15

Large-Scale Hypothesis Testing and False-Discovery Rates

By the final decade of the twentieth century, electronic computation fully dominated statistical practice. Almost all applications, classical or otherwise, were now performed on a suite of computer platforms: SAS, SPSS, Minitab, Matlab, S (later R), and others.

The trend accelerates when we enter the twenty-first century, as statistical methodology struggles, most often successfully, to keep up with the vastly expanding pace of scientific data production. This has been a two-way game of pursuit, with statistical algorithms chasing ever larger data sets, while inferential analysis labors to rationalize the algorithms.

Part III of our book concerns topics in twenty-first-century[1] statistics. The word "topics" is intended to signal selections made from a wide catalog of possibilities. Part II was able to review a large portion (though certainly not all) of the important developments during the postwar period. Now, deprived of the advantage of hindsight, our survey will be more illustrative than definitive.

For many statisticians, *microarrays* provided an introduction to large-scale data analysis. These were revolutionary biomedical devices that enabled the assessment of individual activity for thousands of genes at once—and, in doing so, raised the need to carry out thousands of simultaneous hypothesis tests, done with the prospect of finding only a few interesting genes among a haystack of null cases. This chapter concerns large-scale hypothesis testing and the *false-discovery rate*, the breakthrough in statistical inference it elicited.

[1] Actually what historians might call "the long twenty-first century" since we will begin in 1995.

15.1 Large-Scale Testing

The **prostate** cancer data, Figure 3.4, came from a microarray study of $n = 102$ men, 52 prostate cancer patients and 50 normal controls. Each man's gene expression levels were measured on a panel of $N = 6033$ genes, yielding a 6033×102 matrix of measurements x_{ij},

$$x_{ij} = \text{activity of } i\text{th gene for } j\text{th man.} \tag{15.1}$$

For each gene, a two-sample t statistic (2.17) t_i was computed comparing gene i's expression levels for the 52 patients with those for the 50 controls. Under the null hypothesis H_{0i} that the patients' and the controls' responses come from the same normal distribution of gene i expression levels, t_i will follow a standard Student t distribution with 100 degrees of freedom, t_{100}. The transformation

$$z_i = \Phi^{-1}\left(F_{100}(t_i)\right), \tag{15.2}$$

where F_{100} is the cdf of a t_{100} distribution and Φ^{-1} the inverse function of a standard normal cdf, makes z_i standard normal under the null hypothesis:

$$H_{0i} : z_i \sim \mathcal{N}(0, 1). \tag{15.3}$$

Of course the investigators were hoping to spot some *non-null* genes, ones for which the patients and controls respond differently. It can be †1 shown that a reasonable model for both null and non-null genes is[2] †

$$z_i \sim \mathcal{N}(\mu_i, 1), \tag{15.4}$$

μ_i being the *effect size* for gene i. Null genes have $\mu_i = 0$, while the investigators hoped to find genes with large positive or negative μ_i effects.

Figure 15.1 shows the histogram of the 6033 z_i values. The red curve is the scaled $\mathcal{N}(0, 1)$ density that would apply if in fact *all* of the genes were null, that is if all of the μ_i equaled zero.[3] We can see that the curve is a little too high near the center and too low in the tails. Good! Even though most of the genes appear null, the discrepancies from the curve suggest that there are some non-null cases, the kind the investigators hoped to find.

Large-scale testing refers exactly to this situation: having observed a large number N of test statistics, how should we decide which if any of the null hypotheses to reject? Classical testing theory involved only a single case, $N = 1$. A theory of multiple testing arose in the 1960s, "multiple"

[2] This is model (3.28), with z_i now replacing the notation x_i.

[3] It is $ce^{-z^2/2}/\sqrt{2\pi}$ with c chosen to make the area under the curve equal the area of the histogram.

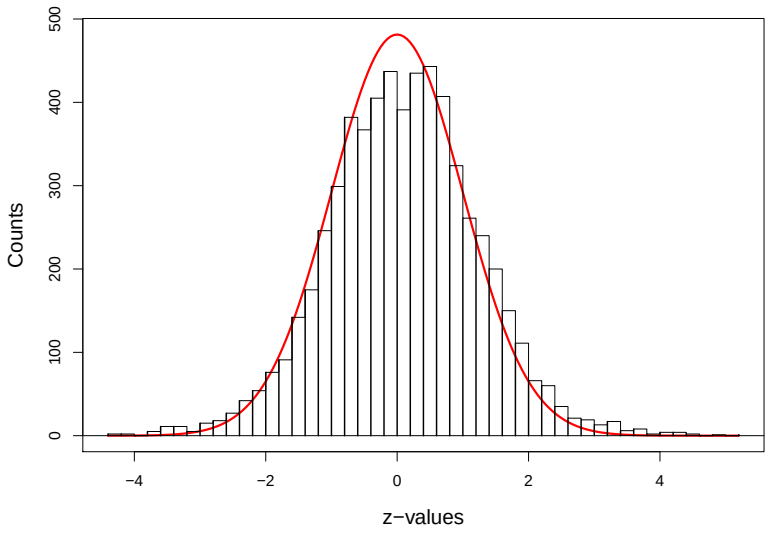

Figure 15.1 Histogram of $N = 6033$ z-values, one for each gene in the prostate cancer study. If all genes were null (15.3) the histogram would track the red curve. For which genes can we reject the null hypothesis?

meaning N between 2 and perhaps 20. The microarray era produced data sets with N in the hundreds, thousands, and now even millions. This sounds like piling difficulty upon difficulty, but in fact there are some inferential advantages to the large-N framework, as we will see.

The most troubling fact about large-scale testing is how easy it is to be fooled. Running 100 separate hypothesis tests at significance level 0.05 will produce about five "significant" results even if each case is actually null. The classical *Bonferroni bound* avoids this fallacy by strengthening the threshold of evidence required to declare an individual case significant (i.e., non-null). For an overall significance level α, perhaps $\alpha = 0.05$, with N simultaneous tests, the Bonferroni bound rejects the ith null hypothesis H_{0i} only if it attains individual significance level α/N. For $\alpha = 0.05$, $N = 6033$, and $H_{0i} : z_i \sim \mathcal{N}(0, 1)$, the one-sided Bonferroni threshold for significance is $-\Phi^{-1}(0.05/N) = 4.31$ (compared with 1.645 for $N = 1$). Only four of the prostate study genes surpass this threshold.

Classic hypothesis testing is usually phrased in terms of *significance levels* and *p-values*. If test statistic z has cdf $F_0(z)$ under the null hypothesis

then[4]

$$p = 1 - F_0(z) \tag{15.5}$$

is the right-sided p-value, larger z giving smaller p-value. "Significance level" refers to a prechosen threshold value, e.g., $\alpha = 0.05$. The null hypothesis is "rejected at level α" if we observe $p \le \alpha$. Table 13.4 on page 245 (where "coverage level" means one minus the significance level) shows Fisher's scale for interpreting p-values.

A level-α test for a single null hypothesis H_0 satisfies, by definition,

$$\alpha = \Pr\{\text{reject true } H_0\}. \tag{15.6}$$

For a collection of N null hypotheses H_{0i}, the *family-wise error rate* is the probability of making even one false rejection,

$$\text{FWER} = \Pr\{\text{reject any true } H_{0i}\}. \tag{15.7}$$

Bonferroni's procedure controls FWER at level α: let I_0 be the indices of the *true* H_{0i}, having say N_0 members. Then

$$\begin{aligned} \text{FWER} &= \Pr\left\{\bigcup_{I_0}\left(p_i \le \frac{\alpha}{N}\right)\right\} \le \sum_{I_0}\Pr\left\{p_i \le \frac{\alpha}{N}\right\} \\ &= N_0\frac{\alpha}{N} \le \alpha, \end{aligned} \tag{15.8}$$

the top line following from Boole's inequality (which doesn't require even independence among the p_i).

The Bonferroni bound is quite conservative: for $N = 6033$ and $\alpha = 0.05$ we reject only those cases having $p_i \le 8.3 \cdot 10^{-6}$. One can do only a little better under the FWER constraint. "Holm's procedure,"† which offers modest improvement over Bonferroni, goes as follows.

†2

- Order the observed p-values from smallest to largest,

$$p_{(1)} \le p_{(2)} \le p_{(3)} \le \cdots \le p_{(i)} \le \cdots \le p_{(N)}, \tag{15.9}$$

  with $H_{0(i)}$ denoting the corresponding null hypotheses.
- Let i_0 be the smallest index i such that

$$p_{(i)} > \alpha/(N - i + 1). \tag{15.10}$$

- *Reject* all null hypotheses $H_{0(i)}$ for $i < i_0$ and *accept* all with $i \ge i_0$.

[4] The left-sided p-value is $p = F_0(z)$. We will avoid two-sided p-values in this discussion.

It can be shown that Holm's procedure controls FWER at level α, while being slightly more generous than Bonferroni in declaring rejections.

15.2 False-Discovery Rates

The FWER criterion aims to control the probability of making even *one* false rejection among N simultaneous hypothesis tests. Originally developed for small-scale testing, say $N \leq 20$, FWER usually proved too conservative for scientists working with N in the thousands. A quite different and more liberal criterion, false-discovery rate (FDR) control, has become standard.

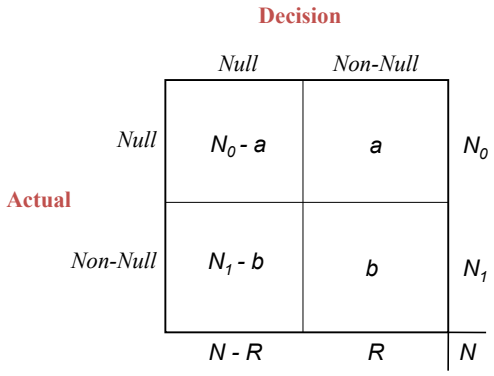

Figure 15.2 A decision rule $\mathcal{D}$ has rejected R out of N null hypotheses; a of these decisions were incorrect, i.e., they were "false discoveries," while b of them were "true discoveries." The false-discovery proportion Fdp equals a/R.

Figure 15.2 diagrams the outcome of a hypothetical decision rule $\mathcal{D}$ applied to the data for N simultaneous hypothesis-testing problems, N_0 null and $N_1 = N - N_0$ non-null. An omniscient oracle has reported the rule's results: R null hypotheses have been rejected; a of these were cases of *false discovery*, i.e., valid null hypotheses, for a "false-discovery proportion" (Fdp) of

$$\text{Fdp}(\mathcal{D}) = a/R. \tag{15.11}$$

(We define Fdp $= 0$ if $R = 0$.) Fdp is unobservable—without the oracle we cannot see a—but under certain assumptions we can control its expectation.

Define

$$\mathrm{FDR}(\mathcal{D}) = E\left\{\mathrm{Fdp}(\mathcal{D})\right\}. \tag{15.12}$$

A decision rule $\mathcal{D}$ controls FDR at level q, with q a prechosen value between 0 and 1, if

$$\mathrm{FDR}(\mathcal{D}) \le q. \tag{15.13}$$

It might seem difficult to find such a rule, but in fact a quite simple but ingenious recipe does the job. Ordering the observed p-values from smallest to largest as in (15.9), define $i_{\max}$ to be the largest index for which

$$p_{(i)} \le \frac{i}{N} q, \tag{15.14}$$

and let $\mathcal{D}_q$ be the rule[5] that rejects $H_{0(i)}$ for $i \le i_{\max}$, accepting otherwise. †3 A proof of the following theorem is referenced in the chapter endnotes.[†]

Theorem (Benjamini–Hochberg FDR Control) *If the p-values corresponding to valid null hypotheses are independent of each other, then*

$$\mathrm{FDR}(\mathcal{D}_q) = \pi_0 q \le q, \qquad \textit{where } \pi_0 = N_0/N. \tag{15.15}$$

In other words $\mathcal{D}_q$ controls FDR at level $\pi_0 q$. The null proportion π_0 is unknown (though estimable), so the usual claim is that $\mathcal{D}_q$ controls FDR at level q. Not much is sacrificed: large-scale testing problems are most often fishing expeditions in which most of the cases are null, putting π_0 near 1, identification of a few non-null cases being the goal. The choice $q = 0.1$ is typical practice.

The popularity of FDR control hinges on the fact that it is more generous than FWER in declaring significance.[6] Holm's procedure (15.10) rejects null hypothesis $H_{0(i)}$ if

$$p_{(i)} \le \mathrm{Threshold(Holm's)} = \frac{\alpha}{N - i + 1}, \tag{15.16}$$

while $\mathcal{D}_q$ (15.13) has threshold

$$p_{(i)} \le \mathrm{Threshold}(\mathcal{D}_q) = \frac{q}{N} i. \tag{15.17}$$

[5] Sometimes denoted "BH$_q$" after its inventors Benjamini and Hochberg; see the chapter endnotes.

[6] The classic term "significant" for a non-null identification doesn't seem quite right for FDR control, especially given the Bayesian connections of Section 15.3, and we will sometimes use "interesting" instead.

In the usual range of interest, large N and small i, the ratio

$$\frac{\text{Threshold}(\mathcal{D}_q)}{\text{Threshold(Holm's)}} = \frac{q}{\alpha}\left(1 - \frac{i-1}{N}\right) i \tag{15.18}$$

increases almost linearly with i.

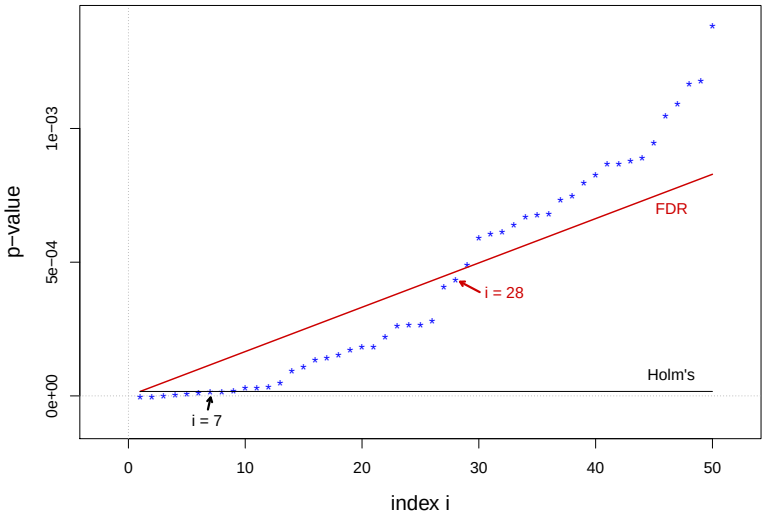

Figure 15.3 Ordered p-values $p_{(i)} = 1 - \Phi(z_{(i)})$ plotted versus i for the 50 largest z-values from the **prostate** data in Figure 15.1. The FDR control boundary (algorithm $\mathcal{D}_q$, $q = 0.1$) rejects $H_{0(i)}$ for the 28 smallest values $p_{(i)}$, while Holm's FWER procedure ($\alpha = 0.1$) rejects for only the 7 smallest values. (The upward slope of Holm's boundary (15.16) is too small to see here.)

Figure 15.3 illustrates the comparison for the right tail of the prostate data of Figure 15.1, with $p_i = 1 - \Phi(z_i)$ (15.3), (15.5), and $\alpha = q = 0.1$. The FDR procedure rejects $H_{0(i)}$ for the 28 largest z-values ($z_{(i)} \geq 3.33$), while FWER control rejects only the 7 most extreme z-values ($z_{(i)} \geq 4.14$).

Hypothesis testing has been a traditional stronghold of frequentist decision theory, with "Type 1" error control being strictly enforced, very often at the 0.05 level. It is surprising that a new control criterion, FDR, has taken hold in large-scale testing situations. A critic, noting FDR's relaxed rejection standards in Figure 15.3, might raise some pointed questions.

1 Is controlling a *rate* (i.e., FDR) as meaningful as controlling a *probability* (of Type 1 error)?
2 How should q be chosen?
3 The control theorem depends on independence among the p-values. Isn't this unlikely in situations such as the prostate study?
4 The FDR significance for gene i_0, say one with $z_{i_0} = 3$, depends on the results of all the other genes: the more "other" z_i values exceed 3, the more interesting gene i_0 becomes (since that increases i_0's index i in the ordered list (15.9), making it more likely that p_{i_0} lies below the $\mathcal{D}_q$ threshold (15.14)). Does this make inferential sense?

A Bayes/empirical Bayes restatement of the $\mathcal{D}_q$ algorithm helps answer these questions, as discussed next.

15.3 Empirical Bayes Large-Scale Testing

In practice, single-case hypothesis testing has been a frequentist preserve. Its methods demand little from the scientist—only the choice of a test statistic and the calculation of its null distribution—while usually delivering a clear verdict. By contrast, Bayesian model selection, whatever its inferential virtues, raises the kinds of difficult modeling questions discussed in Section 13.3.

It then comes as a pleasant surprise that things are different for large-scale testing: Bayesian methods, at least in their empirical Bayes manifestation, no longer demand heroic modeling efforts, and can help untangle the interpretation of simultaneous test results. This is particularly true for the FDR control algorithm $\mathcal{D}_q$ of the previous section.

A simple Bayesian framework for simultaneous testing is provided by the *two-groups model*: each of the N cases (the genes for the prostate study) is either null with prior probability π_0 or non-null with probability $\pi_1 = 1 - \pi_0$; the resulting observation z then has density either $f_0(z)$ or $f_1(z)$,

$$\begin{aligned} \pi_0 &= \Pr\{\text{null}\} && f_0(z) \text{ density if null,} \\ \pi_1 &= \Pr\{\text{non-null}\} && f_1(z) \text{ density if non-null.} \end{aligned} \tag{15.19}$$

For the prostate study, π_0 is nearly 1, and $f_0(z)$ is the standard normal density $\phi(z) = \exp(-z^2/2)/\sqrt{2\pi}$ (15.3), while the non-null density remains to be estimated.

Let $F_0(z)$ and $F_1(z)$ be the cdf values corresponding to $f_0(z)$ and $f_1(z)$,

with "survival curves"

$$S_0(z) = 1 - F_0(z) \quad \text{and} \quad S_1(z) = 1 - F_1(z), \tag{15.20}$$

$S_0(z_0)$ being the probability that a null z-value exceeds z_0, and similarly for $S_1(z)$. Finally, define $S(z)$ to be the mixture survival curve

$$S(z) = \pi_0 S_0(z) + \pi_1 S_1(z). \tag{15.21}$$

The *mixture density*

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z) \tag{15.22}$$

determines $S(z)$,

$$S(z_0) = \int_{z_0}^{\infty} f(z)\, dz. \tag{15.23}$$

Suppose now that observation z_i for case i is seen to exceed some threshold value z_0, perhaps $z_0 = 3$. Bayes' rule gives

$$\begin{aligned} \text{Fdr}(z_0) &\equiv \Pr\{\text{case } i \text{ is null} | z_i \geq z_0\} \\ &= \pi_0 S_0(z_0)/S(z_0), \end{aligned} \tag{15.24}$$

the correspondence with (3.5) on page 23 being $\pi_0 = g(\mu)$, $S_0(z_0) = f_\mu(x)$, and $S(z_0) = f(x)$. Fdr is the "Bayes false-discovery rate," as contrasted with the frequentist quantity FDR (15.12).

In typical applications, $S_0(z_0)$ is assumed known[7] (equaling $1 - \Phi(z_0)$ in the prostate study), and π_0 is assumed to be near 1. The denominator $S(z_0)$ in (15.24) is unknown, but—and this is the crucial point—it has an obvious estimate in large-scale testing situations, namely

$$\hat{S}(z_0) = N(z_0)/N, \qquad \text{where } N(z_0) = \#\{z_i \geq z_0\}. \tag{15.25}$$

(By the definition of the two-group model, each z_i has marginal density $f(z)$, making $\hat{S}(z_0)$ the usual empirical estimate of $S(z_0)$ (15.23).) Plugging into (15.24) yields an empirical Bayes estimate of the Bayes false-discovery rate

$$\widehat{\text{Fdr}}(z_0) = \pi_0 S_0(z_0) \big/ \hat{S}(z_0). \tag{15.26}$$

The connection with FDR control is almost immediate. First of all, from definitions (15.5) and (15.20) we have $p_i = S_0(z_i)$; also for the ith from the largest z-value we have $\hat{S}(z_{(i)}) = i/N$ (15.25). Putting these together, condition (15.14), $p_{(i)} \leq (i/N)q$, becomes

$$S_0(z_{(i)}) \leq \hat{S}(z_{(i)}) \cdot q, \tag{15.27}$$

[7] But see Section 15.5.

or $S_0(z_{(i)})/\hat{S}(z_{(i)}) \leq q$, which can be written as

$$\widehat{\text{Fdr}}(z_{(i)}) \leq \pi_0 q \tag{15.28}$$

(15.26). In other words, the $\mathcal{D}_q$ algorithm, which rejects those null hypotheses having[8] $p_{(i)} \leq (i/N)q$, is in fact rejecting those cases for which the empirical Bayes posterior probability of nullness is too small, as defined by (15.28). The Bayesian nature of FDR control offers a clear advantage to the investigating scientist, who gets a numerical assessment of the probability that he or she will be wasting time following up any one of the selected cases.

We can now respond to the four questions at the end of the previous section:

1 FDR control *does* relate to a probability—the Bayes posterior probability of nullness.
2 The choice of q for $\mathcal{D}_q$ amounts to setting the maximum tolerable amount of Bayes risk of nullness[9] (usually after taking $\pi_0 = 1$ in (15.28)).
3 Most often the z_i, and hence the p_i, will be correlated with each other. Even under correlation, however, $\hat{S}(z_0)$ in (15.25) is still unbiased for $S(z_0)$, making $\widehat{\text{Fdr}}(z_0)$ (15.26) nearly unbiased for $\text{Fdr}(z_0)$ (15.24). There *is* a price to be paid for correlation, which increases the *variance* of $S_0(z_0)$ and $\widehat{\text{Fdr}}(z_0)$.
4 In the Bayes two-groups model (15.19), all of the non-null z_i are i.i.d. observations from the non-null density $f_1(z)$, with survival curve $S_1(z)$. The number of null cases z_i exceeding some threshold z_0 has *fixed* expectation $N\pi_0 S_0(z_0)$. Therefore an increase in the number of observed values z_i exceeding z_0 must come from a heavier right tail for $f_1(z)$, implying a greater posterior probability of non-nullness $\text{Fdr}(z_0)$ (15.24). This point is made more clearly in the *local false-discovery* framework of the next section. It emphasizes the "learning from the experience of others" aspect of empirical Bayes inference, Section 7.4. The question of "Which others?" is returned to in Section 15.6.

Figure 15.4 illustrates the two-group model (15.19). The N cases are

[8] The algorithm, as stated just before the FDR control theorem (15.15), is actually a little more liberal in allowing rejections.

[9] For a case of particular interest, the calculation can be reversed: if the case has ordered index i then, according to (15.14), the value $q = Np_i/i$ puts it exactly on the boundary of rejection, making this its q-value. The 50th largest z-value for the prostate data has $z_i = 2.99$, $p_i = 0.00139$, and q-value 0.168, that being both the frequentist boundary for rejection and the empirical Bayes probability of nullness.

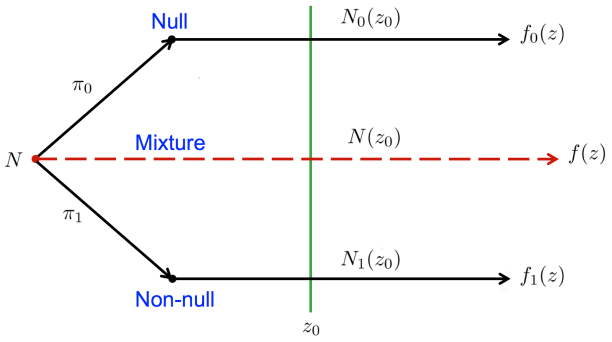

Figure 15.4 A diagram of the two-groups model (15.19). Here the statistician observes values z_i from a mixture density $f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$ and decides to reject or accept the null hypothesis H_{0i} depending on whether z_i exceeds or is less than the threshold value z_0.

randomly dispatched to the two arms in proportions π_0 and π_1, at which point they produce z-values according to either $f_0(z)$ or $f_1(z)$. Suppose we are using a simple decision rule $\mathcal{D}$ that rejects the ith null hypothesis if z_i exceeds some threshold z_0, and accepts otherwise,

$$\mathcal{D}: \begin{cases} \text{Reject } H_{0i} & \text{if } z_i > z_0 \\ \text{Accept } H_{0i} & \text{if } z_i \le z_0. \end{cases} \tag{15.29}$$

The oracle of Figure 15.2 knows that $N_0(z_0) = a$ of the null case z-values exceeded z_0, and similarly $N_1(z_0) = b$ of the non-null cases, leading to

$$N(z_0) = N_0(z_0) + N_1(z_0) = R \tag{15.30}$$

total rejections. The false-discovery proportion (15.11) is

$$\text{Fdp} = \frac{N_0(z_0)}{N(z_0)} \tag{15.31}$$

but this is unobservable since we see only $N(z_0)$.

The clever inferential strategy of false-discovery rate theory substitutes the *expectation* of $N_0(z_0)$,

$$E\{N_0(z_0)\} = N\pi_0 S_0(z_0), \tag{15.32}$$

for $N_0(z_0)$ in (15.31), giving

$$\widehat{\text{Fdp}} = \frac{N\pi_0 S_0(z_0)}{N(z_0)} = \frac{\pi_0 S_0(z_0)}{\hat{S}(z_0)} = \widehat{\text{Fdr}}(z_0), \tag{15.33}$$

using (15.25) and (15.26). Starting from the two-groups model, $\widehat{\text{Fdr}}(z_0)$ is an obvious empirical (i.e., frequentist) estimate of the Bayesian probability $\text{Fdr}(z_0)$, as well as of Fdp.

If placed in the Bayes–Fisher–frequentist triangle of Figure 14.1, false-discovery rates would begin life near the frequentist corner but then migrate at least part of the way toward the Bayes corner. There are remarkable parallels with the James–Stein estimator of Chapter 7. Both theories began with a striking frequentist theorem, which was then inferentially rationalized in empirical Bayes terms. Both rely on the use of indirect evidence—learning from the experience of others. The difference is that James–Stein estimation always aroused controversy, while FDR control has been quickly welcomed into the pantheon of widely used methods. This could reflect a change in twenty-first-century attitudes or, perhaps, only that the $\mathcal{D}_q$ rule better conceals its Bayesian aspects.

15.4 Local False-Discovery Rates

Tail-area statistics (p-values) were synonymous with classic one-at-a-time hypothesis testing, and the $\mathcal{D}_q$ algorithm carried over p-value interpretation to large-scale testing theory. But tail-area calculations are neither necessary nor desirable from a Bayesian viewpoint, where, having observed test statistic z_i equal to some value z_0, we should be more interested in the probability of nullness given $z_i = z_0$ than given $z_i \geq z_0$.

To this end we define the *local false-discovery rate*

$$\text{fdr}(z_0) = \Pr\{\text{case } i \text{ is null} | z_i = z_0\} \tag{15.34}$$

as opposed to the tail-area false-discovery rate $\text{Fdr}(z_0)$ (15.24). The main point of what follows is that reasonably accurate empirical Bayes estimates of fdr are available in large-scale testing problems.

As a first try, suppose that $\mathcal{Z}_0$, a proposed region for rejecting null hypotheses, is a small interval centered at z_0,

$$\mathcal{Z}_0 = \left[z_0 - \frac{d}{2}, z_0 + \frac{d}{2}\right], \tag{15.35}$$

with d perhaps 0.1. We can redraw Figure 15.4, now with $N_0(\mathcal{Z}_0)$, $N_1(\mathcal{Z}_0)$,

and $N(\mathcal{Z}_0)$ the null, non-null, and total number of z-values in $\mathcal{Z}_0$. The local false-discovery proportion,

$$\text{fdp}(z_0) = N_0(\mathcal{Z}_0)/N(\mathcal{Z}_0) \tag{15.36}$$

is unobservable, but we can replace $N_0(\mathcal{Z}_0)$ with $N\pi_0 f_0(z_0)d$, its approximate expectation as in (15.31)–(15.33), yielding the estimate[10]

$$\widehat{\text{fdr}}(z_0) = N\pi_0 f_0(z_0)d/N(\mathcal{Z}_0). \tag{15.37}$$

Estimate (15.37) would be needlessly noisy in practice; z-value distributions tend to be smooth, allowing the use of regression estimates for $\text{fdr}(z_0)$. Bayes' theorem gives

$$\text{fdr}(z) = \pi_0 f_0(z)/f(z) \tag{15.38}$$

in the two-groups model (15.19) (with μ in (3.5) now the indicator of null or non-null states, and x now z). Drawing a smooth curve $\hat{f}(z)$ through the histogram of the z-values yields the more efficient estimate

$$\widehat{\text{fdr}}(z_0) = \pi_0 f_0(z_0)/\hat{f}(z_0); \tag{15.39}$$

the null proportion π_0 can be estimated—see Section 15.5—or set equal to 1.

Figure 15.5 shows $\widehat{\text{fdr}}(z)$ for the prostate study data of Figure 15.1, where $\hat{f}(z)$ in (15.39) has been estimated as described below. The curve hovers near 1 for the 93% of the cases having $|z_i| \le 2$, sensibly suggesting that there is no involvement with prostate cancer for most genes. It declines quickly for $|z_i| \ge 3$, reaching the conventionally "interesting" threshold

$$\widehat{\text{fdr}}(z) \le 0.2 \tag{15.40}$$

for $z_i \ge 3.34$ and $z_i \le -3.40$. This was attained for 27 genes in the right tail and 25 in the left, these being reasonable candidates to flag for follow-up investigation.

The curve $\hat{f}(z)$ used in (15.39) was obtained from a fourth-degree log polynomial Poisson regression fit to the histogram in Figure 15.1, as in Figure 10.5 (10.52)–(10.56). Log polynomials of degree 2 through 6 were fit by maximum likelihood, giving total residual deviances (8.35) shown in Table 15.1. An enormous improvement in fit is seen in going from degree 3 to 4, but nothing significant after that, with decreases less than the null value 2 suggested by (12.75).

[10] Equation (15.37) makes argument (4) of the previous section clearer: having more "other" z-values fall into $\mathcal{Z}_0$ increases $N(\mathcal{Z}_0)$, decreasing $\widehat{\text{fdr}}(z_0)$ and making it more likely that $z_i = z_0$ represents a non-null case.

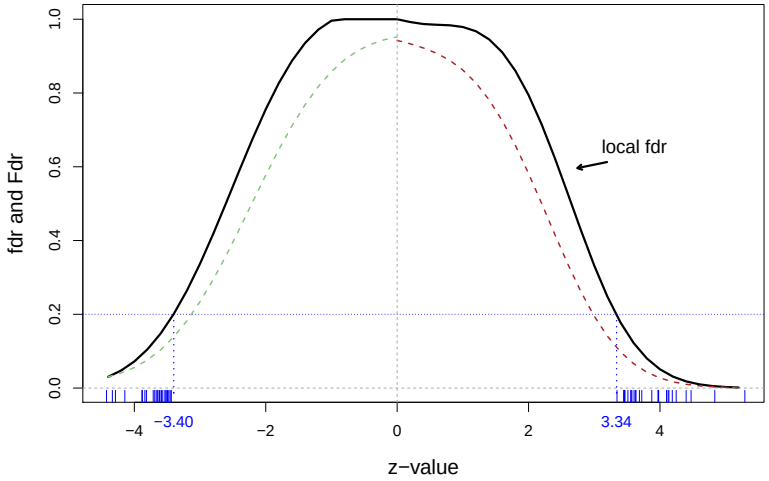

Figure 15.5 Local false-discovery rate estimate $\widehat{\text{fdr}}(z)$ (15.39) for prostate study of Figure 15.1; 27 genes on the right and 25 on the left, indicated by dashes, have $\widehat{\text{fdr}}(z_i) \leq 0.2$; light dashed curves are the left and right tail-area estimates $\widehat{\text{Fdr}}(z)$ (15.26).

Table 15.1 *Total residual deviances from log polynomial Poisson regressions of the prostate data, for polynomial degrees 2 through 6; degree 4 is preferred.*

Degree	2	3	4	5	6
Deviance	138.6	137.1	65.3	64.3	63.8

The points in Figure 15.6 represent the log bin counts from the histogram in Figure 15.1 (excluding zero counts), with the solid curve showing the 4th-degree MLE polynomial fit. Also shown is the standard normal log density

$$\log f_0(z) = -\frac{1}{2}z^2 + \text{constant}. \tag{15.41}$$

It fits reasonably well for $|z| < 2$, emphasizing the null status of the gene majority.

The cutoff $\widehat{\text{fdr}}(z) \leq 0.2$ for declaring a case interesting is not completely arbitrary. Definitions (15.38) and (15.22), and a little algebra, show that it

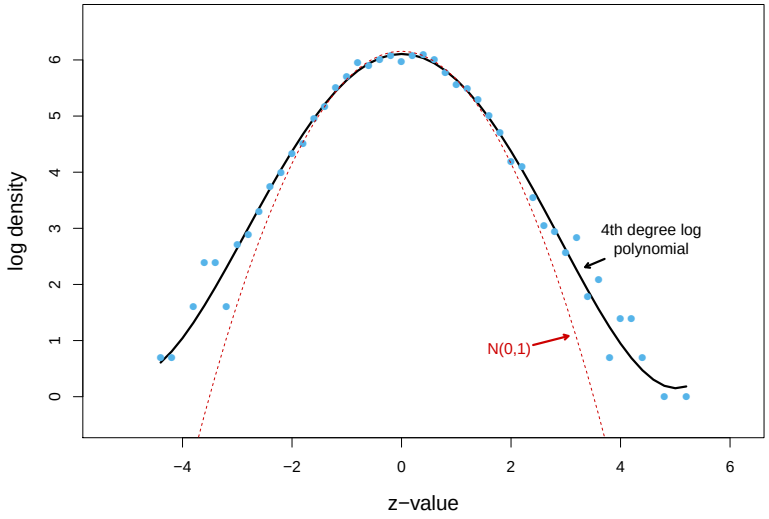

Figure 15.6 Points are log bin counts for Figure 15.1's histogram. The solid black curve is a fourth-degree log-polynomial fit used to calculate $\widehat{\text{fdr}}(z)$ in Figure 15.5. The dashed red curve, the log null density (15.41), provides a reasonable fit for $|z| \le 2$.

is equivalent to

$$\frac{f_1(z)}{f_0(z)} \ge 4\frac{\pi_0}{\pi_1}. \tag{15.42}$$

If we assume $\pi_0 \ge 0.90$, as is reasonable in most large-scale testing situations, this makes the Bayes factor $f_1(z)/f_0(z)$ quite large,

$$\frac{f_1(z)}{f_0(z)} \ge 36, \tag{15.43}$$

"strong evidence" against the null hypothesis in Jeffreys' scale, Table 13.3.

There is a simple relation between the local and tail-area false-discovery rates:† †4

$$\text{Fdr}(z_0) = E\{\text{fdr}(z) | z \ge z_0\}; \tag{15.44}$$

so $\text{Fdr}(z_0)$ is the average value of $\text{fdr}(z)$ for z greater than z_0. In interesting situations, $\text{fdr}(z)$ will be a decreasing function for large values of z, as on the right side of Figure 15.5, making $\text{Fdr}(z_0) < \text{fdr}(z_0)$. This accounts

for the conventional significance cutoff $\widehat{\text{Fdr}}(z) \le 0.1$ being smaller than $\widehat{\text{fdr}}(z) \le 0.2$ (15.40).[†] †5

The Bayesian interpretation of local false-discovery rates carries with it the advantages of Bayesian coherency. We don't have to change definitions as with left-sided and right-sided tail-area $\widehat{\text{Fdr}}$ estimates, since $\widehat{\text{fdr}}(z)$ applies without change to both tails.[11] Also, we don't need a separate theory for "true-discovery rates," since

$$\text{tdr}(z_0) \equiv 1 - \text{fdr}(z_0) = \pi_1 f_1(z_0)/f(z_0) \tag{15.45}$$

is the conditional probability that case i is *non-null* given $z_i = z_0$.

15.5 Choice of the Null Distribution

The null distribution, $f_0(z)$ in the two-groups model (15.19), plays a crucial role in large-scale testing, just as it does in the classic single-case theory. Something different however happens in large-scale problems: with thousands of z-values to examine at once, it can become clear that the conventional theoretical null is inappropriate for the situation at hand. Put more positively, large-scale applications may allow us to empirically determine a more realistic null distribution.

The `police` data of Figure 15.7 illustrates what can happen. Possible racial bias in pedestrian stops was assessed for $N = 2749$ New York City police officers in 2006. Each officer was assigned a score z_i, large positive scores suggesting racial bias. The z_i values were summary scores from a complicated logistic regression model intended to compensate for differences in the time of day, location, and context of the stops. Logistic regression theory suggested the theoretical null distribution

$$H_{0i} : z_i \sim \mathcal{N}(0, 1) \tag{15.46}$$

for the absence of racial bias.

The trouble is that the center of the z-value histogram in Figure 15.7, which should track the $\mathcal{N}(0, 1)$ curve applying to the presumably large fraction of null-case officers, is much too wide. (Unlike the situation for the prostate data in Figure 15.1.) An MLE fitting algorithm discussed below produced the *empirical null*

$$H_{0i} : z_i \sim \mathcal{N}(0.10, 1.40^2) \tag{15.47}$$

[11] Going further, z in the two-groups model could be multidimensional. Then tail-area false-discovery rates would be unavailable, but (15.38) would still legitimately define $\text{fdr}(z)$.

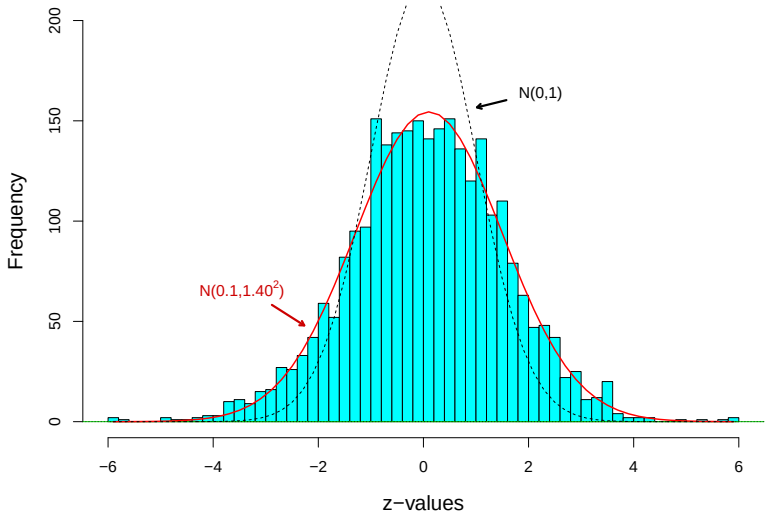

Figure 15.7 Police data; histogram of z scores for $N = 2749$ New York City police officers, with large z_i suggesting racial bias. The center of the histogram is too wide compared with the theoretical null distribution $z_i \sim \mathcal{N}(0, 1)$. An MLE fit to central data gave $\mathcal{N}(0.10, 1.40^2)$ as empirical null.

as appropriate here. This is reinforced by a QQ plot of the z_i values shown in Figure 15.8, where we see most of the cases falling nicely along a $\mathcal{N}(0.09, 1.42^2)$ line, with just a few outliers at both extremes.

There is a lot at stake here. Based on the empirical null (15.47) only four officers reached the "probably racially biased" cutoff $\widehat{\text{fdr}}(z_i) \le 0.2$, the four circled points at the far right of Figure 15.8; the fifth point had $\widehat{\text{fdr}} = 0.38$ while all the others exceeded 0.80. The theoretical $\mathcal{N}(0, 1)$ null was much more severe, assigning $\widehat{\text{fdr}} \le 0.2$ to the 125 officers having $z_i \ge 2.50$. One can imagine the difference in newspaper headlines.

From a classical point of view it seems heretical to question the theoretical null distribution, especially since there is no substitute available in single-case testing. Once alerted by data sets like the police study, however, it is easy to list reasons for doubt:

- *Asymptotics* Taylor series approximations go into theoretical null calculations such as (15.46), which can lead to inaccuracies, particularly in the crucial tails of the null distribution.
- *Correlations* False-discovery rate methods are correct *on the average*,

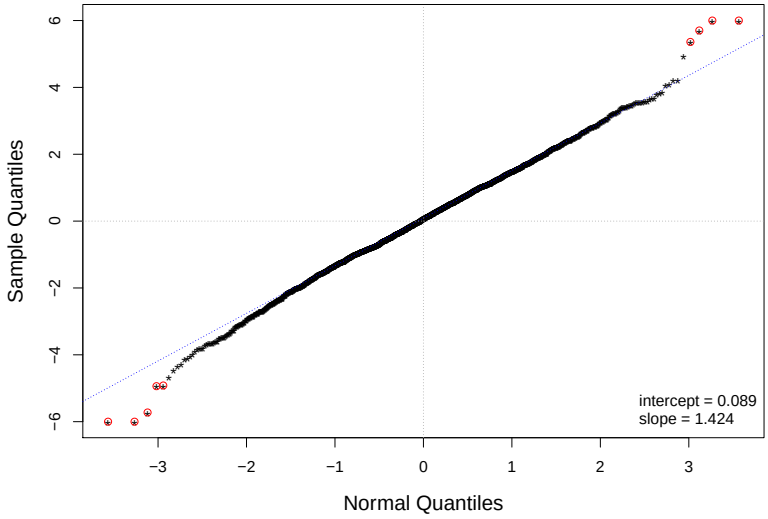

Figure 15.8 QQ plot of police data z scores; most scores closely follow the $\mathcal{N}(0.09, 1.42^2)$ line with a few outliers at either end. The circled points are cases having local false-discovery estimate $\widehat{\text{fdr}}(z_i) \leq 0.2$, based on the empirical null. Using the theoretical $\mathcal{N}(0, 1)$ null gives 216 cases with $\widehat{\text{fdr}}(z_i) \leq 0.2$, 91 on the left and 125 on the right.

even with correlations among the N z-values. However, severe correlation destabilizes the z-value histogram, which can become randomly wider or narrower than theoretically predicted, undermining theoretical null results for the data set at hand.† †6

- *Unobserved covariates* The police study was *observational*: individual encounters were not assigned at random to the various officers but simply observed as they happened. Observed covariates such as the time of day and the neighborhood were included in the logistic regression model, but one can never rule out the possibility of influential unobserved covariates.
- *Effect size considerations* The hypothesis-testing setup, where a large fraction of the cases are truly null, may not be appropriate. An *effect size* model, with $\mu_i \sim g(\cdot)$ and $z_i \sim \mathcal{N}(\mu_i, 1)$, might apply, with the prior $g(\mu)$ *not* having an atom at $\mu = 0$. The nonatomic choice $g(\mu) \sim \mathcal{N}(0.10, 0.63^2)$ provides a good fit to the QQ plot in Figure 15.8.

Empirical Null Estimation

Our point of view here is that the theoretical null (15.46), $z_i \sim \mathcal{N}(0,1)$, is not completely wrong but needs adjustment for the data set at hand. To this end we assume the two-groups model (15.19), with $f_0(z)$ normal but not necessarily $\mathcal{N}(0,1)$, say

$$f_0(z) \sim \mathcal{N}(\delta_0, \sigma_0^2). \tag{15.48}$$

In order to compute the local false-discovery rate $\text{fdr}(z) = \pi_0 f_0(z)/f(z)$ we want to estimate the three numerator parameters $(\delta_0, \sigma_0, \pi_0)$, the mean and standard deviation of the null density and the proportion of null cases. (The denominator $f(z)$ is estimated as in Section 15.4.)

Our key assumptions (besides (15.48)) are that π_0 is large, say $\pi_0 \geq 0.90$, and that most of the z_i near 0 are null cases. The algorithm **locfdr** †begins by selecting a set $\mathcal{A}_0$ near $z = 0$ in which it is assumed that *all* the †7
z_i in $\mathcal{A}_0$ are null; in terms of the two-groups model, the assumption can be stated as

$$f_1(z) = 0 \text{ for } z \in \mathcal{A}_0. \tag{15.49}$$

Modest violations of (15.49), which are to be expected, produce small biases in the empirical null estimates. Maximum likelihood based on the number and values of the z_i observed in $\mathcal{A}_0$ yield the empirical null estimates† $(\hat\delta_0, \hat\sigma_0, \hat\pi_0)$. †8

Applied to the police data, **locfdr** chose $\mathcal{A}_0 = [-1.8, 2.0]$ and produced estimates

$$\left(\hat\delta_0, \hat\sigma_0, \hat\pi_0\right) = (0.10, 1.40, 0.989). \tag{15.50}$$

Two small simulation studies described in Table 15.2 give some idea of the variabilities and biases inherent in the **locfdr** estimation process.

The third method, somewhere between the theoretical and empirical null estimates but closer to the former, relies on permutations. The vector $\boldsymbol{z}$ of 6033 z-values for the **prostate** data of Figure 15.1 was obtained from a study of 102 men, 52 cancer patients and 50 controls. Randomly permuting the men's data, that is randomly choosing 50 of the 102 to be "controls" and the remaining 52 to be "patients," and then carrying through steps (15.1)–(15.2) gives a vector $\boldsymbol{z}^*$ in which any actual cancer/control differences have been suppressed. A histogram of the z_i^* values (perhaps combining several permutations) provides the "permutation null." Here we are extending Fisher's original permutation idea, Section 4.4, to large-scale testing.

Ten permutations of the prostate study data produced an almost perfect

Table 15.2 *Means and standard deviations of* $(\hat\delta_0, \hat\sigma_0, \hat\pi_0)$ *for two simulation studies of empirical null estimation using* `locfdr`. $N = 5000$ *cases each trial with* $(\delta_0, \sigma_0, \pi_0)$ *as shown; 250 trials; two-groups model* (15.19) *with non-null density* $f_1(z)$ *equal to* $\mathcal{N}(3, 1)$ *(left side) or* $\mathcal{N}(4.2, 1)$ *(right side).*

	δ_0	σ_0	π_0	δ_0	σ_0	π_0
true	**0**	**1.0**	**.95**	**.10**	**1.40**	**.95**
mean	.015	1.017	.962	.114	1.418	.958
st dev	.019	.017	.005	.025	.029	.006

$\mathcal{N}(0, 1)$ permutation null. (This is as expected from the classic theory of permutation t-tests.) Permutation methods reliably overcome objection 1 to the theoretical null distribution, over-reliance on asymptotic approxima-
†9 tions, but cannot cure objections 2, 3, and 4.†

Whatever the cause of disparity, the operational difference between the theoretical and empirical null distribution is clear: with the latter, the significance of an outlying case is judged relative to the dispersion of the majority, not by a theoretical yardstick as with the former. This was persuasive for the police data, but the story isn't one-sided. Estimating the null distribution adds substantially to the variability of $\widehat{\text{fdr}}$ or $\widehat{\text{Fdr}}$. For situations such as the prostate data, when the theoretical null looks nearly correct,[12] it is reasonable to stick with it.

The very large data sets of twenty-first-century applications encourage self-contained methodology that proceeds from just the data at hand using a minimum of theoretical constructs. False-discovery rate empirical Bayes analysis of large-scale testing problems, with data-based estimation of $\hat\pi_0$, $\hat f_0$, and $\hat f$, comes close to the ideal in this sense.

15.6 Relevance

False-discovery rates return us to the purview of *indirect evidence*, Sections 6.4 and 7.4. Our interest in any one gene in the prostate cancer study depends on its own z score of course, but also on the other genes' scores—"learning from the experience of others," in the language used before.

The crucial question we have been avoiding is "Which others?" Our tacit answer has been "All the cases that arrive in the same data set," all the genes

[12] The `locfdr` algorithm gave $(\hat\delta_0, \hat\sigma_0, \hat\pi_0) = (0.00, 1.06, 0.984)$ for the prostate data.

in the prostate study, all the officers in the police study. Why this can be a dangerous tactic is shown in our final example.

A **DTI** (diffusion tensor imaging) study compared six dyslexic children with six normal controls. Each **DTI** scan recorded fluid flows at $N = 15{,}443$ "voxels," i.e., at 15,443 three-dimensional brain coordinates. A score z_i comparing dyslexics with normal controls was calculated for each voxel i, calibrated such that the theoretical null distribution of "no difference" was

$$H_{0i} : z_i \sim \mathcal{N}(0, 1) \tag{15.51}$$

as at (15.3).

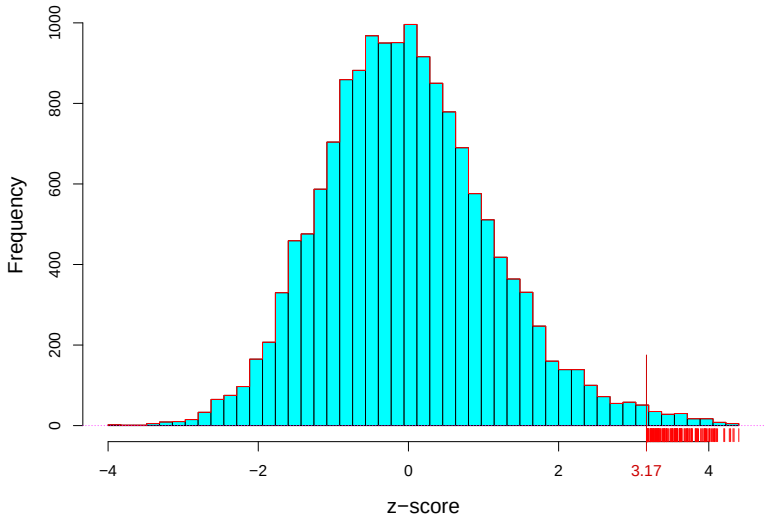

Figure 15.9 Histogram of z scores for the **DTI** study, comparing dyslexic versus normal control children at 15,443 brain locations. A FDR analysis based on the empirical null distribution gave 149 voxels with $\widehat{\text{fdr}}(z_i) \le 0.20$, those having $z_i \ge 3.17$ (indicated by red dashes).

Figure 15.9 shows the histogram of all 15,443 z_i values, normal-looking near the center and with a heavy right tail; **locfdr** gave empirical null parameters

$$\left(\hat{\delta}_0, \hat{\sigma}_0, \hat{\pi}_0\right) = (-0.12, 1.06, 0.984), \tag{15.52}$$

the 149 voxels with $z_i \ge 3.17$ having $\widehat{\text{fdr}}$ values ≤ 0.20. Using the the-

oretical null (15.51) yielded only modestly different results, now the 177 voxels with $z_i \geq 3.07$ having $\widehat{\text{fdr}}_i \leq 0.20$.

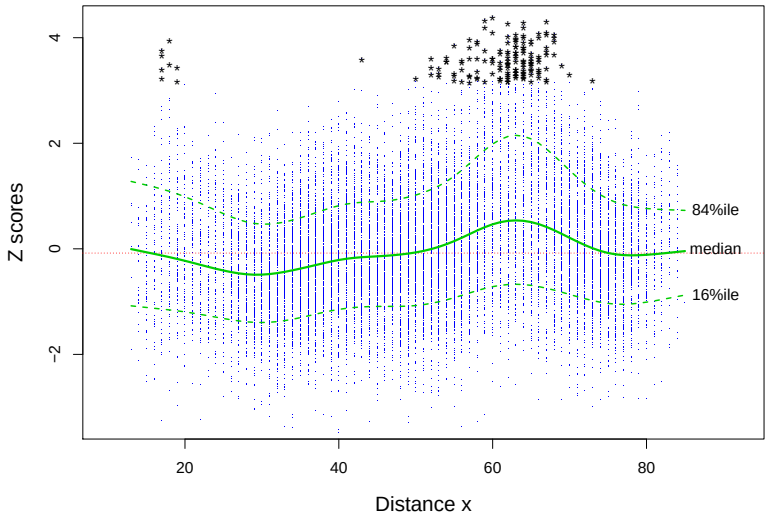

Figure 15.10 A plot of 15,443 z_i scores from a `DTI` study (vertical axis) and voxel distances x_i from the back of the brain (horizontal axis). The starred points are the 149 voxels with $\widehat{\text{fdr}}(z_i) \leq 0.20$, which occur mostly for x_i in the interval $[50, 70]$.

In Figure 15.10 the voxel scores z_i, graphed vertically, are plotted versus x_i, the voxel's distance from the back of the brain. Waves of differing response are apparent. Larger values occur in the interval $50 \leq x \leq 70$, where the entire z-value distribution—low, medium, and high—is pushed up. Most of the 149 voxels having $\widehat{\text{fdr}}_i \leq 0.20$ occur at the top of this wave.

Figure 15.10 raises the problem of fair comparison. Perhaps the 4,653 voxels with x_i between 50 and 70 should be compared only with each other, and not with all 15,443 cases. Doing so gave

$$\left(\hat{\delta}_0, \hat{\sigma}_0, \hat{\pi}_0\right) = (0.23, 1.18, 0.970), \tag{15.53}$$

only 66 voxels having $\widehat{\text{fdr}}_i \leq 0.20$, those with $z_i \geq 3.57$.

All of this is a question of *relevance*: which other voxels i are relevant to the assessment of significance for voxel i_0? One might argue that this is a question for the scientist who gathers the data and not for the statistical analyst, but that is unlikely to be a fruitful avenue, at least not without

a lot of back-and-forth collaboration. Standard Bayesian analysis solves the problem by dictate: the assertion of a prior is also an assertion of its relevance. Empirical Bayes situations expose the dangers lurking in such assertions.

Relevance was touched upon in Section 7.4, where the limited translation rule (7.47) was designed to protect extreme cases from being shrunk too far toward the bulk of ordinary ones. One could imagine having a "relevance function" $\rho(x_i, z_i)$ that, given the covariate information x_i and response z_i for case$_i$, somehow adjusts an ensemble false-discovery rate estimate to correctly apply to the case of interest—but such a theory barely exists.† †10

Summary

Large-scale testing, particularly in its false-discovery rate implementation, is not at all the same thing as the classic Fisher–Neyman–Pearson theory:

- Frequentist single-case hypothesis testing depends on the theoretical long-run behavior of samples from the theoretical null distribution. With data available from say $N = 5000$ simultaneous tests, the statistician has his or her own "long run" in hand, diminishing the importance of theoretical modeling. In particular, the data may cast doubt on the theoretical null, providing a more appropriate empirical null distribution in its place.
- Classic testing theory is purely frequentist, whereas false-discovery rates combine frequentist and Bayesian thinking.
- In classic testing, the attained significance level for case i depends only on its own score z_i, while $\widehat{\text{fdr}}(z_i)$ or $\widehat{\text{Fdr}}(z_i)$ also depends on the observed z-values for other cases.
- Applications of single-test theory usually hope for *rejection* of the null hypothesis, a familiar prescription being 0.80 power at size 0.05. The opposite is true for large-scale testing, where the usual goal is to *accept* most of the null hypotheses, leaving just a few interesting cases for further study.
- Sharp null hypotheses such as $\mu = 0$ are less important in large-scale applications, where the statistician is happy to accept a hefty proportion of uninterestingly small, but nonzero, effect sizes μ_i.
- False-discovery rate hypothesis testing involves a substantial amount of estimation, blurring the line beteen the two main branches of statistical inference.

15.7 Notes and Details

The story of false-discovery rates illustrates how developments in scientific technology (microarrays in this case) can influence the progress of statistical inference. A substantial theory of simultaneous inference was developed between 1955 and 1995, mainly aimed at the frequentist control of family-wise error rates in situations involving a small number of hypothesis tests, maybe up to 20. Good references are Miller (1981) and Westfall and Young (1993).

Benjamini and Hochberg's seminal 1995 paper introduced false-discovery rates at just the right time to catch the wave of large-scale data sets, now involving thousands of simultaneous tests, generated by microarray applications. Most of the material in this chapter is taken from Efron (2010), where the empirical Bayes nature of Fdr theory is emphasized. The police data is discussed and analyzed at length in Ridgeway and MacDonald (2009).

†1 [p. 272] *Model* (15.4). Section 7.4 of Efron (2010) discusses the following result for the non-null distribution of z-values: a transformation such as (15.2) that produces a z-value (i.e., a standard normal random variable $z \sim \mathcal{N}(0,1)$) under the null hypothesis gives, to a good approximation, $z \sim \mathcal{N}(\mu, \sigma_\mu^2)$ under reasonable alternatives. For the specific situation in (15.2), Student's t with 100 degrees of freedom, $\sigma_\mu^2 \doteq 1$ as in (15.4).

†2 [p. 274] *Holm's procedure.* Methods of FWER control, including Holm's procedure, are surveyed in Chapter 3 of Efron (2010). They display a large amount of mathematical ingenuity, and provided the background against which FDR theory developed.

†3 [p. 276] *FDR control theorem.* Benjamini and Hochberg's striking control theorem (15.15) was rederived by Storey *et al.* (2004) using martingale theory. The basic idea of false discoveries, as displayed in Figure 15.2, goes back to Soric (1989).

†4 [p. 285] *Formula* (15.44). Integrating $\text{fdr}(z) = \pi_0 f_0(z)/f(z)$ gives

$$\begin{aligned} E\{\text{fdr}(z) | z \geq z_0\} &= \int_{z_0}^{\infty} \pi_0 f_0(z)\, dz \Big/ \int_{z_0}^{\infty} f(z)\, dz \\ &= \pi_0 S_0(z_0)/S(z_0) = \text{Fdr}(z_0). \end{aligned} \tag{15.54}$$

†5 [p. 286] *Thresholds for Fdr and fdr.* Suppose the survival curves $S_0(z)$ and $S_1(z)$ (15.20) satisfy the "Lehmann alternative" relationship

$$\log S_1(z) = \gamma \log S_0(z) \tag{15.55}$$

for large values of z, where γ is a positive constant less than 1. (This is a reasonable condition for the non-null density $f_1(z)$ to produce larger positive values of z than does the null density $f_0(z)$.) Differentiating (15.55) gives

$$\frac{\pi_0}{\pi_1}\frac{f_0(z)}{f_1(z)} = \frac{1}{\gamma}\frac{\pi_0}{\pi_1}\frac{S_0(z)}{S_1(z)}, \tag{15.56}$$

after some rearrangement. But $\text{fdr}(z) = \pi_0 f_0(z)/(\pi_0 f_0(z) + \pi_1 f_1(z))$ is algebraically equivalent to

$$\frac{\text{fdr}(z)}{1-\text{fdr}(z)} = \frac{\pi_0}{\pi_1}\frac{f_0(z)}{f_1(z)}, \tag{15.57}$$

and similarly for $\text{Fdr}(z)/(1-\text{Fdr}(z))$, yielding

$$\frac{\text{fdr}(z)}{1-\text{fdr}(z)} = \frac{1}{\gamma}\frac{\text{Fdr}(z)}{1-\text{Fdr}(z)}. \tag{15.58}$$

For large z, both $\text{fdr}(z)$ and $\text{Fdr}(z)$ go to zero, giving the asymptotic relationship

$$\text{fdr}(z) \doteq \text{Fdr}(z)/\gamma. \tag{15.59}$$

If $\gamma = 1/2$ for instance, $\text{fdr}(z)$ will be about twice $\text{Fdr}(z)$ where z is large. This motivates the suggested relative thresholds $\widehat{\text{fdr}}(z_i) \le 0.20$ compared with $\widehat{\text{Fdr}}(z_i) \le 0.10$.

†6 [p. 288] *Correlation effects.* The Poisson regression method used to estimate $\hat{f}(z)$ in Figure 15.5 proceeds as if the components of the N-vector of z_i values $\boldsymbol{z}$ are independent. Approximation (10.54), that the kth bin count $y_k \overset{\cdot}{\sim} \text{Poi}(\mu_k)$, requires independence. If not, it can be shown that $\text{var}(y_k)$ increases above the Poisson value μ_k as

$$\text{var}(y_k) \doteq \mu_k + \alpha^2 c_k. \tag{15.60}$$

Here c_k is a fixed constant depending on $f(z)$, while α^2 is the root mean square correlation between all pairs z_i and z_j,

$$\alpha^2 = \left[\sum_{i=1}^{N}\sum_{j\neq i}\text{cov}(z_i, z_j)^2\right] \Big/ N(N-1). \tag{15.61}$$

Estimates like $\widehat{\text{fdr}}(z)$ in Figure 15.5 remain nearly unbiased under correlation, but their sampling variability increases as a function of α. Chapters 7 and 8 of Efron (2010) discuss correlation effects in detail.

Often, α can be estimated. Let $\boldsymbol{X}$ be the 6033×50 matrix of gene expression levels measured for the control subject in the prostate study. Rows

i and j provide an unbiased estimate of $\text{cor}(z_i, z_j)^2$. Modern computation is sufficiently fast to evaluate all $N(N-1)/2$ pairs (though that isn't necessary, sampling is faster) from which estimate $\hat{\alpha}$ is obtained. It equaled 0.016 ± 0.001 for the control subjects, and 0.015 ± 0.001 for the 6033×52 matrix of the cancer patients. Correlation is not much of a worry for the prostate study, but other microarray studies show much larger $\hat{\alpha}$ values. Sections 6.4 and 8.3 of Efron (2010) discuss how correlations can undercut inferences based on the theoretical null even when it is correct for all the null cases.

†7 [p. 289] *The program* **locfdr**. Available from CRAN, this is an R program that provides fdr and Fdr estimates, using both the theoretical and empirical null distributions.

†8 [p. 289] *ML estimation of the empirical null.* Let $\mathcal{A}_0$ be the "zero set" (15.49), $\boldsymbol{z}_0$ the set of z_i observed to be in $\mathcal{A}_0$, $\mathcal{I}_0$ their indices, and N_0 the number of z_i in $\mathcal{A}_0$. Also define

$$\begin{aligned} \phi_{\delta_0,\sigma_0}(z) &= e^{-\frac{1}{2}\left(\frac{z-\delta_0}{\sigma_0}\right)^2} \Big/ \sqrt{2\pi\sigma_0^2}, \\ P(\delta_0,\sigma_0) &= \int_{\mathcal{A}_0} \phi_{\delta_0,\sigma_0}(z)\, dz \quad \text{and} \quad \theta = \pi_0 P(\delta_0,\sigma_0). \end{aligned} \tag{15.62}$$

(So $\theta = \Pr\{z_i \in \mathcal{A}_0\}$ according to (15.48)–(15.49).) Then $\boldsymbol{z}_0$ has density and likelihood

$$f_{\delta_0,\sigma_0,\pi_0}(\boldsymbol{z}_0) = \left[\binom{N}{N_0}\theta^{N_0}(1-\theta)^{N-N_0}\right]\left[\prod_{\mathcal{I}_0}\frac{\phi_{\delta_0,\sigma_0}(z_i)}{P_{\delta_0,\sigma_0}}\right], \tag{15.63}$$

the first factor being the binomial probability of seeing N_0 of the z_i in $\mathcal{A}_0$, and the second the conditional probability of those z_i falling within $\mathcal{A}_0$. The second factor is numerically maximized to give $(\hat{\delta}_0, \hat{\sigma}_0)$, while $\hat{\theta} = N_0/N$ is obtained from the first, and then $\hat{\pi}_0 = \hat{\theta}/P(\hat{\delta}_0, \hat{\sigma}_0)$. This is a partial likelihood argument, as in Section 9.4; **locfdr** centers $\mathcal{A}_0$ at the median of the N z_i values, with width about twice the interquartile range estimate of σ_0.

†9 [p. 290] *The permutation null.* An impressive amount of theoretical effort concerned the "permutation t-test": in a single-test two-sample situation, permuting the data and computing the t statistic gives, after a great many repetitions, a histogram dependably close to that of the standard t distribution; see Hoeffding (1952). This was Fisher's justification for using the standard t-test on nonnormal data.

The argument cuts both ways. Permutation methods tend to recreate the

theoretical null, even in situations like that of Figure 15.7 where it isn't appropriate. The difficulties are discussed in Section 6.5 of Efron (2010).

†10 [p. 293] *Relevance theory.* Suppose that in the **DTI** example shown in Figure 15.10 we want to consider only voxels with $x = 60$ as relevant to an observed z_i with $x_i = 60$. Now there may not be enough relevant cases to adequately estimate $\text{fdr}(z_i)$ or $\text{Fdr}(z_i)$. Section 10.1 of Efron (2010) shows how the complete-data estimates $\widehat{\text{fdr}}(z_i)$ or $\widehat{\text{Fdr}}(z_i)$ can be efficiently modified to conform to this situation.