Lecture 9: Multiple Hypothesis Testing

May 29, 2012
GENOME 560, Spring 2012
Su-In Lee, CSE & GS
suinlee@uw.edu

Goals
- Define the multiple testing problem and related concepts
- Methods for addressing multiple testing (FWER and FDR)
- Correcting for multiple testing in R
- Final course evaluation (15 minutes)

Type I and II Errors

<table>
<thead>
<tr>
<th>Actual Situation “Truth”</th>
<th>H₀ True</th>
<th>H₀ False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Don Not Reject H₀</td>
<td>Correct Decision (True Negative) 1-α</td>
<td>Incorrect Decision (False Negative) Type II Error β</td>
</tr>
<tr>
<td>Reject H₀</td>
<td>Incorrect Decision (False Positive) Type I Error α</td>
<td>Correct Decision (True Positive) 1-β</td>
</tr>
</tbody>
</table>

α = P(Type I Error)  β = P(Type II Error)
Power = 1 - β

Type I and Type II Errors

- Consider the distribution of your test statistic

H₀ is true
Significance level
H₀ is false
Type II Error β
Type I Error α
Test statistic
Why Multiple Testing Matters

- **Genomics: Lots of data, Lots of hypothesis tests**
- A typical microarray experiment might result in performing 10,000 separate hypothesis tests.
- If we use a standard p-value cut-off of $\alpha = 0.05$, we’d expect 500 genes to be deemed “significant” by chance.
- Why 500?

Why Multiple Testing Matters

- In general, if we perform $m$ hypothesis tests, what is the probability of at least 1 false positive?
  - Assume that all the null hypotheses are true
    
    
    $P(\text{Making an error}) = \alpha$
    
    $P(\text{Not making an error}) = 1 - \alpha$
    
    $P(\text{Not making an error in } m \text{ tests}) = (1 - \alpha)^m$
    
    $P(\text{Making at least 1 error in } m \text{ tests}) = 1 - (1 - \alpha)^m$

Probability of At Least 1 False Positive

Counting Errors

- Assume that we are testing $m$ hypotheses: $H^1, \ldots, H^m$
  - $m_0 = \# \text{ of true null hypotheses}$
  - $R = \# \text{ of rejected null hypotheses}$

<table>
<thead>
<tr>
<th></th>
<th>Null True</th>
<th>Alternative True</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Called Significant</td>
<td>$U$</td>
<td>$T$</td>
</tr>
<tr>
<td>Called Significant</td>
<td>$V$</td>
<td>$S$</td>
</tr>
<tr>
<td></td>
<td>$m_0$</td>
<td>$m - m_0$</td>
</tr>
</tbody>
</table>

- $V = \# \text{ Type I errors [false positives]}$
Correcting for Multiple Testing?
- When we say “adjusting p-values for the number of hypothesis tests performed”, what we mean is controlling the Type I error rate.
- Very active area of statistics – many different methods have been described.
- Although these varied approaches have the same goal, they go about it in fundamentally different ways.

Different Approaches to Control Type I Errors
- Per comparison error rate (PCER): the expected value of the number of Type I errors over the number of hypotheses.
  \[ \text{PCER} = \frac{E(V)}{m} \]
- Per-family error rate (PFER): the expected number of Type I errors.
  \[ \text{PFER} = E(V) \]
- Family-wise error rate (FWER): the probability of at least one Type I error.
  \[ \text{FWER} = P(V \geq 1) \]
- False discovery rate (FDR): the expected proportion of Type I errors among the rejected hypotheses.
  \[ \text{FDR} = \frac{E(V/R | R>0)}{P(R>0)} \]
- Positive false discovery rate (pFDR): the rate that discoveries are false.
  \[ \text{pFDR} = \frac{E(V/R | R>0)}{1} \]

Family-Wise Error Rate (FWER)
- Many procedures have been developed to control the Family-Wise Error Rate (the probability of at least one Type I error):
  \[ P(V \geq 1) \]
- Two general types of FWER corrections:
  - Single step: equivalent adjustments made to each p-value
  - Sequential: adaptive adjustment made to each p-value.

Single Step Approach: Bonferroni
- Very simple method for ensuring that the overall Type I error rate of \( \alpha \) is maintained when performing \( m \) independent hypothesis tests.
- Rejects any hypothesis with p-value \( \leq \frac{\alpha}{m} \):
  \[ \tilde{p}_j = \min\{mp_j, 1\} \]
- For example, if we want to have an experiment wide Type I error rate of \( \alpha = 0.05 \) when we perform 10,000 hypothesis tests, we’d need a p-value of \( 0.05/10,000 = 5 \times 10^{-6} \) to declare significance.
Philosophical Objections to Bonferroni Corrections

- "Bonferroni adjustments are, at best, unnecessary and, at worst, deleterious to sound statistical inference" Perneger (1998)
- Counter-intuitive: interpretation of finding depends on the number of other tests performed
- The general null hypothesis (that all the null hypotheses are true) is rarely of interest
- High probability of Type II errors, i.e., of not rejecting the general null hypothesis when important effects exist

FWER: Sequential Adjustments

- Simplest sequential method is Holm’s Method
  - Order the unadjusted p-values such that \( p_1 \leq p_2 \leq ... \leq p_m \)
  - For control of the FWER at level \( \alpha \), the step-down Holm adjusted p-values are
    \[
    \tilde{p}_j = \min[(m - j + 1) \cdot p_j, 1]
    \]
  - The point here is that we don’t multiple every \( p_i \) by the same factor \( m \)
  - For example, when \( m = 10,000 \):
    \[
    \tilde{p}_1 = 10000 \cdot p_1, \quad \tilde{p}_2 = 9999 \cdot p_2, \ldots, \tilde{p}_m = 1 \cdot p_m
    \]

Who Cares About Not Making ANY Type I Errors?

- FWER is appropriate when you want to guard against ANY false positives
- However, in many cases (particularly in genomics) we can live with a certain number of false positives
- In these cases, the more relevant quantity to control is the false discovery rate (FDR)
  \[
  \frac{\text{# falsely rejected}}{\text{# rejected in total}}
  \]

False Discovery Rate

<table>
<thead>
<tr>
<th></th>
<th>Null True</th>
<th>Alternative True</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Called</td>
<td>( U )</td>
<td>( T )</td>
<td>( m-R )</td>
</tr>
<tr>
<td>Called</td>
<td>( V )</td>
<td>( S )</td>
<td>( R )</td>
</tr>
<tr>
<td>Significant</td>
<td>( m_0 )</td>
<td>( m - m_0 )</td>
<td>( m )</td>
</tr>
</tbody>
</table>

- \( V = \# \text{Type I errors [false positives]} \)
- False discovery rate (FDR) is designed to control the proportion of false positives among the set of rejected hypotheses (R) -- \( V/R \)
FDR vs FPR (False Positive Rate)

<table>
<thead>
<tr>
<th></th>
<th>Null True</th>
<th>Alternative True</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Called Significant</td>
<td>$U$</td>
<td>$T$</td>
<td>$m-R$</td>
</tr>
<tr>
<td>Called Significant</td>
<td>$V$</td>
<td>$S$</td>
<td>$R$</td>
</tr>
<tr>
<td></td>
<td>$m_0$</td>
<td>$m - m_0$</td>
<td>$m$</td>
</tr>
</tbody>
</table>

- $V =$ # Type I errors [false positives]
- FDR = $\frac{FP}{FP + TP} = \frac{V}{R}$
- FPR = $\frac{FP}{FP + TN} = \frac{V}{m_0}$

What If $R = 0$?

- Benjamini & Hochberg:
  \[ FDR = E \left[ \frac{V}{R} \right] P(R > 0) \]
- The rate that false discoveries occur
- Story:
  \[ pFDR = E \left[ \frac{V}{R} \right] P(R > 0) \]
- The rate that discoveries false

Benjamini and Hochberg FDR

- To control FDR at level $\delta$:
  1. Order the unadjusted $p$-values: $p_1 \leq p_2 \leq \ldots \leq p_m$
  2. Then find the test with the highest rank, $j$, for which the $p$-value, $p_j$, is less than or equal to $\left(\frac{j}{m}\right) \times \delta$
  3. Declare the tests of rank $1$, $2$, $\ldots$, $j$ as significant

\[ p(j) \leq \delta \frac{j}{m} \]

B&H FDR Example

- Controlling the FDR at $\delta = 0.05$

<table>
<thead>
<tr>
<th>Rank</th>
<th>P-value</th>
<th>$(j/m) \times \delta$</th>
<th>Reject $H_0$ ?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0008</td>
<td>0.005</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0.009</td>
<td>0.010</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0.165</td>
<td>0.015</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0.205</td>
<td>0.020</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0.396</td>
<td>0.025</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0.450</td>
<td>0.030</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0.641</td>
<td>0.035</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0.781</td>
<td>0.040</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0.900</td>
<td>0.045</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0.993</td>
<td>0.050</td>
<td>0</td>
</tr>
</tbody>
</table>
Storey’s Positive FDR (pFDR)

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

Storey: \[ p\text{FDR} = E \left[ \frac{V}{R} \bigg| R > 0 \right] \]

- Since \( P(R>0) \) is \( \sim 1 \) 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

Input Data

- Expression levels of 5419 genes in 32 samples from 16 human individuals
  - There are 2 replicates per individual (e.g. CEU_1_1 & CEU_1_2)
  - 16 individuals are from two populations: CEU (Europe) and YRI (African)

Expressed in the form of a table or matrix showing the expression levels of 5419 probesets across 32 samples from 16 individuals in two populations.