INTERN VISION

MINOR PROJECT

GENOTYPE-PHENOTYPE STATISTICAL ANALYSIS

Internship Name / Organization: Intern Vision

DOMAIN: DATA SCIENCE

Core Focus: Data Science Disciplines (Intro), Pandas, Descriptive Statistics, Probability, T-Test, Chi-Square

Project Goal: Analyze a simulated dataset of patient genotypes and phenotypes to calculate statistical relationships and test simple genetic hypotheses.

SUBMITTED BY: GOLLA NIKITHA

1. Introduction

This minor project focuses on the statistical analysis of a simulated genotype–phenotype dataset using fundamental data science and statistical techniques. The objective of the project is to understand the relationship between genetic variations (genotypes) and observable traits (phenotypes) in patients. A synthetic dataset containing patient information such as genotype, phenotype status, treatment response time, and gender was created to mimic real-world genetic data.

Using the Pandas library, descriptive statistical measures such as mean, median, and standard deviation were calculated to summarize treatment response times across different genotypes. Probability concepts were applied to estimate the likelihood of patients being affected by a simulated genetic disorder, as well as conditional probabilities based on specific genotypes. To further validate genetic hypotheses, inferential statistical tests were performed. A T-Test was used to compare treatment response times between affected and healthy individuals, while a Chi-Square test was conducted to examine the association between genotype carrier status and phenotype.

This project demonstrates the application of basic data manipulation, probability theory, and statistical inference techniques in genetic data analysis, providing a strong foundation for more advanced data science and bioinformatics applications.

2. Objectives

1. **To create and analyze a simulated genotype–phenotype dataset** representing a recessive genetic disorder using structured patient information such as genotype, phenotype status, treatment response time, and gender.

2. **To apply Pandas-based data manipulation and descriptive statistical techniques** to summarize and compare treatment response times across different genotypes.

3. **To compute probability and conditional probability measures** to estimate the likelihood of individuals being affected by the disorder and to assess risk associated with specific genotypes.

4. **To perform inferential statistical testing using a T-Test** in order to determine whether there is a significant difference in treatment response time between affected and healthy patients.

5. **To evaluate the association between genotype carrier status and phenotype** by constructing a contingency table and performing a Chi-Square test for statistical significance.

3. Dataset Description

The dataset used in this project is a **simulated genetic dataset** created to analyze genotype–phenotype relationships using statistical methods. Since real patient genetic data is sensitive and not always publicly available, a synthetic dataset was generated to closely resemble real-world genetic and clinical data for learning and analysis purposes. The dataset contains **more than 100 records** and is stored in **CSV (Comma-Separated Values) format**.

Each row in the dataset represents an individual patient, and the dataset includes the following attributes:

- **PatientID:** A unique identifier assigned to each patient.
- **Genotype:** Represents the genetic makeup related to a simulated recessive disorder, categorized as **AA**, **Aa**, or **aa**.
- **Phenotype:** Indicates the health status of the patient, where **0 = Healthy** and **1 = Affected**.
- **Treatment_Response_Time:** A numerical variable representing the number of days taken by a patient to respond to treatment.
- **Gender:** A categorical variable indicating the patient's gender as **Male (M)** or **Female (F)**.

The dataset was synthetically generated to simulate real-world genetic data for analysis purposes and to enable the application of descriptive statistics, probability calculations, and inferential statistical tests.

4. Tools and Technologies Used

The following tools and technologies were used to successfully complete this minor project on genotype–phenotype statistical analysis. These tools enabled efficient data creation, data manipulation, statistical computation, and interpretation of results.

Python

Python was used as the primary programming language for this project due to its simplicity, flexibility, and strong support for data analysis and statistical computing.

Jupyter Notebook

Jupyter Notebook was used as the development environment to write, execute, and document the Python code in an interactive manner, allowing clear visualization of outputs and results.

Pandas

The Pandas library was used for loading the simulated dataset, performing data manipulation, grouping data, and calculating descriptive statistics such as mean, median, and standard deviation.

NumPy

NumPy was utilized for efficient numerical operations and handling arrays required for statistical calculations.

SciPy

The SciPy library was used to perform inferential statistical tests, including the T-Test and Chi-Square Test, to analyze genetic associations and differences in treatment response.

5. Methodology

Step 1: Dataset Creation

A simulated genotype–phenotype dataset was created in CSV format to represent patient genetic information. The dataset consisted of more than 100 records and included attributes such as Patient ID, genotype, phenotype status, treatment response time, and gender.

Step 2: Data Loading and Preparation

The dataset was loaded into a Pandas DataFrame using Python. Initial data inspection was performed to verify data structure, data types, and completeness, ensuring that the dataset was suitable for analysis.

Step 3: Descriptive Statistical Analysis
Descriptive statistical measures, including mean, median, and standard deviation, were calculated for treatment response time. These statistics were grouped based on genotype to identify variations across different genetic categories.

Step 4: Probability Analysis
Probability calculations were performed to determine the overall probability of patients being affected by the disorder. Conditional probability was also computed to estimate the likelihood of being affected given a specific genotype (aa).

Step 5: Inferential Statistical Testing
A T-Test was conducted to compare the mean treatment response times between affected and healthy patients. Additionally, a Chi-Square test was performed by constructing a contingency table between genotype carrier status and phenotype to assess statistical association.

Step 6: Result Interpretation
The p-values obtained from the statistical tests were analyzed to determine significance. Based on these results, conclusions were drawn regarding the relationship between genotype and phenotype.

6. Results and Observations

The analysis of the simulated genotype–phenotype dataset provided meaningful statistical insights into the relationship between genetic makeup and treatment outcomes.

The descriptive statistical analysis showed noticeable differences in the **mean treatment response time** across different genotypes. Patients with the *aa* genotype generally exhibited higher treatment response times compared to *AA* and *Aa* genotypes, indicating a possible genetic influence on treatment effectiveness.

The **probability analysis** revealed the overall probability of patients being affected by the simulated disorder. Additionally, the **conditional probability of being affected given the genotype aa** was observed to be higher compared to other genotypes, supporting the assumption of a recessive genetic pattern.

The **T-Test** results showed a statistically significant difference in mean treatment response time between affected and healthy patients, with the p-value being less than 0.05. This indicates that the treatment response time varies significantly based on the phenotype status.

The **Chi-Square test** demonstrated a statistically significant association between genotype carrier status and phenotype, suggesting that genetic factors play an important role in determining the occurrence of the disorder.

7. Conclusion

This minor project successfully demonstrated the application of fundamental data science and statistical techniques to analyze a simulated genotype–phenotype dataset. By using synthetic genetic data, the project effectively modeled real-world scenarios where patient genetic variations influence observable traits and treatment outcomes. The use of a simulated dataset ensured ethical handling of data while still allowing meaningful statistical analysis.

Through descriptive statistical analysis, variations in treatment response time across different genotypes were clearly observed. The probability calculations provided valuable insights into the likelihood of patients being affected by the simulated recessive disorder, as well as the increased risk associated with specific genotypes. These results helped reinforce basic genetic principles through quantitative analysis.

Inferential statistical methods further strengthened the findings of the study. The T-Test revealed a statistically significant difference in treatment response times between affected and healthy patients, indicating that phenotype status plays an important role in treatment effectiveness. Additionally, the Chi-Square test confirmed a significant association between genotype carrier status and phenotype, highlighting the influence of genetic factors on disease occurrence.

Overall, this project provided hands-on experience in data manipulation, probability analysis, and hypothesis testing using Python libraries such as Pandas and SciPy. It helped build a strong foundation in statistical inference and data-driven decision-making, which are essential skills in data science and bioinformatics. The methodologies and insights gained from this project can be extended to more complex genetic datasets and advanced analytical techniques in future studies.