

Lecture 10: Multiple Hypothesis Testing & Permutation Tests

May 31, 2012
 GENOME 560, Spring 2012

Su-In Lee, CSE & GS
 suinlee@uw.edu

Goals

- Multiple hypothesis testing
 - Controlling positive FDR
- Permutation tests
 - Paired tests
 - Linear regression
- R instruction
 - Performing permutation tests

Review: Controlling False Positives

- When we say "adjusting p-values for the number of hypothesis tests performed", what we mean is **controlling Type I error rate**

	Null True	Alternative True	Total
Not Called Significant	U	T	$m-R$
Called Significant	V	S	R
	m_0	$m - m_0$	m

- V = # Type I errors [false positives]
- Many procedures have been developed to control the **Family-Wise Error Rate** (the probability of at least one Type I error):
 $P(V \geq 1)$

Review: False Discovery Rate

	Null True	Alternative True	Total
Not Called Significant	U	T	$m-R$
Called Significant	V	S	R
	m_0	$m - m_0$	m

- V = # Type I errors [false positives]
- False discovery rate (FDR) is designed to control the proportion of false positives **among the set of rejected hypotheses** $(R) - V/R$

What If R = 0 ?

- Benjamini & Hochberg:

$$\text{FDR} = E\left[\frac{V}{R} \mid R > 0\right]P(R > 0)$$

- “the rate that false discoveries occur”

- Storey:

$$\text{pFDR} = E\left[\frac{V}{R} \mid R > 0\right]$$

- “the rate that discoveries are false”

5

Storey's Positive FDR (pFDR)

$$\text{BH: FDR} = E\left[\frac{V}{R} \mid R > 0\right]P(R > 0)$$

$$\text{Storey: pFDR} = E\left[\frac{V}{R} \mid R > 0\right]$$

- Since $P(R > 0)$ is ~ 1 in most genomics experiments FDR and pFDR are very similar
- Omitting $P(R > 0)$ facilitates development of a measure of significance in terms of the FDR for each hypothesis

6

FDR in Bayesian terms

- **Theorem:** m identical hypothesis tests are performed with independent statistics T_1, \dots, T_m and rejection area C . A null hypothesis is true with a priori probability $\pi_0 = P(H_0 \text{ is true})$. Then

$$\begin{aligned} \text{pFDR}(C) &= P(H_0 \text{ is true} \mid T \in C) \\ &= \frac{\pi_0 P(T \in C \mid H_0 \text{ is true})}{P(T \in C)} \end{aligned}$$

7

What is a q-value?

- **Definition:** For an observed statistic $T = t$, define the q-value of t to be

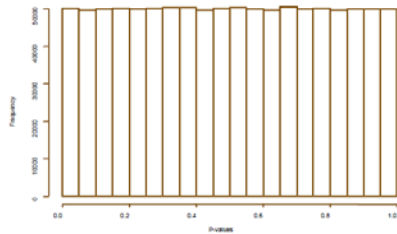
$$q\text{-value}(t) = \min_{\{C: t \in C\}} \text{pFDR}(C)$$

- minimum FDR that can be attained when calling that “feature” significant
- i.e., expected proportion of false positives incurred when calling that feature significant
- The estimated q-value is a function of the p-value for that test and the distribution of the entire set of p-values from the family of tests being considered (Storey and Tibshirani 2003)
- Thus, in an array study testing for differential expression, if gene X has a q-value of 0.013 it means that 1.3% of genes that show p-values at least as small as gene X are false positives

8

Estimating The Proportion of Truly Null Tests

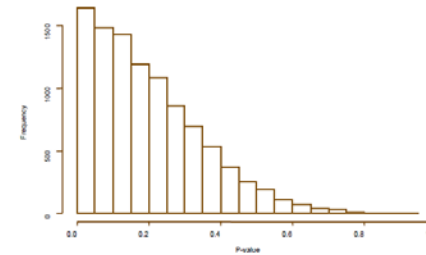
- Under the null hypothesis p-values are expected to be uniformly distributed between 0 and 1



9

Estimating The Proportion of Truly Null Tests

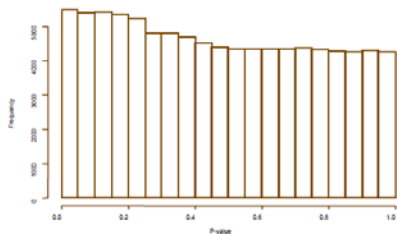
- Under the alternative hypothesis p-values are skewed towards 0



10

Estimating The Proportion of Truly Null Tests

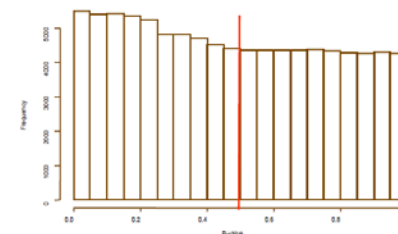
- Combined distribution is a mixture of p-values from the null and alternative hypotheses



11

Estimating The Proportion of Truly Null Tests

- For p-values greater than say 0.5, we can assume they mostly represent observations from the null hypothesis

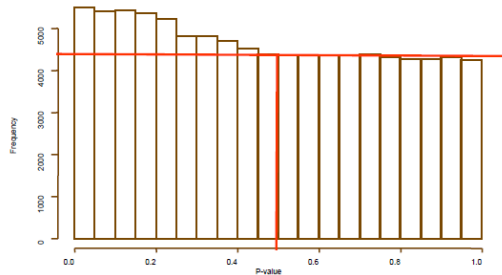


12

Definition of π_0

- The proportion of truly null tests:

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1, 2, \dots, m\}}{m(1 - \lambda)}$$



13

Digression: p-Values

- Implicit in all multiple testing procedures is the assumption that the distribution of p-values is “correct”
- This assumption often is not valid for genomics data where p-values are obtained by asymptotic theory
- Thus, resampling methods are often used to calculate p-values

14

Permutation Tests

- Consider a set of data points in two samples (groups)

$$x_1, x_2, x_3, \dots, x_m \quad y_1, y_2, y_3, \dots, y_n$$

- Under the null hypothesis, any of the $(m+n)$ points could have been in any of the samples
- So, all permutations of the points (shuffling them among samples) are equally likely

$$x_1, y_{11}, y_3, x_4, \dots, x_{m-3}, x_m \quad y_1, x_3, y_2, \dots, y_{n-3}, y_n$$

- Does our sample show more difference than expected, among all these shuffles?

15

Permutation Tests


$$x_1, x_2, x_3, \dots, x_m \quad y_1, y_2, y_3, \dots, y_n$$

- Here is how we test:
 - Compare the difference of means (or some other reasonable statistic) between the two groups
 - Make a large number of random shufflings of the points
 - For each, compute this statistic (means)
 - See whether, out of say 9,999 shuffles, when the true value is added in, it is in the top 5% of these 10,000.
- Note that this test does not assume normality, just that the **points are drawn from the same (unknown) distribution, independently**

16

Permutation Tests: Paired Tests


- There are many variations on permutation tests
 - If the test is a paired test, to see whether the mean difference is zero, shuffle within each pair (i.e. flip each pair the other way with probability 50%)

$x_1, x_2, x_3, \dots, x_n$ Before drug treatment
 $y_1, y_2, y_3, \dots, y_n$ After drug treatment

 d_1, d_2, \dots, d_n After drug treatment
 Difference $d_i = x_i - y_i$

17

Permutation Tests: Paired Tests

- There are many variations on permutation tests
 - If the test is a paired test, to see whether the mean difference is zero, shuffle within each pair (i.e. flip each pair the other way with probability 50%)

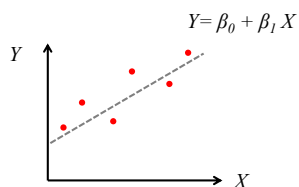
$x_1, x_2, x_3, \dots, x_n$ Before drug treatment
 $y_{11}, y_{30}, y_7, \dots, y_2$ After drug treatment (shuffled)
 Call them z's

 f_1, f_2, \dots, f_n Difference $f_i = x_i - z_i$

18

Permutation Tests: Linear Regression

- There are many variations on permutation tests
 - If it is a regression, and if the Y points are randomly associated with the X points under the null hypothesis, so that the true slope is zero, we can shuffle Y s, associating them with the X s at random. Each time, we compute the slope

x_1, x_2, x_3, x_4, x_5
 y_1, y_2, y_3, y_4, y_5

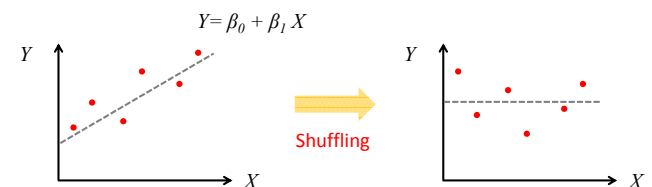


19

Permutation Tests: Linear Regression

- There are many variations on permutation tests
 - If it is a regression, and if the Y points are randomly associated with the X points under the null hypothesis, so that the true slope is zero, we can shuffle Y s, associating them with the X s at random. Each time, we compute the slope

x_1, x_2, x_3, x_4, x_5
 y_3, y_4, y_1, y_5, y_2



20

How To Do Permutation Tests in R

- Let's try something simple first
 - Given two samples called *a* and *b*

```

a <- rnorm(100, mean=0, sd=1)
b <- rnorm(100, mean=-1, sd=2)
mean(a) - mean(b)
m <- length(a)
d <- c(a,b)
e <- sample(d)
a2 <- e[ 1 : m ]
b2 <- e[ (m+1) : (m+n) ]
mean(a2) - mean(b2)
    
```

- Note that sample defaults to `replace=FALSE` and to a number of samples equal to the `length(d)`
- Actually, you might want to try shuffling many times

21

Gene Expression Data

- Let's revisit the gene expression data that we saw on Tuesday with 5194 genes and 32 samples

