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

Big Data Analytics com R e Microsoft Azure Machine Learning

Introdução à Análise Estatística de Dados

Seja Bem-Vindo(a)!

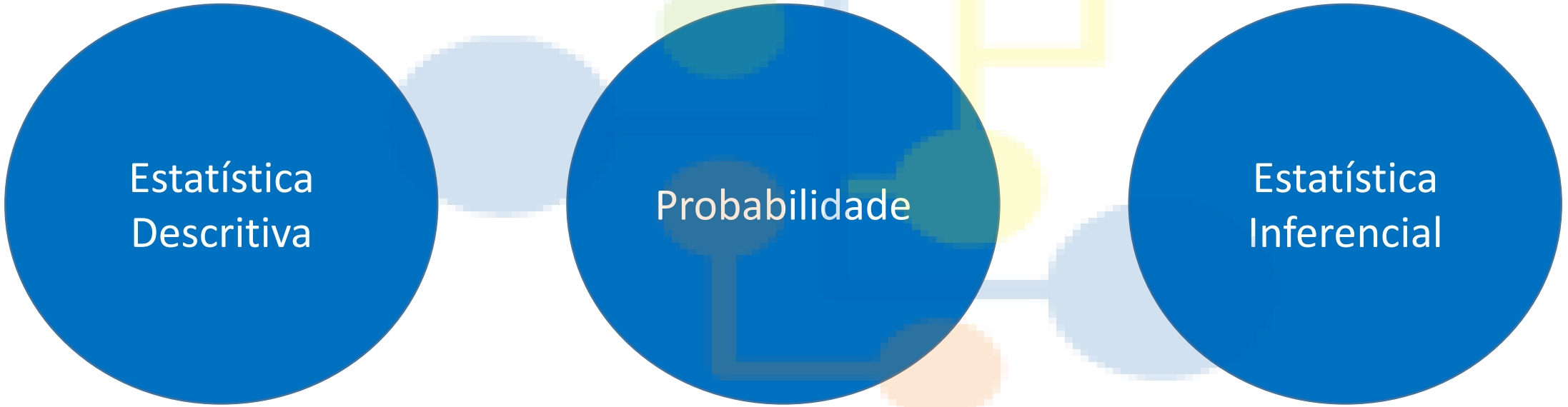
Introdução à Análise Estatística de Dados

Introdução à Análise Estatística de Dados Parte 1

Introdução à Análise Estatística de Dados Parte 2

Introdução à Análise Estatística de Dados Parte 3

Introdução à Análise Estatística de Dados



Estatística Descritiva

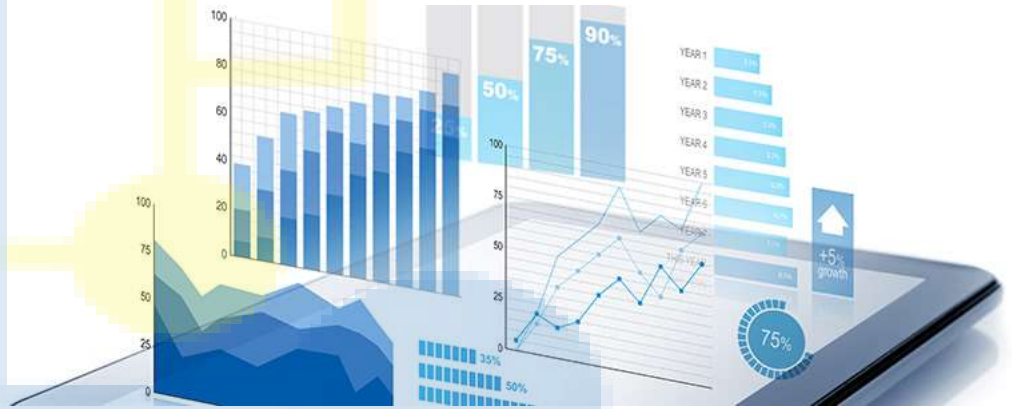
Seja Bem-Vindo(a)!

Estatística Descritiva

A Estatística é um conjunto de técnicas que permite, de forma sistemática, organizar, descrever, analisar e interpretar dados oriundos de estudos ou experimentos, realizados em qualquer área do conhecimento.

Estatística Descritiva

1- Estatística Descritiva
2- Probabilidade
3- Inferência Estatística

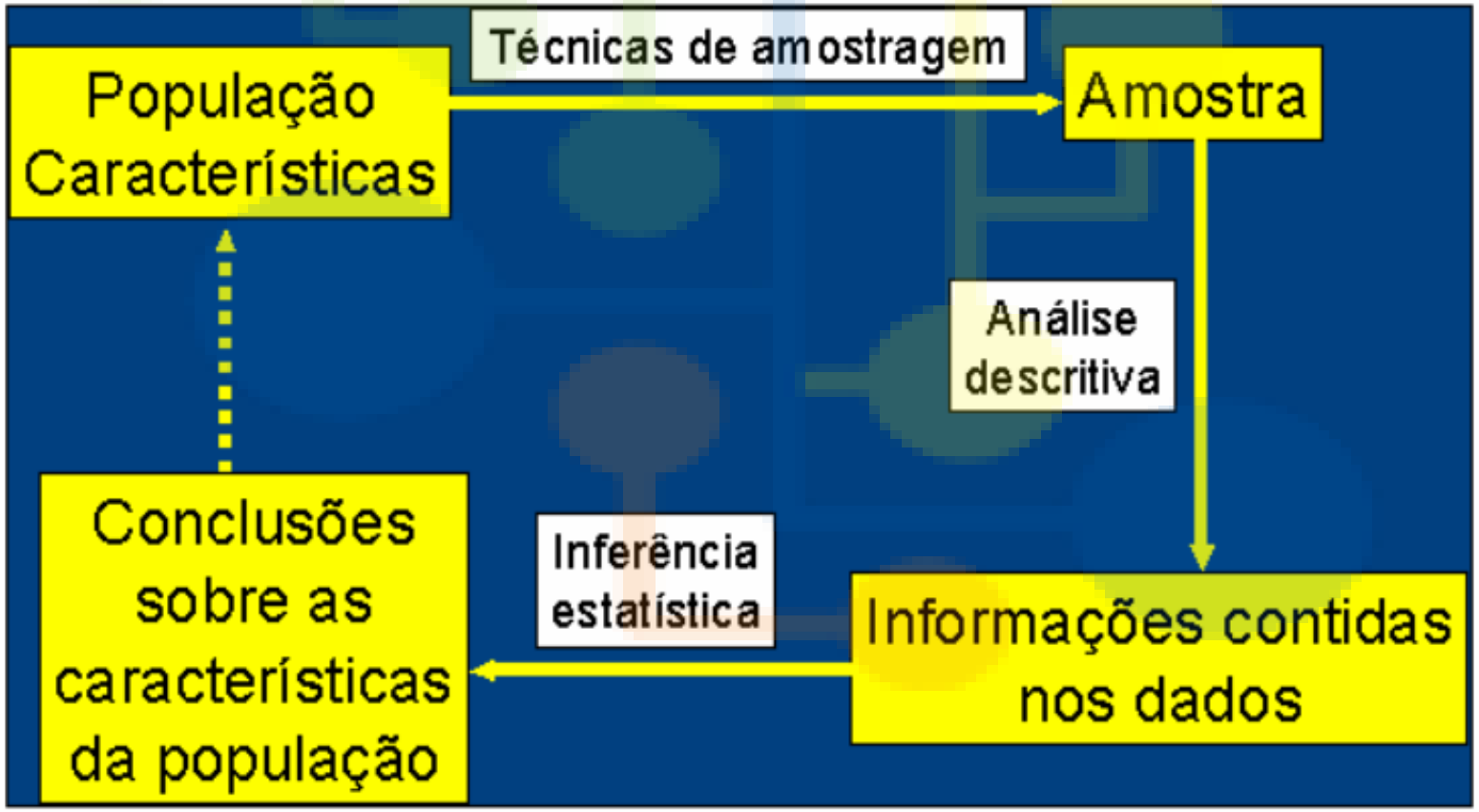


Estatística Descritiva

A Estatística Descritiva é a etapa inicial da análise utilizada para descrever e resumir os dados.

A disponibilidade de uma grande quantidade de dados e de métodos computacionais muito eficientes revigorou está área da estatística.

Estatística Descritiva



Estatística Descritiva

Com a Estatística Descritiva podemos descrever os dados usando 2 tipos principais de medidas:

Medidas de Tendência Central

Média
Mediana
Moda
Valor Máximo e Valor Mínimo
Amplitude

Medidas de Dispersão

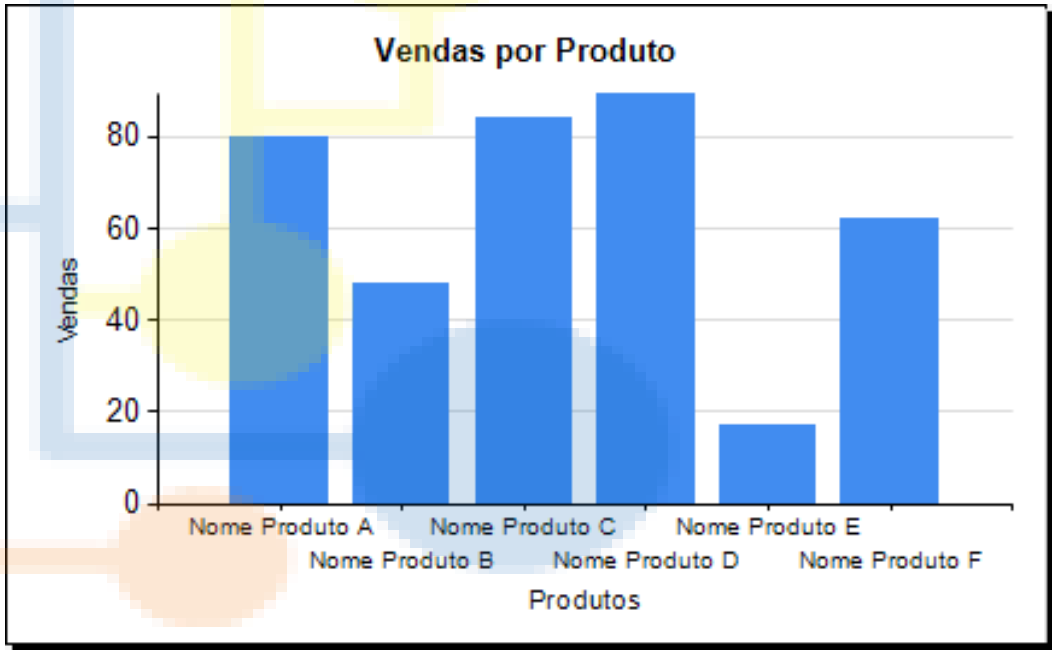
Desvio Padrão
Variância
Coeficiente de Variação

Representação Gráfica dos Dados

Seja Bem-Vindo(a)!

Representação Gráfica dos Dados

	A	B	C	D
1	PRODUTO A		PRODUTO B	
2	R$ 1,50		R$ 10,00	
3	R$ 1,45		R$ 9,95	
4	R$ 1,43		R$ 9,93	
5	R$ 1,52		R$ 10,10	
6	R$ 1,54		R$ 10,75	
7			R$ 10,50	
8				
9				



Representação Gráfica dos Dados

Uma vez que as informações são coletadas dá-se início ao processo de análise dos dados. Um dos passos mais importantes dessa análise é, justamente, a construção e plotagem dos gráficos.

Representação Gráfica dos Dados

A principal função de uma representação gráfica é proporcionar a análise dos dados com maior clareza e fluidez, além de ajudar na solução de problemas e dificuldades à medida que estas aparecem.

Representação Gráfica dos Dados

Um gráfico não precisa, necessariamente, ser elegante nesta fase, o importante é que seja útil.

Introdução à Probabilidade

Seja Bem-Vindo(a)!

Introdução à Probabilidade

Probabilidade é provavelmente um dos tópicos mais interessantes no campo da Estatística, pois permite medir a incerteza.

Introdução à Probabilidade

Mega Sena

Cujo objetivo é acertar 6 números em 60, possui uma probabilidade de vitória (considerando uma única aposta) de 0.00000002 ou aproximadamente 1 em 50 milhões.

Cálculo para eventos possíveis:

$$C_{60,6} = \frac{60!}{6!\ 54!} = 50.063.860 \text{ Possibilidades}$$

Introdução à Probabilidade

Se você comprar um bilhete todos os dias do ano, você poderia levar **136.986** anos para vencer.

Resumindo, a probabilidade é de que você jogue a vida inteira e nunca ganhe!!!

Introdução à Probabilidade

Ser Atingido por um Raio

Segundo os institutos de meteorologia, a probabilidade de ser atingido por um raio é de 1 em 400.000.

Ou seja, é **125 vezes** mais provável que alguém possa ser atingido por um raio, que vencer na loteria.

Introdução à Probabilidade

O mundo atual e enfrenta muitos desafios sobre as incertezas, principalmente no mundo dos negócios.

E a Probabilidade provê uma ferramenta valiosa para quantificar esta incerteza, de forma que os gestores possam tomar melhores decisões.

Introdução à Probabilidade

Probabilidade é um valor numérico que indica a chance, ou probabilidade, de um evento específico ocorrer. Este valor numérico vai estar entre 0 e 1.

Introdução à Probabilidade

Probabilidade é um valor numérico que indica a chance, ou probabilidade, de um evento específico ocorrer. Este valor numérico vai estar entre **0** e **1.**

Se um evento não possui chance de ocorrer, sua probabilidade é **0 (ou 0%)**.

Se temos certeza sobre a ocorrência do evento, sua probabilidade é 1 (ou 100%).

Evento, Experimento e Espaço da Amostra

Seja Bem-Vindo(a)!

Evento, Experimento e Espaço da Amostra

Experimento – é o processo de medir ou observar uma atividade com o propósito de coletar dados.

Exemplo: jogar um dado.

Evento, Experimento e Espaço da Amostra

Espaço da Amostra – todos os possíveis resultados de um experimento.

Exemplo: ao jogar um dado, todos os resultados possíveis são {1, 2, 3, 4, 5, 6}.

Evento, Experimento e Espaço da Amostra

Evento – um ou mais resultados de um experimento.

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

Evento, Experimento e Espaço da Amostra

Evento Simples – um evento com um único resultado na sua forma mais básica, que não pode ser simplificado.

Evento, Experimento e Espaço da Amostra

Experimentos e seus respectivos espaços da amostra.

Experimento	**Espaço da Amostra**
Jogar uma moeda	{cara, coroa}
Responder uma questão de múltipla escolha	{a, b, c, d, e}
Inspecionar um produto	{defeituoso, não defeituoso}
Puxar uma carta de um baralho padrão	{52 cartas de uma baralho padrão}

Probabilidade e Possibilidade São a Mesma Coisa?

Seja Bem-Vindo(a)!

Probabilidade e Possibilidade São a Mesma Coisa?

Não

Probabilidade e Possibilidade São a Mesma Coisa?

Probabilidade é a medida da **possibilidade** de um evento ocorrer.

Em outras palavras, se a chance de chover amanhã é de 40%, há menos possibilidades que chova amanhã, do que não chova.

Probabilidade Clássica

Seja Bem-Vindo(a)!

Probabilidade Clássica

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.

Probabilidade Clássica

Fórmula da Probabilidade Clássica

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

Onde:
s = número de possiveis resultados
n = número resultados possiveis dentro do espaço da amostra

Probabilidade Clássica

Experimento com um Dado:

Um dado possui um espaço de amostra igual a {1, 2, 3, 4, 5, 6}, com 6 possíveis resultados. Qual seria a probabilidade de, ao jogarmos o dado, conseguirmos que o número 5 seja a face em evidência?

Probabilidade Clássica

Resposta:

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

$$P(A) = 1 / 6 = 0.167$$

Ou seja, 16.7% de probabilidade de jogarmos um dado e conseguirmos a face com o número 5.

Probabilidade Clássica

Experimento com um Dado:

Qual a probabilidade de se obter um 3 ou um 4 em uma jogada de um dado equilibrado?

Probabilidade Clássica

Resposta:

Como temos 2 possibilidades, “3 ou 4”.

$$P(A) = \frac{s}{n}$$

2/6 = 0,33

33,33%

Probabilidade Empírica

Seja Bem-Vindo(a)!

Probabilidade Empírica

Quando sabemos os possíveis resultados de um evento, utilizamos a **Probabilidade Clássica.**

Probabilidade Empírica

E quando não sabemos quais os possíveis resultados?

Probabilidade Empírica

Probabilidade Empírica

Para calcularmos a probabilidade empírica, usamos a fórmula:

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

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

Probabilidade Empírica

Experimento da Loja de Livros:

Qual a probabilidade de que uma pessoa que entre na loja faça uma compra?

Probabilidade Empírica

A probabilidade clássica não poderia nos ajudar aqui, pois não temos informação sobre porque as pessoas fazem uma compra e nem quando elas fazem uma compra.

Probabilidade Empírica

Usamos então a **probabilidade empírica**, para contar quantas pessoas que entram na loja finalizam uma compra.

Probabilidade Empírica

Resposta:

Supondo que 100 pessoas entraram na loja e que 15 fizeram uma compra, a probabilidade empírica seria dada pela seguinte fórmula:

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

15/100 = 0.15 = 15%

Probabilidade Subjetiva

Seja Bem-Vindo(a)!

Probabilidade Subjetiva

Usamos **Probabilidade Subjetiva**, quando:

- Probabilidades clássicas ou empíricas não podem ser usadas.
- Dados ou experimentos não estão disponíveis para calcular a probabilidade.
- Nestes casos, confiamos na experiência ou julgamento para estimar as probabilidades.

Probabilidade Subjetiva

Diretor de Marketing:

Um experiente Diretor de Marketing estima que há **50%** de probabilidade de que o maior concorrente da empresa reduza seus preços no mês seguinte.

Regras Básicas da Probabilidade

Seja Bem-Vindo(a)!

5 Regras Básicas que Regem a Teoria da Probabilidade

Regras Básicas da Probabilidade

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

2ª Se P(A) = 0, então podemos garantir que o evento A **NÃO ocorrerá**.

3ª A probabilidade de qualquer evento sempre será entre 0 e 1. Probabilidades nunca podem ser negativas ou maior que 1.

Regras Básicas da Probabilidade

4ª A soma de todas as probabilidades para um evento simples, em um espaço de amostra, será igual a 1 .

5ª 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:

$$P(A) = 1 – P(A'), \text{ onde } P(A') \text{ é o complemento do evento A.}$$

Regras Básicas da Probabilidade Para Mais de Um Evento

Seja Bem-Vindo(a)!

Regras Básicas da Probabilidade Para Mais de Um Evento

No mundo dos negócios, os eventos raramente são simples e frequentemente envolvem dois ou mais eventos.

Regras Básicas da Probabilidade Para Mais de Um Evento

Por exemplo, o gerente de um banco, pode estar interessado em saber a probabilidade de um cliente com histórico de crédito ruim não pagar um empréstimo de cheque especial.

Neste caso, temos **2** eventos:

Evento A – cliente não paga o cheque especial.
Evento B – cliente tem um histórico de crédito ruim.

Regras Básicas da Probabilidade Para Mais de Um Evento

Intersecção de Eventos

União de Eventos

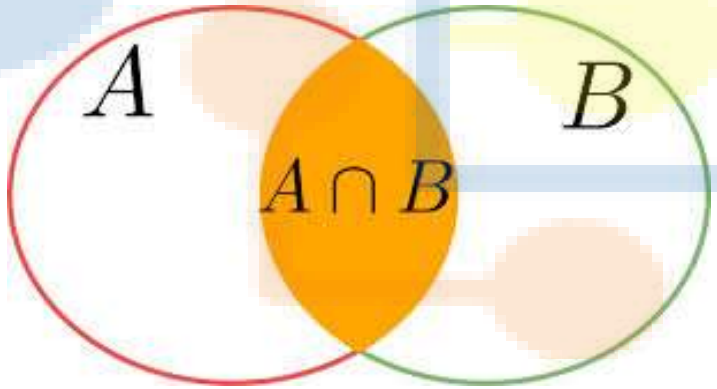
Adição de Eventos

Regras Básicas da Probabilidade Para Mais de Um Evento

Intersecção de Eventos

Regras Básicas da Probabilidade Para Mais de Um Evento

A intersecção de eventos A e B, representa o número de vezes em que os eventos **A e B** ocorrem ao mesmo tempo.



Regras Básicas da Probabilidade Para Mais de Um Evento

Vamos usar uma **tabela de contingência** para exemplificar melhor. A tabela a seguir mostra o número de alunos admitidos em cursos de graduação em Engenharia e Medicina em 3 cidades brasileiras:

Cidade	Engenharia	Medicina	Total
Rio de Janeiro	1500	2300	3800
São Paulo	5600	7500	13100
Porto Alegre	980	1400	2380
Total	**8080**	**11200**	**19280**

Regras Básicas da Probabilidade Para Mais de Um Evento

Vamos definir os eventos sob análise!

Regras Básicas da Probabilidade Para Mais de Um Evento

Evento A – o estudante é da cidade de **São Paulo**.

Cidade	Engenharia	Medicina	Total
Rio de Janeiro	1500	2300	3800
São Paulo	5600	7500	13100
Porto Alegre	980	1400	2380
Total	**8080**	**11200**	**19280**

Regras Básicas da Probabilidade Para Mais de Um Evento

Evento B – o estudante foi admitido em curso de **Medicina.**

Cidade	Engenharia	Medicina	Total
Rio de Janeiro	1500	2300	3800
São Paulo	5600	7500	13100
Porto Alegre	980	1400	2380
Total	**8080**	**11200**	**19280**

Regras Básicas da Probabilidade Para Mais de Um Evento

Vamos Calcular a Probabilidade do **Evento A** ocorrer:

$P(A) = 13100 / 19280 = 0.68 \times 100 = 68\%$

Vamos Calcular a Probabilidade do **Evento B** ocorrer:

$P(B) = 11200 / 19280 = 0.58 \times 100 = 58\%$

Cidade	Engenharia	Medicina	Total
Rio de Janeiro	1500	2300	3800
São Paulo	5600	7500	13100
Porto Alegre	980	1400	2380
Total	**8080**	**11200**	**19280**

Regras Básicas da Probabilidade Para Mais de Um Evento

Vamos Calcular a Probabilidade de um Estudante de **São Paulo**, ser admitido em um curso de **Medicina**.

Para isso, calculamos a intersecção dos eventos A e B.

$$P(A \text{ e } B) = 7500 / 19280 = 0.39 \times 100 = 39\%$$

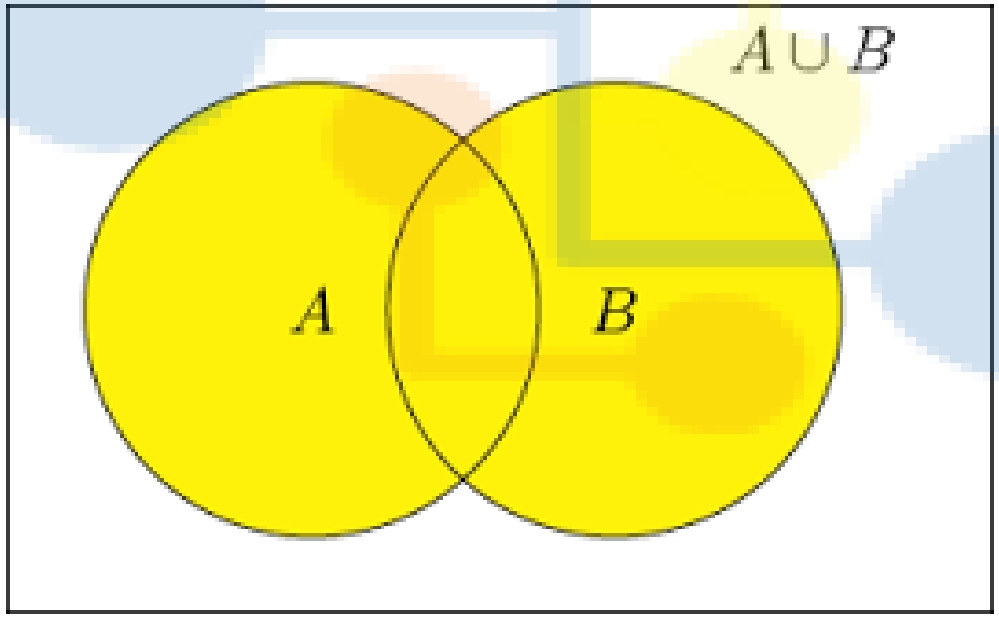
Cidade	Engenharia	Medicina	Total
Rio de Janeiro	1500	2300	3800
São Paulo	5600	**7500**	13100
Porto Alegre	980	1400	2380
Total	**8080**	**11200**	**19280**

Regras Básicas da Probabilidade Para Mais de Um Evento

União de Eventos

Regras Básicas da Probabilidade Para Mais de Um Evento

A união dos eventos A e B representa o número de vezes em que o evento **A** ou o evento **B** ocorrem juntos.



Regras Básicas da Probabilidade Para Mais de Um Evento

Vamos usar uma **tabela de contingência** para exemplificar melhor. A tabela a seguir mostra o número de alunos admitidos em cursos de graduação em Engenharia e Medicina em 3 cidades brasileiras:

Cidade	Engenharia	Medicina	Total
Rio de Janeiro	1500	2300	3800
São Paulo	5600	7500	13100
Porto Alegre	980	1400	2380
Total	**8080**	**11200**	**19280**

Regras Básicas da Probabilidade Para Mais de Um Evento

Vamos definir os eventos sob análise!

Regras Básicas da Probabilidade Para Mais de Um Evento

Evento A – estudante do Rio de Janeiro admitido em curso de Engenharia ou Medicina.

Cidade	Engenharia	Medicina	Total
Rio de Janeiro	1500	2300	3800
São Paulo	5600	7500	13100
Porto Alegre	980	1400	2380
Total	**8080**	**11200**	**19280**

Regras Básicas da Probabilidade Para Mais de Um Evento

Evento B – estudante de qualquer cidade admitido em Engenharia.

Cidade	Engenharia	Medicina	Total
Rio de Janeiro	1500	2300	3800
São Paulo	5600	7500	13100
Porto Alegre	980	1400	2380
Total	**8080**	**11200**	**19280**

Regras Básicas da Probabilidade Para Mais de Um Evento

Como estamos considerando uma **união** dos eventos, tanto um como outro pode ocorrer.

Regras Básicas da Probabilidade Para Mais de Um Evento

Evento A - estudante do Rio de Janeiro admitido em curso de Engenharia ou Medicina.

Evento A = 1500 + 2300 = 3800

Cidade	Engenharia	Medicina	Total
Rio de Janeiro	1500	2300	3800
São Paulo	5600	7500	13100
Porto Alegre	980	1400	2380
Total	**8080**	**11200**	**19280**

Regras Básicas da Probabilidade Para Mais de Um Evento

Evento B - estudante de qualquer cidade admitido em Engenharia.

Evento B = 1500 + 5600 + 980 = 8080

Cidade	Engenharia	Medicina	Total
Rio de Janeiro	1500	2300	3800
São Paulo	5600	7500	13100
Porto Alegre	980	1400	2380
Total	**8080**	**11200**	**19280**

Regras Básicas da Probabilidade Para Mais de Um Evento

Evento A = 1500 + 2300 = 3800

Evento B = 1500 + 5600 + 980 = 8080

A soma dos 2 eventos é 3800 + 8080 = **11880**

Regras Básicas da Probabilidade Para Mais de Um Evento

A probabilidade de A ou B ocorrer, é:

P(A ou B) = 11880 / 19280 = 0.62 x 100 = 62%

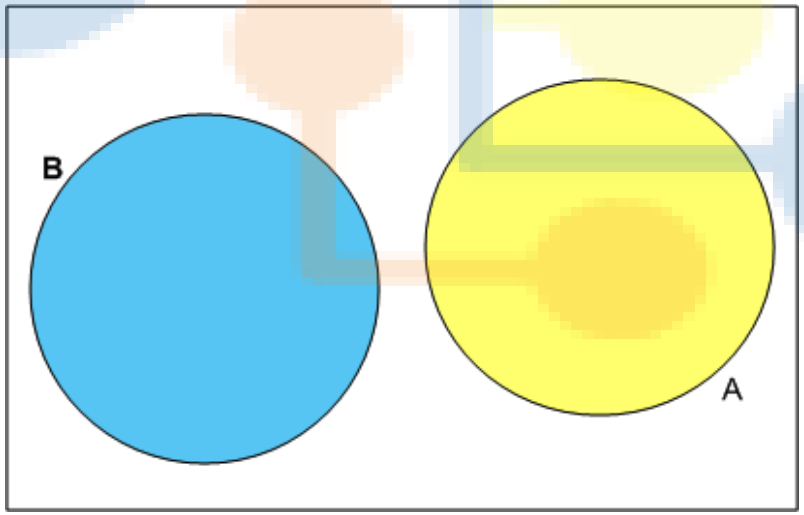
Cidade	Engenharia	Medicina	Total
Rio de Janeiro	1500	2300	3800
São Paulo	5600	7500	13100
Porto Alegre	980	1400	2380
Total	**8080**	**11200**	**19280**

Regras Básicas da Probabilidade Para Mais de Um Evento

Adição de Eventos

Regras Básicas da Probabilidade Para Mais de Um Evento

A **Regra de Adição** em probabilidade é usada para calcular a probabilidade de **união de eventos**, ou seja, a probabilidade do **Evento A mais Evento B** ocorrerem.



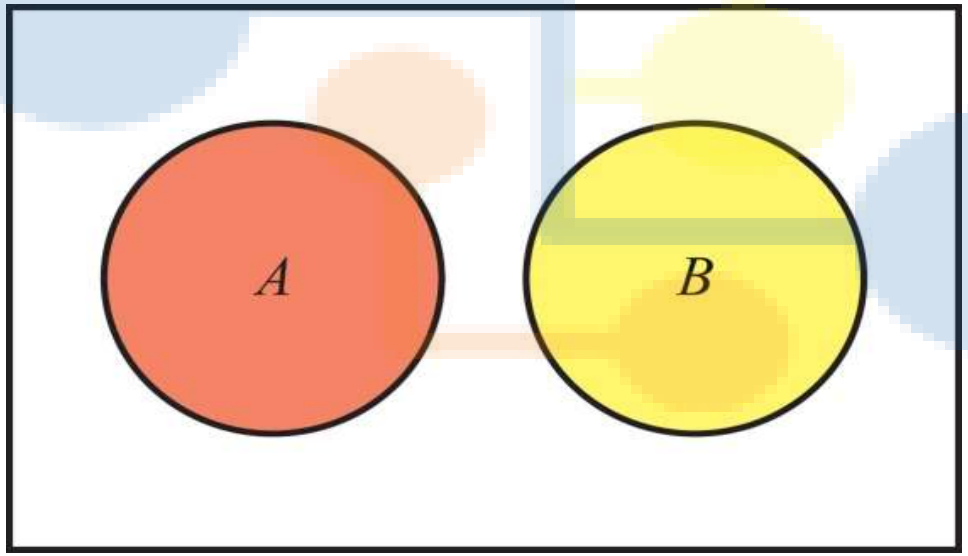
Regras Básicas da Probabilidade Para Mais de Um Evento

Antes, precisamos entender dois conceitos muito importantes:

Eventos Mutuamente Exclusivos
Eventos Não Mutuamente Exclusivos

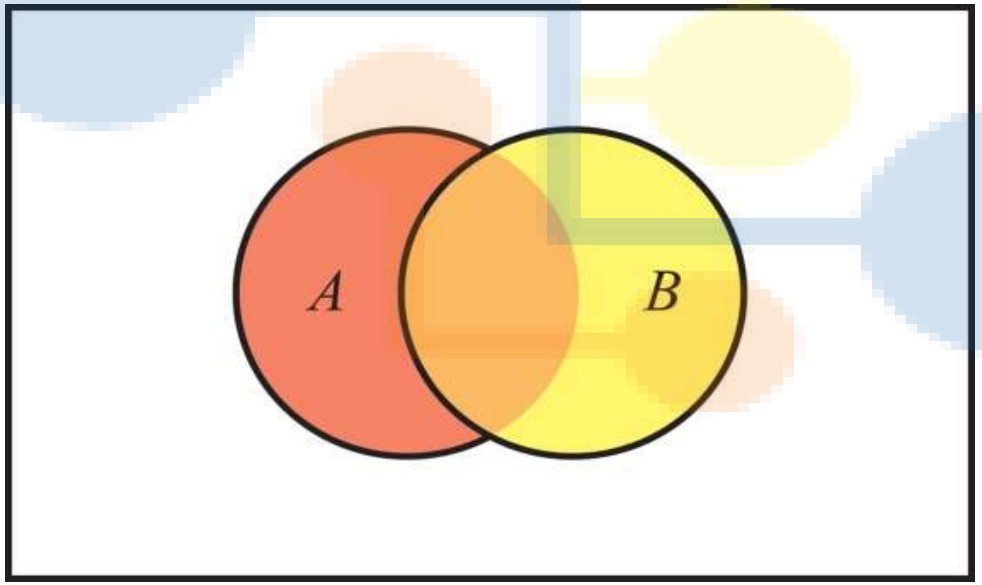
Regras Básicas da Probabilidade Para Mais de Um Evento

Eventos Mutuamente Exclusivos - são aqueles que **não** podem ocorrer ao mesmo tempo durante um experimento.



Regras Básicas da Probabilidade Para Mais de Um Evento

Eventos Não Mutuamente Exclusivos - são aqueles que **podem** ocorrer ao mesmo tempo durante um experimento.



Regras Básicas da Probabilidade Para Mais de Um Evento

A **Regra da Adição** depende se 2 eventos são ou não mutuamente exclusivos.

Nota final no vestibular	Homens	Mulheres	Total
95	60	30	90
90	40	80	120
85	0	40	40
Total	**100**	**150**	**250**

Regras Básicas da Probabilidade Para Mais de Um Evento

Vamos definir os eventos deste experimento:

Regras Básicas da Probabilidade Para Mais de Um Evento

Evento A – estudante com nota final igual a 90.

Evento B – estudante com nota final igual a 85.

Nota final no vestibular	Homens	Mulheres	Total
95	60	30	90
90	40	80	120
85	0	40	40
Total	**100**	**150**	**250**

Regras Básicas da Probabilidade Para Mais de Um Evento

Neste caso, os eventos são **mutuamente exclusivos**, ou seja, um estudante não pode obter notas 90 e 85 no mesmo exame.

Para a regra da adição, usamos a fórmula:

$$P(A \text{ ou } B) = P(A) + P(B)$$

Regras Básicas da Probabilidade Para Mais de Um Evento

Evento A → P(A) = 120 / 250 = 0.48 x 100 = 48%

Evento B → P(B) = 40 / 250 = 0.16 x 100 = 16%

Nota final no vestibular	Homens	Mulheres	Total
95	60	30	90
90	40	80	120
85	0	40	40
Total	**100**	**150**	**250**

Regras Básicas da Probabilidade Para Mais de Um Evento

Evento A → P(**A**) = 120 / 250 = 0.48 x 100 = 48%

Evento B → P(**B**) = 40 / 250 = 0.16 x 100 = 16%

P(A ou B) = P(A) + P(B) = 0.48 + 0.16 = 0.64 x 100 = 64%

64 % é a probabilidade de um estudante ter a nota final igual a 85 ou 90.

Regras Básicas da Probabilidade Para Mais de Um Evento

Mas e se os 2 eventos não forem mutuamente exclusivos?

Regras Básicas da Probabilidade Para Mais de Um Evento

Vamos definir nossos eventos deste experimento de eventos **não** mutuamente exclusivos:

Evento A – estudante com nota final igual a 90.

Evento B – estudante é do sexo feminino.

Regras Básicas da Probabilidade Para Mais de Um Evento

Em casos em que os eventos **não** são mutuamente exclusivos, eles **podem** ocorrer ao mesmo tempo.

Calculamos a Probabilidade da seguinte forma:

$$P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ e } B)$$

Regras Básicas da Probabilidade Para Mais de Um Evento

Evento A – estudante com nota final igual a 90.

Evento B – estudante é do sexo feminino.

Nota final no vestibular	Homens	Mulheres	Total
95	60	30	90
90	40	80	120
85	0	40	40
Total	**100**	**150**	**250**

Regras Básicas da Probabilidade Para Mais de Um Evento

Evento A → P(**A**) = 120 / 250 = 0.48

Evento B → P(**B**) = 150/ 250 = 0.60

P(A e B) = 80 / 250 = 0.32

Nota final no vestibular	Homens	Mulheres	Total
95	60	30	90
90	40	80	120
85	0	40	40
Total	**100**	**150**	**250**

Regras Básicas da Probabilidade Para Mais de Um Evento

Evento A → P(**A**) = 120 / 250 = 0.48

Evento B → P(**B**) = 150/ 250 = 0.60

P(A e B) = 80 / 250 = 0.32

$$P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ e } B)$$

$$P(A \text{ ou } B) = 0.48 + 0.60 - 0.32 = 0.76$$

76% de probabilidade de uma estudante do sexo feminino obter 90 como nota final.

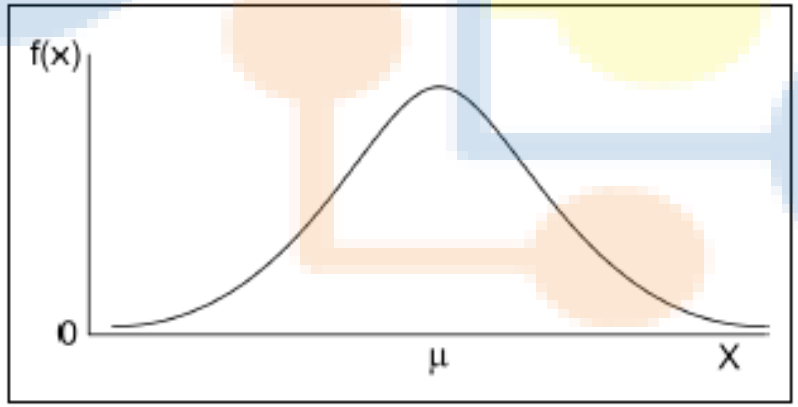
Big Data Analytics com R e Microsoft Azure Machine Learning

Tipos de Distribuição de Probabilidade

Seja Bem-Vindo(a)!

Tipos de Distribuição de Probabilidade

Em estatística, uma **Distribuição de Probabilidade** descreve a **chance** que uma variável pode assumir ao longo de um espaço de valores.



Tipos de Distribuição de Probabilidade

A **Distribuição de Probabilidade** tem por objetivo associar uma probabilidade a cada resultado numérico de um experimento.

Ela é uma função cujo domínio são os valores da variável e cuja imagem são as probabilidades de a variável assumir cada valor do domínio. O conjunto imagem deste tipo de função está sempre restrito ao intervalo entre 0 e 1.

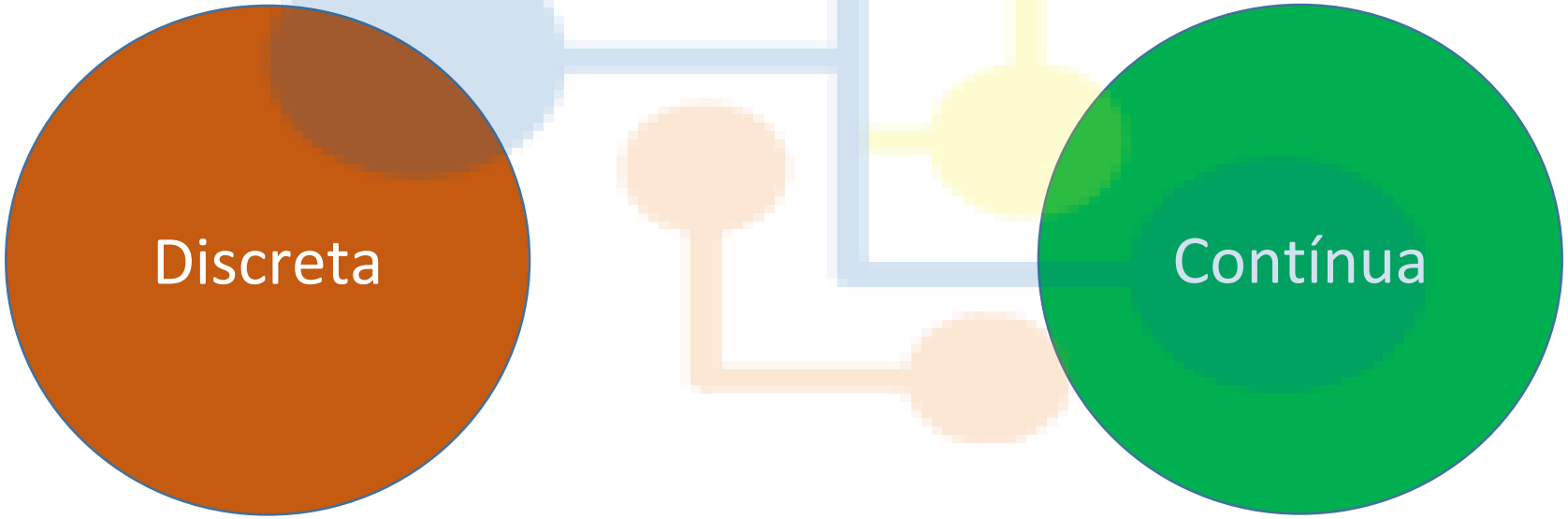
Tipos de Distribuição de Probabilidade

A soma de todos os valores de uma **Distribuição de Probabilidades** deve ser igual a **1.**

A **Probabilidade** de **Ocorrência** de um evento deve ser maior do que **0** e menor do que **1.**

Tipos de Distribuição de Probabilidade

Uma distribuição de probabilidade pode ser:



Tipos de Distribuição de Probabilidade

A distribuição de probabilidade **Discreta**:

Descreve quantidades aleatórias de dados que podem assumir valores **finitos.**

Tipos de Distribuição de Probabilidade

Discreta

Os principais tipos de distribuição de probabilidade para variáveis discretas são:

Binomial
Poisson
Hipergeométrica
Bernoulli

Tipos de Distribuição de Probabilidade

Contínua

A distribuição de probabilidade **Contínua**:

Descreve quantidades aleatórias de dados que podem assumir valores **infinitos.**

Tipos de Distribuição de Probabilidade

Contínua

Os principais tipos de distribuição de probabilidade para variáveis contínuas são:

Uniforme
Exponencial
Gama
Chi-Quadrado

Tipos de Distribuição de Probabilidade

Distribuição Normal

Uma variável randômica contínua que segue uma **Distribuição de Probabilidade Normal** tem uma série de características distintas.

Variáveis Aleatórias Discretas e Contínuas

Seja Bem-Vindo(a)!

Variáveis Aleatórias Discretas e Contínuas

- Registrar o número de clientes que fazem contato telefônico com a central de suporte de um banco.
- Randomicamente selecionar 6 clientes que entram em uma loja de celulares e contar quantos assinam um plano pós-pago.
- Solicitar que cada cliente que deixa um hotel, avalie seu grau de satisfação com o serviço prestado.

Variáveis Aleatórias Discretas e Contínuas

Cada um destes experimentos, vai gerar dados com variáveis aleatórias discretas.

Variáveis Aleatórias Discretas e Contínuas

Variáveis Discretas são, portanto, números inteiros, gerados a partir de resultados de **experimentos**.

Variáveis Aleatórias Discretas e Contínuas

Mas no **mundo dos negócios** normalmente nos deparamos com problemas que requerem medição e cujos valores podem assumir qualquer número em um intervalo.

Variáveis Aleatórias Discretas e Contínuas

- Tempo de duração de voo entre Natal e Maceió.
- Tempo gasto por um cliente ao telefone, com uma companhia de TV a cabo.
- Peso das caixas de biscoito em uma fábrica de alimentos.

Variáveis Aleatórias Discretas e Contínuas

Dados Contínuos normalmente **são medidos** e **não contados**, como no caso dos valores **discretos**.

Variáveis Aleatórias Discretas e Contínuas

Variáveis Contínuas são, portanto, qualquer valor no conjunto de números reais, ou um subconjunto deles.

Variáveis Aleatórias Discretas e Contínuas

Mas tenha em mente, que se optarmos por contar os voos que chegam atrasados ao seu destino, podemos usar **Distribuição Discreta**.

Variáveis Aleatórias Discretas e Contínuas

Por outro lado, se estamos medindo o tempo de voo, estamos na verdade medindo um intervalo de possibilidades (o voo pode durar entre 1 e 2 horas por exemplo, sendo qualquer valor neste intervalo).

Como Construir Uma Distribuição de Probabilidade

Seja Bem-Vindo(a)!

Como Construir Uma Distribuição de Probabilidade

Vamos construir uma Distribuição de Probabilidade para Variável Aleatória Discreta.

Como Construir Uma Distribuição de Probabilidade

Definindo o Problema de Negócio.

O gerente de um restaurante precisa saber qual a probabilidade de grupos de pessoas visitarem o restaurante a fim de organizar melhor as mesas e oferecer uma experiências mais agradável aos clientes.

Como Construir Uma Distribuição de Probabilidade

O experimento foi realizado com 50 observações e os grupos com 2 a 6 pessoas. Um atendente registrou a frequência (ou seja, contou) a quantidade de grupos que chegaram ao restaurante.

Como Construir Uma Distribuição de Probabilidade

Distribuição de Probabilidade

Grupo (x)	Frequência	Frequência Relativa	Probabilidade P(x)
Grupo com 2 pessoas	17	17/50 = 0.34	0.34
Grupo com 3 pessoas	6	6/50 = 0.12	0.12
Grupo com 4 pessoas	16	16/50 = 0.32	0.32
Grupo com 5 pessoas	4	4/50 = 0.08	0.08
Grupo com 6 pessoas	7	7/50 = 0.14	0.14
Total	**50**	**1.00**	**1.0**

Distribuições Contínuas

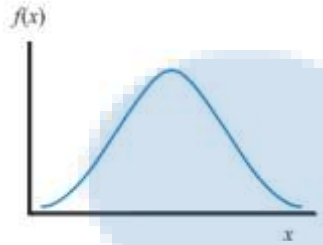
Seja Bem-Vindo(a)!

Distribuições Contínuas

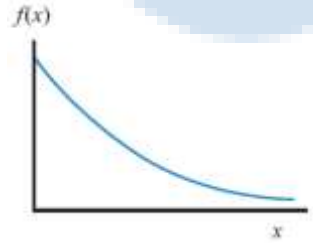
Quando transformadas em **gráficos**, as **Distribuições de Probabilidade Contínua** podem assumir uma variedade de formatos, dependendo dos valores dos dados.

Distribuições Contínuas

Os 3 formatos mais comuns são:



Distribuição Normal

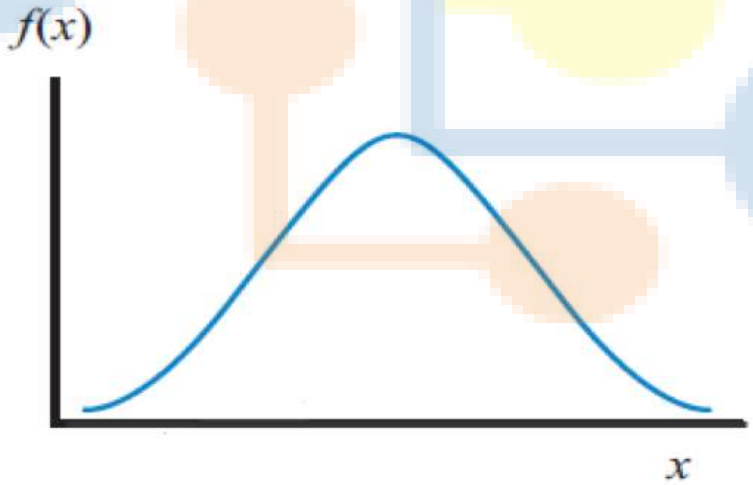


Distribuição Exponencial

Distribuição Uniforme

Distribuições Contínuas

Distribuição Normal



Distribuições Contínuas

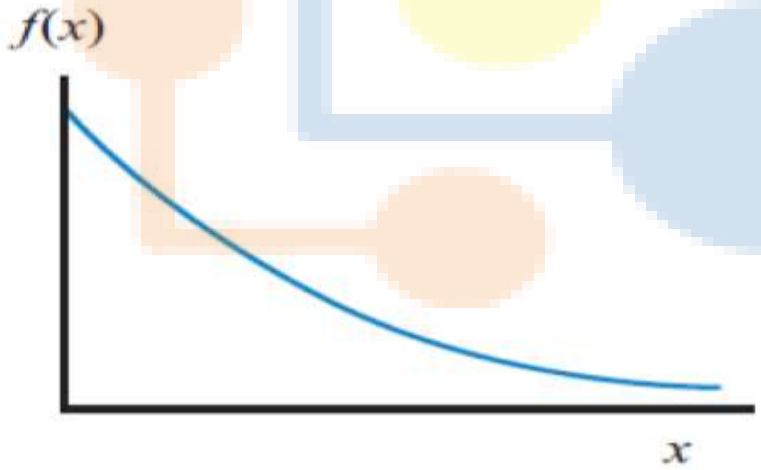
A Distribuição Normal é útil quando os dados tendem a estar próximos ao **centro da distribuição** (**próximos da média**) e quando **valores extremos** (**outliers**) são muito raros.

Distribuições Contínuas

Como a **Distribuição de Probabilidade Normal** é muito comum, ela é a ferramenta usada para calcular diversas estatísticas inferenciais, sendo muito importante em Machine Learning.

Distribuições Contínuas

Distribuição Exponencial



Distribuições Contínuas

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 Contínuas

Distribuição Uniforme

Distribuições Contínuas

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

Big Data Analytics com R e Microsoft Azure Machine Learning

Distribuição Normal

Seja Bem-Vindo(a)!

Distribuição Normal

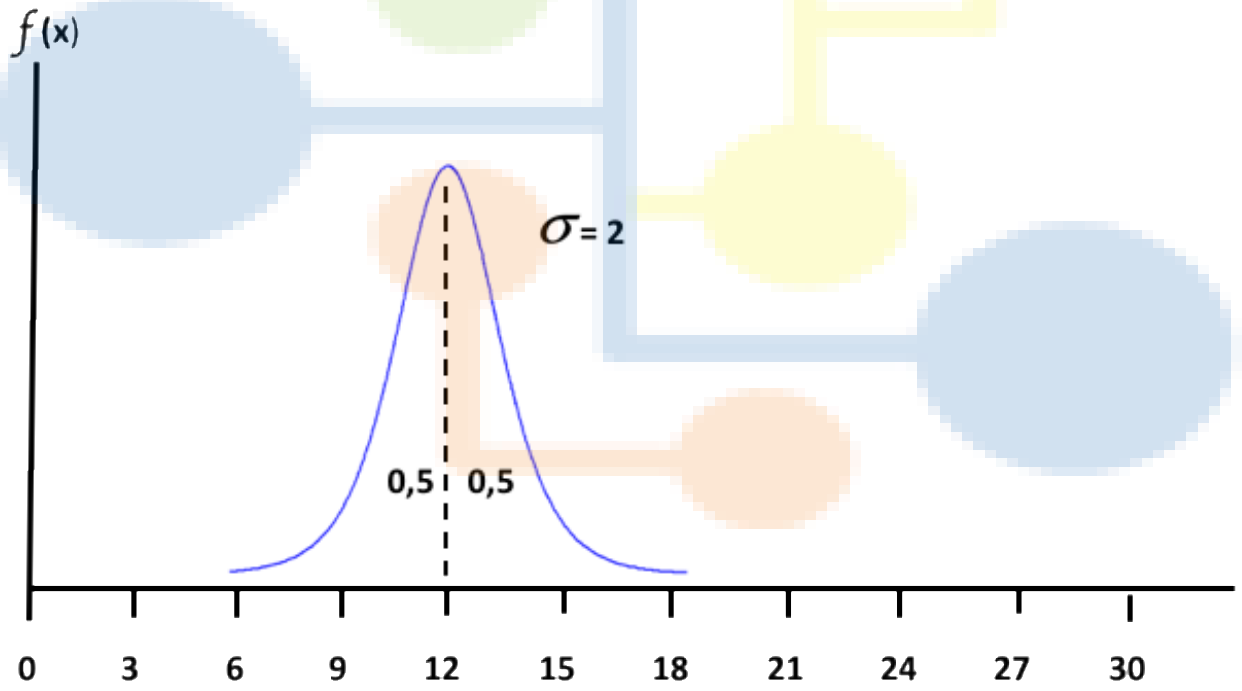
Uma variável randômica contínua que segue uma **Distribuição de Probabilidade Normal** tem uma série de características distintas.

Distribuição Normal

Imagine que o número de minutos que um cliente passa ao telefone com o pessoal de suporte da companhia de TV a cabo, segue uma distribuição normal, com uma média de 12 minutos (μ) e um desvio padrão de 2 minutos (σ).

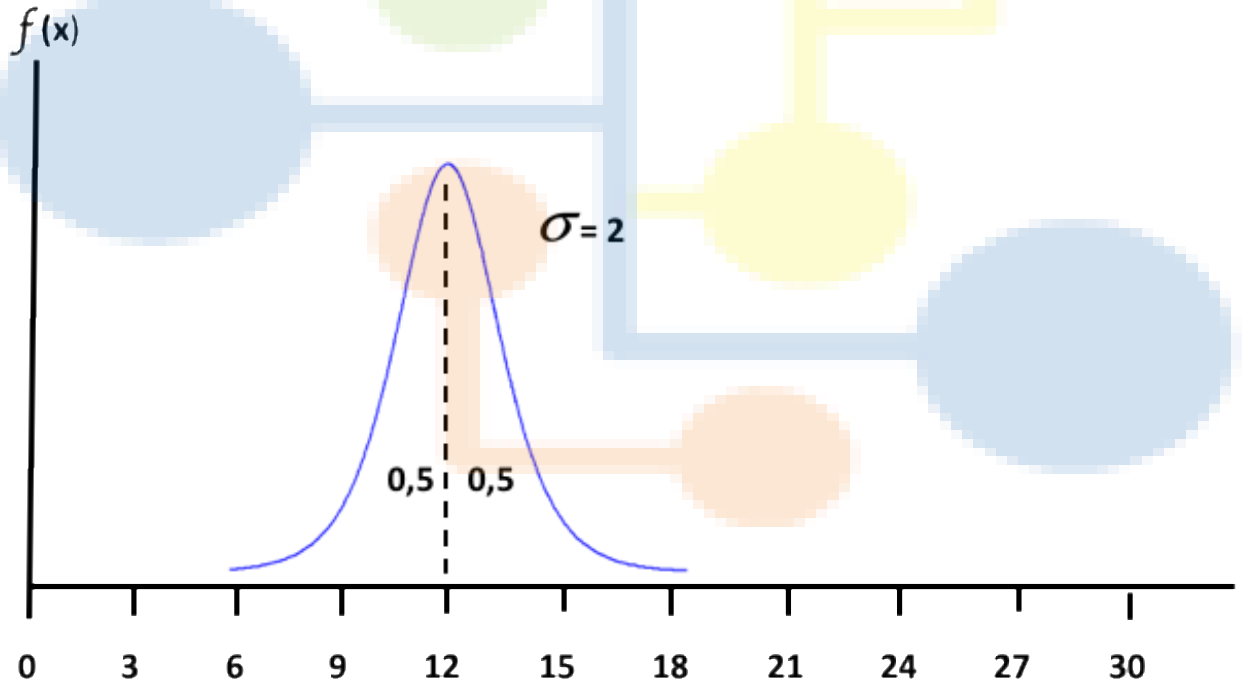
Distribuição Normal

A distribuição de probabilidade desta variável poderia ser representada no gráfico abaixo:

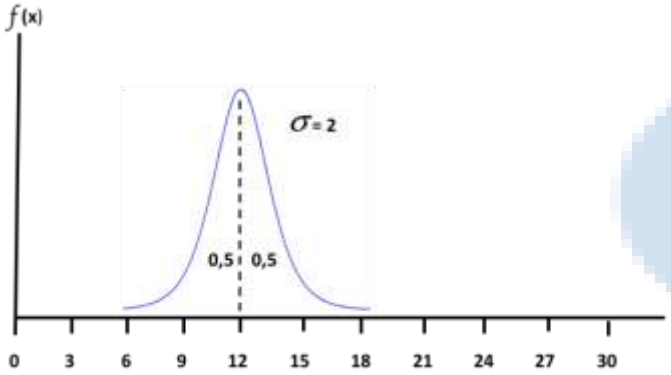


Distribuição Normal

De acordo com o gráfico, podemos fazer as seguintes observações sobre distribuição de probabilidade normal:



Distribuição Normal



A distribuição tem um formato de sino e simétrico em torno da média.

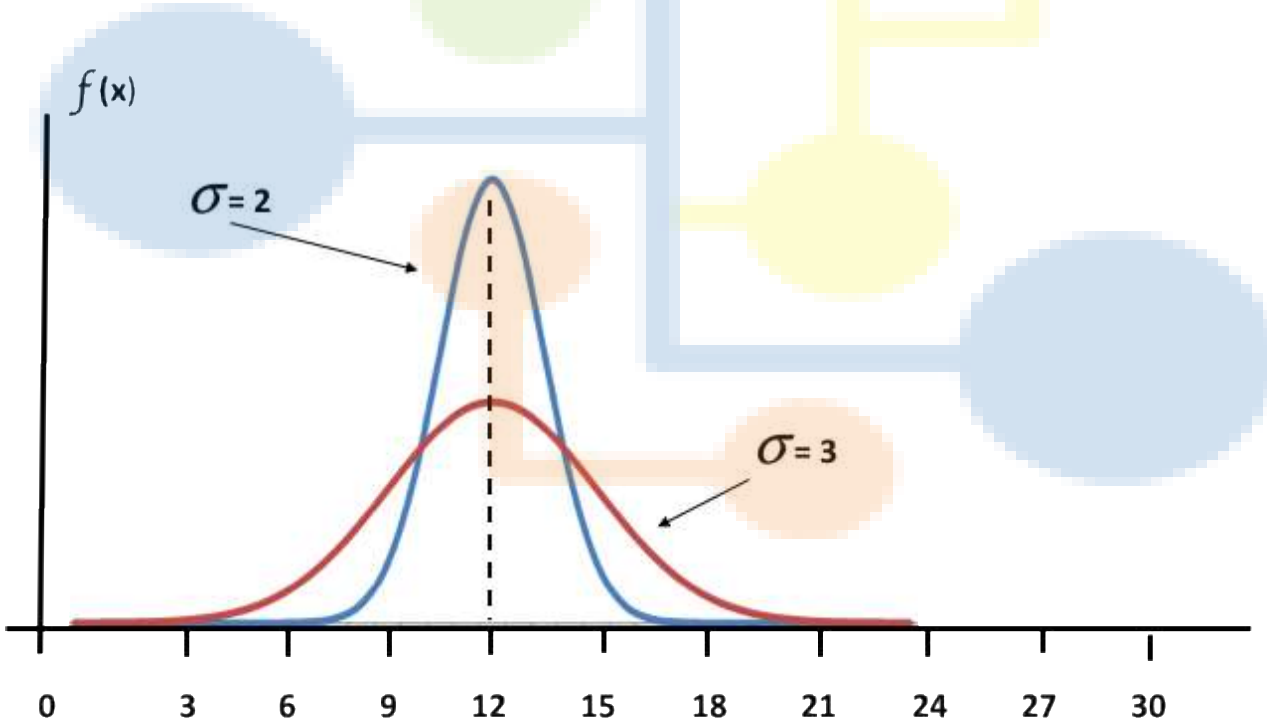
Como o formato da distribuição é simétrico, a média e a mediana possuem o mesmo valor, neste caso, 12 minutos.

Variáveis randômicas em torno da média, na parte mais alta da curva, tem maior probabilidade de ocorrer, que valores situados onde a curva é menor.

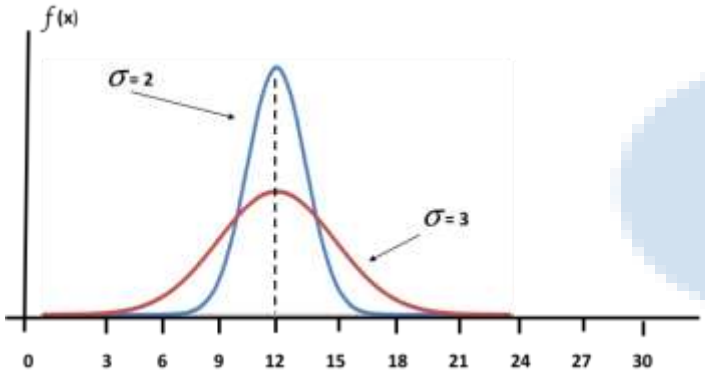
A parte final da curva, tanto do lado direito, quanto do lado esquerdo, em uma distribuição normal, se estende indefinidamente, nunca tocando o eixo x do gráfico.

Distribuição Normal

O **Desvio Padrão** tem uma função importante no formato da **curva** de uma **Distribuição Normal**.



Distribuição Normal



A linha vermelha possui um desvio padrão de 3 ($\sigma = 3$).

A curva ficou mais aberta em relação à média.

O tempo médio das ligações está entre 3 e 21 minutos e não mais entre 6 e 18 minutos, quando o desvio padrão é 2.

Um desvio padrão menor resulta em uma curva mais estreita.

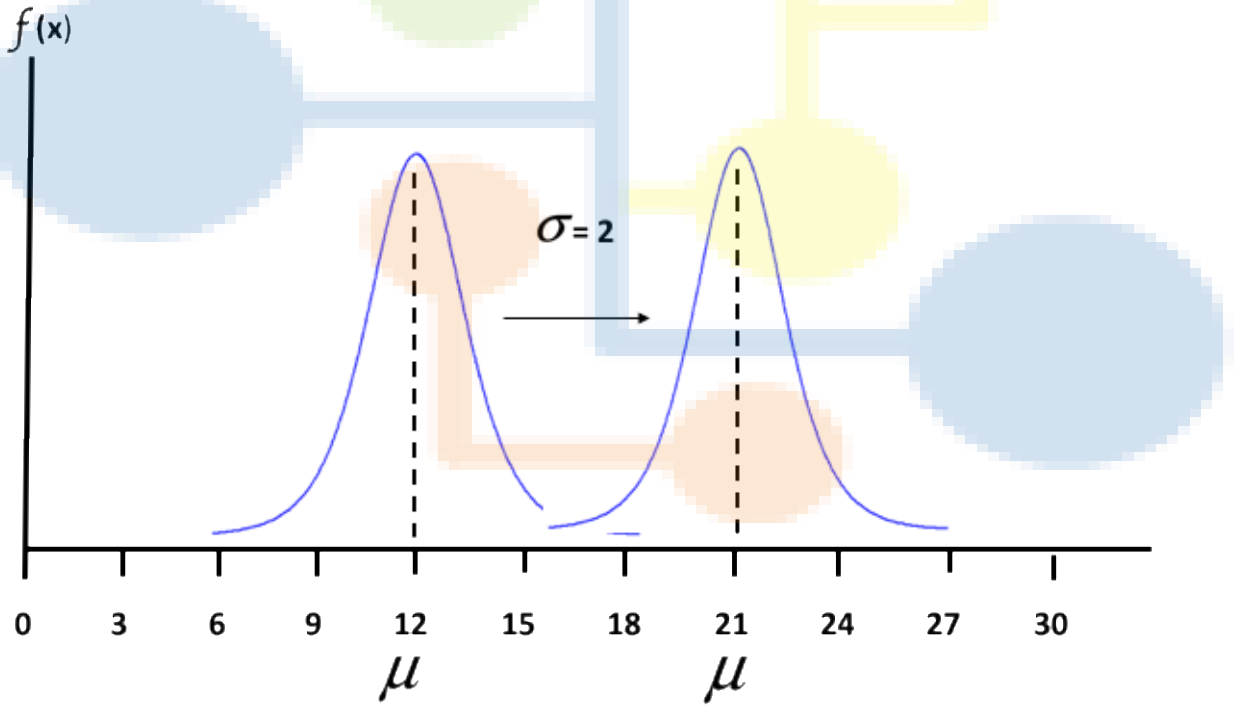
Um desvio padrão maior, faz com que a curva seja mais baixa e mais aberta.

Distribuição Normal

E se mudar a média, de 12 para 21 minutos e manter o desvio padrão de 2?

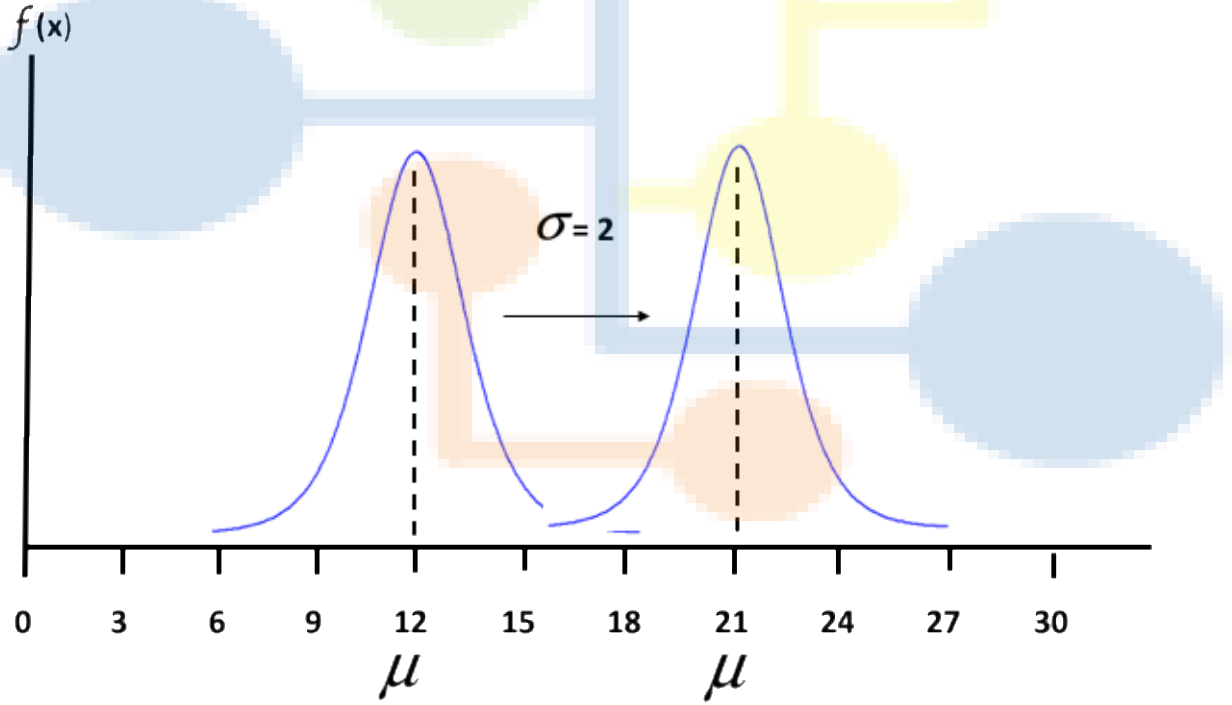
Distribuição Normal

Em cada um dos gráficos apresentados, as características de uma Distribuição de Probabilidade Normal são mantidas.



Distribuição Normal

Em cada caso, os valores de média e desvio padrão, descrevem completamente o formato da distribuição.



Distribuição Normal

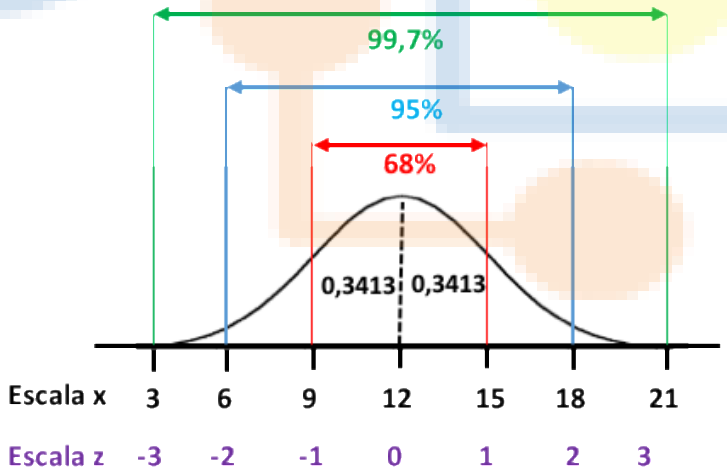
As probabilidades de distribuições normais podem ser calculadas através do uso de fórmulas, tabelas de probabilidade e softwares estatísticos como R, SAS e SPSS ou mesmo com pacotes estatísticos para a linguagem Python.

Distribuição Normal

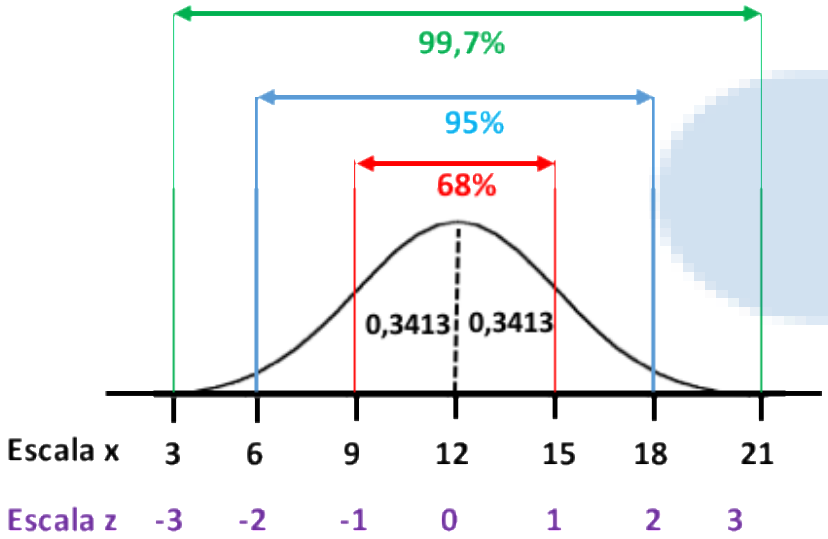
A Regra Empírica

Distribuição Normal

A **Regra Empírica** define o seguinte: se uma distribuição é simétrica e em formato de sino, aproximadamente 68%, 95% e 99% dos dados desta distribuição estarão em 1, 2 e 3 "*desvios padrão*" acima e abaixo da média, respectivamente:



Distribuição Normal

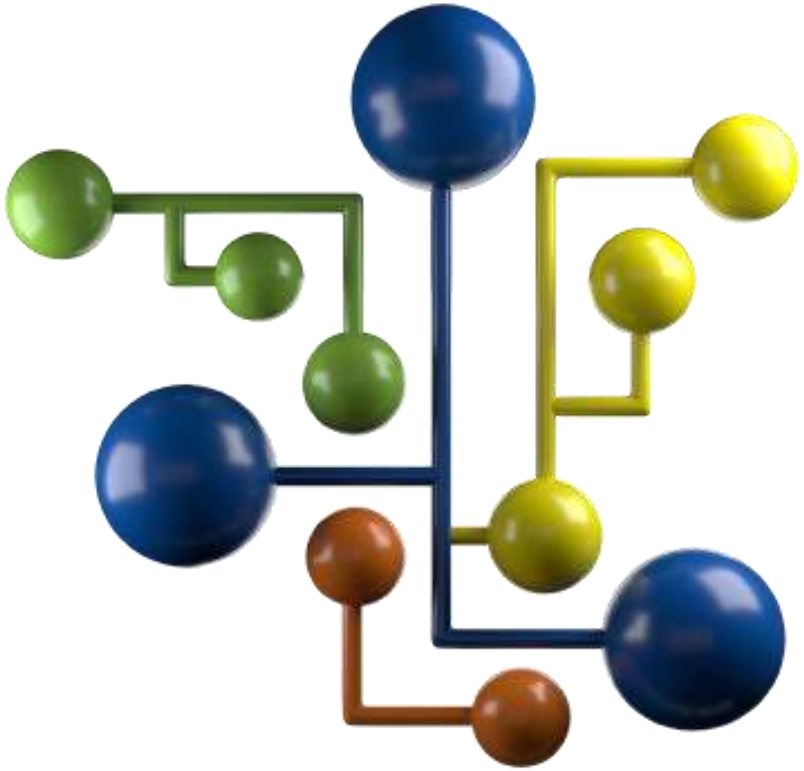


Ou seja, de acordo com a regra empírica, esperamos que:

68% das ligações fiquem entre **9** e **15** minutos,
95% entre **6** e **18** minutos e
99% entre **3** e **21** minutos.

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

Muito Obrigado por Participar!



Tenha uma Excelente Jornada de Aprendizagem.

Equipe Data Science Academy