Introdução à amostragem

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.

[1] Esta secção segue de perto o texto Introdução às Probabilidades e Estatística 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 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:

- **Aquisição de dados**
- **Análise dos dados**
- **Inferência a partir dos dados**

Neste módulo vamos abordar o primeiro tema considerado, ou seja o que diz respeito à Aquisição de Dados, numa perspectiva em que pretendemos obter dados, a partir dos quais seja 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.

2 - Aquisição de dados: sondagens. 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.

De entre as técnicas de aquisição de dados, que se enquadram nesta última situação, distinguem-se as

Sondagens e **Experimentações (aleatoriezadas)**

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*".

2.1 - Sondagens. População e amostra. Parâmetro e Estatística.

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 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.

Dimensão da amostra – número de elementos da amostra.

Dissemos anteriormente que uma População é um conjunto de indivíduos (não necessariamente pessoas), com algumas características comuns, que se pretendem estudar. A uma característica comum, que possa assumir valores diferentes de indivíduo para indivíduo, chamamos *variável*. Sendo então o nosso objectivo o estudo de uma (ou mais) característica da População, vamos identificar População com a *variável* que se está a estudar, dizendo que a População é constituída por todos os valores que a *variável* pode assumir. Por exemplo, relativamente à população constituída pelos portugueses adultos, se o objectivo do nosso estudo for a característica *Peso*, diremos que a População é constituída por todos os valores possíveis para a variável *Peso*. Do mesmo modo identificaremos amostra com os valores observados para a variável em estudo, sobre alguns elementos da População. Assim, na continuação do exemplo referido, os valores 71 kg, 82 kg, 79 kg, 90 kg, 63 kg, 101 kg, obtidos ao pesar 6 homens portugueses, constituem uma amostra da população a estudar.

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

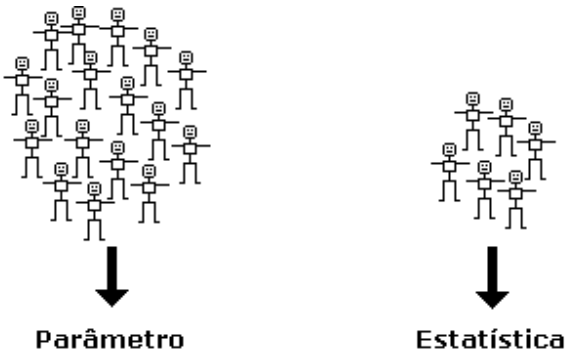
Por exemplo, ao estudar a população constituída por todos os potenciais eleitores para as legislativas, dois parâmetros que podem ter interesse são:

- **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 que se calculam a partir dos valores da amostra. 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.

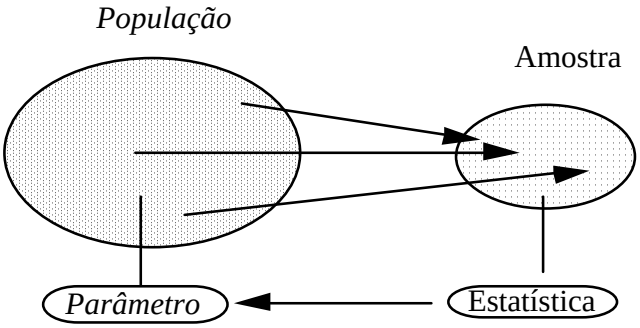
No caso do exemplo anterior, se estivermos interessados em estimar o **parâmetro** ou proporção populacional "*percentagem de eleitores que estão decididos a votar*" através de amostras de dimensão 1000, o **estimador** será a proporção amostral "*percentagem de eleitores, em 1000, que interrogados disserem estar decididos a votar*". Quando se efectivar a recolha de uma amostra (de dimensão 1000) e se, por exemplo, se concluir que 578 eleitores estão decididos a votar, então uma **estimativa** do parâmetro em estudo é 57,8%. À estimativa também se chama **estatística**. Assim, dependerá do contexto, interpretar a palavra estatística como uma função dos valores da amostra (estimador) ou já o valor observado dessa função para uma determinada amostra (estimativa). É nesta perspectiva que se pode dizer que:

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



Estas quantidades são conceptualmente distintas, pois enquanto a característica populacional – *parâmetro*, pode ser considerada um valor exacto, embora (quase sempre) desconhecido, a característica amostral – *estatística*, pode ser calculada,

embora difira de amostra para amostra, mas que todavia pode ser considerada uma estimativa útil da característica populacional respectiva.



Para que os *estimadores* forneçam estimativas úteis, é necessário que as amostras utilizadas, para obter essas estimativas, sejam representativas das populações de onde foram retiradas.

Um **Estimador** é uma função dos elementos da amostra, que se utiliza para estimar parâmetros. Ao valor do estimador calculado para uma amostra que se recolheu, dá-se o nome de **Estimativa.**

Nota – Do que dissemos anteriormente, não esquecer que a palavra **estatística** pode ser utilizada no sentido de estimador ou de estimativa. Tem assim de se tomar atenção ao contexto em que está a ser utilizada.

Exemplos

1. Se estivermos interessados em estudar a média obtida no exame nacional de Matemática, no ano lectivo 2006-2007, então a população a estudar é constituída por todos os alunos que fizeram o exame nacional de Matemática nesse ano lectivo. Estamos interessados em conhecer o valor do parâmetro - **valor médio** da variável *Nota do exame nacional de Matemática*. Para obter uma estimativa deste parâmetro, seleccionam-se alguns alunos que tenham feito o exame, regista-se a nota obtida por cada um e calcula-se a média dessas notas. O valor obtido é uma estimativa do parâmetro desconhecido. Por exemplo, se seleccionarmos 10 alunos e as notas obtidas por esses 10 alunos fossem (numa escala de 0 a 200):

   125, 97, 58, 29, 101, 65, 107, 37, 29, 127

   então uma estimativa para o parâmetro valor médio das notas no exame de Matemática seria $77{,}5=\frac{125+97+58+29+101+65+107+37+29+127}{10}$. O valor 77,5, calculado a partir dos dados da amostra, é uma estimativa. Se seleccionássemos outra amostra de 10 alunos, as notas seriam diferentes e o valor do estimador Média também viria diferente, dando uma estimativa diferente da obtida anteriormente.

2. Suponhamos que em vez da média no exame nacional de Matemática, estávamos interessados em conhecer a proporção de positivas. O parâmetro desconhecido seria agora esta **proporção**. Utilizando a mesma amostra do exemplo anterior, uma estimativa para a proporção (populacional) de positivas no exame de matemática, será a proporção (amostral) de positivas na amostra, ou seja 40%.

3. O gestor de uma agência bancária pretende saber qual o tempo médio que as pessoas esperam para serem atendidas, durante o período de uma hora, após a abertura da agência (entre as 8h30m e as 9h30m). Identificando a população com a variável em estudo, podemos dizer que a população é constituída pelos tempos de espera de todos os possíveis clientes da agência, desde que chegam até serem atendidos, durante aquele período. Para estimar o parâmetro **tempo médio** de espera, pode-se recolher uma amostra de tempos de espera de alguns clientes e calcular a média desses tempos. Por exemplo se os tempos (em minutos) observados em 10 clientes, escolhidos ao acaso, foram

   8, 5, 12, 7, 9, 5, 4, 5, 4, 4

   então uma estimativa para o tempo médio de espera, durante o período considerado, será 6,3 minutos (média dos valores anteriores).

4. A lista X candidata a dirigir a Associação de estudantes de uma dada universidade pretende saber se terá a maioria de votos nas próximas eleições, que se avizinham. Assim, pediu ao Departamento de Estatística da sua Universidade que realizassem uma sondagem que lhes permitisse ter uma ideia do que os esperaria se se candidatassem. O Departamento de Estatística procedeu à recolha de uma amostra de 150 estudantes, potenciais eleitores, a quem perguntou se pensavam votar na lista X. Dos 150 inquiridos, 87 responderam que sim, pelo que uma estimativa para a **proporção** de alunos que pensa votar na lista X é de 58%.

5. O Conselho executivo da escola pretende reivindicar uma melhoria nos transportes públicos, alegando que os alunos esperam muito tempo, na paragem, quando saem da parte da tarde. Assim, encarregou um grupo de alunos para, entre as 15 e as 19 horas, durante alguns dias, registarem os tempos entre passagens sucessivas dos autocarros da carreira que serve a escola. A média dos valores registados fornecerá uma estimativa para o tempo médio entre as passagens dos autocarros da carreira. Se a média obtida for superior ao estipulado pela Carris, então haverá efectivamente lugar para a reivindicação.

3 – 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.

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.

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.

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 decorria uma sondagem de opinião sobre se as pessoas eram a favor da implementação da pena de morte. Esta recolha de opiniões era feita no molde descrito no exemplo anterior, isto é, por resposta voluntária.
  Resultado - A utilização da percentagem de SIM´s, que naturalmente se espera elevada, dá uma indicação errada sobre a opinião da população em geral. As pessoas influenciadas pelo debate e pelo medo da criminalidade serão levadas a telefonar dando indicação de estarem a favor da pena de morte.
- Amostra 3 - Opiniões de alguns leitores de determinada revista técnica, para representar as opiniões dos portugueses em geral.
  Resultado - Diferentes tipos de pessoas lêem diferentes tipos de revistas, pelo que a amostra não é representativa da população. Basta pensar que, de um modo geral, a população feminina ainda não adere às revistas técnicas como a

população masculina. A amostra daria unicamente indicações sobre a população constituída pelos leitores da tal revista.

- Amostra 4 - Utilizar alguns alunos de uma turma, para tirar conclusões sobre o aproveitamento de todos os alunos da escola.
  Resultado - Poderíamos concluir que o aproveitamento dos alunos é pior ou melhor do que na realidade é. As turmas de uma escola não são todas homogéneas, pelo que a amostra não é representativa dos alunos da escola. Poderia servir para tirar conclusões sobre a população constituída pelos alunos da turma.

- Amostra 5 - Utilizar os jogadores de uma equipa de basquete de uma determinada escola para estudar as alturas dos alunos dessa escola.
  Resultado - O estudo concluiria que os estudantes são mais altos do que na realidade são.

Normalmente obtêm-se amostras enviesadas quando existe 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.

Quando se pretende recolher uma amostra de dimensão n, de uma População de dimensão N, podemos recorrer a vários processos de amostragem. Como o nosso 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 e consequentemente "boas" estimativas.

Acontece que as propriedades dos estimadores, como veremos num módulo seguinte, só podem ser estudadas 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.

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 (sem reposição), amostras sistemáticas, amostras estratificadas e amostras aleatórias com reposição.**

3.1 – Amostragem aleatória simples

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

Amostra aleatória simples – Dada uma população de dimensão N, 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 de uma População com dimensão N, se selecciona uma amostra aleatória simples (a.a.s.) de dimensão n, qual a probabilidade de esta amostra ser seleccionada?

De acordo com a definição dada anteriormente para a.a.s., vem que cada amostra tem a mesma probabilidade, igual a $\binom{N}{n}^{-1}$ de ser seleccionada.

A selecção dos elementos da amostra pode ser feita em bloco ou 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. Tendo em consideração as probabilidades de escolher estes elementos (sequencialmente), confirma-se que a probabilidade de cada amostra é $\binom{N}{n}^{-1}$, como se apresenta a seguir:

1º elemento	2º elemento	3º elemento	...	e-nésimo elemento	Probabilidade da amostra
$\frac{n}{N}$ x	$\frac{n-1}{N-1}$ x	$\frac{n-2}{N-2}$ x	...	x $\frac{n-(n-1)}{N-(N-1)}$ =	$\frac{n!(N-n)!}{N!} = \frac{1}{\binom{N}{n}}$

Será que um esquema de amostragem aleatória simples implica que cada elemento da População tenha igual probabilidade de ser seleccionado?

Sim. Um esquema de amostragem aleatória simples, conduz a que cada elemento da População tenha a mesma probabilidade de ser seleccionado para a amostra, podendo-se demonstrar que é igual a $\frac{n}{N}$.

Efectivamente, para demonstrar este resultado, basta fazer o seguinte raciocínio: o número de amostras de n elementos que não contêm um qualquer elemento é $\binom{N-1}{n}$, donde a probabilidade de um qualquer elemento não ser incluído é (número de casos

favoráveis sobre o número de casos possíveis) $\binom{N-1}{n} / \binom{N}{n} = \frac{N-n}{N}$. Então, a probabilidade de um qualquer elemento ser seleccionado é $(1-\frac{N-n}{N}) = \frac{n}{N}$.

Num esquema de amostragem aleatória simples, verifica-se que cada elemento da população tem igual probabilidade de ser seleccionado para a amostra.

Nota – 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 e portanto não é uma amostra aleatória simples. É o que se passa, por exemplo, com a amostragem aleatória sistemática, em determinadas situações particulares, como veremos na secção seguinte.

Exemplo prático 1– Processo para obter uma amostra aleatória simples

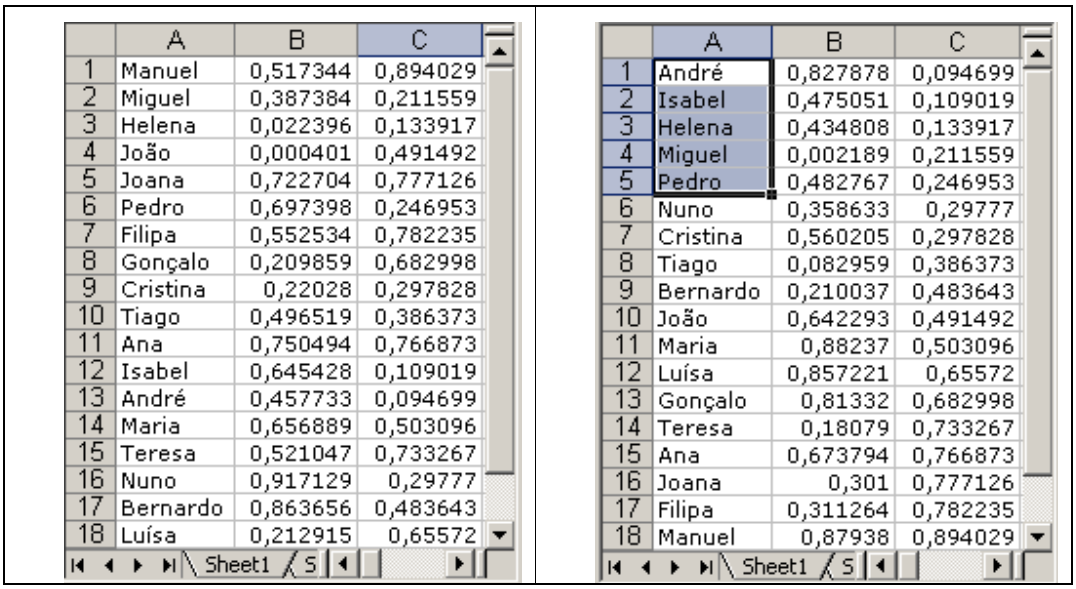
Vamos exemplificar um processo para obter uma amostra aleatória simples. Consideremos a população constituída pelos 18 alunos de uma turma do 10º ano de uma determinada Escola Secundária, em que a variável de interesse a estudar é a *altura* desses alunos, ou mais propriamente, estamos interessados em conhecer o parâmetro "valor médio" da característica *altura*. Uma maneira possível de recolher desta população uma amostra aleatória, seria escrever cada um dos indicadores (nº do aluno, nome, …) dos elementos da população num quadrado de papel, inserir todos esses bocados de papel numa caixa e depois seleccionar tantos quantos a dimensão da amostra desejada. Se os alunos estiverem numerados de 1 a 18, inserem-se numa caixa 18 quadrados de papel, cada um com o seu número e de seguida seleccionam-se tantos quantos a dimensão desejada para a amostra. Aos alunos cujos números foram seleccionados, pergunta-se qual a altura e regista-se. Admitindo que se seleccionou uma amostra de dimensão 5 e que as alturas dos alunos seleccionados foram 144 cm, 134 cm, 148 cm, 150 cm e 139 cm, uma estimativa para a altura média da turma será $\frac{144+134+148+150+139}{5}$ = 143 cm.

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 (desde que se tenha o cuidado de cortar os bocadinhos de papel todos do mesmo tamanho, para ficarem semelhantes, e de os baralhar convenientemente), constituída pelas alturas dos alunos seleccionados. 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 fase não se poder dizer qual das estimativas é "melhor", isto é, qual delas é a melhor aproximação do parâmetro a estimar, já que esse parâmetro é desconhecido (obviamente que nesta população tão pequena seria possível estudar exaustivamente todos os seus elementos, não sendo necessário recolher nenhuma amostra - este exemplo só serve para ilustrar uma situação)!

O processo que acabámos de descrever não é prático se a população a estudar tiver dimensão elevada. Vamos exemplificar um processo expedito, utilizando o Excel.

- 1º passo – inserir os nomes (ou outra identificação, como por exemplo o número) dos alunos numa folha de Excel
- 2º passo – utilizando a função RAND(), atribuir um número aleatório a cada aluno. Para isso basta inserir a função na célula B1 e replicá-la até à célula B18. Como esta função é volátil, isto é, muda quando se recalcula a folha, copiamos os valores gerados e através do *Edit*, fazemos um *Paste Special* - *Values*, para a coluna C, como se apresenta na figura da esquerda (repare-se que os valores que estavam inicialmente na coluna B foram alterados, devido ao facto de a função RAND() ser volátil, como referimos anteriormente):



	A	B	C
1	Manuel	0,517344	0,894029
2	Miguel	0,387384	0,211559
3	Helena	0,022396	0,133917
4	João	0,000401	0,491492
5	Joana	0,722704	0,777126
6	Pedro	0,697398	0,246953
7	Filipa	0,552534	0,782235
8	Gonçalo	0,209859	0,682998
9	Cristina	0,22028	0,297828
10	Tiago	0,496519	0,386373
11	Ana	0,750494	0,766873
12	Isabel	0,645428	0,109019
13	André	0,457733	0,094699
14	Maria	0,656889	0,503096
15	Teresa	0,521047	0,733267
16	Nuno	0,917129	0,29777
17	Bernardo	0,863656	0,483643
18	Luísa	0,212915	0,65572

Sheet1 / S

	A	B	C
1	André	0,827878	0,094699
2	Isabel	0,475051	0,109019
3	Helena	0,434808	0,133917
4	Miguel	0,002189	0,211559
5	Pedro	0,482767	0,246953
6	Nuno	0,358633	0,29777
7	Cristina	0,560205	0,297828
8	Tiago	0,082959	0,386373
9	Bernardo	0,210037	0,483643
10	João	0,642293	0,491492
11	Maria	0,88237	0,503096
12	Luísa	0,857221	0,65572
13	Gonçalo	0,81332	0,682998
14	Teresa	0,18079	0,733267
15	Ana	0,673794	0,766873
16	Joana	0,301	0,777126
17	Filipa	0,311264	0,782235
18	Manuel	0,87938	0,894029

Sheet1 / S

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, embora se comportem como números aleatórios (passam uma bateria de testes destinados a confirmar a sua aleatoriedade);

- 3º passo – ordenar o ficheiro, utilizando como critério a coluna C
- 4º passo – seleccionar para elementos da amostra os primeiros 5 alunos da coluna A. Como se verifica no lado direito da figura anterior, os cinco alunos seleccionados foram o André, a Isabel, a Helena, o Miguel e o Pedro.

Este processo pode ser generalizado para qualquer dimensão da População e qualquer dimensão da amostra.

O número de amostras aleatórias simples, de dimensão 5, que se podem extrair de uma população de dimensão 18 é igual a 8568 ($\binom{18}{5}$). Assim, pode-se utilizar o mesmo processo para obter outras amostras aleatórias simples de dimensão 5, já que a probabilidade de obter 2 amostras iguais é extremamente pequena (≈0,0001).

3.2 - Amostra aleatória sistemática

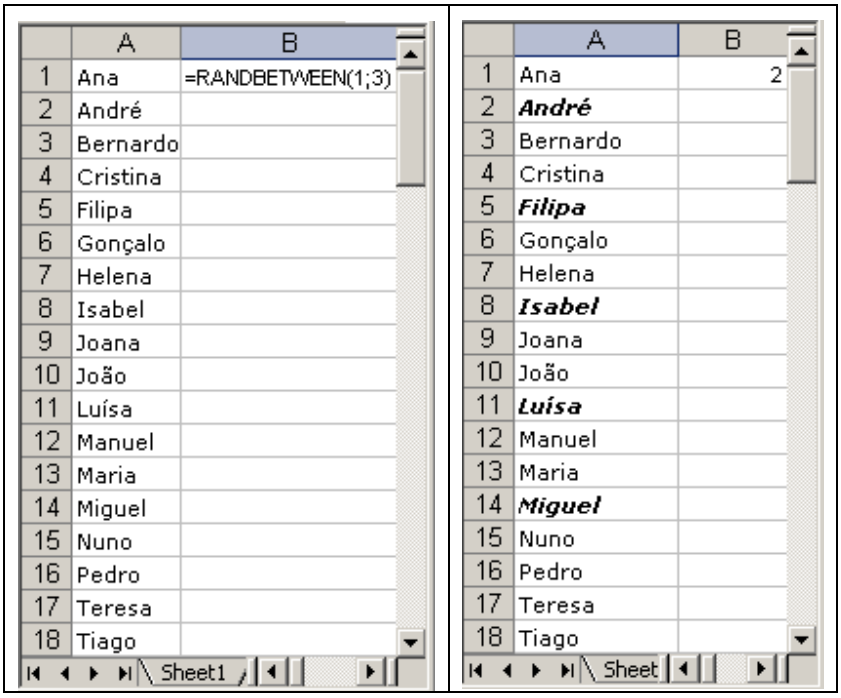
Mesmo considerando a tecnologia, se a dimensão da população for grande o processo anterior torna-se algo trabalhoso. Então uma alternativa é considerar uma amostra aleatória sistemática. Por exemplo, se pretendermos seleccionar uma amostra de 150 alunos de uma Universidade com 6000 alunos, considera-se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

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

A amostra aleatória sistemática não é uma amostra aleatória simples, já que nem todas as amostras possíveis, de dimensão n, têm a mesma probabilidade de serem seleccionadas. No entanto, se o quociente N/n for inteiro, mostra-se que a probabilidade de qualquer elemento ser seleccionado é igual a n/N. Pensemos nos N elementos colocados em círculo e seja N=nk. Comecemos por fixar uma posição inicial j. A probabilidade de um elemento A ser seleccionado é igual a $\sum_{j=1}^{N}$ P(AЄamostra/ posição inicial é j) P(posição inicial ser j) = $\sum_{j=1}^{N} \frac{n}{N} \times \frac{1}{N} = \frac{n}{N}$.

Exemplo prático 2 – Processo para obter uma amostra aleatória sistemática

Consideremos o ficheiro do exemplo anterior de onde pretendemos seleccionar 5 alunos, de forma sistemática. Como a dimensão da população é 18, selecciona-se aleatoriamente 1 elemento de entre os 3 (parte inteira de 18/5) primeiros. Para isso utilizamos a função *RANDBETWEEN* (i;j), que devolve um número aleatório, inteiro, entre i e j. No nosso caso considerámos i=1 e j=3 e o valor devolvido foi o 2 (poderia ter sido também o 1 ou o 3). De seguida seleccionam-se os elementos cujos números estejam espaçados de 3 unidades, até completar a dimensão da amostra:



	A	B
1	Ana	=RANDBETWEEN(1;3)
2	André	
3	Bernardo	
4	Cristina	
5	Filipa	
6	Gonçalo	
7	Helena	
8	Isabel	
9	Joana	
10	João	
11	Luísa	
12	Manuel	
13	Maria	
14	Miguel	
15	Nuno	
16	Pedro	
17	Teresa	
18	Tiago	

Sheet1

	A	B
1	Ana	2
2	***André***	
3	Bernardo	
4	Cristina	
5	***Filipa***	
6	Gonçalo	
7	Helena	
8	***Isabel***	
9	Joana	
10	João	
11	***Luísa***	
12	Manuel	
13	Maria	
14	***Miguel***	
15	Nuno	
16	Pedro	
17	Teresa	
18	Tiago	

Sheet

Nota – Ver na página 18 um processo de seleccionar uma amostra sistemática, utilizando a função *Sampling* do Excel.

3.3 - Amostra estratificada

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

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

A selecção de uma destas amostras não oferece dificuldade, a partir do momento em que os estratos estejam definidos. Os diferentes estratos são considerados como sendo populações distintas, pelo que de cada uma destas populações basta seleccionar uma amostra aleatória simples, utilizando o processo já considerado anteriormente.

3.4 – Amostragem com reposição

Nos esquemas de amostragem anteriormente referidos, utiliza-se a amostragem sem reposição, já que um elemento da população que seja seleccionado para a amostra, não volta a ser reposto, antes de se seleccionar o seguinte. Na amostragem com reposição, sempre que um elemento é seleccionado, é reposto na população.

O tratamento estatístico das propriedades dos estimadores é mais simples na amostragem com reposição do que na amostragem sem reposição, já que existe independência entre os elementos seleccionados. No entanto, como veremos no módulo de **Introdução à Estimação**, a amostragem sem reposição é mais eficiente do que a com reposição. Esta propriedade é de certo modo intuitiva, pois se recolhermos informação sobre elementos que anteriormente já tinham sido recolhidos, não estamos a acrescentar nada de novo.

Veremos também que se a população for "muito grande"[2], as amostragens sem e com reposição são equivalentes. Esta propriedade também é intuitiva, pois se a dimensão da população for muito grande, a probabilidade de o mesmo elemento ser seleccionado 2 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.

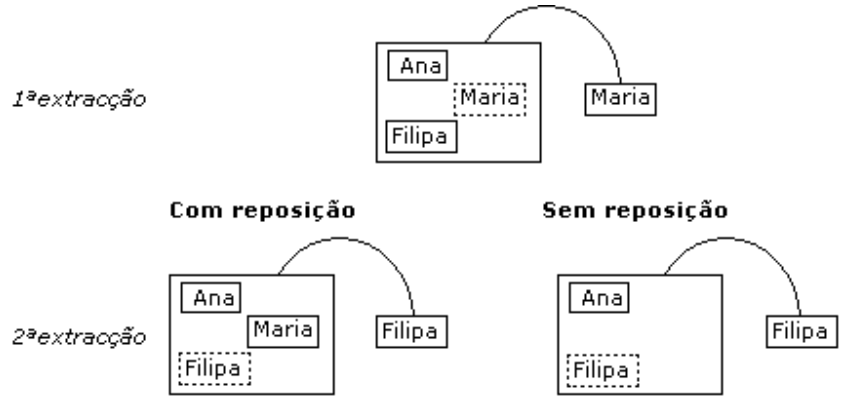
[2] Mais à frente diremos o que se entende por uma população "muito grande".

A probabilidade de cada uma das amostras ser seleccionada é igual a $1/N^n$. Fazendo um esquema idêntico ao considerado para obter a probabilidade de uma amostra aleatória simples, temos, agora para o caso da selecção ser feita com reposição:

1º elemento	2º elemento	3º elemento	...	e-nésimo elemento	Probabilidade da amostra
$\frac{1}{N}$ x	$\frac{1}{N}$ x	$\frac{1}{N}$ x	...	x $\frac{1}{N}$	$= \frac{1}{N^n}$

Exemplo - Selecção com reposição e sem reposição

Colocaram-se (Adaptado de Graça Martins, et al, 1999) numa caixa 3 papéis com o nome de 3 meninas: Ana, Maria e Filipa. Considere a selecção de amostras de dimensão 2, isto é, a experiência aleatória que consiste em retirar da caixa 2 papéis e verificar os nomes que saíram. Quais as amostras possíveis? Para responder a esta questão é necessário saber se a extracção se faz *com reposição*, isto é, se uma vez retirado um papel e verificado o nome se volta a colocar o papel na caixa, antes de proceder à extracção seguinte, ou se a extracção é feita *sem reposição*, isto é, uma vez retirado um papel, ele não é reposto antes de se proceder à próxima extracção. No esquema seguinte procuramos representar as duas situações:



Admitimos que na 1ª extracção saiu o papel com o nome da Maria. Na 2ª extracção, saiu o nome da Filipa nos dois casos, mas *na extracção com reposição* havia uma possibilidade em três de ele sair, tal como na 1ª extracção, enquanto que na *extracção sem reposição* havia uma possibilidade em duas de ele sair. Quer dizer que neste caso havia uma maior probabilidade de sair o nome da Filipa. Os conjuntos de amostras possíveis S_C e S_S correspondentes às duas situações com reposição e sem reposição, são respectivamente:

S_C = {(Ana, Ana), (Ana, Maria), (Ana, Filipa), (Maria, Ana), (Maria, Maria), (Maria, Filipa), (Filipa, Ana); (Filipa, Maria), (Filipa, Filipa)}

S_S = {(Ana, Maria), (Ana, Filipa), (Maria, Ana), (Maria, Filipa), (Filipa, Ana), (Filipa, Maria)}.

Exemplo prático 3 – Processo para obter uma amostra aleatória com reposição

Considere a população constituída pelos deputados da actual Legislatura (X Legislatura), que se pode obter a partir da página da Assembleia da Republica, e que apresentamos em anexo. Uma parte dessa tabela é apresentada a seguir, numa folha de Excel:

	A	B	C	D	E	H
1	**Nome**	**Partido**		**Sexo**	**Data nas.**	**Idade**
2	Abel Lima Baptista	CDS-PP	Viana do	M	13-10-1963	44
3	Adão José Fonseca Silva	PSD	Bragança	M	01-10-1957	50
4	Agostinho Correia Branquinho	PSD	Porto	M	10-08-1956	51
5	Agostinho Moreira Gonçalves	PS	Porto	M	15-07-1952	55
6	Agostinho Nuno de Azevedo Ferreira Lop	PCP	Braga	M	16-11-1944	63
7	Alberto Arons Braga de Carvalho	PS	Setúbal	M	20-09-1949	58
8	Alberto de Sousa Martins	PS	Porto	M	25-04-1945	62
9	Alberto Marques Antunes	PS	Setúbal	M	03-04-1949	58
10	Alcídia Maria Cruz Sousa de Oliveira Lop	PS	Porto	F	09-01-1974	33
11	Alda Maria Gonçalves Pereira Macedo	BE	Porto	F	07-09-1954	53
12	Aldemira Maria Cabanita do Nascimento	PS	Faro	F	04-04-1952	55
13	Ana Catarina Veiga Santos Mendonça M	PS	Setúbal	F	14-01-1973	34
14	Ana Isabel Drago Lobato	BE	Lisboa	F	28-08-1975	32

Sheet1 Sheet2 **Sheet3**

Na tabela anterior a coluna das idades foi acrescentada, tendo a idade de cada deputado sido calculada à data de 31/12/2007.

Admitamos que estamos interessados em estimar o *parâmetro* **idade média** dos deputados, a partir de amostras de dimensão 10. Vamos exemplificar a utilização do Excel, na obtenção de uma amostra aleatória, **com reposição**. Consideraremos dois processos: num dos processos utilizaremos a função *Sampling* e no outro a função *Randbetween*.

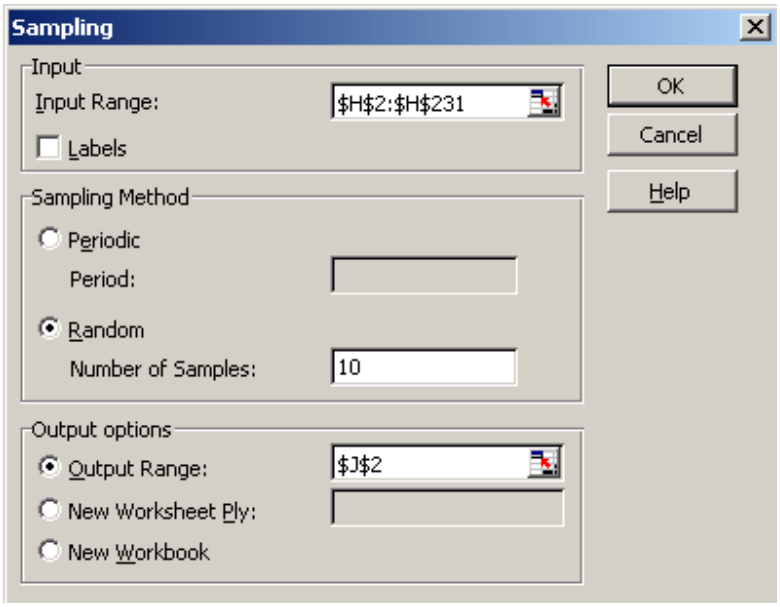
Processo de selecção da amostra aleatória com reposição, utilizando a função *Sampling*

Para utilizar este procedimento tem de se começar por verificar nos *Tools* se existe a opção *Data Analysis*. Caso não exista tem de se instalar, para o que basta aceder ao menu *Tools*, escolher o comando *Add-Ins* e seleccionar a opção *Analysis ToolPack* e clicar *OK*.

Processo de selecção da amostar:

a) Seleccione *Tools* → *Data Analysis* → *Sampling*.

   Na janela que se abre



em Input Range inserimos os endereços da coluna que contêm os elementos da população de onde se vai seleccionar a amostra; em Number of Samples inserimos o número de elementos que queremos seleccionar e em Output Range inserimos o endereço da célula para onde tencionamos colocar o 1º elemento dos elementos seleccionados. O resultado da operação anterior, depois de clicar o OK é:

	H	I	J
1	**Idade**		
2	44		49
3	50		46
4	51		59
5	55		59
6	63		42
7	58		46
8	62		64
9	58		47
10	33		38
11	53		41
12	55		

Observação 1 – A população de onde pretendemos seleccionar os elementos tem de ser constituída por valores numéricos. Este procedimento não serviria para seleccionar uma amostra de 10 nomes de deputados.

Observação 2 – A função *Sampling* também pode ser utilizada para seleccionar uma amostra sistemática, com período k, desde que tenhamos o seguinte cuidado: em Input Range colocamos os endereços das células onde estão os elementos da população, mas a iniciar na posição do 1º elemento que seleccionamos para a amostra, que é um número aleatório entre a posição 1 e k. Por exemplo, admitindo que pretendemos seleccionar uma amostra sistemática de 10 elementos, dos primeiros 23 (=230/10) elementos seleccionamos um ao acaso. Admitamos que saíu o 15. Isto significa que o elemento da população na posição 15 (célula $H$16) é o primeiro elemento a ser seleccionado para a amostra. Colocamo-lo na célula $M$2. Então em Input Range colocamos $H$17:$H$231, em Period escrevemos 23 e em Output Range escrevemos $M$3. Clicando em OK a função *Sampling* selecciona os 9 elementos que

faltavam para a amostra, nomeadamente os elementos das posições 38 = 15+23 (célula $H$39), 61=15+2x23 (célula $H$62), 84=15+3x23 (célula $H$85), ..., 222=15x9x23 (célula $H$223):

A amostra obtida encontra-se na coluna M:

	H	I	J	K	L	M
1	**Idade**					
2	44		49			**46**
3	50		46			46
4	51		59			32
5	55		59			56
6	63		42			69
7	58		46			50
8	62		64			49
9	58		47			62
10	33		38			34
11	53		41			55
12	55					

Sheet1

Processo de selecção da amostra aleatória com reposição, utilizando a função *RANDBETWEEN*

A partir da tabela inicial com a idade dos deputados, construímos uma outra tabela, em que inserimos, à esquerda da coluna dos nomes, uma coluna com um número, de 1 a 230. Para tornar a tabela mais simples, eliminamos as colunas respeitantes ao Partido, Círculo eleitoral, Sexo e Data de nascimento:

	A	B	C
1	**Número**	**Nome**	Idade
2	1	Abel Lima Baptista	44
3	2	Adão José Fonseca Silva	50
4	3	Agostinho Correia Branquinho	51
5	4	Agostinho Moreira Gonçalves	55
6	5	Agostinho Nuno de Azevedo Ferreira Lopes	63
7	6	Alberto Arons Braga de Carvalho	58
8	7	Alberto de Sousa Martins	62
9	8	Alberto Marques Antunes	58
10	9	Alcídia Maria Cruz Sousa de Oliveira Lopes	33
11	10	Alda Maria Gonçalves Pereira Macedo	53
12	11	Aldemira Maria Cabanita do Nascimento Bis	55
13	12	Ana Catarina Veiga Santos Mendonça Menc	34

Sheet1 Sheet2 Sheet3

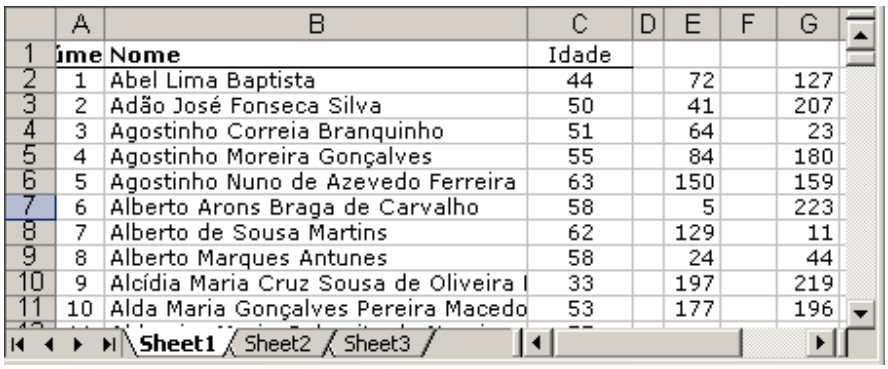
Processo de selecção da amostra:

a) Utilizar a função *RANDBETWEEN* (a;b), com a=1 e b=230, para obter um número aleatório, inteiro, entre 1 e 230;
b) Replicar essa fórmula mais 9 vezes para obter uma amostra de 10 números de deputados. A utilização desta fórmula várias vezes, simula a extracção, com reposição, já que pode sair repetidas vezes o mesmo número:

	A	B	C	E
1	Número	Nome	Idade	
2	1	Abel Lima Baptista	44	=RANDBETWEEN(1;230)
3	2	Adão José Fonseca Silva	50	=RANDBETWEEN(1;230)
4	3	Agostinho Correia Branquinho	51	=RANDBETWEEN(1;230)
5	4	Agostinho Moreira Gonçalves	55	=RANDBETWEEN(1;230)
6	5	Agostinho Nuno de Azevedo Ferreira Lopes	63	=RANDBETWEEN(1;230)
7	6	Alberto Arons Braga de Carvalho	58	=RANDBETWEEN(1;230)
8	7	Alberto de Sousa Martins	62	=RANDBETWEEN(1;230)
9	8	Alberto Marques Antunes	58	=RANDBETWEEN(1;230)
10	9	Alcídia Maria Cruz Sousa de Oliveira Lopes	33	=RANDBETWEEN(1;230)
11	10	Alda Maria Gonçalves Pereira Macedo	53	=RANDBETWEEN(1;230)

Sheet1 Sheet2 Sheet3

c) Uma vez que a função *RANDBETWEEN*(;) é volátil, fazer o *Paste Special* – *Values*, para outras células, dos 10 valores obtidos:

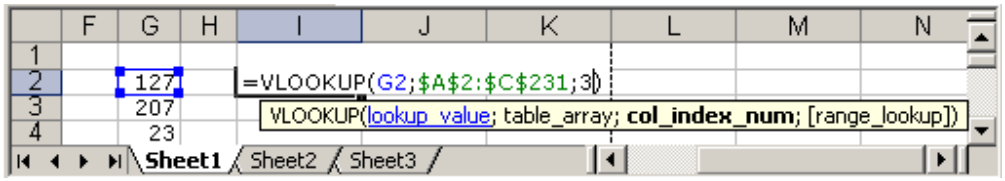


	A	B	C	D	E	F	G
1	ime	Nome	Idade				
2	1	Abel Lima Baptista	44		72		127
3	2	Adão José Fonseca Silva	50		41		207
4	3	Agostinho Correia Branquinho	51		64		23
5	4	Agostinho Moreira Gonçalves	55		84		180
6	5	Agostinho Nuno de Azevedo Ferreira	63		150		159
7	6	Alberto Arons Braga de Carvalho	58		5		223
8	7	Alberto de Sousa Martins	62		129		11
9	8	Alberto Marques Antunes	58		24		44
10	9	Alcídia Maria Cruz Sousa de Oliveira I	33		197		219
11	10	Alda Maria Gonçalves Pereira Macedo	53		177		196

Sheet1 Sheet2 Sheet3

Colámos na coluna G, os 10 valores obtidos na coluna E, e são estes os números dos deputados a quem vamos recolher a informação sobre a Idade. Observe-se que agora os valores obtidos na coluna E, já são outros, pois como se disse, a função *RANDBETWEEN*(;) é volátil e altera, sempre que se recalcula a folha;

d) Para obter as idades dos deputados cujos números foram seleccionados, vamos utilizar função do Excel *VLOOKUP* que, com os argumentos utilizados, devolve o elemento da 3ª coluna (coluna das idades) da matrix constituída pelos dados dos deputados (3 colunas), que corresponde ao número do deputado seleccionado para a amostra (coluna G):

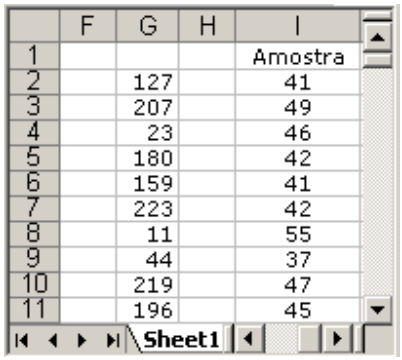


	F	G	H	I	J	K	L	M	N
1									
2		127		=VLOOKUP(G2;$A$2:$C$231;3)					
3		207							
4		23							

VLOOKUP(lookup_value; table_array; **col_index_num**; [range_lookup])

Sheet1 Sheet2 Sheet3

Esta função vai devolver o valor 41, que é a idade do deputado número 127. Replicamos esta fórmula pelas células I3:I11, obtendo as idades dos 10 deputados seleccionados:

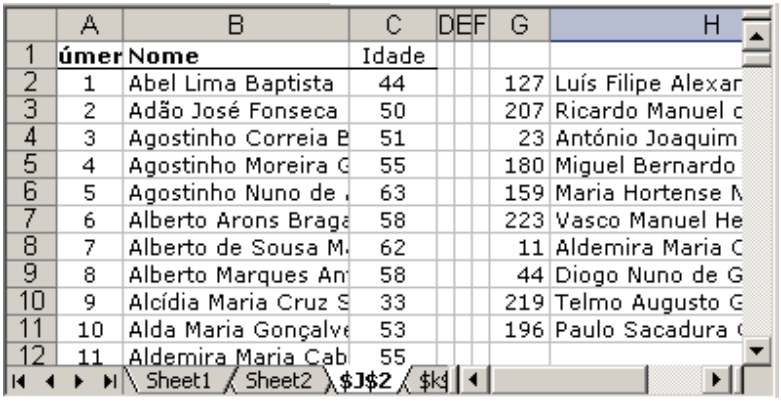


	F	G	H	I
1				Amostra
2		127		41
3		207		49
4		23		46
5		180		42
6		159		41
7		223		42
8		11		55
9		44		37
10		219		47
11		196		45

Sheet1

e) Uma estimativa para a idade média dos deputados, obtém-se calculando a média das idades dos 10 deputados seleccionados anteriormente, e que é 44,5 anos.

A função *VLOOKUP(a, b; c)* pode ser utilizada para seleccionar uma amostra de elementos não numéricos. Por exemplo no caso anterior se estivermos interessados nos nomes dos deputados com os números 127, 207, ..., 196, basta no terceiro argumento da função, ou seja no lugar do c, escrever o 2, para significar que pretendemos seleccionar os elementos na 2ª coluna.



	A	B	C	D	E	F	G	H
1	**úmer**	**Nome**	Idade					
2	1	Abel Lima Baptista	44				127	Luís Filipe Alexar
3	2	Adão José Fonseca	50				207	Ricardo Manuel c
4	3	Agostinho Correia B	51				23	António Joaquim
5	4	Agostinho Moreira G	55				180	Miguel Bernardo
6	5	Agostinho Nuno de	63				159	Maria Hortense M
7	6	Alberto Arons Braga	58				223	Vasco Manuel He
8	7	Alberto de Sousa M	62				11	Aldemira Maria C
9	8	Alberto Marques An	58				44	Diogo Nuno de G
10	9	Alcídia Maria Cruz S	33				219	Telmo Augusto G
11	10	Alda Maria Gonçalve	53				196	Paulo Sacadura
12	11	Aldemira Maria Cab	55					

Sheet1 Sheet2 $J$2 $k9

Não esquecer que se estivessemos interessados em seleccionar uma comissão de deputados para realizarem determinado trabalho, este processo de selecção não deveria ser utilizado, já que o mesmo deputado pode ser seleccionado mais do que uma vez. Nesta situação só teria sentido fazer uma selecção sem reposição.

Exercícios

1.1 - **População, Amostra**, **Variável** de interesse, **Parâmetro** de interesse, **Estatística** utilizada

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;

- População – empregados da empresa.
- Variável de interesse - *Salário* auferido por um empregado, escolhido ao acaso, da população anterior.
- Parâmetro – salário médio dos empregados. Como se recolheu informação sobre o salário de todos os empregados, a média dos valores obtidos dá o valor do salário médio pretendido.

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;

- População – alunos do 10º ano, que realizaram a prova global de Matemática no ano lectivo 2000-2001.
- Variável de interesse - *Nota* obtida por um aluno, escolhido ao acaso, da população anterior.
- Parâmetro – nota média obtida pelos alunos da população anterior. Como se recolheu informação sobre a nota de todos os alunos, a média destas notas dá o valor da nota média pretendida.

c) Pretendia-se averiguar a idade média dos alunos do 10º ano da Escola Secundária Prof. Herculano de Carvalho, no ano lectivo 2007/2008, pelo que se recolheu informação sobre a idade de 45 alunos do 10º ano dessa Escola;

- População - alunos do 10º ano da Escola Secundária Prof. Herculano de Carvalho, no ano lectivo 2007/2008.
- Variável de interesse - *Idade* de um aluno, escolhido ao acaso, da população anterior.
- Parâmetro – idade média dos alunos do 10º ano da Escola Secundária Prof. Herculano de Carvalho, no ano lectivo 2007/2008.
- Amostra - conjunto das idades dos 45 alunos seleccionados (como já foi referido, estamos a identificar os indivíduos da amostra, com os valores observados, da variável de interesse, sobre esses indivíduos .
- Estatística - A média das idades dos 45 alunos é a estatística que se utiliza como estimatíva do parâmetro pretendido, ou seja, da idade média.

d) Pretendia-se averiguar a quantidade de vinho (em litros) 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;

- População – conjunto dos agricultores do Alentejo que produziram vinho em 1999.
- Variável de interesse - *quantidade de vinho* produzida por um agricultor, escolhido ao acaso, da população anterior.
- Parâmetro – quantidade total de litros produzida pelos agricultores do Alentejo no ano de 1999.
- Amostra – quantidades de litros produzidas pelos 10 agricultores seleccionados.
- Estatística - média das quantidades de litros produzidas pelos 10 agricultores, vezes o número total de agricultores da população considerada.

e) Pretendia-se saber o salário médio auferido pelos trabalhadores da indústria têxtil, pelo que se recolheu informação sobre os salários mensais auferidos por 250 desses trabalhadores;

- População – conjunto dos trabalhadores da indústria têxtil.
- Variável de interesse - *salário* auferido por um trabalhador, escolhido ao acaso, da população anterior.
- Parâmetro – salário médio auferido pelos trabalhadores da indústria têxtil.

- Amostra – salários auferidos pelos 250 trabalhadores seleccionados.
- Estatística - média dos valores da amostra.

f) Pretendia-se averiguar a quantidade mensal (em kg) de batata consumida nos lares portugueses, pelo que se recolheu informação sobre as quantidades de batata consumidas mensalmente em 100 lares portugueses;

- População – conjunto dos lares portugueses.
- Variável de interesse – quantidade de quilos de batata, consumidos mensalmente num lar português, escolhido aoa acaso.
- Parâmetro – quantidade média de batata consumida mensalmente, nos lares portugueses.
- Amostra – quantidades de quilos de batata consumidos nos 100 lares seleccionados.
- Estatística - média dos valores da amostra considerada.

g) Pretendia-se estudar a eficácia de um medicamento novo para curar determinada doença, pelo que se seleccionaram 20 doentes padecendo dessa doença;

- População – conjunto dos doentes padecendo da doença em estudo.
- Parâmetro – percentagem de curas que se obtêm, utilizando o medicamento.
- Amostra – conjunto dos 20 doentes seleccionados, a quem se deu o medicamento.
- Estatística – percentagem de curas obtidas, nos 20 doentes seleccionados.

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;

- População – os 5 vendedores do stand de automóveis.
- Parâmetro – total de carros vendidos pelos 5 vendedores. Como se investigou o número de carros que cada vendedor tinha vendido, o total de carros dá o valor do parâmetro pretendido.

i) Pretendia-se averiguar o número de leitores dos jornais diários (editados em Portugal), pelo que se investigou junto de 6 jornais diários, o número de leitores;

- População – conjunto dos jornais diários.
- Parâmetro – número total de leitores de jornais diários.
- Amostra – número de leitores dos 6 jornais diários.
- Estatística – número total de leitores dos 6 jornais seleccionados para a amostra vezes N/6, em que N é o número de jornais diários.

j) Pretendia-se averiguar a percentagem de raparigas que frequentam a FCUL, no ano lectivo de 2007/2008, pelo que se seleccionaram 50 alunos dessa faculdade;

- População – conjunto dos alunos que frequentam a FCUL, no ano lectivo de 2007/2008.
- Parâmetro – percentagem de raparigas na população anterior.
- Amostra – conjunto dos 50 alunos seleccionados.
- Estatística – percentagem de raparigas na amostra anterior.

Parâmetro e Estatística

1.2 - Diga se são verdadeiras ou falsas as seguintes afirmações:

a) Uma estatística é um número que se calcula a partir dos dados da amostra;

Verdadeiro (Chamamos, no entanto, a atenção para o facto de também interpretarmos estatística como uma função que só depende dos valores da amostra e não depende de parâmetros desconhecidos. Ao valor observado desta função, para uma dada amostra que se observou, também é usual dar o nome de estatística. Assim, neste caso, estatística seria um número).

b) Os parâmetros utilizam-se para estimar estatísticas;

Falso.

c) A média populacional é um parâmetro;

Verdadeiro.

d) Um parâmetro é uma característica numérica da variável que se está a estudar na População.

Verdadeiro.

1.3 - 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;

Parâmetro.

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.

Estatística.

c) Dos 230 deputados que compunham a VIII legislatura, **21.3**% eram mulheres.

Parâmetro.

d) Perguntou-se a 80 deputados qual o partido que representavam, tendo-se concluído que **49**% representavam o PS.

Estatística. (A população é constituída por 230 deputados).

e) Perguntou-se a 10 deputados qual a sua idade, tendo-se concluído que a média das idades era de **45** anos.

Estatística.

Amostras enviesadas e amostras aleatórias

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

a) Estes deputados constituem uma amostra ou uma população?

Constituem uma amostra.

b) Quantos deputados, nos 5 seleccionados, pertencem ao círculo eleitoral da sua residência?

c) Suponha que está interessado em estudar o nº médio de anos de serviço dos deputados que constituem a X legislatura. Considera o conjunto de deputados seleccionados representativos da população? Porquê?

Não. Ao seleccionarmos deputados de que já tenhamos ouvido falar, é natural que eles já tenham pertencido a legislaturas anteriores, pelo que os valores obtidos para o número de anos ao serviço, são superiores ao que seria de esperar, se a selecção fosse aleatória.

d) Se calculasse a média dos anos de serviço dos deputados seleccionados esperava obter um valor superior ou inferior ao da média populacional?

Tendo em conta a resposta à alínea anterior, ao calcularmos a média dos números de anos de serviço, é de esperar obter um valor maior do que o 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.

Sim. Como estamos sistematicamente a escolher deputados conhecidos, é de esperar que estejam há mais anos no Parlamento.

f) Se tivesse seleccionado pelo mesmo processo 10 deputados, obteria uma amostra mais representativa do que a constituída pelos 5 deputados? Explique.

Não. Se o processo de selecção da amostra for enviesado, que é o caso, aumentar a dimensão da amostra não elimina o problema.

1.5 - Para que uma amostra seja representativa da população, basta que cada elemento da população tenha igual probabilidade de ser seleccionado?

Não. Pode acontecer que cada elemento da população tenha igual probabilidade de ser seleccionado e no entanto a amostra não ser representativa. Considere por exemplo uma população constituída por um certo número de estratos, com igual número de elementos: por exemplo, uma população constituída por 6 estratos, estrato 1, estrato 2, ..., estrato 6, com igual número de elementos. Lança um dado e se sair a face i, com i=1,...,6, selecciona o estrato i. Depois selecciona todos os elementos deste estrato. A amostra resultante não é representativa da população dada.

Projectos

1 - Numa empresa de 97 trabalhadores, pretende-se seleccionar aleatoriamente 10 trabalhadores para integrarem uma comissão que se encarregará da festa de Natal. Como sugere que se faça a recolha da amostra? Com ou sem reposição? Explique porquê. Obtenha uma dessas amostras.

Trabalhadores da empresa

Nº	Nome	Nº	Nome	Nº	Nome
1	Alexandra Almeida	34	Margarida Simões	67	Paulo Santos
2	Alexandre Carmo	35	M. Adelina Azevedo	68	Paulo Valente
3	Alda Morais	36	M. Alexandra Almeida	69	Pedro Casanova
4	Ana Ribeiro	37	M. Alexandra Ribeiro	70	Pedro Dalo
5	Ana Cristina Santos	38	M. Cristina Carvalho	71	Pedro Martins
6	Ana Cristina Oliveira	39	M. Cristina Freire	72	Pedro Lisboa
7	Anabela Pais	40	M. de Fátima Osório	73	Pedro Sintra
8	António Couto	41	M. Fernanda Rocha	74	Pedro Valente
9	António Fernandes	42	M. Isabel Frade	75	Pedro Viriato
10	António Pinto	43	M. Isabel Santos	76	Rita Amaral
11	Armando Ferreira	44	M. Luísa Faria	77	Rita Bendito
12	Carlos Matos	45	M. Manuel Trindade	78	Rita Évora
13	Carlos Sampaio	46	M. Manuela Lino	79	Rita Seguro
14	Cristina Vicente	47	M. Nazaré Pinto	80	Rita Valente
15	Cristina Zita	48	M. Neusa Lopes	81	Rufo Almeida
16	Dora Ferreira	49	M. Olga Martins	82	Rui André
17	Elsa Sampaio	50	M. Paula Pitarra	83	Rui Martins
18	Fernando Barroso	51	M. Paula Garcês	84	Rui Teixeira
19	Fernando Martins	52	M. Rosário Gomes	85	Rui Vasco
20	Fernando Santos	53	M. Rute Costa	86	Sérgio Teixeira
21	Filomena Silva	54	M. Rute Rita	87	Sílvio Lino
22	Francisco Gomes	55	M. Teresa António	88	Tânia Lopes
23	Isabel Soares	56	M. Teresa Bento	89	Tânia Martins
24	Isabel Silva	57	M. Teresa Garcia	90	Teresa Adão
25	João Morais	58	Mário Martins	91	Teresa Paulo
26	João Sousa	59	Mário Reis	92	Teresa Vasco
27	Luís Horta	60	Nuno Simões	93	Vera Mónica
28	Luís Sousa	61	Nuno Ventura	94	Vera Patrícia
29	Luís Ribeiro	62	Olga Martins	95	Vera Teixeira
30	Manuel Santos	63	Óscar Trigo	96	Vitor Santos
31	Manuel Pereira	64	Osvaldo	97	Vitor Zinc
32	Manuel Teixeira	65	Paulo Nunes		
33	Margarida Almeida	66	Paulo Martins		

A selecção dos 10 trabalhadores deverá ser feita sem reposição, porque se se fizer com reposição o mesmo trabalhador poderia ser seleccionado mais do que uma vez.
Vamos então proceder à selecção de uma amostra aleatória simples, de dimensão 10. Começámos por considerar um ficheiro em Excel, com os números e nomes dos trabalhadores e depois utilizámos a seguinte metodologia:
a) Utilizando a função *RAND()*, atribuímos a cada empregado um número aleatório (pseudo-aleatório) que inserimos na coluna C;
b) Como a função *RAND()* é volátil, utilizando o Paste Special – Values, copiámos os valores obtidos anteriormente, para a coluna D;

	A	B	C	D
1	**Nº**	**Nome**	RAND()	
2	1	Alexandra Almeida	0,118232	0,38824
3	2	Alexandre Carmo	0,791447	0,246146
4	3	Alda Morais	0,316874	0,632173
5	4	Ana Ribeiro	0,825189	0,731376
6	5	Ana Cristina Santos	0,610819	0,787516
7	6	Ana Cristina Oliveira	0,021897	0,635004
8	7	Anabela Pais	0,369577	0,370544
9	8	António Couto	0,069608	0,444863
10	9	António Fernandes	0,117675	0,484795
11	10	António Pinto	0,166421	0,160457
12	11	Armando Ferreira	0,22534	0,139469
13	12	Carlos Matos	0,94335	0,748782
14	13	Carlos Sampaio	0,382802	0,603294
15	14	Cristina Vicente	0,89251	0,29389

Sheet1 Sheet2 Shee

c) Ordenar as 97 linhas que contêm informação sobre os trabalhadores, utilizando como critério de ordenação os valores da coluna D;
d) Seleccionar os primeiros 10 trabalhadores, para integrarem a comissão:

	A	B	C	D
1	**Nº**	**Nome**	RAND()	
2	88	Tânia Lopes	0,127889	0,008246
3	29	Luis Ribeiro	0,882692	0,026596
4	51	M. Paula Garcês	0,405072	0,028246
5	77	Rita Bendito	0,414361	0,029639
6	49	M. Olga Martins	0,449071	0,041879
7	94	Vera Patrícia	0,387723	0,054269
8	17	Elsa Sampaio	0,870379	0,059287
9	28	Luis Sousa	0,319493	0,059884
10	71	Pedro Martins	0,467969	0,076032
11	30	Manuel Santos	0,833374	0,081818

Sheet1 Sheet2 She

e) A comissão é constituída pelos seguintes trabalhadores: Tânia Lopes, Luís Ribeiro, M. Paula Garcês, Rita Bendito, M. Olga Martins, Vera Patrícia, Elsa Sampaio, Luís Sousa, Perro Martins e Manuel Santos.

2 - A Presidente do Conselho Executivo da sua escola secundária encarregou uma comissão de alunos de fazer uma sondagem para averiguar quantas horas, por semana, os alunos gastavam a ver televisão. Admitindo que é um dos elementos da comissão, faça um pequeno relatório onde identifica a População, a característica populacional de interesse, o parâmetro a estudar e descreva o processo de amostragem utilizado e os resultados obtidos.