Chapter 4

False Discovery Rate Control

Applied statistics is an inherently conservative enterprise, and appropriately so since the scientific world depends heavily on the consistent evaluation of evidence. Conservative consistency is raised to its highest level in classical significance testing, where the control of Type I error is enforced with an almost religious intensity. A p-value of 0.06 rather than 0.04 has decided the fate of entire pharmaceutical companies. Fisher's scale of evidence, Table 3.1, particularly the $\alpha = 0.05$ threshold, has been used in literally millions of serious scientific studies, and stakes a good claim to being the 20th century's most influential piece of applied mathematics.

All of this makes it more than a little surprising that a powerful rival to Type I error control has emerged in the large-scale testing literature. Since its debut in Benjamini and Hochberg's seminal 1995 paper, false discovery rate control has claimed an increasing portion of statistical research, both applied and theoretical, and seems to have achieved "accepted methodology" status in scientific subject-matter journals.

False discovery rate control moves us away from the significance-testing algorithms of Chapter 3, back toward the empirical Bayes context of Chapter 2. The language of classical testing is often used to describe Fdr methods (perhaps in this way assisting their stealthy infiltration of multiple testing practice), but, as the discussion here is intended to show, both their rationale and results are quite different.

4.1 True and False Discoveries

We wish to test N null hypotheses

$$H_{01}, H_{02}, \ldots, H_{0N} \tag{4.1}$$

on the basis of a data set X, and have in mind some decision rule $\mathcal{D}$ that will produce a decision of "null" or "non-null" for each of the N cases. Equivalently,[1] $\mathcal{D}$ *accepts* or *rejects* each H_{0i}, $i = 1, 2, \ldots, N$, on the basis of X. X is the 6033×102 matrix of expression values in the prostate data example of Section 2.1, giving the N z-values (2.5), while $\mathcal{D}$ might be the rule that rejects H_{0i} if $|z_i| \geq 3$ and accepts H_{0i} otherwise.

[1] I am trying to avoid the term "significant" for the rejected cases as dubious terminology even in single-case testing, and worse in the false discovery rate context, preferring instead "interesting".

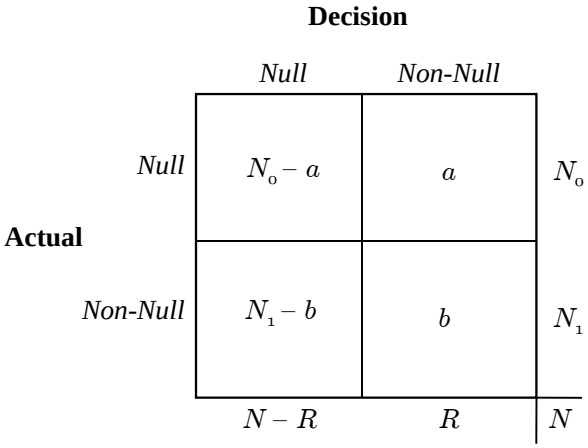

Figure 4.1: A decision rule $\mathcal{D}$ has rejected R out of N null hypotheses (4.1); a of these decisions were incorrect, i.e., they were "false discoveries", while b of them were "true discoveries." The false discovery proportion Fdp equals a/R.

Figure 4.1 presents a hypothetical tabulation of $\mathcal{D}$'s performance from the point of view of an omniscient oracle: N_0 of the N cases were actually null, of which $\mathcal{D}$ called a non-null (incorrectly) and $N_0 - a$ null (correctly); likewise, N_1 were actually non-null, with $\mathcal{D}$ deciding b of them correctly and $N_1 - b$ incorrectly. Of the $R = a + b$ total rejections, a were "false discoveries" and b "true discoveries", in the current terminology. The family-wise error rate of Section 2.2, FWER, equals $\Pr\{a > 0\}$ in terms of the figure.

N equals 1 in the classical single-case testing situation, so either N_0 or N_1 equals 1, with the other 0. Then

$$\Pr\{a = 1 | N_0 = 1\} = \alpha, \tag{4.2}$$

the Type I error rate, or size, of the decision rule, and

$$\Pr\{b = 1 | N_1 = 1\} = \beta, \tag{4.3}$$

the rule's power.

Exercise 4.1. In a multiple testing situation with both N_0 and N_1 positive, show that

$$E\left\{\frac{a}{N_0}\right\} = \bar{\alpha} \quad \text{and} \quad E\left\{\frac{b}{N_1}\right\} = \bar{\beta}, \tag{4.4}$$

$\bar{\alpha}$ and $\bar{\beta}$ being the average size and power of the null and non-null cases, respectively.

Classical Fisherian significance testing is immensely popular because it requires so little from the scientist: only the choice of a test statistic and specification of its probability distribution when the null hypothesis is true. Neyman–Pearson theory adds the specification of a non-null distribution, the reward being the calculation of power as well as size. Both of these are calculated *horizontally* in the figure, that is restricting attention to either the null or non-null row, which is to say that they are frequentist calculations.

Large-scale testing, with N perhaps in the hundreds or thousands, opens the possibility of calculating *vertically* in the figure, in the Bayesian direction, without requiring Bayesian priors. The ratio a/R is what we called the *false discovery proportion* (2.28), the proportion of rejected cases that are actually null. Benjamini and Hochberg's testing algorithm, the subject of the next section, aims to control the expected value of a/R rather than that of a/N_0.

Exercise 4.2. Suppose that $z_1, z_2, \ldots, z_N$ are independent and identically distributed observations from the two-groups model (2.7) and that the decision rule rejects H_{0i} for $z_i \in \mathcal{Z}$, as illustrated in Figure 2.3. Show that a/R has a scaled binomial distribution given R (with $R > 0$),

$$a/R \sim \text{Bi}\left(R, \phi(\mathcal{Z})\right)/R, \tag{4.5}$$

with $\phi(\mathcal{Z}) = \text{Fdr}(\mathcal{Z})$ as in (2.13).

4.2 Benjamini and Hochberg's FDR Control Algorithm

We assume that our decision rule $\mathcal{D}$ produces a p-value p_i for each case i, so that p_i has a uniform distribution if H_{0i} is correct,

$$H_{0i} : p_i \sim \mathcal{U}(0,1). \tag{4.6}$$

Denote the ordered p-values by

$$p_{(1)} \le p_{(2)} \le \cdots \le p_{(i)} \le \cdots \le p_{(N)} \tag{4.7}$$

as in (3.19). Following the notation in Figure 4.1, let $R_{\mathcal{D}}$ be the number of cases rejected, $a_{\mathcal{D}}$ the number of those that are actually null, and $\text{Fdp}_{\mathcal{D}}$ the false discovery proportion

$$\text{Fdp}_{\mathcal{D}} = a_{\mathcal{D}}/R_{\mathcal{D}} \qquad [= 0 \text{ if } R_{\mathcal{D}} = 0]. \tag{4.8}$$

The Benjamini–Hochberg (BH) algorithm uses this rule: for a fixed value of q in $(0,1)$, let $i_{\max}$ be the *largest* index for which

$$p_{(i)} \le \frac{i}{N} q, \tag{4.9}$$

and reject $H_{0(i)}$, the null hypothesis corresponding to $p_{(i)}$, if

$$i \le i_{\max}, \tag{4.10}$$

accepting $H_{0(i)}$ otherwise.

Theorem 4.1. *If the p-values corresponding to the correct null hypotheses are independent of each other, then the rule* $\text{BH}(q)$ *based on the BH algorithm controls the expected false discovery proportion at q,*

$$E\left\{\text{Fdp}_{\text{BH}(q)}\right\} = \pi_0 q \le q \qquad \textit{where } \pi_0 = N_0/N. \tag{4.11}$$

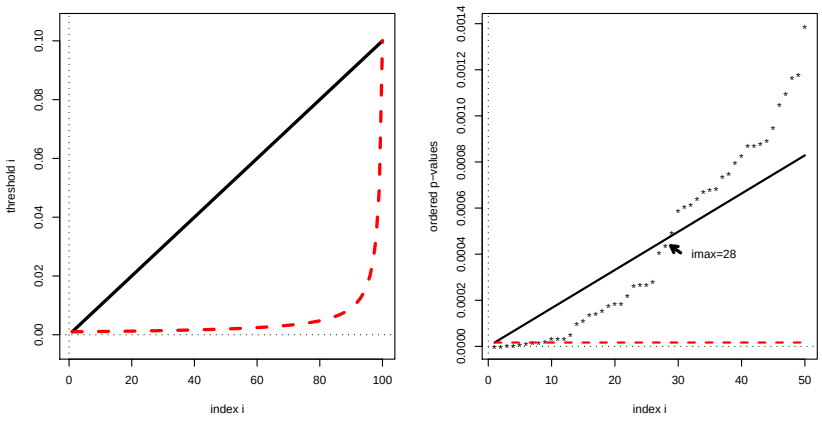

Figure 4.2: *Left panel*: Solid line is rejection boundary (4.9) for FDR control rule BH(q); dashed curve for Hochberg step-up FWER procedure, Exercise 3.8; $\alpha = q = 0.1$, $N = 100$. *Right panel*: Stars indicate p-values for the 50 largest z_i, prostate data (2.6); solid and dashed lines are rejection boundaries, BH(q) and Hochberg, $\alpha = q = 0.1$, $N = 6033$.

A proof of Theorem 4.1 appears at the end of this section. The proportion of null cases $\pi_0 = N_0/N$ is unknown in practice though we will see that it can be estimated, so q is usually quoted as the control rate of BH(q).

There is a practical reason for the impressive popularity of BH(q): it is *much more liberal* in identifying non-null cases than the FWER algorithms of Chapter 3. Figure 4.2 illustrates the point by comparison with Hochberg's step-up procedure (3.30). BH(q) can also be described in step-up form: decrease i starting from $i = N$ and keep accepting $H_{0(i)}$ until the first time $p_{(i)} \le q\, i/N$, after which all $H_{0(i)}$ are rejected. Hochberg's procedure instead uses $p_{(i)} \le \alpha/(N - i + 1)$, Exercise 3.8.

If we set $q = \alpha$, the ratio of the two thresholds is

$$\left(\frac{i}{N}\right) \Big/ \left(\frac{1}{N-i+1}\right) = i \cdot \left(1 - \frac{i-1}{N}\right). \tag{4.12}$$

Usually only small values of i/N will be interesting, in which case BH(q) is approximately i times as liberal as Hochberg's rule.

The left panel of Figure 4.2 makes the comparison for $\alpha = q = 0.1$ and $N = 100$. The two threshold curves are equal at $i = 1$ where both take the Bonferroni value α/N, and at $i = N$ where both equal α. In between, BH(q) allows rejection at much larger values of $p_{(i)}$. The right panel shows the 50 smallest $p_{(i)}$ values for the prostate data, $p_{(i)} = F_{100}(-t_{(i)})$ in (2.5), and also the two rejection boundaries, again with $\alpha = q = 0.1$. Here $i_{\max} = 28$ genes are declared non-null by BH(q)(0.1) compared to 9 for the Hochberg 0.1 test.

Of course, rejecting more cases is only a good thing if they *should* be rejected. False discovery rate control is a more liberal rejecter than FWER: can we still trust its decisions? This is the question we will be trying to answer as we consider, in what follows, the pros and cons of the BH(q) rule and its underlying rationale. Here are a few preliminary comments:

- Theorem 4.1 depends on independence among the p-values of the null cases (the top row in Figure 4.1), usually an unrealistic assumption. This limitation can be removed if the rejection boundary (4.9) is lowered to

$$p_{(i)} \le \frac{i}{N}\frac{q}{l_i} \qquad \text{where } l_i = \sum_{j=1}^{i}\frac{1}{j}. \tag{4.13}$$

  However, (4.13) represents a severe penalty ($l_{28} = 3.93$ for instance) and is not really necessary. The independence condition in Theorem 4.1 can be weakened to *positive regression dependence* (PRD): roughly speaking, the assumption that the null-case z-values have non-negative correlations, though even PRD is unlikely to hold in practice. Fortunately, the empirical Bayes interpretation of BH(q) is *not* directly affected by dependence, as discussed in the next section.

- Theorem 4.1 depends on taking $a/R = 0$ when $R = 0$, that is, defining $0/0 = 0$ in Figure 4.1. Storey's "positive false discovery rate" criterion avoids this by only considering situations with $R > 0$, but doing so makes strict FDR control impossible: if $N_0 = N$, that is, if there are no non-null cases, then all rejections are false discoveries and $E\{\text{Fdp}\,|R > 0\} = 1$ for any rule $\mathcal{D}$ that rejects anything.

- Is it really satisfactory to control an error rate *expectation* rather than an error rate *probability* as in classical significance testing? The next two sections attempt to answer this question in empirical Bayes terms.

- How should q be chosen? The literature hasn't agreed upon a conventional choice, such as $\alpha = 0.05$ for single-case testing, though $q = 0.1$ seems to be popular. The empirical Bayes context of the next section helps clarify the meaning of q.

- The p-values in (4.9) are computed on the basis of an assumed null hypothesis distribution, for example $p_{(i)} = F_{100}(-t_{(i)})$ in the right panel of Figure 4.2, with F_{100} a Student-t cdf having 100 degrees of freedom. This is by necessity in classical single-case testing, where theory is the only possible source for a null distribution. Things are different in large-scale testing: empirical evidence may make it clear that the theoretical null is unrealistic. Chapter 6 discusses the proper choice of null hypotheses in multiple testing.

This last objection applies to all testing algorithms, not just to the Benjamini–Hochberg rule. The reason for raising it here relates to Theorem 4.1: its statement is so striking and appealing that it is easy to forget its limitations. Most of these turn out to be not too important in practice, except for the proper choice of a null hypothesis, which is crucial.

Proof of Theorem 4.1. For t in $(0, 1]$ define

$$R(t) = \#\{p_i \le t\}, \tag{4.14}$$

$a(t)$ the number of null cases with $p_i \le t$, false discovery proportion

$$\text{Fdp}(t) = a(t)/\max\{R(t), 1\}, \tag{4.15}$$

and

$$Q(t) = Nt/\max\{R(t), 1\}. \tag{4.16}$$

Also let

$$t_q = \sup_t \{Q(t) \le q\}. \tag{4.17}$$

Since $R(p_{(i)}) = i$ we have $Q(p_{(i)}) = Np_{(i)}/i$. This implies that the BH rule (4.9) can be re-expressed as

$$\text{Reject } H_{0(i)} \text{ for } p_{(i)} \le t_q. \tag{4.18}$$

Let $A(t) = a(t)/t$. It is easy to see that

$$E\{A(s)|A(t)\} = A(t) \qquad \text{for } s \le t, \tag{4.19}$$

and in fact $E\{A(s)|A(t') \text{ for } t' \ge t\} = A(t)$. In other words, $A(t)$ is a martingale as t decreases from 1 to 0. Then by the optional stopping theorem,

$$E\{A(t_q)\} = E\{A(1)\} = E\{a(1)/1\} = N_0, \tag{4.20}$$

the actual number of null cases.

Finally, notice that (4.16) implies

$$\max\{R(t_q), 1\} = Nt_q/Q(t_q) = Nt_q/q, \tag{4.21}$$

so

$$\text{Fdp}(t_q) = \frac{q}{N}\frac{a(t_q)}{t_q}. \tag{4.22}$$

Then (4.20) gives

$$E\{\text{Fdp}(t_q)\} = \pi_0 q \qquad [\pi_0 = N_0/N] \tag{4.23}$$

which, together with (4.18), verifies Theorem 4.1. ■

Exercise 4.3. Verify (4.19).

4.3 Empirical Bayes Interpretation

Benjamini and Hochberg's BH(q) procedure has an appealing empirical Bayes interpretation. Suppose that the p-values p_i correspond to real-valued test statistics z_i,

$$p_i = F_0(z_i) \qquad i = 1, 2, \ldots, N, \tag{4.24}$$

where $F_0(z)$ is the cdf of a common null distribution applying to all N null hypotheses H_{0i}, for example, F_0 the standard normal cdf in (2.6). We can always take z_i to be p_i itself, in which case F_0 is the $\mathcal{U}(0,1)$ distribution.

Let $z_{(i)}$ denote the ith ordered value,

$$z_{(1)} \le z_{(2)} \le \cdots \le z_{(i)} \le \cdots \le z_{(N)}. \tag{4.25}$$

Then $p_{(i)} = F_0(z_{(i)})$ in (4.7), if we are interested in left-tailed p-values, or $p_{(i)} = 1 - F_0(z_{(i)})$ for right-tailed p-values.

Notice that the empirical cdf of the z_i values,

$$\bar{F}(z) = \#\{z_i \le z\}/N \tag{4.26}$$

satisfies

$$\bar{F}(z_{(i)}) = i/N. \tag{4.27}$$

This means we can write the threshold condition for the BH rule (4.9) as

$$F_0(z_{(i)}) \Big/ \bar{F}(z_{(i)}) \le q \tag{4.28}$$

or

$$\pi_0 F_0(z_{(i)}) \Big/ \bar{F}(z_{(i)}) \le \pi_0 q. \tag{4.29}$$

However, $\pi_0 F_0(z)/\bar{F}(z)$ is the empirical Bayes false discovery rate estimate $\overline{\mathrm{Fdr}}(z)$ (from (2.21) with $\mathcal{Z} = (-\infty, z)$).

We can now re-express Theorem 4.1 in empirical Bayes terms.

Corollary 4.2. *Let $i_{\max}$ be the largest index for which*

$$\overline{\mathrm{Fdr}}(z_{(i)}) \le q \tag{4.30}$$

and reject $H_{0(i)}$ for all $i \le i_{\max}$, accepting $H_{0(i)}$ otherwise. Then, assuming that the z_i values are independent, the expected false discovery proportion of the rule equals q.

Exercise 4.4. Use Theorem 4.1 to verify Corollary 4.2.

Note. With π_0 unknown it is usual to set it to its upper bound 1, giving $\overline{\mathrm{Fdr}}(z_{(i)}) = F_0(z_{(i)})/\bar{F}(z_{(i)})$. This makes rule (4.30) conservative, with $E\{\mathrm{Fdp}\} \le q$. But see Section 4.5.

Returning to the two-groups model (2.7), Bayes rule gives

$$\mathrm{Fdr}(z) = \pi_0 F_0(z)/F(z) \tag{4.31}$$

as the posterior probability that case i is null given $z_i \le z$ (2.13). Section 2.4 shows $\overline{\mathrm{Fdr}}(z_i)$ to be a good estimate of $\mathrm{Fdr}(z_i)$ under reasonable conditions. A greedy empirical Bayesian might select

$$z_{\max} = \sup_z \left\{\overline{\mathrm{Fdr}}(z) \le q\right\} \tag{4.32}$$

and report those cases having $z_i \le z_{\max}$ as "having estimated probability q of being null." Corollary 4.2 justifies the greedy algorithm in frequentist terms: if the z_i values

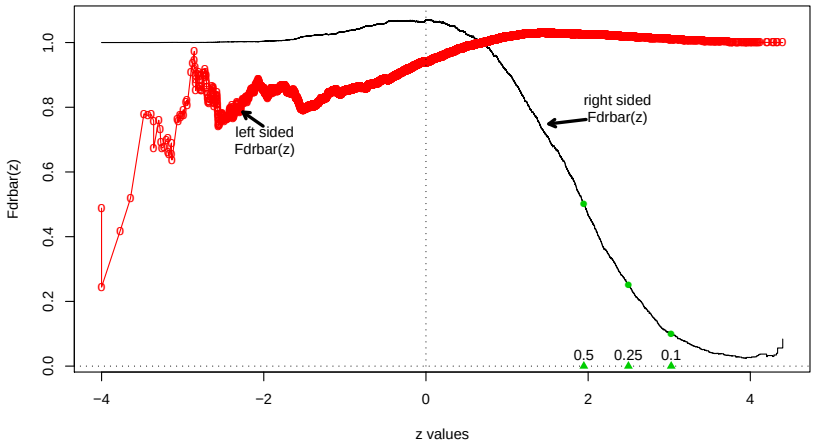

Figure 4.3: Left-sided and right-sided values of $\overline{\text{Fdr}}(z)$ for the DTI data of Section 2.5; triangles indicate values of z having $\overline{\text{Fdr}}^{(c)}(z)$ (4.33) equal 0.5, 0.25 and 0.1; 192 voxels have $z_{(i)}$ exceeding 3.02, the $q = 0.1$ threshold.

are independent, then the expected null proportion of the reported cases will equal q. It is always a good sign when a statistical procedure enjoys both frequentist and Bayesian support, and the BH algorithm passes the test.

Figure 4.3 graphs $\overline{\text{Fdr}}(z)$ and the analogous right-sided quantity

$$\overline{\text{Fdr}}^{(c)}(z) = \pi_0 F_0^{(c)}(z) \Big/ \bar{F}^{(c)}(z) \tag{4.33}$$

for the DTI data of Section 2.5 (setting π_0 to 1 in (4.29) and (4.33)), where $F^{(c)}(z)$ indicates the complementary cdf $1 - F(z)$. There is just the barest hint of anything interesting on the left, but on the right, $\text{Fdr}^{(c)}(z)$ gets quite small. For example, 192 of the voxels reject their null hypotheses, those having $z_i \geq 3.02$ at the $q = 0.1$ threshold.

Exercise 4.5. I set $\pi_0 = 1$ in (4.33). How does that show up in Figure 4.3?

The empirical Bayes viewpoint clarifies some of the questions raised in the previous section.

- *Choice of q* Now q is an estimate of the Bayes probability that a rejected null hypothesis H_{0i} is actually correct. It is easy to explain to a research colleague that $q = 0.1$ means an estimated 90% of the rejected cases are true discoveries. The uncomfortable moment in single-case testing, where it has to be confessed that $\alpha = 0.05$ rejection does *not* imply a 95% chance that the effect is genuine, is happily avoided.

- *Independence assumption* $\overline{\text{Fdr}}(z) = \pi_0 F_0(z)/\bar{F}(z)$ is an accurate estimate of the Bayes false discovery rate $\text{Fdr}(z) = \pi_0 F_0(z)/F(z)$ (2.13) whenever $\bar{F}(z)$, the empirical cdf, is close to $F(z)$. This does *not* require independence of the z-values, as shown in Section 2.4. $\overline{\text{Fdr}}(z)$ is upwardly biased for estimating $\text{Fdr}(z)$, and also for estimating the expected false discovery proportion, and in this sense it is always conservative. Lemma

2.2 shows that the upward bias is small under reasonable conditions. Roughly speaking, $\overline{\text{Fdr}}(z)$ serves as an unbiased estimate of $\text{Fdr}(z)$, and of FDR$= E\{\text{Fdp}\}$, even if the z_i are correlated.

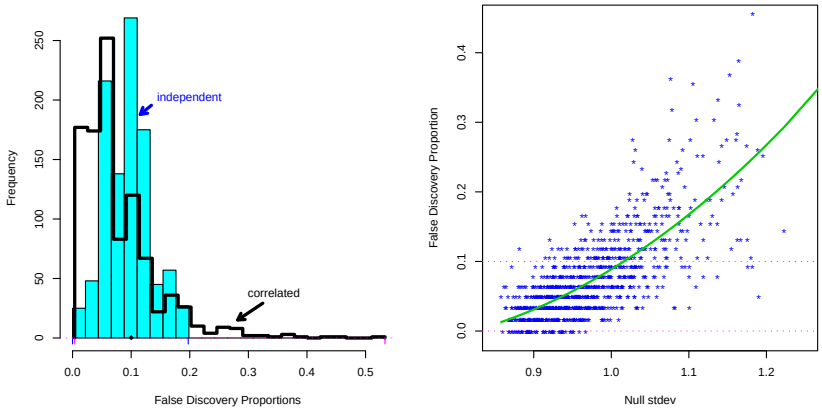

Figure 4.4: *Left panel*: Solid histogram Fdp for BH rule, $q = 0.1$, 1000 simulations of model (4.34) with z_i values independent; line histogram for z_i values correlated, root mean square correlation $= 0.1$. *Right panel*: Fdp values for correlated simulations plotted against $\hat{\sigma}_0$, the empirical standard deviation of the null z_i values; smooth curve is quadratic regression.

The price for correlation is paid in the *variability* of $\overline{\text{Fdr}}(z)$ as an estimate of $\text{Fdr}(z)$, as illustrated in Figure 4.4. Our simulation model involved $N = 3000$ z-values, with

$$\begin{aligned} & z_i \sim \mathcal{N}(0,1), && i = 1, 2, \ldots, 2850 \\ \text{and} \quad & z_i \sim \mathcal{N}(2.5,1), && i = 2851, \ldots, 3000 \end{aligned} \tag{4.34}$$

so $\pi_0 = 0.95$. Two runs of 1000 simulations each were made, the first with the z_i values independent and the second with substantial correlation: the root mean square value of all $3000 \cdot 2999/2$ pairs of correlations equaled 0.1.[2] For each simulation, the rule BH(q), $q = 0.1$, was applied (right-sided) and the actual false discovery proportion Fdp observed.

The left panel of Figure 4.4 compares a histogram of the 1000 Fdp values under independence with that for the correlated simulations. The expected Fdp is controlled below $q = 0.1$ in both cases, averaging 0.095 under independence and 0.075 under correlation (see Table 4.1). Control is achieved in quite different ways, though: correlation produces a strongly asymmetric Fdp distribution, with more very small or very large values. The BH algorithm continues to control the expectation of Fdp under correlation, but $\overline{\text{Fdr}}$ becomes a less accurate estimator of the true Fdr.

In Figure 4.4's right panel, the Fdp values for the 1000 correlated simulations are plotted versus $\hat{\sigma}_0$, the empirical standard deviation of the 2850 null z-values. Correlation greatly increases the variability of $\hat{\sigma}_0$, as discussed in Chapter 7. Fdp tends to be greater or less than the nominal value 0.1 as $\hat{\sigma}_0$ is greater or less than 1.0, *varying by a factor of 10.*

[2]The correlation structure is described in Section 8.2.

Table 4.1: Means and standard deviations (in parentheses) for the simulation experiments of Figure 4.4.

	Fdp	$\hat{\sigma}_0$	# rejected
Uncorrelated:	**.095** (.038)	1.00 (.014)	64.7 (3.7)
Correlated:	**.075** (.064)	.97 (.068)	63.1 (7.5)

In practice, $\hat{\sigma}_0$ isn't observable. However, it is "almost observable" in some situations, in which case the overall control level q can be misleading: if we know that $\hat{\sigma}_0$ is much greater than 1 then there is good reason to believe that Fdp is greater than q. This point in investigated in Chapter 6 in terms of the *empirical null distribution*.

• *FDR control as a decision criterion* The BH algorithm only controls the expectation of Fdp. Is this really sufficient for making trustworthy decisions? Part of the answer must depend upon the accuracy of $\overline{\text{Fdr}}$ as an estimate of Fdr (4.31) or of FDR $= E\{\text{Fdp}\}$. This same question arises in single-case testing where the concept of *power* is used to complement Type I error control. Chapters 5 and 7 discuss accuracy and power considerations for false discovery rate control methods.

• *Left-sided, right-sided, and two-sided inferences* For the DTI data of Figure 4.3, BH$(q)(0.1)$ rejects 0 voxels on the left and 192 voxels on the right. However, if we use two-sided p-values, $p_i = 2 \cdot \Phi(-|z_i|)$, only 110 voxels are rejected by BH$(q)(0.1)$, all from among the 192. From a Bayesian point of view, two-sided testing only blurs the issue by making posterior inferences over larger, less precise rejection regions $\mathcal{Z}$. The *local* false discovery rate (2.14) provides the preferred Bayesian inference. Chapter 5 concerns estimation of the local fdr.

Exercise 4.6. For the prostate data, the left-tailed, right-tailed, and two-tailed BH(q) rules reject 32, 28, and 60 genes at the $q = 0.1$ level. The rejection regions are $z_i \leq -3.26$ on the left, $z_i \geq 3.36$ on the right, and $|z_i| \geq 3.29$ two-sided. Why is two-sided testing less wasteful here than in the DTI example?

• *False negative rates* Looking at Figure 4.1, it seems important to consider the false negative proportion

$$\text{Fnp} = (N_1 - b)/(N - R) \tag{4.35}$$

as well as Fdp. The expectation of Fnp is a measure of Type II error for $\mathcal{D}$, indicating the rule's power. It turns out that the Bayes/empirical Bayes interpretation of the false discovery rate applies to both Fdp and Fnp.

Suppose that rule $\mathcal{D}$ rejects H_{0i} for $z_i \in \mathcal{R}$, and accepts H_{0i} for z_i in the complementary region $\mathcal{A}$. Following notation (2.13),

$$1 - \phi(\mathcal{A}) = \Pr\{\text{non-null} | z \in \mathcal{A}\} \tag{4.36}$$

is the Bayes posterior probability of a Type II error. The calculations in Section 2.3 and Section 2.4 apply just as well to $\overline{\text{Fdr}}(\mathcal{A})$ as $\overline{\text{Fdr}}(\mathcal{R})$: under the conditions stated there, for instance in Lemma 2.2, $1 - \overline{\text{Fdr}}(\mathcal{A})$ will accurately estimate $1 - \phi(\mathcal{A})$, the Bayesian false negative rate. Chapter 5 uses this approach to estimate power in multiple testing situations.

4.4 Is Fdr Control "Hypothesis Testing"?

The Benjamini–Hochberg BH(q) rule is usually presented as a multiple hypothesis testing procedure. This was our point of view in Section 4.2, but not in Section 4.3, where the *estimation* properties of $\overline{\text{Fdr}}$ were emphasized. It pays to ask in what sense false discovery rate control is actually hypothesis testing.

Here we will fix attention on a given subset $\mathcal{R}$ of the real line, e.g., $\mathcal{R} = [3, \infty)$. We compute

$$\overline{\text{Fdr}}(\mathcal{R}) = e_0(\mathcal{R})/R, \tag{4.37}$$

where $e_0(\mathcal{R})$ is the expected number of null cases falling in $\mathcal{R}$ and R is the observed number of z_i in $\mathcal{R}$. We might then follow the rule of rejecting all the null hypotheses H_{0i} corresponding to z_i in $\mathcal{R}$ if $\overline{\text{Fdr}}(\mathcal{R}) \le q$, and accepting all of them otherwise. Equivalently, we reject all the H_{0i} for z_i in $\mathcal{R}$ if

$$R \ge e_0(\mathcal{R})/q. \tag{4.38}$$

It is clear that (4.38) cannot be a test of the FWER-type null hypothesis that at least one of the R hypotheses is true,

$$H_0(\text{union}) = \bigcup_{i:z_i \in \mathcal{R}} H_{0i}. \tag{4.39}$$

Rejecting H_0(union) implies we believe *all* the H_{0i} for z_i in $\mathcal{R}$ to be incorrect (that is, all should be rejected). But if, say, $R = 50$ then $\overline{\text{Fdr}}(\mathcal{R}) = 0.1$ suggests that about 5 of the R H_{0i} are correct.

Exercise 4.7. Calculate the probability that H_0(union) is correct if $R = 50$ and $\phi(\mathcal{R}) = 0.1$, under the assumptions of Lemma 2.2.

In other words, the $\overline{\text{Fdr}}$ rule (4.38) is too *liberal* to serve as a test of H_0(union). Conversely, it is too *conservative* as a test of

$$H_0(\text{intersection}) = \bigcap_{i:z_i \in \mathcal{R}} H_{0i} \tag{4.40}$$

which is the hypothesis that all of the R null hypotheses are correct. Rejecting H_0(intersection) says we believe at least one of the R cases is non-null.

Under the Poisson-independence assumptions of Lemma 2.2, H_0(intersection) implies

$$R \sim \text{Poi}\left(e_0(\mathcal{R})\right). \tag{4.41}$$

The obvious level-α test[3] rejects H_0(intersection) for

$$R \geq Q_\alpha, \tag{4.42}$$

the upper $1-\alpha$ quantile of a Poi($e_0(\mathcal{R})$) variate. A two-term Cornish–Fisher expansion gives the approximation

$$Q_\alpha = e_0(\mathcal{R}) + \sqrt{e_0(\mathcal{R})}z_\alpha + \left(z_\alpha^2 - 1\right)\Big/ 6 \tag{4.43}$$

with z_α the standard normal quantile $\Phi^{-1}(1-\alpha)$. (Increasing (4.43) to the nearest integer makes test (4.42) conservative.) Table 4.2 compares the minimum rejection values of R from (4.38) and (4.42) for $q = \alpha = 0.1$. It is clear that (4.38) is far more conservative.

The inference of $\overline{\text{Fdr}}$ outcome (4.32) lies somewhere between "all R cases are true discoveries" and "at least one of the R is a true discovery." I prefer to think of $\overline{\text{Fdr}}$ as an estimate rather than a test statistic: a quantitative assessment of the proportion of false discoveries among the R candidates.

Table 4.2: Rejection thresholds for R, Fdr test (4.38) and H_0(intersection) test (4.42); $q = \alpha = 0.1$. As a function of $e_0(\mathcal{R})$, the expected number of null cases in R. (Rounding H_0(intersection) upward gives conservative level-α tests.)

$e_0(\mathcal{R})$:	1	2	3	4	6	8
H_0(intersection):	2.39	3.92	5.33	6.67	9.25	11.73
Fdr:	10	20	30	40	60	80

Exercise 4.8. For the DTI data of Figure 4.3, 26 of the 15443 z-values are less than -3.0. How strong is the evidence that at least one of the 26 is non-null? (Assume independence and set π_0 to its upper bound 1.)

4.5 Variations on the Benjamini–Hochberg Algorithm

The BH algorithm has inspired a great deal of research and development in the statistics literature, including some useful variations on its original form. Here we will review just two of these.

• *Estimation of* π_0 The estimated false discovery rate $\overline{\text{Fdr}}(z) = \pi_0 F_0(z)\ /\bar{F}(z)$ appearing in Corollary 4.2 requires knowing π_0, the actual proportion of null cases. Rather than setting π_0 equal to its upper bound 1 as in the original BH procedure, we can attempt to estimate it from the collection of observed z-values.

[3] This test is a form of Tukey's "higher criticism."

Returning to the two-groups model (2.7), suppose we believe that $f_1(z)$ is zero for a certain subset $\mathcal{A}_0$ of the sample space, perhaps those points near zero,

$$f_1(z) = 0 \qquad \text{for } z \in \mathcal{A}_0; \tag{4.44}$$

that is, all the non-null cases must give z-values outside of $\mathcal{A}_0$ (sometimes called the *zero assumption*). Then the expected value of $N_+(\mathcal{A}_0)$, the observed number of z_i values in $\mathcal{A}_0$, is

$$E\{N_+(\mathcal{A}_0)\} = \pi_0 N \cdot F_0(\mathcal{A}_0), \tag{4.45}$$

suggesting the estimators

$$\hat{\pi}_0 = N_+(\mathcal{A}_0)/\left(N \cdot F_0(\mathcal{A}_0)\right) \tag{4.46}$$

and

$$\widehat{\mathrm{Fdr}}(z) = \hat{\pi}_0 F_0(z) \Big/ \bar{F}(z). \tag{4.47}$$

Using $\widehat{\mathrm{Fdr}}(z)$ in place of $\overline{\mathrm{Fdr}}(z) = F_0(z)/\bar{F}(z)$ in Corollary 4.2 increases the number of discoveries (i.e., rejections). It can be shown that the resulting rule still satisfies $E\{\mathrm{Fdp}\} \le q$ under the independence assumption even if (4.44) isn't valid.

We might take $\mathcal{A}_0$ to be the central α_0 proportion of the f_0 distribution on the grounds that all the "interesting" non-null cases should produce z-values far from the central region of f_0. If f_0 is $\mathcal{N}(0,1)$ in (2.7) then $\mathcal{A}_0$ is the interval

$$\mathcal{A}_0 = \left[\Phi^{-1}\left(0.5 - \alpha_0/2\right), \Phi^{-1}\left(0.5 + \alpha_0/2\right)\right] \tag{4.48}$$

with Φ the standard normal cdf. Figure 4.5 graphs $\hat{\pi}_0$ as a function of α_0 for the prostate and DTI data.

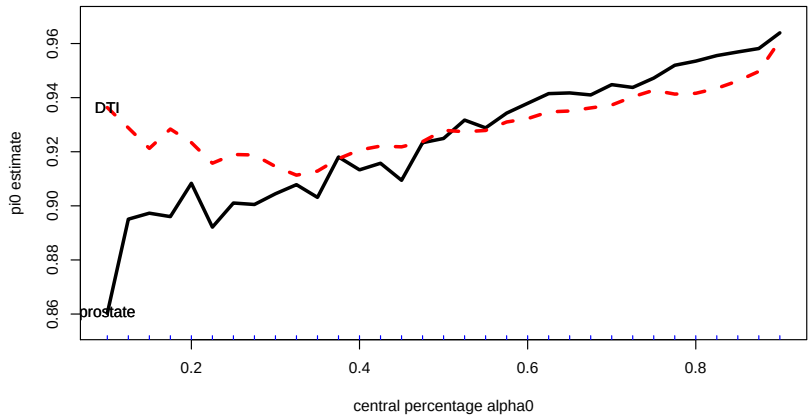

Figure 4.5: Estimated values of π_0 (4.46) for the prostate and DTI data sets; $\mathcal{A}_0$ as in (4.48); α_0 ranging from 0.1 to 0.9.

Nothing in Figure 4.5 suggests an easy way to select the appropriate $\mathcal{A}_0$ region, particularly not for the prostate data. Part of the problem is the assumption that

$f_0(z)$ is the theoretical null $\mathcal{N}(0,1)$ density. The central peak of the prostate data seen in Figure 2.1 is slightly wider, about $\mathcal{N}(0, 1.06^2)$, which affects calculation (4.46). Chapter 6 discusses methods that estimate π_0 in conjunction with estimation of the mean and variance of f_0.

Exercise 4.9. How would the prostate data $\hat{\pi}_0$ values in Figure 4.5 change if we took f_0 to be $\mathcal{N}(0, 1.06^2)$?

The exact choice of $\hat{\pi}_0$ in (4.47) is not crucial: if we are interested in values of Fdr near $q = 0.1$ then the difference between $\hat{\pi}_0 = 0.9$ and $\hat{\pi}_0 = 1$ is quite small. A much more crucial and difficult issue is the appropriate choice of the null density f_0, the subject of Chapter 6.

• *Significance analysis of microarrays* SAM, the significance analysis of microarrays, is a popular Fdr-like program originally intended to identify interesting genes in microarray experiments. Microarray studies have nothing in particular to do with SAM's workings, but we will use them for illustration here. Suppose that X is an $N \times n$ matrix of expression levels as in Section 2.1 that we have used to produce an N-vector of z-values $\boldsymbol{z} = (z_1, z_2, \ldots, z_N)'$. For the sake of definiteness, assume that X represents a two-sample study, say healthy versus sick subjects, and that the z_i are normalized t-values as in (2.2)–(2.5). (SAM actually handles more general experimental layouts and summary statistics: in particular, there need be no theoretical null assumption (2.6).)

1. The algorithm begins by constructing some number B of $N \times n$ matrices X^*, each of which is a version of X in which the columns have been randomly permuted as in (3.40). Each X^* yields an N-vector $\boldsymbol{z}^*$ of z-values calculated in the same way as $\boldsymbol{z}$.

2. Let $\boldsymbol{Z}$ be the ordered version of $\boldsymbol{z}$, and likewise $\boldsymbol{Z}^{*b}$ the ordered version of $\boldsymbol{z}^{*b}$, the bth $\boldsymbol{z}^*$ vector, $b = 1, 2, \ldots, B$. Define

$$\bar{Z}_i = \sum_{b=1}^{B} Z_i^{*b} \Big/ B \tag{4.49}$$

   so $\bar{Z}_i$ is the average of the ith largest values of $\boldsymbol{z}^{*b}$.

3. Plot Z_i versus $\bar{Z}_i$ for $i = 1, 2, \ldots, N$. The upper panel of Figure 4.6 shows the $(\bar{Z}_i, Z_i)$ plot for the prostate data of Figure 2.1. (This amounts to a QQ-plot of the actual z-values versus the permutation distribution.)

4. For a given choice of a positive constant Δ, define

$$\begin{aligned} c_{\text{up}}(\Delta) &= \min\{Z_i : Z_i - \bar{Z}_i \geq \Delta\} \\ \text{and} \quad c_{\text{lo}}(\Delta) &= \max\{Z_i : \bar{Z}_i - Z_i \geq \Delta\}. \end{aligned} \tag{4.50}$$

   In words, $c_{\text{up}}(\Delta)$ is the first Z_i value at which the $(\bar{Z}_i, Z_i)$ curve exits the band $\bar{Z}_i + \Delta$, and similarly for $c_{\text{lo}}(\Delta)$. In the top panel of Figure 4.6, $\Delta = 0.7$, $c_{\text{up}}(\Delta) = 3.29$, and $c_{\text{lo}}(\Delta) = -3.34$.

5. Let $R(\Delta)$ be the number of z_i values outside of $[c_{\rm lo}(\Delta), c_{\rm up}(\Delta)]$,

$$R(\Delta) = \#\{z_i \geq c_{\rm up}(\Delta)\} + \#\{z_i \leq c_{\rm lo}(\Delta)\}, \tag{4.51}$$

and likewise

$$R^*(\Delta) = \#\left\{z_i^{*b} \geq c_{\rm up}(\Delta)\right\} + \#\left\{z_i^{*b} \leq c_{\rm lo}(\Delta)\right\}, \tag{4.52}$$

the sums in (4.52) being over all $N \cdot B$ permutation z-values.

6. Finally, define the false discovery rate corresponding to Δ as

$$\overline{\rm Fdr}(\Delta) = \frac{R^*(\Delta)/NB}{R(\Delta)/N} = \frac{1}{B}\frac{R^*(\Delta)}{R(\Delta)}. \tag{4.53}$$

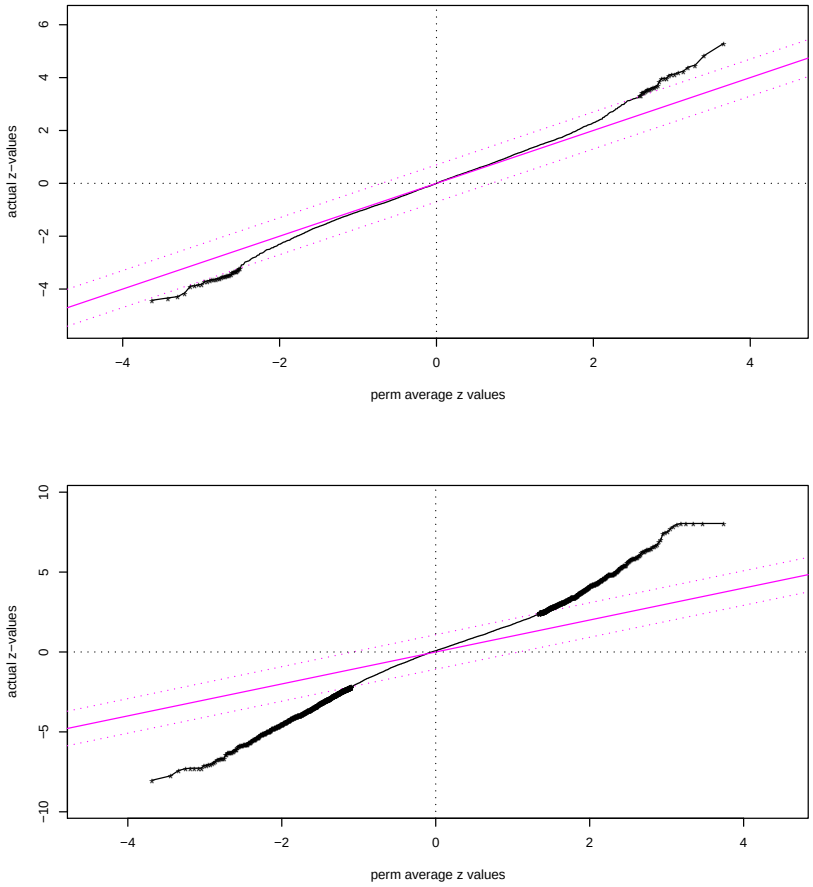

Figure 4.6: SAM plots for the prostate data (top) and the leukemia data (bottom). Starred points indicate "significant" genes at the $q = 0.1$ level: 60 in the top panel, 1660 in the bottom.

The SAM program calculates $\overline{\rm Fdr}(\Delta)$ for a range of Δ values in a search for the Δ that produces a pre-chosen value $\overline{\rm Fdr}(\Delta) = q$. For the prostate data, $\Delta = 0.7$ gave

$\text{Fdr}(\Delta) = 0.1$. $R(\Delta) = 60$ genes were identified as significant, 28 on the right and 32 on the left.

Definition (4.53) of $\overline{\text{Fdr}}(\Delta)$ is equivalent to our previous usage at (4.28) or (2.21): the rejection region

$$\mathcal{R}(\Delta) = \{z \notin [c_{\text{lo}}(\Delta), c_{\text{up}}(\Delta)]\} \tag{4.54}$$

has empirical probability $\bar{F}(\Delta) = R(\Delta)/N$; similarly, $R^*(\Delta)/NB$ is the null estimate $\bar{F}_0(\Delta)$, the proportion of the $N \cdot B$ z_i^{*b}-values in $\mathcal{R}(\Delta)$, so

$$\overline{\text{Fdr}}(\Delta) = \bar{F}_0(\Delta) \Big/ \bar{F}(\Delta), \tag{4.55}$$

which is (4.28), setting $\pi_0 = 1$ and using the permutation z_i^{*b}-values instead of a theoretical distribution to define the nulls.

Despite the disparaged "significance" terminology, the output of SAM is closer to empirical Bayes estimation than hypothesis testing; that is, the statistician gets more than a simple yes/no decision for each gene. The two-sided nature of the procedure is unfortunate from a Bayesian perspective, but this can be remedied by choosing Δ separately for positive and negative z-values.

The bottom panel of Figure 4.6 concerns the *leukemia data*, another microarray study featured in Chapter 6. Here there are $N = 7128$ genes whose expression levels are measured on $n = 72$ patients, 47 with a less severe and 25 with a more severe form of leukemia. Two-sample t-tests have led to z-values as in (2.1)–(2.5) (now with 70 degrees of freedom rather than 100). The SAM plot reveals a serious problem: unlike the prostate panel, the leukemia plot doesn't match the solid 45° line near $z = 0$, crossing it instead at a sharp angle.

We will see in Chapter 6 that the histogram of the 7128 leukemia z-values, unlike Figure 2.1, is much wider at the center than a $\mathcal{N}(0, 1)$ distribution. However, the permutation null distributions are almost perfectly $\mathcal{N}(0, 1)$ in both cases, a dependable phenomenon it turns out. This casts doubt on the appropriateness of $\bar{F}_0(\Delta)$ in the numerator of $\overline{\text{Fdr}}(\Delta)$ (4.55) and the identification of 1660 "significant" leukemia genes. The appropriate choice of a null distribution is the crucial question investigated in Chapter 6.

Exercise 4.10. Suppose the z-value histogram is approximately $\mathcal{N}(0, \sigma_0^2)$ near $z = 0$ while the permutation distribution is $\mathcal{N}(0, 1)$. What will be the angle of crossing of the SAM plot?

4.6 $\overline{\text{Fdr}}$ and Simultaneous Tests of Correlation

When dealing with t-statistics, as in the prostate study (2.2), the false discovery rate estimator $\overline{\text{Fdr}}$ (2.21) has a nice geometrical interpretation in terms of clustering on the hypersphere. This interpretation allows us to use the BH algorithm to answer a different kind of question: Given a case of interest, say gene 610 in the prostate study, which of the other $N-1$ cases is unusually highly correlated with it? "Unusual" has the meaning here of being in the rejection set of a simultaneous testing procedure.

It is easier to describe the main idea in terms of one-sample rather than two-sample t-tests. Suppose that X is an $N \times n$ matrix with entries x_{ij}. For each row $\boldsymbol{x}_i$ of X we compute t_i, the one-sample t-statistic,

$$t_i = \frac{\bar{x}_i}{\hat{\sigma}_i/\sqrt{n}} \qquad \left[\bar{x}_i = \frac{\sum_1^n x_{ij}}{n},\ \hat{\sigma}_i^2 = \frac{\sum_1^n (x_{ij} - \bar{x}_i)^2}{n-1}\right]. \tag{4.56}$$

We wish to test which if any of the N t_i-values are unusually large. (X might arise in a paired comparison microarray study where x_{ij} is the difference in expression levels, Treatment minus Placebo, for gene i in the jth pair of subjects.)

Let

$$\boldsymbol{u} = (1, 1, \ldots, 1)' \Big/ \sqrt{n} \tag{4.57}$$

be the unit vector lying along the direction of the main diagonal in n-dimensional space. The angle θ_i between $\boldsymbol{u}$ and

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})' \tag{4.58}$$

has cosine

$$\cos(\theta_i) = \tilde{\boldsymbol{x}}_i' \boldsymbol{u} \tag{4.59}$$

where

$$\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i / \|\boldsymbol{x}_i\| = \boldsymbol{x}_i \Big/ \left(\sum_{j=1}^n x_{ij}^2\right)^{1/2} \tag{4.60}$$

is the scale multiple of $\boldsymbol{x}_i$ having unit length. A little bit of algebra shows that t_i is a monotonically decreasing function of θ_i,

$$t_i = \sqrt{n-1}\cos(\theta_i) \Big/ \left[1 - \cos(\theta_i)^2\right]^{1/2}. \tag{4.61}$$

Exercise 4.11. Verify (4.61).

The unit sphere in n dimensions,

$$\mathcal{S}_n = \left\{\boldsymbol{v} : \sum_{i=1}^n v_i^2 = 1\right\} \tag{4.62}$$

can be shown to have *area* (i.e., $(n-1)$-dimensional Lebesgue measure)

$$A_n = 2\pi^{n/2}/\Gamma(n/2). \tag{4.63}$$

With $n = 3$ this gives the familiar result $A_3 = 4\pi$. Under the null hypothesis,

$$H_{0i} : x_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_0^2) \qquad j = 1, 2, \ldots, n, \tag{4.64}$$

the vector $\boldsymbol{x}_i$ is known to have spherical symmetry around the origin $\mathbf{0}$; that is, $\tilde{\boldsymbol{x}}_i$ is randomly distributed over $\mathcal{S}_n$, with its probability of falling into any subset $\mathcal{R}$ on $\mathcal{S}_n$ being proportional to the $(n-1)$-dimensional area $A(\mathcal{R})$ of $\mathcal{R}$.

Putting this together, we can calculate[4] p_i, the one-sided p-value of the one-sample t-test for H_{0i}, in terms of θ_i:

$$p_i = A(\mathcal{R}_{\theta_i})/A_N \equiv \tilde{A}(\theta_i). \tag{4.65}$$

Here $\mathcal{R}_\theta$ indicates a spherical cap of angle θ on $\mathcal{S}_n$ centered at $\boldsymbol{u}$, while $\tilde{A}(\theta_i)$ is the cap's area relative to the whole sphere. (Taking $\boldsymbol{u}$ as the north pole on a globe of the Earth, $\mathcal{R}_\theta$ with $\theta = 23.5°$ is the region north of the arctic circle.)

Small values of θ_i correspond to small p-values p_i. If $\theta(n,\alpha)$ defines a cap having relative area α, perhaps $\alpha = 0.05$, then the usual α-level t-test rejects H_{0i} for $\theta_i \le \theta(n,\alpha)$. Intuitively, under the alternative hypothesis $x_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_0^2)$ for $j = 1, 2, \ldots, n$, $\tilde{\boldsymbol{x}}_i$ will tend to fall nearer $\boldsymbol{u}$ if $\mu_i > 0$, rejecting H_{0i} with probability greater than α.

Exercise 4.12. Calculate $\theta(n, 0.05)$ for $n = 5, 10, 20$, and 40. *Hint*: Work backwards from (4.61), using a table of critical values for the t-test.

Getting back to the simultaneous inference problem, we observe N points $\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \ldots, \tilde{\boldsymbol{x}}_N$ on $\mathcal{S}_n$ and wonder which of them, if any, lie unusually close to $\boldsymbol{u}$. We can rephrase the Benjamini–Hochberg procedure BH(q) to provide an answer. Define $\overline{\text{Fdr}}(\theta)$ to be $\overline{\text{Fdr}}(\mathcal{Z})$ in (2.21) with $\mathcal{Z} = \mathcal{R}_\theta$ and let $N_+(\theta)$ denote the number of points $\tilde{\boldsymbol{x}}_i$ in $\mathcal{R}_\theta$. Then

$$\overline{\text{Fdr}}(\theta) = N\pi_0\tilde{A}(\theta)\Big/N_+(\theta) \tag{4.66}$$

as in (2.24).

Corollary 4.2 now takes this form: ordering the θ_i values from smallest to largest,

$$\theta_{(1)} \le \theta_{(2)} \le \cdots \le \theta_{(i)} \le \cdots \le \theta_{(N)}, \tag{4.67}$$

let $i_{\max}$ be the largest index for which

$$\frac{\tilde{A}\left(\theta_{(i)}\right)}{N_+\left(\theta_{(i)}\right)/N} \le q \tag{4.68}$$

and reject $H_{0(i)}$ for $i \le i_{\max}$. Assuming independence of the points, the expected proportion of null cases among the $i_{\max}$ rejectees will be less than q (actually equaling $\pi_0 q$). The empirical Bayes considerations of Section 4.3 suggest the same bound, even under dependence.

Let

$$\hat{\theta}(q) = \theta_{(i_{\max})}, \quad R(q) = N_+\left(\hat{\theta}(q)\right), \quad \text{and} \quad \mathcal{R}(q) = \mathcal{R}\left(\hat{\theta}(q)\right). \tag{4.69}$$

The BH algorithm BH(q) rejects the $R(q)$ cases having $\theta_i \le \hat{\theta}(q)$, that is, those having $\tilde{\boldsymbol{x}}_i$ within the spherical cap $\mathcal{R}(q)$, as illustrated in Figure 4.7.

Exercise 4.13. Show that $R(q)/N$, the observed proportion of points in $\mathcal{R}(q)$, is at least $1/q$ times the relative area $\tilde{A}(\hat{\theta}(q))$. (So if $q = 0.1$ there are at least 10 times as many points in $\mathcal{R}(q)$ as there would be if all the null hypotheses were correct.)

[4] We are following Fisher's original derivation of the Student t-distribution.

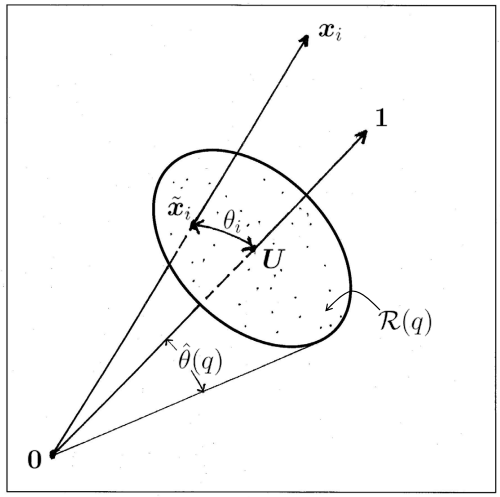

Figure 4.7: Spherical cap rejection region $\mathcal{R}(q)$ for BH procedure BH(q); H_{0i} is rejected since $\theta_i \le \hat{\theta}(q)$. Dots indicate the other rejected cases. The number of points in $\mathcal{R}(q)$ is at least $1/q$ times larger than the expected number if all N null hypotheses were correct.

The same procedure can be used for the simultaneous testing of correlations. Suppose we are interested in a particular case i_0 and wonder which if any of the other $N-1$ cases are unusually highly correlated with case i_0. Define $\boldsymbol{x}_{ij}^\dagger$ to be the centered version of $\boldsymbol{x}_i$,

$$\boldsymbol{x}_{ij}^\dagger = x_{ij} - \bar{x}_i \qquad j = 1, 2, \ldots, n, \tag{4.70}$$

and let $\boldsymbol{x}_{i_0}^\dagger$ play the role of $\mathbf{1}$ in Figure 4.7. Then

$$\begin{aligned} \cos(\theta_i) &= \boldsymbol{x}_i^{\dagger\prime}\boldsymbol{x}_{i_0}^\dagger \Big/ \left[\|\boldsymbol{x}_i^\dagger\| \cdot \|\boldsymbol{x}_{i_0}^\dagger\|\right] \\ &= \widehat{\text{cor}}(i, i_0), \end{aligned} \tag{4.71}$$

the Pearson sample correlation coefficient between $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i_0}$.

Following through definitions (4.67)–(4.68) gives a BH(q) simultaneous test for the $N-1$ null hypotheses

$$H_{0i} : \text{cor}(i, i_0) = 0, \qquad i \ne i_0. \tag{4.72}$$

Thinking of the vectors $\boldsymbol{x}_i^\dagger/\|\boldsymbol{x}_i^\dagger\|$ as points on the $(n-1)$-dimensional sphere

$$\mathcal{S}_{n-1}^\dagger = \left\{ \boldsymbol{v} : \sum_1^n v_i = 0,\ \sum_1^n v_i^2 = 1 \right\}, \tag{4.73}$$

the test amounts to checking for high-density clusters near $\boldsymbol{x}_{i_0}^\dagger/\|\boldsymbol{x}_{i_0}^\dagger\|$. Different choices of i_0 let us check for clusters all over $\mathcal{S}_{n-1}^\dagger$, i.e., for groups of correlated cases. (*Note*: The test can be carried out conveniently by first computing

$$t_i = \sqrt{\nu}\,\widehat{\text{cor}}(i, i_0) \Big/ \left[1 - \widehat{\text{cor}}(i, i_0)^2\right]^{1/2} \quad \text{and} \quad p_i = 1 - F_\nu(t_i) \tag{4.74}$$

with $\nu = n - 2$, the degrees of freedom, and F_ν the Student-t cdf, and then applying BH(q) to the p_i values. Using $p_i = F_\nu(t_i)$ checks for large *negative* correlations.)

Correlation testing was applied to the prostate data of Section 2.1. Definition (4.70) was now modified to subtract either the control or cancer patient mean for gene i, as appropriate, with $\nu = n - 3 = 99$ in (4.74). Gene 610 had the largest z-value (2.5) among the $N = 6033$ genes, $z_{610} = 5.29$, with estimated $\overline{\text{Fdr}}^{(c)}(z_{610}) = 0.0007$ (using (4.33) with $\pi_0 = 1$). Taking $i_0 = 610$ in tests (4.72) produced only gene 583 as highly correlated at level $q = 0.10$; taking $i_0 = 583$ gave genes 637 and 610, in that order, as highly correlated neighbors; $i_0 = 637$ gave 14 near neighbors, etc. Among the cases listed in Table 4.3 only genes 610 and 637 had $\overline{\text{Fdr}}^{(c)}(z_i) \le 0.50$. One might speculate that gene 637, which has low $\overline{\text{Fdr}}^{(c)}$ and a large number of highly correlated neighbors, is of special interest for prostate cancer involvement, even though its z-value 3.29 is not overwhelming, $\overline{\text{Fdr}}^{(c)}(z_{637}) = 0.105$.

Table 4.3: Correlation clusters for prostate data using BH(q) with $q = 0.10$. Taking $i_0 = 610$ gave only gene 583 as highly correlated; $i_0 = 583$ gave genes 637 and 610, etc. Genes are listed in order of $\widehat{\text{cor}}(i, i_0)$, largest values first. Gene 637 with $z_i = 3.29$ and $\overline{\text{Fdr}}^{(c)}(z_i) = 0.105$ is the only listed gene besides 610 with $\overline{\text{Fdr}}^{(c)} \le 0.50$.

$$\begin{array}{l} \mathbf{610} \longrightarrow \mathbf{583} \longrightarrow (\mathbf{637}^*, 610^*) \\ \qquad\qquad\quad \downarrow \\ \qquad\quad (583, \mathbf{837}, 878, 5674, 1048, 1618, 1066, 610^*, \text{ and 5 others}) \\ \qquad\qquad\quad \downarrow \\ \qquad\quad (878, 637^*, 1963, 376, 5674) \end{array}$$

The two-sample t-test has almost the same "points on a sphere" description as the one-sample test: $\boldsymbol{x}_i$ is replaced by $\boldsymbol{x}_i^\dagger = (x_{ij} - \bar{x}_i)$ (4.70), $\mathcal{S}_n$ replaced by $\mathcal{S}_{n-1}^\dagger$ (4.73), and $\mathbf{1}_n$, the vector of n 1's, by

$$\mathbf{1}^\dagger \equiv (-\mathbf{1}_{n_1}/n_1, \mathbf{1}_{n_2}/n_2). \tag{4.75}$$

Everything then proceeds as in (4.65) forward, as illustrated in Figure 4.7 (remembering that $\tilde{A}(\theta)$ now refers to the relative areas on an $(n-1)$-dimensional sphere). The same picture applies to more general regression z-values, as mentioned at the end of Section 3.1.

Notes

The true and false discovery terminology comes from Soric (1989) along with a suggestion of the evocative table in Figure 4.1. Benjamini and Hochberg credit Simes (1986) with an early version of the BH algorithm (4.9) and (3.29), but the landmark FDR control theorem (Theorem 4.1) is original to the 1995 BH paper. The neat martingale proof of Theorem 4.1 comes from Storey, Taylor and Siegmund (2004), as does the result that $\widehat{\text{Fdr}}(z)$ (4.47) can be used to control FDR. Efron et al. (2001) presented an empirical

Bayes interpretation of false discovery rates (emphasizing local fdr) while Storey (2002) developed a more explicitly Bayes approach. The Positive Regression Dependence justification for the BH(q) algorithm appears in Benjamini and Yekutieli (2001). Lehmann and Romano (2005a) develop an algorithm that controls the *probability* that Fdp exceeds a certain threshold, rather than $E\{\text{Fdp}\}$. False Negative Rates are extensively investigated in Genovese and Wasserman (2002). Section 1 of Efron (1969) discusses the geometric interpretation of the one-sample t-test. Donoho and Jin (2009) apply Tukey's higher criticism to large-scale selection problems where genuine effects are expected to be very rare.