



Data Science Academy

Big Data Analytics com R e Microsoft Azure Machine Learning Módulo 6

www.datascienceacademy.com.br



Data Science Academy



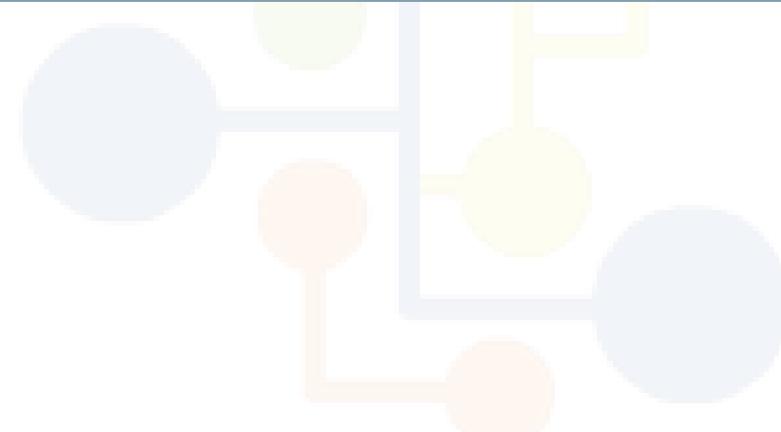
Análise Estatística de Dados



Data Science Academy



Introdução à Estatística



Data Science Academy

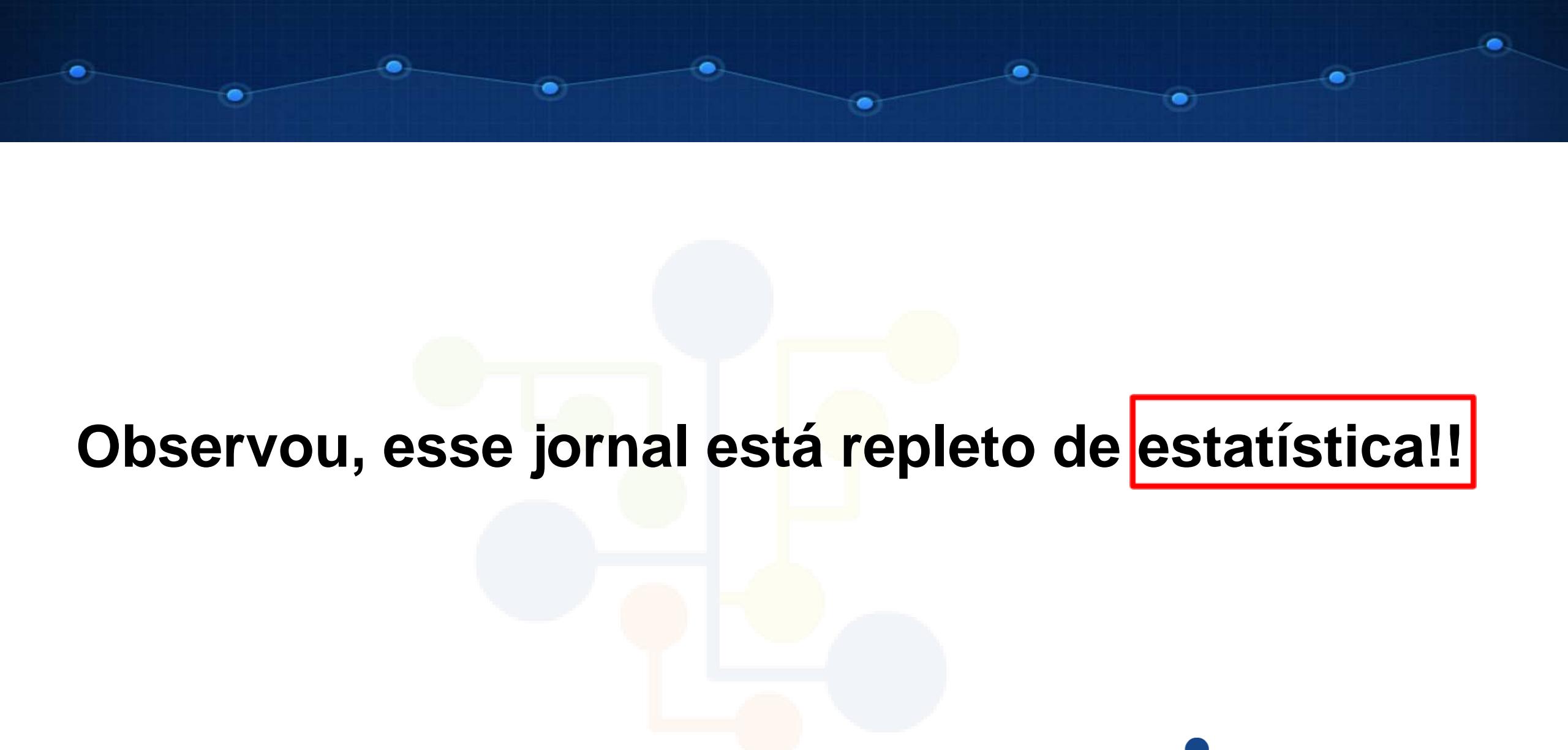
Vendas de imóveis na China avançam 18,2% de janeiro a setembro

Dólar fecha em alta com incerteza política e sobe 3,14% na semana

Indústria de São Paulo demite 18,5 mil em setembro, aponta Fiesp

Valor da produção do grão se aproxima de R\$ 100 bi





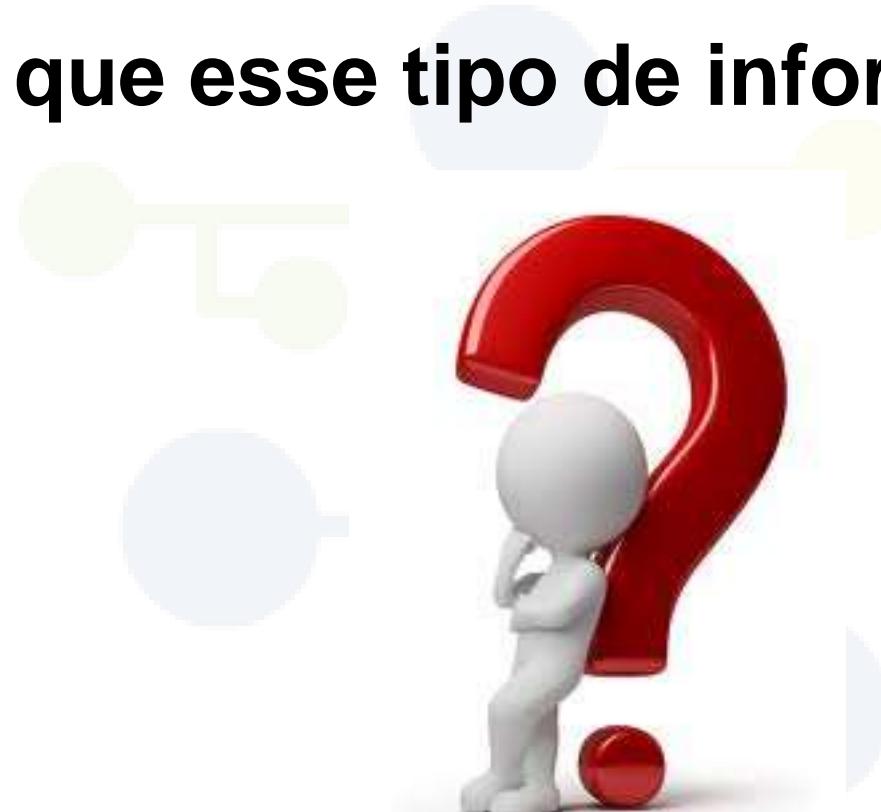
Observou, esse jornal está repleto de estatística!!



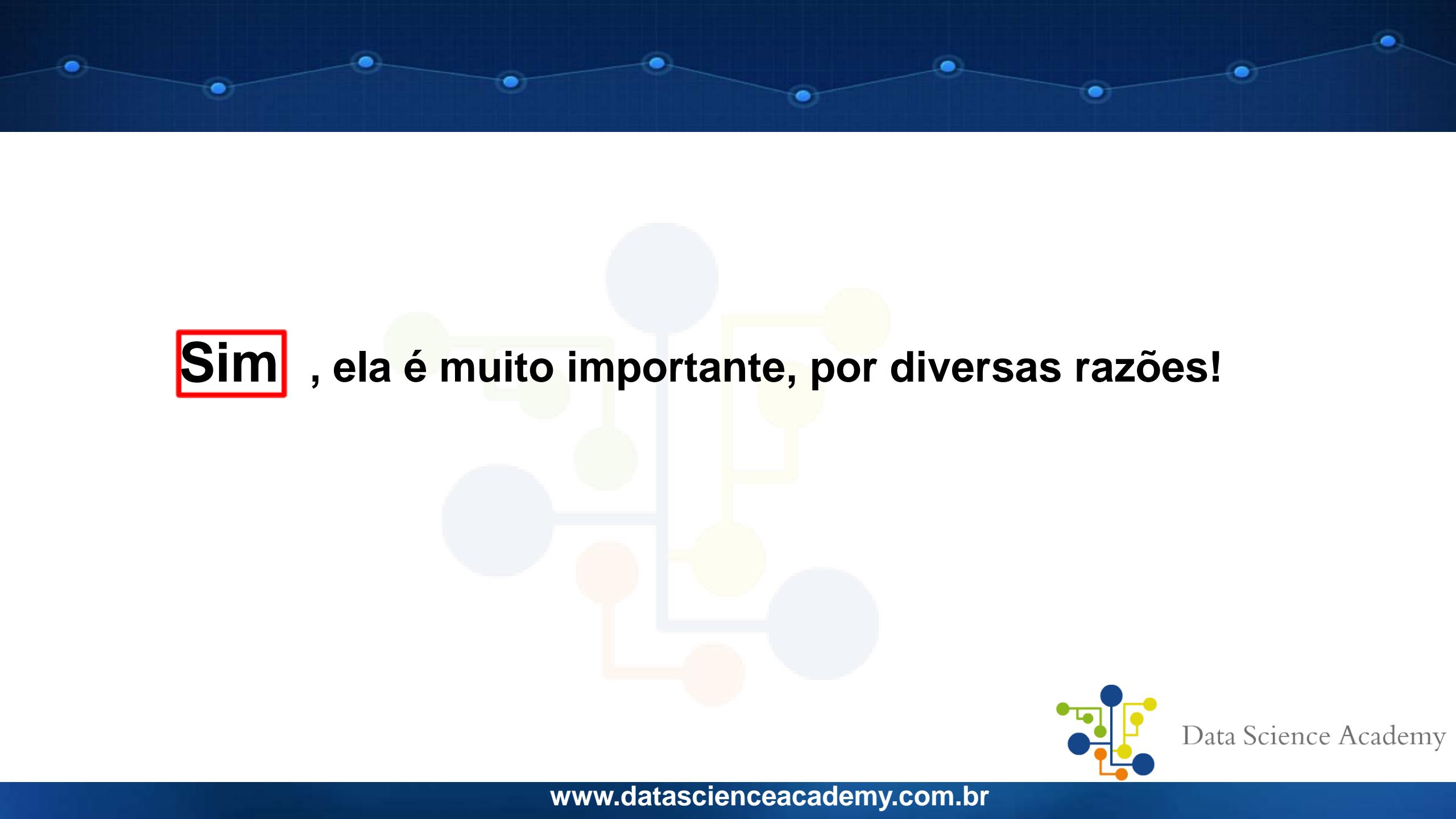
Data Science Academy



Você acredita que esse tipo de informação é importante?



Data Science Academy



Sim, ela é muito importante, por diversas razões!



Data Science Academy

Mas, isso é Estatística?



Data Science Academy

Números sobre finanças?



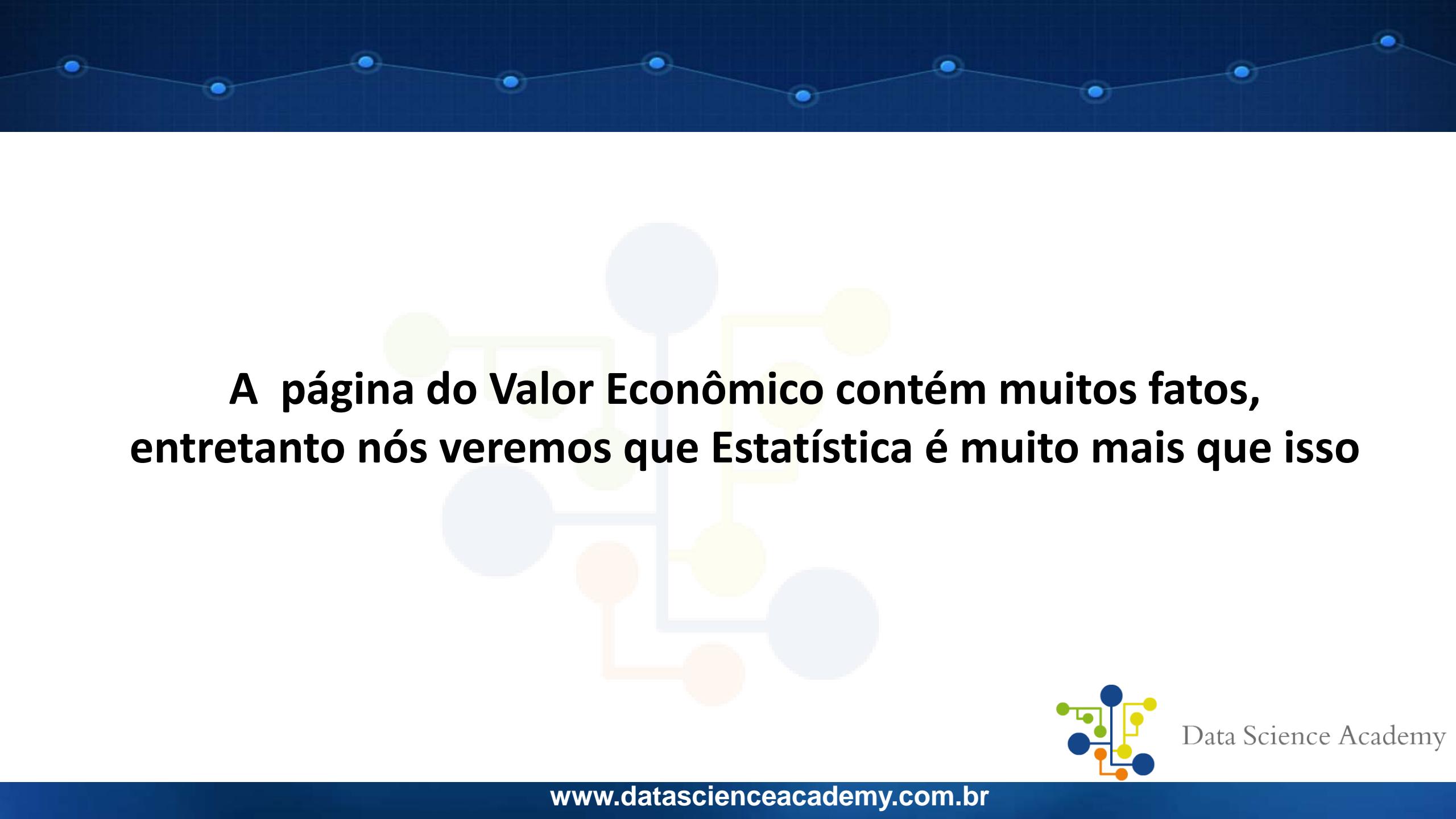
Data Science Academy



Sim e Não



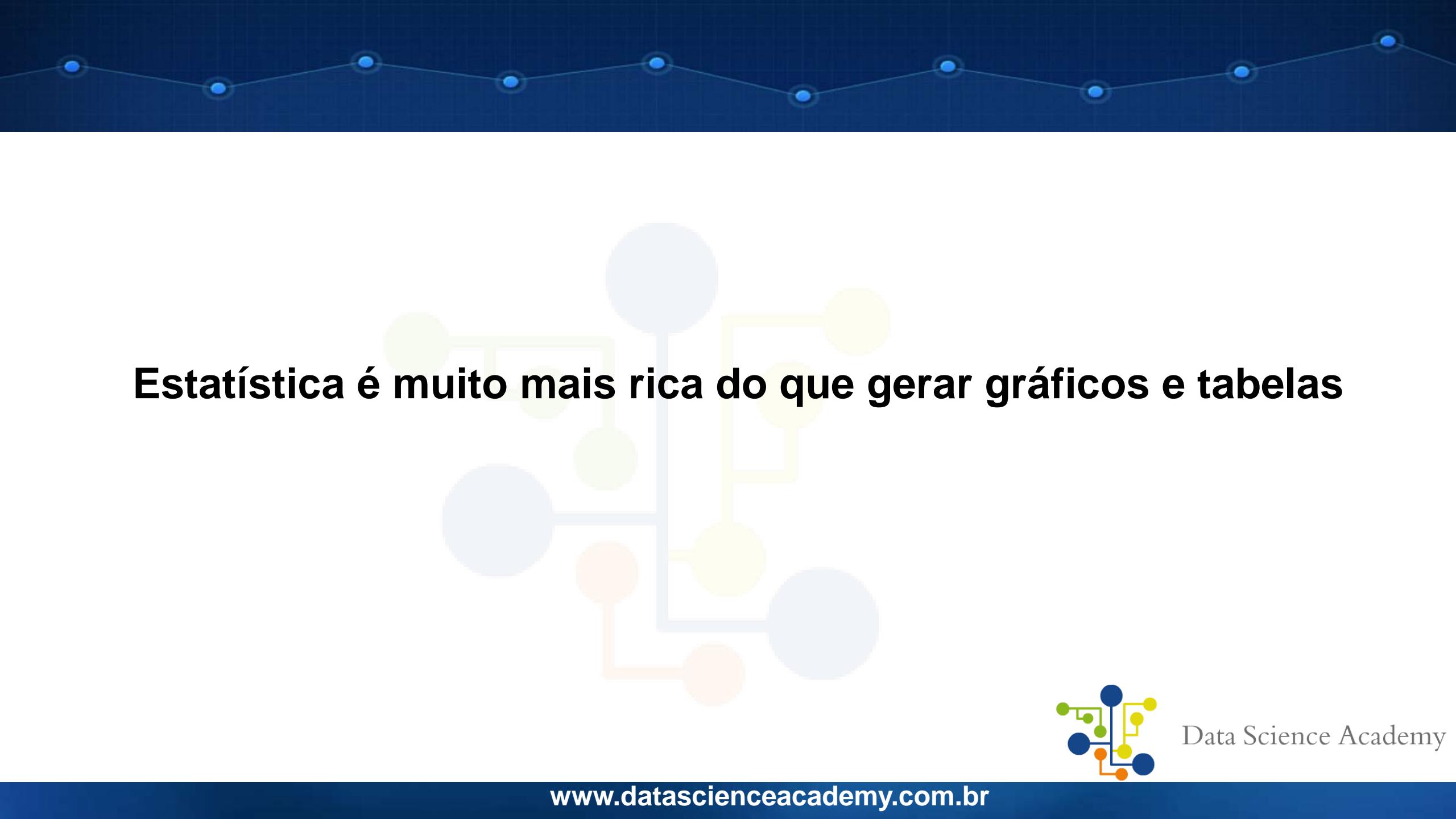
Data Science Academy



**A página do Valor Econômico contém muitos fatos,
entretanto nós veremos que Estatística é muito mais que isso**



Data Science Academy



Estatística é muito mais rica do que gerar gráficos e tabelas



Data Science Academy

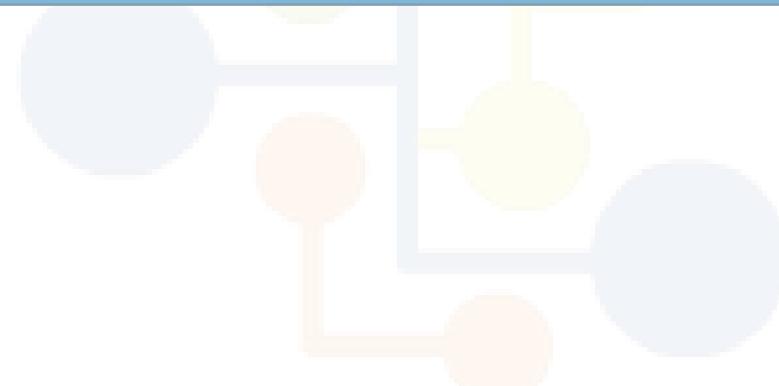
Estatística é muito mais rica do que gerar gráficos e tabelas



Data Science Academy



Como Fazer Análise de Dados



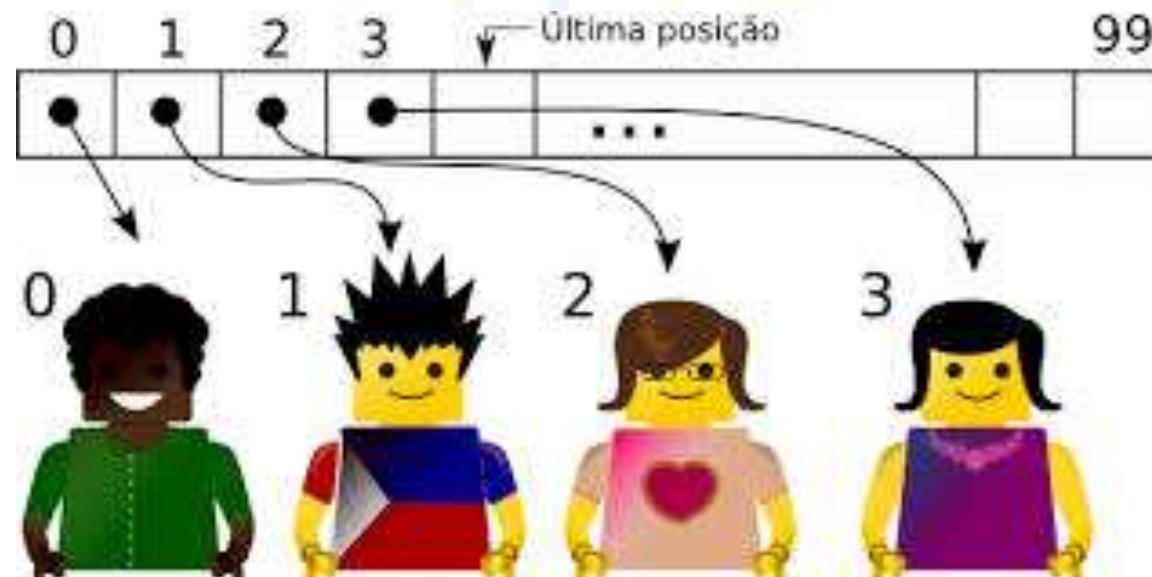
Data Science Academy

Muitas empresas possuem grandes bancos de dados, mas que seriam sem utilidade se estas informações não fossem analisadas...



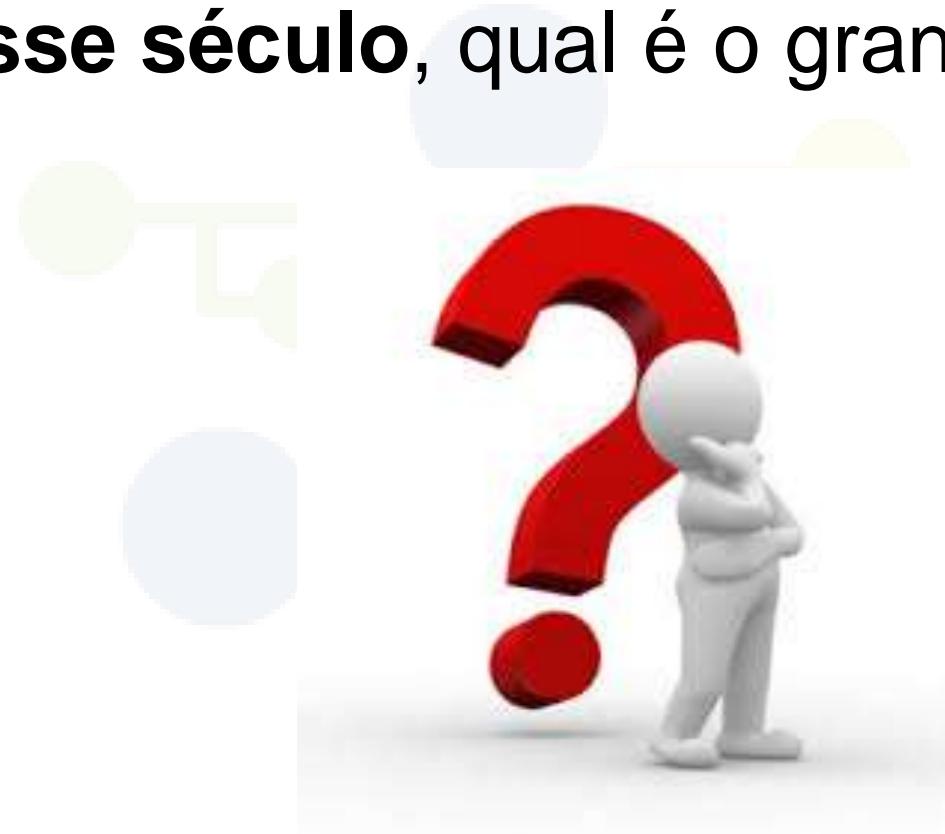
Data Science Academy

No século passado, a sociedade aprendeu como **armazenar** e **indexar** dados, para que informações pudessem ser extraídas.



Data Science Academy

E nesse século, qual é o grande desafio?

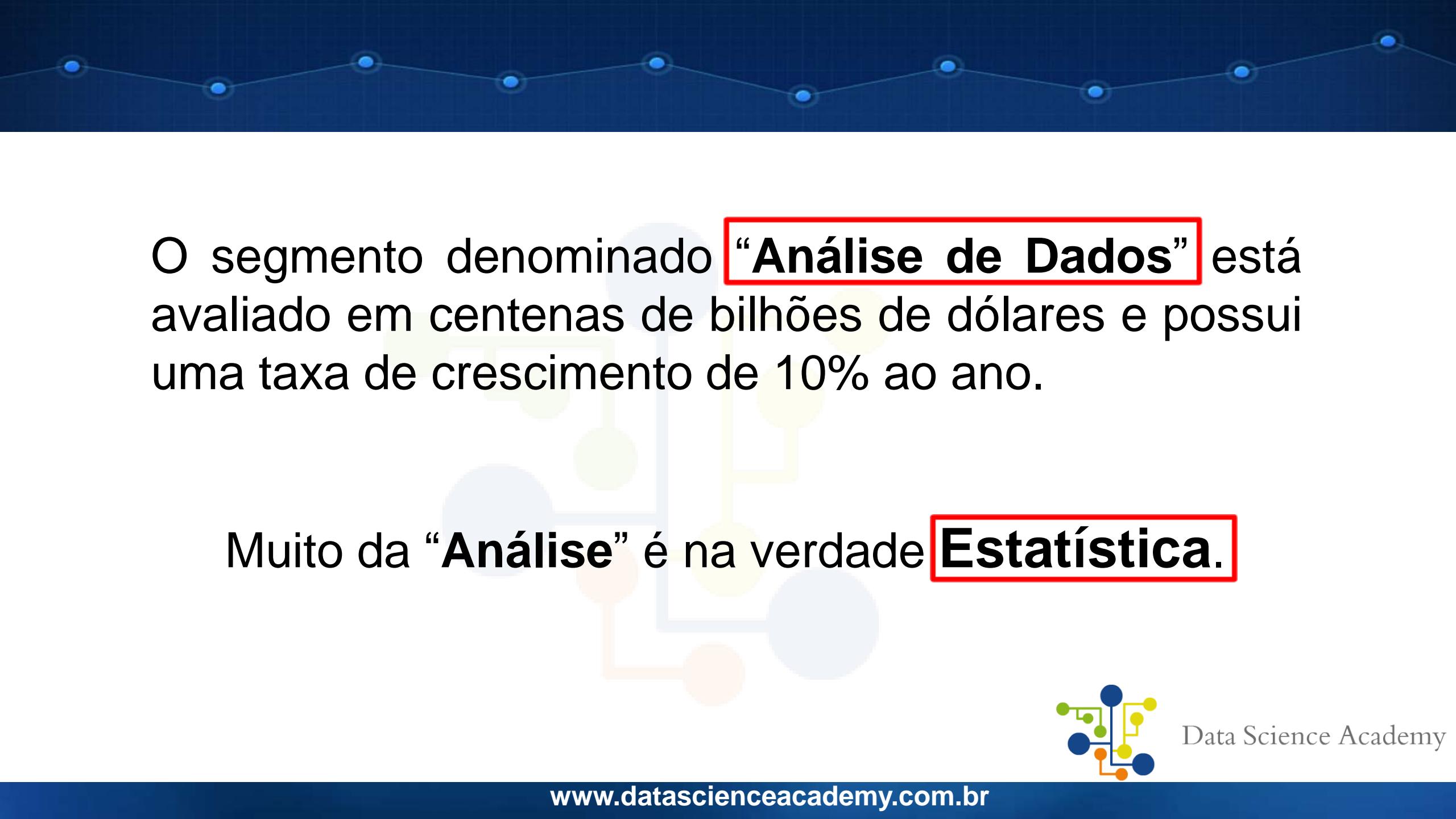


Data Science Academy

Neste século, o grande desafio está em **analisar** estes dados e torná-los **efetivos** na tomada de decisão.



Data Science Academy

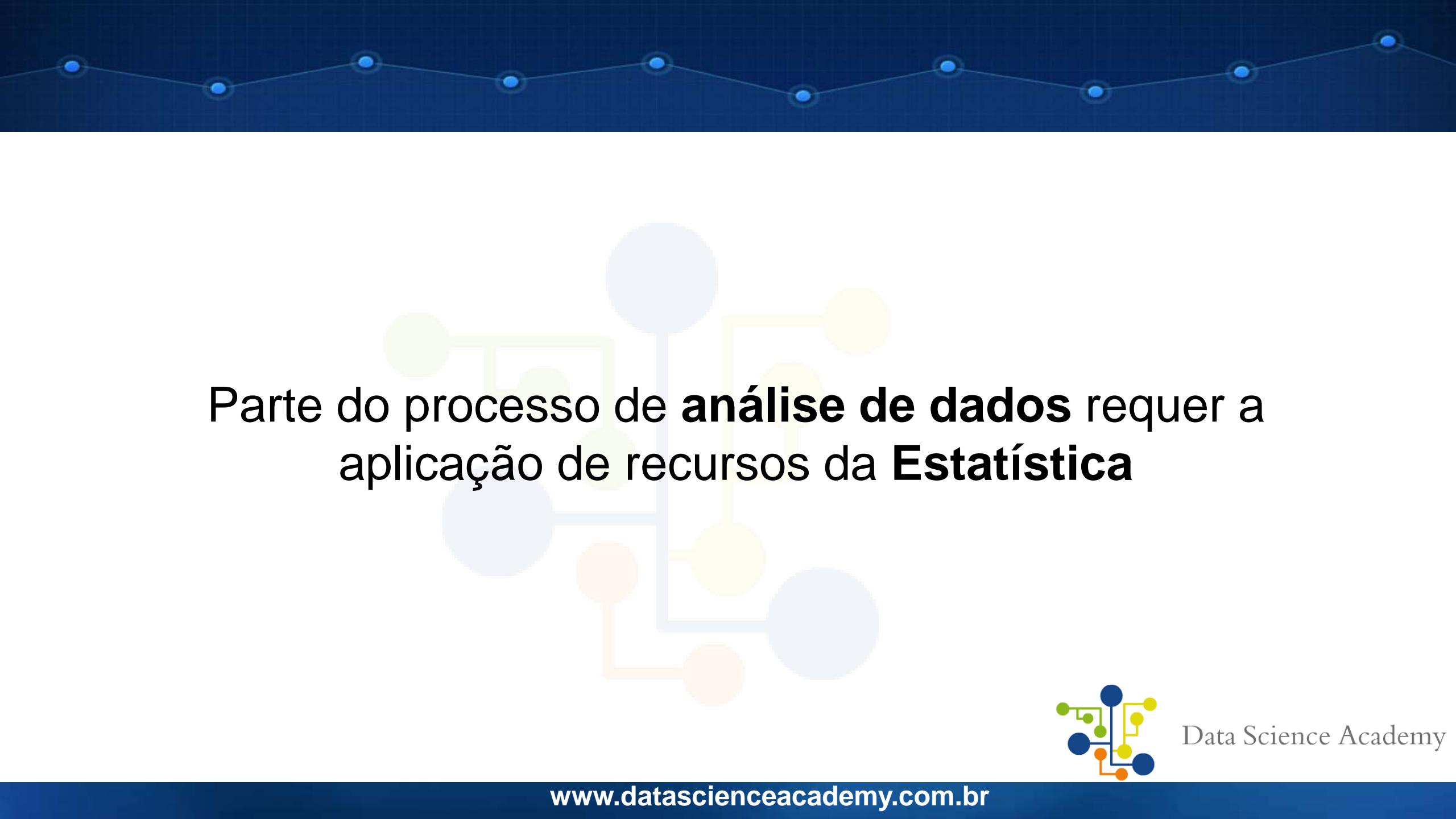


O segmento denominado **“Análise de Dados”** está avaliado em centenas de bilhões de dólares e possui uma taxa de crescimento de 10% ao ano.

Muito da “Análise” é na verdade **Estatística**.



Data Science Academy



Parte do processo de **análise de dados** requer a aplicação de recursos da **Estatística**



Data Science Academy



Parte do processo de **análise de dados** requer a aplicação de recursos da **Estatística**



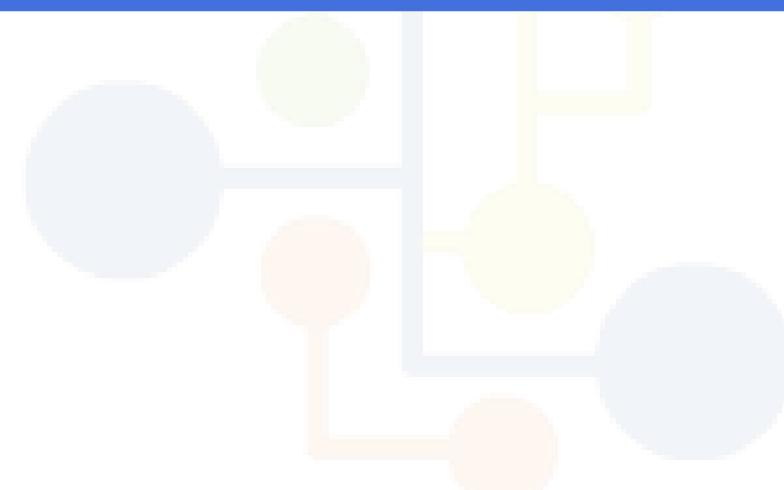
Data Science Academy

Por isso esse curso é tão importante. Pois é quase impossível realizar **análise de dados** profissional sem estatística. Elas estão intrinsecamente interligadas.



Data Science Academy

Exemplo



Data Science Academy

Empresas geram relatórios de vendas uma vez por mês



Data Science Academy

Relatórios são guardados até a próxima reunião da diretoria por mais um, dois ou três meses



ice Academy

Quando estas reuniões ocorrem, os dados já estão obsoletos



Data Science Academy

Como a Estatística pode ajudar a melhorar esse cenário?



Data Science Academy

Hoje as empresas conseguem registrar suas informações no momento da venda



Data Science Academy

Estes dados são incorporados em análises estatísticas através de análise em tempo real



Data Science Academy

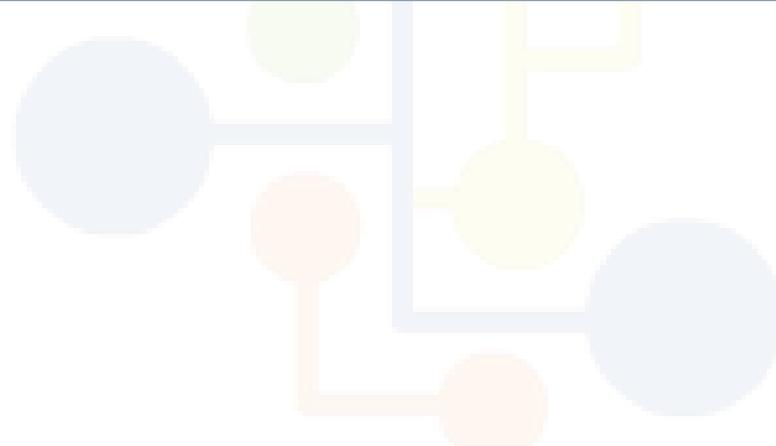
Para que tendências sejam detectadas
imediatamente e fiquem disponíveis a diretores e
executivos para **tomada de decisão**



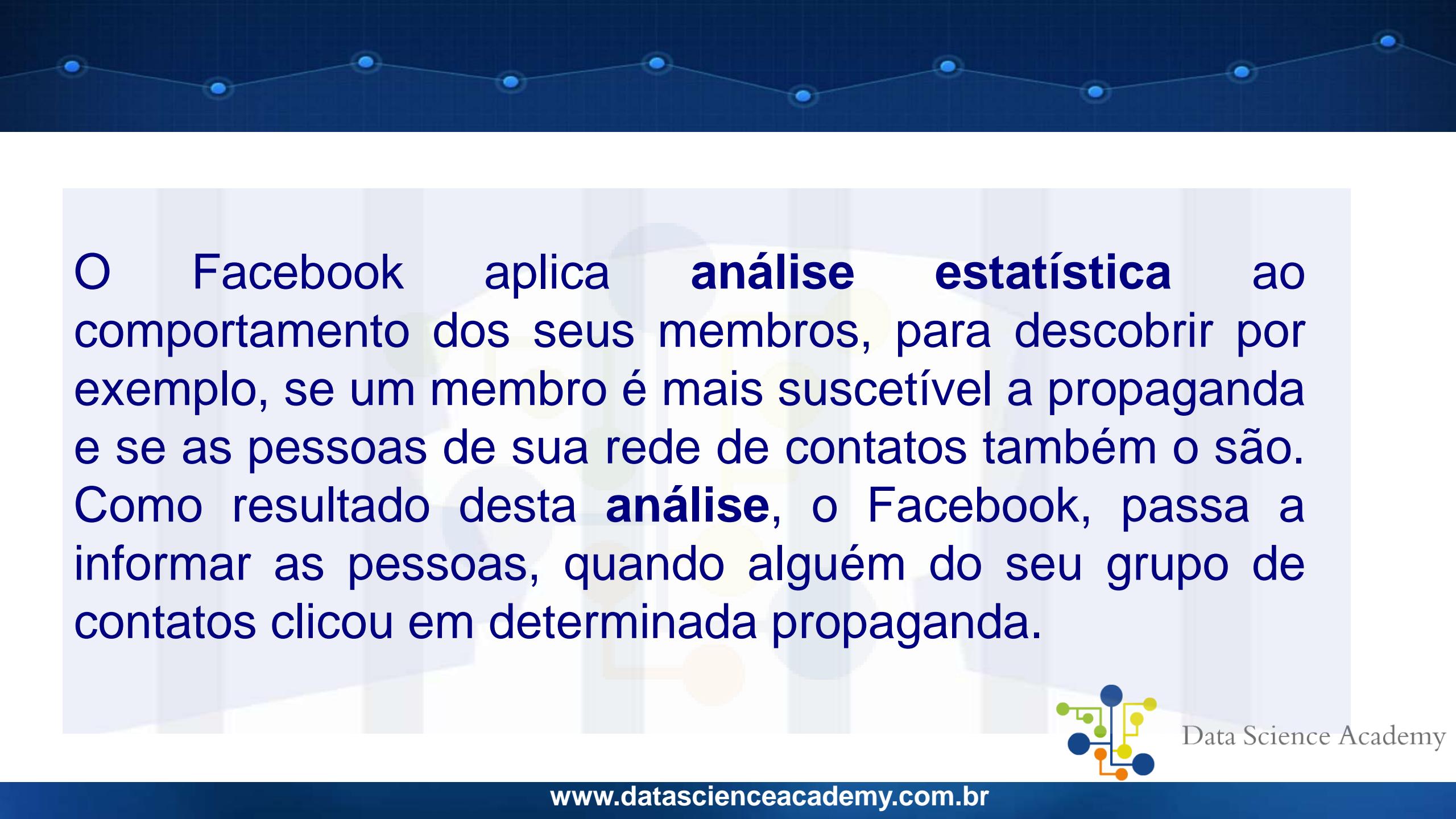
Data Science Academy



Exemplos de Análise de Dados



Data Science Academy

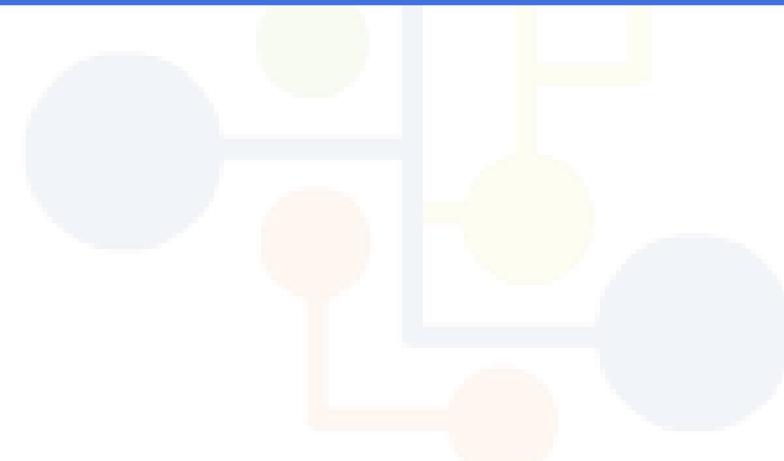


O Facebook aplica **análise estatística** ao comportamento dos seus membros, para descobrir por exemplo, se um membro é mais suscetível a propaganda e se as pessoas de sua rede de contatos também o são. Como resultado desta **análise**, o Facebook, passa a informar as pessoas, quando alguém do seu grupo de contatos clicou em determinada propaganda.



Data Science Academy

Exemplo II



Data Science Academy

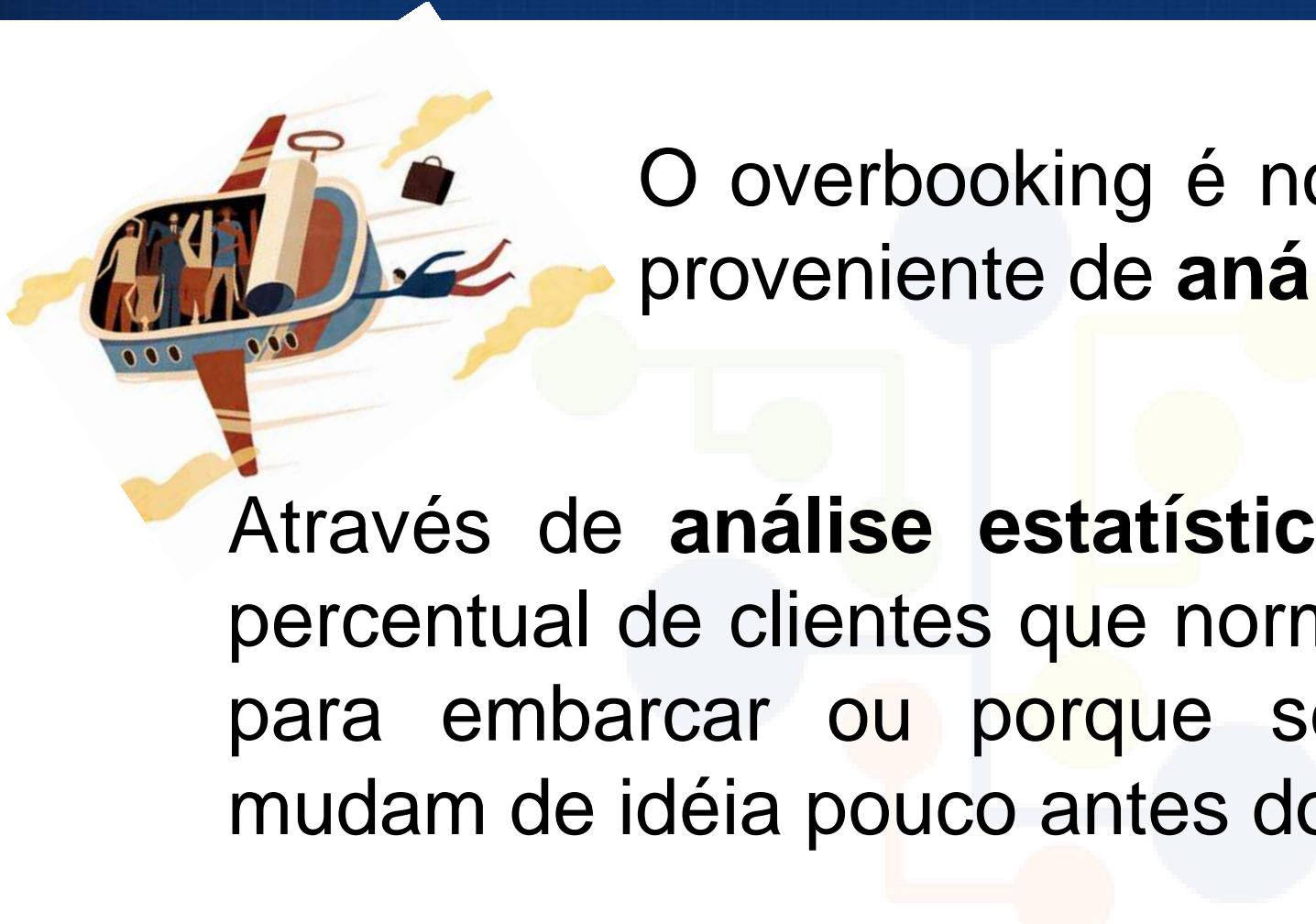
Você já deve ter ouvido falar no termo “overbooking”.



O overbooking é normalmente proposital e proveniente de **análise estatística**.



Data Science Academy



O overbooking é normalmente proposital e proveniente de **análise estatística**.

Através de **análise estatística**, a empresa avalia o percentual de clientes que normalmente não aparecem para embarcar ou porque se atrasam, ou porque mudam de idéia pouco antes do voo.



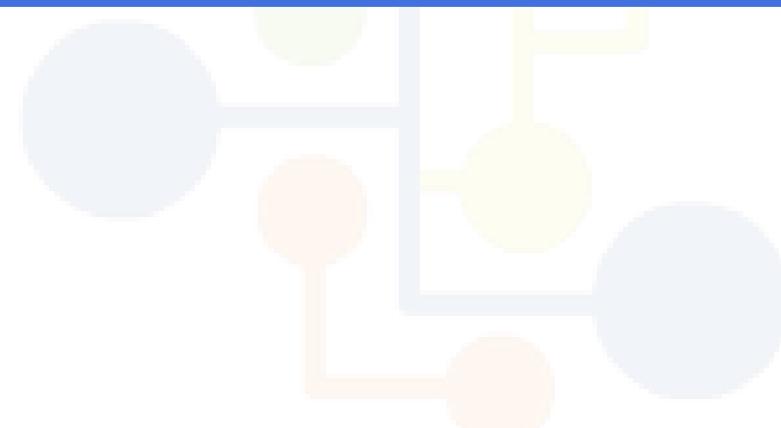
Data Science Academy



Outra análise feita por empresas aéreas nos EUA, mostraram que clientes que são vegetarianos (e que solicitavam este tipo de refeição ao comprar seus bilhetes) dificilmente perdião o voo e nestes voos praticamente não havia overbooking, pois a taxa de ocupação era sempre grande.



Exemplo III



Data Science Academy

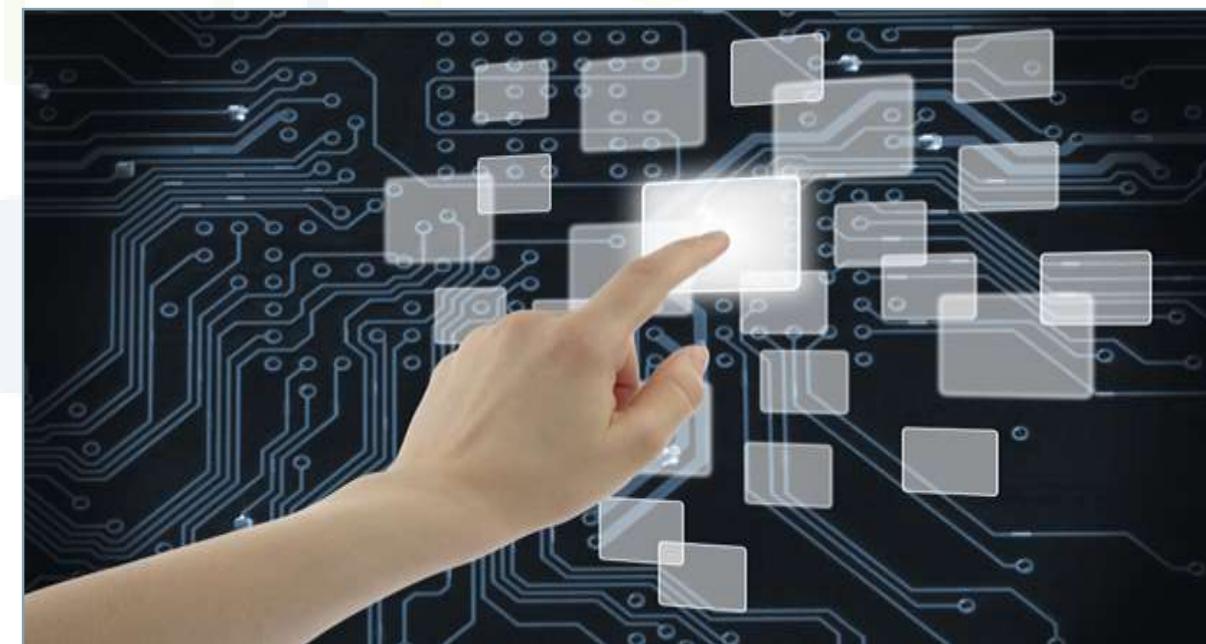
Por conta do inverno rigoroso.

Uma rede de varejo canadense percebeu (após **análises estatísticas**), que alguns dos suprimentos que mais eram vendidos na época de inverno, era um tipo específico de legume, que pode ser consumido sem a necessidade de ser cozido e que tem longo prazo de validade.



Data Science Academy

Nos três exemplos apresentados, perguntas específicas de negócio foram respondidas através da **Análise de Dados e decisões foram tomadas.**



Data Science Academy



Nós temos um desafio muito maior....

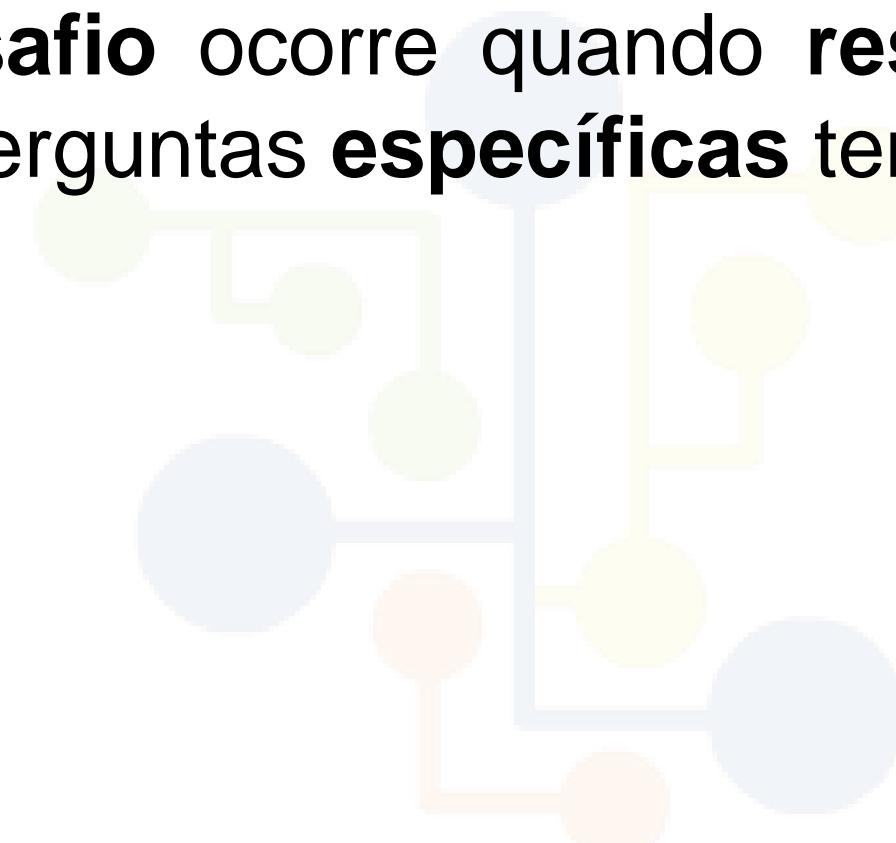


Você consegue perceber qual é esse desafio?



Data Science Academy

O real **desafio** ocorre quando **respostas** são geradas **sem** que perguntas **específicas** tenham sido feitas:



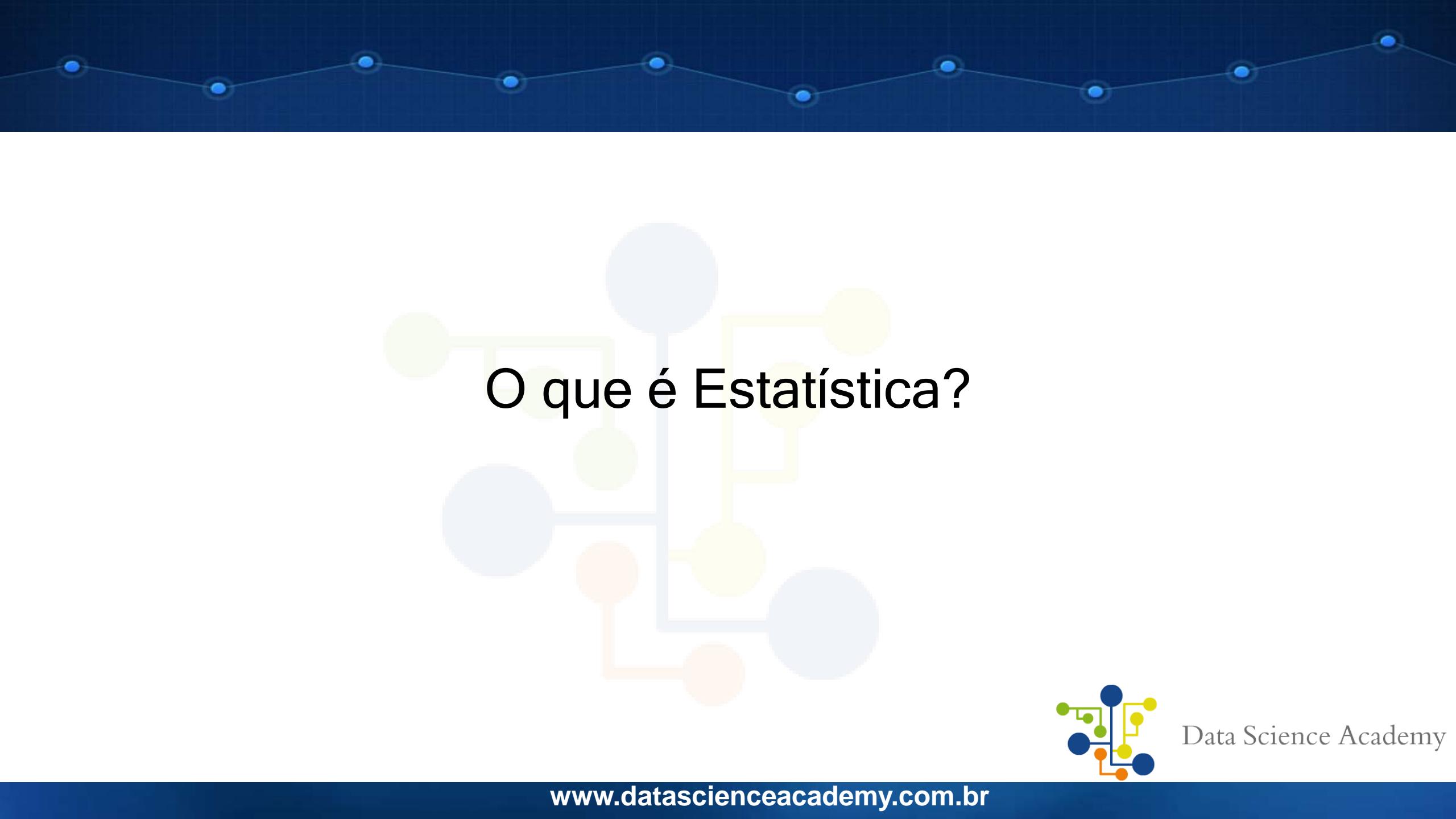
Data Science Academy

“ Como podemos melhorar nosso jeito de fazer negócios, utilizando nosso enorme banco de dados e ligando-o com bancos de dados externos, públicos e disponíveis? ”

“Esta pergunta nos permite aplicar **análises estatísticas** e pensar “**fora da caixa**”.



Data Science Academy



O que é Estatística?



Data Science Academy



Estatística = base da economia global neste
século.



Data Science Academy



Não era essa a resposta que você esperava?

Ou



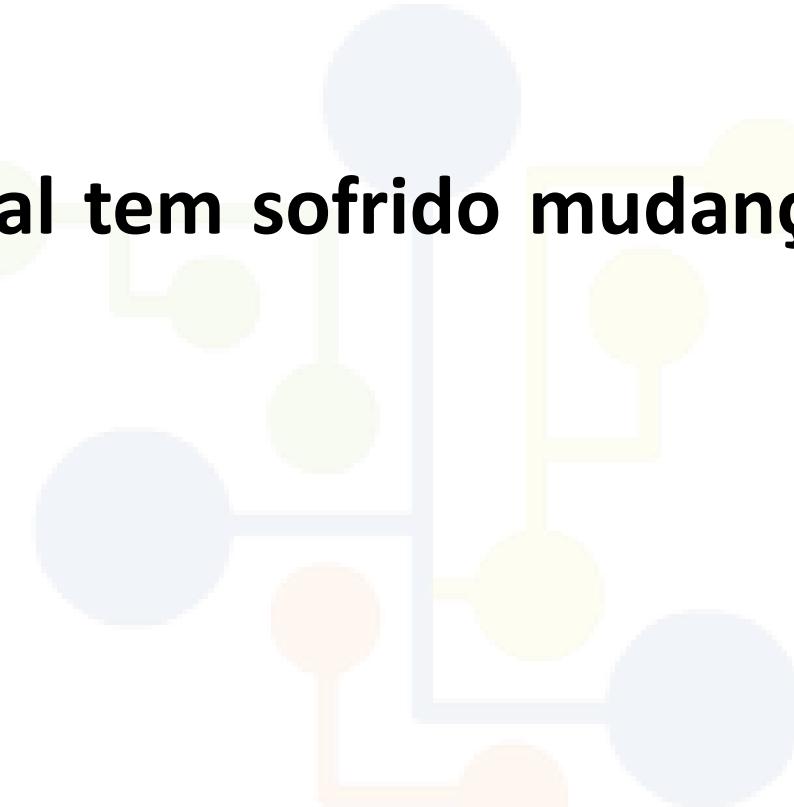
Parece muito genérica?



Data Science Academy



A economia global tem sofrido mudanças drásticas ao longo dos anos, tais como:



Data Science Academy



1

Revolução da Agricultura → atualmente o mundo produz mais alimentos através de cultivos, do que através de caça.

2

Revolução Industrial → fábricas de produção em massa deram ao mundo uma imensa variedade de opções de produtos.

3

Revolução da Informação → a tecnologia nos deu uma grande variedade de produtos eletrônicos, tornou a indústria mais eficiente e aumentou consideravelmente a quantidade de informação a nossa disposição.

4

Revolução da Análise de Dados → estamos no meio desta revolução e o volume de dados gerados pela humanidade, nos traz o desafio de conseguir extrair informação útil. A análise estatística é a chave desta revolução.



Data Science Academy

A Análise Estatística é a chave dessa revolução!!



Data Science Academy

**As três
grandes áreas
da Estatística**

Probabilidade



Estatística Descritiva



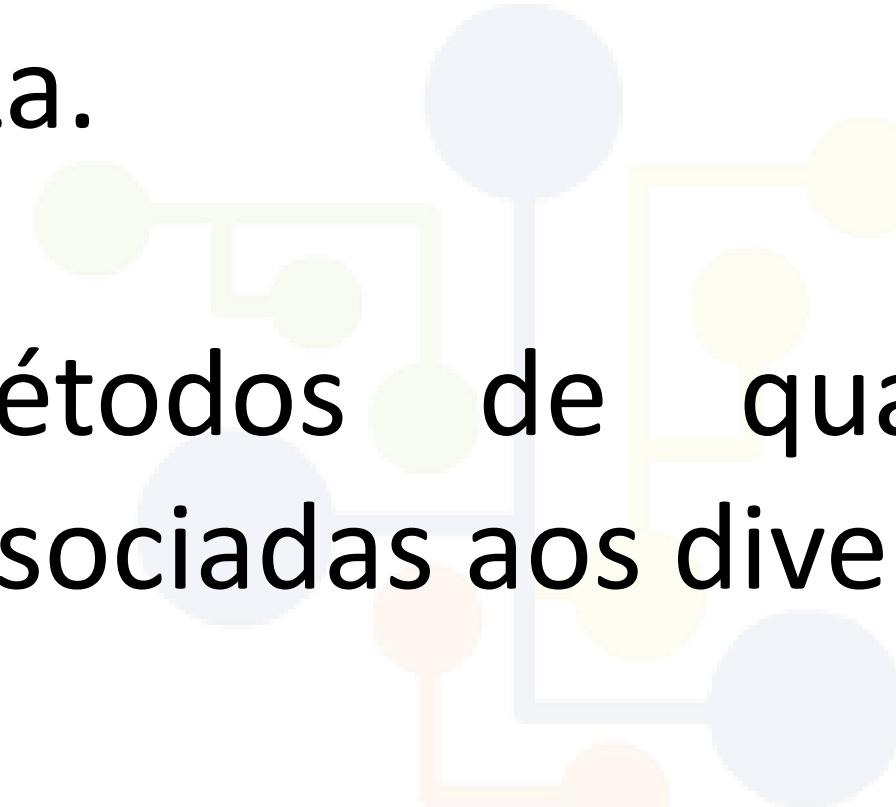
Estatística Inferencial



Data Science Academy



Probabilidade – estudo da aleatoriedade e da incerteza.



Utiliza métodos de quantificação das chances associadas aos diversos resultados.



Data Science Academy

Exemplo



Data Science Academy

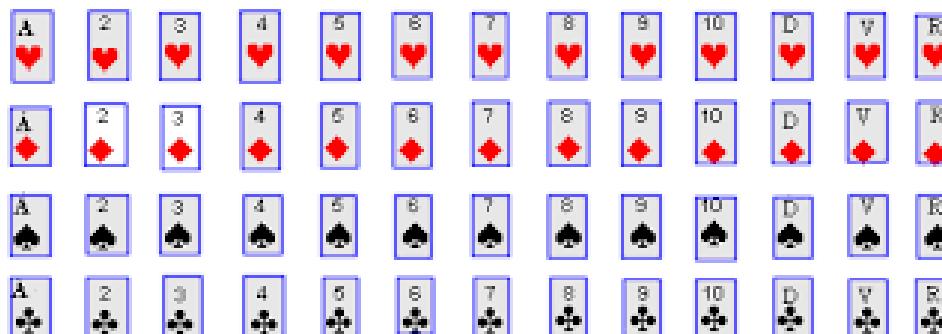
Lançamento da moeda. Qual probabilidade de no lançamento da moeda sair cara ou coroa?



Data Science Academy

Carta do baralho.

Possibilidades num baralho de 52 cartas

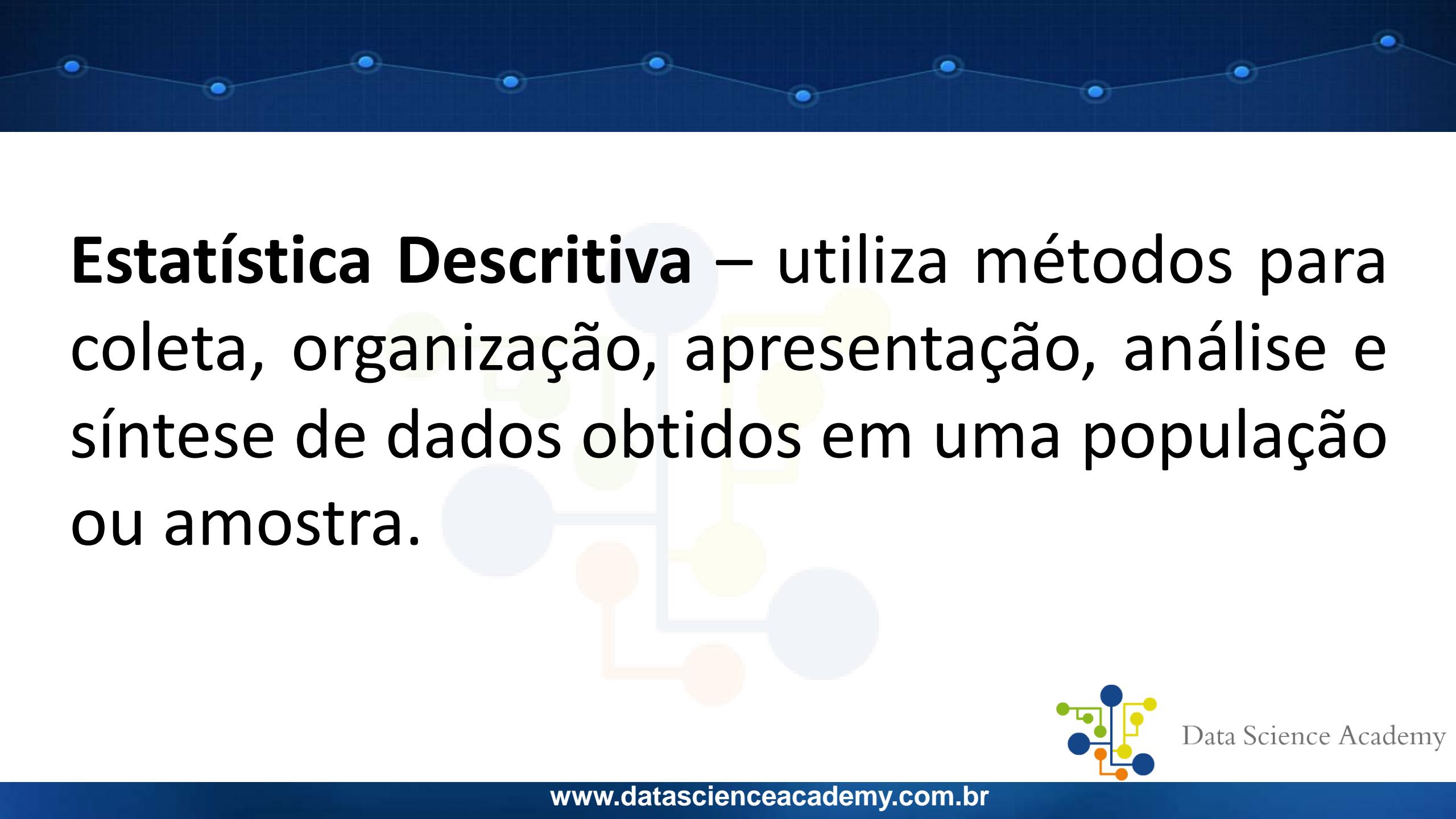


Data Science Academy

Estatística do tempo. Qual a probabilidade de sol no fim de semana?



Data Science Academy



Estatística Descritiva – utiliza métodos para coleta, organização, apresentação, análise e síntese de dados obtidos em uma população ou amostra.



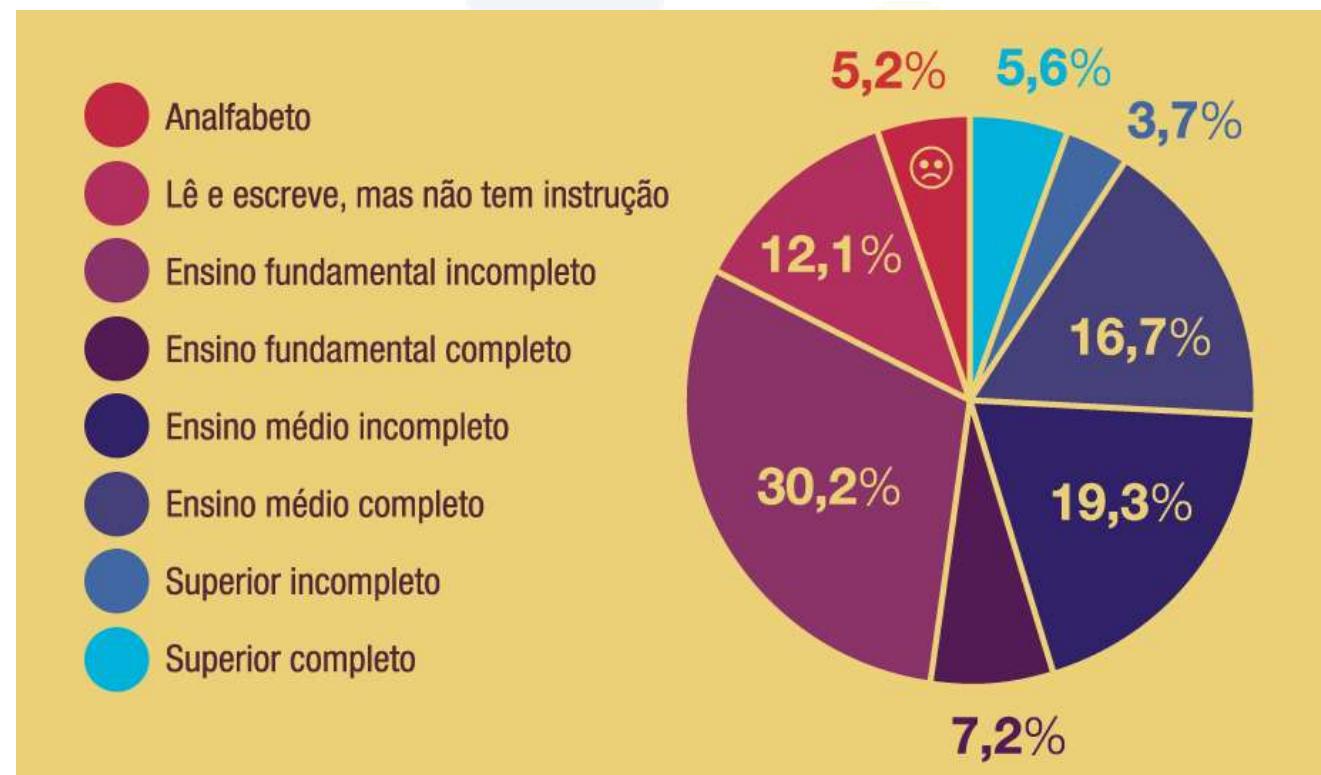
Data Science Academy

Exemplo



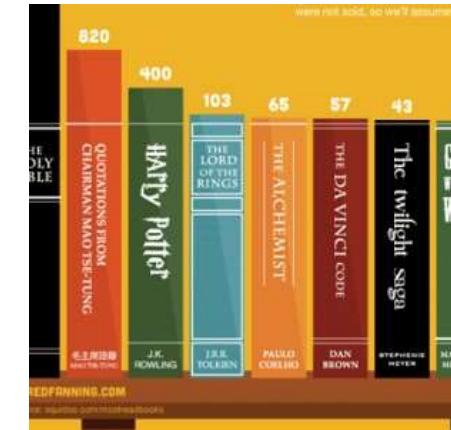
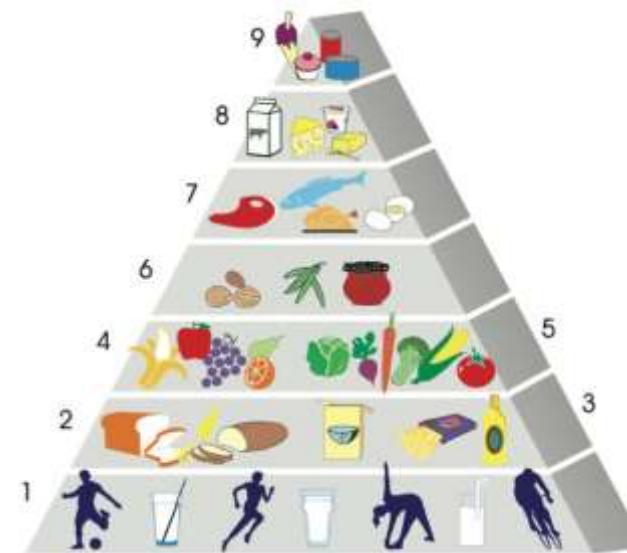
Data Science Academy

Como determinar o grau de escolaridade de uma determinada classe social?



Data Science Academy

Qual o sexo dessa classe social tem maior grau de escolaridade?



Quais os livros mais lidos por essa classe social?

Quais as suas preferências alimentares?

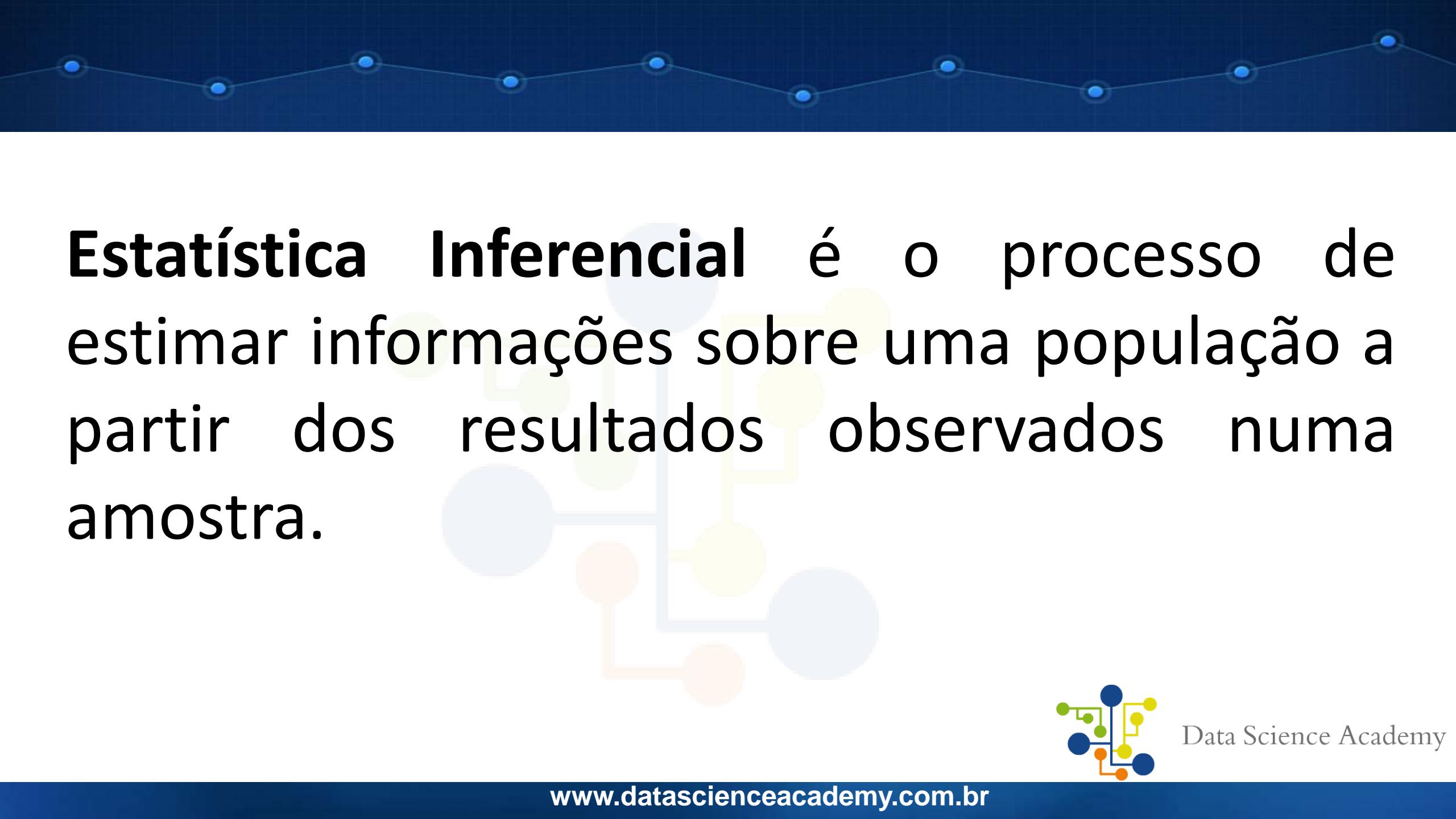
Data Science Academy



Na Estatística Descritiva é preciso estruturar os procedimentos para realizar as análises e obter as informações.



Data Science Academy



Estatística Inferencial é o processo de estimar informações sobre uma população a partir dos resultados observados numa amostra.



Data Science Academy

Exemplo



Data Science Academy

Tempo de conclusão do ensino médio para determinada classe social.

Nesse caso nós aplicamos os princípios da estatística descritiva numa amostra e usamos esse resultado para fazer a estatística inferencial. Você deve procurar entender esses valores e o comportamento da população.

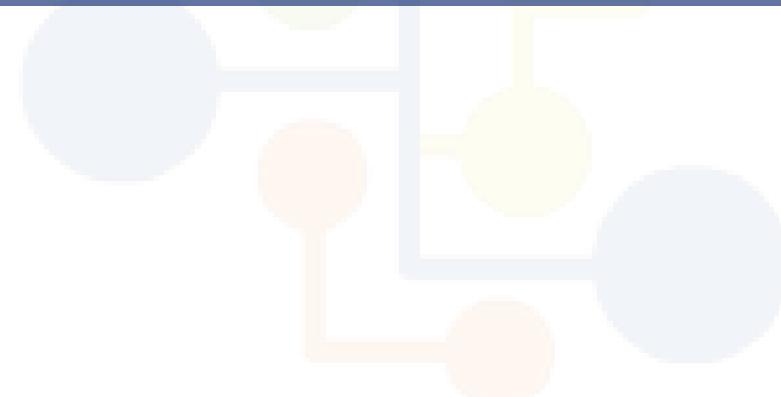
Ao final fazemos o levantamento a partir dos resultados e obtemos o tempo de conclusão.



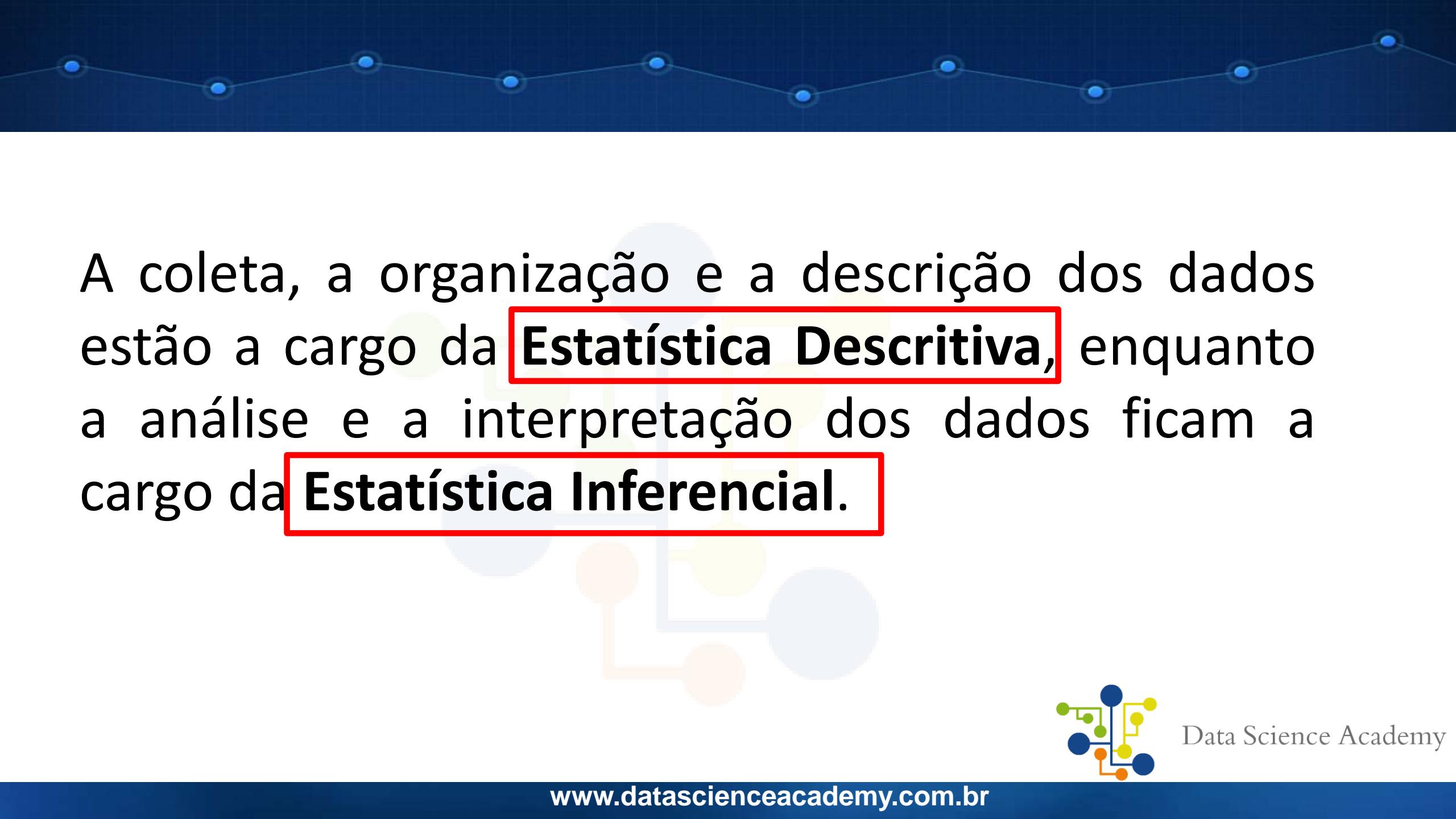
Data Science Academy



Portanto,



Data Science Academy



A coleta, a organização e a descrição dos dados estão a cargo da **Estatística Descritiva**, enquanto a análise e a interpretação dos dados ficam a cargo da **Estatística Inferencial**.



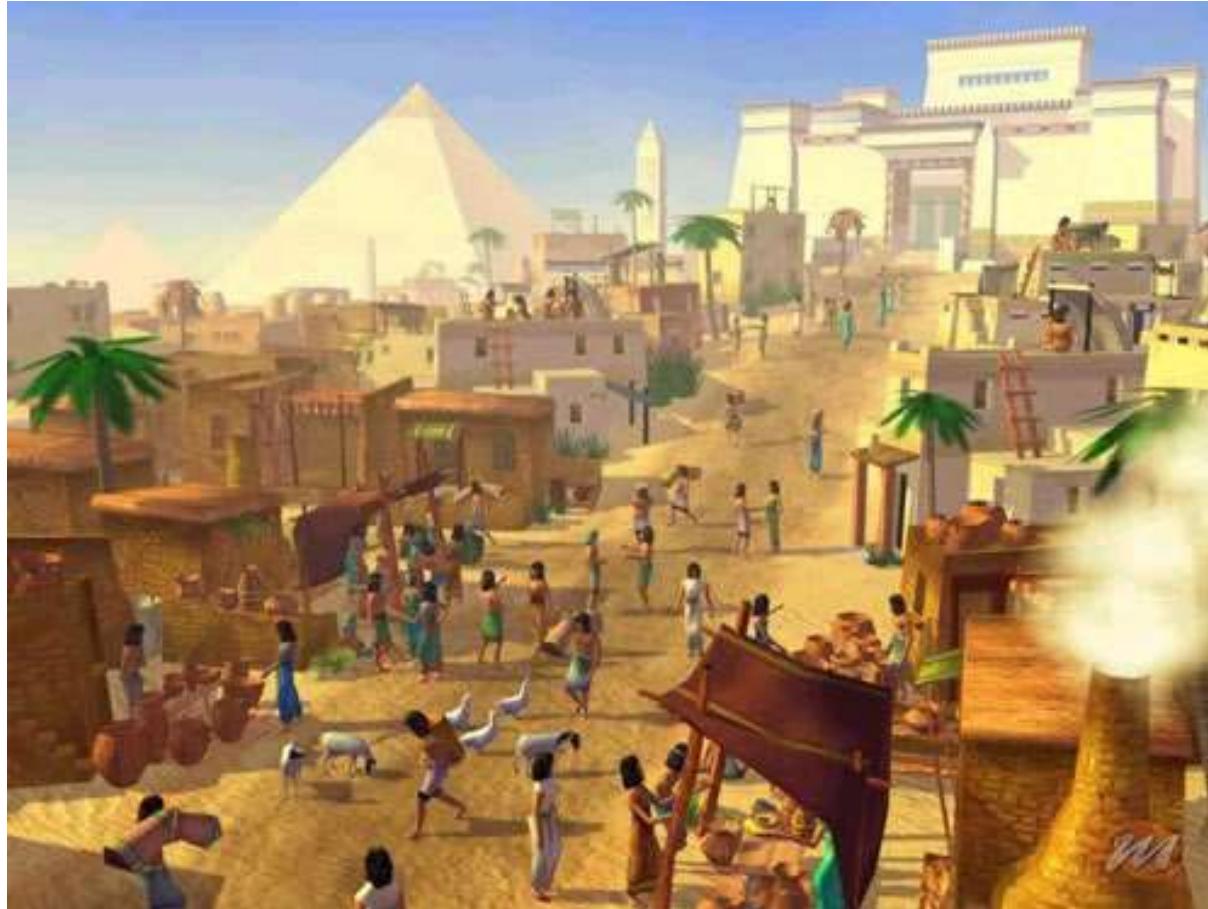
Data Science Academy



Percebeu como a Estatística tem um
papel relevante no mundo atual?



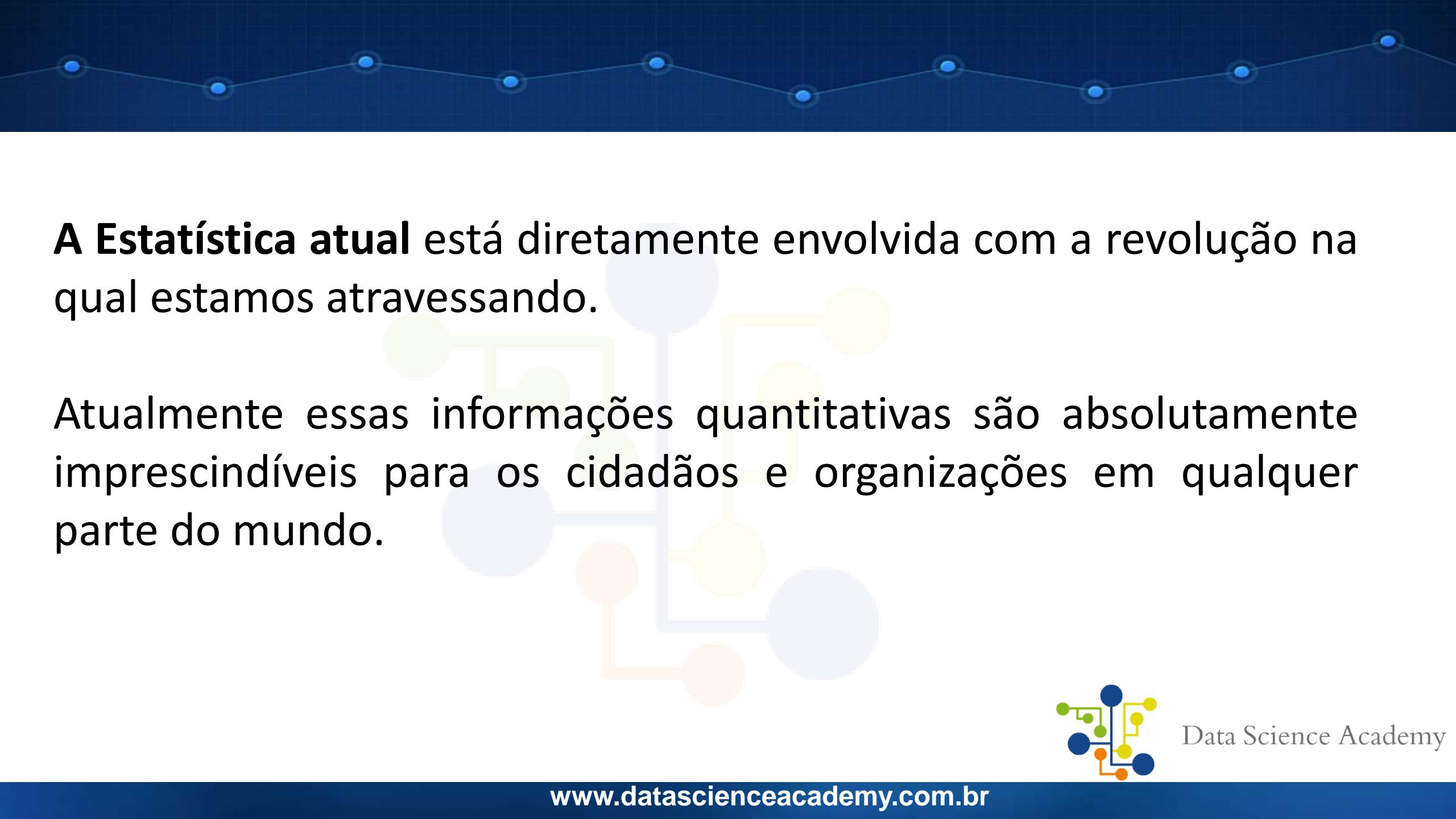
Data Science Academy



Embora a Estatística teve sua origem no antigo Egito, onde os governantes buscavam uma forma de quantificar as informações tais como número de nascimentos, número de óbitos, número de pessoas na família, condição sócio cultural, faixa salarial, religião e sexo.



Data Science Academy

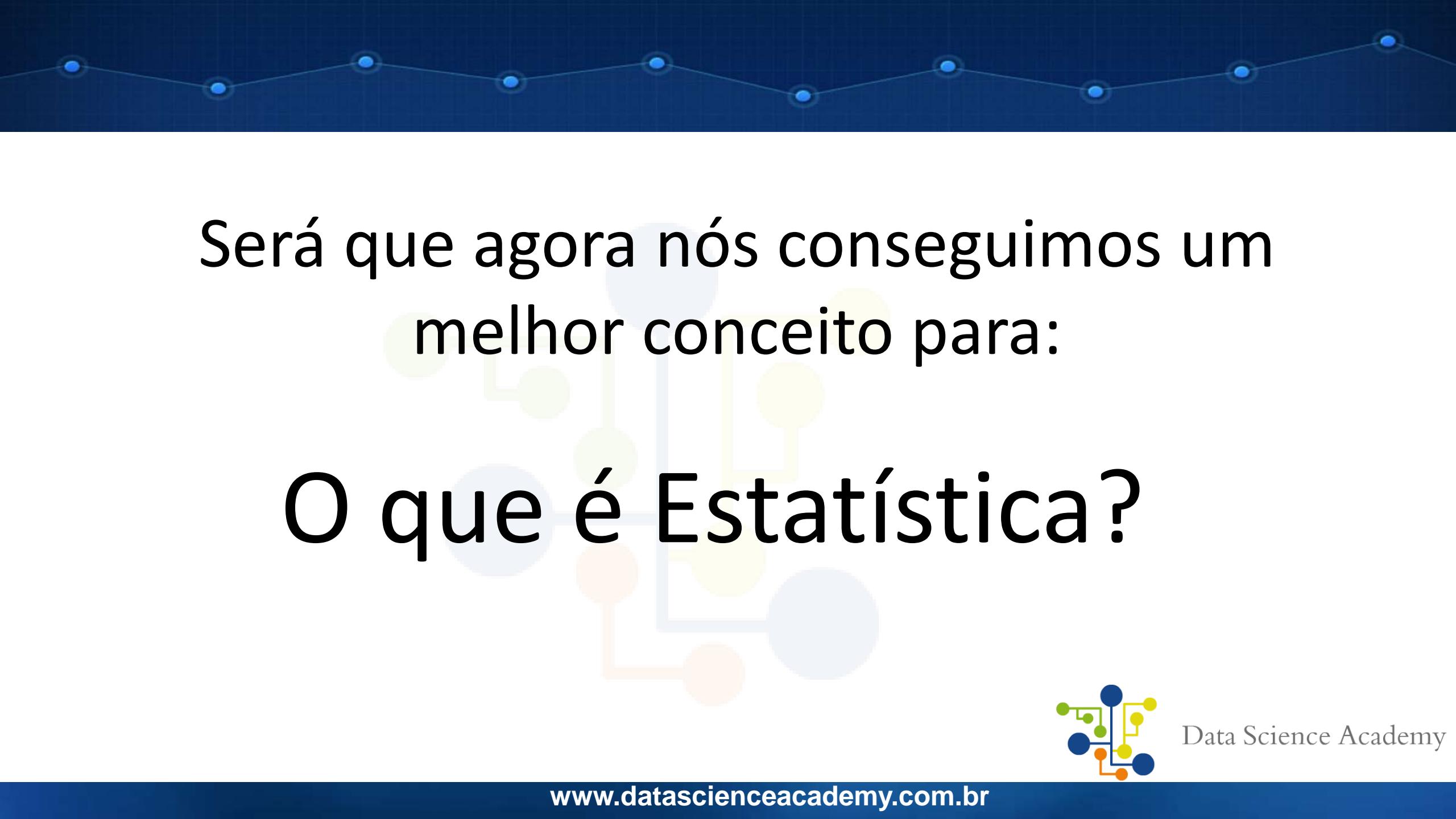


A Estatística atual está diretamente envolvida com a revolução na qual estamos atravessando.

Atualmente essas informações quantitativas são absolutamente imprescindíveis para os cidadãos e organizações em qualquer parte do mundo.



Data Science Academy

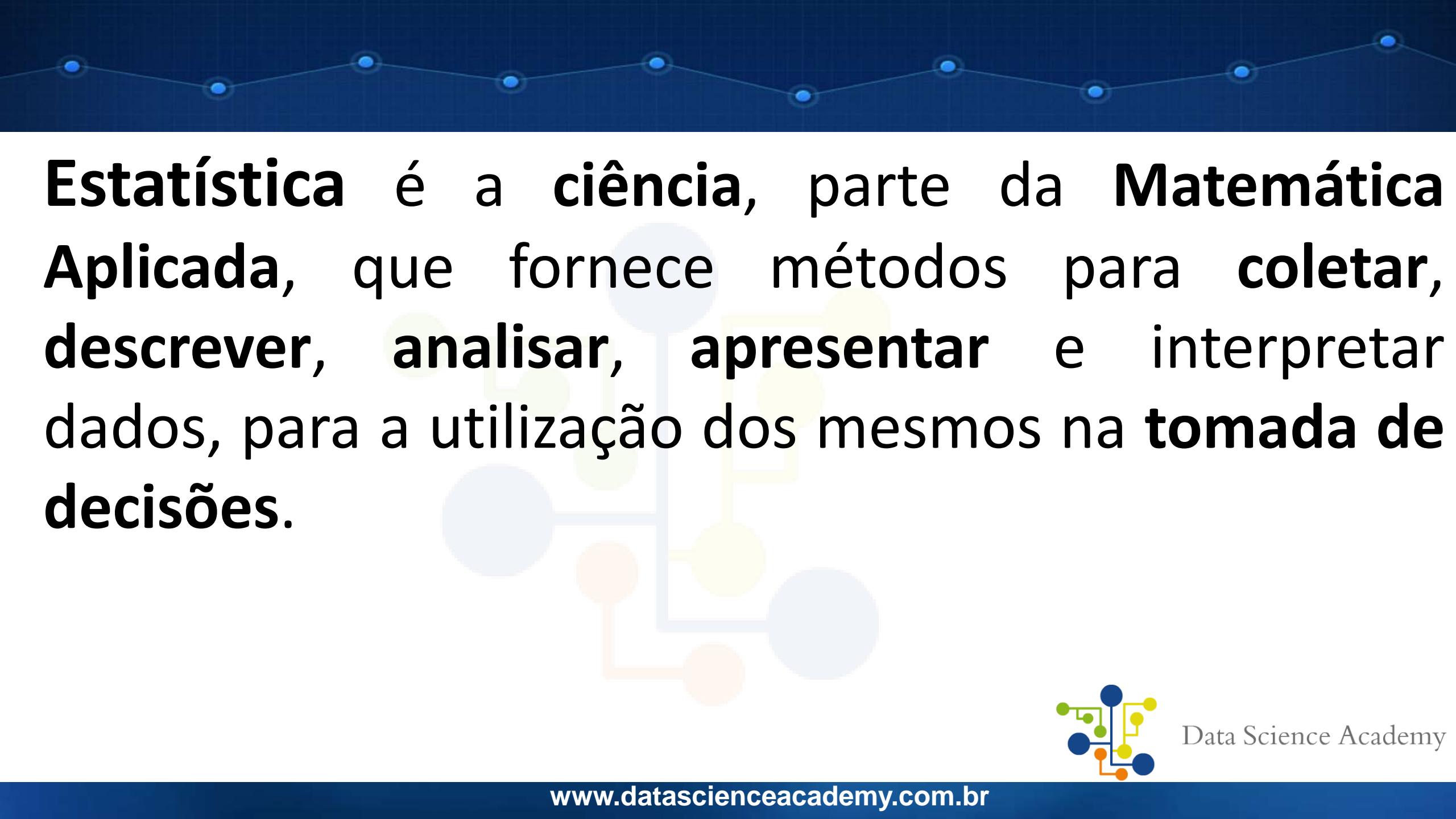


Será que agora nós conseguimos um
melhor conceito para:

O que é Estatística?



Data Science Academy

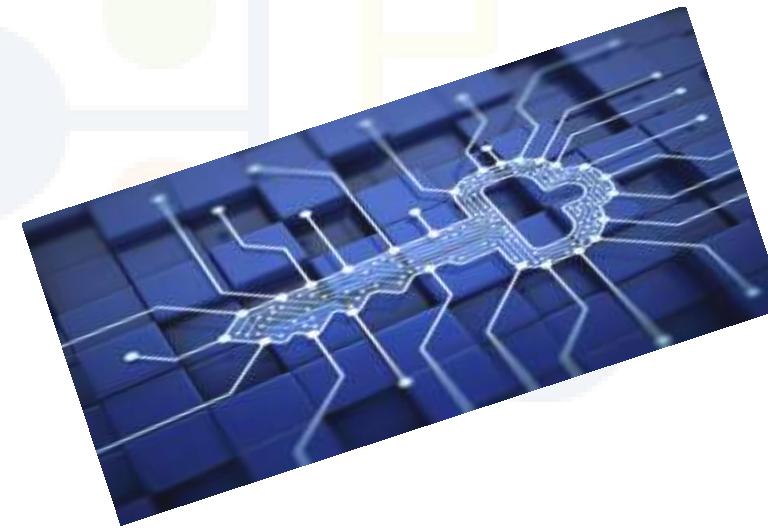


Estatística é a ciência, parte da **Matemática Aplicada**, que fornece métodos para coletar, descrever, analisar, apresentar e interpretar dados, para a utilização dos mesmos na **tomada de decisões**.



Data Science Academy

A **Estatística** está diretamente envolvida com a revolução na qual estamos atravessando e seu conhecimento pode ser a chave que abrirá muitas portas.



Data Science Academy

Big Data Analytics é o termo que se refere a análise estatística de grandes quantidades de dados, para que se possa extrair informação relevante para a compreensão da situação atual e a tomada de decisões.



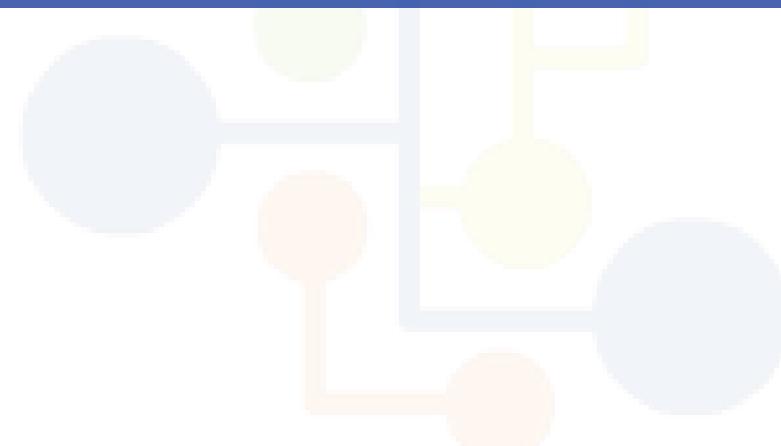
Data Science Academy

Análise de Dados tem se tornado tão importante, que algumas publicações especializadas têm afirmado: “deveríamos inscrever os nossos filhos em cursos de Estatística”.



Data Science Academy

Dicas Importantes



Data Science Academy



Leia o material complementar.



Acesse nosso curso pelo Computador, Tablet ou Smartphone.



Data Science Academy

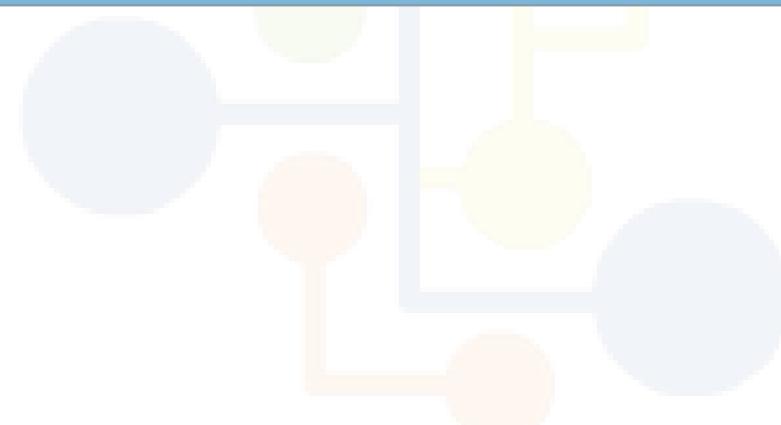
Esse tópico chegou ao final



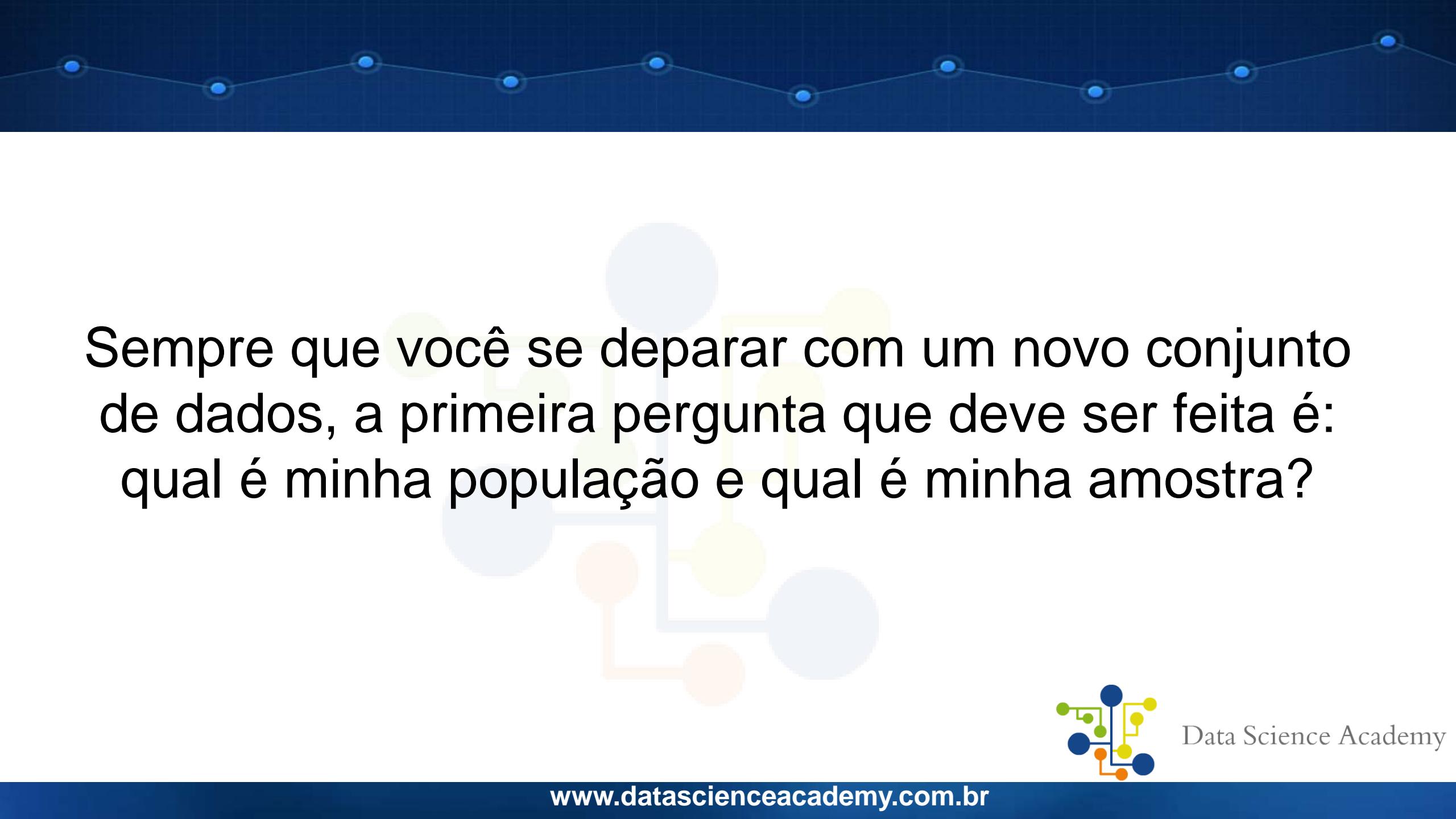
Data Science Academy



População e Amostra



Data Science Academy



Sempre que você se deparar com um novo conjunto de dados, a primeira pergunta que deve ser feita é: qual é minha população e qual é minha amostra?



Data Science Academy



População

- São todos os elementos distintos - indivíduos, itens ou objetos - cujas características estejam sendo estudadas.



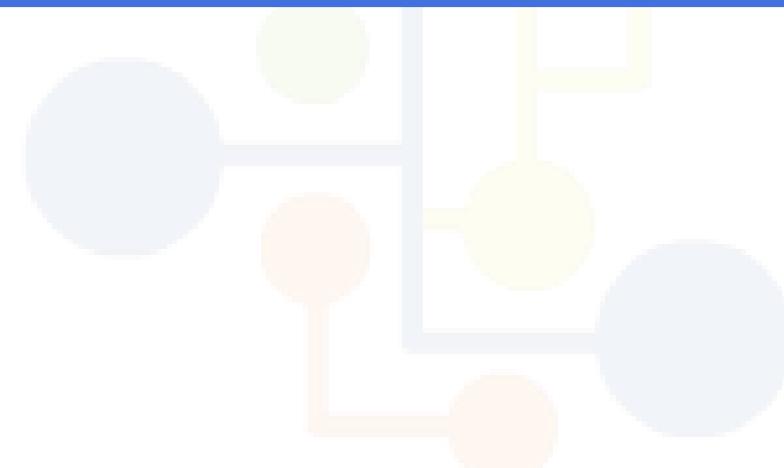
Amostra

- É uma parte da população.
- Ela é coletada a partir da população.



Data Science Academy

Exemplo



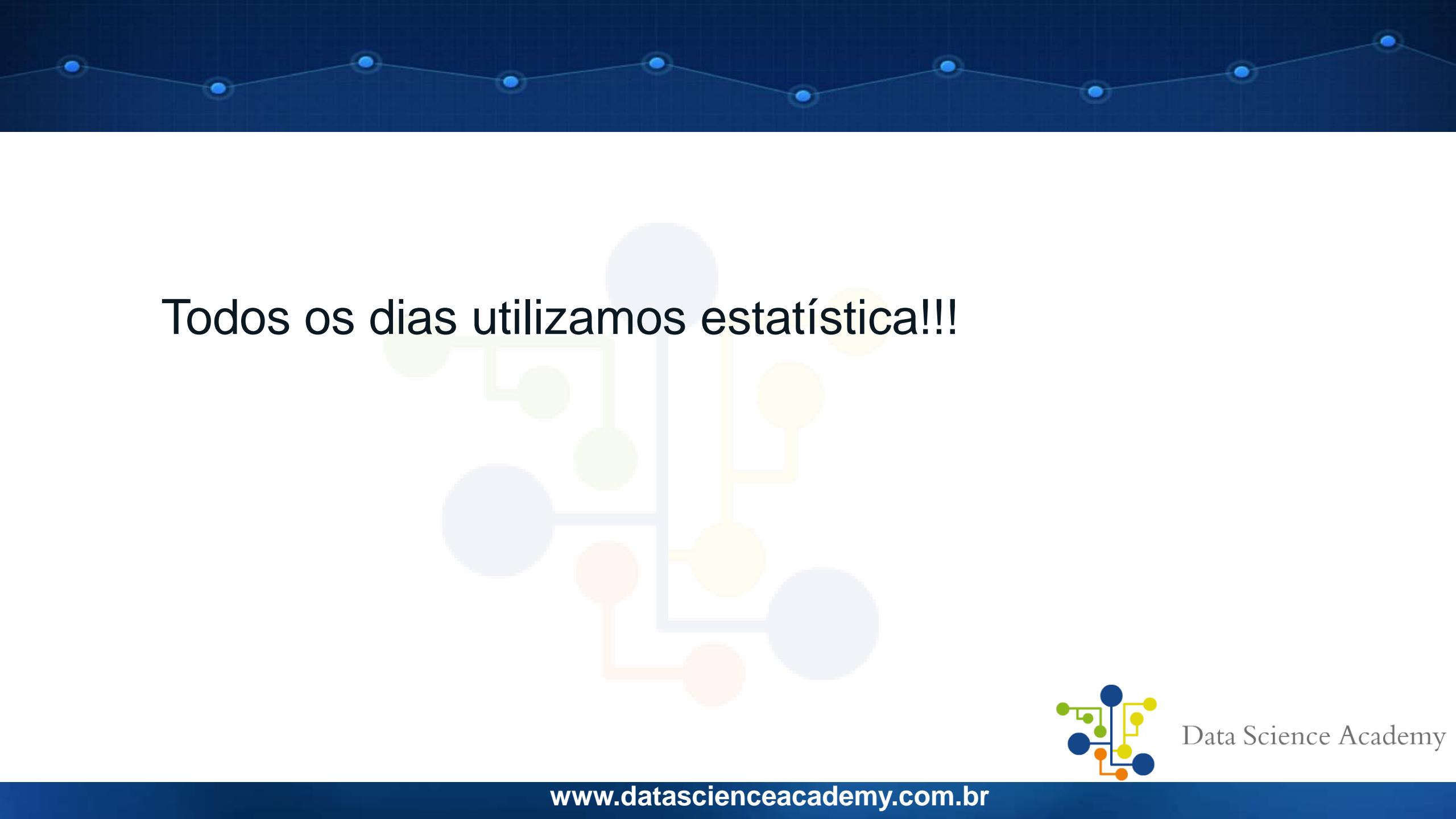
Data Science Academy

Pesquisa Eleitoral

Os institutos de pesquisa, examinam uma amostra e a partir disso, deduzem informações sobre toda a população.



Data Science Academy

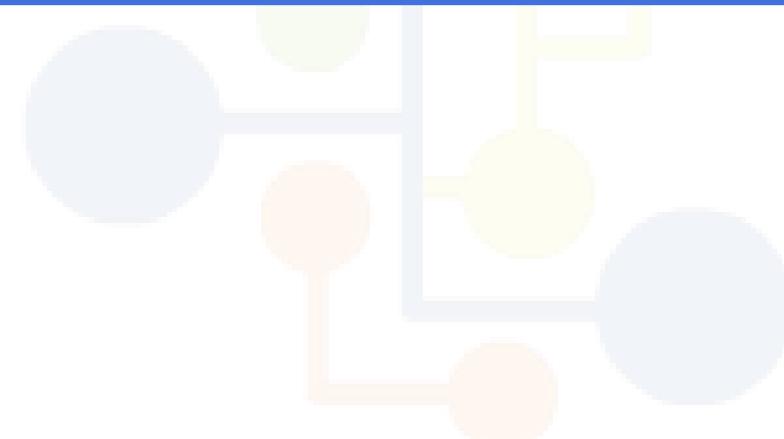


Todos os dias utilizamos estatística!!!



Data Science Academy

Exemplo



Data Science Academy

É Sopa!



Data Science Academy

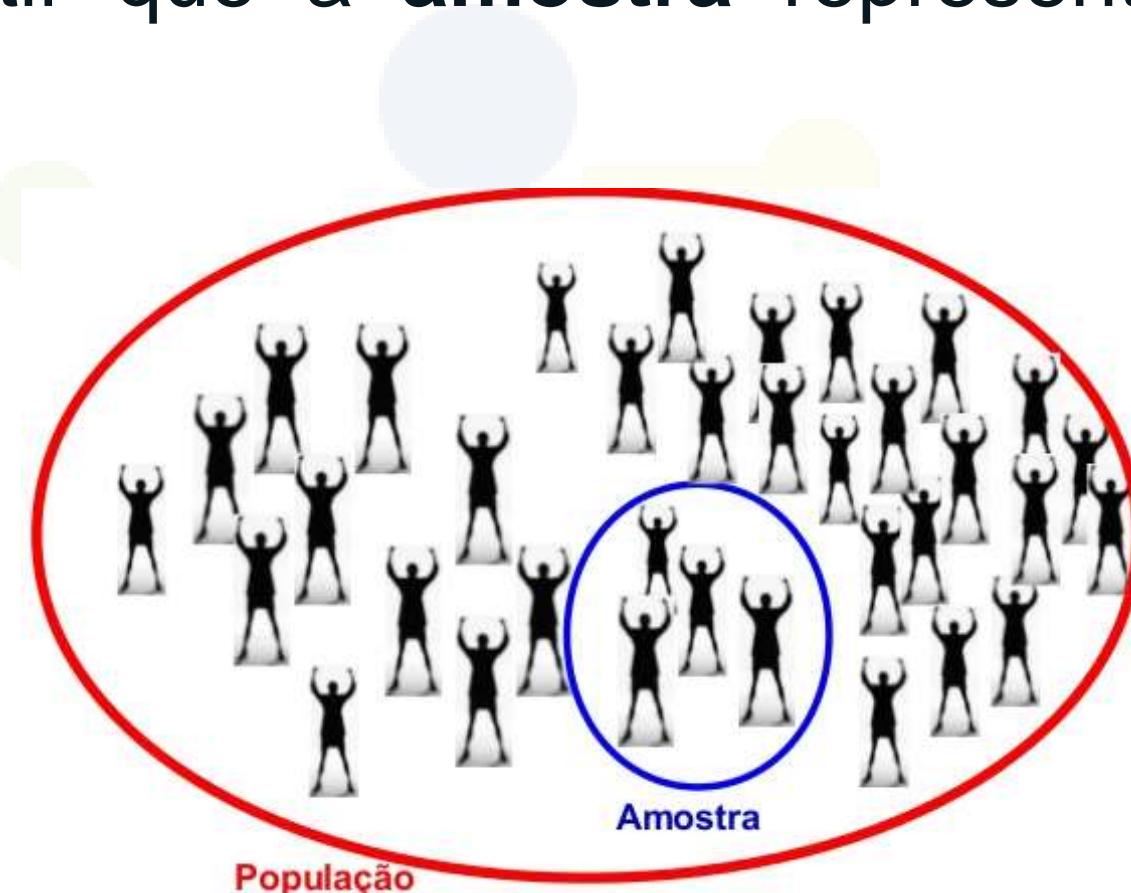


Compreendeu o conceito?



Data Science Academy

Como garantir que a **amostra** representa fielmente a **população**?



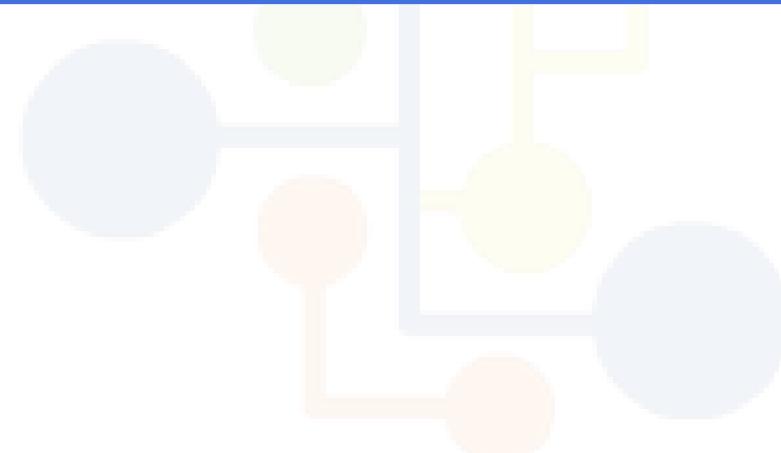
Data Science Academy

É Sopa novamente!



Data Science Academy

Exemplo



Data Science Academy

Randomização

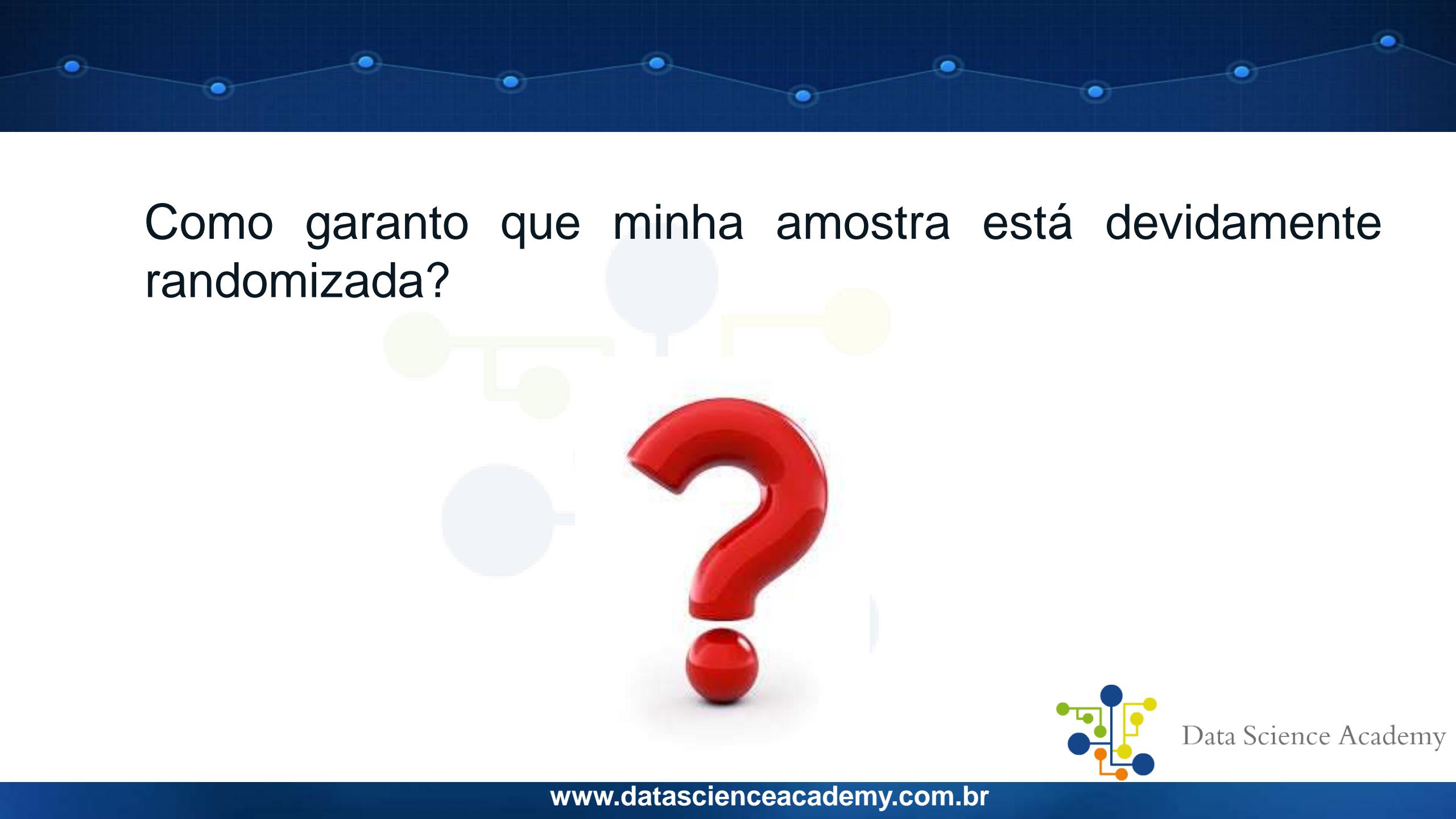


Data Science Academy

Ok, entendi. Mas como eu faço com indivíduos??



Data Science Academy

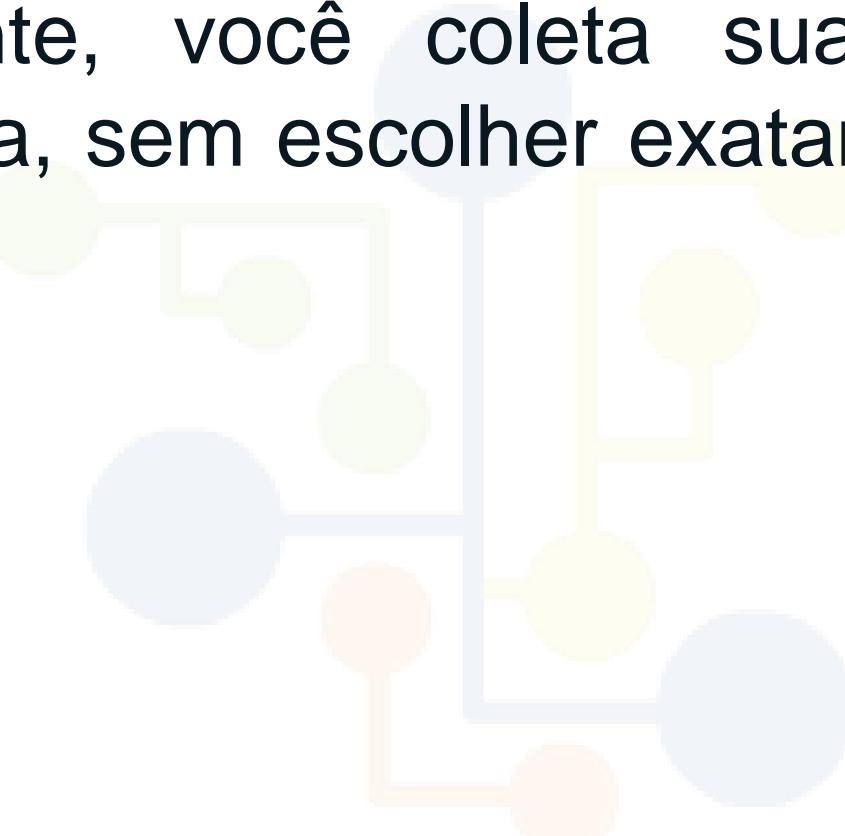


Como garantir que minha amostra está devidamente randomizada?



Data Science Academy

Simplesmente, você coleta sua **amostra** de forma randomizada, sem escolher exatamente quem fará parte da amostra.



Data Science Academy



E qual deve ser o tamanho da **amostra**?



Data Science Academy

Lembra da sopa?



Data Science Academy

Para compreender bem este conceito, precisamos também distinguir entre **parâmetro** e **estatística**:



Data Science Academy

Parâmetro – características sobre a população.

Valores calculados usando dados da população são chamados de **parâmetro**.



Data Science Academy

Estatística – características sobre a amostra.

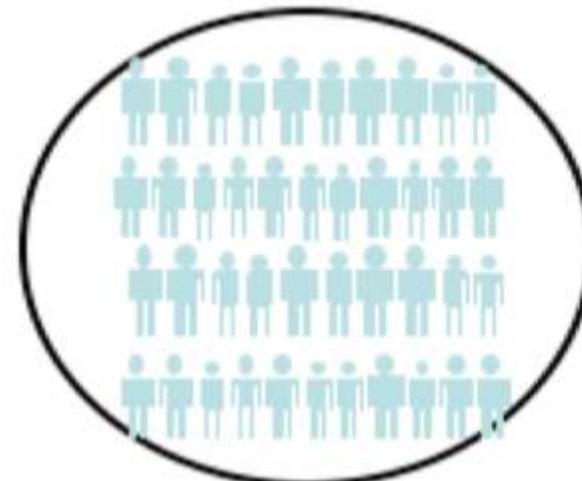
Valores calculados usando dados da amostra são chamados de **estatística**.



Data Science Academy

Estatística Inferencial realiza deduções e conclusões sobre a população, baseados nos resultados obtidos da análise da amostra.

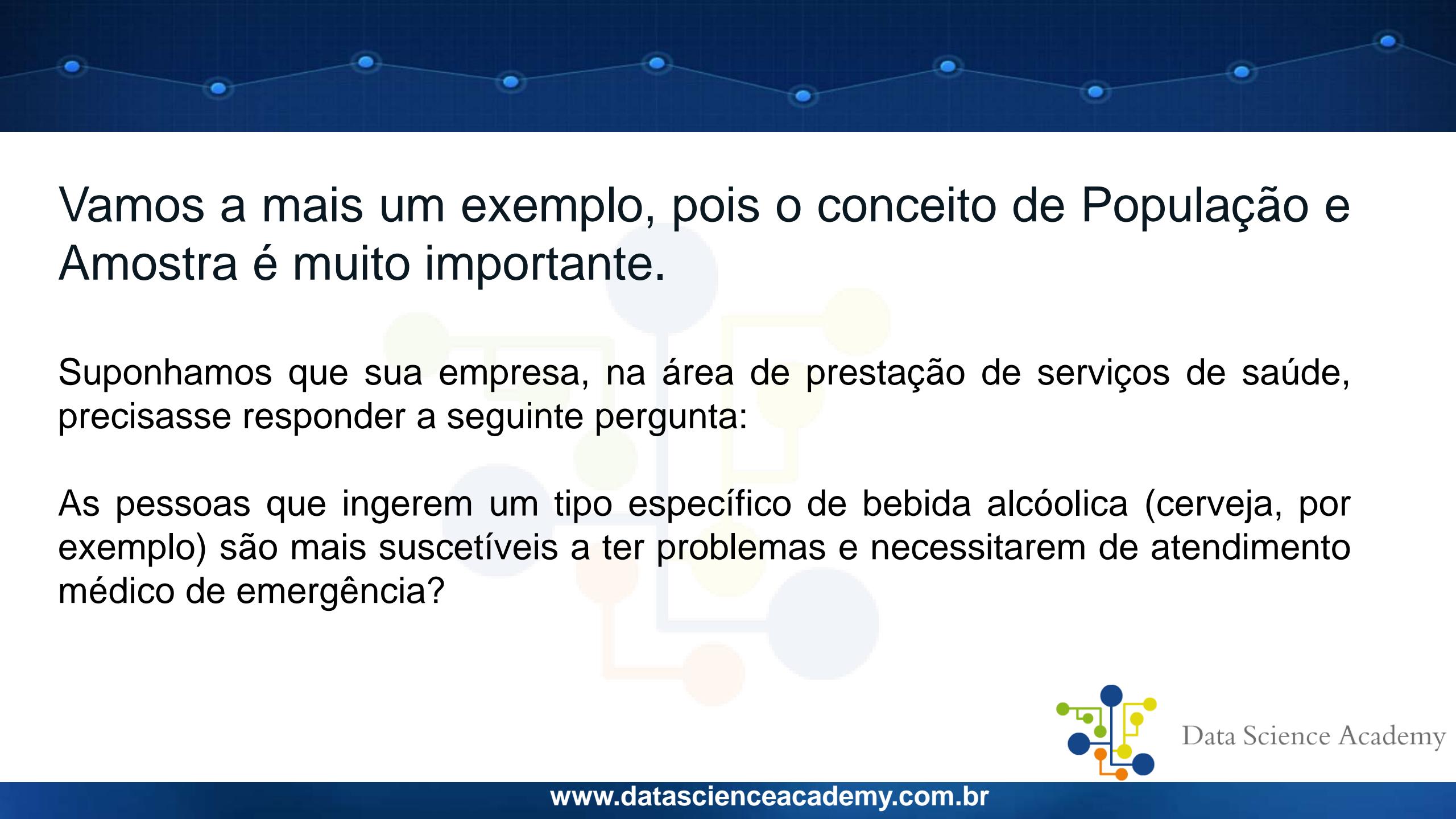
POPULAÇÃO



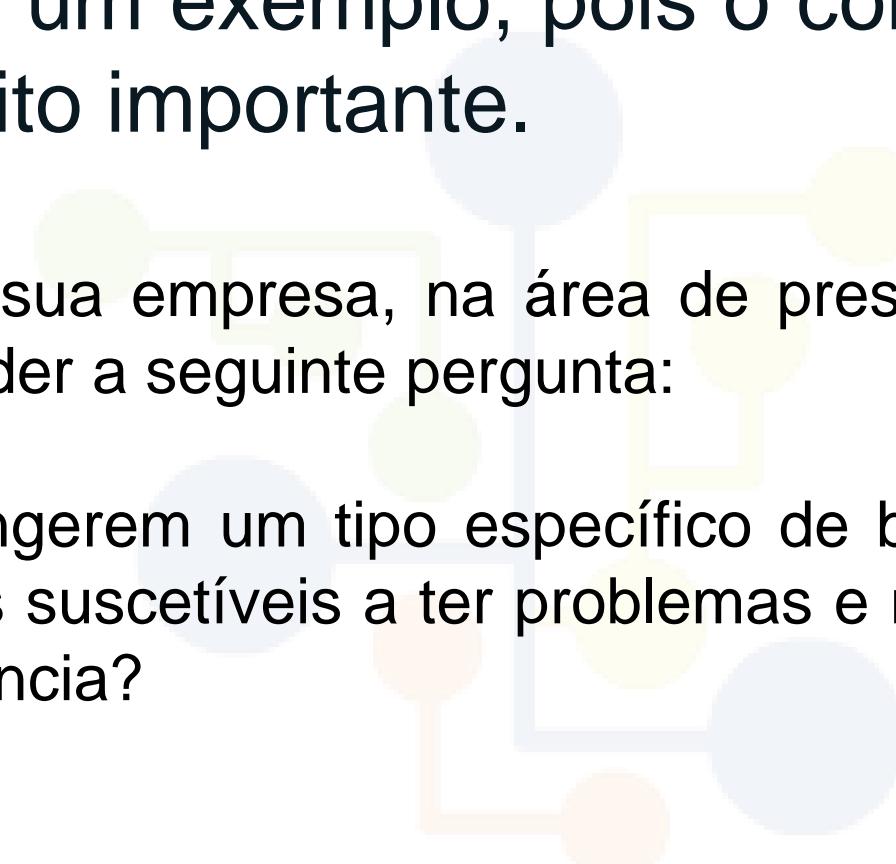
AMOSTRA



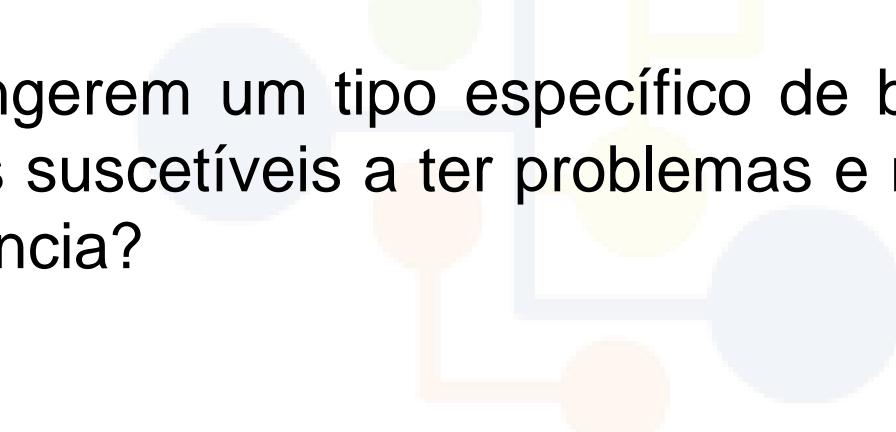
Data Science Academy



Vamos a mais um exemplo, pois o conceito de População e Amostra é muito importante.



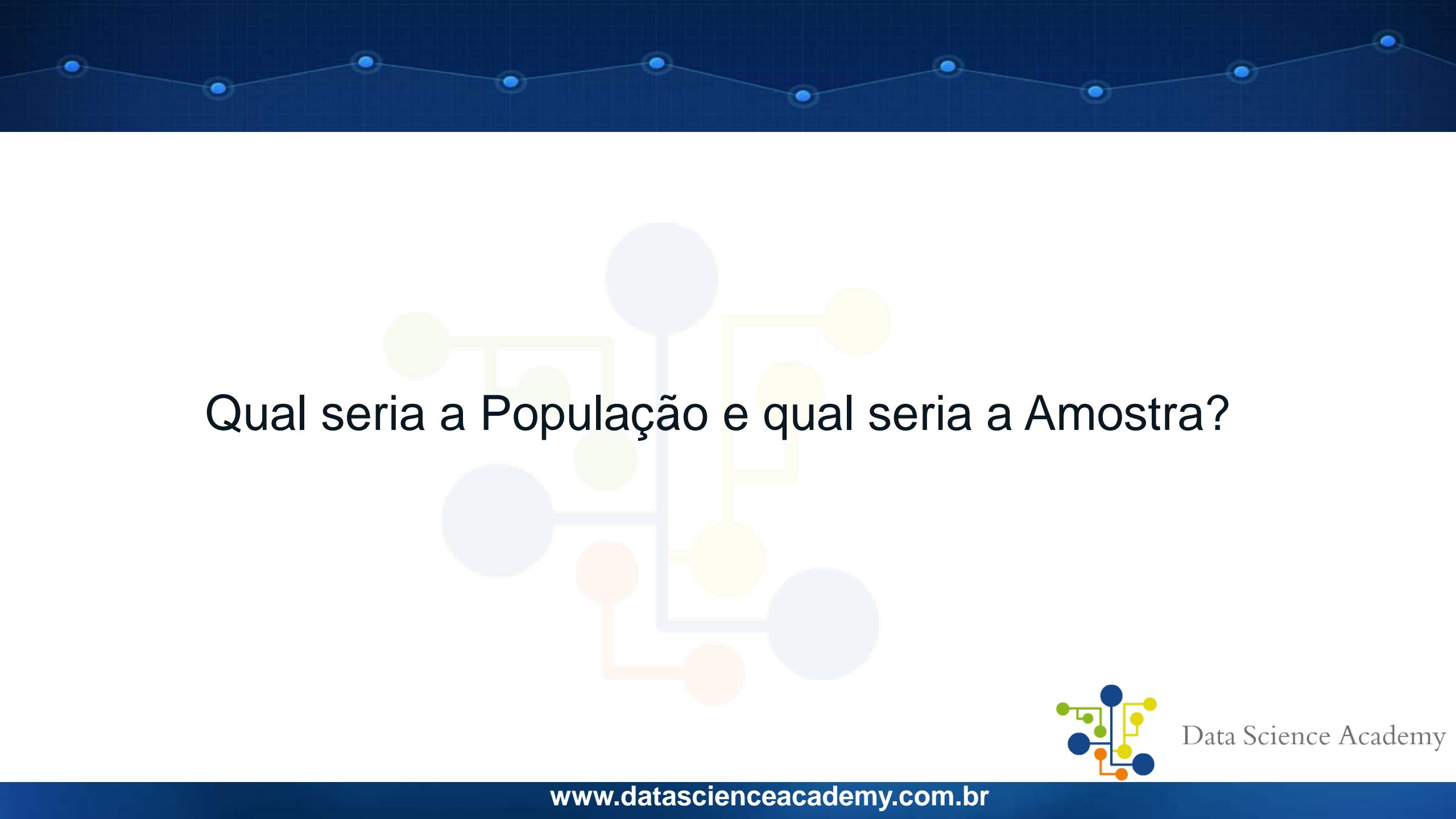
Suponhamos que sua empresa, na área de prestação de serviços de saúde, precisasse responder a seguinte pergunta:



As pessoas que ingerem um tipo específico de bebida alcóolica (cerveja, por exemplo) são mais suscetíveis a ter problemas e necessitarem de atendimento médico de emergência?



Data Science Academy



Qual seria a População e qual seria a Amostra?



Data Science Academy

População – todos os indivíduos
de um país



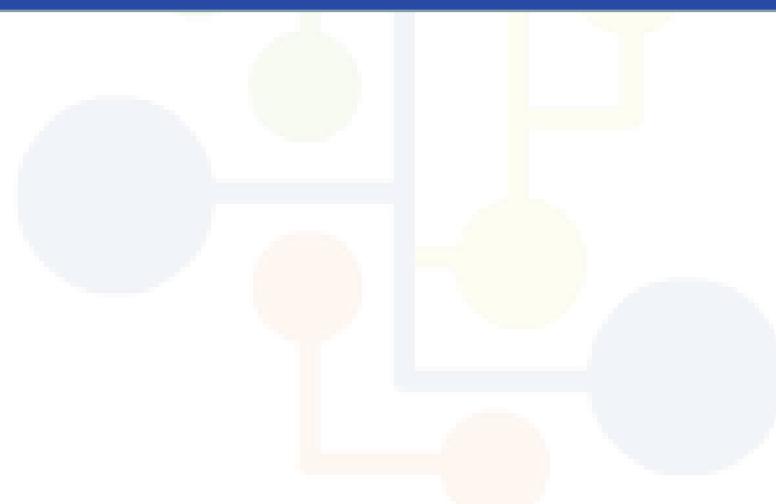
Amostra – pessoas que foram
atendidas em um hospital específico



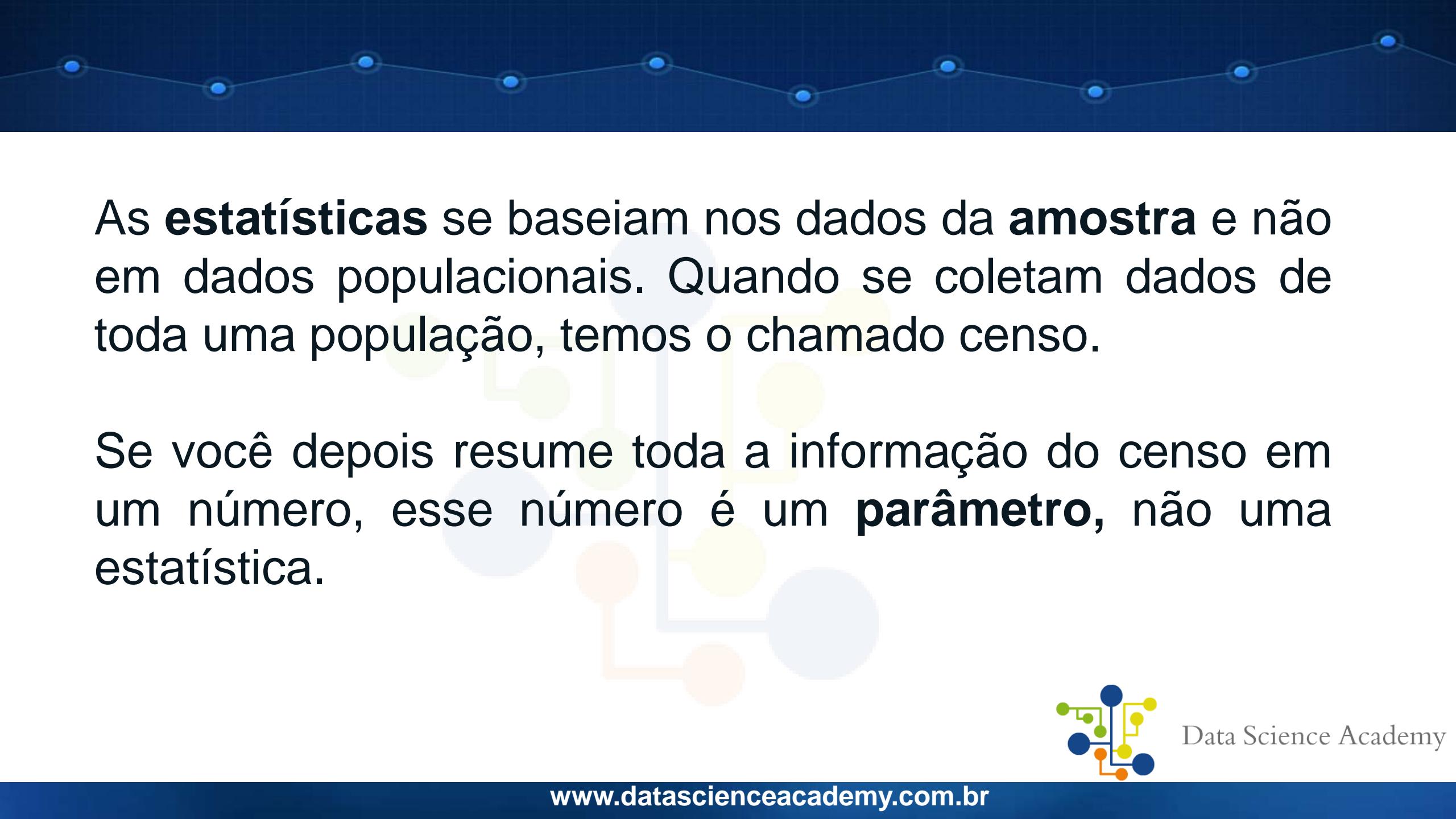
Data Science Academy



Portanto,



Data Science Academy



As **estatísticas** se baseiam nos dados da **amostra** e não em dados populacionais. Quando se coletam dados de toda uma população, temos o chamado censo.

Se você depois resume toda a informação do censo em um número, esse número é um **parâmetro**, não uma estatística.



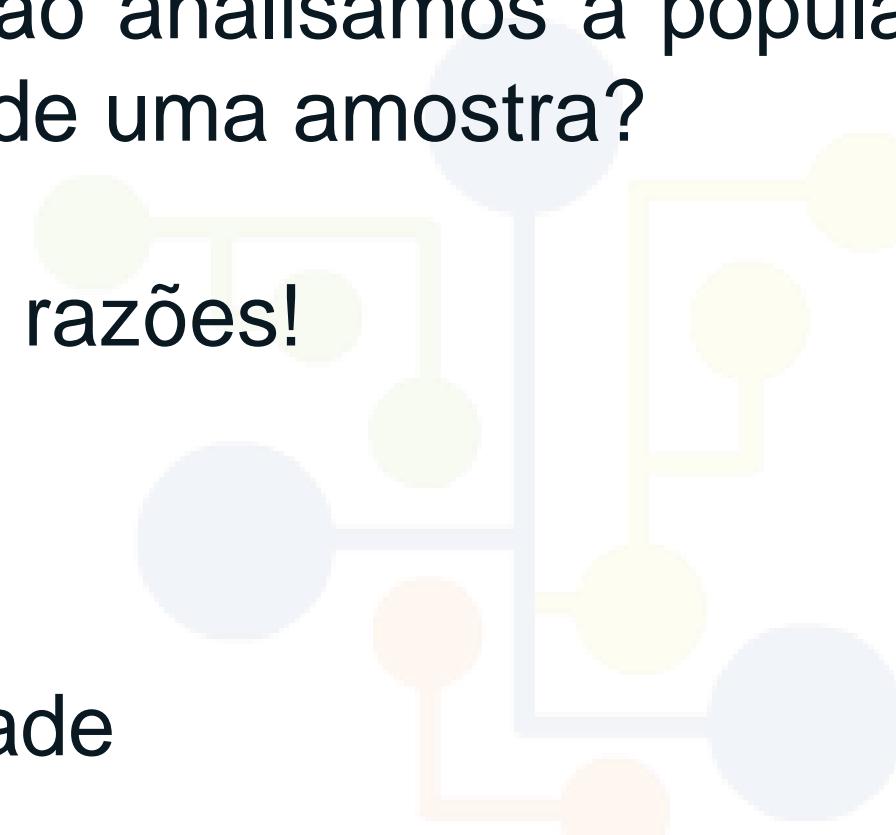
Data Science Academy



E por que não analisamos a população inteira? Por que precisamos de uma amostra?

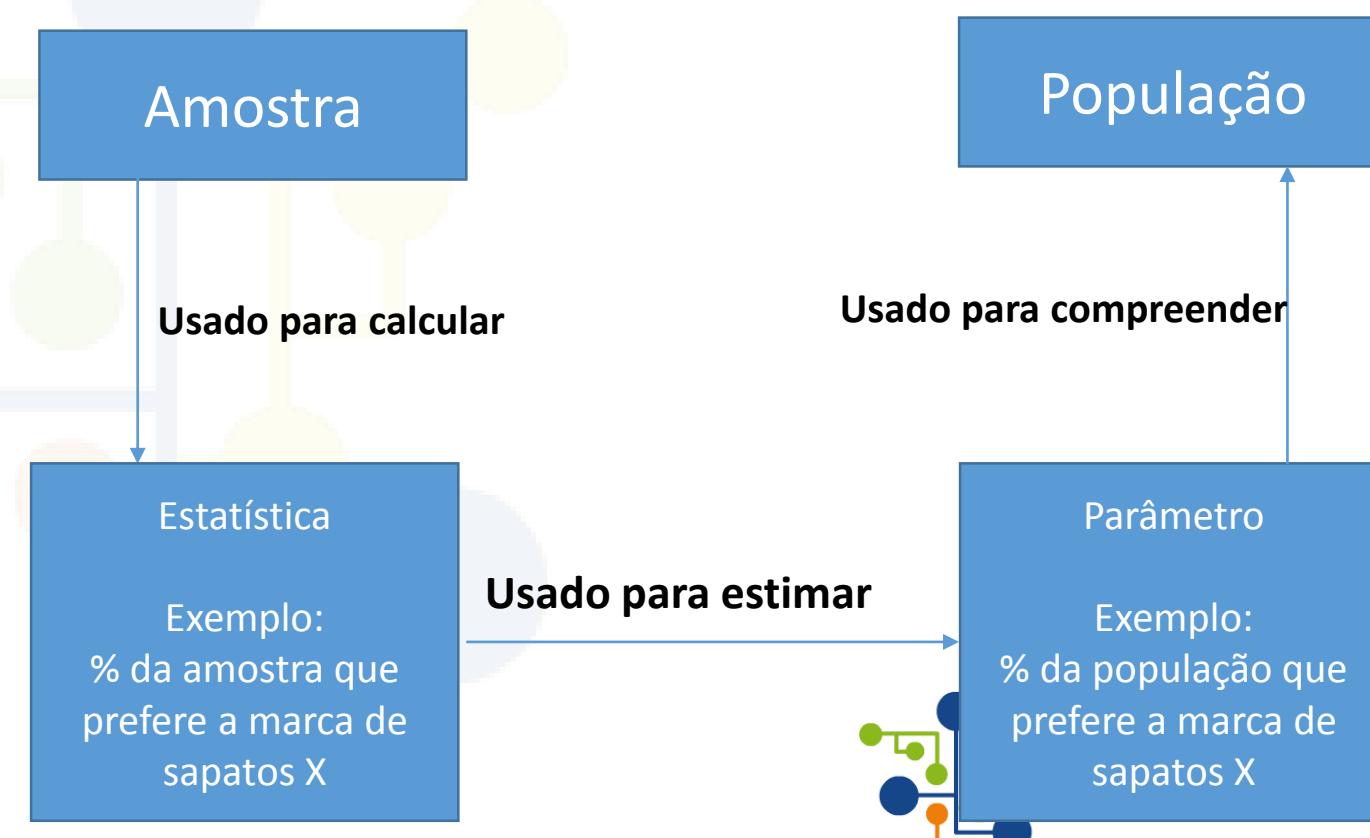
Por diversas razões!

- Custo
- Tempo
- Necessidade



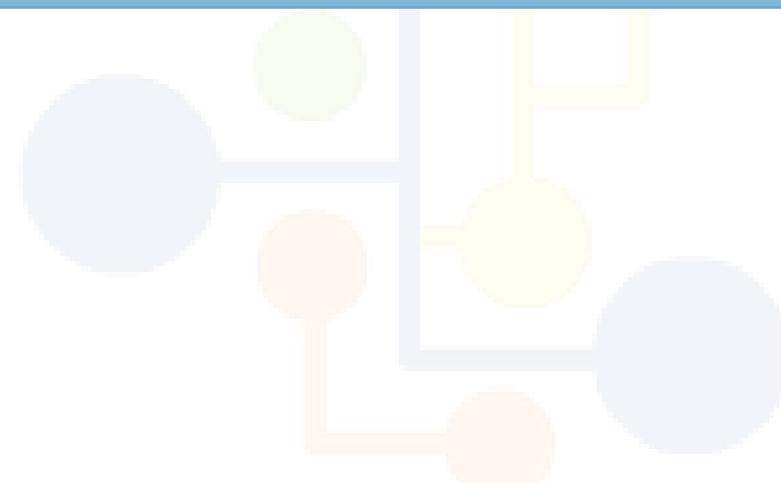
Data Science Academy

A Estatística Inferencial realiza deduções e conclusões sobre a população, baseadas nos resultados obtidos na análise da amostra.





Ética e Estatística



Data Science Academy

As **estatísticas** podem ser utilizadas quando se tenta persuadir um grupo de indivíduos a aceitar um determinado ponto de vista.



Data Science Academy

Isso pode levar a utilização desonesta da **Estatística**.

Ranking de falsidades

Sobre o que as pessoas mais mentem



Homens



Mulheres

Fonte: ParPerfeito



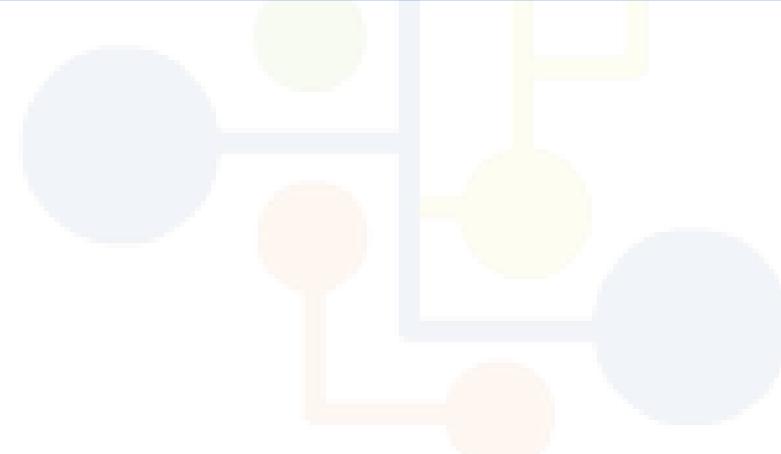
Fonte: Revista Veja
Independente



Data Science Academy



Exemplo I



Data Science Academy

Número de mulheres mortas por aborto clandestino:



Data Science Academy



Estatísticas oficiais divulgadas pelo SUS – DATASUS (portal da saúde):

Números de mulheres que morreram por complicações causadas por abortos provocados (Ano 2013):

Falha de tentativa de aborto – 9 mortes

Aborto NE (não especificado) – 48 mortes

Outros tipos de aborto – 16 mortes

Total: 73 mortes.



Data Science Academy

Exemplo II



Data Science Academy

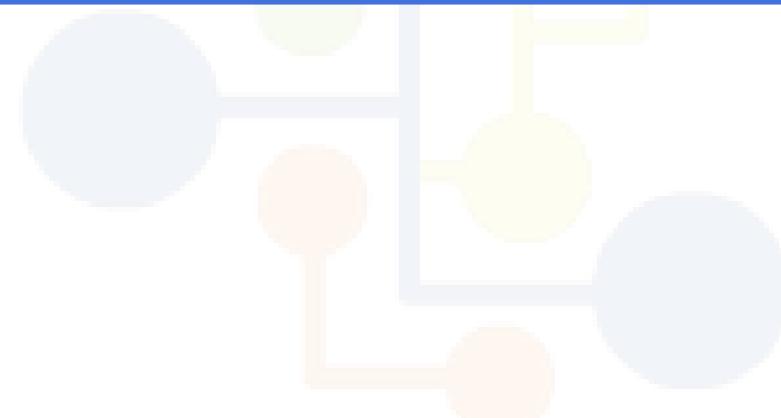
Uma determinada rede de comida vegetariana...

Amostra polarizada: não representa a população



Data Science Academy

Exemplo III



Data Science Academy

Coca-Cola Cherry-Coke.

O Cherry Coke foi um dos maiores fracassos da história da Coca-Cola e a aceitação no mercado foi terrível, levando a empresa a retirar seu produto das prateleiras.



Data Science Academy

Coca-Cola Cherry-Coke.

Ou seja, a amostra estatística não representou a população, levando a conclusões erradas e gerando prejuízo à Coca-Cola.

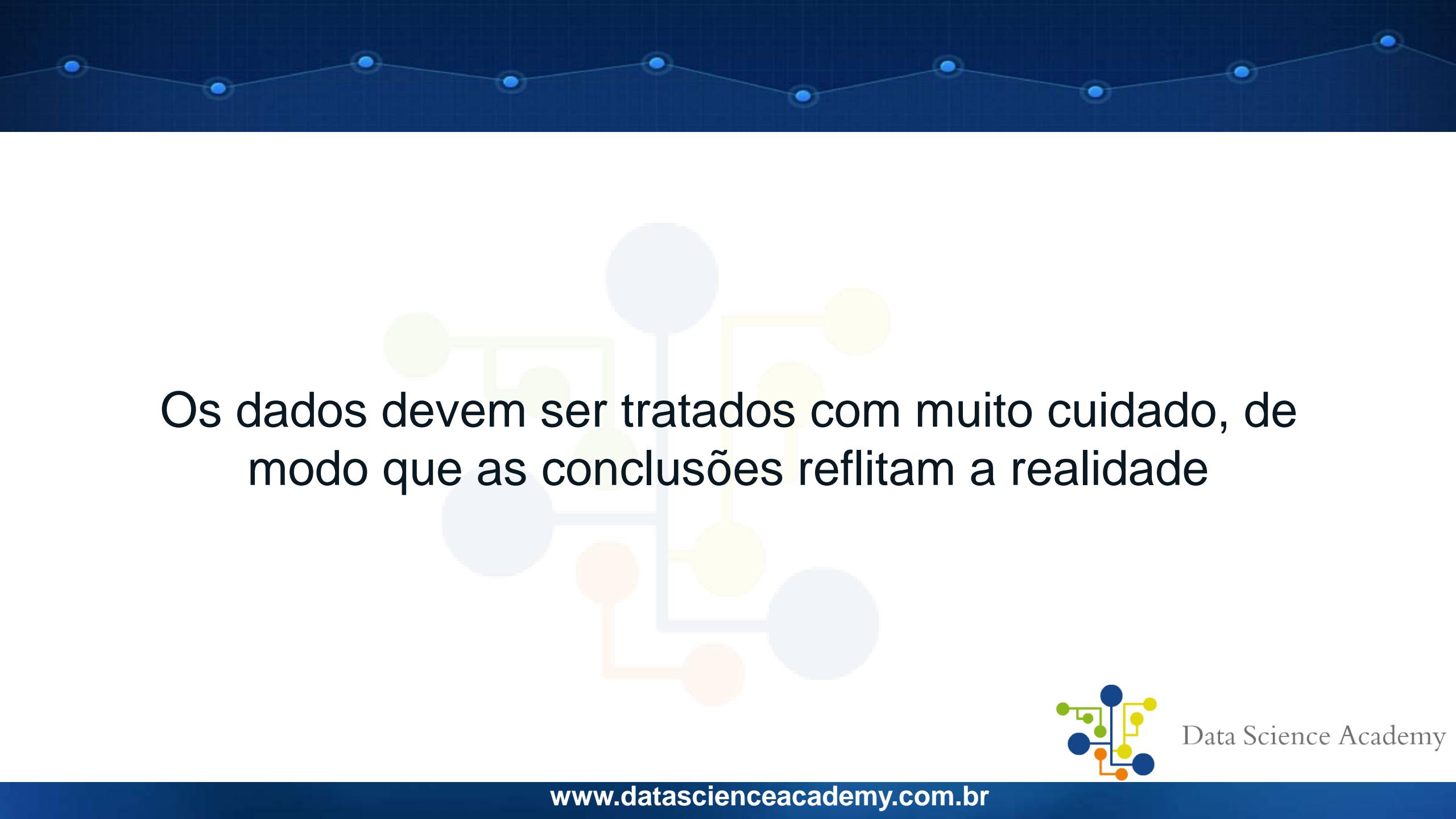


Data Science Academy

Conclusão



Data Science Academy



Os dados devem ser tratados com muito cuidado, de modo que as conclusões refletem a realidade



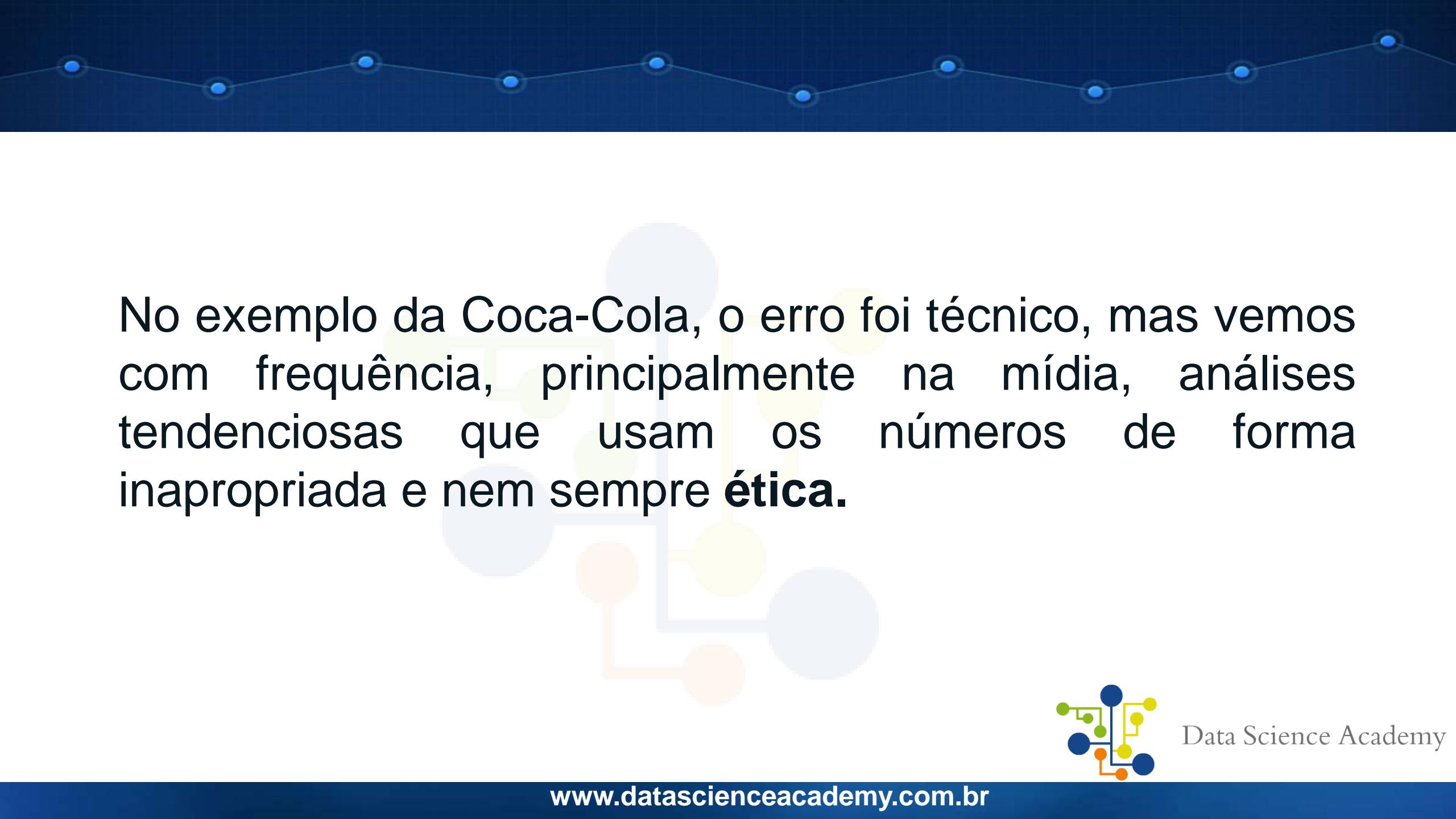
Data Science Academy

Ao escolher uma amostra...

chances iguais a todos os elementos
dados suficientes não distorcidos



Data Science Academy



No exemplo da Coca-Cola, o erro foi técnico, mas vemos com frequência, principalmente na mídia, análises tendenciosas que usam os números de forma inapropriada e nem sempre ética.



Data Science Academy

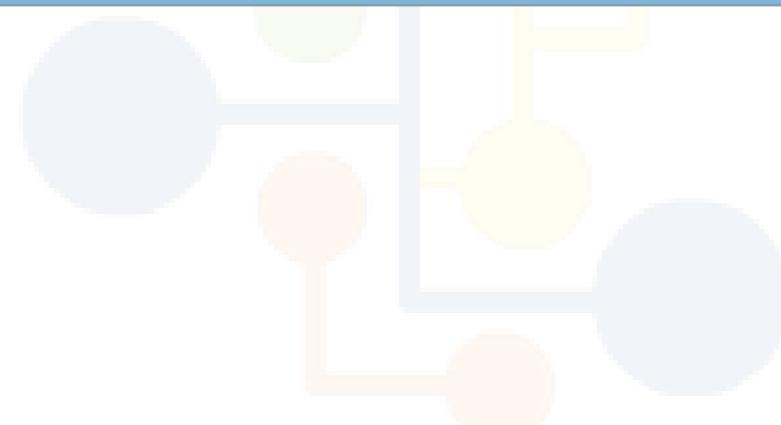
Esse tópico chegou ao final



Data Science Academy



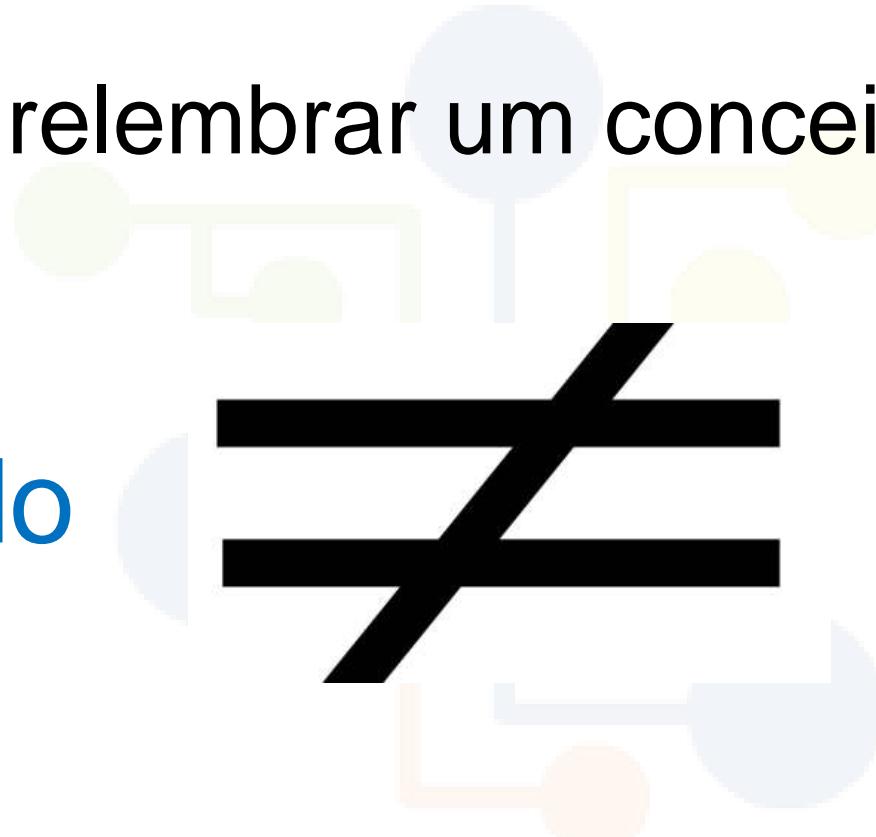
Observações e Variáveis



Data Science Academy

Vamos relembrar um conceito fundamental

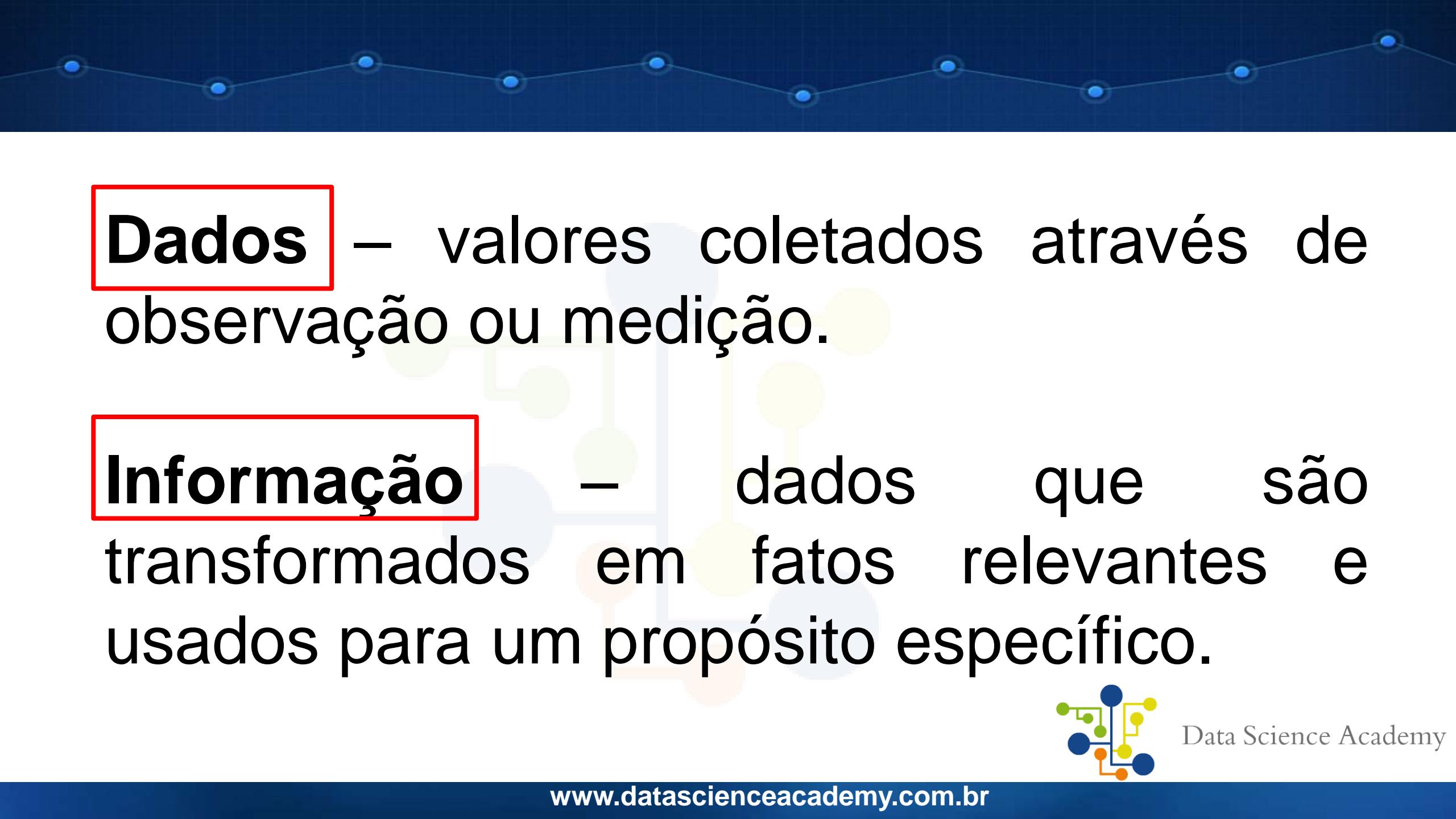
dado



informação



Data Science Academy

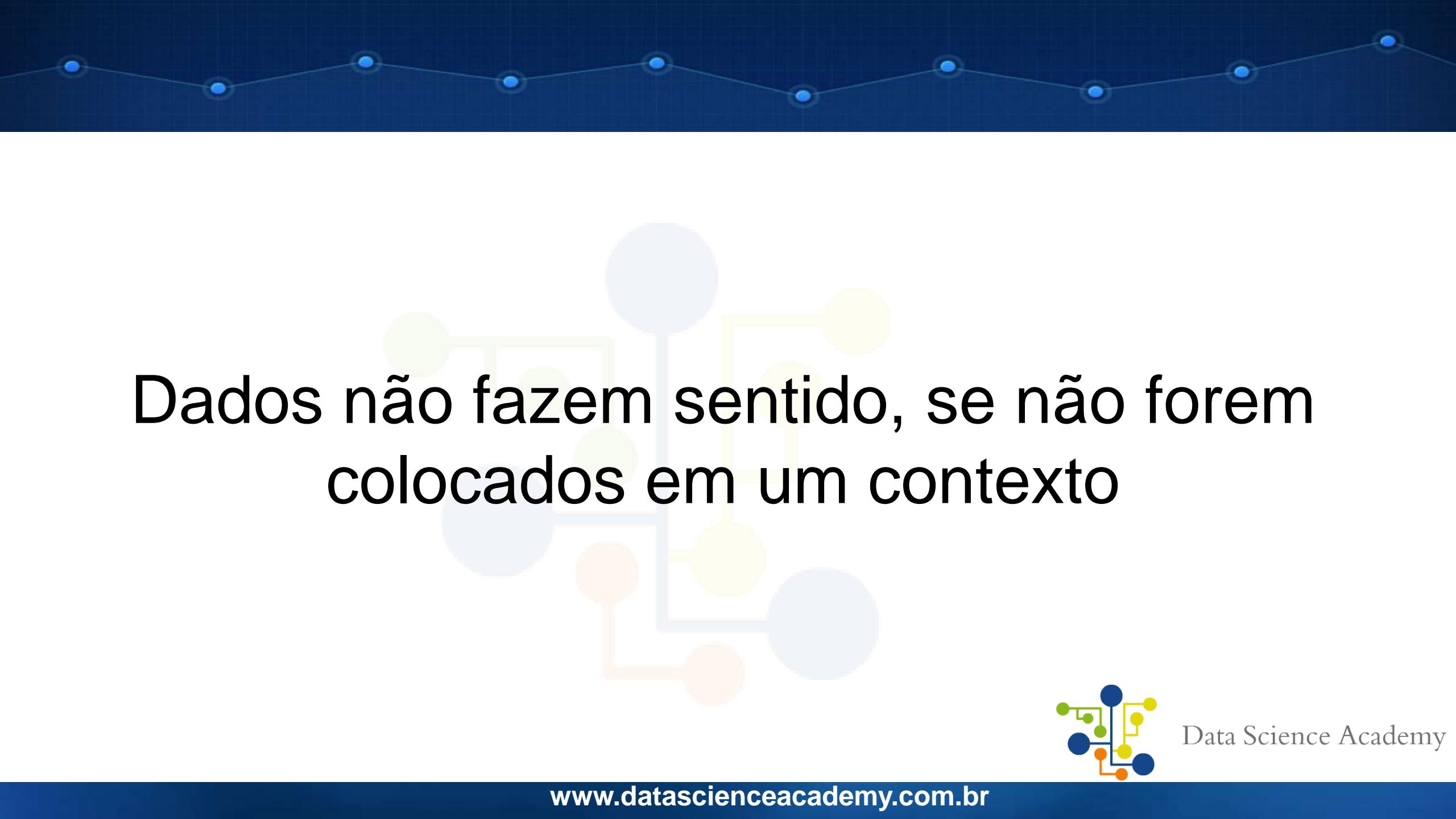


Dados – valores coletados através de observação ou medição.

Informação – dados que são transformados em fatos relevantes e usados para um propósito específico.



Data Science Academy



**Dados não fazem sentido, se não forem
colocados em um contexto**



Data Science Academy

Os dados podem ser obtidos através de duas fontes principais:

Dados Primários

- Coletados por quem faz a análise
- Confiáveis
- Possuem maior controle

Dados Secundários

- Coletados por terceiros
- Não Confiáveis
- Não possuem muito controle



Data Science Academy

Dados Primários

Vantagens

- Confiabilidade
- Qualidade
- Controle das informações
- Acertabilidade nos resultados
- Dados atualizados

- Alto custo
- Demandar tempo maior
- Equipe grande

Desvantagens



Science Academy

Dados Secundários

Vantagens

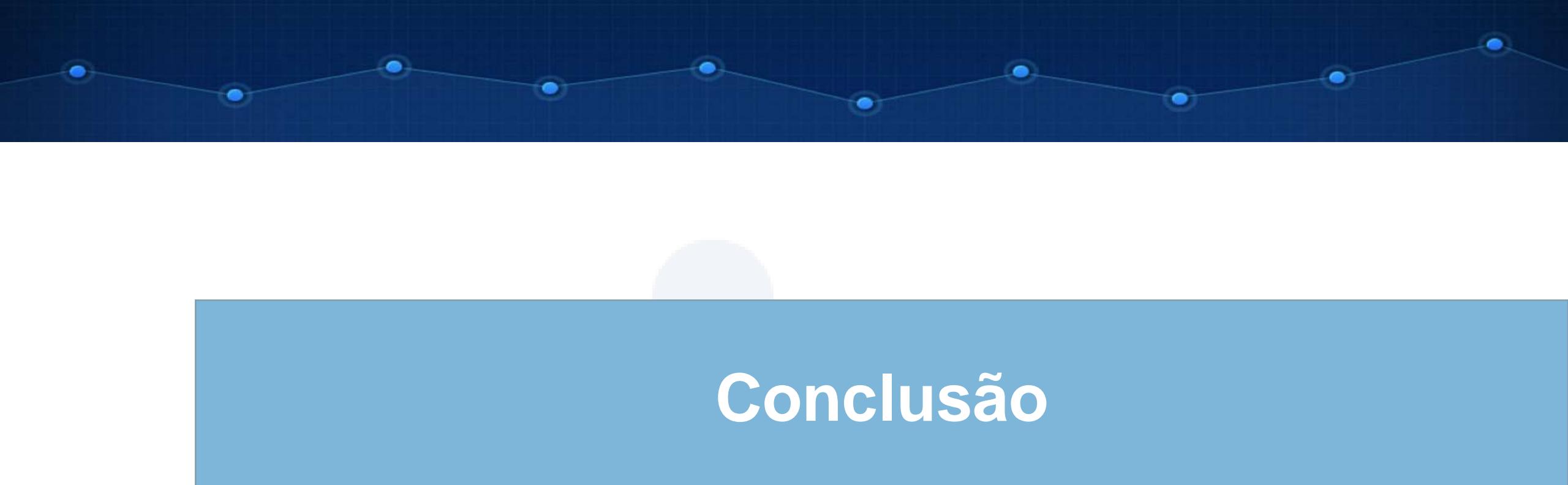
- Baixo custo
- Rapidez
- Existência de diversas fontes
- Diversidade de informações para quantificação de questões

- Falta de controle
- Dados Inadequados
- Diversidade na classificação dos dados
- Dados desatualizados
- Fontes não confiáveis
- Dificuldade de reproduzir um estudo obtendo os mesmos resultados

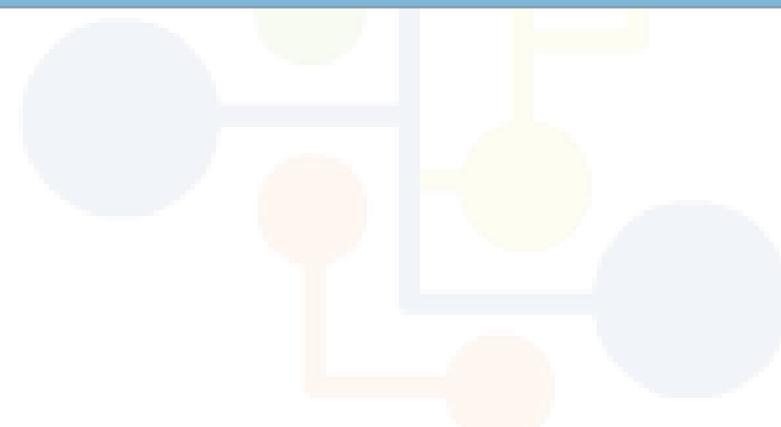
Desvantagens



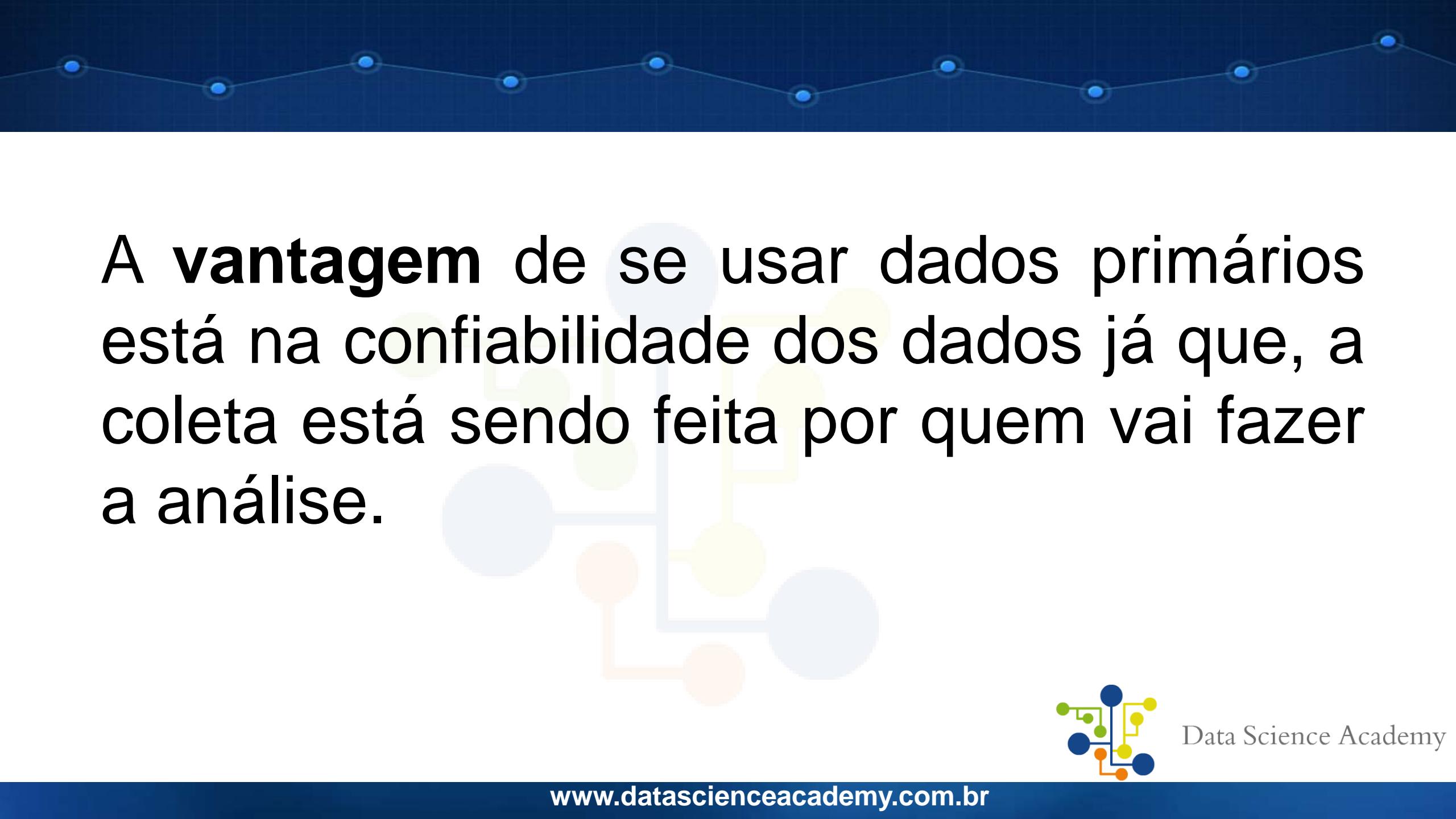
a Science Academy



Conclusão



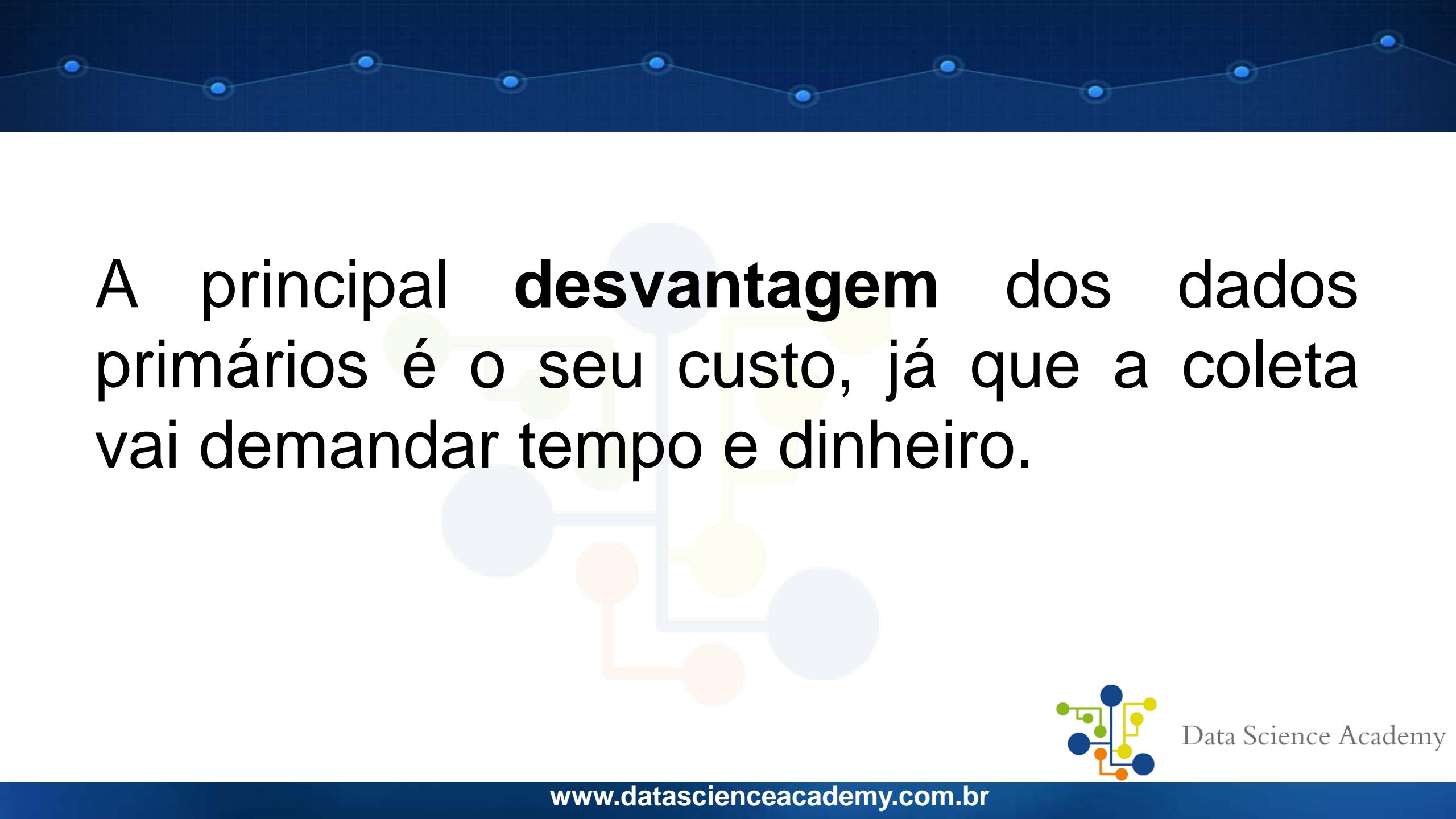
Data Science Academy



A **vantagem** de se usar dados primários está na confiabilidade dos dados já que, a coleta está sendo feita por quem vai fazer a análise.



Data Science Academy



A principal **desvantagem** dos dados primários é o seu custo, já que a coleta vai demandar tempo e dinheiro.



Data Science Academy

Trabalhar com fontes de dados secundários, pode ser mais vantajoso financeiramente, já que os dados já foram coletados, mas a confiança nos dados pode ser comprometida, uma vez que não existe controle na coleta dos dados.



Data Science Academy

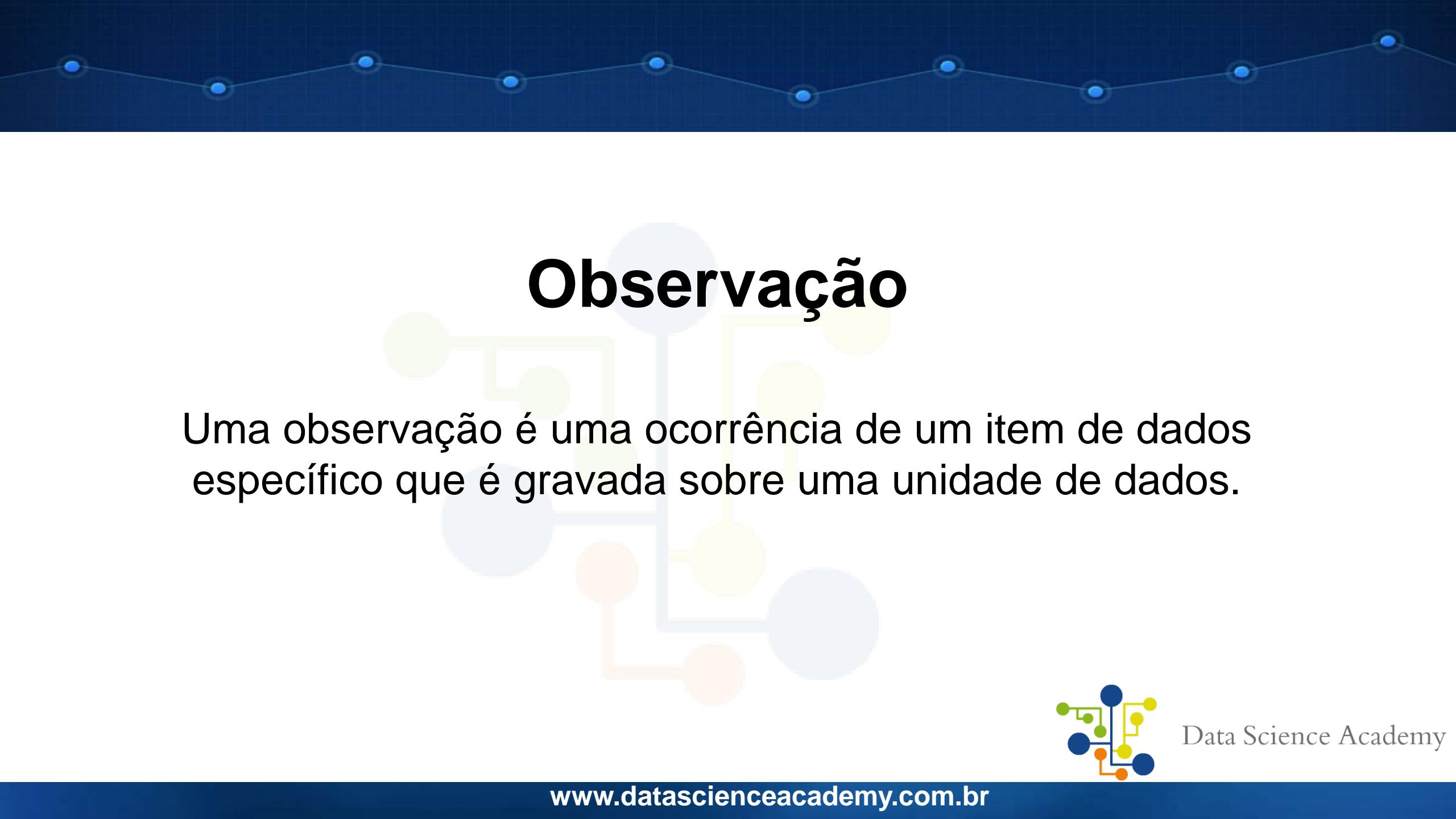


Informação → Conhecimento



Data Science Academy

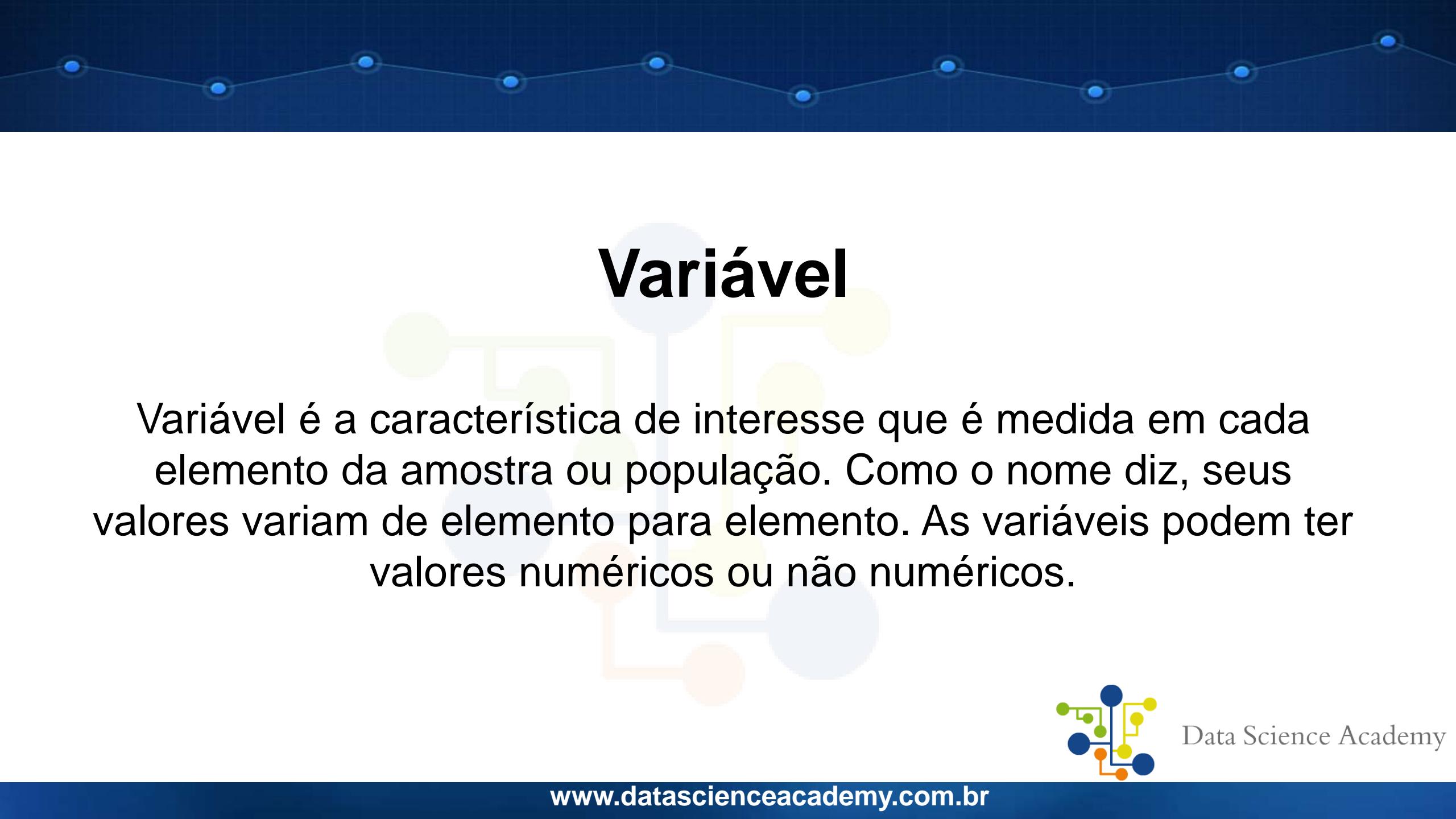
Observação



Uma observação é uma ocorrência de um item de dados específico que é gravada sobre uma unidade de dados.



Data Science Academy



Variável

Variável é a característica de interesse que é medida em cada elemento da amostra ou população. Como o nome diz, seus valores variam de elemento para elemento. As variáveis podem ter valores numéricos ou não numéricos.



Data Science Academy

Observações

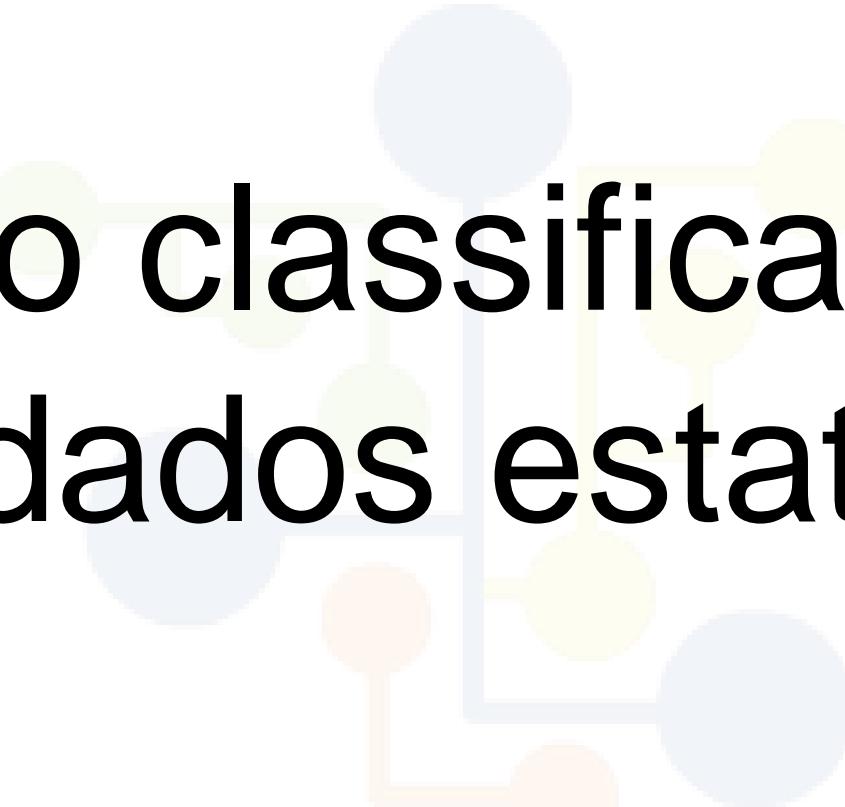
	Idade	Sexo	Peso	Cor dos olhos
Indivíduo 1	42	M	59	Verde
Indivíduo 2	34	M	54	Castanho
Indivíduo 3	56	F	89	Azul
Indivíduo 4	41	M	76	Castanho
Indivíduo 5	23	F	65	Castanho

Variáveis





Como classificar os tipos de dados estatísticos?



Data Science Academy

Os dados podem conter variáveis:

Qualitativas – utilizam termos **descritivos** para descrever algo de interesse. Ex: cor dos olhos, estado civil, religião, sexo, grau de escolaridade, classe social, tipo sanguíneo, cor da pele.



Data Science Academy

Os dados podem conter variáveis:

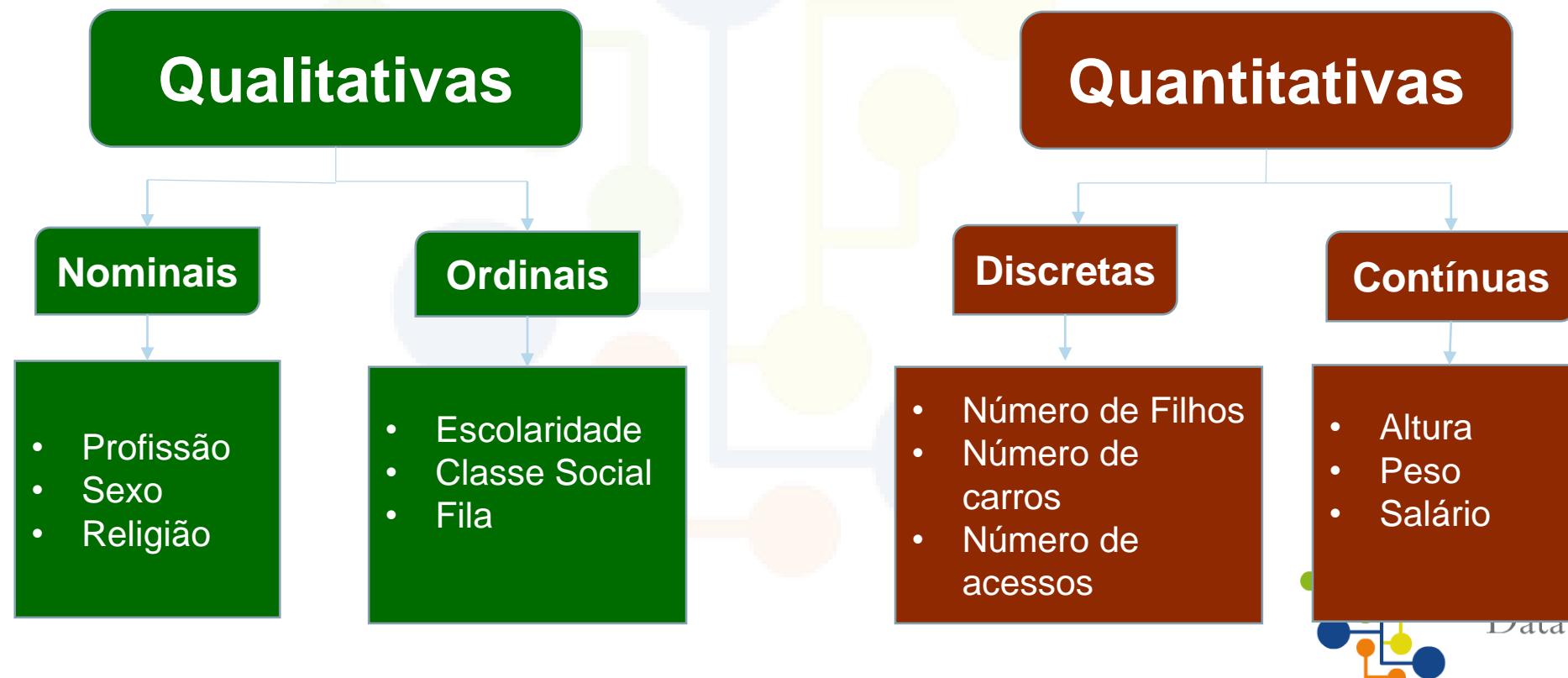
Quantitativas – representados por valores numéricos que podem ser **contados ou medidos**. Ex: número de crianças em uma sala de aula, peso do corpo humano, idade, número de filhos, etc.



Academy



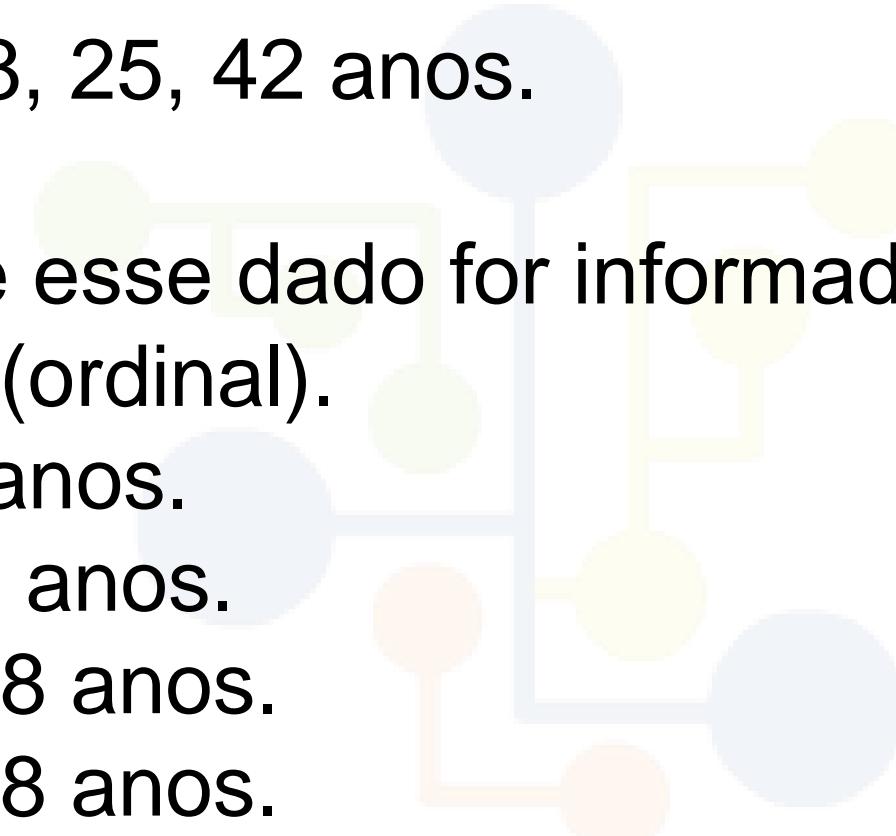
Dentro desta classificação, podemos ter variáveis:





Um dado classificado como "idade" é **quantitativo**

Ex.: 11, 15, 18, 25, 42 anos.



Entretanto, se esse dado for informado por "faixa etária" ele é **qualitativo** (ordinal).

Ex: 0 – 5 anos.

6 – 12 anos.

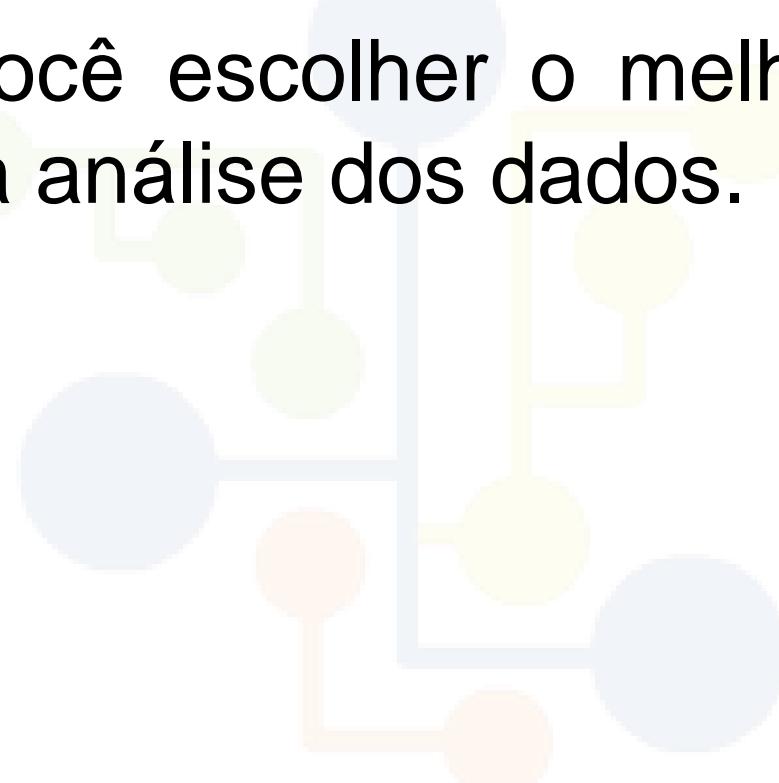
13 – 18 anos.

19 – 28 anos.



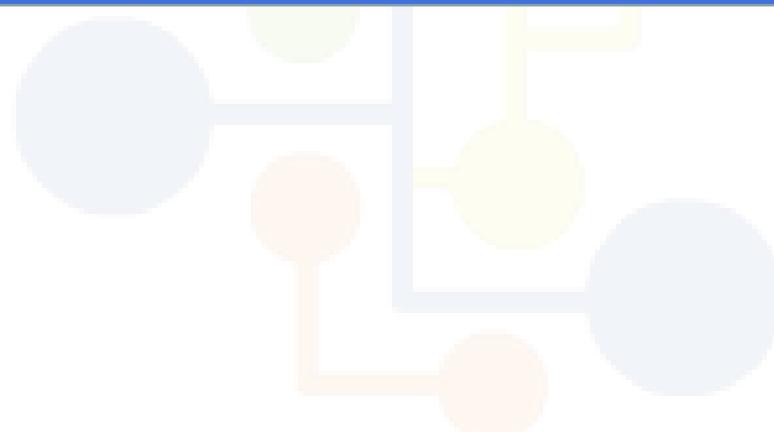
Data Science Academy

É muito importante classificar os dados, pois eles permitirão a você escolher o melhor teste estatístico a ser utilizado na análise dos dados.



Data Science Academy

Exemplo



Data Science Academy

Nível de exercício	Peso (em gramas)
nenhum	3242,82
mudando	3547,59
mudando	3929,22
nenhum	2765,92
baixo/moderado	3134,82
mudando	2693,38
mudando	3144,96
nenhum	3508,47
alto	3728,29
nenhum	4012,09
nenhum	3973,98
mudando	3342,50
mudando	3278,79
mudando	3369,27
baixo/moderado	3583,00
nenhum	2323,93



Data Science Academy

Nível de exercício	Peso (em gramas)
nenhum	3242,82
mudando	3547,59
mudando	3929,22
nenhum	2765,92
baixo/moderado	3134,82
mudando	2693,38
mudando	3144,96
nenhum	3508,47
alto	3728,29
nenhum	4012,09
nenhum	3973,98
mudando	3342,50
mudando	3278,79
mudando	3369,27
baixo/moderado	3583,00
nenhum	2323,93

Compreensão do Problema

- Quanto uma grávida costuma se exercitar?
- O grau do exercício influencia o peso do bebê?
- Que pesos são mais comuns para os bebês?



Data Science Academy

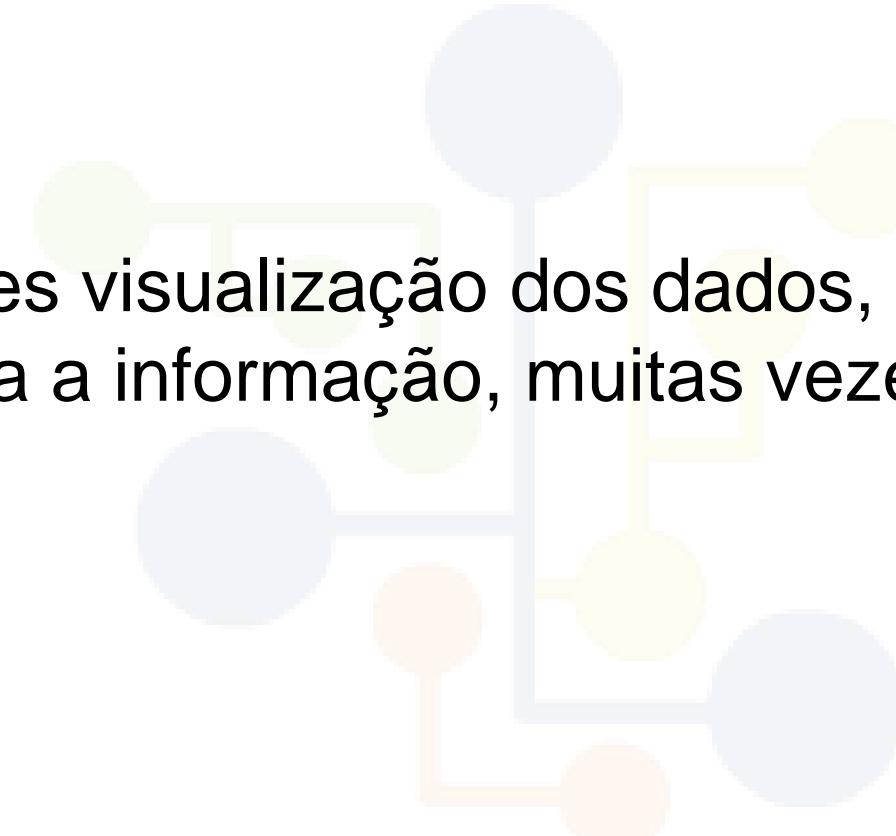
Nível de exercício	Peso (em gramas)
nenhum	3242,82
mudando	3547,59
mudando	3929,22
nenhum	2765,92
baixo/moderado	3134,82
mudando	2693,38
mudando	3144,96
nenhum	3508,47
alto	3728,29
nenhum	4012,09
nenhum	3973,98
mudando	3342,50
mudando	3278,79
mudando	3369,27
baixo/moderado	3583,00
nenhum	2323,93

Compreensão dos Dados

- Estamos interessados apenas no grupo observado ou o grupo deve fornecer informações sobre um conjunto maior de indivíduos (população x amostra)?
- Que tipos de variáveis estão presentes e o que pode ser feito com elas?
- Estes dados ajudam a responder as perguntas ou precisamos de mais dados?



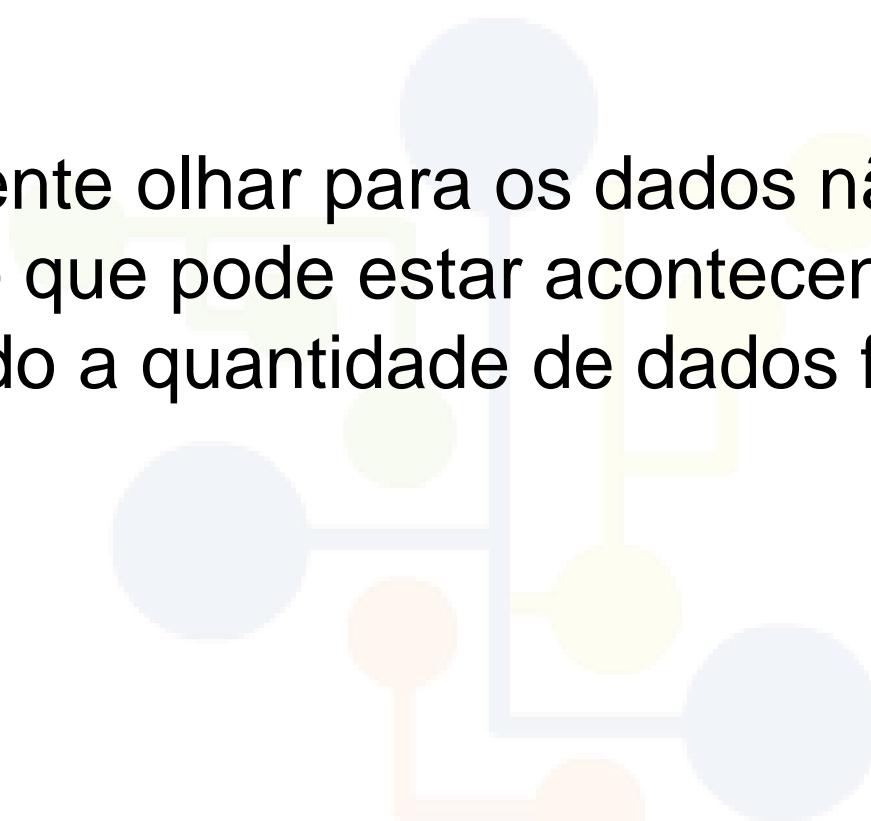
Data Science Academy



A simples visualização dos dados, ainda que contenha
toda a informação, muitas vezes não diz nada



Data Science Academy



Simplesmente olhar para os dados não fornece um quadro claro do que pode estar acontecendo, especialmente quando a quantidade de dados for muito grande.



Data Science Academy

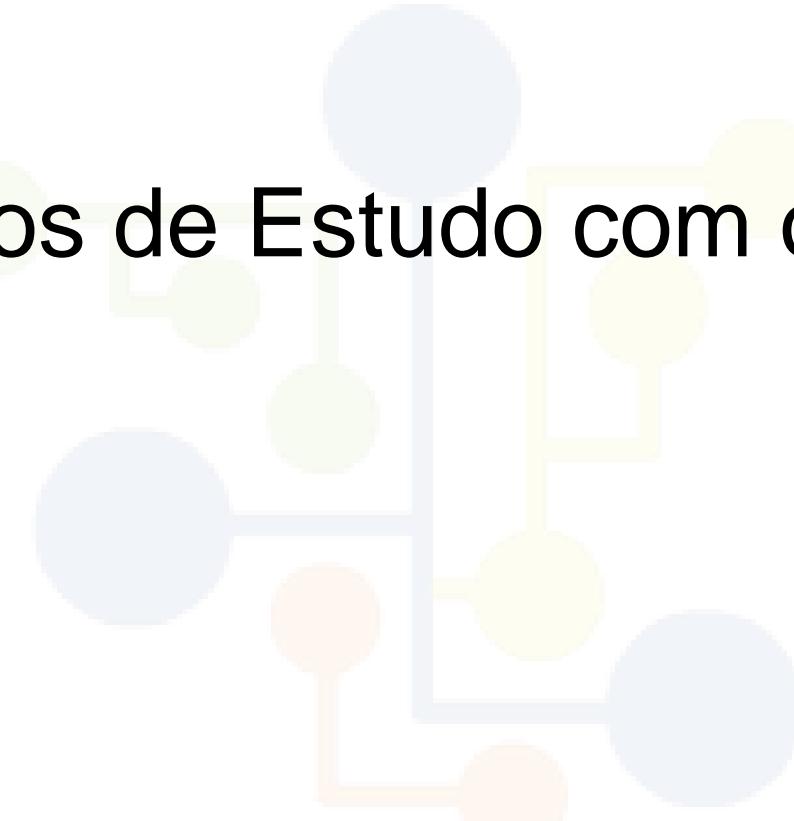
Por isso podemos ensinar algoritmos a fazer isso por nós.
Exatamente onde começa o trabalho do Machine Learning.



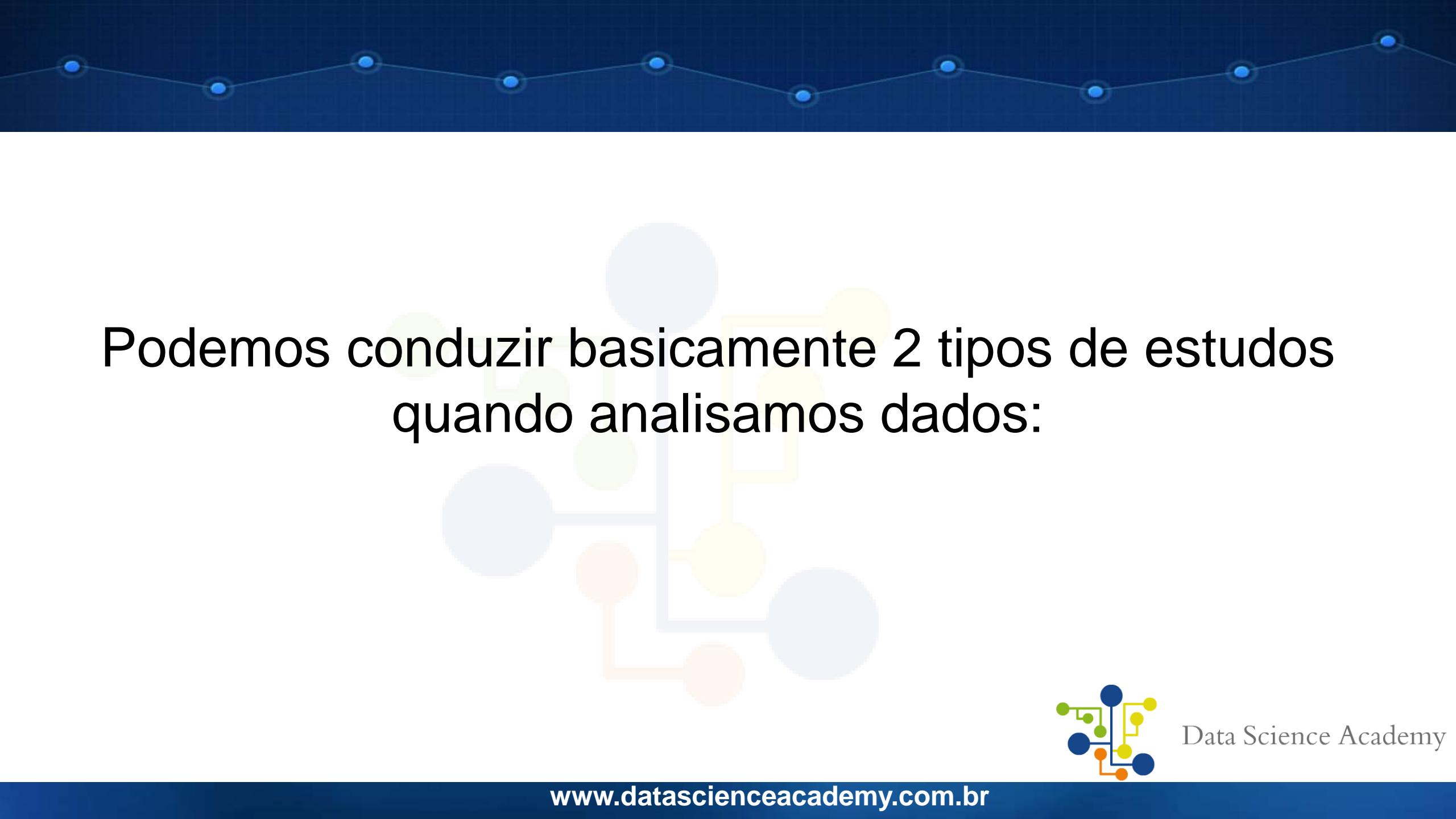
Data Science Academy



Tipos de Estudo com os Dados



Data Science Academy



Podemos conduzir basicamente 2 tipos de estudos quando analisamos dados:



Data Science Academy

Passados =
retrospectiva
Durante estudo
= prospectiva

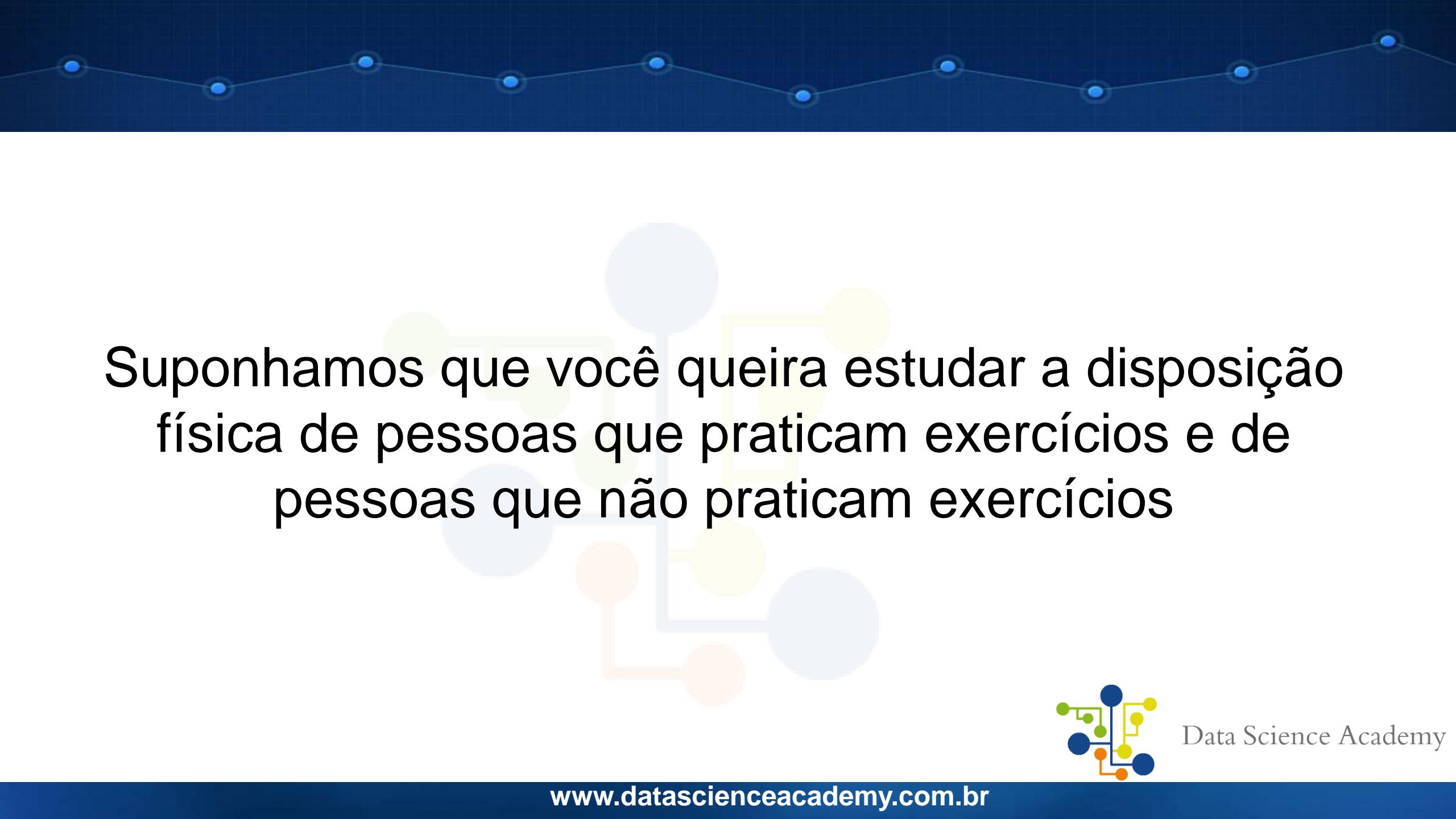
Podemos conduzir basicamente 2 tipos de estudos quando analisamos dados:

Observação

Experimento



Data Science Academy

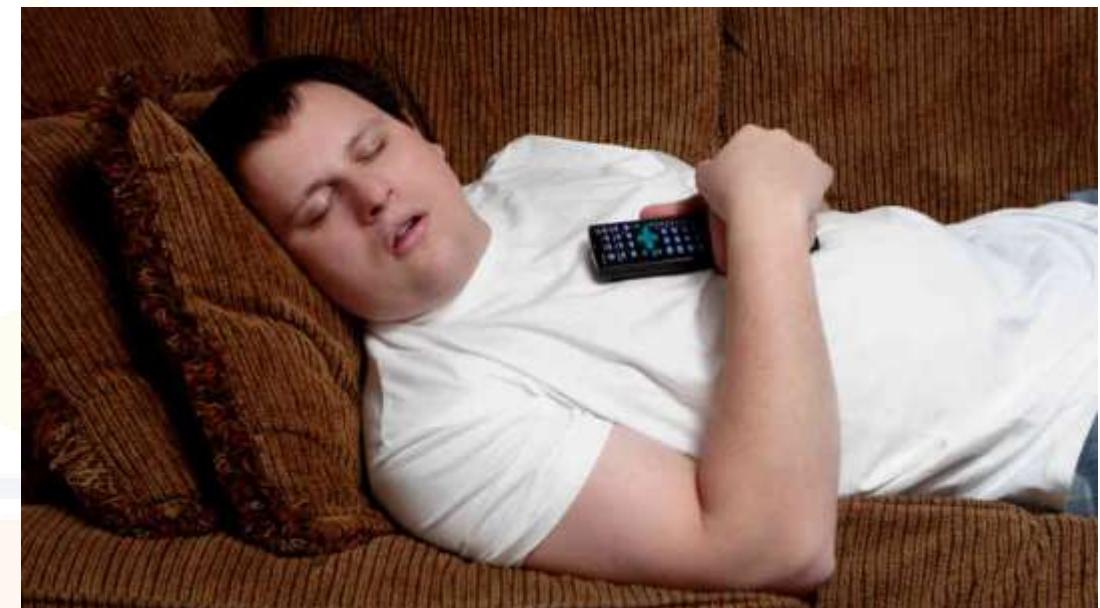


Suponhamos que você queira estudar a disposição física de pessoas que praticam exercícios e de pessoas que não praticam exercícios



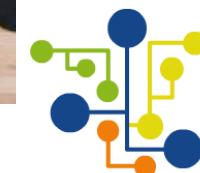
Data Science Academy

Estudo baseado em Observação



Data Science Academy

Estudo baseado em Experimento



Data Science Academy

Esse tópico chegou ao final



Data Science Academy



Medidas de Tendência Central (Média, Mediana, Moda) Medidas de Dispersão (Variância e Desvio Padrão)



Tendência Central



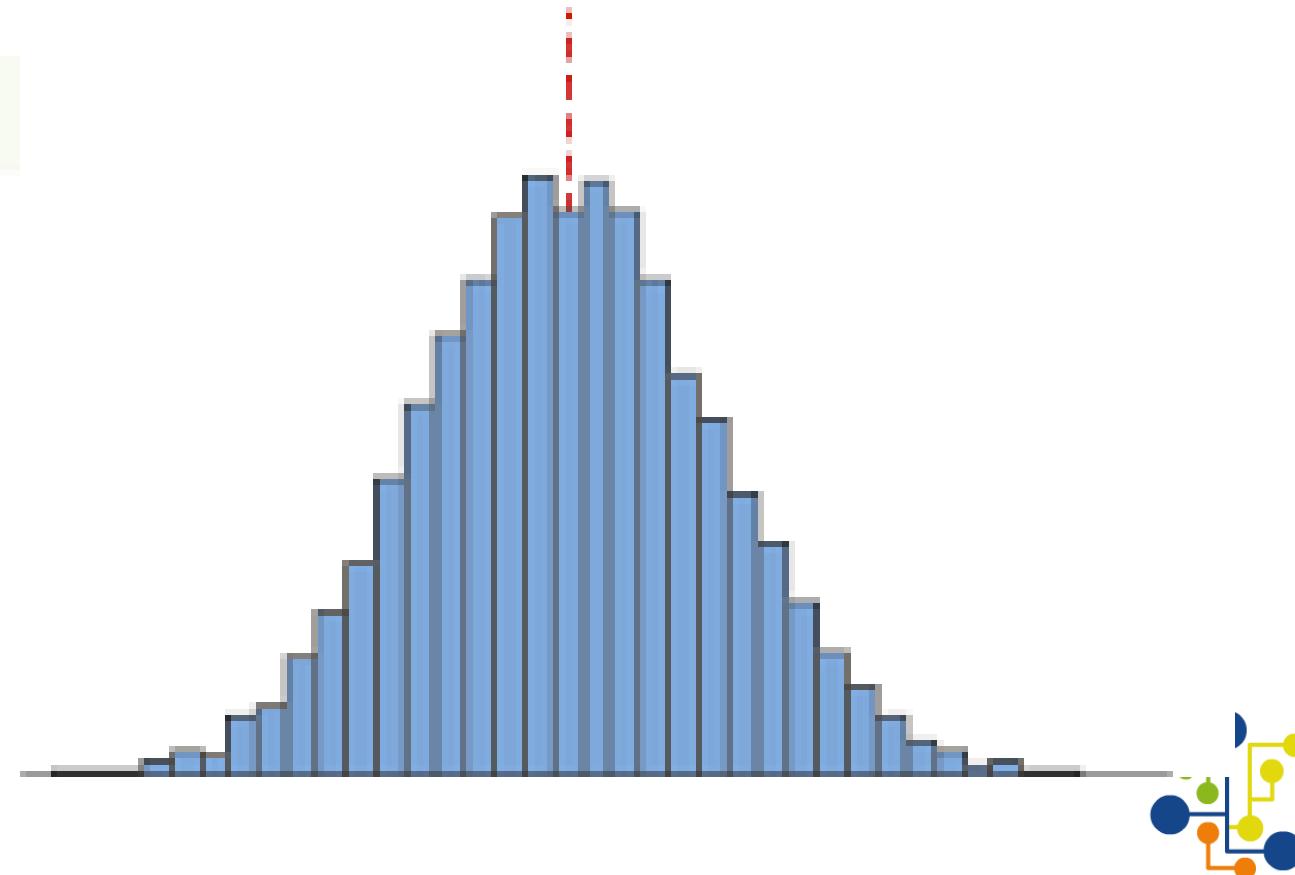
Caracterizar conjunto de valores

Tende a estar no meio



Data Science Academy

Média



Data Science Academy

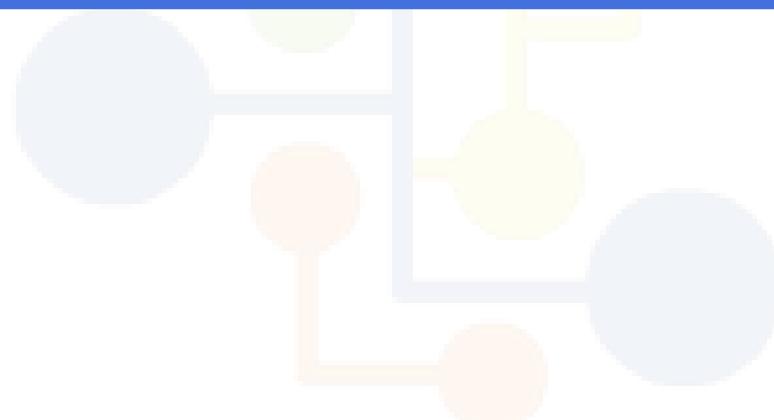


Sem dúvida, **médias** são as formas mais simples de identificar tendências em um conjunto de dados



Data Science Academy

Exemplo



Data Science Academy

Conjunto de dados 1: Preço médio de casas na cidade do Rio de Janeiro em Maio de 2015:



Data Science Academy

Conjunto de dados 1: Preço médio de casas na cidade do Rio de Janeiro em Maio de 2015:

Casa 1 – R\$ 245.000,00

Casa 2 – R\$ 325.000,00

Casa 3 – R\$ 275.000,00

Casa 4 – R\$ 315.000,00

Casa 5 – R\$ 295.000,00

Média = soma dos valores / números de elementos

Média = R\$ 291.000,00



Data Science Academy

Conjunto de dados 2: Preço médio de casas na cidade do Rio de Janeiro em Junho de 2015.

Casa 1 – R\$ 45.000,00

Casa 2 – R\$ 325.000,00

Casa 3 – R\$ 275.000,00

Casa 4 – R\$ 315.000,00

Casa 5 – R\$ 515.000,00

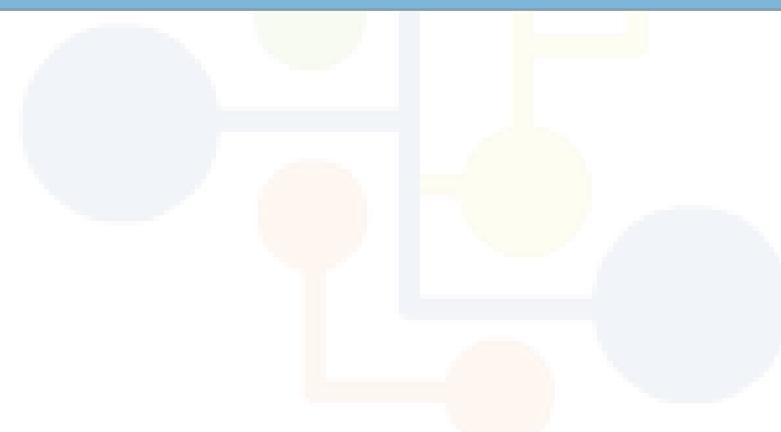
Média = soma dos valores / números de elementos

Média = R\$ 295.000,00



Data Science Academy

Conclusão



Data Science Academy

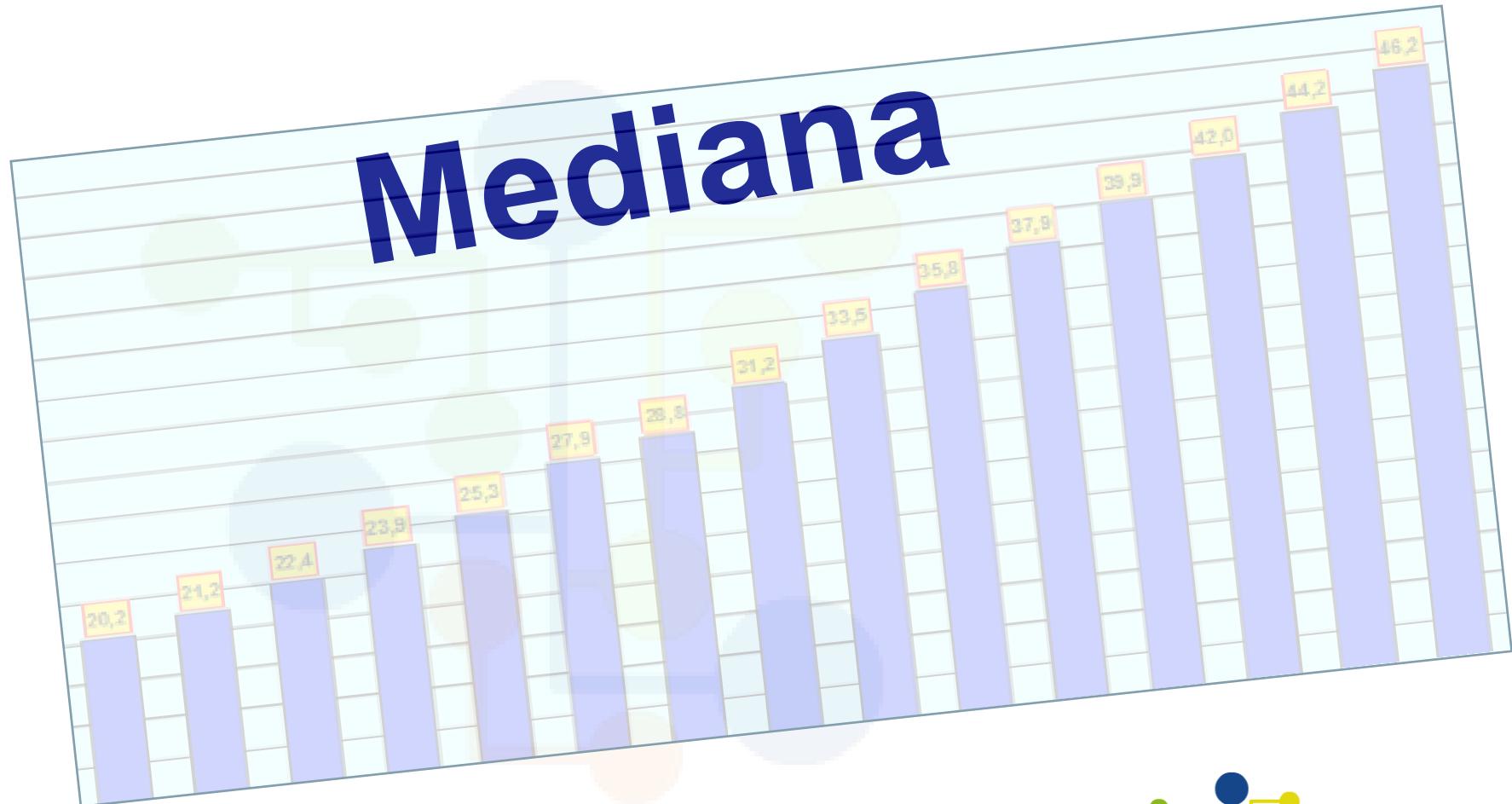
Resultado:

Os 2 conjuntos de dados possuem praticamente a mesma média, entretanto o conjunto de dados 2 possui valores extremos, chamados outliers, que podem levar a conclusão distorcidas.



Data Science Academy

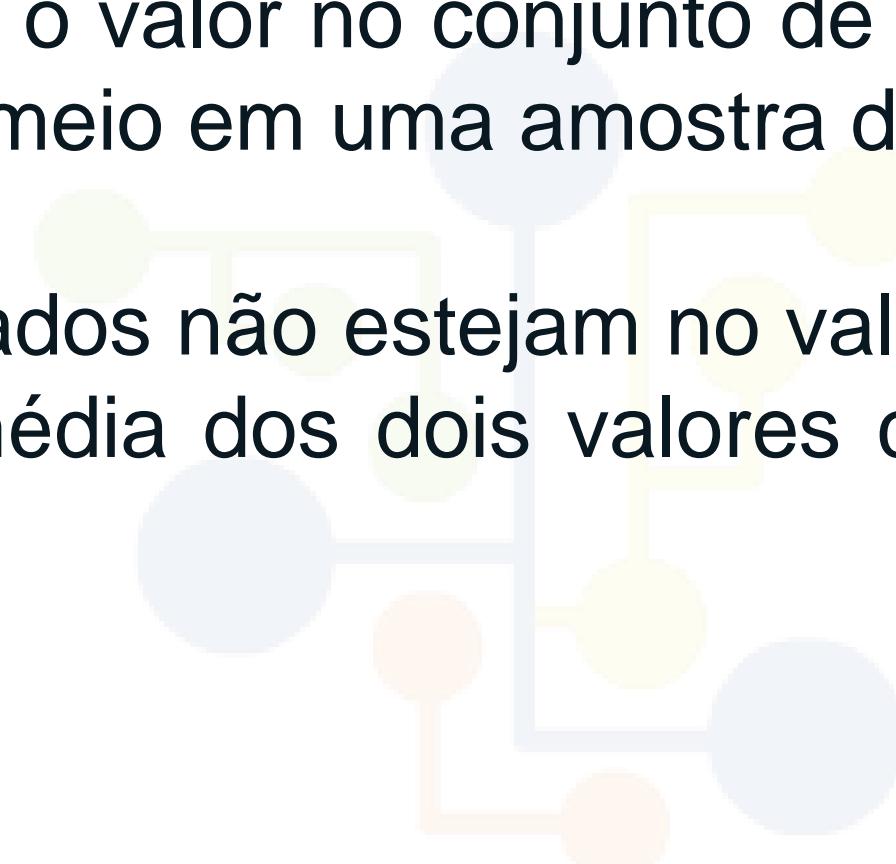
Mediana



Data Science Academy



Mediana é o valor no conjunto de dados, em que mostra o valor do meio em uma amostra de informações.

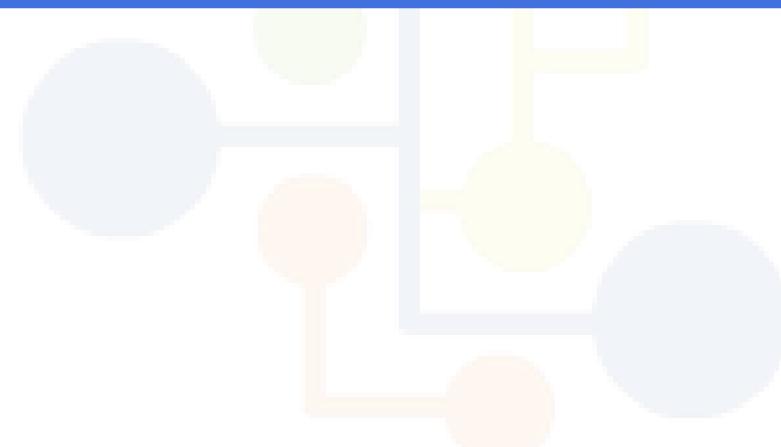


Caso os dados não estejam no valor central é necessário extrair a média dos dois valores centrais para chegar a mediana.



Data Science Academy

Exemplo I



Data Science Academy

Conjunto de dados 1: Preços de casas na cidade do Rio de Janeiro em Maio de 2015:

Casa 1 – R\$ 245.000,00

Casa 2 – R\$ 325.000,00

Casa 3 – R\$ 275.000,00

Casa 4 – R\$ 315.000,00

Casa 5 – R\$ 296.500,00



Data Science Academy



Primeiramente devemos classificar esses preços (valores) em ordem crescente:

Casa 1 – R\$ 245.000,00

Casa 3 – R\$ 275.000,00

Casa 5 – R\$ 296.500,00

Casa 4 – R\$ 315.000,00

Casa 2 – R\$ 325.000,00



Data Science Academy

Resultado: Da mediana dos preços das casas no Rio de Janeiro em Maio de 2015.

Casa 1 – R\$ 245.000,00

Casa 3 – R\$ 275.000,00

Casa 5 – R\$ 296.500,00

Casa 4 – R\$ 315.000,00

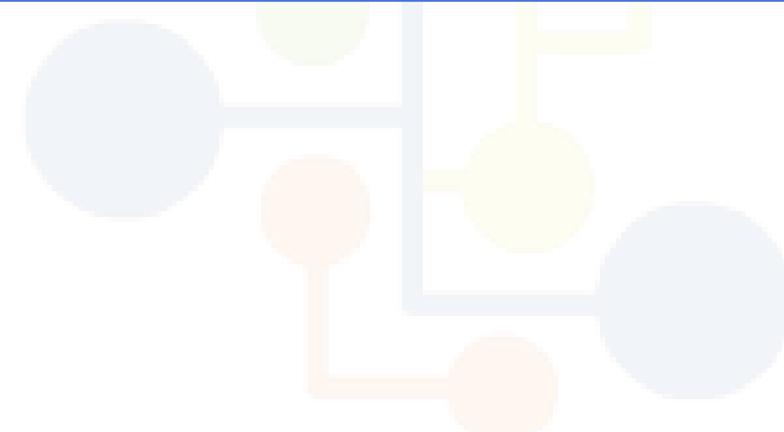
Casa 2 – R\$ 325.000,00

A **mediana** é o preço **R\$ 296.500,00**, ou seja o valor do meio.



Data Science Academy

Exemplo II



Data Science Academy

Conjunto de dados 2: Preços de casas na cidade do Rio de Janeiro em Junho de 2015:

Casa 1 – R\$ 245.000,00

Casa 2 – R\$ 325.000,00

Casa 3 – R\$ 275.000,00

Casa 4 – R\$ 315.000,00

Casa 5 – R\$ 295.000,00

Casa 6 – R\$ 300.000,00



Data Science Academy

Primeiramente devemos classificar esses preços (valores) em ordem crescente:

Casa 1 – R\$ 245.000,00

Casa 3 – R\$ 275.000,00

Casa 5 – R\$ 296.500,00

Casa 6 – R\$ 300.000,00

Casa 4 – R\$ 315.000,00

Casa 2 – R\$ 325.000,00



Data Science Academy

Para: Encontrarmos a mediana.

Casa 1 – R\$ 245.000,00

Casa 3 – R\$ 275.000,00

Casa 5 – R\$ 296.500,00

Casa 6 – R\$ 300.000,00

Casa 4 – R\$ 315.000,00

Casa 2 – R\$ 325.000,00

A média dos valores centrais é =
R\$ 298.250,00

A **mediana** é o preço **R\$ 298.250,00**, ou seja o valor da média dos valores **centrais**.



Data Science Academy



Moda



Data Science Academy

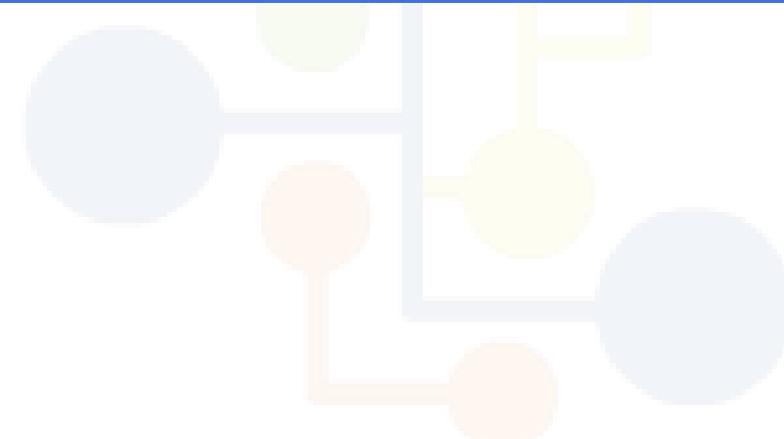


É o valor que aparece mais frequentemente no conjunto de dados.



Data Science Academy

Exemplo



Data Science Academy

Conjunto de dados: Valores de casas na cidade do Rio de Janeiro em Junho de 2015:

Casa 01 – R\$ 245.000,00

Casa 02 – R\$ 325.000,00

Casa 03 – R\$ 300.000,00

Casa 04 – R\$ 300.000,00

Casa 05 – R\$ 300.000,00

Casa 06 – R\$ 245.000,00

Casa 07 – R\$ 515.000,00

Casa 08 – R\$ 278.000,00

Casa 09 – R\$ 300.000,00

Casa 10 – R\$ 488.000,00



Data Science Academy

Conjunto de dados: Valores de casas na cidade do Rio de Janeiro em Junho de 2015:

Casa 01 – R\$ 245.000,00

Casa 02 – R\$ 325.000,00

Casa 03 – R\$ 300.000,00

Casa 04 – R\$ 300.000,00

Casa 05 – R\$ 300.000,00

Casa 06 – R\$ 245.000,00

Casa 07 – R\$ 515.000,00

Casa 08 – R\$ 278.000,00

Casa 09 – R\$ 300.000,00

Casa 10 – R\$ 488.000,00

A **moda** é o valor **R\$ 300.000,00**, ou seja o valor que mais se repete.

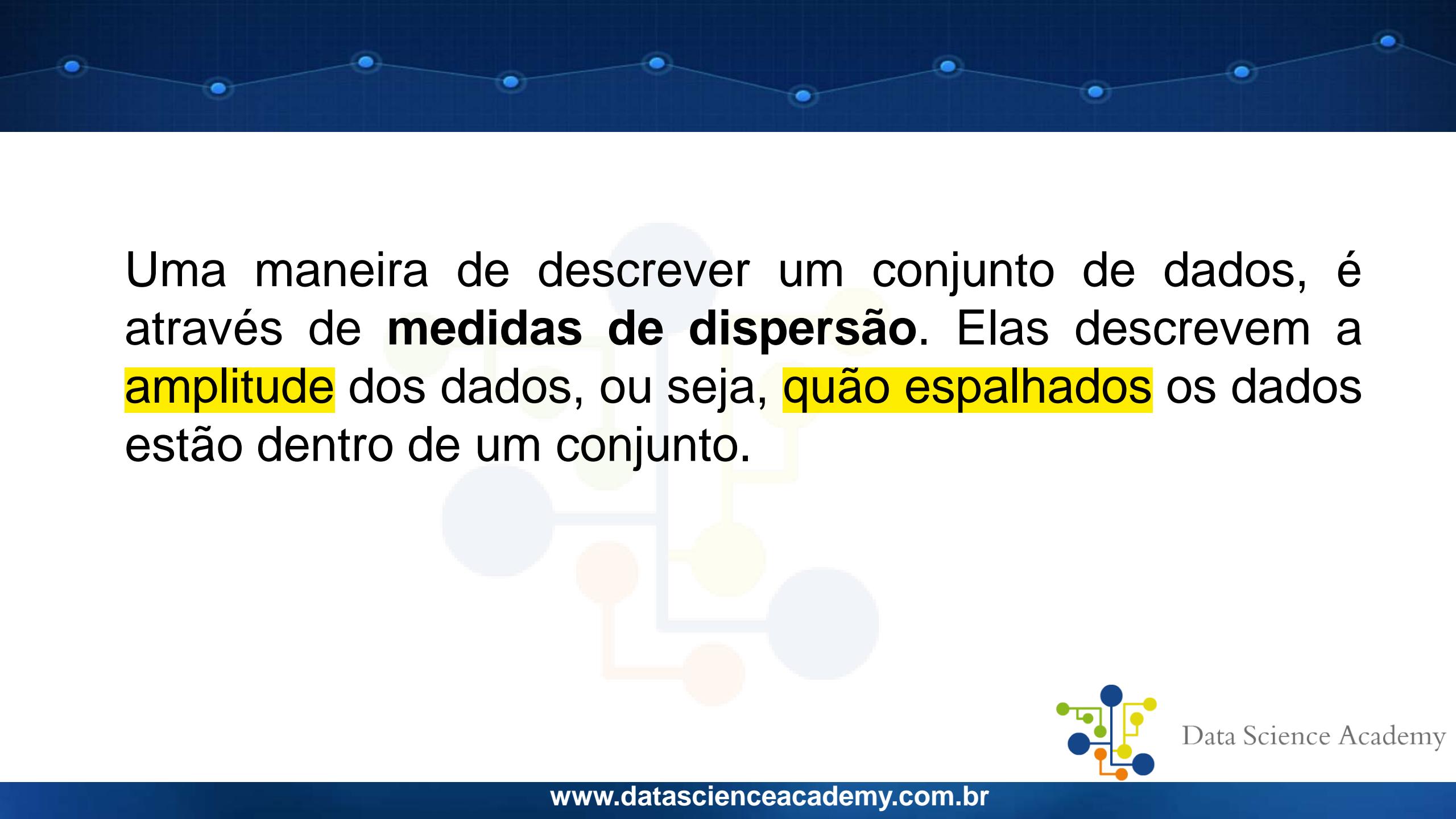


Data Science Academy

Variância e Desvio Padrão



Data Science Academy



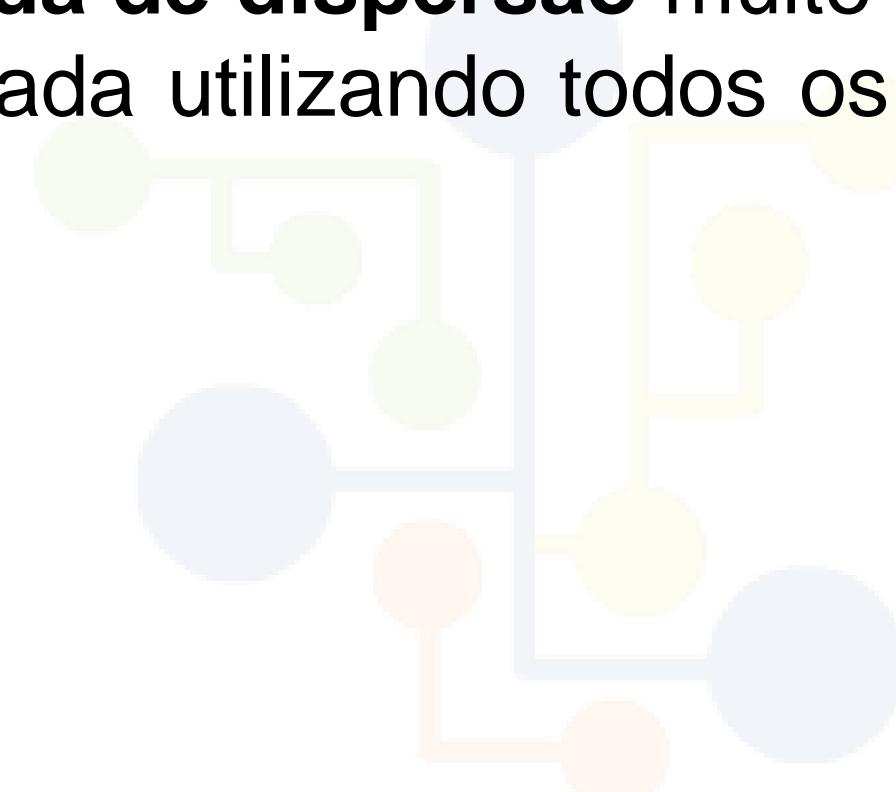
Uma maneira de descrever um conjunto de dados, é através de **medidas de dispersão**. Elas descrevem a **amplitude** dos dados, ou seja, **quão espalhados** os dados estão dentro de um conjunto.



Data Science Academy

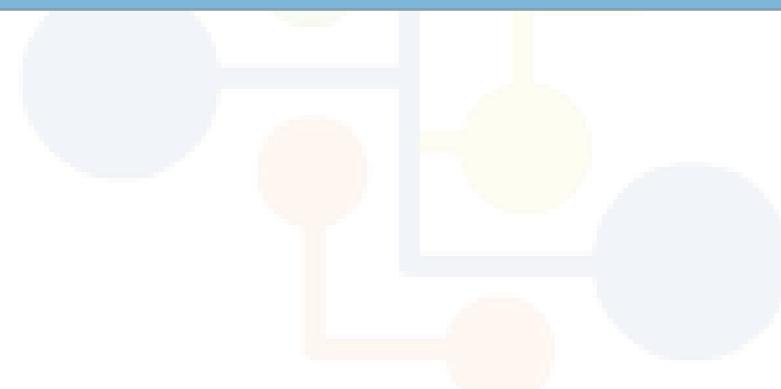


Uma medida de dispersão muito útil, é a **variância**, que é determinada utilizando todos os dados do conjunto de dados.



Data Science Academy

Variância



Data Science Academy



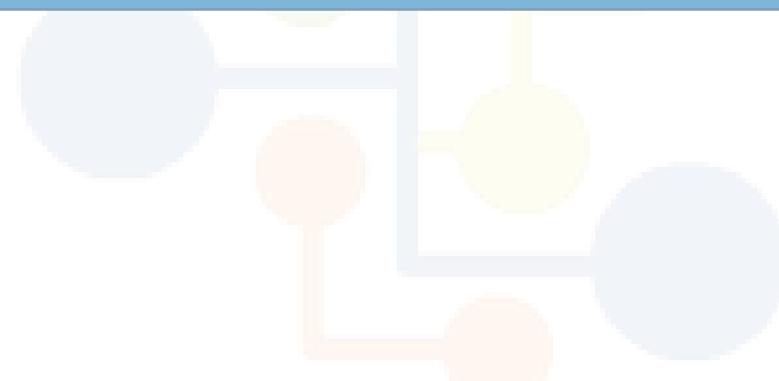
**A variância mede a amplitude (variabilidade) dos dados
em relação à média**



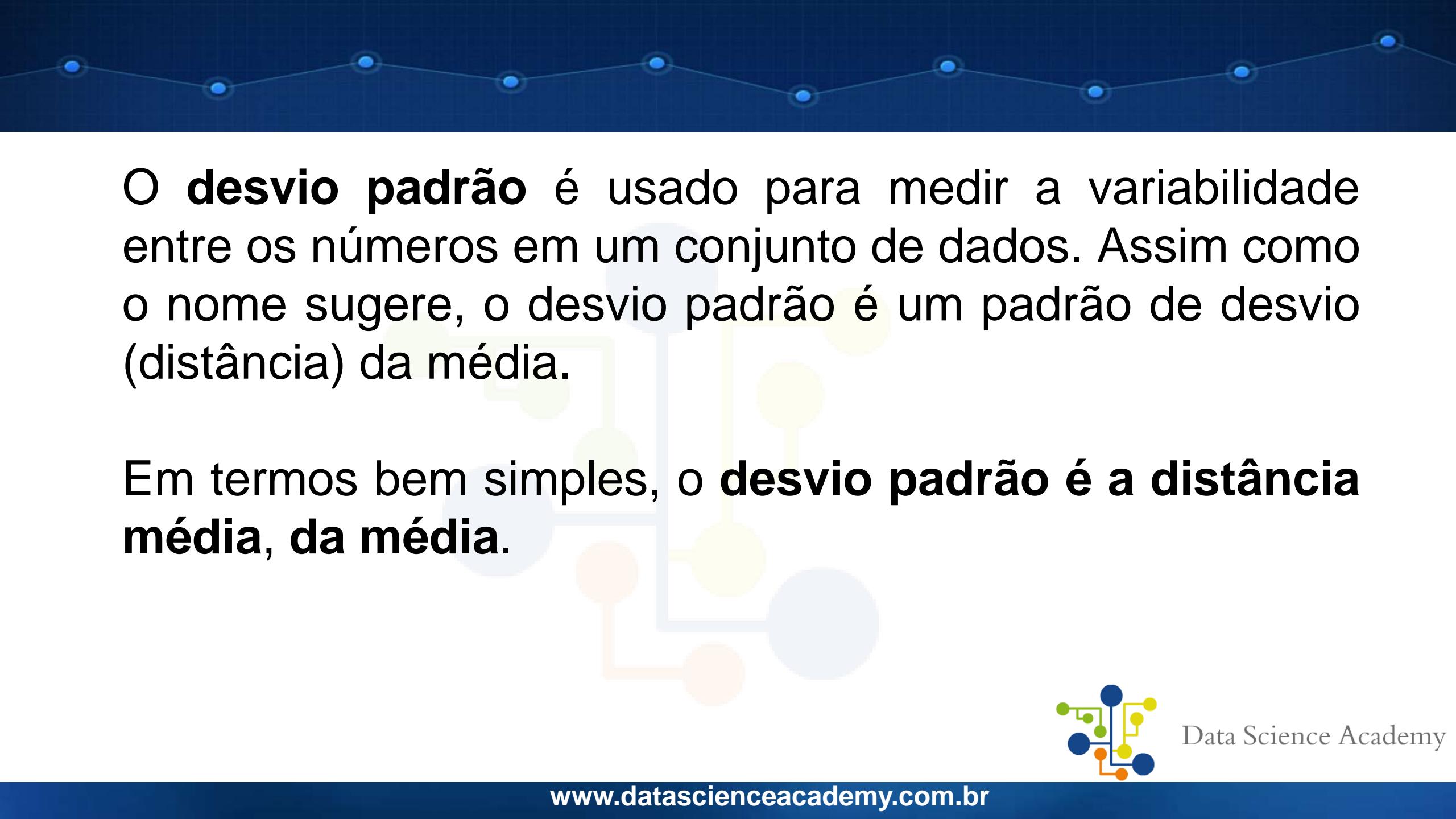
Data Science Academy



Desvio Padrão



Data Science Academy



O **desvio padrão** é usado para medir a variabilidade entre os números em um conjunto de dados. Assim como o nome sugere, o desvio padrão é um padrão de desvio (distância) da média.

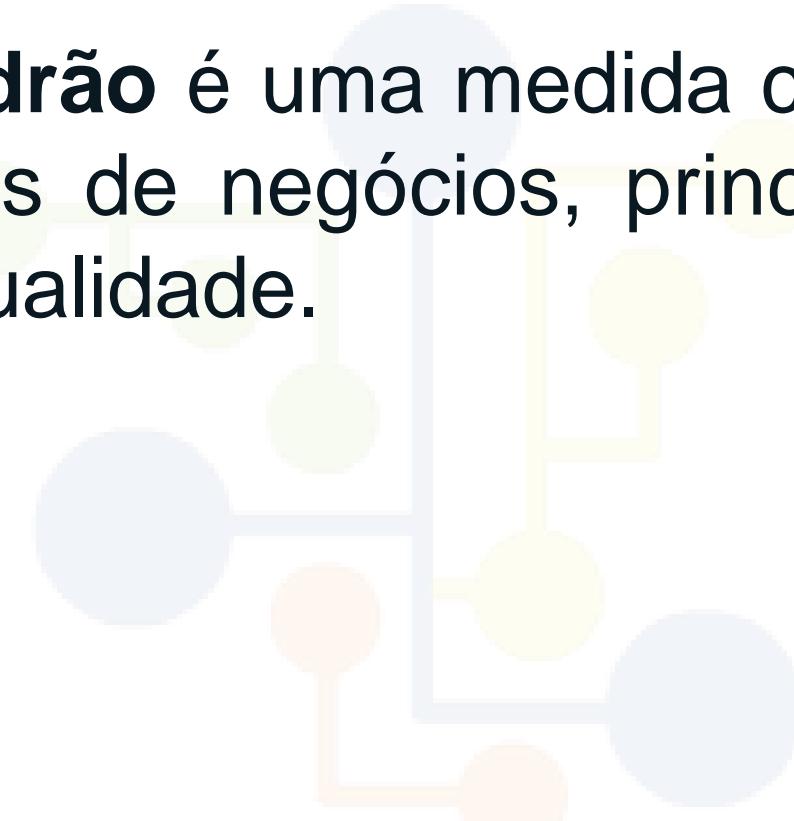
Em termos bem simples, o **desvio padrão** é a distância média, da média.



Data Science Academy

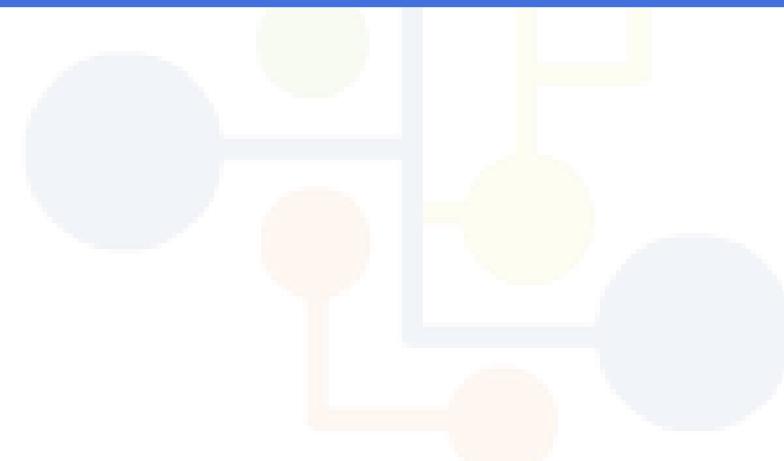


O **desvio padrão** é uma medida comum de consistência em aplicações de negócios, principalmente na área de controle de qualidade.



Data Science Academy

Exemplo



Data Science Academy

Anderson e Patrícia



Data Science Academy

Anderson – cursa 6 disciplinas na faculdade de Estatística e obteve as seguintes notas no exame final:

Disciplinas	Notas
Disciplina 1	100
Disciplina 2	100
Disciplina 3	100
Disciplina 4	50
Disciplina 5	50
Disciplina 6	50



Data Science Academy

Anderson – cursa 6 disciplinas na faculdade de Estatística e obteve as seguintes notas no exame final:

Disciplinas	Notas
Disciplina 1	100
Disciplina 2	100
Disciplina 3	100
Disciplina 4	50
Disciplina 5	50
Disciplina 6	50

Média final = 75



Data Science Academy

Patrícia – também cursa 6 disciplinas na faculdade de Estatística e obteve as seguintes notas no exame final:

Disciplinas	Notas
Disciplina 1	75
Disciplina 2	74
Disciplina 3	76
Disciplina 4	77
Disciplina 5	75
Disciplina 6	74



Data Science Academy

Patrícia – também cursa 6 disciplinas na faculdade de Estatística e obteve as seguintes notas no exame final:

Disciplinas	Notas
Disciplina 1	75
Disciplina 2	74
Disciplina 3	76
Disciplina 4	77
Disciplina 5	75
Disciplina 6	74

Média final = 75

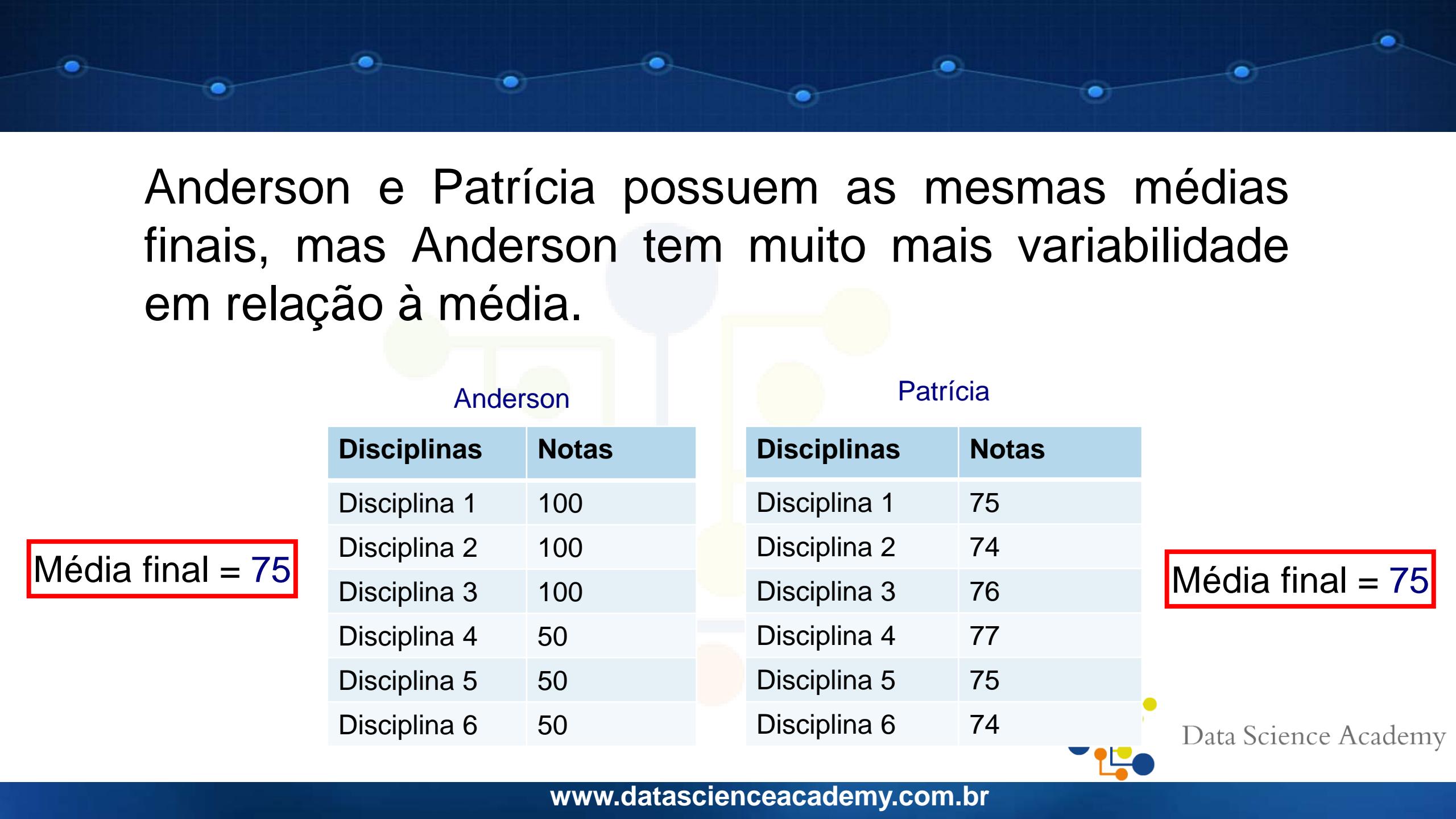


Data Science Academy

Conclusão



Data Science Academy



Anderson e Patrícia possuem as mesmas médias finais, mas Anderson tem muito mais variabilidade em relação à média.

Anderson

Disciplinas	Notas
Disciplina 1	100
Disciplina 2	100
Disciplina 3	100
Disciplina 4	50
Disciplina 5	50
Disciplina 6	50

Média final = 75

Patrícia

Disciplinas	Notas
Disciplina 1	75
Disciplina 2	74
Disciplina 3	76
Disciplina 4	77
Disciplina 5	75
Disciplina 6	74

Média final = 75

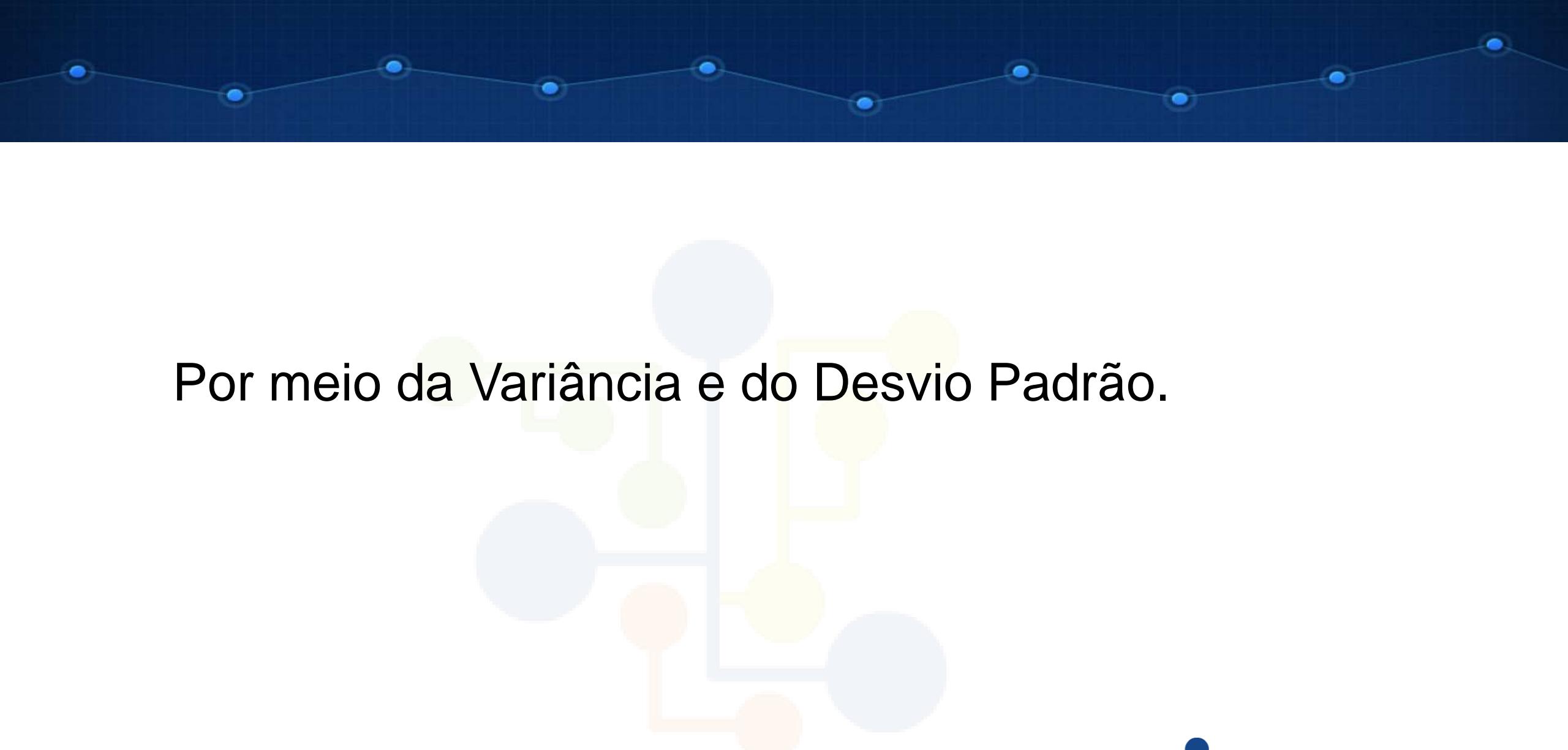




Como diferenciar essas duas distribuições?



Data Science Academy



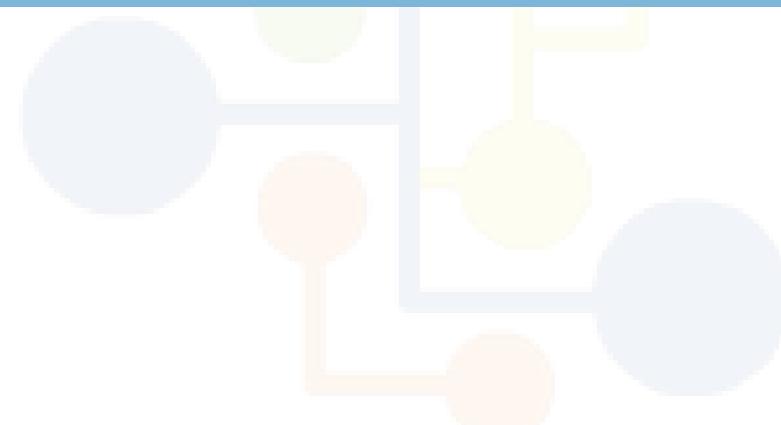
Por meio da Variância e do Desvio Padrão.



Data Science Academy



Análise do Resultado Final



Data Science Academy

Resultado do Anderson:

Média Amostra= **75**

Variância da Amostra = **750**

Desvio Padrão da Amostra = **27.39**

Resultado da Patrícia

Média Amostra= **75**

Variância da Amostra = **1.37**

Desvio Padrão da Amostra = **1.17**



Resultado do Anderson:

Média Amostra =	75
Variância da Amostra =	750
Desvio Padrão da Amostra =	27.39

As notas relevam comportamentos diferentes de estudo.

$$CV = 27,39/75 * 100 \\ = 36,52\%$$

Mais variabilidade nas notas.

O desvio padrão foi alto, entre 27 pontos para mais ou para menos.

Aprovado. Terá de cursar mais 3 disciplinas para receber o diploma final.

Resultado da Patrícia

Média Amostra =	75
Variância da Amostra =	1.37
Desvio Padrão da Amostra =	1.17

As notas relevam comportamentos diferentes de estudo.

$$CV = 1,17/75 * 100 \\ = 1,56$$

Manteve a variabilidade baixa com notas mais uniformes.

O desvio padrão foi baixo, entre 1,17 pontos para mais ou para menos.

Aprovada. Receberá o diploma de curso final nesse semestre.

Exemplo



Data Science Academy

Um fabricante de sabonetes.

Recentemente a empresa recebeu reclamações de consumidores, de que os sabonetes estavam com pesos inferiores ao declarado nas embalagens.



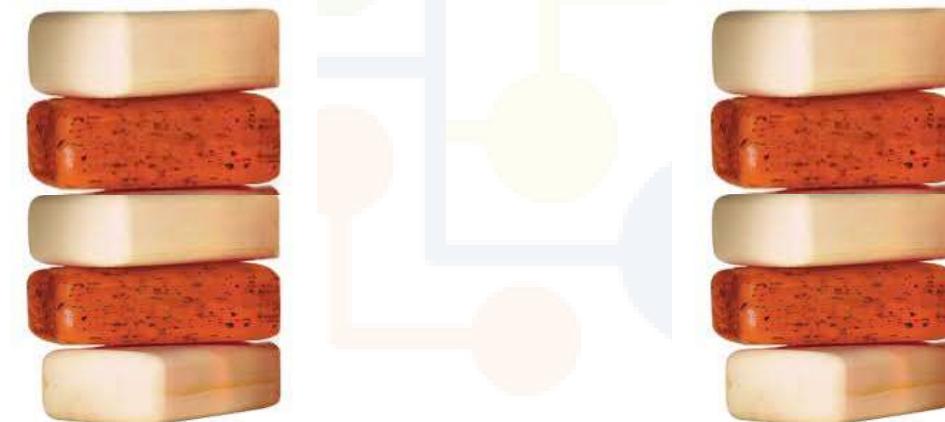
Data Science Academy

O departamento de controle de qualidade precisa agora determinar se as unidades de sabonete estão sendo produzidas com o mesmo peso, a fim de detectar anomalias no processo de produção.

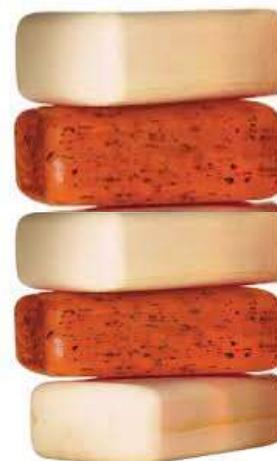


Data Science Academy

Para isso, eles coletam duas amostras de 5 sabonetes cada uma, sendo a amostra 1 no turno da manhã e a amostra 2 no turno da tarde, ao longo de um dia de produção. Cada sabonete deve ser produzido para ter um peso final de 18 gramas.



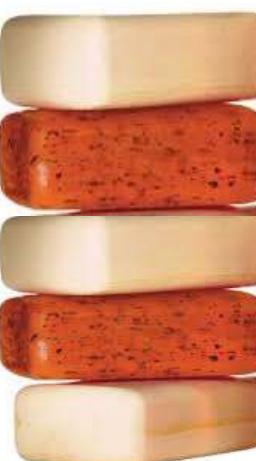
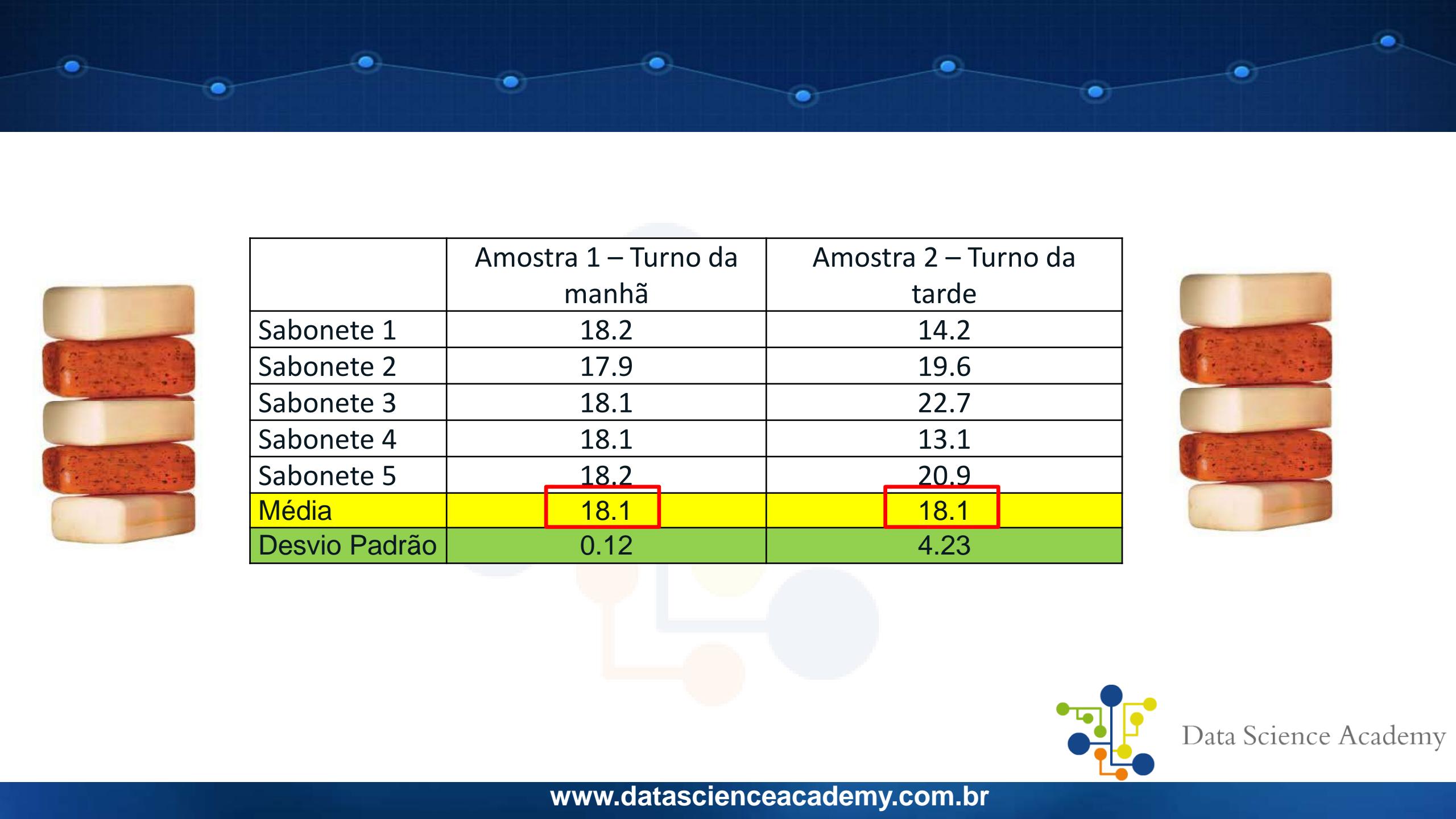
Data Science Academy



	Amostra 1 – Turno da manhã
Sabonete 1	18.2
Sabonete 2	17.9
Sabonete 3	18.1
Sabonete 4	18.1
Sabonete 5	18.2
Média	18.1
Desvio Padrão	0.12



Data Science Academy



	Amostra 1 – Turno da manhã	Amostra 2 – Turno da tarde
Sabonete 1	18.2	14.2
Sabonete 2	17.9	19.6
Sabonete 3	18.1	22.7
Sabonete 4	18.1	13.1
Sabonete 5	18.2	20.9
Média	18.1	18.1
Desvio Padrão	0.12	4.23



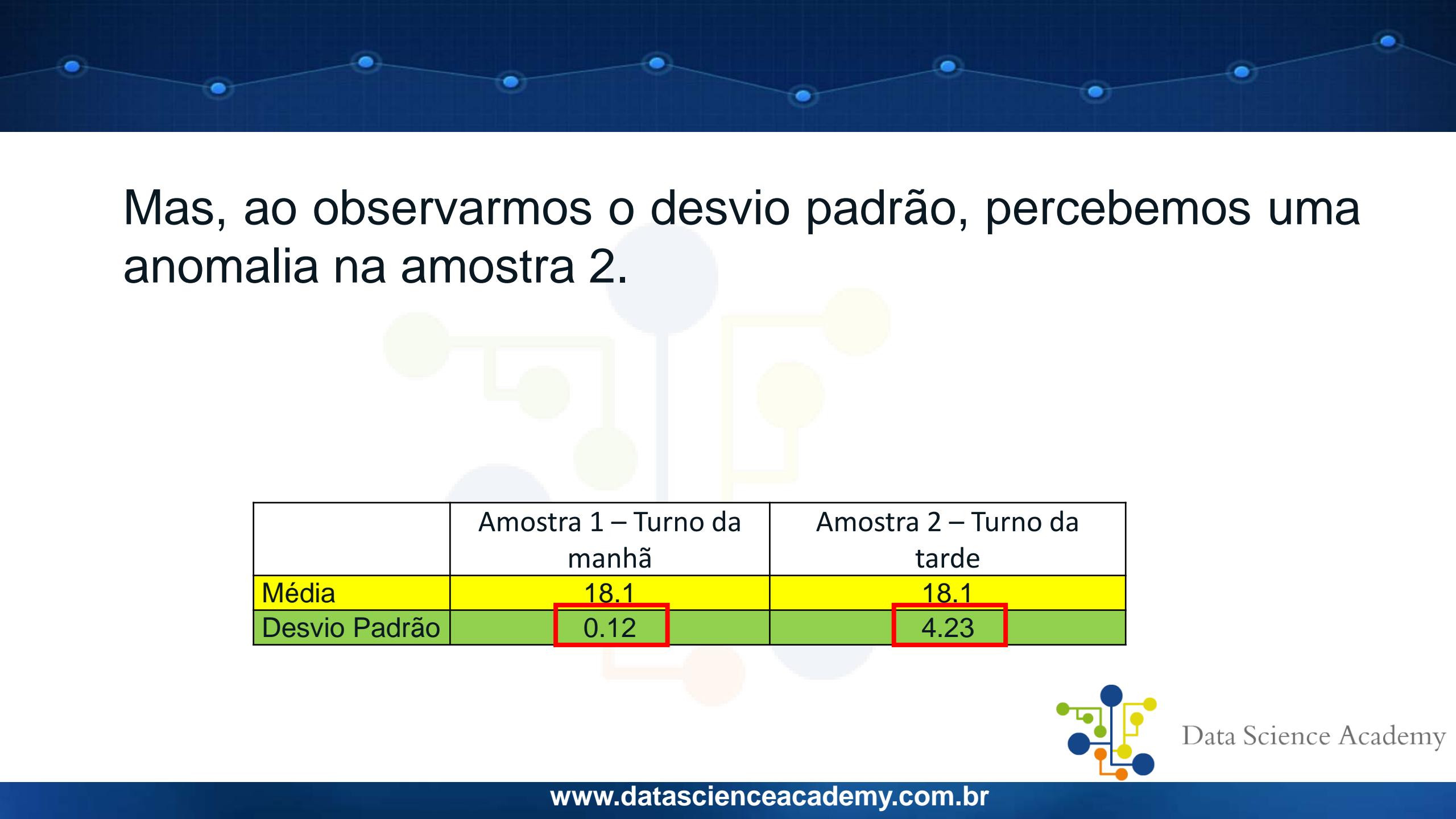
Data Science Academy

Perceba que nos 2 turnos, a média foi a mesma, ou seja, aparentemente todos os sabonetes estão sendo produzidos com o mesmo peso

	Amostra 1 – Turno da manhã	Amostra 2 – Turno da tarde
Média	18.1	18.1



Data Science Academy



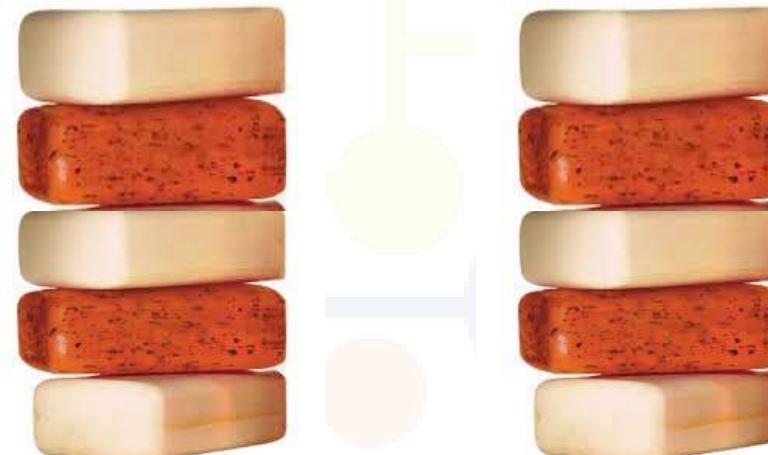
Mas, ao observarmos o desvio padrão, percebemos uma anomalia na amostra 2.

	Amostra 1 – Turno da manhã	Amostra 2 – Turno da tarde
Média	18.1	18.1
Desvio Padrão	0.12	4.23



Data Science Academy

Vale ressaltar, que pudemos fazer esta comparação, porque as amostras são iguais, mas quando as médias são muito diferentes, analisar o desvio padrão pode levar a conclusões incorretas.

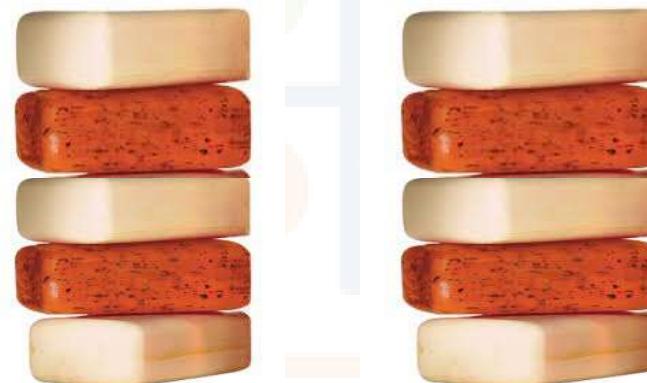


Data Science Academy



Isso ocorre, porque o **desvio padrão** é afetado pelo
tamanho do volume de dados.

Grandes volumes de dados
tendem a gerar maior
desvio padrão



Data Science Academy

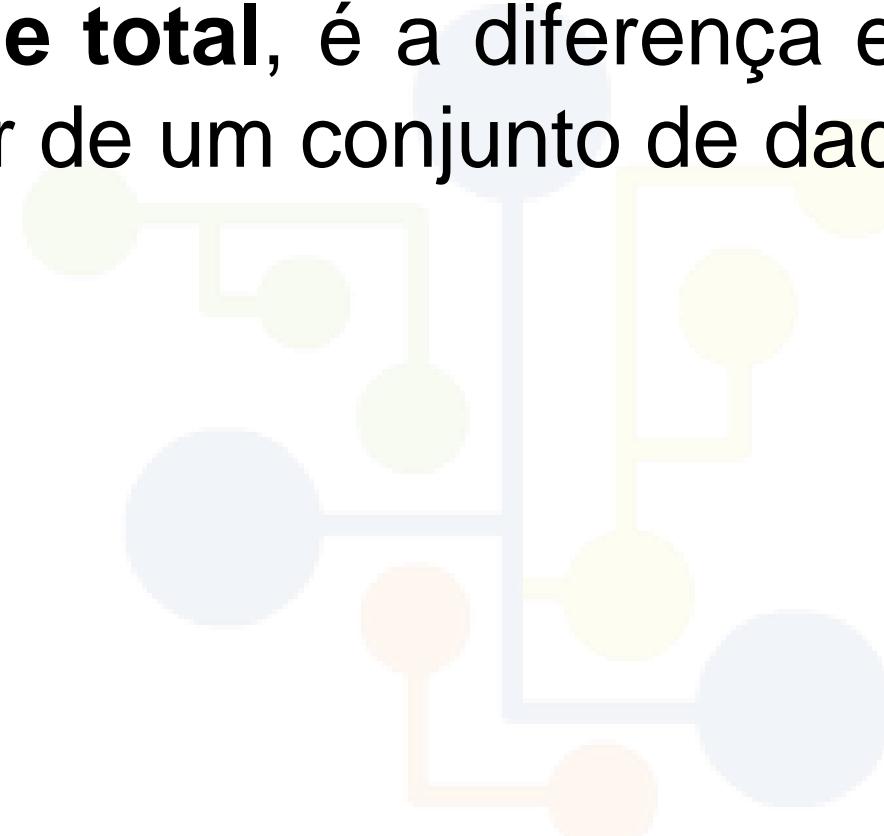
Amplitude Total



Data Science Academy

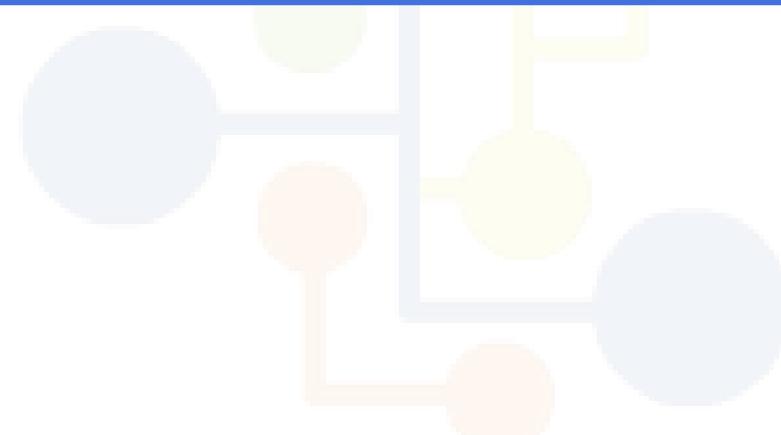


A amplitude total, é a diferença entre o maior valor e o menor valor de um conjunto de dados.



Data Science Academy

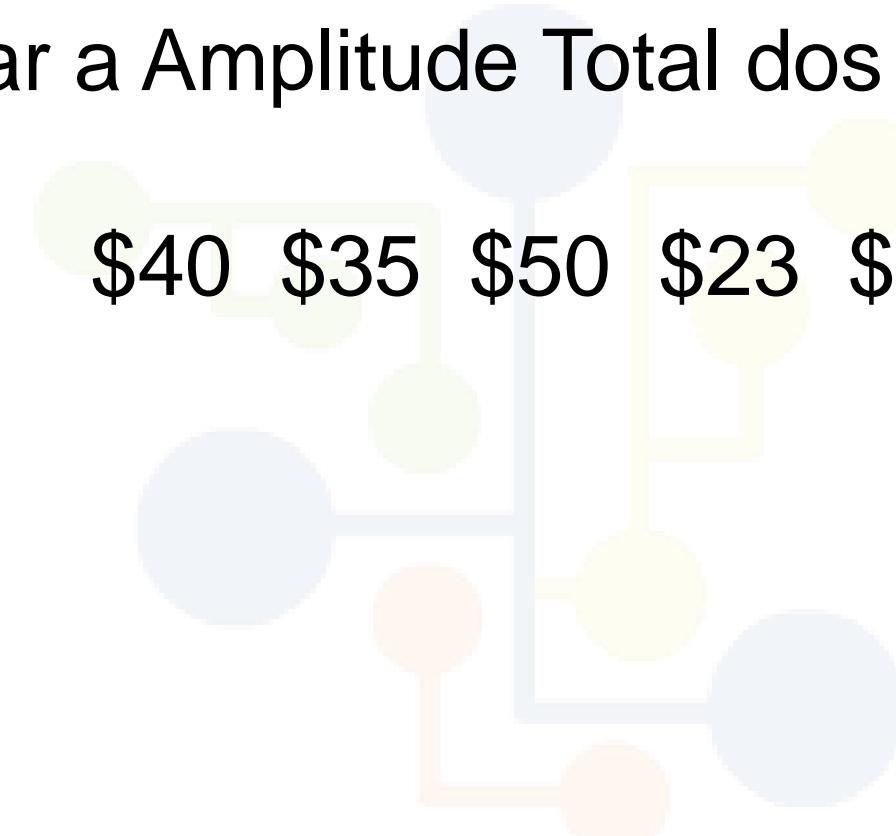
Exemplo



Data Science Academy



Calcular a Amplitude Total dos valores a seguir:



\$40 \$35 \$50 \$23 \$51 \$44



Data Science Academy



\$23 \$35 \$40 \$44 \$50 \$51

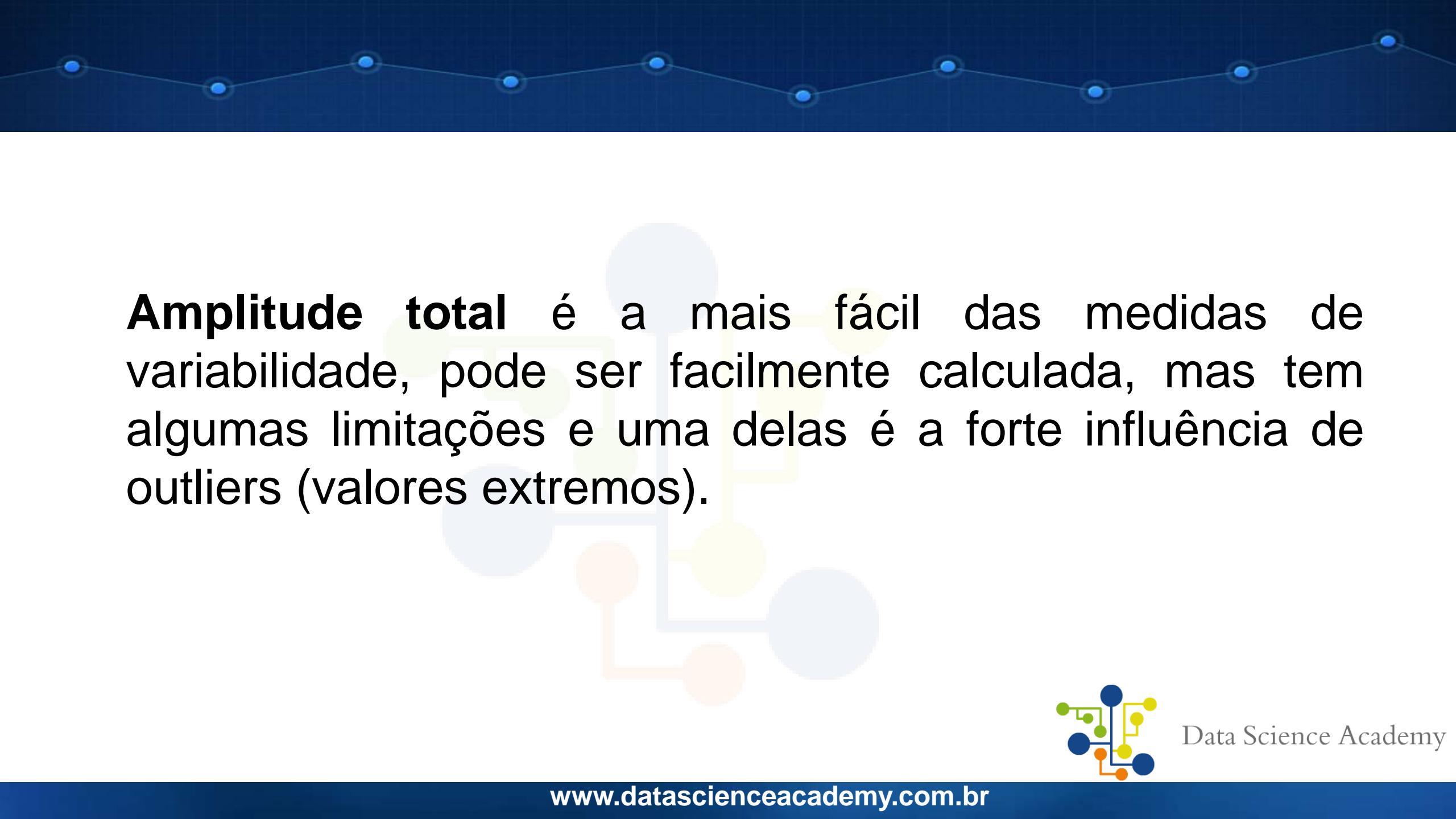
Amplitude Total = Maior valor – Menor valor

$$\$51 - \$23$$

$$\$28$$



Data Science Academy



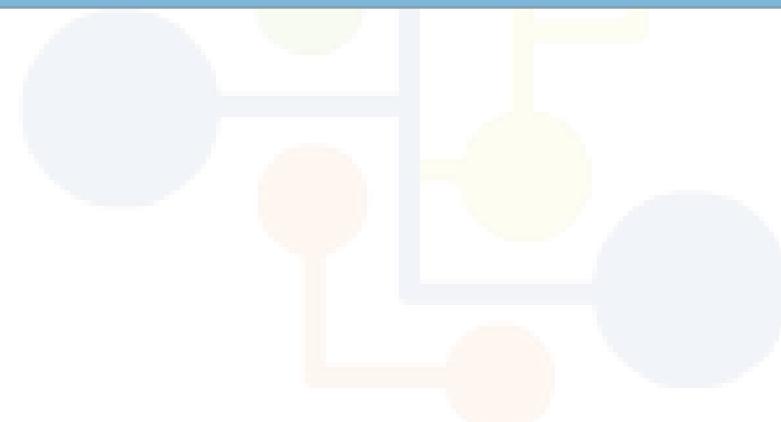
Amplitude total é a mais fácil das medidas de variabilidade, pode ser facilmente calculada, mas tem algumas limitações e uma delas é a forte influência de outliers (valores extremos).



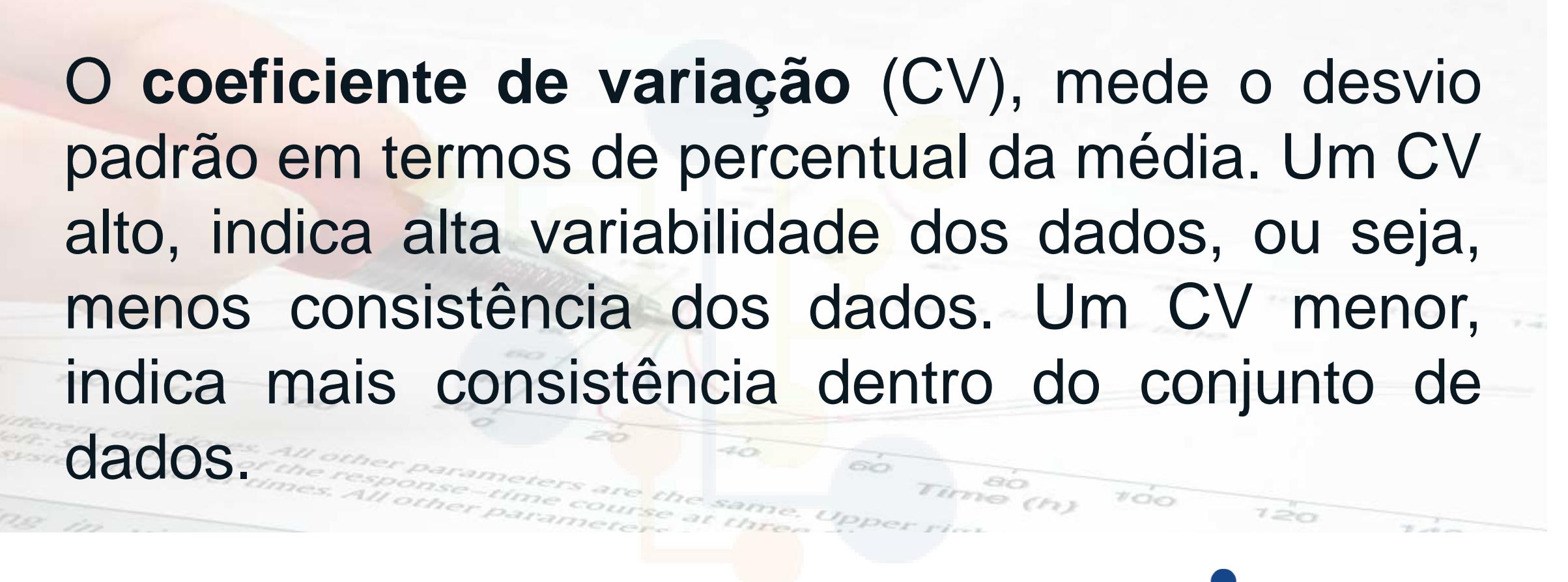
Data Science Academy



Coeficiente de Variação



Data Science Academy



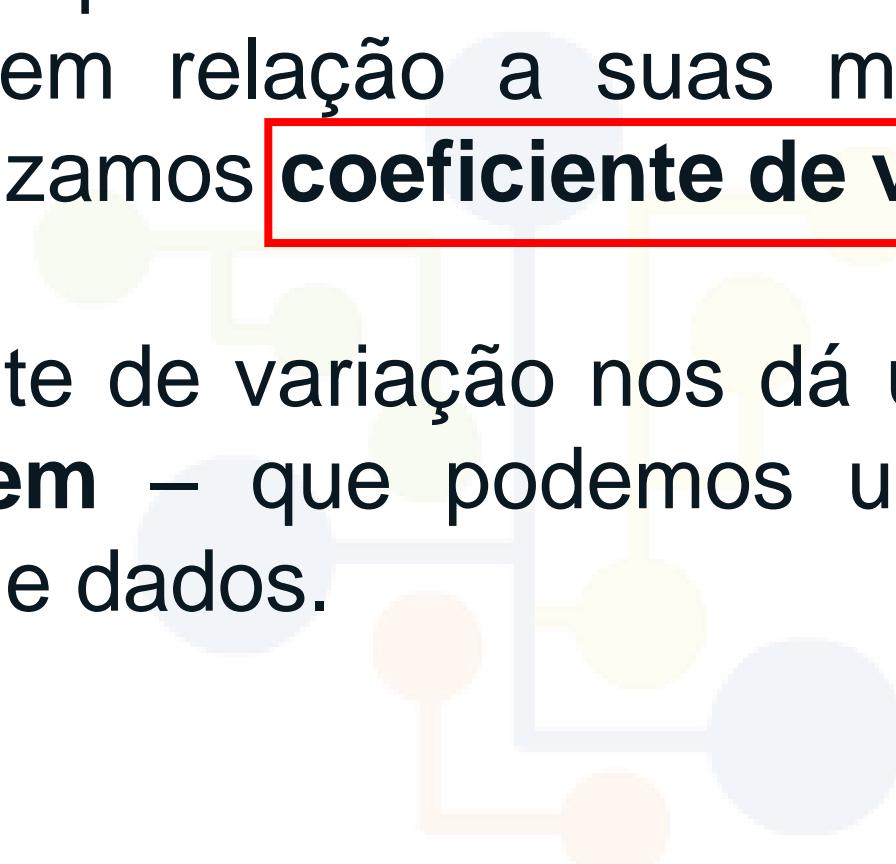
O coeficiente de variação (CV), mede o desvio padrão em termos de percentual da média. Um CV alto, indica alta variabilidade dos dados, ou seja, menos consistência dos dados. Um CV menor, indica mais consistência dentro do conjunto de dados.



Data Science Academy



Quando comparamos a consistência entre 2 conjuntos de dados em relação a suas médias, é melhor feito quando utilizamos **coeficiente de variação.**

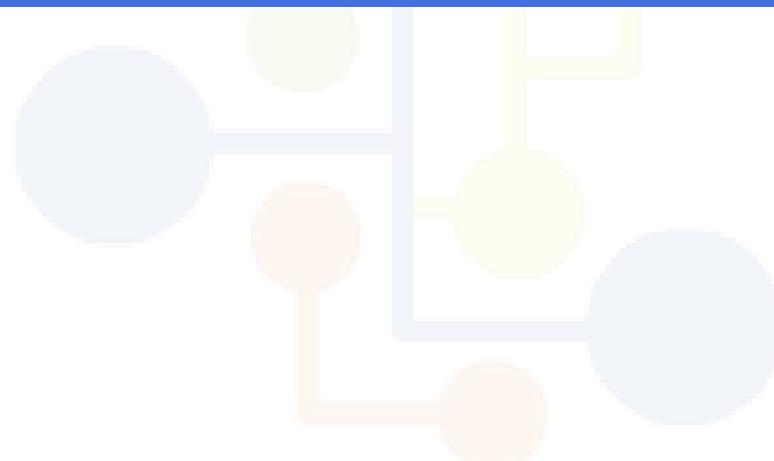


O coeficiente de variação nos dá uma **medida comum-percentagem** – que podemos usar para comparar 2 conjuntos de dados.



Data Science Academy

Exemplo



Data Science Academy



Data Science Academy

www.datascienceacademy.com.br

Como calculamos o Coeficiente de Variação = CV

$$CV = \frac{s}{x} \times 100$$

Onde: **S** = Desvio Padrão
X = Média



Data Science Academy

Os seguintes dados foram coletados:



$$\text{Nike} \rightarrow \text{CV} = S / x \times 100 = \$5.10 / \$55.62 \times 100 = 9.2\%$$



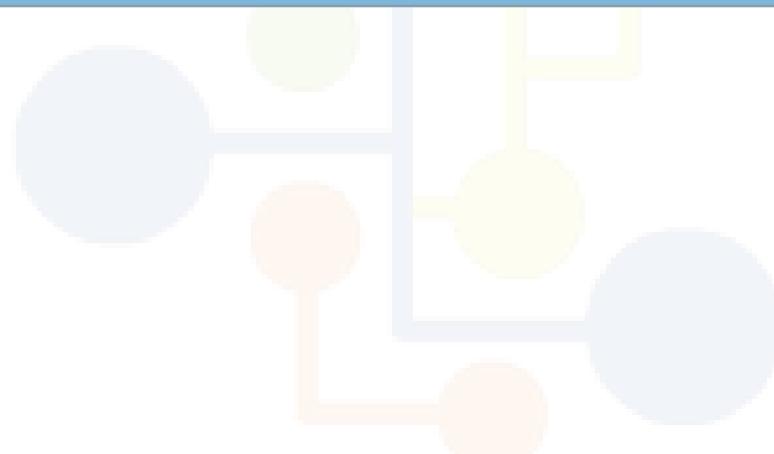
$$\text{Adidas} \rightarrow \text{CV} = S / x \times 100 = \$3.60 / \$24.86 \times 100 = 14.5\%$$



Data Science Academy



Conclusão



Data Science Academy



$$\text{Nike} \rightarrow \text{CV} = S / x (100) = \$5.10 / \$55.62 (100) = 9.2\%$$



$$\text{Adidas} \rightarrow \text{CV} = S / x (100) = \$3.60 / \$24.86 (100) = 14.5\%$$



Data Science Academy

Um investidor se sentiria mais seguro em adquirir ações da **Nike**, pois o preço das ações teria uma variação menor, podendo assim evitar perdas e permitindo ao investidor ter um investimento mais seguro.



Data Science Academy

Esse tópico chegou ao final

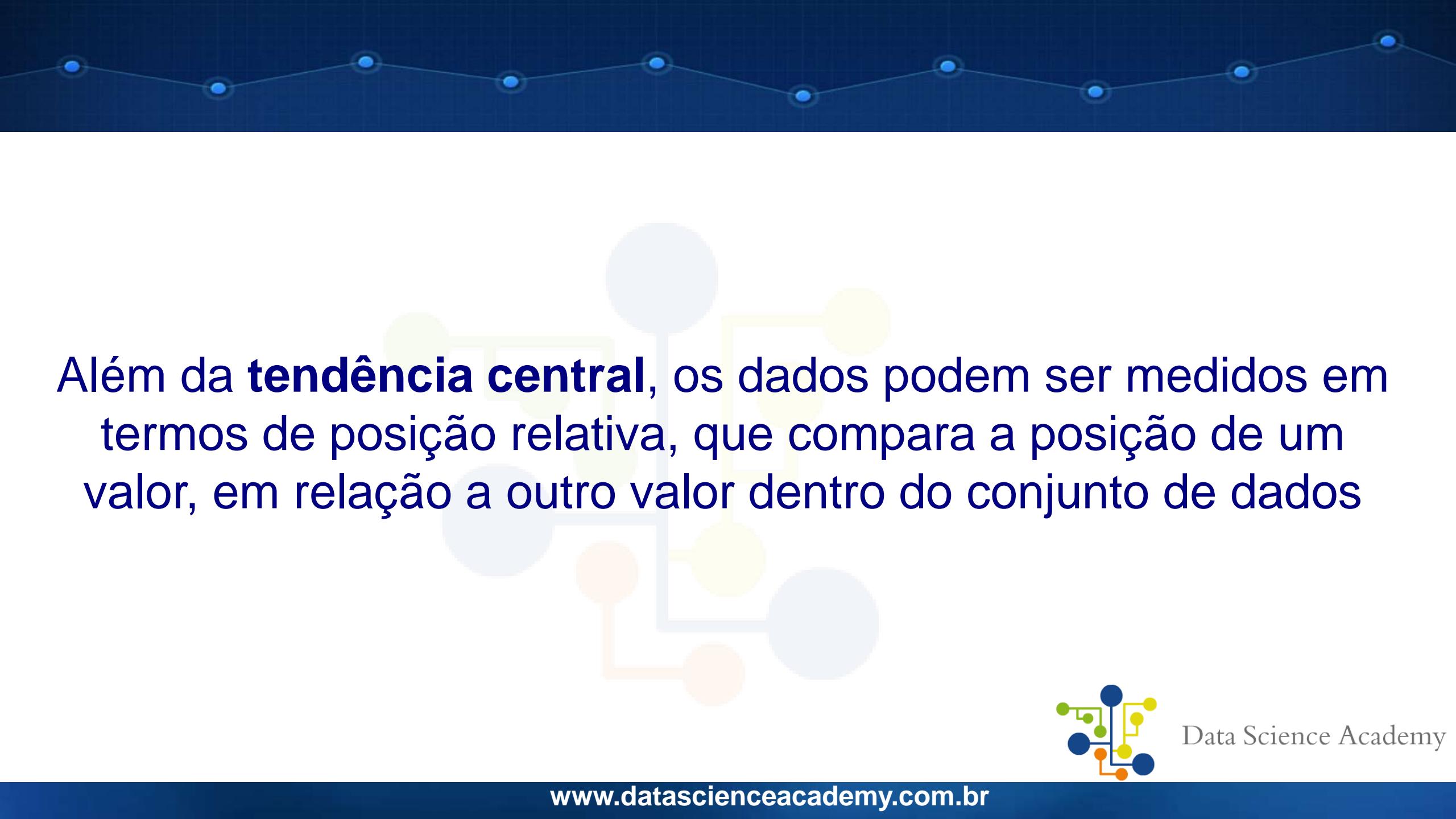


Data Science Academy

Medidas de Posição Relativa



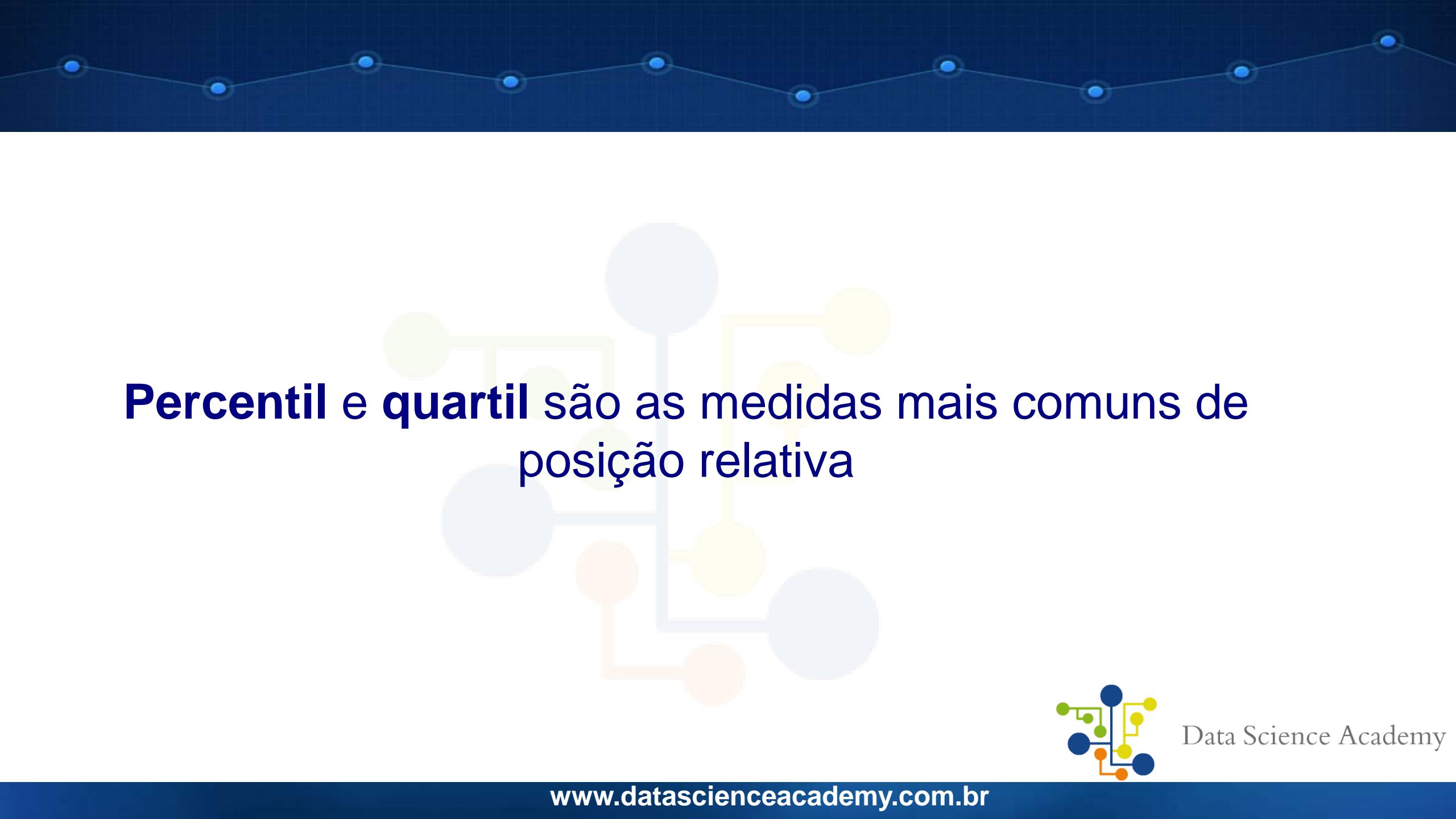
Data Science Academy



Além da **tendência central**, os dados podem ser medidos em termos de posição relativa, que compara a posição de um valor, em relação a outro valor dentro do conjunto de dados



Data Science Academy



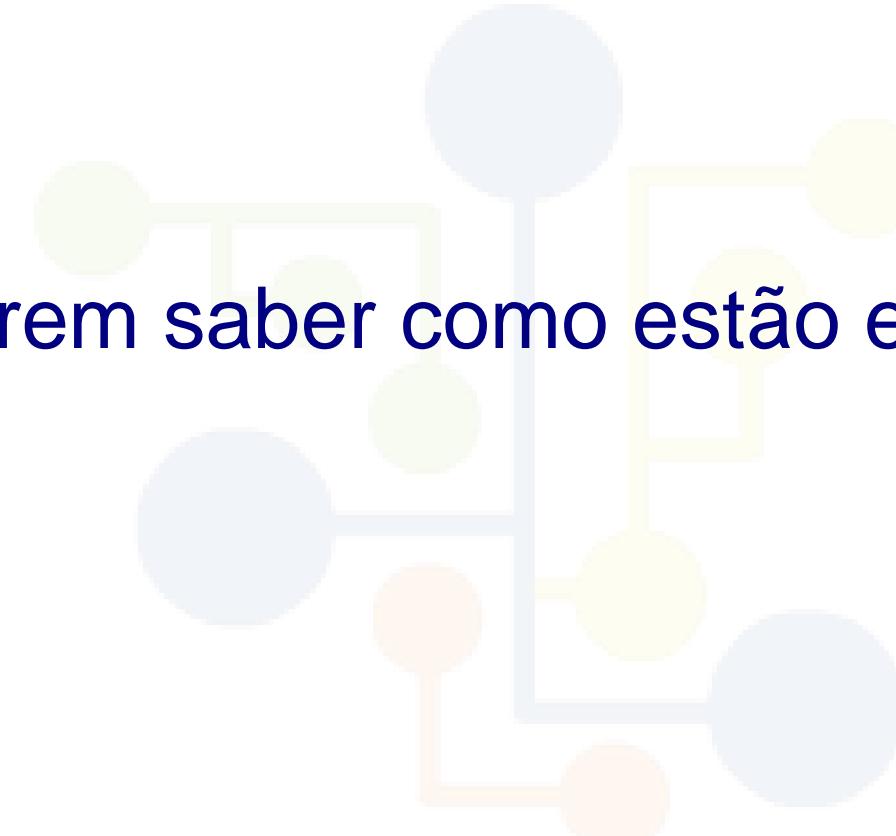
Percentil e quartil são as medidas mais comuns de posição relativa



Data Science Academy



Todos querem saber como estão em relação aos outros

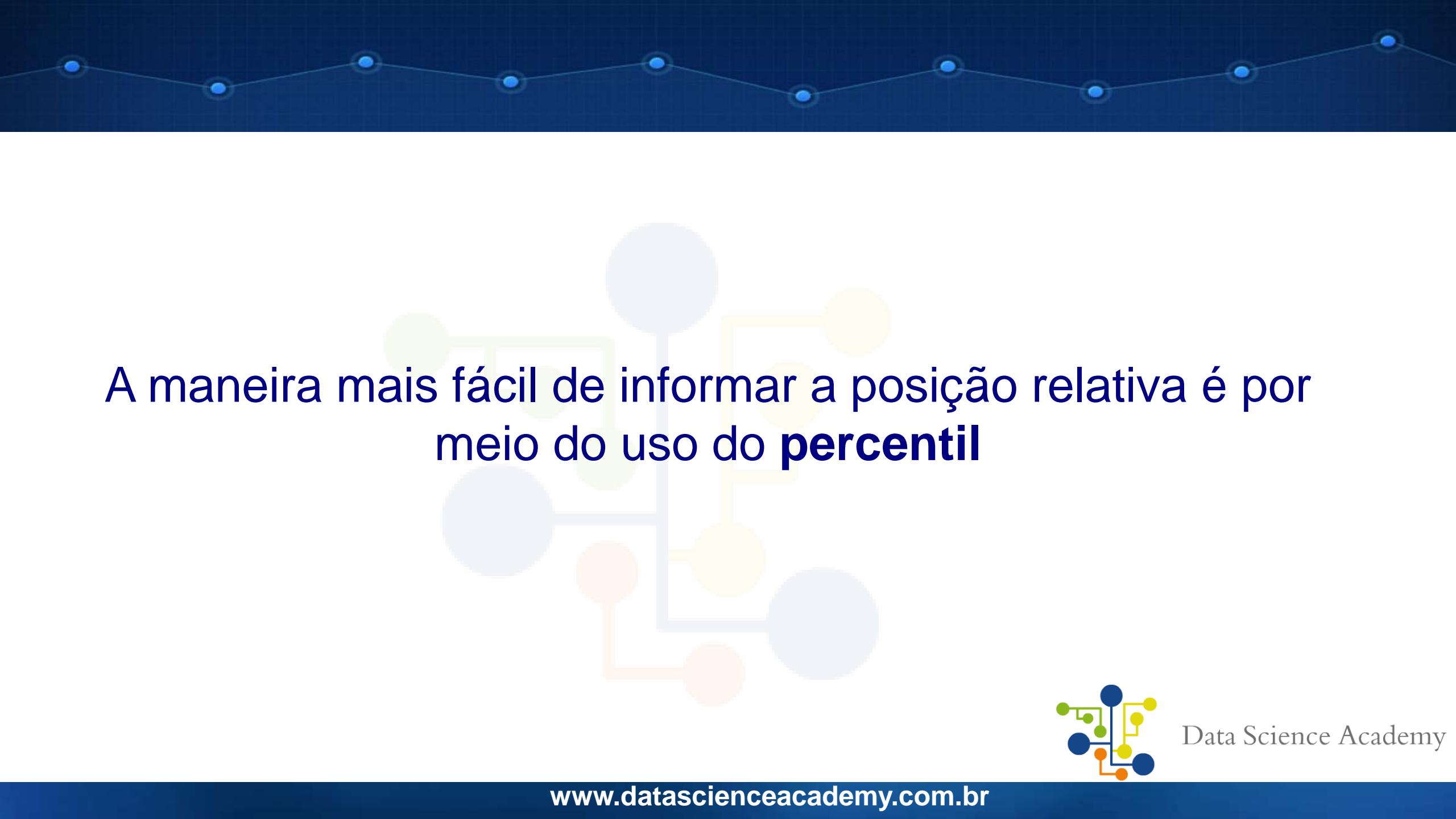


Data Science Academy

Na escola, a nota que um aluno consegue em uma prova não é tão importante, como essa nota se compara com a das outras crianças



Data Science Academy



A maneira mais fácil de informar a posição relativa é por meio do uso do **percentil**



Data Science Academy



Percentil e Porcentagem são a mesma coisa?



Data Science Academy

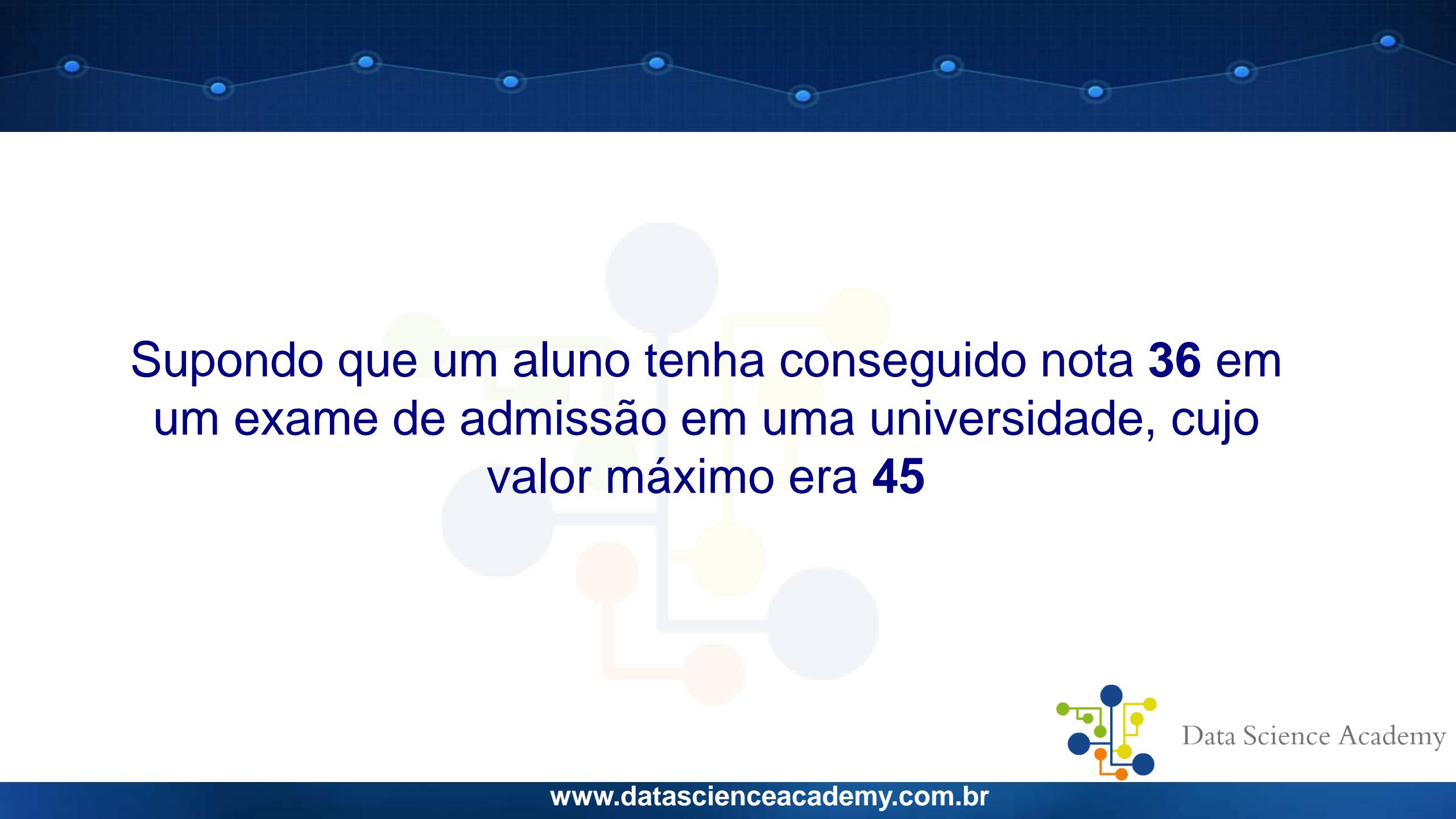
Percentil e Porcentagem **não** são a mesma coisa.

Percentil

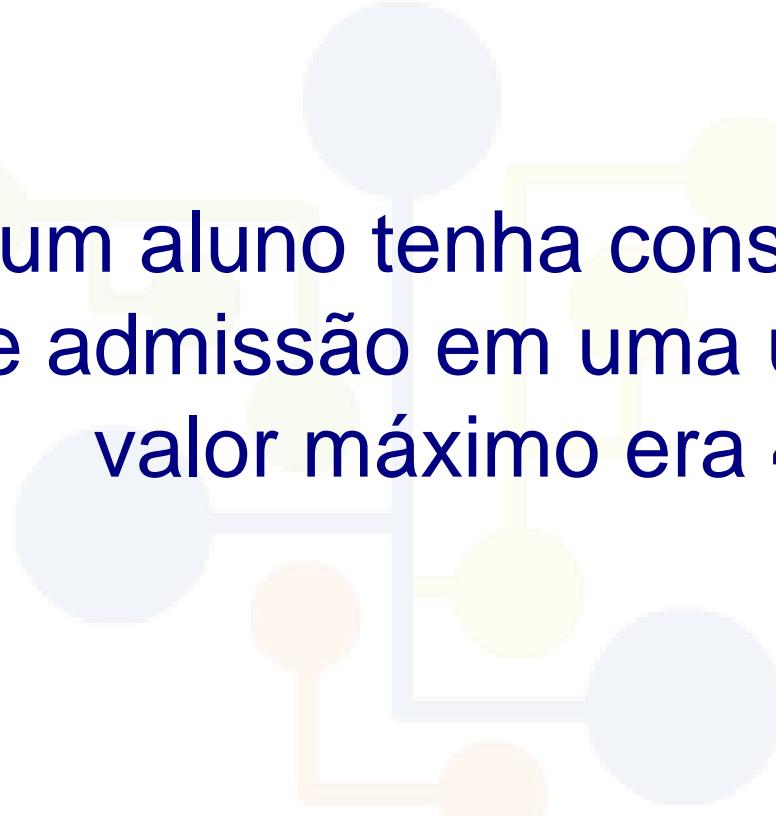
Porcentagem



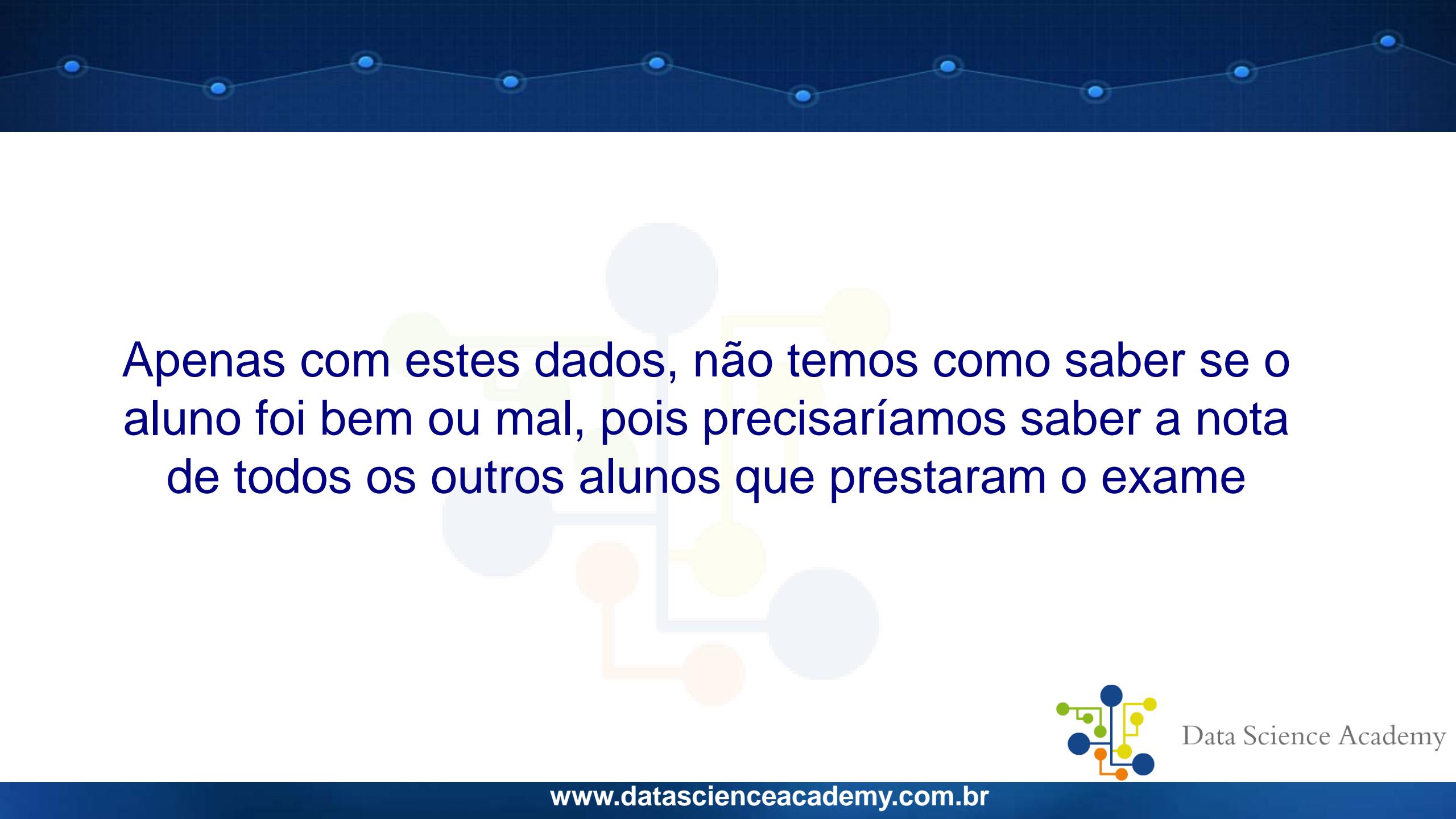
Data Science Academy



Supondo que um aluno tenha conseguido nota **36** em um exame de admissão em uma universidade, cujo valor máximo era **45**



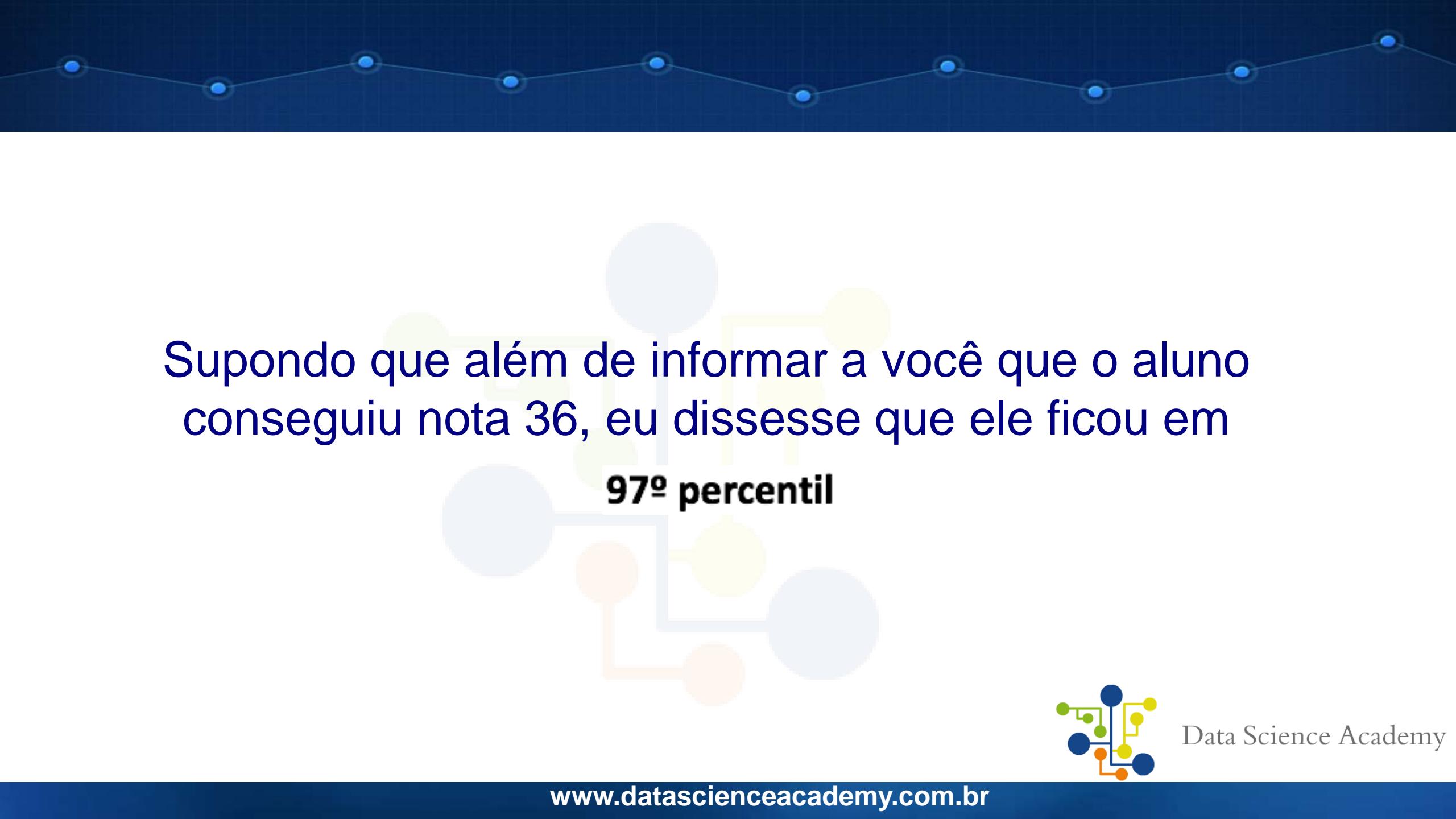
Data Science Academy



Apenas com estes dados, não temos como saber se o aluno foi bem ou mal, pois precisaríamos saber a nota de todos os outros alunos que prestaram o exame



Data Science Academy

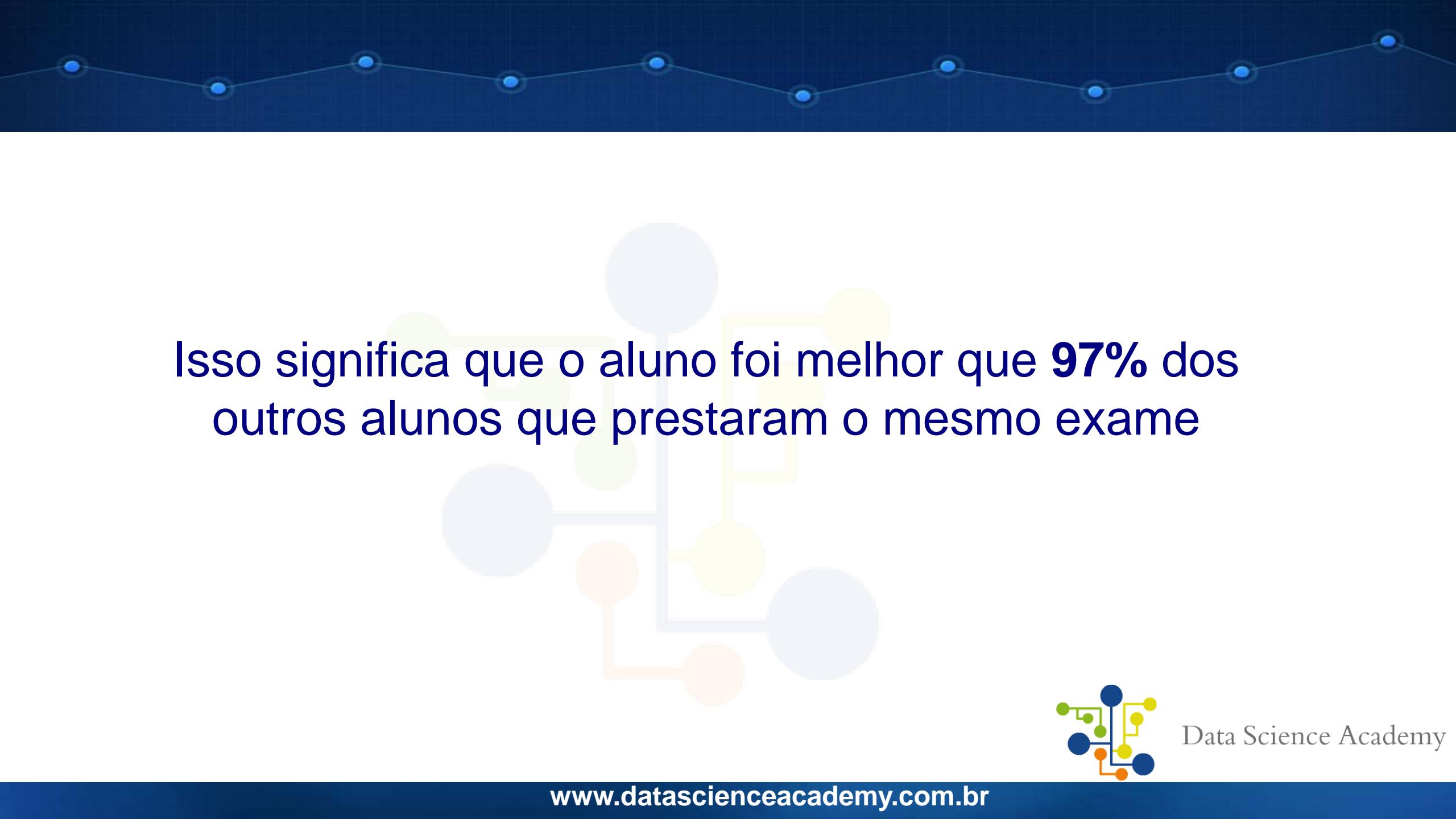


Supondo que além de informar a você que o aluno conseguiu nota 36, eu dissesse que ele ficou em

97º percentil



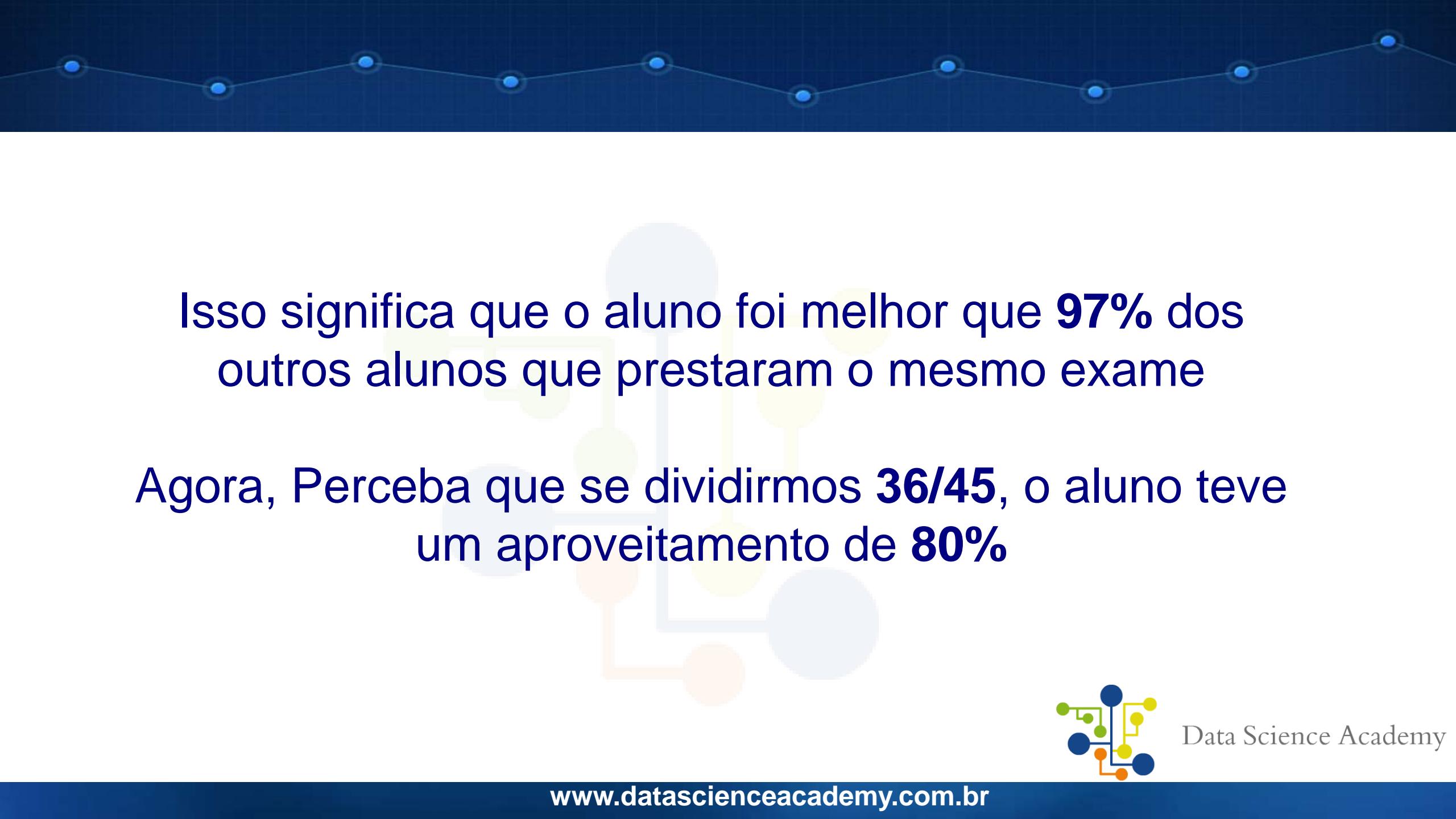
Data Science Academy



Isso significa que o aluno foi melhor que **97%** dos outros alunos que prestaram o mesmo exame



Data Science Academy



Isso significa que o aluno foi melhor que **97%** dos outros alunos que prestaram o mesmo exame

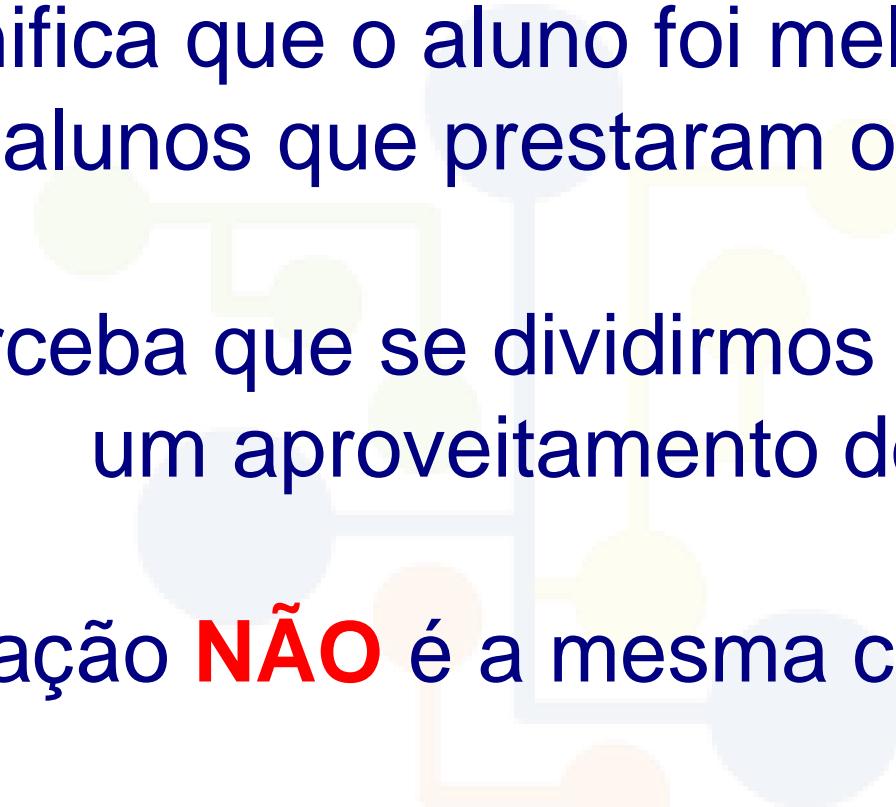
Agora, Perceba que se dividirmos **36/45**, o aluno teve um aproveitamento de **80%**



Data Science Academy



Isso significa que o aluno foi melhor que **97%** dos outros alunos que prestaram o mesmo exame

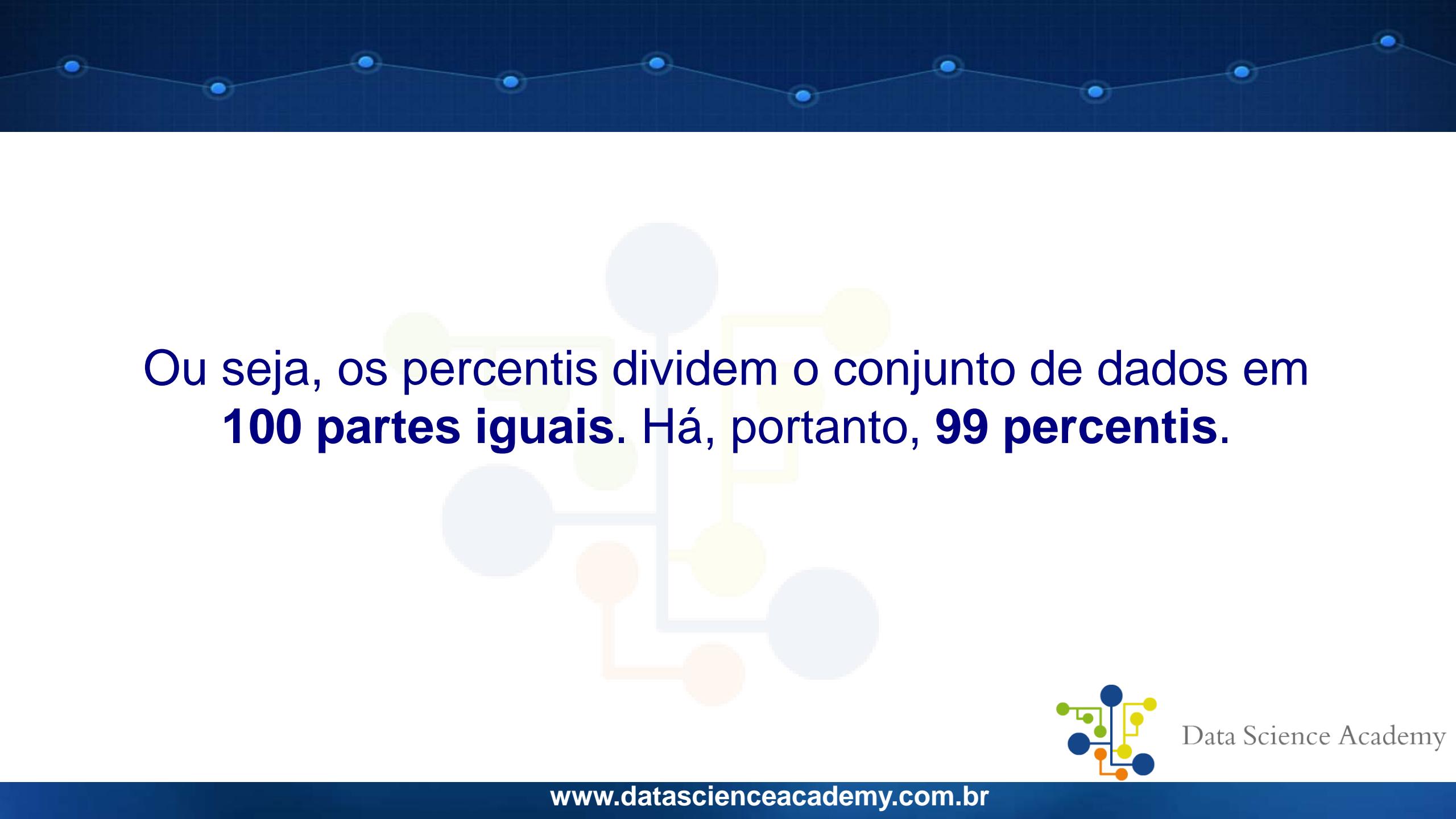


Agora, Perceba que se dividirmos **36/45**, o aluno teve um aproveitamento de **80%**

Esta informação **NÃO** é a mesma coisa que o **percentil**



Data Science Academy

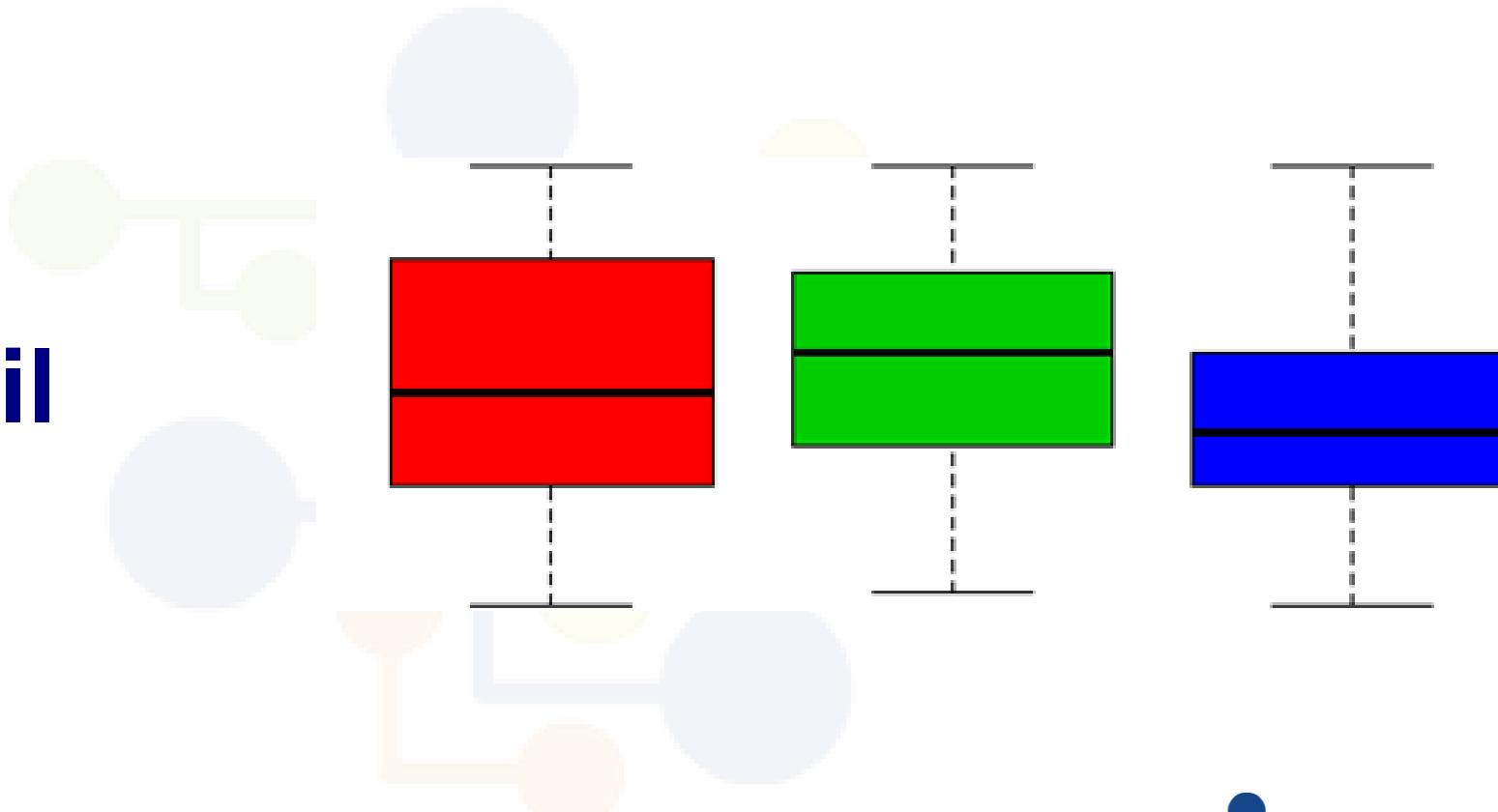


Ou seja, os percentis dividem o conjunto de dados em **100 partes iguais**. Há, portanto, **99 percentis**.

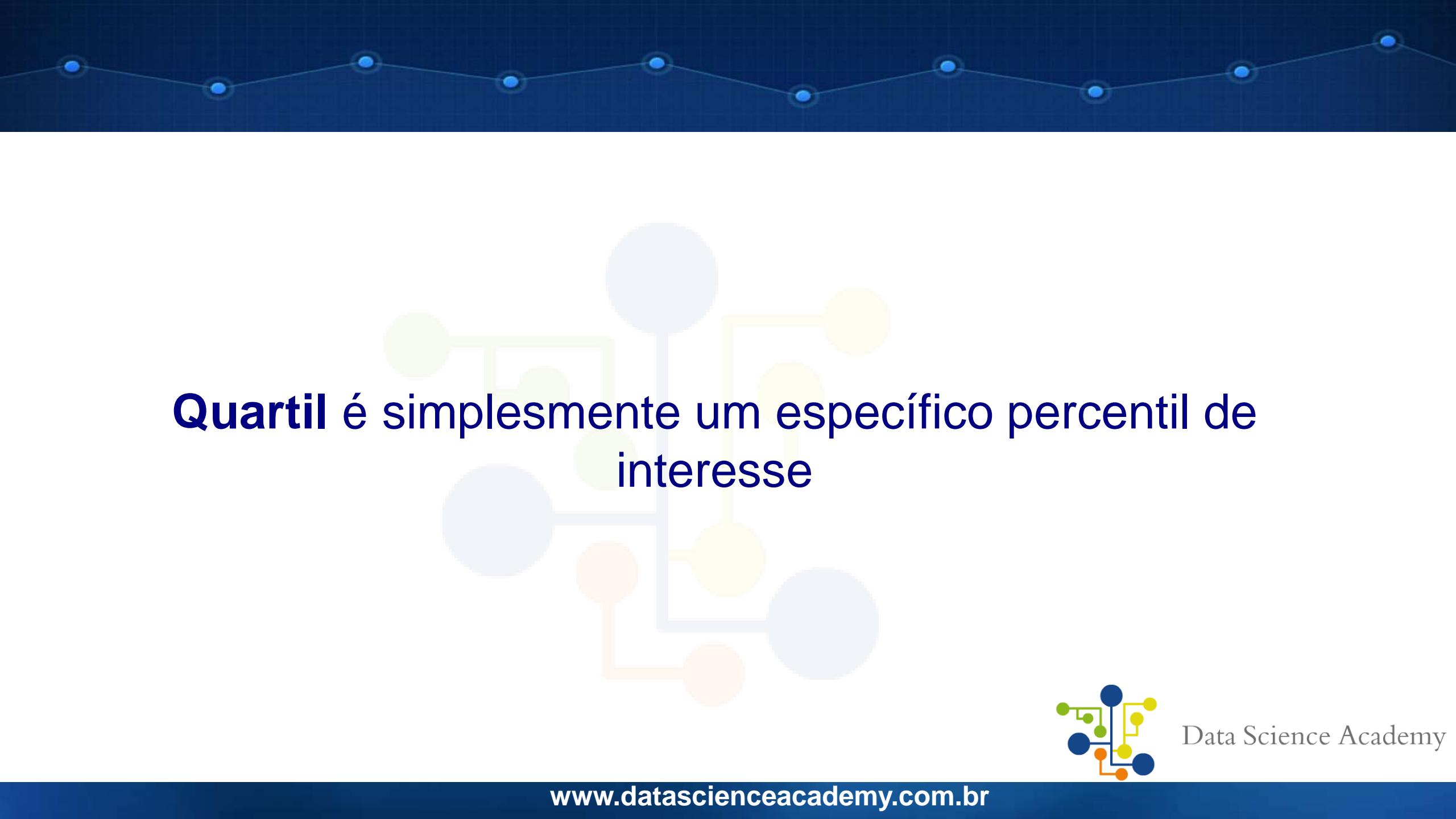


Data Science Academy

Quartil



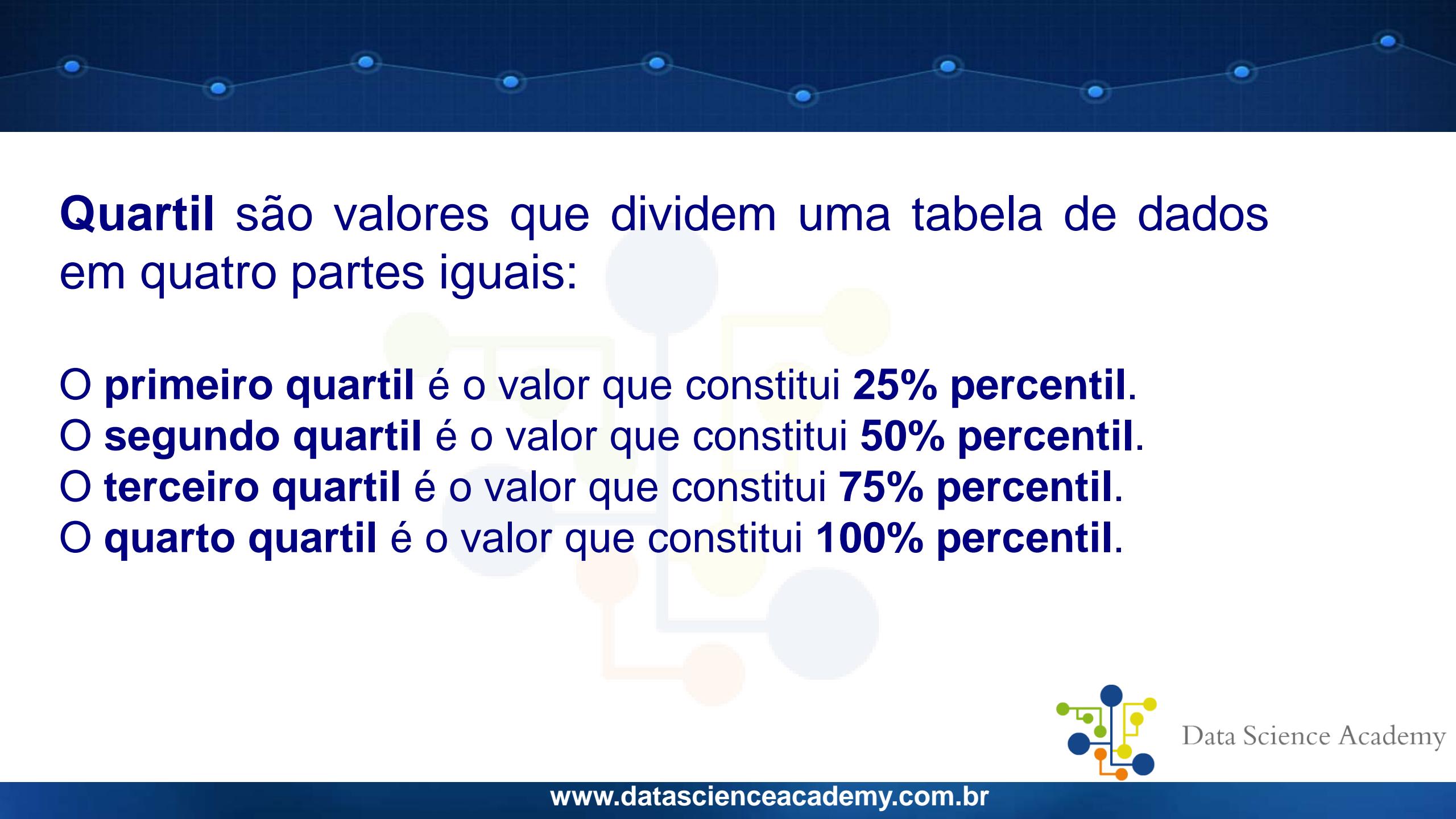
Data Science Academy



Quartil é simplesmente um específico percentil de interesse



Data Science Academy



Quartil são valores que dividem uma tabela de dados em quatro partes iguais:

O **primeiro quartil** é o valor que constitui **25% percentil**.

O **segundo quartil** é o valor que constitui **50% percentil**.

O **terceiro quartil** é o valor que constitui **75% percentil**.

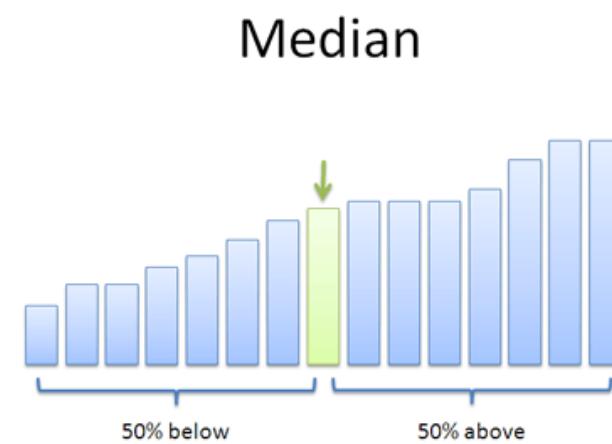
O **quarto quartil** é o valor que constitui **100% percentil**.



Data Science Academy

Perceba que o **segundo quartil** é a **mediana**

Ou seja, **50º percentil**



Data Science Academy

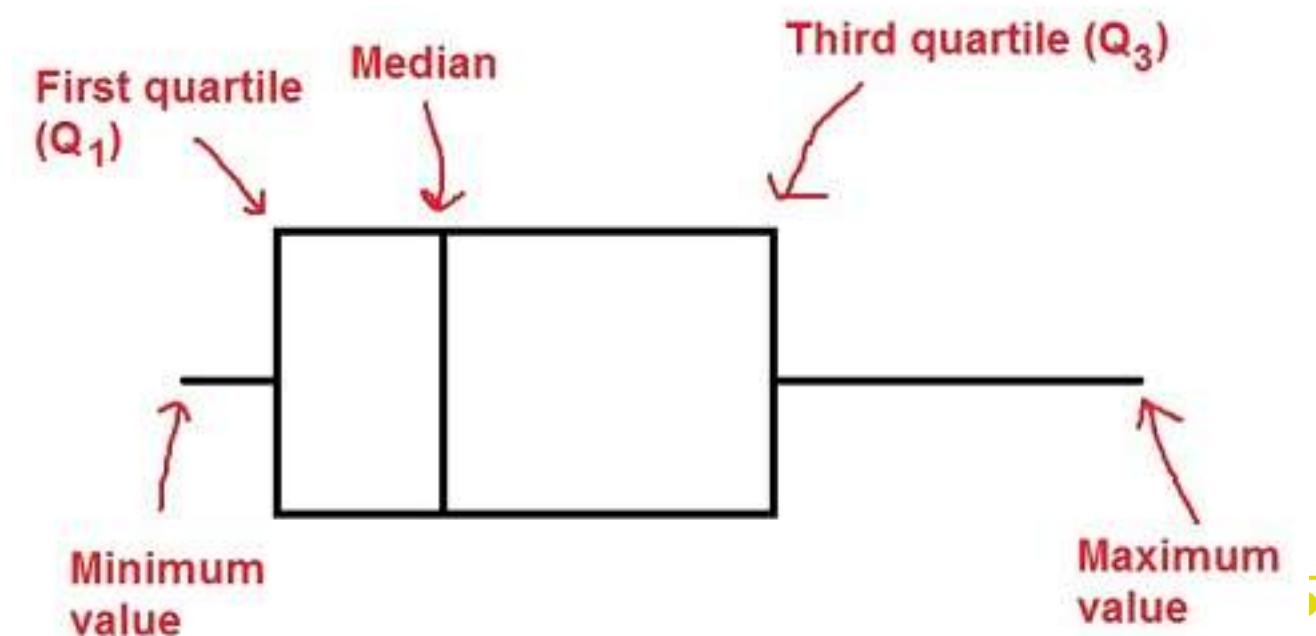
Temos ainda os intervalos interquartis:

- Intervalo interquartil $\rightarrow Q_3 - Q_1$
- Intervalo semi-interquartil $\rightarrow (Q_3 - Q_1)/2$
- Quartil médio $\rightarrow (Q_3 + Q_1)/2$



Data Science Academy

Os intervalos interquartis são fundamentais para saber interpretar um boxplot:



Esse tópico chegou ao final

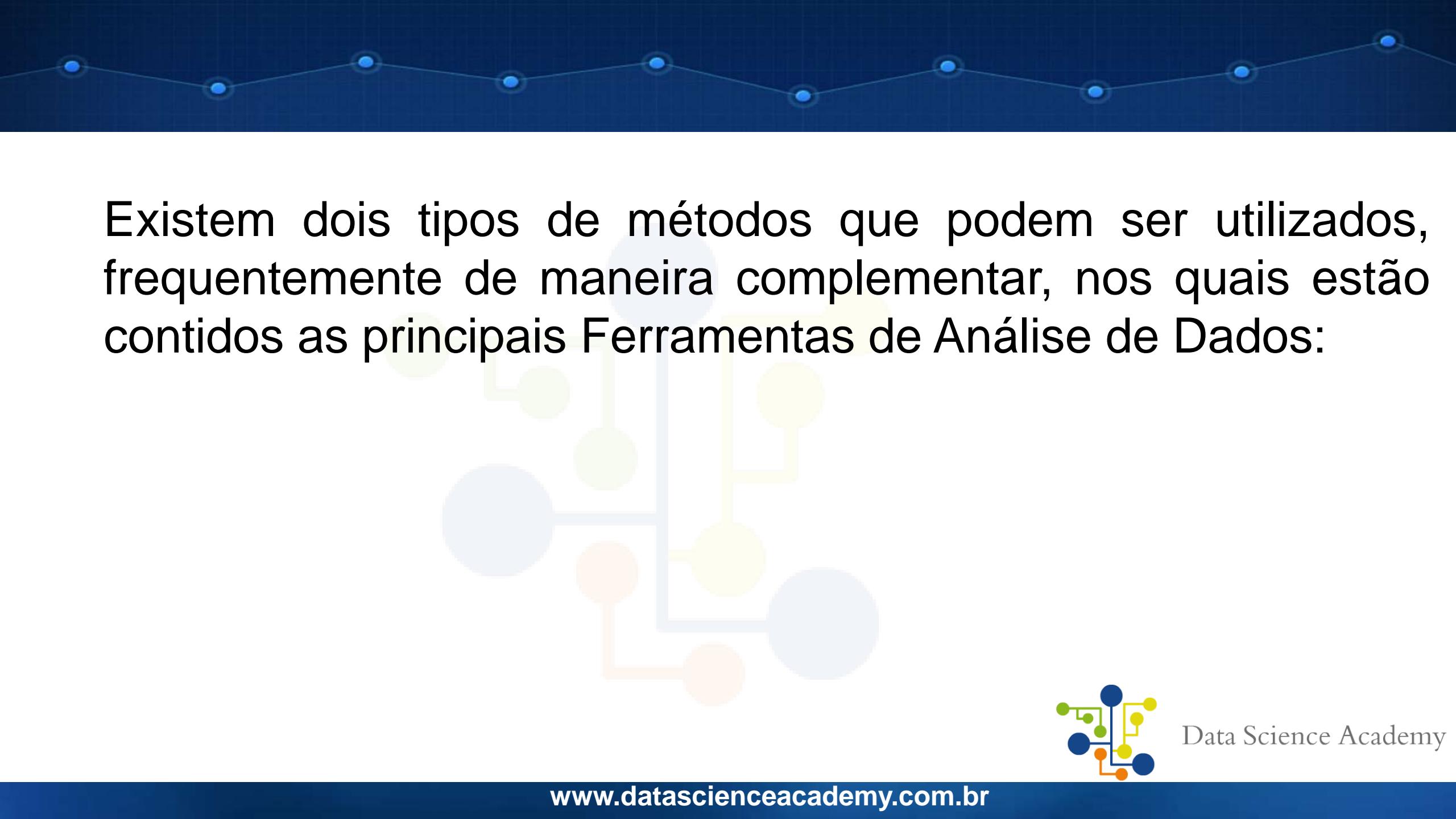


Data Science Academy

Ferramentas de Análise de Dados



Data Science Academy



Existem dois tipos de métodos que podem ser utilizados, frequentemente de maneira complementar, nos quais estão contidos as principais Ferramentas de Análise de Dados:



Data Science Academy

Existem dois tipos de métodos que podem ser utilizados, frequentemente de maneira complementar, nos quais estão contidos as principais Ferramentas de Análise de Dados:



- Métodos Gráficos ou Tabulares

$$CV = \frac{s}{x} \times 100$$

- Métodos Numéricos



Data Science Academy

- Tabela de Frequência
- Tabela de Contingência
- Gráficos de Linhas
- Gráficos de Barras
- Gráfico de Pareto
- Histogramas
- Gráficos de Caixa (box-plots)
- Diagramas de dispersão
- Gráfico Temporal
- Ogiva (frequência cumulativa)
- Ramos e folhas
- Gráficos de Pontos
- Gráfico de Quartis

Métodos Gráficos ou Tabulares



Data Science Academy

Métodos Numéricos

$$CV = \frac{s}{x} \times 100$$

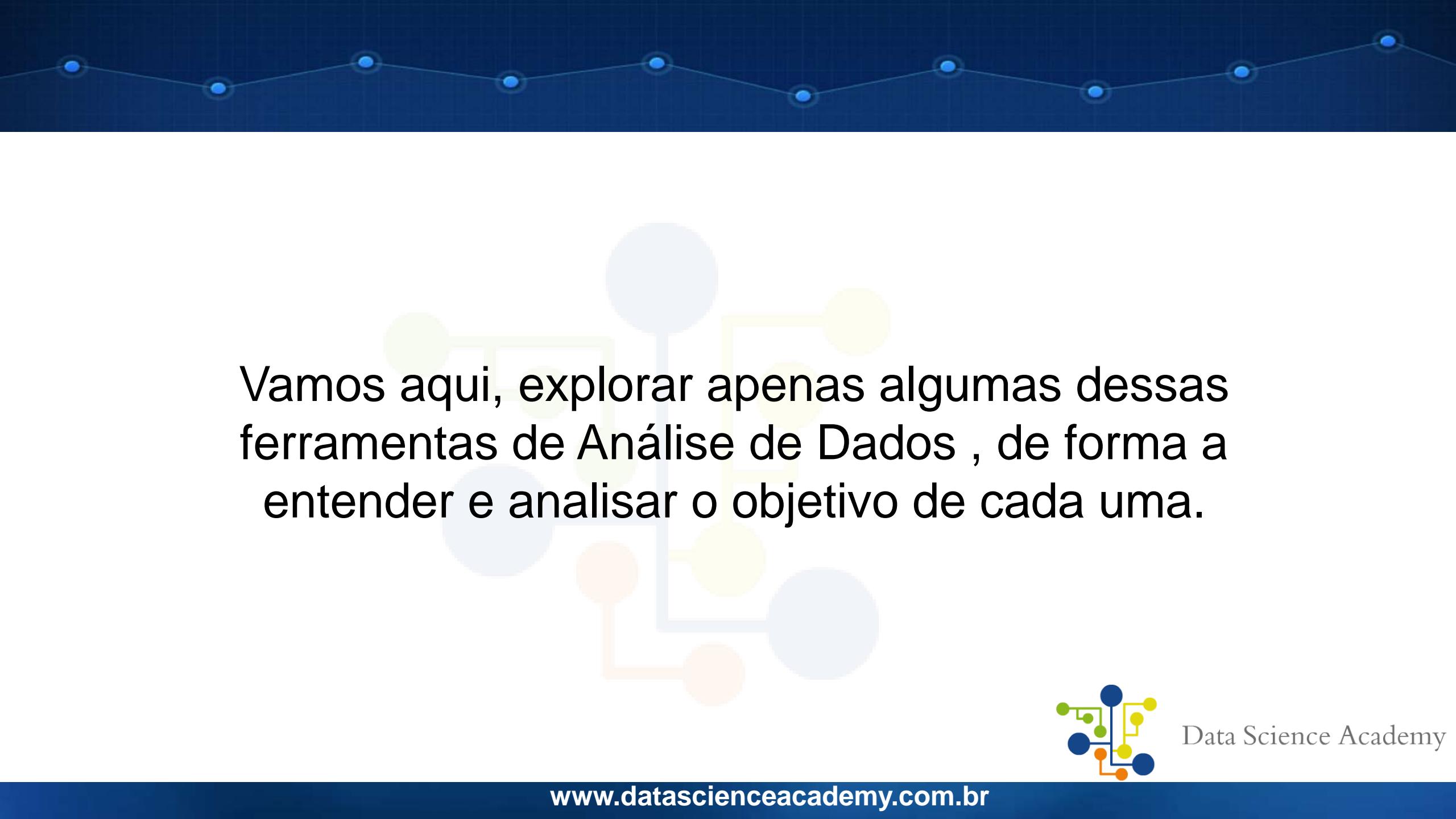
- Média
- Mediana
- Moda
- Quartis
- Desvio Padrão
- Variância
- Intervalo Interquartil
- Coeficiente de Variação
- Coeficiente de Assimetria
- Curtose
- Coeficiente de Correlação Linear
- Covariância
- Coeficientes de Associação



Data Science Academy

A técnica utilizada vai depender do tipo de variável





Vamos aqui, explorar apenas algumas dessas ferramentas de Análise de Dados , de forma a entender e analisar o objetivo de cada uma.



Data Science Academy

Tabela de Frequência

Indica a frequência observada ,ou seja, mostra a frequência com que cada observação aparece nos dados.

- Frequência absoluta
- Frequência relativa



Data Science Academy

Tabela de Frequência

- Frequência absoluta – número de eventos observados.
- Frequência relativa – dada em porcentagem (ou como fração).

<i>Exercício</i>	<i>frequência absoluta</i>	<i>frequência relativa</i>
nenhum	185	40,04%
mudando	213	46,10%
baixo/ moderado	49	10,61%
alto	15	3,25%



Data Science Academy

Tabela de Frequência

- Frequência cumulativa – frequência absoluta ou relativa até um determinado ponto e não apenas em um valor.

<i>Exercício</i>	<i>frequência absoluta</i>	<i>frequência relativa</i>
nenhum	185	40,04%
mudando	213	46,10%
baixo/ moderado	49	10,61%
alto	15	3,25%

<i>Exercício</i>	<i>frequência absoluta</i>	<i>frequência relativa</i>
nenhum	185	40,04%
mudando	398	86,15%
baixo/ moderado	447	96,75%
alto	462	100,00%



Tabela de Contingência

São usadas para analisar dados com mais de uma variável envolvida.

Sexo	Grau de Instrução							Total
	1º Grau	2º Grau	3º Grau Incompleto	3º Grau Completo	Pós-Graduação	Mestrado	Doutorado	
Masculino (M)	200	250	650	740	150	40	15	2045
Feminino (F)	310	560	800	900	270	80	35	2955
Total	510	810	1450	1640	420	120	50	5000



Data Science Academy

Analizando uma Tabela de Contingência

- Conheça a diferença entre as porcentagens e números totais. As porcentagens são frequentemente a estatística mais adequada para a comparação entre diferentes resultados.

Sexo	Grau de Instrução							Total
	1º Grau	2º Grau	3º Grau Incompleto	3º Grau Completo	Pós-Graduação	Mestrado	Doutorado	
Masculino (M)	200	250	650	740	150	40	15	2045
Feminino (F)	310	560	800	900	270	80	35	2955
Total	510	810	1450	1640	420	120	50	5000



Data Science Academy

Analizando uma Tabela de Contingência

- Certifique-se, com relação aos dados numéricos, se os grupos da tabela não se sobrepõem e que estejam divididos de maneira equilibrada para se chegar a uma comparação imparcial.

Sexo	Grau de Instrução							Total
	1º Grau	2º Grau	3º Grau Incompleto	3º Grau Completo	Pós-Graduação	Mestrado	Doutorado	
Masculino (M)	200	250	650	740	150	40	15	2045
Feminino (F)	310	560	800	900	270	80	35	2955
Total	510	810	1450	1640	420	120	50	5000



Data Science Academy

Analizando uma Tabela de Contingência

- Observe atentamente as unidades e como elas estão apresentadas na tabela.

Sexo	Grau de Instrução							Total
	1º Grau	2º Grau	3º Grau Incompleto	3º Grau Completo	Pós-Graduação	Mestrado	Doutorado	
Masculino (M)	200	250	650	740	150	40	15	2045
Feminino (F)	310	560	800	900	270	80	35	2955
Total	510	810	1450	1640	420	120	50	5000



Data Science Academy

Analizando uma Tabela de Contingência

■ Observe o modo como a informação é apresentada. Frequentemente, as tabelas são projetadas para diminuir a importância de certos pontos ao mesmo tempo em que enfatizam apenas os pontos que são convenientes para quem criou a tabela.

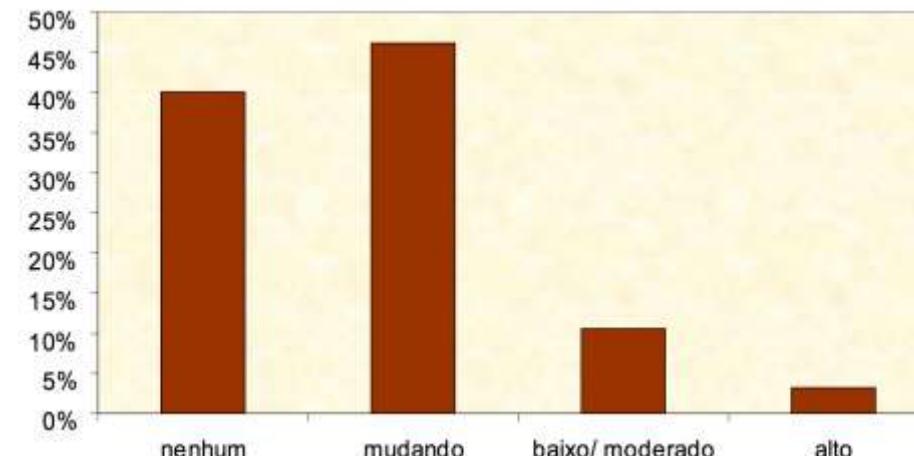
Sexo	Grau de Instrução							Total
	1º Grau	2º Grau	3º Grau Incompleto	3º Grau Completo	Pós-Graduação	Mestrado	Doutorado	
Masculino (M)	200	250	650	740	150	40	15	2045
Feminino (F)	310	560	800	900	270	80	35	2955
Total	510	810	1450	1640	420	120	50	5000



Data Science Academy

Gráfico de Barras

Apresenta a frequência absoluta ou relativa (NÃO cumulativa), ou seja, quantas observações, ou a fração de observações para um dado valor da variável em estudo (ou classe de valores). A altura das barras representa o que foi mais observado.

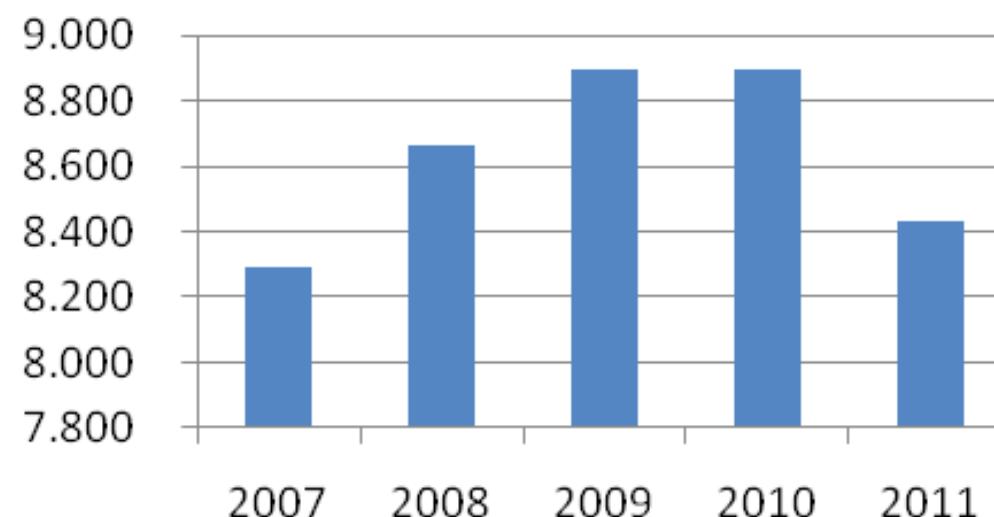


Data Science Academy

Gráfico de Barras

As barras podem ser:

Verticais: também denominado gráfico de colunas.

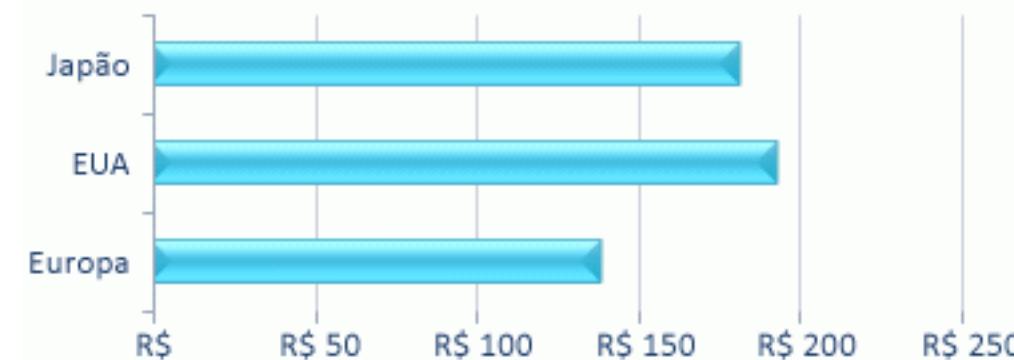


Data Science Academy

Gráfico de Barras

As barras podem ser:

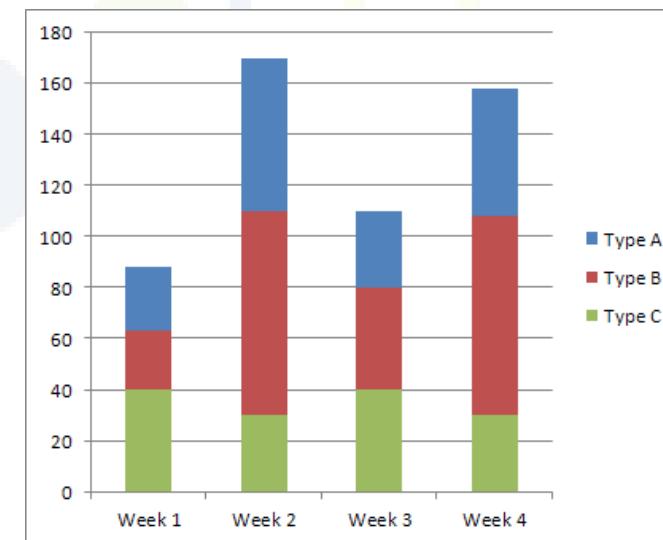
Horizontais: onde os valores ficam localizados no eixo das abscissas.



Data Science Academy

Gráfico de Barras Segmentado

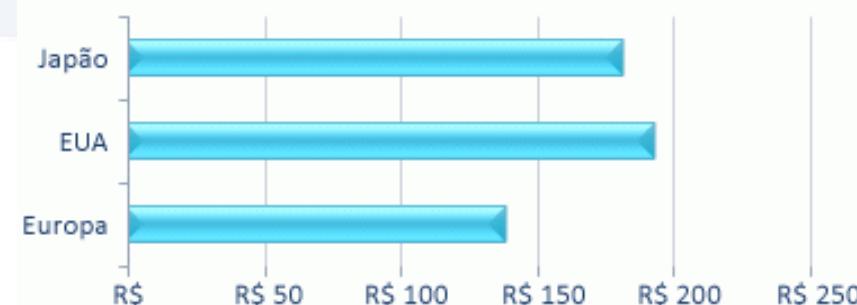
Permite agrupar ou categorizar os dados em subconjuntos. Representam os valores de outra variável que pode ser parte de um total do mesmo intervalo.



Data Science Academy

Analizando o Gráfico de Barras

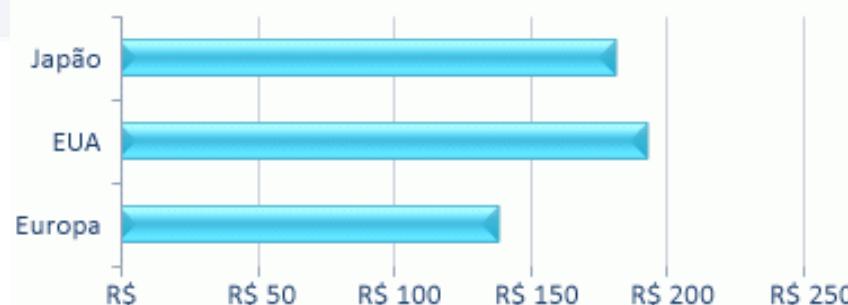
- Fique atento para a escala do gráfico de barras (as unidades na qual a altura das barras está representada) e determine se esta é uma representação apropriada da informação.



Data Science Academy

Analizando o Gráfico de Barras

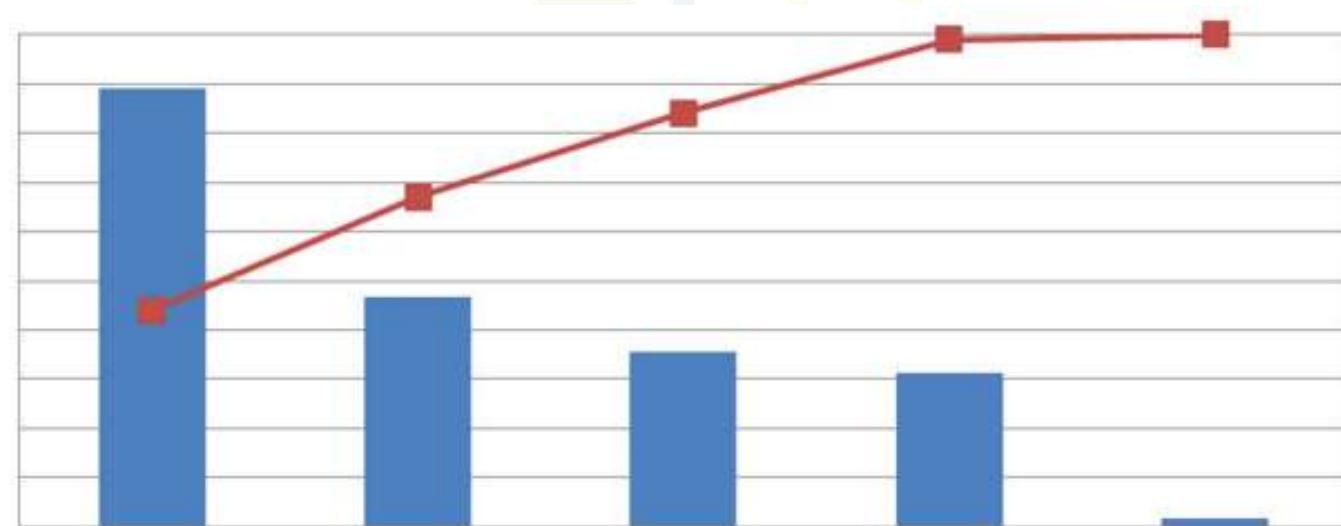
- Não pense que a informação contida em um gráfico de barras é tudo que você precisa saber; esteja pronto para ir mais fundo caso seja necessário.



Data Science Academy

Gráfico de Pareto

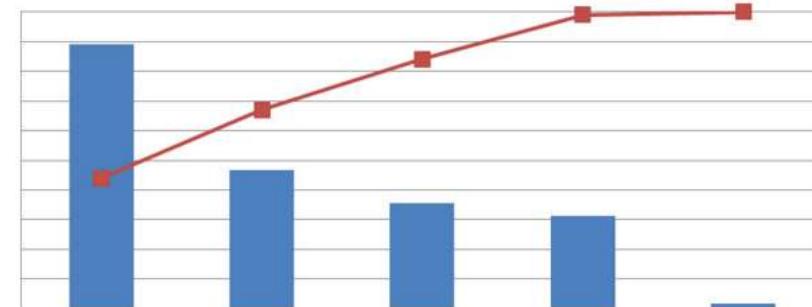
É um tipo de gráfico de barras que permite ordenar as frequências das ocorrências de modo a priorizar os problemas vitais e eliminar futuras perdas.



Data Science Academy

Analizando o Gráfico de Pareto

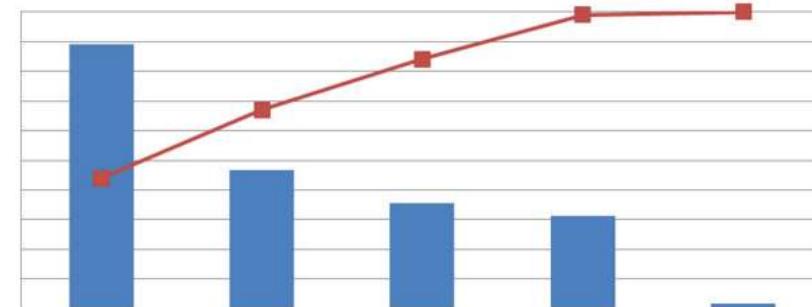
- Permite organizar as linhas na ordem decrescente de importância das causas, ou seja, a causa mais importante primeiro.



Data Science Academy

Analizando o Gráfico de Pareto

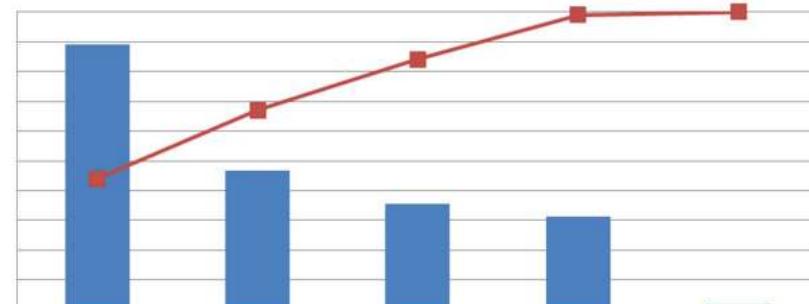
- Permite formar uma tabela listando as causas e sua frequência como uma porcentagem.



Data Science Academy

Analizando o Gráfico de Pareto

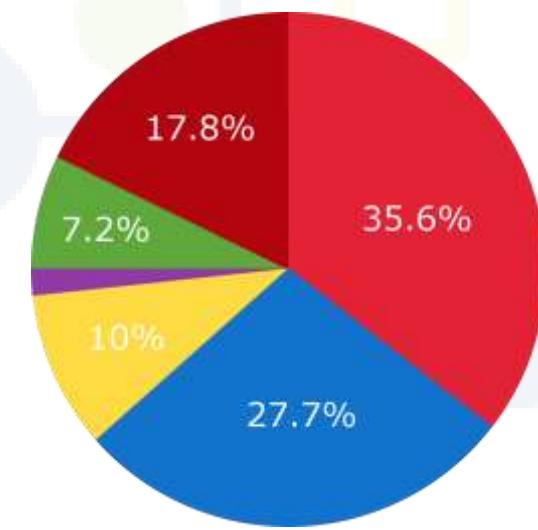
- Geralmente a linha do eixo Y forma uma curva. Esta linha tem a função de separar as causas importantes, que ficam à esquerda da linha, das causas menos importantes ou triviais, que ficam à direita.



Data Science Academy

Gráfico de Pizza

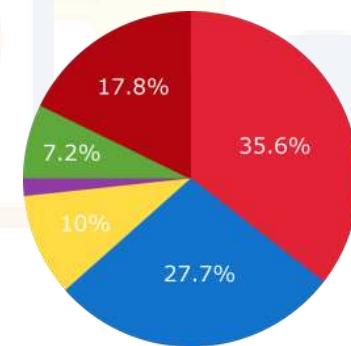
Gráficos de pizza não são ideais para visualizar comparações ou evoluções temporais. Use APENAS para fornecer a visualização de um caso, em um instante!



Data Science Academy

Analizando o Gráfico de Pizza

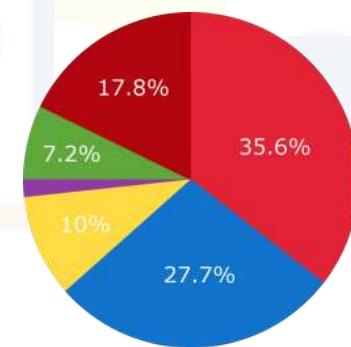
- Verifique se a soma das porcentagens é igual a 100% ou se é próximo desse valor (erros de arredondamentos devem ser muito pequenos).



Data Science Academy

Analizando o Gráfico de Pizza

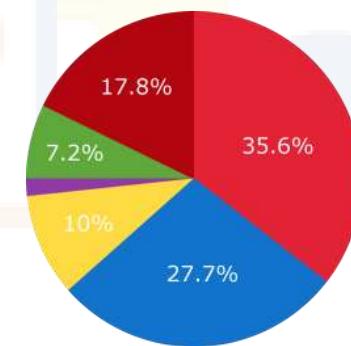
- Fique atento às fatias chamadas “outras/outros” que sejam maiores que várias outras fatias.



Data Science Academy

Analisando o Gráfico de Pizza

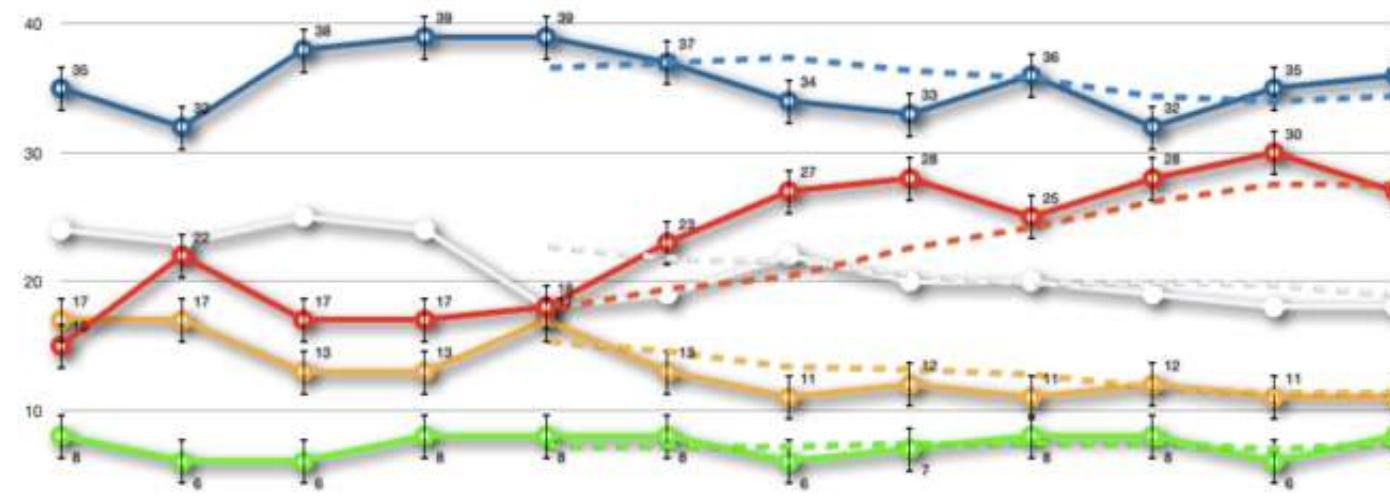
- Procure pela informação do número total de unidades, para que assim você possa determinar qual era o tamanho da pizza antes de ser dividida nas fatias que você está observando.



Data Science Academy

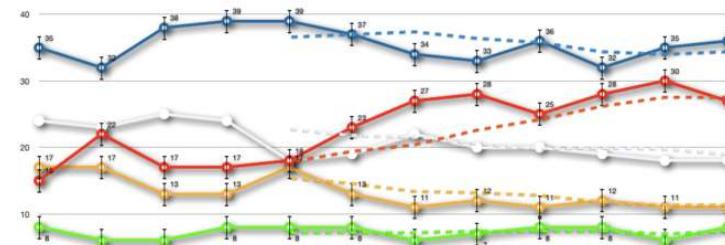
Gráfico de Linha

É um tipo de gráfico construído pela união dos pontos x e y formando uma espécie de reta, esse gráfico permite representar tendências e relacionamento entre variáveis.



Analizando o Gráfico de Linha

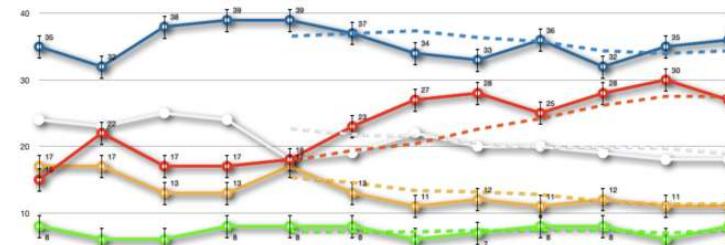
- Examine a escala do eixo vertical assim como o eixo horizontal. Os resultados podem parecer mais ou menos drásticos do que realmente são com apenas uma alteração da escala.



Data Science Academy

Analizando o Gráfico de Linha

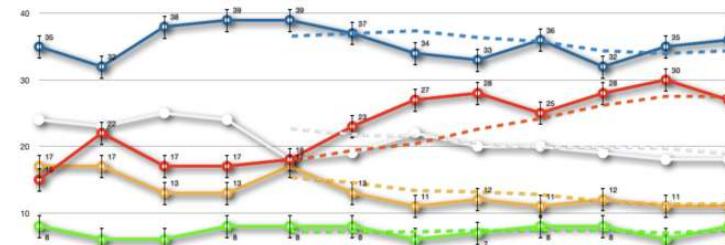
- Leve em consideração as unidades utilizadas no gráfico e assegure-se de que elas sejam apropriadas.



Data Science Academy

Analizando o Gráfico de Linha

- Um gráfico de linhas mostra apenas o que está acontecendo. O porquê de algo estar ocorrendo é uma outra história.

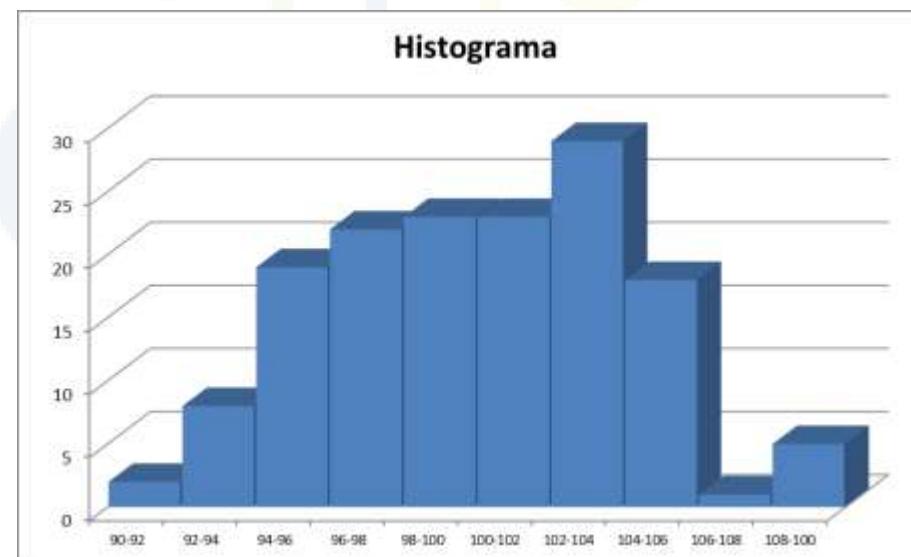


Data Science Academy

Histogramas

Variáveis quantitativas

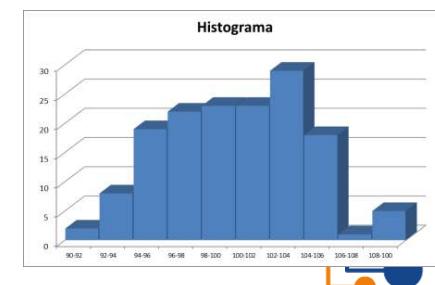
O propósito de um histograma, é oferecer uma descrição geral sobre os dados e não sobre os dados individualmente.



Data Science Academy

Analizando o Histograma

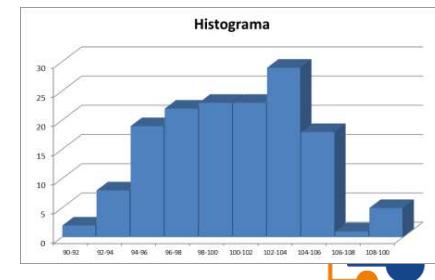
- Basicamente, o histograma pode ser usado para obter três características principais dos dados numéricos: o **modo** como os dados se **distribuem** (simétrico, distorcido para a direita, distorcido para a esquerda), a **variabilidade** encontrada nos dados e onde fica o **centro** dos dados.



Data Science Academy

Analizando o Histograma

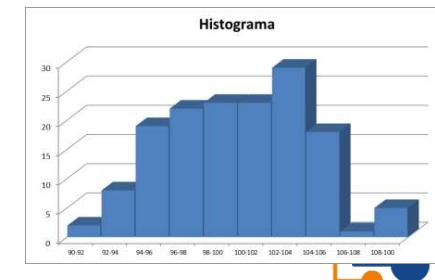
- Examine a escala utilizada para o eixo vertical (frequência ou frequência relativa) e tome cuidado com resultados que pareçam exagerados ou subestimados devido ao uso de escalas inadequadas.



Data Science Academy

Analizando o Histograma

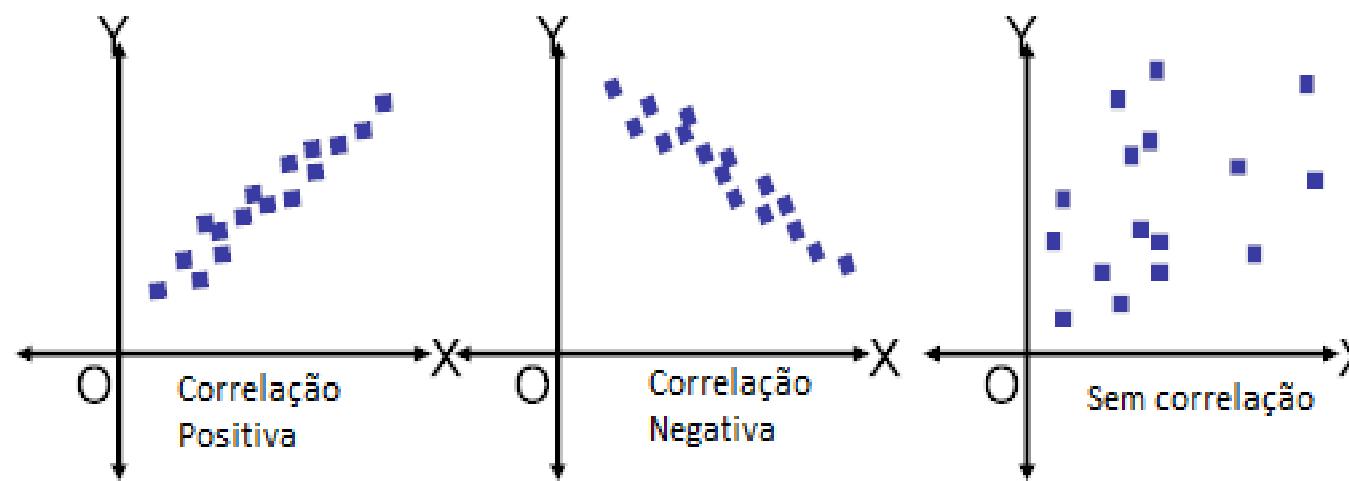
- Observe a escala utilizada para os grupos de variáveis numéricas no eixo horizontal. Se a variação para cada grupo for muito pequena, os dados podem parecer mais suaves do que realmente são.



Data Science Academy

Gráfico de Dispersão (Scatter Plot)

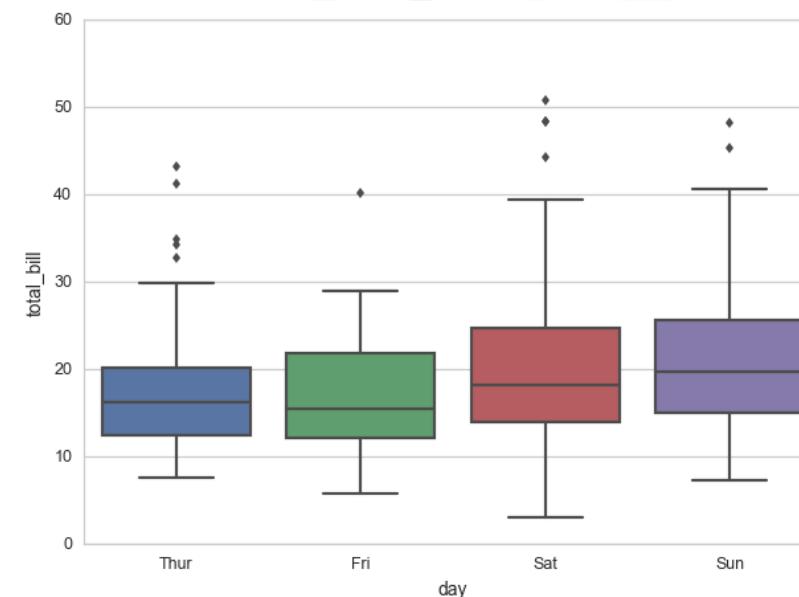
Mostra a relação entre duas **variáveis quantitativas**. Cada par observado de duas variáveis (x, y) é marcado como um ponto a partir de suas coordenadas.



Data Science Academy

Gráfico de Caixa (BoxPlot)

Exibe a distribuição de dados com base no resumo de cinco números: mínimo, primeiro quartil, mediana, terceiro quartil e máximo.



Data Science Academy

Gráfico Temporal ou Sequencial

Mostra a evolução de uma variável ao longo do tempo. É criado da mesma forma que o diagrama de dispersão, afinal é um diagrama de dispersão onde a variável x é o tempo.

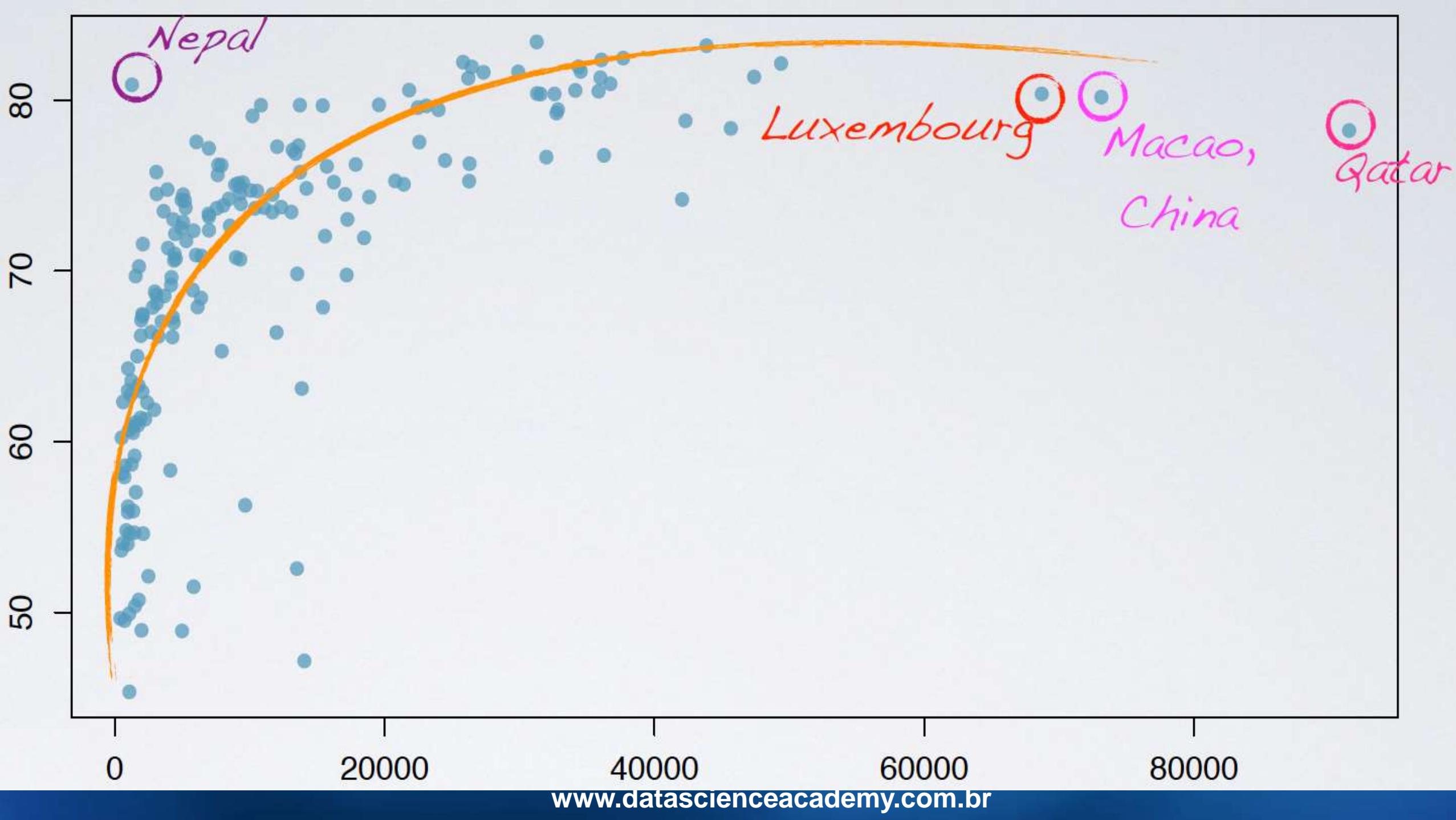


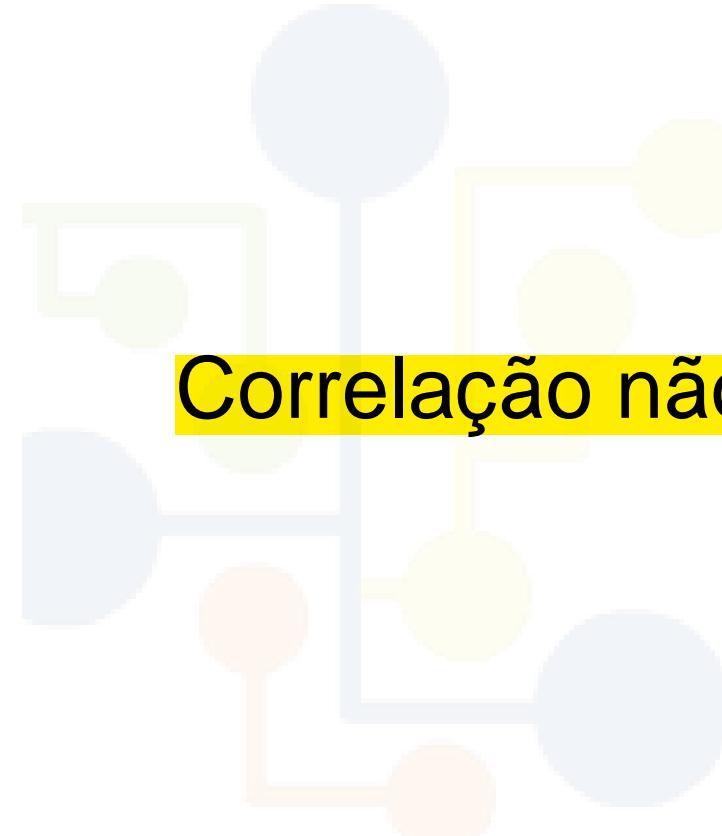
Data Science Academy



data	income per person (\$, 2012)	life expectancy (years, 2012)
Afghanistan	1359.7	60.254
Albania	6969.3	77.185
Algeria	6419.1	70.874
...
Zimbabwe	545.3	58.142

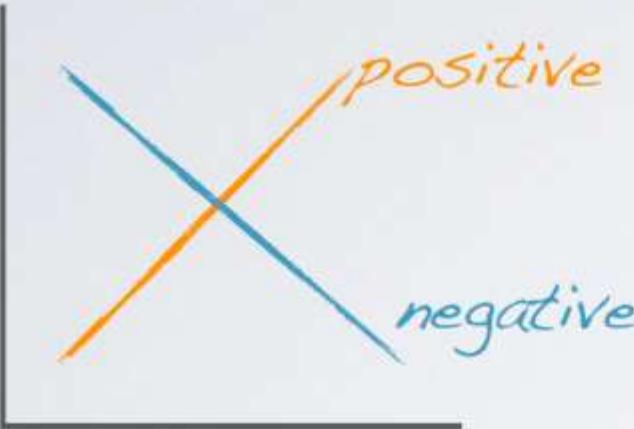




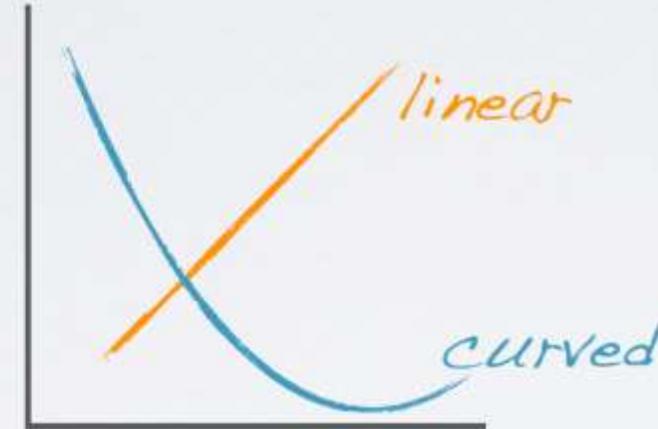


Data Science Academy

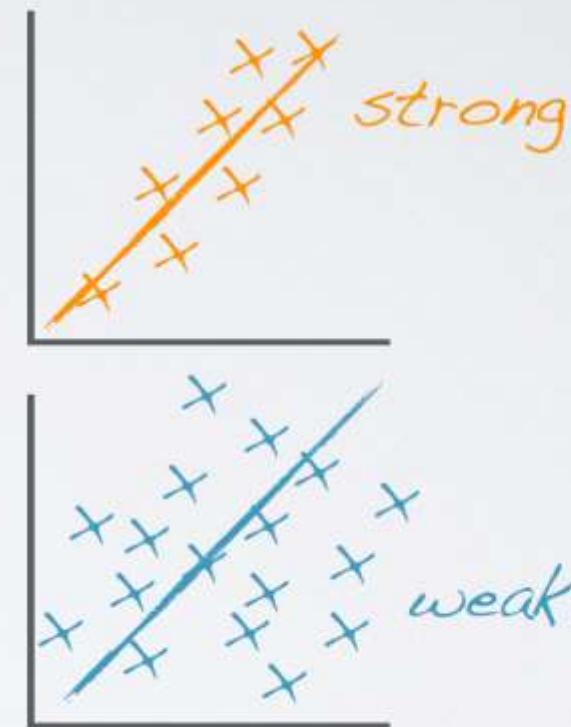
direction



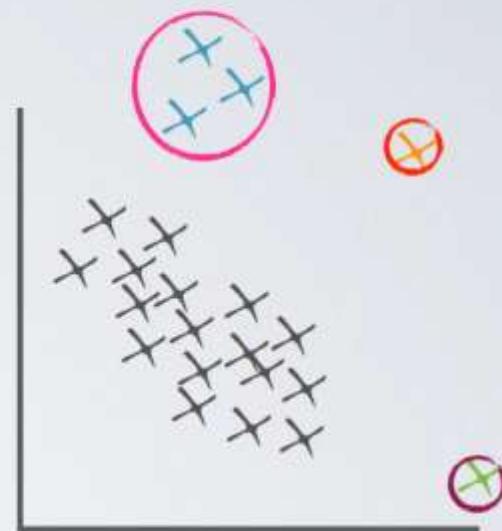
shape



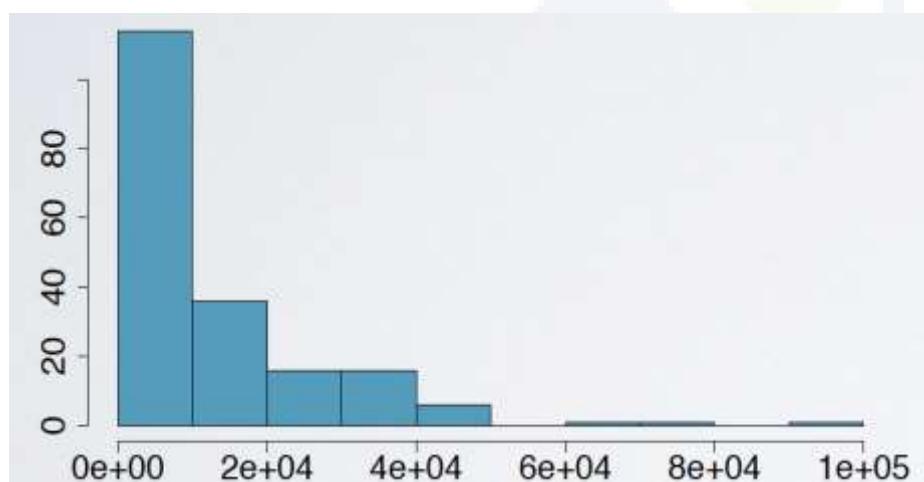
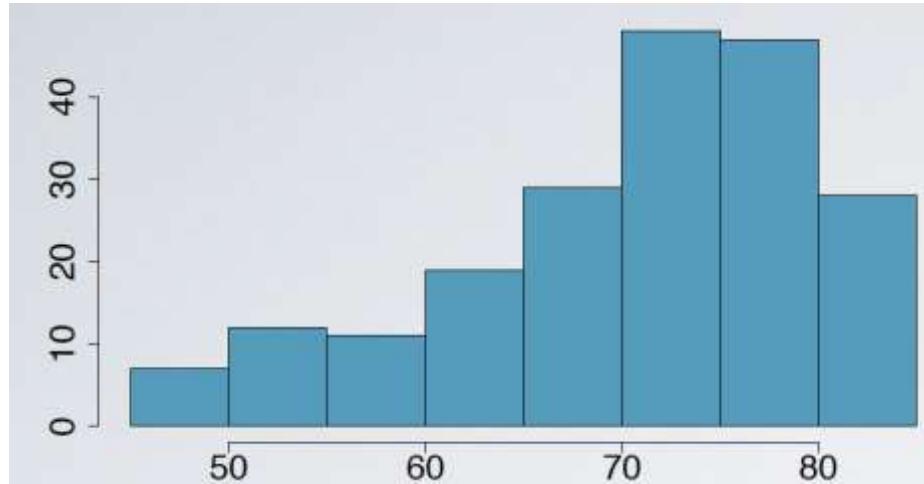
strength



outliers



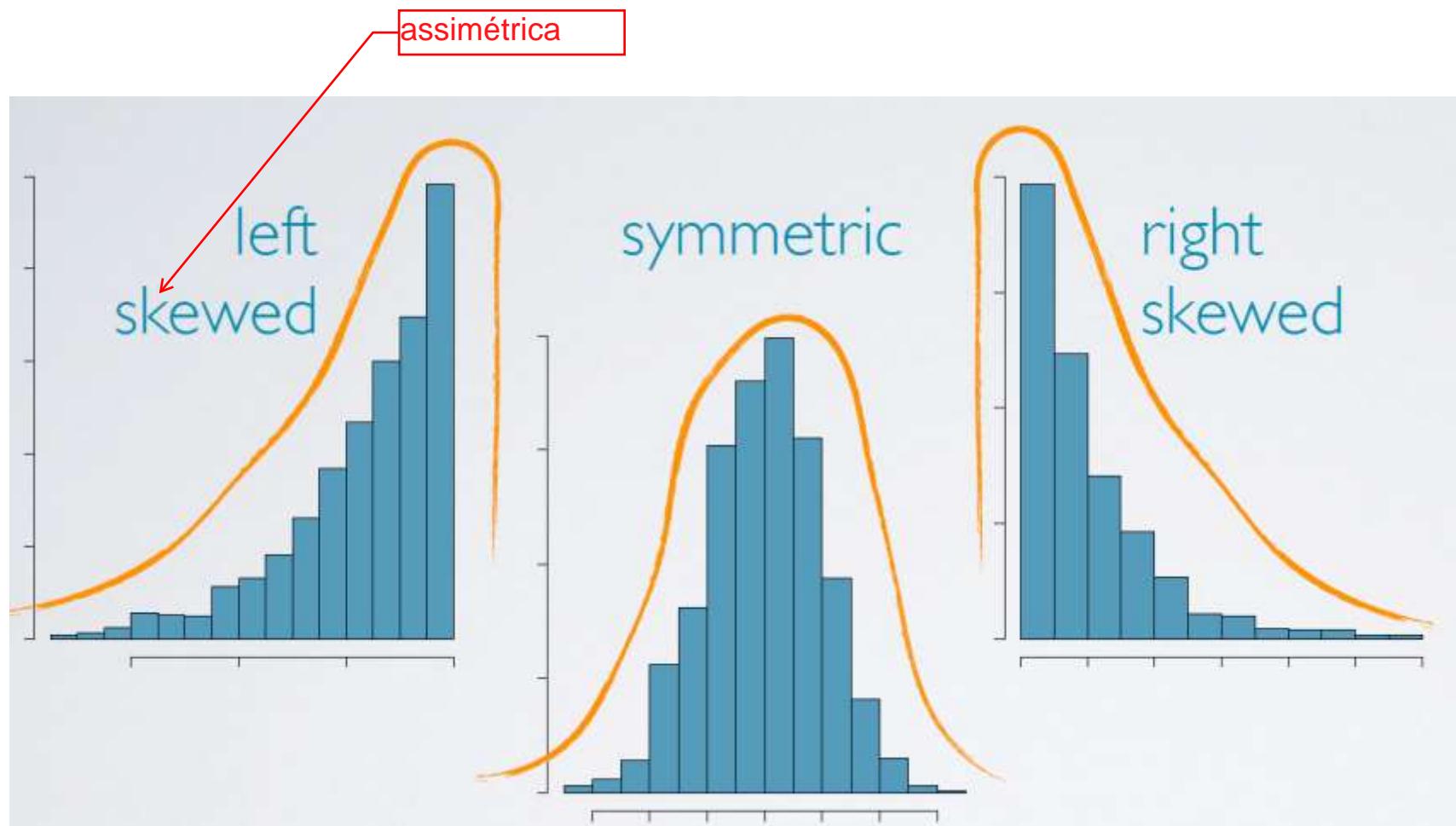
Data Science Academy



O histograma é uma ótima forma
de visualizar a densidade dos
dados...



Data Science Academy

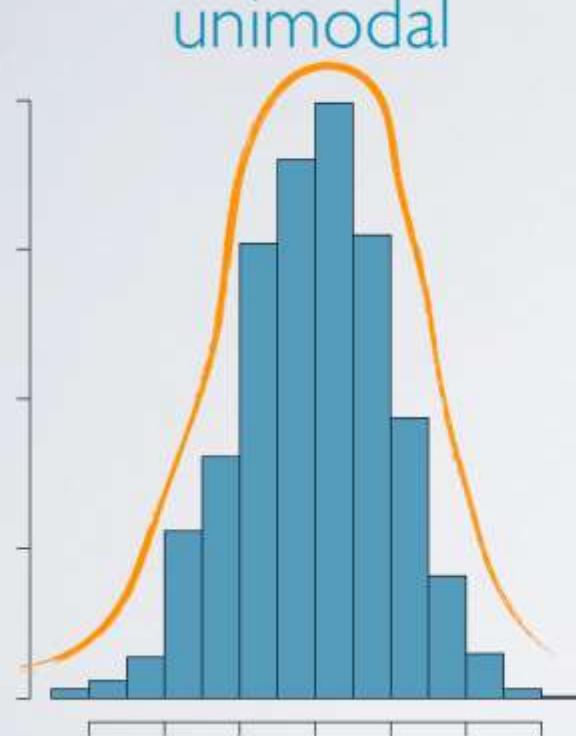


...e muito útil para
descrever o
formato de uma
distribuição

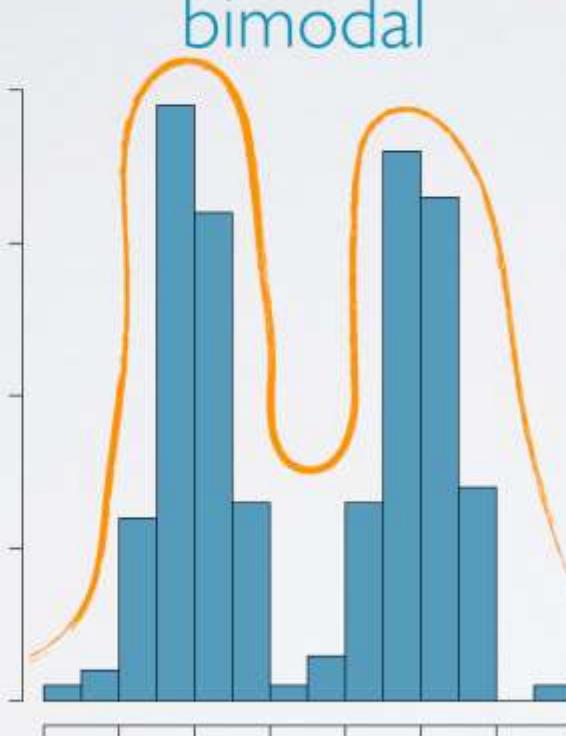


Data Science Academy

unimodal



bimodal



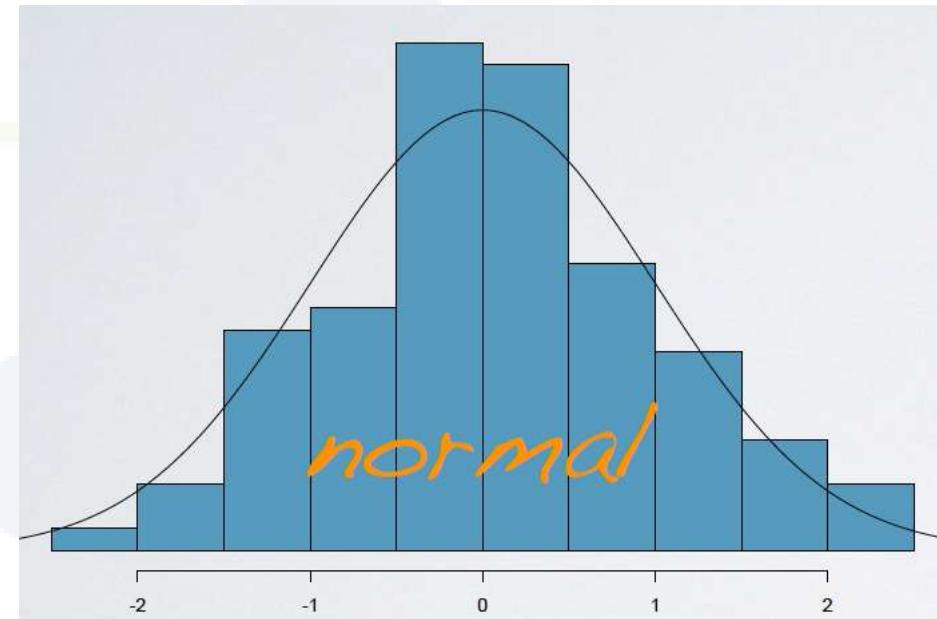
uniform



multimodal

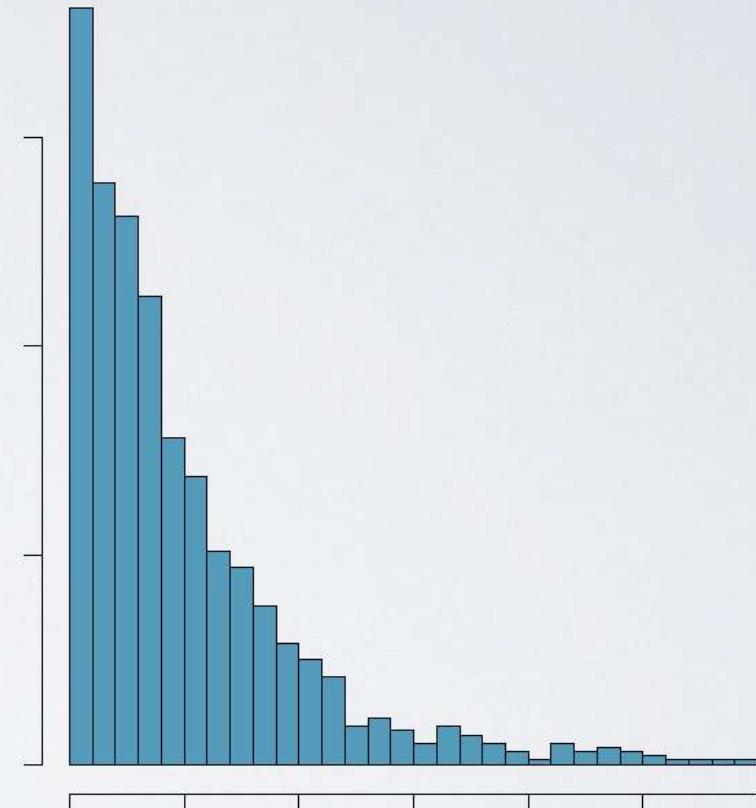
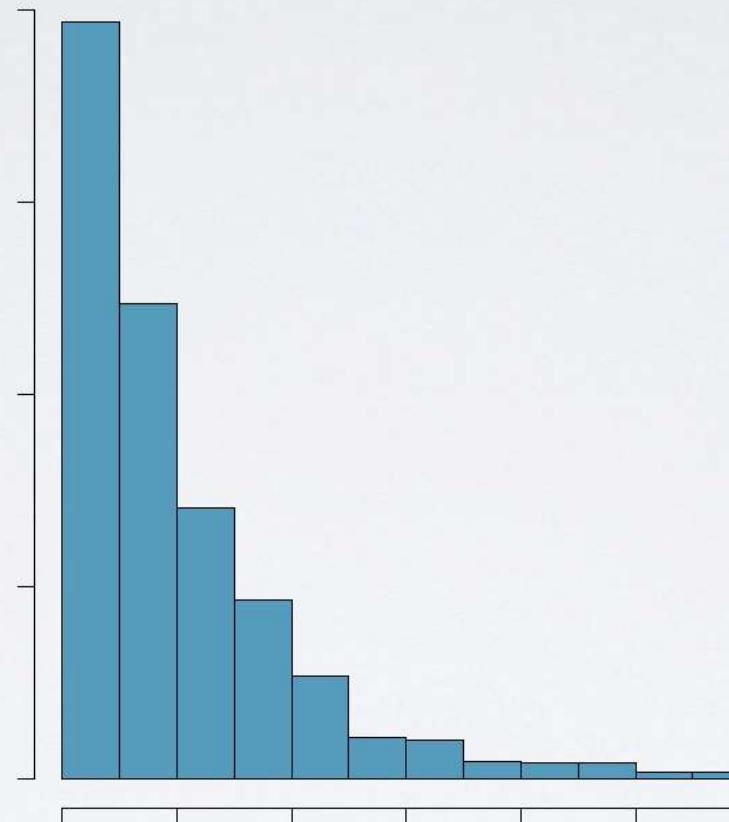
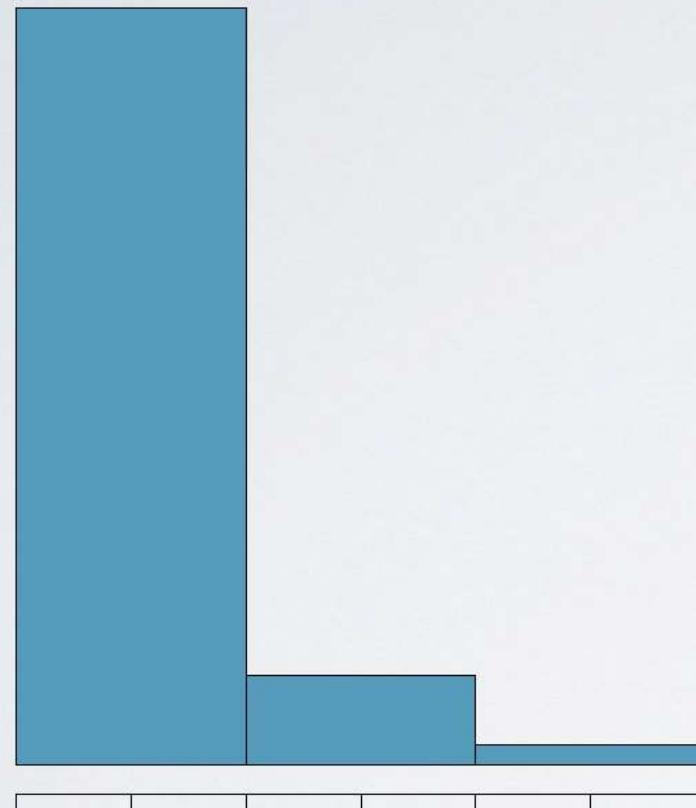


Data Science Academy

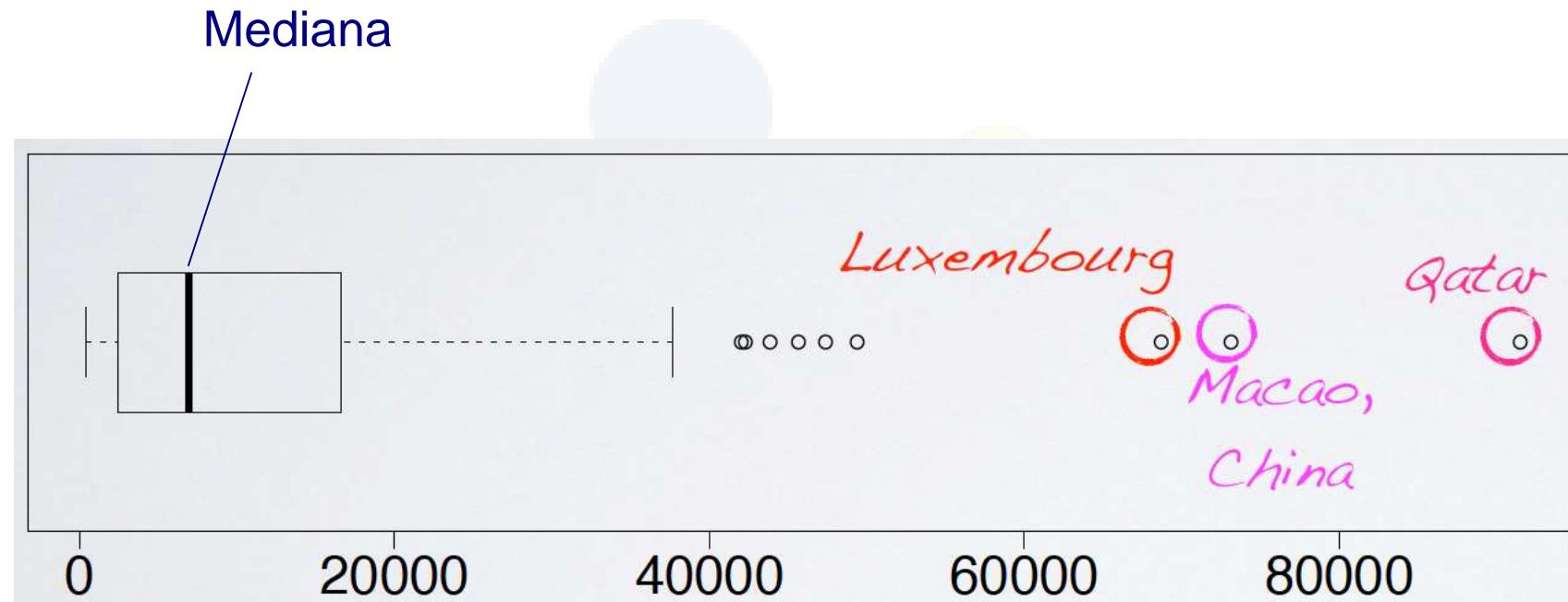


Data Science Academy

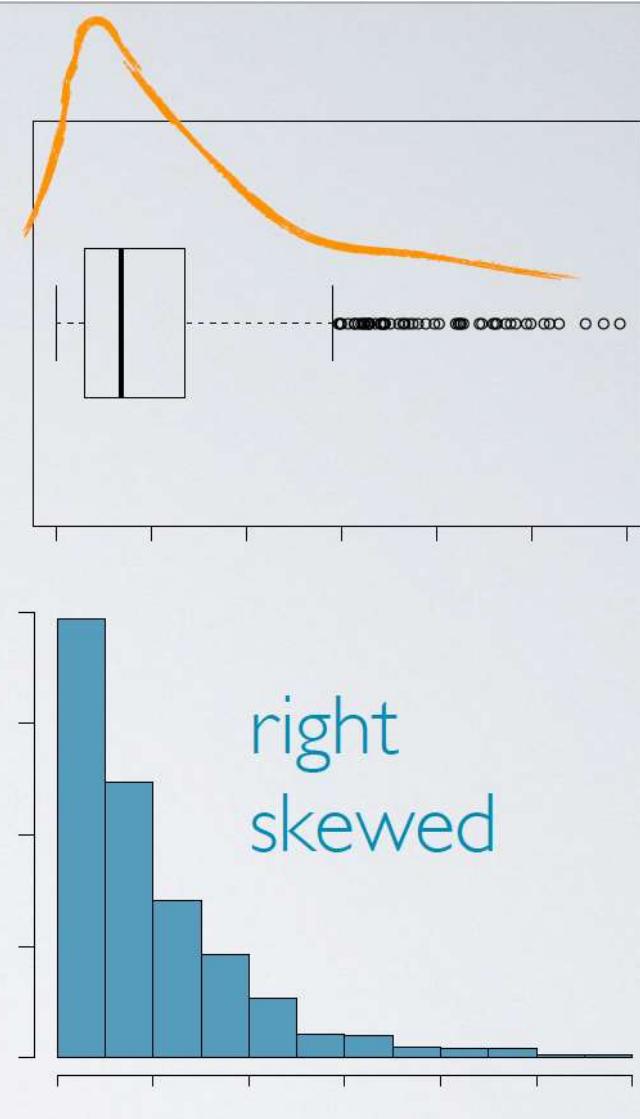
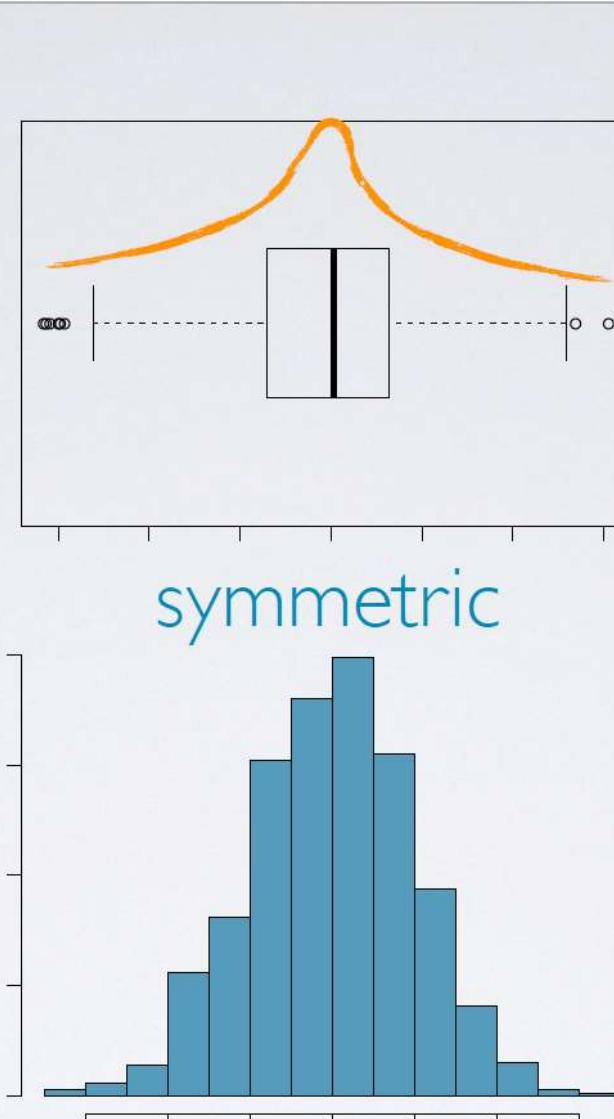
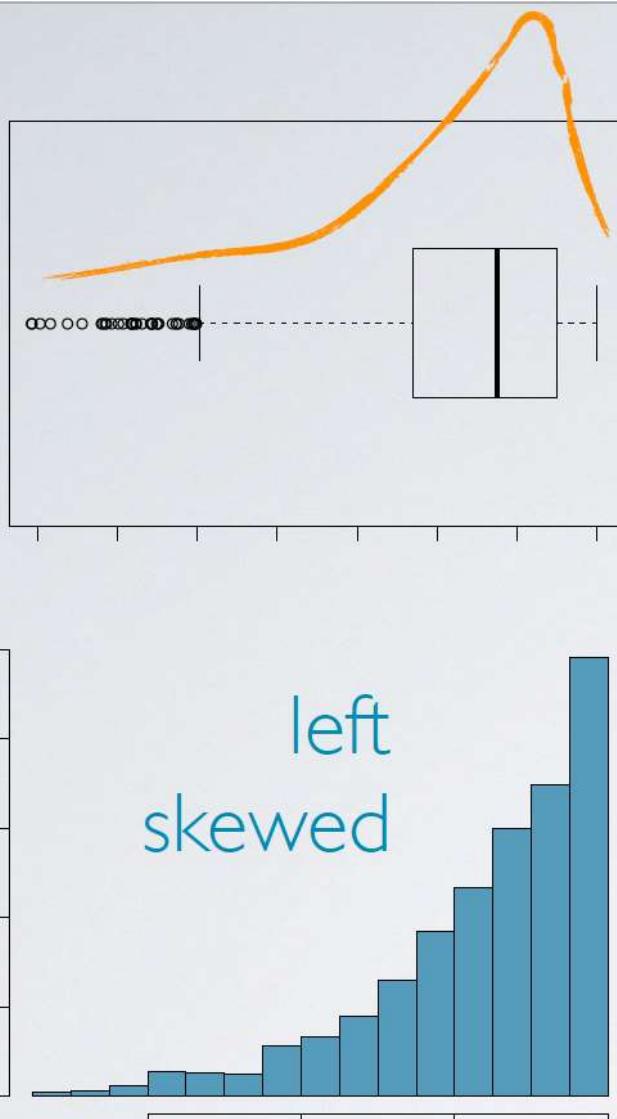
Largura da classe dos dados = bins
Observar isso



Data Science Academy



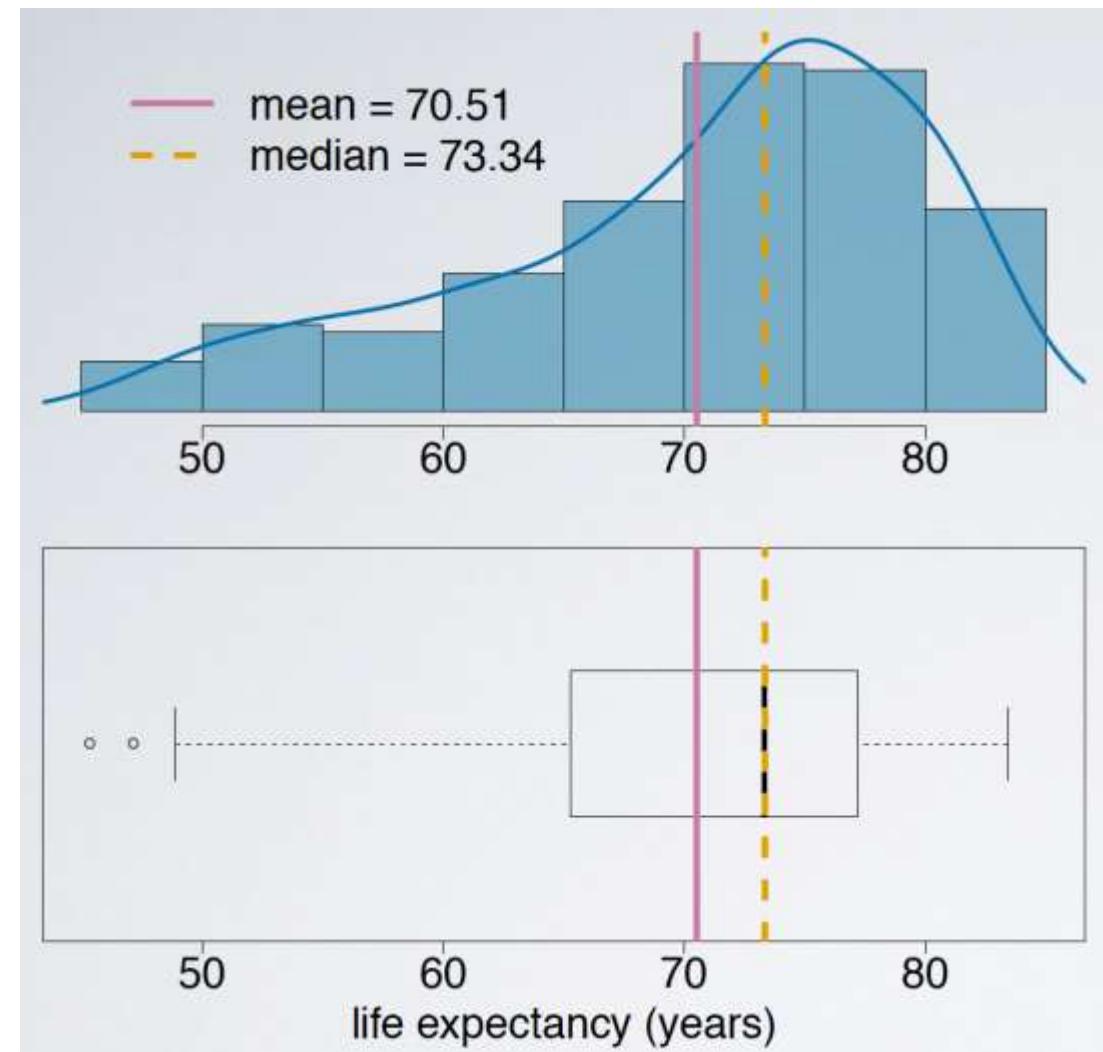
Data Science Academy



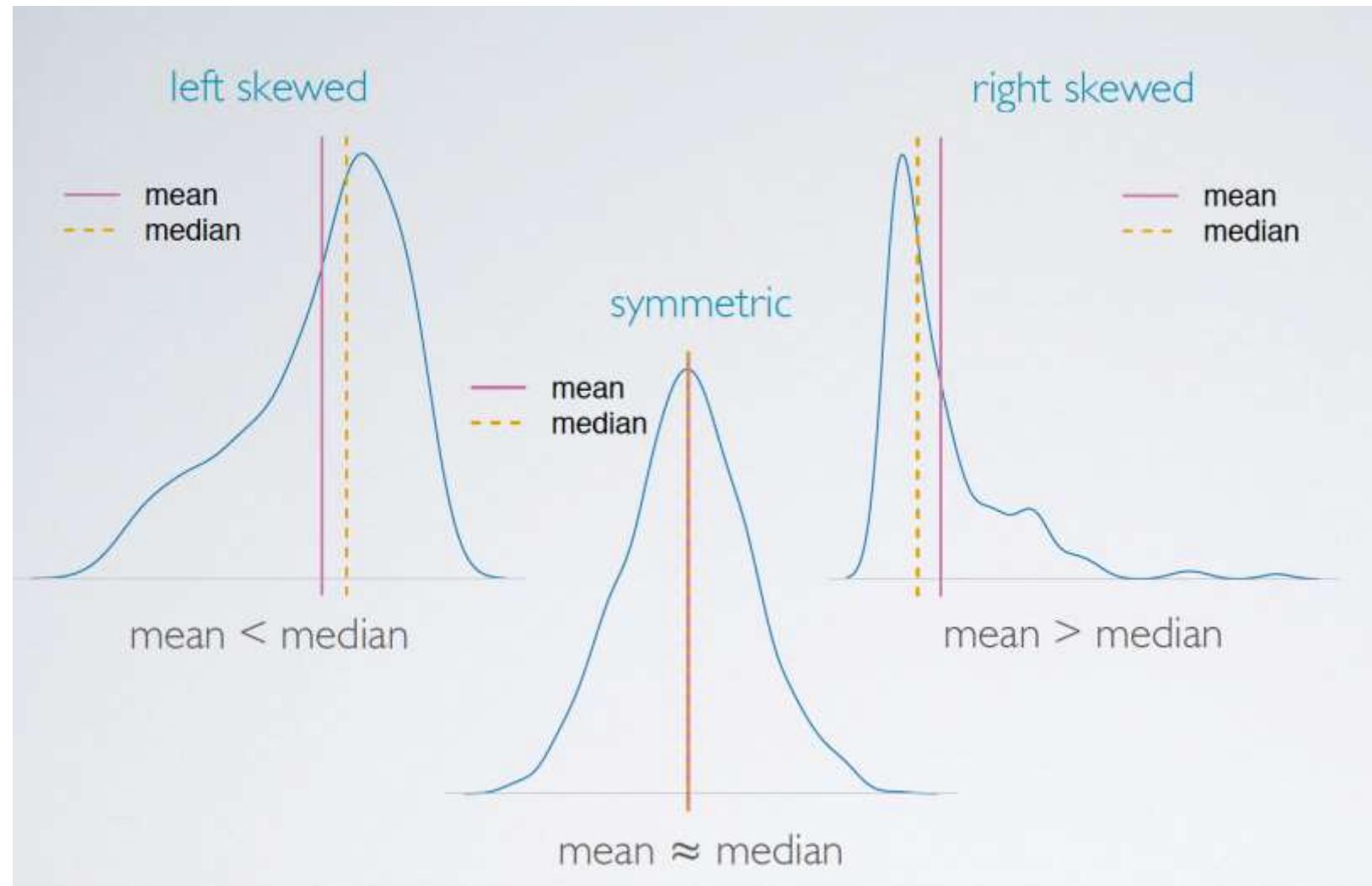


data	income per person (\$, 2012)	life expectancy (years, 2012)
Afghanistan	1359.7	60.254
Albania	6969.3	77.185
Algeria	6419.1	70.874
...
Zimbabwe	545.3	58.142





Data Science Academy



Dados	Média	Mediana
1, 2, 3, 4, 5, 6	3.5	3.5
1, 2, 3, 4, 5, 1000	169	3.5



Data Science Academy

Esse tópico chegou ao final



Data Science Academy



O que é Probabilidade?



Data Science Academy

Probabilidade é provavelmente um dos tópicos de mais difícil compreensão no campo da Estatística



Data Science Academy

Tomada de Decisões

As pessoas normalmente enfrentam dificuldade em tomar decisões racionais, relacionadas a probabilidades, pois é difícil julgar quão provável é, que um evento ocorra.



Data Science Academy

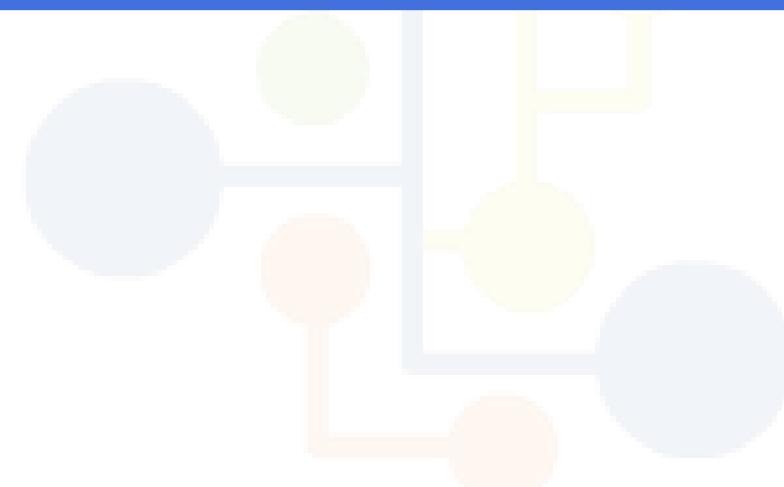
Tomada de Decisões

Muitas pessoas são bastante otimistas sobre ganhar na loteria.



Data Science Academy

Exemplo



Data Science Academy

Mega Sena



Data Science Academy

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.



Data Science Academy

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.

Ex: Cálculo para Eventos possíveis:

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



Data Science Academy

Mega Sena



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



Data Science Academy



www.datascienceacademy.com.br

Ser Atingido por um Raio



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



Data Science Academy

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



Data Science Academy

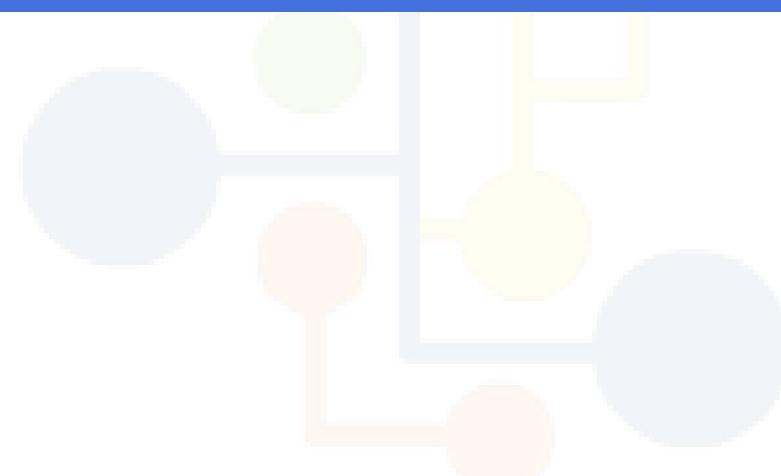
O Mundo Atual

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



Data Science Academy

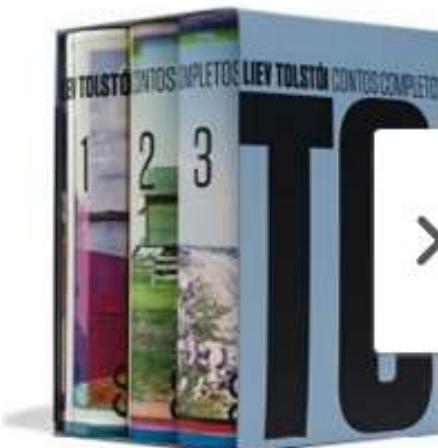
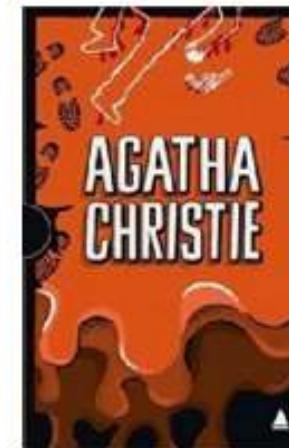
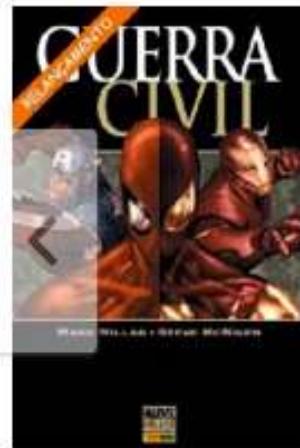
Exemplo



Data Science Academy

Site de venda de livros

Mais Vendidos em Livros [Veja mais](#)



Data Science Academy

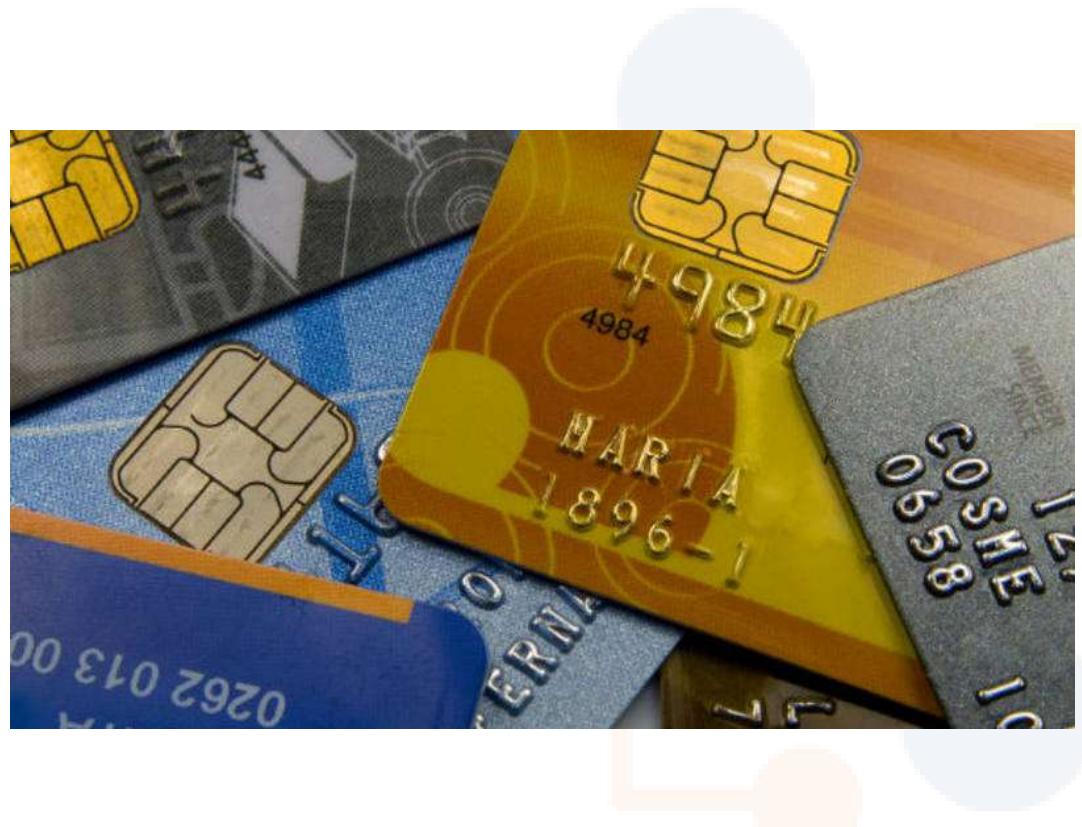
Site de venda de livros

Poderia analisar qual a probabilidade de um cliente fazer uma compra após **10 minutos** navegando pelo site.



Data Science Academy

Executivo de Empresa de Cartão de Crédito



Data Science Academy

Executivo de Empresa de Cartão de Crédito

Poderia analisar qual a probabilidade de um cliente com histórico de atrasos de pagamento, **atrasar** o pagamento da sua próxima fatura.



Data Science Academy

Empresa de Mídia



Data Science Academy

Empresa de Mídia

Poderia analisar a probabilidade de um próximo evento esportivo ter uma audiência superior a **20 milhões** de pessoas.



Data Science Academy

Departamento de Vendas



Data Science Academy

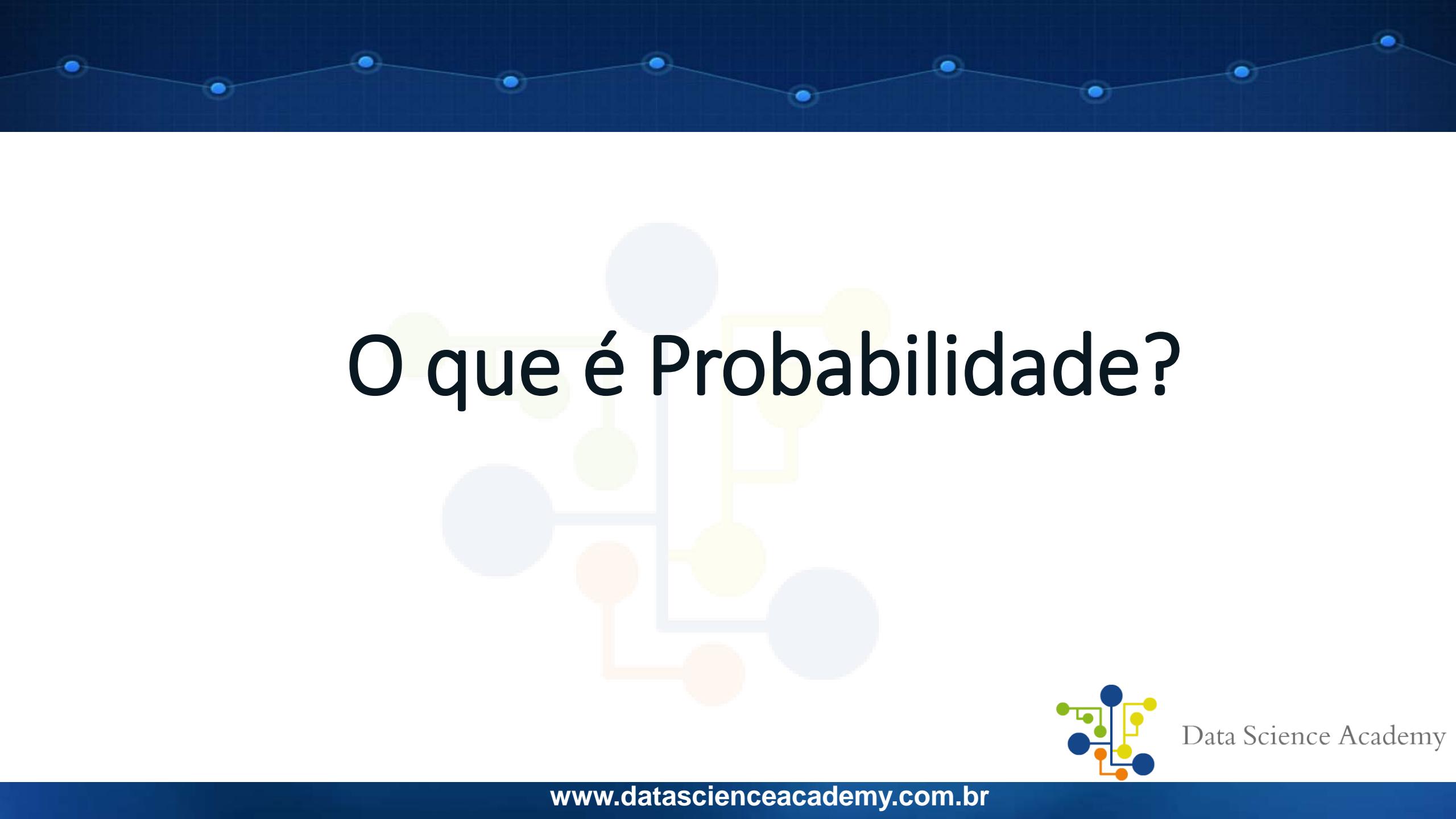
Departamento de Vendas

Poderia analisar a probabilidade de um cliente adquirir uma **garantia estendida**, após comprar um computador.



Data Science Academy

O que é Probabilidade?



Data Science Academy

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**.

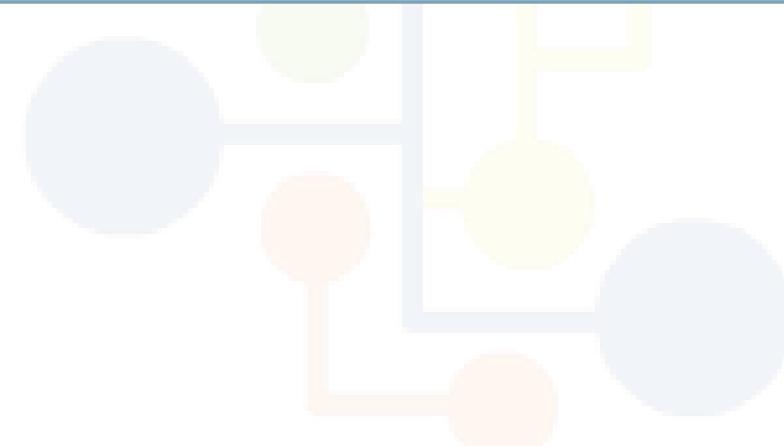
Se um evento não possui chance de ocorrer, sua probabilidade é **0 (ou 0%)**.

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

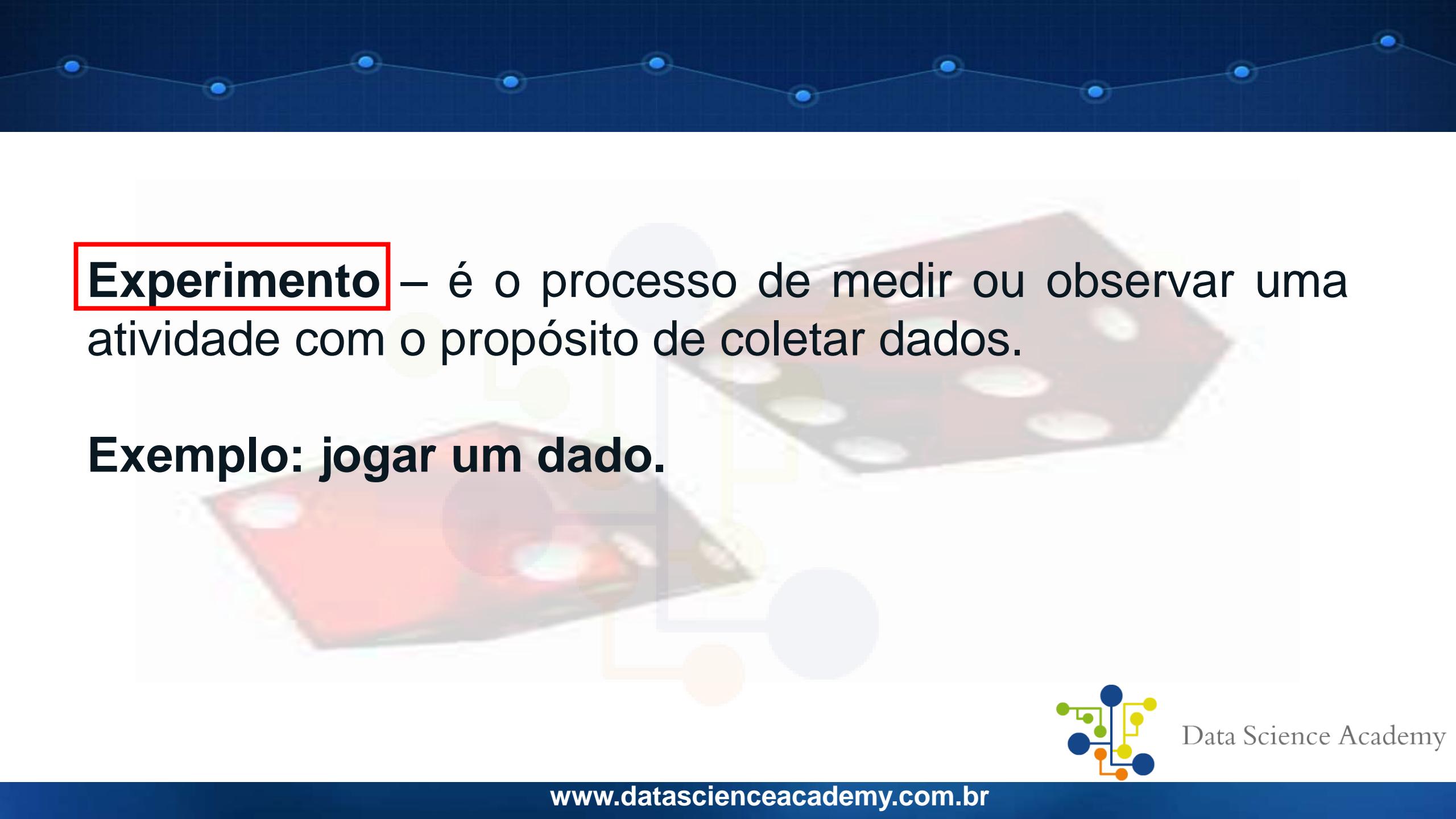


Data Science Academy

Evento e Experimento



Data Science Academy



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

Exemplo: jogar um dado.



Data Science Academy

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

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

Experimento → Espaço da amostra



Data Science Academy

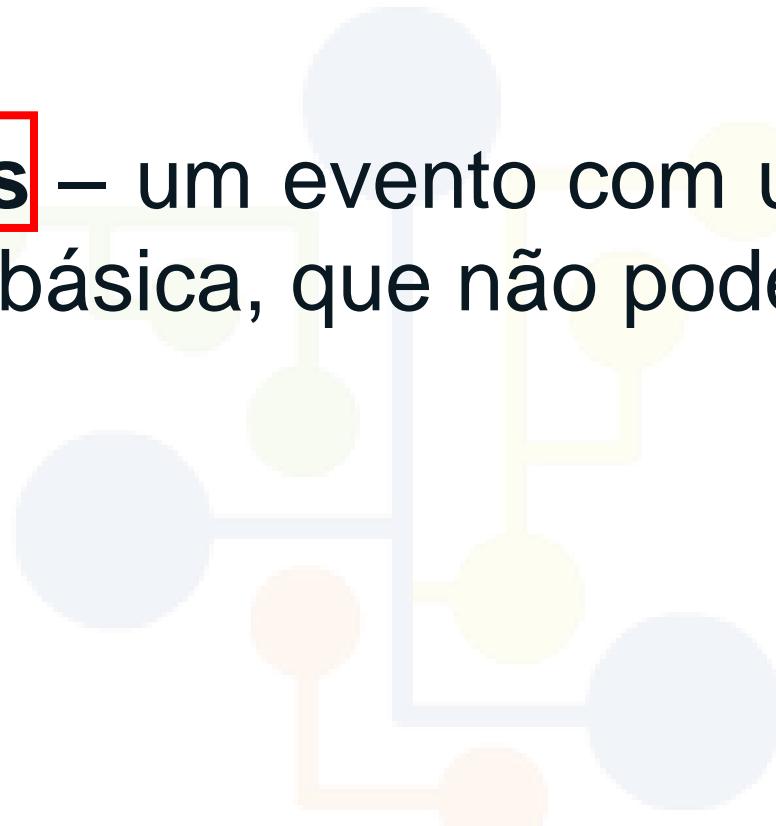
Evento – um ou mais resultados de um experimento.

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

Experimento → Espaço da amostra
↓
Evento



Data Science Academy

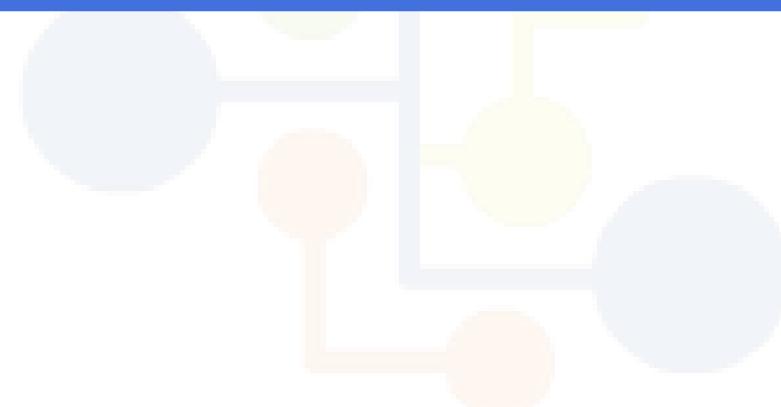


Evento Simples – um evento com um único resultado na sua forma mais básica, que não pode ser simplificado.



Data Science Academy

Exemplo



Data Science Academy

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 um baralho padrão}



Data Science Academy

Probabilidade e Possibilidade são a mesma Coisa?



Data Science Academy



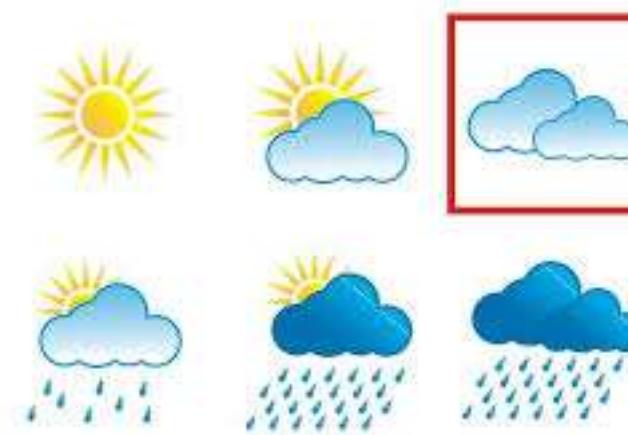
Não



Data Science Academy

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

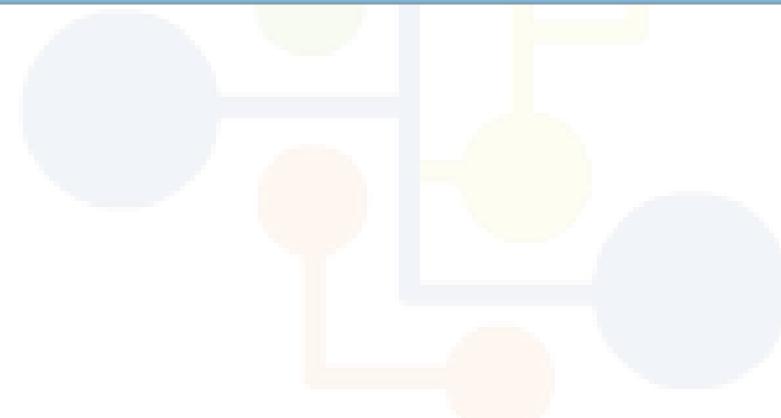
Em outras palavras, se a chance de chover amanhã é de 40%, há menos possibilidades que chova amanhã, do que não chova.



Data Science Academy



Probabilidade Clássica



Data Science Academy

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.



Data Science Academy

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 = resultado de interesse
n = resultados possíveis



Data Science Academy

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?



Data Science Academy

Resposta:

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

um evento, a face 5 no caso

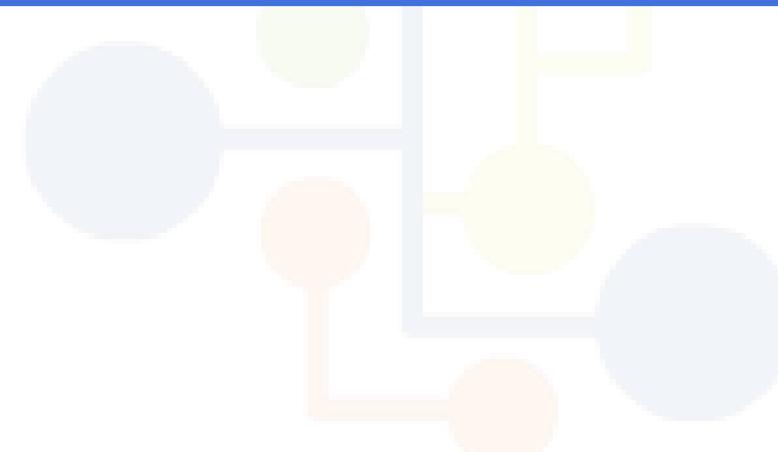
$P(A) = \frac{\text{Número de possíveis resultados do evento A}}{\text{Número total de possíveis resultados dentro do espaço da amostra}}$

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



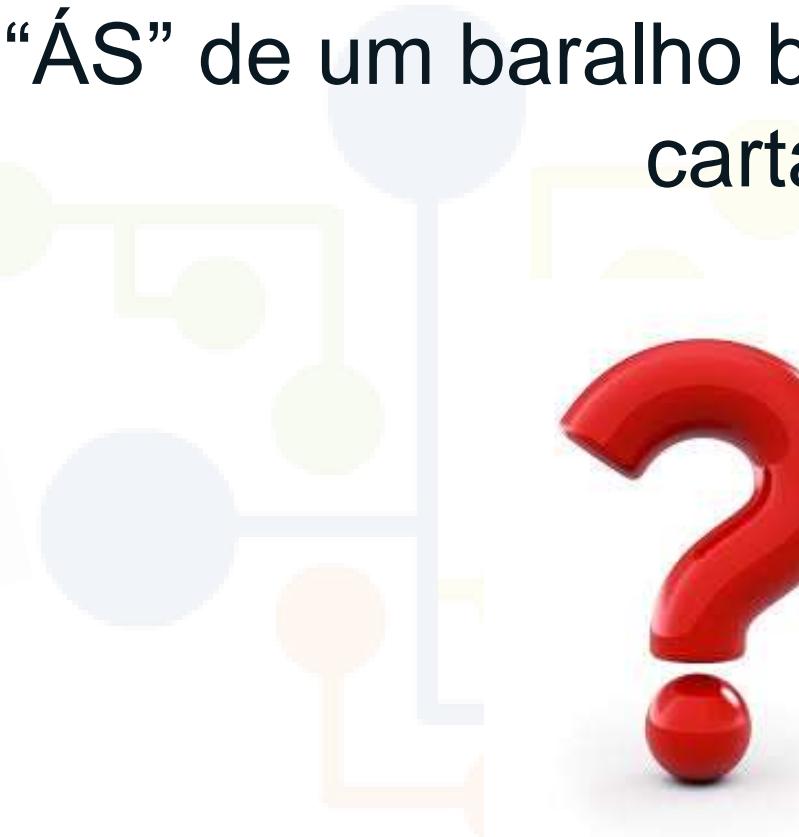
Data Science Academy

Exemplo I



Data Science Academy

Qual a probabilidade de se extrair um
“ÁS” de um baralho bem misturado de 52
cartas?



Data Science Academy



Bem misturado significa que qualquer carta tem a mesma chance de ser extraída.



Data Science Academy



Como temos 4 “Ases” em 52 cartas:



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

número de interesse

número de possíveis resultados

$$\begin{aligned} 4/52 &= 0,08 \\ &8\% \end{aligned}$$



Data Science Academy



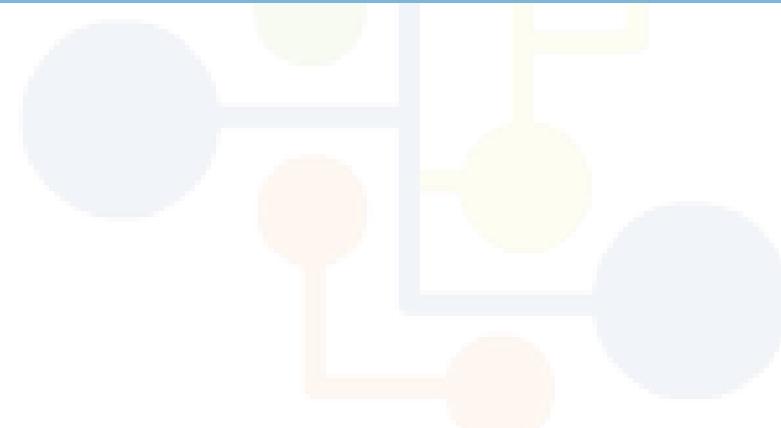
Qual a probabilidade de se extrair um
“ÁS” de um baralho bem misturado de 52
cartas?

8%



Data Science Academy

Conclusão



Data Science Academy



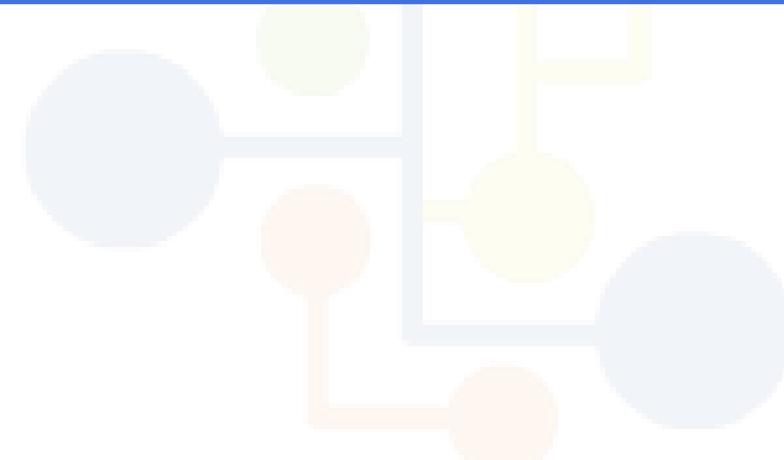
Esse é um problema clássico de probabilidade, uma vez que, todas as cartas tem a mesma chance de ocorrer.

s = sucesso = total de eventos de interesse = 4 Ases
n = total de possíveis retiradas = 52 cartas



Data Science Academy

Exemplo II



Data Science Academy



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



Data Science Academy

Como temos 2 Possibilidades, “3 ou 4”.



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

$$\begin{aligned}2/6 &= 0,33 \\&33,33\%\end{aligned}$$



Data Science Academy



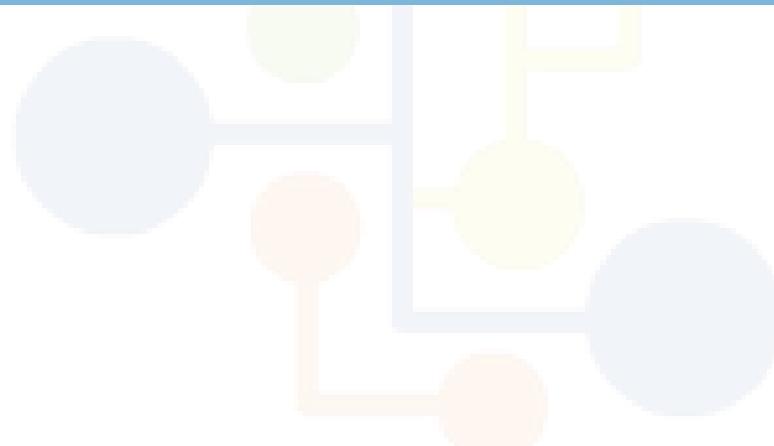
Qual a probabilidade de se obter um 3 ou
um 4 em uma jogada de um dado
equilibrado?

33,33%



Data Science Academy

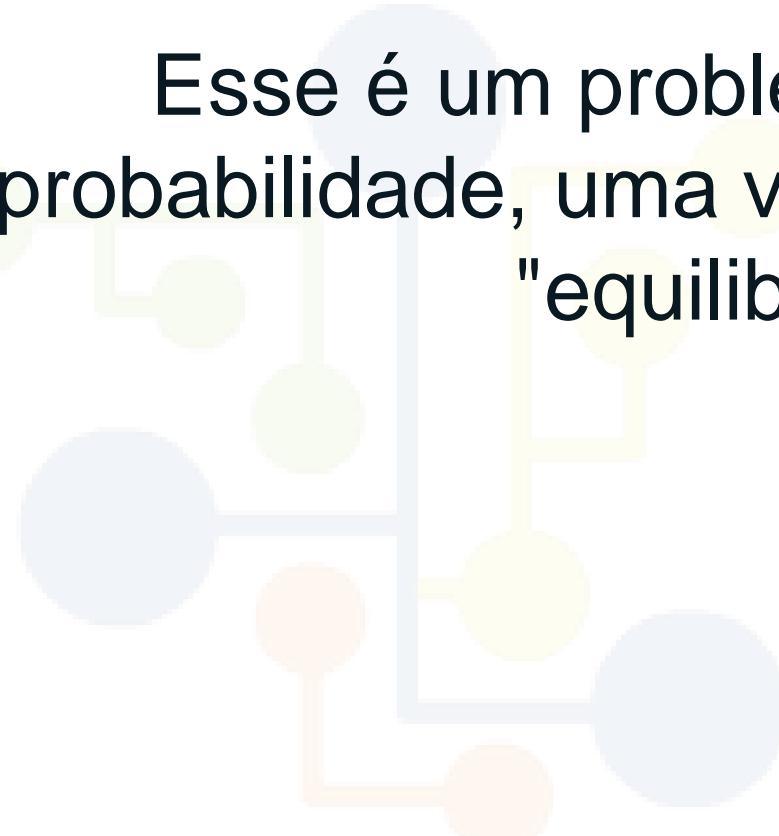
Conclusão



Data Science Academy



Esse é um problema clássico de probabilidade, uma vez que, o dado está "equilibrado".



Data Science Academy



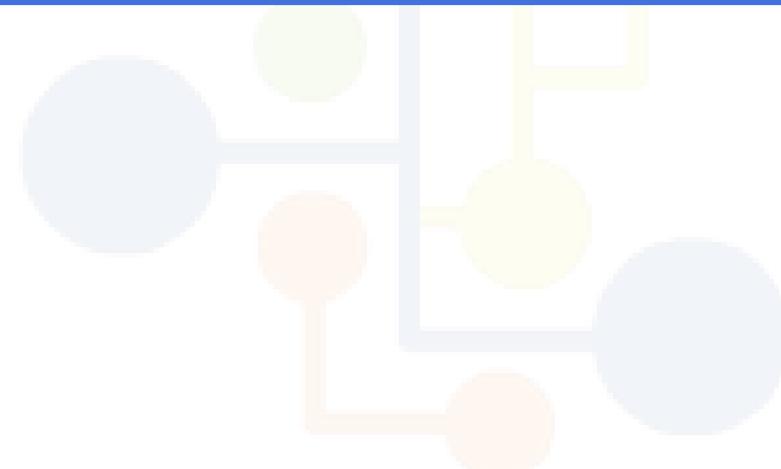
Esse é um problema clássico de probabilidade, uma vez que, o dado está "equilibrado".

s = resultado de interesse = 2 resultados (3 ou 4).
n = resultados possíveis = 6 (1,2,3,4,5,6).



Data Science Academy

Exemplo III



Data Science Academy

Qual a probabilidade de se obter um 7
jogando duas vezes um dado?



Data Science Academy



Qual a probabilidade de se obter um 7
jogando duas vezes um dado?

s = resultado de interesse = 6

6 - 1

1 - 6

2 - 5

5 - 2

3 - 4

4 - 3



Data Science Academy



Qual a probabilidade de se obter um 7
jogando duas vezes um dado?

$n = \text{resultados possíveis} = 36$



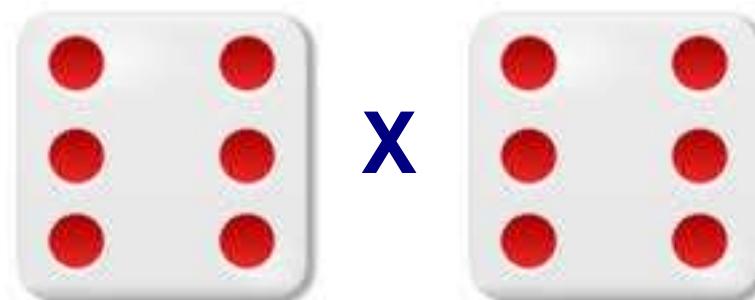
Data Science Academy



Qual a probabilidade de se obter um 7
jogando duas vezes um dado?

n = resultados possíveis = 36

1-1	1-2	1-3	1-4	1-5	1-6
2-1	2-2	2-3	2-4	2-5	2-6
3-1	3-2	3-3	3-4	3-5	3-6
4-1	4-2	4-3	4-4	4-5	4-6
5-1	5-2	5-3	5-4	5-5	5-6
6-1	6-2	6-3	6-4	6-5	6-6



Data Science Academy



Qual a probabilidade de se obter um 7
jogando duas vezes um dado?

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

S número de interesse
n número de possíveis resultados

$$\begin{aligned} 6/36 &= 0,17 \\ &17\% \end{aligned}$$



Data Science Academy



Qual a probabilidade de se obter um 7
jogando duas vezes um dado?

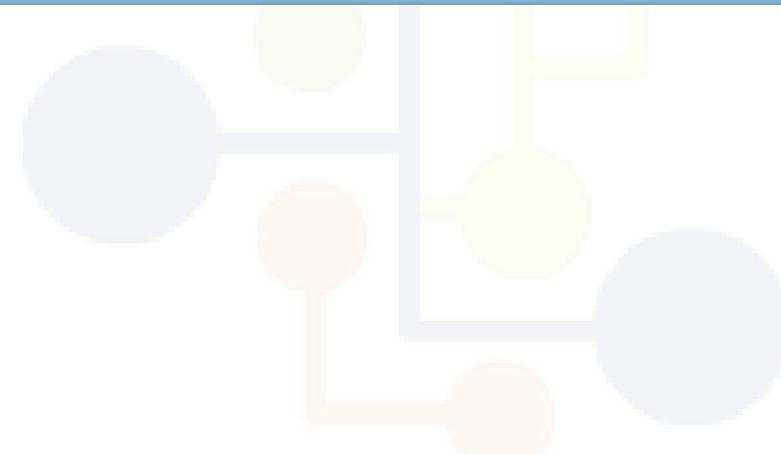
17%



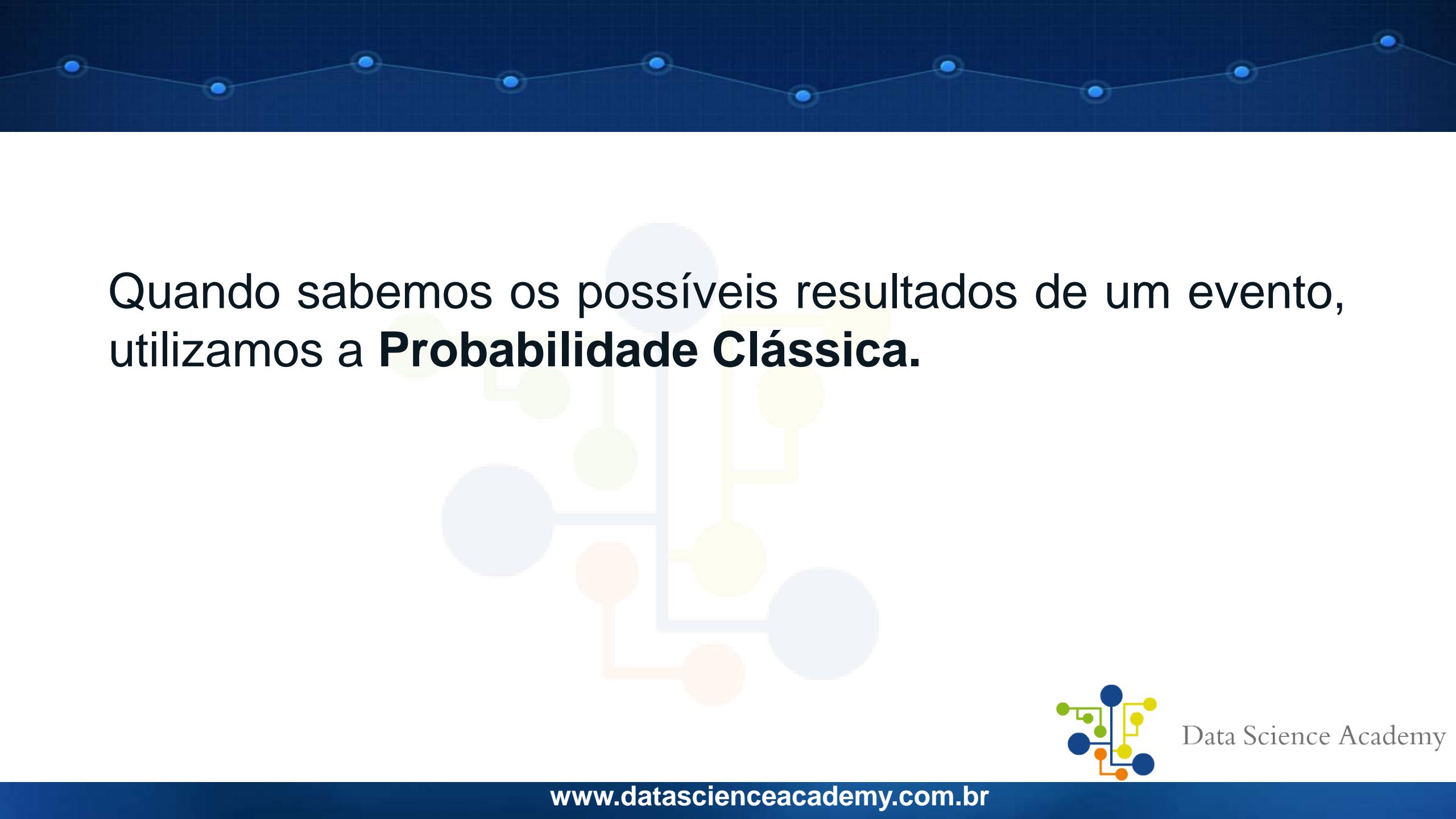
Data Science Academy



Probabilidade Empírica



Data Science Academy

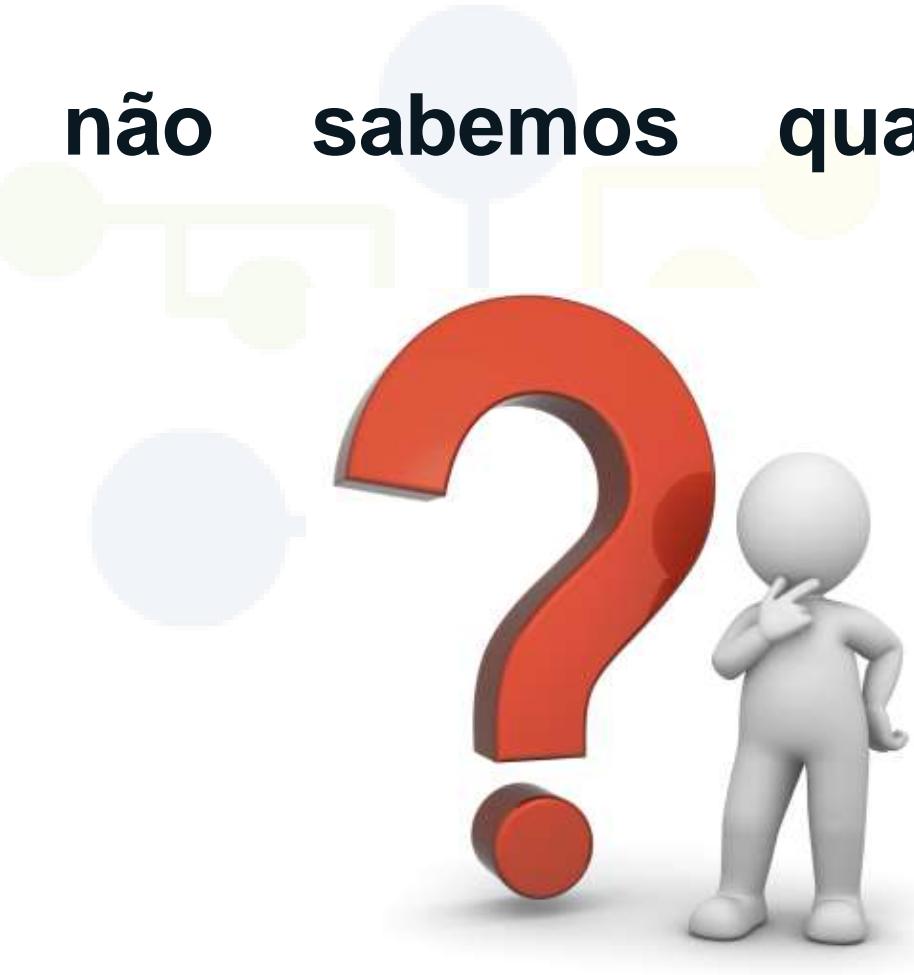


Quando sabemos os possíveis resultados de um evento,
utilizamos a **Probabilidade Clássica**.



Data Science Academy

E quando não sabemos quais os possíveis resultados?



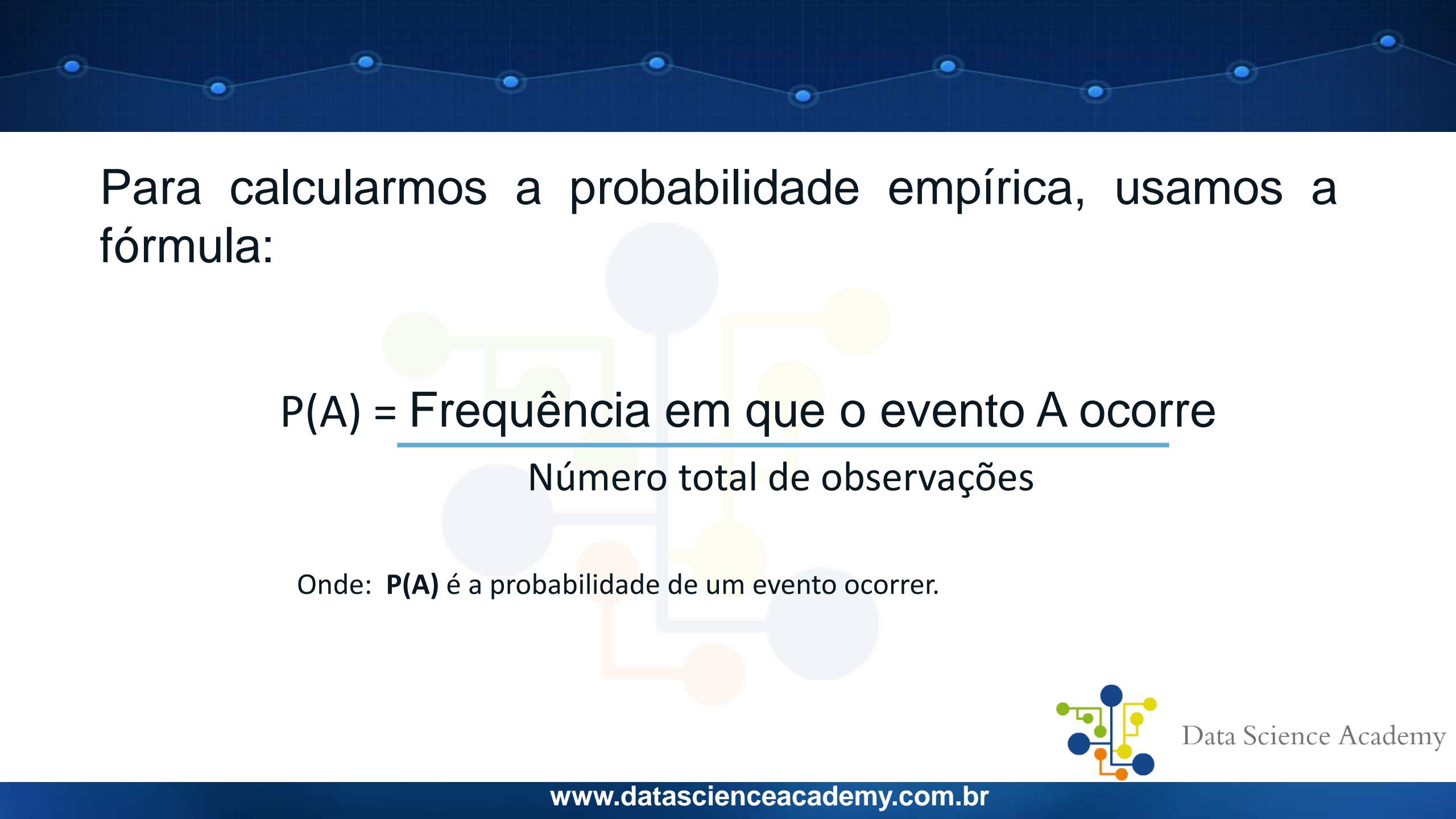
Data Science Academy



Probabilidade Empírica



Data Science Academy



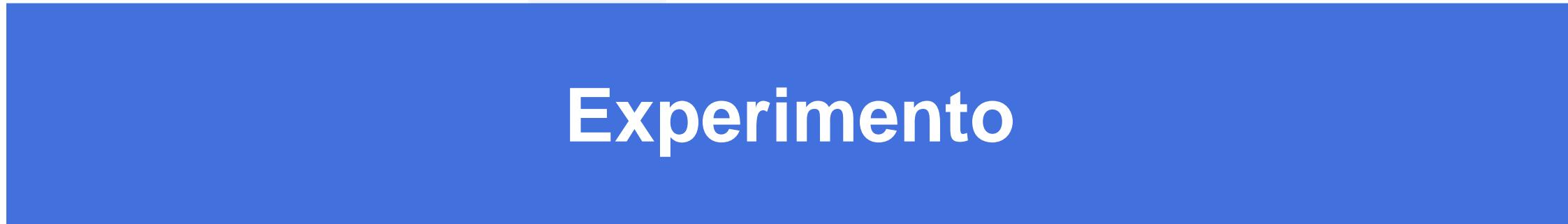
Para calcularmos a probabilidade empírica, usamos a fórmula:

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

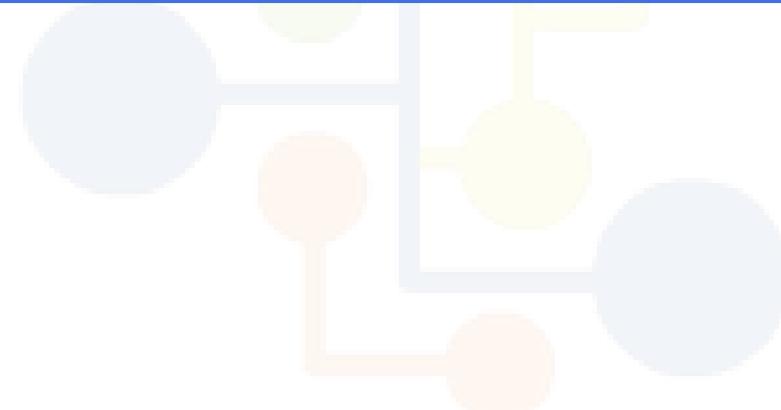
Onde: **P(A)** é a probabilidade de um evento ocorrer.



Data Science Academy



Experimento



Data Science Academy

Experimento da Loja de Livros:

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



Data Science Academy

Resposta:

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



Data Science Academy

Resposta:

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



Data Science Academy

Resposta:

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



Data Science Academy

Resposta:

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

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

A medida que um experimento é conduzido num número considerável de vezes, a probabilidade empírica do processo irá convergir para a probabilidade clássica.

$$15/100 = 0.15$$

15%



Data Science Academy

Experimento da Loja de Livros:

A probabilidade de que uma pessoa que entre na loja, faça uma compra.



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

$$\begin{aligned} 15/100 &= \\ 0.15 \end{aligned}$$

15%



Data Science Academy

Probabilidade Subjetiva



Data Science Academy



Usamos **Probabilidade Subjetiva** quando:

Probabilidades clássicas ou empíricas não podem ser usadas.

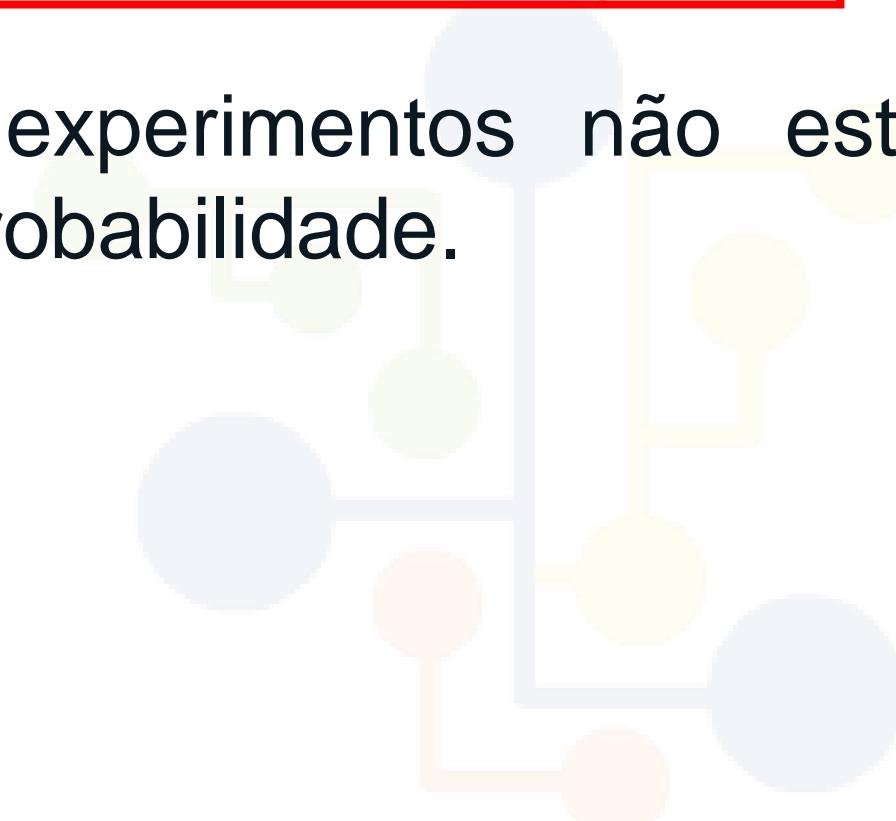


Data Science Academy



Usamos **Probabilidade Subjetiva**, quando:

Dados ou experimentos não estão disponíveis para calcular a probabilidade.

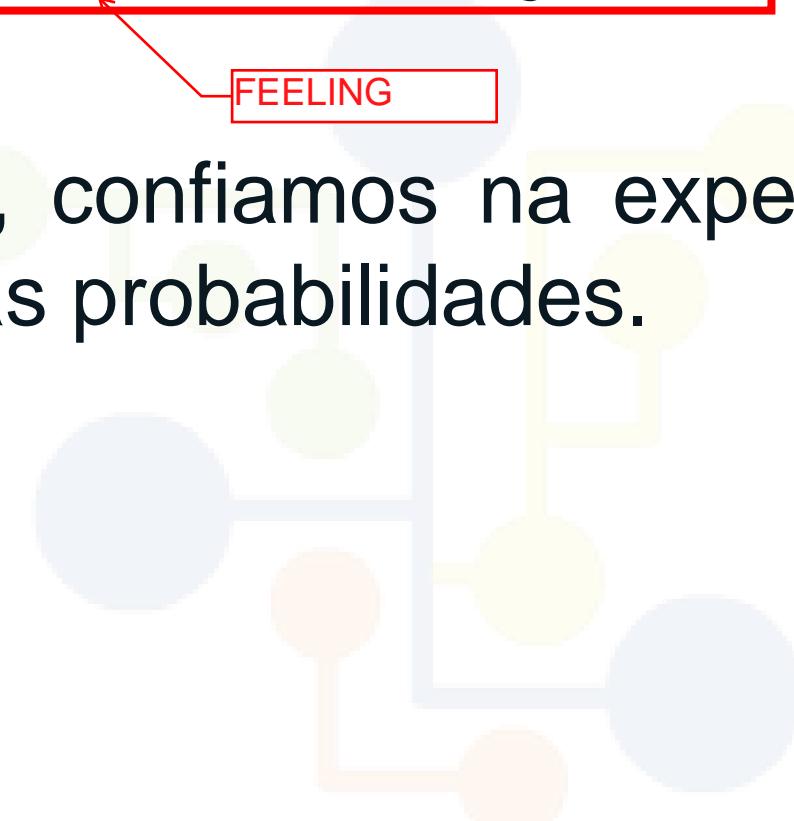


Data Science Academy



Usamos **Probabilidade Subjetiva**, quando:

FEELING

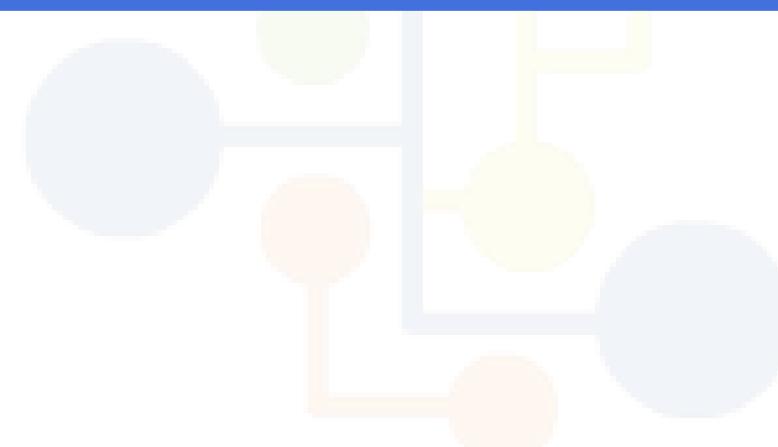


Nestes casos, confiamos na experiência ou julgamento para estimar as probabilidades.



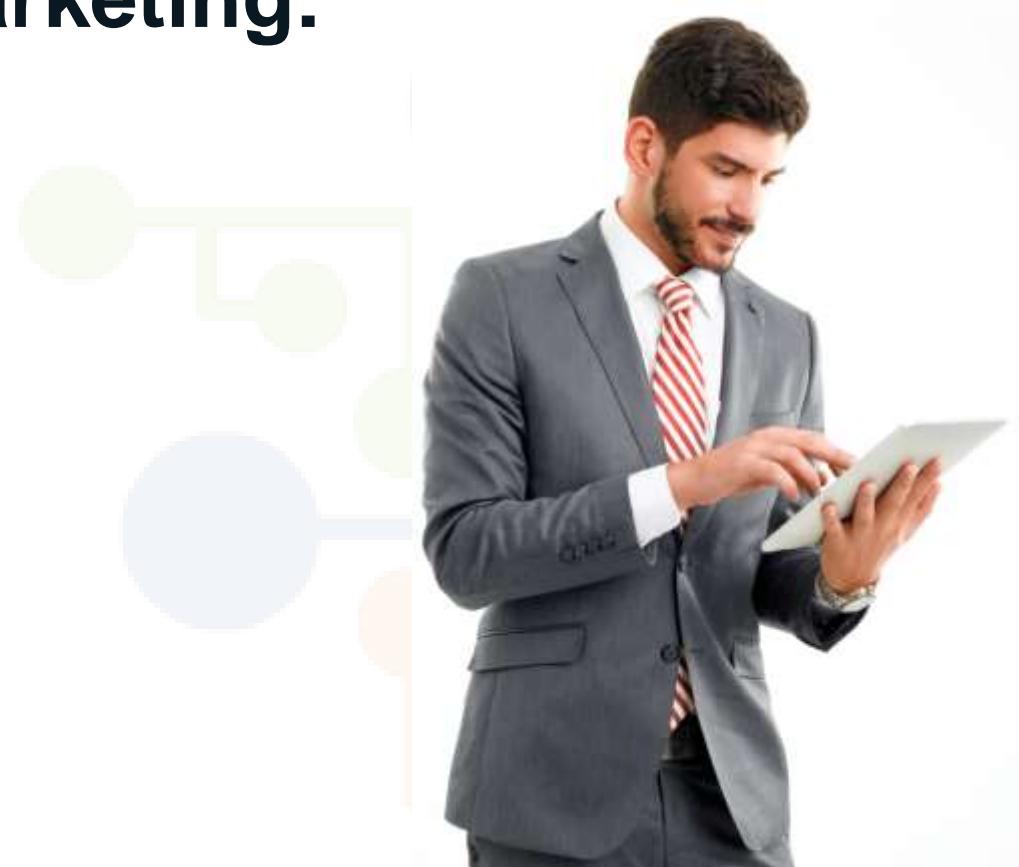
Data Science Academy

Exemplo I



Data Science Academy

Diretor de Marketing:



Data Science Academy

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.



Data Science Academy

Analista Financeiro:



Data Science Academy

Analista Financeiro:

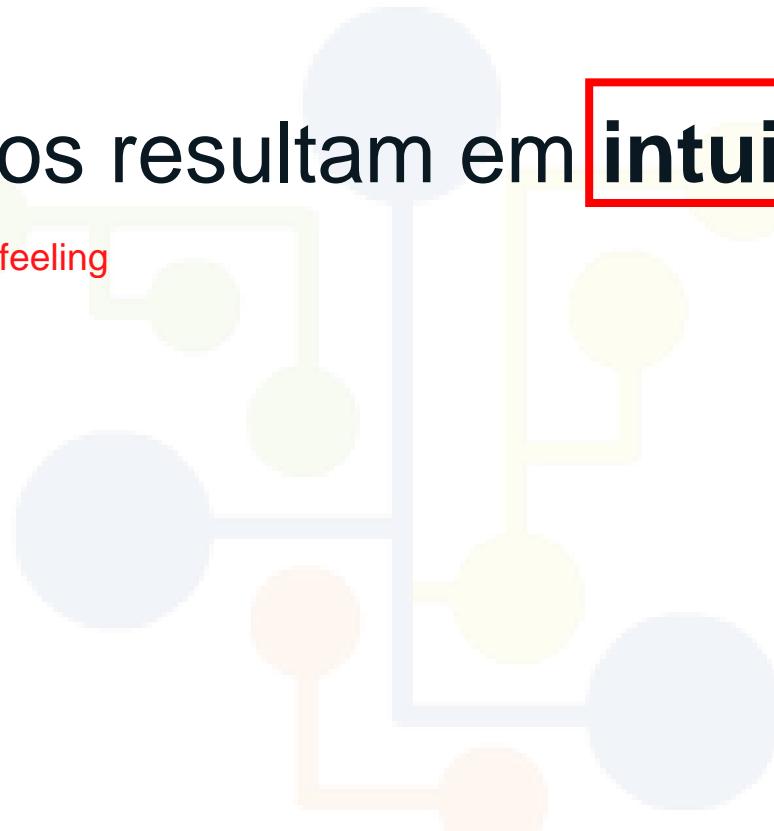
Os analistas estimam que há 40% de chance de uma nova crise ocorrer dentro de 3 anos.



Data Science Academy

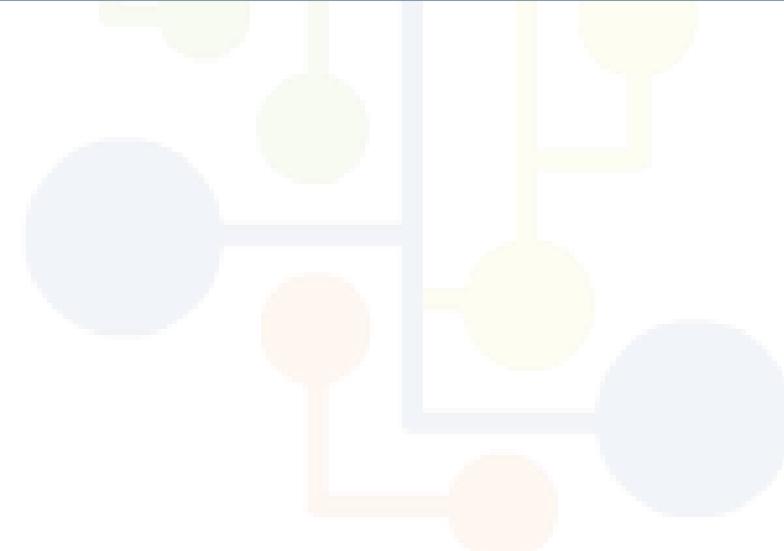
Esses exemplos resultam em **intuição** ou **estimativa**.

É só feeling



Data Science Academy

Regras Básica de Probabilidade



Data Science Academy

1a

Regra

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



Data Science Academy

2^a

Regra

Se $P(A) = 0$, então podemos garantir que o evento A

NÃO ocorrerá.



Data Science Academy

3a

Regra

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



Data Science Academy

4^a

Regra

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



Data Science Academy

5a

Regra

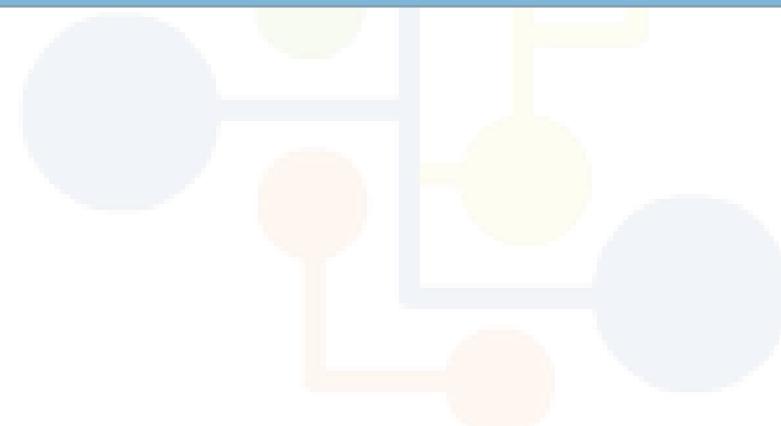
O complemento do evento A é definido como todos os resultados em um espaço de amostra, que **não** fazem parte do evento A. Ou seja:

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



Data Science Academy

Regras Básicas de Probabilidade Para Mais de Um Evento

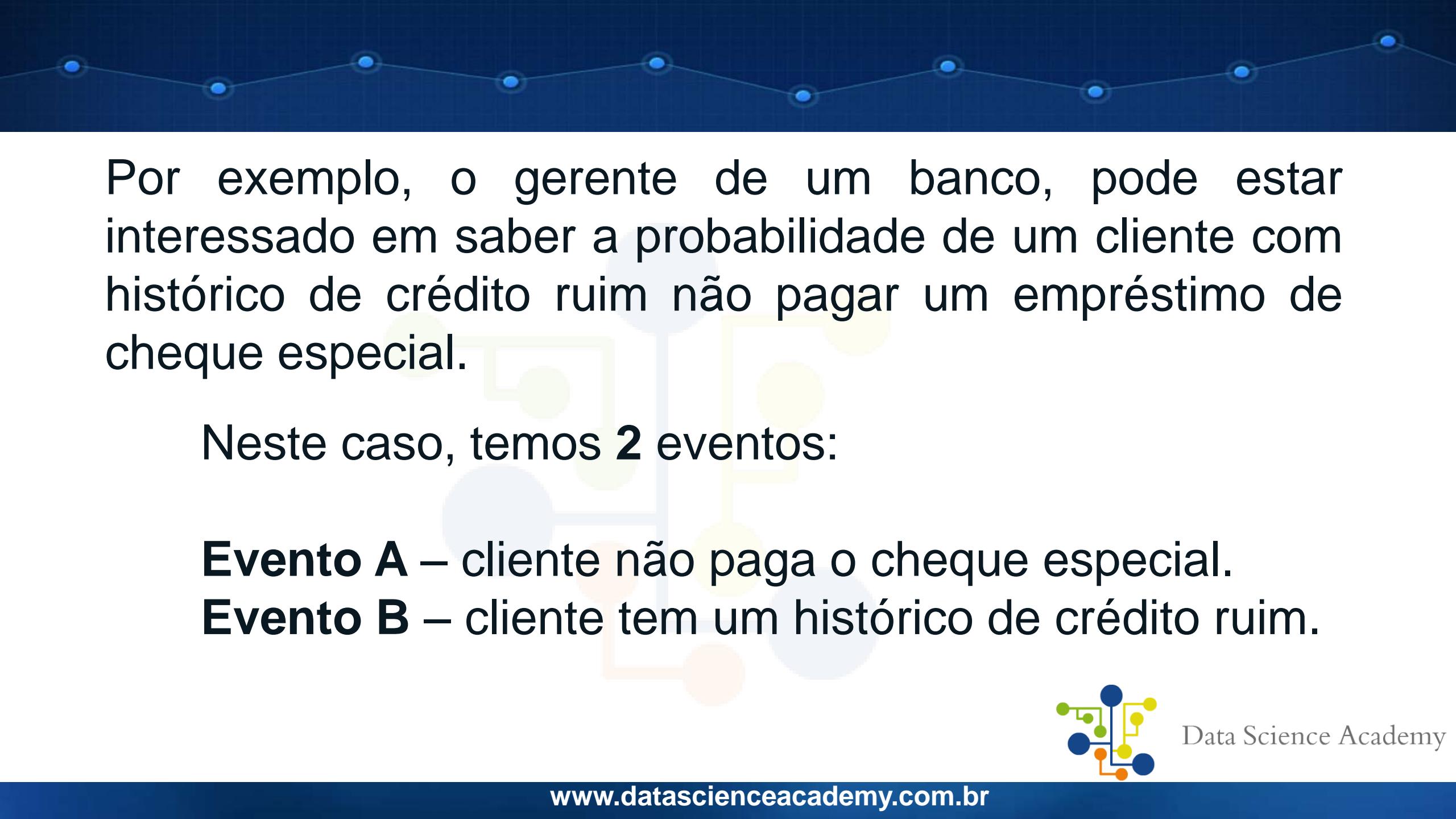


Data Science Academy

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



Data Science Academy



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:

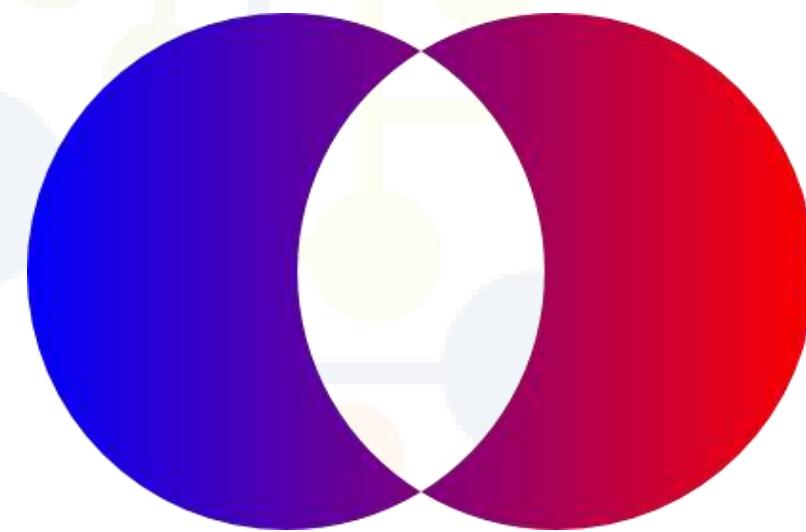
Evento A – cliente não paga o cheque especial.

Evento B – cliente tem um histórico de crédito ruim.



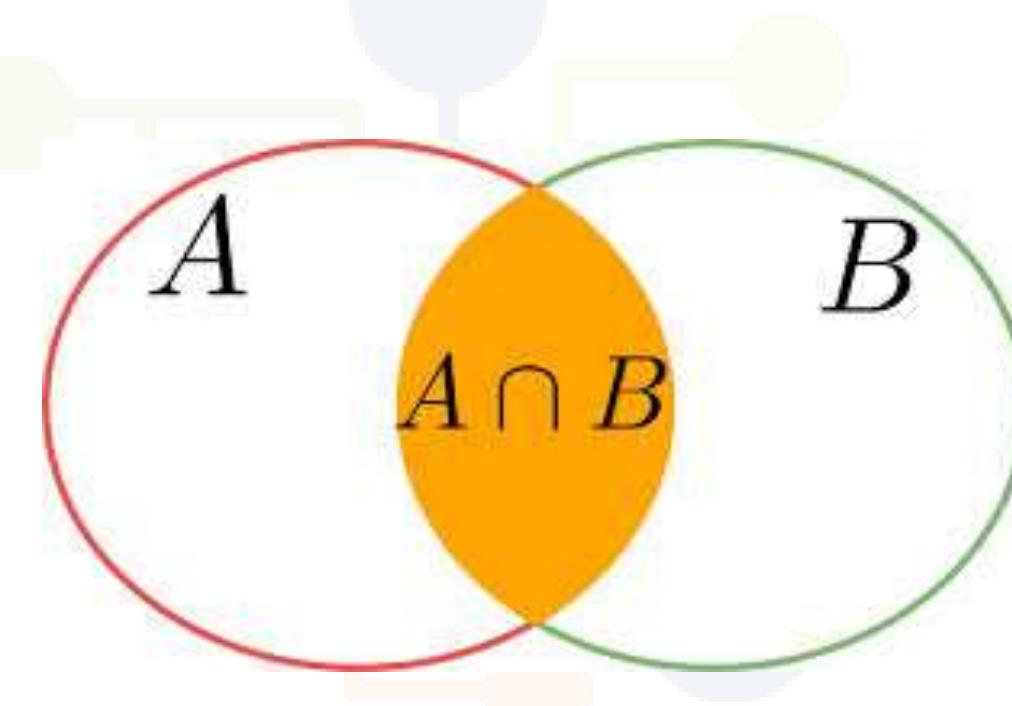
Data Science Academy

Intersecção de Eventos



Data Science Academy

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



Data Science Academy

Vamos usar uma **tabela de contingência** para exemplificar melhor.



Data Science Academy

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



Data Science Academy

Vamos definir os eventos sob análise:



Data Science Academy

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



Data Science Academy

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

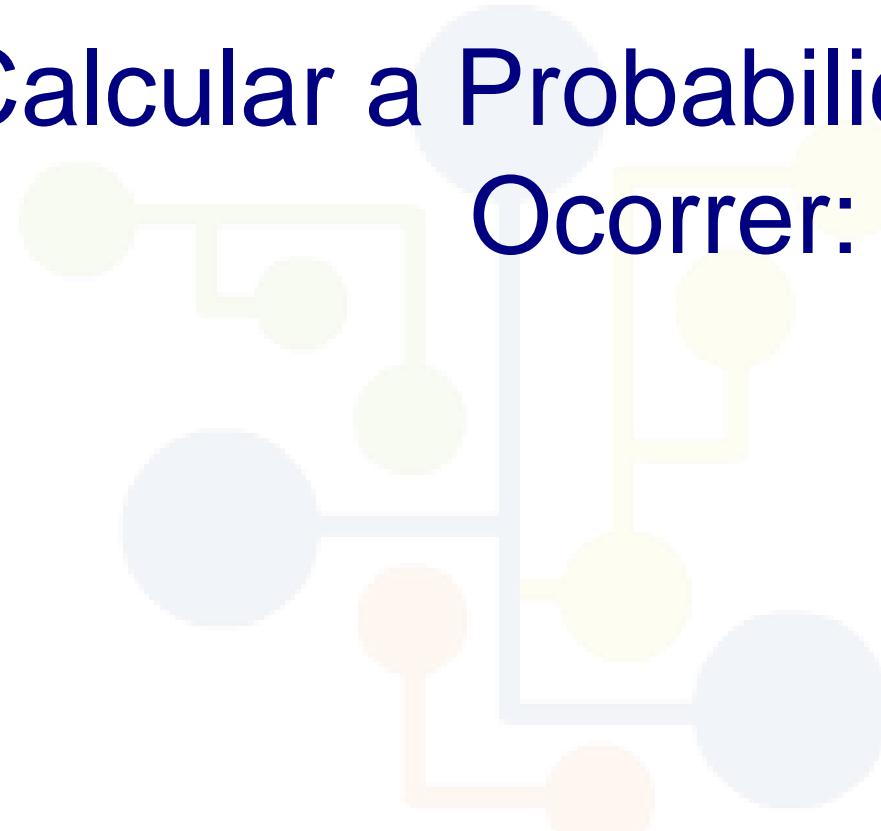
Cidade	Engenharia	Medicina	Total
Rio de Janeiro	1500	2300	3800
São Paulo	5600	7500	13100
Porto Alegre	980	1400	2380
Total	8080	11200	19280



Data Science Academy



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



Data Science Academy

Cidade	Engenharia	Medicina	Total
São Paulo	5600	7500	13100
Total	8080	11200	19280

Resposta: Evento A Ocorrer

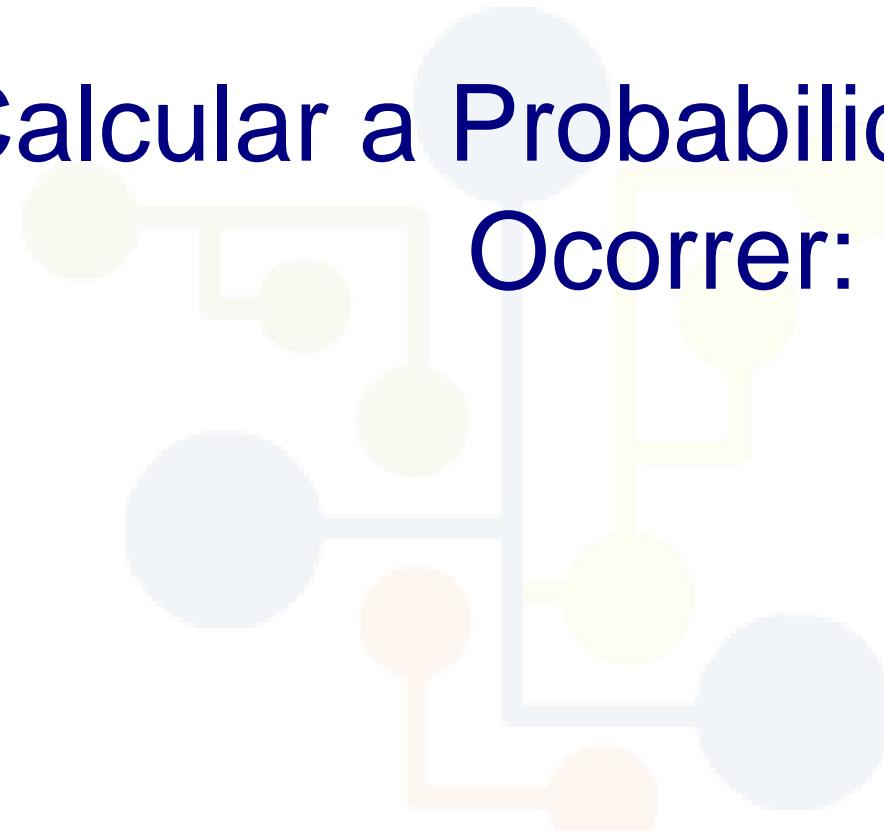
$$P(A) = 13100 / 19280 = 0.68$$
$$0.68 \times 100 = 68\%$$

Ou seja, 68% é a probabilidade de um estudante ser de São Paulo.



Data Science Academy

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



Data Science Academy

Cidade	Engenharia	Medicina	Total
Total	8080	11200	19280

Resposta: Evento B Ocorrer

$$P(B) = 11200 / 19280 = 0.58$$

Ou seja, 58% é a probabilidade de um estudante ser admitido para o curso de Medicina.



Data Science Academy

Vamos Calcular a Probabilidade de um Estudante de São Paulo, ser admitido em um 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



Data Science Academy

Cidade	Engenharia	Medicina	Total
São Paulo	5600	7500	13100
Total	8080	11200	19280

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

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



Data Science Academy

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

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

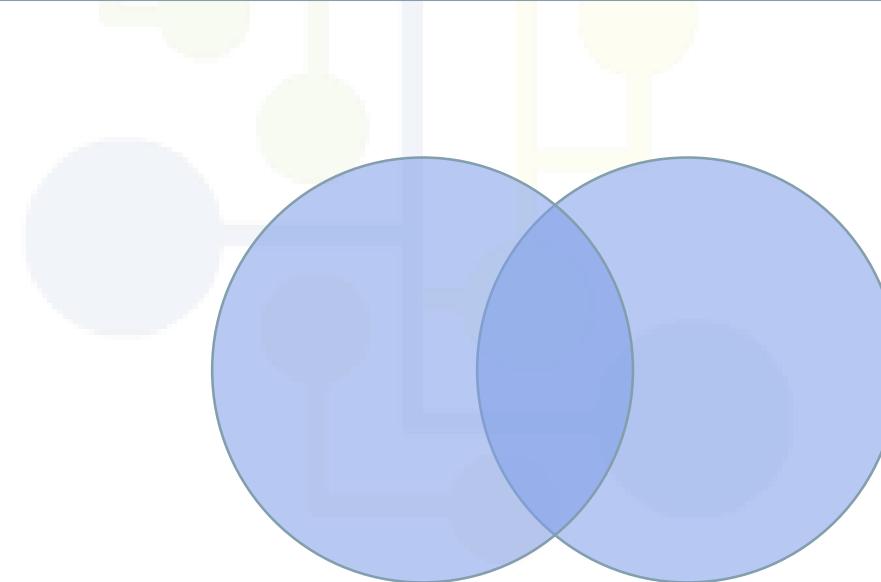
39% é a probabilidade de um estudante de São Paulo ser admitido para o curso de Medicina.



Data Science Academy

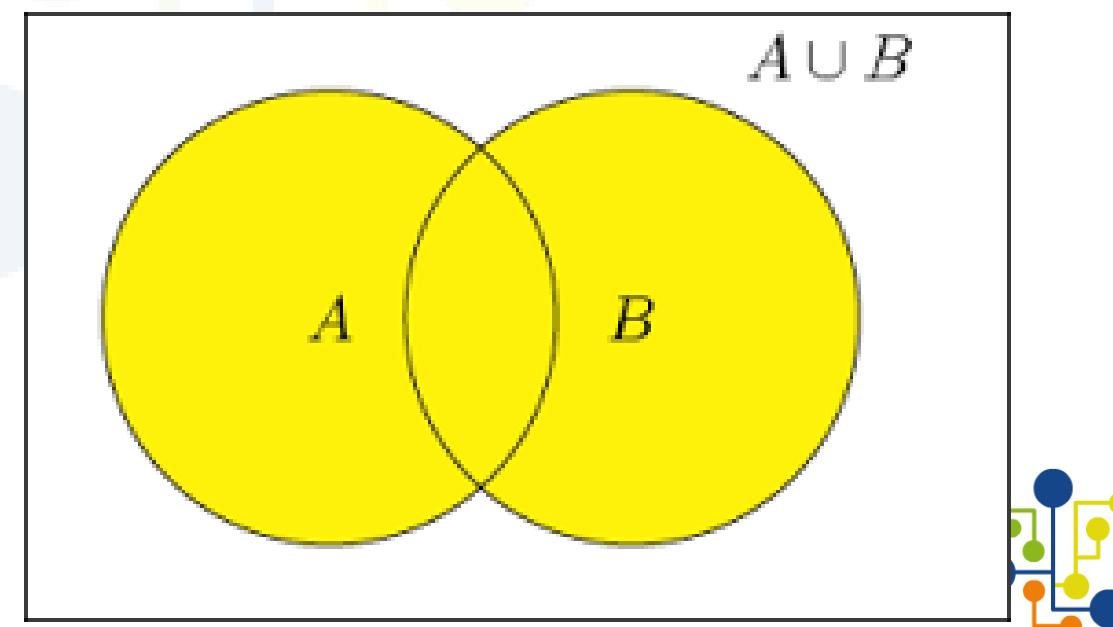


União de Eventos



Data Science Academy

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



Vamos usar uma
exemplificar melhor.

tabela de contingência para



Data Science Academy

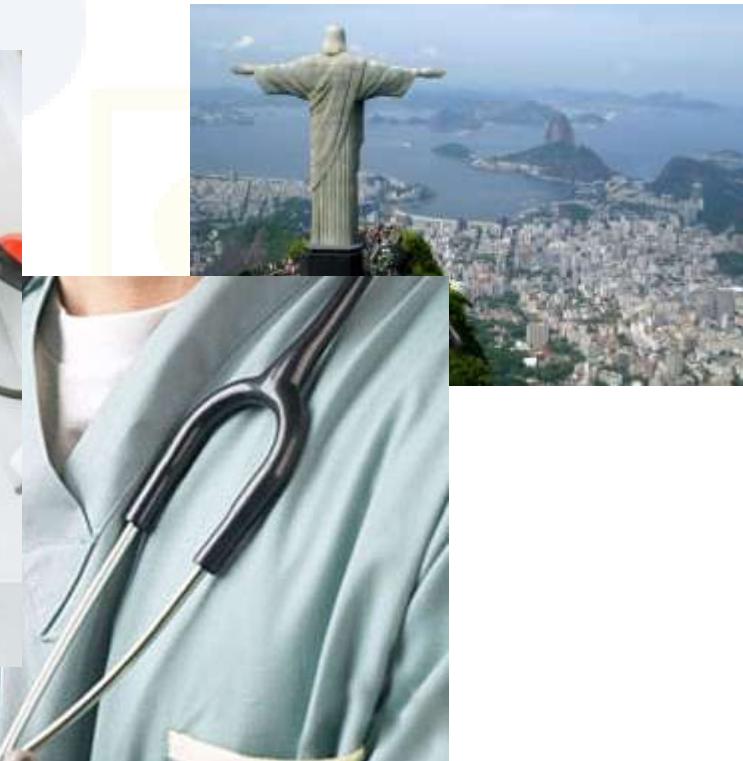
A tabela a seguir mostra o número de alunos admitidos em cursos de graduação em Engenharia e Medicina em 3 cidades brasileiras:

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



Data Science Academy

Vamos definir os eventos sob análise:



Data Science Academy



Vamos definir os eventos sob análise:

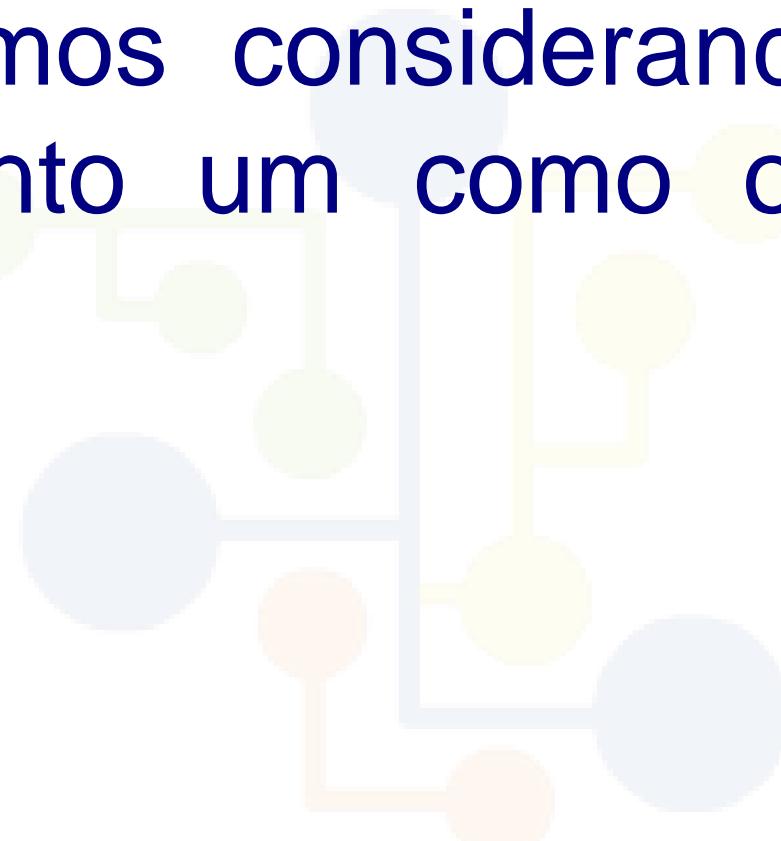
Evento A – estudante do Rio de Janeiro admitido em curso de Engenharia ou Medicina.

Evento B – estudante de qualquer cidade admitido em Engenharia.



Data Science Academy

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



Data Science Academy

Cidade	Engenharia	Medicina	Total
Rio de Janeiro	1500	2300	3800
Total	8080	11200	19280

Resposta:

$$\text{Evento A} = 1500 + 2300 = 3800$$

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



Data Science Academy

Resposta:

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

$$\text{Evento B} = 1500 + 5600 + 980 = \boxed{8080}$$

Evento B – estudante de qualquer cidade admitido em Engenharia.



Data Science Academy

Resposta:

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

$$\text{Evento A} = 1500 + 2300 = 3800$$

$$\text{Evento B} = 1500 + 5600 + 980 = 8080$$

A soma dos 2 eventos é $3800 + 8080 = \boxed{11880}$

$$\text{Evento A} = 1500 + 2300 = 3800$$

$$\text{Evento B} = 1500 + 5600 + 980 = 8080$$

A soma dos 2 eventos é $3800 + 8080 = \boxed{11880}$

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

Resposta:

A probabilidade de A ou B ocorrer, é:

$$P(A \text{ ou } B) = 11880 / 19280 = 0.62$$



Data Science Academy

Resposta:

A probabilidade de A ou B ocorrer, é:

$$P(A \text{ ou } B) = 11880 / 19280 = 0.62$$

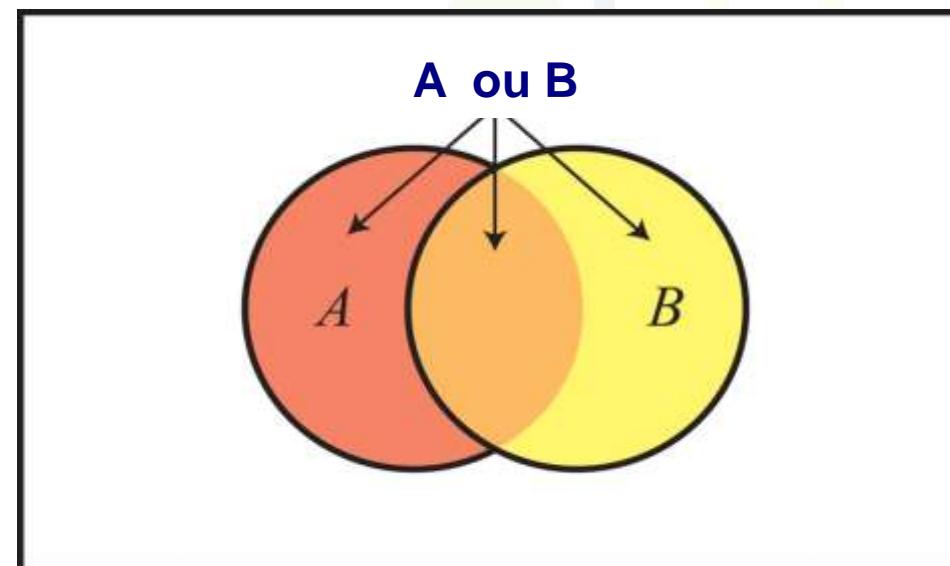


Data Science Academy

Resposta:

A probabilidade de A ou B ocorrer, é:

$$P(A \text{ ou } B) = 11880 / 19280 = 0.62$$



62%



Data Science Academy

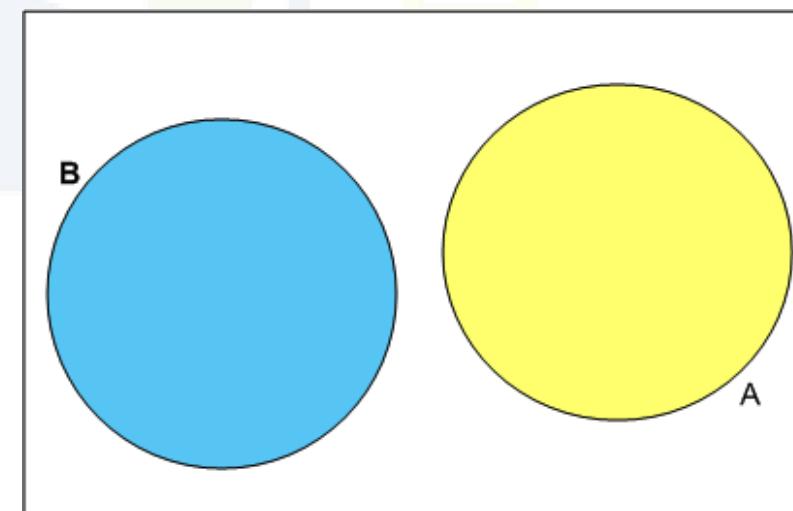
Adição de Eventos

A

B

Data Science Academy

Agora, vamos usar a **Regra de Adição** em probabilidade, para calcular a probabilidade de **união de eventos**, ou seja, a probabilidade do **Evento A, Evento B ou ambos**, ocorrerem.



Data Science Academy



Antes, precisamos entender dois conceitos muito importantes:

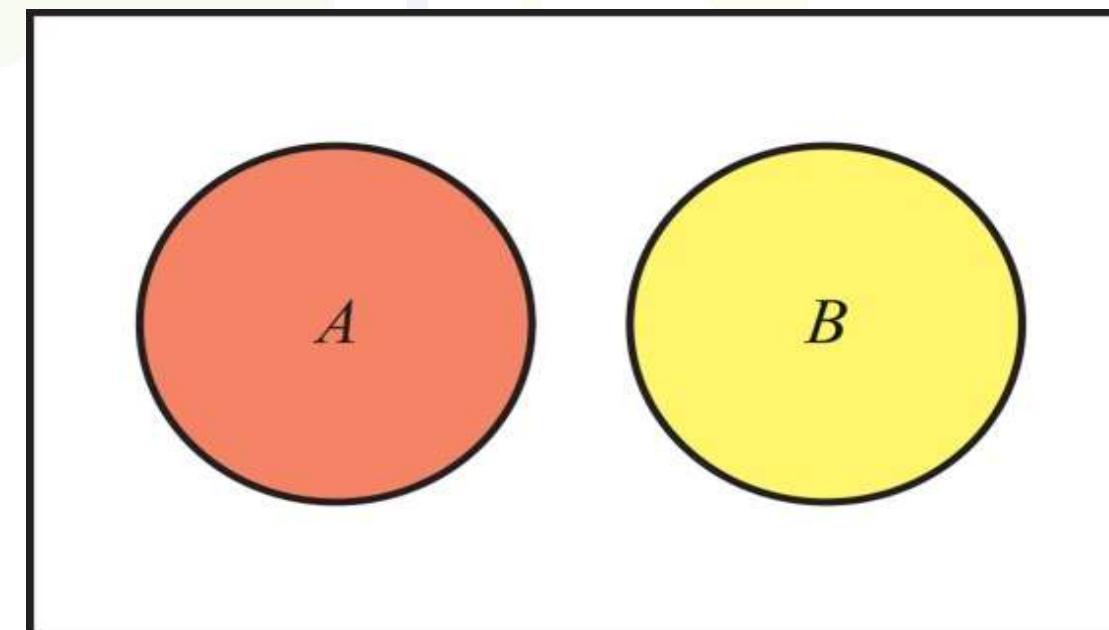
Eventos Mutuamente Exclusivos

Eventos Não Mutuamente Exclusivos



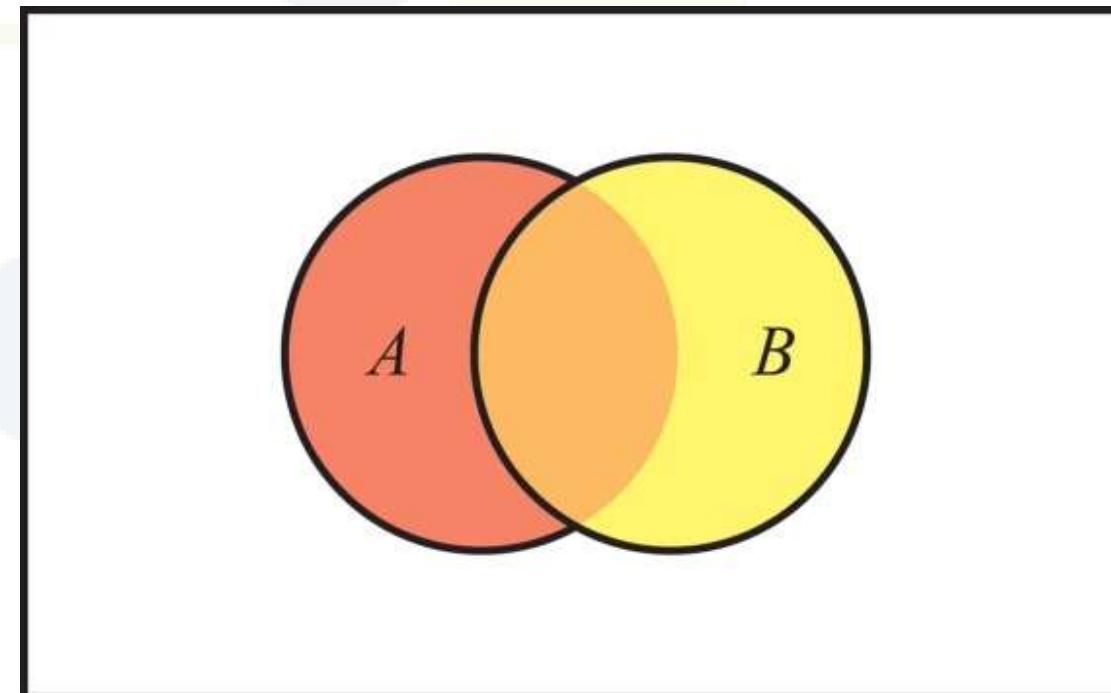
Data Science Academy

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



Data Science Academy

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



Data Science Academy

Regra da adição - depende se 2 eventos são ou não mutuamente exclusivos. Vejamos

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



Data Science Academy

Vamos definir os eventos deste experimento:



Data Science Academy

Vamos definir os eventos deste experimento:

Evento A – estudante com nota final igual a 90.

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



Data Science Academy

Vamos definir os eventos deste experimento:

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



Data Science Academy

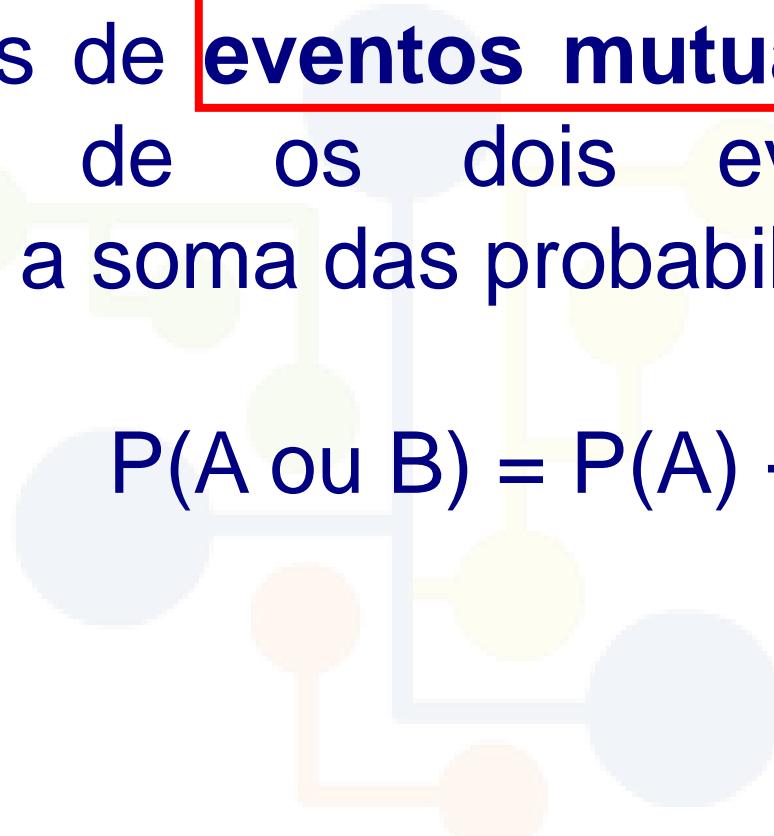
Neste caso, os eventos são **mutuamente exclusivos**, ou seja, um estudante não pode obter notas 90 e 85 no mesmo exame.



Data Science Academy



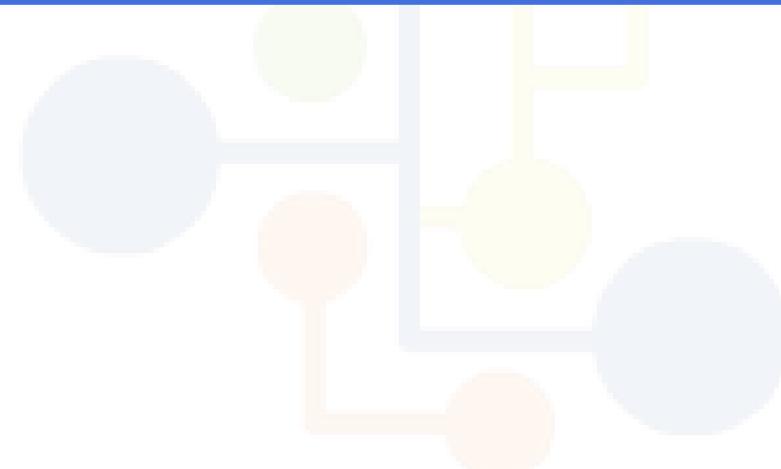
Para os casos de **eventos mutuamente exclusivos**, a probabilidade de os dois eventos ocorrerem é simplesmente a soma das probabilidades individuais.

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




Data Science Academy

Vamos aos Cálculos



Data Science Academy

Resposta:

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

Usando a tabela anterior, temos:

$$P(A) =$$



Data Science Academy

Resposta:

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

Usando a tabela anterior, temos:

$$P(A) = 120 / 250 = 0.48$$

Evento A – estudante com nota final igual a 90.



Data Science Academy

Resposta:

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

Usando a tabela anterior, temos:

$$P(B) =$$



Data Science Academy

Resposta:

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

Usando a tabela anterior, temos:

$$P(B) = 40 / 250 = 0.16$$

Evento B – estudante com nota final igual a 85.



Data Science Academy

Resposta:

$$P(A) = 120 / 250 = 0.48$$

$$P(B) = 40 / 250 = 0.16$$

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

$$\begin{aligned} P(A \text{ ou } B) &= P(A) + P(B) = \\ &0.48 + 0.16 \\ &0.64 \end{aligned}$$

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



Mas e se os 2 eventos não forem mutuamente exclusivos?



Data Science Academy

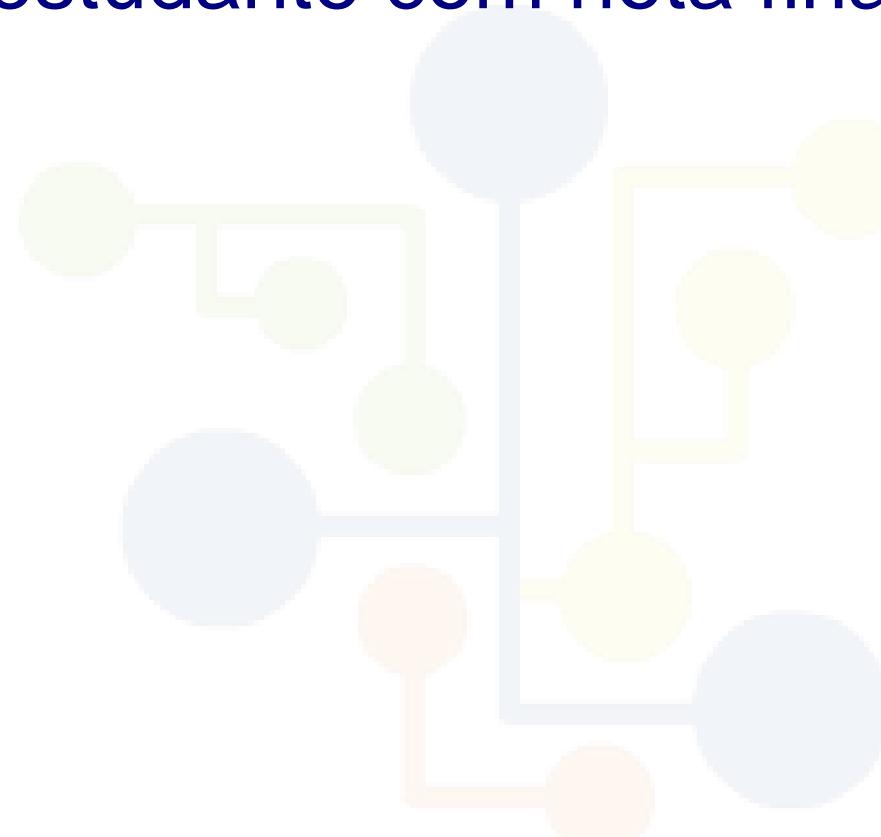
Vamos definir nossos eventos deste experimento de
Eventos **não** mutuamente exclusivos:



Data Science Academy



Evento A – estudante com nota final igual a 90.



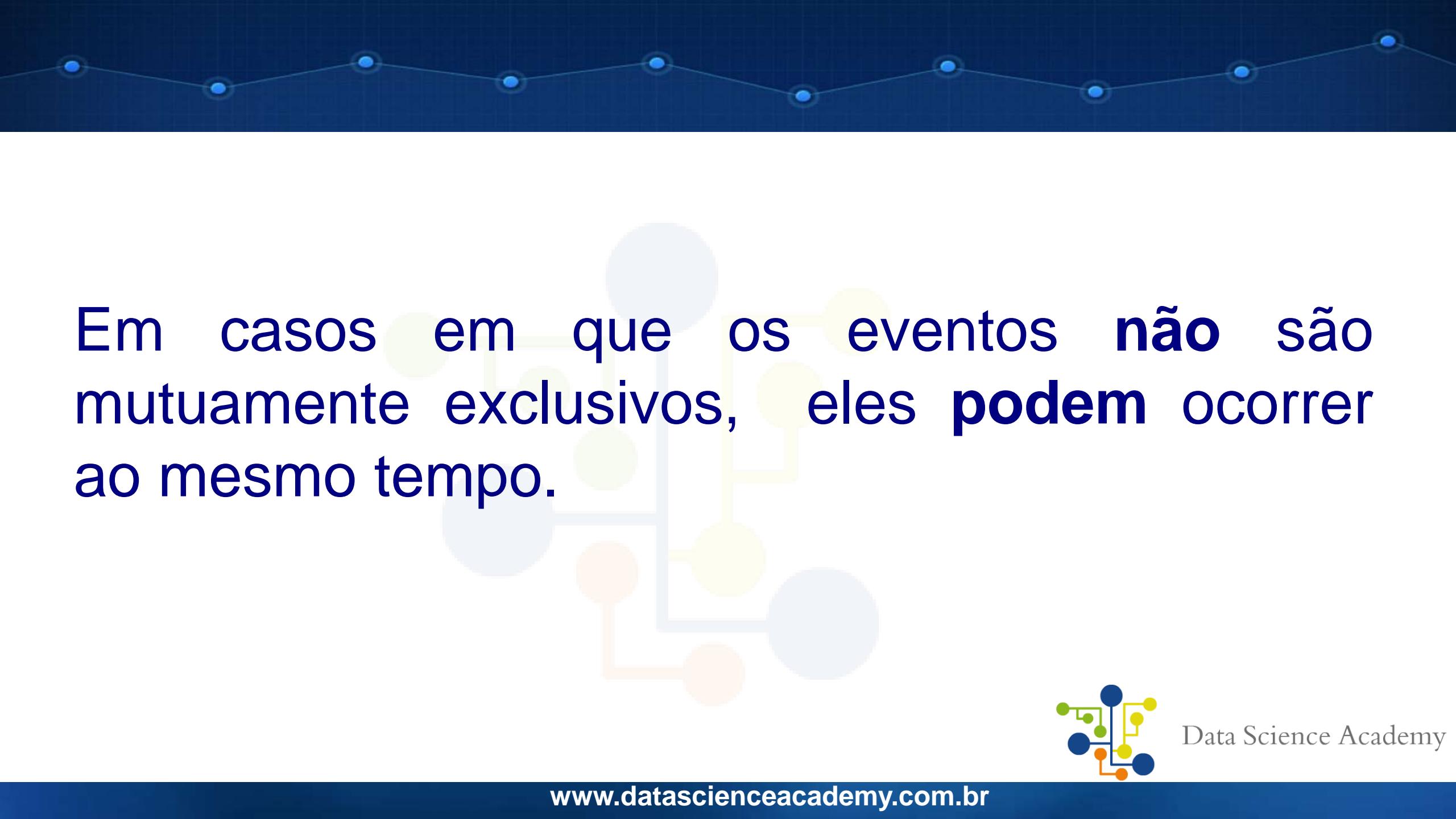
Data Science Academy



Evento B – estudante é mulher.



Data Science Academy



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



Data Science Academy

Resposta:

Calculamos a Probabilidade da seguinte forma:

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



Data Science Academy

Resposta:

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

Usando a tabela anterior, temos:

$$P(A) = 120 / 250 = 0.48$$

Evento A – estudante com nota final igual a 90.



Data Science Academy

Resposta:

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

Usando a tabela anterior, temos:

$$P(B) = 150 / 250 = 0.60$$

Evento B – estudante é mulher.



Data Science Academy

Resposta:

$$P(A) = 120 / 250 = 0.48$$

$$P(B) = 150 / 250 = 0.60$$

$$P(A \text{ e } B) = 80 / 250 = 0.32$$

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

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



Data Science Academy

Resposta:

$$P(A) = 120 / 250 = 0.48$$

$$P(B) = 150 / 250 = 0.60$$

$$P(A \text{ e } B) = 80 / 250 = 0.32$$

$$\begin{aligned} P(A \text{ ou } B) &= P(A) + P(B) - P(A \text{ e } B) \\ &0.48 + 0.60 - 0.32 \end{aligned}$$

0.76



Data Science Academy

Resposta:

$$P(A) = 120 / 250 = 0.48$$

$$P(B) = 150 / 250 = 0.60$$

$$P(A \text{ e } B) = 80 / 250 = 0.32$$

$$\begin{aligned} P(A \text{ ou } B) &= P(A) + P(B) - P(A \text{ e } B) \\ &0.48 + 0.60 - 0.32 \end{aligned}$$

0.76

76%

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



Data Science Academy

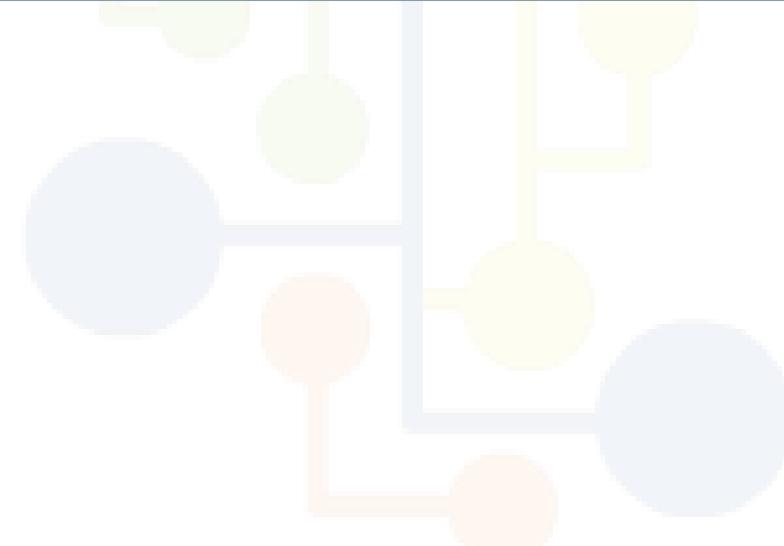
Esse tópico chegou ao final



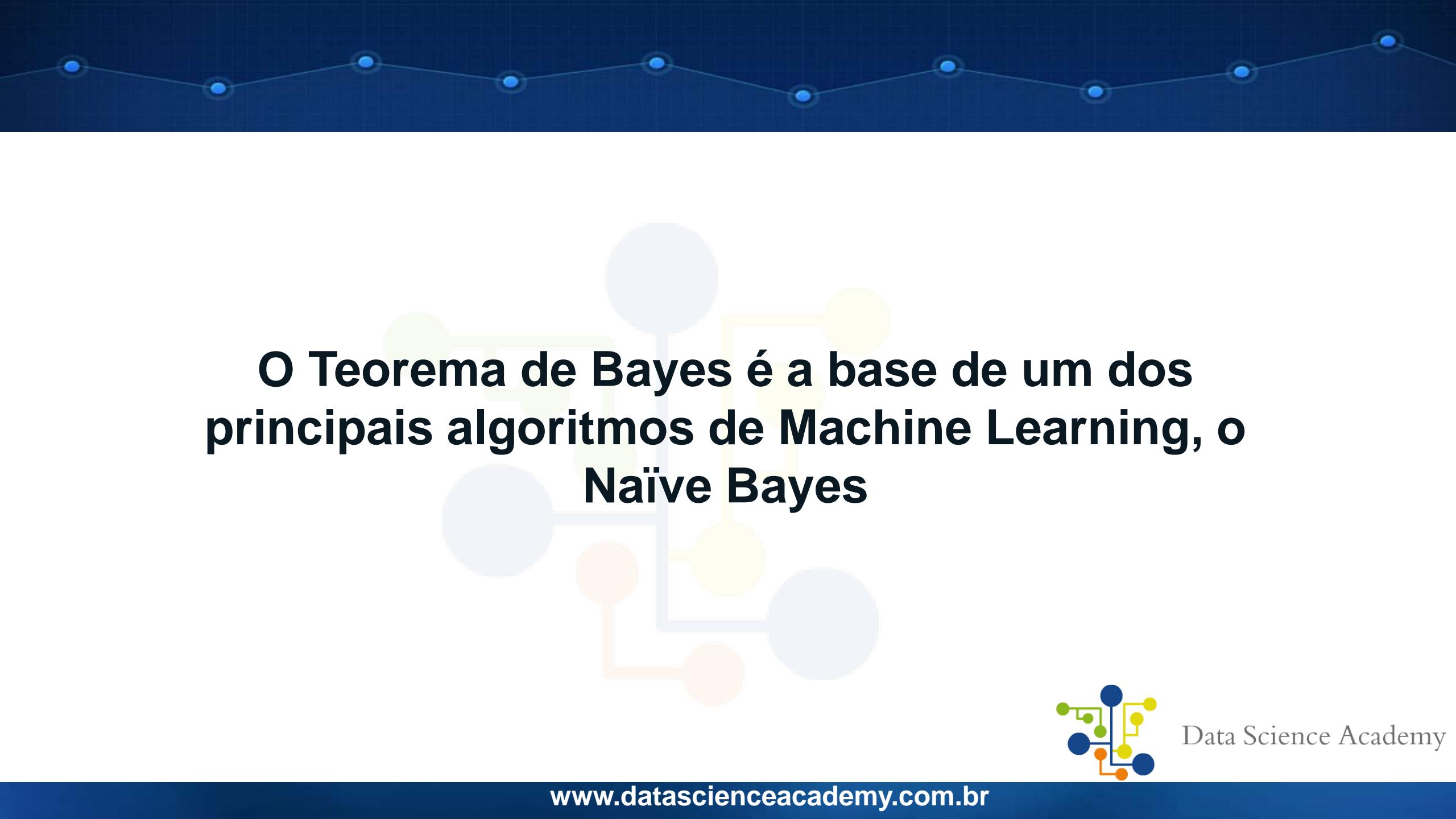
Data Science Academy



Teorema de Bayes



Data Science Academy



O Teorema de Bayes é a base de um dos principais algoritmos de Machine Learning, o Naïve Bayes



Data Science Academy



Thomas Bayes (1701-1761)

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

Desenvolveu a regra matemática que calcula $P(A|B)$ a partir da informação de $P(B|A)$.



Data Science Academy

Ou seja, se soubermos alguma coisa sobre a probabilidade do **evento B** ocorrer, considerando que o evento A já ocorreu, nós podemos calcular o **reverso**, ou seja, podemos calcular a probabilidade do **evento A** ocorrer, considerando que o evento B ocorra.

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$



Data Science Academy

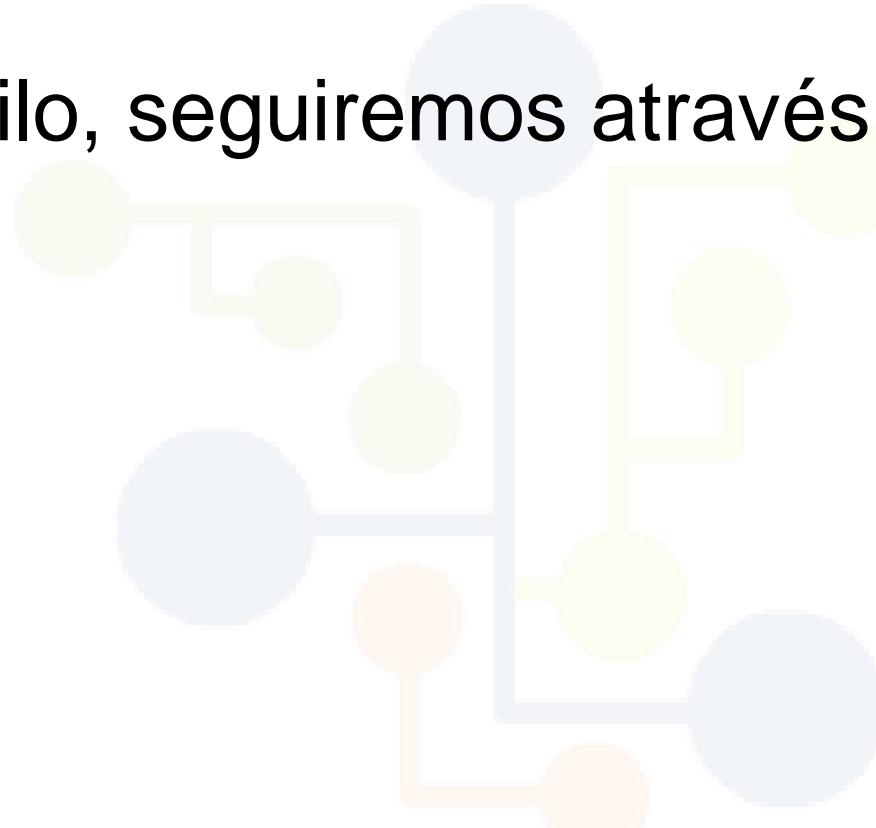
Parece confuso?



Data Science Academy



Fique tranquilo, seguiremos através de exemplos.



Data Science Academy

Exemplos

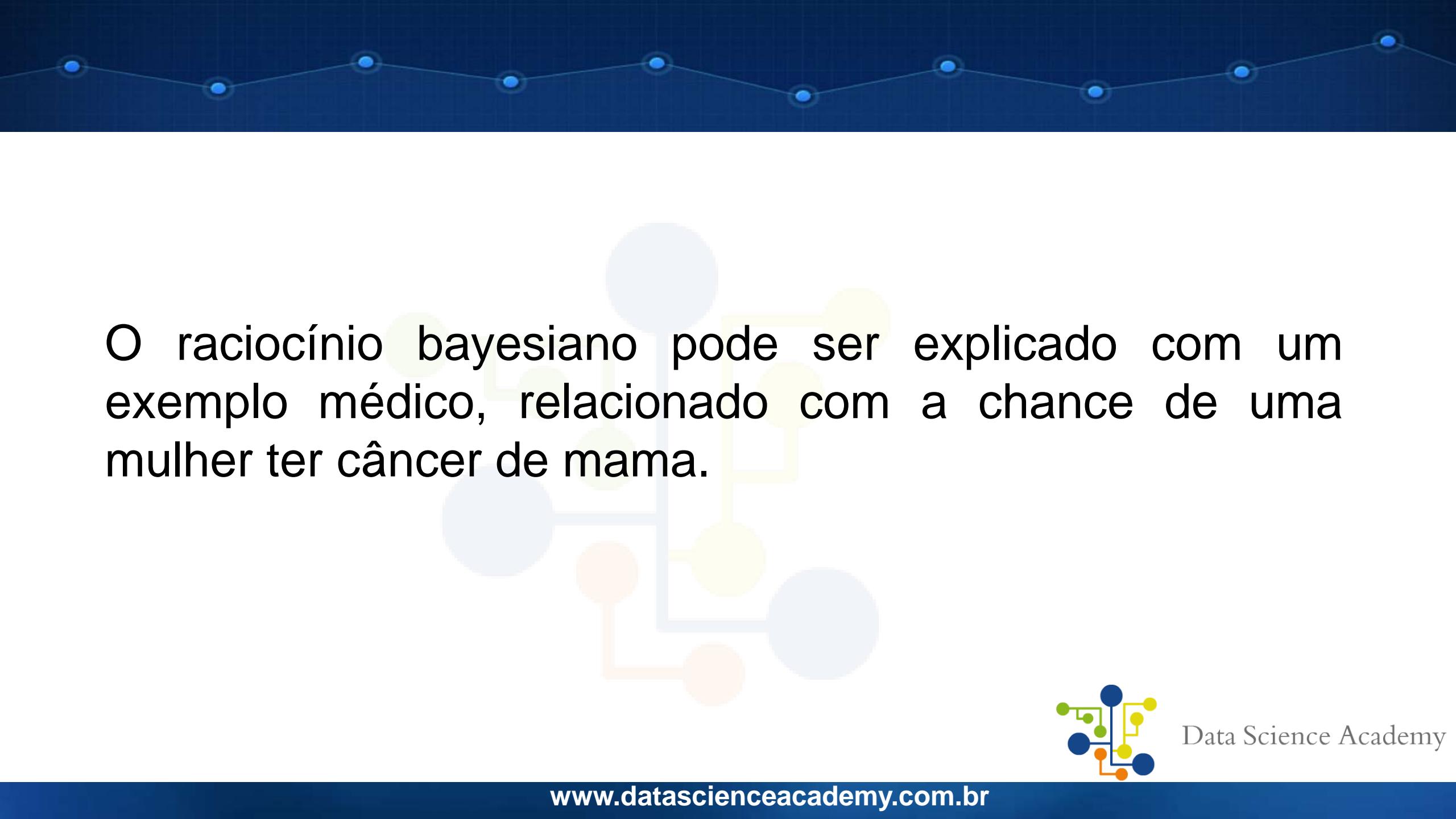


Data Science Academy

O **Teorema de Bayes** permite determinar a probabilidade de identificar um terrorista em um aeroporto através de procedimentos de segurança e avaliar a efetividade de testes de doping em atletas.



Data Science Academy



O raciocínio bayesiano pode ser explicado com um exemplo médico, relacionado com a chance de uma mulher ter câncer de mama.



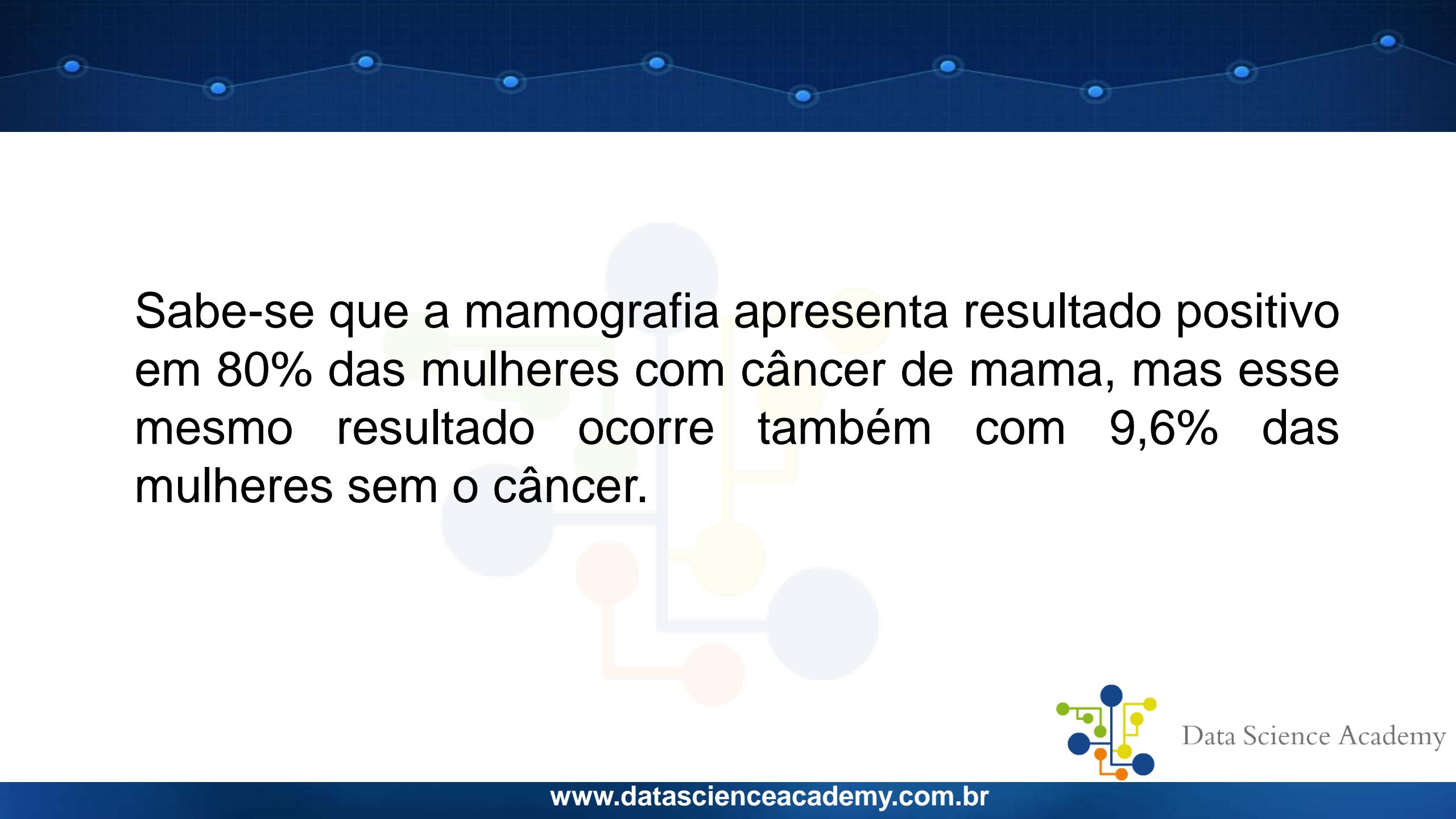
Data Science Academy



Recomenda-se que, a partir dos 40 anos, as mulheres façam mamografias anuais. Nessa idade, 1% das mulheres são portadoras de um tumor assintomático de mama.



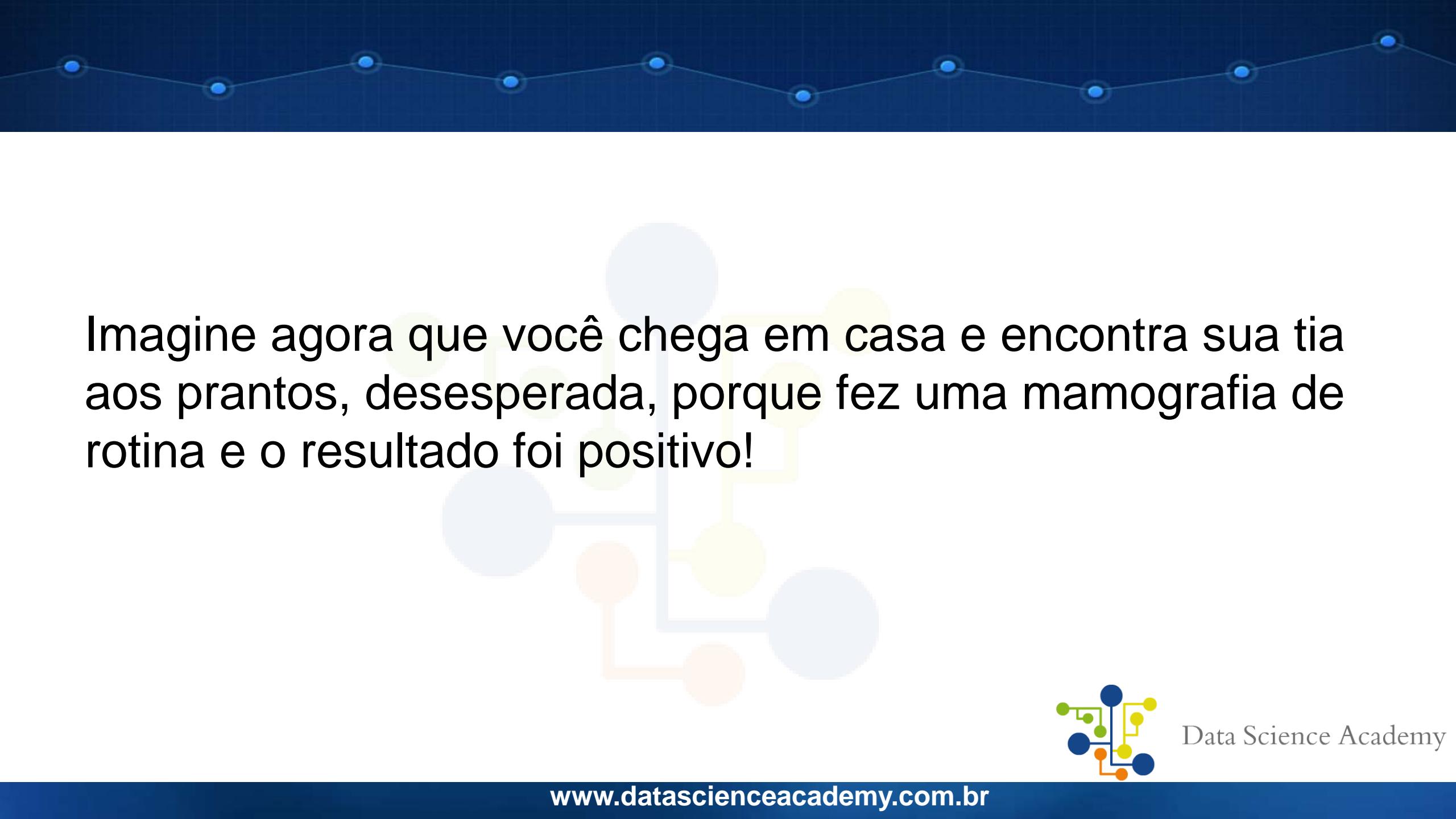
Data Science Academy



Sabe-se que a mamografia apresenta resultado positivo em 80% das mulheres com câncer de mama, mas esse mesmo resultado ocorre também com 9,6% das mulheres sem o câncer.



Data Science Academy



Imagine agora que você chega em casa e encontra sua tia aos prantos, desesperada, porque fez uma mamografia de rotina e o resultado foi positivo!



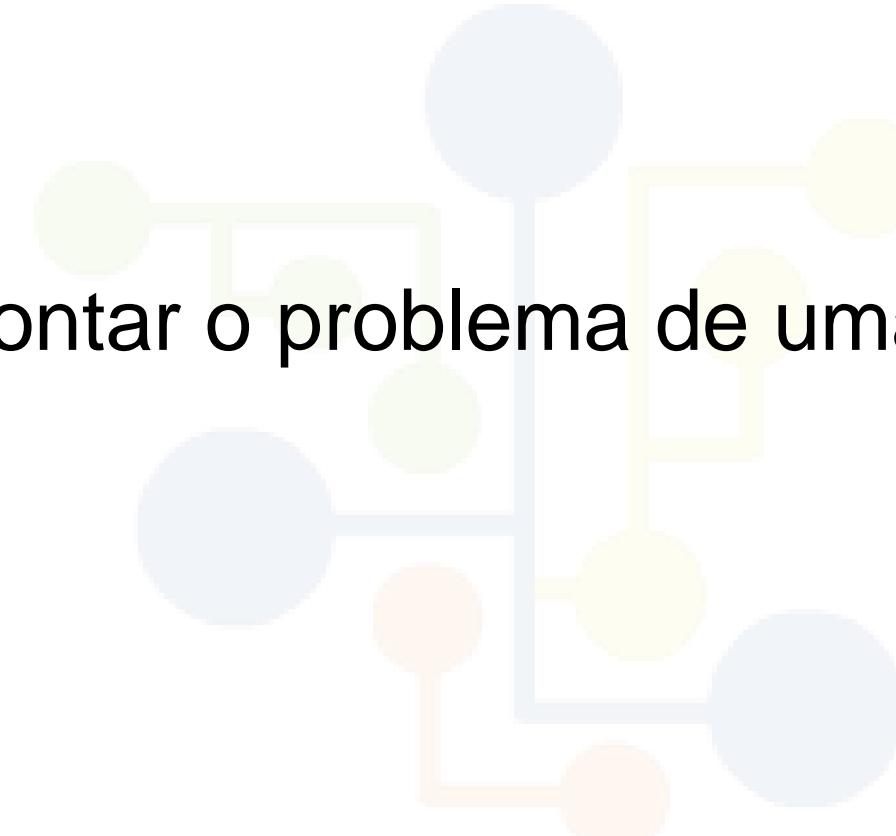
Data Science Academy



Qual seria a probabilidade dela ter um câncer de mama?



Data Science Academy

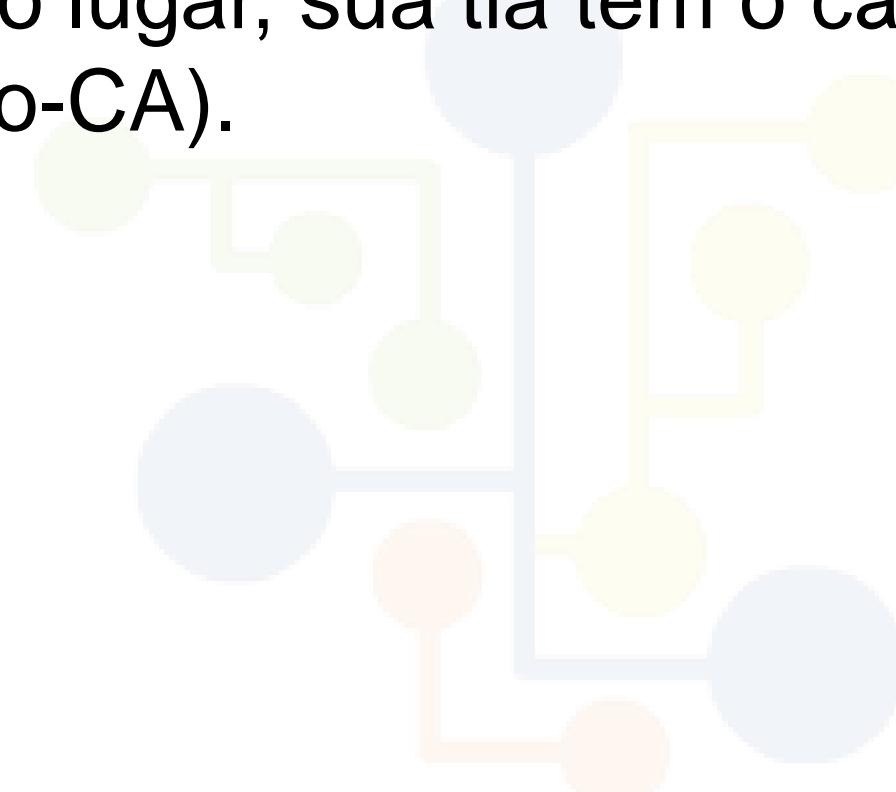


Vamos montar o problema de uma maneira bayesiana.

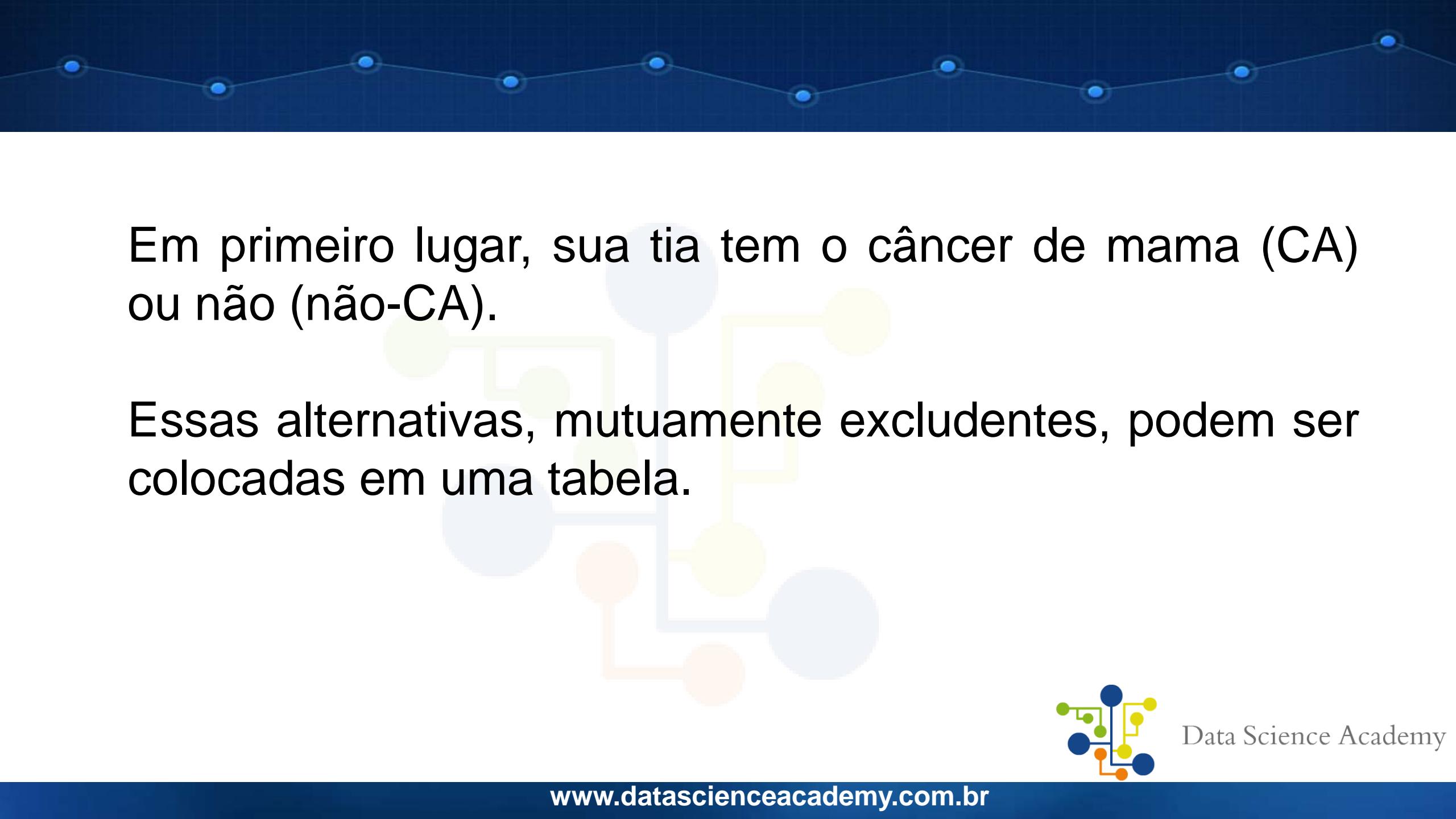


Data Science Academy

Em primeiro lugar, sua tia tem o câncer de mama (CA) ou não (não-CA).



Data Science Academy



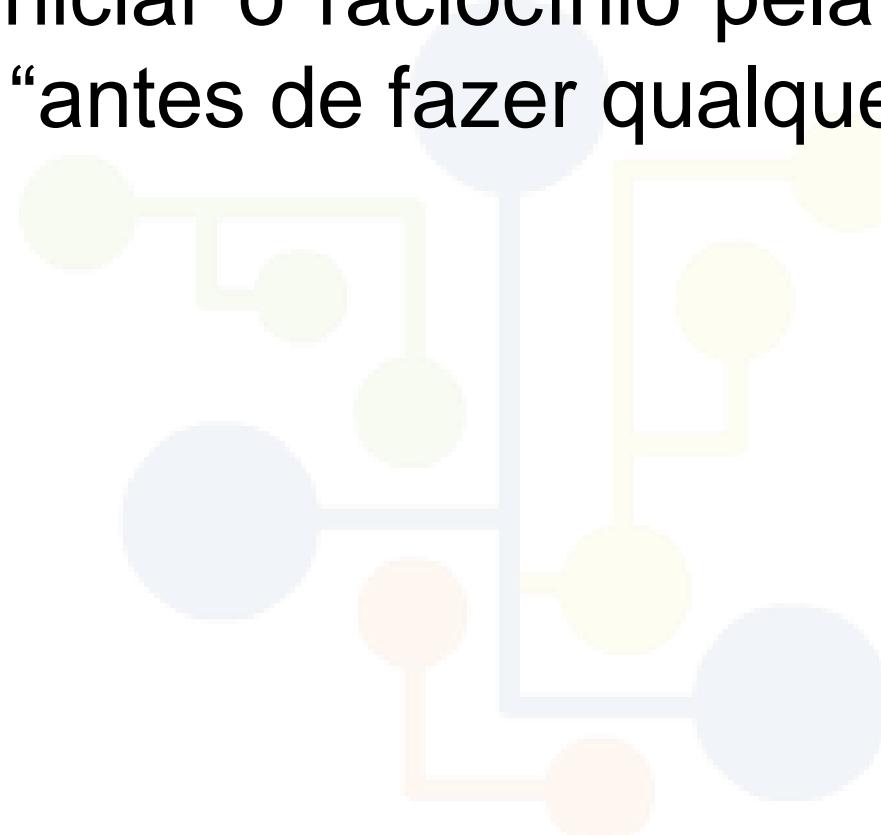
Em primeiro lugar, sua tia tem o câncer de mama (CA) ou não (não-CA).

Essas alternativas, mutuamente excludentes, podem ser colocadas em uma tabela.

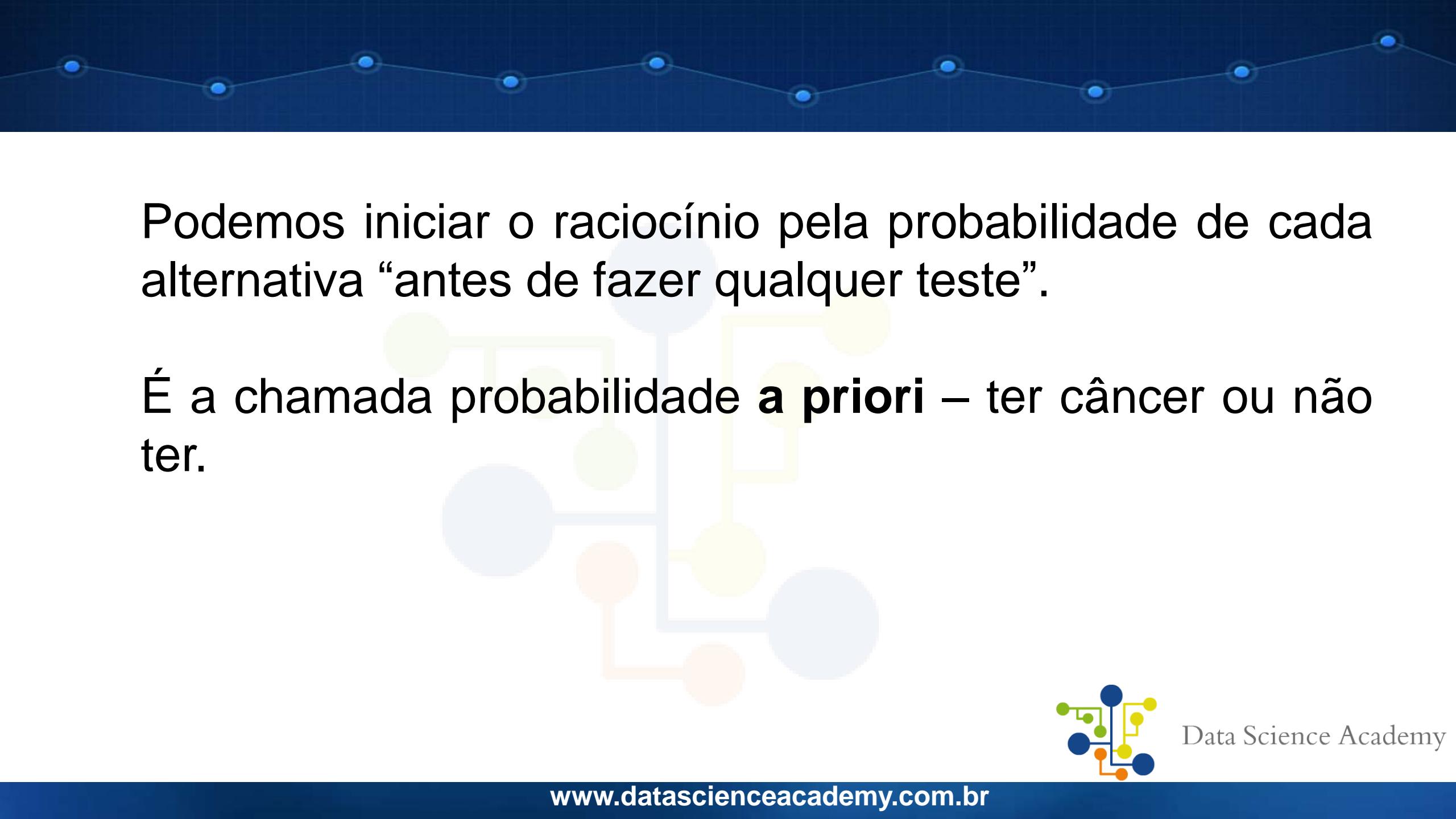


Data Science Academy

Podemos iniciar o raciocínio pela probabilidade de cada alternativa “antes de fazer qualquer teste”.



Data Science Academy



Podemos iniciar o raciocínio pela probabilidade de cada alternativa “antes de fazer qualquer teste”.

É a chamada probabilidade **a priori** – ter câncer ou não ter.



Data Science Academy

Como em média 1% das mulheres de 40 anos têm um tumor de mama, a probabilidade a priori de sua tia ter um câncer é de **1%** (0,01) e de não ter é de **99%** (0,99).

	Tem câncer	Não tem câncer
Probabilidade a priori	0,01	0,99



Data Science Academy

Agora vamos incorporar o resultado da mamografia. Se o câncer de mama está presente, a probabilidade condicional de a mamografia ser positiva é 0,80 (**80%**), e se não está presente é de 0,096 (**9,6%**).

	Tem câncer	Não tem câncer
Probabilidade a priori	0,01	0,99
Probabilidade condicional	0,8	0,096



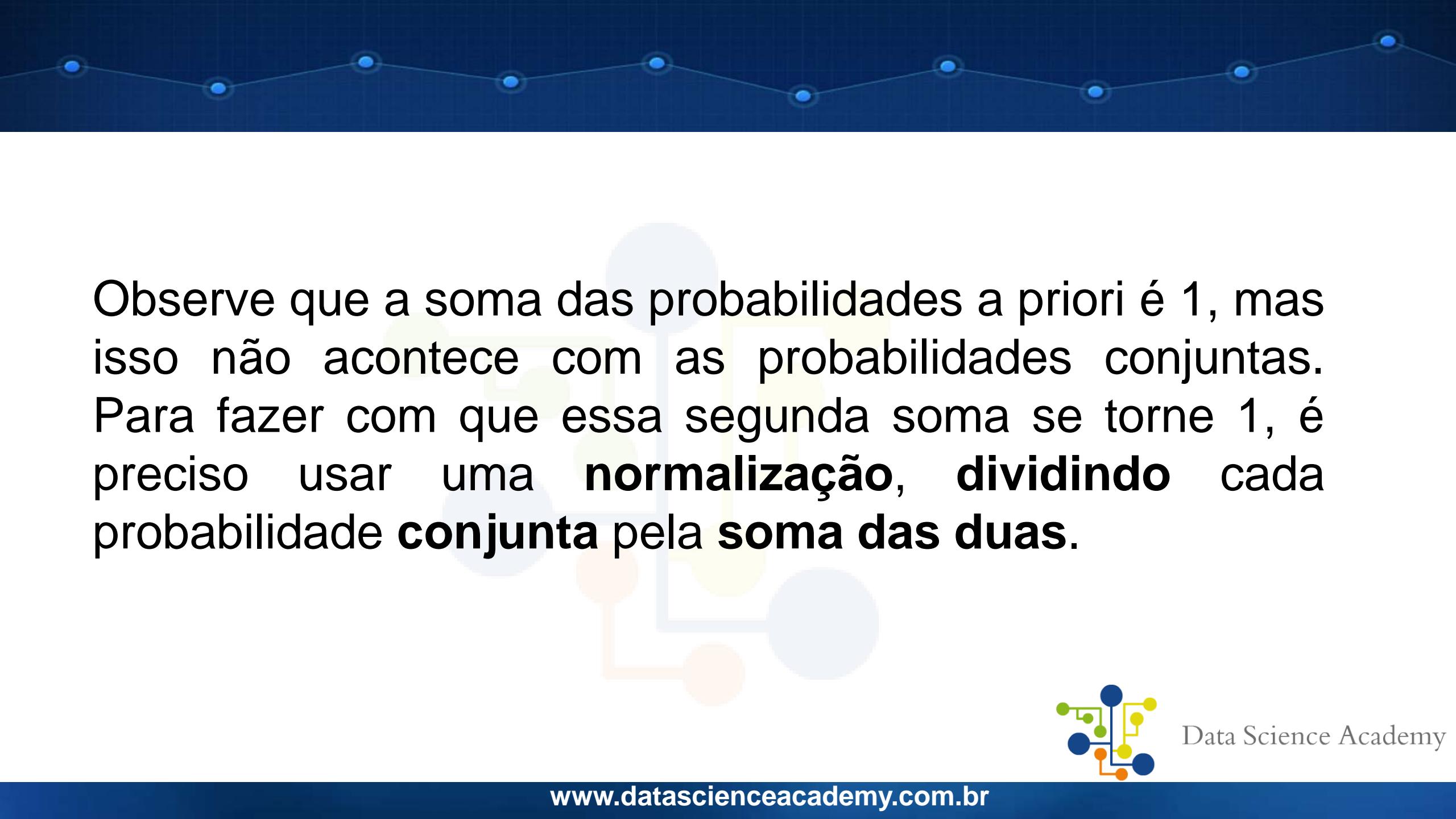
Data Science Academy

Multiplicando a probabilidade **a priori** pela **condicional**, obtemos a **probabilidade conjunta**:

	Tem câncer	Não tem câncer
Probabilidade a priori	0,01	0,99
Probabilidade condicional	0,8	0,096
Probabilidade conjunta	$0,01 \times 0,8 = 0,008$	$0,99 \times 0,096 = 0,0095$



Data Science Academy



Observe que a soma das probabilidades a priori é 1, mas isso não acontece com as probabilidades conjuntas. Para fazer com que essa segunda soma se torne 1, é preciso usar uma **normalização**, dividindo cada probabilidade conjunta pela **soma das duas**.



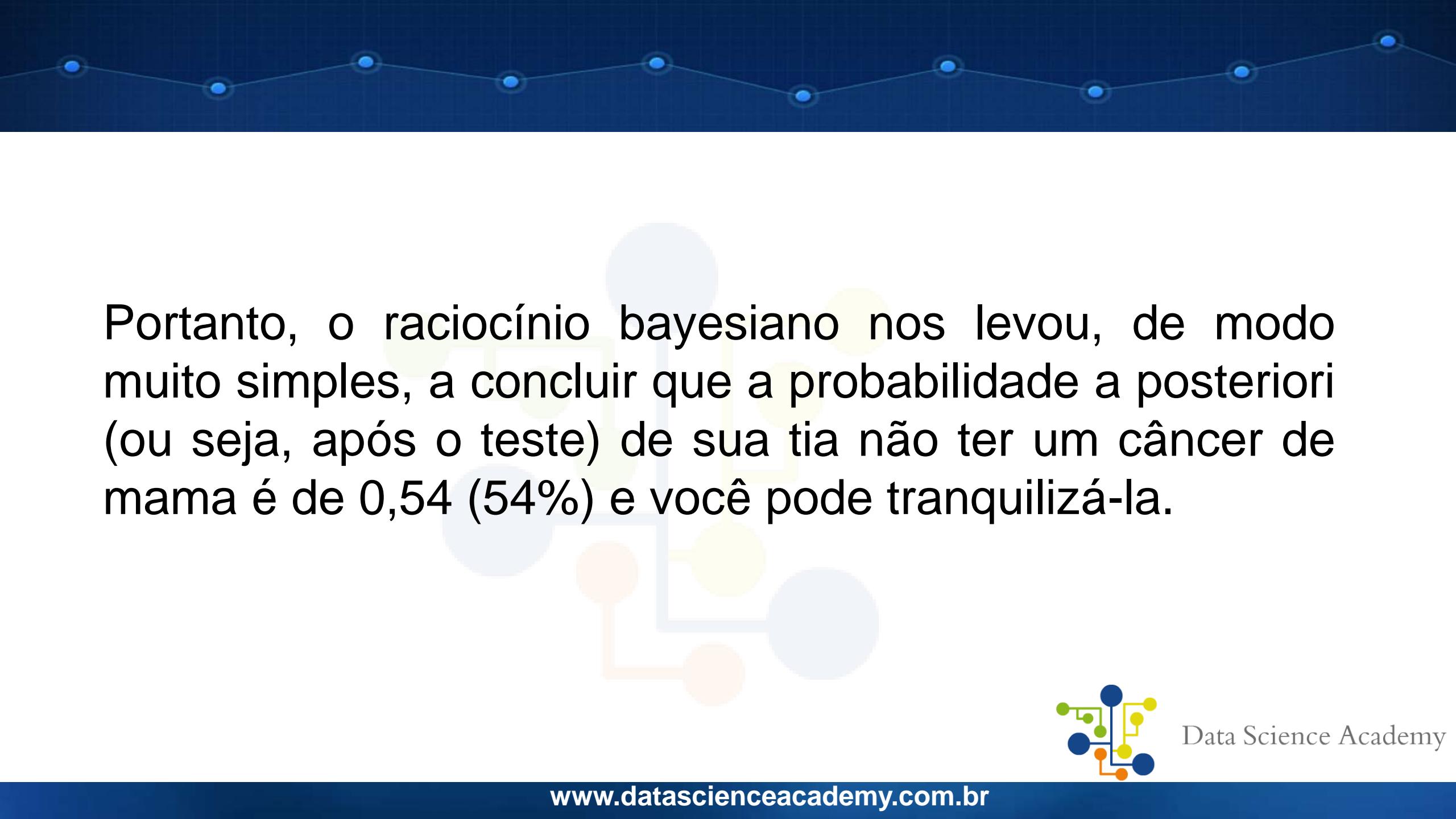
Data Science Academy

Chegamos assim à chamada probabilidade a posteriori.

	Tem câncer	Não tem câncer
Probabilidade a priori	0,01	0,99
Probabilidade condicional	0,8	0,096
Probabilidade conjunta	$0,01 \times 0,8 = 0,008$	$0,99 \times 0,096 = 0,0095$
Normalização	$(0,008 + 0,0095 = 0,0175)$	
Probabilidade a posteriori	$0,008 / 0,0175 = 0,46$	$0,0095 / 0,0175 = 0,54$



Data Science Academy



Portanto, o raciocínio bayesiano nos levou, de modo muito simples, a concluir que a probabilidade a posteriori (ou seja, após o teste) de sua tia não ter um câncer de mama é de 0,54 (54%) e você pode tranquilizá-la.



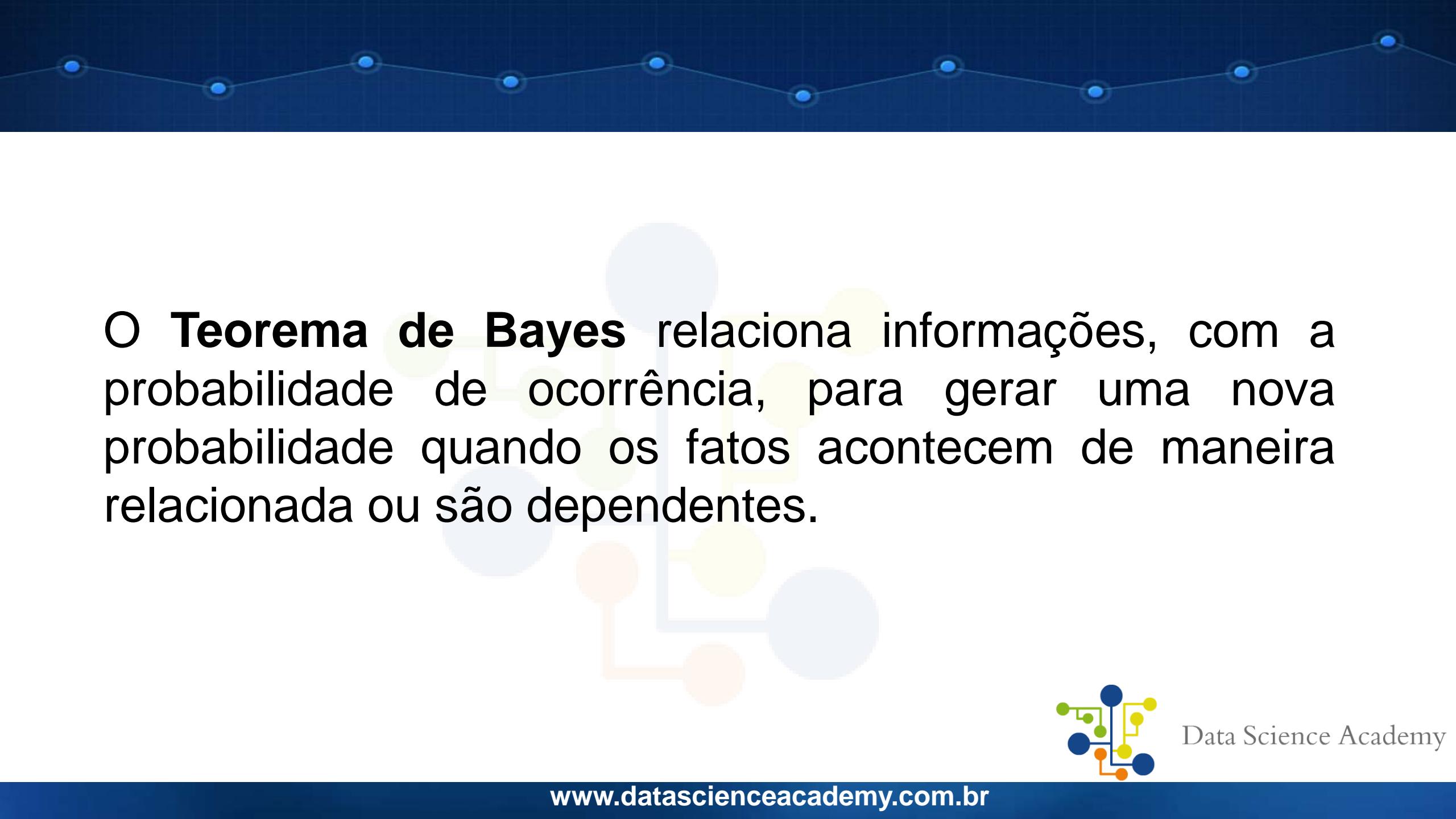
Data Science Academy

Esta é a fórmula para o teorema de Bayes:

$$P(B_i | A) = \frac{P(A | B_i)P(B_i)}{P(A | B_1)P(B_1) + P(A | B_2)P(B_2) + \dots + P(A | B_n)P(B_n)}$$



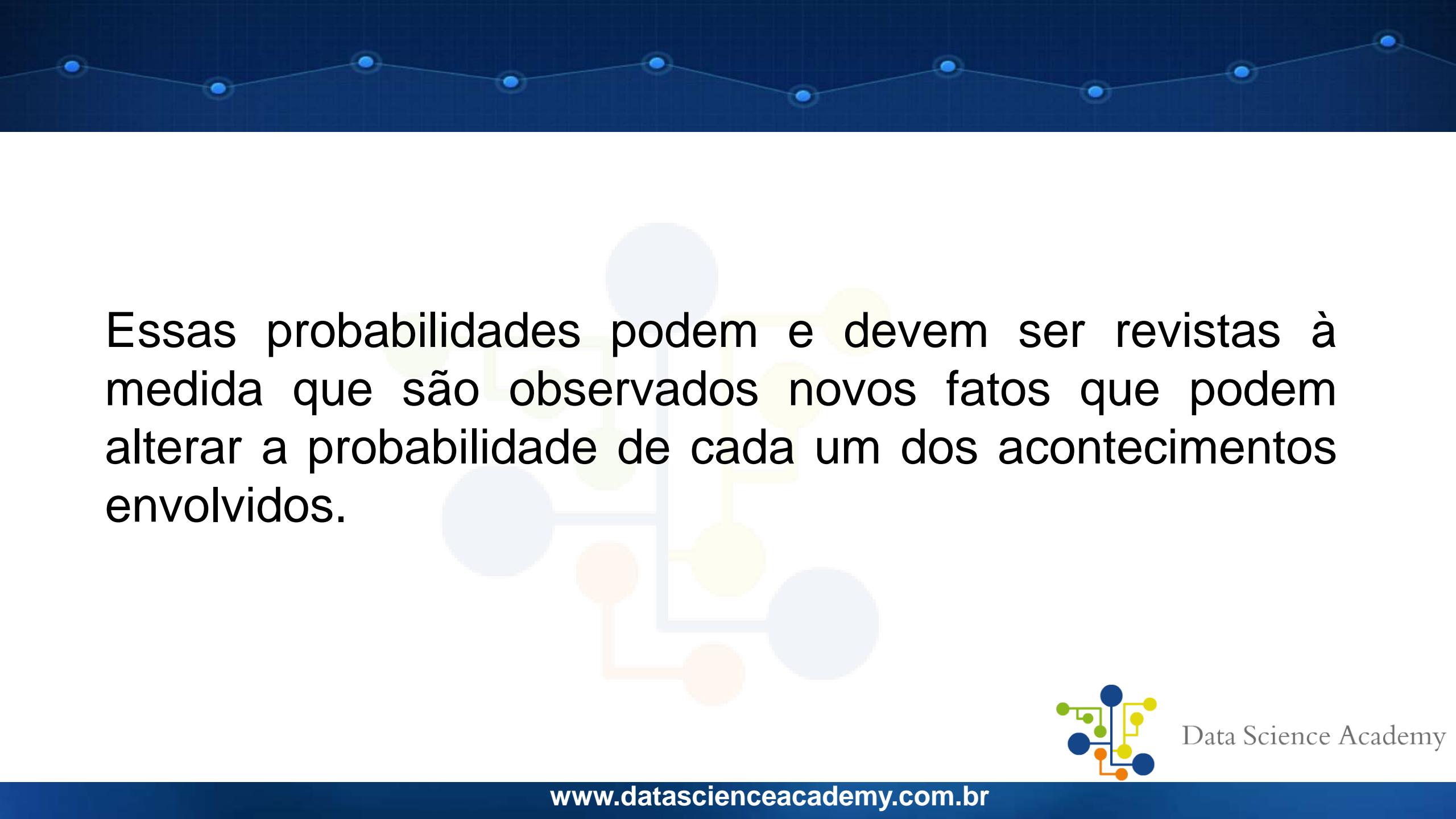
Data Science Academy



O **Teorema de Bayes** relaciona informações, com a probabilidade de ocorrência, para gerar uma nova probabilidade quando os fatos acontecem de maneira relacionada ou são dependentes.



Data Science Academy

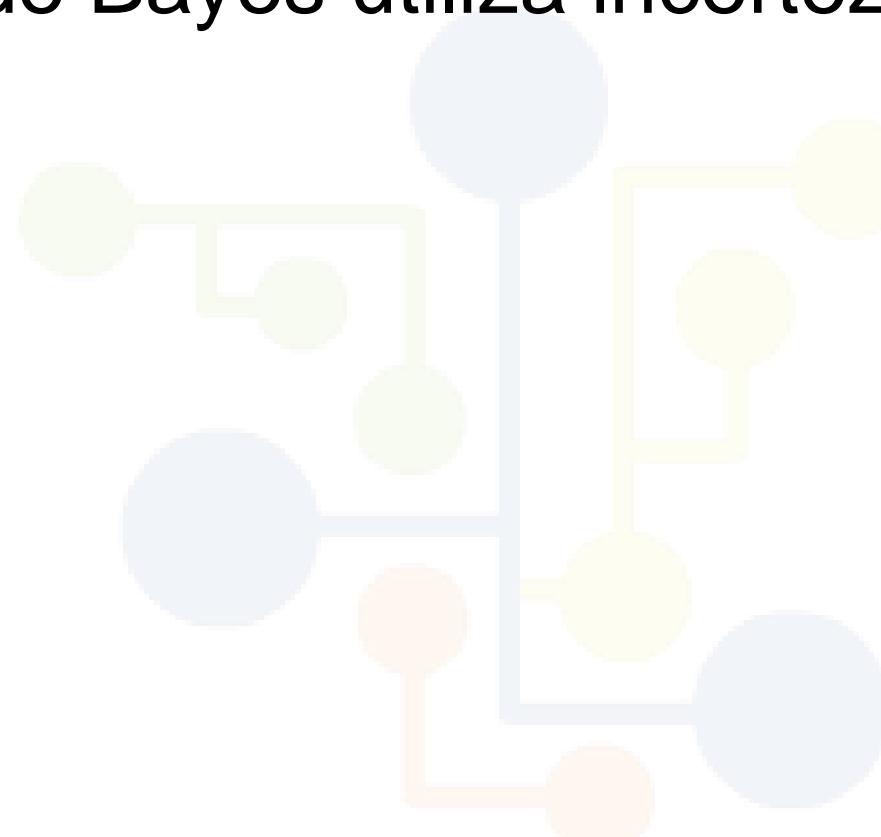


Essas probabilidades podem e devem ser revistas à medida que são observados novos fatos que podem alterar a probabilidade de cada um dos acontecimentos envolvidos.



Data Science Academy

O Teorema de Bayes utiliza incertezas.



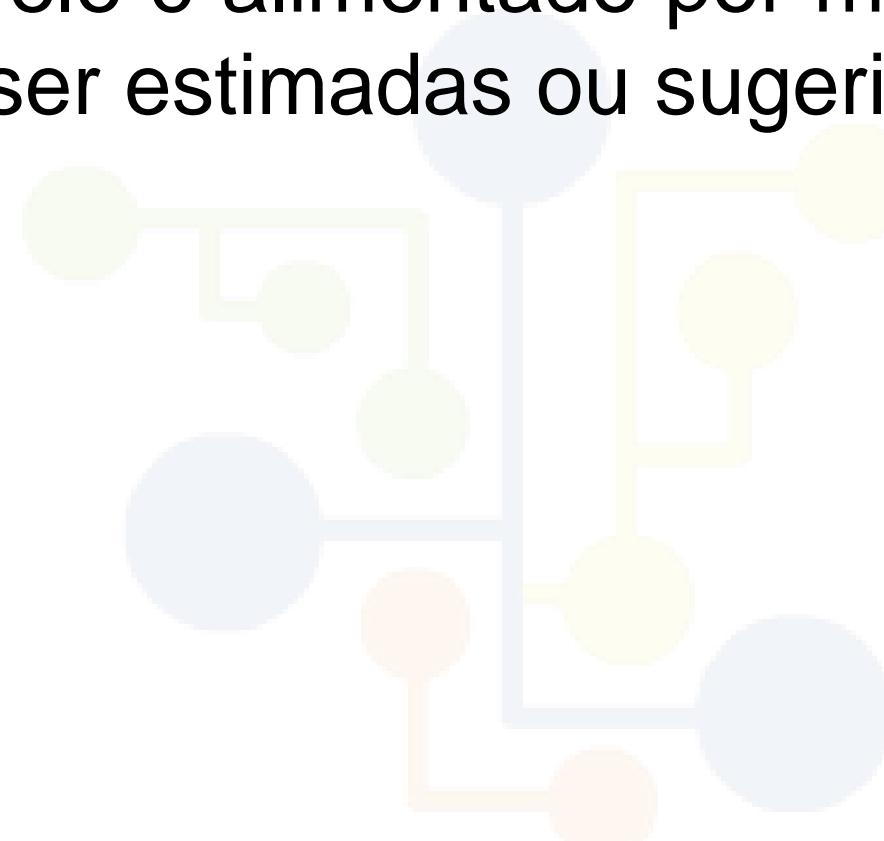
Data Science Academy

Entretanto,

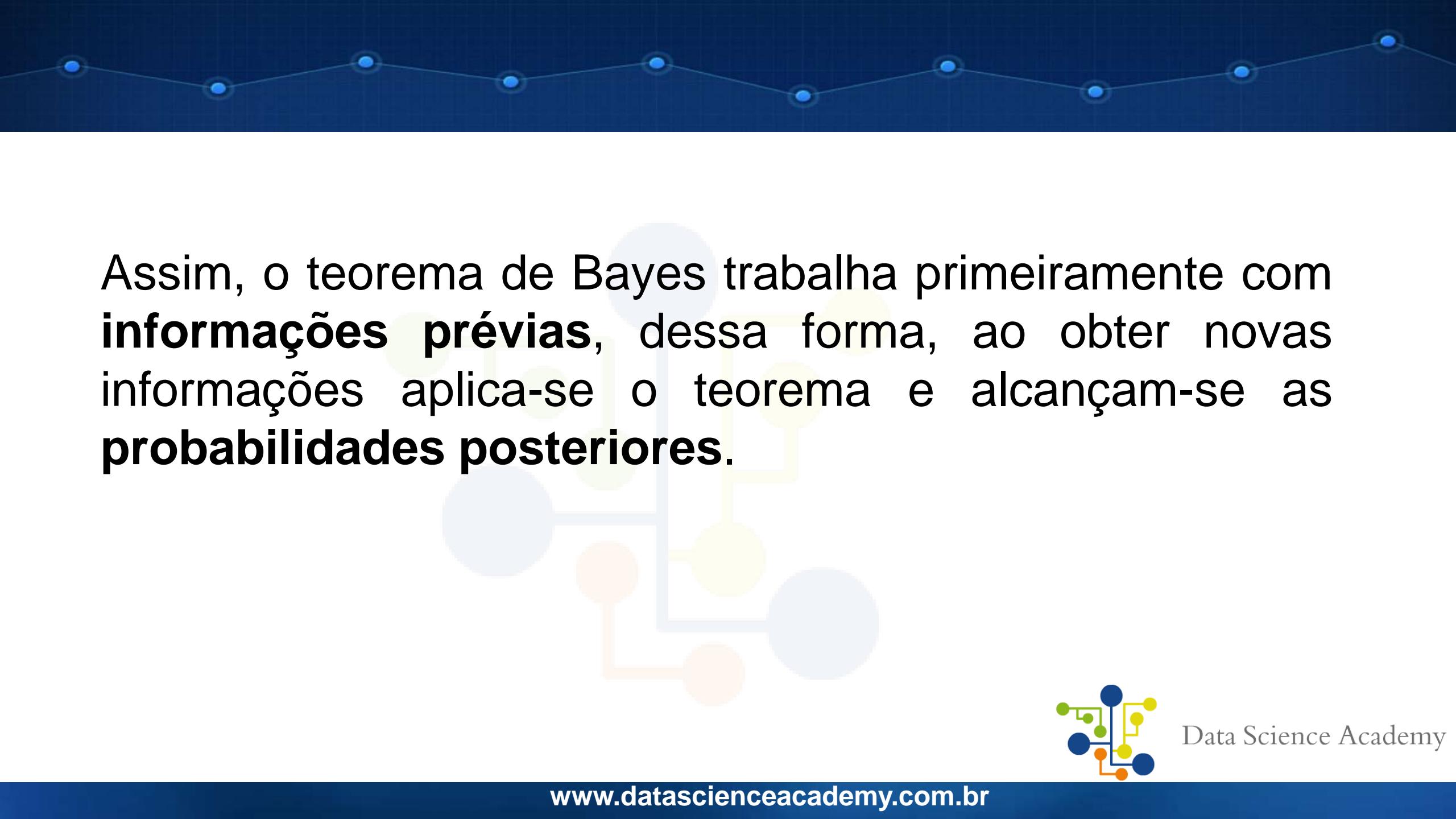


Data Science Academy

Entretanto, ele é alimentado por meio de probabilidades que podem ser estimadas ou sugeridas.



Data Science Academy



Assim, o teorema de Bayes trabalha primeiramente com **informações prévias**, dessa forma, ao obter novas informações aplica-se o teorema e alcançam-se as **probabilidades posteriores**.



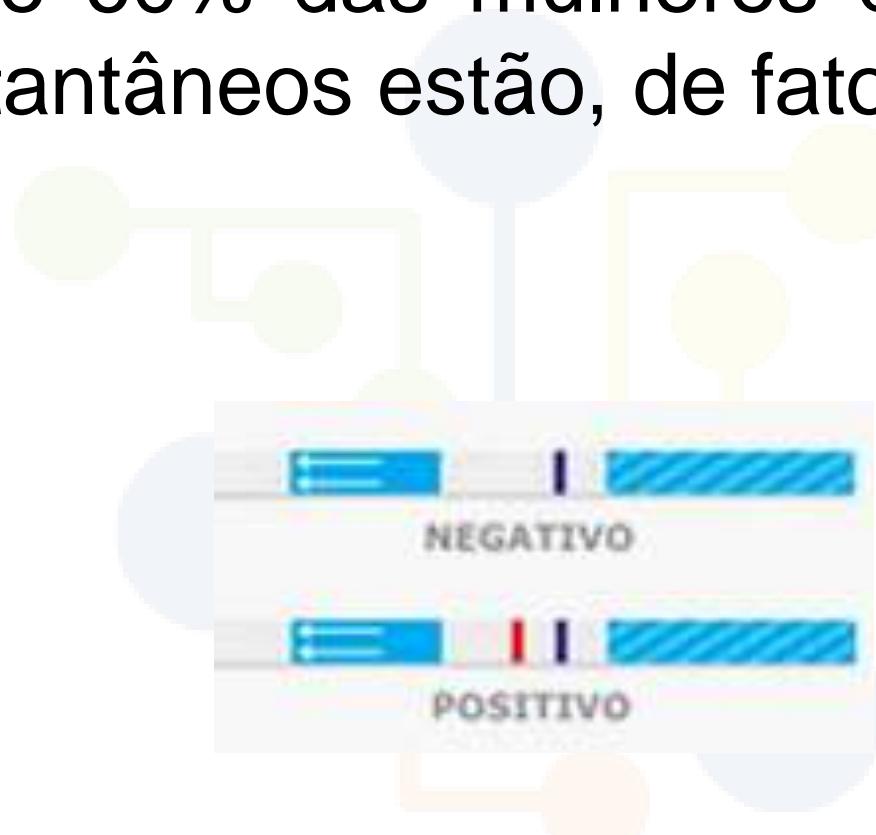
Data Science Academy

Exemplos



Data Science Academy

Suponha que 60% das mulheres que compram kits de gravidez instantâneos estão, de fato, grávidas.



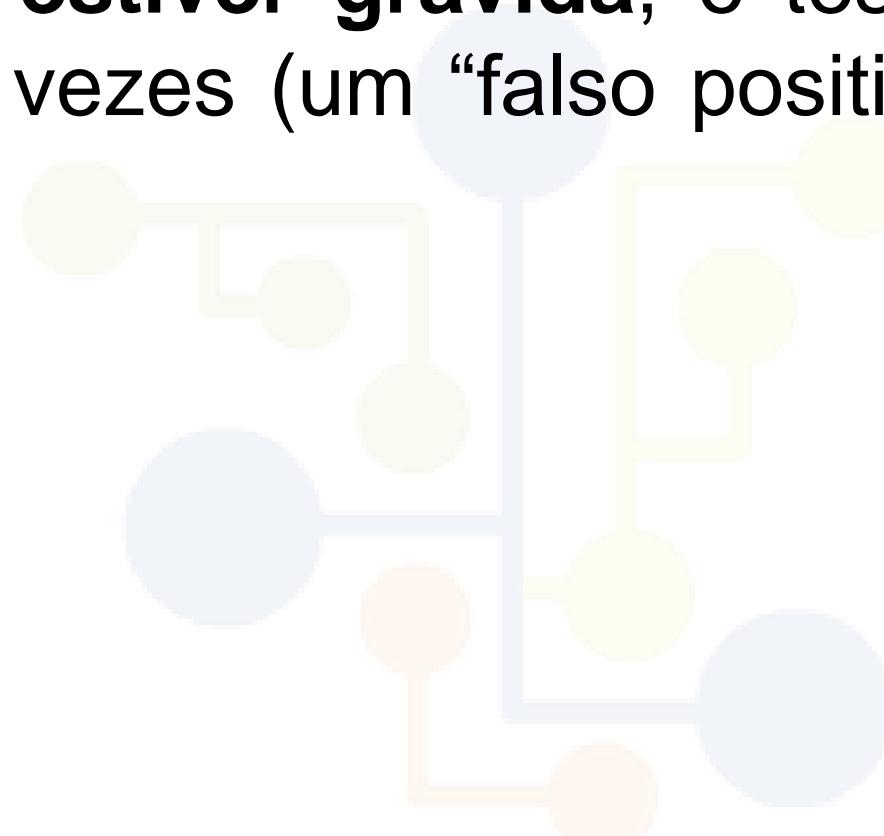
Data Science Academy

Para um kit de uma marca específica, se a mulher estiver grávida, o teste fornecerá resultado positivo **96%** das vezes e negativo **4%** das vezes (um “falso negativo”).



Data Science Academy

Se ela **não estiver grávida**, o teste resultará **positivo** em **1%** das vezes (um “falso positivo”) e **negativo** **99%** das vezes.



Data Science Academy

Suponha que um teste resulte positivo.



Data Science Academy

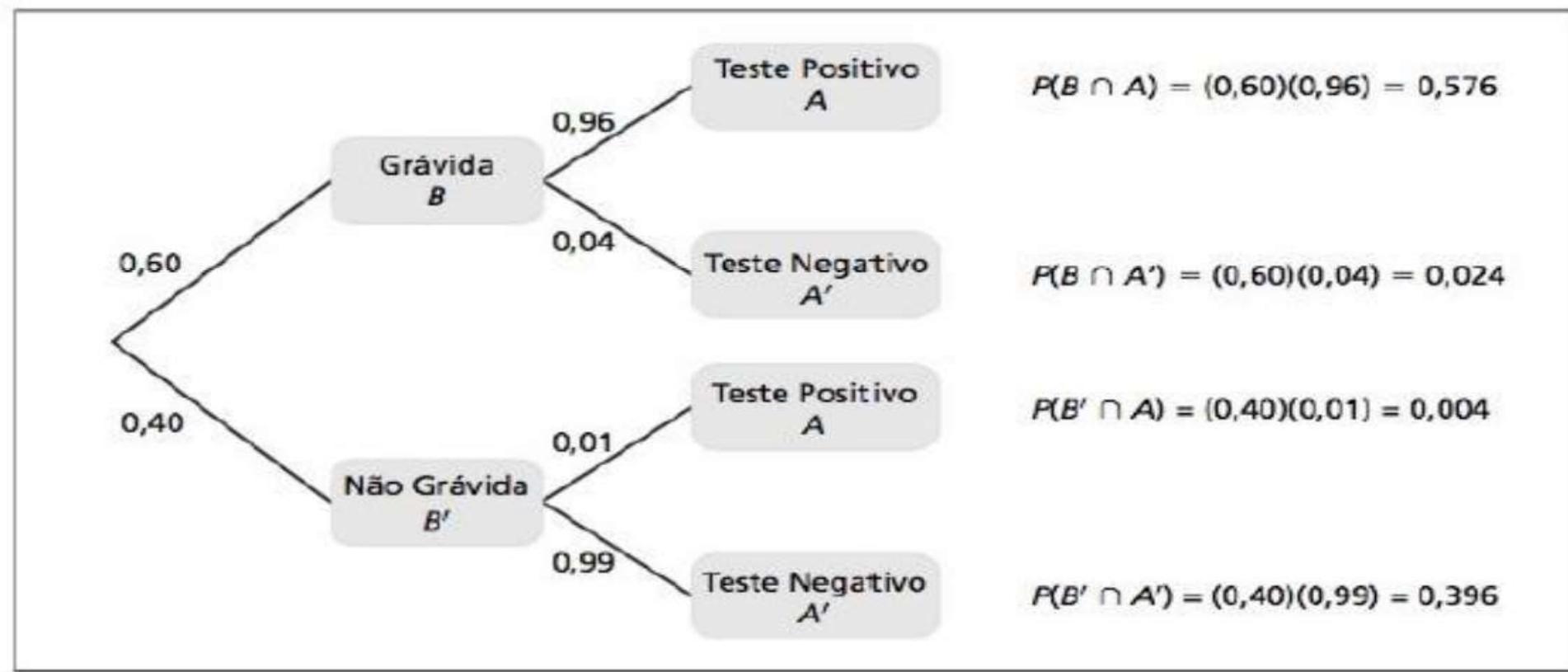


Qual a probabilidade de que a mulher esteja realmente grávida?



Data Science Academy

Qual a probabilidade de que a mulher esteja realmente grávida?





Aplicando o Teorema de Bayes...



Data Science Academy

$P(B|A)$ – probabilidade de grávida e o teste ter dado positivo

$P(A|B)$ – probabilidade de o teste ter dado positivo e estar grávida

$P(B)$ – probabilidade de estar grávida

$$P(B | A) = \frac{P(B)P(A | B)}{P(B)P(A | B) + P(B')P(A | B')}$$

$$P(A | I) = \frac{0,96 * 0,60}{0,96 * 0,60 + 0,01 * 0,40}$$



A probabilidade de que a mulher esteja grávida é
de $P(A|I) = 0.99$ ou 99%



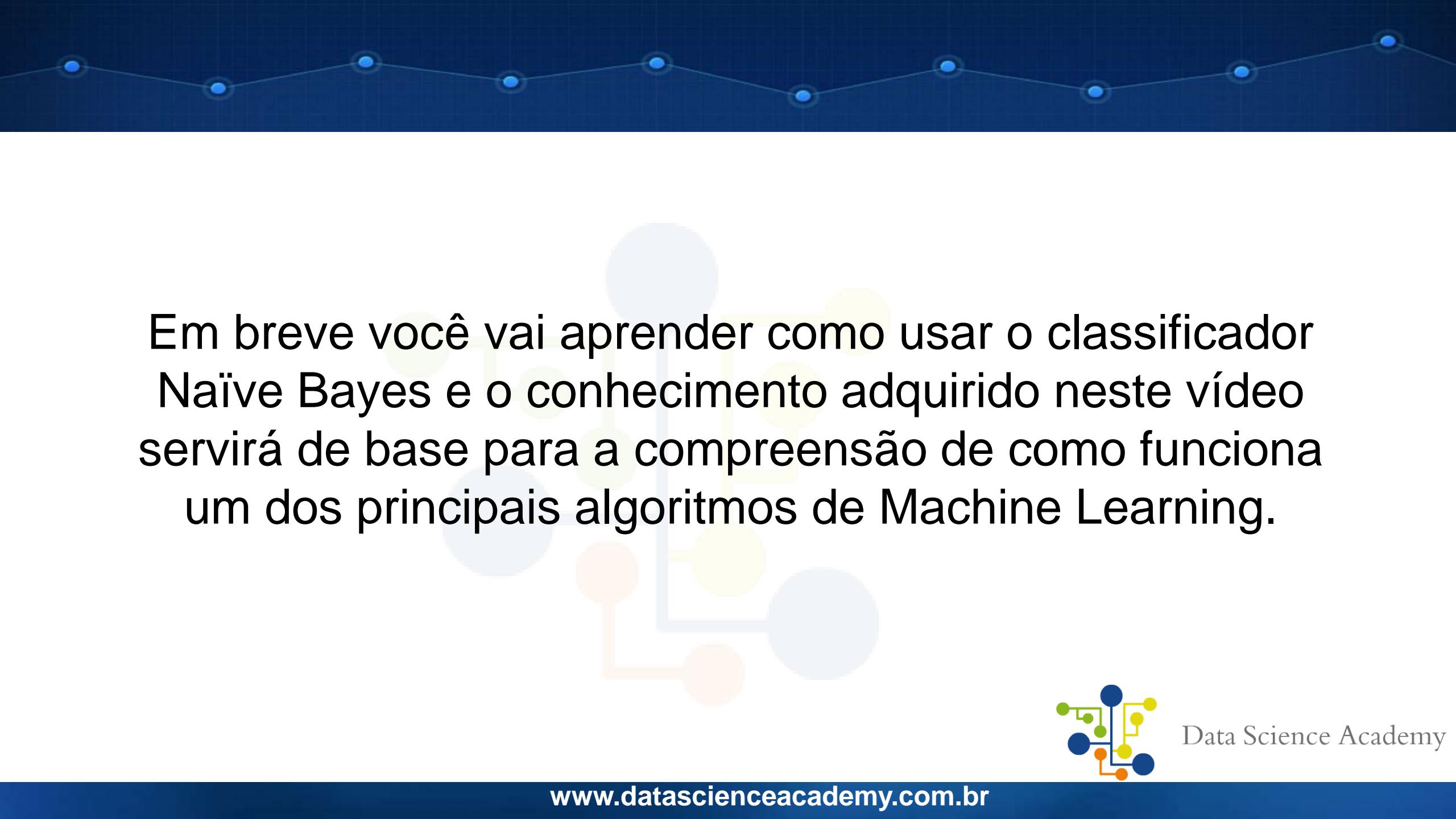
Data Science Academy



Observou como tudo em nossa vida está ligado a
Estatística?



Data Science Academy



Em breve você vai aprender como usar o classificador Naïve Bayes e o conhecimento adquirido neste vídeo servirá de base para a compreensão de como funciona um dos principais algoritmos de Machine Learning.



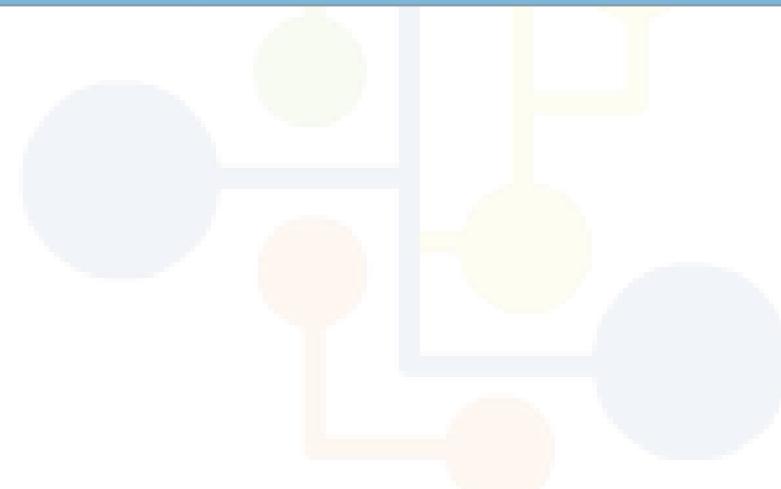
Data Science Academy

Esse tópico chegou ao final



Data Science Academy

Distribuições de Probabilidade



Data Science Academy

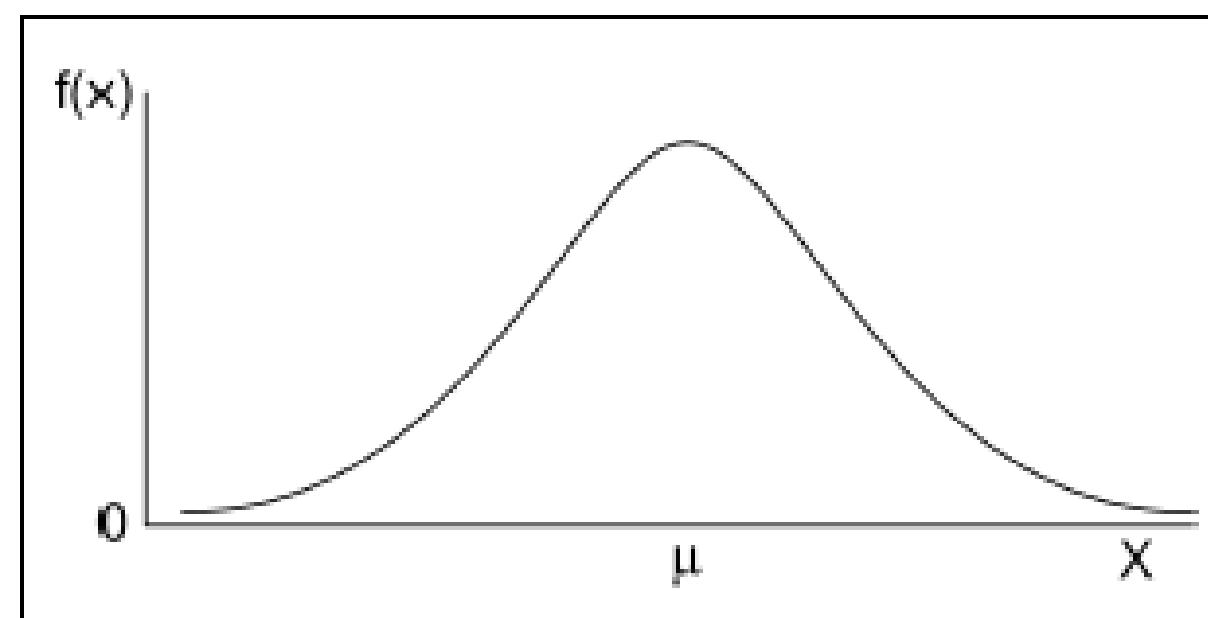
Distribuição de Probabilidade



Data Science Academy

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

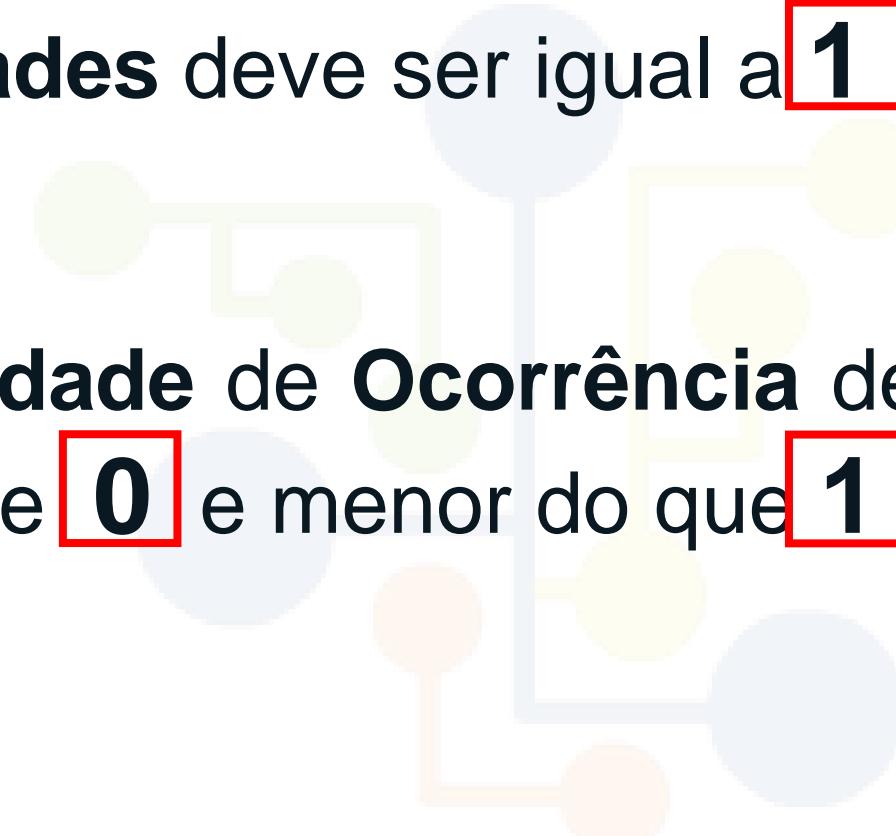
Associa uma probabilidade a cada resultado numérico de um experimento. É uma função cujo domínio são os valores da variável e a imagem são as probabilidades da variável assumir cada valor do domínio.



Data Science Academy



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

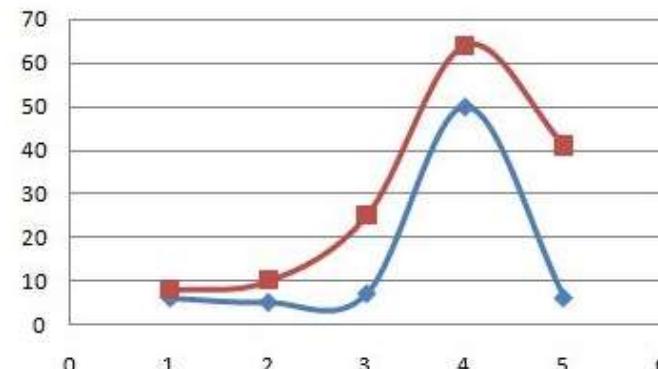


A **Probabilidade de Ocorrência** de um evento deve ser maior do que 0 e menor do que 1



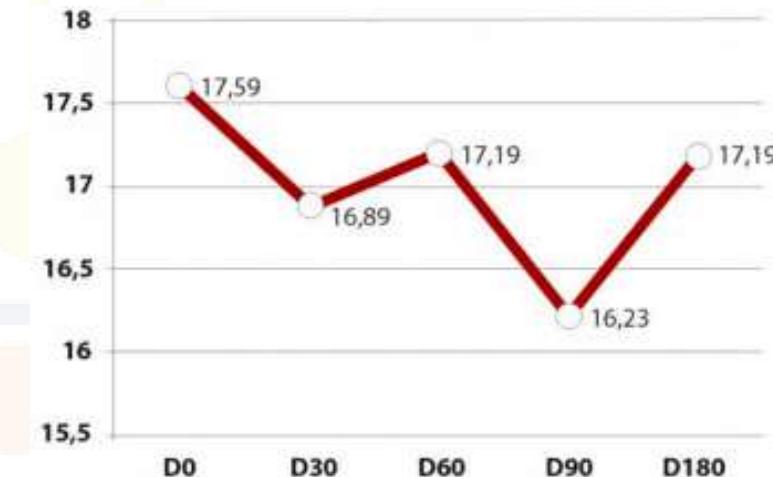
Data Science Academy

Uma distribuição de probabilidade pode ser:



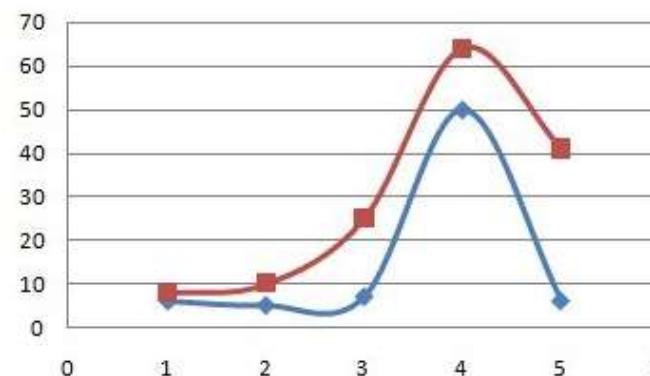
Discreta

Contínua



Data Science Academy

A distribuição de probabilidade **Discreta**:



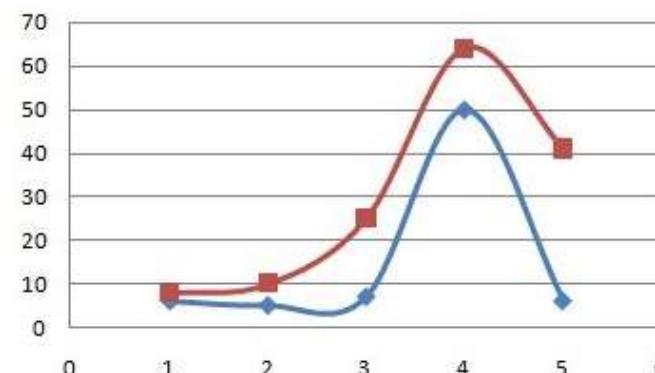
Descreve quantidades aleatórias de dados que podem assumir valores **finitos**



Data Science Academy

A distribuição de probabilidade **Discreta**:

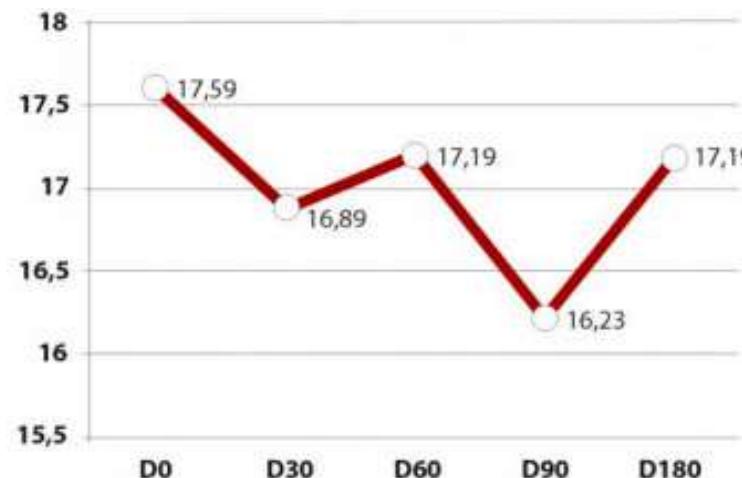
Os principais tipos de distribuição de probabilidade para variáveis discretas são:



Binomial
Poisson
Hipergeométrica
Bernoulli

Data Science Academy

A distribuição de probabilidade **Contínua**:



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

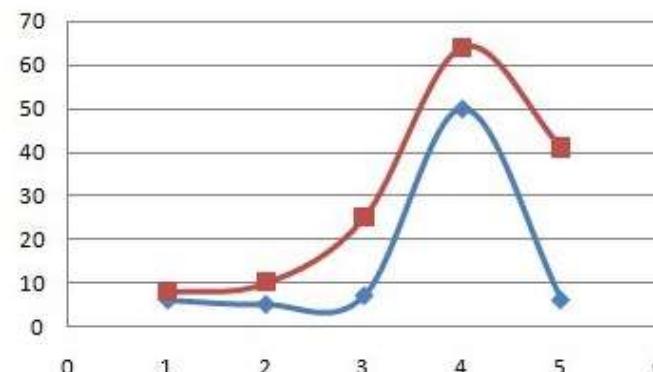


Data Science Academy

A distribuição de probabilidade **Contínua**:

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

Uniforme
Exponencial
Gama
Chi-Quadrado



Data Science Academy

Distribuição Normal

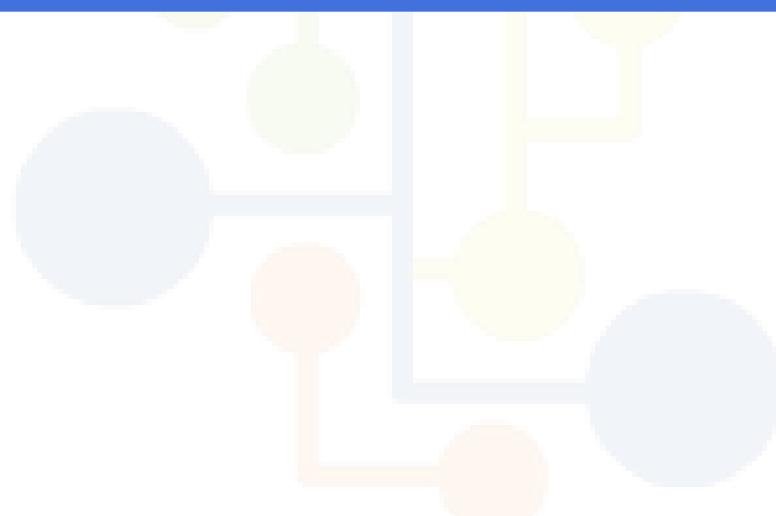
Uma variável randômica contínua que segue uma **Distribuição de Probabilidade Normal** tem uma série de características distintas



Data Science Academy



Exemplos



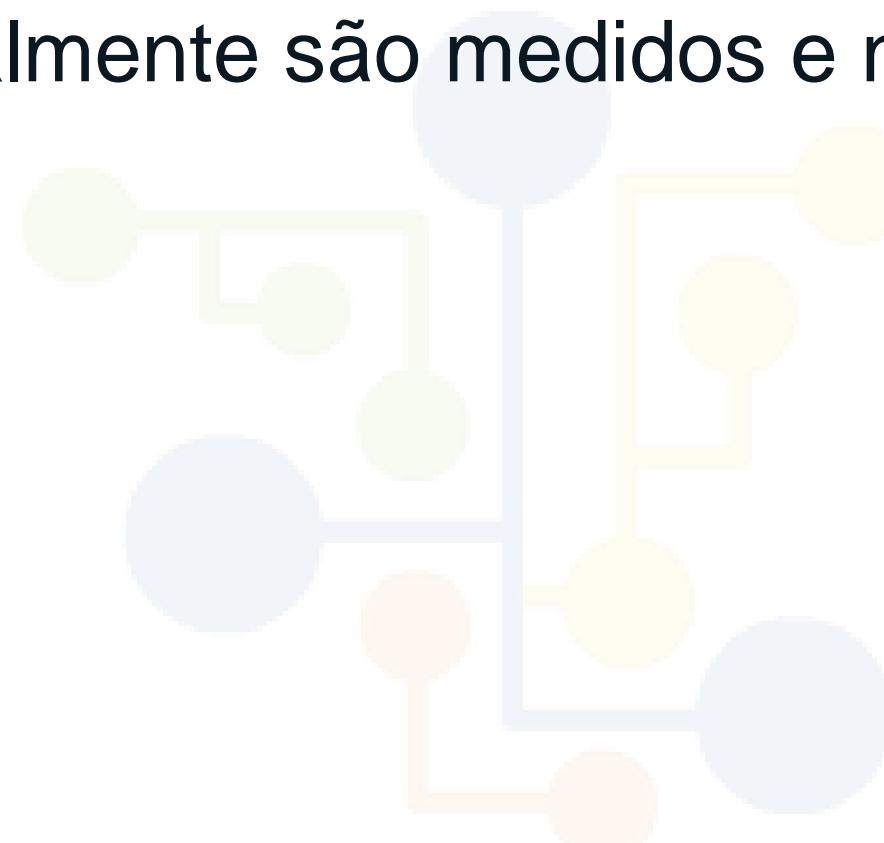
Data Science Academy

Se você decidir **contar o número de pessoas** que visitam uma loja em um dia da semana, você vai estar trabalhando com **dados discretos**.



Data Science Academy

Já os **dados contínuos**, podem ter valores **fracionários** e que normalmente são medidos e não contados.



Data Science Academy

Por exemplo, o peso e altura de uma pessoa.



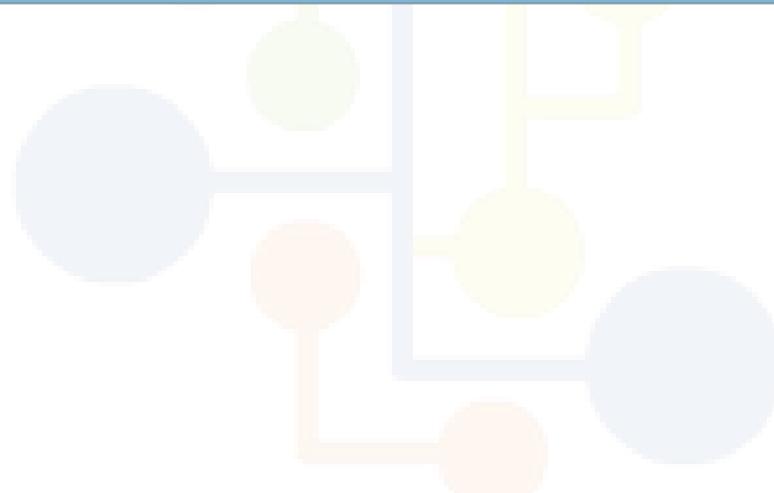
Peso (kg)	Altura (m)
45-52	1,50
44-46	1,52
47-48	1,55
48-51	1,57
49-53	1,60
51-56	1,63
52-29	1,65
53-62	1,68
55-64	1,70
56-67	1,73
58-69	1,75
60-76	1,78
61-77	1,80
63-80	1,83
65-82	1,85



Data Science Academy



Variáveis Discretas



Data Science Academy

Registrar o número de clientes que fazem contato telefônico com a central de suporte de um banco.



Data Science Academy

Randomicamente selecionar 6 clientes que entram em uma loja de celulares e contar quantos assinam um plano pós-pago.



Data Science Academy

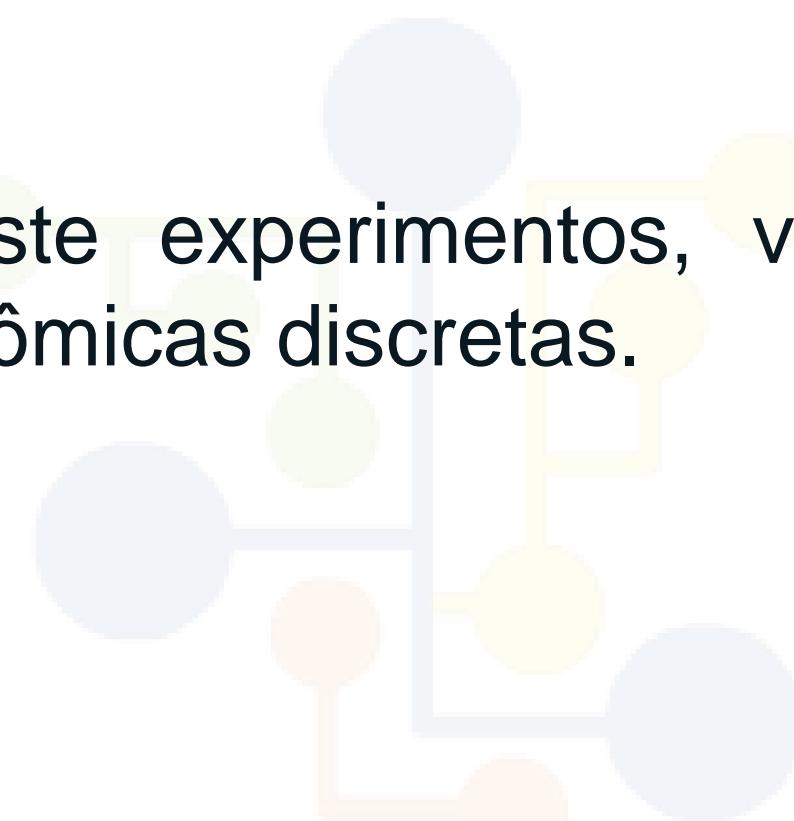
Solicitar que cada cliente que deixa um hotel, avalie seu grau de satisfação com o serviço prestado.



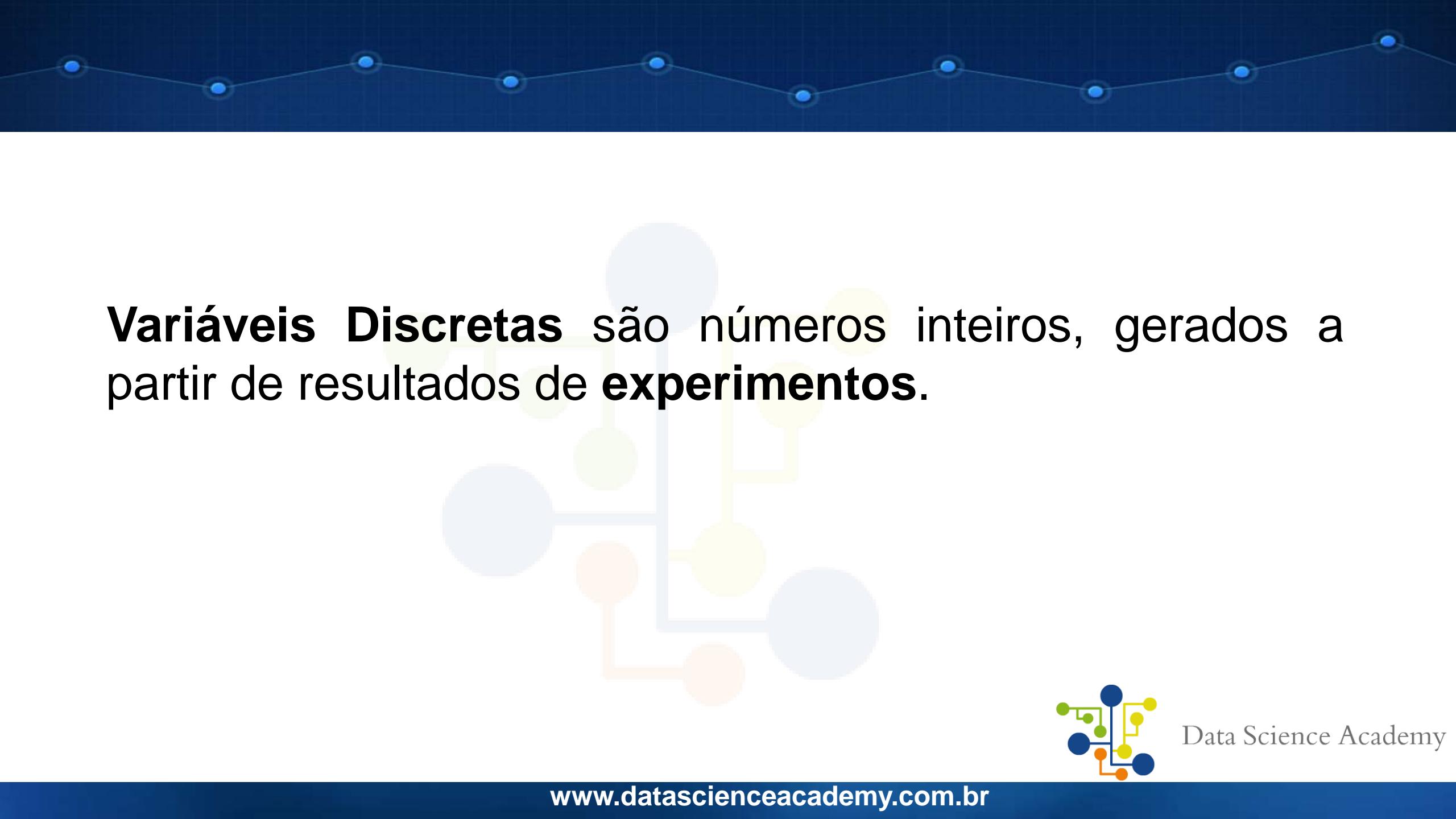
Data Science Academy



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



Data Science Academy



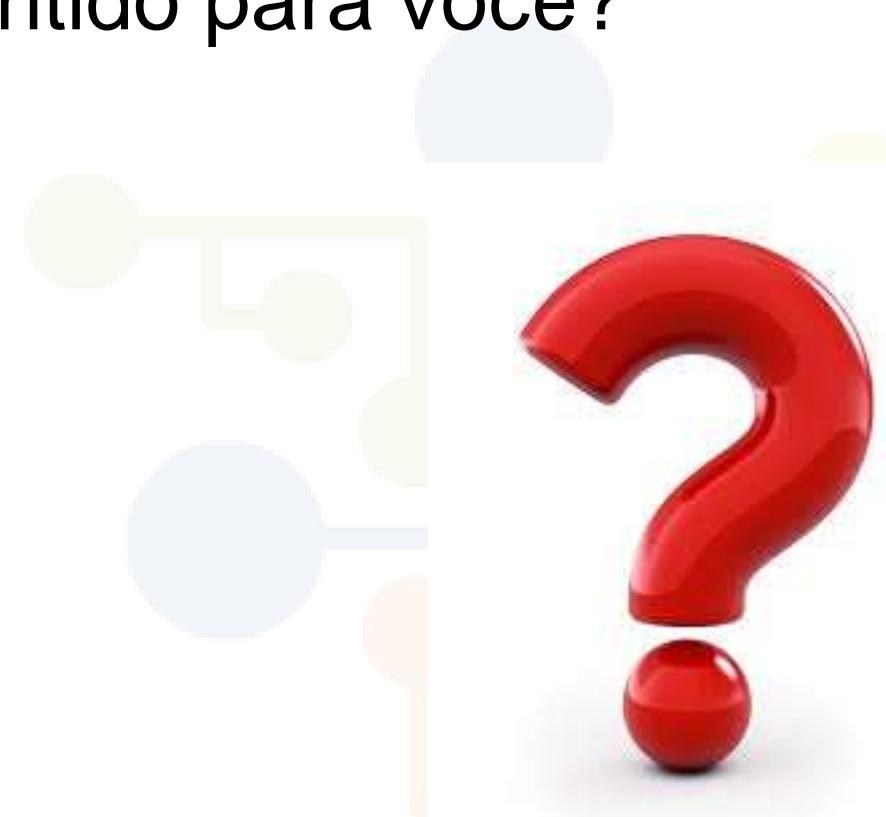
Variáveis Discretas são números inteiros, gerados a partir de resultados de **experimentos**.



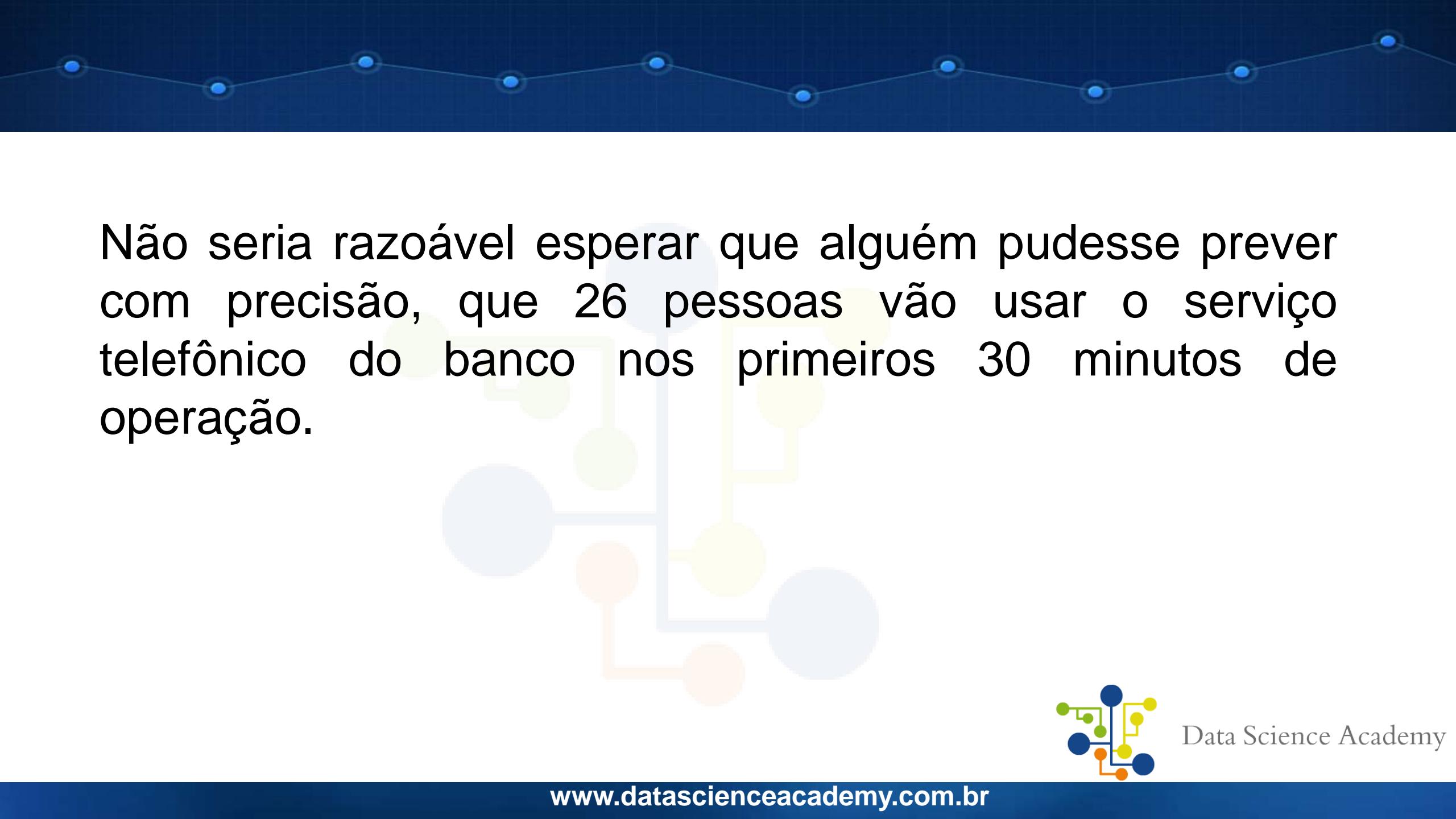
Data Science Academy



E isso faz sentido para você?



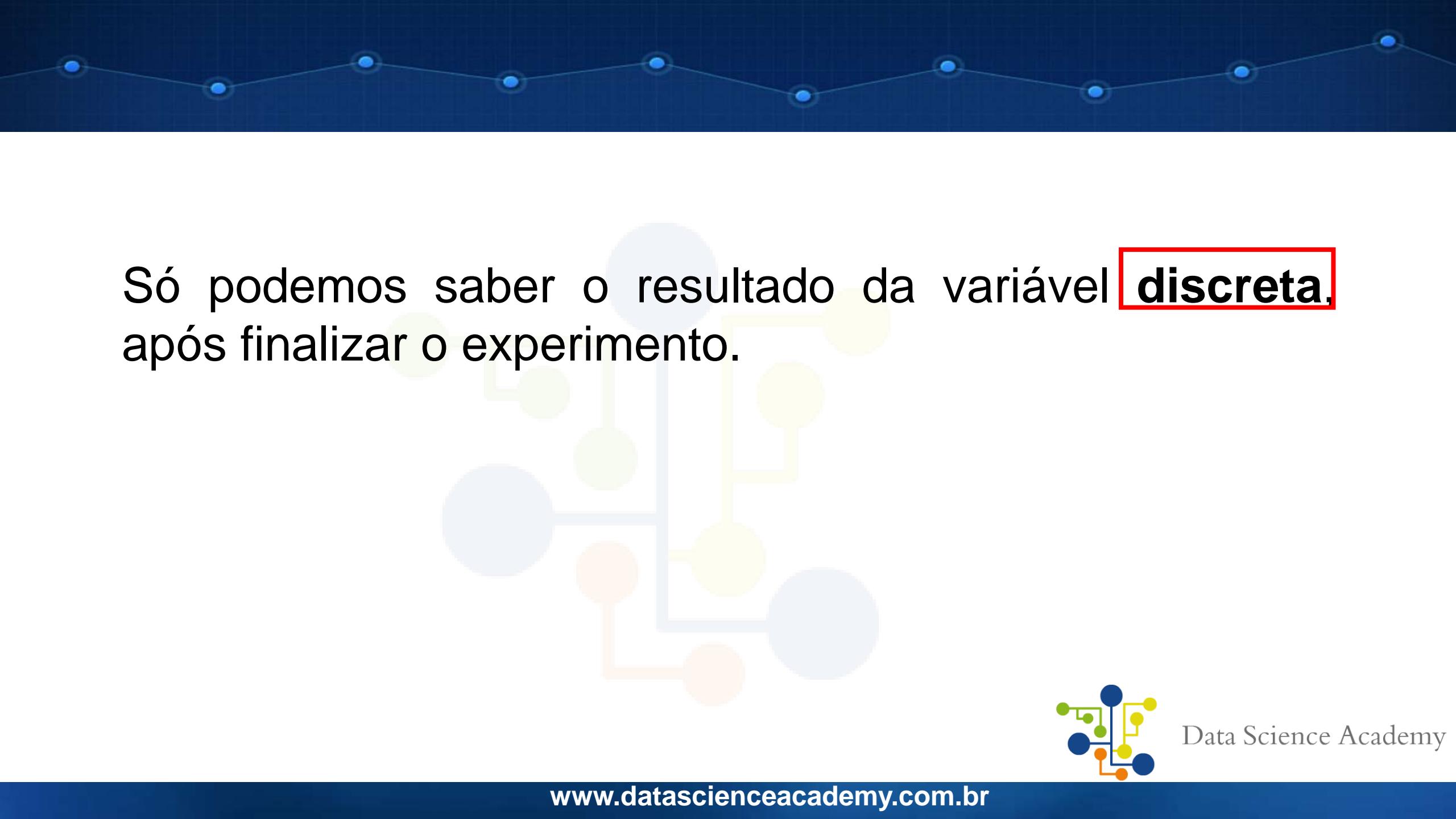
Data Science Academy



Não seria razoável esperar que alguém pudesse prever com precisão, que 26 pessoas vão usar o serviço telefônico do banco nos primeiros 30 minutos de operação.



Data Science Academy



Só podemos saber o resultado da variável **discreta**,
após finalizar o experimento.



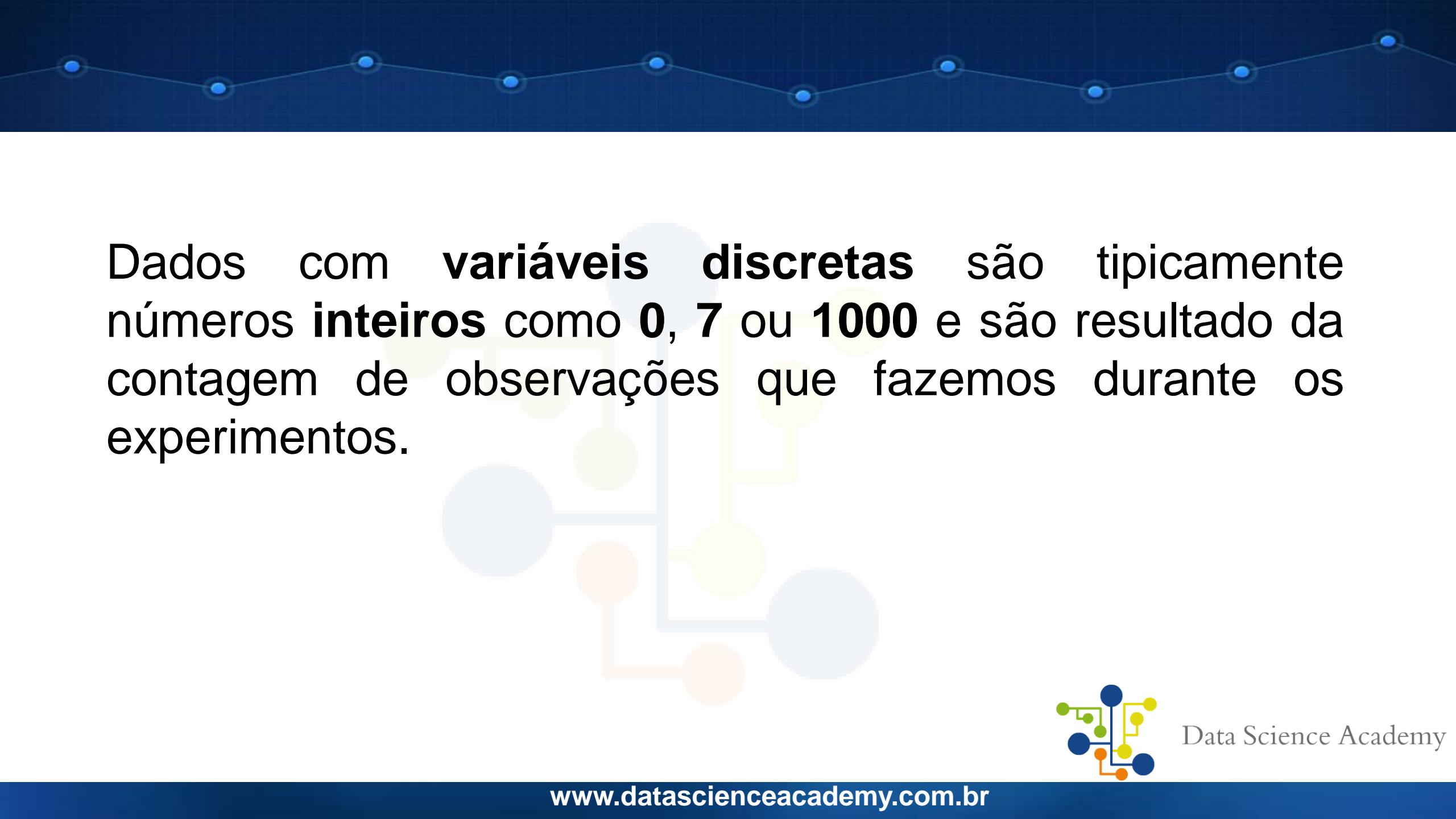
Data Science Academy



Variáveis Contínuas



Data Science Academy



Dados com **variáveis discretas** são tipicamente números **inteiros** como 0, 7 ou 1000 e são resultado da contagem de observações que fazemos durante os experimentos.



Data Science Academy



Variáveis Contínuas



Data Science Academy

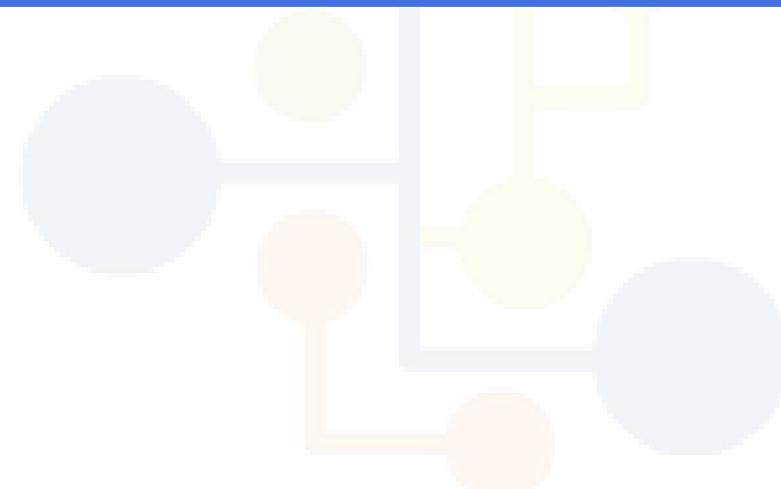
Mas no **mundo dos negócios**, normalmente nos deparamos com conjuntos de dados imensos, que precisam ser medidos.



ence Academy



Mais Exemplos



Data Science Academy

Tempo de duração de voo entre Natal e Maceió.

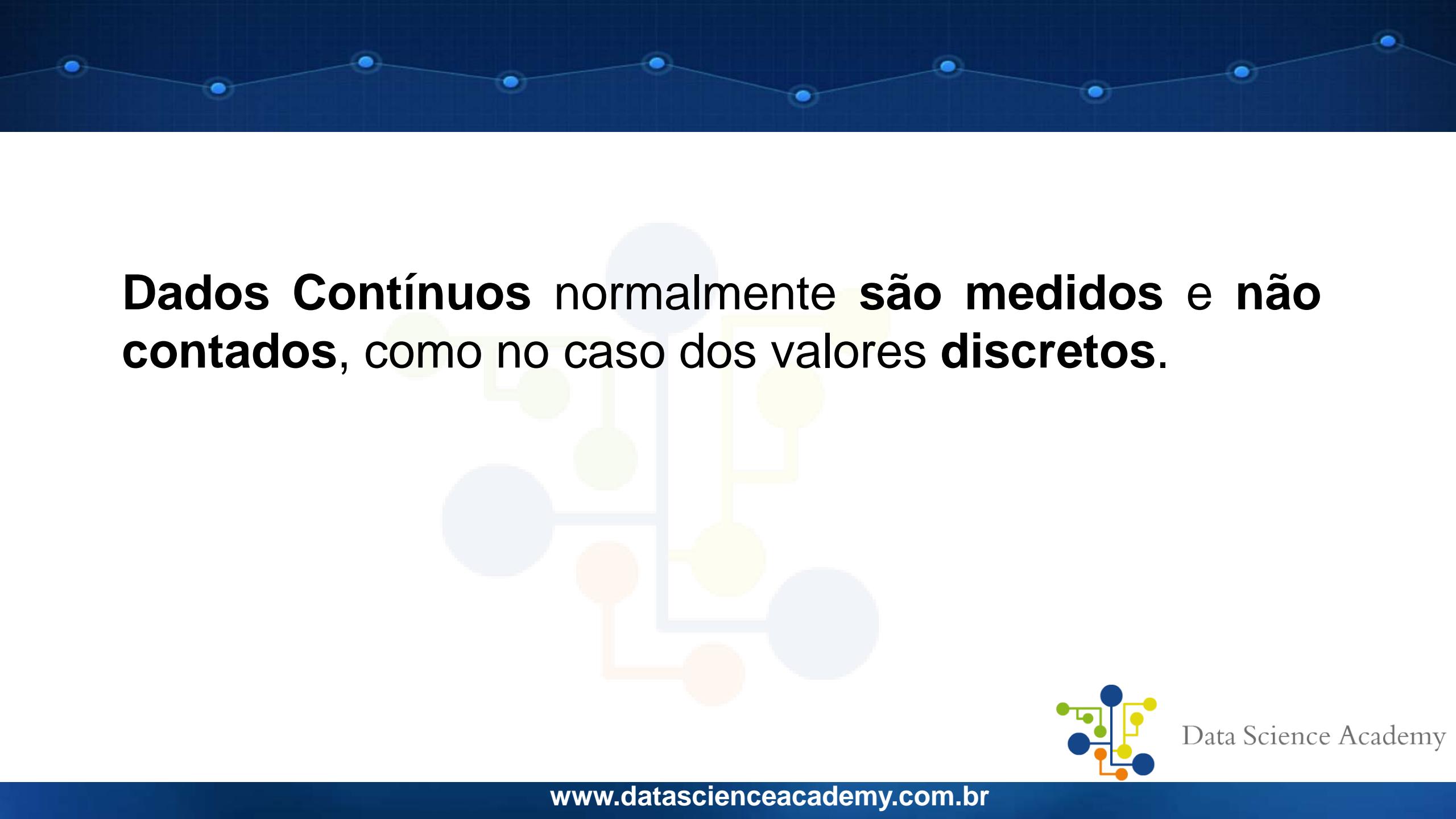


Peso das caixas de biscoito em uma fábrica de alimentos



Tempo gasto por um cliente ao telefone, com uma companhia de TV a cabo.





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



Data Science Academy

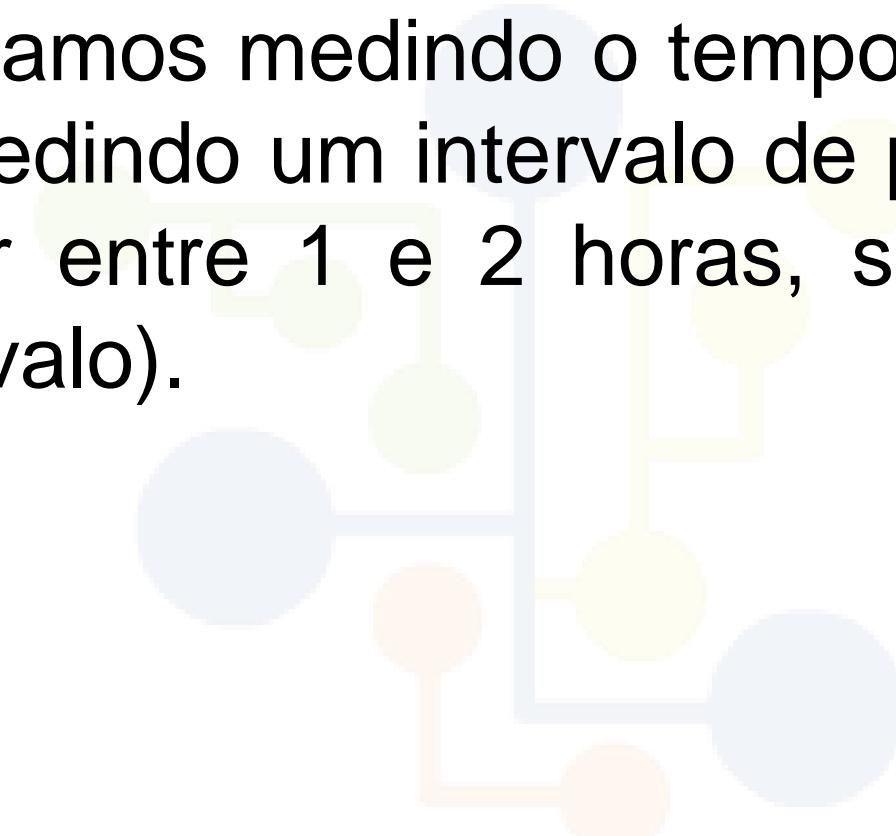
Tenha em mente, que se optarmos por contar os voos que chegam atrasados ao seu destino, podemos usar **Distribuição Discreta**.



Data Science Academy



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



Data Science Academy

Variáveis Contínuas

Extensões
Massa e
Tempo decorrido



Data Science Academy

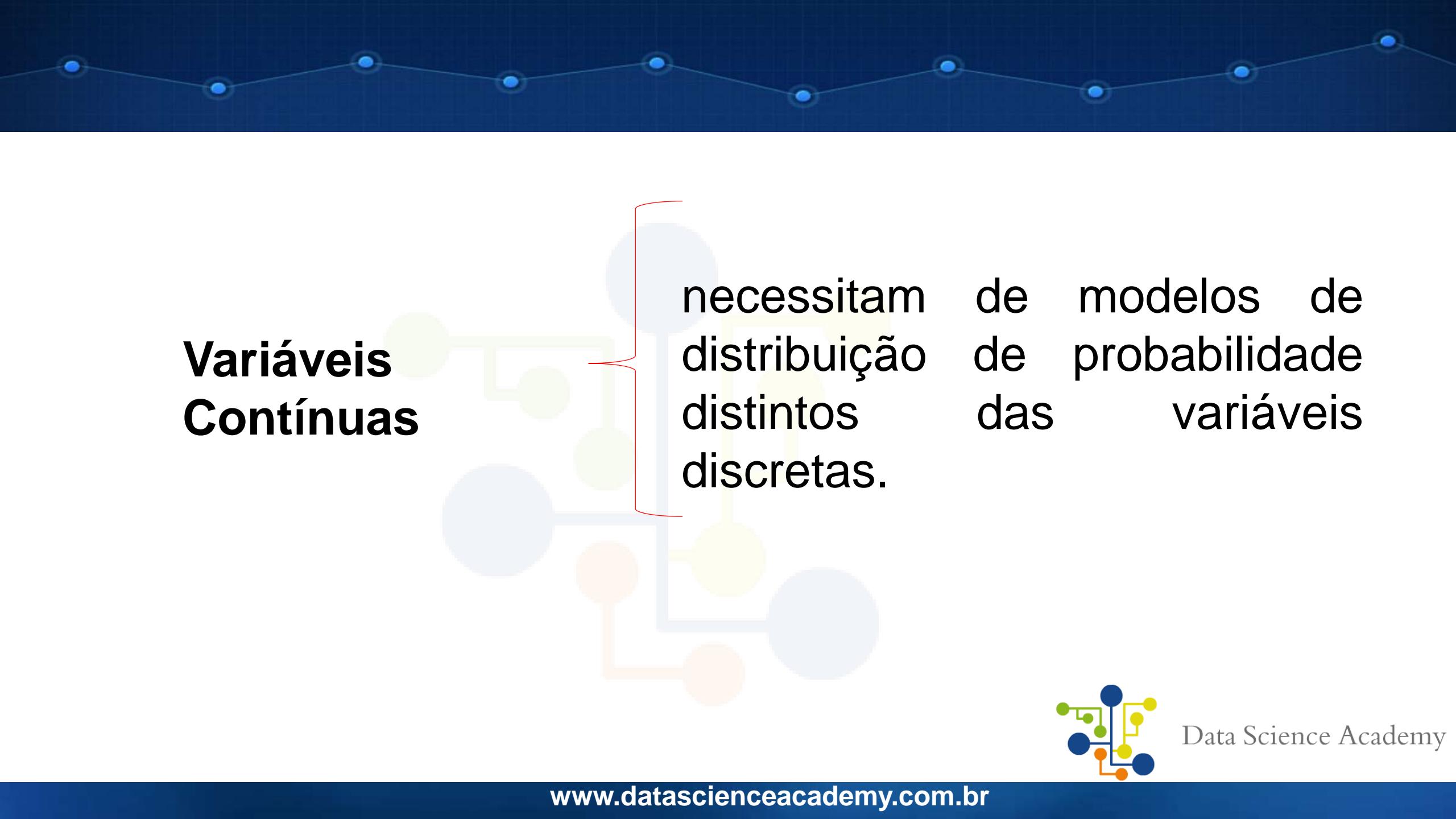
Variáveis Contínuas

Pode-se ter qualquer valor no conjunto de números reais, ou um subconjunto deles.



Data Science Academy

Variáveis Contínuas



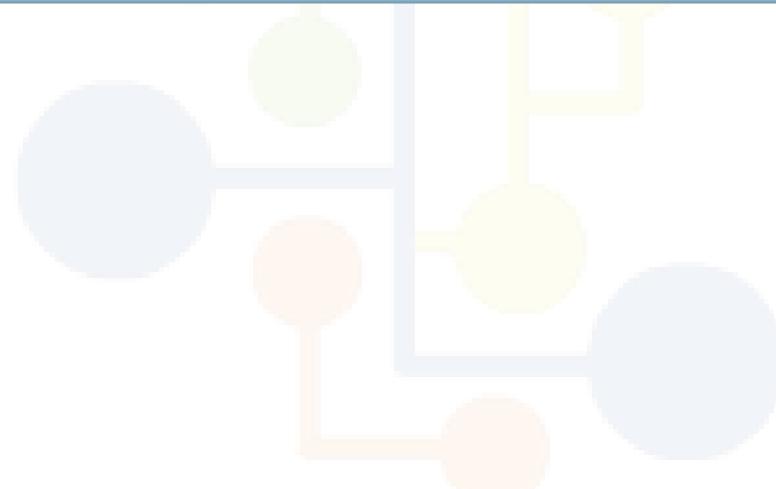
necessitam de modelos de distribuição de probabilidade das variáveis discretas.



Data Science Academy



Média de uma Distribuição de Probabilidade Discreta



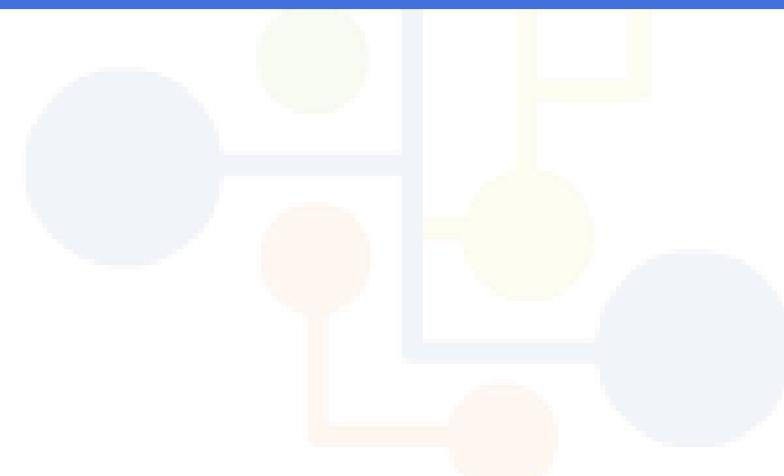
Data Science Academy

A média de uma **Distribuição de Probabilidade Discreta** é simplesmente a média ponderada dos resultados obtidos.



Data Science Academy

Exemplo



Data Science Academy

Vamos calcular a média de frequência com que determinado grupo de pessoas entram em um restaurante.



Data Science Academy

Distribuição de Probabilidade

Restaurante (x)	Frequência	Frequência relativa	Probabilidade $P(x)$	Probabilidade ponderada $P(x)$ x
Grupo com 2 pessoas	17	$17/50 = 0.34$ → 0.34	0.34	0.68
Grupo com 3 pessoas	6	$6/50 = 0.12$ → 0.12	0.12	0.36
Grupo com 4 pessoas	16	$16/50 = 0.32$ → 0.32	0.32	1.28
Grupo com 5 pessoas	4	$4/50 = 0.08$ → 0.08	0.08	0.40
Grupo com 6 pessoas	7	$7/50 = 0.14$ → 0.14	0.14	0.84
Total	50	1.00	1.0	3.56

Somatório das
frequências

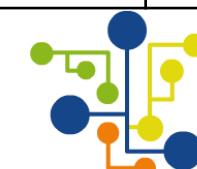
Média da distribuição de
probabilidade discreta



Data Science Academy

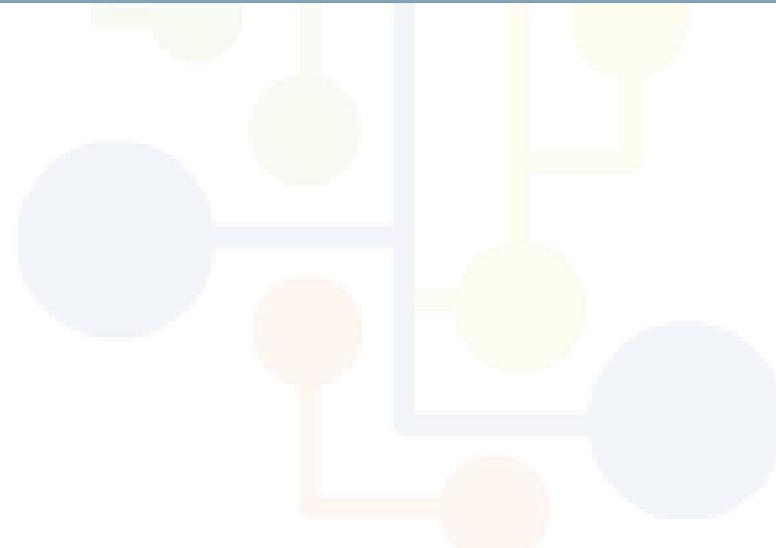
Distribuição de Probabilidade

Restaurante (x)	Frequência	Frequência relativa	Probabilidade $P(x)$	Probabilidade ponderada $P(x)$ x
Grupo com 2 pessoas	17	$17/50 = 0.34$	0.34	0.68
Grupo com 3 pessoas	6	$6/50 = 0.12$	0.12	0.36
Grupo com 4 pessoas	16	$16/50 = 0.32$	0.32	1.28
Grupo com 5 pessoas	4	$4/50 = 0.08$	0.08	0.40
Grupo com 6 pessoas	7	$7/50 = 0.14$	0.14	0.84
Total	50	1.00	1.0	3.56



Data Science Academy

Regras para Distribuição de Probabilidade de Variável Discreta



Data Science Academy



O Analista de Dados utilizará a Distribuição de Probabilidade para:

Verificar a quantidade de reclamações feitas por clientes num ambiente empresarial.



Data Science Academy



O Analista de Dados utilizará a Distribuição de Probabilidade para:

Quantificar o número de televisores ou equipamentos eletrônicos adquiridos por pessoas no Brasil e no mundo.



Data Science Academy



O Analista de Dados utilizará a Distribuição de Probabilidade para:

Quantificar o número de clientes uma empresa poderá receber num determinado período.



Data Science Academy



E inúmeras outras informações, que irão contribuir para que as empresas possam ter um maior controle sobre suas vendas e investir em processos de melhoria e qualidade internas.



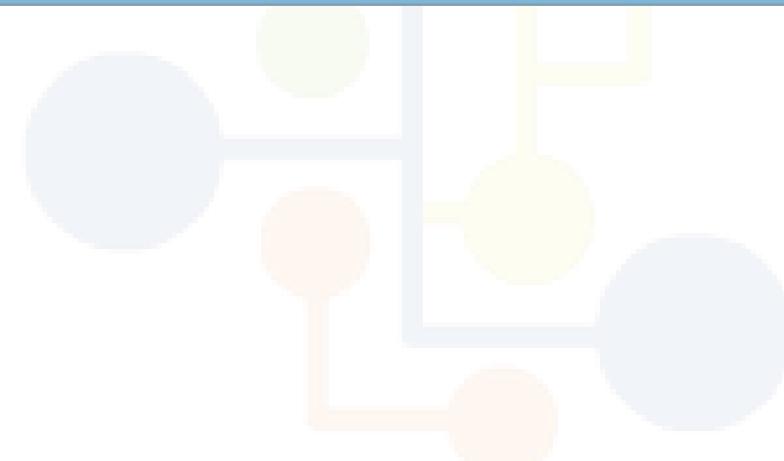
Data Science Academy

Esse tópico chegou ao final



Data Science Academy

Distribuição Binomial



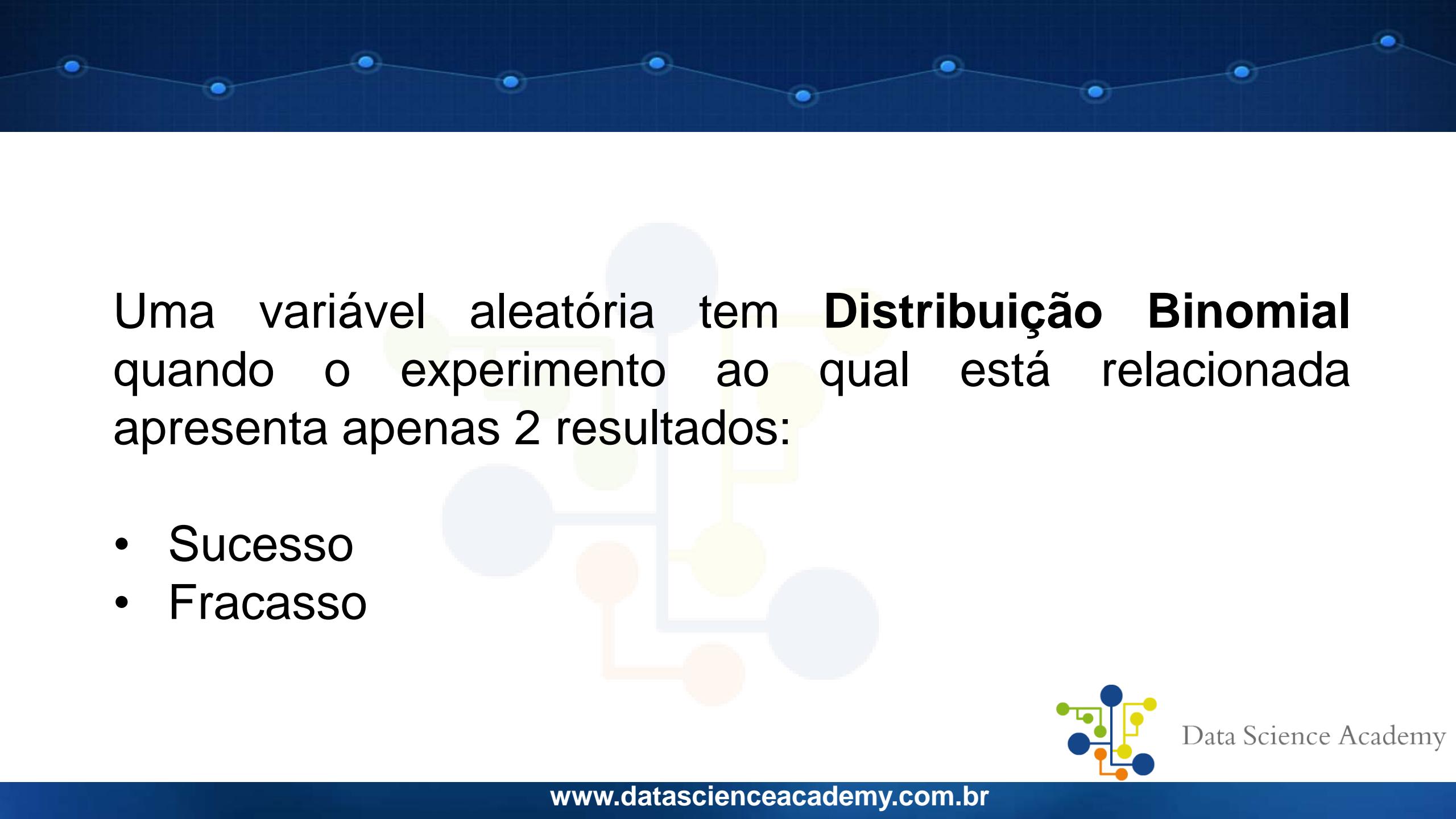
Data Science Academy

Distribuição Binomial

É um tipo de distribuição de probabilidade discreta



Data Science Academy

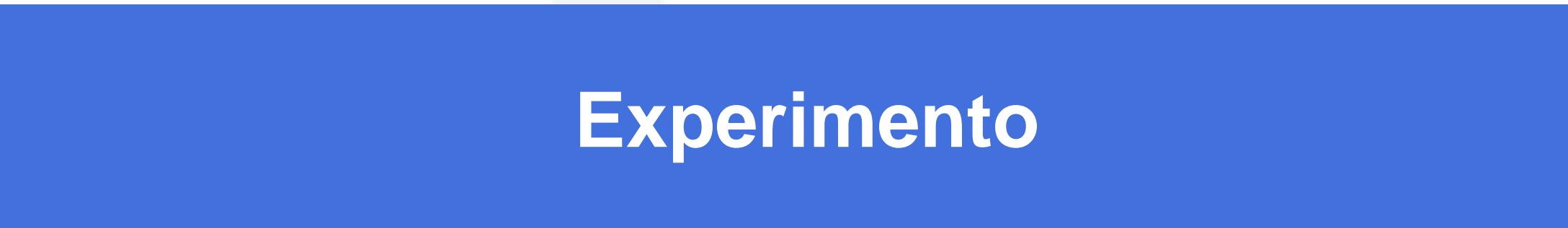


Uma variável aleatória tem **Distribuição Binomial** quando o experimento ao qual está relacionada apresenta apenas 2 resultados:

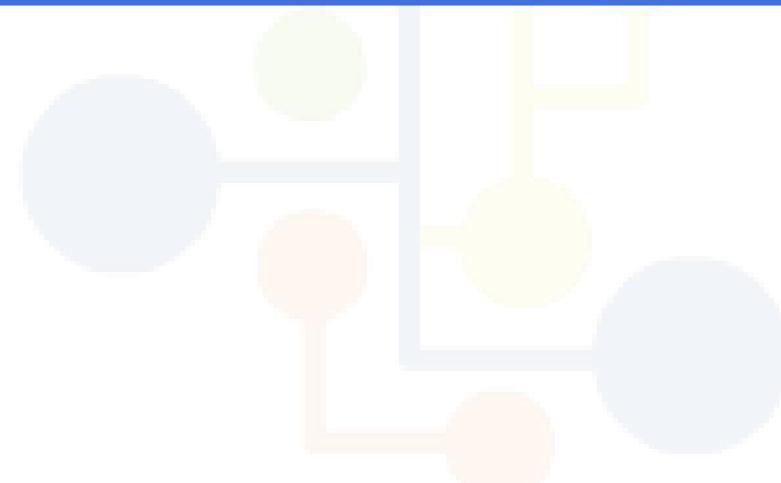
- Sucesso
- Fracasso



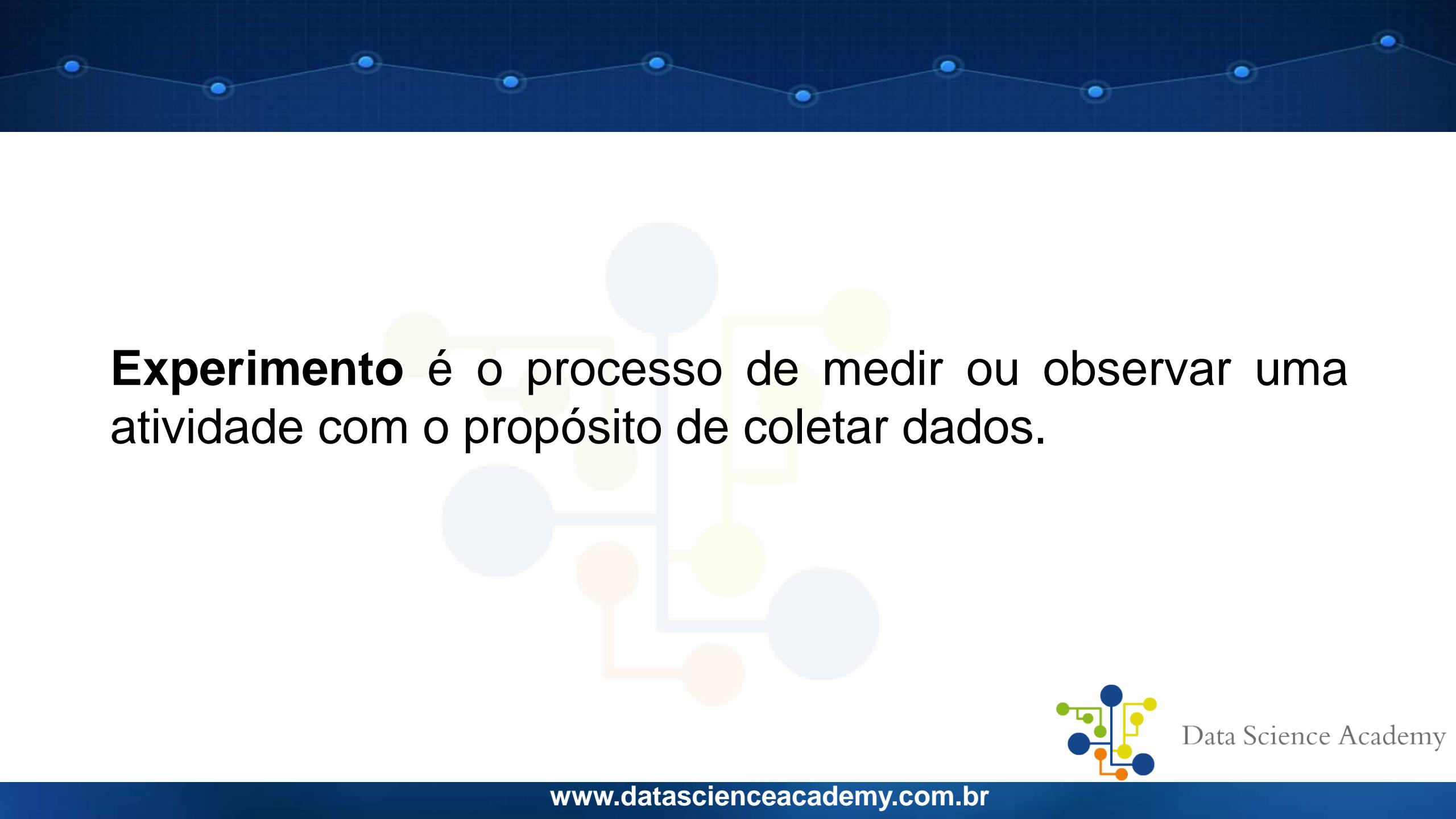
Data Science Academy



Experimento



Data Science Academy

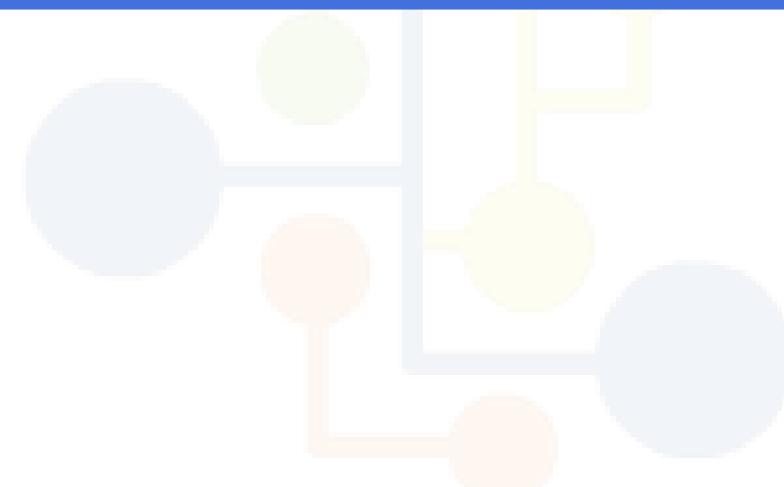


Experimento é o processo de medir ou observar uma atividade com o propósito de coletar dados.



Data Science Academy

Exemplo



Data Science Academy

Vamos imaginar que nosso experimento seja contar quantos clientes que entram em uma loja de celulares, adquirem um plano pós-pago.



Data Science Academy

Para este experimento, temos 2 possibilidades para cada observação: **adquirir** ou **não adquirir** o plano.



Ligações
Ilimitadas
Grátis



Internet
Grátis



R\$ 0,21
por chamada
ilimitada



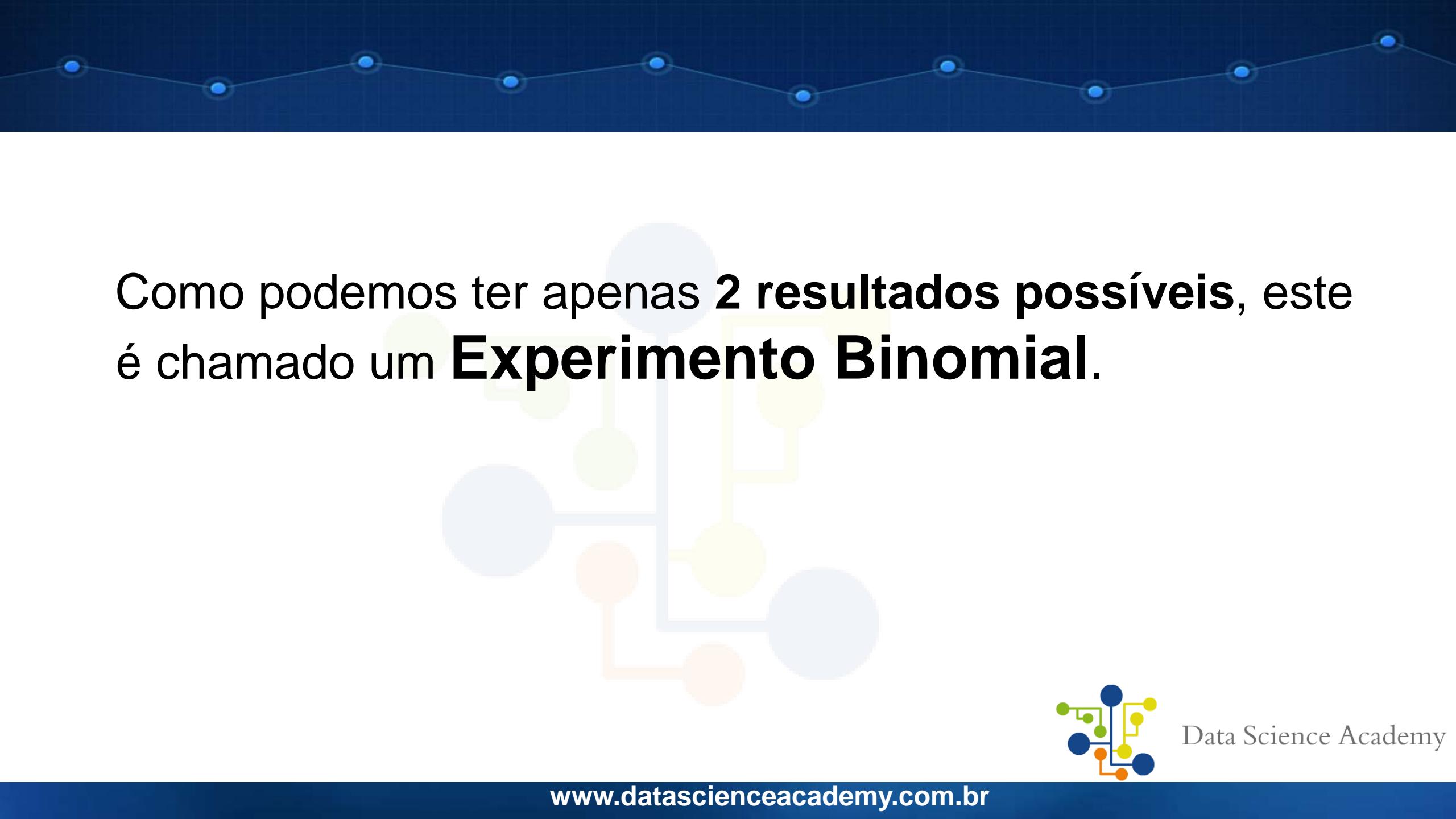
R\$ 0,50
por chamada
ilimitada



R\$ 0,50
por dia



Data Science Academy

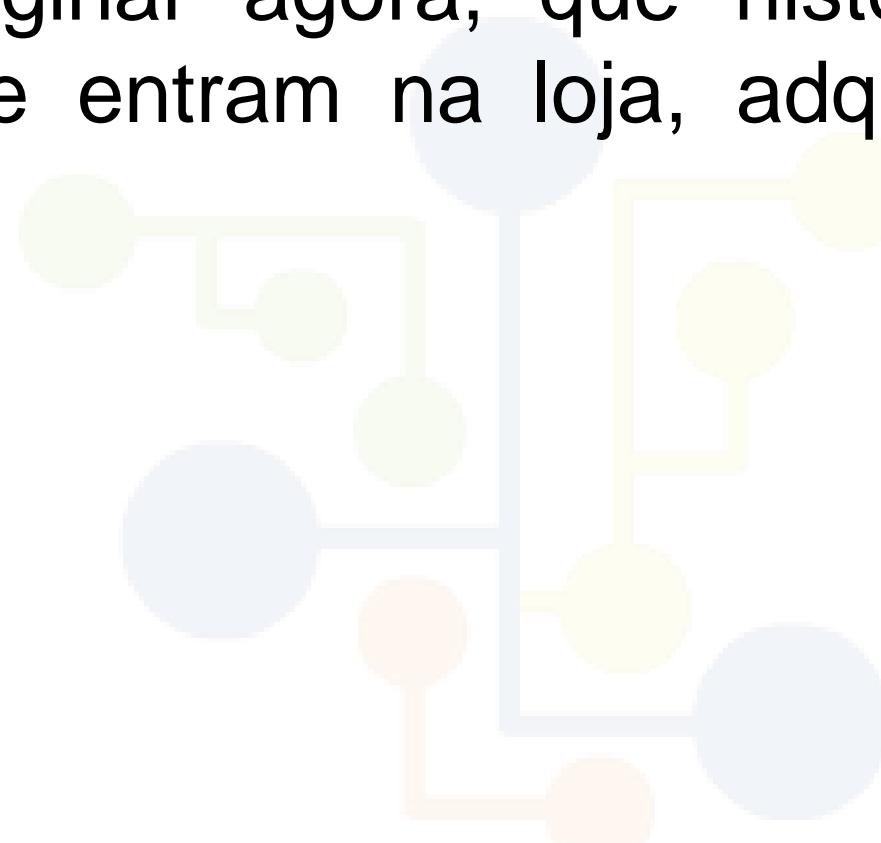


Como podemos ter apenas **2 resultados possíveis**, este é chamado um **Experimento Binomial**.



Data Science Academy

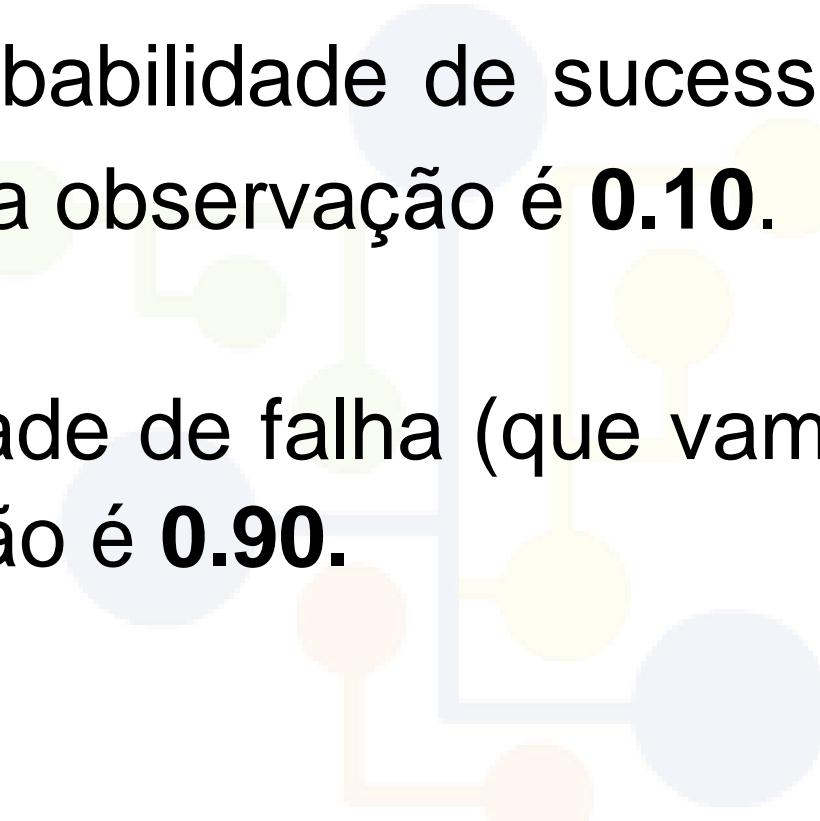
Vamos imaginar agora, que historicamente, **10%** dos clientes que entram na loja, adquirem um plano pós-pago.



Data Science Academy



Portanto, a probabilidade de sucesso (que vamos chamar de **p**) para cada observação é **0.10**.



E, a probabilidade de falha (que vamos chamar de **q**) para cada observação é **0.90**.



Data Science Academy

Ou seja:

$$p = 1 - q$$

Onde:

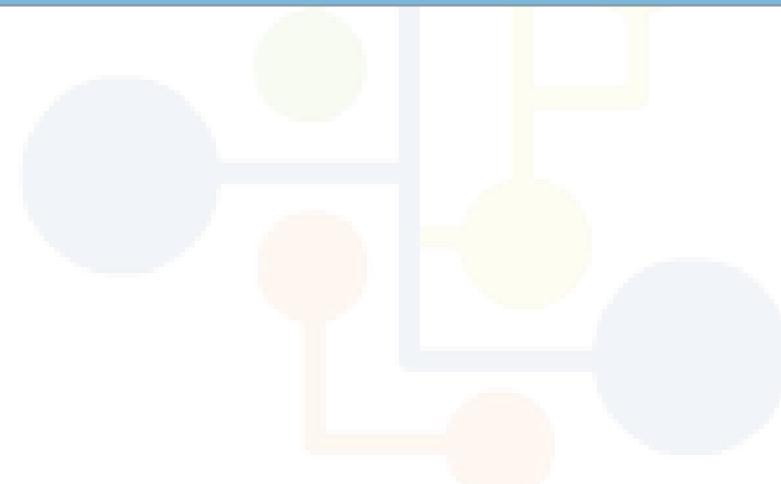
p = probabilidade de sucesso
q = probabilidade de fracasso



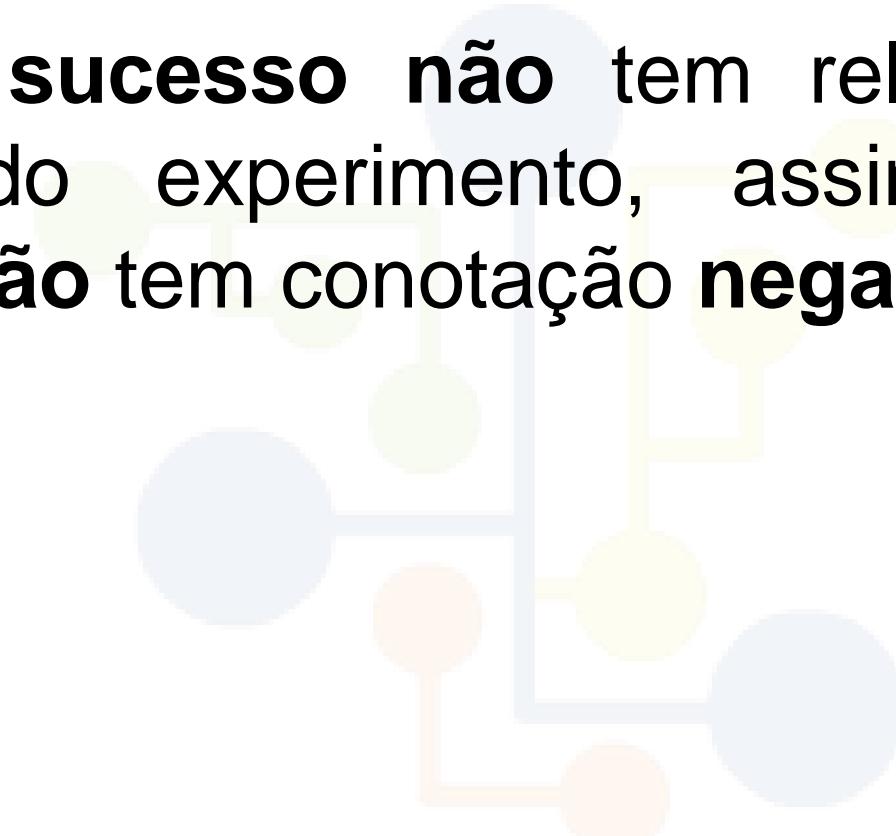
Data Science Academy



Dicas Importantes



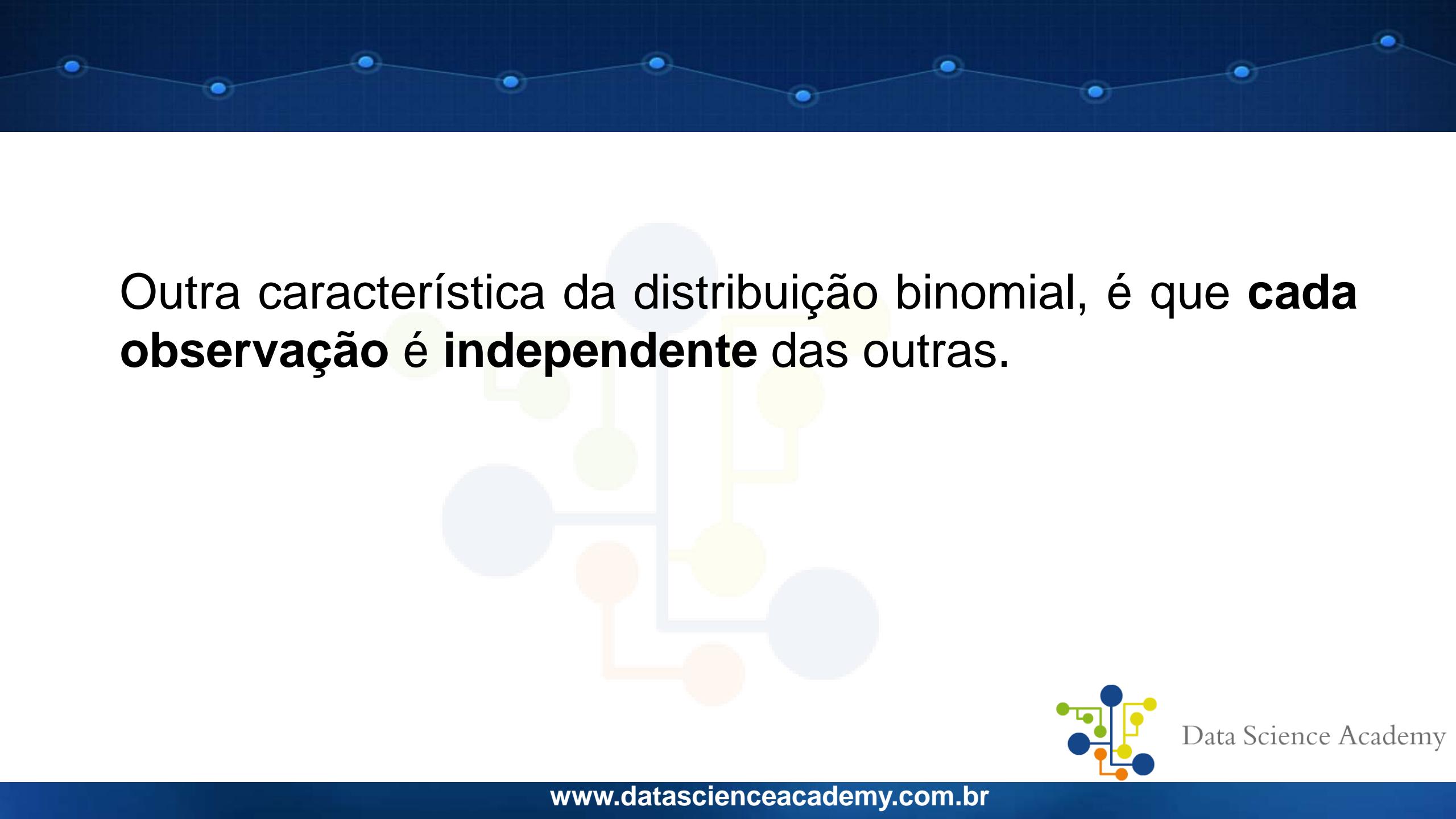
Data Science Academy



A palavra **sucesso** não tem relação com **resultado positivo** do experimento, assim como a palavra **fracasso** não tem conotação **negativa**.



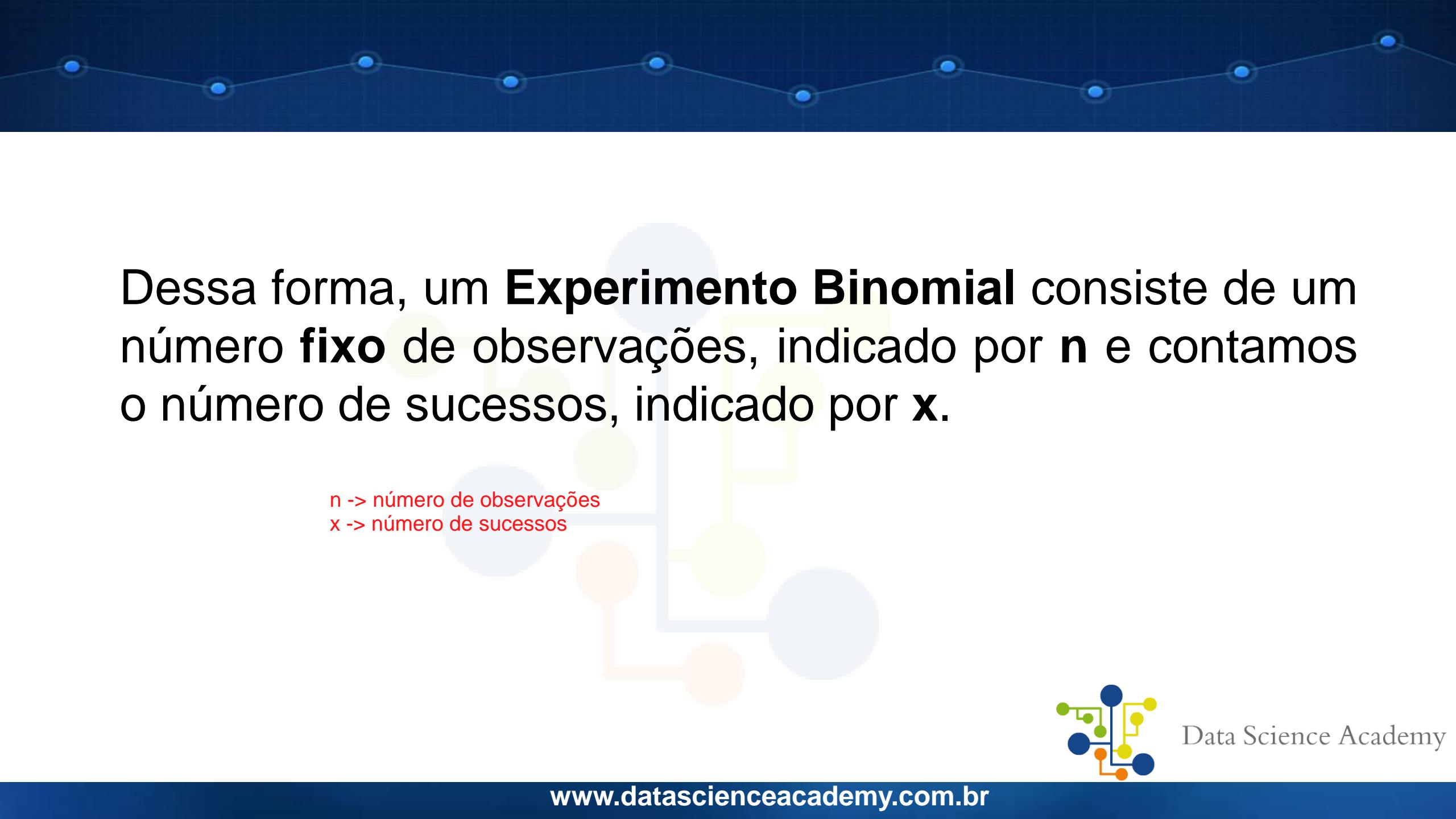
Data Science Academy



Outra característica da distribuição binomial, é que **cada observação é independente das outras.**



Data Science Academy



Dessa forma, um **Experimento Binomial** consiste de um número **fixo** de observações, indicado por n e contamos o número de sucessos, indicado por x .

n -> número de observações
 x -> número de sucessos



Data Science Academy

Exemplo



Data Science Academy

Selecionando randomicamente 6 ($n = 6$) clientes, observamos que 1 assina o contrato do plano pós-pago ($x = 1$).



Data Science Academy

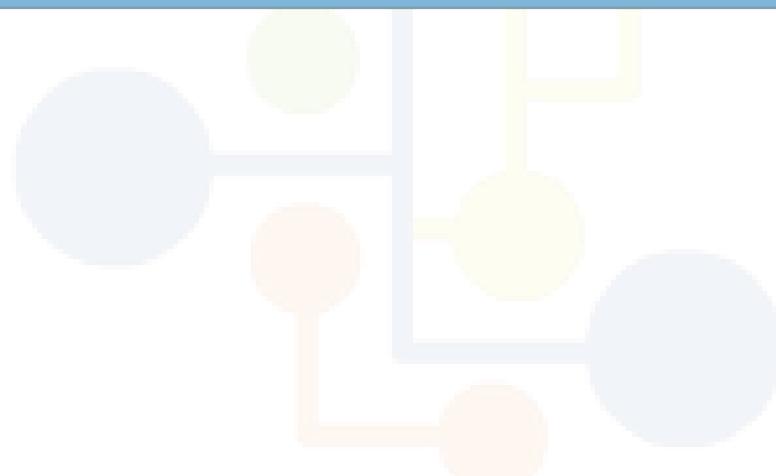
Dada esta informação, podemos calcular a probabilidade de que **1** cliente, a cada **6**, adquira um plano pós-pago.



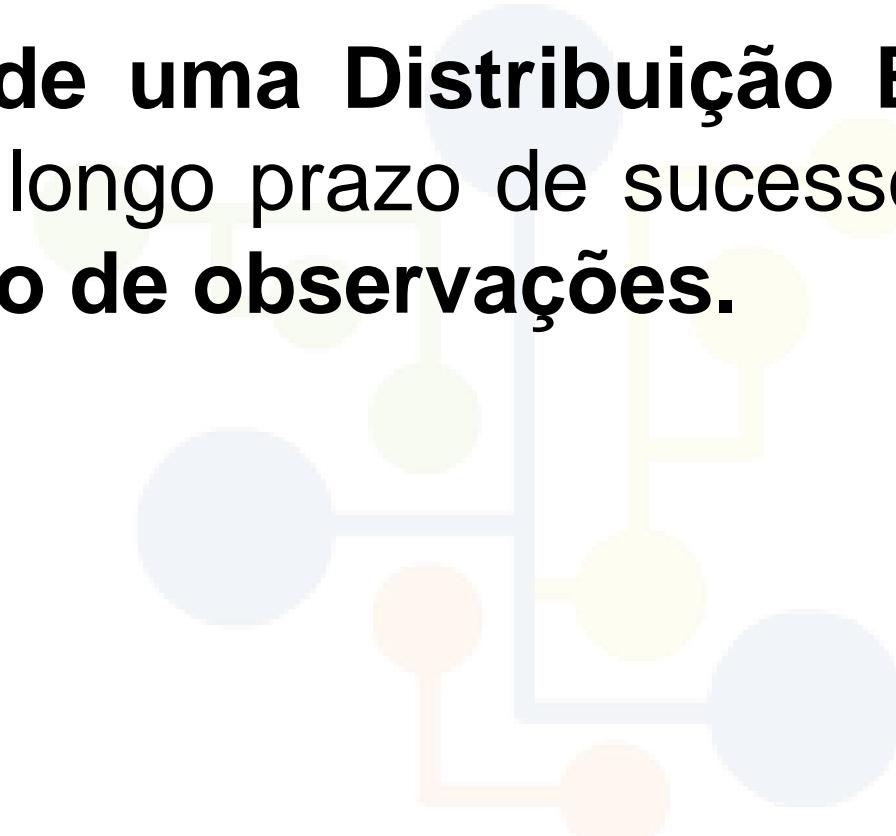
Data Science Academy



Média e Desvio Padrão de uma Distribuição Binomial



Data Science Academy



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



Data Science Academy



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

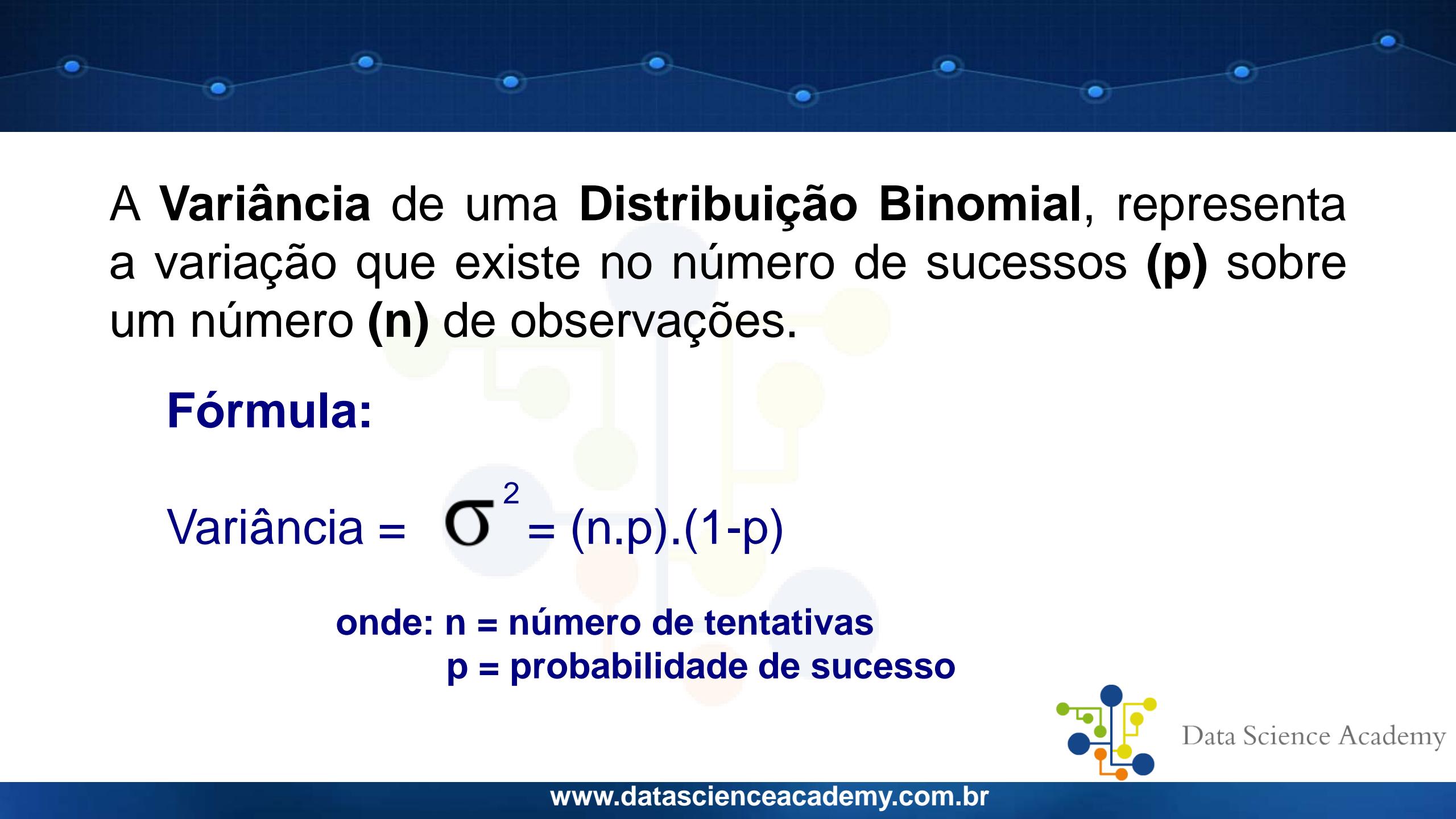
Fórmula:

$$\text{Média} = \mu = n \cdot P$$

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



Data Science Academy



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



Data Science Academy

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

Fórmula:

$$\sigma = \sqrt{\sigma^2}$$

n.p.(1-p)
ver slide
anterior



Data Science Academy



Um loja de Celular está oferecendo um plano de seguro para os clientes que perderam o celular accidentalmente ou por danos. Um representante de vendas, vendeu plano de seguro para 53% de seus clientes que compraram novos celulares. Suponha que o representante tenha vendido 16 celulares nessa semana. Com base nesse cenário o Diretor de Vendas deseja realizar as seguintes análises:



Data Science Academy



- 1) Qual é a probabilidade de que exatamente 10 clientes comprem um plano de seguro esta semana?
- 2) Qual é a probabilidade de que menos de 10 clientes comprem um plano de seguro esta semana?

Você como **Analista de Dados**, deve responder a essas questões e em seguida, construir um histograma para essa distribuição de probabilidade binomial e apresentar ao Diretor de Vendas. Isso vai auxiliar no processo de tomada de decisão, entre contratar ou não mais funcionários.



Data Science Academy

Esse tópico chegou ao final



Data Science Academy



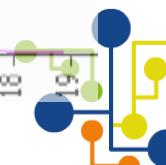
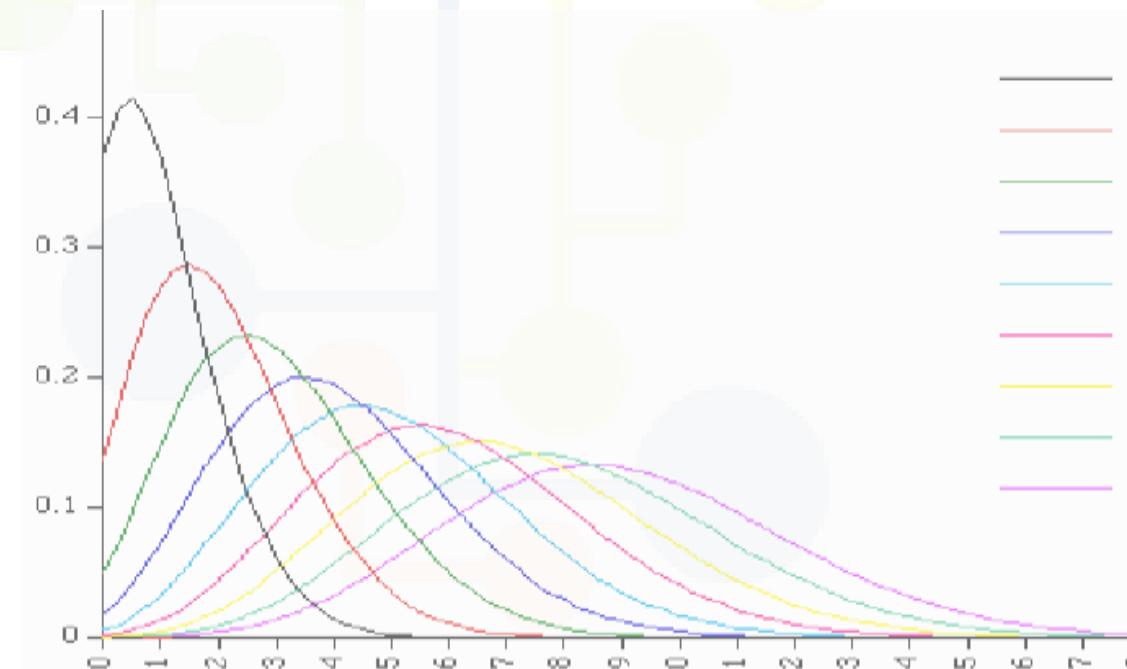
Distribuição Poisson



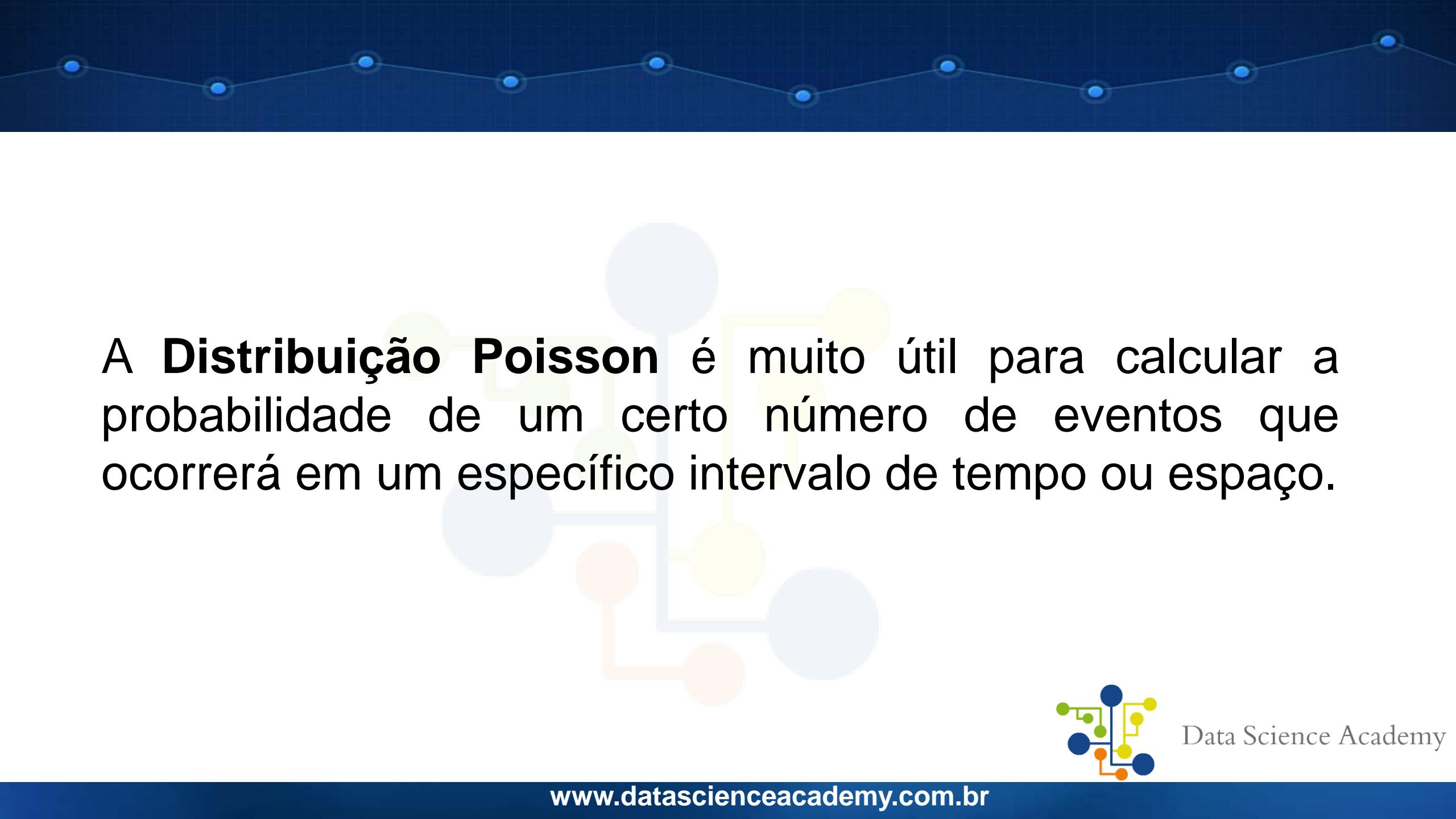
Data Science Academy

Distribuição Poisson

É outra distribuição de probabilidade discreta



Data Science Academy

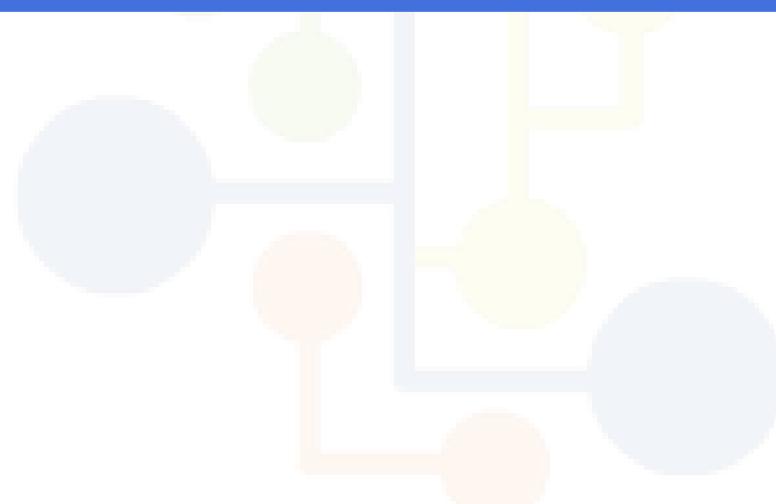


A **Distribuição Poisson** é muito útil para calcular a probabilidade de um certo número de eventos que ocorrerá em um específico intervalo de tempo ou espaço.

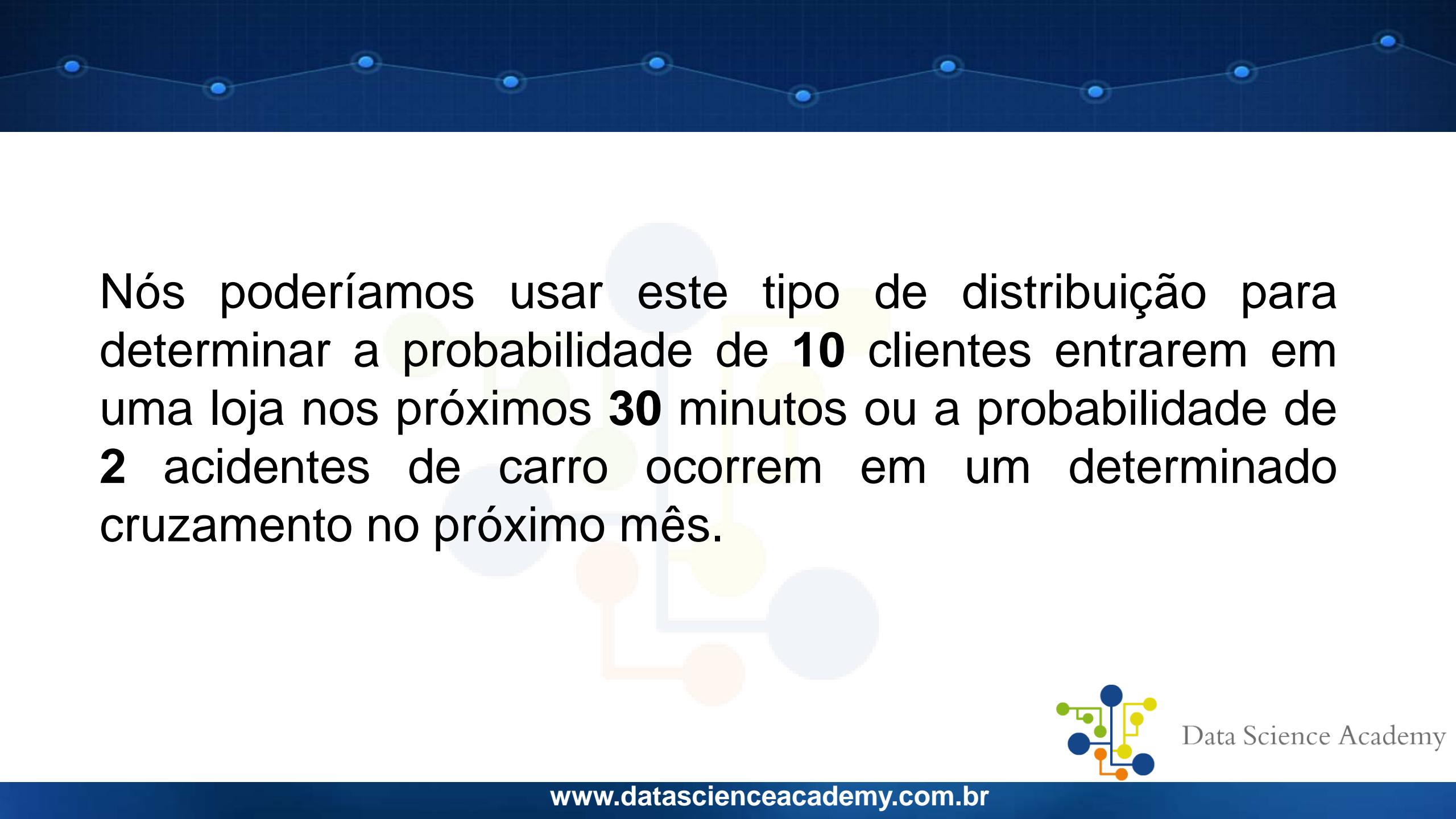


Data Science Academy

Exemplo



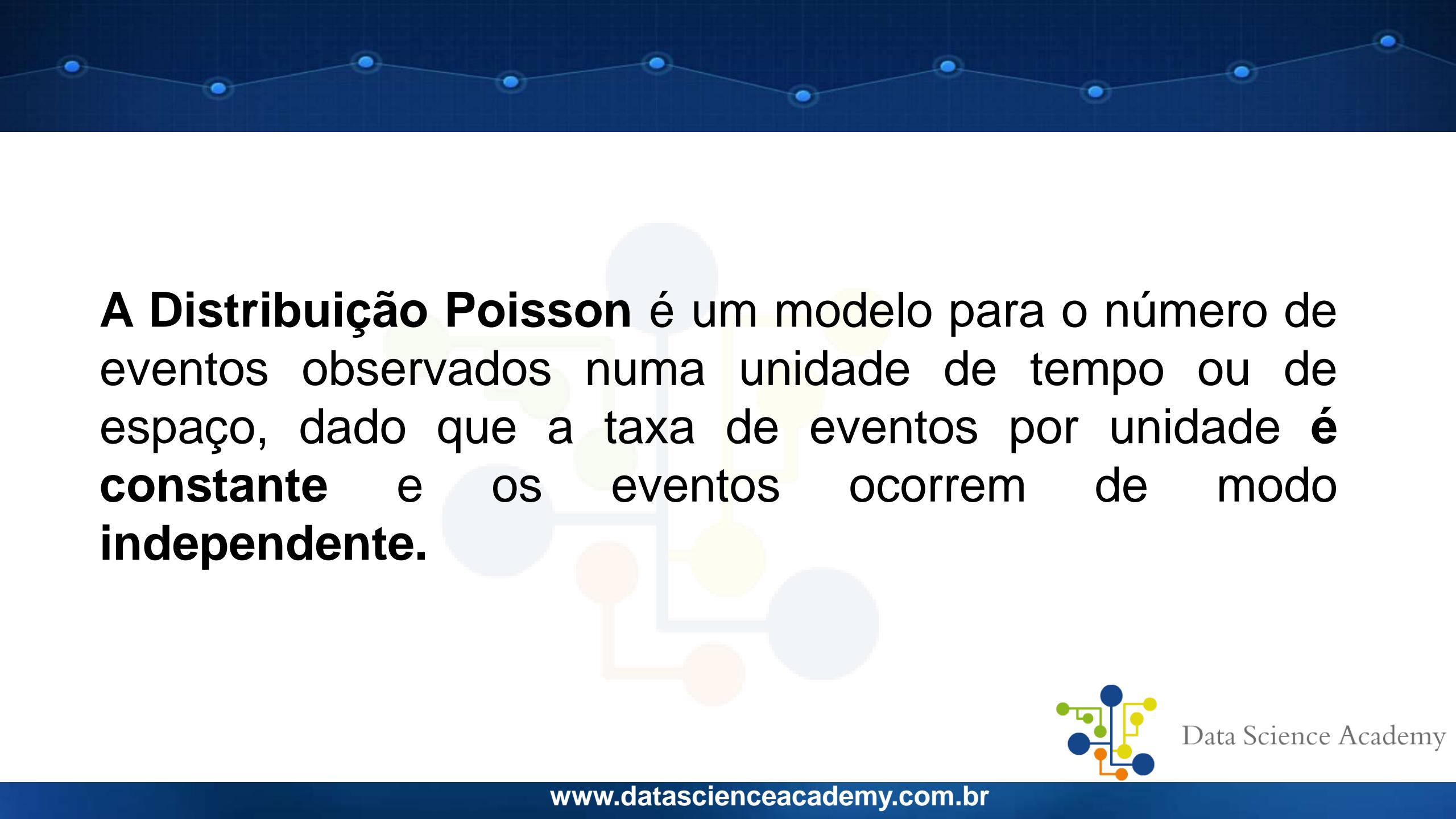
Data Science Academy



Nós poderíamos usar este tipo de distribuição para determinar a probabilidade de **10** clientes entram em uma loja nos próximos **30** minutos ou a probabilidade de **2** acidentes de carro ocorrem em um determinado cruzamento no próximo mês.



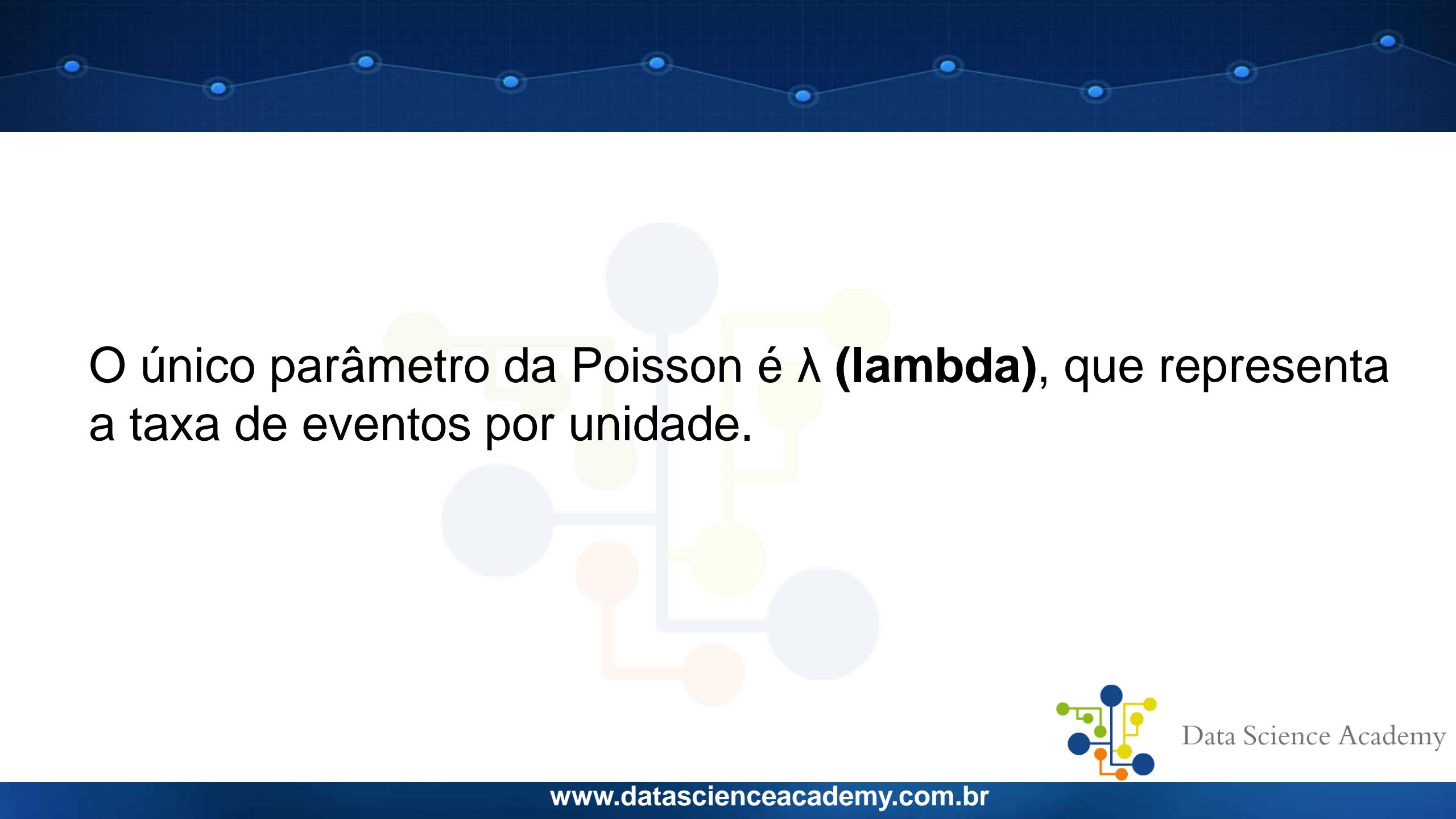
Data Science Academy



A Distribuição Poisson é um modelo para o número de eventos observados numa unidade de tempo ou de espaço, dado que a taxa de eventos por unidade é **constante** e os eventos ocorrem de modo **independente**.



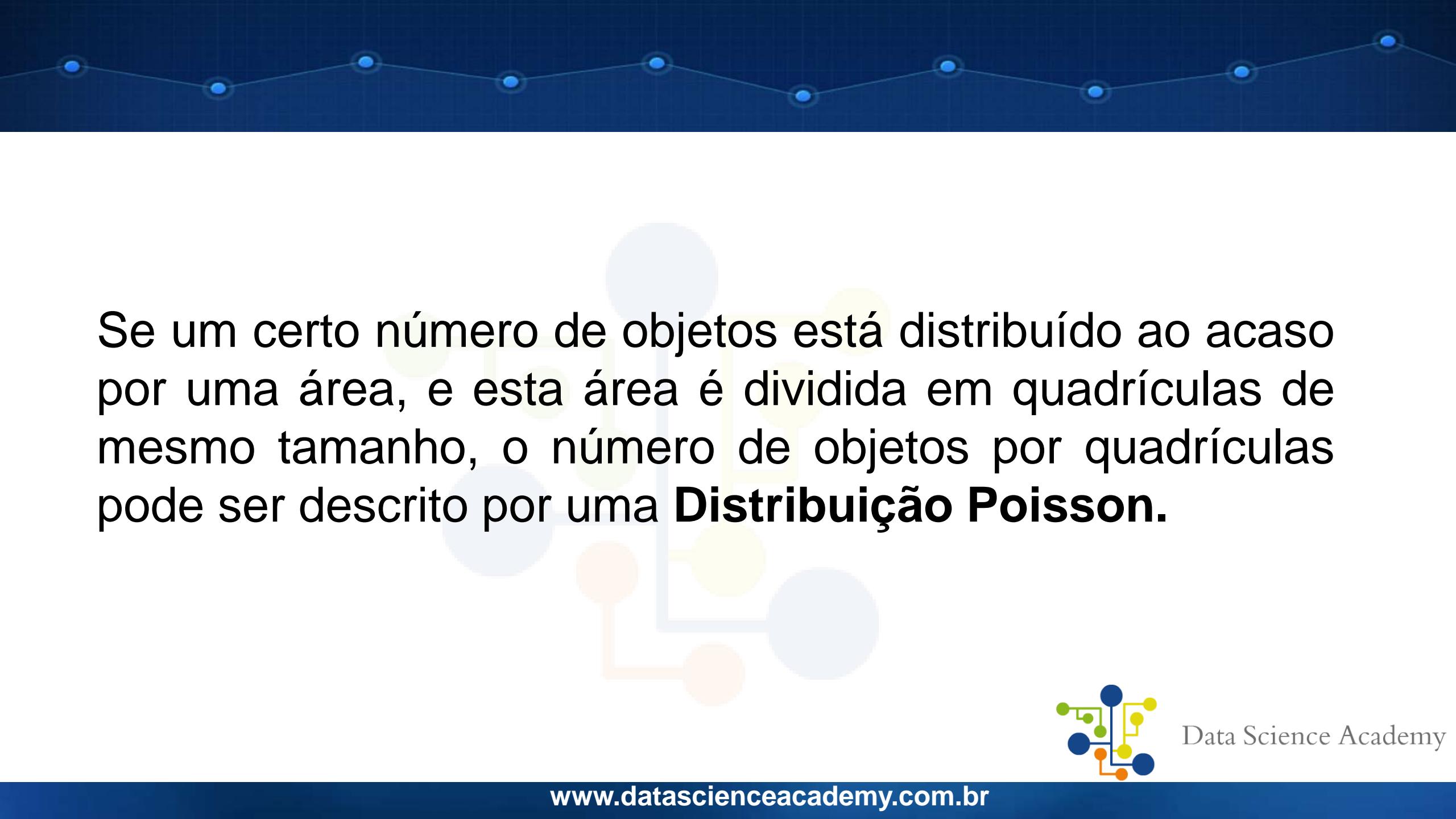
Data Science Academy



O único parâmetro da Poisson é λ (**lambda**), que representa a taxa de eventos por unidade.



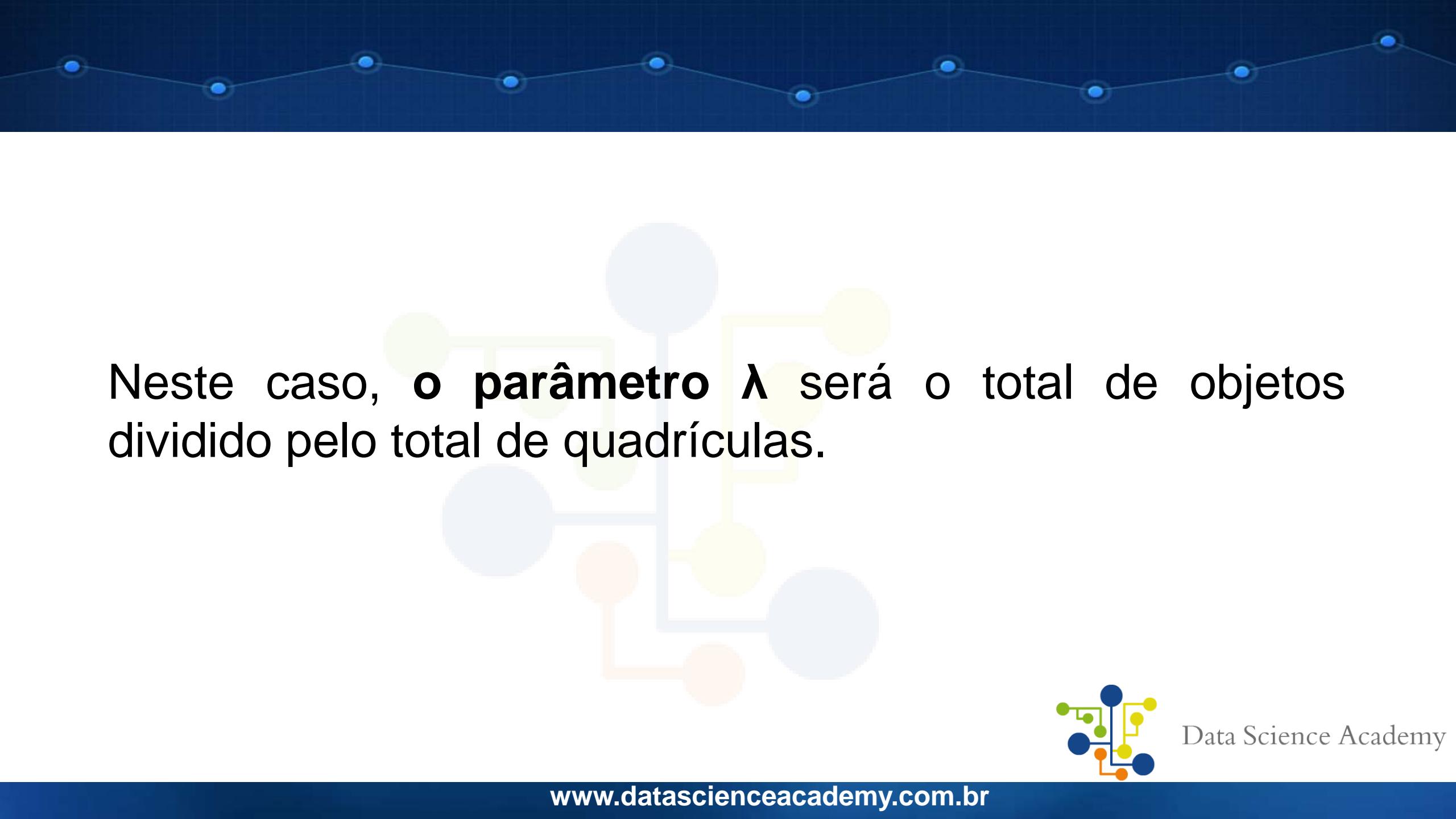
Data Science Academy



Se um certo número de objetos está distribuído ao acaso por uma área, e esta área é dividida em quadrículas de mesmo tamanho, o número de objetos por quadrículas pode ser descrito por uma **Distribuição Poisson**.



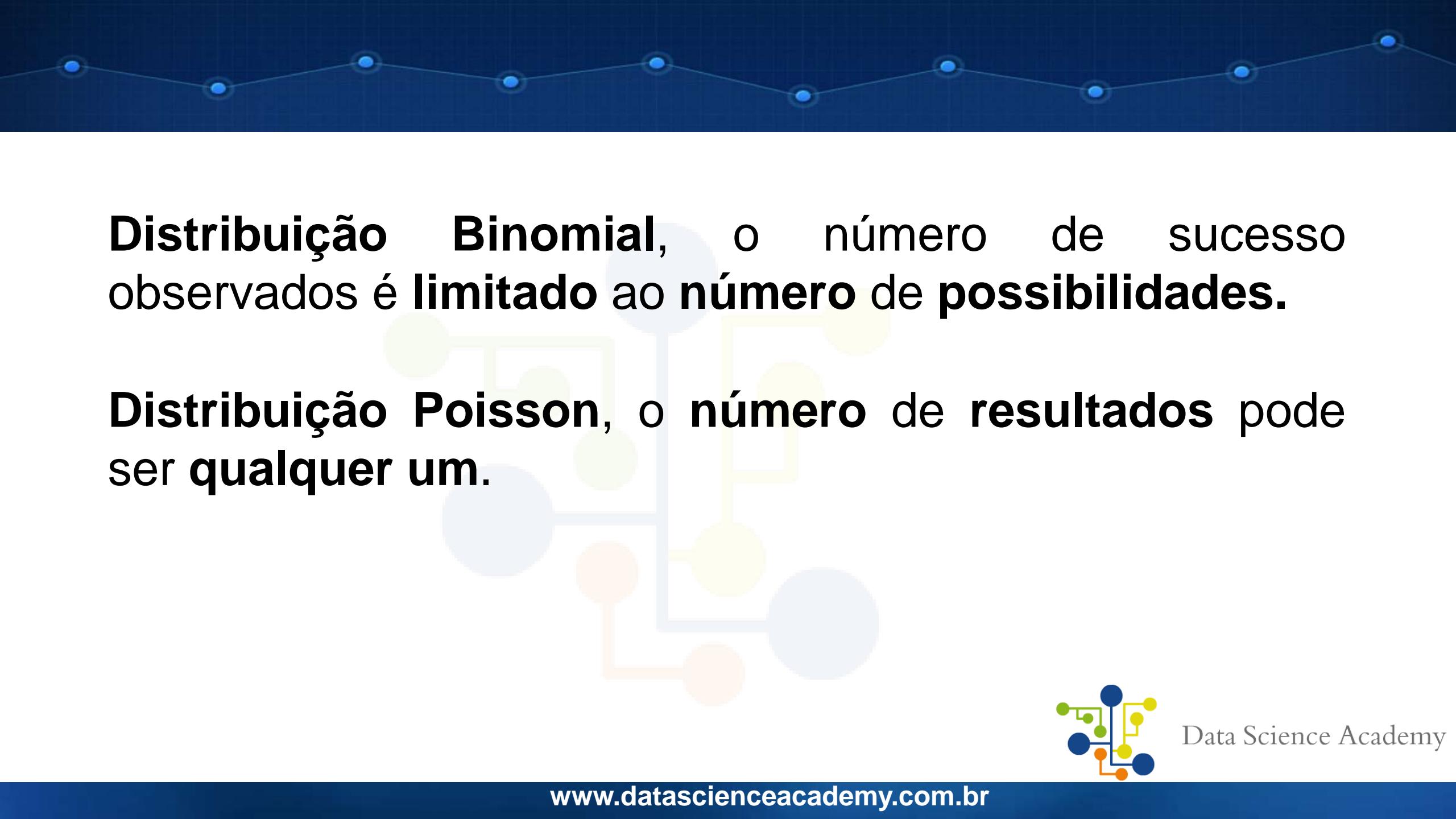
Data Science Academy



Neste caso, o **parâmetro λ** será o total de objetos dividido pelo total de quadrículas.



Data Science Academy



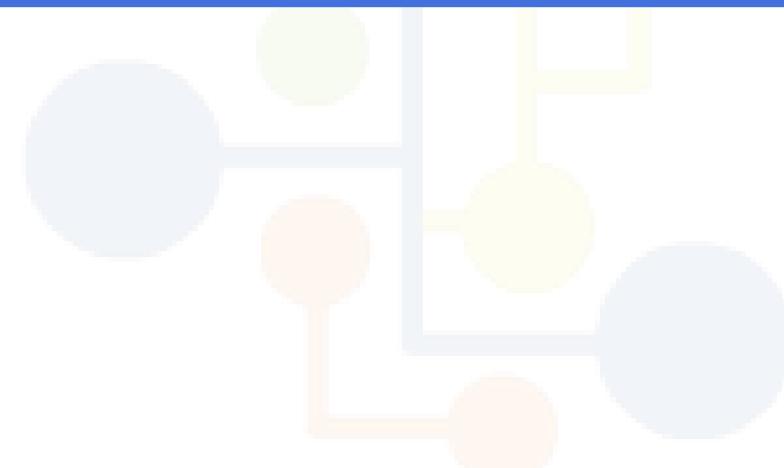
Distribuição Binomial, o número de sucesso observados é **limitado ao número de possibilidades**.

Distribuição Poisson, o número de resultados pode ser **qualquer um**.

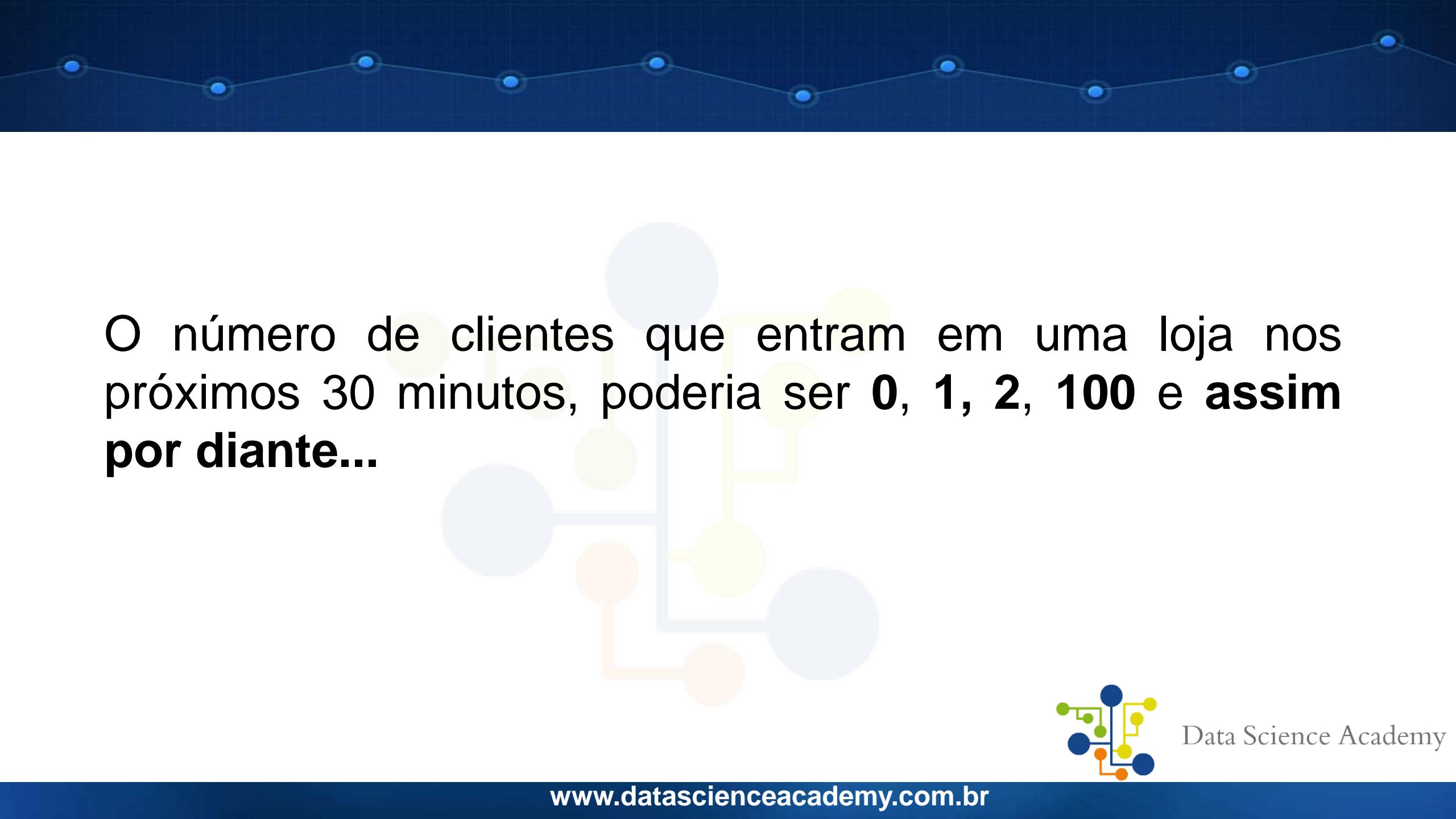


Data Science Academy

Exemplo



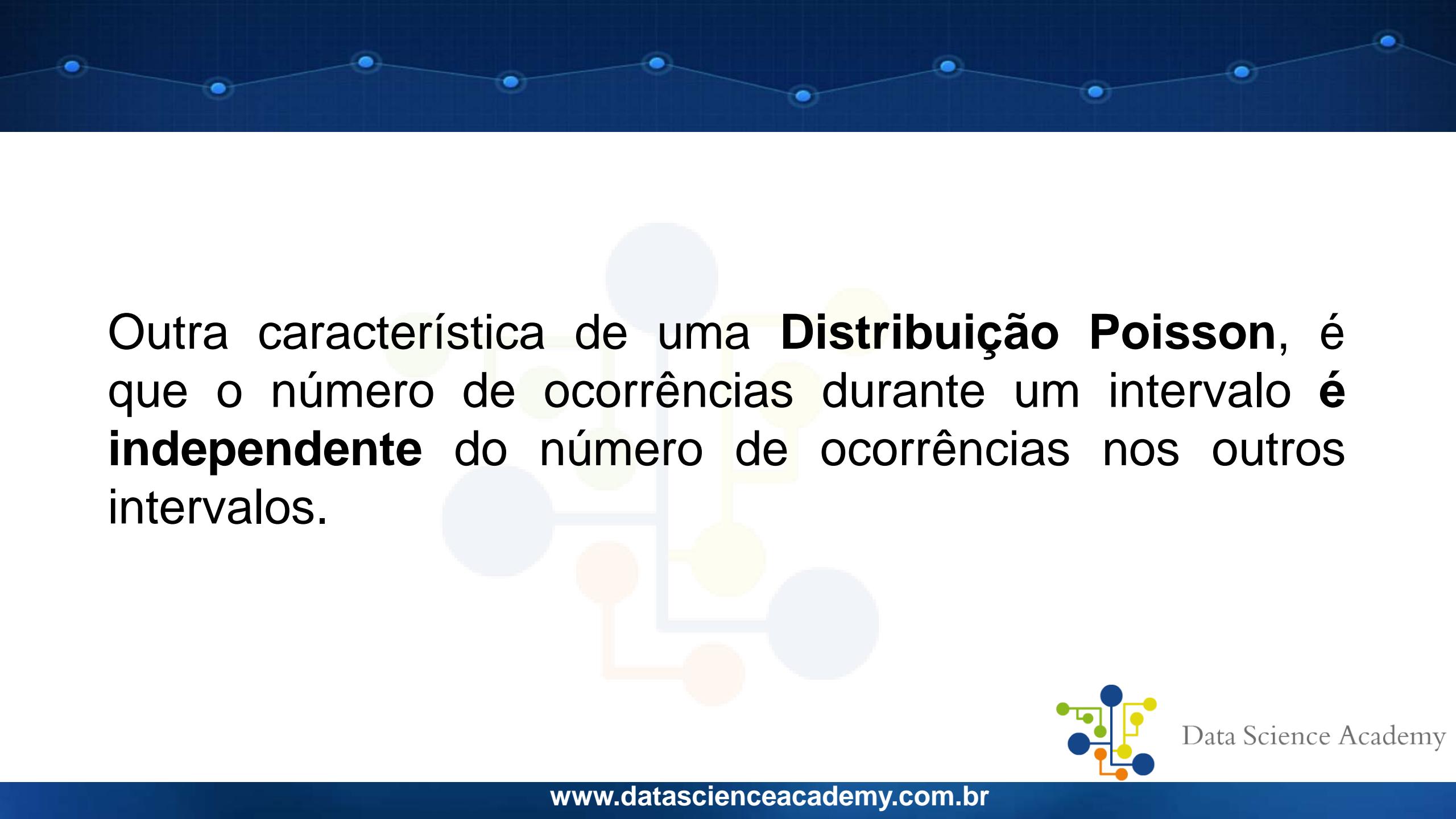
Data Science Academy



O número de clientes que entram em uma loja nos próximos 30 minutos, poderia ser **0, 1, 2, 100 e assim por diante...**



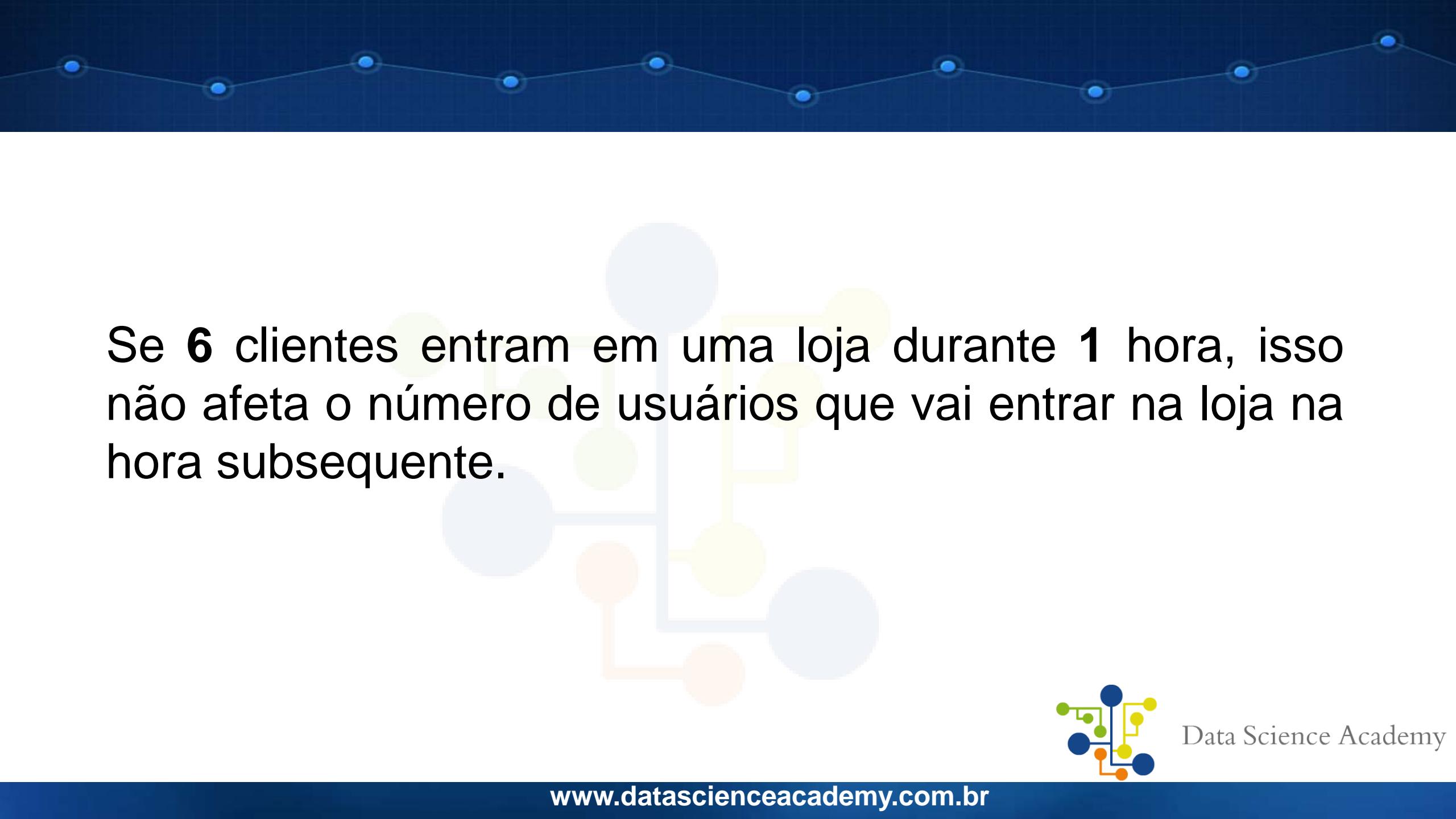
Data Science Academy



Outra característica de uma **Distribuição Poisson**, é que o número de ocorrências durante um intervalo é **independente** do número de ocorrências nos outros intervalos.



Data Science Academy

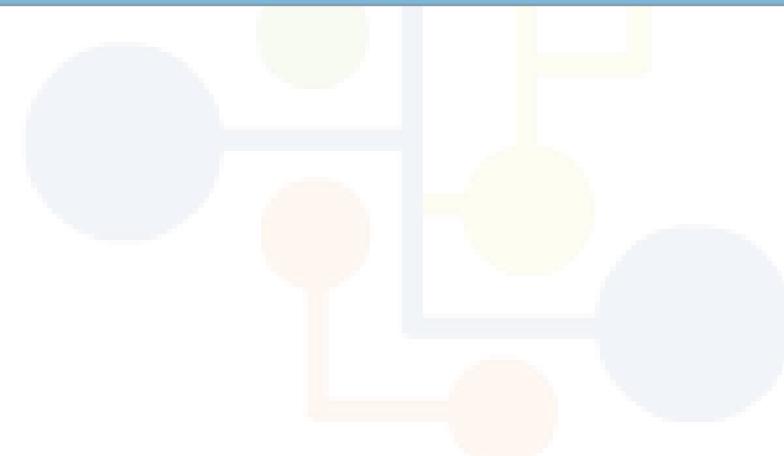


Se **6** clientes entram em uma loja durante **1** hora, isso não afeta o número de usuários que vai entrar na loja na hora subsequente.

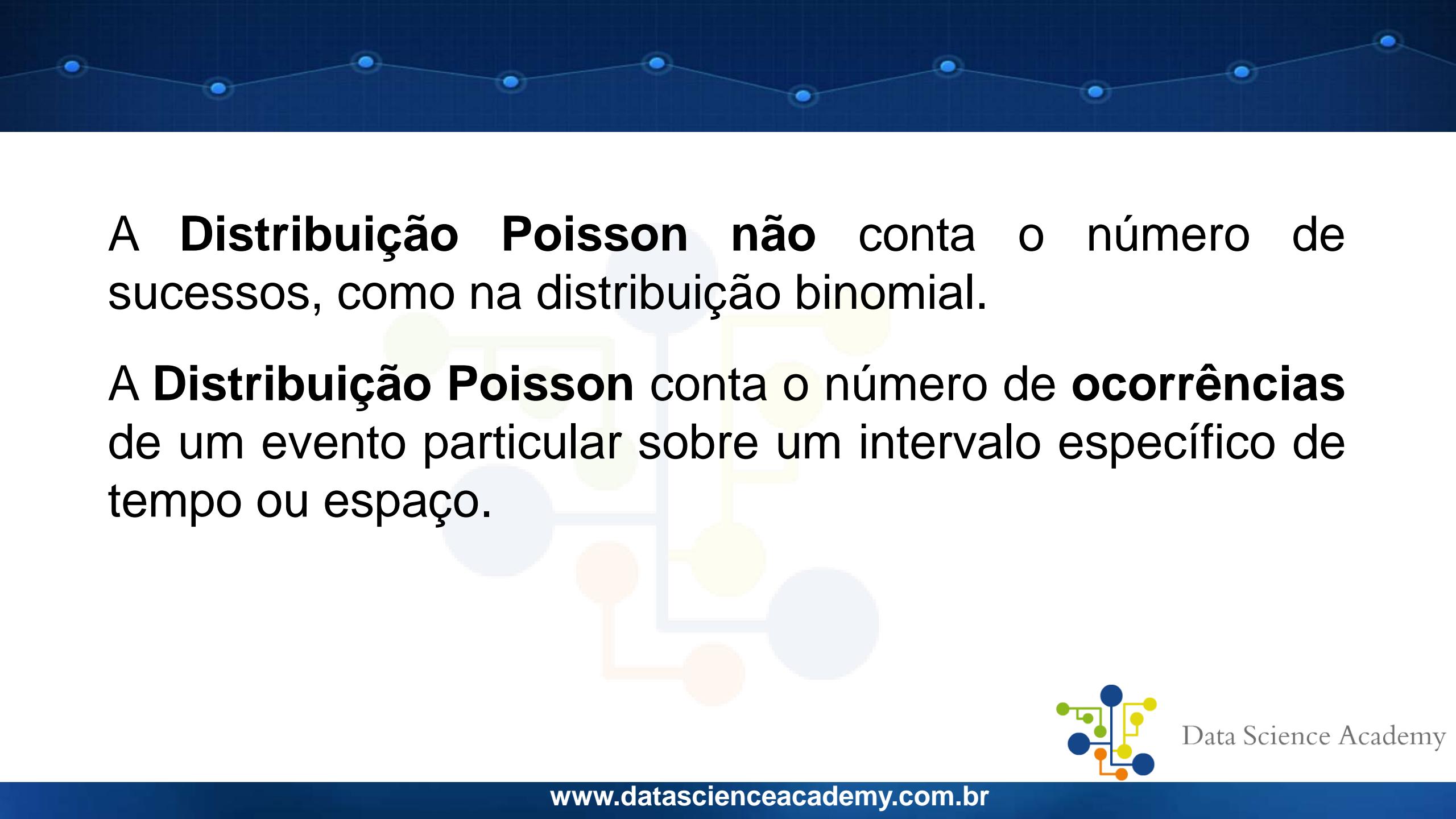


Data Science Academy

Em resumo



Data Science Academy



A **Distribuição Poisson** não conta o número de sucessos, como na distribuição binomial.

A **Distribuição Poisson** conta o número de **ocorrências** de um evento particular sobre um intervalo específico de tempo ou espaço.



Data Science Academy

Um canteiro de Obras de uma construtora gostaria de controlar o número de acidentes ocorridos a cada mês em suas dependências. Suponha que o número de acidentes a cada mês segue uma distribuição de Poisson com uma média de 1.4 acidentes por mês. Você como Analista de Dados, foi contratado para realizar as seguintes análises:

- a) A probabilidade de que exatamente um acidente ocorrerá no próximo mês?
- b) A probabilidade de que menos de 2 acidentes irão ocorrer no próximo mês?
- c) A probabilidade de que 4 ou menos acidentes irão ocorrer no próximo mês?



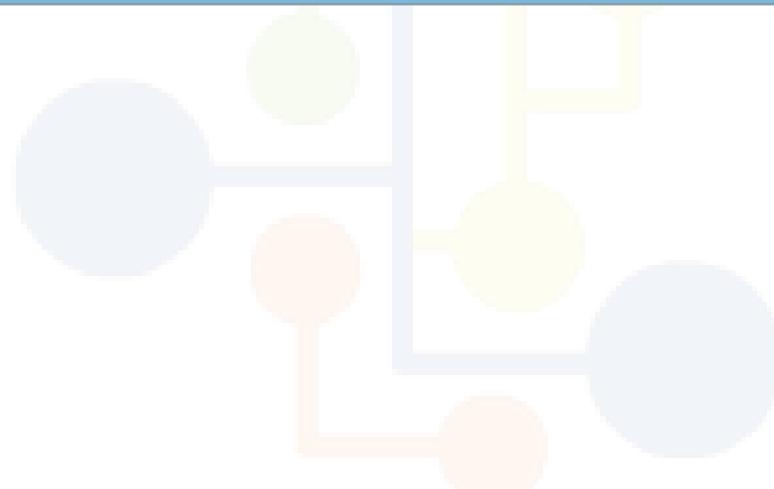
Data Science Academy

Esse tópico chegou ao final

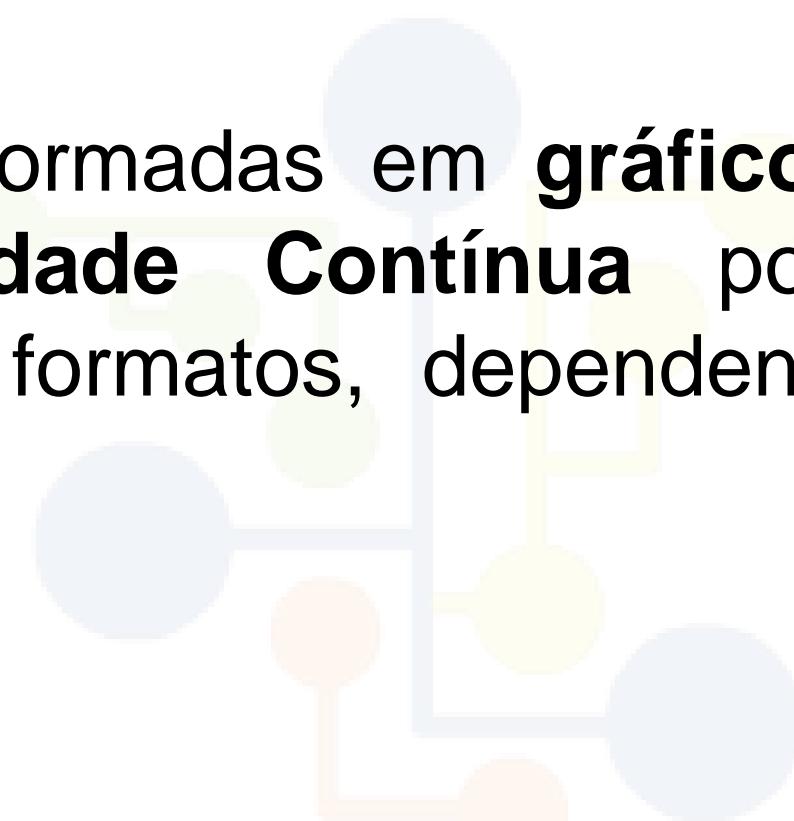


Data Science Academy

Distribuição de Probabilidade Contínua



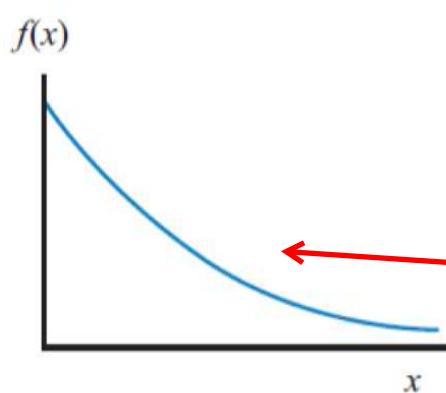
Data Science Academy



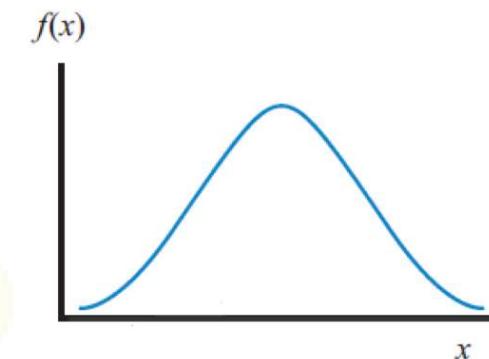
Quando transformadas em **gráficos**, as **Distribuições de Probabilidade Contínua** podem assumir uma variedade de formatos, dependendo dos valores dos dados.



Data Science Academy

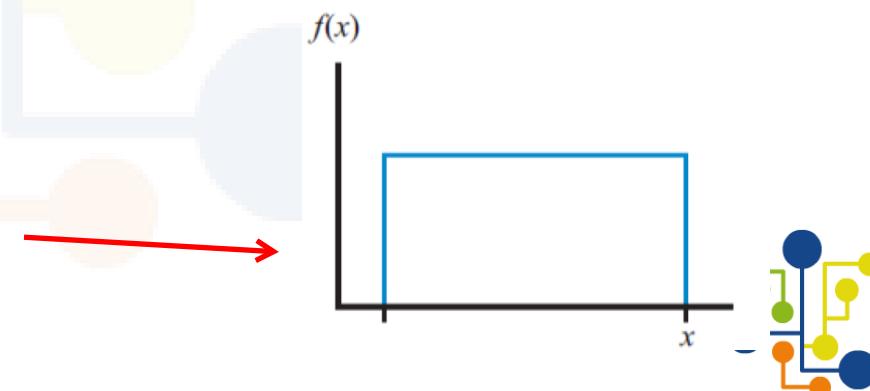


Distribuição Normal:

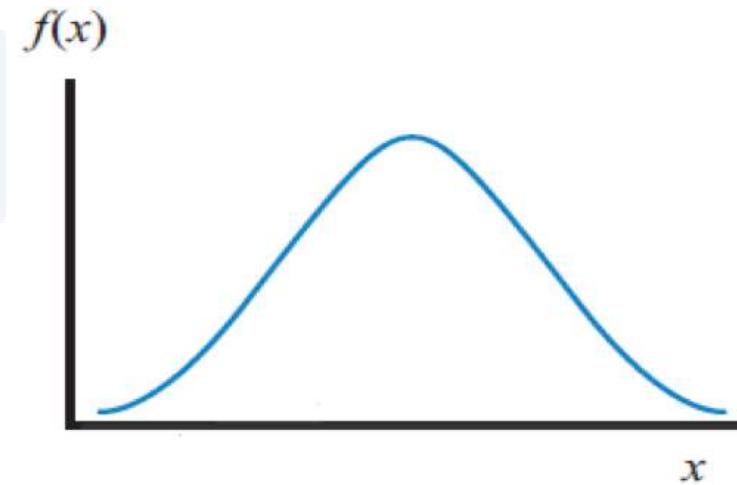


Distribuição Exponencial

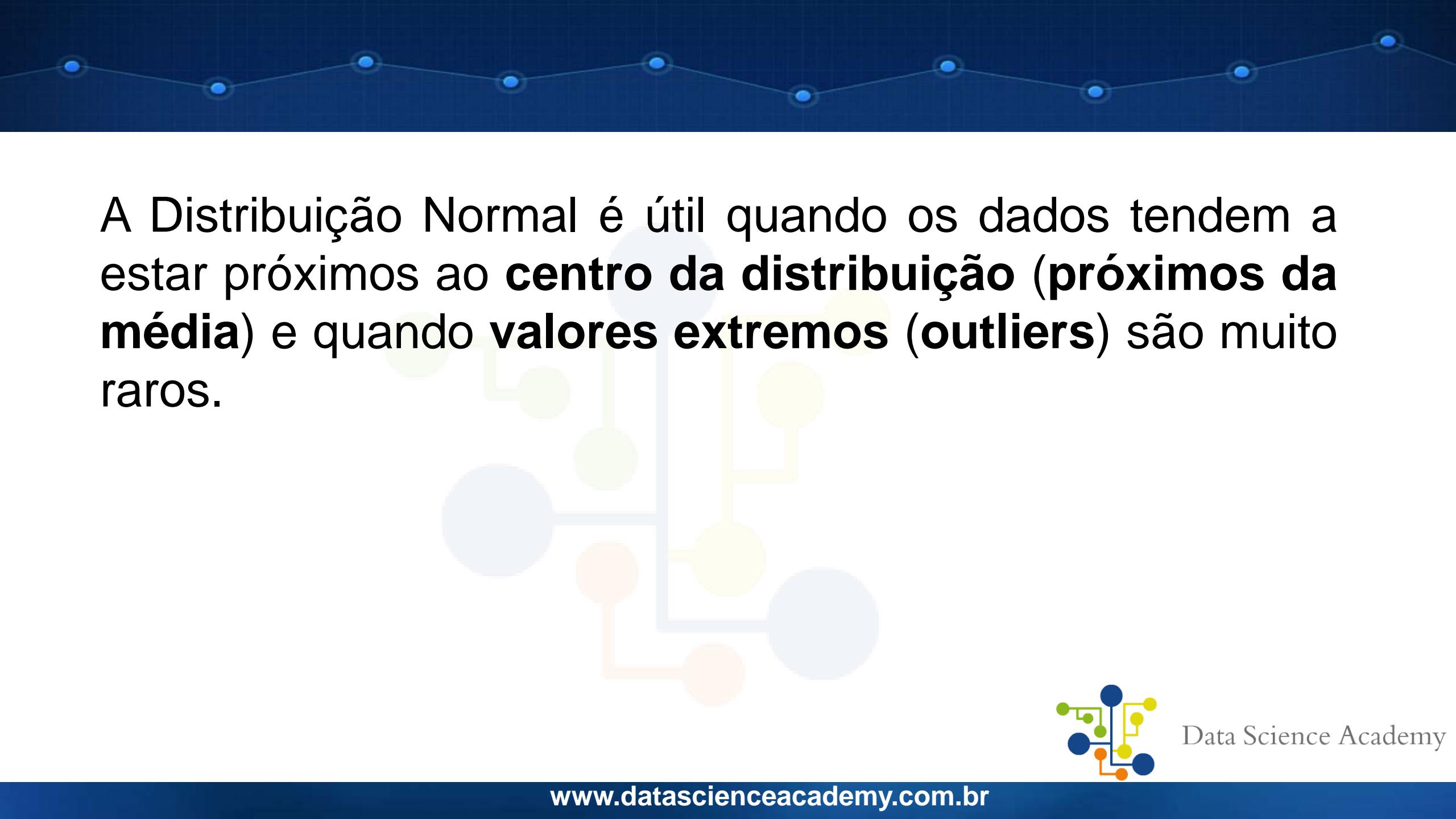
Distribuição Uniforme



Distribuição Normal



Data Science Academy



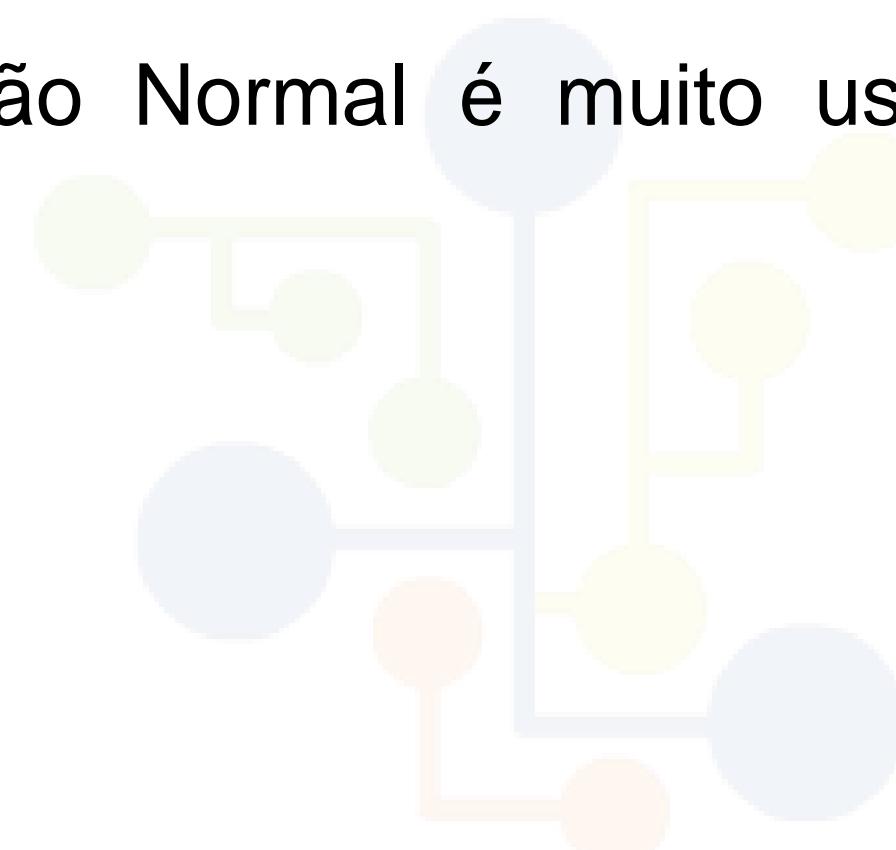
A Distribuição Normal é útil quando os dados tendem a estar próximos ao **centro da distribuição (próximos da média)** e quando **valores extremos (outliers)** são muito raros.



Data Science Academy

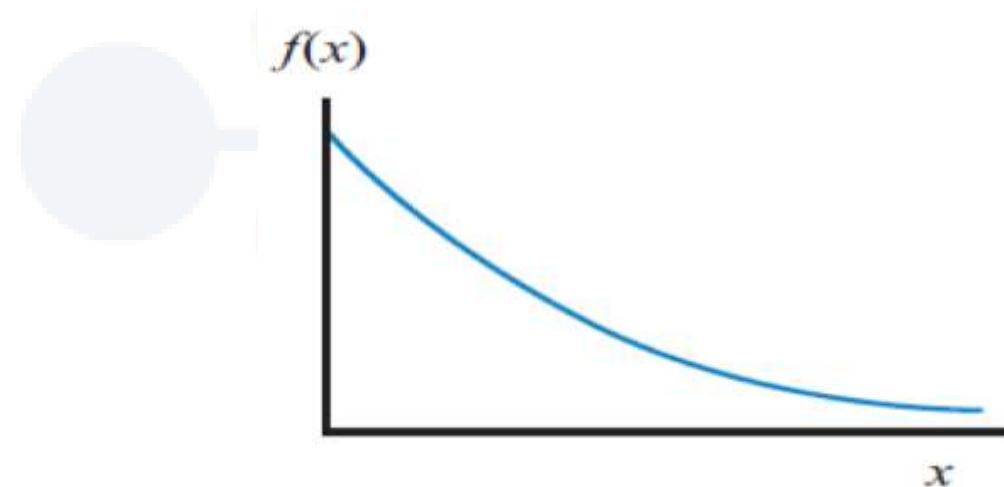


A Distribuição Normal é muito usada em controle de qualidade.



Data Science Academy

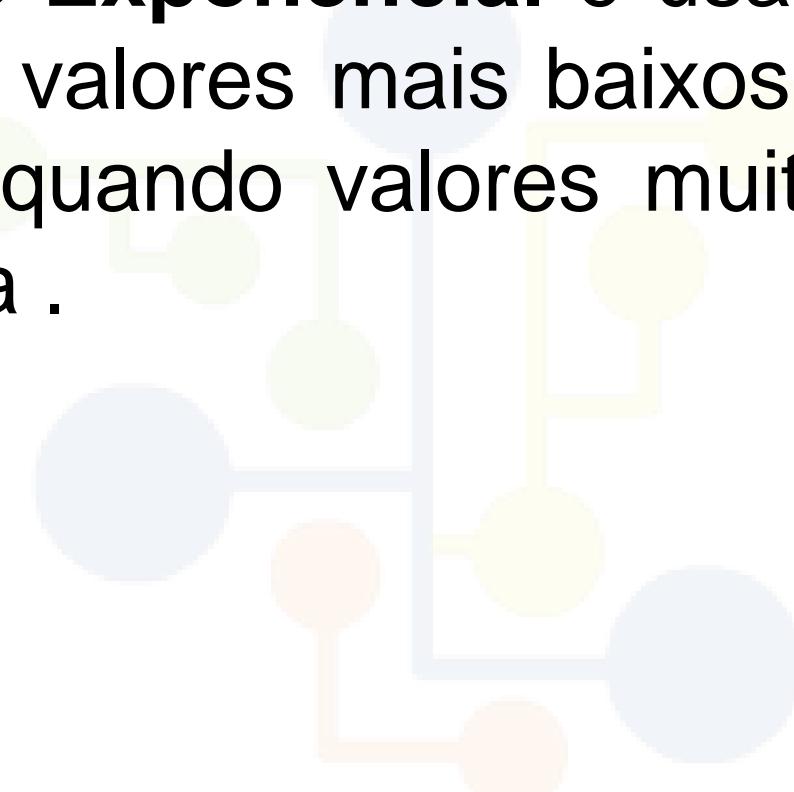
Distribuição Exponencial



Data Science Academy

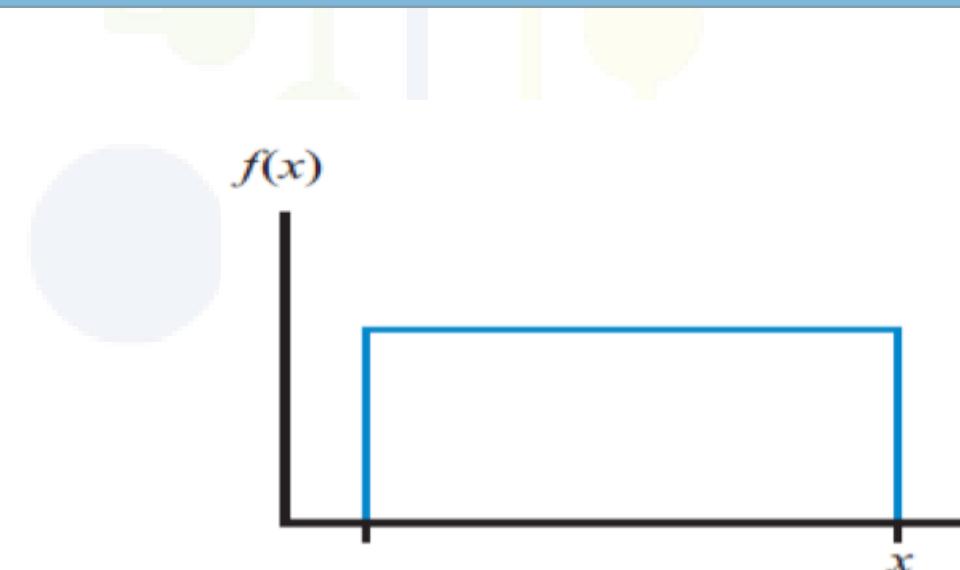


A **Distribuição Exponencial** é usada para descrever os dados quando valores mais baixos tendem a dominar a distribuição e quando valores muito altos não ocorrem com frequência .

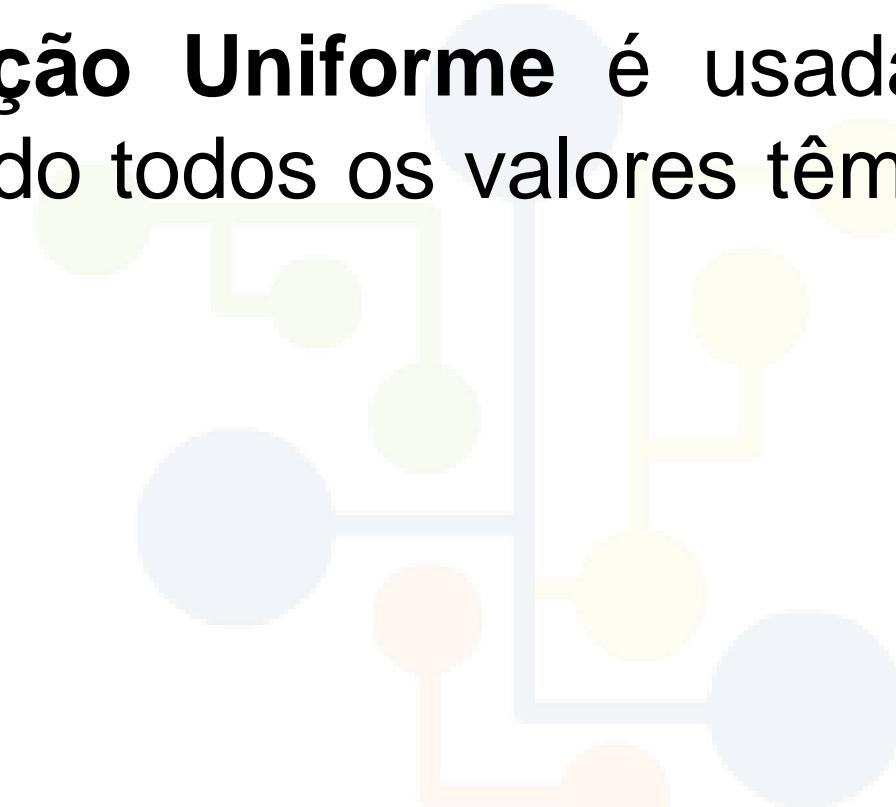


Data Science Academy

Distribuição Uniforme



Data Science Academy



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



Data Science Academy



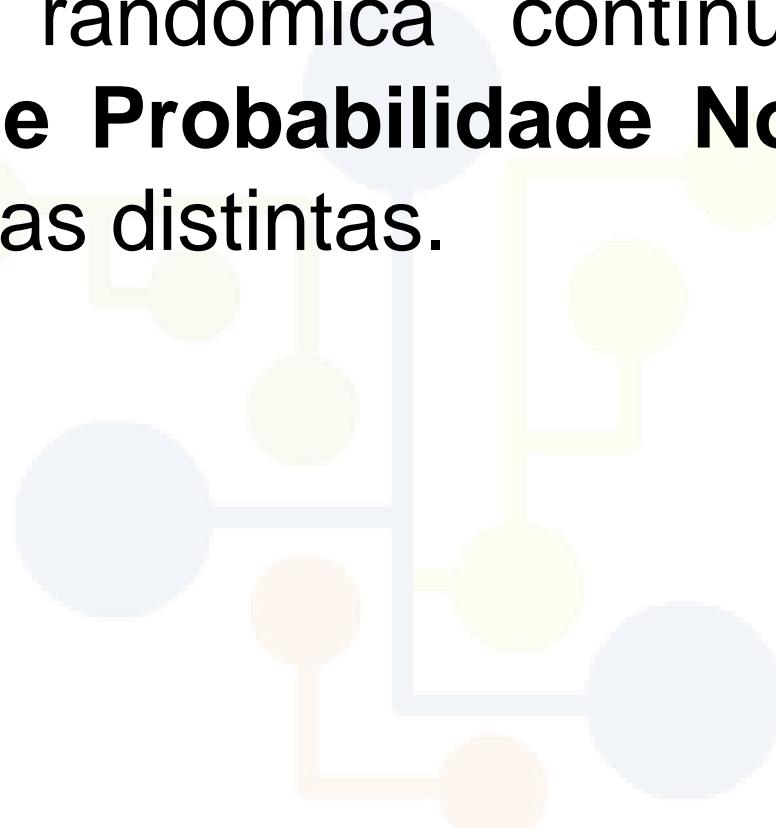
Distribuição Normal



Data Science Academy

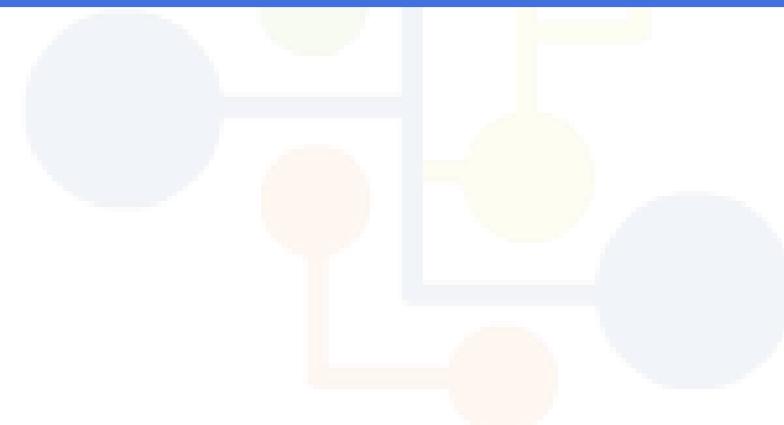


Uma variável randômica contínua que segue uma **Distribuição de Probabilidade Normal** tem uma série de características distintas.



Data Science Academy

Exemplo



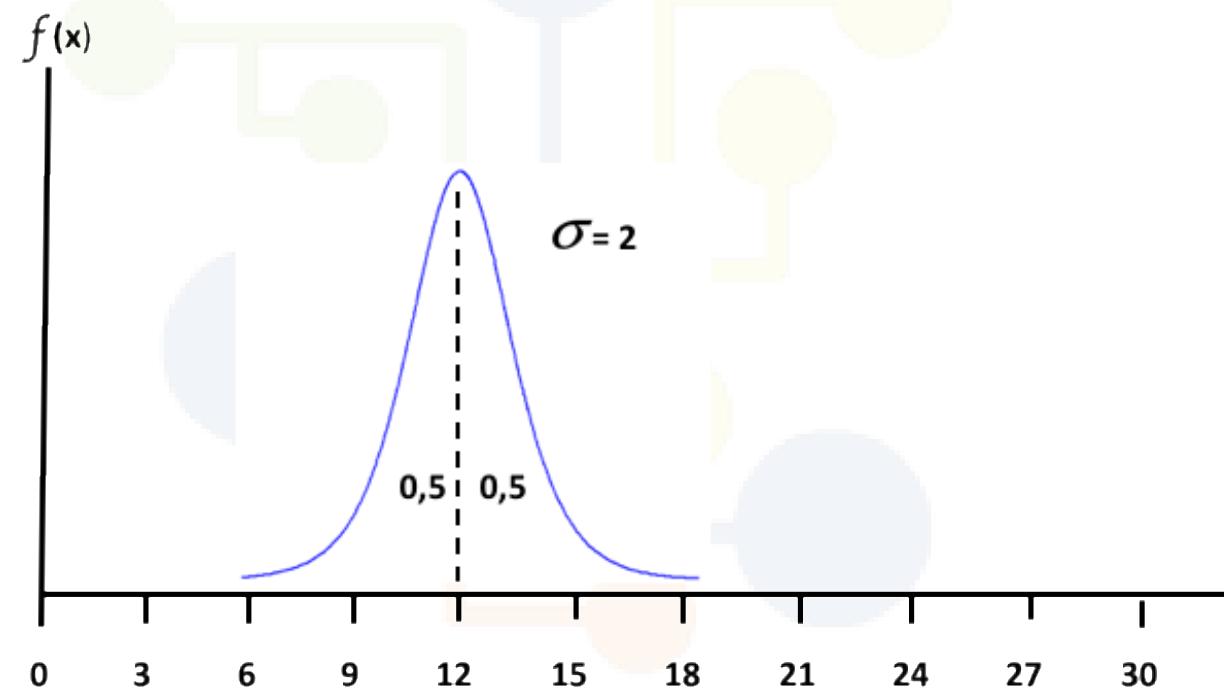
Data Science Academy

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).



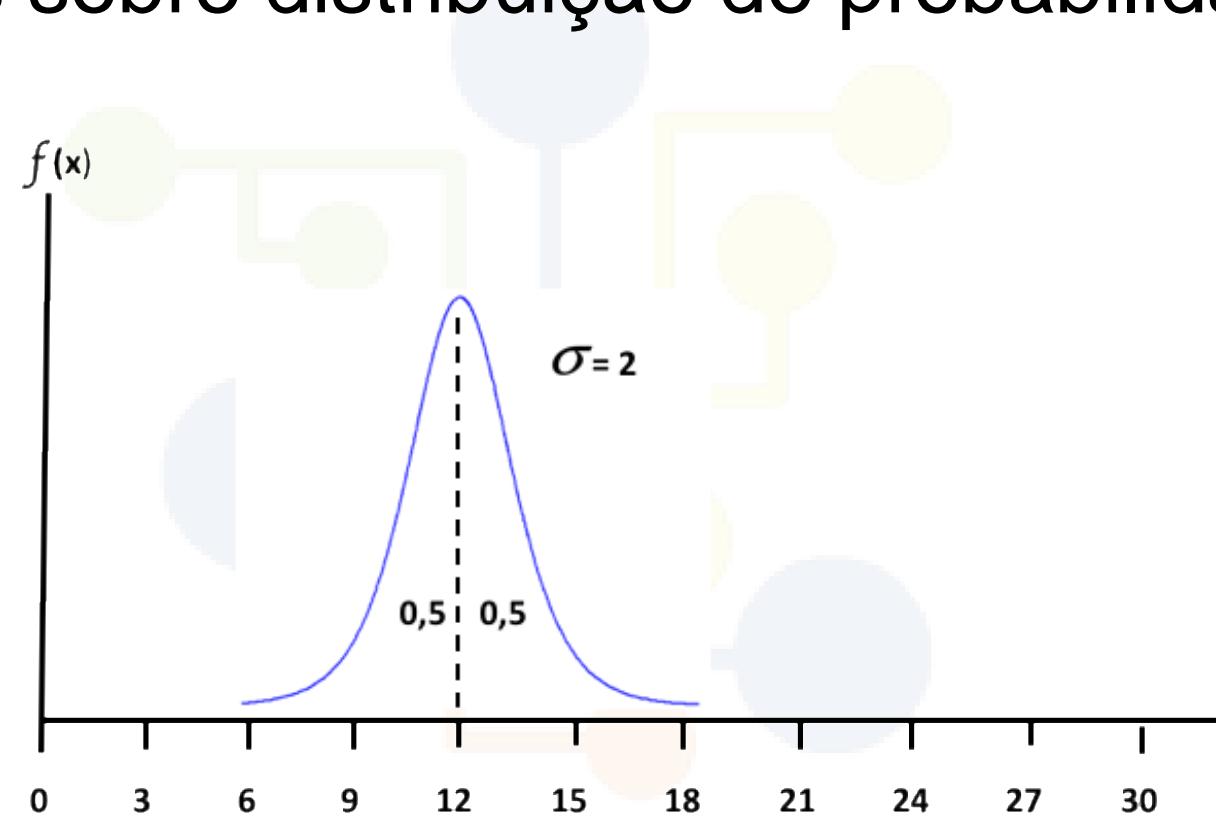
Data Science Academy

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

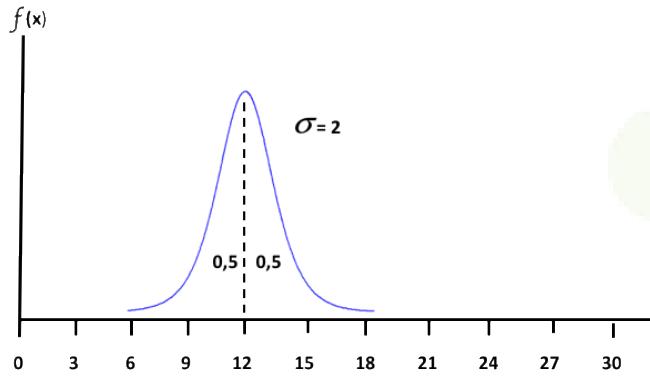


Data Science Academy

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



Data Science Academy



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.

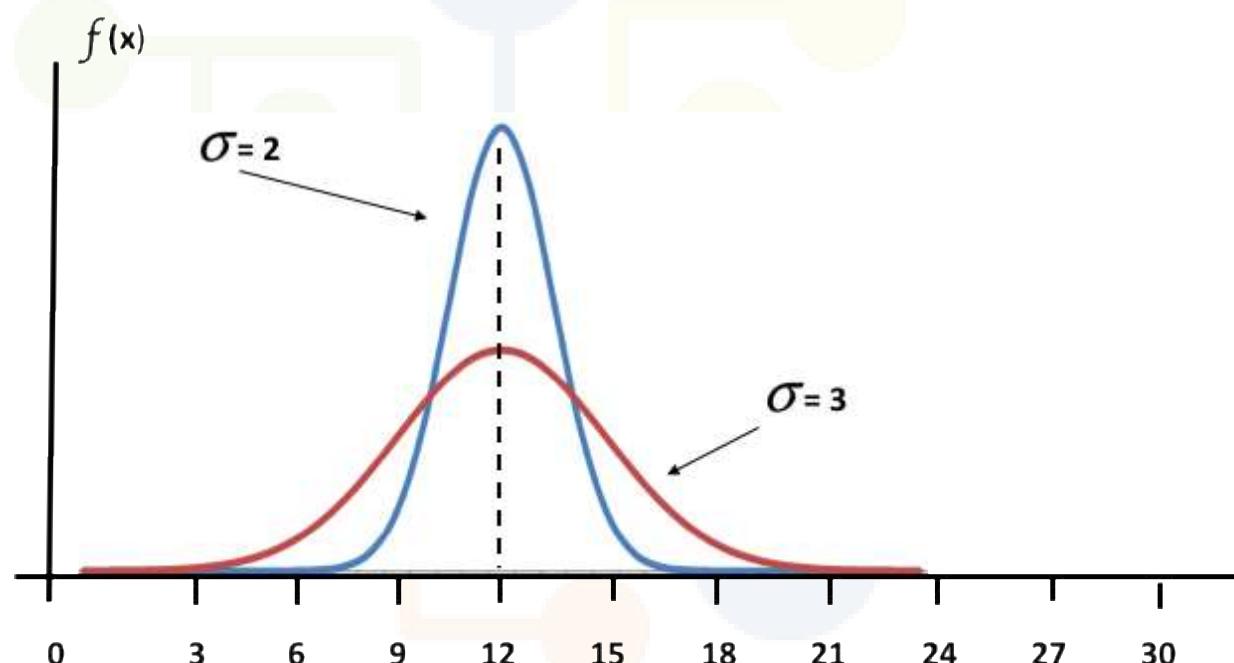
Variáveis randômicas em torno da média, na parte mais alta da curva, tem maior probabilidade de ocorrer, que valores situados onde a curva é menor.

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.

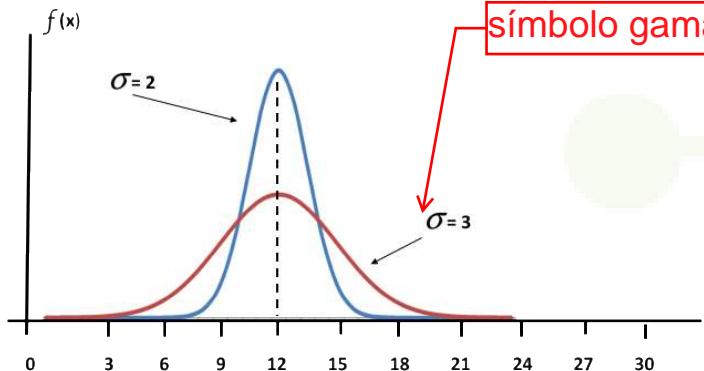


Data Science Academy

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



Data Science Academy



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.



Data Science Academy

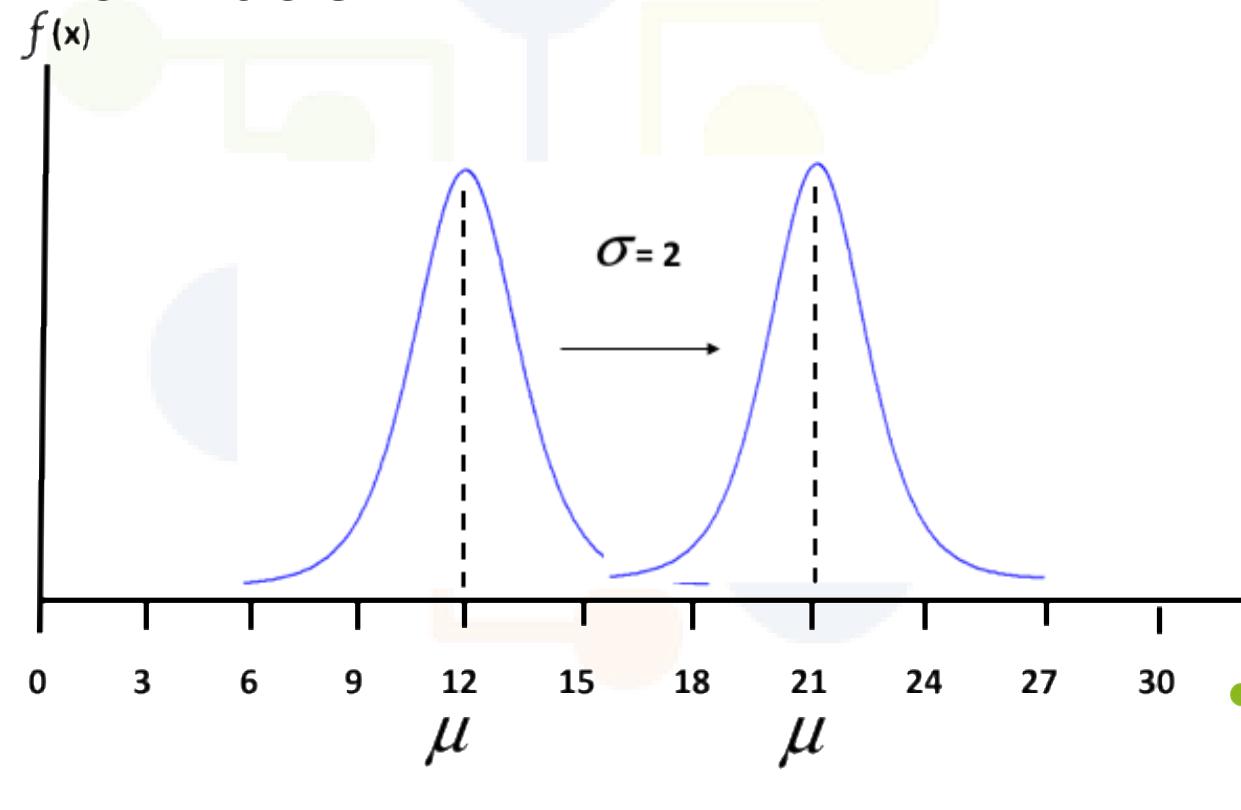


E se mudamos a média, de 12 para 21 minutos e
mantemos o desvio padrão de 2?



Data Science Academy

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



Data Science Academy



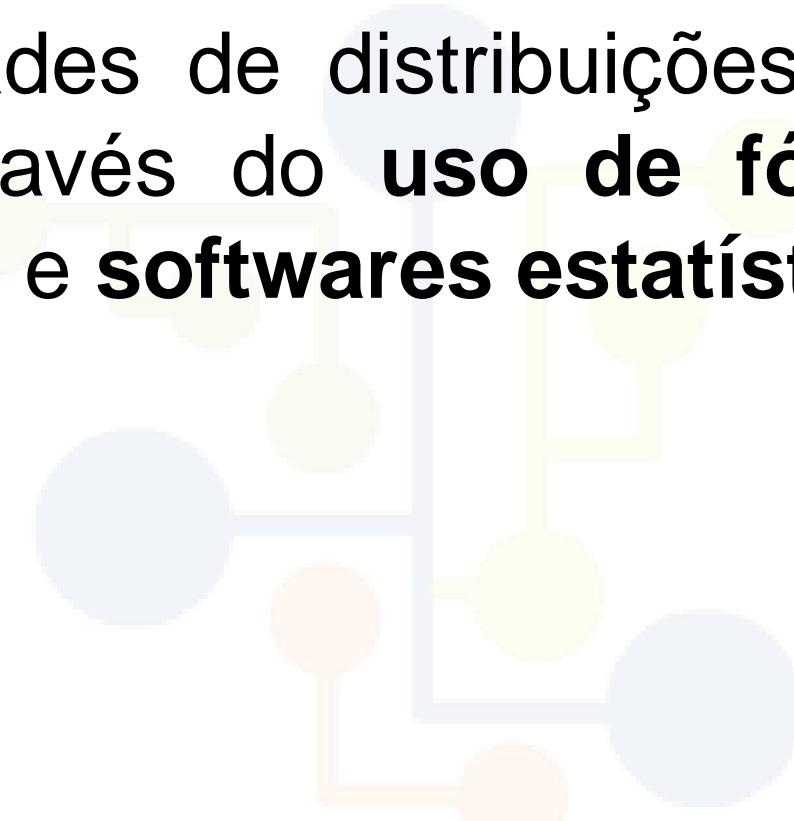
Em cada caso, os valores de média e desvio padrão, descrevem completamente o formato da distribuição.



Data Science Academy



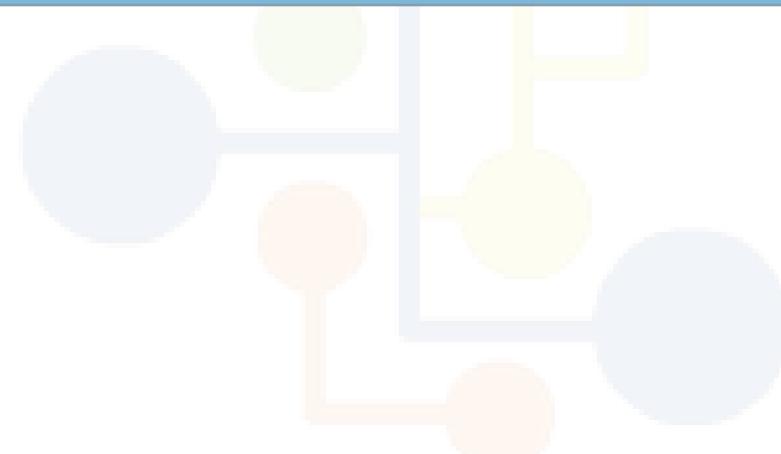
As probabilidades de distribuições normais podem ser calculadas através do **uso de fórmulas, tabelas de probabilidade e softwares estatísticos**



Data Science Academy

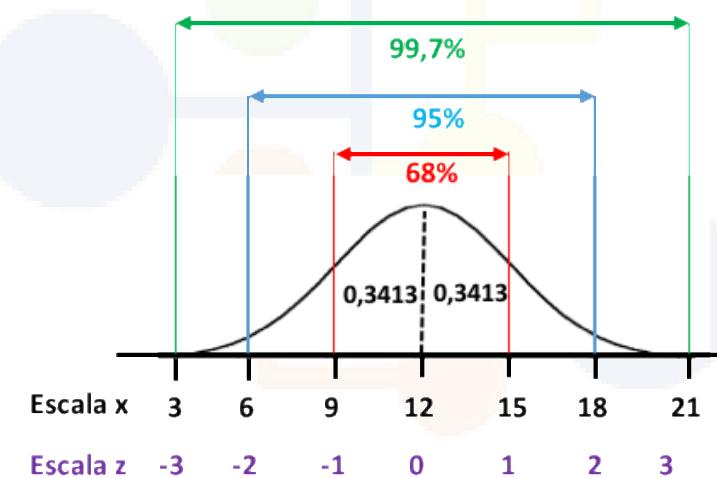


A Regra Empírica

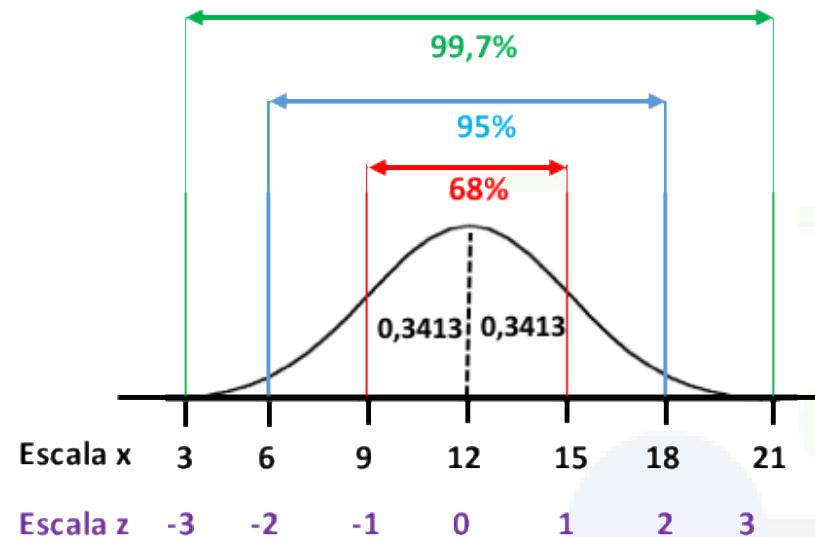


Data Science Academy

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 padrões acima e abaixo da média, respectivamente:



Data Science Academy

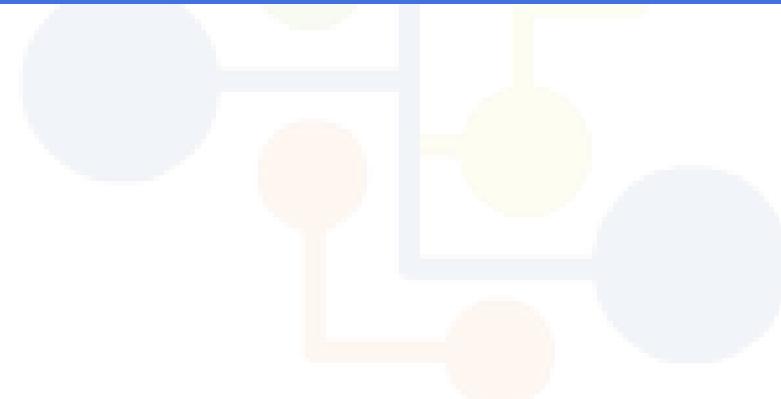


Ou seja, de acordo com a regra empírica, esperamos que **68%** das ligações fiquem entre **9** e **15** minutos, **95%** entre **6** e **18** minutos e **99%** entre **3** e **21** minutos.



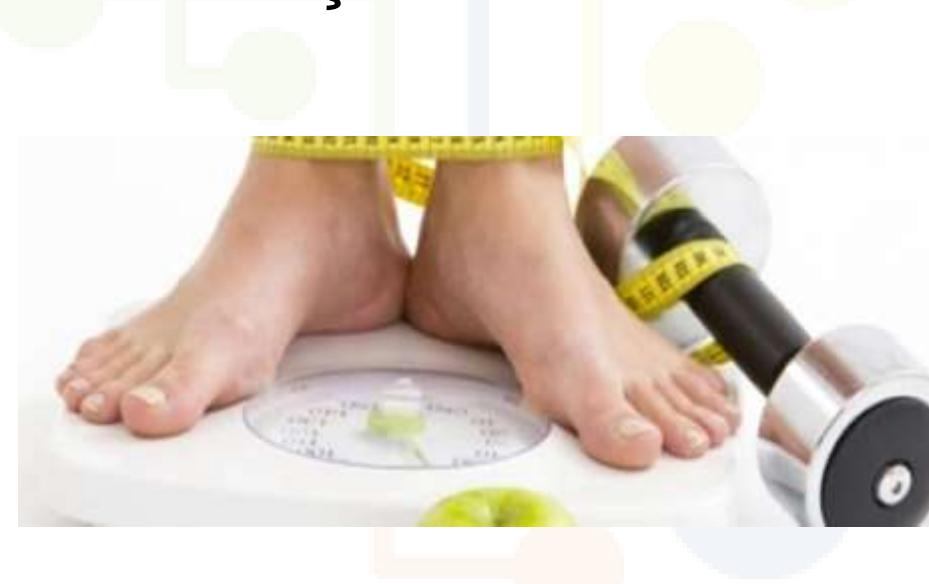
Data Science Academy

Exemplo



Data Science Academy

A Academia Corpus montou um treinamento especial para perda de peso. O coordenador afirma que a média de perda de peso nas 2 primeiras semanas é de **5kg**. Assumindo que a perda de peso resulta da distribuição normal com um desvio padrão de **3kg**.



Data Science Academy



A Academia Corpus montou um treinamento especial para perda de peso. O coordenador afirma que a média de perda de peso nas 2 primeiras semanas é de 5kg. Assumindo que a perda de peso resulta da distribuição normal com um desvio padrão de 3kg.

Analise:

- A probabilidade de uma pessoa perder menos de **7kg** após 2 semanas de treinamento.
- A probabilidade de uma pessoa ganhar **3kg** após 2 semanas de treinamento.



Data Science Academy

Veja, o Analista de Dados deverá estar ciente de que suas tarefas incluem:

- Abrir enormes repositórios de dados, que estão espalhados pelo mundo digital;
- Ter o domínio do uso das ferramentas de análise de dados;
- Ter o conhecimento da aplicação das regras estatísticas para gerar os relatórios de análise de dados para sua empresa, ou empresa para qual você trabalha.



Data Science Academy

Carreiras como analista de dados e cientista de dados, estão entre as de maior relevância até o ano de 2025. Segundo o Fórum Econômico Mundial.



Data Science Academy

Esse tópico chegou ao final



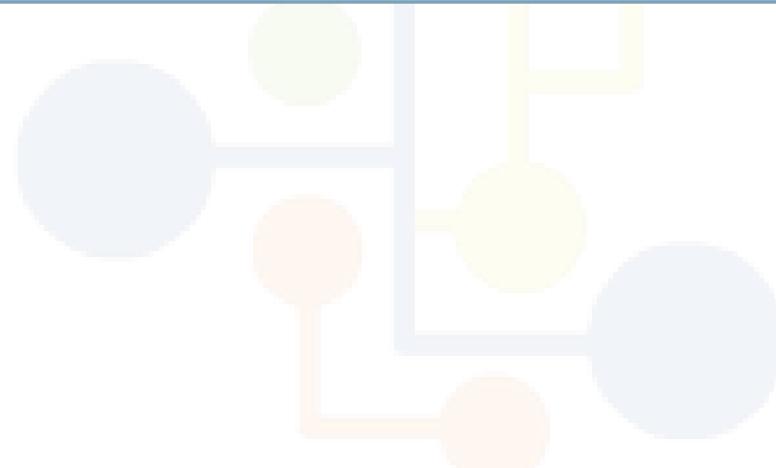
Obrigada



Data Science Academy



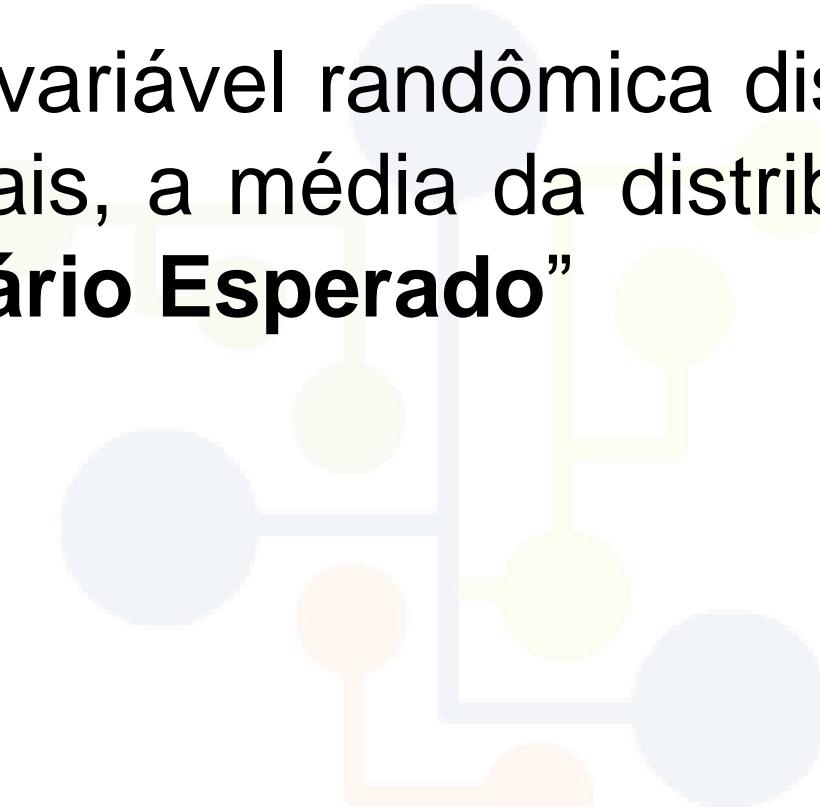
Valor Monetário Esperado



Data Science Academy



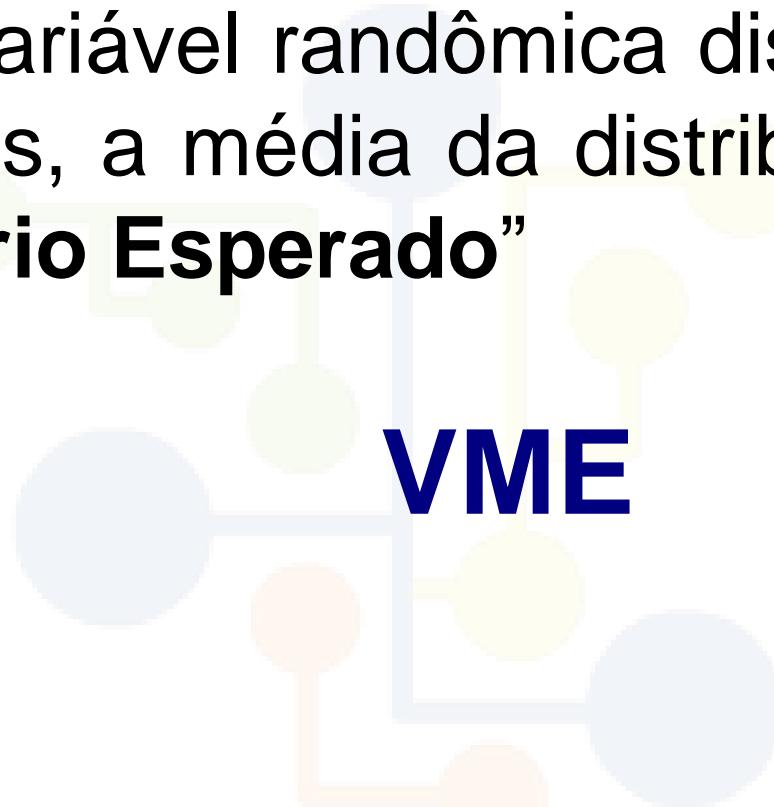
Quando uma variável randômica discreta é expressa em dólares ou reais, a média da distribuição é chamada de **“Valor Monetário Esperado”**



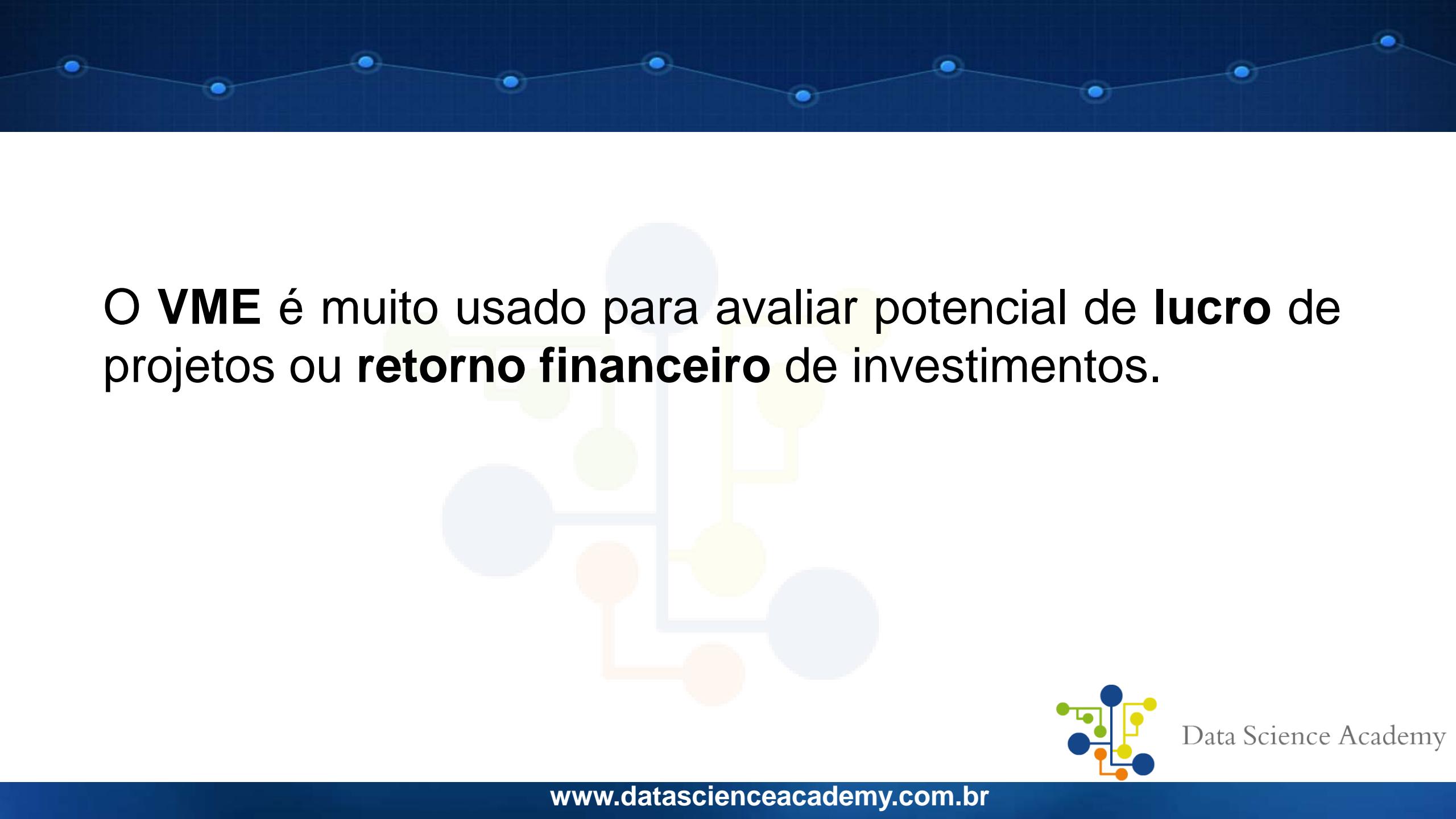
Data Science Academy



Quando uma variável randômica discreta é expressa em dólares ou reais, a média da distribuição é chamada de **“Valor Monetário Esperado”**



Data Science Academy

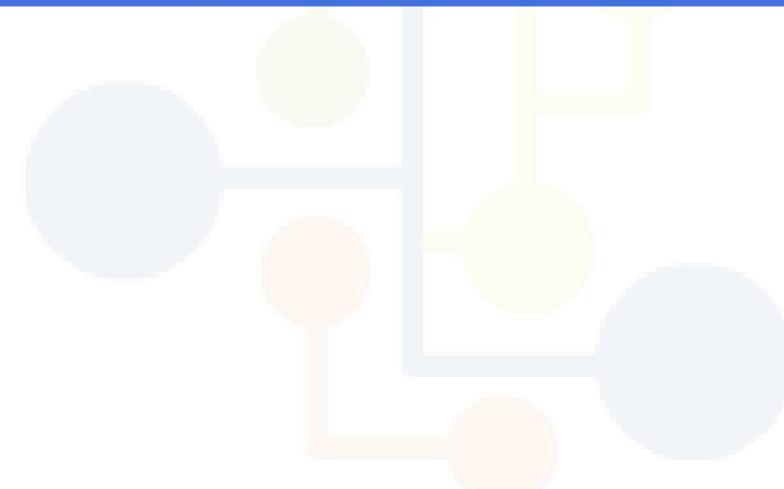


O VME é muito usado para avaliar potencial de **lucro** de projetos ou **retorno financeiro** de investimentos.



Data Science Academy

Exemplo



Data Science Academy

Uma construtora será responsável, por construir uma ponte.



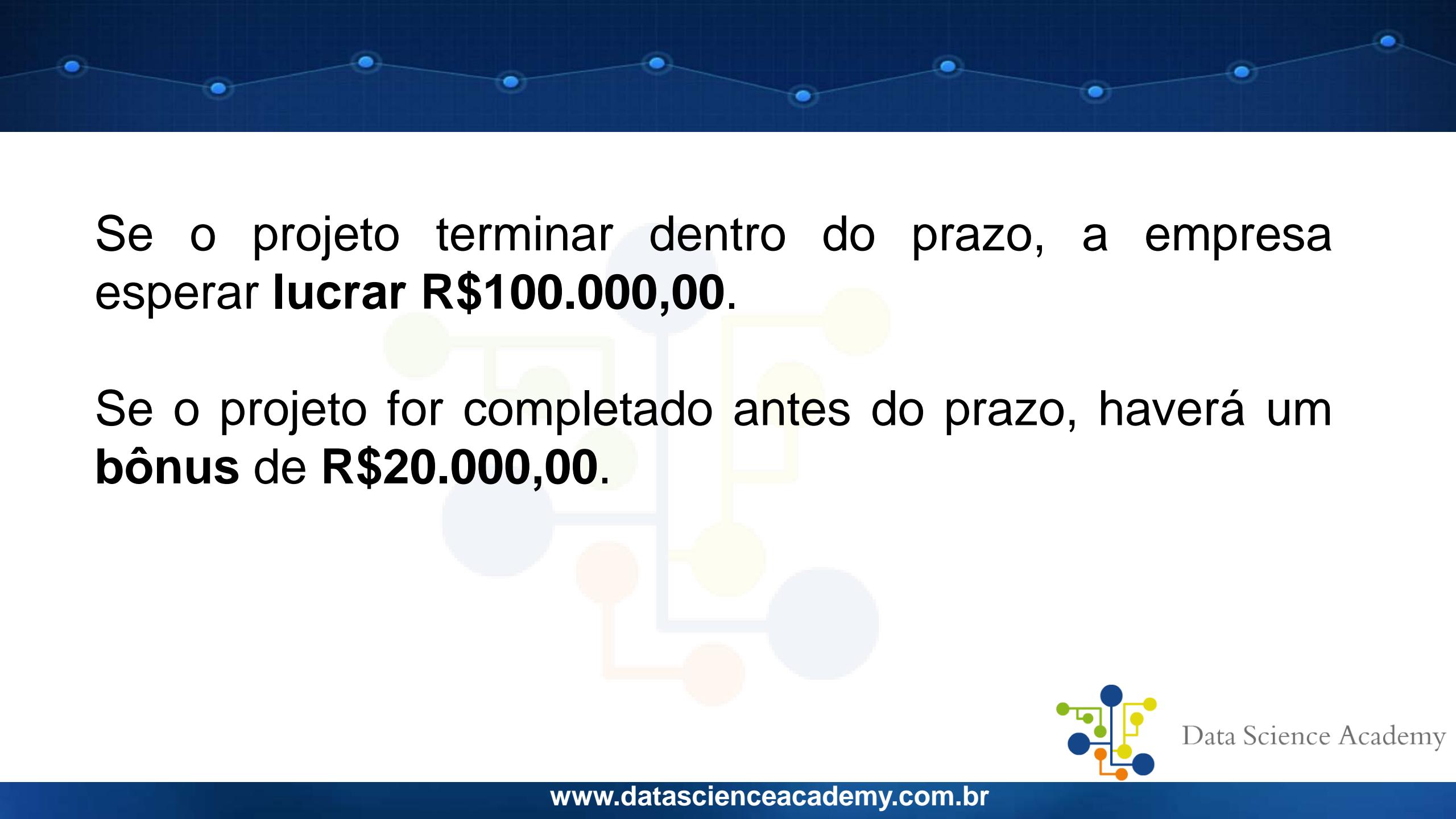
Data Science Academy



Se o projeto terminar dentro do prazo, a empresa
esperar **lucrar R\$100.000,00**.



Data Science Academy

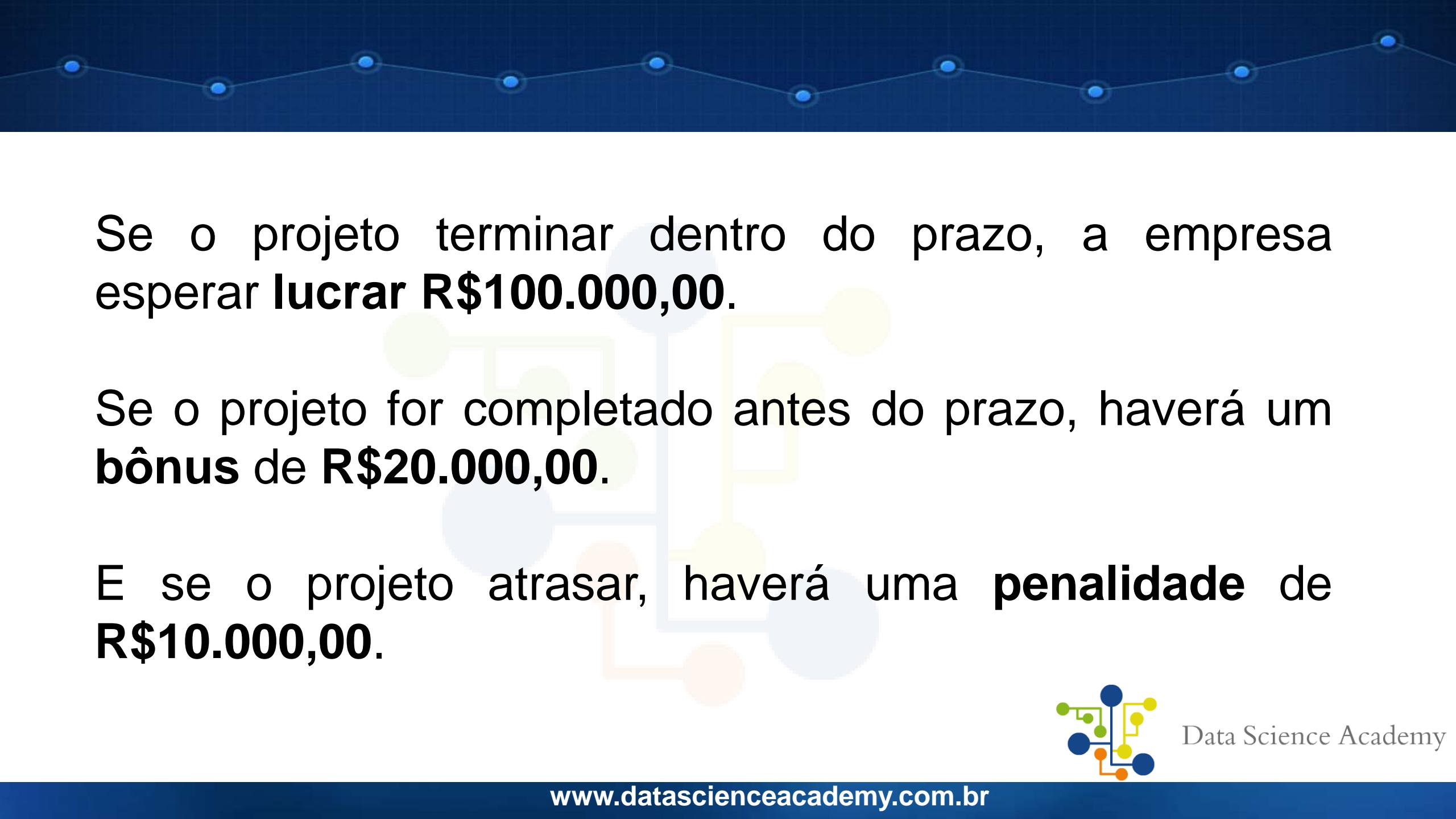


Se o projeto terminar dentro do prazo, a empresa
esperar **lucrar R\$100.000,00**.

Se o projeto for completado antes do prazo, haverá um
bônus de R\$20.000,00.



Data Science Academy



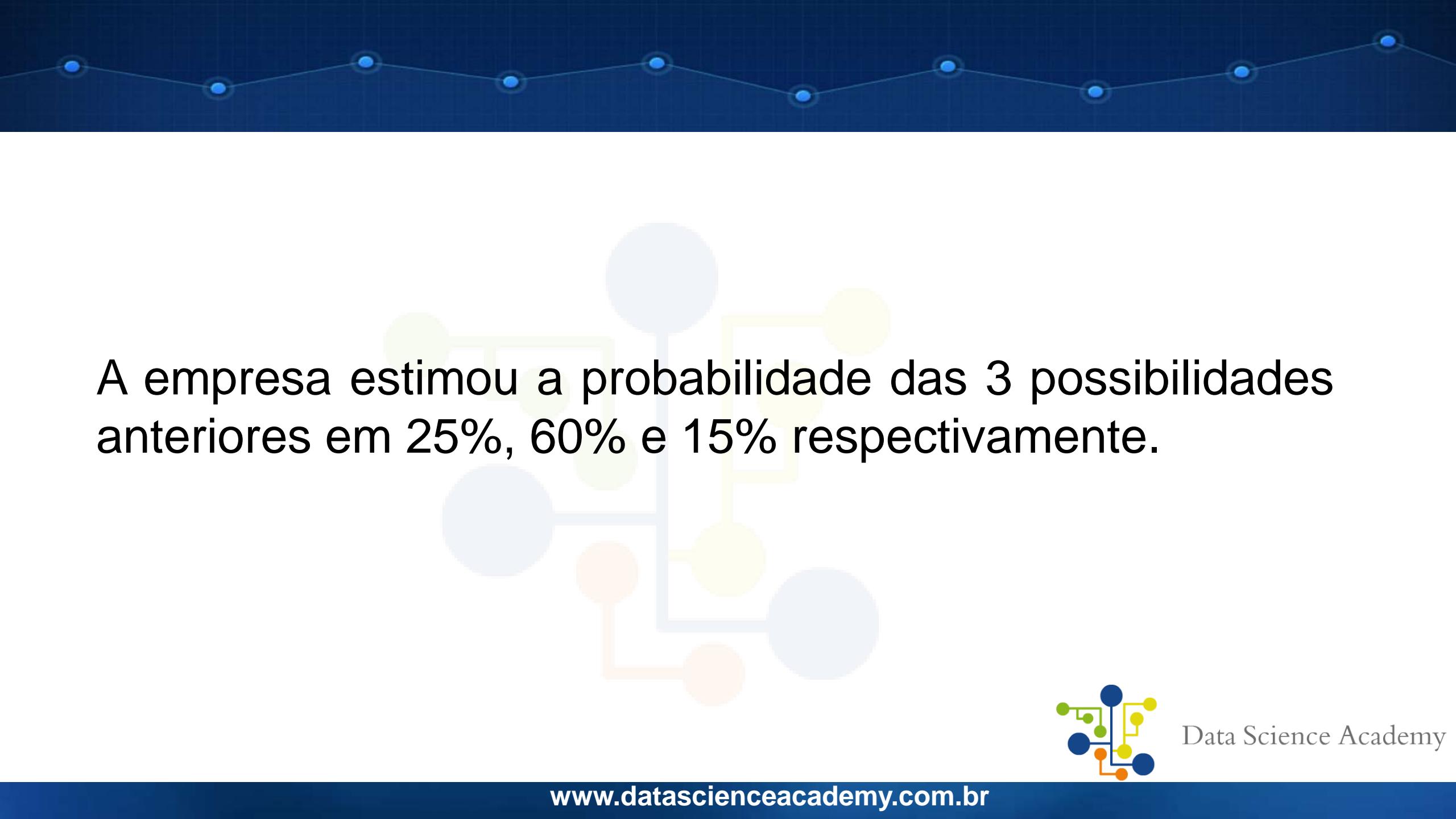
Se o projeto terminar dentro do prazo, a empresa esperar **lucrar R\$100.000,00**.

Se o projeto for completado antes do prazo, haverá um **bônus de R\$20.000,00**.

E se o projeto atrasar, haverá uma **penalidade** de **R\$10.000,00**.



Data Science Academy



A empresa estimou a probabilidade das 3 possibilidades anteriores em 25%, 60% e 15% respectivamente.



Data Science Academy



Vamos tabular isso?



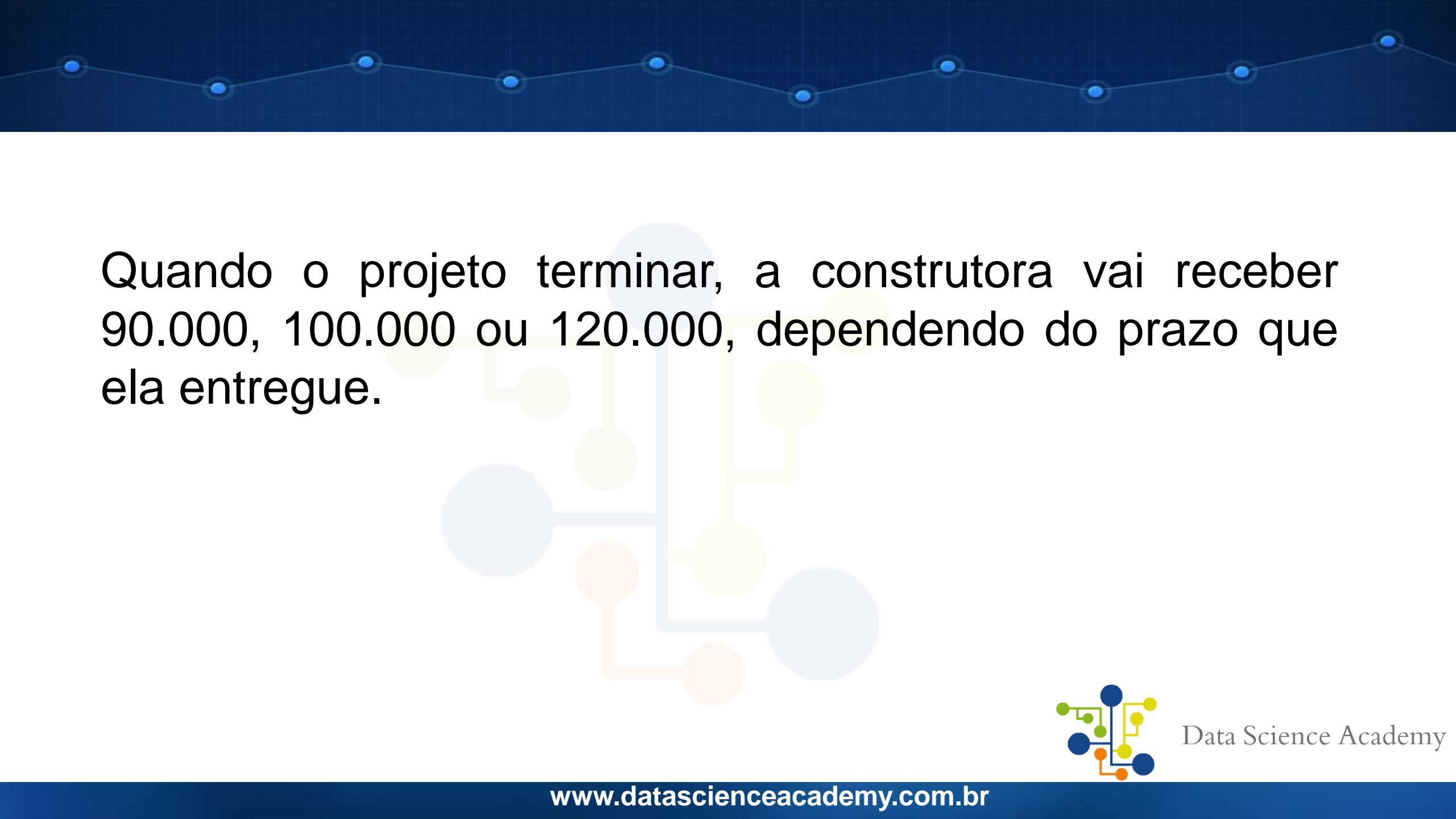
Data Science Academy

Vamos tabular isso?

Término do Projeto	Lucro (x)	Probabilidade P(x)	P(x) x	
Antes do prazo	R\$ 120.000,00	0.25	120.000×0.25	30.000
No prazo	R\$ 100.000,00	0.60	100.000×0.60	60.000
Depois do prazo	R\$ 90.000,00	0.15	90.000×0.15	13.500
VME				R\$ 103.500



Data Science Academy



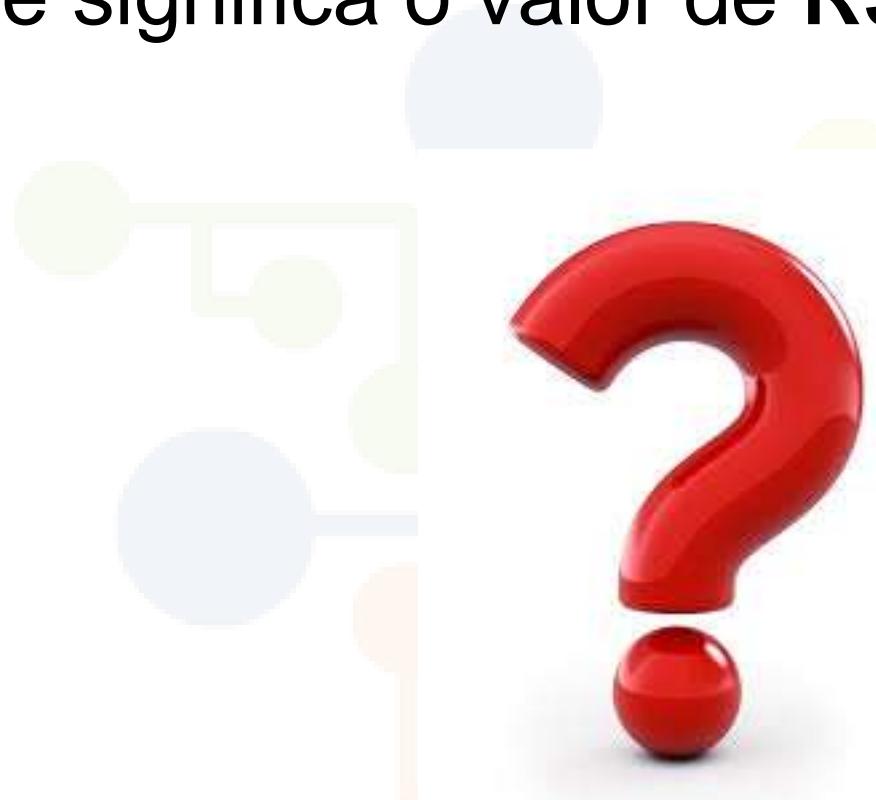
Quando o projeto terminar, a construtora vai receber 90.000, 100.000 ou 120.000, dependendo do prazo que ela entregue.



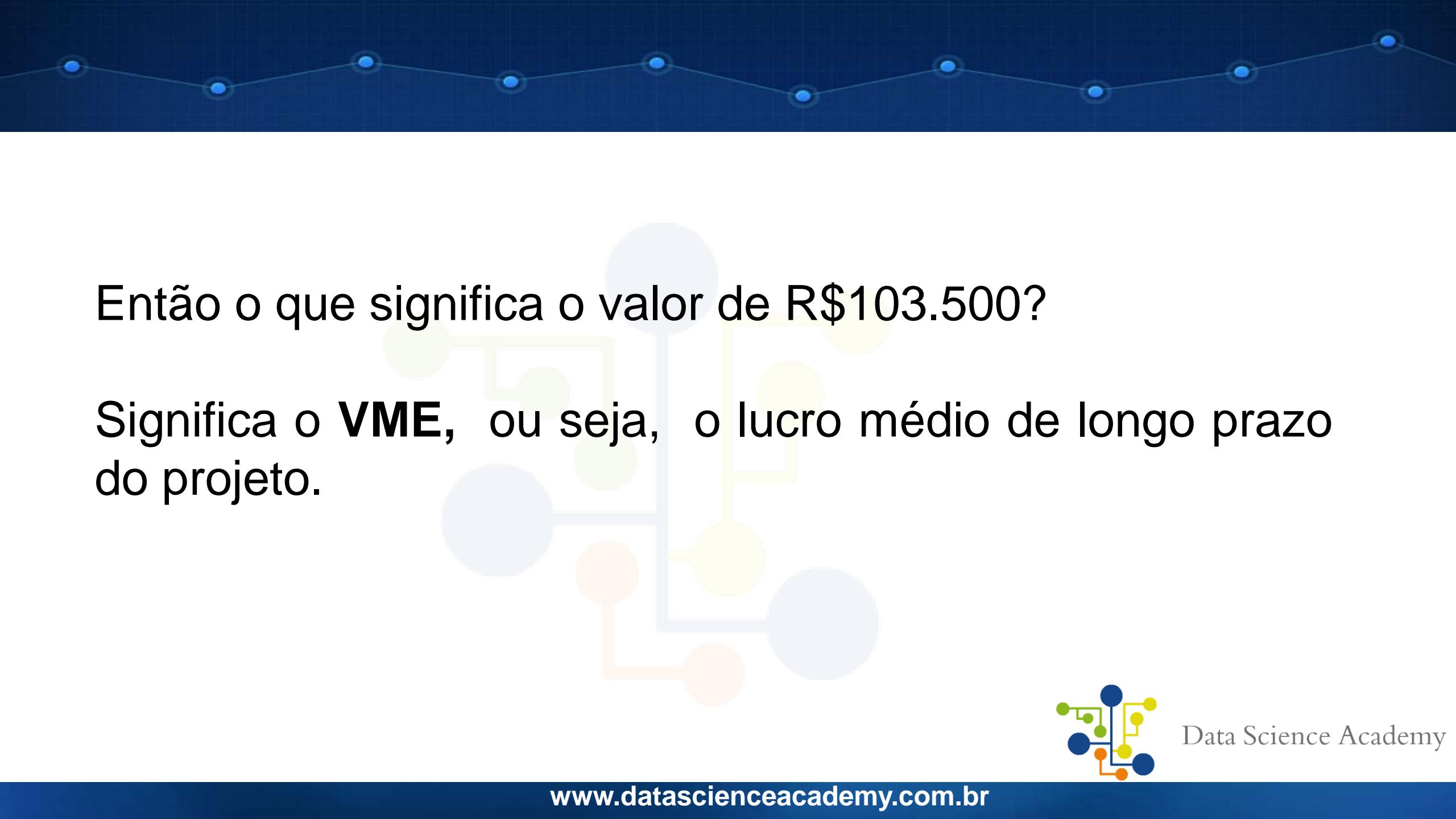
Data Science Academy



Então o que significa o valor de R\$103.500?



Data Science Academy

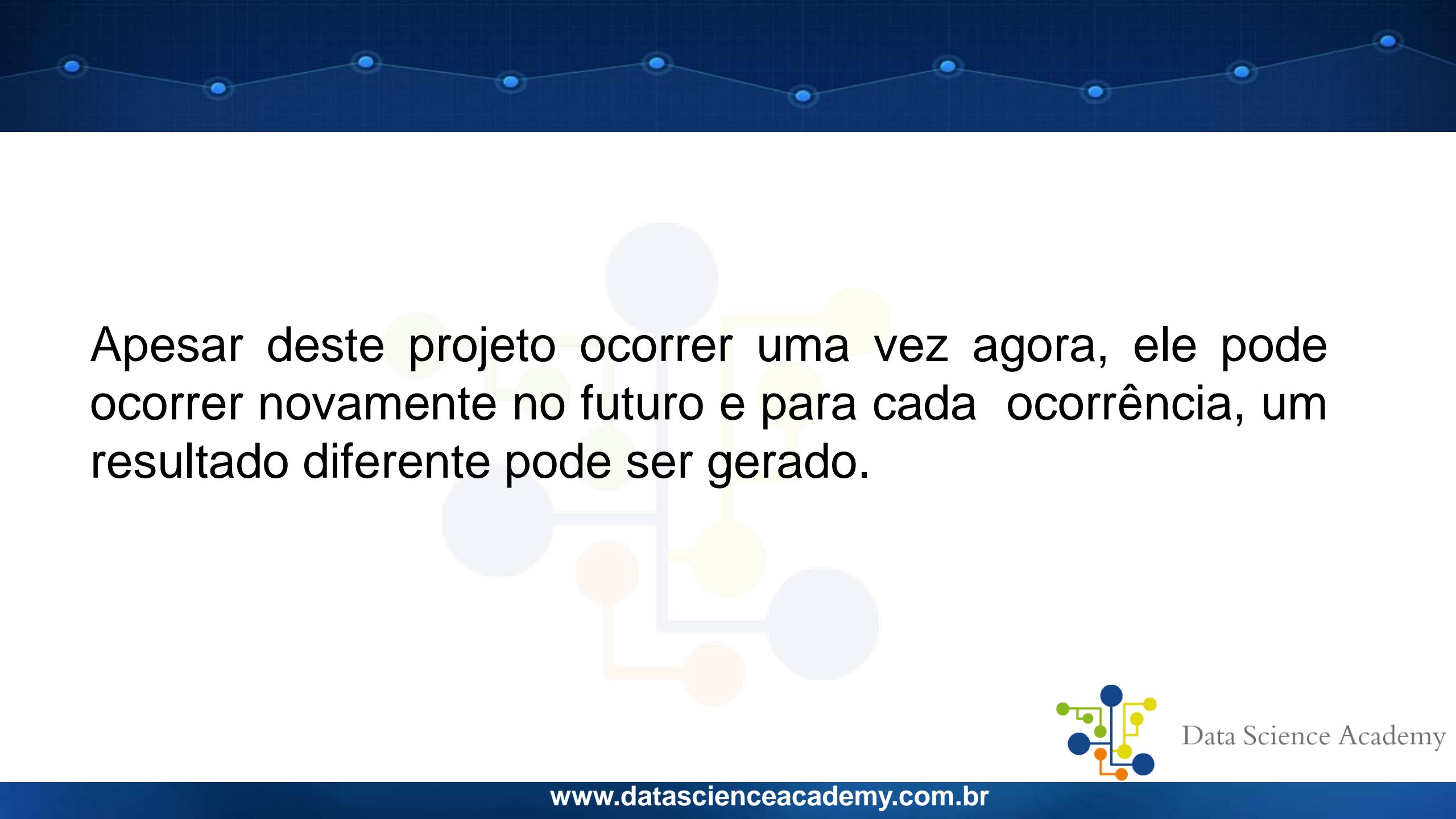


Então o que significa o valor de R\$103.500?

Significa o **VME**, ou seja, o lucro médio de longo prazo do projeto.



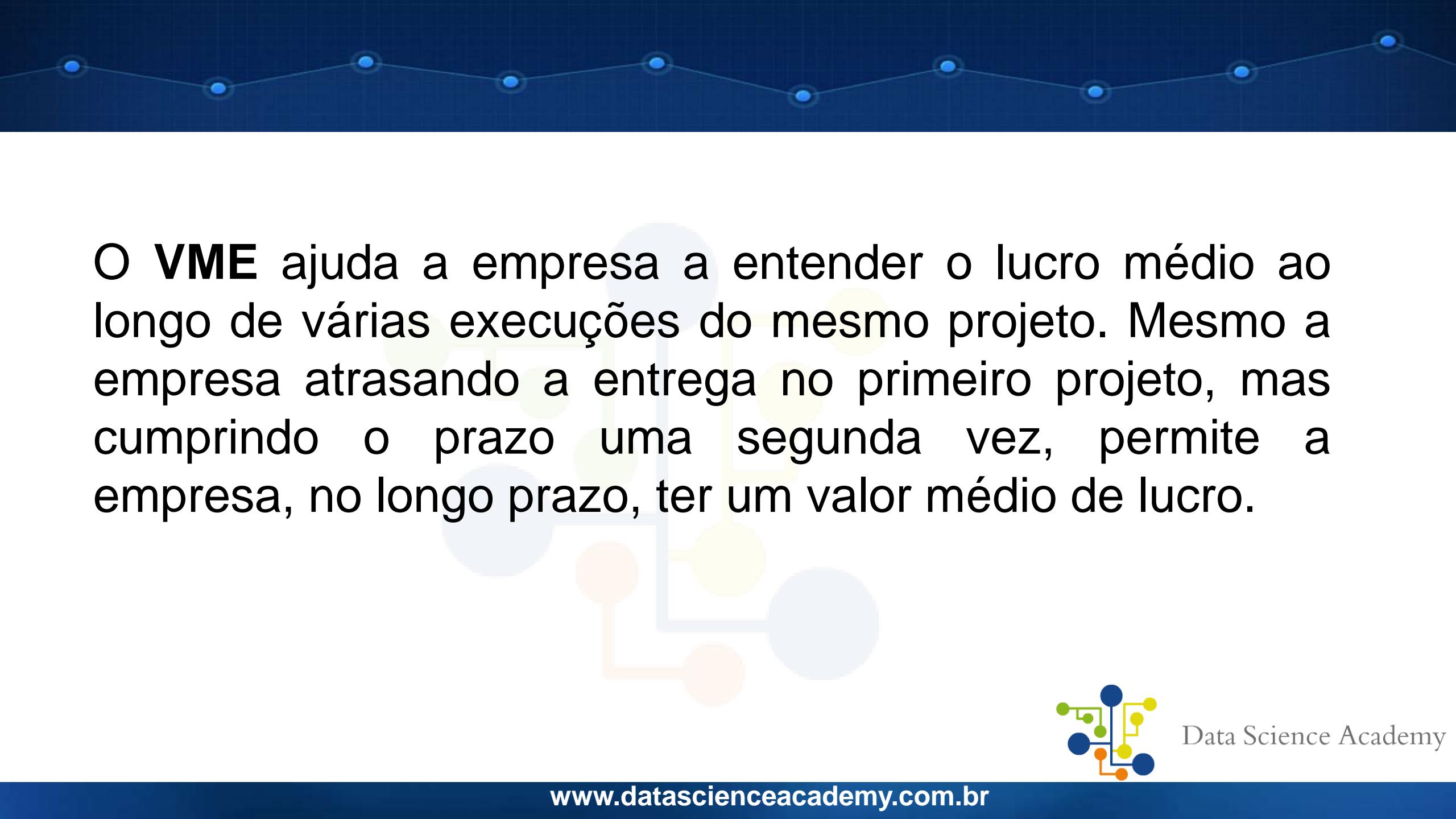
Data Science Academy



Apesar deste projeto ocorrer uma vez agora, ele pode ocorrer novamente no futuro e para cada ocorrência, um resultado diferente pode ser gerado.



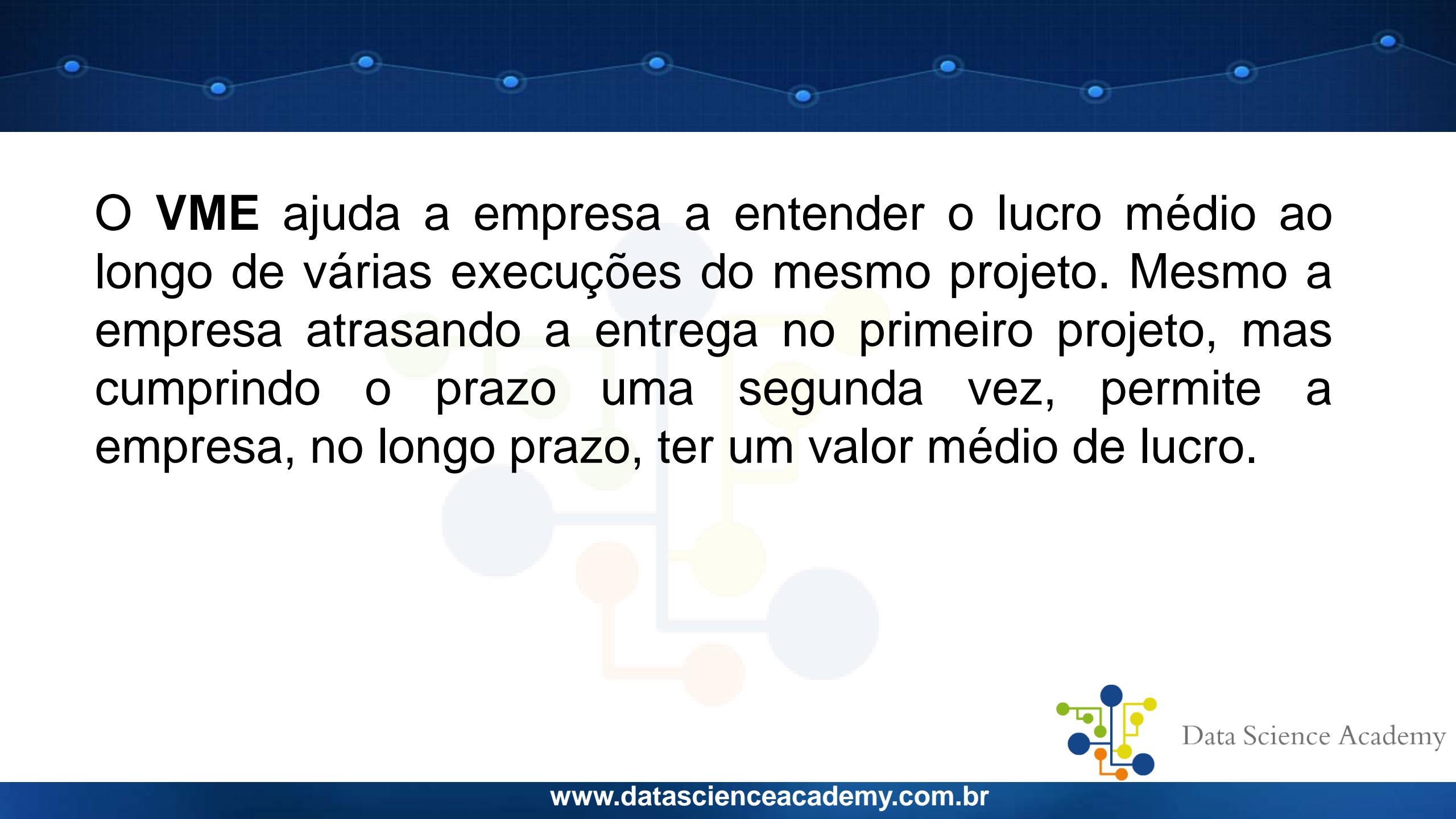
Data Science Academy



O **VME** ajuda a empresa a entender o lucro médio ao longo de várias execuções do mesmo projeto. Mesmo a empresa atrasando a entrega no primeiro projeto, mas cumprindo o prazo uma segunda vez, permite a empresa, no longo prazo, ter um valor médio de lucro.



Data Science Academy



O **VME** ajuda a empresa a entender o lucro médio ao longo de várias execuções do mesmo projeto. Mesmo a empresa atrasando a entrega no primeiro projeto, mas cumprindo o prazo uma segunda vez, permite a empresa, no longo prazo, ter um valor médio de lucro.



Data Science Academy

Esse tópico chegou ao final



Data Science Academy

Introdução à Amostragem



Data Science Academy

Amostragem



Data Science Academy

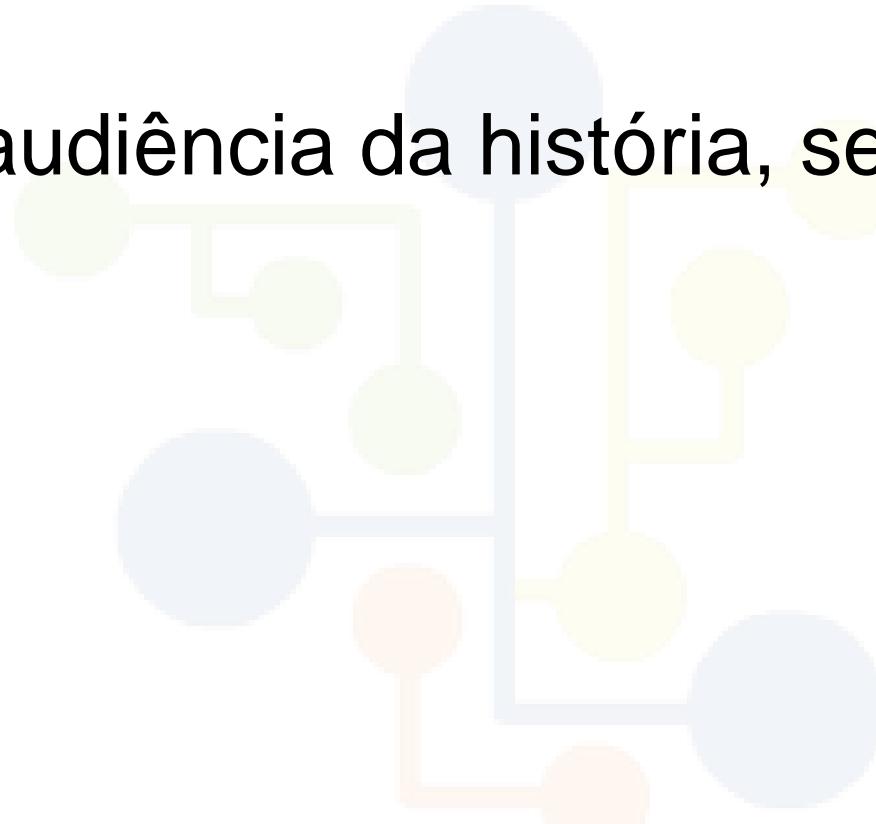
Segundo as TVs americanas, a final da Liga NFL (a liga de futebol Americano) de 2014, foi assistida por **111 milhões** de pessoas.



Data Science Academy



Foi a maior audiência da história, segundo os dirigentes.



Data Science Academy

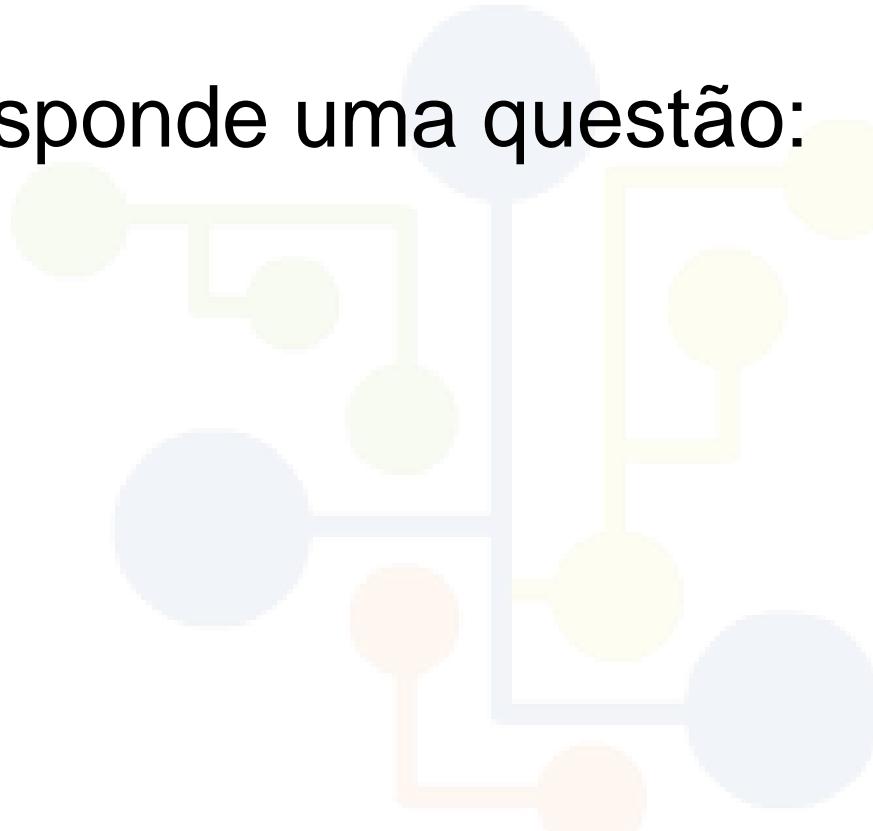
O recorde anterior tinha ocorrido em 2010, com 106,5 milhões de telespectadores.



Data Science Academy



Agora me responde uma questão:



Data Science Academy

Como eles conseguiram contar esse número de pessoas?



Data Science Academy

Para poder responder a esta pergunta, precisaremos de conhecimentos em **Amostragem Estatística**, que aprenderemos neste módulo.



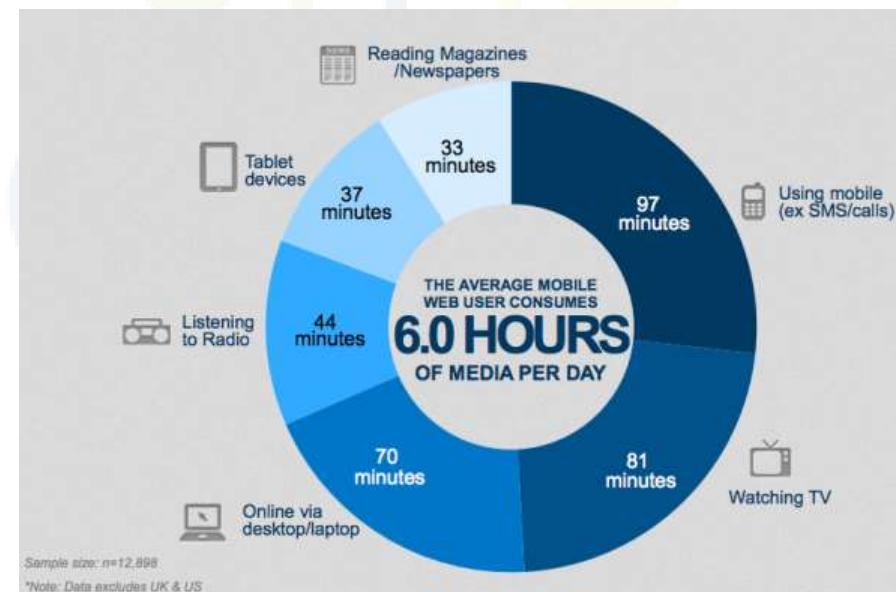
Data Science Academy

Mas, e a resposta?



Data Science Academy

A empresa responsável por gerar este número nos EUA é a empresa que fornece pesquisas estatísticas para redes de canais a cabo e empresas que desejam fazer propagandas nos canais de TV americanos.



Data Science Academy

Para isso, a empresa faz sua pesquisa com **5.000** famílias, dentro do estimado número de **129** milhões de famílias nos EUA (a população americana é de **321** milhões de pessoas).



Data Science Academy

A empresa fornece uma pequena caixa (chamada “medidor de audiência”) para estas 5.000 famílias. Esta caixa é um receptor acoplado à TV (ou TV's, caso exista mais de um televisor na casa), que registra tudo que acontece com a TV: quando ela é ligada, desligada, quanto tempo ela fica ligada, os canais mais acessados, quanto tempo em cada canal, se a pessoa muda de canal durante os comerciais, etc.



Data Science Academy

Essas **5.000 famílias** são cuidadosamente selecionadas,
pois elas devem representar toda a população dos EUA.



Data Science Academy

Baseada no resultado da pesquisa com estas **5.000** famílias, a empresa consegue estimar com precisão os hábitos televisivos de toda a população americana e estimar quantas pessoas assistiram à final da liga NFL, a liga esportiva mais popular entre os americanos.



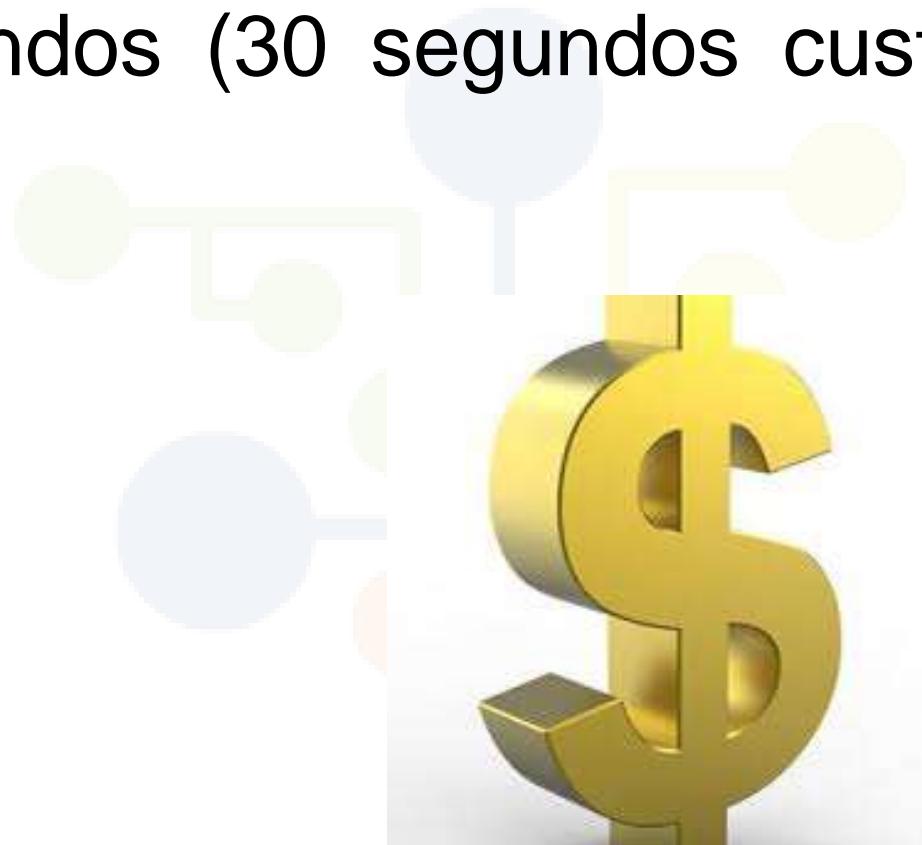
Data Science Academy

Com estes números em mãos, as redes de TV geram o orçamento do valor de cada propaganda durante o intervalo dos jogos.



Data Science Academy

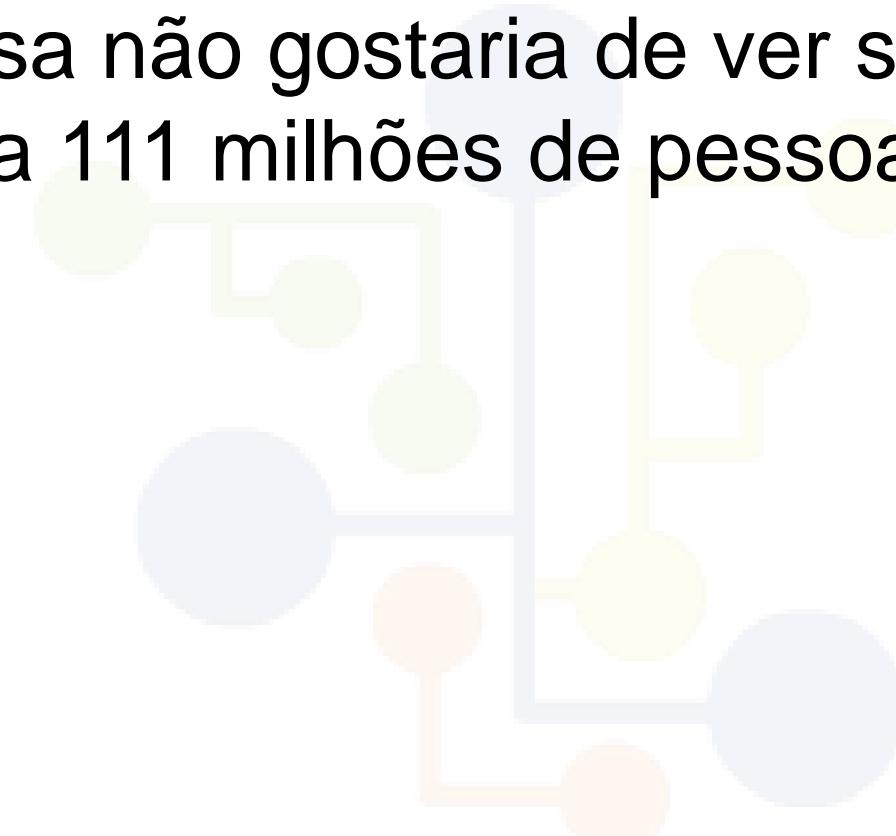
Quanto maior a audiência, maior o valor de uma incursão de 30 segundos (30 segundos custaram 3 milhões de dólares).



Data Science Academy



Qual empresa não gostaria de ver seu produto ou serviço
exposto para 111 milhões de pessoas de uma única vez?



Data Science Academy



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

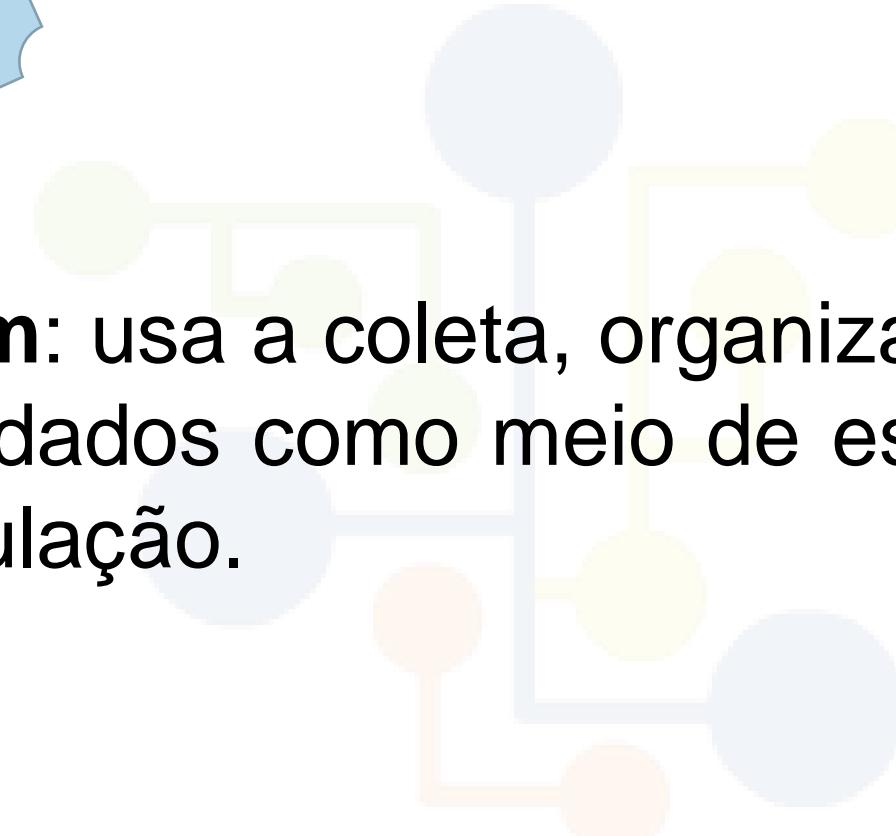


Data Science Academy



Método
Estatístico

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



Data Science Academy



Amostragem: é a técnica que seleciona apenas **alguns** elementos da população quando se realiza uma pesquisa em uma amostra.



Data Science Academy



Diferenças
na
Abordagem

Amostragem: é a técnica que seleciona apenas **alguns** elementos da população quando se realiza uma pesquisa em uma amostra.

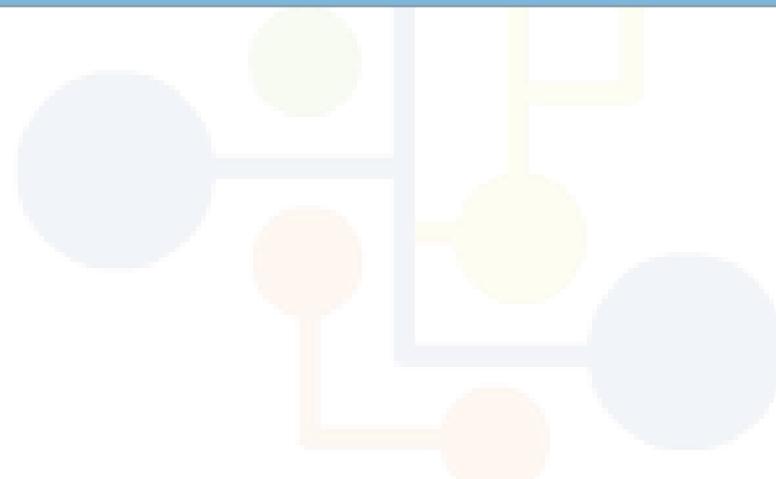
Censo: é a técnica que seleciona e avalia **todos** os elementos da população quando se realiza uma pesquisa.



Data Science Academy



O que é Amostra?



Data Science Academy

Amostra



Data Science Academy

Por que não medir uma **população inteira**, ao invés de medir apenas uma **amostra**?



Science Academy

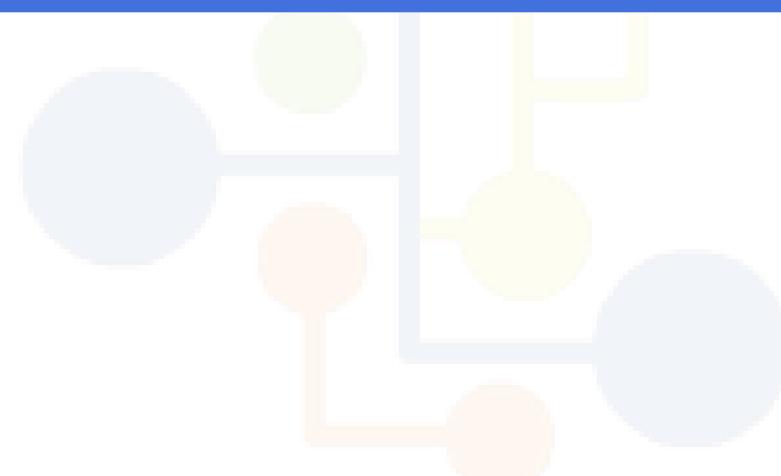


Dependendo das circunstâncias, medir uma população inteira seria **caro demais** ou até mesmo inviável.

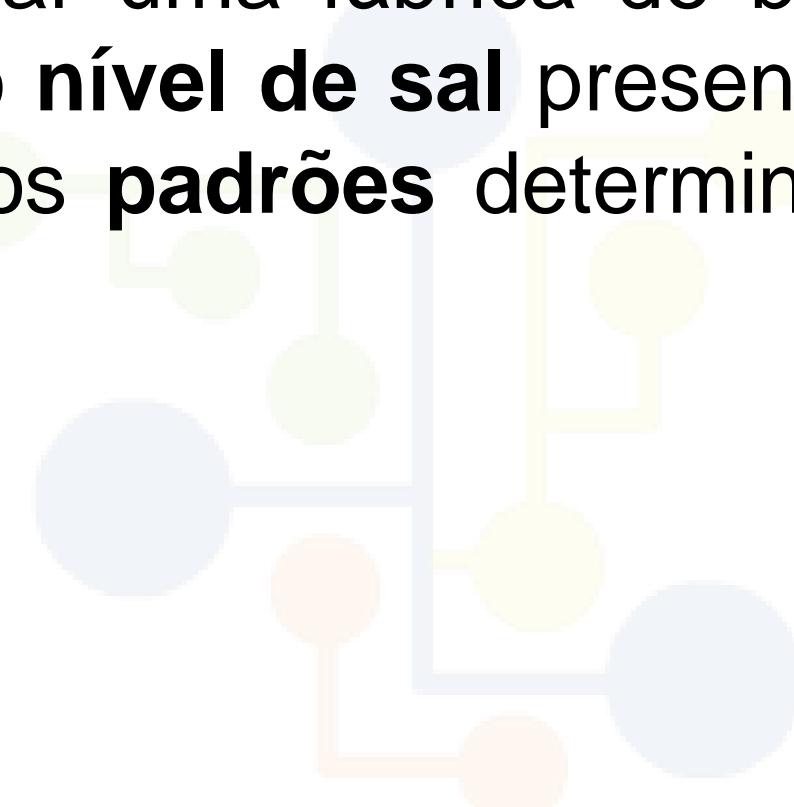


Data Science Academy

Exemplo



Data Science Academy



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**.



Data Science Academy

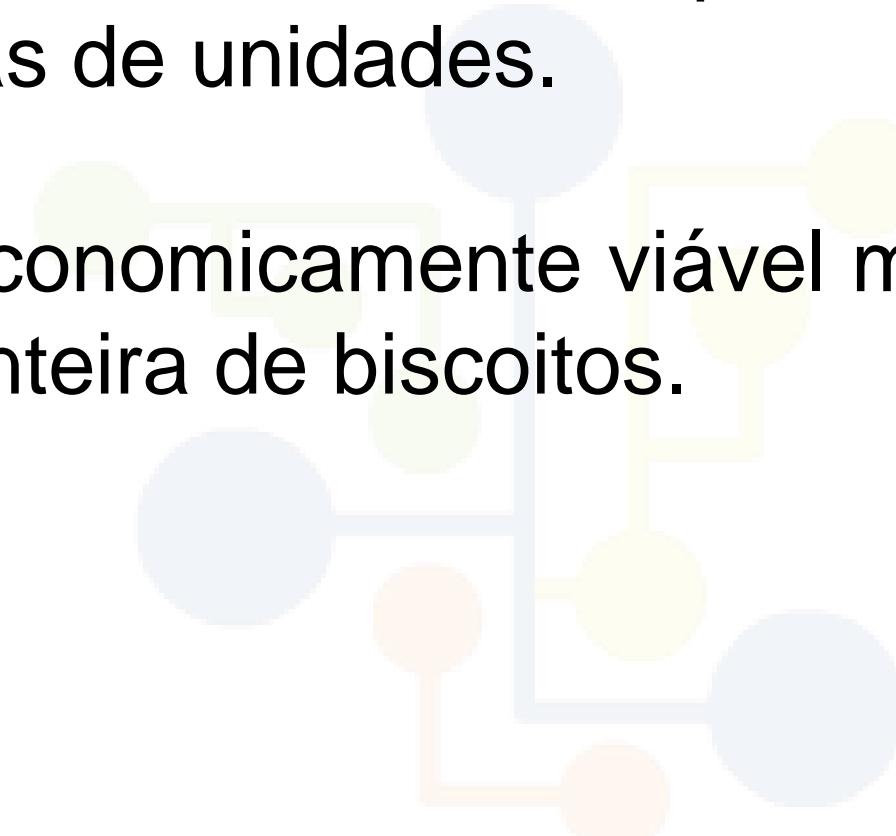
Você acha que seria viável, a empresa medir cada unidade de biscoito produzida?



Data Science Academy



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.



Data Science Academy

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



Data Science Academy

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

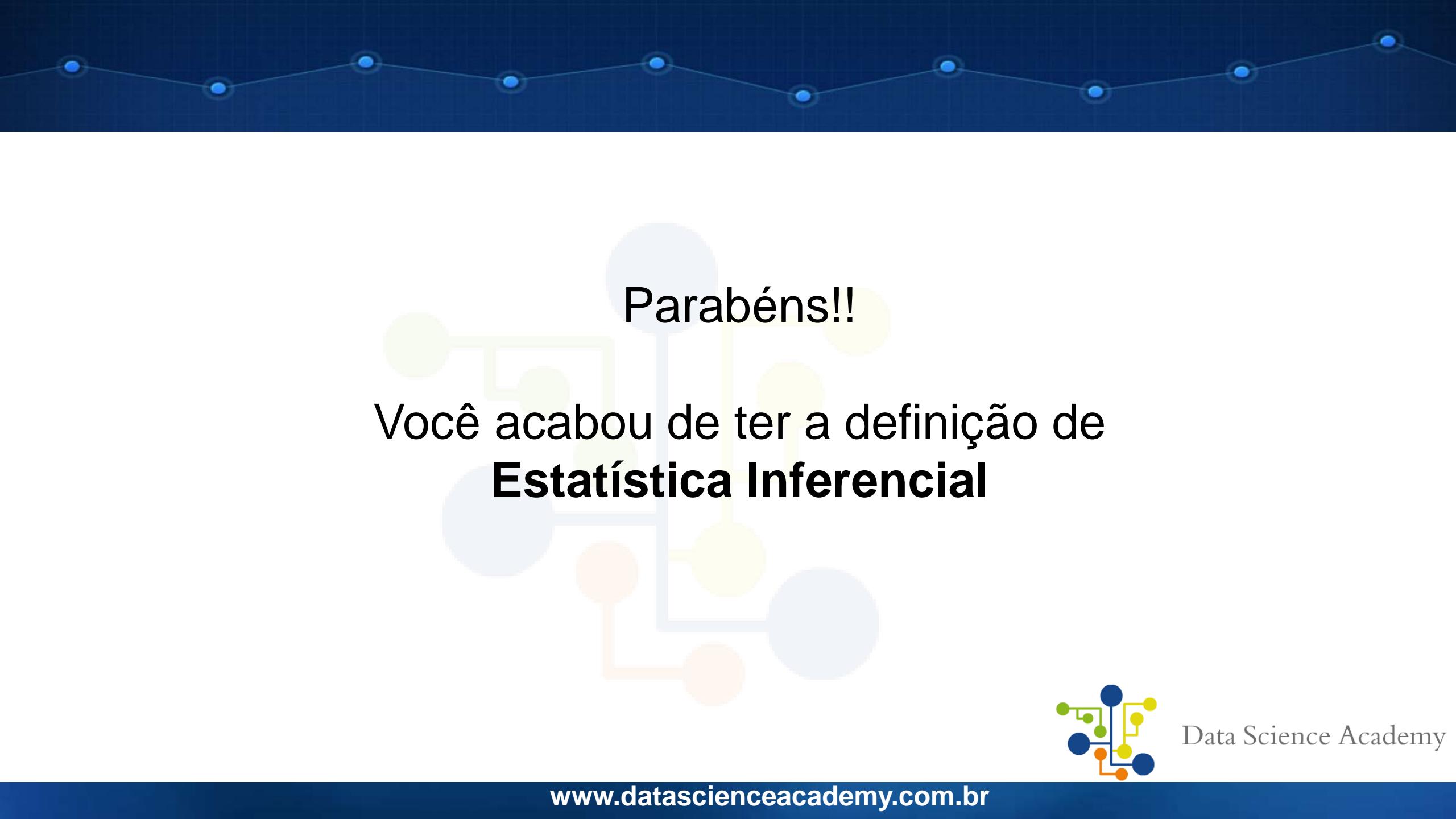


Data Science Academy

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



Data Science Academy

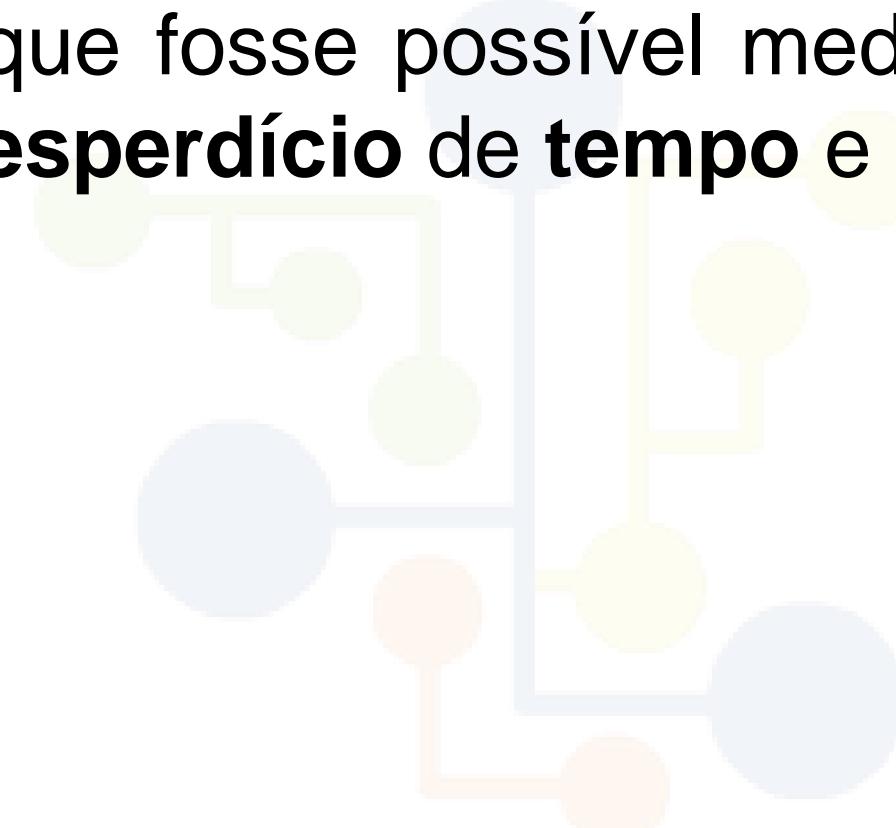


Parabéns!!

Você acabou de ter a definição de
Estatística Inferencial



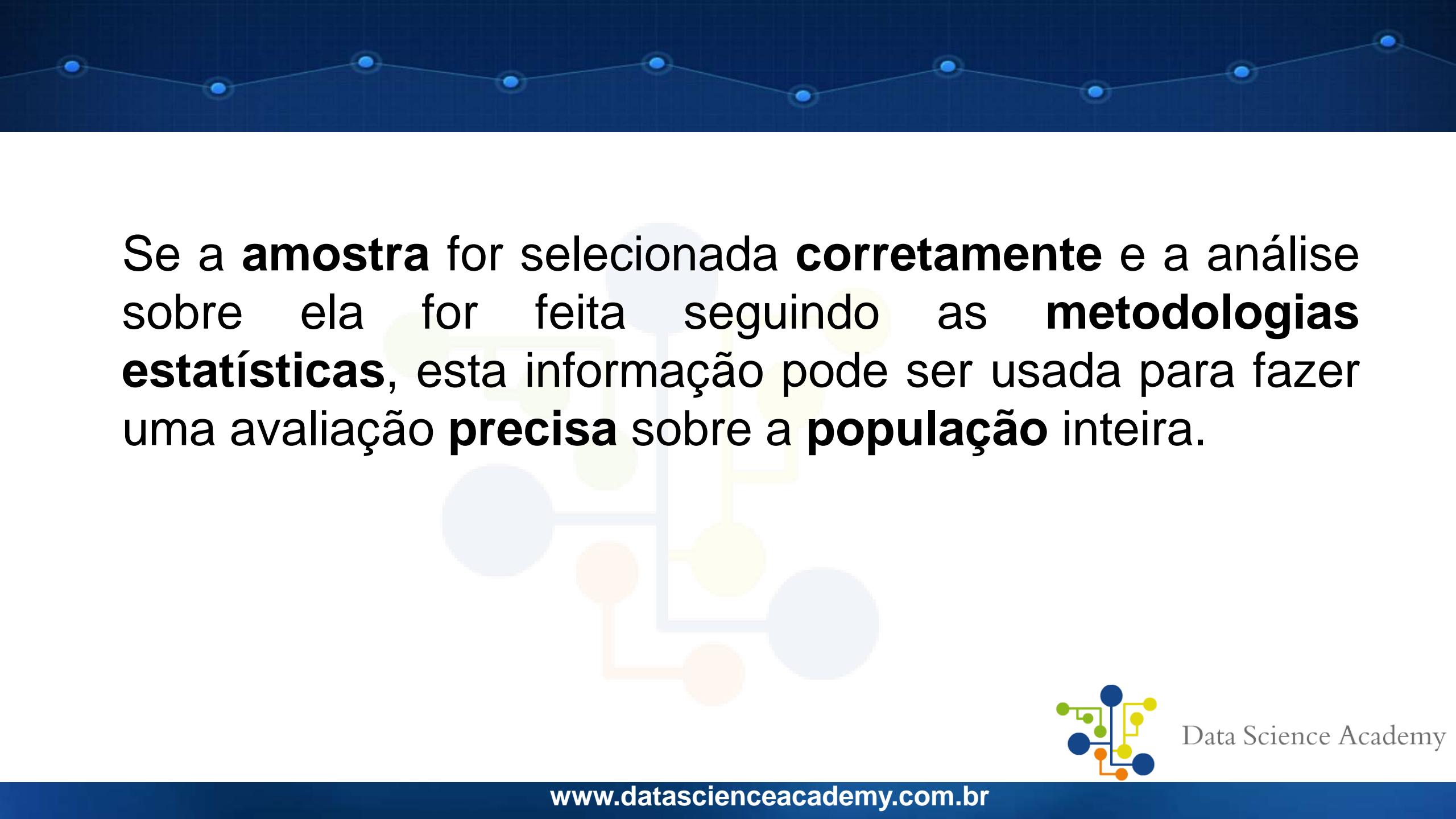
Data Science Academy



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



Data Science Academy



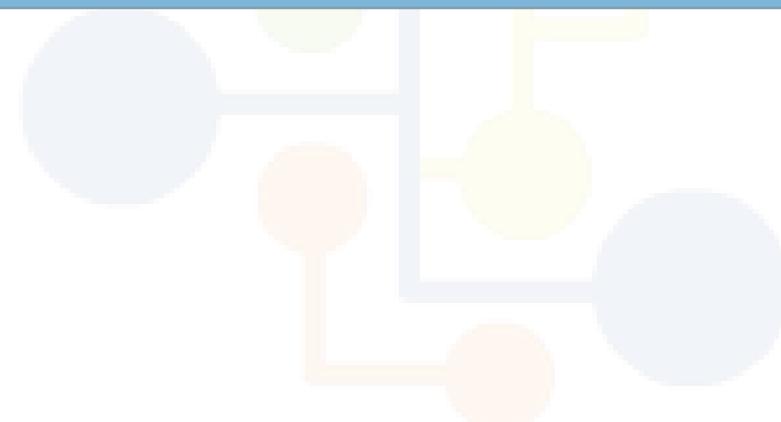
Se a **amostra** for selecionada **corretamente** e a análise sobre ela for feita seguindo as **metodologias estatísticas**, esta informação pode ser usada para fazer uma avaliação **precisa** sobre a **população** inteira.



Data Science Academy

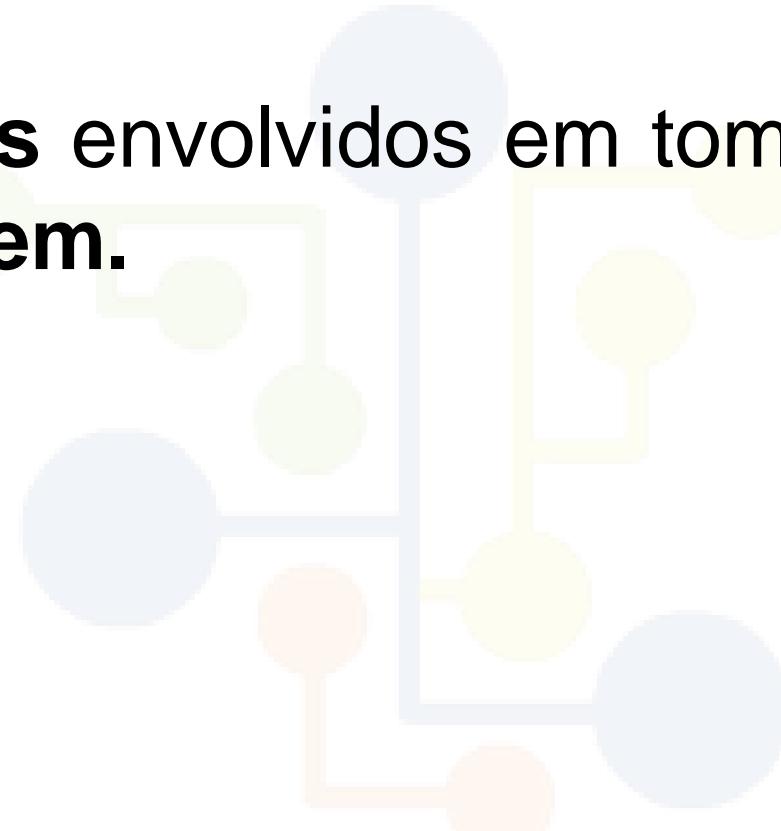


Entretanto,

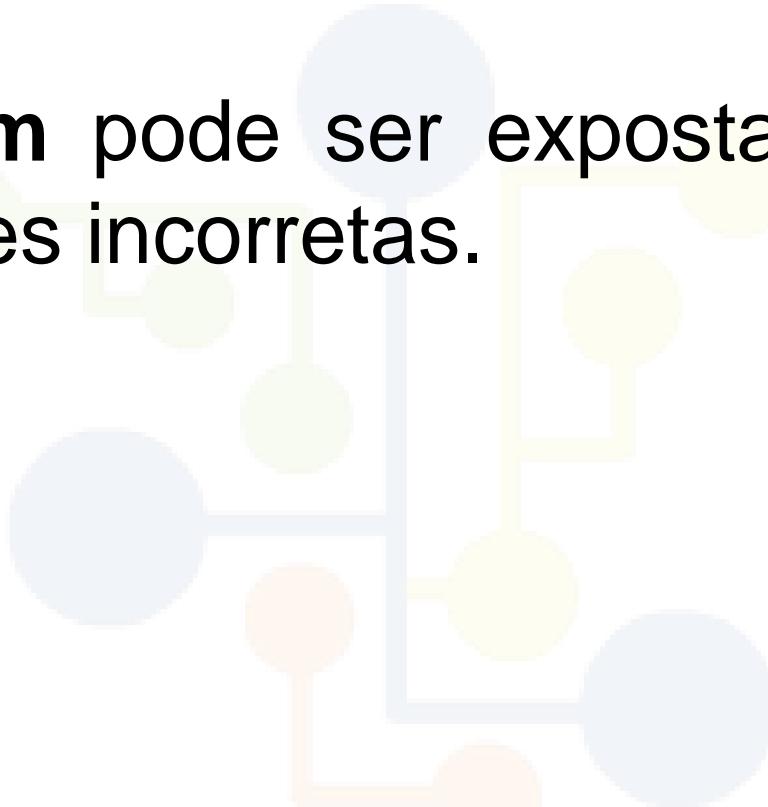


Data Science Academy

Existem **riscos** envolvidos em tomar decisões baseadas em **amostragem**.



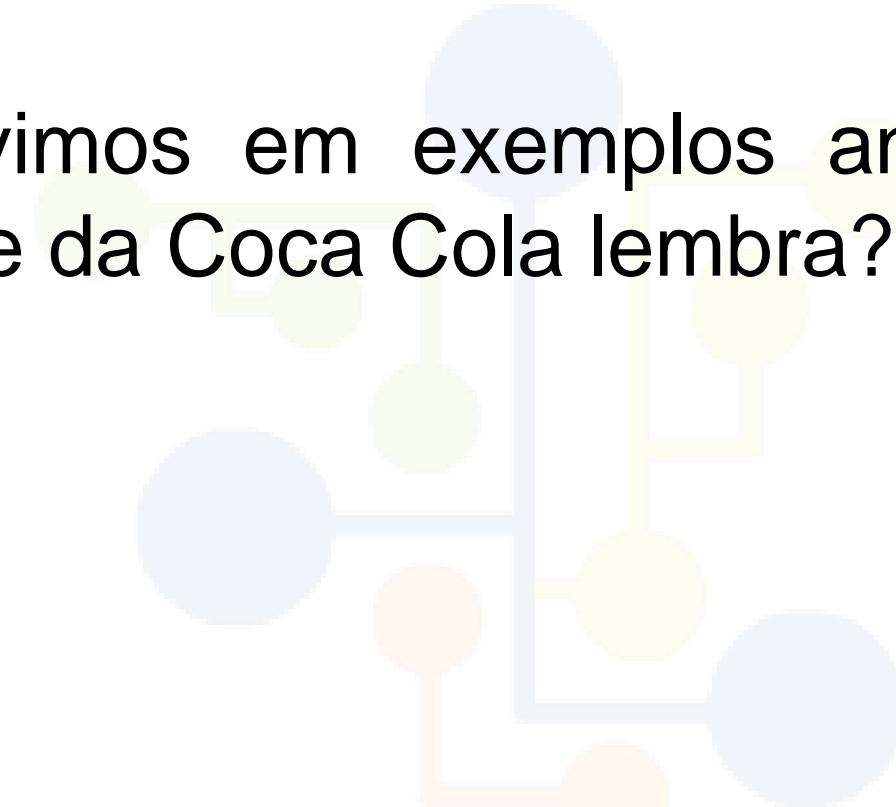
Data Science Academy



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



Data Science Academy

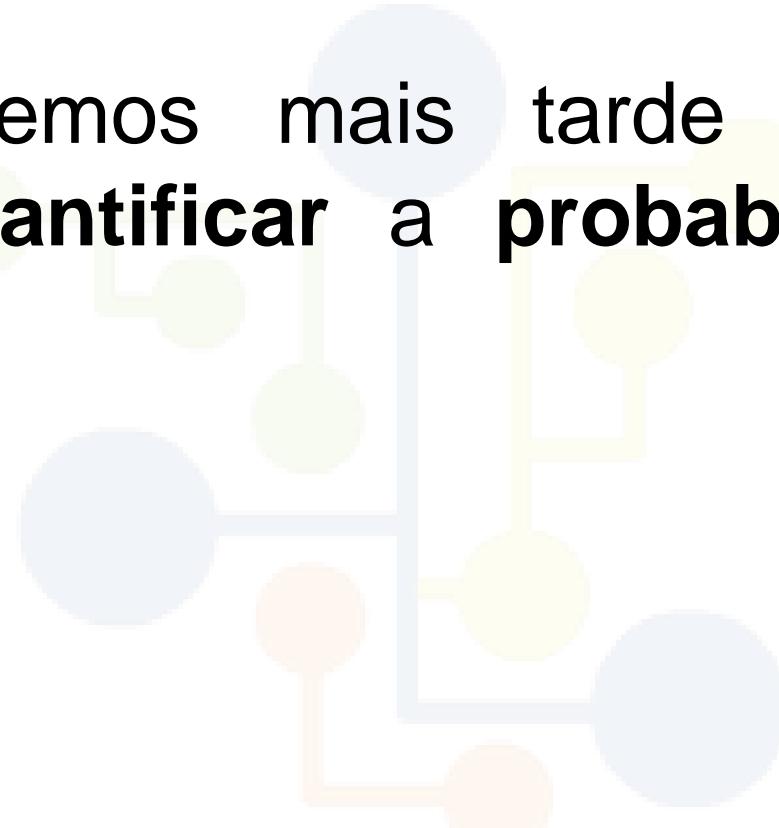


Como já vimos em exemplos anteriores, no caso do Chery Coke da Coca Cola lembra?



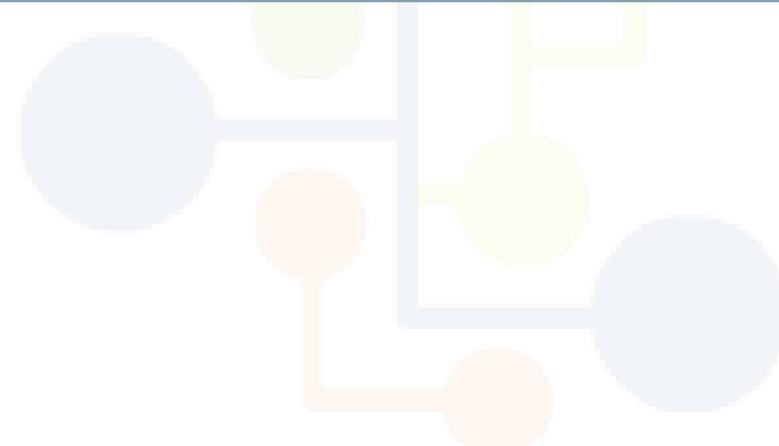
Data Science Academy

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



Data Science Academy

Tipos de Amostragem



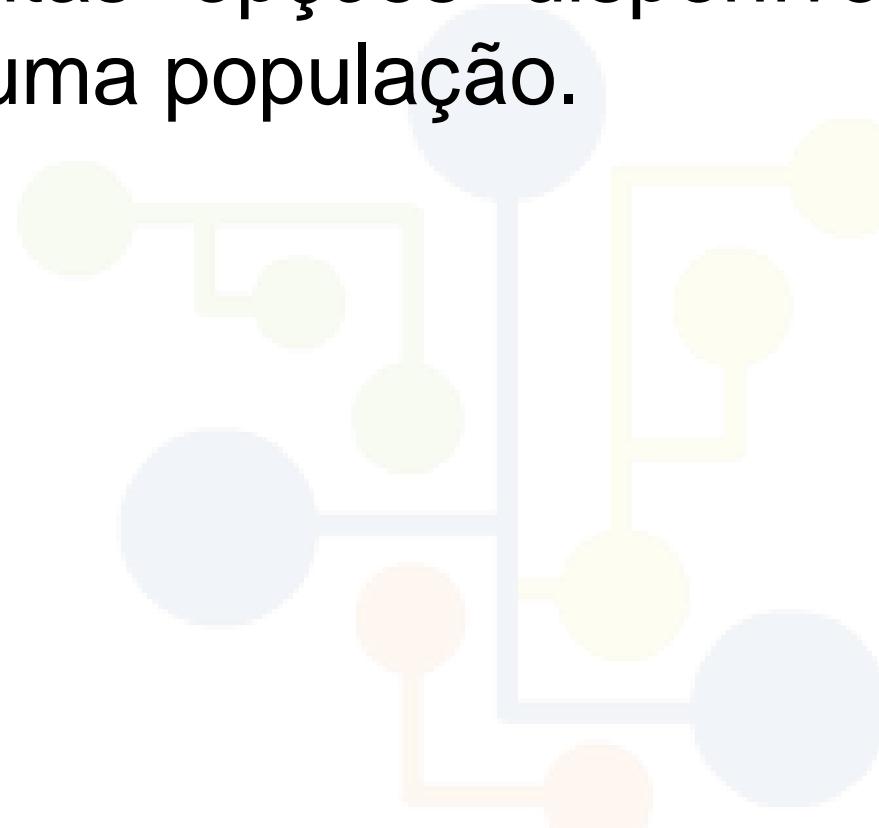
Data Science Academy

Tipos de Amostragem



Data Science Academy

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



Data Science Academy

Os tipos básicos que estudaremos aqui são:

**Amostragem
probabilística.**



Amostragem não-probabilística.



Data Science Academy

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.



Data Science Academy

Amostragem Probabilística

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



Data Science Academy

Na Amostragem Probabilística

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



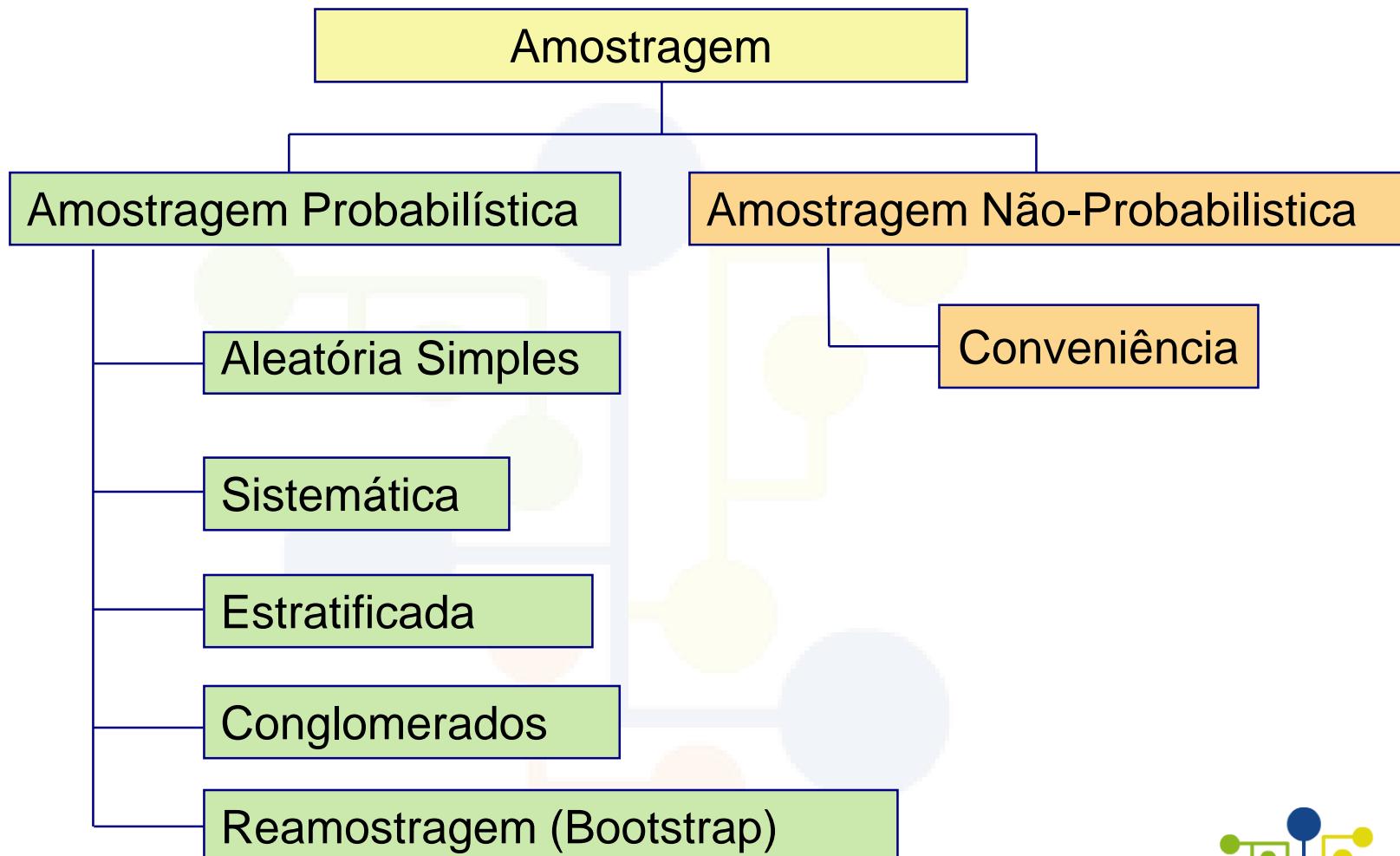
Data Science Academy

Na 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.



Data Science Academy



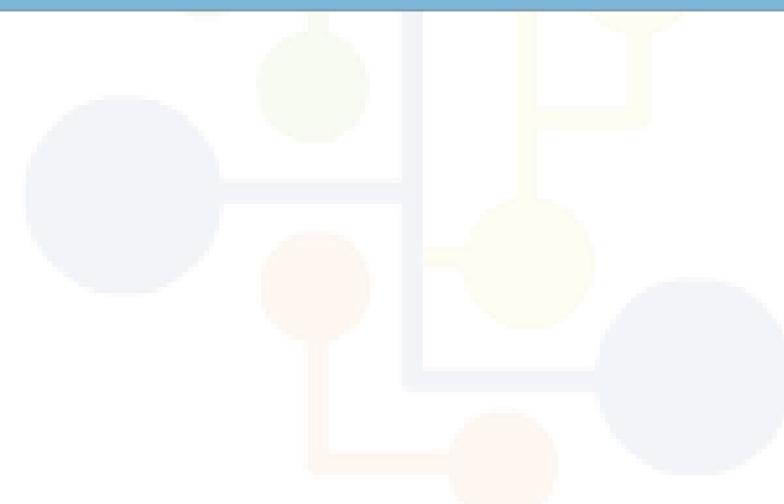
Data Science Academy

Esse tópico chegou ao final



Data Science Academy

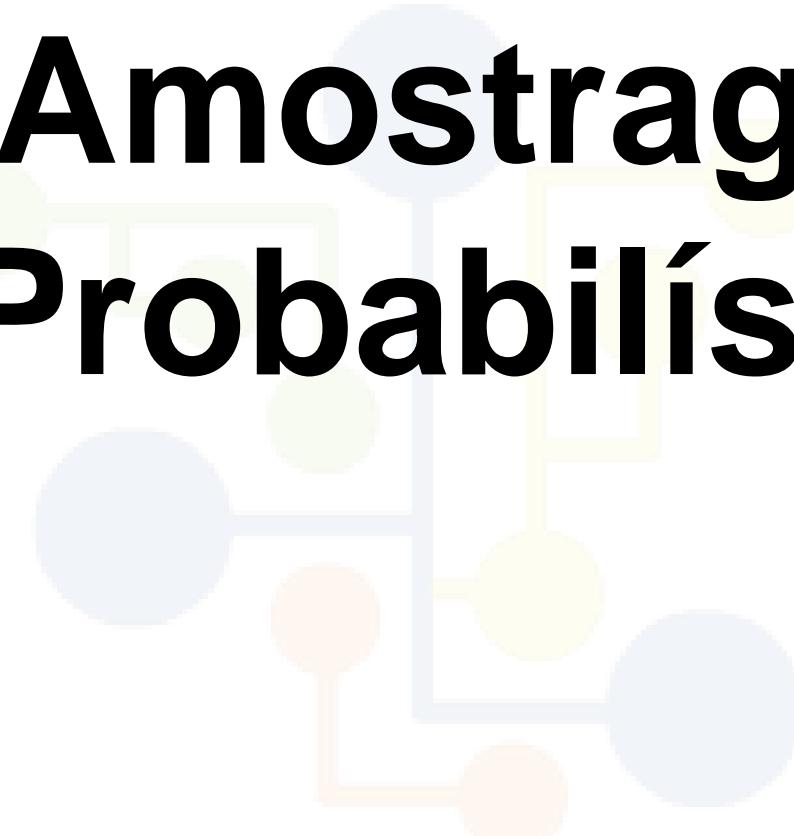
Amostragem Probabilística



Data Science Academy

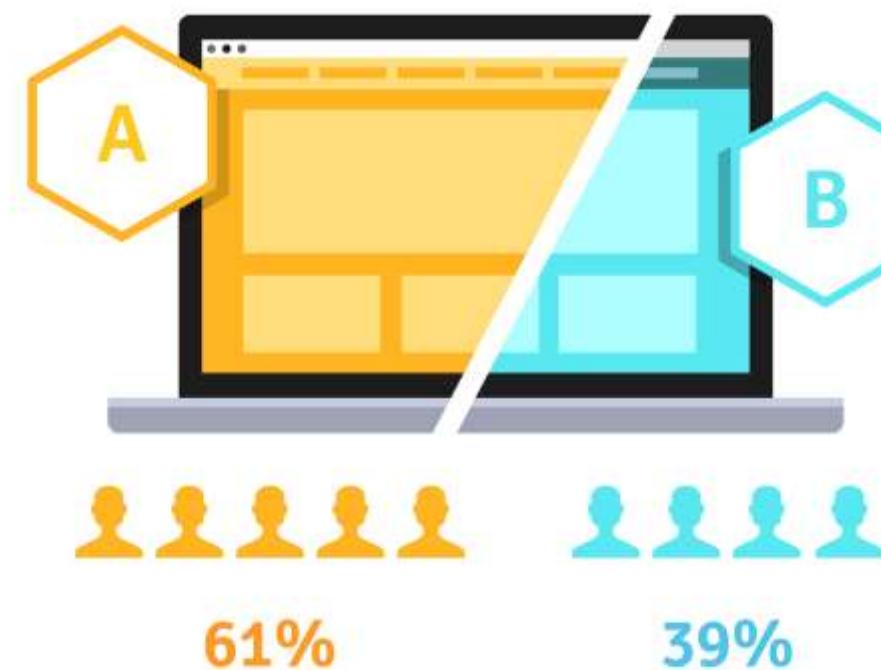


Amostragem Probabilística

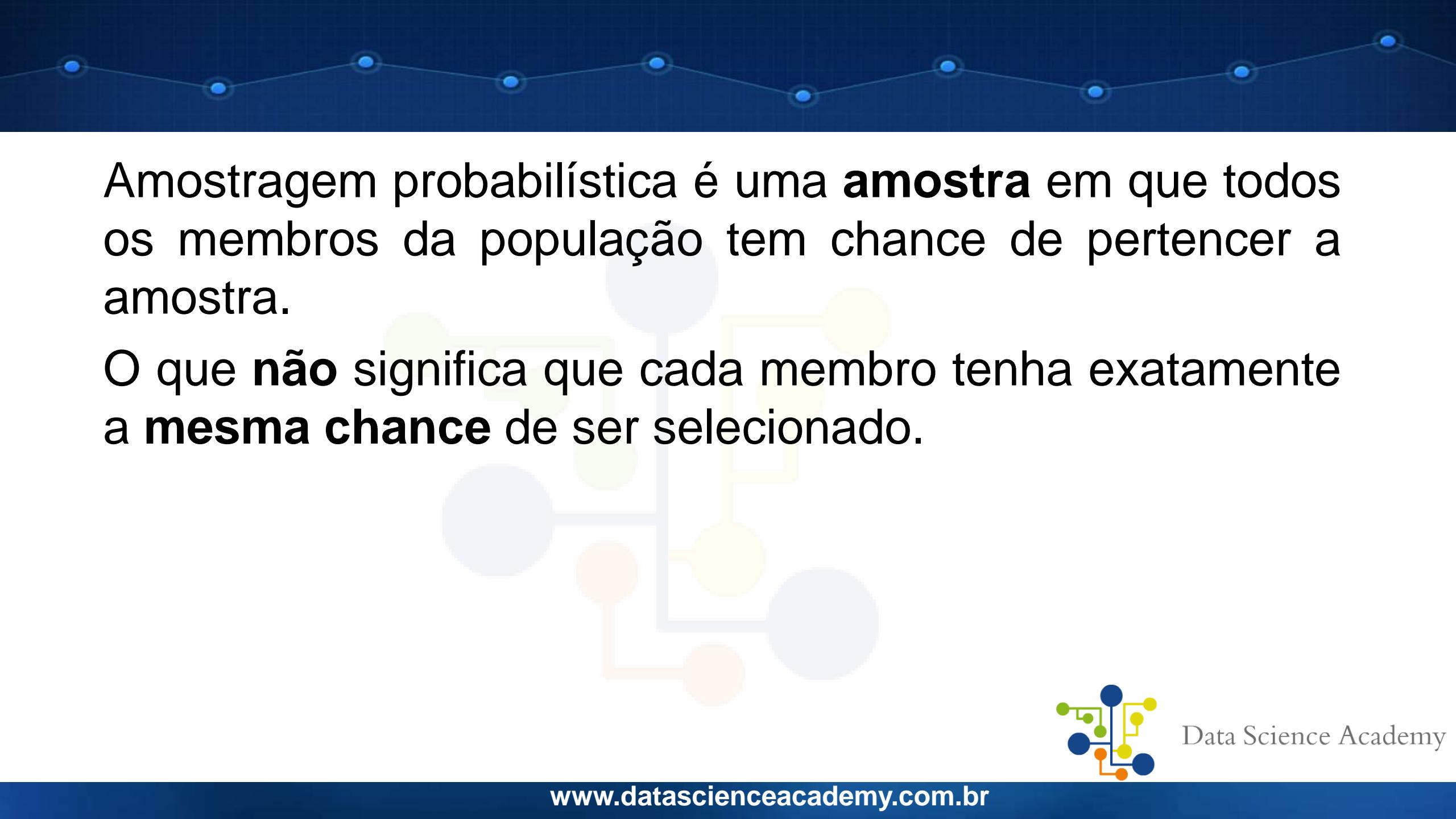


Data Science Academy

Amostragem probabilística é uma **amostra** em que todos os membros da população tem chance de pertencer a amostra.



Data Science Academy



Amostragem probabilística é uma **amostra** em que todos os membros da população tem chance de pertencer a amostra.

O que **não** significa que cada membro tenha exatamente a **mesma chance** de ser selecionado.



Data Science Academy

Com **Amostragem Probabilística**, nós temos a possibilidade de realizar uma variedade de testes de **estatística inferencial**, que nos permitirá extrair **conclusões confiáveis** sobre a **população**.



Data Science Academy

Amostragem Probabilística

Aleatória Simples

Sistemática

Estratificada

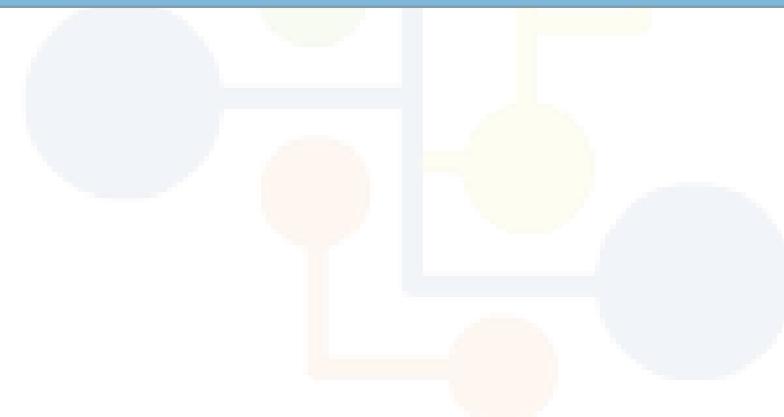
Conglomerados

Reamostragem/
Bootstrap



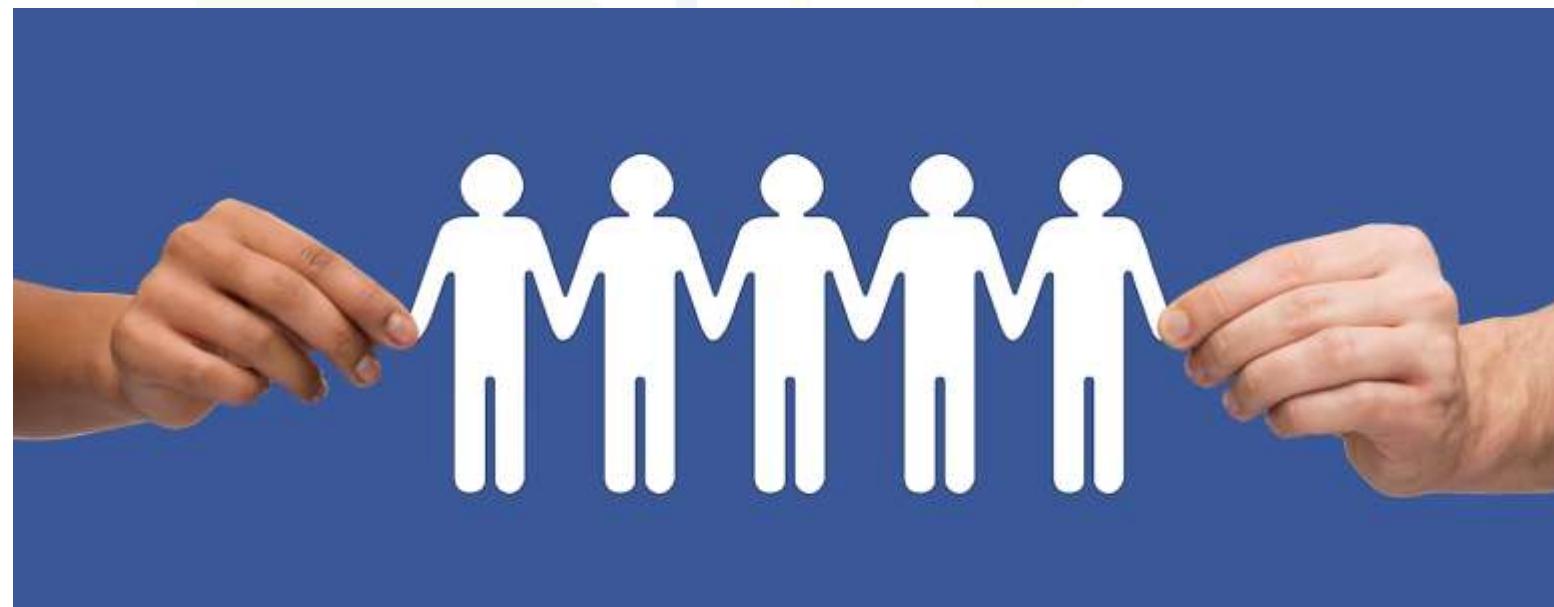
Data Science Academy

Amostragem Aleatória Simples



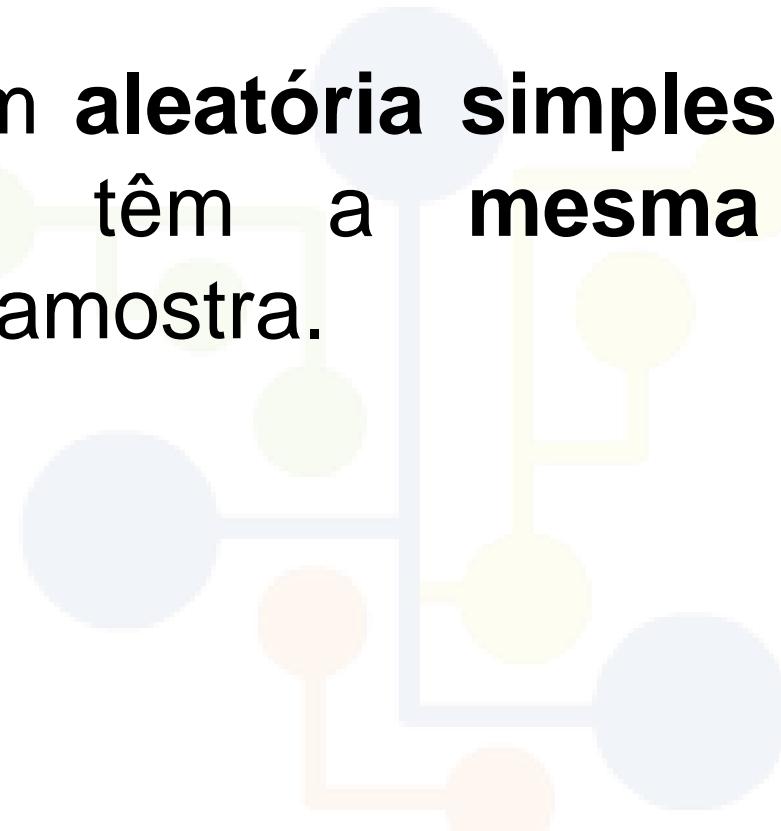
Data Science Academy

Uma **amostra aleatória simples** é a amostra em que cada membro de uma população tem **igual** chance de ser selecionado.



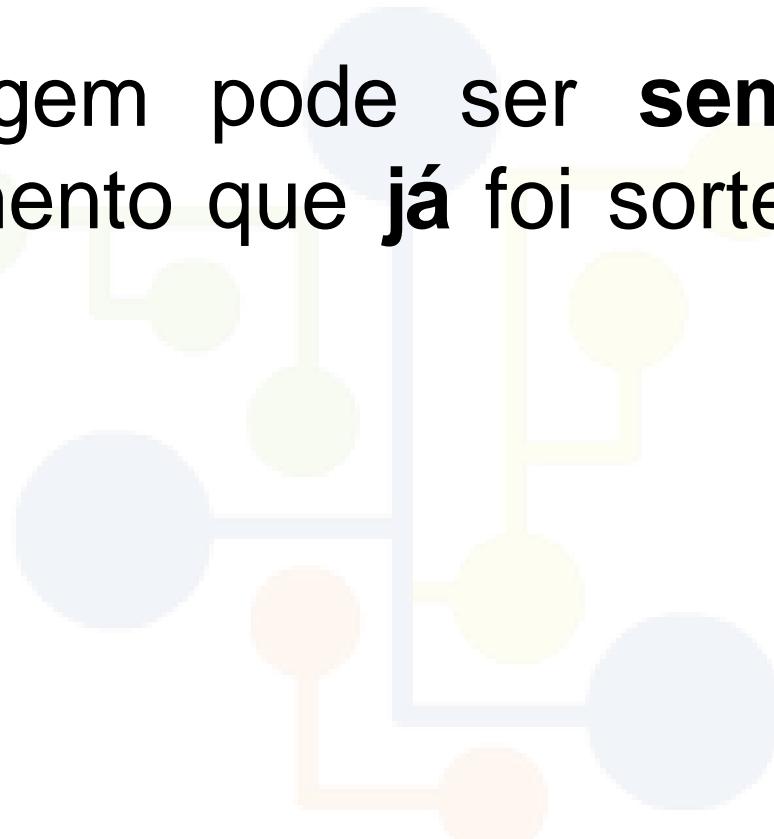
Data Science Academy

Na amostragem **aleatória simples**, todos os elementos da população têm a **mesma probabilidade** de pertencerem à amostra.



Data Science Academy

Essa amostragem pode ser **sem reposição**, que é quando o elemento que já foi sorteado **não** continua no sorteio.

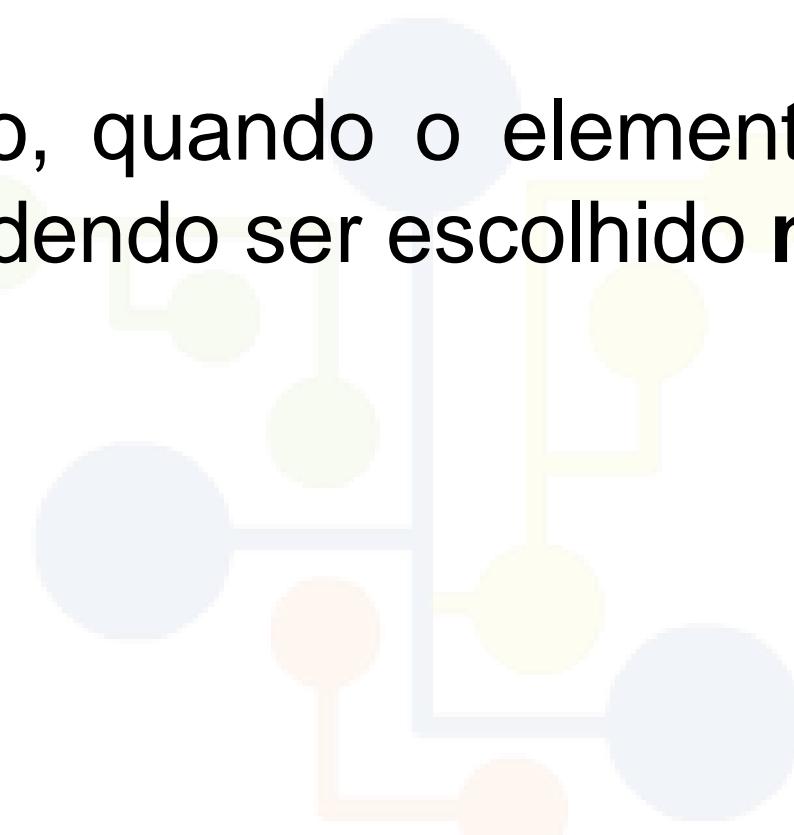


Data Science Academy

OU



Data Science Academy



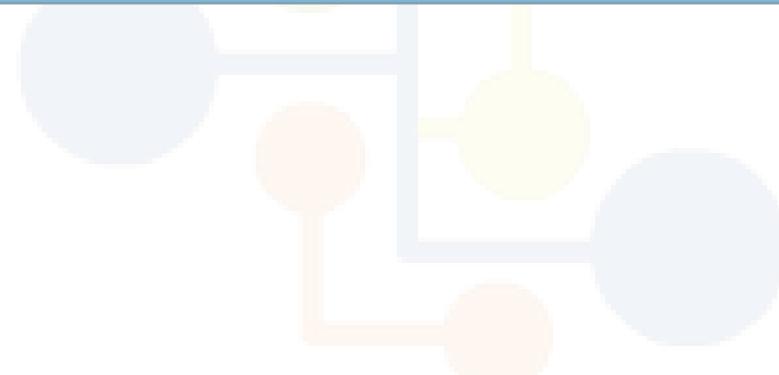
Com reposição, quando o elemento sorteado **continua no sorteio**, podendo ser escolhido **novamente**.



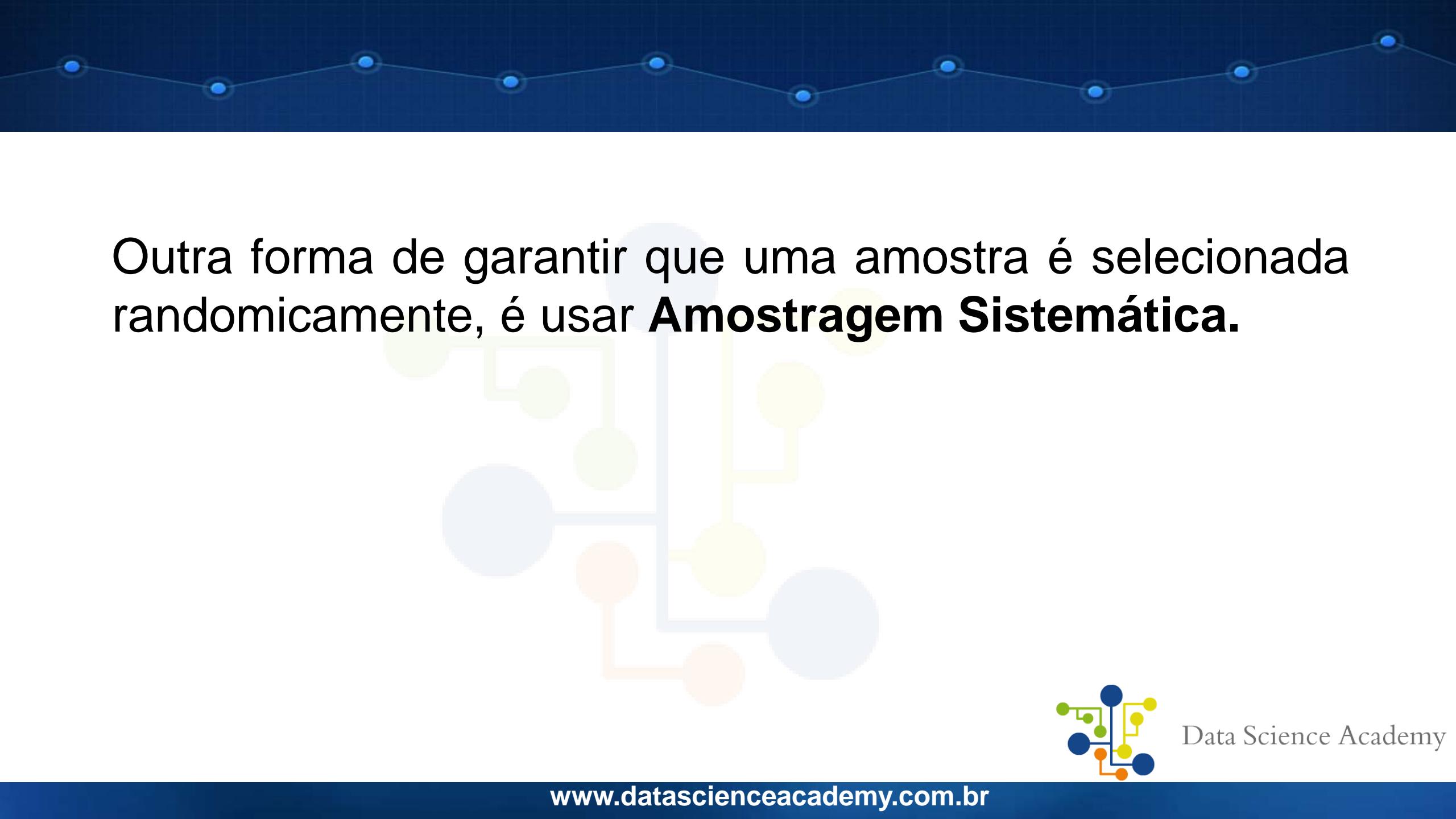
Data Science Academy



Amostragem Sistemática



Data Science Academy



Outra forma de garantir que uma amostra é selecionada randomicamente, é usar **Amostragem Sistemática**.

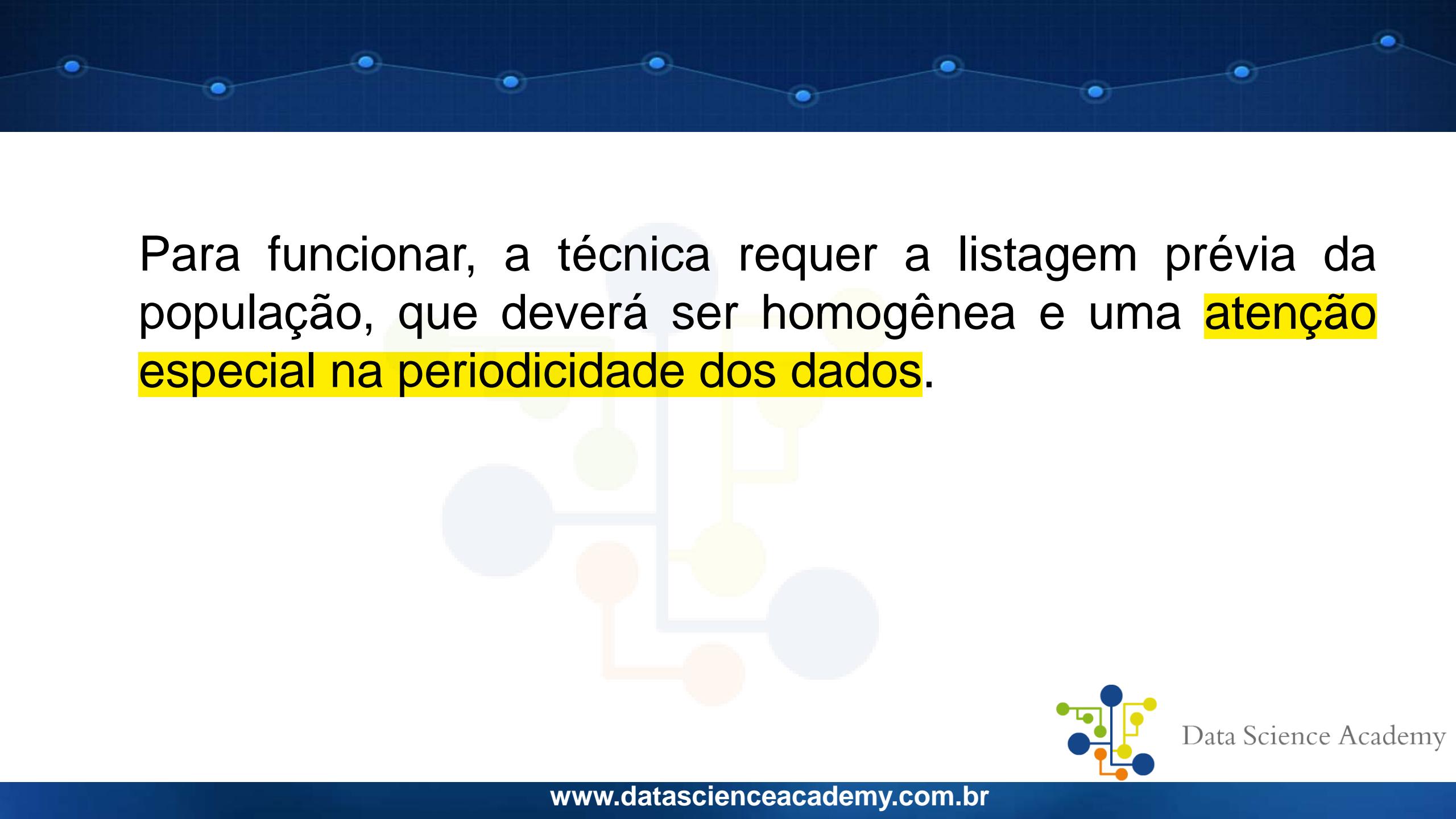


Data Science Academy

A amostragem sistemática consiste em selecionar as unidades elementares da população em intervalos pré-fixados.



Data Science Academy



Para funcionar, a técnica requer a listagem prévia da população, que deverá ser homogênea e uma **atenção especial na periodicidade dos dados.**



Data Science Academy

A amostragem sistemática é semelhante à aleatória simples, mas a listagem é **ORDENADA**.



Data Science Academy

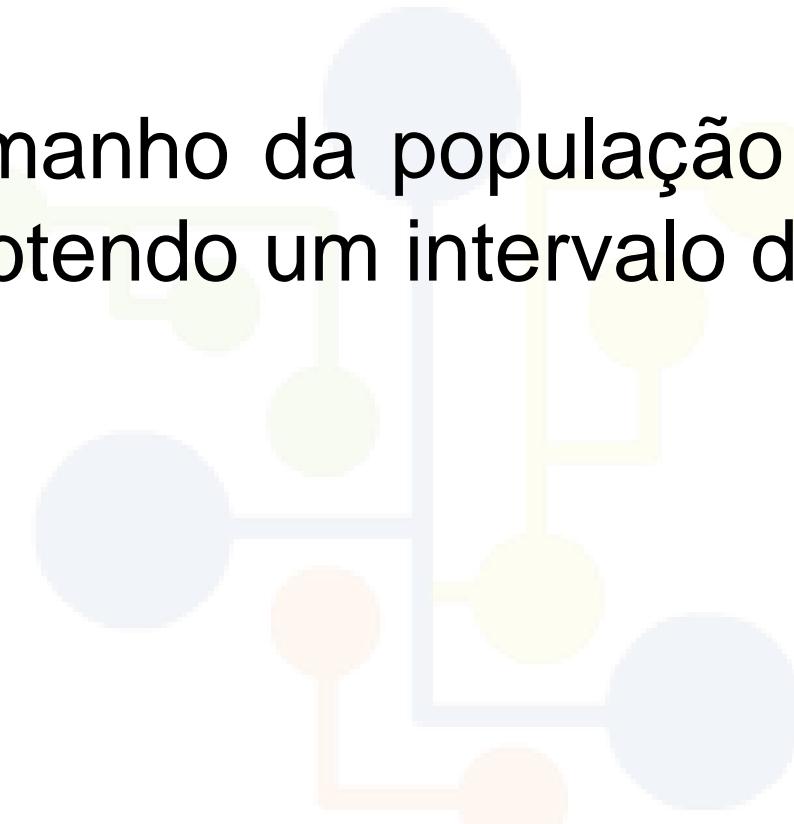
O processo basicamente é:



Data Science Academy

O processo basicamente é:

divide-se o tamanho da população (**N**) pelo tamanho da amostra (**n**), obtendo um intervalo de retirada (**k**).



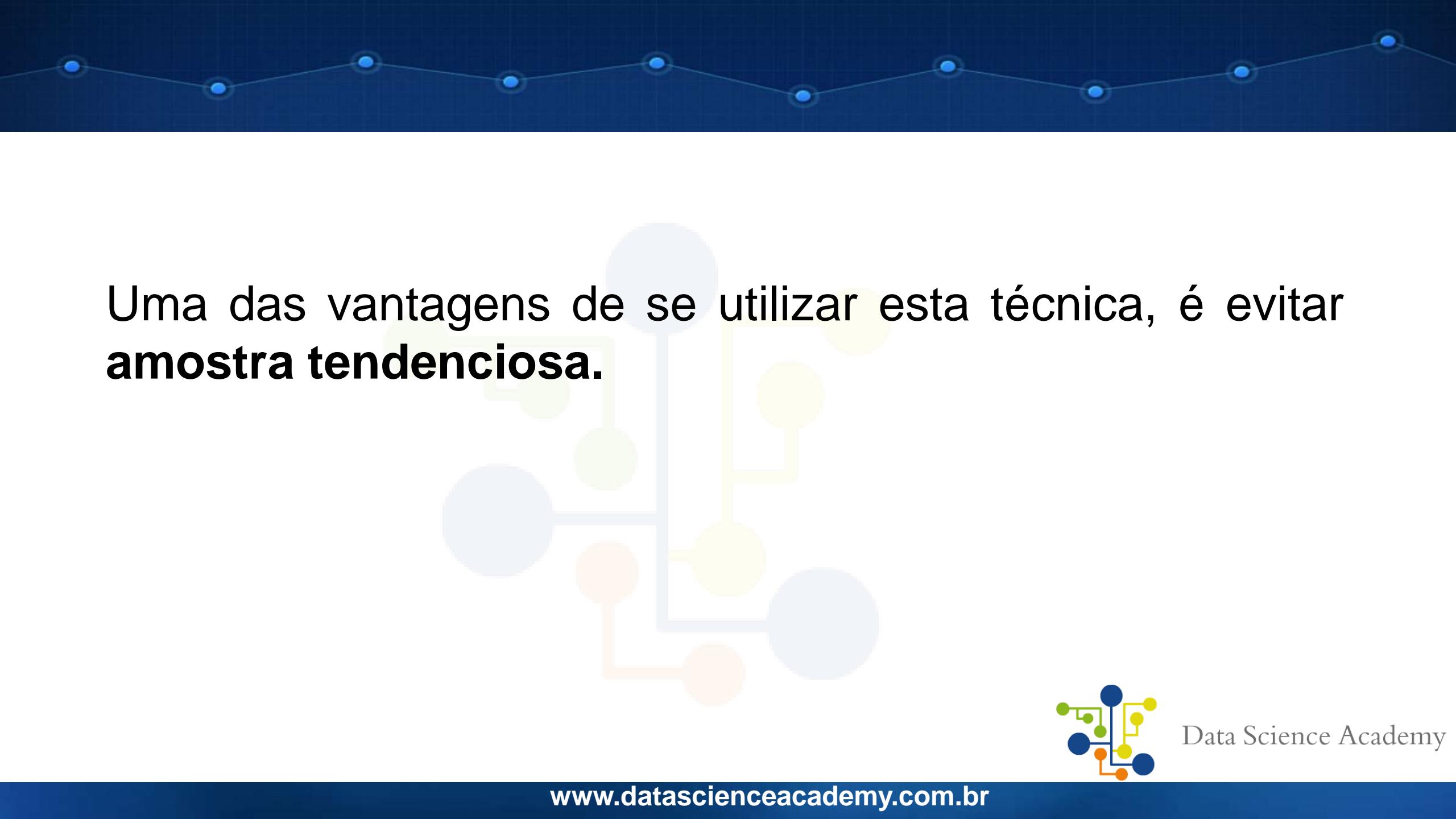
Data Science Academy

O processo basicamente é:

Sorteia-se o ponto de partida. A cada **k** elementos retira-se um para a amostra (**k** pode ser o segundo, quarto, vigésimo elemento, etc...)



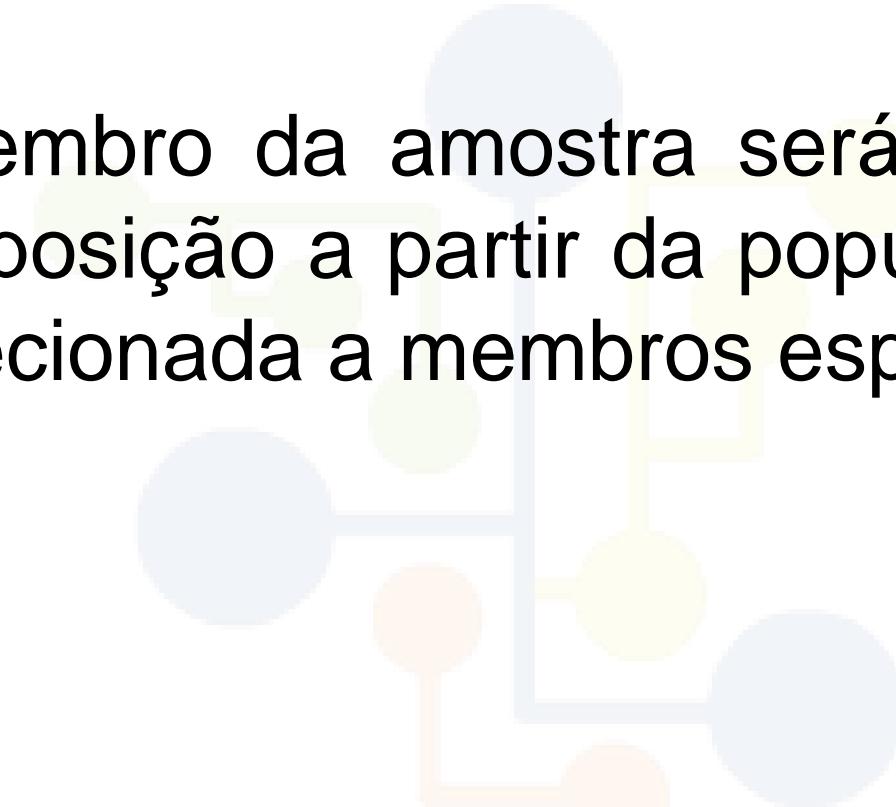
Data Science Academy



Uma das vantagens de se utilizar esta técnica, é evitar amostra tendenciosa.



Data Science Academy



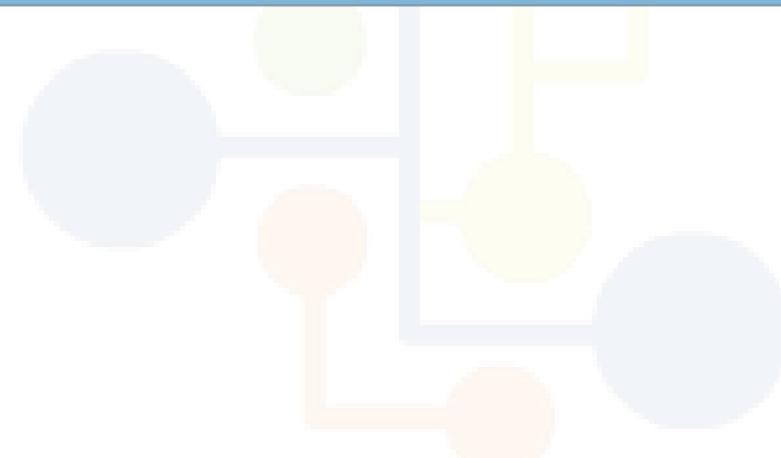
Como o membro da amostra será selecionado sempre na mesma posição a partir da população, evitamos uma seleção direcionada a membros específicos.



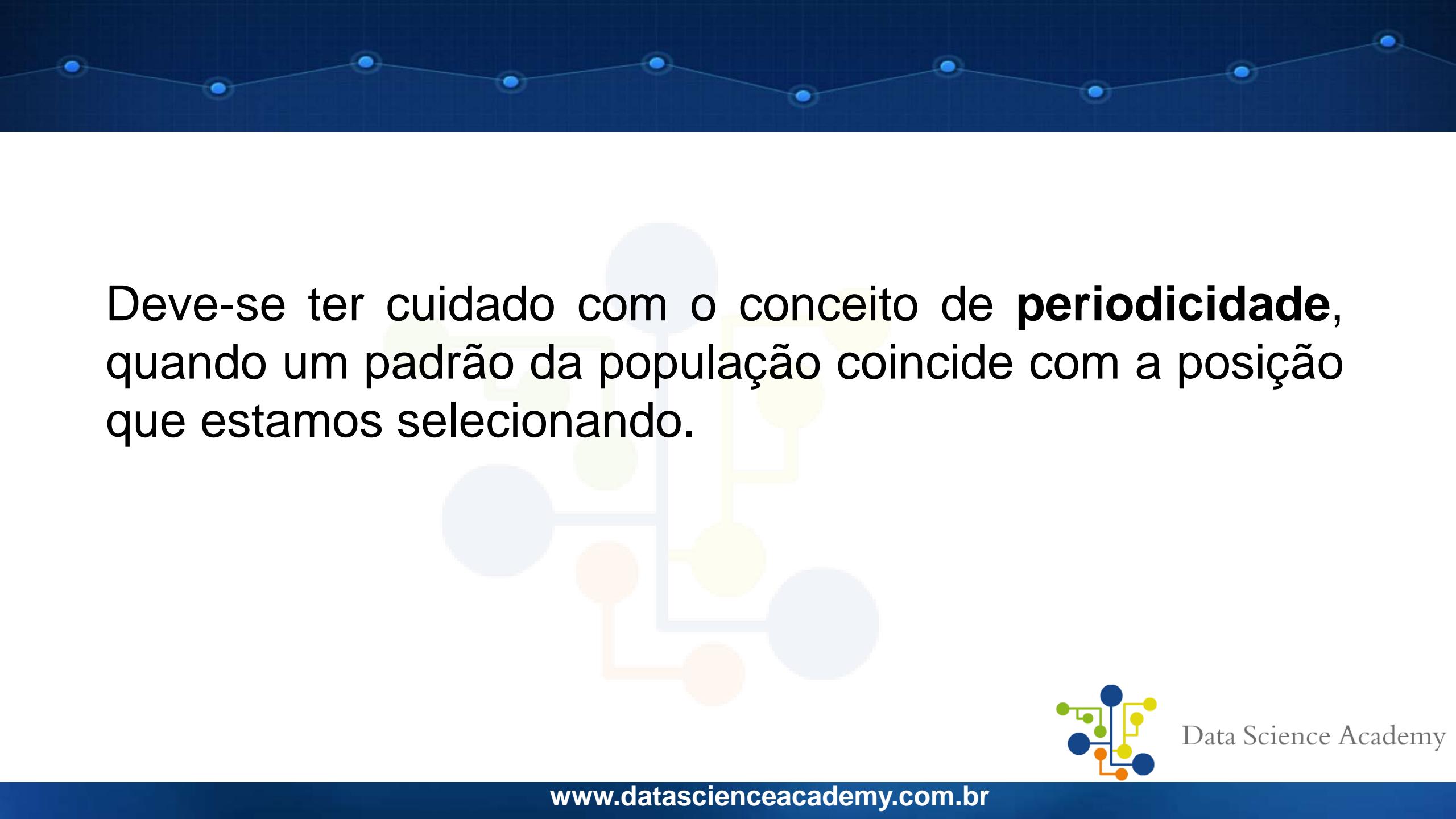
Data Science Academy



Entretanto,



Data Science Academy

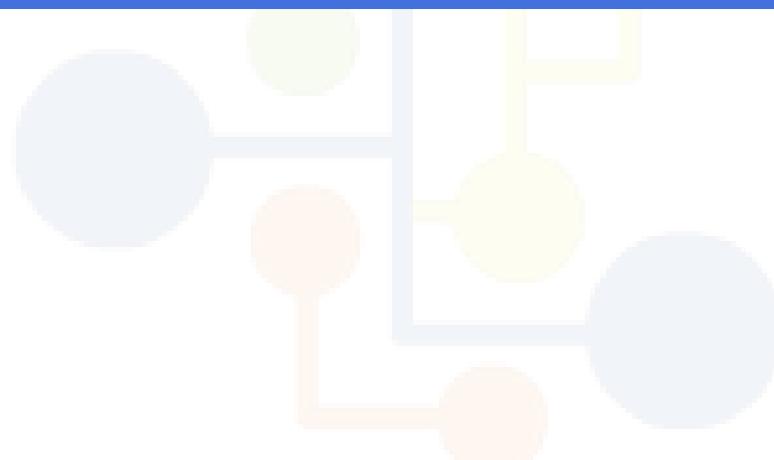


Deve-se ter cuidado com o conceito de **periodicidade**, quando um padrão da população coincide com a posição que estamos selecionando.



Data Science Academy

Exemplo



Data Science Academy

Vamos supor que existe a necessidade de fazer uma pesquisa com estudantes universitários, para descobrir **quantas horas** por semana eles **estudam**.

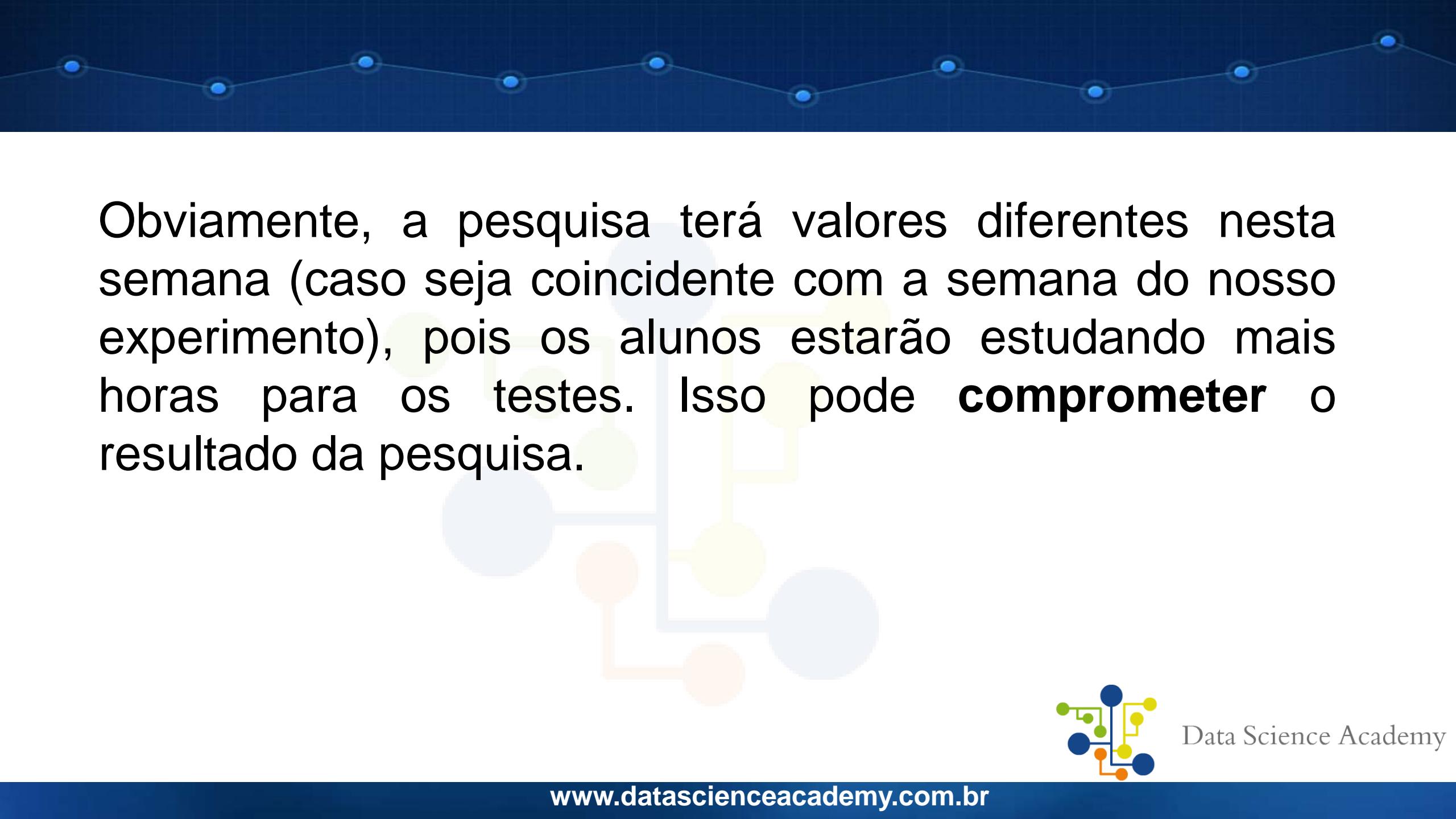


Data Science Academy

Para tal experimento, selecionamos sempre a **quarta semana** de cada mês para a pesquisa. Entretanto, a cada 2 meses a universidade aplica testes e provas, previstos no calendário, na última semana do bimestre.



Data Science Academy



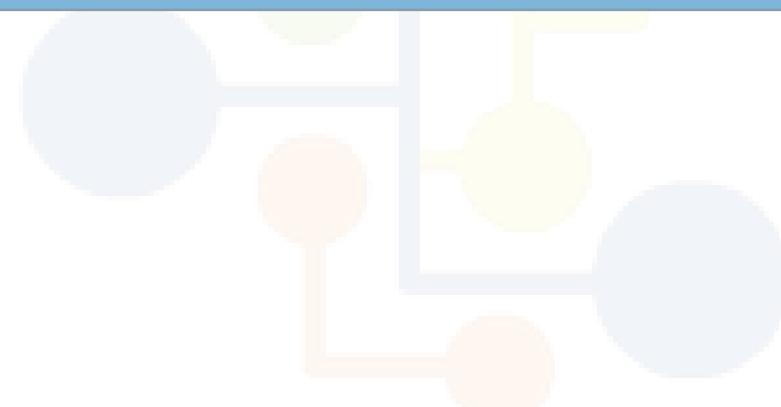
Obviamente, a pesquisa terá valores diferentes nesta semana (caso seja coincidente com a semana do nosso experimento), pois os alunos estarão estudando mais horas para os testes. Isso pode **comprometer** o resultado da pesquisa.



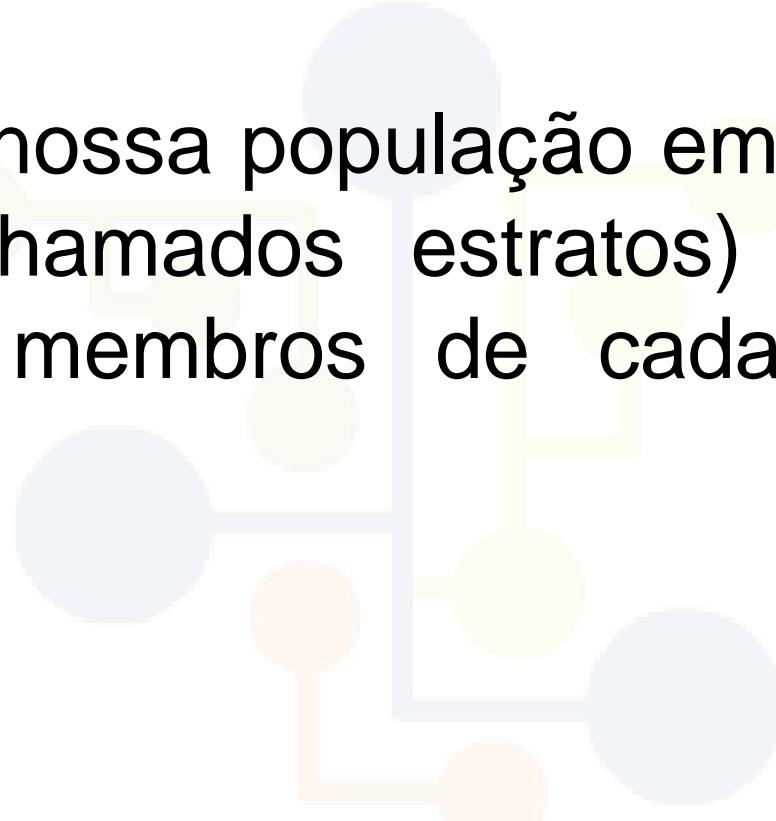
Data Science Academy



Amostragem Estratificada



Data Science Academy



Nós dividimos nossa população em grupos **mutuamente exclusivos** (chamados estratos) e **randomicamente** selecionamos membros de cada grupo para nossa amostra.



Data Science Academy

Exemplo

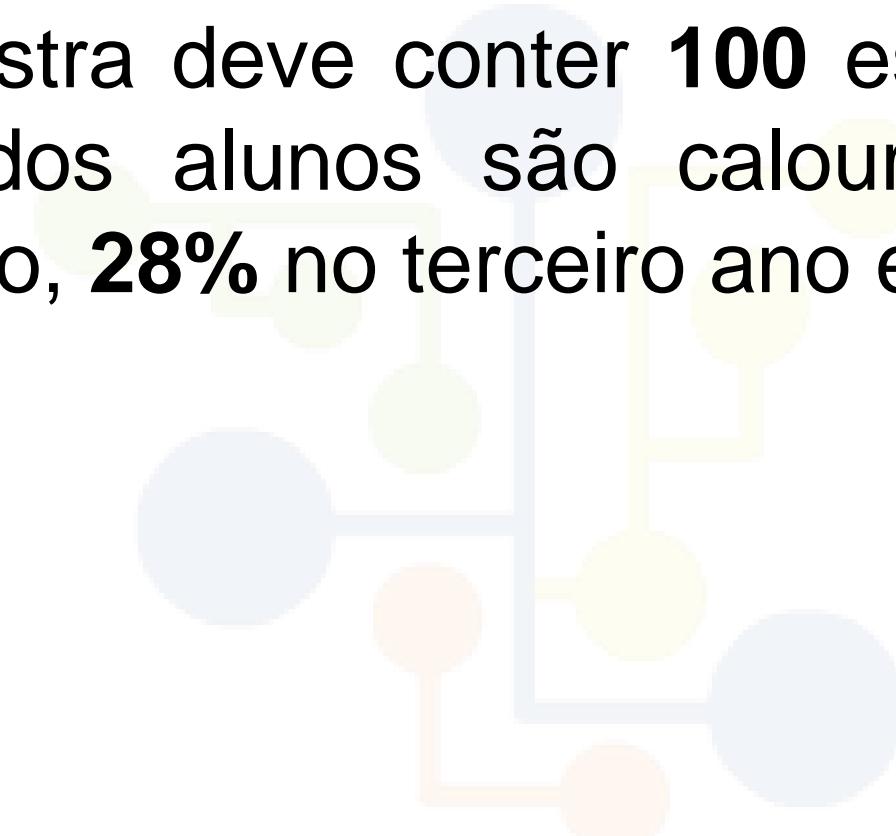


Data Science Academy

Vamos supor que continuamos fazendo um experimento com estudantes universitários, mas desta vez desejamos criar uma **amostra de estudantes**, com a qual faremos nossa pesquisa de **horas de estudo por semana**.



Data Science Academy



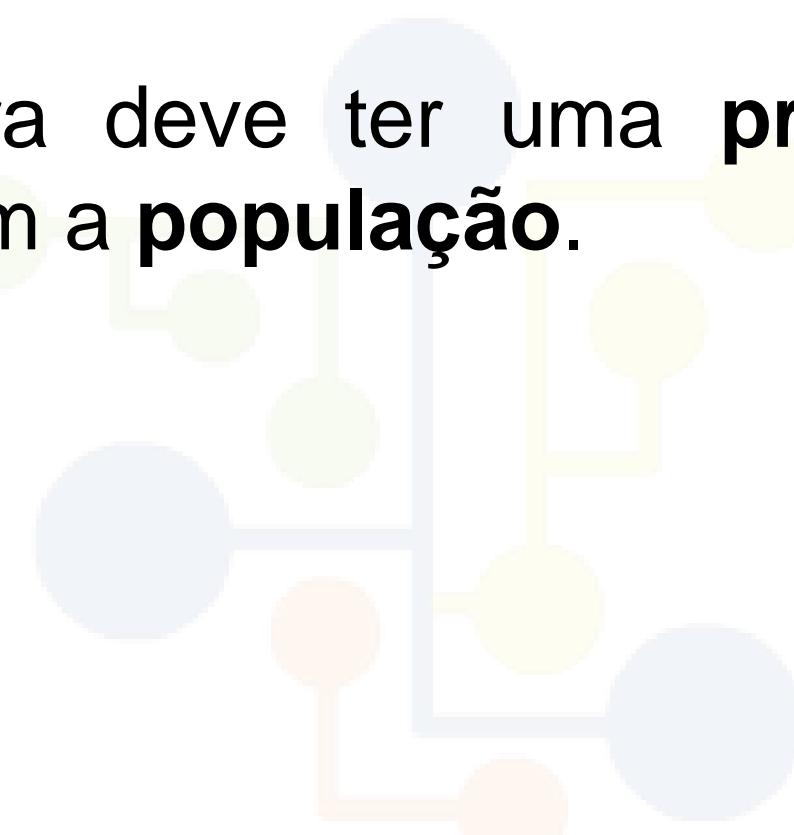
Nossa amostra deve conter **100** estudantes e sabemos que **30%** dos alunos são calouros, **22%** estão no segundo ano, **28%** no terceiro ano e **20%** no quarto ano.



Data Science Academy

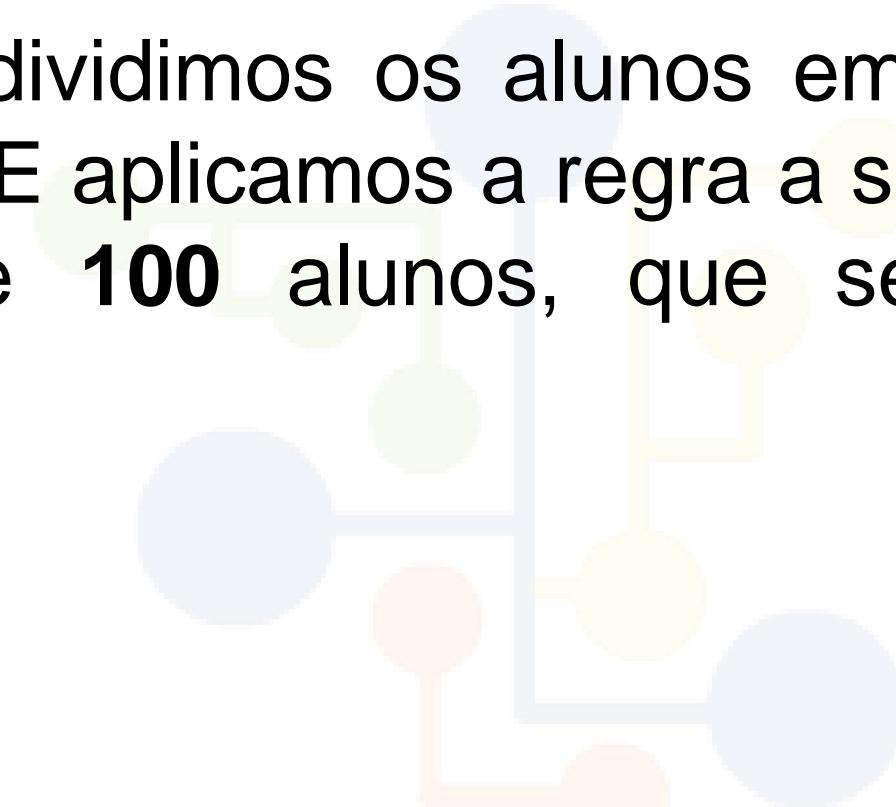


Nossa amostra deve ter uma **proporção** de alunos, condizente com a **população**.



Data Science Academy

Para isso, dividimos os alunos em grupos mutuamente exclusivos. E aplicamos a regra a seguir para criar nossa amostra de **100** alunos, que serão alvo da nossa pesquisa.



Data Science Academy

Teremos uma amostra (100 ← alunos) que é representativa da população.

Grupo	Número de alunos (% da População)	Amostra
Calouros	540 (30%)	$0.30 \times 100 = 30$
Estudantes no segundo ano	396 (22%)	$0.22 \times 100 = 22$
Estudantes no terceiro ano	504 (28%)	$0.28 \times 100 = 28$
Estudantes no quarto ano	360 (20%)	$0.20 \times 100 = 20$
Total	1.800 (100%)	100



Data Science Academy

Teremos uma amostra (**100 alunos**) que é representativa da população.

Grupo	Número de alunos (% da População)	Amostra
Calouros	540 (30%)	$0.30 \times 100 = 30$
Estudantes no segundo ano	396 (22%)	$0.22 \times 100 = 22$
Estudantes no terceiro ano	504 (28%)	$0.28 \times 100 = 28$
Estudantes no quarto ano	360 (20%)	$0.20 \times 100 = 20$
Total	1.800 (100%)	100

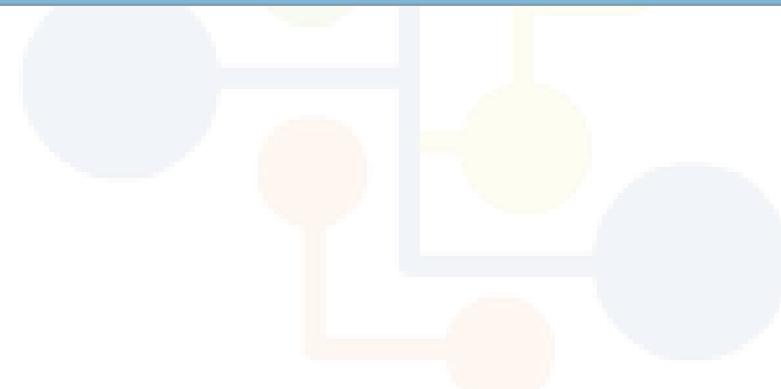
Cada grupo é homogêneo, ou seja, possui as mesmas características em relação a população.



Data Science Academy

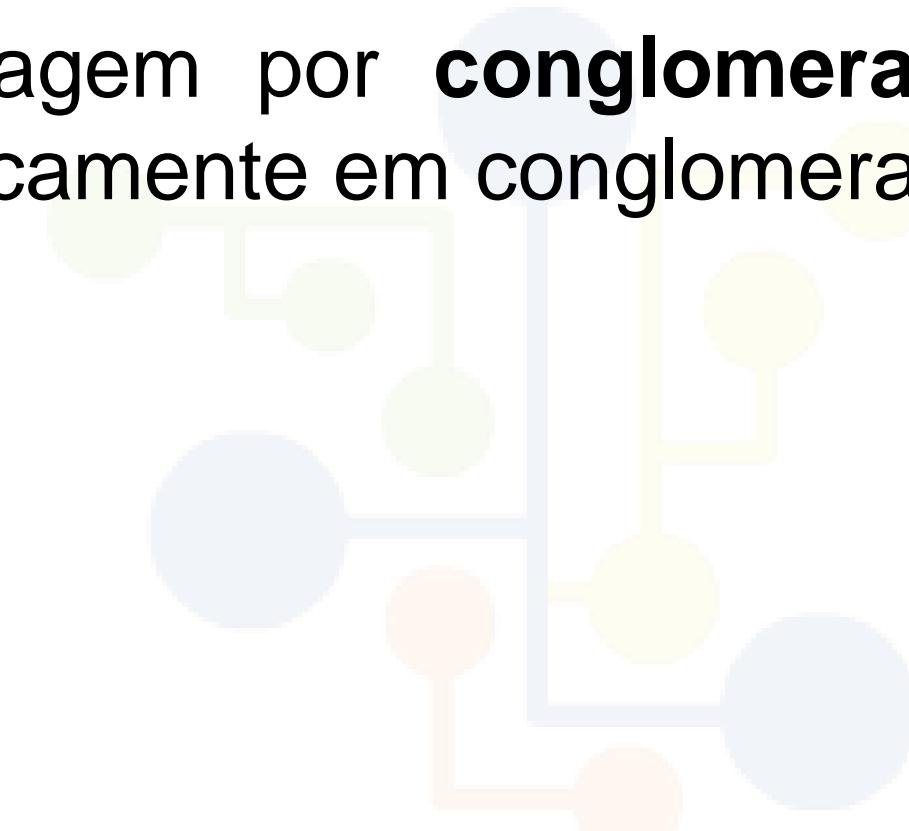


Amostragem Por Conglomerado

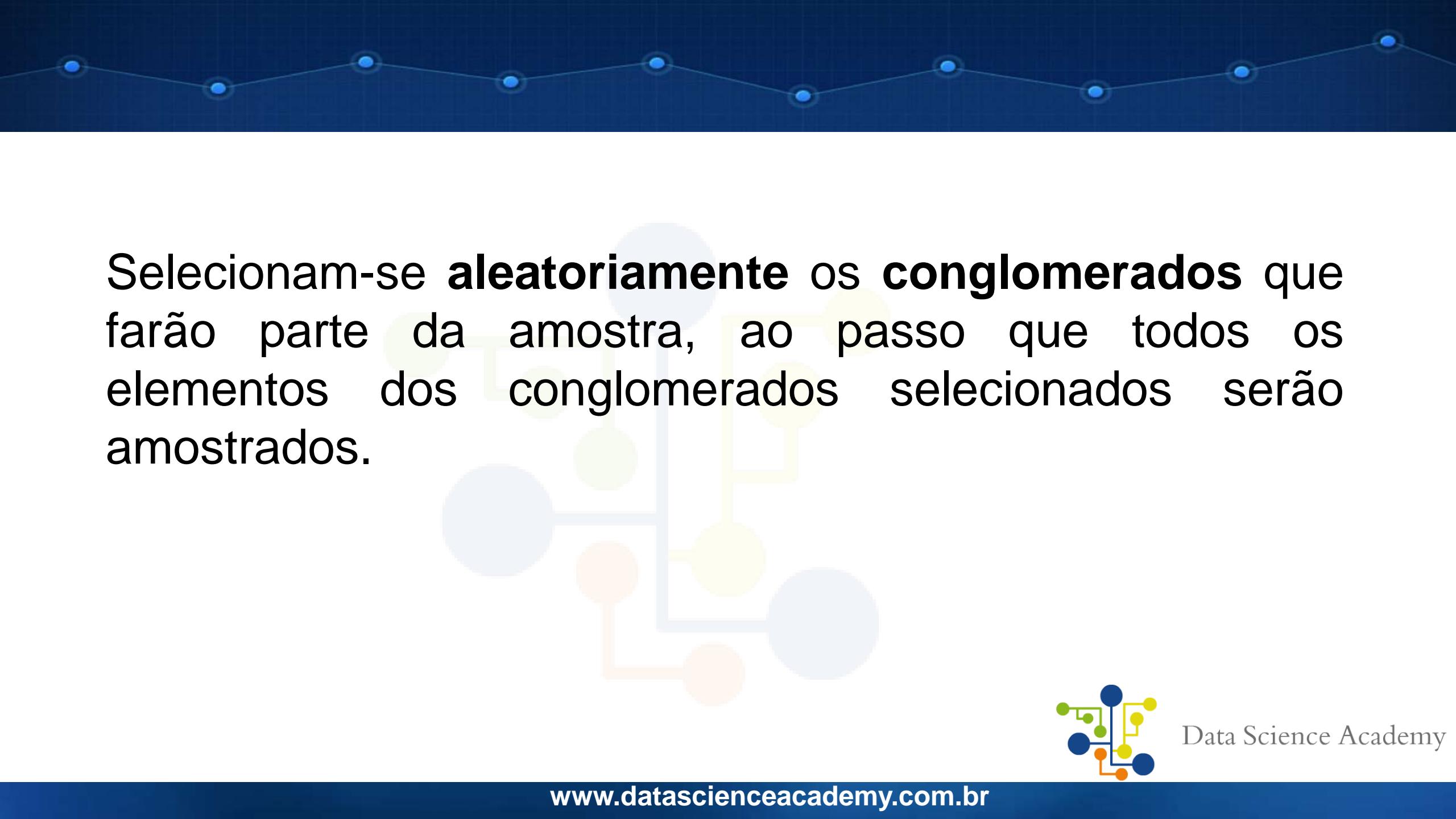


Data Science Academy

Na amostragem por **conglomerados**, a população é dividida fisicamente em conglomerados.



Data Science Academy



Selecionam-se **aleatoriamente** os **conglomerados** que farão parte da amostra, ao passo que todos os elementos dos conglomerados selecionados serão amostrados.



Data Science Academy



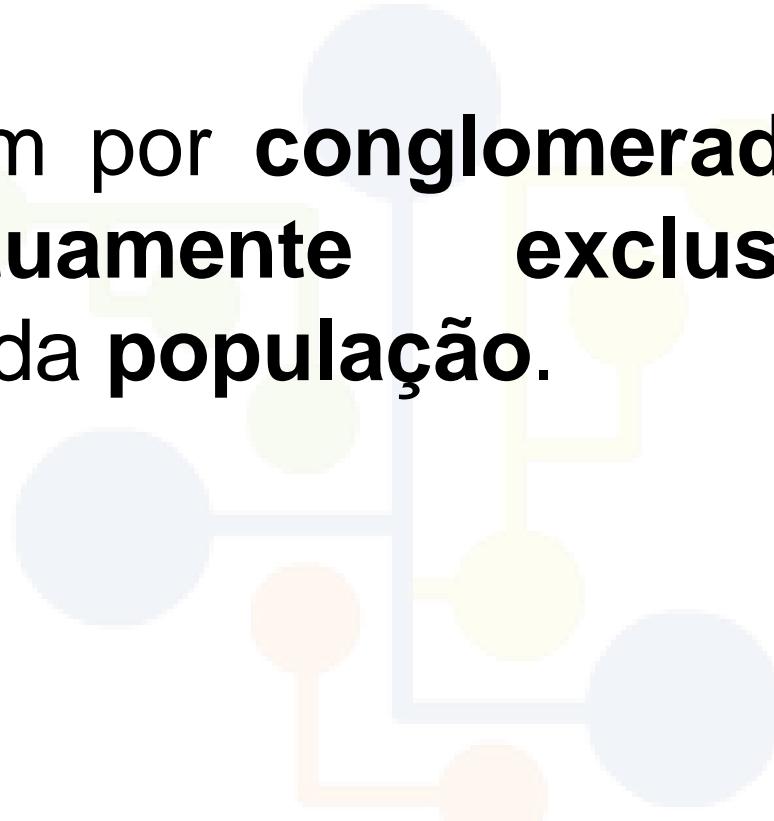
É muito utilizada quando há necessidade de se realizar entrevistas ou observações em **grandes áreas**.



Data Science Academy



Na amostragem por **conglomerados**, também criamos grupos **mutuamente exclusivos**, cada um representativo da **população**.



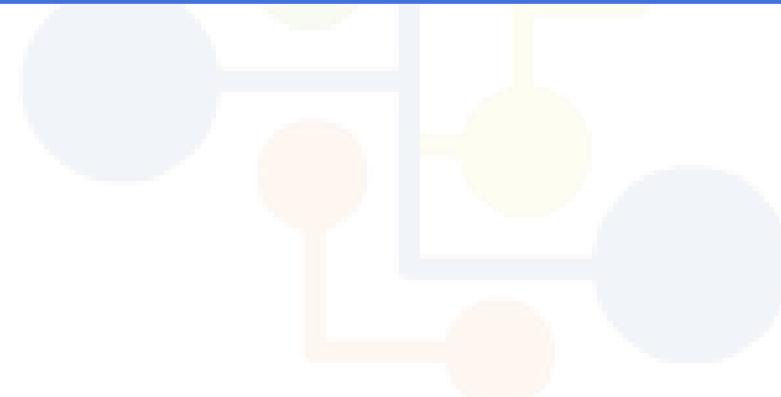
Data Science Academy

Ao invés de selecionarmos membros de cada grupo, selecionamos **randomicamente** grupos inteiros para nossa **amostra**.



Data Science Academy

Exemplo



Data Science Academy

Vamos usar o exemplo anterior, mas desta vez criar uma amostra por conglomerados.



Data Science Academy

Vamos usar o exemplo anterior, mas desta vez criar uma amostra por conglomerados.

Grupo	Tamanho (% da População)	Amostra
Turma de Estatística	540 (30%)	$0.30 \times 100 = 30$
Turma de Filosofia	396 (22%)	$0.22 \times 100 = 22$
Turma de Cálculo I	504 (28%)	$0.28 \times 100 = 28$
Turma de Contabilidade II	360 (20%)	$0.20 \times 100 = 20$
Total	1.800 (100%)	100



Data Science Academy

mutuamente
exclusivo

Grupo	Número de alunos (% da População)	Amostra
Calouros	540 (30%)	$0.30 \times 100 = 30$
Estudantes no segundo ano	396 (22%)	$0.22 \times 100 = 22$
Estudantes no terceiro ano	504 (28%)	$0.28 \times 100 = 28$
Estudantes no quarto ano	360 (20%)	$0.20 \times 100 = 20$
Total	1.800 (100%)	100

não mutuamente
exclusivo

Grupo	Tamanho (% da População)	Amostra
Turma de Estatística	540 (30%)	$0.30 \times 100 = 30$
Turma de Filosofia	396 (22%)	$0.22 \times 100 = 22$
Turma de Cálculo I	504 (28%)	$0.28 \times 100 = 28$
Turma de Contabilidade II	360 (20%)	$0.20 \times 100 = 20$
Total	1.800 (100%)	100

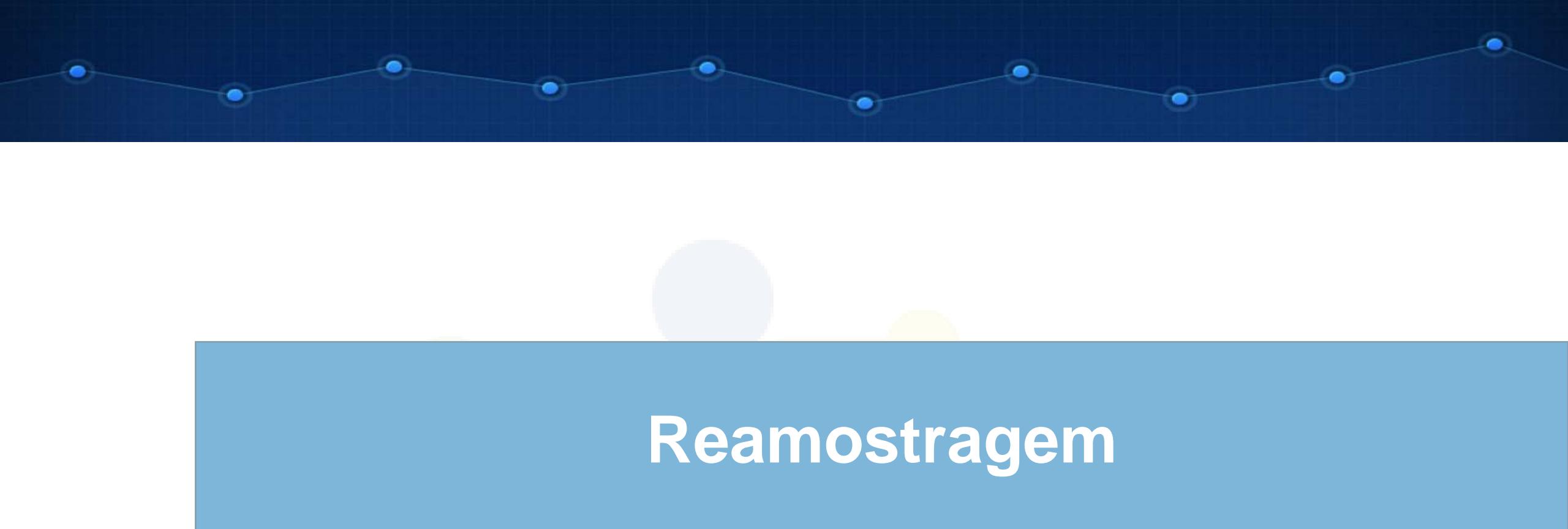


Data Science Academy

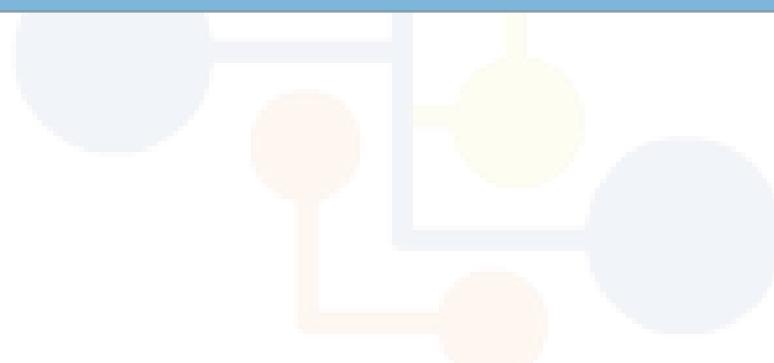
Percebeu que neste caso, os grupos são **heterogêneos**, ou seja, é possível que tenhamos alunos do segundo e terceiro ano na turma de Filosofia. Cada **grupo** representa a **população**. Temos agora nossa amostra de **100 alunos** criada por outra técnica de amostragem.

Grupo	Tamanho (% da População)	Amostra
Turma de Estatística	540 (30%)	$0.30 \times 100 = 30$
Turma de Filosofia	396 (22%)	$0.22 \times 100 = 22$
Turma de Cálculo I	504 (28%)	$0.28 \times 100 = 28$
Turma de Contabilidade II	360 (20%)	$0.20 \times 100 = 20$
Total	1.800 (100%)	100

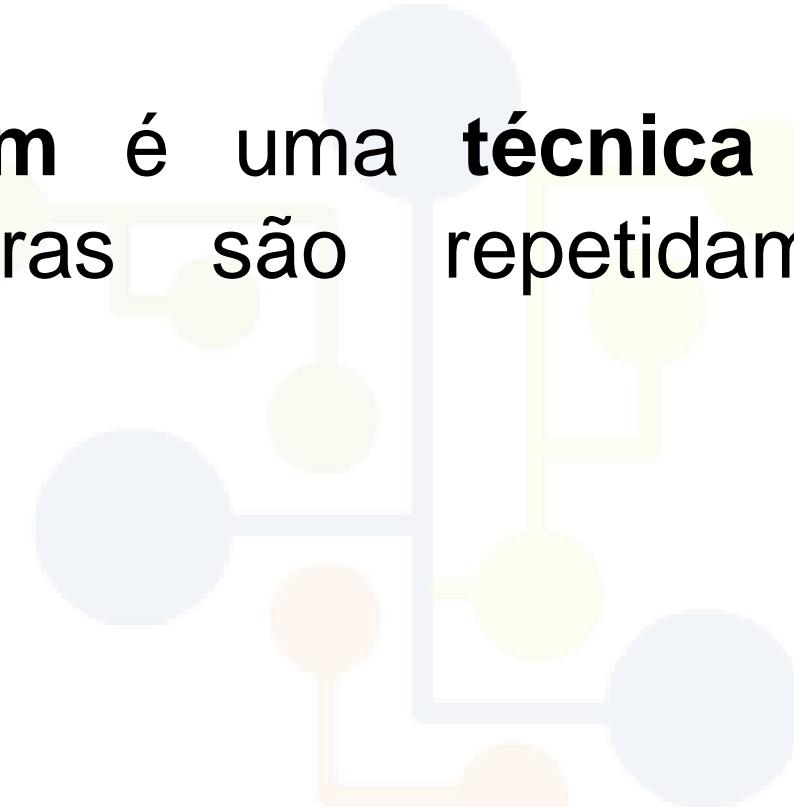




Reamostragem



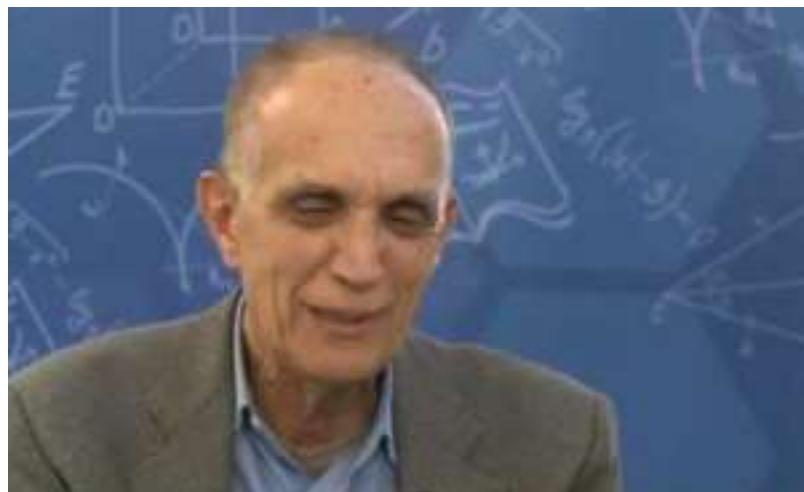
Data Science Academy



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



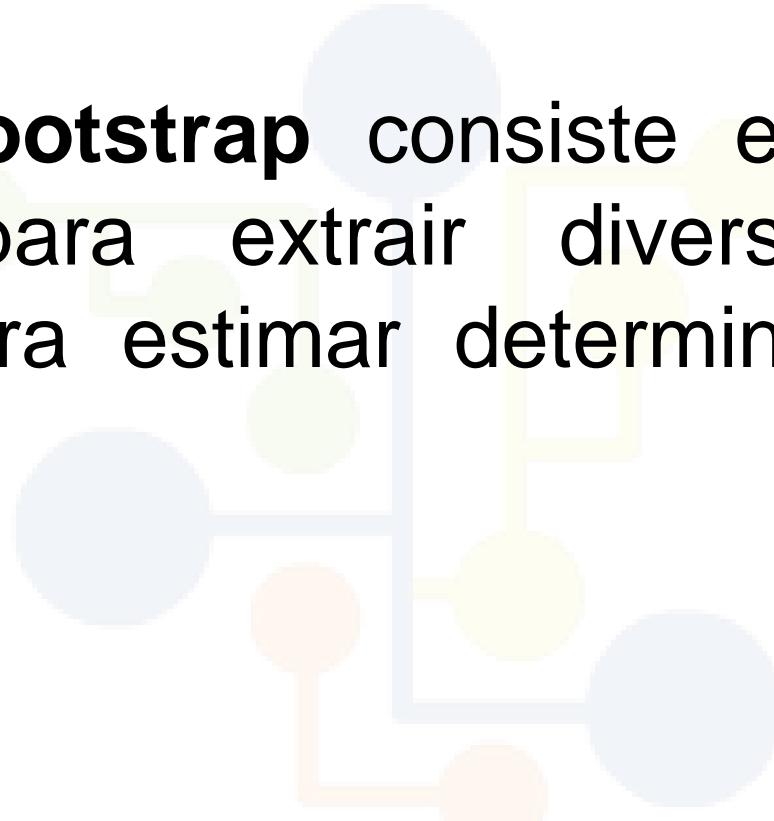
Data Science Academy



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.



Data Science Academy



O método **Bootstrap** consiste em usar software de computador para extrair diversas amostras (com reposição), para estimar determinados parâmetros da população.



Data Science Academy

Tais como: Média e Proporção



Data Science Academy



Estatística está para amostra assim como parâmetro está para população



Data Science Academy

Exemplo

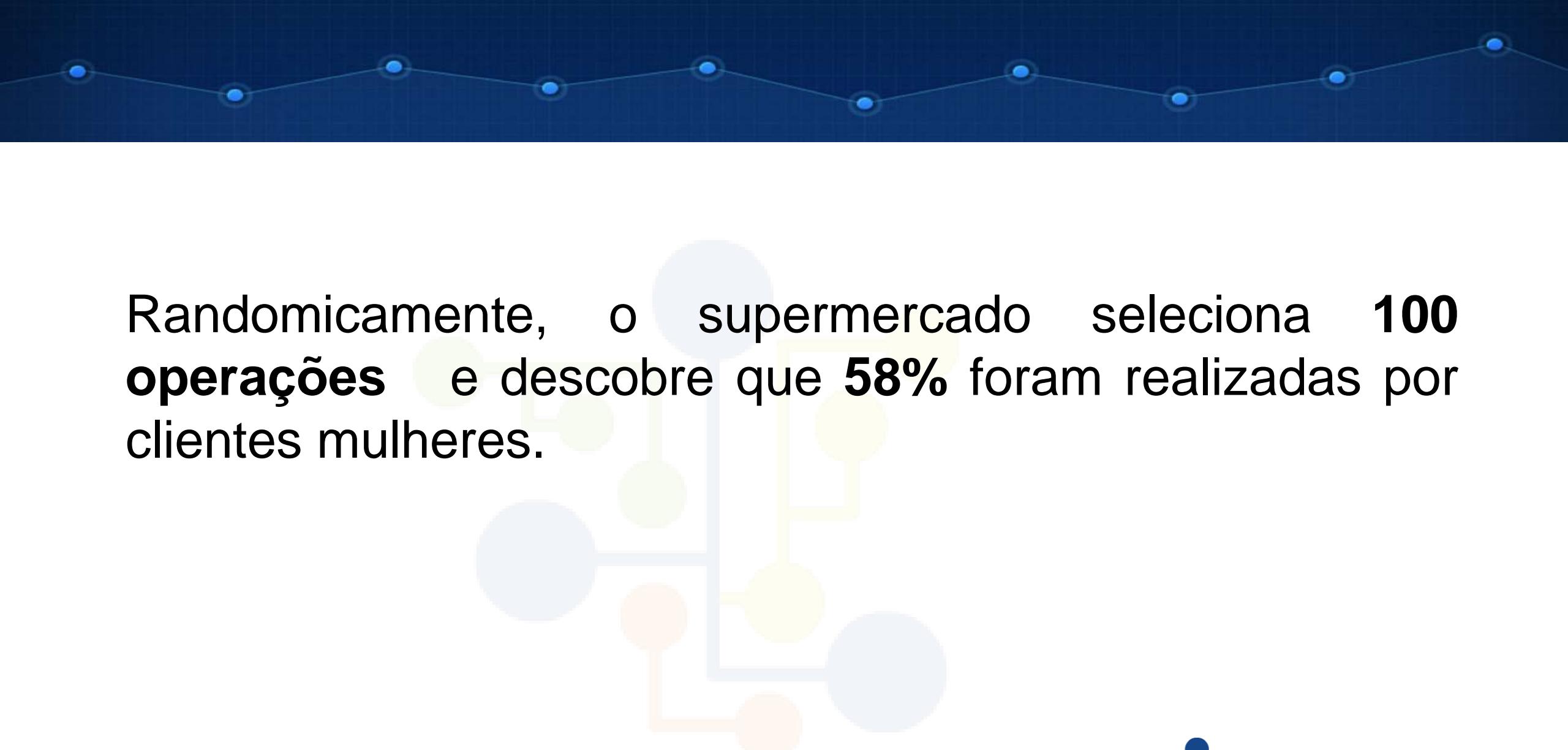


Data Science Academy

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



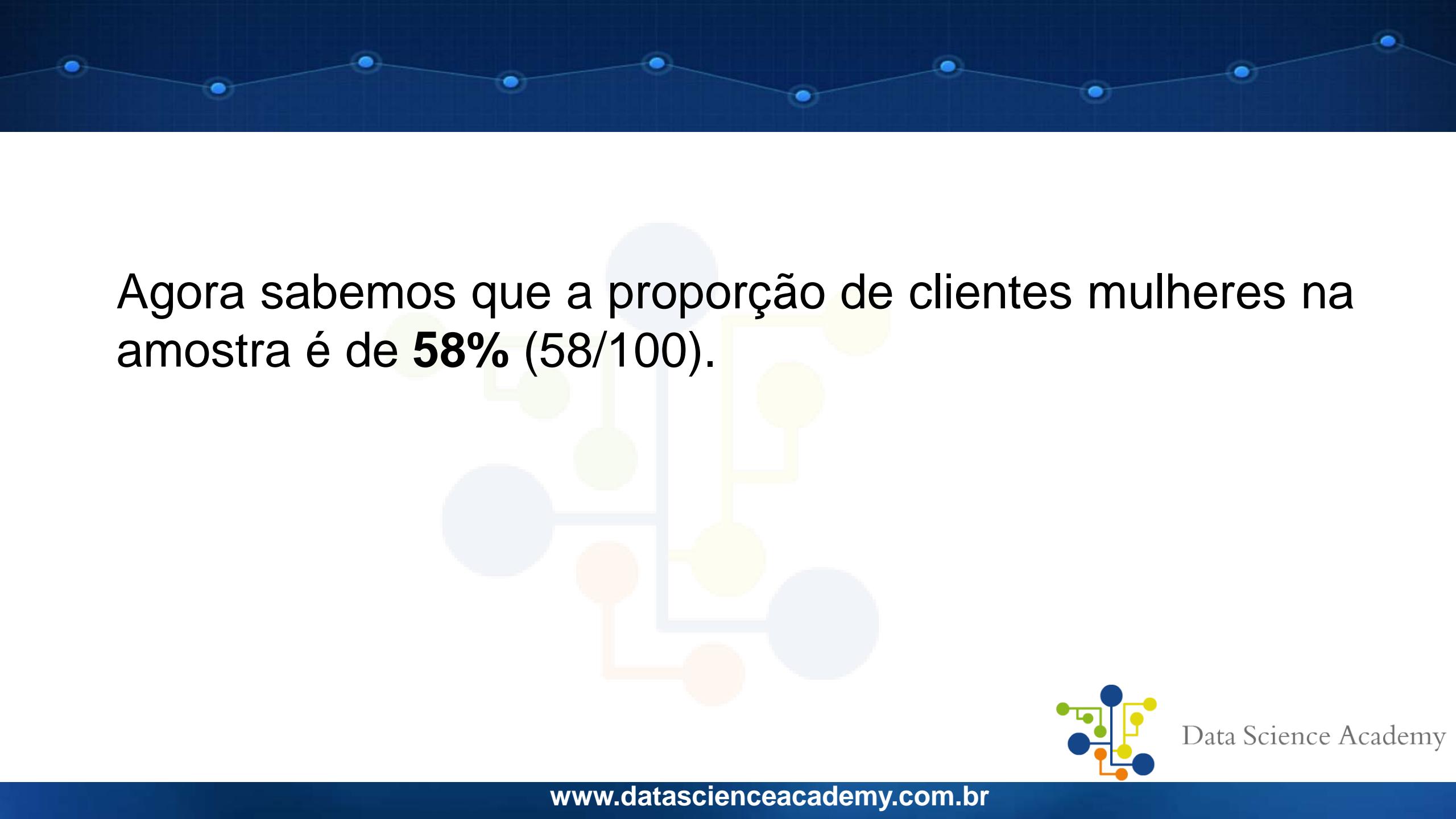
Data Science Academy



Randomicamente, o supermercado seleciona **100 operações** e descobre que **58%** foram realizadas por clientes mulheres.



Data Science Academy



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



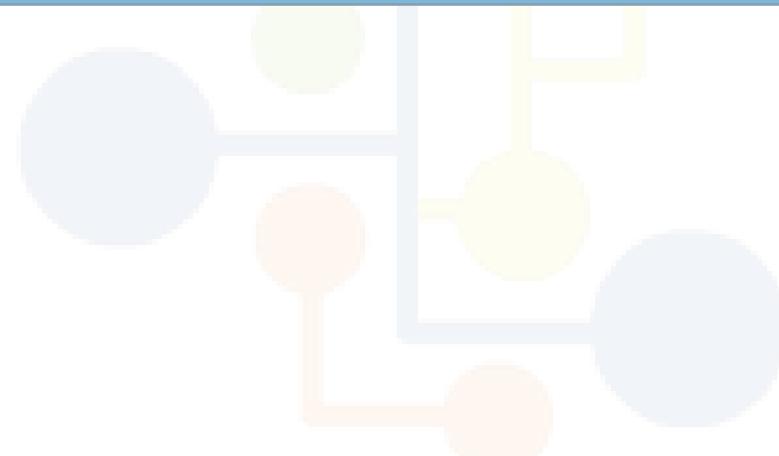
Data Science Academy

Esse tópico chegou ao final

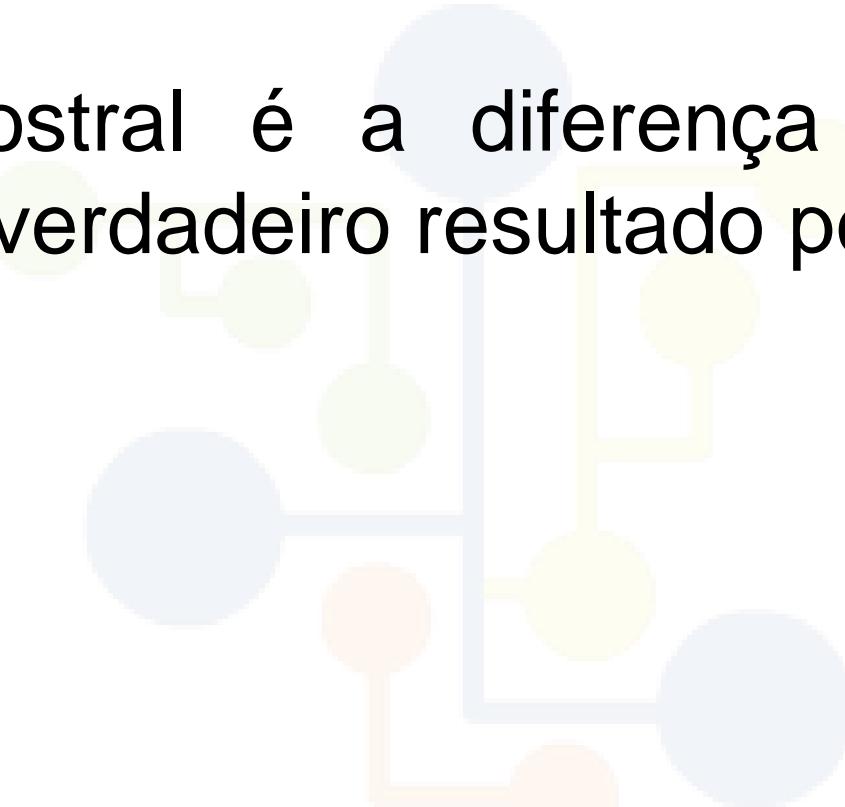


Data Science Academy

Erros de Amostragem



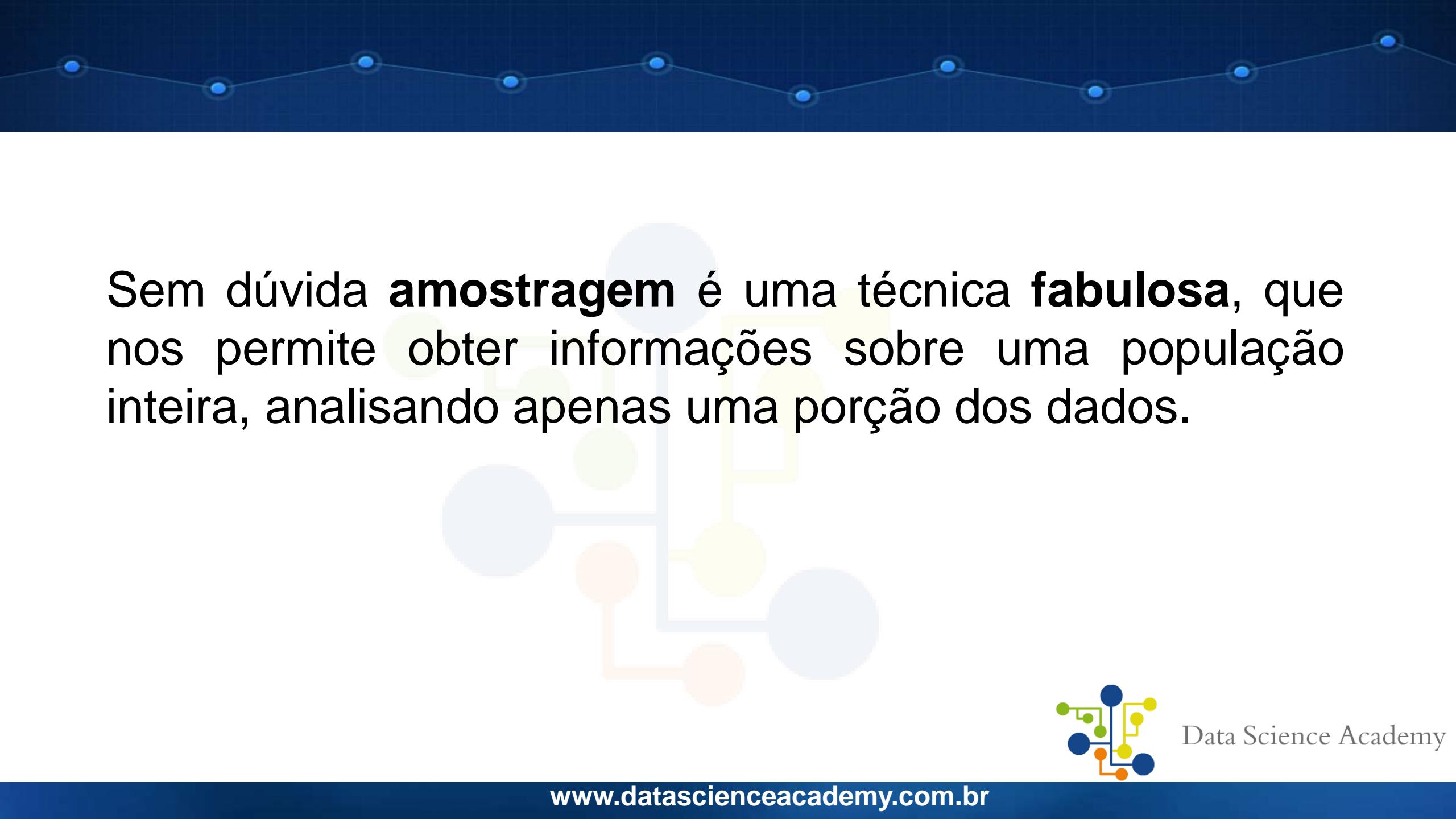
Data Science Academy



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



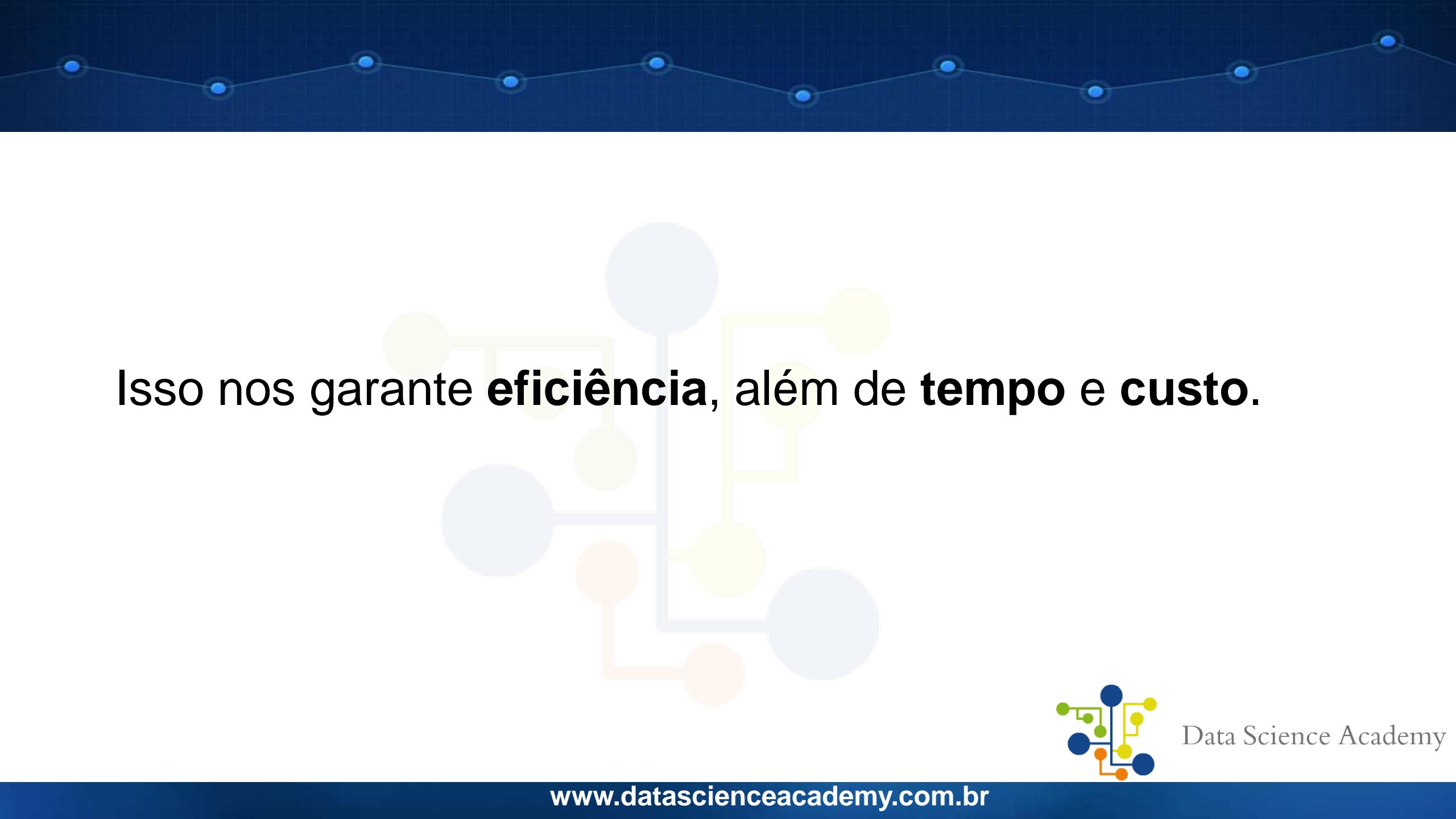
Data Science Academy



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



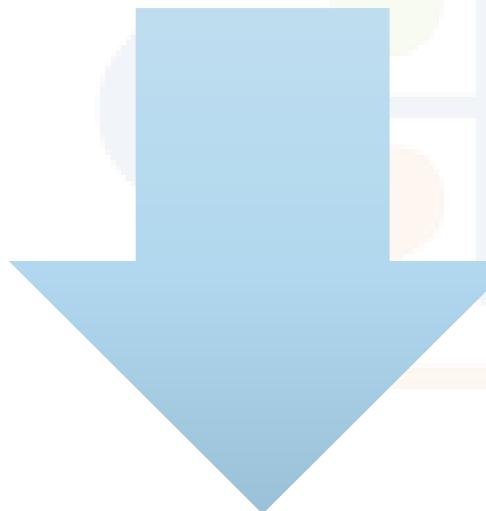
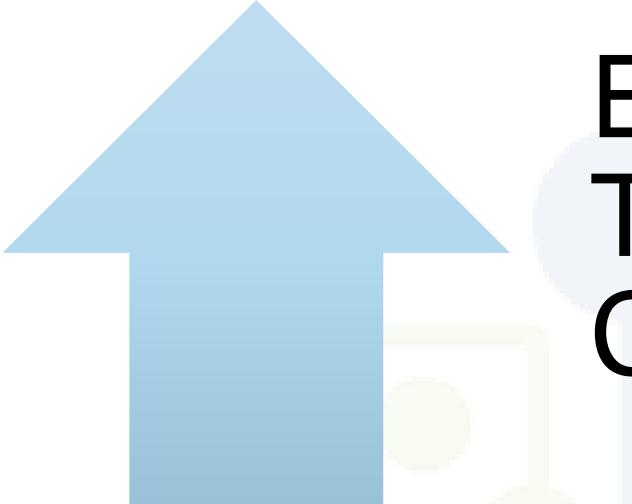
Data Science Academy



Isso nos garante **eficiência**, além de **tempo e custo**.



Data Science Academy



Eficiência ,
Tempo e
Custo

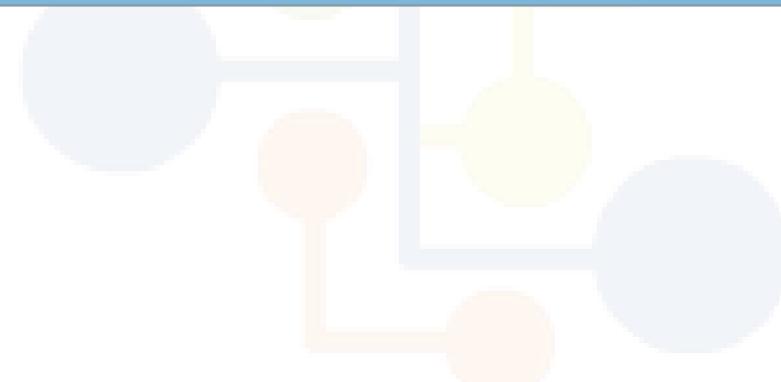
Quanto menor
a amostragem
maior os erros



Data Science Academy



Relembrando Conceitos



Data Science Academy



Parâmetro

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

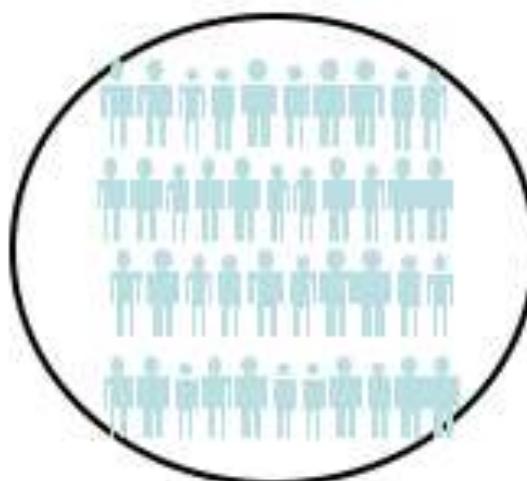
Estatística

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



Data Science Academy

Valores calculados usando dados da população são chamados de **parâmetros**.



População

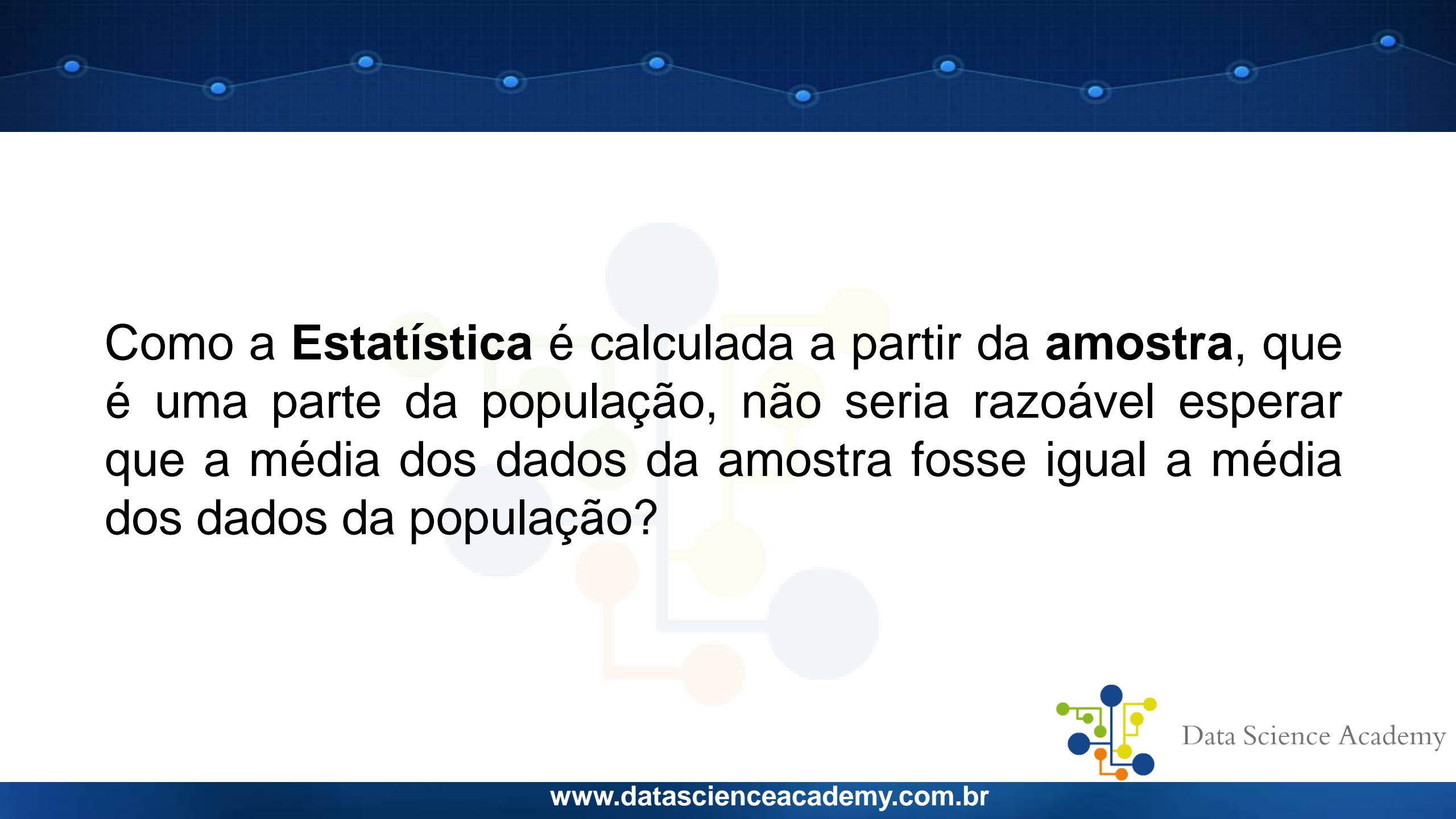
Valores computados de dados da amostra são chamados **estatística**.



Amostra



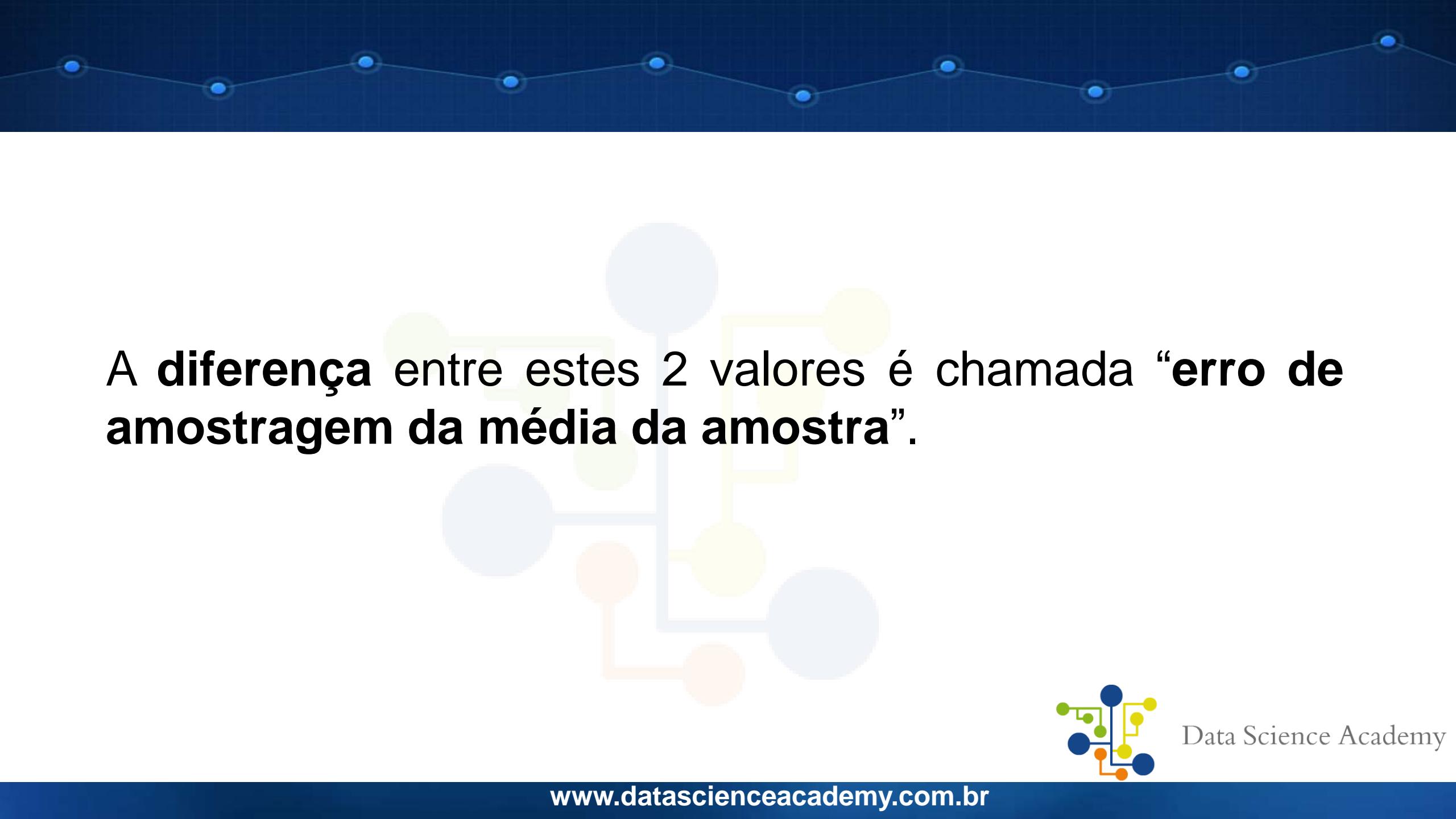
Data Science Academy



Como a **Estatística** é calculada 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?



Data Science Academy



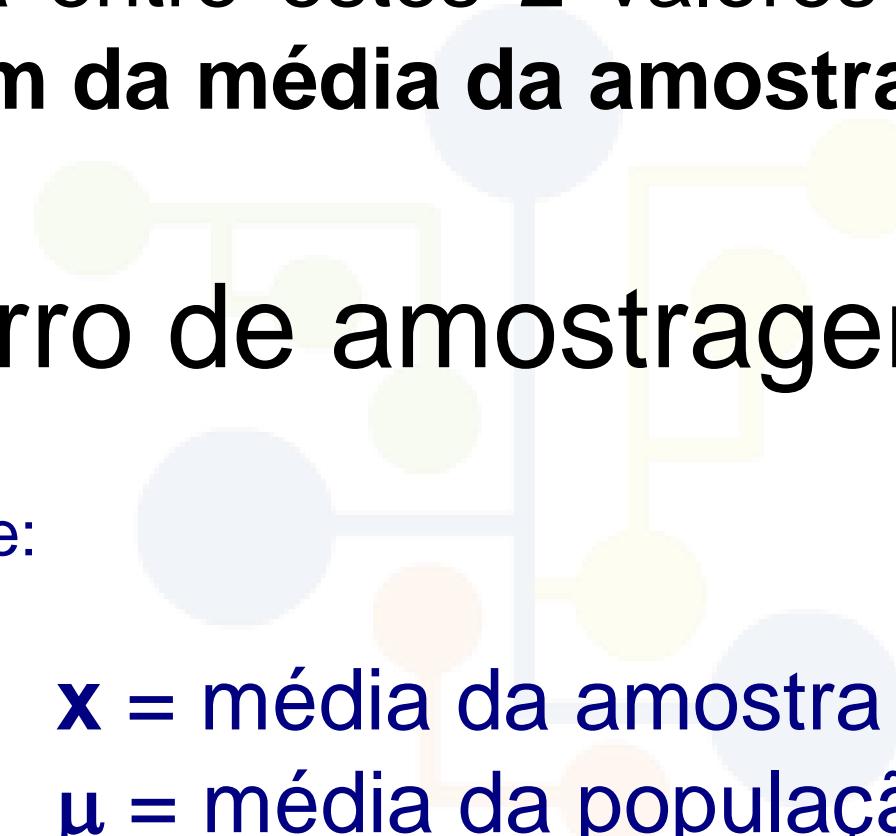
A diferença entre estes 2 valores é chamada “erro de amostragem da média da amostra”.



Data Science Academy



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



Erro de amostragem = $x - \mu$

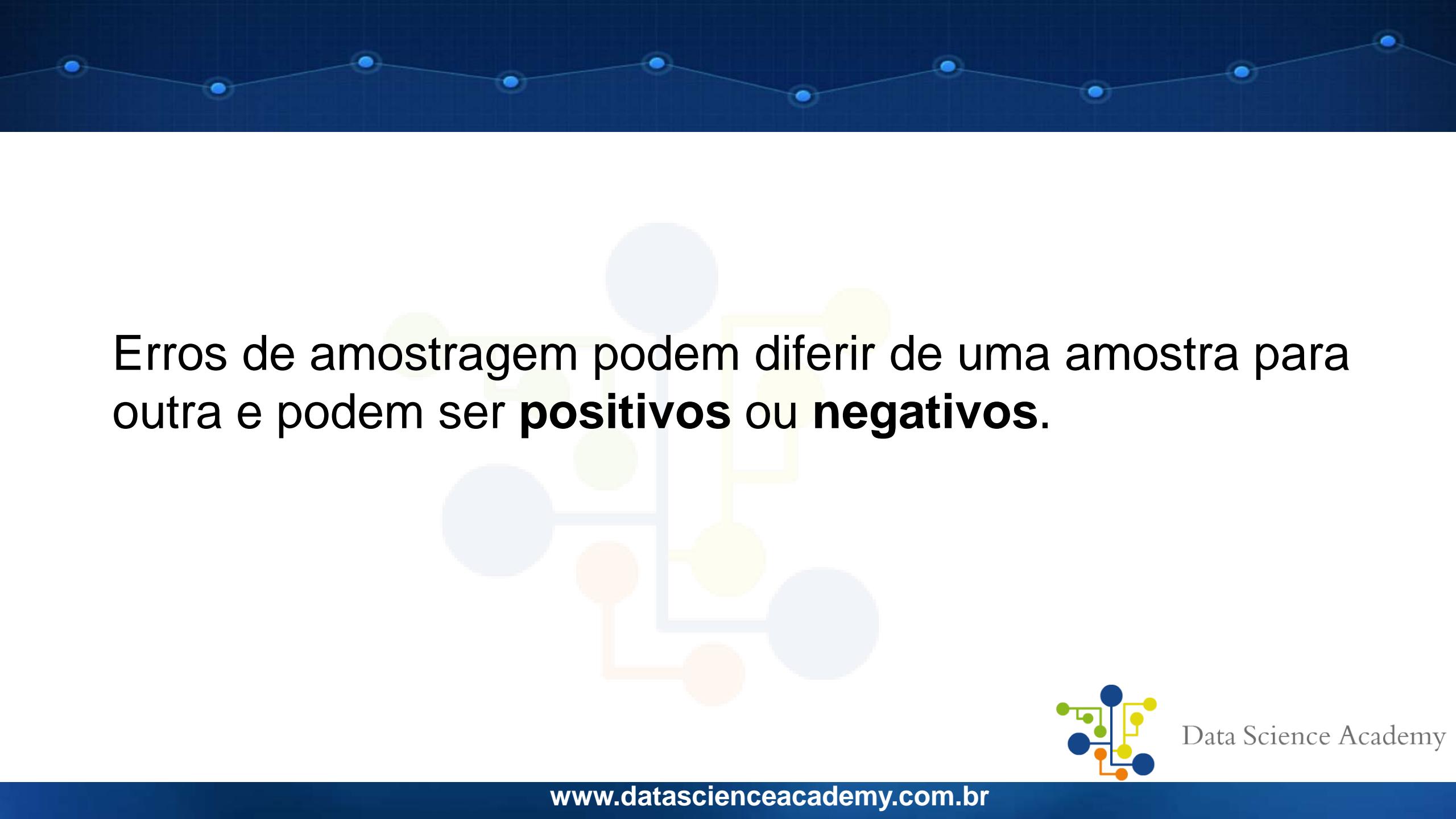
Onde:

x = média da amostra

μ = média da população



Data Science Academy



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



Data Science Academy



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



Data Science Academy

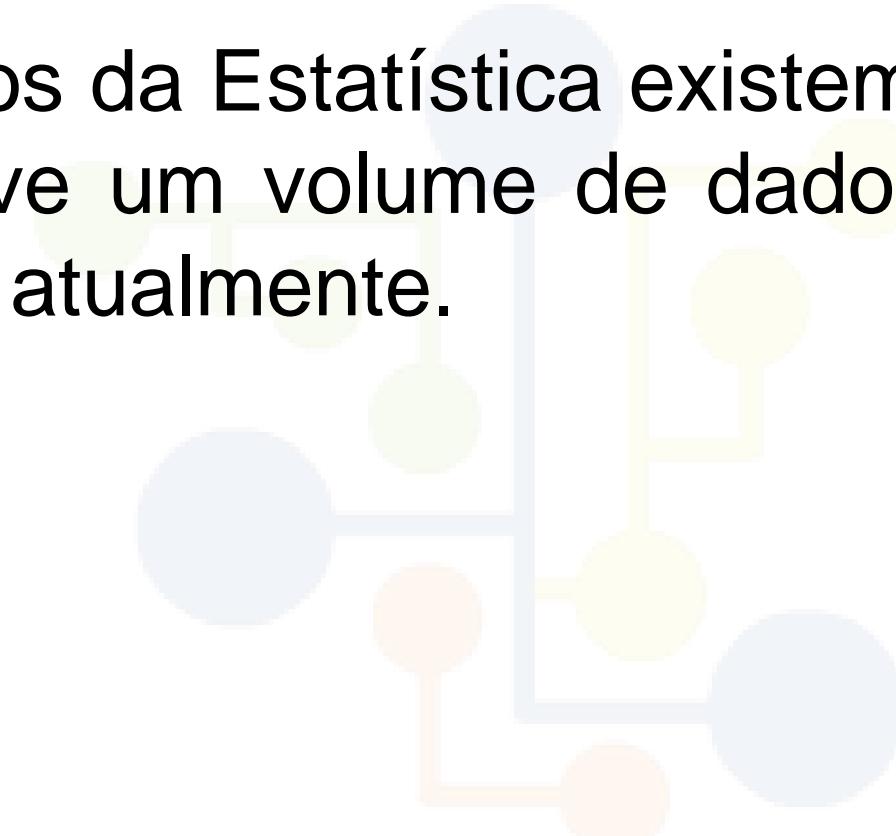
E agora talvez você entenda porque o Big Data é tão fabuloso. Com mais dados a nossa disposição e recursos tecnológicos, podemos coletar e utilizar cada vez mais dados e conduzir testes estatísticos cada vez mais precisos.



Data Science Academy



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



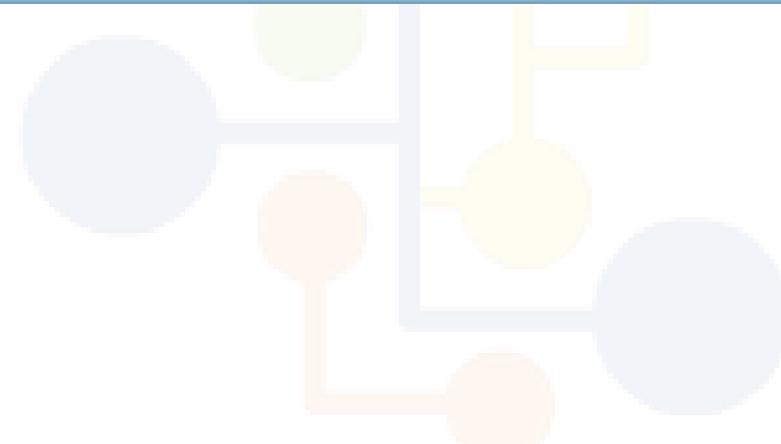
Data Science Academy

Esse tópico chegou ao final

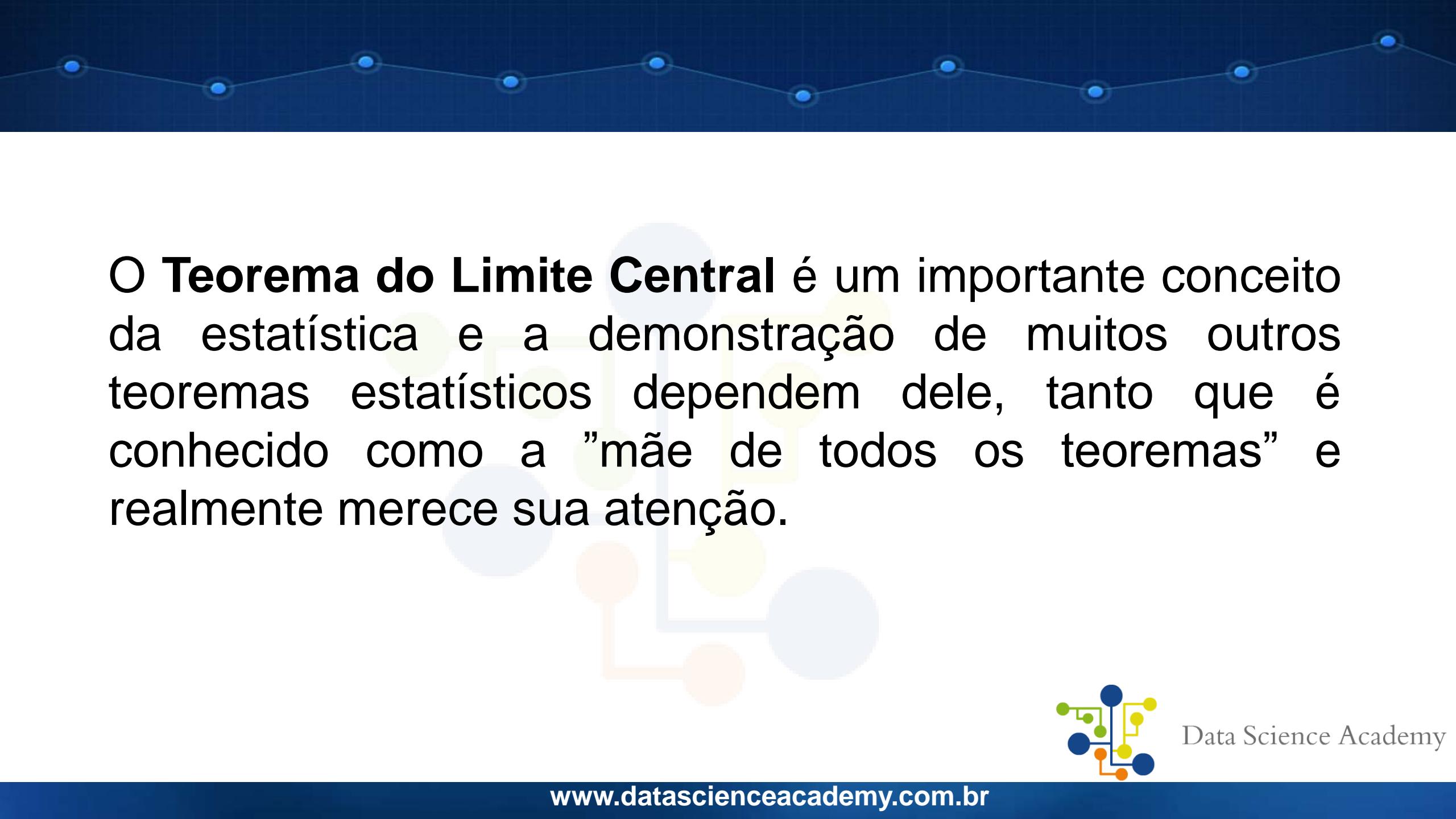


Data Science Academy

Teorema do Limite Central



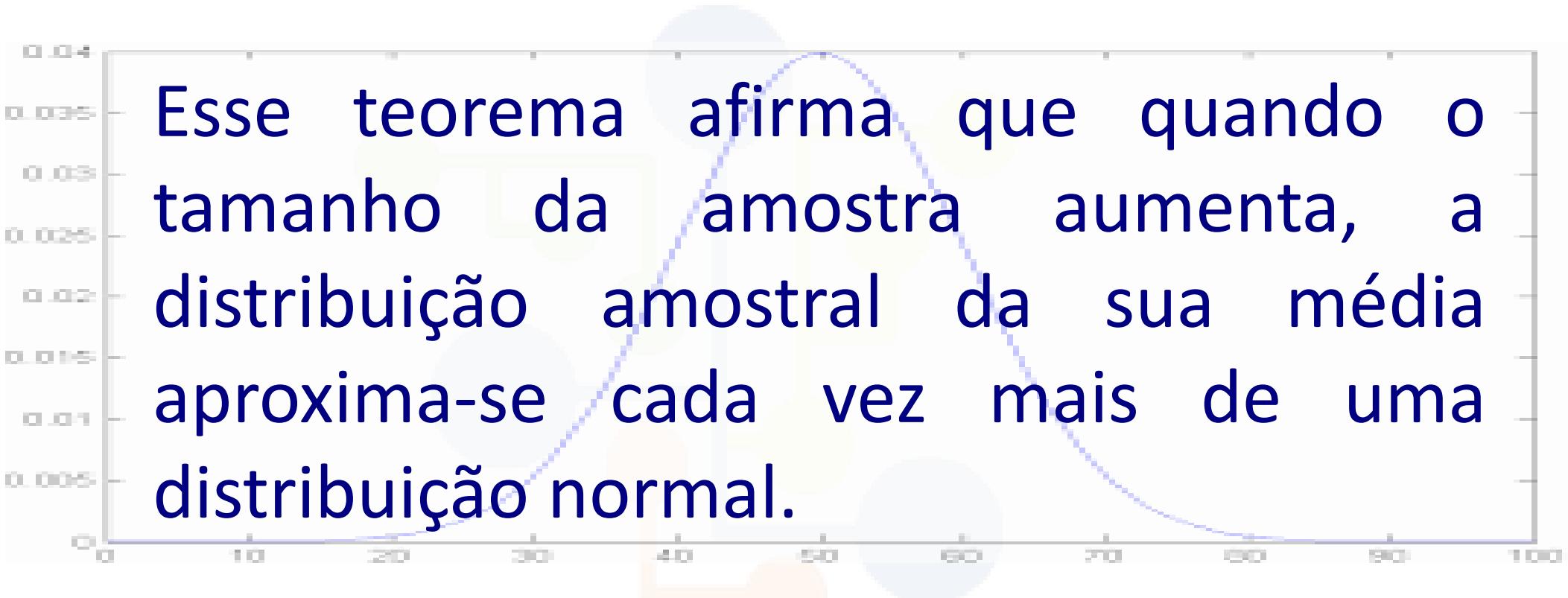
Data Science Academy



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.



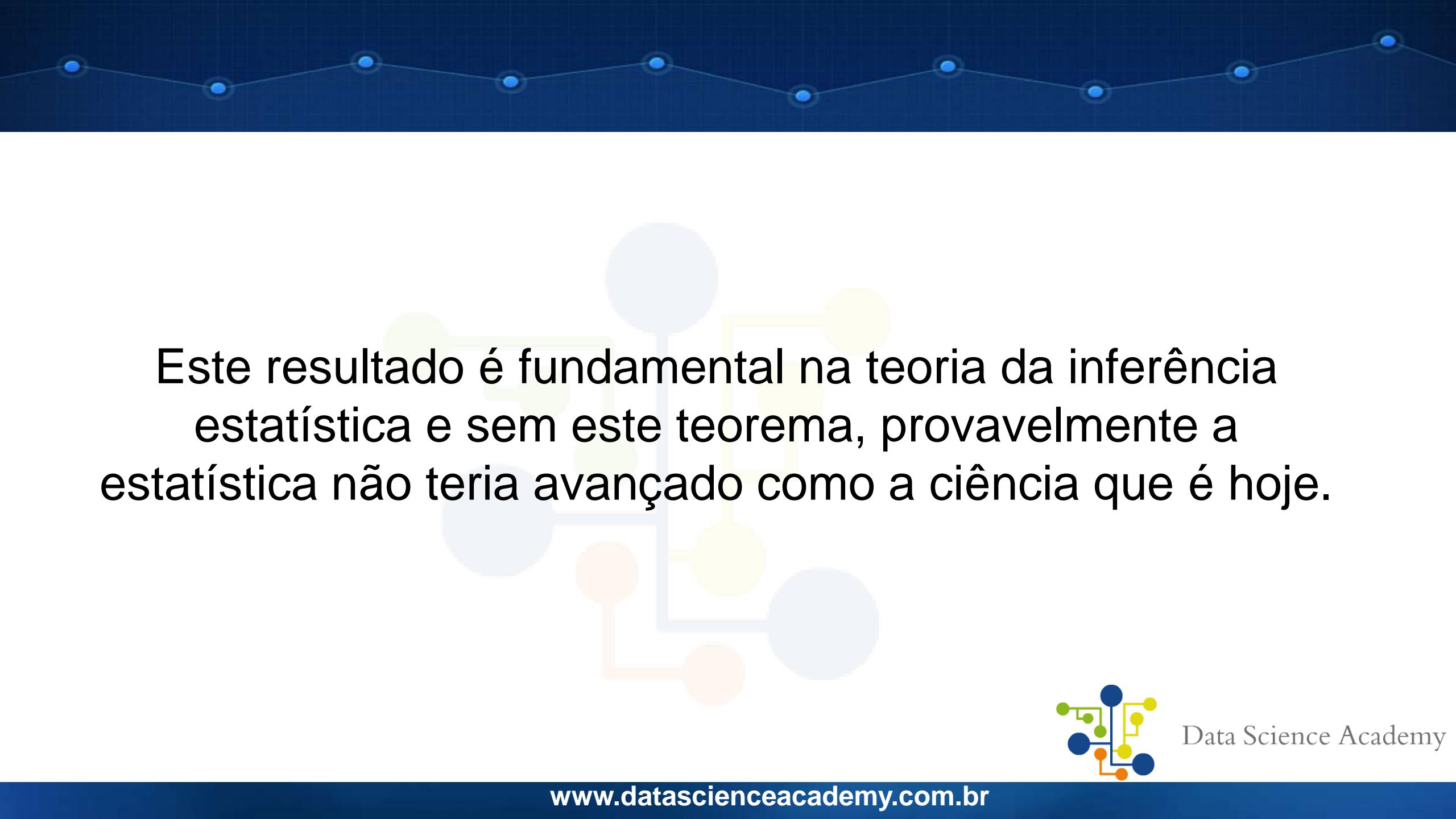
Data Science Academy



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.



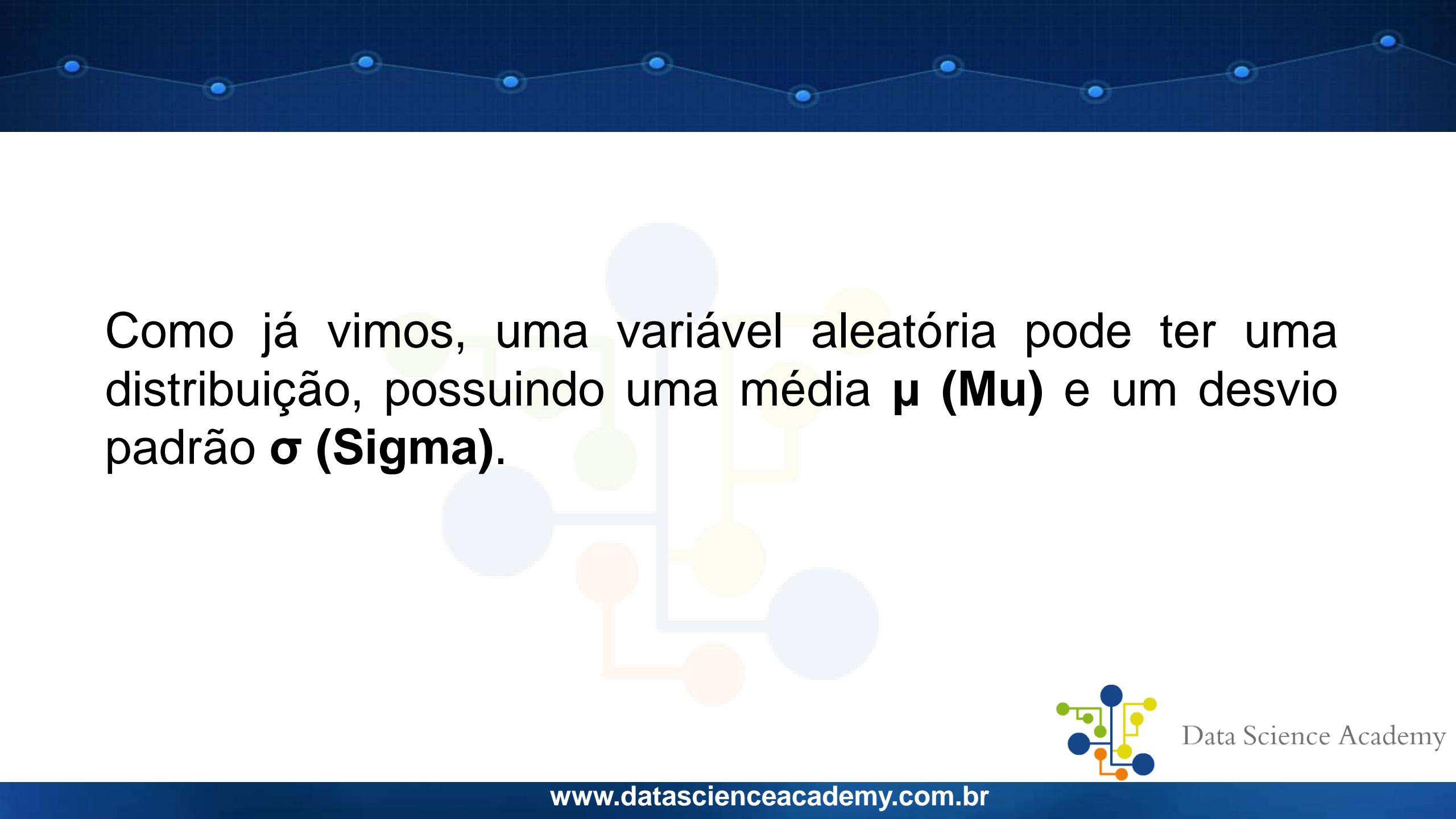
Data Science Academy



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



Data Science Academy



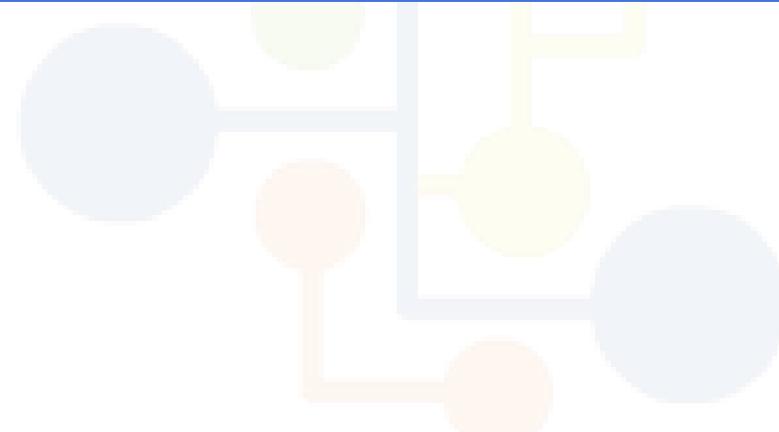
Como já vimos, uma variável aleatória pode ter uma distribuição, possuindo uma média μ (**Mu**) e um desvio padrão σ (**Sigma**).



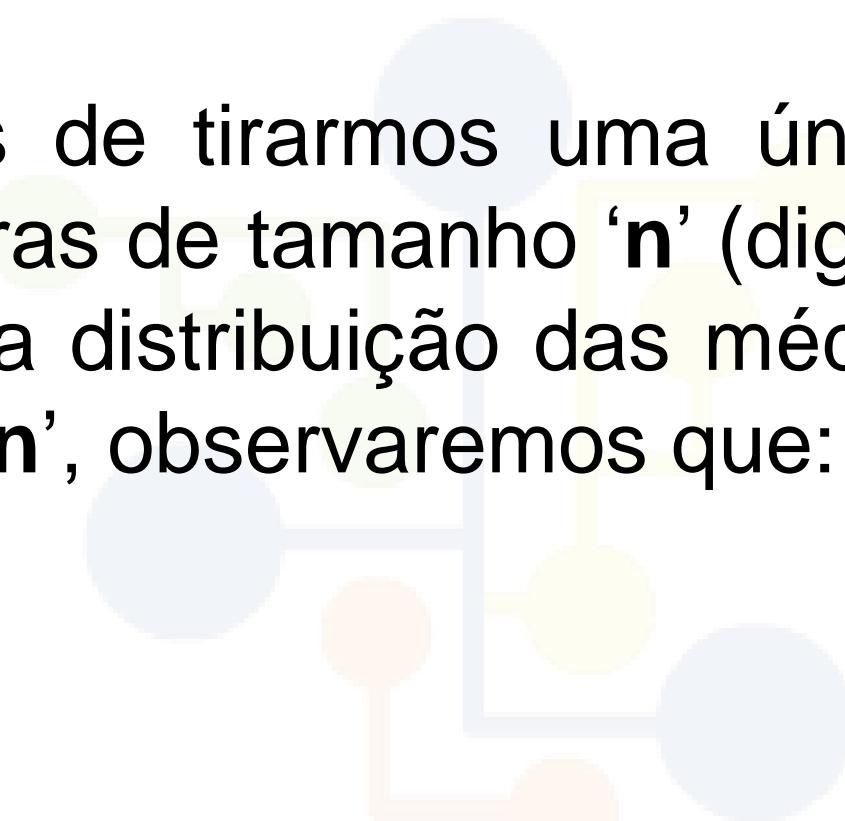
Data Science Academy



Entendendo melhor



Data Science Academy

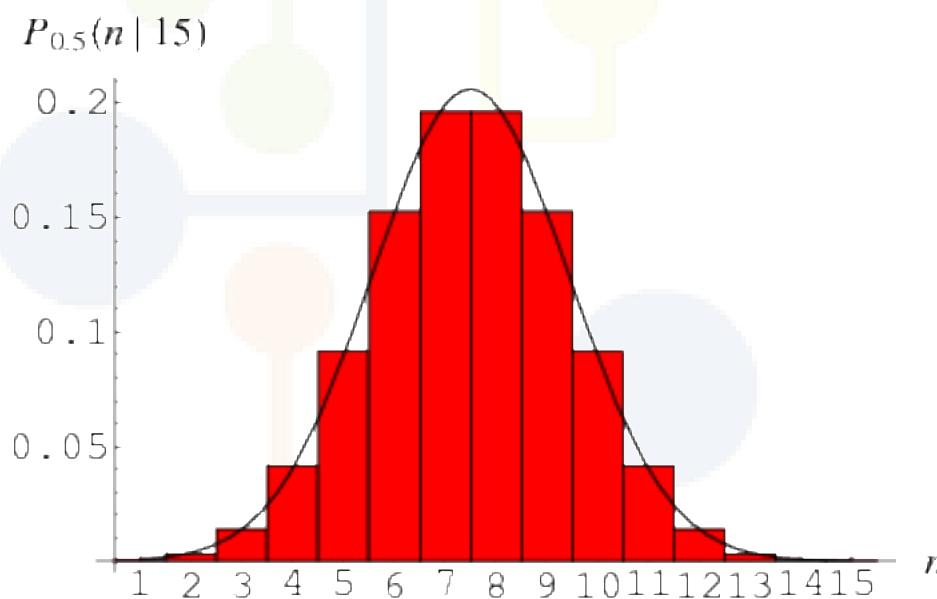


Se, ao invés de tirarmos uma única amostra tirarmos várias amostras de tamanho ' n ' (digamos 20 amostras) e analisarmos a distribuição das médias de cada amostra de tamanho ' n ', observaremos que:



Data Science Academy

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



Data Science Academy

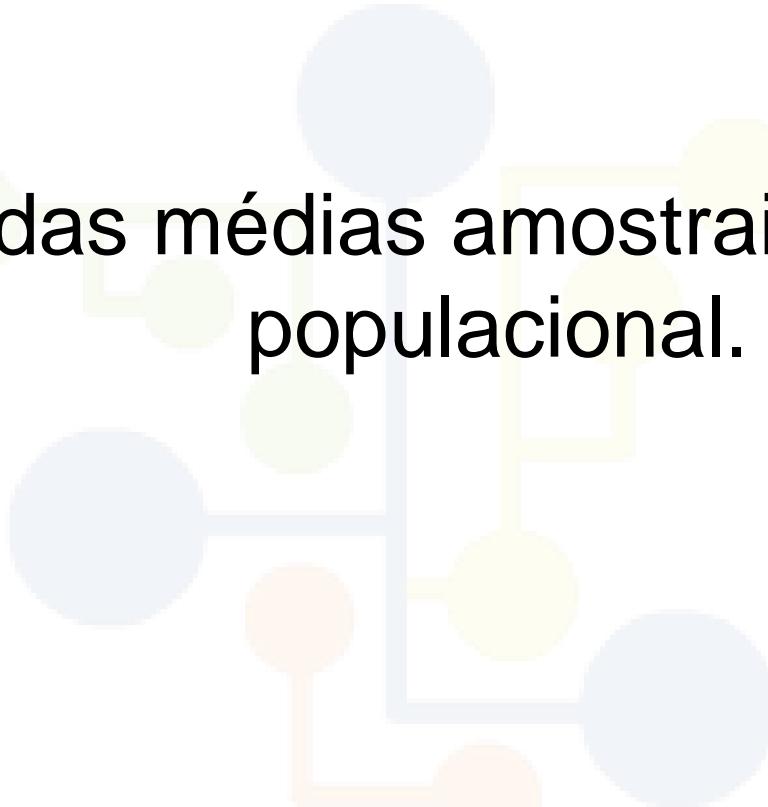
Para $n \geq 30$, a distribuição das médias amostrais pode ser aproximada satisfatoriamente por uma distribuição normal.



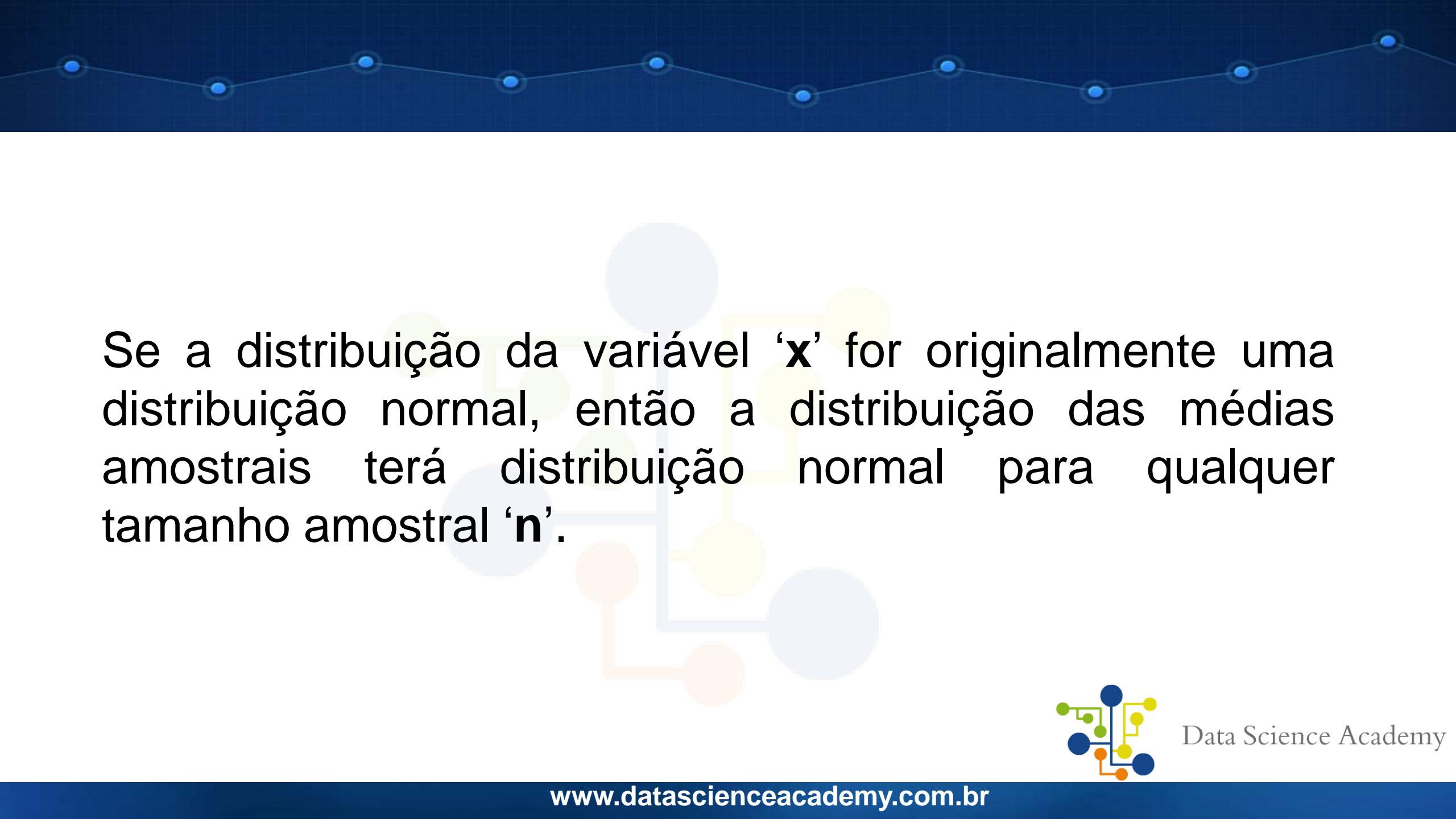
Data Science Academy



A média das médias amostrais será a média populacional.



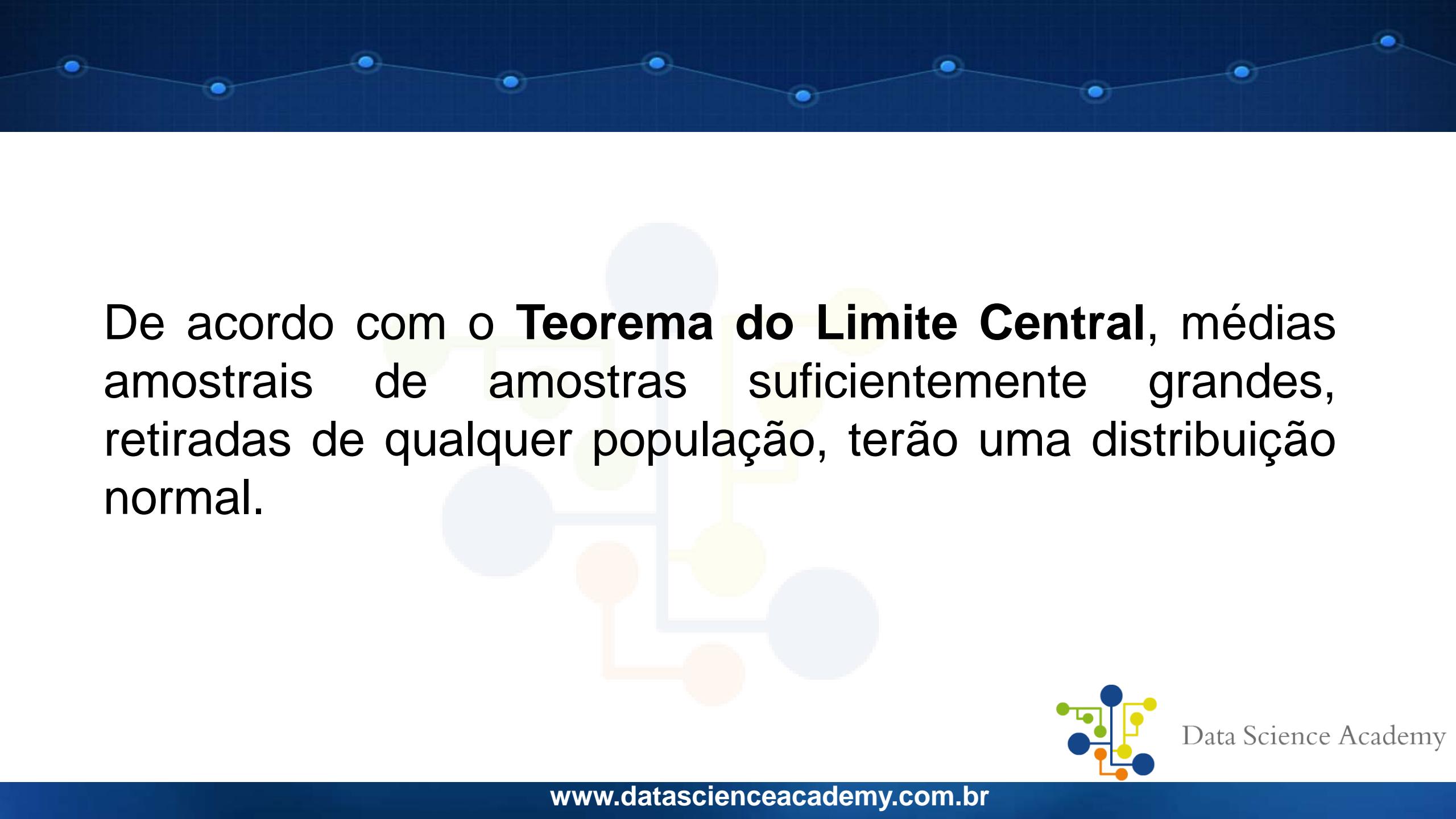
Data Science Academy



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’.



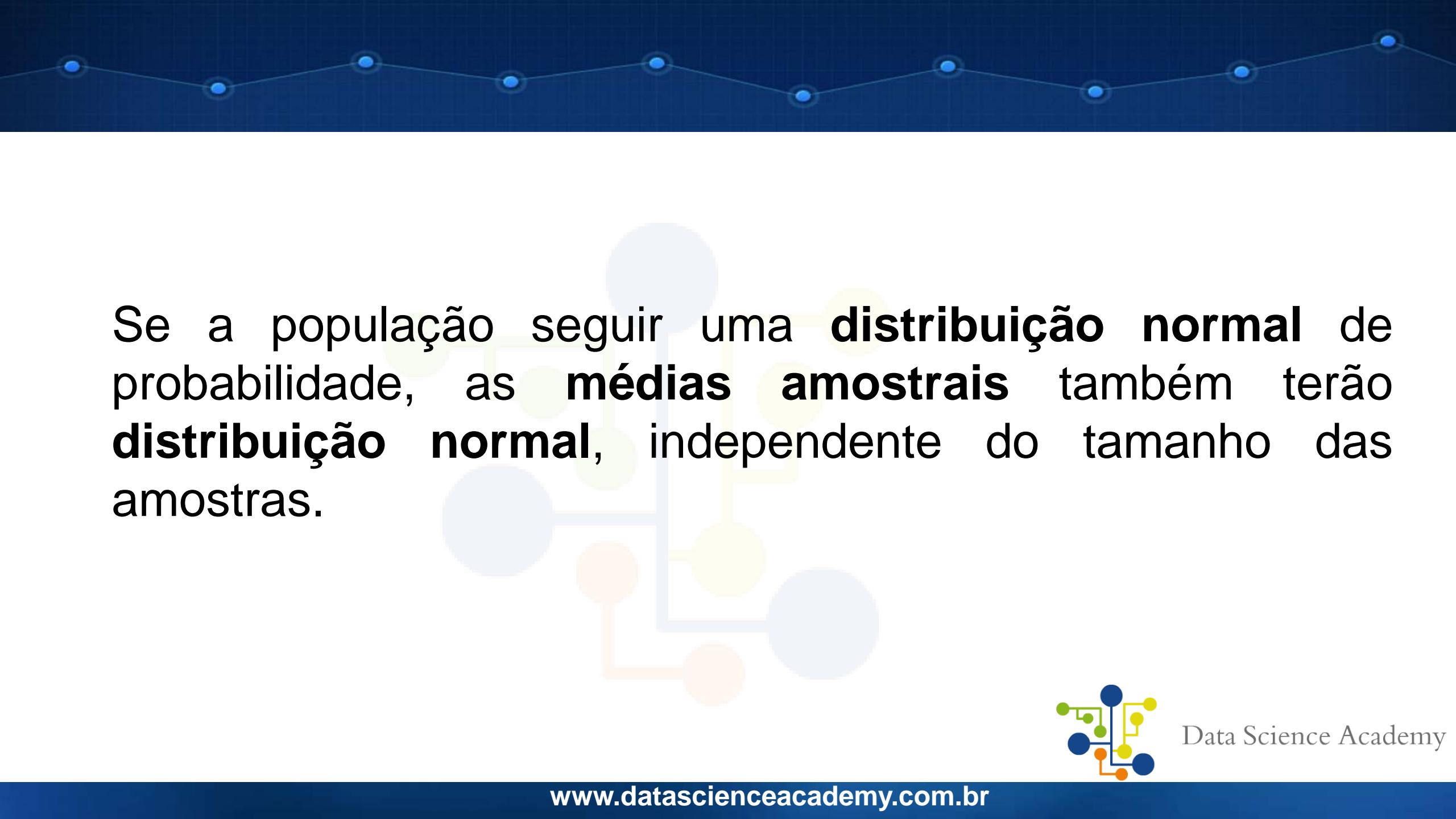
Data Science Academy



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



Data Science Academy



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



Data Science Academy

Exemplo



Data Science Academy

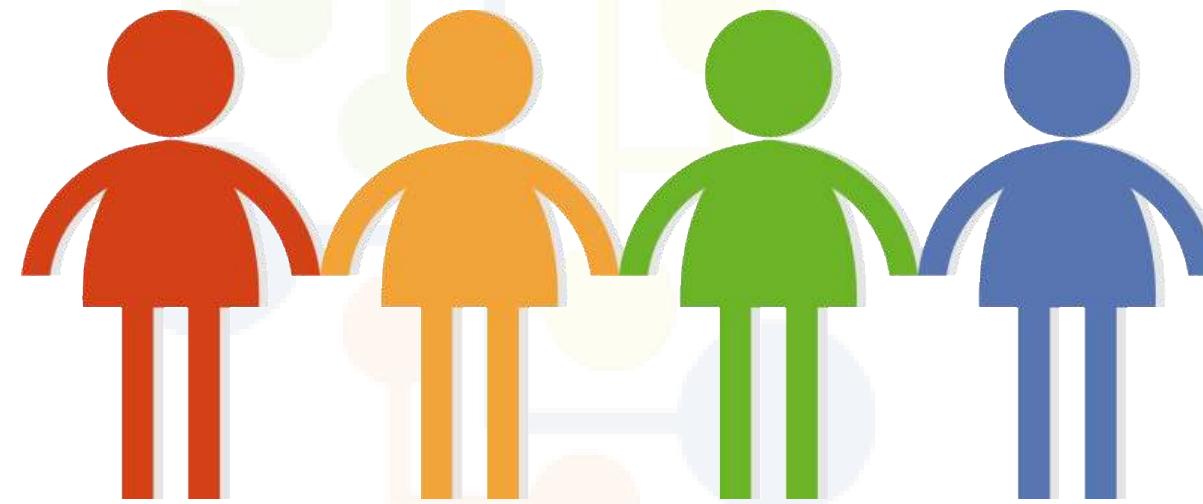


Imagine uma população de 4 pessoas, sendo então $N = 4$.



Data Science Academy

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.



Data Science Academy



➤ População $N = 4$

➤ Variável aleatória x

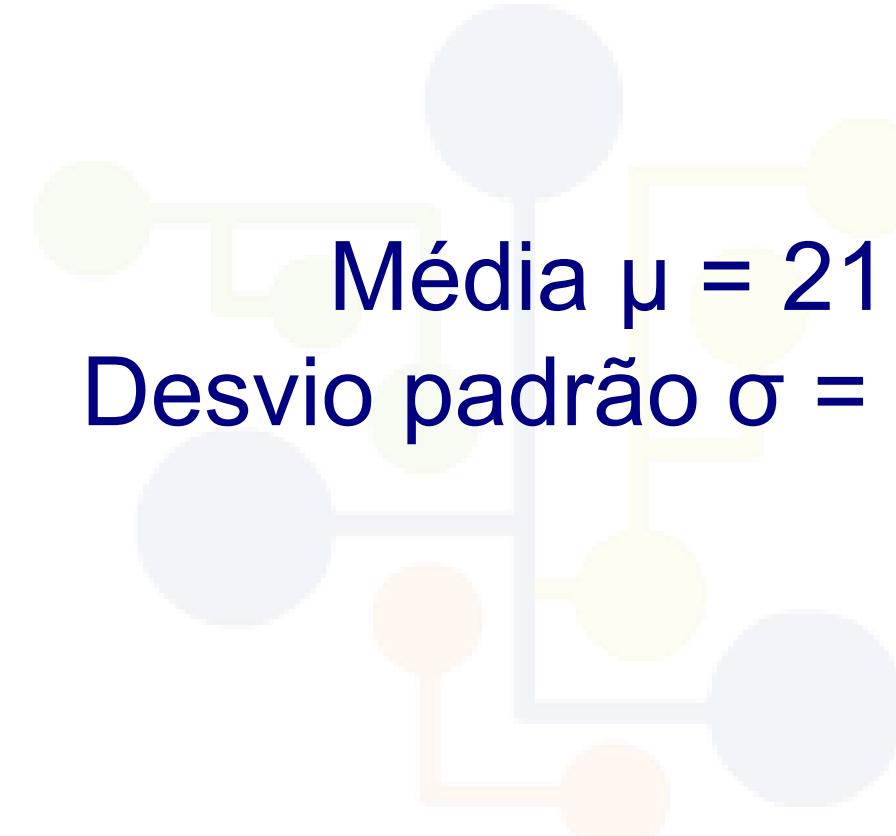
➤ Valores de x :
Idades = 18, 20, 22 e 24 anos



Data Science Academy



Vamos calcular a média e o desvio padrão desta população:

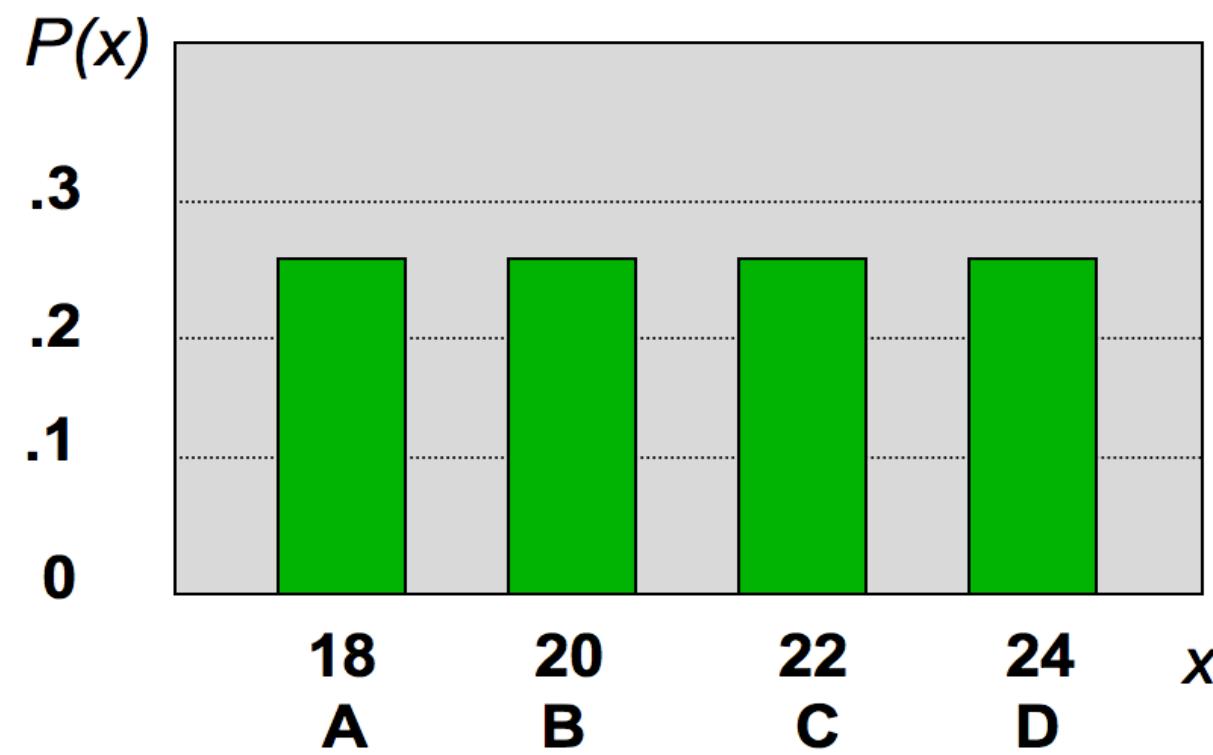


Média $\mu = 21$
Desvio padrão $\sigma = 2.236$

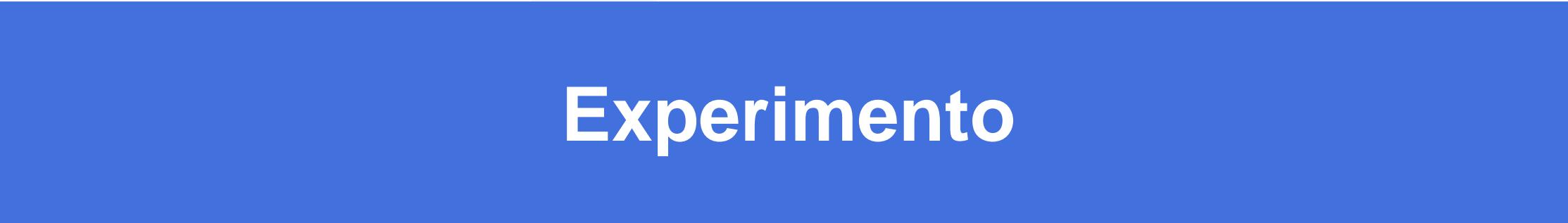


Data Science Academy

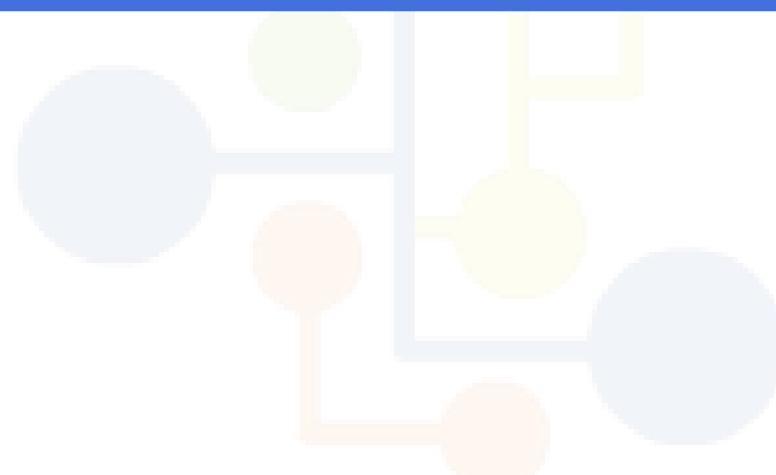
A distribuição de probabilidade desta população é uniforme



Data Science Academy



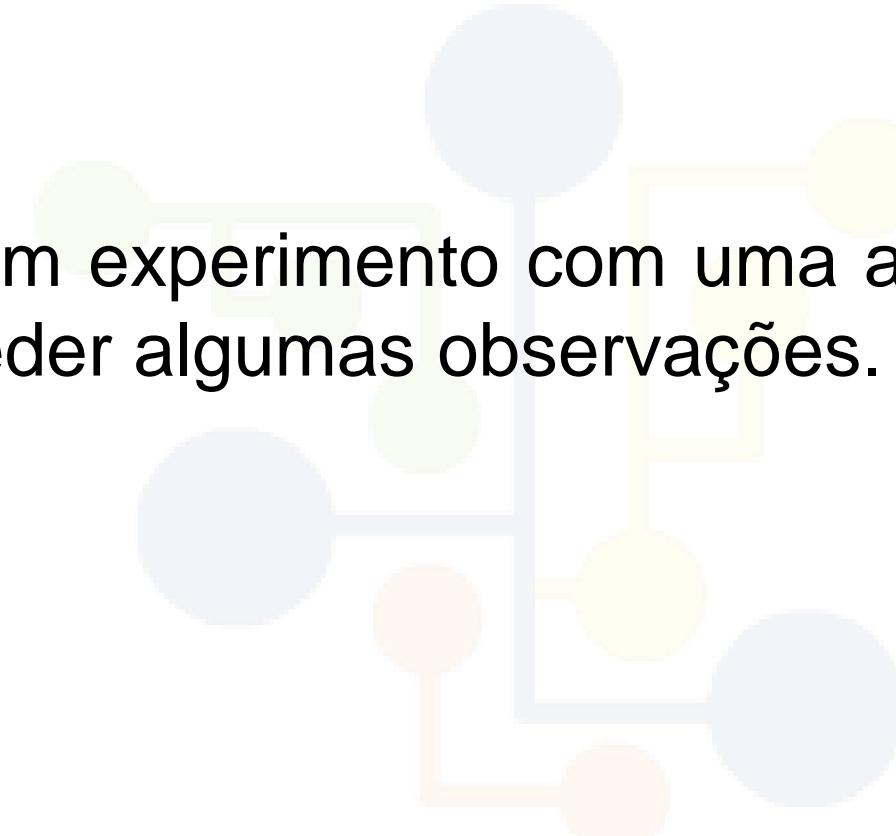
Experimento



Data Science Academy



Vamos fazer um experimento com uma amostra de 2 elementos e vamos proceder algumas observações.



Data Science Academy

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 médias amostrais

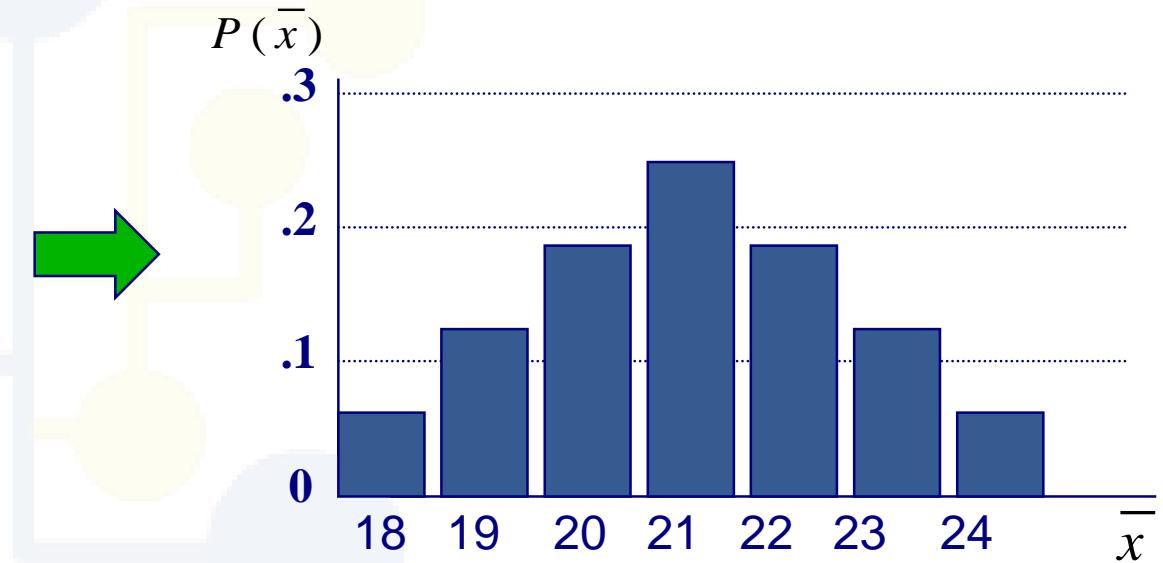
16 amostras possíveis
(amostragem com reposição)

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

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



Data Science Academy

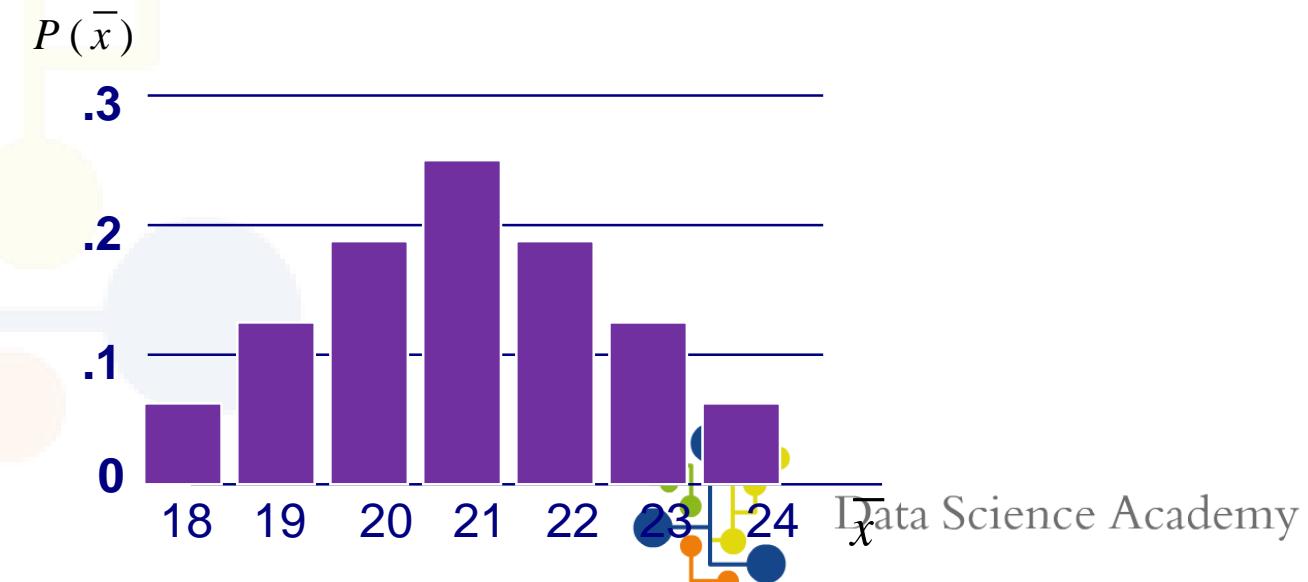
População $N = 4$

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



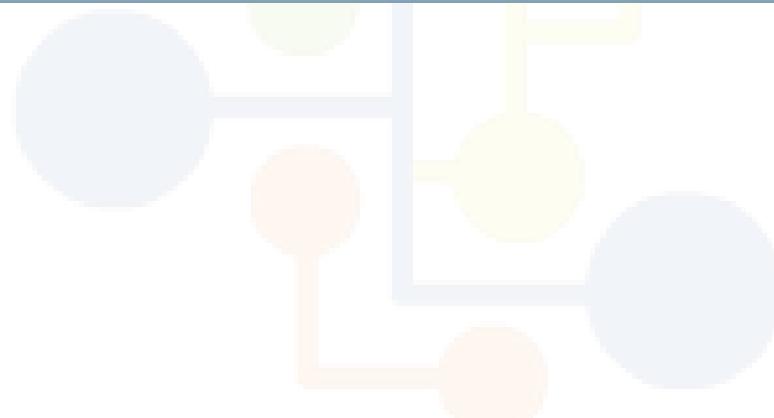
Distribuição amostral da média, $n = 2$

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





Lembre-se dessas Regras



Data Science Academy

1^a

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.

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



Data Science Academy

2^a

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.

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

Também chamado de erro padrão da média.



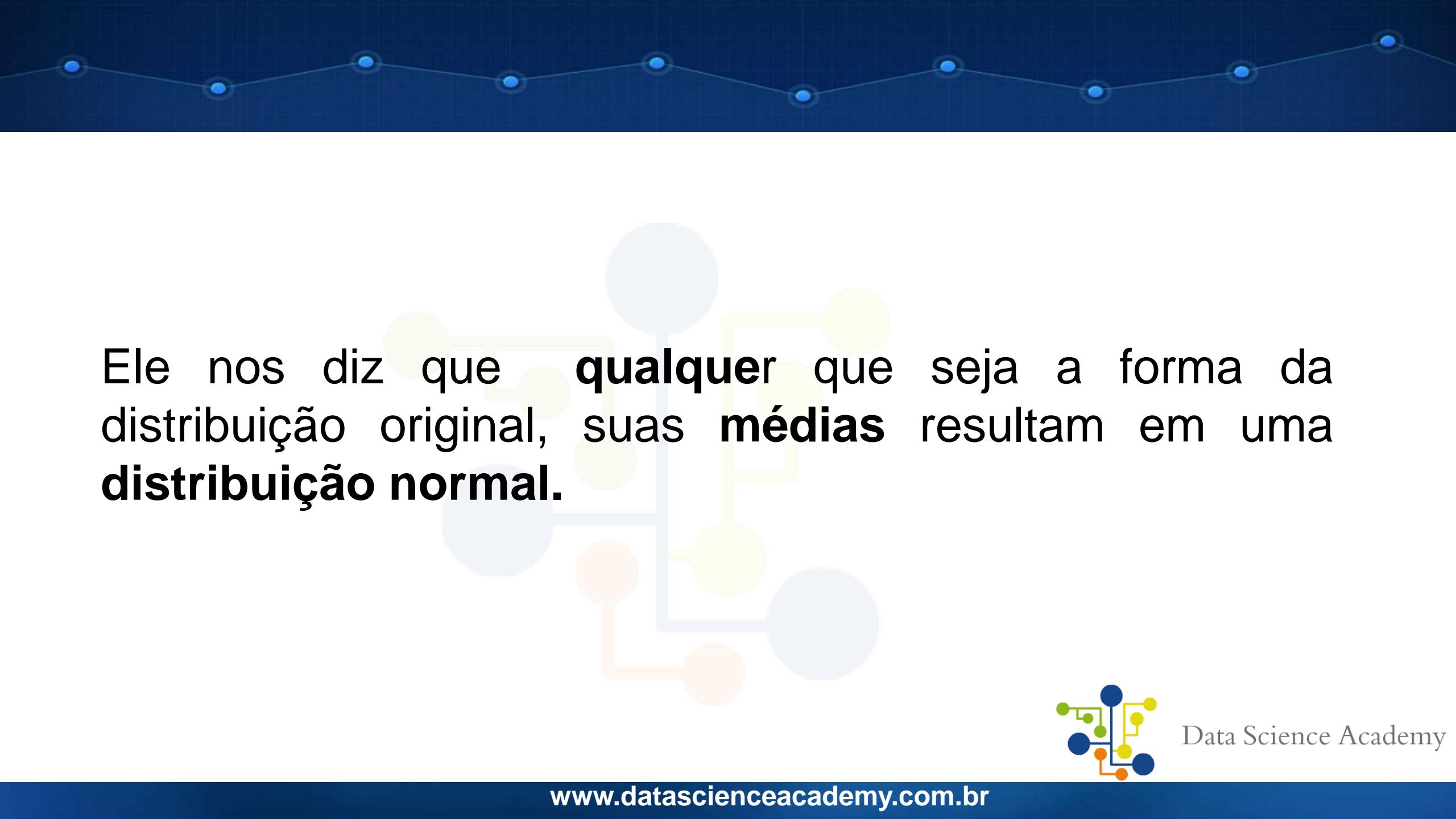
Data Science Academy



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



Data Science Academy

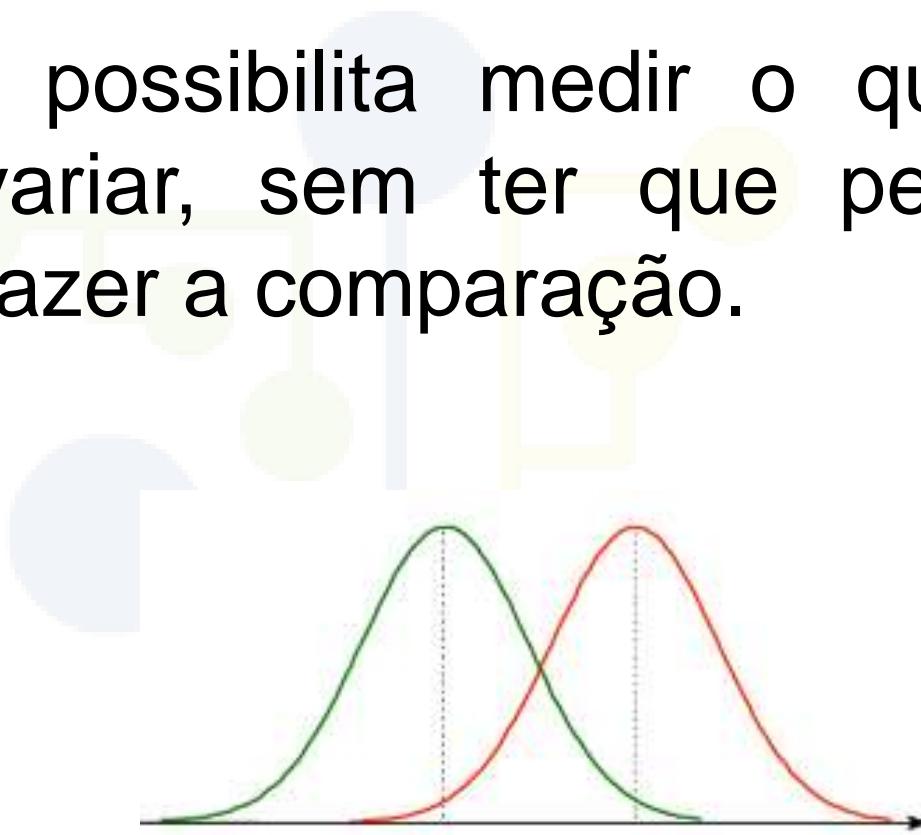


Ele nos diz que **qualquer** que seja a forma da distribuição original, suas **médias** resultam em uma **distribuição normal**.

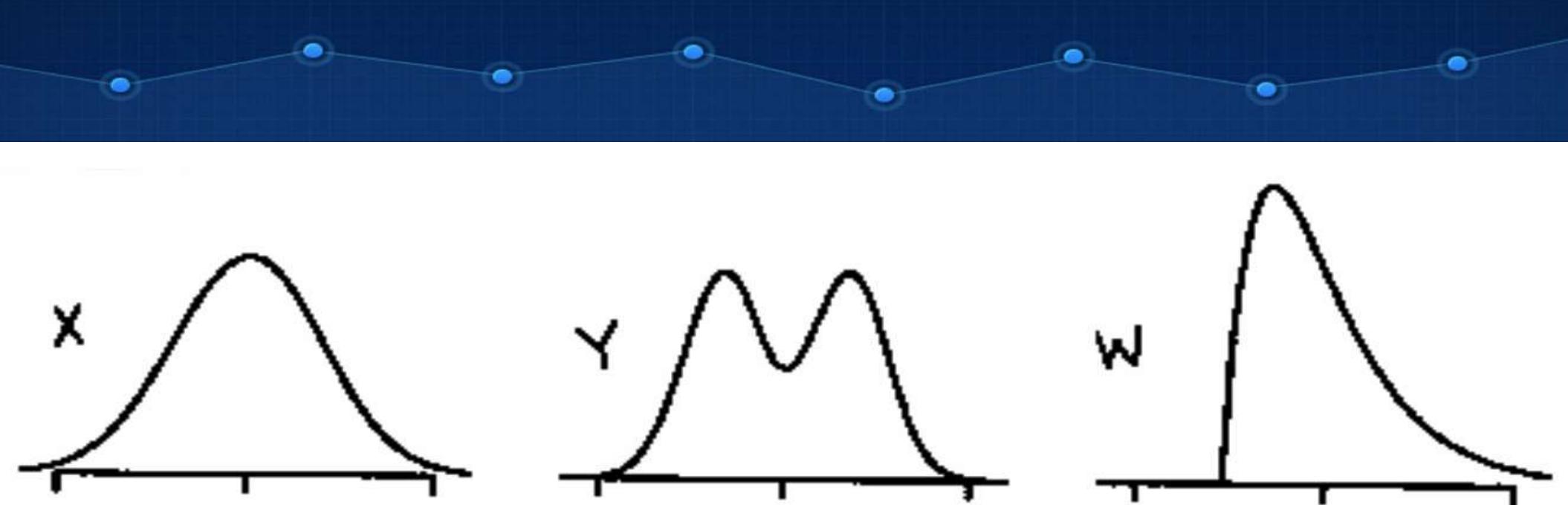


Data Science Academy

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



Data Science Academy



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

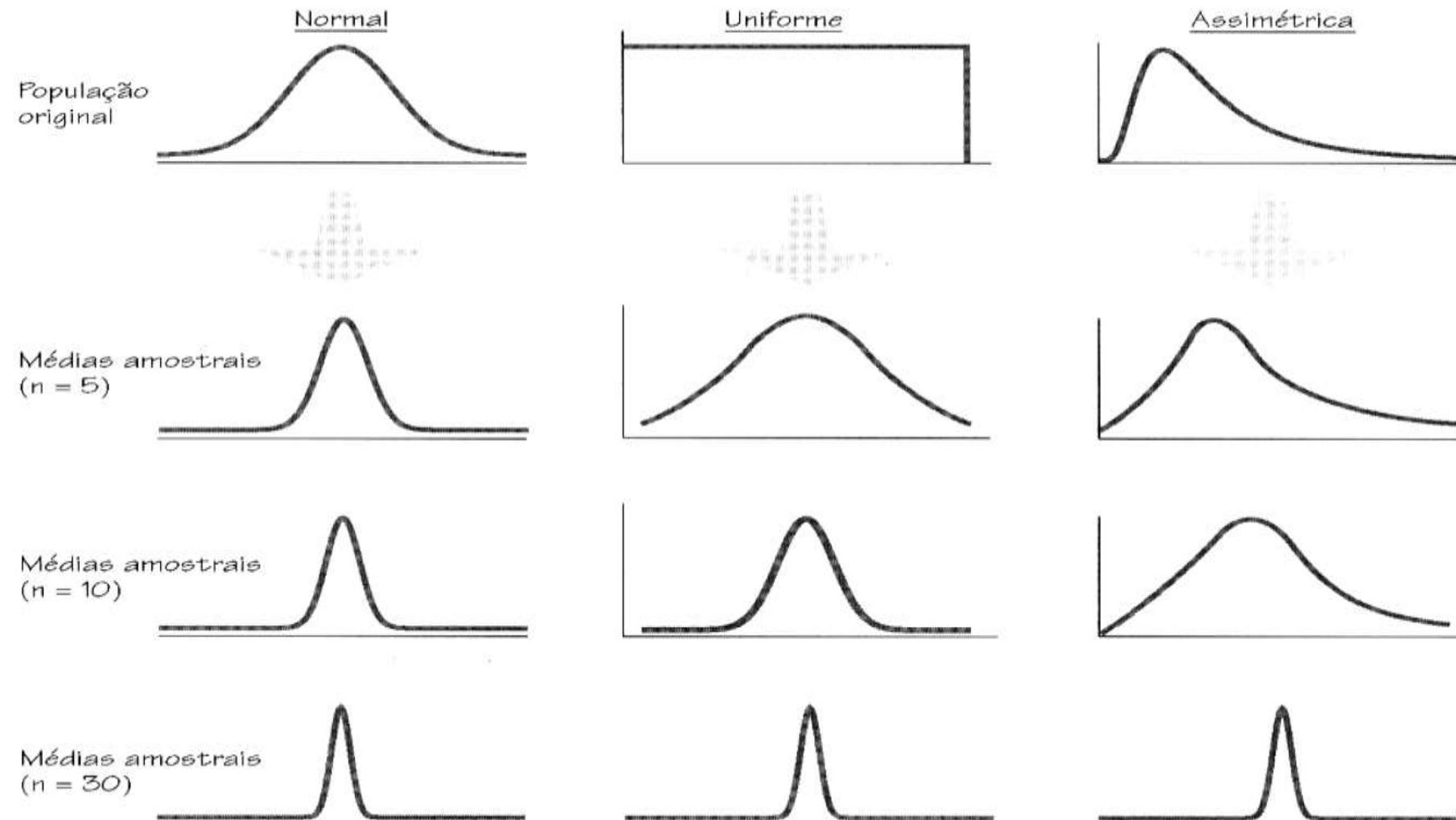


Data Science Academy

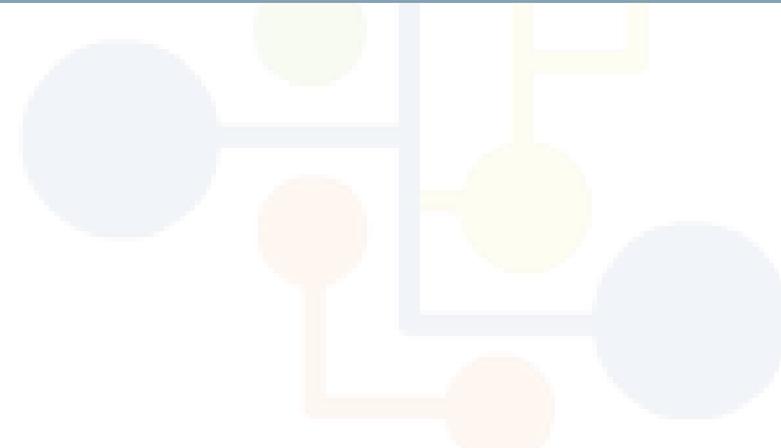
Conclusão



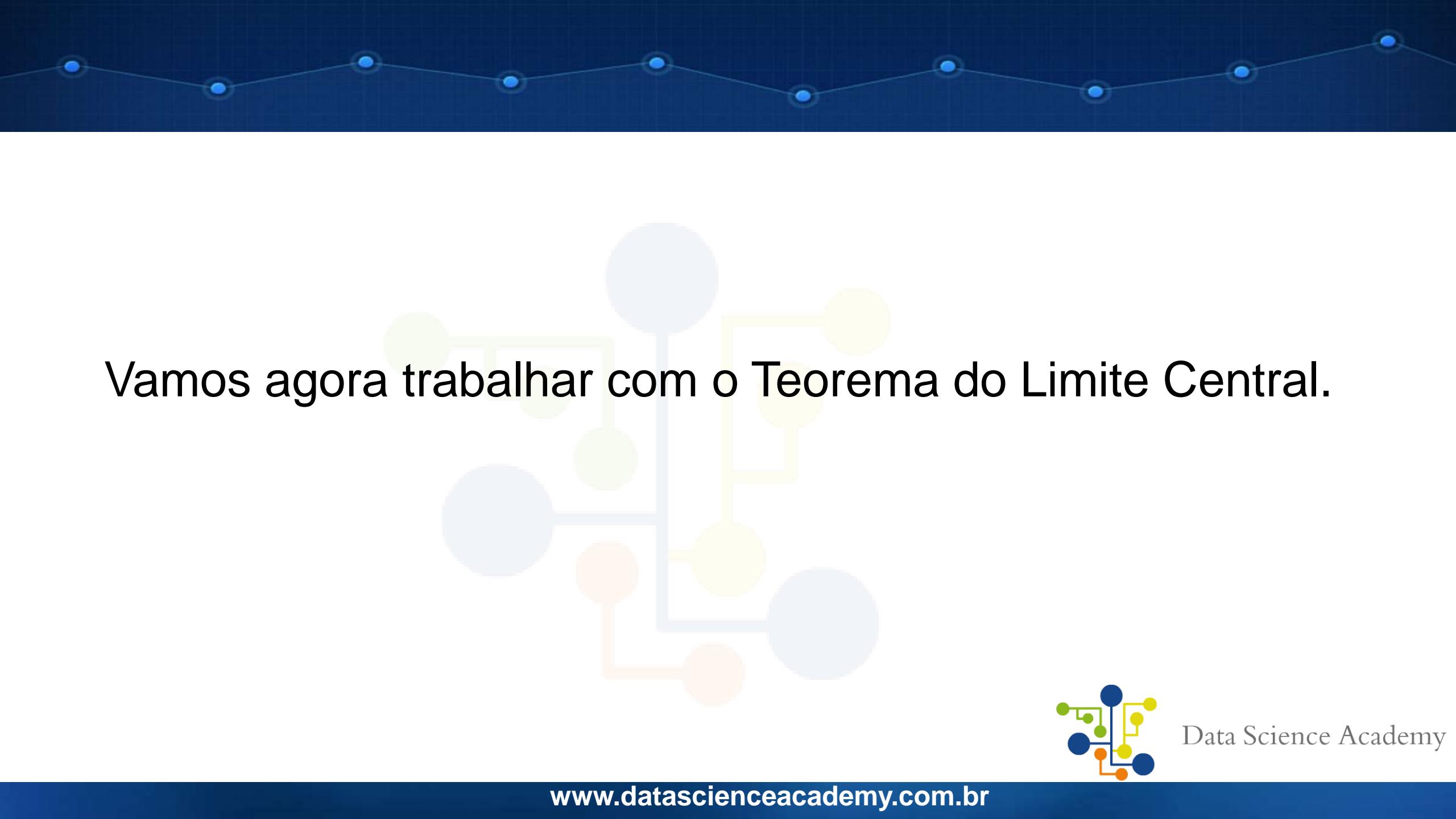
Data Science Academy



Colocando o Teorema do Limite Central para Trabalhar



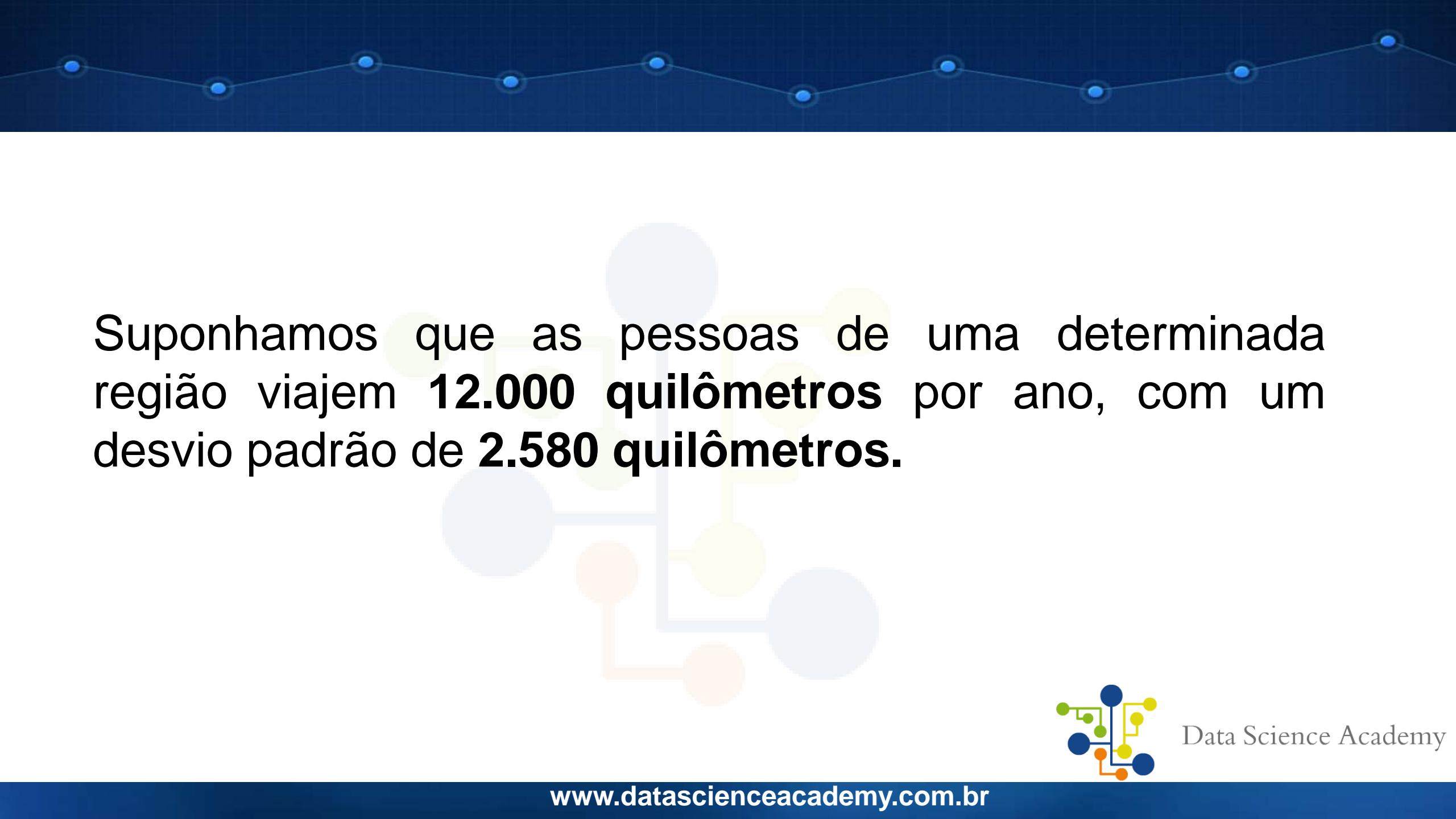
Data Science Academy



Vamos agora trabalhar com o Teorema do Limite Central.



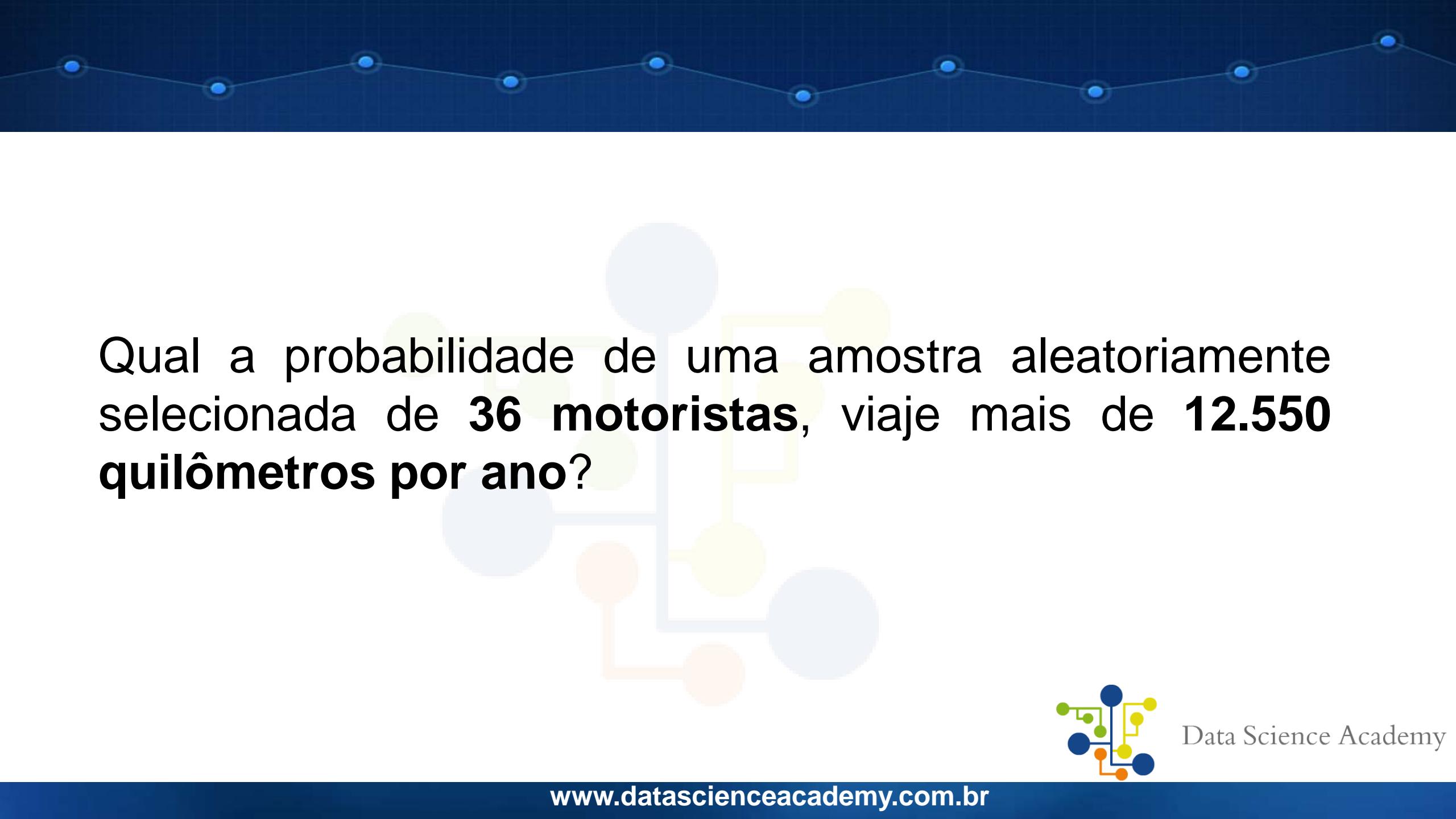
Data Science Academy



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



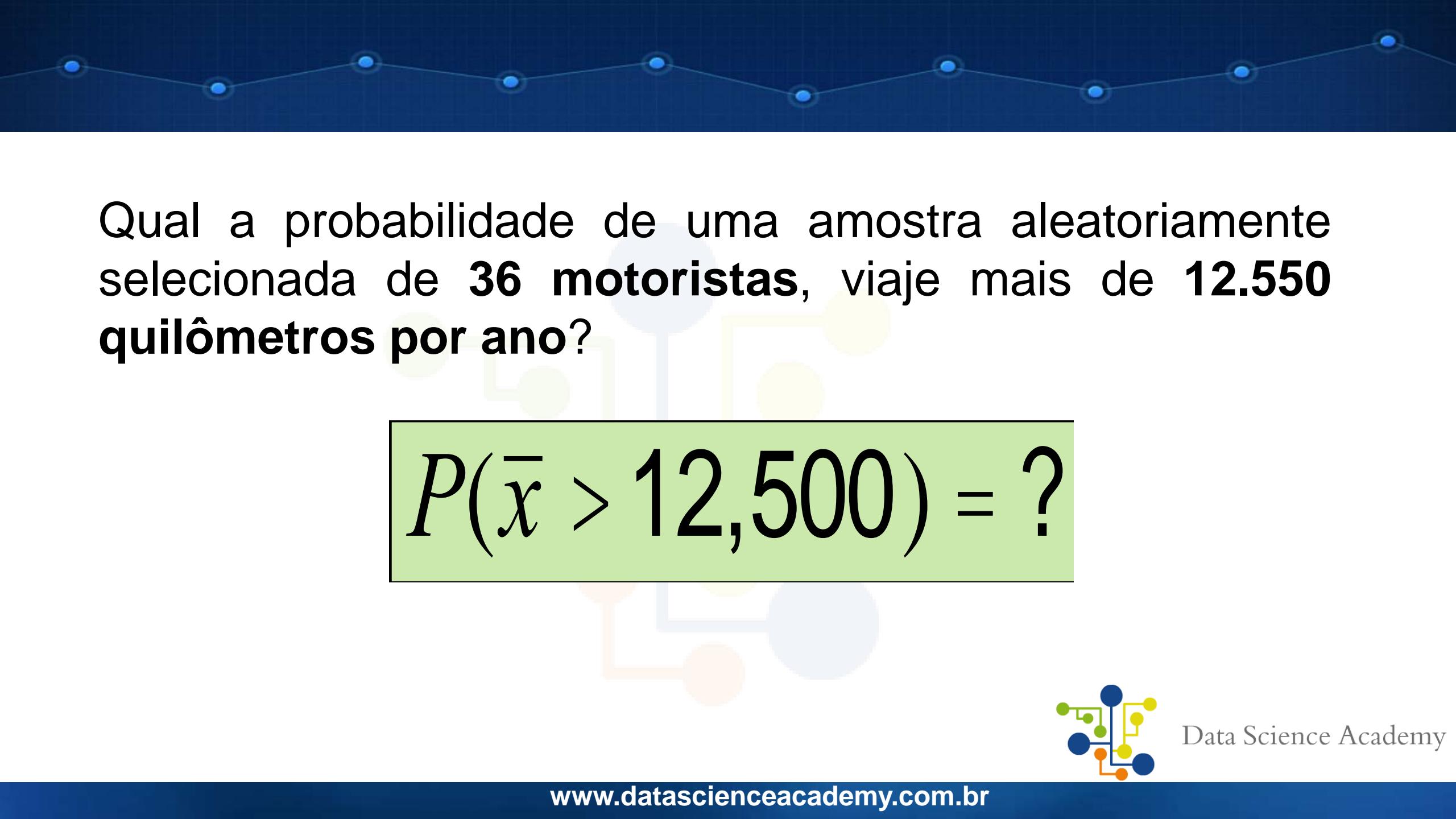
Data Science Academy



Qual a probabilidade de uma amostra aleatoriamente selecionada de **36 motoristas**, viaje mais de **12.550 quilômetros por ano?**



Data Science Academy

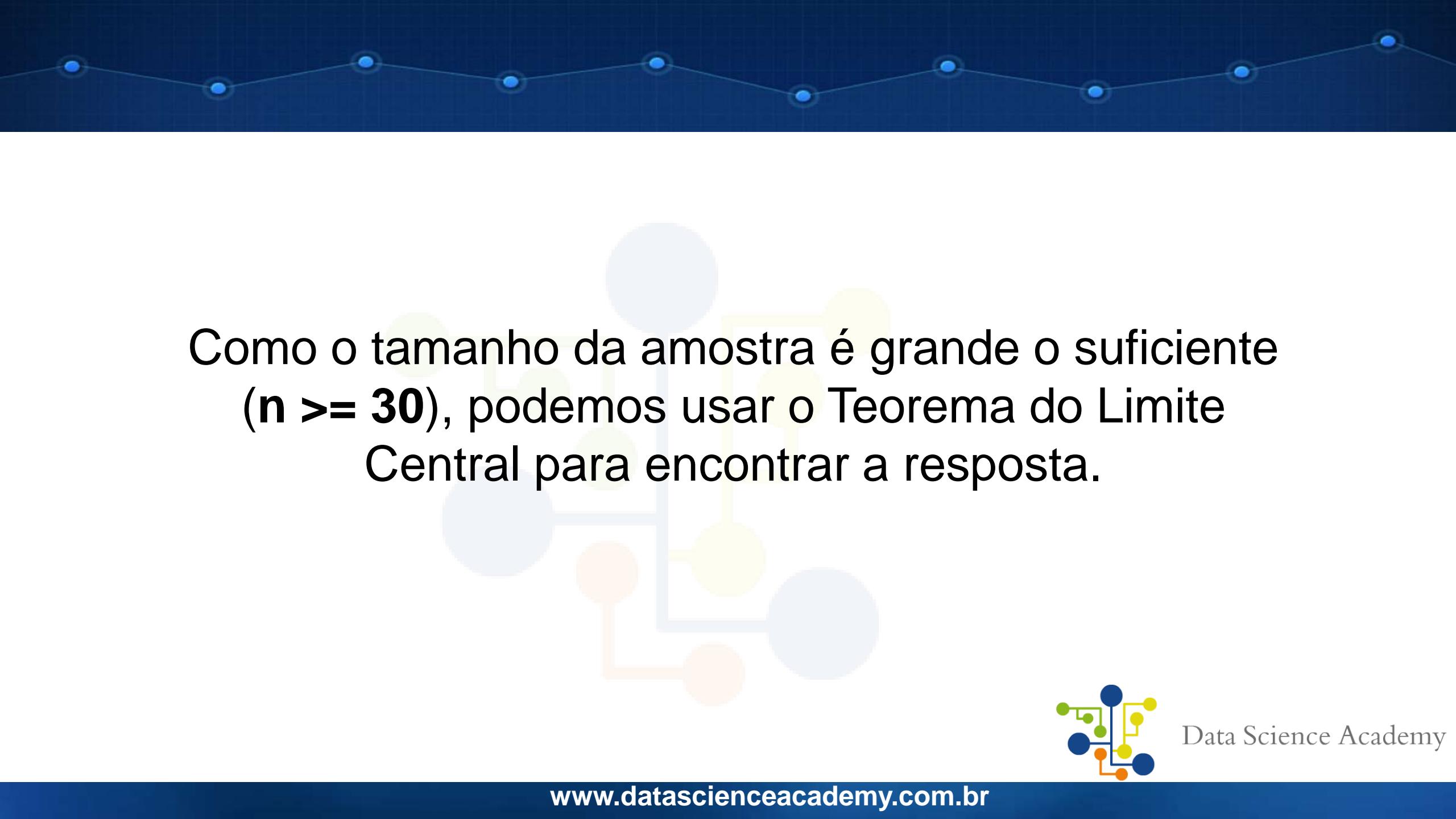


Qual a probabilidade de uma amostra aleatoriamente selecionada de **36 motoristas**, viaje mais de **12.550 quilômetros por ano?**

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



Data Science Academy



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



Data Science Academy

Aplicando as regras vistas anteriormente, teremos:

Desvio padrão
da população

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

Desvio padrão
da amostra

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

Tamanho da
amostra



Data Science Academy

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

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



Data Science Academy

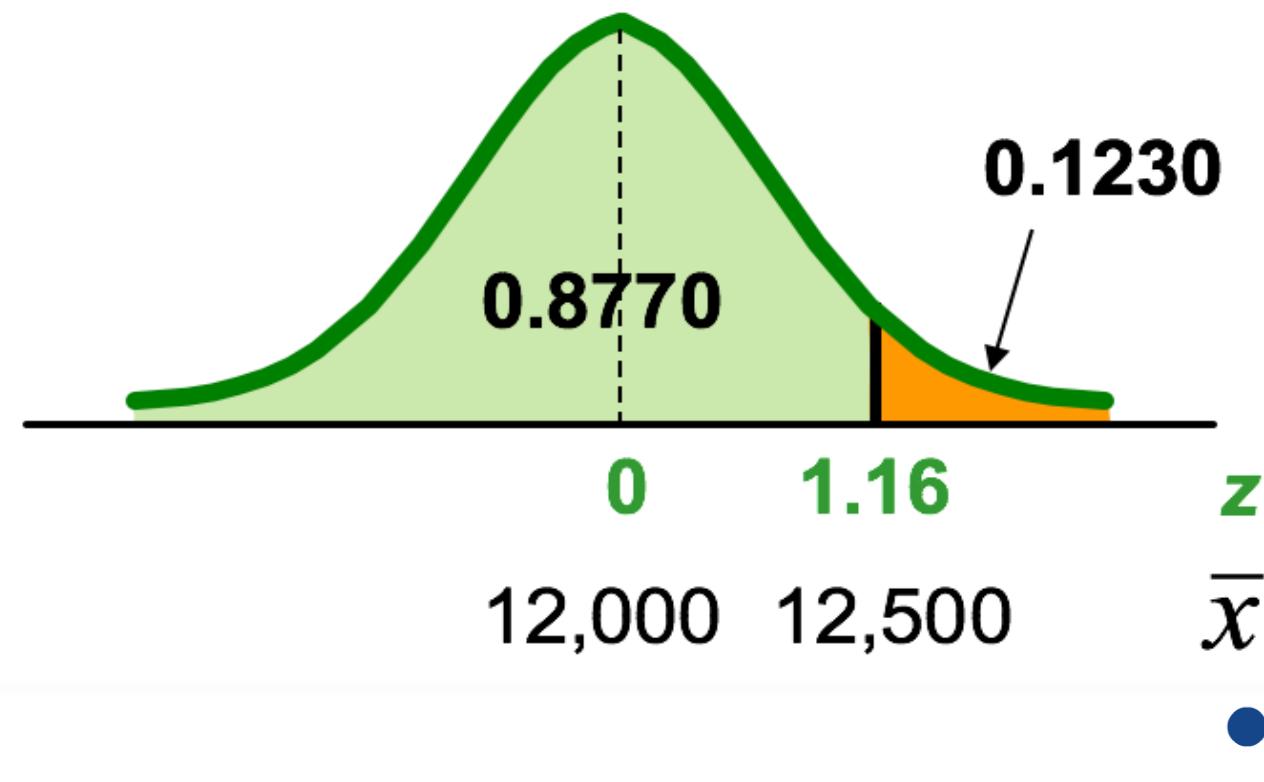
Vamos usar o **Escore_z** para nos ajudar:

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

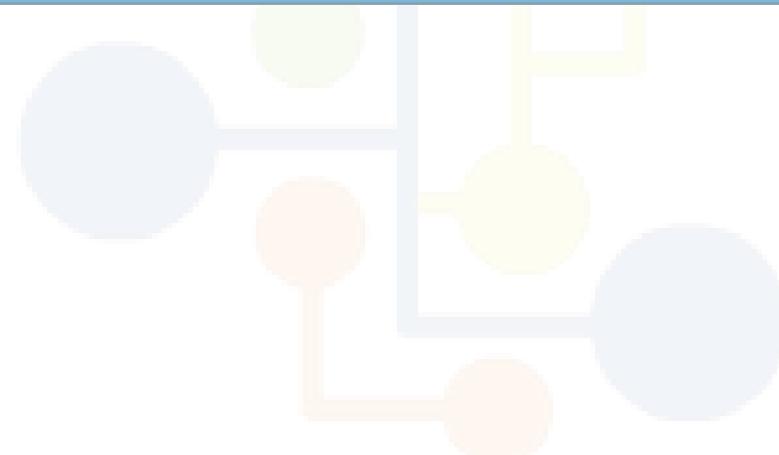


Data Science Academy

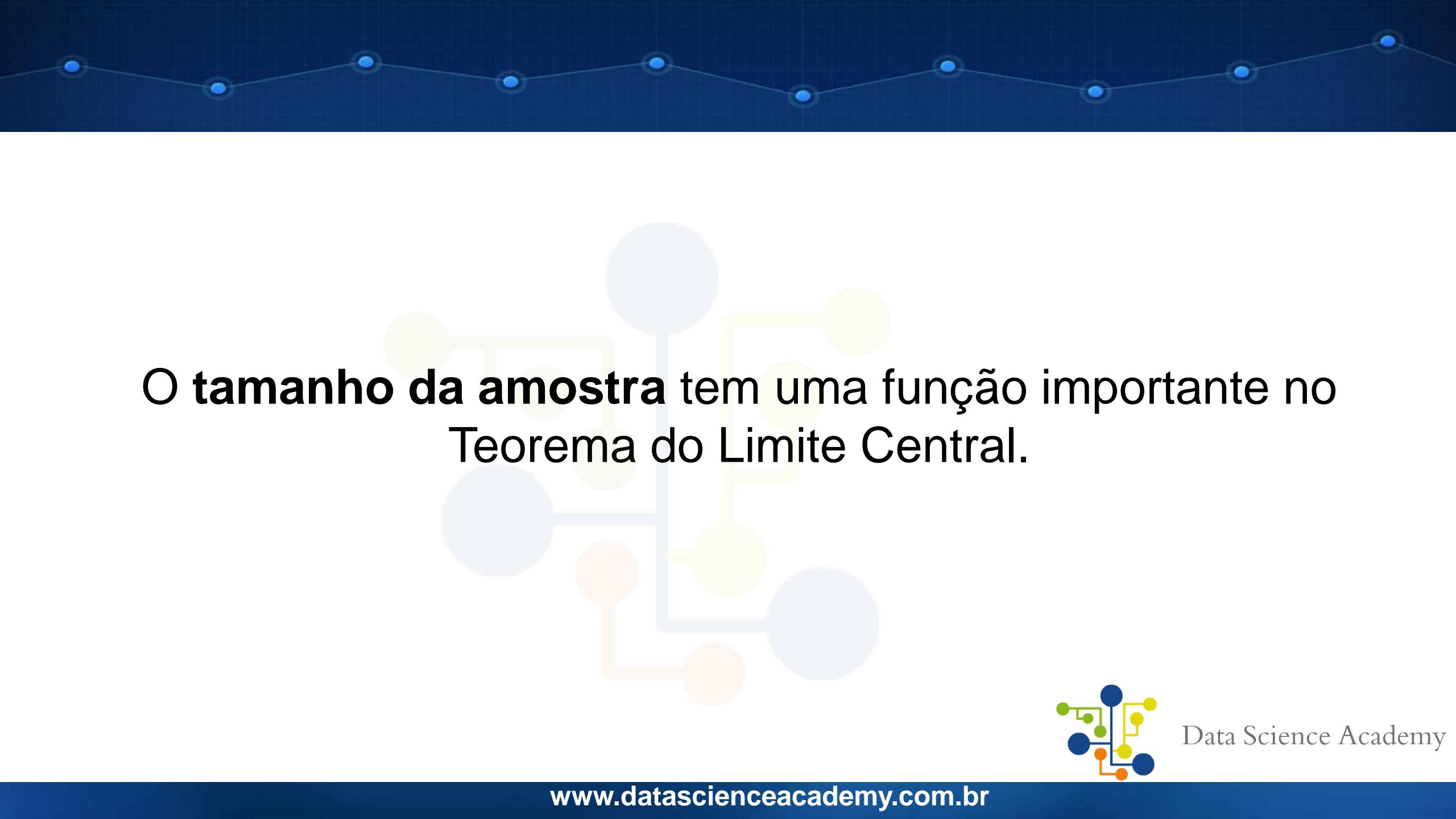
Resposta: A probabilidade de selecionarmos uma amostra aleatória de 36 motoristas que viajem 12.500 quilômetros, é de 12.3%.



Efeito do Tamanho da Amostra na Distribuição da Amostragem



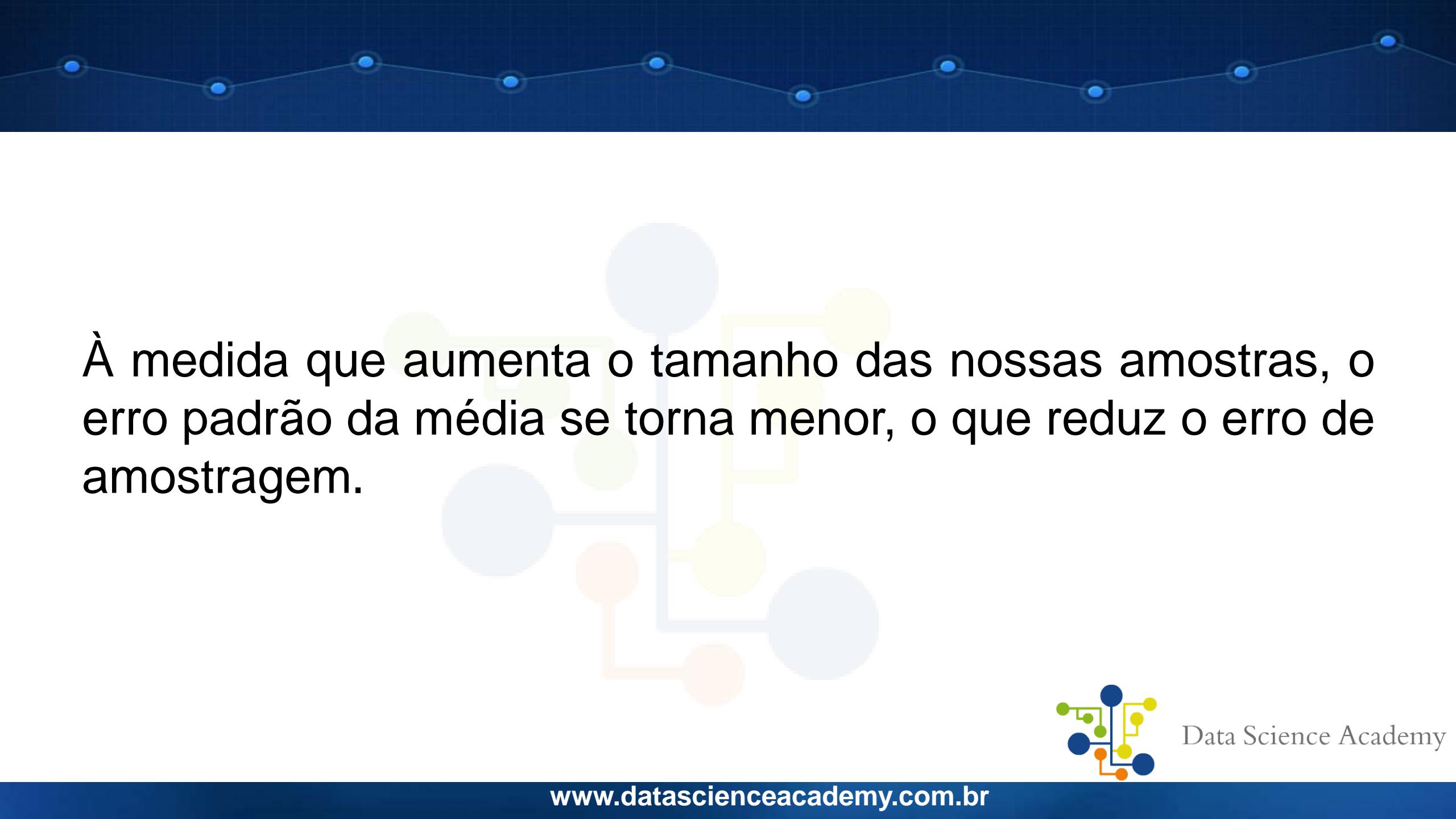
Data Science Academy



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



Data Science Academy



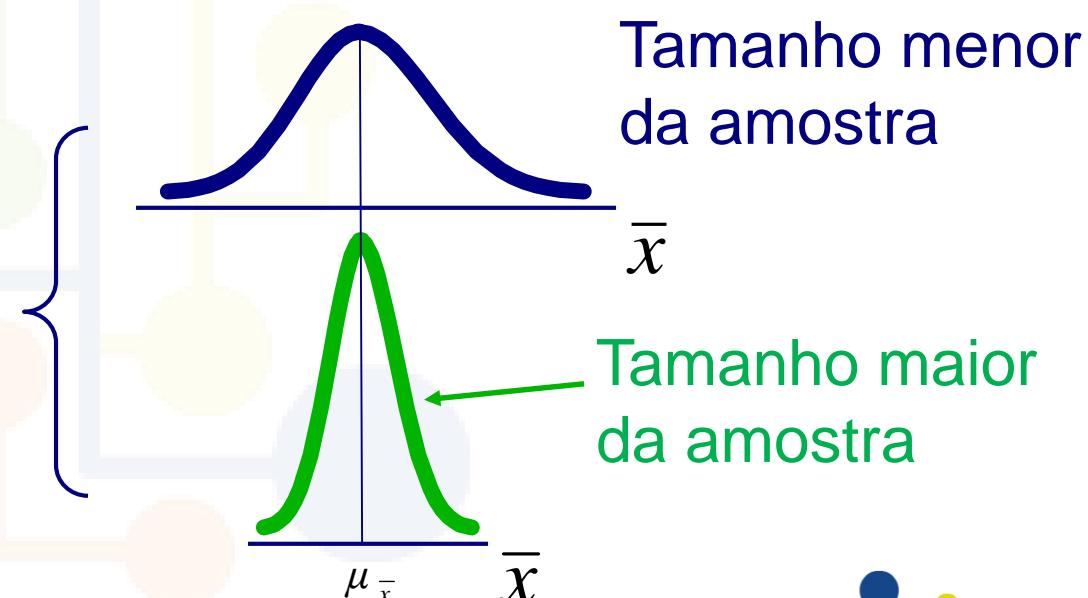
À medida que aumenta o tamanho das nossas amostras, o erro padrão da média se torna menor, o que reduz o erro de amostragem.



Data Science Academy

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

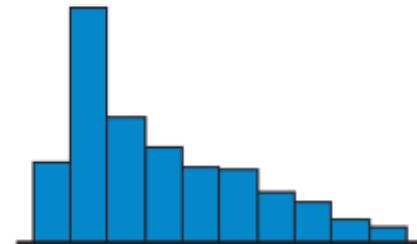
Aumentando o tamanho da amostra, reduz o erro padrão.



Data Science Academy

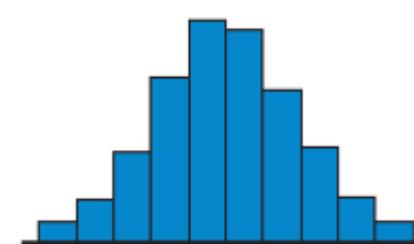
O formato da distribuição da população afetará o formato da distribuição da amostra, assim como o tamanho da amostra

Inclinada à direita



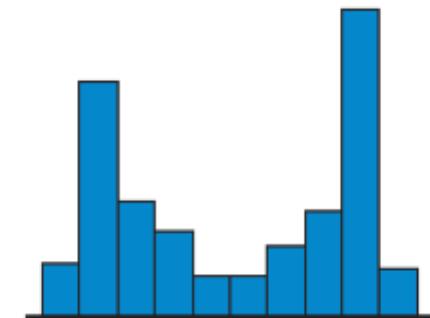
População 1

Normal



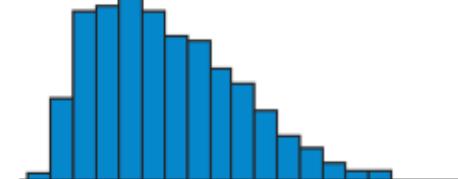
População 2

Em forma de U



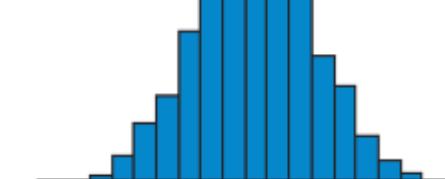
População 3

Inclinada à esquerda



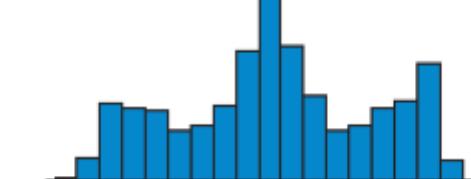
Média da Amostra
 $n=2$

Inclinada à esquerda



Média da Amostra
 $n=2$

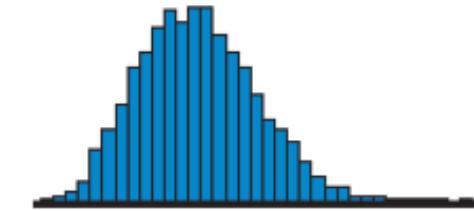
Inclinada à esquerda



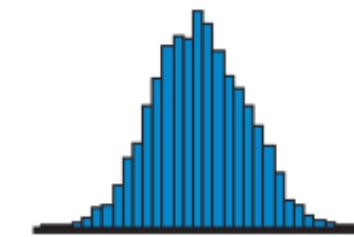
Média da Amostra
 $n=2$

Data Science Academy

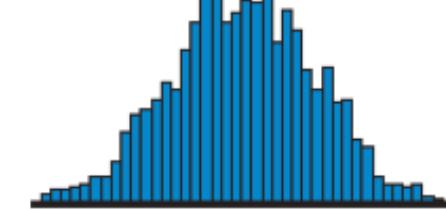
O formato da distribuição da população afetará o formato da distribuição da amostra, assim como o tamanho da amostra



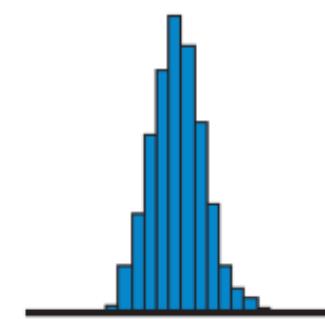
Média da Amostra
 $n=5$



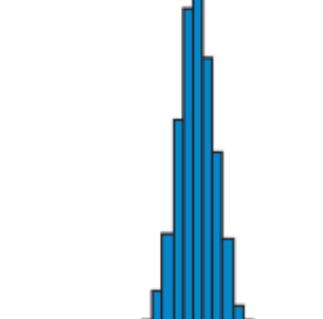
Média da Amostra
 $n=5$



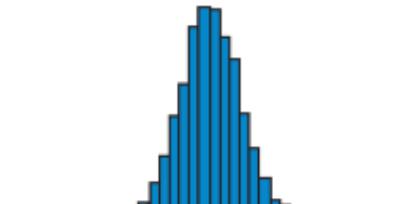
Média da Amostra
 $n=5$



Média da Amostra
 $n=30$



Média da Amostra
 $n=30$



Média da Amostra
 $n=30$



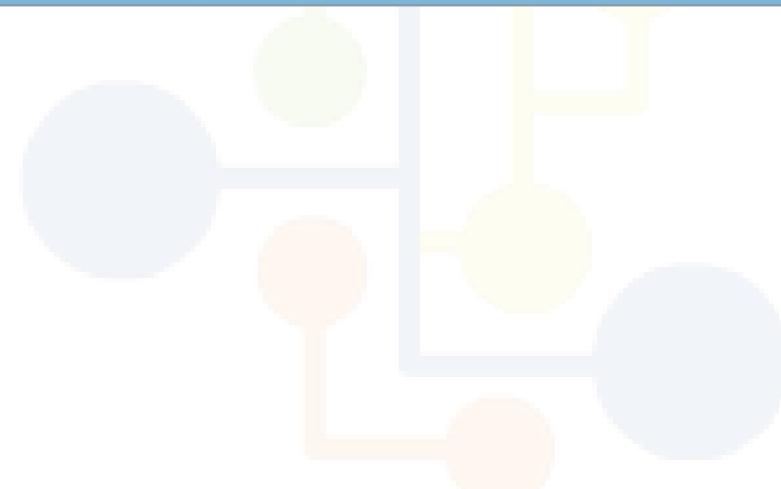
Data Science Academy

Esse tópico chegou ao final

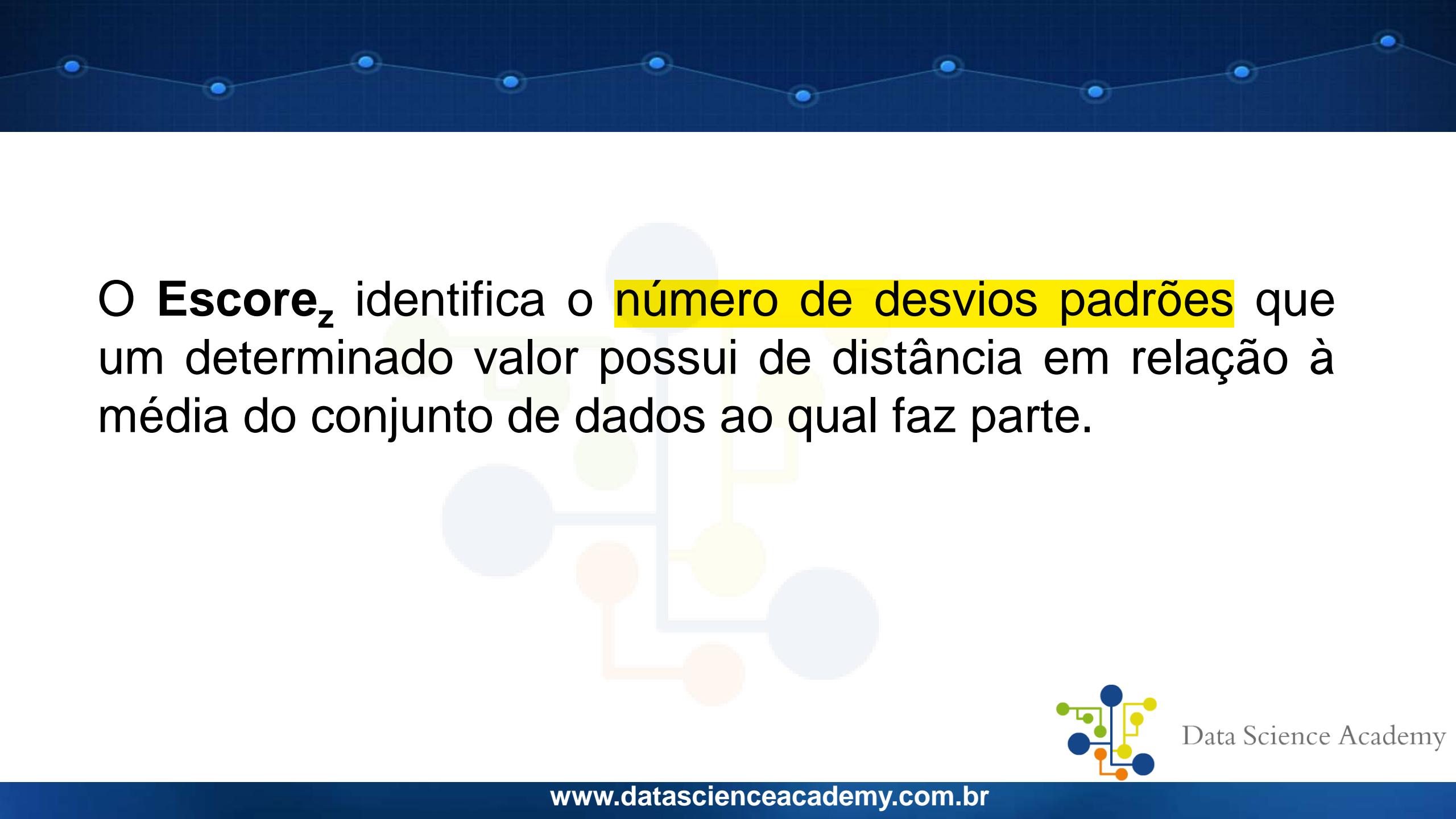


Data Science Academy

Escore z



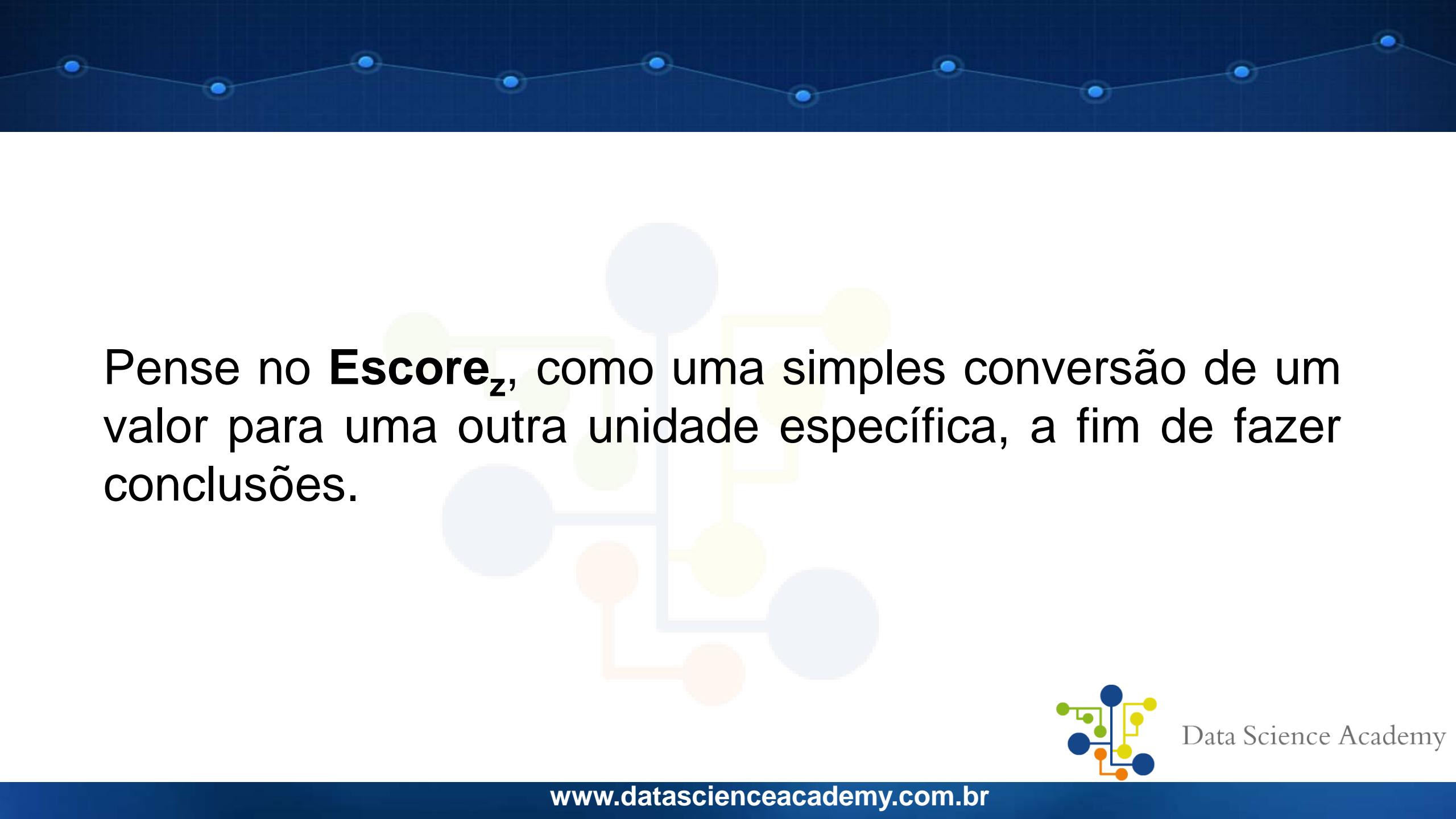
Data Science Academy



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



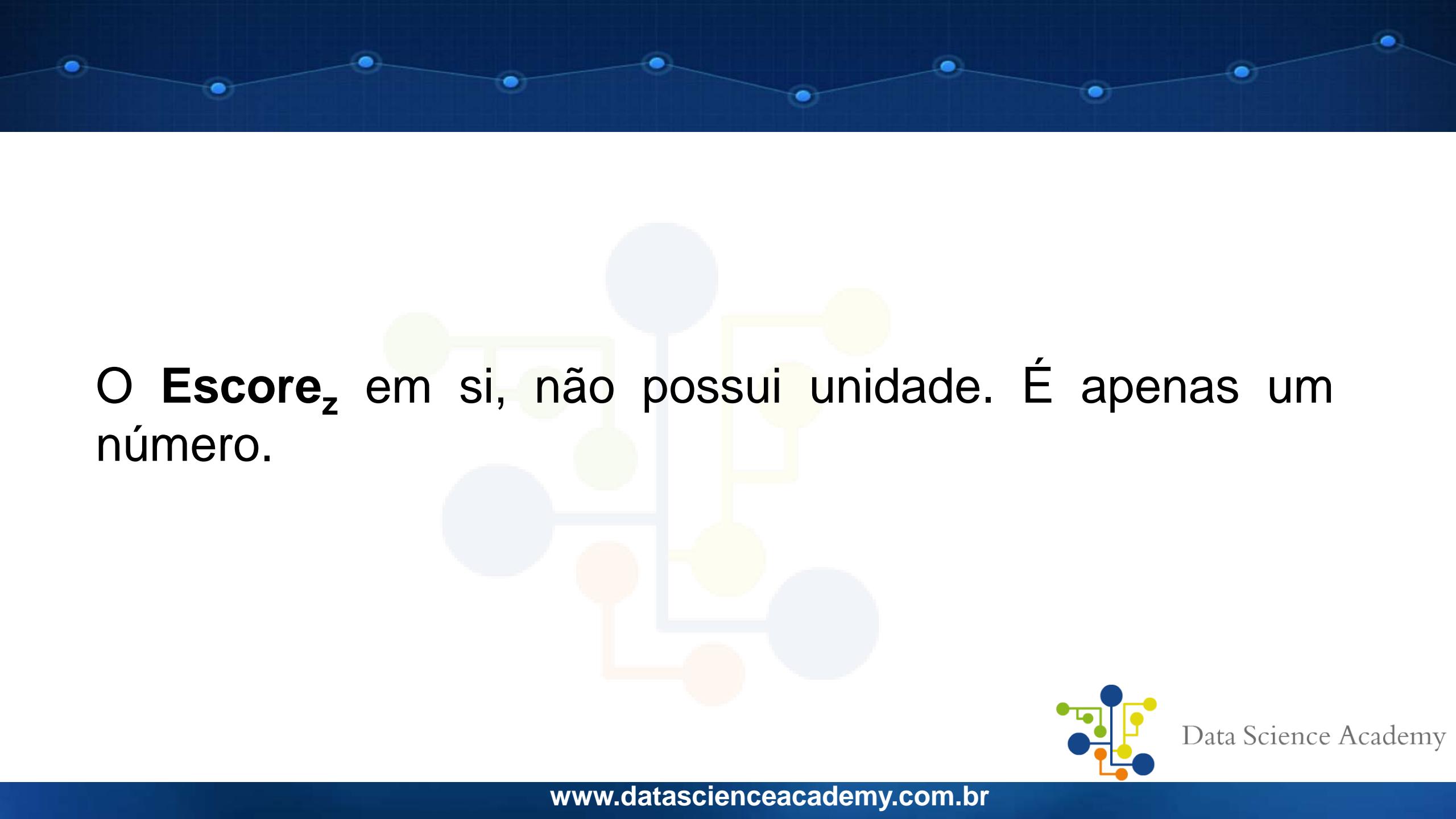
Data Science Academy



Pense no **Escore_z**, como uma simples conversão de um valor para uma outra unidade específica, a fim de fazer conclusões.



Data Science Academy

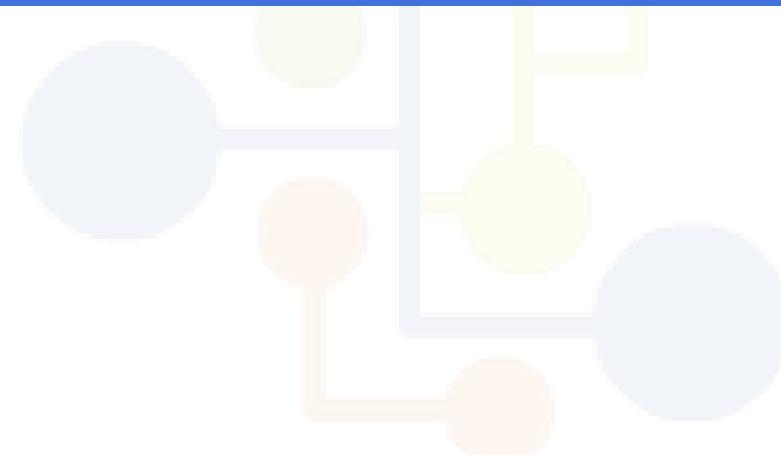


O **Escore_z** em si, não possui unidade. É apenas um número.



Data Science Academy

Exemplo



Data Science Academy

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

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



A tabela anterior, mostrou o número de calorias de sanduíches. Vejamos como podemos :

Calcular o **Escore_z** para o Mega Burger:

$$\text{Escore}_z = (x - x_1) / s = (1.420 - 776,30) / 385,10 = 1,67$$



Data Science Academy

O que 1.67 significa?

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



Data Science Academy

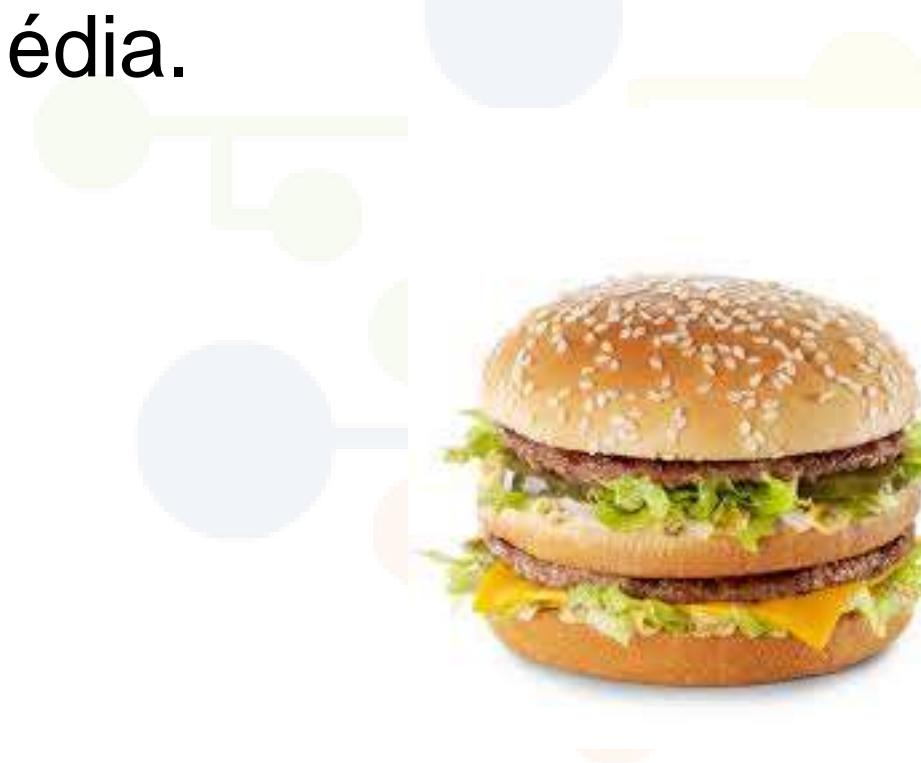
Agora vamos Vamos calcular o Escore_z do **Big Mac**:

$$\text{Escore}_z = (x - \bar{x}) / s = (430 - 776,30) / 385,10 = -0,90$$



Data Science Academy

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



Data Science Academy

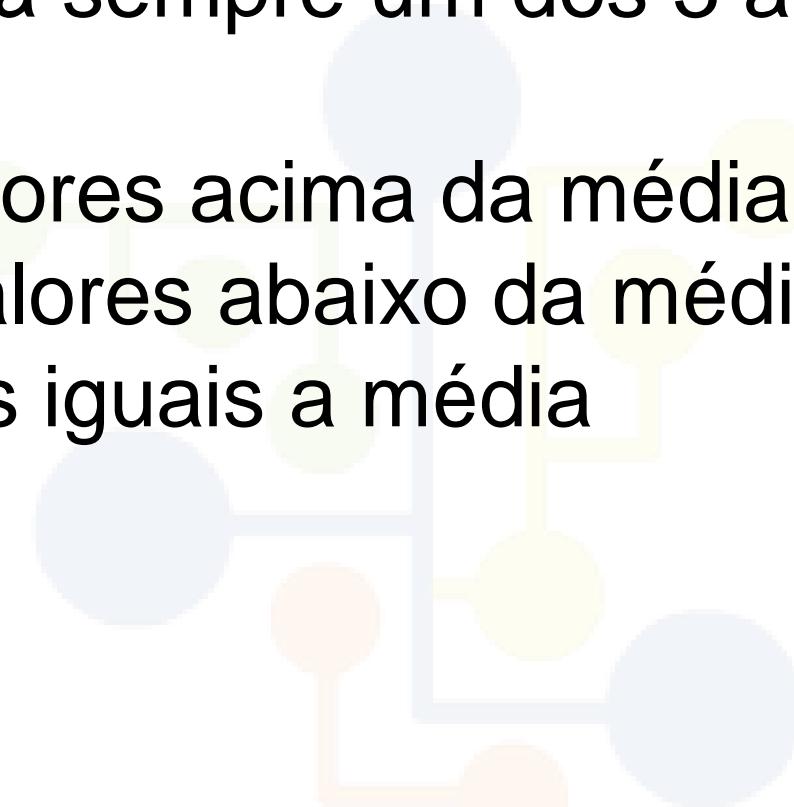


O **Escore_z** terá sempre um dos 3 atributos:

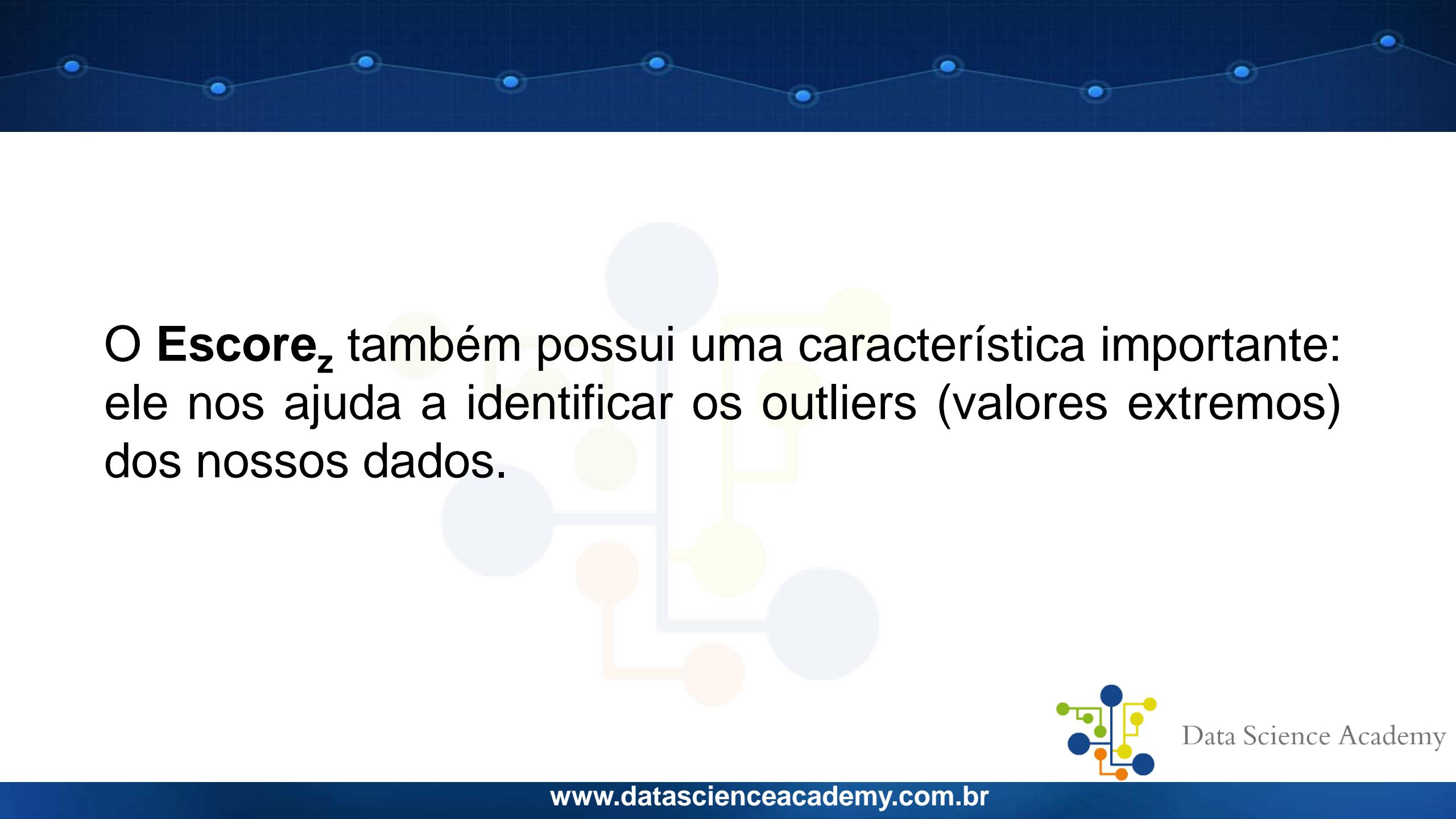
Positivo – valores acima da média

Negativo – valores abaixo da média

Zero – valores iguais a média



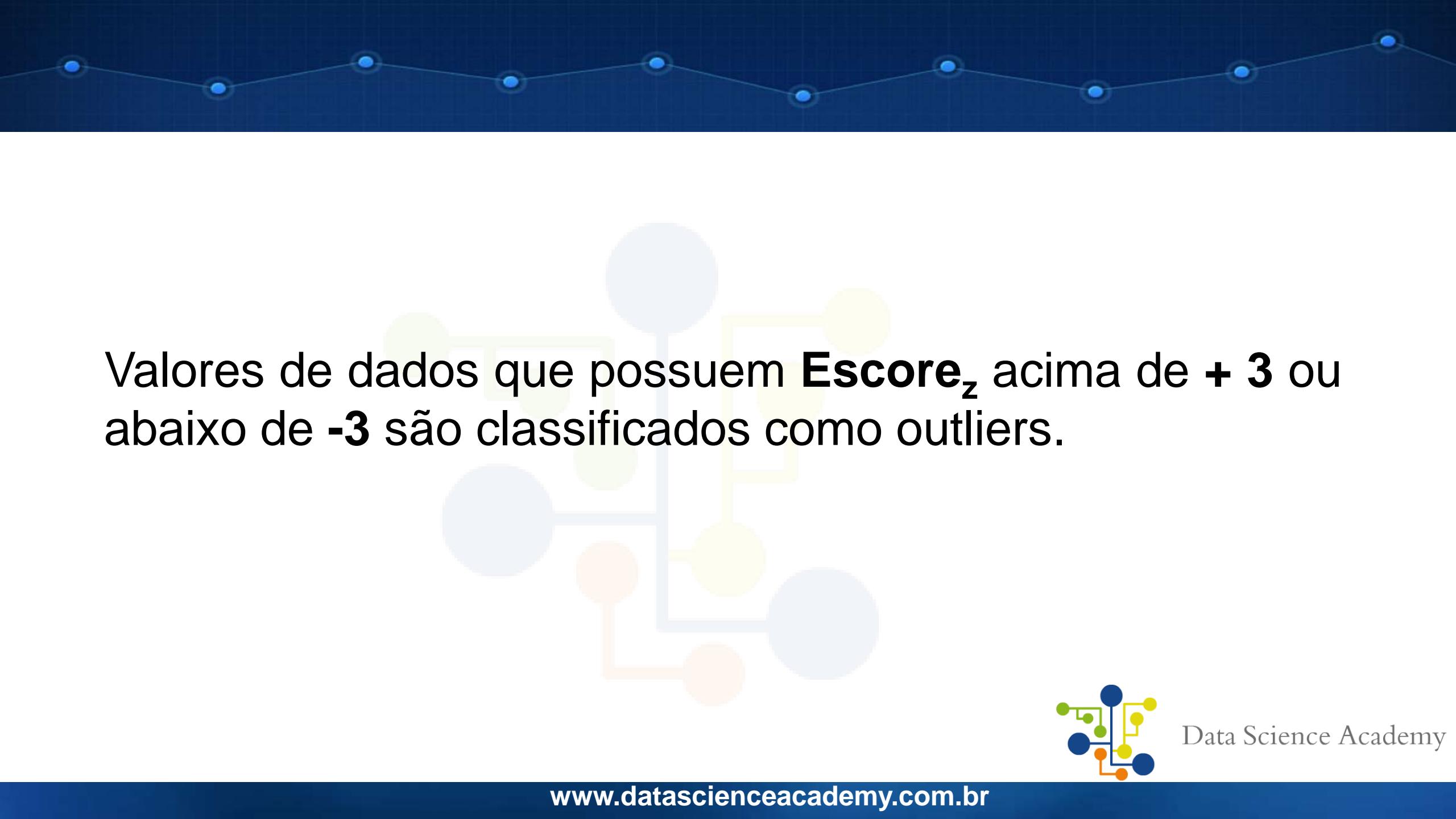
Data Science Academy



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



Data Science Academy



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



Data Science Academy

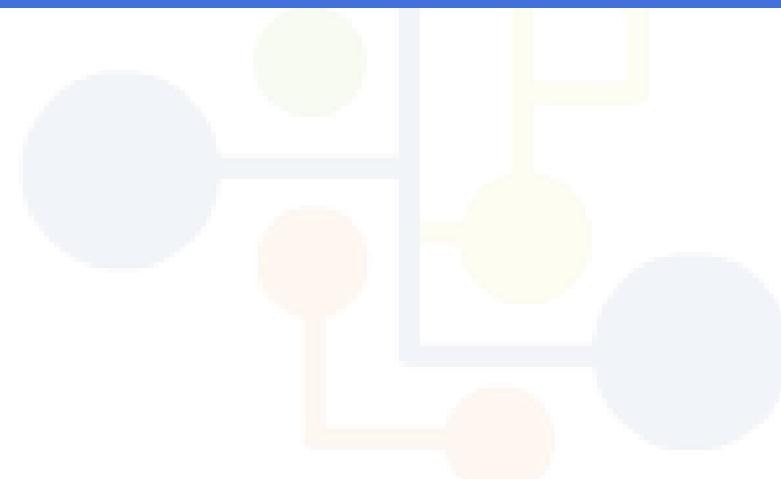


Ainda não ficou muito claro qual a função do **Escore_z**?



Data Science Academy

Exemplo



Data Science Academy



Durante a disputa de uma Olímpiada, como saber qual atleta teve melhor performance, considerando esportes diferentes?



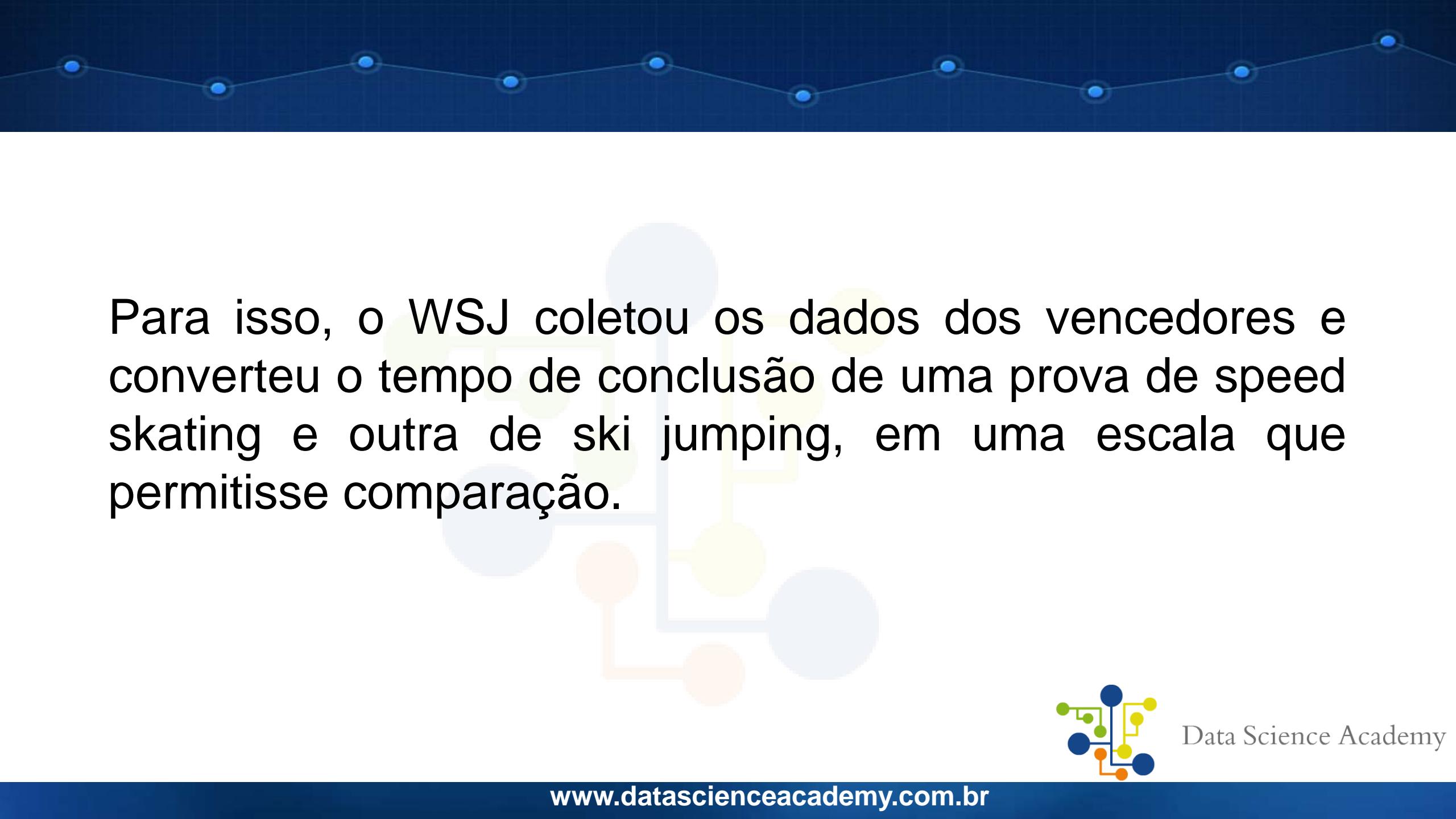
Data Science Academy



Em 2010, nos jogos de inverno de Vancouver, o Wall Street Journal fez uma análise estatística da performance de alguns atletas de 2 esportes, speed skating e o ski jumping.



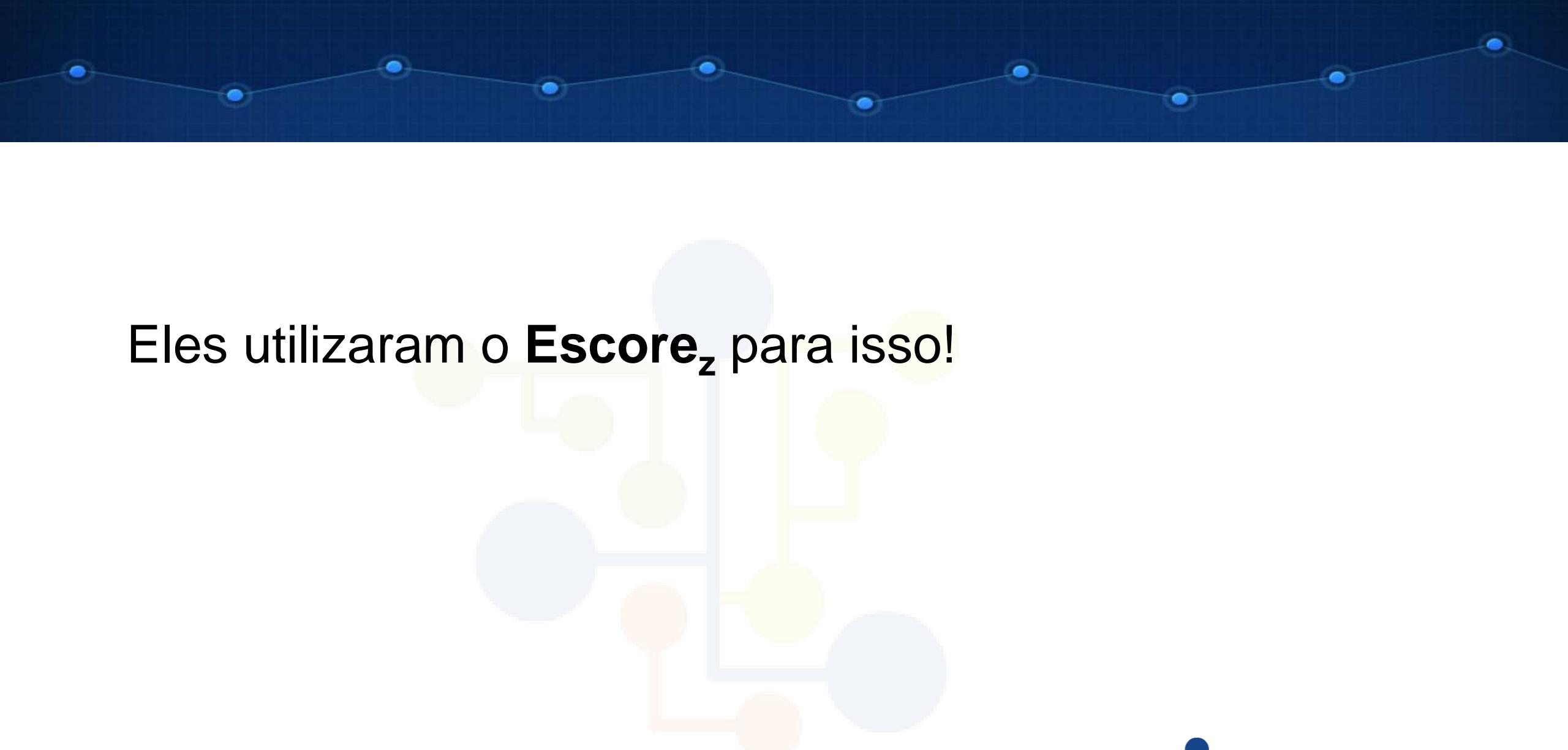
Data Science Academy



Para isso, o WSJ coletou os dados dos vencedores e converteu o tempo de conclusão de uma prova de speed skating e outra de ski jumping, em uma escala que permitisse comparação.



Data Science Academy



Eles utilizaram o **Escore_z** para isso!



Data Science Academy

Primeiro eles calcularam a média de conclusão das provas ao longo das últimas Olímpiadas. Então, calcularam o desvio padrão de cada atleta na competição de 2010.

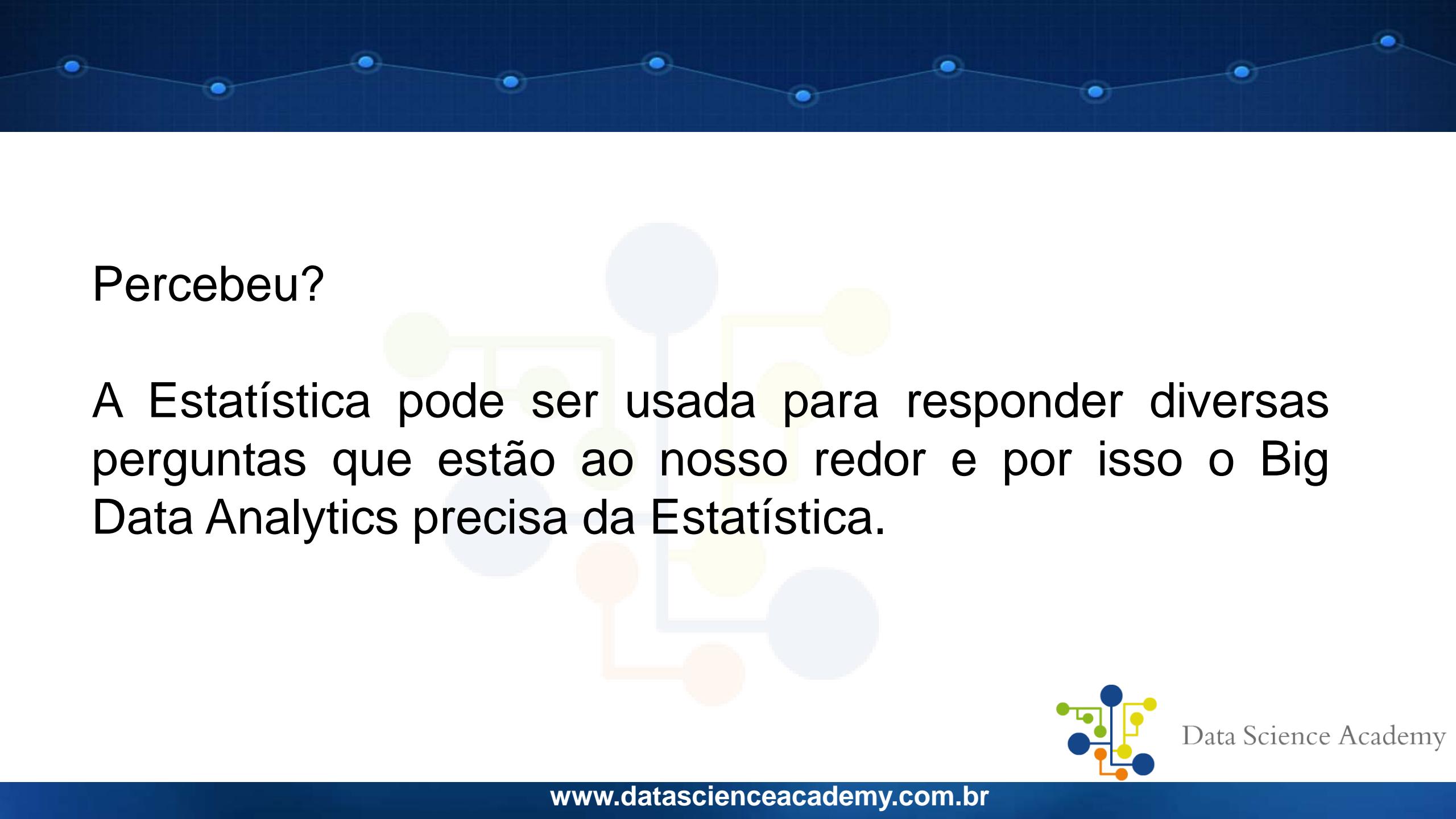


Data Science Academy

Com estes dados em mãos, ficou fácil calcular o **Escore_z**, e verificar os atletas que estavam mais ou menos distantes da média.



Data Science Academy



Percebeu?

A Estatística pode ser usada para responder diversas perguntas que estão ao nosso redor e por isso o Big Data Analytics precisa da Estatística.



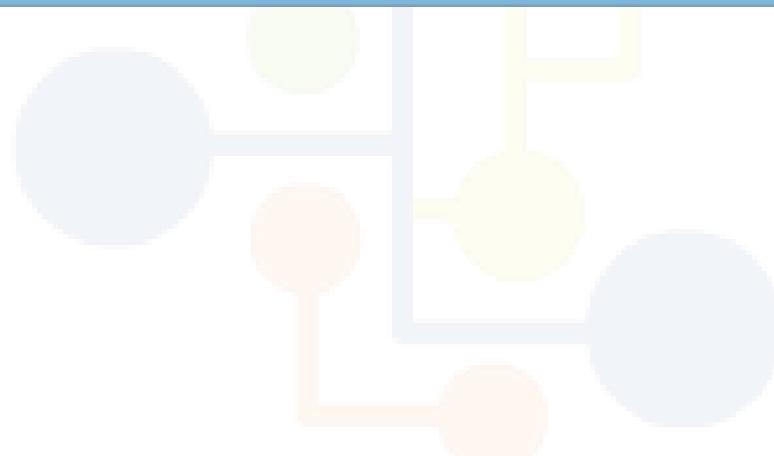
Data Science Academy

Esse tópico chegou ao final



Data Science Academy

Nível de Confiança



Data Science Academy

Imagine um arqueiro atirando em um alvo.

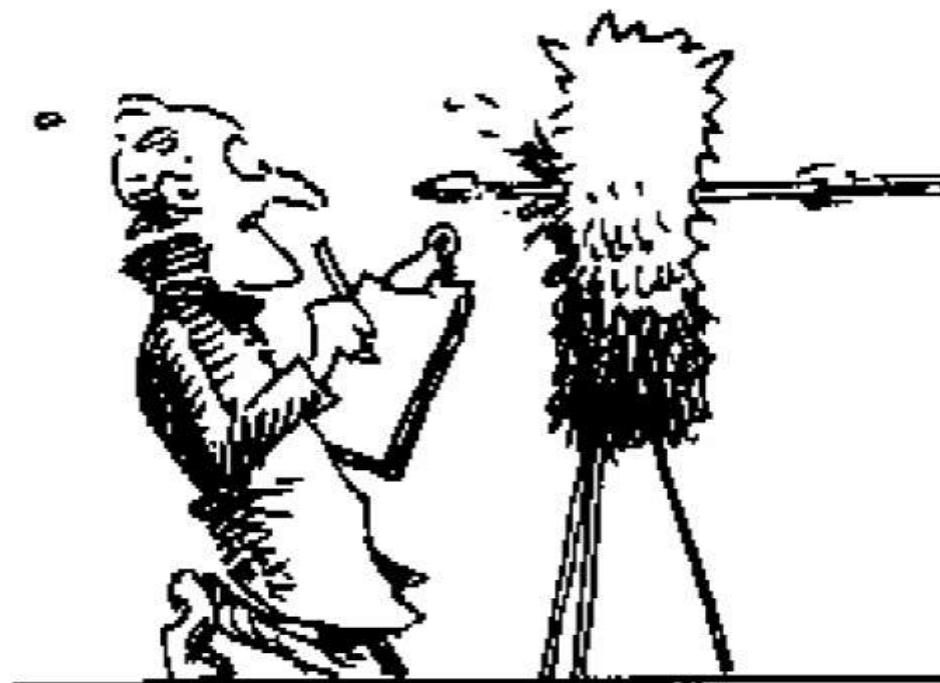


Data Science Academy

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.



Data Science Academy



Atrás do alvo encontra-se um estatístico, que não vê onde está o centro.

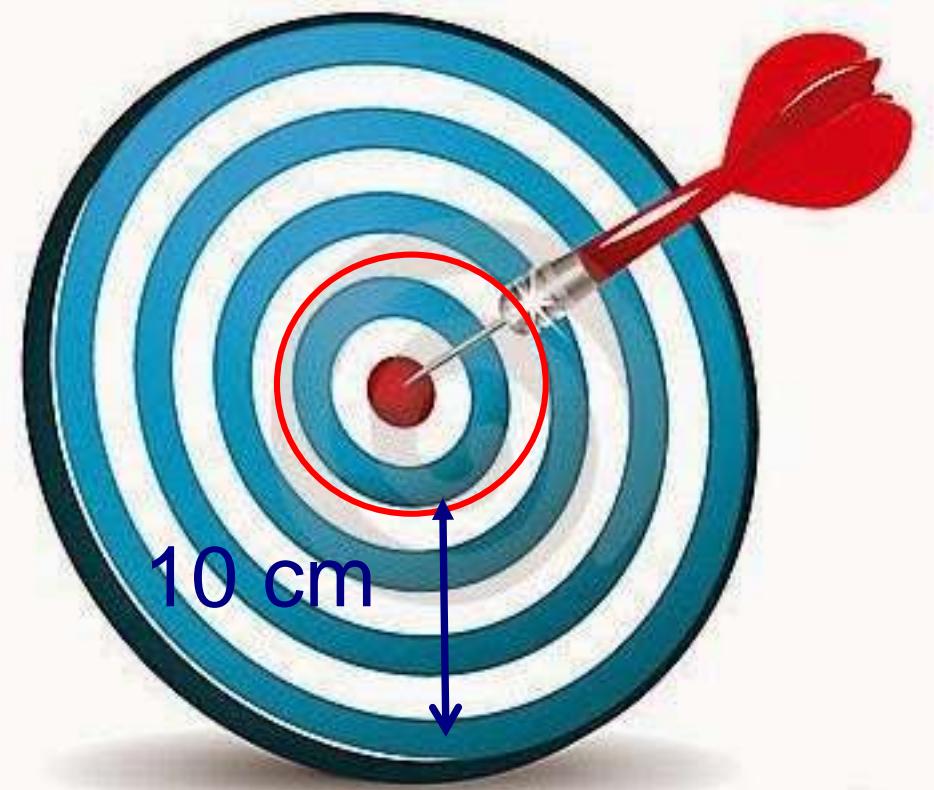
O arqueiro atira a primeira flecha.



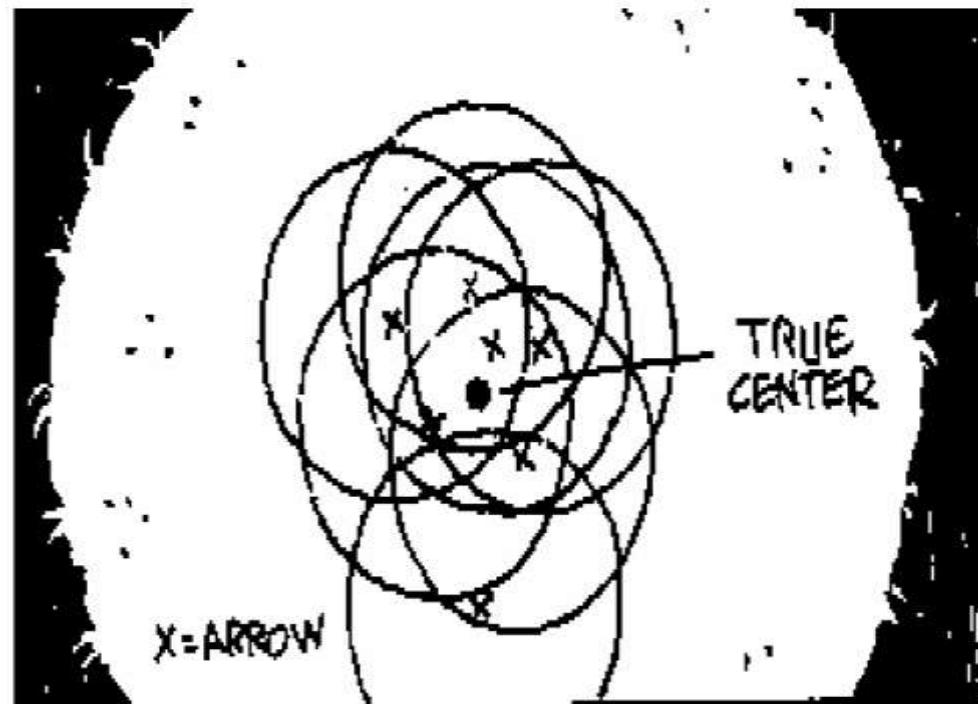
Data Science Academy

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.



Data Science Academy



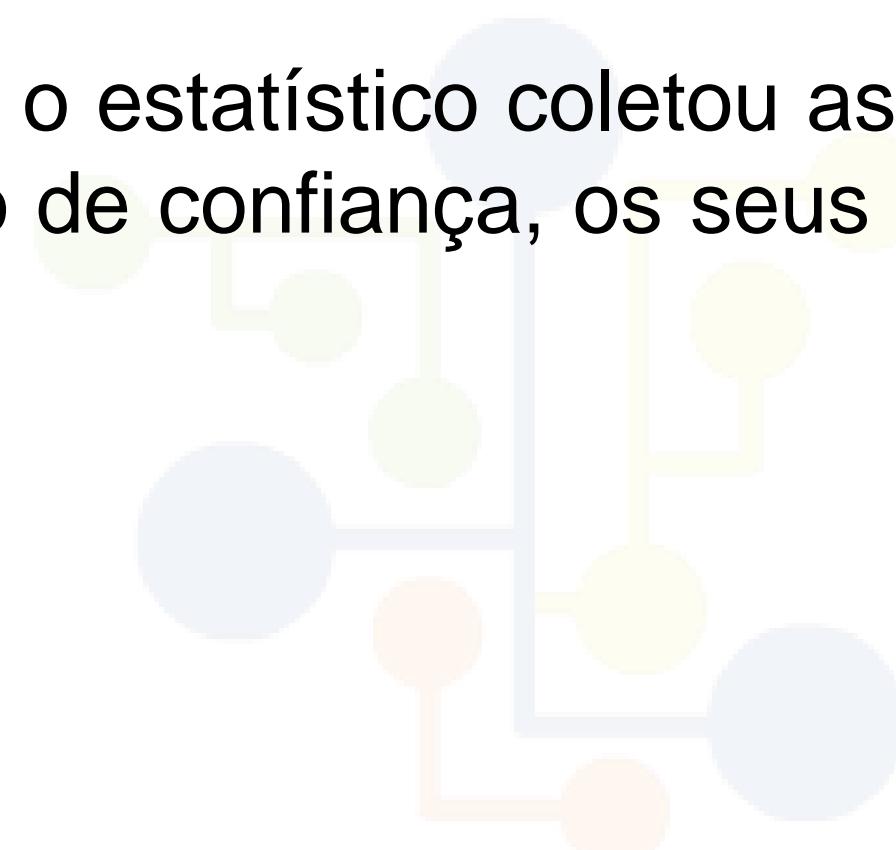
O estatístico raciocinou que se desenhasse círculos com 10 cm de raio ao redor de muitas flechas, os seus círculos incluiriam o centro do alvo em **95%** dos casos.



Data Science Academy



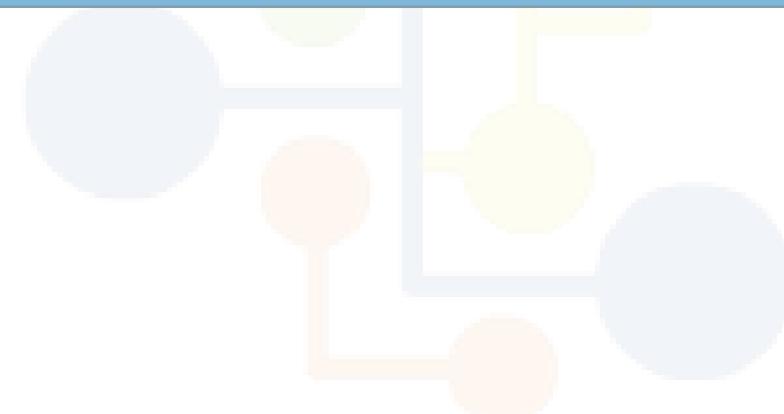
Resumindo, o estatístico coletou as amostras e construiu um intervalo de confiança, os seus círculo incluíram **95%** dos casos.



Data Science Academy



Como Melhorar a Confiança?



Data Science Academy

Considerando o exemplo do Arqueiro atirando no Alvo.



Data Science Academy

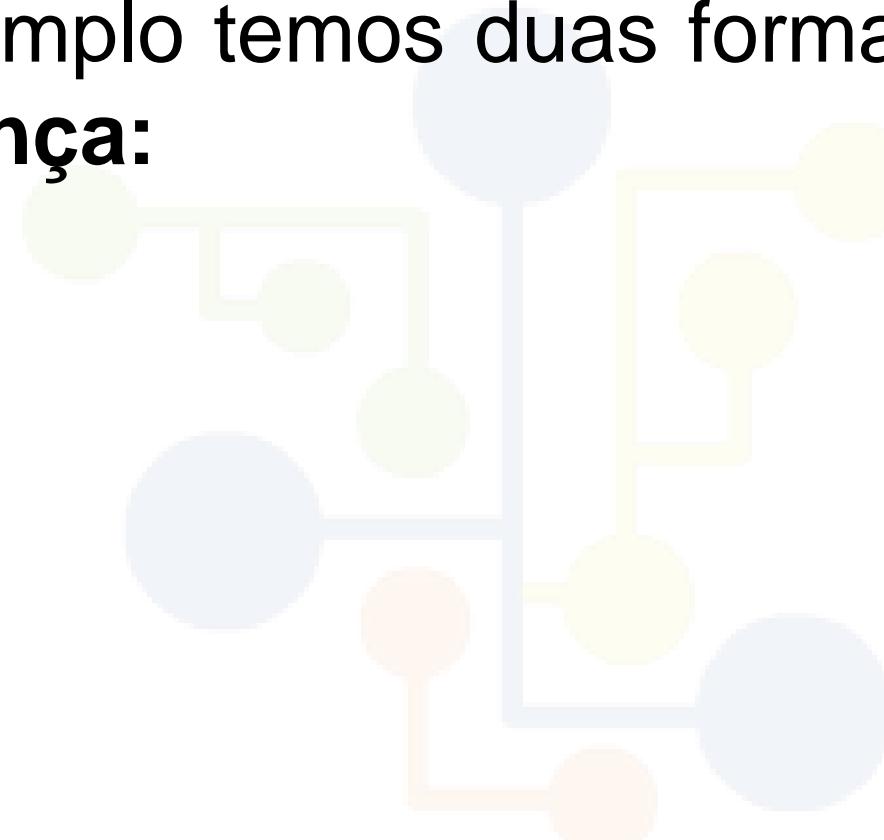


Como podemos aumentar a confiança?



Data Science Academy

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



Data Science Academy

Aumentando o tamanho do círculo

Isso equivale a alargar o intervalo de confiança (de 95% para 99%, por exemplo). Quanto maior a margem de erro, mais certo você estará de que o valor desejado encontra-se no intervalo.

Melhorando a mira do arqueiro

Isso equivale a aumentar o número de observações na amostra.



Data Science Academy

Nível de Confiança



Nível de confiança, é a probabilidade

$$1 - \alpha$$



Data Science Academy

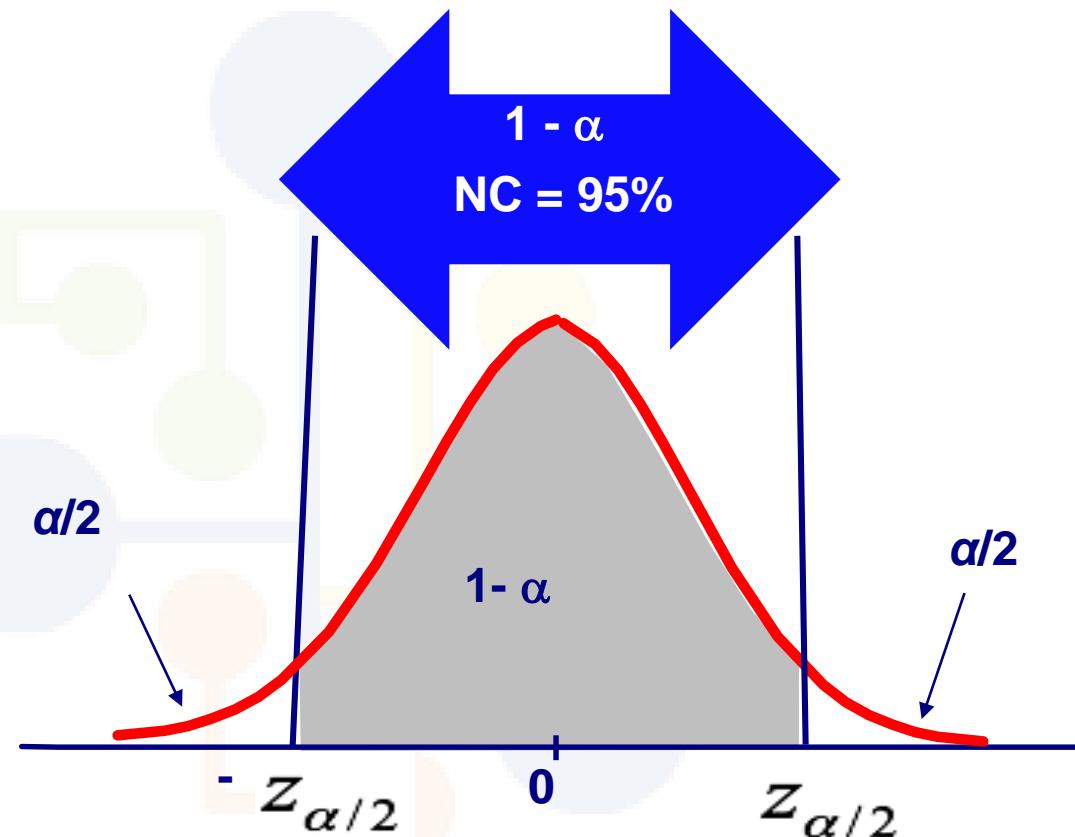
O nível de confiança é expresso percentualmente e por isso usamos:

$$100(1 - \alpha)$$



Data Science Academy

Nível de Confiança:



Data Science Academy

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

90%
 $\alpha = 0,10$

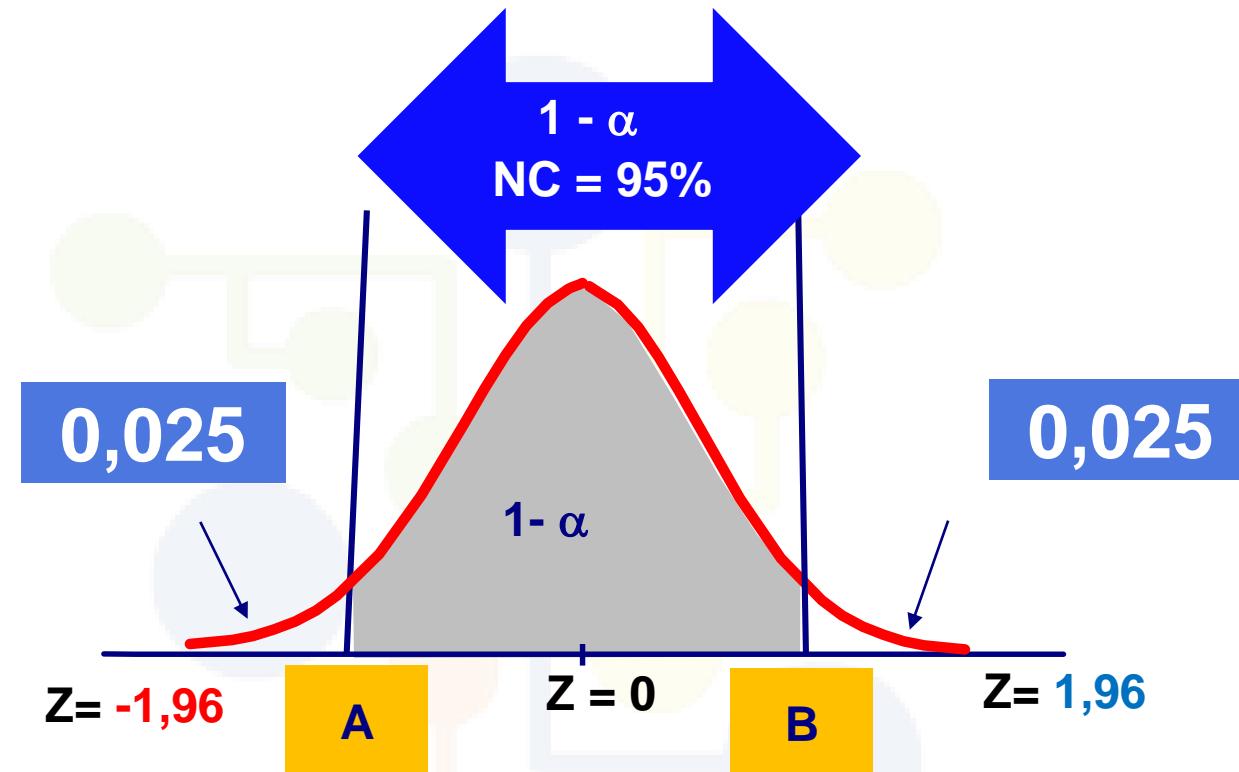
95%
 $\alpha = 0,05$

99%
 $\alpha = 0,01$



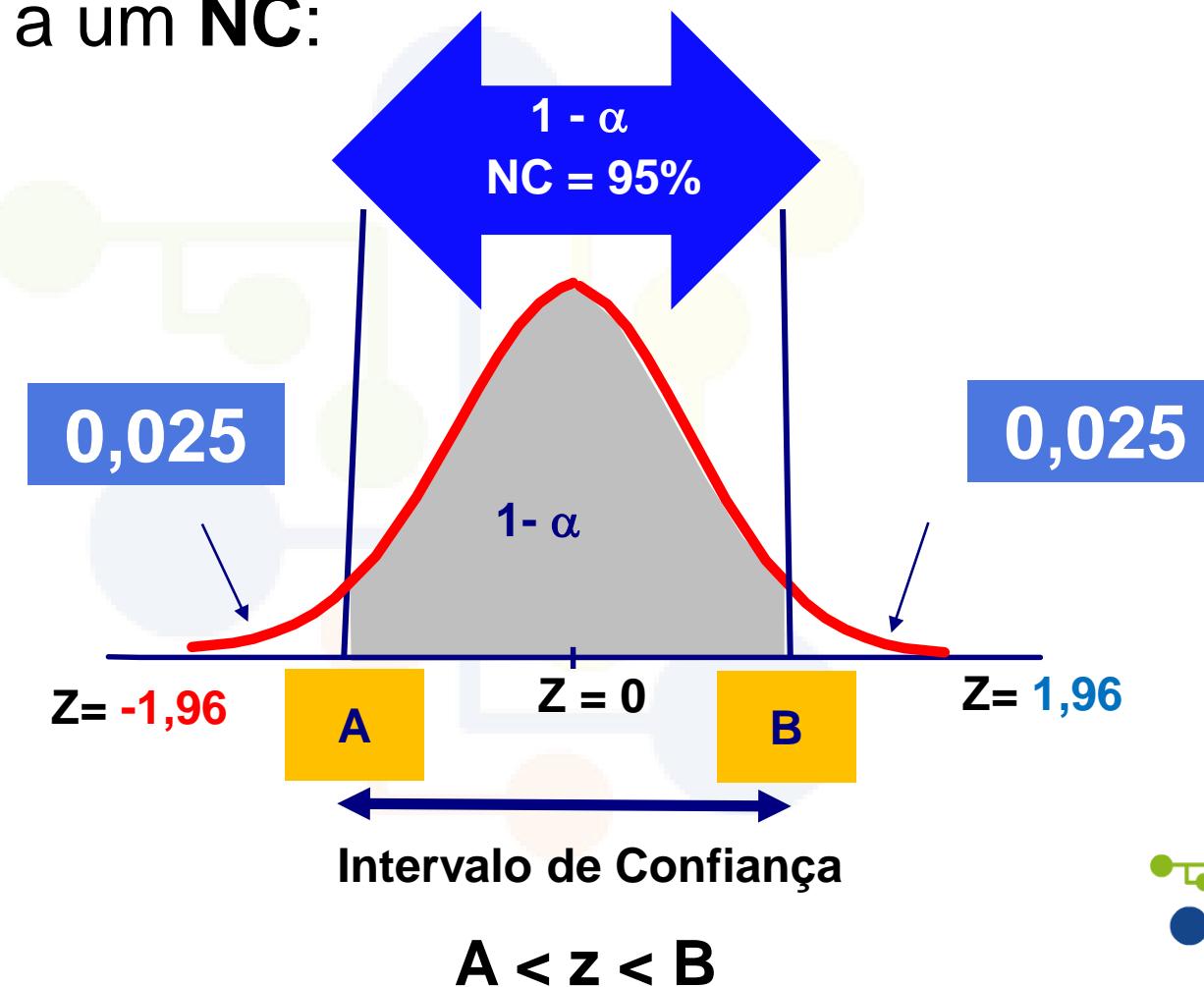
Data Science Academy

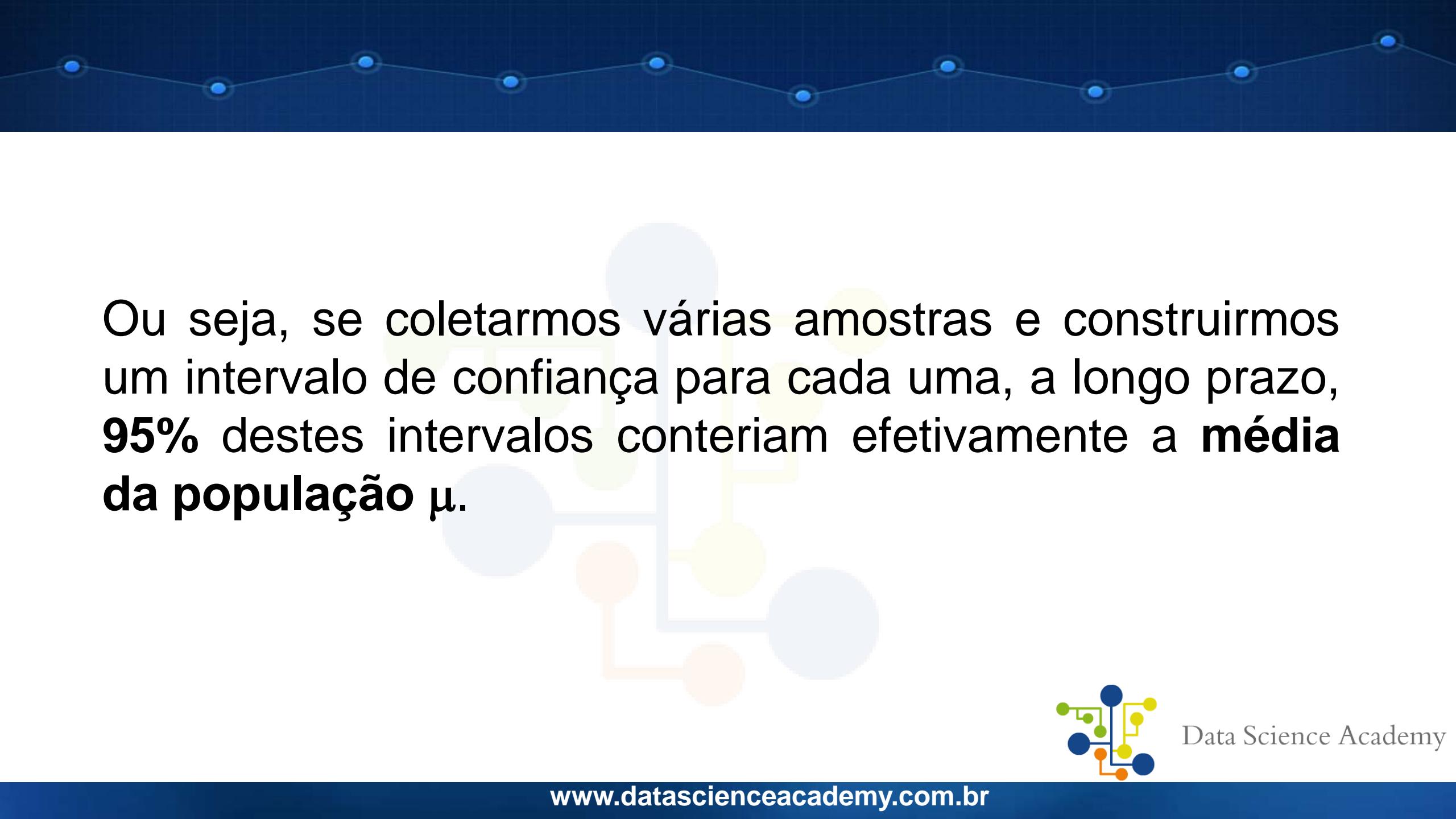
NC



Data Science Academy

O Intervalo de Confiança consiste em um intervalo na escala z e está associado a um NC:





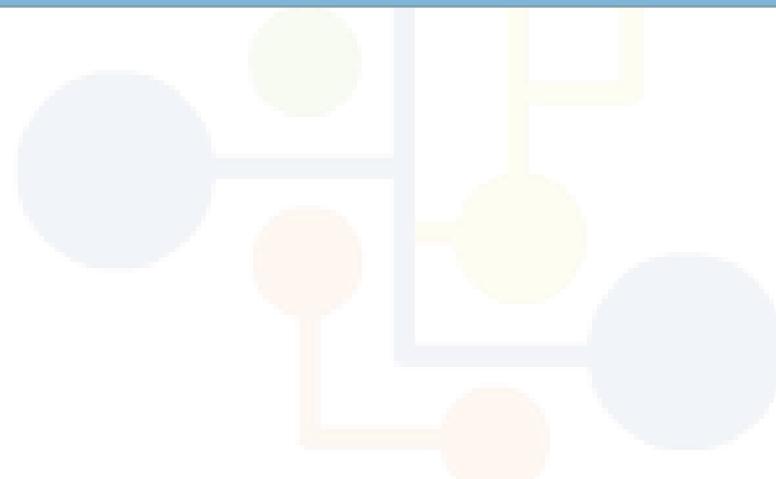
Ou seja, se coletarmos várias amostras e construirmos um intervalo de confiança para cada uma, a longo prazo, **95%** destes intervalos conteriam efetivamente a **média da população μ** .



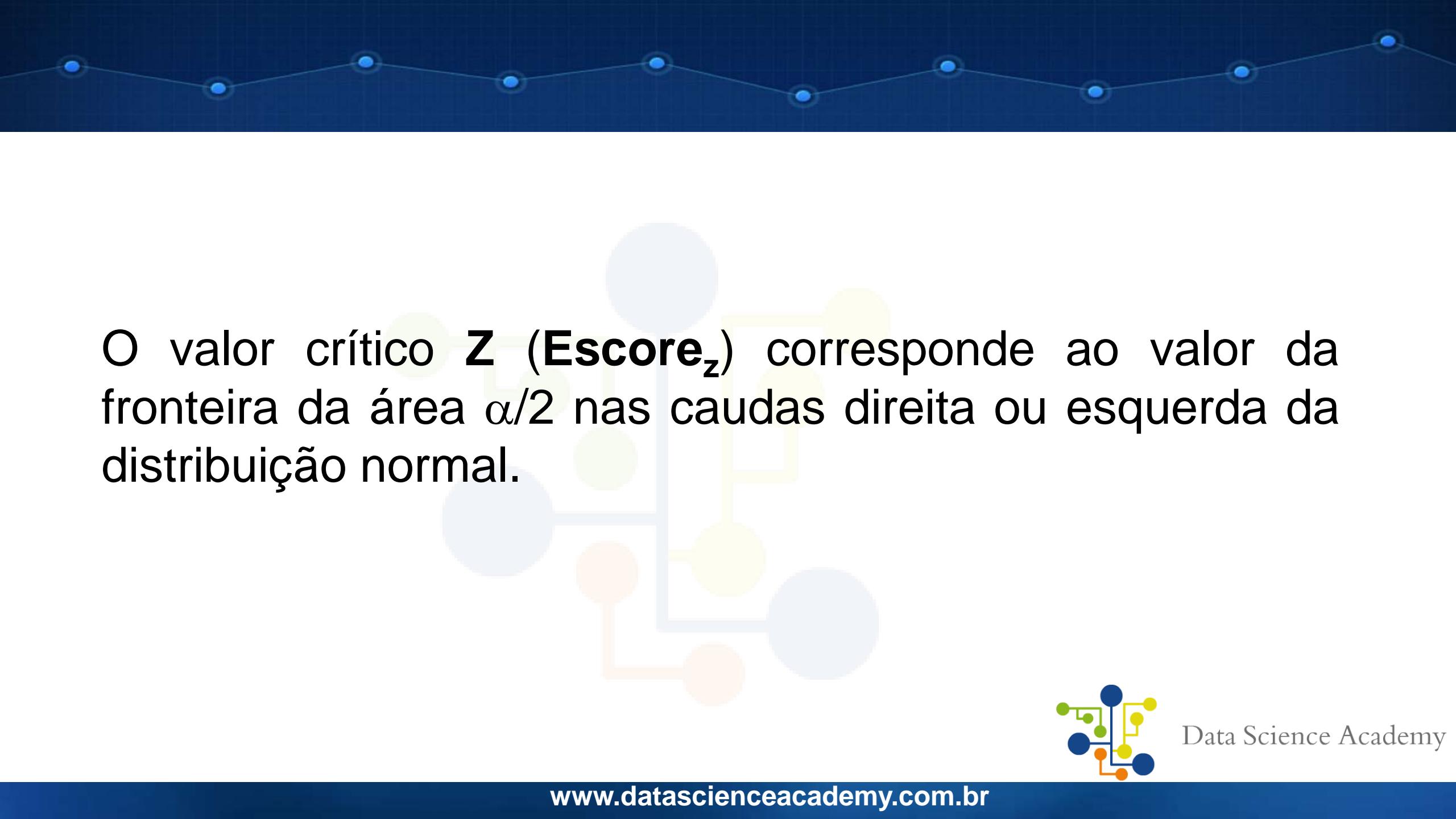
Data Science Academy



Valor Crítico



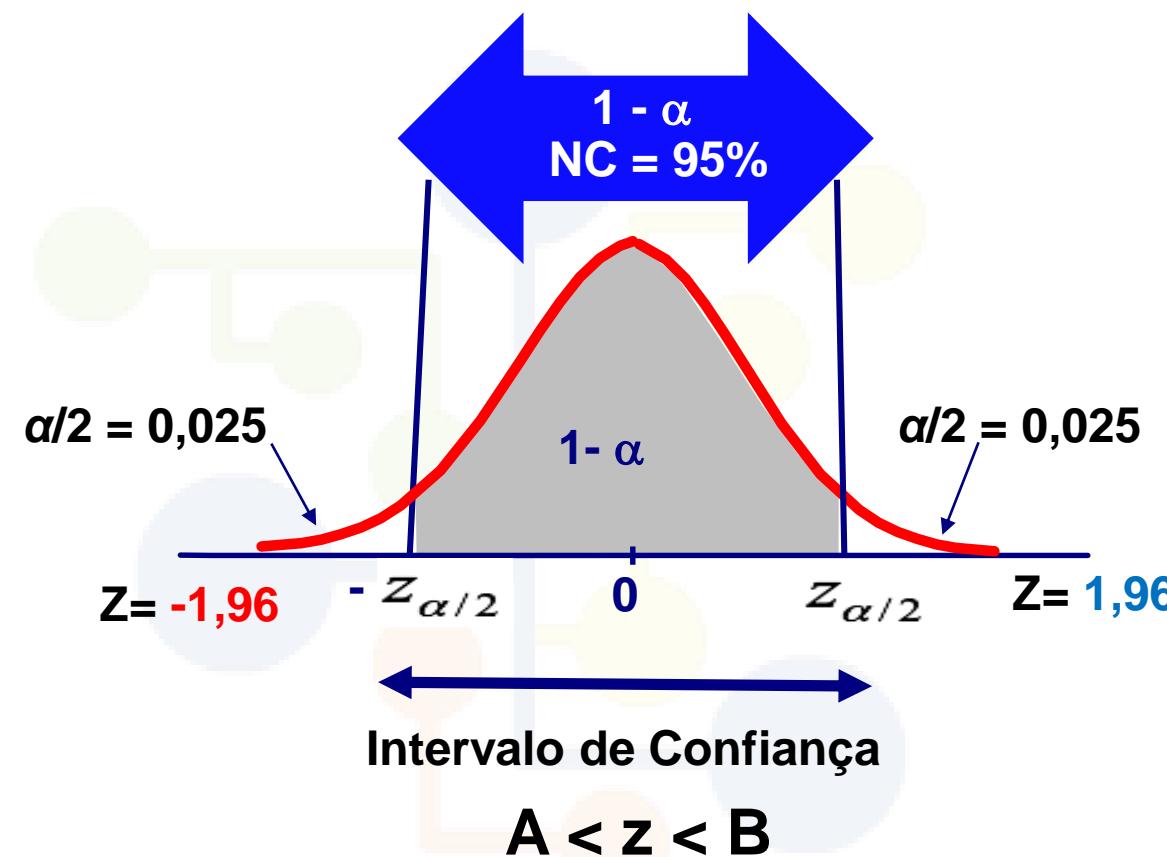
Data Science Academy



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

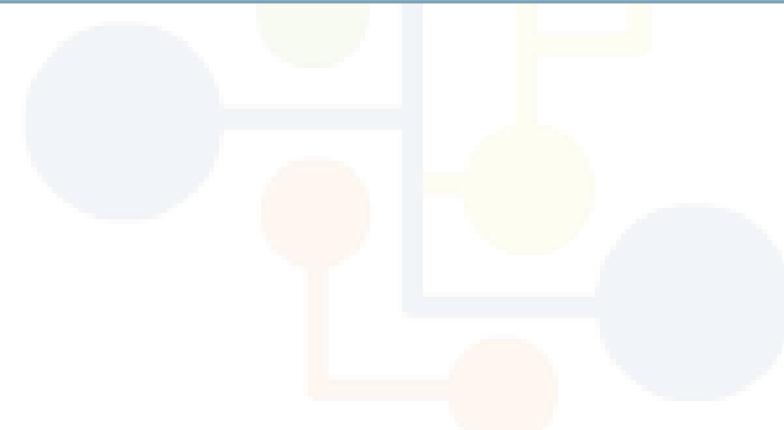


Data Science Academy

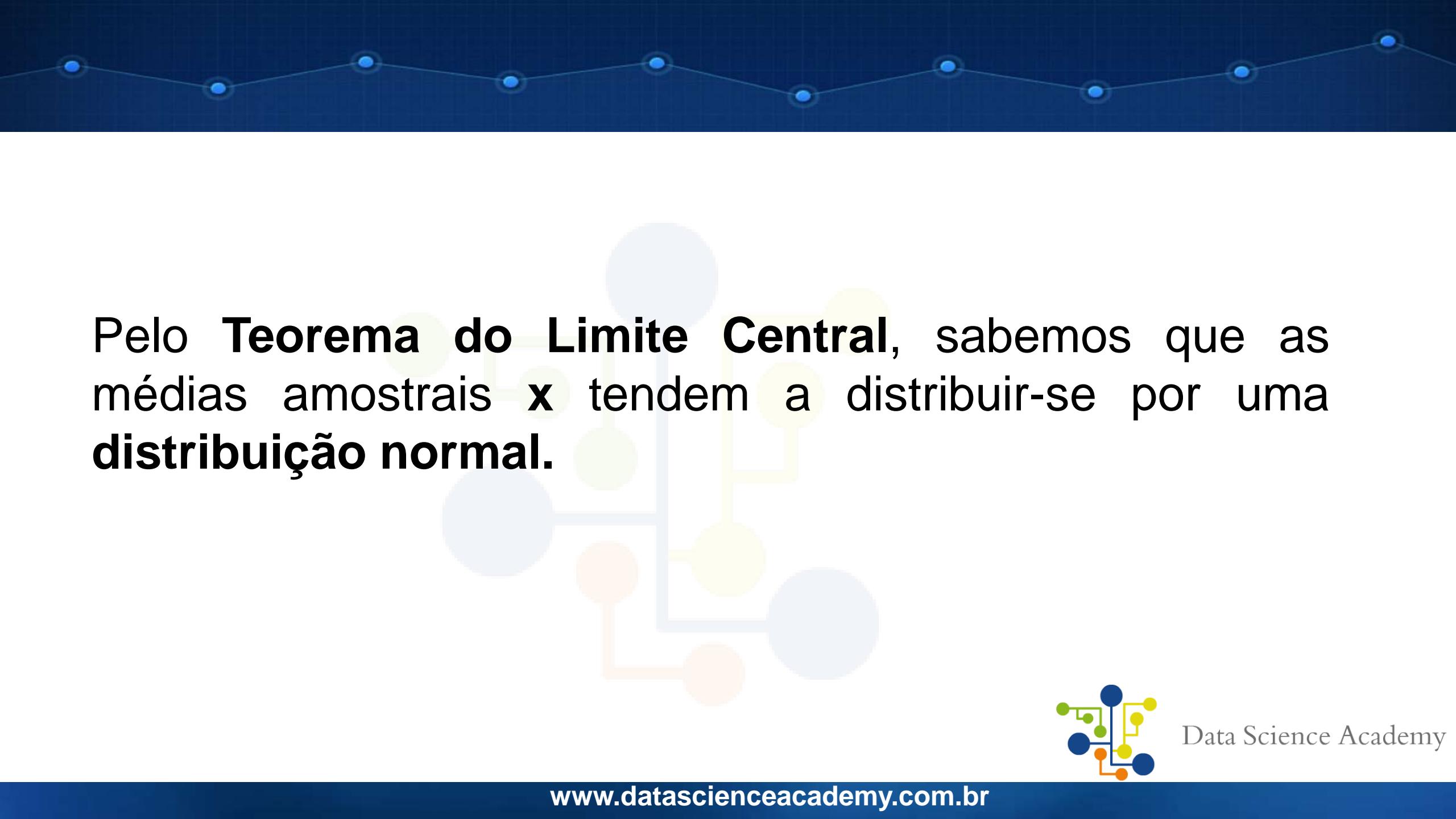


Data Science Academy

Teorema do Limite Central



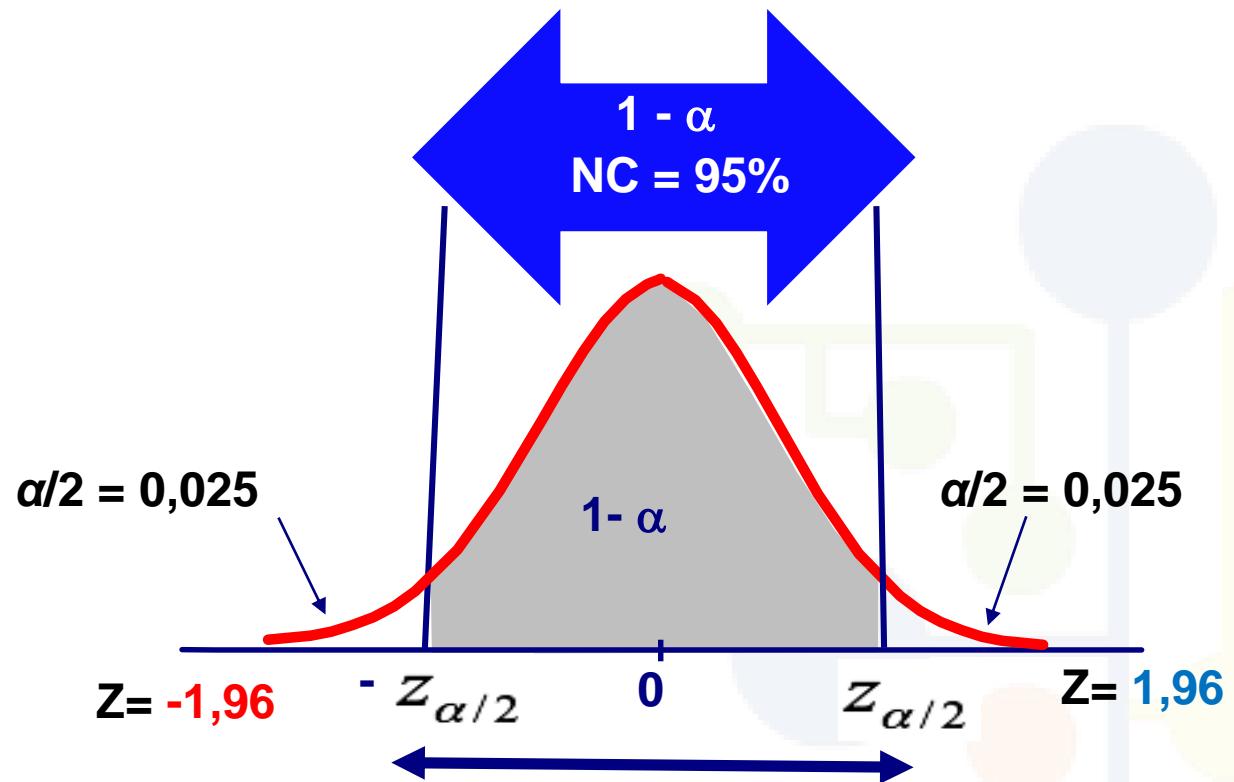
Data Science Academy



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



Data Science Academy

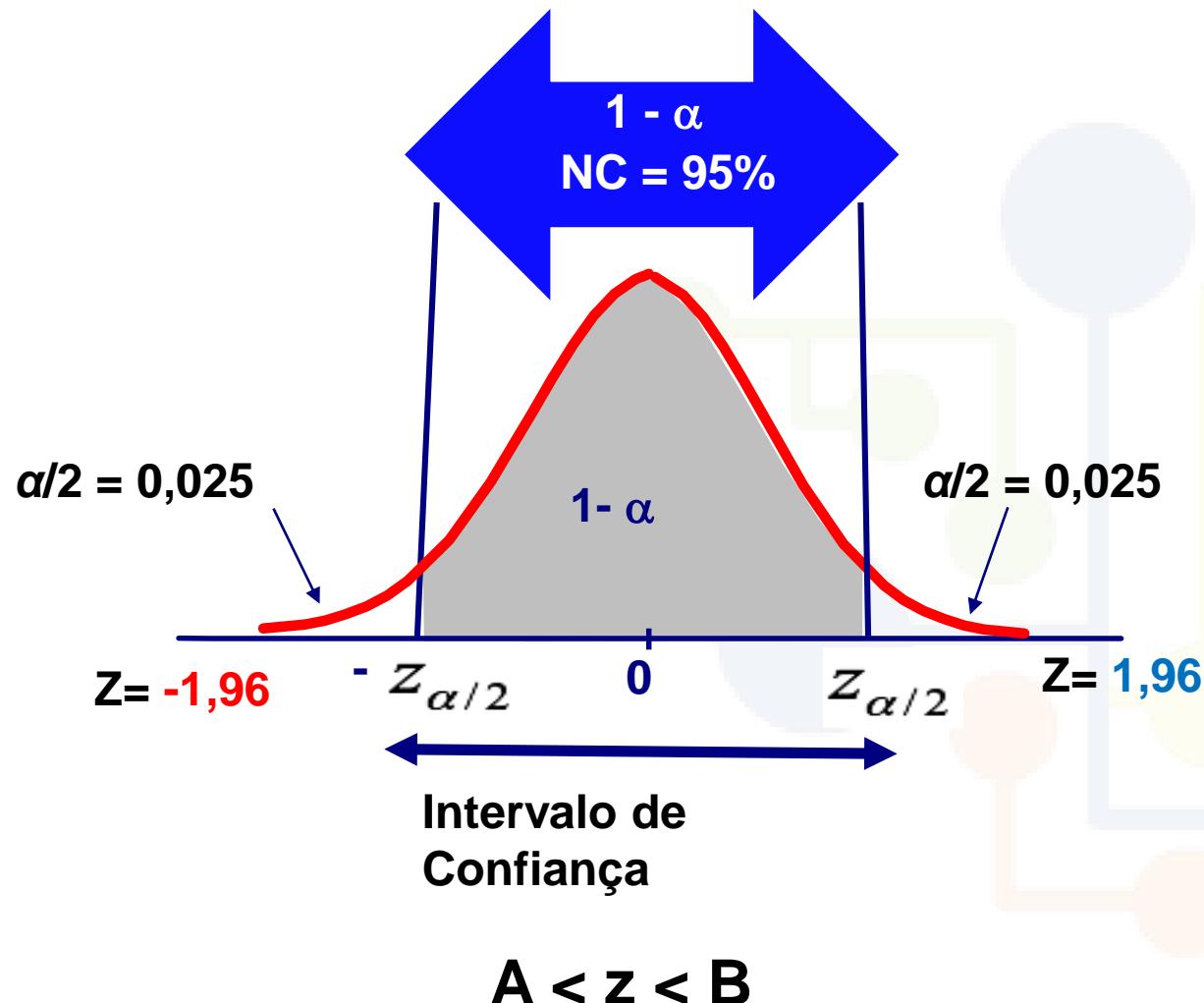


$$A < z < B$$

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



Data Science Academy

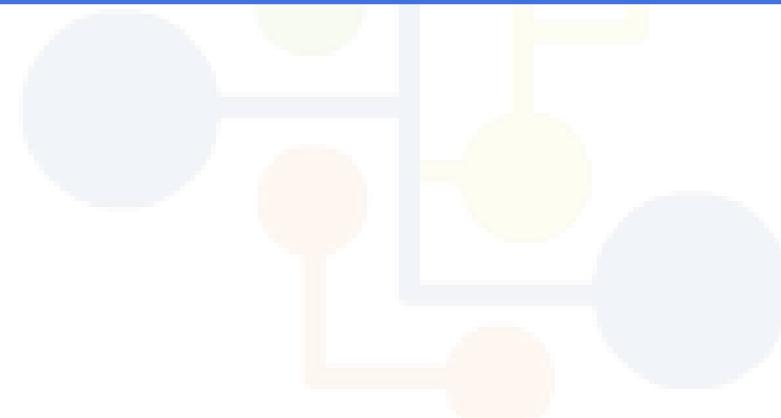


Sendo $\alpha/2$ a área sombreada de cada extremo, há uma possibilidade de α da média amostral estar em um dos **2 extremos**

Pela regra do complemento, há uma probabilidade $1 - \alpha$ da média amostral estar na região **não sombreada**.

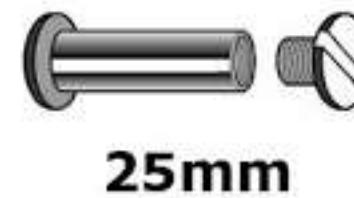


Exemplo



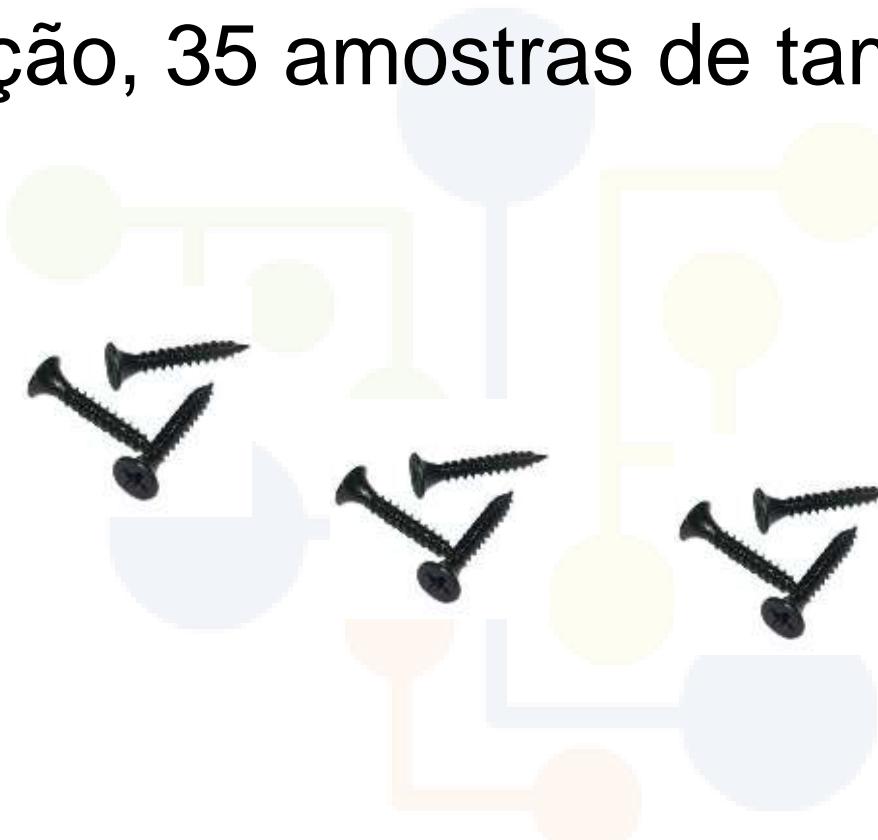
Data Science Academy

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



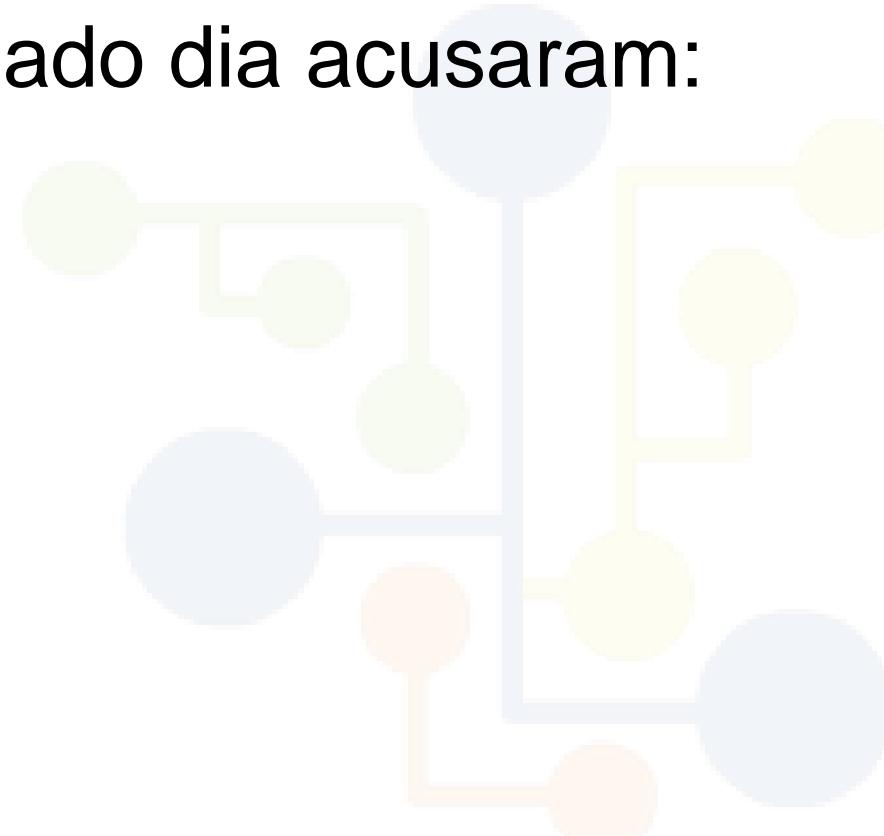
Data Science Academy

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



Data Science Academy

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



Data Science Academy

A média das médias das amostras =

$$\bar{\bar{x}} = 25mm$$

Desvio padrão =

$$\bar{s} = 1,5mm$$



Data Science Academy

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

Grau de Confiança	Nível de Significância	Valor Crítico Z
99%	0,01	2,575



Data Science Academy

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

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

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



Data Science Academy

Resposta:

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



Data Science Academy

Resposta:

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

Ou



Data Science Academy

Resposta:

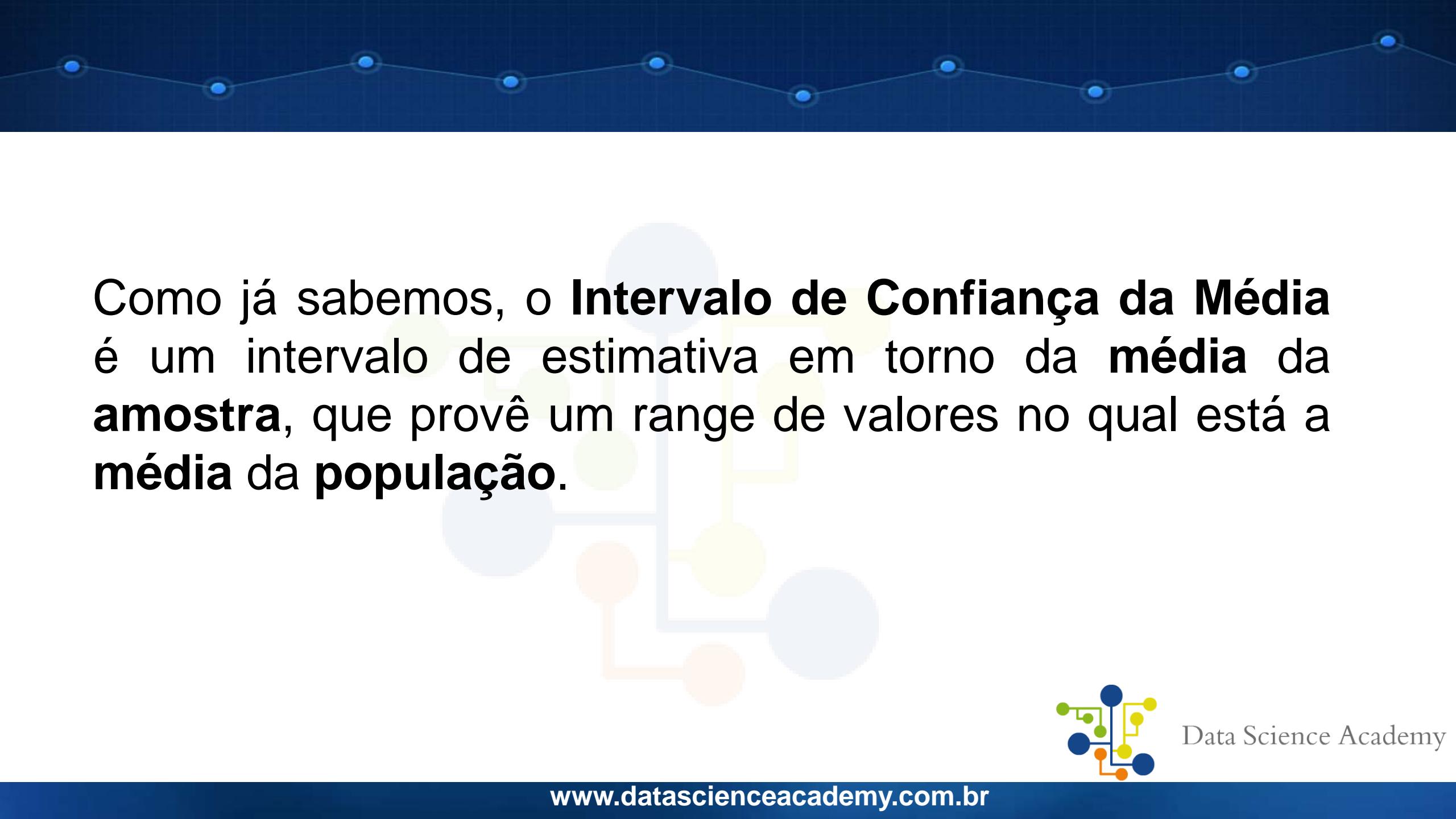
→ A fábrica possui **99%** de chance de produzir lotes de peças com médias entre **21,14 mm** e **28,86 mm**.



21,14 mm
e
28,86 mm



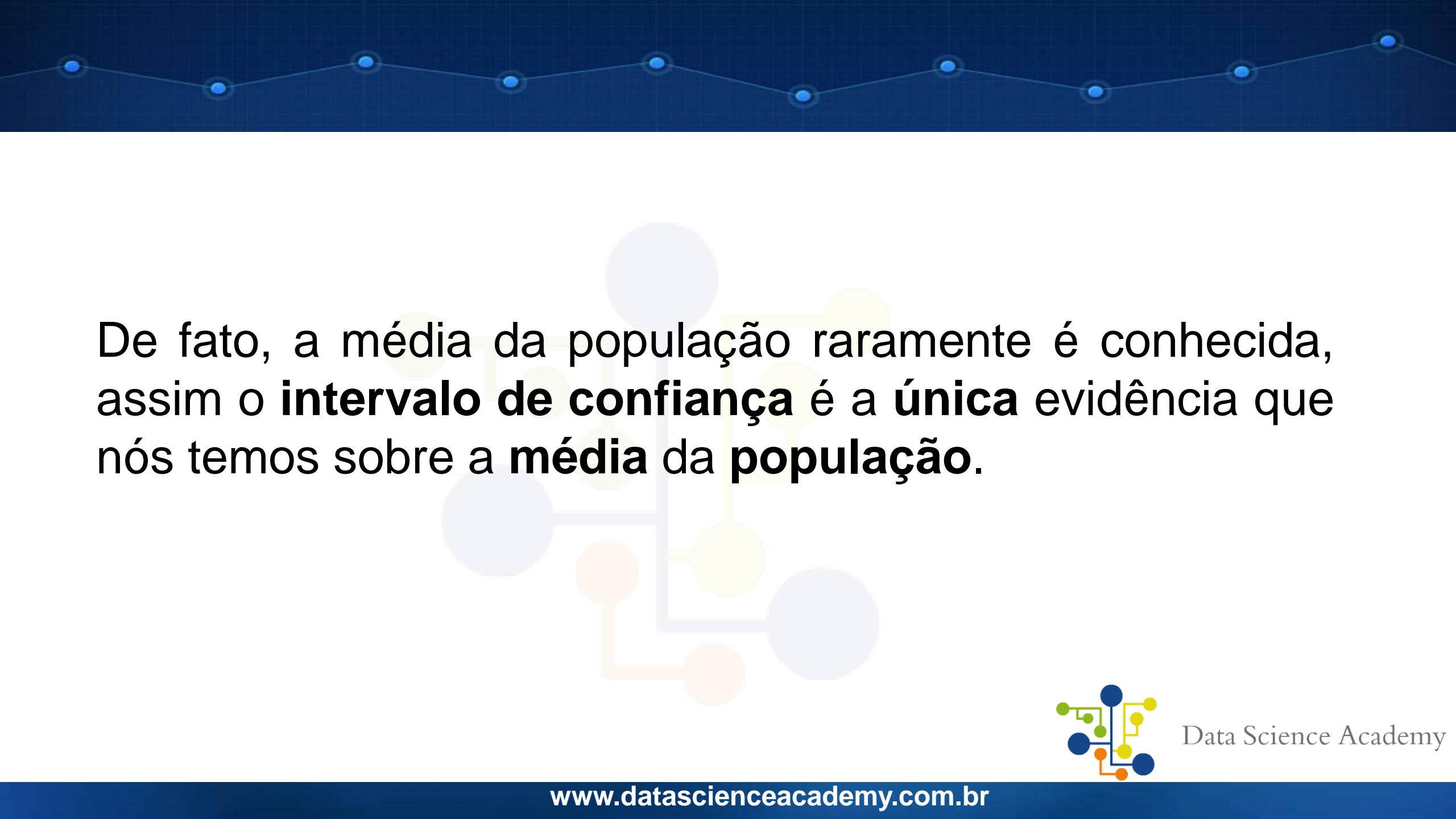
Data Science Academy



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**.



Data Science Academy



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**.



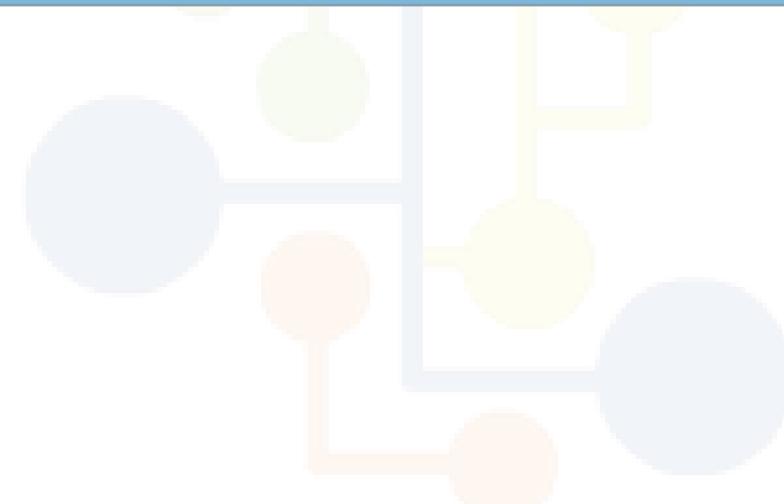
Data Science Academy

Esse tópico chegou ao final



Data Science Academy

O Que é Teste de Hipótese?

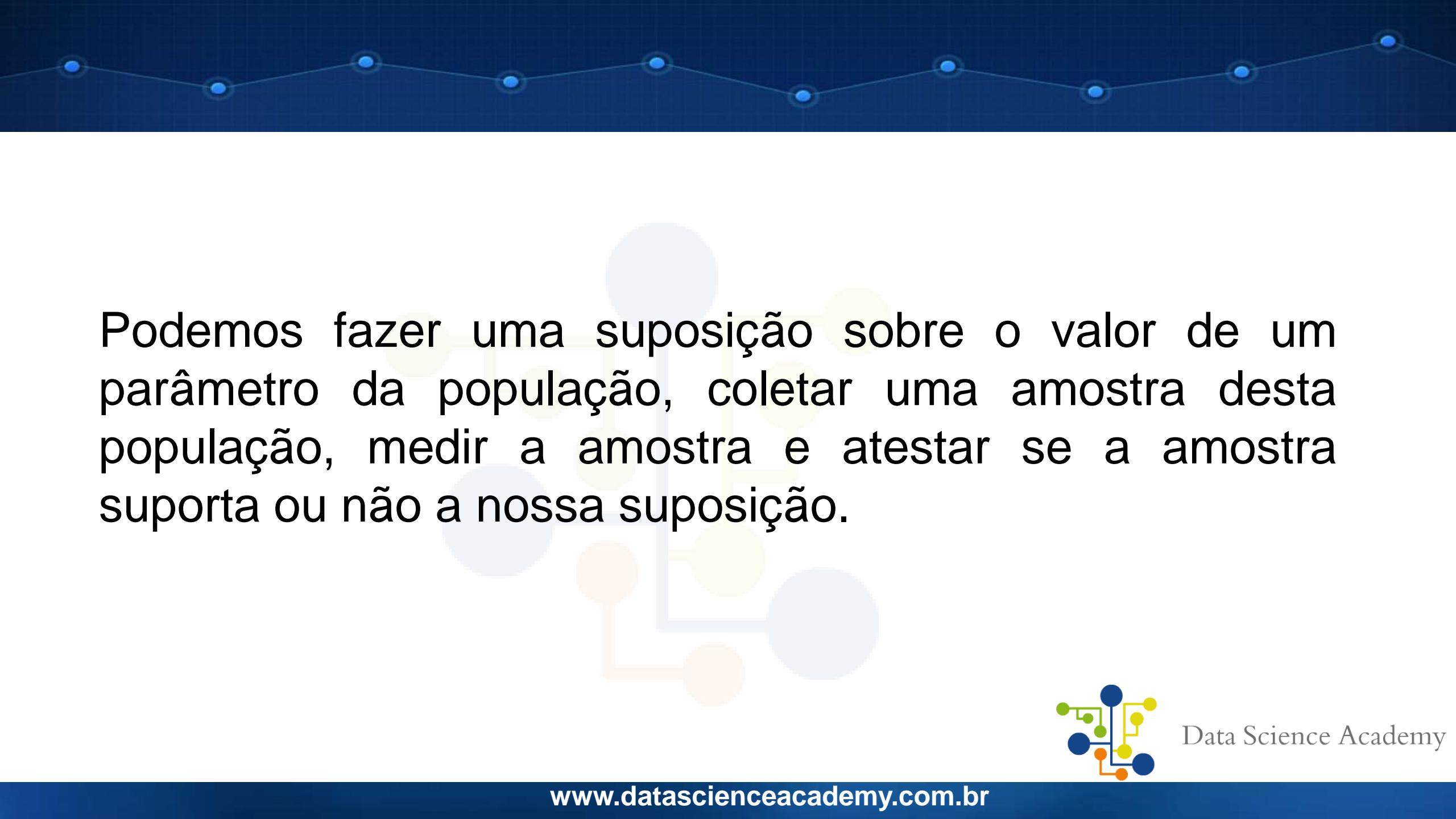


Data Science Academy

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



Data Science Academy



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.

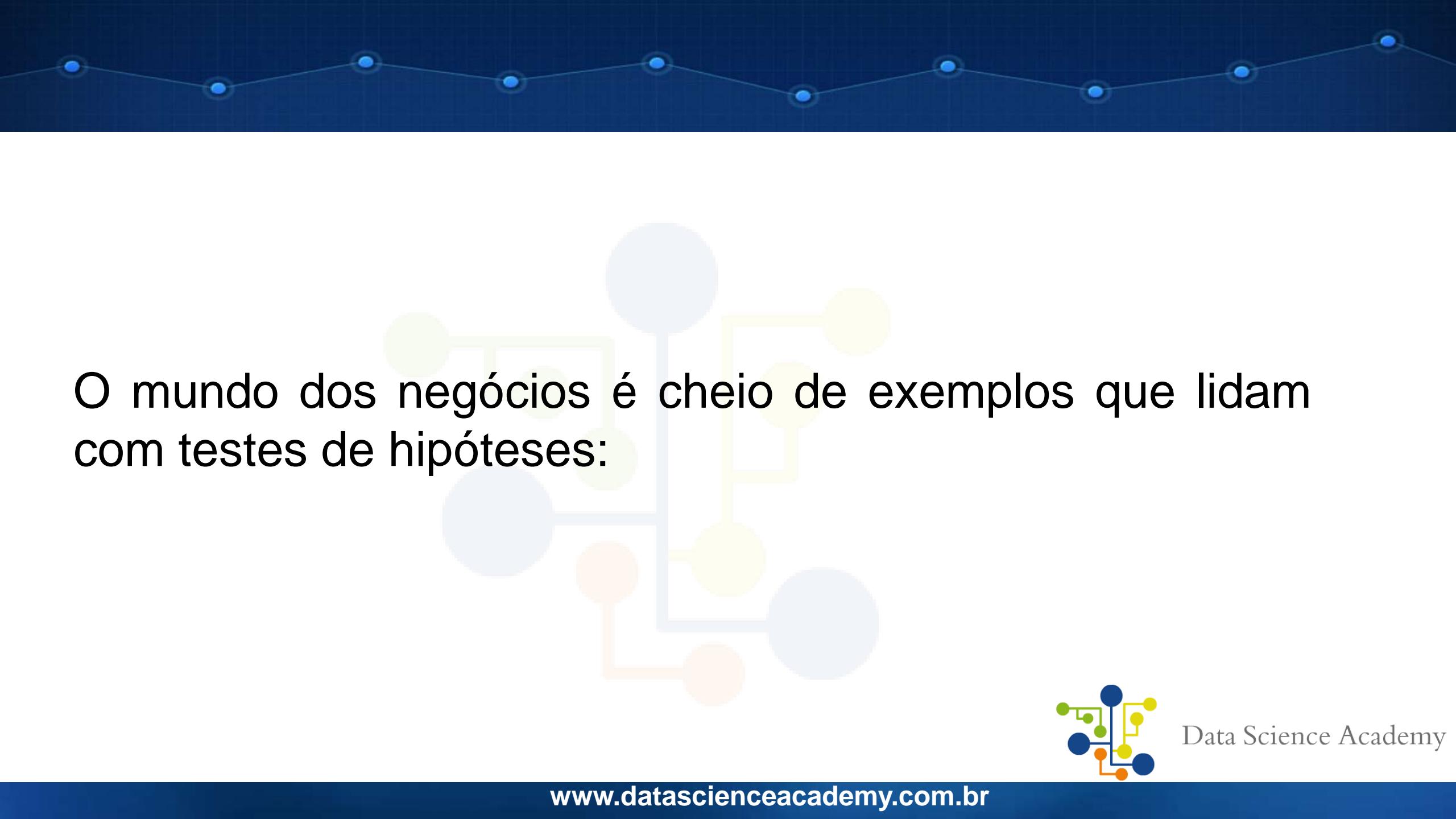


Data Science Academy

Mundos dos Negócios



Data Science Academy



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



Data Science Academy

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



Data Science Academy

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



Um laboratório independente pode testar esta afirmação com um **teste de hipótese**.



Data Science Academy

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



Data Science Academy

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.



Data Science Academy

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

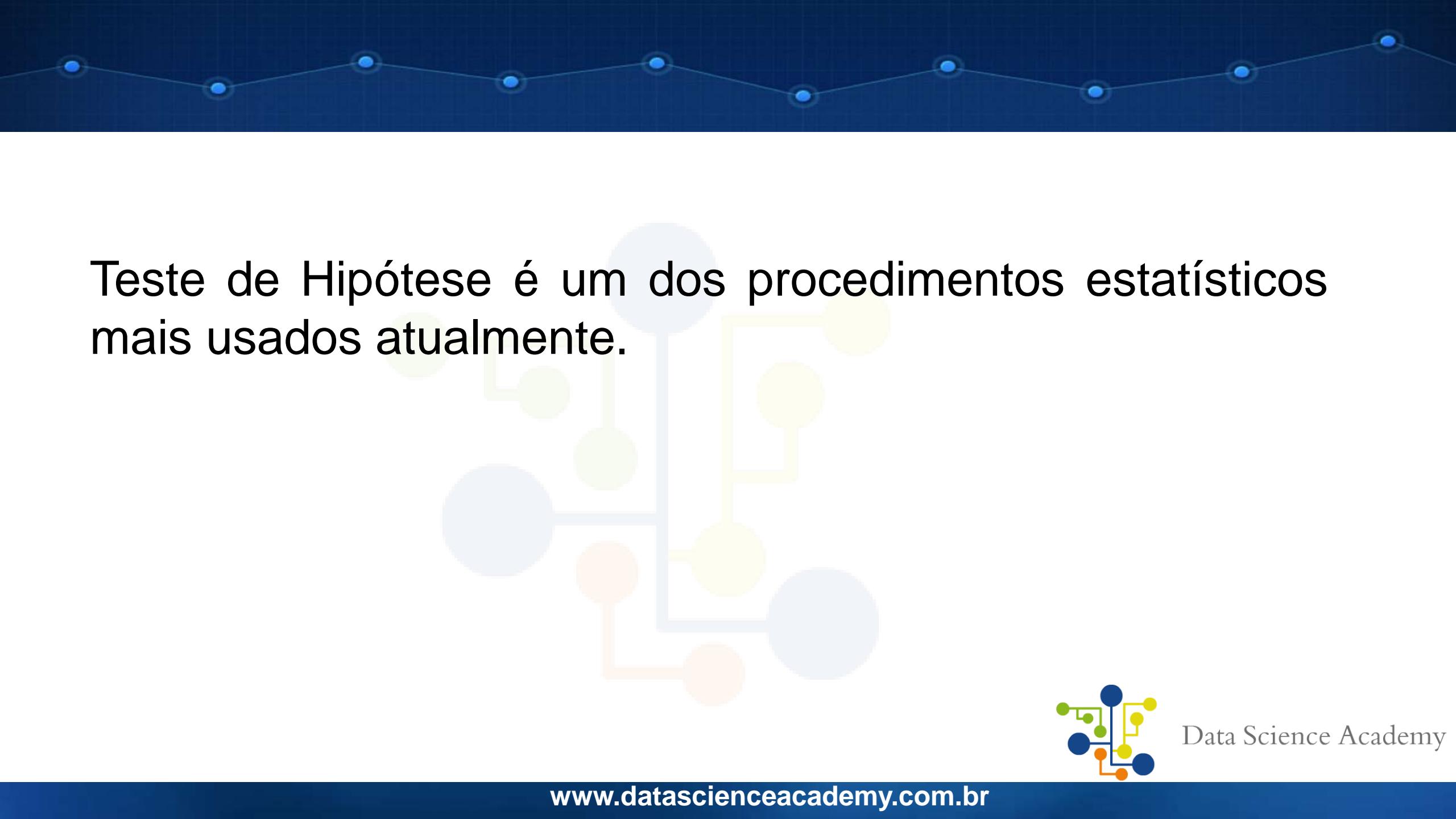


Data Science Academy

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



Data Science Academy



Teste de Hipótese é um dos procedimentos estatísticos
mais usados atualmente.



Data Science Academy



E agora? ficou mais fácil o entendimento acerca do que é
Teste de Hipótese?



Data Science Academy

Então me responda:



Data Science Academy

Qual a diferença entre Teste de Hipótese e Teste de Significância



Data Science Academy

Estimação e Teste de Hipótese = Teste de Significância

Aspectos Principais

Inferência Estatística



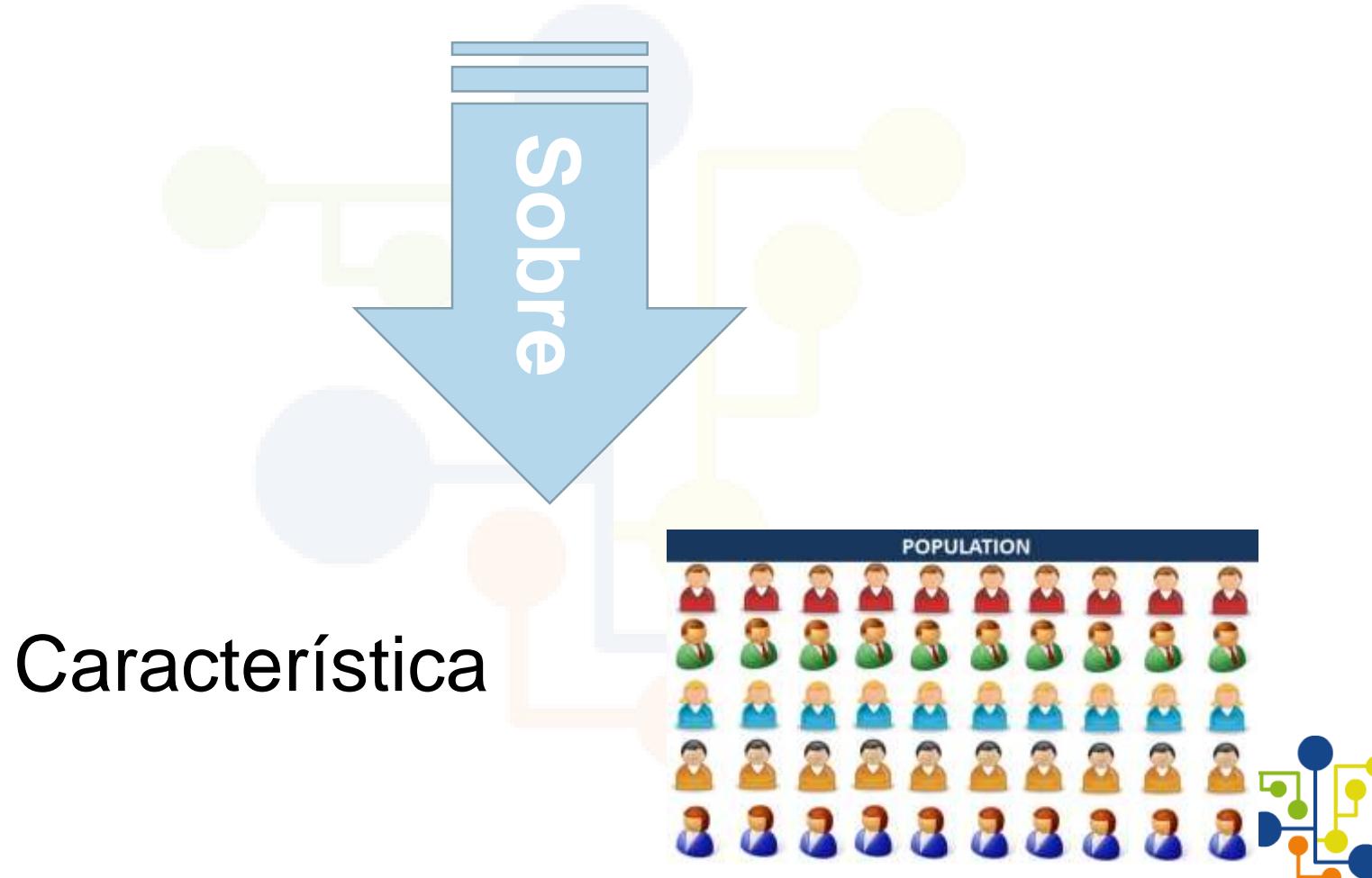
Data Science Academy

Objetivo Teste de Hipótese **é** decidir **se** determinada
afirmação sobre **1** parâmetro populacional
é **OU** não **apoiada** pela **evidência**
Obtida de **dados** amostrais



Data Science Academy

Hipótese **≡** Alegação



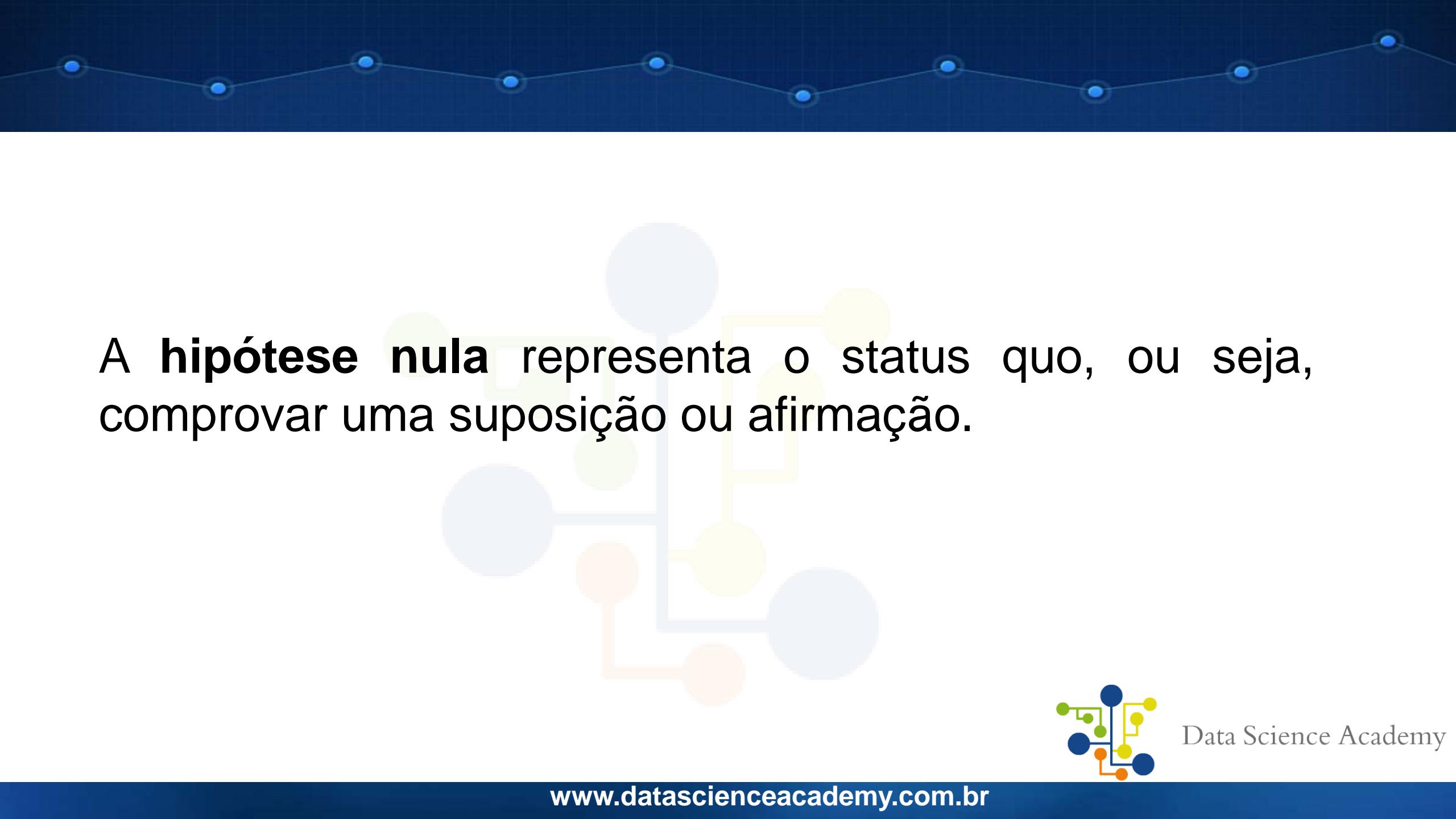
Cada teste de hipótese tem uma hipótese nula e uma hipótese alternativa, representados por:

H_0 - Hipótese nula

H_A - Hipótese alternativa



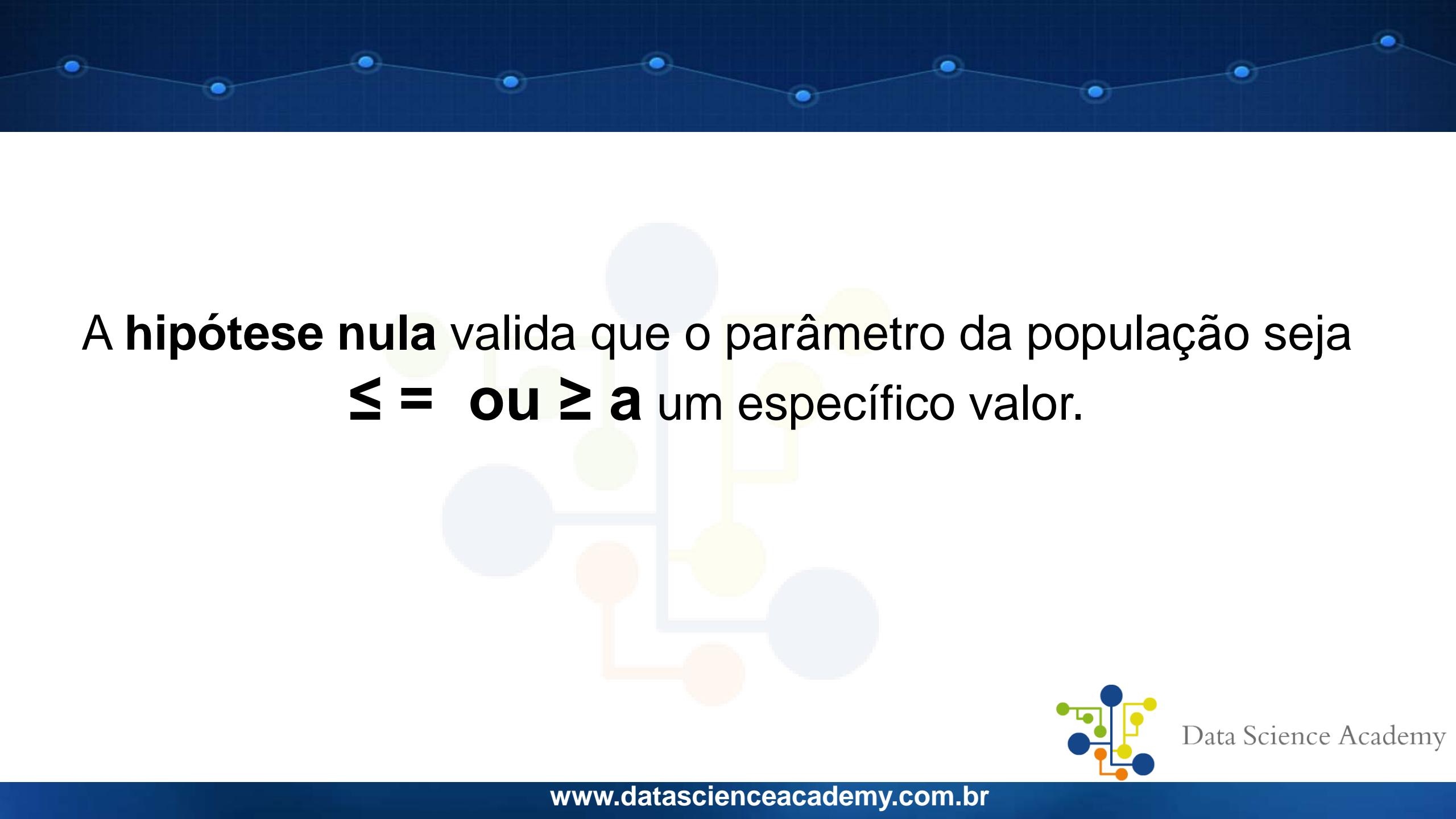
Data Science Academy



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



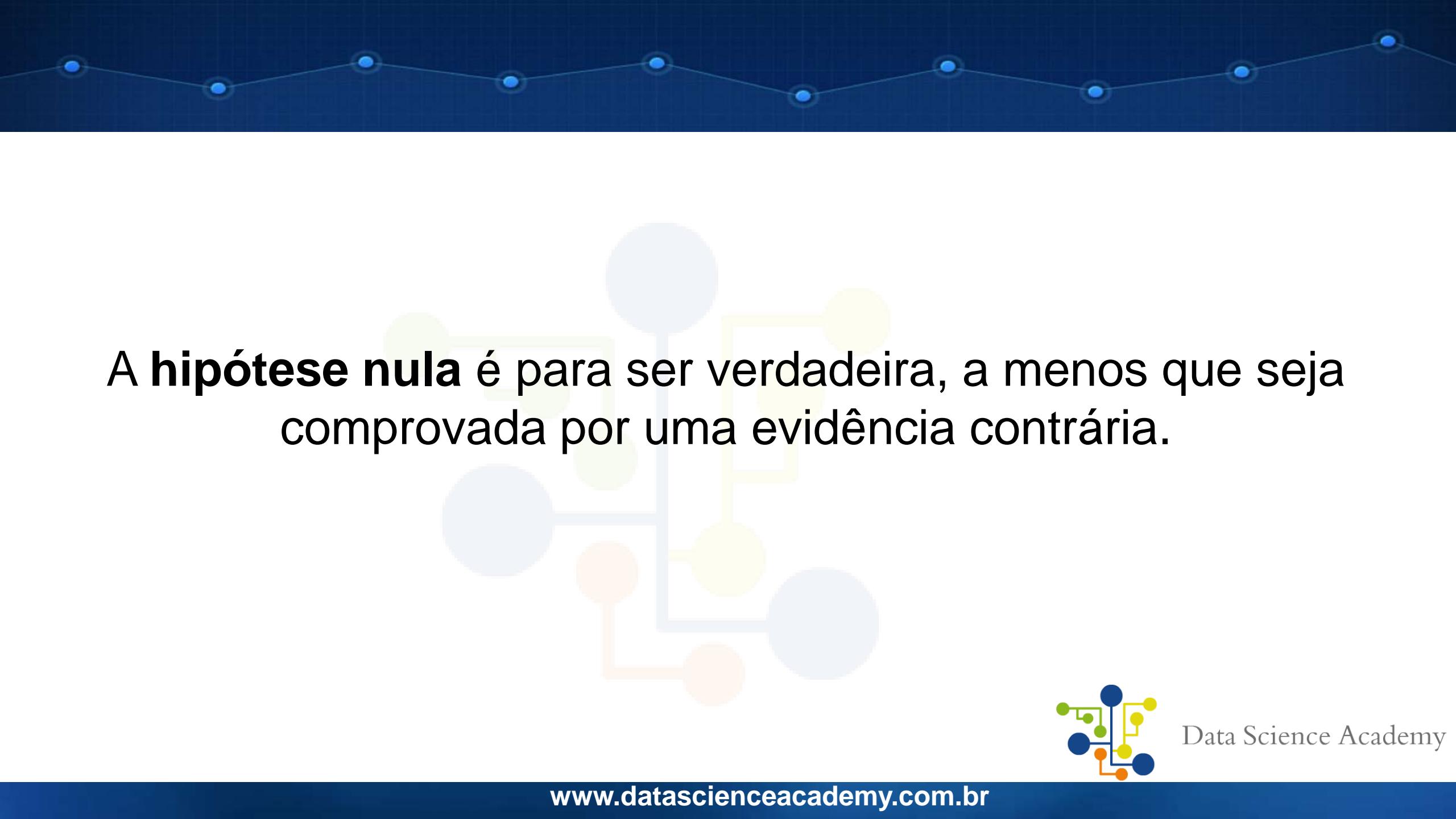
Data Science Academy



A **hipótese nula** valida que o parâmetro da população seja
 \leq ou \geq a um específico valor.



Data Science Academy



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



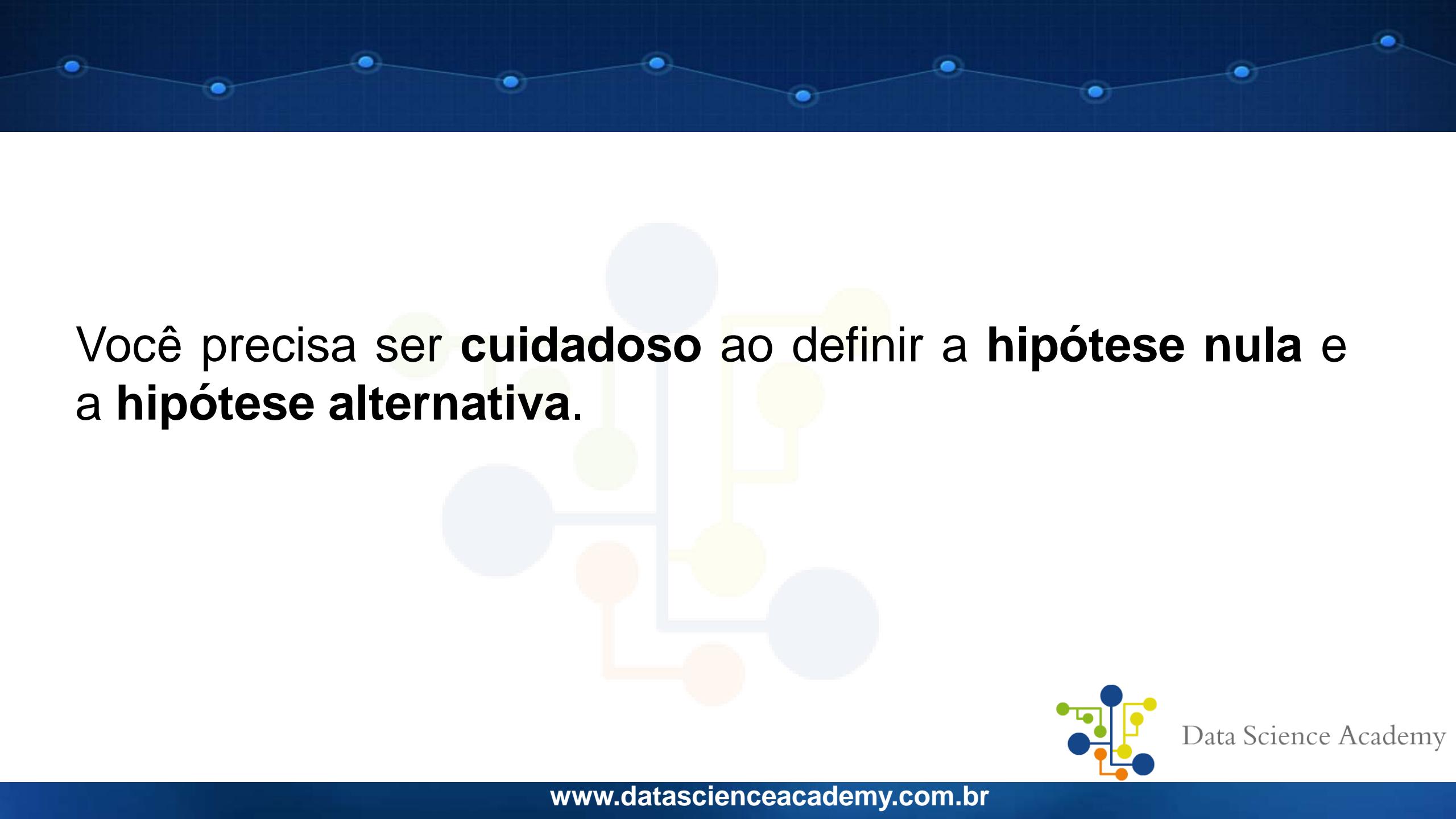
Data Science Academy



A hipótese alternativa representa o oposto da hipótese nula



Data Science Academy

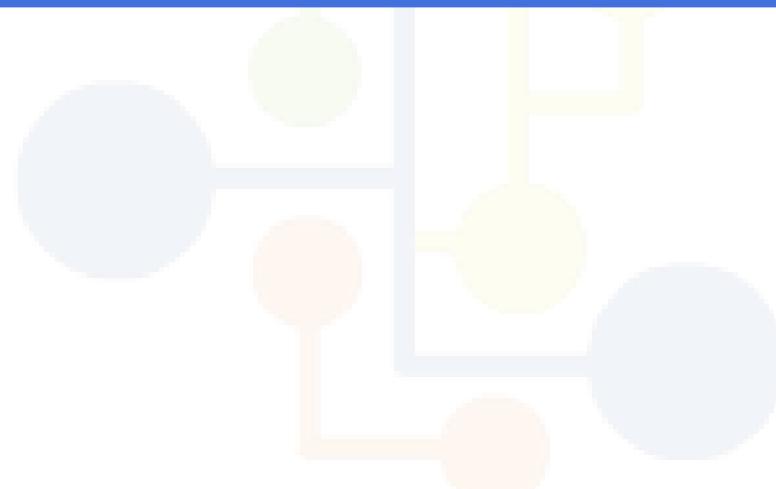


Você precisa ser **cuidadoso** ao definir a **hipótese nula** e a **hipótese alternativa**.



Data Science Academy

Exemplo I



Data Science Academy

Neste exemplo, nós estamos assumindo que os usuários de internet ficam em média 56 segundos em uma web page.



Data Science Academy

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

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

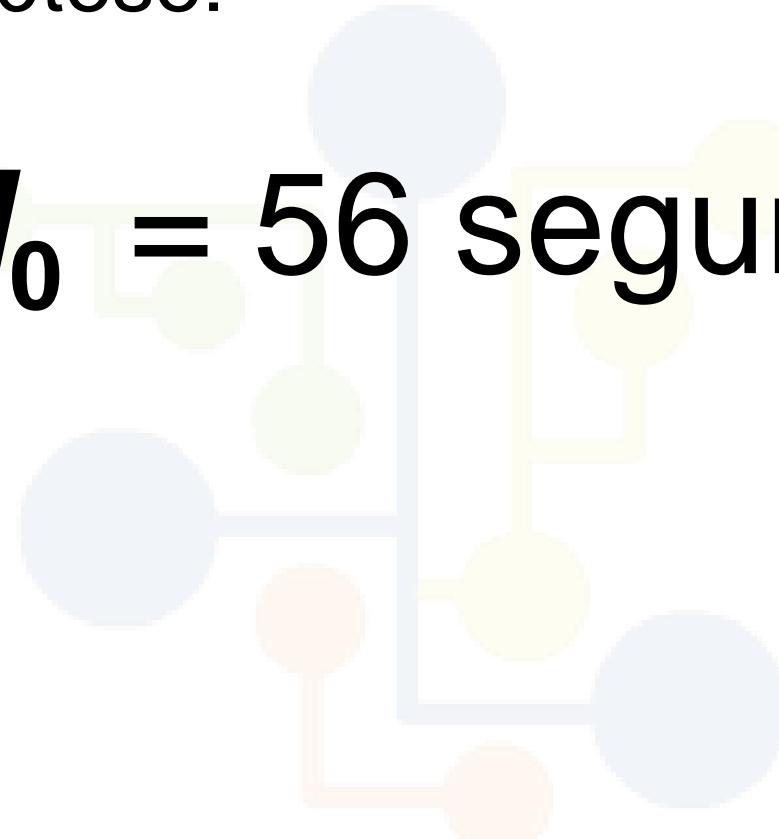
(status quo)



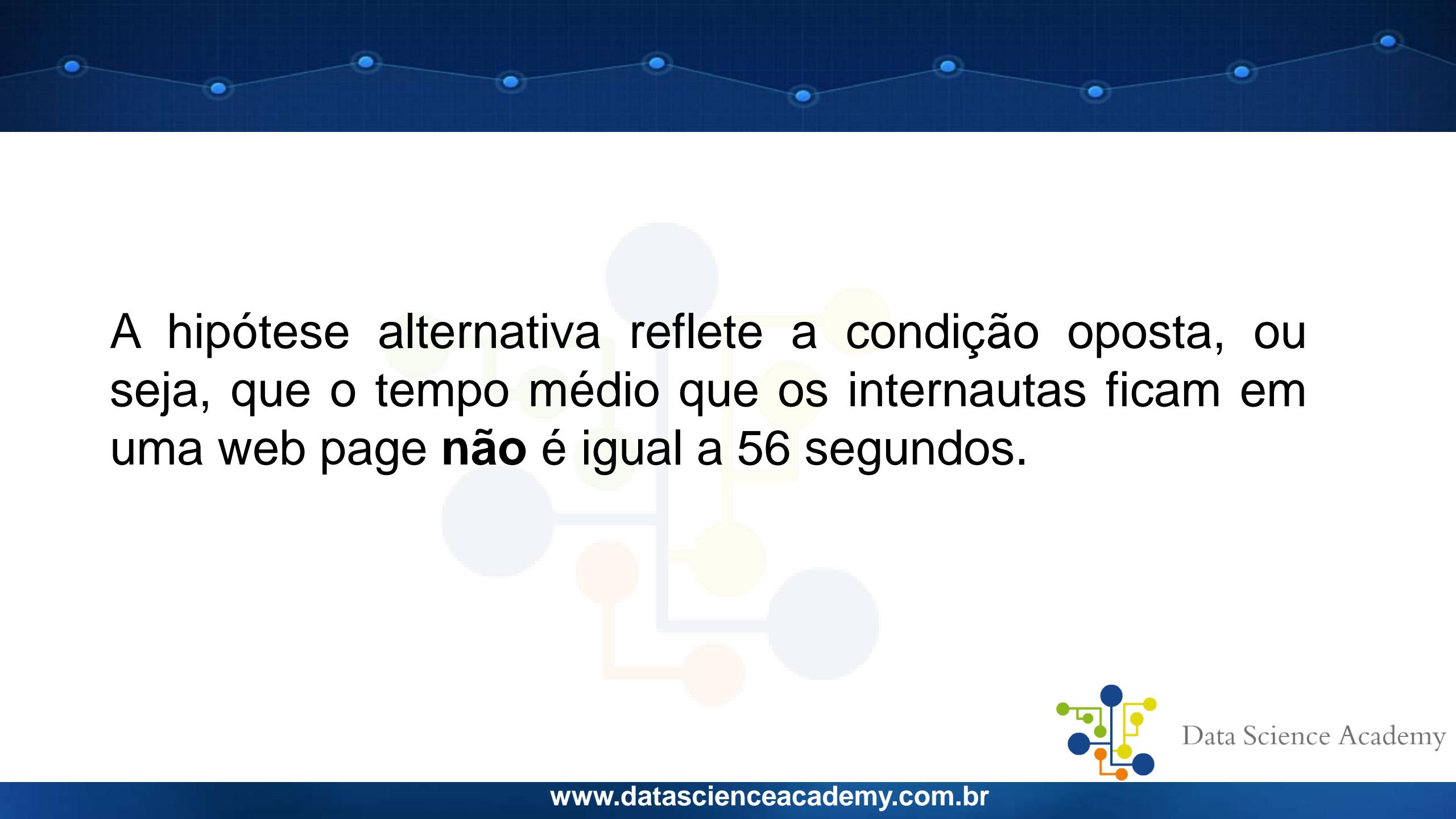
Data Science Academy

Assim nossa hipótese:

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



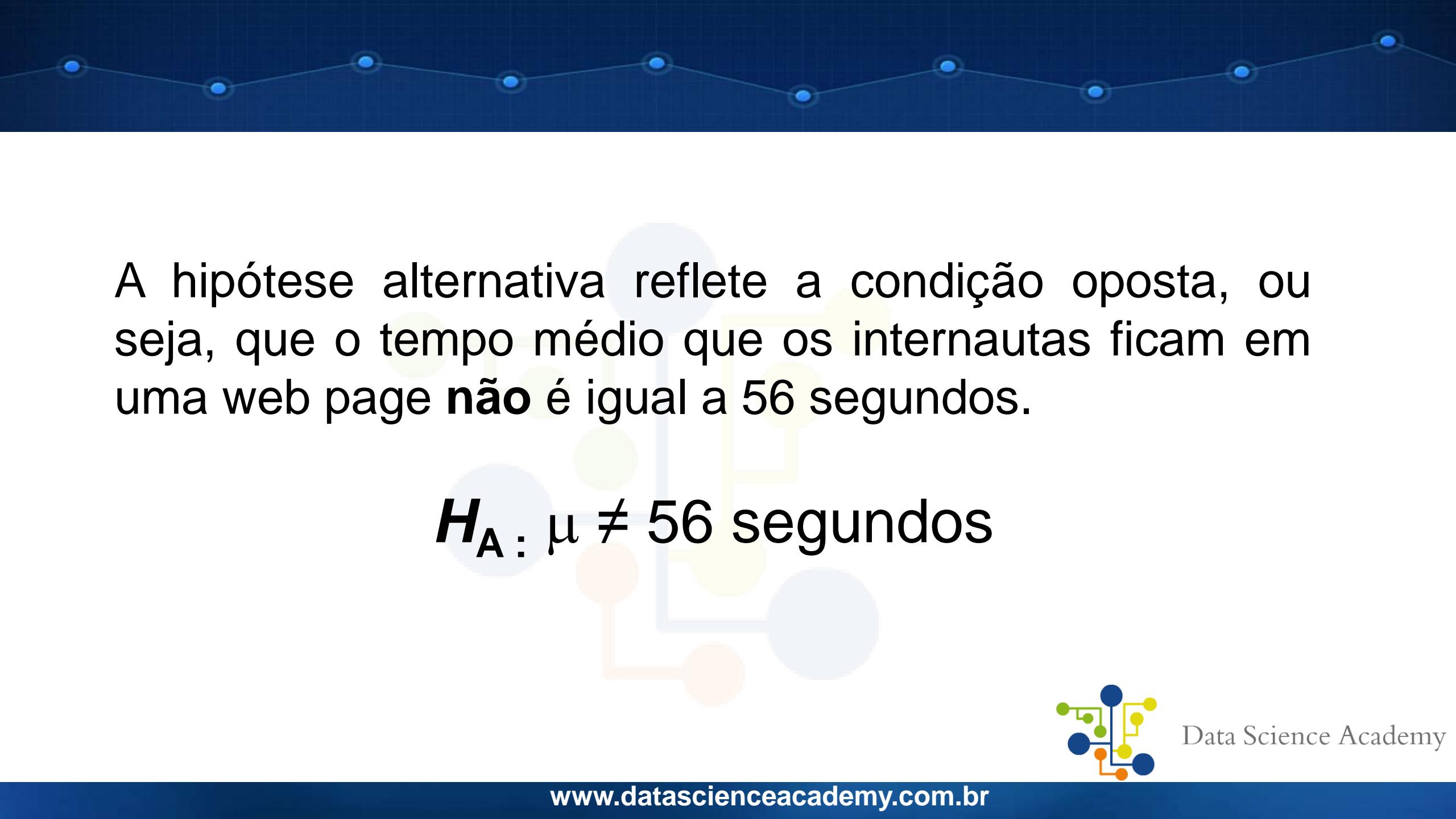
Data Science Academy



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.



Data Science Academy



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.

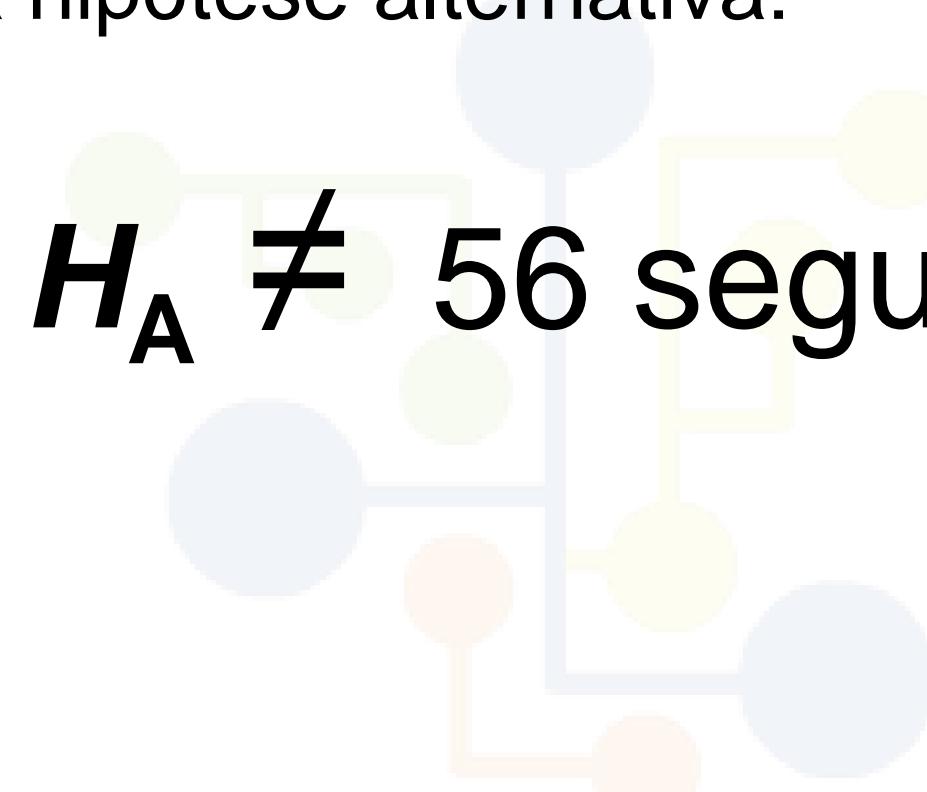
$$H_A: \mu \neq 56 \text{ segundos}$$



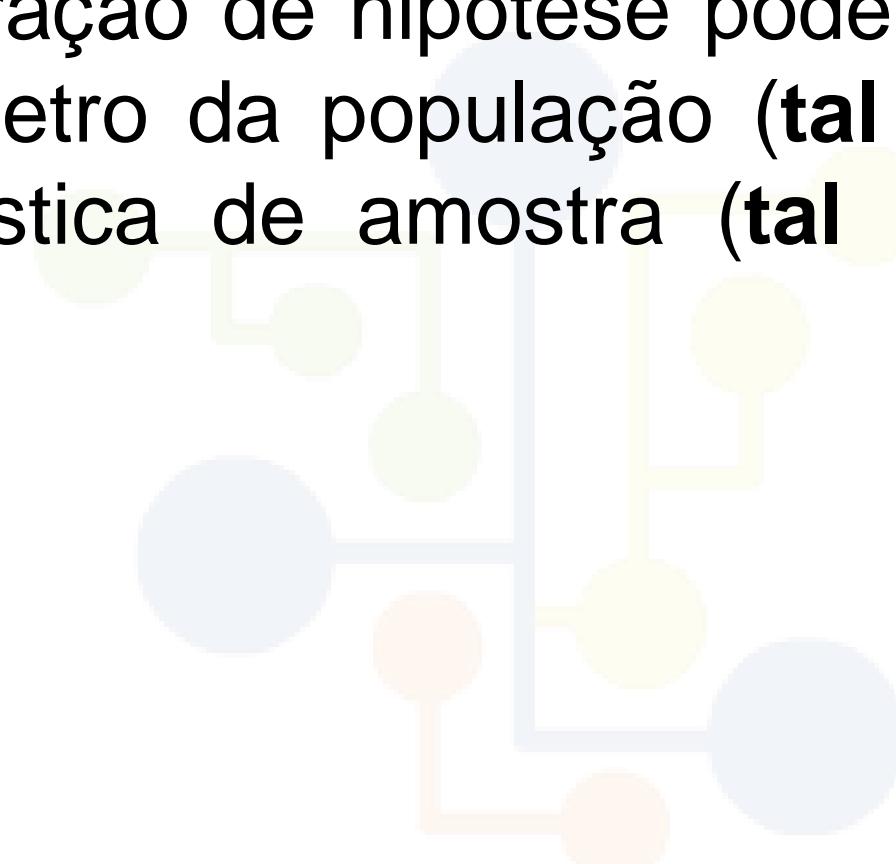
Data Science Academy



Assim nossa hipótese alternativa:


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

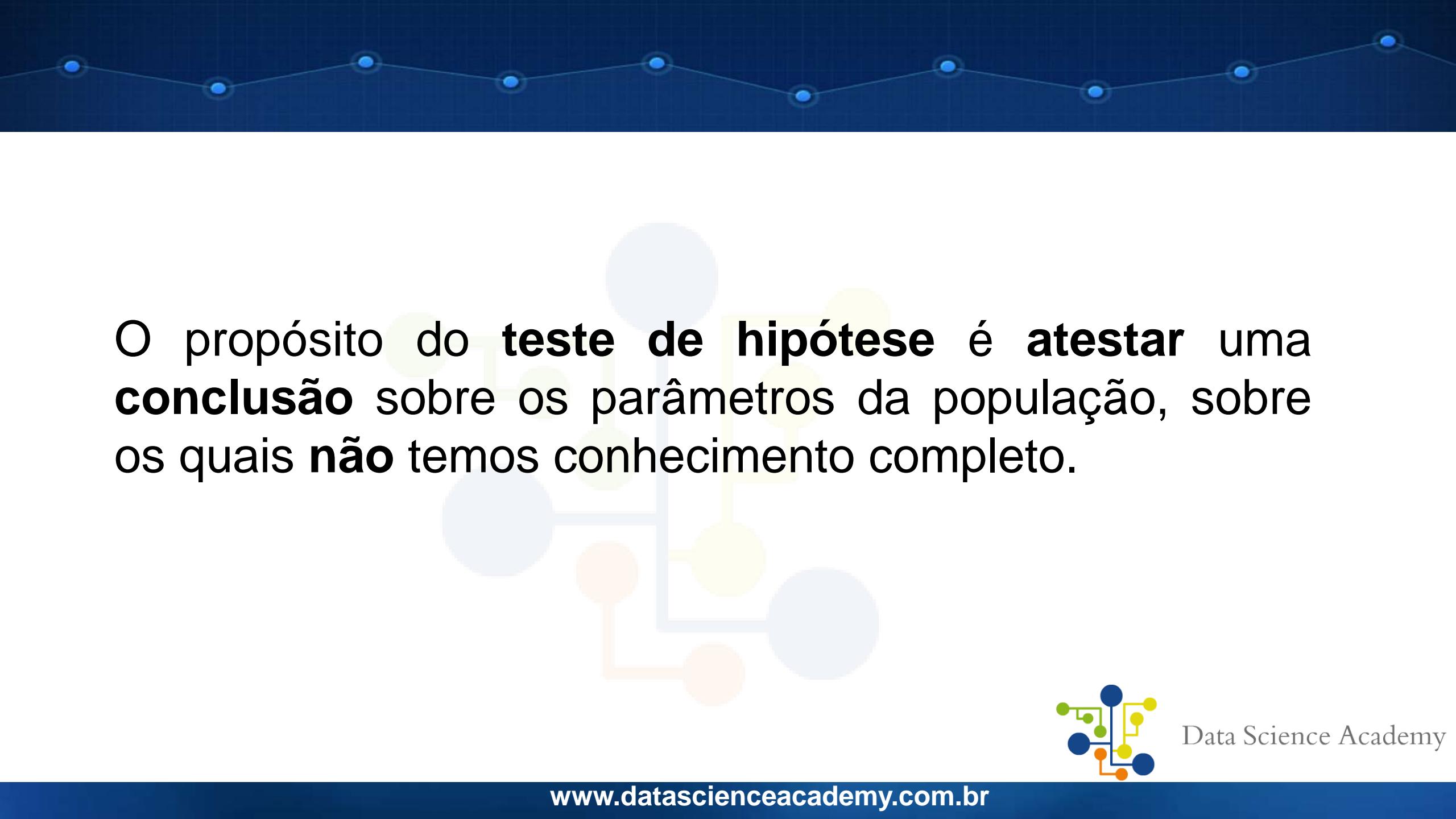

Data Science Academy



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**).



Data Science Academy

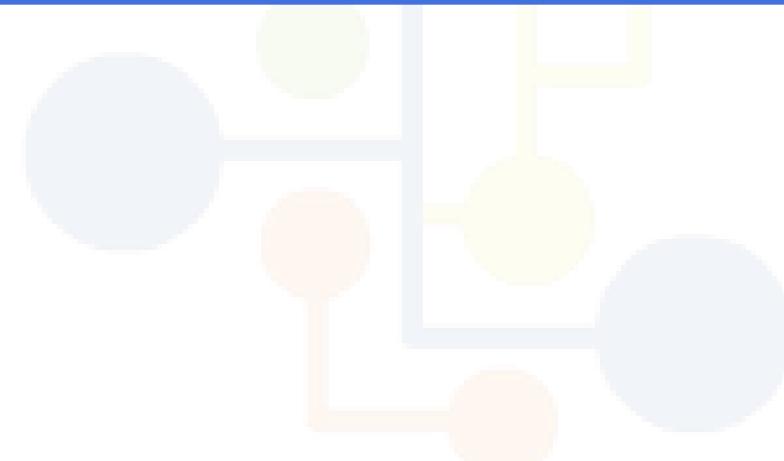


O propósito do **teste de hipótese** é atestar uma **conclusão** sobre os parâmetros da população, sobre os quais **não** temos conhecimento completo.



Data Science Academy

Exemplo II



Data Science Academy

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



Data Science Academy

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

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



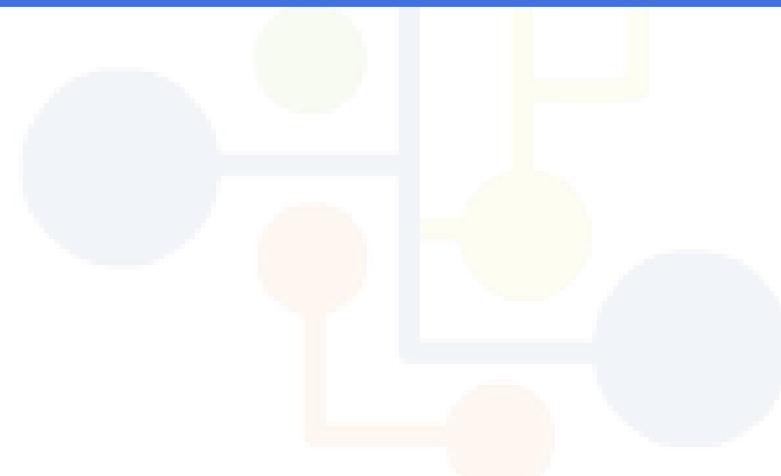
Data Science Academy

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 ao **teste de hipótese nula**.



Data Science Academy

Mais Exemplos



Data Science Academy

Definir as hipóteses nulas e alternativas, 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.

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

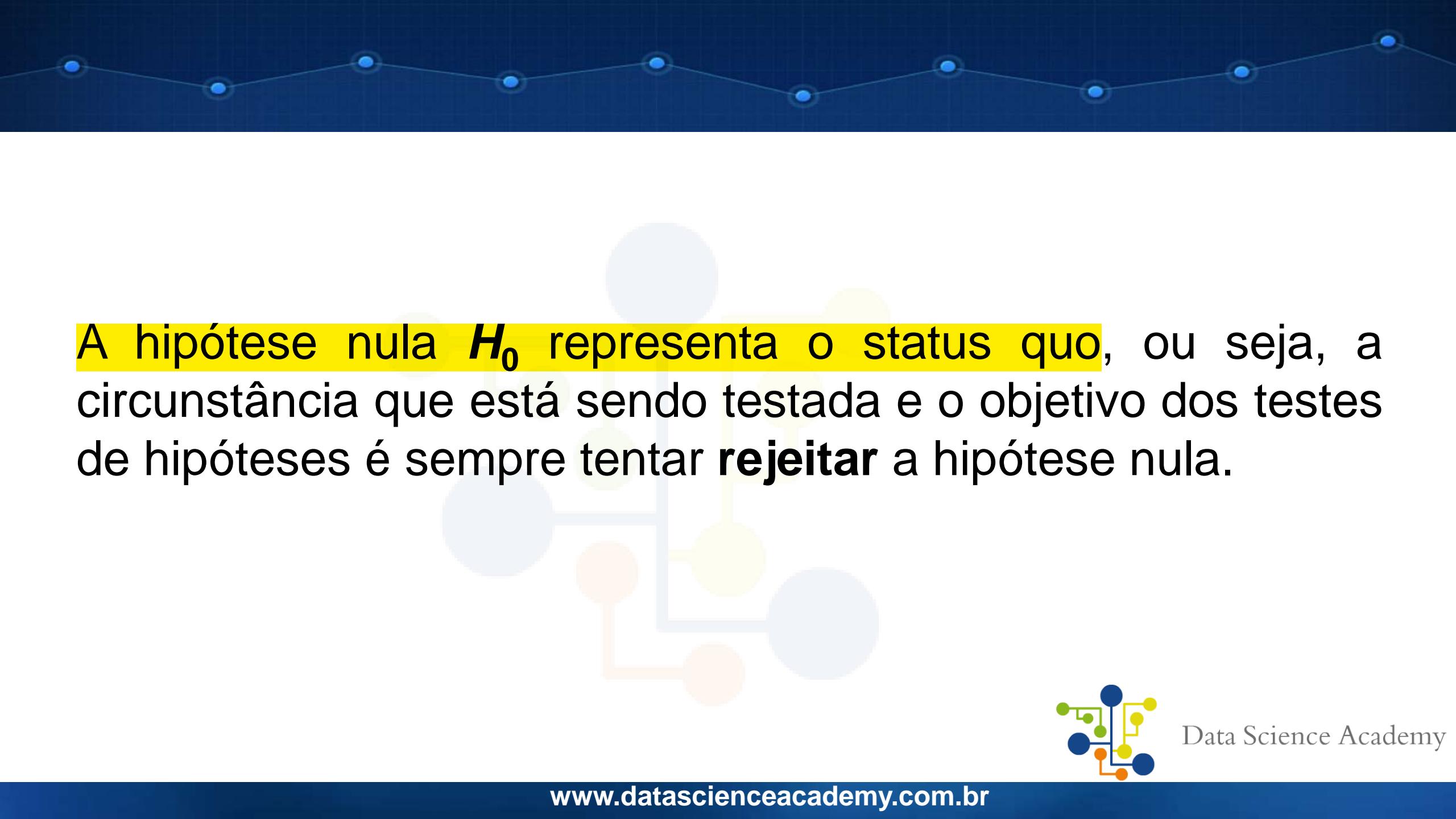
Este teste seria usado por alguém que acredita que o uso de banda larga diminuiu e quer suportar que a média de uso de banda larga é agora menor que 1.8 GB por mês.

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

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.



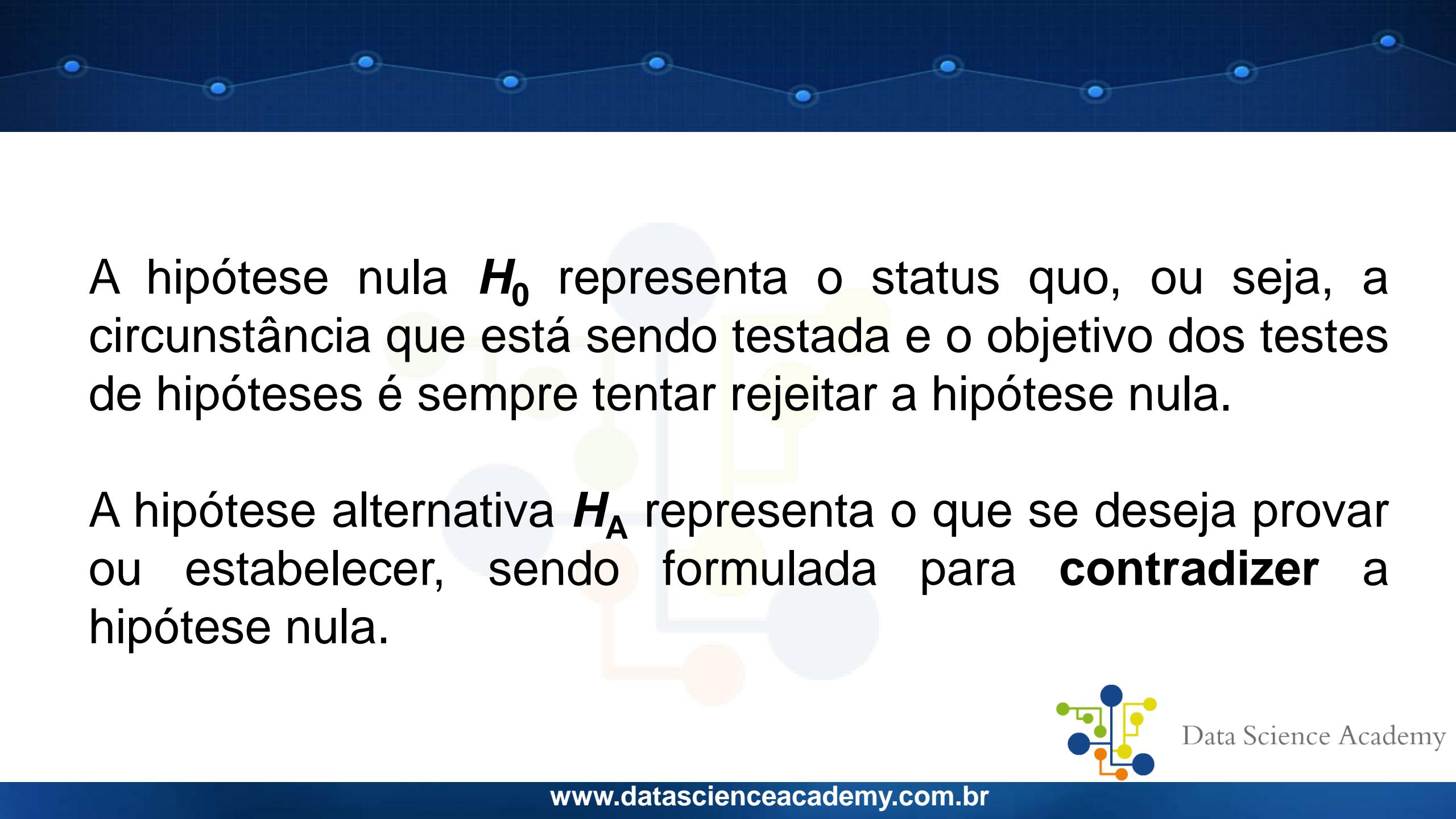
Data Science Academy



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



Data Science Academy



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

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



Data Science Academy



A Lógica do Teste de Hipótese



Data Science Academy

Existem apenas 2 afirmações que podemos fazer sobre a hipótese nula:

Rejeitar

Não Rejeitar

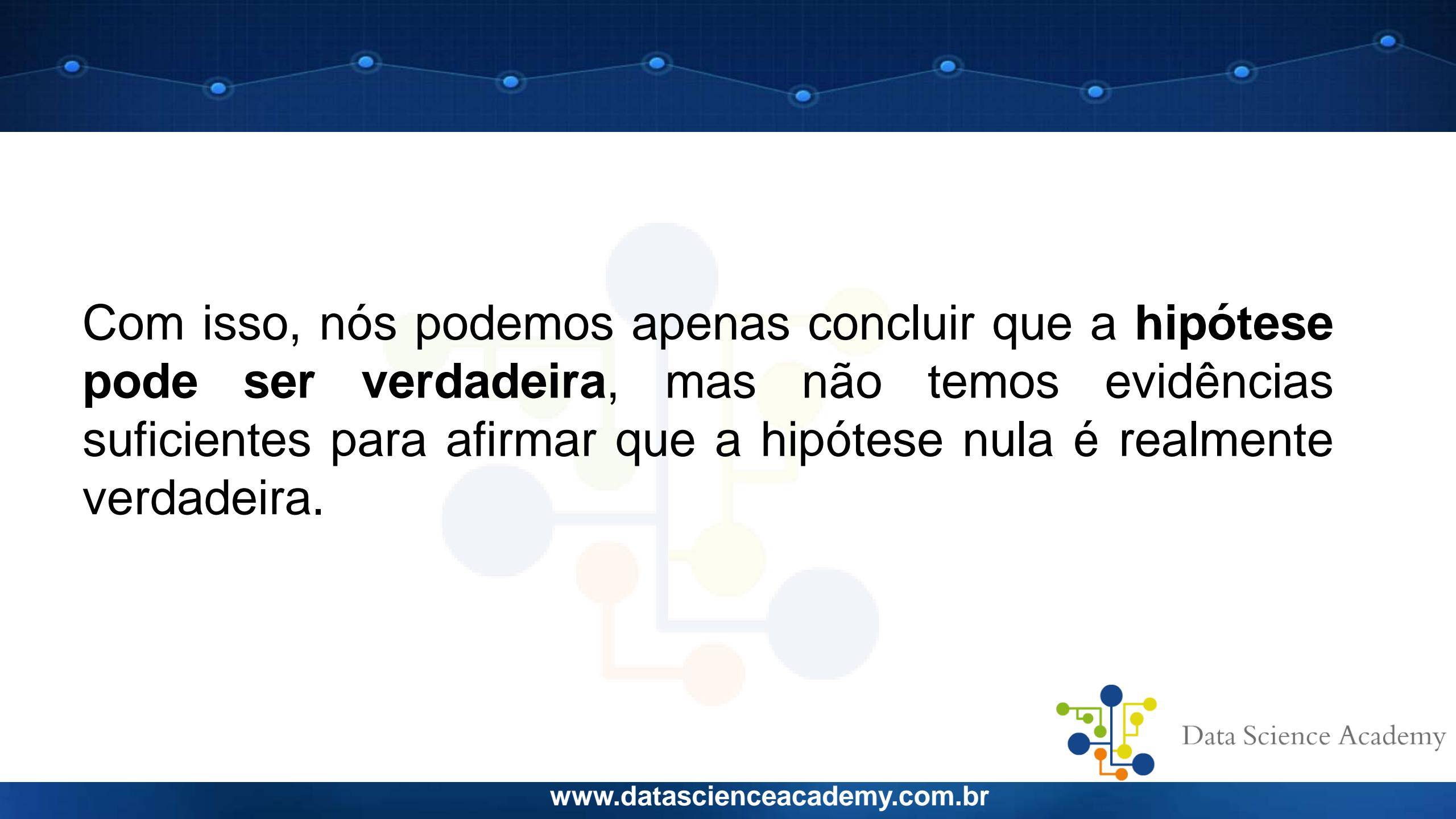


Data Science Academy

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



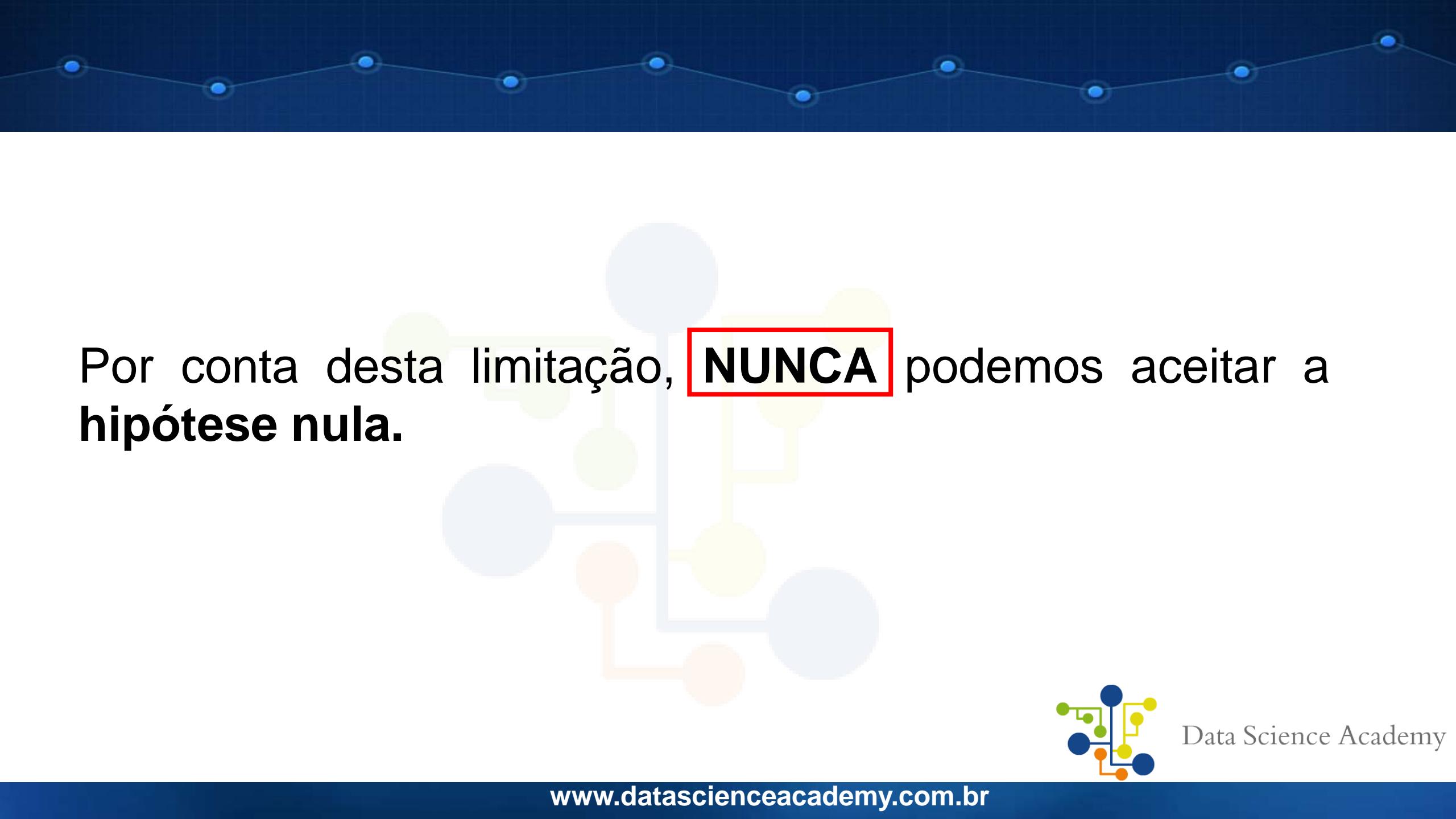
Data Science Academy



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.



Data Science Academy



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



Data Science Academy

Podemos apenas dizer que:

**não há evidências suficientes para rejeitar
a hipótese nula.**



Data Science Academy



Qual das hipóteses devo escolher?

H_0 = nula



H_A = Alternativa



Data Science Academy

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



Data Science Academy

Cabe a você, como pesquisador, a responsabilidade de escolher o teste mais apropriado.



Data Science Academy

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

Exemplo:

$$H_0: \mu = 500$$

$$H_A: \mu \neq 500$$



Data Science Academy

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

Exemplo:

$$H_0: \mu = 1.8$$

$$H_A: \mu < 1.8$$

$$H_A: \mu > 1.8$$

$$H_A: \mu \neq 1.8$$



Média de uso de banda larga é maior, menor ou diferente 1.8 GB por mês.



Data Science Academy

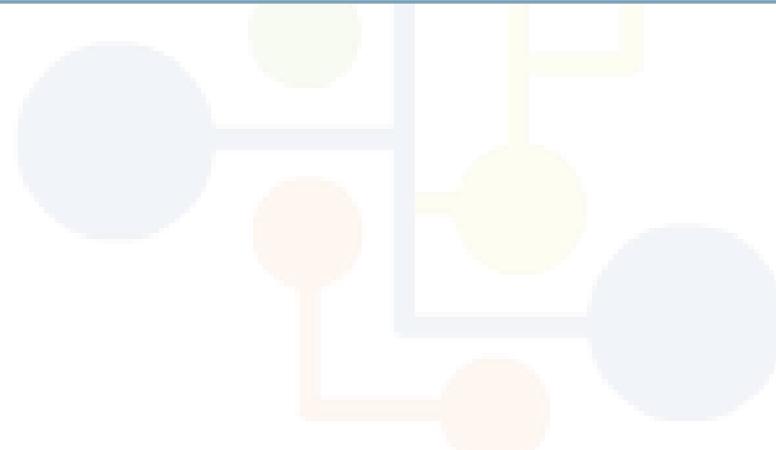
Esse tópico chegou ao final



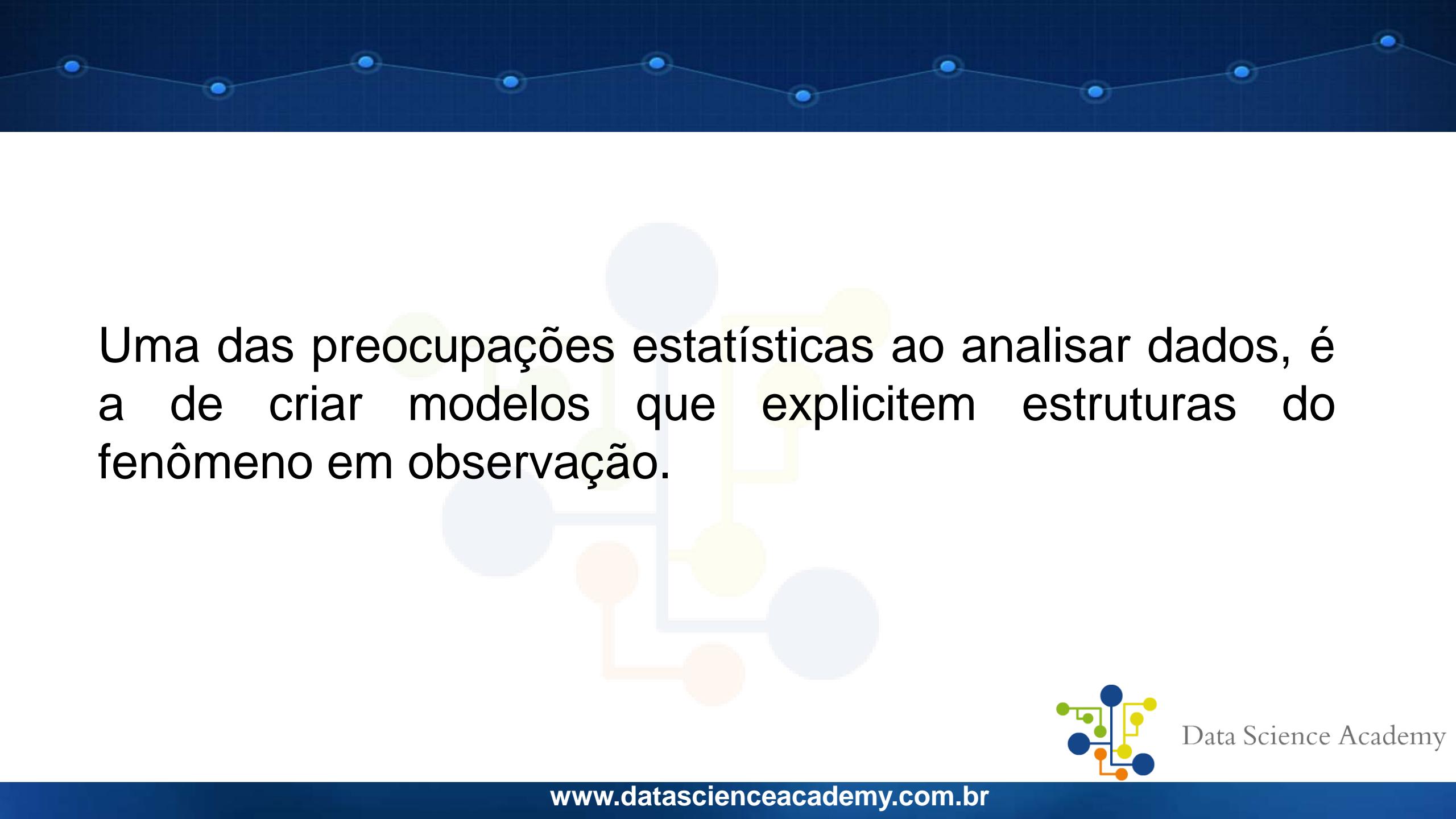
Data Science Academy



Regressão



Data Science Academy

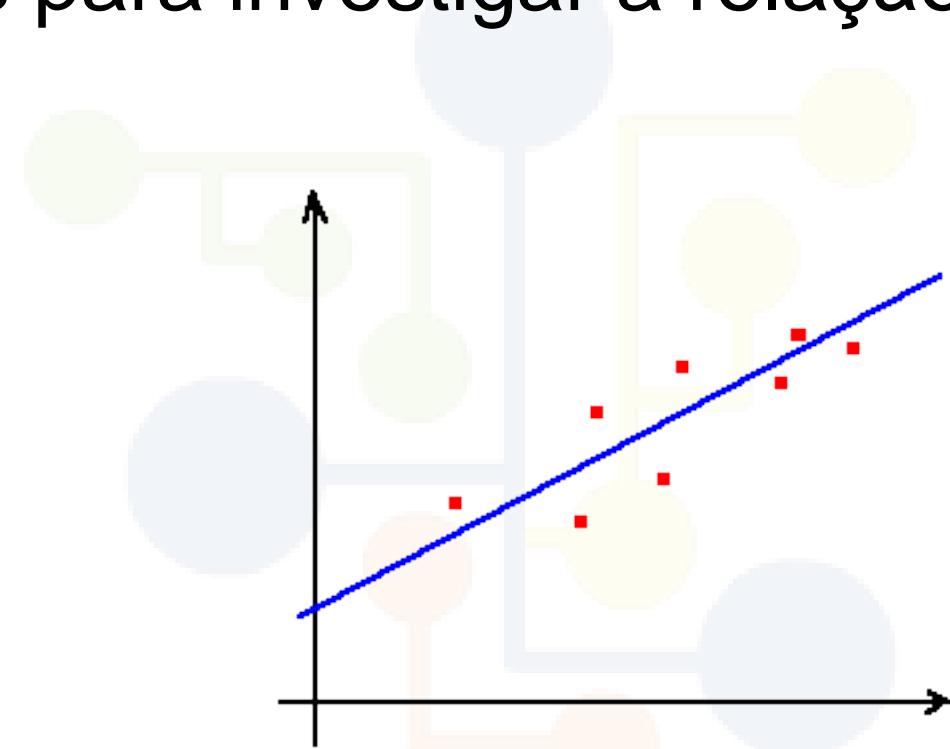


Uma das preocupações estatísticas ao analisar dados, é a de criar modelos que explicitem estruturas do fenômeno em observação.



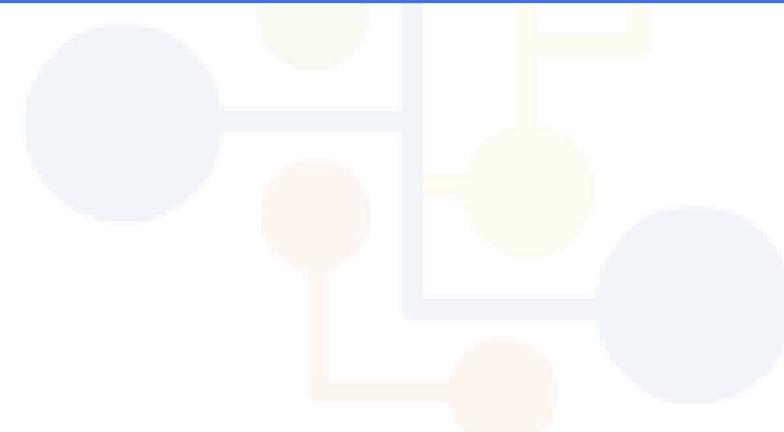
Data Science Academy

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

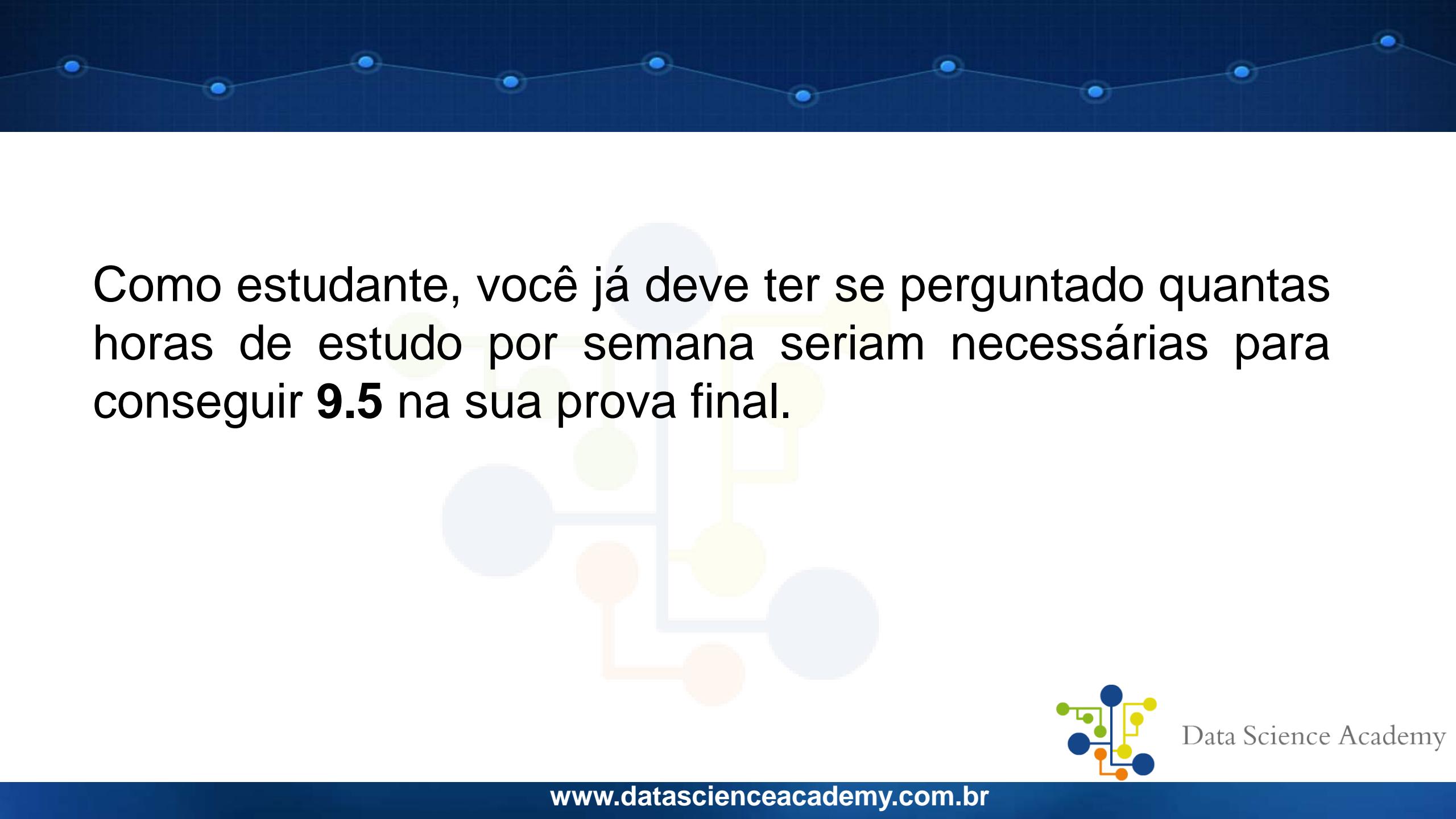


Data Science Academy

Exemplo



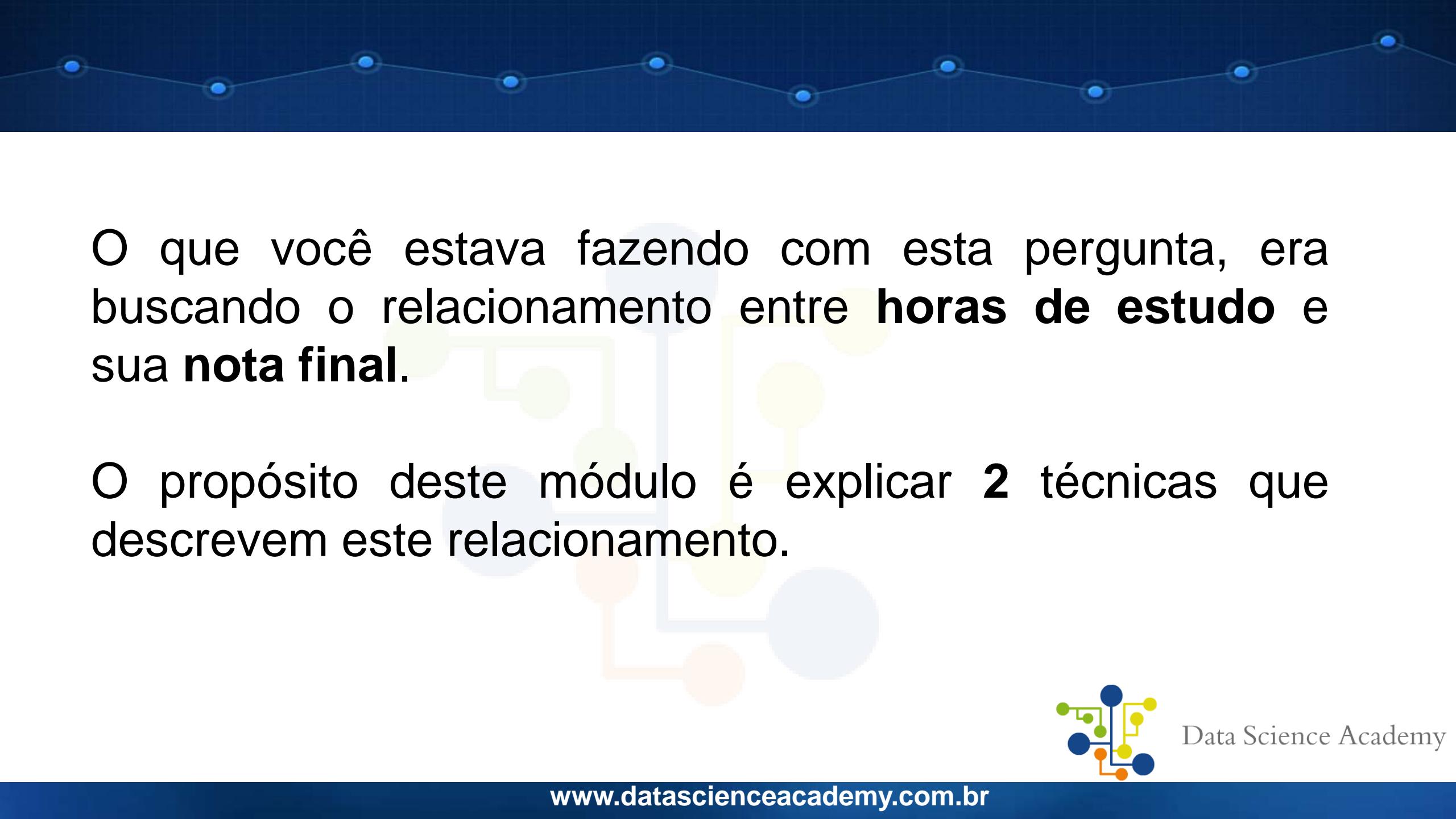
Data Science Academy



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



Data Science Academy



O que você estava fazendo com esta pergunta, era buscando o relacionamento entre **horas de estudo** e sua **nota final**.

O propósito deste módulo é explicar 2 técnicas que descrevem este relacionamento.

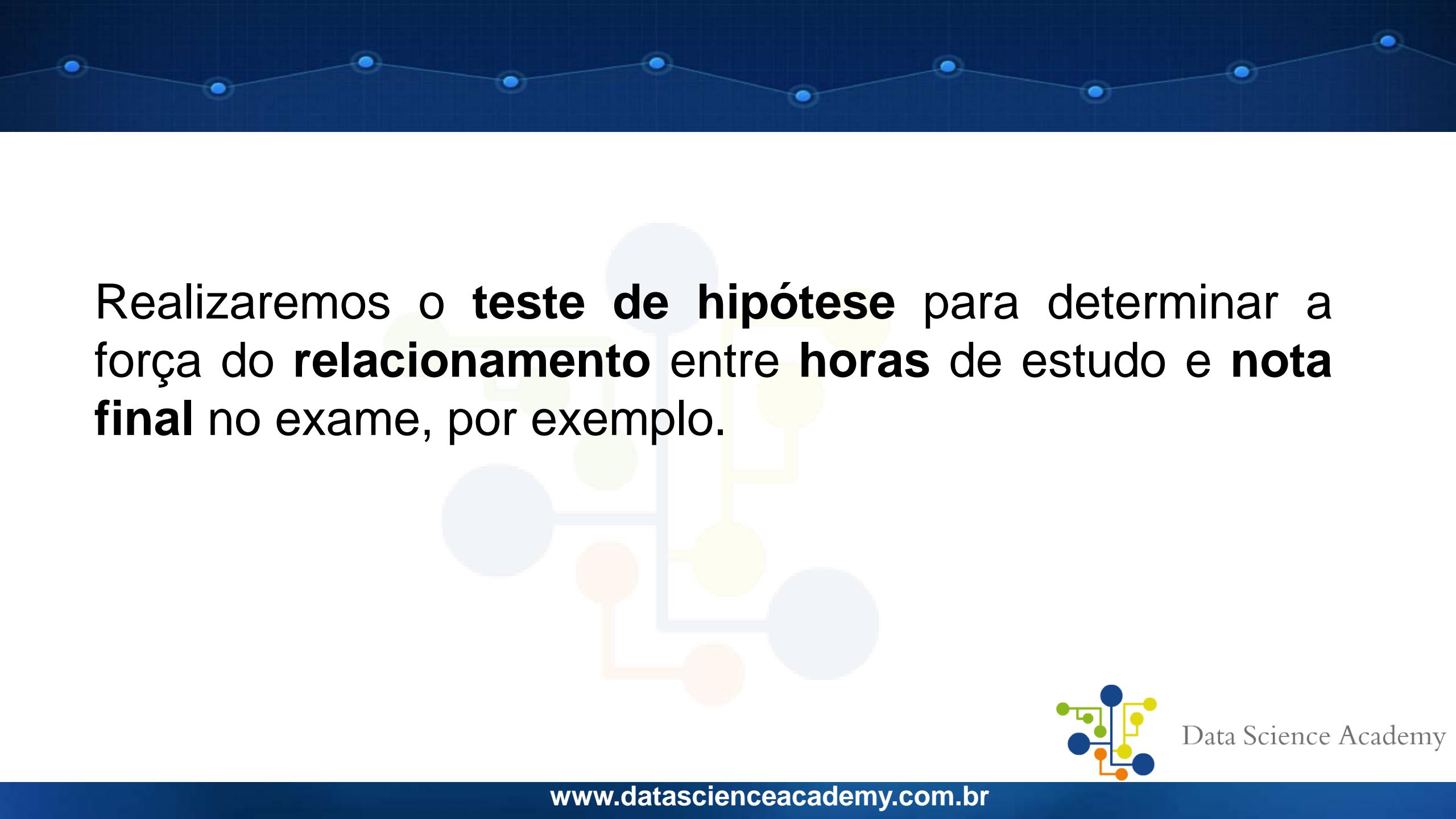


Data Science Academy

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



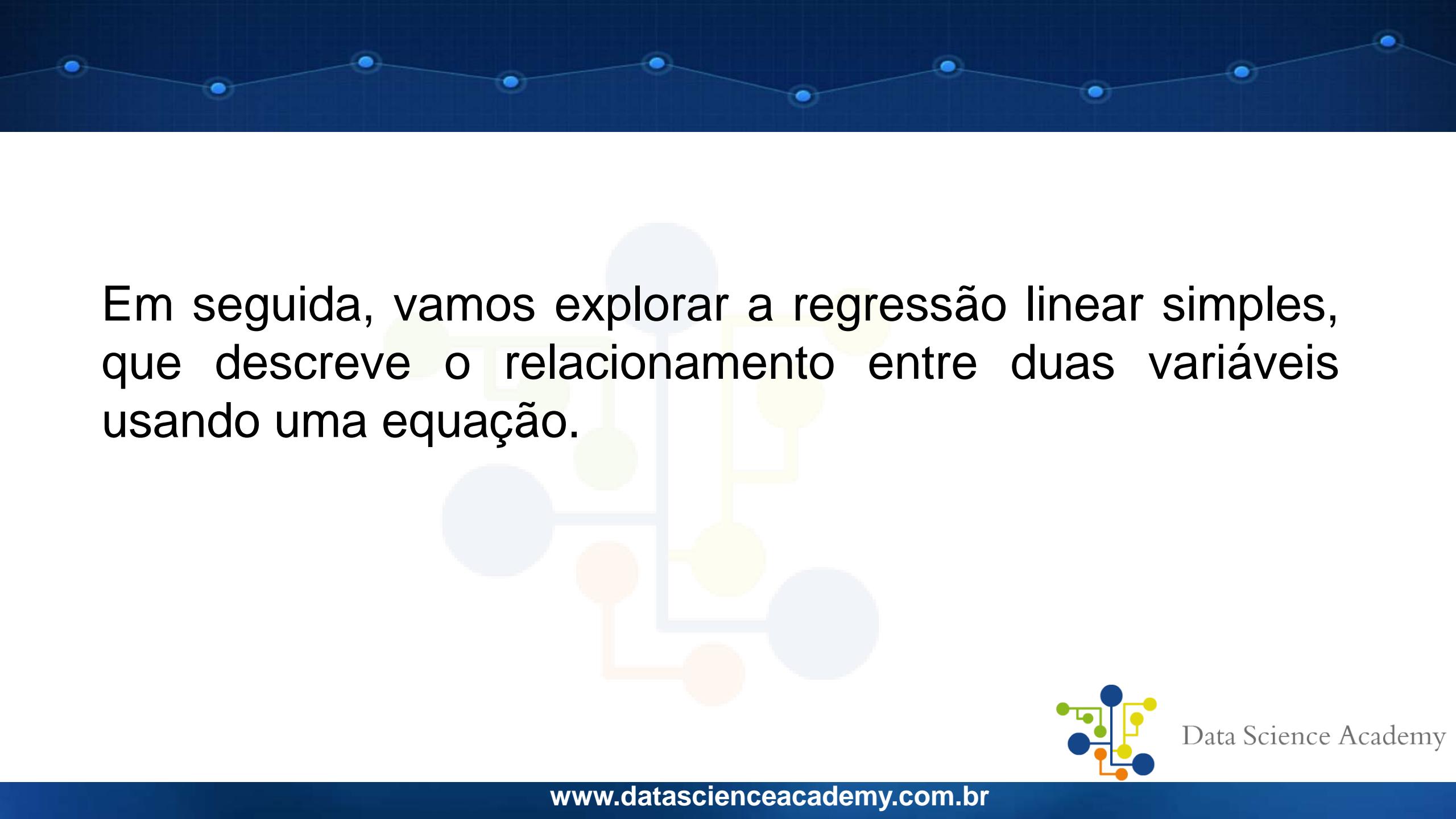
Data Science Academy



Realizaremos o **teste de hipótese** para determinar a força do **relacionamento** entre **horas** de estudo e **nota final** no exame, por exemplo.



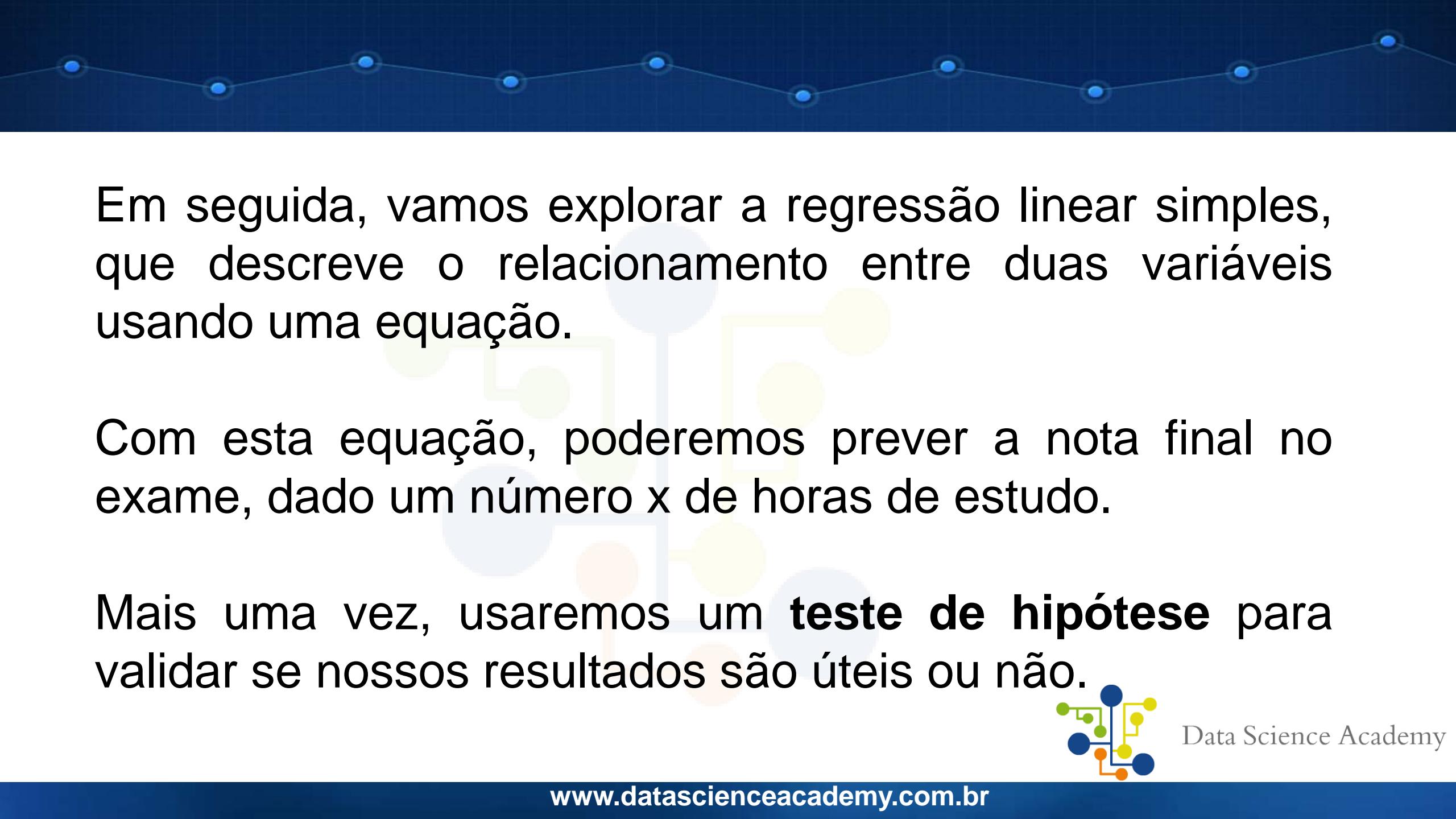
Data Science Academy



Em seguida, vamos explorar a regressão linear simples, que descreve o relacionamento entre duas variáveis usando uma equação.



Data Science Academy



Em seguida, vamos explorar a regressão linear simples, que descreve o relacionamento entre duas variáveis usando uma equação.

Com esta equação, poderemos prever a nota final no exame, dado um número x de horas de estudo.

Mais uma vez, usaremos um **teste de hipótese** para validar se nossos resultados são úteis ou não.



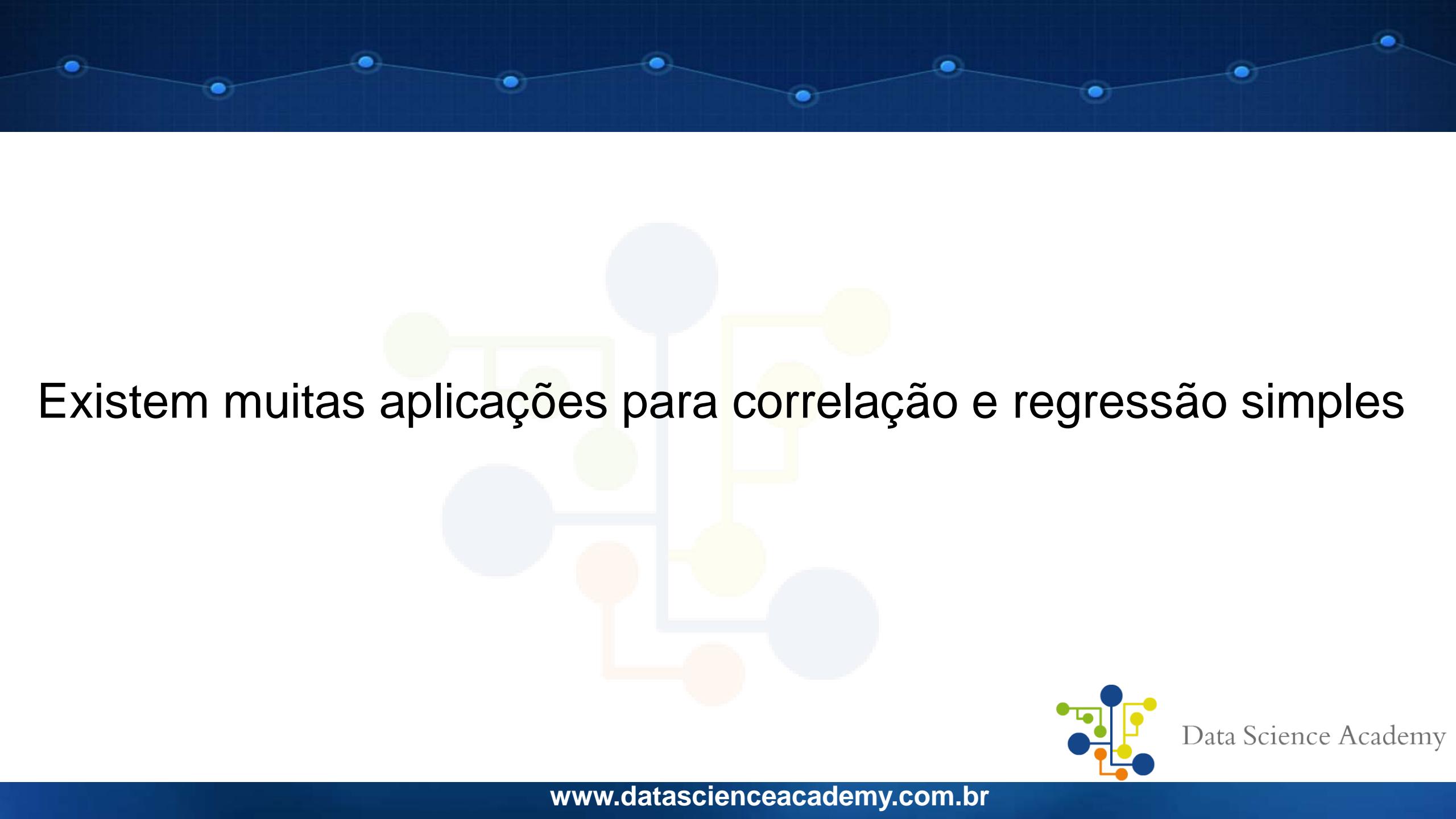
Data Science Academy



Mais Exemplos



Data Science Academy



Existem muitas aplicações para correlação e regressão simples



Data Science Academy



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



Data Science Academy



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.



Data Science Academy

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.



Data Science Academy



Variável Dependente e Independente



Data Science Academy

De forma bem simples: 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)

Esta afirmação ao contrário é falsa

Está

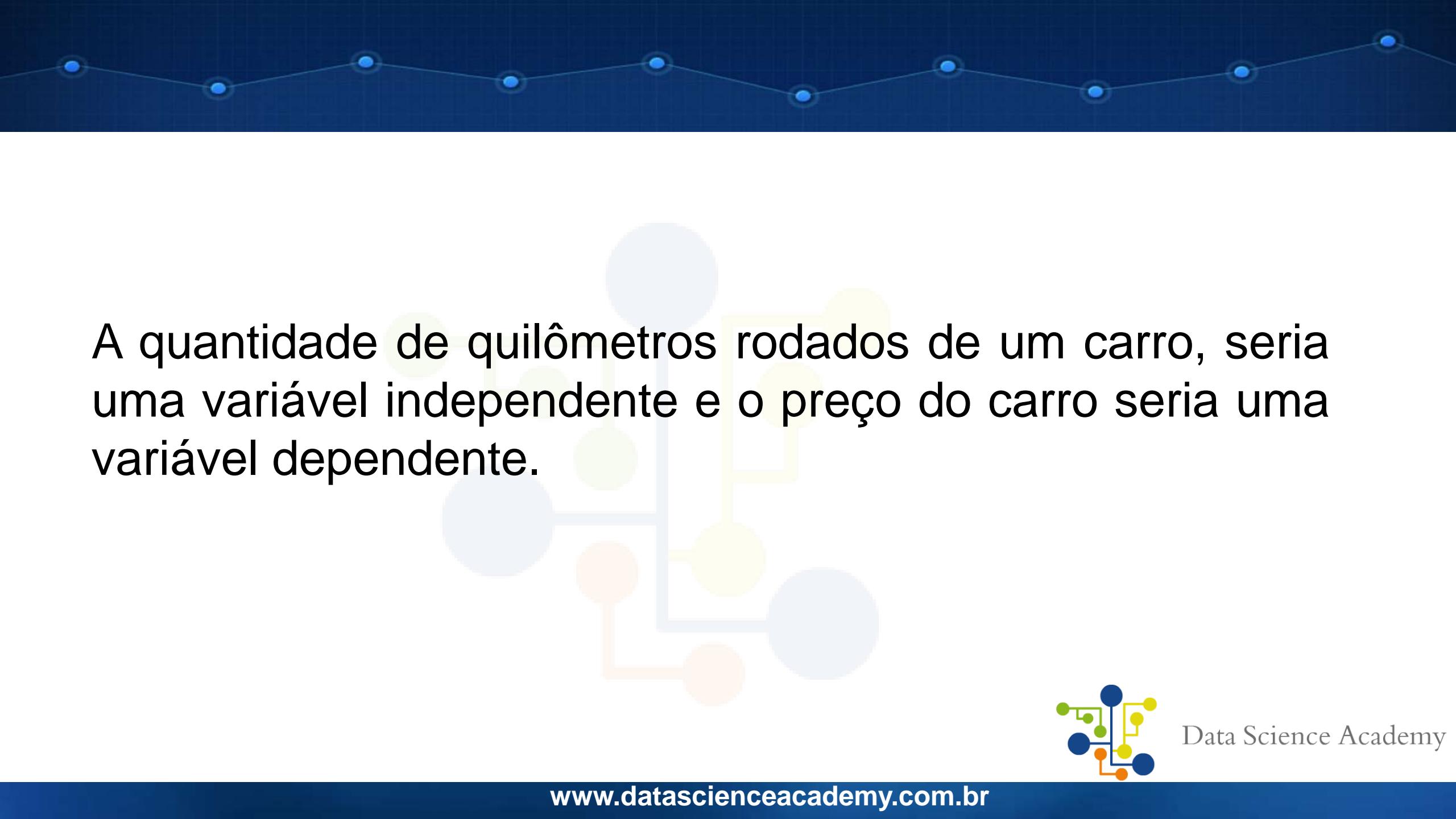


Data Science Academy

Exemplo



Data Science Academy



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



Data Science Academy



= R\$ 50.000,00

Variável
Dependente

85.309 Km

=

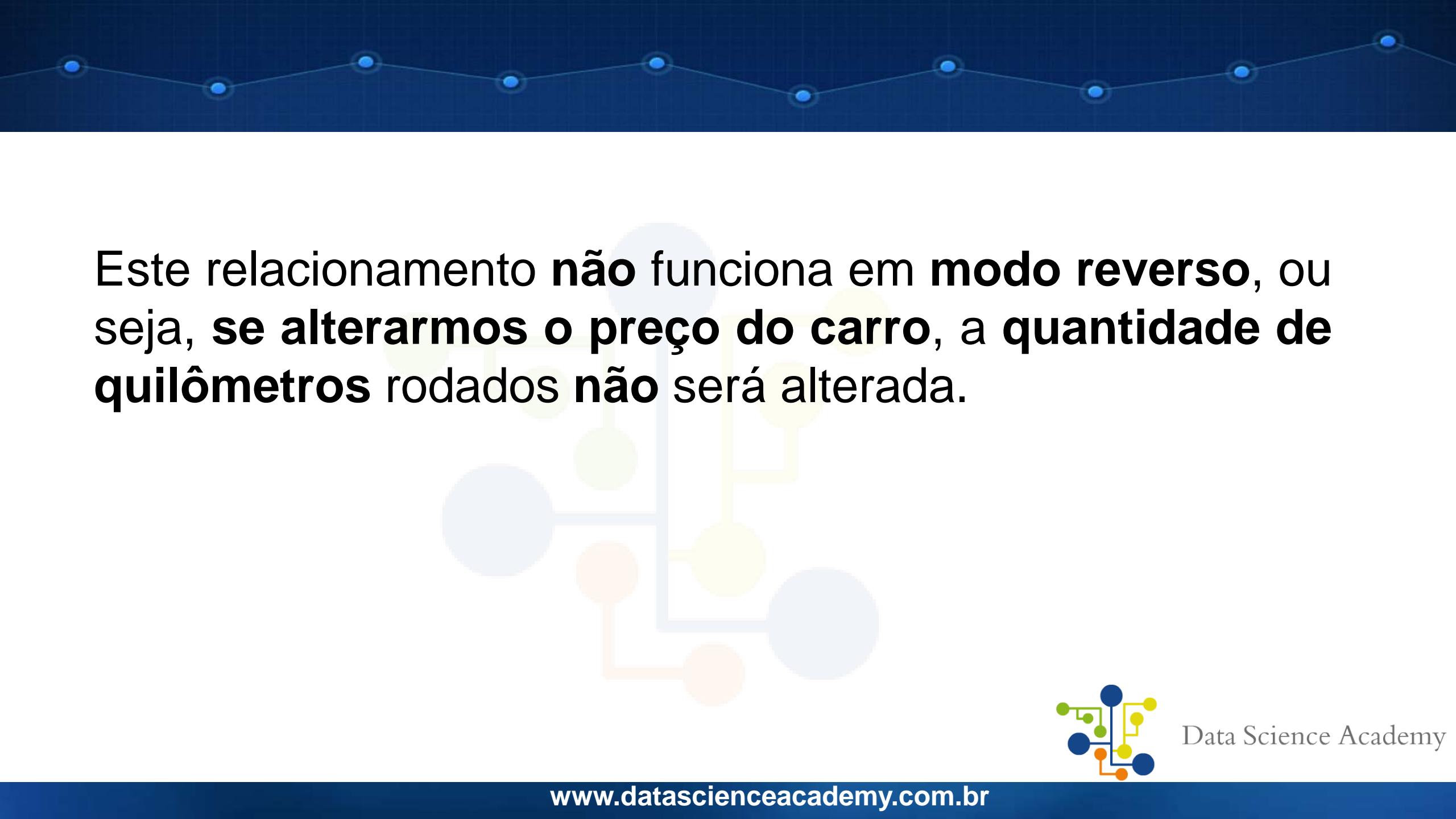


Variável
Independente

À medida que a quantidade de quilômetros rodados do carro aumenta, o preço do carro diminui.



Data Science Academy

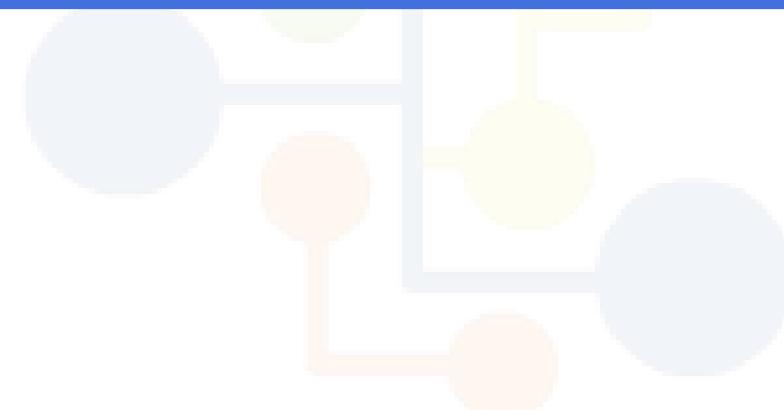


Este relacionamento **não** funciona em **modo reverso**, ou seja, **se alterarmos o preço do carro**, a **quantidade de quilômetros rodados** **não** será alterada.



Data Science Academy

Outros Exemplos



Data Science Academy

Variável independente (X)

O tamanho da tela de um monitor

Variável dependente (Y)

Preço do monitor

Número de visitantes em um web site

Quantidade de vendas no web site

Tempo de experiência profissional

Salário



Data Science Academy

Esse tópico chegou ao final



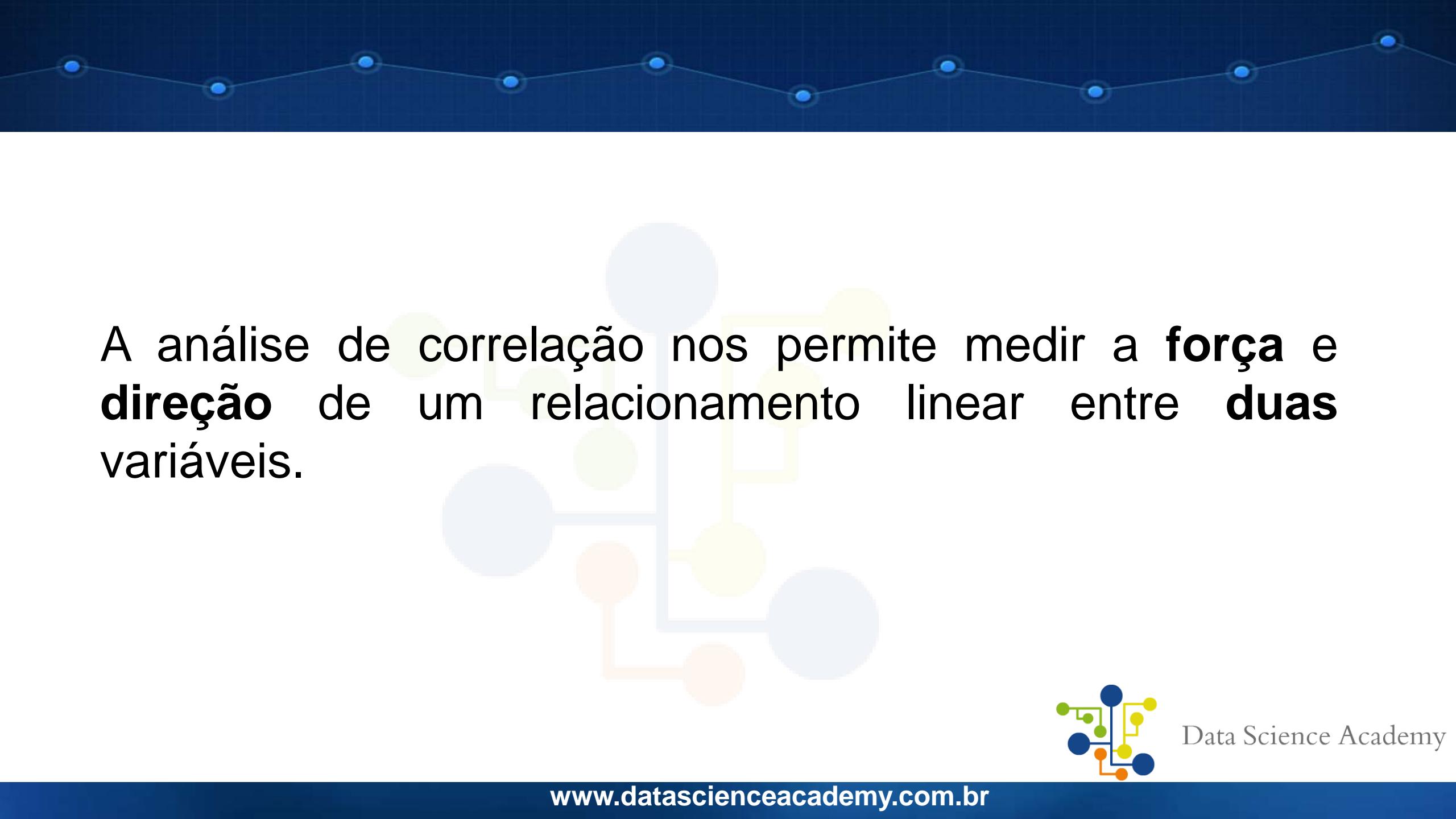
Data Science Academy



Correlação



Data Science Academy



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



Data Science Academy



Neste módulo iremos fazer **duas coisas**:

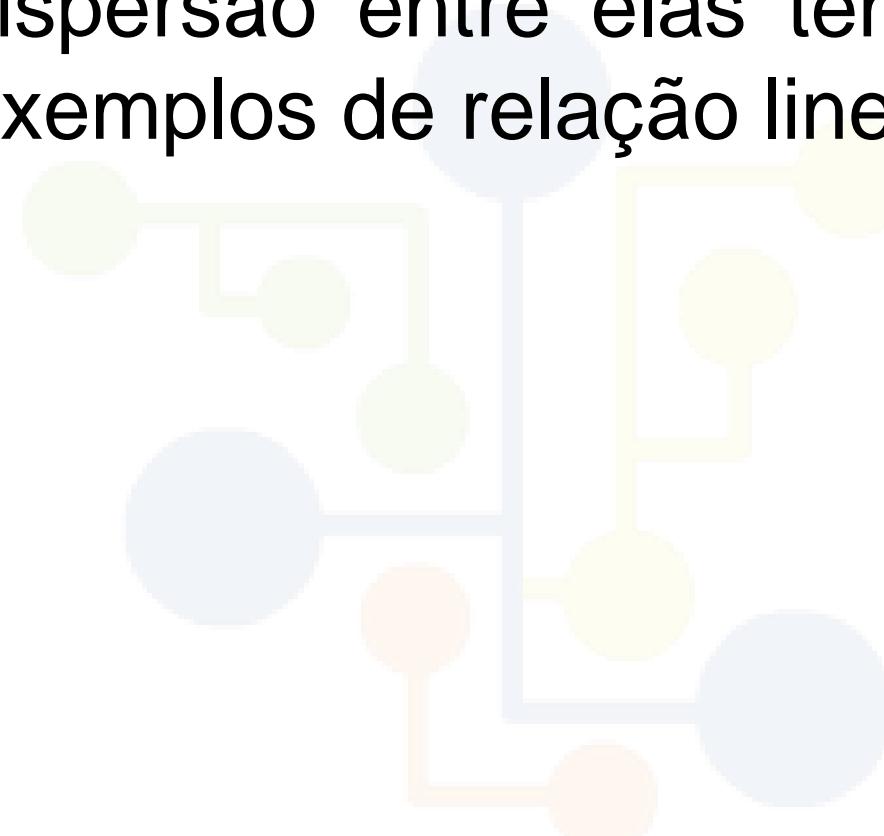
Primeiro vou mostrar como calcular o coeficiente de correlação r , que provê um valor que descreve o relacionamento.

Em seguida, vamos aprender como realizar um teste de hipótese para decidir se o relacionamento entre as duas variáveis é ou não forte suficiente para ser considerado estatisticamente significante.



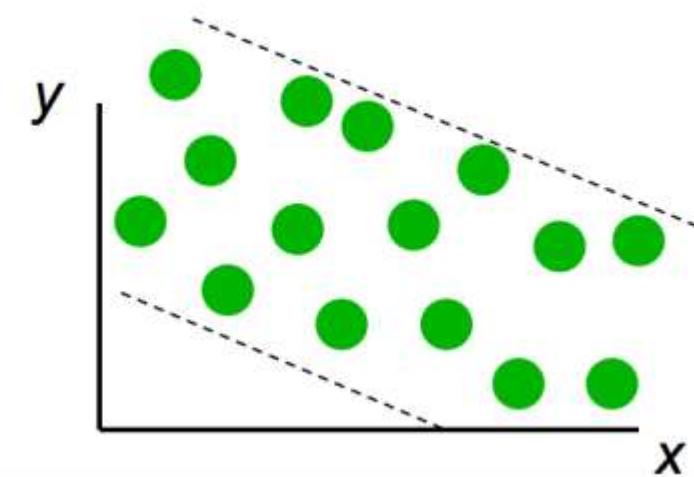
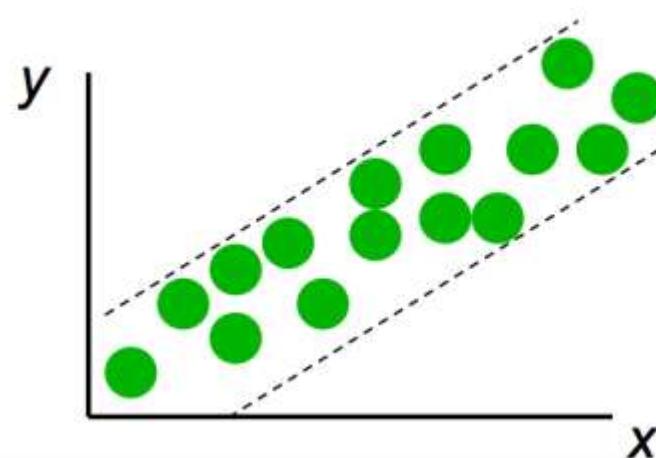
Data Science Academy

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:

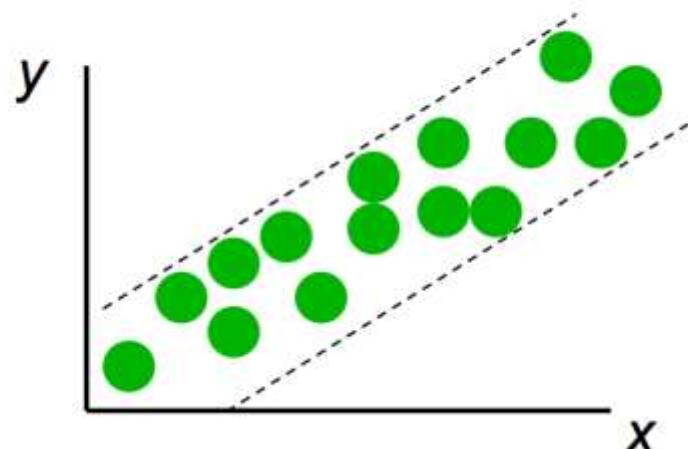


Data Science Academy

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:

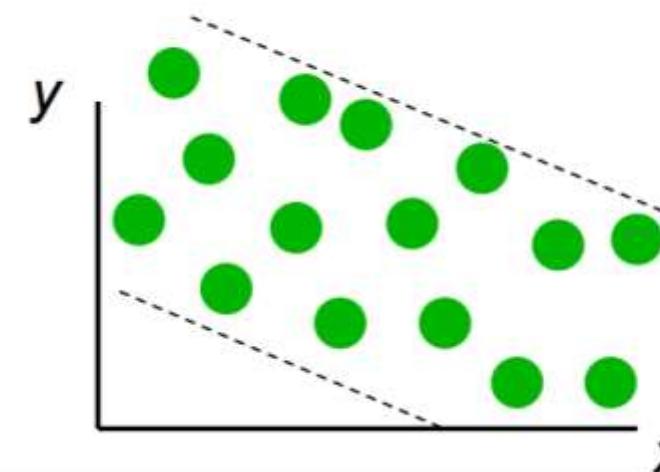


Data Science Academy



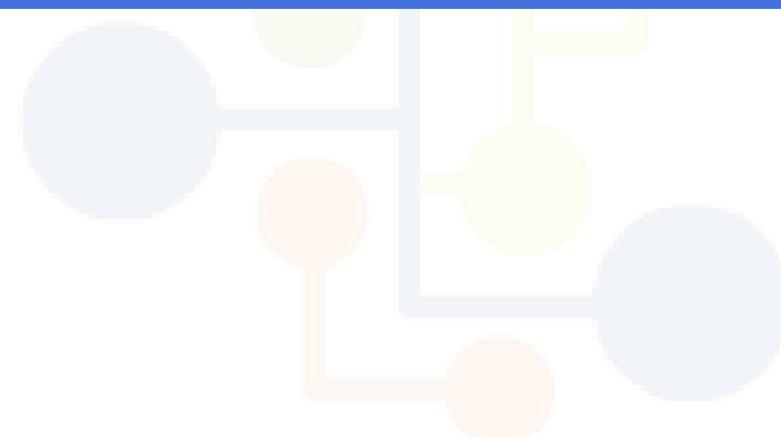
Relacionamento **positivo**,
inclinação se move para cima.

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



Data Science Academy

Exemplo



Data Science Academy

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



Data Science Academy

Espera-se que o número de comerciais de TV por semana (x) afete a venda de carros por semana (y). Perceba que esta relação possui uma única direção. Suponha uma amostra de 6 semanas, com os dados coletados na tabela abaixo:



Data Science Academy

Espera-se que o número de comerciais de TV por semana (x) afete a venda de carros por semana (y). Perceba que esta relação possui uma única direção. Suponha uma amostra de 6 semanas, com os dados coletados na tabela abaixo:

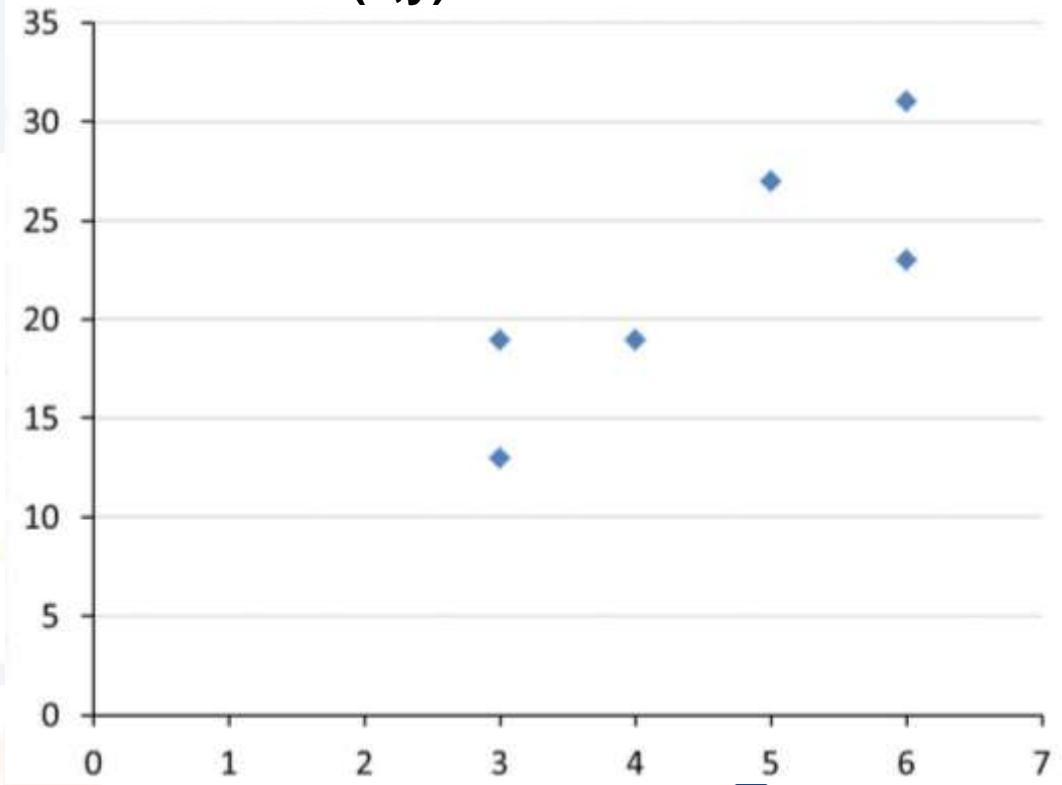
Semana	Número de comerciais x	Número de carros vendidos y
1	3	13
2	6	31
3	4	19
4	5	27
5	6	23
6	3	19



Semana	<i>Número de comerciais</i> <i>x</i>	<i>Número de carros vendidos</i> <i>y</i>
1	3	13
2	6	31
3	4	19
4	5	27
5	6	23
6	3	19

Número de Carros Vendidos por Semana

Gráfico de Dispersão (x,y)



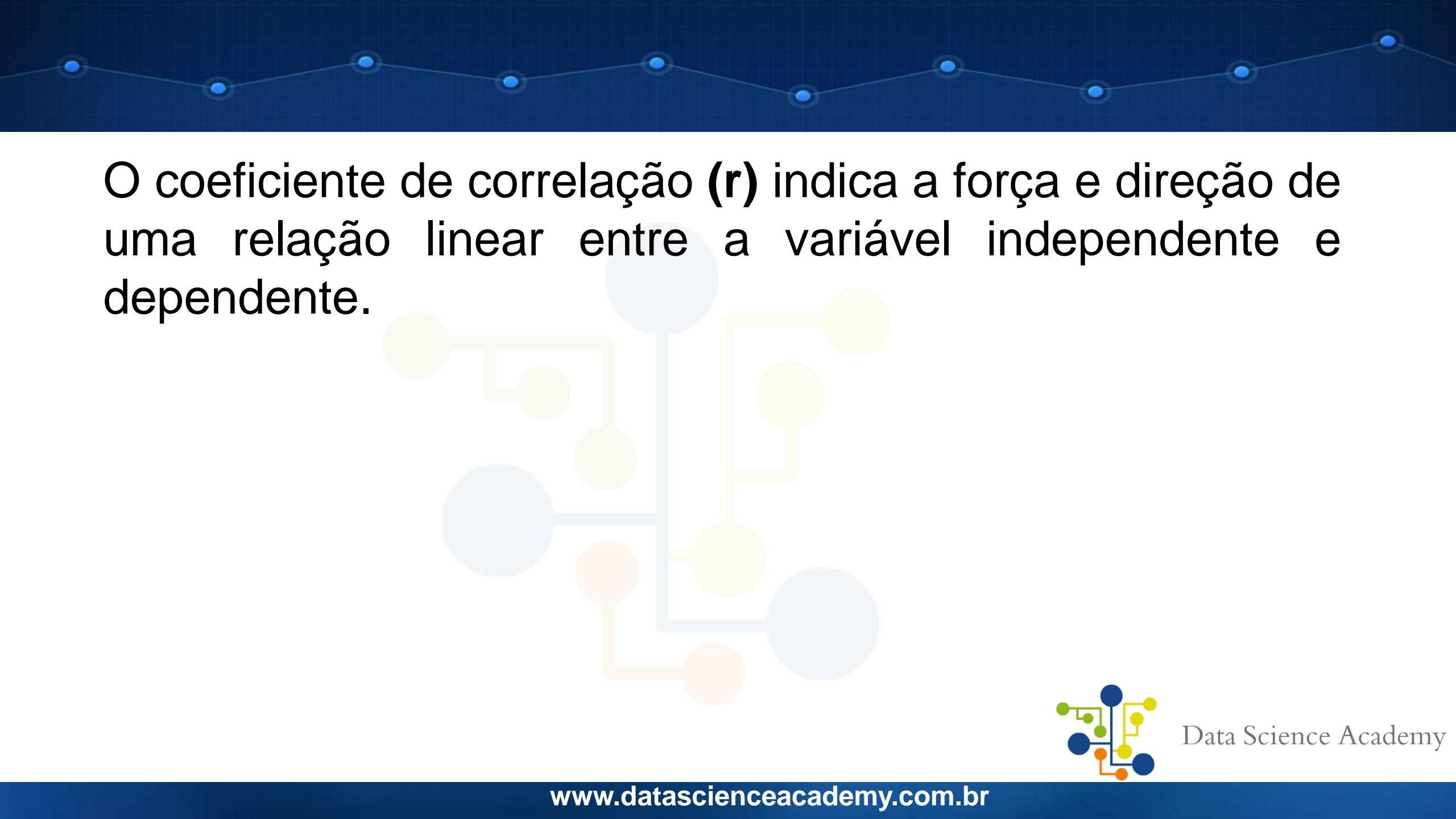
Número de Comerciais por Semana

Data Science Academy

Coeficiente de Correlação



Data Science Academy

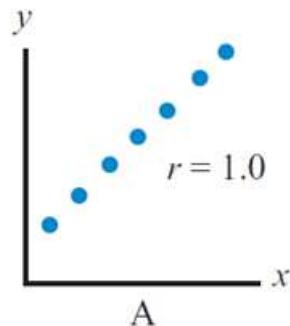


O coeficiente de correlação (r) indica a força e direção de uma relação linear entre a variável independente e dependente.

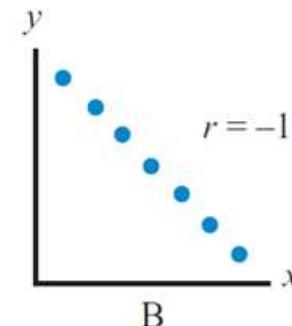


Data Science Academy

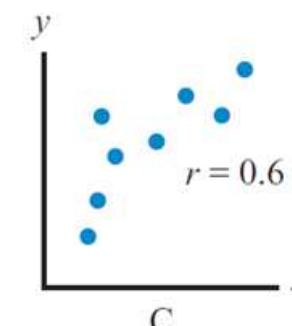
Exemplos de valores de r :



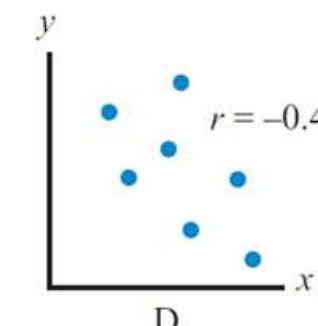
A



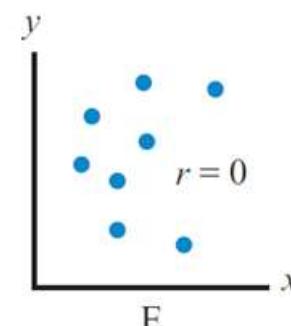
B



C



D



E

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

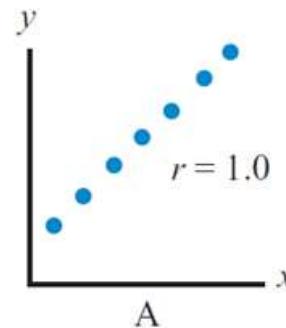
Gráfico D ($r = -0.4$): relação negativa **fraca**: o coeficiente de correlação é próximo de zero ou negativo: y tende a diminuir se x aumenta

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

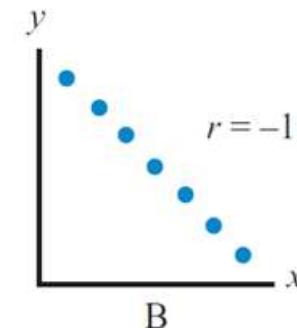


Data Science Academy

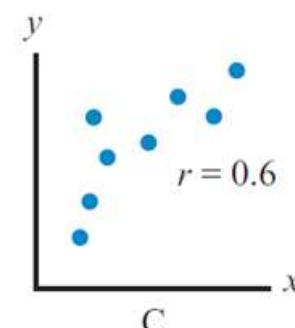
Exemplos de valores de r:



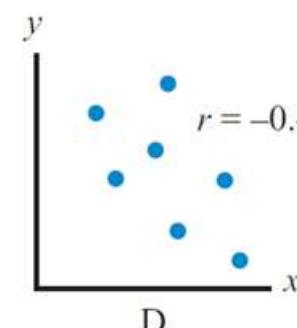
A



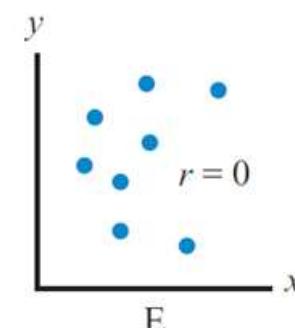
B



C



D



E

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

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

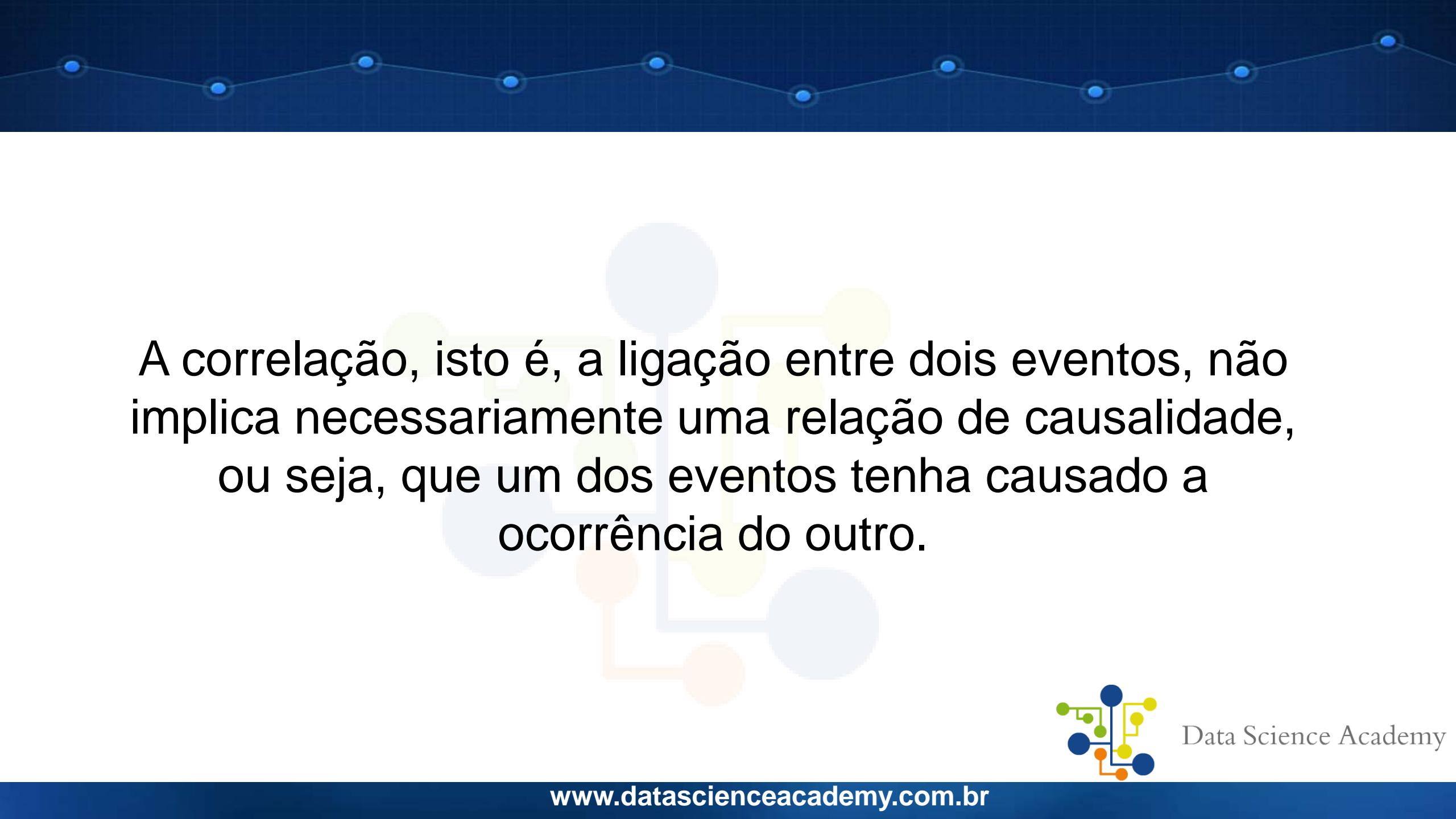
Gráfico C ($r = 0.6$): relação positiva moderada: y tende a aumentar se x aumenta, mas não necessariamente na mesma taxa observada no Gráfico A

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.

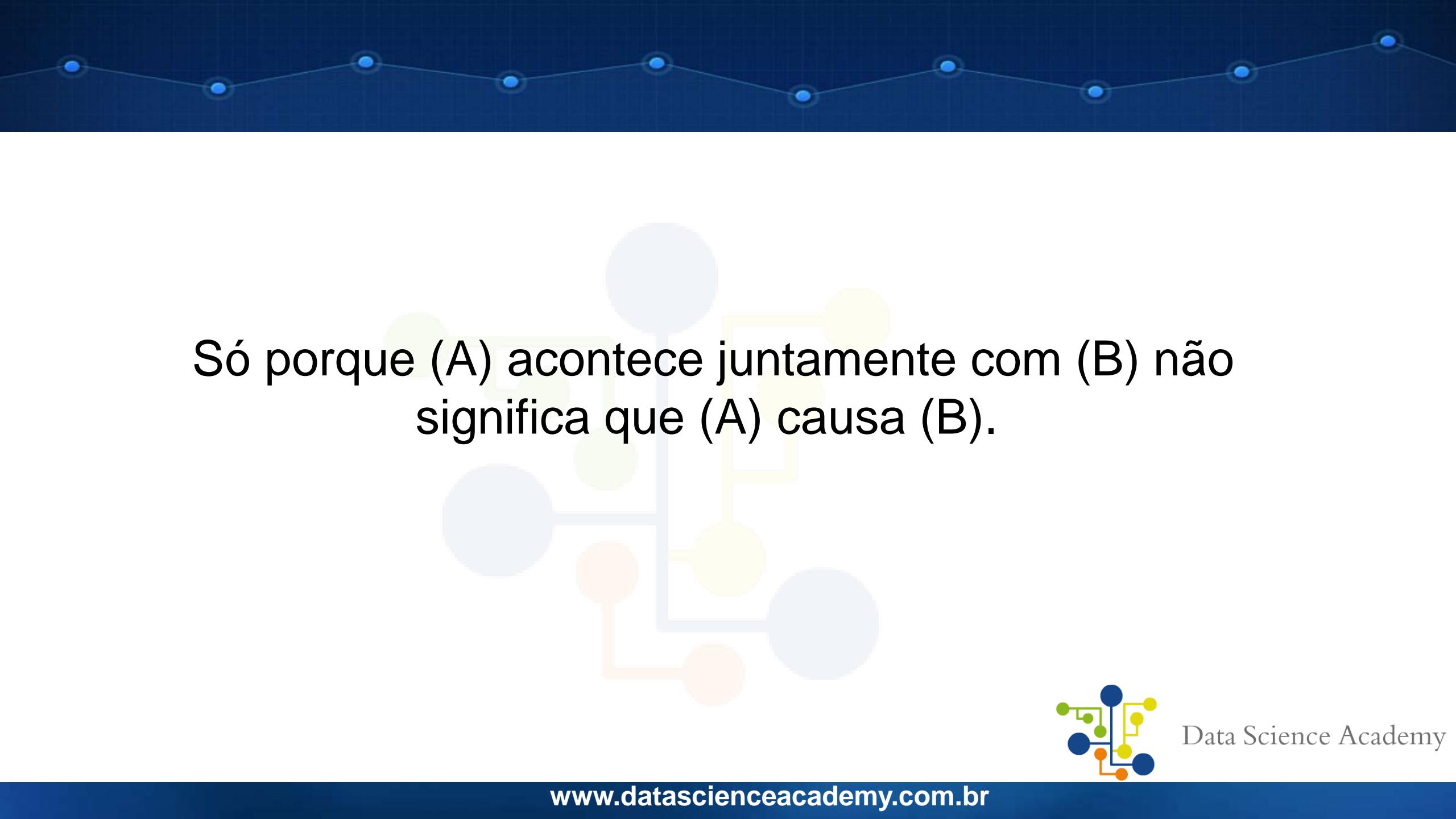




A correlação, isto é, a ligação entre dois eventos, não implica necessariamente uma relação de causalidade, ou seja, que um dos eventos tenha causado a ocorrência do outro.



Data Science Academy

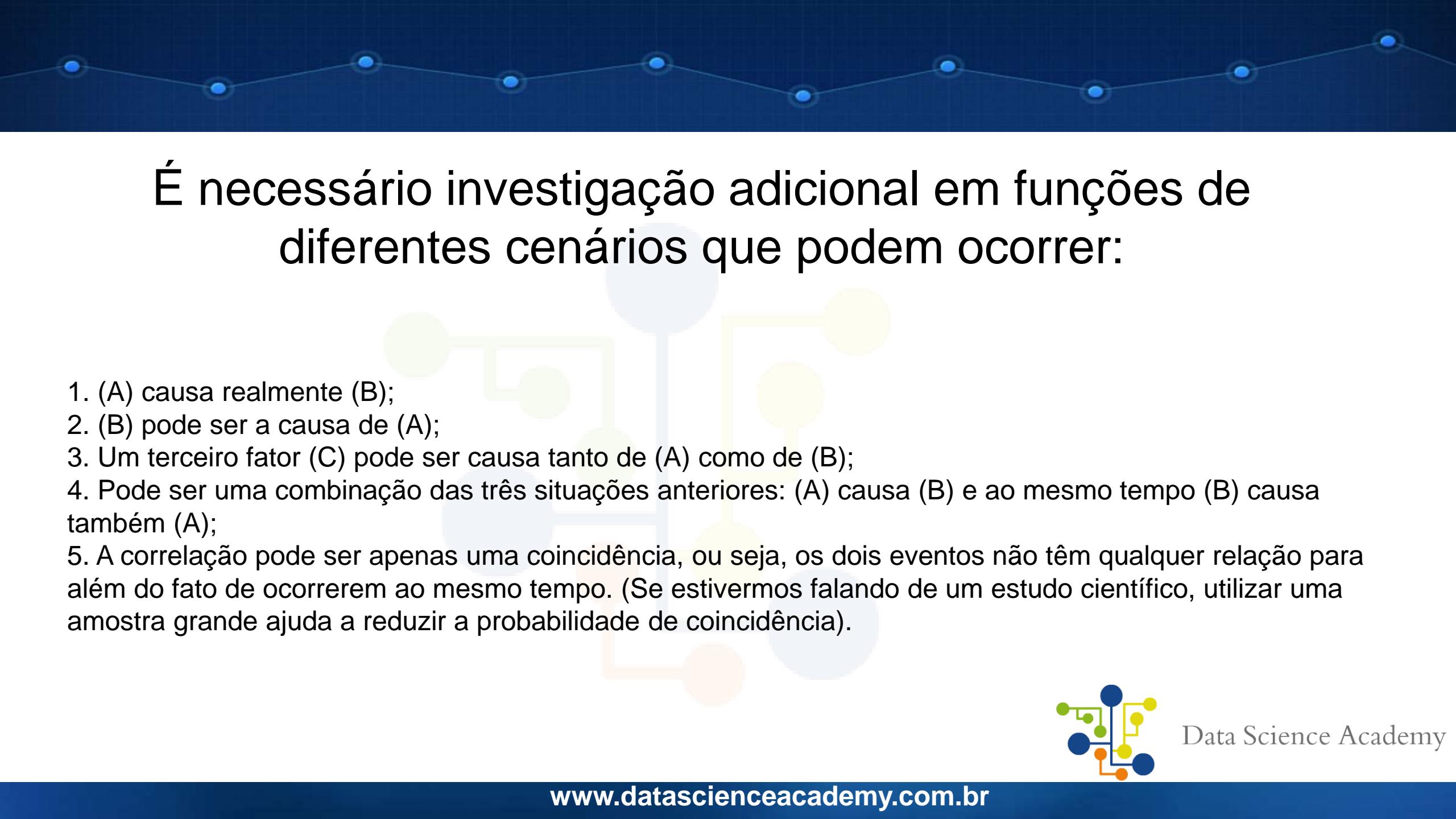


Só porque (A) acontece juntamente com (B) não significa que (A) causa (B).



Data Science Academy

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

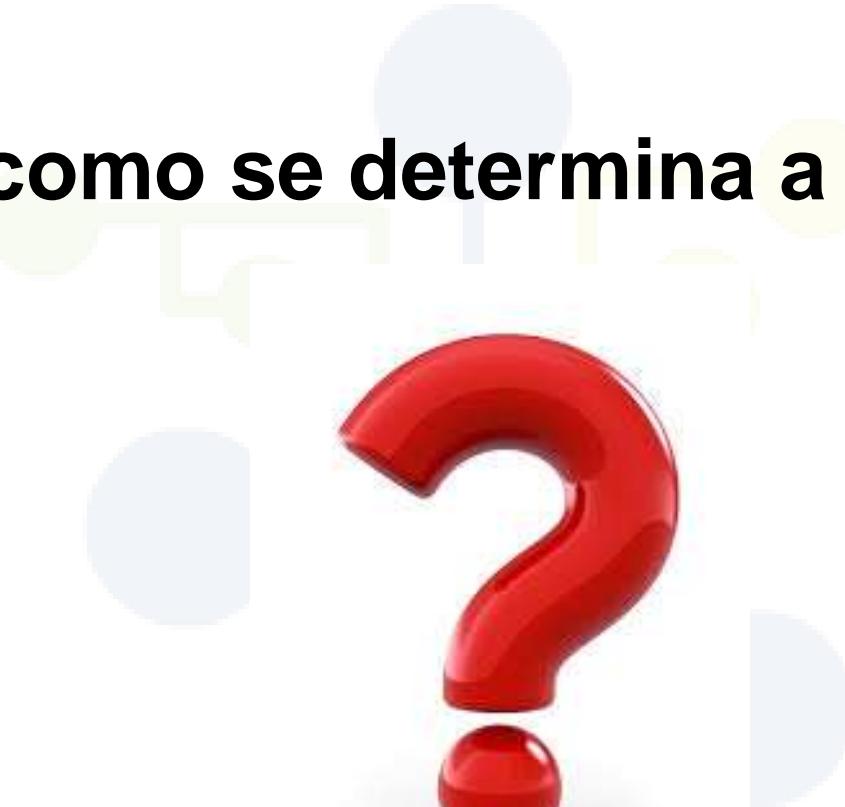
- 
1. (A) causa realmente (B);
 2. (B) pode ser a causa de (A);
 3. Um terceiro fator (C) pode ser causa tanto de (A) como de (B);
 4. Pode ser uma combinação das três situações anteriores: (A) causa (B) e ao mesmo tempo (B) causa também (A);
 5. A correlação pode ser apenas uma coincidência, ou seja, 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).



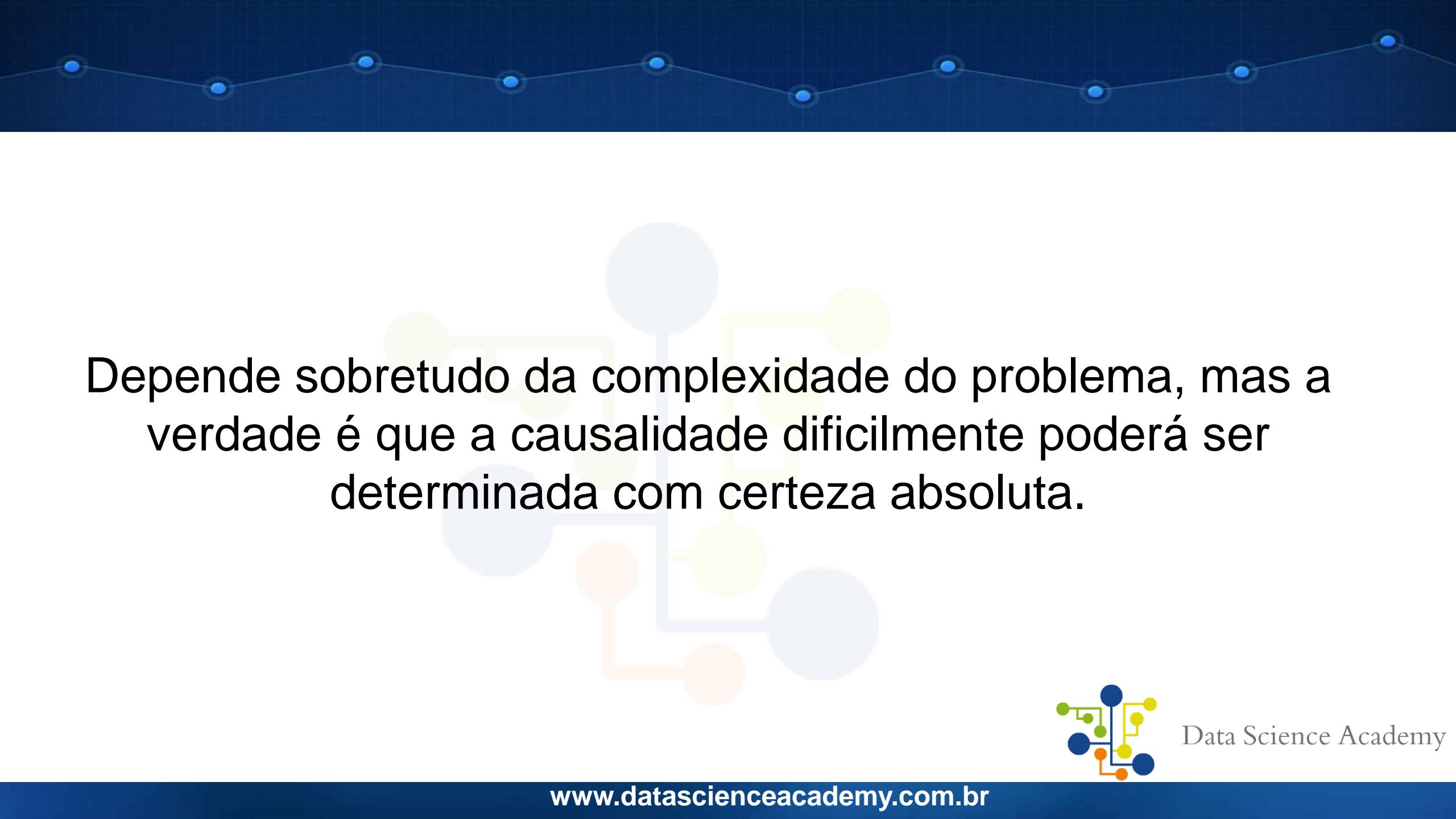
Data Science Academy



Então como se determina a causalidade?



Data Science Academy



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



Data Science Academy

Esse tópico chegou ao final.



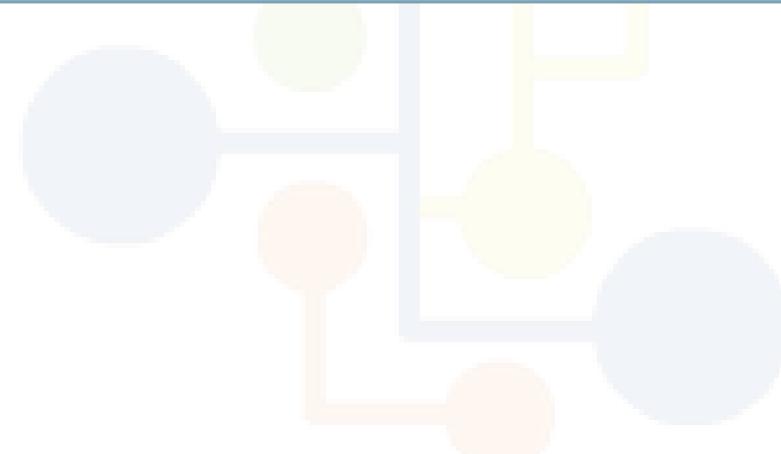
Obrigada



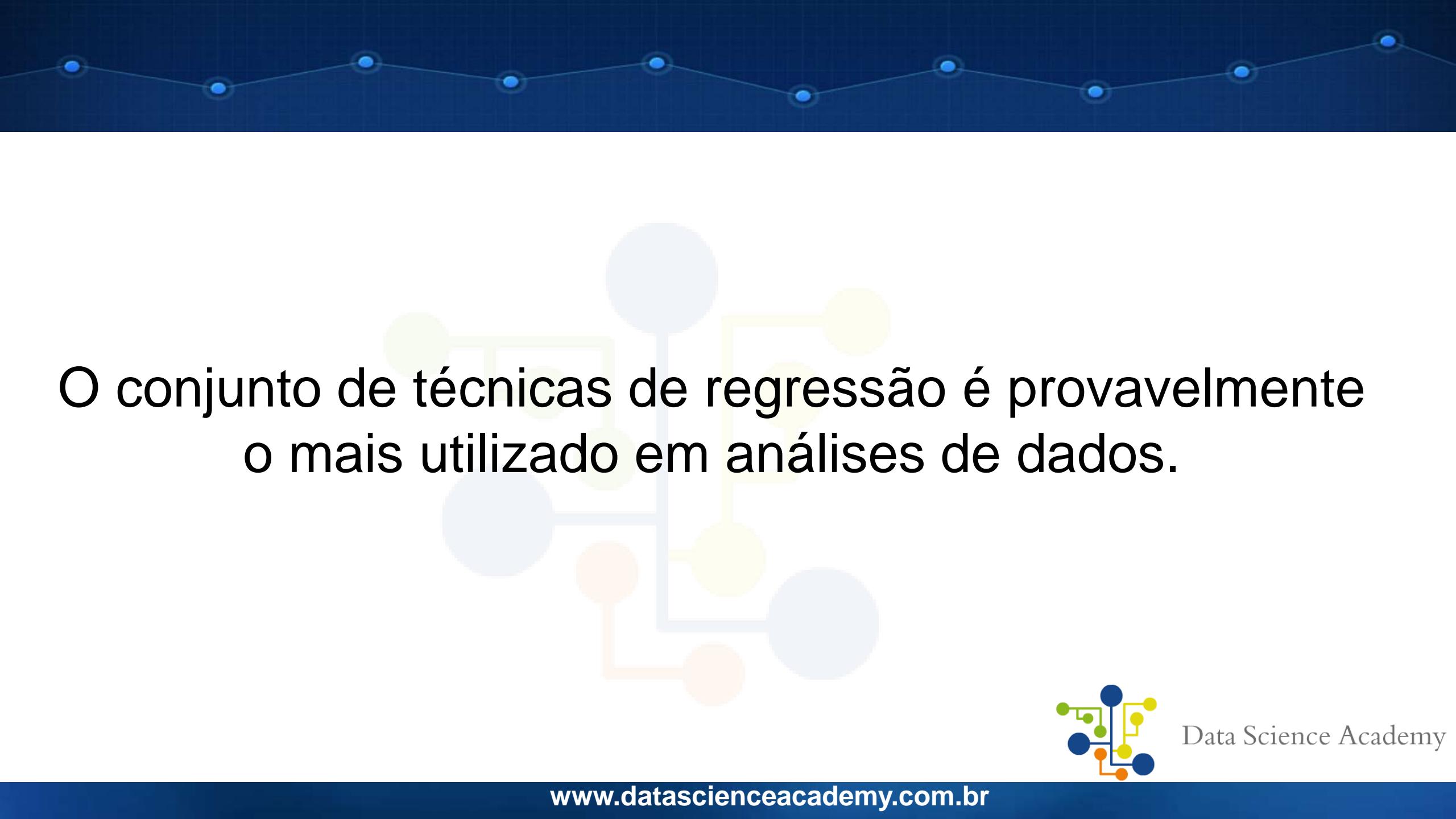
Data Science Academy



Análise de Regressão



Data Science Academy



O conjunto de técnicas de regressão é provavelmente o mais utilizado em análises de dados.



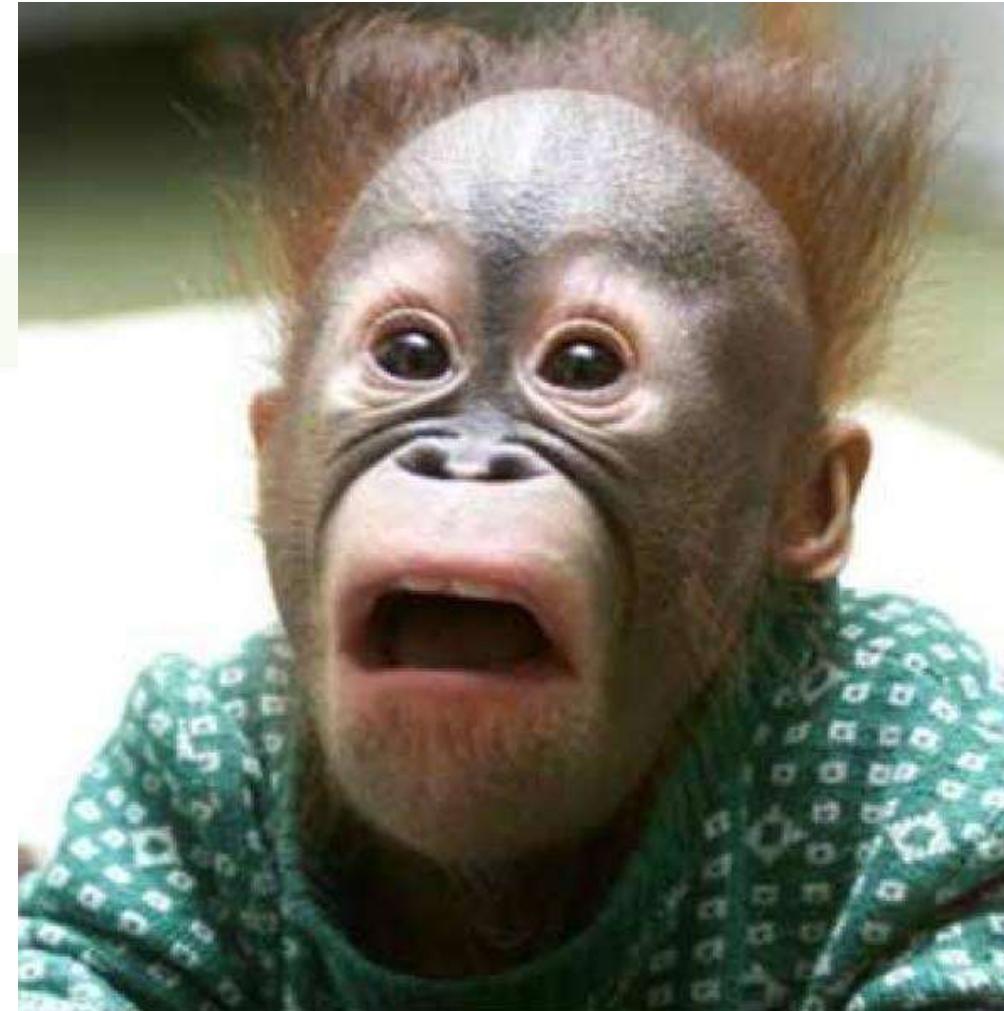
Data Science Academy

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

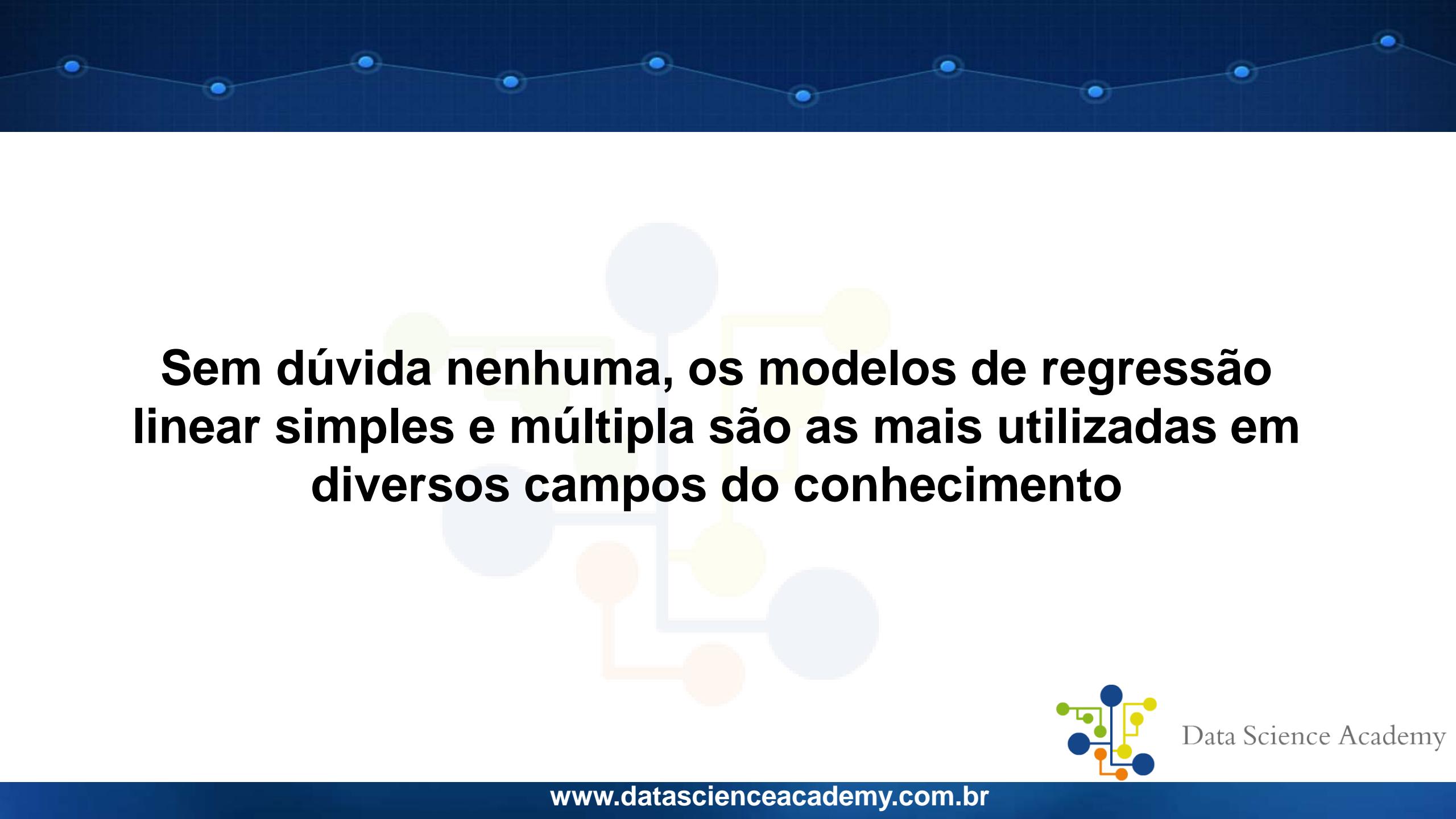


Data Science Academy



Data Science Academy

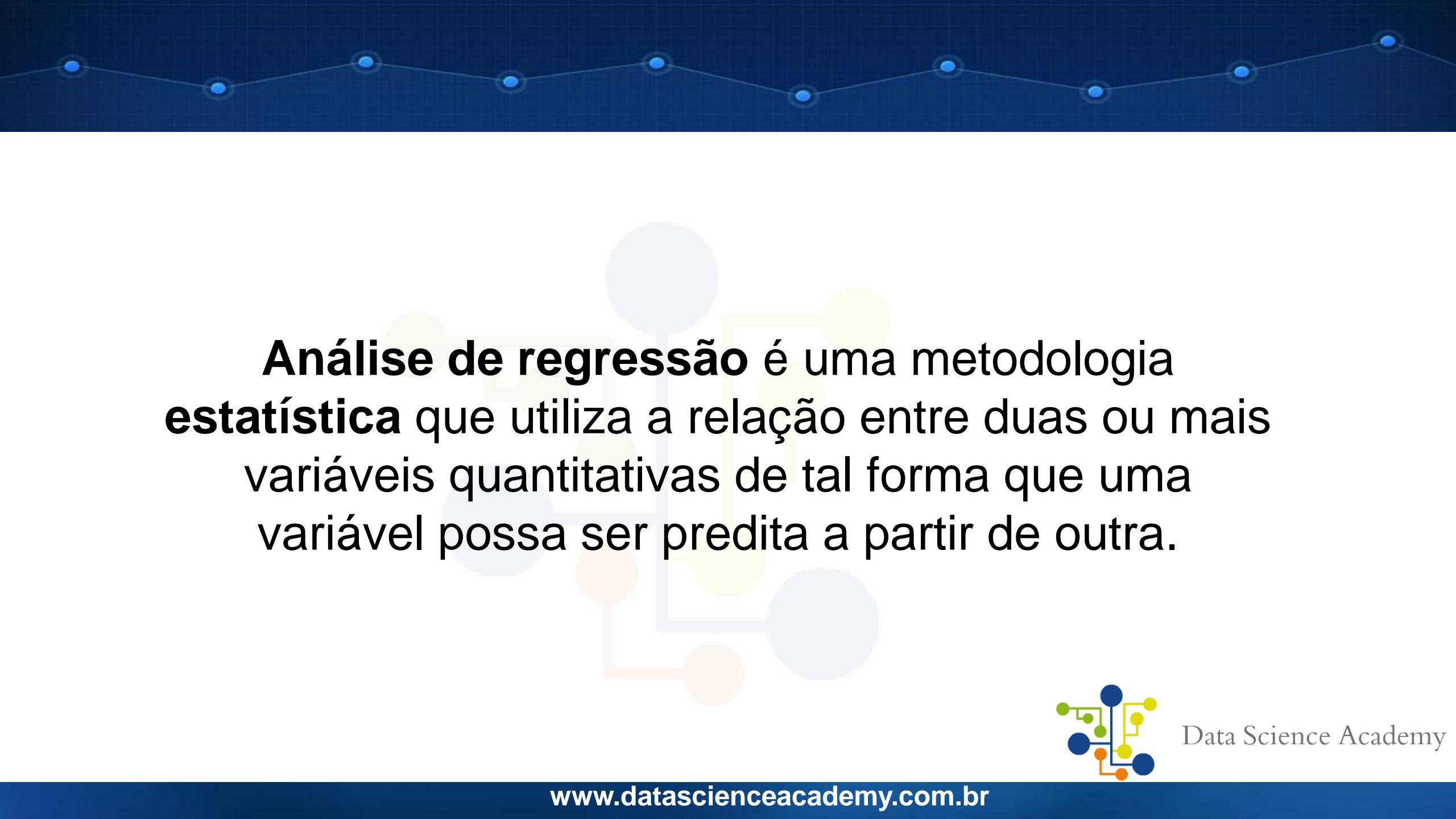
www.datascienceacademy.com.br



Sem dúvida nenhuma, os modelos de regressão linear simples e múltipla são as mais utilizadas em diversos campos do conhecimento



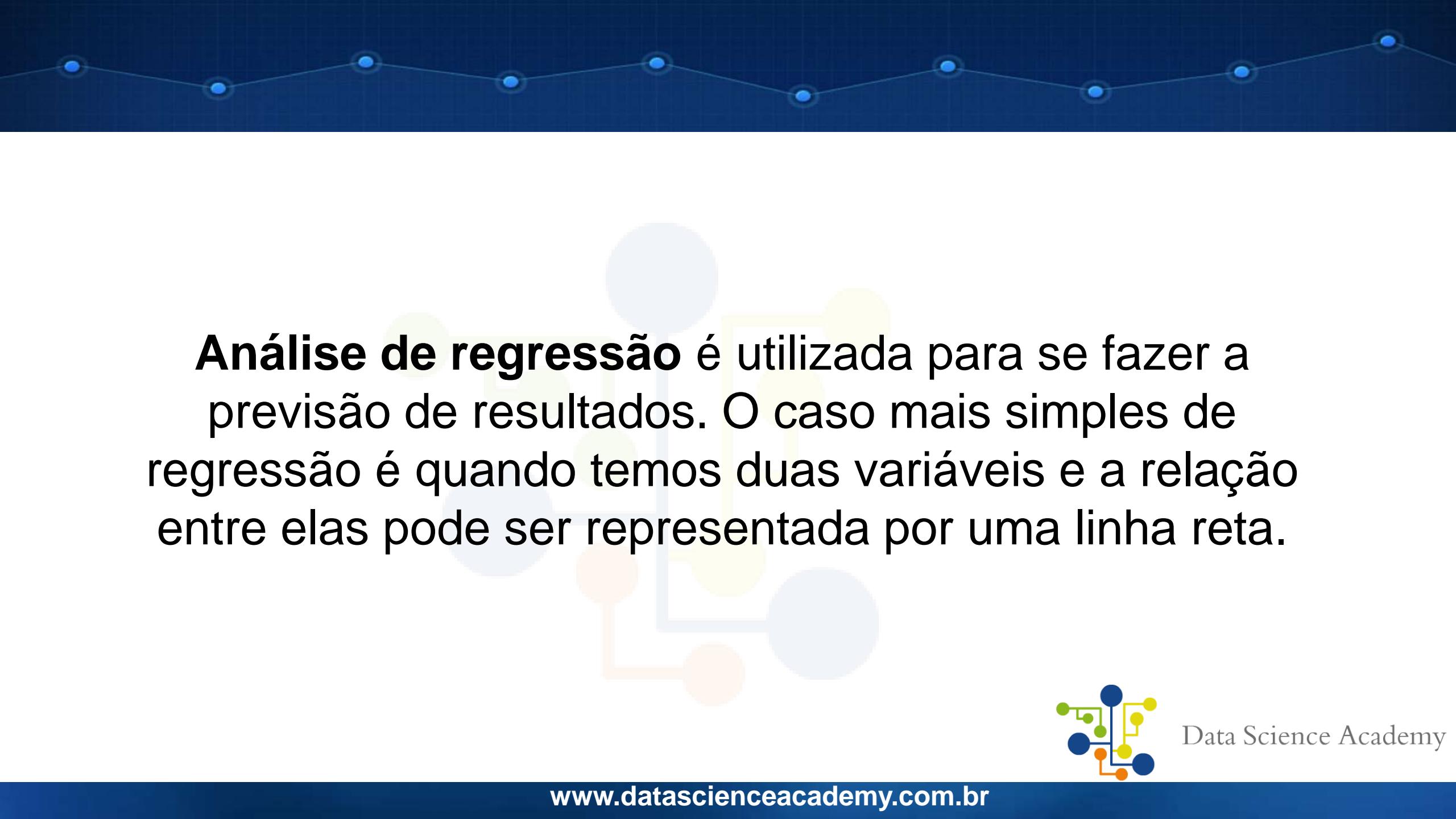
Data Science Academy



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.



Data Science Academy



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.



Data Science Academy

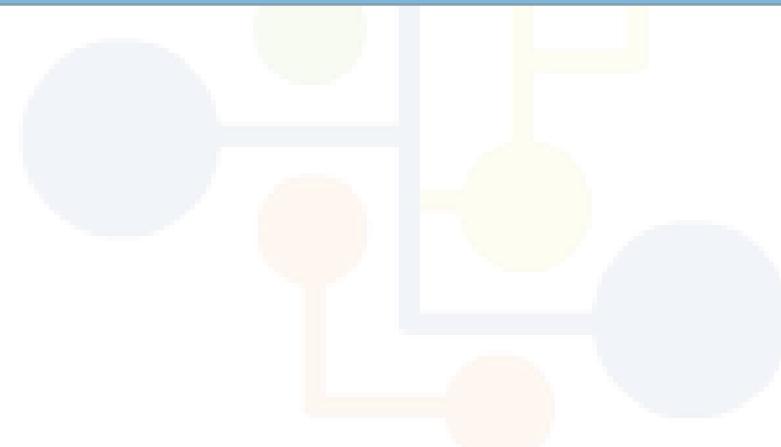


Origem do Modelo Clássico de Análise de Regressão

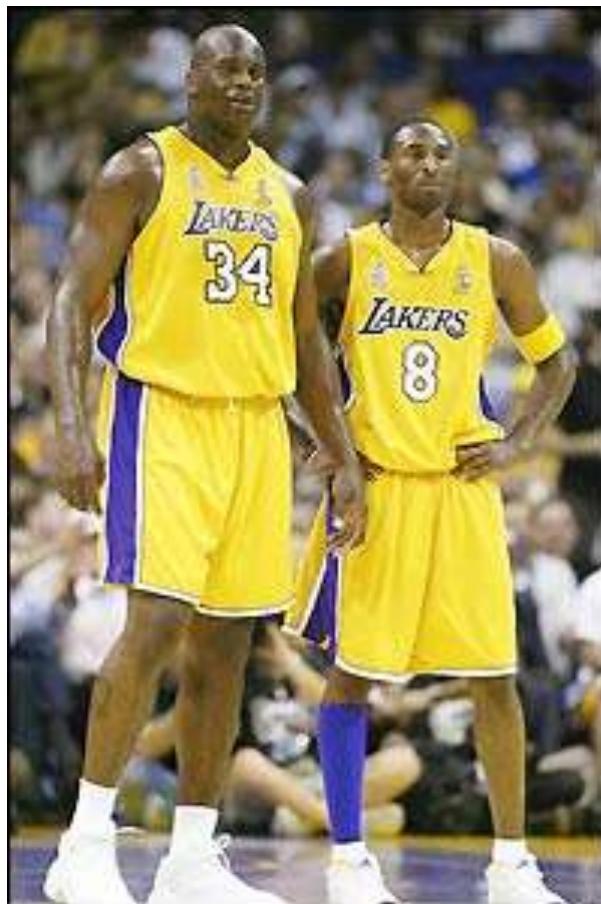
Francis Galton



Exemplo



Data Science Academy

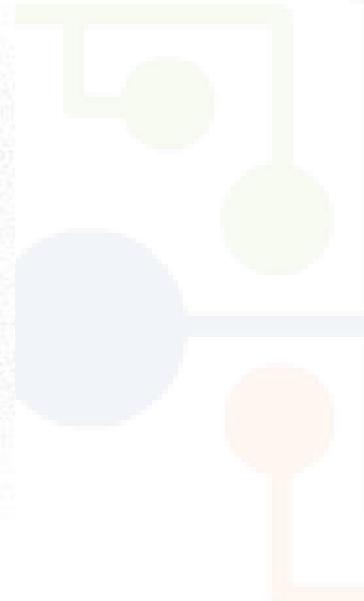


**Shaquille O'Neal
2,16 metros**



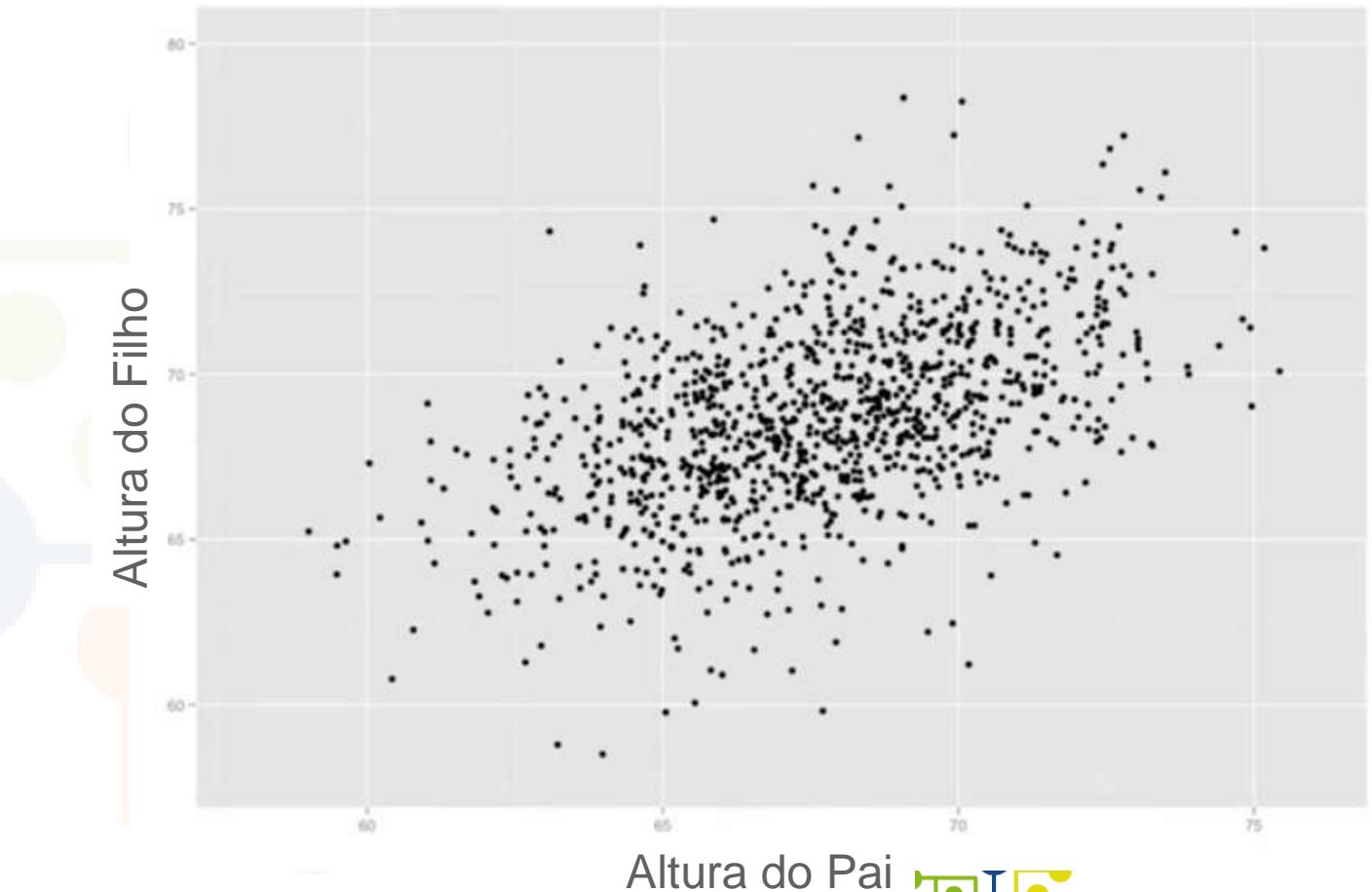
Data Science Academy

Fenômeno de regressão



Data Science Academy

Régressão Linear



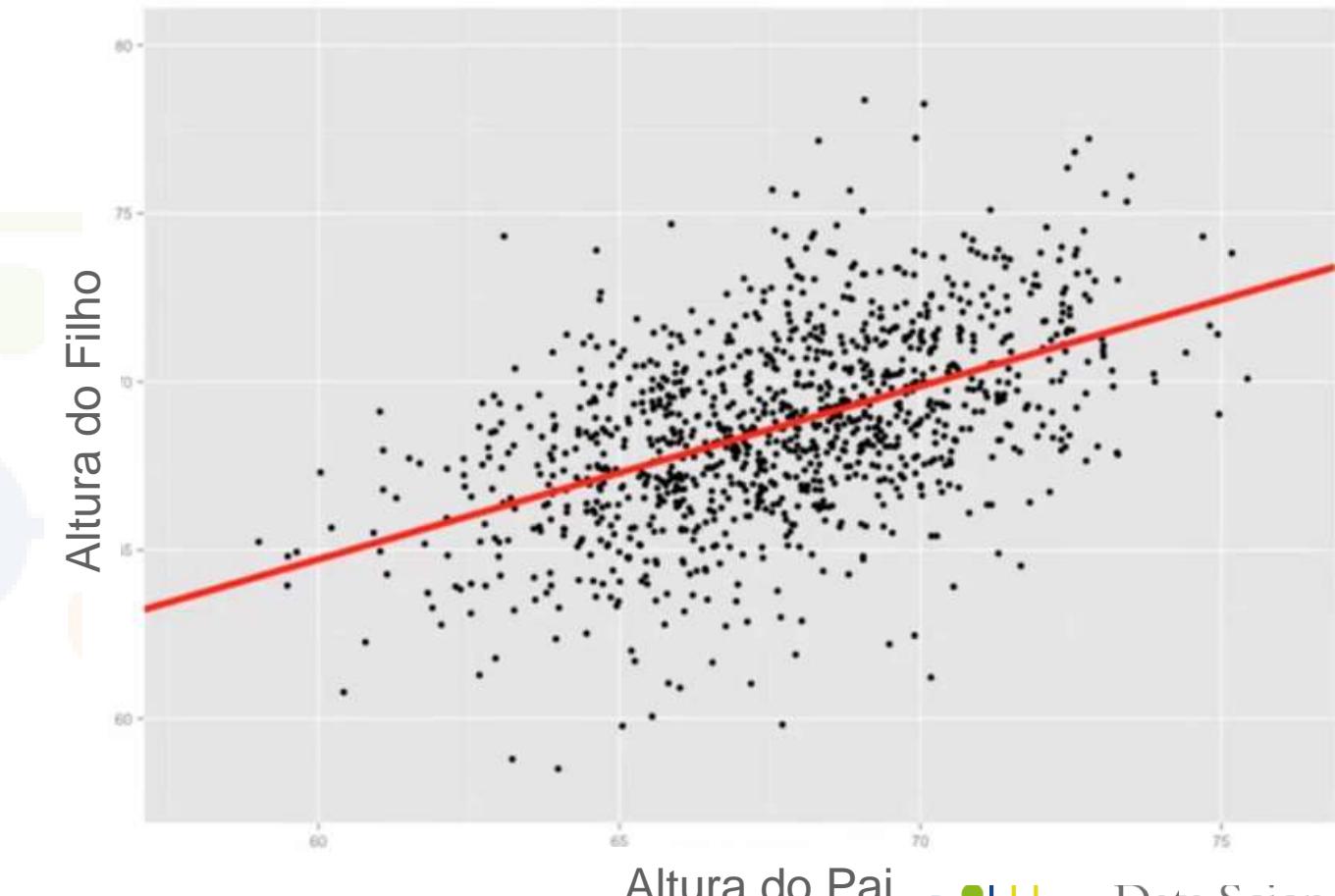
Altura do Pai

Altura do Filho



Data Science Academy

Régressão Linear

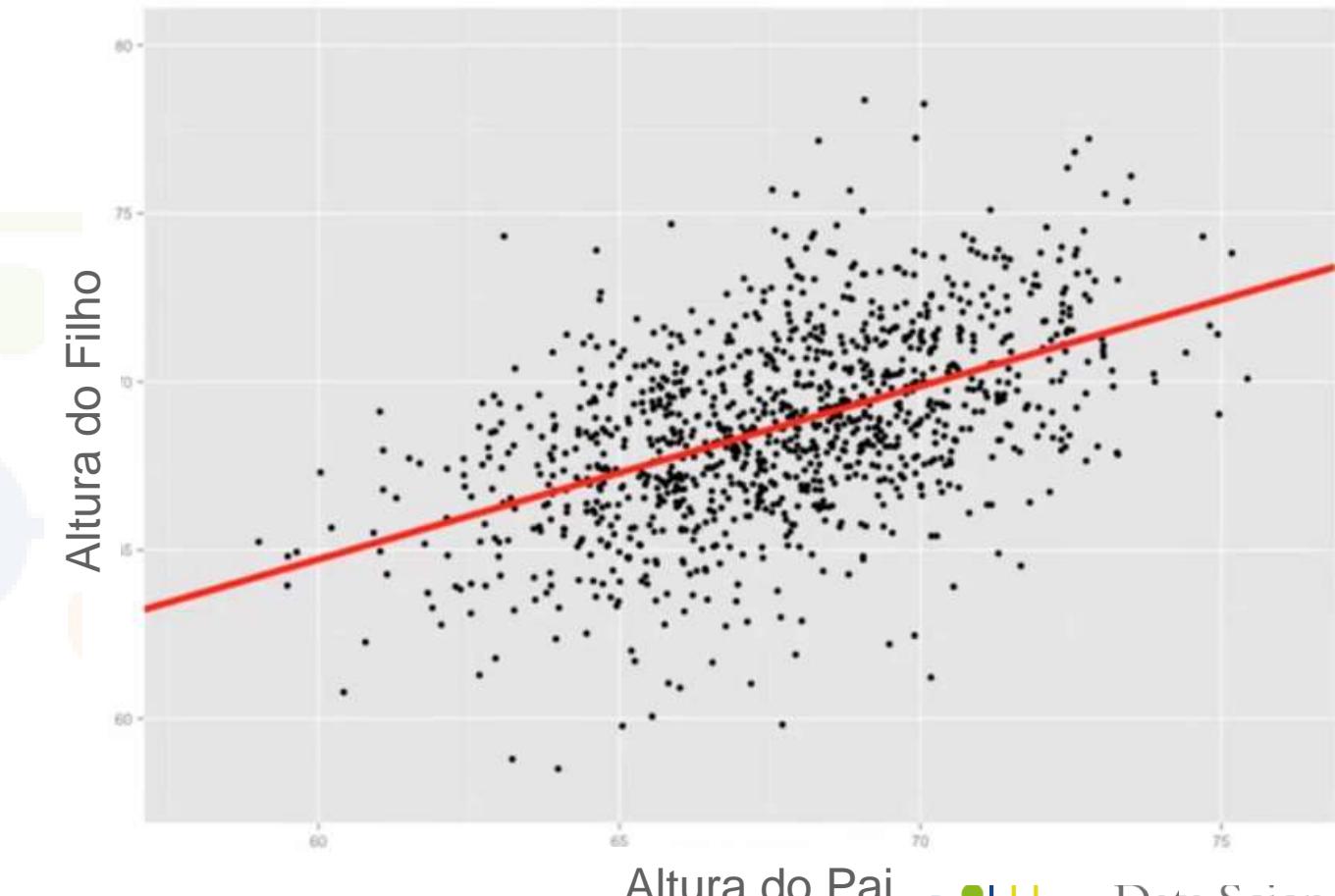


Altura do Pai



Data Science Academy

Régressão Linear



Altura do Pai



Data Science Academy



REGRESSÃO e CORRELAÇÃO

são a mesma coisa?

NÃO



Data Science Academy



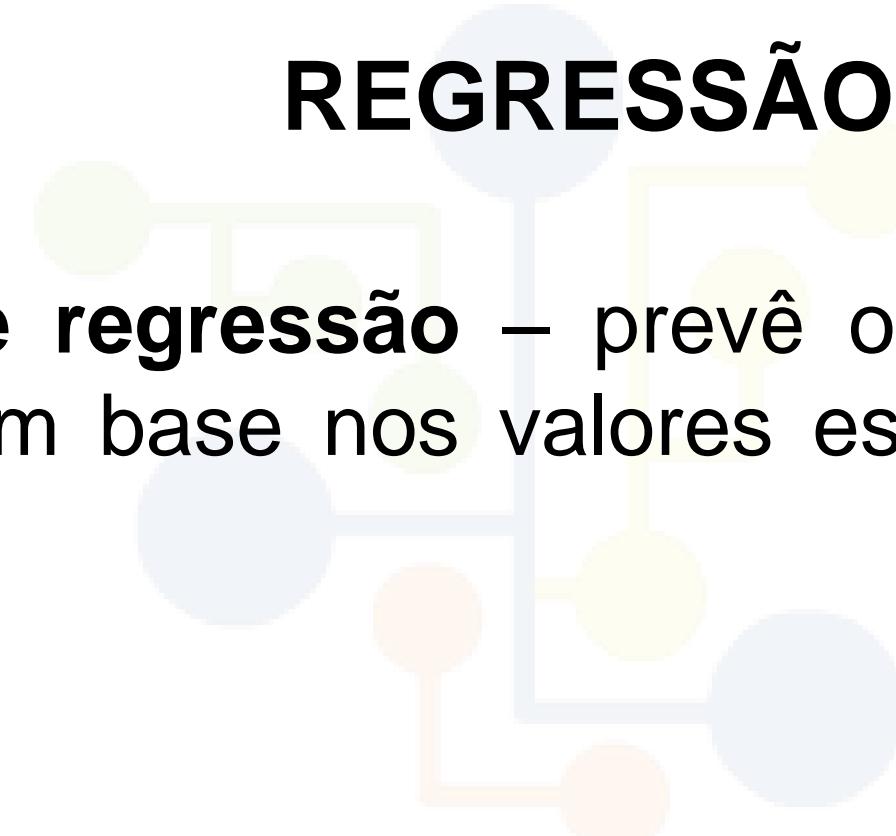
Não



Data Science Academy



REGRESSÃO



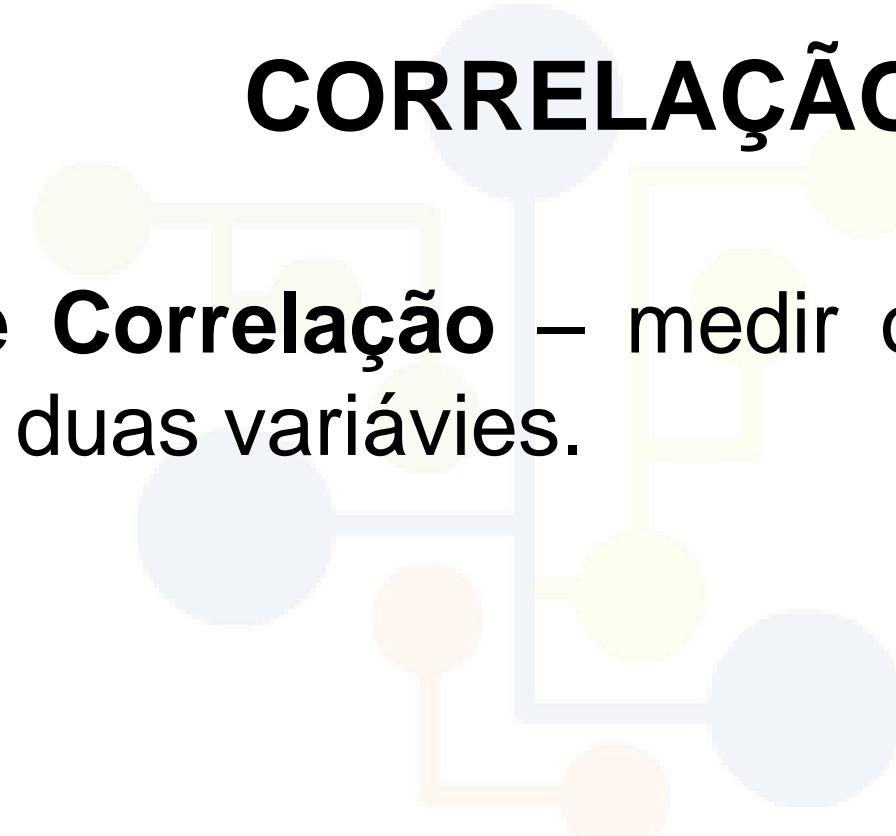
Análise de regressão – prevê o valor médio de uma variável com base nos valores estabelecidos de outras variáveis.



Data Science Academy



CORRELAÇÃO



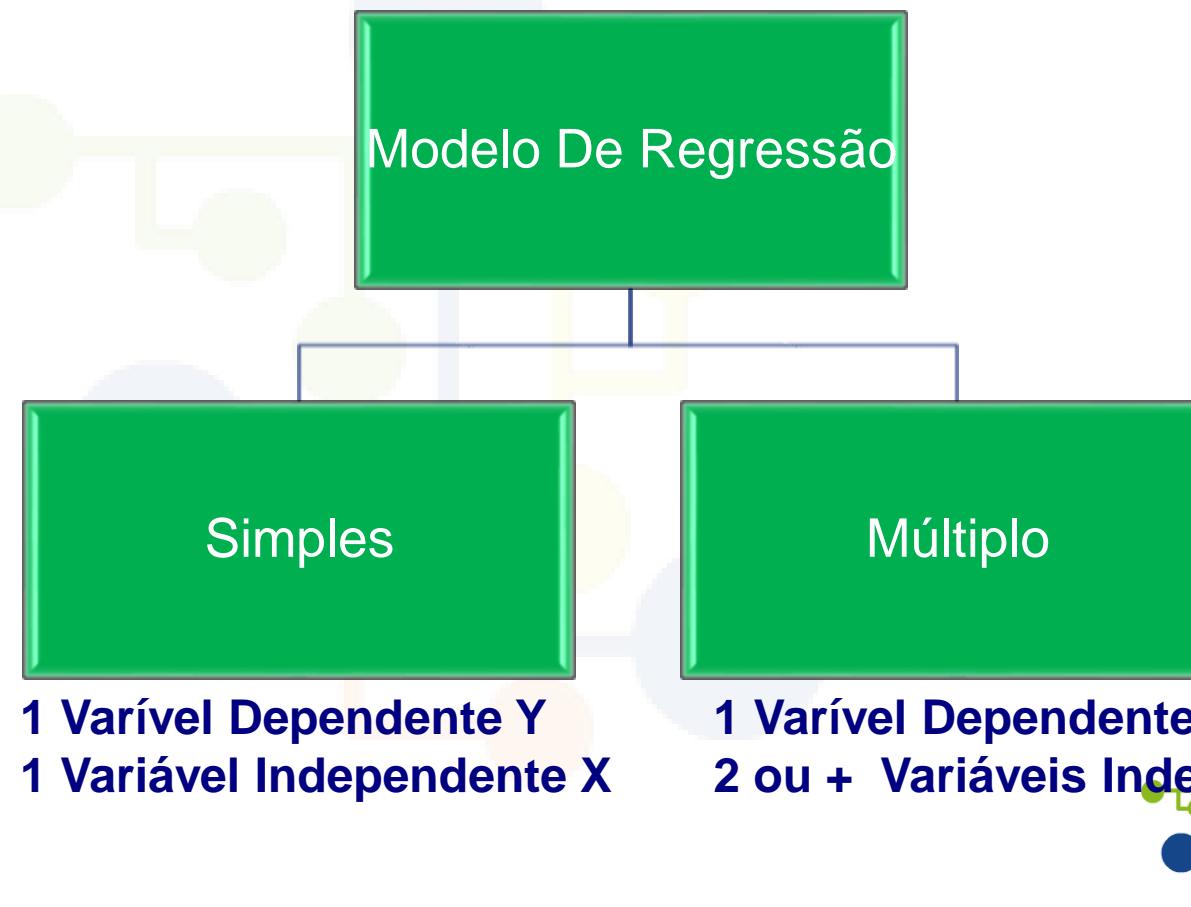
de intensidade

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



Data Science Academy

Tipos de Modelos de Regressão Linear



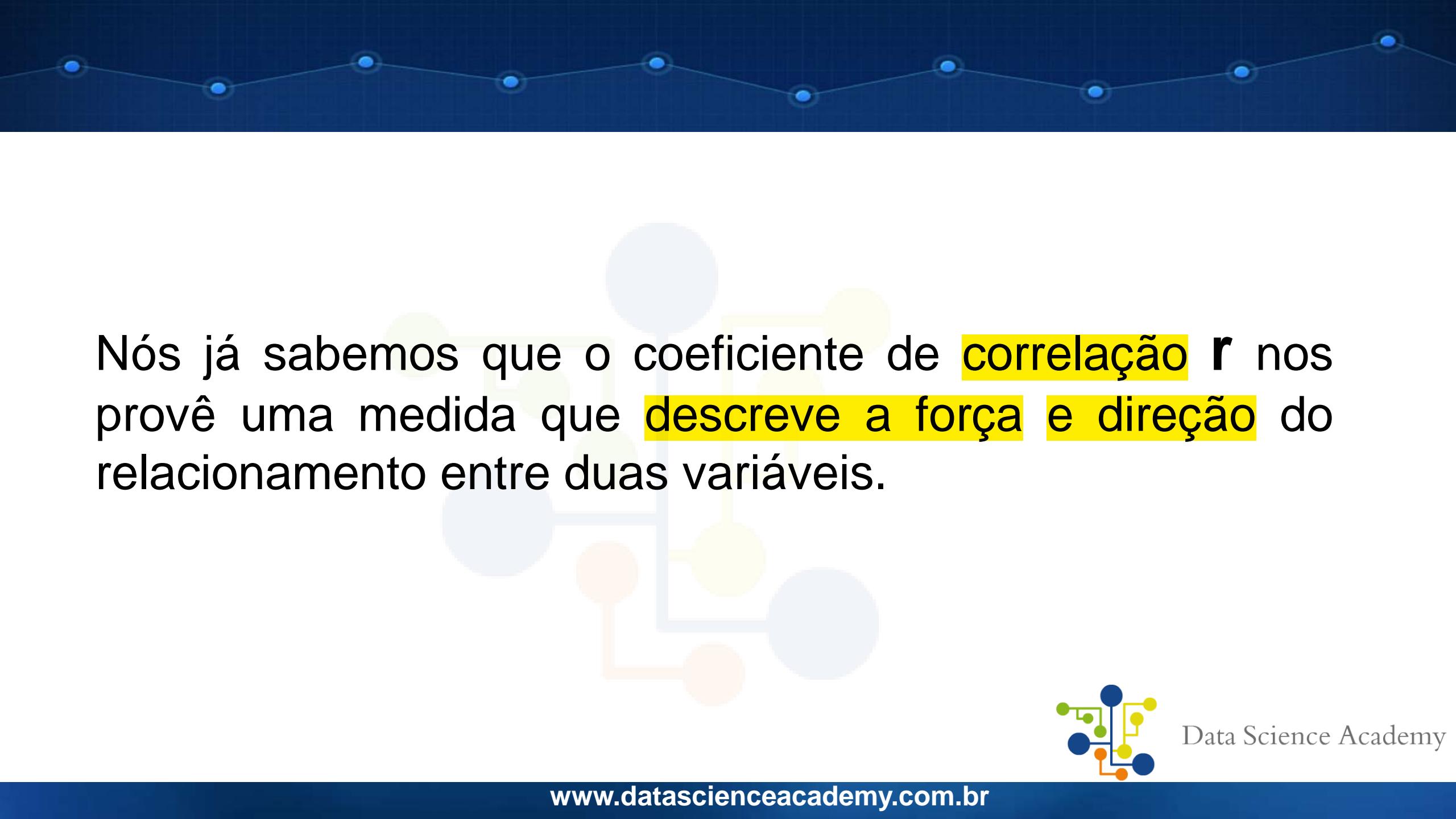
Data Science Academy



Regressão Linear Simples



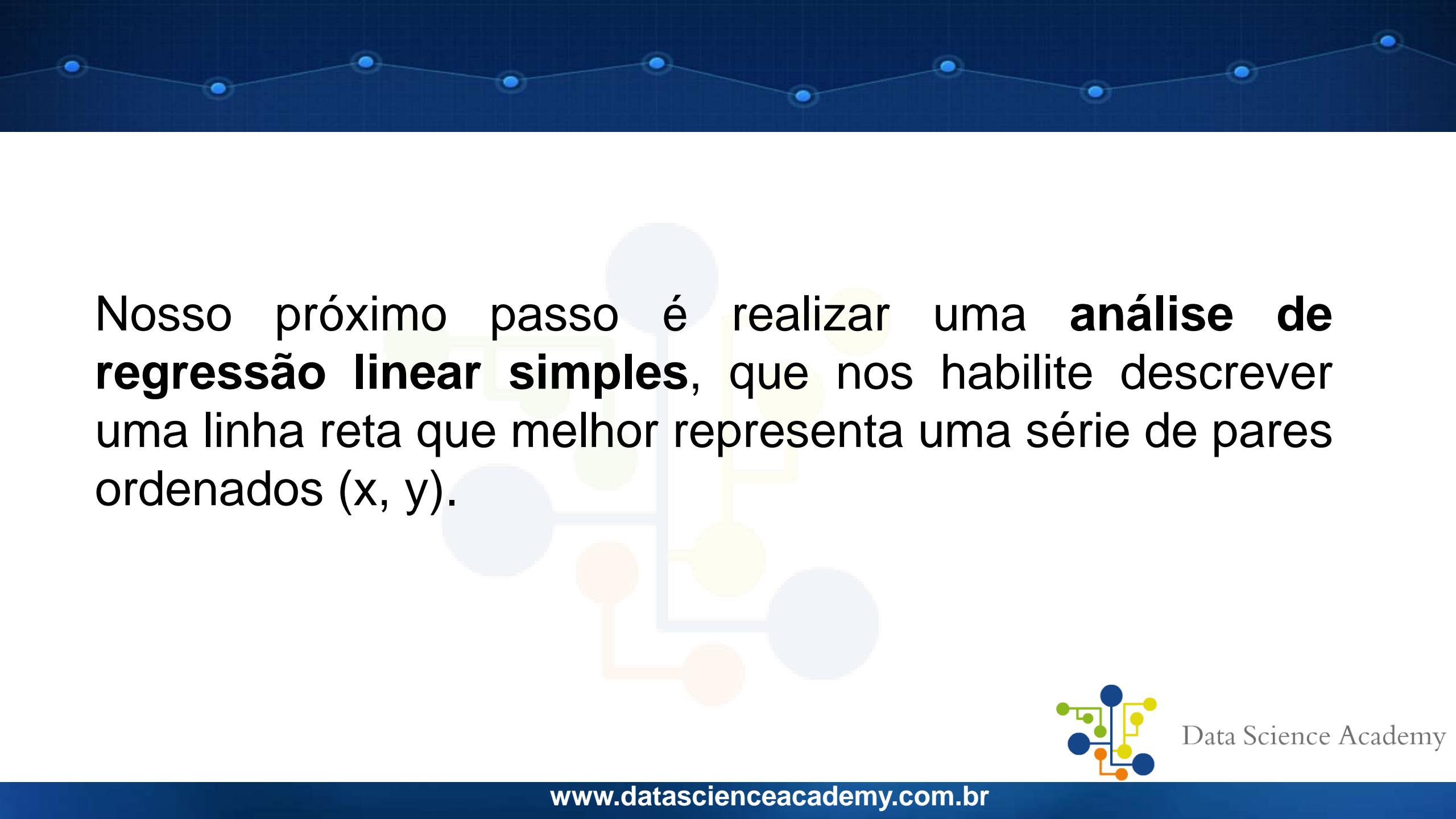
Data Science Academy



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



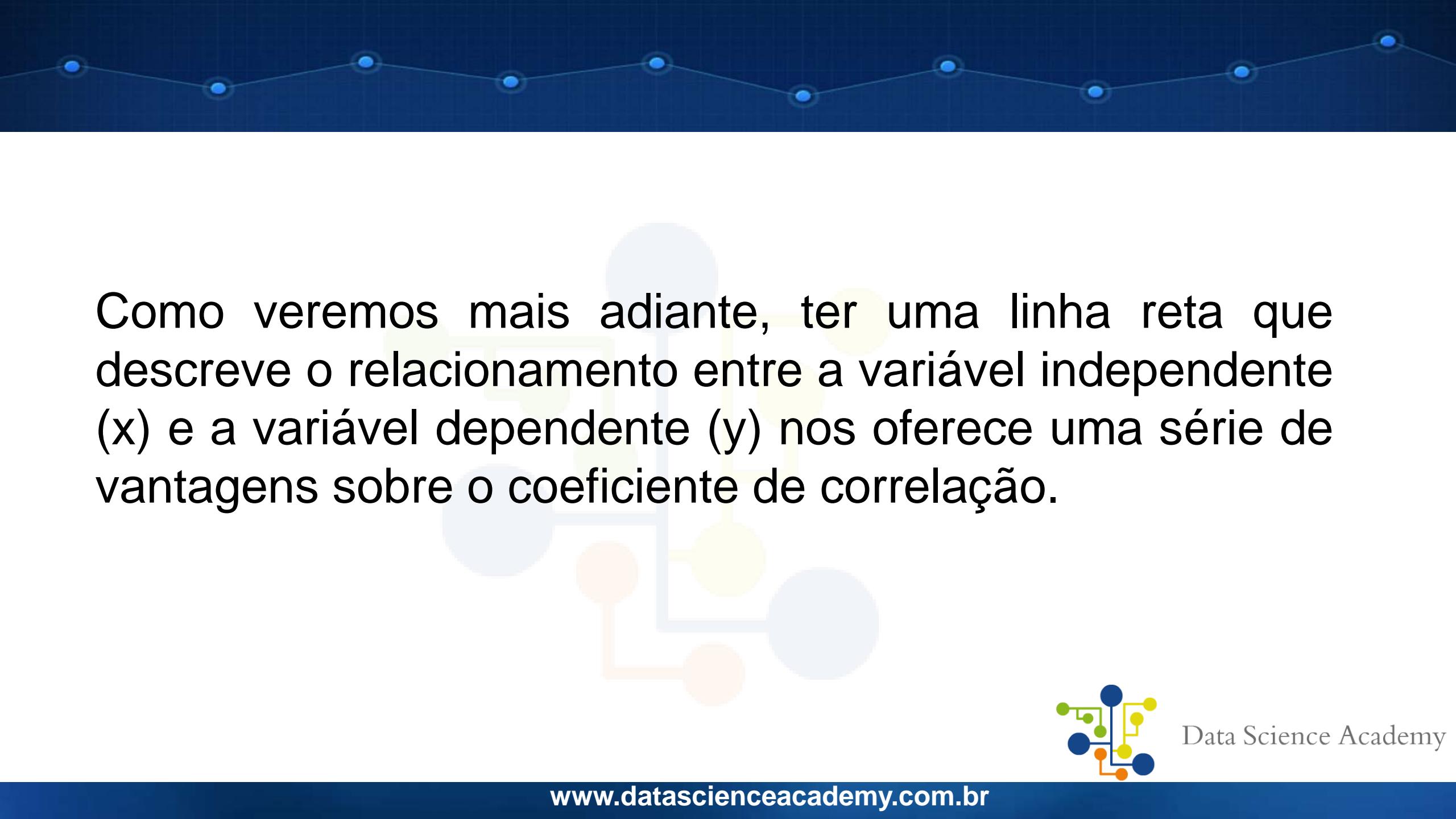
Data Science Academy



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) .



Data Science Academy



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.



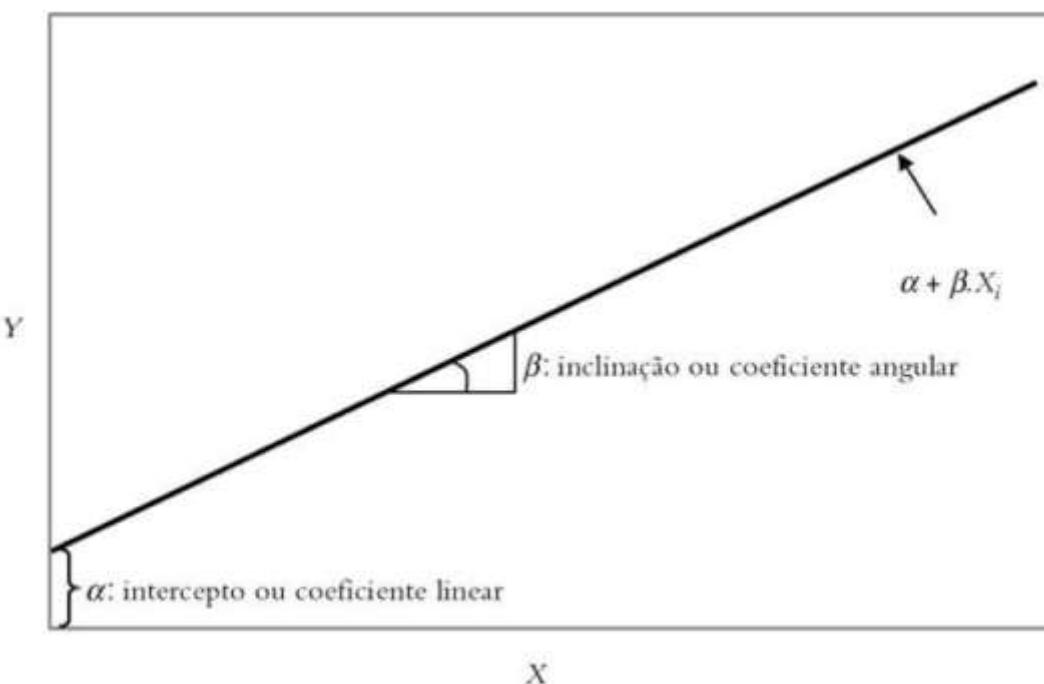
Data Science Academy

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 **explicativas, preditoras**
- a = ponto onde a linha intercepta o eixo y
- b = inclinação da linha reta



Data Science Academy



A diferença entre o valor atual e o valor previsto é conhecido como **residual**, e_i

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

onde:

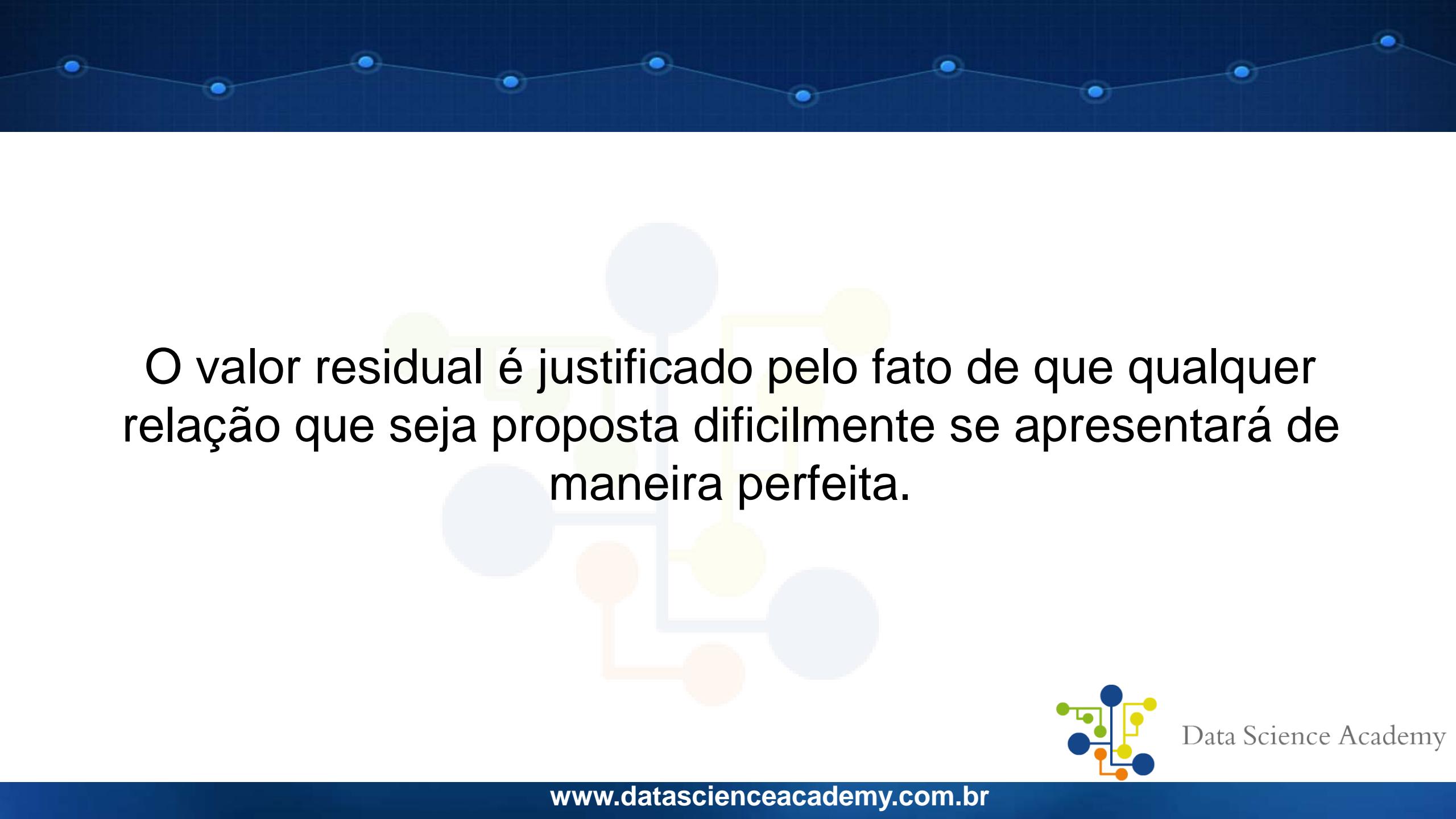
e_i = resíduo da observação *em uma posição específica* na amostra

y_i = o valor atual da variável dependente no ponto *na posição específica*

\hat{y}_i = o valor previsto da variável dependente no ponto *na posição específica*



Data Science Academy

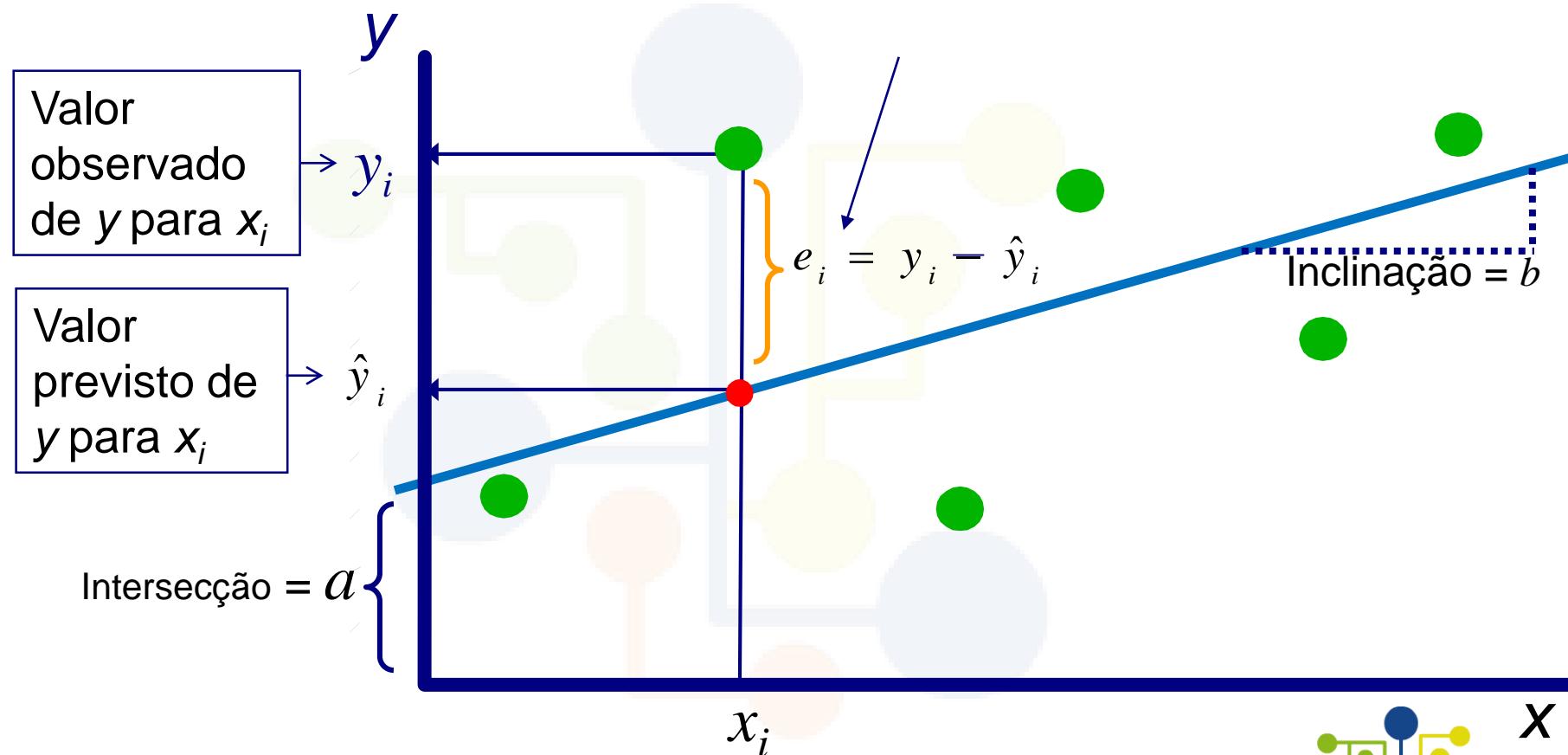


O valor residual é justificado pelo fato de que qualquer relação que seja proposta dificilmente se apresentará de maneira perfeita.



Data Science Academy

Valor residual



Data Science Academy

Regressão Linear Simples (2 variáveis)

$$\hat{Y}_i = \alpha + \beta \cdot X_i$$

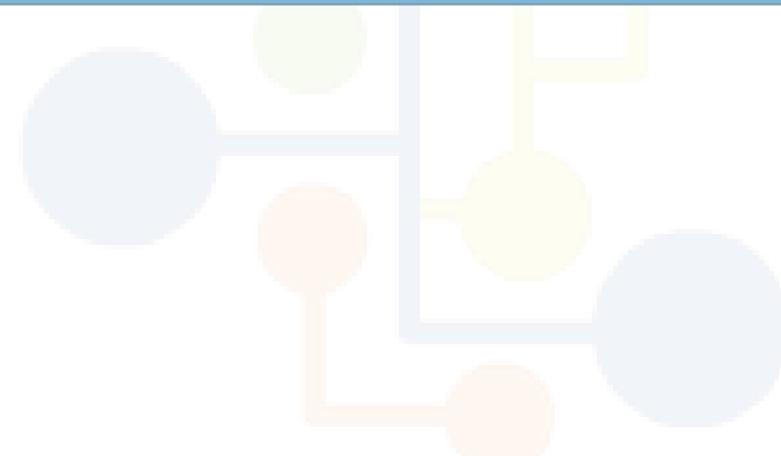
Regressão Linear Múltipla (Mais de 2 variáveis)

$$Y_i = a + b_1 \cdot X_{1i} + b_2 \cdot X_{2i} + \dots + b_k \cdot X_{ki} + u_i$$



Data Science Academy

Mínimos Quadrados Ordinários



Data Science Academy

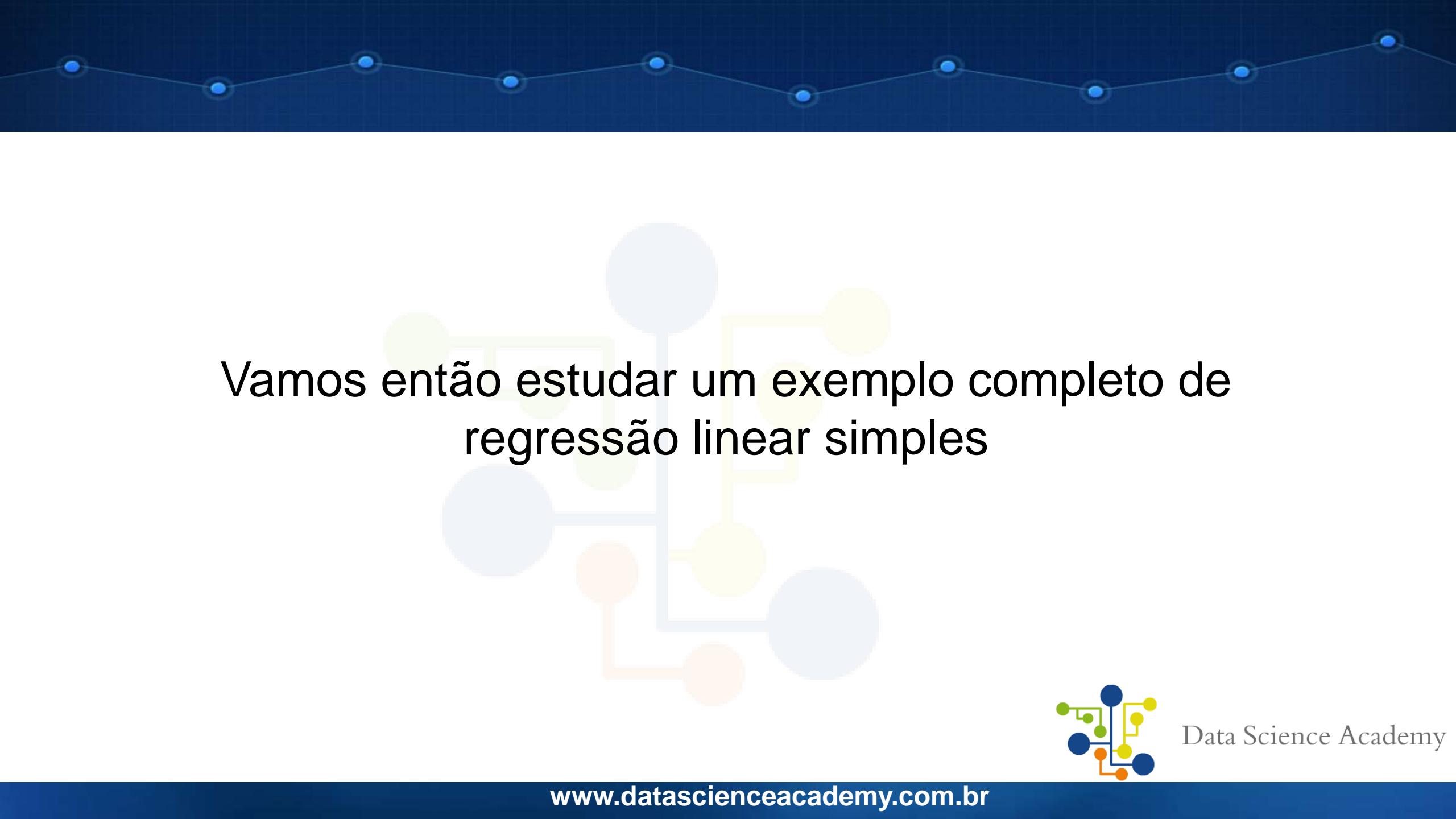


Qual o **objetivo** da Análise de Regressão?

Propiciar condições de avaliar como se comporta uma variável Y com base no comportamento de uma ou mais variáveis X, sem que necessariamente ocorra uma relação de causa e efeito.



Data Science Academy



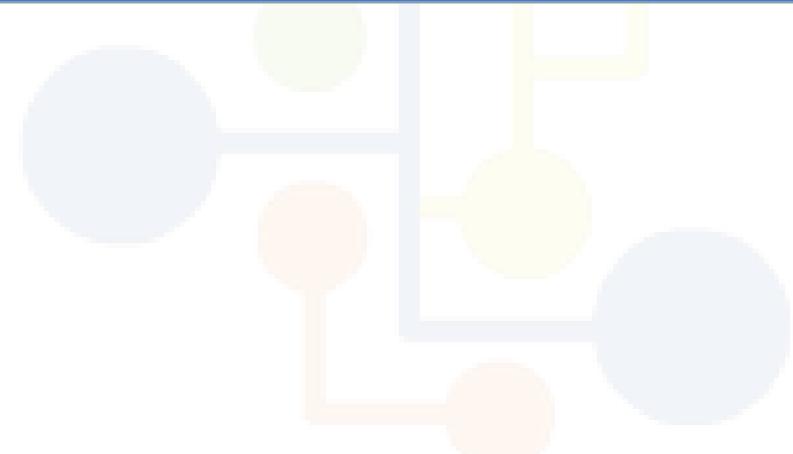
Vamos então estudar um exemplo completo de regressão linear simples



Data Science Academy



Exemplo

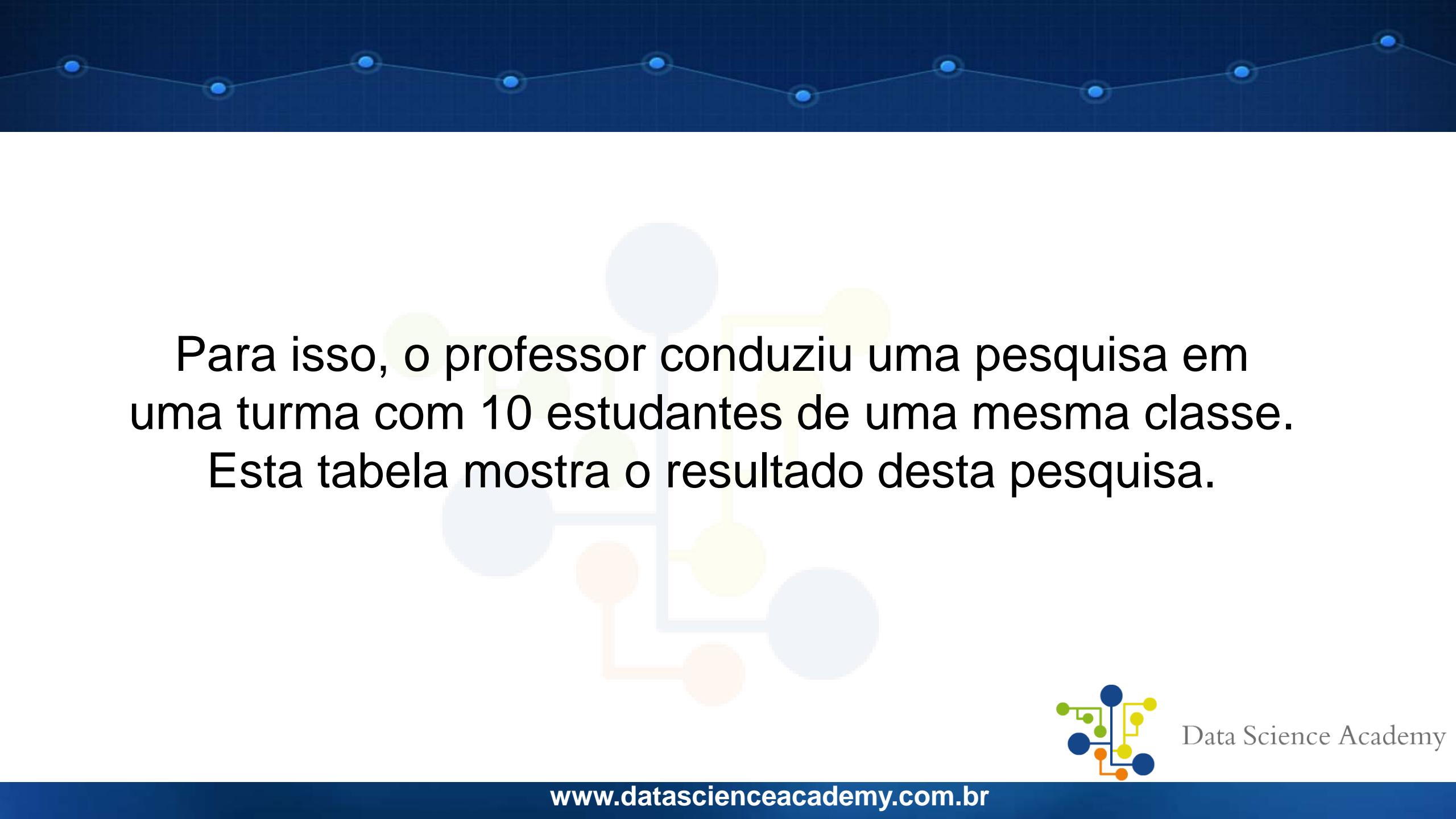


Data Science Academy

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 no exame final:



Data Science Academy



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



Data Science Academy



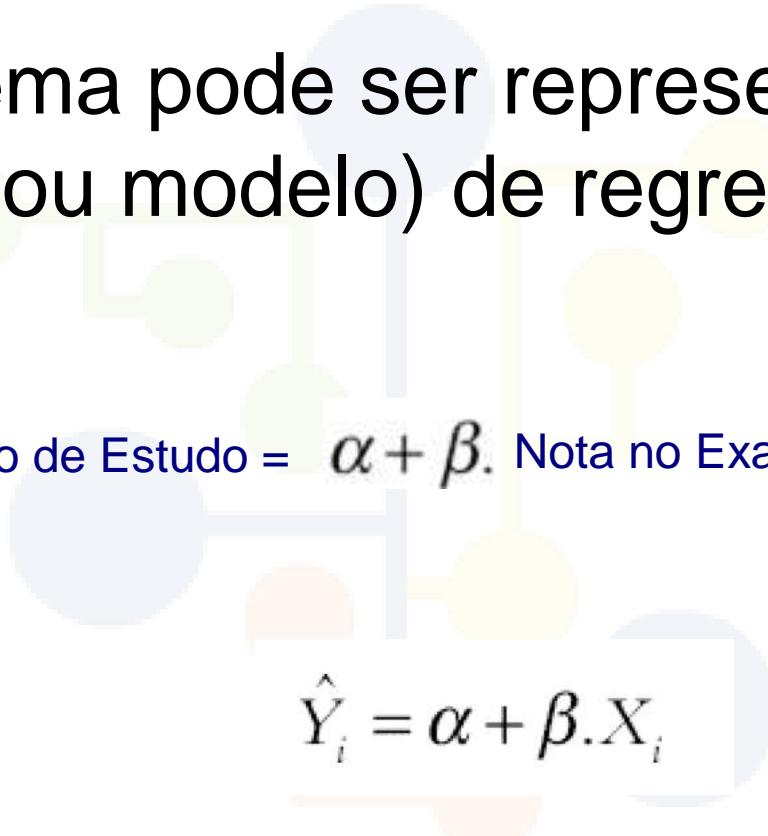
Estudante	Tempo gasto em estudo fora da sala de aula (minutos) (Y)	Nota no exame final (0 a 100) (X)
Marcio	15	24
Tiago	20	18
David	20	45
Nadir	40	60
Leonardo	50	75
Jaime	25	33
Aline	10	15
Dalton	55	96
Flavio	35	84
Henrique	30	60



Academy



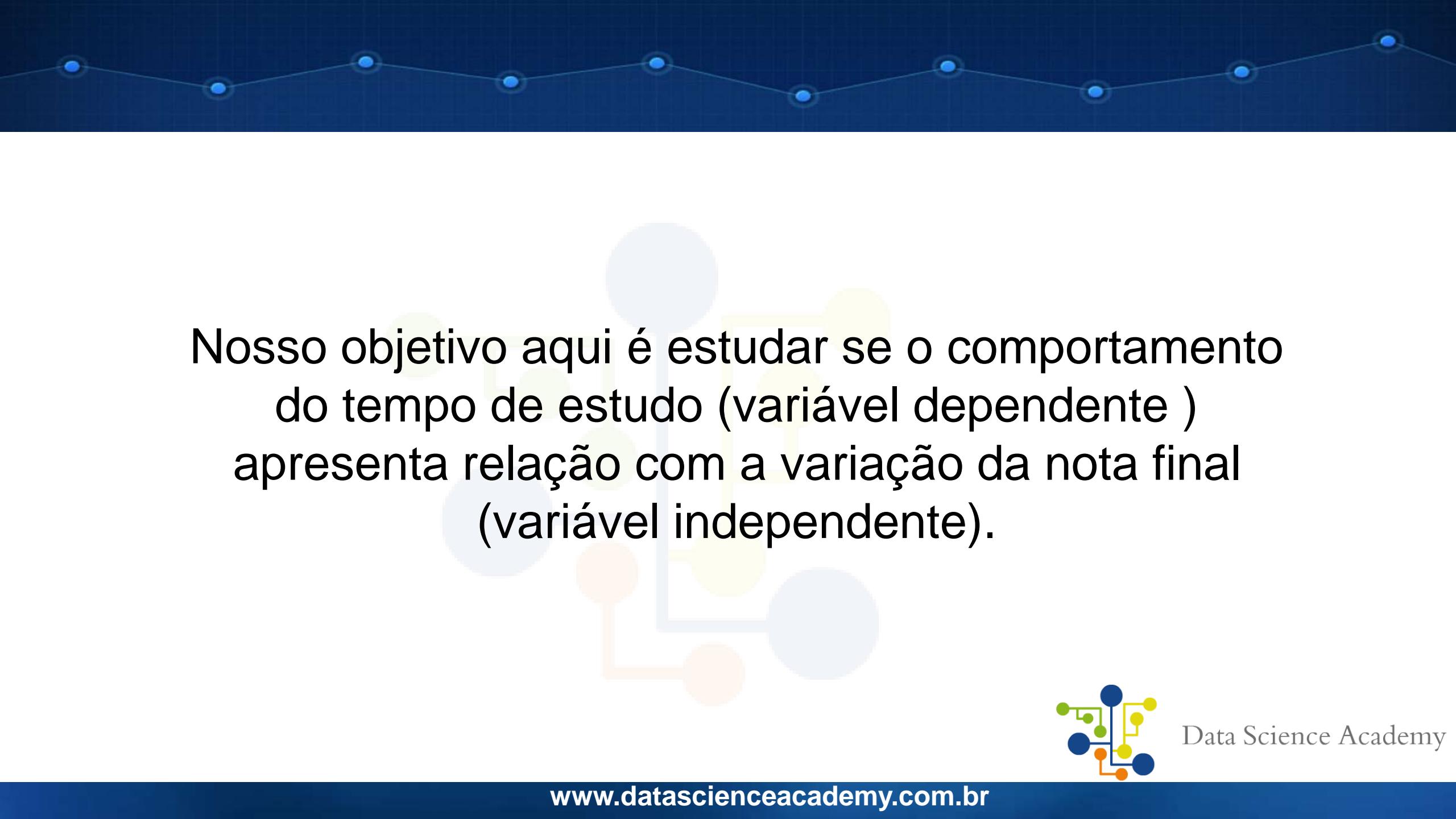
Este problema pode ser representado por esta equação (ou modelo) de regressão simples:


$$\text{Tempo de Estudo} = \alpha + \beta \cdot \text{Nota no Exame} + u_i$$

$$\hat{Y}_i = \alpha + \beta \cdot X_i$$



Data Science Academy

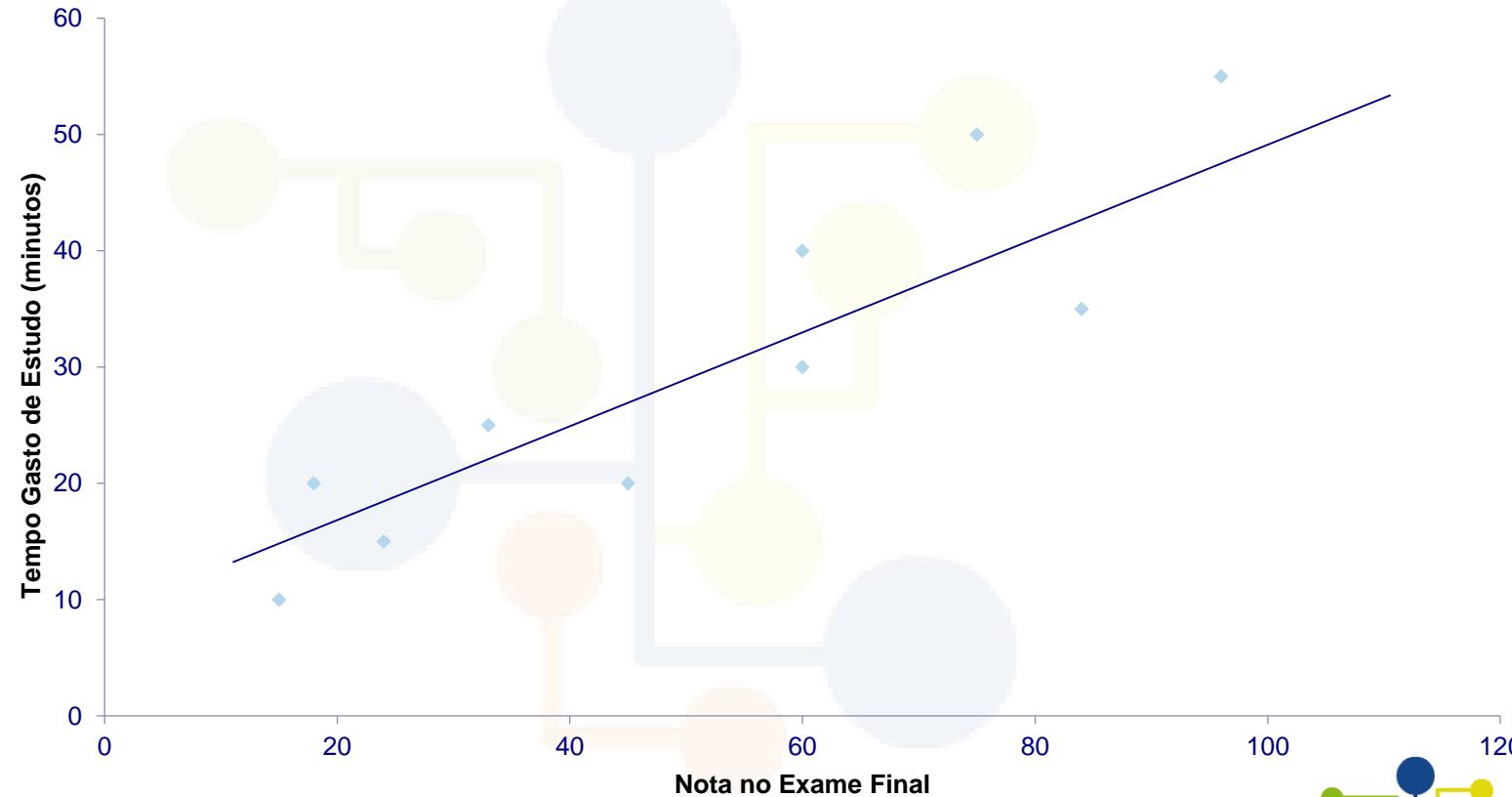


Nosso objetivo aqui é estudar se o comportamento do tempo de estudo (variável dependente) apresenta relação com a variação da nota final (variável independente).

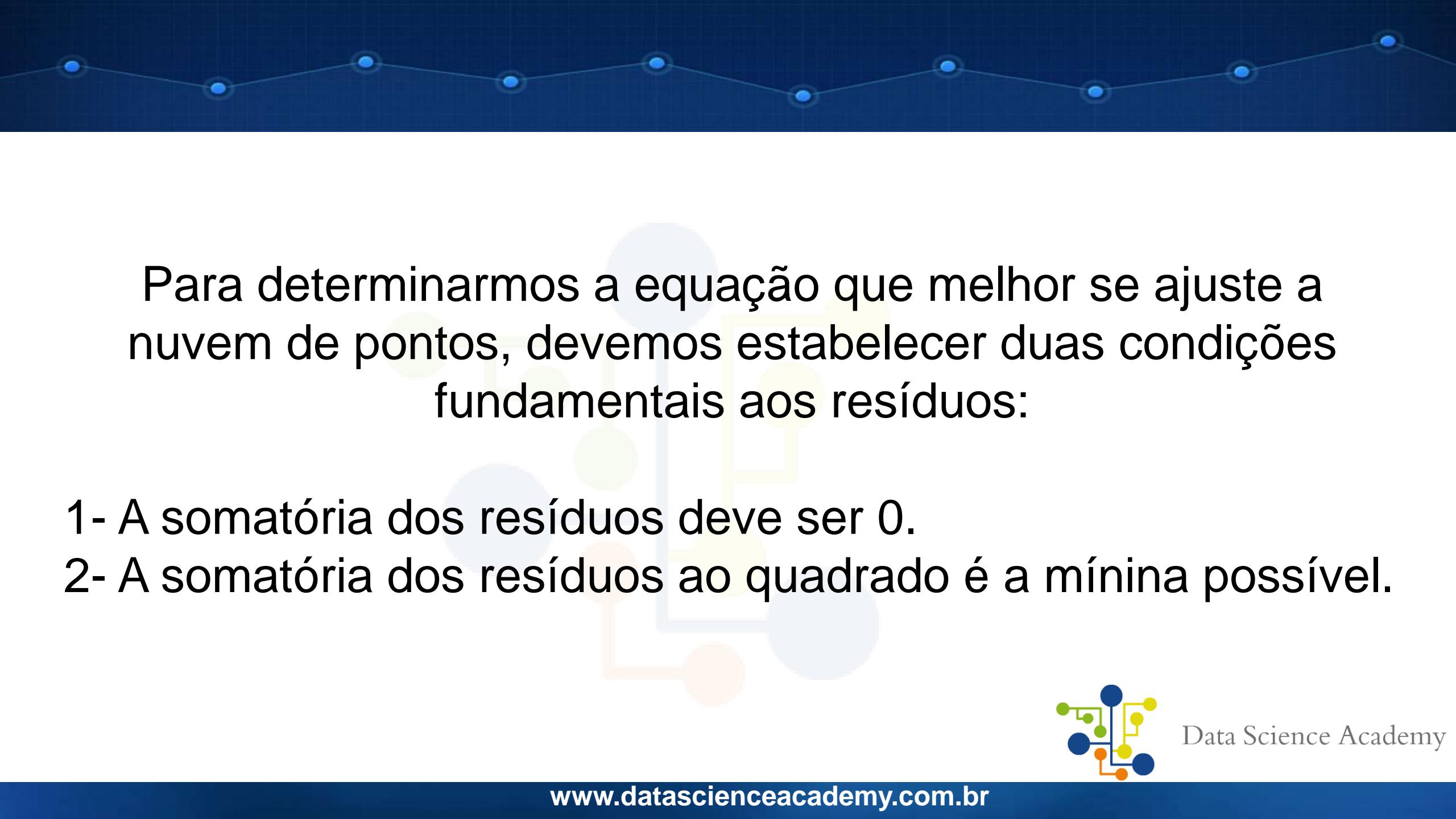


Data Science Academy

ScatterPlot



Data Science Academy

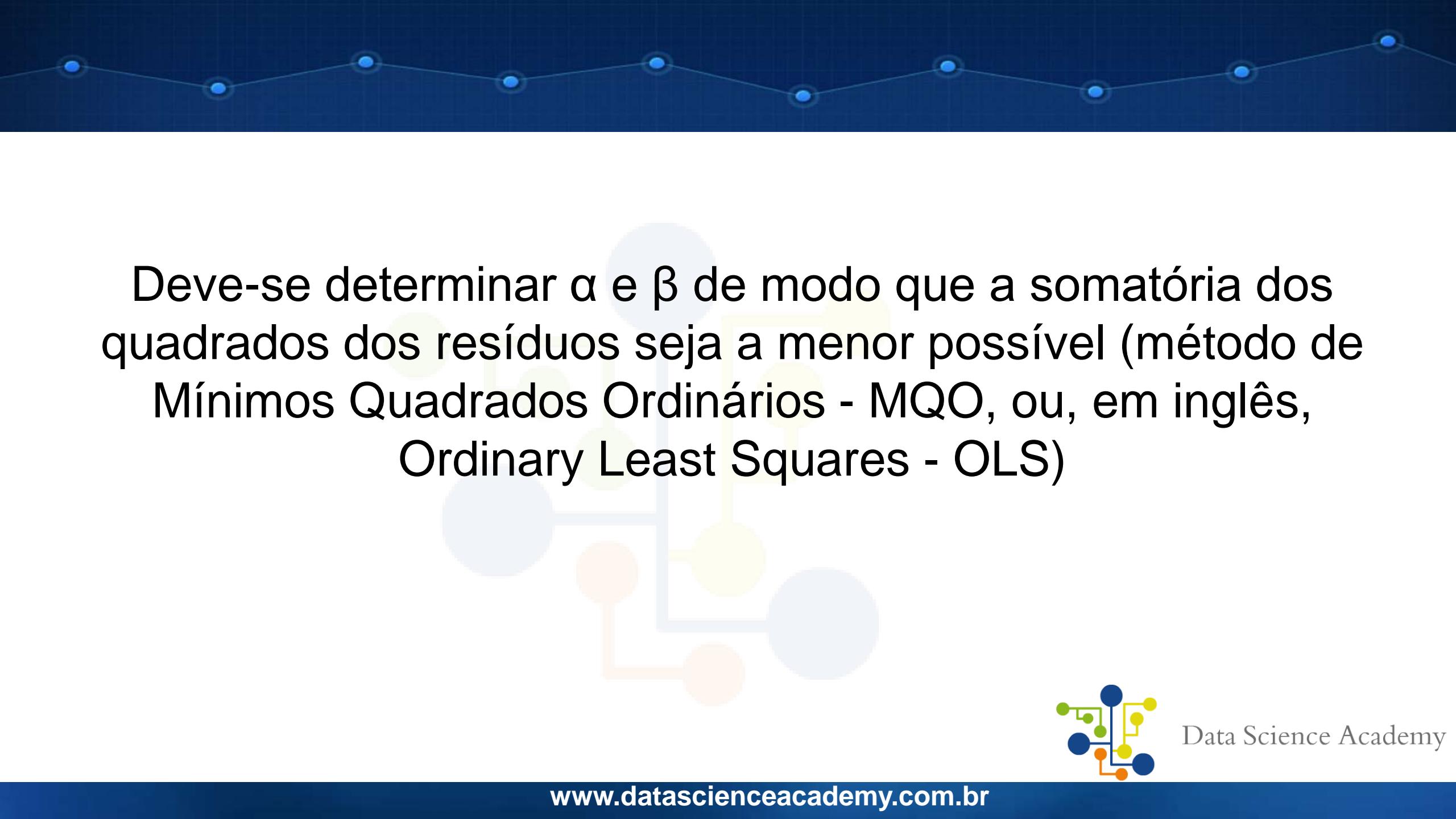


Para determinarmos a equação que melhor se ajuste a nuvem de pontos, devemos estabelecer duas condições fundamentais aos resíduos:

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



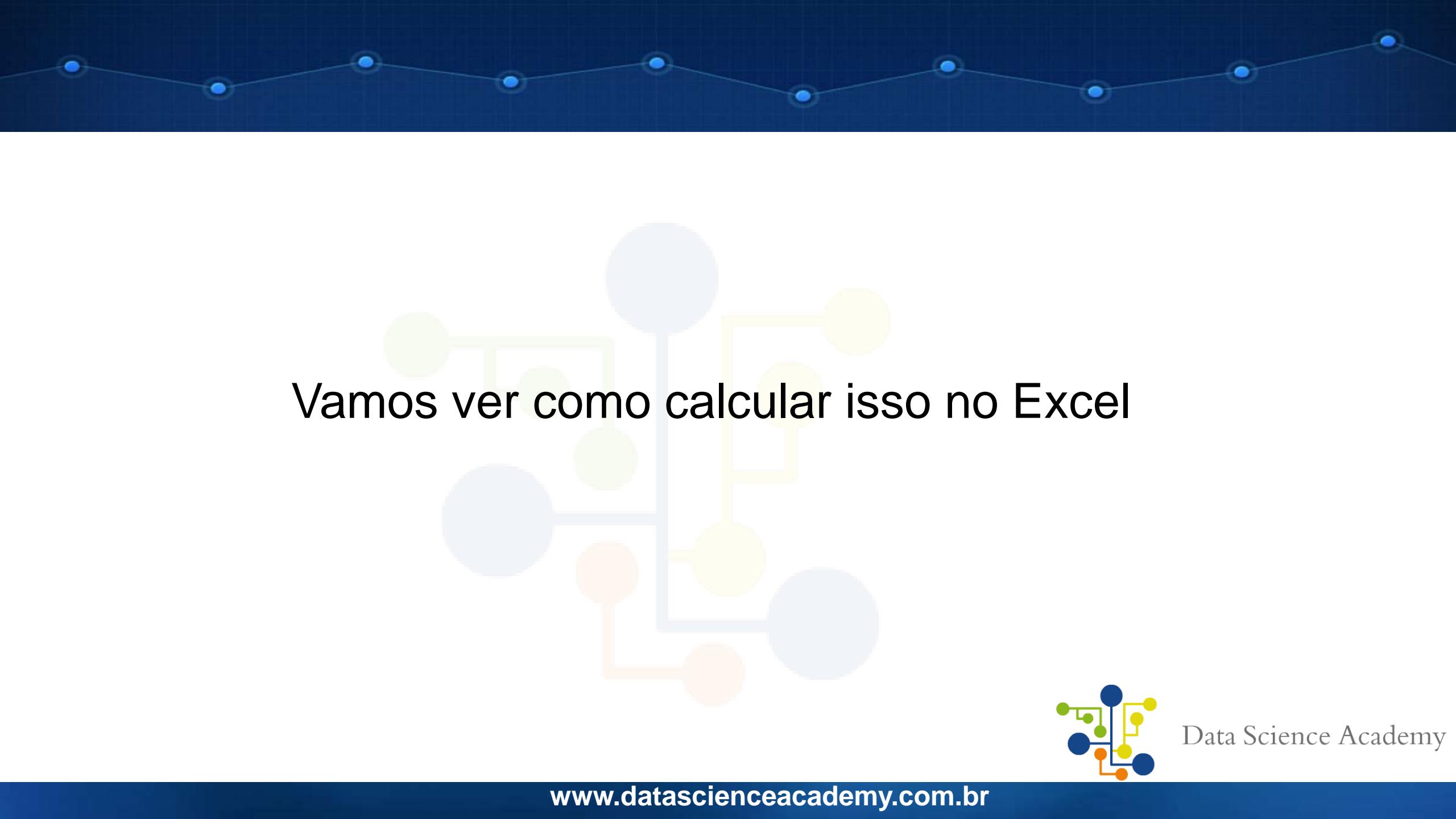
Data Science Academy



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)



Data Science Academy



Vamos ver como calcular isso no Excel



Data Science Academy

Abrindo o Excel



Análise de
Regressão



Data Science Academy

RESUMO DOS RESULTADOS

<i>Estatística de regressão</i>	
R múltiplo	0.905221341
R-Quadrado	0.819425676
R-quadrado ajustado	0.796853885
Erro padrão	6.718897311
Observações	10

RESULTADOS DE RESÍDUOS

Observação	Y previsto	Resíduos
1	17.22972973	-2.22972973
2	14.39189189	5.608108108
3	27.16216216	-7.162162162
4	34.25675676	5.743243243
5	41.35135135	8.648648649
6	21.48648649	3.513513514
7	12.97297297	-2.972972973
8	51.28378378	3.716216216
9	45.60810811	-10.60810811
10	34.25675676	-4.256756757

ANOVA

	gl	SQ	MQ	F	F de significação
Regressão	1	1638.851351	1638.851351	36.303087	0.000314449
Resíduo	8	361.1486486	45.14358108		
Total	9	2000			

	Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Inferior 95.0%	Superior 95.0%
Interseção	5.878378378	4.532327565	1.296988864	0.230788476	-4.57318773	16.32994449	-4.57318773	16.32994449
Variável X 1	0.472972973	0.078499076	6.025204312	0.000314449	0.291953778	0.653992168	0.291953778	0.653992168



	Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Inferior 95.0%	Superior 95.0%
Interseção	5.878378378	4.532327565	1.296988864	0.230788476	-4.57318773	16.32994449	-4.57318773	16.32994449
Variável X 1	0.472972973	0.078499076	6.025204312	0.000314449	0.291953778	0.653992168	0.291953778	0.653992168

Coeficientes	Erro padrão
5.878378378	4.532327565
0.472972973	0.078499076

Coeficientes de Regressão

$$y = a + b \cdot x$$

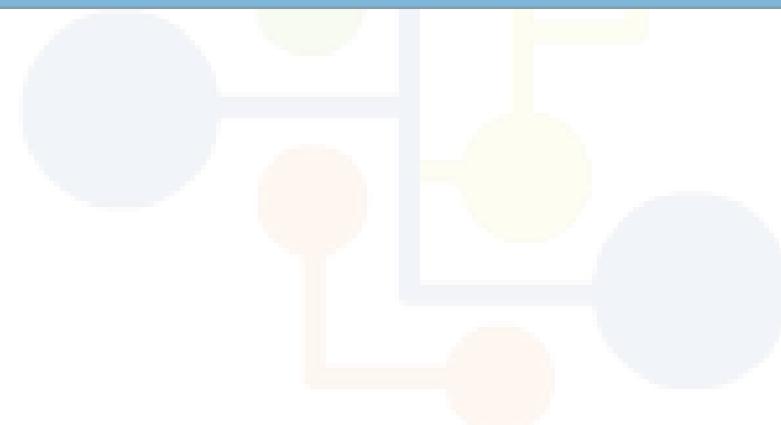
$$y = 5.878378278 + 0.472972973 \cdot x$$



Data Science Academy



Estatísticas de Regressão



Data Science Academy

- Soma Total dos Quadrados (STQ) – Mostra a variação em Y em torno da própria média.
- Soma dos Quadrados de Regressão (SQR) – Oferece a variação de Y considerando as variáveis X utilizadas no modelo.
- Soma dos Quadrados dos Resíduos (SQU) – Variação de Y que não é explicada pelo modelo elaborado.

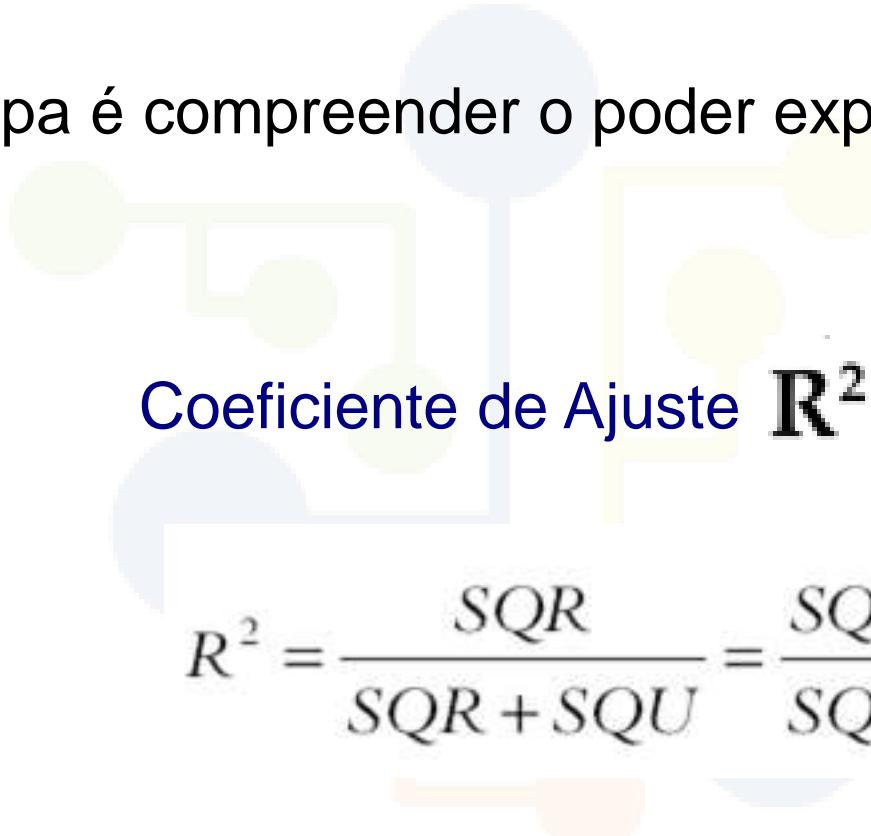
$$STQ = SQR + SQU$$



Data Science Academy



Nossa próxima etapa é compreender o poder explicativo do modelo de regressão

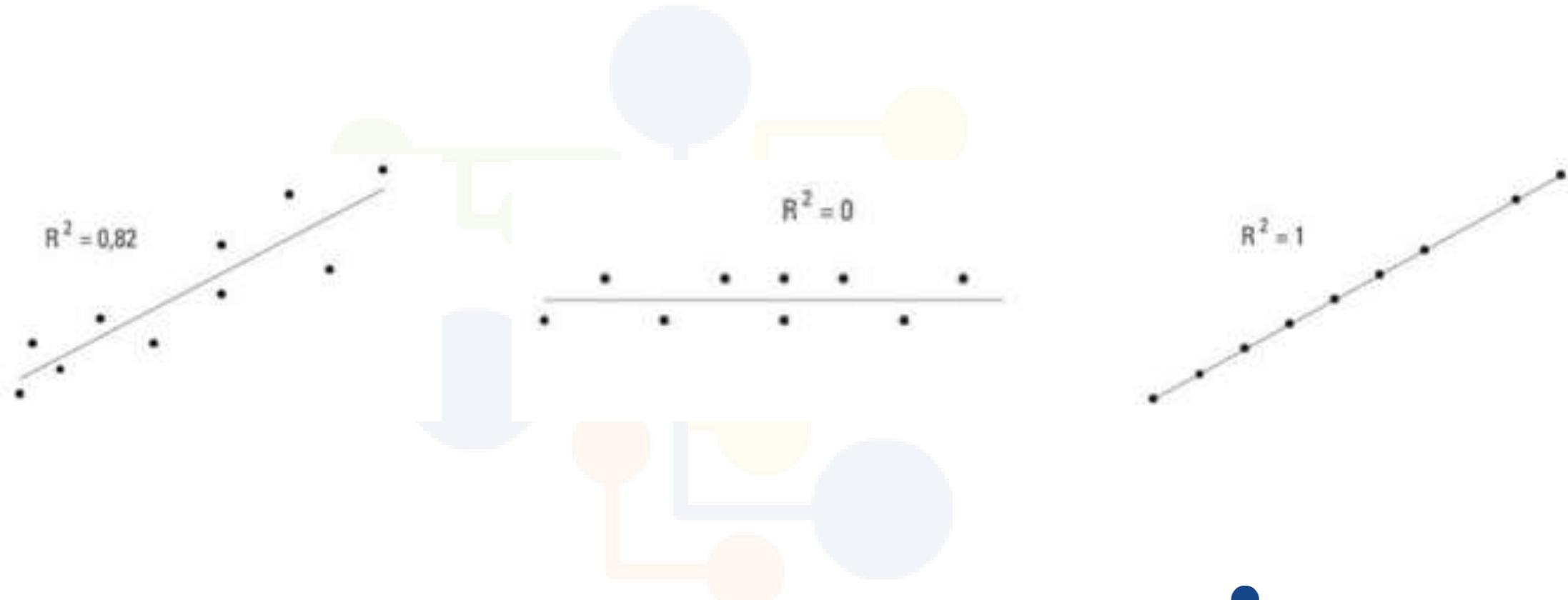


Coeficiente de Ajuste **R²**

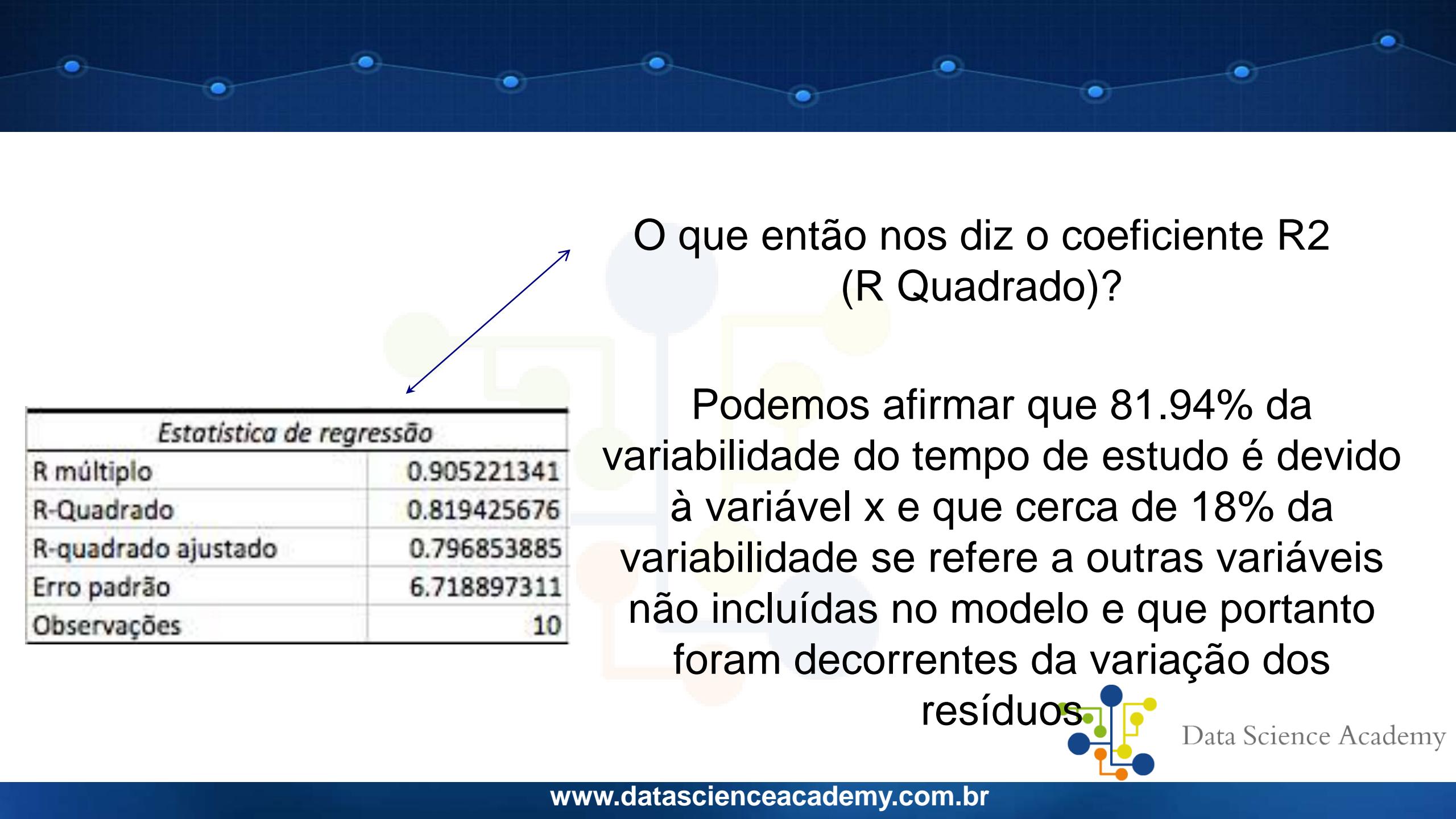
$$R^2 = \frac{SQR}{SQR + SQU} = \frac{SQR}{SQT}$$



Data Science Academy



Data Science Academy

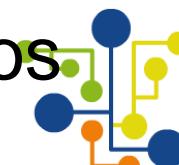


O que então nos diz o coeficiente R² (R Quadrado)?



Estatística de regressão	
R múltiplo	0.905221341
R-Quadrado	0.819425676
R-quadrado ajustado	0.796853885
Erro padrão	6.718897311
Observações	10

Podemos afirmar que 81.94% da variabilidade do tempo de estudo é devido à variável x e que cerca de 18% da variabilidade se refere a outras variáveis não incluídas no modelo e que portanto foram decorrentes da variação dos resíduos.



Data Science Academy

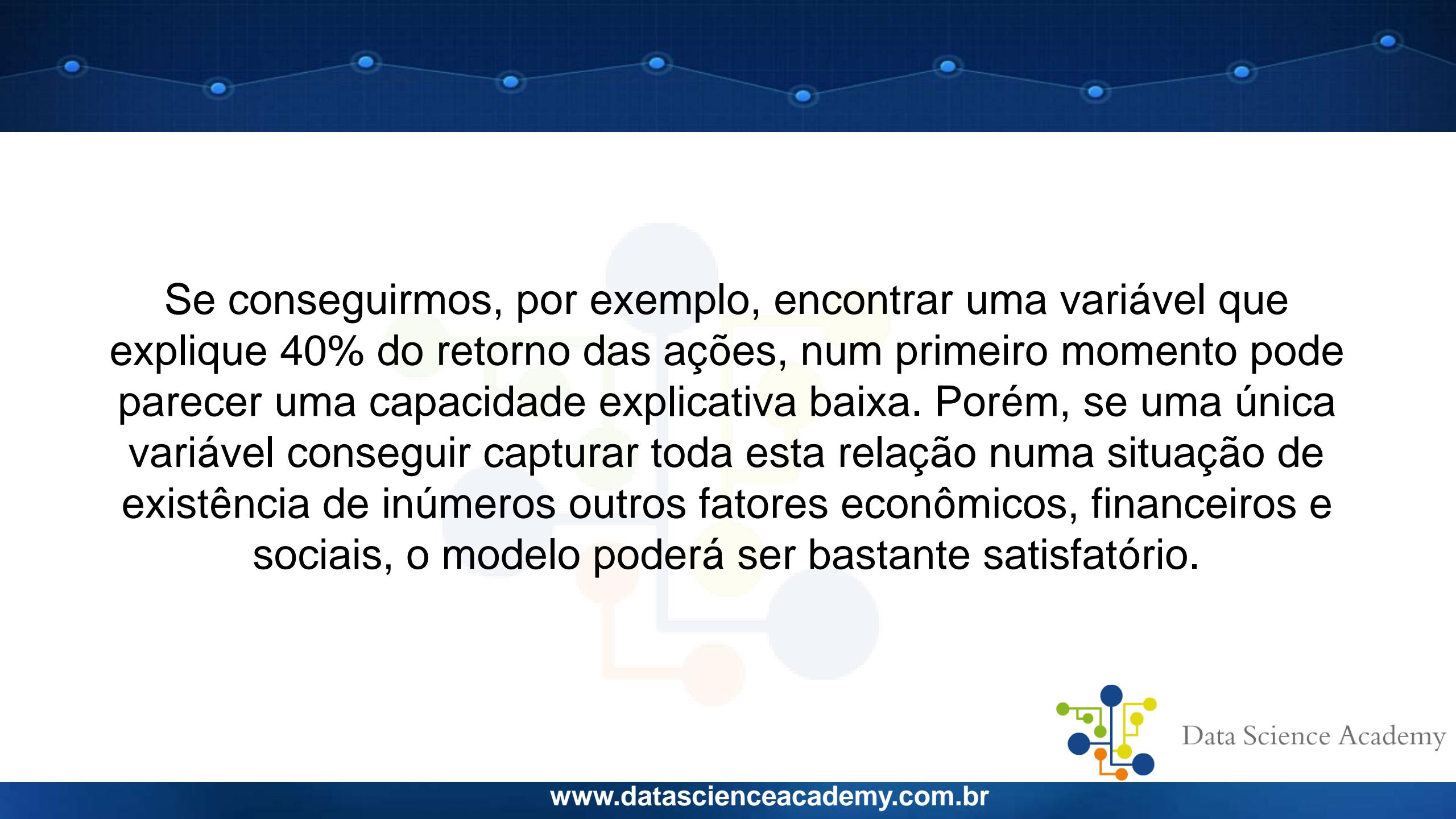


<i>Estatística de regressão</i>	
R múltiplo	0.905221341
R-Quadrado	0.819425676
R-quadrado ajustado	0.796853885
Erro padrão	6.718897311
Observações	10

O coeficiente de ajuste R2 não diz aos analistas se uma determinada variável explicativa é estatisticamente significante e se esta variável é a causa verdadeira da alteração de comportamento da variável dependente.



Data Science Academy



Se conseguirmos, por exemplo, encontrar uma variável que explique 40% do retorno das ações, num primeiro momento pode parecer uma capacidade explicativa baixa. Porém, se uma única variável conseguir capturar toda esta relação numa situação de existência de inúmeros outros fatores econômicos, financeiros e sociais, o modelo poderá ser bastante satisfatório.



Data Science Academy



Análise de Variância (ANOVA)



Data Science Academy

Inicialmente, é de fundamental importância que estudemos a significância estatística geral do modelo estimado. Com tal finalidade, devemos fazer uso do **teste F**, cujas hipóteses nula e alternativa, para um modelo geral de regressão, são, respectivamente:

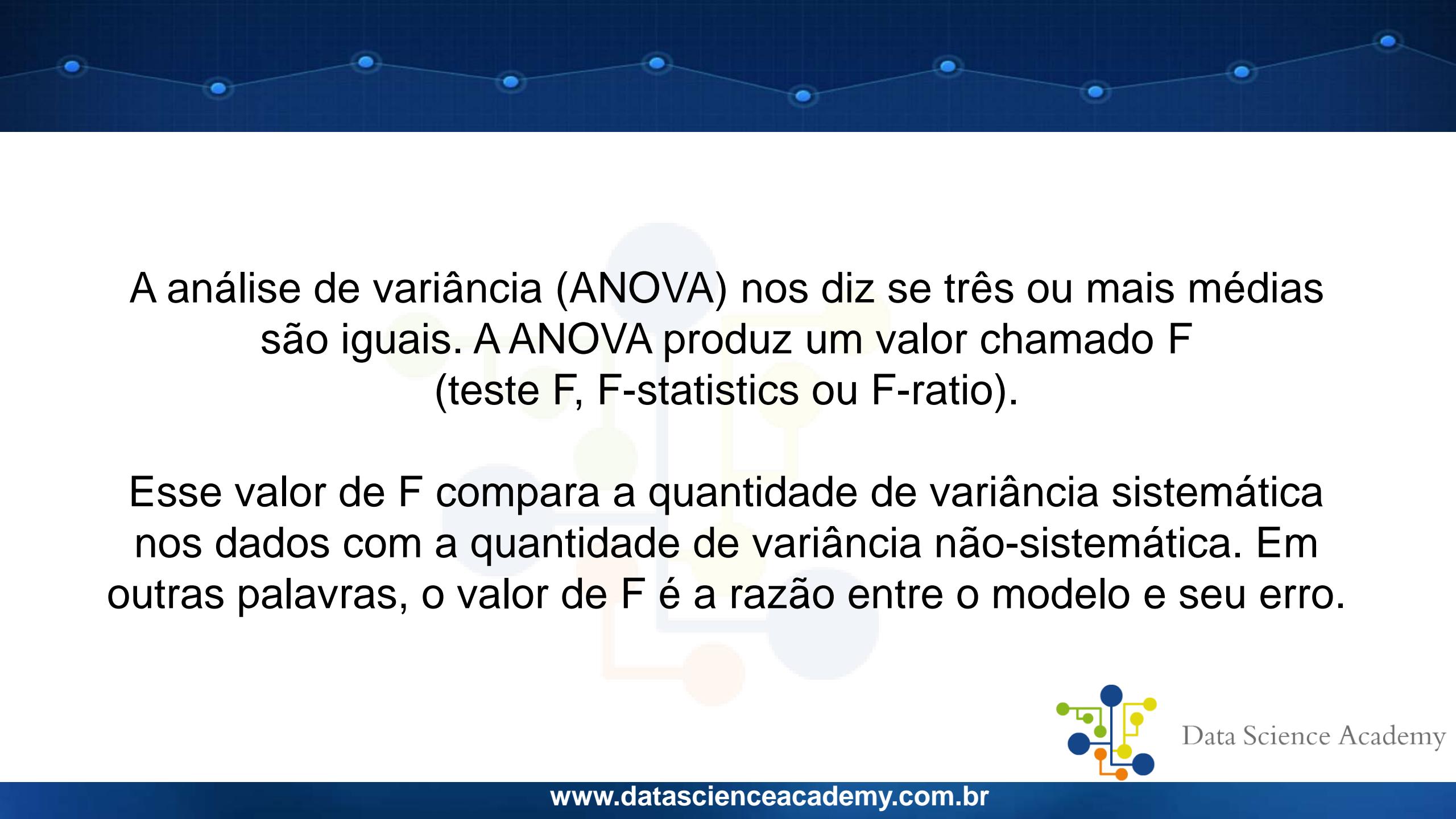
$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$
$$H_1: \text{existe pelo menos um } \beta_j \neq 0$$



$$H_0: \beta = 0$$
$$H_1: \beta \neq 0$$



Data Science Academy



A análise de variância (ANOVA) nos diz se três ou mais médias são iguais. A ANOVA produz um valor chamado F (teste F, F-statistics ou F-ratio).

Esse valor de F compara a quantidade de variância sistemática nos dados com a quantidade de variância não-sistêmática. Em outras palavras, o valor de F é a razão entre o modelo e seu erro.



Data Science Academy

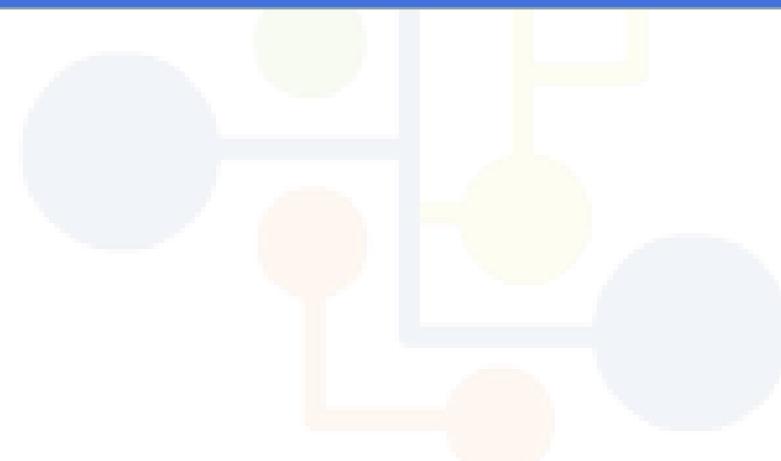
Difícil ainda, né?



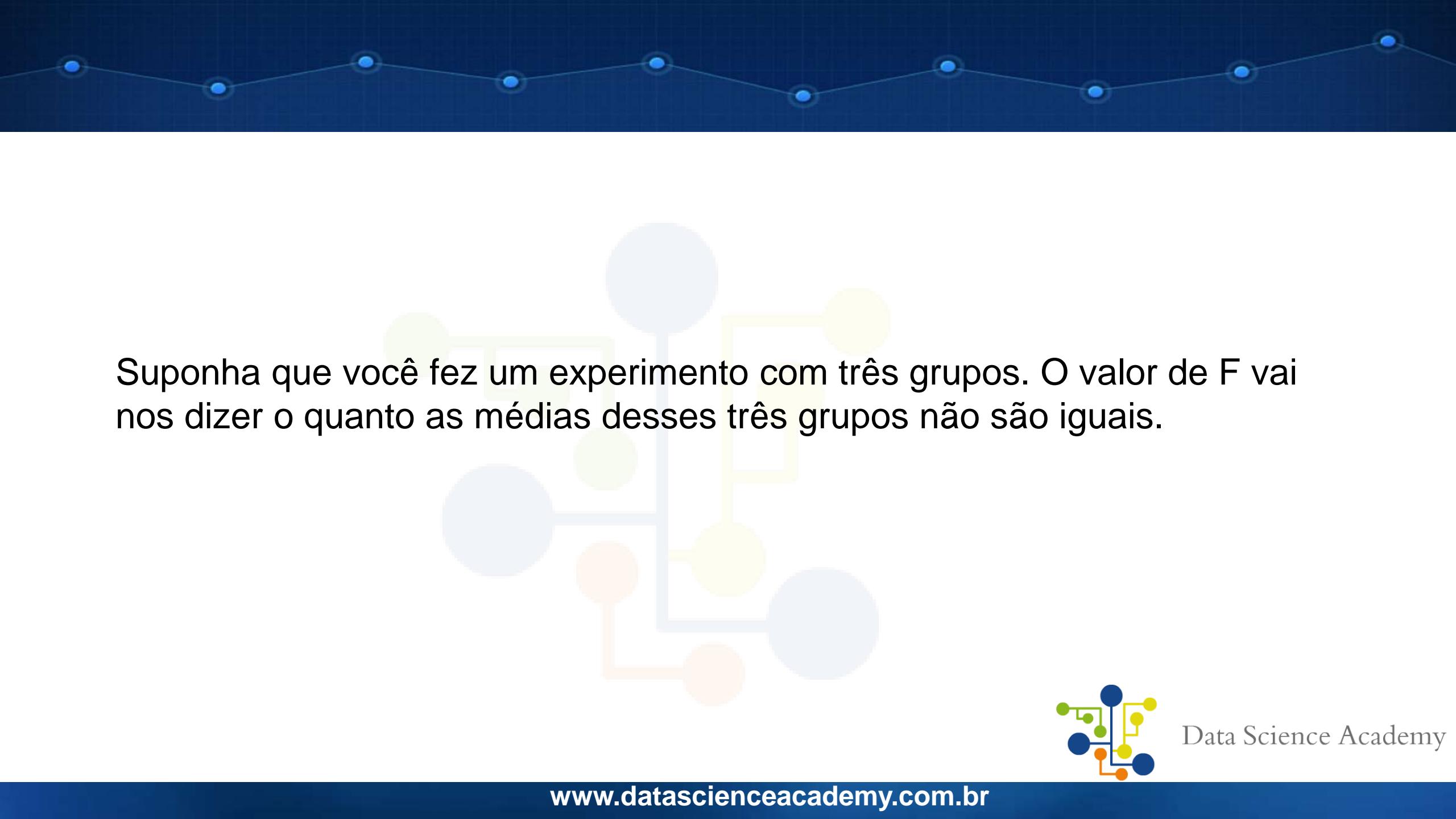
Data Science Academy



Exemplo



Data Science Academy



Suponha que você fez um experimento com três grupos. O valor de F vai nos dizer o quanto as médias desses três grupos não são iguais.



Data Science Academy

ANOVA					
	gl	SQ	MQ	F	F de significação
Regressão	1	1638.851351	1638.851351	36.303087	0.000314449
Resíduo	8	361.1486486	45.14358108		
Total	9	2000			

Voltando ao nosso exemplo:

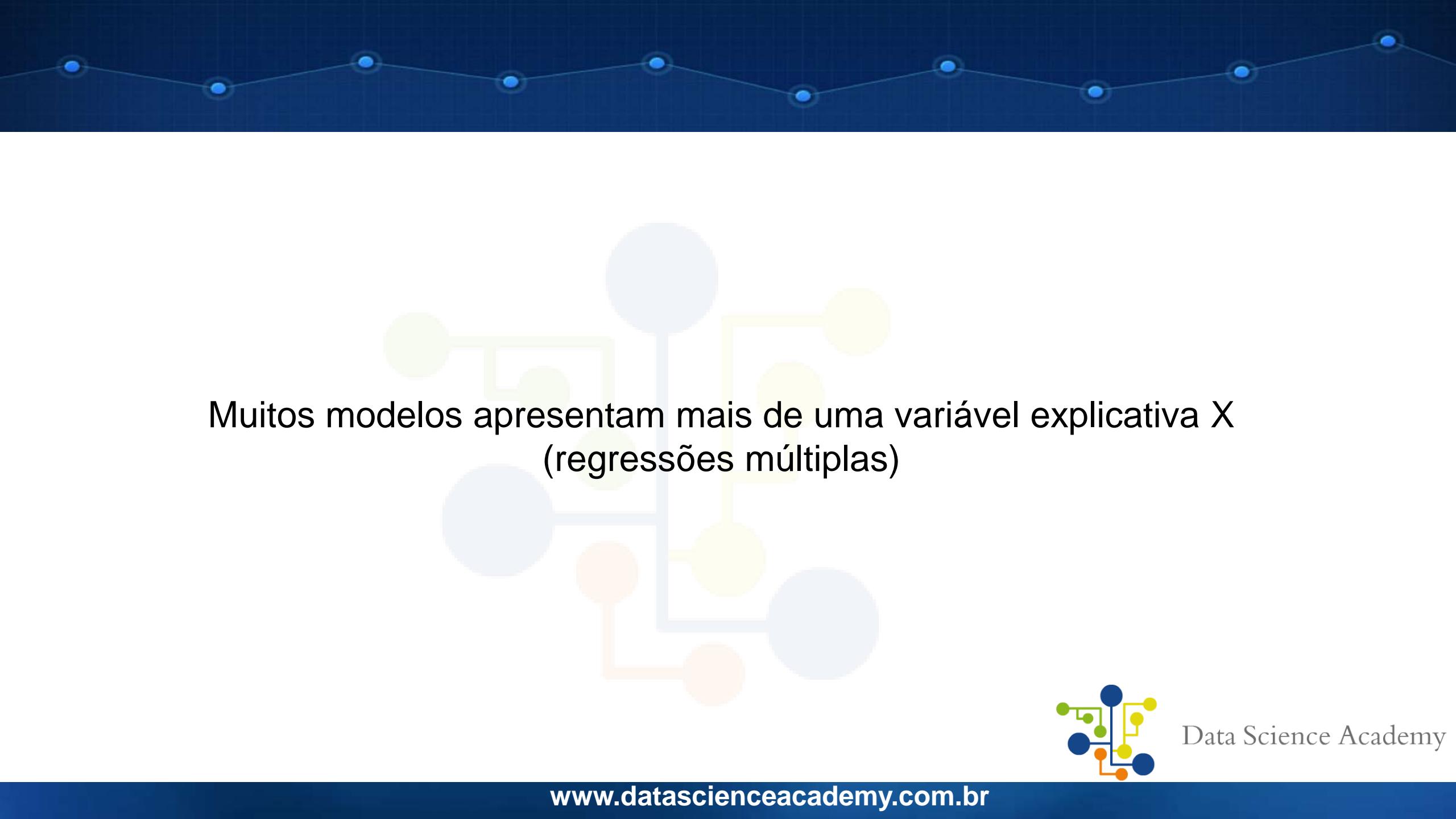
Temos $F = 36.30$ e seu respectivo nível de significância.

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$



Data Science Academy

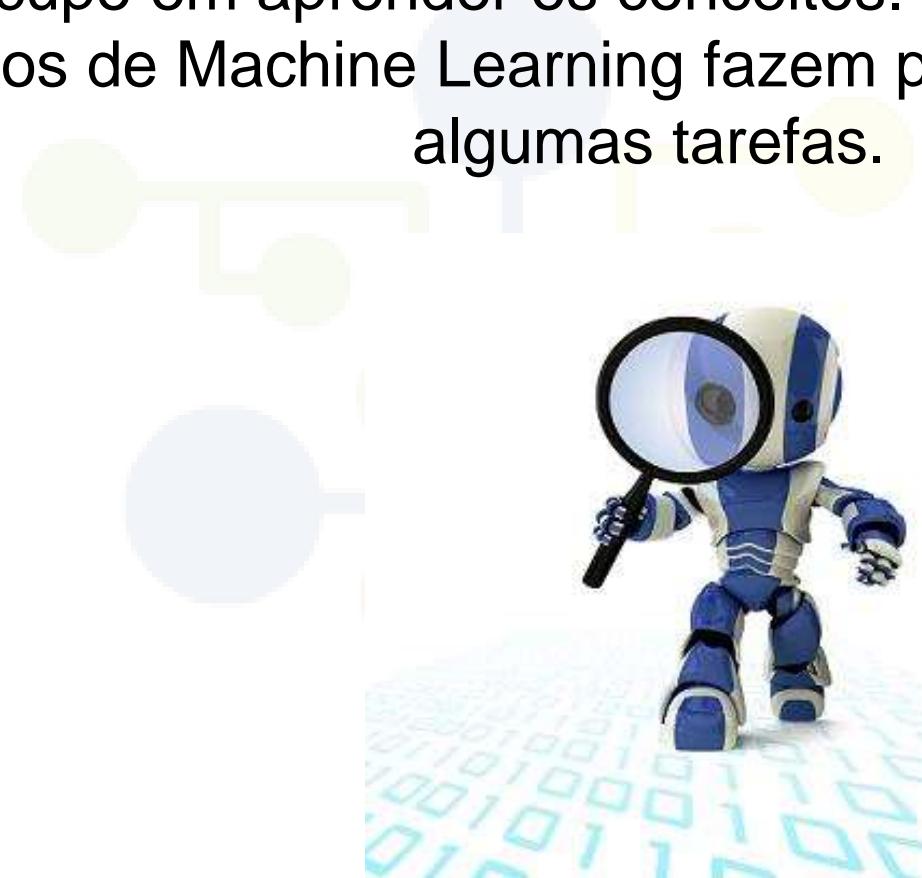


Muitos modelos apresentam mais de uma variável explicativa X
(regressões múltiplas)



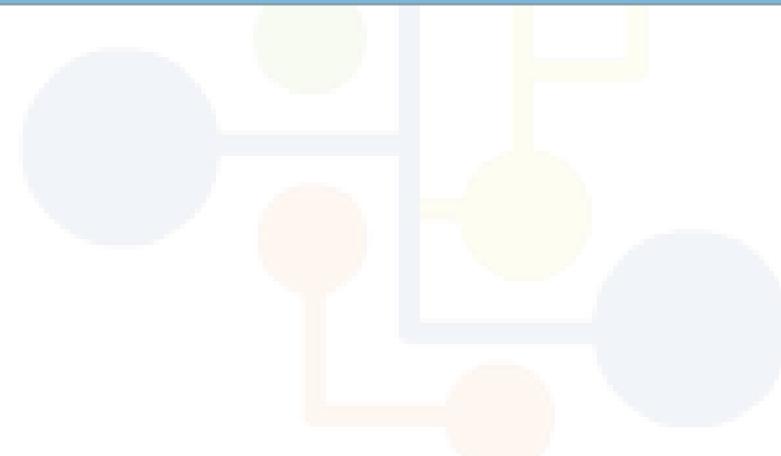
Data Science Academy

Se preocupe em aprender os conceitos. Todos os cálculos, os algoritmos de Machine Learning fazem por nós, além de mais algumas tarefas.



Data Science Academy

Coeficientes de Regressão



Data Science Academy



Estatística t

Significância estatística de cada parâmetro a ser considerado no modelo de regressão e as hipóteses do teste correspondente, o teste T



Data Science Academy

Estatística t

$$H_0: \alpha = 0$$

$$H_1: \alpha \neq 0$$

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

Este teste propicia ao pesquisador uma verificação sobre a significância estatística de cada parâmetro estimado, α e β .



Data Science Academy

	<i>Coeficientes</i>	<i>Erro padrão</i>	<i>Stat t</i>	<i>valor P</i>
Interseção	5.878378378	4.532327565	1.296988864	0.230788476
Variável X 1	0.472972973	0.078499076	6.025204312	0.000314449

valor P

Estatística t

Este teste propicia ao pesquisador uma verificação sobre a significância estatística de cada parâmetro estimado, α e β .



Data Science Academy

Você falou em **teste t**.
O que é isso?



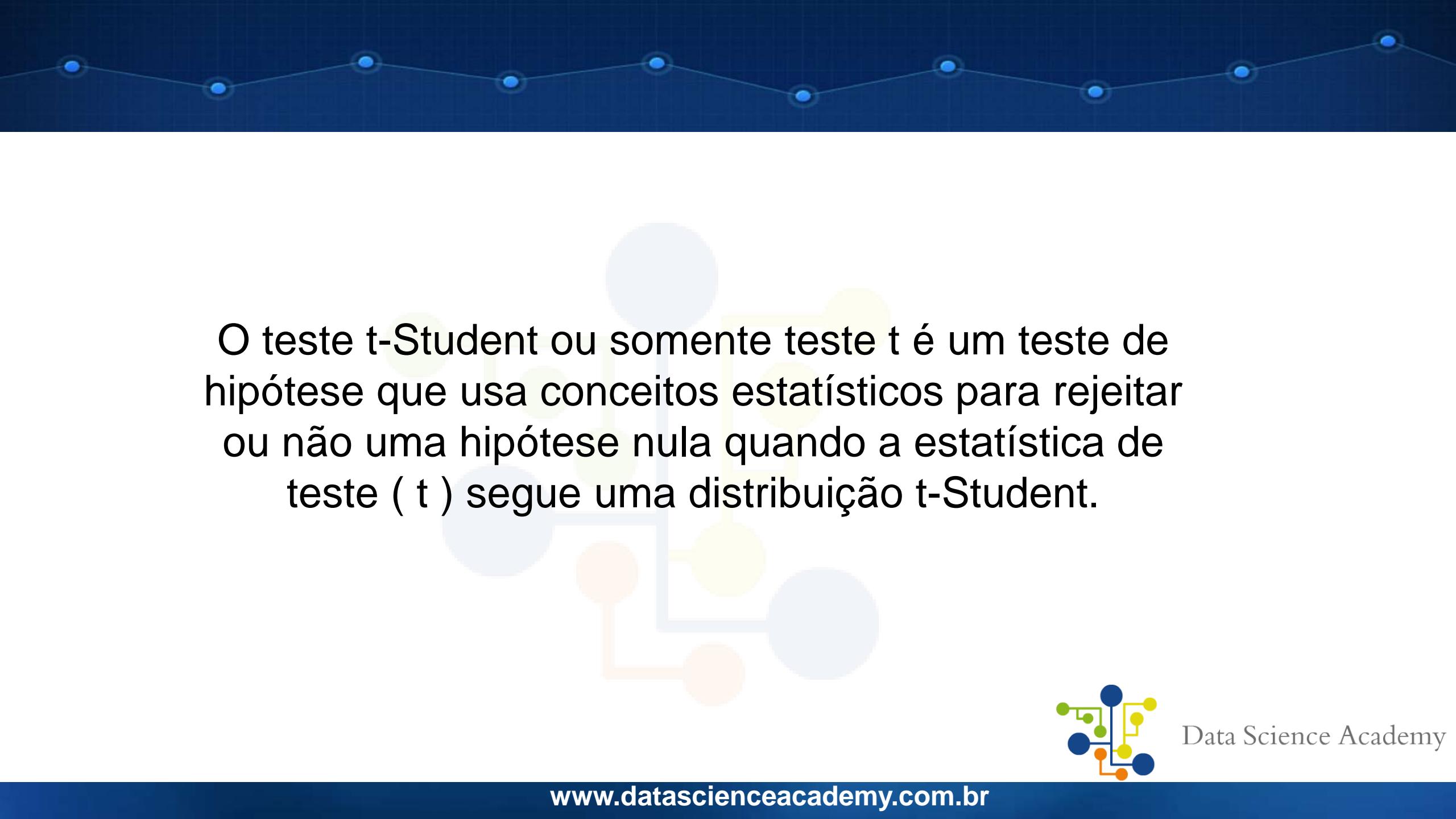
Data Science Academy



Teste t-Student



Data Science Academy



O teste t-Student ou somente teste t é um teste de hipótese que usa conceitos estatísticos para rejeitar ou não uma hipótese nula quando a estatística de teste (t) segue uma distribuição t-Student.



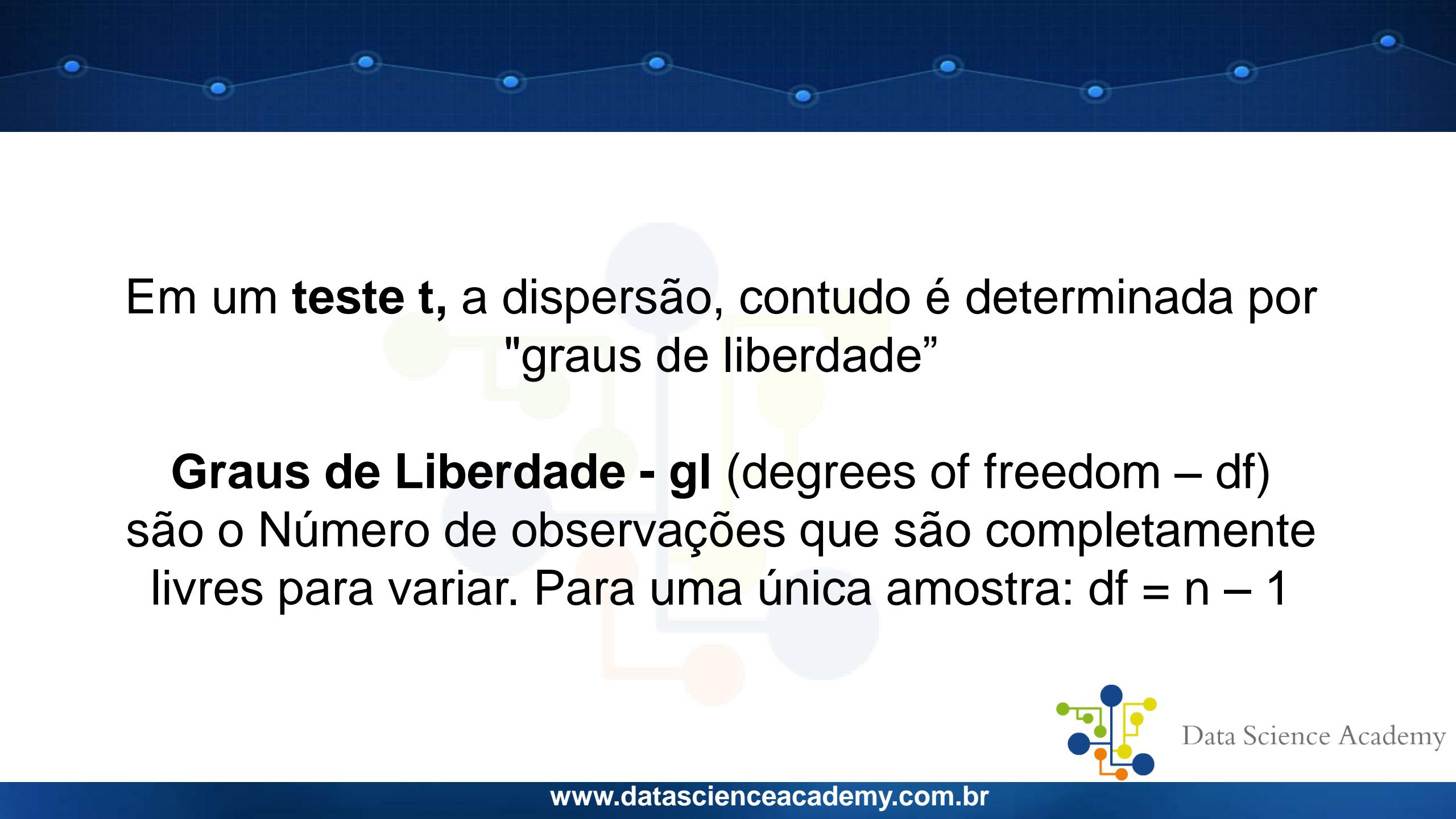
Data Science Academy

Teste t pode ser conduzido para:

- Comparar uma amostra com uma população
- Comparar duas amostras pareadas (mesmos sujeitos em dois momentos distintos)
- Comparar duas amostras independentes



Data Science Academy



Em um **teste t**, a dispersão, contudo é determinada por
"graus de liberdade"

Graus de Liberdade - gl (degrees of freedom – df)
são o Número de observações que são completamente
livres para variar. Para uma única amostra: $df = n - 1$



Data Science Academy

E o p-value?



Data Science Academy

p-value é a probabilidade, quando H_0 é verdadeira, de observar uma amostra tão ou mais diferente/rara (na direção de H_A) do que a amostra que temos

- não é uma suposição de risco
- p simplesmente descreve a “raridade” da amostra que se tem.
- se $p \leq \alpha$, a amostra é suficientemente rara para se rejeitar H_0 .



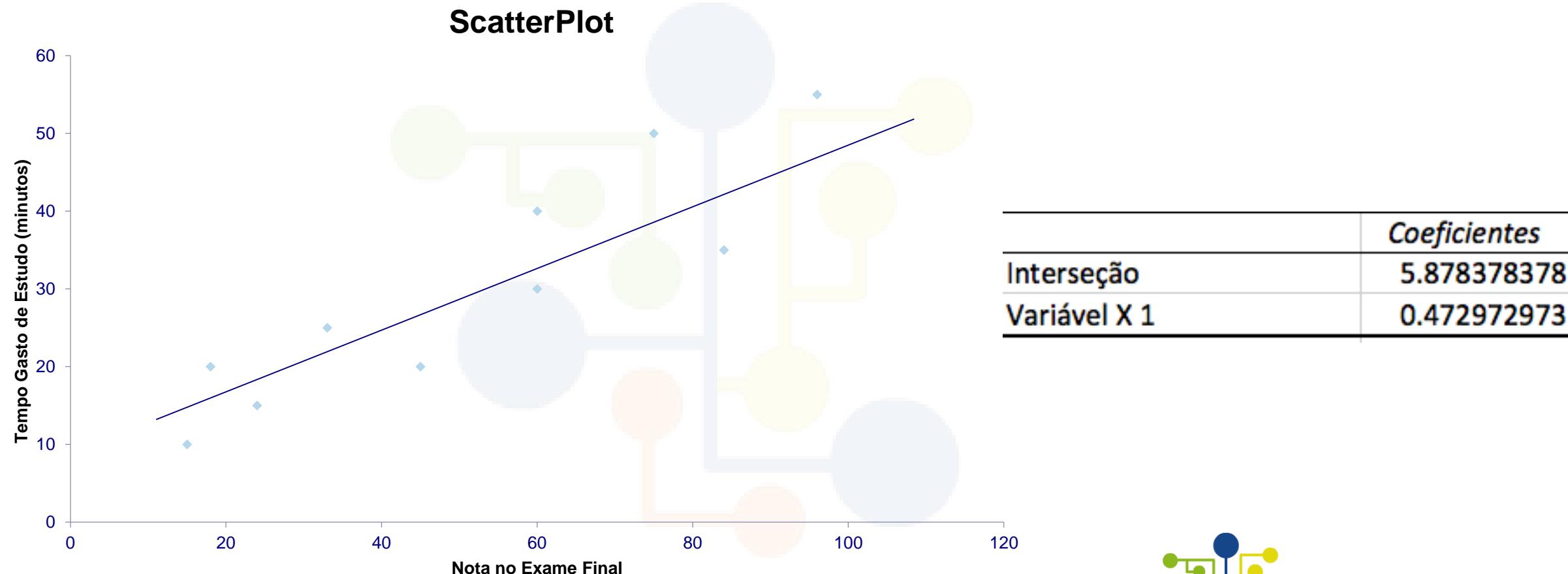
Data Science Academy

Intervalos de Confiança

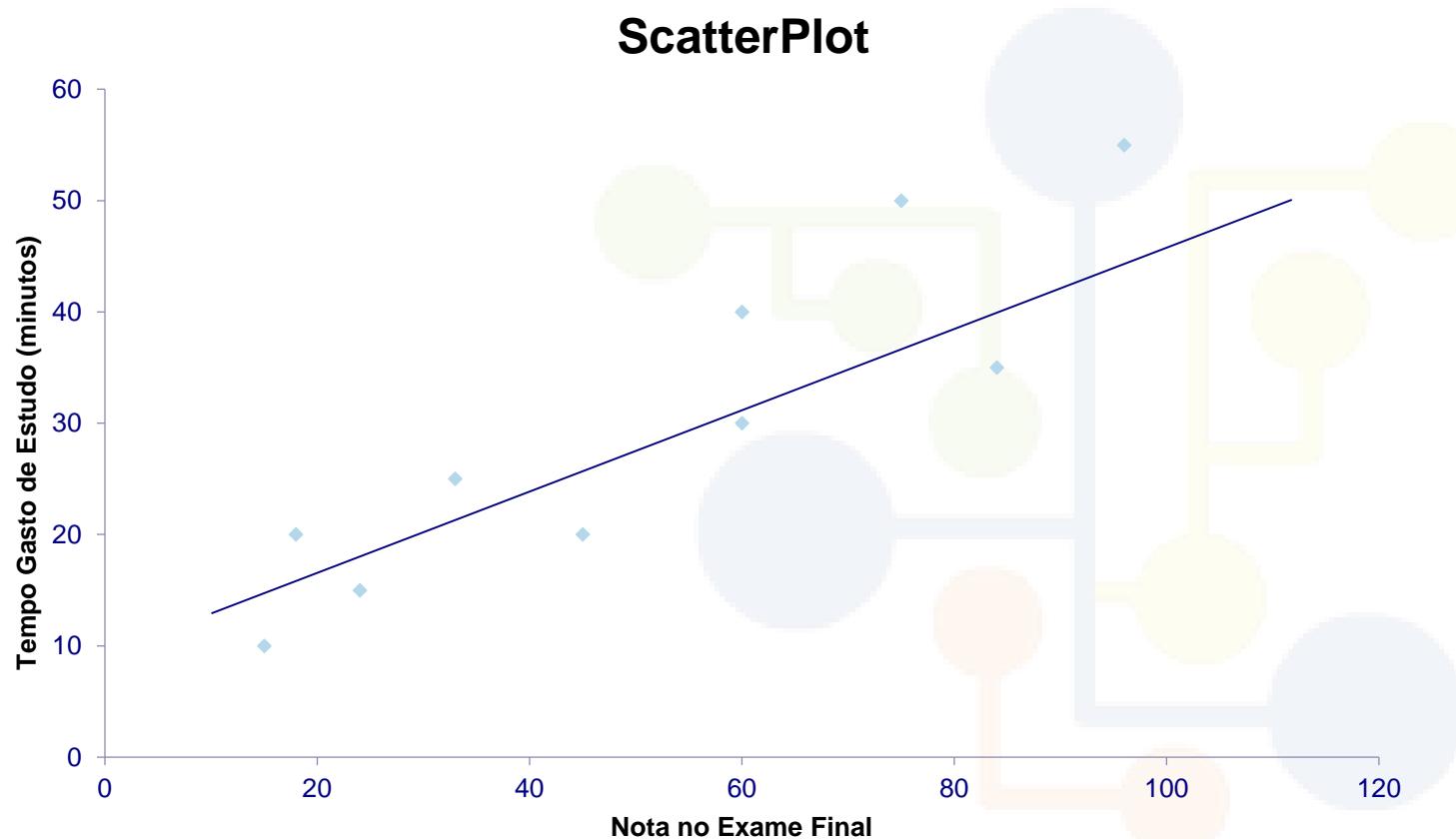
<i>95% inferiores</i>	<i>95% superiores</i>	<i>Inferior 95.0%</i>	<i>Superior 95.0%</i>
-4.57318773	16.32994449	-4.57318773	16.32994449
0.291953778	0.653992168	0.291953778	0.653992168



Data Science Academy



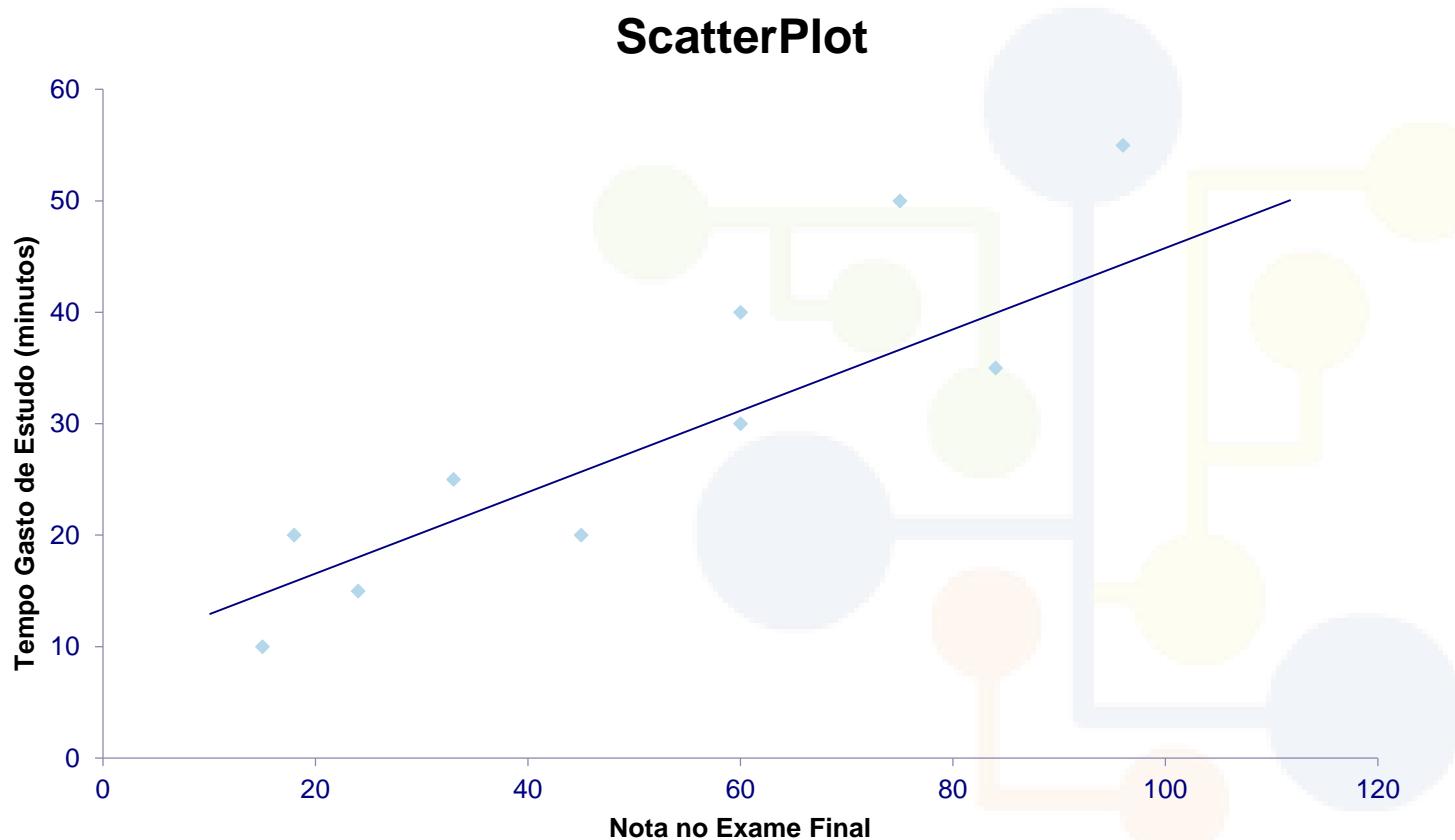
Data Science Academy



Podemos notar que, por mais que o parâmetro α seja positivo e matematicamente igual a 5,8783, não podemos afirmar que ele seja estatisticamente diferente de zero para esta pequena amostra, uma vez que o intervalo de confiança contém o intercepto igual a zero (origem). Uma amostra maior poderia resolver este problema.



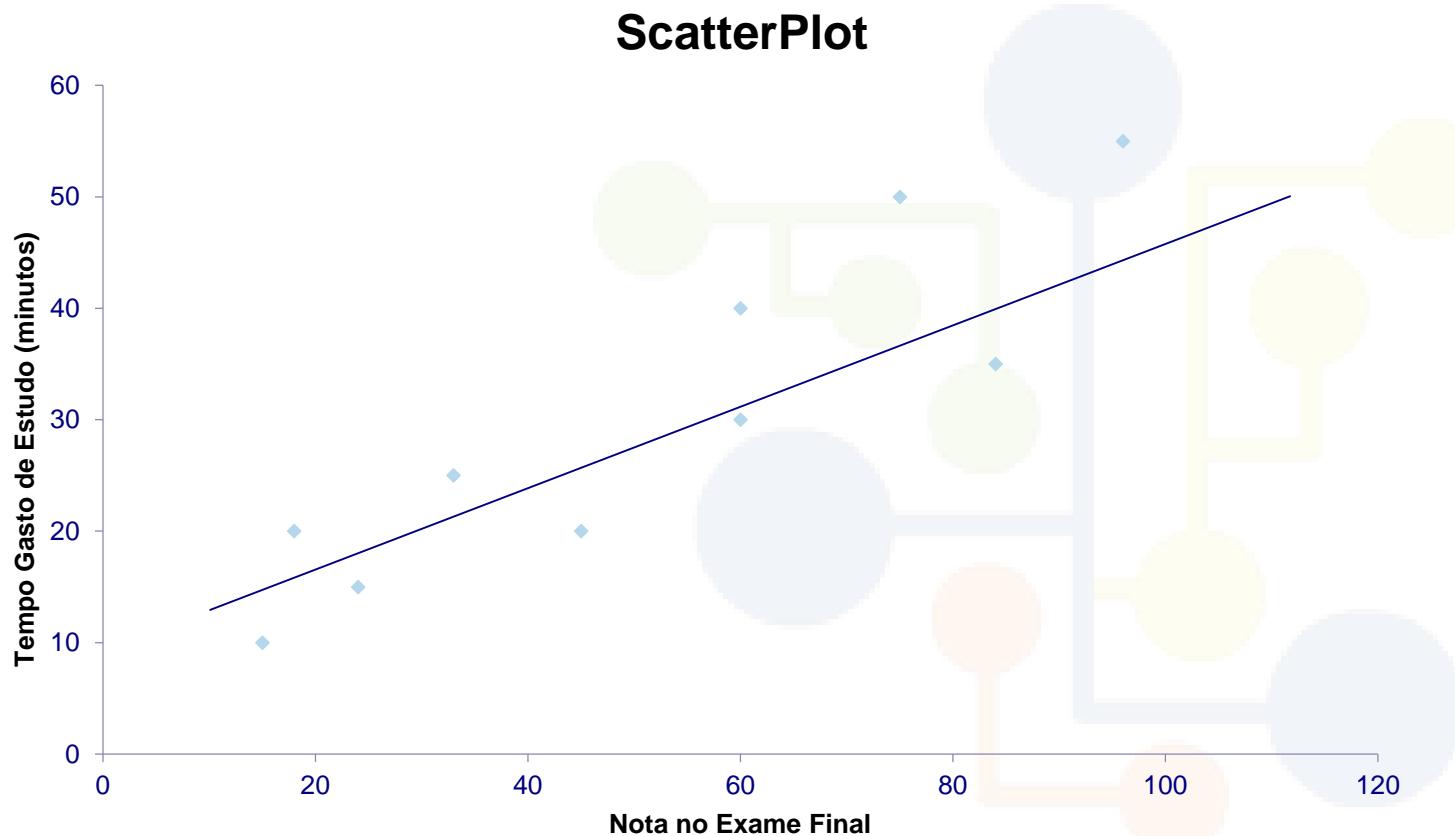
Data Science Academy



Já para o parâmetro β , podemos notar que a inclinação tem sido sempre positiva, com valor médio calculado matematicamente e igual a 0.4729. Podemos visualmente notar que seu intervalo de confiança não contém a inclinação igual a zero.



Data Science Academy

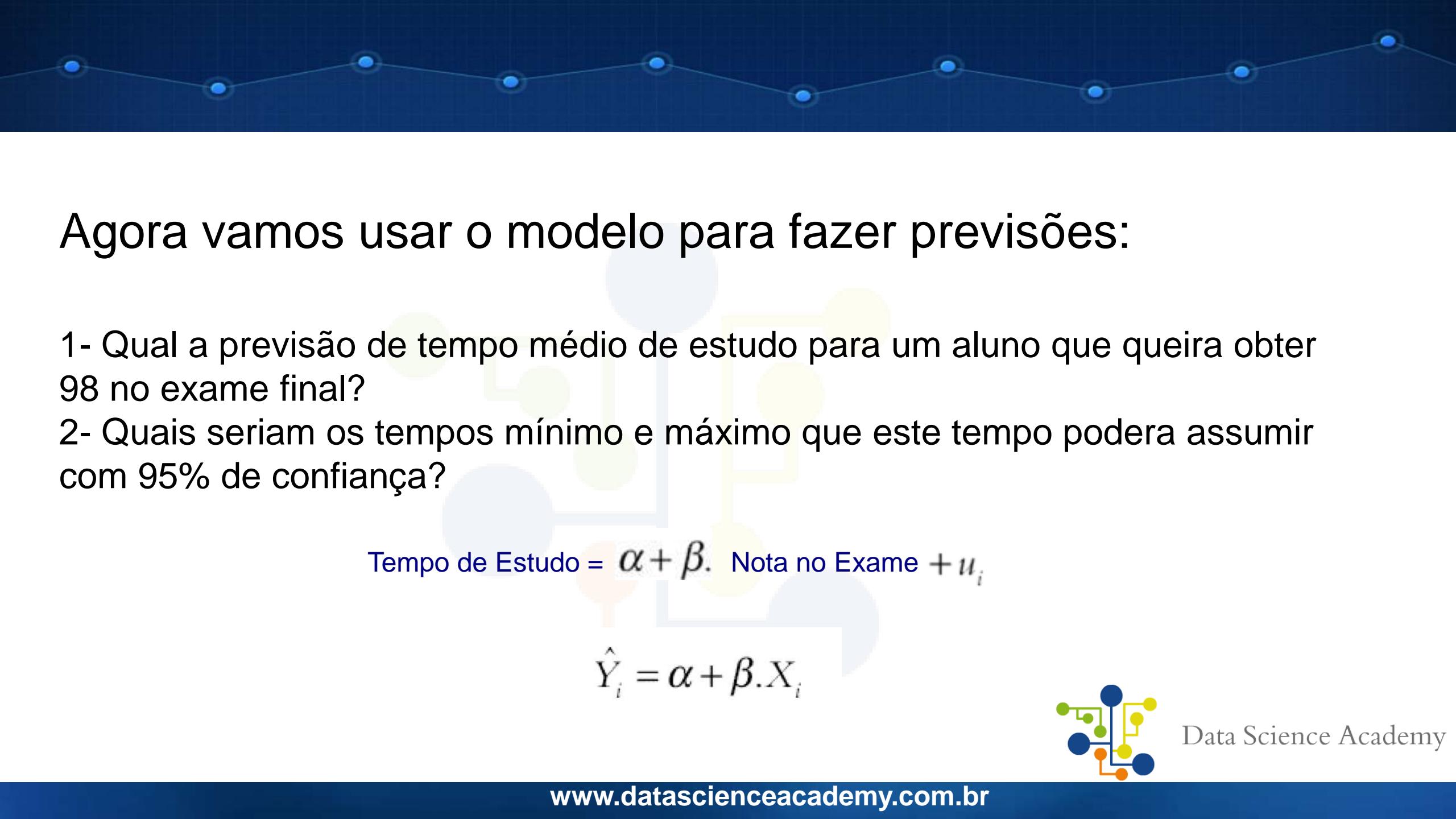


	<i>Coeficientes</i>
Interseção	5.878378378
Variável X 1	0.472972973

Conforme já discutido, a rejeição da hipótese nula para o parâmetro β , a um determinado nível de significância, indica que a correspondente variável X correlaciona-se com a variável Y e, consequentemente, deve permanecer no modelo final.



Data Science Academy



Agora vamos usar o modelo para fazer previsões:

- 1- Qual a previsão de tempo médio de estudo para um aluno que queira obter 98 no exame final?
- 2- Quais seriam os tempos mínimo e máximo que este tempo poderá assumir com 95% de confiança?

$$\text{Tempo de Estudo} = \alpha + \beta \cdot \text{Nota no Exame} + u_i$$

$$\hat{Y}_i = \alpha + \beta \cdot X_i$$



Data Science Academy



Agora vamos usar o modelo para fazer previsões:

1- Qual a previsão de tempo médio de estudo para um aluno que queira obter 98 no exame final?

$$\text{Tempo de Estudo} = \alpha + \beta \cdot \text{Nota no Exame} + u_i$$

$$\text{Tempo de Estudo} = 5.8783 + 0.4729 \times 98$$

$$\text{Tempo de Estudo} = 52.22$$

	<i>Coeficientes</i>
Interseção	5.878378378
Variável X 1	0.472972973

Tempo de Estudo = 52.22 minutos



Data Science Academy



Agora vamos usar o modelo para fazer previsões:

2- Quais seriam os tempos mínimo e máximo que este tempo podera assumir com 95% de confiança?

$$\text{Tempo de Estudo Mínimo} = a + b \cdot (98)$$

$$\text{Tempo de Estudo Mínimo} = -4.5732 + 0.2919 \times 98$$

$$\text{Tempo de Estudo Mínimo} = 24.03$$

95% inferiores

$$-4.57318773$$

$$0.291953778$$



Data Science Academy



Agora vamos usar o modelo para fazer previsões:

2- Quais seriam os tempos mínimo e máximo que este tempo podera assumir com 95% de confiança?

$$\text{Tempo de Estudo M\'aximo} = a + b \cdot (98)$$

$$\text{Tempo de Estudo M\'aximo} = 16.3299 + 0.6539 \times 98$$

$$\text{Tempo de Estudo M\'aximo} = 80.42$$

<u>95% superiores</u>
16.32994449
0.653992168



Data Science Academy



Agora vamos usar o modelo para fazer previsões:

2- Quais seriam os tempos mínimo e máximo que este tempo podera assumir com 95% de confiança?

$$\text{Tempo de Estudo Mínimo} = a + b \cdot (98)$$

$$\text{Tempo de Estudo Mínimo} = -4.5732 + 0.2919 \times 98$$

$$\text{Tempo de Estudo Mínimo} = 24.03$$

95% inferiores

$$-4.57318773$$

$$0.291953778$$

$$\text{Tempo de Estudo Máximo} = a + b \cdot (98)$$

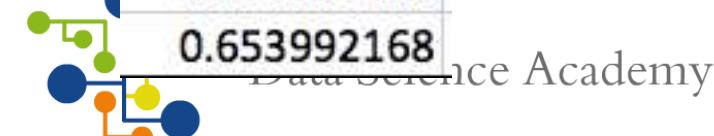
$$\text{Tempo de Estudo Máximo} = 16.3299 + 0.6539 \times 98$$

$$\text{Tempo de Estudo Máximo} = 80.42$$

95% superiores

$$16.32994449$$

$$0.653992168$$



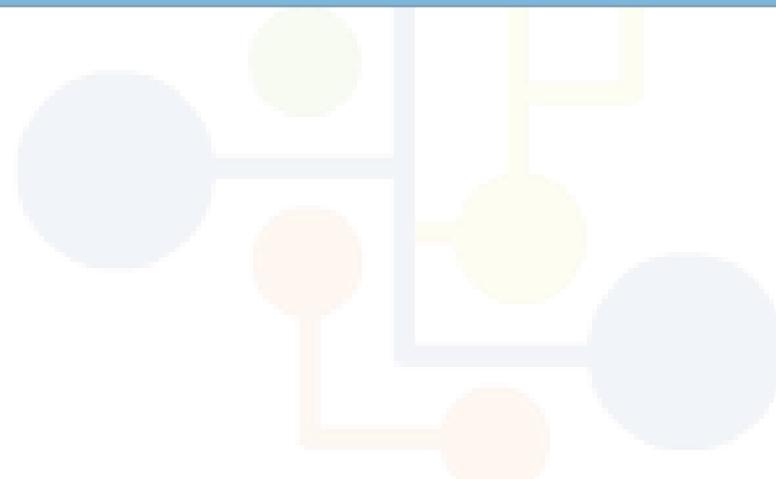
Esse tópico chegou ao final



Data Science Academy



Usando Regressão para Fazer Previsões



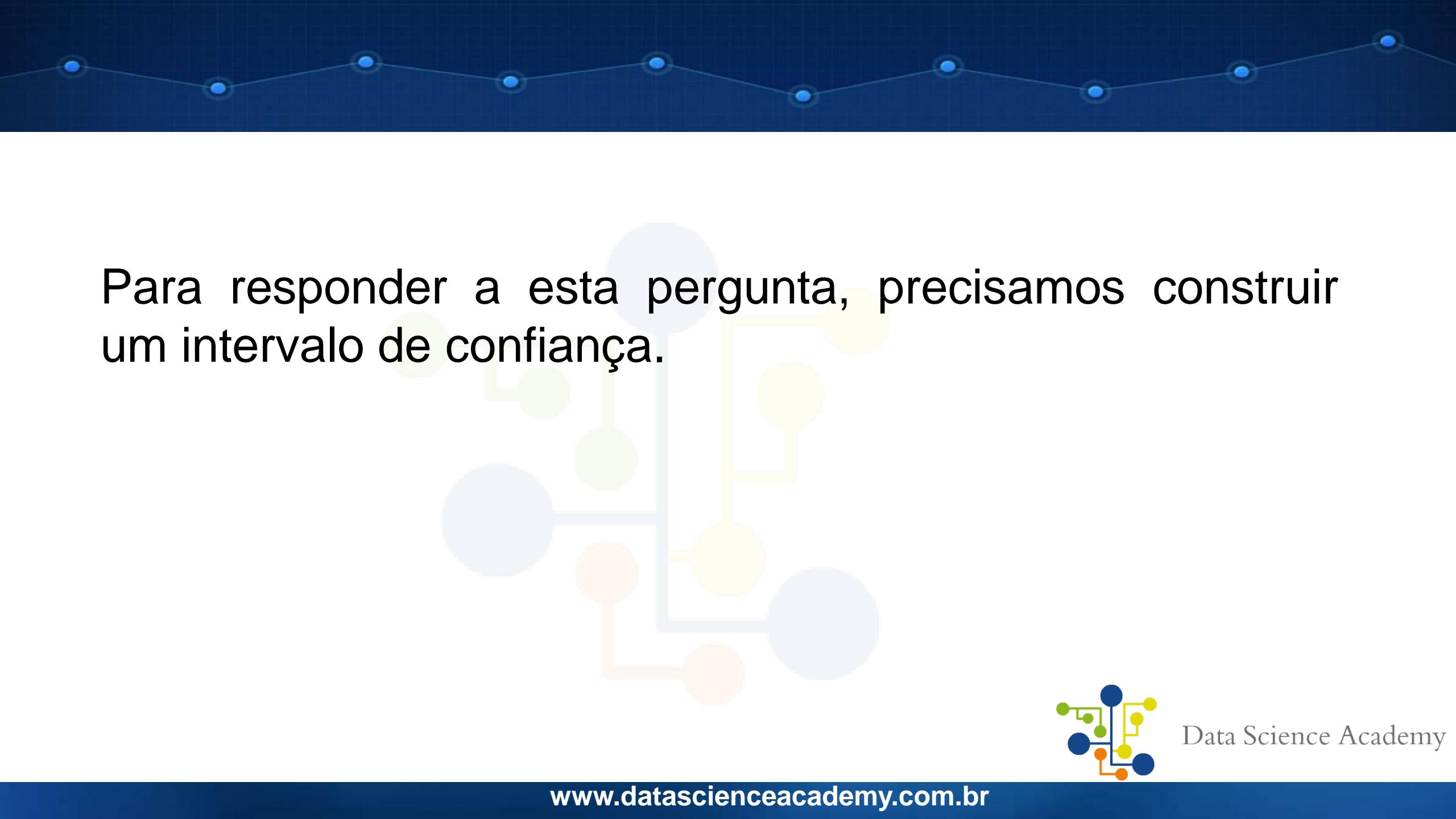
Data Science Academy



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?



Data Science Academy



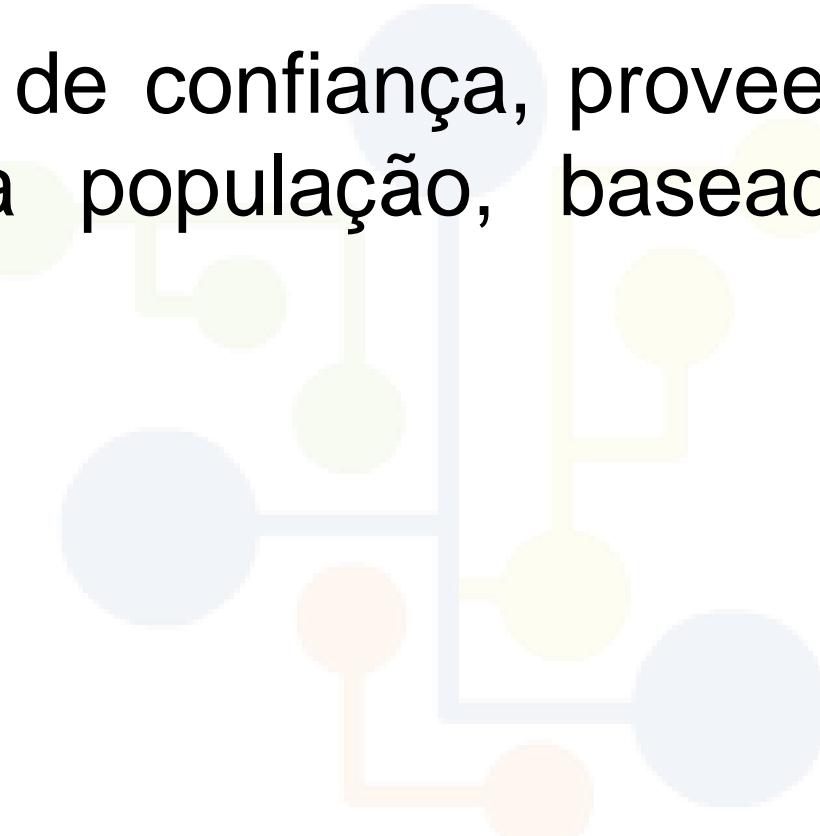
Para responder a esta pergunta, precisamos construir um intervalo de confiança.



Data Science Academy



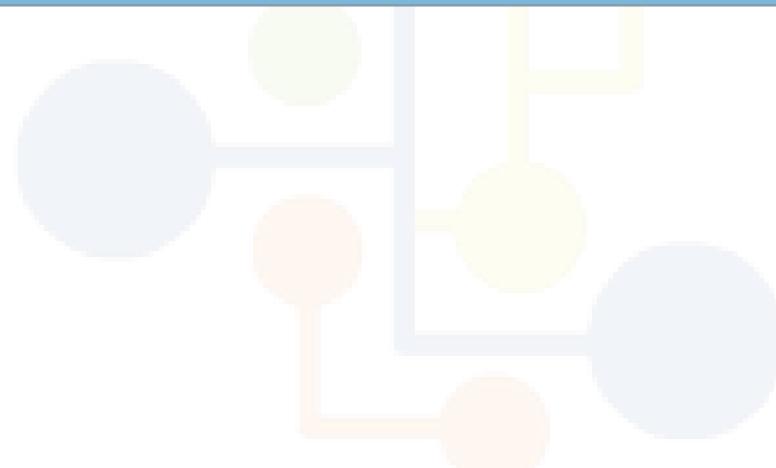
Os intervalos de confiança, proveem uma estimativa do parâmetro da população, baseado na estatística da amostra.



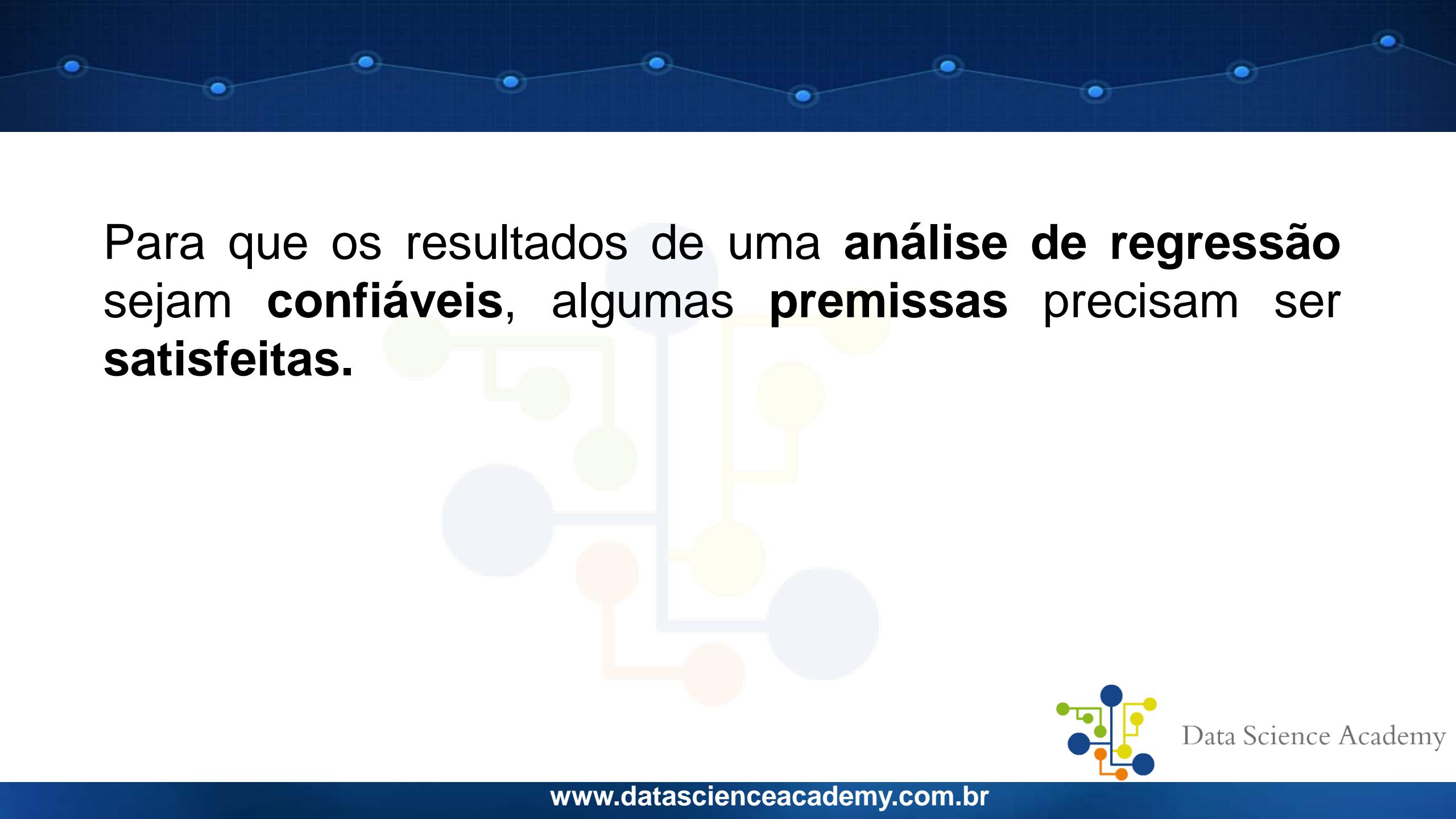
Data Science Academy



Premissas para Análise de Regressão



Data Science Academy



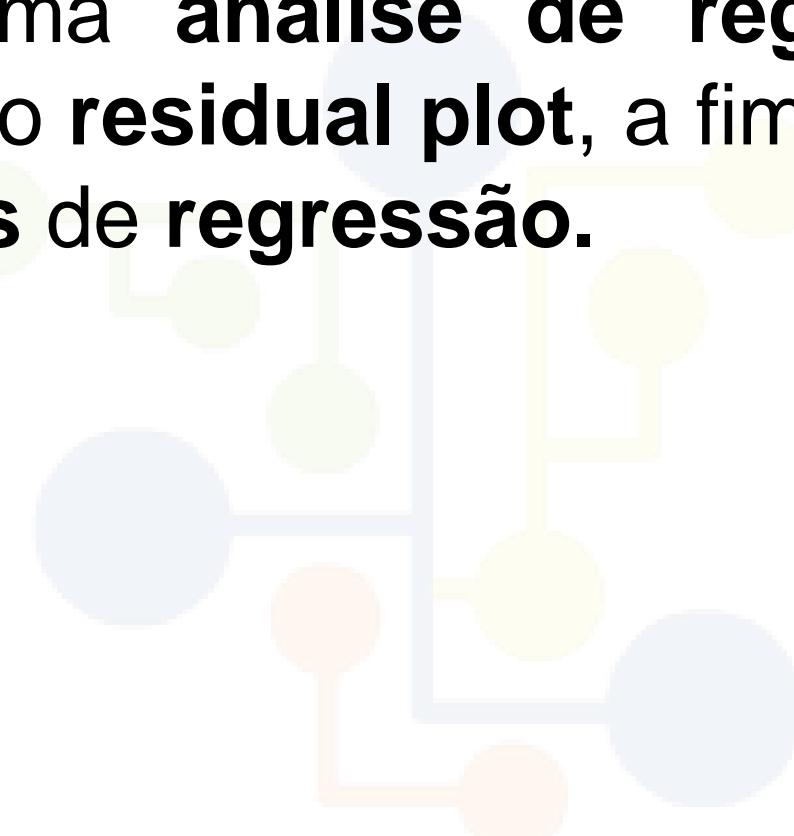
Para que os resultados de uma **análise de regressão** sejam **confiáveis**, algumas **premissas** precisam ser **satisfitas**.



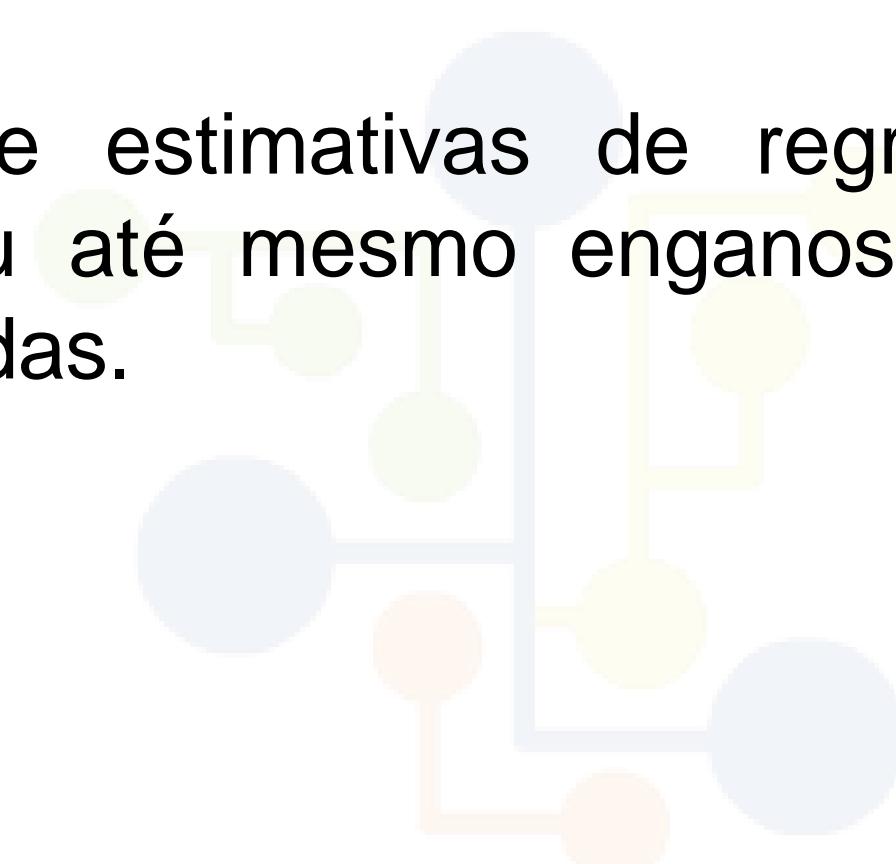
Data Science Academy



Ao realizar uma **análise de regressão** examinar o **scatter plot** e o **residual plot**, a fim de verificar **violação nas premissas de regressão**.



Data Science Academy



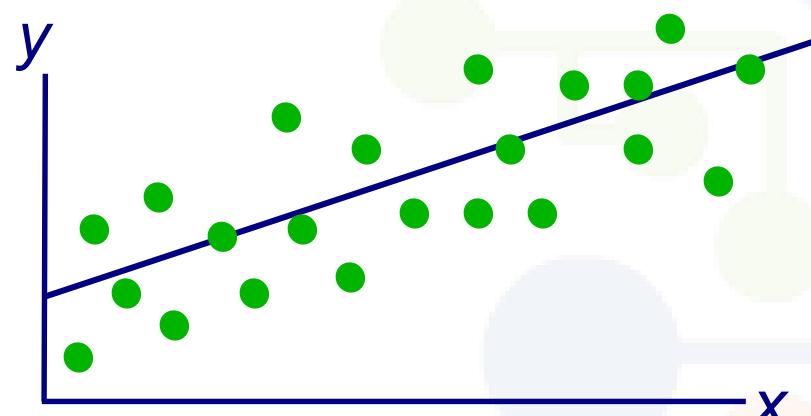
Previsões e estimativas de regressão serão menos precisas ou até mesmo enganosas, se as premissas forem violadas.



Data Science Academy

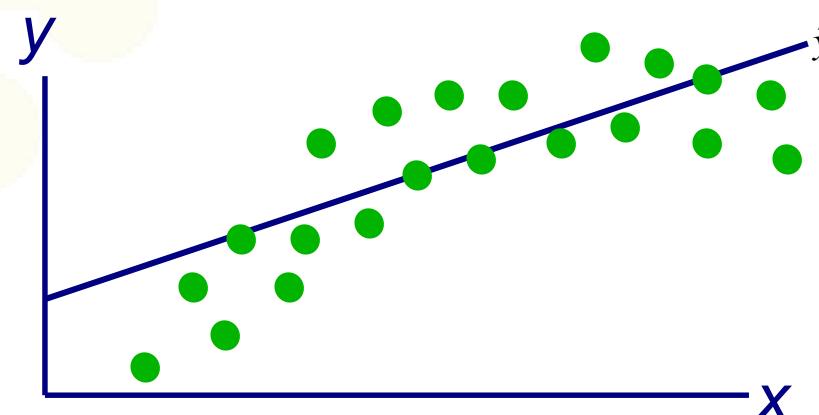
Premissa 1

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



Linear

Os dados seguem um padrão linear



Não Linear

Para x , o estimado \hat{y} é muito alto.
Para x_s , o estimado \hat{y} no centro é muito baixo



Data Science Academy

Premissa 2

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

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



Data Science Academy



Armadilhas da Análise de Regressão



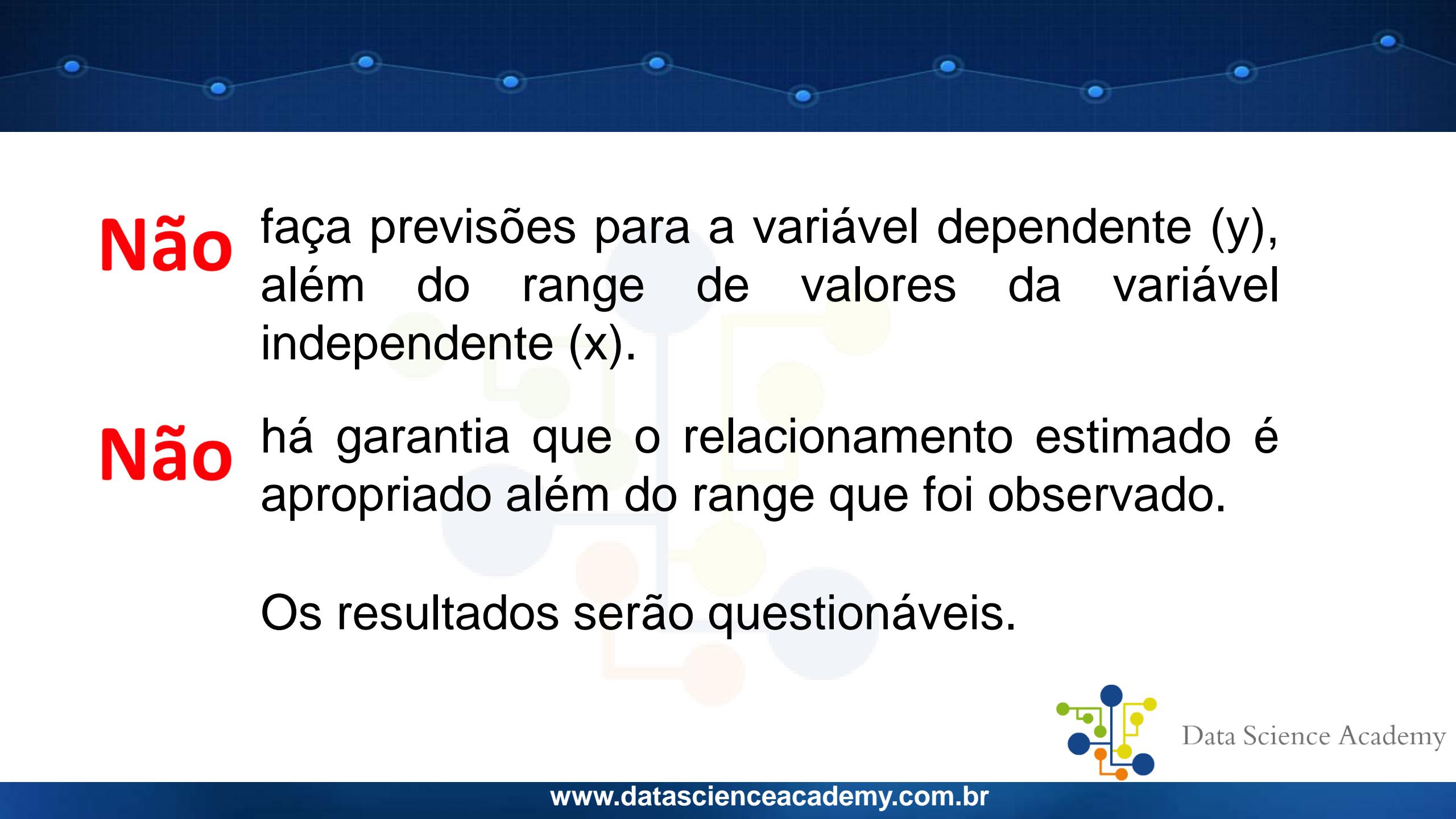
Data Science Academy



Atenção



Data Science Academy



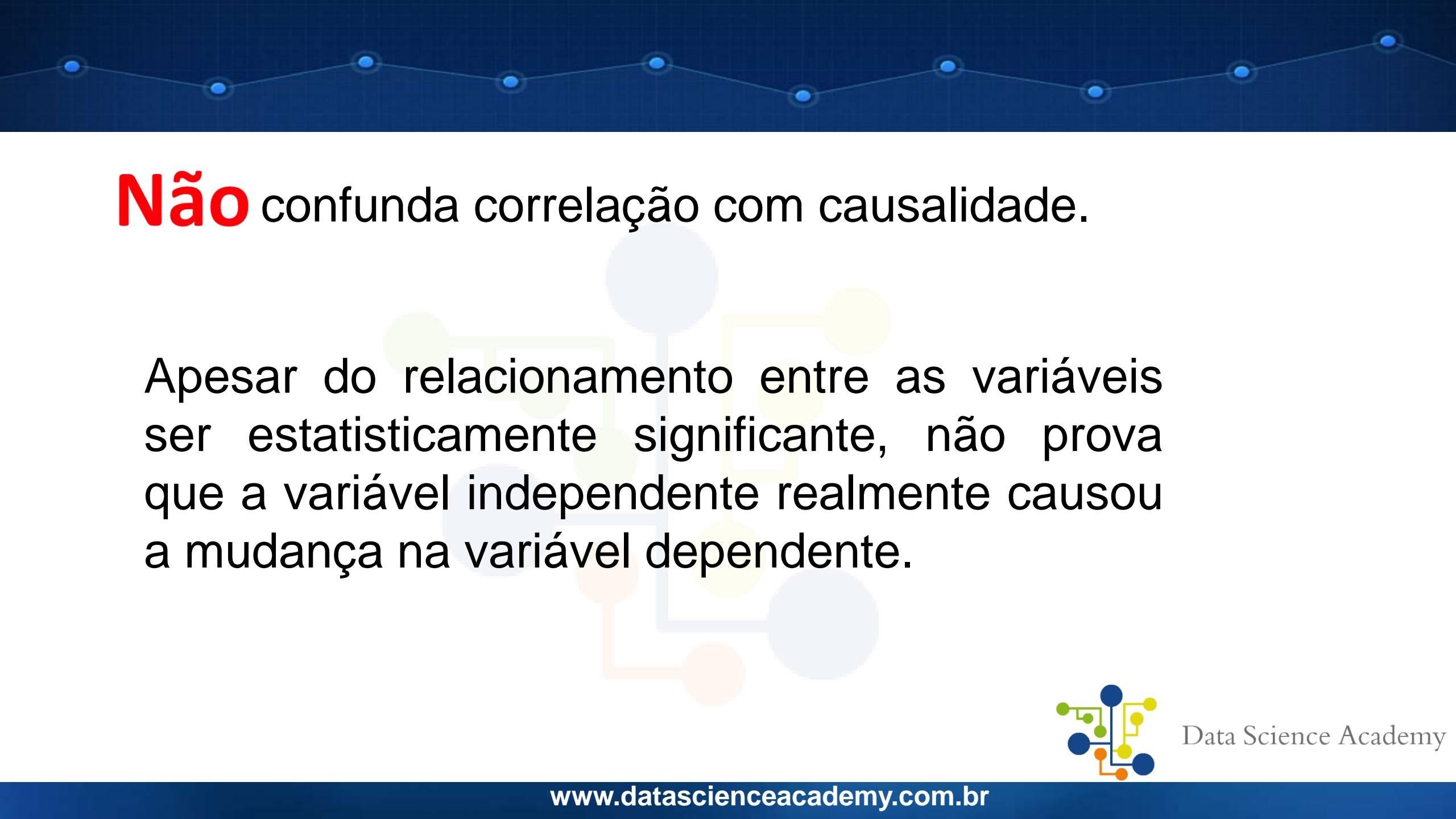
Não faça previsões para a variável dependente (y), além do range de valores da variável independente (x).

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

Os resultados serão questionáveis.



Data Science Academy



Não confunda correlação com causalidade.

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



Data Science Academy

Esse tópico chegou ao final



Data Science Academy

Curta Nossas Páginas nas Redes Sociais

E fique sabendo das novidades em Data Science, Big Data, Internet das Coisas e muito mais...



www.facebook.com/dsacademybr



twitter.com/dsacademybr



www.linkedin.com/company/data-science-academy

www.datascienceacademy.com.br



Data Science Academy



www.datascienceacademy.com.br

www.datascienceacademy.com.br



Data Science Academy