Multiple Testing Problem

Zhang Rui-Yang

Contents

Preface

This notes is of expository nature, and describes the basics of the multiple testing problem. The multiple testing problem arises when we are conducting a large number of hypothesis testing on the same set of data. This is a frequently appearing problem in scientific research and real life scenarios, for example for genetic research (GWAS) and web design (AB testing). This set of notes is not original, and citations have been included (as much as possible) when necessary.

The notes was prepared while the author was working on a summer research project on high-dimensional statistics, supervised by Dr Tengyao Wang. I would like to thank Dr Wang for offering me this opportunity to work with him, and for his immense patience and encouragement.

London, Sep 2021

Chapter 1

Background

1.1 Frequentist Hypothesis Testing

Hypothesis testing is one of the first few concepts in modern Statistics, and it was originally phrased as in a Frequentist fashion. There are two rather different ways of constructing a hypothesis test under the Frequentist category - Fisherian test and Neyman-Pearson Test.

The Fisherian test, created by R. A. Fisher, only involves one hypothesis, the null hypothesis. We will calculate using the collected data, under the assumption that the null hypothesis is true, the probability of the data we receive is at least as extreme as this. This quantity, which is a random variable, is known as the **p-value** p. To decide whether the null hypothesis is rejected or not, we introduce a rather artificial threshold known as the significance level (or α), such that we will reject the hypothesis if $p < \alpha$. This value, under the traditional Fisherian framework, is pre-determined and may vary depending on the situations.

The Neyman-Pearson (NP) test, created by Jerzy Neyman and Egon Pearson, is based on the Neyman-Pearson lemma. It involves a pair of hypotheses, the null hypothesis H_0 and alternative hypothesis H_1. Here, we still have the concept of significance level α, but it is defined differently. Being predetermined, the significance level is the probability of rejecting the null hypothesis when the null hypothesis is in fact true. This is also known as the probability of Type I error. The significance level α is heavily used in this framework, since we construct the rejection region of the observed data R_α based on the value of α such that $\mathbb{P}(R_\alpha|H_0) = \alpha$. This is very different from the Fisherian version where α is merely a threshold for rejection.

The review paper by Ronald (2005) summarised various characteristics of these two methods of test as shown below.

> The basic elements of a Fisherian test are: (1) There is a probability model for the data. (2) Multidimensional data are sum- marized into a test statistic that has a known distribution. (3) This known distribution provides a ranking of the "weird-ness" of various observations. (4) The p-value, which is the probability of observing something as weird or weirder than was actually observed, is used to quantify the evidence against the null hypothesis. (5) α level tests are defined by reference to the p-value.

> The basic elements of an NP test are: (1) There are two hypothesized models for the data: H_0 and H_A. (2) An α level is chosen which is to be the probability of rejecting H_0 when H_0 is true. (3) A rejection region is chosen so that the probability of data falling into the rejection region is α when H_0 is true. With discrete data, this

3

often requires the specification of a randomized rejection region in which certain data values are randomly assigned to be in or out of the rejection region. (4) Various tests are evaluated based on their power properties. Ideally, one wants the most powerful test. (5) In complicated problems, properties such as unbiasedness or invariance are used to restrict the class of tests prior to choosing a test with good power properties.

Though being opposing views initially, the line separating the two methods of hypothesis testing has been greatly blurred and they are all mixed up in today's teaching and applications. Still, it is good to notice certain currently seemingly homogeneous things were, in fact, originated from heterogeneous backgrounds.

Later on, the Bayesian side of Statistics proposed the Bayesian version of hypothesis testing that is different from either of the previously mentioned two. The Bayesian hypothesis test will be described later.

1.2   Single Hypothesis Testing

In a single hypothesis test setting, we would choose to either reject or not reject the null hypothesis H_0 under a particular significance level. This choice may be erroneous, and there are two kinds of errors associating with a test, as shown in Table 1.1.

	H_0 is true	H_0 is false
Reject H_0	Type I Error False Positive α	True Positive $1 - \beta$
Not Reject H_0	True Negative $1 - \alpha$	Type II Error False Negative β

Table 1.1: Single Hypothesis Testing Result

Here, the value $1 - \beta$ is known as the **power** of the test, it measures the probability that the test correctly rejects the null hypothesis when it is indeed false. α is the **significance level** of the test, which is (supposedly) pre-determined and is used to decide the values of test statistics for the null hypothesis to be rejected.

Another concept associated with a hypothesis test is its **p-value**. Assuming the null hypothesis is true, the p-value is the probability that the test statistic is as, or more, extreme than the one we obtained. It is a random variable, and its realisations are also called p-values. When the p-value of a test is smaller than the predetermined significance level α, we would reject the null hypothesis. One thing to notice about the p-value is that if the null hypothesis is true, it follows a Uniform(0,1) distribution.

Now, if we set the significance level α as 0.05, meaning that we will reject our null hypothesis if the p-value is less than 0.05, we will have a false positive result (rejecting the null hypothesis when it is true) 5% of the time. This chance or error leads to the occurrence of p-hacking, a way to pick a particle set of data among many that rejects the null hypothesis when H_0 is true.

The occurrence frequency of false positives depends on the value of α. If we do multiple hypothesis tests to the same set of data while remaining with our significance level α for each individual test, we will have a much more substantial amount of errors in our results. For example, if we do 10000 hypothesis testing each at significance level 5%, we will make 500 false positive decisions on average. This kind of mistakes is known as **multiple testing problem**, or

interchangeably as multiplicity and multiple comparisons problem. This problem does exist in reality. For example, in quantitative trait loci and microarray analyses, the number of hypotheses tested in an experiment reached thousands, which made the issue of multiple testing a more important one (Benjamini, 2010).

1.3 Organisation of Chapters

In Chapter 2, we will discuss the family-wise error rate and introduce two methods to control it. In Chapter 3, we will mention the false discovery rate, the Benjamini-Hochberg procedure, and how we can weaken the dependency condition of test statistics for the standard Benjamini-Hochberg procedure. An alternative set of proofs for this procedure will be provided as well. In Chapter 4, we will provide an alternative formulation of false discovery rate using the language of empirical Bayes. In Chapter 5, we will introduce the concept of e-variable and e-value, as a supplement, or even an alternative, to p-value. We will also be talking about how the e-values can rephrase the Benjamini-Hochberg procedure.

Chapter 2

Family-wise Error Rate

In this chapter, we will be introducing the concept of family-wise error rate, and state some methods that control it.

2.1 FWER

Analogue to *Table 1*, we have the following table to establish the case when we have m null hypotheses denoted by $H_1, H_2, \cdots, H_m$.

	H_0 is true	H_0 is false	Total
Reject H_0	V	S	R
Not Reject H_0	U	T	$m - R$
Total	m_0	$m - m_0$	m

Table 2.1: Multiple Hypothesis Testing Result

Here, the capital letters V, S, U, T and R are all random variables with R being the only observable one among them, and m_0 and m are known values in advanced.

According to Hochberg and Tamhane (1987), the problem of significance arises when we view multiple hypothesis testing as separate inferences rather than related ones. This leads to the introduction of the concept of 'family', which is defined by the authors to be 'any collection of inferences for which it is meaningful to take into account some combined measure of errors'.

Here, the random variable V denotes the number of Type I Errors we made among the m decisions, so we would want to limit it. We will control $\mathbb{P}(V \geq 1)$, the probability of making at least one false positive, and we will call this quantity as **family-wise error rate**, or FWER. So, we have
$$\text{FWER} = \mathbb{P}(V \geq 1) = 1 - \mathbb{P}(V = 0).$$

When controlling the FWER, there are two types of controls, namely the weak control and the strong control. A particular procedure will control the FWER in the weak sense if the FWER control at level α is guaranteed only when all null hypotheses are true. We will call a procedure to control the FWER in the strong sense if the control at level α is guaranteed when at least one null hypothesis is true.

2.2 Bonferroni Correction

One way to control the FWER is by applying the Bonferrnoi correction. If we would want the overall significance level of the family of tests to be α, then we will be setting the significance level for each one of the m test lower, and the Bonferroni correction suggests $\frac{\alpha}{m}$.

To show this correction actually controls the FWER, we first have the Boole's inequality which states that for a countable set of events $A_1, A_2, A_3, \cdots$, we have

$$\mathbb{P}(\bigcup_i A_i) \leq \sum_i \mathbb{P}(A_i),$$

which follows from the sub-additivity property of probability measure.

With this inequality, since we are rejecting each null hypothesis H_i when its p-value $p_i \leq \frac{\alpha}{m}$, we would have

$$
\begin{aligned}
\text{FWER} &= \mathbb{P}(V \geq 1) \\
&= \mathbb{P}\Big(\bigcup_{i=1}^{m_0} (p_i \leq \frac{\alpha}{m}) \Big) \quad\quad \text{since there are } m_0 \text{ of them with true null hypothesis} \\
&\leq \sum_{i=1}^{m_0} \mathbb{P}(p_i \leq \frac{\alpha}{m}) \quad\quad \text{by Boole's inequality} \\
&= \sum_{i=1}^{m_0} \frac{\alpha}{m} \quad\quad \text{since p-value follows Uniform(0,1) when null hypothesis is true} \\
&= \frac{m_0}{m}\alpha \leq \alpha,
\end{aligned}
$$

which indicates that the FWER is controlled under level α.

It should not be too hard to notice that this control is too strict. For example, FWER is actually controlled under $\frac{m_0}{m}\alpha$ instead of α, which could be much lower if m_0 is much smaller than m. Having a strict control for Type I errors implies an increase in Type II errors, which is why Bonferroni correction is not always good. However, since his correction is an easy one to apply, it is still used in practice.

2.3 Holm-Bonferroni Method

As an improvement of the Bonferroni correction, Holm (1979) proposed the following method to proceed the multiple tests.

For the m null hypotheses $H_1, H_2, \cdots, H_m$, we compute their respective p-values $P_1, P_2, \cdots, P_m$ and we rank them such that $P_{(k)}$ denotes the k-th smallest p-value. So, $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$, and we denote the corresponding null hypotheses as $H_{(1)}, H_{(2)}, \cdots, H_{(m)}$. We would want to control the FWER at α.

- Is $P_{(1)} \leq \frac{\alpha}{m}$? If so, reject $H_{(1)}$ and continue. Otherwise, EXIT.
- Is $P_{(2)} \leq \frac{\alpha}{m-1}$? If so, reject $H_{(2)}$ and continue. Otherwise, EXIT.

  $\vdots$

- Is $P_{(k)} \leq \frac{\alpha}{m-k+1}$? If so, reject $H_{(k)}$ and continue. Otherwise, EXIT.

To show that this method does in fact keep the FWER at α, we let I_0 be the set of indices corresponding to the true null hypothesis. This set is unknown to us and has size m_0. Let us assume that we make the first false positive decision at $H_{(h)}$. Based on the procedures stated above, all the decisions for null hypothesis $H_{(1)}, H_{(2)}, \cdots, H_{(h-1)}$ are true positives. Also, we know for a fact that $h - 1 \leq m - m_0$ due to the definition of $m - m_0$. This implies that $m - h + 1 \geq m_0$, and then $\frac{1}{m-h+1} \leq \frac{1}{m_0}$. Now, since $H_{(h)}$ is rejected, we would have $P_{(h)} \leq \frac{\alpha}{m-h+1}$ by definition, so we will then have $P_{(h)} \leq \frac{\alpha}{m-h+1} \leq \frac{\alpha}{m_0}$. This means, if there is any false positive, we have at least one true null hypothesis with p-value less than $\frac{\alpha}{m_0}$. So,

$$
\begin{aligned}
\text{FWER} &= \mathbb{P}(V \geq 1) \\
&= \mathbb{P}\Big(\bigcup_{i \in I_0} (P_i \leq \frac{\alpha}{m_0}) \Big) \\
&\leq \sum_{i \in I_0} \mathbb{P}(P_i \leq \frac{\alpha}{m_0}) \qquad \text{by Boole's inequality} \\
&= m_0 \frac{\alpha}{m_0} \qquad \text{since p-value follows Uniform(0,1) when null hypothesis is true} \\
&= \alpha,
\end{aligned}
$$

which indicates that the FWER is controlled at level α.

By looking at values of significance levels that we used to reject the null hypothesis, we can immediate realise that the Holm-Bonferroni is uniformly more powerful than the Bonferroni correction. So this is, indeed, an improvement of Bonferroni.

2.4   Issues with Controlling FWER

FWER controls the probability of making Type I errors. When we put stronger restrictions on false positives, we will have the problem of having more false negatives, i.e. having a higher probability of making Type II errors. This makes the power of the hypothesis testing decreases, and not always a suitable control.

According to Benjamini and Hochberg (1995), methods that control FWER have flaws in real-life applications.

1. Classical procedures that control the FWER in the strong sense, at levels conventional in single-comparison problems, tend to have substantially less power than the per comparison procedure of the same levels.
2. Often the control of the FWER is not quite needed. The control of the FWER is important when a conclusion from the various individual inferences is likely to be erroneous when at least one of them is. This may be the case, for example, when several new treatments are competing against a standard, and a single treatment is chosen from the set of treatments which are declared significantly better than the standard. However, a treatment group and a control group are often compared by testing various aspects of the effect (different end points in clinical trials terminology). The overall conclusion that the treatment is superior need not be erroneous even if some of the null hypotheses are falsely rejected.

To improve on these aspects, Benjamini and Hochberg proposed to control a different measure of error, the false discovery rate, which will be more ideal to help with the problem of multiplicity. The discussion will be carried out in the next chapter.

Chapter 3

False Discovery Rate

In this chapter, we will be introducing the concept of false discovery rate proposed by Benjamini and Hochberg (1995), and state some of its controlling methods.

3.1 FDR

As mentioned at the end of the previous chapter, controlling FWER is not always ideal. There is a need for a new point of view on the problem of multiplicity. A lot of the times, the number of false positive should be controlled, but we should also consider this number relative to the total number of rejections we are making. For example, among a total of 100 hypothesis testing, if we make 2 false positives among 5 rejections, it is rather serious. However, it we make 2 false positives among 50 rejections, it is more bearable. The seriousness of the loss incurred by false positives is inversely related to the number of hypotheses rejected (Benjamini & Hochberg, 1995). With that in mind, we need a different measure of error that accounts for the proportion of errors among the rejected hypotheses, which is called the **false discovery rate**, or FDR. The word 'discovery' is used since a rejected hypothesis was called a 'statistical discovery' by Soriç (1989).

Using the notations in *Table 2*, we will define a new random variable $Q = V/(V + S)$ where $Q = 0$ when $V + S = 0$. This is the proportion of the false positives over all the rejected null hypotheses. This is unobservable since we do not know V, or S, or their realisations v or s. We will define the FDR as the expectation of Q,

$$\text{FDR} = \mathbb{E}[Q] = \mathbb{E}[V/(V + S)] = \mathbb{E}[V/R].$$

To avoid the division by zero issue, we would have the alternative formula for FDR as

$$\text{FDR} = \mathbb{E}[V/R | R > 0]\mathbb{P}(R > 0).$$

Two properties of FDR can be easily shown. Firstly, if all the null hypotheses are true, FDR = FWER. When $s = 0$ and $v = r$, $Q = 0$ if $v = 0$ and $Q = 1$ if $v > 0$, which means $\mathbb{P}(V \geq 1) = \mathbb{E}[Q]$. This means a control of FDR is a control of FWER in the weak sense. The second property is, when $m_0 < m$, FDR is no bigger than FWER. Given $m_0 < m$, $v > 0$ implies $v/(v + s) \leq 1$, which means $V \geq 1 \implies V \geq v/(v + s)$ and $\mathbb{P}(V \geq 1) \geq \mathbb{E}[Q]$. This means a control of FWER will control FDR. If a procedure controls FDR only, it would be more poewrful than one that controls the FWER. The potential for increase in power is larger when more of the hypotheses are non-true.

One should note that FDR is not the only possible measure to capture the idea of the proportion of false positives among all the rejected null hypotheses. In Benjamini and Hochberg (1995), the authors mentioned two alternatives, that of $\mathbb{E}[V/R|R > 0]$ and $\mathbb{E}[V]/\mathbb{E}[R]$. In the paper, the authors explained their reasons of not choosing these two. However, they turned out to be rather useful later on, and we will be discussed them in detail in later chapters.

3.2 Benjamini-Hochberg Procedure

In order to control the FDR as described above, Benjamini and Hochberg (1995) proposed a procedure that is latter commonly known as the Benjamini-Hochberg procedure, or BH procedure.

For the m null hypotheses $H_1, H_2, \cdots, H_m$, we compute their respective p-values $P_1, P_2, \cdots, P_m$ and we rank them such that $P_{(k)}$ denotes the k-th smallest p-value. So, $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$, and we denote the corresponding null hypotheses as $H_{(1)}, H_{(2)}, \cdots, H_{(m)}$. We let the level of FDR that we would want to control at as α. The procedure works as the following:

- Let k be the largest i for which $P_{(i)} \leq \frac{i}{m}\alpha$.
- Reject all $H_{(i)}$ where $i = 1, 2, \cdots, k$.

We would like to prove that this procedure does in fact control the FDR at level α. So we have the following theorem.

Theorem 1. For **independent** test statistics and for any configuration of false null hypotheses, the above precedure controls the FDR at α.

Proof. (Proof of Theorem) The theorem follows from the following lemma, whose proof will be given after.

Lemma 1. For any m_0 ($0 \leq m_0 \leq m$) independent p-values corresponding to true null hypotheses, and for any values that the $m_1 = m - m_0$ p-values corresponding to the false null hypotheses can take, the BH procedure satisfies the inequality

$$\mathbb{E}[Q|P_{m_0+1} = p_1, \cdots, P_m = p_{m_1}] \leq \frac{m_0}{m}\alpha.$$

With that, whatever the joint distribution of the p-values of the false null hypotheses $P_1'', \cdots, P_{m_1}''$ is, by integrating the above inequality, we would get

$$\mathbb{E}[Q] \leq \frac{m_0}{m}\alpha \leq \alpha.$$

$\square$

Proof. (Proof of Lemma) The proof is completed by an induction on m. When $m = 1$, the lemma is true immediately. Now, assuming that the statement is true for $m = m'$, we would like to show that it holds for $m' + 1$.

If $m_0 = 0$ and $m_1 = m - 0 = m$, all the null hypotheses are false and Q will be 0. This means

$$\mathbb{E}[Q|P_1 = p_1, \cdots, P_m = p_m] = 0 \leq \frac{m_0}{m' + 1}\alpha.$$

If $m_0 > 0$, we denote the p-values corresponding to the true null hypotheses as P_i' with $i = 1, 2, \cdots m_0$ with the largest being $P_{(m_0)}$. These p-values are i.i.d. and follow Uniform(0,1). We will order the p-values of the m_1 false null hypotheses and let them be $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m_1)}$. Next, we will define j_0 to be the largest j in $[0, m_1]$ with

$$p_{(j)} \leq \frac{m_0 + j}{m' + 1}\alpha,$$

and we let $p'' = p_{(j_0)}$.

For the sake of simplicity in notation, we will let A to denote the event $P_{(m_0+1)} = p_{(1)}, \cdots, P_{(m)} = p_{(m_1)}$. So, conditioning on $P_{(m_0)}' = p$ for some variable p,

$$\mathbb{E}[Q|A] = \int_0^1 \mathbb{E}[Q|P_{(m_0)}' = p, A]f_{P_{(m_0)}'}(p)dp$$

$$= \int_0^{p''} \mathbb{E}[Q|P_{(m_0)}' = p, A]f_{P_{(m_0)}'}(p)dp + \int_{p''}^1 \mathbb{E}[Q|P_{(m_0)}' = p, A]f_{P_{(m_0)}'}(p)dp$$

with $f_{P_{(m_0)}'}(p) = m_0 p^{m_0 - 1}$.

In the first term, we have $0 \leq p \leq p''$. This means we are rejecting $m_0 + j_0$ hypotheses, and we would have $Q = m_0/(m_0 + j_0)$. Using the inequality

$$p'' = p_{(j_0)} \leq \frac{m_0 + j_0}{m' + 1}\alpha,$$

we would have

$$\int_0^{p''} \mathbb{E}[Q|P_{(m_0)}' = p, A]f_{P_{(m_0)}'}(p)dp = \int_0^{p''} \frac{m_0}{m_0 + j_0}m_0 p^{m_0 - 1}dp$$

$$= \left[\frac{m_0}{m_0 + j_0}p^{m_0}\right]_0^{p''}$$

$$= \frac{m_0}{m_0 + j_0}(p'')^{m_0}$$

$$\leq \frac{m_0}{m_0 + j_0}\frac{m_0 + j_0}{m' + 1}\alpha(p'')^{m_0 - 1}$$

$$= \frac{m_0}{m' + 1}\alpha(p'')^{m_0 - 1}.$$

For the second term, we will consider separately the values of p when $p_{(j_0)} < p_{(j)} \leq P_{(m_0)}' = p < p_{(j+1)}$, and when $p_{(j_0)} \leq p'' < P_{(m_0)}' = p < p_{(j_0+1)}$. It is important to note that, based on the definition of j_0 and p'', no hypothesis can be rejected due to the values of $p, p_{(j+1)}, p_{(j+2)}, \cdots, p_{(m_1)}$. Therefore, when all hypotheses are considered with ordered p-values, a hypothesis $H_{(i)}$ could be rejected only if there exists k with $i \leq k \leq m_0 + j - 1$, where $p_{(k)} \leq \alpha \cdot k/(m' + 1)$, or

$$\frac{p_{(k)}}{p} \leq \frac{k}{m_0 + j - 1}\frac{m_0 + j - 1}{(m' + 1)p}\alpha.$$

When conditioning on $P_{(m_0)}' = p$, the P_i'/p for $i = 1, 2, \cdots, m_0 - 1$ are i.i.d. random variables following Uniform(0,1), and the p_i/p for $i = 1, 2, \cdots, j$ are numbers corresponding to false null hypotheses between 0 and 1. Then,

$$\mathbb{E}[Q|P_{(m_0)}', A] \leq \frac{m_0 - 1}{m_0 + j - 1}\frac{m_0 + j - 1}{(m' + 1)p}\alpha = \frac{m_0 - 1}{(m' + 1)p}\alpha.$$

11

So,

$$\int_{p''}^{1} \mathbb{E}[Q|P'_{(m_0)} = p, A] f_{P'_{(m_0)}}(p) dp \leq \int_{p''}^{1} \frac{m_0 - 1}{(m' + 1)p} \alpha m_0 p^{m_0 - 1} dp$$

$$= \frac{m_0}{m' + 1} \alpha \int_{p''}^{1} (m_0 + 1) p^{m_0 - 2} dp$$

$$= \frac{m_0}{m' + 1} \alpha (1 - (p'')^{m_0 - 1}).$$

Adding up the two terms, we would have

$$\mathbb{E}[Q|A] \leq \frac{m_0}{m' + 1} \alpha (p'')^{m_0 - 1} + \frac{m_0}{m' + 1} \alpha (1 - (p'')^{m_0 - 1}) = \frac{m_0}{m' + 1} \alpha,$$

which completes the proof. $\qquad\square$

3.3   Dependency of Test Statistics

As stated in bold in Theorem 1, one condition for the BH procedure is that the test statistics need to be independent. This is quite a big restriction, since many hypothesis tests carried out in practice are dependent to each other. There is a need to make some extension of the method on that aspect, and these gaps are filled mostly by Benjamini and Yekutieli (2001). Here, I will state without proof some of the key improvements.

We have already known from the previous section that the BH procedure will control the FDR at $(m_0/m)\alpha$ if test statistics are independent. By Theorem 5.1 of Benjamini and Yekutieli (2001), we have

Independent Test Statistics:
$$\text{FDR} \leq \frac{m_0}{m} \alpha,$$

and Independent and Continuous Test Statistics:
$$\text{FDR} = \frac{m_0}{m} \alpha.$$

The main result of Benjamini and Yekutieli (2001) based on the dependency type *positive regression dependency on each one from a subset*, or PRDS. First, we will call a subset D of Ω as increasing if for some $x \in D$, $y \in \Omega$ and $y \geq x$ imply $y \in D$. For example, the first quadrant of the $\mathbb{R}^2$ plane is an increasing set. Here, if x and y has more than one coordinate, $y \geq x$ means $y_i \geq x_i$ for every coordinate. Now, if test statistics vector $\mathbf{X}$ is PRDS on I_0, it means that for any increasing set D and each $i \in I_0 \subset I$, $\mathbb{P}(\mathbf{X} \in D | X_i = x)$ is non-decreasing as x increases. With this, we can state the main result.

Theorem 2. If the joint distribution of the test statistics is PRDS on the subset of test statistics corresponding to true null hypotheses, the BH procedure controls the FDR at level less than or equal to $(m_0/m)q$.

Proof. Let us define the constants involved in the BH procedure as

$$q_i = \frac{i}{m} q, \quad i = 1, 2, \cdots, m.$$

Let $A_{v,s}$ denote the event that the BH procedure rejects exactly v true and s false null hypotheses. Here, $k = v + s$ hypotheses have been rejected, so all these hypotheses will have p-values $\leq q_{v+s}$. The FDR will then be

$$\text{FDR} = \sum_{s=0}^{m_1} \sum_{v=1}^{m_0} \frac{v}{v+s} \mathbb{P}(A_{v,s}).$$

Here, v starts from 1 since $FDR = 0$ if $v = 0$, and $v + s \neq 0$ for the fraction to make sense.

Now, we would want to find $\mathbb{P}(A_{v,s})$. We claim that

$$\mathbb{P}(A_{v,s}) = \frac{1}{v} \sum_{i=1}^{m_0} \mathbb{P}(\{P_i \leq q_{v+s}\} \cap A_{v,s}).$$

Let us prove this claim. For a fixed v and s, let ω be a subset of $\{1, 2, \cdots, m_0\}$ of size m_0. Then, we let $A_{v,s}^{\omega}$ be the event in $A_{v,s}$ where the v rejected true hypotheses have index ω. Here, the index set $\{1, 2, \cdots, m_0\}$ is for the set of true null hypotheses. Also, the set of $A_{v,s}^{\omega}$ for all possible ω is a partition of $A_{v,s}$ with each of its element being disjoint. So, we would have

$$\mathbb{P}(\{P_i \leq q_{v+s}\} \cap A_{v,s}^{\omega}) = \begin{cases} \mathbb{P}(A_{v,s}^{\omega}) & i \in \omega \\ 0 & i \notin \omega. \end{cases}$$

This is because, if $i \in \omega$, $\{P_i \leq q_{v+s}\}$ means that the hypothesis with index i is rejected. So, $\{P_i \leq q_{v+s}\} \supset A_{v,s}^{\omega}$, and $\mathbb{P}(\{P_i \leq q_{v+s}\} \cap A_{v,s}^{\omega}) = \mathbb{P}(A_{v,s}^{\omega})$. If $i \notin \omega$, i will still be in $\{1, 2, \cdots, m_0\}$ since this is the way i is summed. This implies that H_i becomes a rejected true null hypothesis with index not inside ω, which contradicts with $A_{v,x}^{\omega}$. So, $\{P_i \leq q_{v+s}\}$ and $A_{v,x}^{\omega}$ are disjoint events, making the probability of their intersection 0.

Then, we would have

$$\sum_{i=1}^{m_0} \mathbb{P}(\{P_i \leq q_{v+s}\} \cap A_{v,s}) = \sum_{i=1}^{m_0} \sum_{\omega} \mathbb{P}(\{P_i \leq q_{v+s}\} \cap A_{v,s}^{\omega})$$

$$= \sum_{\omega} \sum_{i=1}^{m_0} \mathbb{P}(\{P_i \leq q_{v+s}\} \cap A_{v,s}^{\omega})$$

$$= \sum_{\omega} \sum_{i=1}^{m_0} \mathbf{1}(i \in \omega) \mathbb{P}(A_{v,s}^{\omega})$$

$$= \sum_{\omega} v \mathbb{P}(A_{v,s}^{\omega})$$

$$= v \mathbb{P}(A_{v,s}),$$

which implies $\mathbb{P}(A_{v,s}) = \frac{1}{v} \sum_{i=1}^{m_0} \mathbb{P}(\{P_i \leq q_{v+s}\} \cap A_{v,s})$.

Combining this with the statement of FDR, we have

$$\text{FDR} = \sum_{s=0}^{m_1} \sum_{v=1}^{m_0} \frac{v}{v+s} \mathbb{P}(A_{v,s})$$

$$= \sum_{s=0}^{m_1} \sum_{v=1}^{m_0} \frac{v}{v+s} \left(\frac{1}{v} \sum_{i=1}^{m_0} \mathbb{P}(\{P_i \leq q_{v+s}\} \cap A_{v,s}) \right)$$

$$= \sum_{i=1}^{m_0} \left(\sum_{s=0}^{m_1} \sum_{v=1}^{m_0} \frac{1}{v+s} \mathbb{P}(\{P_i \leq q_{v+s}\} \cap A_{v,s}) \right).$$

We would want the equation to be independent of v, but we still have to be dependent on it since we have $A_{v,s}$. This means we need to rewrite it using i and $k = v + s$.

For $i = 1, 2, \cdots, m_0$, let $\mathbf{P}^{(i)}$ be the vector of $m - 1$ p-values excluding P_i. Then, we let $C_{v,s}^{(i)}$ be the event in which if P_i is rejected then $v - 1$ true null hypotheses and s false null hypotheses are rejected alongside with it. Thus, we have the equality

$$\{P_i \leq q_{v+s}\} \cap A_{v,s} = \{P_i \leq q_{v+s}\} \cap C_{v,s}^{(i)}.$$

We then denote $C_k^{(i)} = \uplus_{j \leq k} C_{v,s}^{(i)}$. For each i, $C_k^{(i)}$ are disjoint, so we can rewrite the FDR as

$$\text{FDR} = \sum_{i=1}^{m_0} \Big(\sum_{s=0}^{m_1} \sum_{v=1}^{m_0} \frac{1}{v + s} \mathbb{P}(\{P_i \leq q_{v+s}\} \cap A_{v,s}) \Big)$$

$$= \sum_{i=1}^{m_0} \sum_{k=0}^{m} \frac{1}{k} \mathbb{P}(\{P_i \leq q_{v+s}\} \cap C_k^{(i)}),$$

which is what we want.

Now, we will try to use the PRDS property. Let $D_k^{(i)} = \cup_{j \leq k} C_j^{(i)}$ for $k = 1, 2, \cdots, m$. Notice that $D_m^{(i)}$ is the entire space. It is easy to spot that $D_k^{(i)}$ are nondecreasing set. Using PRDS, for $p \leq p'$, we have $\mathbb{P}(D|P_i = p) \leq \mathbb{P}(D|P_i = p')$ for some nondecreasing set D. We would also have, for $j \leq l$, $q_j \leq q_l$ and $\mathbb{P}(D|P_i \leq q_j) \leq \mathbb{P}(D|P_i \leq q_l)$. This means

$$\frac{\mathbb{P}(D_k^{(i)} \cap \{P_i \leq q_k\})}{\mathbb{P}(P_i \leq q_k)} \leq \frac{\mathbb{P}(D_k^{(i)} \cap \{P_i \leq q_{k+1}\})}{\mathbb{P}(P_i \leq q_{k+1})}$$

by the definition of conditional probability and setting $j = k$, $l = k + 1$, and $D = D_k^{(i)}$.

Using the above inequality, we have

$$\frac{\mathbb{P}(D_k^{(i)} \cap \{P_i \leq q_k\})}{\mathbb{P}(P_i \leq q_k)} + \frac{\mathbb{P}(C_{k+1}^{(i)} \cap \{P_i \leq q_{k+1}\})}{\mathbb{P}(P_i \leq q_{k+1})}$$

$$\leq \frac{\mathbb{P}(D_k^{(i)} \cap \{P_i \leq q_{k+1}\})}{\mathbb{P}(P_i \leq q_{k+1})} + \frac{\mathbb{P}(C_{k+1}^{(i)} \cap \{P_i \leq q_{k+1}\})}{\mathbb{P}(P_i \leq q_{k+1})}$$

$$= \frac{\mathbb{P}(D_{k+1}^{(i)} \cap \{P_i \leq q_{k+1}\})}{\mathbb{P}(P_i \leq q_{k+1})},$$

since $D_{j+1}^{(i)} = D_j^{(i)} + C_{j+1}^{(i)}$ for all $k \leq m - 1$. Notice that $C_1^{(i)} = D_1^{(i)}$, we repeat the above inequality for $k = 1, 2, \cdots, m - 1$ and get

$$\sum_{k=1}^{m} \frac{\mathbb{P}(C_k^{(i)} \cap \{P_i \leq q_k\})}{\mathbb{P}(P_i \leq q_k)} \leq \frac{\mathbb{P}(D_m^{(i)} \cap \{P_i \leq q_k\})}{\mathbb{P}(P_i \leq q_k)} = \frac{\mathbb{P}(P_i \leq q_k)}{\mathbb{P}(P_i \leq q_k)} = 1,$$

using the fact that $D_m^{(i)}$ is the entire space.

Thus,

$$
\begin{aligned}
\text{FDR} &= \sum_{i=1}^{m_0} \sum_{k=0}^{m} \frac{1}{k} \mathbb{P}(\{P_i \le q_{v+s}\} \cap C_k^{(i)}) \\
&\le \sum_{i=1}^{m_0} \sum_{k=0}^{m} \frac{q}{m} \frac{\mathbb{P}(\{P_i \le q_{v+s}\} \cap C_k^{(i)})}{\mathbb{P}(P_i \le q_k)} \\
&\quad \text{since } \mathbb{P}(P_i \le q_k) \le q_k = \frac{kq}{m} \implies \frac{1}{\mathbb{P}(P_i \le q_k)} \ge \frac{m}{kq} = \frac{1}{k} \cdot \frac{m}{q} \implies \frac{1}{k} \le \frac{q}{m} \frac{1}{\mathbb{P}(P_i \le q_k)} \\
&= \frac{q}{m} \sum_{i=1}^{m_0} \sum_{k=0}^{m} \frac{\mathbb{P}(\{P_i \le q_{v+s}\} \cap C_k^{(i)})}{\mathbb{P}(P_i \le q_k)} \\
&\le \frac{q}{m} \sum_{i=1}^{m_0} 1 \quad \text{using the above inequality} \\
&= \frac{q}{m} \cdot m_0 = \frac{m_0}{m} q.
\end{aligned}
$$

The proof is completed. $\qquad\square$

Remark. This proof can in fact be used to proof Theorem 1. Given that

$$
\text{FDR} = \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{1}{k} \mathbb{P}(\{P_i \le q_{v+s}\} \cap C_k^{(i)}),
$$

since the test statistics are independent for the BH procedure, we would have independent p-values and that means

$$
\begin{aligned}
\text{FDR} &= \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{1}{k} \mathbb{P}(\{P_i \le \frac{k}{m} q\} \cap C_k^{(i)}) \\
&= \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{1}{k} \mathbb{P}(\{P_i \le \frac{k}{m} q\}) \mathbb{P}(C_k^{(i)}) \\
&= \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{1}{k} \cdot \frac{k}{m} q \mathbb{P}(C_k^{(i)}) \\
&= \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{q}{m} \mathbb{P}(C_k^{(i)}) \\
&= \sum_{i=1}^{m_0} \frac{q}{m} \cdot 1 = \frac{m_0}{m} q.
\end{aligned}
$$

So, Theorem 1 has been shown.

We have shown that the control of FDR is at $(m_0/m)\alpha$ for most of the scenarios in practice. Notice that the coefficient m_0/m is occurring every single time, if we can have an estimate for m_0 (we already know m), we can have a better procedure. This is because, although the level is $(m_0/m)\alpha$, since we do not know m_0 we have to set the level at α which could be too conservative sometimes. If we can estimate m_0, we can control the FDR at α exactly. Many research have been done in order to estimate m_0, and we will be talking about some of those work in the next chapter.

3.4 Alternative Proof for Theorem 1 & 2

The following proofs for the BH procedure control and the version with weaker dependency condition are provided by E. Candés and R. Foygel Barber.

We want to prove the FDR control at α, or FDR $\leq (m_0/m)\alpha \leq \alpha$ where the equality of FDR holds for independent and continuous test statistics. Earlier on, we define FDR to be FDR $= \mathbb{E}[V/R]$ and $V/R = 0$ when $R = 0$. Here, we will use an alternative definition. We define the value false discovery proportion, or FDP, to be

$$\text{FDP} = \frac{V}{R \vee 1} = \begin{cases} V/R & \text{if } R \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Here, $A \vee B = \max(A, B)$. With FDP, we would have

$$\text{FDR} = \mathbb{E}[\text{FDP}]$$

and this is clearly equivalent to the previous definition. Now, to show the proof, we will rewrite FDP in the form of

$$\text{FDP} = \sum_{i \in H_0} \frac{V_i}{R \vee 1}$$

where $i \in H_0$ is the set of index corresponding to true null hypotheses with size m_0, and $V_i = \mathbf{1}\{H_i \text{ is rejected}\}$. We then see that if we claim $\mathbb{E}[V_i/(R \vee 1)] = \alpha/m$

$$\text{FDR} = \mathbb{E}[\text{FDP}] = \mathbb{E}[\sum_{i \in H_0} \frac{V_i}{R \vee 1}] = \sum_{i \in H_0} \mathbb{E}[\frac{V_i}{R \vee 1}] = \sum_{i \in H_0} \frac{\alpha}{m} = \frac{m_0}{m}\alpha$$

This means, to prove the FDR control, we just need to show that $\mathbb{E}[V_i/(R \vee 1)] = \alpha/m$ for the independent case or $\mathbb{E}[V_i/(R \vee 1)] \leq \alpha/m$ for the PRDS case.

Proof. (Theorem 1)
To prove Theorem 1, we only need to prove the claim $\mathbb{E}[V_i/(R \vee 1)] = \alpha/m$. Now, we will have

$$\frac{V_i}{R \vee 1} = \sum_{k=1}^{m} \frac{V_i \mathbf{1}\{R = k\}}{k}.$$

Also, based on the BH procedure, we have two observations. Firstly, when there are k rejections, some H_i will be rejected if and only if $p_i \leq (k/m)\alpha$. So, we have $V_i = \mathbf{1}\{H_i \text{ is rejected}\} = \mathbf{1}\{p_i \leq (k/m)\alpha\}$. Secondly, if we reject some H_i, or we have $p_i \leq (k/m)\alpha$, let us take p_i and set its value to 0, and denote the new number of rejection by $R(p_i \to 0)$. This new number is exactly the same as R. If we reject H_i, since the rejection of hypotheses only take out the k hypotheses and this change of p_i is not affecting the hypotheses we are taking. If we do not reject H_i, V_i will be 0 so the value of R would not matter. Thus, we would have $V_i \mathbf{1}\{R = k\} = V_i \mathbf{1}\{R(p_i \to 0) = k\}$.

Combining the observations and taking the expectation conditional on all p-values except for p_i, i.e. $\mathcal{F}_i = \{p_1, \cdots, p_{k-1}, p_{k+1}, \cdots, p_m\}$, we have

$$\mathbb{E}[\frac{V_i}{R \vee 1}|\mathcal{F}_i] = \sum_{k=1}^{m} \frac{\mathbb{E}[V_i \mathbf{1}\{R = k\}|\mathcal{F}_i]}{k}$$

$$= \sum_{k=1}^{m} \frac{\mathbb{E}[\mathbf{1}\{H_i \text{ is rejected}\}\mathbf{1}\{R(p_i \to 0) = k\}|\mathcal{F}_i]}{k}$$

$$= \sum_{k=1}^{m} \frac{(k/m)\alpha \cdot \mathbf{1}\{R(p_i \to 0) = k\}}{k}$$

where the last equality holds due to the face that $p_i \sim \text{Uniform}(0,1)$ under true null hypotheses and they are independent. Also, given $\mathcal{F}_i$ and $p_i = 0$, $\mathbf{1}\{R(p_i \to 0) = k\}$ is deterministic. This also means $\sum_{k=1}^{m} \mathbf{1}\{R(p_i \to 0) = k\} = 1$. This is because by setting $p_i = 0$, we will have at least one rejection and $R(P_i \to 0) \geq 1$. Also, since $R(P_i \to 0)$ is deterministic and fixed, it will be one of the values from 1 to m. So, we have

$$\mathbb{E}[\frac{V_i}{R \vee 1}|\mathcal{F}_i] = \frac{\alpha}{m}\sum_{k=1}^{m}\mathbf{1}\{R(p_i \to 0)\} = \frac{\alpha}{m}.$$

Using the tower property $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$ for two random variables X and Y, we can conclude that $\mathbb{E}[V_i/(R \vee 1)] = \mathbb{E}[\mathbb{E}[V_i/(R \vee 1)]|\mathcal{F}_i] = \mathbb{E}[\alpha/m] = \alpha/m$. $\qquad\square$

From the above proof, we only use the independence between the true null hypotheses, and the dependency between the false ones are not important to the proof. This means, assuming all test statistics being independent may be too strong of a condition, and we certainly should weaken it. With that, we have Theorem 2 that only assumes PRDS on the true nulls.

Proof. (Theorem 2)
As mentioned earlier, to prove Theorem 2, we only need to prove the claim $\mathbb{E}[V_i/(R \vee 1)] \leq \alpha/m$ under PRDS on true null hypotheses. We have the same setting as the previous proof. Then, we set $q_k = k/m \cdot \alpha$ and have

$$\begin{aligned}
\frac{V_i}{R \vee 1} &= \sum_{k=1}^{m}\frac{\mathbf{1}\{p_i \leq q_k\}\mathbf{1}\{R = k\}}{k} \\
&= \sum_{k=1}^{m}\frac{\mathbf{1}\{p_i \leq q_k\}(\mathbf{1}\{R \leq k\} - \mathbf{1}\{R \leq k-1\})}{k} \\
&= \sum_{k=1}^{m}\frac{\mathbf{1}\{p_i \leq q_k\}\mathbf{1}\{R \leq k\}}{k} - \sum_{k=1}^{m}\frac{\mathbf{1}\{p_i \leq q_k\}\mathbf{1}\{R \leq k-1\}}{k} \\
&= \sum_{k=1}^{m}\frac{\mathbf{1}\{p_i \leq q_k\}\mathbf{1}\{R \leq k\}}{k} - \sum_{k=0}^{m-1}\frac{\mathbf{1}\{p_i \leq q_{k+1}\}\mathbf{1}\{R \leq k\}}{k+1} \\
&= \frac{\mathbf{1}\{p_i \leq q_m\}\mathbf{1}\{R \leq m\}}{m} + \sum_{k=1}^{m-1}\frac{\mathbf{1}\{p_i \leq q_k\}\mathbf{1}\{R \leq k\}}{k} - \sum_{k=1}^{m-1}\frac{\mathbf{1}\{p_i \leq q_{k+1}\}\mathbf{1}\{R \leq k\}}{k+1} \\
&= \frac{\mathbf{1}\{p_i \leq q_m\}\mathbf{1}\{R \leq m\}}{m} + \sum_{k=1}^{m-1}\left[\frac{\mathbf{1}\{p_i \leq q_k\}}{k} - \frac{\mathbf{1}\{p_i \leq q_{k+1}\}}{k+1}\right]\mathbf{1}\{R \leq k\}.
\end{aligned}$$

Notice that
$$\mathbb{E}[\frac{\mathbf{1}\{p_i \leq q_m\}\mathbf{1}\{R \leq m\}}{m}] = \frac{\alpha}{m},$$
since $R \leq m$ is always true and p_i follows Uniform(0,1) under true null. So, if we can show that

$$\mathbb{E}[\sum_{k=1}^{m-1}\left[\frac{\mathbf{1}\{p_i \leq q_k\}}{k} - \frac{\mathbf{1}\{p_i \leq q_{k+1}\}}{k+1}\right]\mathbf{1}\{R \leq k\}] \leq 0,$$

our claim will be proved. Now, for each k, we have

$$\mathbb{E}\Big[\Big[\frac{\mathbf{1}\{p_i \leq q_k\}}{k} - \frac{\mathbf{1}\{p_i \leq q_{k+1}\}}{k+1}\Big]\mathbf{1}\{R \leq k\}\Big]$$

$$= \frac{\mathbb{P}(p_i \leq q_k, R \leq k)}{k} - \frac{\mathbb{P}(p_i \leq q_{k+1}, R \leq k)}{k+1}$$

$$= \frac{\mathbb{P}(R \leq k|p_i \leq q_k)\mathbb{P}(p_i \leq q_k)}{k} - \frac{\mathbb{P}(R \leq k|p_i \leq q_{k+1})\mathbb{P}(p_i \leq q_{k+1})}{k+1}$$

$$\leq \frac{\mathbb{P}(R \leq k|p_i \leq q_{k+1})\mathbb{P}(p_i \leq q_k)}{k} - \frac{\mathbb{P}(R \leq k|p_i \leq q_{k+1})\mathbb{P}(p_i \leq q_{k+1})}{k+1} \quad \text{by PRDS}$$

$$= \mathbb{P}(R \leq k|p_i \leq q_{k+1})\Big[\frac{\mathbb{P}(p_i \leq q_k)}{k} - \frac{\mathbb{P}(p_i \leq q_{k+1})}{k+1}\Big]$$

$$= \mathbb{P}(R \leq k|p_i \leq q_{k+1})\Big[\frac{k\alpha}{m} \cdot \frac{1}{k} - \frac{(k+1)\alpha}{m} \cdot \frac{1}{k+1}\Big]$$

$$= 0.$$

For the inequality, we used the PRDS property and said that

$$\mathbb{P}(R \leq k|p_i \leq q_k) \leq \mathbb{P}(R \leq k|p_i \leq q_{k+1}).$$

By definition of PRDS, for any increasing set D and each $i \in I_0 \subset I$, $P(\mathbf{X} \in D|X_i = x)$ is non-decreasing as x increases. A consequence of this is that if we have $x \leq x'$, we would have

$$\mathbb{P}(\mathbf{X} \in D|X_i \leq x) \leq \mathbb{P}(\mathbf{X} \in D|X_i \leq x').$$

Notice that when p_i increases, we will have less rejections and make $\{R \leq k\}$ increases. So, $\{R \leq k\}$ is indeed an increasing set. $\qquad\square$

Chapter 4

FDR Control using Empirical Bayes

The following content are mostly adapted from the book "Large-Scale Inference" by Efron (2010).

4.1 Bayesian Hypothesis Testing

In the Frequentist setting, or more specifically the Neyman-Pearson setting, we are making a decision between two hypothesis - null hypothesis H_0 and alternative hypothesis H_1. In the Bayesian setting, things are similar but the result of the test is viewed as a random variable H. If the null hypothesis is true, we will have $H = 0$; and if the null hypothesis is false, we will have $H = 1$. Their prior probabilities are $\mathbb{P}(H = 0) = \pi_0$ and $\mathbb{P}(H = 1) = \pi_1 = 1 - \pi_0$.

The test statistics will be denoted as Z, and it has different distributions under true and false null hypothesis. When $H = 0$, we have $Z \sim f_0$ with CDF F_0. When $H = 1$, we have $Z \sim f_1$ with CDF F_1. Combining these two, we have a hierarchical model with $H \sim \text{Bernoulli}(\pi_1)$ and $Z \sim F_H$. So, the pdf of Z is $\mathbb{P}(Z = z) = f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$.

Now, the probability of the null hypothesis being true or false will be updated once we have collected data. Assuming we observe a test statistics value $Z = z$, we would like to know the probabilities $\mathbb{P}(H = 0|z)$ and $\mathbb{P}(H = 1|z)$. If one is bigger than the other, we will know which hypothesis we should pick. To compute these two probabilities, we will use the Bayes rule. This gives us

$$\mathbb{P}(H = 0|z) = \frac{\mathbb{P}(z|H = 0) \cdot \mathbb{P}(H = 0)}{\mathbb{P}(Z = z)} = \frac{\pi_0 \cdot f_0(z)}{f(z)}$$

and

$$\mathbb{P}(H = 1|z) = \frac{\mathbb{P}(z|H = 1) \cdot \mathbb{P}(H = 1)}{\mathbb{P}(Z = z)} = \frac{\pi_1 \cdot f_1(z)}{f(z)}.$$

The comparison between $\mathbb{P}(H = 0|z)$ and $\mathbb{P}(H = 1|z)$, after some computations, is equivalent to the comparison between $\pi_0 \cdot f_0(z)$ and $\pi_1 \cdot f_1(z)$.

The ratio of these two probabilities has its own name, the **Bayes factor**. Normally, we will put the probability for the null hypothesis at the denominator. There are certain existing thresholds for this factor for people to determine how strong the evidence is to choose a hypothesis, just like the frequently used 0.05 and 0.1 significance level for Frequentist tests.

4.2 Bayes FDR

The setting in the above section can be extended to the case where we are doing multiple hypothesis testing. The setting will be identical, and we will let the total number of tests be N.

Now, for a subset A of the real line, we define $\varphi(A) = \mathbb{P}(H = 0 | z \in A)$. Using Bayes rule, we have

$$\varphi(A) = \mathbb{P}(H = 0 | z \in A) = \frac{\mathbb{P}(H = 0) \cdot \mathbb{P}(z \in A | H = 0)}{\mathbb{P}(z \in A)} = \frac{\pi_0 \cdot F_0(A)}{F(A)}$$

where $F_0(A) = \int_A f_0(z) dz$ and $F(A) = \int_A f(z) dz$.

If we set A in such a way that $z \in A$ means the null hypothesis if false, the above probability will be the probability of the null hypothesis actually being false given that we reject it, which is exactly what false discovery rate is about. Thus, the quantity $\varphi(A)$ where $z \in A$ means the null hypothesis if false is known as the **Bayes false discovery rate**, or BFDR. For simplicity, we will also write $\varphi(A)$ as Fdr(A).

This quantity is a random variable and it involves π_0, f_0, and f_1. Here, π_0 is almost known, and is usually near 1 in practice. Since we are normally calculating this quantity in the context of multiple testing like genetic research, we would only have a very small portion of false null hypothesis, making the ratio close to 1. The function f_0 is also known, since we would normally have z-values (or z-scores) as the test statistics, which follows a standard normal distribution $N(0, 1)$. The remaining f_1 is unknown, and it is hard to find out.

To deal with difficulties of this kind, people thought, since we have so many data available, why don't we do an estimate of the distribution of f as that is the only unknown? If we can estimate f, we can estimate the target quantity. This idea of improving Bayesian inference with Frequentist estimations grows out to be a whole branch of thoughts known as the **empirical Bayes**, with this name coined by its founder Herbert Robbins in Robbins (1956).

We let $\bar{F}(A)$ be the empirical distribution of the N z-values and define it to be

$$\bar{F}(A) = \frac{\#\{z \in A\}}{N},$$

i.e. the proportion of the z-values that are being rejected. Now, substituting this value into Fdr, we get

$$\overline{\text{Fdr}}(A) = \bar{\varphi}(A) = \frac{\pi_0 \cdot F_0(A)}{\bar{F}(A)}.$$

For large values of N, we would expect $\bar{F}(A)$ to be a good estimation of $F(A)$, and by extension $\overline{\text{Fdr}}(A)$ being a good estimation of Fdr(A).

4.3 BH Procedure using Empirical Bayes

Recall from earlier chapter that the BH procedure is a method that controls the FDR of the multiple testing. Now that we have Bayes FDR, can we rewrite BH procedure using that?

The BH procedure of n hypothesis tests that controls the FDR at level α will reject all the smallest k p-values with

$$k = \max\{i : p_{(i)} \leq \frac{i}{n}\alpha\}.$$

We can map the p-values with z-values using

$$p_i = F_0(z_i)$$

for $F_0(\cdot)$ being the cdf of Z when the null hypothesis is true, i.e. being $N(0,1)$. This means each $p_{(i)}$ will be mapped to $F_0(z_{(i)})$, and

$$\frac{i}{n} = \frac{\#\{z_j \le z_{(i)}\}}{n} = \bar{F}(z_{(i)}).$$

So, the BH procedure threshold can be transformed, and we have

$$p_{(i)} \le \frac{i}{n}\alpha \implies \frac{p_{(i)}}{i/N} \le \alpha \implies \frac{F_0(z_{(i)})}{\bar{F}(z_{(i)})} \le \alpha \implies \overline{\text{Fdr}}(z_{(i)}) = \pi_0 \cdot \frac{F_0(z_{(i)})}{\bar{F}(z_{(i)})} \le \pi_0 \alpha \le \alpha.$$

Thus, the BH procedure using empirical Bayes controlling FDR at will be rejecting all the k hypothesis with k being the largest value of index i with $\overline{\text{Fdr}}(z_{(i)}) \le \alpha$.

It is not straightforward to understand the rationale behind the BH procedure. Why is the critical line what it is? Why do we reject until the last crossing? It is all a bit mysterious and we can minimally justify BH procedure by simply stating and proving that it works. The empirical Bayes approach shows the power of thinking about multiple testing problems in a Bayesian framework. Using the Bayesian machinery, we are able to arrive at the BH procedure far more simply than by restricting ourselves to the Frequentist framework. We only need to compute the estimated Bayes FDR of each one of the tests and cut off at α level. The procedure become easier to understand.

4.4   Quality of Bayesian FDR estimator

The empirical Bayes BH procedure is phrased using $\overline{\text{Fdr}}$, the Bayesian FDR estimator. How good is this estimator exactly? Is it good enough so that the procedure will yield satisfying results? In this section we will be exploring the quality of the estimator.

To find out how good an estimator $\overline{\text{Fdr}}(A)$ is, we need to adopt some notations. We let $N_0(A) = \#\{i : z_i \in A \text{ and being true null}\}$ denote the number of true nulls being rejected, and $e_0(A) = \mathbb{E}[N_0(A)]$ be its expectation. Furthermore, we let $N_1(A) = \#\{i : z_i \in A \text{ and being false null}\}$ denote the number of false nulls being rejected, and $e_1(A) = \mathbb{E}[N_1(A)]$ be its expectation. So, the total number of rejected hypothesis will be $N_+(A) = \#\{i : z_i \in A\} = N_0(A) + N_1(A)$, and its expectation is $e_+(A) = \mathbb{E}[N_+(A)]$. So, we would have

$$\overline{\text{Fdr}}(A) = \frac{\pi_0 \cdot F_0(A)}{\bar{F}(A)} = \frac{n\pi_0 \cdot F_0(A)}{n\bar{F}(A)} = \frac{e_0(A)}{N_+(A)}$$

and

$$\text{Fdr}(A) = \frac{e_0(A)}{e_+(A)}.$$

In addition, we will need another quantity, the false discovery proportion. The false discovery portion, Fdp, is defined to be

$$\text{Fdp}(A) = \frac{N_0(A)}{N_+(A)}$$

and this is the same quantity as defined in earlier chapter. So, we would also have $\text{Fdr}(A) = \mathbb{E}[\text{Fdp}(A)]$. To discuss the relationships between the above mentioned quantities and illustrate how good $\overline{\text{Fdr}}(A)$, we will state without proof the following two lemmas.

Lemma 2. Suppose $e_0(A)$ is the same as the conditional expectation of $N_0(A)$ given $N_1(A)$. Then the conditional expectation of $\overline{\text{Fdr}(A)}$ and $\text{Fdp}(A)$ given $N_1(A)$ satisfy

$$\mathbb{E}[\overline{\text{Fdr}(A)}|N_1(A)] \geq \varphi_1(A) \geq \mathbb{E}[\text{Fdp}(A)|N_1(A)]$$

where

$$\varphi_1(A) = \frac{e_0(A)}{e_0(A) + N_1(A)}.$$

This lemma says that for every value of $N_1(A)$, the conditional expectation of $\overline{\text{Fdr}(A)}$ exceeds that of $\text{Fdp}(A)$, so that in the sense the empirical Bayes FDR is a conservatively biased estimate of the actual FDP. Taking expectation over $N_1(A)$ and applying Jensen's inequality shows that

$$\varphi(A) \geq \mathbb{E}[\text{Fdp}(A)] = \text{FDR}(A),$$

so that the Bayes FDR is an upper bound of the traditional FDR.

Lemma 3. Let $\gamma(A)$ indicate the squared coefficient of variation of $N_+(A)$,

$$\gamma(A) = \frac{\text{Var}[N_+(A)]}{e_+(A)^2}.$$

Then $\overline{\text{Fdr}(A)}/\varphi(A)$ has approximate mean $1 + \gamma(A)$ and variance $\gamma(A)$.

This lemma quantifies the obvious: the accuracy of $\overline{\text{Fdr}(A)}$ as an estimate of the Bayes FDR depends on the variability of the denominator of $N_+(A)$.

Chapter 5

FDR Control using E-Values

The following content are mostly adapted from the paper by Wang and Ramdas (2020).

5.1 E-Variable and E-Value

As mentioned in earlier chapter, p-value is a random variable of the probability that the test statistics is at least as extreme under the assumption that the null hypothesis is true. This random variable is phrased in terms of probability of an event. Here, we proposed a different random variable, the **e-variable**, which can serve as either as an enhancement or as an alternative to p.

For data X following distribution P and null hypothesis $\mathcal{H}_0$, the e-variable is a nonnegative random variable E for testing null hypotheses $\mathcal{H}_0$ that is defined by

$$\sup_{P_0 \in \mathcal{H}_0} \mathbb{E}_{P_0} E(X) \leq 1.$$

Here, the realisations of e-variables are called e-values to avoid confusion between the random variable and its realisation like that of the p-value.

To illustrate its difference to p-value, let us define p-variable (the random variable version of p-value). A random variable P is called a p-variable for testing $\mathcal{H}_0$ if

$$\sup_{P_0 \in \mathcal{H}_0} \mathbb{P}_{P_0}(P(X) \leq \alpha) \leq \alpha \qquad \forall \alpha \in (0,1).$$

We can see immediately that p-variables control the probability while e-variables control expectation. To relate the two concepts, we have the following theorem.

Theorem 3. Let E be an e-variable. If $P = E^{-1}$, then $\mathbb{P}(P \leq \alpha) \leq \alpha$.

Proof. We fix a $P_0 \in \mathcal{H}_0$. Then, we have

$$
\begin{aligned}
\mathbb{P}_{P_0}(P \leq \alpha) &= \mathbb{P}_{P_0}\left(\frac{1}{E} \leq \alpha\right) \\
&= \mathbb{P}_{P_0}\left(E \geq \frac{1}{\alpha}\right) \\
&\leq \frac{\mathbb{E}_{P_0}(E)}{1/\alpha} \quad \text{using Markov's inequality} \\
&= \alpha \mathbb{E}_{P_0}(E) \leq \alpha \quad \text{by the definition of } E.
\end{aligned}
$$

$\square$

Remark. This E^{-1} is rather conservative, since the bound by Markov's inequality may not be tight.

5.2 e-BH procedure

Earlier on, we have mentioned about the BH procedure, and it involves cutting off at the largest i for which $p_{(i)} \leq \frac{i}{m}\alpha$. Since we have already shown that we could replace P with E^{-1}, we could modify the BH procedure using E instead of P.

For each of the hypothesis $H_1, H_2, \cdots, H_m$, they have the e-values $e_1, e_2, \cdots, e_m$. We will order them in descending order, and have them as $e_{(1)} \geq e_{(2)} \geq \cdots \geq e_{(m)}$. We will reject all the hypotheses with the largest k e-values where k is the largest i for which $\frac{ie_{(i)}}{m} \geq \frac{1}{\alpha}$. This should not be a surprise. Since $p_{(i)} = 1/e_{(i)}$, we have

$$p_{(i)} \leq \frac{i}{m}\alpha \implies \frac{1}{p_{(i)}} \geq \frac{m}{i\alpha} \implies e_{(i)} \geq \frac{m}{i} \cdot \frac{1}{\alpha} \implies \frac{i}{m}e_{(i)} \geq \frac{1}{\alpha}.$$

This modified version of BH procedure using e-values is known as the e-BH procedure. To make sure this procedure really works, we will prove its FDR control.

Theorem 4. The e-BH procedure has FDR at most $m_0\alpha/m$.

Proof. Recall from earlier, we have the definition

$$\text{FDP} = \frac{V}{R \vee 1} = \sum_{i \in H_0} \frac{V_i}{R \vee 1}$$

where we sum over all the indices of true null hypotheses H_0. Now, for any rejected i, we will have

$$\frac{1}{R} \leq \frac{1}{i} \leq \frac{i}{m}e_{(i)}$$

where the first inequality is due to the fact that $i \leq R$. We will take the first and last term of the inequality after multiplying V_i and sum over all true null. This means, we have

$$\text{FDP} = \sum_{i \in H_0} \frac{V_i}{R \vee 1} \leq \sum_{i \in H_0} \frac{V_i i e_{(i)}}{m}$$

and this gives us

$$\text{FDP} \leq \sum_{i \in H_0} \frac{V_i i e_{(i)}}{m} \leq \sum_{i \in H_0} \frac{i e_{(i)}}{m} \leq \frac{\alpha}{m} \sum_{i \in H_0} e_{(i)}.$$

Taking expectation of the above inequality, we have

$$\text{FDR} = \mathbb{E}[\text{FDP}] \leq \mathbb{E}[\frac{\alpha}{m} \sum_{i \in H_0} e_{(i)}] = \frac{\alpha}{m} \sum_{i \in H_0} \mathbb{E}[e_{(i)}] \leq \frac{\alpha}{m} \sum_{i \in H_0} 1 = \frac{m_0}{m}\alpha.$$

$\square$

Bibliography

[1] Benjamini Y., Hochberg Y. (1995) Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B*, **57** No. 1, 289-300.

[2] Benjamini Y., Yekutieli D. (2001) The Control of the False Discovery Rate in Multiple Testing Under Dependency. *The Annals of Statistics*, **29**, No. 4, 1165-1188.

[3] Benjamini Y. (2010) Discovering the False Discovery Rate. *Journal of the Royal Statistical Society, Series B*, **72**, Part 4, 405-416.

[4] Efron, B. (2010) *Large-Scale Inference*. Cambridge University Press.

[5] Hochberg, Y.,Tamhane, A. C. (1987) *Multiple Comparison Procedures*. Wiley.

[6] Holm S. (1979) A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*. **6**, 65-70.

[7] Robert H. (1956) An Empirical Bayes Approach to Statistics, *Berkeley Symposium on Mathematical Statistics and Probability*, **3.1**, 157-163.

[8] Ronald C. (2005) Testing Fisher, Neyman, Pearson, and Bayes, *The American Statistician*, **59**, No.2, 121-126.

[9] Soriç, B. (1989) Statistical "discoveries" and effect size estimations. *Journal of the American Statistical Association*. **84**, 608-610.

[10] Wang, R., Ramdas, A. (2020) False discovery rate with e-values. *arXiv preprint,* arXiv:2009.02824v2.