Introdução à Inferência Estatística

Maria Eugénia Graça Martins

Departamento de Estatística e Investigação Operacional

Faculdade de Ciências da Universidade de Lisboa

Abril de 2006

Introdução à Inferência Estatística

Com este módulo, pretende-se fazer um pequeno curso introdutório à Inferência Estatística. Serão abordados os conceitos necessários para se chegar de uma forma simples, à interpretação e compreensão de certo tipo de informação veiculada pela comunicação social, nomeadamente a que diz respeito às sondagens.

Ao escrever estas folhas pensei particularmente nos professores que leccionam a disciplina de Matemática para as Ciências Sociais, já que correspondem ao conteúdo programático desta disciplina. Não se pretendem que sejam um substituto de outro tipo de informação utilizada por estes professores, mas sim um complemento.

Não pretendo apresentar estas folhas como um produto acabado, mas sim como um passo para um trabalho que possa ser continuamente melhorado com as críticas e sugestões, que desde já agradeço, da parte dos meus colegas professores.

Maria Eugénia Graça Martins
memartins@fc.ul.pt

Índice

Quando a comunicação social, a propósito de uma sondagem, transmite a seguinte notícia[1]:

Sondagem

10% não sabem quem é o Presidente da República

DEZ por cento dos portugueses não sabem quem é o Presidente da República e 9 por cento desconhecem a identidade do primeiro-ministro. Uma sondagem de 2000 inquiridos EXPRESSO/Euroexpansão revela ainda índices mais desoladores para o presidente da Assembleia da República (só identificado por 39 por cento dos inquiridos), para os líderes partidários (desconhecidos de mais de metade do universo) e para os chefes dos grupos parlamentares (ignorados pela quase totalidade da amostra). Os dados da sondagem mostram ainda que os portugueses não distinguem entre António Guterres/ primeiro-ministro e António Guterres/secretário-geral do PS: 91 por cento sabem que ele é o chefe de Governo, mas 52 por cento ignoram que é ele o líder dos socialistas (**ver pág. 7**).

Ficha Técnica

Sondagem efectuada entre os dias 6 e 31 de Janeiro. O universo é constituído pela população de Portugal Continental, com idades entre os 18 e os 74 anos. A amostra é de 1964 indivíduos, entrevistados directamente, nas suas residências, A margem de erro é de 1.3%, para uma confiança é de 95%.

como interpretamos a ficha técnica que a acompanha?

Com este módulo pretendemos responder a esta questão. Estaremos aptos a saber interpretar o resultado de uma sondagem, nomeadamente, sabendo o que se entende por confiança, o que é a margem de erro, porquê uma amostra de 1964 indivíduos, etc.

[1] Exemplo adaptado de uma notícia do Expresso de 15/03/97

1 Introdução

1.1 O que é a Estatística?

A Estatística é uma ciência que estuda a **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.

Tal como refere David Moore, em Perspectives of Contemporary Statistics, podemos considerar três grandes áreas nesta ciência dos dados:

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

O tema da Aquisição de dados, merece relevo especial, pois deverão ser recolhidos numa perspectiva em que será a partir da informação que eles fornecem que iremos responder a determinadas questões, isto é, retirar conclusões para as Populações subjacentes a esses dados – contexto em que tem sentido fazer **Inferência Estatística**.

1.2 Probabilidade e Estatística?

A **Probabilidade** é o instrumento que permite ao Estatístico utilizar a informação recolhida da amostra, para descrever ou fazer inferências sobre a População de onde a amostra foi recolhida. Podemos dizer que os objectivos da Probabilidade e da Estatística são, de certo modo, inversos.

Quando assumimos que a População é conhecida, podemos fazer raciocíonios que vão do geral para o particular, isto é, da População para a Amostra. Quando a População não é conhecida, utilizamos a Estatística no sentido inverso, isto é, para inferir para a População resultados observados na Amostra.

Exemplo – Consideremos a População constituída pelos alunos inscritos na FCUL, no ano lectivo de 2005/2006. Relativamente a esta população, seja p a percentagem de alunos que pratica regularmente desporto. Recolhida uma amostra de 10 alunos, com reposição:

- se conhecermos o valor de p, por exemplo p=0.298, podemos calcular a probabilidade de haver x alunos, a praticar desporto, nos 10 alunos seleccionados. Para calcular esta probabilidade, basta pensar que a variável X, que representa o número de alunos em 10 que pratica desporto, é bem modelada por uma Binomial, neste caso com parâmetros 10 e 0.298. Então, por exemplo, P(X=3) = 0.2668 (Valor calculado no Excel).
- se não conhecermos o valor de p, vamos utilizar o número x de alunos, que praticam desporto, nos 10 seleccionados, para “estimar” p, e temos um problema de Inferência Estatística. Se, por exemplo, x=3, diremos que uma estimativa para p, é 0.3. A partir deste valor temos processos que nos permitem tomar uma decisão sobre o parâmetro p, quantificando ainda o erro cometido ao tomar essa decisão.

No que se segue vamos estudar alguns exemplos de Inferência Estatística, nomeadamente no que diz respeito à estimação de parâmetros, na forma de **Intervalos de Confiança**.

2 Inferência Estatística

2.1 Introdução

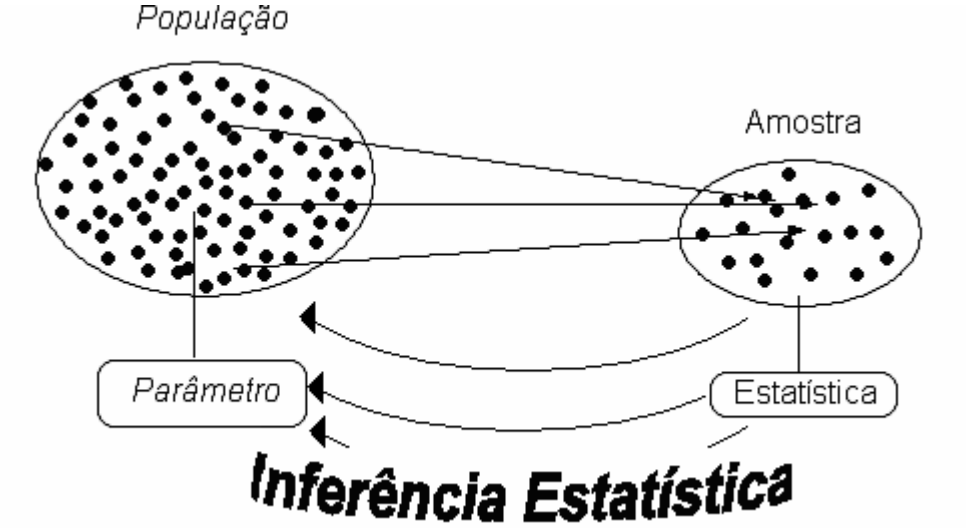
O que é? Quando se utiliza? Para que serve?

- É um processo de raciocínio indutivo, em que se procuram tirar conclusões indo do particular, para o geral. É um tipo de raciocínio contrário ao tipo de raciocínio matemático, essencialmente dedutivo.
- Utiliza-se quando se pretende estudar uma ***população***, estudando só alguns elementos dessa população, ou seja, uma **amostra**.
- Serve para, a partir das propriedades verificadas na **amostra**, inferir propriedades para a ***população***.

2.2 Parâmetro e estatística

Quando se pretende estimar (obter um valor aproximado) um ***parâmetro -*** característica numérica da população, considera-se uma função conveniente, que só dependa dos valores da amostra – **estatística**, a que se dá o nome de **estimador** do parâmetro em estudo. Ao valor desta função a que chamámos estimador, calculada para uma determinada amostra recolhida, chamamos **estimativa**. Também se utiliza o termo estatística como significado de estimativa. Surge assim o conceito de estatística – característica numérica da amostra, por oposição a parâmetro - característica numérica da população.

No seguinte esquema, procuramos traduzir o processo de Inferência Estatística, nomeadamente no que diz respeito à estimação de parâmetros



Embora, neste curso, não abordemos outros temas que os de estimação de parâmetros, a inferência estatística dispõe de instrumentos poderosos que nos permitem tomar decisões de outro tipo. O importante e que convém registar, é que as decisões que tomamos têm inerente um determinado erro, que pode ser quantificado em termos probabilísticos.

2.3 Amostra enviesada. Amostra aleatória e amostra não aleatória. Distribuição de amostragem

2.3.1 Amostra enviesada e amostra aleatória

Como dissemos anteriormente, as decisões que tomamos têm inerente um determinado erro, erro este que é inerente à variabilidade presente na amostra que se recolhe, com o objectivo de tomar decisões, sobre o parâmetro que estamos a estudar. 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.

2.3.2 Distribuição de amostragem

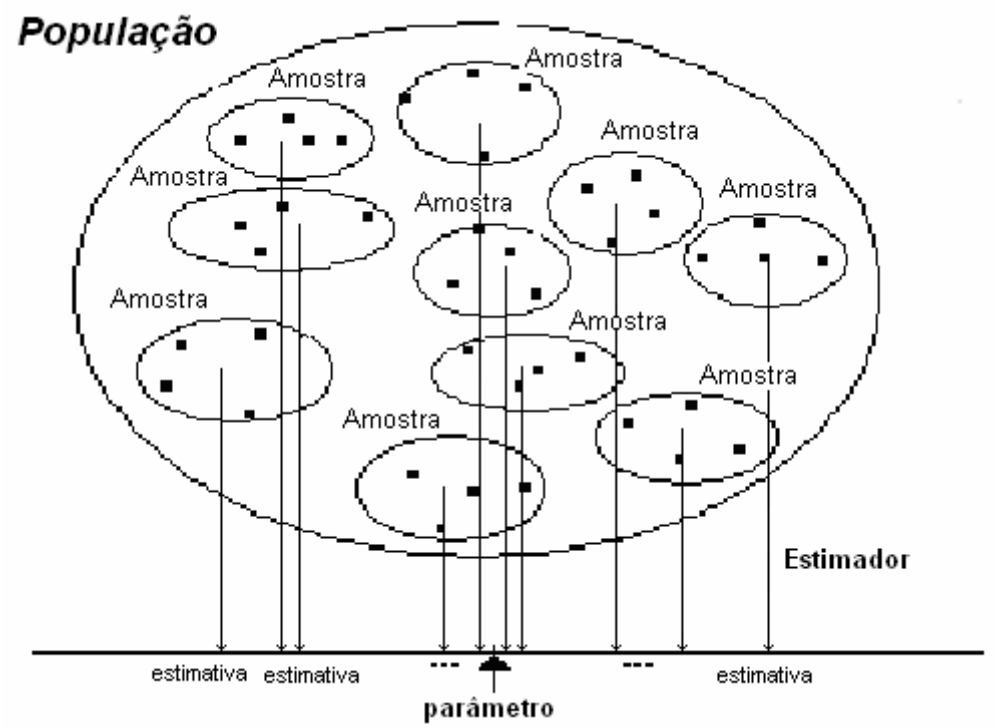
Normalmente obtêm-se amostras enviesadas quando existe a intervenção do factor humano. Com o objectivo de minimizar o enviesamento, no planeamento da escolha da amostra deve ter-se presente o princípio da aleatoriedade de forma a obter uma amostra aleatória.

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 a seguir, 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.

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



Como se comportam todas estas **estimativas**, relativamente ao ***parâmetro***, em estudo?

A resposta é dada estudanto a **distribuição de amostragem** do estimador (não esqueça que o estimador é uma função dos elementos da amostra e que para cada amostra que se recolhe, se obtém um valor dessa função, que se chama estimativa!).

3 Estimador centrado e não centrado (ou enviesado)

Quando é que dizemos que temos um "bom" estimador?

Uma vez escolhido um plano de amostragem aleatório, ao pretendermos estimar um parâmetro, pode ser possível utilizar várias estatísticas (estimadores) diferentes. Por exemplo, quando pretendemos estudar a variabilidade presente numa População, que pode ser medida pela variância populacional σ^2, sabemos que podemos a partir de uma amostra recolhida (x_1, x_2, ..., x_n), obter duas estimativas diferentes para essa variância, a partir das expressões

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

Quais as razões que nos podem levar a preferir uma das estatísticas relativamente à outra? Qual o estimador preferido? S^2 ou S'^2?

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

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

Uma das razões que nos levam a preferir o estimador S^2 para a variância, relativamente a S'^2, é o facto de não apresentar enviesamento (pelo menos para o plano de amostragem que iremos utilizar).

Aparece-nos, novamente a palavra enviesamento, mas noutro contexto. Efectivamente, relacionado com um processo de amostragem e com 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 ou não centrado.

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

A **falta de precisão** 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 problema do enviesamento da amostra se reduz com o recurso a amostras aleatórias, a precisão aumenta-se, aumentando a dimensão da amostra (como veremos).

4 Técnicas de amostragem aleatória

Existem várias técnicas de amostragem aleatória. No entanto, no estudo da distribuição de amostragem dos estimadores *Média* e *Proporção amostral*, utilizados, respectivamente, para estimar os parâmetros **Valor médio** e **Proporção populacional**, vamos limitar-nos a considerar amostras aleatórias obtidas de forma a satisfazerem os seguintes critérios:

- Dada uma população de dimensão N, considera-se que cada elemento da população, deve ter a mesma probabilidade, igual a 1/N, de ser seleccionado para a amostra;
- A selecção é feita com reposição.

Existem outras técnicas de amostragem aleatória, como a amostragem aleatória simples (dada uma população de dimensão N, uma amostra aleatória simples, de n elementos, é aquela tal que, qualquer outro conjunto de n elementos, tem igual probabilidade de ser seleccionado), a amostragem sistemática, a amostragem estratificada, etc. Qualquer uma destas técnicas, aplicadas na recolha das amostras, conduz a que as propriedades dos estimadores utilizados para estimar os mesmos parâmetros, sejam diferentes. Antes de formalizarmos o estudo dos estimadores Média e Proporção amostral, vamos exemplificar um tipo de amostragem **com reposição** e **sem reposição**, e as implicações nas propriedades do estimador, na estimação de uma proporção (Consideramos um exemplo com interesse unicamente teórico, para fins de exemplificação).

Exemplo – No Departamento de Estatística há 5 docentes que são professores associados, dos quais 3 são mulheres – Maria, Ana, Rita e 2 são homens – Pedro e Tiago. Se representarmos por **p** a percentagem de homens que são professores associados, temos que p=2/5 (Numa situação de interesse, a população seria razoavelmente grande e a proporção p seria desconhecida – situação que se verifica quando se pretende averiguar a percentagem de eleitores que pretendem votar num determinado candidato). Suponhamos que pretendíamos estimar esta proporção utilizando a proporção $\hat{\mathbf{p}}$ de homens em amostras de dimensão 2. Então vamos construir todas as amostras desta dimensão para obter a distribuição de amostragem da estatística utilizada:

a) Com reposição

Amostra	$\hat{\mathbf{p}}$	Amostra	$\hat{\mathbf{p}}$
Maria, Maria	0	Rita, Pedro	1/2
Maria, Ana	0	Rita, Tiago	1/2
Maria, Rita	0	Pedro, Maria	1/2
MariaPedro	1/2	Pedro, Ana	1/2
MariaTiago	1/2	Pedro, Rita	1/2
Ana, Maria	0	Pedro, Pedro	2/2
Ana, Ana	0	Pedro, Tiago	2/2
Ana, Rita	0	Tiago, Maria	1/2
Ana, Pedro	1/2	Tiago, Ana	1/2
Ana, Tiago	1/2	Tiago, Rita	1/2
Rita, Maria	0	Tiago, Pedro	2/2
Rita, Ana	0	Tiago, Tiago	2/2
Rita, Rita	0		

A partir da tabela anterior é possível obter a distribuição de amostragem da estatística $\hat{\mathbf{p}}$:

$\hat{\mathbf{p}}$	0	.5	1
Probabilidade	9/25	12/25	4/25

$$E(\hat{\mathbf{p}}) = 2/5 \text{ e } Var(\hat{\mathbf{p}}) = 3/25$$

Repare-se que o valor médio da estatística $\hat{\mathbf{p}}$ coincide com o valor do parâmetro **p** que se está a estimar.

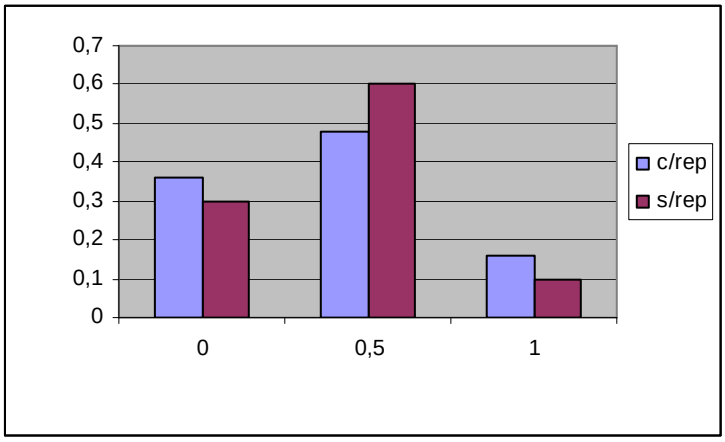
b) Sem reposição(Amostragem aleatória simples)

Amostra	$\hat{\mathbf{p}}$	Amostra	$\hat{\mathbf{p}}$
Maria, Ana	0	Ana, Pedro	1/2
Maria, Rita	0	Ana, Tiago	1/2
Maria, Pedro	1/2	Rita, Pedro	1/2
Maria, Tiago	1/2	Rita, Tiago	1/2
Ana, Rita	0	Pedro, Tiago	1

$\hat{\mathbf{p}}$	0	.5	1
Probabilidade	3/10	6/10	1/10

$$E(\hat{\mathbf{p}}) = 2/5 \text{ e } Var(\hat{\mathbf{p}}) = 9/100$$

Repare-se que, ainda neste caso, o valor médio da estatística (estimador) $\hat{\mathbf{p}}$ coincide com o valor do parâmetro **p** que se está a estimar, mas a variância é inferior à obtida na amostragem com reposição. Comparando as duas distribuições de amostragem, do mesmo estimador, mas para os esquemas de amostragem diferentes, temos



Neste esquema de amostragem aleatória simples (uma amostra aleatória simples de n elementos, é definida como sendo uma amostra tal que qualquer outro conjunto de n elementos da população, tem igual probabilidade de ser seleccionado – os elementos podem ser seleccionados sequencialmente, sem reposição, ou em bloco, todos de uma vez), não interessa a ordem pela qual os elementos são seleccionados, pelo que o número de amostras diferentes é igual a $\binom{5}{2}$=10.

Exemplo (cont) - Suponhamos ainda que, relativamente ao exemplo anterior, estávamos interessados em estimar o parâmetro p, mas através da estatística X - número de homens em amostras de n elementos.

Então o estimador de p, $\hat{\mathbf{p}} = \frac{X}{n}$ onde X é a variável que dá o nº de homens numa amostra de dimensão n. Utilizando ainda amostras de dimensão 2, vamos obter a distribuição de amostragem de $\hat{\mathbf{p}} = \frac{X}{2}$, começando pelo estudo da estatística **X** – nº de homens em amostras de dimensão 2:

a) Com reposição

Amostra	x	Amostra	x
Maria, Maria	0	Rita, Pedro	1
Maria, Ana	0	Rita, Tiago	1
Maria, Rita	0	Pedro, Maria	1
Maria, Pedro	1	Pedro, Ana	1
Maria, Tiago	1	Pedro, Rita	1
Ana, Maria	0	Pedro, Pedro	2
Ana, Ana	0	Pedro, Tiago	2
Ana, Rita	0	Tiago, Maria	1
Ana, Pedro	1	Tiago, Ana	1
Ana, Tiago	1	Tiago, Rita	1
Rita, Maria	0	Tiago, Pedro	2
Rita, Ana	0	Tiago, Tiago	2
Rita, Rita	0		

A partir da tabela anterior é possível obter a distribuição de amostragem da estatística X:

X=x	0	1	2
P(X=x)	9/25	12/25	4/25

Repare-se que a distribuição de amostragem da estatística X não é mais do que a distribuição Binomial de parâmetros 2 e 2/5 (sabemos neste caso que a proporção de homens é 2/5):

$P(X=x)=\binom{2}{x}\left(\frac{2}{5}\right)^{x}\left(\frac{3}{5}\right)^{2-x}$, com x=0, 1, 2

$$E(X) = 4/5 \text{ e } Var(X)= 12/25$$

Então o estimador $\hat{\mathbf{p}}$ é tal que

$\hat{\mathbf{p}}$	0	.5	1
Probabilidade	9/25	12/25	4/25

tal como seria de esperar, pois já havia sido obtido anteriormente.

b) Sem reposição (esquema de amostragem aleatória simples)

Amostra	x	Amostra	x
Maria, Ana	0	Ana, Pedro	1
Maria, Rita	0	Ana, Tiago	1
Maria, Pedro	1	Rita, Pedro	1
Maria, Tiago	1	Rita, Tiago	1
Ana, Rita	0	Pedro, Tiago	2

X=x	0	1	2
P(X=x)	3/10	6/10	1/10

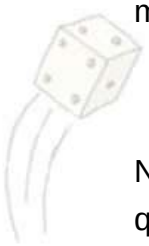
$$E(X) = 4/5 \text{ e } Var(X)= 9/25$$

Repare-se que, agora, para modelar X, já não podemos utilizar o modelo Binomial, mas sim o chamado modelo Hipergeométrico:

$P(X=x)=\dfrac{\binom{2}{x}\binom{3}{2-x}}{\binom{5}{2}}$ com x=0, 1, 2

Então, para $\hat{\mathbf{p}}$ teremos as propriedades já obtidas anteriormente, quando obtivemos a sua distribuição de amostragem directamente.

Observação – Utilizando o mesmo estimador, mas com um esquema diferente de selecção das amostras, temos distribuições de amostragem diferentes. Este exemplo teve como objectivo fazer intervir dois modelos de probabilidade conhecidos – o modelo Binomial e o modelo Hipergeométrico



Não nos debruçaremos sobre esquemas de amostragem sistemática, por estratificação, por quotas, etc, uma vez que não estudaremos as propriedades dos estimadores para estas técnicas de amostragem.

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

Outro problema que se levanta com a recolha da amostra é o de saber qual a **dimensão** desejada para a amostra a recolher.

Este é um problema para o qual, nesta fase, não é possível avançar nenhuma teoria, mas sobre o qual se podem tecer algumas considerações gerais. Pode-se começar por dizer que, para se obter uma amostra que permita calcular estimativas suficientemente precisas dos parâmetros a estudar, a sua dimensão depende muito da variabilidade da população subjacente (como mostraremos mais à frente).

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*".

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!

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

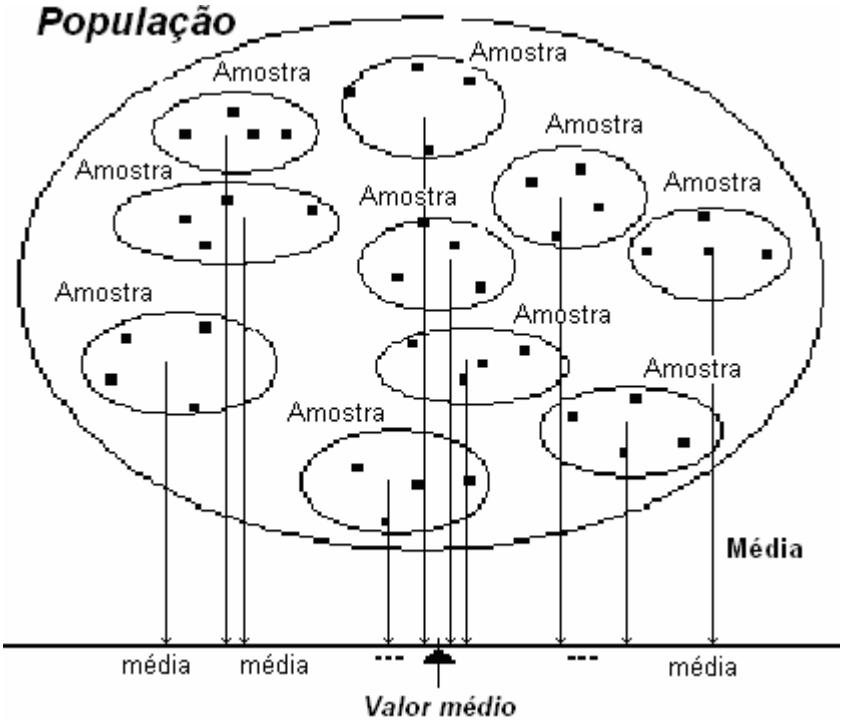
- a estimação do **valor médio** (ou média populacional), pela média (amostral), e
- a estimação da **proporção** (populacional) pela proporção amostral.

6 Estimação do valor médio

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, então é natural considerar como ***estimador*** a função **Média**, que para cada amostra observada dará uma estimativa do parâmetro.

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

Será que para as diferentes amostras que podemos obter da população, as diferentes estimativas são próximas umas das outras e do parâmetro ***valor médio***? Se isso acontecer, temos uma certa garantia que a amostra que seleccionámos, já que na prática só se selecciona uma amostra, nos fornece 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* populacionais 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.

6.1 Distribuição de amostragem da Média, como estimador do valor médio

6.1.1 Distribuição de amostragem exacta da Média

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.

Exemplo – Considere uma população constituída pelos elementos 1, 2, 3, 4 e 5. Pretende estimar o valor médio desta população, utilizando, como estimativa, a média de uma amostra de dimensão 2, obtida com reposição. Obtenha a distribuição de amostragem do estimador utilizado.

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

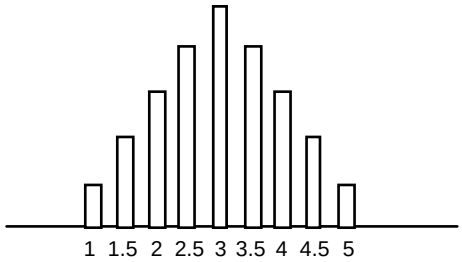
Desvio padrão = $\sqrt{2}$.

A metodologia seguida para obter a distribuição de amostragem consiste em obter 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 para termos presente que se obtém 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

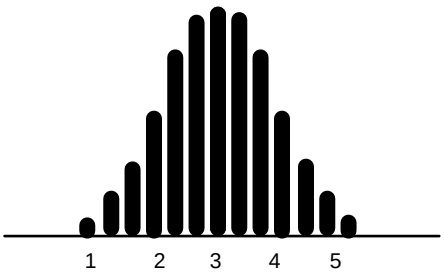
Desvio padrão = 1

Algumas observações:

- O centro da distribuição de amostragem do estimador Média2 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

Desvio padrão = 0.816

Algumas observações:

- O centro 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 ($\sqrt{2}/\sqrt{3}=0.816$).
- 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.

Para melhor comparação dos processos anteriores, resumimos na tabela seguinte algumas características da população e da distribuição de amostragem da Média para amostras de dimensão 2 e 3:

	Valor médio	Desvio padrão
População	3	1.414
Média (amostras dimensão 2)	3	1.000
Média (amostras dimensão 3)	3	0.816

Considerámos um exemplo de uma população muito pequena, em que foi simples obter a distribuição de amostragem da Média. E se a dimensão da população fosse igual a 20 e pretendessemos recolher amostras de dimensão 5, quantas amostras teríamos de recolher para obter a distribuição de amostragem da Média? Nada mais, nada menos que 3 200 000! Como vemos, este processo de obter a distribuição de amostragem da Média seria impraticável, mesmo para populações razoavelmente pequenas. Como proceder então?

Vamos ver que, embora não seja fácil (a maior parte das vezes) obter a distribuição de amostragem exacta, podemos ter a distribuição de amostragem aproximada da Média, que já nos é bastante útil.

6.1.2 Distribuição de amostragem aproximada da Média

No exemplo anterior a população era razoavelmente pequena, pelo que foi possível calcular a distribuição de amostragem exacta da estatística Média, como estimador do valor médio. Vamos considerar agora uma situação ainda de uma população finita, mas suficientemente grande para não ser possível (dentro dos limites do razoável...) obter a distribuição exacta.

Exemplo: Considere a seguinte tabela onde se apresentam os 97 trabalhadores de uma determinada empresa:

Número	Nome	Estado civil	Idade	Altura	Nº filhos
1	Alexandra Almeida	solteira	26	160	0
2	Alexandre Carmo	casado	30	174	2
3	Alda Morais	casada	37	160	3
4	Ana Ribeiro	casada	23	159	1
5	Ana Cristina Santos	casada	26	156	2
6	Ana Cristina Oliveira	solteira	25	153	0
7	Anabela Pais	divorciada	33	156	3
8	António Couto	solteiro	24	177	0
9	António Fernandes	casado	42	161	5
10	António Pinto	casado	51	171	1
11	Armando Ferreira	casado	48	167	1
12	Carlos Matos	casado	37	165	1
13	Carlos Sampaio	casado	40	174	2
14	Cristina Vicente	casada	39	160	2
15	Cristina Zita	casada	27	164	1
16	Dora Ferreira	casada	50	170	4
17	Elsa Sampaio	casada	45	160	4
18	Fernando Barroso	casado	43	164	3
19	Fernando Martins	casado	29	165	1
20	Fernando Santos	divorciado	32	174	2
21	Filomena Silva	solteira	20	165	0
22	Francisco Gomes	casado	26	174	0
23	Isabel Soares	solteira	22	156	0
24	Isabel Silva	casada	34	148	2
25	João Morais	casado	44	171	2
26	João Sousa	solteiro	25	176	0
27	Luis Horta	casado	35	169	2
28	Luis Sousa	casado	37	170	0
29	Luis Ribeiro	casado	49	170	1
30	Manuel Santos	casado	54	175	4
31	Manuel Pereira	divorciado	47	162	3
32	Manuel Teixeira	casado	50	173	2
33	Margarida Almeida	casada	51	166	1
34	Margarida Simões	casada	47	161	4
35	M. Adelina Azevedo	solteira	25	148	0
36	M. Alexandra Almeida	solteira	26	158	0
37	M. Alexandra Ribeiro	casada	39	157	3
38	M. Cristina Carvalho	casada	41	158	2
39	M. Cristina Freire	divorciada	38	161	1
40	M. De Fátima Osório	casada	33	164	1
41	M. Fernanda Rocha	solteira	29	154	0
42	M. Isabel Frade	casada	38	164	2
43	M. Isabel Santos	solteira	26	164	0
44	M. Luisa Faria	casada	35	164	2
45	M. Manuel Trindade	casada	29	167	0
46	M. Manuela Lino	casada	33	159	3
47	M. Nazaré Pinto	solteira	29	162	0
48	M. Neusa Lopes	casada	34	163	2
49	M. Olga Martins	casada	27	165	0
50	M. Paula Pitarra	casada	29	160	3
51	M. Paula Garcês	solteira	25	150	0
52	M. Rosário Gomes	solteira	27	155	0
53	M. Rute Costa	solteira	45	160	0

54	M. Rute Rita	solteira	23	165	0
55	M. Teresa António	casada	46	147	2
56	M. Teresa Bento	casada	54	158	1
57	M. Teresa Garcia	solteira	22	154	0
58	Mário Martins	casado	29	171	1
59	Mário Reis	casado	43	172	0
60	Nuno Simões	casado	43	176	2
61	Nuno Ventura	solteiro	28	175	0
62	Olga Martins	solteira	29	159	0
63	Oscar Trigo	casado	35	169	1
64	Osvaldo	casado	44	172	1
65	Paulo Nunes	casado	38	169	1
66	Paulo Martins	solteiro	41	173	1
67	Paulo Santos	solteiro	51	172	1
68	Paulo Valente	casado	45	168	2
69	Pedro Casanova	casado	46	175	1
70	Pedro Dalo	casado	37	166	1
71	Pedro Martins	casado	39	174	2
72	Pedro Lisboa	casado	44	163	2
73	Pedro Sintra	solteiro	40	170	0
74	Pedro Valente	casado	32	161	0
75	Pedro Viriato	casado	26	169	0
76	Rita Amaral	solteira	23	165	0
77	Rita Bendito	solteira	29	159	0
78	Rita Évora	casada	34	162	1
79	Rita Seguro	solteira	30	163	0
80	Rita Valente	casada	35	170	2
81	Rufo Almeida	solteiro	29	171	0
82	Rui André	solteiro	31	165	0
83	Rui Martins	casado	34	167	0
84	Rui Teixeira	casado	44	166	2
85	Rui Vasco	casado	45	178	2
86	Sérgio Teixeira	divorciado	40	174	2
87	Sílvio Lino	divorciado	44	161	0
88	Tânia Lopes	casada	27	160	0
89	Tânia Martins	solteira	25	162	0
90	Teresa Adão	casada	26	163	1
91	Teresa Paulo	solteira	28	164	0
92	Teresa Vasco	casada	30	157	0
93	Vera Mónica	solteira	25	161	0
94	Vera Patrícia	solteira	26	154	0
95	Vera Teixeira	casada	31	162	1
96	Vitor Santos	casado	37	173	2
97	Vitor Zinc	solteiro	49	169	0

No que diz respeito às variáveis Sexo, Idade, Altura e Número de filhos, a população anterior tem as seguintes características:

Tabela 1

	Freq. abs.	Freq. rel.
Feminino	52	0.536
Masculino	45	0.464
	97	1.000

Tabela 2

Variável	Valor Médio	Desvio padrão	Mínimo	Máximo
Idade	35.19	8.84	20	54
Altura	164.57	7.05	147	178
Nº filhos	1.13	1.21	0	5

Repare-se que para a variável Sexo não calculámos nem a média nem o desvio padrão, já que se trata de uma variável qualitativa.

1 . Estimação da altura média da População constituída pelas alturas dos trabalhadores

a) Amostras de dimensão 15

Utilizando o Excel, seleccionamos 50 amostras de dimensão 15, da população constituída pelas 97 alturas, e para cada amostra calculámos a média:

	A	B	C	D	AX	AY	AZ
17		Am 1	Am 2	Am 3	...	Am 49	Am 50
18		157	171	171		159	156
19		160	162	171		166	173
20		170	162	172		165	172
21		172	171	161		171	150
22		154	154	163		160	174
23		173	163	159		164	176
24		176	162	174		155	171
25		165	160	174		160	150
26		161	174	171		173	169
27		174	174	173		159	164
28		170	174	176		169	170
29		160	174	159		165	164
30		165	170	165		147	174
31		163	162	175		157	160
32		165	162	175		171	165
33	médias	165,7	166,3	169,3	...	162,7	165,87

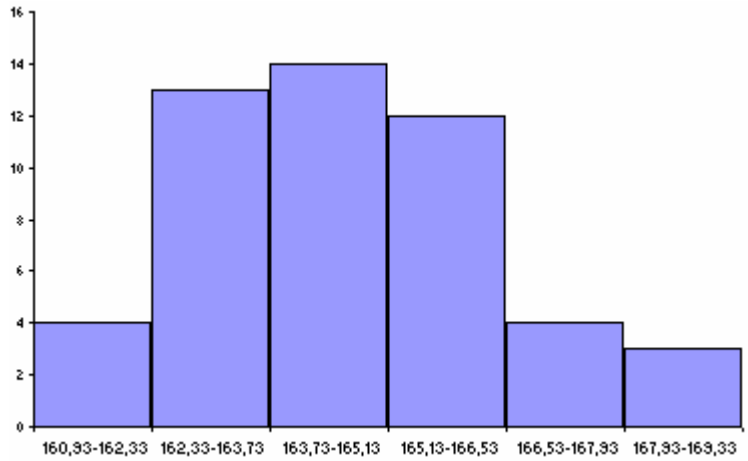
Sheet4 / Sheet1

A redução dos elementos da amostra, constituída pelas 50 médias, através de algumas estatísticas descritivas e da construção do histograma, conduziu aos seguintes resultados:

	B	C	D
37		Mean	164,62
38		Median	164,53
39		Standard Deviation	1,7724
40		Sample Variance	3,1412
41		Minimum	160,93
42		Maximum	169,27
43		Percentil 2,5%	161,74
44		Percentil 5%	162,23
45		Percentil 95%	167,69
46		Percentil 97,5%	168,35
47		Count	50

Sheet4

Count of médias	
médias	Total
160,93-162,33	4
162,33-163,73	13
163,73-165,13	14
165,13-166,53	12
166,53-167,93	4
167,93-169,33	3
Grand Total	50



Algumas conclusões:

- A distribuição da amostra das médias faz-se de forma aproximadamente simétrica em torno do valor 164.6, que é um valor muito próximo do parâmetro em estudo - valor médio da população (variável Altura)
- A distribuição da amostra das médias apresenta uma variabilidade muito pequena, quando comparada com a distribuição da população;
- Da tabela das características amostrais verificamos que 90% dos elementos da amostra das médias estão no intervalo [162.23; 167.69], enquanto que 95% dos elementos da amostra estão no intervalo [161.74; 168.35]. Estes intervalos, de amplitude 5.46 e 6.59 contêm o valor do parâmetro “altura média”.

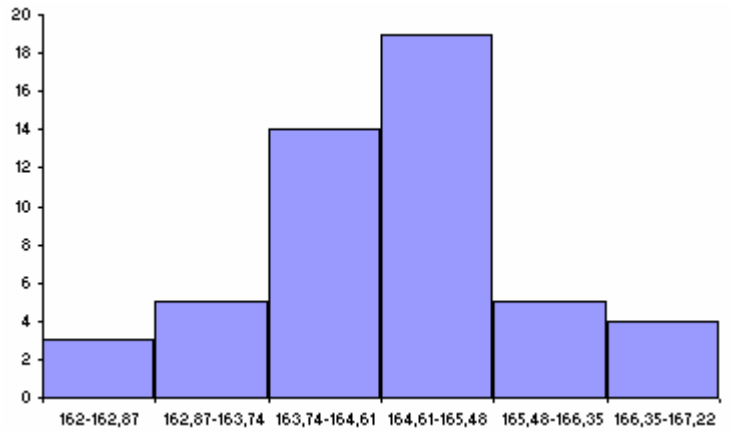
E se em vez de termos seleccionado amostras de dimensão 15, tivessemos seleccionado amostras de dimensão 30?

b) Amostras de dimensão 30

Utilizando ainda o Excel, gerámos 50 amostras de dimensão 30. Uma análise dos dados, idêntica à feita para as anostras de dimensão 15, conduziu aos seguintes resultados:

	D	E
68	Mean	164,70
69	Median	164,80
70	Standard Deviation	1,1192
71	Sample Variance	1,2526
72	Range	5,17
73	Minimum	162,00
74	Maximum	167,17
75	Percentil 2,5%	162,59
76	Percentil 5%	162,85
77	Percentil 95%	166,60
78	Percentil 97,5%	167,03
79	Count	50

Count of Médias	
Médias	Total
162-162,87	3
162,87-163,74	5
163,74-164,61	14
164,61-165,48	19
165,48-166,35	5
166,35-167,22	4
Grand Total	50



Algumas conclusões:

- Do mesmo modo que para as amostras de dimensão 15, a distribuição da amostra das médias das amostras de dimensão 30 também é aproximadamente simétrica em torno do valor 164.7, que é um valor muito próximo do parâmetro em estudo - valor médio da população (característica Altura);
- A distribuição da amostra das médias apresenta uma variabilidade muito pequena, quando comparada com a distribuição da população e é mais pequena do que no caso das amostras de dimensão 15;
- Da tabela das características amostrais verificamos que 90% dos elementos da amostra das médias estão no intervalo [162.85; 166.60], enquanto que 95% dos elementos da amostra estão no intervalo [162.59; 167.03]. Estes intervalos, de amplitude 3.75 e 4.44 contêm o valor do parâmetro “altura média”.

Os resultados anteriores levam-nos a pensar que quanto maior for a dimensão das amostras consideradas menor será a variabilidade entre as médias dessas amostras.

Quando recolhemos as 50 amostras e calculámos a média de cada uma dessas amostras, ficámos com uma ideia do comportamento da ***estatística*** Média, que resumimos no seguinte:

- ❖ Quando consideramos amostras da mesma dimensão, a média varia de amostra para amostra, mas apresenta um comportamento característico, de uma distribuição aproximadamente simétrica, com pequena variabilidade.
- ❖ Quanto maior for a dimensão da amostra, espera-se que seja melhor a estimativa fornecida pela *estatística* “Média” para o *parâmetro* “valor médio” da população que se está a estudar, já que a variabilidade apresentada pelas diferentes estimativas, relativamente ao parâmetro a estimar, diminui.

E se em vez de 50 amostras considerássemos todas as amostras possíveis (diferentes) que se podem extrair da População?

No nosso caso, se quiséssemos amostras de dimensão 30, teríamos de seleccionar 97^{30} amostras! Isto seria muito trabalhoso, mas só assim é que teríamos verdadeiramente a **distribuição de amostragem exacta da Média** para amostras de dimensão 30, isto é, os diferentes valores que a variável

$$\overline{X} = \frac{X_1 + X_2 + ... + X_{30}}{30}$$

pode assumir e a probabilidade de assumir esses valores (Estamos a representar a variável que está a ser estudada por um X, pelo que X_1 representa a 1ª vez que se foi seleccionar um elemento, X_2 representa a 2ª vez que se foi seleccionar um elemento, etc.)

A obtenção da distribuição de amostragem exacta seria uma tarefa árdua, pelo que nos vamos contentar em obter uma aproximação para essa distribuição de amostragem.

Observação 1 - Repare-se que a Média $\bar{x}$ é uma variável aleatória pois os seus valores dependem dos valores das variáveis X_1, X_2, …, X_{30}. Quando observamos um valor de X_1, que representamos por x_1, um valor de X_2, que representamos por x_2, etc, e substituímos esses valores observados na expressão da Média, obtemos um valor observado para a Média, que representamos por $\bar{x}$. Assim, enquanto a variável se representa por letra maiúscula, um valor observado dessa variável representa-se por letra minúscula.

Observação 2 - Aproveitamos para lembrar que a amostragem foi feita **com reposição**, pois cada vez que se selecciona um elemento ele é reposto, antes de seleccionar o seguinte. Esta observação é sobretudo relevante para Populações de dimensão pequena (como a considerada no nosso estudo), em que a composição da População sofre alteração quando se retiram alguns elementos, o que não sucede com Populações de grande dimensão - que é normalmente a situação de interesse em Estatística.

6.1 3 Como obter a distribuição de amostragem da Média?

Então para obter a distribuição de amostragem da Média não é necessário considerar todas as amostras possíveis e depois calcular as respectivas médias?

Felizmente não é necessário estar com tanto trabalho, graças a um dos resultados mais importantes das Probabilidades, conhecido como o Teorema do Limite Central e que nos fornece um modelo matemático para a distribuição de amostragem da Média:

Teorema do Limite Central – Suponhamos que se recolhe uma amostra de dimensão n de uma população X, com valor médio μ e desvio padrão σ. A recolha da amostra deve ter em consideração o seguinte:

- ►Se a população for finita a recolha é feita com reposição;
- ►No caso de a população ter uma dimensão "suficientemente grande", a selecção da amostra pode ser feita sem reposição.

Então, se a dimensão da amostra for suficientemente grande ($n \geq 30$), a distribuição de amostragem da Média pode ser aproximada por uma distribuição Normal. Esta aproximação não depende da forma da distribuição da população.

Outras características da distribuição de amostragem da Média:

Se a população tiver valor médio μ e desvio padrão σ, então a distribuição de amostragem da Média, para amostras de qualquer dimensão n, mas recolhidas nas condições indicadas no enunciado do TLC, tem valor médio μ e desvio padrão $\sigma/\sqrt{n}$. Estas propriedades derivam do facto de a Média $\bar{X}$ ser uma soma (ponderada) de variáveis aleatórias independentes e identicamente distribuídas, e das propriedades do valor médio e da variância, nomedamente:

Se X e Y forem variáveis aleatórias e a e b constantes

- Valor médio (a×X)=a×Valor médio (X)
- Valor médio (X+Y) = valor médio (X) + Valor médio (Y)

Se X e Y forem independentes

- Variância (a×X)= a^2×Variância (X)
- Variância (X+Y) = Variância (X) + variância (Y)

Então, resumindo o que dissémos anteriormente sobre a **distribuição de amostragem da Média** $\bar{X}$, obtida a partir de amostras[2] de dimensão n, de uma população de valor médio μ e variância σ^2, podemos concluir o seguinte:

Valor médio($\bar{X}$) = μ

Variância ($\bar{X}$) = $\frac{\sigma^2}{n}$

Se a dimensão da amostra for suficientemente grande (n≥30), a distribuição de amostragem da Média pode ser aproximada por uma distribuição Normal. Esta aproximação não depende da forma da distribuição da população (Consequência do TLC). (Se se souber que a população tem uma distribuição Normal, já não será necessário invocar o TLC para obter a distribuição aproximada, pois neste caso conhece-se a distribuição exacta da Média, que será Normal se a variância for conhecida, ou será uma t-Student, se a variância for desconhecida).

O que é que significa dizer que se tem uma população de dimensão "suficientemente grande" ou infinita?

Na maior parte dos casos em que é necessário recolher uma amostra, para estudar uma característica da população, não se conhece a sua dimensão N. 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 a fracção de amostragem, isto é, o quociente entre a dimensão n da amostra a recolher e a dimensão N da população, é inferior a 5%, ou dito de outra forma, a dimensão da população é superior a 20 vezes a dimensão da amostra:

$$N \geq 20 \times n$$

[2] Recordamos que a amostragem é com reposição. Se a população for infinita, as conclusões ainda são válidas para amostragem sem reposição.

Algumas consequências práticas das propriedades da distribuição de amostragem da Média:

- A Média $\overline{X}$, como estimador do parâmetro valor médio μ, é um estimador centrado, pois

  Valor médio($\overline{X}$) = μ;

- Como já havíamos referido anteriormente, quanto maior for a dimensão da amostra, menor é a variabilidade apresentada pelo estimador Média, pelo que maior será a precisão do estimador, pois

  Variância ($\overline{X}$) = $\frac{\sigma^2}{n}$;

- Se a dimensão n das amostras for suficientemente grande ($n \geq 30$), podemos utilizar a distribuição Normal para calcular quaisquer probabilidades referentes ao estimador Média;

- Se a amostragem for feita com reposição, ou sem reposição no caso de populações "infinitas", as propriedades do estimador Média não dependem da dimensão da população (repare que nas propriedades da distribuição de amostragem da Média, nunca se faz referência à dimensão N da população);

- A precisão do estimador Média depende da variabilidade presente na população. Quando pretendemos estimar o valor médio de uma população, para obter uma determinada precisão (recorda-se que quando menor for a variabilidade apresentada pelo estimador, maior será a precisão) , a dimensão da amostra terá de ser tanto maior, quanto maior for a variabilidade presente na população (basta ter em conta a expressão da variância da Média).

As propriedades do estimador Média dependem da dimensão da População?

Se a dimensão, N, da população **não for suficientemente grande**, e a amostragem for feita sem reposição, pode-se mostrar que essa dimensão **terá interferência** na precisão da Média, como estimador do valor médio. Mais precisamente, pode-se mostrar que para amostras de dimensão n, suficientemente grande ($n \geq 30$), a distribuição de amostragem da Média pode ser aproximada pela distribuição Normal com valor médio μ e variância $\frac{\sigma^2}{n}\frac{N-n}{N-1}$. Esta expressão para a variância da Média é bastante elucidativa, na medida em que permite concluir que se a **dimensão da população for suficientemente grande**, então a variabilidade do estimador só depende da dimensão da amostra e da variabilidade presente na população e não da sua dimensão, como já havíamos referido anteriormente. Neste caso, os esquemas de amostragem com reposição e sem reposição podem-se considerar equivalentes.

6.2 Intervalo de confiança para o parâmetro valor médio

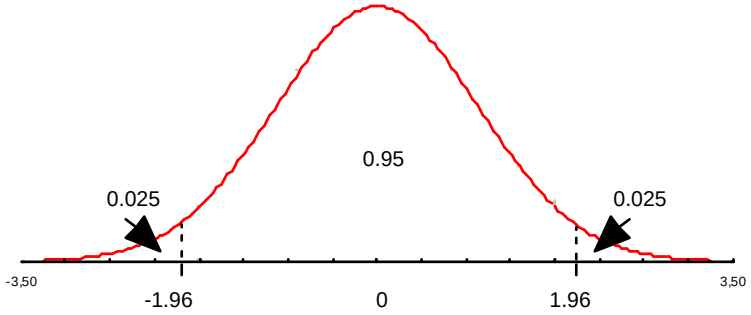
O comportamento da distribuição de amostragem da Média, anteriormente descrito, tem consequências muito importantes, no que diz respeito ao problema da estimação do parâmetro valor médio, já que vamos aproveitá-lo para encarar este problema de um outro ângulo. Em vez de procurarmos um valor – **estimativa 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.

Voltemos ainda a considerar o caso da população X – altura de um indivíduo escolhido ao acaso de entre os 97 indivíduos considerados. Vimos que esta população tinha variância igual a 7.05. Se recolher, com reposição, amostras de dimensão n, igual a 30 ou superior, como espera que seja o comportamento da distribuição de amostragem da Média, para amostras desta dimensão?

De acordo com o Teorema Limite Central, espera-se que a Média tenha uma distribuição de amostragem, que possa ser aproximada por uma Normal.

Então, como se sabe que o valor médio da Média é o valor médio, μ, da população e o desvio padrão da Média é igual a $\sqrt{\frac{7.05^2}{30}}$ = 1.287, quando n=30, podemos, tendo em consideração as propriedades da distribuição Normal, tentar obter o valor de z tal que:

$P(-z \le \frac{\overline{X}-\mu}{1.287} \le z) = 0.95$ ou $P(-z \le Z \le z) = 0.95$ onde $Z=\frac{\overline{X}-\mu}{1.287}$ tem distribuição N(0,1),



O valor de z que satisfaz a condição anterior é 1.96, pelo que a probabilidade anterior se pode escrever

$$P(\overline{X} - 1.96 \times 1.287 \le \mu \le \overline{X} + 1.96 \times 1.287) = .95$$

e o intervalo

$$[\overline{X} - 1.96 \times 1.287 , \overline{X} + 1.96 \times 1.287]$$

diz-se que é **um intervalo de 95% de confiança** para o valor médio μ da Altura, ou Altura média.

Como é que se interpreta esta confiança? O que é que significa?

Consideremos as 50 amostras que recolhemos de dimensão 30 e as respectivas médias. Substituindo essas médias na expressão considerada anteriormente para o intervalo de confiança, obtemos os seguintes intervalos:

	C	D	E	F	G	H
1	lim inf	lim sup	lim inf	lim sup	lim inf	lim sup
2	161,85	166,89	160,38	165,42	**164,61**	**169,65**
3	162,28	167,32	162,78	167,82	162,31	167,35
4	163,58	168,62	160,01	165,05	161,41	166,45
5	161,28	166,32	164,18	169,22	161,48	166,52
6	163,95	168,99	162,18	167,22	161,41	166,45
7	160,28	165,32	162,58	167,62	160,91	165,95
8	162,58	167,62	160,78	165,82	162,91	167,95
9	161,98	167,02	161,21	166,25	162,71	167,75
10	162,58	167,62	161,78	166,82	**165,11**	**170,15**
11	162,48	167,52	162,65	167,69	160,95	165,99
12	163,81	168,85	162,58	167,62	**159,48**	**164,52**
13	163,58	168,62	162,58	167,62	162,65	167,69
14	161,81	166,85	161,48	166,52	162,28	167,32
15	161,95	166,99	162,21	167,25	163,05	168,09
16	161,48	166,52	163,31	168,35	162,28	167,32
17	162,85	167,89	161,48	166,52	161,51	166,55
18	161,31	166,35	162,48	167,52		

Sheet1 Sheet2 Sheet3

Destes 50 intervalos, verifica-se que 47 contêm o valor do parâmetro “Altura média”, que é 164.57, enquanto que 3 – assinalados a escuro , não o contêm. Quando falamos em 95% de confiança, significa que se considerássemos 100 intervalos, esperaríamos que aproximadamente 95 contivessem o valor do parâmetro e 5 não o contivessem.

Como ao fazer um estudo sobre um parâmetro desconhecido, só se recolhe uma amostra, temos *confiança* que a que recolhemos seja uma das “boas”, que vai dar origem a um intervalo que contenha o valor desse parâmetro.

Se mudarmos a probabilidade de 0.95 para 0.90, por exemplo, então em vez de z=1.96, devemos considerar z=1.645. Assim, um intervalo de confiança, com 90% de confiança terá o seguinte aspecto

$$[\bar{x} - 1.645 \times 1.287,\ \bar{x} + 1.645 \times 1.287]$$

De forma análoga ao que fizémos anteriormente, vamos substituir as 50 médias na expressão anterior. Os intervalos obtidos são os seguintes:

	I	J	K	L	M	N
1	lim inf	lim sup	lim inf	lim sup	lim inf	lim sup
2	162,25	166,49	160,78	165,02	**165,01**	**169,25**
3	162,28	167,32	163,18	167,42	162,71	166,95
4	163,58	168,62	160,41	164,65	161,81	166,05
5	161,28	166,32	**164,58**	**168,82**	161,88	166,12
6	163,95	168,99	162,58	166,82	161,81	166,05
7	160,28	165,32	162,98	167,22	161,31	165,55
8	162,58	167,62	161,18	165,42	163,31	167,55
9	161,98	167,02	161,61	165,85	163,11	167,35
10	162,58	167,62	162,18	166,42	**165,51**	**169,75**
11	162,48	167,52	163,05	167,29	161,35	165,59
12	163,81	168,85	162,98	167,22	**159,88**	**164,12**
13	163,58	168,62	162,98	167,22	163,05	167,29
14	161,81	166,85	161,88	166,12	162,68	166,92
15	161,95	166,99	162,61	166,85	163,45	167,69
16	161,48	166,52	163,71	167,95	162,68	166,92
17	162,85	167,89	161,88	166,12	161,91	166,15
18	161,31	166,35	162,88	167,12		

Sheet1 Sheet2 Sheet3

Ao diminuirmos a confiança, aumentámos o número de intervalos que não contêm o parâmetro a estimar (assinalados a preto) – aumentou assim a possibilidade de o intervalo que calcularmos, com a amostra que recolhermos, não conter o parâmetro a estimar.

E o que acontece se aumentarmos a confiança para 99%?

Neste caso o valor de z=2.576 e os intervalos que se obtêm substituindo as médias na expressão

$$[\bar{x} - 2.576 \times 1.287,\ \bar{x} + 2.576 \times 1.287]$$

apresentam-se a seguir:

	I	J	K	L	M	N
1	lim inf	lim sup	lim inf	lim sup	lim inf	lim sup
2	161,05	167,69	159,58	166,22	163,81	170,45
3	161,48	168,12	161,98	168,62	161,51	168,15
4	162,78	169,42	159,21	165,85	160,61	167,25
5	160,48	167,12	163,38	170,02	160,68	167,32
6	163,15	169,79	161,38	168,02	160,61	167,25
7	159,48	166,12	161,78	168,42	160,11	166,75
8	161,78	168,42	159,98	166,62	162,11	168,75
9	161,18	167,82	160,41	167,05	161,91	168,55
10	161,78	168,42	160,98	167,62	164,31	170,95
11	161,68	168,32	161,85	168,49	160,15	166,79
12	163,01	169,65	161,78	168,42	158,68	165,32
13	162,78	169,42	161,78	168,42	161,85	168,49
14	161,01	167,65	160,68	167,32	161,48	168,12
15	161,15	167,79	161,41	168,05	162,25	168,89
16	160,68	167,32	162,51	169,15	161,48	168,12
17	162,05	168,69	160,68	167,32	160,71	167,35
18	160,51	167,15	161,68	168,32		

Sheet1 Sheet2 Sheet

Neste caso, já todos os intervalos contêm o valor do parâmetro a estimar.

Repare-se que ao aumentar a confiança, estamos a aumentar a amplitude do intervalo de confiança, o que, se por um lado é bom, já que aumenta a nossa confiança em que um qualquer intervalo que se construa, contenha o valor do parâmetro que estamos a estimar, por outro lado não é muito bom, pois um intervalo com uma grande amplitude não nos serve para nada!

Então, o que fazer, para termos uma confiança razoável, mas ao mesmo tempo um intervalo com pequena amplitude?

A solução é aumentar a dimensão da amostra, como se verifica imediatamente a partir da expressão genérica de um intervalo de confiança, que apresentaremos a seguir.

Dada uma população com desvio padrão σ, a forma geral do **intervalo de confiança** para o valor médio μ será, tendo em conta as propriedades da Normal

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

onde o valor de **z** dependerá da confiança com que se pretende construir o intervalo. Se o desvio padrão σ da população for desconhecido, utiliza-se o desvio padrão amostral S, para o estimar.

Alguns valores (obtidos a partir da tabela 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

6.2.1 Margem de erro

A metade da amplitude do intervalo de confiança, chama-se margem de erro. Como, de um modo geral, o que se pretende é obter um intervalo de confiança com pequena margem de erro, por exemplo *e*, se pretendermos uma determinada confiança, por exemplo 95%, temos que recolher uma amostra de dimensão

$$n=(\frac{1.96\times\sigma}{e})^2.$$

Então, respondendo à questão

"*O que fazer, para termos uma confiança razoável, mas ao mesmo tempo um intervalo com pequena amplitude?*"

diremos que o que temos a fazer é recolher uma amostra de dimensão suficientemente grande, de forma a satisfazer a precisão exigida.

Repare-se que, para obtermos uma determinada precisão **e**, **quanto maior for a variabilidade** presente na população, **maior terá de ser a dimensão da amostra** a recolher. Recorde-se que já havíamos referido esta propriedade na página 17.

7 Estimação do parâmetro proporção populacional

7.1 Distribuição de amostragem da proporção amostral, como estimador da proporção populacional

Suponhamos que estamos a estudar uma População quanto à presença ou ausência de uma determinada propriedade ou característica, em cada indivíduo dessa População. Admitimos que essa propriedade se verifica na População com uma probabilidade **p** (normalmente desconhecida). Se ao observar o indivíduo verificarmos que tem a propriedade, anotamos um 1, enquanto que se verificarmos que não tem a propriedade anotamos um 0. Então podemos representar a População, quanto a essa propriedade por uma variável X, que pode assumir o valor 1 ou 0, respectivamente com probabilidade p (probabilidade de ter a propriedade) ou (1-p) (probabilidade de não ter a propriedade).

Será que podemos interpretar o parâmetro **p** como um valor médio?

Assim é, de facto, pois **p** é a frequência relativa com que o 1 se verifica na População relativamente à propriedade em estudo, e não é mais do que a **média** do conjunto constituído pelos 0's e 1's.

Analogamente quando recolhemos uma amostra, constituída por 1's e 0's conforme os elementos observados tenham ou não tenham a propriedade, a média desta amostra dá-nos a proporção (amostral) de 1's, ou seja, uma **estimativa pontual** para a ***proporção*** (populacional) ou probabilidade com que a propriedade em estudo se verifica na População.

Do que acabamos de referir, depreende-se que o estudo do parâmetro **p** "proporção de indivíduos da população que verificam determinada propriedade" se reduz ao estudo do parâmetro "valor médio de uma população representada

por 1's e 0's, conforme a propriedade está ou não presente nos indivíduos da população".

Assim, não temos mais que transportar para o estimador proporção amostral, as propriedades verificadas para o estimador Média. Contudo, como veremos a seguir, algumas simplificações serão introduzidas, devido à particularidade da população em estudo ser tão simples, isto é, constituída por 0's e 1´s.

Características da população X

X	0	1
Probabilidade	(1-p)	p

Valor médio(X) = p

Variância(X) = p(1-p)

Como resultado das observações anteriores podemos enunciar o seguinte resultado, para a distribuição de amostragem da **proporção amostral** $\hat{p}$:

Suponhamos que se recolhe uma amostra de dimensão n, com reposição (ou sem reposição se a população for muito grande) de uma população X, em que cada elemento da população tem, ou não, uma determinada propriedade. Seja **p** a proporção de elementos da população com essa propriedade. Então, se a dimensão da amostra for suficientemente grande (n≥30), a distribuição de amostragem da proporção $\hat{\mathbf{p}}$ pode ser aproximada por uma distribuição Normal com valor médio **p** e desvio padrão $\sqrt{\mathbf{p(1-p)}}/\sqrt{\mathbf{n}}$.

7.2 Intervalo de confiança para a proporção populacional p

Já que a proporção populacional **p** é um valor médio e a proporção amostral $\hat{p}$ é uma média, a expressão para o intervalo de confiança da proporção p deduz-se da que se obteve para o intervalo de confiança para o valor médio μ, fazendo as modificações adequadas:

	Onde está	Considera-se
	σ	$\sqrt{p(1-p)}$
ou	s	$\sqrt{\hat{p}(1-\hat{p})}$

Como o valor de p é desconhecido, a expressão para o intervalo de confiança, com uma confiança de 95% vem

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

Dada uma população, em que **p** é a proporção de elementos da população com determinada característica, a forma geral do intervalo de confiança para **p**, a partir de amostras de dimensão n, é

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

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

Alguns valores (obtidos a partir da tabela 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

Exercícios

Exercício 1. 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:

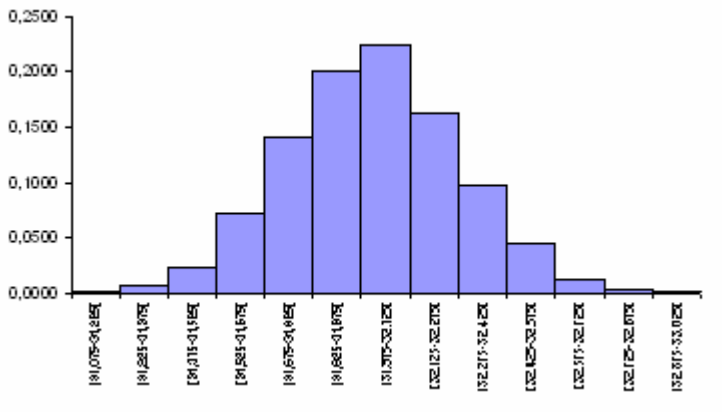
i) Obtenha um intervalo, com uma confiança de 95%, para p;

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%)).

Exercício 2. 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[	0,0015



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) Utlizando 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.

Exercício 3. 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?

Exercício 4. 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?

Exercício 5. No dia 9 de Outubro de 2005 realizar-se-ão as Eleições Autárquicas. Relativamente à cidade de Lisboa, há dois candidatos sobre os quais se criaram mais expectativas, nomedamente Carmona Rodrigues e Manuel Maria Carrilho . 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, que passaremos a representar por X, apresentaram o seguinte resultado: - *O candidato X 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 X?
3. Acha razoável admitir que o candidato X, 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 X era: - *O candidato X 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 X, na forma de um intervalo de confiança, ou na forma de um valor? Explique porquê.

Exercício 6. 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?

Exercício 7. 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 resultado 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.

AtenciosamenteO 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?