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 seleccionar os ingredientes, que serão depois manipulados de acordo com determinada receita. O resultado do cozinhado pode ser desastroso, embora de aspecto agradável. Efectivamente se os ingredientes não estiverem em condições, resulta um prato de aspecto 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.

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

[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.

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 exactamente 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 correcto. 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.

População, 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:

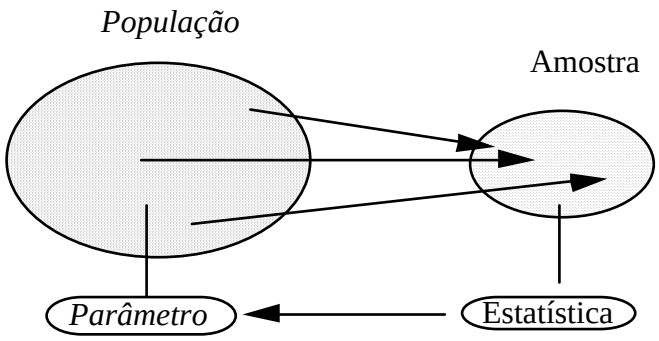
- **idade média** dos potenciais eleitores que estão decididos a votar;
- **percentagem** de eleitores que estão decididos a votar.

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 dissemos 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 mesma 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 negrito, 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 4.º 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 representavam, 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 n.º 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:

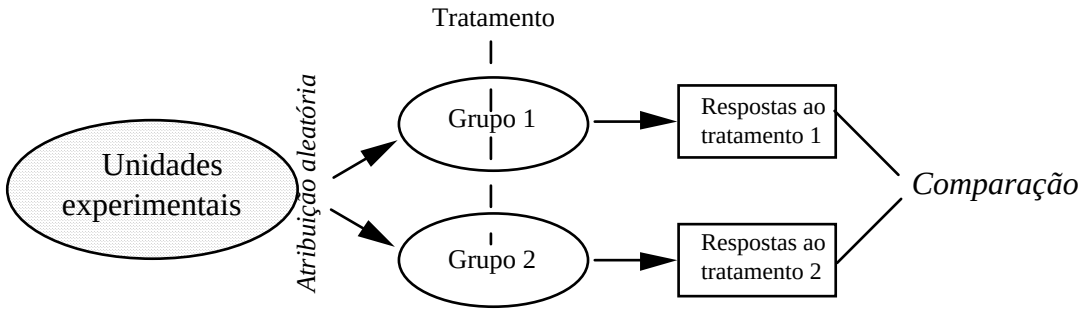
- Unidades experimentais – 22000 médicos

- 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 observáveis, 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 seleccionada. Uma amostra destas pode ser escolhida sequencialmente da população, escolhendo um elemento de cada vez, sem reposição, pelo que em cada selecção cada elemento tem a mesma probabilidade de ser seleccionado. Um esquema de amostragem aleatória simples, conduz a que cada elemento da População tenha a mesma probabilidade de ser seleccionado para a amostra. No entanto existem outros esquemas de amostragem em que cada elemento tem igual probabilidade de ser seleccionado, sem que cada conjunto de n elementos tenha a mesma probabilidade de ser seleccionado. É o que se passa com a amostragem aleatória sistemática, de que falaremos adiante.

Amostragem com reposição

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

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

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

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

Exemplo 1.3.1.1 – Consideremos a população constituída pelos 18 alunos de uma turma do 10.º 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

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 uma tabela de números aleatórios.

Dígitos aleatórios – Uma tabela de dígitos aleatórios é uma listagem dos dígitos 0, 1, 2, 3, 4, 5, 6, 7, 8 ou 9 tal que:
- qualquer um dos destes dígitos tem igual possibilidade de figurar em qualquer posição da lista;
- a posição em que figura cada dígito é independente das posições dos outros dígitos.

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

Linha								
101	19223	95034	05756	28713	96409	12531	42544	82853
102	73676	47150	99400	01927	27754	42648	82425	36290
103	45467	71709	77558	00095	32863	29485	82226	90056
104	52711	38889	93074	60227	40011	85848	48767	52573
105	95592	94007	69971	91481	60779	53791	17297	59335
106	68417	35013	15529	72765	85089	57067	50211	47487
107	82739	57890	20807	47511	81676	55300	94383	14893

108	60940	72024	17868	24943	61790	90656	87964	18883
109	36009	19365	15412	39638	85453	46816	83485	41979

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

Para seleccionar uma amostra de uma população utilizando a tabela procede-se em duas etapas:

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

Exemplo 1.3.1.1 (cont) - Considerando a população do exemplo anterior, constituída por 18 elementos, vamos numerá-los com os números 01, 02, 03, …, 17, 18 (podia ser utilizado qualquer outro conjunto de 18 números de 2 dígitos). Para seleccionar uma amostra de dimensão 4 fixamo-nos numa linha qualquer da tabela, por exemplo a linha 107 e começamos a seleccionar os números de dois dígitos, tendo-se obtido:

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

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

Como obter uma tabela de números aleatórios?

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

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

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 (X) 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). 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. Inserimos ainda uma coluna com identificação do sexo (coluna E). Apresentamos a seguir uma pequena parcela desse ficheiro, a que chamámos *Deputados.xls*. 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 Parl.**	**Circulo Eleitoral**	**Sexo**	**Data nas.**
2	1	Abel Lima Baptista	CDS-PP	Viana do C	M	13-10-1963
3	2	Adão José Fonseca Silva	PSD	Bragança	M	01-10-1957
4	3	Agostinho Correia Branquinho	PSD	Porto	M	10-08-1956
5	4	Agostinho Moreira Gonçalves	PS	Porto	M	15-07-1952
6	5	Agostinho Nuno de Azevedo Ferreira Lo	PCP	Braga	M	16-11-1944
7	6	Alberto Arons Braga de Carvalho	PS	Setúbal	M	20-09-1949
8	7	Alberto de Sousa Martins	PS	Porto	M	25-04-1945
9	8	Alberto Marques Antunes	PS	Setúbal	M	03-04-1949
10	9	Alcídia Maria Cruz Sousa de Oliveira Lo	PS	Porto	F	09-01-1974
11	10	Alda Maria Gonçalves Pereira Macedo	BE	Porto	F	07-09-1954
12	11	Aldemira Maria Cabanita do Nascimento	PS	Faro	F	04-04-1952

Sheet2 Sheet1

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

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 Lima Baptista	=RAND()
3	2	Adão José Fonseca Silva	=RAND()
4	3	Agostinho Correia Branquinho	=RAND()
5	4	Agostinho Moreira Gonçalves	=RAND()
6	5	Agostinho Nuno de Azevedo Ferreira Lopes	=RAND()
7	6	Alberto Arons Braga de Carvalho	=RAND()
8	7	Alberto de Sousa Martins	=RAND()
9	8	Alberto Marques Antunes	=RAND()
10	9	Alcídia Maria Cruz Sousa de Oliveira Lopes	=RAND()
11	10	Alda Maria Gonçalves Pereira Macedo	=RAND()
12	11	Aldemira Maria Cabanita do Nascimento Bi	=RAND()
13	12	Ana Catarina Veiga Santos Mendonça Men	=RAND()
14	13	Ana Isabel Drago Lobato	=RAND()

Sheet2 Sheet1

Para visualizar as fórmulas na folha de Excel, bastou seleccionar:

Tools
Options
View
Formulas
Ok:

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

	A	B	J
1		Nome	
2		Abel Lima Baptista	0,1494229
3		Adão José Fonseca S	0,9789825
4		Agostinho Correia Br	0,339507
5		Agostinho Moreira Gc	0,7098311
6		Agostinho Nuno de A	0,1882448
7		Alberto Arons Braga	0,5993157
8		Alberto de Sousa Mar	0,1543557
9		Alberto Marques Antu	0,1041103
10		Alcídia Maria Cruz So	0,5565095
11		Alda Maria Gonçalves	0,274581
12		Aldemira Maria Caba	0,8644202
13		Ana Catarina Veiga S	0,8629732

Sheet2 She

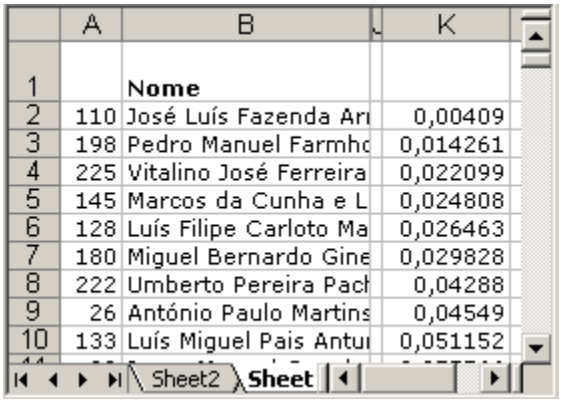
	A	B	J	K
1		Nome		
2		Abel Lima Baptista	0,6117577	0,149423
3		Adão José Fonseca S	0,9961447	0,978983
4		Agostinho Correia Br	0,2126522	0,339507
5		Agostinho Moreira Gc	0,6421697	0,709831
6		Agostinho Nuno de A	0,3761691	0,188245
7		Alberto Arons Braga	0,375953	0,599316
8		Alberto de Sousa Mar	0,462458	0,154356
9		Alberto Marques Antu	0,0394435	0,10411
10		Alcídia Maria Cruz So	0,5813319	0,556509
11		Alda Maria Gonçalves	0,7022381	0,274581
12		Aldemira Maria Caba	0,87367	0,86442
13		Ana Catarina Veiga S	0,9083077	0,862973

Sheet2 Sheet1

Colámos os valores na coluna K e fizemos 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	110	José Luís Fazenda Ari	0,00409
3	198	Pedro Manuel Farmho	0,014261
4	225	Vitalino José Ferreira	0,022099
5	145	Marcos da Cunha e L	0,024808
6	128	Luís Filipe Carloto Ma	0,026463
7	180	Miguel Bernardo Gine	0,029828
8	222	Umberto Pereira Pach	0,04288
9	26	António Paulo Martins	0,04549
10	133	Luís Miguel Pais Antur	0,051152

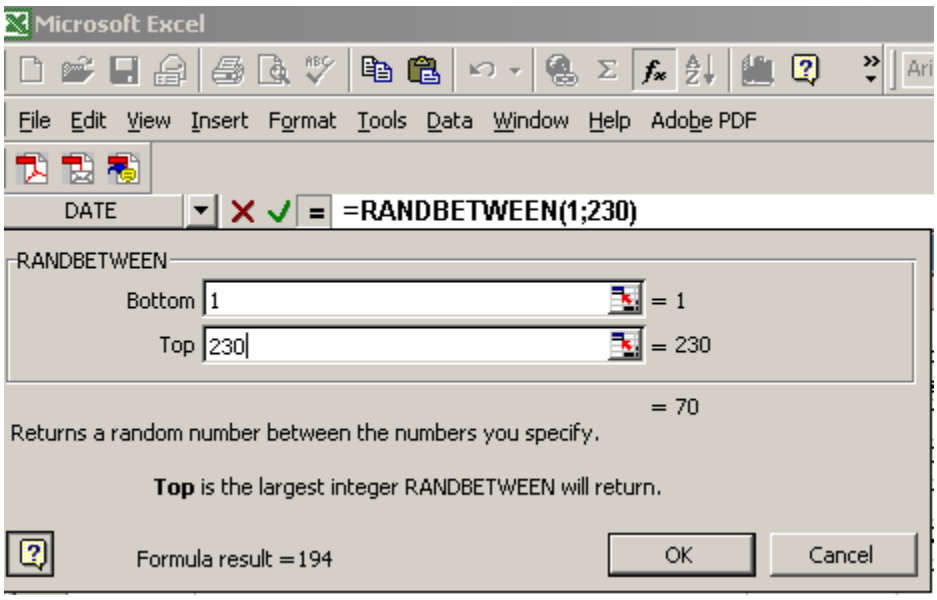
Sheet2 Sheet

Os deputados seleccionados foram os números 110, 198, 225, 145, 128, 180, 222, 26 e 133.

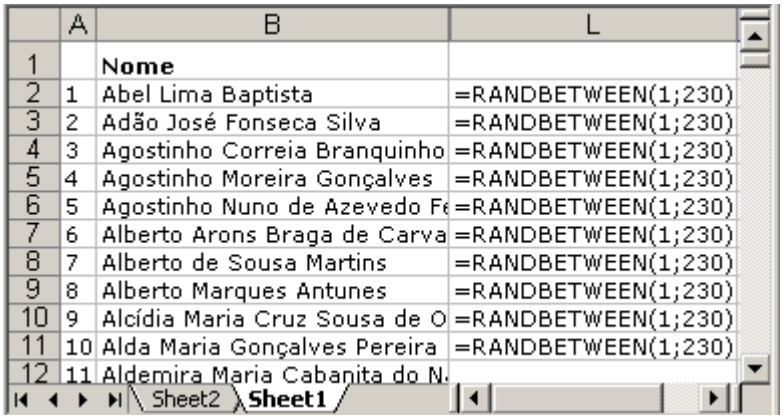
Nota: Embora os números anteriores sejam referidos como aleatórios, convém ter presente que os números que se obtêm são *pseudo-aleatórios*, já que é um mecanismo determinista que lhes dá origem. No entanto comportam-se como números aleatórios (passam uma bateria de testes destinados a confirmar a sua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

Amostra aleatória com reposição

a) *Utilize a função RANDBETWEEN(), para obter números pseudo-aleatórios entre 1 e 230, para simular a extracção de uma amostra aleatória, da população dos deputados*.



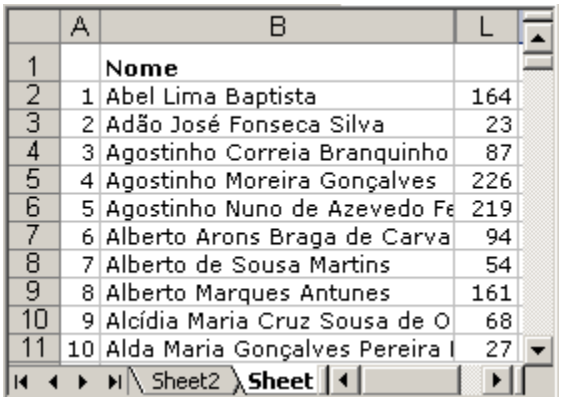
Esta função devolve um número pseudo-aleatório entre os limites especificados nos argumentos. Como pretendemos seleccionar uma amostra de dimensão 10, replicamos a fórmula anterior por 10 células, na coluna L, como se apresenta a seguir:



	A	B	L
1		**Nome**	
2	1	Abel Lima Baptista	=RANDBETWEEN(1;230)
3	2	Adão José Fonseca Silva	=RANDBETWEEN(1;230)
4	3	Agostinho Correia Branquinho	=RANDBETWEEN(1;230)
5	4	Agostinho Moreira Gonçalves	=RANDBETWEEN(1;230)
6	5	Agostinho Nuno de Azevedo Fe	=RANDBETWEEN(1;230)
7	6	Alberto Arons Braga de Carva	=RANDBETWEEN(1;230)
8	7	Alberto de Sousa Martins	=RANDBETWEEN(1;230)
9	8	Alberto Marques Antunes	=RANDBETWEEN(1;230)
10	9	Alcídia Maria Cruz Sousa de O	=RANDBETWEEN(1;230)
11	10	Alda Maria Gonçalves Pereira	=RANDBETWEEN(1;230)
12	11	Aldemira Maria Cabanita do N	

Sheet2 Sheet1

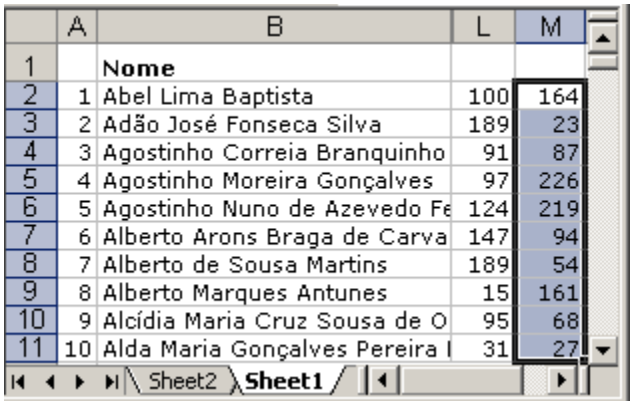
A amostra obtida é constituída pelos deputados com os 10 números nas células L2, ..., L11:



	A	B	L
1		**Nome**	
2	1	Abel Lima Baptista	164
3	2	Adão José Fonseca Silva	23
4	3	Agostinho Correia Branquinho	87
5	4	Agostinho Moreira Gonçalves	226
6	5	Agostinho Nuno de Azevedo Fe	219
7	6	Alberto Arons Braga de Carva	94
8	7	Alberto de Sousa Martins	54
9	8	Alberto Marques Antunes	161
10	9	Alcídia Maria Cruz Sousa de O	68
11	10	Alda Maria Gonçalves Pereira I	27

Sheet2 Sheet

Uma vez que a função *RANDBETWEEN* é uma função volátil, isto é, muda quando se recalcula a folha, para ficar com os valores gerados fomos ao *Edit →Paste Special→Values*, como se indica a seguir:



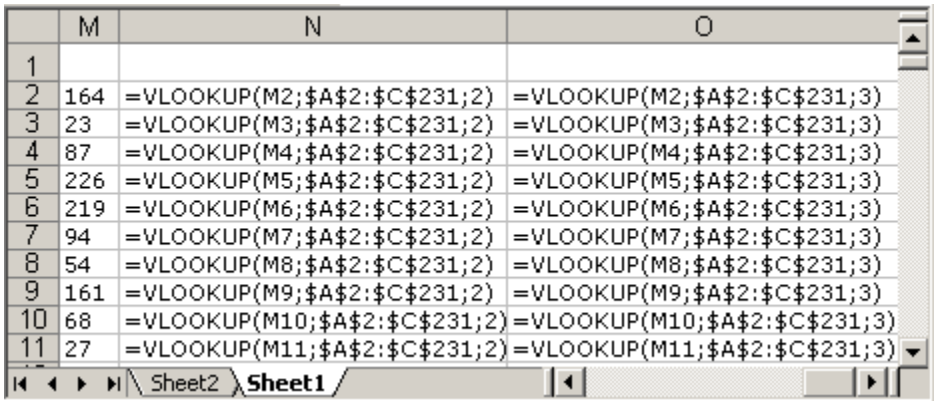
	A	B	L	M
1		**Nome**		
2	1	Abel Lima Baptista	100	164
3	2	Adão José Fonseca Silva	189	23
4	3	Agostinho Correia Branquinho	91	87
5	4	Agostinho Moreira Gonçalves	97	226
6	5	Agostinho Nuno de Azevedo Fe	124	219
7	6	Alberto Arons Braga de Carva	147	94
8	7	Alberto de Sousa Martins	189	54
9	8	Alberto Marques Antunes	15	161
10	9	Alcídia Maria Cruz Sousa de O	95	68
11	10	Alda Maria Gonçalves Pereira I	31	27

Sheet2 Sheet1

Colámos os valores na coluna M e fizemos o *Save*. Repare-se que os valores que estavam inicialmente na coluna L foram alterados, dando origem a uma nova amostra (devido ao facto da função *RANDBETWEEN* ser volátil, como referimos anteriormente):

b) *Da tabela dos deputados, seleccione o nome e o grupo parlamentar dos deputados cujo número seja um dos elementos da amostra obtida anteriormente.*

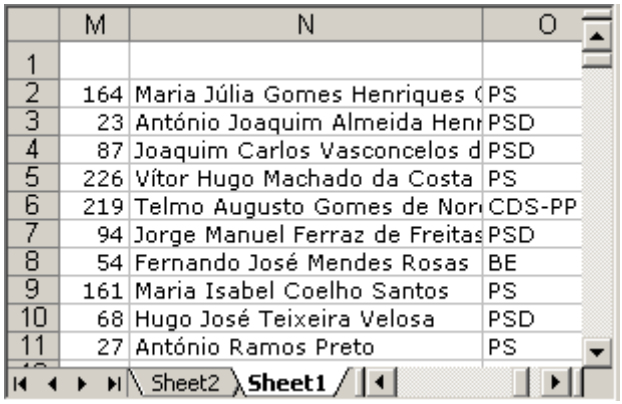
Para seleccionar o nome e o grupo parlamentar dos deputados correspondentes aos 10 números obtidos, vamos utilizar uma função do Excel, a função *VLOOKUP*, do seguinte modo:



	M	N	O
1			
2	164	=VLOOKUP(M2;$A$2:$C$231;2)	=VLOOKUP(M2;$A$2:$C$231;3)
3	23	=VLOOKUP(M3;$A$2:$C$231;2)	=VLOOKUP(M3;$A$2:$C$231;3)
4	87	=VLOOKUP(M4;$A$2:$C$231;2)	=VLOOKUP(M4;$A$2:$C$231;3)
5	226	=VLOOKUP(M5;$A$2:$C$231;2)	=VLOOKUP(M5;$A$2:$C$231;3)
6	219	=VLOOKUP(M6;$A$2:$C$231;2)	=VLOOKUP(M6;$A$2:$C$231;3)
7	94	=VLOOKUP(M7;$A$2:$C$231;2)	=VLOOKUP(M7;$A$2:$C$231;3)
8	54	=VLOOKUP(M8;$A$2:$C$231;2)	=VLOOKUP(M8;$A$2:$C$231;3)
9	161	=VLOOKUP(M9;$A$2:$C$231;2)	=VLOOKUP(M9;$A$2:$C$231;3)
10	68	=VLOOKUP(M10;$A$2:$C$231;2)	=VLOOKUP(M10;$A$2:$C$231;3)
11	27	=VLOOKUP(M11;$A$2:$C$231;2)	=VLOOKUP(M11;$A$2:$C$231;3)

Sheet2 Sheet1

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 que está na coluna M, obtendo-se a seguinte amostra:



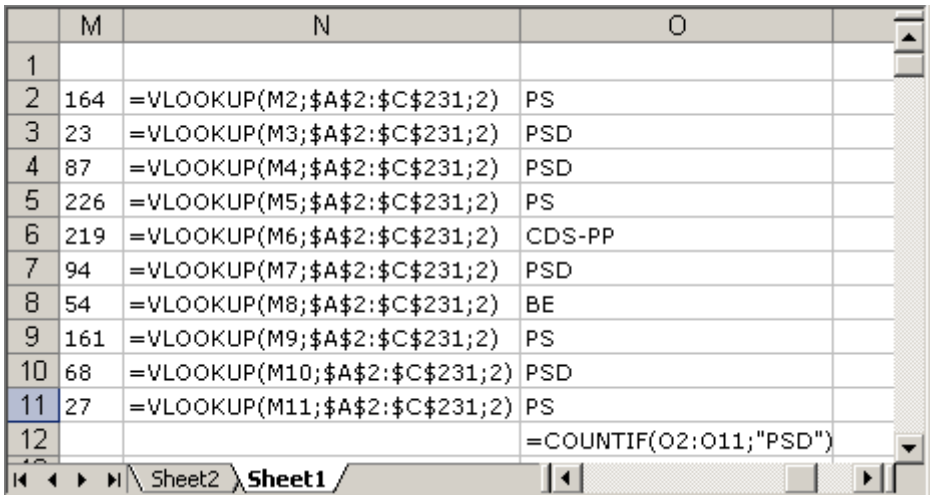
	M	N	O
1			
2	164	Maria Júlia Gomes Henriques (	PS
3	23	António Joaquim Almeida Henr	PSD
4	87	Joaquim Carlos Vasconcelos d	PSD
5	226	Vítor Hugo Machado da Costa	PS
6	219	Telmo Augusto Gomes de Nor	CDS-PP
7	94	Jorge Manuel Ferraz de Freitas	PSD
8	54	Fernando José Mendes Rosas	BE
9	161	Maria Isabel Coelho Santos	PS
10	68	Hugo José Teixeira Velosa	PSD
11	27	António Ramos Preto	PS

Sheet2 Sheet1

2. Parâmetro e Estatística.

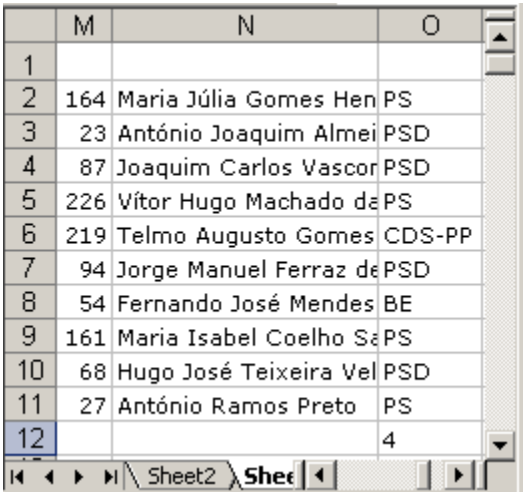
c) *Calcule a percentagem de deputados do grupo parlamentar PSD, na amostra obtida*.

Vamos começar por utilizar a função *COUNTIF*, que inserimos na célula O12, 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":



	M	N	O
1			
2	164	=VLOOKUP(M2;$A$2:$C$231;2)	PS
3	23	=VLOOKUP(M3;$A$2:$C$231;2)	PSD
4	87	=VLOOKUP(M4;$A$2:$C$231;2)	PSD
5	226	=VLOOKUP(M5;$A$2:$C$231;2)	PS
6	219	=VLOOKUP(M6;$A$2:$C$231;2)	CDS-PP
7	94	=VLOOKUP(M7;$A$2:$C$231;2)	PSD
8	54	=VLOOKUP(M8;$A$2:$C$231;2)	BE
9	161	=VLOOKUP(M9;$A$2:$C$231;2)	PS
10	68	=VLOOKUP(M10;$A$2:$C$231;2)	PSD
11	27	=VLOOKUP(M11;$A$2:$C$231;2)	PS
12			=COUNTIF(O2:O11;"PSD")

Sheet2 Sheet1



	M	N	O
1			
2	164	Maria Júlia Gomes Hen	PS
3	23	António Joaquim Almei	PSD
4	87	Joaquim Carlos Vascor	PSD
5	226	Vítor Hugo Machado da	PS
6	219	Telmo Augusto Gomes	CDS-PP
7	94	Jorge Manuel Ferraz de	PSD
8	54	Fernando José Mendes	BE
9	161	Maria Isabel Coelho Sa	PS
10	68	Hugo José Teixeira Vel	PSD
11	27	António Ramos Preto	PS
12			4

Sheet2 Shee

Obtivemos o valor 4 para a frequência absoluta de deputados do PSD. Como o nº de deputados da amostra era 10, a percentagem de deputados do grupo parlamentar do PSD, na amostra é de 40%. Este valor é uma **estatística** – característica numérica da amostra. Utiliza-se como estimativa do **parâmetro** "percentagem de deputados do PSD na população em estudo" – característica numérica da população.

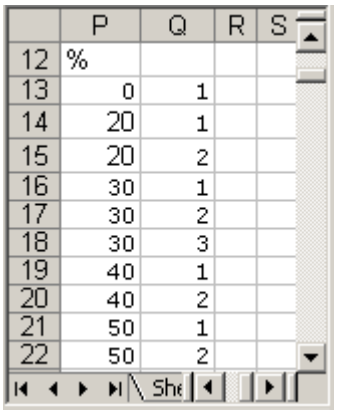
3. Variabilidade amostral

d) Repita 10 vezes o processo descrito nas alíneas anteriores e registe numa tabela os resultados obtidos.

Gerámos 10 amostras e obtivemos 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%	20%	30%	50%	20%	30%	40%	50%	0%	30%

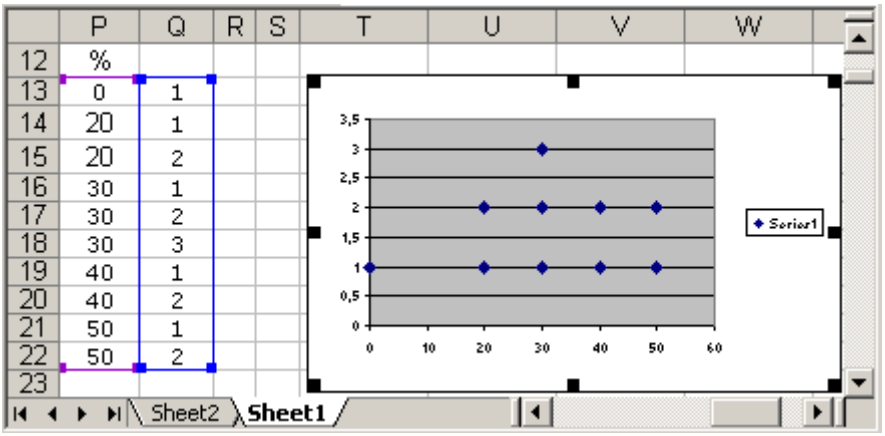
Repare-se na variabilidade apresentada nos resultados obtidos para as diferentes amostras. Os 10 valores obtidos para a percentagem de deputados do PSD existentes em cada uma delas, representam outras tantas estimativas para a verdadeira proporção de deputados existentes na População. Iremos ilustrar 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	R	S
12	%			
13	0	1		
14	20	1		
15	20	2		
16	30	1		
17	30	2		
18	30	3		
19	40	1		
20	40	2		
21	50	1		
22	50	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, dispuséssemos de uma tabela de frequências, para construir esta nova tabela, bastaria repetir cada elemento da amostra, tantas vezes quantas a sua frequência absoluta.

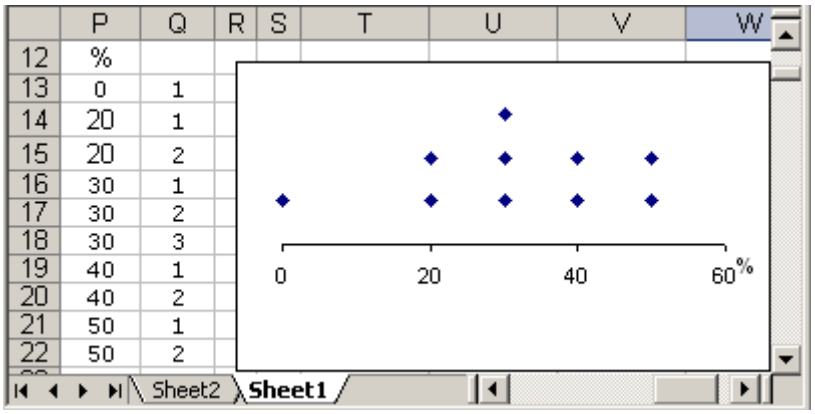
Na folha do Excel, seleccionam-se as duas colunas e no menu *Chart* selecciona-se *Scatter* e o primeiro subtipo desta representação. Obtém-se o diagrama de pontos com o seguinte aspecto:



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

Seleccionar: Legenda e carregar no botão *Delete*;
As linhas e carregar no botão *Delete*;
O fundo cinzento e carregar no botão *Delete*;
O eixo dos YY e carregar no botão *Delete*;

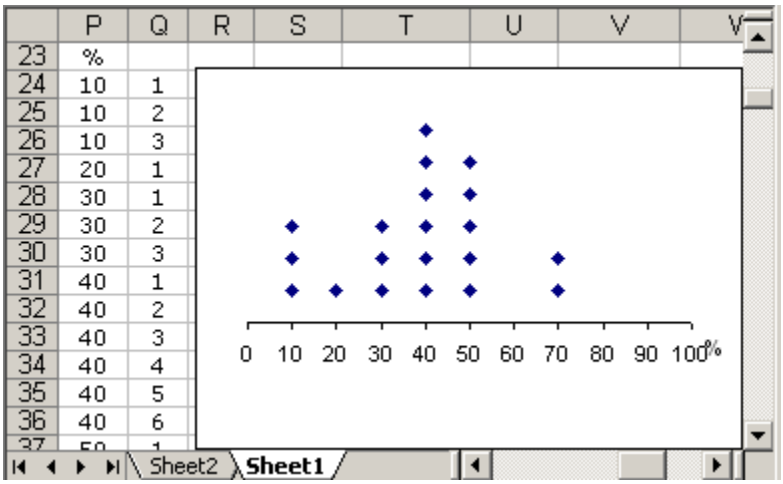
Temos finalmente a seguinte representação:



Da representação gráfica anterior começamos a adivinhar que a distribuição das estimativas apresenta um padrão com uma certa simetria relativamente ao valor de 30%.

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, calculámos a percentagem de deputados do PSD em cada uma delas e com os resultados obtidos construímos a seguinte representação:

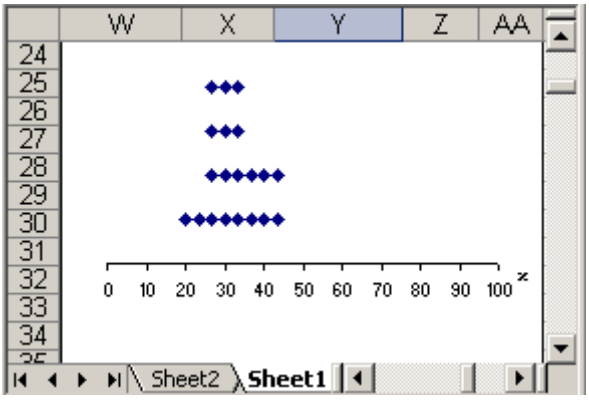


Esta representação é mais elucidativa e reforça a ideia avançada anteriormente, de que o valor do parâmetro em estudo – percentagem de deputados do PSD, se deve situar entre os valores 30% e 40%. Tendo em conta que a verdadeira percentagem de deputados do PSD na população é 32,6%, apesar de o valor apresentado 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

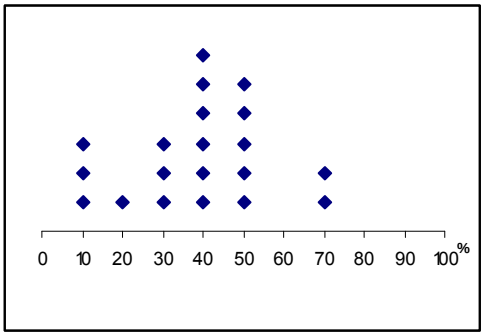
f) Considere agora 20 amostras de dimensão 30, calcule para cada uma o valor da estatística em estudo, e construa o diagrama de pontos dos valores obtidos. Compare a representação obtida, com a que obteve na alínea e).

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:

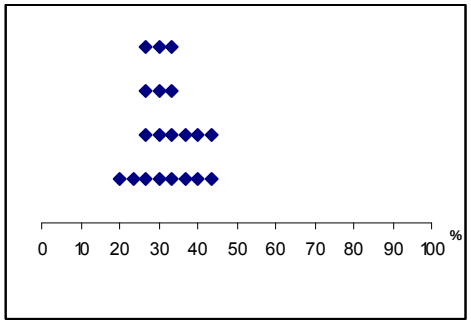


Comparando as duas representações obtidas quando se consideram amostras de dimensão 10 ou de dimensão 30, verificamos que a variabilidade apresentada pelos valores da estatística - percentagem de deputados do PSD, no caso das amostras de maior dimensão, é inferior à apresentada pela estatística no caso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

Percentagem de deputados do PSD em amostras de dimensão 10



Percentagem de deputados do PSD em amostras de dimensão 30



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 apresentar 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 Deputados.xls, 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.

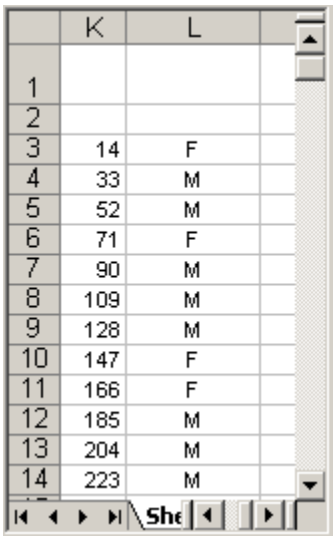
Temos uma população de dimensão 230 e pretendemos obter uma amostra de dimensão 12. Vamos utilizar a seguinte metodologia:

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

Passo 2 – Dos primeiros 19 elementos da lista ordenada dos deputados, vamos seleccionar um elemento ao acaso. Vimos na secção anterior que basta utilizar a função Randbetween(1;19), que inserimos na célula K3. A utilização desta função devolveu-nos o deputado número 14.

Passo 3 – A amostra será constituída pelos deputados números 14, 33, 52, 71, 90, 109, 128, 147, 166, 185, 204, 223, que obtivemos adicionando sucessivamente 19, até obtermos 12 elementos (células K3:K14).

Passo 4 - Utilizando a função VLOOKUP(K3;$A$3:$E$232;5), replicada pelas 12 células L3:L14, obteve-se finalmente a informação solicitada, constituída pelo sexo dos 12 deputados seleccionados para a amostra:



	K	L
1		
2		
3	14	F
4	33	M
5	52	M
6	71	F
7	90	M
8	109	M
9	128	M
10	147	F
11	166	F
12	185	M
13	204	M
14	223	M

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:

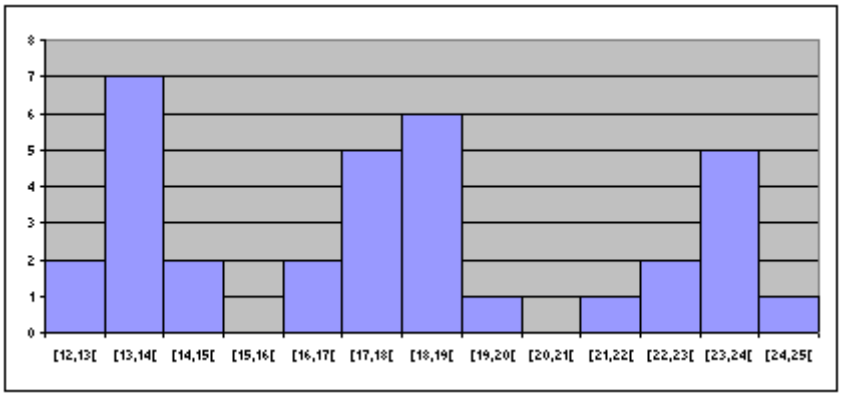
N.º	6	7	8	9	10	26	27	28	29	30
Freq.	4	4	4	4	4	4	4	4	4	4
Cor	Ver	Ver	Ver	Ver	Ver	Preto	Preto	Preto	Preto	Preto

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



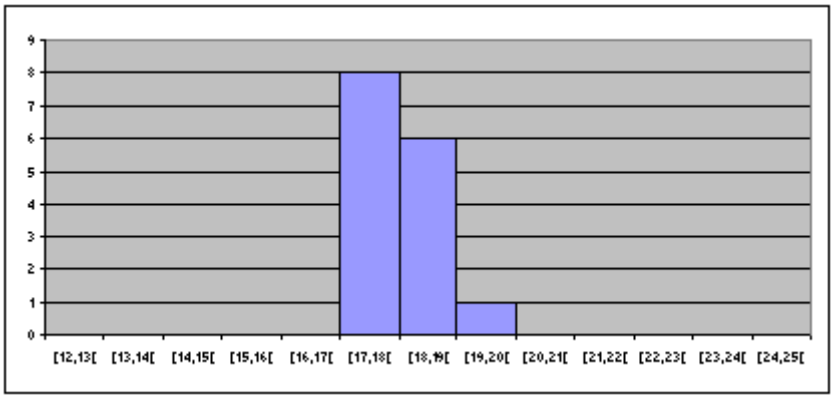
Esta distribuição não nos ajuda muito a dizer qual a estimativa para o valor médio da população (média dos números inscritos). Gostaríamos de ter obtido para a amostra, cujos elementos são as diferentes médias, uma distribuição com pouca variabilidade, para podermos argumentar que a média destes elementos 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*.

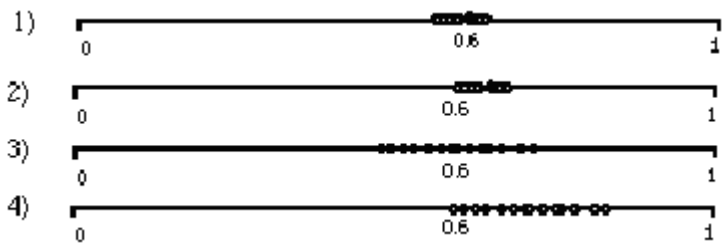
A média das estimativas calculadas a partir da expressão s^2 acima considerada, coincide com a variância σ^2.

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 - *percentagem* 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:



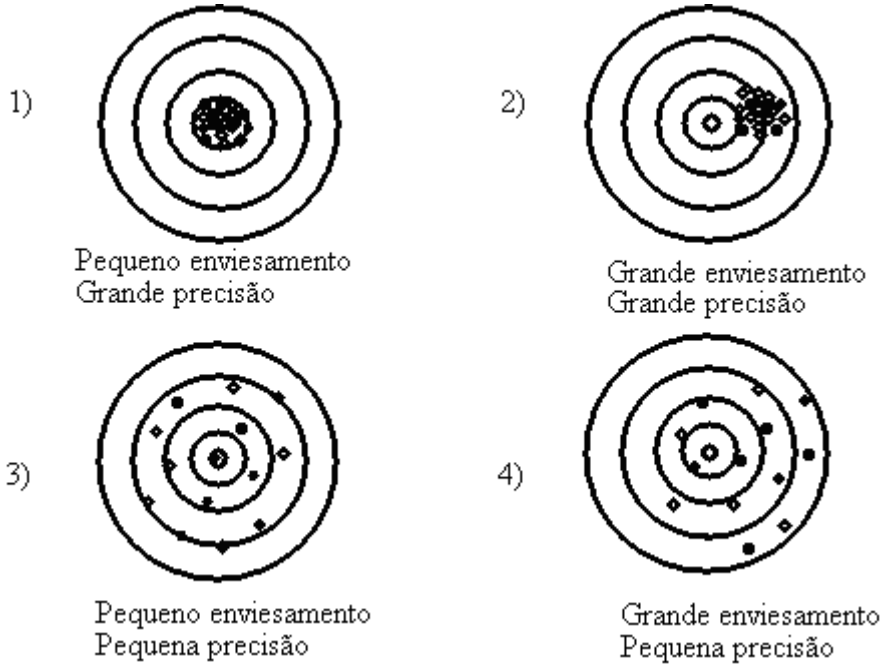
1) reflecte um *pequeno* ou ausência de *enviesamento*, pois os valores para a estatística (proporções obtidas a partir das amostras) situam-se para um e outro lado do valor do parâmetro, e verifica-se ainda a existência de uma pequena variabilidade entre os resultados obtidos para as várias amostras,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

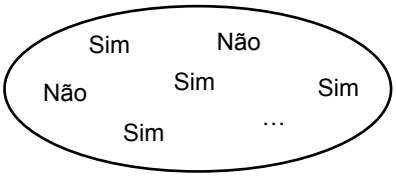
imediatamente concluirá que metade das notas são iguais ou superiores a 15, pelo que se pode considerar um aluno bom. Organizámos os dados através de uma representação gráfica sugestiva, que permitiu realçar a informação desejada. Outro processo seria resumir a informação sob a forma de uma medida que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

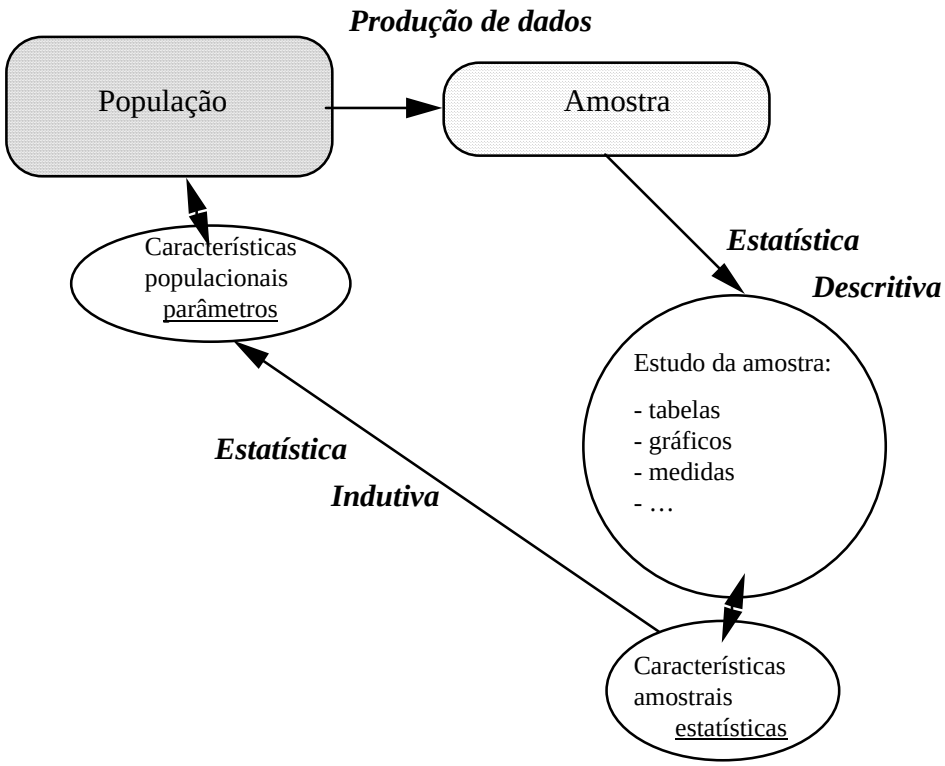
Poderemos agora inferir que 63.5% dos eleitores da cidade do Porto pensam votar no Senhor X? A resposta a esta pergunta nem é sim, nem não, mas talvez. É agora que temos necessidade de utilizar o conceito de Probabilidade, para quantificar a incerteza associada à inferência. Assim, existem processos 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 as 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.