Mundo Tech

ESTATÍSTICA BÁSICA PARA ANÁLISE DE DADOS

SENAI <LAB365>

SUMÁRIO

"Sem dados, você é apenas mais uma pessoa com uma opinião." Willan Edwards Deming (1900-1993)

Neste material, vamos estudar a relação entre *Data Science* e Estatística – muitas pessoas acham que são a mesma coisa. Vamos lá descobrir?

DATA SCIENCE E ESTATÍSTICA

Data science e estatística não são a mesma coisa. Na verdade, *data science* faz uso de diferentes disciplinas para se tornar completa, como o uso da Ciência da Computação, que nos fornece os métodos automatizados para que, utilizando a matemática e a estatística, possamos realizar as análises dos dados a fim de extrair conhecimento nas diferentes áreas de negócios.

Dessa forma, a estatística é um dos importantes pilares necessários para o trabalho do cientista de dados.

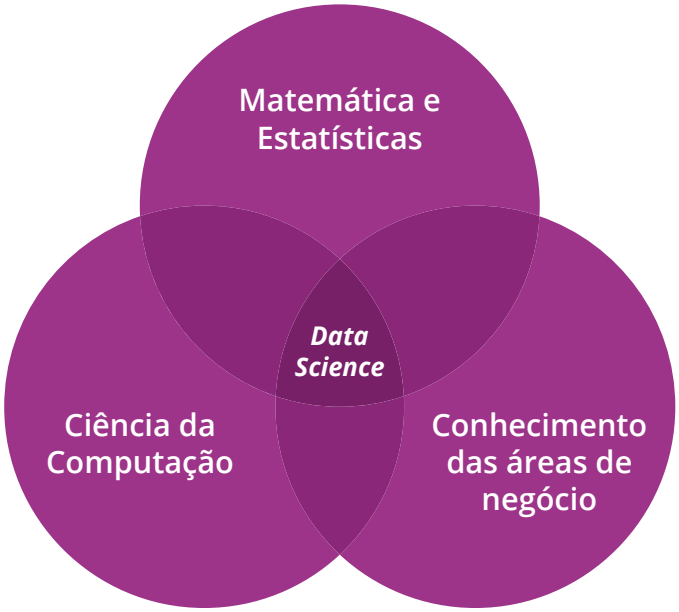


Figura 1 - Disciplinas que envolvem *Data Science*

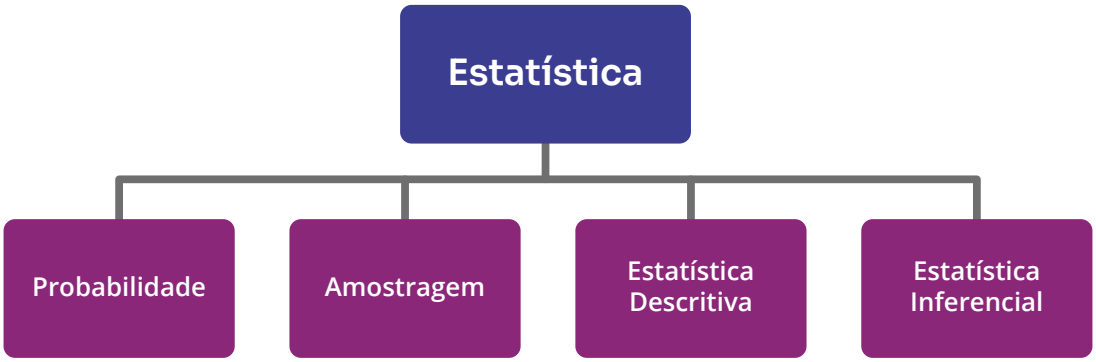
Fonte: Da autora (2022)

Curiosidade

Segundo David Morganstein, Presidente da ASA – Amerian Statistical Association, "a ciência de dados abrange mais do que estatísticas, mas ao mesmo tempo deve-se reconhecer que a ciência estatística desempenha um papel fundamental para o crescimento deste campo." (MORGANSTEIN, 2015, online)

CONHECENDO A ESTATÍSTICA

A estatística tem como função tratar os dados. Os dados se referem a uma variável que pode assumir valores em diferentes unidades, como visto anteriormente. A estatística está organizada em quatro grandes áreas: estatística descritiva, probabilidade, amostragem e estatística inferencial.



Estatística descritiva

É o ramo da estatística que faz uso da organização e descrição dos dados a partir de um *dataset*. Normalmente, essa descrição é apresentada em forma de gráficos ou tabelas. Mostra uma análise do que já aconteceu.

A estatística descritiva é normalmente calculada a partir dos dados brutos e desorganizados. Responde perguntas como:

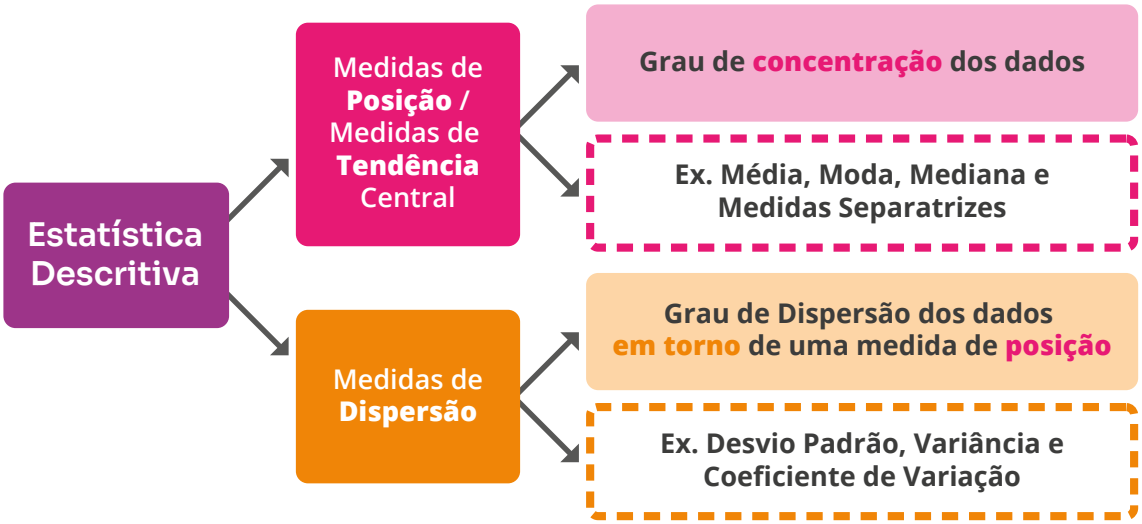
- Qual foi o volume médio de dados no último mês?
- Qual foi variação das notas dos meus alunos no curso de *Data Science*?

A estatística descritiva pode ser dividida em 2 grupos, sendo eles:

Medidas de posição ou tendência central: aqui, estamos interessados no grau de concentração dos dados – ou seja, será que os maiores valores estão concentrados no final da distribuição, no início da distribuição, acima ou abaixo da média? Exemplos: média, moda, mediana.

- Média: o valor médio dos dados.
- Mediana: o valor central se ordenarmos os dados em ordem crescente e dividirmos exatamente pela metade.
- Moda: o valor que ocorre com mais frequência.

Medidas de dispersão: aqui, estamos interessados no quão parecido são meus dados, em como estão meus dados dispersos em uma medida de posição – por exemplo, qual a distância que meus dados estão da média?



Variância

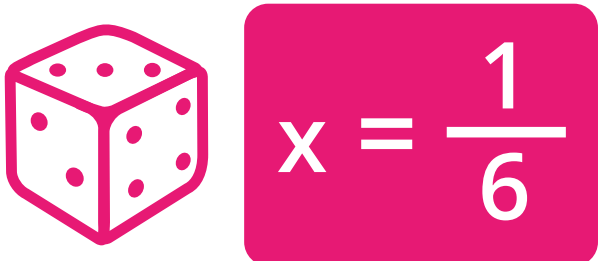
Como já citado, a média é uma medida de tendência central. A variância mede a distância de cada valor no conjunto de dados da média. Essencialmente, é uma medida da dispersão dos números em um conjunto de dados.

O desvio padrão é uma medida comum de variação para dados que têm uma distribuição normal. É um cálculo que fornece um valor para representar a extensão da distribuição dos valores. Um desvio padrão baixo indica que os valores tendem a ficar muito próximos da média, enquanto um desvio padrão alto indica que os valores estão mais dispersos.

Se os dados não seguem uma distribuição normal, outras medidas de variância são usadas. Normalmente, o intervalo interquartil é usado. Essa medida é derivada primeiramente ordenando os valores por classificação e, em seguida, dividindo os pontos de dados em quatro partes iguais, chamadas quartis. Cada quartil descreve onde 25% dos pontos de dados se encontram, de acordo com a mediana. O intervalo interquartil é calculado subtraindo-se a mediana dos dois quartos centrais, também conhecidos como Q1 e Q3.

Probabilidade

É o ramo da estatística que estuda a aleatoriedade e a incerteza. Não temos como garantir que um determinado evento irá ocorrer, mas podemos utilizar ferramentas para prever e auxiliar na tomada de decisões. Por exemplo, ao jogar um dado, não sabemos qual valor irá cair, mas conseguimos contabilizar os valores possíveis. Dessa forma, consigo dar um valor para a incerteza.



Responde perguntas como:

- Qual a probabilidade de eu passar num concurso?
- Qual a probabilidade de chover neste final de semana?

Amostragem

A amostragem estuda técnicas para seleção de uma amostra de uma população. Em estatística, o conjunto formado por todos os dados brutos que você pode ter disponíveis para um teste ou experimento é conhecido como população. Por uma série de razões, não é viável medir os padrões e tendências em toda a população. As estatísticas nos permitem tomar uma amostra, realizar alguns cálculos sobre o conjunto de dados e, usando a probabilidade e algumas suposições, podemos, com um certo grau de certeza, compreender as tendências para toda a população ou prever eventos futuros.

Digamos, por exemplo, que queremos entender a prevalência de uma doença, como o câncer de mama, em toda a população do Brasil. Por razões práticas, não é possível rastrear toda a população. Em vez disso, podemos pegar uma amostra aleatória e medir a prevalência entre esses dados. Supondo que nossa amostra seja suficientemente aleatória e representativa de toda a população, podemos obter uma medida de prevalência e fazer inferências sobre toda a população.

Atenção

População é a coleção de todos os indivíduos que estamos interessados em estudar.

Amostra é um subconjunto da população, é uma parte desses indivíduos.

Estatística Inferencial

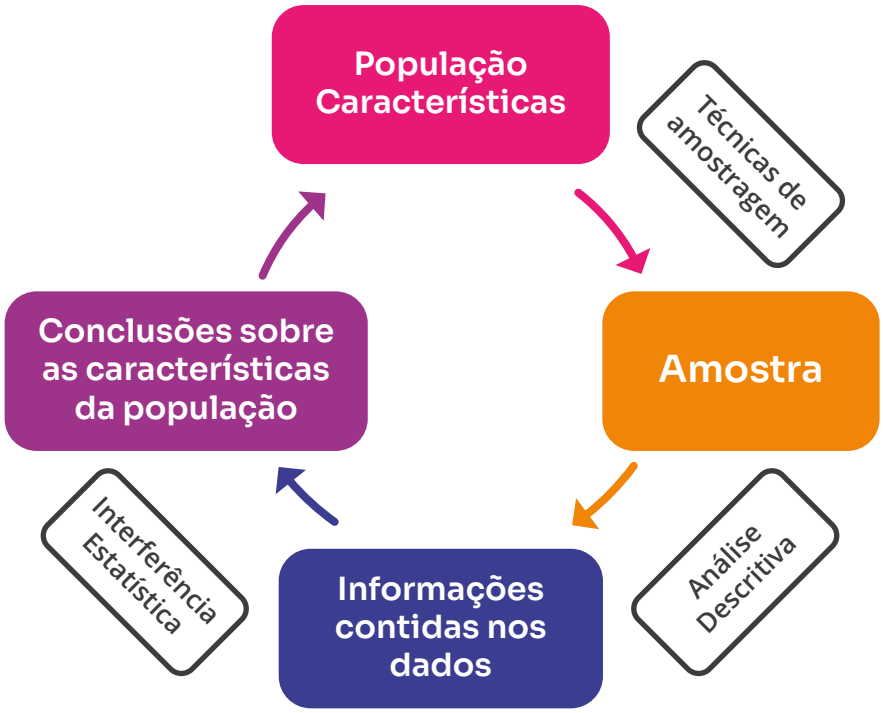
A estatística inferencial busca estimar informações de uma população a partir de resultados observados em uma amostra de dados.

O melhor exemplo é uma pesquisa para candidatos a presidente: não é possível entrevistar todos os eleitores, então, usa-se uma amostragem.

ESTATÍSTICA E *DATA SCIENCE*

Se você observar um problema de *data science*, ele parte de um negócio. Entender os dados do meu negócio é importante e dele vamos extrair uma amostra de dados sobre o meu negócio para realizar a análise. É justamente essa amostra que será o seu *dataset*, que passará pelo processo inicial de análise descritiva e posteriormente uma análise inferencial.

A estatística nos oferece técnicas e procedimentos, e é utilizada em diferentes momentos de um projeto de *data science*.



OUTROS CONCEITOS IMPORTANTES

Distribuições

As estatísticas descritivas são úteis, mas muitas vezes podem ocultar informações importantes sobre o conjunto de dados. Por exemplo, se um conjunto de dados contém vários números que são muito maiores do que os outros, a média pode ser distorcida e não nos dar uma representação verdadeira dos dados.

Uma distribuição pode ser representada por um gráfico, geralmente um histograma, que exibe a frequência com que cada valor aparece em um conjunto de dados. Esse tipo de gráfico nos fornece informações sobre a dispersão e a assimetria dos dados.

Uma distribuição geralmente formará um gráfico semelhante a uma curva, que pode ser inclinada mais para a esquerda ou direita.

Uma das distribuições mais importantes é a distribuição normal, comumente chamada de curva em sino devido ao seu formato. É de forma simétrica, com a maioria dos valores agrupados em torno do pico central e os valores mais distantes distribuídos igualmente em cada lado da curva. Muitas variáveis na natureza formarão uma distribuição normal, como a altura das pessoas e as pontuações de QI. A distribuição normal de uma variável é a suposição de vários algoritmos de *machine learning*.

É importante que tenha ficado claro para você que a estatística nos oferece técnicas e procedimentos, e é utilizada em diferentes momentos de um projeto de *data science*.

Data Science utilizada a estatística para explorar e analisar dados de negócio, tudo isso com o auxílio da Ciência da Computação.

REFERÊNCIAS

MORGANSTEIN, D. The Role of Statistics in Data Science (Statement on 10 Jan. 2015). *In*: WASSERSTEIN, R. **The Role of Statistics in Data Science** – An ASA statement. American Statistical Association (ASA) Community, Alexandria (VA, USA), 2015. Disponível em: https://community.amstat.org/blogs/ronald-wasserstein/2015/10/01/the-role-of-statistics-in-data-science-an-asa-statement. Acesso em: 15 ago. 2022.

SENAI <LAB365>