Implementação de Funções Estatísticas para Mineração de Dados em Python

Anna Beatriz S. Lima[1], Ian Salomão S. Carneiro[1], Valéria S. Santos[1]

[1]Disciplina de Mineração de Dados – Centro Universitário de Excelência (UNEX)
44.085-370 – Feira de Santana– BA – Brazil

ab.annabeatrizlima@gmail.com, iansalomao.ca@gmail.com,
valeriasoressantos2@gmail.com

Abstract. *This paper is a technical and scientific report on the implementation of a series of statistical metrics. We will present the relevance of these statistical implementations, how they were implemented, and the challenges of this development process.*

Resumo. *Este artigo é um relatório técnico científico sobre a implementação de uma série de métricas estatísticas. Apresentaremos a relevância dessa implementações estatísticas, a forma como foram implementadas e os desafios desse processo de desenvolvimento.*

1. Introdução

No contexto atual de tecnologia, dados são o bem mais valioso em diversos setores. Sendo considerado o 'petróleo' da contemporaneidade, eles podem fornecer informações importantes que moldam a tomada de decisões em instituições e empresas. Entretanto, para que todo seu potencial seja aproveitado, é necessária uma análise eficiente desses dados, para que possam ser transformados em informações úteis, e estas em conhecimento aplicável.

Assim, não é equivocado afirmar que a estatística tem um papel intrínseco em como a economia, empresas e instituições do mundo moderno funcionam, se desenvolvem e tomam decisões importantes. São os procedimentos estatísticos que fornecem bases para organizar, resumir e interpretar dados, transformando-os em informações, que por sua vez se transformam em conhecimento, que sustentam a tomada de decisões mais assertivas..

Dessa maneira, este relatório apresenta a implementação das operações estatísticas mais básicas em Python, criadas sem a utilização de bibliotecas externas. A aplicação dessas operações se deu em um projeto simulado de uma empresa de delivery, a *FoodDelivery*, expondo como estas métricas podem ser utilizadas em algoritmos de recomendação.

2. Fundamentação Teórica

Considerando a importância da estatística no processo de transformação de dados em informação, esta seção apresenta os principais conceitos e métricas estatísticas que servem de base para o desenvolvimento deste trabalho.

2.1. Média

A média aritmética é uma medida de tendência central que representa o valor típico de um conjunto de dados. Para calculá-la, somam-se todos os valores do conjunto e divide-se pelo número total de elementos.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

Figura 1. Fórmula da Média

Em que $\bar{x}$ é média aritmética, $x_1 + x_2 + \ldots$ é a soma de todos os valores do conjunto, e n é o número de elementos.

2.2. Mediana

A mediana é o valor central de um conjunto de dados organizados em ordem crescente ou decrescente. Para calculá-la, se o número de elementos *n* for ímpar, a mediana corresponde ao valor central e se *n* for par, é calculada a média aritmética dos dois valores centrais do conjunto ordenado.

2.3. Moda

Moda é a métrica que define quais elementos ocorrem mais frequentemente em um conjunto de dados. Para identificar a(s) moda(s) em um grupo de elementos, é preciso contar com qual frequência cada valor aparece, buscando, posteriormente, o(s) valor(es) de maior frequência e definindo assim a moda daquele conjunto.

2.4. Desvio Padrão

O desvio padrão é uma medida de dispersão que indica o quanto os valores de um conjunto de dados se afastam da média, refletindo a uniformidade dos dados. Quanto mais próximo de zero for o desvio padrão, mais homogêneo é o conjunto de dados, ou seja, os valores estão próximos da média. Por outro lado, quanto maior for o desvio padrão, maior é a irregularidade dos dados, indicando que os valores estão mais espalhados em relação à média.

$$D_p = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

Figura 2. Fórmula do Desvio Padrão

Na fórmula do desvio-padrão (Dp), n representa a quantidade de elementos do conjunto, "xi" são os elementos do conjunto e " $\bar{x}$" é a média do conjunto. O desvio-padrão é calculado como a raiz quadrada da somatória dos quadrados das diferenças entre cada elemento e a média, dividida pelo número total de elementos.

2.5. Variância

A variância é uma medida de dispersão que indica o quanto os valores de um conjunto se afastam, em média, da sua média aritmética. Ela é calculada somando o quadrado das diferenças entre cada valor e a média, e dividindo pelo número total de elementos.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Figura 3. Fórmula da Variância

Onde N representa o número total de elementos da população, x_i é cada valor individual da população, μ corresponde à média da população e σ^2 indica a variância populacional.

2.6. Covariância

A covariância é uma medida estatística que permite a comparação entre duas variáveis, ajudando a entender como elas mudam juntas ou se correlacionam entre si. Basicamente, ela checa se as variáveis aumentam ou diminuem de maneira simultânea, sendo a covariância positiva caso elas se alterem ao mesmo tempo de maneira direta ou negativa caso se alterem de maneira inversa.

$$\mathrm{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)$$

Figura 4. Fórmula da Covariância

Onde ∑ é somatória dos itens seguintes, Xi indica o valor de "x" na posição "i", μx é valor médio de x em todas as posições, Yi indica valor de "y" na posição "i", μy é o valor médio de Y em todas as posições e N é quantidade de posições.

2.7. Itens Únicos

Em estatísticas, itens únicos são os diferentes valores que aparecem em um conjunto de dados, desconsiderando repetições. Cada item único é contado apenas uma vez, independente do número de vezes que ocorre. Essa medida é útil para identificar a diversidade de valores em uma amostra.

2.8. Frequência Absoluta

A frequência absoluta corresponde ao número de vezes que um mesmo elemento aparece em um determinado conjunto de dados. Em outras palavras, trata-se da contagem simples das ocorrências de cada valor. A partir dessa medida, podem ser derivadas outras, como a frequência relativa.

2.9. Frequência Relativa

A frequência relativa representa a proporção da contagem de quantas vezes um valor aparece em um conjunto(frequência absoluta) em relação a quantidade total de dados, transformando essa contagem em porcentagem para melhor entendimento do peso de cada elemento.

$$Frequência\ Relativa = \frac{Frequência\ Absoluta}{Total\ de\ Elementos}$$

Essa métrica pode ajudar a entender a importância ou peso de cada valor dentro de um conjunto de dados, sendo útil em qualquer análise estatística para identificar padrões ou tendências.

2.10. Frequência Acumulada

A frequência acumulada se refere à soma progressiva das frequências dos valores do conjunto de dados, podendo ser absoluta ou relativa. Para obter a absoluta, ordena-se as frequências obtidas pelos valores ordem crescente e soma-se progressivamente cada frequência, obtendo a frequência acumulada de cada valor. Para a relativa, divide-se o valor da absoluta pelo total de dados/observações.

$$FAC_i = \sum_{j=1}^{i} f_j \text{ e } FAR_i = \frac{FAC_i}{N}$$

A frequência acumulada ajuda a entender a distribuição acumulada de dados, podendo identificar os valores mais frequentes em um conjunto e facilitar a tomada de decisões baseadas em tendências.

2.11. Probabilidade Condicional

A probabilidade condicional representa a chance de um evento ocorrer sob a condição de que outro evento já tenha ocorrido. Ela restringe o espaço amostral às situações em que o segundo evento é verdadeiro, permitindo analisar a ocorrência do primeiro evento dentro desse contexto.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{desde que } P(B) \neq 0$$

Figura 5. Fórmula da Probabilidade Condicional

Onde $P(A|B)$ representa a probabilidade de A ocorrer dado que B ocorreu, $P(A \cap B)$ é a probabilidade de A e B ocorrerem simultaneamente, e P(B) indica a probabilidade de B ocorrer.

3. Metodologia

Para fazer a aplicação de todas as operações citadas no relatório, a classe *Statistics* foi criada. Para a inicialização de um objeto Statistics, a função *__init__(self, dataset)* foi criada, com verificações essenciais para garantir que o *dataset*(conjunto de dados) é um dicionário formado por listas do mesmo tamanho com o mesmo tipo de dado em cada lista.

A estrutura de dados de dicionário se mostra como ideal para a utilização dos métodos que foram implementados, pois organiza os dados de forma clara e permite que cada variável seja acessada através do seu nome. Bem como cada valor do dicionário precisa ser uma lista, pois diversas operações precisam percorrer todos os valores de cada tipo de variável. Além disso, as listas precisam ser do mesmo tamanho por conta de métricas que exigem uma correlação de dados, como covariância, que só funciona se obtiver dados correspondentes.

Ademais, as seguintes funções também foram criadas para garantir o funcionamento correto das operações estatísticas implementadas também em métodos criado dentro da classe:

- *_validate_column(self, column)*: Verifica se a coluna cuja operação vai utilizar realmente existe.
- *_validade_numeric_column(self, column):* Verifica se a coluna utilizada para operação possui apenas valores numéricos.

Assim, cada operação foi definida como uma função pertencente à classe *Statistics*, nomeada por seu termo em inglês. A validação da existência da coluna é utilizada para todas as métricas e a validação de dados numéricos é utilizada para métricas cujo dado não poderia ser categórico ou de outro tipo, como média, mediana, variância, covariância e probabilidade condicional.

Para a implementação efetiva de cada operação, foram utilizadas estruturas básicas da linguagem Python, como loops e condicionais, além das estruturas citadas anteriormente, os dicionários e listas. Nenhuma biblioteca externa aos recursos nativos da linguagem foi utilizada.

Dessa maneira, para garantir o funcionamento das funções implementadas, testes foram feitos através de um script de testes automáticos (*tests.py*) fornecido pelo mentor da disciplina. O script instancia um objeto da classe *Statistics* com uma base de dados de 5 colunas com 20 registro. Cada coluna tem um tipo diferente de dados. Dessa forma, funções foram criadas para comparar os resultados obtidos através dos métodos criados por nós com os resultados esperados já calculados para aquele conjunto de dados.

4. Resultados e Discussão

Os resultados obtidos no que tange ao que era esperado após o processamento dos dados de testes foram satisfatórios, tivemos êxito em todos os testes de processamento e validação nos casos de inputs inadequados.

Considerações Finais

O desenvolvimento deste projeto demonstrou-se relativamente desafiador, principalmente no que se refere à compreensão completa das aplicações matemáticas e da compreensão da aplicabilidade dessas aplicações em cenários diversos. A implementação do código em si não nos demandou tanto esforço; porém, devido ao fato de que as implementações foram divididas igualmente entre os membros da equipe, algumas reutilizações poderiam ter sido feitas para melhorar a qualidade do código. Em detrimento da escassez de tempo dos colaboradores, outra questão colocada de lado foi a análise da complexidade de cada algoritmo. Em um cenário em que tivéssemos mais tempo, essa questão seria uma prioridade no momento do desenvolvimento.

7. References

Amazon Web Services, Inc. (2025) “O que é mineração de dados?”. Disponível em: https://aws.amazon.com/pt/what-is/data-mining/.

Bilche, R. (2023) “A importância dos dados na Era Digital”, Rem Soft Sistemas. Disponível em: https://remsoft.com.br/blog/tecnologias/a-importancia-dos-dados-na-era-digital/.

Estatística Fácil. (2025) “O que é frequência acumulada?”. Disponível em: https://estatisticafacil.org/glossario/o-que-e-frequencia-acumulada/.

Khan Academy. (2025) “Revisão: Média, mediana e moda”. Disponível em: https://pt-pt.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/mean-median-basics/a/mean-median-and-mode-review.

Pykes, K. (2023) “A importância dos dados: 5 principais motivos”, DataCamp. Disponível em: https://www.datacamp.com/pt/blog/importance-of-data-5-top-reasons.

Reis, T. (2021) “O que é covariância”, Suno. Disponível em: https://www.suno.com.br/artigos/covariancia/.

Rocha, A. (2025) “Frequência absoluta, relativa, acumulada e relativa acumulada”, Matemática Hoje. Disponível em: https://matematicahoje.blog/frequencia-absoluta-relativa-acumulada-e-relativa-acumulada/.

Waples, J. (2025) “Covariância: Compreendendo a relação entre variáveis”, DataCamp. Disponível em: https://www.datacamp.com/pt/tutorial/covariance.