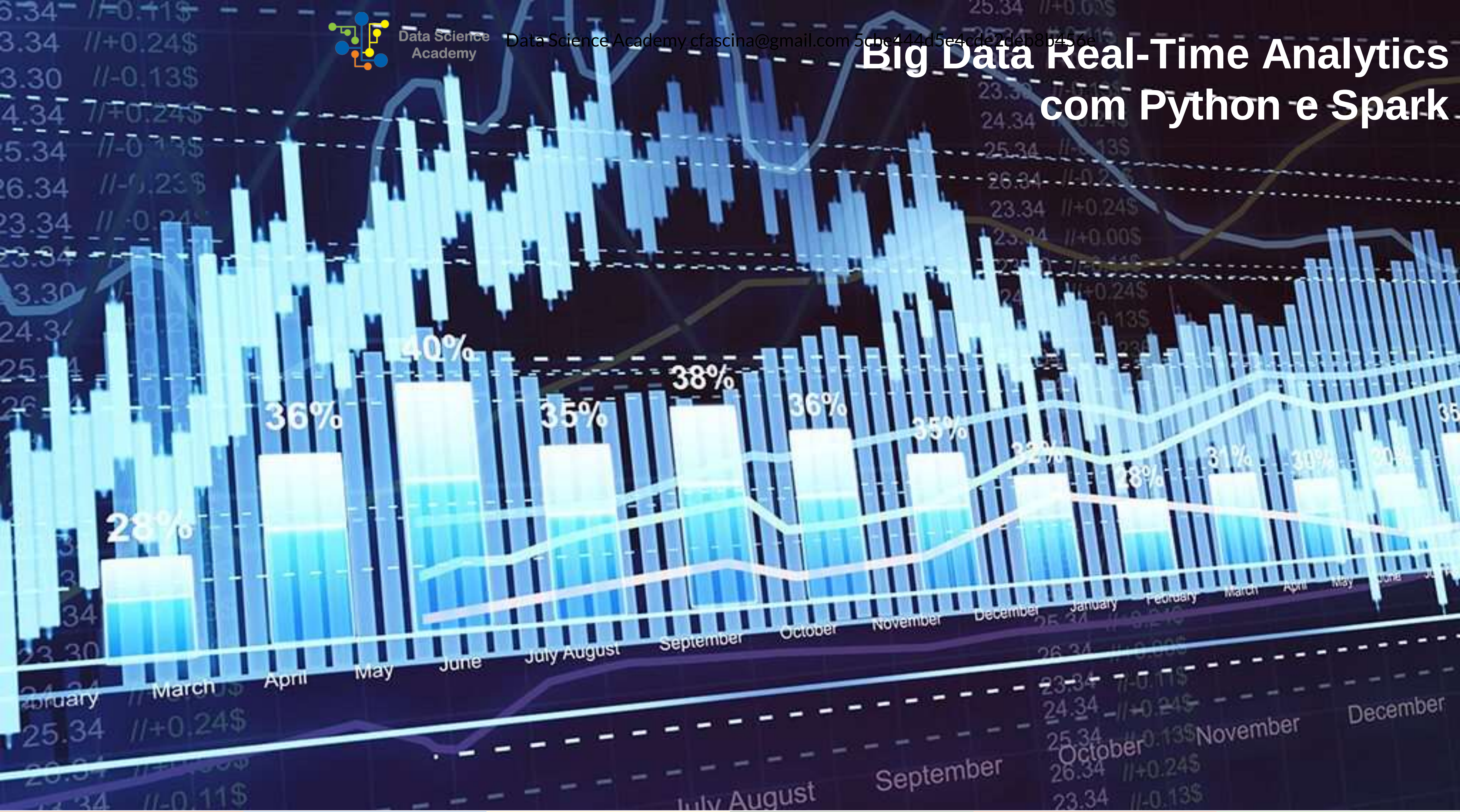
Big Data Real-Time Analytics com Python e Spark



Seja muito bem-vindo(a)!

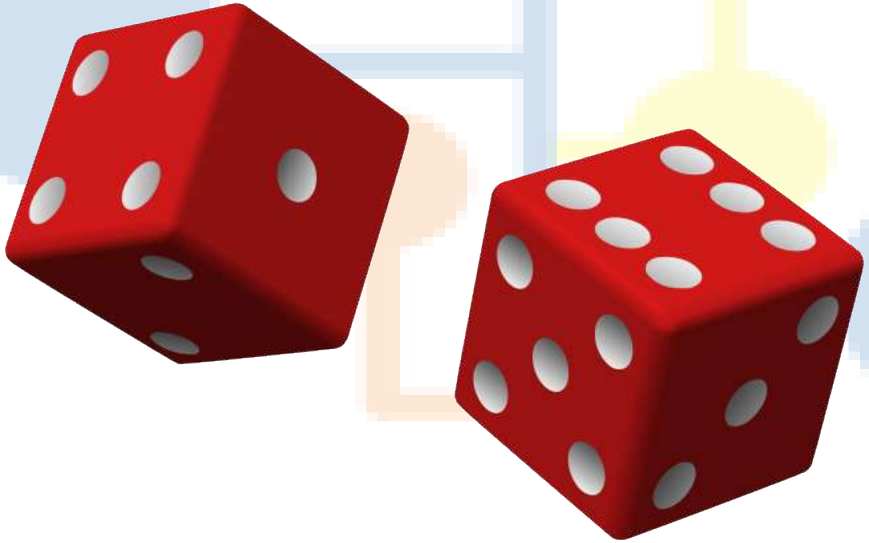
Análise Estatística de Dados

Análise Estatística de Dados

Parte 1

Parte 2

Teoria da Probabilidade



Introdução à Probabilidade

A probabilidade é um número que varia de 0 (zero) a 1 (um) e que mede a chance de ocorrência de um determinado resultado.

Quanto mais próxima de zero for a probabilidade, menores são as chances de ocorrer o resultado e quanto mais próxima de um for a probabilidade, maiores são as chances.

Introdução à Probabilidade

As probabilidades podem ser expressas de diversas maneiras, inclusive decimais, frações e percentagens.

Por exemplo, a chance de ocorrência de um determinado evento pode ser expressa como 10%; 5 em 10; 0,20 ou 1/7.

Introdução à Probabilidade

A teoria da probabilidade consiste em utilizar a intuição humana para estudar os fenômenos do nosso cotidiano. Para isso, vamos utilizar o princípio básico do aprendizado humano que é a ideia de experimento.

Introdução à Probabilidade

Podemos classificar os experimentos em dois tipos:

- Aleatórios (casuais)
- Não aleatórios (determinísticos)

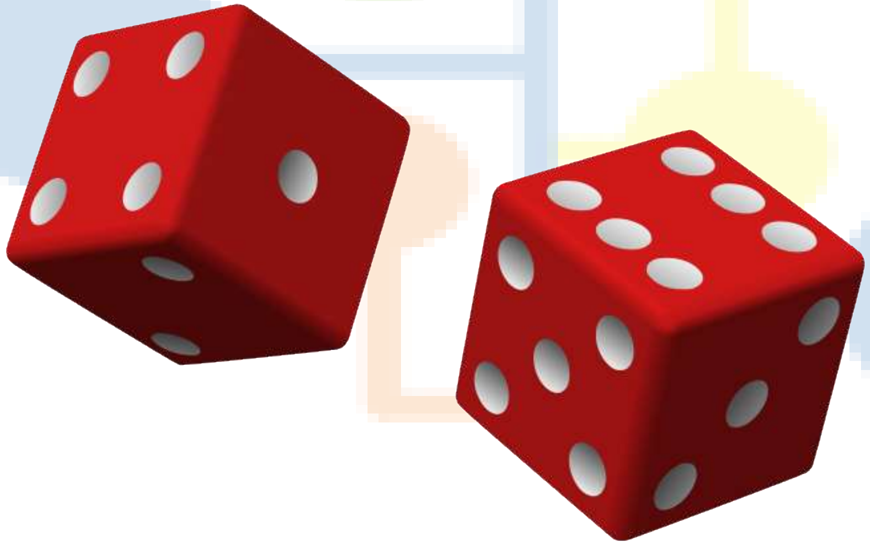
Os experimentos determinísticos são totalmente caracterizados a priori, ou seja, são fenômenos em que o resultado é sabido antes mesmo em que ele ocorra e desta forma, nada temos a fazer.

Introdução à Probabilidade

Os experimentos que iremos estudar são os aleatórios, dos quais não sabemos o resultado a priori.

Um experimento é dito aleatório quando não conseguimos afirmar o resultado que será obtido antes de realizar o experimento.

Experimento Aleatório



Experimento Aleatório

Experimento Aleatório

Experimento aleatório é o fenômeno que, quando repetido inúmeras vezes em processos semelhantes, possui resultados imprevisíveis.

Experimento Aleatório

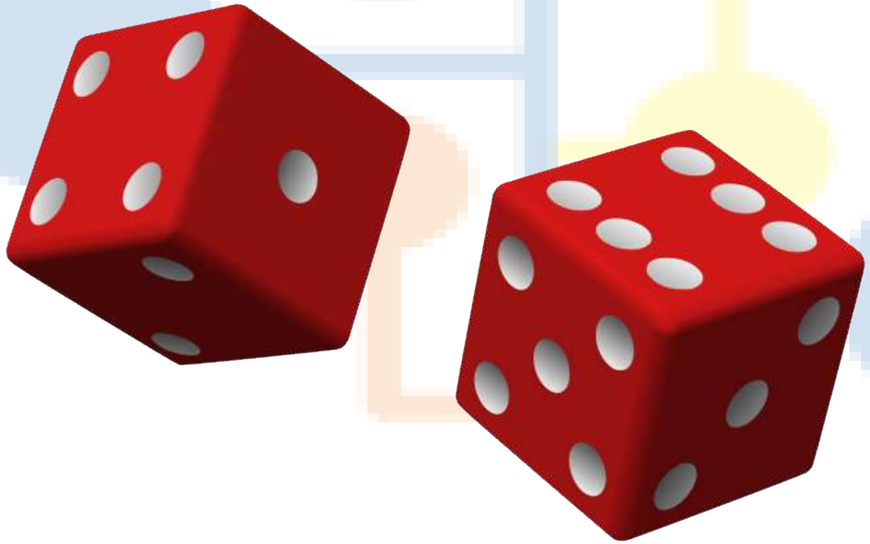
O lançamento de um dado e de uma moeda são considerados exemplos de experimentos aleatórios.

No caso dos dados podemos ter seis resultados diferentes {1, 2, 3, 4, 5, 6} e no lançamento da moeda, dois {cara, coroa}.

Experimento Aleatório

Experimento é qualquer atividade realizada que pode apresentar diferentes resultados. Um experimento é dito aleatório quando não conseguimos afirmar o resultado que será obtido antes de realizar o experimento. Um experimento é dito equiprovável se todos os possíveis resultados possuem a mesma chance de ocorrer.

Tipos de Probabilidade



Tipos de Probabilidade

Evento – um ou mais resultados de um experimento.

O resultado e/ou resultados são um subconjunto do espaço da amostra.

Tipos de Probabilidade

Probabilidade Clásssica

Probabilidade Empírica

Probabilidade Subjetiva

Tipos de Probabilidade

Probabilidade Clássica : é usada quando nós sabemos o número de possíveis resultados do evento de interesse e podemos calcular a probabilidade do evento com a seguinte fórmula:

$$P(A) = \frac{\text{Número de possíveis resultados do evento A}}{\text{Número total de possíveis resultados dentro do espaço da amostra}}$$

Onde: **P(A)** é a probabilidade de um evento ocorrer.

Tipos de Probabilidade

A **Probabilidade Empírica**, envolve conduzirmos um experimento, para observarmos a frequência com que um evento ocorre.

Para calcularmos a probabilidade empírica, usamos a fórmula:

$$P(A) = \frac{\text{Frequência em que o evento A ocorre}}{\text{Número total de observações}}$$

Tipos de Probabilidade

Usamos **Probabilidade Subjetiva**, quando:

Dados ou experimentos não estão disponíveis para calcular a probabilidade.

Regras da Probabilidade

Regras da Probabilidade

1ª Regra

Se P(A) = 1, então podemos garantir que o evento A **ocorrerá.**

Regras da Probabilidade

2ª Regra

Se $P(A) = 0$, então podemos garantir que o evento A **NÃO** ocorrerá.

Regras da Probabilidade

3ª Regra

A probabilidade de qualquer evento sempre será entre **0** e **1**. Probabilidades nunca podem ser **negativas** ou **maior que 1**.

Regras da Probabilidade

4ª Regra

A soma de todas as probabilidades para um evento simples, em um espaço de amostra, será **igual a 1.**

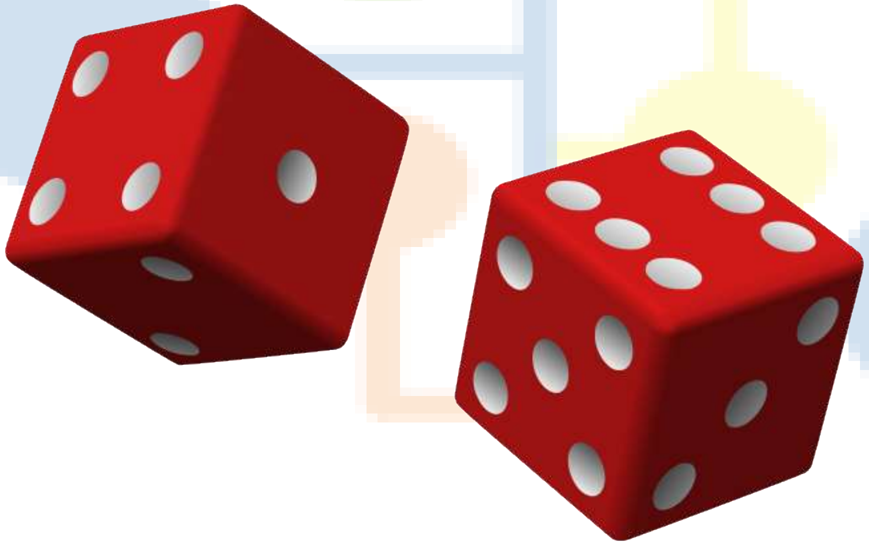
Regras da Probabilidade

5ª Regra

O complemento do evento A é definido como todos os resultados em um espaço de amostra, que **não** fazem parte do evento A. Ou seja:

P(A) = 1 – P(A’), onde P(A’) é o complemento do evento A.

Eventos e Espaço Amostral



Eventos e Espaço Amostral

O primeiro elemento na modelagem de um experimento é o espaço amostral, que consiste no conjunto de todos os possíveis resultados do experimento.

Eventos e Espaço Amostral

Não é possível exibir esta imagem.

Eventos e Espaço Amostral

Não é possível exibir esta imagem.

Eventos e Espaço Amostral

S = {defeituoso, não defeituoso}

S = {Sim, Não}

S = {1, 2, 3, 4, 5, 6}

S = {cara, coroa}

Eventos e Espaço Amostral

Eventos Complementares

Sabemos que um evento pode ocorrer ou não. Sendo **p** a probabilidade de que ele ocorra (sucesso) e **q** a probabilidade de que ele não ocorra (insucesso), para um mesmo evento existe sempre a relação:

$$p + q = 1 \Rightarrow q = 1 - p$$

Eventos e Espaço Amostral

Eventos Independentes (Regra do “e”)

Dizemos que dois eventos são independentes quando a realização ou não realização de um dos eventos não afeta a probabilidade da realização do outro e vice-versa.

Assim, sendo p1 a probabilidade de realização do primeiro evento e p2 a probabilidade do segundo evento, a probabilidade de que tais eventos se realizem simultaneamente é dada por:

$$p = p1 \times p2$$

Probabilidade Conjunta

Escolha de Prêmios				
Sexo	Taco de Golfe	Câmera	Bicicleta	**Total**
Masculino	117	50	60	**227**
Feminino	130	91	30	**251**
Total	**247**	**141**	**90**	**478**

Se o vencedor for escolhido aleatoriamente por esses clientes, a probabilidade de selecionarmos uma mulher é apenas a frequência relativa correspondente (já que temos a mesma probabilidade de selecionar qualquer um dos 478 clientes). Há 251 mulheres nos dados de um total de 478, dando uma probabilidade de:

P(mulher) = 251/478 = 0,525.

Probabilidade Conjunta

Escolha de Prêmios				
Sexo	Taco de Golfe	Câmera	Bicicleta	**Total**
Masculino	117	50	60	**227**
Feminino	130	91	30	**251**
Total	**247**	**141**	**90**	**478**

Isso é chamado de **probabilidade marginal**, porque depende apenas dos totais encontrados nas margens da tabela. O mesmo método funciona para eventos mais complicados.

Por exemplo, qual é a probabilidade de escolher uma mulher cujo prêmio preferido é a câmera? Como 91 mulheres nomearam a câmera como sua preferência, então a probabilidade é:

P(mulher e câmera) = 91/478 = 0,190.

Probabilidade Conjunta

Escolha de Prêmios				
Sexo	Taco de Golfe	Câmera	Bicicleta	**Total**
Masculino	117	50	60	**227**
Feminino	130	91	30	**251**
Total	**247**	**141**	**90**	**478**

Probabilidades como essas são chamadas de **probabilidades conjuntas** porque elas dão a probabilidade de dois eventos ocorrerem juntos.

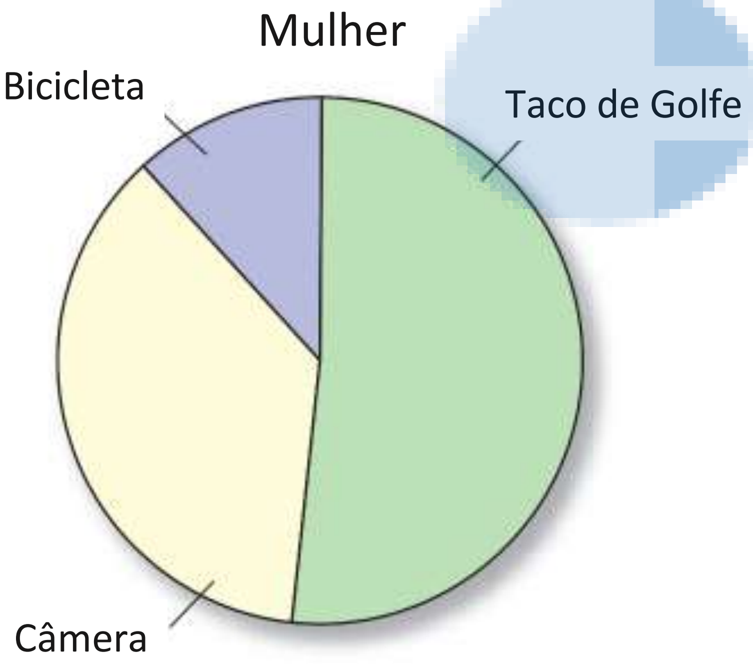
Probabilidade Condicional e Independência

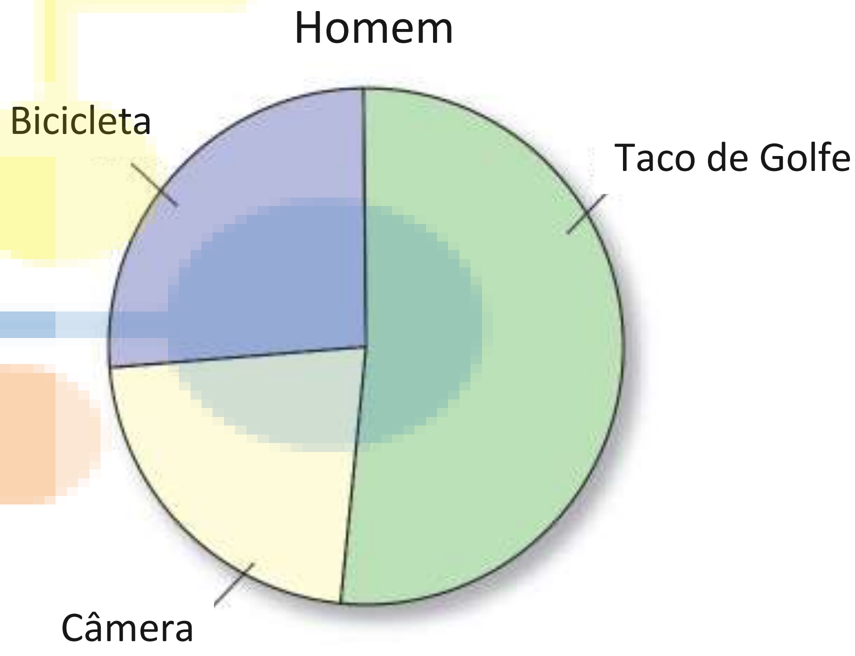
Probabilidade Condicional e Independência

Escolha de Prêmios				
Sexo	Taco de Golfe	Câmera	Bicicleta	**Total**
Masculino	117	50	60	**227**
Feminino	130	91	30	**251**
Total	**247**	**141**	**90**	**478**

Probabilidade Condicional e Independência

Escolha de Prêmios				
Sexo	Taco de Golfe	Câmera	Bicicleta	**Total**
Masculino	117	50	60	**227**
Feminino	130	91	30	**251**
Total	**247**	**141**	**90**	**478**





Probabilidade Condicional e Independência

Escolha de Prêmios				
Sexo	Taco de Golfe	Câmera	Bicicleta	**Total**
Masculino	117	50	60	**227**
Feminino	130	91	30	**251**
Total	**247**	**141**	**90**	**478**

Escrevemos a probabilidade de um cliente selecionado querer uma bicicleta, uma vez que selecionamos uma mulher como:

P(bicicleta | mulher) = 30/251 = 0,120.

Probabilidade Condicional e Independência

Escolha de Prêmios				
Sexo	Taco de Golfe	Câmera	Bicicleta	**Total**
Masculino	117	50	60	**227**
Feminino	130	91	30	**251**
Total	**247**	**141**	**90**	**478**

Para os homens, olhamos para a distribuição condicional de prêmios preferidos dado "Homem" mostrado na linha superior da tabela.

Lá, dos 227 homens, 60 disseram que seu prêmio preferido era uma bicicleta. Então,

P(bicicleta | homem) = 60/227 = 0,264

mais que o dobro da probabilidade das mulheres.

Probabilidade Condicional e Independência

Escolha de Prêmios				
Sexo	Taco de Golfe	Câmera	Bicicleta	**Total**
Masculino	117	50	60	**227**
Feminino	130	91	30	**251**
Total	**247**	**141**	**90**	**478**

Em geral, quando queremos a probabilidade de um evento de uma distribuição condicional, escrevemos P(B | A) e o pronunciamos "a probabilidade de B dado A."

Uma probabilidade que leva em conta uma dada condição como essa é chamada de uma **probabilidade condicional**.

Probabilidade Condicional e Independência

Escolha de Prêmios				
Sexo	Taco de Golfe	Câmera	Bicicleta	**Total**
Masculino	117	50	60	**227**
Feminino	130	91	30	**251**
Total	**247**	**141**	**90**	**478**

Vamos ver o que fizemos. Nós trabalhamos com as contagens, mas também poderíamos trabalhar com as probabilidades. Havia 30 mulheres que selecionaram uma bicicleta como prêmio, e havia 251 mulheres clientes. Então nós achamos a probabilidade de ser 30/251.

Probabilidade Condicional e Independência

Escolha de Prêmios				
Sexo	Taco de Golfe	Câmera	Bicicleta	**Total**
Masculino	117	50	60	**227**
Feminino	130	91	30	**251**
Total	**247**	**141**	**90**	**478**

Para encontrar a probabilidade do evento B dado o evento A, restringimos nossa atenção aos resultados em A. Então, encontramos em que fração desses resultados B também ocorreu.

Formalmente, escrevemos:

$$P(\mathbf{B} \mid \mathbf{A}) = \frac{P(\mathbf{A} \text{ and } \mathbf{B})}{P(\mathbf{A})}$$

Probabilidade Condicional e Independência

Escolha de Prêmios				
Sexo	Taco de Golfe	Câmera	Bicicleta	**Total**
Masculino	117	50	60	**227**
Feminino	130	91	30	**251**
Total	**247**	**141**	**90**	**478**

A fórmula para probabilidade condicional requer uma restrição. A fórmula funciona somente quando o evento que é dado tem uma probabilidade maior que 0.

A fórmula não funciona se P(A) for 0 porque isso significaria que nós tínhamos "dado" o fato de que A é verdade mesmo que a probabilidade de A fosse 0, o que seria uma contradição. Lembra-se da Regra de Multiplicação para a probabilidade de A e B?

P (A e B) = P (A) * P (B) quando A e B são independentes.

Probabilidade Condicional e Independência

Agora podemos escrever uma regra mais geral que não requer independência. Na verdade, já escrevemos. Nós só precisamos reorganizar a equação um pouco.

Probabilidade Condicional e Independência

Não é possível exibir esta imagem.

A equação na definição da probabilidade condicional contém a probabilidade de A e B.

A reorganização da equação fornece a Regra Geral de Multiplicação para eventos compostos que não exigem que os eventos sejam independentes:

P(A e B) = P(A) * P(B | A)

A probabilidade de que dois eventos, A e B, ocorram é a probabilidade de que o evento A ocorra multiplicado pela probabilidade de que o evento B também ocorra - isto é, pela probabilidade de que o evento B ocorra dado que o evento A ocorra.

Probabilidade Condicional e Independência

Resumindo

Ou (Or)	Em Geral	P(A or B) = P(A) + P(B) - P(A and B)
	Se eventos são mutuamente exclusivos	P(A or B) = P(A) + P(B)
E (And)	Em Geral	P(A and B) = P(A) * P(B | A) = P(A | B) * P(B)
	Se eventos são independentes	P(A and B) = P(A) * P(B)

Tabela de Contingência

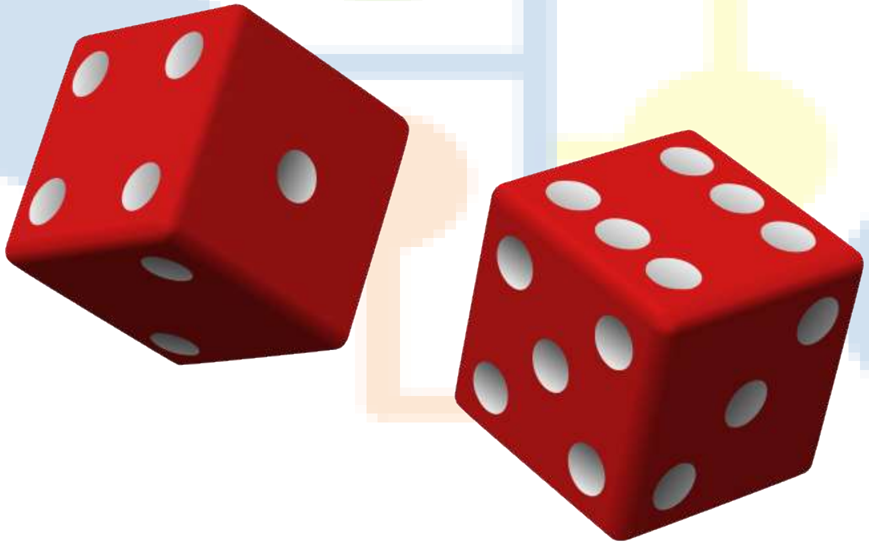


Tabela de Contingência

As **Tabelas de Contingência** são os meios de organizar as informações correspondentes aos dados classificados segundo dois critérios.

	Variável 2				
Variável 1	A	B	...	r	total
A	$n_{1\,1}$	$n_{1\,2}$		$n_{1\,r}$	$n_{1\,\cdot}$
B	$n_{2\,1}$	$n_{2\,2}$		$n_{2\,r}$	$n_{2\,\cdot}$
....	$\vdots$	$\vdots$		$\vdots$	$\vdots$
k	$n_{k\,1}$	n_{k2}		$n_{k\,r}$	$n_{k\,\cdot}$
total	$n_{\cdot\,1}$	$n_{\cdot\,2}$		$n_{\cdot\,r}$	n

Tabela de Contingência

As **Tabelas de Contingência** permitem representar os dados, quer sejam eles qualitativos ou quantitativos.

Variável 1	Variável 2 A	B	...	r	total
A	n_{11}	n_{12}		n_{1r}	$n_{1\cdot}$
B	n_{21}	n_{22}		n_{2r}	$n_{2\cdot}$
....	$\vdots$	$\vdots$		$\vdots$	$\vdots$
k	n_{k1}	n_{k2}		n_{kr}	$n_{k\cdot}$
total	$n_{\cdot 1}$	$n_{\cdot 2}$		$n_{\cdot r}$	n

Tabela de Contingência

Nas **Tabelas de Contingência** podemos ter os dados das linhas representados por um critério e os dados das colunas representados por outro critério totalmente diferente.

	Variável 2				
Variável 1	**A**	**B**	**...**	**r**	**total**
A	$n_{1\,1}$	$n_{1\,2}$		$n_{1\,r}$	$n_{1\,\cdot}$
B	$n_{2\,1}$	$n_{2\,2}$		$n_{2\,r}$	$n_{2\,\cdot}$
....	$\vdots$	$\vdots$		$\vdots$	$\vdots$
k	$n_{k\,1}$	n_{k2}		$n_{k\,r}$	$n_{k\,\cdot}$
total	$n_{\cdot\,1}$	$n_{\cdot\,2}$		$n_{\cdot\,r}$	n

Tabela de Contingência

Nós usamos **Tabela de Contingência** para comparar 2 variáveis. As **Tabelas de Contingência** são muito utilizadas com probabilidades.

	Variável 2				
Variável 1	**A**	**B**	**...**	**r**	**total**
A	$n_{1\,1}$	$n_{1\,2}$		$n_{1\,r}$	$n_{1\,\cdot}$
B	$n_{2\,1}$	$n_{2\,2}$		$n_{2\,r}$	$n_{2\,\cdot}$
....	⋮	⋮		⋮	⋮
k	$n_{k\,1}$	n_{k2}		$n_{k\,r}$	$n_{k\,\cdot}$
total	$n_{\cdot\,1}$	$n_{\cdot\,2}$		$n_{\cdot\,r}$	n

Tabela de Contingência

Uma pesquisa de imóveis na área rural de uma cidade classificou as residências em duas categorias de preço (baixa - menos de R$ 150.000 e alta - acima de R$ 150.000).

A pesquisa também observou se as casas tinham pelo menos dois banheiros ou não (verdadeiro ou falso).

Cerca de 56% das casas tinham pelo menos dois banheiros, 62% das casas tinham um preço baixo e 22% das casas tinham ambos. Isso é informação suficiente para preencher a tabela.

Tabela de Contingência

		Dois Banheiros		
		Verdadeiro	Falso	Total
Preço	Baixo	0.22	0.40	0.62
	Alto	0.34	0.04	0.38
	Total	0.56	0.44	1.00

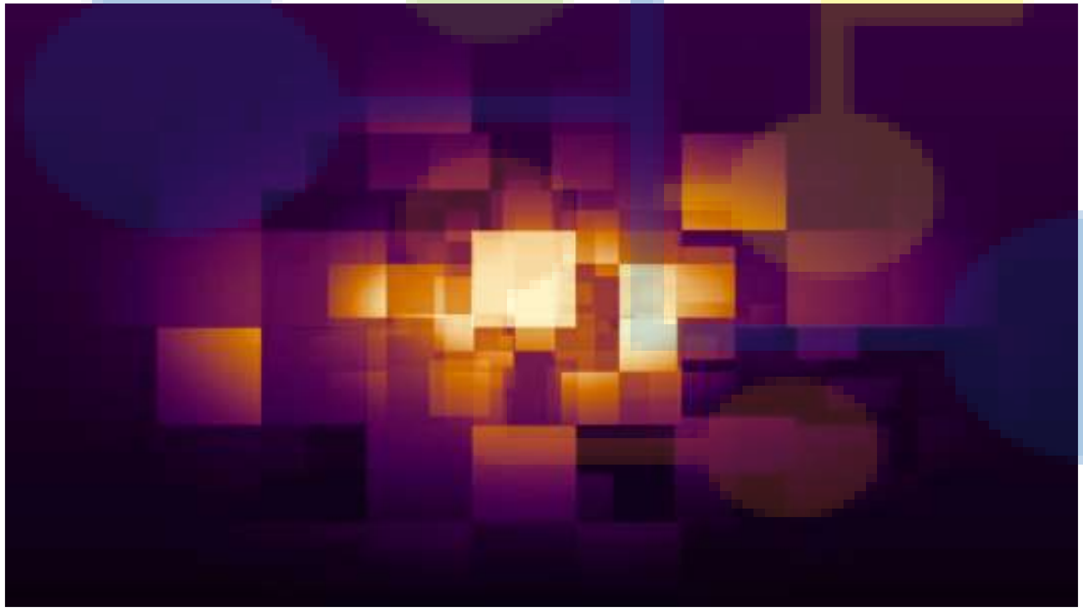
Agora, encontrar qualquer outra probabilidade é simples. Por exemplo, qual é a probabilidade de uma casa de alto preço ter pelo menos dois banheiros?

P (pelo menos dois banheiros | alto preço) = P (pelo menos dois banheiros e alto preço) / P (alto preço)
= 0,34 / 0,38 = 0,895 ou 89,5%.

Distribuições de Probabilidade Discreta e Contínua

Distribuições de Probabilidade Discreta e Contínua

Em Estatística, uma **Distribuição de Probabilidade** descreve a **chance** que uma variável (discreta ou contínua) pode assumir ao longo de um espaço de valores.



Distribuições de Probabilidade Discreta e Contínua

Uma **Distribuição de Probabilidade** é um modelo matemático que relaciona um certo valor da variável de estudo com a sua probabilidade de ocorrência.

Distribuições de Probabilidade Discreta e Contínua

O conjunto de todos os possíveis resultados de um experimento aleatório é denominado espaço amostral.

Distribuições de Probabilidade Discreta e Contínua

O resultado de um experimento de probabilidade geralmente é uma contagem ou uma medida. Quando isso ocorre, o resultado é chamado de **variável aleatória**.

Distribuições de Probabilidade Discreta e Contínua

As variáveis aleatórias podem ser de dois tipos: discretas ou contínuas.

Distribuições de Probabilidade Discreta e Contínua

Discreta

Contínua

Distribuições de Probabilidade Discreta e Contínua

- Tipos de Distribuição de Probabilidade
  - Distribuição Discreta
    - Binomial
    - Poisson
    - Bernoulli
    - Geométrica
    - Hipergeométrica
  - Distribuição Contínua
    - Normal
    - Uniforme
    - Exponencial
    - Gama
    - t de Student

Distribuições de Probabilidade Discreta e Contínua

Distribuições Discretas
Distribuição Binomial

Distribuições de Probabilidade Discreta e Contínua

A **Distribuição Binomial** é utilizada para descrever cenários em que os resultados de uma variável aleatória podem ser agrupados em duas categorias.

Distribuições de Probabilidade Discreta e Contínua

No geral, as duas categorias de uma distribuição binomial são classificadas como:

Sucesso

Falha

Distribuições de Probabilidade Discreta e Contínua

Portanto, a probabilidade de sucesso podemos chamar de **p**.

E a probabilidade de falha vamos chamar de **q.**

Distribuições de Probabilidade Discreta e Contínua

Ou seja:

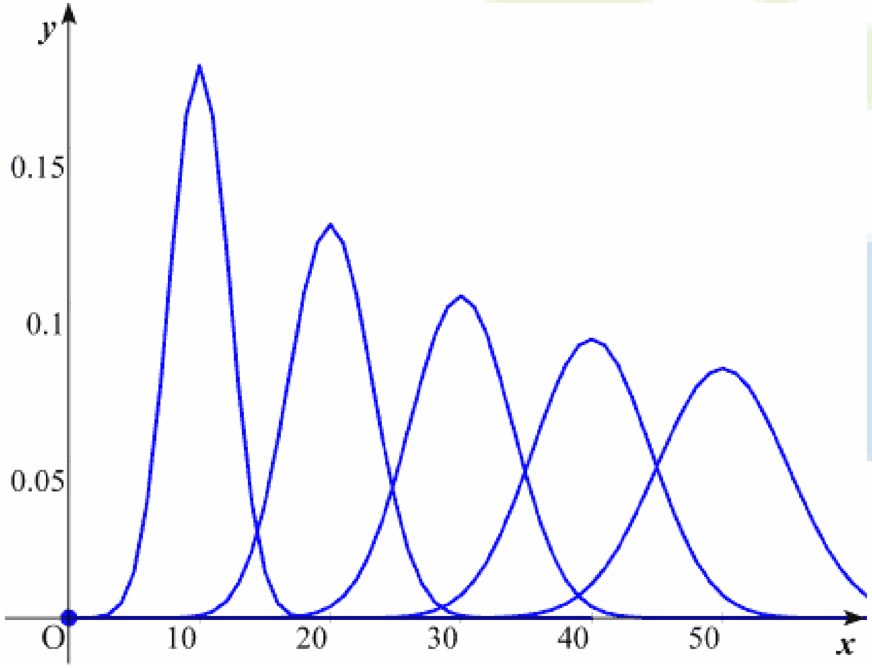
$$p = 1 - q$$

Onde:

p = probabilidade de sucesso

q = probabilidade de fracasso

Distribuições de Probabilidade Discreta e Contínua



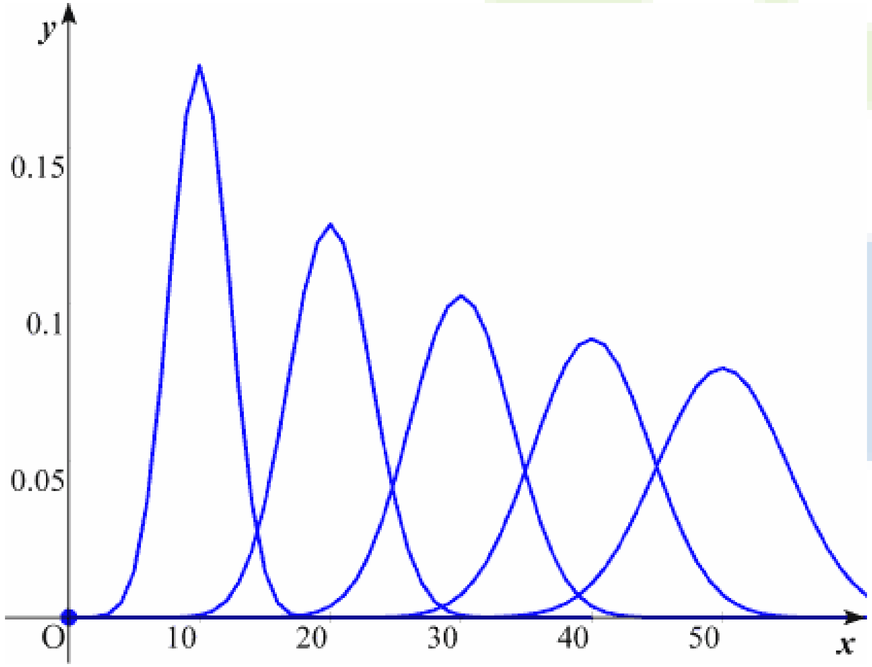
São realizadas **n** repetições no experimento, onde **n** é uma constante.

Só existem dois resultados possíveis em cada repetição, Sucesso e Falha.

A probabilidade de sucesso e a de falha permanencem constantes em todas as repetições.

Todas as repetições são independentes. Os resultados não são influenciados por resultados externos.

Distribuições de Probabilidade Discreta e Contínua



Os parâmetros da Distribuição Binominal são **n** e **p**.

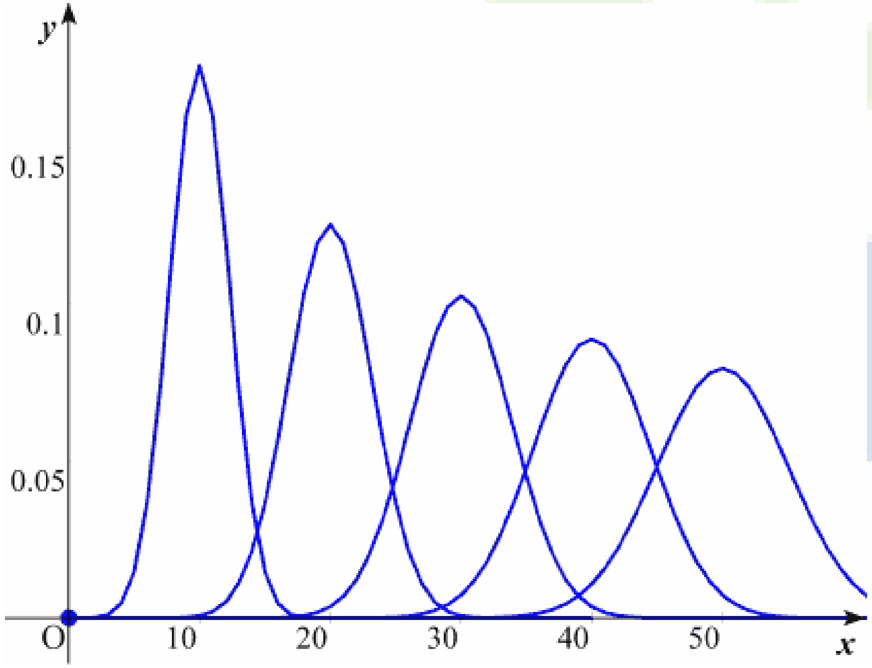
A **Média de uma Distribuição Binomial**, representa a média de longo prazo de sucessos esperados, com base no **número de observações.**

Fórmula:

Média = μ = n . P

Onde: n = número de tentativas
p = probabilidade de sucesso

Distribuições de Probabilidade Discreta e Contínua



A **Variância** de uma **Distribuição Binomial**, representa a variação que existe no número de sucessos **(p)** sobre um número **(n)** de observações.

Fórmula:

$$\text{Variância} = \sigma^2 = (n.p).(1-p)$$

Onde: n = número de tentativas
p = probabilidade de sucesso

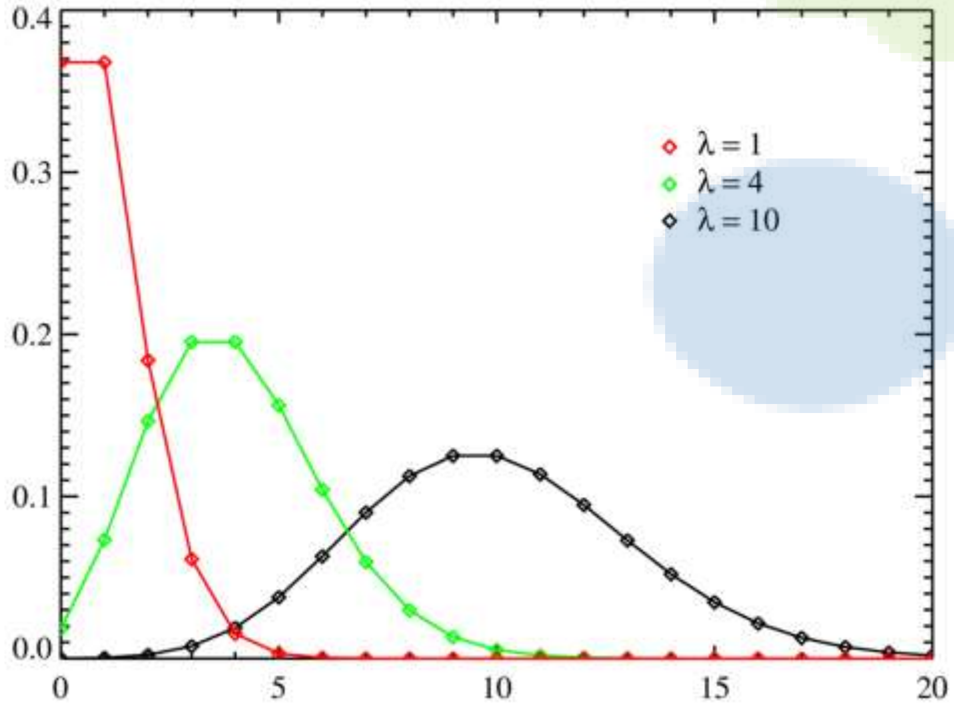
Distribuições de Probabilidade Discreta e Contínua

Distribuições Discretas
Distribuição Poisson

Distribuições de Probabilidade Discreta e Contínua

A **Distribuição Poisson** é utilizada para descrever cenários onde existe a probabilidade de ocorrência do evento em um intervalo contínuo.

Distribuições de Probabilidade Discreta e Contínua



O número de ocorrências depende do tamanho do intervalo.

As ocorrências não interferem sobre as ocorrências de intervalos externos.

A probabilidade de duas ou mais ocorrências acontecerem num mesmo intervalo de tempo é muito pequena.

Distribuições de Probabilidade Discreta e Contínua

A Distribuição Poisson é caracterizada pelo parâmetro único chamado λ **(lambda), que representa a taxa média de ocorrência por unidade de medida.**

$$P(x) = \frac{e^{-\lambda} . \lambda^{x}}{x!}$$

Distribuições de Probabilidade Discreta e Contínua

Distribuições Discretas
Distribuição Hipergeométrica

Distribuições de Probabilidade Discreta e Contínua

Um dos pontos chave das **Distribuições Binomial** e **Poisson** é que os **eventos são independentes** uns dos outros.

Cada amostra de cada experimento é um conjunto novo de dados.

Desta forma, a **probabilidade de sucesso** ou de número de ocorrências, se mantém **constante**.

Distribuições de Probabilidade Discreta e Contínua

A **Distribuição Hipergeométrica** é uma distribuição de probabilidade discreta que descreve o número de sucessos numa sequência de **n** extrações de uma população finita, ou seja, sem reposição.

Distribuições de Probabilidade Discreta e Contínua

Quando a amostragem é sem substituição, a probabilidade de **sucesso muda** durante o processo de amostragem, isso viola os requisitos para uma distribuição de probabilidade binomial.

Nesse caso, use a **Distribuição Hipergeométrica**.

Distribuições de Probabilidade Discreta e Contínua

Fórmula da **Distribuição Hipergeométrica**

$$P(x) = \frac{{}_{N-R}C_{n-x} \times {}_{R}C_{x}}{{}_{N}C_{n}}$$

onde:

N = Tamanho da população
R = O número de sucessos da população
n = Tamanho da Amostra
x = Número de sucessos da amostra

Distribuições de Probabilidade Discreta e Contínua

Assim como as outras distribuições, a **Distribuição Hipergeométrica** também possui **média** e **desvio padrão.**

Distribuições de Probabilidade Discreta e Contínua

$$m = \frac{nR}{N}$$

onde:

N = Tamanho da população
R = O número de sucessos da população
n = Tamanho da Amostra

$$s = \sqrt{\frac{nR(N - R)}{N^2}}\sqrt{\frac{N - n}{N - 1}}$$

onde:

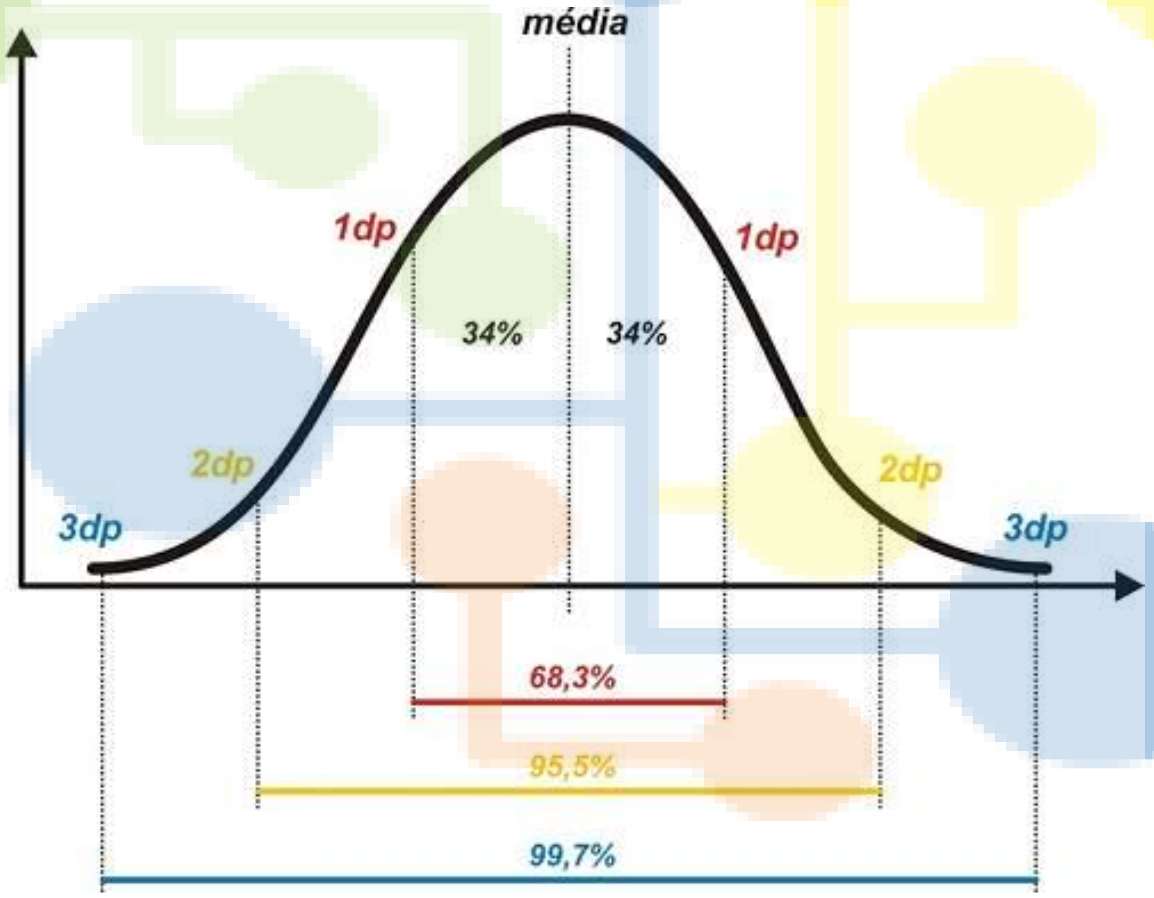
N = Tamanho da população
R = O número de sucessos da população
n = Tamanho da Amostra

Distribuições de Probabilidade Discreta e Contínua

Distribuições Contínuas

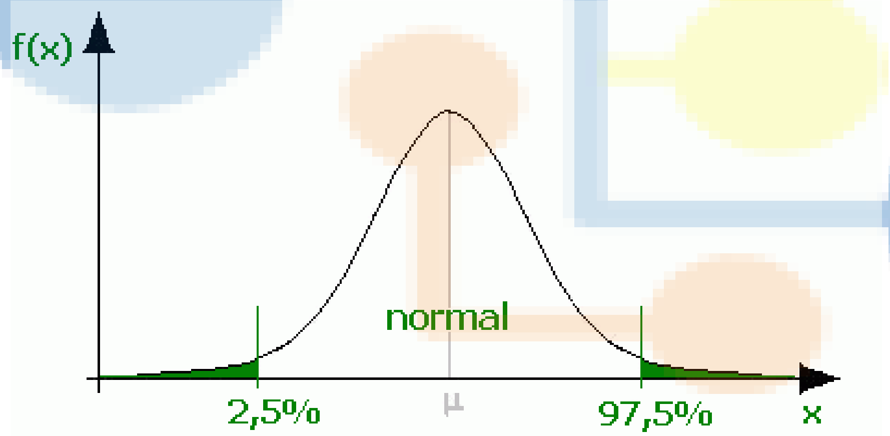
Distribuição Normal

Distribuições de Probabilidade Discreta e Contínua



Distribuições de Probabilidade Discreta e Contínua

A Distribuição Normal é útil quando os dados tendem a estar próximos ao **centro da distribuição** (**próximos da média**) e quando **valores extremos** (**outliers**) são muito raros.

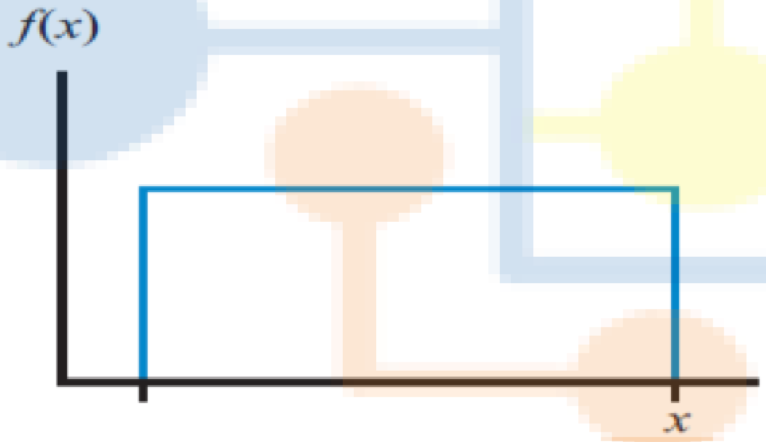


Distribuições de Probabilidade Discreta e Contínua

Distribuições Contínuas
Distribuição Uniforme

Distribuições de Probabilidade Discreta e Contínua

A **Distribuição Uniforme** é usada para descrever os dados quando todos os valores têm a mesma chance de ocorrer.



Distribuições de Probabilidade Discreta e Contínua

A **Distribuição Uniforme** é a distribuição de probabilidades contínua mais simples de conceituar: a probabilidade de se gerar qualquer ponto em um intervalo contido no espaço amostral é proporcional ao tamanho do intervalo, visto que na distribuição uniforme a f(x) é igual para qualquer valor de x no intervalo considerado.

Distribuições de Probabilidade Discreta e Contínua

Outra maneira de se dizer "distribuição uniforme" seria "um número finito de resultados com chances iguais de acontecer".

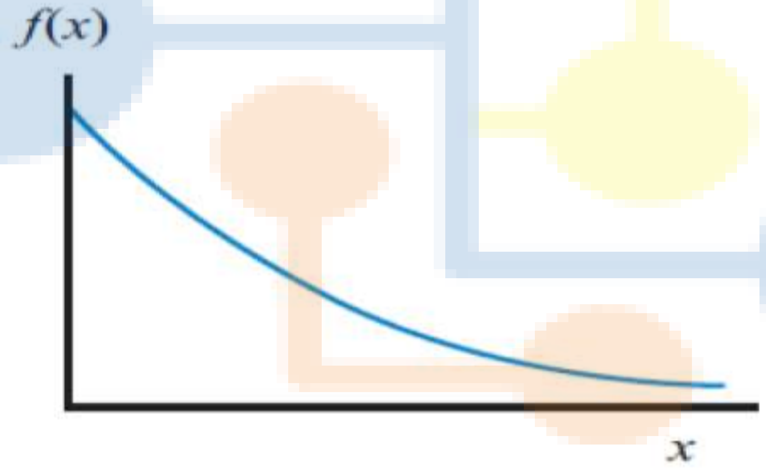
Ela é usada quando assumimos intervalos iguais da variável aleatória que tem a mesma probabilidade.

Distribuições de Probabilidade Discreta e Contínua

Distribuições Contínuas
Distribuição Exponencial

Distribuições de Probabilidade Discreta e Contínua

A **Distribuição Exponencial** é usada para descrever os dados quando valores mais baixos tendem a dominar a distribuição e quando valores muito altos não ocorrem com frequência .



Distribuições de Probabilidade Discreta e Contínua

Na **Distribuição Poisson**, a variável aleatória é definida como o número de ocorrências em determinado período, sendo a média das ocorrências definida como λ (lambda).

Na **Distribuição Exponencial,** a variável aleatória é definida como o tempo entre duas ocorrências, sendo a média de tempo entre as ocorrências de 1 / λ.

Distribuições de Probabilidade Discreta e Contínua

A **Distribuição Exponencial** é amplamente usada no campo da confiabilidade, como um modelo para a distribuição dos tempos até a falha de componentes eletrônicos.

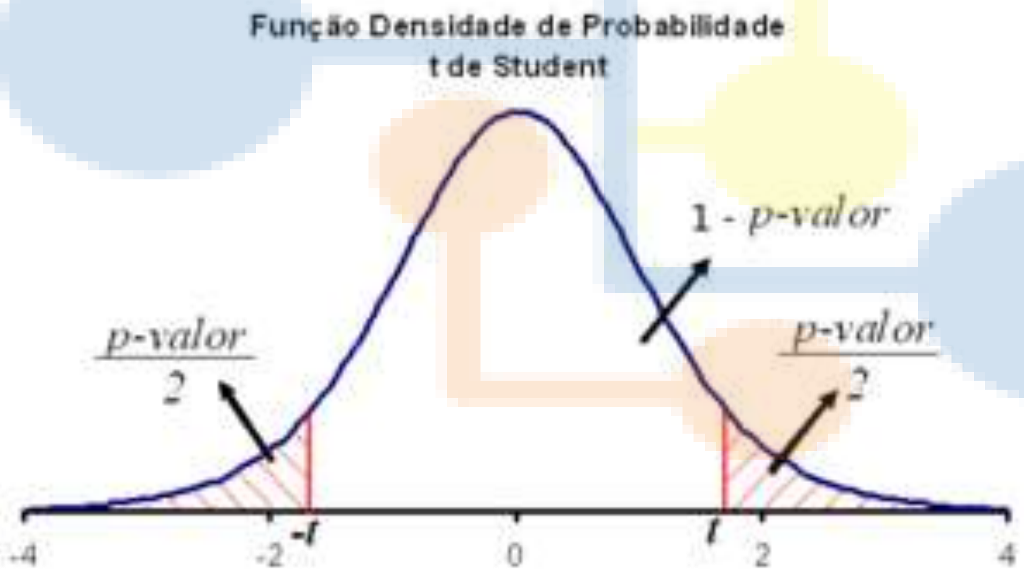
Nessas aplicações, o parâmetro λ representa a taxa de falha para o componente e 1 / λ é o tempo médio até a falha!

Distribuições de Probabilidade Discreta e Contínua

Distribuições Contínuas
Distribuição t de Student

Distribuições de Probabilidade Discreta e Contínua

A **Distribuição t de Student** é uma das principais distribuições de probabilidade, com inúmeras aplicações em inferência estatística.



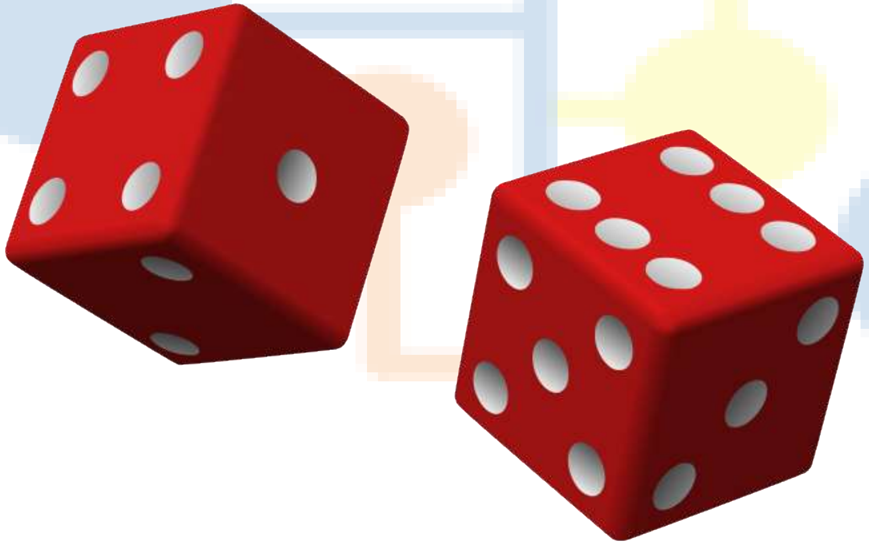
Distribuições de Probabilidade Discreta e Contínua

Resumindo

Na caracterização das distribuições de probabilidade é de grande importância a utilização de medidas que indiquem aspectos relevantes da distribuição, como medidas de posição (média, mediana e moda), medidas de dispersão (variância e desvio-padrão) e medidas de assimetria e curtose.

O entendimento dos conceitos relativos a probabilidade e distribuições de probabilidade auxiliarão o Cientista de Dados no estudo de tópicos sobre inferência estatística, incluindo testes de hipóteses paramétricos e não paramétricos, análise de regressão, etc.

A Distribuição Normal



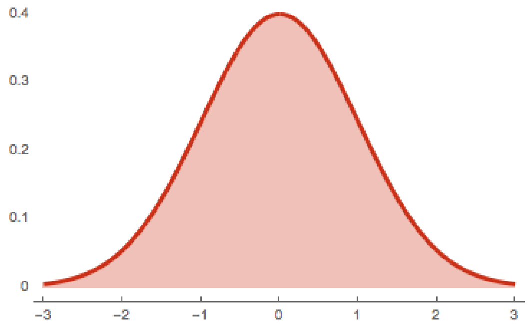
A Distribuição Normal

Uma distribuição estatística é uma função que define uma curva e a área sob essa curva determina a probabilidade de ocorrer o evento por ela correlacionado.

A Distribuição Normal

Você provavelmente já viu a Distribuição Normal antes, e se você já viu uma curva em forma de “bellshaped” (formato de sino), era provavelmente um modelo Normal.

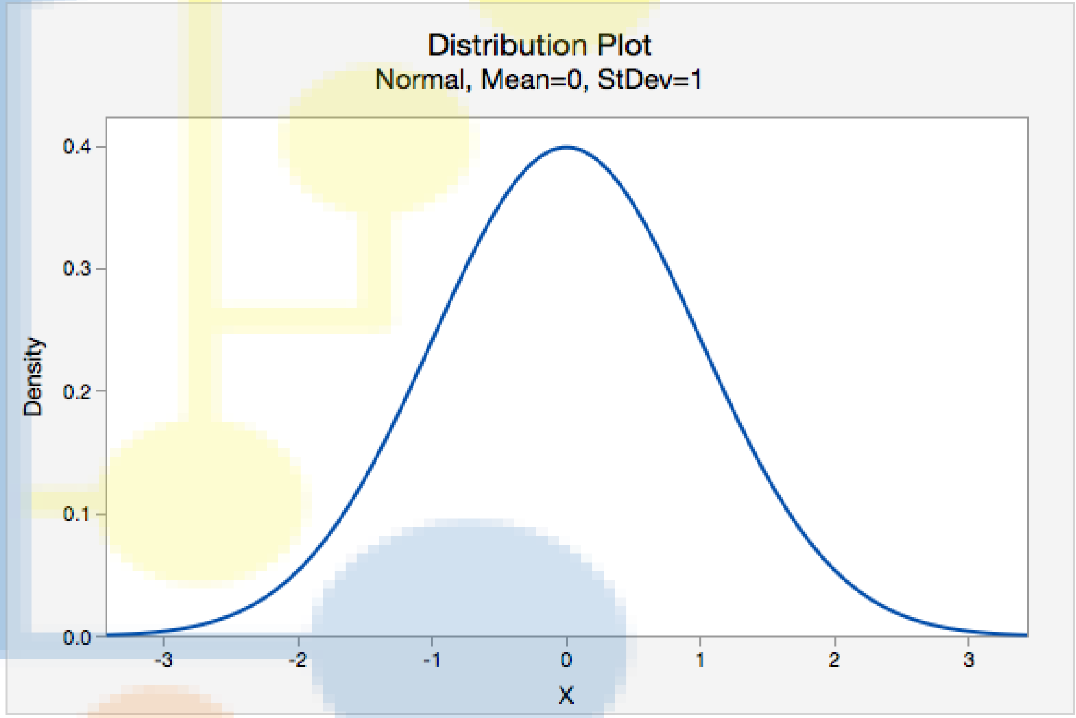
Modelos normais são definidos por dois parâmetros: uma média e um desvio padrão. Por convenção, indicamos os parâmetros com letras gregas. Por exemplo, representamos a média de tal modelo com a letra grega μ, que é o equivalente grego de "m", para média, e o desvio padrão com a letra grega σ, o equivalente grego de "s", para padrão desvio.



A Distribuição Normal

A Distribuição Normal é a mais importante dentre as distribuições estatísticas. Também conhecida como Distribuição Gaussiana, é uma curva simétrica em torno do seu ponto médio, apresentando assim seu famoso formato de sino.

A curva normal representa o comportamento de diversos processos nas empresas e muitos fenômenos comuns, como por exemplo, altura ou peso de uma população, a pressão sanguínea de um grupo de pessoas, o tempo que um grupo de estudantes gasta para realizar uma prova.



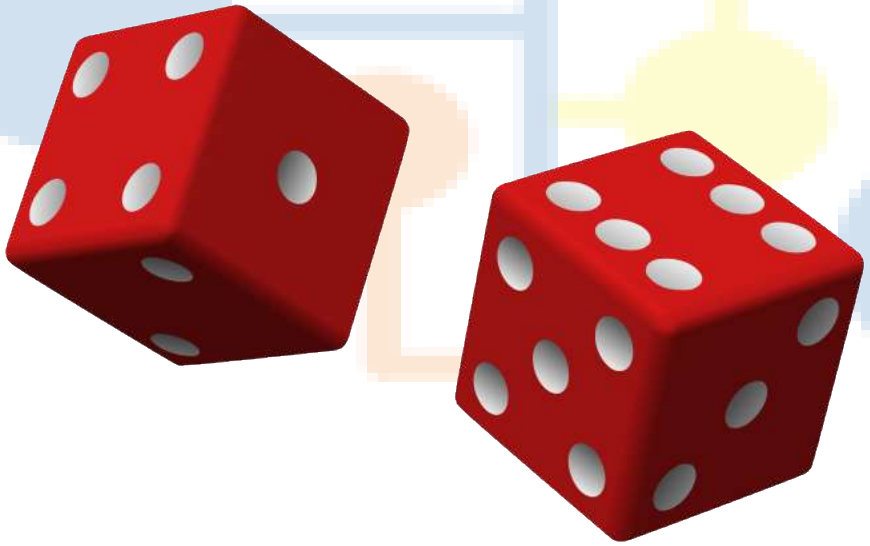
A Distribuição Normal

Há um **Modelo Normal** diferente para cada combinação de m e s, mas se padronizarmos nossos dados primeiro, criando z-scores e subtraindo da média para fazer a média igual a 0 e dividindo pelo desvio padrão para fazer o desvio padrão igual a 1, então precisaremos apenas do modelo com média 0 e desvio padrão 1. Chamamos isso de Modelo Normal Padrão ou Distribuição Normal Padrão (Standard Normal Distribution.).

A Distribuição Normal

Obviamente, não devemos usar um modelo Normal para todos os conjuntos de dados. Se o histograma não for em forma de sino, as pontuações (scores) z não serão serão bem modeladas pelo modelo Normal. E a padronização não ajuda, porque a padronização não altera a forma da distribuição. Portanto, sempre verifique o histograma dos dados antes de usar o modelo Normal.

Como Determinar Se a Distribuição é Normal?



Como Determinar Se a Distribuição é Normal?

Para determinar se uma determinada variável aleatória segue uma distribuição normal, basta verificar se essa segue a função densidade de probabilidade, dada por:

$$f(x) = \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}}$$

Como Determinar Se a Distribuição é Normal?

$$f(x) = \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}}$$

Onde μ é a média e σ² é a variância de x.

A notação N(μ,σ²) é usada para representar tal distribuição. Para calcularmos então a probabilidade de um resultado, basta integrar a função f(x) em relação a x, com os limites de integração representando a faixa de valores que se quer obter a probabilidade. Vale notar que a integral da função densidade de probabilidade normal não possui solução analítica. Sendo assim, seu cálculo deve ser realizado através de um método numérico.

Como Determinar Se a Distribuição é Normal?

$$f(x) = \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}}$$

Onde μ é a média e σ^2 é a variância de x.

Para sanar tal dificuldade a função pode ser padronizada com a substituição dos parâmetros por $\mu = 0$ e $\sigma^2 = 1$. Essa abordagem é dada pela definição de uma nova variável aleatória Z, chamada de variável aleatória normal padronizada. Se x for uma variável aleatória normal com média $E(x) = \mu$ e variância $V(x) = \sigma^2$, a variável aleatória $Z = (x-\mu)/\sigma$ será uma variável aleatória normal, com $E(Z) = 0$ e $V(Z) = 1$.

Como Determinar Se a Distribuição é Normal?

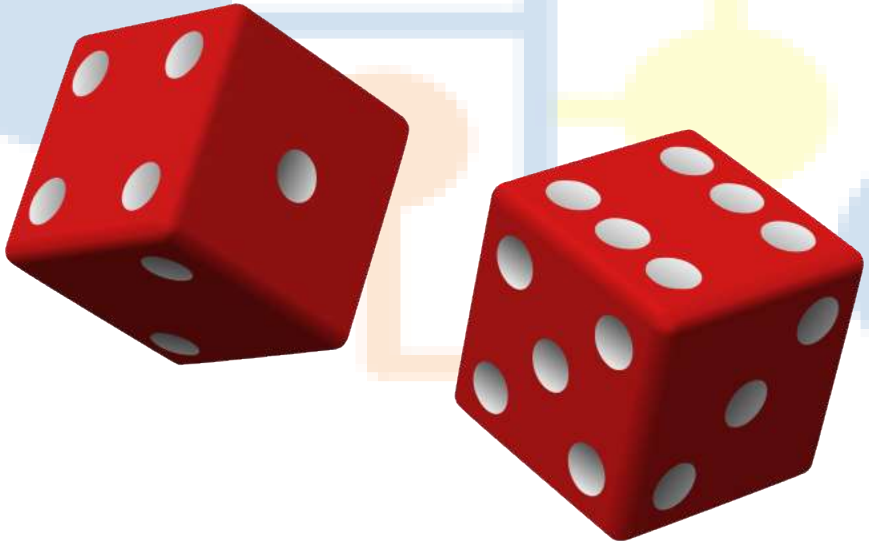
$$f(x) = \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}}$$

Onde μ é a média e σ² é a variância de x.

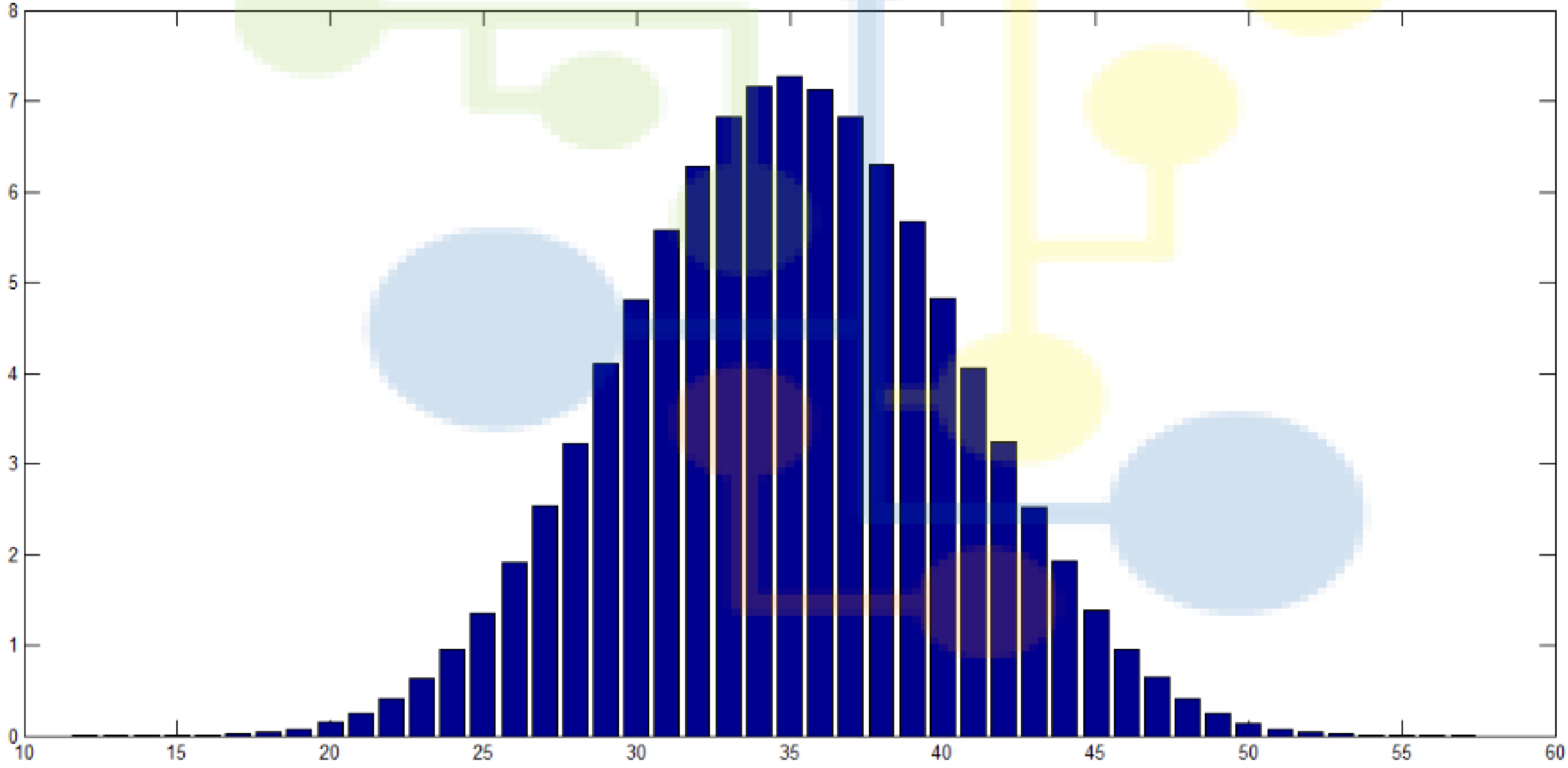
Ou seja, Z é uma variável aleatória normal padrão.

Dessa forma, é possível obter a área sob a curva da normal padrão de forma analítica, e então obter a área entre dois pontos sob a curva, diretamente com o uso de uma tabela de conversão, e essa área representa uma probabilidade.

Teorema do Limite Central



Teorema do Limite Central

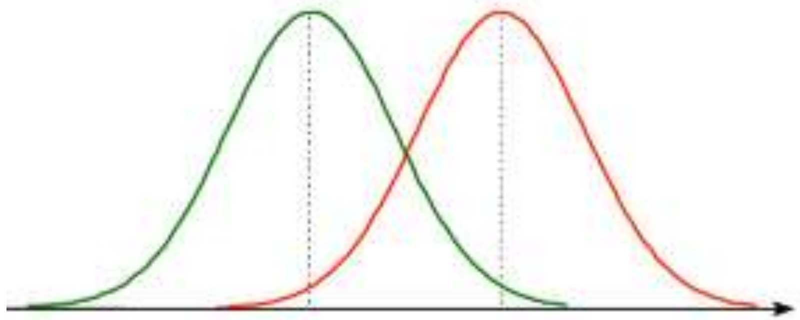


Teorema do Limite Central

O **Teorema do Limite Central** é fundamental para a Estatística, uma vez que diversos procedimentos estatísticos comuns requerem que os dados sejam aproximadamente **normais** e o Teorema do Limite Central permite aplicar esses procedimentos úteis a populações que são fortemente **não-normais.**

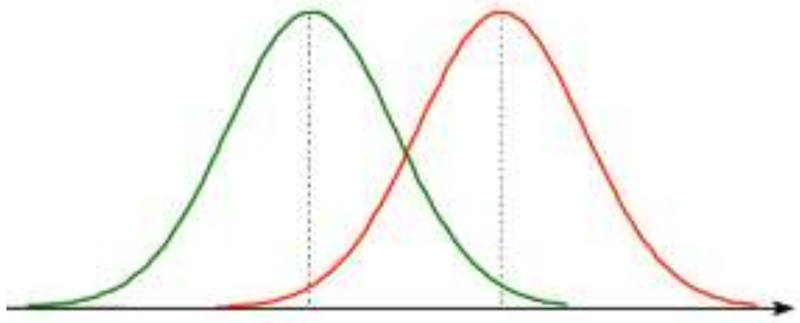
Teorema do Limite Central

Esse teorema possibilita medir o quanto sua média amostral irá variar, sem ter que pegar outra média amostral para fazer a comparação. Ou seja, permite-nos conduzir alguns procedimentos de inferência sem ter qualquer conhecimento de distribuição da população.



Teorema do Limite Central

Esse teorema basicamente diz que sua **média amostral** tem uma **Distribuição Normal**, independente da aparência da distribuição dos dados originais.



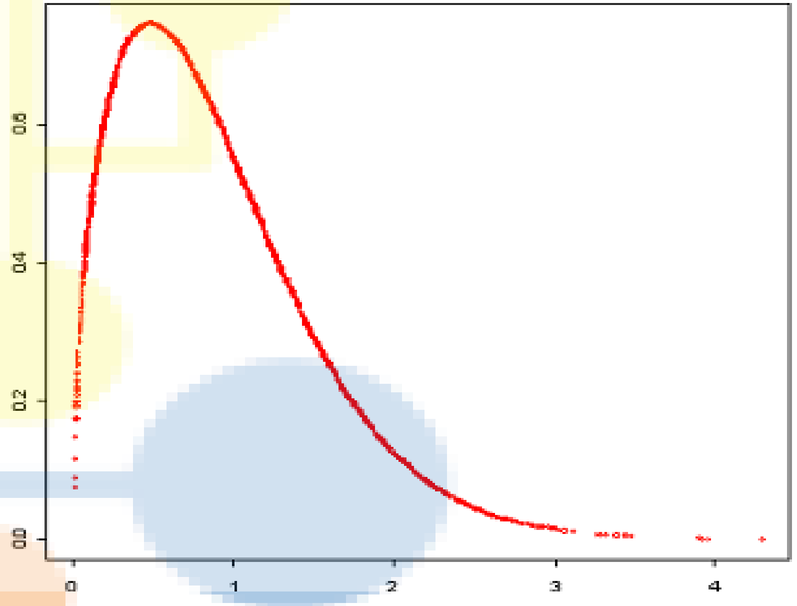
Teorema do Limite Central

Muitos procedimentos pressupõem que uma **Distribuição Normal** é uma **Distribuição Simétrica**.

Teorema do Limite Central

Assimetria indica variação no formato de distribuição.

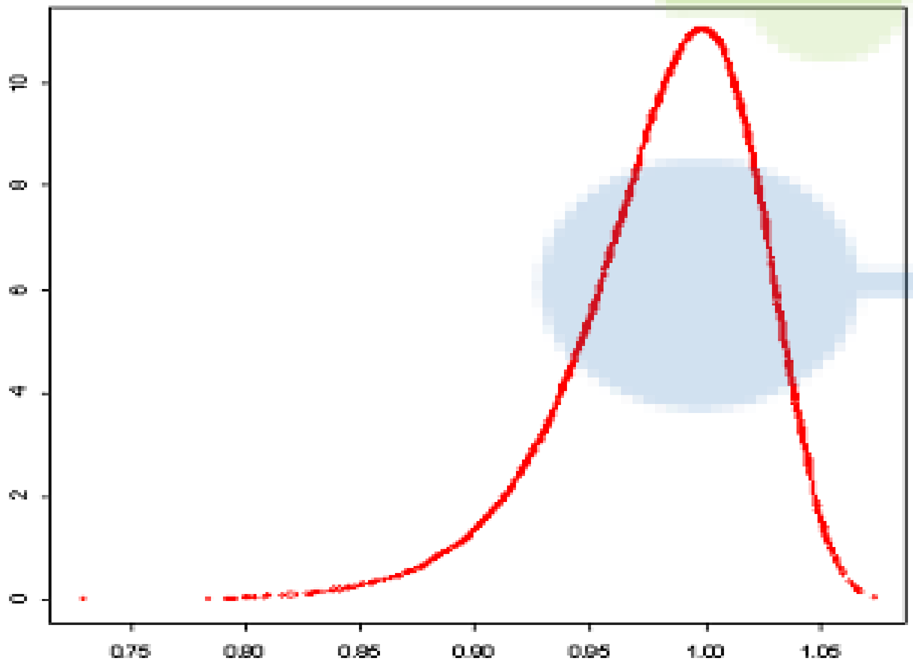
Assimetria Positiva - Implica em uma concentração maior de valores menores, e o gráfico possuirá uma cauda mais longa à direita.



Distribuição Assimétrica Positiva

Teorema do Limite Central

Assimetria indica variação no formato de distribuição.



Distribuição Assimétrica Negativa

Assimetria Negativa - implica em uma concentração de valores maiores, e o gráfico possuirá uma cauda maior à esquerda.

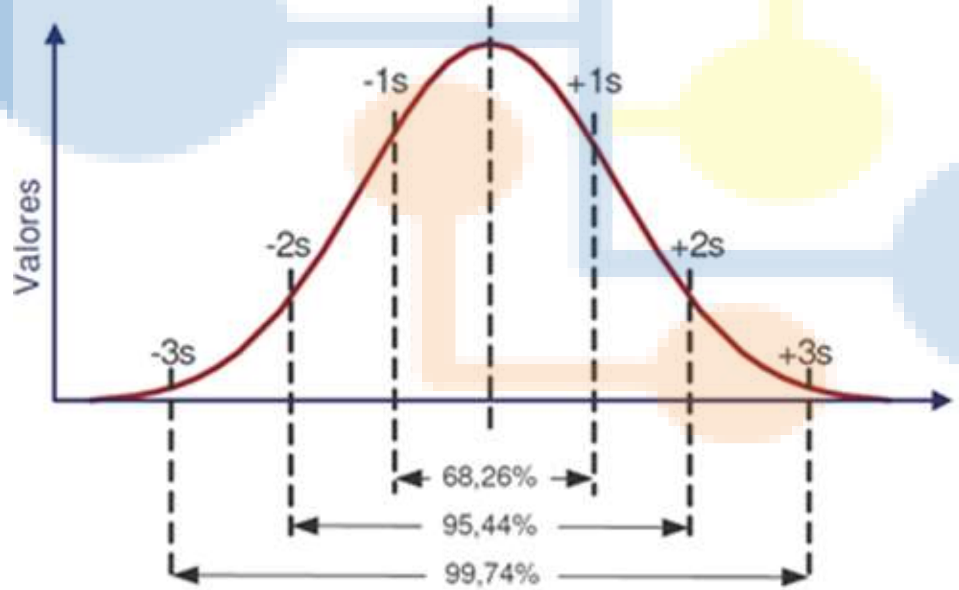
Teorema do Limite Central

Os valores de grandes conjuntos de dados, normalmente se localizam ao redor da média ou da mediana.

Desta forma, um histograma dos dados, mostraria uma curva simétrica bem definida (em forma de sino) em uma Distribuição Normal.

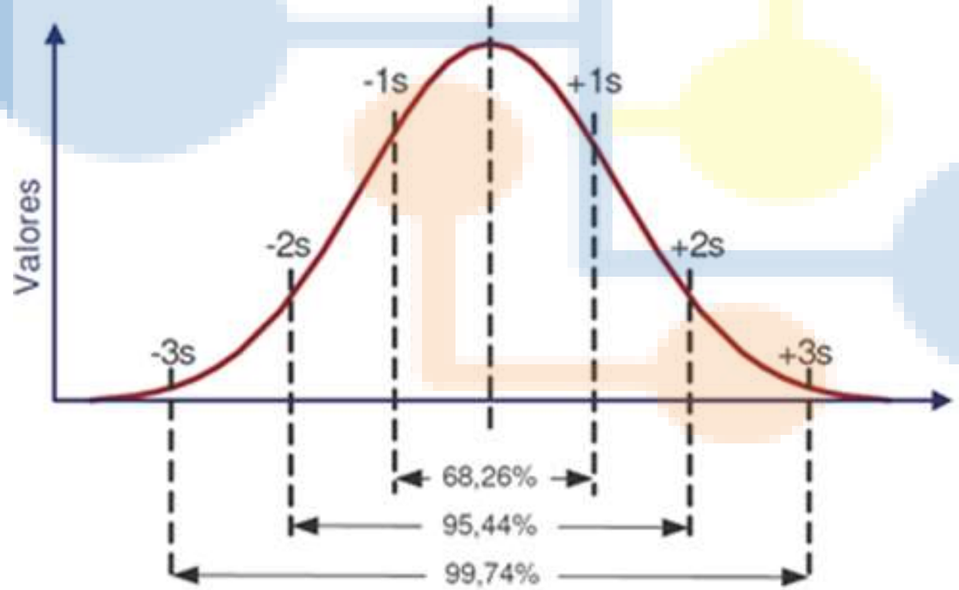
Teorema do Limite Central

Em uma **Distribuição Normal** de dados, simétrica, nós podemos esperar que 68%, 95% e 99.7% dos valores estarão em, respectivamente, 1, 2 e 3 desvios-padrão acima e abaixo da média.



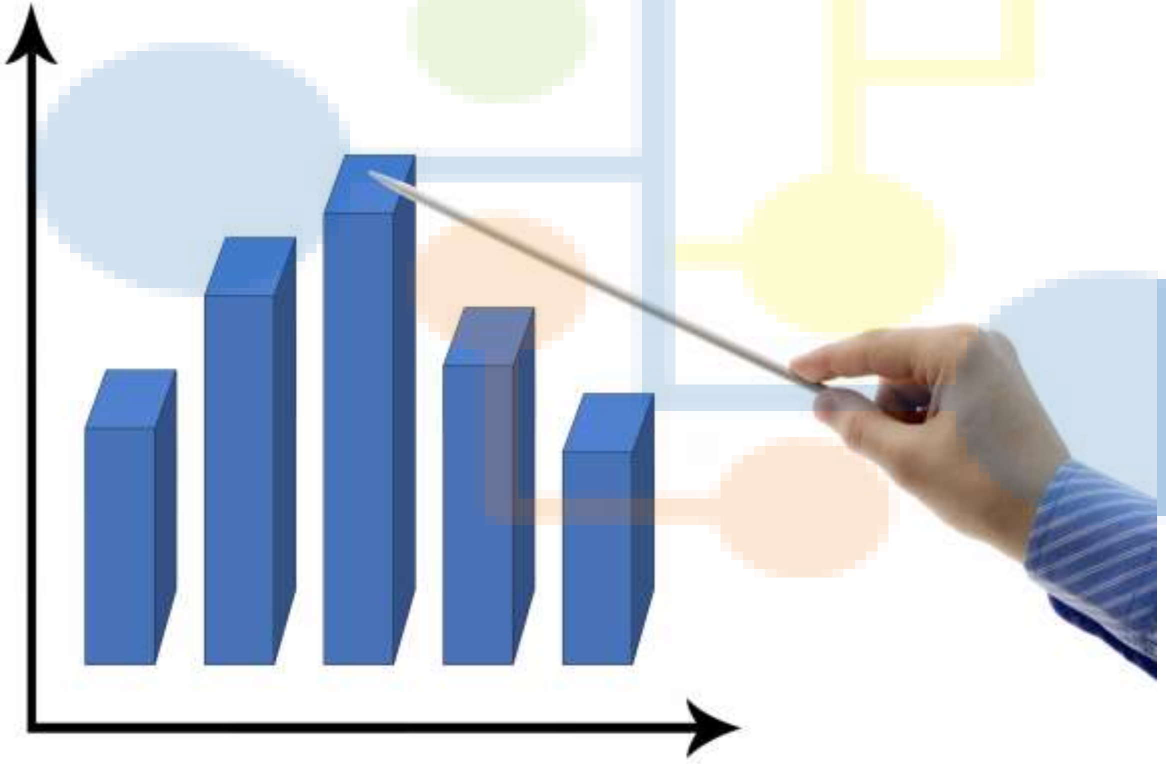
Teorema do Limite Central

Ou seja, em uma curva simétrica dos dados, praticamente todos os dados estarão em até 3 desvios-padrão do centro dos dados (média).



Teorema do Limite Central

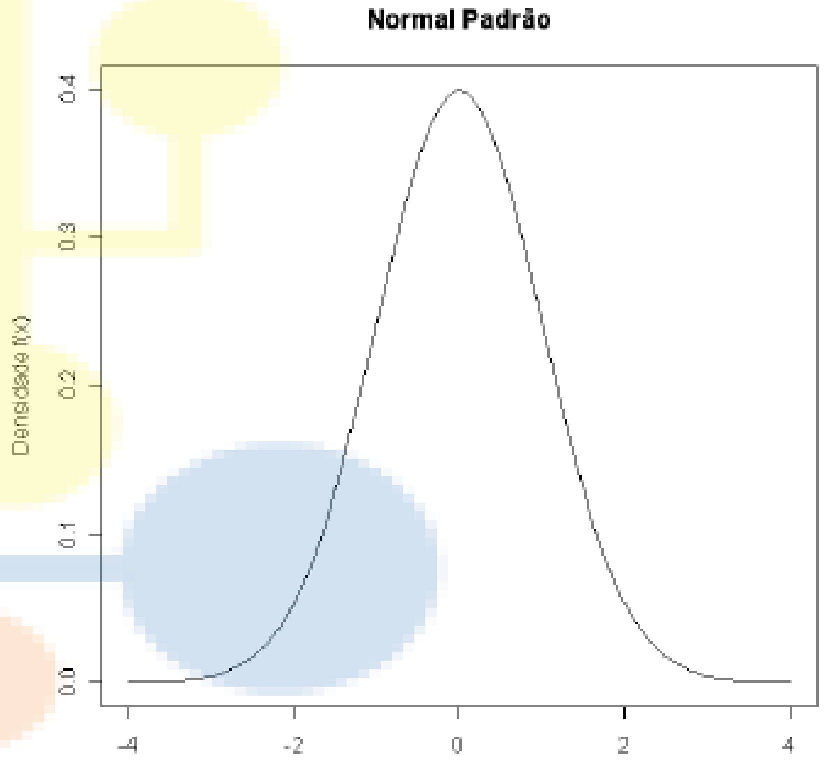
Perceba que este conceito somente se aplica, quando os dados criam um histograma simétrico.



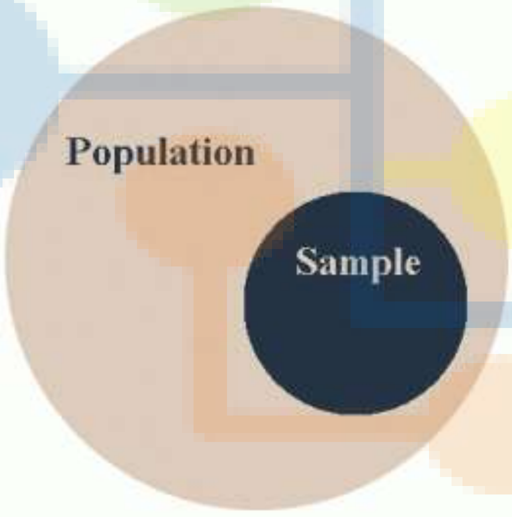
Teorema do Limite Central

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 compreendida entre esses dois pontos.



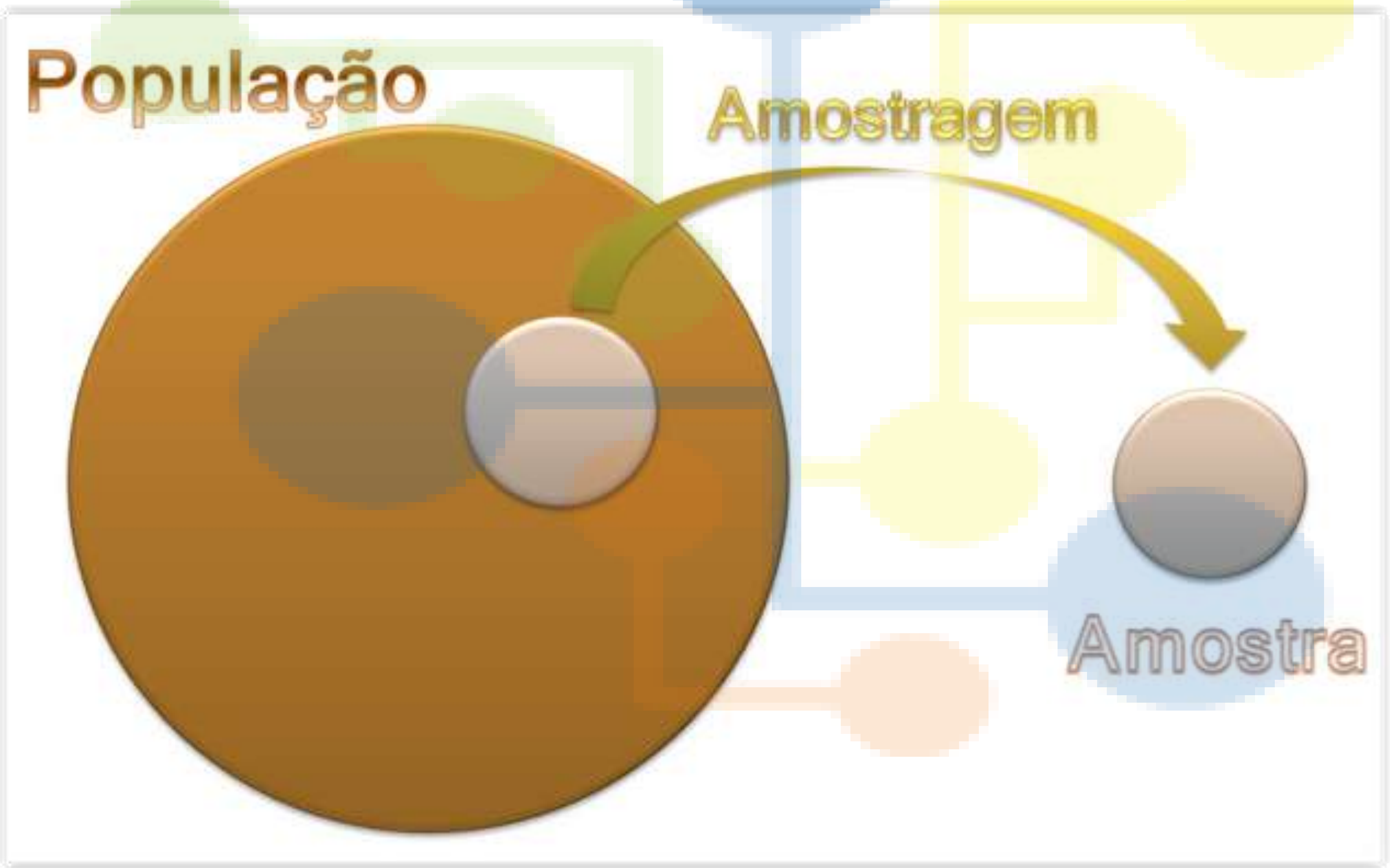
O Que é Estatística Inferencial?



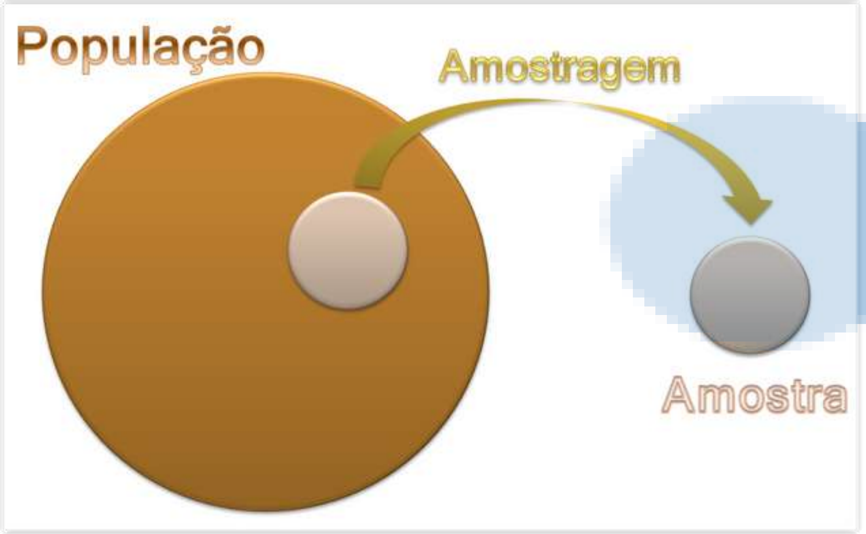
O Que é Estatística Inferencial?

Até aqui estudamos Estatística Descritiva, para descrever como os dados estão organizados e Probabilidade para medir a variabilidade de fenômenos casuais de acordo com a sua ocorrência.

O Que é Estatística Inferencial?

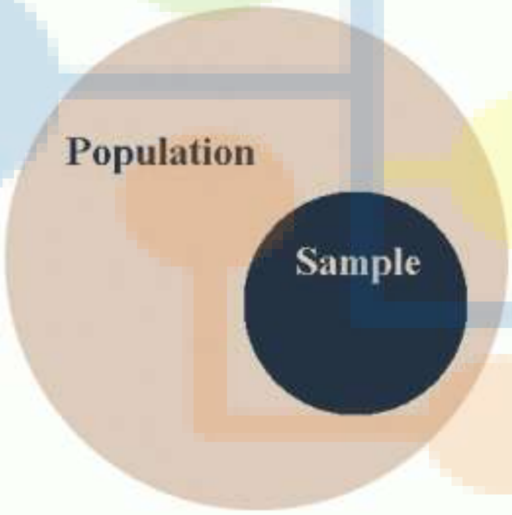


O Que é Estatística Inferencial?



A estatística inferencial tem como objetivo a extrapolação dos resultados (obtidos com a estatística descritiva) para a população.

População e Amostra



População e Amostra

Imagine que você seja convidado a realizar uma pesquisa para medir a durabilidade das lâmpadas produzidas por uma determinada fábrica.

Qual abordagem você usaria?

1- Testar TODAS as lâmpadas produzidas.
2- Obter uma amostra representativa da população de lâmpadas produzidas e então inferir a durabilidade de todas as lâmpadas.

População e Amostra

População

!"#$%&'()*+,-./0123456789:;ÿ
elementos ou resultados sob
investigação.

População e Amostra

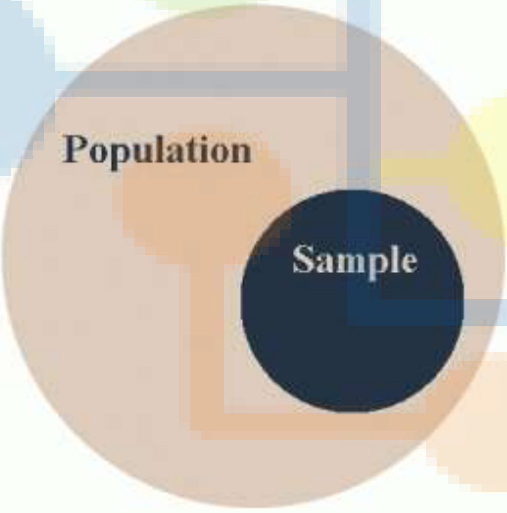
População

População é o conjunto de todos os elementos ou resultados sob investigação.

Amostra

Amostra é qualquer subconjunto da população.

Princípios da Amostragem



Princípios da Amostragem

”Para saber se o bolo de chocolate está bom, basta comer uma fatia.”

Princípios da Amostragem

Amostragem é o processo de determinação de uma amostra a ser pesquisada. A amostra é uma parte de elementos selecionada de uma população.

Princípios da Amostragem

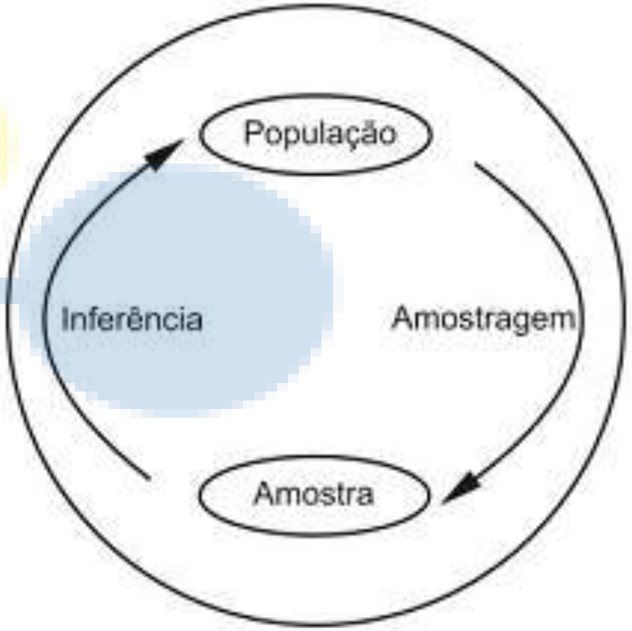
Enquanto que um censo envolve um exame a todos os elementos de um dado grupo, a amostragem envolve um estudo de apenas uma parte dos elementos.

Princípios da Amostragem

A amostragem consiste em selecionar parte de uma população e observá-la com vista a estimar uma ou mais características para a totalidade da população.

Princípios da Amostragem

A teoria da amostragem estuda as relações existentes entre uma população e as amostras extraídas dessa população. E útil para avaliação de grandezas desconhecidas da população, ou para determinar se as diferenças observadas entre duas amostras são devidas ao acaso ou se são verdadeiramente significativas.



Princípios da Amostragem

Exemplos:

- Sondagens à opinião pública que servem para conhecer a opinião da população sobre variadas questões. As mais populares são as sondagens políticas.

- Inspeção de mercado utilizada com o intuito de descobrir as preferências das pessoas em relação a certos produtos. Um dos exemplos mais conhecidos da aplicação desta amostragem é a lista de audiências dos programas de televisão.

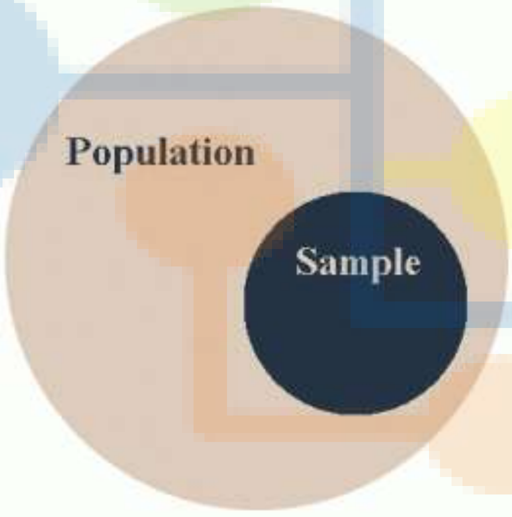
- Para estimar a prevalência de uma doença rara, a amostra pode ser constituída por algumas instituições médicas, cada uma das quais com registros dos pacientes.

Princípios da Amostragem

Termos Básicos da Amostragem

- População
- Unidade
- Amostra
- Variável

Tipos de Amostragem



Tipos de Amostragem

Métodos Aleatórios

Métodos Não Aleatórios

Tipos de Amostragem

- Amostra Intencional
- Amostra "Snowball"
- Amostra por quotas
- Amostra por conveniência

Métodos Não Aleatórios

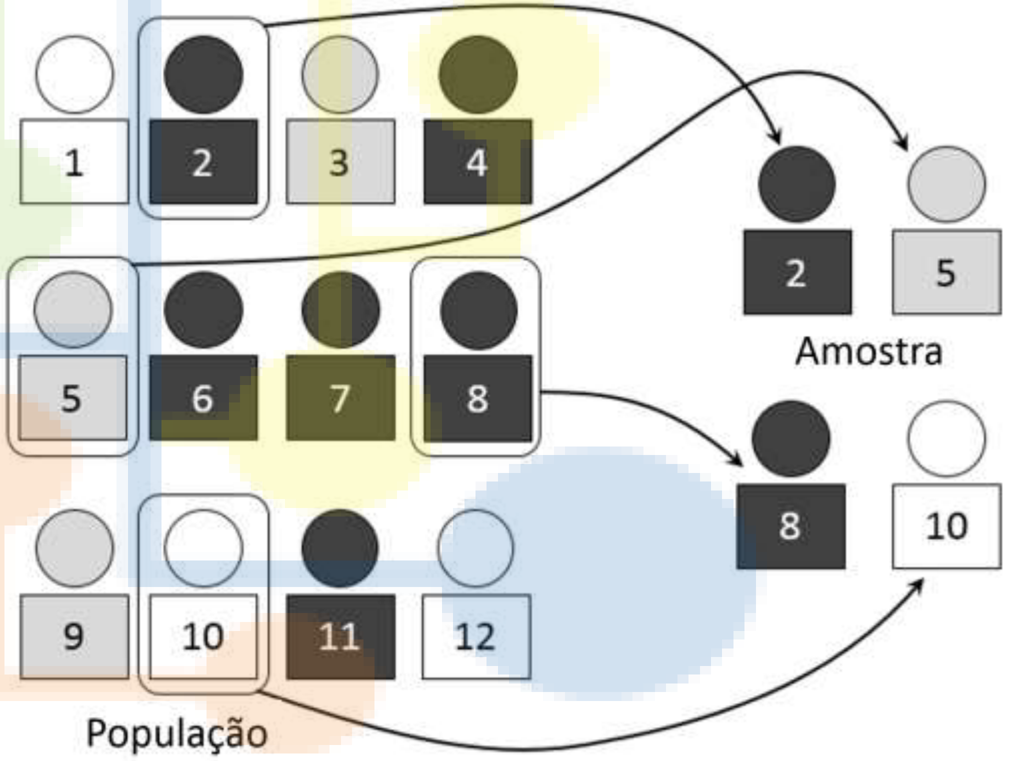
Tipos de Amostragem

Métodos Aleatórios

- Amostragem Aleatória Simples
- Amostragem Sistemática
- Amostragem Estratificada
- Amostragem por Aglomerados
- Amostragem Multi-etapas
- Amostragem Multifásica

Tipos de Amostragem

Amostragem Probabilística ou Aleatória

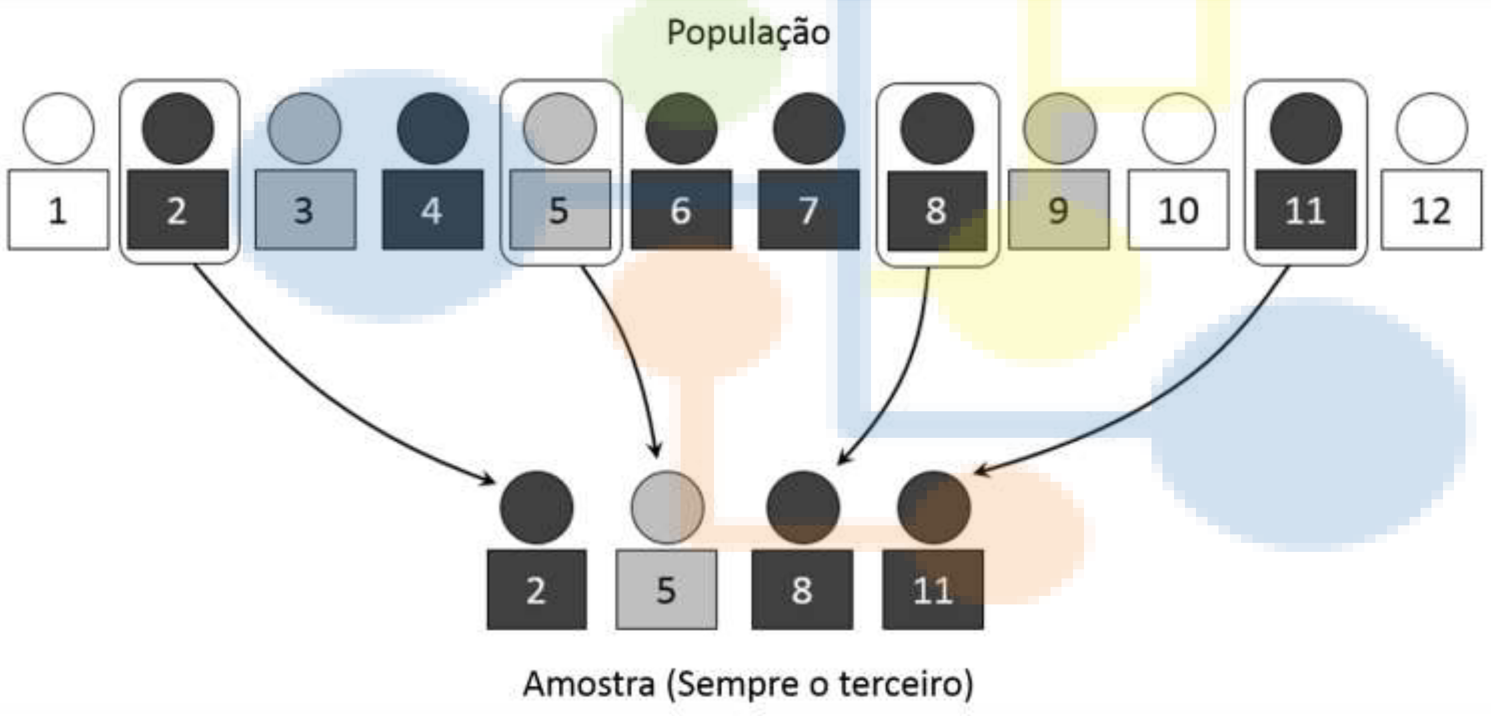


Tipos de Amostragem

- Amostragem Aleatória Simples
- Amostragem Aleatória Simples sem reposição
- Amostragem Aleatória Simples com reposição

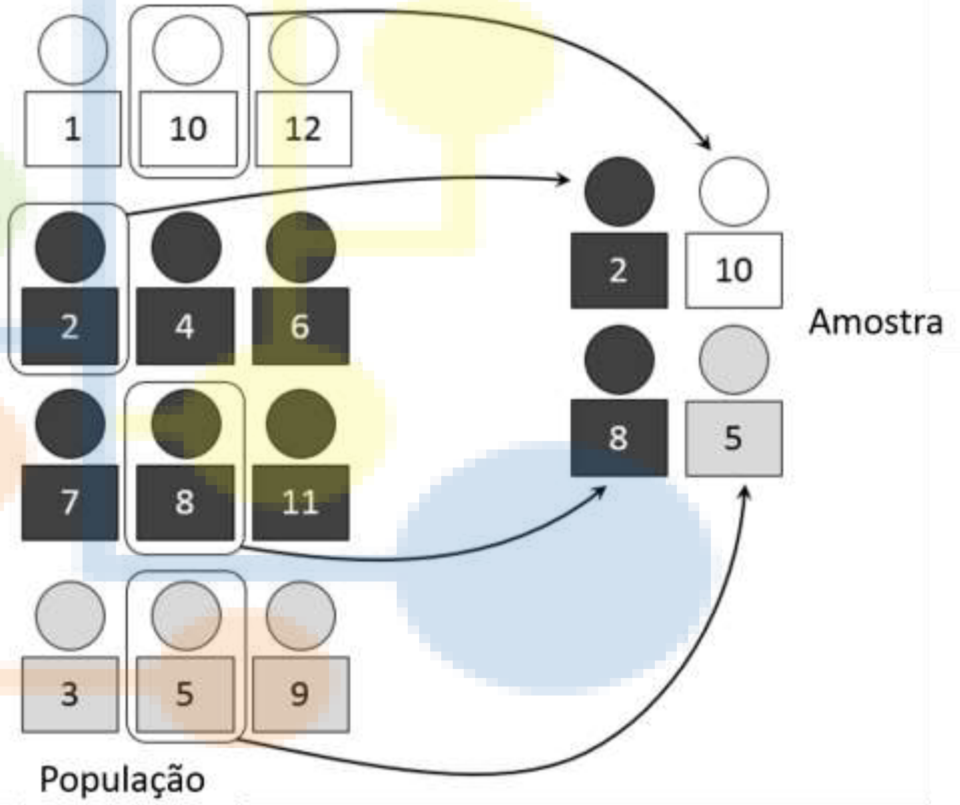
Tipos de Amostragem

Amostragem Sistemática



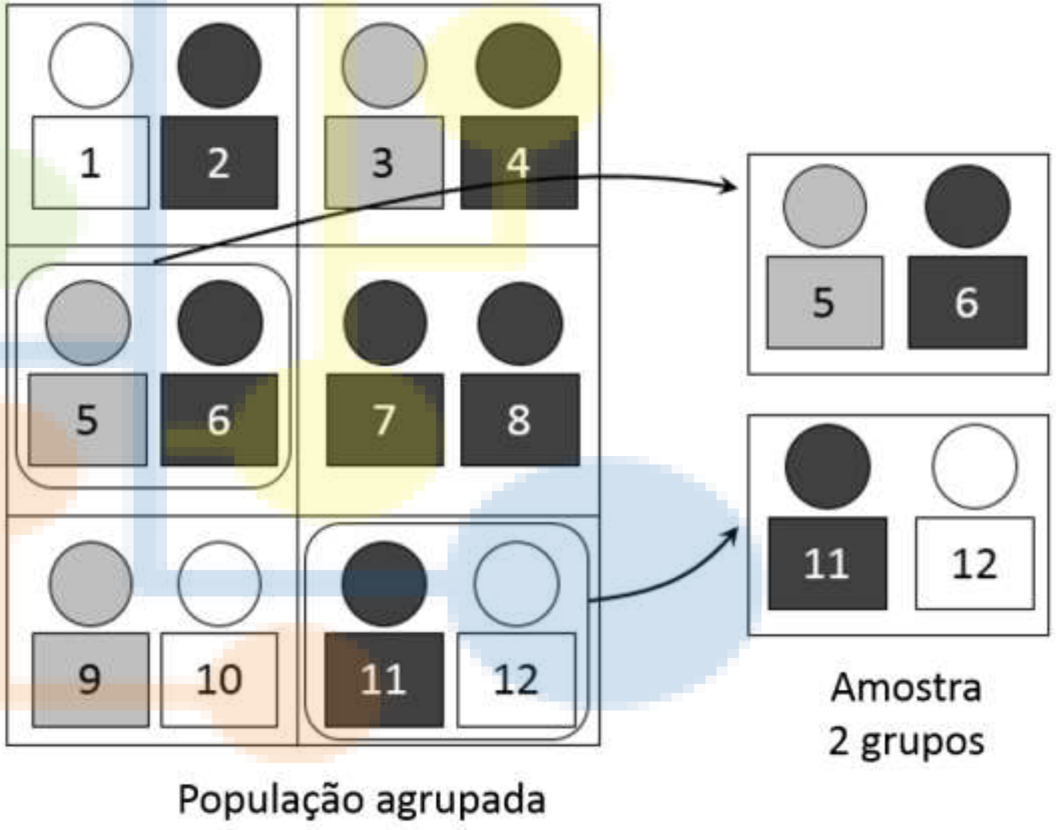
Tipos de Amostragem

Amostragem Estratificada

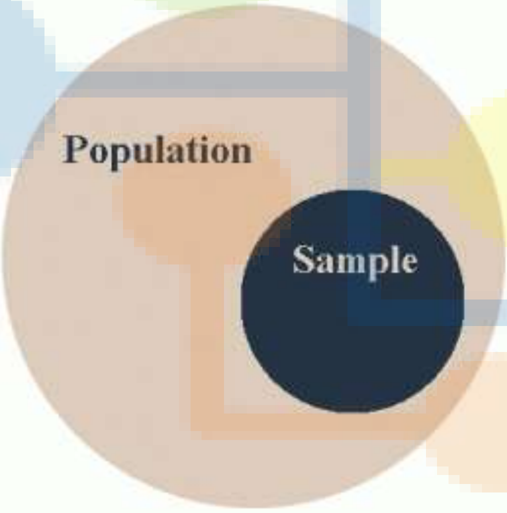


Tipos de Amostragem

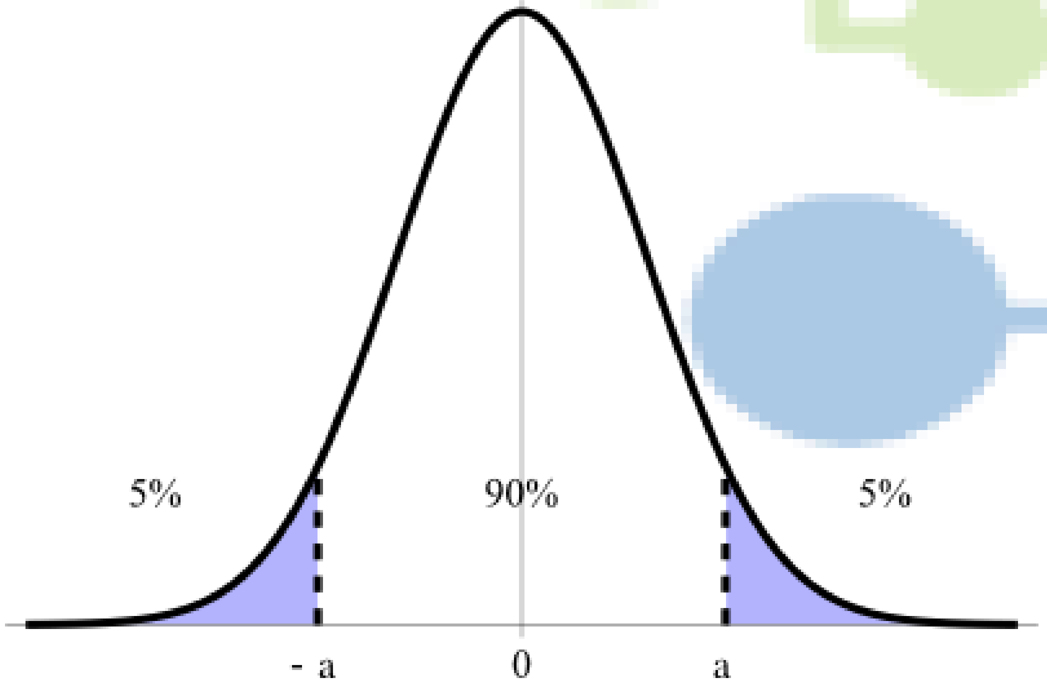
Amostragem Por Conglomerados



Testes de Hipótese



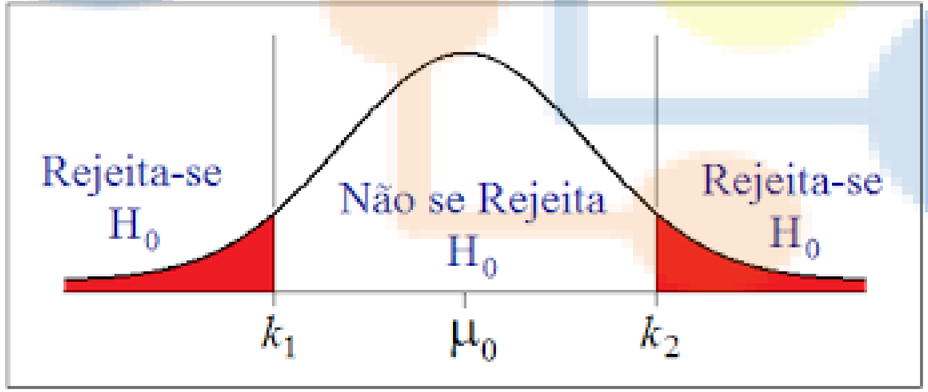
Testes de Hipótese



Uma hipótese estatística é uma suposição sobre um determinado **parâmetro** da população, como média, desvio-padrão, coeficiente de correlação etc. Um teste de hipótese é um procedimento para decisão sobre a veracidade ou falsidade de uma determinada hipótese.

Testes de Hipótese

Um **Teste de Hipótese Estatística** é um procedimento de decisão que nos possibilita decidir entre **H_o** (hipótese nula) ou H_a (hipótese alternativa), com base nas informações contidas na amostra.



Testes de Hipótese

H_o

A hipótese nula afirma que um parâmetro da população (como a média, o desvio padrão, e assim por diante) é igual a um valor hipotético. A hipótese nula é, muitas vezes, uma alegação inicial baseado em análises anteriores ou conhecimentos especializados.

H_a

A hipótese alternativa afirma que um parâmetro da população é menor, maior ou diferente do valor hipotético na hipótese nula. A hipótese alternativa é aquela que você acredita que pode ser verdadeira ou espera provar ser verdadeira.

Testes de Hipótese

Como estamos analisando dados da amostra e não da população, erros podem ocorrer:

Erro Tipo I é a probabilidade de rejeitarmos a hipótese nula quando ela é efetivamente verdadeira.

Erro Tipo II é a probabilidade de rejeitarmos a hipótese alternativa quando ela é efetivamente verdadeira.

Testes de Hipótese

Exemplo:

Um pesquisador tem resultados de exames para uma amostra de alunos que fizeram um curso de formação para um exame nacional. O pesquisador quer saber se os alunos formados obtiveram pontuação acima da média nacional de 78.

Uma hipótese alternativa pode ser usada porque o pesquisador está especificamente levantando a hipótese de que as pontuações para alunos formados são maiores do que a média nacional .

Testes de Hipótese

Exemplo:

Um pesquisador tem resultados de exames para uma amostra de alunos que fizeram um curso de formação para um exame nacional. O pesquisador quer saber se os alunos formados obtiveram pontuação acima da média nacional de 78.

Uma hipótese alternativa pode ser usada porque o pesquisador está especificamente levantando a hipótese de que as pontuações para alunos formados são maiores do que a média nacional . (H_0: $\mu = 78$ e H_a: $\mu > 78$)

Testes de Hipótese

Formular as hipóteses nula e alternativa.

Coletar uma amostra de tamanho n e calcular a média da amostra.

Traçar a média da amostra no eixo x da distribuição da amostra.

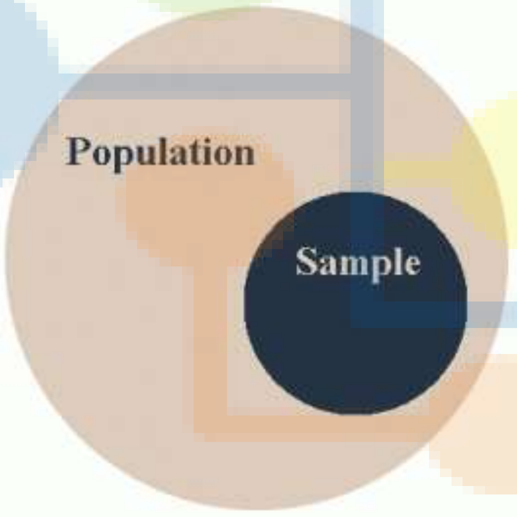
Escolher um nível de significância α com base na gravidade do erro tipo I.

Calcular a estatística, os valores críticos e a região crítica.

Se a média da amostra estiver na região branca do gráfico NÃO rejeitamos a hipótese nula.

Se a média da amostra estiver em uma das caudas nós rejeitamos a hipótese nula.

Teste de Hipótese Unilateral



Teste de Hipótese Unilateral

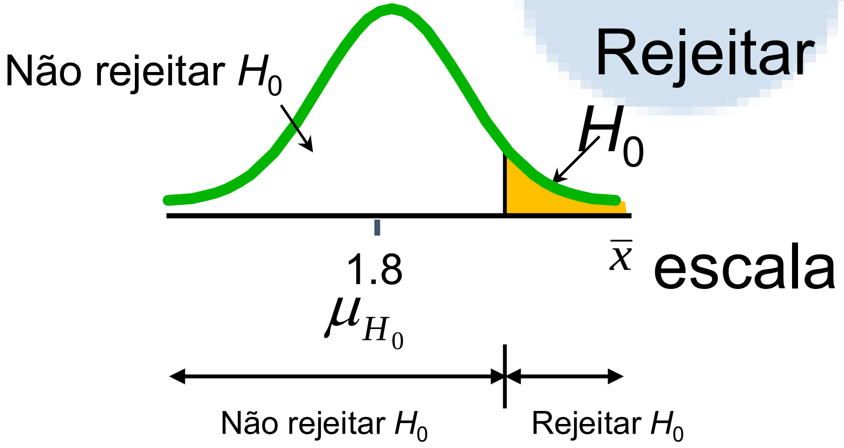
O teste **Unilateral** ou **Unicaudal** é usado quando a hipótese alternativa é expressa como:

< ou **>**

Teste de Hipótese Unilateral

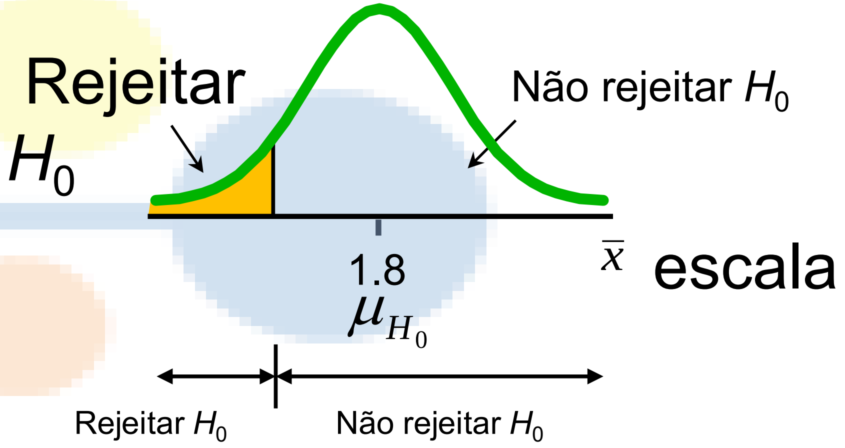
H_0: μ = 1.8
H_A: μ > 1.8

Teste Cauda Superior: nós assumimos que μ = 1.8 a menos que a média da amostra seja maior que the 1.8



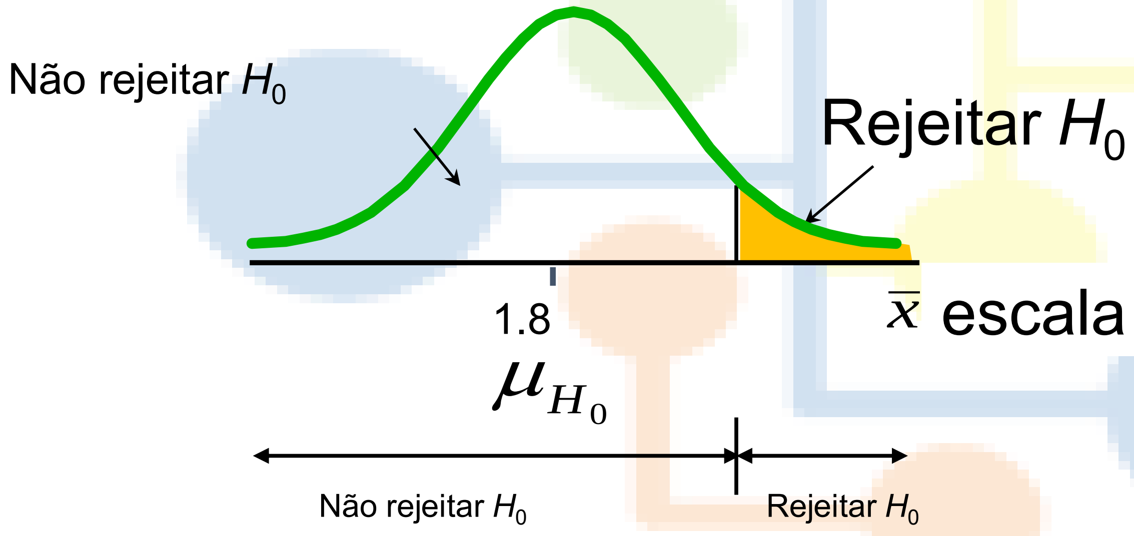
H_0: μ = 1.8
H_A: μ < 1.8

Teste Cauda Inferior: nós assumimos que μ = 1.8 a menos que a média da amostra seja menor que 1.8



Teste de Hipótese Unilateral

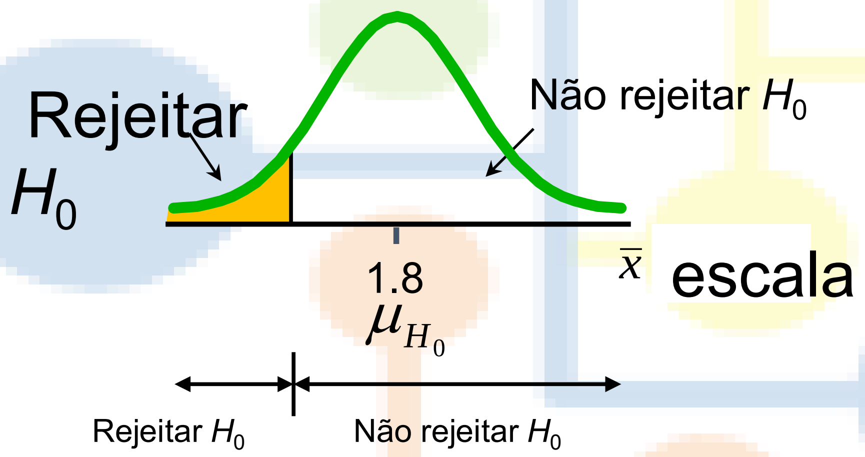
Se a média estiver dentro da região branca do gráfico, não rejeitamos a hipótese nula, caso contrário, a rejeitamos.



Teste Unilateral Direito

Teste de Hipótese Unilateral

Se a média estiver dentro da região branca do gráfico, não rejeitamos a hipótese nula, caso contrário, a rejeitamos.



Teste Unilateral Esquerdo

Teste de Hipótese Unilateral

Teste Unilateral Esquerdo:
(Inferior)

H_0: μ = valor numérico

H_A: μ < valor numérico

Teste Unilateral Direito:
(Superior)

H_0: μ = valor numérico

H_A: μ > valor numérico

Teste de Hipótese Unilateral

Exemplo

Teste de Hipótese Unilateral

Uma escola possui um grupo de alunos (população) considerados obesos. A distribuição de probabilidade do peso dos alunos dessa escola entre 12 e 17 anos é normal com uma média de 80 kgs e desvio padrão de 10 kgs. O diretor da escola propõe uma campanha de tratamento com acompanhamento médico para combater a obesidade. Esse tratamento será composto por dietas, exercícios físicos e mudança de hábito alimentar. O médico afirma que o resultado do tratamento será apresentado em 4 meses. E que os alunos terão seus pesos diminuídos nesse período.

Teste de Hipótese Unilateral

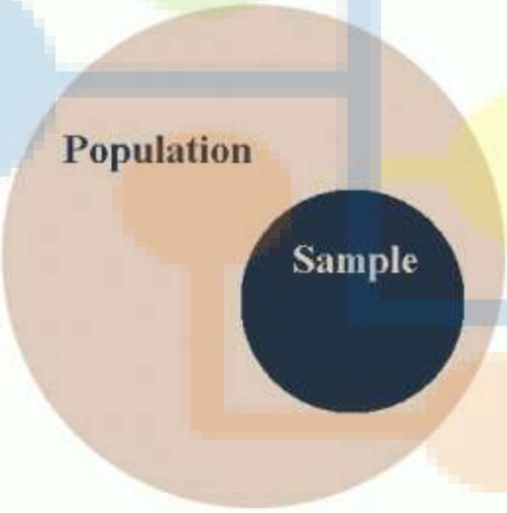
Quais são as hipóteses, nula e alternativa?

$$H_0: \mu = 80$$

$$H_A: \mu < 80$$

Onde: μ = *média dos pesos dos alunos após os 4 meses.*

Teste de Hipótese Bilateral



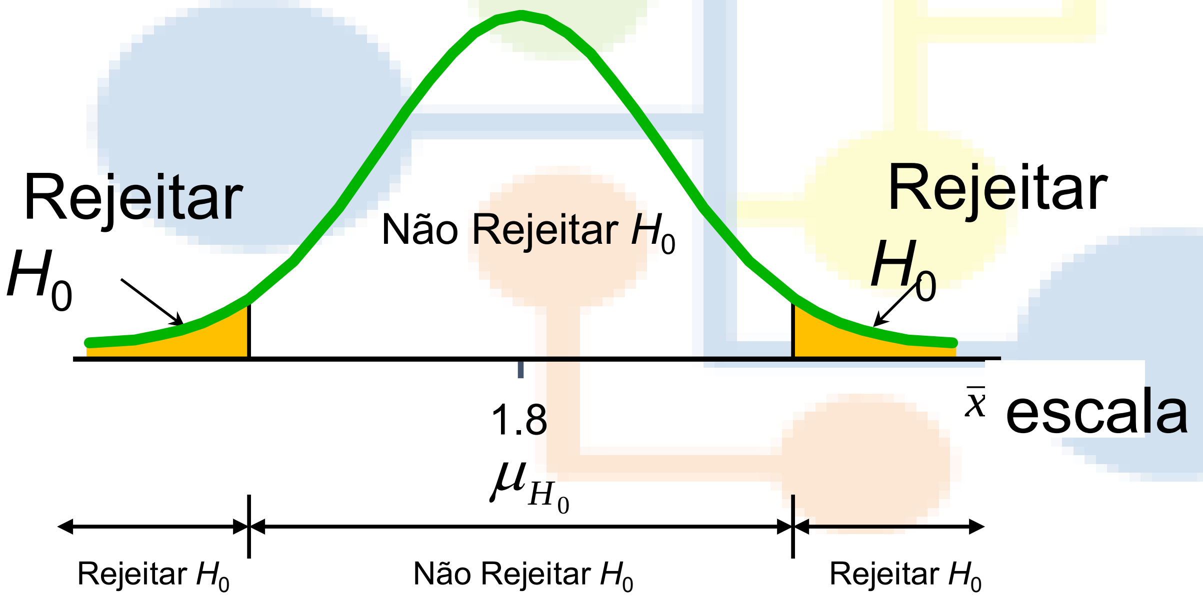
Teste de Hipótese Bilateral

O teste Bilateral é usado sempre que a hipótese alternativa é expressa como **≠ de.**

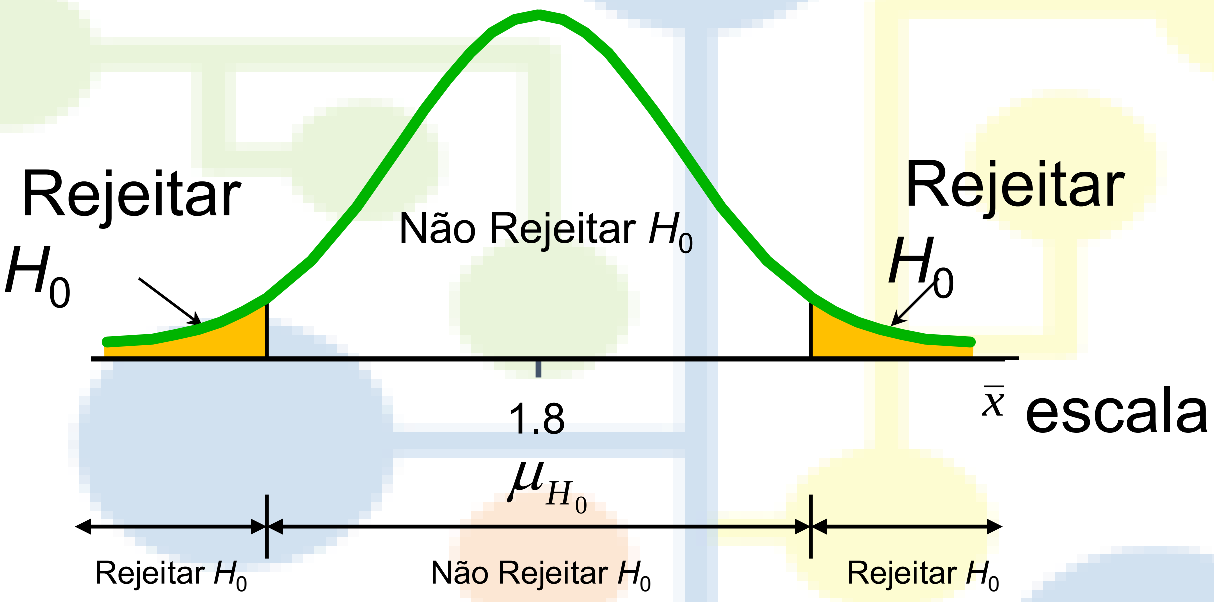
Teste de Hipótese Bilateral

H_0: $\mu = 1.8$
H_A: $\mu \neq 1.8$

Nós assumimos que $\mu = 1.8$ a menos que a média da amostra seja ≠ que 1.8

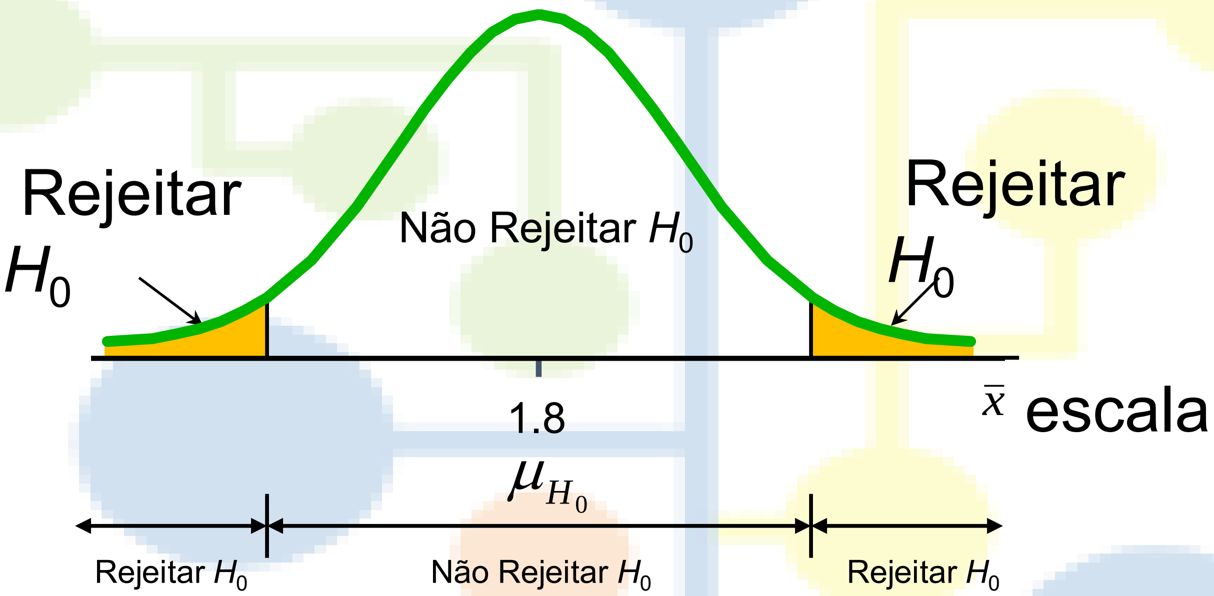


Teste de Hipótese Bilateral



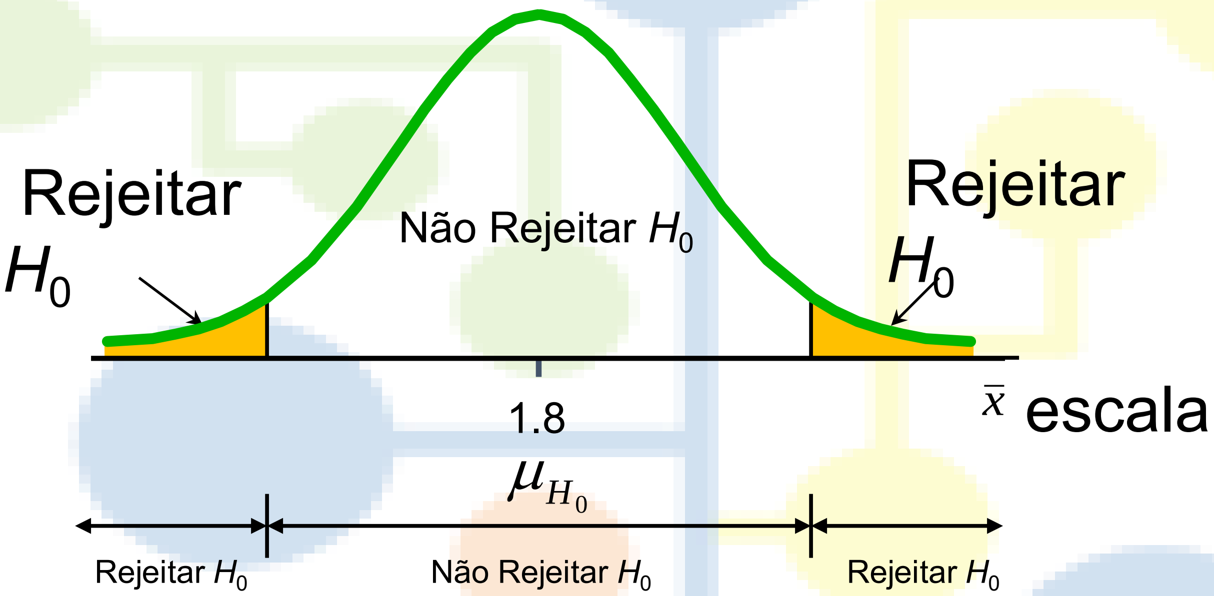
A curva acima representa a distribuição da amostragem da média de utilização de banda larga. Assume-se que a média da população é 1.8 GB, de acordo com a hipótese nula $\boldsymbol{H_0 : \mu = 1.8}$.

Teste de Hipótese Bilateral



Por existirem duas regiões de rejeição no gráfico (regiões em amarelo), este é chamado teste de hipótese **bilateral** ou **bicaudal.**

Teste de Hipótese Bilateral



Como a hipótese nula é expressa como ≠ ela pode ser maior ou menor que, por isso o teste é **bilateral.**

Teste de Hipótese Bilateral

$$H_0: \ \mu \ = \ \text{Valor numérico.}$$

$$H_A: \ \mu \ \neq \ \text{Valor numérico.}$$

Teste de Hipótese Bilateral

Exemplo

Teste de Hipótese Bilateral

Uma fábrica de biscoitos empacota as caixas com peso de 500 gramas. O peso é monitorado periodicamente.

O departamento de qualidade estabeleceu que o peso deve ser mantido em 500 gramas. Qual a condição para que o departamento de qualidade interrompa a produção dos biscoitos?

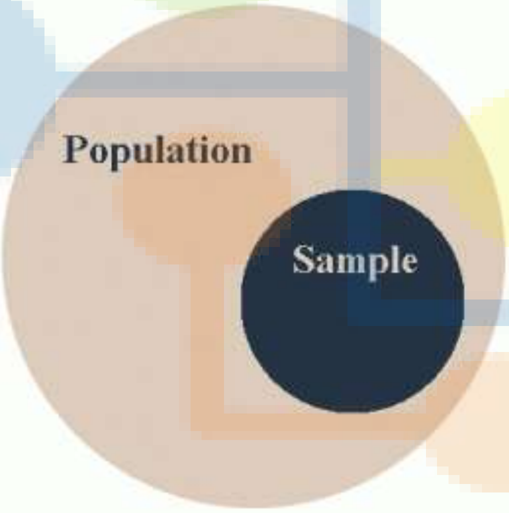
Teste de Hipótese Bilateral

Quais são as hipóteses, nula e alternativa?

$$H_0: \mu = 500$$

$$H_A: \mu \neq 500$$

Erros Tipo I e Tipo II



Erros Tipo I e Tipo II

O propósito do **teste de hipótese** é **verificar** a **validade** de uma **afirmação** sobre um **parâmetro** da **população**, baseado em **amostragem**.

Erros Tipo I e Tipo II

Como estamos tomando **amostra** como base, estamos expostos ao **risco** de conclusões **erradas** sobre a **população**, por conta de **erros de amostragem**.

Erros Tipo I e Tipo II

A **hipótese nula** pode ser verdadeira, caso tenhamos coletado uma amostra que não seja representativa da população.

Ou

talvez, a amostra tenha sido muito pequena.

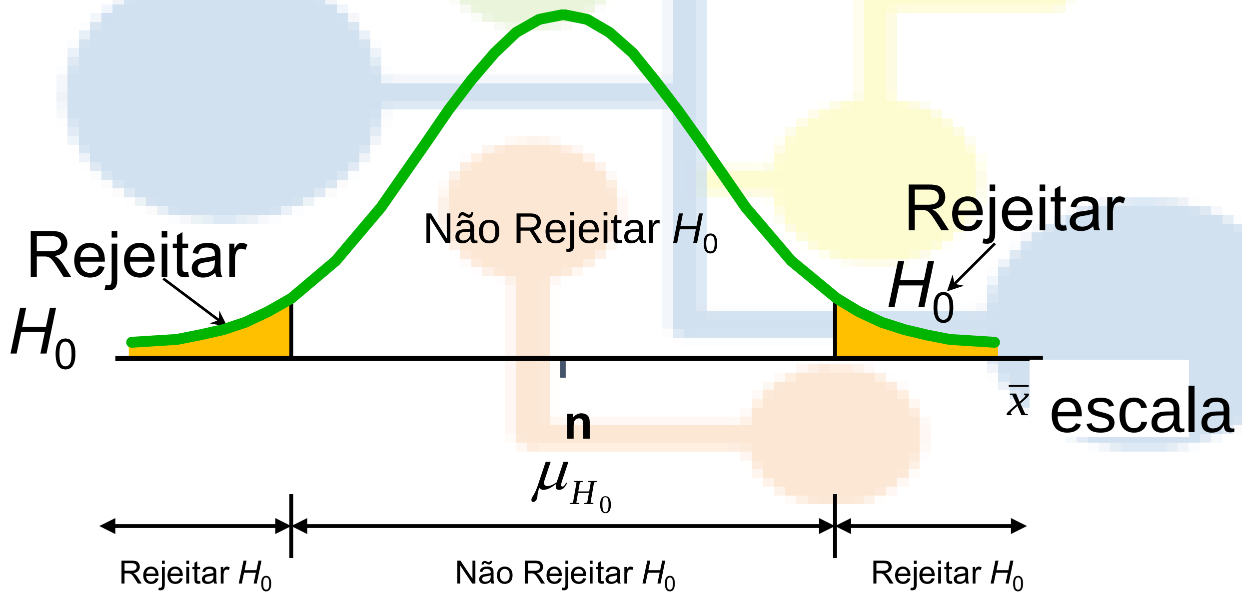
Erros Tipo I e Tipo II

Para testar a H_0, é preciso definir uma regra de decisão com o objetivo de estabelecer uma zona de rejeição da hipótese, ou seja, definir um nível de significância, α , sendo os mais consensuais os alfas 0.10, 0.05 e 0.01.

Grau de Confiança	Nível de Significância
90%	0,10
95%	0,05
99%	0,01

Erros Tipo I e Tipo II

Se o valor do parâmetro da população, defendido pela $\boldsymbol{H_0}$, cair na zona de rejeição, então esse valor é muito pouco provável de ser o valor verdadeiro da população e a $\boldsymbol{H_0}$ será rejeitada em favor da $\boldsymbol{H_A}$.



Erros Tipo I e Tipo II

Pode acontecer, que apesar de rejeitada com base em dados de uma amostra, a $\boldsymbol{H_0}$ de fato seja verdadeira. Nesse caso, estaríamos cometendo um erro de decisão.

Esse erro é chamado de **Erro Tipo I,** cuja probabilidade de ocorrência depende **do alfa escolhido**.

Erros Tipo I e Tipo II

Quando o valor defendido pela $\boldsymbol{H_0}$ cair fora da zona de rejeição, então consideramos que não há evidência para rejeitar $\boldsymbol{H_0}$ em prejuízo da $\boldsymbol{H_A}$. Mas aqui, também podemos estar cometendo um erro se a $\boldsymbol{H_A}$, apesar de descartada pelos dados em mãos, for de fato verdadeira.

Esse erro é chamado **Erro Tipo II.**

Erros Tipo I e Tipo II

Condição		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)

Erros Tipo I e Tipo II

Exemplo

Erros Tipo I e Tipo II

A eficácia de certa vacina após um ano é de 25% (isto é, o efeito imunológico se prolonga por mais de 1 ano em apenas 25% das pessoas que a tomam). Desenvolve-se uma nova vacina, mais cara e deseja-se saber se esta é, de fato, melhor.

Erros Tipo I e Tipo II

A eficácia de certa vacina após um ano é de 25% (isto é, o efeito imunológico se prolonga por mais de 1 ano em apenas 25% das pessoas que a tomam). Desenvolve-se uma nova vacina, mais cara e deseja-se saber se esta é, de fato, melhor.

Que hipóteses devem ser formuladas?
Que erros podemos encontrar?

Erros Tipo I e Tipo II

Hipótese Nula $H_0 : p = 0,25$
Hipótese Alternativa $H_A : p > 0,25$

Erro Tipo I : aprovar a vacina quando, na realidade, ela não tem nenhum efeito superior ao da vacina em uso.

Erro Tipo II : rejeitar a nova vacina quando ela é, de fato, melhor que a vacina em uso.

Erros Tipo I e Tipo II

A probabilidade de se cometer um **Erro Tipo I** depende dos valores dos parâmetros da população e é designada por α (alfa - nível de significância).

Dizemos então que o nível de significância alfa de um teste, é a probabilidade máxima com que desejamos correr o risco de um **Erro Tipo I**.

O valor alfa é tipicamente predeterminado e escolhas comuns são $\alpha = 0.05$ e $\alpha = 0.01$

Erros Tipo I e Tipo II

A probabilidade de se cometer um **Erro Tipo II** é designada por β.

Tenha uma Excelente Jornada de Aprendizagem.

Muito Obrigado por Participar!

Equipe Data Science Academy