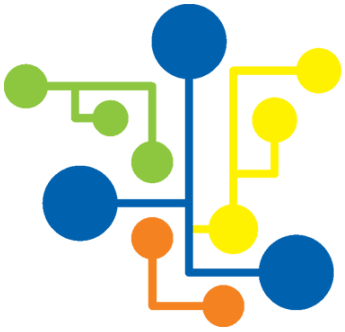

Data Science Academy

www.datascienceacademy.com.br

Microsoft Power BI Para Data Science

Estatística Descritiva
x
Estatística Inferencial

O processo de análise de dados é composto de várias etapas e em cada uma delas devemos empregar as ferramentas corretas a fim de obter o melhor resultado possível. A Estatística Descritiva é o suporte para grande parte do processo de análise e usada principalmente na etapa de análise exploratória, quando você busca compreender os dados a fim de saber como eles devem ser modelados e quais ferramentas utilizar nas etapas posteriores.

Na maioria das vezes, estaremos trabalhando com amostras dos dados e não com a população inteira. Ao criar um modelo para prever se um paciente pode ou não desenvolver uma doença, não usamos dados históricos de toda a população, pois isso seria inviável, com alto custo ou mesmo desnecessário. Neste caso, usamos uma amostra com informações de um grupo de pessoas, amostra essa que deve ser representativa da população, e realizamos nosso trabalho de análise. Mas como nosso interesse está em fazer previsões para a população, usamos a Estatística Inferencial para fazer inferências sobre a população a partir dos resultados obtidos na análise da amostra. Existem várias técnicas e ferramentas para isso e o campo de Estatística Inferencial ajuda a explicar boa parte dos algoritmos de Machine Learning. Antes de avançar no estudo de algumas dessas técnicas e ferramentas, vamos definir o que é Estatística Descritiva e Inferencial.

Estatísticas Descritivas

As estatísticas descritivas são o termo dado à análise de dados que ajuda a descrever, visualizar ou resumir dados de forma significativa, de modo que, por exemplo, os padrões possam emergir dos dados. As estatísticas descritivas, no entanto, nos permitem tirar conclusões além dos dados que analisamos ou chegar a conclusões sobre quaisquer hipóteses que possamos ter feito. Elas são simplesmente uma maneira de descrever nossos dados.

As estatísticas descritivas são muito importantes porque, se simplesmente apresentássemos nossos dados brutos, seria difícil visualizar o que os dados estavam mostrando, especialmente se houvesse muitos. Por conseguinte, as estatísticas descritivas nos permitem apresentar os dados de forma mais significativa, o que permite uma interpretação mais simples dos dados. As principais estatísticas que são usados para descrever dados são:

Medidas de tendência central: são formas de descrever a posição central de uma distribuição de frequência para um grupo de dados. Podemos descrever a posição central usando uma série de estatísticas, incluindo a moda, a mediana e a média.

Medidas de dispersão: são formas de resumir um grupo de dados, descrevendo como são distribuídos os resultados. Medidas de dispersão nos ajudam a resumir como são distribuídos esses resultados. Para descrever esta dispersão, uma série de estatísticas estão disponíveis para nós, incluindo o intervalo, quartis, desvio absoluto, variância e desvio padrão.

Quando usamos estatística descritiva, é útil resumir o nosso grupo de dados utilizando uma combinação de descrição tabulada (isto é, tabelas), descrição gráfica e comentários estatísticos (isto é, uma discussão sobre os resultados).

Estatística Inferencial

Vimos que as estatísticas descritivas fornecem informações sobre nosso conjunto de dados. Por exemplo, podemos calcular a média e desvio padrão de todos os parafusos produzidos por uma empresa e isso nos daria informações preciosas sobre o processo de fabricação. Qualquer conjunto de dados como este, que inclui todos os dados que você está interessado, é chamado de população. Uma população pode ser pequena ou grande, desde que inclua todos os dados que lhe interessam. As estatísticas descritivas são aplicadas às populações, e as propriedades das populações, como a média ou o desvio padrão, são chamadas parâmetros, pois representam toda a população (ou seja, todos os dados nos quais você está interessado).

Muitas vezes, no entanto, você não tem acesso a toda a população que você está interessado em investigar, mas apenas um número limitado de dados. Ou ainda, é inviável analisar toda a população o mesmo desnecessário. No exemplo da produção de parafusos, não é viável medir todos os parafusos produzidos pela empresa todos os dias. Então você precisa medir uma amostra menor de parafusos (por exemplo, um lote de parafusos do turno da manhã), que é usada para representar a população de parafusos produzidos. As propriedades das amostras, como a média ou o desvio padrão, não são chamados de parâmetros, mas sim de estatísticas. As estatísticas inferenciais são técnicas que nos permitem usar essas amostras para fazer generalizações ou inferências sobre as populações das quais as amostras foram obtidas. É, portanto, importante que a amostra represente com precisão a população. O processo de alcançar isso é chamado de amostragem. As estatísticas inferenciais surgem do fato de que a amostragem naturalmente incorre em erro de amostragem e, portanto, não se espera que uma amostra represente perfeitamente a população. Os métodos das estatísticas inferenciais são a estimativa dos parâmetros e teste de hipóteses estatísticas.

Estes tópicos são estudados ao longo de diversos cursos da Formação Cientista de Dados. O objetivo aqui é dar a você uma visão geral de alguns tópicos principais e cada um destes conceitos é amplamente explorado em outros cursos aqui na DSA.

Vejamos agora algumas destas técnicas e conceitos usados em Estatística Inferencial.