ESTATÍSTICA DESCRITIVA E DISTRIBUIÇÕES

Uma das coisas mais importantes em projetos de ciências de dados é fazer uma boa análise exploratória. Entender bem as informações que possui é primordial para saber que tipo de modelo irá criar e como você pode criar valor através de seu trabalho.

A Estatística Descritiva irá te fornecer uma série de técnicas e ferramentas para sumarizar um conjunto de dados. Com isso você irá entender e sintetizar os principais pontos que guiarão suas hipóteses, modelos e, principalmente, a maneira que você olha para um problema que está tentando resolver.

Em linhas gerais: **estatística nos ajuda a descobrir a simplicidade dentro da complexidade.**

Média, mediana e o desvio padrão são alguns termos da Estatística Descritiva que você já deve ter ouvido por aí.

TIPOS DE VARIÁVEIS

Para cada elemento ou observação de um conjunto de dados, tem-se associado uma série de resultados que correspondem às suas características ou algo que podemos usar para diferenciar cada indivíduo. Para cada uma dessas séries damos o nome de Variável.

Vamos começar com os tipos de variáveis que podemos encontrar. Dê uma olhada na tabela abaixo: cada linha é uma observação e cada coluna é uma variável. Cada uma dessas variáveis é de um tipo diferente. Visto isso, que diferenças você nota?

ESTATÍSTICA DESCRITIVA E DISTRIBUIÇÕES

Indivíduo B	SP	Superior completo	3	1,80
Indivíduo C	RJ	Superior incompleto	2	1,62
Indivíduo D	RJ	Mestrado	1	2,05
Indivíduo E	AM	Fundamental	0	1,68

As variáveis **qualitativas** podem ser:

- **Qualitativa nominal:** Quando a variável descreve a observação mas essas descrições não podem ser ordenadas, também chamada de categórica. Como, por exemplo, a naturalidade de uma pessoa, estado civil ou cor preferida. Não podemos dizer que Roxo é maior que Azul.
- **Qualitativa ordinal:** Categorias que podem ser ordenadas. Por exemplo, podemos dizer que uma pessoa com ensino Superior completo teve mais anos de estudo do que uma pessoa com Fundamental. Outro exemplo seria classificar a temperatura de um objeto como Frio, Morno e Quente.

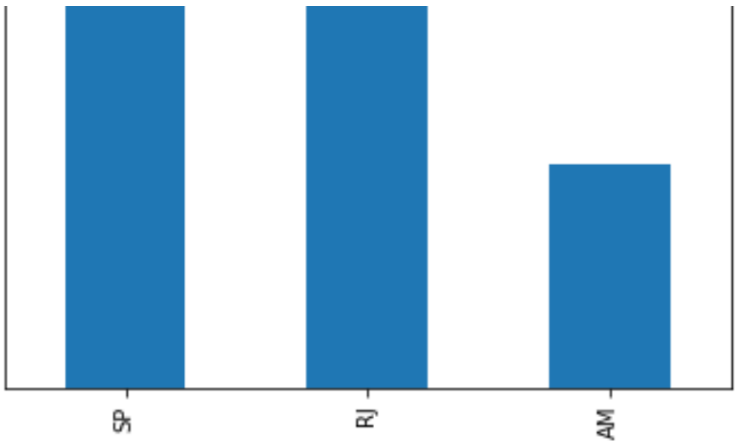
As variáveis **quantitativas** podem ser:

- **Quantitativa discreta:** Em geral são resultado de processos de contagem: Quantidade de smartphones, Número de quartos de uma casa, etc… Não temos uma pessoa com 1.28 smartphones, a pessoa salta de um smartphone para dois. Para essa característica damos o nome de intervalo discreto.
- **Quantitativa contínua:** São valores que varrem uma escala contínua de medição. Altura de uma pessoa, renda, peso de uma carga, etc… Entre uma pessoa com 1,72 de altura e outra de 1,73, por exemplo, existem infinitas possibilidades de medição.

A primeira coisa que podemos fazer para começar a entender melhor nossas variáveis é fazer uma contagem, ou seja, ver quantas vezes os valores se repetem ao longo das observações. Dessa forma fazemos uma consolidação dos números e começamos a dar sentido ao nosso conjunto de dados.

Vamos olhar para a variável Naturalidade da nossa tabela anterior, e contar quantas vezes os elementos se repetem: SP: 2, RJ 2, AM: 1. Podemos plotar um gráfico com esses dados:

ESTATÍSTICA DESCRITIVA E DISTRIBUIÇÕES



Esse é um **histograma**, também conhecido como distribuição de frequências.

AVANÇAR

ESTATÍSTICA DESCRITIVA E DISTRIBUIÇÕES

MEDIDAS DE TENDÊNCIA CENTRAL

Agora, imagine que alguém te chama para uma reunião urgente e pede para resumir em apenas **um número** um conjunto de dados de centenas de milhões de linhas que corresponde a altura de todos os brasileiros. O que você faria?

A melhor escolha nesse caso seria escolher alguma **medida de tendência central**. Que tende a resumir bem um conjunto, assumindo que tenha uma distribuição normal, mas veremos isso mais adiante…

Poderíamos escolher por exemplo a média, moda ou mediana, todas elas são medidas de tendência central. A utilização delas varia de acordo com a informação que você pretende passar. Dessas medidas, a mais famosa é a **Média**.

MÉDIA

Soma de todos os valores dividido pela quantidade de observações. Como cada observação tem o mesmo peso é uma medida que é sensível à valores extremos. Você já deve ter ouvido: *"Se colocar a cabeça no Forno e os pés no Freezer, na média está tudo bem"*.

Mas para uma distribuição com poucos valores discrepantes e não muito extremos, esse número irá te trazer uma boa representação de todos os seus dados.

Vamos ver um exemplo:

Turma A

1	3
2	9
3	5
4	7
5	5
Média	**?**

Qual a média de matemática da Turma A?

(3+9+5+7+5) / 5 = 5,8

A média da turma é de 5,8.

Veja que isso não nos diz se existe alguém que precisa de ajuda na matéria; a ideia da média é ter uma noção da turma como um todo.

Outra forma de olhar é a **Mediana.**

MEDIANA

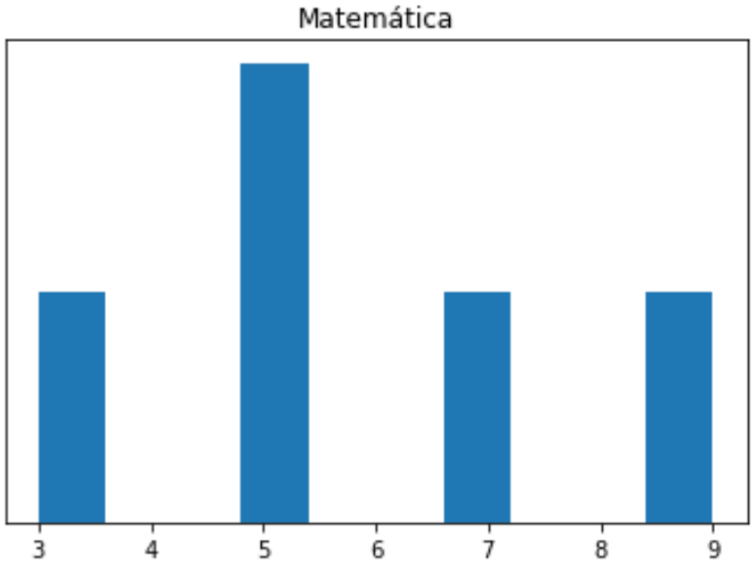
É o valor que separa a metade maior da metade menor de uma variável. Para isso vamos ordenar de maneira crescente as notas da Turma A: [3, 5, 5, 7, 9]. Veja que o terceiro elemento divide bem ao meio nossa amostra. Neste caso a **Mediana da Turma A é 5.**

A principal vantagem da Mediana é que ela é menos sensível a valores extremos (*outliers*).

Imagine uma sala com 3 pessoas. A primeira tem renda mensal de R$1.000, a segunda tem renda de R$2.000 e a terceira de R$1.000.000.

A média de renda de pessoas dessa sala é de: **R$334.333**. Mas a mediana é de apenas **R$2.000.**

também posso enxergar uma medida de centralidade. Por exemplo, ao pensar no histograma, será a barra mais alta. Essa medida é chamada de **Moda**. No caso das notas de matemática da Turma A, qual a moda?



Olhando para este gráfico podemos ver claramente que a moda é **5.**

Portanto para nossa Turma A nós temos: Média 5,8 ; Mediana 5 e Moda 5.

AVANÇAR

ESTATÍSTICA DESCRITIVA E DISTRIBUIÇÕES

MEDIDAS DE DISPERSÃO

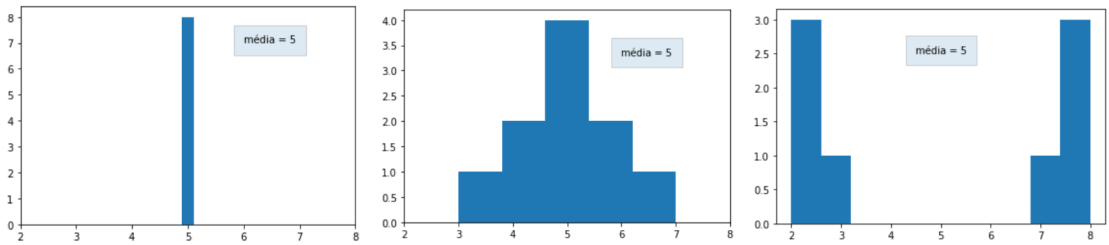
Continuando na nossa investigação estatística a próxima coisa a se observar na suas variáveis é como ela está distribuída. Já começamos a ter uma leve noção quando olhamos para os histogramas, mas agora vamos aprofundar essa noção.

Uma **medida de dispersão** para uma **variável quantitativa** é um indicador do grau de espalhamento dos valores da amostra em torno da **medida de centralidade**.

Bem legal como todos esses conceitos que aprendemos já estão conectados!

Podemos também entender da seguinte forma: A medida de dispersão é como uma as observações se espalham tendo como referência um ponto central, como a média.

Os três gráficos abaixo mostram uma variável com média 5. Mas está claro que são bem diferentes. É aí que as medidas de dispersão vão nos ajudar.



Duas medidas que fazem isso que foi dito acima são a **variância** e o **desvio padrão**.

A variância e o desvio padrão mostram como a variável se distribui em relação à média. Quanto maior forem, mais distantes da média estarão as observações.

Para medirmos a variância nós precisamos somar o quadrado da subtração da média de todos os valores e depois dividir pela quantidade. No exemplo

ESTATÍSTICA DESCRITIVA E DISTRIBUIÇÕES

A variância das notas de Matemática é de 4,16.

*Obs: Quando nós temos todos os valores referente ao nosso objeto de estudo, chamamos isso de **população**, mas na maioria das vezes trabalhamos com uma **amostra**, ou seja, com um recorte da população. Quanto maior nossa amostra e quanto mais amostras temos, mais nós nos aproximamos da representação da população.*

No cálculo da variância, divide-se por ***n*** quando estamos trabalhando com população e dividimos por ***n-1*** quando é uma amostra.

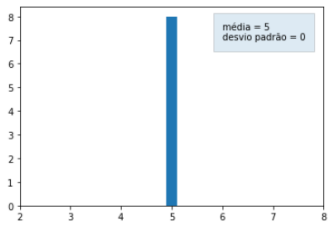
Observe que nós elevamos todas as observações ao quadrado. Se não tivéssemos feito isso essa soma teria dado zero, pois os números abaixo da média iriam "cancelar" os número acima da média.

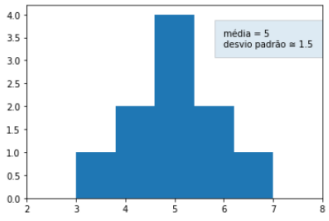
Porém outro efeito que isso causa é que não podemos comparar diretamente esse número com a nossa distribuição, por isso calculamos a raiz quadrada da variância e então temos o **desvio padrão**.

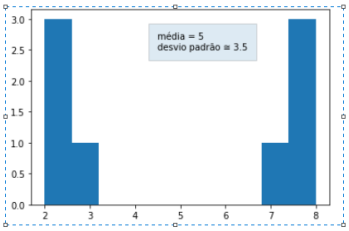
No exemplo das notas de matemática temos uma variância de 4,16 que corresponde a um desvio padrão de: **2.04**

Então, o desvio padrão é o quanto, em média, os valores se desviam da média desse conjunto.

Vamos ver aqueles mesmos gráficos com médias iguais agora com o valor do desvio padrão.







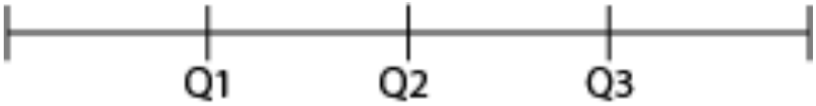
AVANÇAR

ESTATÍSTICA DESCRITIVA E DISTRIBUIÇÕES

MEDIDAS DE POSIÇÃO

Lembra da mediana? Ela reparte nosso conjunto de dados em duas partes com quantidades iguais de elementos. Por isso também chamamos ela de **separatriz**.

Podemos também dividir nossos dados em 4 partes, veja na imagem abaixo:

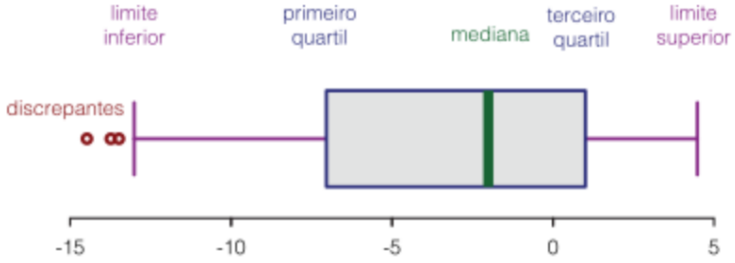


Depois de ordenar os valores você irá dividir os seus dados em 4 partes (**Quartis**).

O primeiro ponto de divisão é chamado de Q1 e 25% dos dados são menores ou iguais a ele. O segundo ponto é o Q2 e 50% dos dados são menores ou iguais a ele (repare que o Q2 é também a mediana). E, por último, temos o Q3, no qual 75% dos dados são menores ou iguais a ele.

Obs: Podemos também dividir em 100 partes (percentis) e aplicar a mesma lógica que nos quartis.

Uma representação gráfica muito legal é o ***boxplot***, também conhecida como diagrama de caixa, que apresenta a distribuição da sua variável através dos seus quartis. Ainda te mostra os valores discrepantes (outliers), fornecendo assim um meio complementar para desenvolver uma perspectiva sobre os seus dados.



O tamanho da caixa é também chamado de intervalo interquartil (IIQ). Você encontra esse valor fazendo Q3 - Q1.

Os valores que estão acima do limite inferior ou acima do limite superior são os valores discrepantes e mais comumente chamados de ***outliers***.

AVANÇAR

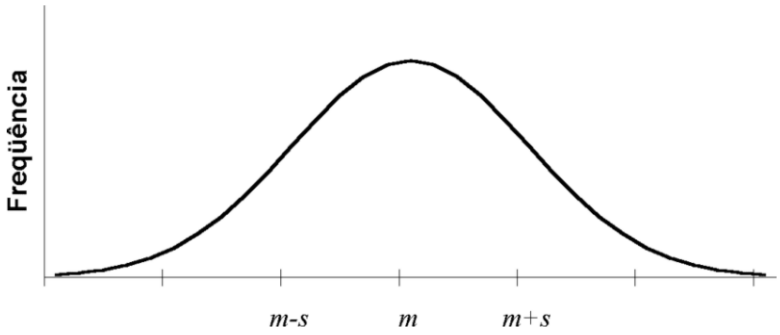
DISTRIBUIÇÕES

Quando olhamos para os histogramas nós vimos como a frequência dos dados se distribui. Agora imagine que esses gráficos são uma pequena amostra de todo o dado que seja possível coletar. Quando generalizamos essas frequências nós temos as distribuições. Uma maneira de entender uma distribuição é olhar pela perspectiva da probabilidade. Quanto mais perto do pico está um valor maior a probabilidade dele aparecer e quanto mais perto das caudas mais raro é o evento. Entender os diferentes tipos de distribuições te facilitará quando precisar inferir características dos seus dados isso será possível tendo uma distribuição com que comparar.

Vamos aprender mais sobre?

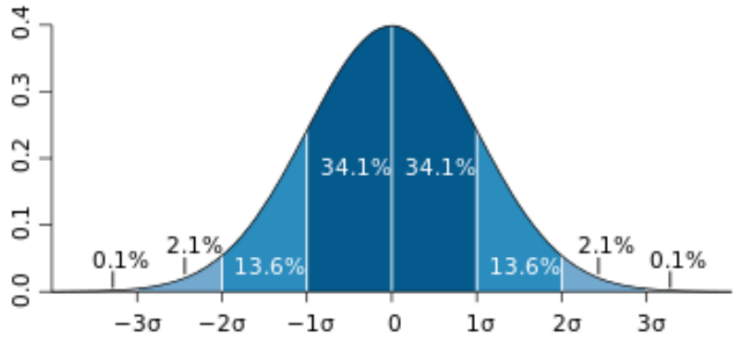
DISTRIBUIÇÃO NORMAL

A distribuição mais conhecida é a **distribuição normal**, ela tem esse formato de sino. Geralmente encontramos ela em eventos naturais e aleatórios. Como por exemplo altura média de uma população: é normal termos uma grande concentração de pessoas com altura em torno da média e pessoas com uma estatura muito baixa ou muito alta são mais difíceis de acontecer, por isso estão nas caudas da distribuição.



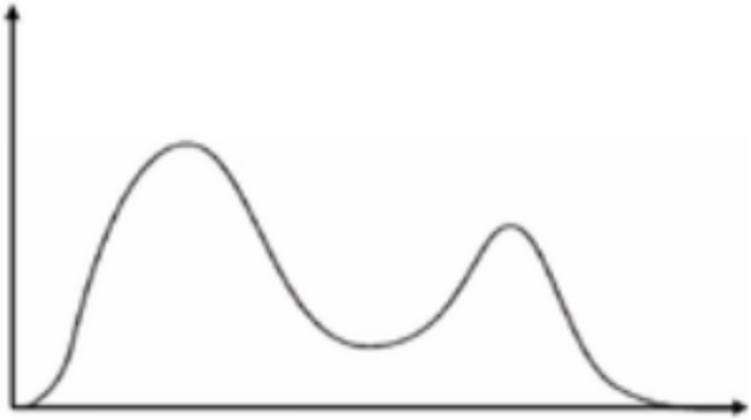
Então quando ouvir que se está assumindo que a variável tem uma distribuição normal, você pode entender que os dados se distribuem conforme o gráfico acima. Assumir que uma distribuição é normal também nos traz muitas vantagens, pois é uma distribuição simétrica (o lado esquerdo da média é igual ao lado direito) e já temos em tabelas

ESTATÍSTICA DESCRITIVA E DISTRIBUIÇÕES



Podemos, por exemplo, assumir que 68% dos nossos dados estão concentrados entre um desvio padrão abaixo da média e um desvio padrão acima da média.

Pode ser que você se depare com um gráfico com dois picos, veja abaixo:



Essa é uma distribuição bimodal, em que a distribuição apresenta dois pontos de alta concentração. Pode ser que de fato uma distribuição seja assim, mas uma investigação mais criteriosa pode mostrar que na verdade existem duas distribuições unimodais juntas. Um bom exemplo seria o desempenho na corrida de São Silvestre, sem separar os amadores dos atletas profissionais.

Vamos ver abaixo outros tipos de distribuição:

DISTRIBUIÇÃO T DE STUDENT

Essa distribuição tem um formato quase igual à normal, só que com caudas mais largas, ou seja, os valores mais extremos são menos raros nessa distribuição.

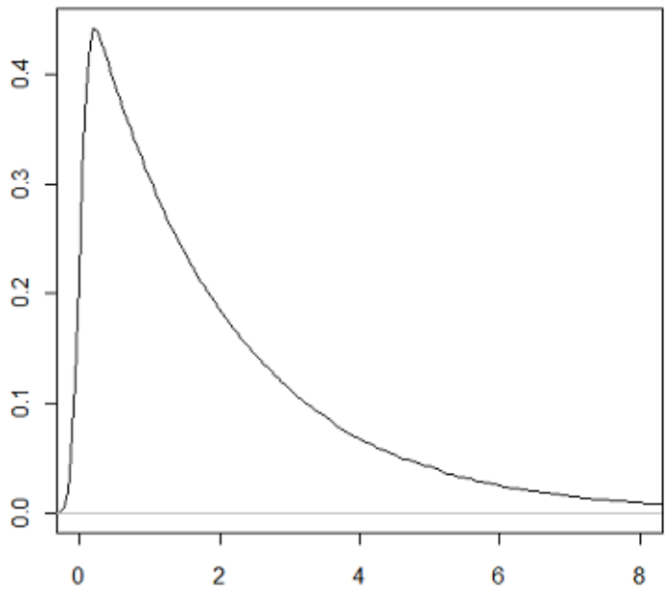
Por exemplo: Imagine que você está fazendo uma análise laboratorial de coletas de sangue de um grupo de estudo e precisa ver se a média da concentração de certo elemento seria igual ao de uma população controle. Seu grupo é composto por 5 pessoas, ou seja, seu n amostral é pequeno e não pode fazer mais coletas. Neste caso você iria fazer um estudo usando a distribuição t de Student.

DISTRIBUIÇÃO QUI-QUADRADO

Como você deve imaginar nem toda distribuição vai ter o formato de uma normal.

A probabilidade da distribuição qui quadrado não é simétrica como a da distribuição normal, essa simetria vai variar de acordo um parâmetro que pode ser passado para a distribuição por isso essa distribuição é usada para verificar o quanto nossa amostra se encaixa em uma certa distribuição.

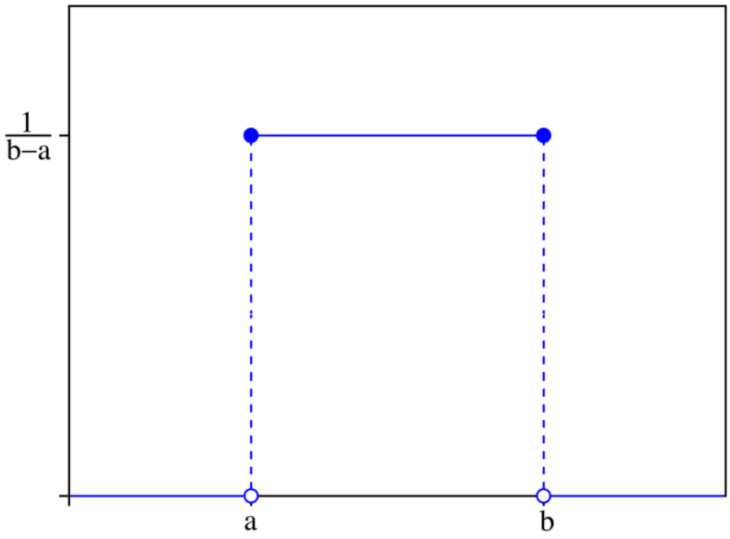
As visualizações de um vídeo novo no Youtube por exemplo: logo que ele é lançado tem muitas visualizações, seja pelo número de inscritos ou porque o tema é mais atual, e logo ele vai perdendo a relevância ou o algoritmo do Youtube já não vai mais recomendá-lo tanto.



DISTRIBUIÇÃO UNIFORME

Imagine que você tem um dado normal, sem truques, de seis lados. Qual a probabilidade de cair 1? e 2? A probabilidade para todos os números é a

Ela vai ter este formato:



Loteria é um exemplo que muitos gostariam que não fosse uma distribuição uniforme, mas é. Pense na Mega-sena: qualquer número de 1 a 60 tem a mesma probabilidade de ser sorteada do que outro. E todo jogo de 6 números tem a mesma probabilidade de sair do que qualquer outro jogo de 6 números. Por mais que você pense que 1-2-3-4-5-6 nunca será sorteado. 😂

OUTRAS DISTRIBUIÇÕES

Existem muitas outras distribuições, mas com essas você já vai entender como as variáveis podem se comportar dentro do seu conjunto de dados.

AVANÇAR

APLICAÇÕES E CONCLUSÃO

APLICAÇÕES

Esse vídeo mostra como é importante a análise de dados em qualquer projeto de Ciência de Dados. Não importa o quão avançado seja o modelo, essa etapa é primordial para todo projeto.

https://www.youtube.com/watch?v=YFvtZfHaPR8

MATERIAIS PARA APROFUNDAMENTO

Estatística é um mundo vasto e com tantos conceitos e aplicações que existem graduação e especializações somente para seu estudo. Por isso deixo abaixo alguns materiais que vão te ajudar neste aprofundamento.

Série de vídeos do Crash Course sobre estatística. São Excelentes e muito visuais!

Curso de estatística da PennState University

Welcome to STAT 200! | STAT 200

StatQuest| Diversos vídeos que explicam os conceitos estatísticos

ESTATÍSTICA DESCRITIVA E DISTRIBUIÇÕES

O QUE ACHOU DESTA AULA?

Deixe seu feedback para continuarmos melhorando sua experiência.

1 MIN

AVALIAR

VOLTAR PARA O CURSO