New statistical approaches for detecting differential expression

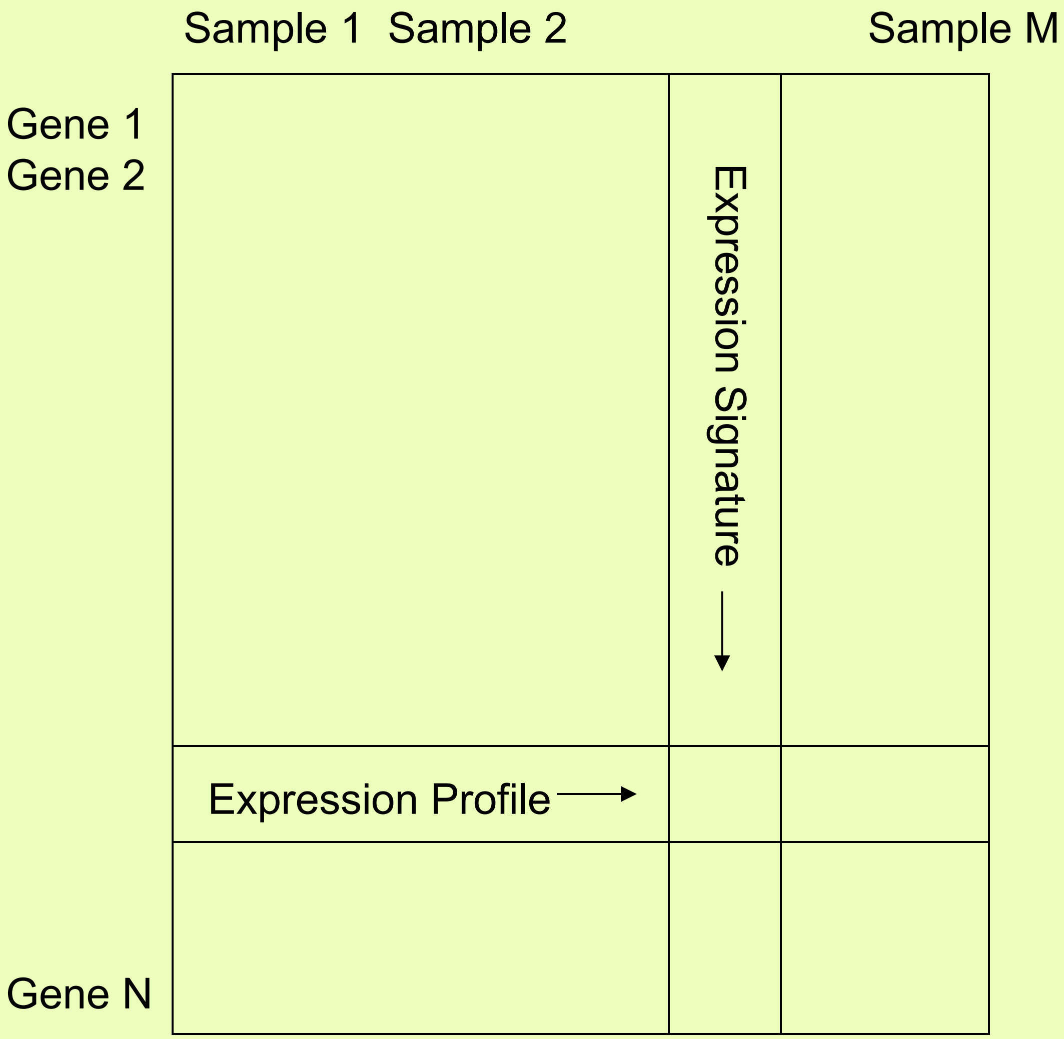
Richard Bean

McLachlan Group

October 5, 2006

Our Group

Microarray data represented as *N* x *M* matrix **Y**



Sample 1 Sample 2 Sample M

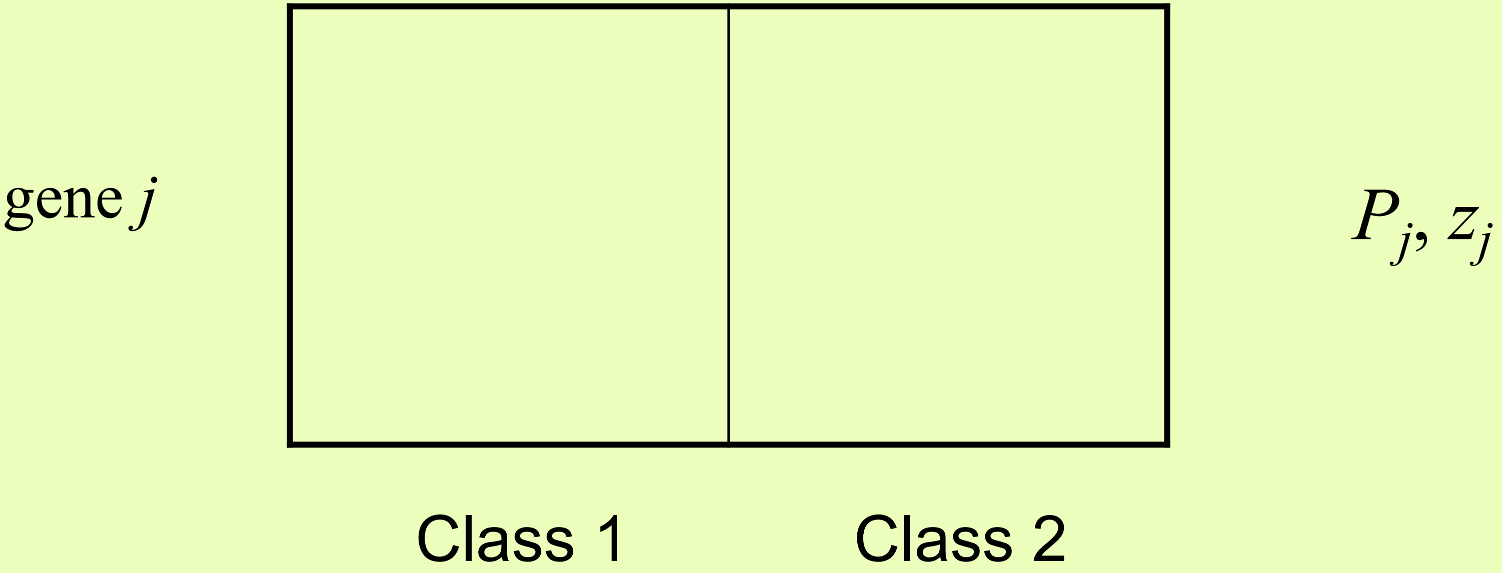
Gene 1

Gene 2

Gene N

Expression Signature

Expression Profile



gene j

P_j, z_j

Class 1

Class 2

$$z_j = \Phi^{-1}\left(1 - P_j\right)$$

Getting a *P*-value: An example of a gene from Hedenfalk et al (2001) breast cancer data

Class 1: BRCA1 (7 tissues)
-0.587 -0.5 -0.0707 -0.265 -0.542 -0.522 0.265

Class 2: BRCA2 (8 tissues)
-0.7 0.377 0.0318 -0.475 -0.627 -0.56 1.39 -0.4

$$\bar{x}_1 = -0.3173, \bar{x}_2 = -0.1203$$

$$s_1^2 = 0.1002, s_2^2 = 0.5066, s^2 = 0.3190$$

$$t_{13} = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = -0.6739$$

$$P = 0.512$$

Requires data to be normal & i.i.d. in each class.

If data are not normally distributed, can use permutation methods.

P=0.511

P=0.512

As Efron (2006) notes

“working inside the **Y** matrix will give more information in some situations – but need assumptions to hold for results to be valid – here aim is to work with a minimum number of assumptions”

Multiple Hypothesis Testing Framework

FDR (False Discovery Rate) of Benjamini & Hochberg (1995)

$$FDR \approx \frac{\#(\text{false positives})}{\#(\text{significant genes})}$$

Can implement a procedure based on $P_1,\ldots,P_N$ to control FDR. But FDR is a global measure.

Three Ideas

1. Use a local FDR measure

2. Estimate other error rates besides FDR e.g. FNR or 1-FNR = sensitivity

3. Use an empirical null distribution in place of the theoretical null distribution

- McLachlan GJ, Bean RW, Ben-Tovim Jones L, Zhu JX. Using mixture models to detect differentially expressed genes. *Australian Journal of Experimental Agriculture* **45** (2005), 859-866.

- McLachlan GJ, Bean RW, Ben-Tovim Jones L. A simple implentation of a normal mixture approach to differential gene expression in multiclass microarrays. Bioinformatics **26** (2006), 1608-1615.

- Efron B et al (2001) Empirical Bayes analysis of a microarray experiment. *JASA* **96,**1151-1160.
- Efron B (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *JASA* **99**, 96-104.
- Efron B (2004) Selection and Estimation for Large-Scale Simultaneous Inference.
- Efron B (2005) Local False Discovery Rates.
- Efron B (2006) Correlation and Large-Scale Simultaneous Significance Testing.
- Efron B (2006) Size, power and false discovery rates.

Local FDR

Lee (2000), Efron et al (2001), Newton et al (2001) proposed a two-component mixture model

$$f(z_j) = \pi_0 f_0(z_j) + (1 - \pi_0) f_1(z_j)$$

$$\begin{aligned}\tau_0(z_j) &= pr\{j\text{th gene is null} \mid z_j\} \\ &= \frac{\pi_0 f_0(z_j)}{f(z_j)} \\ &= \frac{\pi_0 f_0(z_j)}{\pi_0 f_0(z_j) + (1-\pi_0) f_1(z_j)} \quad \text{(by Bayes' theorem)}\end{aligned}$$

Strictly speaking, a real Bayesian would use

$$\tau_{0j} = pr\{j\text{th gene is null} \mid z_1, \ldots, z_N\}$$

An example where local FDR is more informative: Glonek and Solomon (2003)

F_0: N(0,1), π_0=0.9
F_1: N(1,1), π_1=0.1

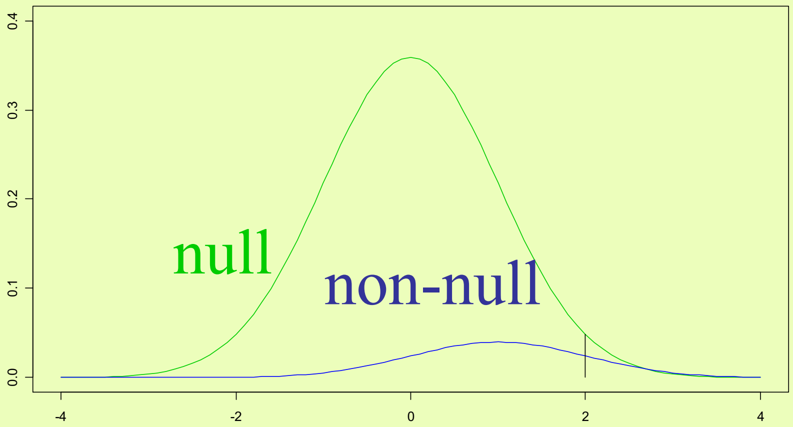
Reject H_0 if $z \geq 2$

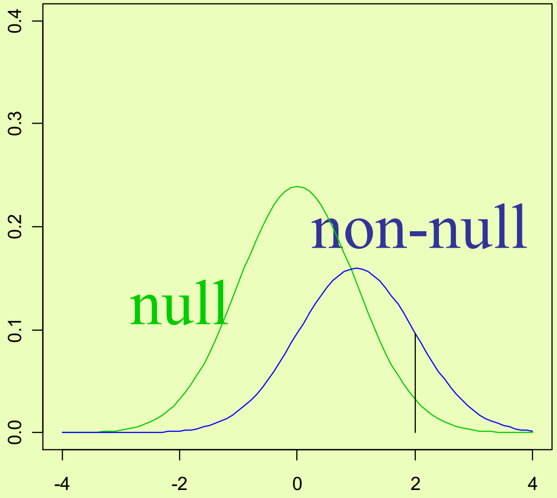
$\tau_0(2) = 0.99972$
but FDR=0.17

F_0: N(0,1), π_0=0.6
F_1: N(1,1), π_1=0.4

Reject H_0 if $z \geq 2$

$\tau_0(2) = 0.251$
but FDR=0.177


0.4
0.3
0.2
0.1
0.0
null
non-null
-4
-2
0
2
4


0.4
0.3
0.2
0.1
0.0
non-null
null
-4
-2
0
2
4

$$\tau_0(z_j) = \frac{\pi_0 f_0(z_j)}{\pi_0 f_0(z_j) + (1 - \pi_0) f_1(z_j)}$$

$$\tau_0(z_j) = \frac{\pi_0 \overset{\mathrm{N}(0,1)}{f_0(z_j)}}{\pi_0 f_0(z_j) + (1-\pi_0) f_1(z_j)}$$

In order to proceed with estimation of π_0 (can easily estimate $f(z_j)$ from $z_1,\ldots,z_N$) we need to make the problem identifiable.

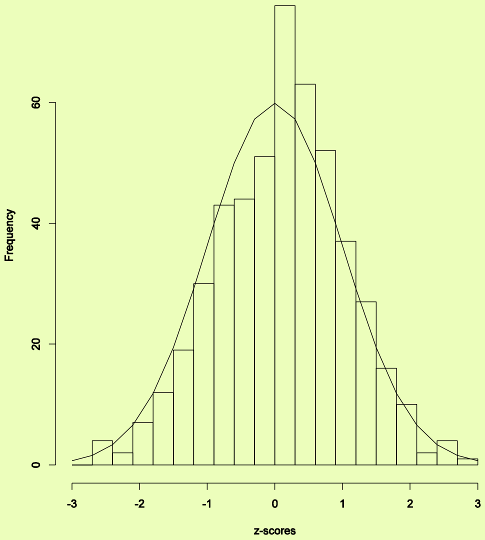
Now $f_0(z_j)$ *is* N(0,1) and we have to assume something about $f_1(z_j)$.

N(0,1)

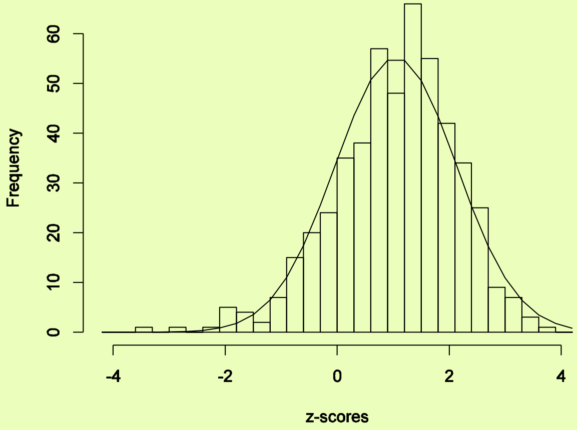
$$\tau_0(z_j) = \frac{\pi_0 f_0(z_j)}{\pi_0 f_0(z_j) + (1-\pi_0) f_1(z_j)}$$

N(0,1)

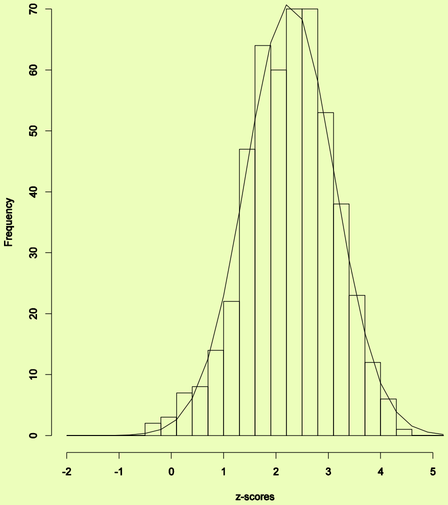
$N(\mu_1,\sigma_1^2)$



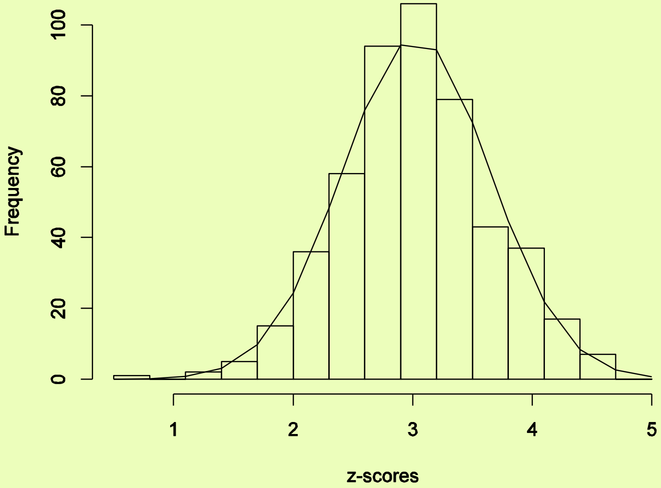
Z-values, null case



Z-values, +1

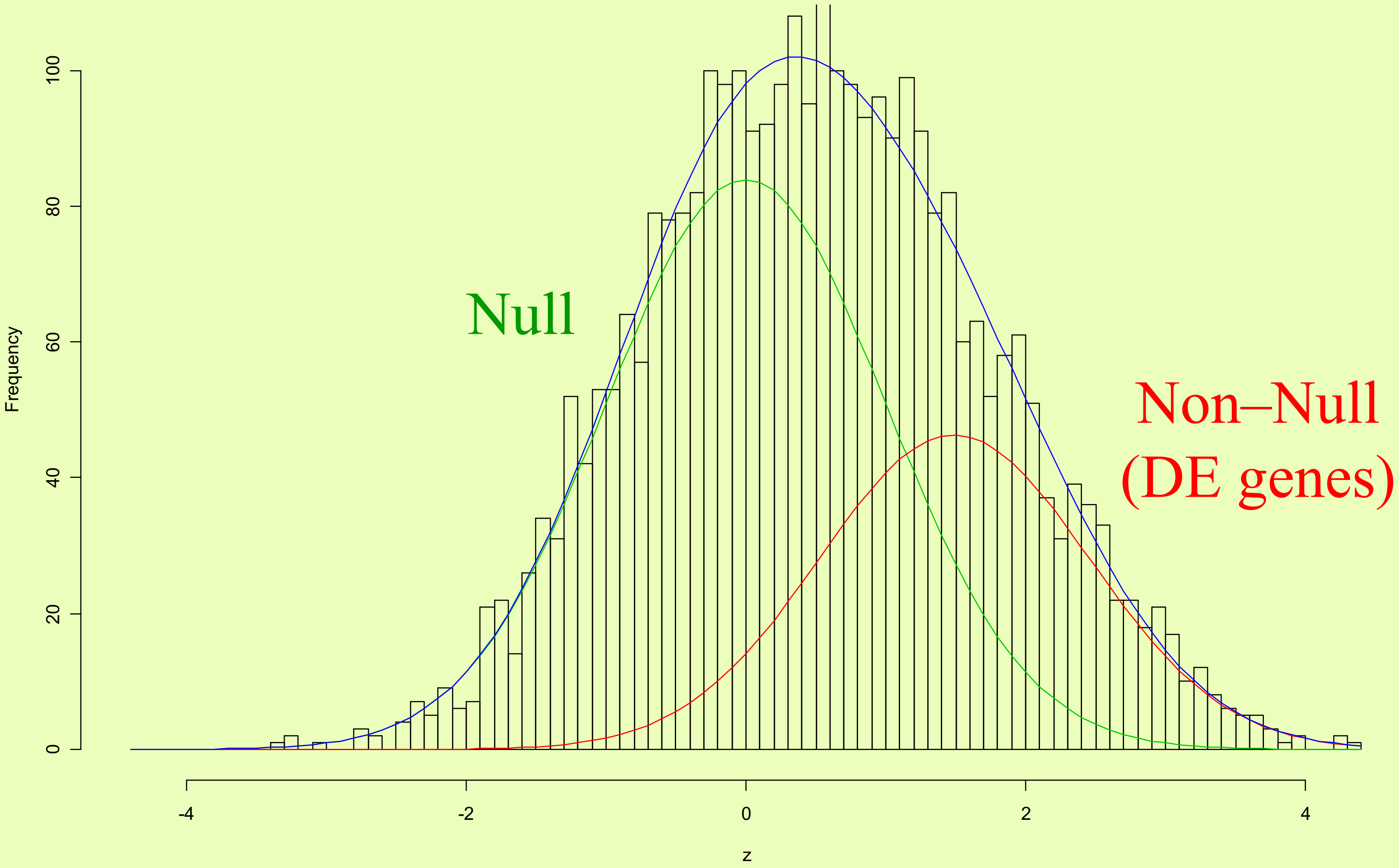


Z-values, +2



Z-values, +3

Fitting two component mixture model to Hedenfalk data

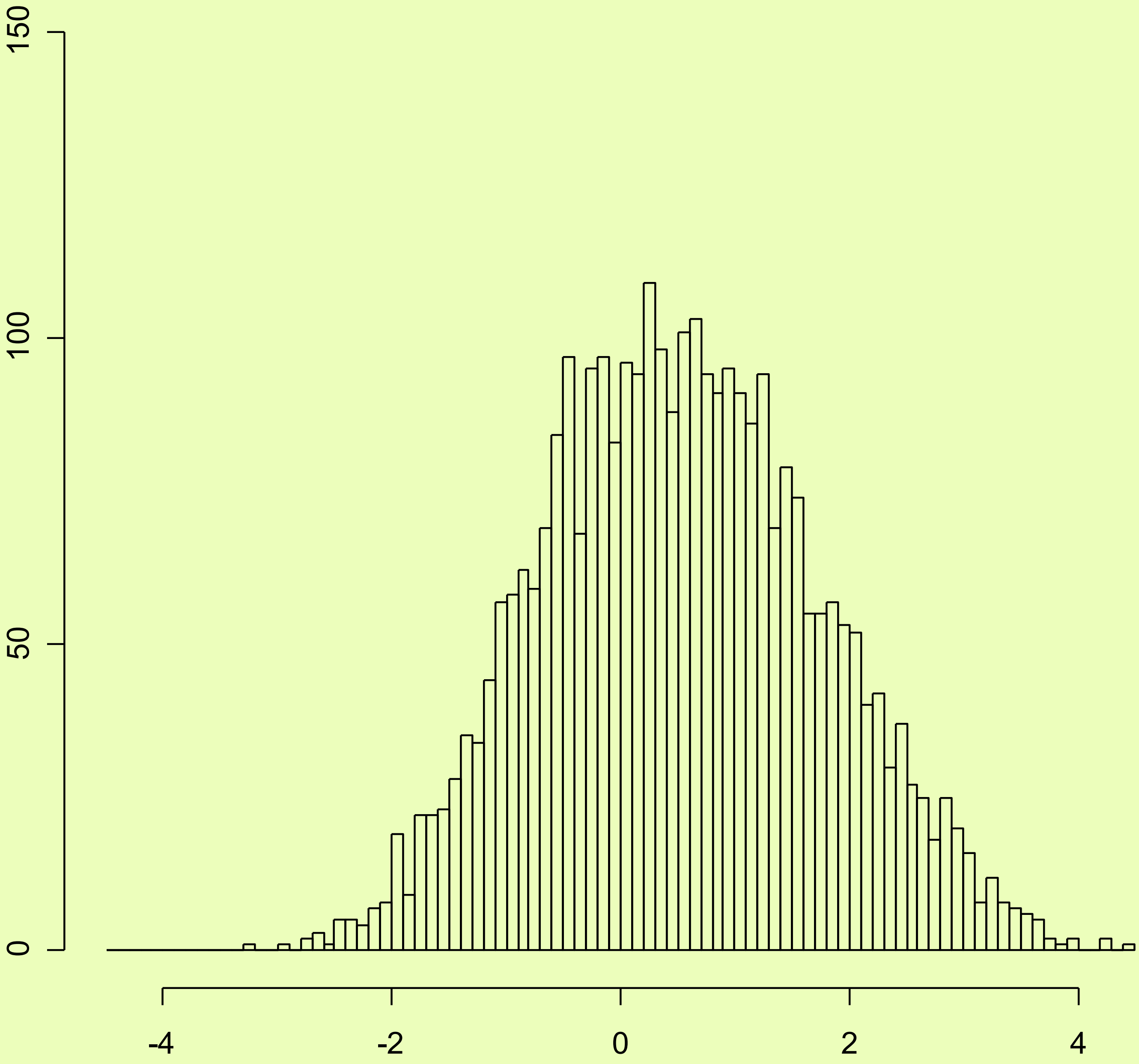


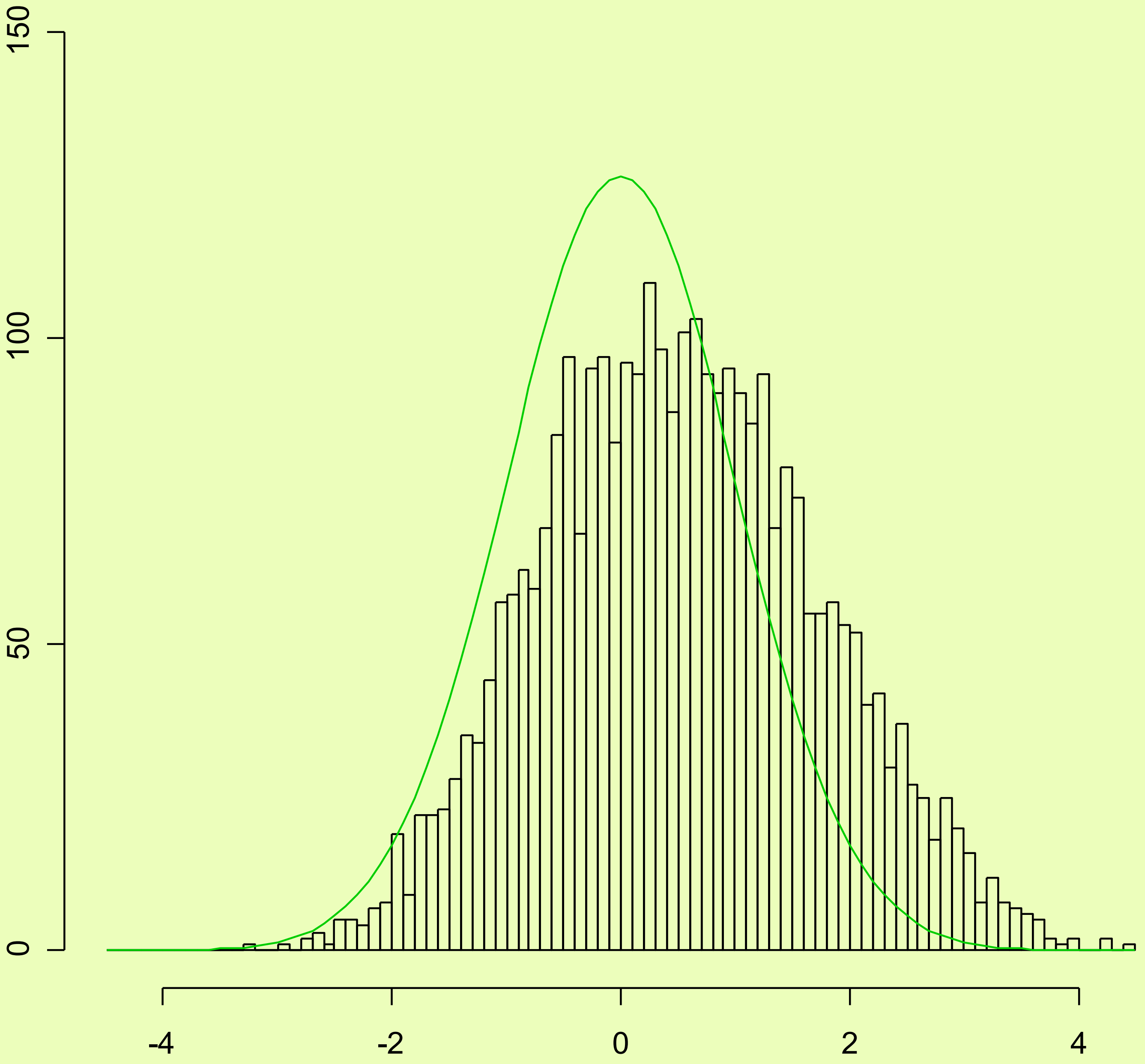
Fit

$$\pi_0 N(0,1) + (1- \pi_0)N(\mu_1,\sigma_1^2)$$

via maximum likelihood.

For given π_0, MLEs of μ_1, σ_1^2 are determined: try various π_0.


150
100
50
0
-4
-2
0
2
4

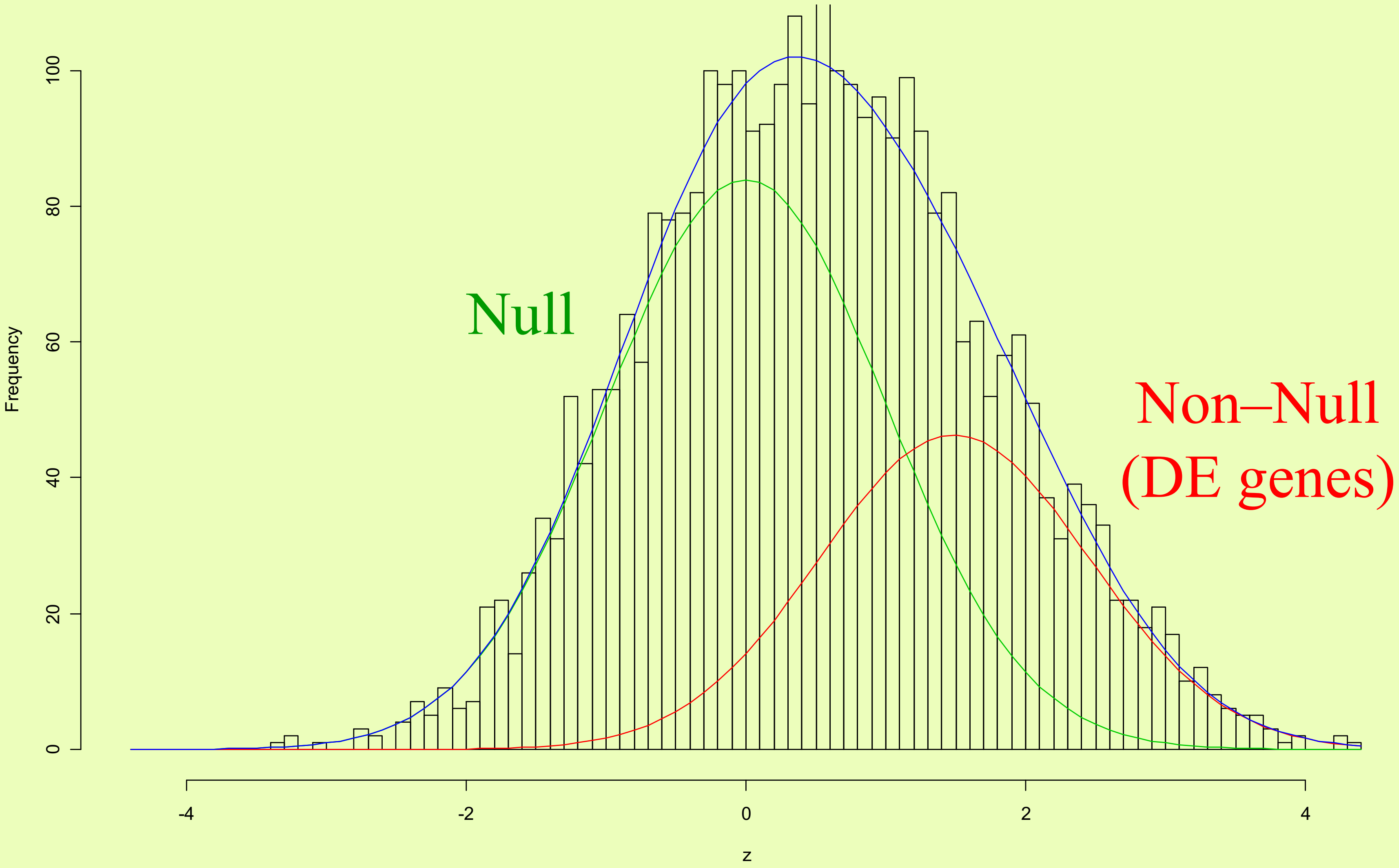

150
100
50
0
-4
-2
0
2
4

Pick a value $\xi < 0$, for example -0.2.

$\pi_0 N$ (area to the left of ξ) $\approx \#(z_j < \xi)$

$\pi_0 \approx \#(z_j < \xi) / N$ (area to the left of ξ)

Fitting two component mixture model to Hedenfalk data



Ranking and Selecting the Genes

Gene j	P_j	z_j	$\hat{\tau}_0(z_j)$
Gene 1			0.002
.			.
.			.
.			.
Gene R			0.1
.			0.12
.			0.18
.			.
.			.
Gene R+R$_1$			0.20
.			
.			
.			
Gene N			

Local FDR

FDR
= Sum/R
= 0.06

$c_0 = 0.1$

Proportion of
False Negatives
= 1 – Sum$_1$/ R$_1$

Estimated FDR

$$\widehat{\mathrm{FDR}} = \sum_{j=1}^{N} \hat{\tau}_0(w_j)\, I_{[0,c_o]}(\hat{\tau}_0(w_j))/N_r$$

where

$$N_r = \sum_{j=1}^{N} I_{[0,c_o]}(\hat{\tau}_0(w_j))$$

Similarly, the false positive rate is given by

$$F\hat{P}R = \sum_{j=1}^{N} \hat{\tau}_0(w_j) I_{[0,c_0]}(\hat{\tau}_0(w_j)) / \sum_{j=1}^{N} \hat{\tau}_0(w_j)$$

and the false non-discovery rate and false negative rate by:

$$FN\hat{D}R = \sum_{j=1}^{N} \hat{\tau}_1(w_j) I_{(c_0,\infty)}(\hat{\tau}_0(w_j)) / (N - N_r)$$

$$F\hat{N}R = \sum_{j=1}^{N} \hat{\tau}_1(w_j) I_{(c_0,\infty)}(\hat{\tau}_0(w_j)) / \sum_{j=1}^{N} \hat{\tau}_1(w_j)$$

Theoretical null may not hold for 4 reasons

1. Failed assumptions
   - Maybe non-normality distorts student's t-distribution
   - Can use permutation methods

2. Correlation across arrays
   - Student-*t* null density assumes independence across arrays
   - Permutation methods cannot help

3. Unobserved covariates (age, weight, stage)
   - Tend to widen null density of the z_j's
   - Permutation methods cannot help

4. Correlation across genes

$$\hat{\tau}_0(z_j) = \pi_0 f_0(z_j) / \hat{f}(z)$$

does not require independence of z_j's

Suppose (1), (2), or (3) is applicable but (4) is not (assume genes independent).

null Z_j may not be ~ N(0,1)

i.e. theoretical null may not hold

Thus: use empirical null

$$\tau_0(z_j) = \frac{\pi_0 f_0(z_j)}{\pi_0 f_0(z_j) + (1-\pi_0) f_1(z_j)}$$

$N(\mu_0,\sigma_0^2)$ $N(\mu_1,\sigma_1^2)$

μ_0, σ_0^2 are now estimated from the data.

Call $N(\mu_0, \sigma_0^2)$ the *empirical null* distribution.

Problem now is to fit

$$\pi_0 N(\mu_0, \sigma_0^2) + (1 - \pi_0) N(\mu_1, \sigma_1^2)$$

1. Specify an initial value of π_0 (try theoretical null estimate and other estimates as before)

2. Rank z_j's and put $N\pi_0$ smallest in null component and remainder in non-null component

3. Work out means/variances as if they are the true groups

Now suppose the z_j's are correlated (4th reason).

Even if theoretical null N(0,1) is correct for an individual z_j of a null gene, the z_j's for the null genes may not behave as N(0,1) variates in the ensemble of $z_1, \ldots, z_N$.

If they don't, then the Benjamini-Hochberg procedure will break down using *P*-values based on theoretical null.

Fit

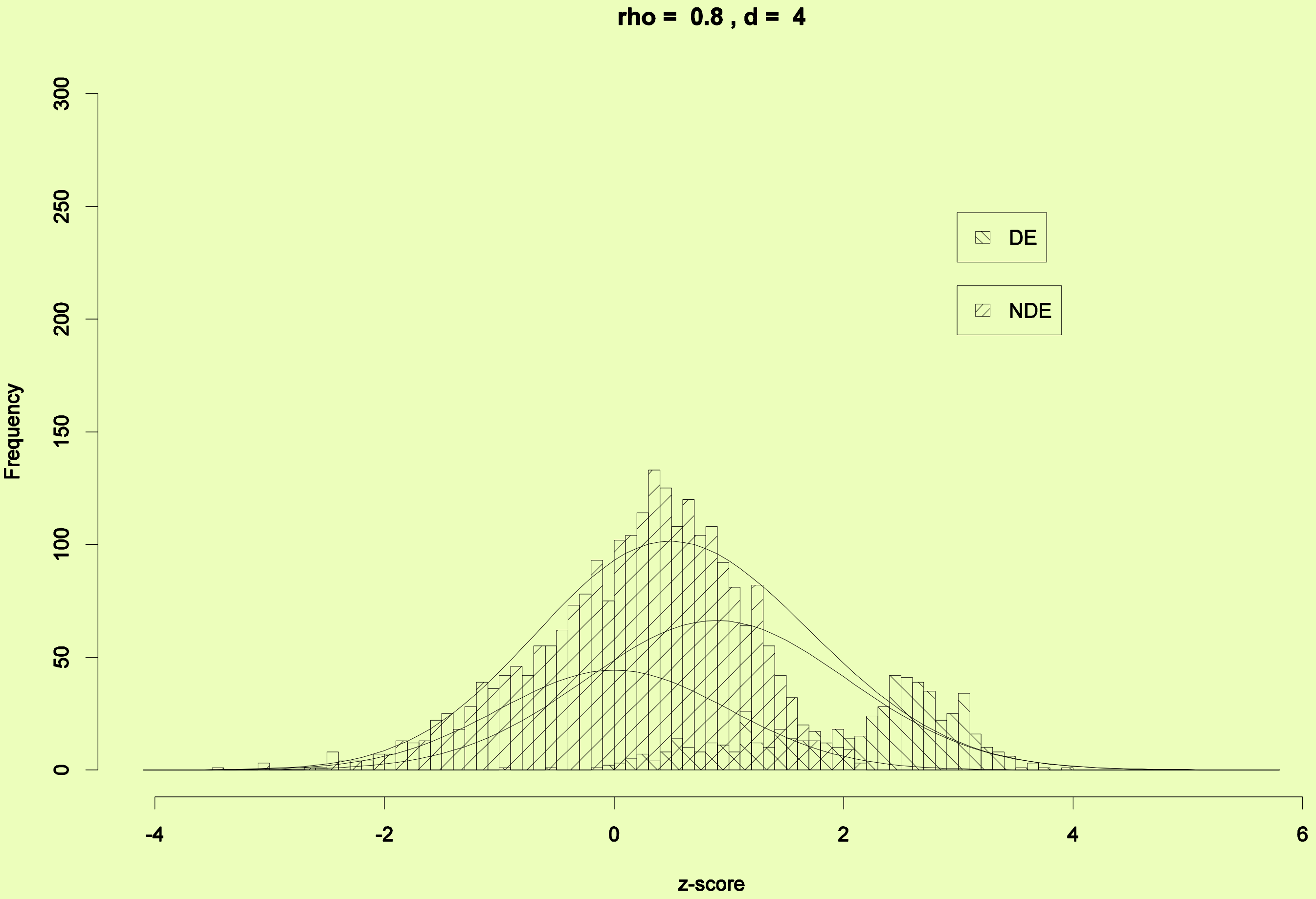
$$\pi_0 N(\mu_0, \sigma_0^2) + (1 - \pi_0) N(\mu_1, \sigma_1^2)$$

Still using maximum likelihood, although the function we are maximizing is no longer the true likelihood due to correlation across the genes.

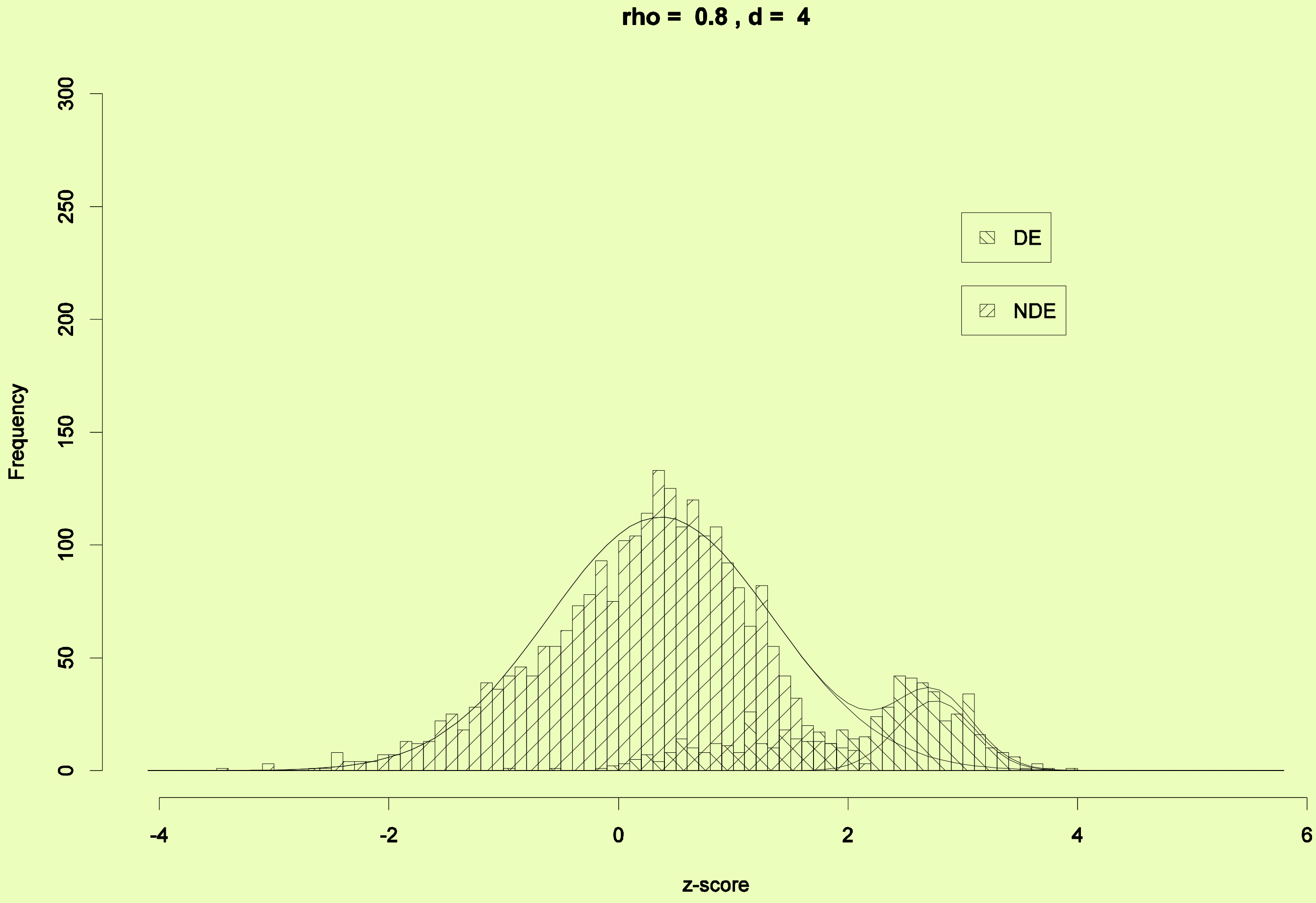
Allison Mice Simulation

Allison et al (CSDA,2002) generated data for 10 mice over 3000 genes. The data are generated in six groups of 500 with a value *ρ* of 0, 0.4, or 0.8 in the off-diagonal elements of the 500 x 500 covariance matrix used to generate each group.

For a random 20% of the genes, a value *d* of 0, 4, or 8 is added to the gene expression levels of the last five mice.



Theoretical null, ρ=0.8, d=4



Empirical null, ρ=0.8, d=4

When we *need* an empirical null in an actual example

e.g. HIV data of van’t Wout et al (2003), analyzed in Gottardo et al (2006)

van’t Wout et al (2003), J Virology **77**, 1392-1402
Gottardo et al (2006), Biometrics **62,** 10-18

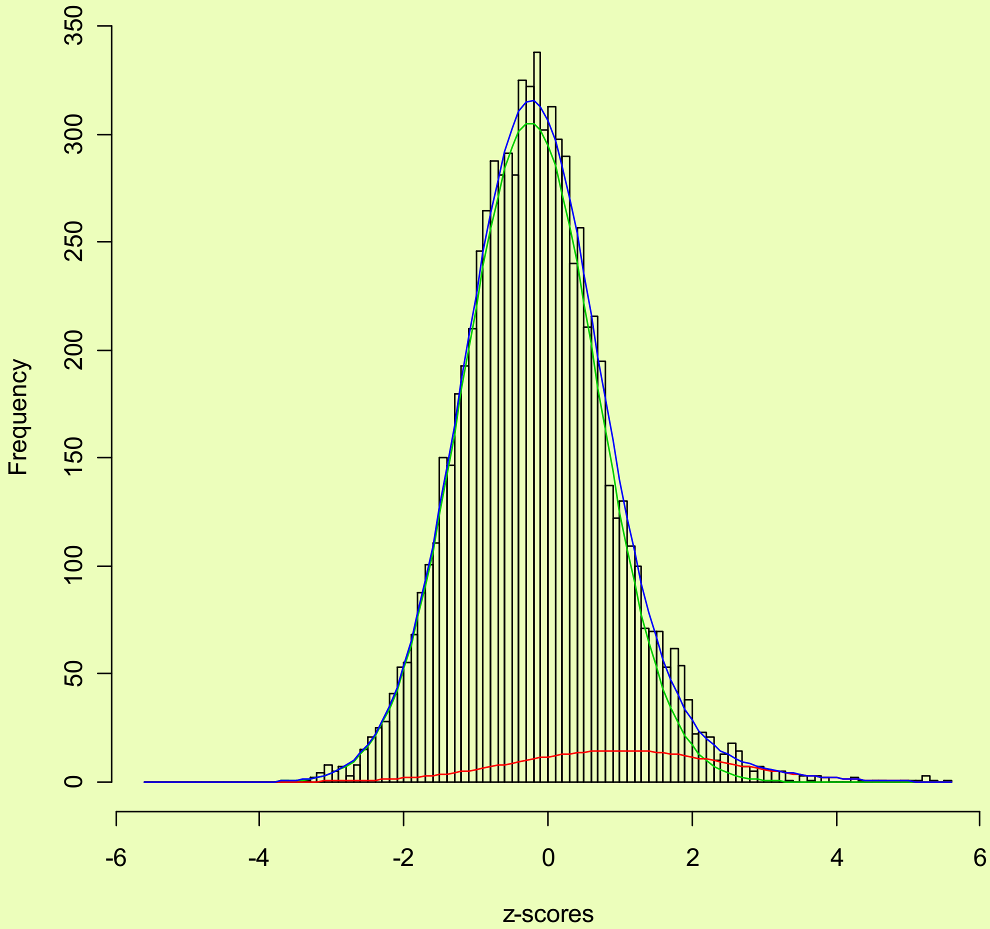
Z_j scores: $N(-0.16, 1.06)$

Fitted two-component model for the HIV Data

$$0.93N(-0.25,0.87)+0.07N(0.99,2.14)$$

j th gene is taken to be differentially expressed if:

$$\hat{\tau}_0(z_j) \leq c_0$$



HIV data: plot of fitted two-component
normal mixture model with empirical null and non-null
components (weighted respectively by the estimated proportion of null and
non-null genes) imposed on histogram of *z*-scores.

Null	# significant genes at c_0=0.1
Theoretical	0
Empirical	35

Can check for need of empirical null in place
of theoretical null by comparing
twice the increase in the log likelihood
when fitting μ_0, σ_0^2 instead of
fixing $\mu_0=0$ and $\sigma_0^2=1$.

Summary

- Mixture model based approach to finding DE genes is effective
- Gives measure of local as well as global FDR; also gives other error rates
- Provides an empirical null for use when theoretical null is misleading