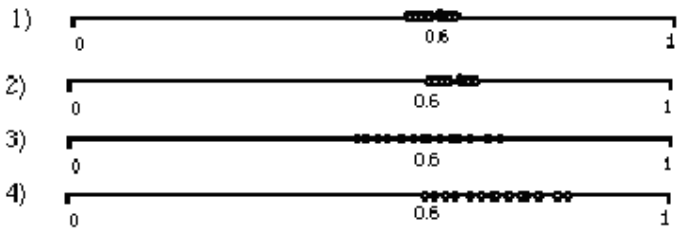
A média das estimativas calculadas a partir da expressão s^2 acima considerada, coincide com a variância σ^2.

Para se evitar o enviesamento, é necessário estarmos atentos, primeiro na escolha do plano de amostragem e depois na escolha do estimador utilizado para estimar o parâmetro desconhecido. O facto de utilizarmos um estimador centrado, não nos previne contra a obtenção de más estimativas, se o plano de amostragem utilizado sistematicamente favorecer uma parte da População (isto é, fornecer amostras enviesadas).

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, e o estimador for centrado, 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 da amostra** são dois tipos de erro com que nos defrontamos num processo de amostragem (mesmo que tenhamos escolhido um "bom" estimador). 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 1.3.4.1 - 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. Considerando como *estimador*, a proporção de elementos na amostra apoiantes do candidato, 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 183. Então a estimativa obtida seria 0.61. A repetição deste processo 15 vezes permitiria obter 15 valores para o estimador, que seriam outras tantas estimativas do parâmetro a estimar - *proporção* de eleitores da cidade, potenciais apoiantes do tal candidato. Representando num eixo os valores obtidos e admitindo que o verdadeiro valor do parâmetro era 0.60, poderíamos deparar-nos com várias situações:



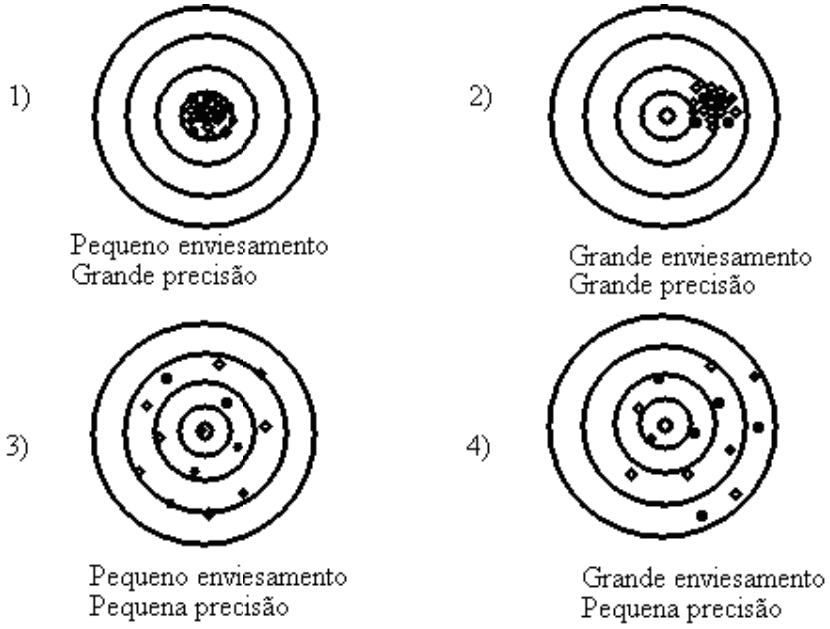
1) reflecte um *pequeno* ou ausência de *enviesamento*, pois os valores para a estatística (proporções obtidas a partir das amostras) situam-se para um e outro lado do valor do parâmetro, e verifica-se ainda a existência de uma pequena variabilidade entre os resultados obtidos para as várias amostras, que se traduz em *grande precisão*.
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. Presume-se

que o esquema de amostragem não seja aleatório, pelo que as amostras só reflectem parte da População.

3) voltamos a ter uma situação de *pequeno enviesamento*, mas de *pequena precisão* devido à grande variabilidade apresentada pelos valores da estatística. Presumimos que as amostras não têm a dimensão suficiente, de forma a garantir uma melhor precisão.

4) a *falta de precisão* da situação 3) é acompanhada de um *grande enviesamento*.

Como sugere Moore (1996), 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



O estudo de um estimador é feito através da sua *distribuição de amostragem*, ou seja, da distribuição dos valores obtidos pelo estimador, quando se consideram todas as amostras possíveis.

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, que se podem seleccionar da população.

A forma da distribuição de amostragem, permite-nos verificar se esses valores se distribuem de forma tal, que a sua média coincide com o parâmetro a estimar – caso em que o estimador é centrado, e além disso se apresenta grande ou pequena variabilidade – o que faz com que o estimador apresente, respectivamente, menor ou maior precisão.

A maior parte das vezes não se consegue obter a distribuição de amostragem exacta, mas tem-se uma distribuição aproximada, considerando um número suficientemente grande de amostras da mesma dimensão e calculando para cada uma delas o valor da estatística que é uma estimativa do parâmetro em estudo.

***1.3.5 - 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, de forma a obter a mesma precisão que no caso anterior, 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 (*Statistics: a Tool for the Social Sciences*, Mendenhall et al., 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*".

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!

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

Além dos problemas relacionados com a amostragem e 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, e como já foi referido anteriormente, muitas vezes a recolha da amostra faz-se de uma população que não é a população que se pretende estudar – *população alvo ou 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.

1.4 - Estatística Descritiva e Inferência Estatística

Uma vez recolhida a amostra procede-se ao seu estudo. Este consiste em resumir a informação contida na amostra construindo *tabelas, gráficos e* calculando algumas *características amostrais* - **estatísticas**. Este estudo descritivo dos dados é o objectivo da *Estatística Descritiva*. Esta fase é a que depende mais da habilidade ou intuição do estatístico (dissemos no início do capítulo que a Estatística além de uma ciência, também é uma arte!). Efectivamente ele vai tentar substituir o conjunto de dados, por um sumário desses dados de forma a realçar a informação que eles contêm. Pense-se o que se passa, por analogia, com um texto comprido e repetitivo em que a pessoa se perde na leitura. Um sumário bem feito do texto, em algumas linhas, dará a informação relevante sobre o texto, que ocupava muito mais linhas. Ao ler o sumário a pessoa fica rapidamente informada sobre o assunto que trata. O mesmo se passa com os dados, sendo necessário que o sumário desses dados seja feito adequadamente de forma a não se perder muita informação, mas também de forma a não sumariar tão pouco que a pessoa seja submergida por tanta informação!

Por exemplo, suponha que perguntou a um aluno se ele foi bom aluno na licenciatura que tirou. Ele responde-lhe com as notas que teve durante os 4 anos que durou a licenciatura:

10	16	11	10	15	17	12	13	17	15	18	14
15	16	12	13	16	11	15	16	12	13	14	14
11	15	17	16	16	13	14	16				

Perante estes dados hesitará um pouco, pois não se vê facilmente qual o tipo de notas que predomina. No entanto se fizer uma representação gráfica muito simples:

10	* *
11	* * *
12	* * *
13	* * * *
14	* * * *
15	* * * * *
16	* * * * * * *
17	* * *
18	*

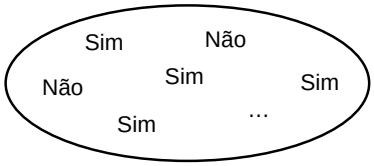
imediatamente concluirá que metade das notas são iguais ou superiores a 15, pelo que se pode considerar um aluno bom. Organizámos os dados através de uma representação gráfica sugestiva, que permitiu realçar a informação desejada. Outro processo seria resumir a informação sob a forma de uma medida que se calculava a partir dos dados (estatística) - a média, que viria igual a 14.2.

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**.

Esta quantificação do erro cometido, ao transportar para a população as propriedades verificadas na amostra, é feita utilizando a Probabilidade. Efectivamente, é nesta fase do processo estatístico que temos necessidade de entrar com este conceito, para quantificar a incerteza associada aos procedimentos aqui considerados. Repare-se que ao transportar

para a população uma propriedade verificada na amostra não podemos dizer que essa propriedade é verdadeira porque não a verificamos em todos os elementos da população, mas também não podemos dizer que é falsa, pois a propriedade foi verificada por alguns elementos da população - a mostra. Assim, estamos numa situação entre o que é verdadeiro e falso, caracterizada por uma incerteza, a qual é medida com a utilização da probabilidade.

Exemplo 1.4.1 - O Senhor X, candidato à Câmara da cidade do Porto, pretende saber, qual a percentagem de eleitores que pensam votar nele nas próximas eleições. Havendo algumas limitações de tempo e dinheiro, a empresa encarregada de fazer o estudo pretendido decidiu recolher uma amostra de dimensão 1000, perguntando a cada eleitor se sim ou não pensava votar no Senhor X. Como resultado da amostragem obteve-se um conjunto de sim's e não's, cujo aspecto não é muito agradável, pois à primeira vista não conseguimos concluir nada:



Procede-se à redução dos dados, resumindo a informação sobre quantos sim's se obtiveram, chegando-se à conclusão que nas 1000 respostas, 635 foram afirmativas. Então dizemos que a percentagem de eleitores que pensam votar no candidato, de entre os inquiridos, é de 63.5%. A função da Estatística Descritiva acabou aqui! (Se toda a População tivesse sido inquirida, este estudo descritivo dar-nos-ia a informação necessária para o fim em vista).

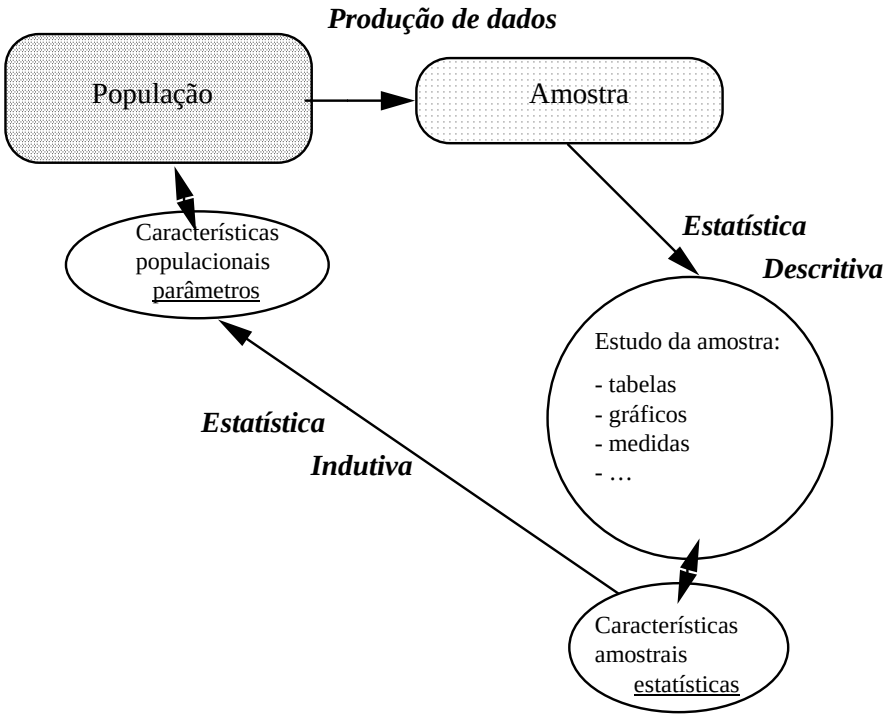
Poderemos agora inferir que 63.5% dos eleitores da cidade do Porto pensam votar no Senhor X? A resposta a esta pergunta nem é sim, nem não, mas talvez. É agora que temos necessidade de utilizar o conceito de Probabilidade, para quantificar a incerteza associada à inferência. Assim, existem processos de inferência estatística que, do resultado obtido a partir da amostra, nos permitirão concluir que o intervalo [60.5%, 66.5%] contém o valor exacto para a percentagem de eleitores da cidade que pensam votar no Senhor X, com uma confiança de 95%.

Observação - A confiança de 95% deve ser entendida no seguinte sentido: se se recolherem 100 amostras, cada uma de dimensão 1000, então poderemos construir 100 intervalos; destes 100 intervalos esperamos que 95 contenham o verdadeiro valor da percentagem (desconhecida) de eleitores da cidade do Porto, que pensam votar no candidato. Como ao fazer um estudo só se recolhe uma amostra, não sabemos se a nossa é uma das que deu origem a um dos intervalos que continha o parâmetro. Estamos confiantes que sim!

Recorde-se a forma como as previsões são dadas, em noite de eleições, sob a forma de intervalos. Por vezes a guerra de audiências faz com que estas previsões tenham pouco sentido, por apresentarem intervalos com uma tão grande amplitude que a sua precisão, como estimativas das percentagens pretendidas, é muito pequena. Esta situação prende-se com o facto de as amostras utilizadas para a construção dos intervalos terem uma dimensão muito reduzida, havendo assim muito pouca informação disponível (recorde-se o que dissemos anteriormente sobre o processo para aumentar a precisão). No entanto, à medida que a noite vai avançando, os intervalos vão diminuindo de amplitude, estando esta

diminuição da amplitude relacionada com a dimensão da amostra que entretanto vai aumentando, até finalmente estarem todos os votos contados. Nesta altura, os intervalos reduzem-se a pontos, que são as percentagens pretendidas - a amostra é constituída por toda a população.

O seguinte esquema pretende resumir as diferentes etapas que normalmente são seguidas num procedimento estatístico:



No esquema anterior a necessidade de utilizar o conceito de probabilidade faz-se sentir ao passarmos das propriedades estudadas na amostra para as propriedades na população, sendo aqui precisamente que vai ser necessário invocar o princípio da aleatoriedade.

Chama-se a atenção para que a compreensão do processo estatístico permitir-nos-á interpretar melhor notícias que, frequentemente, se lêem nos jornais ou ouvem na televisão. Por vezes alguns estudos sobre os mesmos assuntos, apresentam resultados contraditórios! Isto acontece nomeadamente no estudo de certos aspectos do comportamento humano, utilizando testes psicológicos, ou no estudo de certas doenças utilizando cobaias. Muitas das inferências feitas são imperfeitas, a maior parte das vezes por terem como base dados imperfeitos.

2. Representação e redução de dados. Tabelas e gráficos

2.1- Introdução

Num módulo anterior de Estatística, já foram apresentados vários processos de organizar os dados, de forma a realçar as características principais e a estrutura subjacente da população de onde esses dados foram retirados.

Quer estejamos perante uma variável de tipo discreto ou contínuo, o processo de organizar a informação consiste em, de um modo geral, começar por construir tabelas de frequência e proceder a representações gráficas adequadas. Vamos seguidamente utilizar o Excel na construção dessas tabelas de frequência.

2.2 – Utilização do Excel na obtenção de tabelas de frequência

Vamos exemplificar a utilização do Excel na construção de tabelas de frequência a partir do ficheiro *DeputadosXII*, apresentado no capítulo anterior.

2.2.1 – Tabela de dados qualitativos ou quantitativos discretos

O procedimento para a construção das tabelas de frequência é idêntico, quer tenhamos um conjunto de dados qualitativos ou quantitativos discretos, já que as classes que se consideram são as diferentes categorias ou valores que surgem, respetivamente, no conjunto de dados. A seguir apresentamos a construção destas tabelas utilizando a função *COUNTIF*. Numa secção posterior veremos a sua construção utilizando a metodologia da funcionalidade *PivotTables*.

Exemplo 2.2.1 – Considere o ficheiro *DeputadosXII*. Obtenha uma tabela de frequências para a variável Grupo Parlamentar.

Começámos por copiar a coluna correspondente ao Grupo parlamentar para um novo ficheiro. Ordenámos os elementos por ordem crescente e inserimos na coluna **Classes** os diferentes elementos do conjunto de dados. Utilizámos de seguida a função *COUNTIF* para obter as frequências absolutas de deputados de cada um dos grupos parlamentares:

	A	B	C	D	E
1	Grupo parlamentar			Tabela de frequências	
2	BE		Classes	Freq. Abs.	Freq. Rel.
3	BE		BE	=COUNTIF($A$2:$A$231;C3)	=D3/$D$9
4	BE		CDS-PP	=COUNTIF($A$2:$A$231;C4)	=D4/$D$9
5	BE		PCP	=COUNTIF($A$2:$A$231;C5)	=D5/$D$9
6	BE		PEV	=COUNTIF($A$2:$A$231;C6)	=D6/$D$9
7	BE		PS	=COUNTIF($A$2:$A$231;C7)	=D7/$D$9
8	BE		PSD	=COUNTIF($A$2:$A$231;C8)	=D8/$D$9
9	BE			=SUM(D3:D8)	=SUM(E3:E8)

Sheet1 Sheet2 Sheet3

As fórmulas apresentadas anteriormente, deram origem à seguinte tabela:

	A	B	C	D	E
1	Grupo parlamentar		Tabela de frequências		
2	BE		Classes	Freq. Abs.	Freq. Rel.
3	BE		BE	8	0,035
4	BE		CDS-PP	24	0,104
5	BE		PCP	14	0,061
6	BE		PEV	2	0,009
7	BE		PS	74	0,322
8	BE		PSD	108	0,470
9	BE			230	1

Sheet1 Sheet2

2.2.2 – Tabela de dados quantitativos contínuos

Como se viu no módulo anterior de Estatística, no caso de dados contínuos o processo de construção das tabelas é um pouco mais elaborado, já que a definição das classes não é tão imediata. De um modo geral as classes são intervalos com a mesma amplitude, fechados à esquerda e abertos à direita ou abertos à esquerda e fechados à direita. Em certos casos não é conveniente que as classes tenham a mesma amplitude, o que em si não é um problema para a construção da tabela de frequências, mas que implica alguma complicação na construção do histograma associado, quando pretendemos utilizar Excel.

Vamos utilizar ainda o ficheiro *DeputadosXII* para estudar a variável Idade, que é uma variável quantitativa contínua.

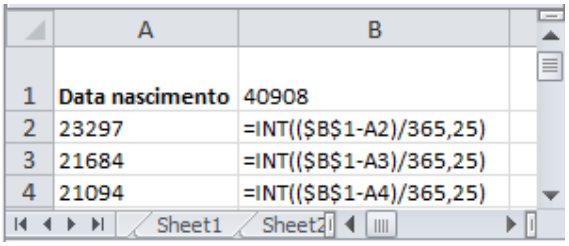
Exemplo 2.2.2 – Utilizando a informação contida no ficheiro *DeputadosXII*, construa uma tabela de frequências para a variável Idade.

Vamos dividir esta tarefa em duas partes: uma primeira parte consistirá na definição das classes e uma segunda parte no cálculo das frequências.

Antes de iniciar o procedimento estatístico, temos de obter a Idade dos deputados, utilizando, por exemplo, o seguinte procedimento:

Copie a coluna "Data de nascimento" para um ficheiro novo com 230 elementos que ocupam as células A2:A231. Para obter a idade em 31/12/2011, podemos utilizar a seguinte metodologia:

- Passo 1 – Inserir na célula B1 a data 31/12/2011;
- Passo 2 – Colocar o cursor na célula B2 e introduzir a expressão: =Int(($B$1-A2)/365,25);
- Passo 3 – Replicar esta função através das células B3 a B231;



	A	B
1	**Data nascimento**	40908
2	23297	=INT(($B$1-A2)/365,25)
3	21684	=INT(($B$1-A3)/365,25)
4	21094	=INT(($B$1-A4)/365,25)

Sheet1 Sheet2

Definição das classes:

a) Determinar a amplitude da amostra, subtraindo o mínimo do máximo;

b) Dividir essa amplitude pelo número K de classes pretendido. 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;

c) Calcular a amplitude de classe h, dividindo a amplitude da amostra por K e tomando para h um valor aproximado por excesso do quociente anteriormente obtido;

d) 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].

Estes passos são representados na figura seguinte:

	B	C	D	E	F	G
1	40908				Classes	
2	=INT(($B$1-A2)/365,25)		Mínimo	=MIN(B2:B231)	Limite inferior	Limite superior
3	=INT(($B$1-A3)/365,25)		Máximo	=MAX(B2:B231)	=E2	=F3+$E$8
4	=INT(($B$1-A4)/365,25)		Amplitude	=E3-E2	=G3	=F4+$E$8
5	=INT(($B$1-A5)/365,25)		n	=COUNT(B2:B231)	=G4	=F5+$E$8
6	=INT(($B$1-A6)/365,25)		k	=INT(LOG(E5;2))+1	=G5	=F6+$E$8
7	=INT(($B$1-A7)/365,25)		Amplitude/k	=E4/E6	=G6	=F7+$E$8
8	=INT(($B$1-A8)/365,25)		h	5,7	=G7	=F8+$E$8
9	=INT(($B$1-A9)/365,25)				=G8	=F9+$E$8

Sheet1 Sheet2 Sheet3

Ready 100%

com os seguintes resultados:

	B	C	D	E	F	G
1	31-12-2011				Classes	
2	48		Mínimo	25	Limite inferior	Limite superior
3	52		Máximo	70	25,0	30,7
4	54		Amplitude	45	30,7	36,4
5	34		n	230	36,4	42,1
6	46		k	8	42,1	47,8
7	47		Amplitude/k	5,625	47,8	53,5
8	67		h	5,7	53,5	59,2
9	64				59,2	64,9
10	66				64,9	70,6

Sheet1 Sheet2 Sheet3

Cálculo das frequências:

Para obter as frequências absolutas das classes anteriormente definidas, vamos utilizar a função *COUNTIF* do seguinte modo:

	H
2	
3	=COUNTIF($B$2:$B$231;"<"&G3)
4	=COUNTIF($B$2:$B$231;"<"&G4)-COUNTIF($B$2:$B$231;"<"&G3)
5	=COUNTIF($B$2:$B$231;"<"&G5)-COUNTIF($B$2:$B$231;"<"&G4)
6	=COUNTIF($B$2:$B$231;"<"&G6)-COUNTIF($B$2:$B$231;"<"&G5)
7	=COUNTIF($B$2:$B$231;"<"&G7)-COUNTIF($B$2:$B$231;"<"&G6)
8	=COUNTIF($B$2:$B$231;"<"&G8)-COUNTIF($B$2:$B$231;"<"&G7)
9	=COUNTIF($B$2:$B$231;"<"&G9)-COUNTIF($B$2:$B$231;"<"&G8)
10	=COUNTIF($B$2:$B$231;"<"&G10)-COUNTIF($B$2:$B$231;"<"&G9)
11	=SUM(H3:H10)

Sheet1 Sheet2 Sheet3

Nota: As classes são fechadas à esquerda e abertas à direita.

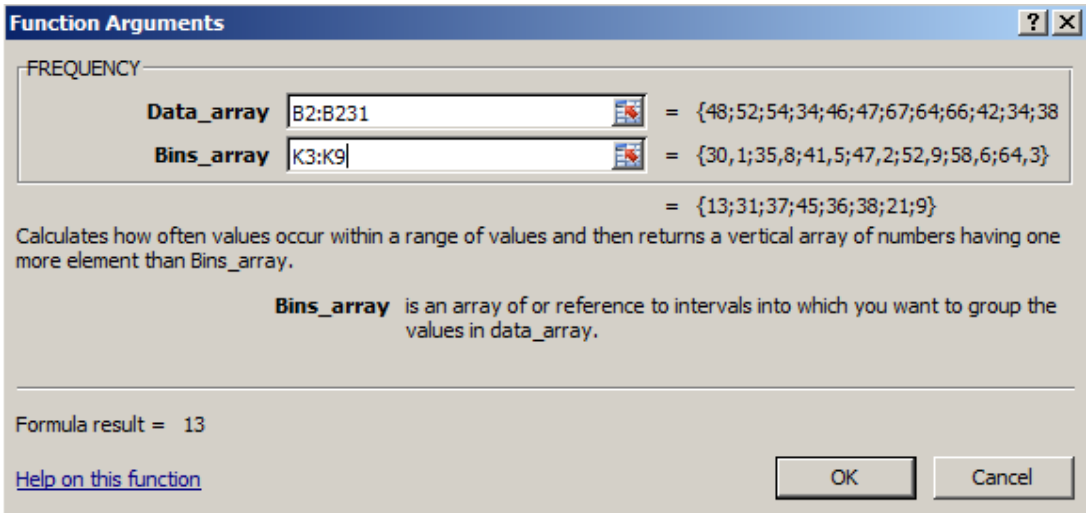
Na tabela anterior inserimos uma coluna com as frequências relativas, tendo-se obtido os seguintes resultados:

	F	G	H	I
1	Classes			
2	Limite inferior	Limite superior	Freq. Absolutas	Freq. Relativas
3	25,0	30,7	13	0,057
4	30,7	36,4	32	0,139
5	36,4	42,1	44	0,191
6	42,1	47,8	37	0,161
7	47,8	53,5	46	0,200
8	53,5	59,2	33	0,143
9	59,2	64,9	16	0,070
10	64,9	70,6	9	0,039
11			230	1,000

Sheet1 Sheet2 Sheet3

2.2.3 - Construção de uma tabela de frequências utilizando a função *Frequency* do Excel

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 no vector **Data_array**, existentes nas classes - cujos limites se encontram em **Bins_array**:



Este vetor **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 k-ésima 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.

Definição das classes:

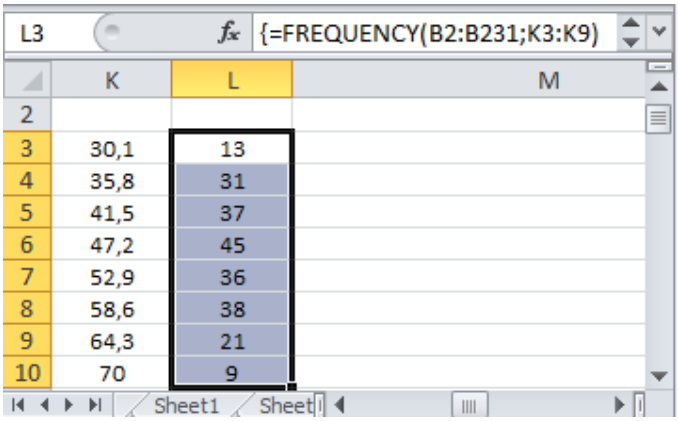
Considerando a amplitude de classe igual a 5,7 (considerada na secção anterior) e tendo em consideração que agora as classes são fechadas à direita, vamos construí-las subtraindo a amplitude de classe ao máximo, como sugerido em 2.2.2, pelo que para a utilização da função *Frequency*, o nosso conjunto de valores para o **Bins_array**, será constituído por {30,1; 35,8; 41,5; 47,2; 52,9; 58,6; 64,3};

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

- Definir a coluna de separadores ou limites das classes, que constituirá o **Bins_array**;
- Selecionar 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 selecionadas 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*

Na figura seguinte apresentamos o resultado deste procedimento:



Verifique que os valores devolvidos pela função *Frequency*, nas células L3: L10, não são iguais às frequências obtidas anteriormente e apresentadas na tabela de frequências já construída, situação previsível devido ao facto de os limites das classes terem sido alterados. Efetivamente, neste caso as classes são]24,4;30,1],]30,1;35,8],]35,8;41,5],]41,5;47,2],]47,2;52,9],]52,9, 58,6],]58,6;64,3],]64,3;70].

Nota – Apesar de na coluna que contém os separadores das classes aparecer o 70, este não é tomado em consideração. Foi utilizado para construir os separadores das células k3:k9.

2.3 – Utilização do Excel na representação gráfica de dados

De forma idêntica à que fizemos para a construção das tabelas de frequências, vamos também considerar separadamente o caso da variável em estudo ser de natureza qualitativa ou quantitativa discreta, ou de natureza quantitativa contínua.

2.3.1 – Variáveis qualitativas ou quantitativas discretas. Diagrama ou gráfico de barras

Neste caso vimos que a construção da tabela de frequências se resume, de um modo geral, a considerar como classes as diferentes categorias ou valores que surgem na amostra. Uma representação gráfica adequada para estes dados, é o diagrama de barras, que já foi introduzido no módulo de Estatística.

Diagrama (ou gráfico) de barras – Representação gráfica que consiste em marcar num sistema de eixos coordenados, no eixo dos xx, pontos representando as categorias ou os valores considerados para as classes na tabela de frequências, e nesses pontos barras verticais de altura igual ou proporcional à frequência absoluta ou à frequência relativa. No

caso dos dados qualitativos, a ordem por que se colocam as barras é qualquer, a não ser que exista alguma ordem subjacente, como nos qualitativos ordinais. Neste caso, respeita-se a ordem, colocando da esquerda para a direita as diferentes categorias, começando na de menor nível para a de maior nível.

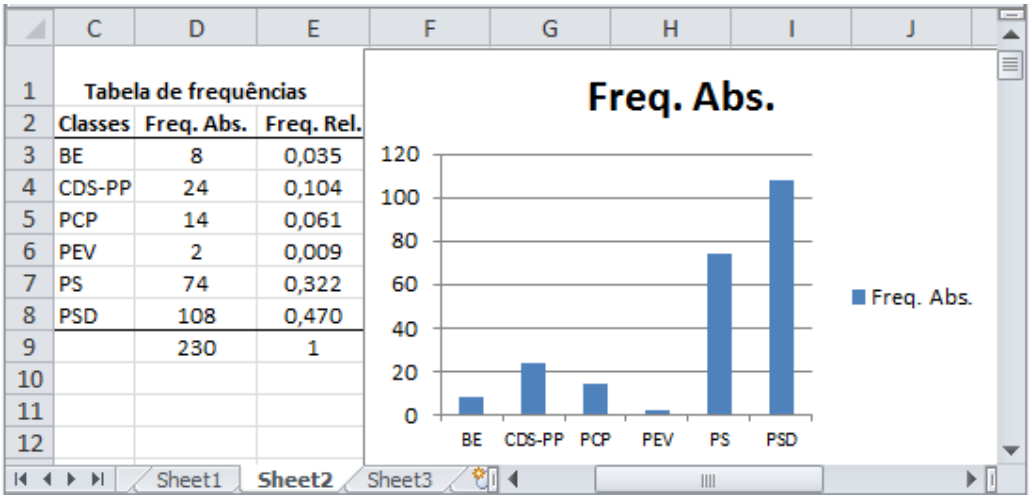
2.3.1.1 - Variável de tipo qualitativo

Exemplo 2.3.1 - Vamos exemplificar a construção de um diagrama de barras de uma variável qualitativa, considerando a tabela de frequências construída em 2.2.1, para estudar a variável Grupo Parlamentar, do ficheiro *DeputadosXII*:

	C	D	E
1	Tabela de frequências		
2	Classes	Freq. Abs.	Freq. Rel.
3	BE	8	0,035
4	CDS-PP	24	0,104
5	PCP	14	0,061
6	PEV	2	0,009
7	PS	74	0,322
8	PSD	108	0,470
9		230	1

A metodologia seguida para construir o diagrama de barras, consiste em, na folha Excel, que contém a tabela:

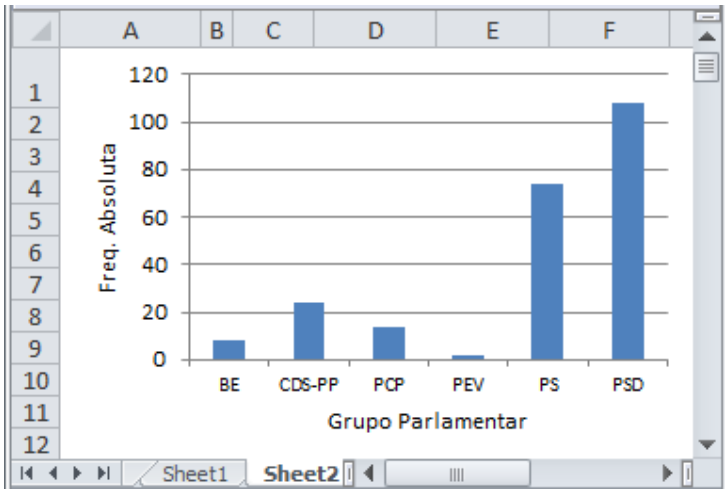
- Selecionar as células que contêm as classes e as frequências absolutas (por exemplo), incluindo (ou não) os títulos;
- Na barra de ferramentas, seleccionar **Insert** e o ícone Chart;
- Selecionar a opção *Column* para o gráfico a 2 dimensões, obtendo-se o gráfico seguinte:



Como não pretendemos este título nem esta legenda, selecionamos as caixas respetivas e fazemos *Delete*. Uma forma de evitar que apareça o título e a legenda, é selecionar unicamente as células C3:D8 ou então deixar em branco as células anteriores aos valores a serem representados, selecionando também essas células (no caso do exemplo teríamos de apagar os conteúdos das células C2 e D2 e selecionar C2:D8.

Nota – Esta metodologia de colocar as células em branco é fundamental no caso de os dados serem quantitativos discretos, considerando como classes os diferentes dados. Consideramo-la aqui também como uma questão de uniformização na construção do gráfico de barras.

Para acrescentar os títulos nos eixos, selecionamos *Layout→Axis Titles* e escrevemos os títulos pretendidos. Finalmente temos o gráfico de barras com o seguinte aspeto:



2.3.1.2 - Variável de tipo quantitativo discreto

2.3.1.2.1 – Diagrama de barras

No caso de dados discretos, para construir a tabela de frequência consideram-se como classes os diferentes valores que surgem na amostra. Estes valores devem ser apresentados, na tabela de frequência, ordenados.

Exemplo 2.3.2 – Suponhamos que para uma amostra de 30 deputados da atual legislatura, se tinha recolhido a informação sobre o número de filhos, tendo-se obtido os seguintes valores:

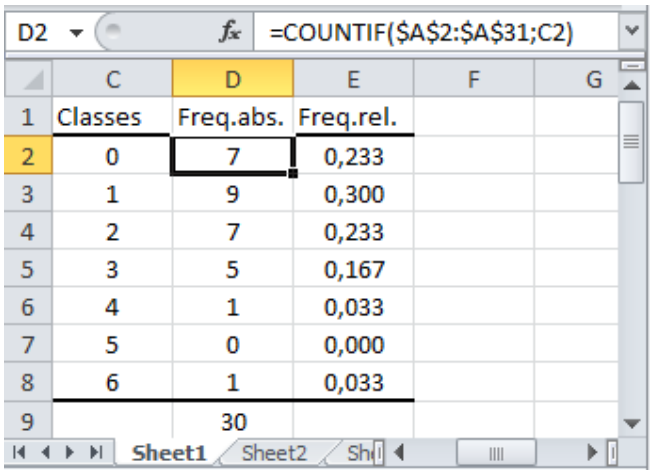
2, 1, 2, 3, 0, 0, 1, 1, 4, 1, 2, 1, 0, 0, 0, 2, 3, 1, 1, 6, 3, 1, 3, 2, 0, 1, 2, 0, 2, 3

Resuma os dados numa tabela de frequências e construa o diagrama de barras associado.

Introduzimos os dados nas células A2:A31 de uma folha de Excel, a que chamámos *Filhos* e a seguir procedemos do seguinte modo:

1ª parte – Procedimento para a construção da tabela de frequências:

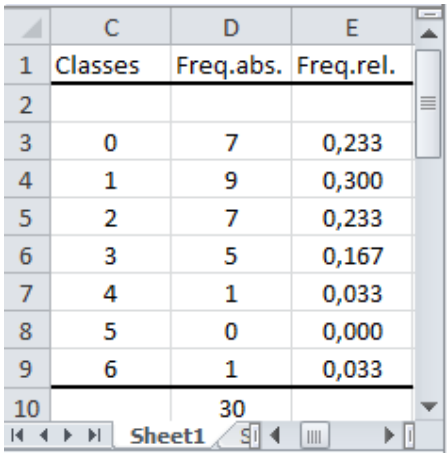
- Selecionar as células que contêm os dados e ordená-los utilizando o botão  da barra de Excel;
- Considerar para classes os diferentes valores que surgem na amostra. Se faltar algum valor entre o máximo e o mínimo, considerá-lo também na tabela de frequências, se a seguir se pretende construir um diagrama de barras; neste exemplo considerámos o valor 5 como classe, com frequência nula;
- Utilizando a função *COUNTIF*, determinar as frequências absolutas das classes consideradas no ponto anterior; calcular a partir destas, as frequências relativas.



	C	D	E	F	G
1	Classes	Freq.abs.	Freq.rel.		
2	0	7	0,233		
3	1	9	0,300		
4	2	7	0,233		
5	3	5	0,167		
6	4	1	0,033		
7	5	0	0,000		
8	6	1	0,033		
9		30			

2ª parte – Procedimento para a construção do diagrama de barras:

- De forma análoga ao que se sugeriu para a representação num diagrama de barras dos dados qualitativos, insira uma célula em branco entre os valores da tabela que se pretendem representar e os títulos das colunas. Assim, a tabela a ser representada é a seguinte:

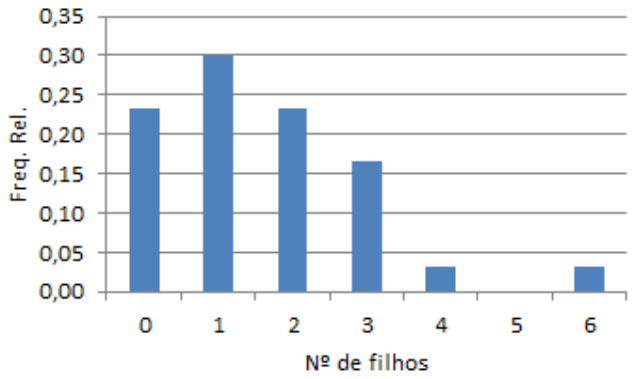


	C	D	E
1	Classes	Freq.abs.	Freq.rel.
2			
3	0	7	0,233
4	1	9	0,300
5	2	7	0,233
6	3	5	0,167
7	4	1	0,033
8	5	0	0,000
9	6	1	0,033
10		30	

- Selecionar as células que contêm as classes e as que contêm as frequências relativas, incluindo as células em branco, ou seja C2:C9 e E2:E9 (para selecionar as células que contêm as frequências relativas, como não são adjacentes às que contêm as classes, depois de selecionar estas, tem que se pressionar a tecla CTRL e com ela pressionada, selecionar as células da coluna E);

- Selecionar na barra de ferramentas a opção *Insert* e aí *Chart* e a seguir a opção *Column*, tal como se fez para os dados de tipo qualitativo;

- Acrescentar os títulos aos eixos:



Nota – Experimente construir o diagrama de barras sem selecionar as células em branco.

2.3.1.2.2 – Função cumulativa

A função cumulativa é uma função definida para todo o valor real x, e que para cada x dá a soma das frequências relativas dos valores da amostra menores ou iguais a x.

Quando temos uma variável de tipo discreto, a função cumulativa é uma função em escada, isto é, é uma função que cresce por degraus, mudando de degrau nos pontos em que a frequência é diferente de 0, e em que a altura do degrau é igual à frequência respetiva. Vamos exemplificar a sua construção com o exemplo apresentado na secção anterior para a construção do diagrama de barras.

Exemplo 2.3.2 (cont) – Construa a função cumulativa para os dados do número de filhos da amostra dos 30 deputados.

Retomando a tabela de frequências do exemplo 2.3.2, vamos acrescentar uma coluna com as frequências relativas acumuladas:

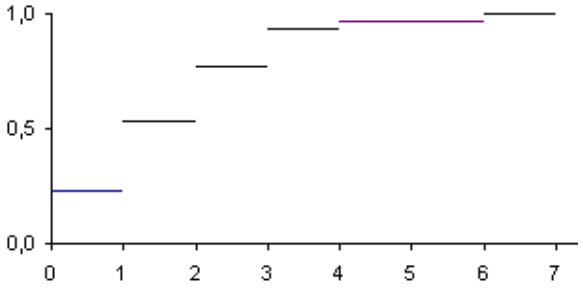
Tabela de frequências

Classes	Freq.abs.	Freq.rel.	Freq.rel.acum.
0	7	0,233	0,233
1	9	0,300	0,533
2	7	0,233	0,767
3	5	0,167	0,933
4	1	0,033	0,967
5	0	0,000	0,967
6	1	0,033	1,000
	30		

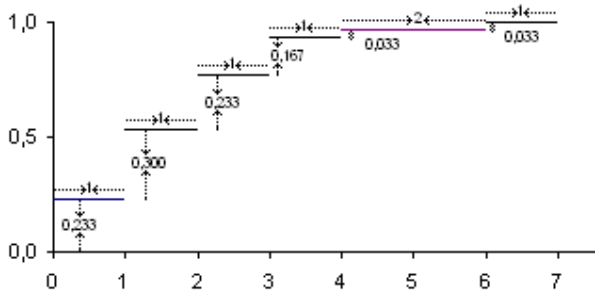
A função cumulativa há-de ser tal que:

- Para valores de x< o, será nula;
- Para valores de o ≤ x < 1, será igual a 0,233;
- Para valores de 1 ≤ x < 2, será igual a 0,533;
- Para valores de 2 ≤ x < 3, será igual a 0,767;
- Para valores de 3 ≤ x < 4, será igual a 0,933;
- Para valores de 4 ≤ x < 6, será igual a 0,967;
- Para valores de x ≥ 6, será igual a 1;

O Excel não dispõe de uma representação imediata para a função anterior, pelo que temos de utilizar um pequeno artifício. Suponhamos, para já, que por algum processo tínhamos conseguido construir o gráfico da função cumulativa, que tem o seguinte aspeto:



Esta função é constituída por 6 degraus, em que a altura do degrau é, em cada ponto, igual à frequência relativa respetiva e a dimensão do patamar é igual à diferença entre os pontos consecutivos, com frequência relativa diferente de zero:



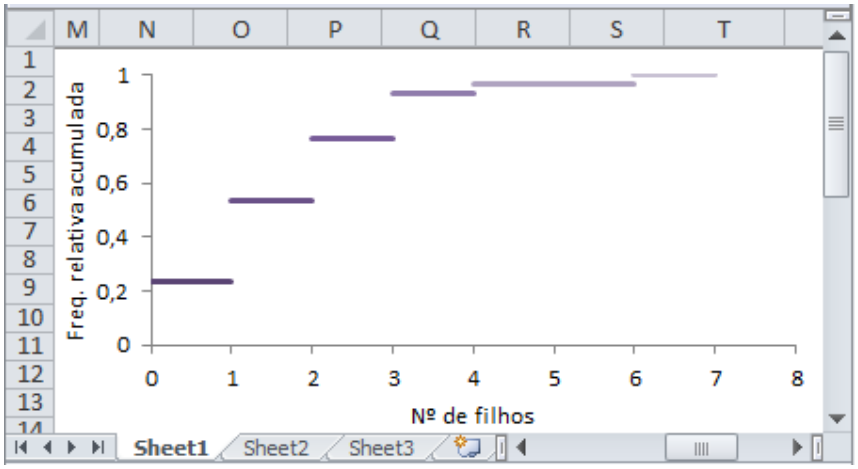
O Excel dispõe de uma representação gráfica, o *Scatter* (Diagrama de dispersão), em que no último subtipo apresentado para as opções, une os pontos, por ordem crescente das abcissas, simultaneamente de tantas séries (conjuntos de pontos) quantas as desejadas.

Vamos utilizar esta função *Scatter* para construir os sucessivos degraus da função cumulativa, em que cada degrau corresponde a uma série - união de dois pontos, e em que temos tantas séries a representar, quantos os degraus. Assim, o artifício está em representar, numa tabela do Excel, os degraus pretendidos através das coordenadas dos pontos, como exemplificamos a seguir:

	M	N	O	P	Q	R	S
1		1º degrau	2º degrau	3º degrau	4º degrau	5º degrau	6º degrau
2							
3	0	0,233					
4	1	0,233					
5	1		0,533				
6	2		0,533				
7	2			0,767			
8	3			0,767			
9	3				0,933		
10	4				0,933		
11	4					0,967	
12	6					0,967	
13	6						1
14	7						1

Sheet1 Sheet2 Sheet

Agora basta selecionar as células M2: S143 e fazer o diagrama de dispersão. Proceda como na construção do diagrama de barras, para retirar a legenda e acrescentar títulos:


M N O P Q R S T
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Freq. relativa acumulada
1
0,8
0,6
0,4
0,2
0
0 1 2 3 4 5 6 7 8
Nº de filhos
Sheet1 Sheet2 Sheet3

2.3.2 – Variáveis quantitativas contínuas

2.3.2.1 – Histograma

2.3.2.1.1 – Tabela de frequências com as classes com a mesma amplitude

No caso de um conjunto de dados contínuos, já vimos anteriormente a forma de obter a tabela de frequências. Como se viu, as classes são intervalos e a representação gráfica adequada é o *histograma*, já apresentado no módulo de Estatística:

Histograma - é um diagrama de áreas, formado por uma sucessão de retângulos adjacentes, tendo cada um por base um intervalo de classe e por área a frequência relativa (ou frequência absoluta). Por conseguinte, a área total coberta pelo histograma é igual a 1 (ou igual a n, a dimensão do conjunto de dados a representar).

Para construir o histograma de forma correta, isto é, de modo a que as áreas dos retângulos sejam iguais às frequências, a altura do retângulo correspondente a determinada classe, deverá ser igual à frequência da classe a dividir pela respetiva amplitude. Contudo, se as classes tiverem todas a mesma amplitude, é usual construir os retângulos com alturas iguais às frequências relativas (absolutas) das respetivas classes, vindo as áreas dos retângulos proporcionais e não iguais às frequências. A constante de proporcionalidade é a amplitude de classe. No entanto, se se pretender comparar amostras através de histogramas, embora o histograma não seja a representação mais adequada para a comparação de amostras, deve-se ter o cuidado de os construir da forma indicada inicialmente, e utilizando as frequências relativas, de modo que a área total ocupada por cada um dos histogramas seja igual a 1.

Exemplificamos, de seguida, a construção de um histograma utilizando o Excel.

Exemplo 2.3.3 – Considerando a tabela de frequências construída em 2.3 para a variável idade, construa o histograma adequado.

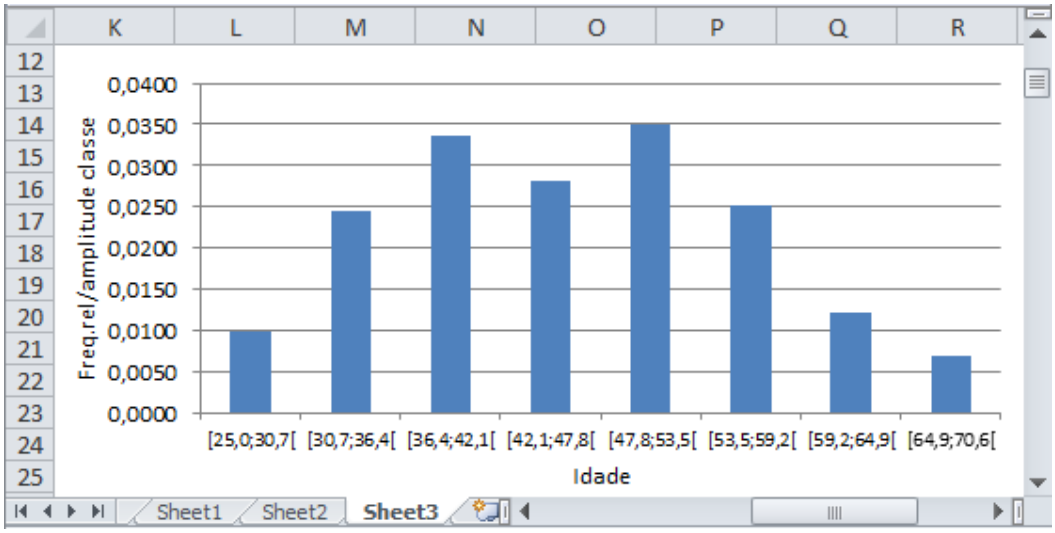
Processo utilizado para obter o histograma:

- Acrescentar, à tabela considerada, uma outra coluna com a frequência relativa a dividir pela amplitude de classe (igual a 5,7). No caso presente, inserimos estas células adjacentes às células que contêm as classes. No entanto, não é necessário ter esta preocupação, já que se pretender selecionar células não adjacentes, basta selecionar as células da primeira coluna e se a coluna seguinte não for adjacente, começar por carregar a tecla *CTRL* e com ela pressionada selecionar, então, as células pretendidas;

	F	G	H	I
13	Classes	Freq. Rel/5,7	Freq. Abs.	Freq. Rel.
14				
15	[25,0;30,7[	0,0099	13	0,057
16	[30,7;36,4[	0,0244	32	0,139
17	[36,4;42,1[	0,0336	44	0,191
18	[42,1;47,8[	0,0282	37	0,161
19	[47,8;53,5[	0,0351	46	0,200
20	[53,5;59,2[	0,0252	33	0,143
21	[59,2;64,9[	0,0122	16	0,070
22	[64,9;70,6[	0,0069	9	0,039

Sheet1 SH

- Selecionar as células F15:G22 (que contêm as classes e as frequências relativas a dividir pela amplitude de classe. Neste caso, como as classes são categorias e não números, não precisamos de selecionar as células em branco;
- Proceder como em 3.1 para construir um diagrama de barras, para obter a figura que se apresenta a seguir;

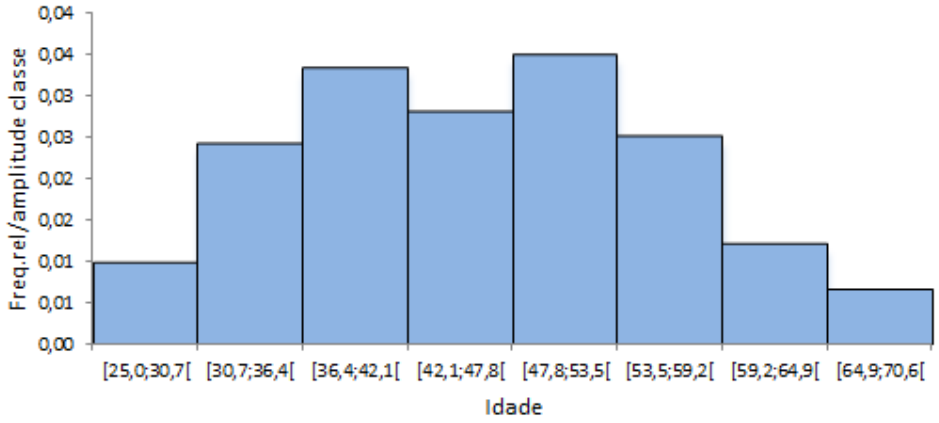


Para obter o histograma, já que o que se nos apresenta na figura anterior não é um histograma pois não tem as barras adjacentes, terá de:

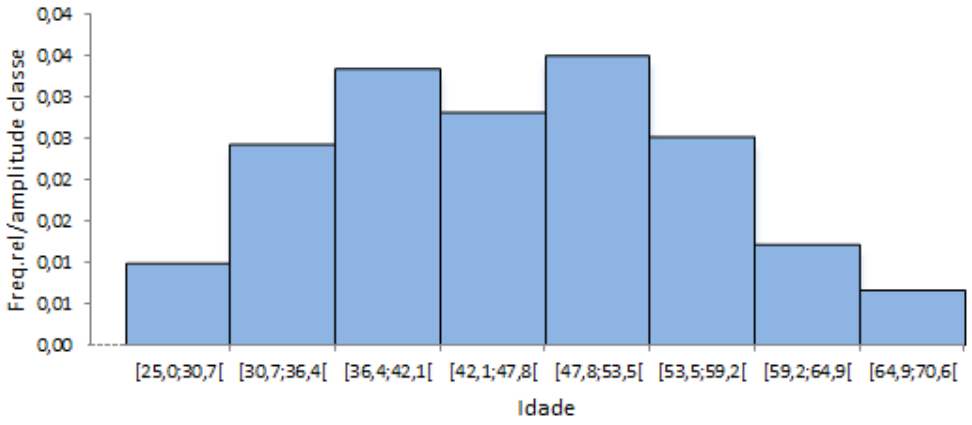
Clicar com o botão direito do rato sobre as barras, de forma a que apareça o menu *Format Data Series* ou *Format data Points*. Aí selecionar *0* em *Gap Whist* e *OK*:

Finalmente pode-se melhorar esteticamente o histograma, diminuindo o número de casas decimais nos valores apresentados no eixo dos YY, retirando as linhas, etc.

Histograma para a idade dos deputados



Nota – Tomar em atenção que as classes aparecem encostadas ao eixo das ordenadas, embora devesse aparecer entre esse eixo e a primeira classe, uma quebra de linha, como se apresenta a seguir (conseguido com o Paint...).

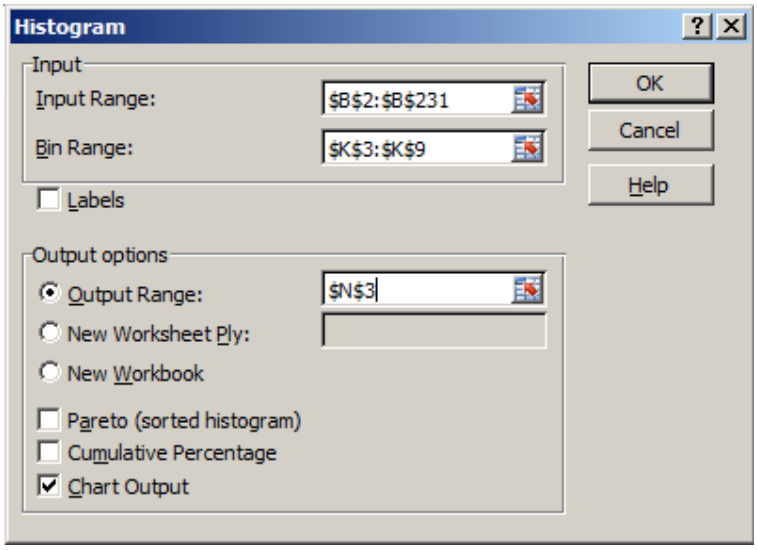


Nota – Da forma como foi construído, a área total ocupada pelo histograma é igual a 1**.**

2.3.2.1.2 – Função *Histogram*

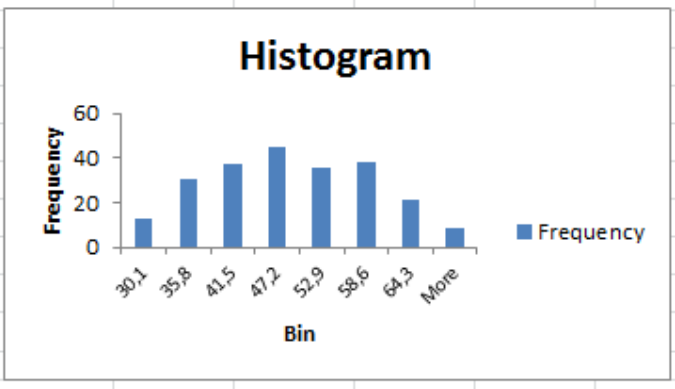
No Excel existe uma função, idêntica à função *Frequency*, a função *Histogram*, a que se acede selecionando *Data→Data Analysis*[3]*→Histogram→Ok*. Vamos exemplificar a sua utilização para o conjunto de dados da variável Idade, anteriormente considerado:

- Definir a coluna de separadores ou limites de classes, que constituirá o *Bin Range*: No nosso caso construimos as classes subtraindo a amplitude de classe sucessivamente ao máximo, obtendo os valores {30,1, 35,8, 41,5, 47,2, 52,9, 58,6, 64,3} (tal como para a função Frequency, as classes são fechadas à direita e abertas à esquerda), que colocámos nas células K3:K9;
- Selecionar *Data→Data Analysis→Histogram→Ok*:



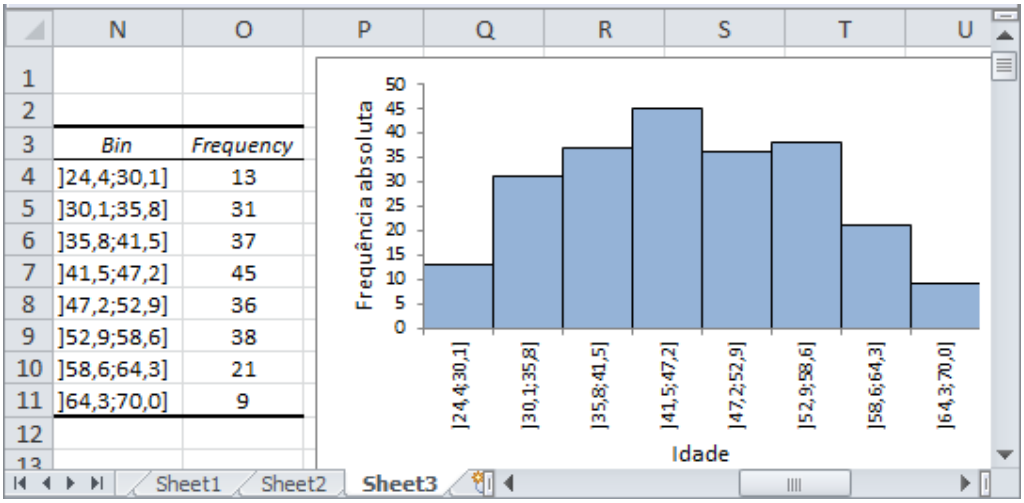
- Em *Input Range*, indicámos o local dos dados e selecionámos ainda a opção Chart Output e clicámos OK. Como resultado obtivemos o seguinte:

Bin	*Frequency*
30,1	13
35,8	31
41,5	37
47,2	45
52,9	36
58,6	38
64,3	21
More	9



- Substituímos 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:

[3] No caso de não estar visível a opção Data Analysis, terá de utilizar a seguinte metodogia para a instalar: Selecionar sucessivamente *File*→Options→ Add-Ins→ Analysis ToolPak→ Go → V



Nota1 - Ao considerar a função *Histogram*, tem a possibilidade de não indicar os separadores de classe, deixando vazio o espaço denominado *Bin Range*, uma vez que serão considerados, por defeito, classes. Contudo, não aconselhamos que se deixe esta escolha ao Excel, uma vez que, por exemplo, a primeira classe que é considerada, é constituída pelos valores menores ou iguais ao mínimo, o que não tem qualquer sentido.

Nota 2 – A área ocupada pelo histograma é igual à amplitude de classe $\times \sum$ frequências absolutas, ou seja, no nosso caso vem 5,7 x 230.

2.3.3.1.3 - Tabela de frequências com as classes com amplitudes diferentes

Por vezes a organização e redução de um conjunto de dados contínuos, através de uma tabela de frequências, pressupõe que os intervalos, que constituem as classes, tenham limites escolhidos pelo utilizador, sem obedecerem a um critério estritamente resultante da aplicação de uma regra matemática. É o caso, por exemplo, da variável idade, em que poderá ser interessante escolher determinadas classes etárias.

Tendo em conta a definição de histograma, como sendo um diagrama de áreas, constituído por uma série de retângulos adjacentes, em que a área de cada retângulo é igual ou proporcional à frequência de classe, no caso de a tabela de frequências não apresentar as classes todas com a mesma amplitude, já o histograma não se pode reduzir a um diagrama de barras, em que as barras tenham a mesma largura e as alturas sejam iguais às frequências.

Não sendo o Excel um software de Estatística, não apresenta uma solução imediata para a construção do histograma nestas condições, sendo necessário recorrer a um artifício. Exemplificaremos a seguir a aplicação de uma técnica possível para a resolução do problema, recorrendo à representação gráfica *Scatter*.

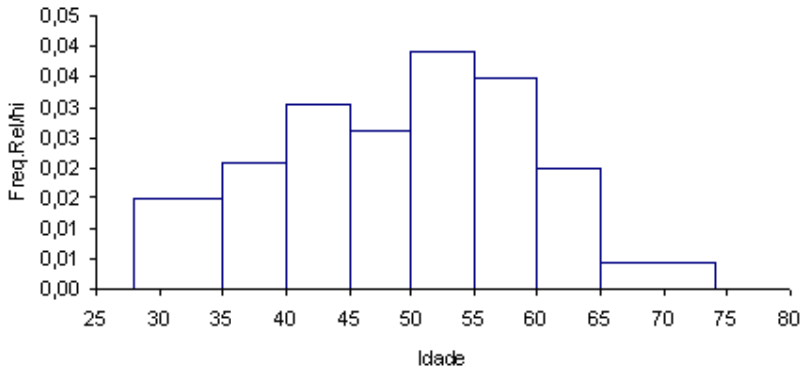
Exemplo 2.3.4 – Consideremos um conjunto de dados organizados numa tabela de frequências, em que se consideraram as seguintes classes [28, 35[, [35, 40[, [40, 45[, [45, 50[, [50, 55[, [55, 60[, [60, 65[, [65, 74[(Este conjunto de dados é também um conjunto de 230 idades, mas não é o conjunto do ficheiro *DeputadosXII*). A esta tabela de frequências acrescentámos uma nova coluna onde, para cada classe, se considera a frequência relativa

(ou absoluta) a dividir pela amplitude de classe. Será esta coluna que irá fornecer as alturas dos retângulos que constituirão o histograma. Com esta precaução, garantimos que as áreas destes retângulos são iguais às frequências relativas (ou absolutas). Apresenta-se a seguir a tabela de frequências obtida, segundo a descrição anterior:

	K	L	M	N	O
2	**Tabela de frequências**				
3	Classes	Amplitude hi	Freq.Rel/hi	Freq. Abs.	Freq. Rel.
4	[28; 35[	7	0,0149	24	0,104
5	[35; 40[	5	0,0209	24	0,104
6	[40; 45[	5	0,0304	35	0,152
7	[45; 50[	5	0,0261	30	0,130
8	[50; 55[	5	0,0391	45	0,196
9	[55; 60[	5	0,0348	40	0,174
10	[60; 65[	5	0,0200	23	0,100
11	[65; 74[	9	0,0043	9	0,039
12				230	1,000

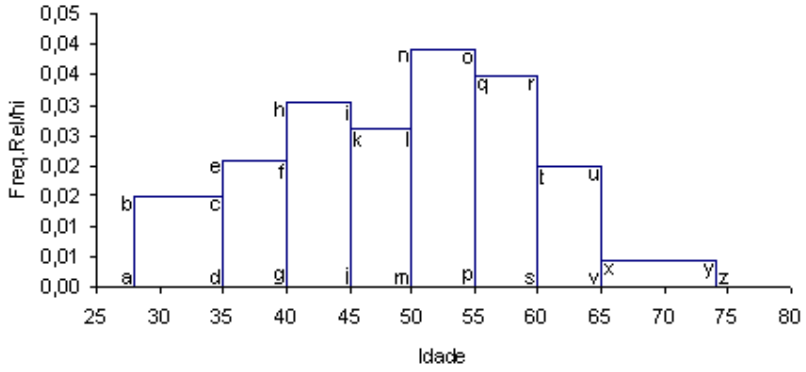
Sheet1 Sheet2

O histograma correspondente a esta tabela de frequências, com cuja construção não nos vamos preocupar para já, terá o seguinte aspeto:

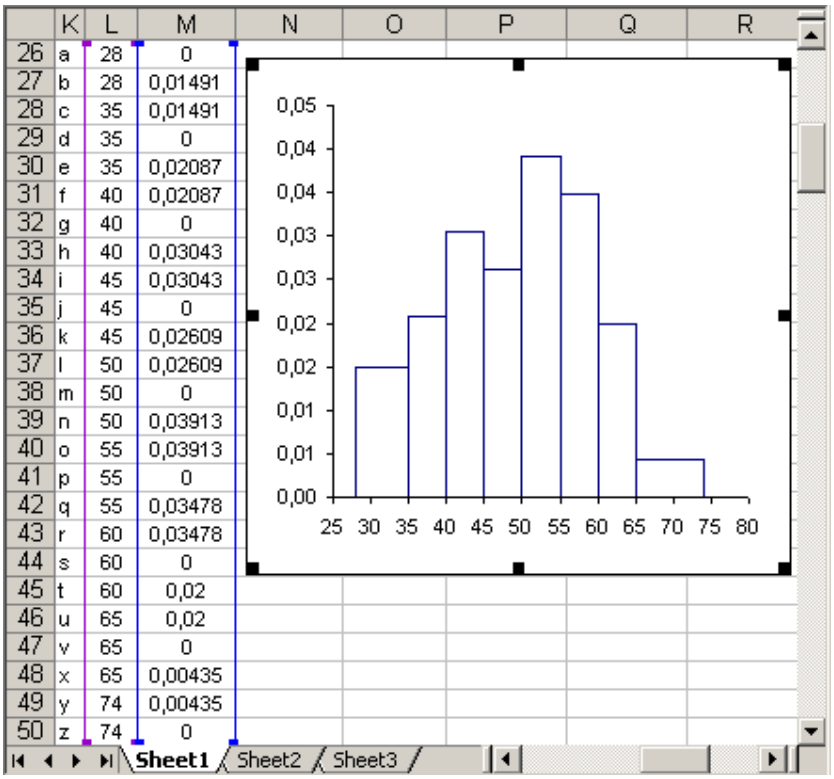


Temos um histograma corretamente construído, em que as áreas dos retângulos são iguais às frequências relativas, ocupando o histograma uma área total igual a 1.

Na figura anterior, vamos marcar alguns pontos com letras:



Repare que se unir o ponto **a** com **b**, de seguida com **c**, até esgotar todos os pontos, obtém o histograma. Então, para obter a representação gráfica desejada, basta construir uma tabela, numa folha de Excel, com as coordenadas dos pontos que pretendemos unir e utilizar a representação *Scatter*, tal como foi feito para representar a função cumulativa em 3.1.2.2:



	K	L	M
26	a	28	0
27	b	28	0,01491
28	c	35	0,01491
29	d	35	0
30	e	35	0,02087
31	f	40	0,02087
32	g	40	0
33	h	40	0,03043
34	i	45	0,03043
35	j	45	0
36	k	45	0,02609
37	l	50	0,02609
38	m	50	0
39	n	50	0,03913
40	o	55	0,03913
41	p	55	0
42	q	55	0,03478
43	r	60	0,03478
44	s	60	0
45	t	60	0,02
46	u	65	0,02
47	v	65	0
48	x	65	0,00435
49	y	74	0,00435
50	z	74	0

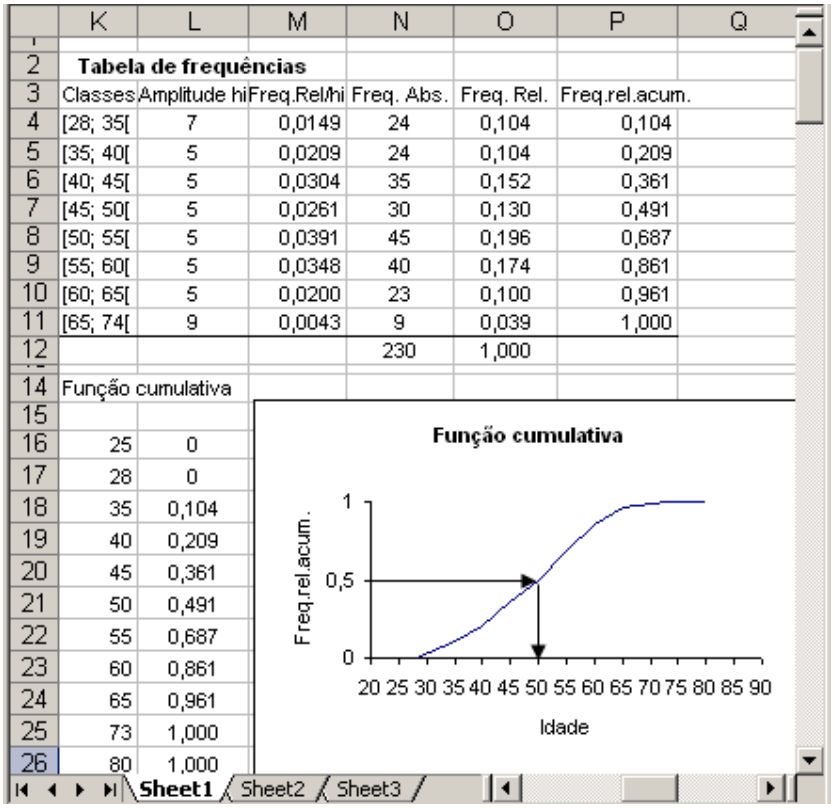
2.3.2.2 – Função cumulativa

Para representar graficamente as frequências acumuladas, considera-se a função cumulativa, que se obtém utilizando a seguinte metodologia:

- Antes do limite inferior da 1ª classe, l_1, a frequência acumulada é nula, pelo que se traça um segmento sobre o eixo dos xx, até esse ponto;
- No limite inferior da 2ª classe, l_2, a frequência acumulada é a frequência da classe anterior, f_1. Admitindo que a frequência se distribui uniformemente no intervalo de classe, unimos os pontos de coordenadas $(l_1,0)$ e (l_2, f_1);

- No limite inferior da 3ª classe, l_3, a frequência acumulada é a soma das frequências das duas classes anteriores, (f_1+f_2). Então unimos os pontos de coordenadas (l_2, f_1) e $(l_3, (f_1+f_2))$;
- Quando chegarmos à última classe, temos a garantia que a frequência acumulada, correspondente ao seu limite superior, é igual a 1, pelo que nesse ponto marcamos 1 e continuamos com um segmento de reta paralelo ao eixo dos xx.

Exemplo 2.3.4 (continuação) – Construa a função cumulativa, a partir da tabela de frequências apresentada no exemplo 2.3.4. Para obter a função cumulativa, basta acrescentar à tabela de frequências uma nova coluna com as frequências relativas acumuladas. De seguida utiliza-se a representação *Scatter*, para unir os pontos, tais como foram definidos nas indicações dadas, anteriormente, para a construção da função cumulativa:



Da maneira como foi construída, a função cumulativa tem algumas propriedades importantes, nomeadamente:

- Está definida para todo o x real (na representação gráfica anterior escolhemos arbitrariamente o valor da abcissa igual a 25 para começar a construir a função cumulativa);
- É sempre não decrescente;
- Só assume valores no intervalo [0, 1];
- Permite obter informação sobre qual o valor da abcissa a que corresponde determinada frequência acumulada.

Vamos explorar um pouco mais esta última propriedade.

Suponhamos que se pretendia saber, a partir da representação gráfica da função cumulativa, obtida para o exemplo anterior, qual o valor aproximado para a idade a que corresponde uma frequência relativa acumulada de 50%. De acordo com a figura, este valor deve estar na classe [50, 55[.

Uma vez que se admite que a frequência se distribui uniformemente sobre a amplitude de classe, isto é a frequência 0,196 (=0,687-0,491) distribui-se uniformemente sobre o intervalo de amplitude 5, através da resolução de uma equação de proporcionalidade, obtém-se o valor que andávamos à procura:

$$\frac{0{,}196}{0{,}009} = \frac{5}{x} \qquad x = \frac{0{,}009 \times 5}{0{,}196} = 0{,}22$$

onde 0,009=0,5-0,491. Então o valor pretendido é 50 + 0,22 = 50,22 anos, ou seja 50 anos.

Ao valor obtido anteriormente, a que corresponde uma frequência acumulada de 50%, chamamos *mediana*. A mediana, que já foi objeto de estudo no módulo de Estatística, divide a distribuição das frequências em duas partes iguais. Recordamos que a técnica utilizada permitiu-nos obter um valor aproximado para a mediana, cujo valor exato só poderia ter sido determinado a partir dos dados originais, antes de proceder ao agrupamento. Aliás, veremos mais à frente a determinação desta e de outras medidas, utilizando o Excel.

Se em vez de pretendermos determinar o valor a que corresponde a percentagem de 50%, procurássemos os valores a que correspondem as percentagens de 25% ou 75%, obteríamos os chamados quartis, respetivamente 1º e 3º quartil, e a metodologia utilizada para os determinar a partir da função cumulativa seria idêntica à utilizada para determinar a mediana.

2.3.3 – Outras representações gráficas

Além das representações gráficas consideradas anteriormente, em que destacamos o diagrama de barras para dados discretos e o histograma para dados contínuos, existem ainda outras representações que podem ser utilizadas para dados qualitativos ou quantitativos – diagrama circular, ou dados quantitativos – caule-e-folhas e diagrama de extremos e quartis. Todas estas representações já foram objeto de estudo no módulo de Estatística, pelo que privilegiaremos aqui a forma de os construir utilizando o Excel.

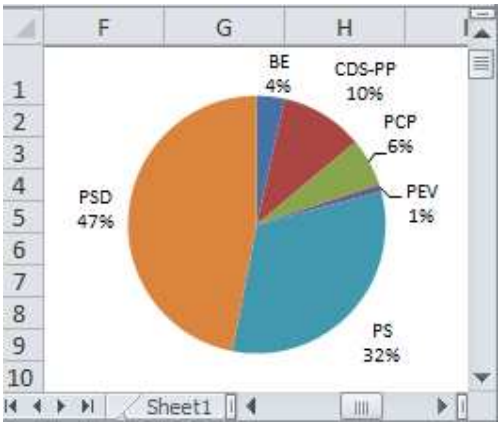
2.3.3.1 – Diagrama circular

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. A representação deste diagrama, em Excel, é imediata, apresentando várias modalidades.

Exemplo 2.3.5 – Apresente sob a forma de um diagrama circular a distribuição dos deputados do ficheiro *DeputadosXII* segundo o grupo parlamentar.

Esta variável já foi objeto de estudo num exemplo anterior, de forma que recorremos à tabela de frequências já calculada, para obter a representação gráfica pretendida. Selecionam-se as células com as classes e as respetivas frequências absolutas ou relativas (neste caso selecionamos C3:C8 e E3:E8, e na barra de ferramentas a opção **Insert,** seguido de *Chart*  e a seguir a primeira opção a duas dimensões do gráfico *Pie*. A representação obtida é a seguinte:

Para tornar o diagrama circular mais informativo, apagamos a legenda e com ele selecionado vamos a *Layout→Data Labels→More Data Labels Options* onde escolhemos a modalidade preferida, que neste caso conduziu ao gráfico com o seguinte aspeto



2.3.3.2 – Caule-e-folha

Esta representação, como se sabe, é uma representação que se pode considerar entre a tabela e o gráfico, uma vez que são apresentados os verdadeiros valores da amostra, mas de forma sugestiva, que faz lembrar um histograma. Antes de abordarmos a forma de construir um caule-e-folhas utilizando o Excel, vamos apresentar um exemplo, que nos poderá ajudar a compreender os passos necessários para essa construção.

Exemplo 2.3.6 – Consideremos a seguinte amostra constituída pela idade de 30 deputados, escolhidos aleatoriamente da tabela de deputados do ficheiro *DeputadosXII*:

63	59	31	51	51	61	42	65	48	63	57	43	54	42
52													
51	57	34	38	44	61	60	56	66	63	52	47	33	46
52													

Uma representação possível em caule-e-folhas é a que se apresenta a seguir:

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

Nesta representação considerámos 4 caules e o intervalo entre caules sucessivos é de 10 unidades. No caule 3 pendurámos todas as folhas deste caule e o mesmo foi feito com todos os outros caules. É como se tivéssemos considerado as classes [30, 40[, [40, 50[, [50, 60[e [60, 70[para agrupar os dados. Suponhamos que em vez de considerar estas classes, de amplitude 10, estávamos interessados em considerar classes de amplitude 5, a saber [30, 35[, [35, 40[, [40, 45[, [45, 50[, [50, 55[, [55, 60[, [60, 65[e [65, 70[. Então a representação anterior teria o seguinte aspeto:

3*	1 3 4
3.	8
4*	2 2 3 4
4.	6 7 8
5*	1 1 1 2 2 2 4
5.	6 7 7 9
6*	0 1 1 3 3 3

6. | 5 6

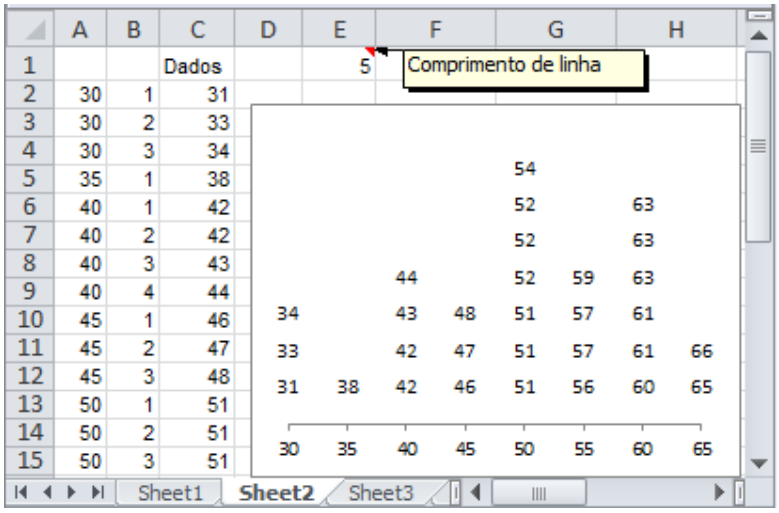
Qualquer que seja a representação considerada, qualquer caule tem sempre a possibilidade de ter penduradas o mesmo número de folhas. No exemplo anterior, no primeiro sub caule 3 (ou 4, ou 5, ou 6) aparecem penduradas as folhas 0, 1, 2, 3 e 4, enquanto que no segundo sub caule 3 (ou 4, ou 5, ou 6) aparecem penduradas as folhas 5, 6, 7, 8 e 9). Uma outra possibilidade seria considerar classes de amplitude 2, fazendo cada caule dividido em 5 sub caules e cabendo a cada sub caule 2 folhas (repare-se com a analogia com a construção do histograma, em que considerámos as classes com igual amplitude). A esta amplitude de classe é usual chamar *comprimento de linha*.

Não existe no Excel uma representação imediata para a construção de um caule-e-folhas, podendo-se desenvolver vários processos que conduzem a uma representação em caule-e-folhas ou idêntica. Transcrevemos a seguir, sem atualizar para o Excel 2010, um processo que inserimos na versão original deste dossiê. A razão que nos leva à não atualização deste método para obter uma representação semelhante a um caule-e-folhas, prende-se com o facto de acharmos mais simples e eficiente o processo que descrevemos em segundo lugar, utilizando as funções REPT e COUNTIF do Excel.

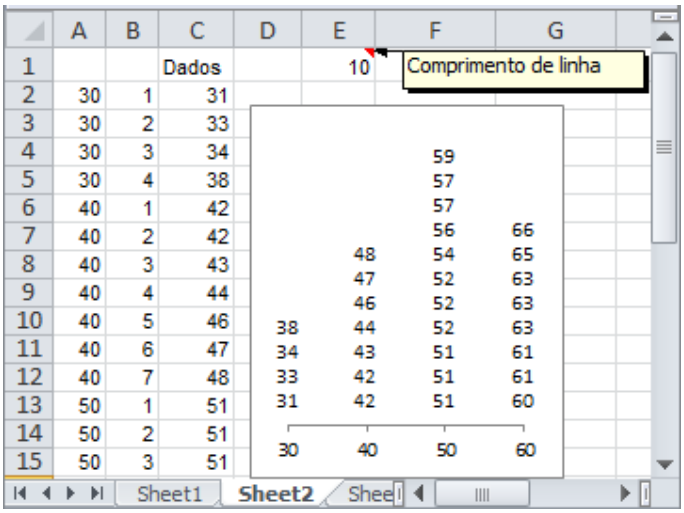
1. Processo desenvolvido por Neville Hunt (Hunt, 2001), para o Excel (versão não atualizada para Excel 2010):

- 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: 10, 5 ou 2 ou uma potência de 10, destes valores;
- 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 – Selecione 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 selecione *Format data Series*) selecione:
  Scale bubble size: 10 e No Fill
  e em Data Labels: Show bubbles sizes e *Center*
  OK;
- 7º passo – Clique numa das linhas horizontais que atravessam o gráfico e apague-as com a tecla Delete. Apague também a legenda.
- 8º passo – Formate convenientemente os eixos.

Vamos exemplificar os passos anteriores com os 30 dados referentes às idades dos deputados.



Na folha de Excel, se mudarmos o valor do comprimento de linha para 10, aparece de imediato a seguinte representação (aparte uma formatação adequada do eixo dos xx):



Repare-se que, embora as notações usadas para os caules e as folhas não sejam idênticos aos da representação inicialmente considerada, feita sem o recurso ao Excel, o aspeto gráfico é o mesmo. Para uma maior semelhança, selecionámos o eixo das ordenadas e fizemos *Delete*.

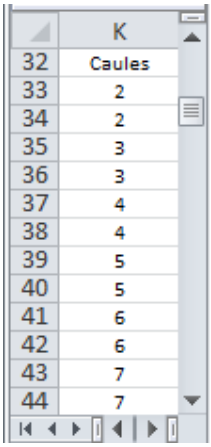
2. Processo utilizando as funções *REPT* e *COUNTIF* do Excel

- 1º passo – Inserir os dados numa coluna do Excel e determinar o mínimo e o máximo;
- 2º passo – Selecionar os valores pretendidos para serem os caules;
- 3º passo – Utilizar a função *REPT* e *COUNTIF*, conforme se exemplifica a seguir.

Vamos exemplificar os passos anteriores com os dados referentes às idades dos 230 deputados do ficheiro *DeputadosXII*, que se encontram nas células B2:B231.

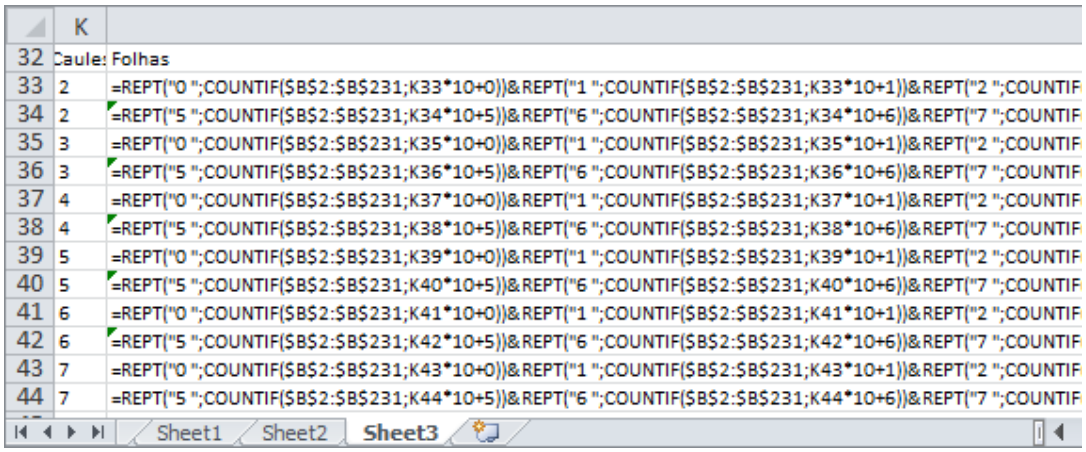
Como relativamente a estes dados já conhecemos o mínimo (=25) e o máximo (=70), vamos escolher como caules os dígitos 2, 3, 4, 5, 6 e 7. Por outro lado, como antecipamos a

existência de muitas folhas penduradas em cada caule, vamos considerar comprimentos de linha igual a 5 em vez de 10, pelo que consideramos cada caule dividido em dois sub-caules:



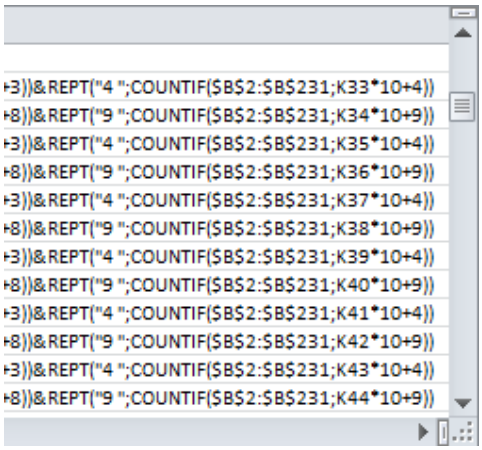
	K
32	Caules
33	2
34	2
35	3
36	3
37	4
38	4
39	5
40	5
41	6
42	6
43	7
44	7

Exemplificamos a seguir o processo de construção do caule-e-folhas:



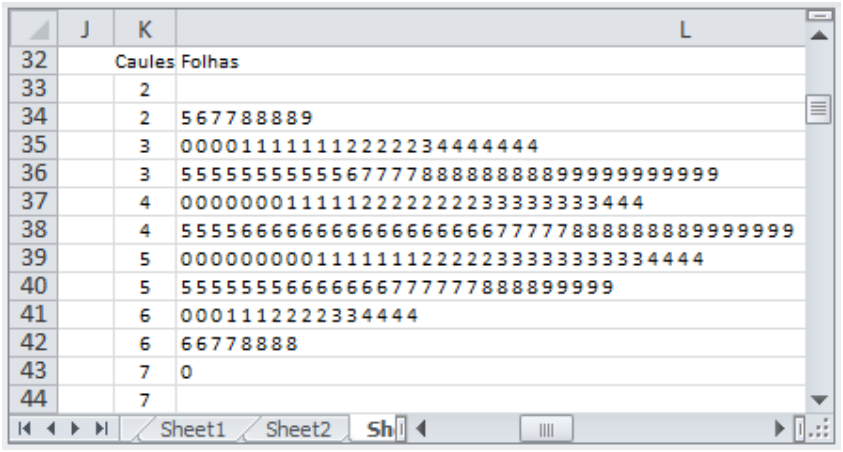
	K	
32	Caule:	Folhas
33	2	=REPT("0 ";COUNTIF($B$2:$B$231;K33*10+0))&REPT("1 ";COUNTIF($B$2:$B$231;K33*10+1))&REPT("2 ";COUNTIF
34	2	=REPT("5 ";COUNTIF($B$2:$B$231;K34*10+5))&REPT("6 ";COUNTIF($B$2:$B$231;K34*10+6))&REPT("7 ";COUNTIF
35	3	=REPT("0 ";COUNTIF($B$2:$B$231;K35*10+0))&REPT("1 ";COUNTIF($B$2:$B$231;K35*10+1))&REPT("2 ";COUNTIF
36	3	=REPT("5 ";COUNTIF($B$2:$B$231;K36*10+5))&REPT("6 ";COUNTIF($B$2:$B$231;K36*10+6))&REPT("7 ";COUNTIF
37	4	=REPT("0 ";COUNTIF($B$2:$B$231;K37*10+0))&REPT("1 ";COUNTIF($B$2:$B$231;K37*10+1))&REPT("2 ";COUNTIF
38	4	=REPT("5 ";COUNTIF($B$2:$B$231;K38*10+5))&REPT("6 ";COUNTIF($B$2:$B$231;K38*10+6))&REPT("7 ";COUNTIF
39	5	=REPT("0 ";COUNTIF($B$2:$B$231;K39*10+0))&REPT("1 ";COUNTIF($B$2:$B$231;K39*10+1))&REPT("2 ";COUNTIF
40	5	=REPT("5 ";COUNTIF($B$2:$B$231;K40*10+5))&REPT("6 ";COUNTIF($B$2:$B$231;K40*10+6))&REPT("7 ";COUNTIF
41	6	=REPT("0 ";COUNTIF($B$2:$B$231;K41*10+0))&REPT("1 ";COUNTIF($B$2:$B$231;K41*10+1))&REPT("2 ";COUNTIF
42	6	=REPT("5 ";COUNTIF($B$2:$B$231;K42*10+5))&REPT("6 ";COUNTIF($B$2:$B$231;K42*10+6))&REPT("7 ";COUNTIF
43	7	=REPT("0 ";COUNTIF($B$2:$B$231;K43*10+0))&REPT("1 ";COUNTIF($B$2:$B$231;K43*10+1))&REPT("2 ";COUNTIF
44	7	=REPT("5 ";COUNTIF($B$2:$B$231;K44*10+5))&REPT("6 ";COUNTIF($B$2:$B$231;K44*10+6))&REPT("7 ";COUNTIF

Sheet1 Sheet2 Sheet3



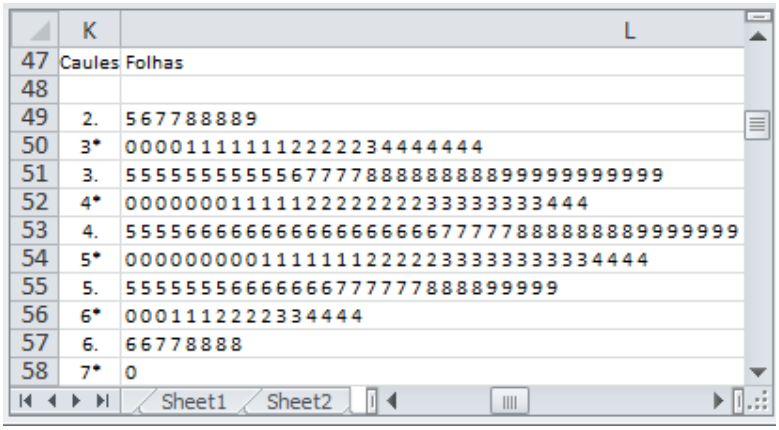
+3))&REPT("4 ";COUNTIF($B$2:$B$231;K33*10+4))
+8))&REPT("9 ";COUNTIF($B$2:$B$231;K34*10+9))
+3))&REPT("4 ";COUNTIF($B$2:$B$231;K35*10+4))
+8))&REPT("9 ";COUNTIF($B$2:$B$231;K36*10+9))
+3))&REPT("4 ";COUNTIF($B$2:$B$231;K37*10+4))
+8))&REPT("9 ";COUNTIF($B$2:$B$231;K38*10+9))
+3))&REPT("4 ";COUNTIF($B$2:$B$231;K39*10+4))
+8))&REPT("9 ";COUNTIF($B$2:$B$231;K40*10+9))
+3))&REPT("4 ";COUNTIF($B$2:$B$231;K41*10+4))
+8))&REPT("9 ";COUNTIF($B$2:$B$231;K42*10+9))
+3))&REPT("4 ";COUNTIF($B$2:$B$231;K43*10+4))
+8))&REPT("9 ";COUNTIF($B$2:$B$231;K44*10+9))

cujo resultado é o seguinte:



	J	K	L
32		Caules	Folhas
33		2	
34		2	567788889
35		3	000011111112222234444444
36		3	55555555555677778888888899999999999
37		4	00000001111122222222333333333444
38		4	55556666666666666666677777888888889999999
39		5	000000000111111122222333333333334444
40		5	55555556666666777777888899999
41		6	0001112222334444
42		6	66778888
43		7	0
44		7	

Para distinguir os dois sub-caules utilizamos um "*" no primeiro sub-caule e um "." no segundo sub-caule. Mas atenção! Antes de procedermos a esta alteração estética foi necessário copiar o diagrama anterior para outras células utilizando o *Paste Special→Values*.



	K	L
47	Caules	Folhas
48		
49	2.	567788889
50	3*	000011111112222234444444
51	3.	55555555555677778888888899999999999
52	4*	00000001111122222222333333333444
53	4.	55556666666666666666677777888888889999999
54	5*	000000000111111122222333333333334444
55	5.	55555556666666777777888899999
56	6*	0001112222334444
57	6.	66778888
58	7*	0

2.3.3.3 – Diagrama de extremos e quartis

Esta representação, muito simples, mas bastante elucidativa ao realçar a informação contida nos dados, no que diz respeito à simetria e variabilidade, pressupõe que se calculem algumas estatísticas necessárias para a sua construção.

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 2.3.7 – Construa um diagrama de extremos e quartis para a variável idade dos deputados do ficheiro DeputadosXII.

Construção do diagrama de extremos e quartis, em Excel:

1. Utilizando o Excel, começam por se calcular as estatísticas necessárias[4], que se apresentam da seguinte forma:

	E	F
1		
2	min	=MIN($B$2:$B$231)
3	Q1	=QUARTILE($B$2:$B$231;1)
4	median	=QUARTILE($B$2:$B$231;2)
5	q3	=QUARTILE($B$2:$B$231;3)
6	max	=QUARTILE($B$2:$B$231;4)

	E	F
1		
2	min	25
3	Q1	38,25
4	median	46
5	q3	53,75
6	max	70

2. A partir destas estatísticas, vão-se calcular alguns valores auxiliares que serão os utilizados para a construção do diagrama de extremos e quartis

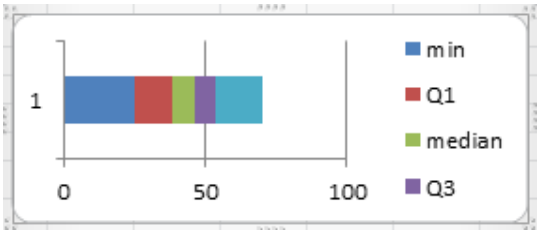
	I	J
1		Valores auxiliares
2	min	=F2
3	Q1	=F3-F2
4	median	=F4-F3
5	Q3	=F5-F4
6	max	=F6-F5

	I	J
1		Valores auxiliares
2	min	25
3	Q1	13,25
4	medi	7,75
5	Q3	7,75
6	max	16,25

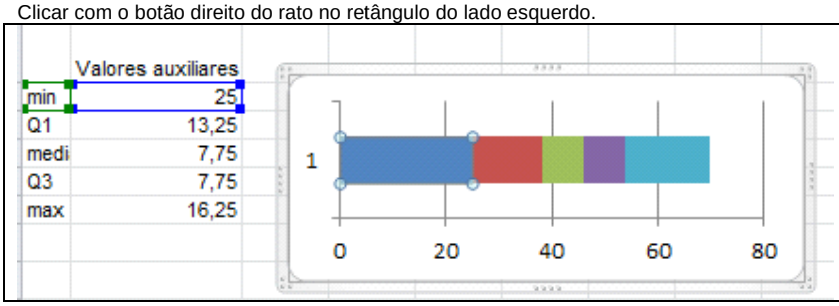
3.

	I	J
1		Valores auxiliares
2	min	25
3	Q1	13,25
4	medi	7,75
5	Q3	7,75
6	max	16,25

Com as células I2:J6 selecionadas como mostra a figura
Selecionar *Insert →Bar Stacked Bar →Switch Row/Column*, obtendo-se

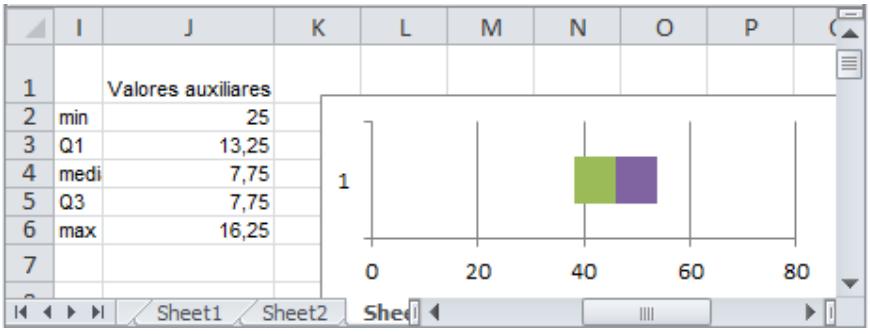


4. Apagar a legenda.
Clicar com o botão direito do rato no retângulo do lado esquerdo.

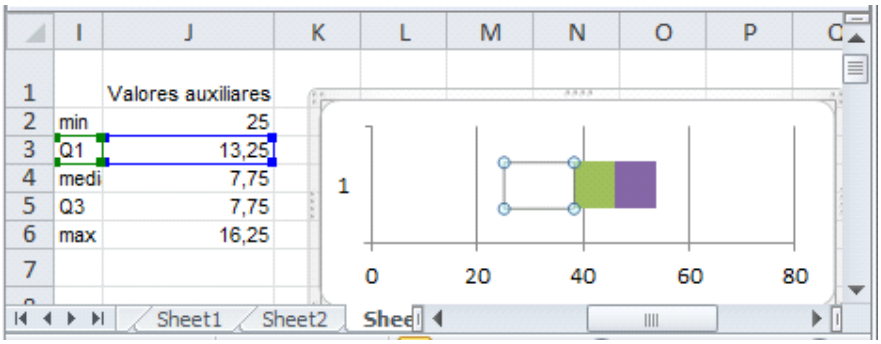
	Selecionar: *Format Data Series→Fill →No Fill*

5. Repetir o procedimento anterior para o retângulo seguinte e para o último retângulo (azul claro). Neste momento temos:

[4] No capítulo 3 abordaremos a determinação das estatísticas descritivas, utilizando o Excel.

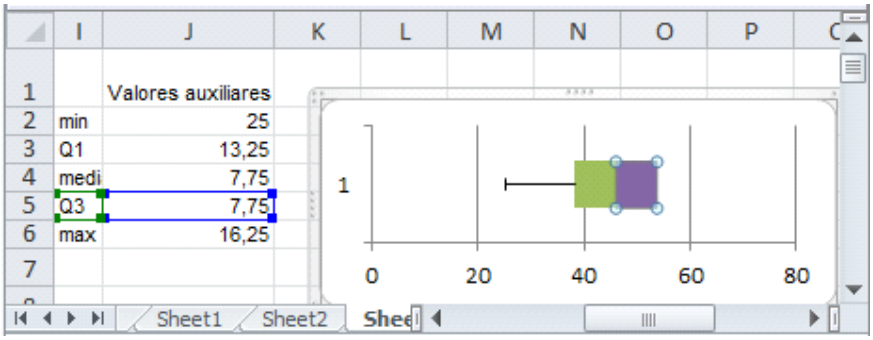


5. Selecionar no gráfico conforme mostra a figura e de seguida



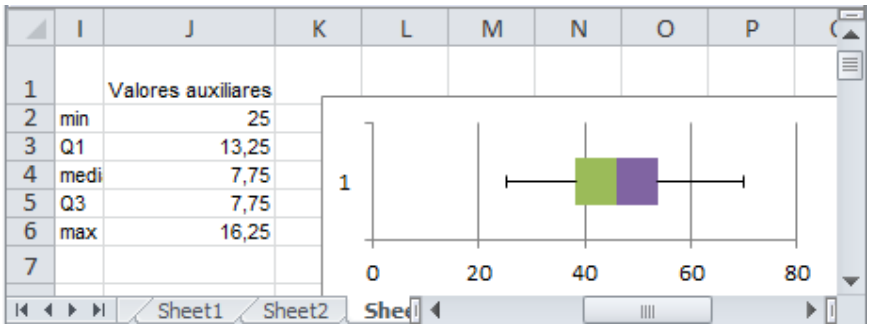
Selecionar
Layout →*Error Bars*
→*More Error Bars Options*
→*Minus*→
Percentage →100%

6. Selecionar no gráfico conforme mostra a figura e de seguida

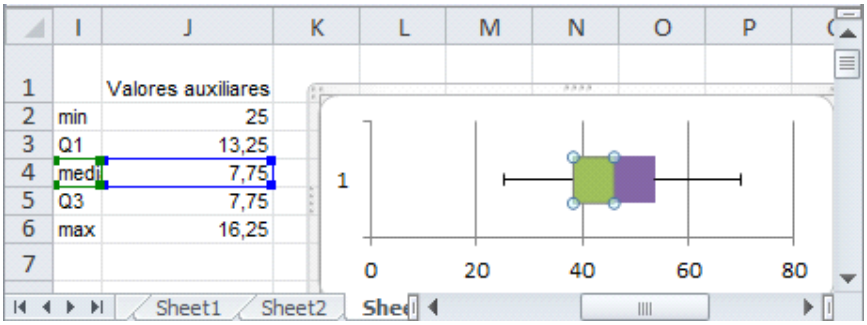


Selecione
Layout →*Error Bars*
→*More Error Bars Options*
→*Plus*→*Custom*
→*Specify value*
→*Positive Error value*
→Inserir o valor da célula J6

Os procedimentos anteriores conduziram à seguinte figura:

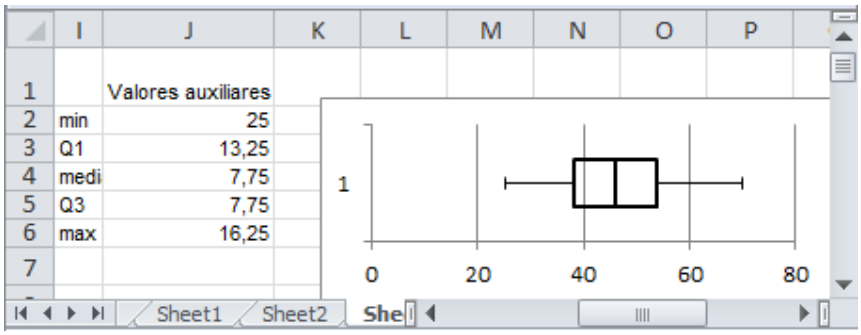


7. Com o botão direito do rato clicar no gráfico conforme mostra a figura e de seguida

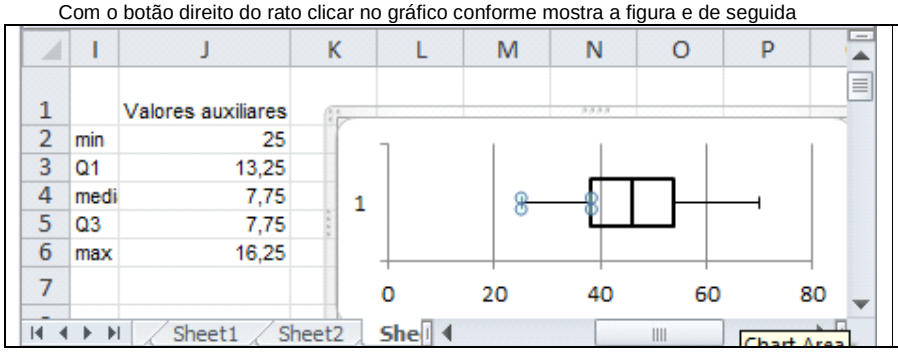


Selecionar
Format Data Series →
Fill→No Fill →
Border Color →Solid Line
(Black) →Border Stylus→2

Proceder como anteriormente para a parte roxa para obter:

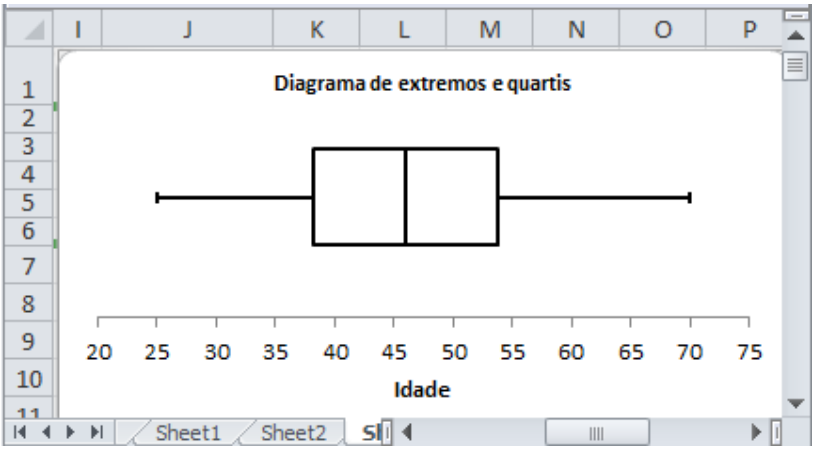


Com o botão direito do rato clicar no gráfico conforme mostra a figura e de seguida

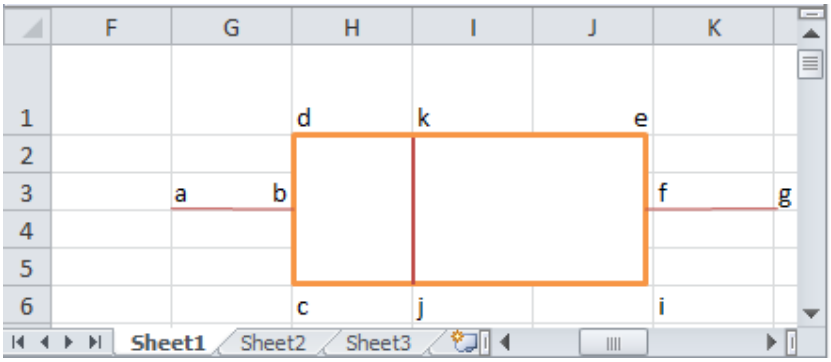


Format Error Bars→
Line Style→Width
→2

Proceder como anteriormente com a linha da direita
Apagar as linhas verticais
Reformular a escala do eixo das abcissas
Apagar o eixo vertical
Introduzir legendas

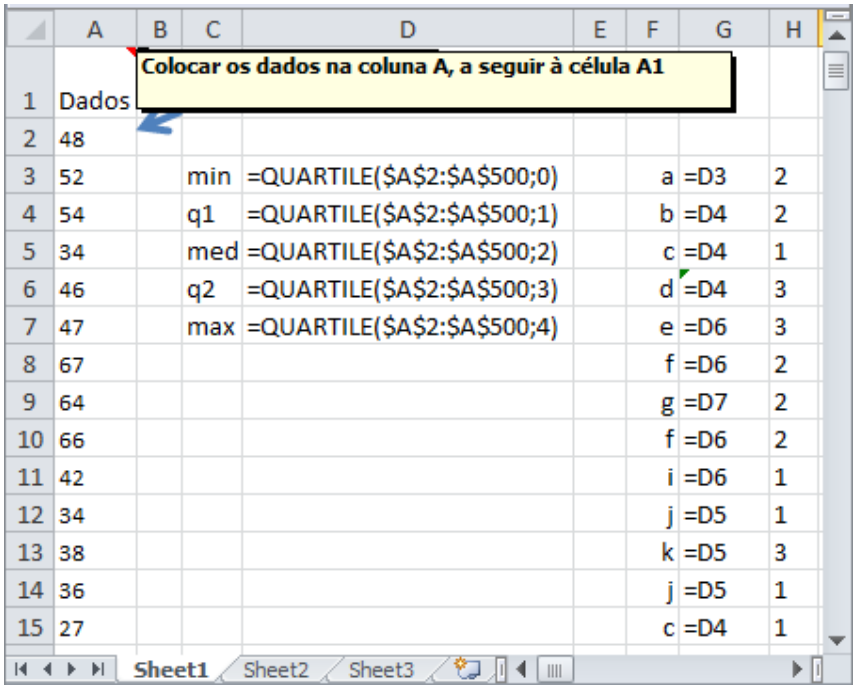


Nota – Utilizando um processo idêntico ao da secção 2.3.3.1.3 em que, utilizando a funcionalidade *Scatter*, se construiu um histograma com classes de amplitude diferente, pode-se construir um diagrama de extremos e quartis. Consideremos então um diagrama de extremos e quartis, em que marcámos alguns pontos com letras:



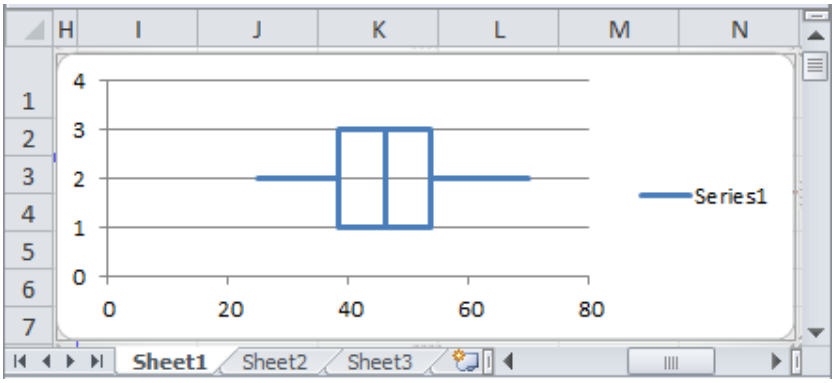
Repare que se unir o ponto **a** com **b**, de seguida com **c**, até esgotar todos os pontos, obtém o diagrama de extremos e quartis. Então, para obter a representação gráfica desejada, basta construir uma tabela, numa folha de Excel, com as coordenadas dos pontos que pretendemos unir e utilizar a representação *Scatter.*

Apresentamos a seguir uma folha de Excel preparada para obter a representação, bastando para isso colocar os dados (até 500 dados) na coluna A:

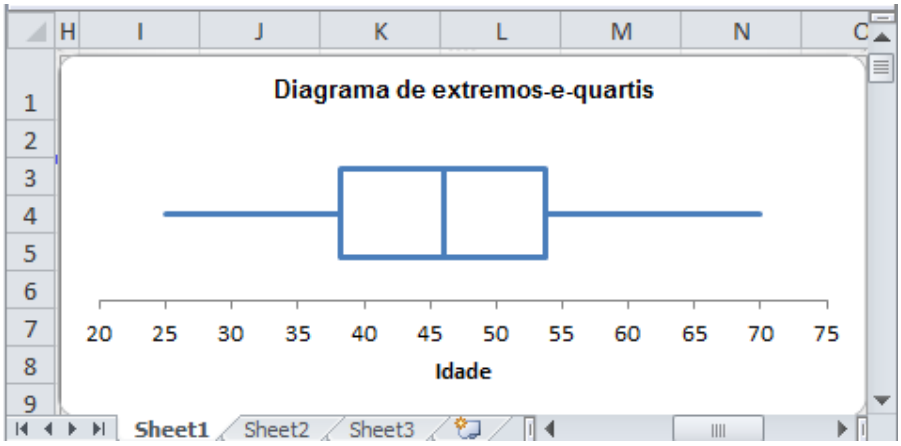


	A	B	C	D	E	F	G	H
1	Dados							
2	48							
3	52		min	=QUARTILE($A$2:$A$500;0)		a	=D3	2
4	54		q1	=QUARTILE($A$2:$A$500;1)		b	=D4	2
5	34		med	=QUARTILE($A$2:$A$500;2)		c	=D4	1
6	46		q2	=QUARTILE($A$2:$A$500;3)		d	=D4	3
7	47		max	=QUARTILE($A$2:$A$500;4)		e	=D6	3
8	67					f	=D6	2
9	64					g	=D7	2
10	66					f	=D6	2
11	42					i	=D6	1
12	34					j	=D5	1
13	38					k	=D5	3
14	36					j	=D5	1
15	27					c	=D4	1

Selecionando as células G3:H15, faça Insert Scatter 5º subtipo



Para obter a representação final, basta apagar a legenda e o eixo vertical, introduzir títulos e mudar a escala:



Sempre que pretender um diagrama de extremos e quartis para qualquer outro conjunto de dados (de dimensão menor ou igual a 500), basta inserir os dados na coluna A, a partir da célula A2.

2.3.3.3.1 Diagramas de extremos e quartis paralelos

A representação de um conjunto de dados, num diagrama de extremos e quartis, é especialmente indicada para comparação de várias amostras, como se exemplifica a seguir:

Exemplo 2.3.8 – Vamos comparar as idades dos deputados dos diferentes grupos parlamentares. Não incluímos o grupo PV por só serem dois deputados.

A exemplo do que foi feito para a construção do diagrama de extremos e quartis para uma coleção de dados, também aqui será necessário calcular algumas características amostrais e a partir delas alguns valores auxiliares.

Construção dos diagramas de extremos e quartis paralelos:

1. Calcular as características amostrais relevantes e a partir dessas calcular os valores auxiliares, utilizados para a construção do gráfico, de acordo com a seguinte metodologia:

Valores originais	Valores auxiliares
min	Q1 - min
Q1	Q1
median	median - Q1
Q3	Q3 - median
max	max - Q3

	E	F	G	H	I	J	L	M	N	O	P	Q
7			Valores originais						Valores auxiliares			
8		BE	CDS-PP	PCP	PS	PSD		BE	CDS-PP	PCP	PS	PSD
9												
10	min	=QUAR	=QUARTI	=QUART	=QUAR	=QUAF	min	=F11-F10	=G11-G10	=H11-H10	=I11-I10	=J11-J10
11	Q1	=QUAR	=QUARTI	=QUART	=QUAR	=QUAF	25th	=F11	=G11	=H11	=I11	=J11
12	median	=QUAR	=QUARTI	=QUART	=QUAR	=QUAF	50th	=F12-F11	=G12-G11	=H12-H11	=I12-I11	=J12-J11
13	Q3	=QUAR	=QUARTI	=QUART	=QUAR	=QUAF	75th	=F13-F12	=G13-G12	=H13-H12	=I13-I12	=J13-J12
14	max	=QUAR	=QUARTI	=QUART	=QUAR	=QUAF	max	=F14-F13	=G14-G13	=H14-H13	=I14-I13	=J14-J13

Sheet1 Sheet2 Sheet3

	E	F	G	H	I	J	K	L	M	N	O	P	Q
7			Valores originais							auxiliares			
8		BE	CDS-PP	PCP	PS	PSD			BE	CDS-PP	PCP	PS	PSD
9													
10	min	32	28	28	27	25		min	5,5	10,5	4,75	15,25	13
11	Q1	37,5	38,5	32,75	42,25	38		25th	37,5	38,5	32,75	42,25	38
12	median	51,5	42	39,5	49,5	46		50th	14	3,5	6,75	7,25	8
13	Q3	55,5	50,25	54	56,75	53		75th	4	8,25	14,5	7,25	7
14	max	60	63	67	68	70		max	4,5	12,75	13	11,25	17

Sheet1 Sheet2 Sheet3

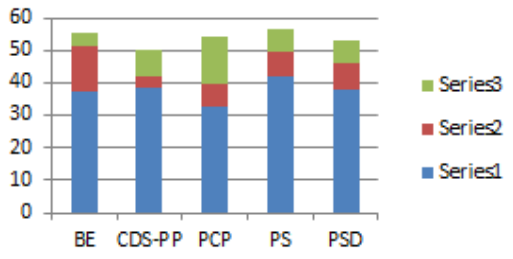
2. Selecionar as células que contêm as etiquetas (M8:Q8) e os valores auxiliares das células M11:Q13 e de seguida

	L	M	N	O	P	Q
8		BE	CDS-PP	PCP	PS	PSD
9						
10	min	5,5	10,5	4,75	15,25	13
11	25th	37,5	38,5	32,75	42,25	38
12	50th	14	3,5	6,75	7,25	8
13	75th	4	8,25	14,5	7,25	7
14	max	4,5	12,75	13	11,25	17

Sheet1

Average: 17,96666667 Count: 20

Selecionar *Insert →Column→ Stacked Column*, obtendo-se



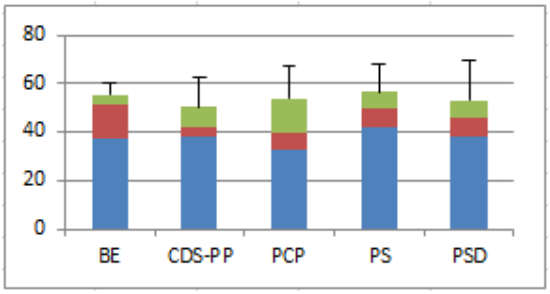
3. Apagar a legenda e selecionar no gráfico conforme mostra a figura

BE	CDS-PP	PCP	PS	PSD
5,5	10,5	4,75	15,25	13
37,5	38,5	32,75	42,25	38
14	3,5	6,75	7,25	8
4	8,25	14,5	7,25	7
4,5	12,75	13	11,25	17

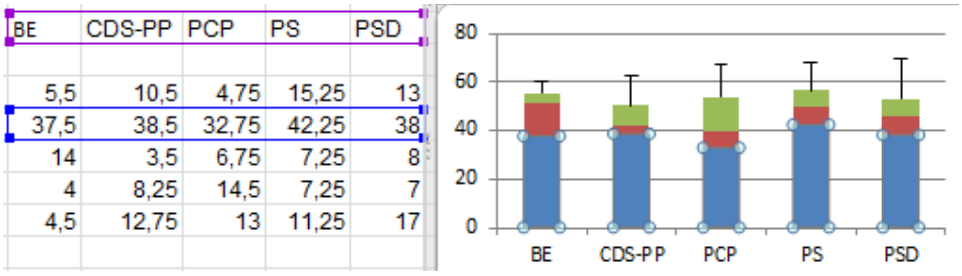
Selecionar

Layout →Error Bars →More Error Bars Options →Plus→Custom →Specify value

→*Positive Error value* →Inserir os valores das células M14:Q14 (ver figura no passo 2)

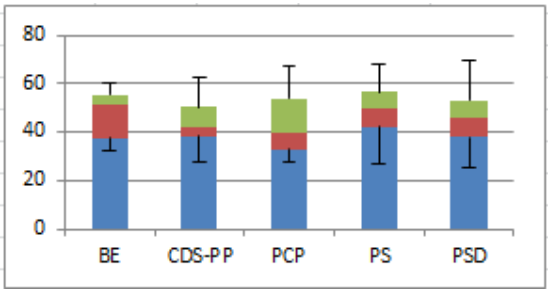


4. Selecionar no gráfico conforme mostra a figura

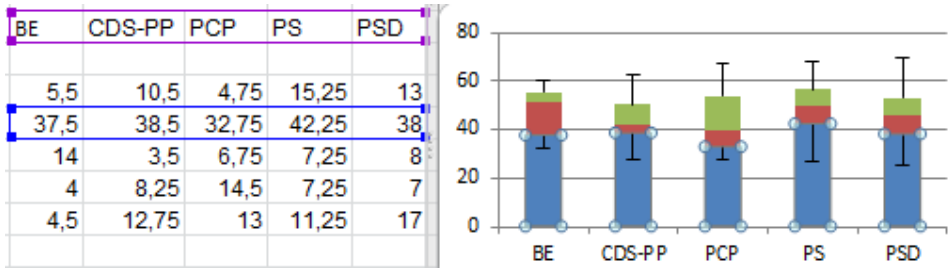


BE	CDS-PP	PCP	PS	PSD
5,5	10,5	4,75	15,25	13
37,5	38,5	32,75	42,25	38
14	3,5	6,75	7,25	8
4	8,25	14,5	7,25	7
4,5	12,75	13	11,25	17

Selecionar
Layout →*Error Bars* →*More Error Bars Options* →*Minus*→*Custom*
→*Specify value*
→*Negative Error value* →Inserir os valores das células M10:Q10 (ver figura no passo 2)

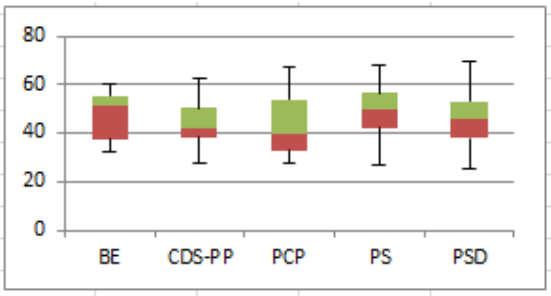


5. Com o botão direito do rato clicar conforme mostra a figura

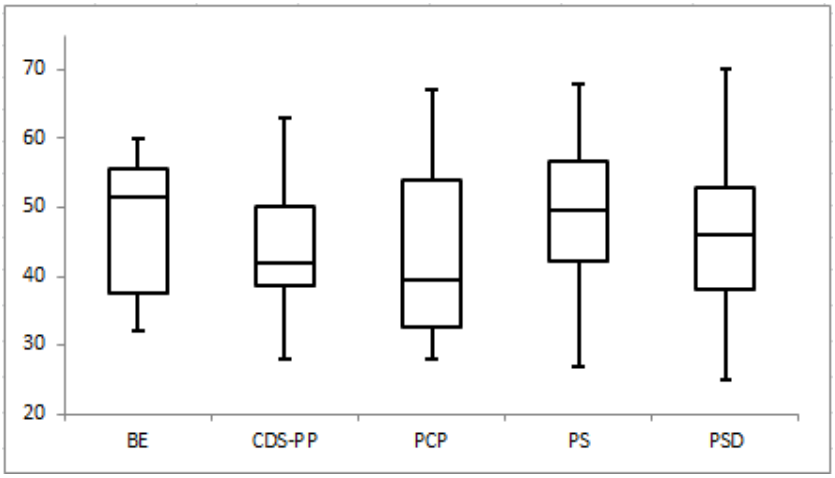


BE	CDS-PP	PCP	PS	PSD
5,5	10,5	4,75	15,25	13
37,5	38,5	32,75	42,25	38
14	3,5	6,75	7,25	8
4	8,25	14,5	7,25	7
4,5	12,75	13	11,25	17

Selecionar
Format Data Series → *Fill*→*No Fill* →
Border Color →*No Border*

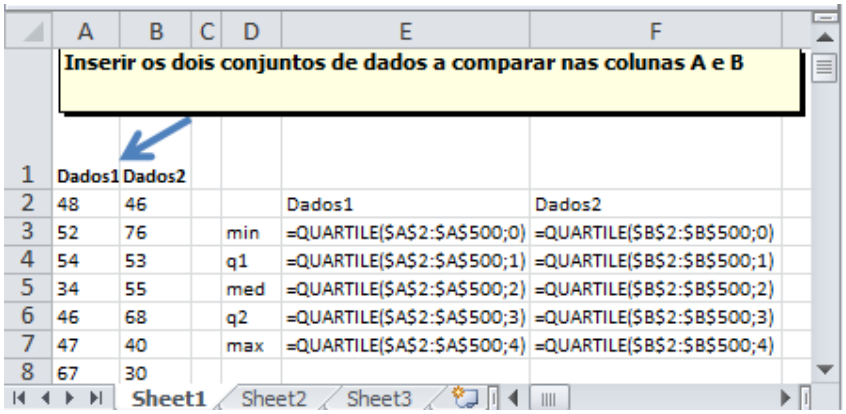


7. Para chegar à forma final dos diagramas de extremos e quartis paralelos, proceder como no passo 7 para o diagrama de extremos e quartis para um único conjunto de dados:



O gráfico anterior é bastante elucidativo na comparação das idades dos deputados dos diferentes grupos parlamentares. É interessante verificar que as distribuições das idades dos deputados do PSD e PS são razoavelmente semelhantes, embora se note que nos deputados do PSD as 50% idades do centro sejam um pouco inferiores às correspondentes do PS. As distribuições dos outros grupos parlamentares apresentam enviesamento, para a direita no caso do CDS-PP e PCP e para a esquerda no caso do BE. Este enviesamento no caso do BE é sintoma de, neste grupo parlamentar, haver uma maior variabilidade de idades nos deputados mais jovens.

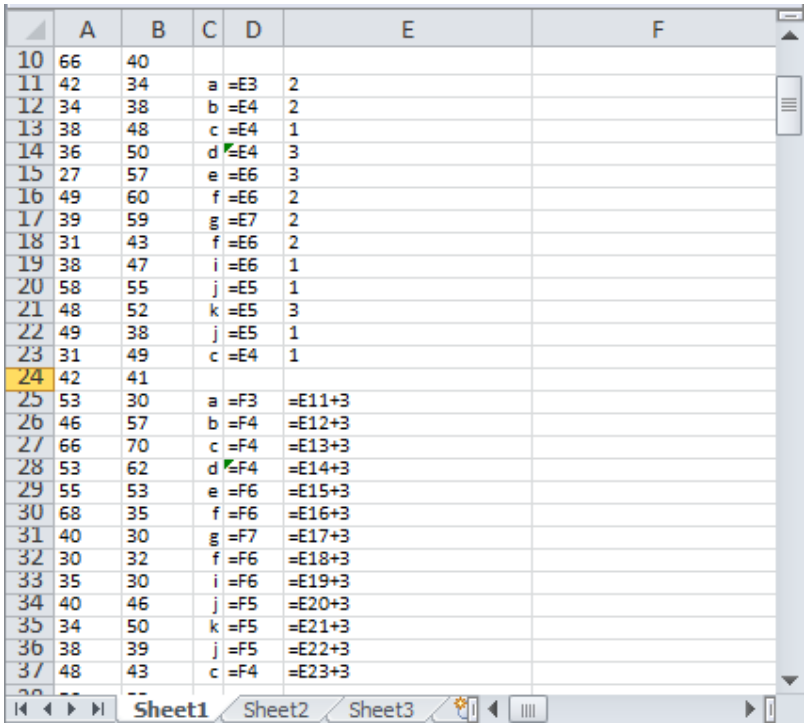
Nota – De forma idêntica à que se fez para a construção do diagrama de extremos e quartis utilizando a funcionalidade Scatter, também se podem construir os diagramas de extremos e quartis utilizando a seguinte metodologia, que vamos mostrar para 2 conjuntos de dados:



	A	B	C	D	E	F
	Inserir os dois conjuntos de dados a comparar nas colunas A e B					
1	Dados1	Dados2				
2	48	46			Dados1	Dados2
3	52	76		min	=QUARTILE($A$2:$A$500;0)	=QUARTILE($B$2:$B$500;0)
4	54	53		q1	=QUARTILE($A$2:$A$500;1)	=QUARTILE($B$2:$B$500;1)
5	34	55		med	=QUARTILE($A$2:$A$500;2)	=QUARTILE($B$2:$B$500;2)
6	46	68		q2	=QUARTILE($A$2:$A$500;3)	=QUARTILE($B$2:$B$500;3)
7	47	40		max	=QUARTILE($A$2:$A$500;4)	=QUARTILE($B$2:$B$500;4)
8	67	30				

Sheet1 Sheet2 Sheet3

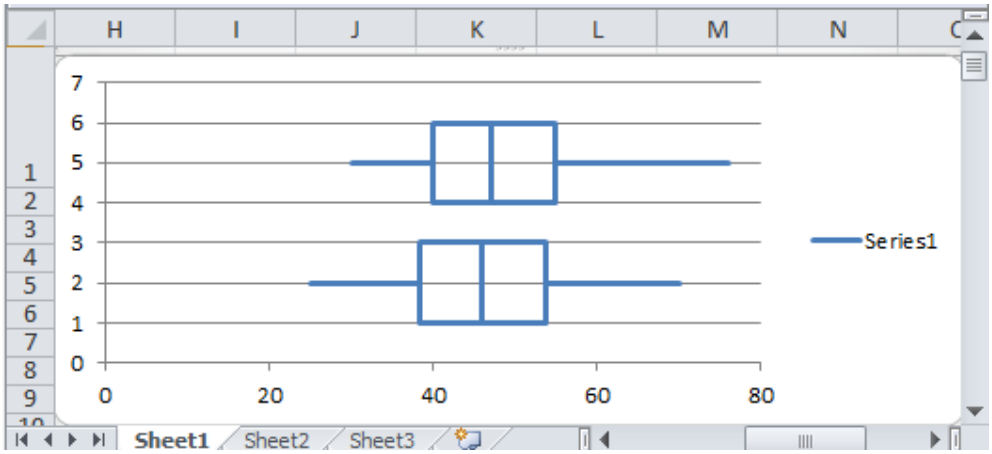
Os pontos correspondentes aos dois diagramas de extremos e quartis que se pretendem construir são colocados nas mesmas colunas, mas deixando 2 células de intervalo (no nosso exemplo as células D24 e E24, para que as duas representações não fiquem ligadas. Repare-se que incrementámos as ordenadas do segundo conjunto de dados, de 3 unidades:



	A	B	C	D	E	F
10	66	40				
11	42	34	a	=E3	2	
12	34	38	b	=E4	2	
13	38	48	c	=E4	1	
14	36	50	d	=E4	3	
15	27	57	e	=E6	3	
16	49	60	f	=E6	2	
17	39	59	g	=E7	2	
18	31	43	f	=E6	2	
19	38	47	i	=E6	1	
20	58	55	j	=E5	1	
21	48	52	k	=E5	3	
22	49	38	j	=E5	1	
23	31	49	c	=E4	1	
24	42	41				
25	53	30	a	=F3	=E11+3	
26	46	57	b	=F4	=E12+3	
27	66	70	c	=F4	=E13+3	
28	53	62	d	=F4	=E14+3	
29	55	53	e	=F6	=E15+3	
30	68	35	f	=F6	=E16+3	
31	40	30	g	=F7	=E17+3	
32	30	32	f	=F6	=E18+3	
33	35	30	i	=F6	=E19+3	
34	40	46	j	=F5	=E20+3	
35	34	50	k	=F5	=E21+3	
36	38	39	j	=F5	=E22+3	
37	48	43	c	=F4	=E23+3	

Sheet1 Sheet2 Sheet3

Selecionando as células D11:E37, obtém-se a seguinte representação:

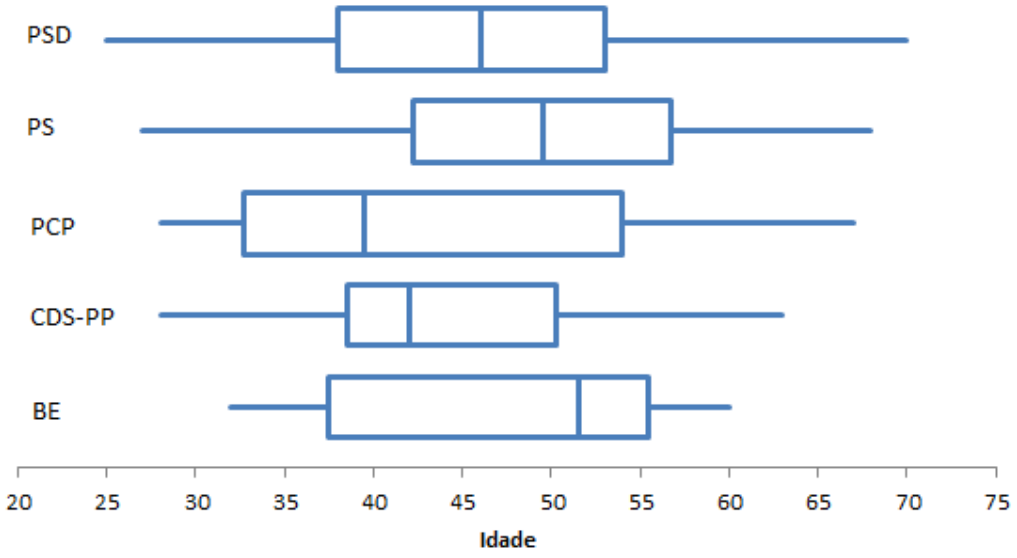


Para comparar as idades dos deputados dos diferentes grupos parlamentares, considerámos os 5 conjuntos de dados referentes às idades respetivas, a partir dos quais se obtiveram as estatísticas do quadro seguinte:

	D	E	F	G	H	I
1		BE	CDS-P	PCP	PS	PSD
2						
3	min	32	28	28	27	25
4	q1	37,5	38,5	32,8	42,3	38
5	med	51,5	42	39,5	49,5	46
6	q2	55,5	50,3	54	56,8	53
7	max	60	63	67	68	70

Sheet1 Sheet

Agora, para obter os 5 diagramas, em vez de 2 séries de pontos como no caso anterior, é necessário considerar 5 séries de pontos, todas separadas de células vazias e incrementando as ordenadas de cada série de 3 unidades (ou outro valor se se pretender os diagramas mais afastados ou mais juntos). O resultado final é o que se apresenta a seguir:



As etiquetas correspondentes aos grupos parlamentares foram introduzidas como texto.

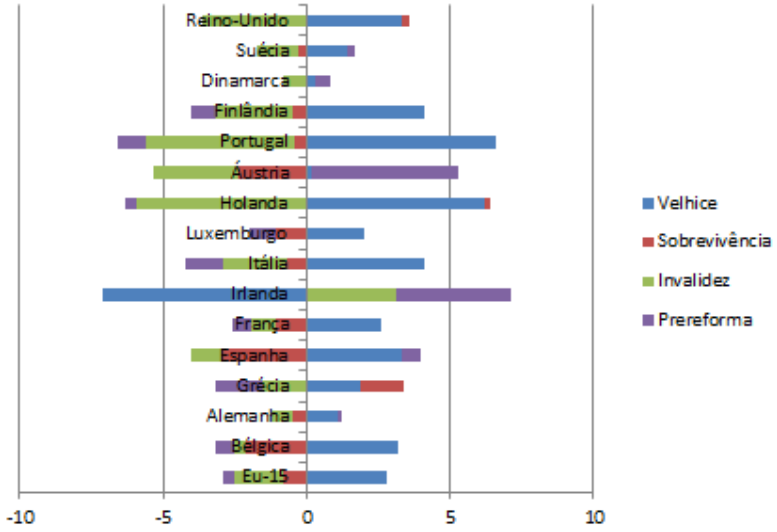
2.4 – Alguns exemplos

A seguir apresentamos alguns exemplos, sobre a forma de projetos, para os quais podemos utilizar vários tipos de representações gráficas, algumas já referidas anteriormente, outras introduzidas pela primeira vez, mas que apresentam realização imediata com o Excel.

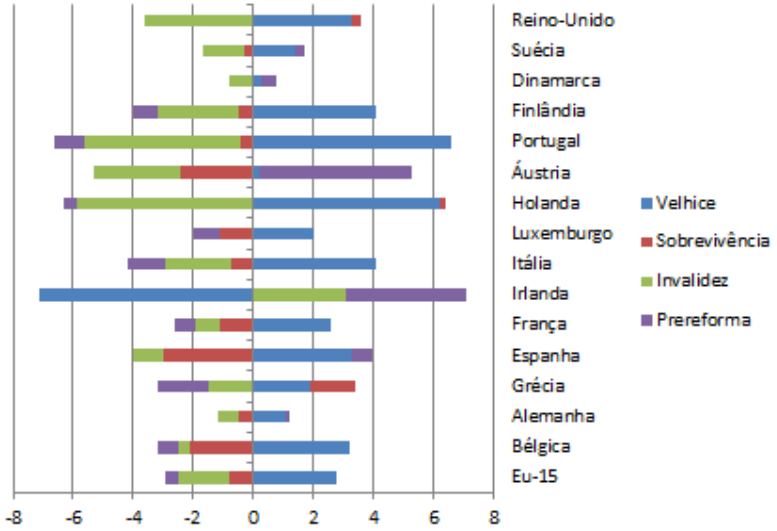
Projeto 1 - Neste projeto são apresentados alguns dados relativamente à Modificação da Estrutura das Categorias de Pensões entre 1993 e 2001 (em pontos percentuais) (Eurostat – Statistiques en bref – Population et conditions sociales, 8/2004):

	Velhice	Sobrevivência	Invalidez	Pré-reforma
Eu-15	2,8	-0,8	-1,7	-0,4
Bélgica	3,2	-2,1	-0,4	-0,7
Alemanha	1,1	-0,5	-0,7	0,1
Grécia	1,9	1,5	-1,5	-1,7
Espanha	3,3	-3	-1	0,7
França	2,6	-1,1	-0,8	-0,7
Irlanda	-7,1	0	3,1	4
Itália	4,1	-0,7	-2,2	-1,3
Luxemburgo	2	-1,1	0	-0,9
Holanda	6,2	0,2	-5,9	-0,4
Áustria	0,2	-2,4	-2,9	5,1
Portugal	6,6	-0,4	-5,2	-1
Finlândia	4,1	-0,5	-2,7	-0,8
Dinamarca	0,3	0	-0,8	0,5
Suécia	1,4	-0,3	-1,4	0,3
Reino-Unido	3,3	0,3	-3,6	0

Uma forma adequada para representar estes dados, é através de um diagrama de barras, nomeadamente barras horizontais, selecionando *Insert→ Bar→ Stacked Bar*:



Vamos fazer alguma "cosmética" na representação gráfica anterior, nomeadamente mudando a escala para –8 a 8 e fazendo com que as legendas não se sobreponham ao gráfico:



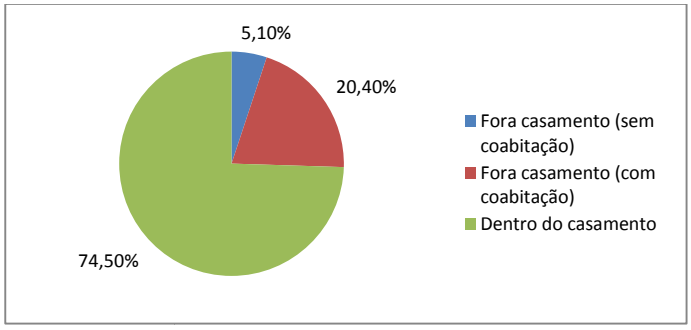
Projeto 2 – Entre os dois últimos recenseamentos da população portuguesa, os Censos 91 e os Censos 2001, realizados, respetivamente, em 15 de Abril de 1991 e 12 de Março de 2001, verificou-se que a população residente no território nacional passou de 9.867.147 para 10.356.117 habitantes, a que corresponde um acréscimo de 4.8%. Na generalidade das regiões verificou-se um aumento da população, com exceção das regiões do Alentejo e Madeira. Partindo dos resultados censitários definitivos, estimou-se a população residente em 31 de Dezembro de 2002 em 10.407.500 indivíduos, dos quais 5.030.200 do sexo masculino.

Apresentam-se a seguir algumas tabelas e gráficos com alguns indicadores (www.ine.pt):

1.Nados-vivos segundo a filiação – 2002

	A	B	C
1			
2	Fora casamento (sem coabitação)	5,10%	
3	Fora casamento (com coabitação)	20,40%	
4	Dentro do casamento	74,50%	

Uma representação adequada para a tabela anterior é o diagrama circular. Assim, vamos selecionar *Insert→Chart* →*Pie* →*1ºsubtipo* →*Layout* → *Data labels*:

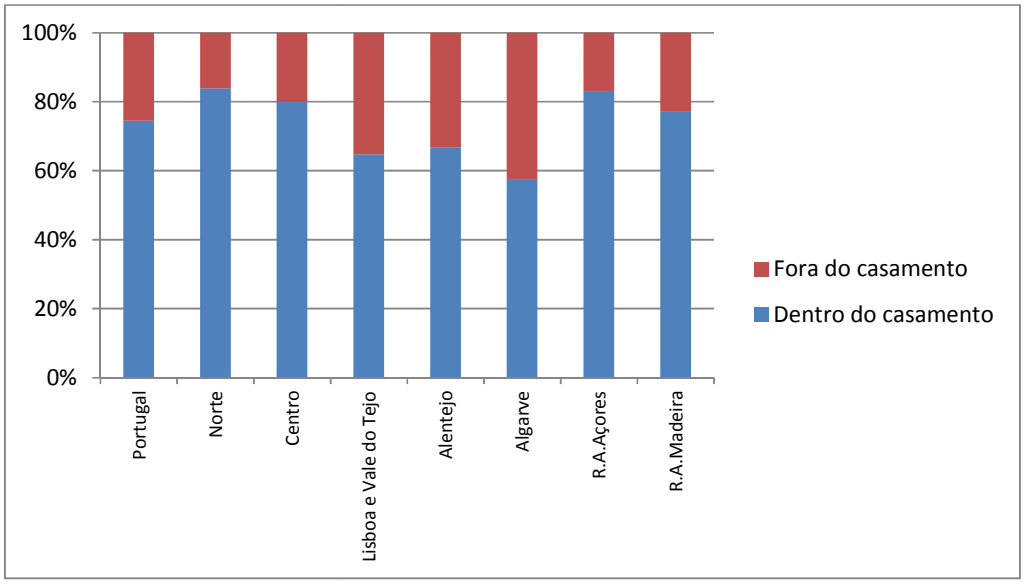


Nados-vivos segundo a filiação, por regiões:

	A	B
9		Dentro do casamento
10	Portugal	74,50%
11	Norte	83,80%
12	Centro	80,00%
13	Lisboa e Vale do Tejo	64,70%
14	Alentejo	66,70%
15	Algarve	57,60%
16	R.A.Açores	83,10%
17	R.A.Madeira	77,30%

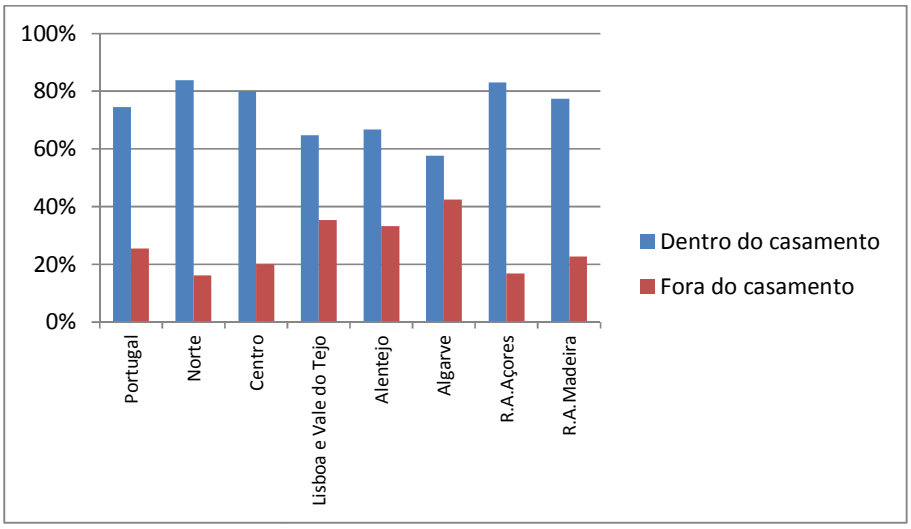
	A	B	C
9		Dentro do casamento	Fora do casament
10	Portugal	74,50%	25,50%
11	Norte	83,80%	16,20%
12	Centro	80,00%	20,00%
13	Lisboa e Vale do Tejo	64,70%	35,30%
14	Alentejo	66,70%	33,30%
15	Algarve	57,60%	42,40%
16	R.A.Açores	83,10%	16,90%
17	R.A.Madeira	77,30%	22,70%

Acrescentámos à tabela do lado esquerdo, retirada da página do INE, uma outra coluna com os filhos fora do casamento e optámos por uma representação em barras verticais:



Observação: Foi possível optarmos pela representação gráfica anterior, uma vez que os dados das duas características em estudo somavam 100%.

Outra representação possível obtém-se selecionando *Insert→Column →1ºsubtipo*:

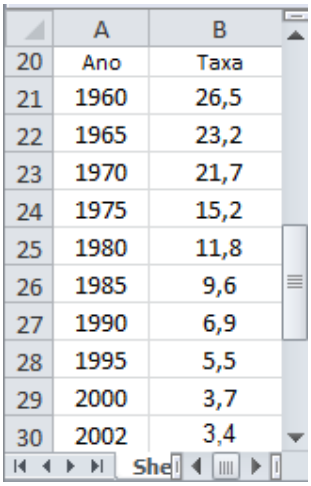


2. Taxa de mortalidade fetal tardia (Taxa mft) (28 ou mais semanas de gestação):

1960	26.5‰
1965	23.2‰
1970	21.7‰
1975	15.2‰
1980	11.8‰
1985	9.6‰
1990	6.9‰
1995	5.5‰
2000	3.7‰

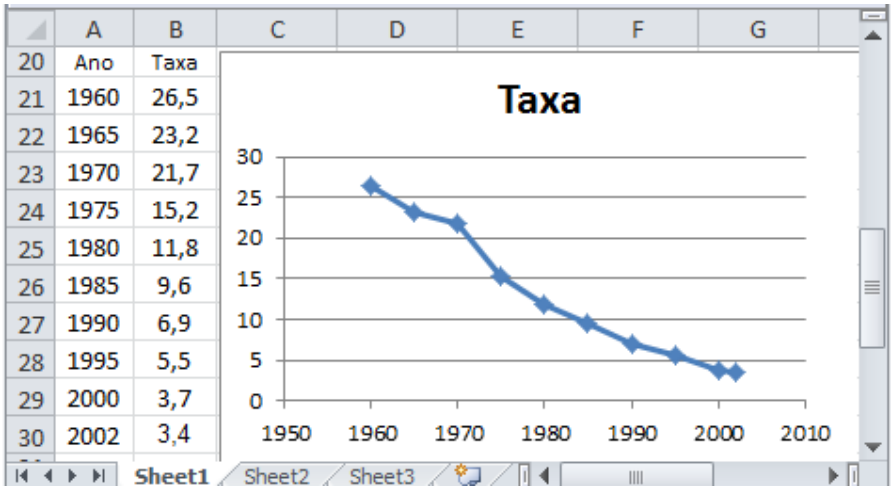
2002	3.4‰

Introduzimos a tabela anterior numa folha de Excel e antes de procedermos a uma representação gráfica passámos os pontos para vírgulas e retirámos a permilagem, não reconhecida no Excel.

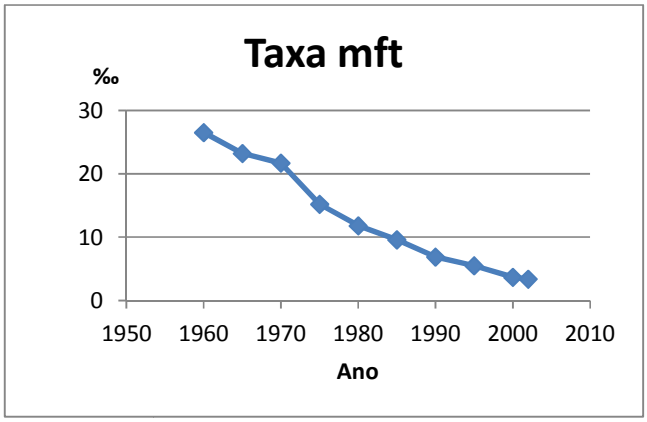


	A	B
20	Ano	Taxa
21	1960	26,5
22	1965	23,2
23	1970	21,7
24	1975	15,2
25	1980	11,8
26	1985	9,6
27	1990	6,9
28	1995	5,5
29	2000	3,7
30	2002	3,4

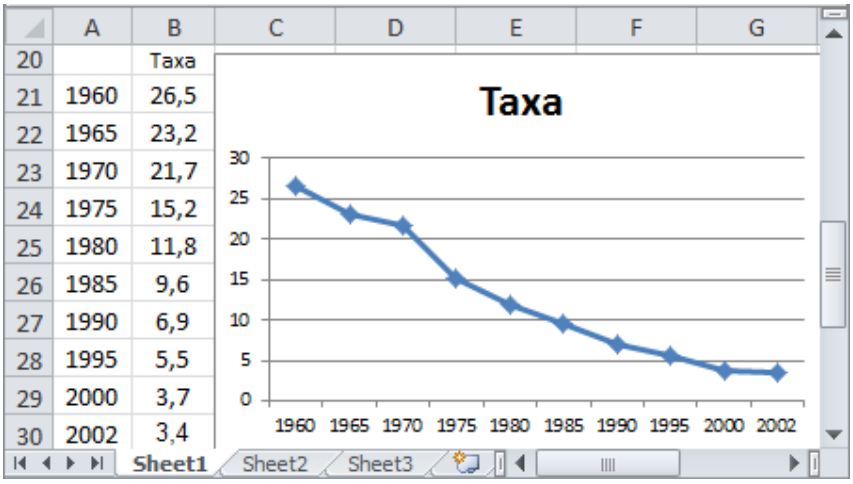
Seguidamente depois de selecionar as células A20 a B30, selecionámos *Insert →Scatter →4ºsubtipo*:



A representação gráfica anterior só ficará completa depois de acrescentar os títulos aos eixos.



Chamamos a atenção para o facto de ser possível obter uma representação aparentemente semelhante à anterior utilizando a opção *Chart* →*Line* →*4ºsubtipo*:



Repare-se, no entanto, que a representação anterior não está correta, pois a variável tempo do eixo dos xx está a ser interpretada como uma variável qualitativa e não quantitativa como deveria ser. Assim, o intervalo entre 1995 e 2000 é igual ao intervalo entre 2000 e 2002, o que obviamente não está correto.

3. Taxa de mortalidade infantil

1960	77.5‰
1965	64.9‰
1970	58.0‰
1975	38.9‰
1980	24.3‰
1985	17.8‰
1990	10.9‰
1995	7.5‰
2000	5.5‰
2002	5.0‰

A representação gráfica dos dados desta tabela pode ser idêntica à do ponto anterior.

4. Casamentos segundo a forma de celebração

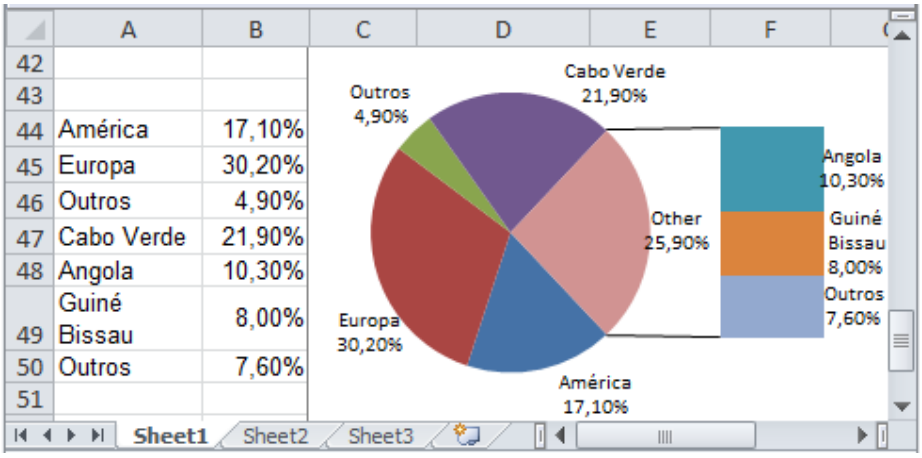
Unidade %	Civil	Católico
1960	9.2	90.8
1965	11.8	88.2
1970	13.4	86.6
1975	20.0	80.0
1980	25.3	74.7
1985	25.9	74.1
1990	27.5	72.5
1995	31.2	68.8
2000	35.2	64.8
2002	37.5	62.5

Para esta tabela pode-se usar uma representação gráfica idêntica à usada no ponto 1, para mostrar a percentagem de filhos dentro e fora do casamento.

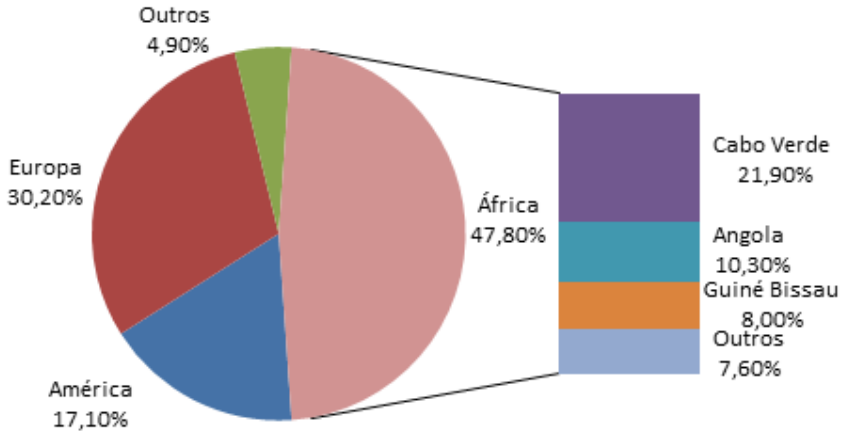
5. População estrangeira com estatuto legal de residente segundo a nacionalidade

América	17,1%			Angola	10,3%
Europa	30,2%		África	Cabo Verde	21,9%
África	47,8%			Guiné Bissau	8,0%
Outros	4,9%			Outros	7,6%

Para fazer uma representação destes dados recorremos a um diagrama em *Pie* (circular), mas num subtipo especial que permite visualizar a forma como África está repartida. Assim considere-se a seguinte tabela em Excel, ocupando as células A44 a B50 e selecione-se *Chart*  *→Pie →4ºsubtipo→Layout→Data labels:*



Para incluir Cabo Verde na parte direita do gráfico carregar com o botão direito do rato em qualquer parte do gráfico e selecionar *Format Data Series → Second plot contais the last: 4 →Close.* Formatar convenientemente o gráfico e por último substituir Other (com 47,8%) por África:

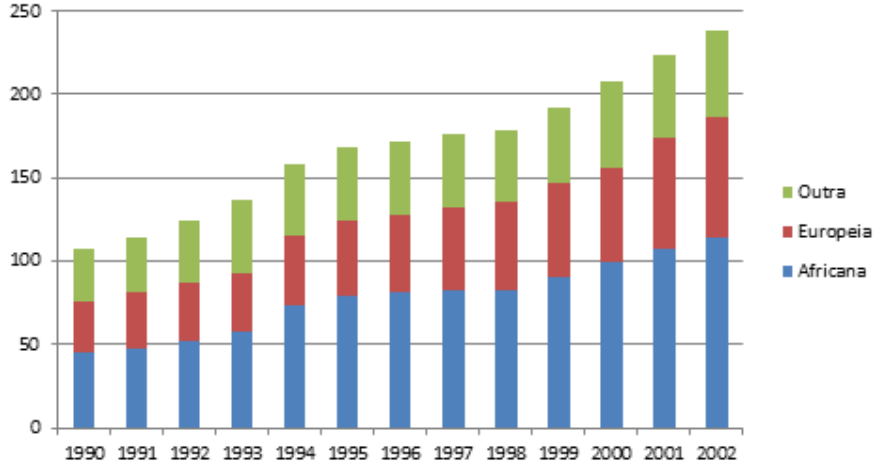


Para representar os dados da tabela seguinte

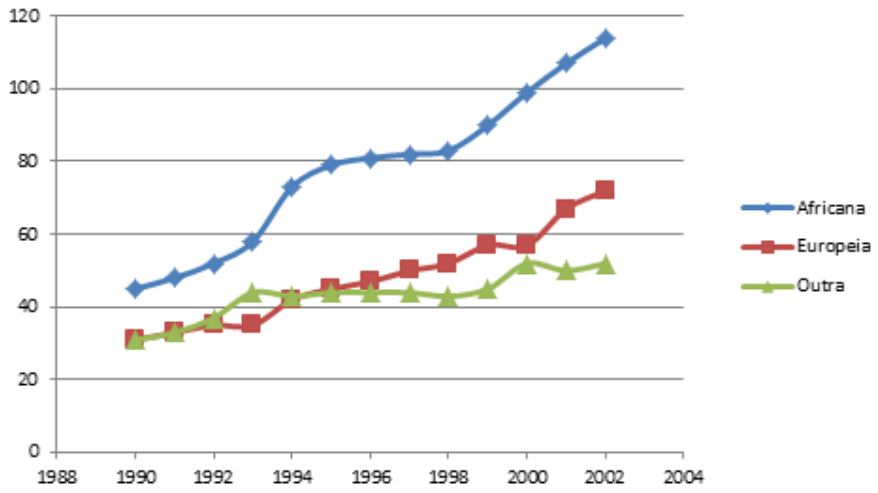
	Africana(1)	Europeia	Outra
1990	45	31	31
1991	48	33	33
1992	52	35	37
1993	58	35	44
1994	73	42	43
1995	79	45	44
1996	81	47	44
1997	82	50	44
1998	83	52	43
1999	90	57	45
2000	99	57	52
2001	107	67	50
2002	114	72	52

(1)Unidade 10^3

podemos considerar o *2º subtipo* de *Column* (chama-se a atenção para que neste caso não seria correto utilizar o *3º subtipo* de *Column*, uma vez que são dadas as frequências absolutas e não as frquências relativas):



ou o *2º subtipo* de *Scatter*, o chamado diagrama ou gráfico de linha:



Como vimos há várias representações gráficas para os dados de uma mesma tabela, umas mais sugestivas ou mais informativas do que outras. Por exemplo, entre as duas representações gráficas anteriores, para os mesmos dados, sugerimos esta última por ser de mais fácil leitura e transmitir melhor a informação que se pretende transmitir no que respeita a evolução ao longo do tempo da variável em estudo, assim como a comparação entre as diferentes coleções de dados.

Projeto 3 – Construção de pirâmides etárias.

A partir dos seguintes dados obtidos em www.pordata.pt construa a pirâmide etária para a população residente do sexo masculino, para comparar os anos de 1960 a 2001.

População residente do sexo masculino, segundo os Censos: total e por grupo etário

Indivíduo

		Grupos etários								
Tempo	Total	0-04	05-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44
1960	4254416	461961	433899	423614	366103	336672	324364	305427	284660	239697
1970	4089165	402170	432445	410865	355490	297945	241340	250355	262665	261040
1981	4737715	404788	439771	435169	433655	385806	337171	307631	268962	273274
1991	4756775	278679	331337	398620	428240	386651	359556	340986	321775	307655
2001	5000141	275969	275199	296385	351422	400087	409243	379363	378783	357528

45-49	50-54	55-59	60-64	65-69	70-74	75+
243551	224227	184394	145362	111672	83660	85153
242785	209280	206185	184055	140065	94250	98230
278017	268382	249183	199108	182049	139169	135580
271665	265623	263265	245150	211990	149226	196357
333382	309484	268899	256179	244230	196615	267373

Fonte de Dados:
INE - X a XIV Recenseamentos Gerais da População

Fonte: PORDATA
Última actualização: 2011-12-23 11:37:17

Da tabela anterior vamos selecionar os dados referentes a 1960 e 2001, que colocamos numa folha de Excel, como se apresenta a seguir:

	A	B	C
1	Idade	Ano 1960	Ano 2001
2	0-04	461961	275969
3	05-09	433899	275199
4	10-14	423614	296385
5	15-19	366103	351422
6	20-24	336672	400087
7	25-29	324364	409243
8	30-34	305427	379363
9	35-39	284660	378783
10	40-44	239697	357528

Sheet1

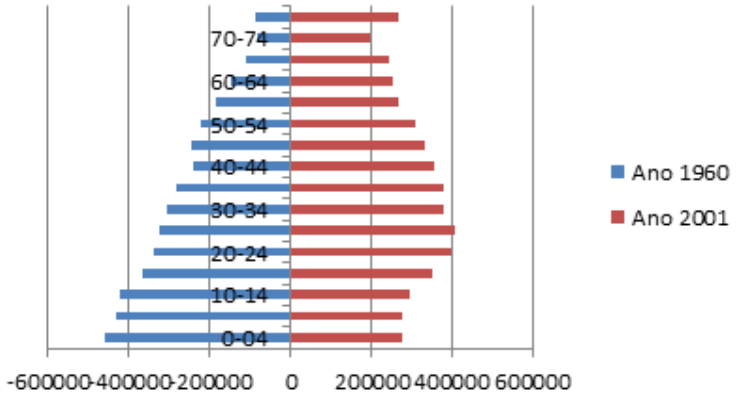
Na pirâmide etária que se vai construir, vamos colocar do lado esquerdo os dados referentes a 1960 e do lado direito os referentes a 2001 (Nota – Numa pirâmide etária para comparação dos sexos masculino e feminino segundo as classes etárias, considera-se o sexo masculino do lado esquerdo e o feminino do lado direito).
Para a construção da pirâmide etária, seguir os seguintes passos:

1. Multiplicar por -1 os dados referentes a 1960;

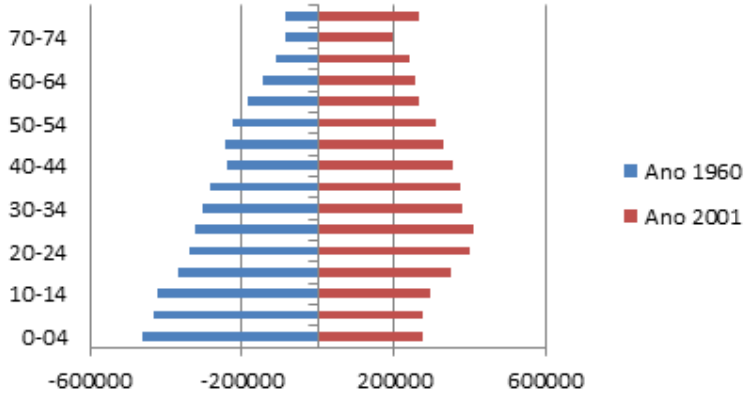
	A	B	C
1	Idade	Ano 1960	Ano 2001
2	0-04	-461961	275969
3	05-09	-433899	275199
4	10-14	-423614	296385
5	15-19	-366103	351422
6	20-24	-336672	400087
7	25-29	-324364	409243
8	30-34	-305427	379363
9	35-39	-284660	378783
10	40-44	-239697	357528

Sheet1

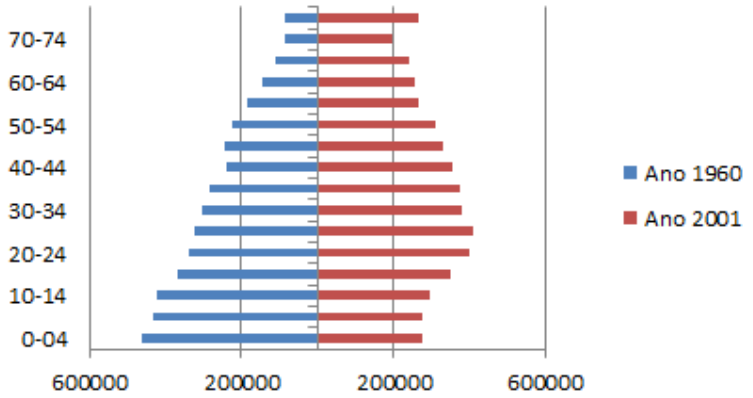
2. Seleccionar as células que contêm os dados e os títulos A1:C17 e fazer *Inserir→Bar→Stacked Bar*



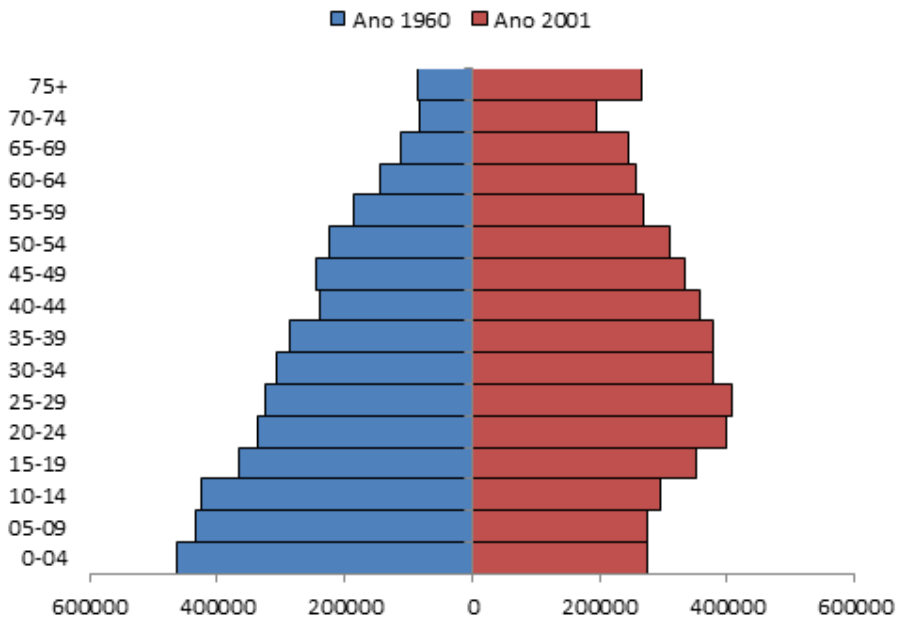
3. Para mover o eixo vertical para a esquerda, com o gráfico seleccionado, fazer *Layout→Axes→Primary Vertical Axes→More Primary Vertical Axes Options→Axis Labels→Low→Close*



4. Para apresentar os números referentes a 1960 positivos, em vez de nagativos, com o gráfico seleccionado, fazer *Layout→Axes→Primary Horizontal Axes→More Primary Horizontal Axes Options→Number→Format code: 0;0→Add→Close*



5. Basta agora juntar as barras e colocar a legenda em cima para obter a pirâmide etária com o seguinte aspecto:



Na representação anterior é nítido o envelhecimento da população residente do sexo masculino. Repare-se que, não só o número de nascimentos tem diminuído, como também a longevidade tem aumentado.

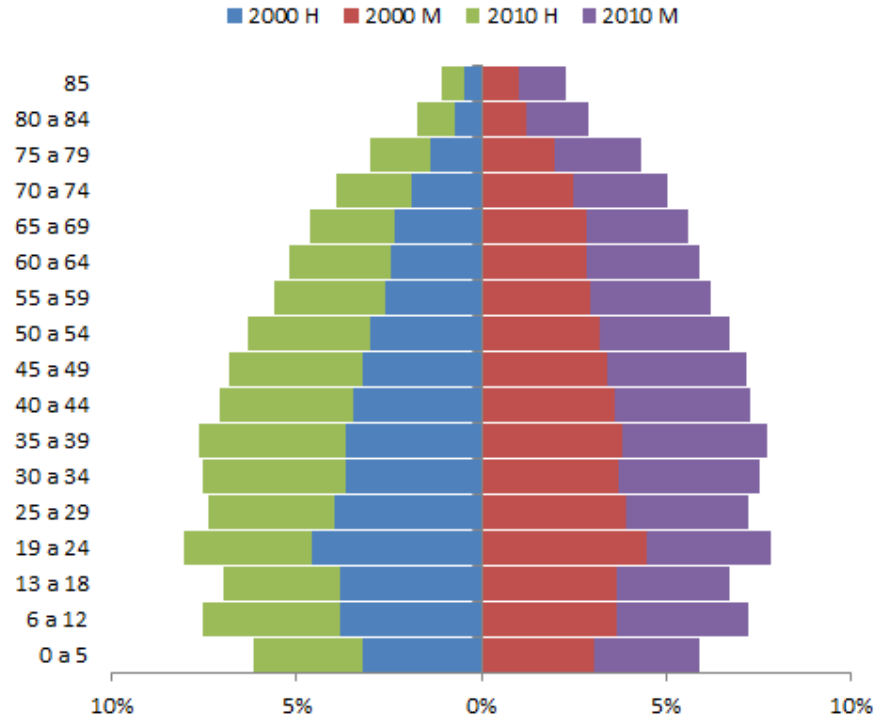
Projecto 4 – Considere os seguintes dados referentes às estimativas anuais da população residente e retirados do relatório *Estado da Educação 2011 – A Qualificação dos Portugueses*, Conselho Nacional de Educação,

www.epatv.pt/.../anolectivo20112012/estado_da_educacao_2011.pdf

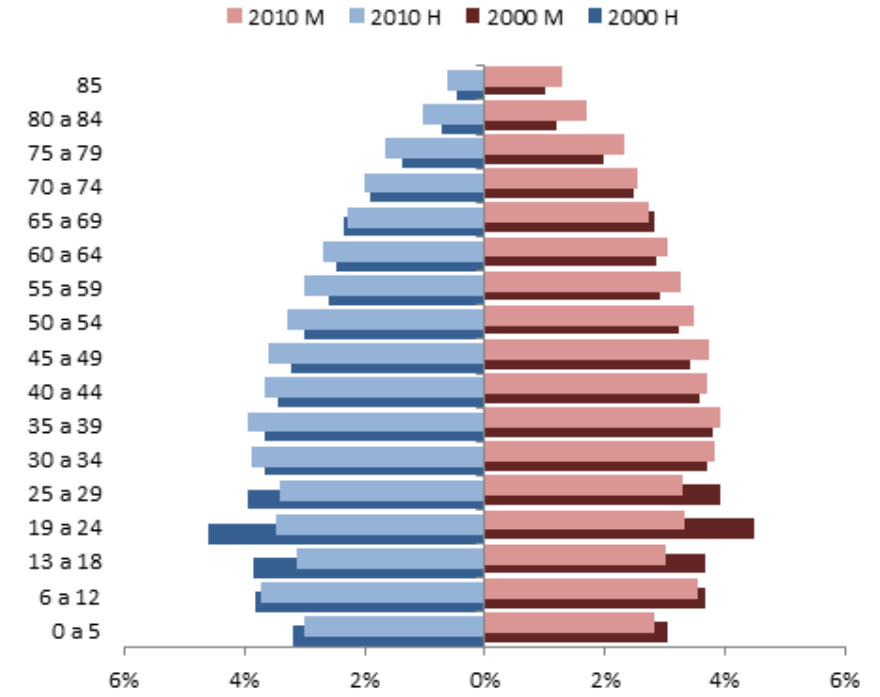
	Ano 2000		Ano 2010	
	Homens	Mulheres	Homens	Mulheres
0 a 5	3,19%	3,04%	2,99%	2,84%
6 a 12	3,82%	3,66%	3,73%	3,54%
13 a 18	3,84%	3,67%	3,14%	3,02%
19 a 24	4,59%	4,48%	3,46%	3,33%
25 a 29	3,95%	3,92%	3,41%	3,31%
30 a 34	3,66%	3,69%	3,88%	3,82%
35 a 39	3,66%	3,79%	3,96%	3,94%
40 a 44	3,45%	3,58%	3,65%	3,70%
45 a 49	3,22%	3,41%	3,59%	3,74%
50 a 54	2,99%	3,22%	3,30%	3,48%
55 a 59	2,60%	2,92%	3,00%	3,27%
60 a 64	2,47%	2,85%	2,70%	3,05%
65 a 69	2,36%	2,84%	2,27%	2,73%
70 a 74	1,90%	2,49%	2,01%	2,54%
75 a 79	1,38%	1,98%	1,65%	2,32%
80 a 84	0,73%	1,21%	1,02%	1,70%
>=85	0,46%	1%	0,63%	1,29%

Nota: escalões entre os 0 e os 24 anos foram divididos de acordo com as idades correspondentes aos níveis de escolaridade

Para comparar a população residente construiu-se a seguinte pirâmide etária, seguindo o processo descrito anteriormente:



Uma alternativa à pirâmide anterior obtém-se considerando *Clustered Bar* em vez de *Stacked Bar*:



3. Características amostrais. Medidas de localização e dispersão

3.1- Introdução

No módulo de Estatística foram apresentadas as medidas ou estatísticas que se utilizam para resumir a informação contida nos dados. Destas medidas, 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.

Neste capítulo não nos debruçaremos sobre as propriedades destas medidas, já apresentadas no módulo referido anteriormente, abordando sobretudo a forma de as calcular, utilizando o Excel. Convém desde já adiantar que este é um trabalho grandemente facilitado pelo facto de existirem funções no Excel que nos dão diretamente estas medidas.

Para facilidade de exposição vamos representar a amostra de dimensão n por

$$x_1, x_2, ..., x_n$$

onde x_1, x_2, ..., x_n representam, respetivamente, os resultados da 1ª observação, da 2ª observação, da n-ésima observação, a serem recolhidas, não pressupondo qualquer ordenação.

3.2 – Medidas de localização

Como medidas de localização, vamos apresentar a média, mediana e quartis.

3.2.1 – Média

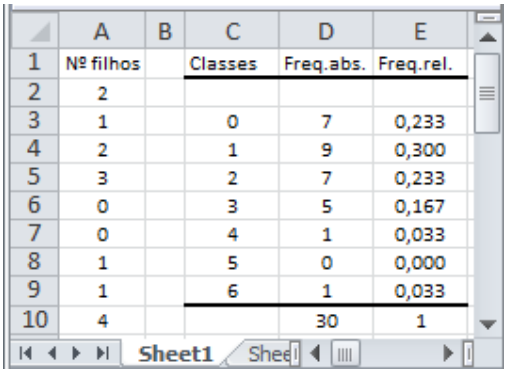
A média é uma medida de localização do centro da distribuição dos dados. Dada a amostra x_1, x_2, ..., x_n, a média representa-se por $\bar{x}$ e obtém-se adicionando todos os elementos e dividindo o resultado por n. Em Excel, determina-se a média através da função *AVERAGE ()*, que retorna a média aritmética dos seus argumentos, que podem ser números ou endereços de células.

Exemplo 3.2.1 – Retomemos a amostra do exemplo 2.3.2, constituída pelo número de filhos de 30 deputados:

2, 1, 2, 3, 0, 0, 1, 1, 4, 1, 2, 1, 0, 0, 0, 2, 3, 1, 1, 6, 3, 1, 3, 2, 0, 1, 2, 0, 2, 3

Calcule a média da amostra.

Considerámos o ficheiro *Filhos*, construído no exemplo 2.3.2, em que os elementos de que se pretende calcular a média ocupam as células A2 a A31:



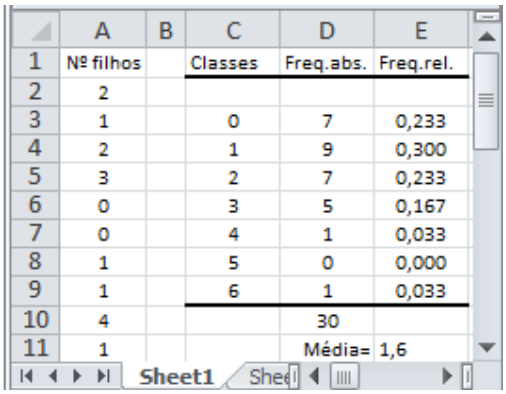
	A	B	C	D	E
1	Nº filhos		Classes	Freq.abs.	Freq.rel.
2	2				
3	1		0	7	0,233
4	2		1	9	0,300
5	3		2	7	0,233
6	0		3	5	0,167
7	0		4	1	0,033
8	1		5	0	0,000
9	1		6	1	0,033
10	4			30	1

Sheet1 Shee

Para calcular a média pretendida, assim como para qualquer outro conjunto de dados de tipo discreto, podemos proceder de dois modos, quer considerando os dados originais, quer agrupados.

1- Cálculo da média, a partir dos dados originais, utilizando a função *AVERAGE()*:

Colocar o cursor na célula onde se pretende colocar a média, por exemplo a célula E11, e inserir a função AVERAGE(A2:A31) – os argumentos desta função são os endereços onde estão os elementos da amostra. Como resultado obtém-se o valor 1,6, que se apresenta na figura seguinte.

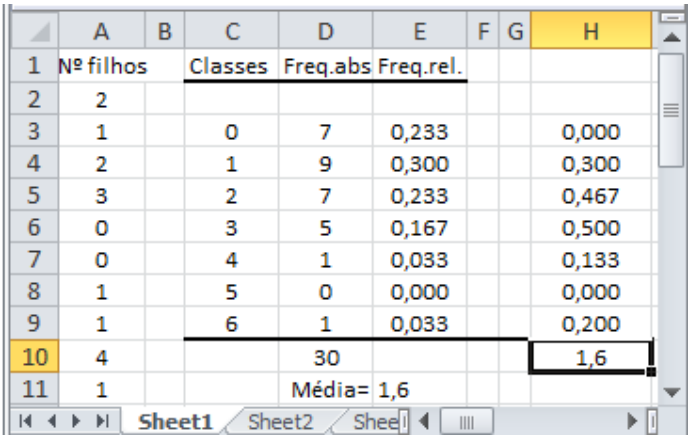


	A	B	C	D	E
1	Nº filhos		Classes	Freq.abs.	Freq.rel.
2	2				
3	1		0	7	0,233
4	2		1	9	0,300
5	3		2	7	0,233
6	0		3	5	0,167
7	0		4	1	0,033
8	1		5	0	0,000
9	1		6	1	0,033
10	4			30	
11	1			Média=	1,6

Sheet1 Shee

2- Cálculo da média, a partir dos dados agrupados:

Adicionar à tabela de frequências uma nova coluna com o produto dos valores que constituem as classes, pelas respetivas frequências relativas (Células H3 a H9) e somar os valores obtidos (Célula H10):



	A	B	C	D	E	F	G	H
1	Nº filhos		Classes	Freq.abs	Freq.rel.			
2	2							
3	1		0	7	0,233			0,000
4	2		1	9	0,300			0,300
5	3		2	7	0,233			0,467
6	0		3	5	0,167			0,500
7	0		4	1	0,033			0,133
8	1		5	0	0,000			0,000
9	1		6	1	0,033			0,200
10	4			30				1,6
11	1			Média=	1,6			

Sheet1 Sheet2 Shee

No caso de dados discretos, como é o caso anterior, o valor da média é o mesmo, quer seja calculada utilizando os dados originais, quer os dados agrupados (utilizando as frequências relativas), em que as classes do agrupamento são os diferentes valores que surgem na amostra. O mesmo não acontece no caso de dados contínuos, como exemplificamos a seguir.

Exemplo 3.2.2 – Calcule a média das idades dos deputados do ficheiro *DeputadosXII*.

Para obter a média das idades procede-se como no primeiro caso do exemplo anterior, a partir dos dados originais. Estes dados encontram-se nas células B2 a B231 pelo que inserindo a função AVERAGE(B2:B231) numa célula, obtemos o valor de 46,24 anos.

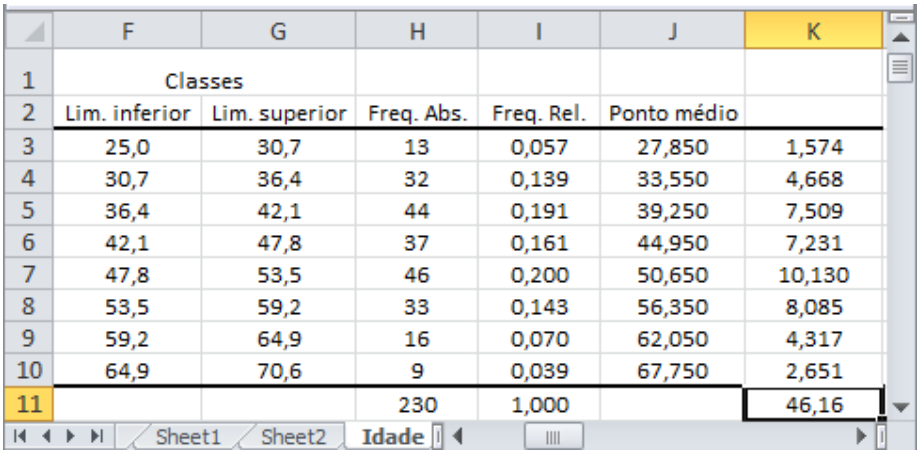
Admitindo que não dispúnhamos dos dados originais, mas apenas de uma tabela de frequências com os dados agrupados, vejamos como obter um valor aproximado para a média.

Reportando-nos ainda ao ficheiro Idade, consideremos a tabela de frequências que serviu para agrupar os dados. Para obter um valor aproximado para a média, procedemos da seguinte forma:

i) Adicionar à tabela de frequências uma nova coluna com os pontos médios dos intervalos de classe, que se obtêm fazendo a semissoma dos limites dos intervalos (células J3 a J10);

ii) Adicionar à tabela uma nova coluna com os produtos dos pontos médios dos intervalos de classe, pelas frequências relativas respetivas (células K3 a K10);

iii) Somar os resultados das células K3 a K10 (célula K11):



	F	G	H	I	J	K
1	Classes					
2	Lim. inferior	Lim. superior	Freq. Abs.	Freq. Rel.	Ponto médio	
3	25,0	30,7	13	0,057	27,850	1,574
4	30,7	36,4	32	0,139	33,550	4,668
5	36,4	42,1	44	0,191	39,250	7,509
6	42,1	47,8	37	0,161	44,950	7,231
7	47,8	53,5	46	0,200	50,650	10,130
8	53,5	59,2	33	0,143	56,350	8,085
9	59,2	64,9	16	0,070	62,050	4,317
10	64,9	70,6	9	0,039	67,750	2,651
11			230	1,000		46,16

Sheet1 Sheet2 Idade

Repare-se que o valor obtido de 46,16 para a média, é muito próximo do verdadeiro valor (=46,24) obtido com os dados originais. Note-se que o valor aproximado obtido para a média depende da escolha das classes feita para agrupar os dados. Outro conjunto de classes daria outro valor diferente para o valor aproximado da média.

3.2.2 – Mediana

Outra medida de localização do centro da distribuição dos dados é a mediana. Ordenados os elementos da amostra, a mediana, **m**, é o valor (pertencente ou não à amostra) que a divide ao meio, isto é, 50% dos elementos da amostra são menores ou iguais a m e os restantes 50% são maiores ou iguais a m. Em Excel, determina-se a mediana através da função *MEDIAN()*, que retorna a mediana dos seus argumentos, que podem ser números ou endereços de células.

Exemplo 3.2.3 – Calcule a mediana das idades dos deputados. Compare com o valor obtido para a média e diga o que poderia concluir da forma como os dados se distribuem.

Voltando ao ficheiro Idade, utilizado no exemplo anterior, inserindo numa célula a função *Median(B2:B231)* obtém-se, como retorno, o valor 46.
O valor obtido para a mediana é aproximadamente igual ao da média, pelo que podemos admitir que a distribuição é aproximadamente simétrica.

Se os dados se apresentarem agrupados, já vimos na secção 3.2.2 do capítulo 2, um processo de obter a mediana através da função cumulativa. No entanto, não é necessário construir esta função para obter um valor aproximado para a mediana, pois este pode ser obtido a partir da tabela de frequências, utilizando ainda o processo de interpolação.

Exemplo 3.2.4 – A partir do agrupamento considerado, no exemplo 2.3.3, para a variável idade, calcule um valor aproximado para a mediana.

Adicionando à tabela de frequências uma nova coluna com as frequências relativas acumuladas, verificamos que a mediana se encontra na classe [42,1; 47,8[, pois a frequência acumulada de 50% é atingida nesta classe:

	F	G	H	I	J
13	Classes	Freq. Rel/5,7	Freq. Abs.	Freq. Rel.	Freq.rel.acum.
14					
15	[25,0;30,7[	0,0099	13	0,057	0,057
16	[30,7;36,4[	0,0244	32	0,139	0,196
17	[36,4;42,1[	0,0336	44	0,191	0,387
18	[42,1;47,8[	0,0282	37	0,161	0,548
19	[47,8;53,5[	0,0351	46	0,200	0,748
20	[53,5;59,2[	0,0252	33	0,143	0,891
21	[59,2;64,9[	0,0122	16	0,070	0,961
22	[64,9;70,6[	0,0069	9	0,039	1,000
23			230	1	

Sheet1 Sheet2

Admitindo que a frequência se distribui uniformemente sobre a amplitude de classe, isto é, a frequência 0,161 se distribui uniformemente sobre o intervalo de amplitude 5,7, resolvendo a equação de proporcionalidade

$$\frac{0{,}161}{0{,}113}=\frac{5{,}7}{x} \qquad x=\frac{0{,}113\times 5{,}7}{0{,}161}=4{,}35$$

onde 0,113=0,5-0,387, obtemos para a mediana o valor aproximado 42,1 + 4,35 = 46,45.

Chamamos a atenção para o seguinte facto: tal como já acontecia para a média, o valor (aproximado) que se obtém para a mediana, depende do agrupamento que se fizer para os dados, pelo que agrupamentos diferentes darão origem a valores diferentes, embora não difiram muito uns dos outros (Lembramos que o valor da mediana apresentado anteriormente de 46 anos foi obtido a partir dos dados não agrupados).

3.2.3 – Quartis

Os quartis, 1º e 3º, definem-se de forma idêntica à mediana, mas considerando em vez da percentagem de 50%, respetivamente 25% para o 1º quartil, Q1, e 75% para o 3º quartil, Q3.

Há vários processos para a determinação dos quartis, nem sempre conduzindo aos mesmos resultados. Este facto não é preocupante, pois de um modo geral nas situações que têm interesse em estatística, as amostras têm dimensão suficientemente elevada de forma que os diferentes processos conduzem a valores próximos.

Em Excel a determinação dos quartis faz-se utilizando a função *QUARTILE(array;quart)*:

Repare que a função *Quartile(array;quart)* tem dois argumentos, em que o primeiro argumento é o endereço das células de que queremos calcular o quartil e o segundo argumento pode tomar vários valores, conforme a medida de localização, de entre as seguintes, que nos interesse calcular:

0 – mínimo
1 – 1º quartil
2 – mediana
3 – 3º quartil
4 – máximo

Assim, esta função, além do 1º e 3º quartis, a que estão associadas as percentagens 25% e 75%, respetivamente, ainda calcula a mediana, a que está associada a percentagem de 50% e o mínimo e máximo com percentagens associadas de 0% e 100%.

Exemplo 3.2.5 – Escolha os primeiros 15 elementos da variável Idade, do ficheiro *Idade*. Obtenha o 1º e 3º quartis.

Os primeiros 15 elementos são os seguintes:

48 52 54 34 46 47 67 64 66 42 34 38 36 27 49

Utilizando a função *QUARTILE(B2:B16;1)* e *QUARTILE(B2:B16;3)*, obtemos Q_1=37 e Q_3=53.

Se utilizar o processo que aprendeu no módulo de Estatística, nomeadamente considerando o 1º quartil como a mediana da primeira parte da amostra, quando esta é dividida pela mediana, depois de ordenar a amostra e tendo em conta que a mediana é 47, temos para 1º quartil o

27 34 34 36 38 42 46 **47** 48 49 52 54 64 66 67

valor 36, se não considerarmos a mediana como pertencente a nenhuma das partes, ou 37 se considerarmos a mediana pertencente às duas partes. Para o 3º quartil obteremos, respetivamente o valor 54 ou 53, utilizando a mesma metodologia.

Exemplo 3.2.5 (cont) – Repita o exemplo anterior, considerando amostras de dimensão 12 e 13.

Considere agora só os primeiros 12 elementos. Como a mediana é 44, o 1º quartil – mediana da 1ª parte da amostra, será (34+36)/2=35, enquanto que o 3º quartil será (48+49)/2=48,5.

27 34 34 36 38 42 46 47 48 49 52 54

Utilizando o Excel, os valores que se obtêm são Q_1=35,5 e Q_3=48,25.

Considere agora os primeiros 13 elementos. Como a mediana é 46, o 1º quartil – mediana da 1ª parte da amostra, será (34+36)/2=35, enquanto que o 3º quartil será (49+52)/2=51,5, não considerando a mediana como pertencente a nenhuma das partes. Caso contrário, teremos Q_1=36 e Q_3=49.

27 34 34 36 38 42 **46** 47 48 49 52 54 64

Utilizando o Excel, os valores que se obtêm são Q_1=36 e Q_3=49.

Obervação: Repare que os valores que se obtêm para os quartis, recorrendo ao Excel não são iguais aos que se obtiveram sem utilizar o Excel. Efetivamente não existe uniformidade na forma de calcular os quartis, como já havíamos referido anteriormente, embora os resultados obtidos satisfaçam a definição de quartis. Exemplificando com a mediana, repare que pela definição de mediana, quando o número de elementos da amostra é par, podemos considerar para mediana qualquer valor compreendido entre os dois elementos médios da amostra ordenada! Não é costume deixar esta opção ao critério de cada um e considera-se a semissoma desses elementos médios.

Voltando aos quartis, pode verificar que, no Excel, o 1º quartil corresponde à observação de ordem (n+3)/4, procedendo-se a uma interpolação, quando necessário (Sugestão – Tente descobrir como é calculado o 3º quartil no Excel).

3.3 – Medidas de dispersão

Continuando na mesma linha de apresentação das medidas de localização, também agora não nos vamos preocupar com as propriedades das medidas de dispersão, pois admitimos que estas já foram estudadas no módulo de Estatística. Debruçar-nos-emos sobre o seu cálculo, utilizando o Excel.

A seguir apresentaremos o cálculo da variância, desvio padrão e amplitude interquartil.

3.3.1 – Variância e desvio-padrão

A variância de um conjunto de dados obtém-se fazendo a média dos quadrados dos desvios dos dados, relativamente à média.

O Excel, tal como as máquinas de calcular, dispõe de duas funções para calcular a variância, conforme estejamos a calcular a variância populacional (parâmetro) ou a variância amostral (estatística).

Resumimos no quadro seguinte a situação de estarmos a calcular parâmetros ou estatísticas.

População de **N** elementos	**Amostra** de **n** elementos
$x_1, x_2, \ldots, x_N$	$x_1, x_2, \ldots, x_n$
Valor médio $\mu = \frac{x_1 + x_2 + \ldots + x_N}{N}$	**Média** $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$
Variância populacional $\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_N - \mu)^2}{N}$	**Variância** amostral $s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$
Desvio padrão populacional σ	Desvio padrão amostral s

Em Excel as funções utilizadas para calcular a variância populacional e amostral, são respetivamente *VAR.P()* e *VAR.S()*. Como argumento utiliza-se a sequência de números de que se quer calcular a variância, ou o endereço das células que os contêm.

Por exemplo, no caso da população dos deputados, que temos vindo a estudar, temos informação completa sobre a variável Idade, pelo que a fórmula que deve ser utilizada para obter a variância é a VAR.P, isto é, esta fórmula dá-nos a variância populacional. Se só dispuséssemos da idade de alguns deputados, isto é, uma amostra da população em estudo, então a fórmula a utilizar seria a VAR.S, que dá a variância amostral. A maneira de calcular as duas variâncias é idêntica, diferindo unicamente no seguinte ponto: enquanto que no caso da variância populacional se divide a soma dos quadrados dos desvios pelo número de parcelas, no caso da variância amostral divide-se a soma dos quadrados dos desvios pelo número de parcelas menos uma.

O desvio padrão obtém-se fazendo a raiz quadrada da variância ou utilizando uma função própria. Como é evidente, existem também duas fórmulas para o calcular, obtendo-se o desvio padrão populacional ou amostral, conforme a fórmula utilizada: STDEV.P() ou STDEV.S() respetivamente.

Exemplo 3.3.1 – A partir do ficheiro Idade, selecione uma amostra aleatória simples de dimensão 40. Calcule a variância e o desvio padrão da amostra obtida. Calcule de seguida a variância da população constituída pelas idades dos 230 deputados e compare com a variância da amostra obtida anteriormente.

Utilizando o processo descrito em 1.3.1.2, selecionámos uma amostra de 40 elementos que posteriormente colocámos nas células F2 a I11, de uma nova folha de Excel. Colocando agora o cursor na célula onde pretendemos colocar a variância, por exemplo na célula K5, inserimos a função *VAR.S(F2:I11)* e a função retorna um valor aproximadamente igual a 115, para a variância da amostra.

Para calcular a variância da população das idades, inserimos na célula J12 a função *VAR.P(B2:B231)*, obtendo-se um valor aproximadamente igual a 106:

	F	G	H	I	J	K
1		Amostra				
2	28	43	51	31		
3	41	47	42	44		
4	26	46	62	30		
5	52	47	53	37	Var. amostral=	114,77
6	61	48	39	42		
7	40	29	40	31		
8	32	68	39	46		
9	70	59	46	50		
10	35	40	52	58		
11	38	38	50	40		
12					Var. populacional=	106,38

Sheet1 Shee

Comparando as variâncias, vemos que não são iguais, o que já seria de esperar, uma vez que a variância amostral foi obtida a partir de 40 dos 230 dados e é uma estimativa da variância populacional. Se recolhermos outra amostra, também de 40 elementos, não esperamos obter o mesmo valor para a estimativa. Esperamos sim, obter valores aproximados.

Para calcular o desvio padrão, ou se calcula a raiz quadrada (positiva) do valor da variância, ou se utilizam as funções *STDEV.S()* ou *STDEV.P()*, conforme se pretenda o desvio padrão amostral ou populacional. No nosso caso os desvios padrões amostral e populacional vêm, respetivamente, aproximadamente iguais a 10,7 e 10,3.

3.3.2 – Amplitude e amplitude interquartil

A amplitude da amostra (não confundir com dimensão da amostra), R, é a medida mais simples para medir a variabilidade, mas tem a grande desvantagem de ser muito sensível à existência na amostra, de uma observação muito pequena ou muito grande. Não existe, no Excel, uma função específica para a calcular, recorrendo-se às funções *MAX()* e *MIN()*. Já tivemos, aliás, oportunidade de utilizar estas funções quando necessitámos de calcular a amplitude de um conjunto de dados, para obter a tabela de frequências, com o objetivo de construir um histograma, com classes de igual amplitude.

Uma medida mais resistente do que a anterior, é a amplitude interquartil que, como o nome indica, se define como a diferença entre os 1º e 3º quartis.

Exemplo 3.3.2 – Calcule a amplitude e a amplitude interquartil da amostra obtida no exemplo anterior.

Como os elementos da amostra se encontram nas células F2 a I11, temos:

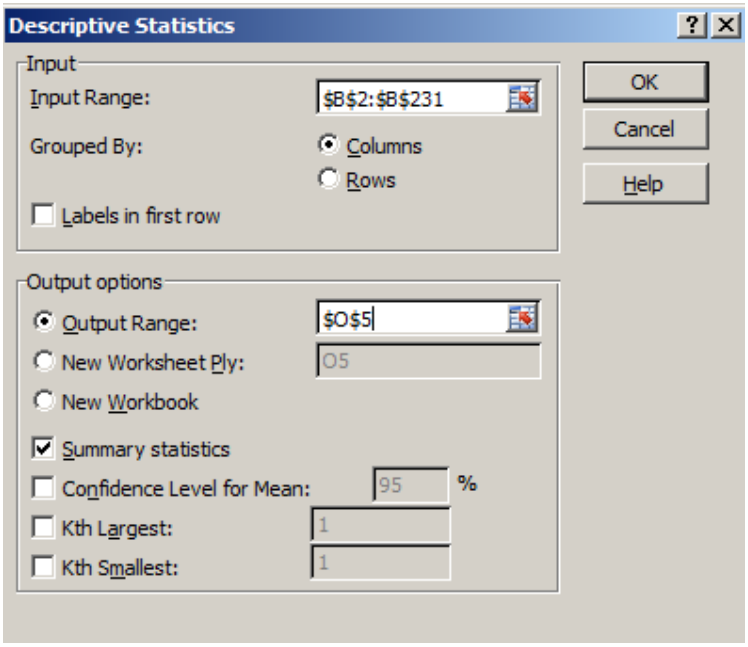
$$R = MAX(F2:I11) - MIN(F2:I11) = 44$$

Recorrendo à terminologia usada quando definimos os quartis, temos:

Amplitude interquartil= *QUARTILE(F2:I11;3) – QUARTILE(F2:I11;1)* = 12,25.

3.4 – Função Descriptive Statistics

O Excel dispõe de uma função a que se acede selecionando *Data →Data Analysis→Descriptive Statistics →OK*



cujo resultado é o que se apresenta a seguir para o ficheiro Idade:

	O	P
5	*Column1*	
6		
7	Mean	46,23913
8	Standard Error	0,681579
9	Median	46
10	Mode	46
11	Standard Deviation	10,33666
12	Sample Variance	106,8465
13	Kurtosis	-0,71646
14	Skewness	0,121479
15	Range	45
16	Minimum	25
17	Maximum	70
18	Sum	10635
19	Count	230

Algumas das funções já são conhecidas das secções anteriores. Chamamos a atenção para o facto de a variância das 230 idades não coincidir com o valor obtido na secção 3.3.1, uma vez que quando se considera um conjunto de dados e se pedem as Estatísticas descritivas, subentende-se que se está perante uma amostra e não da população toda! Por esta razão, a fórmula utilizada para o cálculo da variância é a da variância amostral.

As funções Standard Error, Kurtosis e Skewness saem fora do âmbito estas folhas, pelo que não entraremos em detalhe.

4. Dados bivariados

4.1- Introdução

No módulo de Estatística foi feita referência a dados bidimensionais, de tipo quantitativo. Quando dispomos de uma amostra de dados bivariados, a qual pode ser representada na

forma (x_1, y_1), (x_2, y_2),..., (x_n, y_n), apresentamos esta informação através de uma representação gráfica a que se dá o nome de Diagrama de dispersão:

Diagrama de dispersão – É uma representação gráfica para os dados bivariados, em que cada par de dados (x_i, y_i), é representado por um ponto de coordenadas (x_i, y_i), num sistema de eixos coordenados.

Já vimos no capítulo 2, a forma de representar, em Excel, dados bivariados, utilizando a opção *XY*(*Scatter)*. Não apresenta qualquer dificuldade a construção desta representação gráfica, uma vez que basta proceder da seguinte forma:

- Selecionar as células que contêm os dados, organizados em 2 colunas;
- Selecionar *Insert* →*Scatter* e o sub-tipo pretendido;
- Formatar convenientemente a representação obtida (retirar a legenda, retirar as linhas de grelha, etc).

Quando se trata de dados qualitativos, não tem sentido proceder à representação gráfica dos dados através de um diagrama de dispersão. No entanto, é possível organizar essa informação na forma de tabelas de contingência (que aliás também podem ser usadas para dados quantitativos, quer discretos, quer contínuos, depois de proceder à sua discretização, ou seja, organização em classes).

Vamos, neste capítulo, introduzir uma metodologia que utiliza uma ferramenta do Excel, a *PivotTable*, que além de permitir construir tabelas de contingência, também pode ser utilizada para proceder a agrupamentos de dados qualitativos ou quantitativos.

4.2 – Tabelas de contingência – A funcionalidade PivotTable

Suponhamos que estamos interessados em estudar a associação entre variáveis de tipo qualitativo como, por exemplo, sexo e religião. Uma forma de apresentar os dados é utilizando tabelas de contingência.

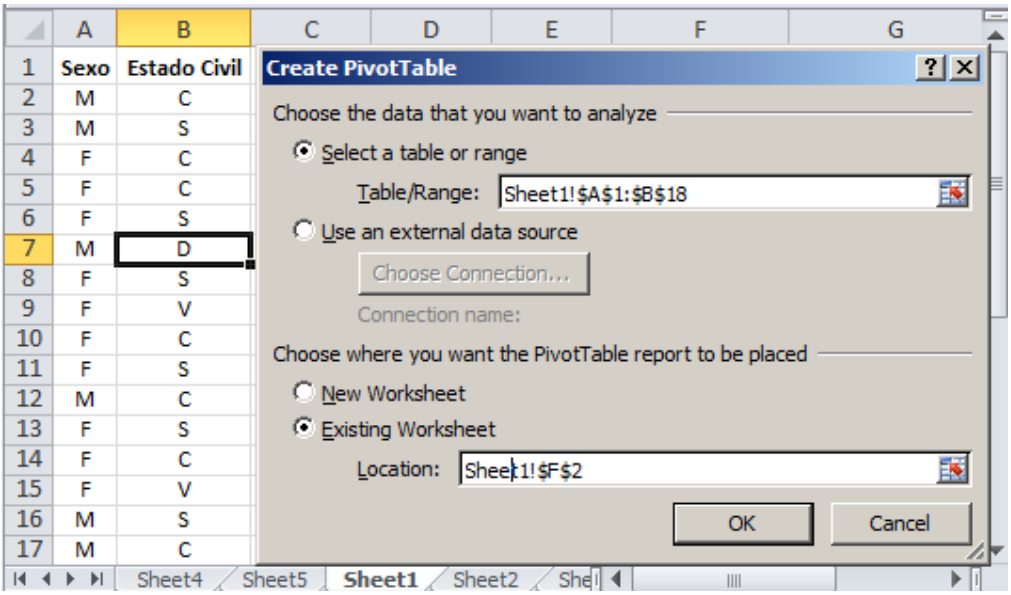
Exemplo 4.2.1 – Uma empresa decidiu estudar o seu pessoal quanto ao estado civil e sexo. Representando por M e F as categorias da variável Sexo, e por C (casado(a)), S (solteiro(a)), D (divorciado(a)) e V (viúvo(a)), obteve a seguinte lista: (M,C), (M,S), (F,C), (F,C), (F,S), (M,D), (F,S), (F,V), (F,C), (F,S), (M,C), (F,S), (F,C), (F,V), (M,S), (M,C), (F,S) (Este exemplo é fictício e serve unicamente para introduzir o estudo das tabelas de contingência, pois os casos interessantes em Estatística envolvem amostras de maior dimensão).

Começámos por introduzir estes dados numa folha de Excel, colocando nas células A1 e B1 os títulos, respetivamente Sexo e Estado Civil, e nas células A2 a A18 a informação sobre o sexo dos 17 elementos e nas células B2 a B18, o respetivo estado civil:

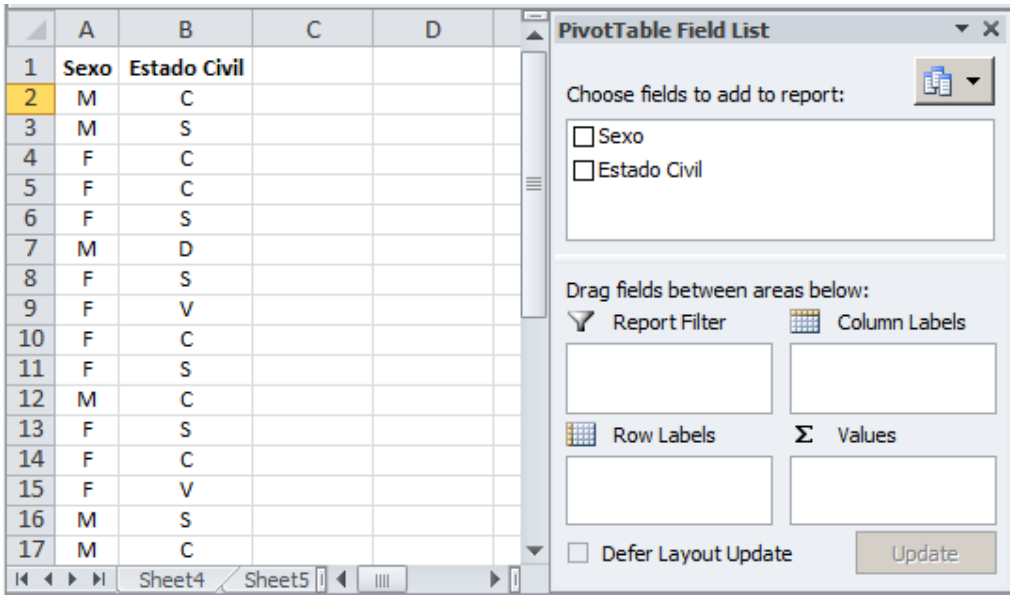
	A	B
1	Sexo	Estado Civil
2	M	C
3	M	S
4	F	C
5	F	C
6	F	S
7	M	D
8	F	S
9	F	V
10	F	C
11	F	S
12	M	C
13	F	S
14	F	C
15	F	V
16	M	S
17	M	C
18	F	S

Para criar uma tabela, utilizando a funcionalidade PivotTabbe, proceder do seguinte modo:

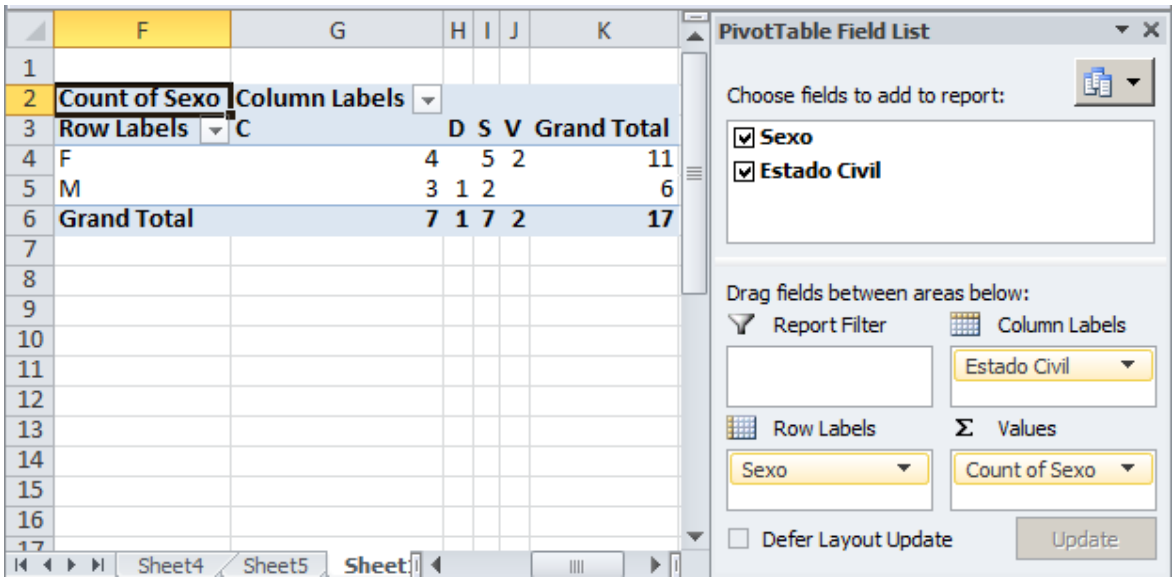
- Colocar o cursor em qualquer célula do interior da tabela e selecionar *Insert→PivotTable* e inserir a informação de acordo com a figura seguinte



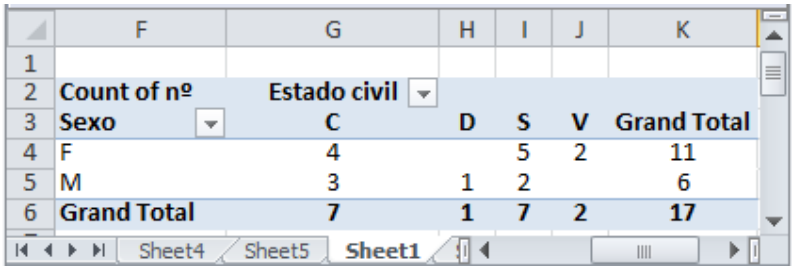
obtendo-se como resultado:



- Arrastar o botão Sexo da barra *PivotTable*, e colocá-lo no campo *Row Labels*; Arrastar o botão Estado civil da barra *PivotTable*, e colocá-lo no campo *Column Labels*; Arrastar qualquer dos botões anteriores e colocá-lo no campo *Values*:



Finalmente temos a tabela de contingência desejada, que nos dá a distribuição conjunta (em valores absolutos) do par (Sexo, Estado civil), permitindo obter o número de indivíduos que satisfazem simultaneamente cada uma das modalidades (feminino(a),casado(a)), (feminino(a),divorciado(a)), ...(masculino(a),viúvo(a)):



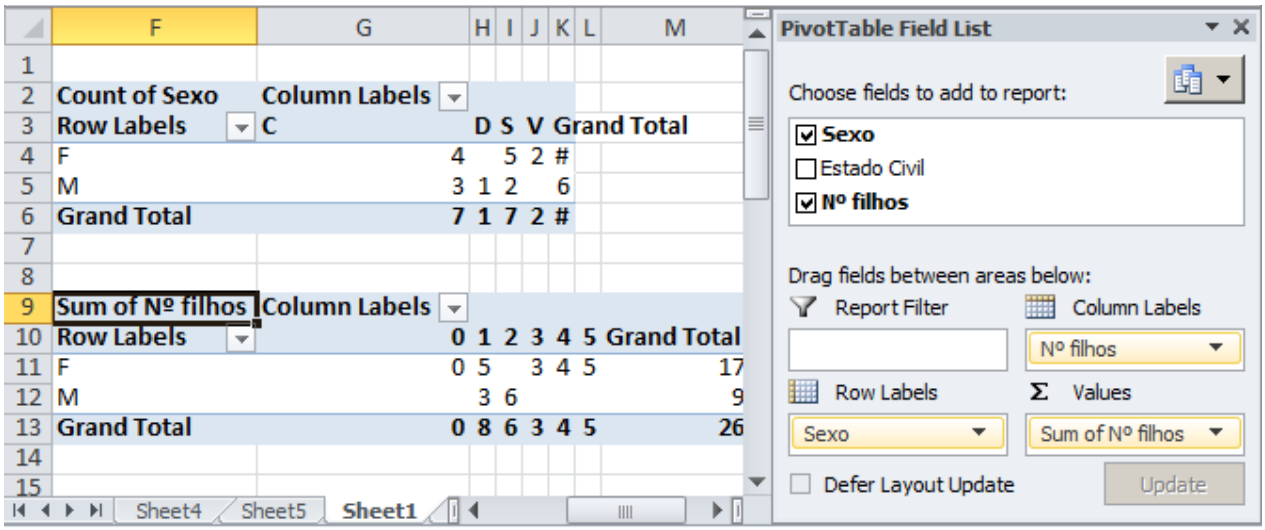
	F	G	H	I	J	K
1						
2	Count of nº	Estado civil				
3	Sexo	C	D	S	V	Grand Total
4	F	4		5	2	11
5	M	3	1	2		6
6	Grand Total	7	1	7	2	17

Sheet4 Sheet5 Sheet1

O facto da célula correspondente às categorias F e D estar vazia, significa que não havia indivíduos do sexo feminino e divorciados. Esta tabela apresenta ainda as distribuições marginais (em valores absolutos) da variável Sexo e Estado civil, respetivamente nas células K4 a K5 e G6 a J6. Efetivamente, através da tabela, pode-se concluir que o número de indivíduos do sexo feminino é 11, enquanto que do sexo masculino é 6. Analogamente, também podemos tirar conclusões sobre o número de indivíduos em cada modalidade da variável Estado civil.

Exemplo 4.2.1 (cont) - Suponhamos que ao recolher a informação, junto de cada indivíduo, sobre o seu estado civil, também se tinha investigado sobre o número de filhos (esta informação é relevante para o serviço de processamento de salários proceder à retenção do IRS). Construa uma tabela de contingência para o par (Sexo, Estado civil).

Inserimos a informação sobre a variável Nº de filhos, e procedemos à construção da tabela de contingência:



	F	G	H	I	J	K	L	M
1								
2	Count of Sexo	Column Labels						
3	Row Labels	C	D	S	V	#	Grand Total	
4	F	4		5	2	#		
5	M	3	1	2		6		
6	Grand Total	7	1	7	2	#		
7								
8								
9	Sum of Nº filhos	Column Labels						
10	Row Labels	0	1	2	3	4	5	Grand Total
11	F	0	5		3	4	5	17
12	M		3	6				9
13	Grand Total	0	8	6	3	4	5	26
14								
15								

Sheet4 Sheet5 Sheet1

PivotTable Field List

Choose fields to add to report:

- Sexo
- Estado Civil
- Nº filhos

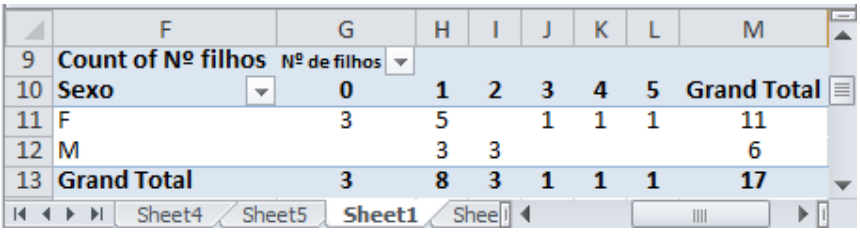
Drag fields between areas below:

Report Filter | Column Labels: Nº filhos

Row Labels: Sexo | Σ Values: Sum of Nº filhos

Defer Layout Update | Update

Esta tabela, que resulta das operações anteriores, não é a que nos interessa, uma vez que apresenta Sum of Nº filhos, em vez de Count of Nº filhos. É agora necessário clicar no campo *Sum of Nº* filhos e selecionar *Value Field Settings→Count.* Nesta 2ª tabela temos a distribuição conjunta do par (Sexo, Nº de filhos):



	F	G	H	I	J	K	L	M
9	Count of Nº filhos	Nº de filhos						
10	Sexo	0	1	2	3	4	5	Grand Total
11	F	3	5		1	1	1	11
12	M		3	3				6
13	Grand Total	3	8	3	1	1	1	17

Sheet4 Sheet5 Sheet1 Shee

Exemplo 4.2.1 (cont) – Proceda como no exemplo anterior, excepto no passo em que o botão da variável que arrasta para o campo Data, é o botão da variável Estado civil.

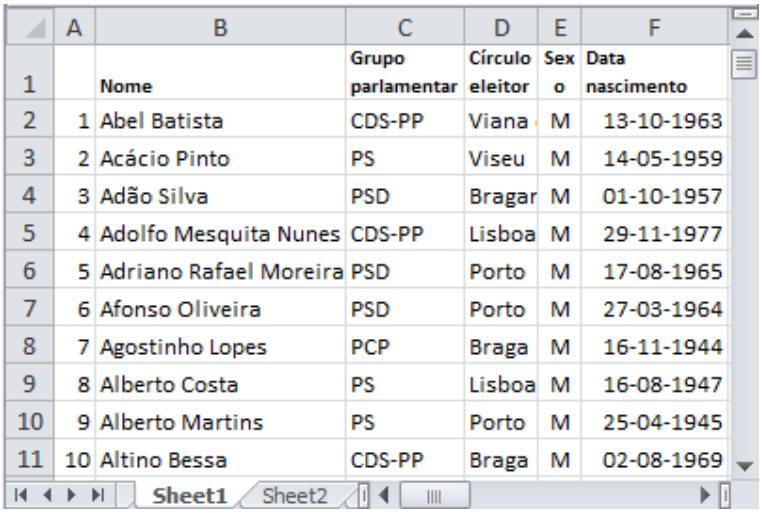
Com este procedimento obteríamos uma tabela igual à anterior, com as contagens, em vez das somas, já que *Count* é a opção que está selecionada, por defeito, quando colocamos no campo *Data* uma variável não numérica.

4.3 – Utilização das *PivotTables* para agrupar dados

Quando temos um conjunto de dados, já vimos no Capítulo 2 a forma de proceder ao seu agrupamento. Vamos agora ver, como essa tarefa pode ser feita através da funcionalidade *PivotTable*.

4.3.1 – Dados de tipo qualitativo

Vamos voltar ao ficheiro *DeputadosXII* (de que apresentamos a seguir uma pequena parte)



	A	B	C	D	E	F
1		Nome	Grupo parlamentar	Círculo eleitor	Sexo	Data nascimento
2	1	Abel Batista	CDS-PP	Viana	M	13-10-1963
3	2	Acácio Pinto	PS	Viseu	M	14-05-1959
4	3	Adão Silva	PSD	Bragar	M	01-10-1957
5	4	Adolfo Mesquita Nunes	CDS-PP	Lisboa	M	29-11-1977
6	5	Adriano Rafael Moreira	PSD	Porto	M	17-08-1965
7	6	Afonso Oliveira	PSD	Porto	M	27-03-1964
8	7	Agostinho Lopes	PCP	Braga	M	16-11-1944
9	8	Alberto Costa	PS	Lisboa	M	16-08-1947
10	9	Alberto Martins	PS	Porto	M	25-04-1945
11	10	Altino Bessa	CDS-PP	Braga	M	02-08-1969

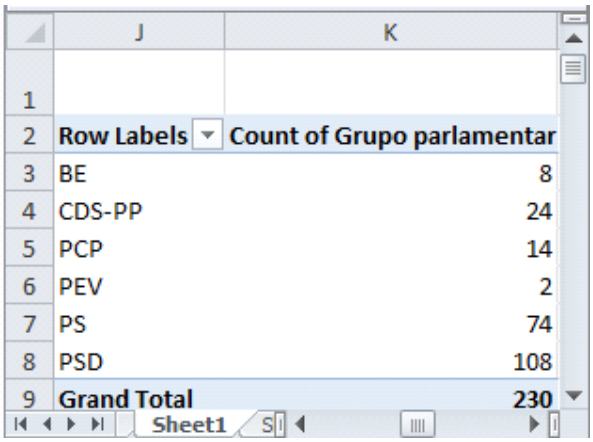
Sheet1 Sheet2

para exemplificar a construção de uma tabela de frequências de uma variável qualitativa, utilizando a *PivotTable*.

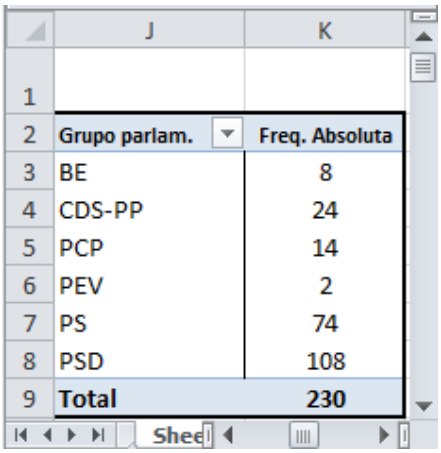
Exemplo 4.3.1 – Utilizando a *PivotTable*, proceda ao agrupamento de dados da variável Grupo parlamentar, do ficheiro *DeputadosXII*.

- Colocar o cursor num ponto qualquer da tabela;
- Selecionar o local onde se pretende inserir a tabela;

- Selecionar *Insert→PivotTable →PivotTable*
- Arrastar o botão Grupo parlamentar da barra *PivotTable*, e colocá-lo no campo *Row*; Arrastar o botão Grupo parlamentar e colocá-lo no campo *Values*:

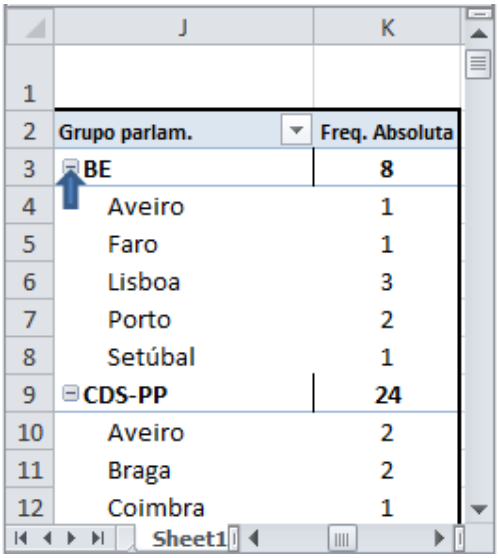


Row Labels	Count of Grupo parlamentar
BE	8
CDS-PP	24
PCP	14
PEV	2
PS	74
PSD	108
Grand Total	230



Grupo parlam.	Freq. Absoluta
BE	8
CDS-PP	24
PCP	14
PEV	2
PS	74
PSD	108
Total	230

O procedimento anterior conduziu-nos à tabela do lado esquerdo da figura anterior, que com algumas transformações estéticas apresenta o aspeto da tabela do lado direito. Se estivermos interessados em saber como se distribuem os deputados de cada grupo parlamentar, pelos diferentes círculos eleitorais, basta arrastar o botão Círculo eleitoral para o campo *Row*:



Grupo parlam.	Freq. Absoluta
BE	8
Aveiro	1
Faro	1
Lisboa	3
Porto	2
Setúbal	1
CDS-PP	24
Aveiro	2
Braga	2
Coimbra	1

Apresentamos só parte da tabela devido à sua extensão. Aqui é possível ver como se distribuem os deputados do BE.

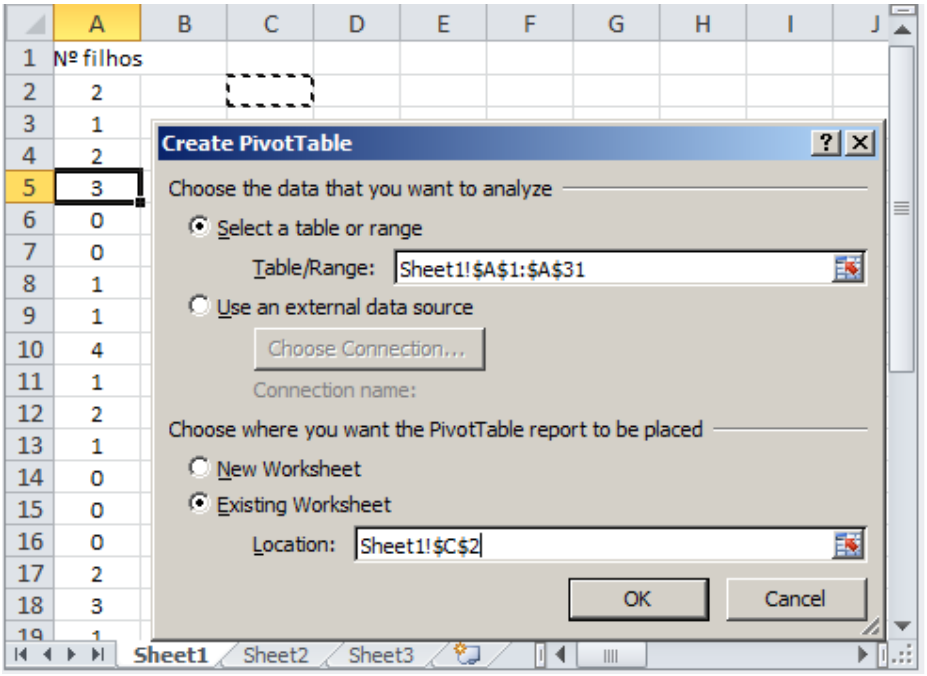
Se carregar com o cursor no sítio indicado pela seta, volta à tabela inicial unicamente com os totais por grupo parlamentar.

4.3.2 – Dados de tipo discreto

A organização de dados discretos numa tabela de frequências, utilizando a *PivotTable*, faz-se do mesmo modo que para os dados de tipo qualitativo. Vamos exemplificar procedendo ao agrupamento da variável Nº de filhos dos dados do ficheiro *Filhos*.

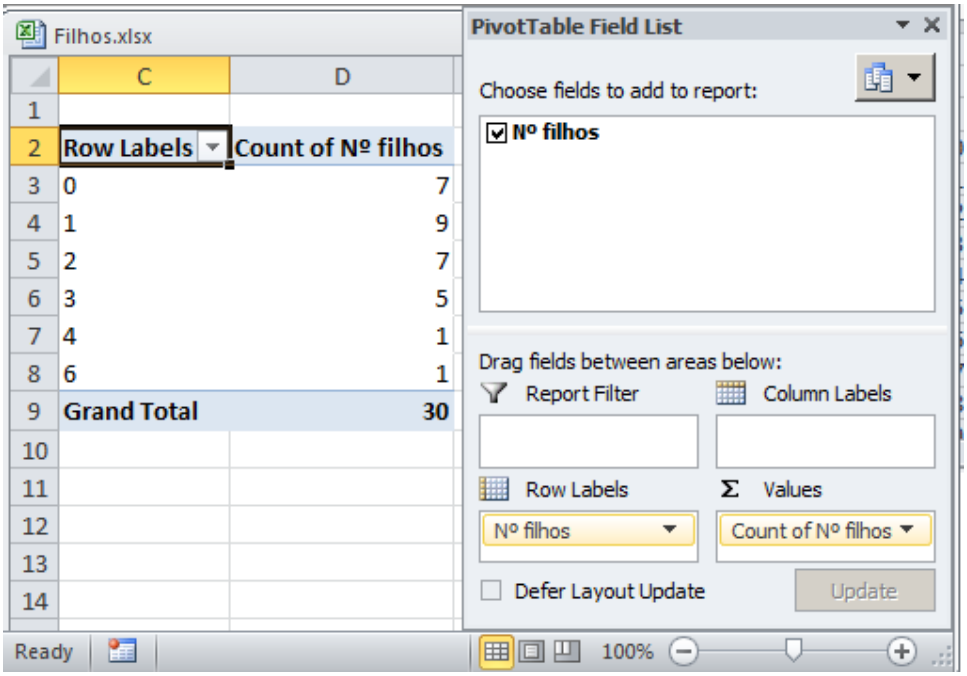
Exemplo 4.3.2 - Utilizando a *PivotTable*, proceda ao agrupamento de dados da variável Nº de filhos, do ficheiro *Filhos*.

- Colocar o cursor em alguma célula da tabela (colocámos na célula A5), e selecionar *Inserir→PivotTable→PivotTable* e proceder de acordo com a figura seguinte:



- Arrastar o botão Nº filhos da barra *PivotTable*, e colocá-lo no campo *Row*; Arrastar o mesmo botão e colocá-lo no campo *Values*;
- Clicar no campo *Sum of Nº* filhos e selecionar *Value Field Settings→Count*

O resultado destas operações é apresentado na figura seguinte:



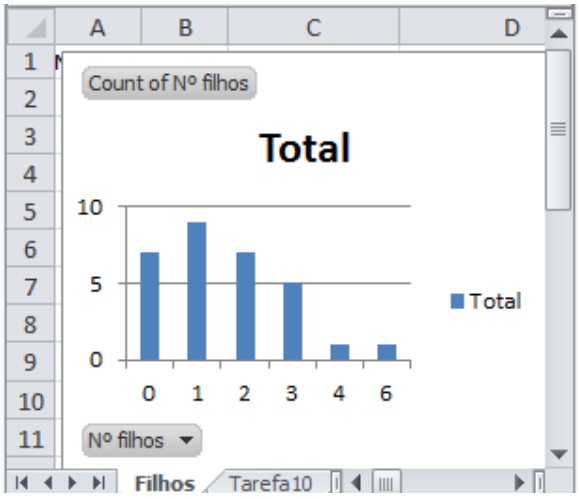
	C	D
1		
2	Row Labels	Count of Nº filhos
3	0	7
4	1	9
5	2	7
6	3	5
7	4	1
8	6	1
9	Grand Total	30
10		
11		
12		
13		
14		

A tabela anterior pode ser modificada de modo a ficar com um aspeto mais usual:

	C	D
1		
2	Nº filhos	Freq. absol.
3	0	7
4	1	9
5	2	7
6	3	5
7	4	1
8	6	1
9	Total	30

Sheet1

Para construir o diagrama de barras associado à tabela anterior basta, com a tabela selecionada, selecionar na barra menu *Options→PivotChart→Column*, obtendo-se a seguinte representação:



Atenção! O gráfico anterior, em que se procura representar num diagrama de barras a distribuição dos dados referente à variável Nº de filhos, não está correto! Num diagrama ou gráfico de barras referente a dados discretos, devem ser consideradas como classes todos os valores entre o mínimo e o máximo, mesmo que tenham frequência nula. Neste caso falta a classe referente ao número 5, pelo que uma solução é copiar os valores dados pela tabela para outras células do Excel, inserir a classe 5 com frequência nula e construir o diagrama de barras como se viu na secção 2.3.1.2.

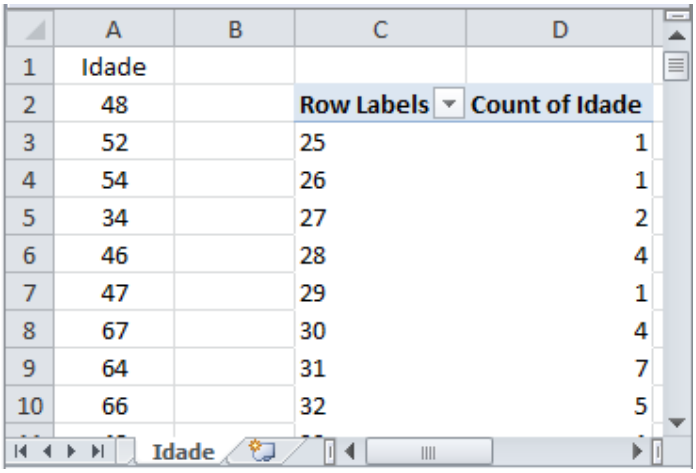
4.3.3 – Dados de tipo contínuo

Vamos exemplificar o agrupamento de uma variável de tipo contínuo, utilizando a *PivotTable*, mas avisamos desde já que, se os dados não forem inteiros, o processo tem de ser utilizado com as devidas precauções. Começaremos por abordar a situação de termos uma variável contínua, mas em que os dados são inteiros.

1ª parte – Dados em formato de inteiro

Exemplo 4.3.3 – Considere o ficheiro Idade, que contém a idade de 230 deputados. Proceda ao agrupamento em classes, utilizando a funcionalidade *PivotTable*.

Considere o ficheiro Idade, em que os dados da variável se encontram nas células A2 a A231 e seguindo o processo utilizado para agrupar os dados referentes à variável Nº de filhos, obtivemos uma tabela, de que mostramos parte, na figura seguinte:

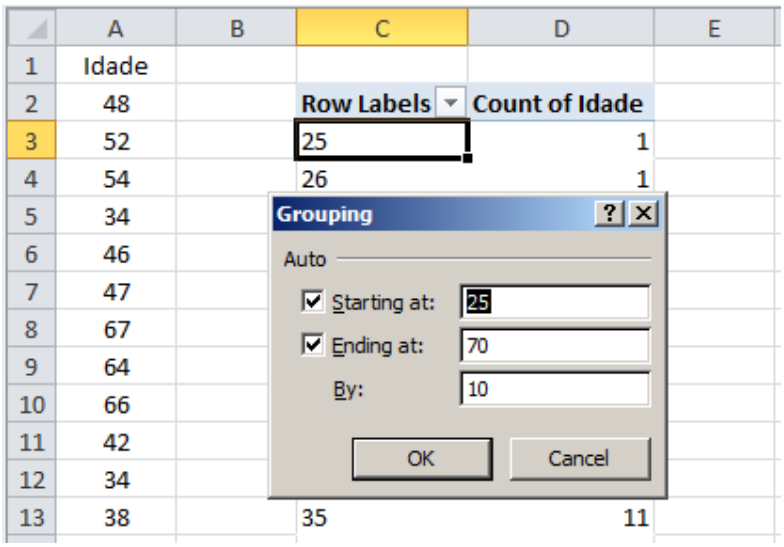


	A	B	C	D
1	Idade			
2	48		Row Labels	Count of Idade
3	52		25	1
4	54		26	1
5	34		27	2
6	46		28	4
7	47		29	1
8	67		30	4
9	64		31	7
10	66		32	5

Idade

A tabela 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:

- Clique em algum dos dados da variável Idade na tabela (clicámos no 25) e selecione *Data →Group→Group*, que faz surgir o seguinte diálogo:



	A	B	C	D	E
1	Idade				
2	48		Row Labels	Count of Idade	
3	52		25	1	
4	54		26	1	
5	34				
6	46				
7	47				
8	67				
9	64				
10	66				
11	42				
12	34				
13	38		35	11	

Por defeito, no diálogo anterior é considerado como "*Starting a*t" e "*Ending at*" respetivamente, 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.

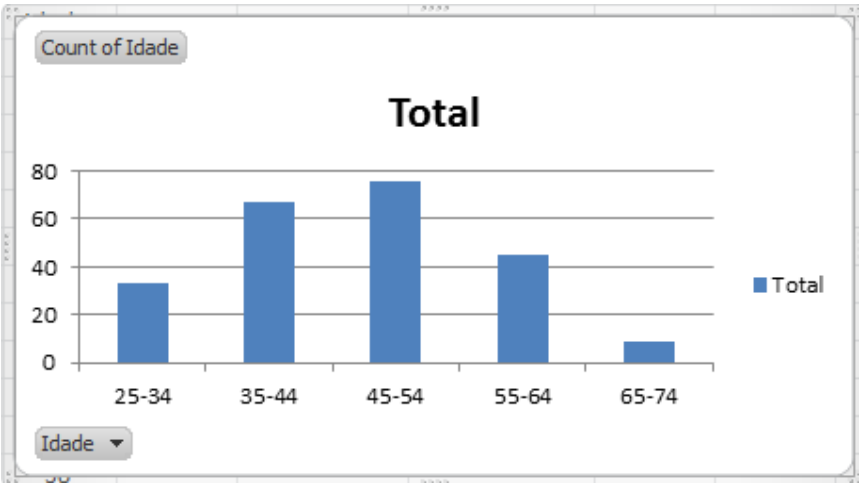
- Clicando em OK, é produzida a seguinte tabela de frequências:

	C	D
1		
2	**Row Labels**	**Count of Idade**
3	25-34	33
4	35-44	67
5	45-54	76
6	55-64	45
7	65-74	9
8	**Grand Total**	**230**

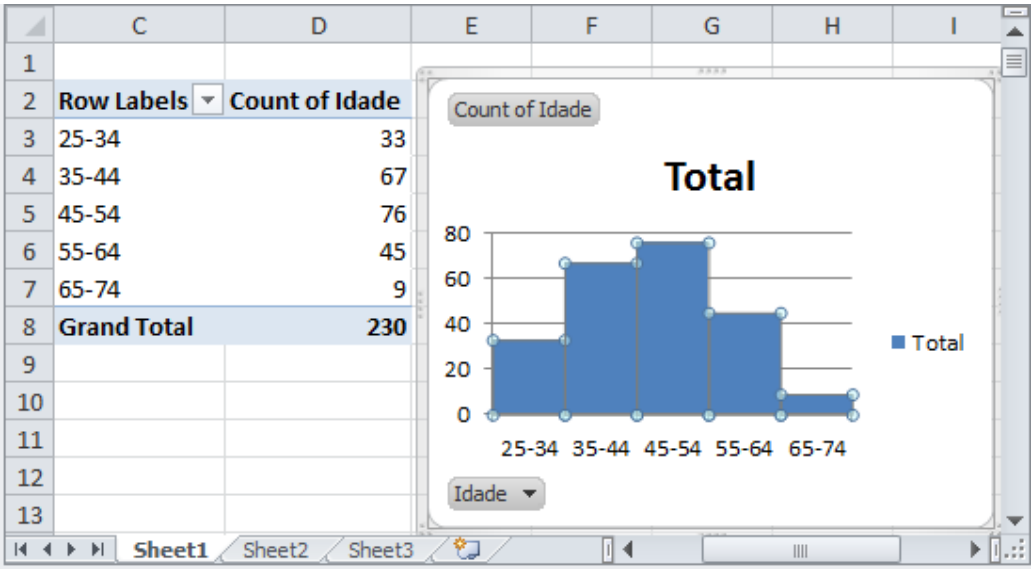
Sheet1

Observação: Repare-se que na construção desta tabela, ao dizer que pretendemos que o agrupamento seja feito *By:10*, não significa que se adicione 10 ao mínimo para formar a 1ª classe e assim por diante. No entanto os intervalos de classe têm, efetivamente, amplitude 10, pois são os seguintes intervalos: [25-35[, [35-45[, [45-55[, [55-65[e [65-75[.

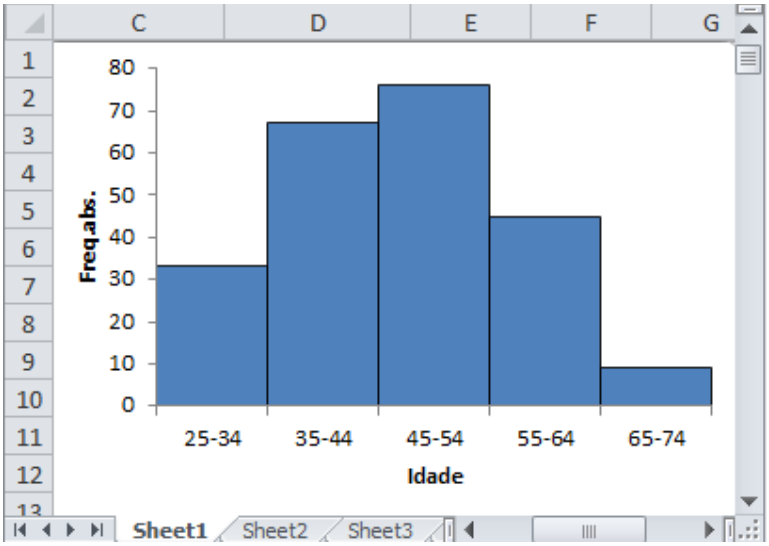
Para construir o histograma associado a esta tabela, basta carregar em alguma parte da tabela e na barra menu *Options→PivotChart→Column:*



Por defeito aparece a construção de um gráfico de barras, com intervalos entre as barras, que podem ser removidas por um processo idêntico ao já utilizado, aquando da construção do histograma no capítulo 2. Assim, temos:



- Finalmente podemos esconder os botões clicando com o lado direito do rato num deles e selecionando *Hide All Field Buttonson Chart* e acrescentando de seguida títulos aos eixos:



2ª parte – Dados em formato decimal

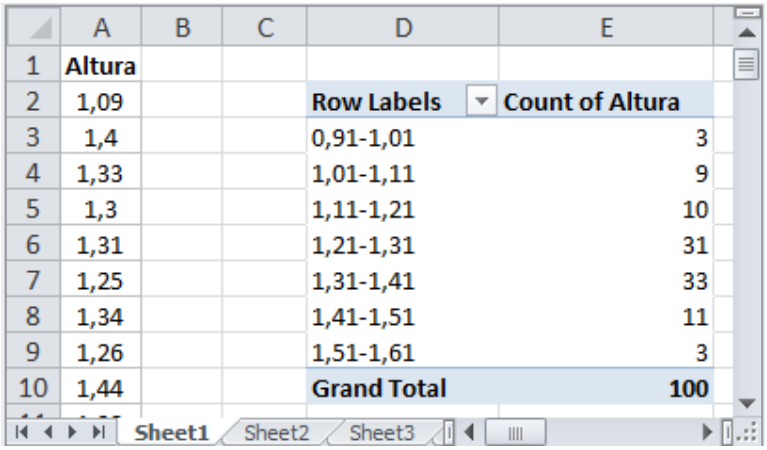
Como vimos na construção das classes da tabela anterior, estas são construídas sem ambiguidade, na medida em que qualquer elemento do conjunto de dados só pode pertencer a uma única classe. O mesmo não acontece se estivermos a trabalhar com dados com casas decimais, como veremos no exemplo seguinte.

Exemplo 4.3.4 – Considere os seguintes dados referentes às alturas de 100 alunos escolhidos aleatoriamente numa escola:

1,09	1,44	1,36	1,31	1,06	1,38	1,27	1,33	1,02	1,23
1,4	1,52	1,38	1,25	1,38	1,4	1,36	1,23	0,95	1,17

1,33	1,4	1,4	1,42	1,2	1,4	1,32	1,17	0,91	1,34
1,3	1,33	1,4	1,35	1,22	1,29	1,31	1,18	1,08	1,27
1,31	1,33	1,42	1,3	1,22	1,3	1,3	1,25	1,08	1,26
1,25	1,34	1,32	1,4	1,26	1,52	1,3	1,3	1,1	1,42
1,34	1,3	1,26	1,42	1,13	1,3	1,51	1,2	0,95	1,4
1,26	1,33	1,32	1,3	1,28	1,42	1,27	1,16	1,15	1,35
1,44	1,45	1,37	1,45	1,22	1,46	1,2	1,02	1,15	1,23
1,32	1,36	1,36	1,41	1,29	1,3	1,21	1,05	1,05	1,29

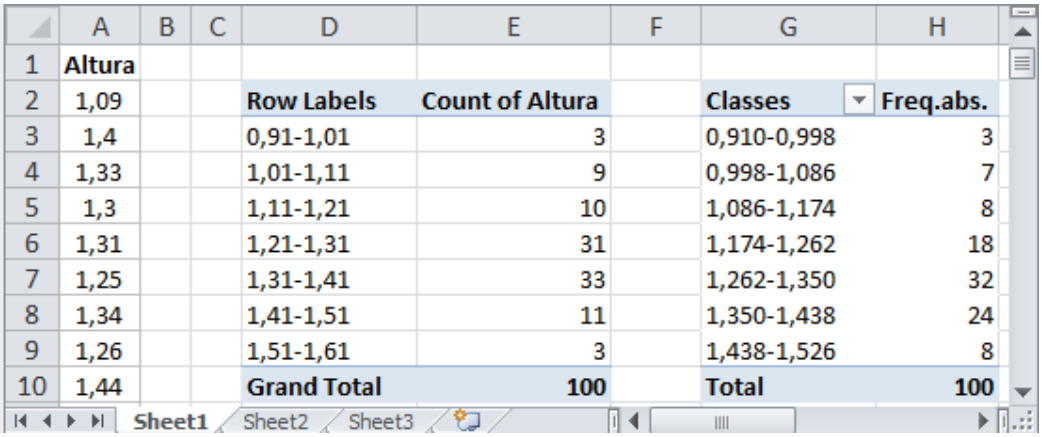
Inserimos os numa folha de Excel, ocupando as células A2 a A101, reservando a A1 para o título Altura. Construímos uma tabela de frequências, utilizando o processo seguido anteriormente. O resultado obtido foi a seguinte tabela:



	A	B	C	D	E
1	Altura				
2	1,09			Row Labels	Count of Altura
3	1,4			0,91-1,01	3
4	1,33			1,01-1,11	9
5	1,3			1,11-1,21	10
6	1,31			1,21-1,31	31
7	1,25			1,31-1,41	33
8	1,34			1,41-1,51	11
9	1,26			1,51-1,61	3
10	1,44			Grand Total	100

Sheet1 Sheet2 Sheet3

Como se verifica, ao contrário do que acontecia com a variável Idade, 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. Na verdade estes intervalos funcionam como se fossem fechados à esquerda e abertos à direita (exceto a última classe que também é fechada à direita), pelo que um valor igual, por exemplo, a 1,01, será contabilizado na classe 1,01-1,11. Este problema pode ser resolvido, a maior parte das vezes, considerando para amplitude de classe um valor decimal, com uma casa decimal a mais do que os dados. No exemplo anterior, vamos escolher para amplitude de classe o valor 0,088, sugerido pela aplicação da regra de Sturges:



	A	B	C	D	E	F	G	H
1	Altura							
2	1,09			Row Labels	Count of Altura		Classes	Freq.abs.
3	1,4			0,91-1,01	3		0,910-0,998	3
4	1,33			1,01-1,11	9		0,998-1,086	7
5	1,3			1,11-1,21	10		1,086-1,174	8
6	1,31			1,21-1,31	31		1,174-1,262	18
7	1,25			1,31-1,41	33		1,262-1,350	32
8	1,34			1,41-1,51	11		1,350-1,438	24
9	1,26			1,51-1,61	3		1,438-1,526	8
10	1,44			Grand Total	100		Total	100

Sheet1 Sheet2 Sheet3

O problema não ficou totalmente resolvido, pois ainda existe a ambiguidade (aparente) relativamente ao valor 1,35. Resolveríamos o problema da ambiguidade considerando como amplitude de classe 0,0879!

Nesta altura convém refletir sobre o que diz Neville Hunt no artigo intitutado “Handling Continuous data in Excel”,na revista *Teaching Statistics (Volume 25, Number 2, Summer 2003)*, página 45, e passamos a citar : ...*After reading this article, some teachers will (not unreasonably) decide that Excel is not fit to be used for this type of analysis. However, the universal popularity and availability of Excel are such that students will inevitably try to use it for this purpose at some stage, so it is important that they should be made aware of its limitations and need for vigilance*.

Esta citação vem ao encontro daquilo que pensamos e já referimos neste texto, de que o Excel não é um software de Estatística, mas ao nível elementar resolve muitas situações, desde que ao utilizá-lo se saiba o que se pretende. Por exemplo, quando se pretende um histograma, e se obtém um diagrama de barras, é necessário ter presente que, embora o histograma seja construído à custa de barras, estas têm que estar unidas.

5. Introdução à simulação

5.1- Introdução

Pretende-se com este Capítulo, dar a conhecer um instrumento poderoso – a simulação, que sobretudo nas duas últimas décadas, com o desenvolvimento e aperfeiçoamento dos meios computacionais, contribuiu de forma decisiva para o estudo das leis de probabilidade e a obtenção da probabilidade associada a determinados acontecimentos. Veremos assim uma forma de imitar o comportamento aleatório, característico dos fenómenos que têm interesse estudar em Probabilidade, isto é, os fenómenos chamados de aleatórios, por oposição aos determinísticos. Na verdade, essa possibilidade de imitação (simulação), baseia-se no facto de ao realizar uma experiência aleatória, repetidamente e em condições semelhantes, os resultados obtidos mostrarem uma regularidade estatística, que é utilizada para obter estimativas das probabilidades dos acontecimentos associados à experiência em causa. Esta regularidade a longo termo é a base da interpretação frequencista de Probabilidade. Simulando várias realizações de uma experiência aleatória, é então possível obter as estimativas consideradas anteriormente.

Por exemplo, ao lançar um dado equilibrado repetidas vezes, registando numa tabela de frequências, a frequência relativa da saída de cada face, verifica-se que à medida que o número de lançamentos aumenta, a frequência relativa da saída de cada face tende a estabilizar à volta do valor 0,167 (aproximadamente 1/6).

Embora não tenhamos chamado explicitamente a atenção para o facto, na verdade já utilizámos o conceito de simulação, quando no capítulo 1, utilizámos a função *Randbetween* do Excel, para "imitar" o comportamento aleatório da extração de uma amostra, de uma certa população.

Vamos ver de seguida, como por simulação se podem obter boas aproximações das probabilidades de acontecimentos, que teoricamente seriam difíceis, ou mesmo impossíveis de obter.

5.2- Obtenção de probabilidades por simulação

Vamos apresentar exemplos simples, que nos servirão para dar uma ideia da utilização e da potencialidade do método da simulação. Vamos utilizar as funções RAND ou RANDBETWEEN, já utilizadas no capítulo 1, que têm por base o conceito de número aleatório, ou mais propriamente pseudoaleatório.

Os algoritmos de geração de números pseudoaleatórios estão concebidos de modo a 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 o algoritmo 100 vezes, é de esperar que caiam 25 dos números gerados em cada quarto do intervalo [0,1]. Na tabela seguinte está listada uma sequência de 100 NPA's obtida através do gerador RAND do software Excel (Graça Martins, M. E e Loura, L., 2001):

0,842050	0,406320	0,848744	0,810469	0,789583
0,965131	0,676239	0,722927	0,825587	0,702971
0,761648	0,552387	0,079614	0,298300	0,087455
0,359825	0,208420	0,098150	0,818893	0,103532
0,054705	0,102768	0,147229	0,557920	0,996667
0,466613	0,493374	0,150888	0,540352	0,480287
0,814300	0,638416	0,086141	0,007840	0,109918
0,449515	0,090759	0,197460	0,209145	0,713230
0,901502	0,552418	0,466389	0,221584	0,623757
0,862762	0,507097	0,613583	0,389183	0,129629
0,395195	0,415666	0,210044	0,379011	0,302539
0,420519	0,469764	0,053714	0,478208	0,444822
0,124664	0,765629	0,737348	0,696311	0,806147
0,537707	0,451921	0,702749	0,683382	0,377823
0,033277	0,523063	0,908485	0,708764	0,196290
0,024371	0,213326	0,442821	0,983754	0,970551
0,558313	0,283191	0,153907	0,655705	0,995760
0,087859	0,429387	0,735276	0,890680	0,569285
0,069915	0,221549	0,358037	0,578713	0,161851
0,774156	0,039495	0,490216	0,755072	0,753139

Como se pode verificar por contagem, esta lista inclui 30 números no intervalo [0,0.25], 24 números nos intervalos]0.25,0.5] e]0.5,0.75] e 22 números no intervalo]0.75,1]. Embora haja métodos estatísticos para avaliar se são ou não significativas as diferenças entre estas frequências observadas e as frequências esperadas (25 – 25 – 25 – 25), facilmente a nossa sensibilidade aceita que estes resultados não contradizem o que se esperaria de uma escolha ao acaso de 100 números do intervalo [0,1]

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).

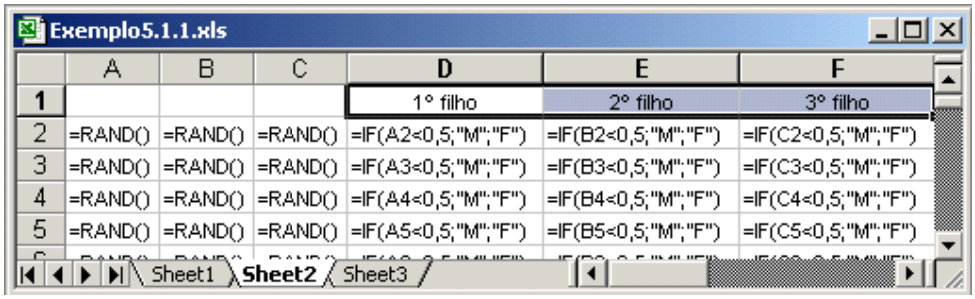
Exemplo 5.1.1 (Adaptado do exemplo 6.2.1 de *Graça Martins et al, 1999*) – Suponha um casal que pretende ter um "casal" de filhos, não desejando mais do que 3 filhos e só tentando o 3º filho se anteriormente tiver tido ou dois rapazes ou duas raparigas. Qual a probabilidade de ter efetivamente o casalinho?

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 optado 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 procedimento 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:



- 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 objetivo guardar os valores gerados anteriormente, pois a função RAND() é volátil, como já referimos nos capítulos anteriores);
- 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.

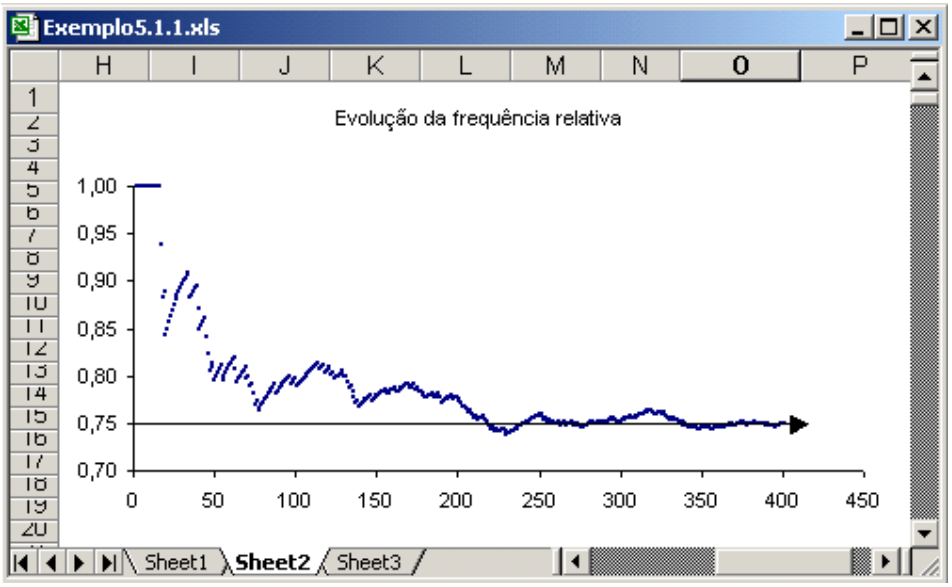
O processo anterior é apresentado na figura seguinte. Por uma questão de espaço só apresentamos a parte inicial e a parte final da tabela:

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 respetivamente 0,790, 0,775 e 0,753. Apresentamos a evolução da frequência relativa na seguinte representação gráfica:

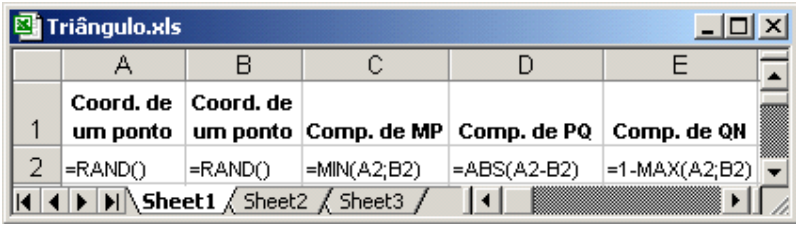


Exemplo 5.1.2 (Ageel, M. I. - Teaching Statistics, Volume 24, Number 2, Summer 2002, pag. 51-54) – Um segmento de linha de comprimento 1 é partido, aleatoriamente, em três pedaços. Qual a probabilidade de as peças resultantes poderem formar um triângulo?

A resolução deste problema prende-se com uma regra que estabelece que a soma dos comprimentos de dois lados de um triângulo é superior ao comprimento do outro lado. Vamos resolver este problema fazendo uma série de simulações e calculando a frequência relativa das situações que dão origem a triângulos. Considera-se então uma folha de cálculo e procede-se da seguinte forma:

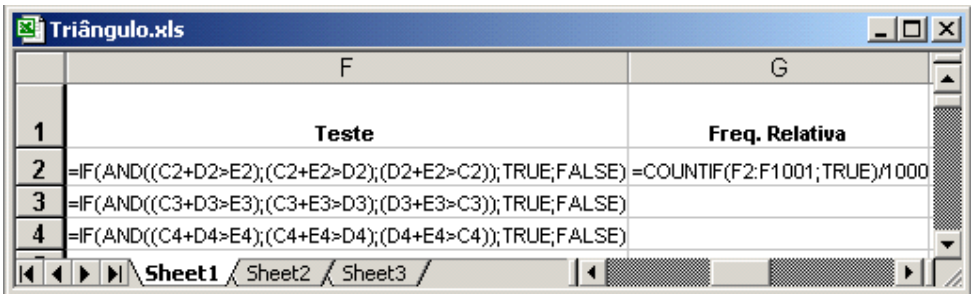
- Nas células A2 e B2 introduz-se a função RAND(), que devolve um número pseudoaleatório entre 0 e 1 (equivalente à função RANDBETWEEN(0;1)). Estes números irão representar os pontos P e Q em que uma linha MN de comprimento 1 fica dividida:

- Considera-se para P o menor dos valores obtidos anteriormente, que será o comprimento de MP – célula C2;
- Calculam-se o comprimentos dos segmentos PQ e QN – células D2 e E2, respetivamente:

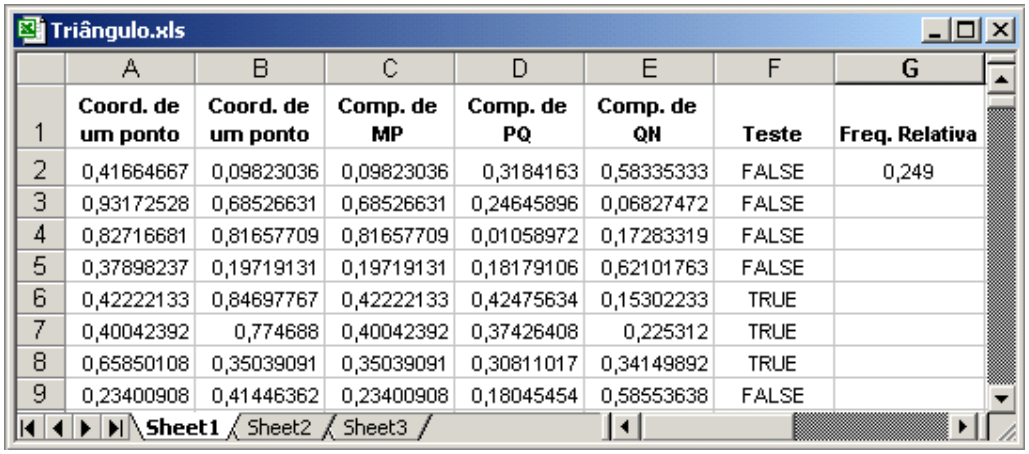


- Testa-se se dos 2 quaisquer dos comprimentos obtidos anteriormente são superior ao terceiro comprimento – célula F2;
- Replica-se as células de A2 a F2 até à linha 1001 (1000 réplicas);

- Calcula-se o número de vezes que o teste anterior deu verdadeiro, ou seja TRUE – célula G2, e divide-se por 1000:



O resultado da simulação anterior deu uma frequência relativa de 0,249, que se pode considerar um valor aproximado para a probabilidade pretendida:



	A	B	C	D	E	F	G
1	Coord. de um ponto	Coord. de um ponto	Comp. de MP	Comp. de PQ	Comp. de QN	Teste	Freq. Relativa
2	0,41664667	0,09823036	0,09823036	0,3184163	0,58335333	FALSE	0,249
3	0,93172528	0,68526631	0,68526631	0,24645896	0,06827472	FALSE	
4	0,82716681	0,81657709	0,81657709	0,01058972	0,17283319	FALSE	
5	0,37898237	0,19719131	0,19719131	0,18179106	0,62101763	FALSE	
6	0,42222133	0,84697767	0,42222133	0,42475634	0,15302233	TRUE	
7	0,40042392	0,774688	0,40042392	0,37426408	0,225312	TRUE	
8	0,65850108	0,35039091	0,35039091	0,30811017	0,34149892	TRUE	
9	0,23400908	0,41446362	0,23400908	0,18045454	0,58553638	FALSE	

Do mesmo modo que a função RANDBETWEEN, também a função RAND é volátil, pelo que qualquer operação na folha de cálculo modifica os números pseudoaleatórios considerados para coordenadas dos pontos e consequentemente a estimativa da probabilidade pretendida. Assim, quantas operações forçar na folha anterior, nomeadamente digitar um valor numa das células em branco consiste numa operação, quantas estimativas obterá para a probabilidade pretendida, ou seja, para a probabilidade de conseguir construir um triângulo com as partes de um segmento de reta de comprimento unitário, dividido aleatoriamente em 3 partes.

Exemplo 5.1.3 - Suponha que em cada minuto a probabilidade de alguém chegar à fila de uma caixa de supermercado é de 75%, enquanto que a probabilidade de abandonar a fila, depois de ser servido é de 30%. Ao fim de 20 minutos qual o tamanho que espera para a fila?

Vamos simular a experiência anterior, simulando a chegada de um cliente à fila sempre que o resultado da função RAND for ≤ 0,75 e a saída de um cliente da fila sempre que a função RAND devolver um resultado ≤ 0,30:

Supermercado.xls

	A	B	C	D	E	F
1	**Minutos**	**Sim. Chegada**	**Chegada fila**	**Sim. Partida**	**Partida fila**	**Estado fila**
2	1	=RAND()	=IF(B2<0,75;1;0)	=RAND()	=IF(D2<0,3;1;0)	=MAX(C2-E2;0)
3	2	=RAND()	=IF(B3<0,75;1;0)	=RAND()	=IF(D3>0,3;0;1)	=MAX(F2+C3-E3;0)
4	3	=RAND()	=IF(B4<0,75;1;0)	=RAND()	=IF(D4>0,3;0;1)	=MAX(F3+C4-E4;0)
5	4	=RAND()	=IF(B5<0,75;1;0)	=RAND()	=IF(D5>0,3;0;1)	=MAX(F4+C5-E5;0)

Sheet1 Sheet2 Sheet3

Para não corrermos os riscos de termos uma fila com um número negativo de pessoas, consideramos a função máximo:

Supermercado.xls

	A	B	C	D	E	F
1	**Minutos**	**Sim. Chegada**	**Chegada fila**	**Sim. Partida**	**Partida fila**	**Estado fila**
2	1	0,011032152	1	0,55809092	0	1
3	2	0,222252425	1	0,95239588	0	2
4	3	0,691275905	1	0,23198885	1	2
5	4	0,573837071	1	0,86520536	0	3
6	5	0,393127845	1	0,11988901	1	3
7	6	0,641297179	1	0,39092097	0	4
8	7	0,446198809	1	0,3298297	0	5
9	8	0,468829371	1	0,58457985	0	6
10	9	0,370038386	1	0,35831157	0	7
11	10	0,189264634	1	0,97929665	0	8
12	11	0,883702743	0	0,0559692	1	7
13	12	0,884816981	0	0,99889606	0	7
14	13	0,36954121	1	0,47555043	0	8
15	14	0,592067357	1	0,90484252	0	9
16	15	0,669707921	1	0,41281699	0	10
17	16	0,48698174	1	0,85900034	0	11
18	17	0,453637211	1	0,38237448	0	12
19	18	0,511048693	1	0,22054477	1	12
20	19	0,181512884	1	0,84936245	0	13
21	20	0,799964271	0	0,66043165	0	13

Sheet1 Sheet2 Sheet3

Ao fim de 20 minutos a fila já tem 13 clientes e com tendência para crescer!

Exemplo 5.1.4 – Suponha uma espécie animal em que as fêmeas têm o seguinte comportamento reprodutor:

- 40% morrem antes de deixar descendência
- 40% têm uma fêmea descendente
- 20% têm duas fêmeas descendentes

Estude o comportamento desta população, nomeadamente se se prevê um crescimento rápido de indivíduos da espécie, a extinção ou uma situação de equilíbrio.

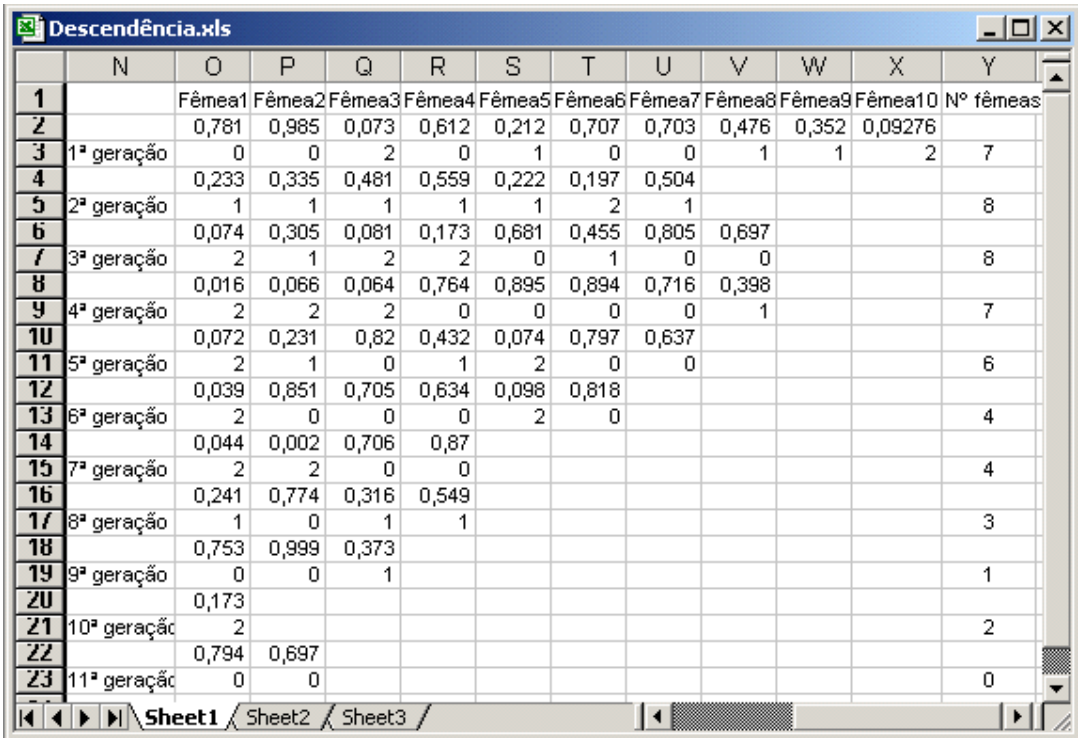
Vamos estudar a evolução da população simulando a descendência de 10 fêmeas, ao longo de algumas gerações. Para cada fêmea, geramos um número pseudoaleatório, cujo resultado será interpretado da seguinte forma:

Se o número for inferior a 0,20, a fêmea deixa 2 descendentes fêmeas;

Se o número estiver compreendido entre 0,2 e 0,6, a fêmea deixa 1 descendente fêmea;

Se o número estiver compreendido entre 0,6 e 1, a fêmea morre sem descendência.

A presentamos a seguir uma simulação da experiência com as 10 fêmeas:



Descendência.xls

	N	O	P	Q	R	S	T	U	V	W	X	Y
1		Fêmea1	Fêmea2	Fêmea3	Fêmea4	Fêmea5	Fêmea6	Fêmea7	Fêmea8	Fêmea9	Fêmea10	Nº fêmeas
2		0,781	0,985	0,073	0,612	0,212	0,707	0,703	0,476	0,352	0,09276	
3	1ª geração	0	0	2	0	1	0	0	1	1	2	7
4		0,233	0,335	0,481	0,559	0,222	0,197	0,504				
5	2ª geração	1	1	1	1	1	2	1				8
6		0,074	0,305	0,081	0,173	0,681	0,455	0,805	0,697			
7	3ª geração	2	1	2	2	0	1	0	0			8
8		0,016	0,066	0,064	0,764	0,895	0,894	0,716	0,398			
9	4ª geração	2	2	2	0	0	0	0	1			7
10		0,072	0,231	0,82	0,432	0,074	0,797	0,637				
11	5ª geração	2	1	0	1	2	0	0				6
12		0,039	0,851	0,705	0,634	0,098	0,818					
13	6ª geração	2	0	0	0	2	0					4
14		0,044	0,002	0,706	0,87							
15	7ª geração	2	2	0	0							4
16		0,241	0,774	0,316	0,549							
17	8ª geração	1	0	1	1							3
18		0,753	0,999	0,373								
19	9ª geração	0	0	1								1
20		0,173										
21	10ª geração	2										2
22		0,794	0,697									
23	11ª geração	0	0									0

Sheet1 Sheet2 Sheet3

Na tabela anterior considerámos:

- Nas células O2:X2, 10 números pseudoaleatórios para simular a descendência das 10 fêmeas com que iniciámos a nossa experiência;
- Na célula Y3, o número de fêmeas obtidas ao fim da primeira geração – neste caso 7;
- Nas células O4:U4, 7 números pseudoaleatórios para simular a descendência das 7 fêmeas obtidas na geração anterior;
- Na célula Y5, o número de fêmeas obtidas ao fim da segunda geração – neste caso 8;
- Repetimos o processo anterior, até não haver descendência de fêmeas.

Como se verifica, a população tem tendência a extinguir-se, pois ao fim da 11ª geração já não há descendentes das 10 fêmeas com que iniciámos o estudo.

Repita a experiência admitindo que:

- 20% morrem antes de deixar descendência

- 40% Têm uma fêmea descendente
- 40% têm duas fêmeas descendentes

Um outro exemplo interessante e que tem levantado bastante polémica é o seguinte exemplo de decisão estratégica.

Exemplo 5.1.5 (Graça Martins, M. E. e Loura, L., 2001) - Num concurso é dada a escolher ao concorrente uma de 3 portas. Atrás de uma delas está um carro e atrás de cada uma das outras duas está uma ovelha. O concorrente escolhe uma das portas (sem a abrir) e o apresentador, que sabe exatamente qual é a porta que esconde o carro, abre, de entre as duas portas que restam, uma onde está uma ovelha. Nesse momento pergunta ao concorrente se deseja ou não trocar a porta que escolheu pela outra porta que ainda está fechada. O primeiro pensamento que ocorre é que não há qualquer vantagem em trocar, pois temos agora apenas duas portas e o carro tanto pode estar atrás de uma como da outra. No entanto, se se calcular teoricamente a probabilidade do concorrente ganhar o carro, trocando de porta, verifica-se que esta é igual a 2/3. Para os mais reticentes uma simulação talvez os faça reconsiderar a sua posição inicial. Não há qualquer dúvida de que ao escolher uma porta ao acaso a probabilidade de ela esconder o carro é igual a 1/3.

Para simular o decorrer de 100 destes concursos vamos então considerar que o concorrente escolheu a boa porta sempre que o valor do número pseudoaleatório (NPA) estiver entre 0 e 1/3. Nestes casos, quando ele trocar de porta, ficará com a “ovelha” mas, em compensação, ficará com o carro em todos os outros casos (se ele tiver escolhido inicialmente a “ovelha”, a porta que resta terá obrigatoriamente o carro pois o apresentador encarregou-se de eliminar a outra porta que também tinha “ovelha”!...)

Eis o resultado da simulação obtida a partir de 100 números pseudoaleatórios gerados numa folha de Excel:

NPA	O que ganha não trocando	O que ganha trocando	NPA	O que ganha não trocando	O que ganha trocando	NPA	O que ganha não trocando	O que ganha trocando
0,842	Ovelha	Carro	0,406	Ovelha	Carro	0,849	Ovelha	Carro
0,965	Ovelha	Carro	0,676	Ovelha	Carro	0,723	Ovelha	Carro
0,762	Ovelha	Carro	0,552	Ovelha	Carro	0,080	Carro	Ovelha
0,360	Ovelha	Carro	0,208	Carro	Ovelha	0,098	Carro	Ovelha
0,055	Carro	Ovelha	0,103	Carro	Ovelha	0,147	Carro	Ovelha
0,467	Ovelha	Carro	0,493	Ovelha	Carro	0,151	Carro	Ovelha
0,814	Ovelha	Carro	0,638	Ovelha	Carro	0,086	Carro	Ovelha
0,450	Ovelha	Carro	0,091	Carro	Ovelha	0,197	Carro	Ovelha
0,902	Ovelha	Carro	0,552	Ovelha	Carro	0,466	Ovelha	Carro
0,863	Ovelha	Carro	0,507	Ovelha	Carro	0,614	Ovelha	Carro
0,395	Ovelha	Carro	0,416	Ovelha	Carro	0,210	Carro	Ovelha
0,421	Ovelha	Carro	0,470	Ovelha	Carro	0,054	Carro	Ovelha
0,125	Carro	Ovelha	0,766	Ovelha	Carro	0,737	Ovelha	Carro
0,538	Ovelha	Carro	0,452	Ovelha	Carro	0,703	Ovelha	Carro
0,033	Carro	Ovelha	0,523	Ovelha	Carro	0,908	Ovelha	Carro
0,024	Carro	Ovelha	0,213	Carro	Ovelha	0,443	Ovelha	Carro
0,558	Ovelha	Carro	0,283	Carro	Ovelha	0,154	Carro	Ovelha
0,088	Carro	Ovelha	0,429	Ovelha	Carro	0,735	Ovelha	Carro
0,070	Carro	Ovelha	0,222	Carro	Ovelha	0,358	Ovelha	Carro
0,774	Ovelha	Carro	0,039	Carro	Ovelha	0,490	Ovelha	Carro
0,810	Ovelha	Carro	0,709	Ovelha	Carro	0,713	Ovelha	Carro
0,826	Ovelha	Carro	0,984	Ovelha	Carro	0,624	Ovelha	Carro
0,298	Carro	Ovelha	0,656	Ovelha	Carro	0,130	Carro	Ovelha
0,819	Ovelha	Carro	0,891	Ovelha	Carro	0,303	Carro	Ovelha
0,558	Ovelha	Carro	0,579	Ovelha	Carro	0,445	Ovelha	Carro
0,540	Ovelha	Carro	0,755	Ovelha	Carro	0,806	Ovelha	Carro
0,008	Carro	Ovelha	0,790	Ovelha	Carro	0,378	Ovelha	Carro
0,209	Carro	Ovelha	0,703	Ovelha	Carro	0,196	Carro	Ovelha
0,222	Carro	Ovelha	0,087	Carro	Ovelha	0,971	Ovelha	Carro
0,389	Ovelha	Carro	0,104	Carro	Ovelha	0,996	Ovelha	Carro
0,379	Ovelha	Carro	0,997	Ovelha	Carro	0,569	Ovelha	Carro
0,478	Ovelha	Carro	0,480	Ovelha	Carro	0,162	Carro	Ovelha
0,696	Ovelha	Carro	0,110	Carro	Ovelha	0,753	Ovelha	Carro
0,683	Ovelha	Carro						

Como se verifica, nas 100 realizações simuladas deste concurso o concorrente ganharia o carro em 67 dessas realizações, se se decidisse por trocar de porta!...

Lista de algumas funções usadas no Excel:

Inglês	Português	
And()	*E()*	Devolve verdadeiro se todos os argumentos forem verdadeiros e devolve falso se algum dos argumentos for falso
Average()	*Media()*	Calcula a média dos valores existentes num conjunto de células
Count()	*Contar()*	Conta as células com valores numéricos, incluindo datas e fórmulas cujos resultados são numeros
Counta()	*Contar.val()*	Conta todas as células não vazias
Countblank()	*Contar.vazio()*	Conta as células vazias
Countif()	*Contar.se()*	Conta as ocorrências verificadas num conjunto de célula, que obedecem a um critério
Frequency()	*Frequência*	
If()	*Se()*	Executa uma de duas acções possíveis, em função do resultado da condição
Int()	*Int()*	Devolve a parte inteira de um número
Max()	*Maximo()*	Devolve o maior valor de um conjunto de células
Min()	*Minimo()*	Devolve o menor valor de um conjunto de células
Mod()	*Resto()*	Devolve o resto de uma divisão
Or()	*Ou()*	Devolve verdadeiro se um dos argumentos for verdadeiros e devolve falso se todos os argumentos forem falsos
Pie		
Product()	*Produto()*	Multiplica os valores de um conjunto de células, ignorando as células vazias e/ou com texto
Rand()	*Aleatório()*	Devolve um número pseudoaleatório (no intervalo (0,1))
Randbetween()	*Aleatórioentre()*	Devolve um número pseudoaleatório no intervalo especificado
Round()	*Arred()*	Devolve um número arredondado, na posição indicada
Rounddown()	*Arred.para.baixo()*	Devolve um número arredondado, por defeito, na posição indicada
Roundup()	*Arred.para.cima()*	Devolve um número arredondado, por excesso, na posição indicada
Scatter		
Sum()	*Soma()*	Soma os valores de um conjunto de células
Sumif()	*Soma.se()*	Soma as ocorrências verificadas num conjunto de células que obedecem a um critério
Sumproduct()	*Somarproduto()*	Multiplica dois conjuntos de células e devolve a soma total dos produtos

Vlookup()	*Procv()*	Procura um valor na coluna mais à esquerda de uma tabela e devolve um valor na mesma linha na coluna indicada

Bibliografia / Outros Recursos

BARNETT, V. (1997) – *Sample Survey: Principles & Methods*, Arnold, London.

GRAÇA MARTINS, M.E. et al (1999) – *Introdução às Probabilidades e à Estatística*, Edição da Universidade Aberta.

GRAÇA MARTINS, M.E. (2005) – *Introdução à Probabilidade e à Estatística* – Com complementos de Excel. Edição da Sociedade Portuguesa de Estatística. www.arquivoescolar.org

GRAÇA MARTINS, M.E. et al (2001) – *Estatística* – 10º ano de escolaridade, Edição do Ministério da Educação – Departamento do Ensino Secundário.

GRAÇA MARTINS, M.E. e Loura, L. (2001) – Matemática para as Ciências Sociais – Anexo para apoio à interpretação do programa.

GRAÇA MARTINS, M. E., LOURA, L., MENDES, F. (2007) – *Análise de dados*, Texto de apoio para os professores do 1º ciclo, Ministério da Educação, DGIDC. ISBN-978-972-742-261-6. Depósito legal 262674/07

GRAÇA MARTINS, M. E., PONTE, J. P. (2010) – *Organização e tratamento de dados*, http://area.dgidc.min-edu.pt/materiais_NPMEB/matematicaOTD_Final.pdf

MOORE, D. (1992) – What is Statistics in *Perspectives on Contemporary Statistics*, Edição de David Hoaglin e David Moore, The Mathematical Association of America.

MOORE, D. ET AL (1996) – *Introduction to the Practice of Statistics*, Freeman, New York.

MOORE, D. (1996) – *The Basic Practice of Statistics*, Freeman, New York.

MOORE, D. (1997) – *Statistics – Concepts and Controversies*, Freeman, New York.

MURTEIRA, B. (1993) – Análise Exploratória de Dados. Estatistica Descritiva, McGraw-Hill.

COMAP, (2000) – *For all Practical Purposes: Mathematical Literacy in Todays World*, Freeman and Company, New York.

ROSSMAN, A. et al (2001) –*Workshop Statistics – Discovery with data,* Key College Publishing.

TANNENBAUM. P. et al (1998) – Excursions in modern Mathematics, Prentice Hall.

VICENTE, P., REIS, E., FERRÃO, F. (1996) – *Sondagens*, Edições Sílabo.

Artigos da revista TEACHING STATISTICS

AGEEL, M.I. – Spreadsheets as a Simulation Tool for Solving Probability Problems, Vol 24, 2, 51-54.

Hodgson, T., and Borkowski, J. - Why Stratify? 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

ESCOLA SECUNDÁRIA TOMAZ PELAYO E INSTITUTO NACIONAL DE ESTATÍSTICA

PROJETO 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://www.stat.fi/tup/maanum/index_en.html

Tem informação das mais diversas áreas, tais como população e estatísticas vitais, cultura, religiões, emprego, consumo, etc., relativa a todos os países do mundo.

Anexo – Ficheiro de Deputados da XII Legislatura

	Nome	Círculo eleitoral	Grupo parlamentar	Data nascimento
1	Abel Batista	Viana do Castelo	CDS-PP	13-10-1963
2	Acácio Pinto	Viseu	PS	14-05-1959
3	Adão Silva	Bragança	PSD	01-10-1957
4	Adolfo Mesquita Nunes	Lisboa	CDS-PP	29-11-1977
5	Adriano Rafael Moreira	Porto	PSD	17-08-1965
6	Afonso Oliveira	Porto	PSD	27-03-1964
7	Agostinho Lopes	Braga	PCP	16-11-1944
8	Alberto Costa	Lisboa	PS	16-08-1947
9	Alberto Martins	Porto	PS	25-04-1945
10	Altino Bessa	Braga	CDS-PP	02-08-1969
11	Amadeu Soares Albergaria	Aveiro	PSD	16-01-1977
12	Ana Catarina Mendonça Mendes	Setúbal	PS	14-01-1973
13	Ana Drago	Lisboa	BE	28-08-1975
14	Ana Oliveira	Coimbra	PSD	10-04-1984
15	Ana Paula Vitorino	Porto	PS	25-04-1962
16	Ana Sofia Bettencourt	Lisboa	PSD	24-03-1972
17	Andreia Neto	Porto	PSD	04-08-1980
18	Ângela Guerra	Guarda	PSD	20-06-1973
19	António Braga	Braga	PS	21-04-1953
20	António Filipe	Santarém	PCP	28-01-1963
21	António José Seguro	Braga	PS	11-03-1962
22	António Leitão Amaro	Lisboa	PSD	02-04-1980
23	António Prôa	Lisboa	PSD	26-10-1969
24	António Rodrigues	Lisboa	PSD	30-05-1958
25	António Serrano	Santarém	PS	16-01-1965
26	Arménio Santos	Viseu	PSD	22-11-1945
27	Artur Rêgo	Faro	CDS-PP	24-04-1958
28	Assunção Esteves	Lisboa	PSD	15-10-1956
29	Basílio Horta	Leiria	PS	16-11-1943
30	Bernardino Soares	Lisboa	PCP	15-09-1971
31	Bruno Coimbra	Aveiro	PSD	21-04-1981
32	Bruno Dias	Setúbal	PCP	19-10-1976
33	Bruno Vitorino	Setúbal	PSD	15-05-1971
34	Carina Oliveira	Santarém	PSD	22-07-1977
35	Carla Rodrigues	Aveiro	PSD	18-01-1973
36	Carlos Abreu Amorim	Viana Castelo	PSD	01-10-1963
37	Carlos Alberto Gonçalves	Europa	PSD	20-10-1961
38	Carlos Costa Neves	Castelo Branco	PSD	16-06-1954
39	Carlos Enes	Açores	PS	10-03-1951
40	Carlos Páscoa Gonçalves	Fora da Europa	PSD	09-02-1952
41	Carlos Peixoto	Guarda	PSD	13-02-1968
42	Carlos Santos Silva	Lisboa	PSD	07-10-1964
43	Carlos São Martinho	Castelo	PSD	18-07-1956

		Branco		
44	Carlos Zorrinho	Évora	PS	28-05-1959
45	Catarina Martins	Porto	BE	07-09-1973
46	Cecília Honório	Faro	BE	01-07-1962
47	Clara Marques Mendes	Braga	PSD	30-04-1970
48	Cláudia Monteiro de Aguiar	Madeira	PSD	08-04-1982
49	Conceição Bessa Ruão	Porto	PSD	28-08-1954
50	Correia de Jesus	Madeira	PSD	16-12-1941
51	Couto dos Santos	Aveiro	PSD	18-05-1949
52	Cristóvão Crespo	Portalegre	PSD	01-09-1958
53	Cristóvão Norte	Faro	PSD	06-08-1976
54	Cristóvão Simão Ribeiro	Porto	PSD	13-05-1986
55	Duarte Cordeiro	Setúbal	PS	23-02-1979
56	Duarte Marques	Santarém	PSD	09-05-1981
57	Duarte Pacheco	Lisboa	PSD	25-11-1965
58	Eduardo Cabrita	Setúbal	PS	26-09-1961
59	Eduardo Teixeira	Viana Castelo	PSD	25-06-1972
60	Elsa Cordeiro	Faro	PSD	16-06-1968
61	Elza Pais	Viseu	PS	22-11-1958
62	Emídio Guerreiro	Braga	PSD	23-05-1965
63	Emília Santos	Porto	PSD	16-05-1972
64	Eurídice Pereira	Setúbal	PS	20-10-1962
65	Fernando Jesus	Porto	PS	04-06-1950
66	Fernando Marques	Leiria	PSD	01-06-1958
67	Fernando Medina	Viana Castelo	PS	10-03-1973
68	Fernando Negrão	Braga	PSD	29-11-1955
		Castelo		
69	Fernando Serrasqueiro	Branco	PS	20-06-1951
70	Fernando Virgílio Macedo	Porto	PSD	28-12-1965
71	Ferro Rodrigues	Lisboa	PS	03-11-1949
72	Filipe Neto Brandão	Aveiro	PS	09-09-1968
73	Francisca Almeida	Braga	PSD	06-11-1983
74	Francisco de Assis	Porto	PS	08-01-1965
75	Francisco Lopes	Setúbal	PCP	29-08-1955
76	Francisco Louçã	Lisboa	BE	12-11-1956
77	Glória Araújo	Porto	PS	04-01-1976
78	Graça Mota	Braga	PSD	24-08-1955
79	Guilherme Silva	Madeira	PSD	16-07-1943
80	Helder Amaral	Viseu	CDS-PP	08-06-1967
81	Hélder Sousa Silva	Lisboa	PSD	21-07-1965
82	Heloísa Apolónia	Setúbal	PEV	26-06-1969
83	Honório Novo	Porto	PCP	24-10-1950
		Castelo		
84	Hortense Martins	Branco	PS	21-09-1966
85	Hugo Lopes Soares	Braga	PSD	02-03-1983
86	Hugo Velosa	Madeira	PSD	18-04-1948
87	Idália Salvador Serrão	Santarém	PS	16-05-1964
88	Inês de Medeiros	Lisboa	PS	15-04-1968
89	Inês Teotónio Pereira	Lisboa	CDS-PP	14-12-1971

90	Isabel Alves Moreira	Lisboa	PS	02-04-1976
91	Isabel Galriça Neto	Lisboa	CDS-PP	10-07-1961
92	Isabel Oneto	Porto	PS	14-09-1959
93	Isabel Santos	Porto	PS	12-02-1968
94	Isilda Aguincha	Santarém	PSD	03-04-1966
95	Jacinto Serrão	Madeira	PS	16-02-1969
96	Jerónimo de Sousa	Lisboa	PCP	13-04-1947
97	Joana Barata Lopes	Lisboa	PSD	16-03-1985
98	João Figueiredo	Viseu	PSD	18-05-1967
99	João Galamba	Santarém	PS	04-08-1976
100	João Gonçalves Pereira	Lisboa	CDS-PP	07-05-1977
101	João Lobo	Braga	PSD	01-01-1952
102	João Oliveira	Évora	PCP	09-07-1979
103	João Paulo Pedrosa	Leiria	PS	29-09-1965
104	João Paulo Viegas	Setúbal	CDS-PP	30-06-1970
105	João Pinho de Almeida	Porto	CDS-PP	11-09-1976
106	João Portugal	Coimbra	PS	01-10-1977
107	João Prata	Guarda	PSD	04-06-1963
108	João Ramos	Beja	PCP	13-10-1972
109	João Rebelo	Lisboa	CDS-PP	02-02-1970
110	João Semedo	Porto	BE	20-06-1951
111	João Serpa Oliva	Coimbra	CDS-PP	26-06-1948
112	João Soares	Faro	PS	29-08-1949
113	Joaquim Ponte	Açores	PSD	06-06-1956
114	Jorge Fão	Viana Castelo	PS	04-11-1957
115	Jorge Lacão	Lisboa	PS	04-09-1954
116	Jorge Machado	Porto	PCP	20-05-1976
117	Jorge Paulo Oliveira	Braga	PSD	28-12-1965
118	José de Matos Correia	Lisboa	PSD	08-05-1963
119	José de Matos Rosa	Lisboa	PSD	04-09-1959
120	José Junqueiro	Viseu	PS	28-06-1953
121	José Lello	Porto	PS	18-05-1944
122	José Lino Ramos	Lisboa	CDS-PP	21-03-1969
123	José Luís Ferreira	Lisboa	PEV	27-08-1962
124	José Manuel Canavarro	Coimbra	PSD	17-04-1965
125	José Manuel Rodrigues	Madeira	CDS-PP	13-07-1960
126	José Ribeiro e Castro	Porto	CDS-PP	24-12-1953
127	Laura Esperança	Leiria	PSD	19-02-1958
128	Laurentino Dias	Braga	PS	04-02-1954
129	Lídia Bulcão	Açores	PSD	19-04-1974
130	Luís Campos Ferreira	Porto	PSD	26-11-1961
131	Luís Fazenda	Lisboa	BE	08-10-1957
132	Luís Leite Ramos	Vila Real	PSD	29-10-1961
133	Luís Menezes	Porto	PSD	20-12-1980
134	Luís Montenegro	Aveiro	PSD	16-02-1973
135	Luís Pedro Pimentel	Vila Real	PSD	16-04-1970
136	Luís Pita Ameixa	Beja	PS	13-10-1960
137	Luís Vales	Porto	PSD	23-08-1979
138	Luísa Salgueiro	Porto	PS	02-01-1968

139	Manuel Isaac	Leiria	CDS-PP	26-01-1960
140	Manuel Pizarro	Porto	PS	02-02-1964
141	Manuel Seabra	Porto	PS	28-07-1962
142	Marcos Perestrello	Lisboa	PS	23-08-1971
143	Margarida Almeida	Porto	PSD	17-01-1955
144	Margarida Neto	Santarém	CDS-PP	24-10-1964
145	Maria Antónia Almeida Santos	Lisboa	PS	14-02-1963
146	Maria Conceição Pereira	Leiria	PSD	08-12-1950
147	Maria da Conceição Caldeira	Lisboa	PSD	13-07-1952
148	Maria das Mercês Borges	Setúbal	PSD	15-07-1957
149	Maria de Belém Roseira	Lisboa	PS	28-07-1949
150	Maria Ester Vargas	Viseu	PSD	28-03-1955
151	Maria Gabriela Canavilhas	Braga	PS	29-03-1961
152	Maria Helena André	Aveiro	PS	29-10-1960
153	Maria João Ávila	Fora da Europa	PSD	26-07-1956
154	Maria José Castelo Branco	Porto	PSD	28-07-1958
155	Maria José Moreno	Bragança	PSD	22-06-1963
156	Maria Manuela Tender	Vila Real	PSD	09-02-1971
157	Maria Paula Cardoso	Aveiro	PSD	24-08-1972
158	Mariana Aiveca	Setúbal	BE	03-02-1954
159	Mário Magalhães	Porto	PSD	07-02-1965
160	Mário Ruivo	Coimbra	PS	23-03-1960
161	Mário Simões	Beja	PSD	08-08-1970
162	Maurício Marques	Coimbra	PSD	28-12-1958
163	Mendes Bota	Faro	PSD	04-08-1955
164	Michael Seufert	Porto	CDS-PP	15-04-1983
165	Miguel Coelho	Lisboa	PS	04-07-1952
166	Miguel Frasquilho	Porto	PSD	12-11-1965
167	Miguel Freitas	Faro	PS	25-12-1960
168	Miguel Laranjeiro	Braga	PS	13-08-1965
169	Miguel Santos	Porto	PSD	24-03-1971
170	Miguel Tiago	Lisboa	PCP	27-08-1979
171	Miranda Calha	Porto	PS	17-11-1947
172	Mónica Ferro	Lisboa	PSD	07-11-1972
173	Mota Amaral	Açores	PSD	15-04-1943
174	Mota Andrade	Bragança	PS	25-11-1955
175	Nilza de Sena	Coimbra	PSD	21-11-1976
176	Nuno André Figueiredo	Porto	PS	12-02-1976
177	Nuno Encarnação	Coimbra	PSD	24-12-1972
178	Nuno Filipe Matias	Setúbal	PSD	26-01-1977
179	Nuno Magalhães	Setúbal	CDS-PP	04-03-1972
180	Nuno Reis	Braga	PSD	23-09-1978
181	Nuno Sá	Braga	PS	02-04-1976
182	Nuno Serra	Santarém	PSD	28-07-1973
183	Odete João	Leiria	PS	03-01-1958
184	Odete Silva	Lisboa	PSD	18-09-1971
185	Paula Santos	Setúbal	PCP	29-09-1980

186	Paulo Batista Santos	Leiria	PSD	03-12-1968
187	Paulo Cavaleiro	Aveiro	PSD	08-12-1973
188	Paulo Mota Pinto	Lisboa	PSD	18-11-1966
189	Paulo Pisco	Europa	PS	22-08-1961
190	Paulo Ribeiro de Campos	Guarda	PS	07-04-1965
191	Paulo Rios de Oliveira	Porto	PSD	19-02-1965
192	Paulo Sá	Faro	PCP	12-07-1965
193	Paulo Simões Ribeiro	Setúbal	PSD	28-03-1969
194	Pedro Alves	Viseu	PSD	30-12-1972
195	Pedro Delgado Alves	Lisboa	PS	12-12-1980
196	Pedro do ó Ramos	Setúbal	PSD	04-10-1974
197	Pedro Farmhouse	Lisboa	PS	27-06-1961
198	Pedro Filipe Soares	Aveiro	BE	15-02-1979
199	Pedro Jesus Marques	Portalegre	PS	01-08-1976
200	Pedro Lynce	Évora	PSD	06-02-1943
201	Pedro Nuno Santos	Aveiro	PS	13-04-1977
202	Pedro Pimpão	Leiria	PSD	21-07-1980
203	Pedro Pinto	Lisboa	PSD	24-10-1956
204	Pedro Roque	Faro	PSD	10-01-1963
205	Pedro Silva Pereira	Vila Real	PS	15-08-1962
206	Ramos Preto	Lisboa	PS	19-01-1956
207	Raúl de Almeida	Aveiro	CDS-PP	08-05-1968
208	Renato Sampaio	Porto	PS	03-05-1952
209	Ricardo Baptista Leite	Lisboa	PSD	31-05-1980
210	Ricardo Rodrigues	Açores	PS	01-06-1958
211	Rita Rato	Lisboa	PCP	05-01-1983
212	Rosa Arezes	Viana Castelo	PSD	22-05-1967
213	Rosa Maria Bastos Albernaz	Aveiro	PS	04-09-1947
214	Rui Jorge Santos	Vila Real	PS	24-01-1969
215	Rui Paulo Figueiredo	Lisboa	PS	09-10-1972
216	Rui Pedro Duarte	Coimbra	PS	31-08-1984
217	Sérgio Azevedo	Lisboa	PSD	30-10-1981
218	Sérgio Sousa Pinto	Aveiro	PS	29-07-1972
219	Sónia Fertuzinhos	Braga	PS	12-01-1973
220	Telmo Correia	Braga	CDS-PP	04-02-1960
221	Teresa Anjinho	Aveiro	CDS-PP	03-10-1974
222	Teresa Caeiro	Lisboa	CDS-PP	14-02-1969
223	Teresa Costa Santos	Viseu	PSD	10-03-1966
224	Teresa Leal Coelho	Porto	PSD	29-03-1961
225	Ulisses Pereira	Aveiro	PSD	04-02-1954
226	Valter Ribeiro	Leiria	PSD	19-06-1974
227	Vasco Cunha	Santarém	PSD	23-03-1965
228	Vera Rodrigues	Porto	CDS-PP	08-02-1981
229	Vieira da Silva	Setúbal	PS	14-02-1953
230	Vitalino Canas	Lisboa	PS	14-07-1959