Análise de dados

A disciplina de **Análise de Dados** é um campo multidisciplinar que envolve a extração, transformação e interpretação de dados para a tomada de decisões informadas. Fundamentada em estatística, ciência da computação e teoria da informação, a análise de dados tem aplicações em diversas áreas, como negócios, saúde, finanças e inteligência artificial.

Davenport e Harris (2007), no livro *Competing on Analytics*, destacam que organizações orientadas por dados obtêm vantagens competitivas ao transformar informações em insights acionáveis. Já Provost e Fawcett (2013), em *Data Science for Business*, enfatizam a importância dos dados como um ativo estratégico, destacando métodos estatísticos e de aprendizado de máquina para análise preditiva.

Outra contribuição essencial vem de Tukey (1977), que, em *Exploratory Data Analysis*, reforça a necessidade da exploração visual e interativa dos dados antes da aplicação de modelos estatísticos, influenciando a abordagem moderna de análise de dados.

Com o avanço das tecnologias de big data e inteligência artificial, a disciplina de análise de dados evoluiu para incorporar técnicas avançadas de aprendizado de máquina e mineração de dados, conforme descrito por Hastie, Tibshirani e Friedman (2009) em *The Elements of Statistical Learning*.

Dessa forma, a análise de dados se estabelece como um campo essencial para transformar grandes volumes de informação em conhecimento útil, permitindo uma tomada de decisão baseada em evidências.

Tópicos previsto para disciplina

Estatística

A estatística é uma área da matemática que se dedica à coleta, organização, análise, interpretação e apresentação de dados. Seu objetivo principal é extrair informações relevantes e conclusões significativas a partir de dados coletados de amostras ou populações. A estatística é fundamental em diversos campos, como ciências sociais, economia, saúde, educação, e muitos outros, pois permite tomar decisões informadas baseadas em evidências.

Representações gráficas

Representações gráficas são uma ferramenta visual usada para apresentar dados de forma clara e acessível. Através delas, podemos identificar padrões, tendências, distribuições e relações entre variáveis de maneira intuitiva. Exemplos incluem gráficos de barras, histogramas, diagramas de dispersão, gráficos de linha e setores. Essas representações são essenciais para a interpretação rápida e eficiente de grandes volumes de dados.

Medidas de tendência central - Média Aritmética, Média Geométrica, Média Harmônica

As medidas de tendência central são utilizadas para descrever o valor típico ou central de um conjunto de dados.

- **Média Aritmética**: É a soma de todos os valores dividida pelo número de observações. É a medida de tendência central mais comum e é útil quando os dados não possuem grandes variações extremas.
- **Média Geométrica**: É o valor médio calculado através do produto dos dados e da raiz enésima do resultado, onde "n" é o número de dados. Essa medida é utilizada principalmente em situações que envolvem taxas de crescimento ou variações percentuais, como o cálculo do retorno de investimentos ao longo do tempo.
- **Média Harmônica**: Utilizada em situações onde se deseja ponderar mais os valores menores, como na média de velocidades. Ela é calculada como o inverso da média aritmética dos inversos dos dados. Esse tipo de média é comum em áreas como física e economia.

Medidas de tendência central - Moda e Mediana

Além das médias, existem outras duas importantes medidas de tendência central: a moda e a mediana.

- **Moda**: É o valor ou valores que ocorrem com maior frequência em um conjunto de dados. Quando os dados apresentam uma grande concentração de valores iguais, a moda fornece uma boa representação do que é mais comum ou recorrente no conjunto.
- **Mediana**: É o valor que separa a metade superior e a metade inferior de um conjunto de dados ordenados. Em um conjunto de dados ímpar, a mediana é o valor do meio; em um conjunto par, a mediana é a média dos dois valores centrais. A mediana é especialmente útil quando se lida com dados assimétricos ou com outliers, pois não é influenciada por valores extremos.

Medidas separatrizes

As medidas separatrizes são usadas para dividir um conjunto de dados em partes menores, com o objetivo de analisar a distribuição e a dispersão dos dados. Os principais tipos são:

- **Quartis**: Dividem os dados em quatro partes iguais, sendo o primeiro quartil (Q1) o valor abaixo do qual 25% dos dados estão, o segundo quartil (Q2) corresponde à mediana, e o terceiro quartil (Q3) divide os dados de forma que 75% dos dados estão abaixo dele.
- **Percentis**: Dividem os dados em 100 partes iguais, permitindo uma análise ainda mais detalhada da distribuição dos dados. Cada percentil indica o valor abaixo do qual uma determinada porcentagem dos dados se encontra.

Medidas de dispersão

As medidas de dispersão são utilizadas para analisar a variação ou a dispersão dos dados em relação à média. Elas ajudam a entender o quanto os dados estão espalhados ou concentrados. As principais medidas de dispersão são:

- **Amplitude**: A diferença entre o maior e o menor valor de um conjunto de dados.
- **Desvio padrão**: Mede a média dos desvios de cada valor em relação à média do conjunto de dados. Quanto maior o desvio padrão, maior a dispersão dos dados.
- **Variância**: É o quadrado do desvio padrão e também mede a dispersão, mas de forma mais sensível a valores extremos.

Introdução à teoria da amostragem

A amostragem é um processo que envolve a seleção de um subconjunto representativo de uma população para fazer inferências sobre ela. A teoria da amostragem é fundamental para a estatística, pois permite que, a partir de uma amostra, se obtenha uma estimativa para parâmetros populacionais, como a média ou a proporção. A amostragem pode ser probabilística (onde cada elemento tem uma chance conhecida de ser selecionado) ou não probabilística.

Fatorial, Permutação, Arranjo, Combinação

Esses conceitos estão relacionados à contagem e organização de elementos em um conjunto, sendo essenciais para o cálculo de probabilidades.

- **Fatorial (n!)**: Representa o produto de todos os números inteiros positivos até "n". É utilizado em problemas que envolvem permutações ou arranjos de elementos.
- **Permutação**: Refere-se à disposição de elementos em uma ordem específica. O número de permutações de "n" elementos é dado por n!.
- **Arranjo**: É uma seleção de "k" elementos a partir de um conjunto de "n" elementos, levando em conta a ordem.
- **Combinação**: Diferente da permutação, nas combinações não se considera a ordem dos elementos selecionados. O número de combinações de "n" elementos tomados de "k" em "k" é dado pela fórmula (\frac{n!}{k!(n-k)!}).

Probabilidade de um evento - Definições básicas

Probabilidade é uma medida da chance de ocorrência de um evento. A probabilidade de um evento (E) é calculada pela razão entre o número de resultados favoráveis e o número total de resultados possíveis. Em termos matemáticos, (P(E) = \frac{\text{número de resultados favoráveis}}{\text{número total de resultados possíveis}}). A probabilidade varia de 0 a 1, sendo 0 impossível e 1 certeza.

Teoremas de cálculo de probabilidade

Vários teoremas são utilizados para calcular as probabilidades de eventos compostos ou independentes. Alguns dos principais teoremas incluem:

- **Teorema da probabilidade total**: Calcula a probabilidade de um evento a partir de várias possibilidades mutuamente exclusivas.
- **Teorema de Bayes**: Permite calcular a probabilidade de um evento com base em informações anteriores, sendo essencial em contextos como a atualização de probabilidades à medida que novas informações são obtidas.

Probabilidade condicional e eventos independentes

- **Probabilidade condicional**: Refere-se à probabilidade de um evento ocorrer dado que outro evento já ocorreu. A probabilidade condicional de um evento (A) dado que (B) ocorreu é dada por $P(A|B) = \frac{P(A \cap B)}{P(B)}$.
- **Eventos independentes**: São eventos em que a ocorrência de um evento não afeta a probabilidade de ocorrência do outro. Se dois eventos (A) e (B) são independentes, então $P(A \cap B) = P(A) \cdot P(B)$.

Variáveis aleatórias discretas - Distribuição equiprovável e distribuição de Bernoulli

Variáveis aleatórias discretas assumem valores contáveis, e sua distribuição descreve a probabilidade associada a cada valor.

- **Distribuição equiprovável**: Todos os resultados possíveis têm a mesma probabilidade. Exemplo clássico é o lançamento de uma moeda honesta, onde a probabilidade de cara ou coroa é a mesma.
- **Distribuição de Bernoulli**: Relacionada a experimentos com dois resultados possíveis (sucesso ou fracasso). A probabilidade de sucesso é (p) e a de fracasso é (1 - p).

Distribuição binomial e distribuição de Poisson

- **Distribuição binomial**: Modela o número de sucessos em um número fixo de experimentos independentes, com dois resultados possíveis e a mesma probabilidade de sucesso em cada experimento.
- **Distribuição de Poisson**: Utilizada para modelar o número de ocorrências de um evento em um intervalo fixo de tempo ou espaço, quando esses eventos ocorrem de maneira independente e a uma taxa constante.

Distribuição geométrica e distribuição hipergeométrica

- **Distribuição geométrica**: Descreve o número de tentativas até o primeiro sucesso em uma sequência de experimentos independentes de Bernoulli.
- **Distribuição hipergeométrica**: Semelhante à distribuição binomial, mas é utilizada quando os experimentos são realizados sem reposição, como na seleção de uma amostra de uma população

finita.

Distribuição de Pascal (ou Distribuição Binomial Negativa)

A distribuição de Pascal, ou binomial negativa, é uma generalização da distribuição geométrica. Ela descreve o número de falhas antes de um número fixo de sucessos em experimentos independentes de Bernoulli.

Variáveis aleatórias contínuas

Variáveis aleatórias contínuas podem assumir qualquer valor dentro de um intervalo. Para variáveis contínuas, a probabilidade de um valor específico ocorrer é 0, mas a probabilidade de um valor cair dentro de um intervalo é dada pela área sob a curva da distribuição de probabilidade.

Distribuição uniforme

Na distribuição uniforme contínua, todos os valores dentro de um intervalo têm a mesma probabilidade de ocorrer. Sua função de densidade é uma linha reta no intervalo considerado.

Distribuição exponencial
A distribuição exponencial modela o tempo entre ocorrências de um evento em um processo de Poisson, ou seja, descreve o tempo até o próximo evento ocorrer em um processo contínuo e aleatório.

Distribuição normal
A distribuição normal, ou gaussiana, é uma das distribuições mais importantes da estatística. Sua curva é simétrica e em forma de sino, e é definida por dois parâmetros: a média ((\mu)) e o desvio padrão ((\sigma)). A distribuição normal é amplamente utilizada para modelar variáveis naturais e fenómenos de erro aleatório.

Em resumo, todos esses tópicos estão interconectados para proporcionar uma compreensão abrangente de como trabalhar com dados, calcular probabilidades, e interpretar variáveis aleatórias. A combinação desses conceitos forma a base para análise estatística, que é crucial para a tomada de decisões e inferências em diversas áreas.