Doing Thousands of Hypothesis Tests at the Same Time

Bradley Efron

Stanford University

Simultaneous Hypothesis Testing

1980: "Simultaneous Statistical Inference" (Rupert Miller)

- 2, 3, …, 20 simultaneous tests

Today: Several thousand tests

- *High Throughput Devices*:
  Microarrays, fMRI, proteomics, large-scale surveys

- **Love/Hate** Classical single-test theory

- frequentist, Bayes, empirical Bayes

A Microarray Example: The Prostate Data
(Singh et al. 2002)

- *102 Subjects*: 50 normal, 52 cancer
- $N = 6033$ genes : X 6033×102
- *Which genes are "non-null"?*

  i.e. expressed differently in cancer vs normal subjects?

t-statistics and *z*-scores

- i^{th} row of X

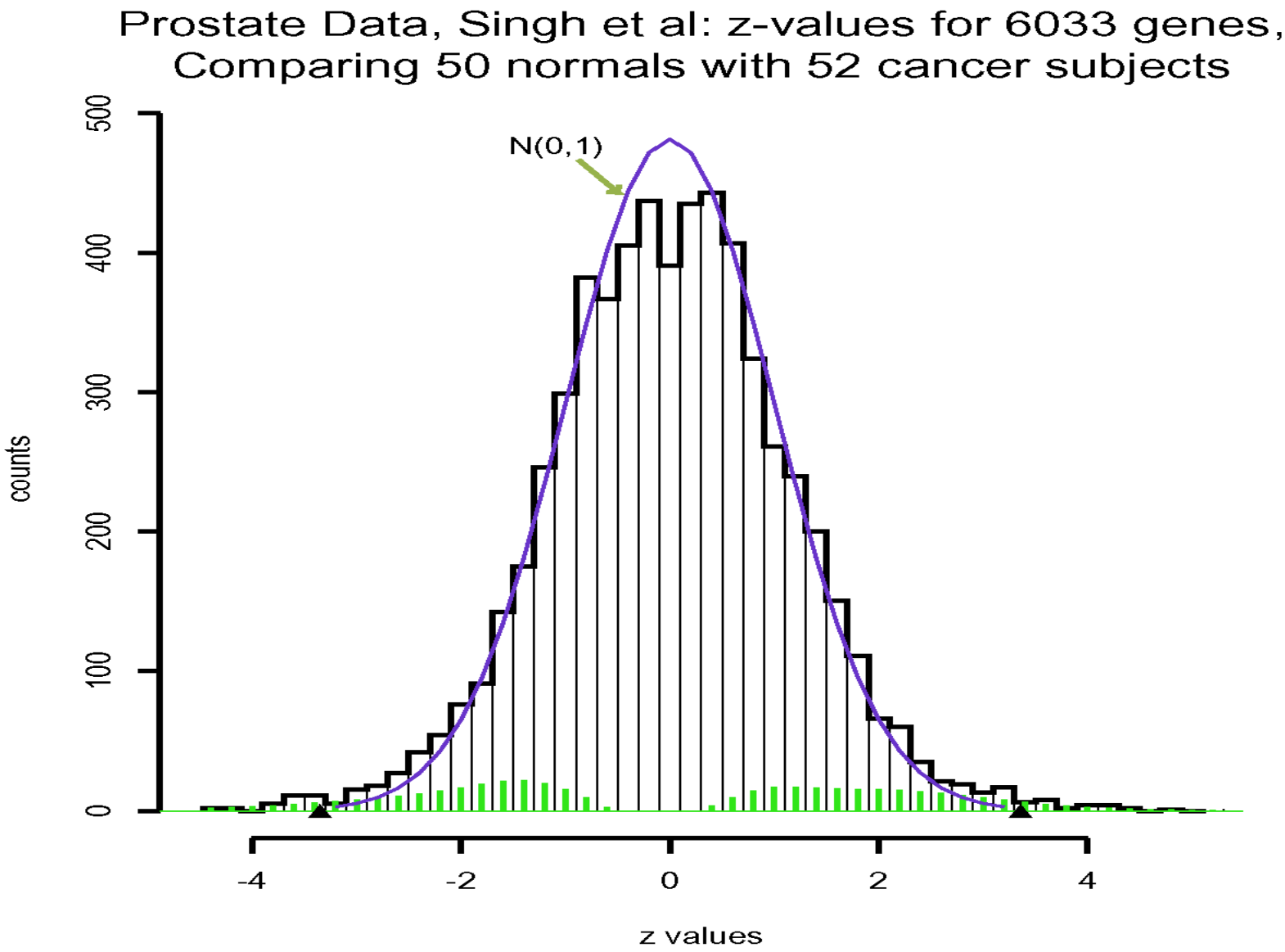
  normals $(x_{i1}, x_{i2}, \ldots, x_{i50})$
  cancer $(x_{i51}, x_{i52}, \ldots, x_{i102})$ } "t_i"

- t_i = two-sample *t*-stat, cancer vs normals

z-scores $\boxed{z_i = \Phi^{-1}(F_{100}(t_i))}$

where ϕ is $N(0,1)$ cdf, and F_{100} is cdf for t_{100}.

- **Theoretical Null** $z_i \sim N(0,1)$


Prostate Data, Singh et al: z-values for 6033 genes,
Comparing 50 normals with 52 cancer subjects
N(0,1)
counts
0
100
200
300
400
500
-4
-2
0
2
4
z values

The Two-Groups Model

Two Classes of Genes "null", "non-null"

- $\begin{cases} p_0 = \text{Prob \{null\}}, & f_0(z) \text{ density if null} \\ p_1 = \text{Prob \{non-null\}}, & f_1(z) \text{ density if non-null} \end{cases}$
- p_0 *large* (≥ 0.90)
- **Theoretical Null** $f_0 \sim N(0,1)$

(fits center of histogram)

- $f_1(z)$ "long tailed" (z_i's far from zero)

False Discovery Rates
(Efron 2006)

- *Mixture density*

$$f(z) = p_0 f_0(z) + p_1 f(z)$$

Bayes Rule "Local false discovery rate"

$$\boxed{\begin{aligned} fdr(z) &= Prob\{gene\ i\ null | z_i = z\} \\ &= p_0 f_0(z)/f(z) \end{aligned}}$$

$Fdr(z)$ Replace densities with cdfs

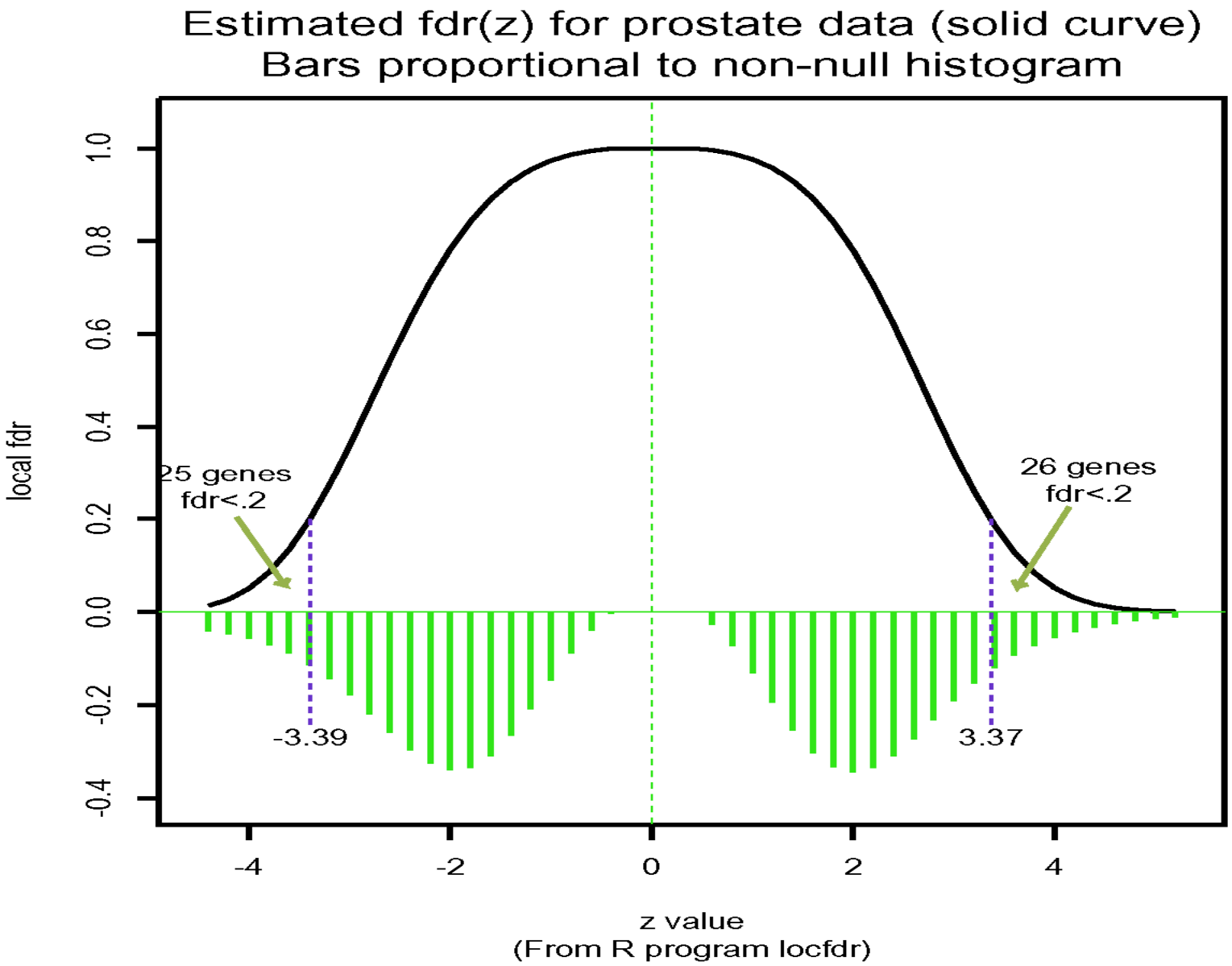
Empirical Bayes

- Estimate mixture density *f(z)* from observed z-values $z_1, z_2, \ldots, z_N$:

$$\widehat{fdr}(z) = p_0 f_0(z) / \widehat{f}(z)$$

- $BH : p_0 = 1 \quad (\widehat{p}_0 = 0.94.)$

- *Don't need*: z_i's independent, *t*-tests …


Estimated fdr(z) for prostate data (solid curve)
Bars proportional to non-null histogram
local fdr
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
25 genes
fdr<.2
26 genes
fdr<.2
-3.39
3.37
-4
-2
0
2
4
z value
(From R program locfdr)

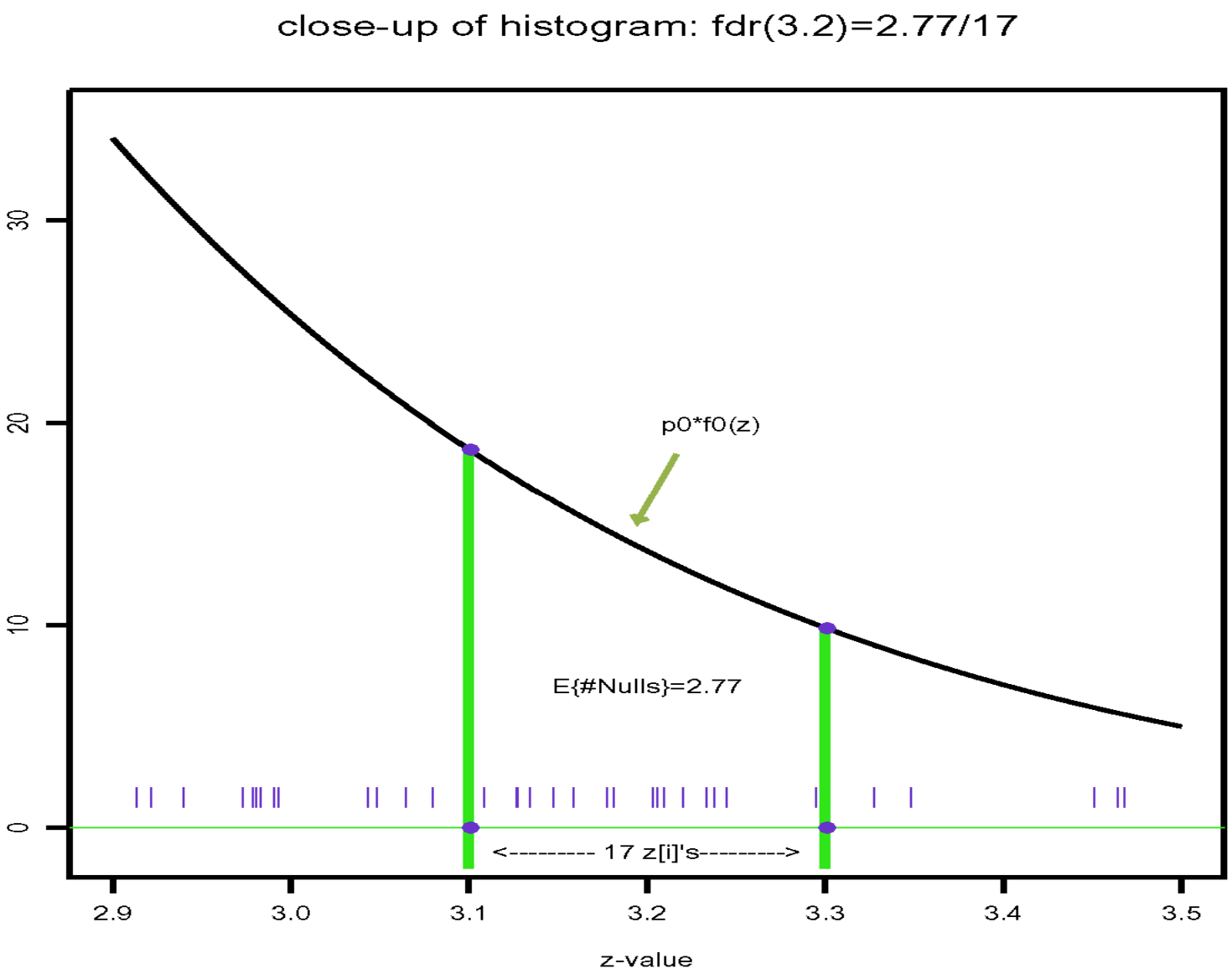
Basic Fdr Idea

- Histogram has 49 bins, width $\Delta = .2$
- $\text{bin}_{39} = [3.1, 3.3]$; count $y_{39} = 17$
- $y_{39}^{(o)} = \#\{\text{null genes in bin}_{39}\} = ??$
- $\widehat{y}_{39}^{(o)} = E\#\{\text{null genes in bin}_{39}\} = 2.77$

  $(= 6033 \cdot 0.94 \cdot \varphi(3.2) \cdot 0.1)$

- $\widehat{fdr}(3.2) = 2.77/17 = .16$

- *About one sixth of the 17 genes in* bin_{39} *are false discoveries*


close-up of histogram: fdr(3.2)=2.77/17
30
20
10
0
p0*f0(z)
E{#Nulls}=2.77
<--------- 17 z[i]'s--------->
2.9
3.0
3.1
3.2
3.3
3.4
3.5
z-value

The Non-Null Counts

- $y_k = \text{observed count in } k^{th} bin$
- $1 - fdr(z_i) = Prob\{gene\ i \text{ non-null} | z_i = z\}$
- So estimated number non-nulls in bin_k is

$$\widehat{y}_k^{(1)} = [1 - \widehat{fdr}_k] y_k$$

where $\widehat{fdr}_k = \widehat{fdr}(z = \text{midpoint bin}_k)$.

- Plotted bars used smoothed version.

Power Diagnostics
(Efron 2006, Section 3)

Good Power: $\widehat{fdr}_k$ small where $\widehat{y}_k^{(1)}$ big

Expected Non-Null fdr

$$\widehat{Efdr}^{(1)} = \sum \widehat{fdr}_k \cdot \widehat{y}_k^{(1)} / \sum \widehat{y}_k^{(1)}$$

Prostate Data $\widehat{Efdr}^{(1)} = 0.68$ (Bad!)

- "Why aren't our favorite genes on your list of non-null cases?"

Increased Sample Size

- Multiply number microarrays by "c"

  ($c = 2$: 100 non-tumor men, 104 tumor)

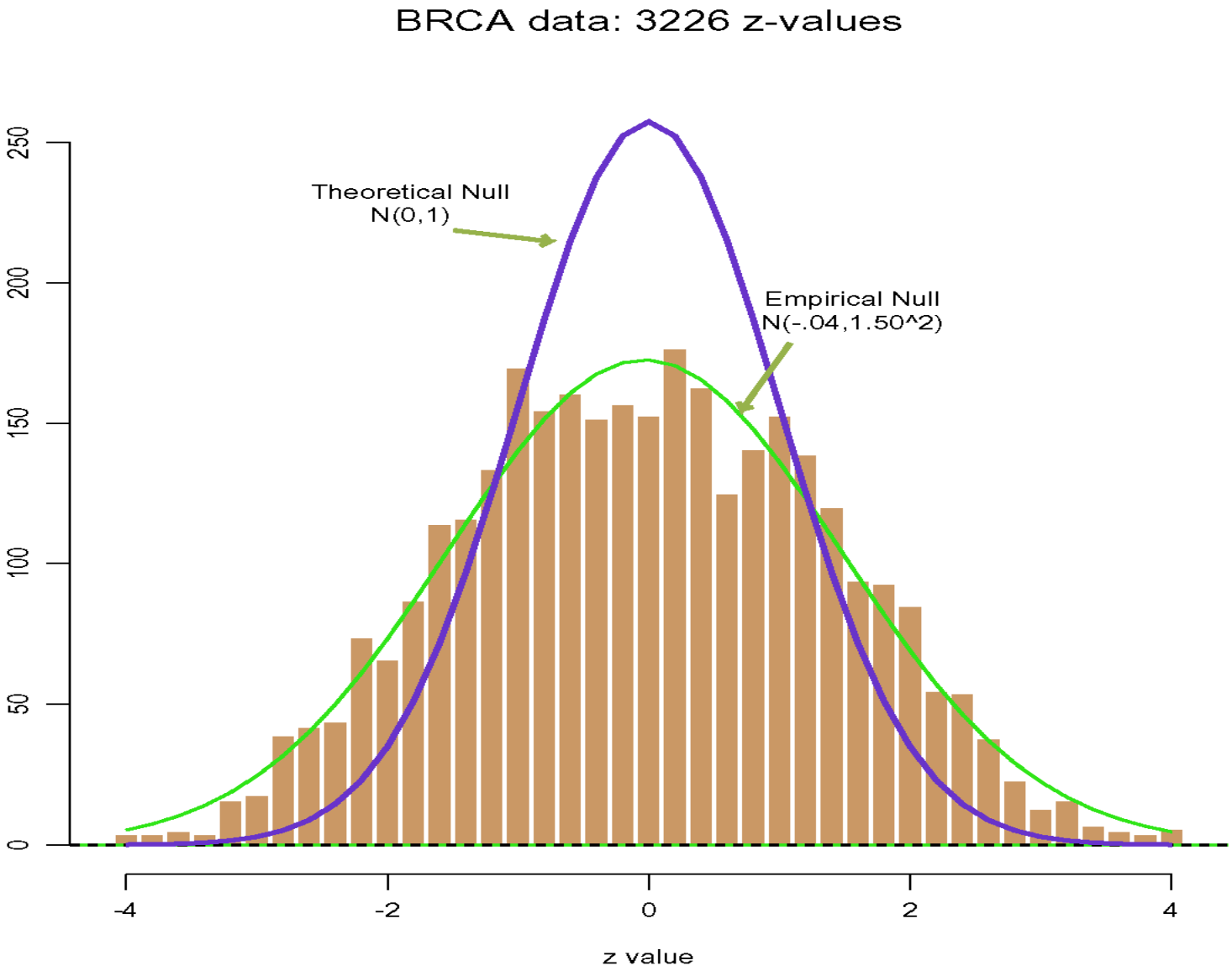
- Can estimate improvement in $\widehat{E}fdr^{(1)}$:

c:	1	1.5	2	2.5	3
$\widehat{E}\mathrm{fdr}^{(1)}$	.68	.54	.44	.38	.34

The BRCA Data
(Hedenfalk et al. 2001)

- *Microarray study* comparing tumors from women with BRCA1 or BRCA2 mutations
- *15 microarrays:* 7 BRCA1, 8 BRCA2,

  same 3226 genes: $X \;\; 3226 \times 15$
- $t_i =$ two-sample t stat for gene i, BRCA2 vs BRCA1
- $z_i = \Phi^{-1}(F_{13}(t_i))$
- *Theoretical Null* $z_i \sim N(0,1)$


BRCA data: 3226 z-values
Theoretical Null
N(0,1)
Empirical Null
N(-.04,1.50^2)
250
200
150
100
50
0
-4
-2
0
2
4
z value

Four Arguments Against the Theoretical Null

- Central histogram doesn't match theoretical $N(0,1)$ null

  $(\text{requires } p_0 \leq 0.7).$

- Central hist matches "empirical null" $N(-.04, 1.50^2)$

- Four Reasons Why null z_i's *not* $N(0,1)$

- **Reason 1** *Failed Assumptions:*

- Maybe nonnormality of microarray measurements Distorts student's *t* distribution for "t_i"

- *Permutation Null*: Scramble the 102 microarrays (gives $z_i \dot{\sim} N(0,1)$ for prostate data)

Reason 2: Unobserved Covariates
(Efron 2004, Section 4)

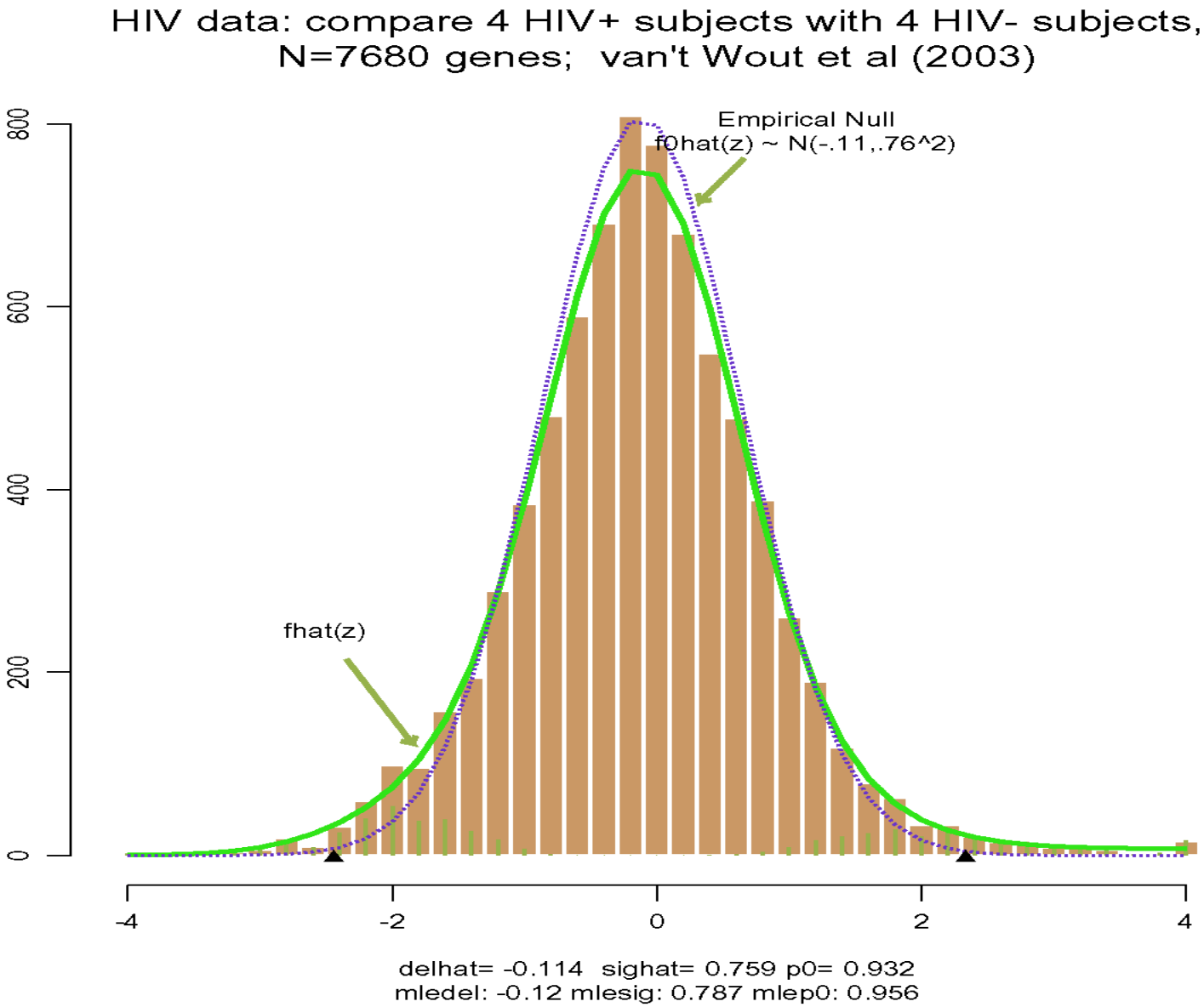
- *BRCA Study Observational* Unobserved Covariates
  Age, Wt, Stage, Race …
- *If observed* would be factored out of z_i's
- Tends to *widen* null density $f_0(z)$
- Could account for BRCA histogram
- *Won't show up in permutation distribution*

Reason 3: Correlation Across Arrays

- $Gene\ i\ data:\ (x_{i1}, x_{i2}, ..., x_{i7}), (x_{i8}, x_{i9}, ..., x_{i15})$
- Student-*t* null density assumes independence across microarrays
- Principle Component Analysis showed correlation
  among $(x_{i8}, x_{i9}, x_{i10}, x_{i11})$, likewise $(x_{i12}, x_{i13}, x_{i14}, x_{i15})$
- "Less than 13 df"
- Not detectable from permutations

Reason 4: Correlation Across Genes (Efron 2007)

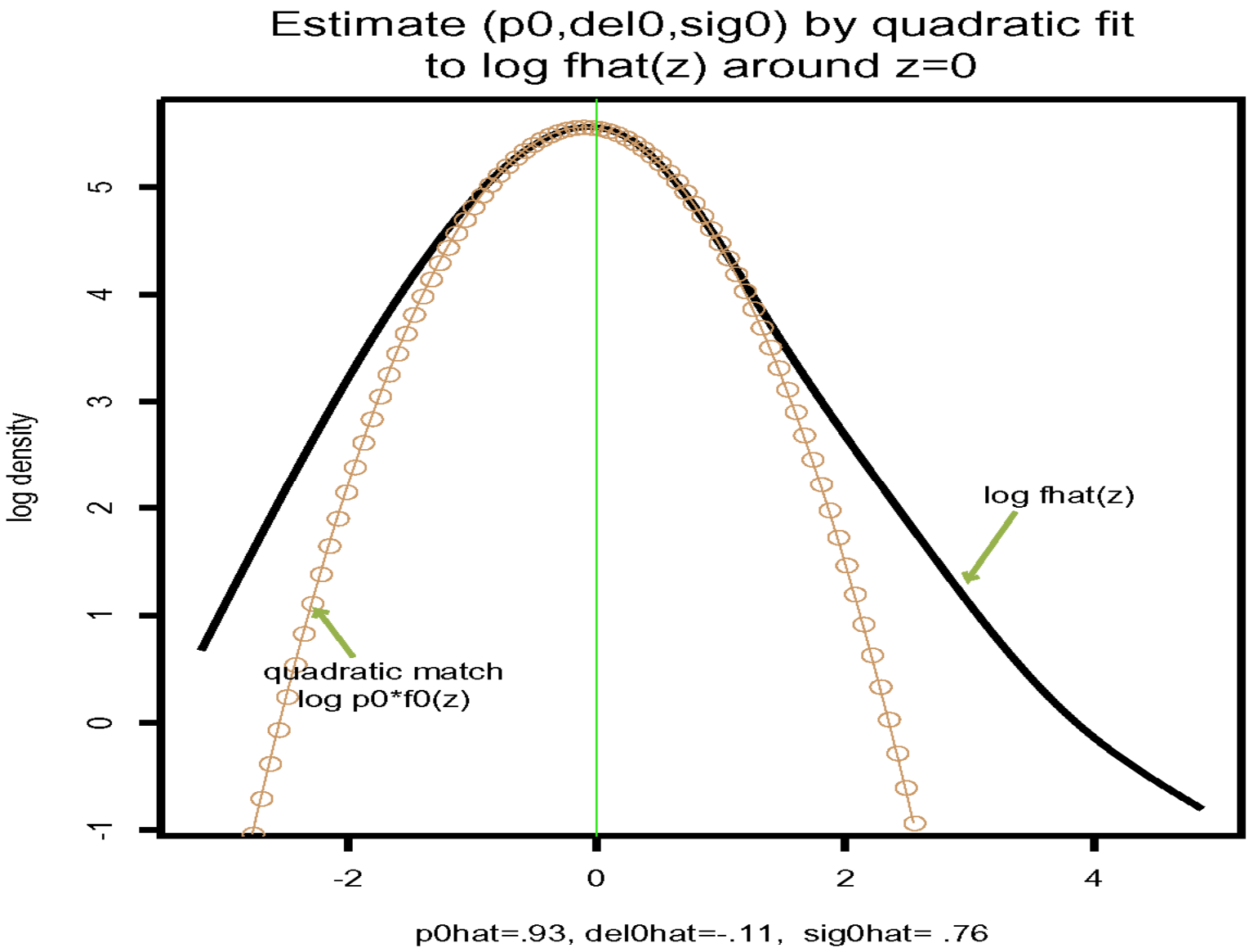
- $\widehat{fdr}(z) = p_0 f_0(z)/\widehat{f}(z)$ does **not** require independence of gene measurements.
- *However*: gene-wise correlations affect "f_0 "
- *BRCA:* 5 million pairwise correlations
  rms correlation = 0.15
- Even if $f_0(z_i) \sim N(0,1)$ individually, can behave as $N(\theta, \sigma_0^2)$ collectively, with $.75 < \sigma_0 \leq 1.25$.
- Not detectable from permutations.


HIV data: compare 4 HIV+ subjects with 4 HIV- subjects,
N=7680 genes; van't Wout et al (2003)
Empirical Null
f0hat(z) ~ N(-.11,.76^2)
fhat(z)
800
600
400
200
0
-4
-2
0
2
4
delhat= -0.114 sighat= 0.759 p0= 0.932
mledel: -0.12 mlesig: 0.787 mlep0: 0.956

Empirical Null Estimation

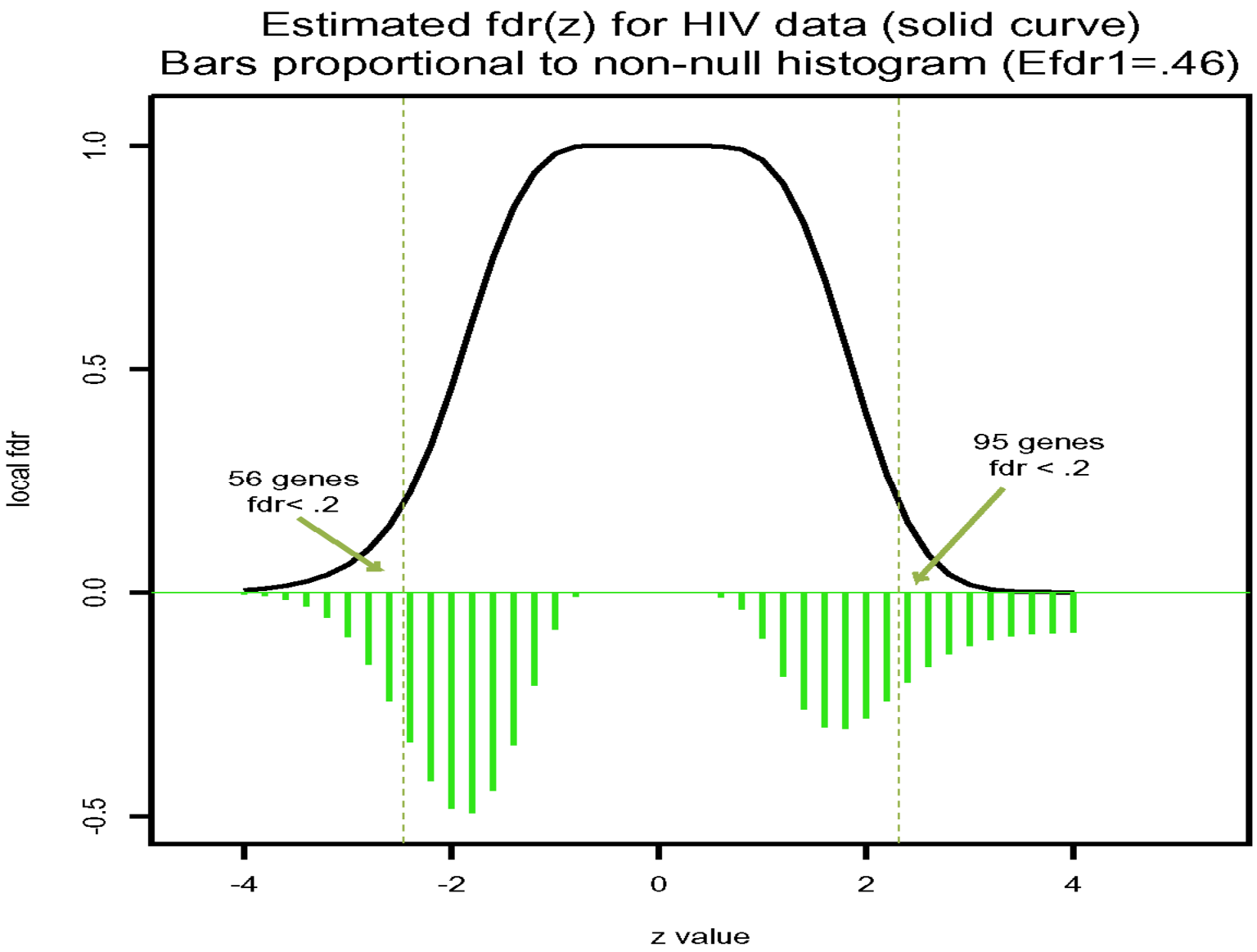
$$\text{fdr}(z) = p_0 f_0(z)/f(z)$$

- Theoretical $N(0,1)$ may not fit z-value histogram central peak
- *Idea* Assume $f_0(z) \sim N(\delta_0, \sigma_0^2)$; fit p_0, δ_0, σ_0 from histogram counts near $z = 0$. ["zero assumption"]
- *Central Matching* :
  (1) Plot $\log \widehat{f}(z)$
  (2) Find best quadratic match near $z = 0$
  (3) coeffs of match $\Rightarrow (\widehat{p}_0, \widehat{\delta}_0, \widehat{\sigma}_0)$
- Nearly unbiased for $(\widehat{\delta}_0, \widehat{\sigma}_0)$, Efron (2004).


Estimate (p0,del0,sig0) by quadratic fit
to log fhat(z) around z=0
log density
-1
0
1
2
3
4
5
-2
0
2
4
log fhat(z)
quadratic match
log p0*f0(z)
p0hat=.93, del0hat=-.11, sig0hat= .76

Direct Maximum Likelihood Estimation of $(p_0, \delta_o, \sigma_0)$

- Assume all the z_i's in $[-x_0, x_0]$ are from $f_0(z)$, the null density
- Let $\mathbf{Z}_0 = (Z_1, Z_2, ..., Z_{N_0})$ be them.
- Then $\mathbf{Z}_0$ follows a truncated $N(\delta_0, \sigma_0^2)$ distribution: can estimate $(p_0, \delta_0, \sigma_0)$.
- More biased, less variable than central matching method


Estimated fdr(z) for HIV data (solid curve)
Bars proportional to non-null histogram (Efdr1=.46)
local fdr
1.0
0.5
0.0
-0.5
56 genes
fdr< .2
95 genes
fdr < .2
-4
-2
0
2
4
z value

Large-Scale Simultaneous Testing

- *Not just a lot of classical single tests*
  - Multiplicities
  - Empirical Bayes
- Can learn things you didn't want to know
- Permutation methods not cure-alls
- **Modelling**: Better to minimize
  - Model inside of X ?
  - Big data sets should supply own models

References

Efron (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *JASA* **99**, 96–104.

Efron (2006). Microarrays, empirical Bayes, and the two-groups model. http://www-stat.stanford.edu/~brad/papers/twogroups.pdf

Efron (2007). Correlation and large-scale simultaneous significance testing. *JASA* **102**, 93–103.

Singh et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1:302–309.

Hedenfalk et al. (2001). Gene expression profiles in hereditary breast cancer. *N. Engl. J. Med.* 539–548.

Van't Wout et al. (2003). Cellular human gene expression upon human immunodeficiency versus type 1 infection of CDS + T-Cell lines. *J. Virol.* 1392–1402.

"locfdr" R program, available on CRAN on Efron site above.