Family-wise error rates

Patrick Breheny

January 25
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 does it mean when 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.

This portion of the course will not use our textbook, but instead Bradley Efron’s *Large-Scale Inference*.
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 microarrays 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.
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
I will make the data sets for this course available online in the following format.

- All data sets will be saved R objects (i.e., .RData files).
- 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.
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 \( u \),

\[
P_0\{P \leq u\} \leq u,
\]

where \( P \) is the \( p \)-value and \( P_0 \) denotes the probability under the null hypothesis; note that \( P \) is the random variable here in the sense that it depends on the data.

Thus, for any continuous null distribution,

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

under the null hypothesis.
Sometimes, it is more useful to work with $z$-values than $p$-values:

$$z_j = \Phi^{-1}(p_j),$$

where $\Phi^{-1}$ is the inverse of the standard normal CDF.

- Under $H_0$, $Z \sim N(0, 1)$

- One advantage of $z$-values for two-tailed tests 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
- 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 $\text{FWER} \leq \alpha$.

- FWER control was the first rigorous approach to assessing significance in the presence of multiple comparisons.
The simplest and most well-known FWER control procedure is the *Bonferroni correction*, in which we reject all hypotheses for which

\[ p_j \leq \frac{\alpha}{h}, \]

where \( h \) is the number of hypotheses being tested.

**Theorem:** The Bonferroni correction controls the FWER at level \( \alpha \).

Note that the above theorem makes no assumptions concerning independence between tests; it is valid for any dependence among the \( h \) tests.
Another way of thinking about FWER control procedures is in terms of \textit{adjusted} \( p \)-\textit{values}.

The adjusted \( p \)-value for hypothesis \( j \) is defined as

\[ \tilde{p}_j = \inf \{ \alpha : H_0^j \text{ rejected at FWER} \leq \alpha \} \]

For the Bonferroni correction,

\[ \tilde{p}_j = hp_i; \]

by convention, with an upper bound of 1.
FWER for leukemia study

- 7129 hypothesis tests
- 2106 have \( p_j \leq .05 \)
- 130 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?
- In particular, 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 $\alpha/h$; 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 $\alpha$
- 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

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

- The Holm procedure is known as a *step down* procedure; there are a variety of other stepwise approaches to FWER control
The Bonferroni and Holm procedures are both implemented (along with many others) in the R function `p.adjust`:

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

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

- For the Leukemia data, at FWER 0.05,
  - 130 genes are declared significant using the Bonferroni approach
  - 131 genes are declared significant using the Holm approach
- These results are not atypical: the Holm approach is more powerful than the Bonferroni approach, but the difference is not as dramatic as you might imagine
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.
The basic idea of the Westfall-Young procedure is to permute the class labels \( y \), then reapply the test in question. Doing this a large number of times allows us to estimate

\[
\pi(j) = \mathbb{P}_0 \left\{ \min_{k \in R_j} P_k \leq p(j) \right\},
\]

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.
The main idea is that by permuting $y$, we force independence between $y$ and $x_j$ for all $j$; i.e., we force the \textit{complete null hypothesis} to be true.

However, by keeping the rows of $X$ 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 \textit{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 $\alpha$ 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.
The Westfall-Young procedure allows us to identify 153 differentially expressed genes at a FWER of 5%.