Tweedie's Formula and Selection Bias

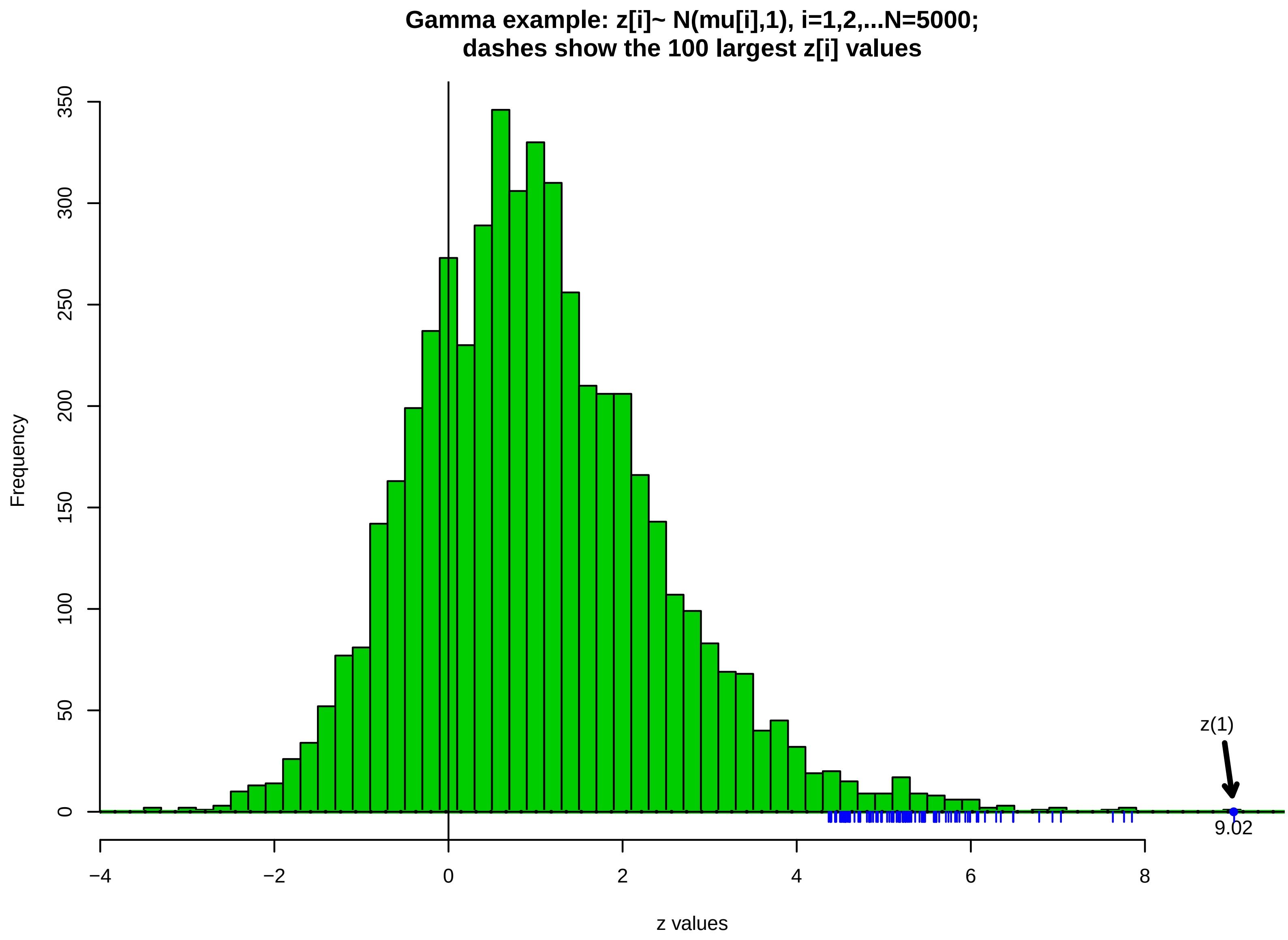
Bradley Efron
Stanford University

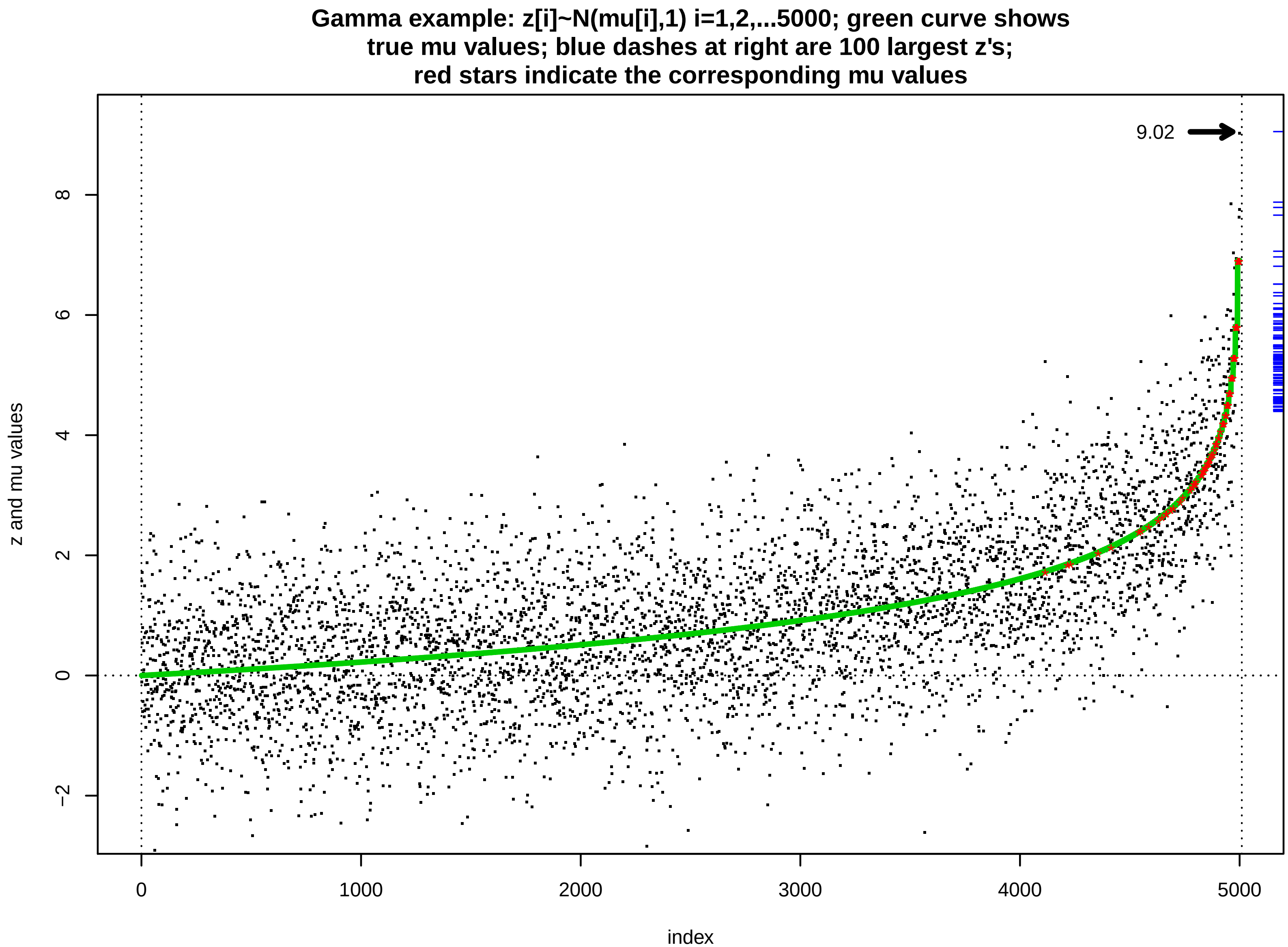
Selection Bias

- Observe $z_i \sim \mathcal{N}(\mu_i, 1)$ for $i = 1, 2, \ldots, N$
- Select the m biggest ones:

$$z_{(1)} > z_{(2)} > z_{(3)} > \cdots > z_{(m)}$$

- Question: What can we say about their corresponding μ values?
- *Selection Bias* The μ's will usually be smaller than the selected z's.


Gamma example: z[i]~ N(mu[i],1), i=1,2,...N=5000;
dashes show the 100 largest z[i] values
Frequency
0
50
100
150
200
250
300
350
z(1)
9.02
−4
−2
0
2
4
6
8
z values


Gamma example: z[i]~N(mu[i],1) i=1,2,...5000; green curve shows
true mu values; blue dashes at right are 100 largest z's;
red stars indicate the corresponding mu values
9.02
z and mu values
−2
0
2
4
6
8
0
1000
2000
3000
4000
5000
index

Tweedie's Formula

(Robbins, 1956)

- Bayes Model $\quad \mu \sim g(\cdot) \quad$ and $\quad z|\mu \sim \mathcal{N}(\mu, 1)$
- *Marginal Density* $\quad f(z) = \int_{-\infty}^{\infty} \varphi(z-\mu) g(\mu)\, d\mu \qquad \left[\varphi = \dfrac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}}\right]$
- Tweedie's Formula (Normal version)

$$\boxed{E\{\mu|z\} = z + l'(z)} \qquad \left[l'(z) = \frac{d}{dz} \log f(z)\right]$$

$$\uparrow \qquad \uparrow$$

MLE $\qquad$ Bayes correction

- Advantage $\quad$ Only need f, not g.

Higher Moments

- *Cumulant Generating Function* of μ given z:

$$\text{cgf}(z) = \frac{1}{2}z^2 + l(z) \qquad [l(z) = \log f(z)]$$

$$\text{so } \operatorname{var}\{\mu|z\} = 1 + l''(z), \quad \text{skew}\{\mu|z\} = \frac{l'''(z)}{[1 + l''(z)]^{3/2}}, \text{ etc.}$$

$$\boxed{\mu|z \sim (z + l', 1 + l'')}$$

- Bayes Risk (C.-H. Zhang, 1997)

$$\mu^\dagger = E\{\mu|z\} : E\left\{\left(\mu^\dagger - \mu\right)^2\right\} = 1 - E\left\{l'(z)^2\right\}$$

Empirical Bayes Estimates

- Use $\boldsymbol{z} = (z_1, z_2, \ldots, z_N)$ to get estimate $\hat{f}(z)$, $\hat{l}(z) = \log \hat{f}(z)$; take

$$\boxed{\mu_i | z_i \sim \Big(\underbrace{z_i + \hat{l}'_i}_{\hat{\mu}_i}, \underbrace{1 + \hat{l}''_i}_{\widehat{\mathrm{var}}_i} \Big)}$$

- Idea: Bayes estimates are immune to selection bias — maybe empirical Bayes estimates $\hat{\mu}_i$ are, too!

Maximum Likelihood Estimation of $f(z)$

- *Parametric Model*

  Suppose $l(z) = \log f(z) = \sum_{j=0}^{J} \beta_j z^j$, with MLE

$$\hat{l}(z) = \sum_{j=0}^{J} \hat{\beta}_j z^j.$$

- *Lindsey's Method*

  Histogram bins $\mathcal{Z} = \bigcup_{k=1}^{K}, \quad y_k = \#\{z_i \in \mathcal{Z}_k\}$.

  $\hat{l}$ from $\boxed{\texttt{glm}(\mathbf{y} \sim \texttt{poly}(\mathbf{x}, \texttt{J}),\ \texttt{Poisson})}$

  where $\mathbf{x}$ is vector of bin centers.

  [Slide 2: $K = 63$ bins of width 0.2]

The James–Stein Estimator

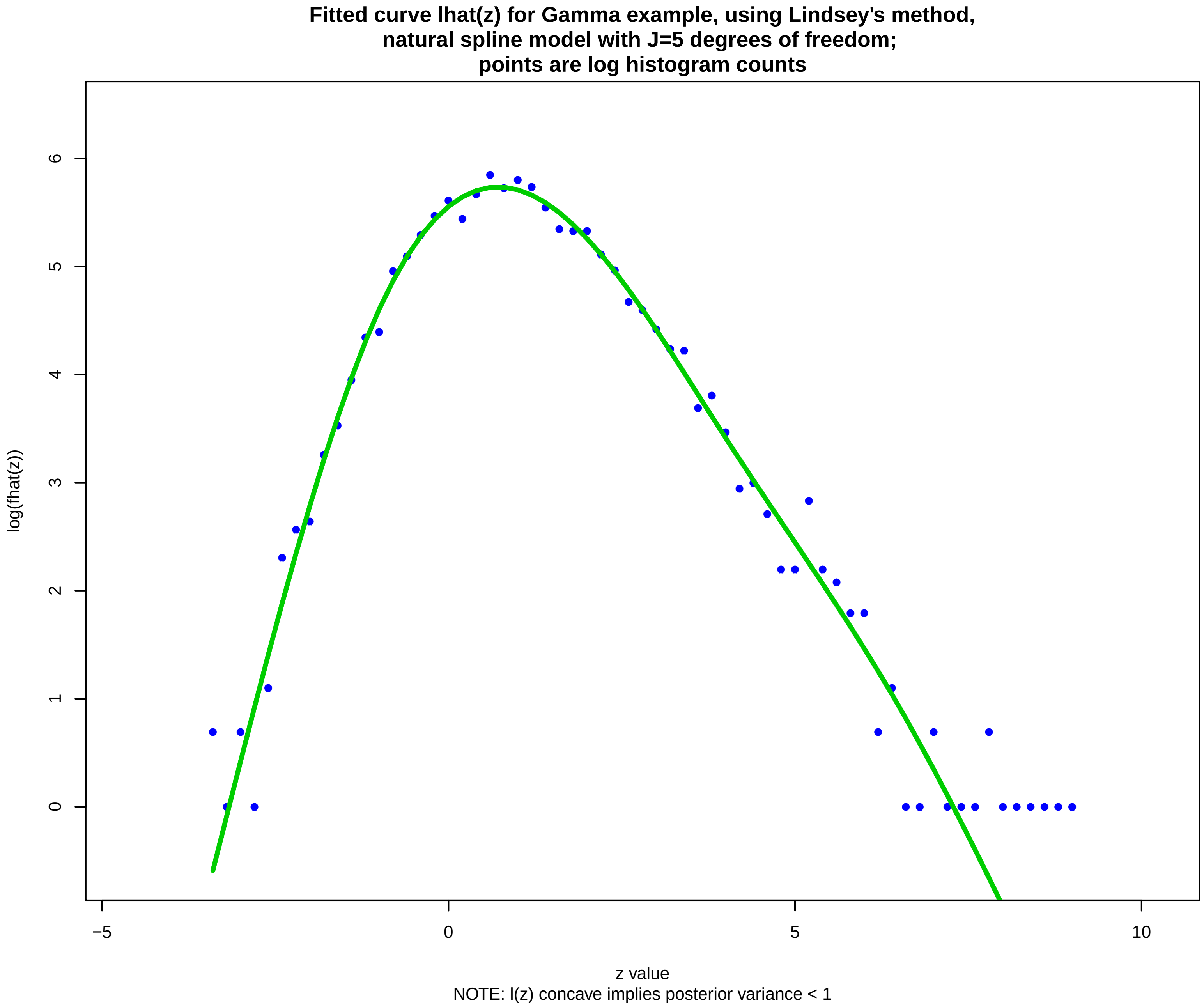
- If $\mu_i \sim \mathcal{N}(0, A)$ and $z_i | \mu_i \sim \mathcal{N}(\mu_i, 1)$,
  then $z_i \sim \mathcal{N}(0, V = A + 1)$.

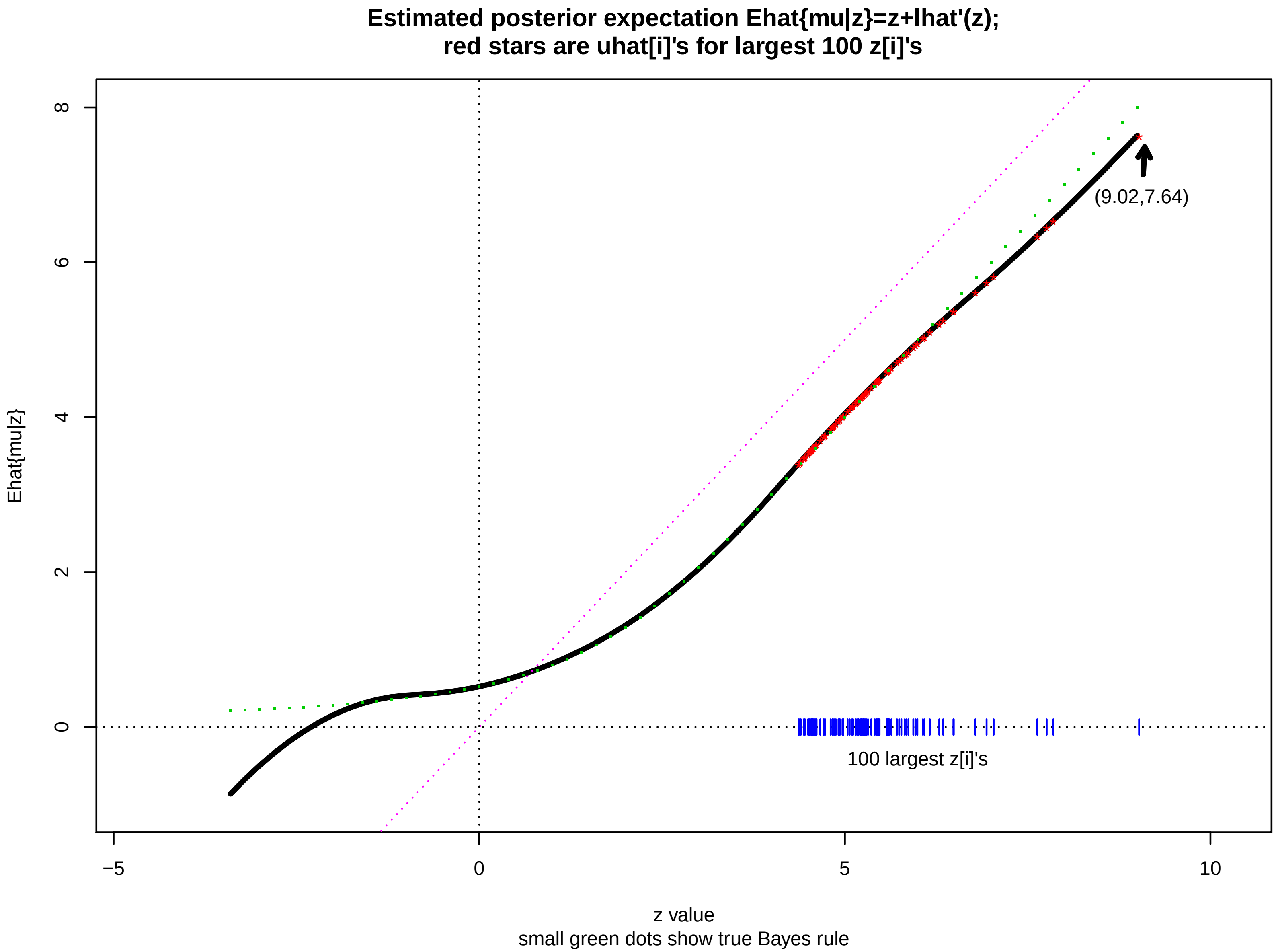
- log marginal density $l(z_i) = -z_i^2 \big/ 2V$ Quadratic ($J = 2$),

$$\boxed{E\{\mu_i | z_i\} = z_i - \frac{z_i}{V}.}$$

- James–Stein $\quad \hat{\mu}_i = z_i - \dfrac{z_i}{\hat{V}} \qquad$ where $\hat{V}^{-1} = \dfrac{(N-2)}{\|\boldsymbol{z}\|^2}$

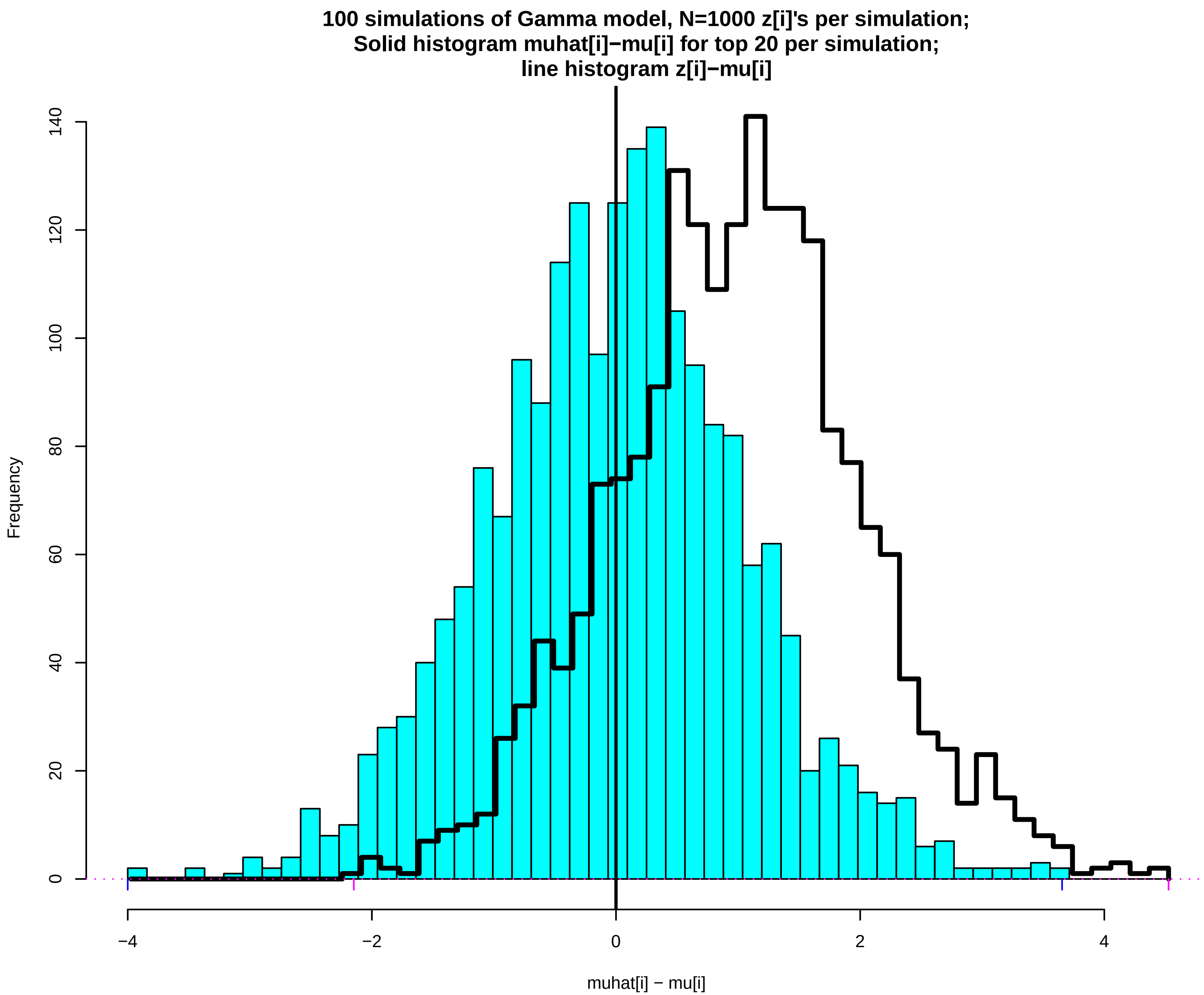
- Tweedie estimates are a generalization of James–Stein.


Fitted curve lhat(z) for Gamma example, using Lindsey's method,
natural spline model with J=5 degrees of freedom;
points are log histogram counts
log(fhat(z))
z value
NOTE: l(z) concave implies posterior variance < 1

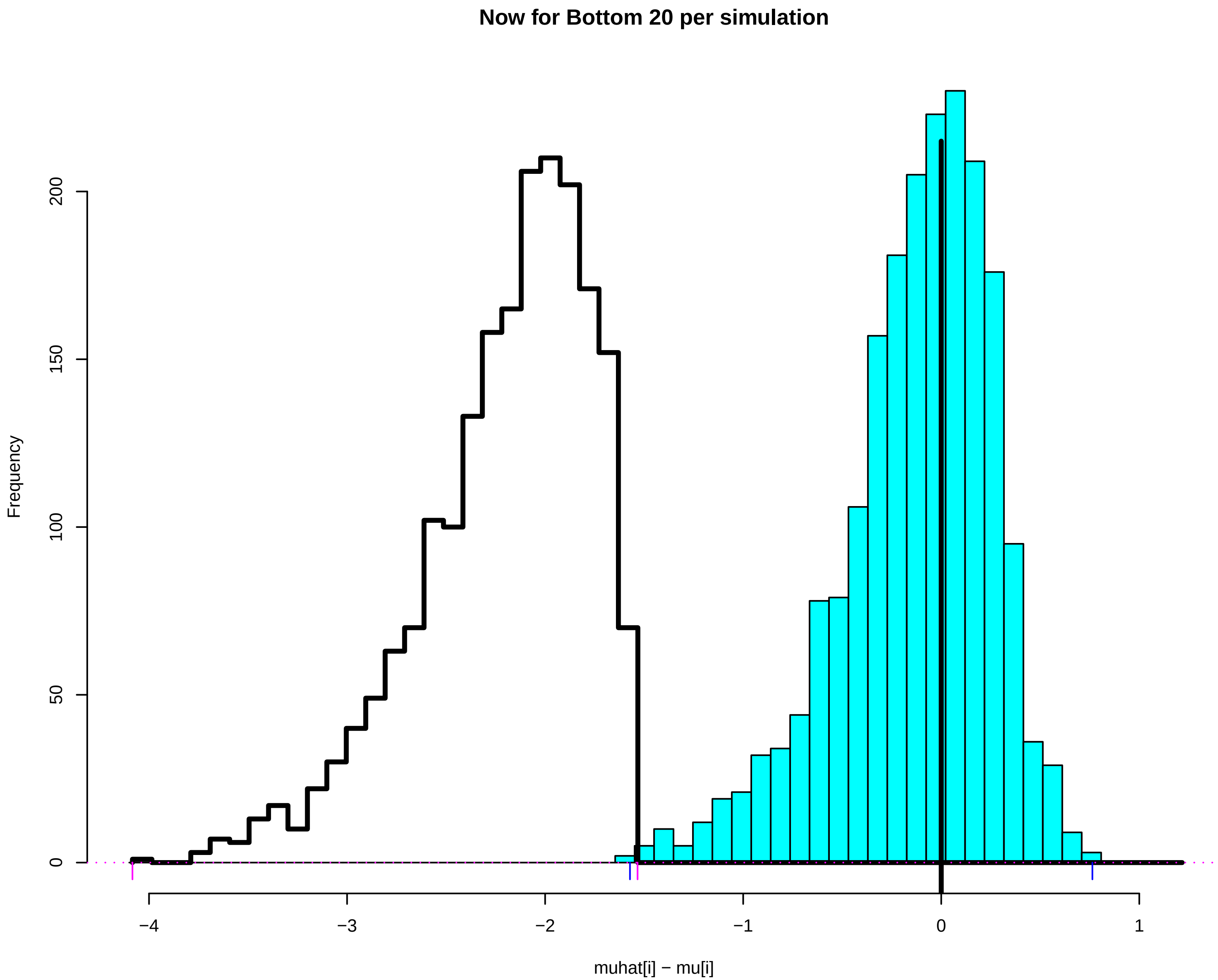

Estimated posterior expectation Ehat{mu|z}=z+lhat'(z);
red stars are uhat[i]'s for largest 100 z[i]'s
Ehat{mu|z}
8
6
4
2
0
−5
0
5
10
(9.02,7.64)
100 largest z[i]'s
z value
small green dots show true Bayes rule

Does “Tweedie” Cure Selection Bias?

- Simulations $\mu_i \sim$ Gamma, and $z_i \sim \mathcal{N}(\mu_i, 1)$ for $i = 1, 2, \ldots, N = 1000$
- 100 Simulations: Select top 20 and bottom 20 z_i's each time.
- Next histograms show $\hat{\mu}_i - \mu_i$.
- Tweedie's formula works well even for $N = 200$.



100 simulations of Gamma model, N=1000 z[i]'s per simulation;
Solid histogram muhat[i]−mu[i] for top 20 per simulation;
line histogram z[i]−mu[i]
Frequency
0
20
40
60
80
100
120
140
−4
−2
0
2
4
muhat[i] − mu[i]


Now for Bottom 20 per simulation
Frequency
0
50
100
150
200
−4
−3
−2
−1
0
1
muhat[i] − mu[i]

Bayes Regret
(Muralidharan, 2009; Zhang, 1997)

- Regret $\quad \mathrm{Reg}(z_0) \equiv E\left[\mu - \hat{\mu}_{\boldsymbol{z}}(z_0)\right]^2 - E\left[\mu - \mu^{\dagger}(z_0)\right]^2$
  with $\boldsymbol{z} = (z_i, \ldots, z_N)$ and $\mu | z_0$ random, z_0 fixed.

- $\mathrm{Reg}(z_0)$ depends on accuracy of $\hat{l}'_{\boldsymbol{z}}(z_0)$ as estimate of $l'(z_0) = \left.\dfrac{d}{dz} \log f(z)\right|_{z_0}$

$$\boxed{\mathrm{Reg}(z_0) = E\left[\hat{l}'_{\boldsymbol{z}}(z_0) - l'(z_0)\right]^2}$$

- Asymptotically $\quad \mathrm{Reg}(z_0) \approx \mathrm{var}\left\{\hat{l}'_{\boldsymbol{z}}(z_0)\right\} \doteq c(z_0)/N$

RMS Regret for Gamma Example

%ile	.9	.95	.99	.999
z-value	2.8	3.5	4.2	7.2
$N = 250$	.18	.17	.37	6.4
$N = 500$	.11	.10	.20	3.5
$N = 1000$	.09	.08	.15	1.5
$c(z_0)/1000$	.08	.07	.13	.6

Empirical Bayes Information

- As N increases $\hat{\mu}_i$ goes from MLE z_i ($N = 1$) to Bayes estimate $\mu_i^\dagger$ ($N = \infty$)
- Posterior Variability

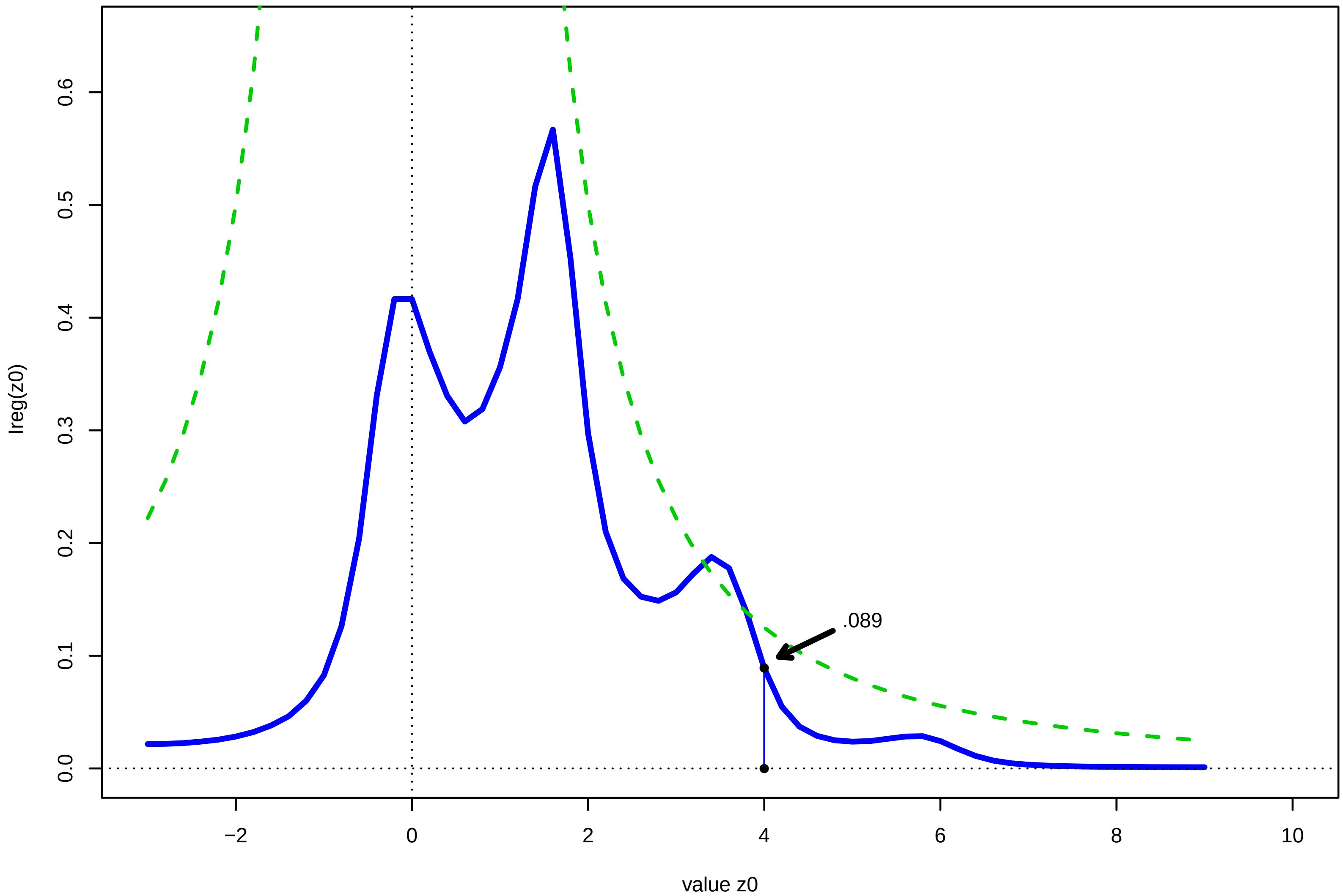
$$E\left[\mu_i - \hat{\mu}_i\right]^2 = \mathrm{var}_i^\dagger + \mathrm{Reg}(z_i) \qquad \left(\mathrm{var}_i^\dagger \approx 1\right)$$

- $\hat{\mu}_i$ undependable for most extreme few z_i's
- *Empirical Bayes Information* (per "other" observation):

$$\boxed{\mathcal{I}(z_0) = \lim_{N \to \infty} \frac{1}{(N \cdot \mathrm{Reg}(z_0))} = \frac{1}{c(z_0)}}$$

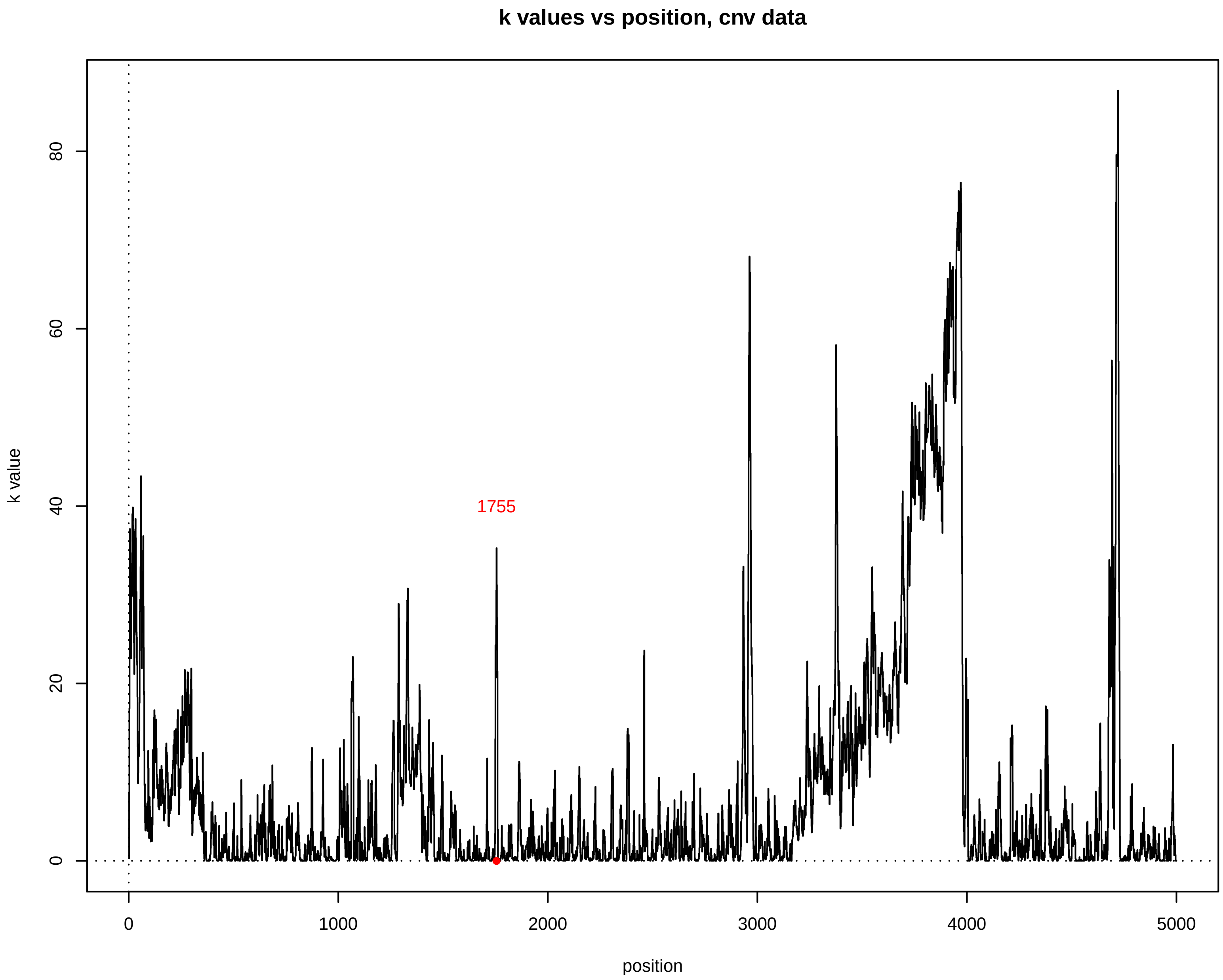
$$\left[\mathrm{Reg}(z_0) \approx \frac{1}{N\mathcal{I}(z_0)}\right]$$

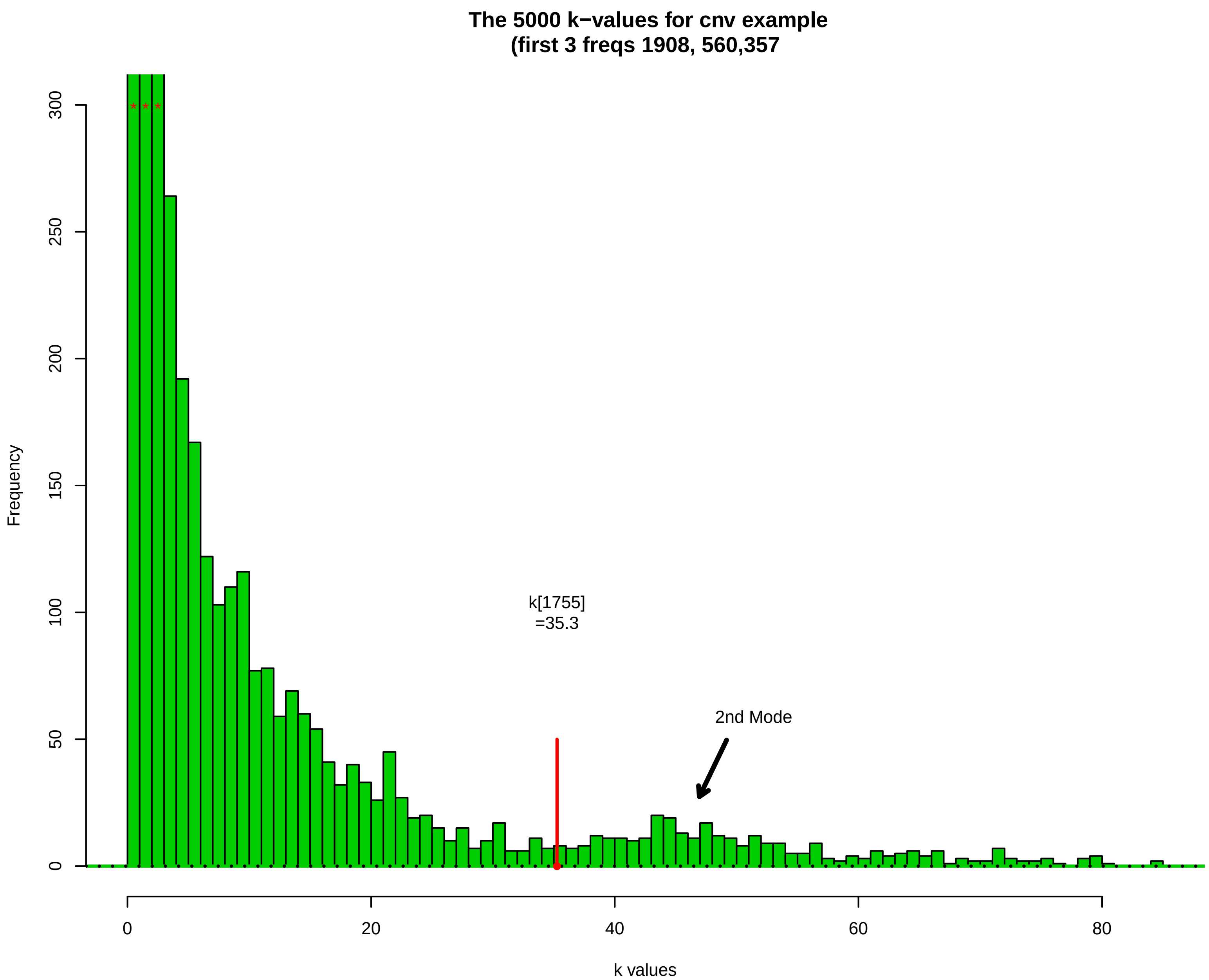
Empirical Bayes information per observation Ireg(z0) for Gamma model. (Dashed curve for James−Stein)

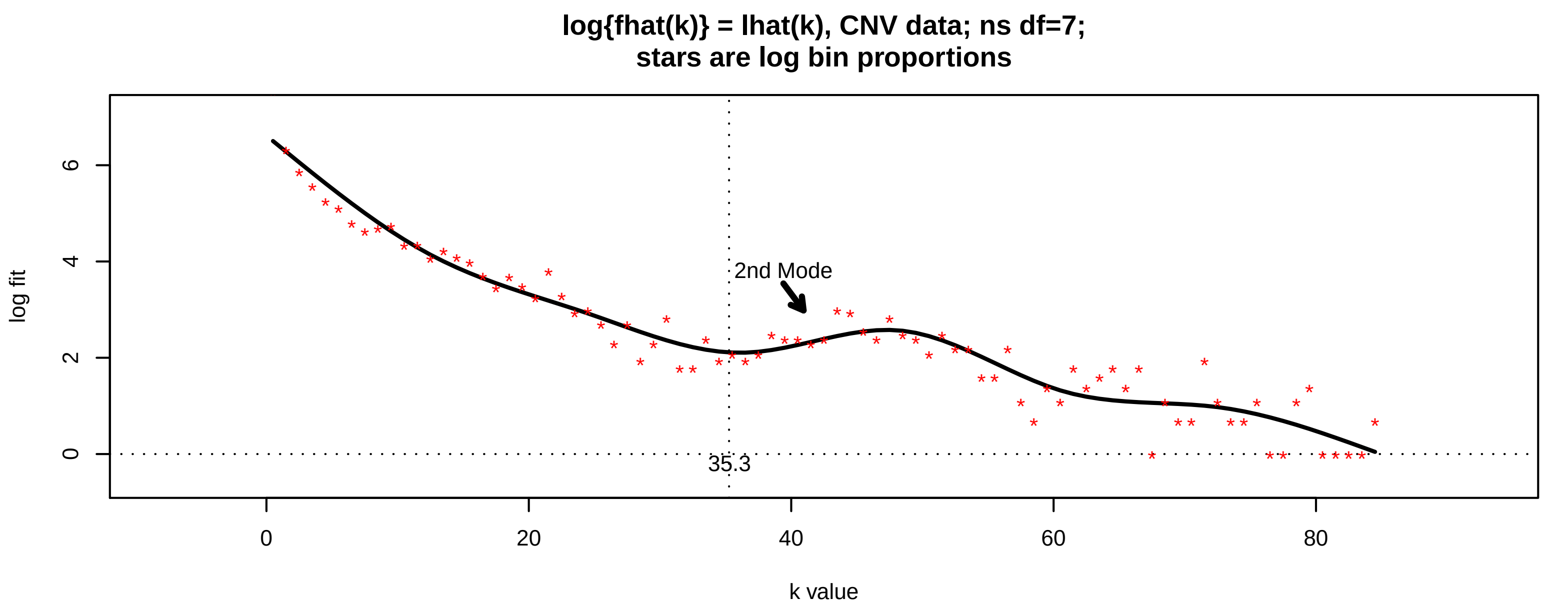


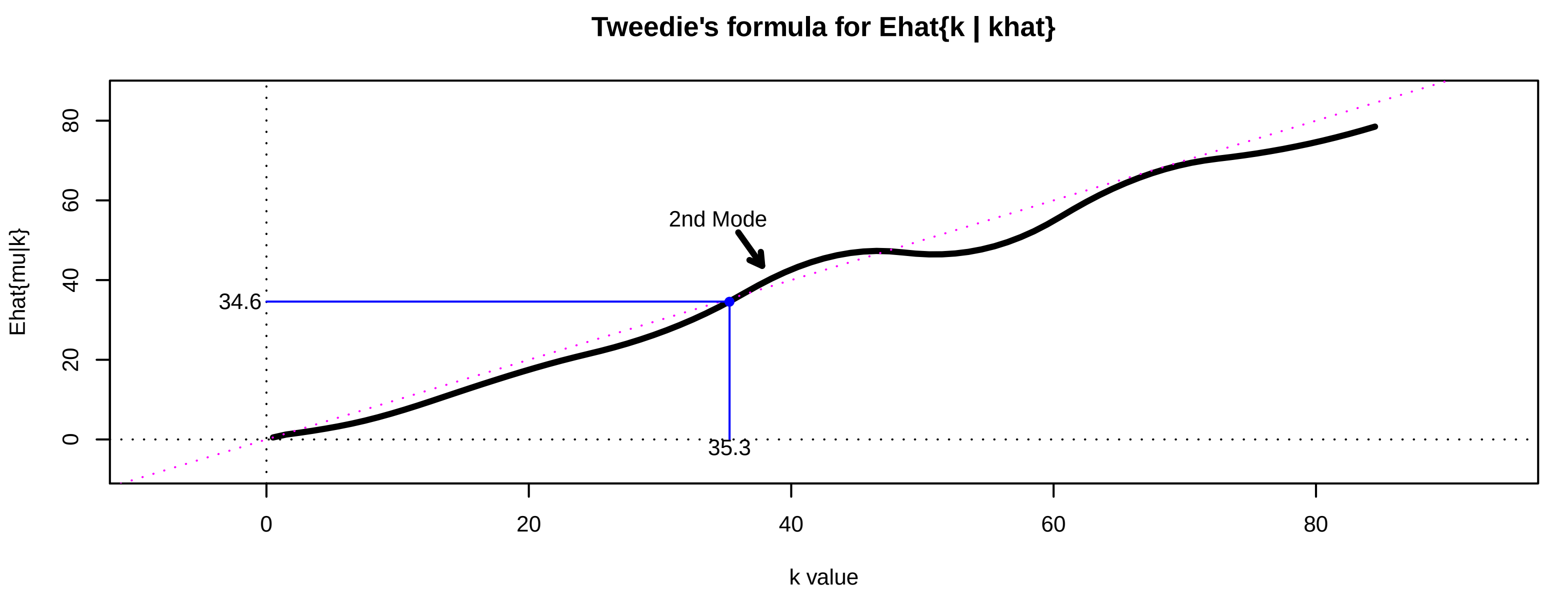
CNV Data
(Efron and Zhang, 2010)

- Copy Number Variation   150 healthy controls measured at 5000 positions
- $\hat{k}_i$ = estimated number of cnv subjects at position i, $i = 1, 2, \ldots, N = 5000$
- Bootstrapping Subjects $\Longrightarrow \hat{k}_i \dot{\sim} \mathcal{N}\left(k_i, \sigma_i^2\right)$ with σ_i^2 increasing in k_i
- *Selected*   position $i = 1755$ with $\hat{k}_i = 35.3$ $\left(\sigma_i^2 \doteq 6.5\right)$
- Empirical Bayes inference for k_i? (Don't need independence!)


k values vs position, cnv data
k value
position
0
20
40
60
80
0
1000
2000
3000
4000
5000
1755


The 5000 k−values for cnv example
(first 3 freqs 1908, 560,357
Frequency
0
50
100
150
200
250
300
k[1755]
=35.3
2nd Mode
0
20
40
60
80
k values


log{fhat(k)} = lhat(k), CNV data; ns df=7;
stars are log bin proportions
log fit
0
2
4
6
2nd Mode
35.3
0
20
40
60
80
k value

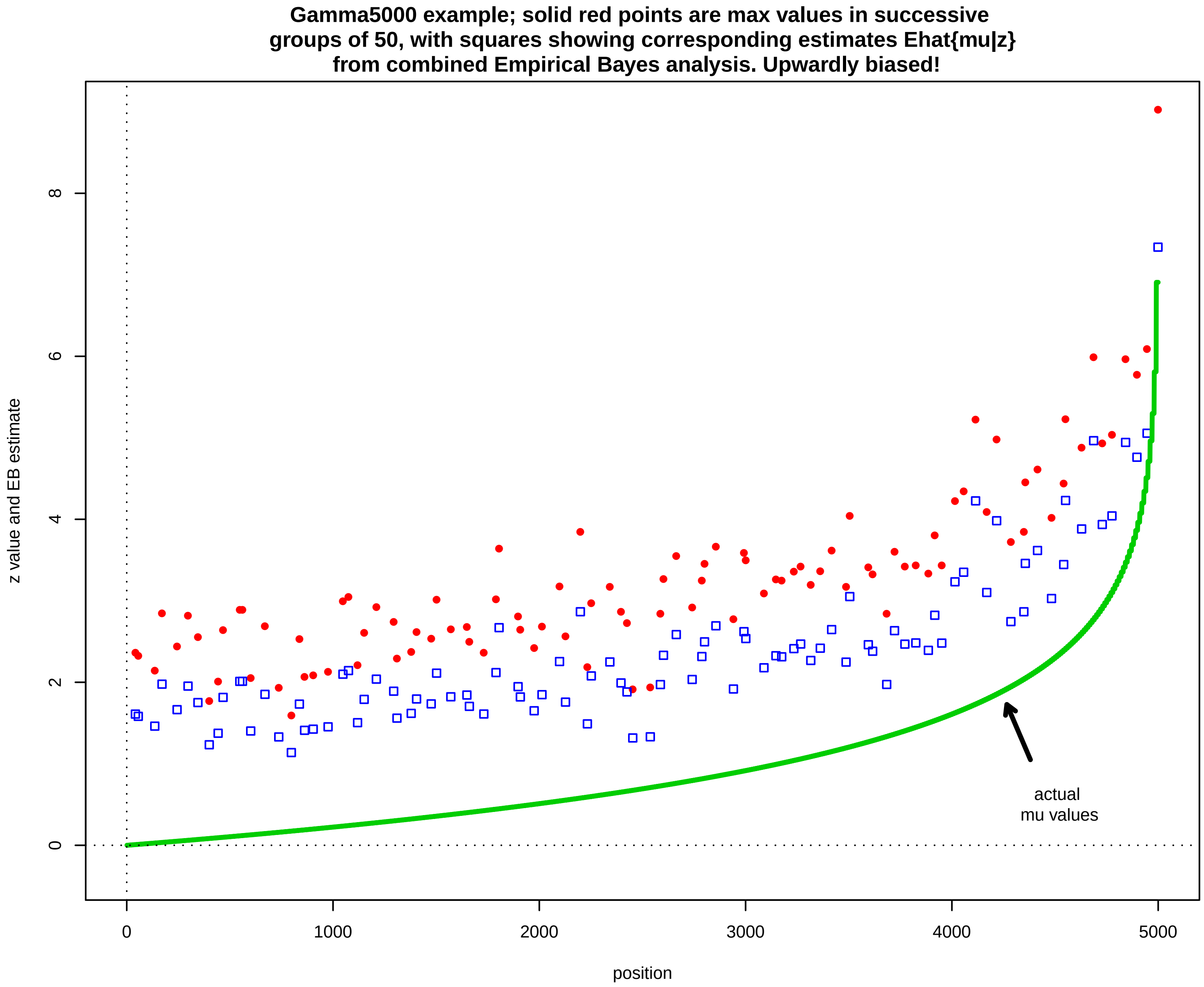

Tweedie's formula for Ehat{k | khat}
Ehat{mu|k}
0
20
40
60
80
2nd Mode
34.6
35.3
0
20
40
60
80
k value

Which "Other Cases" are Relevant?

- EB estimate of k_{1755} depends on which other $\hat{k}_i$'s thought relevant:

Relevant Cases	1:5000	1:2500	1000:2000
$\hat{\mu}_{1755}$	34.6	32.3	29.4

- Large $\hat{k}$'s in 3000:5000 pull up estimate $\hat{\mu}_{1755}$
- *Bayes Problem* Relevant prior $g(k)$ may depend on position



Gamma5000 example; solid red points are max values in successive
groups of 50, with squares showing corresponding estimates Ehat{mu|z}
from combined Empirical Bayes analysis. Upwardly biased!
z value and EB estimate
0
2
4
6
8
0
1000
2000
3000
4000
5000
position
actual
mu values

Relevance

- Let $\rho_0(i)$ be relevance of case i to target case i_0

- Examples

  (1) $\rho_0(i) = 1$ if $i \in i_0 \pm 500$; 0 otherwise

  (2) $\rho_0(i) = \exp\{-|i - i_0|/500\}$

- *Extended Tweedie Formula*

$$\boxed{\hat{\mu}_{i_0} = z_{i_0} + \hat{l}'(z_{i_0}) + \frac{d}{dz} \log \hat{R}_0(z)\Big|_{z_{i_0}}}$$

  where $R_0(z) = E\{\rho_0(i)|z\}$ $\qquad$ $\left[\text{Regress } \rho_0(i) \text{ on } z_i \text{ to get } \hat{R}_0(z).\right]$

- Using (1) or (2) cures bias on previous panel.

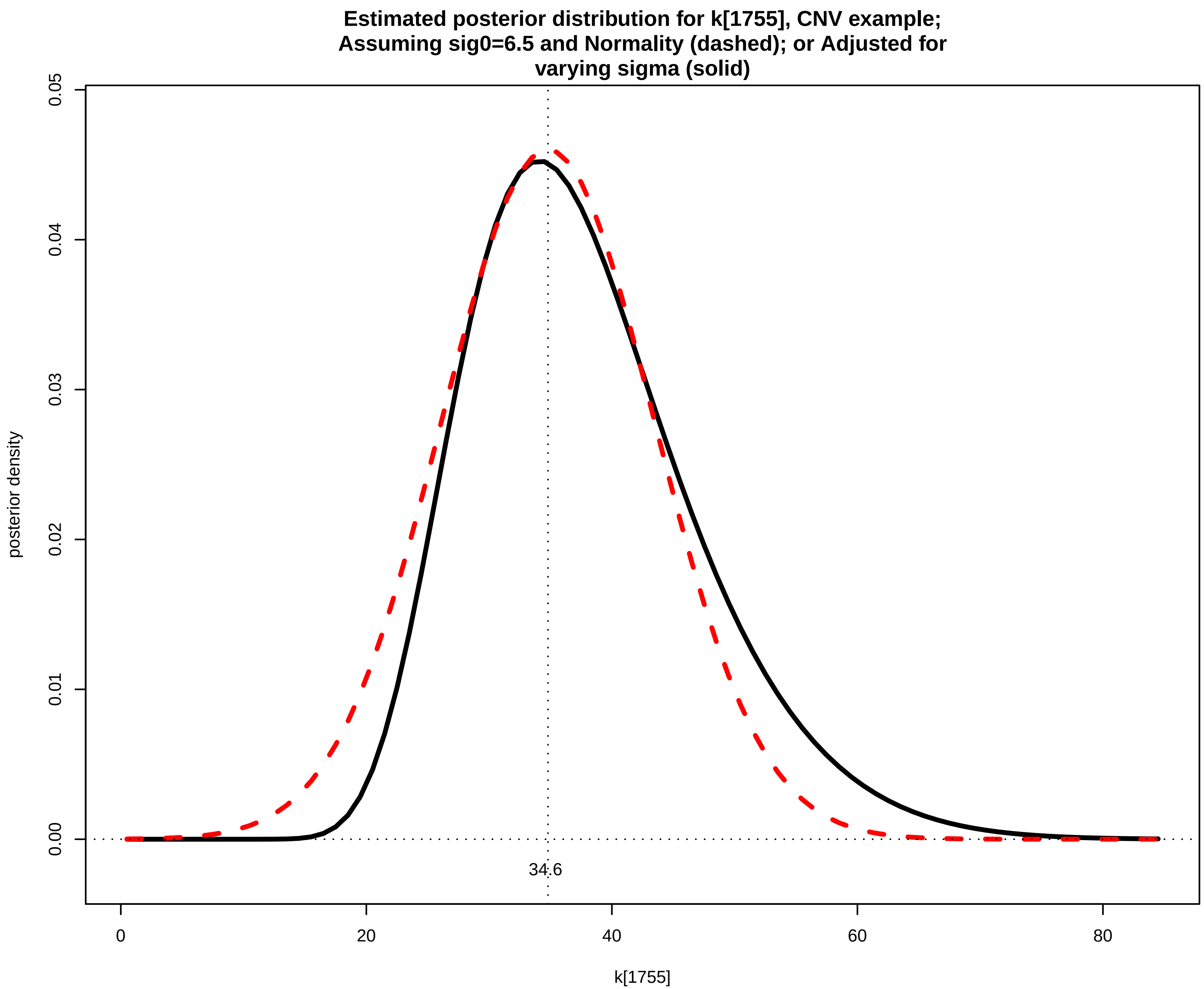
Non-Constant Variance

- If $z \sim \mathcal{N}(\mu, \sigma_0^2)$ then $E\{\mu|z\} = z + \sigma_0^2 l'(z)$
  (took $\sigma_0 = 6.5$ for cnv)

- Suppose $z \sim \mathcal{N}(\mu, \sigma_\mu^2)$

  Theorem $$\boxed{\frac{g(\mu|z_0)}{g_0(\mu|z_0)} = c_0 \lambda_\mu e^{-\frac{1}{2}(\lambda_\mu^2 - 1)\Delta_\mu^2}} \qquad \begin{cases} \lambda_\mu = \sigma_0/\sigma_\mu \\ \Delta_\mu = (\mu - z_0)\big/\sigma_0 \end{cases}$$

  where $\sigma_0 = \sigma_{\mu = z_0}$ and $g_0(\mu|z_0)$ is distribution assuming constant $\sigma = \sigma_0$.

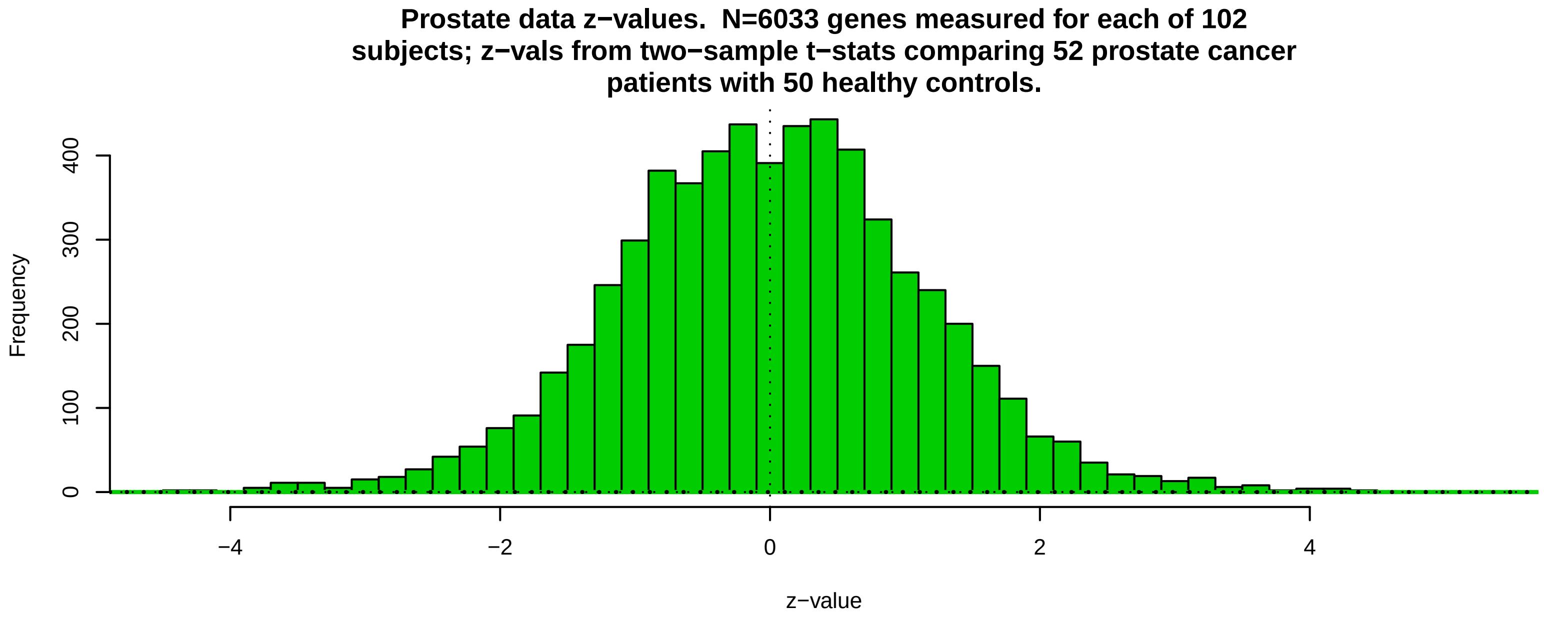

Estimated posterior distribution for k[1755], CNV example;
Assuming sig0=6.5 and Normality (dashed); or Adjusted for
varying sigma (solid)
posterior density
0.00
0.01
0.02
0.03
0.04
0.05
34.6
0
20
40
60
80
k[1755]

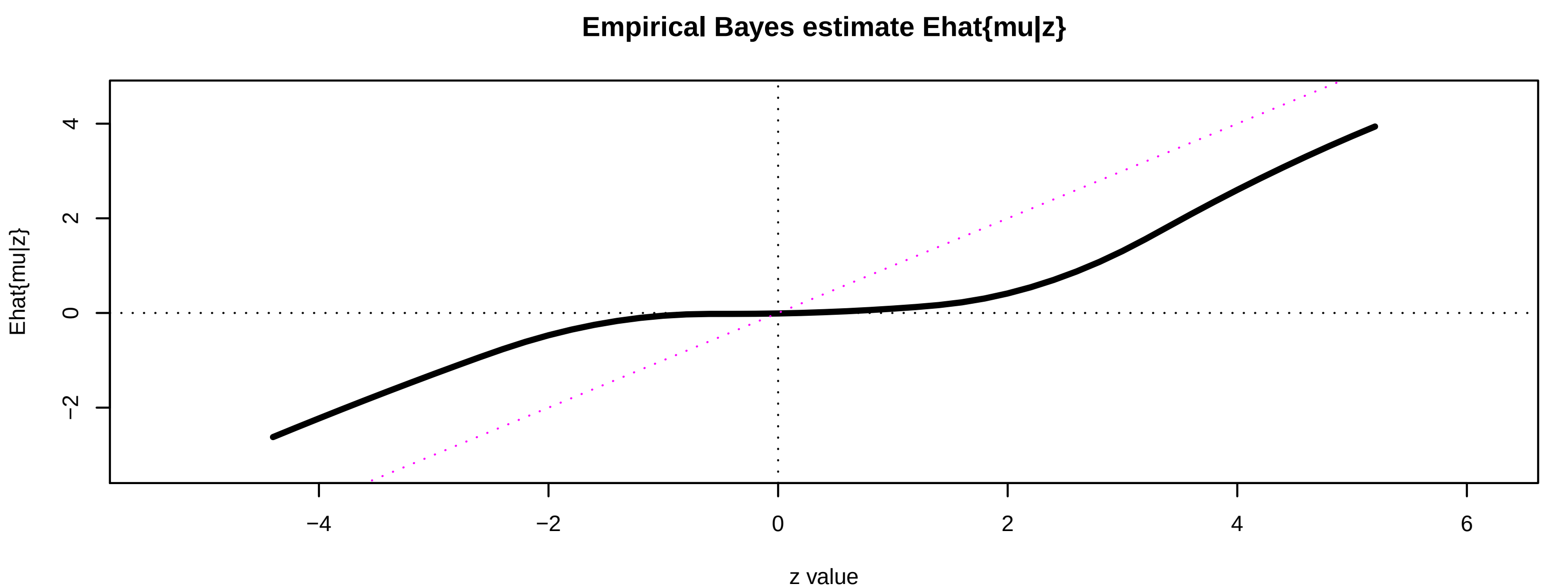
False Discovery Rates

- $\mu \sim g(\cdot) = \pi_0 \delta_0(\cdot) + \pi_1 g_1(\cdot)$ and $z|\mu \sim \mathcal{N}(\mu, 1)$
- $\pi_0 =$ prior $\Pr\{\text{null}\}$
- Then $\text{fdr}(z) = \Pr\{\text{null}|z\} = \pi_0 \varphi(z) \Big/ f(z)$
  ("local false discovery rate")

$$\boxed{-\frac{d}{dz} \log \text{fdr}(z) = z + l'(z) = E\{\mu|z\}}$$

- *Tail Area Fdr* $\text{Fdr}(z_0) = \Pr\{\text{null}|z \leq z_0\} = \pi_0 \Phi(z_0) \Big/ F(z)$
  where $F(z)$ is cdf of mixture density $f(z)$


Prostate data z−values. N=6033 genes measured for each of 102
subjects; z−vals from two−sample t−stats comparing 52 prostate cancer
patients with 50 healthy controls.
Frequency
0
100
200
300
400
−4
−2
0
2
4
z−value

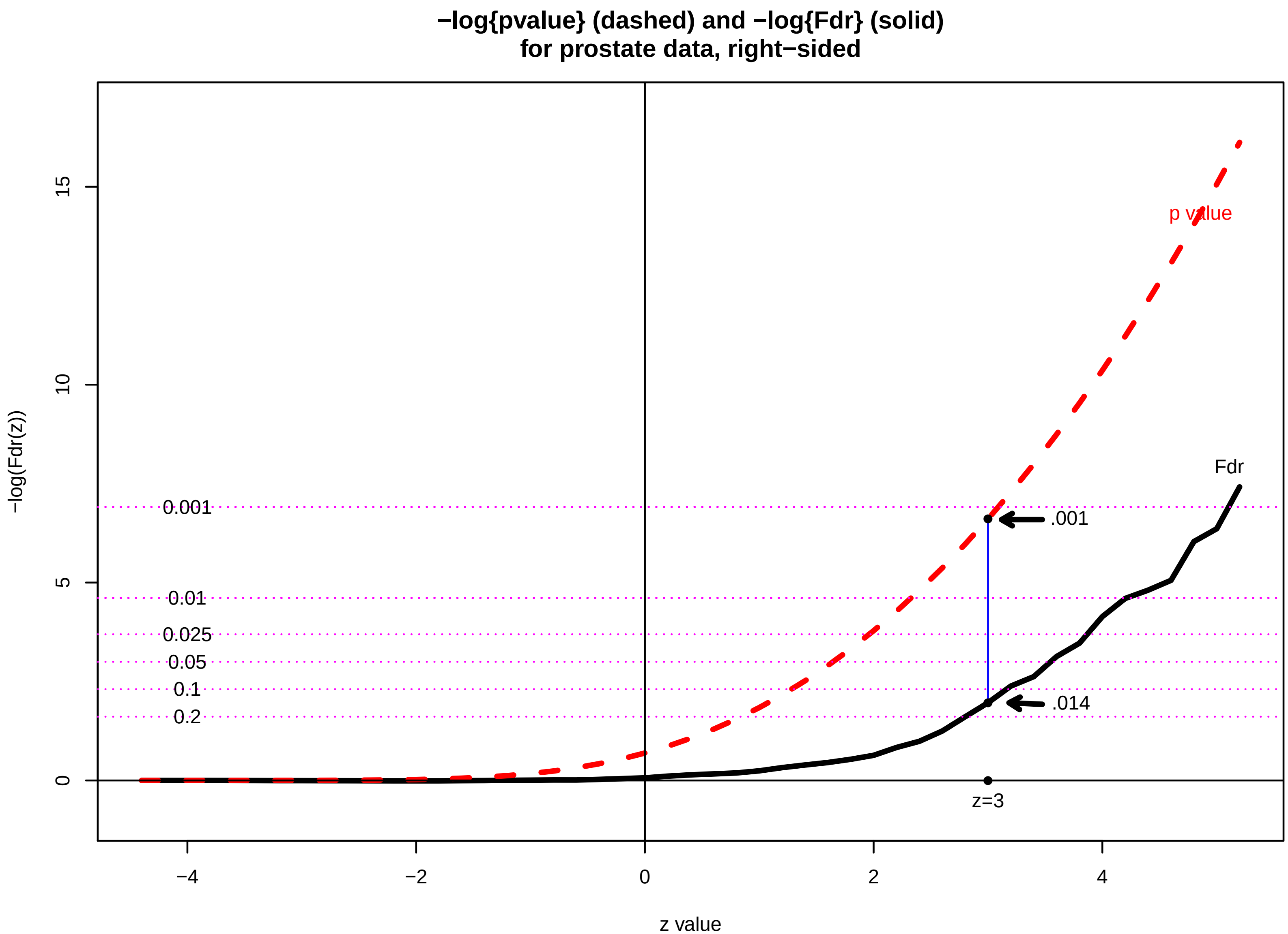

Empirical Bayes estimate Ehat{mu|z}
Ehat{mu|z}
−2
0
2
4
−4
−2
0
2
4
6
z value

p-Value Selection Bias

- Observe $z_1, z_2, \ldots, z_N$ and select smallest values of $p_i = \Phi(z_i)$
- Significance?
- *Benjamini–Hochberg* Reject small values of $\widehat{\mathrm{Fdr}}_i = \pi_0 p_i \Big/ \hat{F}(z_i)$, or large values of

$$\boxed{-\log\left(\widehat{\mathrm{Fdr}}_i\right) = -\log(p_i) + \log\left(\hat{F}(z_i)\Big/\hat{\pi}_0\right)} \quad \text{(usually } \hat{\pi}_0 = 1\text{)}$$

↑ EB estimate ↑ frequentist estimate ↑ EB correction


−log{pvalue} (dashed) and −log{Fdr} (solid)
for prostate data, right−sided
−log(Fdr(z))
z value
p value
Fdr
0.001
0.01
0.025
0.05
0.1
0.2
.001
.014
z=3
0
5
10
15
−4
−2
0
2
4

References

Efron, B. and Zhang, N. (2010). False discovery rates and copy number variation, http://stat.stanford.edu/∼brad/.

Muralidharan, O. (2009). High dimensional exponential family estimation via empirical Bayes, under review.

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proc. 3rd Berkeley Symposium*. Berkeley: UC Press, 157–163.

Zhang, C.-H. (1997). Empirical Bayes and compound estimation of normal means. *Statist. Sinica* 7: 181–193.