1 Multiple Testing

Often, we want to test multiple hypotheses from a dataset, rather than one hypothesis.

In the Neyman-Pearson framework, for a hypothesis, we are controlling the Type-I error rate $\mathbb{P}_{H_0}(X \in R)$ to be less than or equal to α. If we test multiple hypotheses, each at level α, then still each one individually will have Type-I error rate less than or equal to α. The problem is that with multiple hypotheses, the probability of *more than one error* will in general be higher than α.

This can be handled by controlling the *combined* error rate among the multiple tests. Let $H_1, \ldots, H_m$ be a family of m hypotheses. We can summarize the possibilities for rejections/"acceptances" of the hypotheses as follows:

	Declared Non-Significant	Declared Significant	Total
True Null Hyp.	U	V	m_0
False Null Hyp.	T	S	$m - m_0$
	$m - R$	R	m

In this course, we will cover two notions of controlling the error.

- **Family-wise Error Rate (FWER)**: probability of a single false rejection

$$FWER = \mathbb{P}(V > 0).$$

- **False Discovery Rate (FDR)**: expected proportion of rejections that are false

$$\mathbb{E}(V/R).$$

Here are the most popular methods to correct each one.

- **Bonferroni's Correction**: Simply divide α by the number of hypotheses m, and reject hypothesis j if its p-value $p_j \leq \alpha/m$.
- **Benjamini-Hochberg Algorithm** Benjamini and Hochberg invented FDR and also created the following algorithm that controls it:
  (a) Order the p-values $p_{(1)} < \ldots < p_{(m)}$.
  (b) Select $J = \max\{j : p_{(j)} < \frac{j\alpha}{m}\}$, when the p-values are independent.
  (c) Reject all null hypotheses for which the p-value $\leq p_{(J)}$.

2 Replicability Crisis in Science

In many studies, the hypotheses are set up so that it is preferable to reject H_0 in favor of H_1. For example, a medical researcher could be testing

$$H_0: \text{ their new drug has the same effect as the existing drug}$$
$$H_1: \text{ their new drug has a better effect than the existing drug.}$$

Other examples are where H_1 might be considered the "surprising" or "interesting" result, e.g.,

$$H_0: \text{ there is no association between musical genre preference and personality}$$
$$H_1: \text{ there is an association between musical genre preference and personality.}$$

This can lead to a bias in publishing, in that "negative" results (insignificant results) may not be published. This can also lead to incorrect uses of hypothesis testing (whether intentional or unintentional), such as p-hacking, data dredging/snooping, etc. Here is a non-exhaustive list of incorrect uses:

- Looking at the data before proposing a hypothesis test
- Performing a hypothesis test at level α, failing to reject it, then changing the significance level to be higher
- Performing a large number of tests until you find something significant and interesting, then only reporting the one test and (unadjusted) p-value
- Describing the results in a misleading way, e.g., incorrectly claiming causality or using tricky wording like, "the results are approaching significance"

Sometimes these invalid results are reported on by news outlets. It is also possible that a study with valid research methods is misrepresented by news outlets.

In general, it is good if a study's results can be replicated – if so, there is more evidence that the result is not occurring by random chance. In recent history, there have been growing concerns in all fields of science about replicability. If a large number of studies are not replicable, then it seems to call into question the integrity of science.

About a decade ago, it was found that a large number of scientific results were *not* replicable. This was dubbed the "reproducibility crisis". There were a number of papers redacted for over-confident reporting of results, p-hacking, and other misuses of hypothesis testing. Some of these cases were quite high-profile. Here are a few examples of the erroneous, non-replicable findings:

- As discussed by Benjamini (2020)[1], a study sought to identify associations between risk of prostate cancer and more than a hundred different types of food. Their conclusion was that pizza and tomato sauce are protective against prostate cancer.
- One of Gelman's blog posts[2] discusses several examples. Probably the most well-publicized example is of the "power-pose". The claim was that striking a "powerful" posture before, e.g., an important meeting, can lead to improved outcomes.
- Another example from Gelman is of a paper that claimed that ovulation leads to differences in women's political belief, voting patterns, and religiosity.

In response to the replicability crisis, many researchers proposed different methods to improve the quality of research. Here are a few examples:

- Greater education on hypothesis testing and statistical procedures in general
- Pre-registration of studies
- Greater transparency in describing all of one's methods in a study – this might involve sharing code too
- Allowing negative results to be published in some journals
- Collectively, changing the "publish or perish" culture

1. `https://hdsr.mitpress.mit.edu/pub/l39rpgyc/release/3`
2. `https://statmodeling.stat.columbia.edu/2016/09/21/what-has-happened-down-here-is-the-`

3 Example of Multiple Testing & Correction in R

See R code.