

Agenda

Aula 2



01

Distribuições

Contínuas

02

Estimação e Teoria

do Limite Central

03

Amostragem

04

Testes de Hipóteses

e Significância

Sessão Síncrona nº 2 – 14 de Setembro de 2021 (21h00-23h00)

Distribuições Contínuas Univariadas

- Distribuição Uniforme
- Distribuição Exponencial
- Distribuição normal
- Distribuição Qui-Quadrado
- Distribuição t-student

Estimação e Testes de Hipóteses

- Tipos de Amostragem
- Estimação Pontual
- Intervalos de Confiança
- Testes de Hipóteses



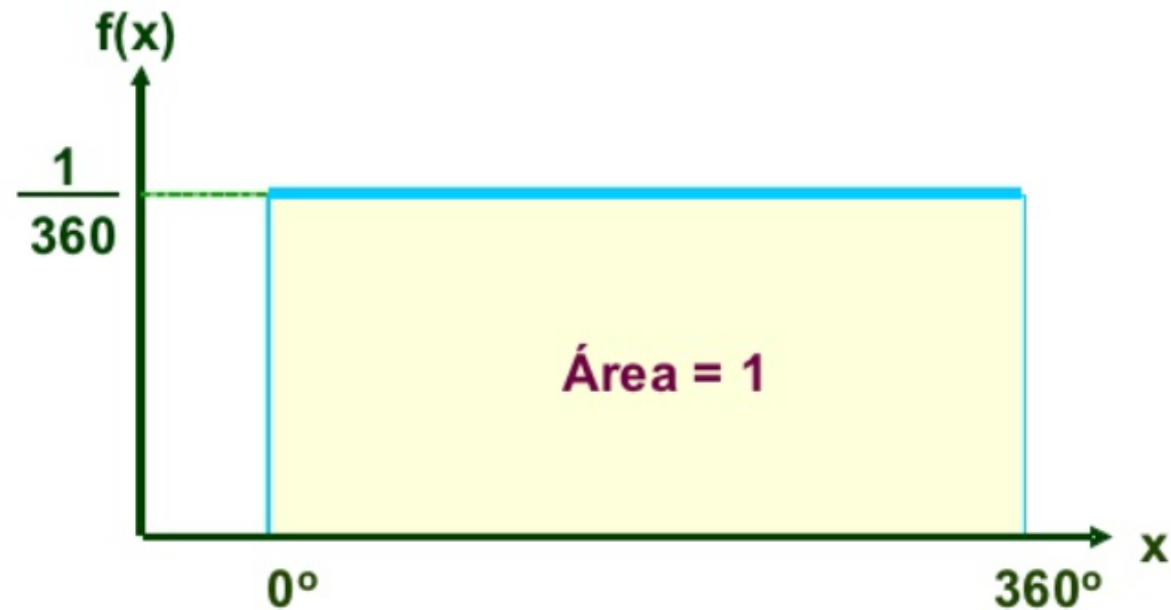
01

Distribuições Contínuas



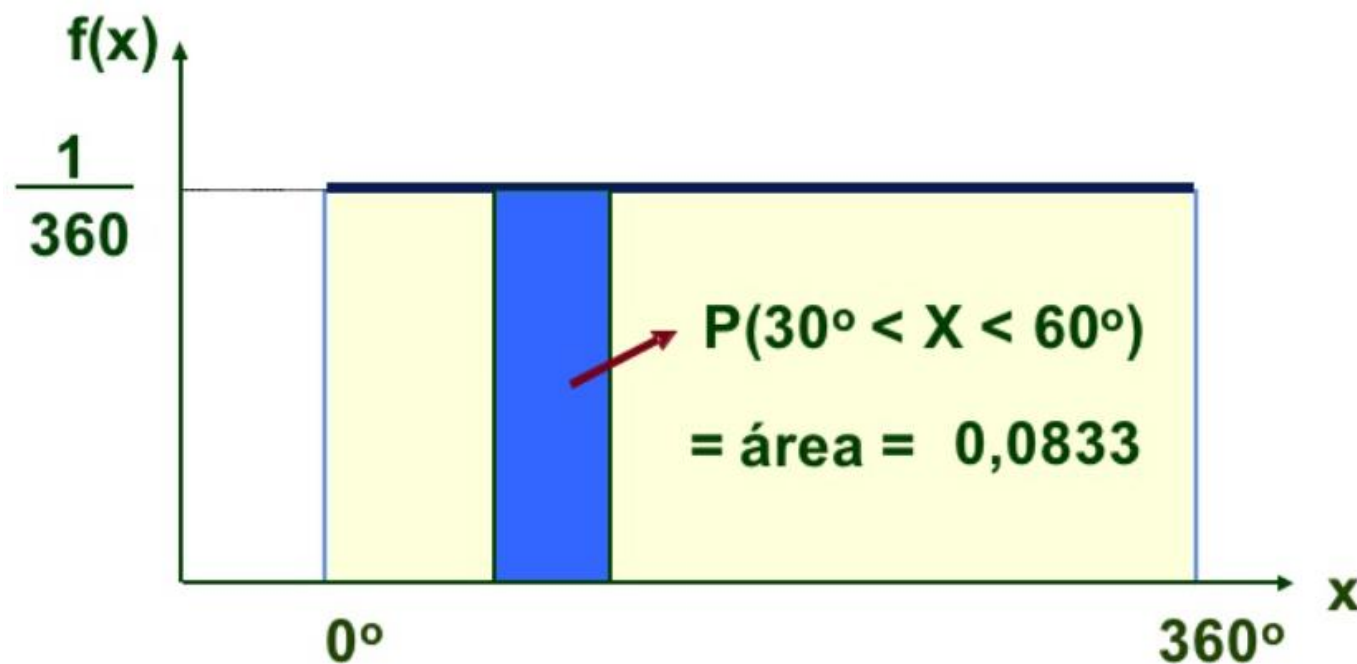
Distribuição Uniforme

X - Variável aleatória que indica o ângulo formado



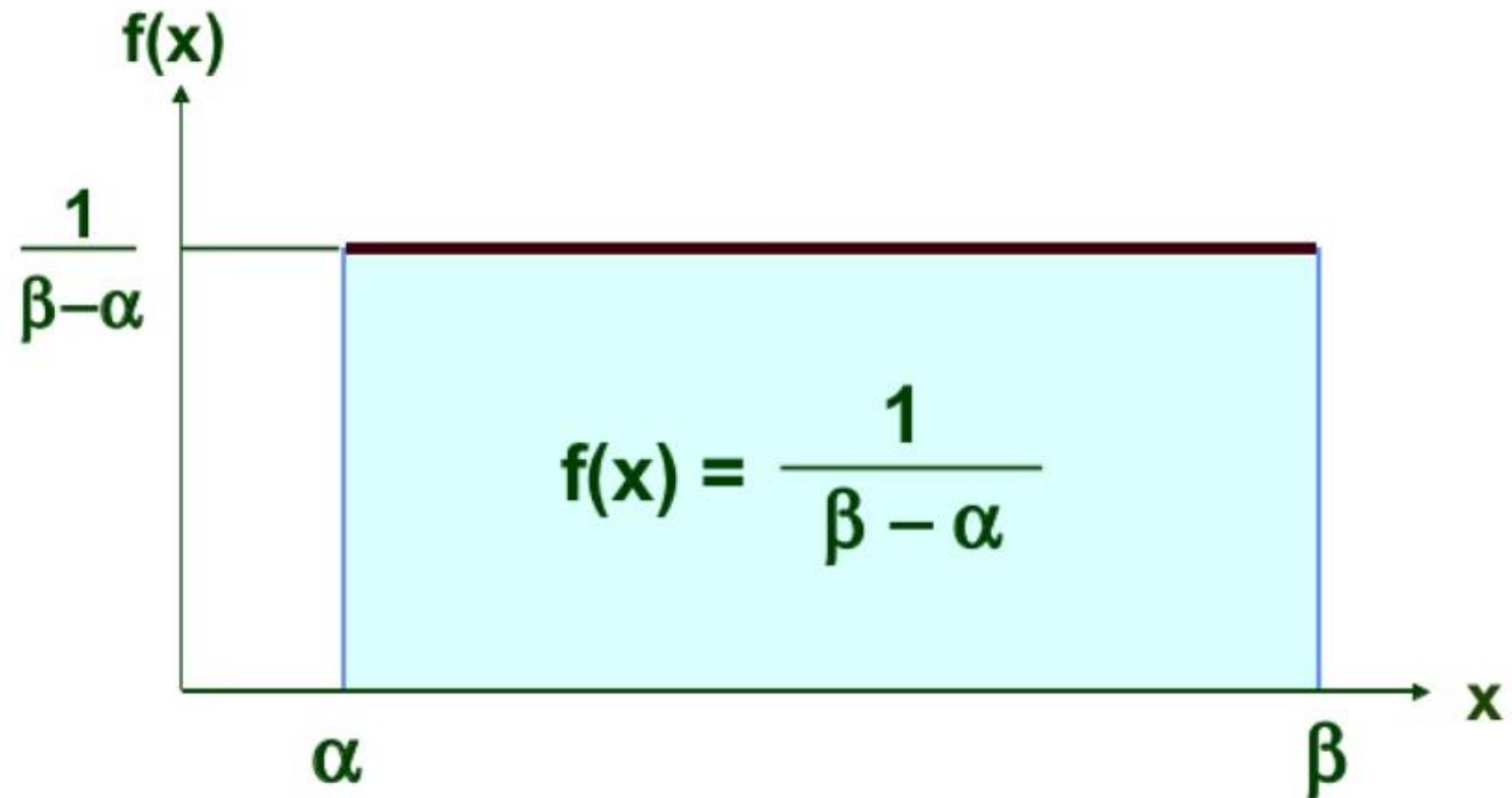
Distribuição Uniforme

Qual é a probabilidade de obter um ângulo entre 30 e 60 graus?

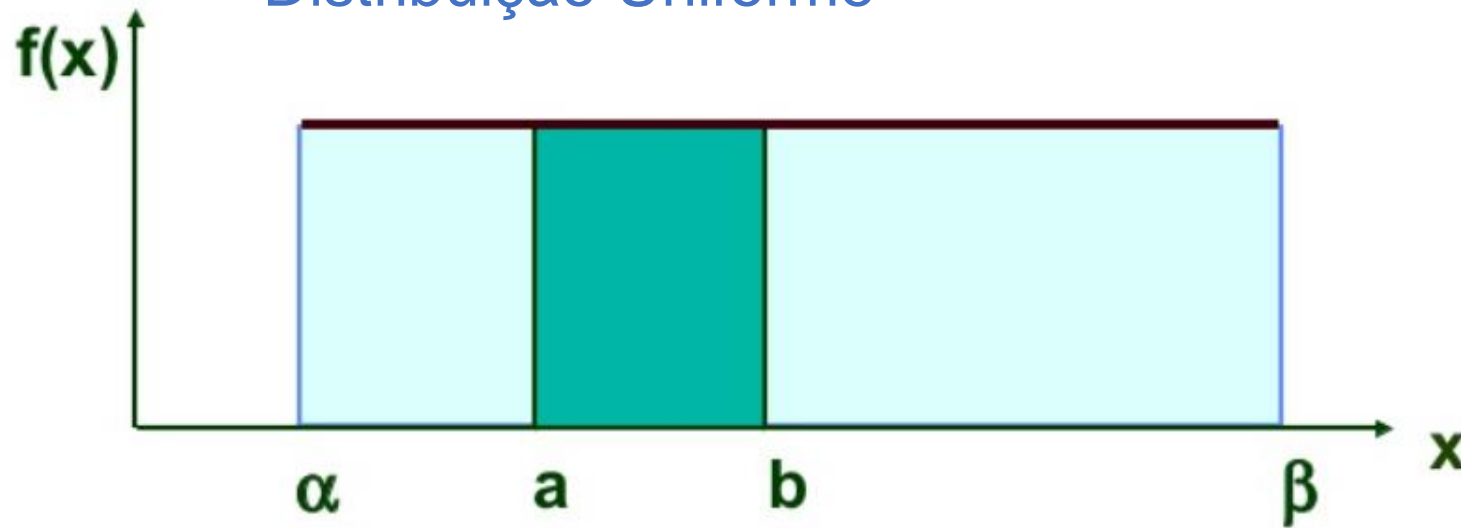


Distribuição Uniforme

Distribuição Uniforme



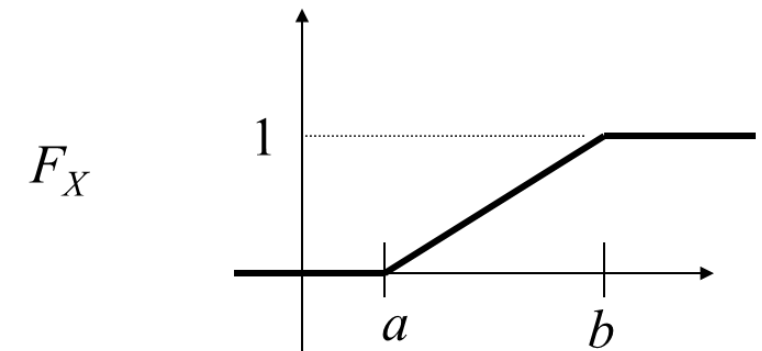
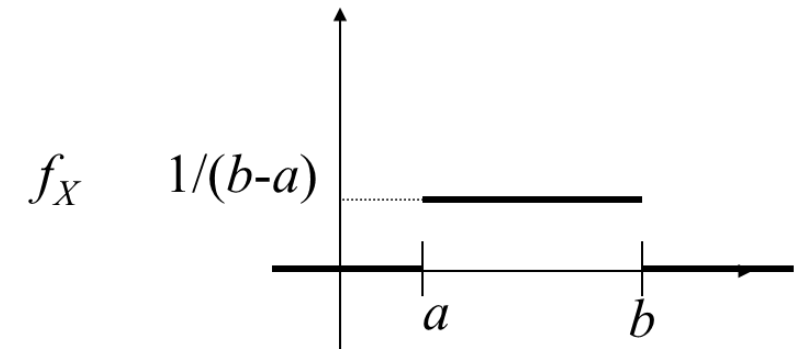
Distribuição Uniforme



$$P(a < X < b) = \frac{b - a}{\beta - \alpha}$$

$$E(X) = \frac{\alpha + \beta}{2}$$

$$Var(X) = \frac{(\beta - \alpha)^2}{12}$$



Distribuição Uniforme

Um ponto é escolhido ao acaso no segmento de reta $[0,2]$. Qual será a probabilidade de que o ponto escolhido esteja entre 1 e $3/2$?

$$f(x) = \frac{1}{b-a} = \frac{1}{2-0} = \frac{1}{2} \quad , \text{ para } 0 \leq x \leq 2$$

$$P(1 \leq x \leq \frac{3}{2}) = \int_1^{1,5} \frac{1}{2} dx = \frac{1}{4} \quad \frac{(1,5-1)}{(2-0)}$$

Distribuição Exponencial

É muito útil para descrever o tempo que se leva para completar uma tarefa.

Exemplos:

*tempo entre chegada de pessoas a uma fila,
tempo de vida de material eletrônico,
tempo de atendimento de um pedido de suprimento de materiais,
tempo entre chegadas de arquivos num servidor*

Distribuição Exponencial

Existe uma relação entre a distribuição de Poisson e a Exponencial.

Na distribuição de Poisson, a variável aleatória é definida como o número de ocorrências em determinado período, sendo a média das ocorrências no período definida como λ .

Na distribuição Exponencial a variável aleatória é definida como o tempo entre ocorrências, sendo a média de tempo entre ocorrências de $1/\lambda$.

Por exemplo, se a média de atendimentos no caixa bancário é de $\lambda = 6/\text{min}$, então o tempo médio entre atendimentos é $1/\lambda = 1/6$ de minuto ou 10 segundos.

Distribuição Exponencial

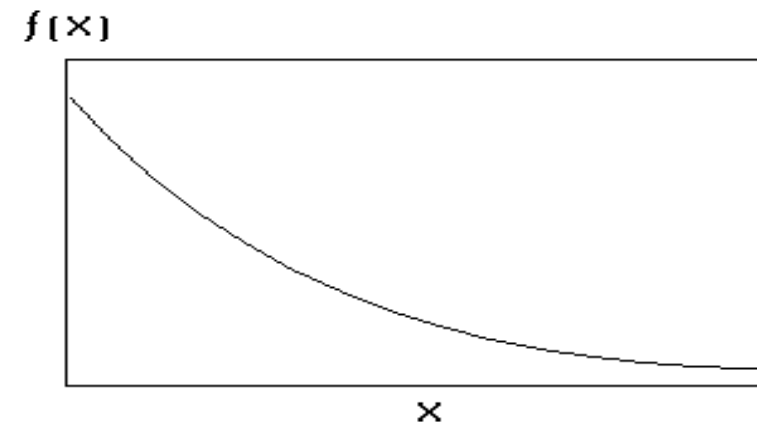
O modelo da distribuição Exponencial é o seguinte:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{se } x \geq 0 \\ 0 & \text{se } x < 0 \end{cases}$$

onde $\lambda > 0$ é uma constante.

A média e o desvio padrão da distribuição exponencial são calculados usando:

$$\mu = \frac{1}{\lambda}$$
$$\sigma = \frac{1}{\lambda^2}$$



Distribuição Exponencial

O cálculo de probabilidade acumulada até determinado ponto é dado por:

$$F(x) = P(X \leq x) = 1 - e^{-\lambda x}, \quad x \geq 0.$$

A distribuição Exponencial é largamente utilizada no campo da confiabilidade, como um modelo para a distribuição dos tempos até à falha de componentes eletrônicos.

Nessas aplicações o parâmetro λ representa a taxa de falha para o componente, e $1/\lambda$ é o tempo médio até ocorrer a falha.

A probabilidade num intervalo (a:b) é dada por:

$$P(a \leq x \leq b) = e^{-\lambda \cdot a} - e^{-\lambda \cdot b}$$

Distribuição Exponencial

Por exemplo, suponha que uma máquina falhe em média uma vez a cada dois anos $\lambda=1/2=0,5$. Calcule a probabilidade da máquina falhar durante o próximo ano.

$$P(t) = P(T \leq 1) = 1 - e^{(-0,5 \cdot 1)} = 1 - 0,607 = 0,393$$

A probabilidade de falhar no próximo ano é de 0,393 e de não falhar no próximo ano é de $1 - 0,393 = 0,607$.

Ou seja, se forem vendidos 100 máquinas 39,3% irão falhar no período de um ano.

Conhecendo-se os tempos até ocorrer uma falha de um produto é possível definir os períodos de garantia.

Distribuição Exponencial

Os defeitos de um tecido seguem a distribuição de Poisson com média de um defeito a cada 400m. Qual a probabilidade de que o intervalo entre os dois defeitos consecutivos seja entre 800m e 1000m?

$$\text{Logo, } \lambda = 1/400$$

$$P(800 \leq t \leq 1000) = P(t \geq 800) - P(t \geq 1000)$$

$$= e^{-\frac{800}{400}} e^{-\frac{1000}{400}} = e^{-2} - e^{-2,5} = 0,0532 = 5,32\%$$

Distribuição Normal



Distribuição Normal

A distribuição Normal é a mais importante das distribuições estatísticas, tanto na teoria como na prática:

- Representa a distribuição de frequência de muitos fenómenos naturais;
- Serve como aproximação da distribuição Binomial, quando n é grande;
- As médias e as proporções de grandes amostras seguem a distribuição Normal (Teorema do Limite Central).

Distribuição Normal

A distribuição Normal é em forma de sino, unimodal, simétrica em relação à sua média e tende cada vez mais ao eixo horizontal à medida que se afasta da média.

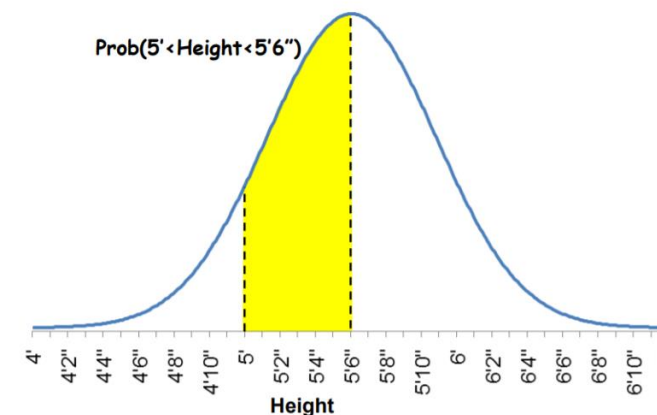
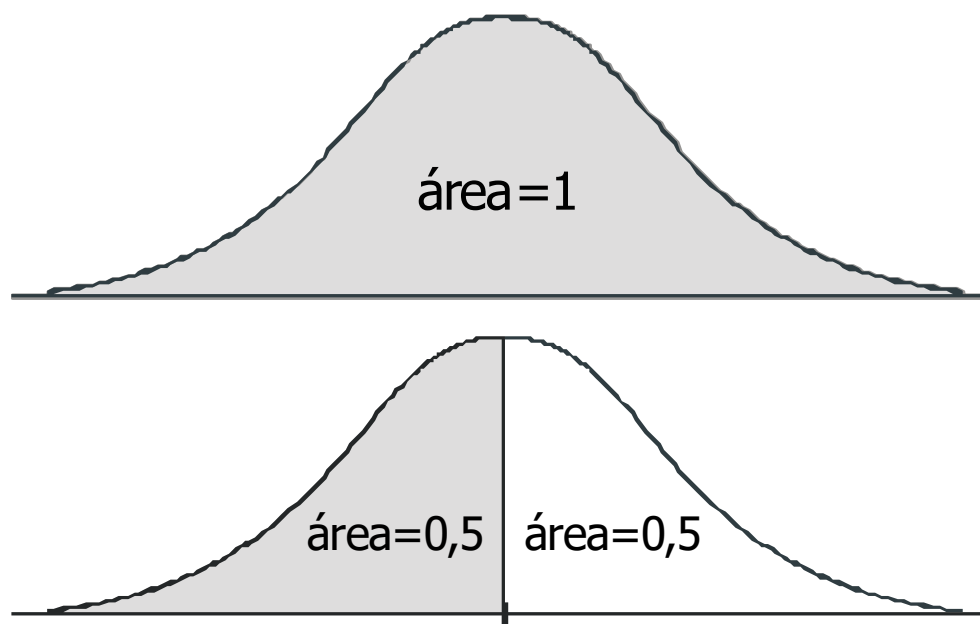
Ou seja, teoricamente os valores da variável aleatória podem variar de $-\infty$ até $+\infty$.

A área abaixo da curva Normal representa 100% de probabilidade associada a uma variável.

A probabilidade de uma variável aleatória tomar um valor entre dois pontos quaisquer é igual à área debaixo da curva compreendida entre esses dois pontos.

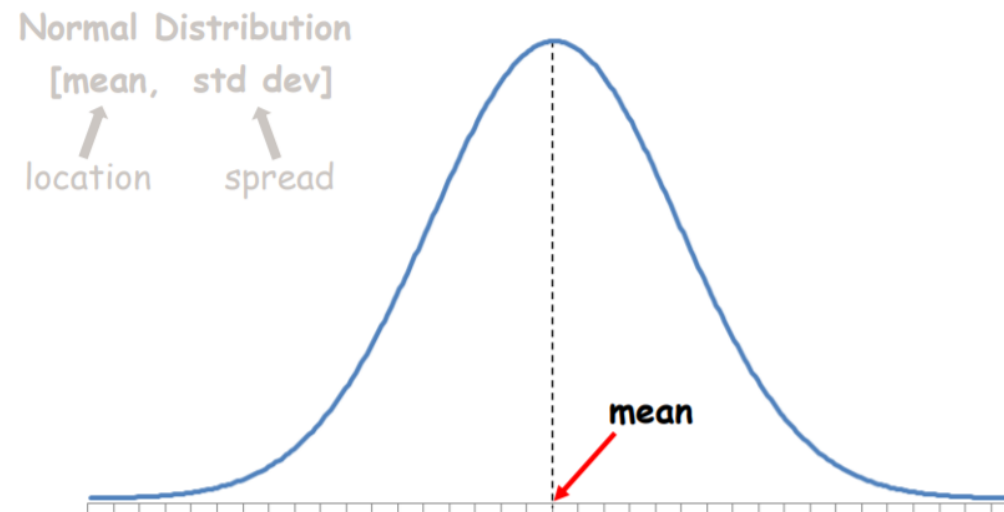
Distribuições Normal

A área total abaixo da curva é considerada como 100%. Isto é, a área total abaixo da curva é 1.



Distribuições Normal

A caracterização da distribuição normal assenta em dois parâmetros:

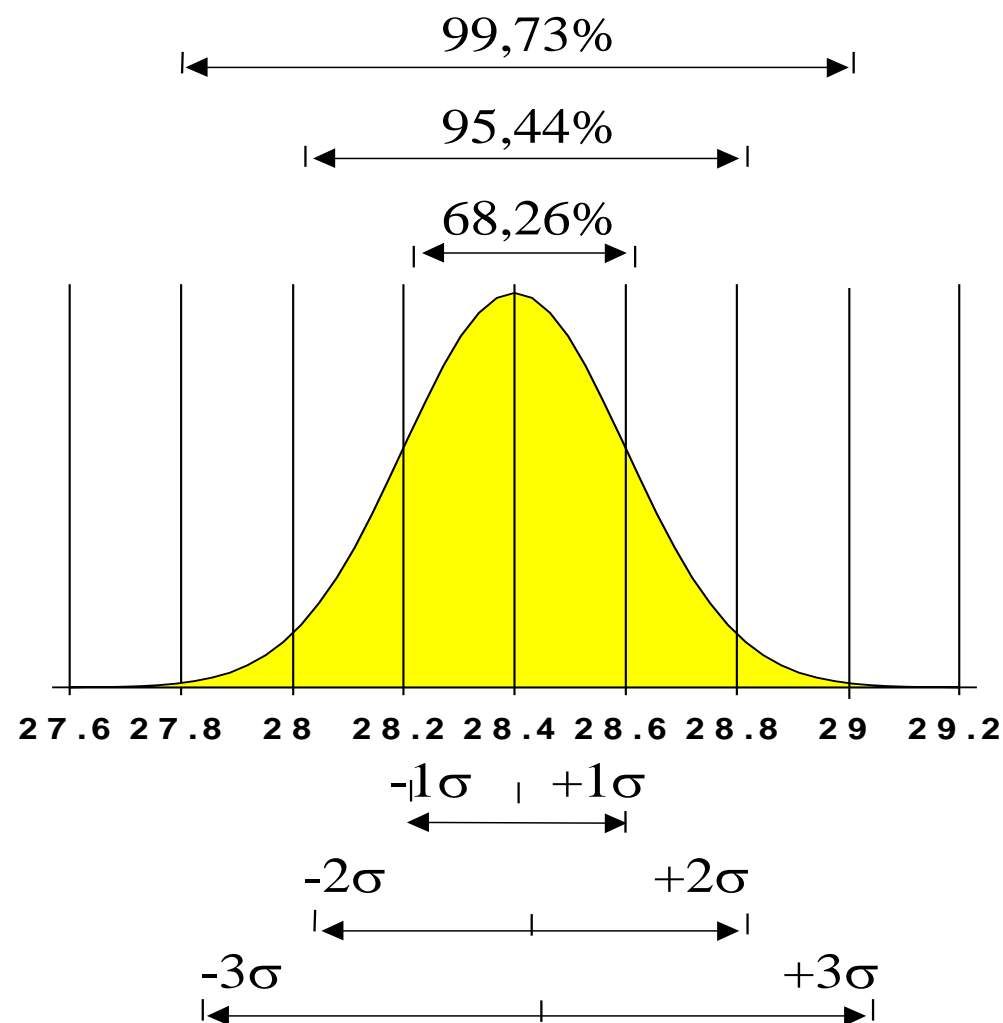


PG_MKT & Business Technologies

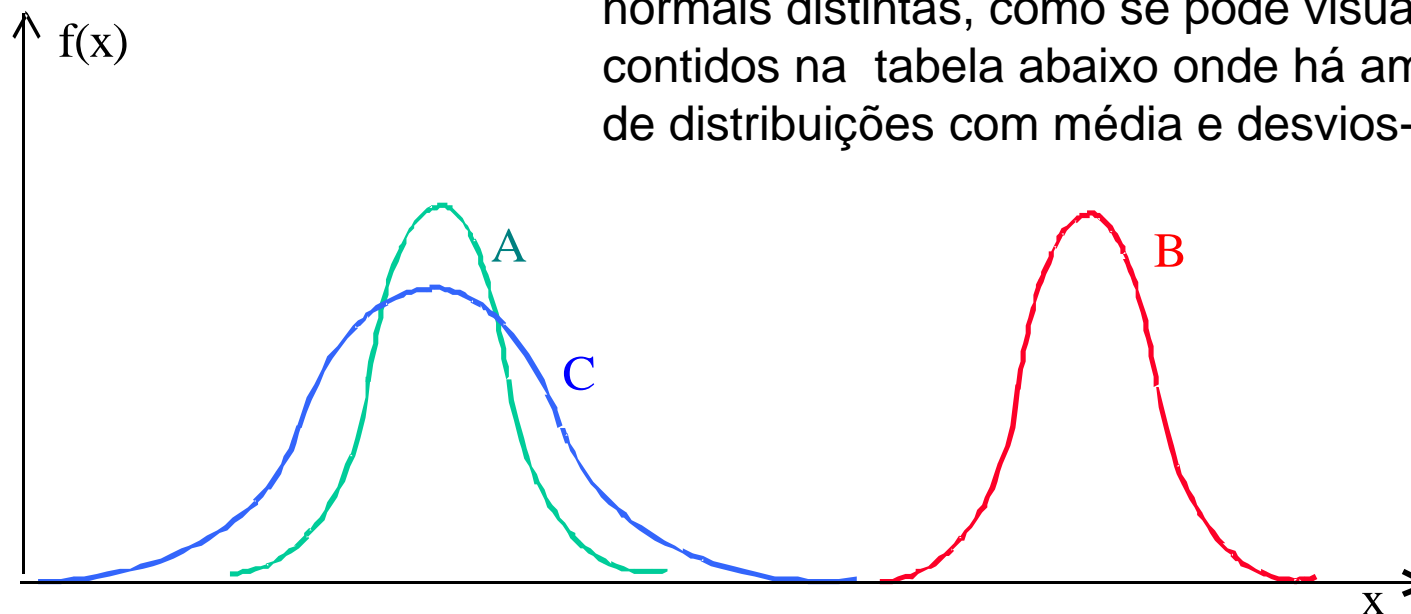
Models and Decision in Business Analytics

... e o desvio padrão

Percentuais da distribuição
Normal:



Ou seja, diferentes médias e desvio-padrão originam curvas normais distintas, como se pode visualizar nos exemplos contidos na tabela abaixo onde há amostras provenientes de distribuições com média e desvios-padrão distintos.



- a) da distribuição A para B muda a tendência central, mas a variabilidade é constante;
- b) da distribuição A para C muda a variabilidade, mas a tendência central é constante;
- c) da distribuição B para C muda a tendência central e a variabilidade.

Importante a normalização da curva

Uma consequência importante do fato de uma distribuição Normal ser completamente caracterizada por sua média e desvio-padrão é que a área sob a curva entre um ponto qualquer e a média é função somente do número de desvios-padrão a que o ponto está da média.

Como existem uma infinidade de distribuições normais (uma para cada média e desvio-padrão), transformamos a unidade estudada seja ela qual for (peso, espessura, tempo, etc.) na unidade Z , que indica o número de desvios-padrão a contar da média.

Dessa forma, o cálculo de probabilidades (área sob a curva) pode ser realizado através de uma distribuição Normal padronizada, onde o parâmetro é a variável reduzida Z .

A distribuição Normal pode ser representada por uma equação matemática dada por:

Probability Density Function

$$\int_a^b \frac{1}{\sqrt{2\pi * \text{std}^2}} e^{-(x-\text{mean})^2/2*\text{std}^2} dx$$

A distribuição Normal acumulada é obtida calculando a probabilidade de X ser menor que um dado valor x :

$$P(X \leq x) = F(x) = \int_{-\infty}^x f(x) dx$$

A solução está apresentada em tabelas da distribuição Normal padronizada onde se entra com a variável reduzida Z (número de desvios-padrão distantes da média) e encontra-se $F(Z)$ ou vice-versa.

$$P\{X \leq x\} = P\left\{Z \leq \frac{x - \mu}{\sigma}\right\} = F(Z) \Rightarrow \text{Tabelado}$$

A variável reduzida mede a magnitude do desvio em relação à **média**, em unidades de **desvio padrão**.

$Z = 1,5$ significa uma observação está desviada 1,5 desvios padrão à direita da média.

A variável reduzida é muito útil para comparar distribuições e detectar dados atípicos.

Dados são considerados atípicos quando $Z > 3$.

$$Z = \frac{X - \bar{X}}{S}$$

Exemplo

Suponha que o peso de um rolo de arame seja normalmente distribuído com média 100 e desvio-padrão 10. Este desvio padrão significa que o peso está em media em torno de 100 a uma distância às vezes maior, às vezes menor do que 10.

Queremos saber qual a probabilidade que um rolo, escolhido ao acaso da produção, possuir peso menor ou igual a 110:

$$P(x < 110) = P(Z < 1) = 0,8413$$

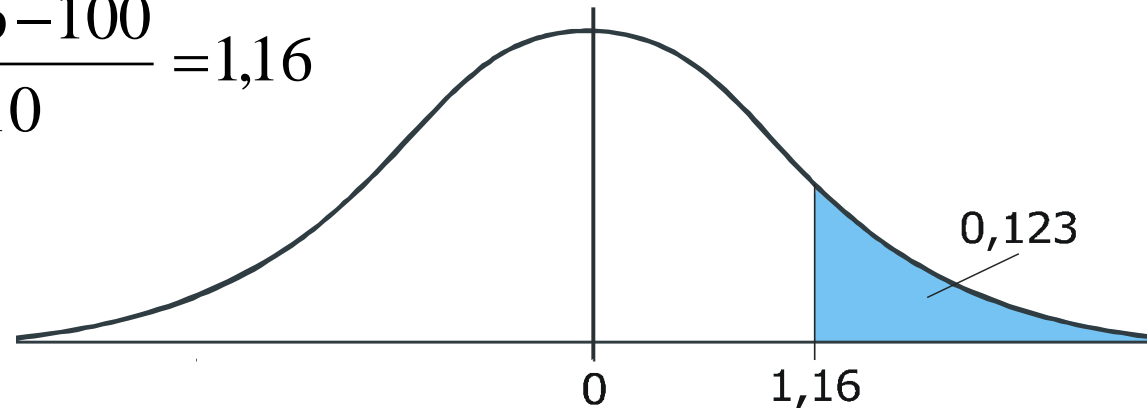
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9278	0.9292	0.9306	0.9319

Probabilidade de ocorrência de valores abaixo de Z

Exemplo

Se pretendessemos saber a probabilidade do peso do rolo ser maior que 111,6, iniciamos calculando o valor de Z:

$$Z = \frac{111,6 - 100}{10} = 1,16$$



Encontramos o valor de probabilidade 0,8770.

$$P(Z > 1,16) = 1 - P(Z < 1,16) = 1 - 0,8770 = 0,123$$

Exemplo

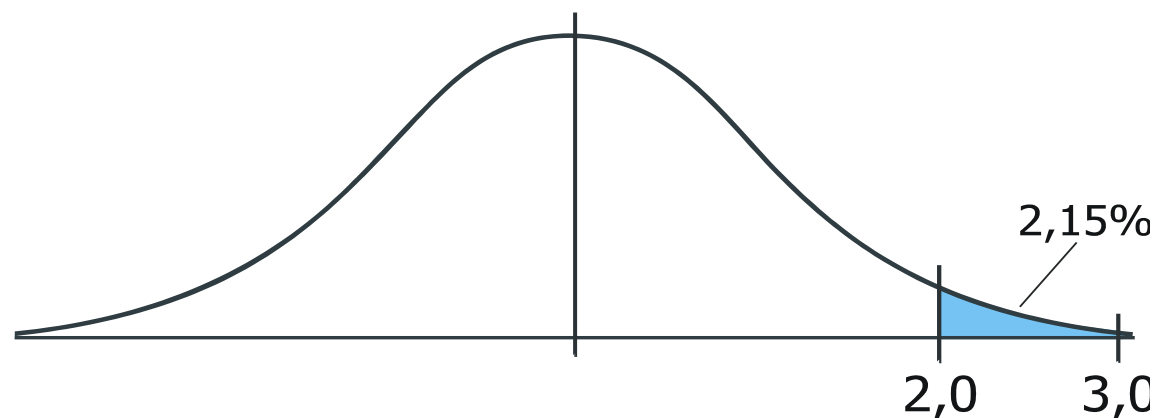
Da mesma forma, se pretendessemos a probabilidade do peso estar entre 120 e 130, teríamos que fazer o seguinte raciocínio:

$$P(120 < X < 130) = P(X < 130) - P(X < 120) =$$

$$P(Z < 3) - P(Z < 2) =$$

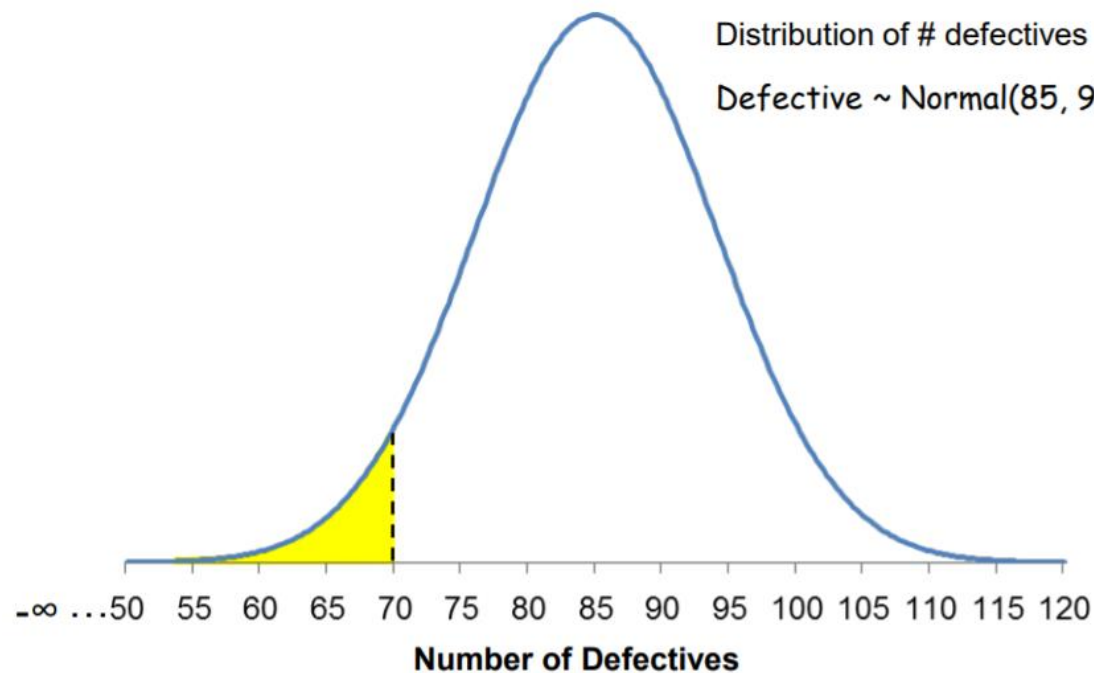
$$0,9987 - 0,9772 = 0,0215$$

ou seja, 2,15% de chance de um rolo pesar entre 120 e 130



Utilização da Função NORMA.DIST

Considere uma distribuição normal com média 85 e desvio padrão 9.



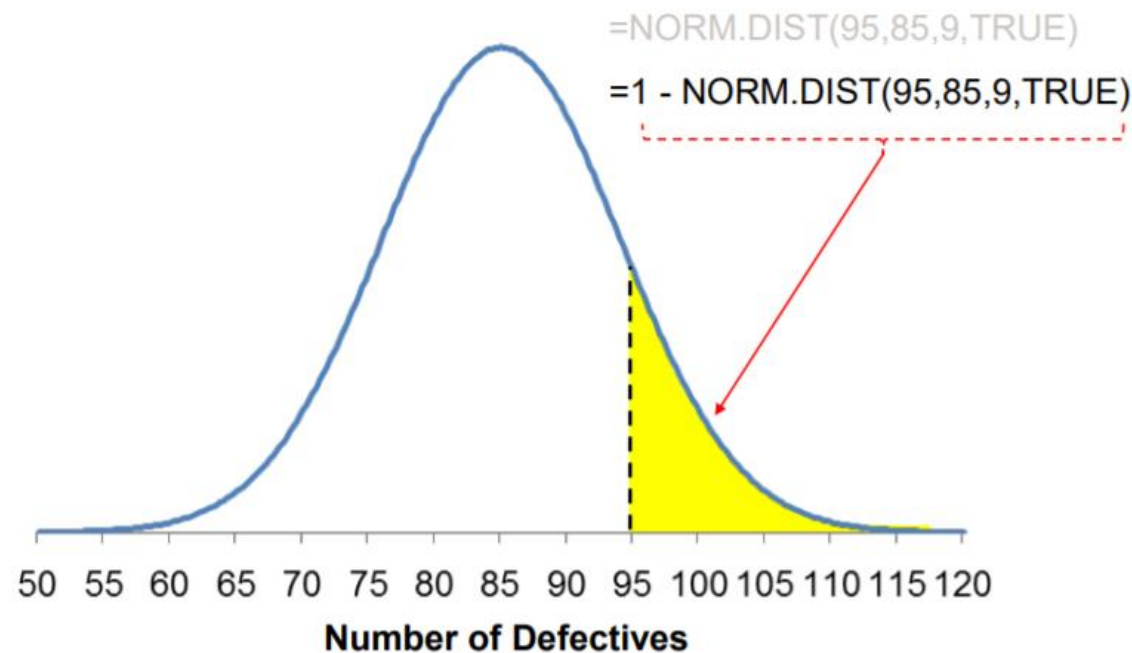
Qual a Probabilidade de < 70 ?

`=NORM.DIST(x, mean, std, TRUE)`

`=NORM.DIST(70, 85, 9, TRUE)`

Utilização da Função NORMA.DIST

Considere uma distribuição normal com média 85 e desvio padrão 9.



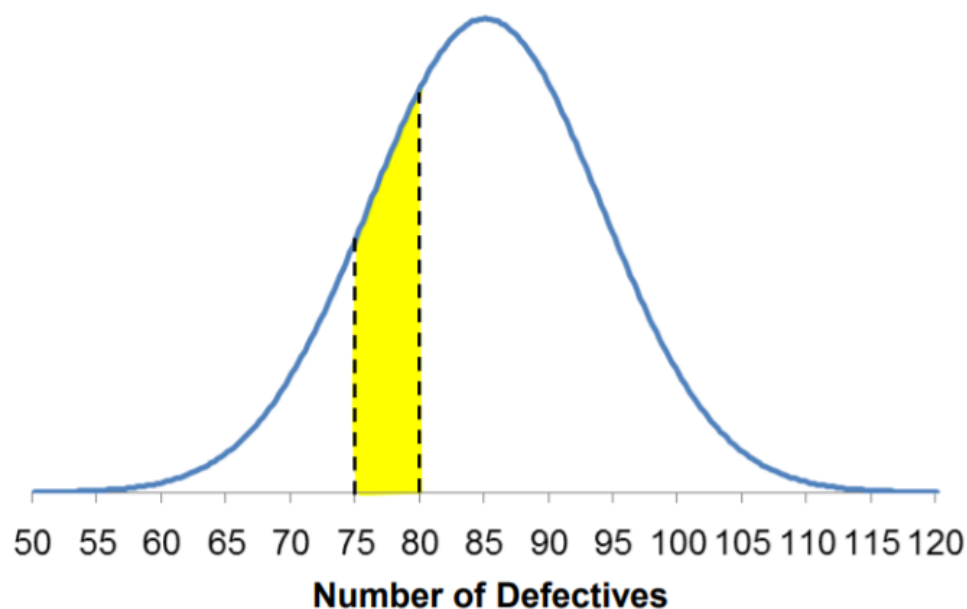
Qual a probabilidade > 95 ?

Utilização da Função NORMA.DIST

Considere uma distribuição normal com média 85 e desvio padrão 9.

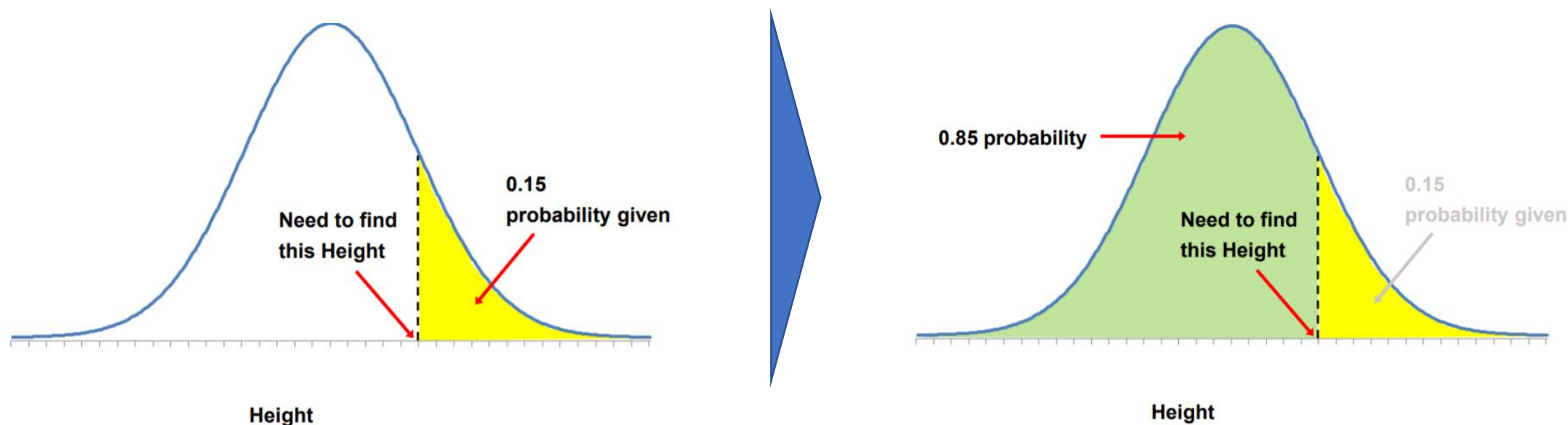
$\text{Prob}(75 < \text{Defective} < 80) = \text{NORM.DIST}(80, 85, 9, \text{TRUE}) - \text{NORM.DIST}(75, 85, 9, \text{TRUE})$

Qual a probabilidade entre 75 e 80?



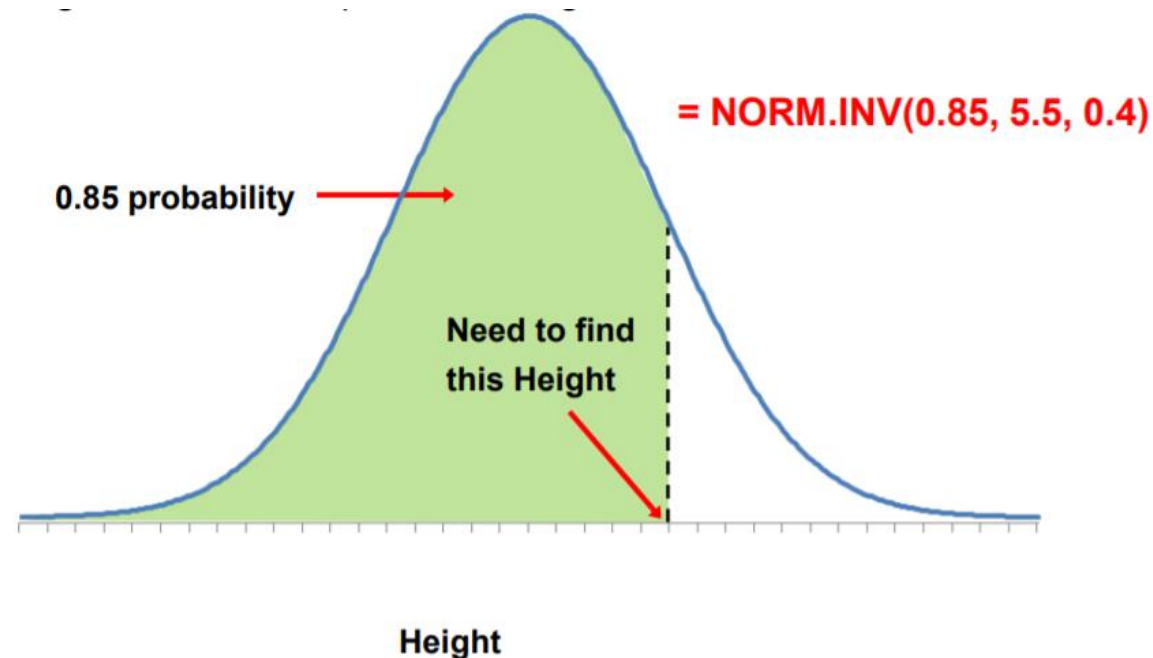
Utilização da Função NORMA.INV (função inversa)

Útil quando queremos saber qual o valor da variável que causa determinada probabilidade. Por exemplo imagine que quer saber qual o valor que gera uma probabilidade de 15%?



Utilização da Função NORMA.INV (função inversa)

Útil quando queremos saber qual o valor da variável que causa determinada probabilidade. Considere uma distribuição normal com média 5,5 e desvio padrão de 0,4). Por exemplo imagine que quer saber qual o valor que gera uma probabilidade de 15%?



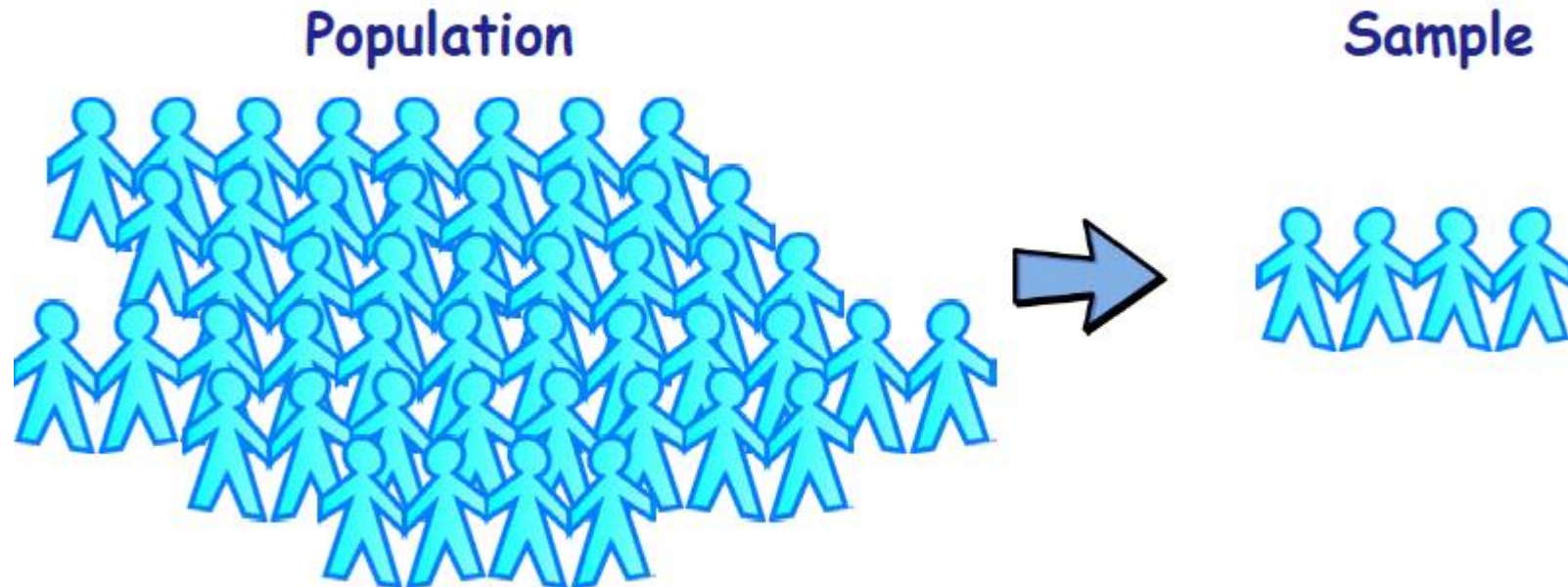


02

Estimação e Teoria do Limite Central

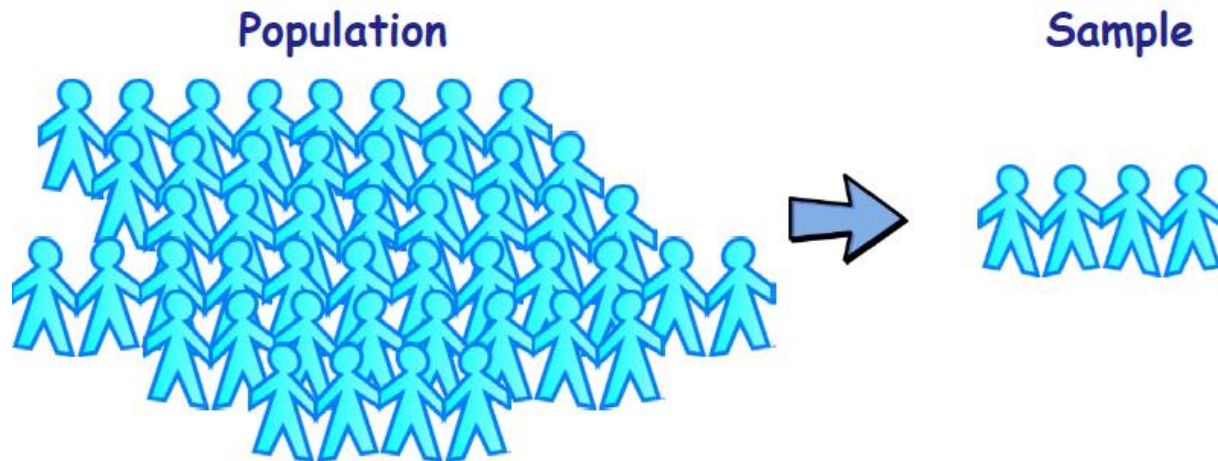


População e Amostra



- Populations and Samples are context dependent
- Sample is a subset of the population
- Sample should be representative of the population

População e Amostra



- Populations and Samples are context dependent
- Sample is a subset of the population
- Sample should be representative of the population

Population Data

Population Mean: μ

Population Standard Deviation: σ

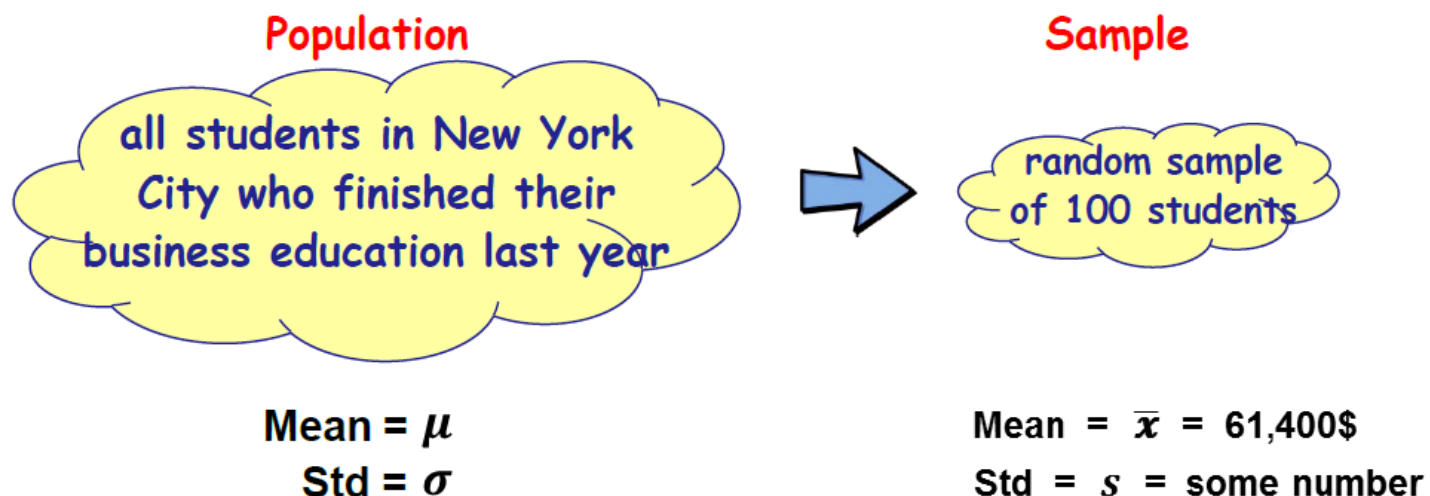
Sample Data

Sample Mean: \bar{x}

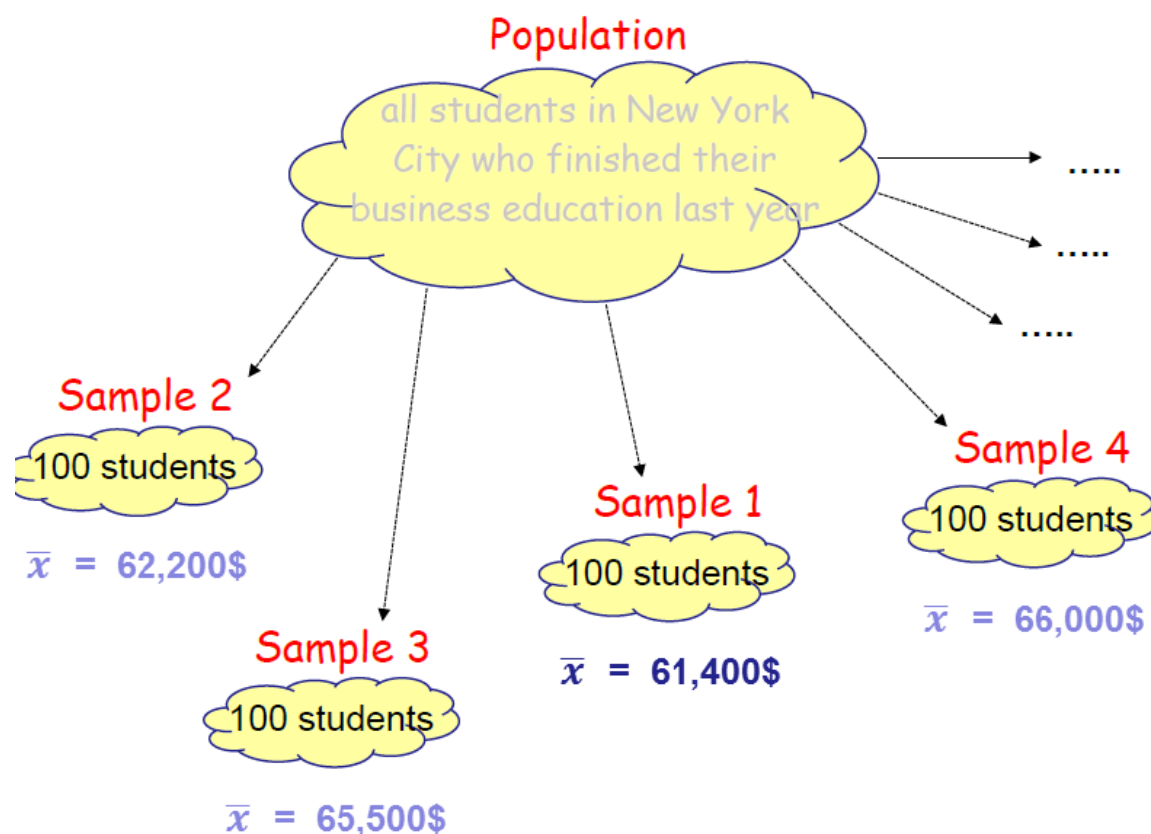
Sample Standard Deviation: s

Exemplo

Imaginem que queremos saber qual o salário médio do primeiro emprego de todos os estudantes de Nova Iorque?



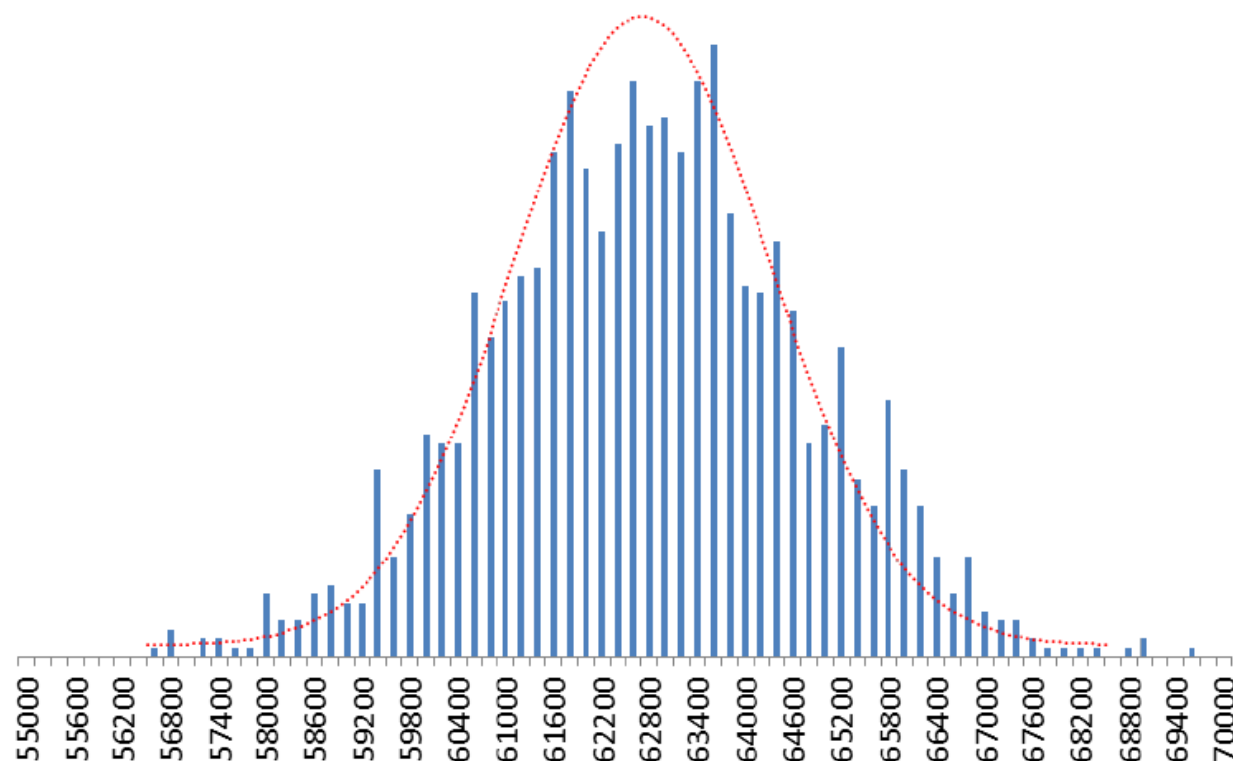
Exemplo



Se formos fazer a distribuição de frequências das médias vamos reparar que ela se aproxima de uma distribuição normal com os seguintes parâmetros:

$$\bar{x} \sim \text{Normal} \left(\mu, \frac{\sigma}{\sqrt{n}} \right)$$

Teoria do Limite Central



Distribuição das médias

$$\bar{x} \sim \text{Normal} \left(\mu, \frac{\sigma}{\sqrt{n}} \right)$$



Considerados para estimar os valores da população

Outro exemplo:

A distribuição de probabilidade da variável resultante do lançamento de um dado segue a distribuição uniforme, ou seja, qualquer valor (1,2,3,4,5,6) tem a mesma probabilidade ($1/6$) de ocorrer.

No entanto, se ao invés de lançar um dado, sejam lançados dois dados e calculada a média, a média dos dois dados seguirá uma distribuição aproximadamente Normal.



PG_MKT & Business Technologies

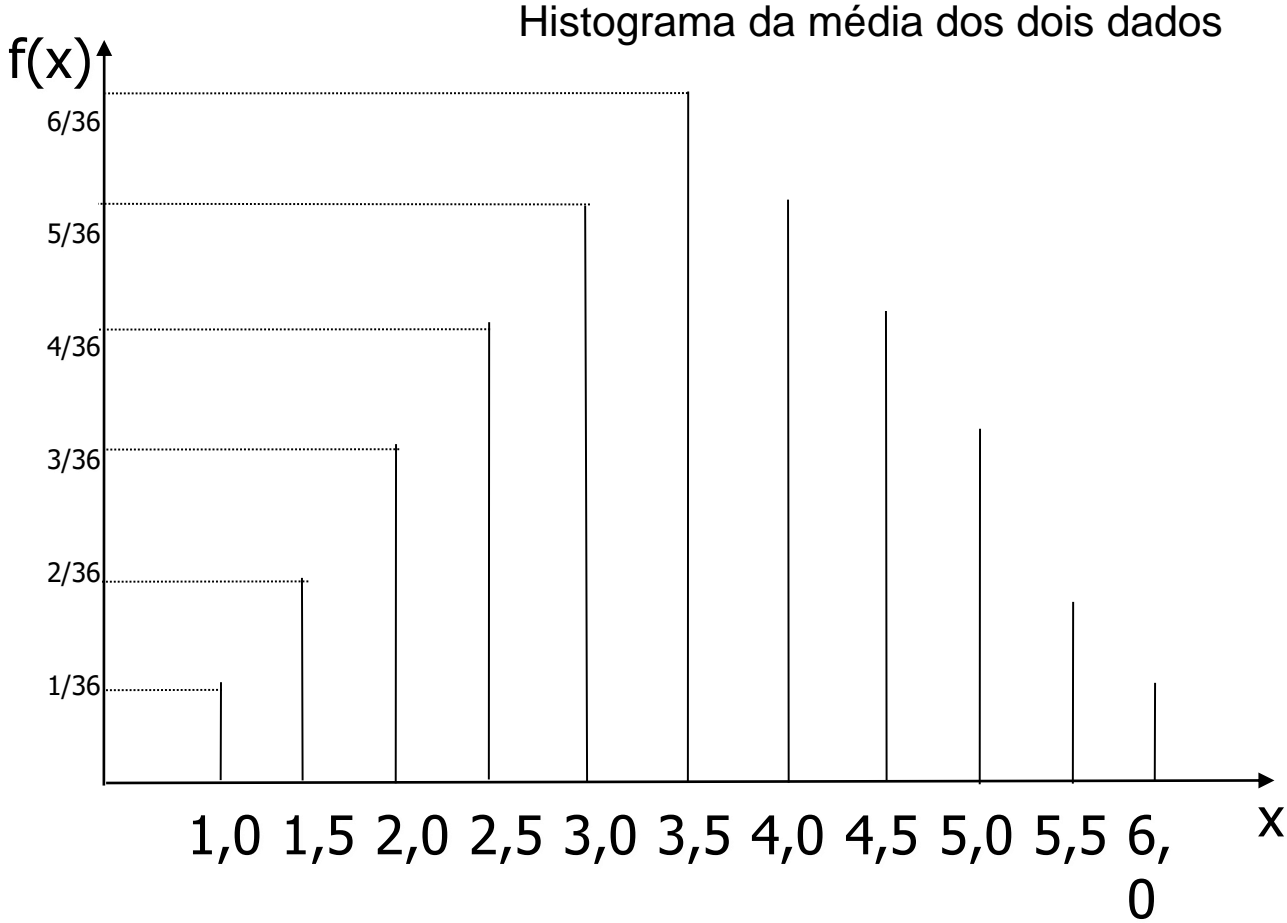
Models and Decision in Business Analytics

Teorema do Limite Central

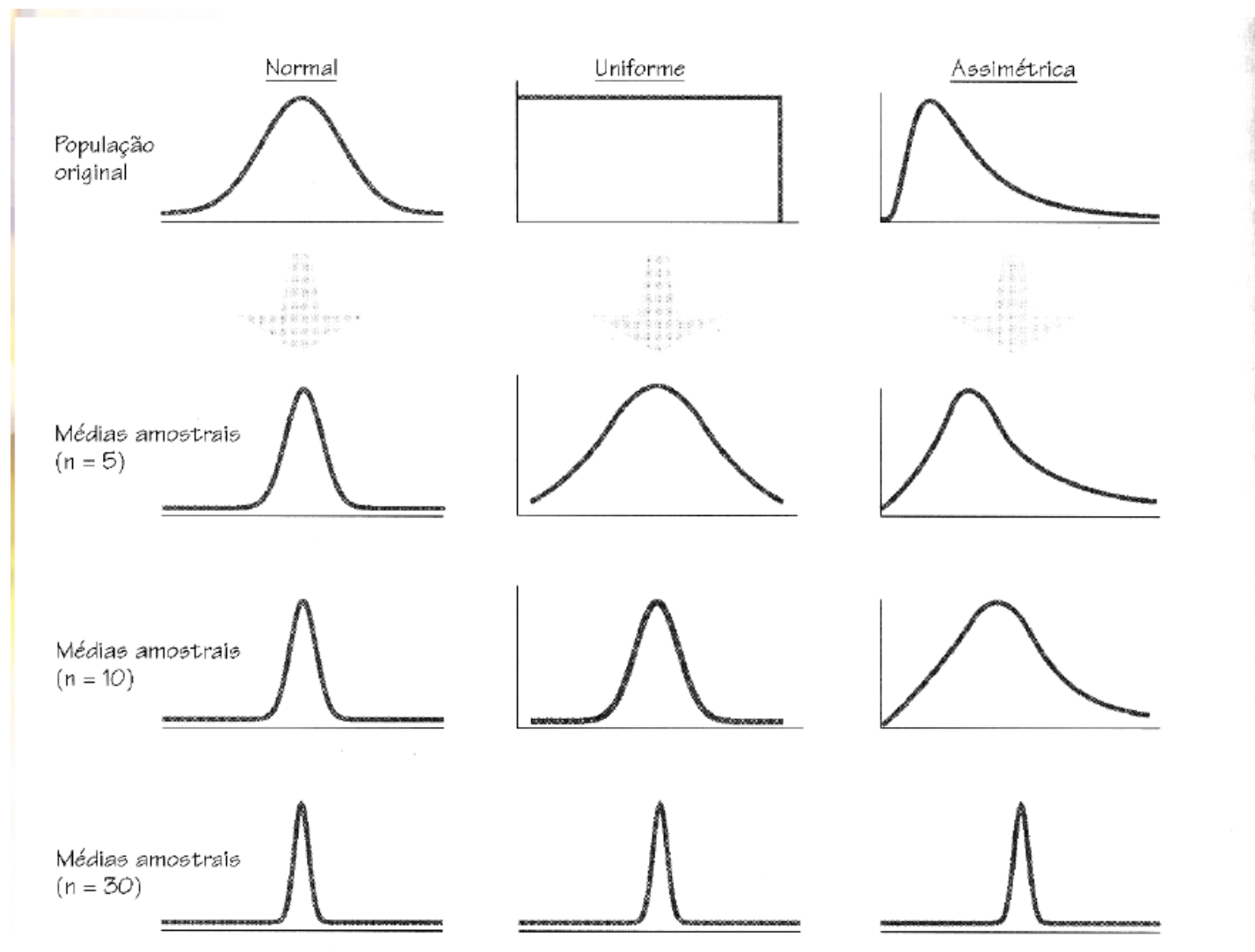
1º dado	2º dado	Soma	Média	1º dado	2º dado	Soma	Média
1	1	2	1,0	5	2	7	3,5
1	2	3	1,5	3	4	7	3,5
2	1	3	1,5	4	3	7	3,5
1	3	4	2,0	2	6	8	4,0
3	1	4	2,0	6	2	8	4,0
2	2	4	2,0	3	5	8	4,0
1	4	5	2,5	5	3	8	4,0
4	1	5	2,5	4	4	8	4,0
3	2	5	2,5	3	6	9	4,5
2	3	5	2,5	6	3	9	4,5
1	5	6	3,0	4	5	9	4,5
5	1	6	3,0	5	4	9	4,5
2	4	6	3,0	4	6	10	5,0
4	2	6	3,0	6	4	10	5,0
3	3	6	3,0	5	5	10	5,0
1	6	7	3,5	5	6	11	5,5
6	1	7	3,5	6	5	11	5,5
2	5	7	3,5	6	6	12	6,0

Tabela de frequência da média dos dois dados

Média de dois dados	Frequência
1,0	1
1,5	2
2,0	3
2,5	4
3,0	5
3,5	6
4,0	5
4,5	4
5,0	3
5,5	2
6,0	1



Teorema do Limite Central



Teorema do Limite Central, estimadores da população

Um investigador deseja saber a média da idade dos alunos de uma pós-graduação. Supondo que a população dos alunos seja:

25	35	24	43	35	22	49	56
34	26	35	52	40	35	35	25
61	42	58	56	45	40	38	45
33	53	22	35	23	25	36	39

$$\mu = \frac{\sum x_i}{N} = \frac{25 + \dots + 39}{32} = 38,19$$

$$\sigma = \sqrt{\frac{(x_i - \mu)^2}{N}} = \sqrt{\frac{(25 - 38,19)^2 + \dots + (39 - 38,19)^2}{32}} = 11,11$$

PG_MKT & Business Technologies

Models and Decision in Business Analytics

Supondo que não fosse possível analisar a população inteira, e os dados fossem recolhidos por amostras cada uma de tamanho $n=4$

	1		2		3		4		5		6		7		8
	25		35		24		43		35		22		49		56
	34		26		35		52		40		35		35		25
	61		42		58		56		45		40		38		45
	33		53		22		35		23		25		36		39
Média (x)	38,25		39		34,75		46,5		35,75		30,5		39,5		41,25
Desvio (S)	15,69		11,40		16,52		9,40		9,43		8,43		6,45		12,92

$$\bar{x} = \frac{\sum \bar{x}_i}{k} = \frac{38,25 + \dots + 41,25}{8} = 38,18$$

$$\bar{\bar{x}} = 38,18 \cong \mu = 38,19$$

$$\begin{aligned}\hat{\sigma}_{\bar{x}} &= \sqrt{\frac{\sum (\bar{x}_i - \bar{x})^2}{k-1}} = \\ &= \sqrt{\frac{(38,25 - 38,18)^2 + \dots + (41,25 - 38,18)^2}{8-1}} = 4,75\end{aligned}$$

$$\sigma_{\bar{x}} = 4,75 \cong \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{11,11}{\sqrt{4}} = 5,55$$

Observações importantes:

- Quando maior o tamanho das amostras, a distribuição das médias será mais próxima de uma distribuição normal.
- Regra prática: para $n > 30$, a distribuição das médias amostrais pode ser aproximada satisfatoriamente por uma distribuição normal.
- Se a distribuição da variável 'x' for originalmente uma distribuição normal, então a distribuição das médias amostrais terá distribuição normal para qualquer tamanho amostral 'n'.

Estimativa média populacional

Em geral a média amostral do conjunto de dados é a melhor estimativa de uma média populacional.

Observação:

Uma estimativa é um valor específico, ou um intervalo de valores usados para aproximar um parâmetro populacional.

Um estimador é uma característica da amostra (Ex:), utilizado para obtermos uma aproximação do parâmetro populacional

Para proceder à estimação do intervalo de confiança:

É uma amplitude (ou um intervalo) de valores que tem a probabilidade de conter o valor verdadeiro da população.

Observa-se que, na definição de intervalo de confiança, está associado uma probabilidade.

A esta probabilidade chamamos de: Nível de Confiança, Grau de Confiança ou Coeficiente de Confiança.

Intervalo de
confiança

$$\text{Probabilidade}\{c_1 \leq \mu \leq c_2\} = 1 - \alpha$$

O intervalo (c_1, c_2) é chamado de intervalo de confiança da média da população.

α é o nível de significância.

100 $(1 - \alpha)$ é o nível de confiança em %.

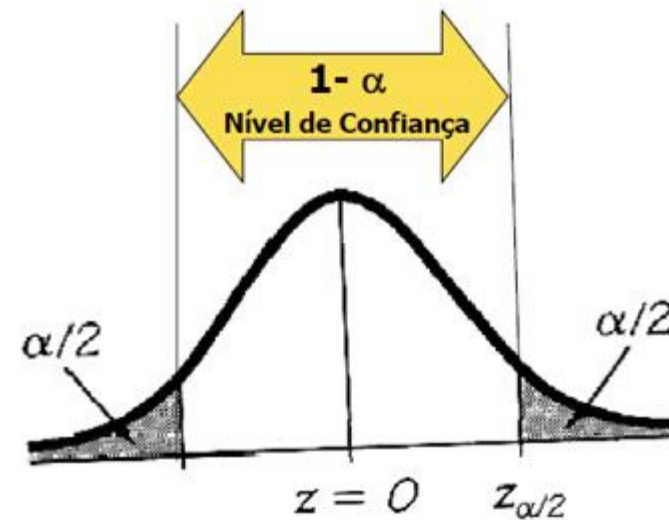
$1 - \alpha$ é o coeficiente de confiança.

Nível de Confiança

É a probabilidade $1-\alpha$ (comumente expressa percentualmente) do intervalo de confiança conter o valor verdadeiro, isto é, o parâmetro respeitante à população populacional

Construção dos limites do intervalo:

$$z_{\alpha/2} = \frac{x - \bar{\bar{x}}}{\sigma/\sqrt{n}}$$





Exercicio:

O processo de produção mancais para unidades de caixa de um determinado tipo de motor foi modificado recentemente. Antes da modificação, os dados históricos indicavam que os diâmetros do orifício dos mancais nas caixas eram distribuídos normalmente com $\sigma=0,100\text{mm}$. Acredita-se que a modificação no processo não tenha alterado a distribuição ou o desvio padrão, mas o valor do diâmetro médio pode ter mudado.

Selecionou-se uma amostra de 40 caixas e mede-se o diâmetro do orifício para cada uma, resultando num diâmetro médio de 5,426mm. Calcule um Intervalo de Confiança para o diâmetro médio real (populacional) do orifício usando um nível de confiança de 90%.

$$\Rightarrow 1,645 = \frac{x_s - 5,426}{0,100/\sqrt{40}} \therefore 0,026 = x_s - 5,426 \therefore x_s = 5,452$$

$$\Rightarrow -1,645 = \frac{x_i - 5,426}{0,100/\sqrt{40}} \therefore -0,026 = x_i - 5,426 \therefore x_i = 5,400$$

$$\mu = 5,426 \pm 0,026 \text{ ou } 5,400 < \mu < 5,452$$

Existe 90% de probabilidade do intervalo de 5,400mm a 5,452mm conter a média populacional do diâmetro do orifício do mancal.

Exercicio:

Na engenharia de determinados produtos é importante considerar os pesos das pessoas, de modo a evitar sobrecargas (aviões, elevadores) ou falhas (cadeiras que se quebram).

Dado que a população de homens dos EUA (ano desconhecido) tem pesos distribuídos normalmente com média 78,47Kg e desvio-padrão 13,61Kg, determinar a probabilidade de:

- (a) um homem escolhido aleatoriamente pesar mais de 81,65Kg.
- (b) em 36 homens escolhidos aleatoriamente, o peso médio ser superior a 81,65Kg.

a) um homem escolhido aleatoriamente pesar mais de 81,65Kg.

Como se trata de um valor individual proveniente de uma população com distribuição normal, calcular o valor de z diretamente:

$$z = \frac{x - \mu}{\sigma} = \frac{81,65 - 78,47}{13,61} = 0,2337$$

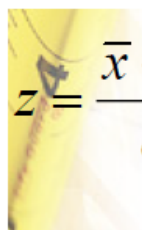
A área correspondente a $z=0,2337$ é **0,5910 (tabela acumulada)**. A probabilidade desejada é, pois:

$$P(z > 0,2337) = 1 - 0,5910 = \mathbf{0,4090}$$

(b) em 36 homens escolhidos aleatoriamente, o peso médio ser superior a 81,65Kg.

Como estamos lidando com a média para um grupo de 36 valores, usamos o *Teorema do Limite Central*

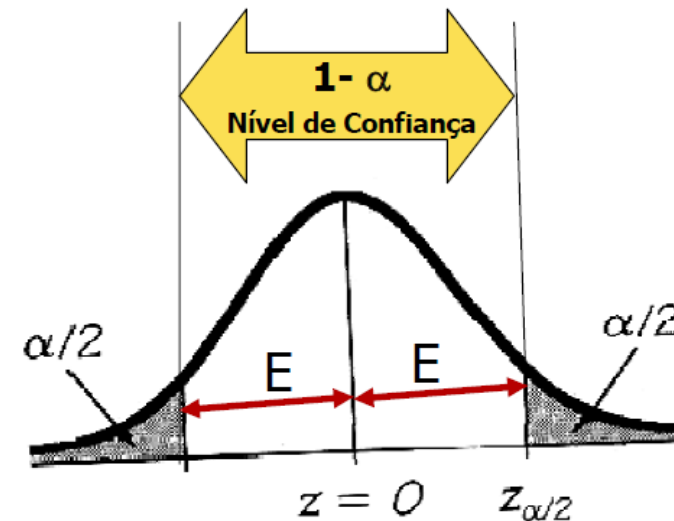
$$\mu_{\bar{x}} = \mu = 78,47$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{13,61}{\sqrt{36}} = 2,2683$$


$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{81,65 - 78,47}{\frac{13,61}{\sqrt{36}}} = 1,4019 \Rightarrow$$

$$P(z > 1,4019) = 1 - 0,9192 = \mathbf{0,0808}$$

Ou seja:

Há uma probabilidade de $1-\alpha$ de uma média amostral conter um erro não superior a E , e uma probabilidade de α de uma média amostral conter um erro superior a E .



Exemplo:

Numa pesquisa, foram coletadas 106 amostras de temperatura, obtendo-se uma média de 98,20 °F e desvio padrão $s=0,62$ °F

Para um nível de confiança de 95%, determine:

- (a) A margem de erro da estimativa
- (b) O Intervalo de confiança para μ

$$NC=95\% \Rightarrow \alpha=0,05 \Rightarrow Z_{\alpha/2} = 1,96$$

$$(a) E = Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 1,96 \cdot \frac{0,62}{\sqrt{106}} = 0,12$$

$$(b) \text{ Como } \bar{x} = 98,20 \text{ e } E = 0,12;$$

$$\bar{x} - E < \mu < \bar{x} + E$$

$$98,2 - 0,12 < \mu < 98,2 + 0,12$$

$$98,08 < \mu < 98,32$$

$$\mu = 98,20 \pm 0,12$$

OU

$$(98,08; 98,32)$$

**Formulas da média da população
com base na amostra**

$$\boxed{\bar{x} - E < \mu < \bar{x} + E} \text{ ou}$$

$$\boxed{\mu = \bar{x} \pm E}$$

$$\boxed{(\bar{x} - E; \bar{x} + E)}$$



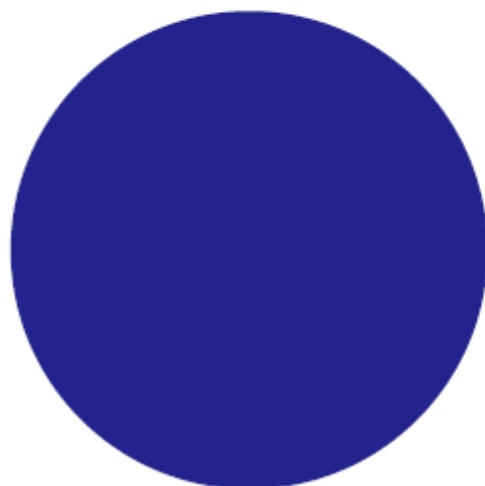
03

A m o s t r a g e m



População e Amostra

Population



Sample from that Population



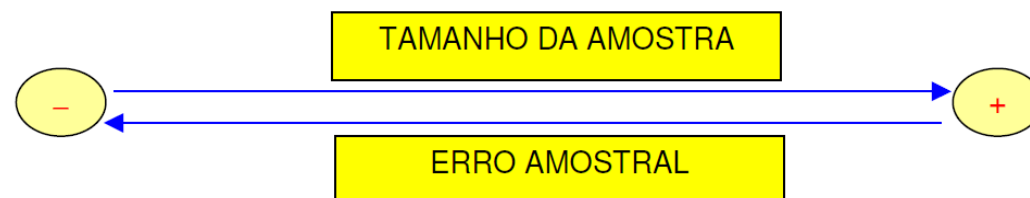
Métodos de Amostragem Probabilística são os que selecionam os indivíduos da população de forma que todos tenham as mesmas hipóteses de participar da Amostra (aleatória).

População e Amostra

Não há dúvida de que uma amostra não representa perfeitamente uma população. Ou seja, a utilização de uma amostra implica na aceitação de uma margem de erro que denominaremos ERRO AMOSTRAL.

Erro Amostral é a diferença entre um resultado amostral e o verdadeiro resultado populacional; tais erros resultam de flutuações amostrais aleatórias

Não podemos evitar a ocorrência do ERRO AMOSTRAL, porém podemos limitar seu valor através da escolha de uma amostra de tamanho adequado. Obviamente, o ERRO AMOSTRAL e o TAMANHO DA AMOSTRA seguem sentidos contrários



Margem de Erro

Quando utilizamos dados amostrais para estimar uma média populacional μ , a margem de erro (E) é a diferença máxima provável (com probabilidade $1-\alpha$) entre a média amostral observada e a verdadeira média da população (μ)

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Grau de Confiança	α	Valor Crítico $Z_{\alpha/2}$
90%	0,10	1,645
95%	0,05	1,96
99%	0,01	2,575

Determinação do Tamanho da Amostra

Uma das perguntas mais importantes numa análise estatística é determinar qual o melhor tamanho de amostras que devemos ter.

- Amostras muito grandes são dispendiosas e demandam mais tempo de manipulação e estudo
- Amostras pequenas são menos precisas e pouco confiáveis

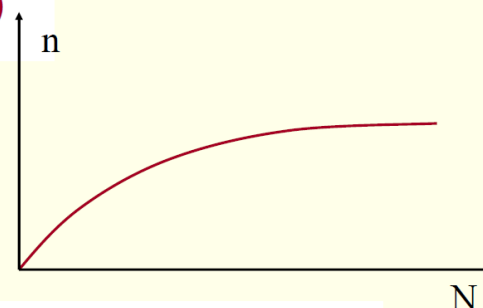
DETERMINAÇÃO DO TAMANHO DE UMA AMOSTRA COM BASE NA ESTIMATIVA DA MÉDIA POPULACIONAL (População muito grande)

Pode-se estimar o melhor tamanho da amostra pela fórmula:

$$n = \left[\frac{Z_{\alpha/2} \cdot \sigma}{E} \right]^2$$

Observa-se que o tamanho da amostra depende do grau de confiança desejado, da margem de erro pretendida e do σ .

Tamanho da amostra (n) e tamanho da população (N)



A fórmula exige que se substitua por algum valor o desvio-padrão populacional (σ), mas se este for desconhecido, devemos poder utilizar um valor preliminar obtido por processos como:

- Realizar um estudo piloto iniciando o processo de amostragem. Com base na primeira coleção de pelo menos 31 valores amostrais selecionados aleatoriamente, calcular o desvio padrão amostral 's' e utilizá-lo em lugar de σ . Este valor pode e deve ser refinado com a obtenção de mais dados amostrais.

Exemplo

Queremos estimar a renda média no primeiro ano de um profissional. Quantas recolhas devemos realizar se queremos 95% de confiança em que a média esteja a menos que R\$1.000,00 da renda média verdadeira da população. Suponha σ conhecido e igual a R\$3.000,00.

Com erro de R\$1.000

$$\left[\frac{1,96 \cdot 3000}{1000} \right]^2 = 34,54 \Rightarrow 35 \text{ amostras}$$

Com erro de R\$2.000

$$\left[\frac{1,96 \cdot 3000}{2000} \right]^2 = 8,64 \Rightarrow 9 \text{ amostras}$$

DETERMINAÇÃO DO TAMANHO DE UMA AMOSTRA COM BASE NA ESTIMATIVA DA MÉDIA POPULACIONAL - POPULAÇÃO FINITA (quando n é $\geq 5\%$ de N)

A fórmula para determinação do tamanho da amostra que vimos até agora trabalha com a idéia de que a população de onde se retirava a amostra era tão grande, que poderíamos considerá-la infinita. Entretanto, a maior parte das populações não é tão grande em comparação com as amostras.

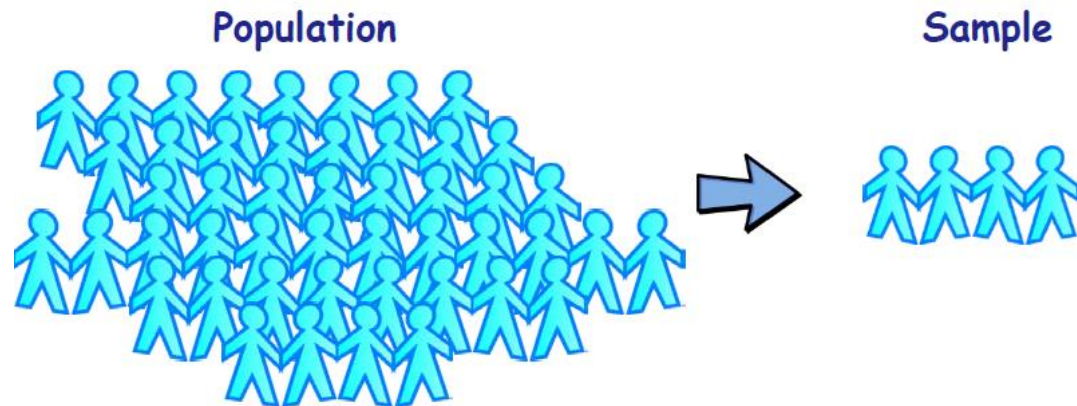
$$n = \frac{N \cdot \sigma^2 \cdot (Z_{\alpha/2})^2}{(N-1) \cdot E^2 + \sigma^2 \cdot (Z_{\alpha/2})^2}$$

$$e = z \cdot \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

Tipos de Amostragem

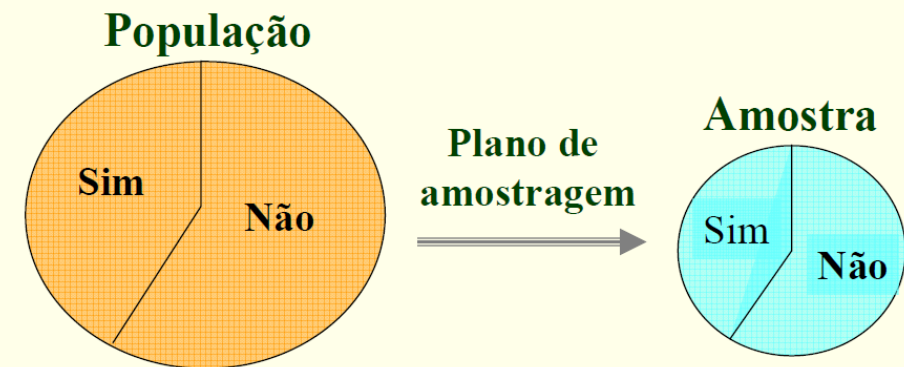
PG_MKT & Business Technologies

Models and Decision in Business Analytics



Amostragem

A amostra deve ser *representativa*!



Técnicas de amostragem:

Amostragem probabilística (aleatória) - a probabilidade de um elemento da população ser escolhido é conhecida.

Amostragem não probabilística (não aleatória) - Não se conhece a probabilidade de um elemento da população ser escolhido para participar da amostra.

Tipologias de amostragem:

- Amostragem aleatória simples
- Amostragem sistemática
- Amostragem estratificada
- Amostragem por conglomerados

Amostragem aleatória simples

Cada subconjunto da população com o mesmo n^o de elementos tem a mesma chance de ser incluído na amostra.

$$p = n / N$$

É equivalente a um sorteio aleatório. Nesse tipo de amostragem é necessário que os elementos da população sejam numerados. Quando o número de elementos da amostra é muito grande, esse tipo de sorteio torna-se muito trabalhoso.

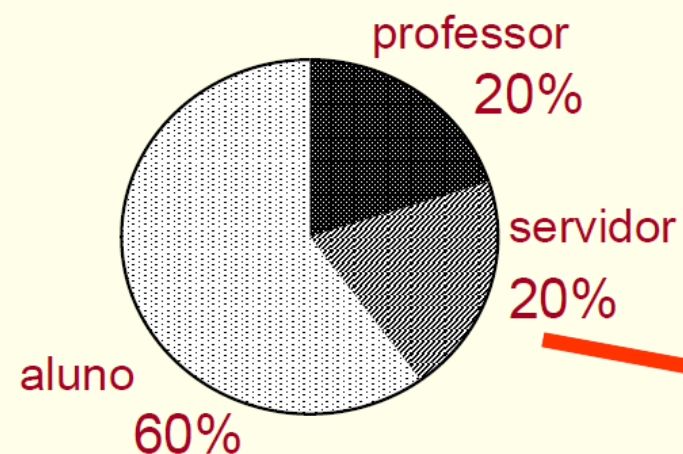
Amostragem sistemática

Os elementos da população apresentam-se ordenados e são retirados periodicamente (de cada k elementos, um é escolhido)

Amostragem Estratificada:

POPULAÇÃO:

comunidade da escola



AMOSTRA: parte da
comunidade da escola

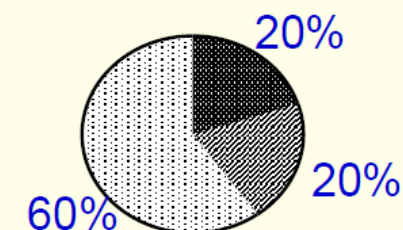
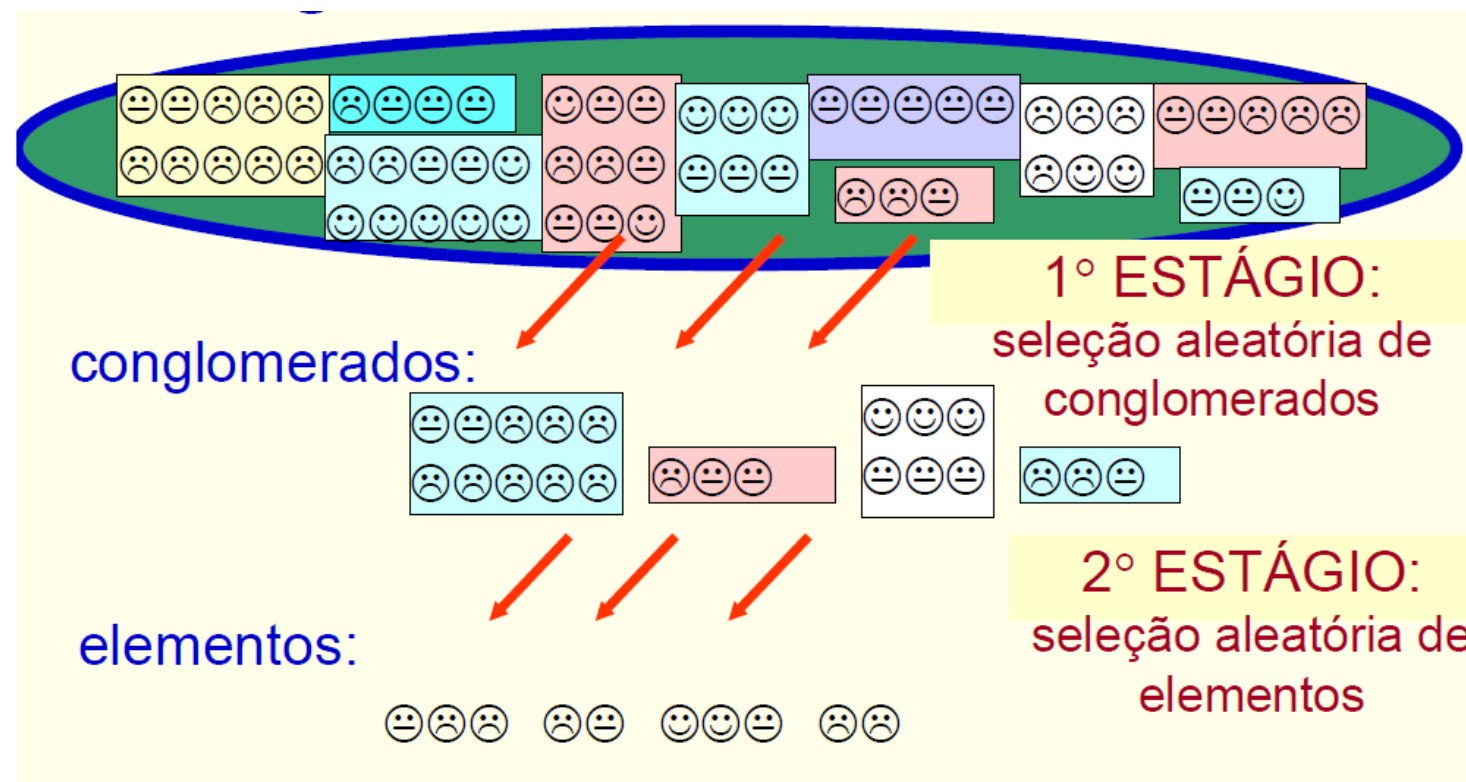


Ilustração de uma amostragem estratificada proporcional.

Amostragem por conglomerados:





04

Testes de Hipóteses e Significância



Testes de Hipóteses

Decidir se determinada afirmação sobre um parâmetro populacional é, ou não, apoiada pela evidência obtida de dados amostrais.

Em estatística, uma hipótese é uma alegação, ou afirmação, sobre uma característica de uma população:

- Investigadores médicos afirmam que a temperatura média do corpo humano não é igual a 37°C ;
- Um novo fertilizante utilizado no cultivo de hortaliças aumenta a produtividade.

Exemplo:

Estudos prévios indicam que a temperatura do corpo humano é 98,60°F. Investigadores médicos de Maryland recolheram dados amostrais com média = 98,20°F e distribuição aproximadamente normal.

Estes dados amostrais constituem evidência suficiente para rejeitar a crença comum de que $\mu = 98,6^\circ\text{F}$?

Testes de Hipóteses: metodologia

A hipótese nula H_0 é uma afirmação que diz que o parâmetro populacional é tal como especificado (isto é, a afirmação é correta).

$$H_0 : \mu = 98,6$$

A hipótese alternativa H_1 é uma afirmação que oferece uma alternativa à alegação (isto é, o parâmetro é maior/menor/diferente que o valor alegado).

$$H_1 : \mu \neq 98,6$$

Testes de Hipóteses: metodologia

A hipótese nula H_0 representa o *status quo*, ou seja, a circunstância que está sendo testada, e o objetivo dos testes de hipóteses é sempre tentar rejeitar a hipótese nula.

A hipótese alternativa H_1 representa o que se deseja provar ou estabelecer, sendo formulada para contradizer a hipótese nula.

Testes de Hipóteses: metodologia

Teste Bilateral:

$H_0 : \mu = \text{valor numérico}$

$H_1 : \mu \neq \text{valor numérico}$

Teste Unilateral Superior:

$H_0 : \mu = \text{valor numérico}$

$H_1 : \mu > \text{valor numérico}$

Teste Unilateral Inferior:

$H_0 : \mu = \text{valor numérico}$

$H_1 : \mu < \text{valor numérico}$

Tipo de Erros:

		O Verdadeiro Estado da Natureza	
		A hipótese nula é verdadeira.	A hipótese nula é falsa.
Decisão	Decidimos rejeitar a hipótese nula.	Erro tipo I (rejeição de uma hipótese nula verdadeira)	Decisão correta
	Não rejeitamos a hipótese nula.	Decisão correta	Erro tipo II (Não rejeição de uma hipótese nula falsa)

Exemplo:

Uma máquina automática enche pacotes de café segundo uma distribuição normal com média μ e desvio-padrão 20g. A máquina foi regulada para $\mu = 500$ g e meia em meia hora tiramos uma amostra de 16 pacotes para verificar se o empacotamento está sob controle, isto é, se $\mu = 500$ g.

Se uma dessas amostras apresentasse $x = 492$ g, você pararia ou não o empacotamento para verificar se o ajuste da máquina está correto?

Exemplo:

Passo 1: Indicamos por X o peso de cada pacote, então X é uma normal com média μ e $\sigma = 20$. As hipóteses que nos interessam são:

Hipótese nula: $H_0 : \mu = 500 \text{ g}$

Hipótese alternativa: $H_1 : \mu \neq 500 \text{ g}$ (BILATERAL)

pois a máquina pode desregular para mais ou para menos.

Exemplo:

Passo 2: Escolher a distribuição amostral

Se o desvio padrão populacional é conhecido:

- Distribuição NORMAL (Caso deste exemplo típico)

Se o desvio é desconhecido e a amostra é pequena ($n < 30$):

- Distribuição de STUDENT

Passo 3: Escolher o nível de significância

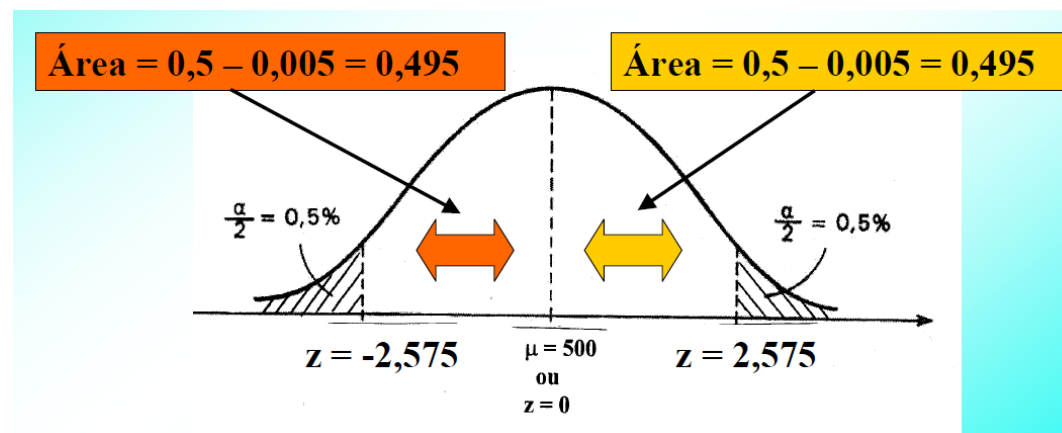
Pela situação descrita no problema, podemos fazer $\alpha = 0,01$

Passo 4: Calcular a estatística de teste, valores e região crítica

$$\text{estatística de teste} = \frac{\text{média amostral} - \text{média alegada}}{\text{desvio padrão da distribuição amostral}}$$

$$z_{\text{teste}} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \quad \text{ou} \quad t_{\text{teste}} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

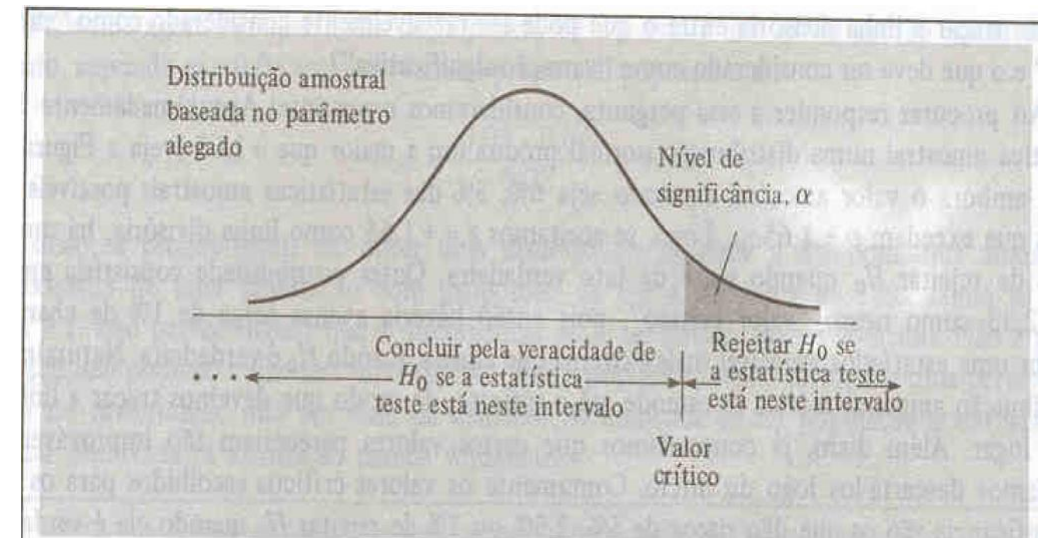
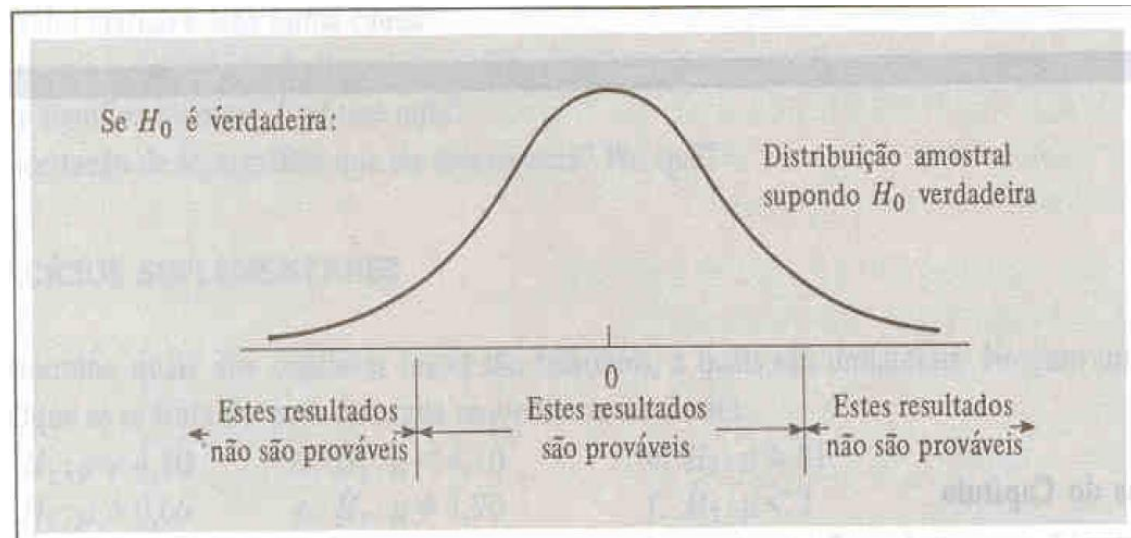
$$z = \frac{x - \mu}{\sigma / \sqrt{n}} = \frac{492 - 500}{20 / \sqrt{16}} = \frac{-8}{5} = -1,6$$



Passo 5:

A informação da amostra é que $x = 492$ g (o que fornece $z = -1,6$)

Como $x \notin$ Região Crítica, nossa conclusão será não rejeitar H_0



Obrigado!
(luisflcosta@sapo.pt)