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

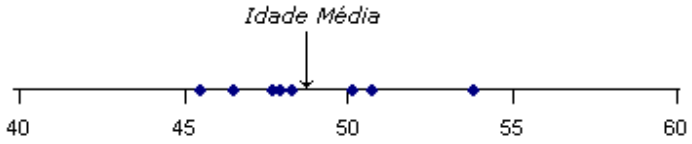
a estimativa que se obtém para a amostra que se recolhe (na prática só recolhemos uma amostra!) daria um valor aproximado do parâmetro.

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:

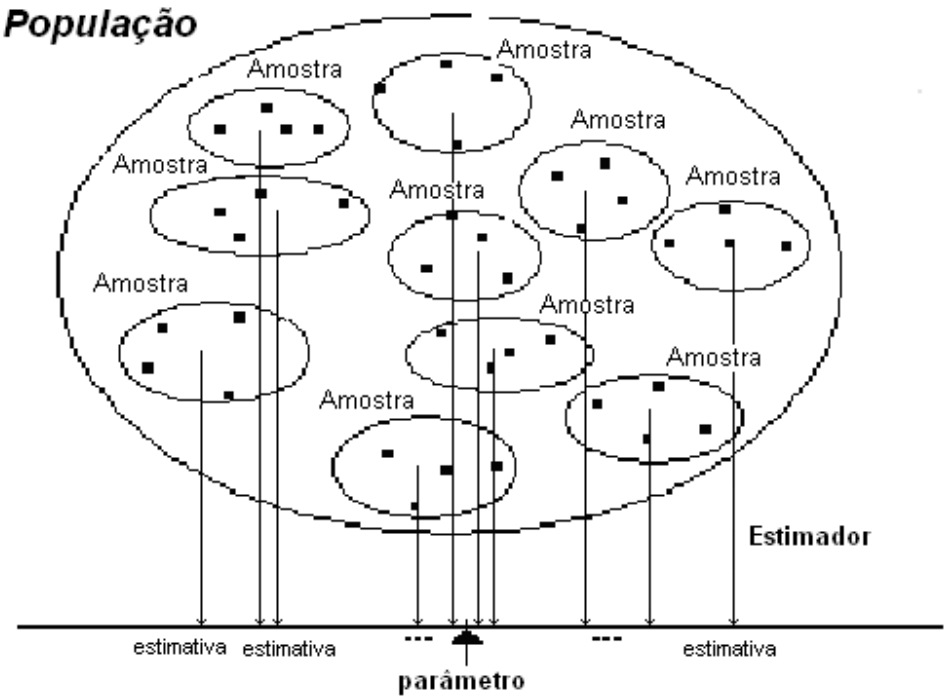
- As médias obtidas distribuem-se para um e outro lado do parâmetro e
- A variabilidade apresentada pelas estimativas é relativamente pequena, isto é, as diferentes estimativas estão próximas do parâmetro a estimar.

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.

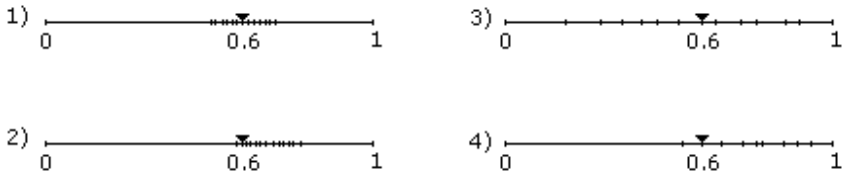
A **falta de precisão** e o problema do **enviesamento da amostra** são dois tipos de erro com que nos defrontamos num processo de amostragem (mesmo que tenhamos escolhido um "bom" estimador). Não se devem, contudo, confundir. Enquanto o enviesamento se manifesta por um desvio nos valores da estatística,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



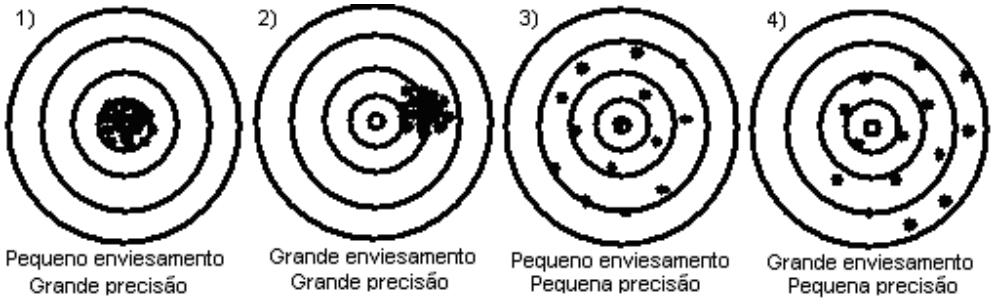
Se admitirmos que o valor do parâmetro é 0.60, então a situação:

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

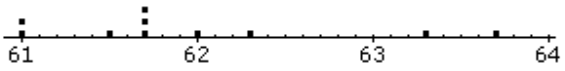
A média das alturas observadas é **61.7 cm**, que é uma estimativa da altura média da População.

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.

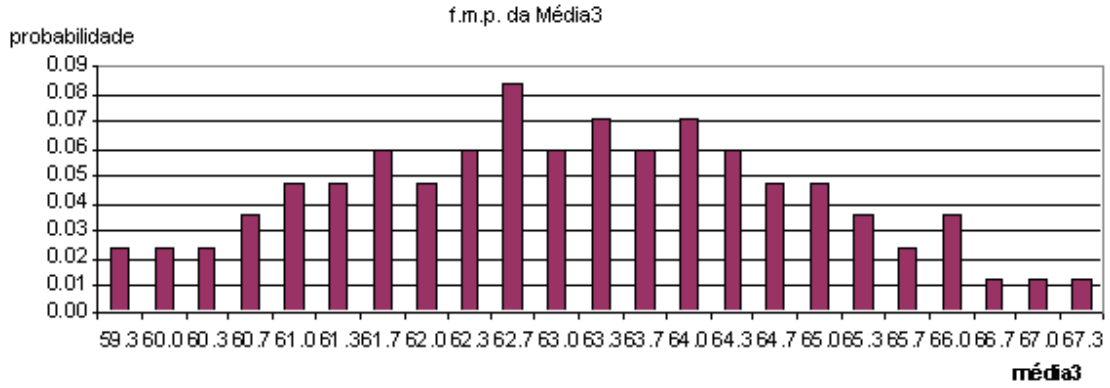
Considerando então todas as amostras aleatórias simples, diferentes, de dimensão 3, obtemos:

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).

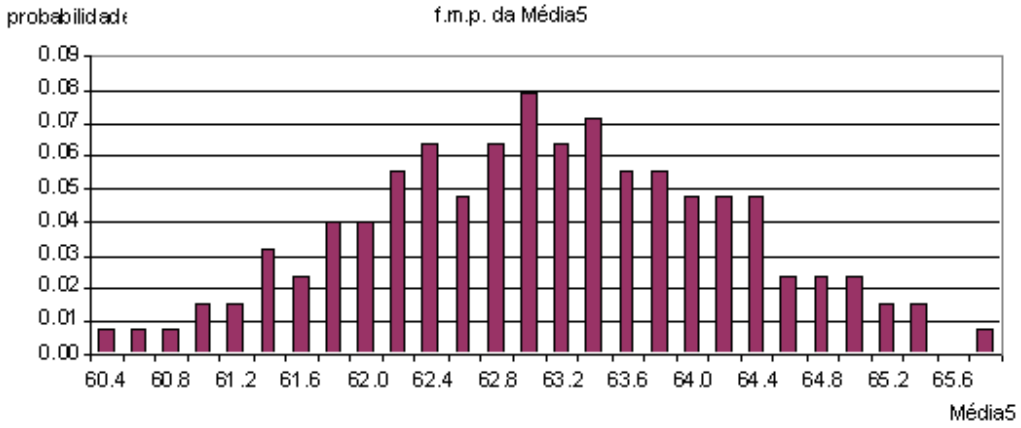
Se utilizarmos amostras de maior dimensão o que é que ganhamos?

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:

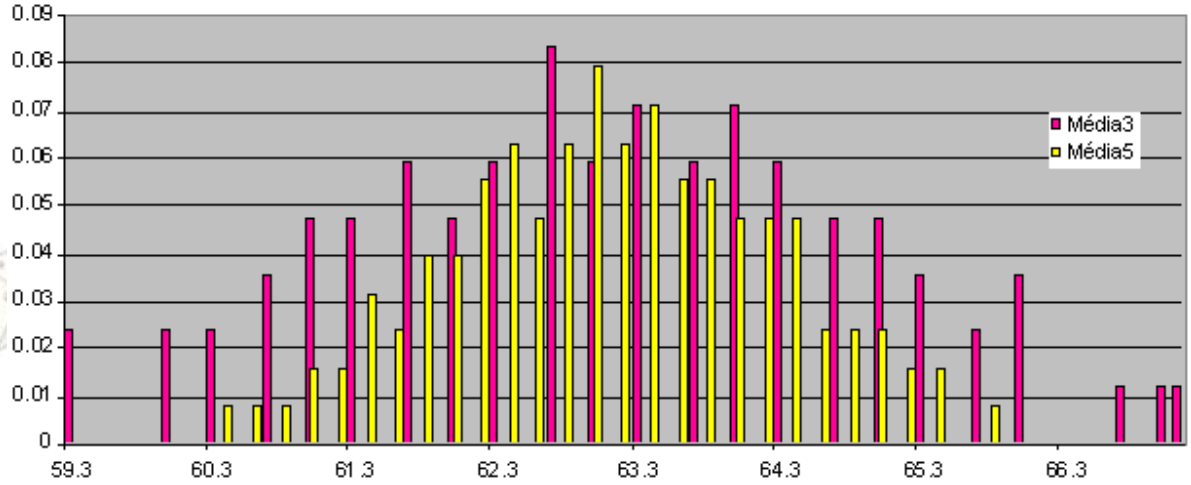
	Valor médio	Desvio padrão	Mínimo	Máximo	Mediana
Média5	63.11	1.13	60.4	65.8	63.1

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** = $\sqrt{2}$.

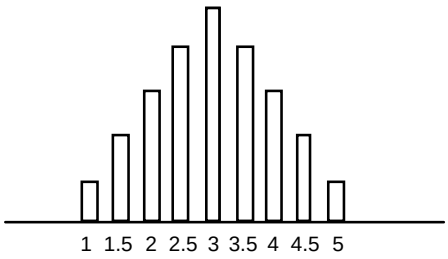
[2] Propriedade do valor médio, segundo a qual o valor médio de uma soma de variáveis aleatórias é igual à soma dos valores médios das parcelas.

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:

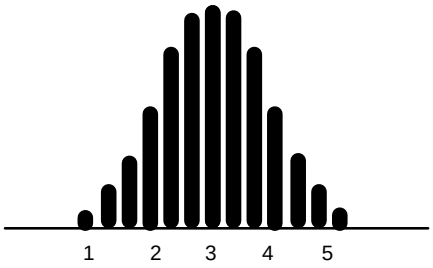
Valor médio = 3 e **Desvio padrão = 1**

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

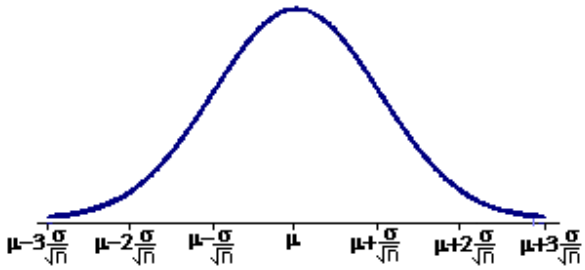
seja, quando a fracção de amostragem for menor que 5%. Então, na prática, temos o seguinte resultado:

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.

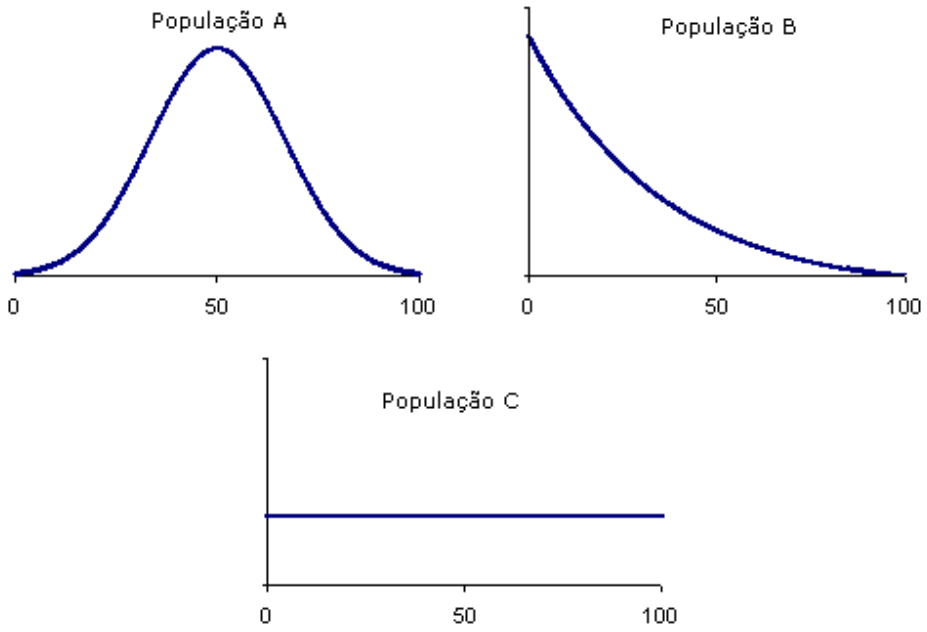
a) O valor de 540 euros é um parâmetro ou uma estatística?
b) Identifique a população em estudo e o parâmetro de interesse.
c) O facto de ter obtido 540 euros para a média, significa que o gasto médio da população nas compras de Natal tenha de ser necessariamente 540 euros?
d) Acharia "razoável" ter 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

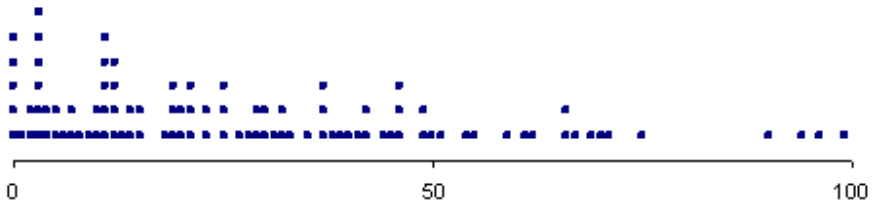
a) Calcule a média dos pesos obtidos. Esse valor é um parâmetro ou uma estatística?
b) O facto de não ter obtido um valor para a média igual a 150 gramas, significa que o peso médio dos pacotes não possa ser de 150 gramas? Justifique a sua resposta.
c) Calcule o desvio padrão (amostral) dos pesos obtidos. 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:

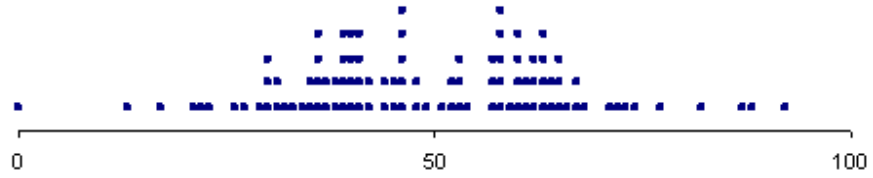


Os seguintes diagramas de pontos representam a distribuição dos valores de 3 amostras de dimensão 100, cada uma extraída de uma das 3 populações anteriores:

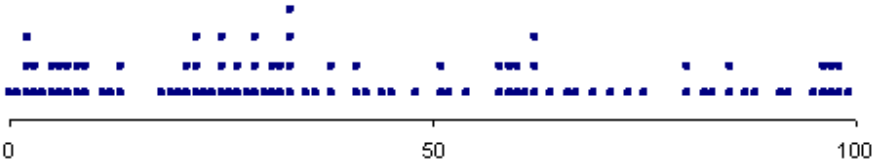
Amostra 1



Amostra2



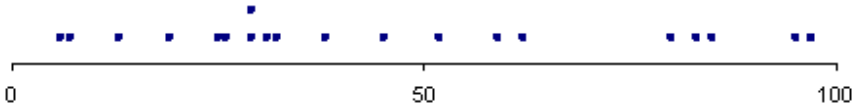
Amostra3



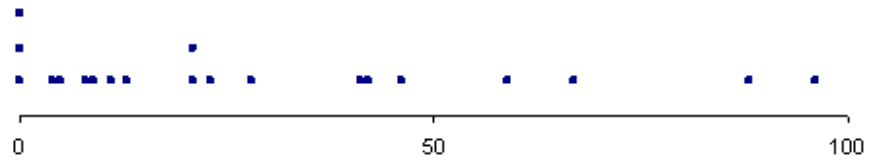
a) Identifique a população de onde foi seleccionada cada uma das amostras anteriores.

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?

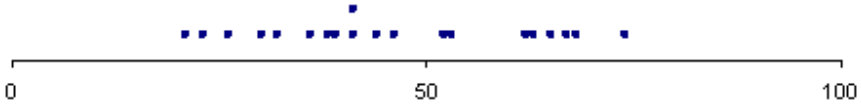
Amostra 1



Amostra 2



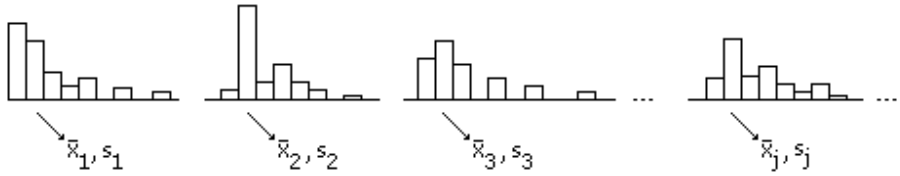
Amostra 3



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:

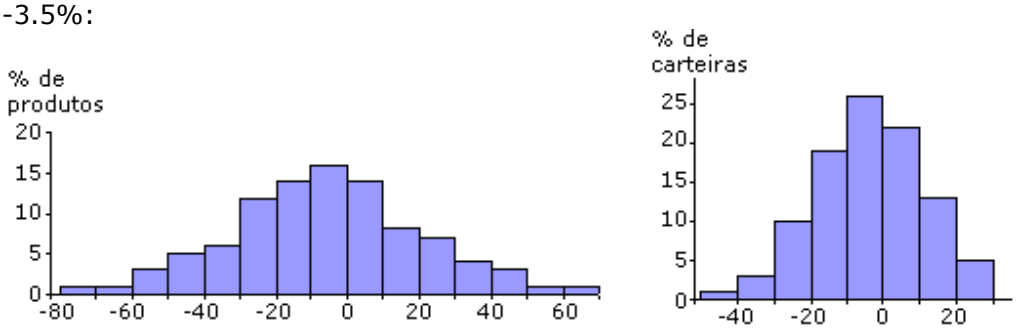


a) Faça um esquema de um histograma que represente adequadamente todas as médias que obteve.
b) Qual o modelo que pode utilizar para modelar o histograma que obteve na alínea anterior?
   i) Teve de fazer algumas hipóteses para utilizar o modelo proposto?
   ii) Quais os parâmetros valor médio e desvio 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:

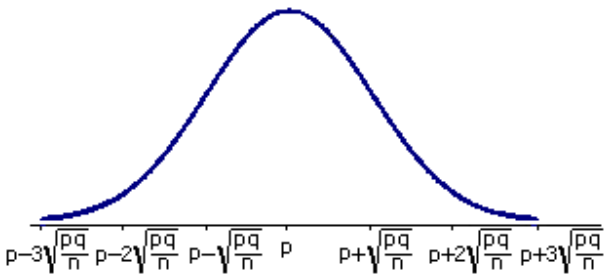
- O estimador $\hat{p}$ de p é um estimador centrado, já que o seu valor médio coincide com p, $E(\hat{p}) = p$
- $Var(\hat{p}) = \frac{\mathbf{p(1-p)}}{\mathbf{n}}(\frac{\mathbf{N-n}}{\mathbf{N-1}})$ no esquema de <u>amostragem aleatória simples</u>

  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:

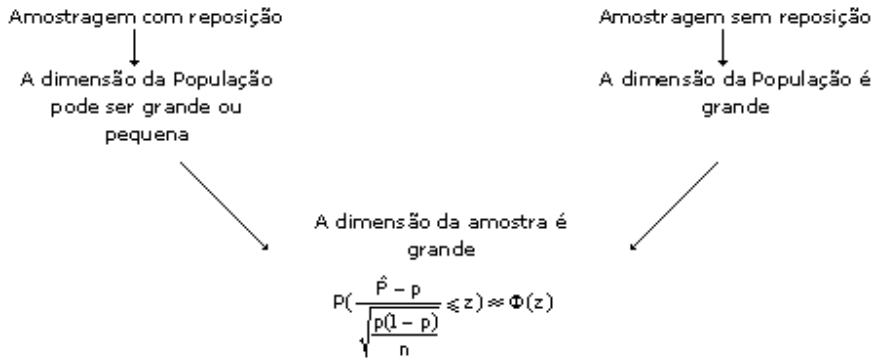
a) Qualquer que seja o processo de selecção da amostra aleatória, isto é, com reposição ou sem reposição, a proporção amostral $\hat{p}$ é sempre um estimador centrado da proporção populacional **p**, isto é, o valor médio da sua distribuição de amostragem é **p**.

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.

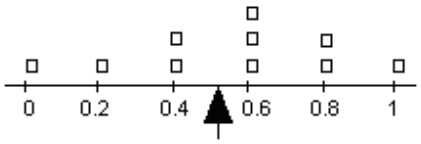
a) O valor 25% é um parâmetro ou uma estatística?
b) Obtenha uma estimativa para a probabilidade e em 50 alunos, 30 ou mais terem computador.

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.

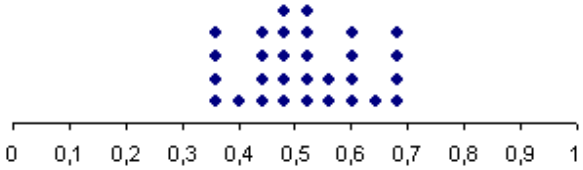
a) Recolha uma amostra de 25 pastilhas e registe o número e a proporção de pastilhas de cada uma das cores:

	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.

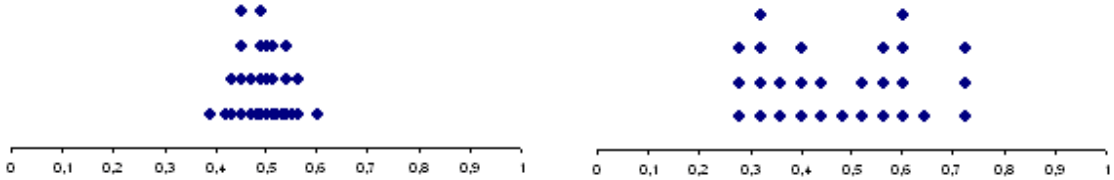
f) Suspeita que todos os seus 29 colegas obtiveram a mesma proporção de pastilhas amarelas, na amostra de 25 pastilhas que cada um recolheu?
g) Admita que no seguinte diagrama de pontos se apresentam as proporções de pastilhas amarelas obtidas por si e pelos seus colegas (também em amostras de dimensão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

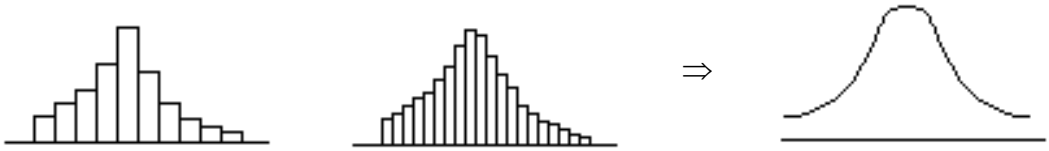
e) Represente por p a proporção de votos recebidos por um certo candidato numa eleição. Admita que selecciona repetidamente amostras de dimensão 100 e calcula a proporção de eleitores que votariam nesse candidato.
   i) Calcule os desvios padrões das distribuições de amostragem das proporções amostrais 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.

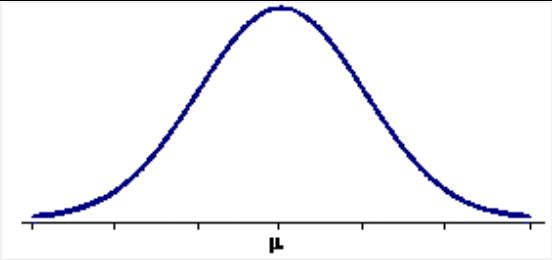
A **distribuição Normal**, das distribuições contínuas, a mais conhecida, foi obtida matematicamente por Gauss, como a distribuição dos erros de medidas, tendo-lhe dado o nome sugestivo de "lei normal dos erros". A partir daí, astrónomos, físicos e mais tarde, cientistas de outros campos, que manipulavam 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

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

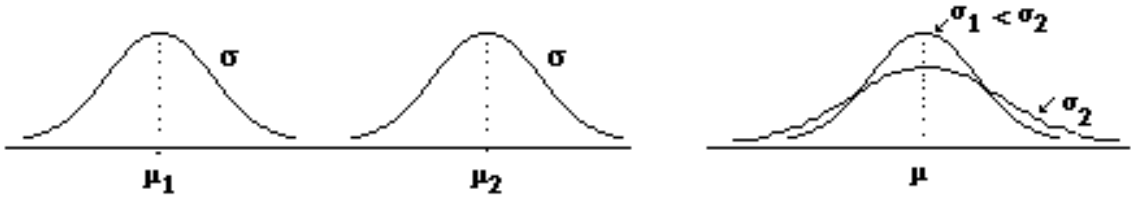
Pode-se mostrar que:

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

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.

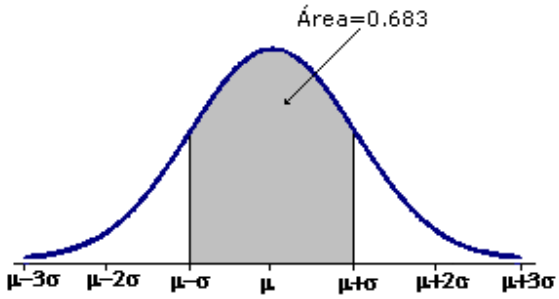


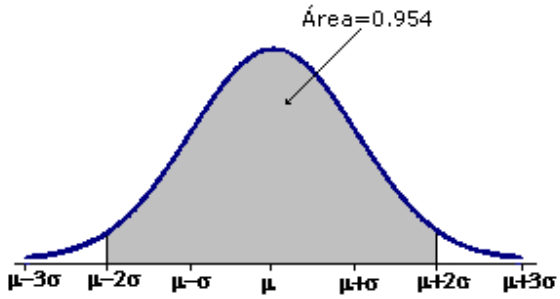
Para dar uma ideia da **concentração** da distribuição normal, em torno do seu valor médio, apresentamos seguidamente algumas probabilidades:

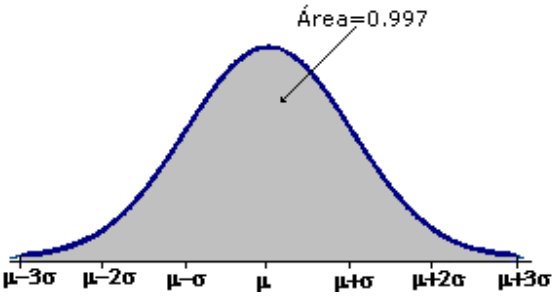
$$P(\mu - \sigma \le X \le \mu + \sigma) = .683$$

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

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







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

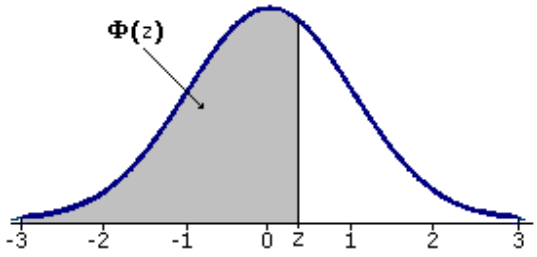
$$\mathbf{Z \cap N(0,1)}$$

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

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

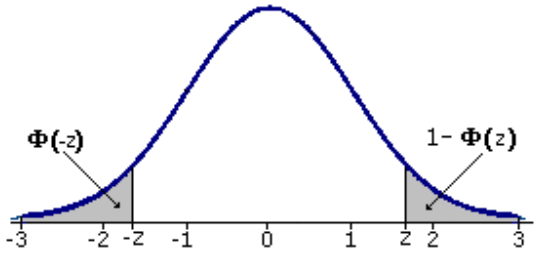
A função distribuição da normal reduzida, tem uma notação especial. Assim, se Z for uma v.a. normal reduzida, representamos

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



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

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

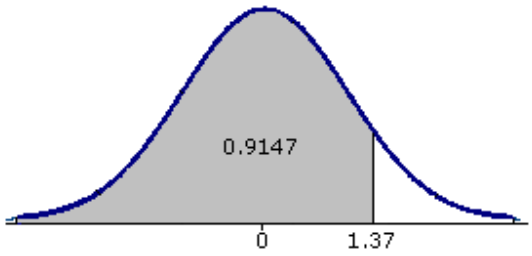


Hoje em dia o cálculo das probabilidades correspondentes à distribuição Normal não oferece qualquer dificuldade, pois pode ser feito com as máquinas de calcular ou a folha de Excel do computador. Até há bem pouco tempo, só dispunhamos de tabelas extensivas da função distribuição da normal standard, que 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

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

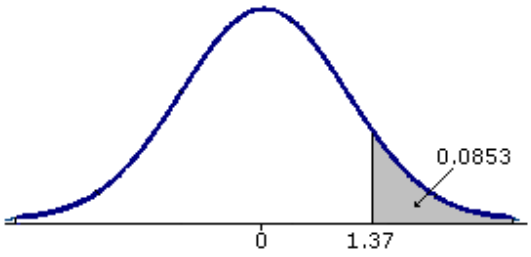
$= .9147$



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

$= 1 - .9147$

$= .0853$

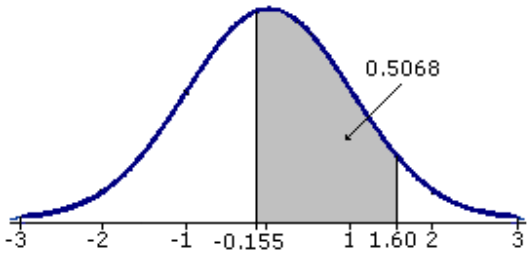


$P(-.155 < Z < 1.60) = \Phi(1.60) - \Phi(-.155)$

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

$= .9452 - 1 + .5616$

$= .5068$

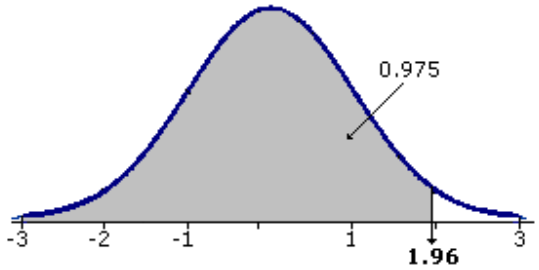


Determinar o valor de z, tal que

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

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

$= 1.96$

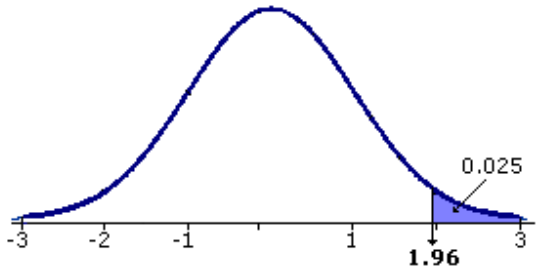


Determinar o valor de z tal que

$P(Z > z) = .025$

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

$= 1.96$



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:

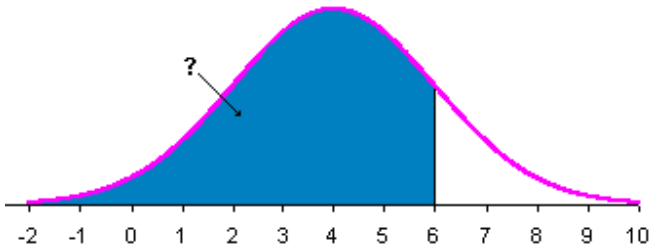
$$\mathbf{X \cap N(\mu, \sigma)} \quad \Leftrightarrow \quad \mathbf{Z = \frac{X - \mu}{\sigma} \cap N(0,1)}$$

donde:

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

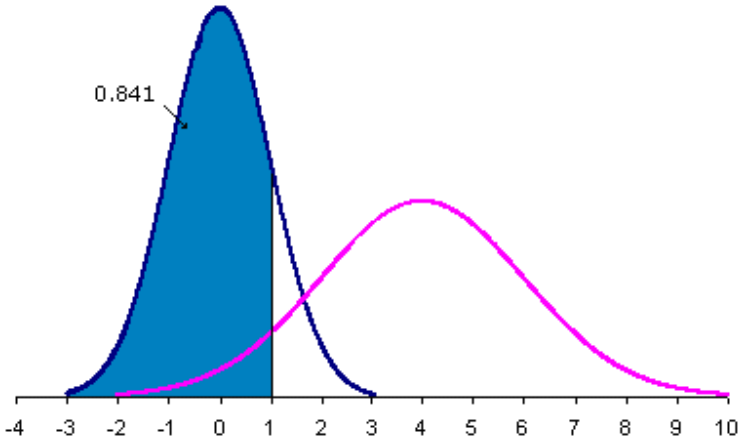
Se X∩N(4,2) calcular P(X≤6)

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



$= \Phi(1)$

$= 0.841$



Se X∩N(μ, σ) calcular

P(μ-1.96σ≤X≤μ+1.96σ)

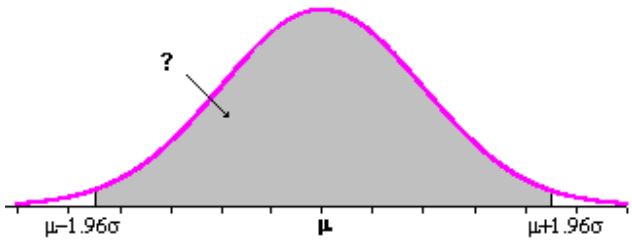
P((μ-1.96σ ≤X≤μ+1.96σ) =

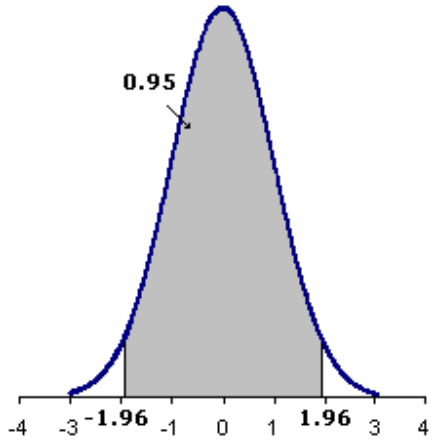
$P(\frac{\mu - 1.96\sigma - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{\mu + 1.96\sigma - \mu}{\sigma})$

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

$= 0.975 - 0.025$

$= .0.95$





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.

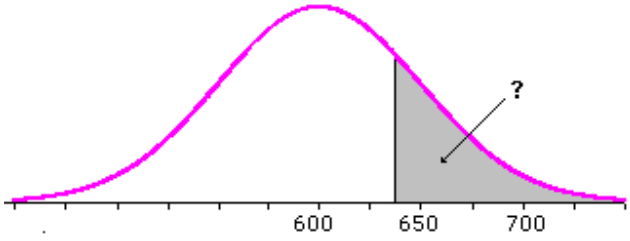
b) Qual deveria ser o stock, de modo que a probabilidade de ruptura fosse de .01?

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)

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

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

$s = 693kg$