DS 102 Data, Inference, and Decisions **Spring 2020**

Lecture 4: False Discovery Rate Control

Lecturer: Moritz Hardt

4.1 Multiple Hypothesis Testing

Multiple hypothesis testing is a problem that arises when we want to test many hypotheses while controlling an appropriate error rate. The Neyman-Pearson testing framework does not give satisfactory guarantees when applied to many hypotheses. In particular, controlling the probability of (falsely) discovering a null under α is vacuous if there are many tested nulls. For example, if there are m of them, the probability of falsely discovering at least one null might be as high as $m\alpha$! This is a useless guarantee if m is large.

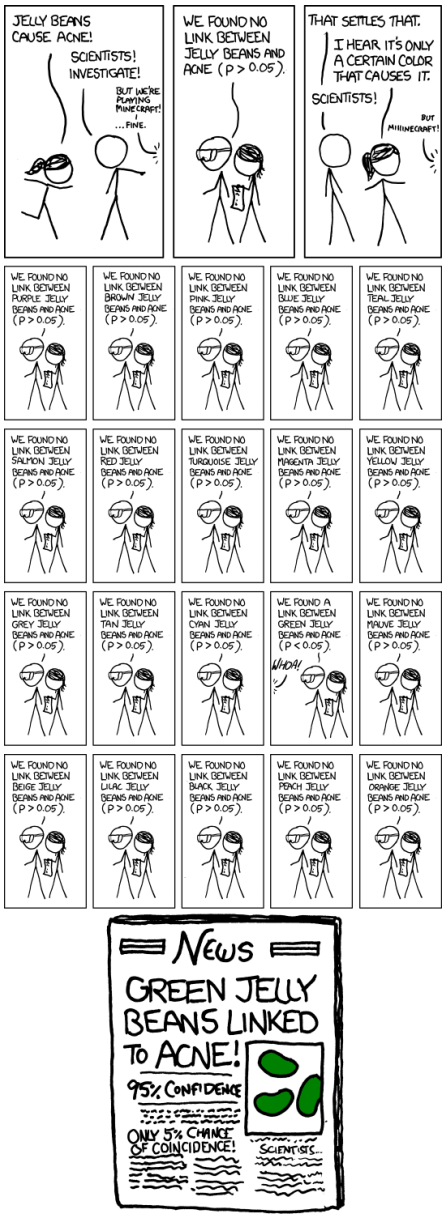

Figure 4.1: The multiple testing problem.

4.2 Reproducibility Crisis

The reproducibility crisis is one poignant example of the importance of controlling for false discoveries in multiple hypothesis testing. "Reproducibility crisis" refers to the increasing concern that many published scientific findings may actually be false. Over the last couple of decades, metastudies demonstrating an inability to replicate scientific studies have become more common. One such metastudy found that most research claims are more likely to be false than true (Ioannidis, 2005).

The reproducibility crisis is intimately linked to the problem of multiple hypothesis testing. For example, the term "p-hacking" derives from p-values resulting from hypothesis testing, and has recently gained prevalence to describe the phenomenon in which a scientist tests many hypotheses in sequence until obtaining a "significant" p-value, without reporting or adjusting for all of the preceding hypotheses. Indeed, a 2012 metastudy by Masicampo & Lalande found a "peculiar prevalence of p values just below .05." This phenomenon is likely influenced by p-hacking, as well as the fact that scientific journals may be less likely to consider results with p-values above 0.05 for publication.

Although many factors contribute to the reproducibility crisis, hypothesis testing plays an important part. Despite this controversy, hypothesis testing can be a useful tool when well-understood and carefully applied. Today, we will focus on proper analysis of multiple hypothesis tests.

4.3 Recap of P-values

In hypothesis testing we often use **p-values** to guide our decisions. The p-value is the probability that the **hypothetical data from the null hypothesis** would be equal to, or more extreme than, the actual observed samples. This concept should be familiar to you from previous classes, but we review it here for convenience.

Let $\mathcal{P}_\theta$ be shorthand for the likelihood under θ. Further, $\mathcal{S} \sim \mathcal{P}_\theta$ be the observed data set, and let $T(\mathcal{S})$ be some real-valued summary statistic obtained from $\mathcal{S}$. For example, $\mathcal{S}$ could be a set of n i.i.d. samples $X_1, \ldots, X_n$, and $T(\mathcal{S})$ could be their average $\frac{1}{n}\sum_{i=1}^n X_i$.

Let the null hypothesis be $\theta \in \Theta_0$, and let the alternative be $\theta \in \Theta_1$. Suppose that for every $\theta_0 \in \Theta_0$, we can compute the distribution of $T(\mathcal{S}_0)$, where $\mathcal{S}_0 \sim \mathcal{P}_{\theta_0}$. Then, the p-value is defined as:

$$P := \sup_{\theta_0 \in \Theta_0} \mathbb{P}(T(\mathcal{S}_0) \geq T(\mathcal{S}) \mid T(\mathcal{S})),$$

where the "hallucinated" sample $\mathcal{S}_0$ is independent from $\mathcal{S}$. We proclaim a discovery if P is less than or equal to the significance level α.

Denote by $F(\cdot)$ the *tail* CDF of $T(\mathcal{S}_0)$:

$$F(t) = \mathbb{P}(T(\mathcal{S}_0) > t).$$

Notice that the p-value can be written as

$$P = F(T(\mathcal{S})),$$

where $T(\mathcal{S})$ is the test statistic computed from data.

Going forward, we will use a few basic properties of p-values. First, they take values in $[0, 1]$, which is not surprising given that they are a probability. Second, if the null is true, they are *uniformly distributed*:

$$\mathbb{P}(P \leq u) = \mathbb{P}(F(T(\mathcal{S})) \leq u) = \mathbb{P}(T(\mathcal{S}) > F^{-1}(u)) = F(F^{-1}(u)) = u, \text{for all } u \in [0, 1],$$

where we assume certain regularity condition, like F is invertible, and $T(\mathcal{S})$ has a continuous distribution. The second equality follows because F^{-1} is a monotone decreasing transformation, due to F being decreasing. The second to last equality uses the assumption that the ground truth is null. If the alternative is true, on the other hand, we do not know anything about the distribution of P, but it is reasonable to expect P to be more "biased" toward smaller values.

4.4 Bonferroni Correction

The initial approach to multiple testing was to control the probability of making at least one false discovery, also known as the **family-wise error rate** FWER:

$$\text{FWER} := \mathbb{P}(\text{at least one false discovery is made}).$$

A simple procedure that ensures FWER control is the **Bonferroni correction**. Instead of testing each hypothesis under level α, the idea is to test each hypothesis under level α/N, where N is the total number of hypotheses. Denote by $\{E_i = 1\}$ the event that a false discovery is made on the i-th test. By a union-bound argument, it follows that the FWER is controlled under α:

$$\text{FWER} = \mathbb{P}(\cup_{i=1}^{N}\{E_i = 1\}) \leq \sum_{i=1}^{N} \mathbb{P}(E_i = 1) \leq \sum_{i=1}^{N} \alpha/N = \alpha,$$

where the last inequality uses the fact that hypothesis tests by design have probability of false discovery controlled under the chosen significance level, which in this case is α/N.

The Bonferroni correction is unfortunately too stringent, and often does not make any discoveries; as N grows large, α/N is too small. And indeed, this makes intuitive sense: if we test a million hypotheses, preventing the possibility of making a *single* false discovery in a million tests necessarily implies that we have to be extremely conservative and only proclaim a discovery if there is overwhelmingly strong signal of something interesting.

4.5 False Discovery Proportion

Recall from Lecture 1, when we are testing N different hypotheses, we can talk about the the number of true positives, true negatives, false positives, and false negatives as counts, as in Table 4.1.

		decision		
		null (0)	*non-null (1)*	
reality	*null*	n_{00}	n_{01}	$n_{00} + n_{01}$
	non-null	n_{10}	n_{11}	$n_{10} + n_{11}$
		$n_{00} + n_{10}$	$n_{01} + n_{11}$	N

Table 4.1: Different ground truth and decision relationships in multiple testing.

Further recall, we defined the **false discovery proportion** (FDP) as:

$$\text{FDP} = \frac{n_{01}}{n_{01} + n_{11}}.$$

This quantity can be thought of as being an estimate of a conditional probability $\mathbb{P}(H = 0 \mid D = 1)$, where D is a random variable denoting the decision, and H is a random variable denoting the ground truth about the hypothesis. Unlike sensitivity and specificity, this quantity is dependent on the prevalence (prior probability of null, i.e. $\mathbb{P}(H = 0)$).

By Bayes' theorem, we can rewrite this conditional probability as

$$\mathbb{P}(H = 0 \mid D = 1) = \frac{\mathbb{P}(D = 1 \mid H = 0)\, \mathbb{P}(H = 0)}{\mathbb{P}(D = 1)} := \frac{\mathbb{P}(\text{type I error})\, \pi_0}{\mathbb{P}(D = 1)}.$$

We can simply bound π_0 by 1; this even makes a reasonable assumption, because in practice most things we test are truly null. In both frequentist and Bayesian thinking, $\mathbb{P}(\text{type I error})$ is assumed known, because we assume we know how the data behave under the null. The denominator $\mathbb{P}(D = 1)$ is equal to

$$\mathbb{P}(D = 1) = \pi_0 \mathbb{P}(D = 1 | H = 0) + (1 - \pi_0)\mathbb{P}(D = 1 | H = 1),$$

so this term does indeed depend on the prevalence. However, a fortunate circumstance is that it can easily be estimated from data. An "obvious" estimate of this quantity is

$$\widehat{\mathbb{P}}(D = 1) = \frac{n_{01} + n_{11}}{N}.$$

Therefore, it seems reasonable to find a procedure that ensures

$$\mathbb{P}(H = 0 \mid D = 1) \approx \frac{\mathbb{P}(\text{type I error})}{\widehat{\mathbb{P}}(D = 1)} \leq \alpha.$$

We will return to this idea in the next section.

4.6 False Discovery Rate Control

The **false discovery rate** is defined as the expectation of the false discovery proportion:

$$\text{FDR} = \mathbb{E}[\text{FDP}] = \mathbb{E}\left[\frac{n_{01}}{n_{01} + n_{11}}\right].$$

Since data are random, we cannot hope to control the FDP under a target level α with probability one. For example, if all of our hypotheses are null, there is a possibility that all corresponding p-values are extremely small, small enough to be rejected (maybe even under the Bonferroni correction). This event indeed happens with very small probability, however it prevents us from controlling FDP across *all* events. For this reason, we are happy if we can control the FDP *on average*;

in other words, we want to design decision rules which will guarantee that the FDR is kept under a target level α (e.g. 0.05). Because the FDR is not random but is simply a number, this will be manageable.

FDR was proposed as an appropriate error metric in multiple testing by Benjamini and Hochberg, who argued that FWER is too stringent for modern testing practices. They also proposed the first procedure for FDR control, usually called the Benjamini-Hochberg (BH) procedure. It takes a target FDR level α, as well as a set of p-values, and outputs which of the p-values correspond to discoveries. It does so in a way that guarantees FDR $\leq \alpha$. The explicit procedure statement is given below.

Algorithm 1 The Benjamini-Hochberg Procedure

input: FDR level α, set of N p-values $P_1, \ldots, P_N$
Sort the p-values $P_1, \ldots, P_N$ in non-decreasing order $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(N)}$
Find $K = \max\{i \in \{1, \ldots, N\} : P_{(i)} \leq \frac{\alpha}{N} i\}$
Reject the null hypotheses (declare discoveries) corresponding to $P_{(1)}, \ldots, P_{(K)}$

Notice that it makes sense to sort p-values and start rejecting from the smallest one; small p-values are generally indicative of an interesting finding, and null p-values are very small with tiny probability (recall that they are uniformly distributed).

The picture below visualizes how the BH procedure works.

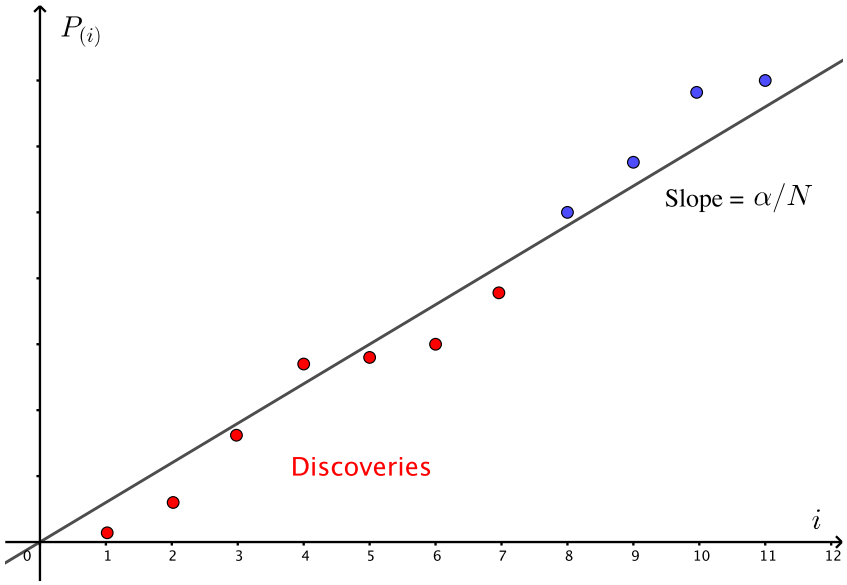

Figure 4.2: Illustration of the BH procedure.

Recall our discussion in Section 5, where we said that ensuring $\frac{\mathbb{P}(\text{type I error})}{\widehat{\mathbb{P}}(D=1)} \leq \alpha$ seems like a reasonable false discovery guarantee. In that case, we would like to pick the "least conservative" rule that will satisfy this. Suppose that we discover all p-values less than some threshold, and imagine that this threshold is equal to the K-th largest p-value, denoted $P_{(K)}$, for some fixed K. Under this decision rule, $\mathbb{P}(D = 1) = \frac{K}{N}$ by construction. On the other hand, the probability of a false positive $\mathbb{P}(\text{type I error})$ is equal to the probability that a null p-value is less than or equal to $P_{(K)}$. Since null p-values are uniform, this probability is exactly equal to $P_{(K)}$. Therefore, the condition $\frac{\mathbb{P}(\text{type I error})}{\widehat{\mathbb{P}}(D=1)} \leq \alpha$ can now be written as $\frac{P_{(K)}}{K/N} \leq \alpha$, i.e. $P_{(K)} \leq \frac{\alpha}{N}K$. Since we want our decision rule to be the least conservative one, we pick the maximum K such that $P_{(K)} \leq \frac{\alpha}{N}K$. Notice that this exactly corresponds to the BH procedure! To conclude, we have given a reinterpretation of the BH procedure, which says that it could actually be thought of as controlling an estimate of $\mathbb{P}(H = 0 \,|D = 1)$, the probability of a null given that we have made a discovery.

4.7 Adaptivity and Inference after Selection

All of the multiple hypothesis testing approaches that we have discussed today make a crucial assumption: all the hypotheses must be independent. Specifically, all of our analyses required that the the p-values are uniformly distributed under the null, regardless of the other hypotheses.

Often in practice, this independence assumption is violated.

Example 4.1. Suppose you collect a large dataset. As you begin to play around with the data, you clean it and select some reasonable variables for downstream analysis, discarding some others that seem irrelevant. Then, you run a single hypothesis test and obtain a p-value of 0.0001. Is this significant finding legitimate or should you do a correction? If you choose to do a correction, for what should you correct?

Computing p-values after making data-dependent choices generally breaks independence assumptions (and therefore the assumption the null distribution of your p-values is uniform). This is known as Freedman's paradox, and is now studied as "inference after selection" in Statistics and as "adaptive data analysis" in Computer Science.

Although wasteful in terms of sample efficiency, one of the best ways to adjust for data-dependent choices is to select new data from the same distribution and check your hypothesis on the fresh data. More sophisticated approaches exist which improve on sample efficiency, but are not yet very practical.

Data-dependent choices can include implicit comparisons on the researcher's part. Implicit comparisons are extremely difficult to correct for, as they often happen without being recognized or recorded. One proposed solution is known as pre-registration: the researcher specifies the entire experimental set-up before data collection, runs the set-up as specified once data is obtained, and reports the result regardless of the outcome. Initial metastudies suggest that pre-registration helps correct the peculiar distribution of p-values recognized as part of the reproducibility crisis.