Genome Wide Association Study (GWAS)
Student Name: Prit Desai
UTA ID: 1002170533
Course: CSE 5370 - Bioinformatics
Assignment: Homework 1

1.2 Fisher's Exact Test

The null hypothesis (H) for the Fisher's Exact Test states that there is no association between the presence of the C-allele at a given SNP and the complex genetic trait being studied. In other words, the odds ratio between case and control groups for the C-allele is equal to 1. Rejecting this null hypothesis suggests that the SNP may be associated with the trait.

Choice of Alternative Hypothesis: The test uses the alternative="two-sided" option because the analysis aims to detect any association between SNP presence and the trait, whether positive or negative. This ensures that SNPs with either increased or decreased associations are identified.

Number of Significant SNPs:

- **Significant SNPs under Original Threshold (5×10):** 11
- **Significant SNPs under Bonferroni Correction (5×10^{11}):** 3

1.3 Bonferroni Correction

The Bonferroni correction addresses the multiple hypothesis testing issue by adjusting the significance threshold. Since 1000 SNPs are tested, the corrected p-value threshold is:

$$\frac{5 \times 10^{-8}}{1000} = 5 \times 10^{-11}$$

This correction reduces the likelihood of Type I errors but increases the chance of Type II errors, making the test more conservative and resulting in fewer significant SNPs compared to the original threshold.

1.4 Manhattan Plot Interpretation

The Manhattan plot visualizes the -log(p-values) of each SNP locus. Higher points indicate stronger evidence against the null hypothesis.

Key Observations:

- **Red dashed line:** Original significance threshold (5×10)
- **Blue dashed line:** Bonferroni-corrected threshold (5×10^{11})
- **Points above the lines:** Statistically significant SNPs
- A cluster of significant SNPs between loci 450 and 470 indicates a potential genomic region associated with the trait.

[width=0.8]manhattan$_p lot.png$

Figure 1: Manhattan Plot for GWAS showing the SNPs with their p-values. The red and blue dashed lines represent the original and Bonferroni-corrected significance thresholds, respectively.

Files Included in Submission

- 1002170533.csv: Generated dataset
- results.csv: SNP names, p-values, and significance indicators
- manhattan_plot.png: Visualization of GWAS results
- GWAS_Writeup.pdf: This document
- readme.txt: Instructions to run the code
- Source Code:
  - datasetGenerator.py: Dataset generation
  - HW1.py: GWAS analysis (Fisher's test and Manhattan plot)

Instructions to Run Code

1. Install Python (version 3.8) and required libraries: numpy, scipy, matplotlib.
2. Activate the virtual environment:

   ```
   .\venv\Scripts\Activate
   ```

3. Run the dataset generator:

   ```
   python datasetGenerator.py --ID 1002170533
   ```

4. Run the analysis script:

   ```
   python HW1.py
   ```

5. Outputs:
   - results.csv: Contains p-values and significance flags
   - manhattan_plot.png: Displayed and saved automatically

Difficulty & Time Spent

- **Estimated Time:** 10 hours
- **Challenges:** Understanding the Bonferroni correction and interpreting the plot required extra time.