INTRODUÇÃO À INFERÊNCIA ESTATÍSTICA

Módulo 1 – Introdução à Amostragem

Módulo 2 – Introdução à Estimação – Estimação pontual

Módulo 3 – Introdução à Inferência Estatística – Estimação Intervalar ou Intervalos de confiança

Maria Eugénia Graça Martins (memartins@fc.ul.pt)
Departamento de Estatística e Investigação Operacional
Faculdade de Ciências da Universidade de Lisboa

Statistical Thinking will be one day as necessary for efficient citizenship as the ability to read and write. - HG Wells **...**

Nota prévia

Este pequeno curso de Introdução à Inferência Estatística destina-se essencialmente aos professores que leccionam a disciplina de Matemática Aplicada às Ciências Sociais.

No âmbito desta disciplina foram introduzidos alguns conceitos novos, nomeadamente os que dizem respeito à Inferência Estatística, pelo que é necessário disponibilizar algum material por onde os professores possam actualizar os seus conhecimentos.

É neste contexto que foram feitas estas folhas que, espero, possam ajudar a esclarecer algumas dúvidas que nos surgem sempre, quando nos propomos dar conteúdos novos.

Sei, pela minha experiência de professora, que só conseguimos transmitir conhecimentos, quando estes estão tão bem interiorizados dentro de nós, de modo que, mesmo para um assunto um pouco mais complicado, consigamos encontrar as palavras simples e certas. As palavras certas que façam um "clique" na motivação dos nossos alunos. Caso contrário, só se transmite mais alguma informação, entre tanta a que estão (estamos) sujeitos e esta facilmente se esquece....

Caros colegas: espero que tenham tanto gosto em ler este texto, como eu tive em fazê-lo.

Janeiro, 2008

Maria Eugénia Graça Martins

Índice

Módulo 1 – Introdução à Amostragem
1 – Introdução 1
2 - Aquisição de dados: sondagens. População e amostra. Parâmetro e Estatística . 3
2.1 - Sondagens. População e amostra. Parâmetro e Estatística 4
3 – Amostra enviesada. Amostra aleatória e amostra não aleatória 8
3.1 – Amostragem aleatória simples 10
3.2 - Amostra aleatória sistemática 13
3.3 - Amostra estratificada 15
3.4 – Amostragem com reposição 15
Exercícios 22

Módulo 2 - Introdução à Estimação – estimação pontual
1 – Introdução 27
2 - Distribuição de amostragem. Estimador centrado e não centrado. Precisão 27
3 - Estimação do valor médio 34
3.1 - Estimação do valor médio utilizando amostras aleatórias simples (sem reposição) 34
3.1.1 - Distribuição de amostragem da média, como estimador do valor médio de uma População finita 34
3.1.2 - Distribuição de amostragem aproximada da média, como estimador do valor médio de uma População finita, mas de dimensão suficientemente grande 40
3.2 – Distribuição de amostragem da média, em amostragem com reposição 41
Exercícios 47
4 - Estimação da proporção 51
4.1 - Distribuição de amostragem da proporção amostral, como estimador da proporção populacional 51
Exercícios 54
5 - O modelo Normal (ou Gaussiano) 58

Módulo 3 - Introdução à Inferência estatística – estimação intervalar ou intervalos de Confiança
1 – Introdução 65
2 – Intervalo de confiança para o valor médio 65
3 – Intervalo de confiança para a proporção 71
Exercícios 78
Bibliografia 85

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**.

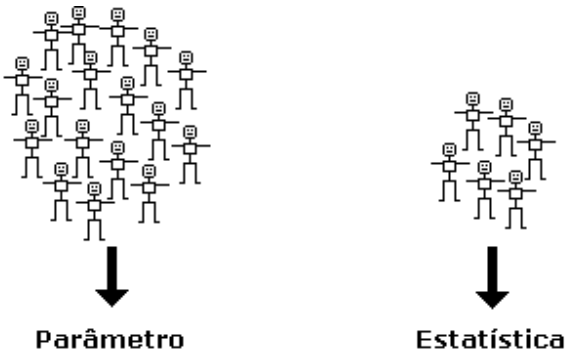
Por exemplo, ao estudar a população constituída por todos os potenciais eleitores para as legislativas, dois parâmetros que podem ter interesse são:

- **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.

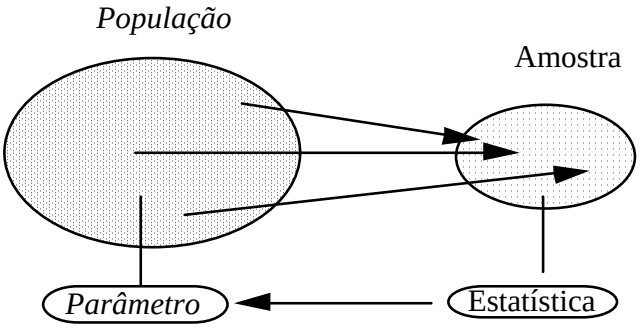
No caso do exemplo anterior, se estivermos interessados em estimar o **parâmetro** ou proporção populacional "*percentagem de eleitores que estão decididos a votar*" através de amostras de dimensão 1000, o **estimador** será a proporção amostral "*percentagem de eleitores, em 1000, que interrogados disserem 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

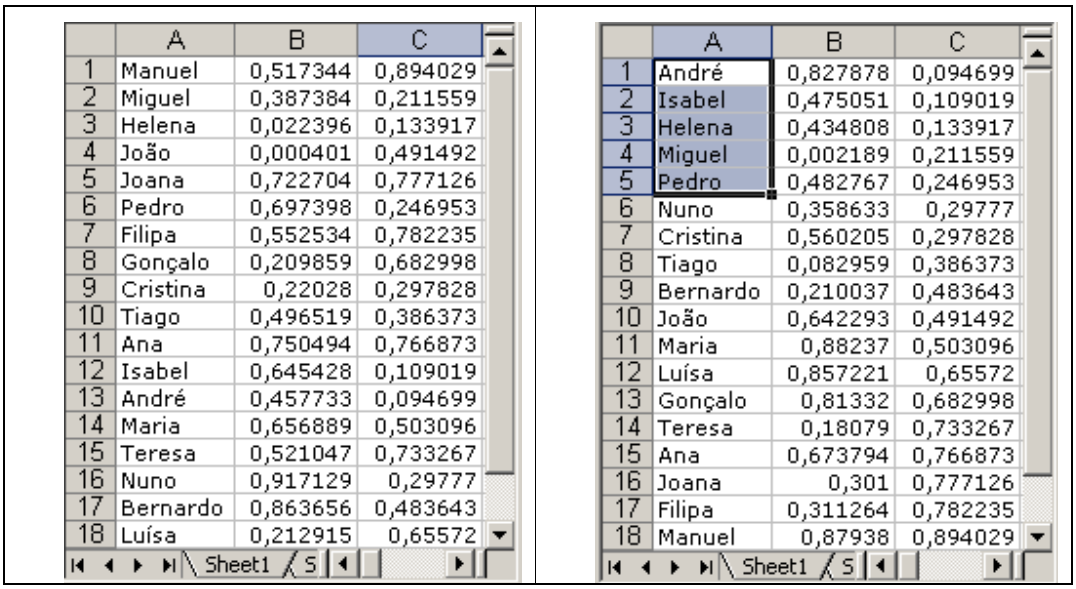
Vamos exemplificar um processo para obter uma amostra aleatória simples. Consideremos a população constituída pelos 18 alunos de uma turma do 10º ano de uma determinada Escola Secundária, em que a variável de interesse a estudar é a *altura* desses alunos, ou mais propriamente, estamos interessados em 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

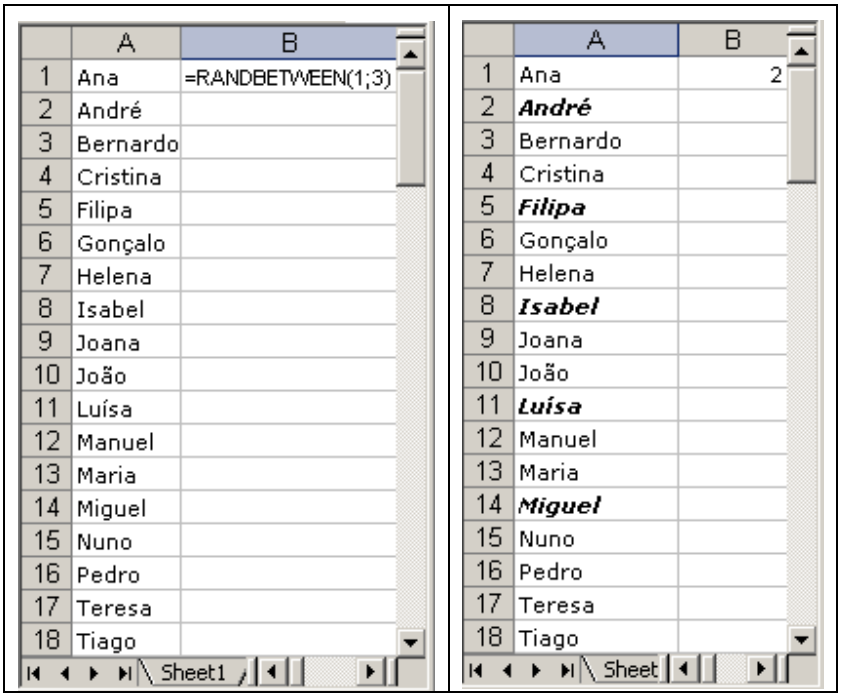
Mesmo considerando a tecnologia, se a dimensão da população for grande o processo anterior torna-se algo trabalhoso. Então uma alternativa é considerar uma amostra aleatória sistemática. Por exemplo, se pretendermos seleccionar uma amostra de 150 alunos de uma Universidade com 6000 alunos, considera-se 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.

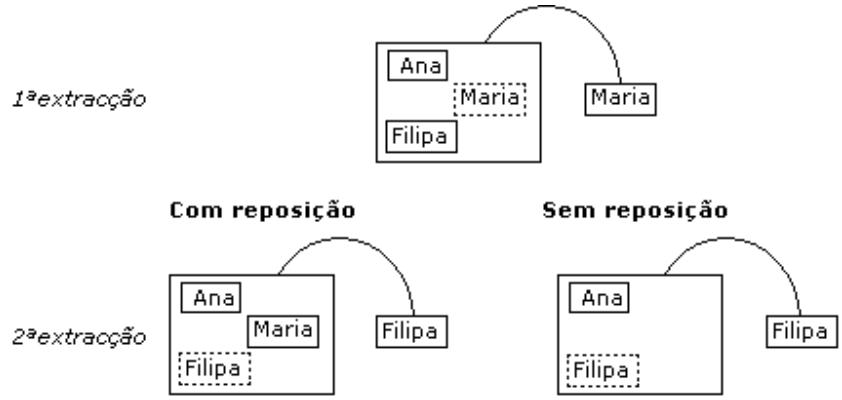
[2] Mais à frente diremos o que se entende por uma população "muito grande".

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*.

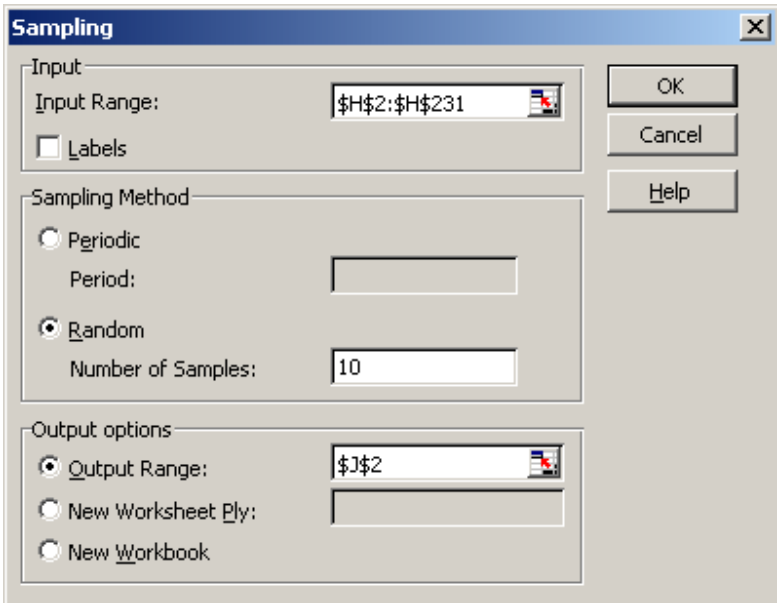
Processo de selecção da amostra aleatória com reposição, utilizando a função *Sampling*

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

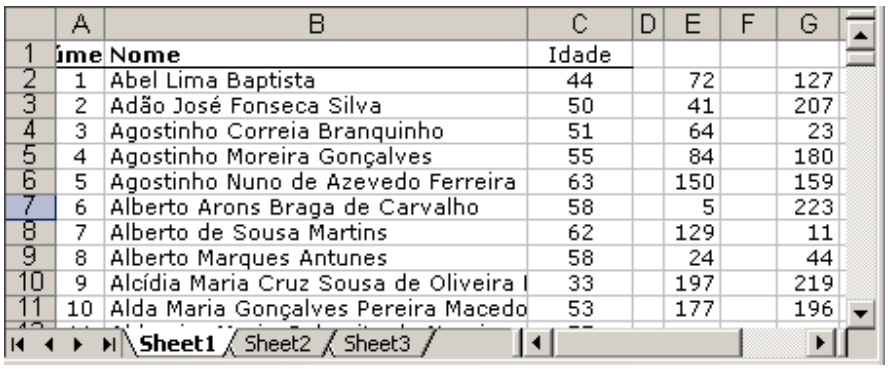
Processo de selecção da amostra:

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:

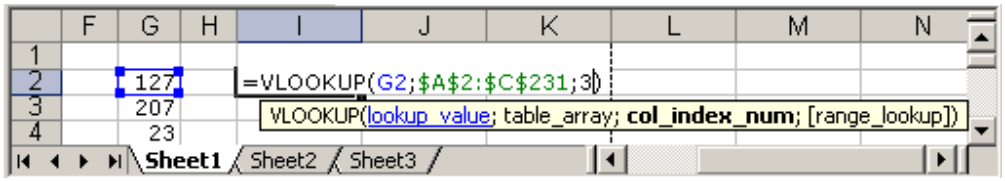


	A	B	C	D	E	F	G
1	ime	Nome	Idade				
2	1	Abel Lima Baptista	44		72		127
3	2	Adão José Fonseca Silva	50		41		207
4	3	Agostinho Correia Branquinho	51		64		23
5	4	Agostinho Moreira Gonçalves	55		84		180
6	5	Agostinho Nuno de Azevedo Ferreira	63		150		159
7	6	Alberto Arons Braga de Carvalho	58		5		223
8	7	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):

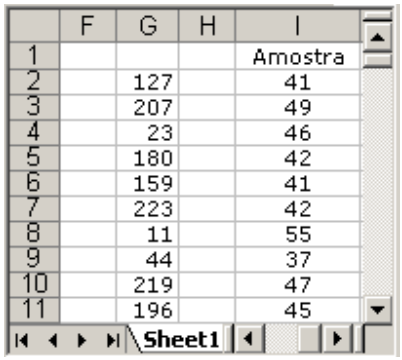


	F	G	H	I	J	K	L	M	N
1									
2		127		=VLOOKUP(G2;$A$2:$C$231;3)					
3		207							
4		23							

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:

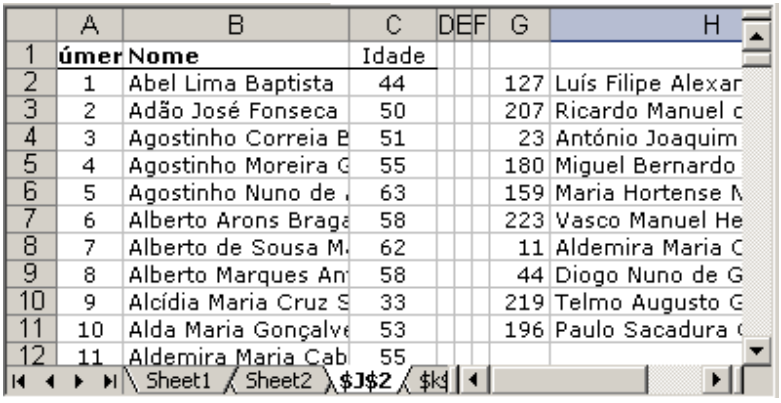


	F	G	H	I
1				Amostra
2		127		41
3		207		49
4		23		46
5		180		42
6		159		41
7		223		42
8		11		55
9		44		37
10		219		47
11		196		45

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.

Introdução à Estimação – estimação pontual

1 - Introdução

No módulo 1 – Introdução à Amostragem, foi tratado o problema da amostragem. Este é um problema de grande importância, pois como foi referido na altura, o nosso objectivo é, a partir das propriedades estudadas na amostra, *inferir* propriedades para a População, nomeadamente estimar os parâmetros desconhecidos, pelo que é necessário utilizar processos de amostragem que dêem origem a "bons" estimadores e consequentemente "boas" estimativas, ou seja valores "próximos" dos parâmetros a estimar.

Acontece que as propriedades dos estimadores, como veremos neste módulo, 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.

2 - Distribuição de amostragem. 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 X, que pode ser medida pela variância populacional σ^2, sabemos que podemos depois de recolher uma amostra, obter duas estimativas diferentes para essa variância, substituindo os valores da amostra nas expressões dos estimadores

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} \quad \text{ou} \quad S'^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$$

onde representamos por X_1, X_2, ..., X_n, variáveis independentes, com distribuição idêntica à de X e por $\overline{X}$ a média, ou seja, $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$.

Quais as razões que nos podem levar a preferir um dos estimadores relativamente ao outro? Qual o que fornece, de um modo geral, "melhores estimativas"? Intuitivamente desejaríamos que as diferentes estimativas fornecidas por um estimador, para diferentes amostras, da mesma dimensão, não estivessem "muito afastadas" do parâmetro que estamos a estimar! Se assim fosse teríamos uma certa garantia de que

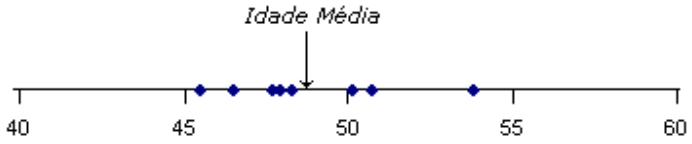
a estimativa que se obtém para a amostra que se recolhe (na prática só recolhemos uma amostra!) daria um valor aproximado do parâmetro.

Exemplo – Consideremos o exemplo utilizado no módulo 1 – Introdução à Amostragem, da população constituída pelos deputados da X Legislatura, e suponhamos que se pretende estimar a idade média ou o valor médio da característica *Idade* (média das idades de todos os deputados). Vamos seleccionar 8 amostras aleatórias simples, de dimensão 10, e calcular as médias das idades dos deputados das amostras seleccionadas:

	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
1	Amostra1		Amostra2		Amostra3		Amostra4		Amostra5		Amostra6		Amostra7		Amostra8	
2	nº	idade	nº	idade	nº	idade	nº	idade	nº	idade	nº	idade	nº	idade	nº	idade
3	65	46	117	50	55	57	57	47	2	50	119	50	43	32	39	55
4	129	34	79	70	105	49	33	55	32	31	28	68	66	42	129	34
5	71	54	123	36	44	37	201	30	55	57	89	64	201	30	178	43
6	225	48	161	39	170	58	212	52	130	50	102	59	33	55	91	42
7	14	40	202	55	161	39	41	54	78	52	99	55	79	70	19	39
8	127	41	144	56	182	73	78	52	131	50	52	58	44	37	95	64
9	108	40	149	60	201	30	91	42	136	40	227	55	100	53	220	38
10	207	49	100	53	89	64	152	44	170	58	220	38	65	46	165	58
11	41	54	43	32	2	50	118	48	1	44	135	33	73	60	57	47
12	138	71	133	50	37	50	146	31	58	51	6	58	98	54	147	45
13	média	47,7	média	50,1	média	50,7	média	45,5	média	48,3	média	53,8	média	47,9	média	46,5

Sheet1 Sheet2 Sheet3

Repare-se que as 8 médias obtidas são diferentes umas das outras, mas estão relativamente próximas:



No esquema anterior assinalámos a posição do parâmetro em estudo (48,7 anos), já que neste caso a dimensão da população é razoavelmente pequena, e facilmente se calculou a média das idades dos 230 deputados. Da figura anterior sobressai o seguinte:

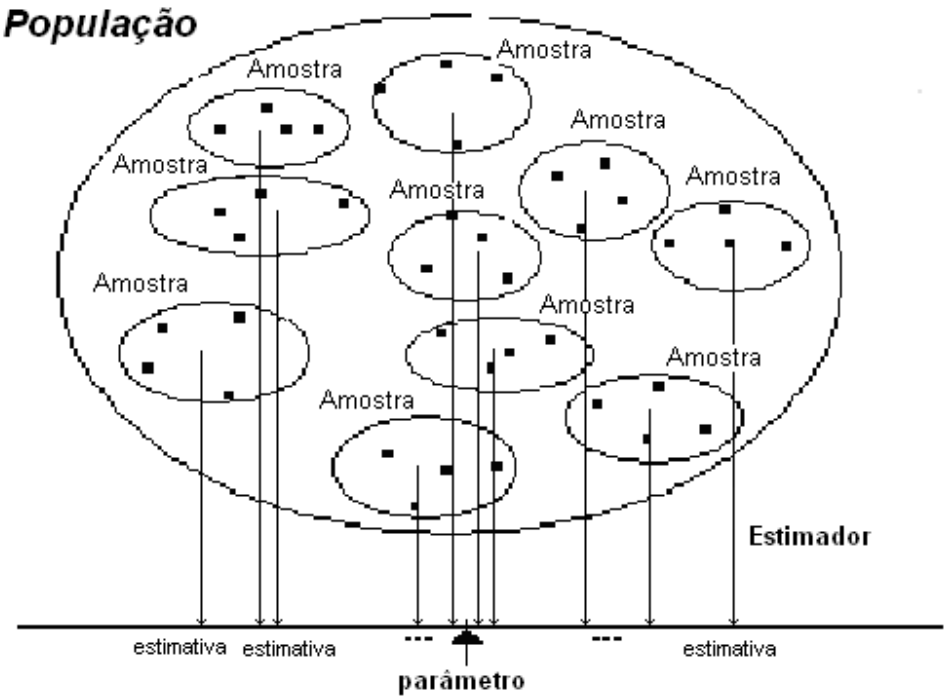
- As médias obtidas distribuem-se para um e outro lado do parâmetro e
- A variabilidade apresentada pelas estimativas é relativamente pequena, isto é, as diferentes estimativas estão próximas do parâmetro a estimar.

Numa situação que tenha interesse, sob o ponto de vista estatístico, o parâmetro em estudo é desconhecido e não é fácil de o calcular, como no caso deste exemplo, pelo que terá de ser estimado. Então a pergunta que devemos fazer e para a qual vamos procurar dar resposta, é a seguinte:

Como se comportam, relativamente ao parâmetro em estudo, todas as estimativas fornecidas por um dado estimador, para todas as amostras possíveis?

Como veremos a seguir, 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, da mesma dimensão, que se podem extrair da População.



Distribuição de amostragem – Distribuição de amostragem de um estimador é a distribuição dos valores que o estimador assume para todas as possíveis amostras, da mesma dimensão, da População.

A maior parte das vezes não se consegue obter a distribuição de amostragem exacta, pois não está dentro dos "limites do razoável", considerar todas as amostras possíveis, mas tem-se uma distribuição aproximada, considerando um número suficientemente grande de amostras da mesma dimensão e calculando para cada uma delas o valor do estimador (problema a estudar posteriormente).

O que é que se entende por um "bom" estimador?

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 o valor médio da sua distribuição de amostragem for igual ao parâmetro a estimar, ou seja, 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*.

No início desta secção questionámos quais as razões que nos poderiam levar a preferir, para a variância populacional, o estimador S^2 relativamente a S'^2. Neste momento podemos dizer que é o facto de S^2 não apresentar enviesamento (a demonstração desta propriedade sai fora do âmbito deste curso).

Aparece-nos, novamente a palavra enviesamento, que já nos tinha surgido no módulo 1 – Introdução à Amostragem, mas agora noutro contexto. Efectivamente, relacionado com um processo de amostragem e com a escolha de um estimador, temos dois tipos de **enviesamento**:

- O associado com o *processo de amostragem*, isto é, com a recolha da amostra, em que uma amostra enviesada é o resultado do processo de amostragem não ser aleatório;
- O associado com o *estimador* escolhido, para estimar o parâmetro em estudo. Se o estimador não for centrado, diz-se que é enviesado.

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).

Por outro lado, temos que ter outra preocupação com o estimador escolhido, que diz respeito à precisão:

Precisão – Quando utilizamos um estimador para estimar um parâmetro, e calculamos o seu valor para várias amostras, obtêm-se outras tantas estimativas. Estas estimativas não são iguais devido à *variabilidade* presente na amostra. Se, no entanto, os diferentes valores obtidos para o estimador 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 estimar.

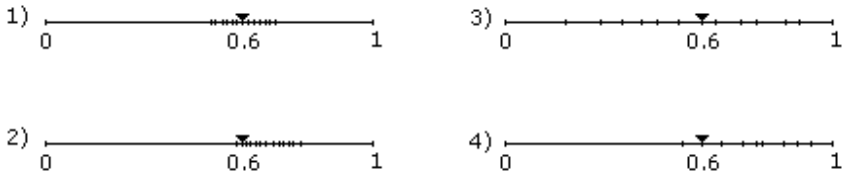
A **falta de precisão** e o problema do **enviesamento da amostra** são dois tipos de erro com que nos defrontamos num processo de amostragem (mesmo que tenhamos escolhido um "bom" estimador). Não se devem, contudo, confundir. Enquanto o enviesamento se manifesta por um desvio nos valores da estatística, 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, como veremos mais tarde. 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!

Aliás, são bem conhecidos alguns desastres, provocados por más amostras, de que o caso seguinte é um exemplo:

A sondagem de 1936 do Literary Disgest (Tannenbaum, 1998)
Nas eleições presidenciais de 1936 nos EUA, defrontaram-se Alfred Landon, o governador republicano do Kansas, e o presidente em exercício Franklin D. Roosevelt. Na altura da eleição a nação não tinha ainda recuperado da Grande Depressão. O Literary Digest, um dos jornais mais respeitados da época, conduziu uma sondagem durante duas semanas antes da eleição. Baseado nesta sondagem o jornal previu que Landon obteria 57% dos votos, contra 43% de Roosevelt. Os resultados da eleição foram 62% para Roosevelt contra 38% para Landon. Como foi possível uma discrepância destas? Na realidade a sondagem levada a cabo pelo Literary Digest foi uma das maiores e mais caras jamais conduzidas, baseada numa amostra de aproximadamente 2.4 milhões de pessoas. Para a mesma eleição a Gallup (Gallup Organization, www.gallup.com) baseada numa amostra muito mais pequena de aproximadamente 50000 pessoas, conseguiu prever a vitória de Roosevelt.
Como foi isto possível?
Comentário: A amostra do Literary Digest foi extraída de uma lista enorme constituída a partir do ficheiro de utentes de telefones, da listagem dos subscritores de jornais e revistas e dos membros das associações profissionais. A partir daí foi criada uma lista de 10 milhões de nomes, tendo sido enviado a cada um, um boletim de voto que deveria ser enviado para o jornal depois de preenchido. Na sua edição de 22 de Agosto de 1936, o Literary Digest apregoava: *Once again, [we are] asking more than then millions voters – one out of four, representing every county in the United States – to settle November's election in October. Next week, the first answers from these ten million will begin the incoming tide of marked ballots, to be triple-checked, verified, five-times cross-classified and totaled. When the last figure has been totted and checked, if past experience is a criterion, the country will know to within a fraction of 1 percent the actual popular vote of forty million (voters).*
A realidade foi bem mais dura! Após a eleição, com a credibilidade completamente desfeita e as vendas em baixo, o Literary Digest foi obrigado a fechar as portas, vítima de um passo em falso estatístico. A primeira coisa que estava errada nesta sondagem foi o processo de selecção para os nomes da lista a quem foi posta a questão, já que esta lista ficou constituída sobretudo por nomes de pessoas das classes média e alta. Em 1936 o telefone ainda era um luxo, assim como o era ser assinante de um jornal ou membro de uma associação profissional, numa altura em que havia 9 milhões de desempregados. Assim a amostra era grandemente *enviesada* e não era de modo nenhum representativa da população. Outro problema a considerar foi o facto de 10 milhões de pessoas terem sido contactadas e só cerca de 2.4 milhões terem respondido. Este problema da não resposta provoca um novo enviesamento, que é muito difícil de corrigir, já que num país livre não se pode obrigar as pessoas a responder, mesmo pagando, o que não melhoraria a situação, pois introduziria outras fontes de enviesamento.
Moral: É preferível utilizar uma amostra boa, ainda que dimensão pequena, do que uma grande amostra, mas má.

Exemplo - Suponhamos que ao pretender estudar a percentagem de eleitores que votariam favoravelmente num candidato à Câmara de determinada cidade, se recolhia uma amostra de 300 eleitores, dos quais 175 responderam que sim. Então uma estimativa para a proporção pretendida seria 0.58. Se considerássemos outra amostra de 300 eleitores, suponhamos que o valor obtido para o número de sim's tinha sido 181. Então o valor obtido para a estatística seria 0.60. A repetição deste processo 15

vezes permitiria obter 15 valores para a estatística, que seriam outras tantas estimativas do parâmetro a estimar - percentagem de eleitores da cidade, potenciais apoiantes do tal candidato. Representando num eixo os valores obtidos, poderíamos deparar-nos com várias situações:



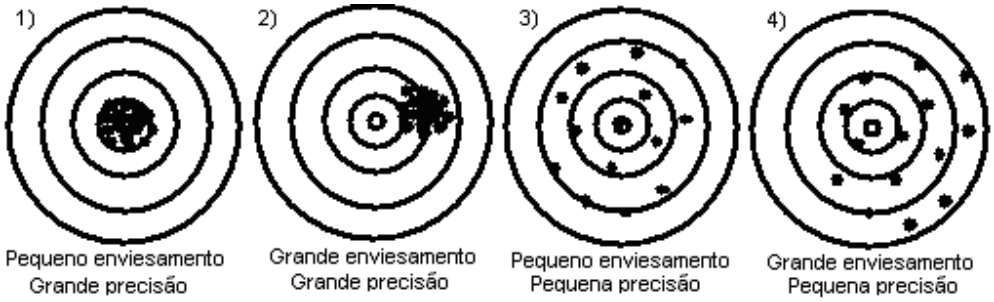
Se admitirmos que o valor do parâmetro é 0.60, então a situação:

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

A situação 2) deste exemplo poderia ter sido obtida se a selecção dos elementos para a amostra fosse feita em eleitores do mesmo partido que o candidato à Câmara, já que as amostras seriam enviesadas, e dariam origem a proporções amostrais superiores ao que seria de esperar com amostras seleccionadas de entre todos os eleitores possíveis. Mesmo que o estimador utilizado, ou seja, a *proporção amostral* seja um "bom" estimador do parâmetro *proporção populacional*, e mesmo que recolhêssemos amostras de dimensão razoável, os resultados sobrevalorizariam o valor do parâmetro a estimar.

Por outro lado uma selecção de amostras aleatórias, mas de pequena dimensão, poderia conduzir à situação 3), que apresenta grande variabilidade.

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



Qual a dimensão que se deve considerar para a amostra?

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

Chama-se a atenção para a existência de técnicas que permitem obter valores mínimos para as dimensões das amostras a recolher e que garantem estimativas com uma determinada **precisão** exigida à partida (como veremos mais à frente). 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*".

A seguir vamos ver dois casos importantes de estimação de parâmetros, nomeadamente:

- a estimação do **valor médio** (ou média populacional), pela **média** (amostral), e
- a estimação da **proporção populacional** pela **proporção amostral**.

3 - Estimação do valor médio

3.1 - Estimação do valor médio utilizando amostras aleatórias simples (sem reposição)

Quando se pretende estimar um **parâmetro**, uma vez definido o esquema de amostragem, considera-se uma ***estatística*** conveniente, isto é, uma função adequada das observações, função esta que para cada amostra observada dará uma ***estimativa*** do parâmetro que se pretende estimar. Quando o parâmetro a estimar é o valor médio ou média populacional, que se representa por μ, então é natural considerar como ***estimador*** a função ***média***, que se representa por $\overline{\mathbf{X}}$, e que para cada amostra observada dará uma estimativa $\bar{x}$ do valor médio μ.

Como é que podemos saber se a média é um "bom" estimador para o valor médio?

Será que para as diferentes amostras que podemos seleccionar da população, as diferentes médias dessas amostras são próximas umas das outras e do parâmetro valor médio? Se isso acontecer, temos uma certa garantia que a amostra que seleccionarmos, nos fornecerá uma estimativa razoável. A resposta à questão anterior é dada construindo a **distribuição de amostragem** da média.

São as distribuições de amostragem das *estatísticas* que nos vão permitir fazer inferências sobre os *parâmetros* correspondentes.

A aleatoriedade presente no processo de selecção das amostras, faz com que se possa utilizar a distribuição de amostragem de uma estatística para descrever o comportamento dessa estatística, quando se utiliza para estimar um determinado parâmetro. Podemos dizer que é através da distribuição de amostragem que introduzimos a probabilidade num procedimento estatístico, em que a partir das propriedades estudadas na amostra, procuramos tirar conclusões para a população.

3.1.1 - Distribuição de amostragem da média, como estimador do valor médio de uma População finita

Distribuição de amostragem exacta

Seguidamente vamos exemplificar o processo de obtenção da distribuição de amostragem da Média, e consequente estudo das suas propriedades como estimador do valor médio de uma População finita. Vamos considerar uma População de dimensão suficientemente pequena, para que o problema possa ser tratado dentro dos limites do razoável. Consideremos a seguinte população constituída pelos 9 alunos de uma classe infantil, sobre os quais se recolheram alguns dados:

Nº	Aluno	Peso (kg)	Altura (cm)	Nº irmãos
1	Maria	12.5	65	0
2	Teresa	11.6	68	1
3	Tiago	13.4	61	0
4	David	14.1	64	1
5	Rita	12.0	59	2
6	Ana	10.8	69	1
7	Joana	11.9	58	0
8	Bernardo	12.7	61	1
9	Leonor	9.6	63	1

Algumas características numéricas desta população são:

	Val. médio	Desvio padrão	Mín.	Máx.	Mediana
Peso	12.07	1.34	9.6	14.1	12
Altura	63.11	3.57	58	69	63
Nº irmãos	0.78	0.67	0	2	1

Esta população é tão pequena, que para a estudar não tivemos necessidade de recorrer a amostras para estimar alguns parâmetros desconhecidos, tais como altura média, peso médio, etc. Vamos, no entanto utilizá-la para exemplificar como se pode estimar a altura média a partir da média de amostras de dimensão 3. Como a nossa População tem dimensão 9, vamos utilizar a máquina de calcular para seleccionar números entre 1 e 9, tendo os elementos seleccionados sido o 5, o 2 e o 7, sobre os quais vamos recolher a informação relevante ou seja a altura:

Nº	Nome	Altura
5	Rita	59
2	Teresa	68
7	Joana	58

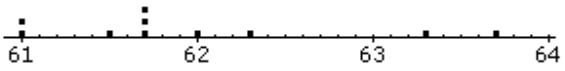
A média das alturas observadas é **61.7 cm**, que é uma estimativa da altura média da População.

Como neste caso conhecemos o valor do parâmetro, podemos dizer que a estimativa está razoavelmente próxima do parâmetro a estimar. Obviamente que se recolhermos outras amostras, obteremos outras estimativas. Então vamos seleccionar mais 9 amostras de dimensão 3, com o auxílio da máquina de calcular:

Amostra	1		2		3		4		5		6		7		8		9		10	
	5	59	1	65	8	61	7	58	2	68	1	65	8	61	6	69	3	61	5	59
	2	68	3	61	9	63	4	64	7	58	8	61	4	64	7	58	9	63	9	63
	7	58	8	61	3	61	6	69	4	64	5	59	7	58	5	59	5	59	2	68

Na obtenção das amostras anteriores tivemos o cuidado de fazer a selecção **sem reposição**, o que significa que ao obter cada amostra, um elemento seleccionado não poderia voltar a ser seleccionado. Também tivemos o cuidado de verificar se as amostras eram todas distintas (constituídas por pelo menos um elemento diferente). Os valores obtidos para as médias das 10 amostras foram:

Amostra	1	2	3	4	5	6	7	8	9	10
	61.7	62.3	61.7	63.7	63.3	61.7	61.0	62.0	61.0	61.5



Obtivemos vários valores diferentes como estimativas, sendo esta variabilidade resultado da variabilidade presente na amostra. Os valores apresentados pelas médias das 10 amostras, não diferem muito entre si, nem do valor do parâmetro. Mas como é que podemos ter a garantia que se recolhermos outra amostra, não vamos obter como estimativa do valor médio da altura, um valor muito diferente do verdadeiro valor do parâmetro? Por outras palavras, gostaríamos de poder responder à seguinte questão:

Para este processo de amostragem, como é que podemos concluir que a média é um "bom" estimador do valor médio (média populacional)?

Teremos de estudar a distribuição de amostragem da média, que neste caso consiste em estudar como se comporta a distribuição das médias obtidas para as $\binom{9}{3} = 84$ amostras diferentes, de dimensão 3, que se podem extrair da População.

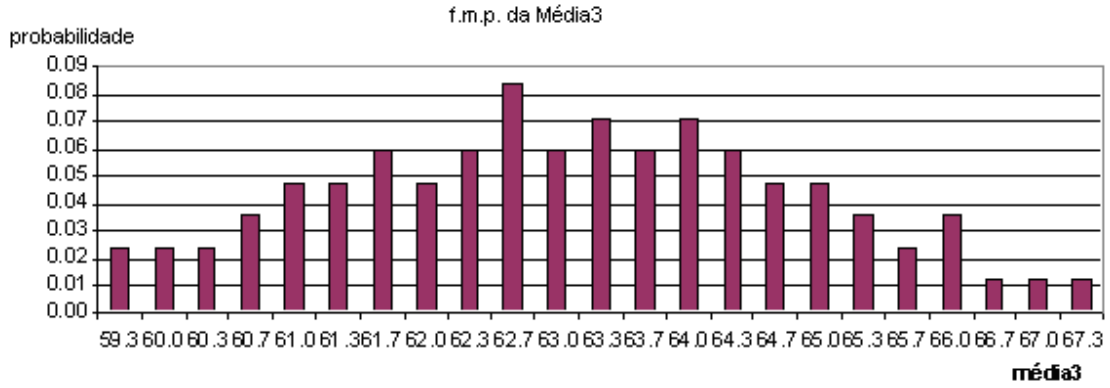
Considerando então todas as amostras aleatórias simples, diferentes, de dimensão 3, obtemos:

Am.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
	65	65	65	65	65	65	65	65	65	65	65	65	65	65	65	65	65	65	65	65	65
	68	68	68	68	68	68	68	61	61	61	61	61	61	64	64	64	64	64	59	59	59
	61	64	59	69	58	61	63	64	59	69	58	61	63	59	69	58	61	63	69	58	61
média	64.7	65.7	64.0	67.3	63.7	64.7	65.3	63.3	61.7	65.0	61.3	62.3	63.0	62.7	66.0	62.3	63.3	64.0	64.3	60.7	61.7
Am.	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**
	65	65	65	65	65	65	65	68	68	68	68	68	68	68	68	68	68	68	68	68	68
	59	69	69	69	58	58	61	61	61	61	61	61	61	64	64	64	64	64	59	59	59
	63	58	61	63	61	63	63	64	59	69	58	61	63	59	69	58	61	63	69	58	61
média	62.3	64.0	65.0	65.7	61.3	62.0	63.0	64.3	62.7	66.0	62.3	63.3	64.0	63.7	67.0	63.3	64.3	65.0	65.3	61.7	62.7
Am.	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**	**61**	**62**	**63**
	68	68	68	68	68	68	68	61	61	61	61	61	61	61	61	61	61	61	61	61	61
	59	69	69	69	58	58	61	64	64	64	64	64	59	59	59	59	69	69	69	58	58
	63	58	61	63	61	63	63	59	69	58	61	63	69	58	61	63	58	61	63	61	63
média	63.3	65.0	66.0	66.7	62.3	63.0	64.0	61.3	64.7	61.0	62.0	62.7	63.0	59.3	60.3	61.0	62.7	63.7	64.3	60.0	60.7
Am.	**64**	**65**	**66**	**67**	**68**	**69**	**70**	**71**	**72**	**73**	**74**	**75**	**76**	**77**	**78**	**79**	**80**	**81**	**82**	**83**	**84**
	61	64	64	64	64	64	64	64	64	64	64	59	59	59	59	59	59	69	69	69	58
	61	59	59	59	59	69	69	69	58	58	61	69	69	69	58	58	61	58	58	61	61
	63	69	58	61	63	58	61	63	61	63	63	58	61	63	61	63	63	61	63	63	63
média	61.7	64.0	60.3	61.3	62.0	63.7	64.7	65.3	61.0	61.7	62.7	62.0	63.0	63.7	59.3	60.0	61.0	62.7	63.3	64.3	60.7

Uma vez que o plano de amostragem considerado, foi a **amostragem aleatória simples**, **cada amostra tem igual probabilidade** (=1/84) de ser seleccionada, pelo que podemos considerar os diferentes valores obtidos para a variável Média, assim como as respectivas probabilidades – ou seja, estamos em condições de considerar a seguinte função massa de probabilidade para a variável Média, que vamos designar por Média3, para realçar o facto de as amostras a partir das quais se obtiveram os seus valores, terem dimensão 3:

Distribuição de Amostragem da Média para amostras de dimensão 3

Média3	**59.3**	**60.0**	**60.3**	**60.7**	**61.0**	**61.3**	**61.7**	**62.0**	**62.3**	**62.7**	**63.0**	**63.3**
Prob.	2/84	2/84	2/84	3/84	4/84	4/84	5/84	4/84	5/84	7/84	5/84	6/84
Média3	**63.7**	**64.0**	**64.3**	**64.7**	**65.0**	**65.3**	**65.7**	**66.0**	**66.7**	**67.0**	**67.3**	
Prob.	5/84	6/84	5/84	4/84	4/84	3/84	2/84	3/84	1/84	1/84	1/84	



Algumas propriedades da distribuição de amostragem da variável Média3 são:

	Valor médio	Desvio padrão	Mínimo	Máximo	Mediana
Média3	63.11	1.79	59.3	67.3	62.83

Repare-se que:

- o valor médio da variável Média3 (=63.11 cm) coincide com o valor médio da População – Altura (=63.11 cm), de onde se recolheram as amostras;
- o desvio padrão da variável Média3 (=1.79 cm) é bastante menor que o da População - Altura (=3.57 cm).

As propriedades anteriores permitem-nos concluir que a Média3, como estimador do parâmetro – valor médio da Altura, é um ***estimador centrado***, já que o seu valor médio, ou seja a média de todas as estimativas, para todas as amostras possíveis, coincide com o parâmetro a estimar.

A partir da distribuição de probabilidade da Média3, podemos ainda concluir que a probabilidade de obtermos estimativas no intervalo [61.3 cm, 65.3 cm] é de 0.75 (=63/84), assim como a probabilidade de obtermos essas estimativas no intervalo [60.0 cm, 66.7 cm] é superior a 0.95 (=80/84) ou 95%. Este resultado significa que, ao recolhermos uma amostra de dimensão 3 e ao calcularmos a partir dela uma estimativa para o valor médio, estamos ***confiantes***, com uma *confiança* superior a 95%, de que essa estimativa não se afasta do parâmetro a estimar de uma distância superior a 3.6 cm, aproximadamente (63.1–60.0=*3.1*; 66.7–63.1=*3.6*).

Chamamos a atenção para que a confiança anterior, não nos dá a garantia de que a estimativa que nós calculamos, para a amostra seleccionada, esteja naquele intervalo. Temos "fé" que sim! Já seria "azar" a amostra que nós seleccionamos ser uma das 4 que dá origem a estimativas fora do intervalo [60.0 cm, 66.7 cm]. Efectivamente, cerca de 5% das estimativas (=4/84) distam do parâmetro mais de 3.6 cm (2 distam *3.81*=63.11-59.3, 1 dista *3.89*=67-63.11 e 1 dista *4.19*=67.3-63.11).

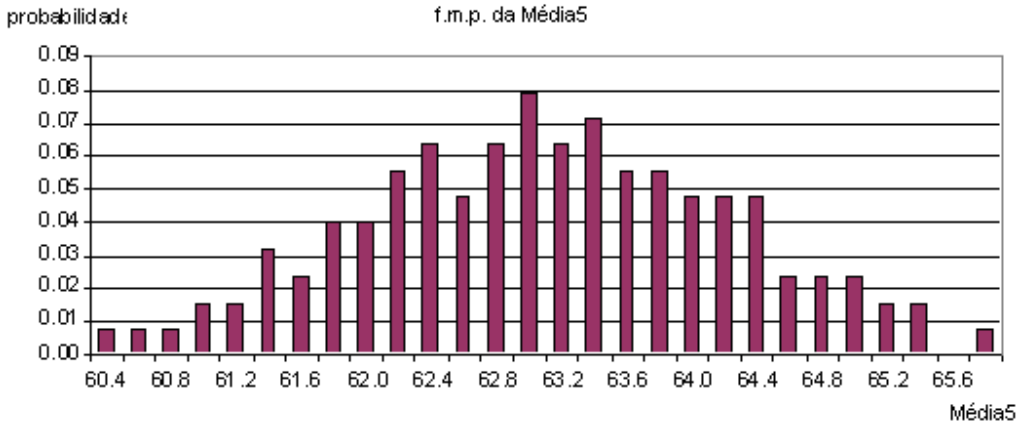
Se utilizarmos amostras de maior dimensão o que é que ganhamos?

Repetindo o processo anterior, mas agora para amostras de dimensão 5, será que a variabilidade apresentada pelo estimador diminui? Já que temos mais informação, é de esperar algum "ganho" na precisão do estimador!

Vamos então considerar a distribuição de amostragem da média para amostras de dimensão **5**. O processo é em tudo idêntico ao considerado anteriormente, mas agora será um pouco mais trabalhoso já que o número de amostras distintas, de dimensão 5, que podemos extrair da População de dimensão 9 é $\binom{9}{5} = 126$.

Os resultados obtidos para a distribuição de amostragem da média, para amostras de dimensão 5, foram:

Média5	60.4	60.6	60.8	61.0	61.2	61.4	61.6	61.8	62.0	62.2	62.4	62.6	62.8	63.0
Probab	0.008	0.008	0.008	0.016	0.016	0.032	0.024	0.040	0.040	0.056	0.063	0.048	0.063	0.079
Média5	63.2	63.4	63.6	63.8	64.0	64.2	64.4	64.6	64.8	65.0	65.2	65.4	65.8	
Probab	0.063	0.071	0.056	0.056	0.048	0.048	0.048	0.024	0.024	0.024	0.016	0.016	0.008	



Algumas propriedades da distribuição de amostragem da variável Média5 são:

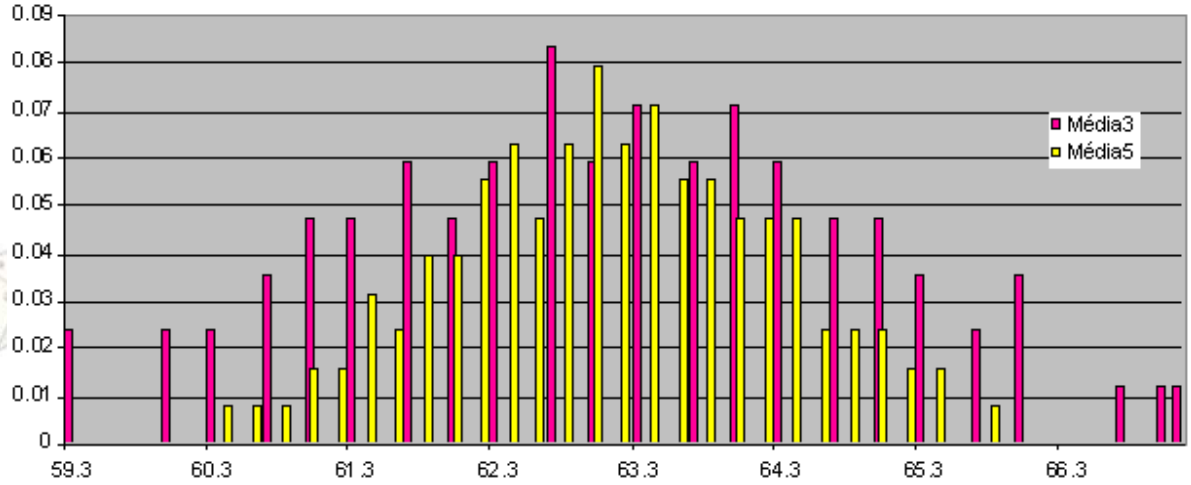
	Valor médio	Desvio padrão	Mínimo	Máximo	Mediana
Média5	63.11	1.13	60.4	65.8	63.1

Repare-se que:

- o valor médio da variável Média5 coincide com o valor médio da População – Altura, de onde se recolheram as amostras;
- o desvio padrão da variável Média5 (=1.13) é bastante menor que o da variável Altura (=3.57) e é ainda inferior ao da variável Média3 (=1.79).

Conclusão: a ***precisão*** do estimador aumenta, à medida que se aumenta a dimensão da amostra (Recordamos que quanto menor for a variabilidade apresentada pelo estimador, maior é a precisão).

Na figura seguinte apresentamos as distribuições de amostragem da Média3 e da Média5:



Como se verifica, a variabilidade é maior na distribuição de amostragem da média quando se consideram amostras de menor dimensão.

Resultado teórico (a demonstração do resultado seguinte, está fora do âmbito deste curso):

Dada uma População de dimensão N, de valor médio μ e variância σ^2, quando se considera um plano de amostragem aleatória simples, e como estimador de μ a Média, calculada a partir de amostras de dimensão **n**, então:

- O valor médio da Média é **μ**, isto é, a Média como estimador do valor médio é um estimador **centrado**;
- A variância da Média é igual a $\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$

A expressão obtida para a variância é muito interessante pela informação que contém. Nomeadamente:

- Confirma o que já havíamos esperado, no sentido de que ao **aumentar a dimensão** da amostra, **aumentamos a precisão** do estimador (na medida em que diminui a sua variabilidade).
- Permite-nos ainda concluir que, **para obter a mesma precisão**, quando estimamos o valor médio de Populações da mesma dimensão**, a dimensão da amostra terá de ser tanto maior, quanto maior for a variabilidade** presente na População.
- Mas mais interessante, embora menos intuitivo, permite-nos concluir **que se a dimensão da População for substancialmente maior que a da amostra**, então **a precisão do estimador não depende da dimensão dessa População**, mas unicamente da variabilidade aí presente (pois (N-n) / (N-1) ≈1).

Distribuição de amostragem aproximada

Os exemplos tratados anteriormente só têm interesse para exemplificar o processo de obter a distribuição de amostragem exacta da média, já que os valores considerados para a dimensão da população e da amostra são "ridiculamente" pequenos. Contudo, o processo utilizado deixa-nos adivinhar o trabalho árduo que teríamos se pretendêssemos fazer o mesmo com populações e amostras de dimensões razoáveis! Normalmente nas situações de interesse não se consegue obter a distribuição de amostragem exacta da média. Contudo o problema não é grave, já que, quando se faz a amostragem sem reposição, existem algumas condições necessárias e suficientes para que se possa aproximar a distribuição da média pela distribuição Normal. Não vamos apresentar essas condições, embora admitamos que elas estão satisfeitas e enunciamos o seguinte resultado:

Suponhamos que uma amostra aleatória simples é seleccionada de uma População de dimensão N, com valor médio μ[1] e variância σ^2. Então, se a dimensão n da amostra for suficientemente grande (um valor que é usual considerar como suficientemente grande é 30), a distribuição de amostragem da média pode ser aproximada pela distribuição Normal com valor médio $\boldsymbol{\mu}$ e variância $\frac{\sigma^2}{\mathbf{n}}(\frac{\mathbf{N-n}}{\mathbf{N-1}})$. A aproximação verifica-se para amostras de dimensão suficientemente grande, independentemente da forma da distribuição da População(1).

(1) Ao fazer esta afirmação, não podemos deixar de referir que a forma da distribuição da população subjacente tem alguma influência, no seguinte sentido: se tivermos duas populações, uma aproximadamente simétrica e outra apresentando um grande enviesamento, para amostras da mesma dimensão, a aproximação é melhor, quando estamos a estimar o valor médio da população simétrica. Para obtermos o mesmo grau de precisão da aproximação, no caso da outra população, seria necessário recolher uma amostra de maior dimensão.

3.1.2 - Distribuição de amostragem aproximada da média, como estimador do valor médio de uma População finita, mas de dimensão suficientemente grande

Na maior parte dos casos em que é necessário recolher uma amostra para estudar uma característica de uma População, não se conhece a dimensão desta. Então costuma-se assumir que é suficientemente grande de modo que se diz que se tem uma População de dimensão infinita. Em termos práticos costuma-se considerar que se tem uma população de dimensão infinita quando **N>20n**. Nestas condições o factor (N-n)/(N-1) que aparece na expressão da variância da Média toma um valor aproximadamente igual a 1,

$$(\frac{N-n}{N-1}) \approx 1 \quad \rightarrow \quad \frac{\sigma^2}{\mathbf{n}}(\frac{\mathbf{N-n}}{\mathbf{N-1}}) \approx \frac{\sigma^2}{\mathbf{n}}$$

[1] Estamos a identificar a População com a variável em estudo. Por essa razão dizemos que a População tem valor médio μ e variância σ^2.

pelo que temos o seguinte resultado, conhecido como **Teorema Limite Central** (TLC), de que o resultado anterior é uma versão para Populações finitas (que não possa ser assumida infinita, segundo as condições indicadas):

Suponhamos que uma amostra aleatória simples é seleccionada de uma População de dimensão grande, em que a variável em estudo tem valor médio μ e variância σ^2. Então, se a dimensão n da amostra for suficientemente grande (um valor que é usual considerar como suficientemente grande é 30), a distribuição de amostragem da média pode ser aproximada pela distribuição Normal com valor médio $\boldsymbol{\mu}$ e variância $\frac{\boldsymbol{\sigma^2}}{\mathbf{n}}$. A aproximação verifica-se para amostras de dimensão suficientemente grande, independentemente da forma da distribuição da População subjacente às amostras.

Mais uma vez chamamos a atenção para as seguintes propriedades, já anteriormente referidas:

- quanto maior for a dimensão da amostra, menor é a variabilidade apresentada pelo estimador. Ao desvio padrão da média dá-se o nome de **erro padrão**. Assim, esta propriedade pode ser enunciada do seguinte modo: quanto maior for a dimensão da amostra, menor será o erro padrão $\frac{\boldsymbol{\sigma}}{\sqrt{\mathbf{n}}}$;
- além disso, também concluímos que, para Populações de dimensão suficientemente grande, esta não tem influência sobre a variabilidade do estimador.

Em conclusão, a precisão de um estimador, para Populações de **grande dimensão**, não depende do tamanho da População, mas sim da variabilidade aí presente. **Quando pretendemos estimar um parâmetro da População, para obter uma determinada precisão, a dimensão da amostra terá de ser tanto maior, quanto maior for a variabilidade existente na População**. No entanto, se a dimensão da População já não for suficientemente grande, essa dimensão terá interferência na precisão do estimador, como vimos na secção anterior.

3.2 – Distribuição de amostragem da média, em amostragem com reposição

Será interessante estudarmos a distribuição de amostragem da Média, quando se faz amostragem **com reposição**, de uma População com dimensão N e comparar com o que se passa na amostragem **sem reposição**, tratada anteriormente.

Agora, cada elemento da População tem uma probabilidade constante e igual a 1/N de ser seleccionado para pertencer à amostra, já que quando um elemento é seleccionado, uma vez a informação recolhida, ele é novamente reposto na População. Este processo é equivalente a seleccionarmos uma amostra aleatória de dimensão n de uma população **uniforme discreta** no conjunto dos valores da característica a estudar da População, que podemos representar por x_1, x_2, …, x_N. Então cada vez que se selecciona um elemento da População é como se obtivéssemos um valor da variável

aleatória X que assume os valores x_i considerados anteriormente, com probabilidade 1/N. Seleccionar uma amostra de dimensão n significa seleccionar n variáveis X_1, X_2, …, X_n, independentes e com distribuição idêntica à de X. Então a Média

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

será uma variável aleatória, tal que:

- O valor médio da Média é μ [2], pelo que a Média, como estimador do valor médio μ, é um estimador **centrado**;
- A variância da Média é igual a $\frac{\sigma^2}{n}$, onde σ^2 é a variância da População.

Resumindo, se tivermos uma população finita de dimensão N, valor médio μ e variância σ^2, algumas características para a distribuição de amostragem da **Média** (de amostras de dimensão n) são:

	Sem reposição	Com reposição
Valor médio	μ	μ
Variância	$\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$ <	$\frac{\sigma^2}{n}$

Comparando os resultados anteriores, conclui-se que a **amostragem sem reposição é mais eficiente**, quando se pretende estimar o valor médio da População, uma vez que produz um estimador com uma variância mais pequena, isto é, que apresenta menor variabilidade.

Exemplo – Considere uma população constituída pelos elementos 1, 2, 3, 4 e 5. Pretende estimar o valor médio desta população, pelo que decide recolher uma amostra de dimensão 2, com reposição e calcular a sua média. Obtenha a distribuição de amostragem do estimador utilizado para estimar o valor médio da população.
Resolução: A População anterior é constituída pelos elementos 1, 2, 3, 4 e 5, tendo cada um uma probabilidade constante e igual a 1/5 de ser seleccionado para pertencer a uma amostra:

População X	1	2	3	4	5
Probabilidade	1/5	1/5	1/5	1/5	1/5

Propriedades da População:

Valor médio = 3 e **Desvio padrão** = $\sqrt{2}$.

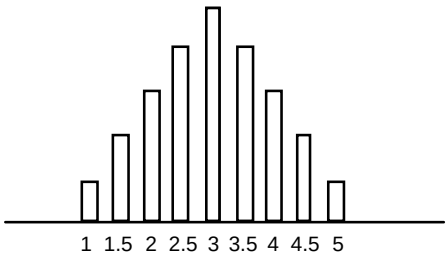
[2] Propriedade do valor médio, segundo a qual o valor médio de uma soma de variáveis aleatórias é igual à soma dos valores médios das parcelas.

A metodologia seguida para obter a distribuição de amostragem consiste em seleccionar todas as amostras de dimensão 2, com reposição, calcular o valor da estatística média para cada uma delas e depois representar a distribuição dos valores obtidos:

Amostras	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)
		(2,1)	(2,2)	(2,3)	(2,4)	(3,4)	(4,4)	(5,4)	
			(3,1)	(3,2)	(3,3)	(4,3)	(5,3)		
				(4,1)	(4,2)	(5,2)			
					(5,1)				
média	**1**	**1.5**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**

De acordo com a tabela anterior obtemos a seguinte distribuição de amostragem para o estimador Média2 (assim representado por se obter a partir de amostras de dimensão 2)

Média2	1	1.5	2	2.5	3	3.5	4	4.5	5
Probabilidade	1/25	2/25	3/25	4/25	5/25	4/25	3/25	2/25	1/25



Características da distribuição de amostragem da Média para amostras de dimensão 2:

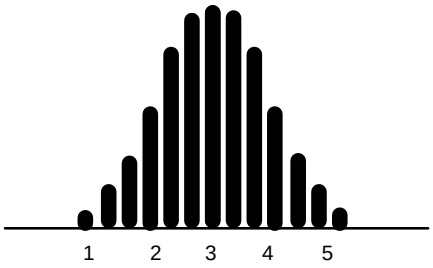
Valor médio = 3 e **Desvio padrão = 1**

Algumas observações:

- O valor médio da distribuição de amostragem do estimador Média, utilizado para estimar o valor médio da população (igual a 3), coincide com o parâmetro a estimar .
- O desvio padrão da população inicial é igual a $\sqrt{2}$, enquanto que o desvio padrão da média, calculada a partir de amostras de dimensão 2 é 1 ($\sqrt{2}/\sqrt{2}=1$ – resultado considerado anteriormente).

Se repetirmos a metodologia seguida no processo do exemplo anterior, considerando agora amostras de dimensão 3, o problema torna-se mais trabalhoso, já que o número de amostras possíveis é $5^3=125$. Assim, abstemo-nos de apresentar todas essas amostras, limitando-nos a apresentar a distribuição de amostragem da Média3:

Média3	1	1.33	1.67	2	2.33	2.67	3	3.33	3.67	4	4.33	4.67	5
Proba.	.008	.024	.048	.080	.120	.144	.152	.144	.120	.080	.048	.024	.008



Características da distribuição de amostragem:

Valor médio = 3 e **Desvio padrão = 0.816**

Algumas observações:

- O valor médio da distribuição de amostragem do estimador Média3 utilizado para estimar o valor médio da população (igual a 3), coincide com o parâmetro a estimar .
- O desvio padrão da população inicial é igual a $\sqrt{2}$, enquanto que o desvio padrão da Média3, calculada a partir de amostras de dimensão 3 é 0.816 ($\sqrt{2}/\sqrt{3}=0.816$ – o que condiz com o resultado apresentado anteriormente, de que a variância da Média é σ^2/n).
- A variabilidade apresentada pela distribuição de amostragem é inferior à obtida quando se consideram amostras de dimensão 2. Este resultado indicia que quanto maior for a dimensão da amostra, menor é a variabilidade apresentada pela distribuição de amostragem.

O que acontece se a dimensão da população for grande?

Se a dimensão da População for razoavelmente grande, a probabilidade de extrairmos o mesmo elemento duas vezes é extremamente pequena (Por exemplo, numa população de dimensão 1000, a probabilidade de extrairmos amostras de dimensão 2, com elementos iguais, seria 0,001). Assim, os dois processos de amostragem, **com reposição** *e* **sem reposição***, são praticamente equivalentes*, quando estamos a estimar o valor médio.

Esta conclusão vai de encontro com a que se pode obter também se tomarmos atenção às variâncias das Médias de amostras de dimensão n, quando se faz extracção **com** e **sem** reposição. Efectivamente o factor

$$\frac{N-n}{N-1} = \frac{N}{N-1} \times \left(1-\frac{n}{N}\right)$$

que aparece na expressão da variância num processo de **amostragem aleatória simples** (sem reposição) assume um valor próximo de 1, quando N é razoavelmente grande e n é razoavelmente pequeno, quando comparado com N. Ao quociente $\frac{n}{N}$ costuma-se chamar **fracção de amostragem**.

Já apontámos anteriormente que, em termos práticos, se considera uma População "grande" se a sua dimensão for cerca de 20 vezes superior à dimensão da amostra, ou

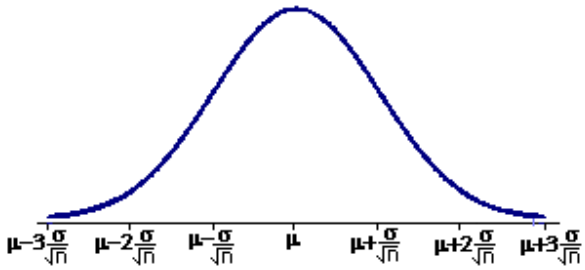
seja, quando a fracção de amostragem for menor que 5%. Então, na prática, temos o seguinte resultado:

Se a população tiver **dimensão grande**, é praticamente indiferente fazer a recolha da **amostra com reposição** ou **sem reposição**, quando se estão a estudar as propriedades da média, como estimador do valor médio!

No que diz respeito à forma da distribuição de amostragem da média, invocando mais uma vez o **Teorema Limite Central** (TLC), temos:

Suponhamos que uma amostra aleatória, de dimensão n, é seleccionada, com reposição (se a População tiver dimensão N, grande, e N>20xn, a selecção pode ser feita sem reposição), de uma população em que a variável em estudo tem valor médio μ e variância σ^2. Então, se a dimensão n da amostra for suficientemente grande (um valor que é usual considerar como suficientemente grande é 30), a distribuição de amostragem da média pode ser aproximada pela distribuição Normal com valor médio $\boldsymbol{\mu}$ e variância $\frac{\boldsymbol{\sigma^2}}{\mathbf{n}}$. A aproximação verifica-se para amostras de dimensão suficientemente grande, independentemente da forma da distribuição da População subjacente às amostras.

Assim, o modelo Normal, centrado em μ e com desvio padrão $\sigma/\sqrt{n}$, é um bom modelo para o conjunto das médias de todas as amostras aleatórias (ver na caixa anterior as condições), de dimensão n, que se podem seleccionar de uma população com valor médio μ e desvio padrão σ:



Propriedades:

- quanto maior for a dimensão da amostra, menor é a variabilidade apresentada pelo estimador;
- além disso, também concluímos que, para Populações de dimensão suficientemente grande, esta não tem influência sobre a variabilidade do estimador.

Em conclusão, a precisão de um estimador, para Populações de **grande dimensão**, não depende do tamanho da População, mas sim da variabilidade aí presente. **Quando pretendemos estimar um parâmetro da População, para obter uma determinada precisão, a dimensão da amostra terá de ser tanto maior, quanto maior for a variabilidade existente na População**. No entanto, se a dimensão da

População já não for suficientemente grande, essa dimensão terá interferência na precisão do estimador, como vimos na secção anterior.

O teorema limite central dá-nos uma justificação teórica para a grande utilização da distribuição Normal, como modelo de fenómenos aleatórios. Quantidades tais como alturas e pesos de uma população relativamente homogénea, podem ser consideradas como somas de um grande número de causas genéticas e efeitos devido ao meio ambiente, mais ou menos independentes entre si, cada um contribuindo com uma pequena quantidade para a soma.

O que é que se entende por um valor de n suficientemente grande?

Uma questão que se pode pôr é a seguinte: quando queremos aplicar o teorema do limite central "*qual o valor de n, para que se possa utilizar a distribuição Normal, como uma "boa" aproximação para a distribuição de amostragem pretendida"*?

Este valor de n depende, em certa medida, da distribuição subjacente à amostra e será tanto maior quanto mais enviesada for a distribuição da população (o termo enviesado aplica-se como contrário a simétrico). No entanto é usual referir que um valor igual ou superior a 30 já permite fazer a aproximação com uma precisão razoável.

Nota (Moore et al. 1993): o facto de a média de várias medições apresentar menor variabilidade do que uma única medição, é bastante importante em ciência. Quando Simon Newcomb mediu a velocidade da luz, fez repetidas vezes a medição do tempo necessário para um raio de luz percorrer determinada distância. As 64 observações que ele reteve,

28	22	36	26	28	28	26	24	32	30	27	24	33	21	36	32
31	25	24	25	28	36	27	32	34	30	25	26	26	25	23	21
30	33	29	27	29	28	22	26	27	16	31	29	36	32	28	40
19	37	23	32	29	24	25	27	24	16	29	20	28	27	39	23

podem ser consideradas valores de 64 variáveis aleatórias independentes, cada uma com uma distribuição de probabilidade que descreve a população de todas as medições feitas utilizando o procedimento de Newcomb. Se este processo de Newcomb estiver correcto, a média populacional μ é o verdadeiro valor do tempo que a luz leva a percorrer a distância escolhida (ir do seu laboratório no Protomac Rives até um espelho na base do Washington Monument e voltar, numa distância total de cerca de 7400 metros). A variabilidade populacional reflecte a variação aleatória nas medições, devida a pequenas modificações no meio envolvente, no equipamento, e no procedimento. Suponha que o desvio padrão desta população é σ=5 segundos $\times 10^{-9}$ (unidade utilizada nas medições).

Se Newcomb tivesse feito uma única medida, o desvio padrão do resultado seria 5, pelo que uma outra medição poderia ter um valor substancialmente diferente. Tomando as 64 medições como ele fez, a média tem um desvio padrão de $5/\sqrt{64.}$=0.625. O valor de 27.75 que Newcomb obteve para a média das suas 64 observações é muito mais fiável que o obtido a partir de uma única observação.

Exercícios

2.3.1 –Considere a população dos deputados da X Legislatura e considere os dados referentes à variável Idade.

a) Calcule o valor médio e o desvio padrão das idades.
b) Organize os dados na forma de uma tabela de frequências, considerando como classes os diferentes valores obtidos para as idades (Chama-se a atenção para o facto da variável Idade ser de natureza contínua e quando falamos, por exemplo, na classe dos 45 anos, estamo-nos a referir a um intervalo, representado pelo valor 45, que inclui todas as idades dos indivíduos que acabaram de fazer 45 anos, mas ainda não fizeram 46). Calcule a probabilidade de um deputado, escolhido ao acaso, ter 40 ou menos anos.
c) Obtenha uma estimativa para a probabilidade da média das idades de 30 deputados, seleccionados aleatoriamente (com reposição), ser igual ou menor que 45 anos.
d) Obtenha uma estimativa para a probabilidade da média das idades de 50 deputados, seleccionados aleatoriamente (com reposição), ser igual ou menor que 45 anos.
e) Compare os valores obtidos nas 3 alíneas anteriores e tire conclusões. Justifique as conclusões a que chegou.

2.3.2 – Quando pretende estimar o valor médio de uma população e aumenta a dimensão da amostra, a probabilidade de obter uma média com maior precisão:

Aumenta? Diminui? Fica na mesma?

2.3.3 – Diga se a seguinte afirmação é Verdadeira ou falsa: Diminui de metade o erro padrão, aumentando para o dobro a dimensão da amostra.

2.3.4 – Perguntou-se a 835 portugueses adultos quanto esperavam gastar nas prendas de Natal do ano em curso. A média obtida foi de 540 euros.

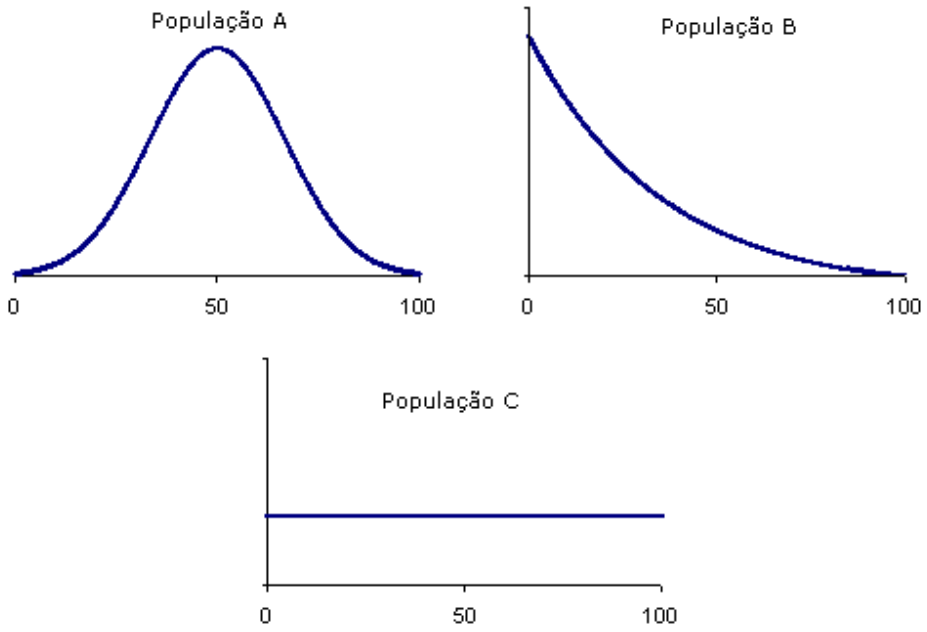
a) O valor de 540 euros é um parâmetro ou uma estatística?
b) Identifique a população em estudo e o parâmetro de interesse.
c) O facto de ter obtido 540 euros para a média, significa que o gasto médio da população nas compras de Natal tenha de ser necessariamente 540 euros?
d) Acharia "razoável" ter obtido como média 540 euros, se o gasto médio da população, nas compras de Natal, fosse 550 euros? E se fosse 750 euros? Justifique a sua resposta.
e) Admitindo que o gasto médio da população nas compras de Natal é de 550 euros e o desvio padrão dos gastos é de 150 euros:
    i) qual a distribuição de amostragem aproximada para a média de amostras de dimensão 835?
    ii) calcule um valor aproximado para a probabilidade de obter um valor para a média menor ou igual a 540 euros.
f) Tendo em consideração os dados da alínea anterior, calcule a probabilidade de uma pessoa, escolhida ao acaso, gastar nas suas compras 540 ou menos, euros.
g) Admita agora que o gasto médio nas compras de natal é de 550 euros, mas o desvio padrão dos gastos é igual a 250 euros.
    i) Qual a distribuição de amostragem aproximada para a média de amostras de dimensão 835?
    ii) Qual a probabilidade de obter para a média um valor menor ou igual a 540 euros?
    iii) Compare o valor obtido na alínea anterior com o que obteve na alínea e) ii). Justifique as conclusões a que chegou.

2.3.5 – Com o objectivo de verificar se efectivamente os pacotes de 150 gramas de batatas fritas da marca Fri-Fri, tinham o peso anunciado, recolheram-se aleatóriamente 30 pacotes que se pesaram, tendo-se obtido os seguintes resultados:

148	158	146	139	148	147
143	139	141	147	148	159
151	148	140	147	156	154
156	155	145	150	149	161
155	144	146	148	149	146

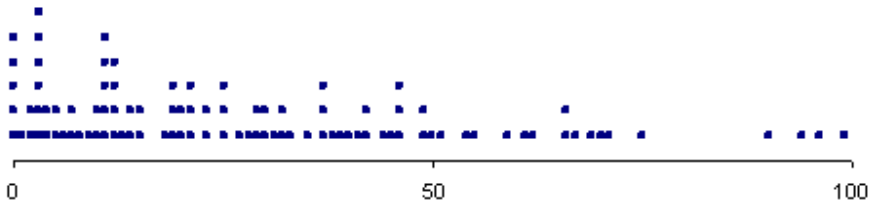
a) Calcule a média dos pesos obtidos. Esse valor é um parâmetro ou uma estatística?
b) O facto de não ter obtido um valor para a média igual a 150 gramas, significa que o peso médio dos pacotes não possa ser de 150 gramas? Justifique a sua resposta.
c) Calcule o desvio padrão (amostral) dos pesos obtidos. O valor que obteve é uma estimativa de quê?
d) Obtenha a distribuição de amostragem aproximada da média de amostras dimensão 30, de pesos de pacotes de batatas fritas.
e) Admitindo que o peso médio dos pacotes de batatas fritas é efectivamente 150 gramas, calcule um valor aproximado para a probabilidade da média dos pesos de 30 pacotes ser inferior a 150 gramas. Poderia calcular o valor exacto para essa probabilidade?
f) Obtenha um valor aproximado para a probabilidade da média dos pesos de 30 pacotes, seleccionados aleatoriamente, se afastar do pressuposto peso médio de 150 gramas, de 2 gramas.

2.3.6 – (Sugerido por Rossman, 2001) Considere as seguintes funções densidades que modelam as populações constituídas pelos resultados obtidos (numa escala de 0 a 100) em 3 testes:

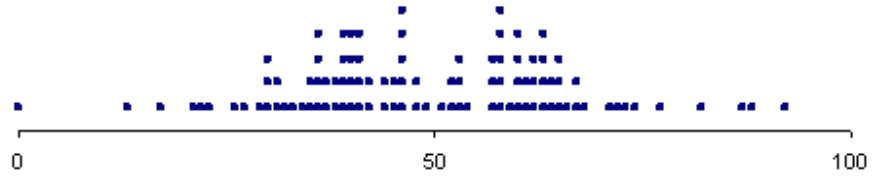


Os seguintes diagramas de pontos representam a distribuição dos valores de 3 amostras de dimensão 100, cada uma extraída de uma das 3 populações anteriores:

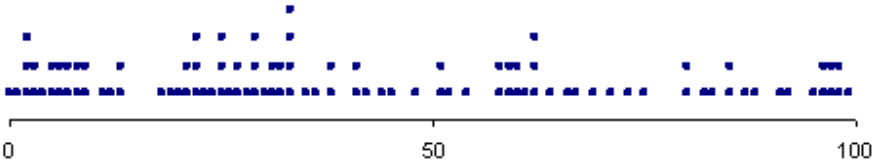
Amostra 1



Amostra2



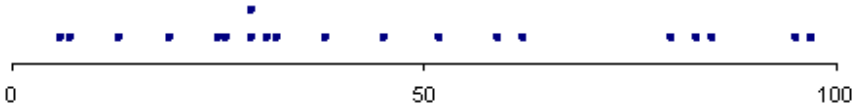
Amostra3



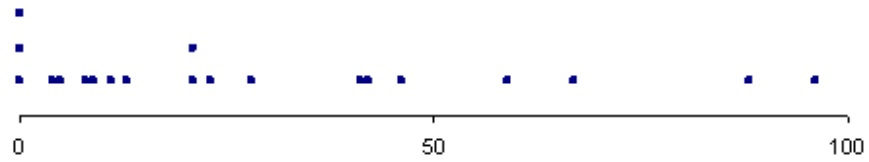
a) Identifique a população de onde foi seleccionada cada uma das amostras anteriores.

b) Os diagramas de pontos seguintes apresentam a distribuição de 3 amostras, de dimensão 20, seleccionadas também das 3 populações dadas inicialmente. Tem agora a mesma facilidade, em distinguir de que população se seleccionou cada uma delas?

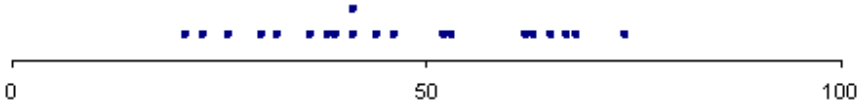
Amostra 1



Amostra 2



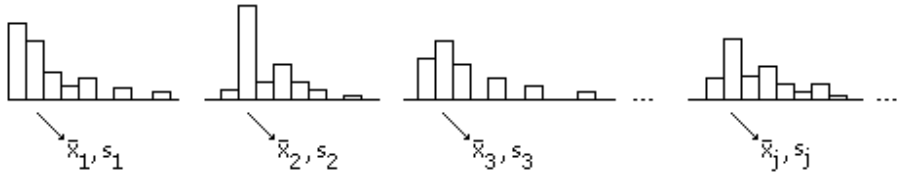
Amostra 3



Concorda que com uma amostra de maior dimensão se consegue visualizar melhor a estrutura da população subjacente?

c) Suponha que pretende estimar o valor médio da população a que corresponde o modelo normal. Para cada uma das amostras seleccionadas da população normal, calculou a média e obteve os valores 46,05 e 49,05. Qual destes valores pensa que foi obtido a partir da amostra de maior dimensão? Justifique a sua resposta.

2.3.7 – Considere a população B do exemplo anterior e represente por μ e σ, respectivamente o seu valor médio e desvio padrão. Suponha que selecciona várias amostras de dimensão n, representa-as graficamente, e para cada uma dessas amostras calcula a média e o desvio padrão:

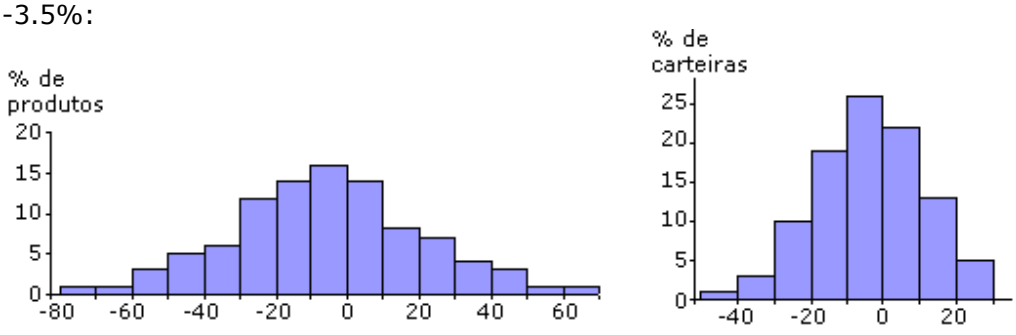


a) Faça um esquema de um histograma que represente adequadamente todas as médias que obteve.
b) Qual o modelo que pode utilizar para modelar o histograma que obteve na alínea anterior?
   i) Teve de fazer algumas hipóteses para utilizar o modelo proposto?
   ii) Quais os parâmetros valor médio e desvio padrão do modelo considerado?
   iii) Se não conhecer o desvio padrão, σ, da população, como é que o pode estimar?
c) O facto de a população de onde se recolhem as amostras, ter uma distribuição enviesada, tem algumas consequências na obtenção da distribuição de amostragem (aproximada) da média?
d) Explique por algumas palavras, qual a diferença entre distribuição de amostragem e distribuição da amostra.

2.3.8 – (Adaptado de Velleman, 2004) Suponha que a duração de uma gravidez (humana) pode ser bem modelada por uma Normal com valor médio igual a 266 dias e desvio padrão igual a 16 dias.

a) Qual a percentagem de mulheres cuja gravidez tem uma duração entre 270 e 280 dias?
b) Suponha que um obstrecta presta assistência regularmente a 60 grávidas. Qual a distribuição de amostragem da média do tempo de gravidez de 60 grávidas? Especifique o modelo, o seu valor médio e o seu desvio padrão.
c) Qual a probabilidade de que a média do tempo de gravidez de 60 mulheres, seja inferior a 160 dias?

2.3.9 – Um princípio básico quando se quer investir na bolsa, é o de diversificar os investimentos de modo a reduzir o risco. A figura a seguir, à esquerda, mostra a distribuição dos retornos para todos os 1815 produtos da bolsa no ano de 1987. Este foi um ano "negro" para os investidores. O retorno médio para os investimentos foi de -3.5%:



Na figura da direita apresenta-se a distribuição dos retornos para todas as possíveis carteiras que tenham investido iguais quantidades em 5 produtos. Uma carteira é uma amostra de 5 produtos e o retorno é dado pela média dos 5 produtos escolhidos. Embora o retorno médio seja o mesmo, qual a vantagem de investir em carteiras?

4 - Estimação da proporção

4.1 - Distribuição de amostragem da proporção amostral, como estimador da proporção populacional

Anteriormente estudámos a estimação do valor médio e vamos, neste capítulo, ver como os resultados que se obtiveram podem ser traduzidos para o estudo da estimação do parâmetro *proporção* de elementos da População, que satisfazem determinada propriedade ou pertencem a determinada categoria.

Consideremos então uma população de dimensão N e seja **p** a proporção (desconhecida) de elementos da população que pertencem à categoria em estudo. Na metodologia que vamos utilizar, no estudo da estimação da proporção, começamos por verificar que uma proporção é uma média de 0's e 1's em que atribuímos o valor **1** a um elemento da população que pertença à categoria em estudo e o valor **0** a um elemento que não pertença a essa categoria. Assim, a ***proporção p*** não é mais do que o ***valor médio*** desta população cujos elementos são 0's e 1's, pelo que o estudo feito para a estimação do valor médio será facilmente adaptado para a estimação da proporção.

Para esta população tão particular, constituída por 0's e 1's, em que a proporção populacional é a média populacional, a **proporção amostral** também será a **média** (amostral), que será assim, o estimador intuitivo para a proporção populacional.

Como no capítulo anterior estudámos a distribuição de amostragem da média, tendo concluído que a média é um "bom" estimador para o valor médio, imediatamente concluímos que a **proporção amostral** é um "**bom**" estimador para a **proporção populacional**.

A fim de utilizar os resultados enunciados para a distribuição de amostragem da média, vejamos a que é igual a variância de uma população constituída por 0's e 1's em que a percentagem de 1's é **p**.

Valor médio μ e variância σ^2 da população em estudo:

Dada uma população de N elementos, em que cada elemento ou é 0 ou é 1, sendo p a percentagem de elementos 1's (elementos pertencentes à categoria em estudo)

Classe	Freq.abs.	Freq.rel.
1	Np	p
0	N(1-p)	1-p
Total	N	1

Para esta população, é imediato que o valor médio μ é igual a p (=1*p+0*(1-p)), e a partir da expressão da variância, temos que

$$\sigma^2 = (1-p)^2*p+(0-p)^2*(1-p)$$

$$\sigma^2 = p\,(1-p)$$

O valor médio e a variância de uma população constituída por 0's e 1's, em que a proporção de 1's é p, é igual a **p** e a **p(1-p)**, respectivamente.

As conclusões a que chegámos no capítulo anterior, permitem-nos agora enunciar os seguintes resultados (obtidos a partir dos resultados obtidos para a média):

Dada uma população de dimensão N, em que **p** é a percentagem de elementos da população que verificam determinada característica, quando se considera um esquema de **amostragem aleatória simples**, ou um esquema de amostragem **com reposição**, e como estimador do parâmetro **p**, a proporção amostral $\hat{p}$, isto é a proporção de elementos pertencentes à categoria em estudo, existente em amostras de dimensão n, então:

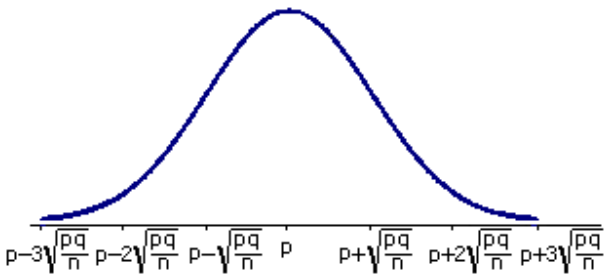
- O estimador $\hat{p}$ de p é um estimador centrado, já que o seu valor médio coincide com p, $E(\hat{p}) = p$
- $Var(\hat{p}) = \frac{\mathbf{p(1-p)}}{\mathbf{n}}(\frac{\mathbf{N-n}}{\mathbf{N-1}})$ no esquema de <u>amostragem aleatória simples</u>

  e $Var(\hat{p}) = \frac{\mathbf{p(1-p)}}{\mathbf{n}}$ num esquema de <u>amostragem aleatória com reposição</u>.

O resultado teórico conhecido como Teorema Limite Central permite-nos, agora, apresentar o seguinte resultado:

Suponhamos que se selecciona uma **amostra aleatória simples** de uma População de **dimensão grande**, ou que se selecciona uma amostra aleatória, **com reposição** de uma população de dimensão qualquer, em que a característica em estudo está presente numa proporção **p** (desconhecida). Então, se a **dimensão n da amostra for suficientemente grande** (um valor que é usual considerar como suficientemente grande é 30), a distribuição de amostragem da proporção amostral $\hat{p}$ pode ser aproximada pela distribuição Normal com valor médio p e variância $\frac{p(1-p)}{n}$.

Assim, o modelo Normal, centrado em p e com desvio padrão $\sqrt{\frac{pq}{n}}$, onde representamos por q=1-p, é um bom modelo para o conjunto das proporções obtidas a partir de todas as amostras aleatórias (ver na caixa anterior as condições), de dimensão n, que se podem seleccionar da população, em que a característica em estudo existe com uma proporção p:



Recordamos algumas das condições para aplicar o modelo anterior:

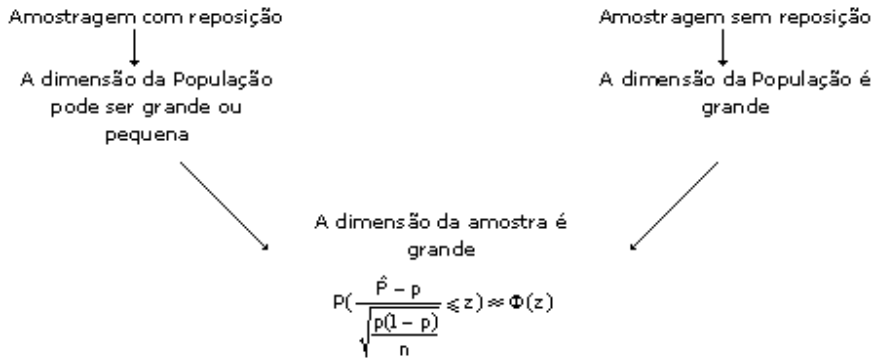
a) Qualquer que seja o processo de selecção da amostra aleatória, isto é, com reposição ou sem reposição, a proporção amostral $\hat{p}$ é sempre um estimador centrado da proporção populacional **p**, isto é, o valor médio da sua distribuição de amostragem é **p**.

b) Se a amostragem se fizer com reposição, então existe independência entre a selecção dos elementos que vão constituir a amostra, na medida em que a probabilidade de um qualquer elemento ser seleccionado, não depende do elementos que já tiverem sido seleccionados. A variância do estimador vem $\frac{p(1-p)}{n}$.

c) Se a amostragem se fizer sem reposição, então já a dimensão N, da população pode interferir nas propriedades do estimador, a não ser que essa dimensão seja "grande", isto é, N>20n, pois neste caso a probabilidade de um mesmo elemento ser seleccionado 2 vezes é muito pequena. Se efectivamente N for grande, podemos ainda utilizar para a variância da proporção amostral a expressão $\frac{\mathbf{p(1-p)}}{\mathbf{n}}$ e não depende da dimensão da população.

d) Se na amostragem sem reposição, N for grande e a dimensão da amostra for suficientemente grande, podemos aproximar a distribuição de amostragem da proporção, pela distribuição Normal.

Resumindo o que acabámos de dizer, temos



Notação: Não esqueça que a notação para parâmetro e estatística é diferente. Assim

	Parâmetro (população)	**Estatística** (amostra)
Proporção	p	$\hat{p}$
Valor médio	μ	$\bar{x}$
Desvio padrão	σ	s

Exercícios

2.4.1 – Na sua escola pretende-se averiguar qual a proporção de alunos que gostaria que se realizasse uma festa de Natal. Recolhem-se aleatoriamente 2 amostras, uma de dimensão 50 e outra de dimensão 100. Tem a garantia que a proporção de respostas favoráveis à realização da festa, obtida a partir da amostra de 100 alunos, esteja mais perto da verdadeira proporção de alunos favoráveis à festa, do que a obtida a partir da amostra de dimensão 50? Explique a sua resposta.

2.4.2 – O Conselho Directivo da escola pretende averiguar qual a percentagem de alunos que utilizaria regularmente (3 ou mais vezes por semana) a cantina, para almoçar, no caso de esta começar a fornecer almoços. Encarregou uma comissão de alunos de fazer um estudo sobre este problema. Esta comissão recolheu informação junto de 100 alunos da escola e elaborou a seguinte tabela de frequências:

Quantas vezes almoçarias na Escola?	0	1	2	3	4	5
Nº de respostas	10	12	18	40	12	8

a) Identifique o parâmetro em estudo.
b) Qual a percentagem de alunos que pensam utilizar regularmente a cantina?
c) A proporção obtida anteriormente é uma estatística ou um parâmetro?
d) Se em vez de uma amostra de dimensão 100, recolhesse uma amostra de dimensão 150, teria uma maior probabilidade de obter um valor para a proporção amostral, mais perto do parâmetro? Explique.

2.4.3 – Uma máquina de leitura óptica tem uma probabilidade de 1% de cometer erro(s) na leitura de uma folha. Esta máquina é utilizada para processar a informação contida nas fichas dos alunos (uma ficha por cada aluno) de uma Universidade. De cada vez que a máquina é utilizada, processa a leitura de lotes de 100 fichas.

a) Como se distribui a proporção de fichas lidas erradamente pela máquina, por lote?
b) A máquina será rejeitada se a probabilidade de ler 3 ou mais fichas erradas, em cada utilização, for superior a 5%. Será de rejeitar a máquina?

2.4.4 – Pretende-se averiguar qual a percentagem de alunos da escola que têm computador. Recolheu-se uma amostra de 50 alunos e verificou-se que 25% dos alunos tinham computador.

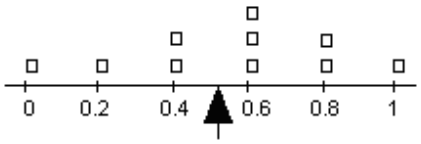
a) O valor 25% é um parâmetro ou uma estatística?
b) Obtenha uma estimativa para a probabilidade e em 50 alunos, 30 ou mais terem computador.

2.4.5 – Considere a população de deputados da X Legislatura, que se encontra em Anexo. Seleccione aleatóriamente 5 deputados, e registe as observações na seguinte tabela:

Número	Nome	Sexo	Partido	Circulo Eleitoral	Idade em 31/12/2001

a) Algum dos deputados seleccionados é seu conhecido?
b) Qual a proporção de deputados do PS na sua amostra? Esta proporção é igual a 52.6% (proporção de deputados do PS na X Legislatura)?

c) Se a sua resposta à questão anterior foi "não" isso significa que a amostra é enviesada?
d) Peça aos seus colegas que seleccionem também 5 deputados e registe o valor obtido para a proporção de deputados do PS para cada amostra seleccionada. Marque os valores obtidos num gráfico.
Sugestão: Sugere-se uma representação gráfica como a que se apresenta a seguir



onde se verifica que, por exemplo, de 10 amostras consideradas, se obteve: 1 vez, uma percentagem de 0 deputados do PS; 1 vez, uma proporção de 0.2 deputados do PS, 2 vezes uma proporção de 0.4 deputados do PS, etc. A seta indica a posição da percentagem de deputados do PS na população considerada.
e) Quantos dos seus colegas obtiveram uma percentagem de deputados do PS superior a 52.6%? E quantos obtiveram uma percentagem inferior?
f) Tendo em consideração os resultados obtidos na alínea anterior, pensa que a proporção amostral é um estimador centrado da proporção populacional?

Sugestão: Sugere-se uma representação gráfica como a que se apresenta a
Reparou que: o valor obtido pela estatística varia de amostra para amostra? A esta propriedade chamamos variabilidade amostral.

2.4.6 - Considere a população constituída pelas pastilhas XPTO fabricadas pela XPTO Lda. Estas pastilhas podem ser de 3 cores, e pretende-se averiguar algo sobre a forma como se distribuem as cores.

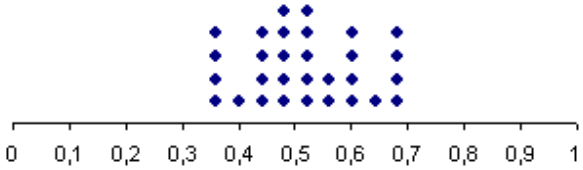
a) Recolha uma amostra de 25 pastilhas e registe o número e a proporção de pastilhas de cada uma das cores:

	Vermelha	Amarela	Azul
Freq. Abs.			
Freq. Rel.			

b) A proporção de pastilhas amarelas, de entre as 25 que seleccionou, é um *parâmetro* ou uma *estatística*?
c) A proporção de pastilhas amarelas fabricadas pela fábrica XPTO Lda é um parâmetro ou uma estatística?
d) Conhece o valor da proporção de pastilhas amarelas fabricadas pela XPTO Lda?
e) Conhece o valor da proporção de pastilhas amarelas existentes nas 25 pastilhas que seleccionou?

Estas questões que acabamos de pôr realçam o facto de facilmente se poder calcular o valor de uma estatística, mas só raramente se conhecer o valor de um parâmetro. É por essa razão, que um dos principais objectivos pelos quais de recolhe uma amostra, é para estimar o valor do parâmetro, com base no valor da estatística.

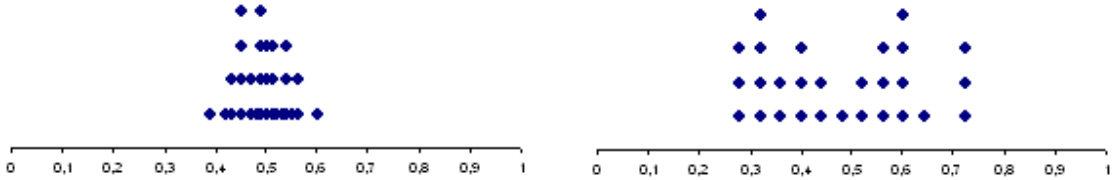
f) Suspeita que todos os seus 29 colegas obtiveram a mesma proporção de pastilhas amarelas, na amostra de 25 pastilhas que cada um recolheu?
g) Admita que no seguinte diagrama de pontos se apresentam as proporções de pastilhas amarelas obtidas por si e pelos seus colegas (também em amostras de dimensão 25):



Sugira um valor para a proporção de pastilhas amarelas produzidas pela fábrica XPTO Lda.

h) Admitindo que o valor da proporção de pastilhas amarelas produzidas pela fábrica era 0.5, algumas das proporções obtidas não estão "razoavelmente" próximas desse valor. Como explica este facto?

i) Suponha que em vez de 25 pastilhas, recolhia 100 pastilhas e os seus colegas faziam o mesmo. Qual dos dois diagramas de pontos seguintes, esperaria obter para representar as proporções de pastilhas amarelas obtidas nas amostras de dimensão 100?



Explique convenientemente a sua resposta.

j) As proporções de pastilhas amarelas obtidas nas 30 amostras recolhidas foram as seguintes:

0,39	0,45	0,49	0,51	0,54
0,42	0,45	0,49	0,51	0,54
0,43	0,47	0,49	0,51	0,55
0,43	0,47	0,5	0,52	0,56
0,45	0,48	0,5	0,53	0,56
0,45	0,49	0,5	0,54	0,6

Calcule a média e o desvio padrão dos valores anteriores.

k) Admita que a população de onde esteve a seleccionar as amostras anteriores é constituída por elementos pertencentes a uma de duas categorias: pastilhas amarelas e pastilhas não amarelas. Se uma pastilha for amarela, representa-a por 1. Caso contrário representa-a por 0. Seja 0,5 a proporção de pastilhas amarelas. Calcule o valor médio o desvio padrão desta população.

l) Compare os valores obtidos nas duas alíneas anteriores. Comente.

2.4.7 – (Rossman, 2001) – Em 1996, nas eleições presidenciais nos Estados Unidos, Bill Clinton recebeu 49% dos votos, enquanto Bob Dole recebeu 41% e Ross Perot 8%. Suponha que selecciona uma amostra de 100 eleitores das eleições referidas anteriormente e pergunta a cada eleitor, em quem votou.

a) Tem a certeza que nos 100 eleitores vai obter 49 a favor de Clinton, 41 a favor de Dole e 8 a favor de Perot?

b) Suponha que selecciona várias amostras de 100 eleitores. Vai obter a mesma proporção de eleitores adeptos de Clinton, em todas as amostras?

c) Suponha que conseguia seleccionar todas as amostras possíveis, de dimensão 100, da população de eleitores e calculava para cada uma delas a proporção de eleitores favoráveis a Clinton. Que nome dá à distribuição dos valores obtidos anteriormente? Qual a média e desvio padrão que esperaria obter para o conjunto de todos esses valores?

d) Admita que selecciona amostras de dimensão n, com n= 50, 100, 200, 400, 500, 800, 1000, 1600, 2000 e calcula a proporção de eleitores favoráveis a Clinton.
   i) Calcule os desvios padrões das distribuições de amostragem da proporção de eleitores que votam Clinton, para cada uma das dimensões de amostras consideradas (Não esqueça que o valor obtido por Clinton, nas eleições referidas, foi de 49%).
   ii) Construa um diagrama de dispersão dos valores dos desvios padrões, versus os valores das dimensões das amostras consideradas.
   iii) Para que o desvio padrão reduza de metade, de quanto é que tem de aumentar a dimensão da amostra?

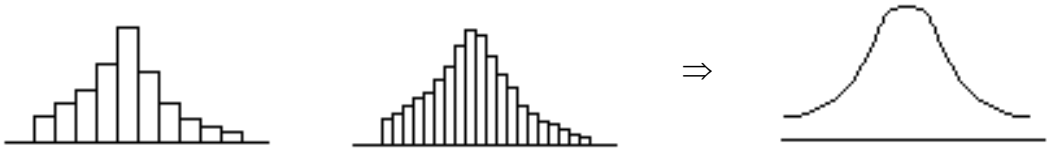
e) Represente por p a proporção de votos recebidos por um certo candidato numa eleição. Admita que selecciona repetidamente amostras de dimensão 100 e calcula a proporção de eleitores que votariam nesse candidato.
   i) Calcule os desvios padrões das distribuições de amostragem das proporções amostrais obtidas, admitindo que o parâmetro p assume cada um dos seguintes valores: 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.
   ii) Construa um diagrama de dispersão dos desvios padrões obtidos, versus o valor do parâmetro ou proporção polpulacional p.
   iii) Para que valor de p obtém maior variabilidade nas proporções amostrais?
   iv) Para que valor de p obtém menor variabilidade nas proporções amostrais? Comente.

5 - O modelo Normal (ou Gaussiano)

Como vimos nas secções anteriores, o modelo Normal é adequado para descrever, em determinadas condições, as distribuições de amostragem das estatísticas Média $\overline{\mathbf{X}}$ e Proporção Amostral $\hat{\mathbf{P}}$, utilizadas para estimar o valor médio μ e a proporção populacional **p**, respectivamente. Assim, está na base de técnicas de inferência estatística largamente utilizadas, nomeadamente as que dizem respeito à estimação intervalar ou construção de intervalos de confiança, como veremos no módulo 3 – Intervalos de confiança.

Independentemente da população que esteja a ser objecto de estudo, vimos nas secções anteriores que, se a dimensão da amostra for suficientemente grande, o Teorema Limite Central dá-nos legitimidade para utilizarmos o modelo Normal, na aproximação da distribuição de amostragem da Média ou da Proporção, sempre que não for conhecida a distribuição exacta.

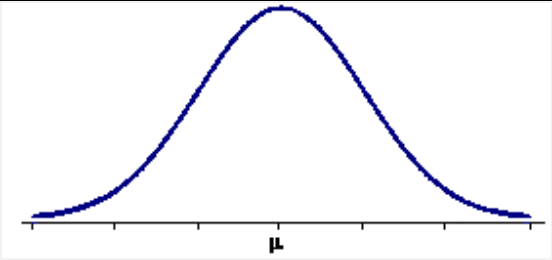
A **distribuição Normal**, das distribuições contínuas, a mais conhecida, foi obtida matematicamente por Gauss, como a distribuição dos erros de medidas, tendo-lhe dado o nome sugestivo de "lei normal dos erros". A partir daí, astrónomos, físicos e mais tarde, cientistas de outros campos, que manipulavam dados, verificaram que muitos dos histogramas que construíam apresentavam a característica seguinte: começavam a crescer gradualmente, até atingirem um ponto máximo, a partir do qual decresciam de forma simétrica:



Este aspecto apresentado pelos histogramas, foi o suficiente para desencadear um entusiasmo pela distribuição (População) Normal, com função densidade em forma de sino, a qual se admitia como subjacente aos dados. Chegou-se ao ponto de duvidar de dados, cujos histogramas não tinham aquele comportamento!

Desfeito o mito da distribuição normal, podemos dizer que ela tem ainda hoje um papel importante em estatística, já que muitos dos processos de inferência estatística clássica, têm por base, precisamente a distribuição **Normal**.

Ao falarmos na distribuição **Normal**, estamos na realidade a referir-nos a uma família de distribuições, indexadas pelos parâmetros μ e σ. Assim, para cada par de valores destes parâmetros temos uma distribuição normal, cuja função densidade de probabilidade tem o seguinte aspecto:



Uma v.a. **X** com distribuição **Normal** de parâmetros μ e σ representa-se por

$$X \cap N(\mu,\sigma)$$

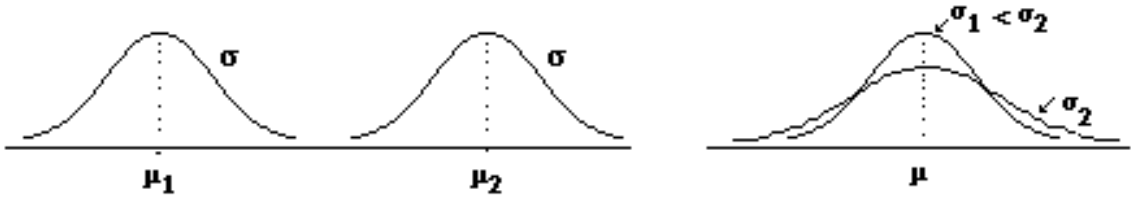
Pode-se mostrar que:

$$E(X) = \mu \qquad e \qquad Var(X) = \sigma^2$$

Vejamos algumas propriedades, relativamente à representação gráfica, da função densidade normal, que se deduzem da sua expressão analítica

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2], \; x \varepsilon R:$$

- é simétrica relativamente ao seu valor médio μ, de modo que duas curvas correspondentes a duas distribuições com o mesmo desvio padrão têm a mesma forma, diferindo unicamente na localização.
- é tanto mais achatada, quanto maior for o valor de σ, de modo que duas curvas correspondentes a duas distribuições com o mesmo valor médio, são simétricas, relativamente ao mesmo ponto, diferindo no grau de achatamento.

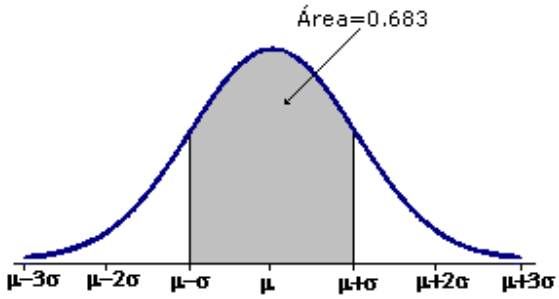


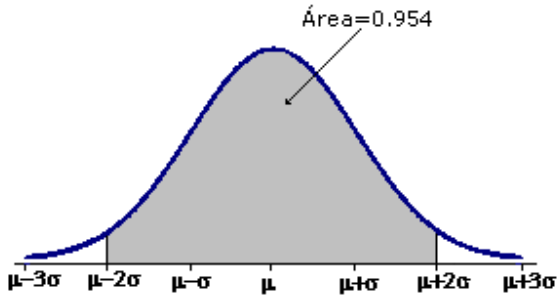
Para dar uma ideia da **concentração** da distribuição normal, em torno do seu valor médio, apresentamos seguidamente algumas probabilidades:

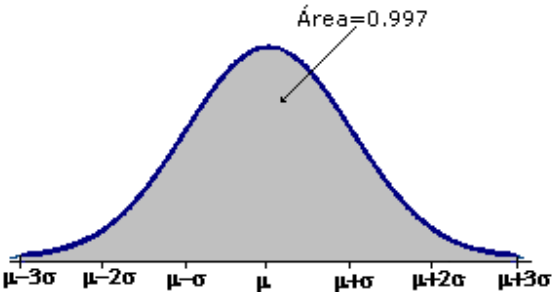
$$P(\mu - \sigma \le X \le \mu + \sigma) = .683$$

$$P(\mu - 2\sigma \le X \le \mu + 2\sigma) = .954$$

$$P(\mu - 3\sigma \le X \le \mu + 3\sigma) = .997$$







À distribuição normal que tem valor médio 0 e desvio padrão 1 chamamos distribuição "*standard* " ou *reduzida*, e representamos por

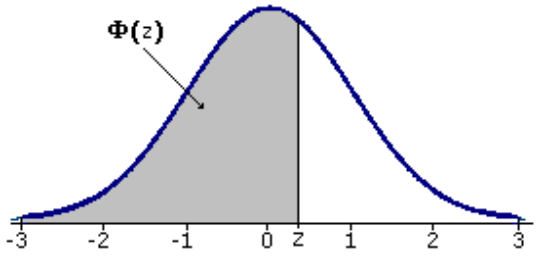
$$\mathbf{Z \cap N(0,1)}$$

Se a v.a. **X** tiver valor médio μ e desvio padrão σ, então a v.a. $\mathbf{Z} = \dfrac{X-\mu}{\sigma}$, tem valor médio 0 e desvio padrão 1. Assim

$$\mathbf{X \cap N(\mu,\sigma)} \Leftrightarrow \quad \mathbf{Z} = \frac{X-\mu}{\sigma} \cap \mathbf{N(0,1)}$$

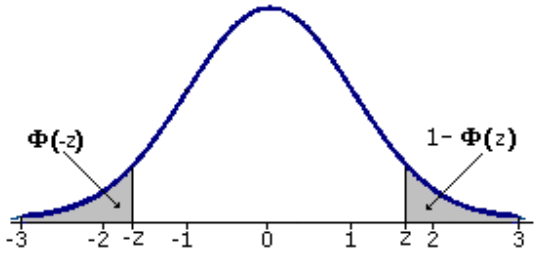
A função distribuição da normal reduzida, tem uma notação especial. Assim, se Z for uma v.a. normal reduzida, representamos

$$P(Z \leq z) = \Phi(z)$$



Da simetria da curva normal, deduz-se imediatamente a seguinte propriedade:

$$\Phi(-z) = 1 - \Phi(z)$$

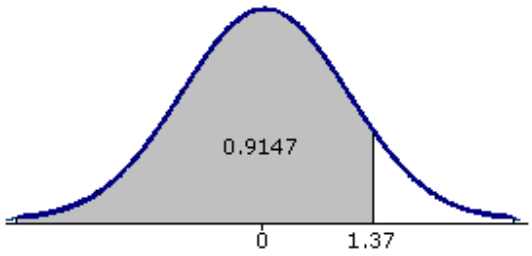


Hoje em dia o cálculo das probabilidades correspondentes à distribuição Normal não oferece qualquer dificuldade, pois pode ser feito com as máquinas de calcular ou a folha de Excel do computador. Até há bem pouco tempo, só dispunhamos de tabelas extensivas da função distribuição da normal standard, que permitiam o cálculo de quaisquer probabilidades, referentes à v.a. Z (veremos mais adiante a utilização do computador para o cálculo das probabilidades da Normal). A propriedade enunciada anteriormente também permite concluir, que bastava haver tabelas para os valores de $z \geq 0$ ou de $z \leq 0$.

Alguns exemplos

$P(Z \leq 1.37) = \Phi(1.37)$

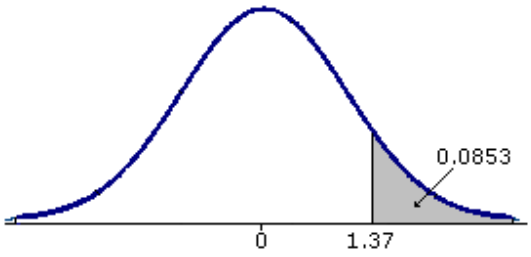
$= .9147$



$P(Z > 1.37) = 1 - P(Z \leq 1.37)$

$= 1 - .9147$

$= .0853$

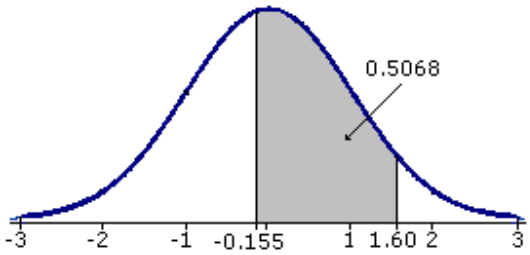


$P(-.155 < Z < 1.60) = \Phi(1.60) - \Phi(-.155)$

$= \Phi(1.60) - 1 + \Phi(.155)$ (a tabela disponível só tinha os valores positivos)

$= .9452 - 1 + .5616$

$= .5068$

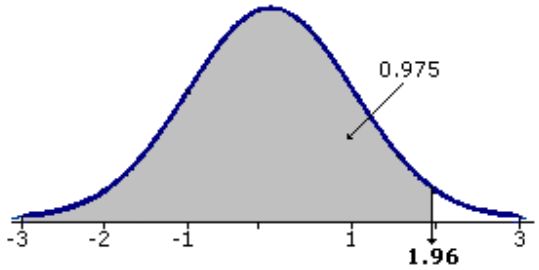


Determinar o valor de z, tal que

$P(Z \leq z) = .975$

$\Phi(z) = .975 \Rightarrow z = \Phi^{-1}(.975)$

$= 1.96$

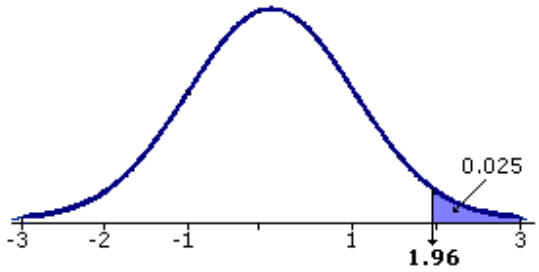


Determinar o valor de z tal que

$P(Z > z) = .025$

$1 - \Phi(z) = .025 \Rightarrow z = \Phi^{-1}(.975)$

$= 1.96$



Mas se a Normal não tiver valor médio nulo e desvio padrão 1, como fazer para ainda ser possível utilizar as tabelas?

Para o cálculo das probabilidades correspondentes a uma distribuição normal de parâmetros **μ** e **σ**, vamo-nos servir das tabelas da normal reduzida, tendo em atenção a seguinte relação, já apresentada anteriormente:

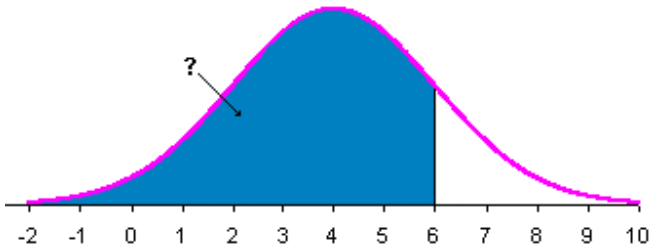
$$\mathbf{X \cap N(\mu, \sigma)} \quad \Leftrightarrow \quad \mathbf{Z = \frac{X - \mu}{\sigma} \cap N(0,1)}$$

donde:

$$P(a \leq X \leq b) = P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) \Leftrightarrow P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

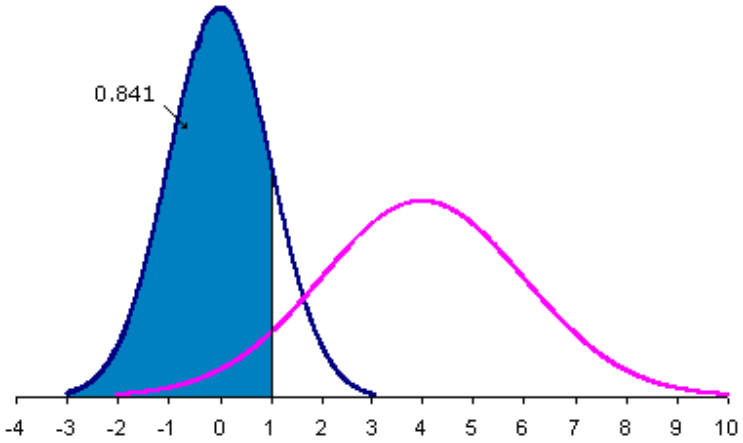
Se X∩N(4,2) calcular P(X≤6)

$P(X \leq 6) = \Phi(\frac{6-4}{2})$



$= \Phi(1)$

$= 0.841$



Se X∩N(μ, σ) calcular

P(μ-1.96σ≤X≤μ+1.96σ)

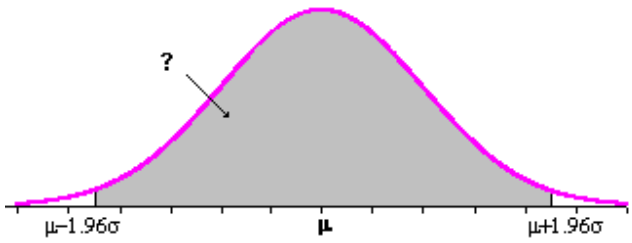
P((μ-1.96σ ≤X≤μ+1.96σ) =

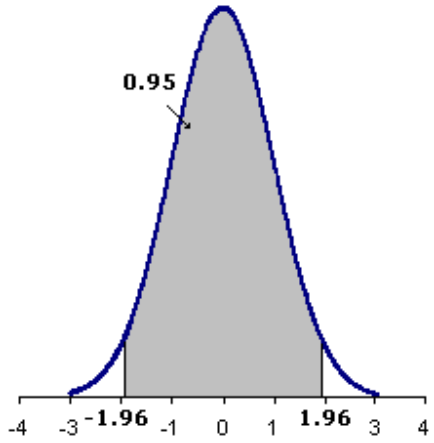
$P(\frac{\mu - 1.96\sigma - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{\mu + 1.96\sigma - \mu}{\sigma})$

$= \Phi(1.96). - \Phi(-1.96)$

$= 0.975 - 0.025$

$= .0.95$





Exemplo - Na pastelaria "Gulosa" a quantidade de farinha F utilizada semanalmente, é uma variável aleatória com distribuição normal de valor médio 600kg e desvio padrão 40kg. Havendo no início de determinada semana, um armazenamento de 634kg e não sendo possível receber mais farinha durante a semana:

a) Determine a probabilidade de ruptura do stock de farinha.

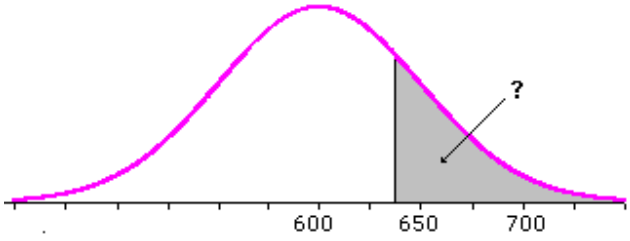
b) Qual deveria ser o stock, de modo que a probabilidade de ruptura fosse de .01?

Resolução:

a) Pretende-se calcular a probabilidade de ruptura do stock, isto é, P(F>634), com F∩N(600,40)

$P(F > 634)=1 - P(F \leq 634)=1 - P(Z \leq \frac{634-600}{40}) = 1 - \Phi(.85)$

$= 1- .8023 =.1977$

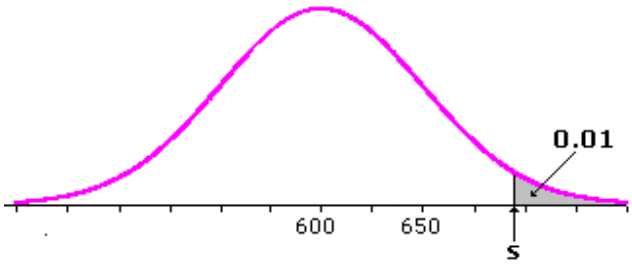


b)

$P(F > s)= .01 \Rightarrow 1-\Phi(\frac{s-600}{40})=.01$

$\Phi(\frac{s-600}{40}) = .99 \Rightarrow \frac{s-600}{40} = 2.326$

$s = 693kg$



Introdução à Inferência estatística – estimação intervalar ou intervalos de confiança

1 – Introdução

Nas secções anteriores estudámos o comportamento da Média e da Proporção amostral, como estimadores, respectivamente, do valor médio de uma População e da proporção com que os elementos da População verificam determinada característica. Verificámos que, quando se consideram amostras diferentes, embora da mesma dimensão, a média ou a proporção variam de amostra para amostra, mas apresentam um **comportamento característico**, de uma distribuição *aproximadamente simétrica*, com *pequena variabilidade*, acentuando-se estas características à medida que a dimensão da amostra aumenta.

Por exemplo, no caso da estimação do valor médio, o facto de a média variar de amostra para amostra, não nos permite saber, recolhida uma amostra, se a média dessa amostra é uma boa estimativa do valor médio da população subjacente à amostra (como temos feito várias vezes, estamos a identificar população com a variável em estudo, cujo valor médio se pretende conhecer), isto é, não podemos atribuir nenhuma "confiança" a essa estimativa do valor médio.

2 – Intervalo de confiança para o valor médio

No estudo da distribuição de amostragem da média, concluímos ainda que quando se faz amostragem sem reposição e as Populações têm dimensão razoavelmente grande, ou no caso de a amostragem ser com reposição, Populações com qualquer dimensão, e as amostras também têm dimensão grande (maior ou igual a 30), a distribuição de amostragem da Média pode ser aproximada pela distribuição Normal (Teorema Limite Central).

Este comportamento da distribuição de amostragem da Média tem consequências muito importantes, no que diz respeito ao problema da estimação do parâmetro valor médio, já que vamos aproveitá-lo para encarar este problema de um outro ângulo. Em vez de procurarmos um valor **– estimativa pontual**, como aproximação do valor do parâmetro desconhecido, vamos procurar obter um intervalo – estimativa intervalar **ou intervalo de confiança**, que com uma determinada confiança contenha o valor do parâmetro.

Vamos então procurar um intervalo aleatório [A, B] que, com uma "grande probabilidade", por exemplo 0.95, contenha o parâmetro μ:

$$P([A, B] \text{ conter } \mu) = 0.95$$

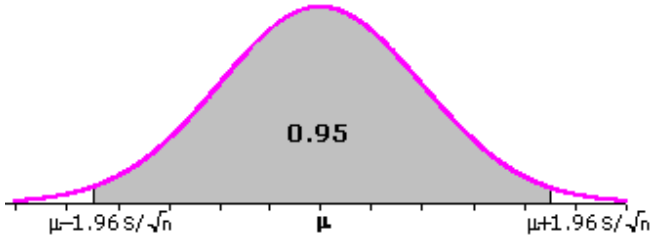
Ora, é precisamente na construção destes intervalos de confiança, que vamos aproveitar o facto de a distribuição de amostragem da Média poder ser aproximada

pelo modelo Normal, com valor médio igual ao valor médio **μ** da População (parâmetro que estamos a estimar) e desvio padrão igual a **σ/√n**, onde **σ** é o desvio padrão da população. Como o desvio padrão da População é quase sempre desconhecido, vamos também estimá-lo a partir do desvio padrão amostral, **s**, pelo que um valor aproximado para o desvio padrão da média, também conhecido como **erro padrão**, é s/√n.

Então, tendo em consideração as propriedades da distribuição Normal, podemos escrever:

$$P(-1.96 \leq \frac{\overline{X} - \mu}{S/\sqrt{n}} \leq 1.96) \approx 0.95 \qquad (1)$$

O valor 1.96 pode ser obtido consultando uma tabela, a calculadora ou a folha de Excel.



De (1) vem

$P(\mu - 1.96\ S/\sqrt{n} \leq \overline{X} \leq \mu + 1.96\ S/\sqrt{n}) \approx 0.95$

ou

$$P(\overline{X} - 1.96\ S/\sqrt{n} \leq \mu \leq \overline{X} + 1.96\ S/\sqrt{n}) \approx 0.95$$

Então o intervalo aleatório que andávamos à procula é

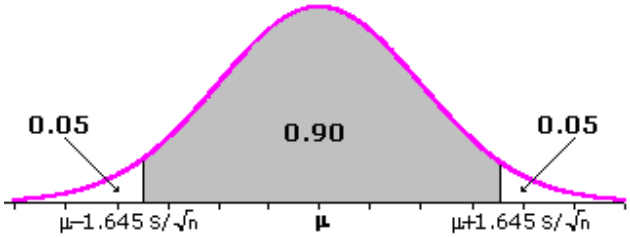
$$\mathbf{[\overline{X} - 1.96 \times S/\sqrt{n},\ \overline{X} + 1.96 \times S/\sqrt{n}]}$$

Repare-se que o intervalo anterior é aleatório, já que o valor da média e do desvio padrão variam, dependendo da amostra que se recolher. Se recolhermos duas amostras diferentes, ambas da mesma dimensão, vamos obter valores diferentes para a média e para o desvio padrão. Dizemos que este intervalo é um **intervalo de confiança**, com uma **confiança ou um nível de confiança de 95%.**

Afinal, o que significa um intervalo de 95 % de confiança?

Significa que se recolhermos muitas amostras de dimensão n, calcularmos as médias e os desvios padrões dessas amostras e construirmos os intervalos de confiança respectivos, utilizando a expressão anterior, cerca de 95% desses intervalos conterão o valor médio μ, enquanto que os restantes 5% não conterão o parâmetro μ. Não temos a certeza que um dado intervalo, em particular, contenha o parâmetro desconhecido, mas estamos confiantes que assim aconteça, isto é estamos **95% confiantes** que o intervalo que calculámos a partir da amostra seleccionada (na prática só seleccionamos uma amostra), contenha o valor do parâmetro.

Se na expressão (1) da probabilidade, mudarmos a probabilidade de 0.95 para 0.90, por exemplo, então em vez de 1.96, devemos considerar 1.645:



Assim, um intervalo de confiança, com 90% de confiança terá o seguinte aspecto

$$[\bar{X} - 1.645 \times S/\sqrt{n},\ \bar{X} + 1.645 \times S/\sqrt{n}]$$

A forma geral do intervalo de confiança será,

$$[\bar{X} - \mathbf{z} \times S/\sqrt{n},\ \bar{X} + \mathbf{z} \times S/\sqrt{n}]$$

onde o valor de **z** dependerá da confiança com que se pretende construir o intervalo.

Alguns valores (obtidos a partir da distribuição da Normal(0,1)), incluindo os já considerados anteriormente, são:

Confiança	**z**
90%	1.645
95%	1.960
97.5%	2.326
99%	2.576
99.5%	3,090
99.9%	3.291
99.95%	3.891
99.995%	4.417

Como se verifica a partir da tabela anterior, quanto *maior for a confiança*, maior é o valor de **z**, pelo que *maior será a amplitude* do intervalo.

Como diminuir a amplitude de um intervalo de confiança?

De um modo geral pretende-se construir um intervalo com **pequena amplitude**, pois nos dá uma **maior precisão**. Como se depreende da forma desse intervalo, para diminuir a sua amplitude, que é dada por $2 \times z \times \frac{s}{\sqrt{n}}$ podemos fazê-lo de duas maneiras:

- ou diminuir a confiança (o que faz com que diminua o valor de z), o que não é aconselhável;
- ou aumentar a dimensão da amostra considerada para calcular o intervalo. Por exemplo, se aumentar 4 vezes a dimensão da amostra, a amplitude do intervalo reduz-se a metade.

Nas considerações anteriores estamos a admitir que a dimensão da amostra inicial já é suficientemente grande, de modo que a estimativa **s** para o desvio padrão da população não se altera significativamente quando utilizamos mais informação (uma amostra de maior dimensão) para a calcular.

Como casos extremos de intervalos de confiança, temos:

- o intervalo de confiança, com uma **confiança 0%,** que se reduz a um ponto, que não é mais do que a estimativa pontual do valor médio, ou seja a média calculada a partir da amostra considerada;
- e temos ainda o intervalo com uma **confiança de 100%**, que é a recta real (porque vem o valor de z igual a infinito).

Obviamente que nenhum destes intervalos é de grande utilidade!

Margem de erro

A metade da amplitude de um intervalo de confiança, é costume chamar *margem de erro*.

Exemplo – Considerando a população dos deputados da X Legislatura, suponhamos que estávamos interessados em estimar o parâmetro idade média da população. Seleccionou-se uma amostra aleatória (com reposição) de dimensão 30 e registaram-se as idades dos elementos seleccionados. Os valores obtidos apresentam-se na seguinte tabela:

46	50	47
34	70	55
54	36	30
48	39	52
40	55	54
41	56	52
40	60	42
49	53	44
54	32	48
71	50	31

A média e o desvio padrão das idades anteriores são, respectivamente, 47.8 e 10.2 anos. Então, um intervalo de 95% de confiança para a idade média da população é

$$[47.8 - 1.96 \times 10.2/\sqrt{30}\,,\ 47.8 + 1.96 \times 10.2/\sqrt{30}\,]$$

ou seja [44.2, 51.4], é um intervalo com uma confiança de 95%. Repare-se que o intervalo anterior contém o parâmetro em estudo (esta é uma situação de excepção, em que a população é tão pequena, que facilmente se obtém o valor do parâmetro valor médio da Idade).

Chamamos a atenção para que, se não conhecêssemos o valor do parâmetro em estudo, não poderíamos garantir que o intervalo que calculámos anteriormente o contivesse. Apenas estamos **confiantes** em que isso acontecesse, pois se calculássemos 100 amostras de dimensão 30, como a anterior, esperávamos que cerca de 95 dos intervalos que se poderiam construir com as médias e desvios padrões dessas amostras, contivessem o parâmetro em estudo.

Exemplo – Considere a população constituída pelos empregados da empresa X, em Anexo. Suponha que estamos interessados em estudar o parâmetro **altura média**.
a) Seleccione uma amostra de dimensão 30 e calcule um intervalo de 95% de confiança para o parâmetro em estudo.
b) Seleccione mais 99 amostras de dimensão 30, e a partir de cada uma delas construa um intervalo de 95% de confiança.
c) Quantos dos intervalos considerados anteriormente contêm o valor do parâmetro? Comente.

Resolução:

a) A seguir apresenta-se a amostra seleccionada pelo processo de amostragem com reposição

153	160	163	173	174	173	169	154	164	165	156	169	159	157	173
154	177	160	170	161	161	174	165	170	165	157	158	160	160	170

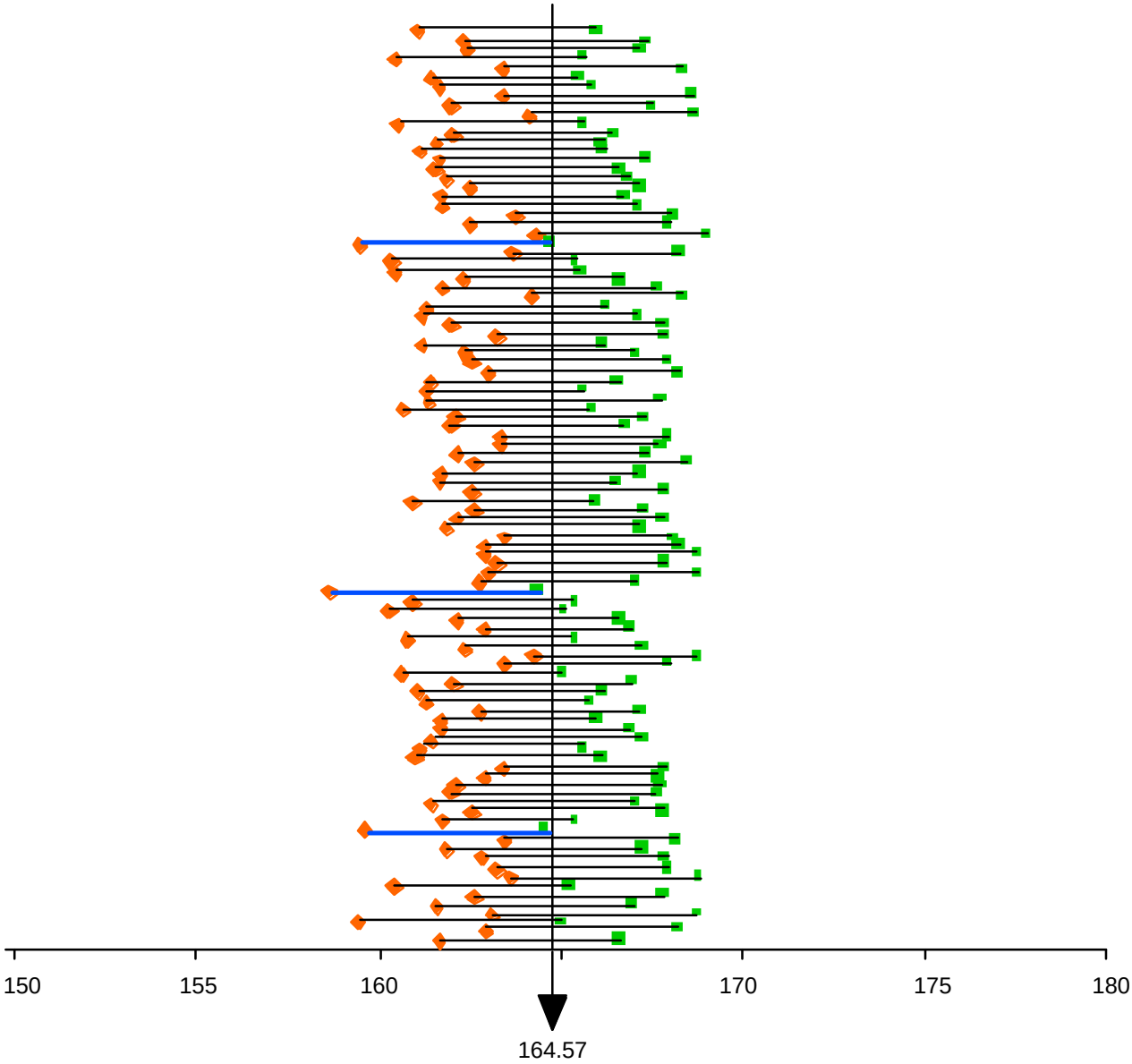
Média = 164.13

Desvio padrão = 6.95

Intervalo de 95% de confiança $[164.13 - 1.96 \times \frac{6.95}{\sqrt{30}}, 164.13 + 1.96 \times \frac{6.95}{\sqrt{30}}]$

$$[161.65, 166.62]$$

b) Seleccionámos mais 99 amostras aleatórias, para as quais calculámos a média e o desvio padrão e os respectivos intervalos de confiança, pelo mesmo processo que o calculado na alínea anterior. Apresentamos esses intervalos graficamente, na figura seguinte:



Na figura anterior, a seta indica a posição do valor do parâmetro a estimar, ou seja a altura média da população.

c) Verificamos que três dos intervalos construídos não contêm o valor do parâmetro (Esperarávamos encontrar um valor próximo de 5).

3 – Intervalo de confiança para a proporção

No estudo da distribuição de amostragem da proporção, concluímos que quando se faz amostragem sem reposição e as Populações têm dimensão razoavelmente grande, ou no caso de a amostragem ser com reposição, Populações com qualquer dimensão, e as amostras também têm dimensão grande (maior ou igual a 30), a distribuição de amostragem da Proporção amostral pode ser aproximada pela distribuição Normal (Teorema Limite Central). Este comportamento da distribuição de amostragem da Proporção, tal como vimos anteriormente para a Média, tem consequências muito importantes, no que diz respeito ao problema da estimação do parâmetro proporção populacional, já que vamos aproveitá-lo para encarar este problema de um outro ângulo. Em vez de procurarmos um valor **– estimativa pontual**, como aproximação do valor do parâmetro desconhecido, vamos procurar obter um intervalo – estimativa intervalar **ou intervalo de confiança**, que com uma determinada confiança contenha o valor do parâmetro.

Representando por $\hat{p}$ a proporção amostral, estimador do parâmetro **p**, sabemos do módulo 2 – Introdução à Estimação que, se a recolha da amostra for feita com reposição de uma População de dimensão qualquer, ou sem reposição de uma população de grande dimensão, e se a dimensão, n, da amostra for grande, então

$$P\left(\frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}} \le z\right) \approx \Phi(z)$$

Por um processo perfeitamente idêntico ao considerado para obter o intervalo de confiança para o valor médio, em que no caso em que a variância σ^2 da população, é desconhecida, a substituímos pela variância amostral, também aqui, substituímos na variância da população p(1-p), o p por $\hat{p}$. Temos assim o intervalo de confiança para a proporção

$$\left[\hat{p} - \mathbf{z}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + \mathbf{z}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Fazendo a analogia com o que se passa com o intervalo de confiança para o valor médio, no intervalo anterior, a proporção $\hat{p}$, substituiu a média, e considerou-se $\hat{p}(1-\hat{p})$ como estimador da variância populacional p(1-p).

Observação: Ao contrário do que é usual, em que se considera a variável aleatória com letra maiúscula e um seu valor observado com letra minúscula, no caso da proporção não é costume fazer essa distinção. Assim, representa-se indiferentemente por $\hat{p}$ tanto a variável aleatória como um seu valor observado, dependendo do contexto em que está a ser utilizado a sua interpretação como variável aleatória ou valor dessa variável aleatória.

O valor de **z** depende da confiança com que se quer construir o intervalo, como vimos para o caso do valor médio.

No caso particular de um intervalo de 95% de confiança, temos

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

Em que condições é que se pode utilizar o intervalo anterior?

Dissemos anteriormente que era necessário que a dimensão da amostra fosse suficientemente grande. No entanto, também já vimos que quanto maior for a variabilidade presente na população de onde se recolhe a amostra, maior terá de ser a dimensão dessa amostra. Uma regra empírica aconselha-nos a considerar

$$n\hat{p} \geq 10 \text{ e } n(1-\hat{p}) \geq 10$$

O intervalo anterior, obtido a partir de uma amostra de dimensão n, tem amplitude igual a

$$2 \times 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Como já referimos para o intervalo de confiança para o valor médio, a metade da amplitude do intervalo, ou seja à quantidade $1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, chamamos *margem de erro* da sondagem.

Exemplo – Suponha que para a população dos empregados da empresa X se pretende estimar a proporção de mulheres casadas. Seleccione uma amostra de dimensão 30 e obtenha uma estimativa pontual e uma estimativa intervalar ou intervalo de 95% de confiança para essa proporção.

Resolução:

Para facilitar o estudo, utilizámos uma folha de Excel e convertemos as categorias da característica populacional estado civil, da seguinte forma:

Casada	1	Solteira	0
Casado	0	Solteiro	0
Divorciada	0	Divorciado	0

Obtivemos uma população de 0's e 1's, em que um elemento da população assumia o valor 1 sempre que a característica em estudo se verificava.

Seleccionada uma amostra aleatória, com reposição, de dimensão 30, obtivemos os seguintes valores:

0 1 0 0 1 0 0 1 0 0 0 1
0 0 0 1 0 0 0 1 0 0 0 1
0 0 0 0 0 0

Proporção de mulheres casadas na amostra= 7/30 = 0.233

Este valor é *uma estimativa pontual* da proporção p de mulheres casadas na população, cujo valor é 0.2268 (Mais uma vez estamos numa situação em que foi fácil calcular o valor do parâmetro, atendendo a que a população tinha uma dimensão muito pequena)..

Um intervalo de 95% de confiança para a proporção de mulheres casadas na população é

$$[0.233 - 1.96 \times \sqrt{\frac{0.233\,(1-0.233)}{30}}\,,\ 0.233 + 1.96 \times \sqrt{\frac{0.233\,(1-0.233)}{30}}\,],$$

ou seja [0.082, 0.384]. O intervalo anterior tem uma margem de erro de 0.151.

Exemplo – O Diário de Notícias na sua edição do dia 25 de Fevereiro de 1998, referia, relativamente a uma sondagem realizada em colaboração com a TSF/Universidade Moderna, que a confiança dos portugueses é "açambarcada" pelos docentes, logo seguidos pelos médicos, adiantando ainda que os políticos e sindicalistas partilham os últimos lugares. Apresentamos de seguida um excerto desse artigo.

Barómetro de profissões

Em relação às seguintes profissões com importância na vida nacional, diga se tem ou não confiança na sua acção

Profissão	Não tem confiança	Tem confiança	Não sabe/Não responde
Professores	9.1%	85.1%	5.8%
Médicos	9.7%	85.0%	5.3%
Juízes	26.9%	64.7%	8.4%
Militares	30.9%	61.6%	7.4%
Jornalistas	30.8%	58.2%	11.1%
Padres	31.8%	55.2%	13.0%
Empresários	33.5%	53.6%	13.0%
Polícias	38.6%	51.8%	9.6%
Sindicalistas	49.6%	39.4%	11.0%
Políticos	57.5%	30.9%	11.7%

Ficha técnica: Esta sondagem foi encomendada pelo DN e pela TSF ao Centro de Sondagens da Universidade Moderna. O trabalho de campo decorreu entre os dias 12 e 17 de Fevereiro de 1998. Os inquéritos foram realizados em 25 freguesias de Portugal continental. A amostra foi seleccionada aleatoriamente e, para cada uma das freguesias, foi feito um estudo demográfico com base nos dados do Censos 91 e do Stape. Foram validados 1303 inquéritos. O processo de informação utilizado foi a recolha directa (porta a porta), através de um inquérito estruturado onde estava anexado o boletim com o nome de várias profissões com importância na vida nacional. Foi feita uma primeira validação pelos monitores, no local, a 10 por cento dos inquéritos e, posteriormente, foram validados outros 20 por cento telefonicamente. O

erro máximo, para um nível de confiança de 95%, é de 2.71. A análise dos resultados é da responsabilidade do DN.

a) O valor de 85.1% apresentado para os Professores é uma estatística ou um parâmetro?

b) Construa um intervalo de 95% de confiança para a percentagem da população que tem confiança nos Professores.

c) O intervalo anterior contém necessariamente a percentagem de indivíduos da população que têm confiança nos Professores?

d) Calcule a margem de erro dos intervalos de confiança, para a confiança associada às diferentes profissões consideradas.

e) Qual o valor máximo obtido para as margens de erro obtidas na alínea anterior? Isso estará de acordo com o que vem especificado na ficha técnica?

Resposta alínea d)

85.1%	0.0193
85.0%	0.0194
64.7%	0.0259
61.6%	0.0264
58.2%	0.0268
55.2%	0.0270
53.6%	0.0271
51.8%	0.0271
39.4%	0.0265
30.9%	0.0251

Reparou que: a margem de erro é máxima para um valor da proporção próximo de 0.5? Efectivamente se se fizer o estudo da função $f(\hat{p}) = 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, verifica-se que ela assume um valor máximo para $\hat{p} = 0.5$.

Qual a dimensão da amostra que se deve recolher para obter um intervalo com um nível de confiança de 95% e com uma determinada precisão?

Pretende-se que a margem de erro do intervalo de confiança seja menor ou igual que um valor d. Então, temos de resolver a seguinte desigualdade, em relação a n:

$$1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq d$$

de onde vem

$$n \geq \left(\frac{1.96}{d}\right)^2 \hat{p}(1-\hat{p}) \qquad (1)$$

Como, de um modo geral, não se conhece o valor da proporção amostral, antes de recolher a amostra, considera-se o valor máximo para a expressão $\hat{p}(1-\hat{p})$, que se obtém quando o valor da proporção é 0.5. Vem então

$$n \geq \left(\frac{1.96}{2d}\right)^2 \qquad (2)$$

No caso do exemplo anterior, a margem de erro relativamente aos portugueses que confiam nos professores, é de 1.93%. Se se pretendesse uma margem de erro não superior a 1.5%, teríamos de considerar uma amostra de dimensão **2165**, se entrarmos em consideração com o valor da estimativa 0.851, na fórmula (1).

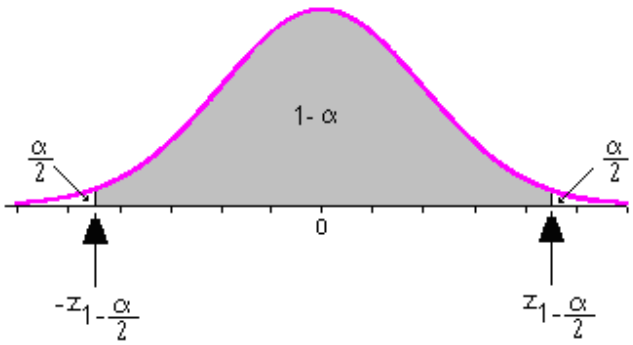
Se desconhecêssemos uma estimativa para a proporção, entraríamos com a fórmula (2) para calcular o valor da dimensão da amostra e obteríamos como valor necessário para a dimensão da amostra **4269**!

Qual a dimensão da amostra que se deve recolher para obter um intervalo com uma determinada precisão, com um nível de confiança de 100(1-α)%?

A confiança de um intervalo costuma exprimir-se na forma anterior, onde α é uma probabilidade relativamente pequena. Assim, se α=5%, temos um intervalo de 95% de confiança. Se representarmos a função densidade da Normal(0,1), o valor de z que aparece no intervalo de confiança genérico

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

não é mais do que o quantil de probabilidade 1-α/2, como se apresenta a seguir:



pelo que se se pretende um intervalo de 100(1-α)% de confiança, com uma precisão não inferior a **d**, a expressão (2) anterior, para a dimensão **n** da amostra necessária, toma a forma:

$$n \geq \left(\frac{z_{1-\alpha/2}}{2d}\right)^2$$

Ainda sobre intervalos de confiança

A interpretação do que é um intervalo de confiança nem sempre é feita correctamente. Vamos aproveitar o seguinte diálogo para ficarmos com as ideias um pouco mais claras sobre este assunto.

Suponhamos que um candidato à Câmara de Lisboa, o Dr. Gentil Alves, pretendia saber qual a percentagem **p**, de eleitores (lisboetas) que pensavam votar nele. Encomendou um estudo à empresa Sondagem, tendo esta questionado 785 lisboetas, escolhidos aleatoriamente, e verificado que a percentagem destes eleitores que pensavam votar no candidato era 56%. Se este valor dava um certo alento ao Dr. Gentil Alves para se candidatar, não o deixava, no entanto, descansado! Ele sabia que se fosse recolhida outra amostra, embora da mesma dimensão, quase de certeza obteria outro valor como estimativa de **p** e quem é que lhe garantia que não era um valor inferior a 50%, o que o deixaria infelicíssimo! Como interpretar este valor de 56%? O Prof. Amável, um amigo estatístico do Dr. Gentil Alves, ajudou-o nesta tarefa. Relatamos a seguir a conversa que se passou entre ambos.

Dr. Gentil Alves – Bom dia Amável, estás bem? Olha, ando um pouco preocupado com esta questão da candidatura à Câmara de Lisboa. Numa sondagem realizada ontem, deram-me uma percentagem de 56% de eleitores a votarem em mim. Mas com que confiança é que eu posso interpretar este resultado? Posso estar seguro que tenho a maioria?

Prof. Amável – Para te ser franco, a confiança que podes ter nesse resultado é nula! Tu próprio sabes que se tivessem sido outros eleitores escolhidos para a sondagem, quase certamente não obterias 56%. Mas não fiques muito preocupado, pois eu vou adiantar-te mais alguma coisa. O valor de 56% vai-me servir para obter um intervalo de 95% de confiança. Deixa-me fazer aqui umas contas que já te telefono.

Dr. Gentil Alves – Está bem. Muito obrigado.

Prof. Amável – Cá estou eu novamente. Com esse valor que me adiantaste construí o intervalo (52.5%; 59.5%), que é um intervalo de 95% de confiança para a percentagem de lisboetas que pensam votar em ti. Estás contente?

Dr. Gentil Alves – Significa isso que existe uma probabilidade de 95% desse intervalo conter essa percentagem (**p**) de eleitores que pensam votar em mim?

Prof. Amável – Nada disso!

Dr. Gentil Alves – Então 95% é a probabilidade de **p** estar contido no intervalo?

Prof. Amável – Que horror! Porventura o **p** é uma variável aleatória? Nem o p nem o intervalo que eu te dei. Assim não podemos falar na probabilidade do **p** estar contido no intervalo, nem do intervalo conter o **p**! Os 95% de confiança significam o seguinte: o processo que se utiliza para calcular os intervalos, como o que te apresentei, é um processo tal que se o utilizasse com todas as amostras possíveis (da mesma dimensão) que posso seleccionar da população, cerca de 95% das vezes produziria intervalos que contêm o p e cerca de 5% das vezes intervalos que não o contêm. No que diz respeito a um intervalo particular, como o que te dei, ficaremos sempre na

dúvida se é um dos que contém p ou não! Temos "fé" que sim, pois já era preciso ter "azar" irmos obter um dos poucos intervalos que não contêm p.

Dr. Gentil Alves – Muito bem. Compreendi o que disseste, mas então porque é que não construo intervalos com, por exemplo, 99% de confiança? Assim, só 1% dos intervalos possíveis de construir é que não conteriam o p, não é verdade?

Prof. Amável – Muito bem observado! Mas nunca ouviste dizer que "sem ovos não se fazem omeletes" ou "que não há almoços grátis"? Pois é! A contrapartida para, com a mesma dimensão da amostra, termos intervalos de 99% de confiança, é que a margem de erro vem maior, isto é, vamos ter intervalos com maior amplitude, o que significa uma menor precisão. Em último caso construiríamos intervalos com uma confiança de 100%! Sabes ao que chegávamos? A R! Não tens nenhuma dúvida de que o intervalo está em R, pois não? Não nos adianta é nada! Já agora, com o valor de 56% obtido na amostra que a Sondagem recolheu, um intervalo de 99% de confiança seria (51.4%; 60.6%). Assim, enquanto que com o primeiro intervalo temos uma margem de erro de 3.5%, agora a margem de erro passou para 4.5%. Ficaste esclarecido?

Dr. Gentil Alves – Penso que sim. Só mais uma questão. Haveria algum processo de, com a confiança de 99%, obter um intervalo com a margem de erro que obtive para o intervalo de 95% de confiança?

Prof. Amável – Mais uma vez estás a colocar uma questão interessante. Efectivamente, podemos, mantendo a confiança, diminuir a margem de erro, agora à custa de recolhermos uma amostra de maior dimensão. Nada se faz sem custos, como estás a ver. Por exemplo, admitindo que a percentagem de lisboetas, que pensam votar em ti, não se alteraria muito se se recolhesse uma amostra de maior dimensão, então teria de ser recolhida uma amostra de 1335 lisboetas, em vez de 785 (estou a considerar que a proprção de votos a teu favor, obtida ao questionar os 1335 lisboetas, é aproximadamente igual a 56%).

Dr. Gentil Alves – Muito obrigada por estes esclarecimentos. Vou mesmo avançar com a minha candidatura.

Passados 8 dias realizaram-se as eleições. O Dr. Gentil Alves é o novo presidente da Câmara de Lisboa.

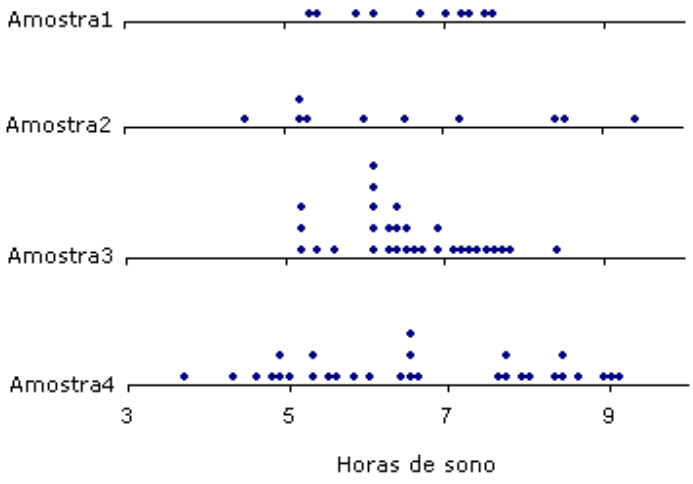
Exercícios

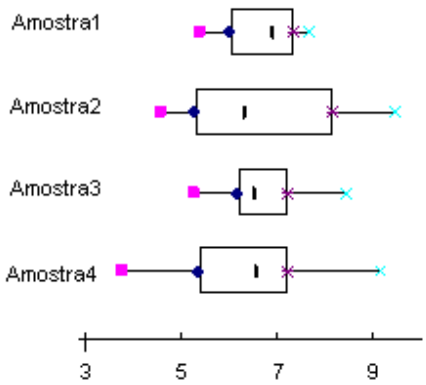
3.1 – Para cada uma das seguintes situações diga se o parâmetro de interesse é um valor médio ou uma proporção:

a) Numa sondagem perguntou-se a cada um de 100 estudantes quantas horas por semana, gastavam a ver televisão.
b) Numa sondagem perguntou-se a cada um de 100 estudantes se passavam mais de 8 horas por semana, a ver televisão.
c) Numa sondagem, perguntou-se a 100 agregados familiares qual a percentagem do orçamento familiar que era gasto com a renda da casa.
d) Num estudo sobre o consume de bebidas alcoólicas, investigou-se junto de 50 restaurantes, qual a percentagem de bebidas alcoólicas, de entre as bebidas consumidas por semana.
e) Junto dos mesmos restaurantes da alínea anterior, verificou-se que 35% dos restaurantes vendiam semanalmente mais bebidas alcoólicas, que não alcoólicas.

3.2 – Num Censo, em que a dimensão da amostra é igual à dimensão da população, o erro padrão da média (ou da proporção amostral) é igual a zero. Explique porquê.

3.3 – Suponha (Adaptado de Rossman, 2001) que pretende conhecer o tempo médio de sono que os alunos da sua escola dormiram, na última noite. Considere os seguintes diagramas que apresentam os tempos de sono de alunos da escola, referentes a 4 amostras:





a) As seguintes estatísticas descritivas foram calculadas com base nas amostras anteriores. Complete a tabela.

Amostra nº	Dimensão da amostra	Média	Desvio padrão amostral
	30	6.6	0.82
	10	6.6	0.82
	10	6.6	1.59
	30	6.6	1.59

b) O que é que todas as amostras têm em comum?
c) Qual a característica que sobressai quando comparamos as distribuições correspondentes às amostras 1 e 2?
d) Qual a característica que sobressai quando comparamos as distribuições correspondentes às amostras 1 e 3?

Na seguinte tabela apresentamos os intervalos de confiança para as amostras de dimensão 30 (consegue obter os intervalos de confiança correspondentes às amostras de dimensão 10?):

Amostra nº	Dimensão da amostra	Média	Desvio padrão amostral	Int. confiança
	30	6.6	0.82	(6.31; 6.89)
	10	6.6	0.82	-
	10	6.6	1.59	-
	30	6.6	1.59	(6.03; 7.17)

Qual das 2 amostras produz uma estimativa para o tempo médio de sono, mais precisa? Qual a influência da variabilidade apresentada pela amostra, para a amplitude do intervalo de confiança?

3.4 – Suponha que na sua escola, cada um dos 50 alunos de Matemática para as Ciências Sociais, foi encarregue de recolher informação junto de 10 adultos, se eram a favor do referendo da Constituição Europeia. O histograma construído com as 50 proporções obtidas terá um aspecto que faz lembrar o modelo normal? Justifique.

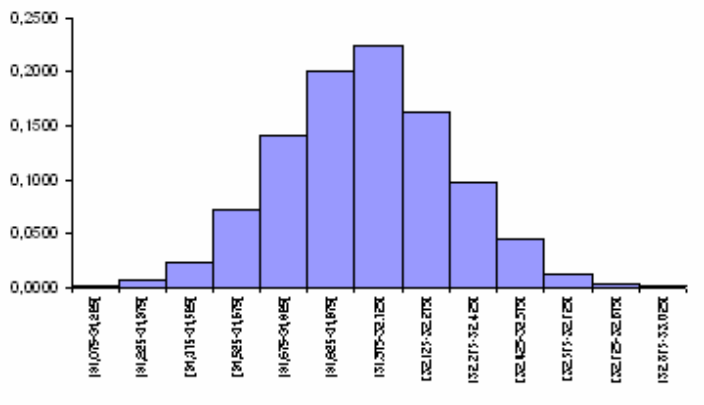
3.5 – Suponha que na sua escola, cada um dos 45 alunos de Matemática A. Foi encarregue de recolher informação, junto de 30 alunos de outras escolas, se eram a favor do "Novo Estatuto para o Aluno". O histograma construído com as 45 proporções obtidas terá um aspecto que faz lembrar o modelo normal? Justifique.

3.6 – Na correcção de certo tipo de exames, feitos a nível nacional, em que cada exame é constituído por uma parte fechada e uma parte aberta, utiliza-se um leitor óptico para corrigir a parte fechada. Cada exame tem 50 questões, e a probabilidade de a máquina ler erradamente uma destas questões é p, a qual é constante de questão para questão e de exame para exame. Desconhece-se este valor de p.

a) Admitindo que em 10 destes exames, a máquina leu erradamente 15 questões, obtenha uma estimativa pontual para p.
b) Utilizando o resultado da alínea anterior:
   i) Obtenha um intervalo, com uma confiança de 95%, para p;
   ii) Qual a margem de erro do intervalo que obteve?
c) A empresa que vende as máquinas de leitura óptica diz que a percentagem de erros que a máquina comete, anda à volta de 1%. Tendo em conta o intervalo de confiança obtido na alínea anterior, pensa que a empresa tem razão no que afirma? Justifique a sua resposta. (Se na alínea anterior não conseguiu determinar o intervalo de confiança pretendido, admita o seguinte intervalo (1.5%; 4.5%)).

3.7 - Uma fábrica de calçado para adultos, pretende começar a produzir sapatos para criança. Encarregou uma empresa de sondagens, de lhe fazer um estudo sobre qual seria o tamanho médio (em cm) do pé de crianças de determinada classe etária. Mesmo antes da empresa apresentar as conclusões, o dono da fábrica (que há muitos anos tinha tido uma disciplina de Estatística) teve acesso à seguinte tabela de frequências e correspondente histograma, dos valores calculados para as médias de 500 amostras, de dimensão 30, recolhidas pela empresa:

Classes	**Freq.rel.**
[31,075-31,225[	0,0020
[31,225-31,375[	0,0075
[31,375-31,525[	0,0250
[31,525-31,675[	0,0735
[31,675-31,825[	0,1410
[31,825-31,975[	0,2005
[31,975-32,125[	0,2250
[32,125-32,275[	0,1635
[32,275-32,425[	0,0990
[32,425-32,575[	0,0445
[32,575-32,725[	0,0130
[32,725-32,875[	0,0040
[32,875-33,025[	0,001 5



Então, na posse destes elementos, pediu ao filho, que tinha frequentado a disciplina de MACS do 11º ano, que lhe respondesse às seguintes questões:

a) Este histograma pretende representar a distribuição de amostragem, aproximada, de uma certa variável. Que variável?

b) Utilizando a tabela anterior, obtenha um valor aproximado para o valor médio da distribuição de amostragem da Média, para amostras de dimensão 30 (considere o valor aproximado às unidades).

c) Tendo em consideração que a estatística Média $\overline{X}$, é um estimador centrado do valor médio da população X, de onde se retiram as amostras, sugira um valor para o valor médio μ, da população X, constituída pelo tamanho do pé, das crianças da classe etária considerada.

d) Sabendo que o desvio padrão de $\overline{X}$, é igual a $\frac{\sigma}{\sqrt{30}}$, onde σ é o desvio padrão da população X, utilize a tabela dada para sugerir um valor para este desvio padrão σ.

e) Como o histograma anterior sugere, e o Teorema Limite Central justifica, a distribuição de amostragem da Média pode ser aproximada por uma distribuição Normal (para amostras de dimensão n, suficientemente grande, ou seja, $n \geq 30$). Admitindo que um dos valores obtidos para a média de uma das 500 amostras de dimensão 30 consideradas, foi 32.125, obtenha um intervalo de 95% de confiança para o valor médio do comprimento do pé. (Se na alínea d) não conseguiu determinar o valor de σ, admita que é igual a 1.5).

f) Admitindo que a população X tem distribuição normal, com o valor médio e desvio padrão obtidos, respectivamente, nas alíneas c) e e), calcule a probabilidade de uma criança, escolhida ao acaso, da classe etária em estudo, ter um comprimento do pé superior a 32.5 cm. (Se não resolveu as alíneas c) e e) considere os valores 32 cm e 1.5 cm, respectivamente para valor médio e desvio padrão de X.

3.8 – Nas últimas eleições legislativas, passada uma hora do fecho das mesas de voto, apareceram os resultados para o concelho de Sintra, dando uma percentagem de votos para JS e FS, respectivamente de 39% e 42%, com uma margem de erro de 3.5% e uma confiança de 95%.

a) O locutor afirmou, ao apresentar aqueles resultados, que os candidatos estavam empatados tecnicamente. Explique, por palavras suas, o que quereria o locutor dizer.
b) Passadas duas horas a margem de erro, diminuiu para 2.5%. Admitindo que a confiança era a mesma, dê uma explicação para a diminuição da margem de erro.
c) Numa sondagem realizada antes das eleições, JS tinha encomendado uma sondagem, que lhe dava a vitória, quando afinal veio a perder as eleições. Teremos que deixar de acreditar nas sondagens?

3.9 – Uma sondagem da TSF/DN publicada na edição do DN de 2 de Julho de 2004, dizia:

Portugueses querem referendo

Maioria mostra-se favorável à eleição de um presidente e de um governo da União Europeia. E também quer exército comum

Os portugueses manifestam tendência para o federalismo europeu: a maioria defende um presidente e um governo europeus, eleitos pelos cidadãos. São igualmente favoráveis à criação de um exército da União Europeia (UE). E, na análise que fazem sobre o futuro comunitário, dizem ainda que querem referendar a próxima reforma institucional da UE. A maioria já ouviu falar do Tratado de Nice, mas está longe de saber o que ele contempla. Talvez por isso, a larga maioria não sabe se o documento deve ou não ser aprovado pelos deputados.

O Barómetro de Junho do DN/TSF/Marktest não incluiu qualquer pergunta directa sobre o federalismo europeu, mas os portugueses acabaram por pronunciar-se nesse sentido. Senão vejamos: 62 por cento dos inquiridos mostrou-se favorável à eleição de um presidente da UE e 53 por cento disse também estar a favor de um governo europeu. É uma tese defendida equitativamente por mulheres e homens no que diz respeito à eleição de um presidente europeu.

Nota-se, contudo, alguma diferença quando a questão é a eleição de um governo europeu. Aqui, já são os homens que se mostram mais favoráveis. Sobre um e outro assunto é, claramente, a classe média a maior defensora de um executivo europeu.

Quando questionados sobre a criação de um exército na UE, uma questão que até aqui tem levantado alguma polémica, 45 por cento dos inquiridos afirmam ser defensores desta ideia. Embora o número daqueles que se opõem não seja muito inferior - 36 por cento. Significativa é também a percentagem dos que não sabem o que responder - 19 por cento. Esta hipótese acolhe mais adeptos entre os entrevistados do sexo masculino (53 por cento) e na faixa etária que poderá ser contemplada pelas incorporações (igualmente 53 por cento). E se a maioria dos portugueses refere já ter ouvido falar do Tratado de Nice, também são peremptórios a afirmar que não fazem a mais pequena ideia das suas linhas gerais: 65 por cento sublinha que não sabe o que está consagrado no documento. Uma resposta que justifica a elevada percentagem (62 por cento) daqueles que não sabe se os deputados devem ou não aprovar o Tratado.

A larga maioria dos inquiridos (60 por cento) defende, por outro lado, que as mudanças na organização da União Europeia devem ser referendadas no nosso País. O que não deixa de ser curioso, já que as duas experiências anteriores (aborto e regiões) revelaram uma grande falta de participação dos cidadãos. Só 18 por cento tem opinião contrária e 22 por cento optou por não responder a esta questão. O alargamento da União Europeia aos países do Centro e de Leste do continente merece o acordo da maioria (64 por cento), que se mostram convencidos de que essa reestruturação

interna vai tirar poderes a Portugal no seio da UE (46 por cento). Mais de dois terços (67 por cento) considera também que o processo de alargamento poderá reduzir a atribuição de fundos comunitários para Portugal.
Embora não seja referido no artigo anterior, segundo a notícia da TSF, a sondagem envolveu **813** indivíduos adultos, dos quais 421 eram mulheres e foi realizada via telefone. É referido no artigo que **62**% dos **inquiridos** se mostra favorável à eleição de um presidente da UE.

a) Este valor de 62% é uma *estatística* ou um *parâmetro*?
b) Seria possível ter obtido este valor, se a percentagem de portugueses adultos que se mostra favorável à eleição de um presidente da UE fosse 65%?
c) Tendo em conta o resultado obtido pela sondagem da TSF/DN, acha plausível que a proporção de portugueses que se mostra favorável à eleição de um presidente da UE seja 68%? Porquê?

3.10 – No dia x do mês y do ano z realizar-se-ão as Eleições Autárquicas. Relativamente à cidade de Lisboa, há dois candidatos sobre os quais se criaram mais expectativas, nomeadamente TT e MM. Suponha que, no dia das eleições, passado uma hora sobre o fecho das urnas, altura em que começam a contar os votos para cada candidato, surgiram os primeiros resultados nos canais televisivos. Relativamente a um daqueles candidatos, o candidato TT, apresentaram o seguinte resultado: - *O candidato TT tem, neste momento, uma percentagem de 48.4%, com um erro máximo de 3.45% e uma confiança de 95%.*
1. Explique, por palavras suas, o que significa o resultado anterior.
2. Qual a amplitude do intervalo de confiança, que pode construir com os resultados apresentados no enunciado do problema, para a percentagem de lisboetas que votaram no candidato TT?
3. Acha razoável admitir que o candidato TT, ao ouvir aquele resultado, pense que tem alguma "Chance" de ganhar a Câmara de Lisboa, admitindo que para ganhar essa Câmara eram necessários, pelo menos, 50% de votos favoráveis?
4. Passadas três horas do fecho das urnas, o resultado anunciado para o candidato TT era: - *O candidato TT tem, neste momento, uma percentagem de 49.8%, com um erro máximo de 1.23% e uma confiança de 95%.*
a) Compare a amplitude do intervalo de confiança considerado na alínea 2, com a amplitude do intervalo de confiança, que pode construir com os resultados agora anunciados.
b) Como é que interpreta o resultado a que chegou na alínea anterior?
5. Quando todos os votos tiverem sido escrutinados, obtém o resultado para a percentagem de eleitores que votaram no candidato TT, na forma de um intervalo de confiança, ou na forma de um valor? Explique porquê.

3.11 - Numa altura em que se discutia o problema dos touros de morte, em Portugal, nomeadamente por causa das festas de Barrancos, uma conhecida estação de televisão propôs a seguinte questão aos telespectadores, no final do telejornal de uma 6ª feira:

- Se é a favor dos touros de morte, em Portugal, envie uma mensagem para 7771
- Se é contra os touros de morte, em Portugal, envie uma mensagem para 7772

No telejornal do dia seguinte, sábado, apresentaram a seguinte notícia, como sendo o resultado da sondagem efectuada: *72% dos portugueses são a **favor** dos touros de morte, em Portugal, enquanto que 28% são **contra***!
Acontece que o jornal Expresso, desse sábado, publicou o seguinte resultado de uma sondagem, encomendada a uma conceituada empresa de sondagens: *81% dos portugueses são **contra** os touros de morte, em Portugal*!

1. Alguma das amostras consideradas para obter os resultados anteriores, pode ser considerada enviesada? Isso poderá explicar a discrepância obtida, nas duas sondagens, relativamente às percentagens obtidas para os portugueses, que são contra os touros de morte?
2. Qual dos resultados anteriores, 28% ou 81%, estará mais perto da percentagem de portugueses que são contra os touros de morte em Portugal? Explique porquê
3. Admitindo que o resultado obtido pela empresa de sondagens, foi baseado numa amostra aleatória de dimensão 150, obtenha um intervalo de 95% de confiança para a percentagem de portugueses que são contra os touros de morte, em Portugal.
4. Calcule a margem de erro do intervalo obtido anteriormente. O que é que aconselharia a alguém, que lhe perguntasse como poderia obter um intervalo de confiança, com uma margem de erro inferior?

3.12 – O Sr. Silva, fabricante de camisas para homem, recebeu uma encomenda proveniente de Macau. Ficou um pouco preocupado, pois quando visitou este território, na sua viagem de lua-de-mel, apercebeu-se que os homens tinham, de um modo geral, os braços mais curtos. Sendo assim, não poderia utilizar os moldes habituais. Pediu, então, a uma empresa de sondagens que lhe fornecessem uma estimativa do comprimento médio dos braços dos naturais de Macau. A empresa apresentou um estudo, que se pode resumir da seguinte forma:

Sr. Silva

Apresentando os nossos cumprimentos, vimos apresentar os resultados do nosso estudo: recolhemos uma amostra de dimensão 70, de outros tantos indivíduos adultos, do sexo masculino, a quem medimos o tamanho do braço, tendo obtido como média dos 70 valores observados, o valor 52 cm.

Reiterando os nossos cumprimentos, aproveitamos para dizer que segue, em anexo, a factura do trabalho prestado.

Atenciosamente o gerente (assinatura irreconhecível)

O Sr. Silva ficou um pouco menos preocupado, mas continuava sem saber o que fazer:

1. Efectivamente, qual a confiança que poderia atribuir à estimativa obtida? Se tivesse sido outra a amostra obtida, seria de esperar obter o mesmo valor para a média? Explique porquê.
2. O Sr. Silva resolveu questionar a empresa e esta forneceu-lhe os seguintes intervalos de confiança para o tamanho médio do braço dos naturais de Macau, com uma confiança de 50% e 75%, respectivamente, e obtidos a partir da mesma amostra: [51.4, 52.6] e [51.0, 53.0].
    a. Qual a margem de erro dos intervalos anteriores?
    b. Se fosse o Sr. Silva, qual o intervalo que escolhia? O de menor amplitude ou o de maior amplitude? Explique porquê.

3.13 – Lançou-se uma moeda 50 vezes e saiu cara 20 vezes. Tem a certeza que a moeda não é equilibrada? Justifique. Sugestão: Construa um intervalo de 95% de confiança para a probabilidade de sair cara.

3.14 – Pretende-se determinar um intervalo de confiança para a proporção p, de peças defeituosas produzidas por determinada máquina. Pretende-se que a precisão seja grande, pelo que não queremos que a margem de erro seja superior a 3%. Tem-se a informação de que a máquina em questão costuma produzir cerca de 2% de peças defeituosas, mas não se tem a garantia que este valor não esteja um pouco alterado. Qual a dimensão da amostra que deve recolher para construir o intervalo pretendido?

3.15 (continuação de 3.14)– Considere de novo o exercício 3.14, mas admita que a percentagem de peças defeituosas que a máquina costuma produzir, anda à volta de 20%. Qual a dimensão da amostra que se deve recolher?

3.16 (continuação de 3.14 e 3.15) – Admita agora que não tinha qualquer informação sobre a percentagem de peças defeituosas produzidas pela máquina. Qual a dimensão da amostra que teria de recolher? Compare o valor obtido para a dimensão da amostra com os valores obtidos em 3.14 e 3.15. Comente os resultados obtidos.

3.18 – Suponha que num curso com 350 raparigas e 150 rapazes quer seleccionar uma amostra de 50 alunos.

a) Pode-se considerar que a população de onde está a seleccionar a amostra, tem dimensão suficientemente grande para poder ser considerada uma população infinita?
b) Considere a amostragem com reposição. Calcule o valor médio e o desvio padrão do estimador da proporção amostral de raparigas. Pode continuar a utilizar o mesmo valor médio e/ou o mesmo desvio padrão para o estimador da proporção amostral de raparigas, se a amostragem for feita sem reposição?
c) Qual a distribuição de amostragem, aproximada, da proporção de raparigas numa amostra de dimensão 50?
d) Qual a probabilidade, aproximada, do número de raparigas na amostra, estar entre 25 e 35?
e) Se quisesse garantir na amostra, com uma probabilidade de 0.5, 25 raparigas, quantos alunos devia seleccionar?

Lista de algumas funções usadas no Excel:

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

Bibliografia

Graça Martins, M. E. (2005) – *Introdução à Probabilidade e à Estatística*. Sociedade Portuguesa de Estatística.

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

Graça Martins. M. E. and al (1999) – *Introdução às Probabilidades e à Estatística*. Universidade Aberta.

Graça Martins, M.E. e Loura, M. E. (2001) – *Matemática para as Ciências Sociais* – Anexo para apoio à interpretação do programa.

Moore, D. (1977) – *Statistics. Concepts and Controversies*. W.H. Freeman and Company, New York.

Moore, D. and al (1992) – Perspectives in Contemporary Statistics. The Mathematical Association of America.

Moore, D. and al (1993) – *Introduction to the Practice of Statistics*. W.H. Freeman and Company, New York.

Murteira, B. (1993) – *Análise Exploratória de Dados. Estatística Descritiva*. McGraw-Hill de Portugal.

Rossman, A. and al (2001) – *Workshop Statistics. Discovery with data*. Key College Publishing/Spinger-Verlag. New York, Inc.

Tannenbaum, P. and al (1998) – *Excursions in Modern Mathematics*. Prentice Hall.

Velleman, P. and al (2004) – *Intro Stats*. Pearson Education, Inc.

Artigos da revista TEACHING STATISTICS

AGEEL, M.I. – Spreadsheets as a Simulation Tool for Solving Probability Problems, Vol 24, 2, 51-54.
Hodgson, T., and Borkowski, J. - Why Stratify? Vol 20, 1, 68-71.
NEVILLE, H. – Handling Continuous Data in Excel, Vol 25, 2, 42-45.
NEVILLE, H. – Charts in Excel, Vol 26, 2, 49-53.

Páginas na Internet

INSTITUTO NACIONAL DE ESTATÍSTICA E ESCOLA SECUNDÁRIA TOMAZ PELAYO
PROJECTO ALEA – http://www.alea.pt
(Desta página recomenda-se, em especial, o dossier didáctico "ESTATÍSTICA COM EXCEL" da autoria de Luís Miguel Cunha).
INSTITUTO NACIONAL DE ESTATÍSTICA – www.ine.pt/
Tem informação sobre Portugal, ao nível da freguesia.
EUROSTAT – europa.eu.int/comm/eurostat/
Tem informação relativa aos diversos países da Europa.