Statistical Considerations of Multiple Testing

BE561

Single Hypothesis Testing

Two Types of Error

Type I Error: False Discovery

Type II Error: Missed Discovery

	Not Reject	**Reject**
H_0 **True**	TN	FP (Type I Error)
H_0 **False**	FN(Type II Error)	TP

Multiple Hypothesis Testing Problems in Genomics

- High-throughput microarray gene expression experiments

Multiple Hypothesis Testing Problems in Genomics

- High-throughput microarray gene expression experiments
  - ⇒ Identification of differentially expressed genes by testing for associations between gene expression measures and clinical covariates and outcomes

Multiple Hypothesis Testing Problems in Genomics

- High-throughput microarray gene expression experiments
  - ⇒ Identification of differentially expressed genes by testing for associations between gene expression measures and clinical covariates and outcomes
  - ⇒ Identification of co-expressed genes by testing for associations in the expression measures of sets of genes across biological samples

Multiple Hypothesis Testing Problems in Genomics

- High-throughput microarray gene expression experiments
  - ⇒ Identification of differentially expressed genes by testing for associations between gene expression measures and clinical covariates and outcomes
  - ⇒ Identification of co-expressed genes by testing for associations in the expression measures of sets of genes across biological samples
- Biological annotation metadata analysis

Multiple Hypothesis Testing Problems in Genomics

- High-throughput microarray gene expression experiments
  - ⇒ Identification of differentially expressed genes by testing for associations between gene expression measures and clinical covariates and outcomes
  - ⇒ Identification of co-expressed genes by testing for associations in the expression measures of sets of genes across biological samples
- Biological annotation metadata analysis
  - ⇒ Tests of association between gene expression measures and biological annotation metadata
    e.g.Gene Ontology(GO, `www.geneontology.org` annotation.

Multiple Hypothesis Testing Problems in Genomics

- ChIP-chip experiments. Identification of transcription factor binding sites in ChIP-chip experiments, where chromatin immunoprecipitation (ChIP) of transcription factor bound DNA is followed by microarray (chip) hybridization of the IP-enriched DNA

  Test of association between probe intensity measures and target sample (TF ChIP vs. control sample)

Multiple Hypothesis Testing Problems in Genomics

- ChIP-chip experiments. Identification of transcription factor binding sites in ChIP-chip experiments, where chromatin immunoprecipitation (ChIP) of transcription factor bound DNA is followed by microarray (chip) hybridization of the IP-enriched DNA

  Test of association between probe intensity measures and target sample (TF ChIP vs. control sample)
- Protein sequence analysis. Tests of association between phenotypes and codon/amino acid mutations.

  e.g. Association between viral replication capacity and HIV-1 sequence variation.

Multiplicity Problem

- Now assume we are carrying out multiple tests
  Test1: H_1 vs A_1 with p-value p_1
  Test2: H_2 vs A_2 with p-value p_2
  ...
  Testm: H_m vs A_m with p-value p_m

Multiplicity Problem

- Now assume we are carrying out multiple tests
  Test1: H_1 vs A_1 with p-value p_1
  Test2: H_2 vs A_2 with p-value p_2
  ...
  Testm: H_m vs A_m with p-value p_m
- If we knew which null hypotheses were true and if we had a procedure to accept/reject each test (p-value $< \alpha$), then we would have a table as follows:

	Not Significant	**Significant**	**Total**
Null is TRUE	U	V	m_0
Null is FALSE	T	S	$m-m_0$
	m-R	R	m

Multiplicity Problem

- Now assume we are carrying out multiple tests
  Test1: H_1 vs A_1 with p-value p_1
  Test2: H_2 vs A_2 with p-value p_2
  ...
  Testm: H_m vs A_m with p-value p_m
- If we knew which null hypotheses were true and if we had a procedure to accept/reject each test (p-value $< \alpha$), then we would have a table as follows:

	Not Significant	**Significant**	**Total**
Null is TRUE	U	V	m$_0$
Null is FALSE	T	S	m-m$_0$
	m-R	R	m

- Note that V is the number of total Type I Errors, and T is the number of Type II Errors.

Multiplicity Problem

- Now assume we are carrying out multiple tests
  Test1: H_1 vs A_1 with p-value p_1
  Test2: H_2 vs A_2 with p-value p_2
  ...
  Testm: H_m vs A_m with p-value p_m
- If we knew which null hypotheses were true and if we had a procedure to accept/reject each test (p-value $< \alpha$), then we would have a table as follows:

	Not Significant	**Significant**	**Total**
Null is TRUE	U	V	m$_0$
Null is FALSE	T	S	m-m$_0$
	m-R	R	m

- Note that V is the number of total Type I Errors, and T is the number of Type II Errors.
- m is known, R (number of rejected null hypotheses) is observed. U,T,V,and S are all unobservable random variables.

Multiplicity Problem

- Assume we are looking at each hypothesis in isolation, rejecting the null hypothesis H_i if $p_i < \alpha$. The probability of making a Type I Error for a single test is α.

Multiplicity Problem

- Assume we are looking at each hypothesis in isolation, rejecting the null hypothesis H_i if $p_i < \alpha$. The probability of making a Type I Error for a single test is α.
- For multiple tests, the probability of making at least one Type I Error in m tests is:

$$1 - (1 - \alpha)^m$$

Multiplicity Problem

- Assume we are looking at each hypothesis in isolation, rejecting the null hypothesis H_i if $p_i < \alpha$. The probability of making a Type I Error for a single test is α.
- For multiple tests, the probability of making at least one Type I Error in m tests is:

$$1 - (1 - \alpha)^m$$

- m $= 1000$, $\alpha = 0.01$, $P(TypeIErrors \geq 1) = 0.9999568$!

Multiplicity Problem

- Assume we are looking at each hypothesis in isolation, rejecting the null hypothesis H_i if $p_i < \alpha$. The probability of making a Type I Error for a single test is α.
- For multiple tests, the probability of making at least one Type I Error in m tests is:

$$1 - (1 - \alpha)^m$$

- m $= 1000$, $\alpha = 0.01$, $P(TypeIErrors \geq 1) = 0.9999568$!
- We need to adjust for multiple hypothesis testing.

Family-wise Error Rate

- Definition: The family-wise error rate is the probability of at least one FP (i.e. Type I error), that is: $FWER = P(\#FP \geq 1)$

Family-wise Error Rate

- Definition: The family-wise error rate is the probability of at least one FP (i.e. Type I error), that is: $FWER = P(\#FP \geq 1)$
- FWER is said to be controlled at level α if FWER $\leq \alpha$.

Compute FWER

- Let $H_1, H_2, ..., H_m$ be independent hypotheses.

Compute FWER

- Let $H_1, H_2, ..., H_m$ be independent hypotheses.
- Assume the first m are true, the others false:

Compute FWER

- Let $H_1, H_2, ..., H_m$ be independent hypotheses.
- Assume the first m are true, the others false: $FWER = P(\#FP \geq 1) = 1 - P(\#FP = 0)$

Compute FWER

- Let $H_1, H_2, ..., H_m$ be independent hypotheses.
- Assume the first m are true, the others false:
  $FWER = P(\#FP \geq 1) = 1 - P(\#FP = 0)$
  $P(FP = 0) = P(\textit{ not reject } H_1, ..., H_m)$
  $P(FP = 0) = (1 - \alpha_1)...(1 - \alpha_m)$ where $\alpha_j = P(\textit{reject } H_j)$

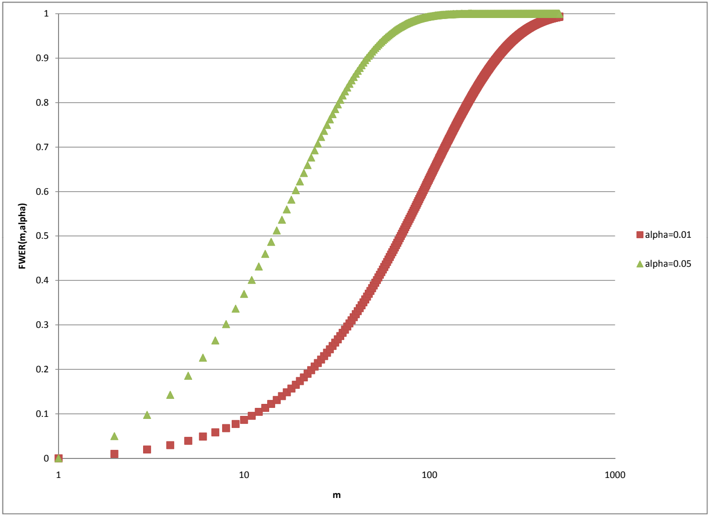
Compute FWER

- Let $H_1, H_2, ..., H_m$ be independent hypotheses.
- Assume the first m are true, the others false:
  $FWER = P(\#FP \geq 1) = 1 - P(\#FP = 0)$
  $P(FP = 0) = P(\text{ not reject } H_1, ..., H_m)$
  $P(FP = 0) = (1 - \alpha_1)...(1 - \alpha_m)$ where $\alpha_j = P(\text{reject } H_j)$
  $FWER = 1 - \prod_{j=1}^{m} (1 - \alpha_j)$

Compute FWER

- e.g. 10 hypotheses, $\alpha_j = 0.05, FWER(m) = 1 - 0.95^m$
  FWER(0)=0%, FWER(1)=5%,
  FWER(2)$\approx$ 9.8%, $FWER(10) \approx 40.1\%$
- e.g. 10 hypotheses, $\alpha_j = 0.167, FWER(m) = 1 - 0.83^m$
  FWER(0)=0%, FWER(1)=16.7%,
  FWER(2)$\approx$ 31.1%, $FWER(10) \approx 84.5\%$

Compute FWER



Adjusting p-value for FWER control

- Keep FWER below α

Adjusting p-value for FWER control

- Keep FWER below α
- Bonferroni Correction:
  Reject any hypothesis with $p - value \leq \frac{\alpha}{m}$
  Bonferroni adjusted p-values: $p_{Bonferroni} = min(m.p_j, 1)$

Adjusting p-value for FWER control

- Keep FWER below α
- Bonferroni Correction:
  Reject any hypothesis with $p - value \leq \frac{\alpha}{m}$
  Bonferroni adjusted p-values: $p_{Bonferroni} = min(m.p_j, 1)$
- Bonferroni Correction controls FWER.

Adjusting p-value for FWER control

- Keep FWER below α
- Bonferroni Correction:
  Reject any hypothesis with $p - value \leq \frac{\alpha}{m}$
  Bonferroni adjusted p-values: $p_{Bonferroni} = min(m.p_j, 1)$
- Bonferroni Correction controls FWER.
- There are also other methods that control FWER:
  - ⇒ Holm(1979) based on the order of raw p-values
  - ⇒ Westfall-Young (1993) step-up/step-down methods use order and joint distribution of raw p-values.

False Discovery Rate

- Alternative measure for multiple testing error introduced by Benjamini-Hochberg (1995):

  $$FDR = E\left[\frac{FP}{max(R, 1)}\right]$$

  where R is the number of rejected hypotheses.

False Discovery Rate

- Alternative measure for multiple testing error introduced by Benjamini-Hochberg (1995):

$$FDR = E\left[\frac{FP}{max(R,1)}\right]$$

  where R is the number of rejected hypotheses.
- FDR is the expected proportion of false positives among rejected hypotheses.

False Discovery Rate

- Alternative measure for multiple testing error introduced by Benjamini-Hochberg (1995):

$$FDR = E\left[\frac{FP}{max(R,1)}\right]$$

  where R is the number of rejected hypotheses.
- FDR is the expected proportion of false positives among rejected hypotheses.
- Why introduce another quantity to control?

False Discovery Rate

- Alternative measure for multiple testing error introduced by Benjamini-Hochberg (1995):

  $$FDR = E\left[\frac{FP}{max(R,1)}\right]$$

  where R is the number of rejected hypotheses.
- FDR is the expected proportion of false positives among rejected hypotheses.
- Why introduce another quantity to control?
  - $\Rightarrow$ Bonferroni adjustment is too strict for many applications. It was originally developed for well-crafted experiments with well designed follow-up questions. It works well for those...

False Discovery Rate

- Alternative measure for multiple testing error introduced by Benjamini-Hochberg (1995):

$$FDR = E\left[\frac{FP}{max(R,1)}\right]$$

  where R is the number of rejected hypotheses.
- FDR is the expected proportion of false positives among rejected hypotheses.
- Why introduce another quantity to control?
  - ⇒ Bonferroni adjustment is too strict for many applications. It was originally developed for well-crafted experiments with well designed follow-up questions. It works well for those...
  - ⇒ In theory-poor observational studies(i.e.microarray, ChIP-chip studies), the strategy is to test everything in sight.

False Discovery Rate

- Alternative measure for multiple testing error introduced by Benjamini-Hochberg (1995):

$$FDR = E\left[\frac{FP}{max(R,1)}\right]$$

  where R is the number of rejected hypotheses.
- FDR is the expected proportion of false positives among rejected hypotheses.
- Why introduce another quantity to control?
  - ⇒ Bonferroni adjustment is too strict for many applications. It was originally developed for well-crafted experiments with well designed follow-up questions. It works well for those...
  - ⇒ In theory-poor observational studies(i.e.microarray, ChIP-chip studies), the strategy is to test everything in sight.
  - ⇒ For genomics experiments, controlling the probability of one or more Type I errors is too severe but doing nothing at all is also unacceptable. FDR is a compromise.

Benjamini-Hochberg Procedure to Control FDR

To control FDR at level α:

- Let $p_1, p_2, ..., p_n$ the p-values of the m tests we carried out.

Benjamini-Hochberg Procedure to Control FDR

To control FDR at level α:

- Let $p_1, p_2, ..., p_n$ the p-values of the m tests we carried out.
- Order these p-values from smallest to largest: $p_{(1)}, p_{(2)}, \cdots, p_{(n)}$

Benjamini-Hochberg Procedure to Control FDR

To control FDR at level α:

- Let $p_1, p_2, ..., p_n$ the p-values of the m tests we carried out.
- Order these p-values from smallest to largest: $p_{(1)}, p_{(2)}, ..., p_{(n)}$
- Calculate the threshold value for each p-value:

$$\frac{k.\alpha}{m}$$

Benjamini-Hochberg Procedure to Control FDR

To control FDR at level α:

- Let $p_1, p_2, ..., p_n$ the p-values of the m tests we carried out.
- Order these p-values from smallest to largest: $p_{(1)}, p_{(2)}, \cdots, p_{(n)}$
- Calculate the threshold value for each p-value:

$$\frac{k.\alpha}{m}$$

- Let $k^{'} = max\left\{k : p_{(k)} \leq \frac{k.\alpha}{m}\right\}, k = 1, 2, ..., m$. If it turns out that $k^{'} = 0$ for all k then take $p_{(k)} \geq \frac{k.\alpha}{m}$.

Benjamini-Hochberg Procedure to Control FDR

To control FDR at level α:

- Let $p_1, p_2, ..., p_n$ the p-values of the m tests we carried out.
- Order these p-values from smallest to largest: $p_{(1)}, p_{(2)}, ..., p_{(n)}$
- Calculate the threshold value for each p-value:

$$\frac{k.\alpha}{m}$$

- Let $k^{'} = max\left\{k : p_{(k)} \leq \frac{k.\alpha}{m}\right\}$,$k = 1, 2, ..., m$. If it turns out that $k^{'} = 0$ for all k then take $p_{(k)} \geq \frac{k.\alpha}{m}$.
- Decision rule:

Benjamini-Hochberg Procedure to Control FDR

To control FDR at level α:

- Let $p_1, p_2, ..., p_n$ the p-values of the m tests we carried out.
- Order these p-values from smallest to largest: $p_{(1)}, p_{(2)}, \cdots, p_{(n)}$
- Calculate the threshold value for each p-value:

$$\frac{k.\alpha}{m}$$

- Let $k^{'} = max\left\{k : p_{(k)} \leq \frac{k.\alpha}{m}\right\}, k = 1, 2, ..., m$. If it turns out that $k^{'} = 0$ for all k then take $p_{(k)} \geq \frac{k.\alpha}{m}$.
- Decision rule:
  - $\Rightarrow$ If $k^{'} \geq 1$, then reject the hypotheses corresponding to $p_{(1)}, p_{(2)}, ..., p_{(k)}$.

Benjamini-Hochberg Procedure to Control FDR

To control FDR at level α:

- Let $p_1, p_2, ..., p_n$ the p-values of the m tests we carried out.
- Order these p-values from smallest to largest: $p_{(1)}, p_{(2)}, ..., p_{(n)}$
- Calculate the threshold value for each p-value:

$$\frac{k.\alpha}{m}$$

- Let $k^{'} = max\left\{k : p_{(k)} \leq \frac{k.\alpha}{m}\right\}$,$k = 1, 2, ..., m$. If it turns out that $k^{'} = 0$ for all k then take $p_{(k)} \geq \frac{k.\alpha}{m}$.
- Decision rule:
  - $\Rightarrow$ If $k^{'} \geq 1$, then reject the hypotheses corresponding to $p_{(1)}, p_{(2)}, ..., p_{(k)}$.
  - $\Rightarrow$ If $k^{'} = 0$, don't reject anything.

Benjamini-Hochberg Procedure to Control FDR

- Benjamini-Hochberg procedure controls FDR at level α assuming that the test statistics from each hypothesis is independent. You are guaranteed that the false discovery rate for the k hypotheses you have rejected is not bigger than α.

Benjamini-Hochberg Procedure to Control FDR

- Benjamini-Hochberg procedure controls FDR at level α assuming that the test statistics from each hypothesis is independent. You are guaranteed that the false discovery rate for the k hypotheses you have rejected is not bigger than α.
- FDR is a global (for all hypotheses) measure of significance. It is the expected proportion of false positives among significant hypotheses.

Benjamini-Hochberg Procedure to Control FDR

- Benjamini-Hochberg procedure controls FDR at level α assuming that the test statistics from each hypothesis is independent. You are guaranteed that the false discovery rate for the k hypotheses you have rejected is not bigger than α.
- FDR is a global (for all hypotheses) measure of significance. It is the expected proportion of false positives among significant hypotheses.
- This is not the only way to control FDR or other quantities. See:
  Genomics, Prior Probability, and Statistical Tests of Multiple Hypotheses, Genome Res. 2004 Jun;14(6):997-1001.

Acknowledgement

- Slides are due to Zhiping Weng, BU