# Multiple Hypothesis Testing and False Discovery Rates

Collin Nolte

March 6, 2018

# Table of Contents

# The Null Hypothesis

- Formally, a (usually dichotomous) statement that is testable on the basis of observing a process modeled by a set of random variables, or a random event
- Testable in that we can use observed data to quantify the strength of the statement, given observed data or outcome
- Can be used to test a particular statement of interest, or used (in the same manner) as a set of diagnostics for variable coefficients in a regression model
- The alternate hypothesis is usually a negation of the null, although could be one-sided

$$Y = X\beta, \quad \beta \in \mathbb{R}^p$$

$$\begin{cases} H_{0_{(i)}} & : \beta_i = 0 \\ H_{A_{1(i)}} & : \beta_i \neq 0 \\ H_{A_{2(i)}} & : \beta_i > 0 \end{cases}$$

# p-values

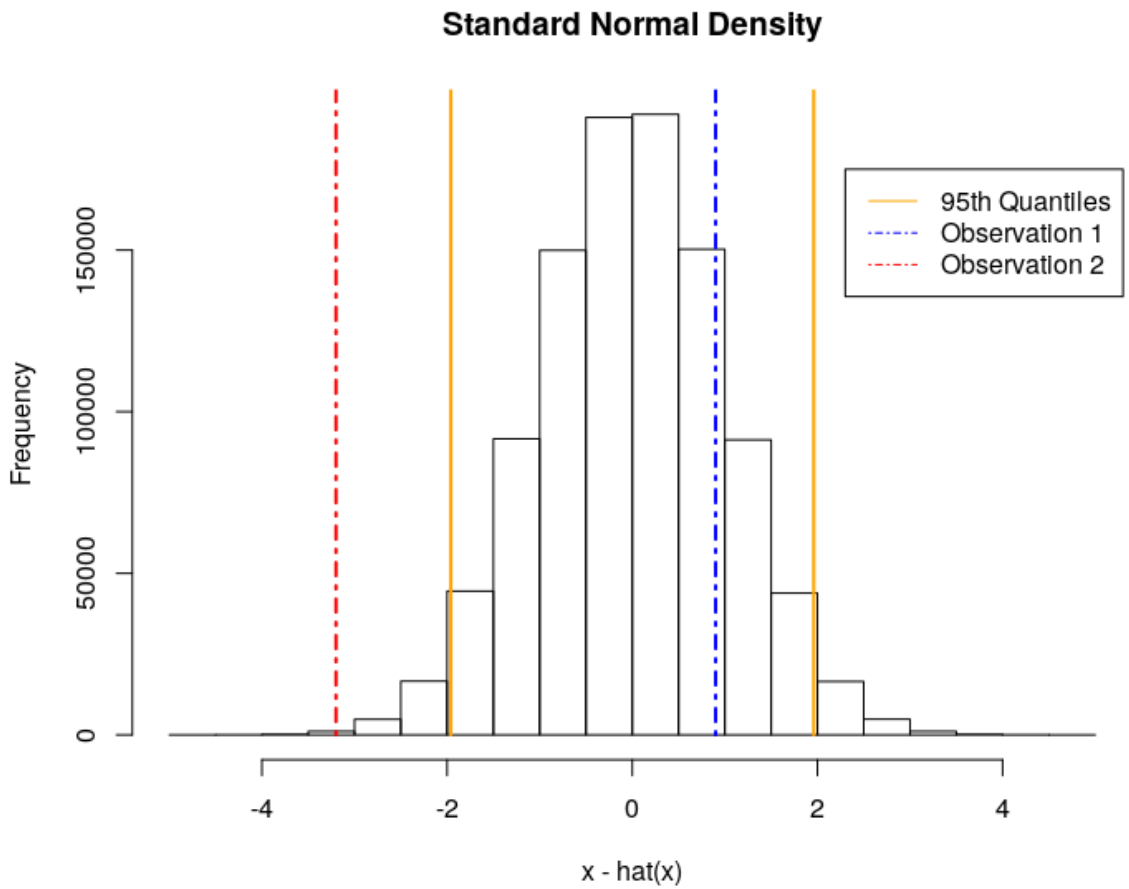
- The quantification of the strength of the null hypothesis is the p-value, which is the probability of the observed data, assuming that the null hypothesis is true
- Suppose we have observed data $x$ from a random process with standard error $SE$, and we hypothesize that it's true value is equal to $\hat{x}_0$, that is, $E(x) = \hat{x}$. By the central limit theorem, we have

$$\frac{x - \hat{x}}{SE} \stackrel{def}{=} z \sim N(0, 1)$$

$$\begin{cases} H_0 : x - \hat{x} = z = 0 \\ H_A : x - \hat{x} = z \neq 0 \end{cases}$$

# p-values cont

We can then consider the value of $\frac{x-\hat{x}}{SE}$ in the context of its assumed distribution

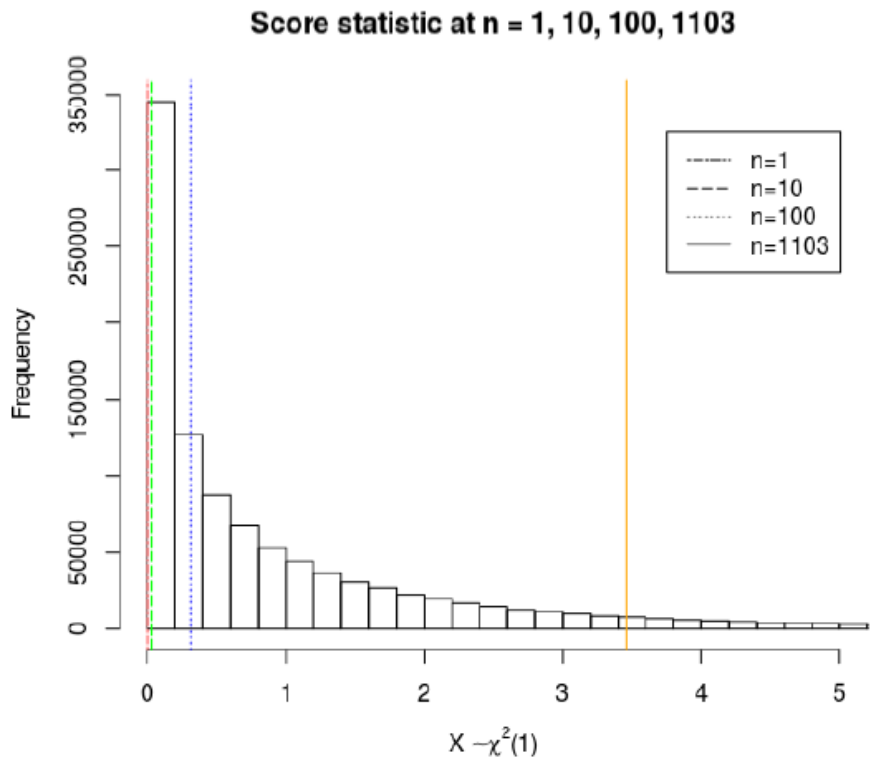


# Some limitations

- p-values are rooted with errors in terms of interpretability
- Consider a historical example in which 1103 seedlings of self fertilized plants, of which 854 were green and 249 were yellow. The hypothesis in question was that the ratio of green to yellow seedlings was 3:1, or, given a binary outcome (green/not green), the probability of a seedling being green was $\pi = 0.75$
- Note that the observed value was $\hat{\pi} = 854/1103 = 0.774$

# Limitations continued

The following is a plot of the observed value against the null distribution, with varying levels of $n$



# Type I and Type II Errors

- p-value represents a probability, rather than certainty
- By chance alone, uninteresting outcomes could be considered significant
- The Type 1 error rate is called the significance level, the probability of rejecting a null hypothesis when it is true
- Similarly, the power of a test is the probability of not committing a type II error, or falsely determining significance when there is none

| | Null Hypothesis | |
|---|---|---|
| | $H_0$ True | $H_0$ False |
| $H_0$ Rejected | Type 1 Error | Correct |
| $H_0$ Not Rejected | Correct | Type II Error |

# Refresher in Probability

- For events $A$ and $B$, we have (Bayes Rule) that

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A \cap B) = P(B|A)P(A) \leq P(B)P(A)$$

  where, under independence of events, $P(B|A) = P(B)$
- In general, then, the assumption that $A$ and $B$ are independent results in a larger probability of them both occurring than if they had non-empty intersection

# Type 1 Error for Multiple Tests

- Suppose, then, that we have $k$ sets of hypothesis tests, with a predetermined Type 1 Error rate (significance level) of $\alpha$
- The probability that any particular test does incorrectly reject a true null hypothesis is $1 - \alpha$
- Under the conservative assumption that each hypothesis test is in independent of one another, the probability of $k$ tests not incorrectly rejecting a true null is

$$P(\text{no Type 1 Errors}) = \prod^{k}(1-\alpha) = (1-\alpha)^k$$

- Consider a situation in which $k = 20$, with predetermined significance of $\alpha = 0.05$. The probability of committing no errors now becomes $(1 - 0.95)^{20} = 0.3584$

# Family-Wise Error Rates

- This can be restated as the Family-Wise Error Rate (FWER), or the probability of committing at least one error

$$\begin{aligned} FWER &= P(\text{at least 1 error}) \\ &= 1 - P(\text{no errors}) \\ &= 1 - \prod_{i=1}^{k}(1-\alpha) \\ &= 1 - (1-\alpha)^k \end{aligned}$$

- We can use this to specify $\alpha$ to control our FWER:

$$\alpha = 1 - (1 - FWER)^{\frac{1}{k}}$$

Slide modified from BIOS 5720, Spring 2017

# Corrections and procedures

- Sidac Correction, holds under independence, otherwise conservative

$$\alpha = 1 - (1 - FWER)^{1/k}$$

- Bonferroni Correction modifies nominal value of $\alpha$, dependent on $k$

$$P(\text{Type 1 Error}) = P\left(\bigcup_{i=1}^{k} z_i \leq \frac{\alpha}{k}\right)$$

$$\leq \sum_{i=1}^{k} P\left(z_i \leq \frac{\alpha}{k}\right)$$

$$\leq k\left(\frac{\alpha}{k}\right) = \alpha$$

- Alternatively, we could choose a different partition of $\alpha$, the modifications can be thought of as weights 'spent' over each of the $k$ tests

# Issues

- These tests tend to be overly conservative
- What if $k$ is really really big?
- Microarray gene data can have anywhere from 5,000 to 50,000 genes on which we are conducting a hypothesis
- We can no longer make crude adjustments and expect to find results

# West Gene Dataset Example, ER Status

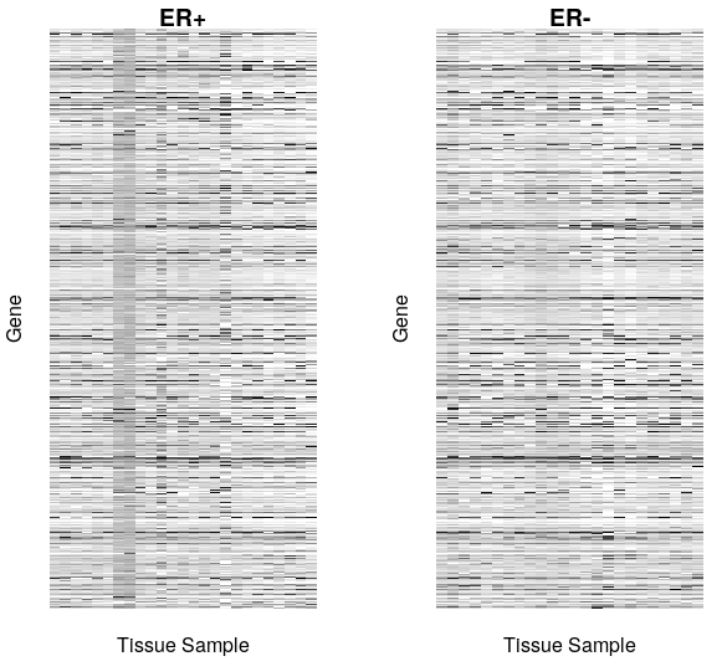
- West data set is a collection of 49 breast cancer tissue samples measuring the expression level of 7129 genes
- The samples were classified according to it's estrogen receptor (ER) status, a marker explaining several characteristics about the tumor
- Goal was to identify a set of genes that may be significant in determining the classification of a sample
- For each gene, significance was tested by comparing the mean expression of the gene in subjects who were $ER_+$ against those who were $ER_-$
- The null hypothesis is that an individual gene is not significant, i.e., there is no difference in mean expression level, i.e

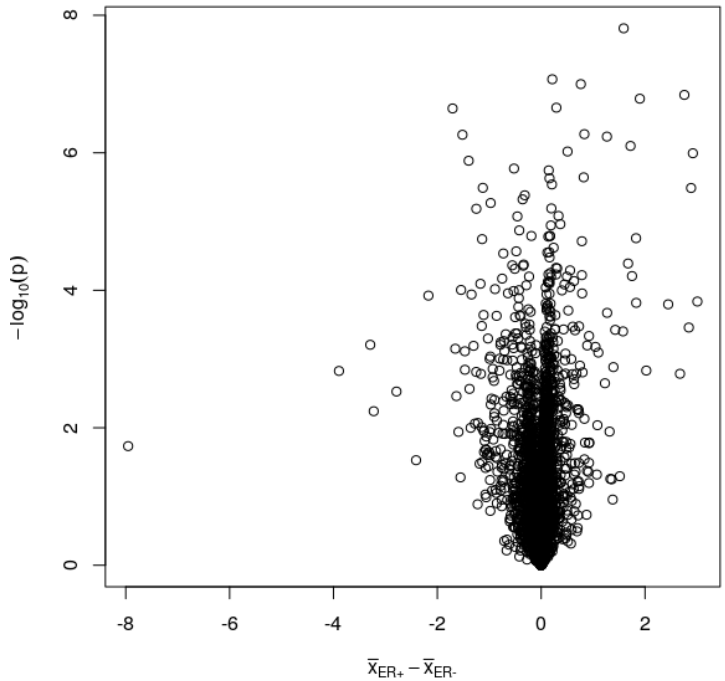
$$H_0 : \mu_{ER_+} = \mu_{ER_-} \quad \Rightarrow \quad H_0 : \mu_{ER_+} - \mu_{ER_-} = 0$$

# West Gene Dataset

```
## Create vector for ER status for each sample (1 = ER+, 0 = ER-)
er <- clinical$ER
table(er)
er
 0  1
24 25

## Perform two-sample t-tests for each gene and save all p-values
pval <- NULL
stat <- NULL
m <- nrow(chip.norm)
for(i in 1:m) {
   result <- t.test(chip.norm[i,] ~ er)
   pval <- c(pval, result$p.value)
   est <- result$estimate
   stat <- c(stat, est[2] - est[1])
}

## Number of p-values significant at the 5% level
> sum(pval < 0.05)
[1] 1325

## Bonferroni adjustment
> sum(pval < 0.05 / length(pval))
[1] 26
```

# West Gene Data Continued




Images and code modified from BIOS 6720, Spring 2018

# False Discovery Rates

- By chance, we expect $\alpha\%$ of the $M$ genes to be considered significant
- In table below, $V$ now represents the occurence of a type 1 error amongst the $M$ tests, and $P(V \geq 1)$ would be our FWER. This value is unobserved
- $R$ represents the total number of genes declared significant, regardless of the true value. This value is observed and known

| | Null Hypothesis | | |
|---|---|---|---|
| | $H_0$ True | $H_0$ False | Total |
| $H_0$ Rejected | V | S | R |
| $H_0$ Not Rejected | U | T | M-R |
| Total | $M_0$ | $M_1$ | M |

# False Discovery Rates

- Define the False Discovery Rate (FDR) to be

$$FDR = E(V/R)$$

- Regardless of independence or the distribution of the p-values, it holds that

$$FDR = E(V/R) \leq \frac{V+U}{M}\alpha = \frac{M_0}{M}\alpha \leq \alpha$$

- Important to note that the FDR is not the same as the FWER, which is $E(V)$

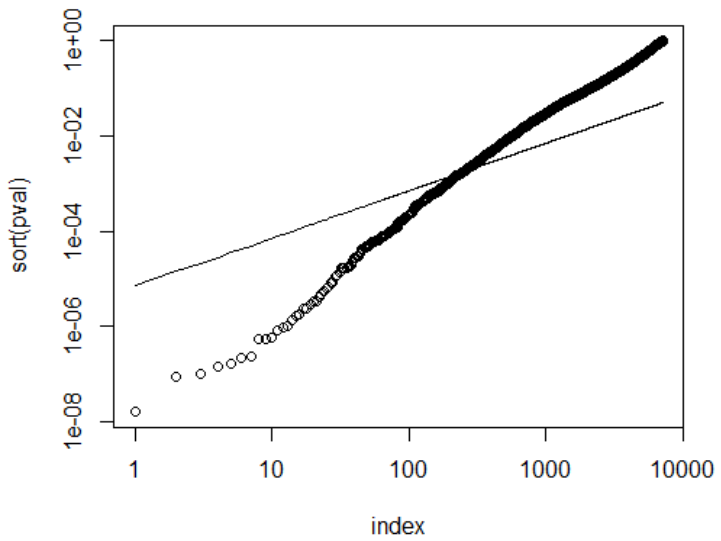
| | Null Hypothesis | | |
|---|---|---|---|
| | $H_0$ True | $H_0$ False | Total |
| $H_0$ Rejected | V | S | R |
| $H_0$ Not Rejected | U | T | M-R |
| Total | $M_0$ | $M_1$ | M |

# FDR Formulation (Benjamini-Hochberg

- Begin by fixing the false discovery rate $\alpha$, and let $p_{(1)} \leq p_{(2)} \leq \ldots p_{(n)}$ denote the ordered p values
- Define

$$L = \max \left\{ j \ : \ p_{(j)} \leq \alpha \frac{j}{M} \right\}$$

- Reject all hypothesis $H_{0_j}$ for which $p_{(j)} \leq p_{(L)}$



# Intuitive FDR

- As ER status is binary, for each gene, we could consider an array of $\{0, 1\}$, representing the subject's ER status, along with their gene expression
- Consider the p value $p_j$ associated with gene $j$. If gene $j$ is significant, relative to all other genes, then considering any random permutation of the $\{0, 1\}$ array would result in p values much larger than what was observed
- The idea, then, would be to consider the result of the hypothesis test for each possible permutation, and count the number of those randomly assigned that appear more significant than what would have been observed

# Intuitive FDR continued

- Letting $t_j$ denote the observed statistic for the $j$th gene, and $t_j^k$ denoting the observed statistic for the $kth$ permutation of the $j$th gene, we can define a new p value for gene $j$ to be

$$p_j = \frac{1}{K} \sum_{k=1}^{K} I(|t_j^k| > |t_j|)$$

- Now, for a desired $\alpha$, we consider a range of cutoff values $C$, and define

$$R_{obs} = \sum_{i=1}^{M} I(|t_j^k| > C), \quad \widehat{E(V)} = \frac{1}{K} \sum_{j=1}^{M} \sum_{k=1}^{K} I(|t_j^k| > C)$$

# Intuitive FDR continued

- We then define our estimate of FDR to be

$$\widehat{FDR} = \widehat{E(V)}/R_{obs}$$

- It can be shown that for the previously defined $p_j$, implementing the Benjamini-Hochberg Method asymptotically produces the same result. That is, the value of $C$ producing the desired cutoff value will be nearly equivalent to the cutoff value associated with $p_{(L)}$

# Conclusion

- Using the R package for estimating the FDR, along with user implemented simulation of the permutation method, we find $p = 247$ and $p = 306$ genes considered significant respectively
- Discrepancy likely comes from overly conservative nature of R package, which uses standard p-values instead of permutation p values
- This represents a huge improvement from the original 1325 significant genes found from p-values alone, and the 26 genes found using the Bonferroni adjustment

# Sources

- Course notes BIOS 5270
- Course notes BIOS 6270
- "Elements of Statistical Learning", Chapter 18, Trevor Hastie, Robert Tibshirani, Jerome Friedman.
- "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing", Yoav Benjamini, Yosef Hochberg, Jan 1993.