Bayesians, empirical Bayesians and Frequentists

Brad Efron

"Foundations of Statistics"

October 2007

"Why do some of the best Bayesian ideas come from Frequentists?"

— B. Efron, October 2007

- More a comparison of different attitudes than of different philosophies

- *Compromise Attitude* Empirical Bayes

Stein Estimation (1955-1961)

- Observe

$$x_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_i, 1) \qquad i = 1, 2, \ldots, N \quad (N \geq 3)$$

- Wish to estimate $\boldsymbol{\mu}$ with squared error loss

$$L(\boldsymbol{\mu}, \boldsymbol{\delta}(\boldsymbol{x})) = ||\boldsymbol{\delta}(\boldsymbol{x}) - \boldsymbol{\mu}||^2 = \sum_{i=1}^{N} (\delta_i(\boldsymbol{x}) - \mu_i)^2$$

- *MLE* $\qquad \boldsymbol{\delta}^0(\boldsymbol{x}) = \boldsymbol{x}$

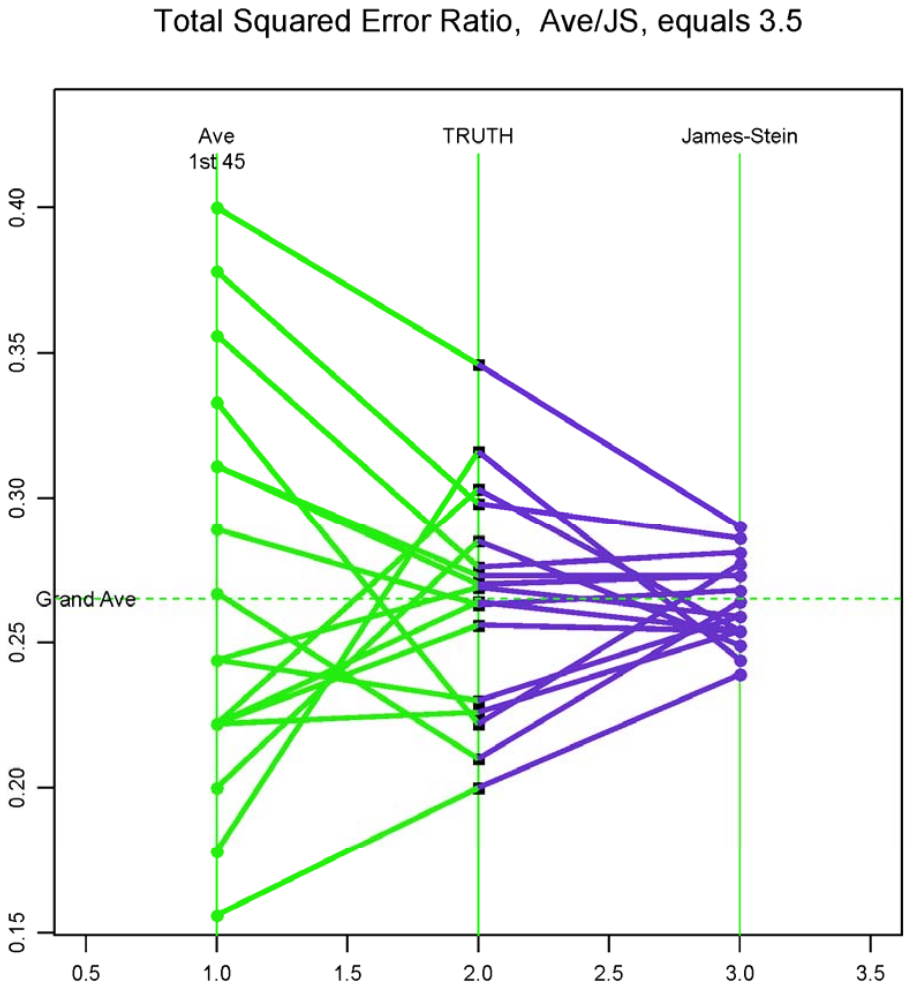
  James-Stein $\qquad \boldsymbol{\delta}^1(\boldsymbol{x}) = \left[1 - \dfrac{N-2}{\sum x_i^2}\right] \boldsymbol{x}$

- *Theorem* (James & Stein, 1961)
  For every $\boldsymbol{\mu}$

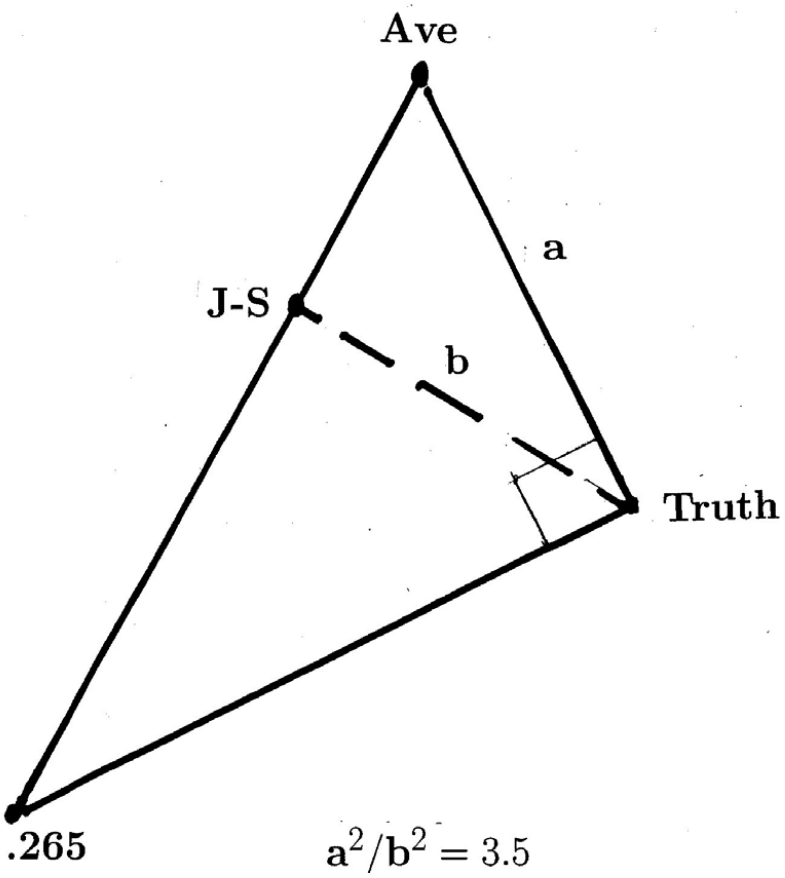
$$E_{\boldsymbol{\mu}} L(\boldsymbol{\mu}, \boldsymbol{\delta}^1(x)) < E_{\boldsymbol{\mu}} L(\boldsymbol{\mu}, \boldsymbol{\delta}^0(\boldsymbol{x}))$$

Eighteen Baseball Players

Name	hits/AB	Observed Ave	**"TRUTH"**	James-Stein
Clemente	18/45	.400	**.346**	0.290
F Robinson	17/45	.378	**.298**	0.286
F Howard	16/45	.356	**.276**	0.281
Johnstone	15/45	.333	**.222**	0.277
Berry	14/45	.311	**.273**	0.273
Spencer	14/45	.311	**.270**	0.273
Kessinger	13/45	.289	**.263**	0.268
L Alvarado	12/45	.267	**.210**	0.264
Santo	11/45	.244	**.269**	0.259
Swoboda	11/45	.244	**.230**	0.259
Unser	10/34	.222	**.264**	0.254
Williams	10/45	.222	**.256**	0.254
Scott	10/45	.222	**.303**	0.254
Petrocelli	10/45	.222	**.264**	0.254
E Rodriguez	10/45	.222	**.226**	0.254
Campaneris	9/45	.200	**.286**	0.249
Munson	8/45	.178	**.316**	0.244
Alvis	7/45	.156	**.200**	0.239
Grand Average		.265	**.265**	0.265


Total Squared Error Ratio, Ave/JS, equals 3.5
Ave
1st 45
TRUTH
James-Stein
Grand Ave
0.40
0.35
0.30
0.25
0.20
0.15
0.5
1.0
1.5
2.0
2.5
3.0
3.5

Pythagoras and Stein



Bayesian Interpretation (Lindley, 1962?)

- **Bayes Model**

$$\begin{array}{r} \boldsymbol{\mu} \sim \mathcal{N}_N(0, AI) \\ \boldsymbol{x}|\boldsymbol{\mu} \sim \mathcal{N}_N(\boldsymbol{\mu}, I) \end{array} \Rightarrow \begin{array}{c} \boldsymbol{x} \sim \mathcal{N}_N(0, (A+1)I) \\ \boldsymbol{\mu}|\boldsymbol{x} \sim \mathcal{N}_N\Big(\dfrac{A}{A+1}\boldsymbol{x}, \dfrac{A}{A+1}I\Big) \end{array}$$

- **Bayes Rule**

$$\boldsymbol{\delta}^A(\boldsymbol{x}) = \frac{A}{A+1}\boldsymbol{x} = \Big[1 - \frac{1}{A+1}\Big]\boldsymbol{x}$$

- **Empirical Bayes** Don't know A:

if $\boldsymbol{x} \sim \mathcal{N}_N(0, (A+1)I)$ then

$$E_A\left\{\frac{N-2}{\|\boldsymbol{x}\|^2}\right\} = \frac{1}{A+1}$$

$$\Rightarrow \hat{\boldsymbol{\delta}}^A(\boldsymbol{x}) = \left[1 - \frac{N-2}{\|\boldsymbol{x}\|^2}\right]\boldsymbol{x} \quad (= \text{JS!})$$

Relative Savings Loss (Efron & Morris, 1972)

- $\boldsymbol{\mu} \sim \mathcal{N}_N(0, AI) \quad \boldsymbol{x}|\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, I)$
  - $\boldsymbol{\delta}^A = \frac{A}{A+1}\boldsymbol{x} \quad \boldsymbol{\delta}^0 = \boldsymbol{x} \quad \boldsymbol{\delta}^1 = \text{JS}$
- $R^{\text{Bayes}} = E_A\|\boldsymbol{\delta}^A - \boldsymbol{\mu}\|^2 = N\frac{A}{A+1}$
  - $R^0 = N$
  - $R^0 - R^A = \frac{N}{A+1}$
- **Theorem** $\frac{R^1 - R^A}{R^0 - R^A} = \frac{2}{N}$
- "JS loses $2/N$ possible savings of Bayes Rule."
- *Question* Which estimation problems should be combined?

Example of Stein Estimation

- $N = 10$ independent cases, with

$$x[i] \sim \mathcal{N}(\mu[i], 1) \qquad i = 1, 2, 3, \ldots, 10$$

- 1000 simulations: table shows average squared error per trial for each case, and also the total.
- Stein is better overall but ...

μ	MSEmle	MSEstein
-0.81	0.95	0.61
-0.39	1.04	0.62
-0.39	1.03	0.62
-0.08	0.99	0.58
0.69	1.06	0.67
0.71	0.98	0.63
1.28	0.95	0.71
1.32	1.04	0.77
1.89	1.00	0.88
4.00	1.08	2.04 !!
Total sqerr	10.12	8.13

- **Relevance Functions:** (Efron & Morris, 1972) Let individual coords opt out of the joint estimation procedure if far away from the others.

Large-Scale Hypothesis Testing (Robbins 1951-1956)

Problem

- Observe $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, 1) \quad i = 1, 2, \ldots, N$
  (Not necessarily independent)

- Simultaneously test all null hypotheses
  $H_{0i} : \mu_i = 0.$

Bonferroni

- p-value $p_i = \text{Prob}_0\{X_i \geq x_i\} = 1 - \Phi(\boldsymbol{x})$

- Reject all H_{0i} with $p_i \leq \frac{\alpha}{N}$

  $\Rightarrow \quad \text{Prob}\{\text{reject any true } H_{0i}\} \leq \alpha.$

Robbins

- Asymptotically achieve Bayes risk as if you knew true $\{\mu\}$ distribution.

Microarray Example (Efron 2006)

- *Prostate Study:* $n = 102$ men,
  50 controls and 52 prostate cancer patients;
  $N = 6033$ genes in each microarray

- $X_{N\times n} \rightarrow$ two-sample t-stats "t_i",
  $i = 1, 2, \ldots, N$

- *z-values:* $z_i = \Phi^{-1}(F_{100}(t_i))$
  [F_{100} is cdf of t_{100} dist.]

- **Theoretical Null Hypothesis**

$$z_i \sim N(0, 1)$$

- *Simple Model:*

$z_i \sim \mathcal{N}(\mu_i, 1)$ $H_{0i} : \mu_i = 0$ $[z_i \Leftrightarrow$ "x_i"$]$

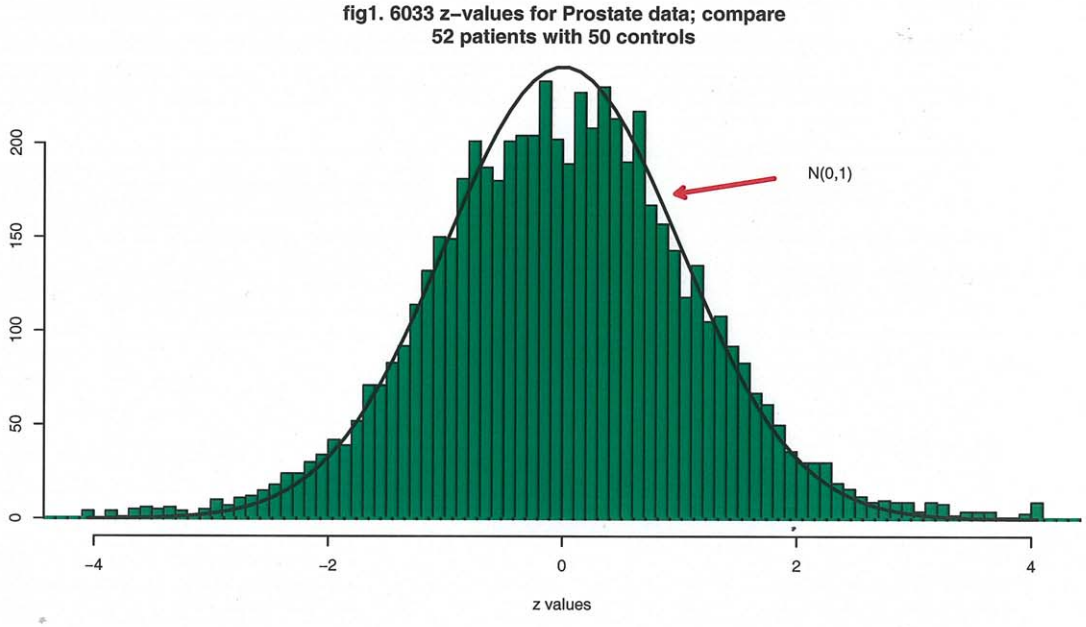
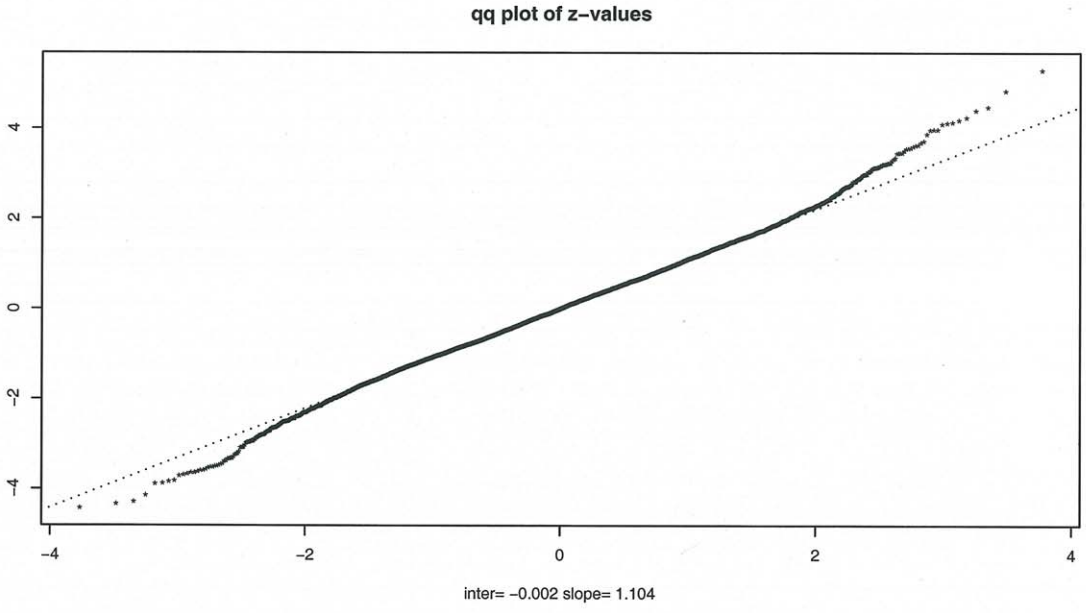


fig1. 6033 z–values for Prostate data; compare
52 patients with 50 controls
N(0,1)
0
50
100
150
200
−4
−2
0
2
4
z values


qq plot of z–values
−4
−2
0
2
4
inter= −0.002 slope= 1.104

Bayesian Two-Groups Model

- Each case (gene) either "null" or "non-null", prior probs

$p_0 = \text{Prob}\{\text{null}\}$ $f_0(z)$ density if null
$p_1 = \text{Prob}\{\text{non}-\text{null}\}$ $f_1(z)$ density if non-null

- *Simple Model:* $f_0(z) = \varphi(z)$,

$$f_1(z) = \int_{-\infty}^{\infty} \varphi(z-\mu) g_1(\mu) d\mu$$

where $g_1(\mu)$ prior density of non-zero μ's

- **Bayes Rule**

$$\text{Prob}\{\text{null}|z_i = z\} = p_0 f_0(z)/f(z)$$

where $f(z) = p_0 f_0(z) + p_0 f_1(z)$,
the *mixture density*.

False Discovery Rates (Benjamini & Hochberg, 1995)

- $F_0(z) = \text{Prob}\{Z \geq z\} = 1 - \Phi(z)$
- $\hat{F}(z) = \#\{z_i \geq z\}/N$
- $\widehat{\text{Fdr}}(z) = p_0 F_0(z)/\hat{F}(z)$
- **BH Rule** Reject all H_{0i} for $z_i \geq z_0$ where

$$z_0 = \min_z\{\widehat{\text{Fdr}}(z) \leq q\}$$

- *Theorem* Expected proportion falsely rejected nulls less than q.

	Accept	Reject	
True Nulls	a	b	N_{True}
False	c	d	
	N_{Reject}		

$\alpha = E\{b/N_{True}\}$

$\text{Fdr} = E\{b/N_{Reject}\}$

Empirical Bayes Interpretation

- *Bayesian Fdr:*

$$\mathrm{Fdr}(z) = \mathrm{Prob}\{\text{null} \,|z_i \geq z\} = p_0 F_0(z)/F(z)$$

- $\widehat{\mathrm{Fdr}}(z)$ estimates $\mathrm{Fdr}(z)$

- *BH rule:* Reject H_{0i} if estimated

$$\mathrm{Prob}\{\text{null} \,|z_i \geq z\} \text{ is } \leq q.$$

- $\widehat{\mathrm{Fdr}}(z) = p_0 F_0(z)/\hat{F}(z)$ only depends on *proportion* z_i's $\geq z$.

$\widehat{\mathrm{Fdr}}$ Example, Prostate Study

- $N(3.3) = 49$ of the genes have $z_i \geq 3.3$

- Expected number of nulls ≥ 3.3 is

  $Np_0(1 - \Phi(3.3)) = 6033 \cdot .93 \cdot .000483 = 7.6$

- $\widehat{\mathrm{Fdr}}(3.3) = 7.6/49 = .16$ so

  $\boxed{\frac{5}{6} \text{ of the 49 are non-null}}$

- **Exchangeability**

  Don't know which $\frac{1}{6}$ are false discoveries:
  assign probability 1/6 to all 49

Relevance

- *Hidden Assumption:* All N cases equally relevant to inference for any particular case.

- **Brain Study**

  $n = 12$ children, 6 dyslexic and 6 controls

- Each: DTI brain image giving response

  at $N = 15443$ voxels

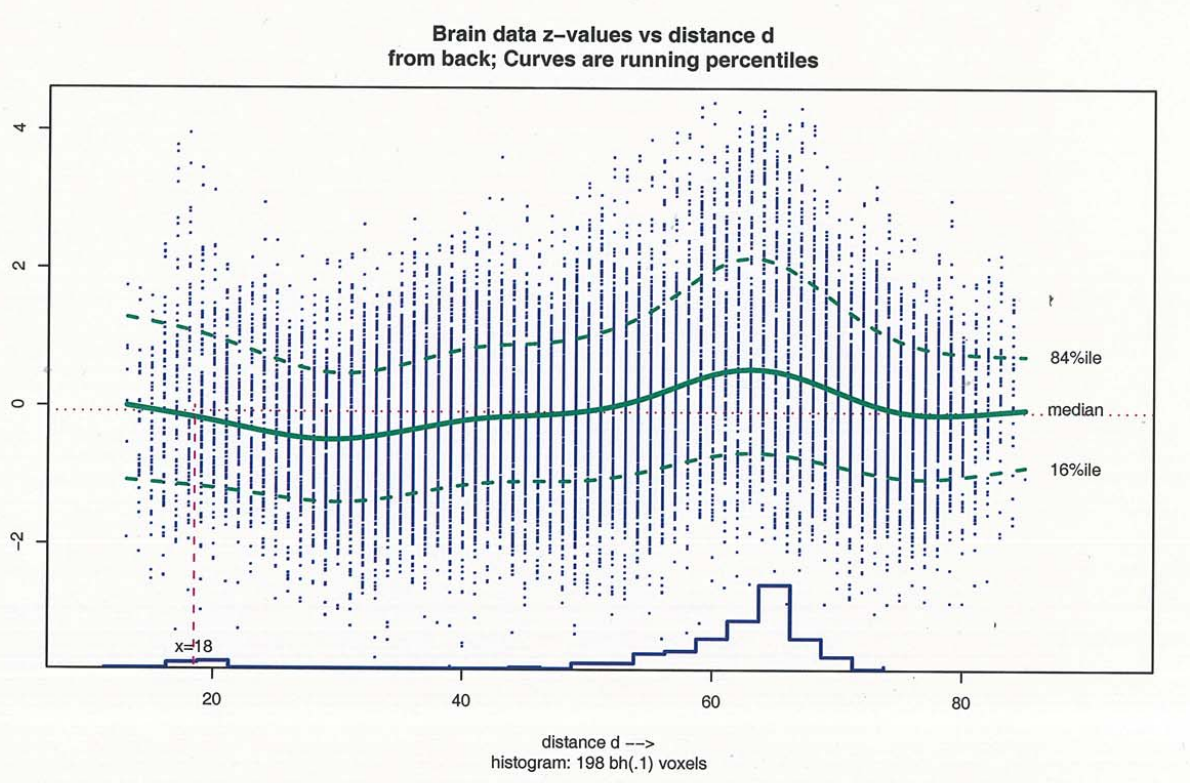
- t_i = two-sample t-stat for ith voxel

  $\Rightarrow z_i = \Phi^{-1}(F_{10}(t_i))$

  (Schwartzman, et al., 2005)

- **Next Figure** plots z_i vs d_i, distance voxel from back of brain

- Waves!


Brain data z–values vs distance d
from back; Curves are running percentiles
84%ile
median
16%ile
x=18
4
2
0
−2
20
40
60
80
distance d −−>
histogram: 198 bh(.1) voxels

Relevance Functions (Efron 2007)

- $\rho_i(d) \in [0, 1]$: relevance of voxel at d to one at d_i

- E.g., $\rho_i(d) = [1 + |d - d_i|/10]^{-1}$

- **Lemma** $\mathrm{Fdr}(z_i) = \mathrm{Fdr}(z)\frac{E_0\{\rho_i(D)|Z \geq z\}}{E\{\rho_i(D)|Z \geq z\}}$

- **Frequentists** worry a lot about individual bad situations.

Wellcome SNP Study (2007)

- " . . . we do not subscribe to the view that one should correct significance levels for the number of tests performed . . ."
- *Bayes:* $\mathrm{Prob}\{\text{null}\,|Z > z\}$
- *Empirical Bayes:* $\widehat{\mathrm{Fdr}}(z) = p_0(1-\Phi(z))/\hat{F}(z)$
- EB only depends on *proportion*

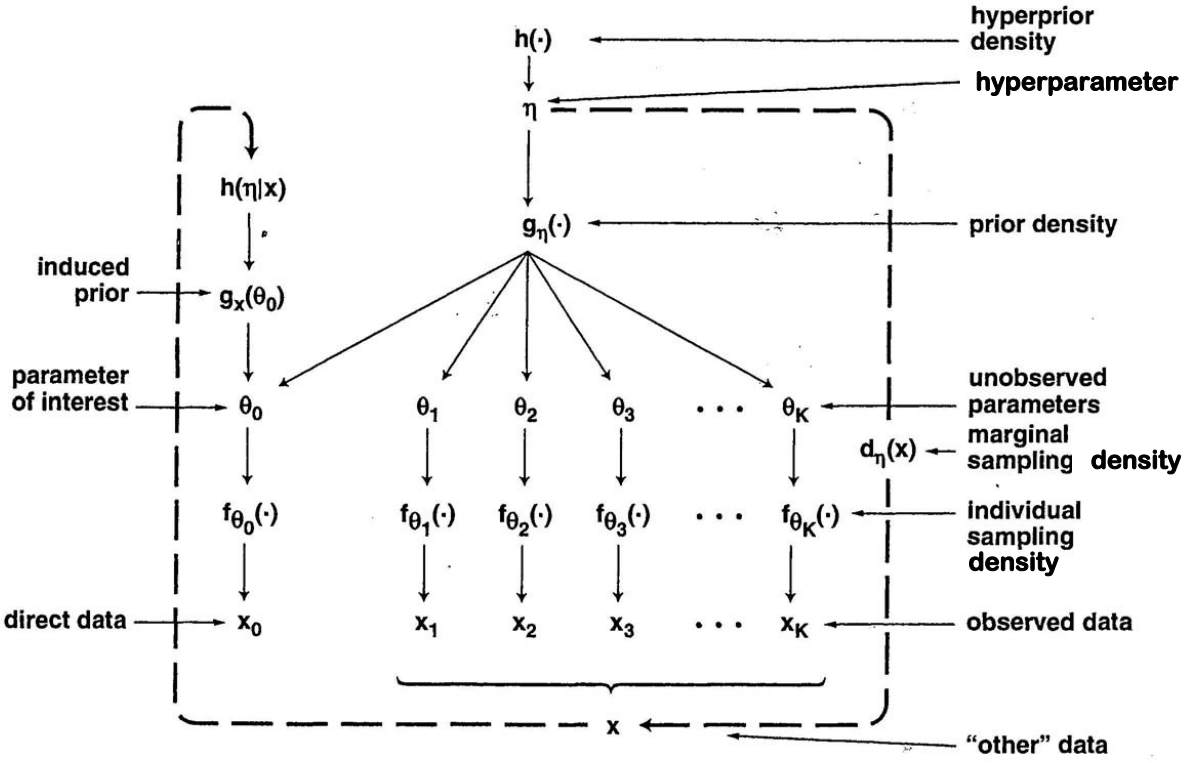
$$\hat{F}(z) = \#\{z_i \geq z\}/N$$

- Suppose $z_1 < z_2 < \cdots < z_N$:

$$\begin{aligned}\widehat{\mathrm{Fdr}}(z_N) &= p_0(1 - \Phi(z_N)/\hat{F}(z_N) \\ &= p_0 \cdot \mathrm{pval}(z_N)/(1/N)\end{aligned}$$

so BH significant if $\widehat{\mathrm{Fdr}}(z_N) \leq q$ or

$$\boxed{\mathrm{pval}(z_N) \leq \frac{q}{p_0}\,\frac{1}{N}} \qquad \text{Bonferroni!}$$



The hierarchical Bayes model hyperparameter η sampled from hyperprior density $h(\cdot)$. (*JASA*, 1996: 538–565)

Estimation compared to Hypothesis Testing (Efron 2006)

- **Common Model**

$$\mu_i \sim g(\mu) \text{ and } z_i \,|\, \mu_i \sim \mathcal{N}(0,1)$$

- *Estimation:* $g(\mu)$ smooth, e.g., $\mathcal{N}(0, A)$

- *Hypothesis Testing:* $g(\mu)$ bumpy, e.g.,

$$p_0\delta_0(\mu) + (1-p_0)\mathcal{N}(0,A)$$

References

Benjamini & Hochberg (1995). *JRSS-B* **57**, 289–300.

Efron (2007b). Microarrays, empirical Bayes, and the two-groups model. Available at `http://www-stat.stanford.edu/~brad/papers/twogroups.pdf`.

Efron (2007c). Simultaneous inference: When should hypothesis testing problems be combined? Available at: `http://www-stat.stanford.edu/~brad/papers/combinationpaper.pdf`.

Efron & Morris (1972). *JASA* **67** 130–139.

James & Stern (1961). *Proc. 4th Berkeley Symp.* **1**, 361–379.

Robbins (1956). *Proc. 3rd Berkeley Symp.* **1**, 157–163.

Schwartzman, Dougherty & Taylor (2005). *Mag. Res. Med.* **53**, 1423–1431.

Wellcome Trust (2007). *Nature* **447**, 661–678. Available at `http://www.nature.com/nature/journal/v447/n7145/abs/nature05911.html`.