CHL5224 Final Project:

Over-fitting, Subtle Multiple Hypothesis Testing and Possible Remedies

Hainan Xu

December 18, 2023

Contents

1 Introduction

Genome-wide association studies (GWAS) [5] aim to identify the associations of genotypes with phenotypes. In GWAS, researchers conduct hundreds of thousands of tests on genetic variants across multiple genomes to identify statistically significant associations with specific traits or diseases. However, when performing genome-wide association studies, performing multiple hypotheses simultaneously often leads to increase in the probability of false positives, which in turn affects the validity of the inferences based on the hypothesis tests. Therefore, an appropriate way of correcting the false positives that occurred in the analyses is needed.

The false positive rate referred to the type I error rate as well, which is defined as the probability of rejecting the null hypothesis when it is actually true.

In this project, we explore the behaviour of p-values under a subtle multiple hypothesis scenario and aim to find a possible remedy to adjust the type I error rate resulting from multiple hypothesis testing.

2 Methods

2.1 The Data

The data being analyzed is simulated under the null hypothesis, which means no genotype (G) has a significant effect on phenotype Y. In each interaction, the continuous phenotype Y is generated from a normal distribution with a mean of 170 and a variance of 49. The genotypes are generated under the Hardy-Weinberg Equilibrium assumption with a minor allele frequency of 0.2. The sample size of each dataset generated is 1000. We run 10,000 simulations in total.

2.2 Four Genetic Models and the Minimum P-value Approach

Suppose there are two alleles at an SNP locus, A and a. Specifically, we are interested in studying type I errors that occurred in multiple hypothesis testing with the minimum p approach, where the minimum p approach refers to taking the minimum p-value of all the p-values obtained from four different genetic models. The four different genetic models refer to additive, dominant, recessive and genotypic, and each of them is defined as follows:

Additive Model: We suppose individuals with genotype AA have a higher trait value than Aa, and Aa individuals have a higher trait value than aa in the additive model.

Dominant Model: We assume A is dominant to a, and therefore Aa and AA have the same trait value, and aa has a different and smaller trait value compared with Aa and AA.

Recessive Model: We assume A is recessive to a. Therefore, Aa and aa have the same trait value, and AA has a different and smaller trait value compared to Aa and aa.

Genotypic Model: We assume that the trait values are distinct for each genotype. Individuals with genotype AA, Aa, and aa have different trait values.

The type I error rate α is defined as:

$$\alpha = P(\text{reject } H_0 | H_0 \text{ is true})$$

Therefore, to assess the type I error, we generate datasets under the null hypothesis. Then, we fit four different genetic models and obtain the p-values for each model. The type I error occurs when the p-value falls below the prespecified significance level. For each iteration, we return the p-value that is the smallest among the four models.

2.3 Type I Error Correction Methods

We consider two different type I error correction methods. One is by changing the significance level. The other one is by adjusting the p-values (or the significance level) using Bonferroni correction [3].

2.3.1 Adjusting Significance Level

Under the significance level of 0.05, we identify an inflated Type I error when it exceeds 5%. To address this inflation, we establish a new significance level, ensuring that only 5% of the observations are erroneously rejected. In this project, the new significance level is determined by ranking the p-values from the lowest to the highest, with the position of the new significance level corresponding to the value associated with the 5% threshold in the dataset.

2.3.2 Bonferroni Correction

Bonferroni corrections is a corrections that made to p values when several dependent or independent statistical tests are being performed simultaneously on a single dataset. To perform the Bonferroni correction, the new significance level (α) is derived by the original significance level divided by the number of comparisons being made. For example, in this project, n hypotheses are being tested, then the new significance level would be

$$\alpha_{new} = \frac{\alpha_{old}}{\text{n}} = \frac{0.05}{4} = 0.0125$$

Instead of adjusting significance levels, one can also consider adjusting the p-values, which is simply multiplying the current p-values by the number of tests being performed, if the outcome p-value exceeds 1 then cap the p-value at 1. The resulting type I error is the same as adjusting the significance level using Bonferroni correction.

2.3.3 Other Methods

We also consider other methods that correct the family-wise error rate (FWER) [6] and family discovery rate (FDR), For example, Holm's [4] method that corrects the FWER. The Holm's method dominates the Bonferroni correction, and is also valid under arbitrary assumptions. For this specific problem, the result of Holm's method is the same as the Bonferroni correction.

Methods implemented to correct FDR are also considered. Compared with FWER, the FDR is a less stringent condition. So the methods that are applied to FDR are more powerful than the FWER ones. For example, Benjamini & Hochberg's method [2] and Benjamini, Hochberg, and Yekutieli's method [1].

3 Results

With the p-values obtained from the minimum p-value approach, we observe an inflated type I error rate of 10.89% under the significance level of 0.05.

Implementing the method described in 2.3.1, we can obtain a new significance level. With the new significance level of 0.022, the type 1 error rate is reduced from 10.89% to exactly 5%.

Figure 1 shows the distribution of unadjusted p-value obtained using the minimum p-value approach. The distribution is not uniform, and we observe decreasing frequencies as the p-value increases. The histogram strengthens the fact that the type I error is inflated with the stated minimum p-value approach. In addition, the red line represents the unadjusted significance level ($\alpha = 0.05$), and the green line represents the adjusted significance level ($\alpha \approx 0.022$).

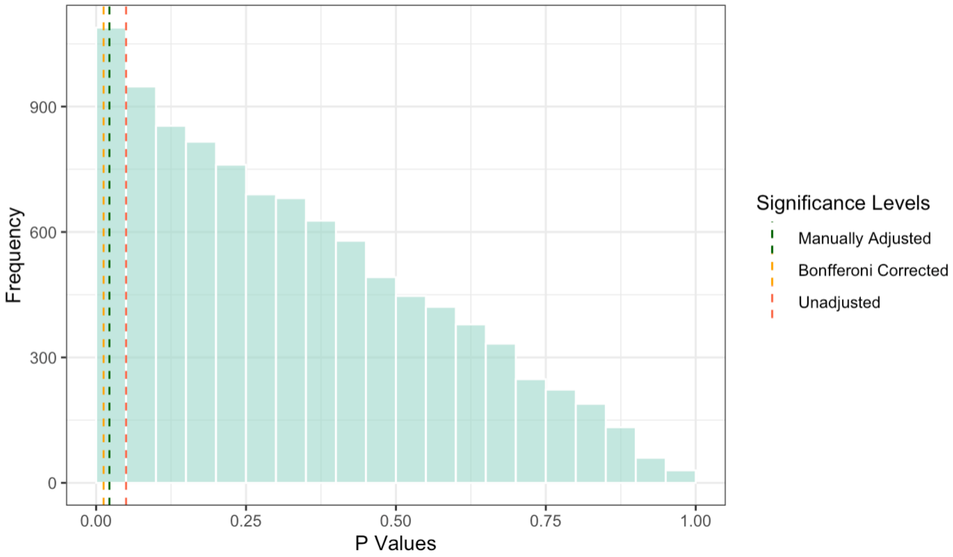

Figure 1: Histogram of P-values with Different Thresholds

Using the Bonferroni correction method mentioned in 2.3.2, we can obtain a significance level of 0.0125, which is denoted as an orange dashed line in figure 1. With the Bonferroni adjusted significance level, the type I error reduced from 10.89% to 2.93%, which is within the desired 5% type I error rate. The top left histogram in Figure 2 displays the distribution of adjusted p-values using Bonferroni correction. After the corrections, most p-values are clustered around 1. The type I error rate is 2.93%.

Figure 2 shows the different methods of correcting p-values. The red dashed line represents the significance level of $\alpha = 0.05$. The Holm corrected p-values are the same as the Bonferroni corrected p-values, resulting the same type I error, 2.93%. Under the same significance level, Benjamin & Hochberg's method gives a type I error rate of 3.58%, whereas the Benjamini, Hochberg, and Yekutieli's method gives a type I error rate of 1.8%.

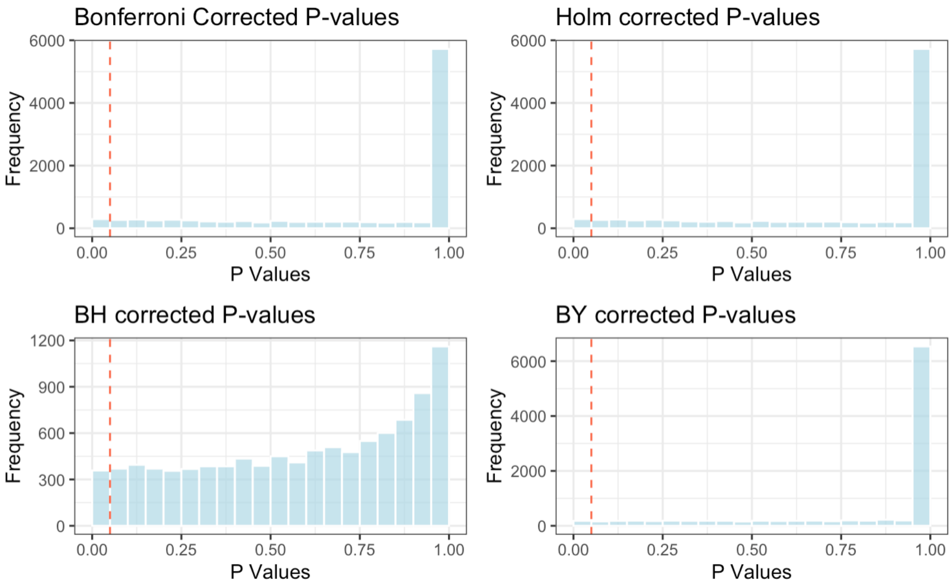

Figure 2: Histogram of Corrected P-values with Different Correction Methods

4 Conclusion and Discussion

In this project, we identified and corrected inflated Type I errors in a specific multiple hypothesis testing scenario, namely, the minimum p-value approach. We detected the inflated Type I errors through an empirical distribution of p-values. Subsequently, the Type I error rate was corrected by adjusting either the significance level or the p-values using multiple methods.

However, it is worth noticing that there is a trade-off between Type I error and Type II error. Therefore, while the methods implemented in this project effectively reduced Type I error, they may lead to an inflation in Type II error.

The type I error analysis in this project provides a foundation for future analysis for the minimum p-approach. Further analyses should be undertaken to conduct a thorough power analysis, thereby enhancing our understanding of the statistical properties and performance of the minimal p approach and corresponding correction methods.

5 Appendix - Checking the Uniformity of the P values

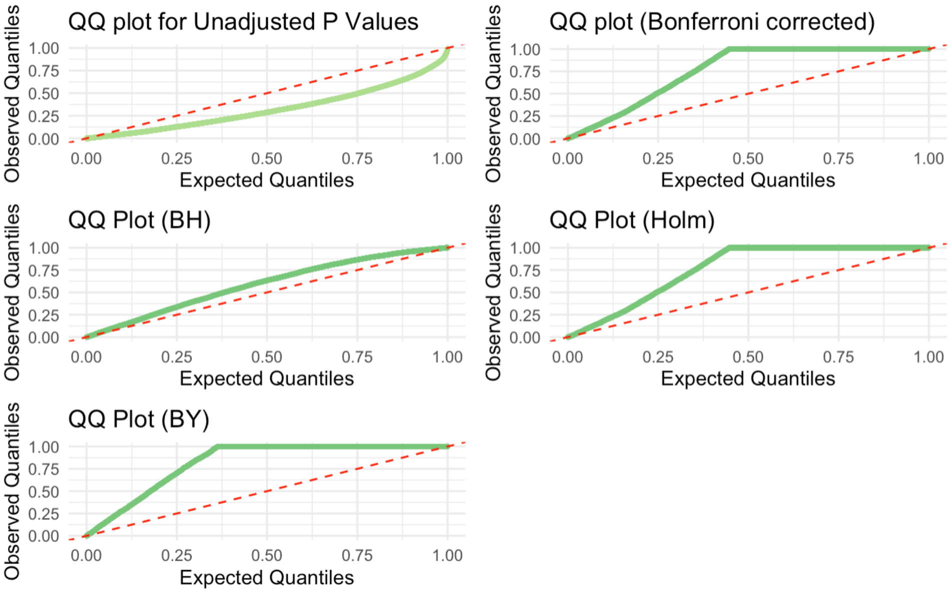

Figure 3: Uniform QQ plots of P-values btained from Different Methods

References

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[2] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[3] William S. Bush and Jason H. Moore. Chapter 11: Genome-Wide Association Studies. *PLoS Comput Biol*, 8(12):e1002822, December 2012.

[4] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[5] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nat Rev Methods Primers*, 1(1):1–21, August 2021. Number: 1 Publisher: Nature Publishing Group.

[6] Qiong Yang, Jing Cui, Irmarie Chazaro, L. Adrienne Cupples, and Serkalem Demissie. Power and type I error rate of false discovery rate approaches in genome-wide association studies. *BMC Genetics*, 6(1):S134, December 2005.