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

Actual Situation "Truth"

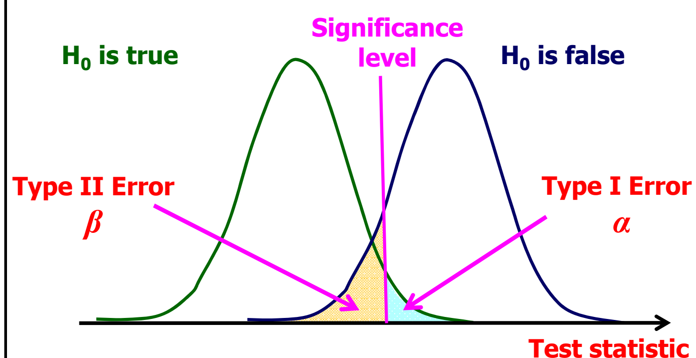
Decision	H_0 True	H_0 False
Don Not Reject H_0	Correct Decision (True Negative) $1-\alpha$	Incorrect Decision (False Negative) Type II Error β
Reject H_0	Incorrect Decision (False Positive) Type I Error α	Correct Decision (True Positive) $1-\beta$

α = P(Type I Error) β = P(Type II Error)

Power = $1 - \beta$

Type I and Type II Errors

- Consider the distribution of your 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?

5

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

$$P(\text{Making an error}) = \alpha$$

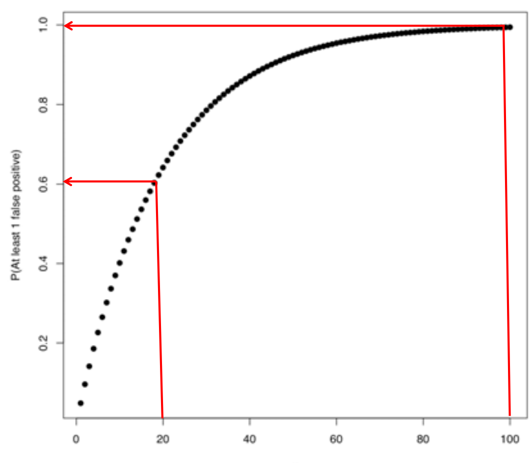
$$P(\text{Not making an error}) = 1 - \alpha$$

$$P(\text{Not making an error in } m \text{ tests}) = (1 - \alpha)^m$$

$$P(\text{Making at least 1 error in } m \text{ tests}) = 1 - (1 - \alpha)^m$$

6

Probability of At Least 1 False Positive



7

Counting Errors

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

	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]

8

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
  PCER = E(V)/m
- **Per-family error rate (PFER):** the expected number of Type I errors
  PFE = E(V)
- **Family-wise error rate (FWER):** the probability of at least one Type I error
  FWER = P(V≥1)
- **False discovery rate (FDR):** the expected proportion of Type I errors among the rejected hypotheses
  FDR = E(V/R | R>0) P(R>0)
- **Positive false discovery rate (pFDR):** the rate that discoveries are false
  pFDR = E(V/R | R>0)

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≥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 of α is maintained when performing m independent hypothesis tests
- Rejects any hypothesis with p-value ≤ α/m :

$$\tilde{p}_j = \min[mp_j, 1]$$

- For example, if we want to have an experiment wide Type I error rate of α = 0.05 when we perform 10,000 hypothesis tests, we'd need a p-value of 0.05/10,000 = 5 x 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 $p_1 \le p_2 \le \ldots \le p_m$
  - For control of the FWER at level α, 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:

$$\tilde{p}_1 = 10000 \cdot p_1, \quad \tilde{p}_2 = 9999 \cdot p_2, \cdots, \tilde{p}_m = 1 \cdot p_m$$

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

	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

FDR vs FPR (False Positive 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]

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TP}} = \frac{V}{R} \qquad \mathrm{FPR} = \frac{FP}{FP+TN} = \frac{V}{m_0}$$

What If R = 0 ?

- Benjamini & Hochberg:

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

  - The rate that false discoveries occur

- Story:

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

  - The rate that discoveries false

Benjamini and Hochberg FDR

- To control FDR at level δ:
  - 1. Order the unadjusted p-values: $p_1 \le p_2 \le \dots \le p_m$
  - 2. Then find the test with the highest rank, j, for which the p-value, p_j, is less than or equal to (j/m) x δ
  - 3. Declare the tests of rank 1, 2, …, j as significant

$$p(j) \le \delta \frac{j}{m}$$

B&H FDR Example

- Controlling the FDR at δ = 0.05

Rank (j)	P-value	(j/m) x δ	Reject H_0 ?
1	0.0008	0.005	1
2	0.009	0.010	1
3	0.165	0.015	0
4	0.205	0.020	0
5	0.396	0.025	0
6	0.450	0.030	0
7	0.641	0.035	0
8	0.781	0.040	0
9	0.900	0.045	0
10	0.993	0.050	0

Storey's Positive FDR (pFDR)

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

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

- 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

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)