Trabalhando com Estatística Básica – Implementação de Funções Estatísticas em Python

Lucas F. Lan[1], Herick Marcio M. Brito[2], Juary dos Santos L. Júnior[3]

[1]Mineração de Dados – Centro Universitário de Excelência (UNEX) Caixa Postal 44.085-370 – Feira de Santana – BA – Brasil

Lucaslan10@hotmail.com, herickmarcio356@gmail.com, juaryjuniorr@gmail.com

Abstract. *This paper addresses the implementation of statistical functions using the Python programming language. The work was carried out in the context of the Data Mining course, with the objective of applying fundamental statistical concepts and deepening the theoretical understanding of the main statistical measures through their practical application.*

Resumo. *Este artigo aborda a implementação de funções estatísticas utilizando a linguagem Python. O trabalho foi feito no contexto da disciplina de Mineração de Dados, com o objetivo de aplicar conceitos fundamentais de estatística e aprofundar o entendimento teórico das principais medidas estatísticas por meio de sua aplicação prática.*

1. Introdução

Vivemos na era digital, onde grandes volumes de dados são gerados por segundo através de sistemas digitais, redes sociais, sensores e transações comerciais. No entanto, essa quantidade massiva de dados brutos, por si só, não apresenta muito valor. A verdadeira importância está na capacidade de extrair informações e padrões a partir desses dados brutos. É neste contexto que a estatística se revela uma ferramenta fundamental.

Desse modo, a estatística desempenha um papel essencial na análise e interpretação de dados, pois permite transformar dados brutos em conhecimento útil para a tomada de decisões. O conhecimento de conceitos estatísticos básicos é fundamental em diversas áreas, uma vez que possibilita identificar padrões, realizar previsões e validar hipóteses.

Em síntese, esse relatório descreve o desenvolvimento de uma classe ‘Statistics’ para o cálculo de métricas básicas de estatística descritiva, com o objetivo principal de aprender e aplicar conhecimentos teóricos, por meio da implementação pura sem utilizar bibliotecas externas.

2. Fundamentação Teórica

As métricas estatísticas adotadas nesse trabalho são essenciais para a análise de dados. A seguir, são apresentadas as métricas utilizadas, juntamente com seu conceito, equações e comportamentos conhecidos.

2.1. Média Aritmética (Mean)

A média é uma medida de tendência central, ou seja, o valor central de um conjunto de dados. Ela é calculada pela soma de todos os valores dividida pelo número de observações, sendo muito sensível a valores extremos (outliers).

Média = $\sum X_i$ / n, onde (X_i) são os valores do conjunto de dados e (n) o número de observações

2.2. Mediana (Median)

A mediana é o valor que ocupa a posição central de um conjunto de dados, dividindo o conjunto em duas partes iguais, não sendo afetada por outliers. Para calcular a mediana é preciso ordenar os dados em ordem crescente ou decrescente, caso o número de observações seja ímpar, a mediana é o valor central, se for par, a mediana será a média dos dois valores centrais.

2.3. Moda (Mode)

A moda é o valor que mais ocorre em um conjunto de dados. A moda pode não ser única e não é afetada por outliers. Para descobrir a moda, basta contar a frequência de cada valor.

2.4. Desvio Padrão Populacional (Standard Deviation)

O desvio padrão mede a dispersão dos dados em relação à média. O desvio padrão pode ser calculado pelo quadrado da diferença entre cada ponto de dados e a média, em seguida, o somatório dessas diferenças é dividido pelo número de observações, após isso, é feito a raiz quadrada. O desvio padrão também pode ser calculado pelo quadrado da variância.

$\sigma = \sqrt{[\ \sum(xi - \mu)^2 / N\ ]}$

2.5. Variância Populacional (Variance)

A variância mede a dispersão dos dados em relação à média, ou seja, ela indica o quanto os valores de um conjunto se afastam do valor médio. Ela é calculada através do quadrado da diferença entre cada ponto de dados e a média, em seguida, o somatório dessas diferenças é dividido pelo número de observações. A variância quanto maior indicam maior dispersão e são sensíveis a outliers.

$\sigma^2 = \Sigma(x_i - \mu)^2 / N$

2.6. Covariância (Covariance)

A covariância mede como duas variáveis variam em conjunto, ou seja, se a covariância for positiva, quando um aumenta o outro tende a aumentar, enquanto negativa quando um aumenta o outro tende a diminuir. Para calculá-la basta calcular a média de cada conjunto e para cada par de valores, subtrair da média do seu conjunto e multiplicar o resultado dessas diferenças, a seguir some todos os resultados e divida por o número de observações.

$Cov(X, Y) = \sum(xi - \bar{X}) * (yi - \bar{Y}) / n$

2.7. Itens Únicos (Itemset)

Essa função retorna itens únicos do conjunto de dados. Sendo muito útil para identificação de inconsistências e anomalias, além de evitar processamento redundante.

2.8. Frequência Absoluta (Absolute Frequency)

A frequência absoluta é o número de vezes que cada valor aparece no conjunto de dados, para calculá-la basta contar a frequência de cada valor no conjunto de dados.

2.9. Frequência Relativa (Relative Frequency)

A frequência relativa é a proporção de vezes que cada valor aparece em relação ao número total de dados. Ela pode ser calculada pela divisão entre a frequência absoluta e o número total de elementos do conjunto de dados.

2.10. Frequência Acumulada (Cumulative Frequency)

A frequência acumulada pode ser absoluta ou relativa, ela é calculada pela soma das frequências (absolutas ou relativas) de um determinado valor com todas as frequências dos valores anteriores.

2.11. Probabilidade Condicional (Conditional Probability)

A probabilidade condicional mede a chance de um evento A acontecer dado que outro evento B já ocorreu. A probabilidade condicional é calculada pela probabilidade de ambos os eventos A e B ocorrerem juntos dividido pela probabilidade do evento B.

$P(A|B) = P(A \cap B) / P(B)$,

3. Metodologia

O projeto foi desenvolvido na linguagem Python orientado a objetos, onde implementamos os métodos da classe “Statistics”. Primeiramente, inicializamos o objeto Statistics, passando no construtor da classe verificações sobre o dataset recebido, como: se o dataset passado é um dicionário, se os valores no dicionário do dataset são listas e se todas as colunas do dataset tem o mesmo tamanho antes de armazenar o dataset na instância da classe.

Nos métodos, são feitas verificações de falhas com os dados para evitar resultados falsos. Desse modo, é feito validações se a coluna passada existe, além disso se as colunas não tiverem dados, o método retorna 0.0 como resultado imediatamente.

A classe conta com 11 métodos, cada um com sua implementação própria e autêntica seguindo suas fórmulas, portanto, algumas funções que chamam atenção são as de mediana, moda e itens único, assim, vamos comentar sobre os recursos utilizados. No método median, utilizamos um algoritmo de ordenação simples chamado de Bubble Sort, onde ele faz uma comparação de cada elemento com seu adjacente, se ele for maior, troca as posições, se não, passa sem alterações, logo em seguida é feito uma verificação com o operador módulo para verificar se o número de elementos da lista é par ou ímpar e calcular a mediana corretamente.

Ademais, no método de mode, percorremos o conjunto com um loop for e fazemos uma verificação se o item já foi contado ou não, para a contagem utilizamos a lógica de armazenar a frequência dos valores em um dicionário, onde a chave é o valor do conjunto e o valor é um inteiro que representa sua frequência, a partir disso verificamos qual o valor máximo que mais se repete, ou seja, qual a frequência máxima e verificamos quais valores tem essa frequência e retornamos em uma lista

Para finalizar nos itens únicos, utilizamos um conjunto, obtido a partir do método set() do Python, que remove duplicatas e retorna apenas itens únicos.

4. Resultados

A partir de um conjunto de dados conhecidos fornecidos no arquivo tests.py, foi feito o teste de todos os métodos implementados. A partir dos testes, todos os métodos, exceto o de probabilidade condicional, se comportaram da maneira esperada, passando por todos os testes e validações feitos.

O método de probabilidade condicional, apresentou um resultado diferente do esperado no teste, enquanto o teste esperava o resultado 0.5, o meu método implementado retornou o valor 0.5714285714285714, entretanto, o resultado esperado no teste consideramos apenas um arredondamento para um float com uma casa decimal, pois os outros testes de probabilidade passaram sem erros, portanto alteramos o valor do teste para 0.5714285714285714 para poder prosseguir com os outros testes adiantes.

5. Referências

MUNDO EDUCAÇÃO. *Frequência absoluta: o que é e como calcular?* Disponível em: https://mundoeducacao.uol.com.br/matematica/frequencia-absoluta.htm. Acesso em: 19 ago. 2025.

MUNDO EDUCAÇÃO. *Probabilidade condicional.* Disponível em: https://mundoeducacao.uol.com.br/matematica/probabilidade-condicional.htm. Acesso em: 19 ago. 2025.

DATA AVENUE – Desenvolvimento de Dados. *Principais métricas estatísticas utilizadas em ciência e análise de dados.* Disponível em: https://dataavenue.com.br/blog/principais-metricas-estatisticas-utilizadas-em-ciencia-e-analise-de-dados/. Acesso em: 23 ago. 2025.

MUNDO EDUCAÇÃO. *Frequência relativa: como calcular e exercícios.* Disponível em: https://mundoeducacao.uol.com.br/matematica/frequencia-relativa.htm. Acesso em: 23 ago. 2025.

APRENDER ESTATÍSTICA FÁCIL. *O que é: Métricas Estatísticas – Definição e Importância.* Disponível em: https://estatisticafacil.org/glossario/o-que-e-metricas-estatisticas-definicao-e-importancia/. Acesso em: 23 ago. 2025.