The Multiple Comparison Problem and FDR

Acknowlegement: Rafael A. Irizarry John Storey
Yongchao Ge,
Sandrine Dudoit, Terry Speed

Hypothesis testing

- Once you have a score for each gene, how do you decide on a cut-off?
  p-values are popular. Are they appropriate?
- Test for each gene null hypothesis: no differential expression.
- Notice that if you have look at 10,000 genes for which the null is true you expect to see 500 attain p-values of 0.05
- This is called the multiple comparison problem. Statisticians fight about it. But not about the above.
- Main message: p-values can't be interpreted in the usual way

What do we do?

- Bonferoni correction (too conservative)
  - Adjusted p: $p^* = p \times n$ or $\alpha^* = \alpha/n$
- Give list of genes and report:
  - Family-Wise Error Rate: probability of including at least one non-differentially expressed gene
  - False discovery rate (FDR): expected proportion of Type I errors among the rejected hypotheses
  - pFDR: Expected proportion of false discoveries among the genes in your list conditioning on at least one gene is included in the differential list.
- Bayesian inference

p-values

The ***p*-value** or **observed significance level** *p* is the chance of getting a test statistic as or more extreme than the observed one, under the null hypothesis *H* of no differential expression.

Null distribution of *p*-value should be unif(0,1).

The ***nominal p-value*** is a statistic calculated from the data. Sometimes the nominal p-value may not be real p-value: when you have an invalid test. For example, when assumptions are severely violated.

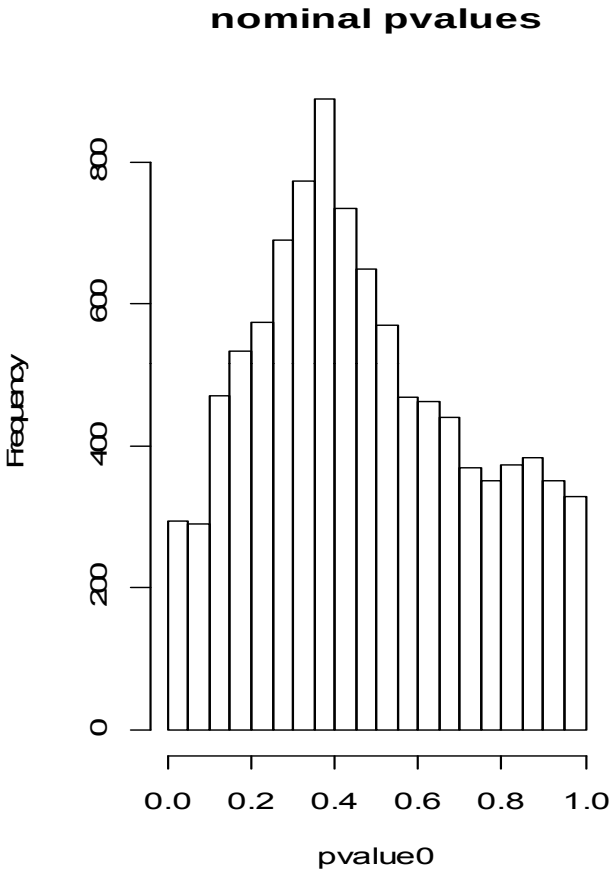
Example of invalid test

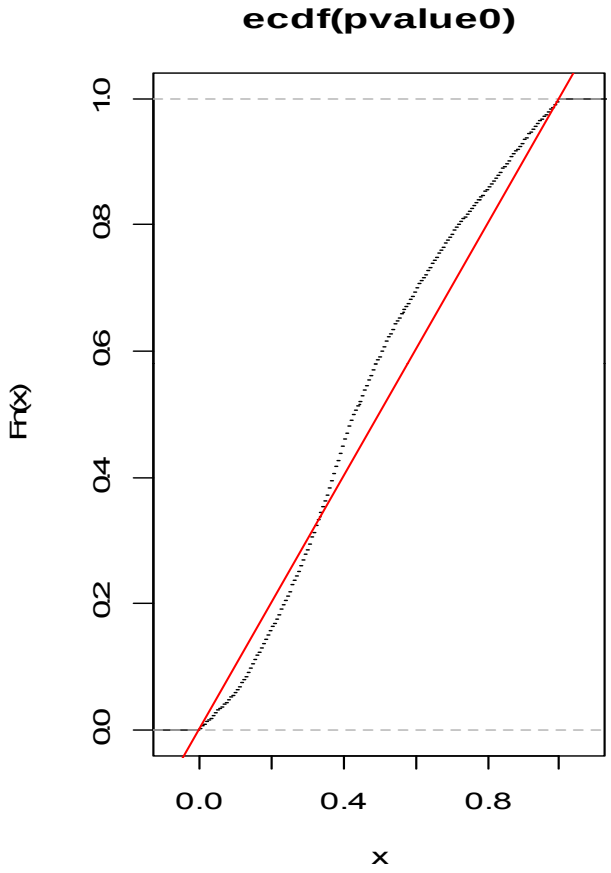
You have seen in your homework simulations that student t-test is rather robust. Even when normality assumption is violated and replaced by exponential distribution or with t-distribution, using student t-test still resulted in reasonable p-values.

When t-test breaks down:

```
set.seed(0227)
n1=n2=3;mu=rgamma(10000,6,.6)
X=matrix( exp(rnorm(10000*10,mu,1.2)),10000,10)
myt.test <-function(x)
  t.test(x[1:n1], x[(n1+1):(n1+n2)], var.equal = TRUE)$p.value
pvalue0 <- apply(X,1,myt.test)
par(mfrow=c(1,2))
hist(pvalue0, main = "nominal pvalues")
plot(ecdf(pvalue0))
abline(0,1,col=2)
```

Nominal p-values and actual type I error






```
> mean(pvalue0<.05)
[1] 0.0295
> mean(pvalue0<.01)
[1] 0.0065
```

p-values by permutations

We focus on one gene only. For the *b*th iteration, $b = 1, \cdots, B$;

1. Permute the n data points for the gene (x). The first n_1 are referred to as "treatments", the second n_2 as "controls".
2. For each gene, calculate the corresponding two sample t-statistic, t_b.

After all the B permutations are done;

3. Put $p = \#\{b: |t_b| \geq |t_{observed}|\}/B$.

P-value by bootstrapping

- $Y=X\beta+\varepsilon$
- Under null and assuming equal variance for all genes the residuals follow iid distribution
- If there is gene specific variance, then the scaled residuals may still follow iid distribution
- Resample the residuals ε^* with replacement
- Create new data set by $Y^* = X\hat{\beta} + \varepsilon^*$
- Repeat analysis and get a new test-statistic or nominal p-value
- Put $p = \#\{b: |t_b| \geq |t_{observed}|\}/B$

.

Multiple Hypothesis Testing

	Called Significant	Not Called Significant	Total
Null True	***V***	$m_0 - V$	m_0
Altern.True	***S***	$m_1 - S$	m_1
Total	***R***	$m - R$	m

Multiple Hypothesis Test
Error Controlling Procedure

- Suppose *m* hypotheses are tested with p-values $p_1, p_2, \ldots, p_m$
- A multiple hypothesis error controlling procedure is a function $T(\mathbf{p}; \alpha)$ such that rejecting all nulls with $p_i \leq T(\mathbf{p}; \alpha)$ implies that *Error* $\leq \alpha$
- *Error* is a population quantity (not random)

Weak and Strong Control

- If $T(\mathbf{p}; \alpha)$ is such *Error* ≤ α only when $m_0 = m$ *(complete null)*, then the procedure provides ***weak control*** of the error measure

- If $T(\mathbf{p}; \alpha)$ is such *Error* ≤ α for any value of m_0, then the procedure provides ***strong control*** of the error measure – note that m_0 is <u>not</u> an argument of $T(\mathbf{p}; \alpha)$.

- In microarray setting, $m=m_0$ is very unlikely. It is particularly important to have strong control.

Error Rates

- **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

  $PFER = E(V)$

- **Family-wise error rate**: the probability of at least one Type I error

  $FEWR = Pr(V \geq 1)$

- **False discovery rate** (FDR) rate that false discoveries occur

  $FDR = E(V/R; R>0) = E(V/R \mid R>0)Pr(R>0)$

- **Positive false discovery** rate (pFDR): rate that discoveries are false

  $pFDR = E(V/R \mid R>0)$.

Bonferroni Procedure

$$T(\mathbf{p};\alpha) = \max\left\{ p_i : p_i \le \frac{\alpha}{m} \right\}$$

Provides strong control…..

$$\Pr(V \ge 1) \le \Pr\left(\min_i \; p_i \le \frac{\alpha}{m} \mid H_0^C \right)$$

$$\le \sum_{i=1}^{m} \Pr\left(p_i \le \frac{\alpha}{m} \mid H_0^i \right)$$

$$= m \cdot \frac{\alpha}{m}$$

Holm Procedure

Order the p-values $p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)}$

$$T(\mathbf{p};\alpha) = \min\left\{ p_{(i)} : p_{(i)} > \frac{\alpha}{m-i+1} \right\}$$

$$T(\mathbf{p};\alpha) = \min\left\{ p_{(i)} : p_{(i)} > 1-\left(1-\alpha\right)^{1/(m-i+1)} \right\}$$

Requires independence for strong control…

Hochberg Procedure

$$T(\mathbf{p};\alpha) = \max\left\{ p_{(i)} : p_{(i)} \leq \frac{\alpha}{m-i+1} \right\}$$

...the step-up analogue of Holm

Simes/BH Procedure

$$T(\mathbf{p};\alpha) = \max\left\{ p_{(i)} : p_{(i)} \leq \frac{i \cdot \alpha}{m} \right\}$$

- Weak controls the FWER (Simes 1986)
- Strongly controls FDR (Benjamini & Hochberg 1995)
- Both require the null p-values to be independent

False Discovery Rate

- **The** "false discovery rate" **measures the *EXPECTED* proportion of false positives among all genes called significant:**

$$E\left[\frac{\#\,\text{false positives}}{\#\,\text{called significant}}\right] = E\left[\frac{V}{V+S}\right] = E\left[\frac{V}{R}\right]$$

- **This is usually appropriate because one wants to find as many truly differentially expressed genes as possible with relatively few false positives**
- **The false discovery rate gives the rate at which further biological verification will result in dead-ends**

False Positive Rate versus False Discovery Rate

- False positive rate **is the rate at which truly null genes are called significant**

$$\text{FPR} \approx \frac{\#\,\text{false positives}}{\#\,\text{truly null}} = \frac{V}{m_0}$$

- False discovery rate **is the rate at which significant genes are truly null**

$$\text{FDR} \approx \frac{\#\,\text{false positives}}{\#\,\text{called significant}} = \frac{V}{R}$$

- Notice we wrote $\mathrm{FDR} \approx \frac{V}{R}$

- The FDR is the Expected (average) proportion. In a particular experiment, it is not guaranteed that the false discovery *proportion* is the same.

False Positive Rate and P-values

- The *p-value* is a measure of significance in terms of the false positive rate (aka Type I error rate)
- P-value is defined to be the minimum **false positive rate** at which the statistic can be called significant
- Can be described as the probability a truly null statistic is “as or more extreme” than the observed one

False Discovery Rate and Q-values

- The *q-value* is a measure of significance in terms of the false discovery rate
- Q-value is defined to be the minimum **false discovery rate** at which the statistic can be called significant
- Can be described as the probability a statistic “as or more extreme” is truly null

Bayesian interpretation

- Suppose m hypothesis tests are performed with independent statistics $X_1, \ldots, X_m$ and significance region Γ.
- Let $H_i = 0$ if null hypothesis i is true, and $H_i = 1$ if it is false. Assume $\Pr(H_i = 0) = \pi_0$ and $\Pr(H_i = 1) = \pi_1$.
- Assume each statistic comes from the mixture distribution, $X_i \sim (1 - H_i) \cdot F_0 + H_i \cdot F_1$, where F_0 is the null and F_1 is the alternative.

Theorem: (Storey 2001)

$$\mathrm{pFDR}(\Gamma) = \mathrm{E}\left[\frac{V(\Gamma)}{R(\Gamma)} \middle| R(\Gamma) > 0\right] = \frac{\pi_0 \cdot \Pr(X \in \Gamma | H = 0)}{\Pr(X \in \Gamma)} = \Pr(H = 0 | X \in \Gamma).$$

Power/Type I error decomposition

- **Under the mixture model assumptions ...**

$$\text{pFDR}(\Gamma) = \frac{\pi_0 \cdot \Pr(X \in \Gamma | H = 0)}{\pi_0 \cdot \Pr(X \in \Gamma | H = 0) + \pi_1 \cdot \Pr(X \in \Gamma | H = 1)}$$

$$= \frac{\pi_0 \cdot \text{Type I error rate}}{\pi_0 \cdot \text{Type I error rate} + \pi_1 \cdot \text{Power}}$$

q-values

- **In general, for a nested set of significance regions $\{\Gamma\}$, the p-value of an observed statistic x is defined to be**

$$\textbf{p-value}(x) = \inf_{x \in \Gamma} \Pr(X \in \Gamma | H = 0)$$

- **Likewise, under the independent mixture model,**

$$\textbf{q-value}(x) = \inf_{x \in \Gamma} \text{pFDR}(\Gamma) = \inf_{x \in \Gamma} \Pr(H = 0 | X \in \Gamma).$$

Bayesian Connections

- **This allows Bayesians to estimate FDR as well:**

$$\mathrm{pFDR}(\Gamma) = \int \Pr(H = 0 \mid X = x)\, f(x \mid x \in \Gamma)\, dx$$

- **This motivates the name "q-value" directly:**

$$\mathrm{p-value}(x_i) = \Pr\left(\mid X \mid \geq \mid x_i \mid \,\middle|\, H = 0 \right)$$

$$\mathrm{q-value}(x_i) = \Pr\left(H = 0 \,\middle|\, \mid X \mid \geq \mid x_i \mid \right)$$

- **All the estimation presented below can be viewed as an "empirical Bayes" approach**

SAM Version

$$\widehat{\mathrm{FDR}}(\Delta) = \frac{\hat{\pi}_0 \sum_{b=1}^{B} \#\left\{d_i^{0b} : d_i^{0b} \le \ell(\Delta) \text{ or } d_i^{0b} \ge r(\Delta)\right\} / B}{\#\left\{d_i : d_i \le \ell(\Delta) \text{ or } d_i \ge r(\Delta)\right\}}$$

$$= \frac{\hat{\pi}_0 \cdot \text{avg no. nulls called significant}}{\text{no. observed called significant}}$$

$$\hat{\pi}_0(\Delta') = \frac{\#\left\{d_i : d_i > \ell(\Delta') \text{ or } d_i < r(\Delta')\right\}}{\sum_{b=1}^{B} \#\left\{d_i^{0b} : d_i^{0b} > \ell(\Delta') \text{ or } d_i^{0b} < r(\Delta')\right\} / B}$$

- The test a method uses may not be “valid”
- From the same data, different methods report different p-values. What is reported is a “nominal p-value”, which could be too liberal or too conservative.
- The same applies to FDR.
    - Nominal FDR is not necessarily actual FDR
    - Even actual FDR is not the same as the false discovery proportion.
- There are a lot of theoretical work on FDR, but most methods implement a simple version. The most widely used is B-H procedure.

In your practice, simulation under your own setting will provide some information on how good the reported FDR is.