Family-wise error rates

Patrick Breheny

January 27

Introduction

- We will begin by discussing the topic of high-dimensional data from a multiple testing perspective
- The basic issue is this: a p-value of 0.03 has a certain interpretation when we test a single hypothesis – we would tend to think of this as significant evidence
- But what if we've tested 100 or 1,000 hypotheses?
- We will explore three fundamentally different answers to that question in the coming lectures: family-wise error rates, false discovery rates, and local false discovery rates
- Note: The "Large scale testing" portion of the course will not use our textbook; material is based in part on *Large-Scale Inference*, by Bradley Efron

Leukemia data

- To illustrate these ideas, we will use data from one of the earliest and most well-known high-dimensional studies: a gene expression study of leukemia patients
- The study used a technology called a microarray to measure the expression of $7,129$ genes for 72 patients
- Of the 72 patients,
  - 47 patients had acute lymphoblastic leukemia (ALL)
  - 25 patients had acute myeloid leukemia (AML)

  Of the two diseases, AML has a considerably worse prognosis: only 26% survive at least 5 years following diagnosis, compared to 68% for ALL

Analysis goals

- The analysis could be approached from one of two perspectives:
  - Testing whether the expression of each gene differs between the two types of cancer, in the hopes of identifying genes that may be affected differently by the two diseases
  - Using the gene expression data to explain/predict the type of cancer
- For this unit, we are focusing on the first goal; for most of the rest of the course, we will focus on the second

Data format

I will make the data sets for this course available online in the following format:

- All data sets will be saved R objects in the `.rds` format; use `readRDS()` to read them into R
- Each data set will contain (at least) two objects:
    - `y`, a vector (here, the disease status); in regression problems, this would be the response, or outcome
    - `X`, a matrix (here, the gene expression data) with the same number of rows as `y` has elements, and many columns

p-values

- For the leukemia data, let's carry out 7,129 two-sample t-tests, obtaining the set of p-values $\{p_j\}_{j=1}^{7{,}129}$
- A critical property of p-values is that for any value u,

$$\mathbb{P}_0\{P \le u\} \le u,$$

  where P is the p-value and $\mathbb{P}_0$ denotes the probability under the null hypothesis; note that P is a random variable here in the sense that it depends on the data
- For a continuous null distribution, we have

$$P \sim \text{Unif}(0,1)$$

  under the null hypothesis

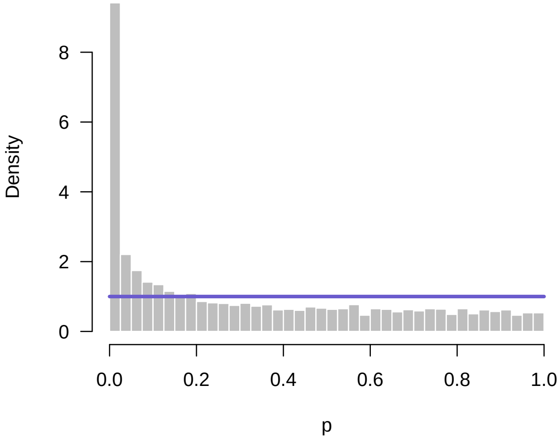
z-values

- Sometimes, it is more useful to work with z-values than p-values:

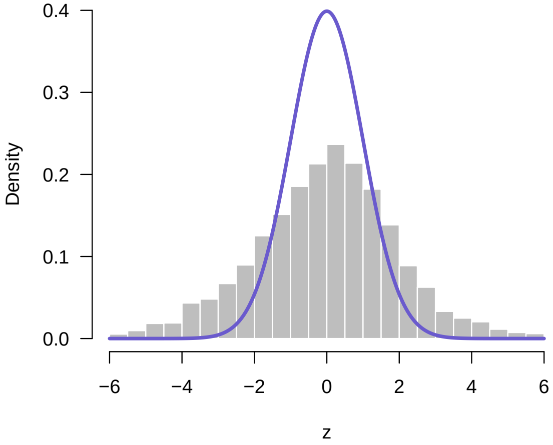
$$\begin{aligned} z_j &= \Phi^{-1}(p_j) && \text{(one-sided)}, \\ z_j &= -s_j \Phi^{-1}(p_j/2) && \text{(two-sided)} \end{aligned}$$

  where Φ^{-1} is the inverse of the standard normal CDF and s_j is the sign of the jth test
- Under H_0, $Z \sim \mathrm{N}(0, 1)$
- One advantage of z-values for two-tailed tests is that they retain the sign information; in the present context, the z-value tells us whether expression was higher in ALL or AML patients, while the p-value does not

p-values: Leukemia data



z-values: Leukemia data



FWER

- The *family-wise error rate* (FWER) is defined as the probability of making at least one false rejection in a family of hypothesis-testing problems
- A *FWER control procedure* is a method for taking a set of p-values and deciding which null hypotheses to reject, subject to the requirement that FWER $\leq \alpha$
- FWER control was the first rigorous approach to assessing significance in the presence of multiple comparisons

Bonferroni correction

- The simplest and most well-known FWER control procedure is the *Bonferroni correction*, in which we reject all hypotheses for which

$$p_j \leq \alpha/h,$$

  where h is the number of hypotheses being tested
- **Theorem:** The Bonferroni correction controls the FWER at level α
- Note that the above theorem makes no assumptions concerning independence between tests; it is valid for any dependence among the h tests

Adjusted p-values

- Another way of thinking about FWER control procedures is in terms of *adjusted p-values*
- The adjusted p-value for hypothesis j is defined as

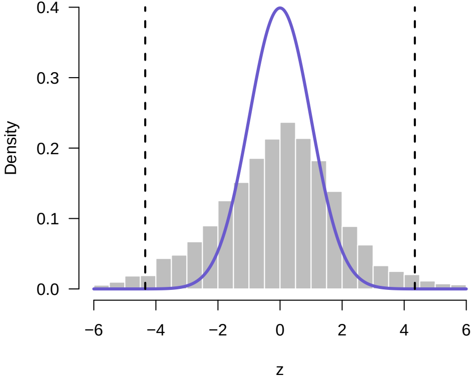
$$\tilde{p}_j = \inf\{\alpha : H_{0j} \text{ rejected at } \mathrm{FWER} \leq \alpha\}$$

- For the Bonferroni correction,

$$\tilde{p}_j = hp_j;$$

by convention, with an upper bound of 1

FWER for leukemia study



- 7,129 hypothesis tests
- 2,071 have $p_j \leq .05$
- 260 have $\tilde{p}_j \leq .05$ using the Bonferroni approach

Bonferroni: Too conservative?

- One concern with the Bonferroni approach is that the upper bound it provides may be loose; could it be improved upon?
- For example, if we knew the number of true null hypotheses, we could divide by that number instead of h
- In a sense, this is the motivation behind a clever modification of the Bonferroni approach proposed by Sture Holm

Holm procedure

Letting $p_{(1)}, p_{(2)}, \ldots, p_{(h)}$ denote the p-values, sorted from smallest to largest, the Holm procedure is as follows:

(1) Compare $p_{(1)}$ to α/h; if $p_{(1)} > \alpha/h$, do not reject any hypotheses; if $p_{(1)} \leq \alpha/h$, reject the corresponding hypothesis and move on to $p_{(2)}$

(2) Compare $p_{(2)}$ to $\alpha/(h-1)$; if $p_{(2)} > \alpha/(h-1)$, do not reject any additional hypotheses; if $p_{(2)} \leq \alpha/(h-1)$, reject the corresponding hypothesis and move on to $p_{(3)}$

(3) Continue in this manner until no more hypotheses can be rejected

Properties and remarks

- **Theorem:** The Holm procedure controls the FWER at level α
- As with the Bonferroni approach, note that we have made no assumptions regarding dependence between tests
- Note that the Holm procedure is always more powerful than the Bonferroni procedure, since

$$\frac{\alpha}{h-j+1} \geq \frac{\alpha}{h} \qquad \text{for all } j$$

- The Holm procedure is known as a *step down* procedure; there are a variety of other stepwise approaches to FWER control

R code

- The Bonferroni and Holm procedures are both implemented (along with many others) in the R function `p.adjust`:

```
p.adjust(p, method='bonferroni')
p.adjust(p, method='holm')          # Default
```

- The above code returns the adjusted p-values; by comparing $\tilde{p}$ to α, we determine which hypotheses may be rejected at FWER α

Leukemia results

- For the Leukemia data, at FWER 0.05:
  - 260 genes are declared significant using the Bonferroni approach
  - 262 genes are declared significant using the Holm approach
- These results are typical: the Holm approach is more powerful than the Bonferroni approach, but the difference is not as dramatic as you might imagine

Motivation

- The appeal of the Holm and Bonferroni approach is that they work for any dependency structure among the hypotheses
- The disadvantage, however, is that for many types of dependence, we can achieve better bounds on the FWER if we use this information
- So, let's cover one more FWER control procedure, proposed by Westfall and Young, who use a permutation-based approach to preserve the dependency among the features

Permutation tests

- To briefly review the general idea of permutation tests, suppose we observe values $\{x_i\}_{i=1}^{n_1}$ in group 1 and $\{x_i\}_{i=n_1+1}^{n_1+n_2}$ in group 2
- Under the null hypothesis, these values all come from the same distribution and any partition of the x values into sets of size n_1 and n_2 should be equally likely
- We can therefore carry out a test by randomly permuting the x values, calculating the test statistic $T(\mathbf{x})$, and calculating the fraction of random permutations that are less than the observed value of $T(\mathbf{x})$
- Ideally, we would do this for all possible permutations, but unless the sample size is small, this is not feasible from a computational perspective

Westfall-Young procedure

- This is the basic idea of the Westfall-Young procedure: permute the class labels $\mathbf{y}$, then reapply the test in question
- Doing this a large number of times allows us to estimate

$$\pi(j) = \mathbb{P}_0\Big\{ \min_{k \in R_j} P_k \leq p_{(j)} \Big\},$$

  where $R_j = \{k : p_k \geq p_{(j)}\}$
- The adjusted p-value is then

$$\tilde{p}_{(i)} = \max_{j \leq i} \hat{\pi}(j),$$

  where $\hat{\pi}$ is the empirical mean over all the permutations

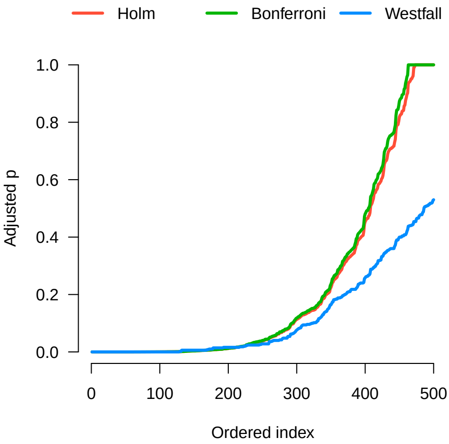
Remarks

- The main idea is that by permuting $\mathbf{y}$, we force independence between $\mathbf{y}$ and $\mathbf{x}_j$ for all j; i.e., we force the *complete null hypothesis* to be true
- However, by keeping the rows of $\mathbf{X}$ intact, we preserve the correlation structure between the features (here, genes)
- It is reasonably clear, then, that the Westfall-Young procedure controls FWER in the *weak* sense: if all the null hypotheses are true

Strong vs. weak control

- *Strong* control of the FWER means that the FWER is bounded by α regardless of which null hypotheses are true and which are false
- Strong control is obviously more desirable, but harder to demonstrate, at least without added assumptions
- In the case of the Westfall-Young procedure, to prove strong FWER control, we require an assumption of *subset pivotality*: that the vector $(P_i : H_{0i} \text{ true})$ always follows the same distribution

Leukemia data: Comparison



The Westfall-Young procedure allows us to identify 291 differentially expressed genes at a FWER of 5% (compared to 260 for Bonferroni)