False discovery rates

Patrick Breheny

January 24

Introduction

- Last time, we saw how FWER can be used to address the question of statistical significance in light of multiple testing
- However, especially in high dimensions, FWER seems like a rather extreme condition to satisfy
- For example, in our leukemia data set, we could reject 262 hypotheses with only a 5% chance of a single false rejection among those 262 . . . seems like we could probably reject a few more and still have a lot of confidence in our results, right?

True and false discoveries

Suppose we arrange the outcomes of all the tests we conduct into a 2×2 table on the basis of our decision to reject the null hypothesis or not (known, random) and whether the null hypothesis, in reality, is true or not (fixed, unknown):

		Decision		
		"Don't reject"	"Discovery"	Total
Reality	Null true	$h_0 - A$	A	h_0
	Null false	$h_1 - B$	B	h_1
	Total	$h - R$	R	h

"Horizontal" and "vertical" rates

- Classical frequentist statistics is entirely preoccupied with the "horizontal" proportions in the previous table
  - Type I error: A/h_0
  - Power: B/h_1
- Our focus for today, however, is a "vertical" proportion:
  - False discovery proportion: A/R
- To prove anything about this proportion, we need to consider its expected value, or rate; thus, we define the *false discovery rate* as $\mathbb{E}(A/R)$, and so on for the Type I error rate, etc.

False discovery rates and high-dimensional data

- The false discovery rate has a much more direct interpretation than the Type I error rate, in that it explicitly tells what fraction of the discoveries we are claiming might simply be due to chance
- This is, of course, appealing in the low-dimensional case as well, but can't be done (at least, not outside of a Bayesian framework) for reasons that we will discuss shortly
- With high-dimensional data, however, we can estimate and control false discovery rates without the requirement of priors

Benjamini & Hochberg

- In 1995, Yoav Benjamini and Yosef Hochberg published a paper demonstrating a procedure for rejecting hypotheses in the multiple comparison setting while controlling the false discovery rate
- The procedure was not necessarily new, nor was the term "false discovery rate", but they were the first to prove that the procedure controlled the FDR
- The paper has gone on to become extraordinarily influential, with over 60,000 citations – one of the most highly cited papers in the history of statistics

The BH procedure

The Benjamini-Hochberg procedure is as follows:

- For a fixed value q, let $i_{\max}$ denote the largest index for which

$$p_{(i)} \leq \frac{i}{h} q$$

- Then reject all hypotheses $H_{0(i)}$ for $i = 1, 2, \ldots, i_{\max}$

(Note: the Holm and Westfall-Young procedures we discussed last time are "step-down" procedures; BH is a "step-up" procedure)

Background: Martingales

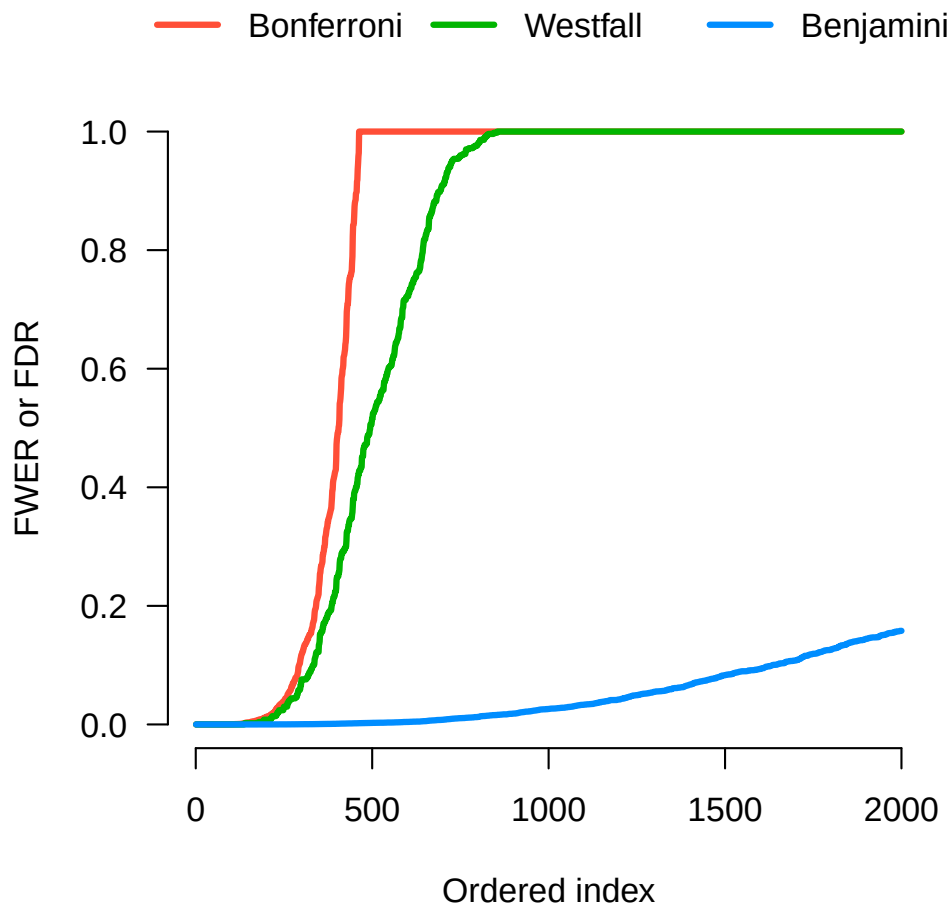
- Our proof will use martingale theory; first, a brief background
- A sequence of random variables forms a *martingale* if $\mathbb{E}(X_{n+1}|X_1, X_2, \ldots, X_n) = X_n$
- The most remarkable fact about martingales is the *optional stopping theorem*, which states that if T is a random stopping time that depends only on the past and present (i.e, on $X_1, X_2, \ldots, X_n$ but not $X_{n+1}, \ldots$), then $\mathbb{E}(X_T) = \mathbb{E}(X_1)$
- For example, suppose that the families in a population decide to have children until they reach the point where they have one more son than daughter
- One might imagine that this would skew the sex ratio in the population, but the optional stopping theorem says no, this does not happen

FDR control

- **Theorem:** For independent test statistics and for any configuration of true and false null hypotheses, the BH procedure controls the FDR at q
- Remark #1: The above theorem depends on taking A/R to be 0 when $R = 0$; typically, this is a minor concern in high dimensions, but seriously distorts the meaning of FDR for, say, $h = 1$
- Remark #2: Our proof assumed independent tests (as did Benjamini and Hochberg); later efforts have extended the results to tests that are weakly dependent

Comparison with FWER

For the leukemia data, FDR control is *much* more liberal than FWER control; at 10%, we can reject 335 hypotheses using the Westfall-Young approach, compared with 1,635 using the Benjamini-Hochberg approach



Remarks

- With FWER, we want to limit the probability of making *even a single mistake*
- With FDR, not only do we allow ourselves to make mistakes, in the leukemia case, we're allowing ourselves to make well over a hundred mistakes
- Although FDR has become a widely accepted methodology, there is no conventional standard for FDR cutoffs the way there is for p-values
- Part of the reason for this may be that FDR, being more directly interpretable, is in less need of a standard: an investigator can immediately weigh the costs of failing to reproduce the findings in 20% of discoveries vs. 5%

q-values

- As with FWER and adjusted p-values, it is often convenient to quantify the significance of each test by obtaining a value that may be simply compared with, say, .1 to find the tests that can be rejected with a FDR control of 10%
- In the FDR literature, this is known as the q *value*:

$$q_j = \inf\{q : H_{0j} \text{ rejected at } \mathrm{FDR} \leq q\}$$

- In `R`, this can be obtained with

```
p.adjust(p, method='BH')
```

although keep in mind that the interpretation of false discovery rates is very different from p-values

Fraction of null hypotheses

- In our proof of the Benjamini-Hochberg theorem, we saw that their proposed procedure was conservative: its actual FDR is

$$\mathbb{E}(A/R) = \frac{h_0}{h} q$$

- Letting $\pi_0 = h_0/h$ denote the fraction of hypotheses that are truly null, one potential improvement to the BH procedure is to estimate π_0
- Given such an estimate, we can simply replace h with $\hat{h}_0 = h\hat{\pi}_0$ everywhere it appears in the BH procedure

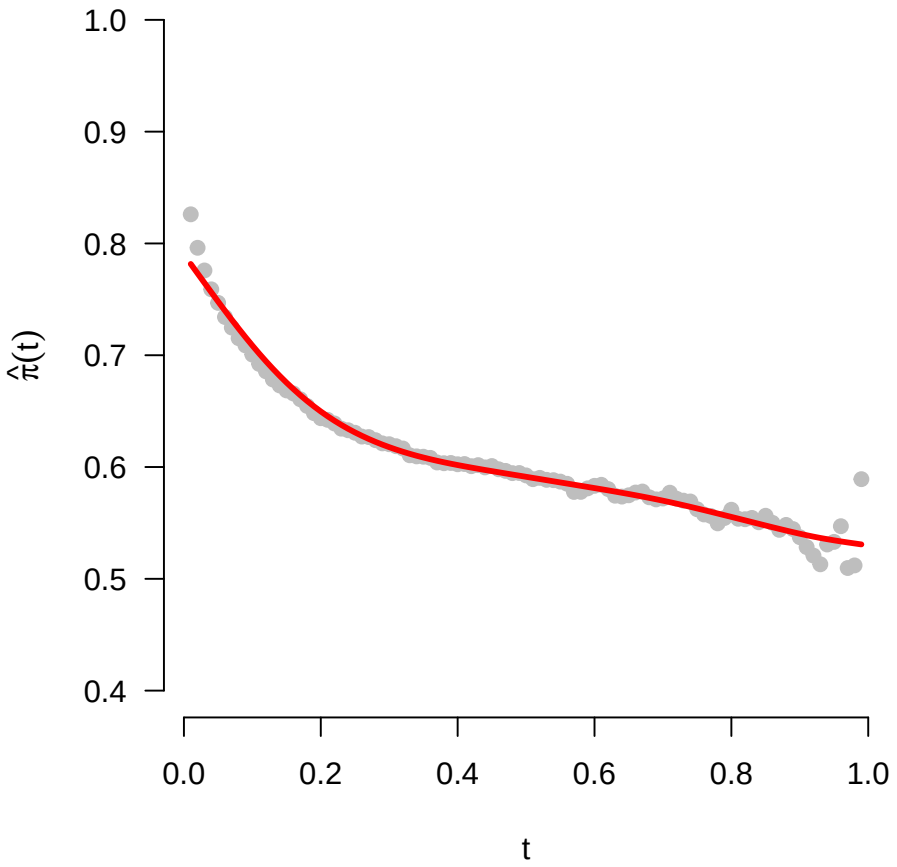
$\hat{\pi}(t)$

- Consider the following straightforward estimator for π_0, originally proposed by John Storey:

$$\hat{\pi}_0(t) = \frac{\#\{p_i > t\}}{h(1-t)}$$

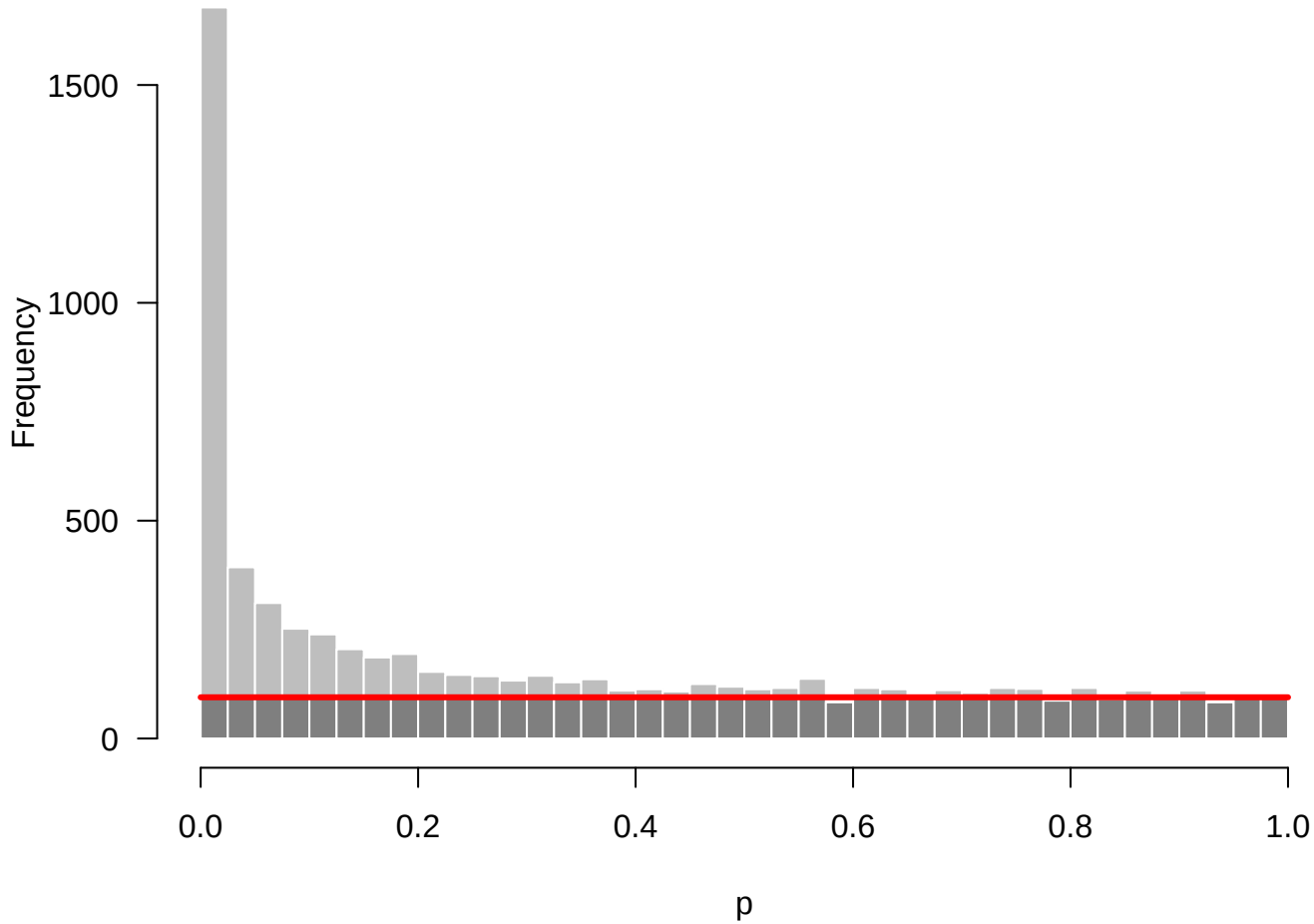
- The idea behind the estimator is that most of the high p-values should be coming from the population of null features; the estimator is simply a method-of-moments estimator under the assumption that only the null hypotheses will have p-values above t
- There is a bias-variance tradeoff at play here: for low t, we are likely including non-null hypotheses, while at high t the sample size is small

The bias-variance tradeoff



- Fitting a spline offers a way to balance this tradeoff, giving $\hat{\pi}_0 = .53$; thus, we estimate that 47% of the genes being tested differ between ALL and AML
- This idea is implemented in the package `qvalue` (on Bioconductor)

$\hat{\pi}_0$ and the p-value histogram



Empirical Bayes setup

- The preceding development of FDR has adopted a purely frequentist outlook: proposing a procedure and then proving something about its frequentist properties with respect to some error rate
- The same formula, however, can be motivated from an empirical Bayes treatment of the problem as well
- Suppose that the z-values come from a mixture of two groups: the null group with probability π_0 and density $f_0(z)$, and the non-null group with probability π_1 and density $f_1(z)$

Bayes' rule

- Consider a region $\mathcal{Z}$ and let $F_0(\mathcal{Z})$ denote the probability, for a feature in the null group, of $z \in \mathcal{Z}$, with

$$F(\mathcal{Z}) = \pi_0 F_0(\mathcal{Z}) + \pi_1 F_1(\mathcal{Z})$$

  denoting the marginal probability of $z \in \mathcal{Z}$
- Suppose we observe $z \in \mathcal{Z}$ and wish to know the group it belongs to; applying Bayes' rule,

$$\mathbb{P}(\text{Null}|z \in \mathcal{Z}) = \frac{\pi_0 F_0(\mathcal{Z})}{F(\mathcal{Z})}$$

- This requires three quantities: $F_0(\mathcal{Z}), \pi_0$, and $F(\mathcal{Z})$

Empirical distribution function

- Assuming $z \sim \mathrm{N}(0, 1)$ holds, we have $F_0(\mathcal{Z}) = \Phi(\mathcal{Z})$
- We could estimate π_0, as we have seen, or we could just use 1 as an upper bound
- Finally, since we observe a large number, h, of z-values, we can use their empirical distribution to estimate $F(\mathcal{Z})$:

$$\hat{F}(\mathcal{Z}) = \frac{\#\{z_j \in \mathcal{Z}\}}{h}$$

- Letting $\pi_0 = 1$,

$$\mathbb{P}(\mathrm{Null}|z \geq z_{(i)}) = \frac{p_{(i)}}{i/h}$$

for the ith ranked z-value; comparing this quantity to q is the same inequality checked by the BH procedure

Remarks

- Note that the FDR has a nice interpretation here: whereas in frequentist statistics, a common misconception is that $p = 0.02$ means that $\mathbb{P}(H_0|\text{Data}) = 2\%$, here the FDR actually *does* mean that (at least, in the aggregate sense)
- From the empirical Bayes perspective, the FDR methodology is not a testing procedure with error rates to be controlled, but an estimation problem
- The biggest consequence of this is with respect to correlated tests: this poses a considerable challenge to FDR control, but as an estimate remains reasonably accurate even in the presence of correlated tests

Remarks (cont'd)

- The accuracy of $\hat{\pi}_0 F_0(\mathcal{Z})/\hat{F}(\mathcal{Z})$ depends primarily on the accuracy of $\hat{F}$
- Regardless of whether the z-values are correlated or not, the empirical distribution function is an unbiased estimate of $F(\mathcal{Z})$
- However, it can have a substantial impact on the variance
- Correlated tests, therefore, introduce little bias into our FDR estimate, but diminish our confidence in its accuracy