Estatística Descritiva

Mentora: Beatriz Yumi

Para começar: preciso saber estatística?

É fundamental aprender estatística em seu nível conceitual. Isso significa que não é preciso conseguir reproduzir sozinho os cálculos matemáticos, mas aprender a usar a estatística como ferramenta de abstração dos conteúdos complexos que abordamos na análise de dados.

É através desta ferramenta que conseguimos coletar, analisar, interpretar e apresentar os dados com os quais estamos trabalhando.

Neste módulo, vamos focar na estatística descritiva, onde vamos entender como descrever e sumarizar nossos dados.

Essas habilidades vão nos permitir fazer análises ricas e conseguir interpretar os resultados que são trazidos até nós com maior embasamento teórico e confiança.

Tipos de variáveis

As variáveis podem descrever uma qualidade ou atributo do que se está analisando. Neste caso, são chamadas de **variáveis qualitativas**.

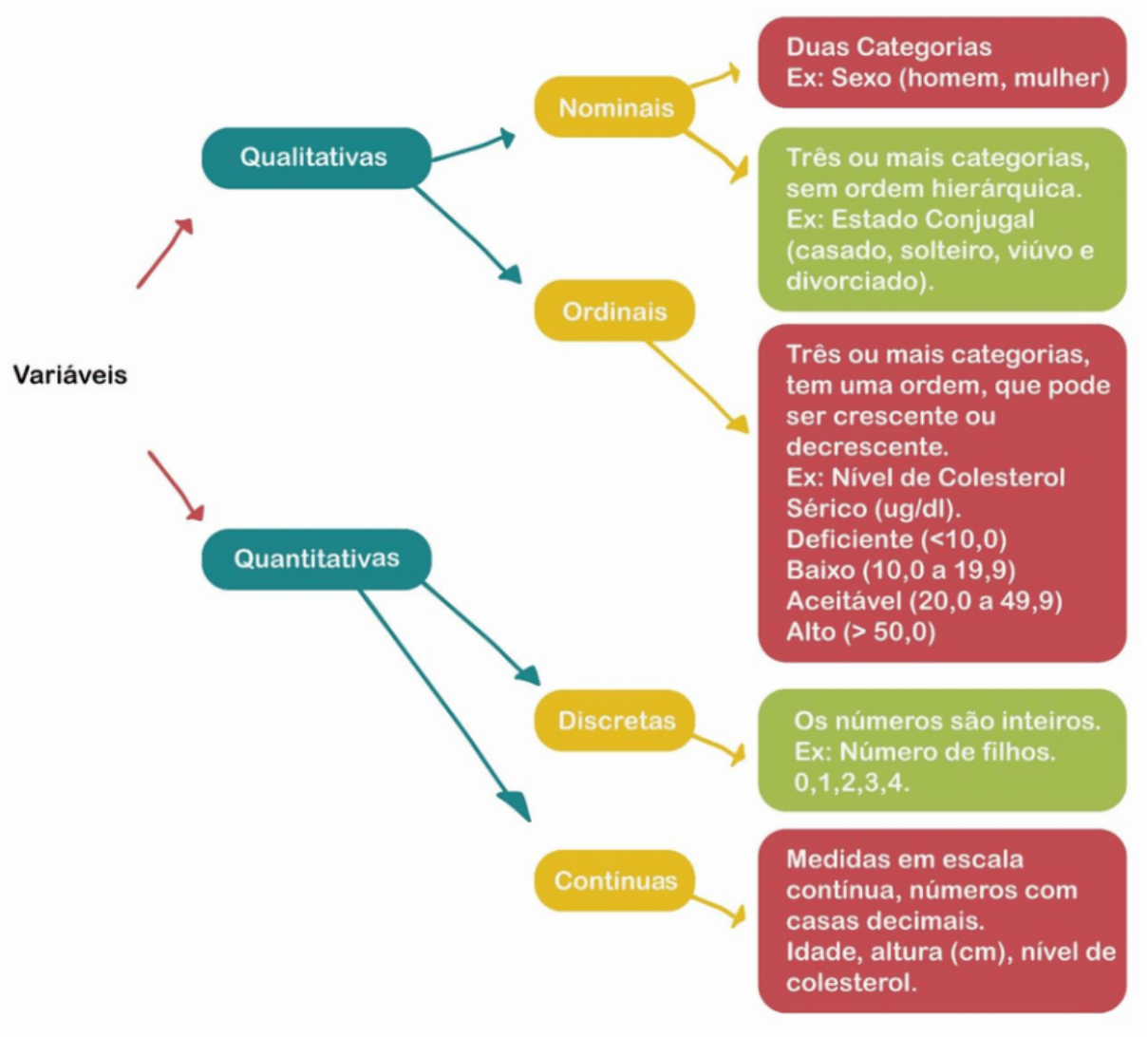
Se, neste tipo de atributo existir alguma ordem, as variáveis são chamadas de **ordinal** - por exemplo, quando estamos analisando o grau de instrução de um grupo de indivíduos - pois existe uma ordem que começa no ensino fundamental, passa pelo ensino médio e vai até o superior.

Outro exemplo seria a classe social, onde podemos dividi-la em baixa, média e alta.

Já quando não há uma ordem no atributo, chamamos estas variáveis de **nominais**. Um exemplo é o gênero dos indivíduos, outro seria seu estado civil.

As variáveis também podem representar números resultantes de uma contagem ou mensuração, sendo então chamadas de **variáveis quantitativas**.

Quando resultam de um processo de contagem, chamamos de **discretas**, como quantidade de filhos ou quantos veículos o indivíduo possui. Quando resultantes de um processo de mensuração, chamamos de **contínuas**, como estatura e peso de um indivíduo.



Após entender os diferentes tipos de dados, no próximo bloco conheceremos diferentes formas de medir nossos dados e estudaremos os usos dos histogramas, que são a representação gráfica de um conjunto de dados.

Média, mediana (ou separatriz) e moda

Estes três conceitos servem quando queremos representar um conjunto de dados em apenas um valor. Vamos entender melhor cada um deles?

Média

Essa, de todas, é a medida mais conhecida. É o resultado da soma das observações dividida pelo número delas. Devemos ter muito cuidado ao utilizar a média, pois, com a ocorrência de *outliers* a média pode conduzir a erros de interpretação.

Por exemplo, se quase todos os integrantes de um grupos tem entre 20 a 30 anos, mas dois possuem 90, a média pode levar a entender que os integrantes são mais velhos do que realmente são.

Mediana ou Separatriz

É a posição central da série de observações, quando ordenadas em ordem crescente. Na maioria das situações é a medida mais adequada por ser resistente a mudanças de uma pequena porção de dados e por não se afetar por *outliers*.

Moda

É a realização mais frequente de um conjunto de valores observados.

Medidas de dispersão: Variância e desvio padrão

Para conseguir entender como ocorre a dispersão dos dados em torno de sua média, temos duas medidas que são muito utilizadas: o **desvio padrão** e a **variância**.

Basicamente, o que fazemos é verificar como se desviam as observações em relação à média dessas mesmas observações.

Fórmula da variância:

$$V = \sum_{i=1}^{n} \frac{(x_i - M_a)^2}{n}$$

Sendo,

∑: símbolo de somatório. Indica que temos que somar todos os termos, desde a primeira posição (i=1) até a posição n

xi: valor na posição **i** no conjunto de dados

MA: média aritmética dos dados

n: quantidade de dados

A variância calcula o quadrado da distância de cada um dos pontos até a média (x - x barra) e divide pela quantidade de elementos observados na amostra.

Como fica em uma medida diferente da amostra, por exemplo a amostra está em cm, a variância estará em cm², utilizamos muito o **desvio padrão**, que é **sua raiz quadrada positiva**.

Fórmula do desvio padrão:

$$DP = \sqrt{\sum_{i=1}^{n} \frac{(x_i - M_a)^2}{n}}$$

Sendo,

∑: símbolo de somatório. Indica que temos que somar todos os termos, desde a primeira posição (i=1) até a posição n

xi: valor na posição **i** no conjunto de dados

MA: média aritmética dos dados

n: quantidade de dados

Amplitude

A amplitude é a diferença entre o maior e o menor elemento de um conjunto de dados. Ela não leva em consideração como os dados estão distribuídos, portanto não são tão utilizadas.

Fórmula da amplitude:

A = Xmaior - Xmenor

Medidas de posição: quartis e outliers

Pode ser que usar apenas a variância e o desvio padrão não nos dê uma noção de representação de dados porque são afetados por valores extremos, e também não nos trazem a ideia da simetria da distribuição destes dados.

Para isso, nós utilizamos os quantis de ordem q(p), onde p é uma proporção qualquer, em que todas as observações sejam menores que q(p). O que nós mais utilizamos são alguns quantis bem famosos. São eles:

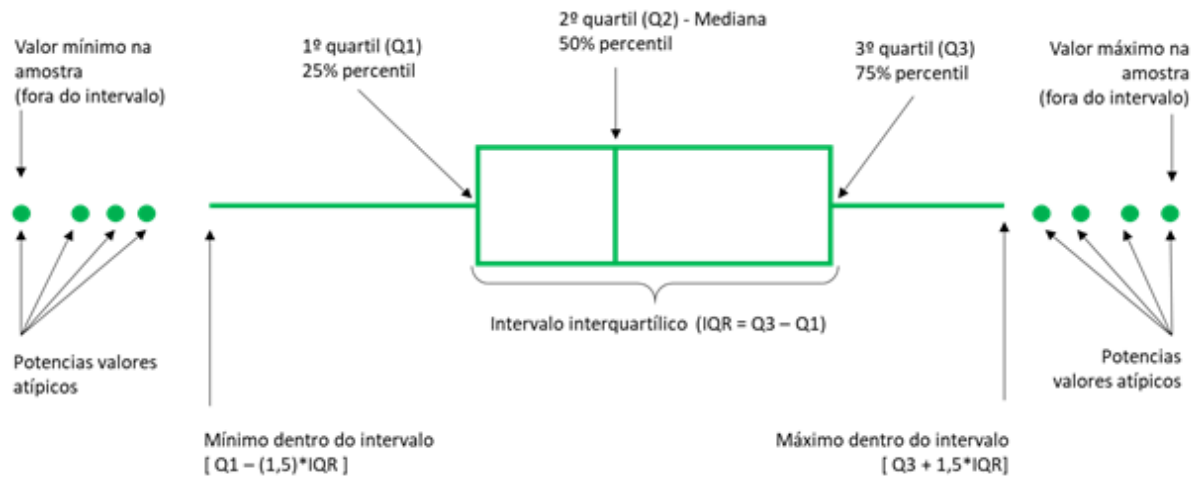
- q(0,25): 1º Quartil

- q(0,50): 2º Quartil - a Mediana ou Separatriz dos dados
- q(0,75): 3º Quartil

Um jeito bem comum de representar esses quartis é com o gráfico de boxplot, ou gráfico "caixa-de-bigodes" (gente eu juro que é esse o nome).

Aqui, a gente vê um gráfico boxplot explicando como funcionam as divisões. A caixa representa o intervalo interquartílico ou dispersão, ou seja, entre o 1º Quartil e o 3º Quartil, o que quer dizer que a caixa contém 50% dos dados analisados.

A caixa fica dentro do limite inferior e limite superior, que contempla também os valores adjacentes. O que fica fora deste intervalo é chamado de ***outlier*** ou de ***valores atípicos.***



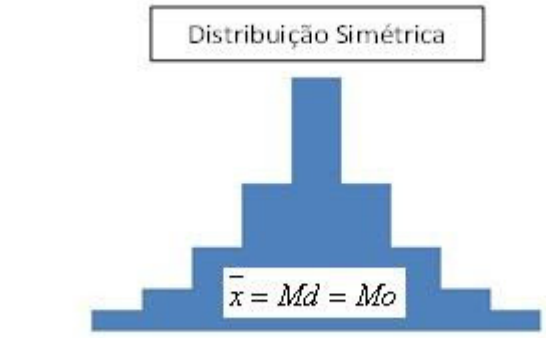
Para calcular as medidas do boxplot, você pode seguir os seguintes passos:

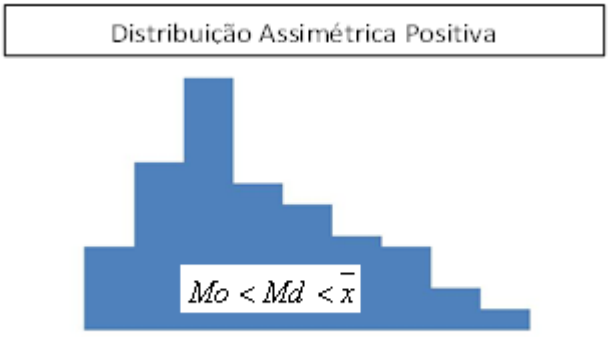
Simetria

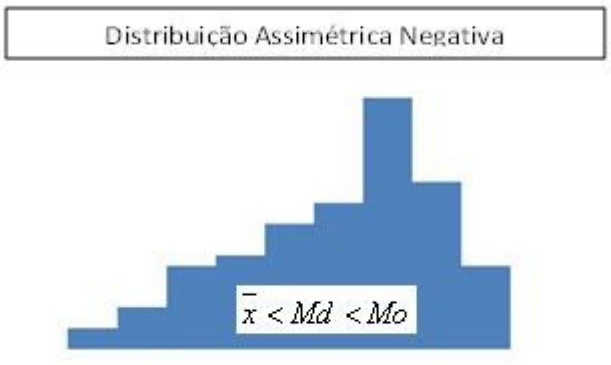
Para entender como os dados indicam a distribuição ao longo do intervalo, podemos observar sua simetria.

Caso a média, a mediana e a moda coincidem, temos uma distribuição simétrica. Caso isso não aconteça, a média puxa a cauda da distribuição para o seu lado, pois é sensível aos valores extremos.

A moda e a mediana puxam a parte mais alta da distribuição para o seu lado. Portanto, podemos pensar na inclinação da simetria de acordo com a diferença da média e da moda.





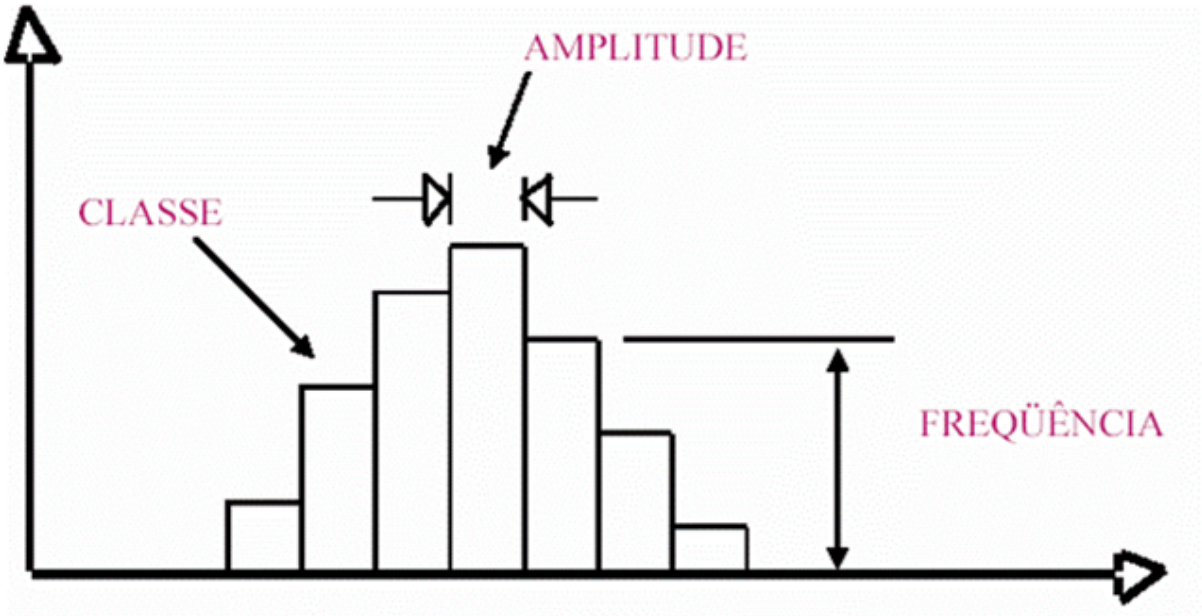


Histograma

Histogramas são uma ótima forma de visualizarmos como nossos dados estão divididos. Eles funcionam em cima da **contagem** de elementos.

Distribuimos no eixo x faixas de intervalo e contamos quantos elementos ficam em cada faixa, resultando em uma distribuição de frequência. Ele ajuda a ter uma ideia de qual distribuição utilizar para prever futuros inputs de data.

Só é importante ressaltar que nem sempre as amplitudes de faixa que vem pré-definidas são as ideais, depende muito dos dados que você está observando. Isso pode tornar seu histograma inútil.



Após assimilar o conhecimento deste capítulo, no próximo bloco falaremos sobre distribuições estatísticas.

Distribuições estatísticas

Normal

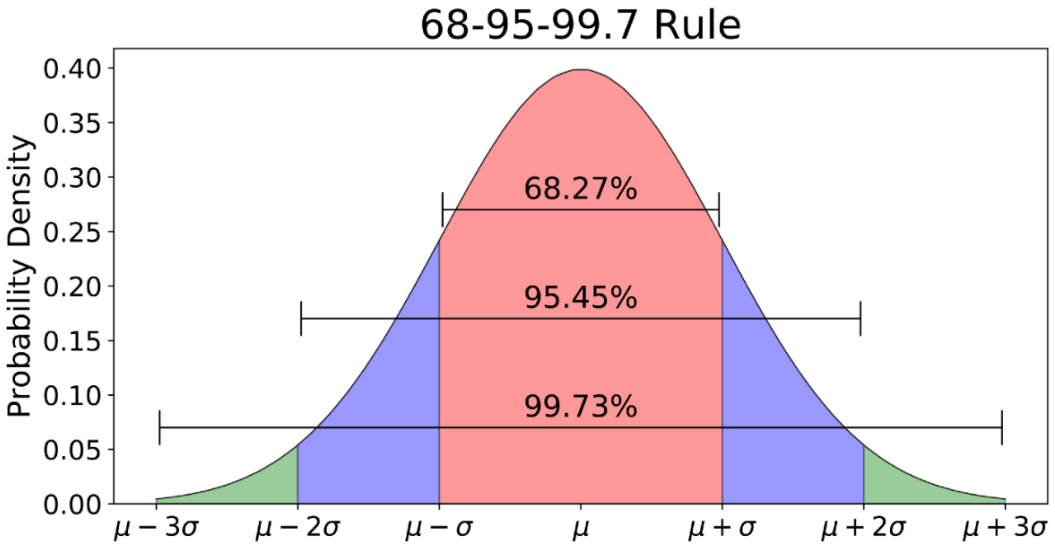
Conhecida como distribuição normal, distribuição Gaussiana ou curva de sino, é uma distribuição simétrica, sempre com o centro na média dos dados.

A curva pode ser bem alta ou bem baixa, dependendo do intervalo de dados que é dado pelo desvio padrão. Quanto menor o desvio padrão, mais alta a curva, quanto maior o desvio padrão mais achatada a curva.

Saber o desvio padrão é muito importante, pois em distribuições normais 95% dos dados ficam entre +-2 desvios padrões da média.

Algumas distribuições normais bem conhecidas são a altura do ser humano e tamanho de bicos de passarinhos, mas você também sabia que o modelo de precificação de derivativos Black Scholes usa a distribuição normal? Ou que a distribuição normal é utilizada para calibrar a balística de armamentos?

Abaixo deixei um gráfico que demonstra melhor como se comportam os dados em distribuições normais.



Curiosidade: Você sabia que a distribuição normal, também chamada de distribuição Gaussiana, tem origem em Gauss e seus trabalhos sobre erros de observações astronômicas realizados em 1810?

Distribuição t de Student

A distribuição de Student é importante para estudos de inferências de médias populacionais, onde não se sabe qual é a média ou o desvio padrão da população, mas ela deve ser normal.

É uma distribuição de densidade de probabilidade que, por ser amplamente utilizada, consta com tabelas prontas para consulta. É uma curva simétrica, que lembra a curva normal, só que tem caudas mais longas.

Sua forma varia conforme seu grau de liberdade, conceito também importante para o cálculo da distribuição qui quadrado.

O grau de liberdade é o número de determinações independentes (dimensão da amostra) menos o número de parâmetros estatísticos a serem avaliados na população.

Suponha que:

- Z é uma variável aleatória de distribuição normal padrão com média 0 e variância 1;

- V é uma variável aleatória com distribuição Chi-quadrado com *v* graus de liberdade.
  Se Z e V são independentes, então a distribuição da variável aleatória t será:

$$t = \frac{Z}{\sqrt{V/\nu}}$$

Curiosidade: Você sabia que o autor da distribuição de Student usava este pseudônimo porque não podia publicar trabalhos em seu nome, William Sealy Gosset, enquanto trabalhava na cervejaria da Guiness?

Distribuição qui quadrado

A distribuição de qui quadrado ou chi square é amplamente utilizada na estatística. Ela serve para realizar o teste de x^2, onde avaliamos quantitativamente a relação entre o resultado de um experimento e a distribuição esperada para o fenômeno.

Ela nos diz com quanta certeza os valores observados podem ser aceitos como regidos pela teoria em questão. Ela também é utilizada em outros testes de hipótese.

A distribuição de qui quadrado é uma distribuição de densidade de probabilidade, que representa a chance de x^2 estar em um determinado intervalo, por exemplo, entre eventos observados e eventos esperados.

É uma distribuição não simétrica e para que seja possível torná-la mais simétrica devemos aumentar seu grau de liberdade.

É bem comum utilizarmos uma tabela para consultar os resultados de qui quadrado que nos apresentam intervalos de confiança para cada grau de liberdade.

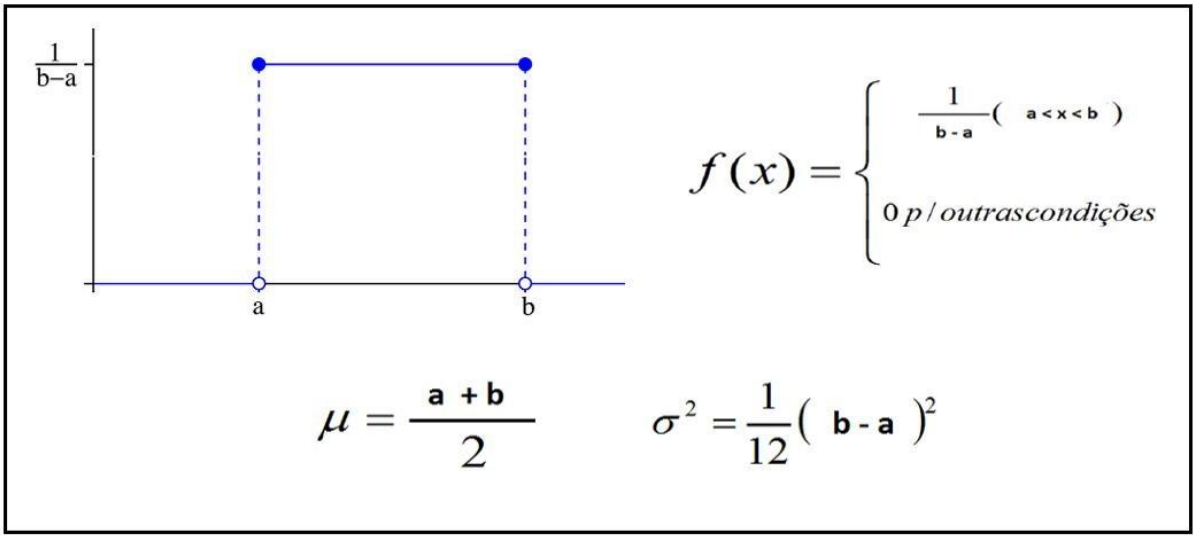
Aqui deixei um gráfico que muda conforme aumentamos o grau de liberdade da distribuição qui quadrado.

Distribuição uniforme

Na distribuição uniforme nós sempre sabemos qual a cara da função da probabilidade. Ela é 1 dividida pela diferença do intervalo, caso o evento esteja no intervalo, ou 0 para os demais casos.

A maioria das linguagens de programação, pacotes estatísticos ou planilhas de cálculo possuem um gerador de números aleatórios, que gera a partir de uma distribuição uniforme, com valores entre 0 e 1.

Esse número é chamado de pseudo-aleatório, porque é possível repetir a mesma sequência a partir de uma mesma semente (valor inteiro).



E já aviso que você utilizará muitas seeds ao longo de sua vida de data scientist.

Saiba Mais

Abaixo, compartilho algumas dicas para que você possa aprofundar seus conhecimentos no assunto.

Bibliografia sugerida

Recomendação de livro:

- Estatística Básica do Morettin

Recomendações de vídeos:

- O que é uma distribuição estatística
- Histogramas
- Média, Mediana e Moda
- Variância e Desvio Padrão
- Quantis e Percentis
- Boxplot
- Distribuição normal
- Distribuição T de Student
- Teste Qui Quadrado
- Distribuição Uniforme

Checklist

Ao final desta aula, você deve sentir segurança para:

- Conceituar Estatística descritiva;
- Diferenciar a estatística descritiva de estatística inferencial;
- Reconhecer os tipos de dados quantitativos (discreto e contínuo);
- Reconhecer os tipos de dados qualitativos (nominais e ordinais);
- Calcular os parâmetros descritivos de medidas de posição (média, mediana e moda);
- Identificar os parâmetros descritivos de medidas de dispersão (amplitude, variância, desvio padrão, e percentis);
- Compreender a visualização de percentis através do boxplot;
- Compreender a medida de assimetria (Skew);
- Identificar as principais distribuições (Normal, Binomial, t de student, Qui quadrado e uniforme);
- Identificar os parâmetros descritivos em datasets e exemplos reais.