False discovery rates

Patrick Breheny

January 27

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 131 hypotheses with only a 5% chance of a single false rejection among those 131 . . . this seems like an overwhelming success story, but FWER says we are right at the limit of what is allowed

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		
		Null	"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” proportions:
  - False discovery proportion: A/R
- To prove anything about these proportions, we need to consider their expected values, or rates; 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 we can expect to be mere coincidences
- This is, of course, appealing in the low-dimensional case as well, but it isn't possible to make claims along the lines of "there is a 95% probability the null hypothesis is true, given the data" without specifying Bayesian priors
- 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 30,000 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 $i = 1, 2, \ldots, i_{\max}$

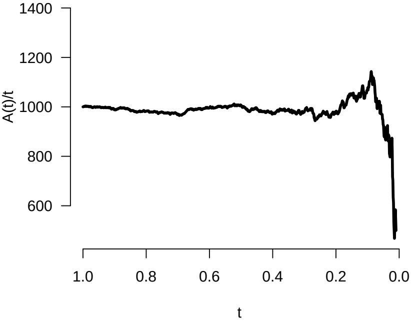
Note that, unlike the Holm and Westfall-Young procedures we discussed yesterday, this is not a step-down procedure; rather, it would be a "step-up" procedure, although that is not how I describe it above

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: The original theorem was proved only for the case of independent tests; later efforts have extended the results to tests that are weakly dependent

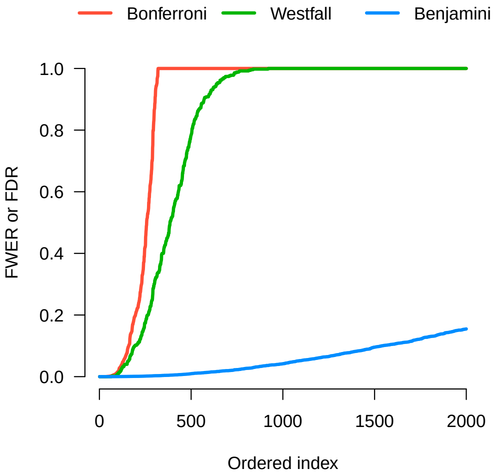
Proof: Illustration

Benjamini & Hochberg's original proof was somewhat long and tedious; a more elegant proof uses the idea of martingales and the optional stopping theorem with respect to the decision rule $p_i \leq t$



Comparison with FWER

For the leukemia data, FDR control is *much* more liberal than FWER control; at 10%, we can reject 192 hypotheses using the Westfall-Young approach, compared with 1,537 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 desirable 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

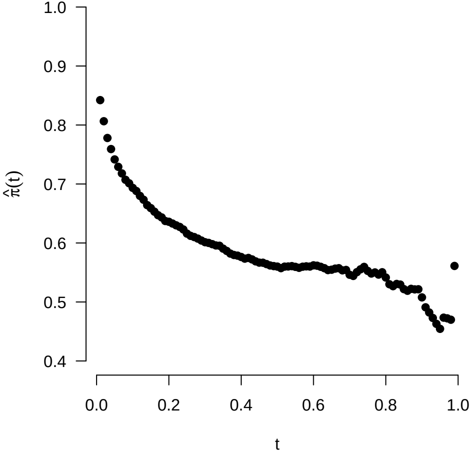
$\hat{\pi}(t)$

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

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

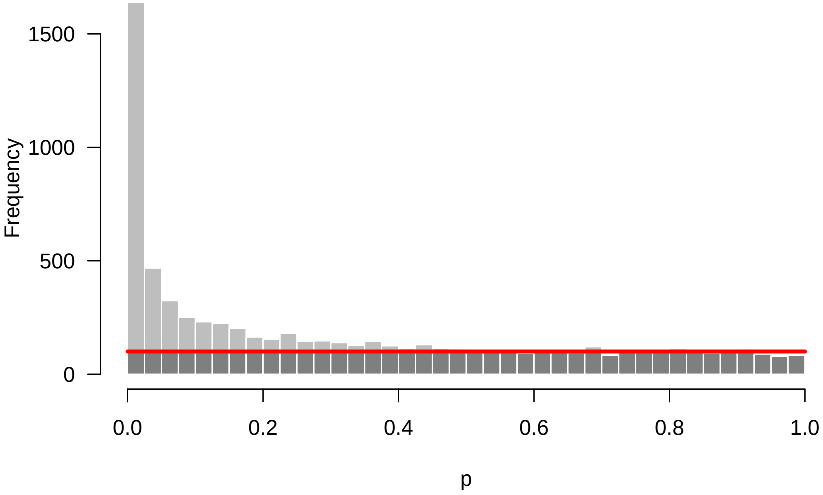
- The idea behind the estimator is that most of the high p-values should be coming from the population of null features; the estimator is simply the observed number divided by the amount you would expect in the region is all hypotheses were null
- There is a bias-variance tradeoff at play here: for low t, we are likely including non-null hypotheses, while at high t the sample size is small

The bias-variance tradeoff



Somewhere around $t = 0.6$ seems reasonable, with $\hat{\pi}(0.6) = .56$; thus, we estimate that 44% of the genes being tested differ between ALL and AML

$\hat{\pi}_0$ 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 estimator, 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 we believe in the theoretical null, $F(\mathcal{Z}) = \Phi(\mathcal{Z})$
- 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 $F(\mathcal{Z})$:

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

- Substituting, we have that for the ith ranked z-value,

$$\mathbb{P}(\text{Null}|z \in \mathcal{Z}) = \frac{p_{(i)}}{i/h},$$

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 $\hat{F}$
- Correlation among the z-values introduces little or no bias to the empirical distribution function as an estimate of $F(\mathcal{Z})$
- However, it can have a substantial impact on the variance
- This insight offers the clearest picture of how dependence between tests affects FDR: the estimate remains essentially unbiased, but our confidence in its accuracy is diminished