INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO

CAMPUS SÃO PAULO

YHASMIN SOUZA E SILVA

Relatório Final do Projeto de Estatística e Probabilidade

São Paulo - SP

2025

Yhasmin Souza e Silva

Relatório Final do Projeto de Estatística e Probabilidade

Trabalho apresentado à disciplina de Probabilidade e Estatística do Instituto Federal de Educação, Ciência e Tecnologia Campus São Paulo como nota parcial para aprovação na disciplina do curso tecnólogo em Análise e Desenvolvimento de Sistemas.
Professorª: Josceli Maria Tenorio

São Paulo – SP

2025

SUMÁRIO

Tema: Análise Integrada de Casos de Síndrome Gripal e Níveis de Cianobactérias em Municípios Paulistas (2024)

Introdução

Este projeto teve como finalidade realizar uma análise estatística e probabilística integrando duas bases de dados públicas: a primeira, referente às notificações de síndrome gripal no estado de São Paulo durante o ano de 2024; e a segunda, contendo os resultados de medições de cianobactérias em águas, conforme os registros do sistema SISAGUA. A proposta central foi investigar a possibilidade de existirem correlações temporais ou geográficas entre os indicadores ambientais e a incidência de doenças respiratórias. O estudo também buscou aplicar de forma prática os conceitos fundamentais trabalhados na disciplina, incluindo estatística descritiva, distribuições de probabilidade, teorema de Bayes e testes de hipótese.

Metodologia

O tratamento inicial das bases consistiu na padronização de colunas, conversão adequada de datas e eliminação de registros inconsistentes ou incompletos. A base de síndrome gripal foi normalizada com foco nas colunas "município", "data de notificação" e "classificação final", gerando um novo arquivo com os dados tratados. Já a base de cianobactérias teve atenção especial para a variável "Resultado", que indicava a concentração da substância, e foi convertida para tipo numérico após substituição correta de vírgulas por pontos.

Ambas as bases foram agrupadas por município e por ano-mês, o que possibilitou uma análise temporal agregada. Em seguida, realizou-se uma junção entre essas duas tabelas, resultando em uma base única denominada "base_agregada.csv", que combinava a média de concentração de cianobactérias com a quantidade de casos de síndrome gripal por município e período.

Os scripts foram desenvolvidos em Python, com uso das bibliotecas pandas, matplotlib, seaborn e scipy.stats, e organizados em arquivos distintos. O script tratamento_dados.py ficou responsável por toda a parte de pré-processamento e normalização. O grafico_ciano.py, grafico_sg.py e grafico_comparativo.py foram criados para gerar os gráficos necessários. Por fim, teste_hipotese.py reuniu os cálculos probabilísticos e testes de inferência estatística.

Resultados e Análises

No que diz respeito à síndrome gripal, observou-se uma distribuição assimétrica nos dados, com elevação significativa de notificações nos meses de abril e maio, possivelmente associada à sazonalidade e queda de temperatura. O gráfico de linha

temporal reforçou esses padrões, enquanto o boxplot revelou maior variabilidade nos municípios mais populosos. Já para os dados de cianobactérias, a maioria dos registros apresentou baixos níveis de concentração, com poucos outliers. A tentativa de aplicação de discretizações automáticas por quantis esbarrou na baixa variabilidade dos dados, que chegou a impedir o uso do qcut em alguns momentos. Mesmo assim, o histograma indicou uma leve assimetria à direita, com tendência à presença esporádica de valores elevados.

Quando as duas bases foram cruzadas, algumas coincidências de picos em determinados períodos chamaram a atenção, porém a correlação linear entre as variáveis foi baixa. Isso sugeriu que, apesar de compartilharem sazonalidade, não havia uma relação direta evidente entre os níveis de cianobactérias e os casos de síndrome gripal. Foi considerada a possibilidade de fatores ambientais mais amplos influenciaram ambas as variáveis de forma indireta, como temperatura, umidade e poluição.

Probabilidade e Inferência Estatística

Aplicou-se o Teorema de Bayes com base na categorização dos dados em níveis "alto" e "baixo", sempre que a distribuição permitiu esse tipo de classificação. A ideia era calcular a probabilidade de um município apresentar alto nível de cianobactérias, dado que apresentasse um número elevado de notificações de síndrome gripal. No entanto, os resultados não indicaram uma associação estatisticamente significativa entre os eventos, sugerindo independência probabilística.

Para confirmar esses achados, foi realizado um teste de hipótese do tipo qui-quadrado, com variáveis categorizadas previamente. Em geral, os testes não rejeitaram a hipótese nula de independência, o que significa que os níveis de cianobactérias e os casos de síndrome gripal não apresentaram dependência estatística no contexto das amostras analisadas. As condições para aplicação dos testes foram respeitadas, com frequência esperada adequada nas categorias.

Conclusão

Este projeto permitiu uma aplicação concreta de diversos conceitos de estatística e probabilidade, desde o pré-processamento dos dados até a análise inferencial. A ausência de relação estatística forte entre os indicadores analisados não desvaloriza o trabalho; pelo contrário, contribui para reforçar a importância de dados confiáveis e do uso rigoroso das ferramentas estatísticas para embasar conclusões. As bases de dados, apesar de públicas e valiosas, apresentaram algumas limitações, como registros ausentes em certos municípios e concentrações estáticas em parte dos dados.

Os scripts desenvolvidos são reutilizáveis e servem como base para estudos futuros, inclusive com outros conjuntos de dados da área da saúde pública ou do meio ambiente. A análise pode ser ampliada futuramente com uso de regressões múltiplas ou técnicas de machine learning, e o cruzamento com dados climáticos e socioeconômicos também pode revelar padrões mais complexos.

Apêndices

Os arquivos do projeto foram organizados conforme a estrutura abaixo:

- Dados tratados: data/processed/*.csv
- Scripts utilizados: scripts/*.py
- Gráficos gerados: imagens/*.png
- Repositório GitHub: https://github.com/Yhasmin01/projeto_esta_yhasmin.git