

Bootstrap Methods and the Accuracy of Large-Scale Estimators

Bradley Efron

Stanford University

Correlation and Accuracy

- **Modern Scientific Studies** N cases (genes, SNPs, pixels, ...) each with its own summary statistic “ z_i ”, $i = 1, 2, \dots, N$
- $N \sim 10,000$
- *Estimate of interest* $\hat{\theta} = s(\mathbf{z})$ [e.g., $\hat{\theta} = \#\{z_i > 3\}/N$]
- **Question** How accurate is $\hat{\theta}$?
- Easy answer if z_i 's independent (but usually not!)
- Troubles for the bootstrap

Leukemia Microarray Study

(Golub et al., 1999)

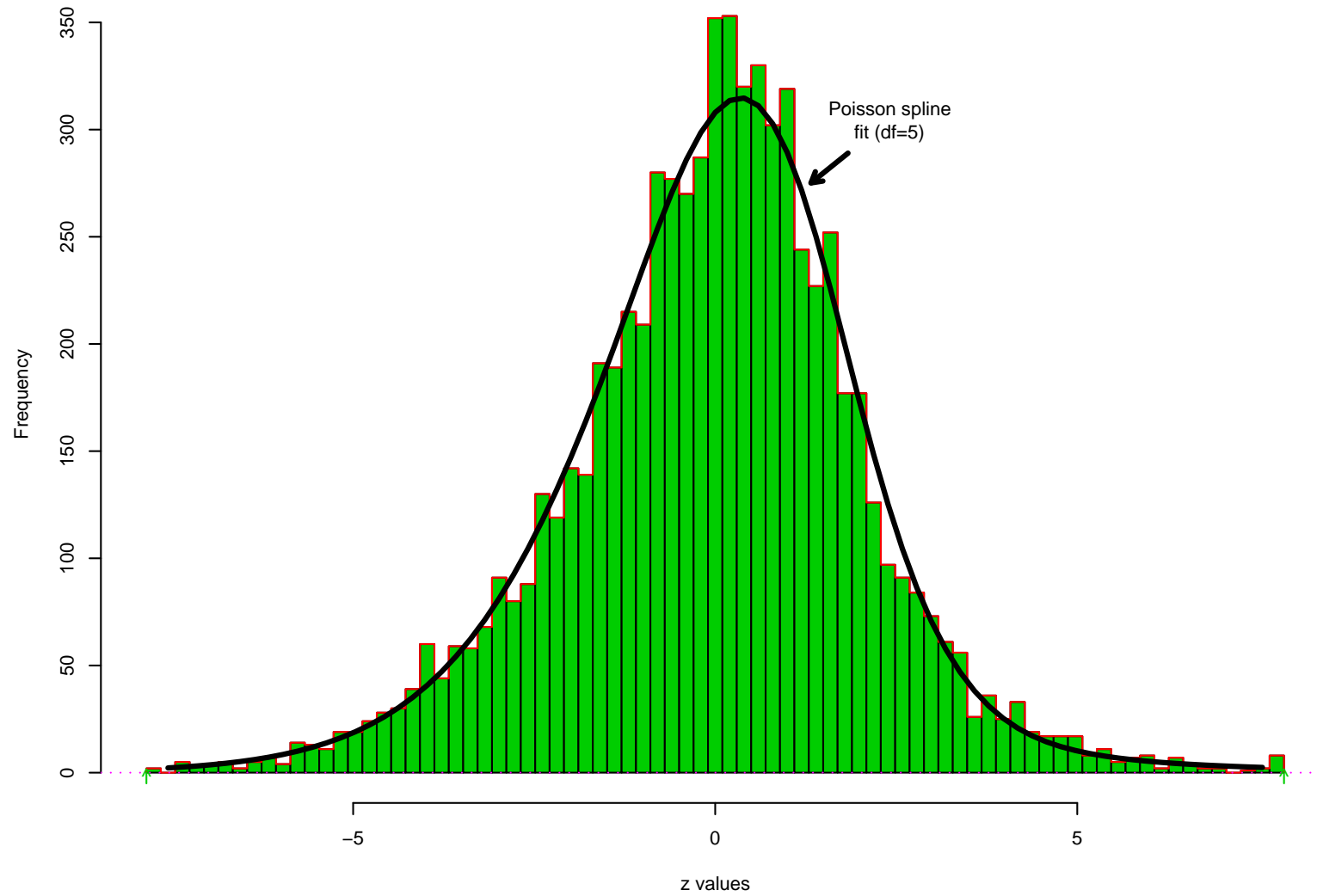
- 72 leukemia patients: $n_1 = 47$ “ALL”, $n_2 = 25$ “AML”
- $N = 7128$ genes
- Data matrix \mathbf{X} 7128×72
- \mathbf{X} has independent columns but correlated rows

rms correlation $\hat{\alpha} = .11$

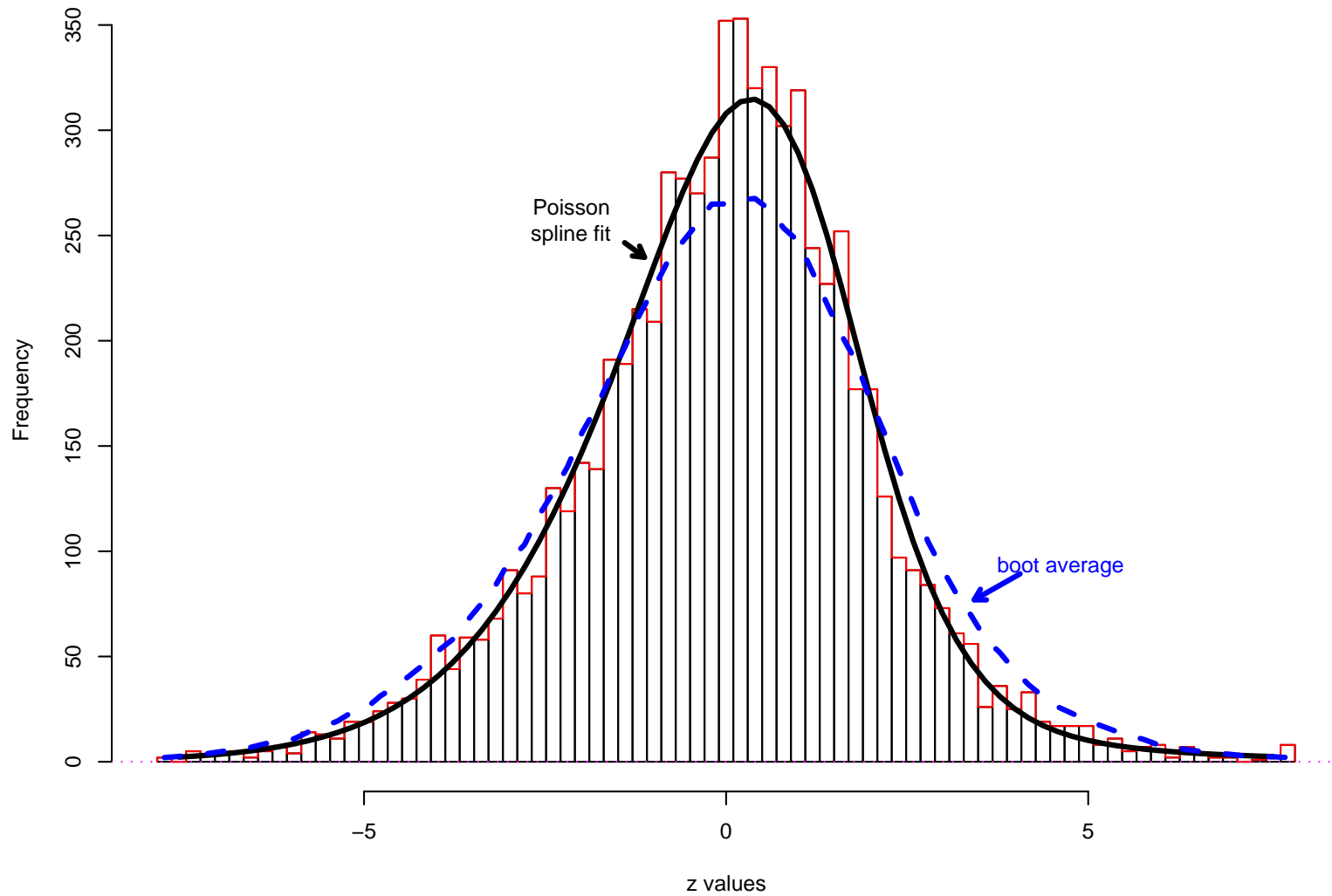
- $t_i =$ two-sample z -statistic, AML vs. ALL for gene i
- $z_i = \Phi^{-1}(F_{70}(t_i))$ [Φ, F_{70} cdfs $\mathcal{N}(0, 1), t_{70}$]

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

Leukemia data: N=7128 z-values, 47 ALL versus 25 AML patients;
RMS correlation =.11; Emp Null $\sim N(.10, 1.68^2)$



Leukemia z-value histogram and average 100 bootstrap z* histograms.
[Two-sample Nonparametric Boots: resample Columns of X]



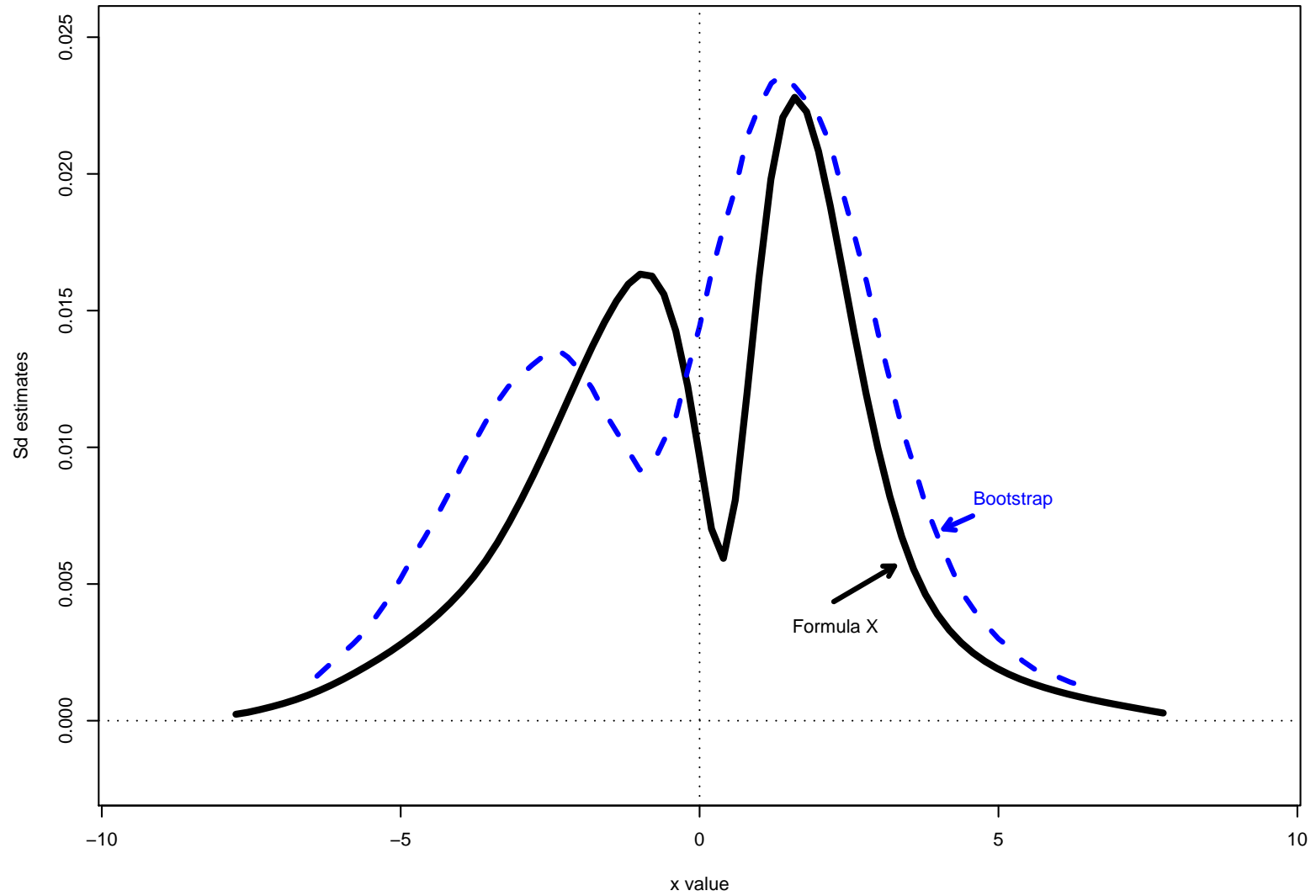
Bootstrap Dilation

- $\mathbf{x}_i = i$ th row of \mathbf{X} (n equals $72 = 47 + 25$)
- $\mathbf{x}_i \rightarrow z_i$
- $\mathbf{x}_i^* \rightarrow z_i^* \sim z_i + \mathcal{N}(0, \sigma_i^2)$
- *Bootstrap histogram* has extra component of variance:

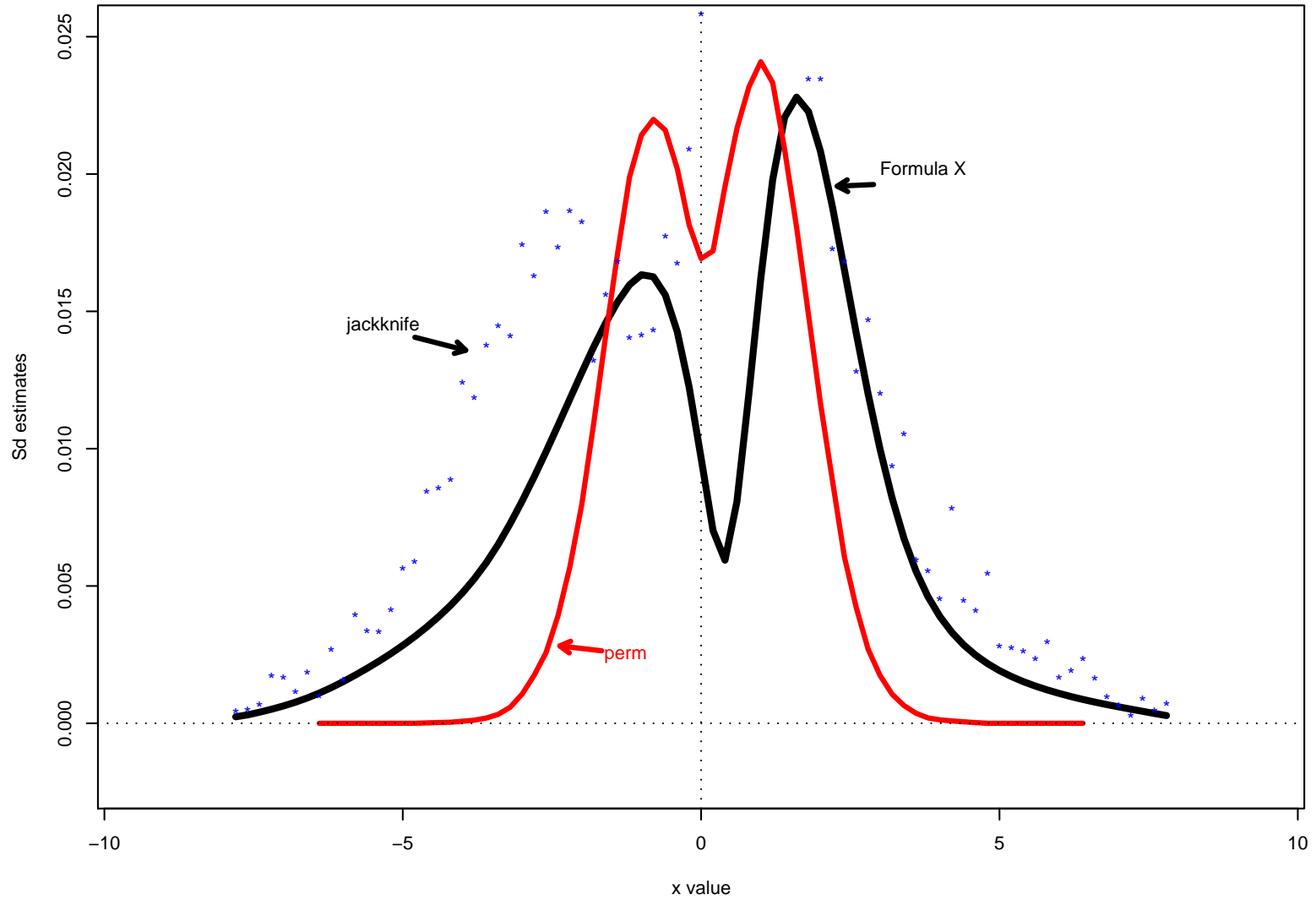
$$E_* \left\{ \sum_1^N z_i^{*2} / N \right\} = \sum_1^N z_i^2 / N + \sum_1^N \sigma_i^2 / N$$

- **Next:** Boot stdev estimates for $\hat{F}(x) = \#\{z_i \geq x\} / N$

Bootstrap Stdev for empirical cdf of Leukemia z-values,
compared with Formula X



Now permutation and jackknife ests of $sd\{\text{empirical cdf}\}$
compared with Formula X



Formula X

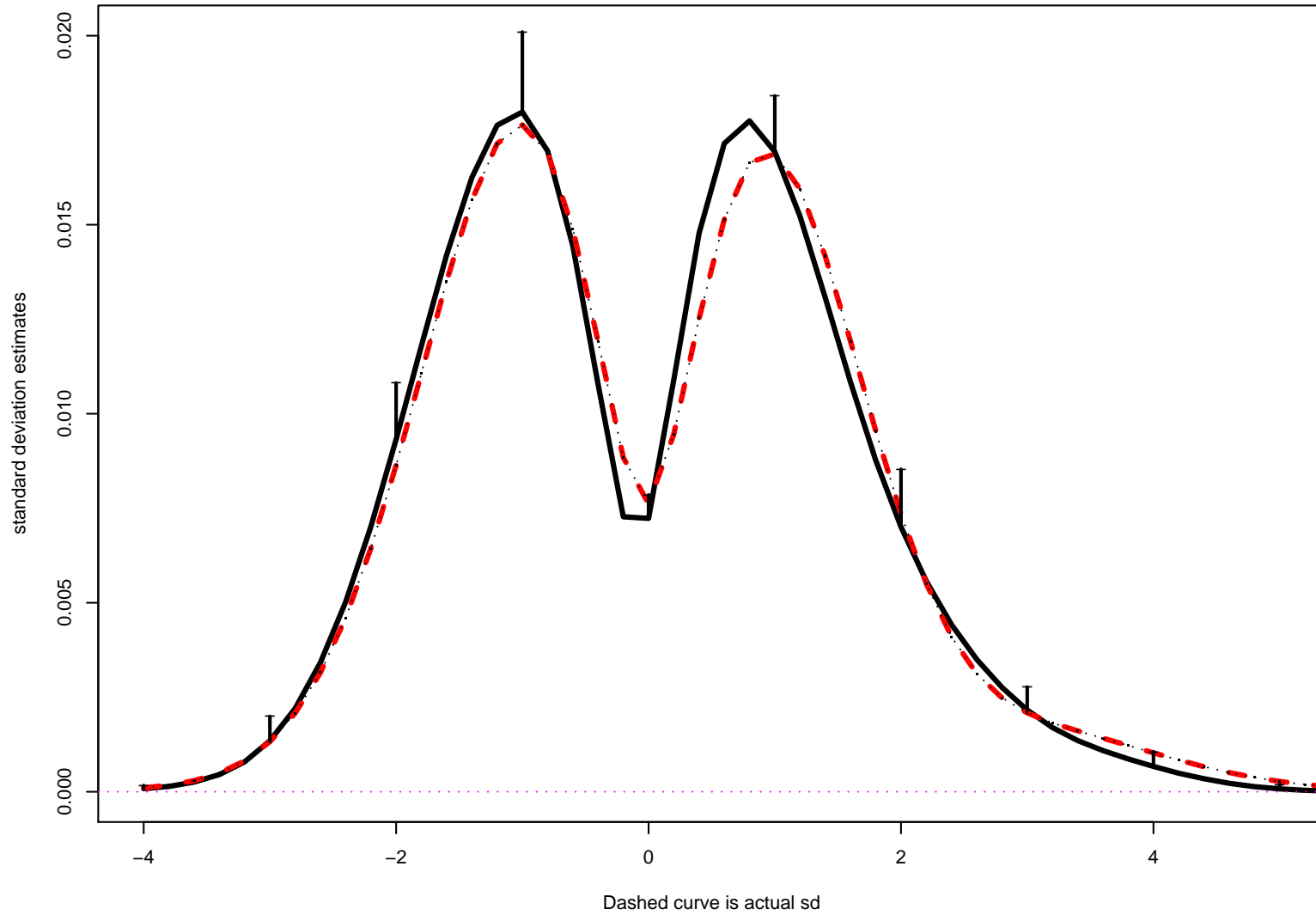
$$\text{Var} \left\{ \hat{F}(x) \right\} \doteq \underbrace{\left\{ \frac{\hat{F}(x)(1-\hat{F}(x))}{N} \right\}}_{\text{independence}} + \underbrace{\left\{ \frac{\hat{\sigma}_0^2 \hat{\alpha} \hat{f}^{(1)}(x)}{\sqrt{2}} \right\}^2}_{\text{correlation penalty}}$$

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

Formula X for Leukemia Data

$x:$	1	2	3	4	5
$\hat{F}(x)$.29	.13	.057	.025	.010
\hat{sd}	.017	.022	.010	.004	.002
\hat{sd}_0	.005	.004	.003	.002	.001

Simulation: $\text{sd}\{\hat{F}(x)\}$ from Formula X; $N=6000$, $n=20+20$, $\alpha=.10$;
Solid Curve and bars are mean and stdev of sdhat values, 100 sims



Multi-Class Normal Model

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

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

- $N_c = \# \{C_c\}$, $p_c = N_c/N$ [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

- Assume g_{cd} all equal $g(z)$

$$g(z) = \text{empirical density all } \binom{N}{2} \text{ correlations}$$

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

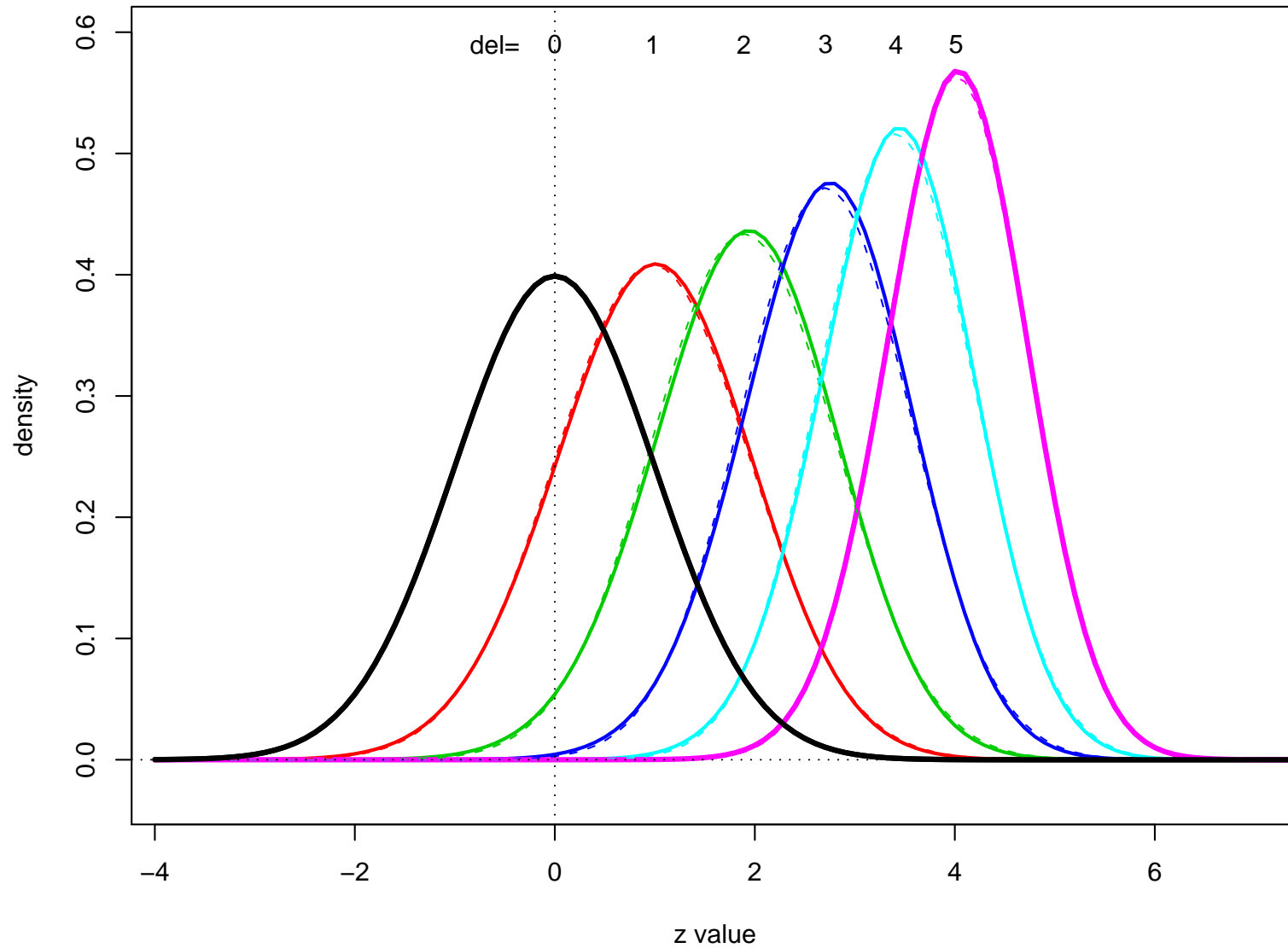
$$\sigma^2 = 1 + O\left(1/n^{1/2}\right)$$

- 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 δ , $df = \nu$]
- $H_0 : \delta = 0$
- $z = \Phi^{-1}F_\nu(t)$ [F_ν central t cdf, $df = \nu$]
so under H_0 , $z \sim \mathcal{N}(0, 1)$
- What if $\delta \neq 0$?

Densities for $z = \Phi^{-1}(F_{\nu}(t))$, $t \sim t(\text{del}, \nu=20)$, for $\text{del}=0,1,2,3,4,5$; Dotted dashed lines are matching $N(M,SD)$



The Count Vector \mathbf{y}

- Partition range \mathcal{Z} of z into K bins: $\mathcal{Z} = \bigcup_{k=1}^K \mathcal{Z}_k$
- Each bin of width “ Δ ”
- Bin centers “ x_k ”, $k = 1, 2, \dots, K$
(Leukemia histogram: $\mathcal{Z} = [-7.9, 7.9]$, $\Delta = .2$, $K = 79$)
- Counts $y_k = \#\{z_i \in \mathcal{Z}_k\}$ • $\mathbf{y} = (y_1, y_2, \dots, y_K)'$
- Count vector \mathbf{y} is discretized order statistic of z
(most statistics of interest of form $\hat{\theta} = m(\mathbf{y})$)

Mehler's Identity (Lancaster, 1958)

- $\varphi_\rho(u, v)$ = standard normal bivariate density
- **Mehler** $\lambda_\rho(u, v) = \frac{\varphi_\rho(u, v)}{\varphi(u)\varphi(v)} - 1 = \sum_{j \geq 1} \frac{\rho^j}{j!} h_j(u) h_j(v)$

where h_j is j th Hermite polynomial

- *Crucial quantity*: $\Lambda(u, v) = \int_{-1}^1 \lambda_\rho(u, v) g(\rho) d\rho$

$$= \sum_{j \geq 1} \frac{\alpha_j}{j!} h_j(u) h_j(v) \quad \text{where } \alpha_j = \int_{-1}^1 \rho^j g(\rho) d\rho$$

Exact Covariance of \mathbf{y}

- $z_i \sim \mathcal{N}(\mu_c, \sigma_c^2)$ for $z_i \in \mathcal{C}_c$
- $N_c = \#\mathcal{C}_c$, $p_c = N_c/N$

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

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

where $\pi_{ck} = \Pr_c\{z_i \in \text{bin}_k\}$, $\boldsymbol{\pi}_c = (\cdots \pi_{ck} \cdots)'$,

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

and $B_{cd}(k, l) = \pi_{ck} \pi_{dl} \Lambda \left(\frac{x_k - \mu_c}{\sigma_c}, \frac{x_l - \mu_d}{\sigma_d} \right)$.

Four Simplifications of \mathbf{cov}_1

- Drop N term
- Microarray standardization methods make $\alpha_1 \doteq 0$
- Mehler expansion: $\alpha_2 = \int_{-1}^1 \rho^2 g(\rho)$ is the lead term
- Higher terms ignorable if α_2 small

Simplified Formula (almost Formula X):

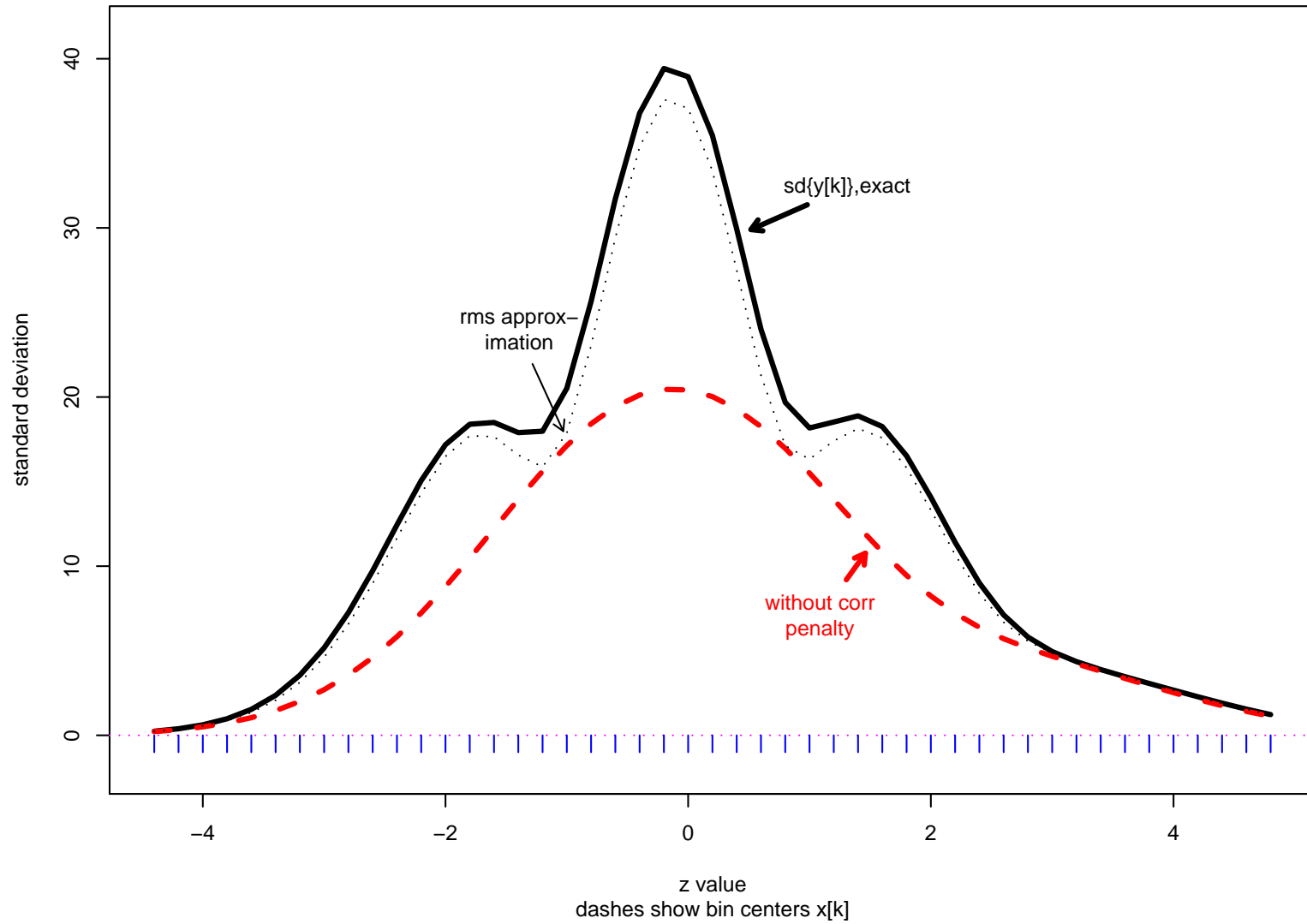
Letting $\alpha = \alpha_2^{\frac{1}{2}}$ and $\bar{\phi}_k^{(2)} = \sum_c p_c \varphi^{(2)}\left(\frac{x_{kc} - \mu_c}{\sigma_c}\right) / \sigma_c$

$$\mathbf{cov}_1 \doteq (N\Delta\alpha)^2 \bar{\phi}^{(2)} \bar{\phi}^{(2)'} / 2 \quad [\text{rms approximation}]$$

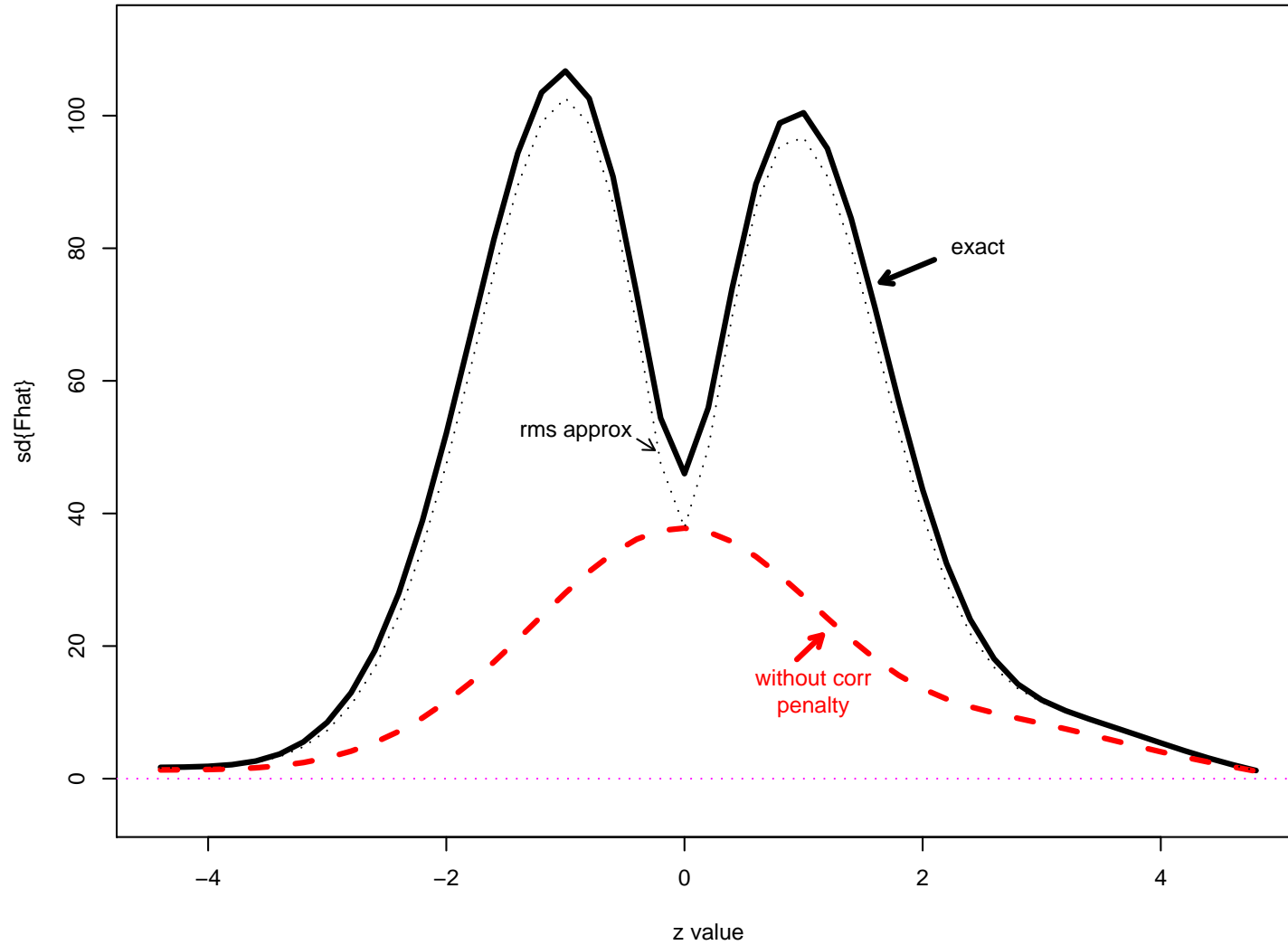
Numerical Comparison

- $N = 6000, \alpha = .1$
- 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 $\mathbf{cov}(\mathbf{y})$ & rms approximation

Compare $\text{sd}\{y[k]\}$ from exact formula (solid) with rms approx (dashed);
 $N=6000$, $\alpha=.1$, $(p_0, \mu_0, \text{sig}_0)=(.95, 0, 1)$ and $(.05, 2.5, 1)$



Same numerical example, now $sd\{\hat{F}_k\}$
[$\hat{F}_k = \sum_{l \geq k} y[l] / N$]



Estimation of RMS Correlation α

- $\hat{\rho}_{ii'}$ = empirical correlation, rows i, i' of \mathbf{X} ,
 $N \times n$ expression matrix
- $\{\hat{\rho}_{ii'}\}$ has mean and variance (m, v)
[leukemia = (.00, .19²)]

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

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

More General Accuracy Estimates

- “ Q ” q -dimensional statistic of interest: $Q = Q(\mathbf{y})$
- *Influence Function*

$$dQ = \hat{D} d\mathbf{y} \quad \left[\hat{D}_{jk} = \partial Q_j / \partial y_k \right]$$

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

Example: Accuracy of \hat{f}

- $z \rightarrow y \rightarrow \hat{f}$ by Poisson GLM
of counts y_k on polynomial (x_k)
- $Q = \log(\hat{f}) = (\dots \log f(x_k) \dots)'$
- $\hat{D} = M [M' \text{diag}(\hat{f}) M] M' / N\Delta$

with M the GLM structure matrix

Local False Discovery Rate

- $$\begin{cases} p_0 = \text{prior Pr null} \\ p_1 = \text{prior Pr non-null} \end{cases} \quad z \sim \begin{cases} f_0(z) \\ f_1(z) \end{cases}$$

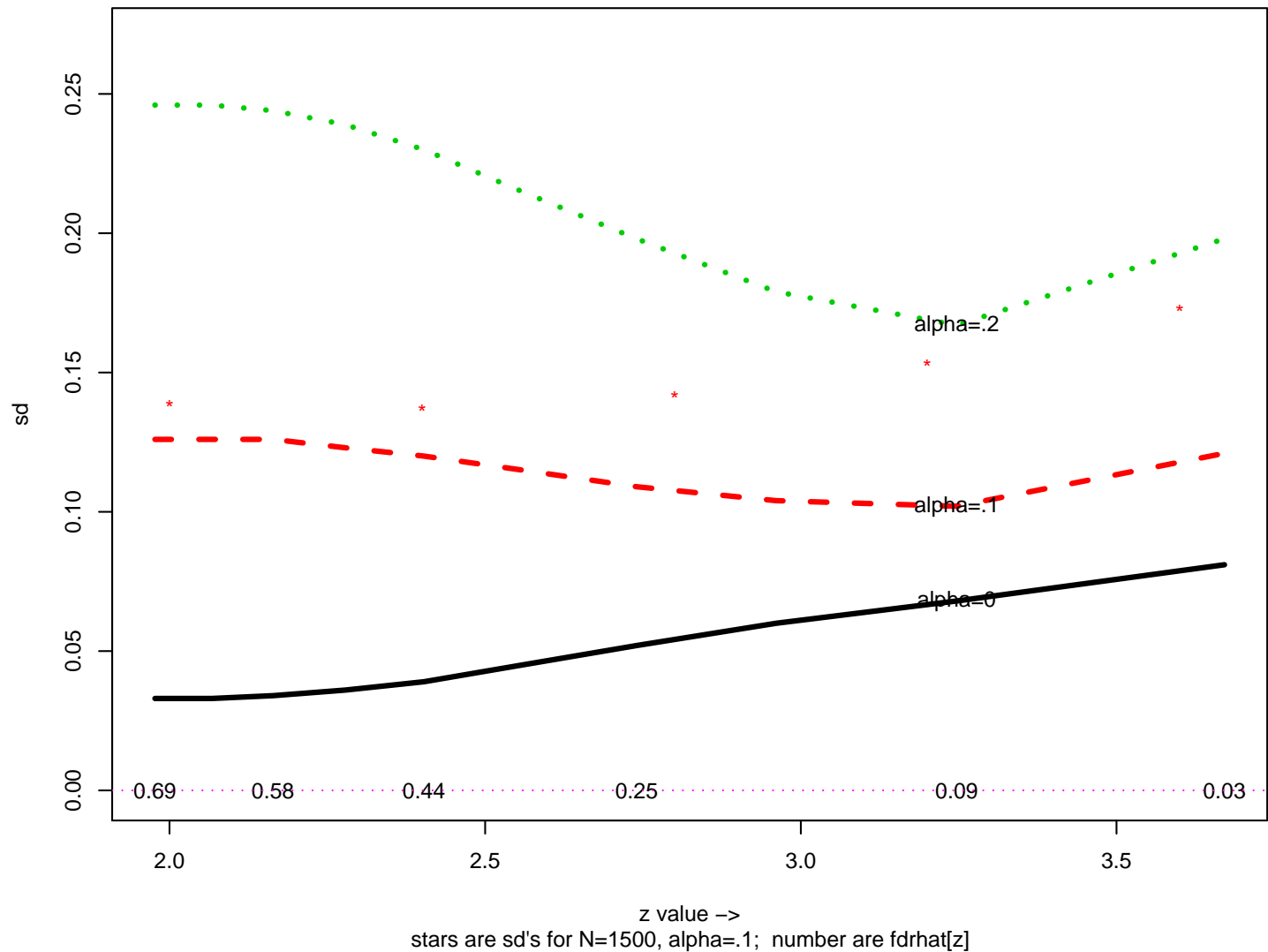
- *Mixture* $f(z) = p_0 f_0(z) + p_1 f_1(z)$

- Estimated local false discovery rate

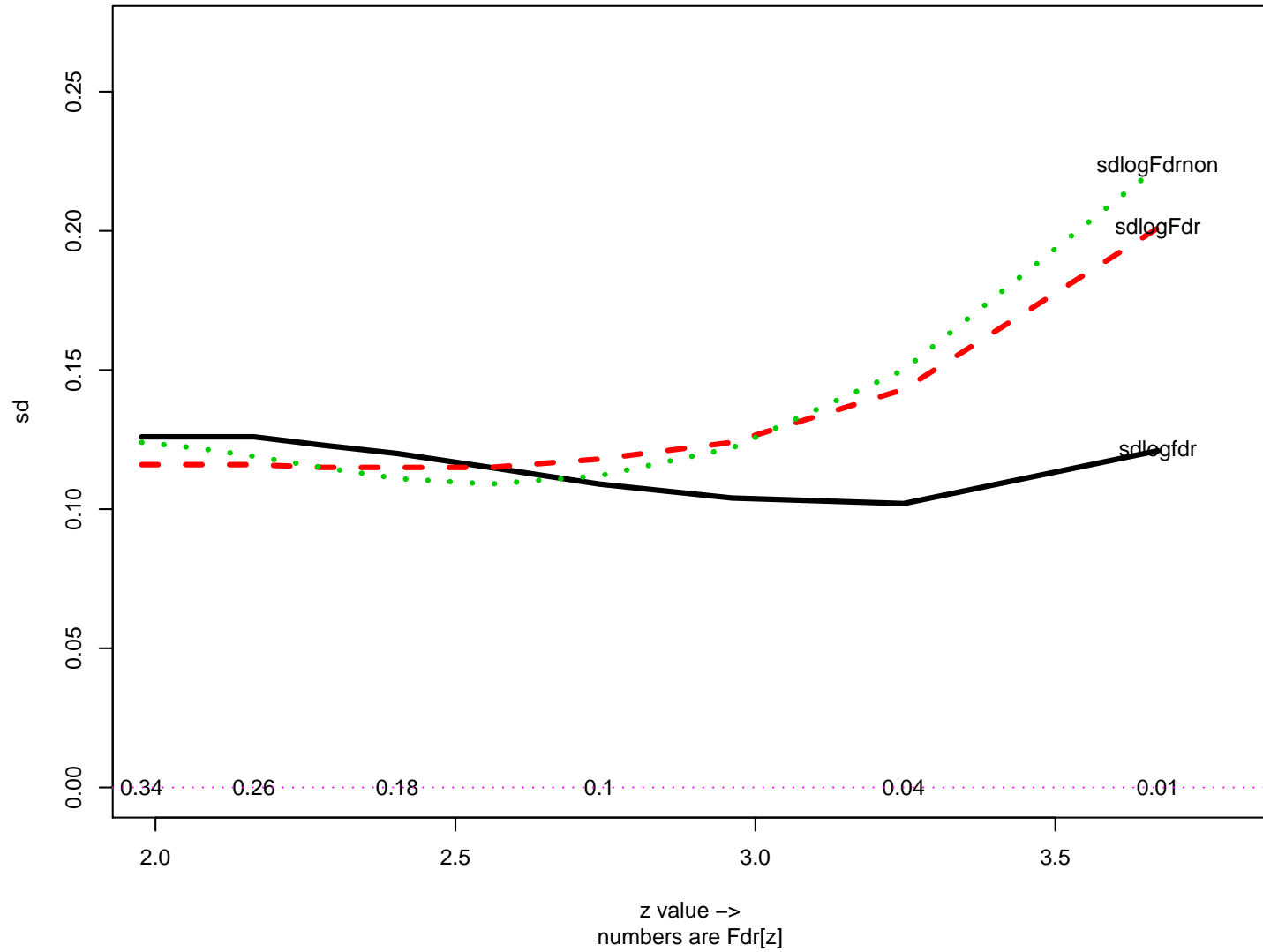
$$\widehat{\text{fdr}}(z) = \widehat{\text{Pr}}\{\text{null}|z\} = p_0 f_0(z) / \hat{f}(z)$$

- $\text{cov} \{ \log \widehat{\text{fdr}} \} \doteq \text{cov} \{ \log \hat{f} \}$

sd{log fdrhat(z)} ; N=6000, alpha=0, .1, and .2,
 (p0,mu,sig) = (.95,0,1) and (.05,2.5,1)



Now compare sd's for $\log\{\text{fdrhat}\}$ and $\log\{\text{Fdrhat}\}$,
 $\alpha=.1$



Poisson Bootstrap

- **Null Case** All $z_i \sim \mathcal{N}(0, 1)$
- Let $A \sim \mathcal{N}(0, \alpha^2)$ and $w = N\pi_0 h_2 / \sqrt{2}$
- **Hierarchical Poisson Resampling:** $u = N\pi_0 + Aw$

and

$$y_k \stackrel{\text{ind}}{\sim} \text{Poi}(u_k) \quad k = 1, 2, \dots, K$$

- i.i.d. case if $\alpha = 0$
- like rms formula if $\alpha = 0$

References

- 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.
- Efron, B. (2009). Correlated z -values and the accuracy of large-scale statistical estimates. *J. Amer. Statist. Assoc.* To appear (<http://stat.stanford.edu/~brad/papers>).
- Golub, T. R., 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, [the leukemia data].

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.