Unsupervised Analysis: High-Dimensional Hypothesis Testing

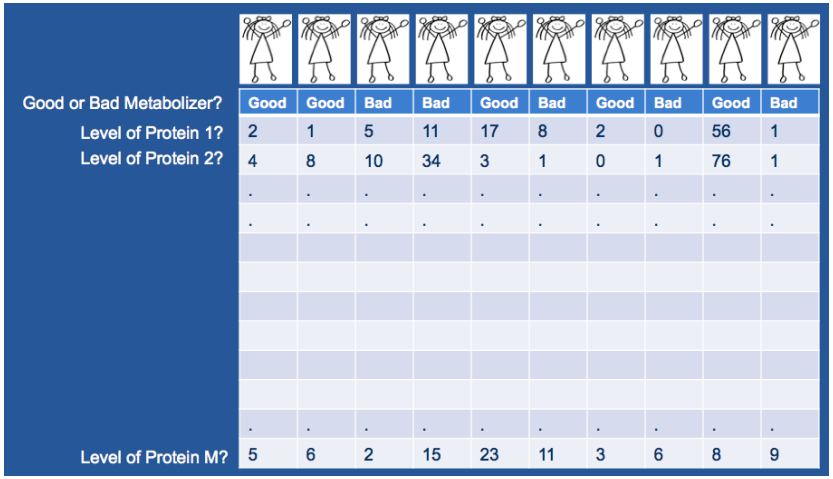
Hypothesis Testing

- High dimensional hypothesis testing
  - challenges of high dimensional hypothesis testing
  - controlling family-wise error rate (FWER)
  - controlling false discovery rate (FDR)
  - significance analysis of microarrays (SAM)

Hypothesis Testing

- This is very different from various methods for unsupervised learning in high-dimensional settings (clustering, dimension reduction, graphical models)
- It is a seemingly much simpler topic
- But hypothesis testing in high dimensions has its challenges!

Hypothesis Testing



Good or Bad Metabolizer?	Good	Good	Bad	Bad	Good	Bad	Good	Bad	Good	Bad
Level of Protein 1?	2	1	5	11	17	8	2	0	56	1
Level of Protein 2?	4	8	10	34	3	1	0	1	76	1
	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.
Level of Protein M?	5	6	2	15	23	11	3	6	8	9

Hypothesis Testing

- We have M features, each of which is measured in n observations.
- (In this lecture only, M is number of features, rather than p.)
- We also have a response vector of length n. Could be
  - blood pressure
  - tumor size
  - survival time
  - cancer subtype
- We wish to test the null hypothesis

$$H_{0j} : j\text{th feature is not associated with the response}$$

for $j = 1, \ldots, M$.

Testing One Hypothesis

Suppose we had only one hypothesis to test:

$$H_0 : \text{ feature is not associated with the response.}$$

The type of test that we use will depend on the type of response:

- binary response (e.g. cancer versus normal): two-sample t-test
- categorical response (e.g. cancer type 1 versus cancer type 2 versus cancer type 3): F-statistic for one-way ANOVA
- survival response (e.g. time to recurrence): score statistic for Cox proportional hazards model
- quantitative response (e.g. blood pressure): t-statistic for regression coefficient

Regardless of the response type, we get a test statistic and a p-value for the association between the feature and the response.

Quick Review: Test Statistics & P-Values

- Suppose we have measured the expression level of gene A in n patients, and want to test its association with the response, cancer versus normal.
- We compute a two-sample t-statistic quantifying its association with the response.
- Large (absolute) t-statistic: strong association with response.
- How big is big?
- The p-value is the probability of observing a t-statistic at least this large, under the null hypothesis of no association between the feature and the response.
- A p-value of 0.02 means that the probability of seeing such a strong association between the response and the feature by chance, under the null hypothesis, is 0.02.
- A small p-value indicates strong evidence for association: i.e. reject the null hypothesis.

Type I Error

- When we reject the null hypothesis if a p-value is below 0.05, we are controlling Type I Error at level $\alpha = 0.05$.
- Type I error occurs when we reject the null hypothesis when the null hypothesis actually holds.
- We are not required to control Type I error at level $\alpha = 0.05$. For instance, we could use $\alpha = 0.01$ instead. Then we'd reject the null hypothesis if the p-value is below 0.01.
- Physicists control Type I error at around $\alpha = 5 \times 10^{-7}$. (Then again, atoms are cheaper than patients and mice.)

Multiple Testing

- Now, instead of testing one feature's association with the response, suppose we test the association of M features with the response.
- We can compute a p-value for the association between each feature and the response.
- Probability that the jth p-value will be less than 0.05, assuming the null hypothesis holds for the jth feature, is 0.05.
- But the probability that at least one of the M p-values will be less than 0.05, assuming that all of the null hypotheses hold, is much greater than 0.05.
- In other words, we will erroneously reject the null hypothesis if we test a lot of hypotheses and reject the null hypothesis for any p-values less than the usual threshold, like 0.05.

Multiple Testing

- The probability of falsely rejecting the jth null hypothesis is α.
- The probability of falsely rejecting at least one of M null hypotheses is $1 - (1 - \alpha)^M$.
- Let $\alpha = 0.05$, and assume that the tests are all independent.
  - $M = 1$: probability of falsely rejecting at least one null hypothesis is 0.05.
  - $M = 2$: probability of falsely rejecting at least one null hypothesis is 0.0975.
  - $M = 10$: probability of falsely rejecting at least one null hypothesis is 0.40.
  - $M = 200$: probability of false rejecting at least one null hypothesis is 0.9999649.

Hypothesis Testing

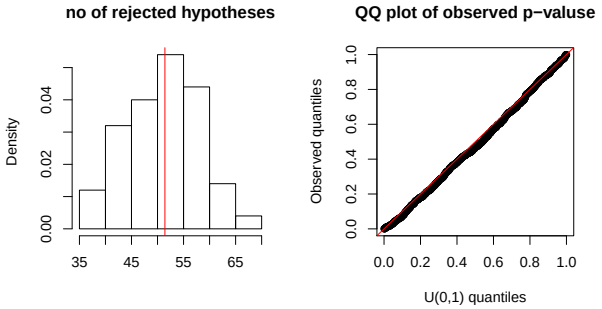
We will get a lot of false positives if we use 0.05 as a cut-off for rejecting the null hypothesis.

Protein	T-stat	P-Value
1	3.2	0.00137
2	-1.8	0.0718
3	5.8	6e-9
4	13.2	0
5	1.4	0.1615
6	-0.2	0.8414
.	.	.
.	.	.
.	.	.
M	4	6e-5

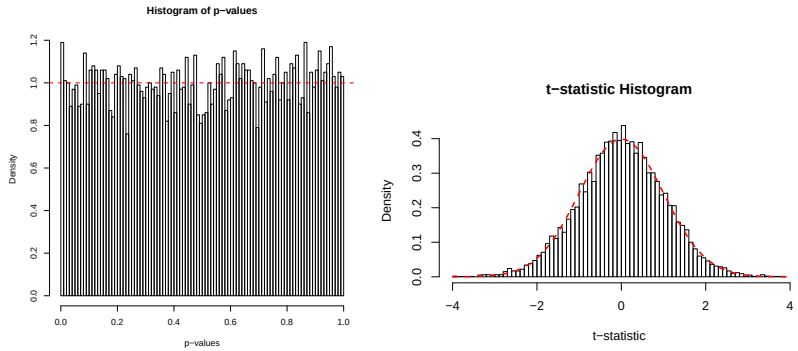
A Simple Illustration

- Consider 1000 hypotheses & 50 samples in two groups (25 treatment, 25 control) (e.g. mRNA expression for 1000 genes) from $N(0, 1)$: nothing significant!
- 10 smallest p-values:

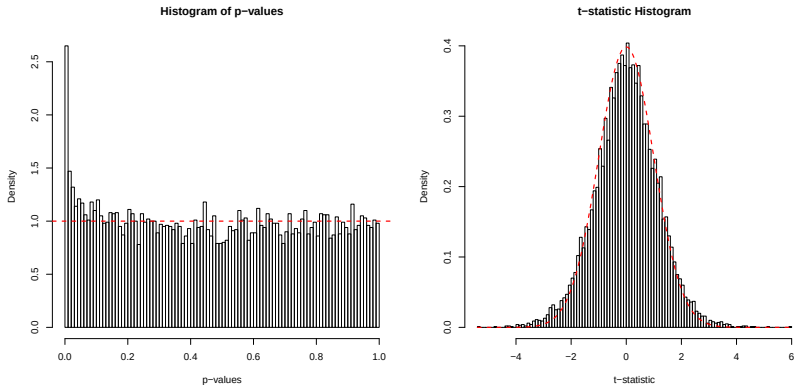
  0.0005, 0.0069, 0.0076, 0.0087, 0.0123, 0.0124, 0.0132, 0.0160, 0.0168, 0.0176
- All of these will be rejected by chance at the level of 5%!!



Example 1: All Null Hypotheses Hold



Example 2: Not All Null Hypotheses Hold



Bonferroni Correction

- Simple solution: A Bonferroni correction.
- Rather than rejecting the jth null hypothesis is less than 0.05, we reject the jth null hypothesis if the jth p-value is less than $0.05/M$.
- More generally, to control the overall Type I error – probability of falsely rejecting the null hypothesis – at level α, we must reject the jth null hypothesis if its p-value is below α/M.

Why does the Bonferroni correction control Type I Error?

Consider test statistics T_j for tests of M features (actually T_j here is the absolute value of the test statistic, so we reject test j if T_j is large)

$$\begin{aligned} P(\max T_j > t^*) &= P(\text{at least one } T_j > t^*) \\ &= P(\{T_1 > t^*\} OR \{T_2 > t^*\} OR \ldots OR \{T_M > t^*\}) \\ &\leq P(T_1 > t^*) + P(T_2 > t^*) + \cdots + P(T_M > t^*) \\ &= M \times P(T > t^*)^{\dagger} \end{aligned}$$

So, to get

$$P(\max T_j > t^*) \leq 0.05$$

we can find t^* such that

$$M \times P(T_j > t^*) \leq 0.05 \;\Leftrightarrow P(T_j > t^*) \leq 0.05/M$$

† considering the setting that there is no effect!

Why does the Bonferroni correction control Type I Error?

- Let A denote the event that at least one null hypothesis is falsely rejected, and let A_j be the event that the jth null hypothesis is falsely rejected. So

$$P(A) = P(\bigcup_{j=1}^{M} A_j) \leq \sum_{j=1}^{M} P(A_j).$$

  This means that if we keep $P(A_j)$ below α/M, then $P(A)$ will be below α.
- And keeping $P(A_j)$ below α/M means that we reject the jth null hypothesis if the p-value is below α/M.

Bonferroni Bottom Line

- To summarize, to control the overall Type I error at level α, we reject the jth null hypothesis if its p-value is below α/M.
- So if we are testing each SNP for association with a phenotype in a GWAS with $M = 10^6$ SNPs, then we reject the jth null hypothesis if its p-value is below 0.05×10^{-6}.
- This is very conservative! We will hardly ever reject the null hypothesis.
- Slightly less conservative alternatives to Bonferroni, that also control overall Type I error, are also available (Holm (1979), Hochberg (1988), Hommel (1988)).
- Holm's method dominates Bonferroni under arbitrary assumptions.

A Less Conservative Notion of Error Control

- The Bonferroni correction allows us to be confident that we will almost never falsely reject the null hypothesis.
- But in the analysis of omics data, this approach can be a little too conservative.
- If an investigator spends a lot of money measuring DNA methylation levels at millions of sites, he or she might be willing to sometimes falsely reject the null hypothesis, in exchange for correctly rejecting the null hypothesis more frequently.
- In other words: willing to have some false positives, in exchange for more true positives.
- A new notion of error control: false discovery rate.
- Proposed in 1995, but still a very active area of research!

Possible Outcomes from M Hypothesis Tests

	Called Not Significant	Called Significant	Total
H_0 True	U	V	M_0
H_0 False	T	S	M_1
Total	$M - R$	R	M

- V is number of false positives.
- T is number of false negatives.
- Type I error is $E(V)/M_0$.
- Type II error is $E(T)/M_1$.
- The power is $1 - E(T)/M_1$.
- The Bonferroni correction guarantees that $P(V \geq 1) \leq \alpha$.

False Discovery Rate

$$FDR = E\left(\frac{\text{number of null hypotheses falsely rejected}}{\text{number of null hypotheses rejected}}\right).$$

For instance, in a gene expression experiment,

$$FDR = E\left(\frac{\text{number of genes incorrectly declared significant}}{\text{number of genes declared significant}}\right).$$

In other words, this is the fraction of discoveries that are false positives.

False Discovery Rate

	Called Not Significant	Called Significant	Total
H_0 True	U	V	M_0
H_0 False	T	S	M_1
Total	$M - R$	R	M

- FDR: V/R, which is random... we often control $E(V/R)$.
- By controlling the FDR, we are guaranteeing that we don't have too many false discoveries.
- If we use an FDR threshold of 0.2, then no more than 20% of the null hypotheses that we reject were actually true. (Unfortunately, we don't know which ones!)
- For instance, in a GWAS, if we reject the null hypothesis of no association for SNPs with $FDR \leq 0.2$, then we expect no more than 20% of those SNPs to be false positives.

Benjamini-Hochberg Algorithm for FDR Control

1. Fix the false discovery rate, α.
2. Let $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(M)}$ denote the ordered p-values for the M hypothesis tests.
3. Define
$$L = \max\left\{k : p_{(k)} < \alpha\frac{k}{M}\right\}.$$
4. Then, reject the jth null hypothesis if $p_j \leq p_{(L)}$, the Benjamini-Hochberg rejection threshold.
5. In other words, find the maximum order statistic (k) such that
$$\frac{M \times p_{(k)}}{k} \leq \alpha$$
Then reject all tests for which $p_j \leq p_{(k)}$

FDR vs Bonferroni Control

- Benjamini & Hochberg:
  - Find the maximum order statistic (k) such that
$$p(k) \leq \frac{\alpha k}{M}$$
  - Reject all tests j with $p_j < p(k)$.
- Bonferroni:
  - Reject test j if
$$p_j \leq \frac{\alpha}{M}$$

Example in R: All Null Hypotheses Hold

```
x <- matrix(rnorm(1000*50),ncol=50)
y <- sample(c(0,1),50,rep=TRUE)
ps <- NULL
for(i in 1:1000) ps <- c(ps,
 t.test(x[i,y==0],x[i,y==1])$p.value)
cat("Around 5% of p-values are below 0.05:",
mean(ps<.05),fill=TRUE)
fdrs.bh <- p.adjust(ps, method="BH")
plot(ps,fdrs.bh)
plot(fdrs.bh)
```

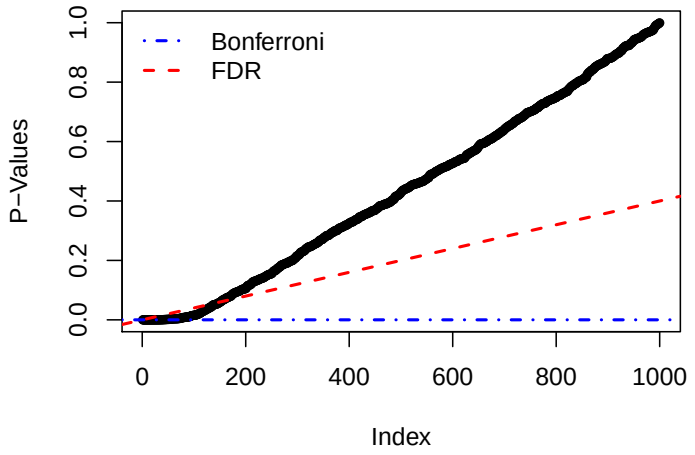

Example in R: Not All Null Hypotheses Hold

```
x <- matrix(rnorm(1000*50),ncol=50)
y <- sample(c(0,1),50,rep=TRUE)
x[1:100,y==0] <- x[1:100,y==0] + 1
ps <- NULL
for(i in 1:1000) ps <- c(ps,
t.test(x[i,y==0],x[i,y==1])$p.value)
cat("Way more than 5% of p-values are below 0.05:",
mean(ps<.05),fill=TRUE)
fdrs.bh <- p.adjust(ps, method="BH")
plot(ps,fdrs.bh)
plot(fdrs.bh)
```


Example, Continued

```
cat("Number of Tests with FDR below 0.4:",
sum(fdrs.bh<0.4), fill=TRUE)
cat("Compute the BH FDR Directly:",
max(which(sort(ps,decreasing=FALSE) < .4*(1:1000)/1000)),
                                    fill=TRUE)
plot(sort(ps,decreasing=FALSE),ylab="P-Values")
abline(a=0, b=0.4/1000,col="red")
```


Output From R



Correlated Hypotheses

- So far, we have not discussed the effect of correlation in the data
- In real data examples, genes, proteins etc work together, and are correlated
- How would methods of multiple testing adjustment work if tests are correlated?

Correlated Hypotheses

- Bonferroni correction works regardless of the correlation structure (however, we know that in general, Bonferroni is not very appealing)
- Benjamini & Hochberg can handle positive dependence: think of this as all genes being positively correlated with each other
- However, in practice genes may have both positive (inducers) and negative (inhibitors) correlations with each other, so this assumption may not hold
- This is still an active area of research

FDR Correction Under Dependence

- A control under dependence due to Benjamini & Yakutielli
  - Find the maximum order statistic (k) such that

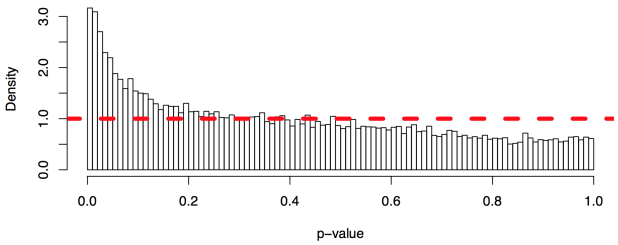
$$p(k) \leq \frac{\alpha k}{M \times \left(\sum_{j=1}^{M} 1/j\right)}$$

  - Reject all j with $p_j \leq p(k)$.
- $\sum_{j=1}^{M} 1/j \approx \log(M)$, so we pay an additional price of $\log(M)$, which makes this procedure more conservative than Benjamini & Hochberg (but then again, there is no free lunch!)
- In R, `p.adjust(ps, method=''BY'')`

A Real Data Example

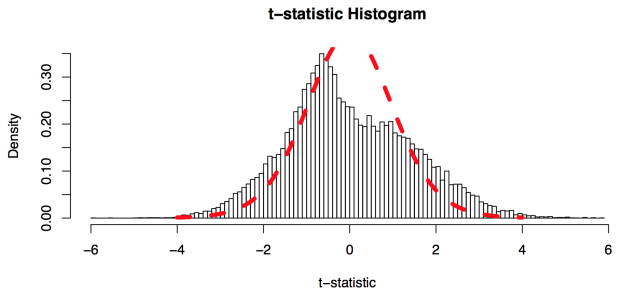
- Mechanisms for Cancer immortality! (TELO vs ALT)
- Paper by Lafferty-whyte and co claiming to build a gene signature to differentiate mechanisms

Let's look at the p-values



A Real Data Example

And, test statistics



What is going on??

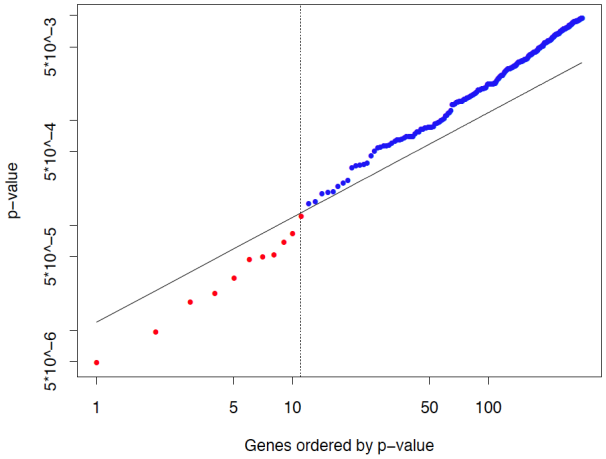
A Real Data Example

- With dependence correction, 0 rejections for FDR control ¡ 1
- Unclear how significant findings are...
  - Report K (10, 100, ...) most significant effects
  - Give FDR estimates from both "BH" and "BY"
  - Small sample-size and strange histogram shape add extra skepticism
  - We also need to check for batch effects and other potentially damaging factors

Example: Radiation Treatment Sensitivity

- Microarray data set on sensitivity of cancer patients to ionizing radiation treatment.
- Citation: Rieger, Hong et al, PNAS, 2004
- $M = 12,625$ genes, $n = 58$ samples: 44 samples with a normal reaction, 14 patients with severe reaction to radiation.
- Compute two-sample t-statistic for each gene's association with response to radiation.
- NO genes are declared significant after Bonferroni correction at level $\alpha = 0.05$.
- 11 genes had FDR below $\alpha = 0.15$.
- For each gene we can get a q-value, which is like a p-value for that gene, but quantifies the FDR associated with that gene.

Example: Radiation Treatment Sensitivity



Permutation Approach to FDR Estimation

- Another way to estimate FDRs: by permutation.
- Easy and natural – unlike Benjamini-Hochberg, can explain it pretty simply to a non-statistician.
- Does not require computing p-values, which can be helpful in settings where p-values are (1) difficult to compute, or (2) unreliable.
- Also known as "plug-in estimate for FDR".

Permutation Approach

1. Compute $t_1, \ldots, t_M$, the test statistic for each of the M features.
2. Create K permutations of the responses, and for each permutation from $k = 1, \ldots, K$, and for each feature $j = 1, \ldots, M$, compute $t_j^1, \ldots, t_j^K$.
3. For a range of values of the cut-point C, let

$$R_{obs} = \sum_{j=1}^{M} 1_{(|t_j| > C)},$$

and

$$\widehat{E(V)} = \frac{1}{K} \sum_{j=1}^{M} \sum_{k=1}^{K} 1_{(|t_j^k| > C)}.$$

4. Estimate the FDR by $\widehat{FDR} = \widehat{E(V)}/R_{obs}$.

Permutation Approach

- Based upon the approximation

$$E(V/R) \approx E(V)/E(R).$$

- We estimate $E(R)$ using R_{obs} – the actual number of test statistics that exceed the threshold.
- We estimate $E(V)$ by permuting the response and calculating the number of test statistics that exceed the threshold, in the absence of any real relationship between the features and the response.
- Actually, $\widehat{E(V)}$ estimates $(M/M_0)E(V)$, but since $M > M_0$, this over-estimate is conservative.

Permutation Approach

- An advantage of the permutation approach: can estimate FDR even when we don't know how to compute p-values.
- For instance, suppose we have protein level measurements for 100 proteins, as well as associated response measurements, in both pregnant and non-pregnant women.
- We want to develop a test statistic that combines the info from the pregnant women and from the non-pregnant women.
- For the jth gene, can compute $t_j^{pregnant}$, the test statistic for pregnant women, and $t_j^{non-pregnant}$, the test statistic for non-pregnant women.
- Can use $|t_j^{pregnant}| + |t_j^{non-pregnant}|$ as an overall test statistic.
- How to get p-value for this new test statistic? Who knows!
- But estimating FDR using a permutation approach is pretty straightforward.

At What Level to Control FDR?

- In biology, we typically control Type I Error at level $\alpha = 0.05$ or $\alpha = 0.01$. (We reject the null hypothesis if the p-value is less than 0.05 or 0.01.)
- However, there is no analogous default cut-off for rejecting the null hypothesis using FDR control.
- We might want to follow up on genes whose FDR is below 10% or 20% in a gene expression experiment.
- In general, the FDR threshold that we use will depend on the data set as well as the number of genes with small FDRs... as well as the number of rejected null hypotheses that we can afford to follow up on in the lab!

P-Values Using Permutations

- We could also compute p-values using a permutation-type approach: the p-value for the jth feature is given by

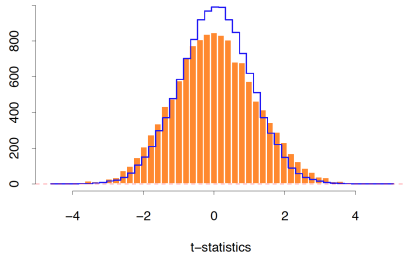
$$p_j = \frac{1}{MK} \sum_{k=1}^{K} \sum_{j'=1}^{M} 1_{(|t_{j'}^k| > |t_j|)}.$$

- Can also use the permutation p-values only

$$p_j = \frac{1}{K} \sum_{k=1}^{K} 1_{(|t_{j'}^k| > |t_j|)}.$$

P-Values Using Permutations

- Example on radiation sensitivity data: orange shows true t-statistics, and blue shows t-statistics for permuted data.

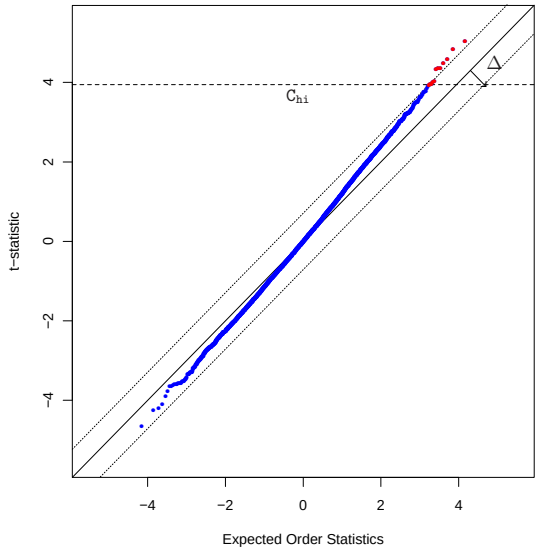


Asymmetric Cut-Points

- In previous examples, we used the absolute value of the test statistic, $|t_j|$, and hence the same cut-points for both positive and negative values
- In some cases, (most) differentially expressed genes change in the positive (or negative) direction. In such cases, it may be better to use different cut-points for negative and positive test statistics
- SAM (Significance Analysis of Microarrays) is such a procedure (Tusher et al, 2001):

SAM (Significance Analysis of Microarrays)

Tusher et al (2001)

- uses the values of test-statistics instead of p-values
- based on a permutation approach to determine the "expected" test-statistics
- uses a QQ-plot to find test statistics that are far from expected
- the cutoff value is determined by finding a band of length Δ around the expected test-statistics
- is implemented in the R-package `samr`: `SAM(x,y=NULL,...)`

SAM