# Estimation Accuracy in Multiple Testing Problems

James Long
jplong@mdanderson.org
Rice STAT 533 / GSBS 1283

April 23, 2020

# Announcements

- HW 9: Due today at 5:00pm, email TA Scott Liang at ricestat533@gmail.com
- Lectures: Today is final class
- Take home exam
  - Send out tomorrow
  - Due April 29 at 5:00pm (if you need more time, let me know)
  - Similar structure to Exams 1 and 2
  - Same policies as Exam 2
  - Strong focus on content in Efron, last third of course
- Lecture Format
  - Slides (plots / analyses in R)
  - .pdf and .R available on course website

# Outline

Simulation

Correlation in Test Statistics

Data Examples

# Two Group Model

- Hypotheses $H_{01}, \ldots, H_{0N}$
- $\pi_0 =$ proportion of true nulls
- $\pi_1 = 1 - \pi_0 =$ proportion of true alternatives
- $y_i$ is indicator $H_{1i}$ is true
  - $y_i \sim Bernoulli(\pi_1)$
- $z_i$ (or $p_i$) drawn from distribution:

$$f_0(z) \text{ if } y_i = 0 \text{ (i.e. } H_{0i} \text{ is true)}$$
$$f_1(z) \text{ if } y_i = 1 \text{ (i.e. } H_{1i} \text{ is true)}$$

- The marginal distribution of $z_i$ is

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$$

# Outline

Simulation

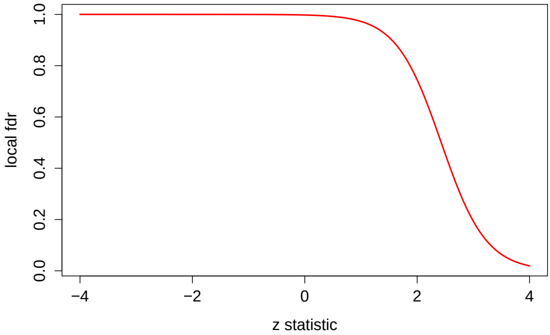
Correlation in Test Statistics

Data Examples

# Simulation

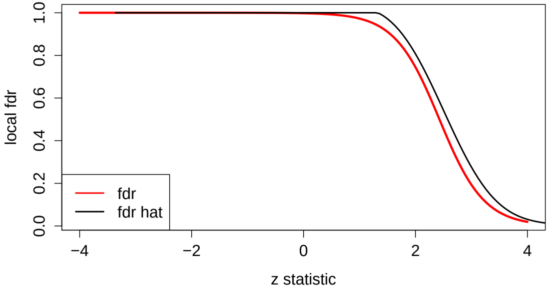
- $\pi_0 = 0.95$
- $f_0 = N(0, 1)$
- $f_1 = N(2.5, 1)$
- Then

$$\text{fdr}(z) = \frac{\pi_0 f_0(z)}{f(z)} = \frac{\pi_0 \phi(z)}{\pi_0 \phi(z) + (1 - \pi_0)\phi(z - 2.5)}$$



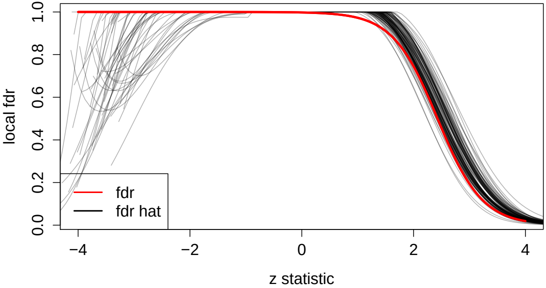
# Generate Data from Model

- Generate $N = 6000$ z-statistics from model assuming independent test statistics
- Compute $\widehat{\text{fdr}}(z)$



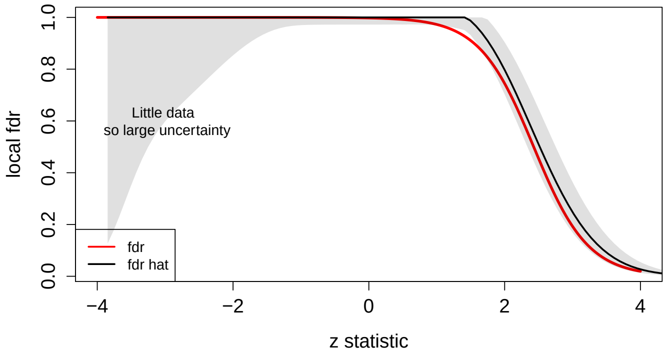

**Message:** $\widehat{\text{fdr}}(z)$ (black line) is an estimate of the true local fdr (red line).

# Run Simulation 100 Times




- Some gross errors on left side.
- But mostly care about uncertainty when $\text{fdr}(z) < 0.3$
- This is an approximation to sampling distribution of $\widehat{\text{fdr}}$ which cannot be observed.

# Standard Errors




- Black line: $\widehat{\text{fdr}}(z)$ from one simulation run
- Grey region: 95% confidence interval (pointwise)
- **Today:** Discuss how to compute these uncertainties.

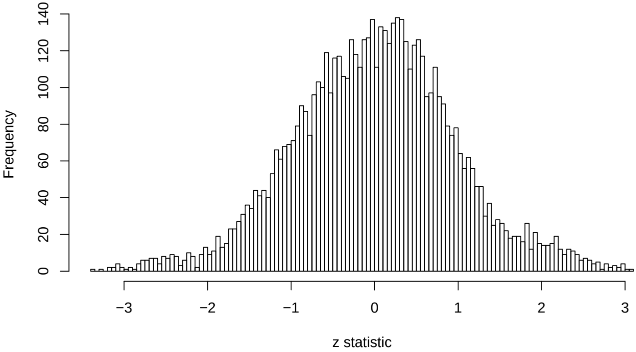
# Dependent z-statistics

- Test statistics may be dependent
  - Genes with similar function will have similar expression
  - Test statistics will be similar for these genes
- Correlated Simulation
  - Divide $z$ in $J = 60$ blocks of length $H = N/J$
  - For $h$ element in block $j$

$$z_{hj} = \frac{\gamma U_j + V_{hj}}{(1+\gamma^2)^{1/2}}$$

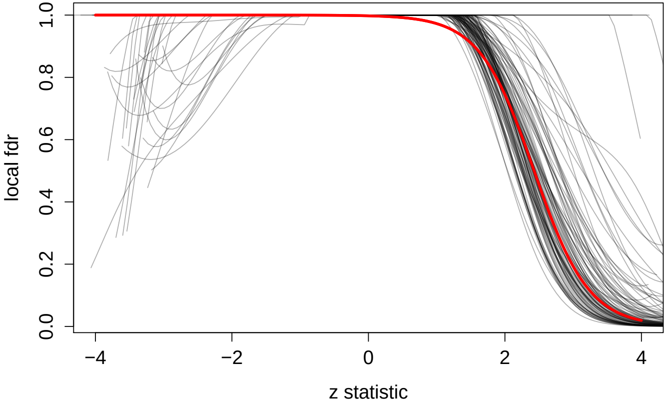
    where $U_j$ and $V_{hj}$ are $N(0,1)$ all independent
  - $\gamma$ controls degree of correlation
  - Let $\rho_{ii'}$ be correlation between $z_i$ and $z_{i'}$
  - $\gamma$ chosen such that $\alpha = \sqrt{M^{-1} \sum \rho_{ii'}^2} = 0.1$ where $M = \binom{N}{2}$

# Histogram of Test Statistics




Modes may appear near $U_j$.

# Local fdr with Correlated z–statistics




- Larger variance when using correlated test statistics.
- Need methodology which accounts for this.

# Outline

# Correlated Data Matrix

- $X_{ij}$ is data matrix
  - $i = 1, \ldots, N$ indexes hypotheses
  - $j = 1, \ldots, n$ indexes cases
- $X_{\cdot j} \sim_{iid} N(\mu, \Sigma)$
  - $\mu \in \mathbb{R}^N$ is mean vector
  - $\Sigma$ is $N \times N$ covariance
  - $\Sigma_{ii'}$ large says genes $i$ and $i'$ are highly correlated
- Assumes independence across columns (usually people / tissue samples)
- $t_1, \ldots, t_N$ are test–statistics from t-tests
- How are $t_1, \ldots, t_N$ correlated?

**Theorem 8.5:**

$$cor(t_i, t_{i'}) = \rho_{ii'} + O(n^{-1})$$

where $\rho_{ii'}$ is correlation between genes $i$ and $i'$ (computed from $\Sigma$)

# Root Mean Square Correlation

- Test statistics $z_1, \ldots, z_N$
- $\rho_{ii'} = Cor(z_i, z_{i'})$
- Define the root mean square correlation:

$$\alpha = \sqrt{\frac{1}{\binom{N}{2}} \sum_{i<i'} \rho_{ii'}^2}$$

- Measure of overall correlation in test statistics
- Based on theorem, can estimate from data matrix X

$$\widehat{\alpha} = \sqrt{\frac{1}{\binom{N}{2}} \sum_{i<i'} Cor(X_{i\cdot}, X_{i'\cdot})^2}$$

# Approach to Estimating local fdr Uncertainty

- Estimate uncertainty in

$$\widehat{\mathsf{fdr}}(z) = \frac{\pi_0 f_0(z)}{\widehat{f}(z)}$$

- For simplicity (sanity), we assume $\pi_0$ and $f_0$ are known
- Consider the log local fdr

$$\widehat{\mathsf{lfdr}}(z) = \log(\pi_0) + \log(f_0(z)) - \log(\widehat{f}(z))$$

- **Estimate** $s.d.(\log(\widehat{f}(z)))$
- Under asymptotic normality

$$\left[e^{\widehat{\mathsf{lfdr}}(z) - 2s.d.(\log(\widehat{f}(z)))}, e^{\widehat{\mathsf{lfdr}}(z) + 2s.d.(\log(\widehat{f}(z)))}\right]$$

is a 95% CI for $\widehat{\mathsf{fdr}}(z)$

# Approach to Estimating local fdr Uncertainty

- $\widehat{f}(z)$ is Efron's Poisson regression estimate based on binned data
- Let $y = (y_1, \ldots, y_K)$ be counts in equal sized bins spread across test statistic domain
- $Cov(y) = Cov_0(y) + Cov_1(y)$ (Lemma 7.1 in Efron)
  - $Cov_0$ is multinomial covariance based on independent $z_i$
  - $Cov_1$ is covariance resulting from dependence in $z_i$.
    - Depends on $Cor(z_i, z_{i'}) = \rho_{ii'}$
    - Not directly estimable from $z_i$ and $z_{i'}$ because only have one realization. But can estimate with $Cor(X_{i\cdot}, X_{i'\cdot})$
    - Size of $Cov_1$ depends on root mean square correlation $\alpha$ (7.38 in Efron)
- Express $\widehat{f}$ as a smooth functional of $y$.
- Use delta method to determine $s.d.(\log(\widehat{f}(z)))$

$$\widehat{Cov}(\log(\widehat{f})) = \left(\frac{d\log\widehat{f}}{dy}\right)\widehat{Cov}(y)\left(\frac{d\log\widehat{f}}{dy}\right)$$

All terms $K \times K$ matrices.

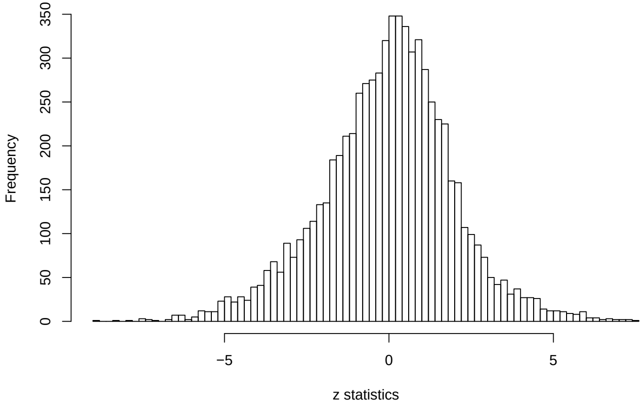
# Outline

Simulation

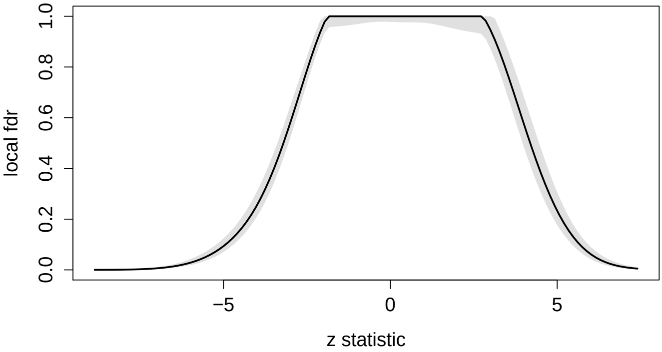
Correlation in Test Statistics

Data Examples

# Leukemia Data

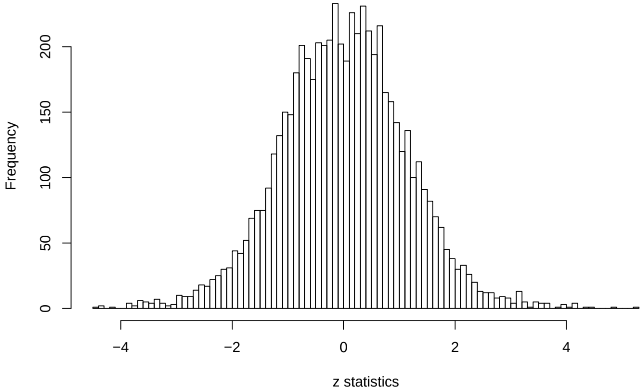


# Leukemia Data local fdr with Uncertainties

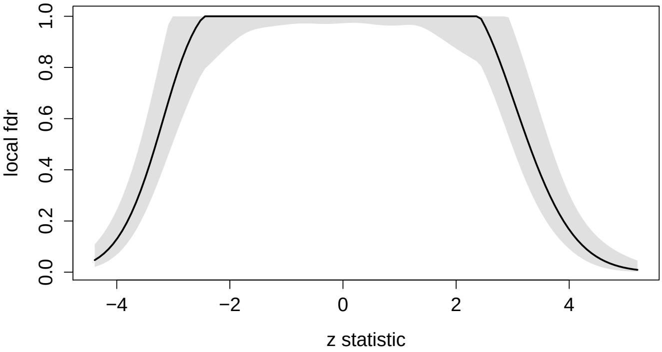



Black line is $\widehat{\text{fdr}}$. Grey region is 95% confidence set.

# Prostate Data



# Prostate Data local fdr with Uncertainties




Black line is $\widehat{\text{fdr}}$. Grey region is 95% confidence set.

# Coding

`locfdr` package on CRAN does these calculations.

```
> a <- locfdr(zstats,nulltype=1,
+             pct0=c(0.25,0.75),plot=0)
> plot(a$mat[,1],a$mat[,2],type='l',
+      lwd=2,cex.lab=1.3,cex.axis=1.3,
+      xlab="z statistic",ylab="local fdr")
> ldfr_up <- pmin(exp(log(a$mat[,2]) +
+                       2*a$mat[,10]),1)
> ldfr_low <- pmin(exp(log(a$mat[,2]) -
+                        2*a$mat[,10]),1)
> polygon(c(a$mat[,1],rev(a$mat[,1])),
+         c(ldfr_low,rev(ldfr_up)),
+         col="#00000020",border=NA)
```

# Closing

- Reminders:
  - HW9 due at 5:00 today
  - Exam 3 sent out tomorrow
- Thank you for sticking with course over tough semester.
- I wish you all best of luck going forward:
  - Stay safe.
  - Follow health guidelines.