Simultaneous Inference: When Should Hypothesis Testing Problems be Combined?

Bradley Efron
Stanford

Large-Scale Simultaneous Hypothesis Testing

New Technology: microarrays, fMRI,
satellite imaging, proteomics

Simultaneous Inference: hundreds or thousands of
problems presented at the same time

Methodology: Bonferroni, FDR, FWER, step-down...

Question: Should all the problems be analyzed together,
or analyzed in separate subgroups of some sort?

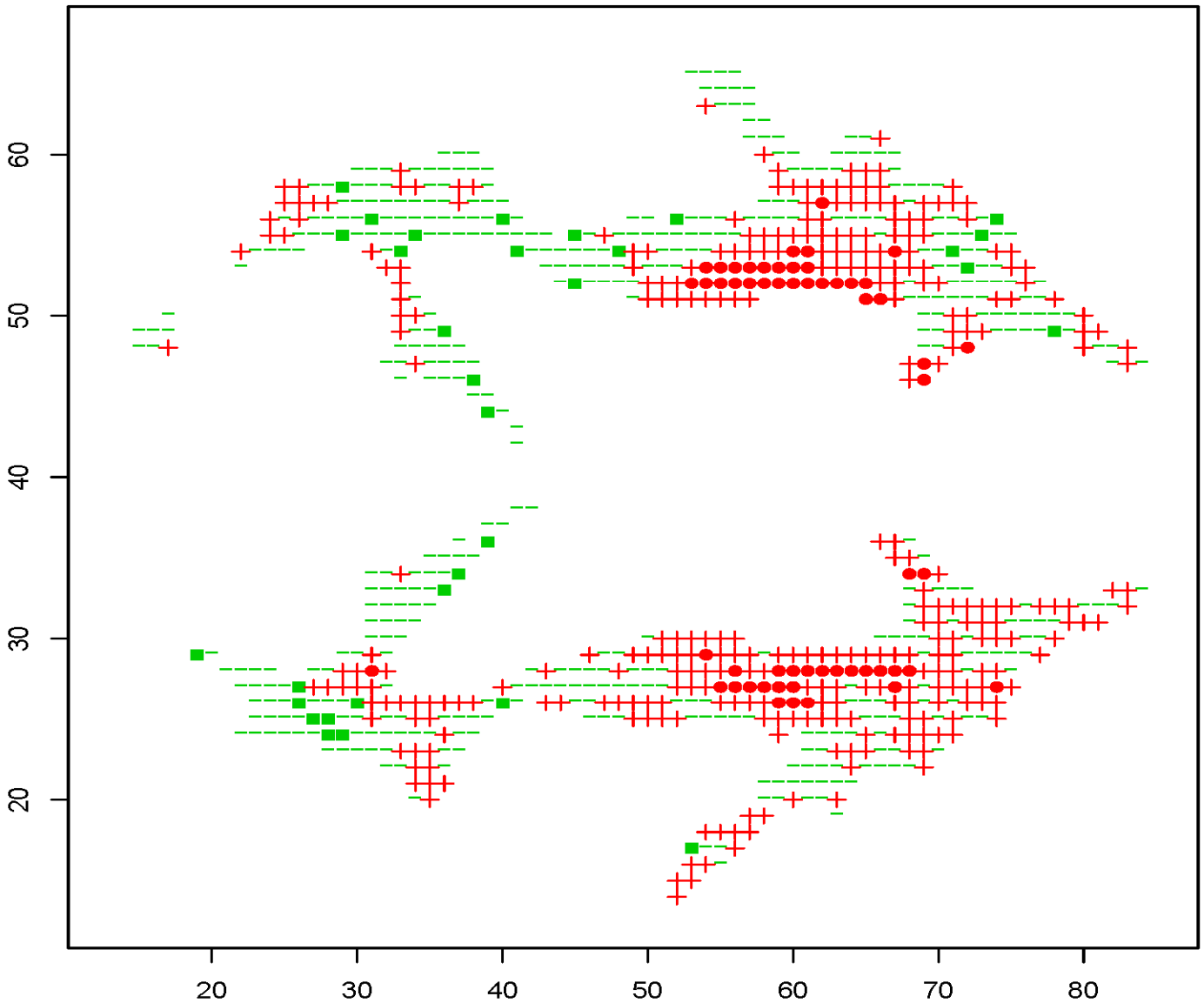
The Brain Data

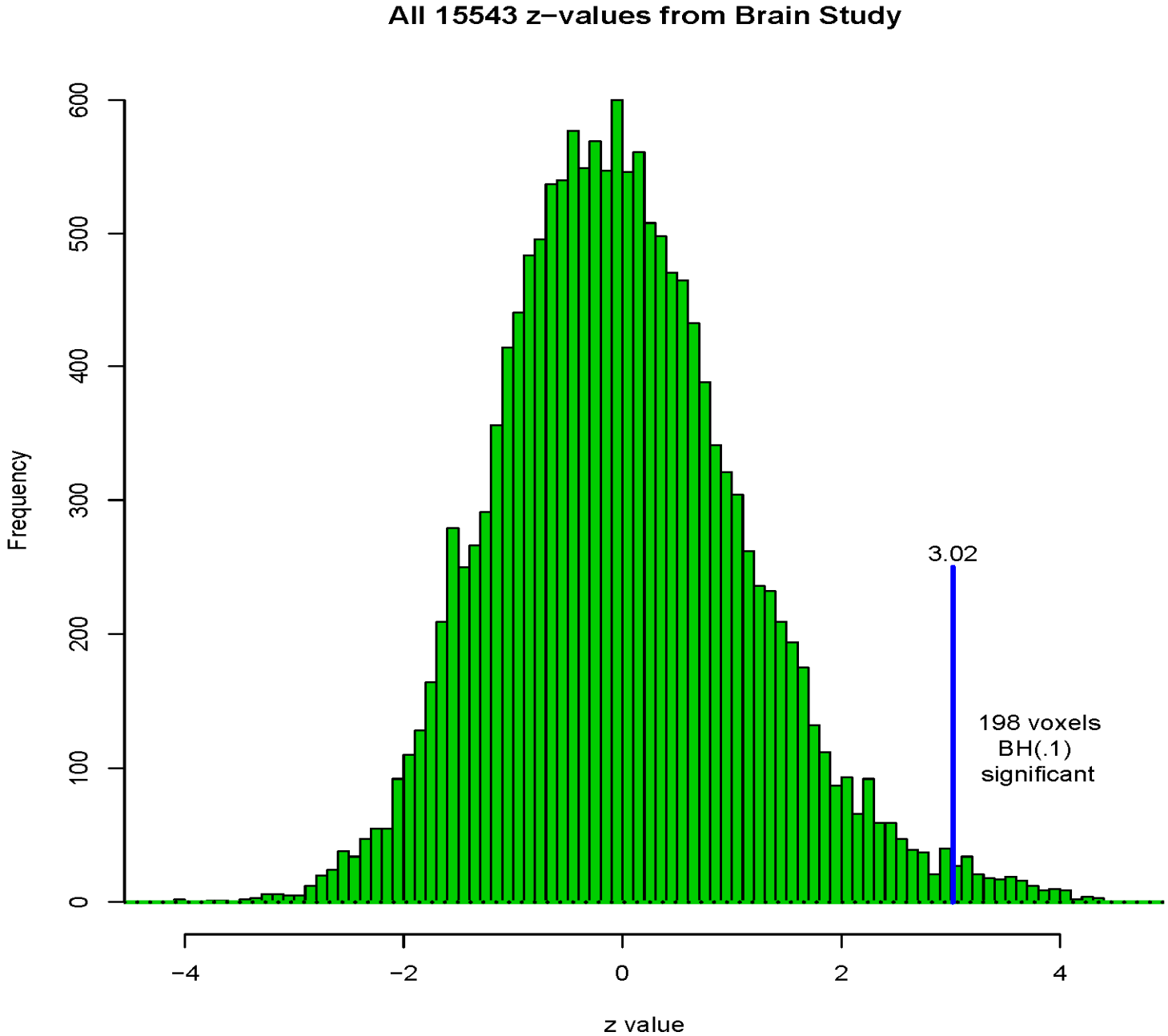
(Schwartzman et al. 2005)

- *12 children*, 6 dyslexic, 6 normal
- *Diffusion Tensor Image*, $N = 15445$ voxels for each child
- **z-values** "z_i" comparing dyslexics vs normals

  for each voxel, $i = 1, 2, \ldots, N$.

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


Brain data: z−values comparing 6 normals vs 6 dyslexics;
Red >0, Green <0; solid >2; 848 of 15443 voxels
60
50
40
30
20
20
30
40
50
60
70
80
Midbrain Horizontal Slice; front −−>
Schartzman et al (2004)


All 15543 z−values from Brain Study
Frequency
0
100
200
300
400
500
600
3.02
198 voxels
BH(.1)
significant
−4
−2
0
2
4
z value

FDR Analysis

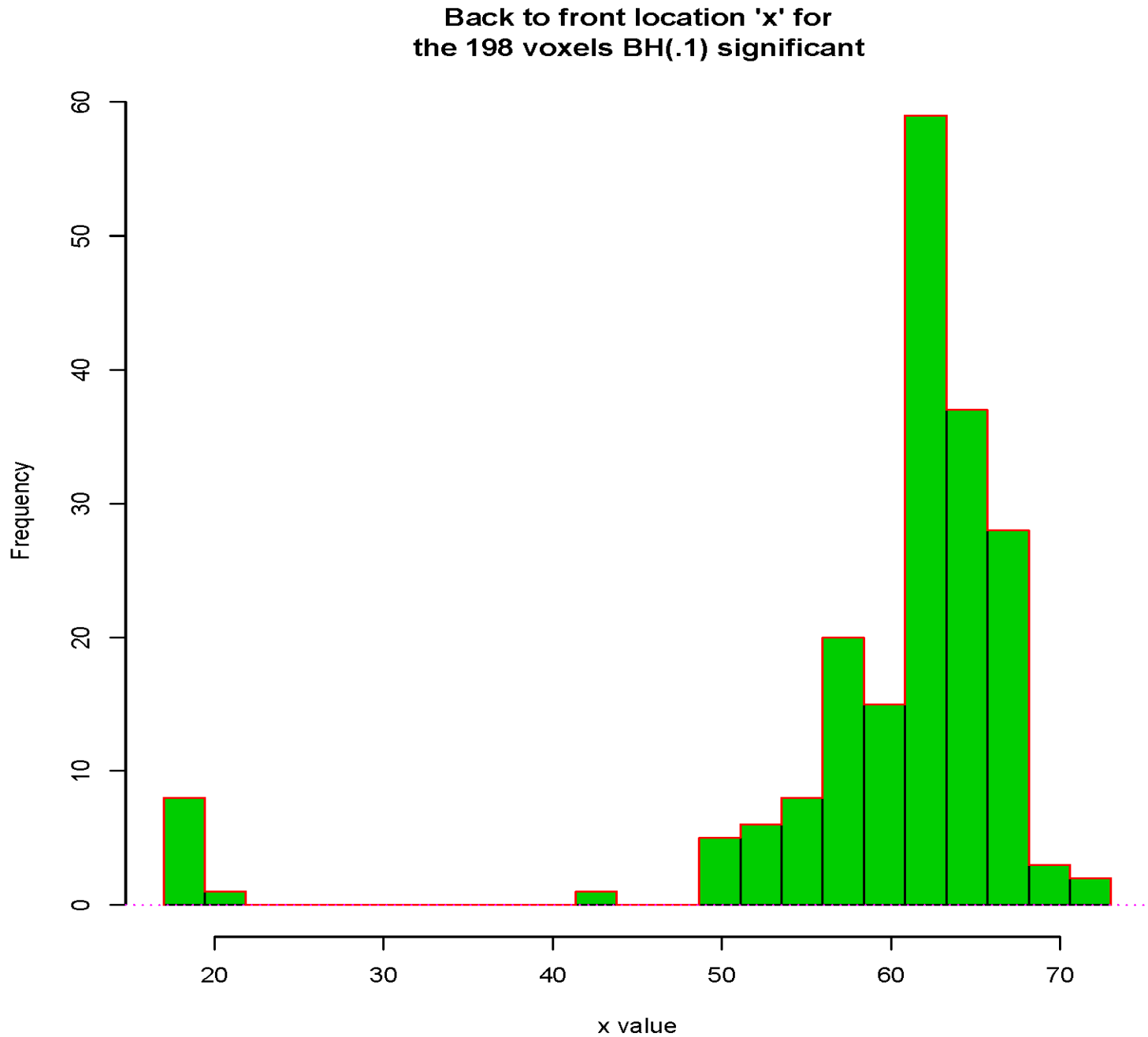
- Standard Benjamini-Hochberg FDR analysis identified 198 "significant" voxels at control level $q = 0.1$.
- Cutoff value

$$\boxed{z_i \geq 3.02}$$

- Most of 198 near distance

$$x = 65$$

from back of brain (red circle voxels in panel 4).

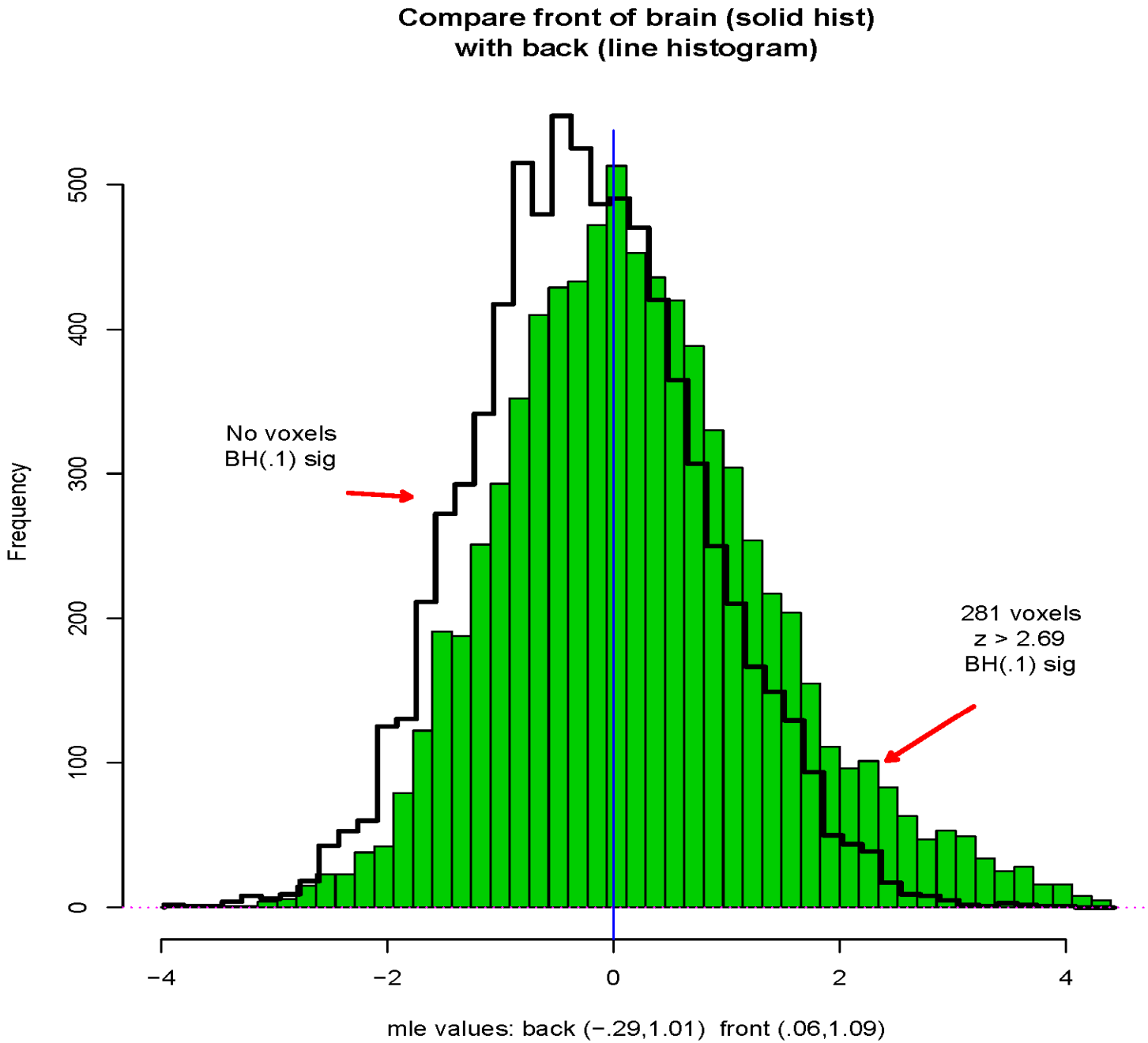

Back to front location 'x' for
the 198 voxels BH(.1) significant
Frequency
0
10
20
30
40
50
60
20
30
40
50
60
70
x value

Combined or Separate Analyses?

- Second Thoughts Maybe front and back half of brain are much different, and should be analyzed separately.

- **Separate BH(.1) FDR Analyses**:

Back half	No significant voxels
Front half	281 voxels with $z_i \geq 2.69$


Compare front of brain (solid hist)
with back (line histogram)
Frequency
0
100
200
300
400
500
No voxels
BH(.1) sig
281 voxels
z > 2.69
BH(.1) sig
−4
−2
0
2
4
mle values: back (−.29,1.01) front (.06,1.09)

Are Separate Analyses Legitimate?

- **Bonferroni** Reject Null for p-value $\leq \alpha/N$,

  so separate analyses with $N/2$ are less conservative,

  (Need $\alpha \to \alpha/2$.)

- **False Discovery Rates**

  Separate $BH(q)$ analyses $\Rightarrow$ combined $BH(q)$

Rest of talk: Fdr calculations (mostly Bayesian)

The Two-Groups Model
(Efron 2006)

- *N cases* (voxels, genes, etc.), each with z-value z_i
- Each case either null or non-null, prior probability p_0 or p_1
- z_i has density $\left\{ \begin{matrix} f_0(z) & \text{if null} \\ f_1(z) & \text{if non-null} \end{matrix} \right\} f(z) = p_0 f_0(z) + p_1 f_1(z)$

False Discovery Rates

- *Local false discovery rate:* $\text{fdr}(z) = p_0 f_0(z)/f(z)$

  $= \text{Prob}\{\text{null}|z\text{-value} = z\}$

- *Tail Area Fdr* F_0, F_1 cdfs for f_0, f_1, and

  $F(z) = p_0 F_0(z) + p_1 F_1(z)$; Bayes False Discovery Rate is

  $\text{Fdr}\ (z) = p_0 F_0(z)/F(z) = \text{Prob}\{\text{null}|z\text{-value} \leq z\}$

Bayes and Empirical Bayes Fdr's

- *Bayesian* $\quad \mathrm{Fdr}(z) = p_0 F_0(z)/F(z)$ has empirical Bayes est

  $\overline{\mathrm{Fdr}}(z) = p_0 F_0(z)/\bar{F}(z) \qquad [\bar{F}(z)$ empirical cdf of z_i's$]$

- **BH rule** Reject H_0 for z_i with $\overline{\mathrm{Fdr}}(z_i) \le q$.

  (Usually taking $p_0 = 1, F_0 =$ theoretical null.)

- **Coefficient of Variation**

  $N(z) \equiv \#\{z_i \le z\}: \quad CV(\overline{\mathrm{Fdr}}(z)) \doteq 1/\sqrt{E\{N(z)\}}$

- Need $E\{N(z)\} \ge 10$ for $CV \le 0.3$

  · Perhaps $N \ge 1000$ (≥ 3000 if $p_0 F_0(z)$ also estimated.)

Estimating local fdr(z)

("locfdr", CRAN)

- *Local* $\text{fdr}(z) = p_0 f_0(z)/f(z)$ estimated by

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

where $\widehat{f}(z)$ is smooth estimate based on all the z_i's,

$f_0(z)$ theoretical null, often $p_0 \doteq 1$. Need $N \geq 1000$

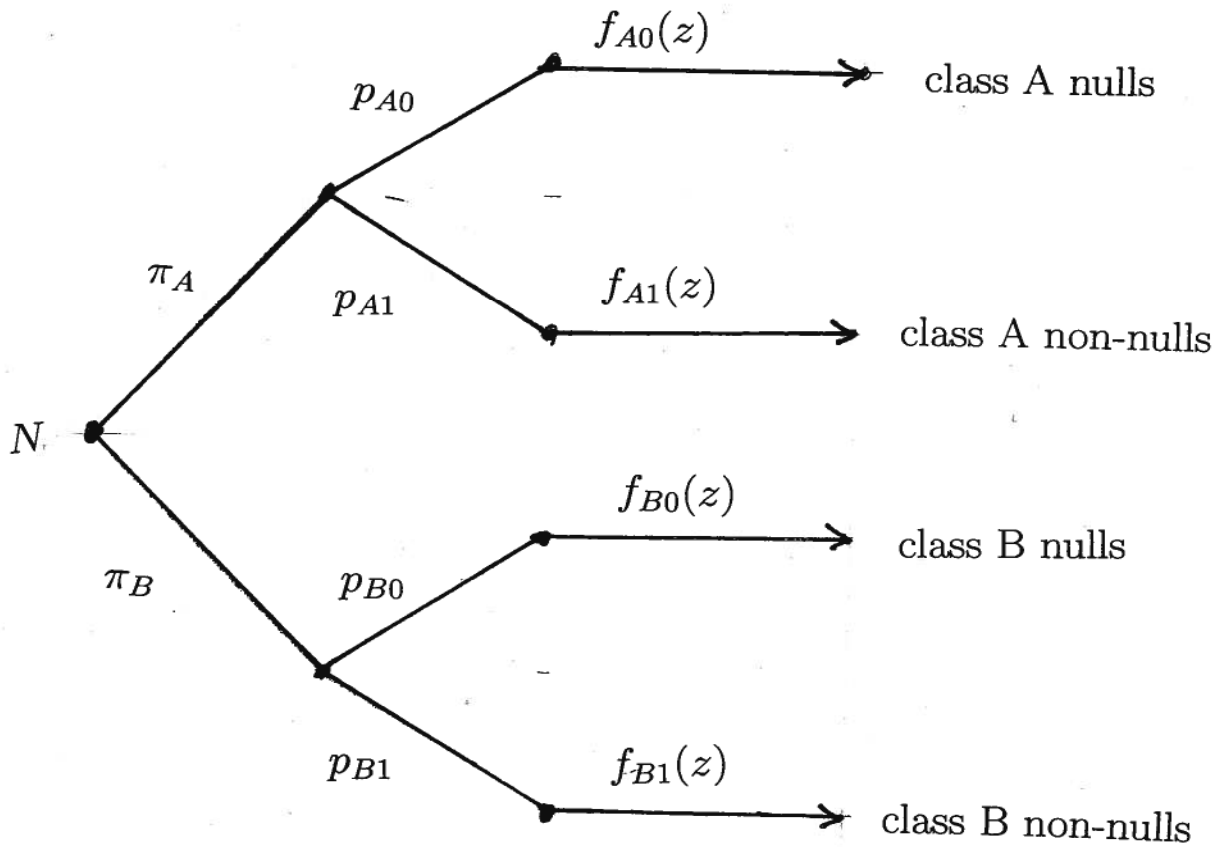
- Can also estimate p_0, f_0 from z_i's near 0,

$$\widehat{\text{fdr}}(z) = \widehat{p}_0 \widehat{f}_0(z)/\widehat{f}(z),$$

Need $N \geq 3000$.

The Two-Class Two-groups Model

- Classes "A" and "B" ("front" and "back"), prior probs π_A and π_B



fdr Relationship

- Let $\text{fdr}_A(z)$ be local fdr for class A by itself,

  and $\text{fdr}(z)$ marginal fdr ignoring class.

- **Theorem** $\text{fdr}_A = \text{fdr}(z) \cdot R_A(z), \;\; R_A(z) = \pi_{A0}(z)/\pi_A(z),$

where

$$\pi_A(z) = \text{Prob}\{\text{class } A | z\}$$
$$\pi_{A0}(z) = \text{Prob}\{\text{class } A | \text{null}, z\}$$

- $\pi_A(z)$ easy to estimate by logistic regression

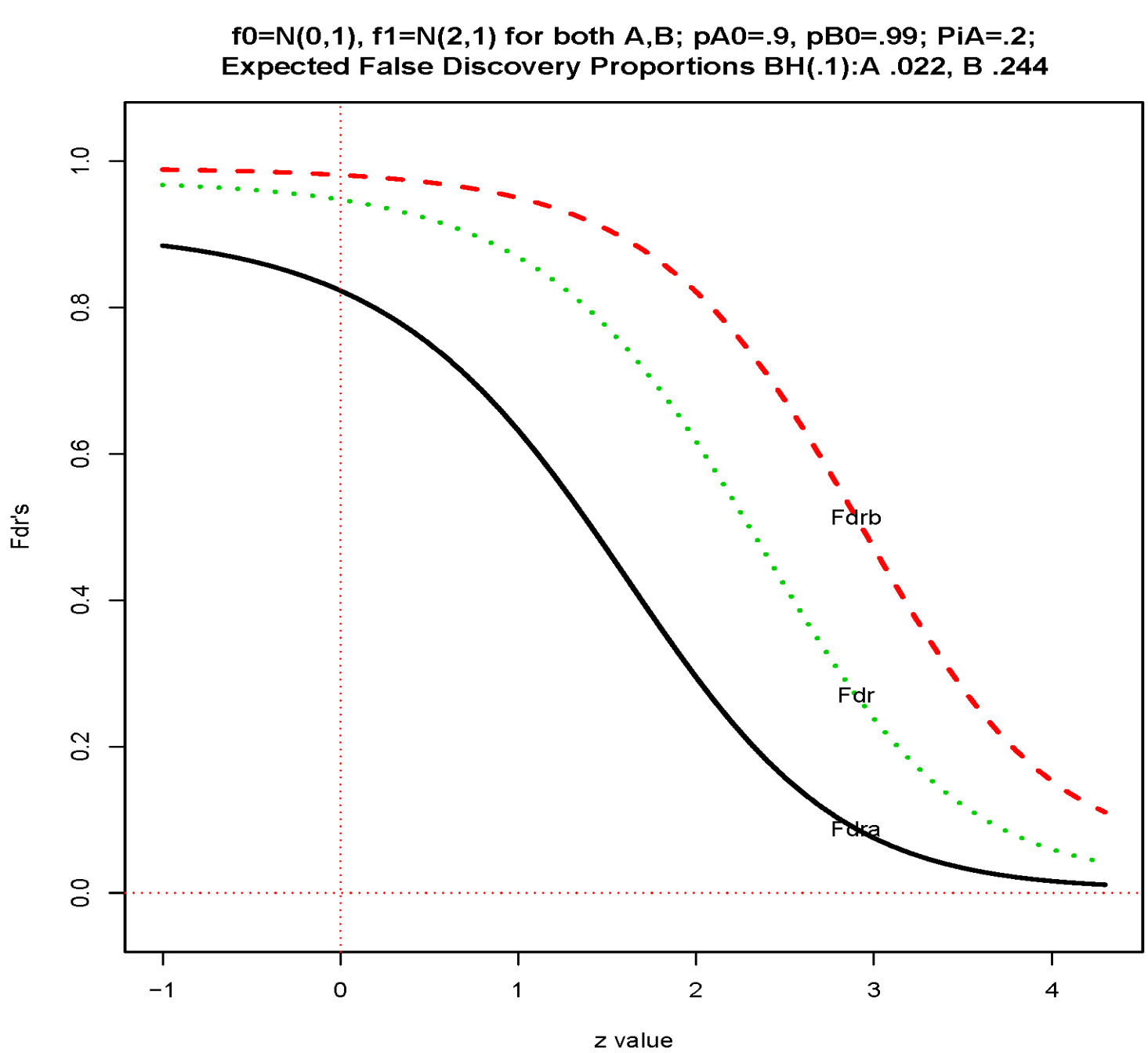
- If $p_{A0} f_{A0}(z) = p_{B0} f_{B0}(z)$ then $\pi_{A0}(z) = \pi_A$

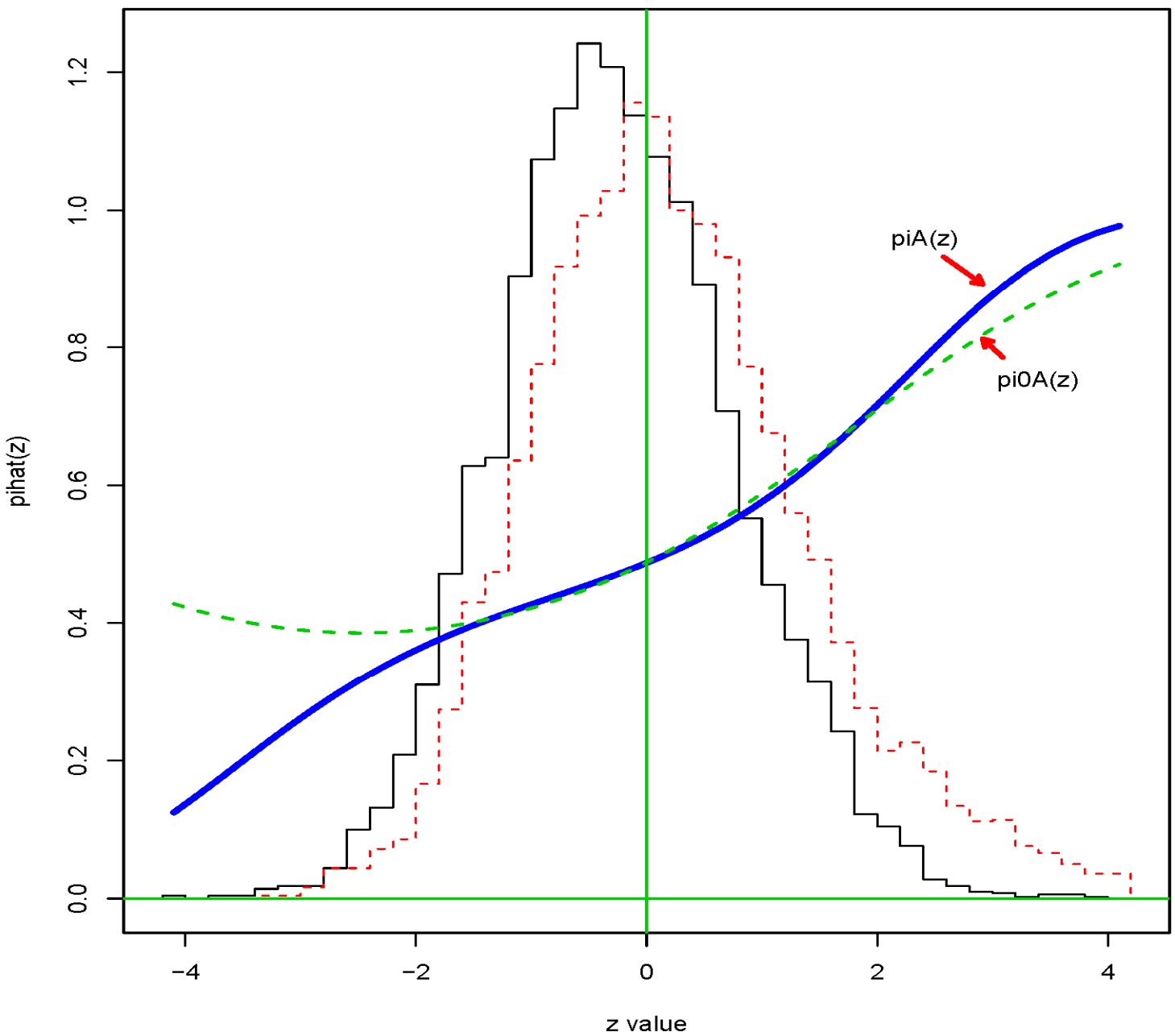
Dangers of Combination

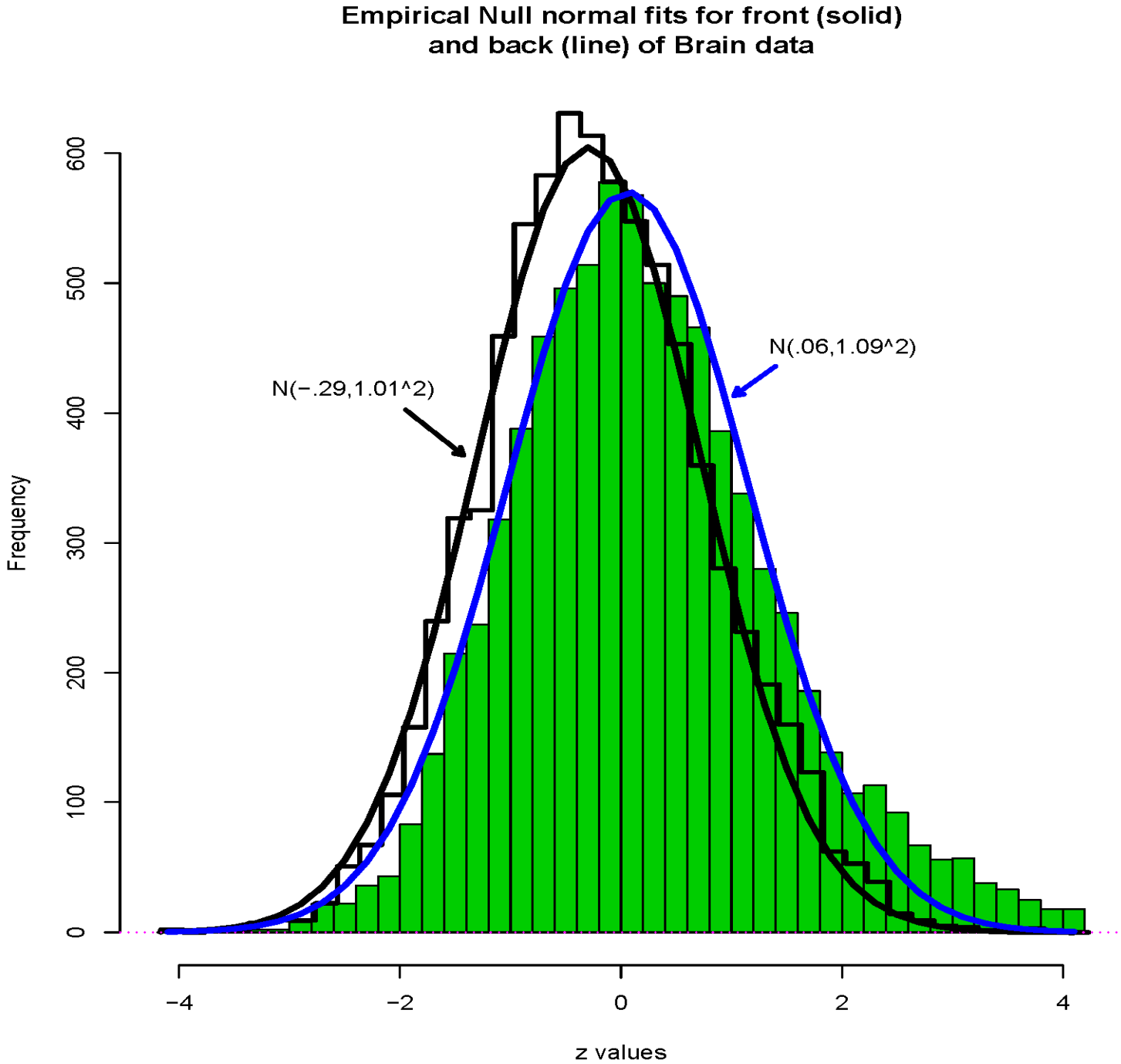
- Suppose "A" combined with all others *without using*

$$\text{fdr}_A(z) = \text{fdr}(z) \cdot R(z)$$

- *"A" rich in non-null cases*: $R_A(z) < 1$: $\text{fdr}(z) > \text{fdr}_A(z)$
- *"A" poor in non-null cases*: $R_A(z) > 1$: $\text{fdr}(z) < \text{fdr}_A(z)$
- Likewise for $\text{Fdr}_A(z)$, BH rule.


f0=N(0,1), f1=N(2,1) for both A,B; pA0=.9, pB0=.99; PiA=.2;
Expected False Discovery Proportions BH(.1):A .022, B .244
Fdr's
z value
Fdrb
Fdr
Fdra
0.0
0.2
0.4
0.6
0.8
1.0
−1
0
1
2
3
4


Estimate of piA(z) for A=front half brain voxels;
cubic logistic regression; Dashed is piA0(z)
piA(z)
pi0A(z)
pihat(z)
z value
0.0
0.2
0.4
0.6
0.8
1.0
1.2
−4
−2
0
2
4


Empirical Null normal fits for front (solid)
and back (line) of Brain data
Frequency
0
100
200
300
400
500
600
N(−.29,1.01^2)
N(.06,1.09^2)
−4
−2
0
2
4
z values

Estimating $\pi_{A0}(z)$

- **Empirical Null** In two-groups model: assume

  $f_0(z) \sim N(\delta_0, \sigma_0^2)$, estimate

  p_0, δ_0, σ_0 from histogram heights around $z = 0$, (Efron 2006)

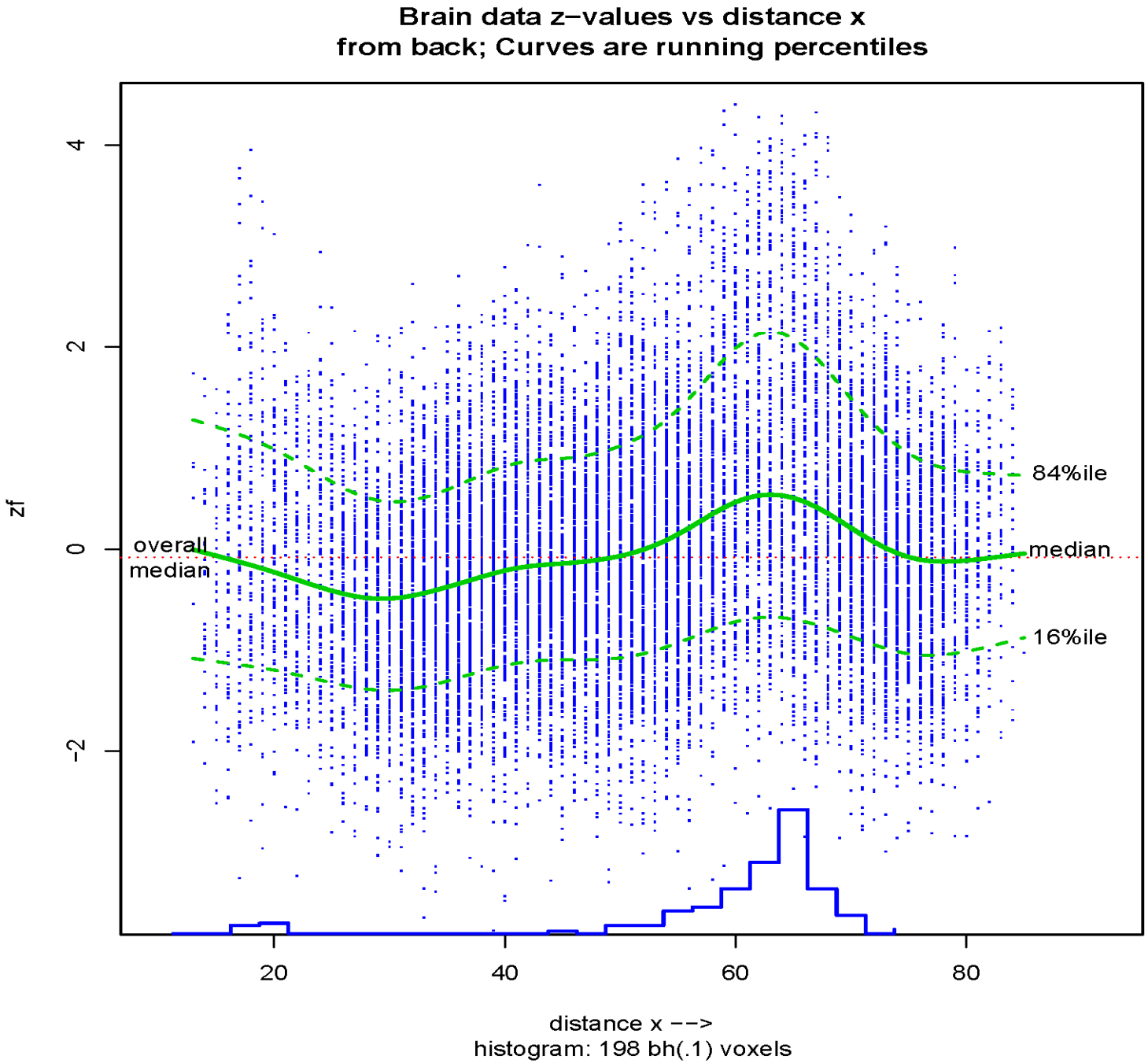
- *Brain Data*

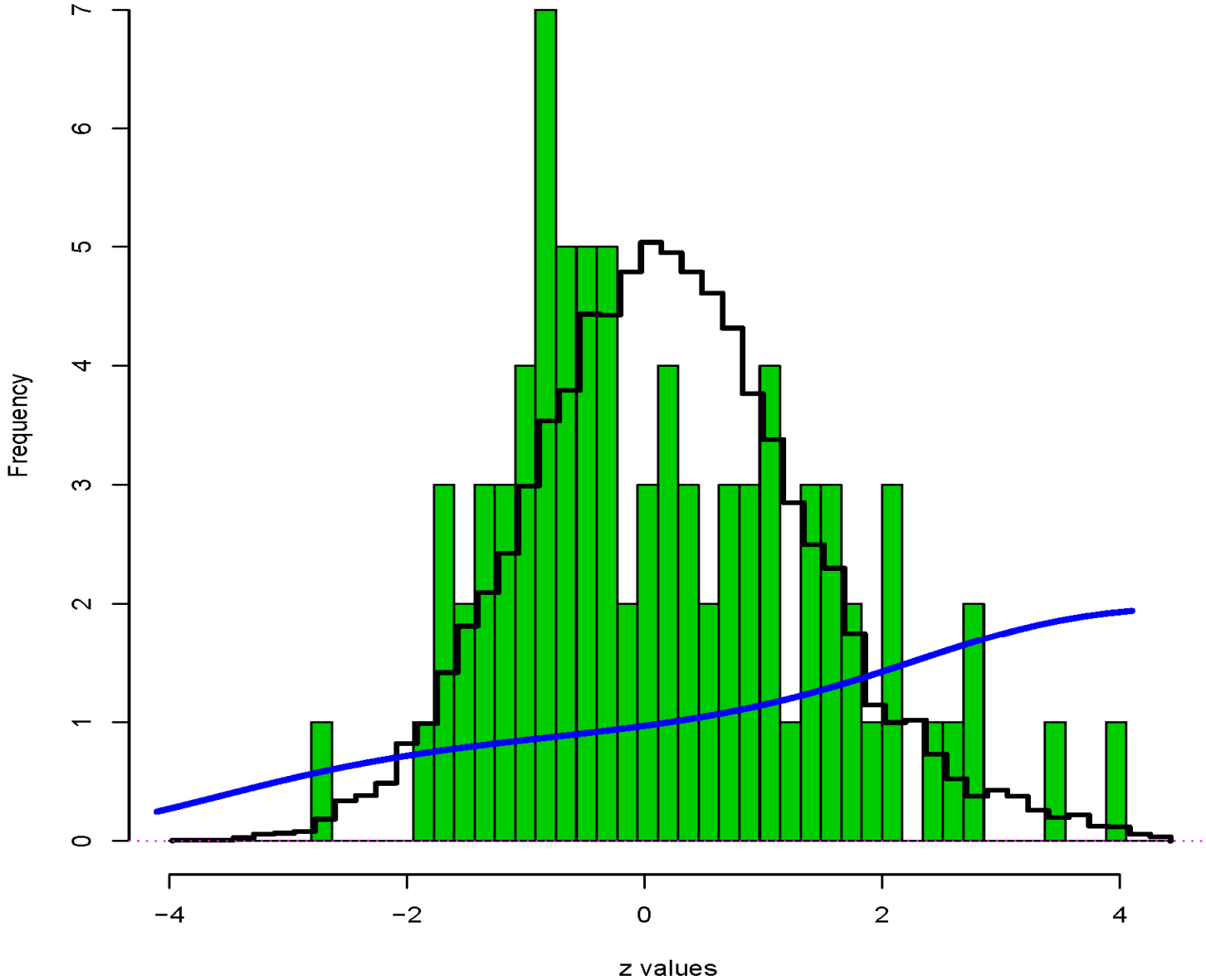
	$\widehat{p}_0$	$\widehat{\delta}_0$	$\widehat{\sigma}_0^2$	(π)
A (front):	.97	.06	1.01	(.50)
B (back):	1.00	-.29	1.09	(.50)

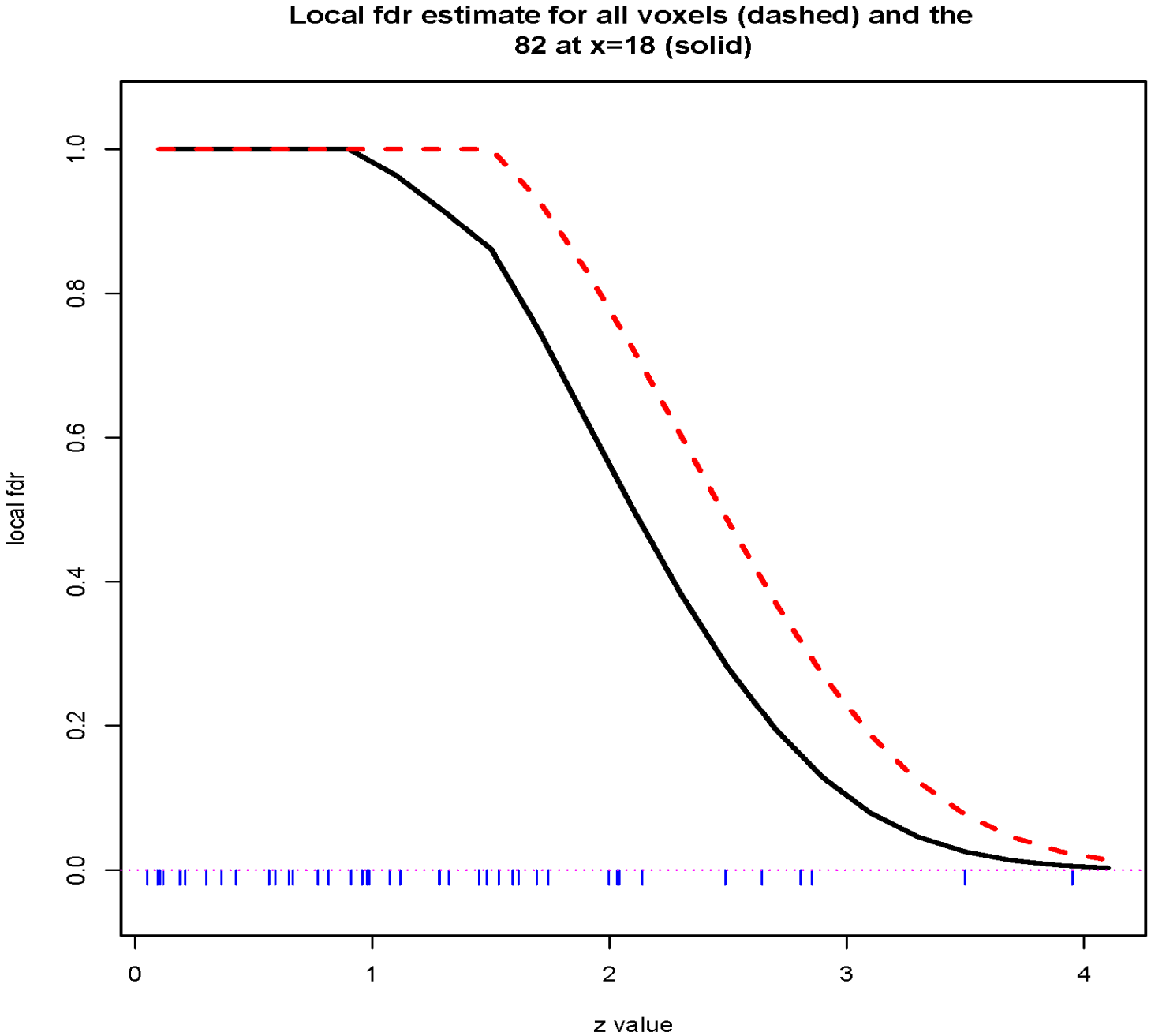
- $\widehat{\pi}_{A0}(z)$ not much different than $\widehat{\pi}_A(z)$

- (Not a coincidence that $\widehat{\pi}_{A0}(z) \doteq \widehat{\pi}_A(z)$ for z near 0.)

- $R_A(z) \doteq .94$ for $z \geq 3$


Brain data z−values vs distance x
from back; Curves are running percentiles
zf
4
2
0
-2
overall
median
84%ile
median
16%ile
20
40
60
80
distance x −−>
histogram: 198 bh(.1) voxels


82 voxels at distance x=18 (solid hist),compared to
all others; heavy line piA(z)/piA
Frequency
0
1
2
3
4
5
6
7
−4
−2
0
2
4
z values


Local fdr estimate for all voxels (dashed) and the
82 at x=18 (solid)
local fdr
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
z value

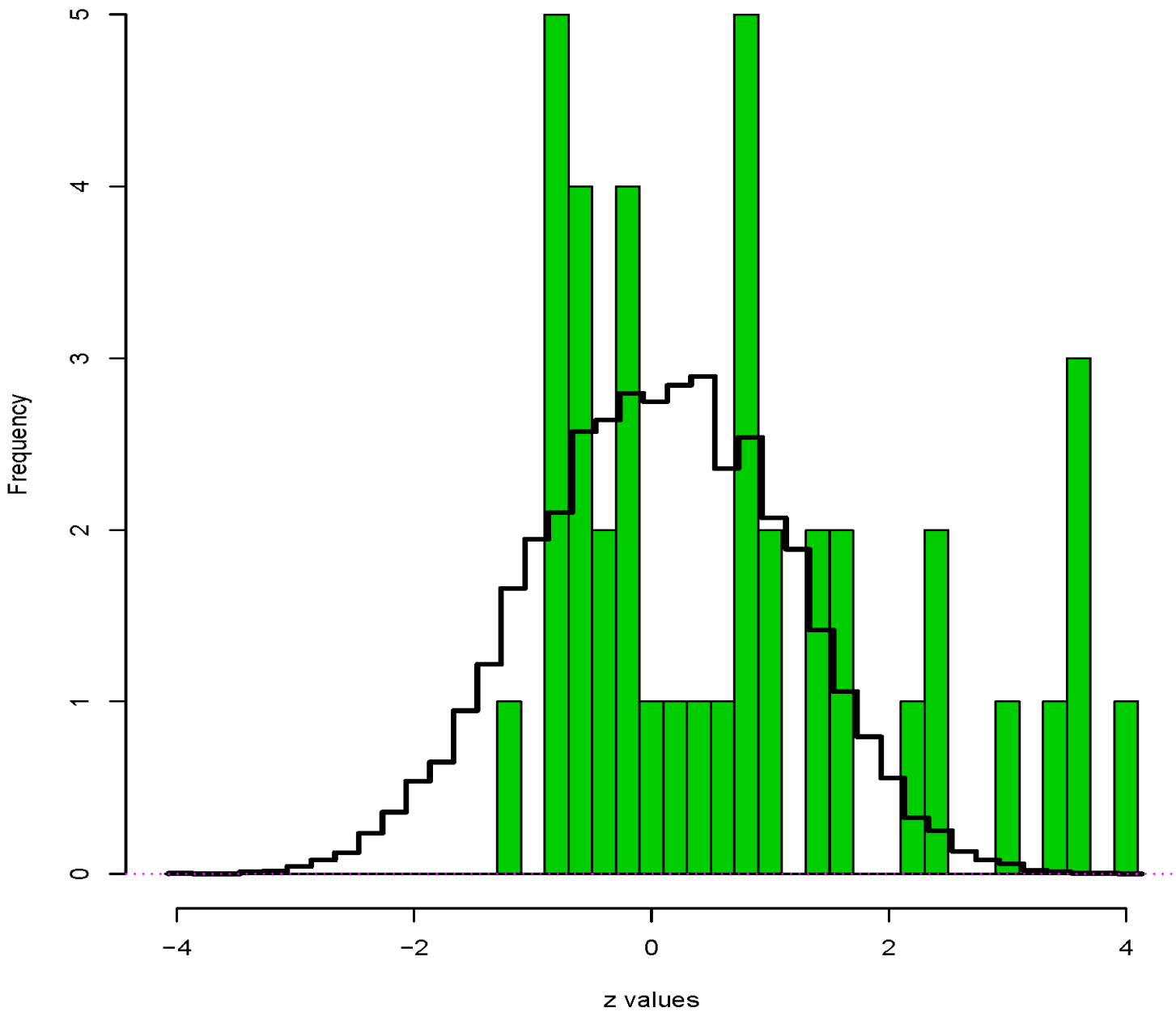
Estimating $\mathbf{fdr_A(z)}$ for small classes

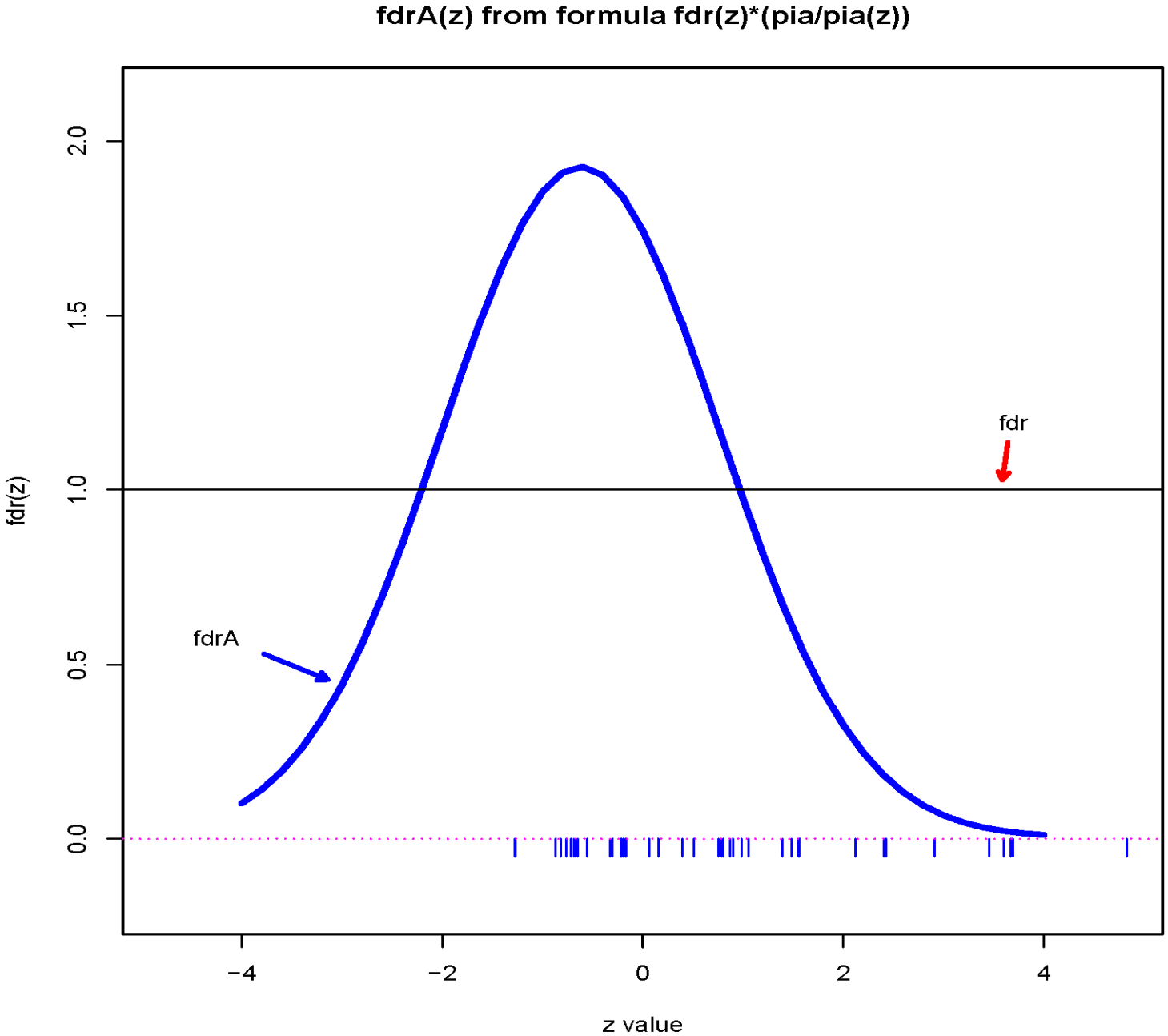
- **Theorem** $\log(\text{fdr}_A(z)) = \log(\text{fdr}(z)) - \log(\pi_A(z)) + \log(\pi_A)$

Estimation $\text{Var}\{\log \widehat{\text{fdr}}_A(z)\} \doteq \text{Var}\{\log \widehat{\text{fdr}}(z)\} + \text{Var}\{\log \widehat{\pi}_A(z)\}$

(orthogonal parameters)

- Formulas for 1st term in Efron (2005)
- Second term from standard logistic regression formulas.
- **More efficient** than estimating $\widehat{\text{fdr}}_A(z)$ from just "A" data


p53 data, N=10100 genes: compare geneset 'P52_UP',Na=40,
with all others; Efron and Tibshirani 2007
Frequency
0
1
2
3
4
5
−4
−2
0
2
4
z values


fdrA(z) from formula fdr(z)*(pia/pia(z))
fdr(z)
0.0
0.5
1.0
1.5
2.0
fdrA
fdr
−4
−2
0
2
4
z value

Fold-Change or t-statistics?

(Guo et al. 2006)

Fold Change (log scale) $\bar{x}_i - \bar{y}_i$ more consistent than

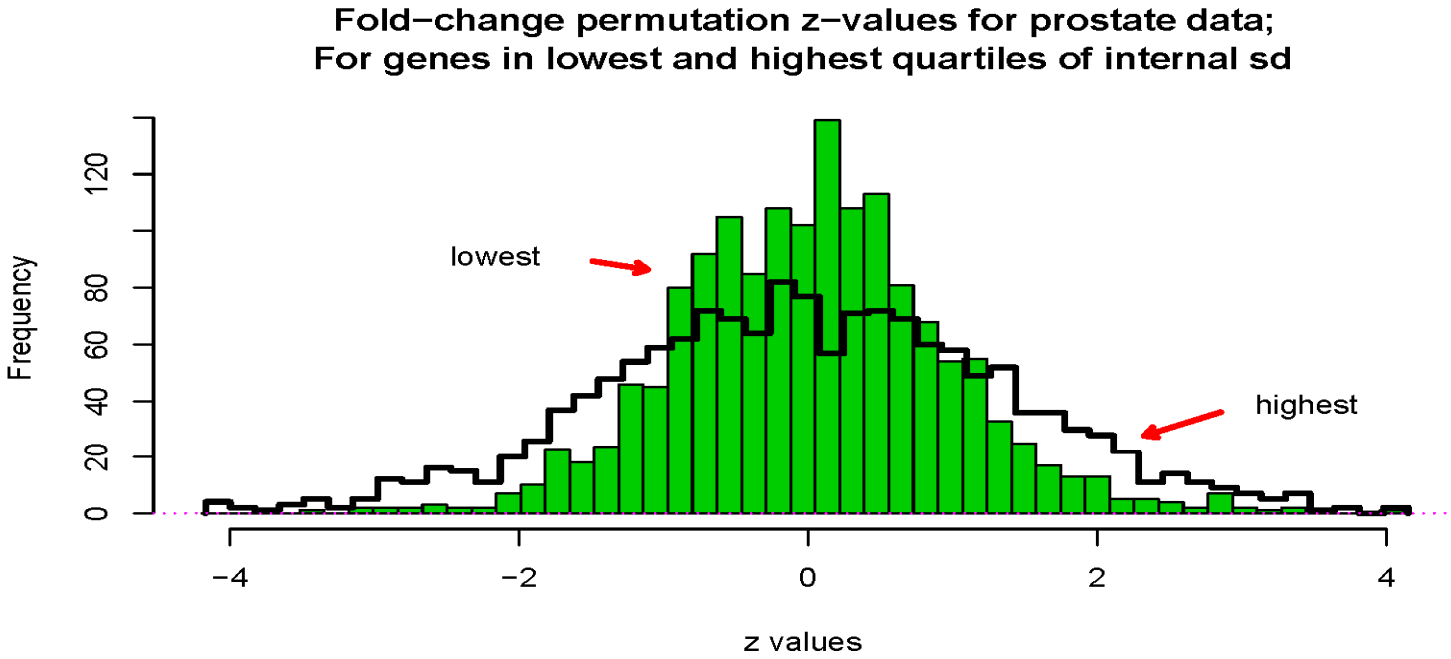
t-stat $(\bar{x}_i - \bar{y}_i)/\widehat{\text{sd}}_i \Rightarrow$ use Fold-change for Fdr, FWER etc.?

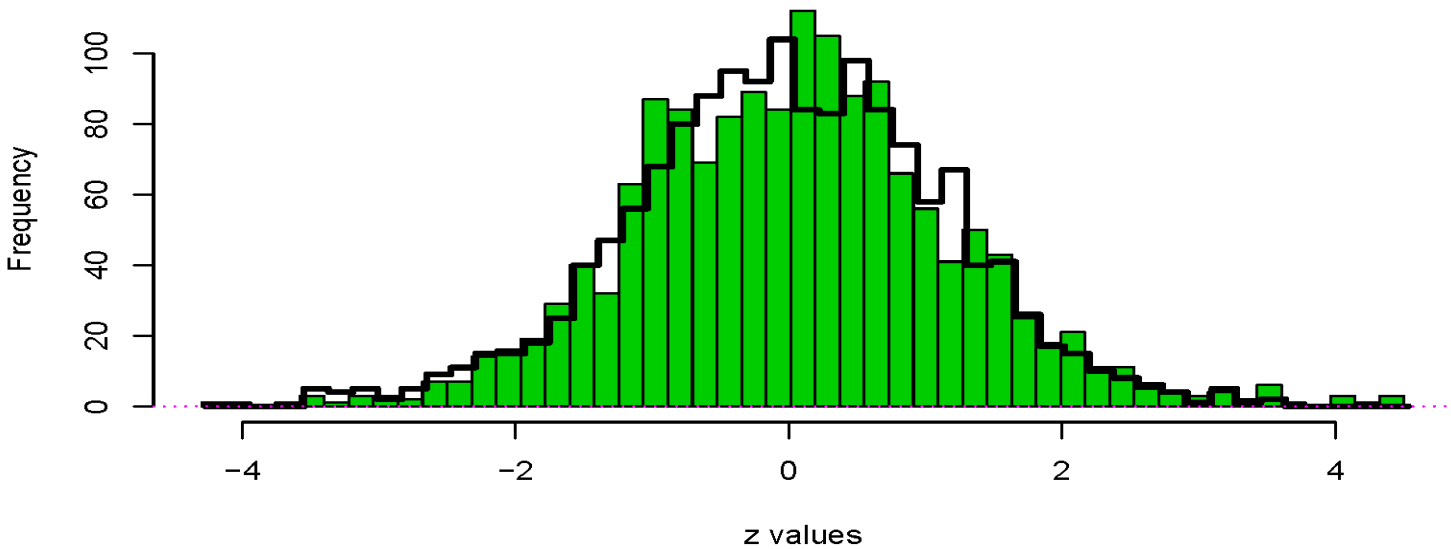
Prostate Data (Singh et al. 2002): 102 microarrays

50 controls, 52 prostate cancer, $N = 6033$ genes

For Each Gene $\widehat{\text{sd}}_i, S_i = \bar{x}_i - \bar{y}_i, T_i = (\bar{x}_i - \bar{y}_i)/\widehat{\text{sd}}_i$

(Actually use permutation z-values)


Fold−change permutation z−values for prostate data;
For genes in lowest and highest quartiles of internal sd
Frequency
0
20
40
60
80
120
lowest
highest
−4
−2
0
2
4
z values


Same comparison for permuation t−test
Frequency
0
20
40
60
80
100
−4
−2
0
2
4
z values

Conclusions

- Combining hypothesis tests requires comparable text statistics "z_i"
- z_i's should be plotted versus covariates (such as "x")
- Big covariate effects $\Rightarrow$ separate analyses (easiest with Fdr)
- Histograms differing near $z = 0 \Rightarrow$ different null densities "f_0"
- Logistic reg $\widehat{\pi}_A(z)$ convenient diagnostic

$$(\widehat{\text{fdr}}_A(z) = \widehat{\text{fdr}} \cdot \pi_A/\widehat{\pi}_A(z) \text{ for } A \text{ small.})$$

References

Schwartzman, Dougherty, and Taylor (2005) *Magnetic Resonance in Medicine* 1423-31.

Efron (2006) “Microarrays, Empirical Bayes, and the Two-groups Model” (2006)

www-stat.stanford.edu/~brad/papers/twogroups.pdf

Efron and Tibshirani (2007) “On testing the significance of sets of genes” www-stat.stanford.edu/~brad/papers/genesetpaper.pdf (to appear AOAS)

Guo et al. (2006) “Rat toxicogenomic study reveals analytical consistency across microarray platforms” *Nature Biotechnology* 1162-69.

Singh et al. (2002) “Gene expression correlates of clinical prostate cancer behavior” *Cancer Cell* 302-09.

locfdr local and tail-area fdr calculations, available in R on CRAN.