

Tutorial on Statistical Methods and Software for the Analysis of Microarray Experiments

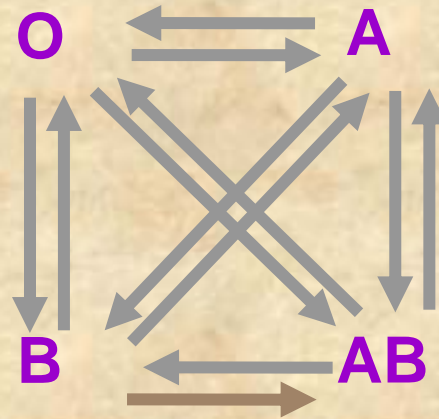


Nicholas P. Jewell
Sandrine Dudoit

September 23, 2004

Class 4
Analysis of Microarray Experiments:
Multiple Comparisons

Experimental Design



Combining data across slides

Data on G genes for n hybridizations



$G \times n$ genes-by-arrays data matrix

Arrays

		Array1	Array2	Array3	Array4	Array5	...
Genes	Gene1	0.46	0.30	0.80	1.51	0.90	...
	Gene2	-0.10	0.49	0.24	0.06	0.46	...
	Gene3	0.15	0.74	0.04	0.10	0.20	...
	Gene4	-0.45	-1.03	-0.79	-0.56	-0.32	...
	Gene5	-0.06	1.06	1.35	1.09	-1.09	...

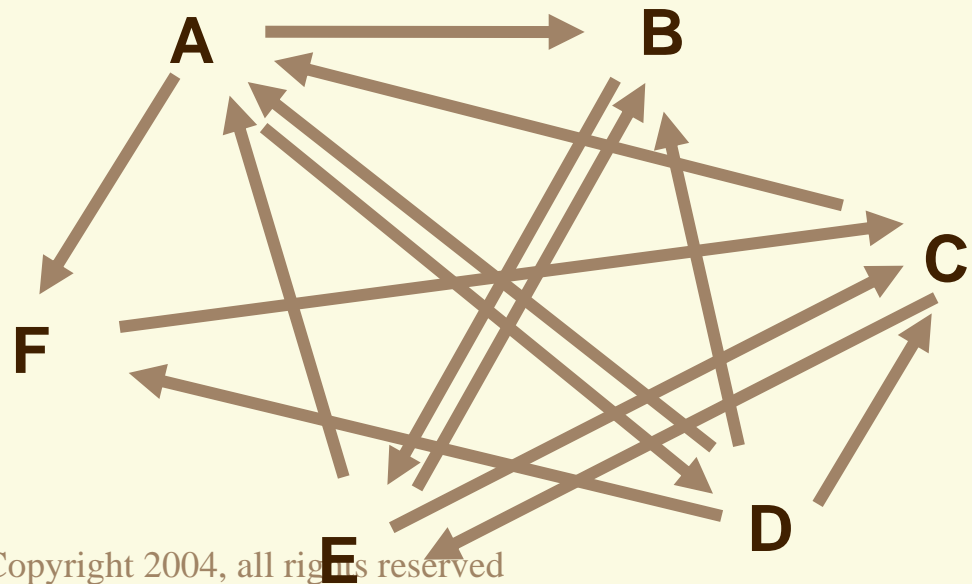
$$\mathbf{M} = \log_2(\text{Red intensity} / \text{Green intensity})$$

Combining data across slides

... but columns have **structure**

How can we design experiments and combine data across slides to provide accurate estimates of the effects of interest?

Linear models



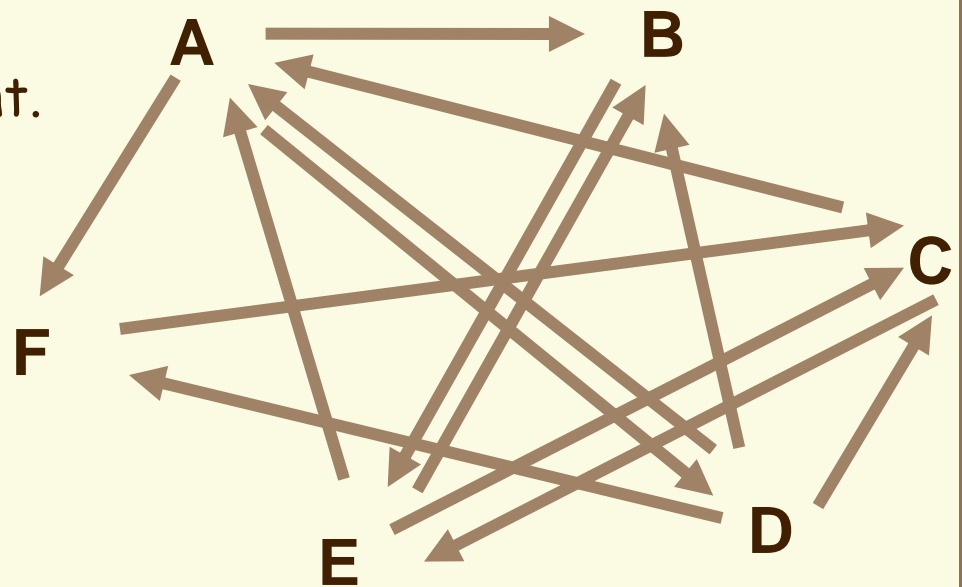
Experimental design

- ❖ Design of the array itself
 - which cDNA probe sequences to print;
 - whether to use replicated probes;
 - whether to use control sequences;
 - how many and where these should be printed.
- ❖ Allocation of mRNA samples to the slides
 - pairing of mRNA samples for hybridization;
 - dye assignments;
 - type and number of replicates.

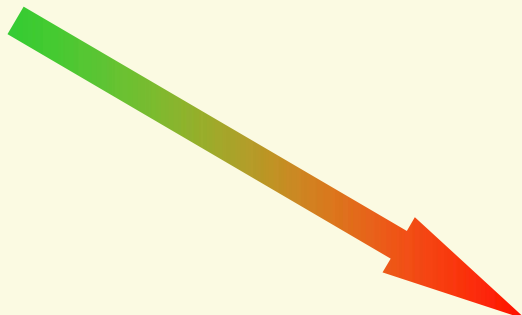
Graphical representation

Multi-digraph

- ❖ *Vertices*: mRNA samples;
- ❖ *Edges*: hybridization;
- ❖ *Direction*: dye assignment.



Cy3 sample



Cy5 sample

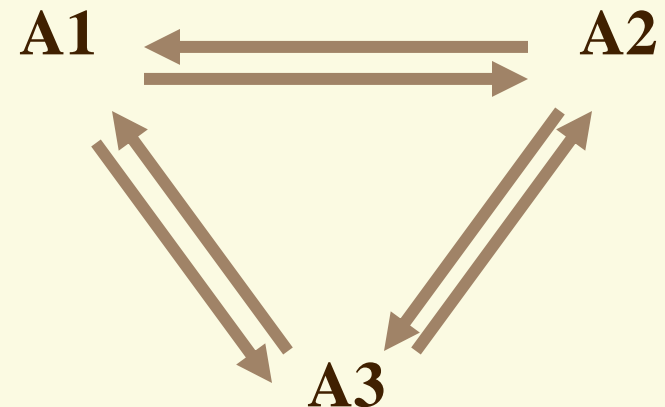
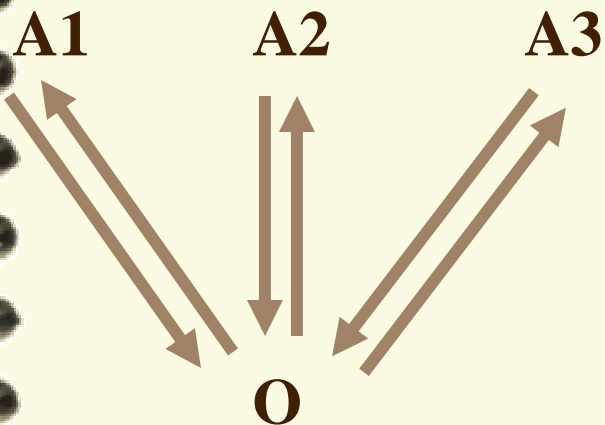
A design for 6 types of mRNA samples

Graphical representation

- ❖ The structure of the graph determines which effects can be estimated and the precision of the estimates.
 - Two mRNA samples can be compared only if there is a **path** joining the corresponding two vertices.
 - The precision of the estimated contrast then depends on the **number of paths** joining the two vertices and is inversely related to the **length of the paths**.
- ❖ Direct comparisons **within slides** yield more precise estimates than indirect ones between slides.

Comparing K treatments

(i) Common reference design (ii) All-pairs design



Question: Which design gives the most precise estimates of the contrasts A1-A2, A1-A3, and A2-A3?

Comparing K treatments

- ❖ **Answer:** The all-pairs design is better, because comparisons are done **within slides**.

For the same precision, the common reference design requires three times as many hybridizations as the all-pairs design.

- ❖ In general, for K treatments

Relative efficiency

$$= 2K/(K-1) = 4, 3, 8/3, \dots \rightarrow 2.$$

For the same precision, the common reference design requires $2K/(K-1)$ times as many hybridizations as the all-pairs design.

Experimental design

- ❖ In addition to experimental constraints, design decisions should be guided by the knowledge of which effects are of greater interest to the investigator.
E.g. which main effects, which interactions.
- ❖ The experimenter should thus decide on the comparisons for which he wants most precision and these should be made **within slides** to the extent possible.

Issues in experimental design

- ❖ Replication.
- ❖ Type of replication:
 - *within or between* slide replicates;
 - *biological or technical* replicates (different vs .same extraction): generalizability vs. reproducibility.
- ❖ Sample size and power calculations.
- ❖ Dye assignments.
- ❖ Combining data across slides and sets of experiments:
regression analysis ... next.

2 x 2 factorial experiment

Study the **joint** effect of two treatments (e.g. drugs), A and B, say, on the gene expression response of tumor cells.

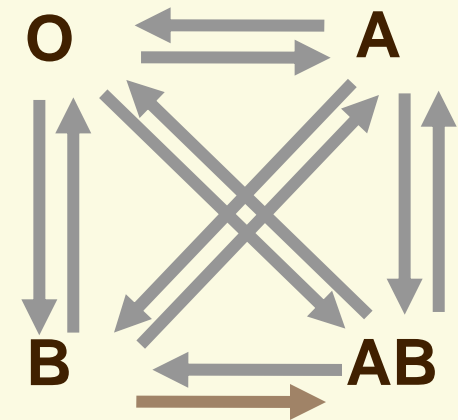
There are four possible treatment combinations

AB: both treatments are administered;

A : only treatment A is administered;

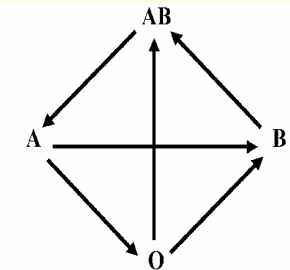
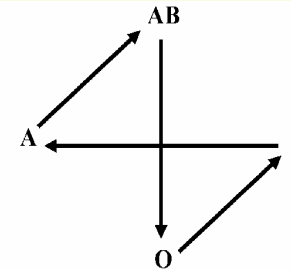
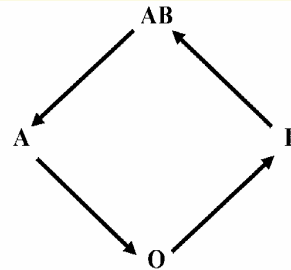
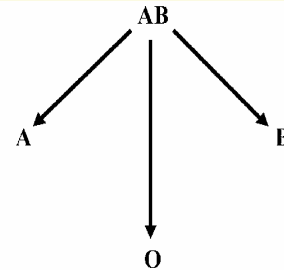
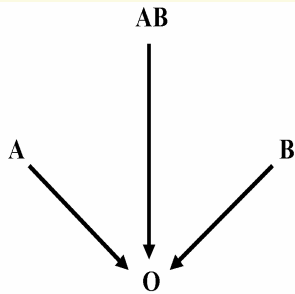
B : only treatment B is administered;

O : cells are untreated.



2 x 2 factorial experiment

two factors, two levels each

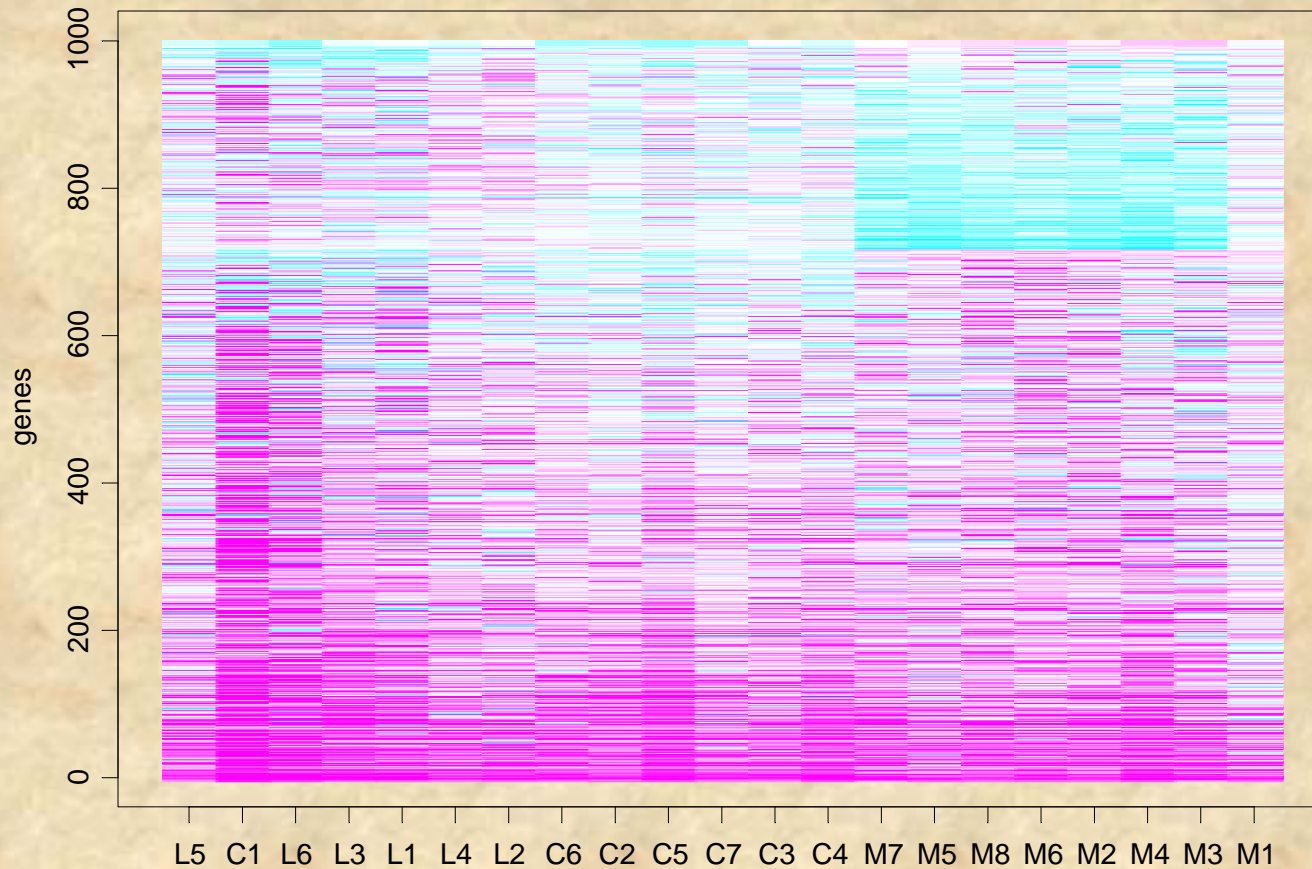


(1) Common ref. (2) Common ref. (3) Connected (4) Connected (5) All-pairs

Scaled variances of estimated effects

	(1)	(2)	(3)	(4)	(5)
Main Effect A	1	2	1	4/3	1
Main Effect B	1	2	1	1	1
Interaction AB	3	3	4/3	8/3	2
Contrast A-B	2	2	4/3	1	1

Analysis of Microarray Data



Overview

- ❖ Differential gene expression: identify genes whose expression levels are associated with an outcome or covariate
 - Estimation of parameter of interest for the joint distribution of microarray results and responses/covariates
 - For each gene, test a null hypothesis concerning a gene-specific parameter of interest
 - Estimating sensitivity/specificity of method used to select genes.
- ❖ Parameters of interest include means, differences in means, interactions, correlations, slope coefficients, survival probabilities etc

Selecting Differentially Selected Genes

- ❖ Typically one of the first goals of analyzing microarray data is to find the set of differentially expressed genes.
- ❖ This includes "*significantly*" over or under-expressed (relative to comparison tissue results).
- ❖ The first step then is to decide what is "significant"
 - Serious multiple hypothesis testing problem

Example—Alizadeh et al. (2000)

- ❖ Using gene expression profiling to distinguish types of B-cell lymphoma
- ❖ Expression levels of 13,412 "genes", relative to a pooled control, were measured in blood samples of 40 patients using cDNA arrays
- ❖ Patients represent two molecularly distinct disease groups: Activated (n=21) and Germinal Center (n=19)
- ❖ Publicly available data is logged (base 2), values over 20 (fold over expression) are reassigned the value $\log_2 20$, and missing values imputed.
- ❖ Goal: Identify "genes" with significantly different mean expression levels between the Activated and

Data Notation

- ❖ Suppose there are p genes (typically as large as 5,000)
- ❖ Let the gene expression (W_i) and covariate/outcome (Z_i) for the i^{th} person be given

by
$$\vec{X}_i = (W_i, Z_i) = (X_{1i}, X_{2i}, \dots, X_{pi})$$

- ❖ The overall data then for a sample of n individuals is

$$\vec{X}_1, \vec{X}_2, \dots, \vec{X}_n$$

- ❖ The units are different subjects (or plants, cells, etc). These replicates are assumed to be *independent*.

Parameters of Interest

- ❖ Parameters are functions of the unknown data generating distribution
 - Means: $E(W_{ji}) = (\mu_j : j = 1, \dots, p)$
 - Regression coefficients of W_i on Z_i , e.g slope coefficients
 - Logistic regression coefficients of Z_i on W_i , if Z_i is a binary group indicator
 - Cox proportional hazards coefficient if Z_i is a (censored) survival time

Log Transforming the Data

- ❖ Because relative expression data is typically skewed to the right, the data are typically log-transformed.
- ❖ Occurs because outcomes for under-expressed genes are squeezed between 0 and 1, whereas over-expressed genes vary from 0 to infinity.
- ❖ Logging the data makes the resulting distribution symmetric (closer to normal)
- ❖ Also, data is also usually truncated to remove large + and - values after truncation.

Example Data

- ❖ 10 breast cancer lines, 9,600 genes
- ❖ Log relative expression (to referent cell line).

Log of the Relative Expression of Breast Cancer Cell Lines								
Gene Number	Line 1	Line 2	Line 3	Line 4	Line 5	Line 6	Line 7	Line 8
13	-0.47804	-0.45459	-0.28768	-0.45459	-2.20728	-0.45459	-0.31471	-1.07881
15	-0.04789	-0.04789	-0.04789	2.075685	-0.69315	-0.04789	-1.30933	0.285179
17	1.735189	-0.00493	-0.00493	-1.42712	-0.00493	-0.00493	-0.73397	-1.30933
19	0.139164	0.139164	-0.73397	0.139164	0.139164	0.139164	1.202972	-1.77196
21	-1.10866	-0.77673	-0.77673	-0.77673	-0.77673	-0.77673	-0.84397	-1.51413
31	-0.4943	0.29267	-0.53253	-0.53253	-3.91202	-0.15082	-0.53253	0.398776
32	1.20896	-0.23954	-0.23954	-1.66073	-0.84397	-0.8675	-0.23954	0.262364
34	0.71784	0.158342	0.158342	-0.21072	0.158342	0.158342	0.158342	-0.15082
37	0.530628	0.031922	0.173953	0.031922	-2.40795	0.031922	-1.83258	3.135494
38	-0.47804	1.558145	0.65752	0.63485	0.63485	0.63485	0.457425	-0.27444

Simple Gene Selection Rules

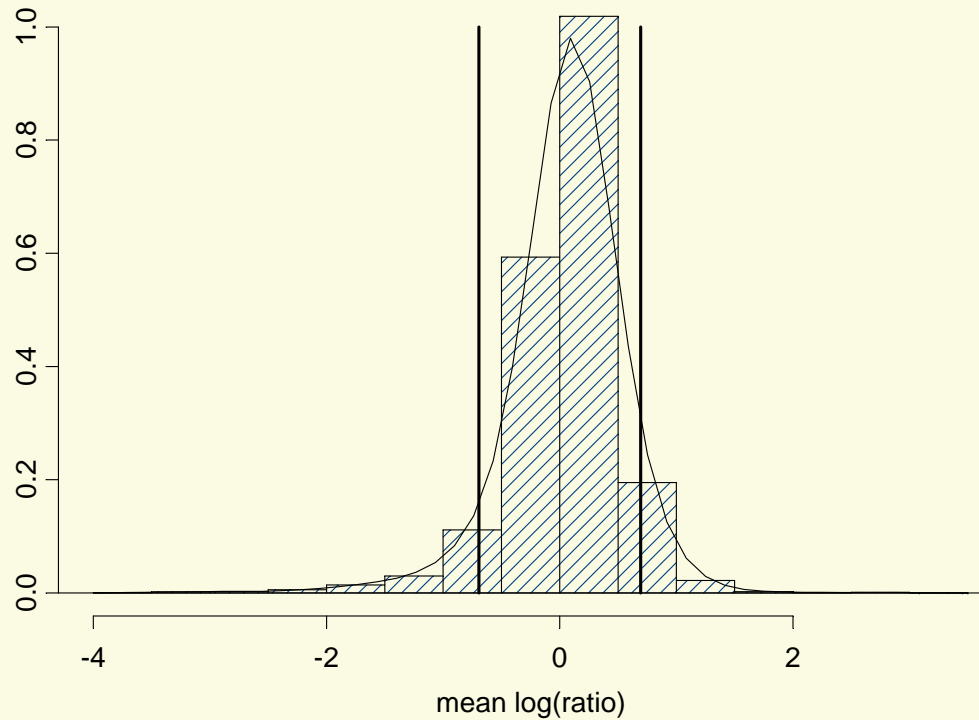
- ❖ Let $Y_{ji} = \log(W_{ji})$.
- ❖ Simplest rules use the average relative expression across units to decide on whether a gene is differentially expressed (from now on, we will assume the data has been logged).
- ❖ Let μ_j be the average relative expression for the j th gene, $j=1, \dots, p$
- ❖ Then, $\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^n Y_{ji}$ is the average relative gene expression for gene j across the n units.

Simple Gene Selection Rules

- ❖ Let $\hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_p)$ be the list of averages of each gene across units.
- ❖ Then, choose simply those genes, j , with absolute expression value $>$ than a chosen cut-off, e.g.

$$|\hat{\mu}_j| > \log(2)$$

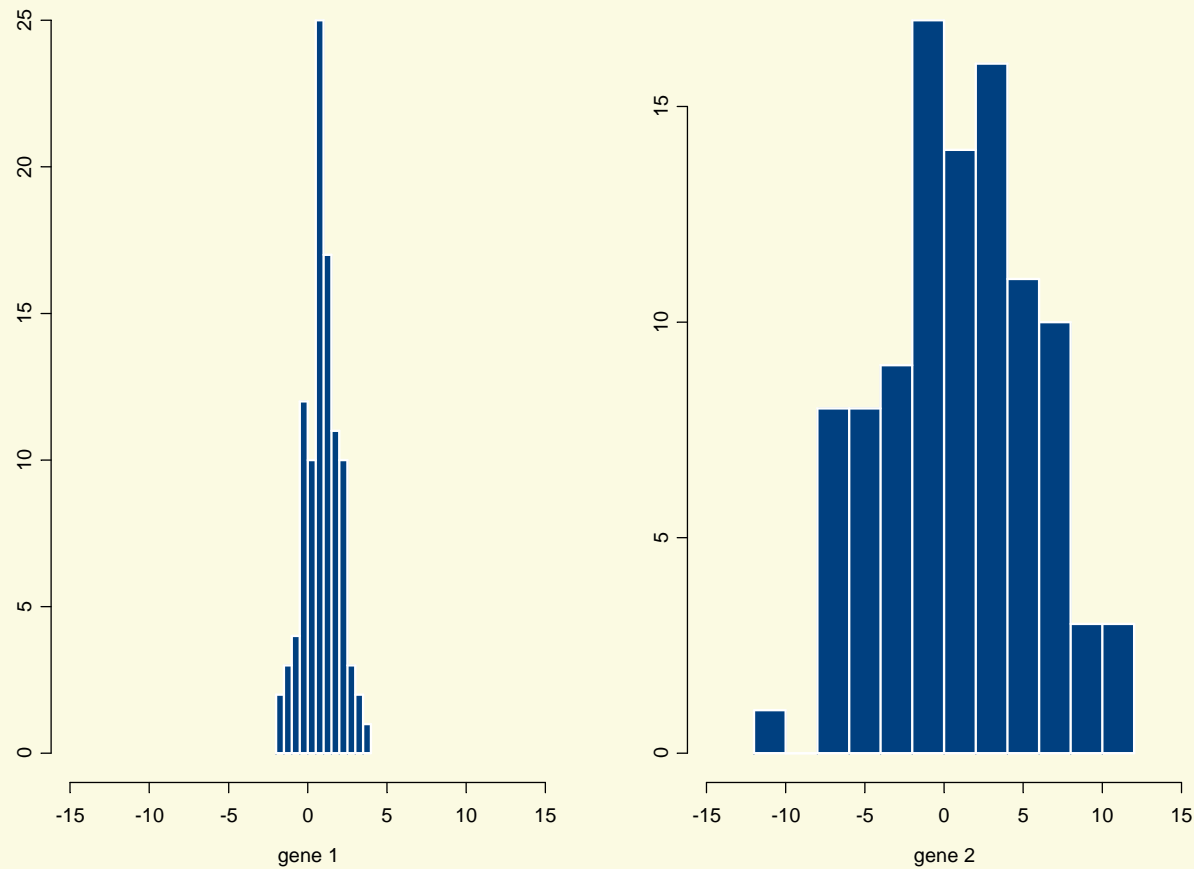
Selecting Differentially Expressed Genes: fixed cut-off



Accounting for Variability

- ❖ One problem is that just using the average expression ignores the between unit variability.
- ❖ Thus, one treats equivalently genes with the same average expression even if the variability across the units is very different.
- ❖ Intuitively, one should select for further research, with greater probability, those genes that are differentially and consistently expressed (less variability) across units.

Equivalent average expression with different variability



Using z-statistics to subset genes

- ❖ One way to incorporate variability is, for each gene, to calculate the z-statistic.
- ❖ Let $\hat{\sigma}_j$ be the sample standard deviation of the jth gene across the n units:

$$\hat{\sigma}_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (Y_{ij} - \hat{\mu}_j)^2}$$

Using z-statistics to subset genes

- ❖ The z-score is (for $H_0: \mu_j=0$) is (two-sided test):

$$z_j = \frac{\sqrt{n}\hat{\mu}_j}{\hat{\sigma}_j}$$

- ❖ Test statistic that relative log-expression is 0 or relative expression is 1.
- ❖ Choose a cut-off based on the standard Normal distribution and a particular error-rate, α , for a *single* test.
- ❖ $\alpha = P(\text{Type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true})$
 $= P(\text{false positive}).$

Choosing α , the Error rate

- ❖ α is set by the researcher deciding how many Type I errors they are willing to accept.
- ❖ If one wishes to be more aggressive in investigating potentially differentially expressed genes, then choose α relatively large (e.g., 0.20).
(will have implications later)
- ❖ Conversely, if one wishes to be more conservative, then choose α to be small (e.g., 0.001).

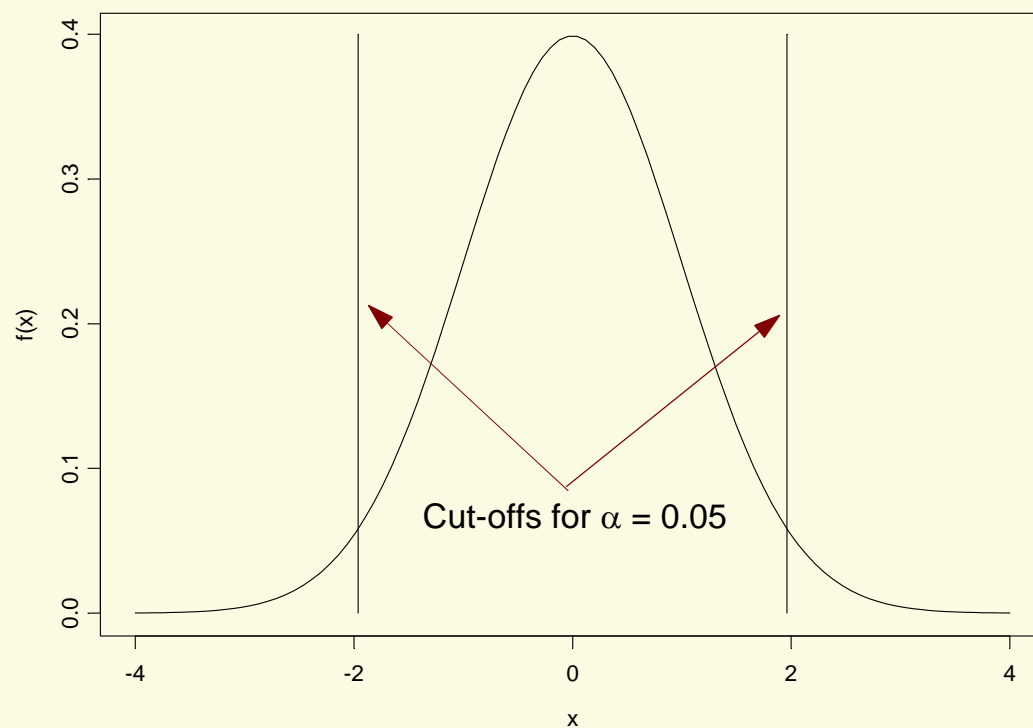
Statistical Subsetting Rule

- ❖ Rule is to reject null (accept as differentially expressed) if:

$$\frac{\sqrt{n}|\hat{\mu}_j|}{\hat{\sigma}_j} > z_{1-\alpha/2}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard Normal distribution

Standard Normal Distribution



Multiple Hypothesis Testing

- ❖ Large multiplicity problem: thousands of hypotheses are tested simultaneously!
 - Increased chance of false positives,
 - e.g. chance of at least one p-value $< \alpha$ for m independent tests is $1 - (1 - \alpha)^m \approx m\alpha$ for small α , even when all nulls are true, and converges to one as m increases.
For $m=1,000$ and $\alpha = 0.01$, this chance is 0.9999568!
 - Individual p-values of 0.01 no longer necessarily correspond to significant findings.
- ❖ Need to **adjust for multiple testing** when assessing the statistical significance of findings, or, more pragmatically, reduce the number of false positives. Have to consider the joint distribution of the test statistics

Multiple Testing-Bonferroni Method

❖ Specifically, set the individual type I error rates, say the α_j 's, to control the probability of making any type I errors (FWER).

❖ Suppose

$$P(E_j) = P(\text{Type I error is not made in } j\text{th test}) = 1 - \alpha_j$$

❖ No false positives:

$$1 - \alpha_0 = P\left(\bigcap_{j=1}^p E_j\right) \geq 1 - \sum_{j=1}^p P(\bar{E}_j) = 1 - \sum_{j=1}^p \alpha_j$$

Multiple Testing-Bonferroni Method

- ❖ Therefore, if $\alpha_j \equiv \alpha$, $\alpha_0 \leq 1-(1-p\alpha)=p\alpha$
- ❖ So, using $\alpha_j = \alpha = \alpha_0/p$ will achieve desired overall level α_0 .
- ❖ Thus, reject the null: $H_0: \mu_j=0$ if:

$$\frac{\sqrt{n}|\hat{\mu}_j|}{\hat{\sigma}_j} > z_{1-\frac{\alpha}{2p}}$$

- ❖ Conservative if genes are statistically dependent.

Multiple testing

- ❖ 1. Define the appropriate set of null hypotheses. What are the parameters of interest?
- ❖ 2. Compute an appropriate test statistic for each null hypothesis
- ❖ 3. Select an appropriate **Type I error** or **false positive rate** (now many possibilities: FWER, PCER, FDR).
- ❖ 4. Develop a multiple testing procedure based on the joint distribution of the test statistics that
 - provides **control** of the selected error rate,
 - are **powerful** (few false negatives),
- ❖ 5. Use **resampling** (bootstrap or permutation) methods to estimate the unknown (null) joint distribution of the test statistics.
- ❖ 6. Report **adjusted p-values** for each gene which reflect the **overall Type I error rate** for the experiment.

1. Parameters of Interest

❖ Location parameters: means, medians, differences in means etc

- e.g. assess (differential) expression for p genes
 $\mu_1(j)$ is mean expression level in Population 1

Then $\mu(j) = \mu_2(j) - \mu_1(j)$ is difference in mean expression level for gene j between Populations 1 & 2

- Scale parameters: covariances, correlations. For example, pairwise correlations for $p(p-1)/2$ pairs of genes

1. Parameters of Interest

- Regression parameters: slopes, main effects, interactions

e.g. $\mu(j)$ = parameter of univariate logistic regression coefficient for gene j on outcome (expression is explanatory variable here)

$\mu(j)$ = interaction effect of two drugs on expression level of gene j (expression is outcome variable here)

$\mu(j)(w) = \Pr(\text{survival} > 2 \mid \text{expression of gene } j = w)$

2. Test Statistics/Null Hypotheses

- ❖ H_{0j} is a null hypothesis relative to gene j
 H_{Aj} is an alternative hypothesis relative to gene j

- e.g. single parameter hypotheses

$$H_{0j} : \mu(j) = 0 \quad H_{Aj} : \mu(j) \neq 0$$

- multiple parameter hypotheses

$$H_{0j} : \mu_1(j) = \mu_2(j) = \cdots \mu_K(j) \quad K > 2 \text{ populations}$$

$$H_{Aj} : \text{at least one of } \mu_1(j), \mu_2(j), \dots, \mu_K(j) \text{ is different}$$

2. Test Statistics/Null Hypotheses

- ❖ To “build” a test statistic, we use the strategy of estimating the parameter of interest and then comparing the estimate with the hypothesized null value

So, for $H_{0j} : \mu(j) = 0$

- Difference statistics $T_n(j) = \sqrt{n}(\hat{\mu}_n(j) - \mu_0(j)) = \sqrt{n}\hat{\mu}_n(j)$

- *t*-statistics $T_n(j) = \frac{\sqrt{n}(\hat{\mu}_n(j) - \mu_0(j))}{\sigma_n(j)} = \frac{\sqrt{n}\hat{\mu}_n(j)}{\sigma_n(j)}$

where $\sigma_n(j)/\sqrt{n}$ is an estimate of the sd of $T_n(j)$

2. Test Statistics/Null Hypotheses

- ❖ Rejection decisions are based on a p -vector of test statistics, one component for each hypothesis: $\{T_n(j) : j = 1, \dots, p\}$ where we assume that large values of $T_n(j)$ provide evidence against H_{0j}
 - e.g. Reject H_{0j} if $T_n(j) > c_j$ or if $|T_n(j)| > c_j$ for two-sided tests
 - multiple parameter hypotheses
 H_{Aj} : at least one of $\mu_1(j), \mu_2(j), \dots, \mu_K(j)$ is different

2. Examples of Test Statistics

❖ Population mean parameters:

$\mu(j) = E(Y_j)$, the mean (log) expression level for gene j

$$H_{0j} : \mu(j) = 0 \quad \longrightarrow \quad T_n(j) = \frac{\sqrt{n}\bar{Y}_n(j)}{\sigma_n(j)}$$

$$\bar{Y}_n(j) = \frac{1}{n} \sum_i Y_{ji}$$

$$\sigma_n^2(j) = \frac{1}{n} \sum_i (Y_{ji} - \bar{Y}_n(j))^2$$

2. Examples of Test Statistics

- ❖ Two-sample t -statistics (comparing mean gene expression levels across two populations):

$$H_{0j} : \mu_1(j) = \mu_2(j) \longrightarrow T_n(j) = \frac{\bar{Y}_{2,n}(j) - \bar{Y}_{1,n}(j)}{\sqrt{\sigma_{1,n}^2(j)/n_1 + \sigma_{2,n}^2(j)/n_2}}$$

2. Examples of Test Statistics

- ❖ K -sample F -statistics (comparing mean gene expression levels across K populations):

$$H_{0j} : \mu_1(j) = \mu_2(j) = \cdots \mu_K(j)$$

→
$$T_n(j) = \frac{\frac{1}{K-1} \sum_{k=1}^K n_k (\bar{Y}_{k,n}(j) - \bar{Y}_n(j))^2}{\frac{1}{n-K} \sum_{k=1}^K \sum_{i=1}^{n_k} (Y_{k,i}(j) - \bar{Y}_{k,n}(j))^2}$$

3. Type I Error (False Positive) Rates

- ❖ A multiple testing procedure produces a set S_n of rejected hypotheses that estimates the set of actual false nulls
- ❖ S_n depends on
 - The data $\vec{X}_1, \vec{X}_2, \dots, \vec{X}_n$ through the test statistics $T_n(j)$
 - The cut-off values c_j
 - The desired level of false positive error rate α

3. Type I and Type II Errors

	Non-rejected hypotheses	Rejected hypotheses	
True null hypotheses	T	V Type I error	M_0
False null hypotheses	U Type II error	W	$p - M_0$
	p-r	r	p

Upper case: unobservable parameter

3. False Positives and Associated Error Rates

- ❖ V_n is a random variable—the (unobserved) number of false positives—with distribution F_n , the properties of which determine various error rates (nb this distribution depends, of course, on the true distribution of the data)
 - **Per-comparison error rate (PCER).** The PCER is defined as the expected value of (# false positives / # of hypotheses), i.e.,

$$\text{PCER} = E(V_n)/p.$$

3. False Positives and Associated Error Rates

- Family-wise error rate (FWER). The FWER is defined as the probability of at least one false positive, i.e.,

$$\text{FWER} = \text{pr}(V_n > 0) = 1 - F_n(0).$$

- Generalized family-wise error rates (gFWER) is the probability of at least $k+1$ false positive

$$\text{gFWER} = \text{pr}(V_n > k) = 1 - F_n(k).$$

3. False Discovery Rate

- ❖ False discovery rate (FDR). The FDR of Benjamini & Hochberg (1995) is the expected proportion of false positives among the rejected hypotheses, i.e.,

$$\text{FDR} = E(V_n / r_n),$$

where by definition

$$V_n / r_n = 0 \text{ if } r_n = 0$$

- The FDR is not a property of F_n because it also depends on r_n , the random number of rejections

4. Choosing Cut-Offs to Control Error Rate

- ❖ Given our preferred error rate, we now want to choose the cut-off values c_j so that our rate $\leq \alpha$ (preferably = α to maximize power)
- ❖ But we don't know the distribution of V_n , in fact we don't even observe its value!
- ❖ We therefore create an "approximate" world where we can observe V_n , and prove that this is conservative, that is, that the error rate in the approximate world is greater than in the true world

4. Choosing Cut-Offs to Control Error Rate

- ❖ A suitable "approximate" world (where we can observe V_n) is the case where all null hypotheses are true and the rest of the distribution of $\{T_n(j) : j = 1, \dots, p\}$ is the same as the true world
- ❖ Let's call this "null" distribution Q_0
- ❖ We still don't know Q_0 so we have to estimate it somehow

4. Choosing Cut-Offs by Estimating Q_0

❖ Methods for estimating Q_0

- Bootstrap null distribution
- Parametric limit distribution, e.g. $Q_0 = N(0, \Sigma)$
 - Still have to estimate Σ
- Data generating null distribution for Y , e.g. permutation distribution---not recommended

5. Using the Bootstrap to Define Cut-off

- ❖ Another method of determining statistical cut-offs uses a simulation technique called bootstrapping.
- ❖ Bootstrapping works by re-sampling the original vectors of genes $(\vec{Y}_1, \vec{Y}_2, \dots, \vec{Y}_n)$ at random and with replacement.
- ❖ It uses the re-sampled data to determine quantities that can't be easily estimated directly from the data itself.

5. Using the Bootstrap to Select Cut-off Values

- ❖ Generate B bootstrap samples: $(\vec{Y}_1^b, \vec{Y}_2^b, \dots, \vec{Y}_n^b)$
- ❖ For each bootstrap sample, compute a p -vector of test statistics $\{T_n^b(j) : j = 1, \dots, p\}$
arranged in a matrix where rows correspond to the p hypotheses and columns to the B bootstrap samples
- ❖ Compute row means of this matrix to obtain estimate $\hat{E}(T_n(j))$

5. Using the Bootstrap to Select Cut-off Values

- ❖ Row shift the matrix by computing

$$T_n(j) - \hat{E}(T_n(j))$$

so that each row now has average 0 across the bootstrap samples (the null distribution)

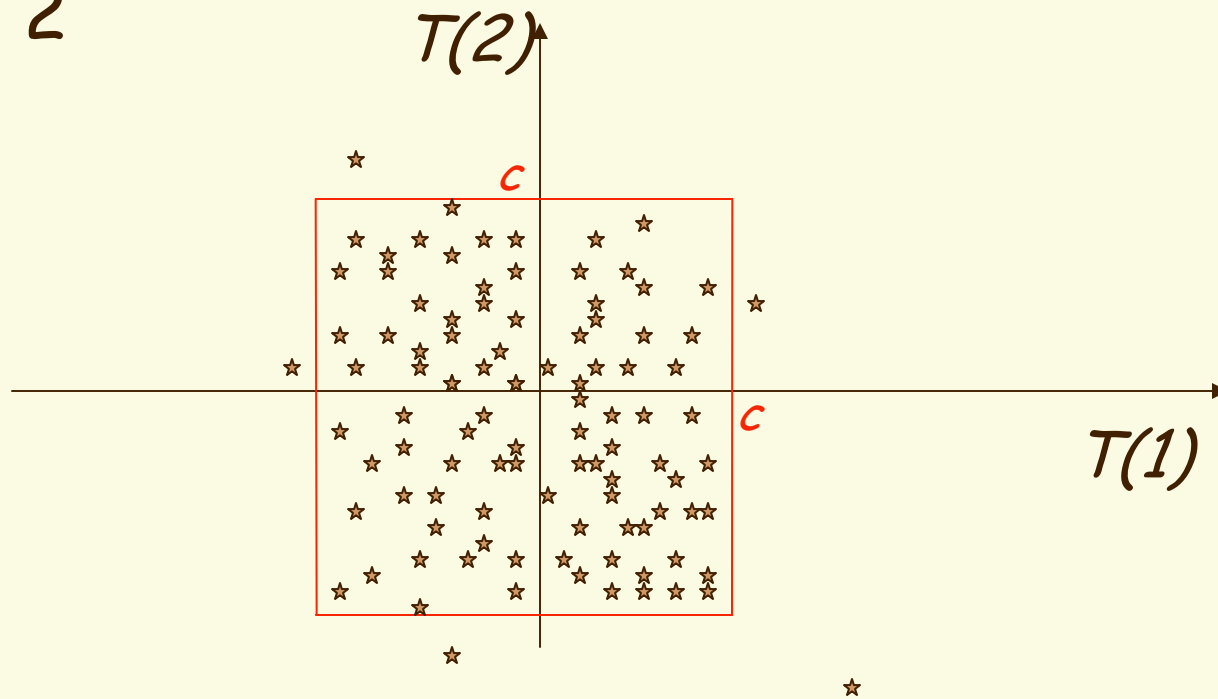
- ❖ The bootstrap estimate of the null distribution Q_0 is now just the empirical distribution of the columns of the matrix of bootstrapped values $\{T_n^b(j) : j = 1, \dots, p\}$

5. Common Cut-Offs

- ❖ Remember that the cut-offs play the following role:
 - Reject H_{0j} if $|T_n(j)| > c_j$ for two-sided tests
- ❖ For control of *FWER* we can select all $c_j \equiv c$ with c having the property that
$$c_0 = \min\{c: \Pr(V_0 > 0) \leq \alpha\}$$
where V_0 is # of (false) positives based on bootstrap estimate of Q_0

5. Common Cut-Offs

$p = 2$



★ points of mass of bootstrapped Q_0

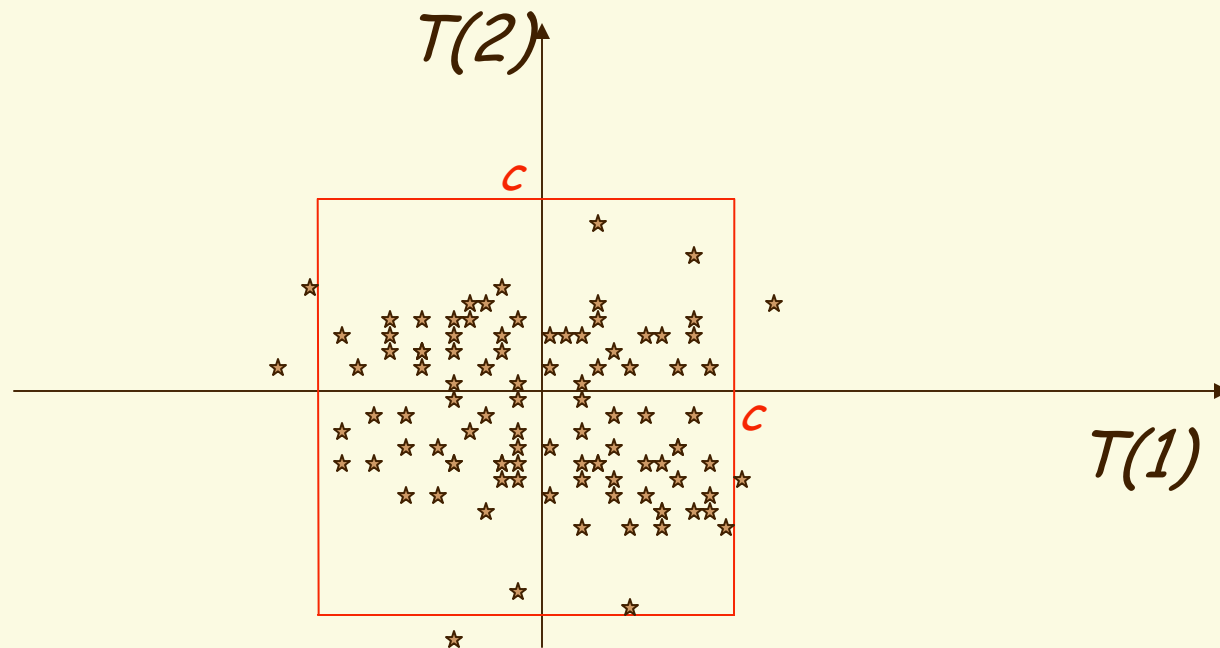
5. Common Cut-Offs

- ❖ In fact to select common cut-offs with FWER, all we need is an estimate of the null distribution of the *maximum* of $\{T_n(j) : j = 1, \dots, p\}$; in particular the $100(1-\alpha)\%$ -tile.
- ❖ From the bootstrapped samples, we can take this common cut-off c to be the $100(1-\alpha)\%$ -tile of the bootstrapped

$$V_b = \max |T_n^b(j)| \text{ over } j = 1, \dots, p$$

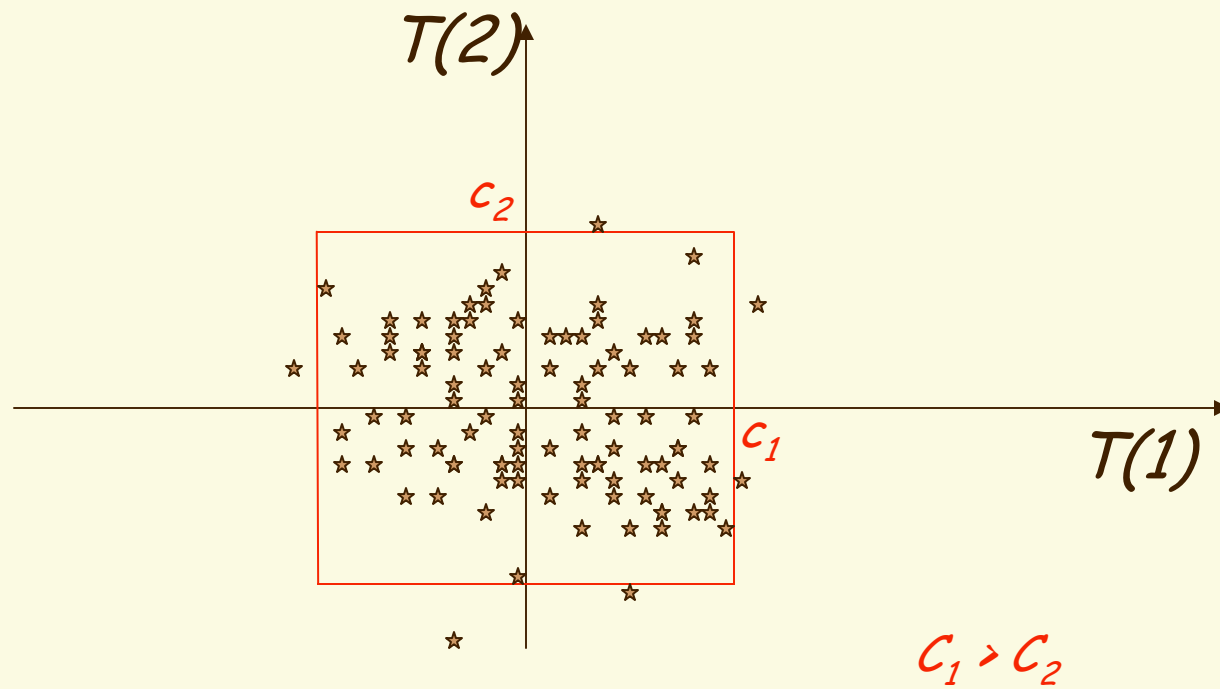
5. Common Cut-Offs

$p = 2$ (different variances)



common cut-offs not so sensible

5. Common Quantiles



$c_j = 1-\delta$ quantile of j^{th} marginal of Q_0

6. Adjusted p-values

- ❖ Given any test procedure, the **adjusted p-value** for a single gene g can be defined as the level of the **entire** test procedure at which the j^{th} test (e.g. for whether gene j is differentially expressed) would just be rejected.
- ❖ Adjusted p-values reflect for each gene the **overall experiment Type I error rate** when genes with a smaller p-value are declared differentially expressed.
- ❖ Can be estimated by **resampling** (bootstrap).

6. Adjusted p -values

❖ Common Cut-Offs

- Use common cut-offs $(c_1, \dots, c_p) = (T_n(j), \dots, T_n(j))$

❖ Common Quantiles

- Calculate the percentile corresponding to $T_n(j)$ for the j^{th} test using the j^{th} marginal of the estimated null distribution Q_0 . For example, $T_n(j)$ might be the 98%-tile. Then choose c_1 to be the same (e.g. 98%-tile) of the first marginal of Q_0 etc.

Step-Down Methods

- ❖ So far, choice of cut-offs only depend on (estimated) null distribution Q_0 and level α of our selected error rate
- ❖ Step-down (and step-up) methods allow the cut-offs to also depend on the test statistics $\{T_n(j) : j = 1, \dots, p\}$
 - Step-down: start with most significant hypothesis; as soon as one fails to reject the next most significant hypothesis, you stop

Step-Down Methods

- ❖ Most easy to get idea by considering the step-down method applied to Bonferroni method
 - Order test statistics by size of their unadjusted p-values
 - Starting with the test statistic with the smallest p-value, reject this null hypothesis if the p-value $\leq \alpha/p$
 - If rejection, go to the test statistic with the next smallest p-value and reject if this is $\leq \alpha/(p-1)$
 - Keep going until a hypothesis is accepted and then stop
- ❖ Advantage: more power—has FWER exactly α

Step-Down Methods

❖ Common cut-offs:

- Order the test statistics $T_n(1) \leq \dots \leq T_n(p)$
- Take c_1 to be the $100(1-\alpha)\%$ -tile of the maximum of $\{T_n(j) : j = 1, \dots, p\}$
- If the first test is rejected based on $T_n(1)$, then, take c_2 to be the $100(1-\alpha)\%$ -tile of the maximum of $\{T_n(j) : j = 2, \dots, p\}$
- If the second test is rejected based on $T_n(2)$, then ... etc, otherwise stop.

❖ Can calculate adjusted p-value based on this procedure step-down common quantile approach

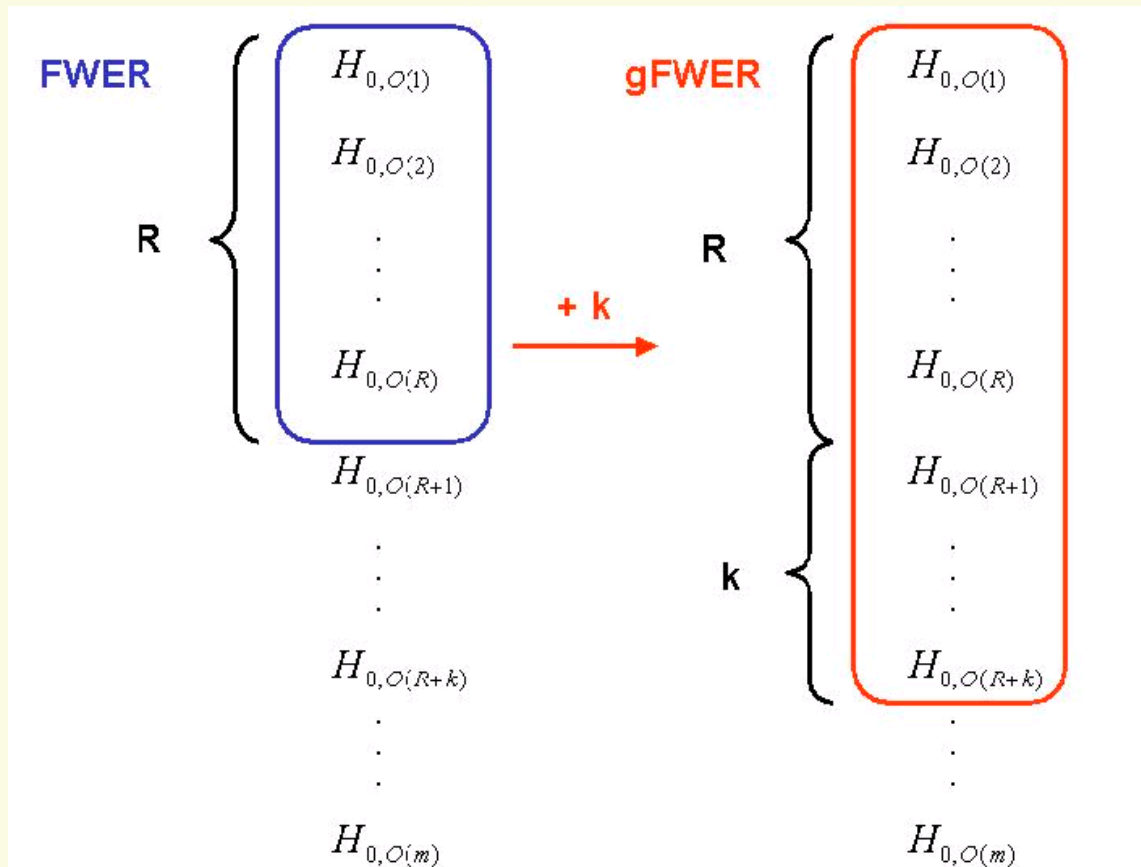
Augmentation Procedures for *g*FWER and PFP Control

- ❖ Take the set of m rejected nulls under FWER at error rate α .
 - For *g*FWER, take the k hypotheses with the next lowest adjusted p-values (under FWER)
 - For PFP at error rate q , keep adding the next most significant p-values but stop just before

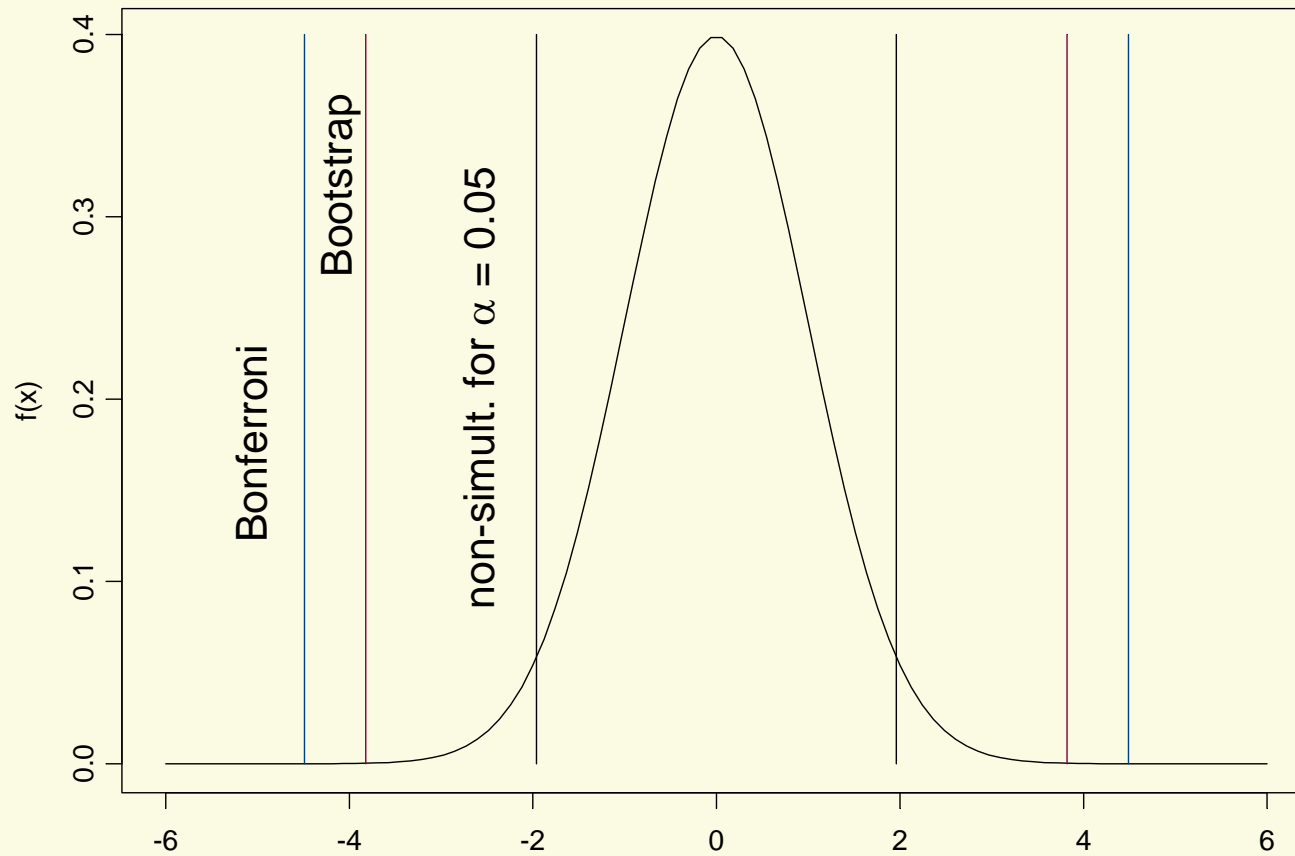
$$\frac{\text{\# of augmented rejections}}{\text{\# of augmented rejections} + m} > q$$

- ❖ Easy to calculate adjusted p-values for these procedures

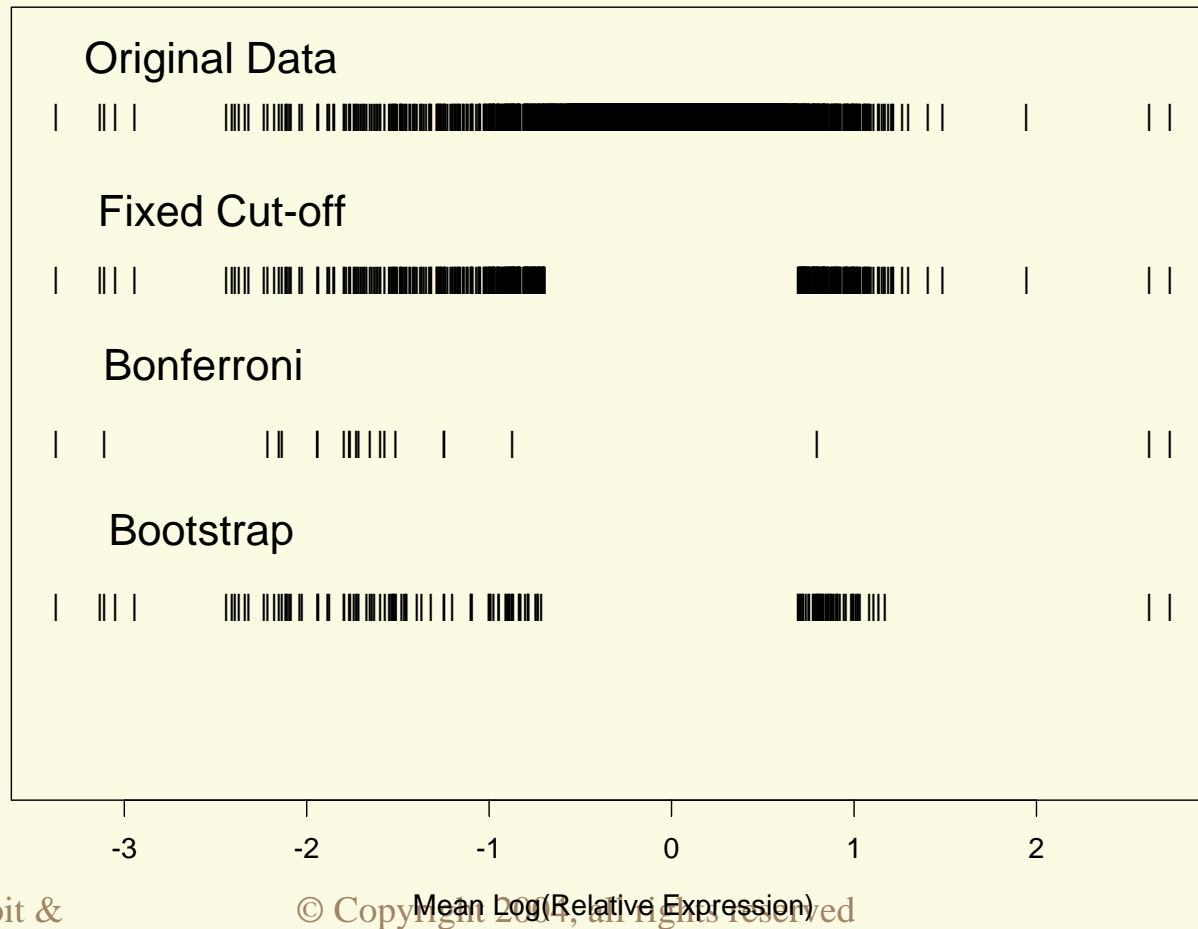
Control of gFWER using control of FWER



Comparing Methods on Standard Normal Curve



Comparing Methods



Example—Alizadeh et al. (2000)

- ❖ Goal: Identify “genes” with significantly different mean expression levels between the Activated and GC groups.

- $H_{0j} : \mu_1(j) = \mu_2(j); \quad j = 1, \dots, 13412$

- ❖ Test statistics: two-sample t-statistics

- $$T_n(j) = \frac{\bar{Y}_{2,n}(j) - \bar{Y}_{1,n}(j)}{\sqrt{\sigma_{1,n}^2(j)/n_1 + \sigma_{2,n}^2(j)/n_2}}$$

Example—Alizadeh et al. (2000)

❖ Error rate: Family-wise error rate (FWER).

$$\text{FWER} = \text{pr}(V_n > 0)$$

❖ Results of methods:

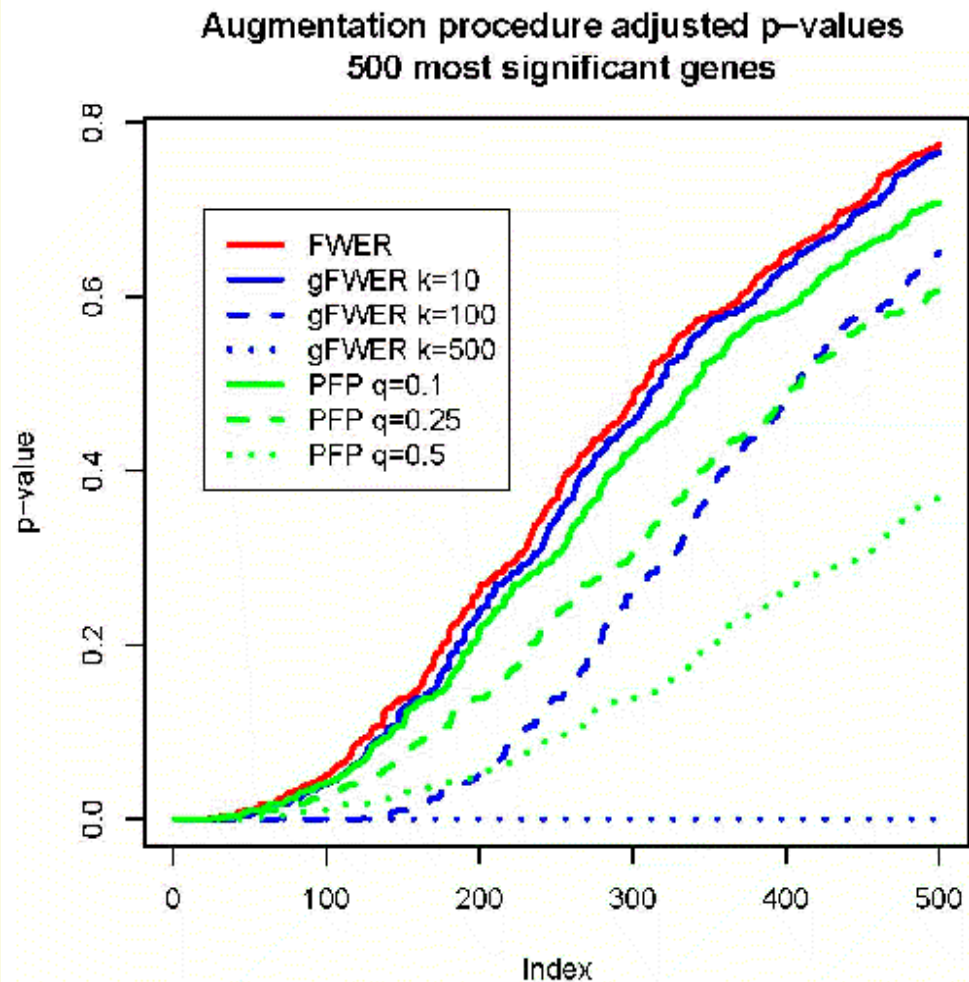
- Common quantiles with bootstrap null distribution (B=1000)
 - 186 rejections
- Bonferroni with t null distribution
 - 32 rejections (all contained within the 187 rejections above)

Example—Alizadeh et al. (2000)

*Common cut-off
 $B=5000$*

*Augmentation
procedures for
both gFWER and
PFP*

Sandrine Dudoit &
Nicholas P. Jewell



Sensitivity and Specificity of Selection Procedure

- ❖ Another potential use of the bootstrap is to estimate the repeatability of gene selection procedure.
- ❖ Two possible measures of this are sensitivity and specificity.
- ❖ Sensitivity = probability of selecting a gene as differentially expressed given that it is "truly" differentially expressed (true positive).
- ❖ In short hand, $P(+|DE)$, where + means procedure selects gene and DE = differentially expressed.
- ❖ Specificity is $P(- | \text{not DE})$, i.e., probability of not selecting gene given it is not differentially expressed (true negative).

Sensitivity and Specificity of Selection Procedure

- ❖ Want both sensitivity and specificity to be high (few false positives and few false negatives).
- ❖ Can not estimate them directly from data.
- ❖ Bootstrapping offers some hope.

Formal Definition of Sensitivity

❖ For gene j , let

$S(j) = \text{DE}$ if j differentially expressed

$S(j) = \text{not DE}$ if j not differentially expressed.

$\hat{S}(j) = +$ if selection rule chooses j as DE

$\hat{S}(j) = -$ if rule chooses j as not DE

❖ **Sensitivity is:**

$P(\hat{S}(j) = + \mid S(j) = \text{DE})$ is estimated by

$\frac{\text{\# of genes with both } \hat{S}(j) = + \text{ and } S(j) = \text{DE}}{\text{\# of genes } S(j) = \text{DE}}$

Using Bootstrapping to Estimate Sensitivity

- ❖ Select the units $(\vec{Y}_1, \vec{Y}_2, \dots, \vec{Y}_n)$ at random and with replacement.
- ❖ For each realization, use a procedure to select differentially expressed genes (e.g., fixed cut-off).

$\hat{S}^b(j)$ = bootstrap sample b assignment of gene j

- ❖ Estimate, for each bootstrap sample, the sensitivity as:

$$P(\hat{S}^b(j) = + \mid \hat{S}(j) = +) =$$

$$\frac{\text{\# of genes with both } \hat{S}^b(j) = + \text{ and } \hat{S}(j) = +}{\text{\# of genes } \hat{S}(j) = +}$$

Using Bootstrapping to Estimate Sensitivity

- ❖ This is repeated B times ($b=1,2,\dots,B$), typically something like 1000.
- ❖ Sensitivity is then estimated as the average of all the bootstrap estimates, or

$$\frac{1}{B} \sum_{b=1}^B P(\hat{S}^b(j) = + | \hat{S}(j) = +)$$

Using Bootstrapping to Estimate Specificity

- ❖ Equivalent to the calculation of sensitivity.

$$P(\hat{S}^b(j) = - | \hat{S}(j) = -) =$$

$$\frac{\text{\# of genes with both } \hat{S}^b(j) = - \text{ and } \hat{S}(j) = -}{\text{\# of genes } \hat{S}(j) = -}$$

- ❖ Specificity is then estimated as the average of all the bootstrap estimates, or

$$\frac{1}{B} \sum_{b=1}^B P(\hat{S}^b(j) = - | \hat{S}(j) = -)$$

Example with Breast Cancer Data

- ❖ Using fixed cut-off of $\log(2)$, the sensitivity was estimated to be only around 10% whereas the specificity was nearly 99%.
- ❖ Thus, even the fixed cut-off is a conservative procedure for this data set.

Software for multiple testing

- ❖ Bioconductor R **multtest** package
- ❖ Multiple testing procedures for controlling
 - FWER: Bonferroni, Holm (1979), Hochberg (1986), Westfall & Young (1993) maxT and minP.
 - FDR: Benjamini & Hochberg (1995), Benjamini & Yekutieli (2001).
- ❖ Tests based on t- or F-statistics for one- and two-factor designs.
- ❖ Permutation procedures for estimating adjusted p-values.
- ❖ Documentation: tutorial on multiple testing.

Further Applications

- ❖ We have considered mean gene expressions here. Can also look at:
 - Which genes predict certain outcome
 - Outcomes might include survival data
- ❖ Extend to gene expressions in different structured settings
 - Over time
 - In space

Summary---Lessons Learned

- ❖ Multiple comparisons issue relevant to selection of genes and interpretation.
- ❖ Bootstrapping procedure very effective in adding statistical inference and estimates of repeatability to microarray analyses.
 - Select differentially expressed genes.
 - Estimate sensitivity and specificity of selection procedures.