Family-Wise Error Rate & False Discovery Rate

Collin Nolte

March 22, 2022

Review

Up until now, we have concerned ourselves with performing a single statistical test on a collection of observed data

This has involved finding a test statistic, creating confidence intervals, and then computing an associated *p*-value, the probability of observing the data (or something more extreme) under the null hypothesis

Our conclusions have centered around controlling the Type I error rate, typically at $\alpha = 0.05$, where we have rejected nulls with *p*-values below the specified α

Types of Errors

Recall from a previous lecture that there were two ways in which we might make an error

Test Result	True State of Nature	
	H_0 True	H_0 False
Fail to reject H_0	Correct $(1-\alpha)$	Incorrect Type II Error (β)
Reject H_0	Incorrect Type I Error (α)	Correct $(1-\beta)$

Our goals in statistical inference are simultaneously to control the Type I error rate with as much power $(1-\beta)$ as possible

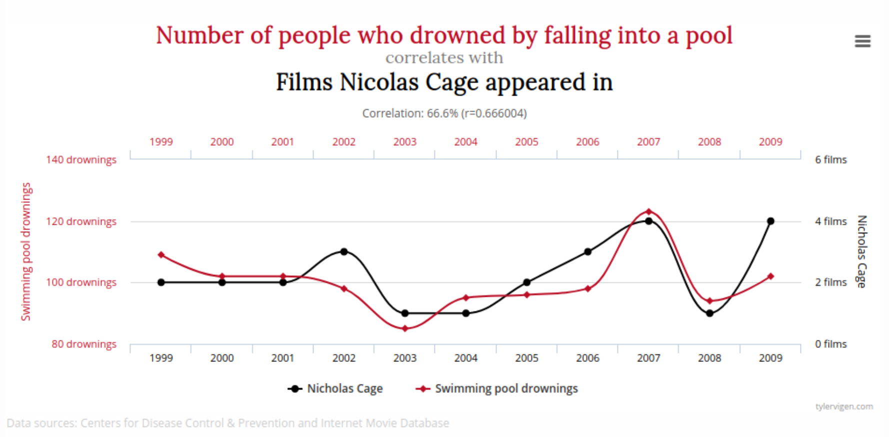
Multiple Comparisons

The issue of multiple comparisons arises when a set of statistical inferences are considered simultaneously (relevant xkcd)

There are two main flavors of this phenomenon. Most relevant in genetics is the case of having multiple (possibly independent) tests, such as considering differential expression between genes on a microarray

Also common is the case of multiple testing. Reachers may consider a multitude of similar models for explaining an outcome while failing to account for the fact that effectively, multiple tests were done

A more odious example of the latter is known as *data dredging*: a motivated researcher can often employ any number of statistical tests until finding something justifying their own hypotheses



http://www.tylervigen.com/spurious-correlations

FWER

Here, we will primarily focus on the case of conducing multiple hypothesis tests at once – for example, performing a t-test between all SNPs studied by genomic position

If we performed tests on 10,000 *independent* SNPs, the Type I error rate for any particular SNP would be 0.05. However, the probability of making 0 errors for all $10,000$ SNPs would be $(1 - 0.05)^{10,000} = 1.722 \times 10^{-223}$

When controlling the error rate for a group of tests, we are instead going to be interested in controlling the *family-wise error rate (FWER)*, which is the probability of making one or more false discoveries

Classification of Tests

	Null is True	Alternative is True	Total
Test significant	V	S	R
Test non-significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

- m is total test
- m_0 is the number of *true* null hypotheses (unknown)
- $m - m_0$ is the number of true alternative hypotheses
- V is the number of false positives (Type I error)
- S is the number of true positives
- U is the number of true negatives
- T is the number of false negatives (Type II error)
- $R = V + S$ is the number of rejected nulls (discoveries, true or false)

FWER

In other words, we might define the FWER as

$$FWER = P(V \geq 1)$$

That is, if we control FWER at $\alpha = 0.05$, the probability of making a single false discovery is controlled at 5%

There are a number of FWER techniques available, each slightly different in terms of assumptions and power

Bonferroni Adjustment

The simplest of the FWER procedures is the *Bonferroni correction*.

Assume that we have m total tests, with hypotheses H_i for $i = 1, \dots, m$. We will reject H_i if

$$p_i \leq \frac{\alpha}{m}$$

While this will effectively control the FWER at $\alpha = 0.05$, it will severely reduce the power; in the case with 10,000 SNPs, a single test would be considered significant only if it has a p-value of

$$p \leq 0.05/10,000 = 5 \times 10^{-6}$$

If the number of tests were to increase (often the case in genomic studies), this value would become prohibitively small, potentially resulting in us failing to reject a case in which the alternative hypothesis were true (loss of power)

Bonferroni

Let $\alpha = 0.05$ and

$$p_1 = 0.001 \quad p_2 = 0.01 \quad p_3 = 0.04 \quad p_4 = 0.05 \quad p_5 = 0.1.$$

As we have $m = 5$ tests, our Bonferroni adjustment gives us a cutoff of $\alpha/m = 0.01$

From this, we see that controlling FWER at $\alpha = 0.05$, we would reject p_1 and p_2 while failing to reject the rest

Holm's Procedure

For illustration, we can also consider Holm's procedure. For m tests, we rank the p-values from smallest to largest, $p_{(1)}, \dots, p_{(m)}$

For a given α, we then consider for each p-value, $p_{(k)}$, the expression

$$p_{(k)} > \frac{\alpha}{m - k + 1}$$

For the smallest k for which this is true, we then reject the null hypotheses $H_{(1)}, \dots, H_{(k-1)}$

Holm's Procedure

Let $\alpha = 0.05$ and

$$p_{(1)} = 0.001 \quad p_{(2)} = 0.01 \quad p_{(3)} = 0.04 \quad p_{(4)} = 0.05 \quad p_{(5)} = 0.1.$$

Then

1. $p_{(1)} = 0.001 > \frac{0.05}{5-1+1} = 0.01$ ✗
2. $p_{(2)} = 0.01 > \frac{0.05}{5-2+1} = 0.0125$ ✗
3. $p_{(3)} = 0.04 > \frac{0.05}{5-3+1} = 0.016$ ✓

As $k = 3$ was the smallest value for which this inequality holds, we would *reject* $H_{(1)}$ and $H_{(2)}$ with the FWER controlled at $\alpha = 0.05$. Note that although the cutoffs changed, the same hypotheses were rejected in both cases

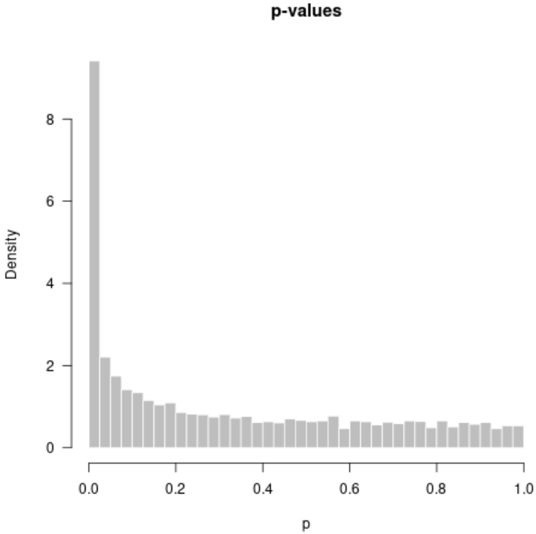
Example

To see this in action, we will consider data from a high-dimensional study on leukemia patients.

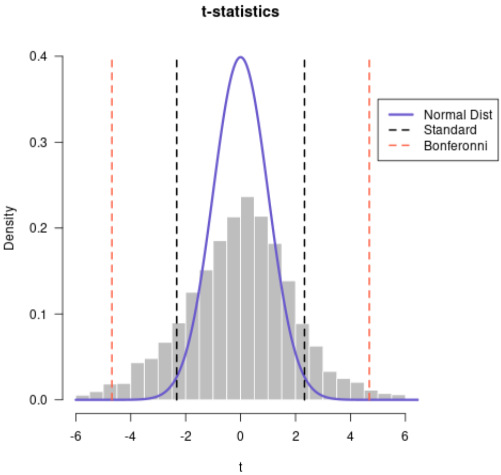
This studied used a microarray to measure the expression of 7,129 genes in 72 patients. Of these, 47 patients had acute lymphoblastic leukemia (ALL) and 25 had acute myeloid leukemia (AML). Of the two, AML is considerably worse, with a 26% 5-year survival rate, compared with 66% for ALL

Our interest here will be to test whether the expression of each gene differs between the two types of cancer

p-values from t-test



test statistics



- 7,129 hypothesis tests
- 2,071 genes with $p_j \leq 0.05$
- 260 genes considered significant with Bonferonni correction
- 262 genes significant with Holm adjustment

FWER vs FDR

As previously alluded to, controlling the FWER can be prohibitive when the number of tests performed is large, as is often the case with genomic data

A major consequence of this is a loss of power – while we may protect against false discoveries, we drastically limit our ability to discover cases when the alternative hypothesis is true

One alternative to controlling FWER is to instead focus on the *false discovery rate (FDR)* which has greater power at the price of increased Type I errors

False Discovery Rate

The FDR works to control Type I errors by controlling the proportion of false positives to the total number of positives

	Null is True	Alternative is True	Total
Test significant	V	S	R
Test non-significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

Where R is the total number of significant tests in a family of tests, we define the false discovery rate as

$$FDR = \frac{\#\text{ of false positives}}{\#\text{ of significant tests}} = \frac{V}{V+S} = \frac{V}{R}$$

FDR

The goal of the FDR is to provide a reasonable balance between the number of false positives and true positives. That is, in a clinical setting, it may prove more fruitful to weed through a larger collection of positives than to conjure relevant genes from an empty list

The false discovery rate was introduced by Yoav Benjamini and Yosef Hochberg in 1995, and is one of the most widely cited publications in statistics with over 50,000 citations

Motivating examples from Storey & Tibshirani (2003) include

- Microarray experiments – detection of differential gene expression
- Idenfitication of exonic splicing enhancers
- Genetic dissection of transcriptional regulation
- Finding binding sites of transcriptional regulators

Estimate of FDR

We can approach the task of using FDR in two separate ways. The first involves using our specified level of α, and determining of these, what percentage we expect to be false. Here, m is the total number of tests, and R is the total number of tests found significant

$$FDR = \frac{m \times \alpha}{R}$$

For our leukemia example, we had 7,129 genes, with 2,071 having $p_j \leq 0.05$. Our estimate of FDR here would be

$$\widehat{FDR} = \frac{7,129 \times 0.05}{2,071} = 0.172$$

That is, of the 2,071 tests considered significant at $\alpha = 0.05$, we would expect about 17% to be false positives

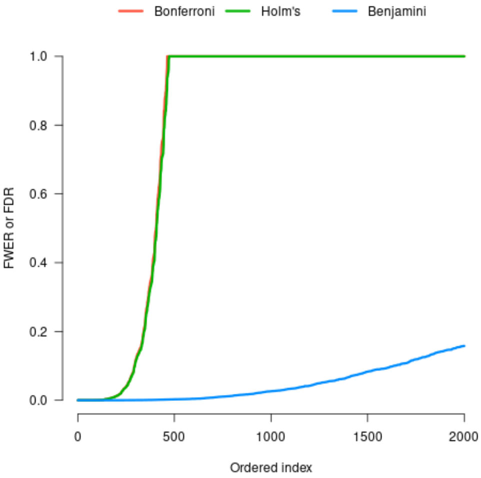
q-values

As an alternative to estimating the FDR for a given α, Storey & Tibshirani (2003) proposed the use of the *q*-value, analogous to the *p*-value, but instead representing significance in terms of FDR.

The interpretation here is a little different: whereas the *p* value gives the probability of observing the data or something greater, given the null, the *q* value of a test gives the expected proportion of false positives for rejecting all hypotheses with an equal or smaller *q*-value

Typical FDR cutoff values range from 5% to 20%, giving the expected proportion of positives to be false

FWER vs FDR



- 7,129 hypothesis tests
- 2,071 genes with $p_j \leq 0.05$
- 260 genes considered significant with Bonferonni correction
- 262 genes significant with Holm adjustment
- 1,635 using Benjamini-Hochberg at FDR of 10% (of which 1472 are expected to be true positives)

FDR Review

To reiterate, we have two methods here in which we might use FDR.

The first involves using a specified α, and then determining an estimate of the FDR

The second involves transforming *p*-values to *q*-values, and then selecting all tests with *q*-values below the specified FDR threshold.

However, both are describing the same thing: had we used an FDR cutoff of 17.211% for Benjamini-Hochberg on the previous slide, we would have found 2,074 genes to be significant. Here, the 17.211% was our estimate of FDR based on $\alpha = 0.05$ (slide 17)

Conclusions

Both approaches considered today seek to address the issue with multiple testing, a particularly relevant problem in genomic studies

Whereas the FWER seeks to limit the possibility of making a single false discovery, the FDR assumes they exist by definition, with the previous example allowing nearly 200 false discoveries to exist

It will often be on investigators to weigh the pros and cons of each approach, though for high dimensional studies, the false discovery rate is most commonly employed. What remains to be determined is the acceptable number of Type I errors

References

- Patrick Breheny High-Dimensional Data Analysis (BIOS 7240) lecture notes
- Deborah Dawson GENE 6234 lecture notes
- Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, and Lander ES (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286:531-537.
- Storey, John D. and Tibshirani, Robert (2003). Statistical significance for genomewide studies. National Academy of Sciences, 10.1073/pnas.1530509100
- Benjamini, Yoav and Hochberg, Yosef (1995). Controlling the False Discovery Rate: A Practice and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society, 57:289-300