# Correlated $z$-Values and the Accuracy of Large-Scale Statistical Estimates

Bradley Efron
*Stanford University*

# Correlation and Accuracy

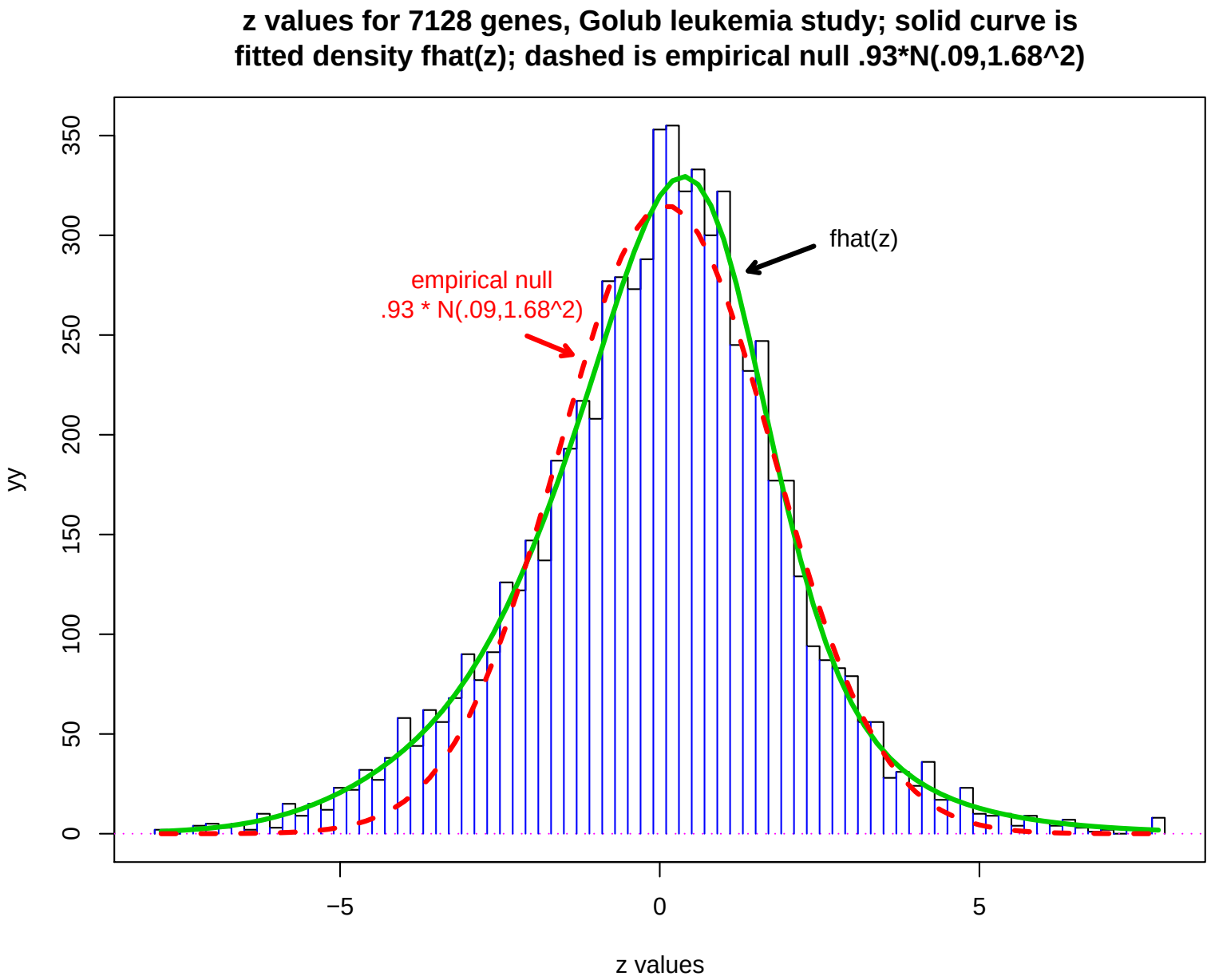
- **Modern Scientific Studies** $N$ cases (genes, SNPs, pixels, ...) each with its own summary statistic "$z_i$", $i = 1, 2, \ldots, N$
- *Estimate of interest* $\hat{\theta} = s(\boldsymbol{z})$ $\left[\text{e.g., } \hat{\theta} = \#\{z_i > 3\}/N\right]$
- **Question** How accurate is $\hat{\theta}$?
- Easy answer if $z_i$'s independent (but usually not!)
- Practical answer possible if $z_i$'s are correlated normal variates

# Leukemia Microarray Study

(Golub et al., 1999)

- 72 leukemia patients: $n_1 = 47$ "ALL", $n_2 = 25$ "AML"
  - Acute Lymphoblastic Leukemia
  - Acute Myeloid Leukemia
- $N = 7128$ genes
- $t_i$ = two-sample $z$-statistic comparing AML with ALL patient expression levels for gene $i$
- $z_i = \Phi^{-1} F_{70}(t_i)$ $\quad$ [$\Phi, F_{70}$ cdf's $\mathcal{N}(0,1), t_{70}$ distributions]

$$\boxed{H_0 : z_i \sim \mathcal{N}(0,1)} \quad \text{("theoretical null")}$$


z values for 7128 genes, Golub leukemia study; solid curve is
fitted density fhat(z); dashed is empirical null .93*N(.09,1.68^2)
empirical null
.93 * N(.09,1.68^2)
fhat(z)
yy
0
50
100
150
200
250
300
350
−5
0
5
z values

# Accuracy of Empirical CDF

- $\hat{F}(x) = \#\{z_i > x\}/N$
- How accurate?

| $x$: | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $\hat{F}(x)$ | .29 | .13 | .057 | .025 | .010 |
| $1000 \cdot \widehat{\mathrm{sd}}$ | 17.1 | 21.5 | 10.1 | 4.0 | 1.9 |
| $1000 \cdot \widehat{\mathrm{sd}}_0$ | 5.4 | 4.0 | 2.7 | 1.8 | 1.2 |
| $1000 \cdot \widehat{\mathrm{sd}}_{\mathrm{perm}}$ | 20.9 | 9.9 | 1.4 | 0.1 | .000 |

# CDF Accuracy Formula

$$\text{Var}\left\{\hat{F}(x)\right\} \doteq \underbrace{\left\{\frac{\hat{F}(x)\big(1-\hat{F}(x)\big)}{N}\right\}}_{\text{independence}} + \underbrace{\left\{\frac{\hat{\sigma}_0^2\hat{\alpha}\hat{f}^{(1)}(x)}{\sqrt{2}}\right\}^2}_{\text{correlation penalty}}$$

- $\hat{\sigma}_0 = 1.68$ from empirical null
- $\hat{\alpha} = .11$ estimated RMS correlation
- $\hat{f}^{(1)}(x)$ first derivative of estimate $\hat{f}(x)$
- Depends on normality: $z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$

# Digression: The Non-Null Distribution of $z$-Values

- $z$-value is a test statistic $\sim \mathcal{N}(0,1)$ under $H_0$
- *Theorem* *Under reasonable conditions the non-null distribution of $z$ is*

$$z \sim \mathcal{N}(\mu, \sigma^2) + O_p(1/n)$$

*where*

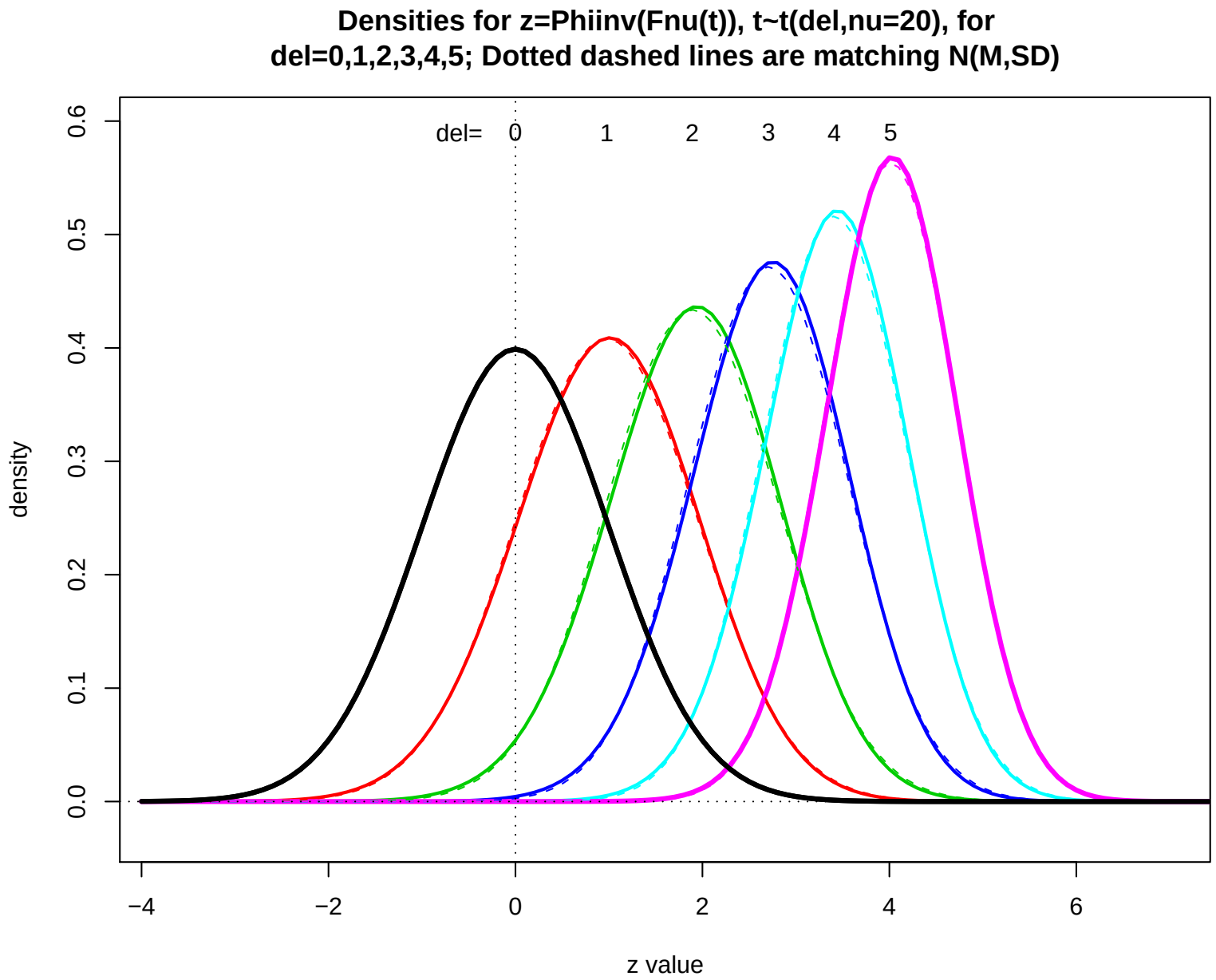
$$\sigma^2 = 1 + O\left(1/n^{\frac{1}{2}}\right)$$

- Normality degrades more slowly than unit standard deviation
- Helps justify model $z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$

# Student-$t$ $z$-Values

- $t \sim t_\nu(\delta)$ [noncentral-$t$, noncentrality $\delta$, $df = \nu$]
- $H_0 : \delta = 0$
- $z = \Phi^{-1} F_\nu(t)$ [$F_\nu$ *central* $t$ cdf, $df = \nu$]

  so under $H_0$, $z \sim \mathcal{N}(0, 1)$
- What if $\delta \neq 0$?

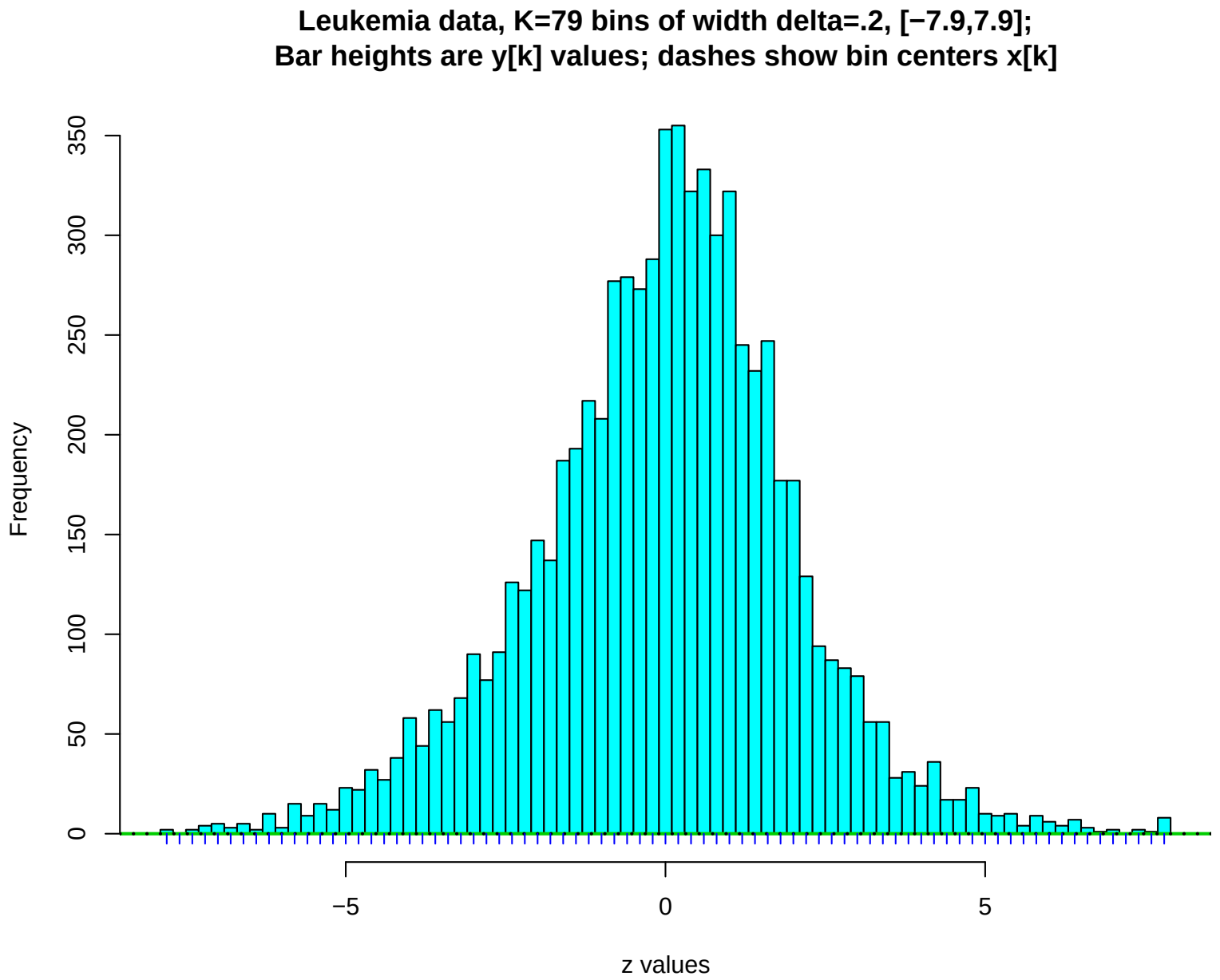

Densities for z=Phiinv(Fnu(t)), t~t(del,nu=20), for
del=0,1,2,3,4,5; Dotted dashed lines are matching N(M,SD)
del= 0 1 2 3 4 5
density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
−4
−2
0
2
4
6
z value

# The Count Vector $\boldsymbol{y}$

- Partition range $\mathcal{Z}$ of $z$ into $K$ bins: $\mathcal{Z} = \bigcup_{k=1}^{K} \mathcal{Z}_k$
- Each bin of width "$\Delta$"
- Bin centers "$x_k$", $k = 1, 2, \ldots, K$

  (Leukemia histogram: $\mathcal{Z} = [-7.9, 7.9], \Delta = .2, K = 79$)
- *Counts* $y_k = \#\{z_i \in \mathcal{Z}_k\}$
- $\boldsymbol{y} = (y_1, y_2, \ldots, y_K)'$
- Count vector $\boldsymbol{y}$ is discretized order statistic of $\boldsymbol{z}$

  (most statistics of interest of form $\hat{\theta} = m(\boldsymbol{y})$)


Leukemia data, K=79 bins of width delta=.2, [−7.9,7.9];
Bar heights are y[k] values; dashes show bin centers x[k]
Frequency
0
50
100
150
200
250
300
350
−5
0
5
z values

# Multi-Class Model

- Suppose $z_i$'s are in "classes" $C_1, C_2, \ldots, C_C$, with

$$z_i \sim \mathcal{N}(\mu_c, \sigma_c^2) \quad \text{for } z_i \in C_c$$

- $N_c = \#\{C_c\}, \quad p_c = N_c/N \qquad [\text{so } \sum_c N_c = N,\ \sum_c p_c = 1]$
- *Correlation distribution* $g_{cd}(\rho)$ = empirical density of $N_c \cdot N_d$ correlations between members of $C_c$, $C_d$
- Likewise $g_{cc}(\rho)$ for $N_c(N_c - 1)/2$ in $C_c$
- *Assume* $g_{cd}$ all equal $g(z)$

  [actually only need 2nd moments]

# Crucial Definitions

- $x_{kc} = \dfrac{x_k - \mu_c}{\sigma_c}$ "Adjusted bin center"
- $\pi_{kc} = \Delta \cdot \varphi(x_{kc}) \Big/ \sigma_c \doteq \text{Prob}_c\{z_i \in \mathcal{Z}_k\}$ $\left[\varphi(u) = e^{-u^2/2} \Big/ \sqrt{2\pi}\right]$
- *Expectation of* $\boldsymbol{y}$: $E\{\boldsymbol{y}\} = N \sum_c p_c \boldsymbol{\pi}_c$ $[\boldsymbol{\pi}_c = (\ldots, \pi_{kc}, \ldots)']$
- $\varphi_\rho(u, v) =$ standard bivariate normal density, correlation $\rho$
- $\lambda_\rho(u, v) = \dfrac{\varphi_\rho(u, v)}{\varphi(u)\varphi(v)} - 1$ $[= 0 \text{ if } \rho = 0]$

$$\lambda(u, v) = \int_{-1}^{1} \lambda_\rho(u, v) g(\rho)\, d\rho$$

# Exact Covariance of $\boldsymbol{y}$

*Theorem* *Under the multi-class model,*

$$\text{cov}(\boldsymbol{y}) = \mathbf{cov}_0 + \mathbf{cov}_1,$$

$$\mathbf{cov}_0 = N \sum_c p_c \{\text{diag}(\boldsymbol{\pi}_c) - \boldsymbol{\pi}_c \pi_c'\} \qquad [\textit{independence}]$$

*and*

$$\boxed{\mathbf{cov}_1 = N^2 \sum_c \sum_d p_c p_d \boldsymbol{B}_{cd} - N \sum_c p_c \boldsymbol{B}_{cc}} \quad \langle 1 \rangle \; [\textit{corr penalty}]$$

*where* $B_{cd}(k,l) = \pi_{kc}\pi_{ld}\lambda(x_{kc}, x_{ld})$.

The second term of $\langle 1 \rangle$ is negligible for $N$ large.

## Mehler's Identity (Lancaster, 1958)

- $\lambda_\rho(u,v) = \sum_{j\geq 1} \dfrac{\rho^j}{j!} h_j(u)h_j(v)$ $\quad [h_j = j\text{th Hermite polynomial}]$
- So $\displaystyle\int_{-1}^{1} \lambda_\rho(u,v)g(\rho)\,d\rho = \sum_{j\geq 1} \frac{\alpha_j}{j!} h_j(u)h_j(v) \quad \left[\alpha_j = \int_{-1}^{1} \rho^j g(\rho)\,d\rho\right]$

  and

$$B_{cd}(x_{kc}, x_{ld}) = \pi_{kc}\pi_{ld} \sum_{j\geq 1} \frac{\alpha_j}{j!} h_j(x_{kc})h_j(x_{ld})$$

$$= \Delta^2 \sum_{j\geq 1} \frac{\alpha_j}{j!} \varphi^{(j)}(x_{kc})\varphi^{(j)}(x_{ld}) \Big/ \sigma_c\sigma_d \qquad \langle 2 \rangle$$

- $\varphi^{(j)}$ the $j$th derivative of $\varphi$ $\qquad [\text{using } \varphi h_j = (-1)^j \varphi^{(j)}]$

# Three Simplifications of **cov**

- Drop second term of $\langle 1 \rangle$
- Microarray standardization methods make $\alpha_1 \doteq 0$ in $\langle 2 \rangle$
- Now $\alpha_2 = \int_{-1}^{1} \rho^2 g(\rho)$ is the lead term
- Higher terms ignorable if $\alpha_2$ small

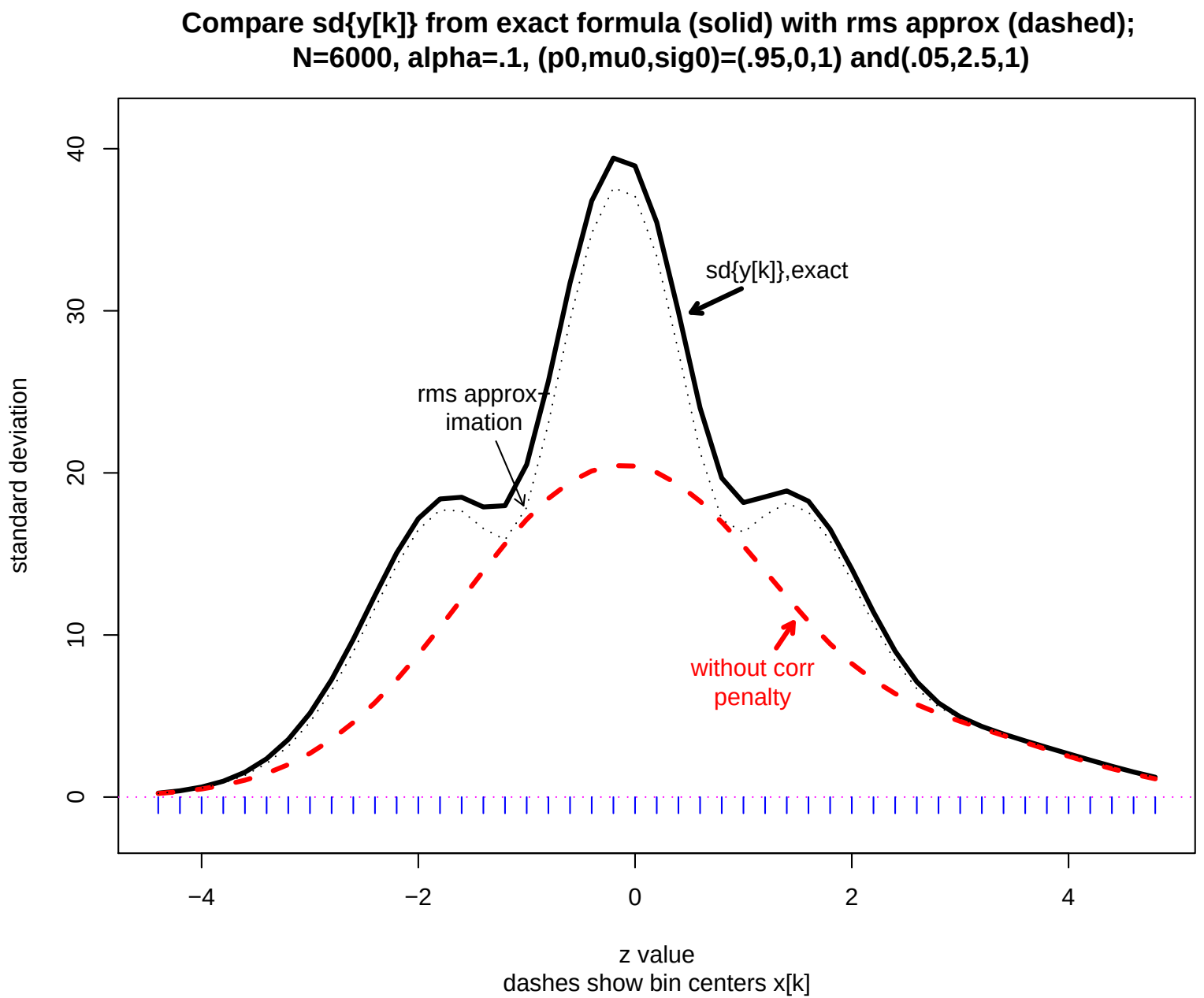
**Simplified Formula**

Letting $\alpha = \alpha_2^{\frac{1}{2}}$ and $\bar{\phi}_k^{(2)} = \sum_c p_c \varphi^{(2)}\left(\frac{x_{kc}-\mu_c}{\sigma_c}\right) \Big/ \sigma_c$

$$\boxed{\mathbf{cov}_1 \doteq (N\Delta\alpha)^2 \bar{\boldsymbol{\phi}}^{(2)} \bar{\boldsymbol{\phi}}^{(2)\prime} \Big/ 2}$$

# Numerical Comparison

- $N = 6000,\ \alpha = .1$
- Two classes: $(p_c, \mu_c, \sigma_c) = \begin{cases} (.95, 0, 1) \\ (.05, 2.5, 1) \end{cases}$
- Next figure compares standard deviations (square roots diagonal elements) of exact $\mathbf{cov}(\boldsymbol{y})$ and rms approximation


Compare sd{y[k]} from exact formula (solid) with rms approx (dashed);
N=6000, alpha=.1, (p0,mu0,sig0)=(.95,0,1) and(.05,2.5,1)
standard deviation
0
10
20
30
40
sd{y[k]},exact
rms approx-
imation
without corr
penalty
−4
−2
0
2
4
z value
dashes show bin centers x[k]

# CDF Accuracy (Owen, 2005)

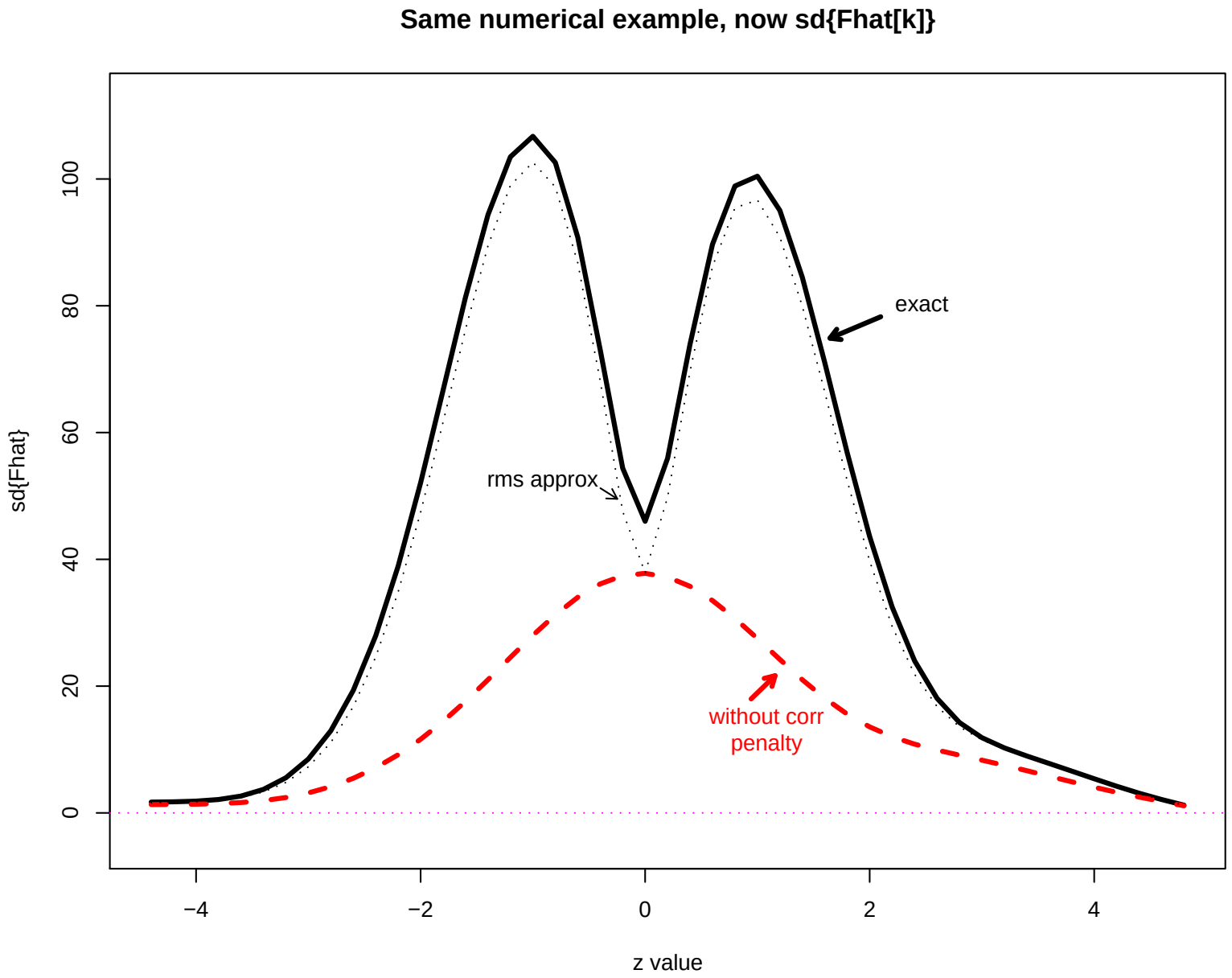
- Let $\hat{F}_k = \#\{z_i \text{ in bins} \geq k\}/N$ [right-sided empirical cdf]

  so $\quad \hat{\boldsymbol{F}} = \dfrac{1}{N}\boldsymbol{M}\boldsymbol{y}$ where $M_{jk} = \begin{cases} 1 & \text{if } j \leq k \\ 0 & \text{if } j > k \end{cases}$

- Then $\mathbf{Cov}(\hat{\boldsymbol{F}}) = \mathbf{Cov}_0 + \mathbf{Cov}_1$ where $\mathbf{Cov}_0$ from usual multinomial, and

$$\boxed{\mathbf{Cov}_1 \doteq \frac{\alpha^2}{2}\bar{\boldsymbol{\varphi}}^{(1)}\bar{\boldsymbol{\varphi}}^{(1)\prime}}$$

$$\left[\bar{\varphi}_k^{(1)} = \sum_c p_c \varphi^{(1)}\left(\frac{x_k - \mu_c}{\sigma_c}\right),\ \varphi^{(1)}(u) = -\varphi(u)u\right]$$


Same numerical example, now sd{Fhat[k]}
sd{Fhat}
0
20
40
60
80
100
−4
−2
0
2
4
z value
exact
rms approx
without corr
penalty

# Practical Implementation

- $\mathbf{Cov}_1$ requires estimating $\alpha$ and $\bar{\varphi}^{(1)}$
- $\alpha$ = rms correlation: not difficult in microarray studies
- $\bar{\varphi}_k^{(1)} = \sum_c p_c \varphi^{(1)}\left(\frac{x_k - \mu_c}{\sigma_c}\right)$: seems to require class parameters $(p_c, \mu_c, \sigma_c)$, but can be finessed in some situations

# Estimation of RMS Correlation $\alpha$

- $\hat{\rho}_{ii'}$ = empirical correlation, rows $i$, $i'$ of $\boldsymbol{X}$,
  $N \times n$ expression matrix
- $\{\hat{\rho}_{ii'}\}$ has mean and variance $(m, v)$
  [leukemia = $(.00, .19^2)$]

$$\hat{\alpha}^2 = \frac{n}{n-1}\left(v - \frac{1}{n-1}\right)$$

| | ALL | AML | Both |
|---|---|---|---|
| $\hat{\alpha}$: | .121 | .109 | .114 |

# Estimation of $\bar{\varphi}_k^{(1)}$

- *Multi-Class Model*: $f(x_k) = \sum_c p_c \varphi\left(\frac{x_k - \mu_c}{\sigma_c}\right)\Big/\sigma_c$, so

$$f^{(1)}(x_k) = \sum_c p_c \varphi^{(1)}\left(\frac{x_k - \mu_c}{\sigma_c}\right)\Big/\sigma_c^2$$

- *If* $\sigma_c \equiv \sigma_0$ $\quad \boxed{\sigma_0^2 f^{(1)}(x_k) = \sum_c p_c \varphi'\left(\frac{x_k - \mu_c}{\sigma_c}\right) = \bar{\varphi}_k^{(1)}}$

- $\boldsymbol{y} \to \hat{\boldsymbol{f}} \to \hat{\boldsymbol{f}}^{(1)} \to \bar{\boldsymbol{\varphi}}^{(1)}$ [★ formula correlation penalty]

# More General Accuracy Estimates

- "$\boldsymbol{Q}$" $q$-dimensional statistic of interest: $\boldsymbol{Q} = Q(\boldsymbol{y})$
- *Influence Function*

$$d\boldsymbol{Q} = \hat{\boldsymbol{D}}\, d\boldsymbol{y} \qquad \left[\hat{D}_{jk} = \partial Q_j / \partial y_k\right]$$

$$\boxed{\widehat{\mathbf{cov}}(\boldsymbol{Q}) = \hat{D}\mathbf{cov}(\boldsymbol{y})\hat{D}'}$$

# Example: Accuracy of $\hat{f}$

- $z \to y \to \hat{f}$ by Poisson GLM

  of counts $y_k$ on polynomial $(x_k)$

- $Q = \log(\hat{f}) = (\ldots \log f(x_k) \ldots)'$

- $$\hat{D} = M \left[ M' \text{diag}\left(\hat{f}\right) M \right] M' \Big/ N\Delta$$

  with $M$ the GLM structure matrix (Efron, 2007b)

# The Local False Discovery Rate

- *Bayes*: $\left\{\begin{array}{ll} p_0 = \Pr\{\text{null}\} & f_0(z) = \text{null density} \\ p_1 = \Pr\{\text{non-null}\} & f_1(z) = \text{non-null density} \end{array}\right\}$

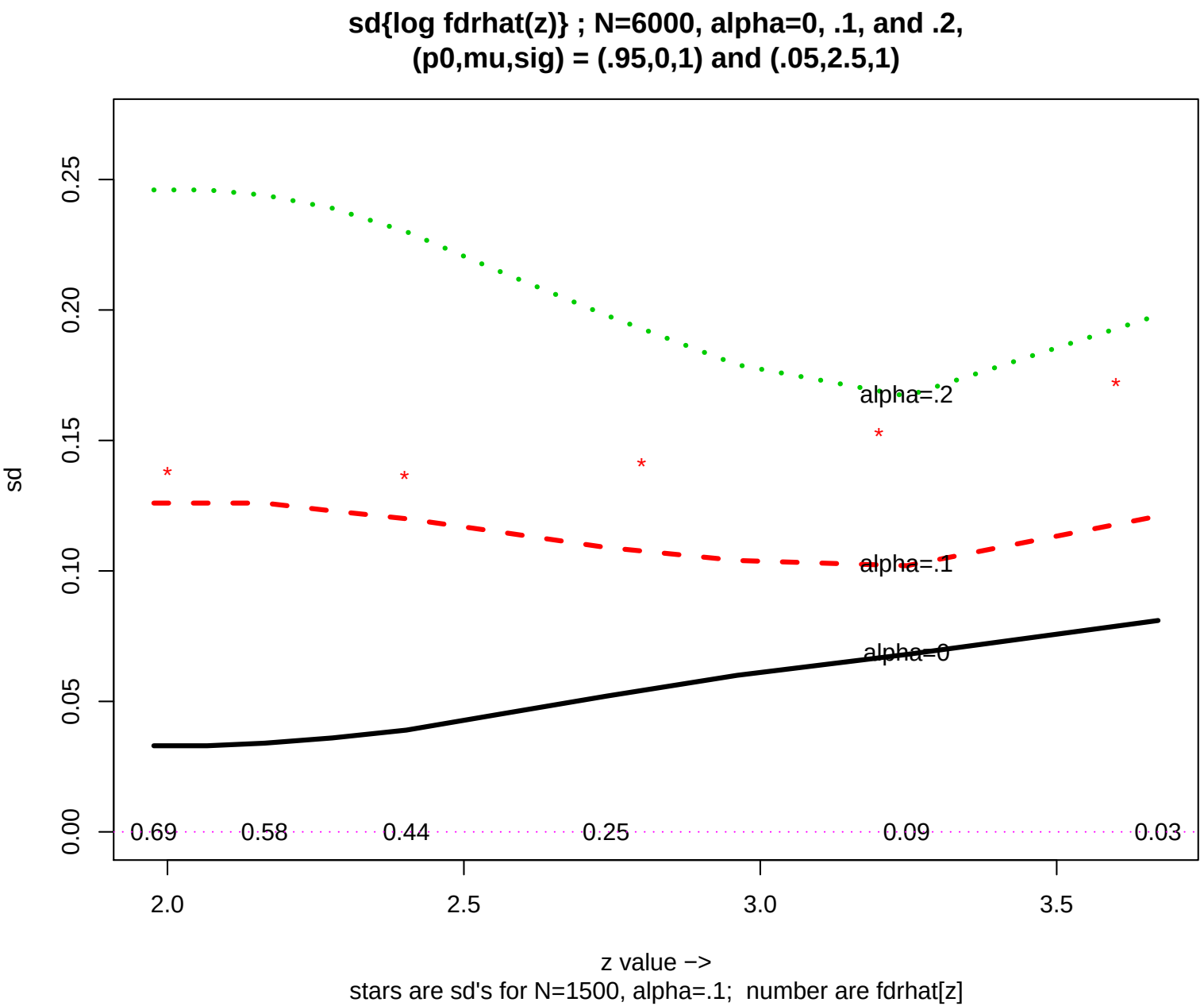
$$f(z) = p_0 f_0(z) + p_1 f_1(z)$$

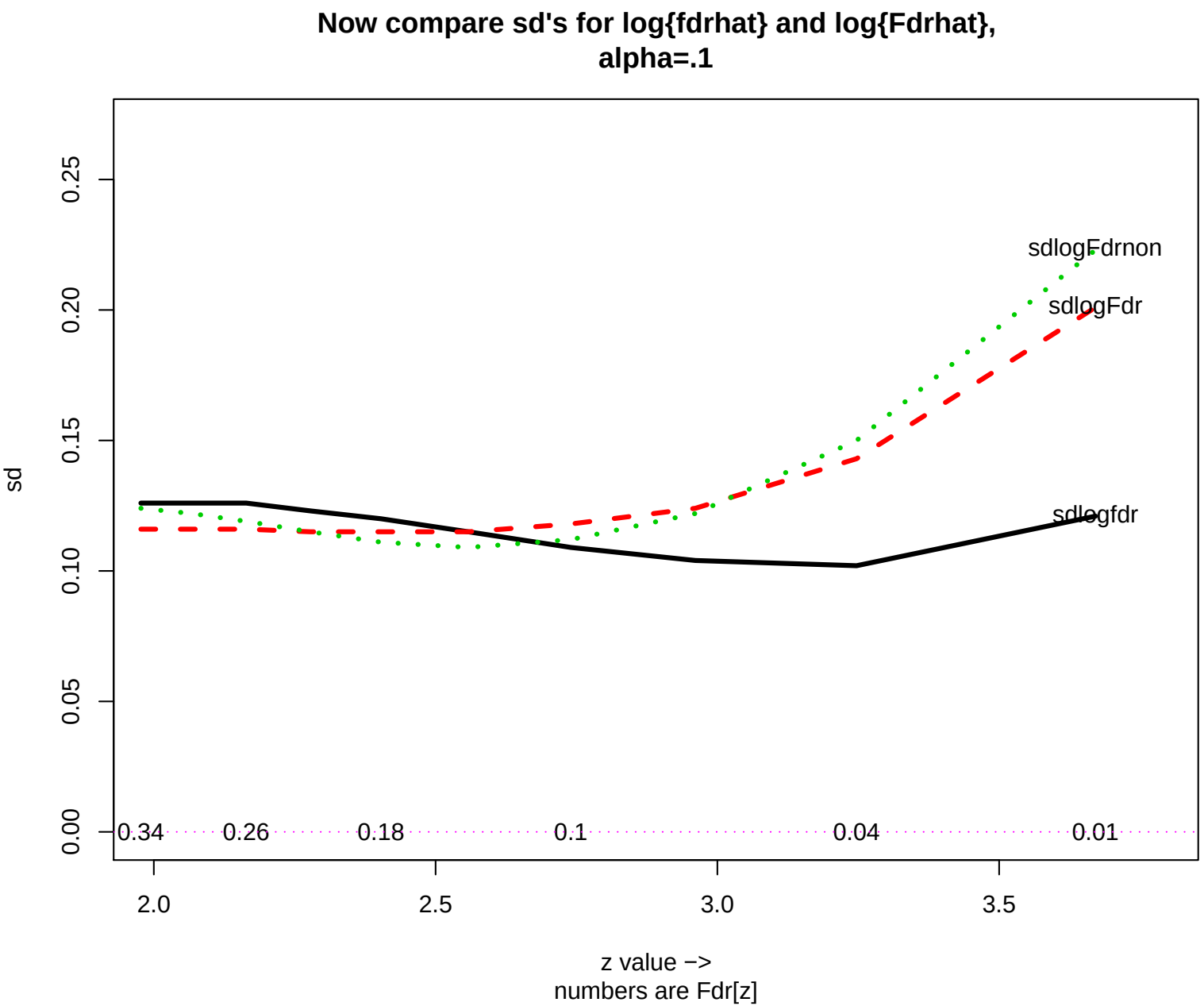
- *Local False Discovery Rate*

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

- $\widehat{\textbf{fdr}} = p_0 \boldsymbol{f}_0/\hat{\boldsymbol{f}}$ $\qquad$ $\widehat{\textbf{lfdr}} = \log(\widehat{\textbf{fdr}}) = \log(p_0 \boldsymbol{f}_0) - \log \hat{\boldsymbol{f}}$

- If $p_0 f_0$ known then $\widehat{\textbf{cov}}(\widehat{\textbf{lfdr}}) = \widehat{\textbf{cov}}(\log \hat{\boldsymbol{f}})$


sd{log fdrhat(z)} ; N=6000, alpha=0, .1, and .2,
(p0,mu,sig) = (.95,0,1) and (.05,2.5,1)
sd
0.00
0.05
0.10
0.15
0.20
0.25
alpha=.2
alpha=.1
alpha=0
0.69
0.58
0.44
0.25
0.09
0.03
2.0
2.5
3.0
3.5
z value −>
stars are sd's for N=1500, alpha=.1; number are fdrhat[z]


Now compare sd's for log{fdrhat} and log{Fdrhat},
alpha=.1
sd
0.00
0.05
0.10
0.15
0.20
0.25
sdlogFdrnon
sdlogFdr
sdlogfdr
0.34
0.26
0.18
0.1
0.04
0.01
2.0
2.5
3.0
3.5
z value −>
numbers are Fdr[z]

# Accuracy When $p_0 f_0(z)$ Must Be Estimated

- Program `locfdr` estimates $p_0$ and $f_0$ from central histogram counts, the "empirical null"
- Section 5 of Efron (2007b) gives influence functions $\hat{D}$
- *Next Table* $\text{sd}\left\{\log \widehat{\text{fdr}}(z)\right\}$ using "central matching" method, for $\alpha = .1$ case, as in Figures

## $\text{sd}\{\log \widehat{\text{fdr}}(z)\}$

| $z$: | 1.98 | 2.16 | 2.40 | 2.74 | 3.25 |
|---|---|---|---|---|---|
| empirical null: | .18 | .26 | .36 | .54 | .83 |
| theoretical null: | .13 | .13 | .12 | .11 | .10 |
| fdr: | .69 | .58 | .44 | .25 | .09 |

- Still worse for $\log \widehat{\text{Fdr}}$
- Better using MLE option
- Some of increase is "signal" rather than noise (Efron, 2007a)

# References

Csörgő, S. and Mielniczuk, J. (1996). The empirical process of a short-range dependent stationary sequence under Gaussian subordination. *Probab. Theory Related Fields* 104: 15–25.

Efron, B. (2007a). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* 102: 93–103.

Efron, B. (2007b). Size, power and false discovery rates. *Ann. Statist.* 35: 1351–1377.

Golub, T., Slonim, D. K. and Tamayo, P. et al. (1999). Molecular Classification of Cancer: Class Discovery and

Class Prediction by Gene Expression Monitoring. *Science* 286: 531–537.

Lancaster, H. O. (1958). The structure of bivariate distributions. *Ann. Math. Statist.* 29: 719–736.

Owen, A. B. (2005). Variance of the number of false discoveries. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67: 411–426.

Qiu, X., Klebanov, L. and Yakovlev, A. (2005). Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. *Stat. Appl. Genet. Mol. Biol.* 4: Art. 34, 32 pp. (electronic).