Multiple Testing and Error Control: A Short Overview

Elly Lee

Introduction

This Directed Reading Program (DRP), **Multiple Testing**, was mentored by Ethan Ancell and Kayla Irish. The project centered on understanding how standard statistical guarantees—such as control of Type I error—break down when an analyst considers many statistical inferences simultaneously, and how the field of multiple testing addresses this challenge. In modern applications ranging from genomics to social science, researchers routinely test dozens, thousands, or even millions of hypotheses at once. Without appropriate adjustment, classical hypothesis testing procedures lead to inflated false positive rates and poor replicability.

Throughout the quarter, we read and discussed both foundational ideas and modern research papers in multiple testing, with optional opportunities to explore simulations and computational aspects. While the overall theme of the DRP was broad, each mentee was free to choose a specific focus for their presentation. My presentation—and this writeup—focuses on a high-level overview of what multiple testing is, why it matters, and how common error control methods such as FWER and FDR are defined and achieved.

Why Multiple Testing Is a Problem

When many hypotheses are tested at once, even a small per-test error rate accumulates. For example, if 20 independent null hypotheses are all true and each is tested at level $\alpha = 0.05$, the probability of making *at least one* false rejection is already about 64%. This effect undermines the reliability and replicability of scientific findings if no adjustment is made. Therefore, valid inference in modern data analysis requires error control that accounts for the full *family* of hypotheses being tested.

Error Rates: FWER and FDR

Two major frameworks are commonly used to formalize error control in multiple testing:

- **Family-Wise Error Rate (FWER)** is the probability of making at least one false rejection among all tests. Procedures that control FWER are conservative but appropriate in settings where even a single false positive would be costly, such as confirmatory clinical trials.
- **False Discovery Rate (FDR)** is the expected proportion of false rejections among all rejected hypotheses. FDR control is less stringent and allows for higher power, making it particularly useful in large-scale exploratory studies such as gene expression analysis.

Understanding the difference between these two goals is essential for choosing an appropriate multiple testing method.

Methods for Controlling FWER

The simplest FWER-controlling method is the **Bonferroni correction**, which divides the desired significance level α by the number of tests m. A hypothesis is rejected only if its p-value satisfies $p \leq \alpha/m$. While Bonferroni control is valid under any dependence structure, it is often overly conservative.

More powerful alternatives include stepwise procedures. **Holm's step-down procedure** orders p-values from smallest to largest and compares them to increasingly relaxed thresholds. It uniformly improves upon Bonferroni while still controlling FWER under arbitrary dependence. **Hochberg's step-up procedure** is even more powerful but requires additional assumptions, such as independence or positive dependence among tests.

Methods for Controlling FDR

The most widely used FDR-controlling method is the **Benjamini–Hochberg (BH) procedure**. BH orders p-values and compares them to adaptive thresholds of the form $(i/m)\alpha$. All hypotheses with sufficiently small p-values are rejected. Unlike FWER methods, BH allows some false positives but guarantees that, on average, their proportion among discoveries is bounded by α. This balance between error control and power explains why BH has become a standard tool in modern statistical practice.

Conclusion

Multiple testing is unavoidable in contemporary data analysis, and ignoring it leads to misleading and irreproducible results. Through this project, I learned how different error rates—FWER and FDR—capture distinct scientific priorities, and how classical and modern procedures address these goals. Bonferroni, Holm, and Hochberg provide strong guarantees against false positives, while Benjamini–Hochberg offers a powerful and flexible alternative when large numbers of hypotheses are tested. Together, these methods form a core toolkit for responsible statistical inference in the presence of multiplicity.