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.

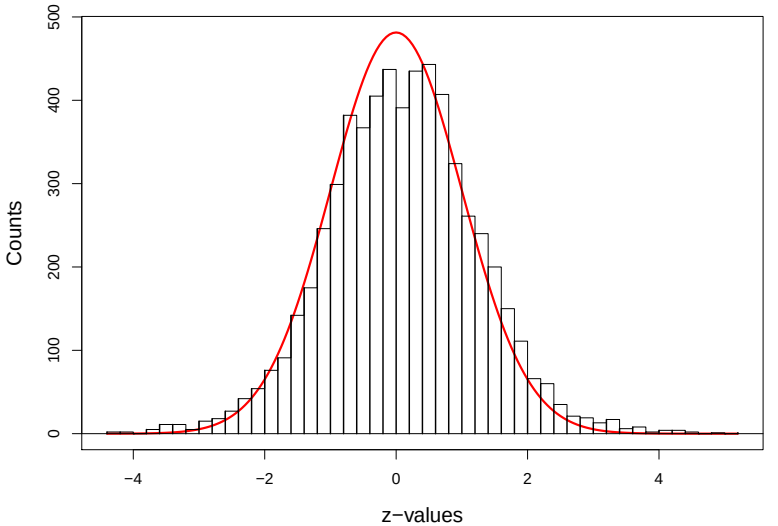

Figure 15.1 Histogram of $N = 6033$ z-values, one for each gene in the prostate cancer study. If all genes were null (15.3) the histogram would track the red curve. For which genes can we reject the null hypothesis?

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.

- 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.

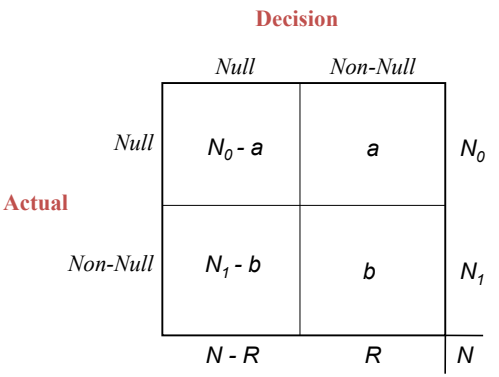

Figure 15.2 A decision rule $\mathcal{D}$ has rejected R out of N null hypotheses; a of these decisions were incorrect, i.e., they were "false discoveries," while b of them were "true discoveries." The false-discovery proportion Fdp equals a/R.

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

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

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.

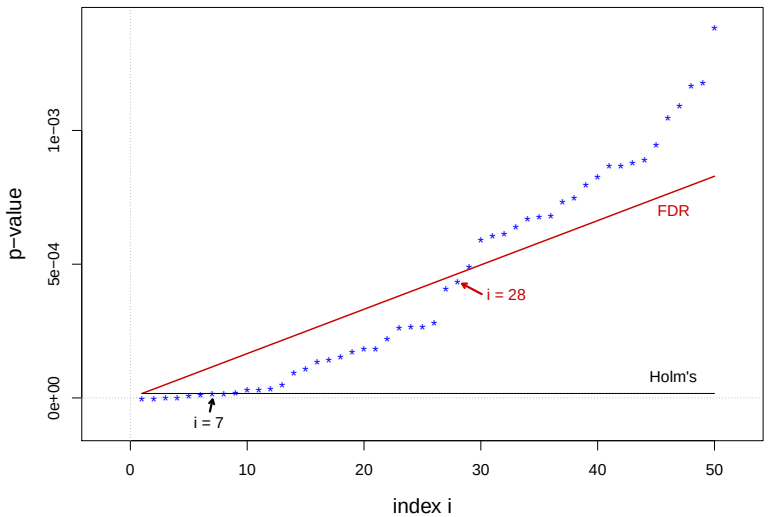

Figure 15.3 Ordered p-values $p_{(i)} = 1 - \Phi(z_{(i)})$ plotted versus i for the 50 largest z-values from the `prostate` data in Figure 15.1. The FDR control boundary (algorithm $\mathcal{D}_q$, $q = 0.1$) rejects $H_{0(i)}$ for the 28 smallest values $p_{(i)}$, while Holm's FWER procedure ($\alpha = 0.1$) 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 $f_1(z)$,