Contents

1 Multiple Testing **2**
1.1 Discourse on Multiple Testing 2
1.2 The p-value 3
1.3 Family-wise Error Rate 4
1.4 False Discovery Rate 4

1
Multiple Testing

1.1 Discourse on Multiple Testing

In many statistical studies, a large number of hypotheses are tested simultaneously. For example, a genome-wide association study (GWAS) involves scanning loci across the complete sets of DNA or genomes of a sample of subjects in order to find genetic variations in the hopes of associating particular loci with a given trait or disease. A locus is a specific location of a gene or some other part of the DNA sequence, such as a Single Nucleotide Polymorphism, or SNP. The study may comprise diseased subject and controls, or subjects with different degrees of severity of some disease.

For each of many thousands of loci in the DNA sequence we set a null hypothesis that it is not related to the disease in question. Rejecting this null amounts to a discovery of a gene or a SNP which is associated with the disease. If we test for m loci at level of significance $\alpha = 0.01$, and $m = 20,000$, say, which is roughly the number of protein-coding human genes, then we should expect about 200 'discoveries' even if the disease has no genetic component. Such a number of false discoveries is a reason for concern, which we address in this chapter.

For another example consider the well-known issue of disease clusters which appear occasionally in certain areas and groups, and are sometimes ascribed to some source of contamination. How does one prove that a high incidence of a certain disease in a particular group is caused by the alleged source, and is not due to random variation? If a large number of different groups of people who at some period in the past lived in the same region,

say, get together and search for common diseases from large lists of diseases, it is inevitable that there will arise groups with a higher than average incidence in certain diseases. They may then believe, and claim, that their observed higher incidence is significant and not random, and that pollution or some other problem in the region in question is the cause.

While the GWAS example above is a planned scientific study with a well-defined search space and a *family of hypotheses*, it is usually not so in disease cluster examples. In fact, various types of exploratory data-mining in medical records may clearly lead to groups and diseases that may appear to be suspicious clusters, and these may then presented to the public and to courts without ever defining the extent of the search. In studies having clear search procedure and a list of hypotheses, it is not clear how to formally determine statistical significance and control the error rate of the combined inference.

In cases where the search and the associated family of hypotheses are well defined, one can take the large number of hypotheses into consideration in performing confirmatory data analysis that controls the error rates in various ways. For a reference on multiple testing procedures,see, for example, Hochberg and Tamhane (1987).

1.2 The p-value

Given data X, a statistic $T = T(X)$, a hypothesis H_0 and rejection region A, the probability $P_{H_0}(T \in A)$ is well defined when H_0 is a simple hypothesis. For H_0 composite we will take $P_{H_0}(T \in A)$ as $\sup_H P(T \in A)$, where the sup is taken over all the hypotheses H in H_0. This notation is used in the following definition of the p-value, which is not the most general, but suffices for most purposes.

Definition 1.2.1 *Given data X, a statistic $T = T(X)$, a simple or composite hypothesis H_0 and a procedure that rejects H_0 for $T \geq c$ for some c, set $G(t) = P_{H_0}(T \geq t)$. The p-value associated with an observed value of T is defined to be $G(T)$, and will be denoted by P.*

In words, the p-value is the probability of getting a value at least as extreme as the one associated with the observed T. A small p-value suggests that T is extreme and H_0 should be rejected. In particular a p-value less than some small α implies rejection of H_0 at significance level α.

For a given significance level $\alpha \in (0, 1)$, define

$$c_\alpha = \inf\{c : P_{H_0}(T \geq c) \leq \alpha\},$$

the critical value of the test statistic for level α.

Proposition 1.2.2 *We have $P = \inf\{\alpha : T \geq c_\alpha\}$.*

Proof: Note that $c_\alpha = \inf\{c : G(c) \leq \alpha\}$. Given T, let α satisfy $T \geq c_\alpha$. This implies $G(T) \leq \alpha$ and thus the p-value P satisfies $P \leq \alpha$. This proves that $P \leq \inf\{\alpha : T \geq c_\alpha\}$. If $P < \inf\{\alpha : T \geq c_\alpha\}$ then for some α_0 we have $G(T) = P = \alpha_0$ and $T < c_{\alpha_0}$. The latter inequality implies $G(T) > \alpha_0$, a contradiction. □

Proposition 1.2.3 *For all $u \in (0, 1)$ it holds that $P_{H_0}(P \leq u) \leq u$.*

Proof: From the definition of c_u we have that $P_{H_0}(T \geq c_u) \leq u$. This implies $P_{H_0}(P \leq u) = P_{H_0}(\inf\{\alpha : T \geq c_\alpha\} \leq u) = P_{H_0}(T \geq c_u) \leq u$. □

If T is a continuous variable and H_0 is simple then $P \sim \mathcal{U}[0, 1]$; see Exercise 1.4.1.

1.3 Family-wise Error Rate

Family-wise error rate (FWER) is the probability of making one or more false discoveries, that is, rejection of a true H_0^i, when hypotheses $H_0^1, \ldots, H_0^m$ are being tested simultaneously. If V denotes the number of false rejections, then when all hypotheses are true, the probability of rejecting at least one of them is

$$\text{FWER} = P(V \geq 1) \tag{1.1}$$

where P is any probability distribution for which all H_0^i are true; the FWER criterion requires FWER $\leq \alpha$ for a prescribed value of α.

Proposition 1.3.1 *Let P_i denote the p-values associated with m null hypotheses, $H_i, i = 1, \ldots, m$. Then the procedure that rejects the hypotheses for which $P_i \leq \alpha/m$ satisfies* FWER $\leq \alpha$.

Proof: Let $G_i(t) = P_{H_i}(T_i \geq t)$. By Proposition 1.2.3, $P_{H_0^i}(P_i \leq \alpha/m) \leq \alpha/m$. Now FWER $\leq P(\bigcup_{i=1}^m \{P_i \leq \alpha/m\})$, where the probability P is any probability under the assumption that all H_0^i hold. The inequality between FWER and the probability of the union follows from the fact that if some null hypotheses are untrue, they can be omitted from the union. By Bonferroni's inequality FWER $\leq \sum_{i=1}^m P_{H_0^i}(P_i \leq \alpha/m) \leq \alpha$.

1.4 False Discovery Rate

The term discovery refers to the fact that in many situations, the rejection of a null hypothesis amounts to a discovery. For example, a typical null hypothesis is that two treatments, and old one and a proposed new one, are equal, and a discovery would be that the new treatment is better.

Suppose tests are performed on m null hypotheses, each producing a p-value. To achieve a type I error level α over the collection of tests, the

Bonferroni procedure rejects all those hypotheses having p-values no greater than α/m. Though this procedure indeed has Type 1 error no more than α, it may be overly conservative. Following hoc and Hochberg and Benjamini (1990) we discuss a criterion that leads to a different procedure based on these same p-values.

Let R be the total number of null hypotheses rejected, sometimes called discoveries, and of those, V, the number false rejections. We set

$$Q = V/R,$$

the proportion of rejections that were false; we set $Q = 0$ when $R = 0$. The *False Discovery Rate*, or *FDR*, is defined as

$$\text{FDR} = \text{E}[Q],$$

represents the expected proportion of false rejections out of the total number of rejections. Thus, the FDR represents the expected number of false discoveries relative to the total number of claimed discoveries. A procedure that satisfies the constraint

$$\text{FDR} \leq \alpha \tag{1.2}$$

allows a scientist to err in proportion to the total number of discoveries in a given study.

For example, if in a genetic study we wish to discover which of m DNA loci are associated with a certain phenotypical trait, such as the height of a plant, then the FDR criterion (1.2) allows an average proportion of α of the declared discoveries to be false, that is, 'discoveries' of loci not associated with the given trait. This criteria should be compared with the Family Wise Error Rate approach which sets a bound on the probability of making one or more false discoveries, which will only be small when with high probability all discoveries are true ones, and thus errors are allowed only with small probability.

Let $P_{(1)} \leq \cdots \leq P_{(m)}$ be the ordered p-values corresponding to the m tested hypotheses. Define

$$K = \max\{i : P_{(i)} \leq i\alpha/m\}.$$

The *Benjamini Hochberg* procedure rejects the null hypotheses corresponding to $P_{(1)}, \ldots, P_{(K)}$. Note that among the rejected hypotheses the largest p-value must satisfy $P_{(i)} \leq i\alpha/m$ for $i = K$, but for $i < K$ this inequality need not hold. The thresholds $i\alpha/m$ in the BH procedure against which the p-values are tested are never smaller than α/m, the threshold of the Bonferroni procedure, leading to more rejected hypotheses, that is, more claimed discoveries.

Here is a simple example. Suppose $m = 3$ and $P_3 < P_1 < P_2$. Then, writing these values in increasing order as $P_{(1)} < P_{(2)} < P_{(3)}$, if the largest of them, $P_{(3)} = P_2$ is strictly smaller than $3\alpha/3 = \alpha$ then all three hypotheses

are rejected. If not, but $P_{(2)} = P_1 < 2\alpha/3$ then the hypotheses corresponding to $P_{(2)} = P_1$ and $P_{(1)} = P_3$, that is, the first and third hypotheses are rejected, and so on.

To study the behavior of the BH procedure we make the following definition.

Definition 1.4.1 *The random variables $X_1, \ldots, X_n$ are PRDS (Positive Regression Dependence on each one from a Subset) for $S \subseteq \{1, \ldots, n\}$ if, for any coordinatewise nondecreasing function g, and for each $i \in S$, $E[g(X_1, \ldots, X_n) \mid X_i = x]$ is nondecreasing in x. If $S = \{1, \ldots, n\}$ we say that the variables are PRDS*

Independent random variables are clearly PRDS. Note that PRDS implies PRDS on any S, so PRDS is a stronger assumption. We will actually require a weaker property than PRDS, as defined in the following.

Proposition 1.4.2 *If the random variables $X_1, \ldots, X_n$ are PRDS on $S \subseteq \{1, \ldots, n\}$ then for any coordinatewise nondecreasing function g, and for each $i \in S$, $E[g(X_1, \ldots, X_n) \mid X_i \leq x]$ is nondecreasing in x.*

Proof: The claim follows readily from the fact that if $h(x)$ is nondecreasing in x then $\int_{-\infty}^{x} h(t)dF(t)/F(x)$ is nondecreasing in x for any distribution function F. This fact and the rest of the proof is left to the reader, see Exercises 1.4.2 and 1.4.3, with the hint that

$$E[g(X_1, \ldots, X_n) \mid X_i \leq x] = \int_{-\infty}^{x} E[g(X_1, \ldots, X_n) \mid X_i = t]dF_{X_i}(t)/F_{X_i}(x).$$

Theorem 1.4.3 *Suppose of m null hypotheses being tested that exactly m_0 are true, and let $P_1, \ldots, P_m$ be PRDS on the set of indices corresponding to the true null hypotheses. Then the BH procedure guarantees $E[Q] \leq \alpha m_0/m$. Equality holds when $P_1, \ldots, P_m$ are independent, the statistics T_i are continuous, and the null hypotheses are simple.*

Proof: Let I_i denote the indicator of the event that under the BH procedure, H_0^i is rejected. Recalling that R is the number of rejected hypotheses, we first claim that

$$\{R = r, I_i = 1\} = \{R = r, P_i \leq r\alpha/m\} \quad \text{for all } i = 1, \ldots, m. \tag{1.3}$$

The verification is left as Exercise 1.4.4.

Assuming without loss of generality that the true m_0 hypotheses are $H_0^1, \ldots, H_0^{m_0}$, we have $Q = \sum_{i=1}^{m_0} I_i/R$. Since

$$E(I_i/R) = \sum_{r=1}^{m} \frac{1}{r} P(R = r, I_i = 1),$$

we obtain, first by using (1.3),

$$
\begin{aligned}
EQ &= \sum_{i=1}^{m_0}\sum_{r=1}^{m}\frac{1}{r}P(R=r, I_i=1) \\
&= \sum_{i=1}^{m_0}\sum_{r=1}^{m}\frac{1}{r}P(R=r \mid P_i \le r\alpha/m)P(P_i \le r\alpha/m) \\
&\le \frac{\alpha}{m}\sum_{i=1}^{m_0}\sum_{r=1}^{m}P(R=r \mid P_i \le r\alpha/m) = \frac{\alpha}{m}\sum_{i=1}^{m_0}\Big[P(R\ge 1 \mid P_i \le \alpha/m) \\
&\quad + \sum_{r=2}^{m}P(R\ge r \mid P_i \le r\alpha/m) - P(R \ge r \mid P_i \le (r-1)\alpha/m)\Big] \\
&\le \frac{\alpha}{m}\sum_{i=1}^{m_0}P(R\ge 1 \mid P_i \le \alpha/m) = \alpha m_0/m, \quad (1.4)
\end{aligned}
$$

where the first inequality holds by Proposition 1.2.3, and the second follows by PRDS which implies

$$
P(R \ge r \mid P_i \le r\alpha/m) - P(R \ge r \mid P_i \le (r-1)\alpha/m) \le 0, \qquad (1.5)
$$

see exercise 1.4.5.

Equality holds in the first inequality if $H_0^1, \ldots, H_0^{m_0}$ are simple and T has a continuous distribution function, see remark after Proposition 1.2.3. The second inequality is an equality under independence of the P_i's, since given $P_i \le \kappa\alpha/m$ for $\kappa \le r$ the events $\{R \ge r\}$ holds if and only if for some $j \ge r$ there exists $j-1$ p values, excluding P_i, satisfying $P_{(k)} \le k\alpha/m$, a condition that does not involve P_i. We apply this to the terms in (1.5) with $\kappa = r$ and $\kappa = r-1$, respectively. □

The above proof is due to Finner et al. (2009) Note that it suffices to assume PRDS on the set of true H_0^i's. This is not very useful, since in general we don't know which hypotheses are true, of course.

In the case that $m_0 = m$, that is, all null hypothesis are true, we have $V = R$ since any rejection of a null hypothesis is false. In this case $Q = 1$ if $R > 0$ and 0 otherwise and Theorem 1.4.3 implies that $P(R > 0) \le \alpha$ so that the BH procedure guarantees FWER $\le \alpha$.

Exercise 1.4.1 *If T is a continuous variable and H_0 is simple then $P \sim \mathcal{U}[0,1]$.*

Exercise 1.4.2 *We recall that for two random variables U and V we say that U is stochastically dominated by V, and write $U \le_{\mathrm{st}} V$, when*

$$
E[h(U)] \le E[h(V)] \quad \textit{for all non-decreasing functions } h.
$$

Prove that $U \leq_{\rm st} V$ if and only if there exists a joint distribution for the pair (U, V) having the given marginal distributions, and satisfying

$$P(U \leq V) = 1.$$

Hint: Consider applying the inverse probability integral transformation.

Exercise 1.4.3 *For any random variable X and $u \in \mathbb{R}$ let X_u be such that*

$$P(X_u \leq x) = P(X \leq x | X \leq u) \quad \textit{for all } x \leq u.$$

Prove, perhaps by using Exercise 1.4.2, that the distributions $X_u, u \in \mathbb{R}$ are stochastically increasing in u, that is,

$$X_s \leq_{\rm st} X_t \quad \textit{for all } s \leq t.$$

Hint: Note that $p = P(X_t \leq s)$ and $q = P(s < X_t \leq t)$ are nonnegative and sum to one. Consider generating X_s and X_t in a way such that $X_s \leq X_t$ with probability one by flipping a coin with success probablity p, and adopting different strategies for generating the variables depending on whether the coin comes up 'heads' or 'tails'.

Now show how the claim in Proposition 1.4.2 follows, that is, that $\int_{-\infty}^{x} h(t)dF(t)/F(x)$ is nondecreasing in x whenever $h(x)$ is nondecreasing in x.

Exercise 1.4.4 *Show 1.3.*

Exercise 1.4.5 *Show 1.5.*

References

(????), "A sharper b."

Finner, Helmut, Thorsten Dickhaus, Markus Roters, et al. (2009), "On the false discovery rate and an asymptotically optimal rejection curve." *The Annals of Statistics*, 37, 596–618.

Hochberg, Yosef and Yoav Benjamini (1990), "More powerful procedures for multiple significance testing." *Statistics in medicine*, 9, 811–818.

Hochberg, Yosef and Ajit C Tamhane (1987), *Multiple comparison procedures*. John Wiley & Sons, Inc.