Big Data Analytics com R e Microsoft Azure Machine Learning

Introdução à Análise Estatística de Dados

Seja Bem-Vindo(a)!

Introdução à Análise Estatística de Dados

Introdução à Análise Estatística de Dados Parte 1

Introdução à Análise Estatística de Dados Parte 2

Introdução à Análise Estatística de Dados Parte 3

Amostragem

Seja Bem-Vindo(a)!

Amostragem

Vamos imaginar uma fábrica de biscoitos, que gostaria de **medir** se o **nível de sal** presente nos seus produtos, está dentro dos **padrões** determinados pelo **Ministério da Saúde.**

Amostragem

Você acha que seria viável, a empresa medir cada unidade de biscoito produzida?

Amostragem

São produzidos milhares de pacotes por dia, cada um com dezenas de unidades.

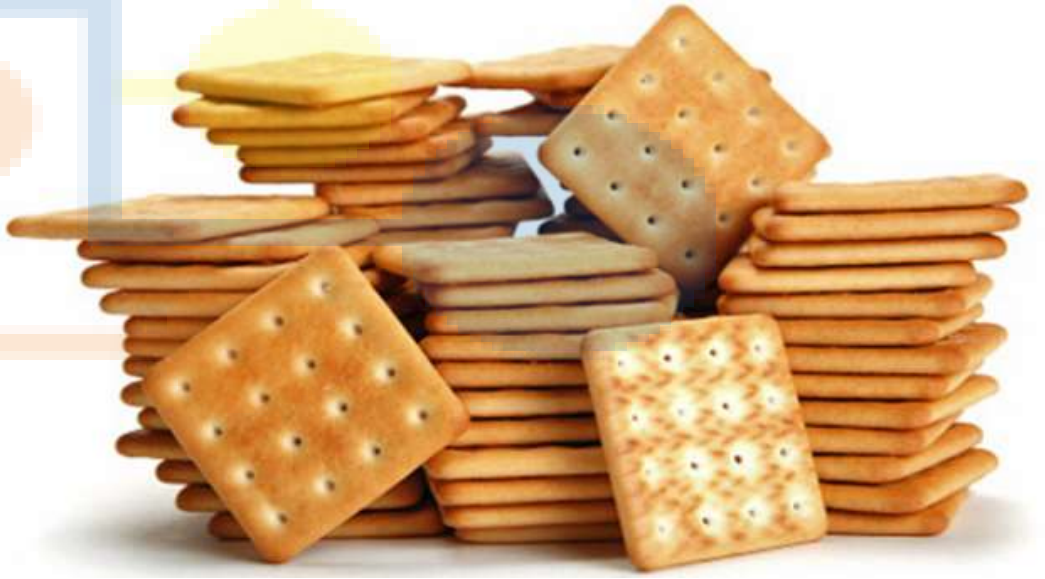
Não seria economicamente viável medir o nível de sal da população inteira de biscoitos.

Amostragem

A solução então, seria selecionar de forma randômica, mas rotineira, pequenas amostras de biscoitos, que fossem representativas da população.

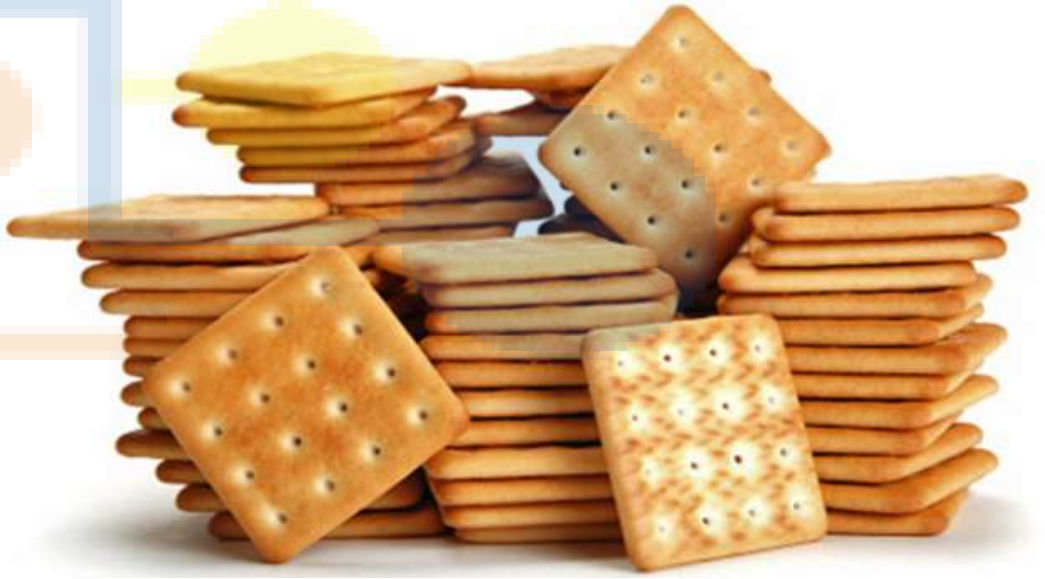
Amostragem

A análise da quantidade de sal na amostra, permitiria fazer inferências sobre toda a população de biscoitos.



Amostragem

Trabalhando com dados representativos na amostra, podemos inferir o que está acontecendo na população como um todo.



Amostragem

Parabéns!!

Você acabou de ter a definição de
Estatística Inferencial

Amostragem

Amostragem: é a técnica, processo ou pesquisa que podem ser realizadas para obter uma amostra.

Amostragem

Amostragem: usa a coleta, organização, apresentação e análise dos dados como meio de estudar os parâmetros de uma população.

Amostragem

Amostragem: é a técnica que seleciona apenas **alguns** elementos da população para se obter uma amostra.

Amostragem

Censo: é a técnica que seleciona e avalia **todos** os elementos da população quando se realiza uma pesquisa.

Amostragem

E mesmo que fosse possível medir a população inteira, seria um **desperdício** de **tempo** e **dinheiro**.

Amostragem

Se a **amostra** for selecionada **corretamente** e a análise sobre ela for feita seguindo as **metodologias estatísticas**, esta informação pode ser usada para fazer uma avaliação **precisa** sobre a **população** inteira.

Amostragem

Entretanto, existem **riscos** envolvidos em tomar decisões baseadas em **amostragem.**

A **amostragem** pode ser exposta a erros, que podem levar a decisões incorretas.

E como veremos mais tarde neste capítulo, **nós podemos quantificar** a **probabilidade** destes **erros ocorrerem**.

Tipos de Amostragem

Seja Bem-Vindo(a)!

Tipos de Amostragem

Existem muitas opções disponíveis para coletar uma amostra de uma população.

Tipos de Amostragem

Os tipos básicos que estudaremos aqui são:

Amostragem Probabilística

Amostragem Não-Probabilística

Tipos de Amostragem

Amostragem Não-Probabilística

Amostragem Não-Probabilística é **subjetiva**, pois é influenciada pela pessoa que está conduzindo a pesquisa. Ela se baseia nas decisões pessoais do pesquisador.

Tipos de Amostragem

Amostragem Probabilística

Amostragem Probabilística é **objetiva**, pois **não** é influenciada pela pessoa que está conduzindo a pesquisa.

Tipos de Amostragem

Amostragem Probabilística

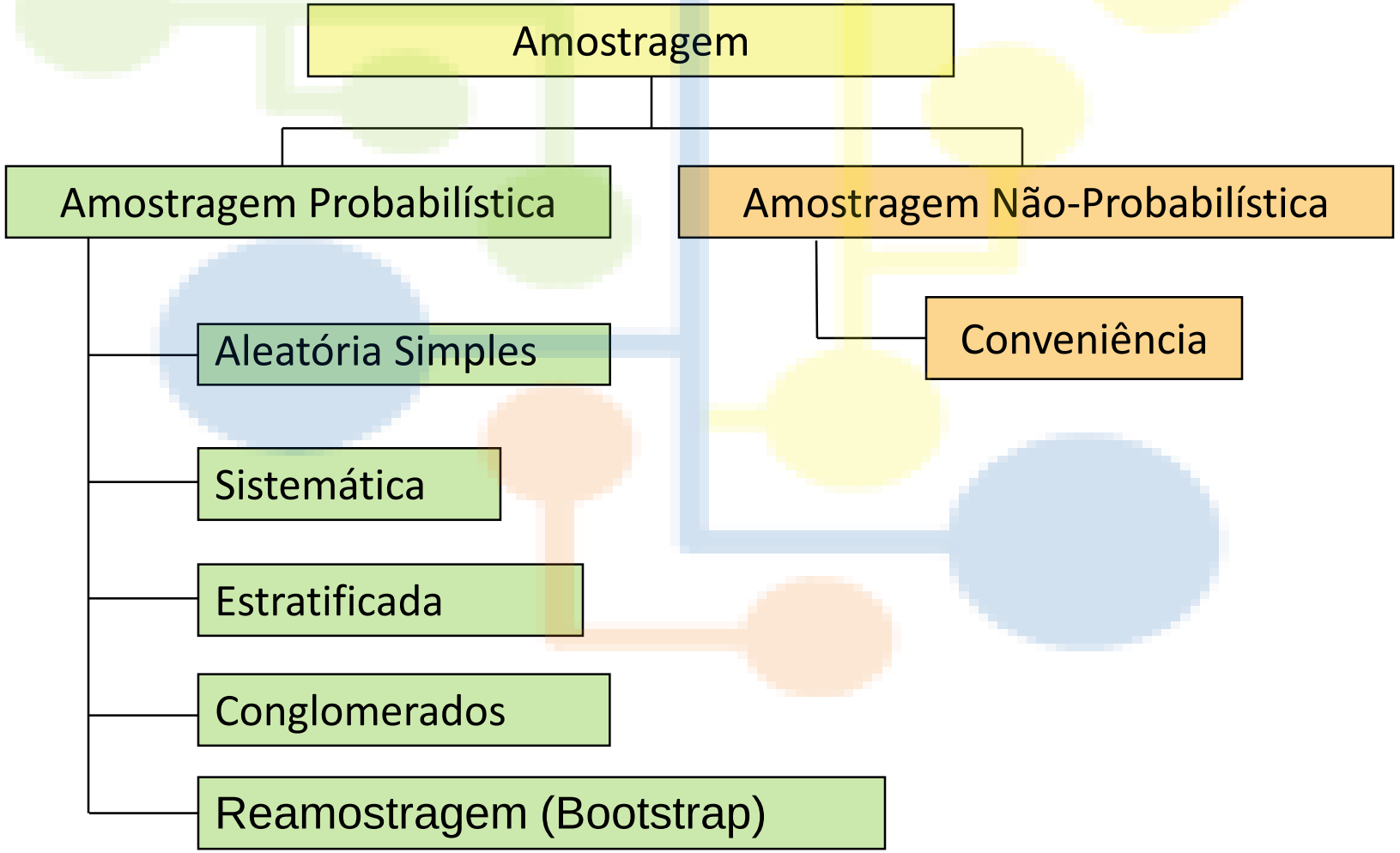
Os elementos da amostra são selecionados aleatoriamente e todos eles possuem probabilidade conhecida de serem escolhidos.

Tipos de Amostragem

Amostragem Probabilística

Tal seleção ocorre através de uma forma de sorteio não viciado, como o sorteio em uma urna ou por números gerados por computador.

Tipos de Amostragem



Amostragem Probabilística (Bootstrapping)

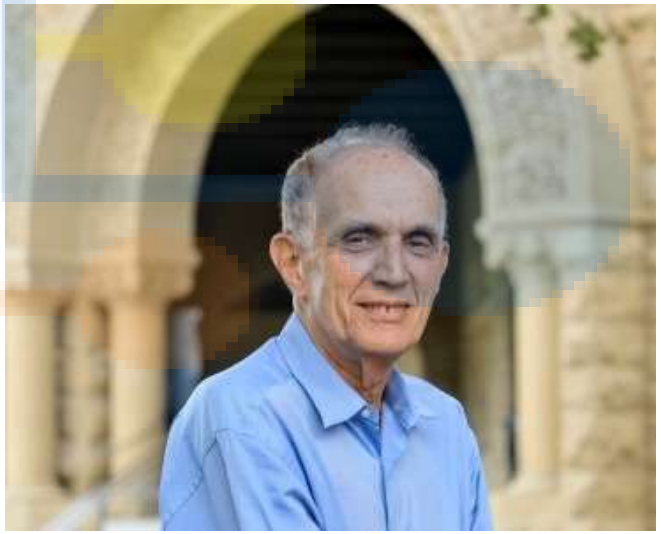
Seja Bem-Vindo(a)!

Amostragem Probabilística (Bootstrapping)

Reamostragem é uma **técnica estatística** em que várias amostras são repetidamente extraídas da **população**.

Amostragem Probabilística (Bootstrapping)

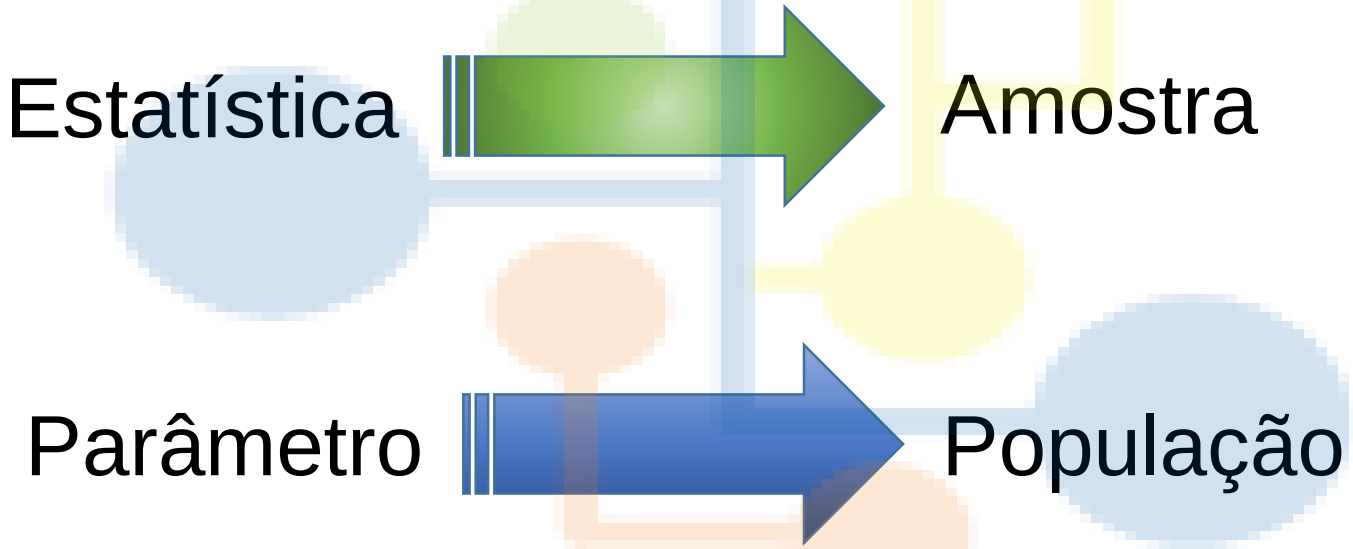
Um tipo específico de técnica de **Reamostragem** é conhecido como método **Bootstrap**, desenvolvido por Bradley Efron, membro do Departamento de Estatística da Universidade de Stanford na Califórnia nos EUA.



Amostragem Probabilística (Bootstrapping)

O **Método Bootstrap** consiste em usar software de computador para extrair diversas amostras (com reposição), para estimar determinados parâmetros da população, tais como média e proporção.

Amostragem Probabilística (Bootstrapping)



Amostragem Probabilística (Bootstrapping)

Exemplo

Amostragem Probabilística (Bootstrapping)

Vamos supor que uma rede de supermercados queira estimar a **proporção** de **clientes do sexo feminino** em suas lojas:

Amostragem Probabilística (Bootstrapping)

Randomicamente, o supermercado seleciona 100 operações (compras de produtos) e descobre que **58%** foram realizadas por clientes mulheres.

Amostragem Probabilística (Bootstrapping)

Agora sabemos que a proporção de clientes mulheres na amostra é de **58%** (58/100).

Erros de Amostragem

Seja Bem-Vindo(a)!

Erros de Amostragem

Parâmetros

Valores que descrevem características da **população**, como **média** e **mediana** da **população.**

Estatísticas

Valores calculados a partir da **amostra**, como **média** e **mediana** da **amostra.**

Erros de Amostragem

Como as **Estatísticas** são calculadas a partir da **amostra**, que é uma parte da população, não seria razoável esperar que a média dos dados da amostra fosse igual a média dos dados da população?

Erros de Amostragem

A **diferença** entre estes 2 valores é chamada “**erro de amostragem da média da amostra**”.

Erros de Amostragem

$$\text{Erro de amostragem} = x - \mu$$

Onde:

x = média da amostra

μ = média da população

Erros de Amostragem

Sem dúvida **amostragem** é uma técnica **fabulosa**, que nos permite obter informações sobre uma população inteira, analisando apenas uma porção dos dados.

O Erro Amostral é a diferença entre um resultado amostral e o verdadeiro resultado populacional.

Erros de Amostragem

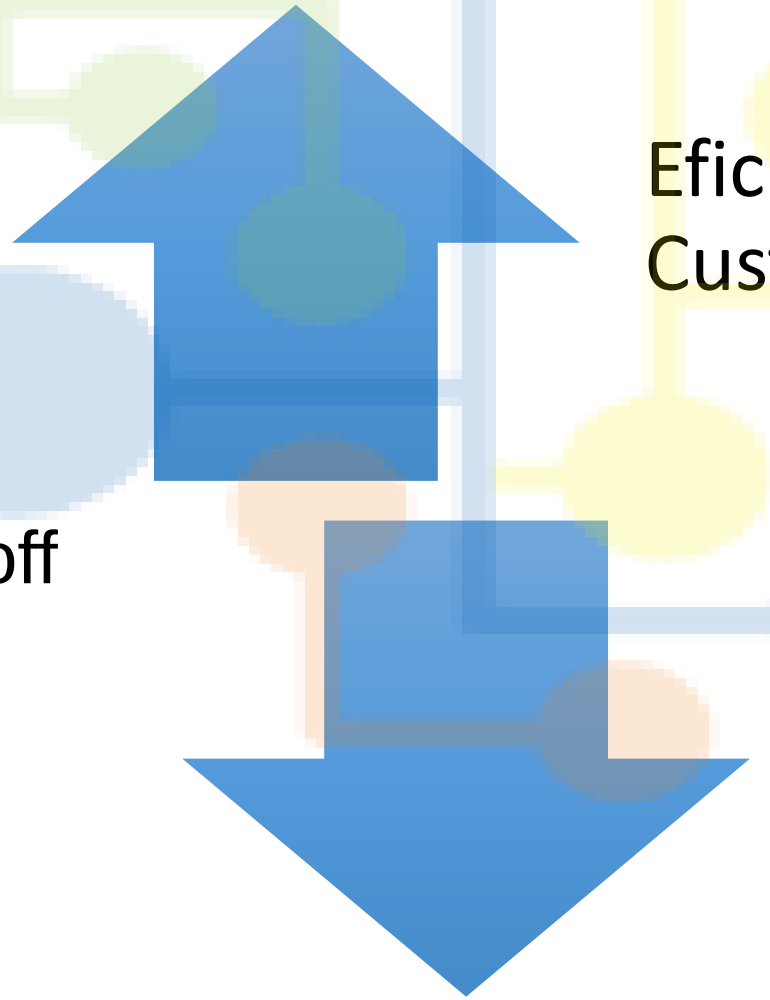
Erros de amostragem podem diferir de uma amostra para outra e podem ser **positivos** ou **negativos**.

Erros de Amostragem

Como regra geral, quanto **maior** o tamanho da **amostra**, **menor** será o **erro** de **amostragem**.

Erros de Amostragem

Agora temos um trade-off (fazer uma escolha)



Eficiência , Tempo e Custo

Quanto menor a amostragem maior possibilidade de erros

Erros de Amostragem

E agora talvez você fique mais claro porque o **Big Data** é tão fabuloso. Com mais dados a nossa disposição e recursos tecnológicos, podemos coletar e utilizar cada vez mais dados e conduzir testes estatísticos cada vez mais precisos.

Erros de Amostragem

Os conceitos da Estatística existem há muito tempo, mas nunca houve um volume de dados tão grande, como o que vemos atualmente.

Teorema do Limite Central

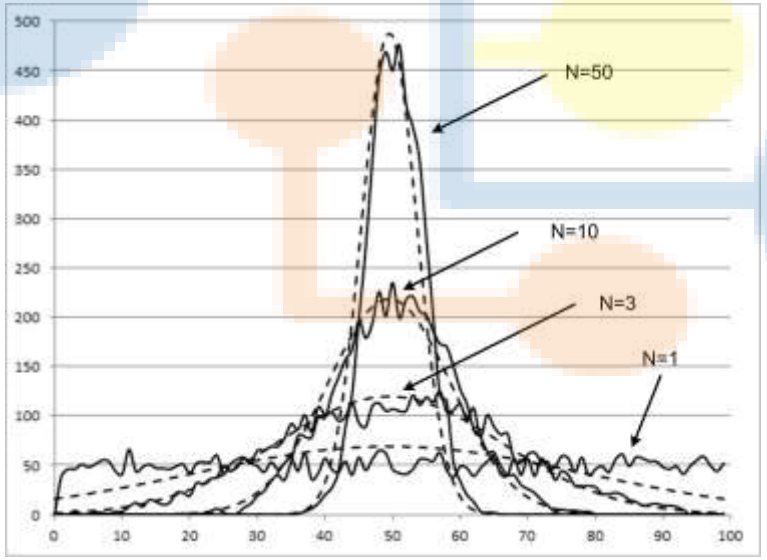
Seja Bem-Vindo(a)!

Teorema do Limite Central

O **Teorema do Limite Central** é um importante conceito da estatística e a demonstração de muitos outros teoremas estatísticos dependem dele, tanto que é conhecido como a ”mãe de todos os teoremas” e realmente merece sua atenção.

Teorema do Limite Central

Esse teorema afirma que quando o tamanho da amostra aumenta, a distribuição amostral da sua média aproxima-se cada vez mais de uma distribuição normal.

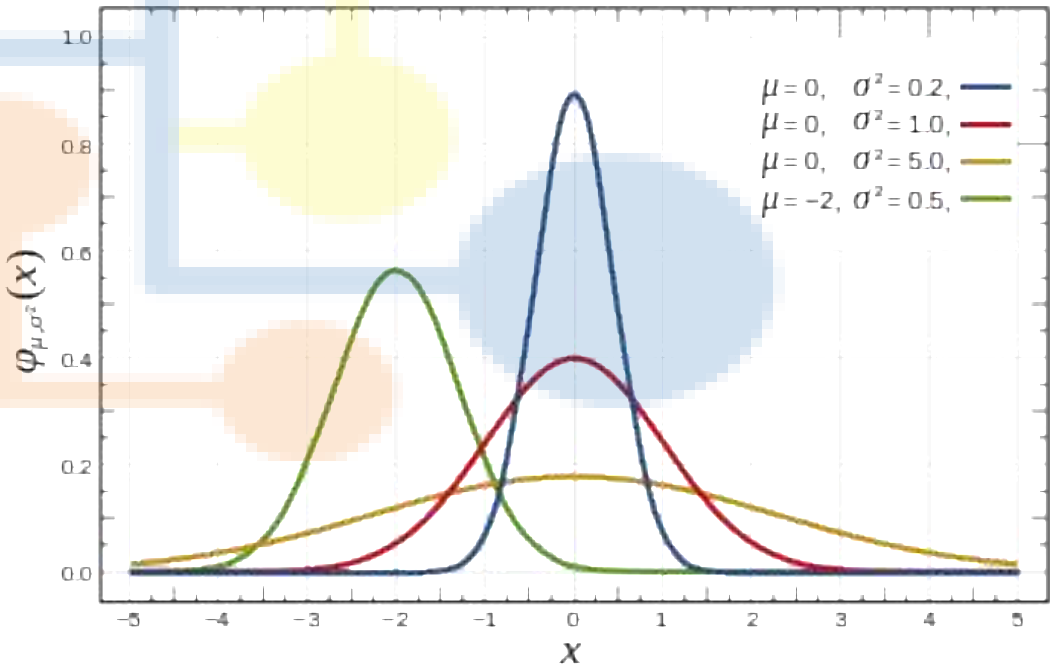


Teorema do Limite Central

Este teorema é fundamental na teoria da inferência estatística e sem este teorema, provavelmente a estatística não teria avançado como a ciência que é hoje.

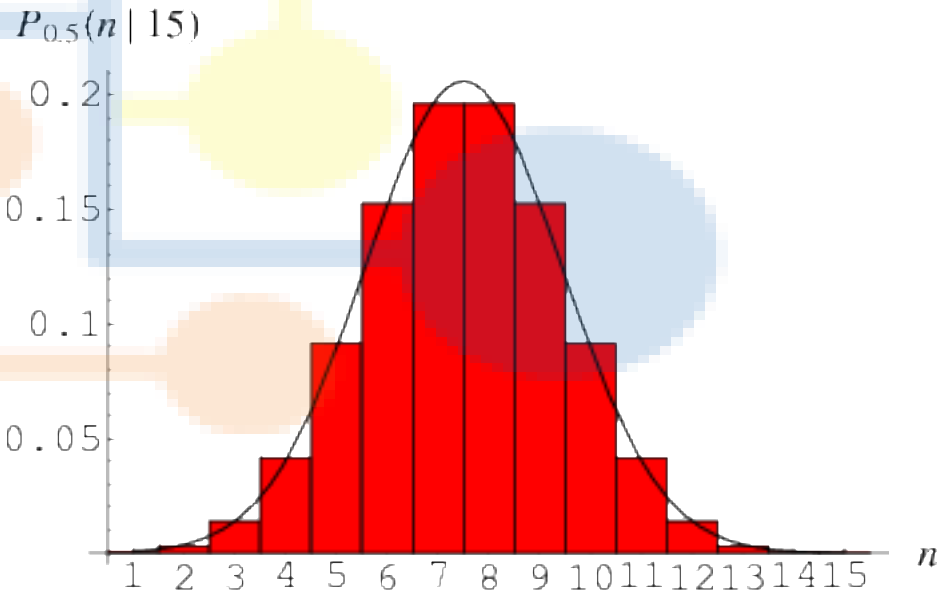
Teorema do Limite Central

Como já vimos, uma variável aleatória pode ter uma distribuição, possuindo uma média **μ (Mu)** e um desvio padrão **σ (Sigma)**.

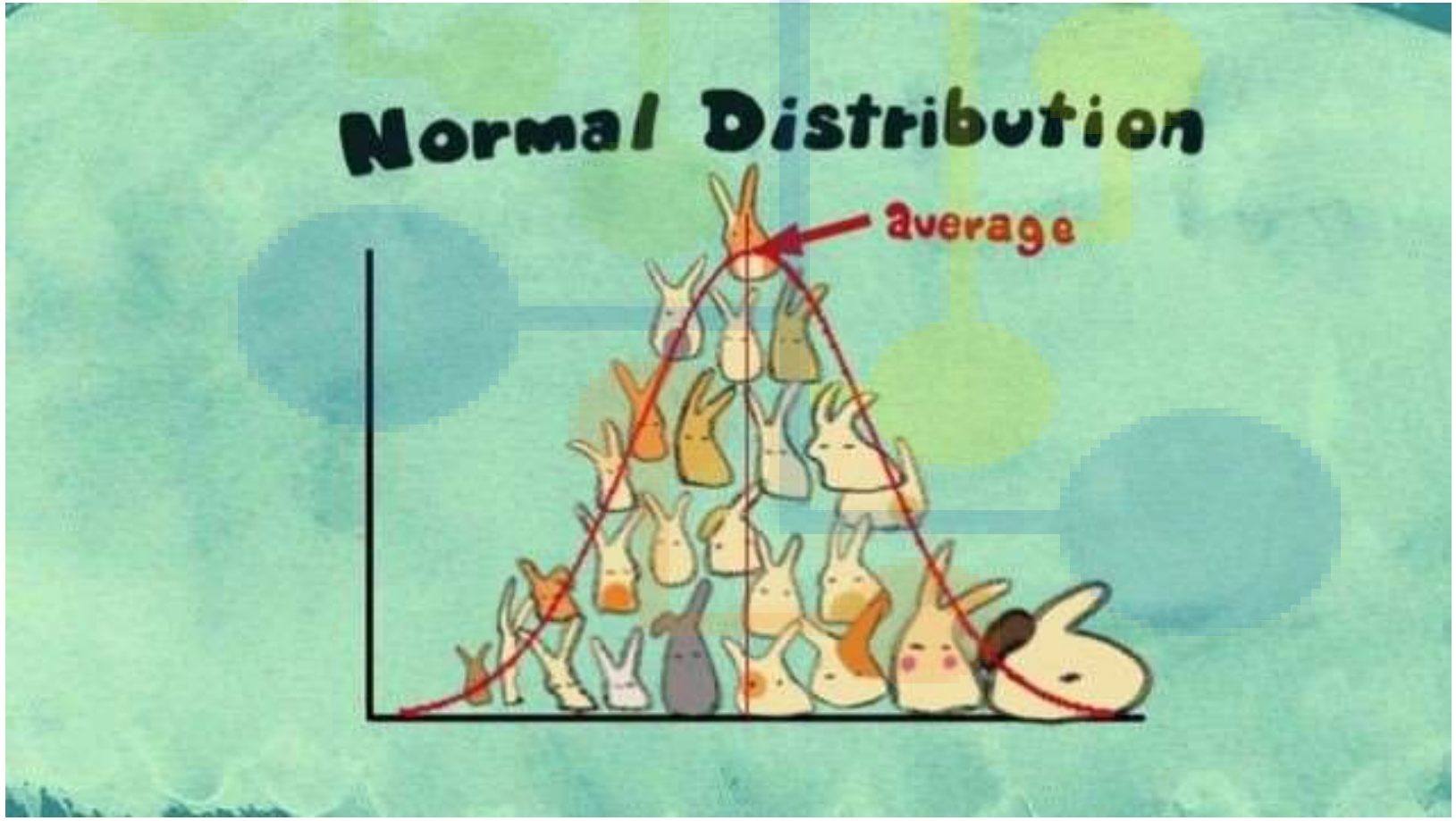


Teorema do Limite Central

À medida que o tamanho '**n**' da amostra aumenta, a distribuição das médias amostrais tende a uma distribuição normal.



Teorema do Limite Central



Teorema do Limite Central

Para **n >= 30**, a distribuição das médias amostrais pode ser aproximada satisfatoriamente por uma distribuição normal.

A **média das médias amostrais** será a média populacional.

Teorema do Limite Central

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**'.

Teorema do Limite Central

De acordo com o **Teorema do Limite Central**, médias amostrais de amostras suficientemente grandes, retiradas de qualquer população, terão uma distribuição normal.

Teorema do Limite Central

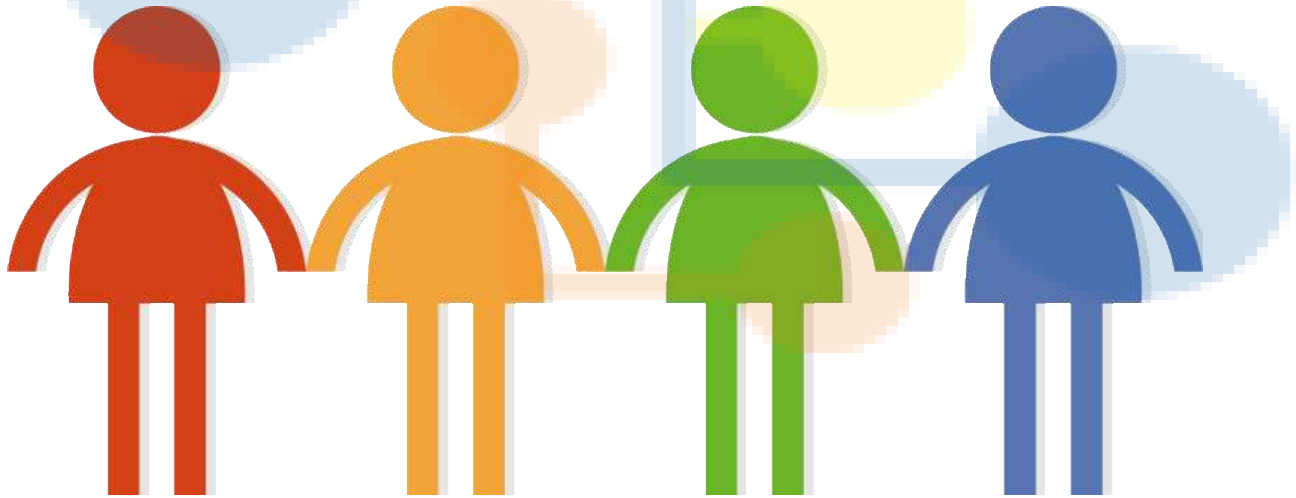
Se a população seguir uma **distribuição normal** de probabilidade, as **médias amostrais** também terão **distribuição normal**, independente do tamanho das amostras.

Teorema do Limite Central – Exemplo

Seja Bem-Vindo(a)!

Teorema do Limite Central - Exemplo

Imagine uma população de 4 pessoas, sendo então N = 4.



Teorema do Limite Central - Exemplo

A variável em consideração é a idade dos indivíduos e vamos chamar esta variável de **x.** Vamos imaginar a idade dos indivíduos (**x**) sendo 18, 20, 22 e 24 anos.

Teorema do Limite Central - Exemplo

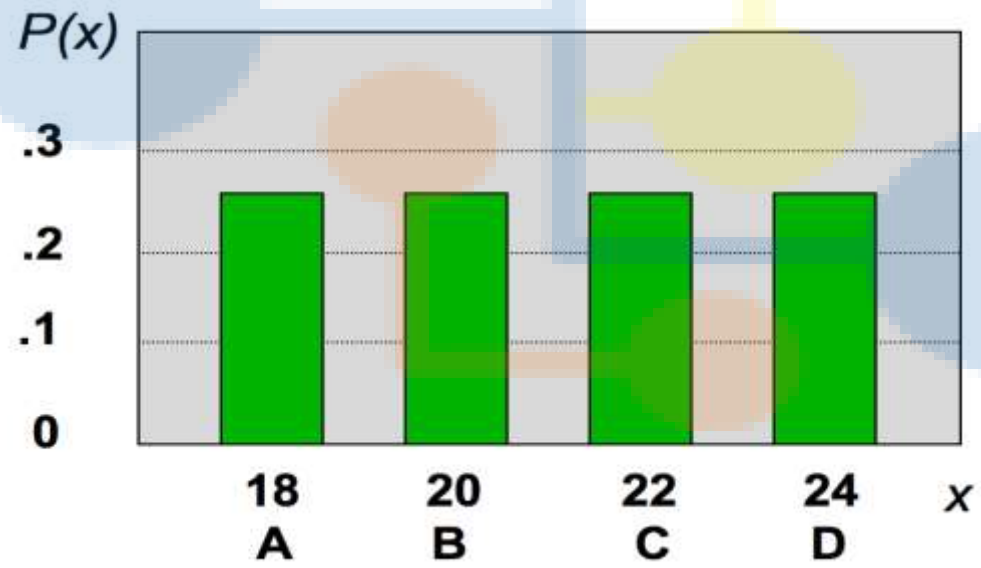
- População N = 4
- Variável aleatória x
- Valores de x:
  Idades = 18, 20, 22 e 24 anos

Vamos calcular a média e o desvio padrão desta população:

Média μ = 21
Desvio padrão σ = 2.236

Teorema do Limite Central - Exemplo

A distribuição de probabilidade desta população é uniforme.



Teorema do Limite Central - Exemplo

Obs 1	Observação 2			
	18	20	22	24
18	18,18	18,20	18,22	18,24
20	20,18	20,20	20,22	20,24
22	22,18	22,20	22,22	22,24
24	24,18	24,20	24,22	24,24

Média =(18+18)/2

16 amostras possíveis (amostragem com reposição)

16 médias amostrais

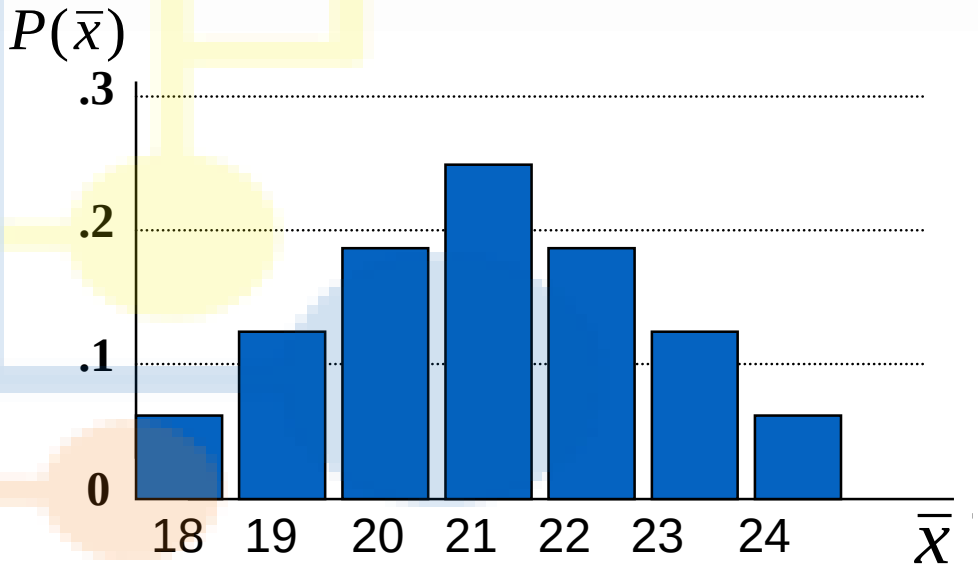
Obs 1	Observação 2			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

Teorema do Limite Central - Exemplo

16 médias amostrais

Obs 1	Observação 2			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

Distribuição amostral da média, $n = 2$



Distribuição Normal

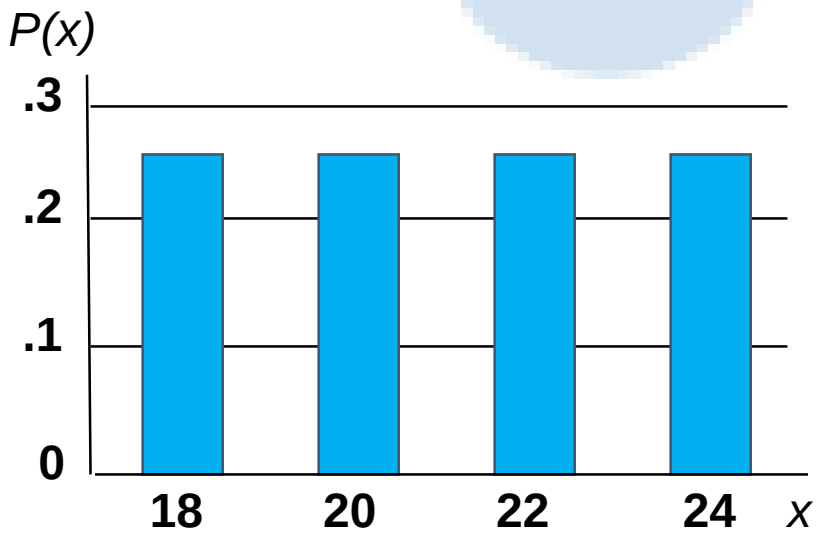
Teorema do Limite Central - Exemplo

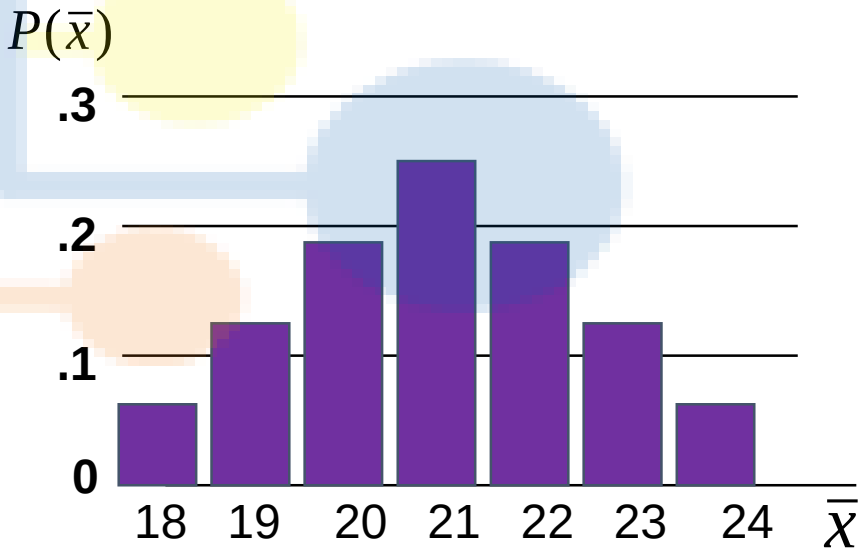
População
$N = 4$

$\mu = 21 \quad \sigma = 2.236$

Distribuição amostral da média, $n = 2$

$\mu_{\bar{x}} = 21 \quad \sigma_{\bar{x}} = 1.58$





A Importância do Tamanho da Amostra no Teorema do Limite Central

Seja Bem-Vindo(a)!

A Importância do Tamanho da Amostra no Teorema do Limite Central

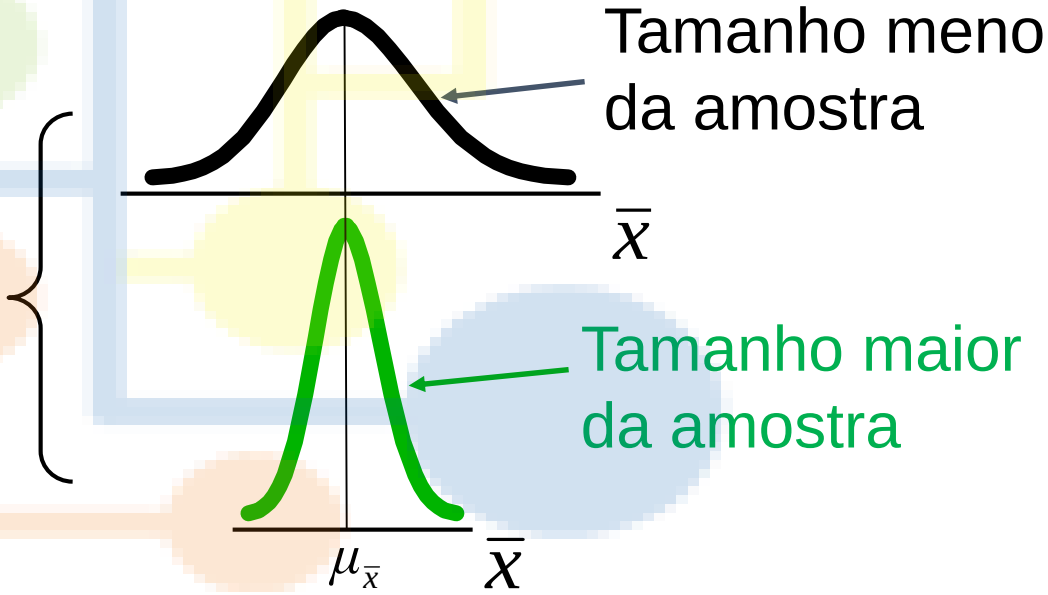
O **tamanho da amostra** tem uma função importante no Teorema do Limite Central.

A Importância do Tamanho da Amostra no Teorema do Limite Central

À medida que aumenta o tamanho das nossas amostras, o erro padrão da média se torna menor, o que reduz o erro de amostragem.

A Importância do Tamanho da Amostra no Teorema do Limite Central

Aumentando o tamanho da amostra, reduz o erro padrão.



A Importância do Tamanho da Amostra no Teorema do Limite Central

1ª Regra – Para qualquer população

O valor médio de todas as médias de amostras possíveis, a partir de um dado tamanho da população, é igual a média da população.

$$\mu_{\bar{x}} = \mu$$

A Importância do Tamanho da Amostra no Teorema do Limite Central

2ª Regra – Para qualquer população

O desvio padrão das médias das amostras de tamanho *n*, é igual ao desvio padrão da população dividido pela raiz quadrada do tamanho da amostra.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

A Importância do Tamanho da Amostra no Teorema do Limite Central

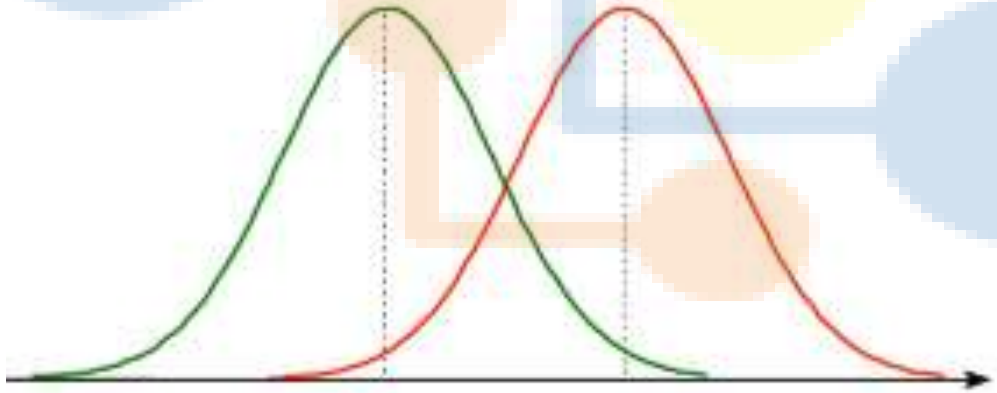
Mas afinal, o que há de extraordinário no Teorema do Limite Central?

A Importância do Tamanho da Amostra no Teorema do Limite Central

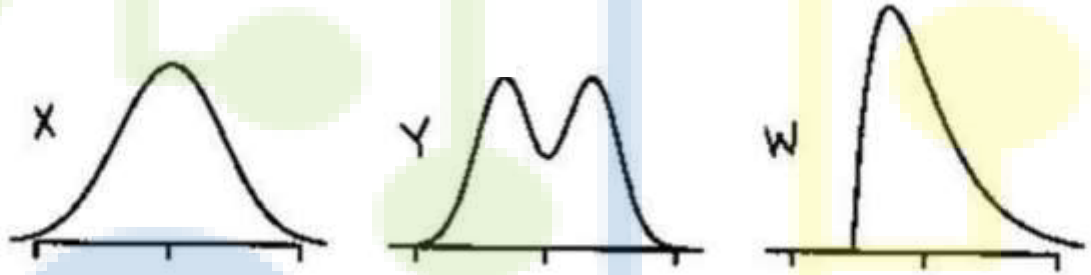
Ele nos diz que **qualquer** que seja a forma da distribuição original, suas **médias** resultam em uma **distribuição normal.**

A Importância do Tamanho da Amostra no Teorema do Limite Central

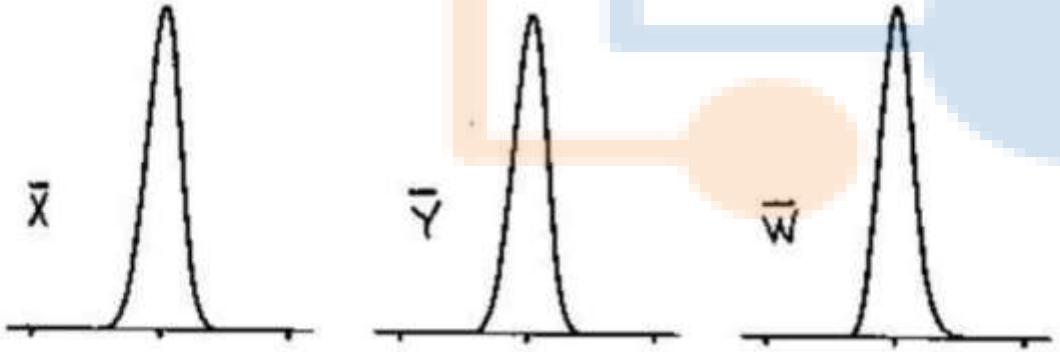
Esse teorema possibilita medir o quanto sua média amostral irá variar, sem ter que pegar outra média amostral para fazer a comparação.



A Importância do Tamanho da Amostra no Teorema do Limite Central



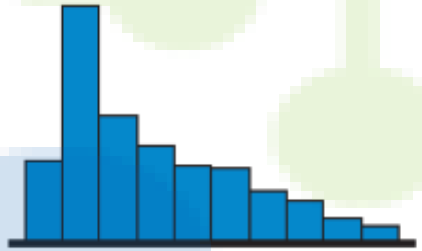
Todas as 3 densidades acima têm a mesma média e desvio padrão, apesar de suas formas diferentes. Mas as distribuições das médias das amostras, abaixo, são praticamente idênticas.



A Importância do Tamanho da Amostra no Teorema do Limite Central

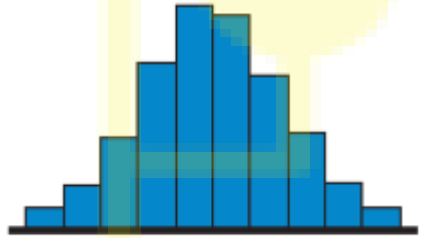
O formato da distribuição da população afetará o formato da distribuição da amostra, assim como…

Inclinada à direita



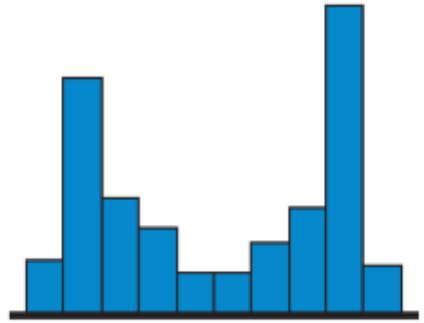
População 1

Normal

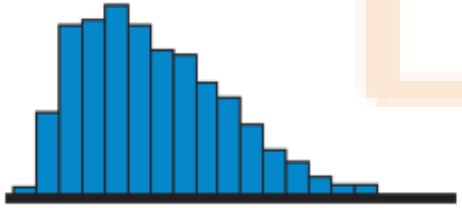


População 2

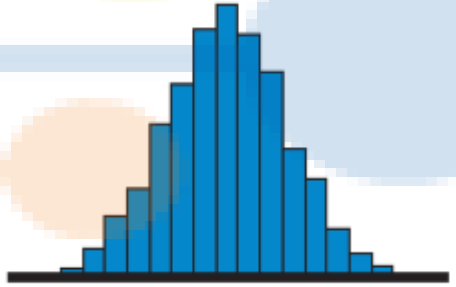
Em forma de U



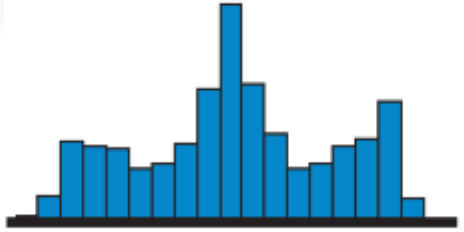
População 3



Média da Amostra n=2



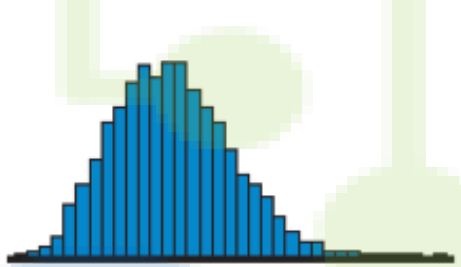
Média da Amostra n=2



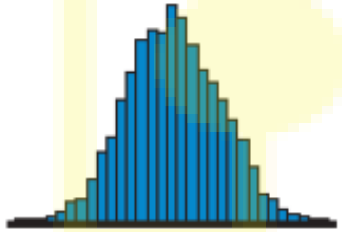
Média da Amostra n=2

A Importância do Tamanho da Amostra no Teorema do Limite Central

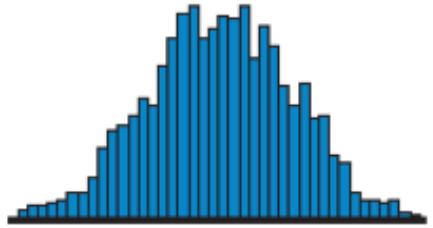
…o tamanho da amostra.



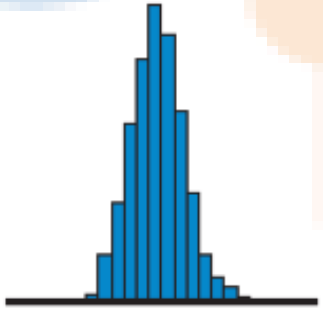
Média da Amostra n=5



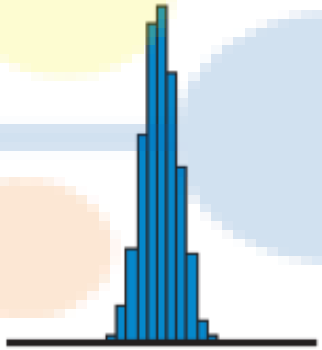
Média da Amostra n=5



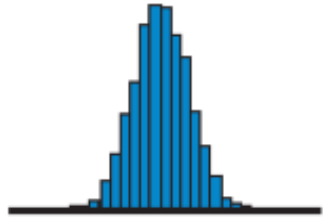
Média da Amostra n=5



Média da Amostra n=30



Média da Amostra n=30



Média da Amostra n=30

A Importância do Tamanho da Amostra no Teorema do Limite Central

Colocando o Teorema do Limite Central para Trabalhar

A Importância do Tamanho da Amostra no Teorema do Limite Central

Vamos agora trabalhar com o Teorema do Limite Central.

A Importância do Tamanho da Amostra no Teorema do Limite Central

Suponhamos que as pessoas de uma determinada região viajem **12.000 quilômetros** por ano, com um desvio padrão de **2.580 quilômetros.**

A Importância do Tamanho da Amostra no Teorema do Limite Central

Qual a probabilidade de que uma amostra aleatoriamente selecionada de **36 motoristas**, viaje mais de **12.500 quilômetros por ano**?

A Importância do Tamanho da Amostra no Teorema do Limite Central

Qual a probabilidade de que uma amostra aleatoriamente selecionada de **36 motoristas**, viaje mais de **12.500 quilômetros por ano**?

$$P(\bar{x} > 12,500) = ?$$

A Importância do Tamanho da Amostra no Teorema do Limite Central

Como o tamanho da amostra é grande o suficiente (**n >= 30**), podemos usar o Teorema do Limite Central para encontrar a resposta.

A Importância do Tamanho da Amostra no Teorema do Limite Central

Aplicando as regras vistas anteriormente, teremos:

$$\mu_{\bar{x}} = \mu = 12,000$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2,580}{\sqrt{36}} = 430$$

A Importância do Tamanho da Amostra no Teorema do Limite Central

Vamos usar o **Escore$_z$** para nos ajudar:

$$z_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

Onde:

$\bar{x}$ = Média da amostra

$\mu_{\bar{x}}$ = Média das médias amostrais

$\sigma_{\bar{x}}$ = Desvio padrão da média

$$z_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{12{,}500 - 12{,}000}{430} = 1.16$$

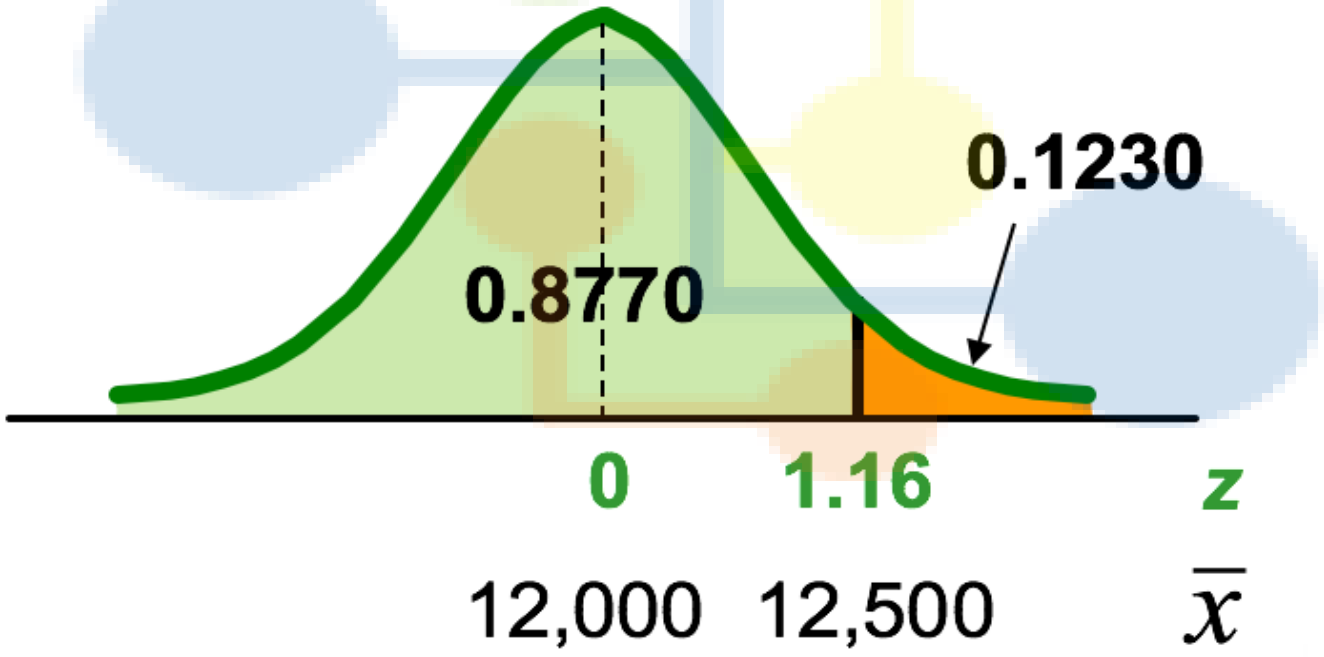
A Importância do Tamanho da Amostra no Teorema do Limite Central

Vamos usar o **Escore$_z$** para nos ajudar:

$$P(\bar{x} > 12{,}500) = P(z_{\bar{x}} > 1.16)$$
$$= 1 - P(z_{\bar{x}} \leq 1.16)$$
$$= 1 - 0.8770$$
$$= 0.1230$$

A Importância do Tamanho da Amostra no Teorema do Limite Central

Resposta: A probabilidade de selecionarmos uma amostra aleatória de 36 motoristas que viajem 12.500 quilômetros, é de 12.3%.



Escore z

Seja Bem-Vindo(a)!

Escore z

O **Escore$_z$** identifica o número de *desvios padrão* que um determinado valor possui de distância em relação à média do conjunto de dados ao qual faz parte.

Escore z

Pense no **Escore$_z$**, como uma simples conversão de um valor para uma outra unidade específica, a fim de chegar a conclusões.

Escore z

O **Escore$_z$** em si, não possui unidade.
É apenas um número.

Escore z

Exemplo

Escore z

A tabela abaixo, mostra o número de calorias de sanduíches. Vejamos:

Hamburguer	Restaurante	Calorias
Cheeseburguer	McDonalds's	300
Big Mac	McDonalds's	430
Whopper	Burger King	540
Double Cheeseburguer	Bob's	670
Chicken Burger	McDonalds's	780
Bacon Burger	Bob's	840
Quarteirão	McDonalds's	1.230
Mega Burger	Burger King	1.420
Média da amostra		776,30
Desvio padrão da amostra		385,10

Escore z

Hamburguer	Restaurante	Calorias
Cheeseburguer	McDonalds's	300
Big Mac	McDonalds's	430
Whopper	Burger King	540
Double Cheeseburguer	Bob's	670
Chicken Burger	McDonalds's	780
Bacon Burger	Bob's	840
Quarteirão	McDonalds's	1.230
Mega Burger	Burger King	1.420
Média da amostra		776,30
Desvio padrão da amostra		385,10

Calculando o **Escore$_z$** para o Mega Burger:

$$Z_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

Escore$_z$ = (x – x1) / s = (1.420 – 776,30) / 385,10 = **1.67**

Escore z

O que 1.67 significa?

Significa que as calorias do **Mega Burger**, são 1.67 desvio padrão acima da média.

Escore z

Hamburguer	Restaurante	Calorias
Cheeseburguer	McDonalds's	300
Big Mac	McDonalds's	430
Whopper	Burger King	540
Double Cheeseburguer	Bob's	670
Chicken Burger	McDonalds's	780
Bacon Burger	Bob's	840
Quarteirão	McDonalds's	1.230
Mega Burger	Burger King	1.420
Média da amostra		776,30
Desvio padrão da amostra		385,10

Agora vamos calcular o $Escore_z$ do **Big Mac**:

$$Escore_z = (x - x1) / s = (430 - 776,30) / 385,10 = \mathbf{-0.90}$$

Escore z

O valor negativo, indica que as calorias do **Big Mac**, estão abaixo da média, neste caso, exatamente 0.90 abaixo da média.

Escore z

O **Escore$_z$** terá sempre um dos 3 atributos:

- **Positivo** – valores acima da média
- **Negativo** – valores abaixo da média
- **Zero** – valores iguais a média

Escore z

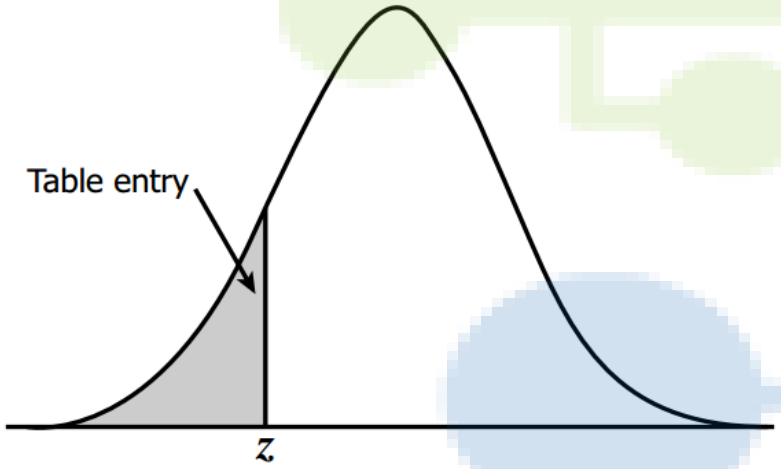
O **Escore$_z$** também possui uma característica importante: ele nos ajuda a identificar os outliers (valores extremos) dos nossos dados.

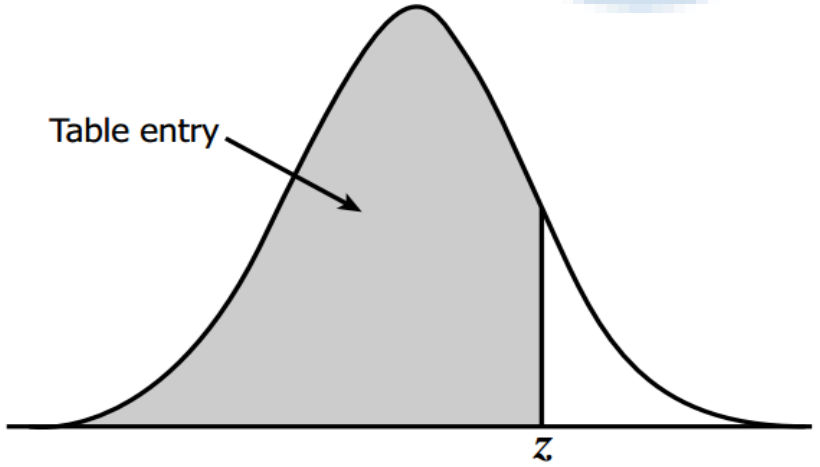
Escore z

Valores de dados que possuem **Escore$_z$** acima de **+3** ou abaixo de **-3** são classificados como outliers.

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

Escore z





z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
–3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
–2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
–2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
–2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
–2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
–2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
–2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
–2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
–2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
–2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
–2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
–1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
–1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
–1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
–1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
–1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
–1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
–1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
–1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
–1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
–1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
–0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
–0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
–0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
–0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
–0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
–0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
–0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
–0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
–0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
–0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Intervalo de Confiança

Seja Bem-Vindo(a)!

Intervalo de Confiança

Imagine um arqueiro atirando em um alvo.

Intervalo de Confiança

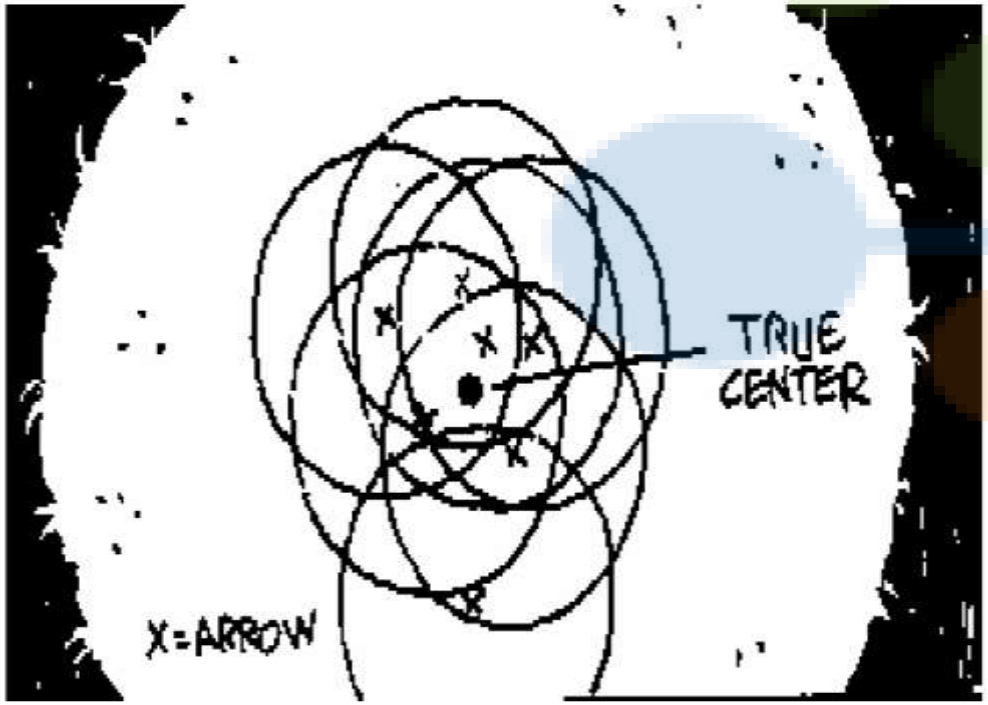
Suponha que ele acerte no centro do raio de 10 cm 95% das vezes.
Ou seja, ele erra apenas uma vez a cada 20 tentativas.

Intervalo de Confiança

Conhecendo o nível de habilidade do arqueiro, o estatístico desenha um círculo com 10 cm de raio ao redor da flecha.

Ele tem **95% de confiança** de que o círculo inclui o centro do alvo.

Intervalo de Confiança



O estatístico raciocinou que se desenhasse círculos com 10 cm de raio ao redor de muitas flechas, os seus círculos incluiriam o centro do alvo em **95%** dos casos.

Intervalo de Confiança

Resumindo, o estatístico coletou as amostras e construiu um intervalo de confiança, os seus círculos incluíram **95%** dos casos.

Intervalo de Confiança

Como Melhorar a Confiança?

Intervalo de Confiança

Considerando o exemplo do arqueiro atirando no alvo.

Intervalo de Confiança

Como podemos aumentar a confiança?

Intervalo de Confiança

Nesse exemplo temos duas formas de aumentar o **Nível de Confiança:**

Aumentando o tamanho do círculo	**Melhorando a mira do arqueiro**
Isso equivale a alargar o intervalo de confiança (de 95% para 99%, por exemplo). Quanto maior o intervalo, mais certo você estará de que o valor desejado encontra-se no intervalo.	Isso equivale a aumentar o número de observações na amostra.

Intervalo de Confiança

Entendi o que é o Intervalo de Confiança, mas o que é exatamente o **Nível de Confiança?**

Nível de Confiança

Seja Bem-Vindo(a)!

Nível de Confiança

Nível de confiança é a probabilidade!

$$1 - \alpha$$

One alfa é o que chamamos de Nível de Significância.

Nível de Confiança

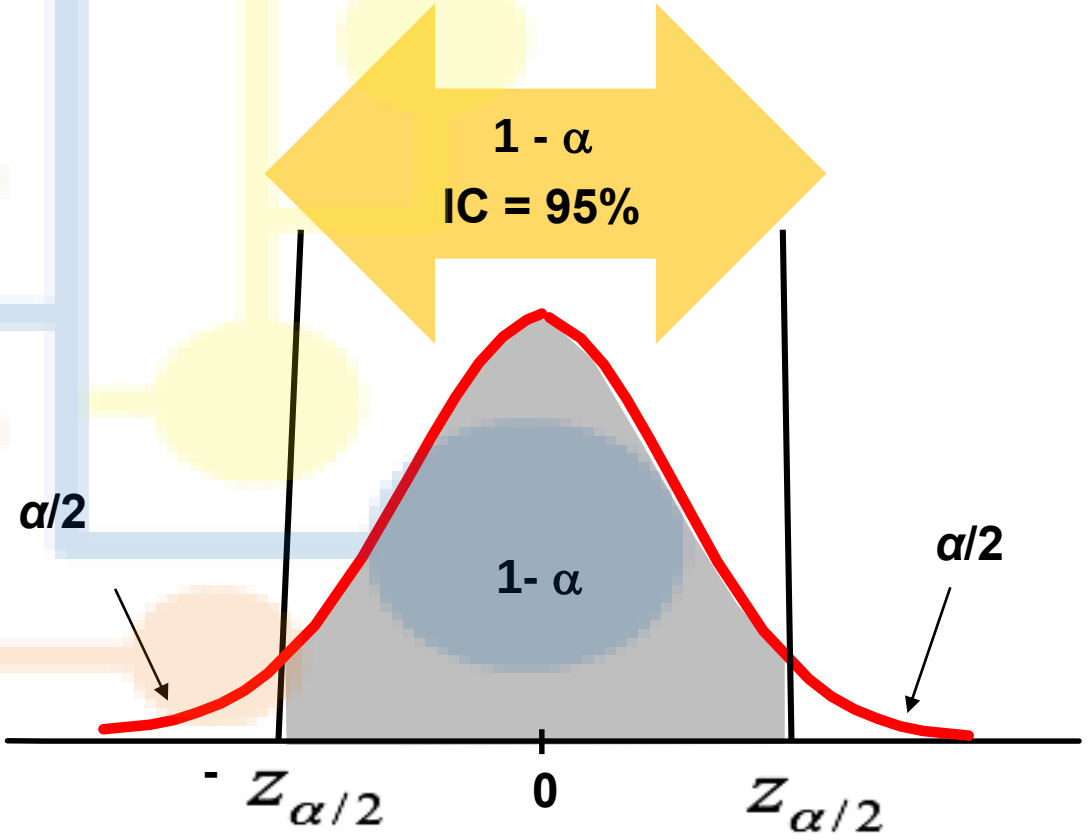
O **Nível de Confiança** é expresso percentualmente e por isso usamos:

100 (1 - α)

Nível de Confiança

O **Nível de Confiança** é expresso percentualmente e por isso usamos:

100 (1 - α)



Nível de Confiança

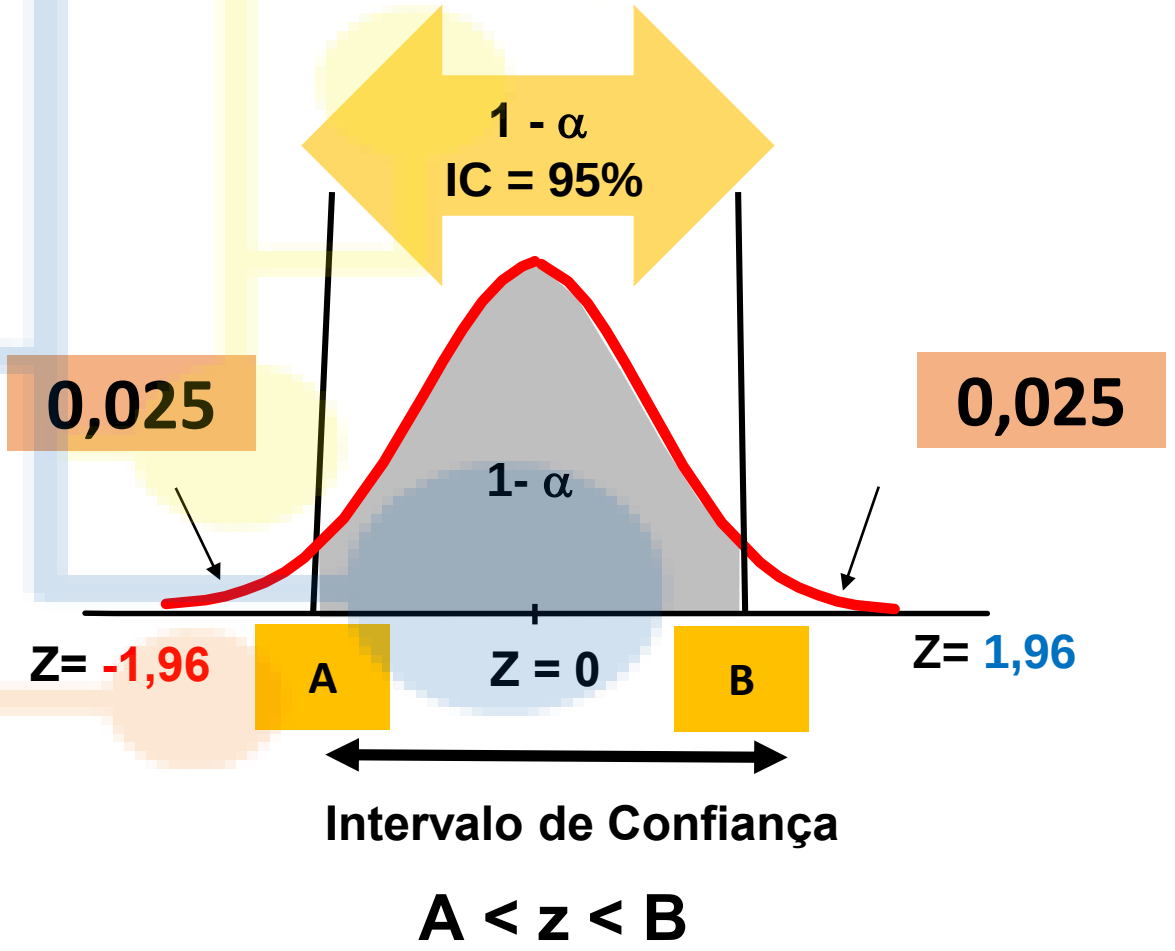
Normalmente utiliza-se **Nível de Confiança** (**NC**) de:

IC = 90%	IC = 95%	IC = 99%
$\alpha = 0,10$	$\alpha = 0,05$	$\alpha = 0,01$

Nível de Confiança

O **Intervalo de Confiança** consiste em um intervalo na **escala z (Escore z)** e está associado a um **NC.**

Ou seja, se coletarmos várias amostras e construirmos um intervalo de confiança para cada uma, a longo prazo, **95%** destes intervalos conteriam efetivamente a **média da população** μ.



Nível de Confiança

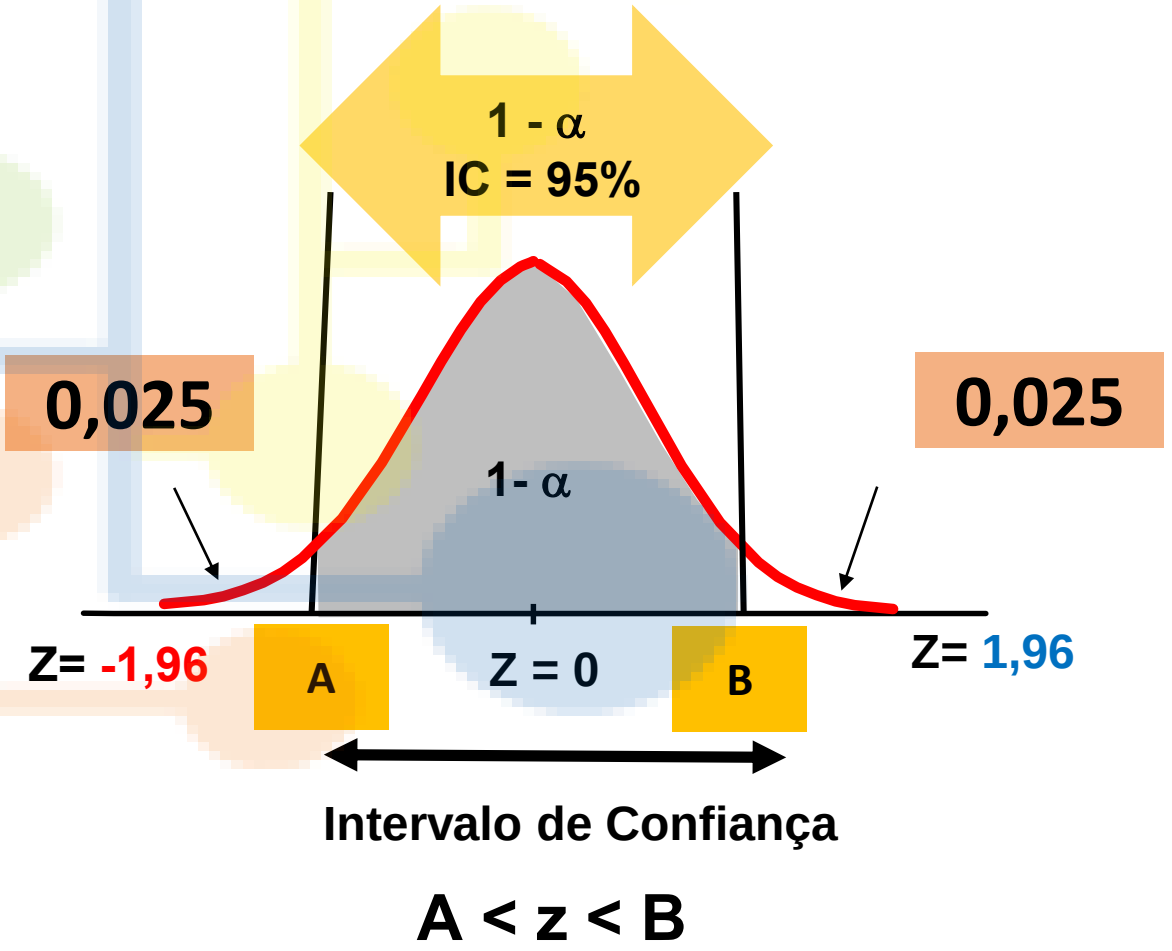
E qual a relação entre o Nível de Confiança e o Escore z?

Valor Crítico

Seja Bem-Vindo(a)!

Valor Crítico

O valor crítico **Z** (**Escore$_z$**) corresponde ao valor da fronteira da área $\alpha/2$ nas caudas direita ou esquerda da distribuição normal.



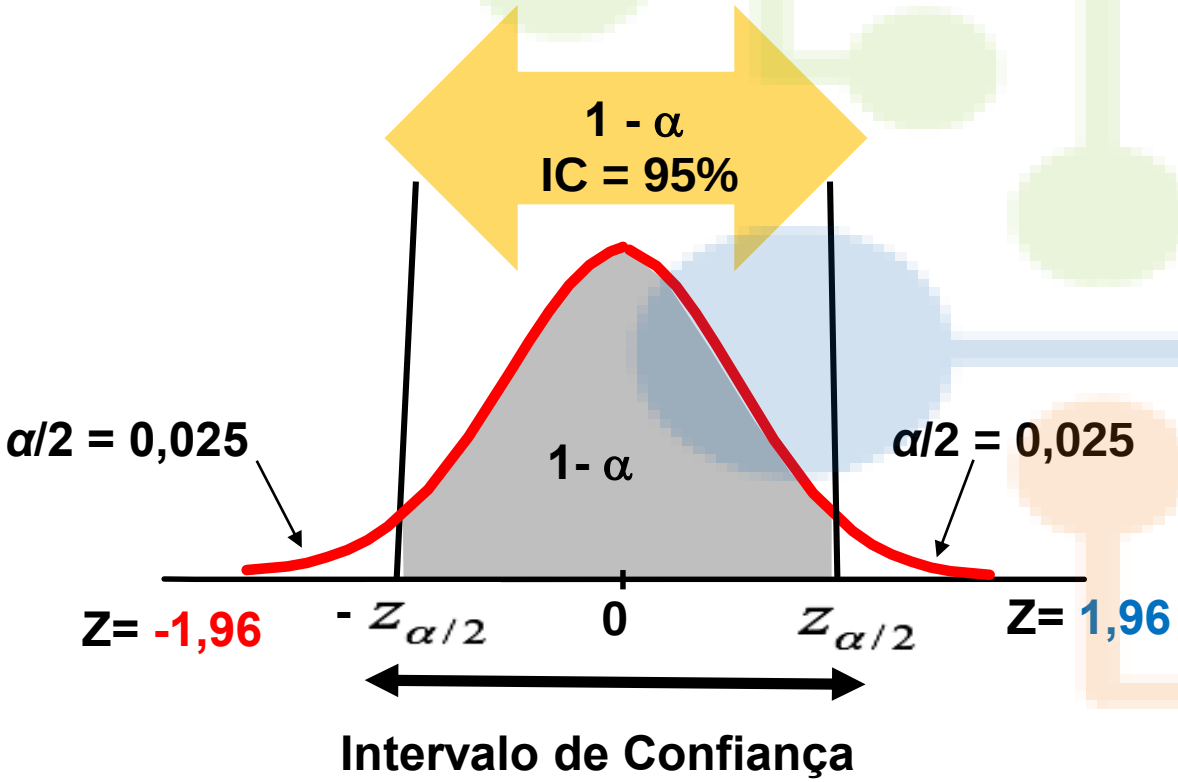
Valor Crítico

E por que isso é importante?

Valor Crítico

Pelo **Teorema do Limite Central**, sabemos que as médias amostrais **x** tendem a distribuir-se por uma **distribuição normal.**

Valor Crítico



A < z < B

Sendo **α/2** a área de cada extremo, há uma possibilidade de α da média amostral estar em um dos **2 extremos.**

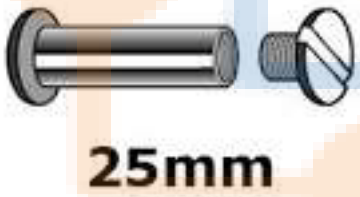
Pela regra do complemento, há uma probabilidade da média estar na região **não** sombreada.

Valor Crítico

Exemplo

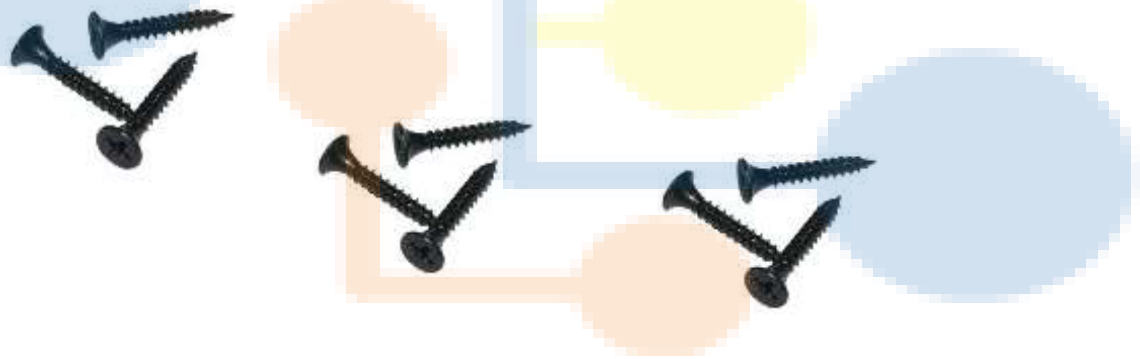
Valor Crítico

Uma fábrica de parafusos tem por especificação fabricá-los com diâmetro médio de 25 mm.



Valor Crítico

Para controle de seu processo, coleta-se ao longo de um dia de operação, 35 amostras de tamanho 30 (n = 30).



Valor Crítico

A média das médias das amostras e o desvio padrão de um determinado dia acusaram:

A média das médias das amostras = **25 mm**

Desvio padrão = **1,5 mm**

Valor Crítico

Considerando um grau de confiança de **99%**, calcule os valores críticos $Z_{\alpha/2}$ desta distribuição:

Grau (Intervalo) de Confiança	Nível de Significância (alfa)	Valor Crítico Z
99%	0,01	2,575

Valor Crítico

Considerando um grau de confiança de **99%**, calcule os valores críticos $Z_{\alpha/2}$ desta distribuição:

$$Z_{\alpha/2} = \frac{x - \bar{\bar{x}}}{\bar{s}} \Rightarrow \mathbf{2,575} = \frac{x - 25}{1,5} \Rightarrow 3,8625 = x - 25 \Rightarrow x_1 = \mathbf{28,86\ mm}$$

$$Z_{\alpha/2} = \frac{x - \bar{\bar{x}}}{\bar{s}} \Rightarrow \mathbf{-2,575} = \frac{x - 25}{1,5} \Rightarrow -3,8625\ x - 25 \Rightarrow x_2 = \mathbf{21,14\ mm}$$

Valor Crítico

Resposta:

Existe 99% de probabilidade do intervalo de **21,14 e 28,86 mm** conter a média populacional de diâmetro de parafuso.

Ou

A fábrica possui **99%** de chance de produzir lotes de peças com médias entre **21,14 mm e 28,86 mm**.

Valor Crítico

Como já sabemos, o **Intervalo de Confiança da Média** é um intervalo de estimativa em torno da **média** da **amostra**, que provê um range de valores no qual está a **média** da **população**.

Valor Crítico

De fato, a média da população raramente é conhecida, assim o **Intervalo de Confiança** é a **única** evidência que nós temos sobre a **média** da **população**.

Teste de Hipótese

Seja Bem-Vindo(a)!

Teste de Hipótese

No mundo da **Estatística**, uma **hipótese** é uma **suposição** sobre um **parâmetro** específico de uma população, tal como média, proporção ou desvio padrão.

Teste de Hipótese

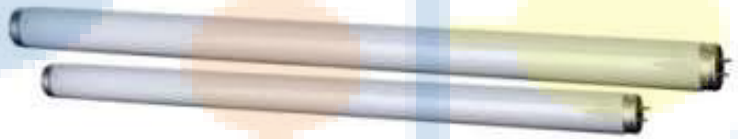
Podemos fazer uma suposição sobre o valor de um parâmetro da população, coletar uma amostra desta população, medir a amostra e atestar se a amostra suporta ou não a nossa suposição.

Teste de Hipótese

O mundo dos negócios é cheio de exemplos que lidam com testes de hipóteses!

Teste de Hipótese

Uma indústria de lâmpadas que fabrica lâmpadas fluorescentes afirma que suas lâmpadas usam 75% menos energia e duram 10 vezes mais que as lâmpadas comuns.



Um laboratório independente poderia testar essa afirmação com um **teste de hipótese**.

Teste de Hipótese

Um artigo recente de um grande jornal, afirmou que o excesso de tempo em redes sociais poderia afetar a capacidade intelectual das pessoas.

Um pesquisador poderia validar esta afirmação usando um teste de hipótese.

Teste de Hipótese

Antes da crise, os bancos cobravam em média R$40 em taxas administrativas de contas correntes de pessoas físicas.

Um órgão regulador do governo poderia testar a hipótese de os bancos estarem cobrando mais de R$40 em taxas.

Teste de Hipótese

Teste de Hipótese é um dos procedimentos estatísticos mais usados atualmente, com muitas aplicações em Data Science, como em Testes A/B aplicados por equipes de Marketing por exemplo.

Teste de Hipótese

O objetivo de um teste de hipótese é decidir se determinada afirmação sobre um parâmetro populacional é ou não apoiada pela evidência obtida de dados amostrais.

Teste de Hipótese

Cada teste de hipótese tem uma hipótese nula e uma hipótese alternativa, representados por:

H_0 - Hipótese nula

H_A - Hipótese alternativa

Teste de Hipótese

A **hipótese nula** representa o *status quo*, ou seja, comprovar uma suposição ou afirmação.

A **hipótese nula** valida que o parâmetro da população seja

<= ou => a um específico valor.

A **hipótese nula** é para ser verdadeira, a menos que seja comprovada por uma evidência contrária.

Teste de Hipótese

A **hipótese alternativa** representa o oposto da hipótese nula.

Você precisa ser **cuidadoso** ao definir a **hipótese nula** e a **hipótese alternativa**.

As decisões sobre as hipóteses vão depender da **natureza do teste** e da **pessoa** que o está conduzindo.

Teste de Hipótese – Aplicação

Seja Bem-Vindo(a)!

Teste de Hipótese - Aplicação

Exemplo I

Teste de Hipótese - Aplicação

Vamos assumir que usuários de internet ficam em média 56 segundos em uma web page.

Teste de Hipótese - Aplicação

Suponhamos que o propósito do teste seja determinar se a média da população é igual a um valor específico. Definiríamos assim nossas hipóteses:

$$H_0 : \mu = 56 \text{ segundos}$$

(status quo)

$$H_0 = 56 \text{ segundos}$$

Teste de Hipótese - Aplicação

A hipótese alternativa reflete a condição oposta, ou seja, que o tempo médio que os internautas ficam em uma web page **não** é igual a 56 segundos.

$$H_A : \mu \neq 56 \text{ segundos}$$

$$H_A \neq 56 \text{ segundos}$$

Teste de Hipótese - Aplicação

Uma declaração de hipótese pode somente ser usada com parâmetro da população (**tal como μ**), não com uma estatística de amostra (**tal como a média da amostra**).

Teste de Hipótese - Aplicação

O propósito do **teste de hipótese** é **atestar** uma **conclusão** sobre os parâmetros da população, sobre os quais **não** temos conhecimento completo.

Teste de Hipótese - Aplicação

Exemplo II

Teste de Hipótese - Aplicação

Vamos imaginar um fabricante de lâmpadas que afirma que desenvolveu um novo produto cujo tempo médio de vida **supera a média da indústria de 8.000 horas**.

Teste de Hipótese - Aplicação

Para testarmos esta afirmação, definimos o seguinte teste de hipótese:

$$H_0 : \mu \leq 8.000 \text{ horas}$$
$$H_A : \mu > 8.000 \text{ horas}$$

Teste de Hipótese - Aplicação

Perceba que a hipótese alternativa foi usada para representar a afirmação feita pelo fabricante. As **8.000** horas de tempo de vida em **média** é considerado ser verdadeiro (status quo) e por isso foi atribuído à **hipótese nula.**

Teste de Hipótese - Aplicação

A Lógica do Teste de Hipótese

Teste de Hipótese - Aplicação

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

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

Teste de Hipótese - Aplicação

Podemos fazer apenas **2** afirmações sobre a **hipótese nula**:

Rejeitar

Não Rejeitar

Teste de Hipótese - Aplicação

A razão pela qual estamos limitados a apenas 2 conclusões possíveis é que o teste de hipótese se baseia em "**provar contradições".**

Teste de Hipótese - Aplicação

Com isso, nós podemos apenas concluir que a **hipótese pode ser verdadeira**, mas não temos evidências suficientes para afirmar que a hipótese nula é realmente verdadeira.

Teste de Hipótese - Aplicação

Por conta desta limitação, **NUNCA** podemos aceitar a **hipótese nula.**

Teste de Hipótese - Aplicação

Podemos apenas dizer que:

Não há evidências suficientes para rejeitar a hipótese nula.

Teste de Hipótese - Aplicação

Qual das hipóteses devo escolher?

H_0 = nula

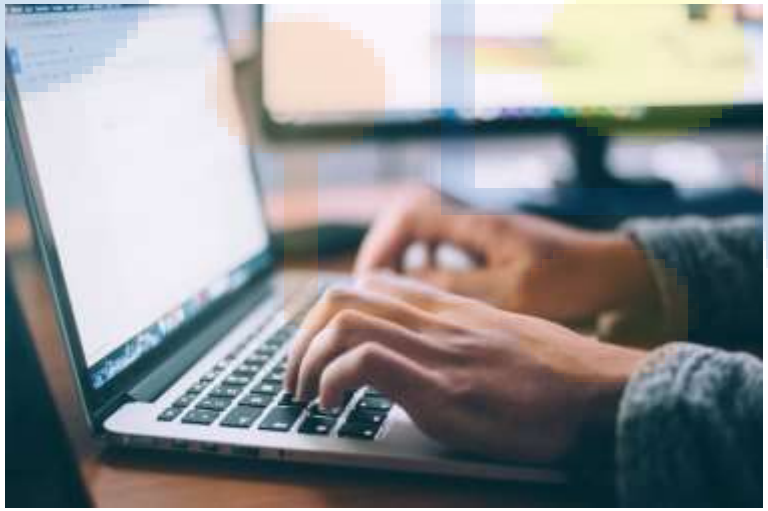
H_A = Alternativa

Teste de Hipótese - Aplicação

Para iniciar um teste de hipótese é importante que as hipóteses nula e alternativa sejam escolhidas corretamente.

Teste de Hipótese - Aplicação

Cabe a você Cientista de Dados, a responsabilidade de escolher o teste mais apropriado.



Teste de Hipótese - Aplicação

Se você deseja testar uma situação pré-estabelecida ou uma determinada afirmação, esta afirmação deverá ser a **Hipótese nula**, ou seja H_0.

Teste de Hipótese - Aplicação

Se você deseja obter uma evidência para suportar uma afirmação feita por você, então, você deve escolher a **Hipótese alternativa**, ou seja, H_A.

Teste de Hipótese - Aplicação

Exercício

Teste de Hipótese - Aplicação

Um provedor quer validar se a média de uso de banda larga é maior, menor ou diferente de 1.8 GB por mês. Quais seriam as hipóteses nula e alternativa?

Teste de Hipótese - Aplicação

Um provedor quer validar se a média de uso de banda larga é maior, menor ou diferente de 1.8 GB por mês. Quais seriam as hipóteses nula e alternativa?

H_0: $\mu = 1.8$

H_A: $\mu < 1.8$

H_A: $\mu > 1.8$

H_A: $\mu \neq 1.8$

Teste de Hipótese - Aplicação

Definir as hipóteses nula e alternativa, depende da natureza do teste e da pessoa que o está conduzindo.

H_0: $\mu = 1.8$ H_A: $\mu > 1.8$	Este teste seria usado por alguém que acredita que o uso de banda larga aumentou e quer suportar que a média de uso de banda larga é agora maior que 1.8 GB por mês.
H_0: $\mu = 1.8$ H_A: $\mu < 1.8$	Este teste seria usado por alguém que acredita que o uso de banda larga diminuiu e quer suportar que a média de uso de banda larga é agora menor que 1.8 GB por mês.
H_0: $\mu = 1.8$ H_A: $\mu \neq 1.8$	Este teste seria usado por alguém que não possui uma expectativa específica, mas quer testar a suposição que a média de uso de banda larga é 1.8 GB por mês.

Análise de Regressão

Seja Bem-Vindo(a)!

Análise de Regressão

Uma imobiliária precisa estabelecer a relação entre o tamanho de uma casa e seu preço de venda.

Análise de Regressão

O gerente de uma loja de eletrônicos gostaria de saber o efeito de **reduzir** o preço de uma impressora em **R$ 10,00** e a demanda pela impressora na semana seguinte.

Análise de Regressão

A Coca-Cola gostaria de prever se o aumento do tempo de seus comerciais em horário nobre, de **30** para **45** segundos, resultaria em aumento de vendas dos seus produtos.

Análise de Regressão

O que temos em comum nesses exemplos de problemas de negócio?

Variável Dependente e Independente

Análise de Regressão

De forma bem objetiva:

Uma **variável independente x** explica a variação em outra variável, que é chamada **variável dependente y**. Este relacionamento existe em apenas uma direção:

variável independente (x) → variável dependente (y)

Análise de Regressão

Por exemplo: A quantidade de quilômetros rodados de um carro, seria uma variável independente e o preço do carro seria uma variável dependente.

Análise de Regressão

= R$ 60.000,00

Variável Dependente

78.239 Km =

Variável Independente

Análise de Regressão

Este relacionamento **não** funciona em **modo reverso**, ou seja, **se alterarmos o preço do carro**, a **quantidade de quilômetros** rodados **não** será alterada.

Análise de Regressão

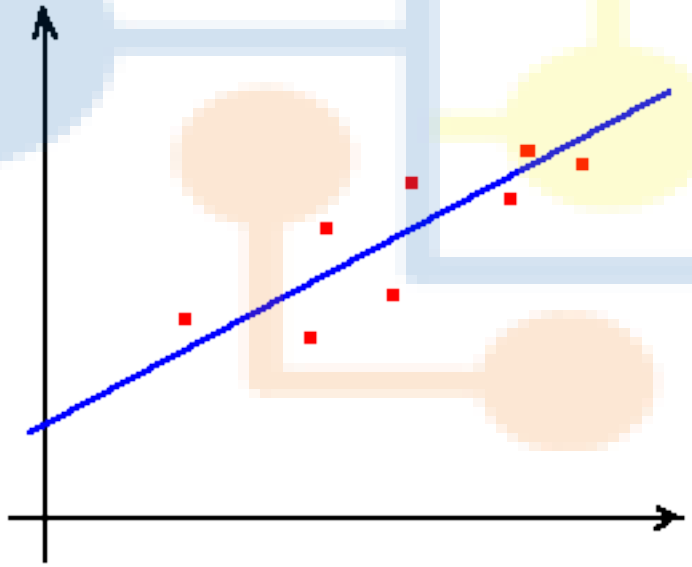
Variável independente (X)	Variável dependente (Y)
O tamanho da tela de um monitor	Preço do monitor
Número de visitantes em um web site	Quantidade de vendas no web site
Tempo de experiência profissional	Salário

Análise de Regressão

Uma das preocupações estatísticas ao analisar dados é a de criar modelos que expliquem estruturas do fenômeno em observação.

Análise de Regressão

E o **modelo de regressão** é um dos métodos estatísticos mais usados para investigar a relação entre variáveis.



Análise de Regressão

Como estudante, você já deve ter se perguntado quantas horas de estudo por semana seriam necessárias para conseguir **9.5** na sua prova final.

Com esta pergunta, você estava buscando o relacionamento entre **horas de estudo** e sua **nota final**.

Análise de Regressão

E temos algumas técnicas de análise estatística principais para estudar este relacionamento:

Teste de hipótese para validar uma suposição no relacionamento entre variáveis.

Análise de correlação, que determina a força e direção do relacionamento entre **duas** variáveis.

Regressão linear simples, que descreve o relacionamento entre duas variáveis usando uma equação.

Análise de Regressão

Seja Bem-Vindo(a)!

Análise de Regressão

Correlação

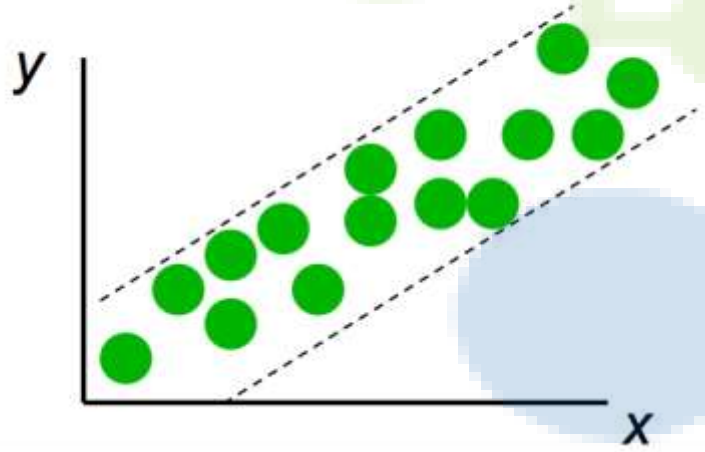
Análise de Regressão

A análise de correlação nos permite medir a **força** e **direção** de um relacionamento linear entre **duas** variáveis.

Análise de Regressão

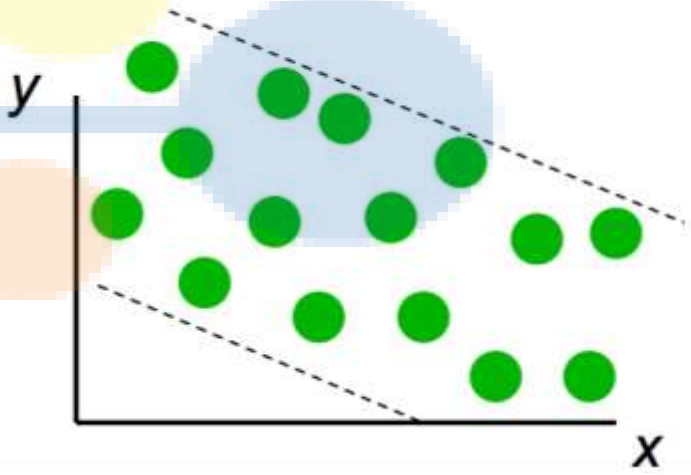
O relacionamento entre duas variáveis **é linear**, se o gráfico de dispersão entre elas tem o **padrão de uma linha reta**. Exemplos de relação linear:

Análise de Regressão



Relacionamento **positivo**, inclinação se move para cima.

Relacionamento **negativo**, inclinação se move para baixo.



Análise de Regressão

Exemplo

Análise de Regressão

Uma revendedora de automóveis gostaria de examinar a relação entre a **quantidade** de **comerciais** de TV por semana e a **venda de carros** por semana.

Análise de Regressão

Espera-se que o número de comerciais de TV por semana (**x**) afete a venda de carros por semana (**y**).

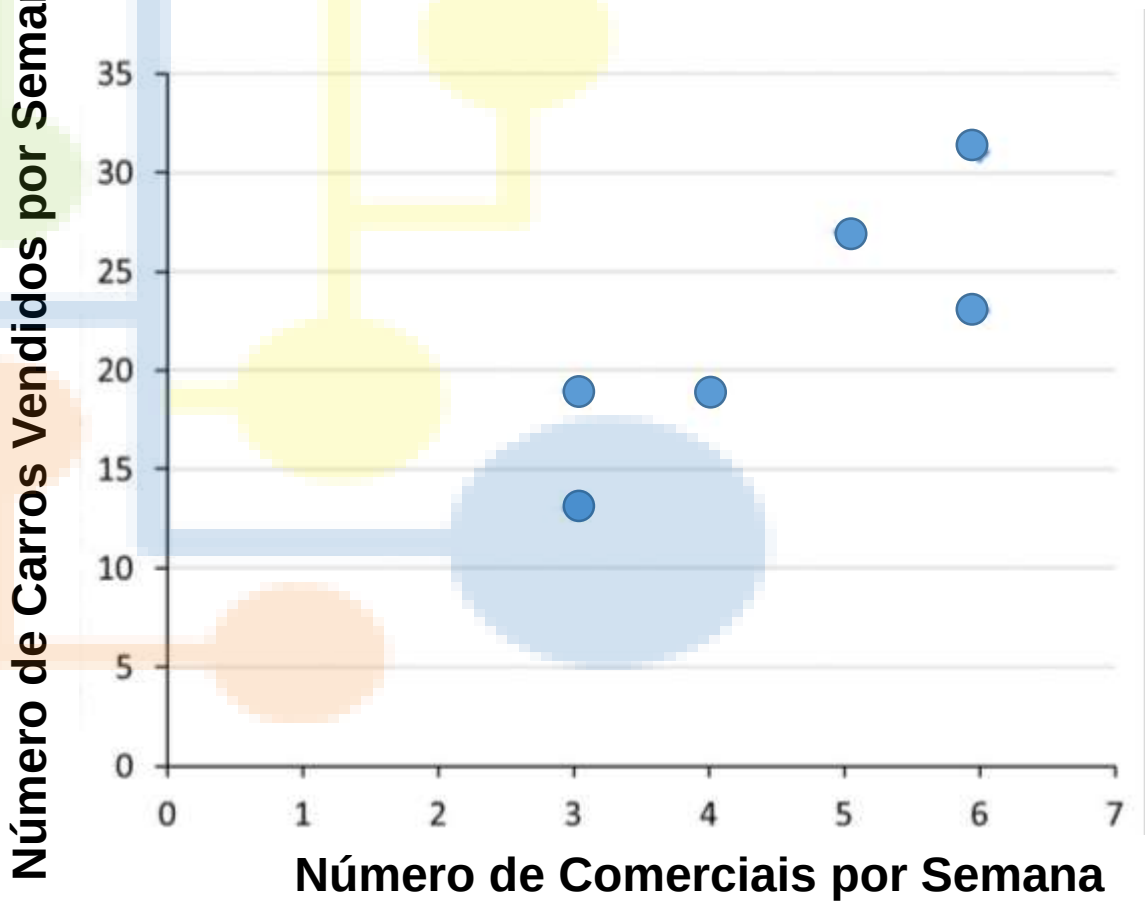
Análise de Regressão

Perceba que esta relação possui uma única direção. Suponha uma amostra de 6 semanas, com os dados coletados na tabela ao lado:

Semana	*Número de comerciais* *x*	*Número de carros vendidos* *y*
1	3	13
2	6	31
3	4	19
4	5	27
5	6	23
6	3	19

Análise de Regressão

Semana	*Número de comerciais* *x*	*Número de carros vendidos* *y*
1	3	13
2	6	31
3	4	19
4	5	27
5	6	23
6	3	19



Análise de Regressão

O coeficiente de correlação **(r)** indica a força e direção de uma relação linear entre a variável independente e dependente.

Análise de Regressão

Exemplos de valores de **r:**

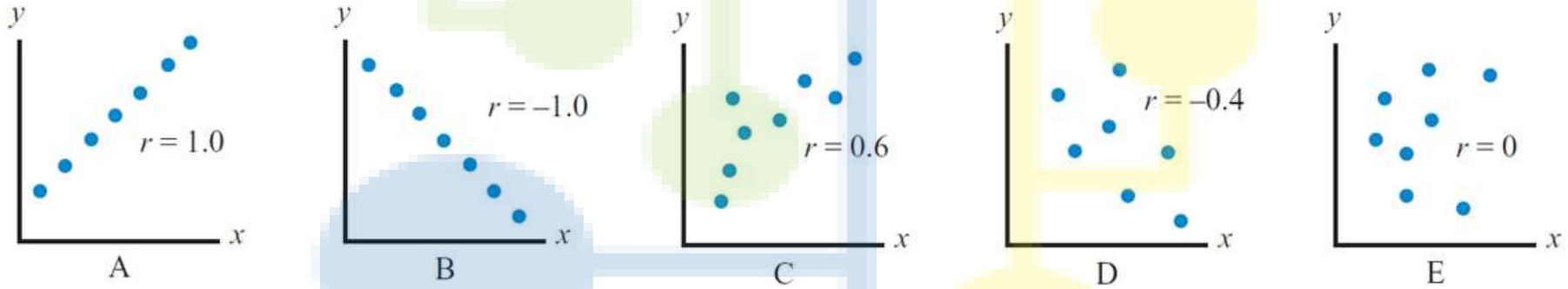


Gráfico A (*r* = 1.0): correlação positiva perfeita entre *x* e *y*

Gráfico B (*r* = -1.0): correlação negativa perfeita entre *x* e *y*

Gráfico C (*r* = 0.6): relação positiva moderada: y tende a aumentar se x aumenta, mas não necessariamente na mesma taxa observada no Gráfico A

Gráfico D (*r* = -0.4): relação negativa fraca: o coeficiente de correlação é próximo de zero ou negativo: y tende a diminuir se x aumenta

Gráfico E (*r* = 0): Sem relação entre x e y

Os valores de **r** variam entre **-1.0** (uma forte relação negativa) até **+1.0**, uma forte relação positiva.

Análise de Regressão

A correlação, isto é, a ligação entre dois eventos, não implica necessariamente uma relação de causalidade, ou seja, que um dos eventos tenha causado a ocorrência do outro.

Análise de Regressão

A correlação pode no entanto indicar possíveis causas ou áreas para um estudo mais aprofundado, ou seja, a correlação pode ser uma pista. A ideia oposta, de que correlação prova automaticamente causalidade, é uma falácia lógica.

Obviamente, dois eventos que possuam de fato uma relação de causalidade deverão apresentar também uma correlação. **O que constitui a falácia é o salto imediato para a conclusão de causalidade**, sem que esta seja devidamente demonstrada.

Análise de Regressão

Só porque (A) acontece juntamente com (B) não significa que (A) causa (B).

Análise de Regressão

É necessário investigação adicional em função de diferentes cenários que podem ocorrer:

1. (A) causa realmente (B);
2. (B) pode ser a causa de (A);
3. Um terceiro fator (C) pode ser causa tanto de (A) como de (B);
4. Pode ser uma combinação das três situações anteriores: (A) causa (B) e ao mesmo tempo (B) causa também (A);
5. A correlação pode ser apenas uma coincidência, ou seja, os dois eventos não têm qualquer relação para além do fato de ocorrerem ao mesmo tempo. (Se estivermos falando de um estudo científico, utilizar uma amostra grande ajuda a reduzir a probabilidade de coincidência).

Análise de Regressão

Então como se determina a causalidade?

Análise de Regressão

Depende sobretudo da complexidade do problema, mas a verdade é que a causalidade dificilmente poderá ser determinada com certeza absoluta.

Análise de Regressão

Seja Bem-Vindo(a)!

Análise de Regressão

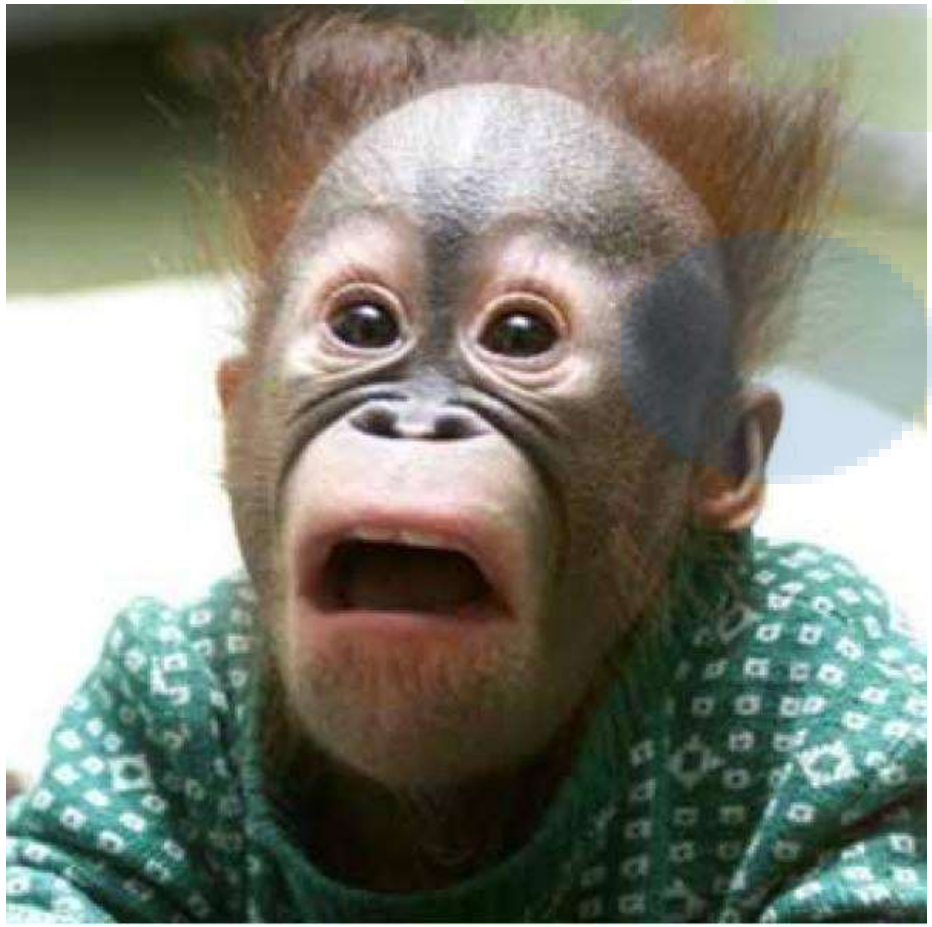
O conjunto de técnicas de regressão é provavelmente um dos mais utilizados em análise de dados.

Análise de Regressão

Existem diversos modelos de regressão:

- Regressão Linear Simples e Múltipla
- Regressão Logística Binária
- Regressão Logística Multinomial
- Regressão Poisson
- Regressão Binomial
- Regressão Ridge
- Regressão Lasso
- Regressão ElasticNet

Análise de Regressão



Compreendeu porque não temos como estudar todo esse conteúdo de forma adequada em apenas um curso e por isso temos a Formação Cientista de Dados?

Análise de Regressão

Por hora, vamos definir o que é a Regressão Linear Simples e mais a frente aqui mesmo no curso, vamos estudar e aplicar com Linguagem R e Azure Machine Learning.

Análise de Regressão

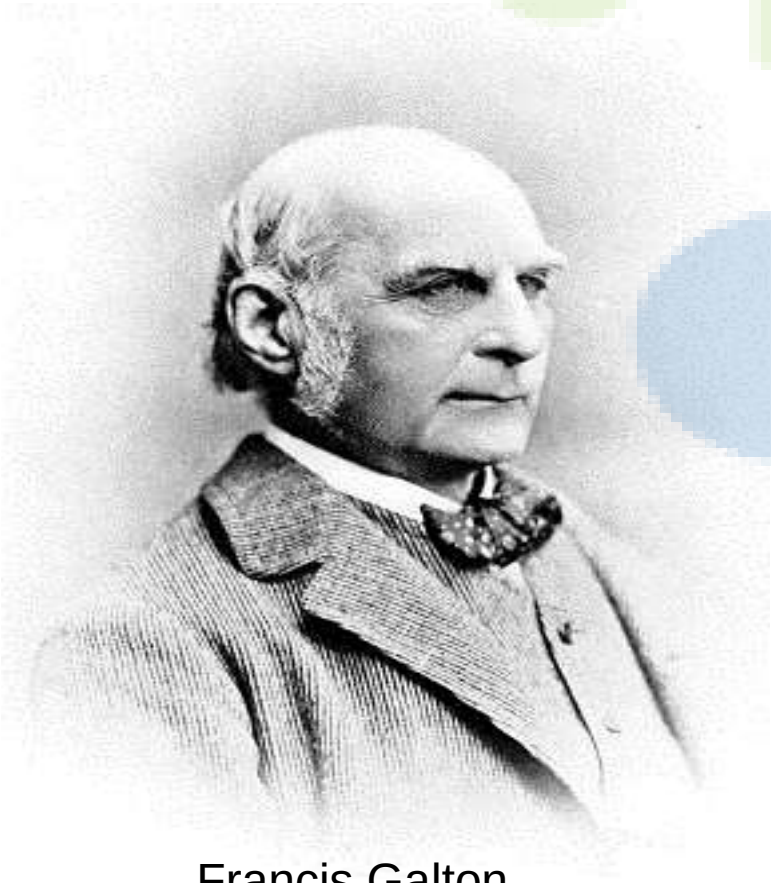
Os modelos de regressão linear simples e múltipla são os mais utilizadas em diversos campos do conhecimento.

Análise de Regressão

Análise de regressão é uma metodologia **estatística** que utiliza a relação entre duas ou mais variáveis quantitativas de tal forma que uma variável possa ser predita a partir de outra.

Análise de Regressão

Análise de regressão é utilizada para se fazer a previsão de resultados. O caso mais simples de regressão é quando temos duas variáveis e a relação entre elas pode ser representada por uma linha reta.



Francis Galton

Origem do Modelo Clássico de Análise de Regressão

Análise de Regressão

Shaquille O'Neal
2,16 metros

Análise de Regressão

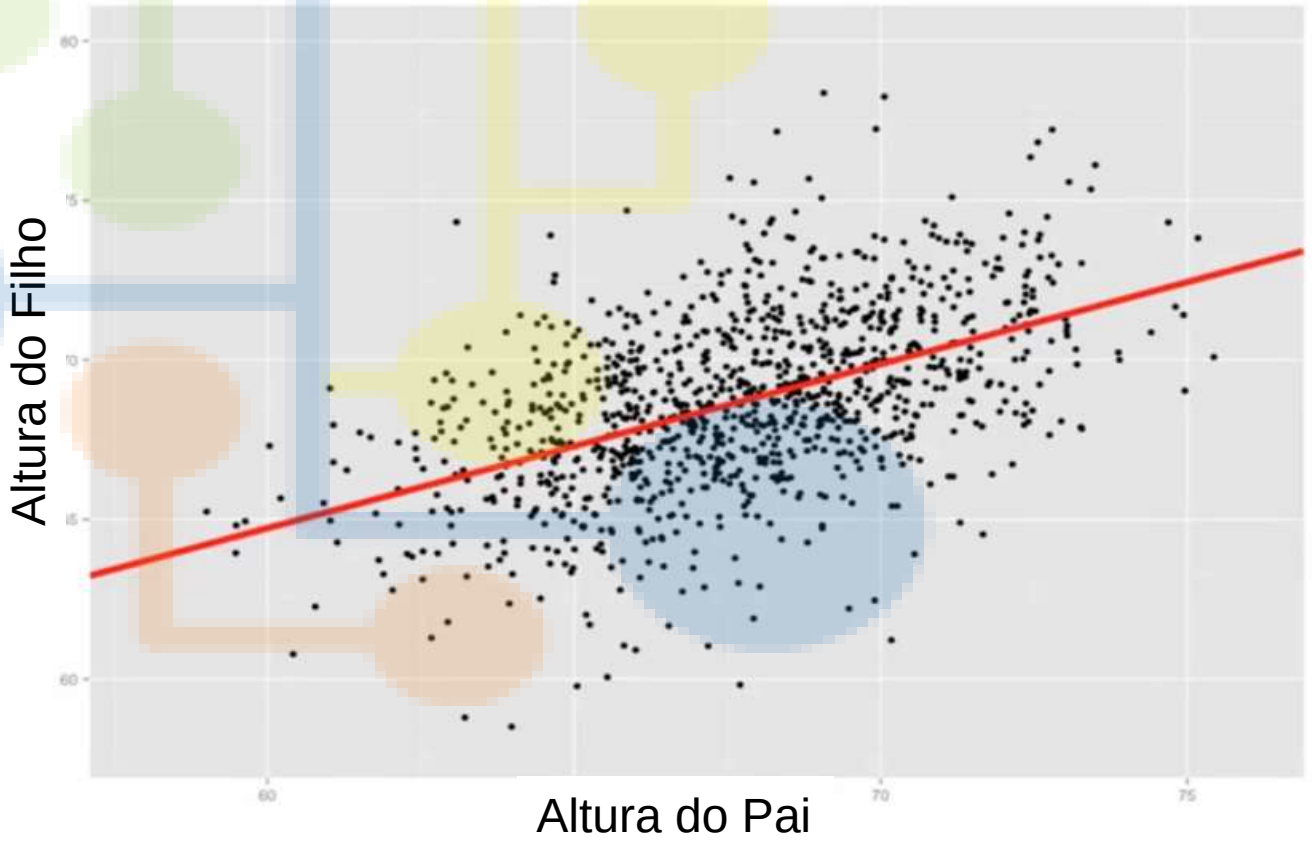
A interpretação moderna da regressão é diferente – ocupa-se do estudo da dependência de uma variável (chamada variável endógena, resposta ou dependente), em relação a uma ou mais variáveis, (chamadas variáveis explicativas ou exógenas), com o objetivo de estimar e/ou prever a média (da população) ou valor médio de uma variável dependente em termos dos valores conhecidos ou fixos (em amostragem repetida) das variáveis independentes.

Análise de Regressão

Fenômeno de Regressão

Análise de Regressão

Regressão Linear



REGRESSÃO e CORRELAÇÃO são a mesma coisa?

Análise de Regressão

Não

REGRESSÃO

Análise de Regressão – prevê o valor médio de uma variável com base nos valores estabelecidos de uma ou mais variáveis.

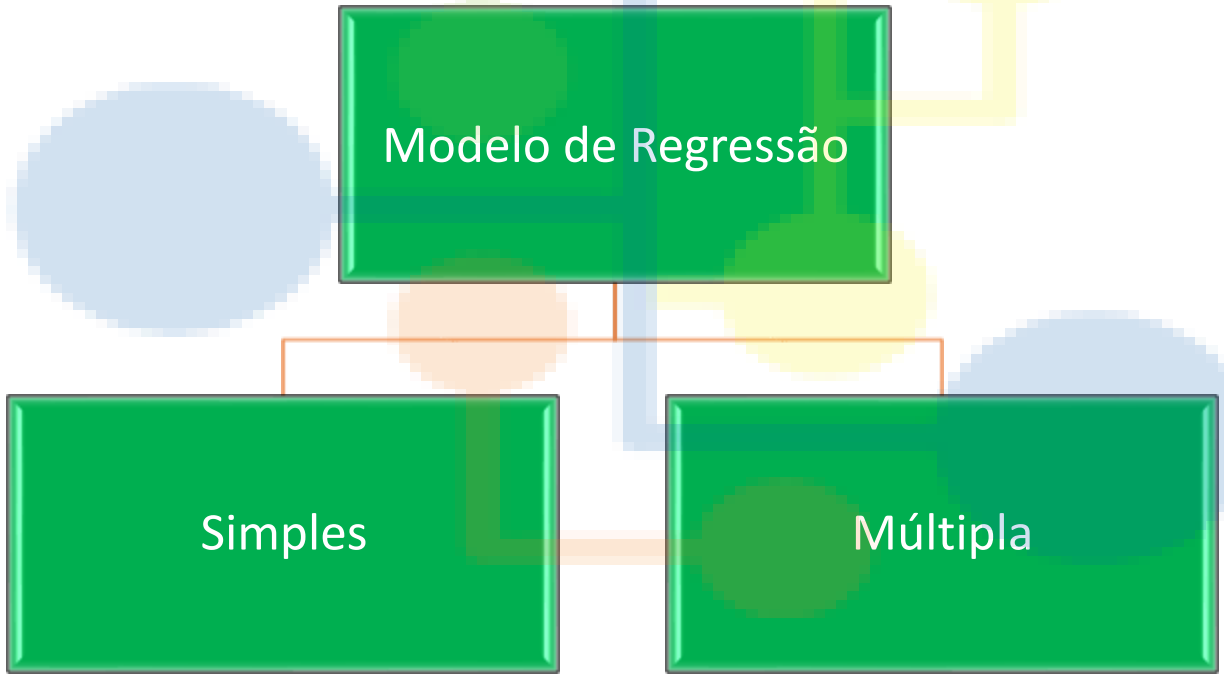
CORRELAÇÃO

Análise de Correlação – tem como objetivo medir o grau de associação linear entre duas variáveis.

Análise de Regressão

Ou seja, usamos a correlação para medir o grau de relação entre duas variáveis e depois usamos regressão para estudar o relacionamento entre elas.

Tipos de Modelos de Regressão Linear



1 Variável Dependente Y
1 Variável Independente X

1 Variável Dependente Y
2 ou + Variáveis Independentes X, X_i

Análise de Regressão

Regressão Linear Simples

Análise de Regressão

Nós já sabemos que o coeficiente de correlação **r** nos provê uma medida que descreve a força e direção do relacionamento entre duas variáveis.

Análise de Regressão

Nosso próximo passo é realizar uma **análise de regressão linear simples**, que nos habilite descrever uma linha reta que melhor representa uma série de pares ordenados (x, y).

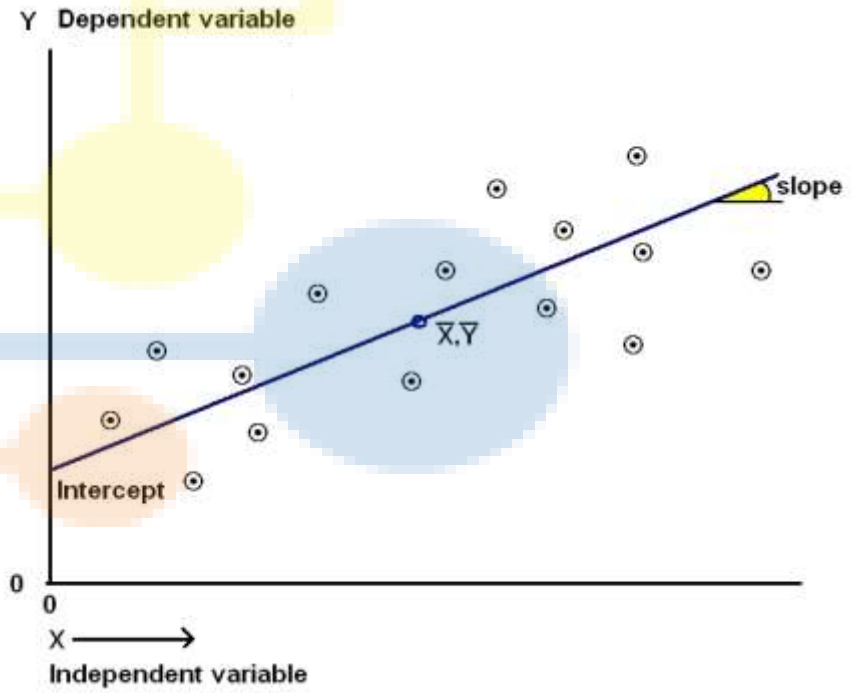
Análise de Regressão

Como veremos mais adiante, ter uma linha reta que descreve o relacionamento entre a variável independente (x) e a variável dependente (y) nos oferece uma série de vantagens sobre o coeficiente de correlação.

Análise de Regressão

Fórmula para a equação que descreve uma linha reta através de um par ordenado:

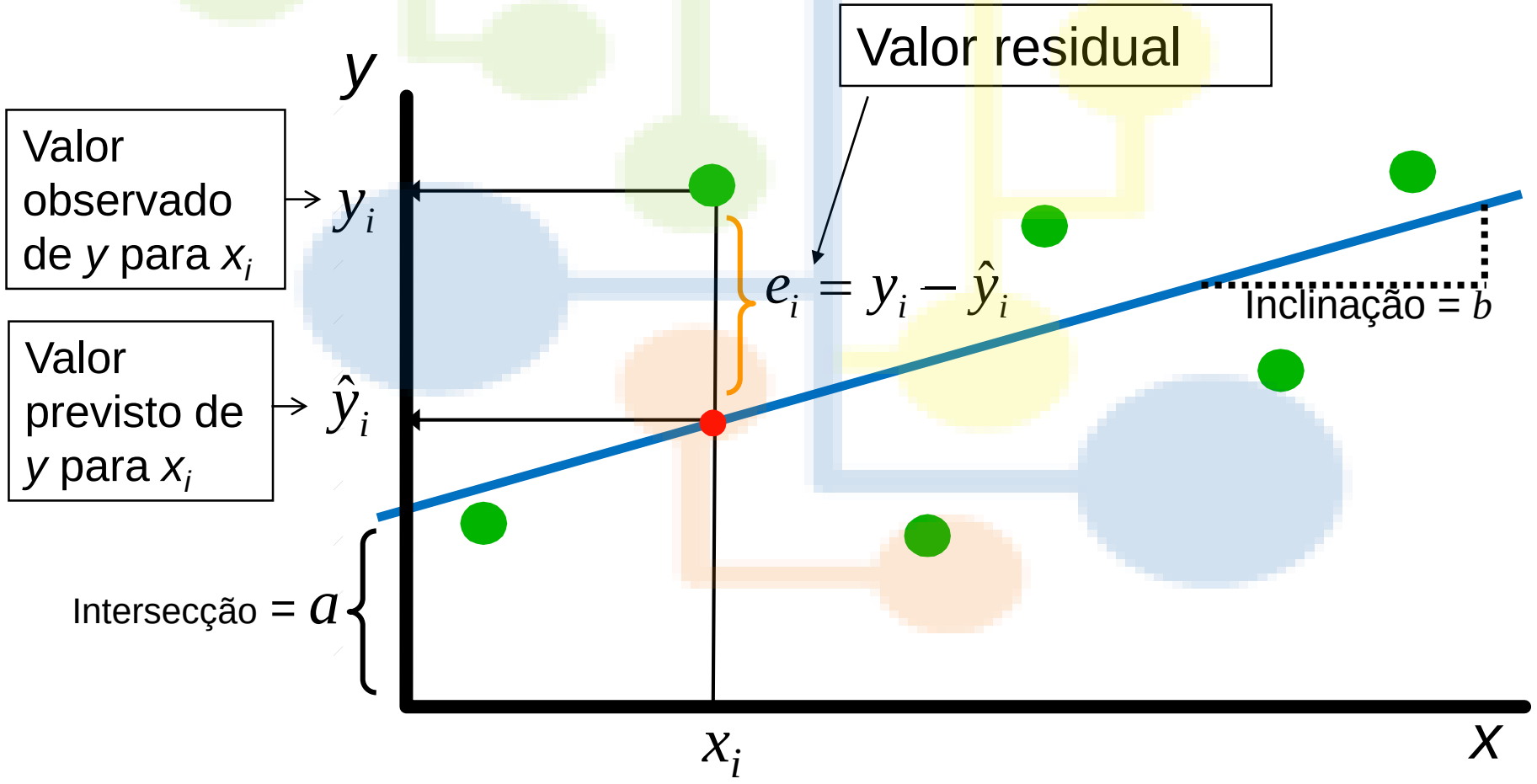
$$\hat{y} = a + bx$$



Onde:

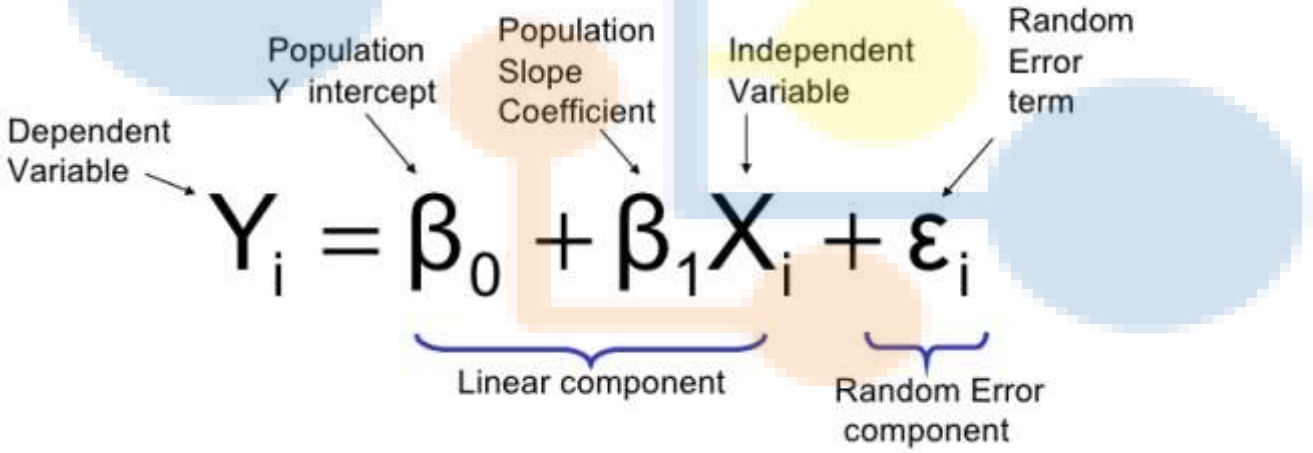
- $\hat{y}$ = valor previsto de *y* dado um valor para *x*
- ***x*** = variável independente
- ***a*** = ponto onde a linha intercepta o eixo y
- ***b*** = inclinação da linha reta

Análise de Regressão



Análise de Regressão

Regressão Linear Simples (2 variáveis)



Análise de Regressão

Regressão Linear Múltipla (Mais de 2 variáveis)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

Análise de Regressão

Qual o objetivo da Análise de Regressão?

Análise de Regressão

Vamos então estudar um exemplo completo de regressão linear simples

Análise de Regressão

Exemplo

Análise de Regressão

Vamos imaginar que um professor tenha interesse em saber a relação entre as horas de estudo fora da sala de aula e nota final dos alunos em um exame.

Análise de Regressão

Para isso, o professor conduziu uma pesquisa em uma turma com 10 estudantes de uma mesma classe. Esta tabela mostra o resultado desta pesquisa.

Análise de Regressão

Estudante	Tempo gasto em estudo fora da sala de aula (minutos)	Nota no exame final (0 a 100)
Marcio	15	24
Tiago	20	18
David	20	45
Nadir	40	60
Leonardo	50	75
Jaime	25	33
Aline	10	15
Dalton	55	96
Flavio	35	84
Henrique	30	60

Análise de Regressão

Este problema pode ser representado por esta equação (ou modelo) de regressão simples:

$$\hat{y} = a + bx$$

Análise de Regressão

Este problema pode ser representado por esta equação (ou modelo) de regressão simples:

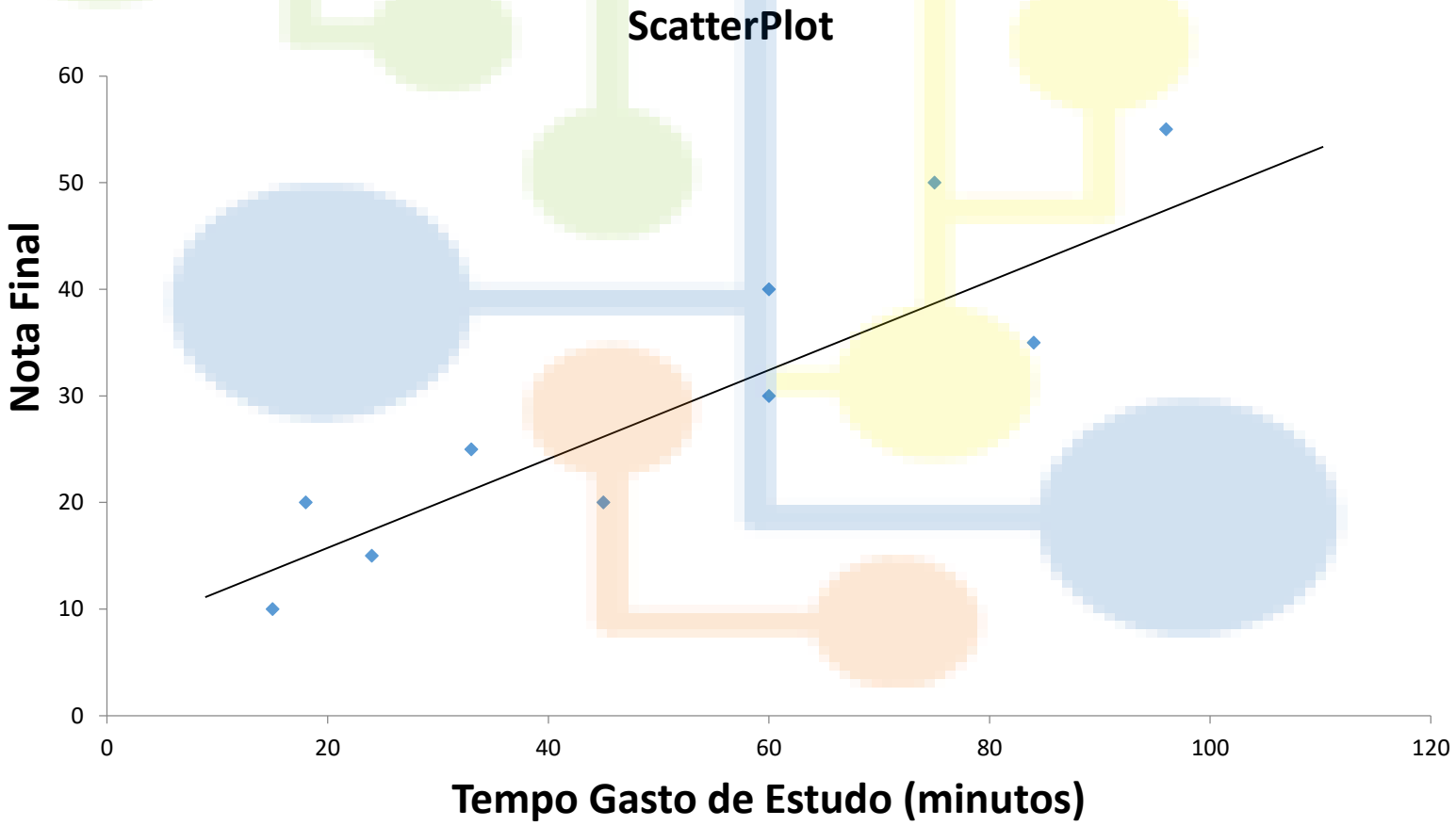
$$\hat{y} = a + bx$$

Nota Final = A + B x Tempo de Estudo

Análise de Regressão

Nosso objetivo aqui é estudar se o tempo de estudo (variável independente) apresenta relação com a variação da nota final (variável dependente).

Análise de Regressão



Análise de Regressão

Para determinarmos a equação que melhor se ajusta a nuvem de pontos, devemos estabelecer duas condições fundamentais aos resíduos:

1- A somatória dos resíduos deve ser 0.
2- A somatória dos resíduos ao quadrado é a mínima possível.

Análise de Regressão

Deve-se determinar α e β de modo que a somatória dos quadrados dos resíduos seja a menor possível (método de Mínimos Quadrados Ordinários - MQO, ou, em inglês, Ordinary Least Squares - OLS)

Análise de Regressão

No próximo capítulo, você vai trabalhar em um projeto de regressão do início ao fim em Linguagem R.

Premissas da Análise de Regressão

Seja Bem-Vindo(a)!

Premissas da Análise de Regressão

Como estamos trabalhando com dados da amostra para fazer previsões sobre a população, como podemos saber qual é o nível de precisão nas predições que fazemos usando regressão linear?

Premissas da Análise de Regressão

Para responder a esta pergunta, precisamos construir um intervalo de confiança.

Os intervalos de confiança oferecem uma estimativa do parâmetro da população, baseado na estatística da amostra.

Premissas da Análise de Regressão

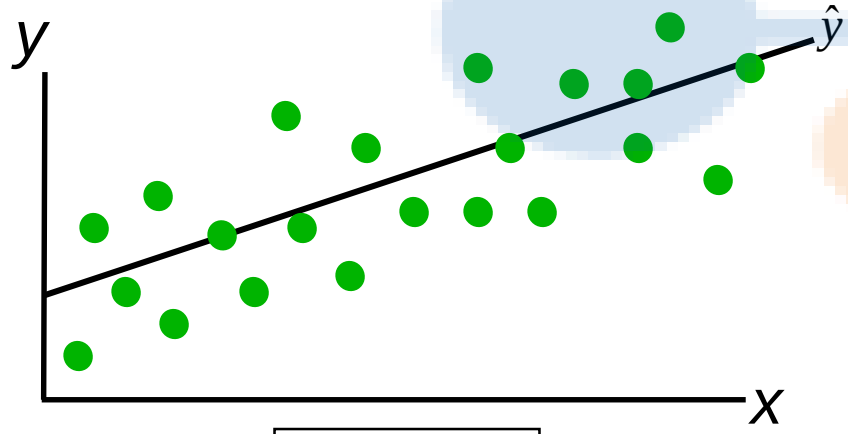
Premissas para Análise de Regressão

Premissas da Análise de Regressão

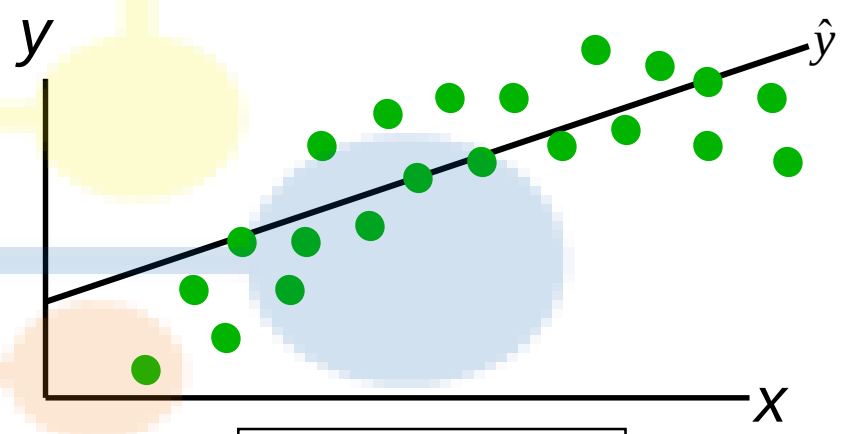
Para que os resultados de uma **análise de regressão** sejam **confiáveis**, algumas **premissas** precisam ser **satisfeitas.**

Premissas da Análise de Regressão

Premissa 1 O relacionamento entre a variável independente e a variável dependente deve ser linear.



Linear



Não Linear

Premissas da Análise de Regressão

Premissa 2 O valor residual não deve exibir um padrão através da variável independente.

$$e_i = y_i - \hat{y}_i$$

Premissas da Análise de Regressão

Armadilhas da Análise de Regressão

Premissas da Análise de Regressão

Não faça previsões para a variável dependente (y), além do range de valores da variável independente (x).

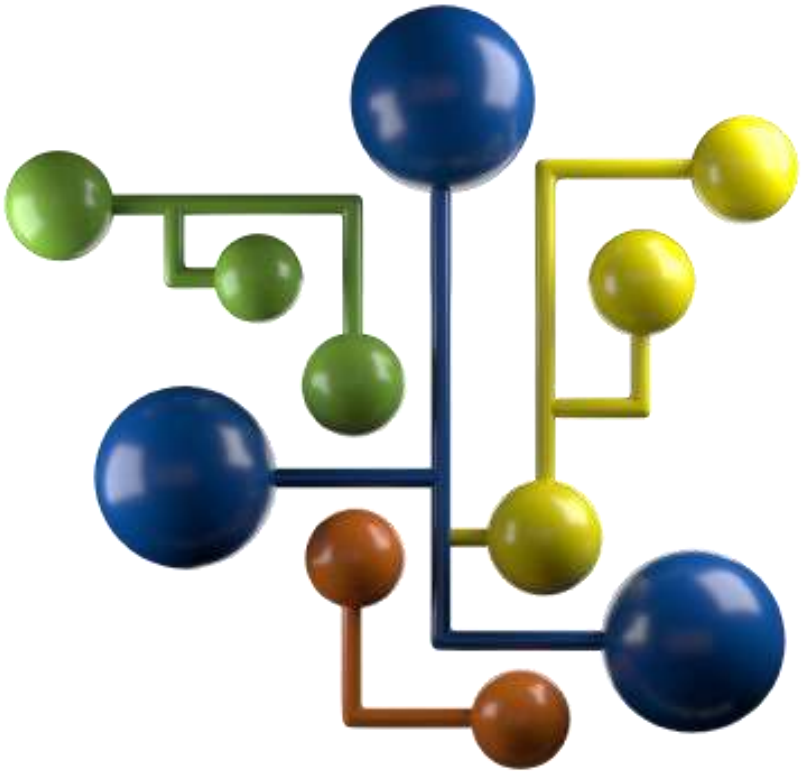
Não há garantia que o relacionamento estimado é apropriado além do range que foi observado.

Não confunda correlação com causalidade.

Premissas da Análise de Regressão

Apesar do relacionamento entre as variáveis ser estatisticamente significante, não prova que a variável independente realmente causou a mudança na variável dependente.

Muito Obrigado por Participar!



Tenha uma Excelente Jornada de Aprendizagem.

Equipe Data Science Academy