The Future of Indirect Evidence

Bradley Efron
Stanford University

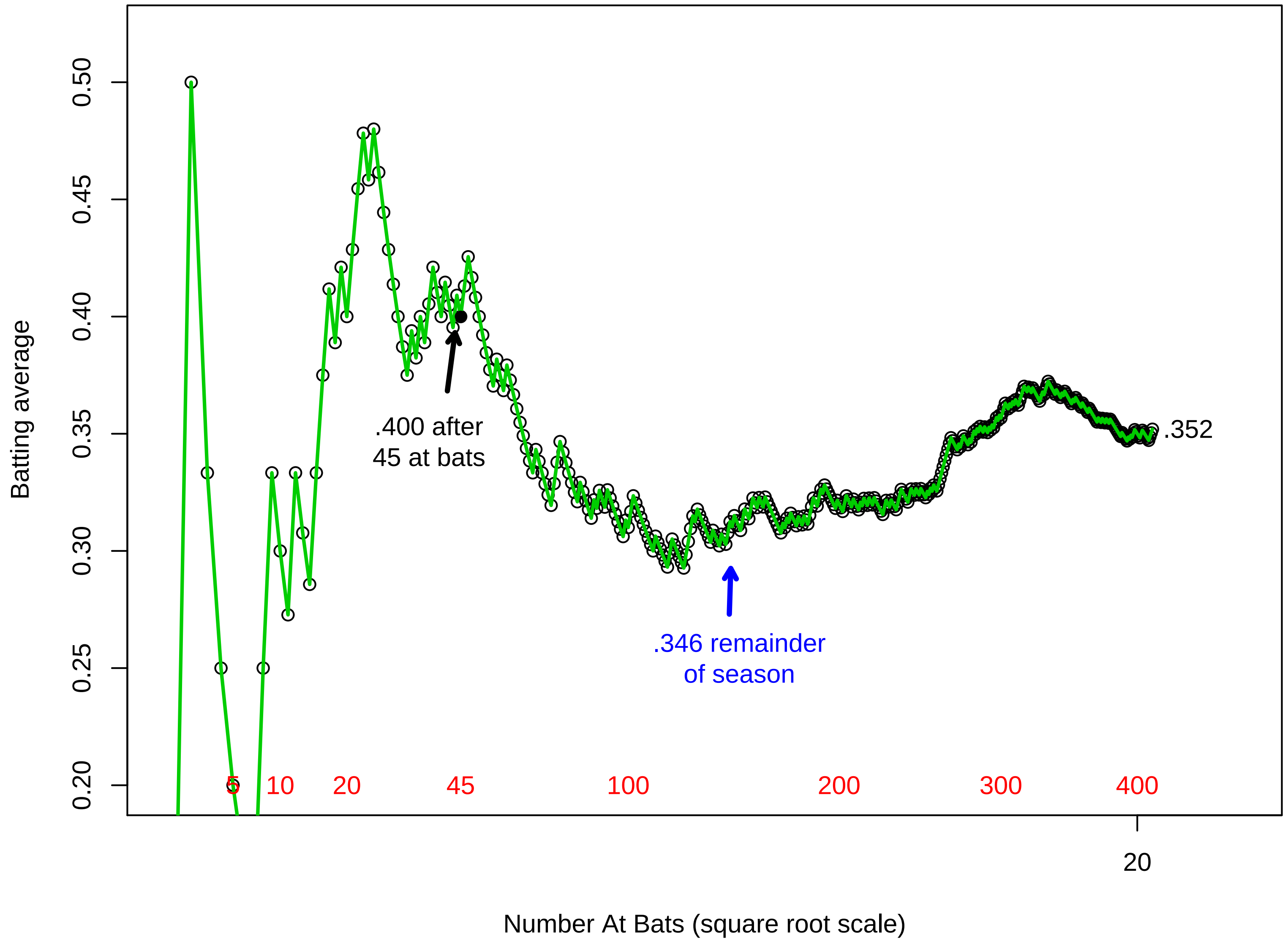
What is Statistics?

- The theory of *learning from experience* when experience arrives a little bit at a time

- Collecting and combining many small pieces of sometimes contradictory evidence

- *Direct* and *Indirect* statistical evidence

'Clemente' batting averages over 1970 season:
.400 after 45 at bats; .346 for remainder ; .352 overall



The Puzzled Physicist

- *Ultrasound*: "Twin Boys"
- *Doctor*: Proportion of twins identical $= \dfrac{1}{3}$
- *Physicist*: "Probability **my** twins identical?"

Bayes' Rule (1763)

- *Prior Odds* $\frac{\text{Prob}\{\text{Ident}\}}{\text{Prob}\{\text{Not}\}} = \frac{1/3}{2/3} = \frac{1}{2}$
- *Likelihood Ratio* $\frac{\text{Prob}\{\text{Twin Boys}|\text{Ident}\}}{\text{Prob}\{\text{Twin Boys}|\text{Not}\}} = 2$
- *Bayes' Rule*

$$\text{Posterior Odds} = (\text{Prior Odds}) \cdot (\text{Likelihood Ratio})$$

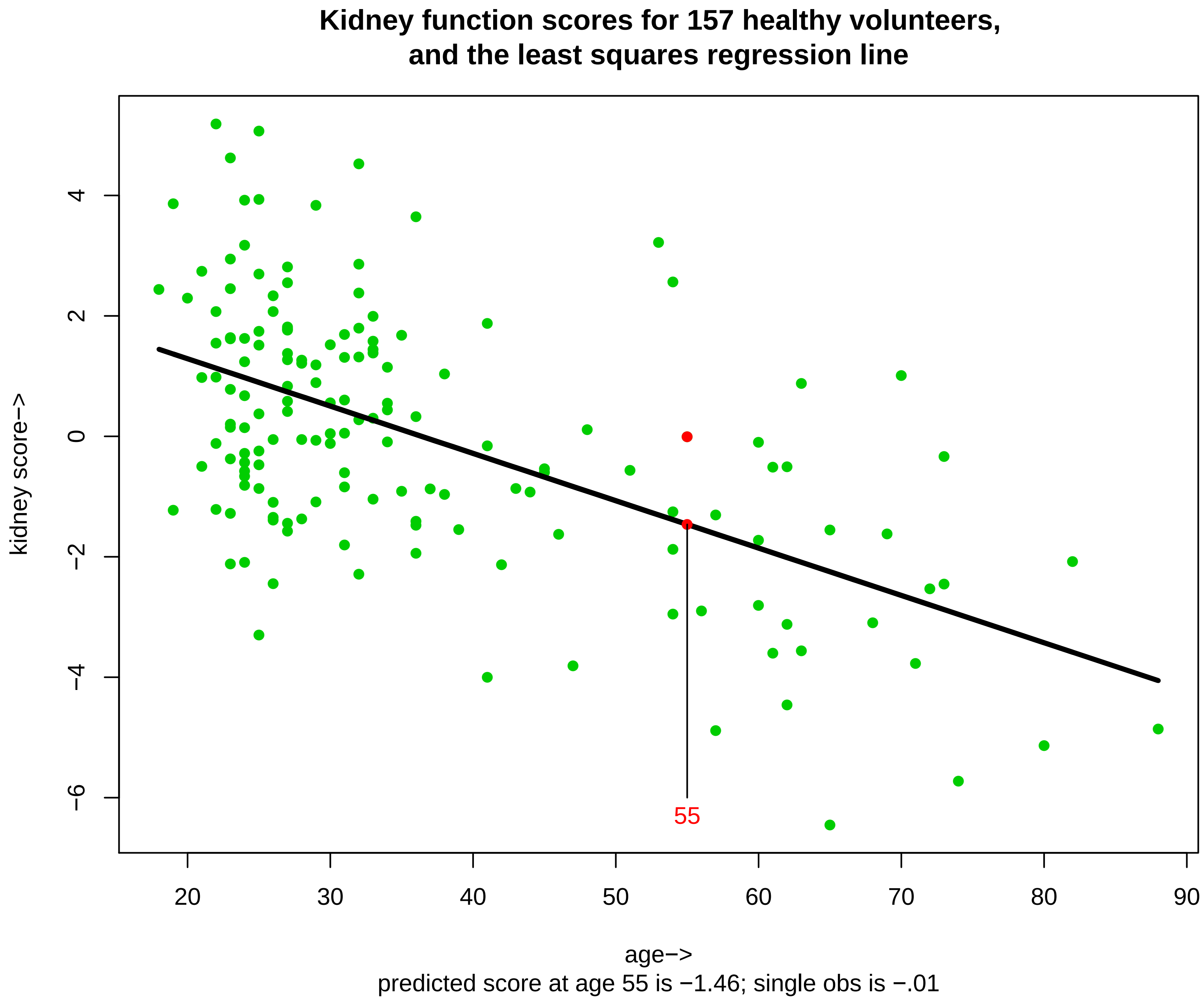
$$= \frac{1}{2} \cdot 2 = 1.$$

- *Answer to Physicist*: "50–50"
- *Crucial Ingredient*

Prior Odds: "Bayesian Prior Distribution"

Learning from Experience

- *Clemente*: Learning from his own experience [direct evidence, "frequentist"]
- *Physicist*: Learning from her own experience (sonogram) and also from the experience of others (prior distribution) [indirect evidence, "Bayesian"]
- *Holy Grail of Statisticians*: Use the experience of others without needing a (subjective) prior distribution
- *Fisher*: "Enjoy the Bayesian omelette without breaking the Bayesian eggs."



Kidney function scores for 157 healthy volunteers,
and the least squares regression line
kidney score−>
age−>
4
2
0
−2
−4
−6
20
30
40
50
60
70
80
90
55
predicted score at age 55 is −1.46; single obs is −.01

Large-Scale Regression Algorithms

- LARS, Lasso, Boosting, Bagging, CART, . . .

- *Data Mining*

- *Naïve Frequentism*

  "Chipper Jones has 3 hits in 16 tries vs. Clemens"

Eighteen Baseball Players

(Efron and Morris, 1977)

Name	hits/AB	Observed Avg	**“TRUTH”**	James–Stein
1. Clemente	18/45	.400	.346	0.290
2. F. Robinson	17/45	.378	**.298**	0.286
3. F. Howard	16/45	.356	**.276**	0.281
4. Johnstone	15/45	.333	**.222**	0.277
⋮	⋮	⋮	⋮	⋮
14. Petrocelli	10/45	.222	**.264**	0.254
15. E. Rodriguez	10/45	.222	**.226**	0.254
16. Campaneris	9/45	.200	**.286**	0.249
17. Munson	8/45	.178	.316	0.244
18. Alvis	7/45	.156	**.200**	0.239
Grand Average		.265	**.265**	0.265

James–Stein Estimation

- *Observe* $x_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, 1), \quad i = 1, 2, \ldots, n \, (\geq 4)$
- MLE Estimate $\hat{\mu}_i^{\text{MLE}} = x_i \quad E\left\{\sum_{i=1}^{n} \left(\hat{\mu}_i^{\text{MLE}} - \mu_i\right)^2\right\} = n$
- Bayes Estimate Assume $\mu_i \overset{\text{ind}}{\sim} \mathcal{N}(M, A)$

$$\hat{\mu}_i^{\text{Bayes}} = M + B(x_i - M), \quad B = \frac{A}{A+1}$$

- James–Stein "Empirical Bayes"

$$\hat{\mu}_i^{\text{JS}} = \hat{M} + \hat{B}\left(x_i - \hat{M}\right) \begin{cases} \hat{M} = \bar{x} \\ \hat{B} = 1 - (n-3) \Big/ \sum (x_i - \bar{x})^2 \end{cases}$$

Stein's Paradox (1956)

- *Theorem* $\hat{\mu}^{\text{JS}}$ **always** beats $\hat{\mu}_i^{\text{MLE}}$ in terms of total expected squared error

  (factor of 3.5 less for baseball data)

- If $\mu_i \overset{\text{ind}}{\sim} \mathcal{N}(M, A)$: (Bayes risk) (EB penalty)

$$E\left\{\sum_{i=1}^{n}\left(\hat{\mu}_i^{\text{MLE}} - \mu_i\right)^2\right\} = Bn + 3(1 - B)$$

- **Paradox** Why should Clemente's good performace raise our prediction for Munson? (Indirect evidence!)

- JS is tough on Clemente.

Large-Scale Multiple Inference

- R. G. Miller, *Simultaneous Statistical Inference* (1966):

  $N = 2$ to 10 (frequentist)

- *Microarrays* (1995+): $N = 1000$ to $10,000$

  SNP chips: $N = 500,000+$

- Indirect evidence too important to ignore

- Problems for both frequentists and Bayesians

A Microarray Example

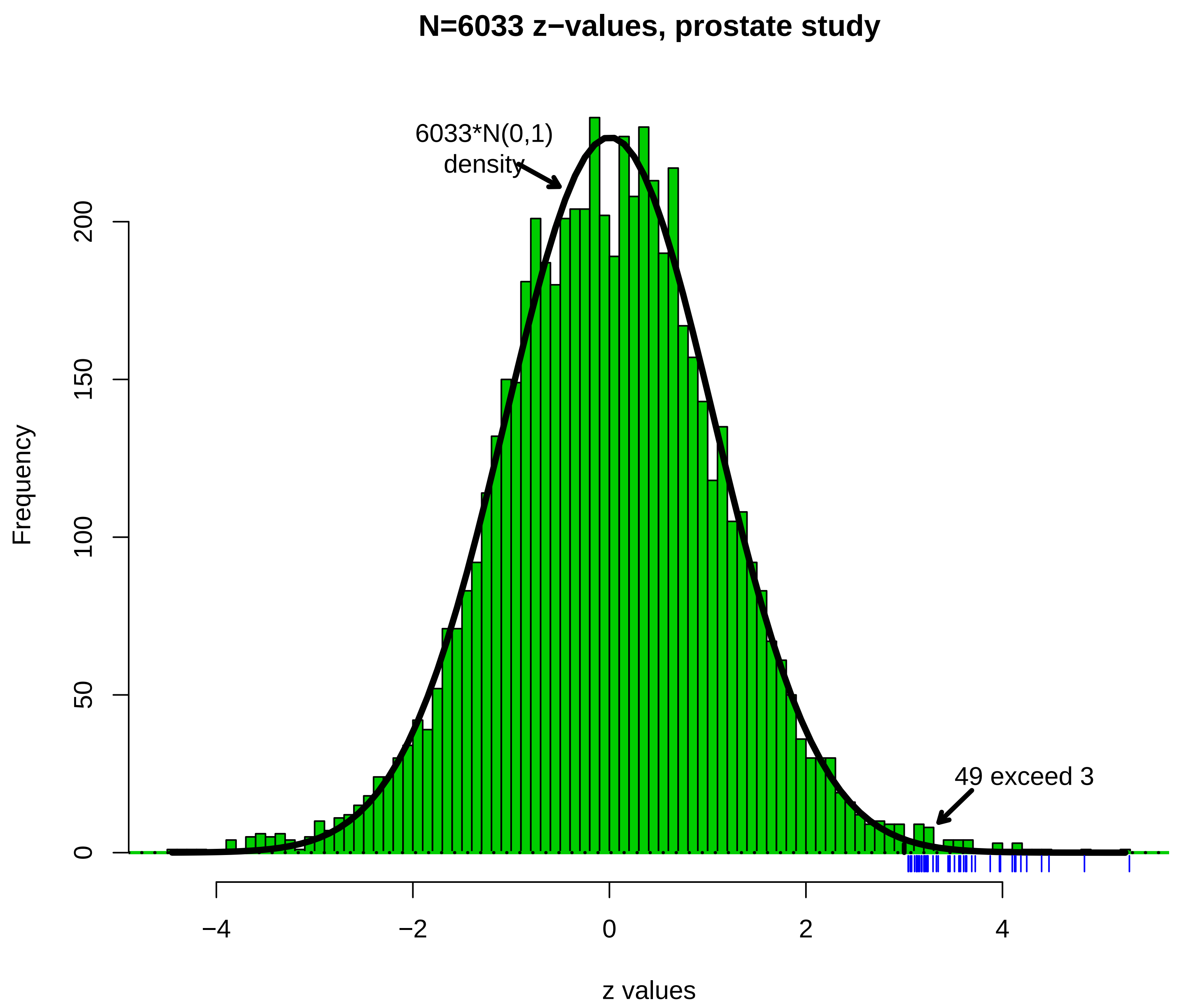
Prostate Data (Singh et al., 2002)

- $N = 6033$ genes $\begin{cases} n_1 = 50 & \text{healthy controls} \\ n_2 = 52 & \text{prostate cancer patients} \end{cases}$

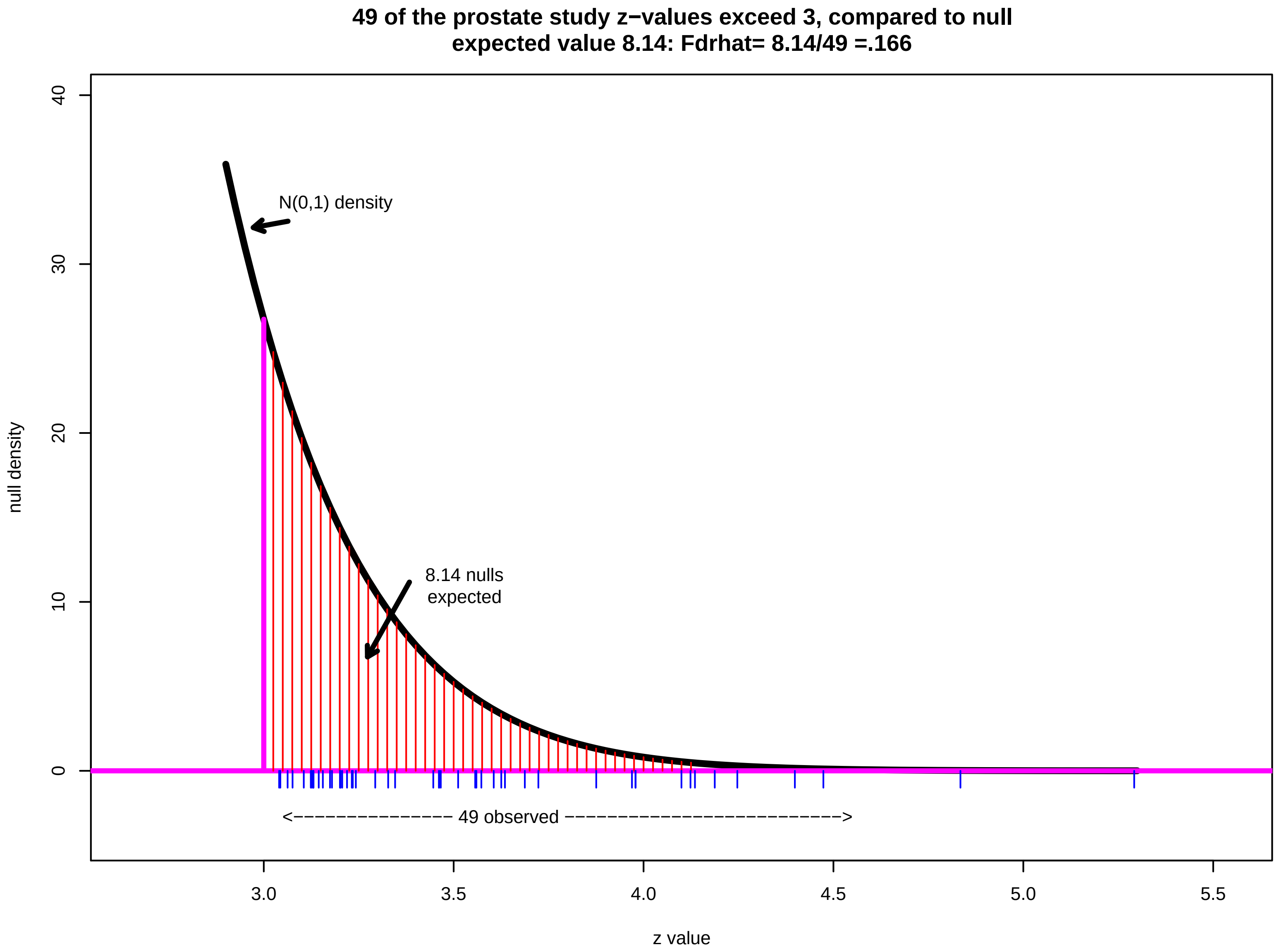
- t_i = two-sample t-stat: patients vs. controls, gene i

- $z_i = \Phi^{-1} F_{100}(t_i)$ (transform to normality)

$$\boxed{H_0 : z_i \sim \mathcal{N}(0, 1)} \qquad (\text{“theoretical null”})$$



N=6033 z−values, prostate study
6033*N(0,1)
density
49 exceed 3
Frequency
0
50
100
150
200
−4
−2
0
2
4
z values


49 of the prostate study z−values exceed 3, compared to null
expected value 8.14: Fdrhat= 8.14/49 =.166
N(0,1) density
8.14 nulls
expected
<−−−−−−−−−−−−−−− 49 observed −−−−−−−−−−−−−−−−−−−−−−−−−−>
null density
z value
0
10
20
30
40
3.0
3.5
4.0
4.5
5.0
5.5

False Discovery Rates

(Benjamini and Hochberg, 1995)

- $\widehat{\mathrm{Fdr}}(z) = \dfrac{E_0(z)}{N(z)} \begin{cases} E_0(z) = \text{expected number } z_i\text{'s} > z \text{ under } H_0 \\ N(z) = \text{observed number } z_i\text{'s} > z \end{cases}$

- $\widehat{\mathrm{Fdr}}(z) = \dfrac{8.14}{49} = \dfrac{1}{6}$

- *Theorem* For a chosen "control value" q let $z_q = \min_z \left\{\widehat{\mathrm{Fdr}}(z) \leq q\right\}$. Then the expected proportion of nulls among the $N(z_q)$ is $\leq q$. [frequentist]

- Prostate Data $q = 1/6 : z_q = 3,\ N(z_q) = 49$

Indirect Evidence and Fdr's

- The "significance" of a given $z_i > 3$ depends on how many other z_i's exceed 3.

- *Exchangeability*: If $\widehat{\text{Fdr}}(z) = 1/6$ then we report that all 49 have probability $1/6$ of being null.

- *Local Fdr*: Instead of $\mathcal{Z} = (3, \infty)$ consider $\mathcal{Z} = (3.0, 3.1)$.

Empirical Bayes Interpretation

- *Bayes*: Prior $\begin{cases} p_0 \text{ null} & z \sim f_0(z) \;\; (= \mathcal{N}(0,1)) \\ p_1 = 1 - p_0 \text{ non-null} & z \sim f_1(z) \end{cases}$

- Right cdf's $F_0(z),\; F_1(z): \; F(z) = p_0 F_0(z) + p_1 F_1(z)$

$$\boxed{\text{Fdr}(z) \equiv \text{Prob}\{\text{gene}_i \text{ null} \,|\, z_i > z\} = p_0 F_0(z)/F(z)}$$

- Empirical Bayes

$$\widehat{\text{Fdr}}(z) = p_0 F_0(z)/\hat{F}(z) \qquad \text{where } \hat{F}(z) = \#\{z_i > z\}/N$$

$$= N_0(z)/N(z) \qquad \text{as before}$$

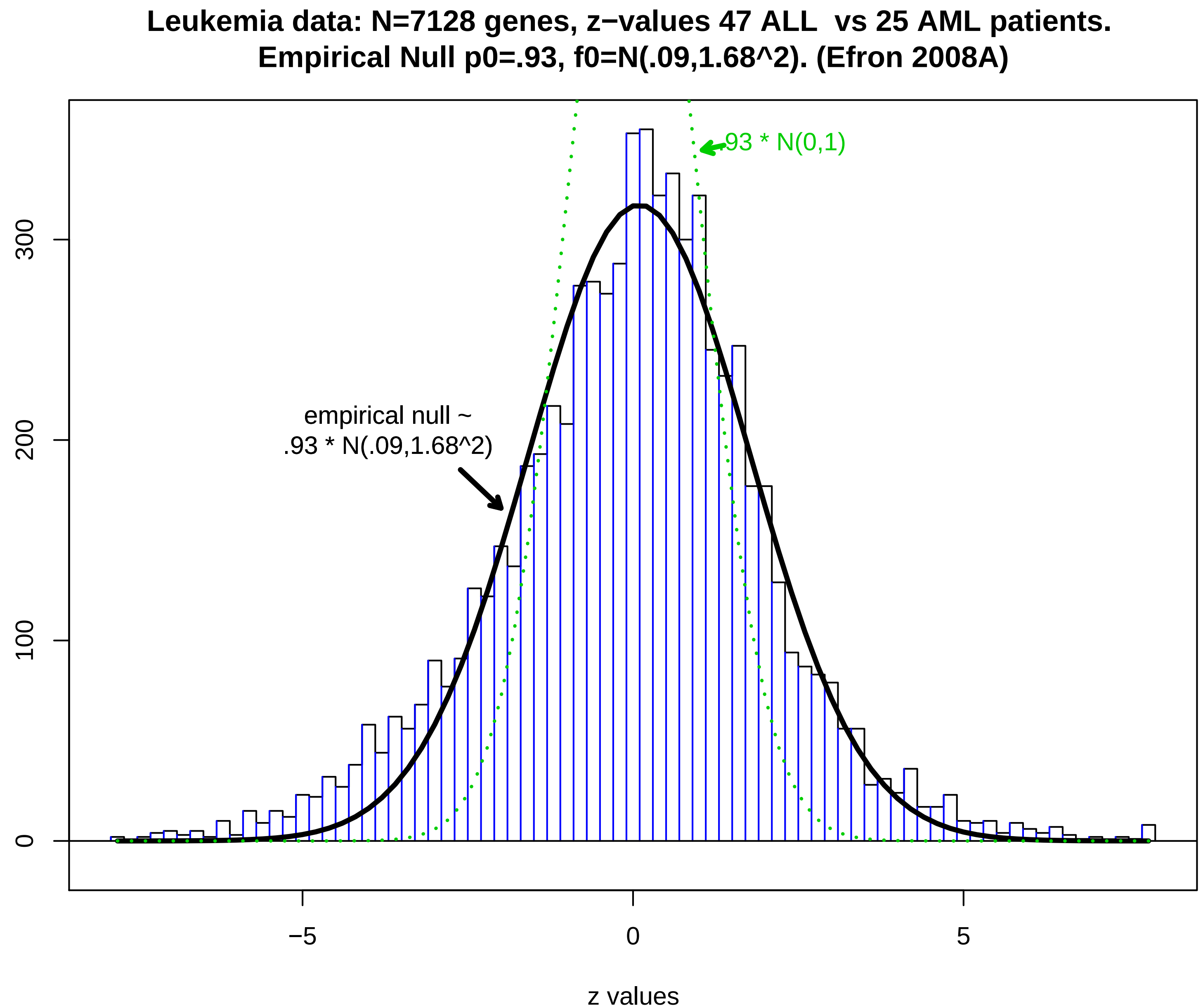
- BH Rule Reject "null" for genes with posterior Prob $\leq q$.

The Proper Use of Indirect Evidence

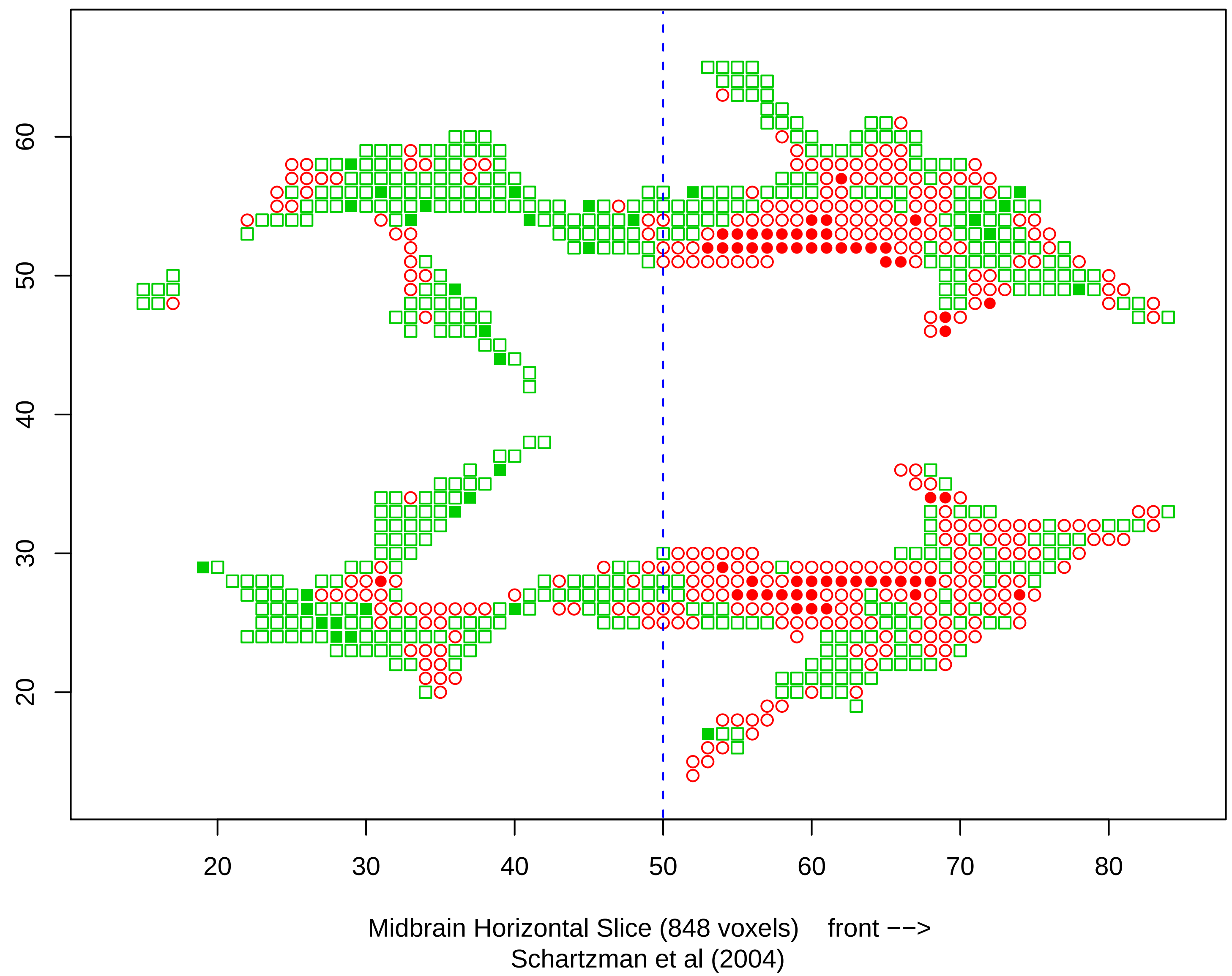
- *Difficulties and Opportunities*

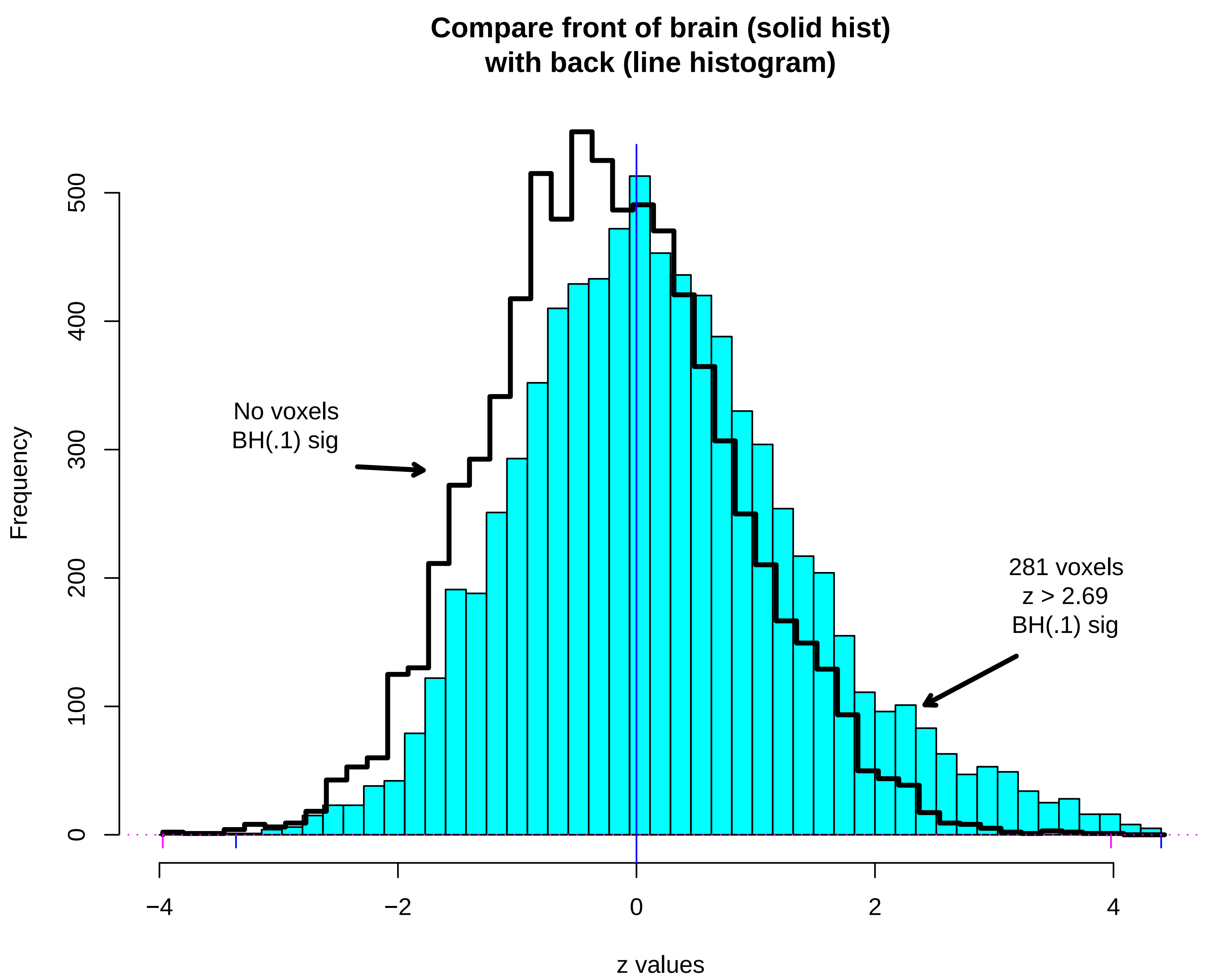
- *The Clemente Problem*: How to protect atypical cases from too much indirect evidence

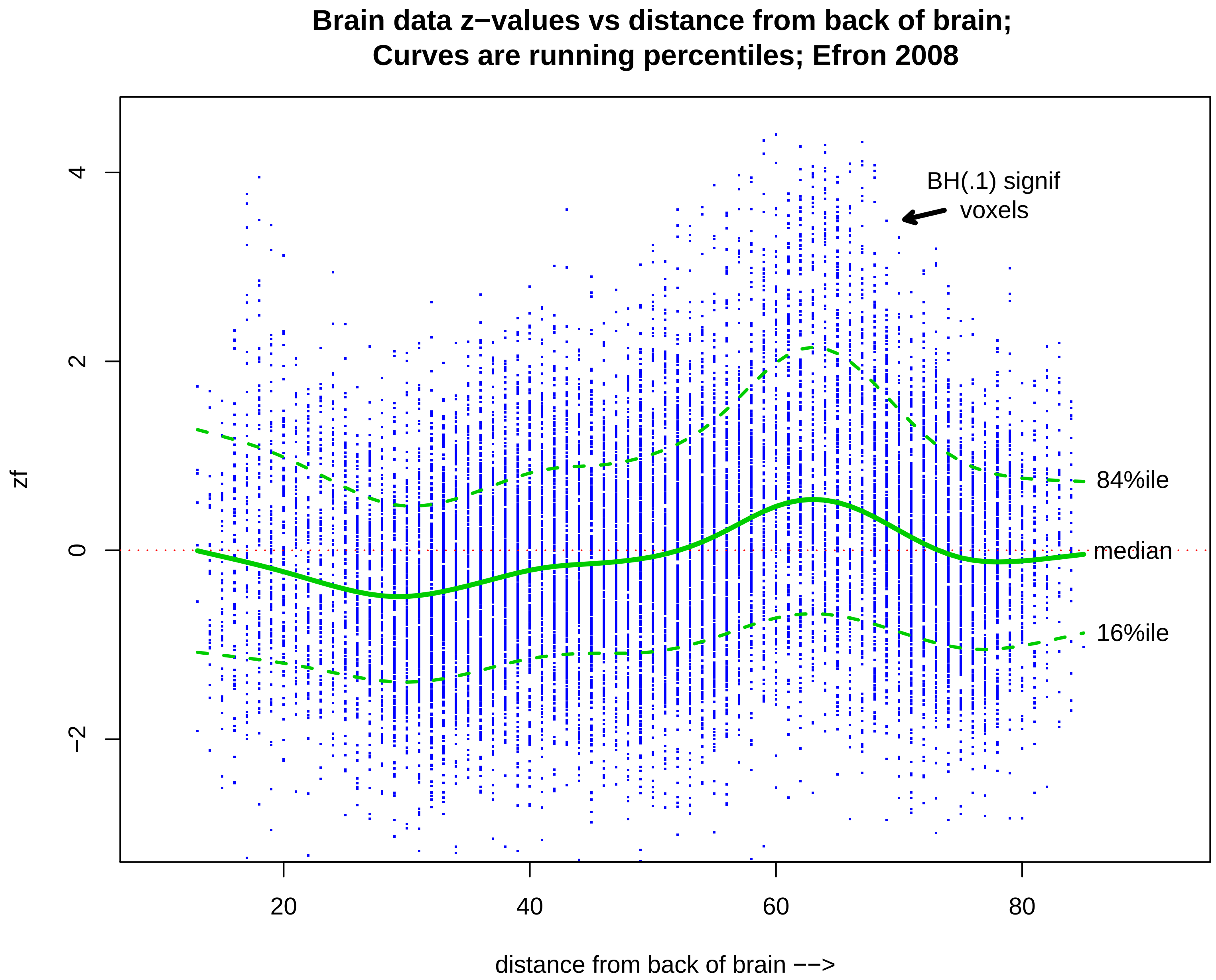
- *Fdr Theory*: With thousands of z-values, sometimes we can see, and correct, defects in the methodology.


Leukemia data: N=7128 genes, z−values 47 ALL vs 25 AML patients.
Empirical Null p0=.93, f0=N(.09,1.68^2). (Efron 2008A)
.93 * N(0,1)
empirical null ~
.93 * N(.09,1.68^2)
300
200
100
0
−5
0
5
z values

Brain data:15443 z−values comparing 6 normals vs 6 dyslexics;
Red >0, Green <0; solids show abs(z)>2

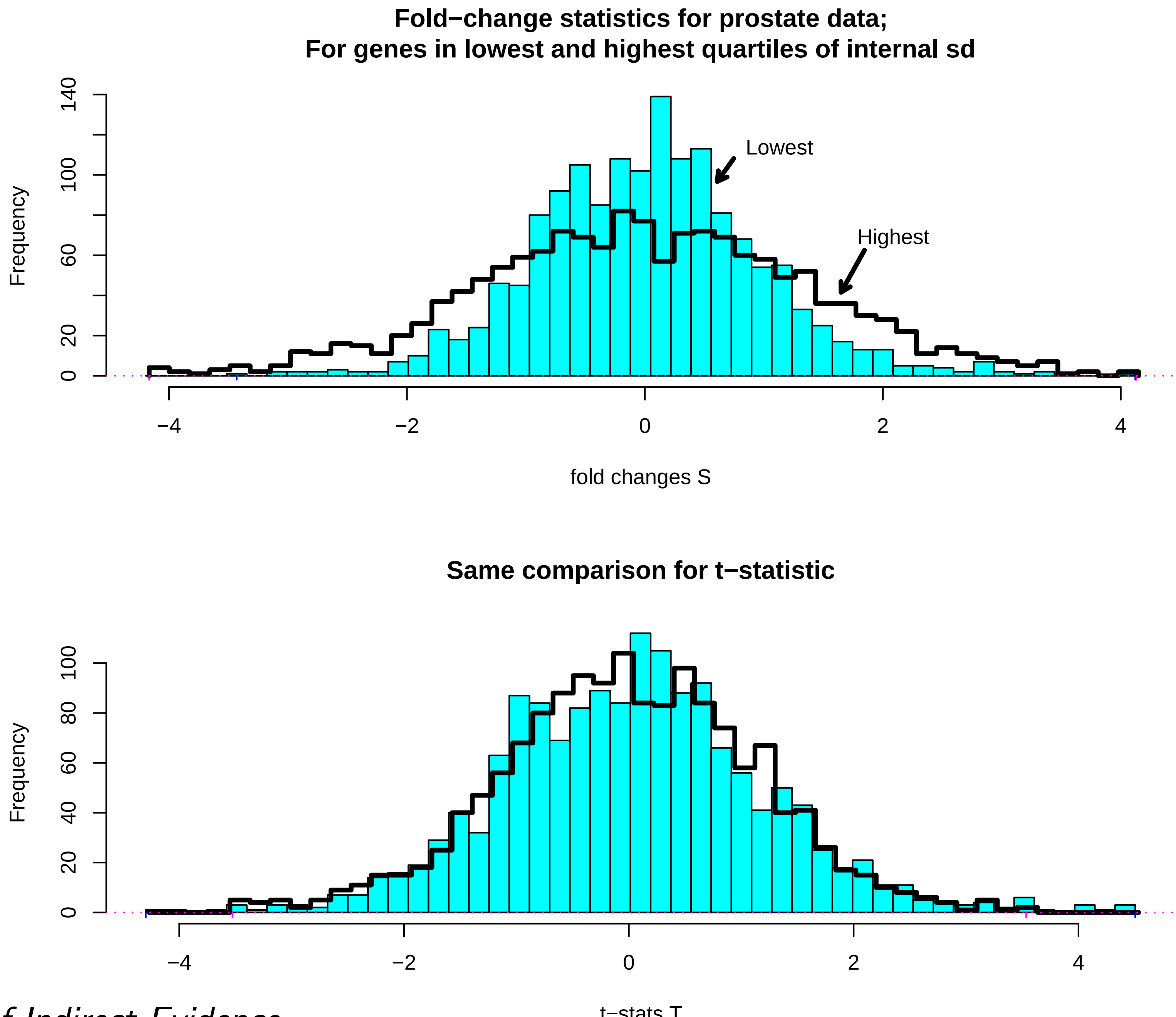



Compare front of brain (solid hist)
with back (line histogram)
No voxels
BH(.1) sig
281 voxels
z > 2.69
BH(.1) sig
Frequency
0
100
200
300
400
500
−4
−2
0
2
4
z values


Brain data z−values vs distance from back of brain;
Curves are running percentiles; Efron 2008
BH(.1) signif
voxels
84%ile
median
16%ile
zf
4
2
0
−2
20
40
60
80
distance from back of brain −−>

Fold-Change or t-Statistics?

- Two-sample Problems $(\bar{x}_i, \bar{y}_i, \hat{\sigma}_i)$ for each gene
- Fold Change (log scale) $S_n = \bar{y}_i - \bar{x}_i$
- t-statistic $T_i = (\bar{y}_i - \bar{x}_i) \big/ \hat{\sigma}_i$
- "Fold change more consistent"
- Use for Fdr, etc.?
- Prostate Data Compare genes in lowest, highest quartiles of $\hat{\sigma}_i$


Fold−change statistics for prostate data;
For genes in lowest and highest quartiles of internal sd
Lowest
Highest
Frequency
0
20
60
100
140
−4
−2
0
2
4
fold changes S
Same comparison for t−statistic
Frequency
0
20
40
60
80
100
−4
−2
0
2
4
t−stats T

The Normal Structure Model

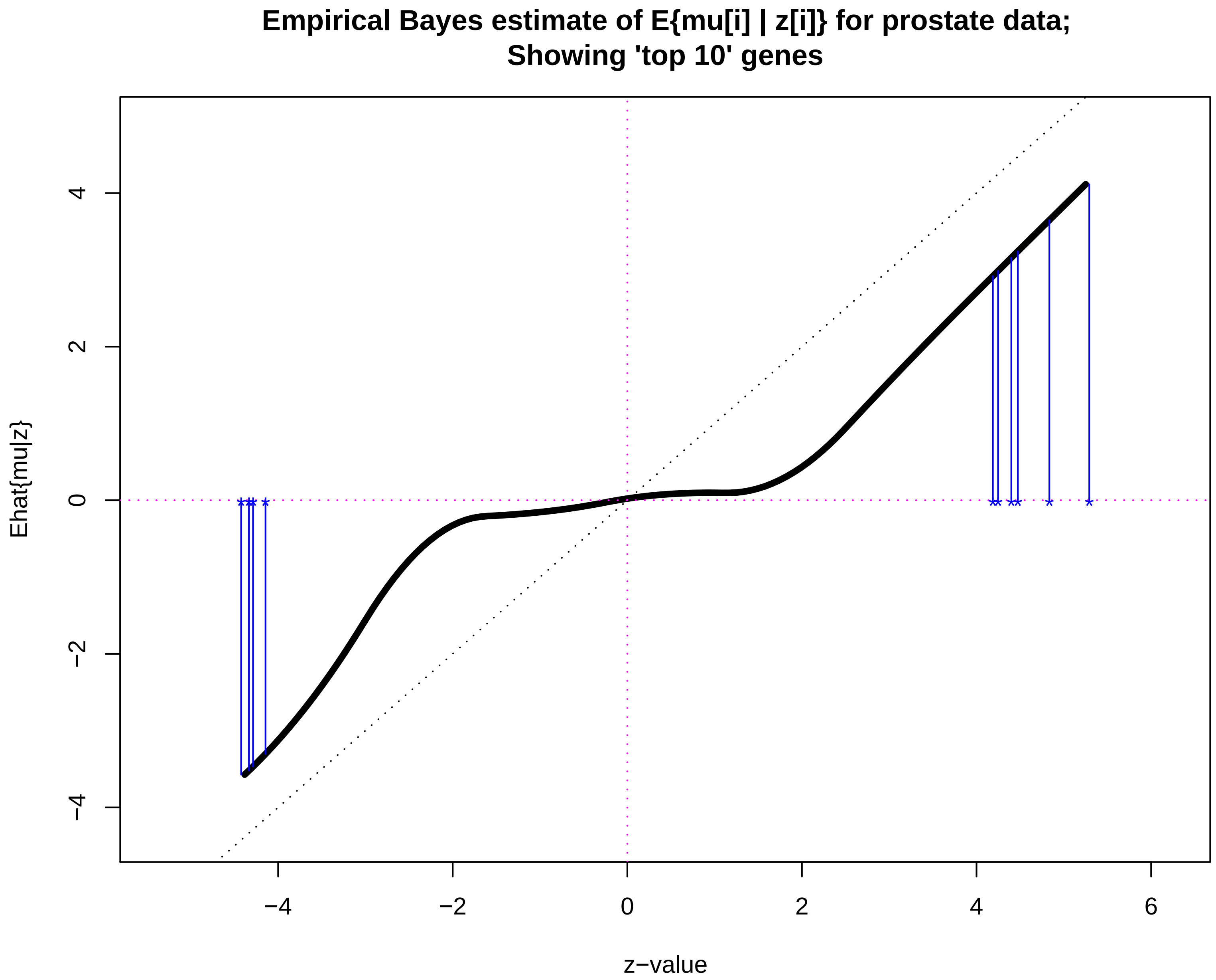
(Brown, 1971; Stein, 1981)

- *Model* $\mu \sim g(\cdot)$ and $z|\mu \sim \mathcal{N}(\mu, 1)$
- *James–Stein* $g = \mathcal{N}(M, A)$ *Fdr* $g = p_0\delta_0 + p_1 g_1$

  Theorem $\boxed{E\{\mu|z\} = z + f'(z)/f(z)}$

  where $f(z) = \int_{-\infty}^{\infty} \varphi(z-\mu)g(\mu)d\mu$ (marginal density)
- *Empirical Bayes*

$$\boldsymbol{z} \rightarrow \hat{f}(z) \rightarrow \hat{E}(\mu_i|z_i) = z_i + \hat{f}'(z_i)\Big/\hat{f}(z_i)$$



Empirical Bayes estimate of E{mu[i] | z[i]} for prostate data;
Showing 'top 10' genes
Ehat{mu|z}
z−value
−4
−2
0
2
4
6

The Top Ten Genes

	gene	z-value	$\hat{E}$
1	610	5.29	4.11
2	1720	4.83	3.65
3	332	4.47	3.24
4	364	−4.42	−3.57
5	914	4.40	3.16
6	3940	−4.33	−3.52
7	4546	−4.29	−3.47
8	1068	4.25	2.99
9	579	4.19	2.92
10	4331	−4.14	−3.30
Sum	Squares	200	116

- $\sum_1^{10} \mu_i^2$ determines prediction accuracy

Selection Bias and Bayes Estimation

- Selection Bias $E\left\{\sum_1^{10} z_i^2\right\} \gg \sum_1^{10} \mu_i^2$

- Bayes if $\tilde{\mu}_i \equiv E\{\mu_i | z_i\}$ then $E\left\{\sum_1^{10} \tilde{\mu}_i^2\right\} \approx \sum_1^{10} \mu_i^2$

  (Bayes estimates immune to selection bias)

- Empirical Bayes ?? Depends on N • Bayes: $N = \infty$

Learning from the Experience of Others

(*Which* others?)

- *Frequentist* Too conservative (stand-alone inferences)
- *Bayesian* Too bold ($N = \infty$)
- *Fisherian* (MLE) Too low-dimensional
- *Empirical Bayes* Nice compromise, but not yet a coherent theory Information? Bias?

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate .... *J. Roy. Statist. Soc. Ser. B* 57: 289–300.

Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and .... *Ann. Math. Statist.* 42: 855–903.

Efron, B. (2008a). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* 23: 1–22.

Efron, B. (2008b). Simultaneous inference: When should hypothesis testing .... *Ann. Appl. Statist.* 2: 197–223.

Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American* 236: 119–127.

Guo, L., Lobenhofer, E. K. and Wang, C. et al. (2006). Rat toxicogenomic study reveals .... *Nature Biotechnology* 24: 1162–1169.

Miller, R. G., Jr. (1981). *Simultaneous Statistical Inference*. New York: Springer-Verlag, 2nd ed.

Schwartzman, A. and Dougherty, R. et al. (2005). Cross-subject comparison of .... *Magn. Reson. Med.* 53: 1423–1431.

Singh, D., Febbo, P. G. and Ross, K. et al. (2002). Gene expression correlates of .... *Cancer Cell* 1: 203–209.

Stein, C. M. (1956). Inadmissibility of the usual estimator .... In *Proceedings of the 3rd Berkeley Symposium*. University of California Press, 197–206.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* 9: 1135–1151.