Multiple Testing

Hoang Tran

Department of Statistics, Florida State University

Large-Scale Testing

Examples:

- Microarray data: testing differences in gene expression between two traits/conditions
- Microbiome data: which bacteria are **differentially expressed** between two traits/conditions?

Prostate cancer data: $n = 102$ subjects (52 cases and 50 controls) and $N = 6033$ genes. How do we test for differences in gene expression?

Large-Scale Testing

1. For the jth gene: compute the two-sample t statistic comparing gene expression between cases and controls (t_j).
2. Test H_{0j}: gene j's expression levels are the same between the two groups at significance level α.
3. Repeat for all $N = 6033$ genes.

What's the problem?

Multiple Comparisons

- Running 100 separate hypothesis tests at $\alpha = 0.05$ will produce about 5 "significant" results even if each case is actually null.
- Examples:
  - Efficacy of a drug in terms of reduction of disease symptoms.
  - Two methods of teaching writing are used on students. Students in the groups are compared in terms of grammar, spelling, etc.
- Increased likelihood of **type I errors**

Bonferroni Correction

- Family-Wise Error Rate: the probability of rejecting **any** true null hypothesis.
- Test each individual hypothesis at α/N.
- Let J_0 be the indices of the true H_{0j} with $|J_0| = N_0$.

$$\begin{aligned}\text{FWER} = P\left\{\bigcup_{J_0}\left(p_j \leq \frac{\alpha}{N}\right)\right\} &\leq \sum_{J_0} P\left\{p_j \leq \frac{\alpha}{N}\right\} \\ &= N_0\frac{\alpha}{N} \leq N\frac{\alpha}{N} = \alpha\end{aligned}$$

- The **Bonferroni Correction** ensures FWER $\leq \alpha$.

Bonferroni Correction

- No requirement of independence of p_i's, but it's perhaps too *conservative*.
- $N = 6033$ and $\alpha = 0.05$: only reject when $p_j \leq 8.3 \times 10^{-6}$!
- We want to control type I errors, but we also want to find interesting/significant genes.

Holm's Procedure

- Order the p-values:

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}$$

  with $H_{0(j)}$ the null hypotheses.
- Let j_0 be the smallest index j such that

$$p_{(j)} > \alpha/(N - j + 1)$$

- Reject all $H_{0(j)}$ for $j < j_0$ and accept all with $j \geq j_0$.

Satisfies FWER $\leq \alpha$ but not as conservative as Bonferroni (more **rejections**).

Stepdown Procedures

- Holm's procedure: look at most "significant" test first and continue rejecting hypotheses if p-values are small.
- An improvement: incorporate the dependence structure of individual tests.

Generic Stepdown Method

- $K \subset \{1, \ldots, N\}$, H_K: *intersection* hypothesis that all H_{0j} with $j \in K$ are true.
- T_j is the jth test statistic. $T_{(1)} \leq \cdots \leq T_{(N)}$ and $H_{0(j)}$.

Generic Stepdown Method

1. Let $K_1 = \{1, \ldots, N\}$. If $T_{(N)} \leq \hat{c}_{K_1}(1-\alpha)$ then accept all hypotheses and stop; otherwise, reject $H_{0(N)}$ and continue.
2. Let K_2 be the indices of the hypotheses not previously rejected. If $T_{(N-1)} \leq \hat{c}_{K_2}(1-\alpha)$ then accept all hypotheses and stop; otherwise, reject $H_{0(N-1)}$ and continue.

Generic Stepdown Method

- How do we find the critical values $\hat{c}_K(1-\alpha)$?
- Under certain conditions, FWER $\leq \alpha$ (Lehmann and Romano 2005).

$$\text{FWER} \leq P(\max_{j \in J_0} T_j > \hat{c}_{J_0}(1-\alpha)) \leq \alpha$$

- Critical values are the $(1-\alpha)$ quantile of $\max_{j \in K} T_j$ under H_K.

Not as **conservative** as Holm's procedure in general.

The Hypothesis of Homogeneity

Consider testing

$$H_{i,j} : \mu_i = \mu_j, \quad i < j$$

for all $\binom{N}{2}$ pairs $i < j$.

- $\{H_{1,2}, H_{2,3}\}$ cannot be the set of all true hypotheses
- Previous methods allow acceptance of $H_{1,2} : \mu_1 = \mu_2$ and $H_{2,3}$ but rejection of $H_{1,3}$

A Holm type approach

Setup ($N = 6$):

- Normal random variables with common variance σ^2.
- $\bar{X}_{(1)} \leq \cdots \leq \bar{X}_{(N)}$ and $\mu_{(j)}$.
- $\hat{p}_{(i),(j)}$: the p-value for testing $\mu_{(i)} = \mu_{(j)}$.

Procedure:

1. If $\hat{p}_{(6),(1)} \geq \alpha/\binom{N}{2}$, accept all hypotheses and terminate. Otherwise, reject $\mu_{(1)} = \mu_{(6)}$ and continue.
2. Test the largest of $|\bar{X}_{(6)} - \bar{X}_{(2)}|$ and $|\bar{X}_{(5)} - \bar{X}_{(1)}|$ by comparing $\hat{p}_{(6),(2)}$ or $\hat{p}_{(5),(1)}$ with $\alpha/(\binom{N}{2} - 1)$.

FWER is controlled at α.

An improvement

- Suppose we are at step 2 ($\mu_{(1)} = \mu_{(6)}$ has been rejected).
- $\mu_{(1)} = \mu_{(2)}$ or $\mu_{(2)} = \mu_{(6)}$ must be false.
- Possible true hypotheses: $\binom{6}{2} - 5 = 10 < \binom{6}{2} - 1 = 14$.
- No violation of FWER and **more** rejections.

An improvement

Table 9.1.
Possible Number of True Hypotheses

s	Total # of Hypotheses $H_{i,j}$	Possible Number of True Hypotheses
3	3	0, 1, 3
4	6	0-3, 6
5	10	0-4, 6, 10
6	15	0-4, 6, 7, 10, 15
7	21	0-7, 9, 10, 11, 15, 21
8	28	0-13, 15, 16, 21, 28
9	36	0-13, 15, 16, 18, 21, 22, 28, 36
10	45	0-18, 20, 21, 22, 24, 28, 29, 36, 45

FWER Summary

- **Family-Wise Error Rate**: the probability of rejecting any true null hypothesis.
- Bonferroni, Holm, and Generic Stepdown method all control FWER
  - Holm and Generic Stepdown method have at *least* as much power as Bonferroni
- We can exploit the structure of pairwise tests to **improve** the Holm procedure's power for these situations

False-Discovery Rates

- FWER probably too conservative for very large N, such as $N \geq 20$.
  - For N in the thousands/millions, the issue is exacerbated.
- A more **liberal** criterion: False-Discovery Rates.

False-Discovery Rates

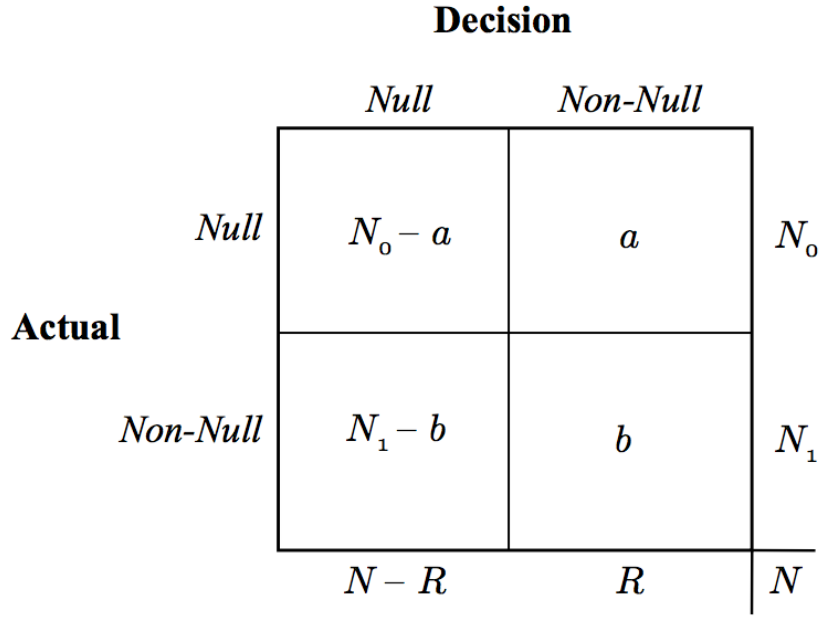


Figure 4.1 A decision rule $\mathcal{D}$ has rejected R out of N null hypotheses (4.1); a of these decisions were incorrect, i.e., they were "false discoveries", while b of them were "true discoveries." The false discovery proportion Fdp equals a/R.

False-Discovery Rates

- The number of false discoveries (a) is unobservable.
- We want to minimize $\text{Fdp} = a/R$.
- Define $\text{FDR}(\mathcal{D}) = E(\text{Fdp}(\mathcal{D}))$.
- We can't observe Fdp but we can *control* FDR.
- Decision rule $\mathcal{D}$ controls FDR at $q \in (0, 1)$ if

$$\text{FDR}(\mathcal{D}) \leq q$$

Benjamini-Hochberg Procedure for FDR Control

1. For given q, let $j_{\max}$ be the largest j such that

$$p_{(j)} \leq \frac{j}{N} q$$

2. Let $\mathcal{D}_q$ be the rule that rejects $H_{0(j)}$ for all $j \leq j_{\max}$.

If p-values are independent:

$$\mathrm{FDR}(\mathcal{D}_q) = \frac{N_0}{N} q \leq q.$$

FDR is more *generous* than FWER.

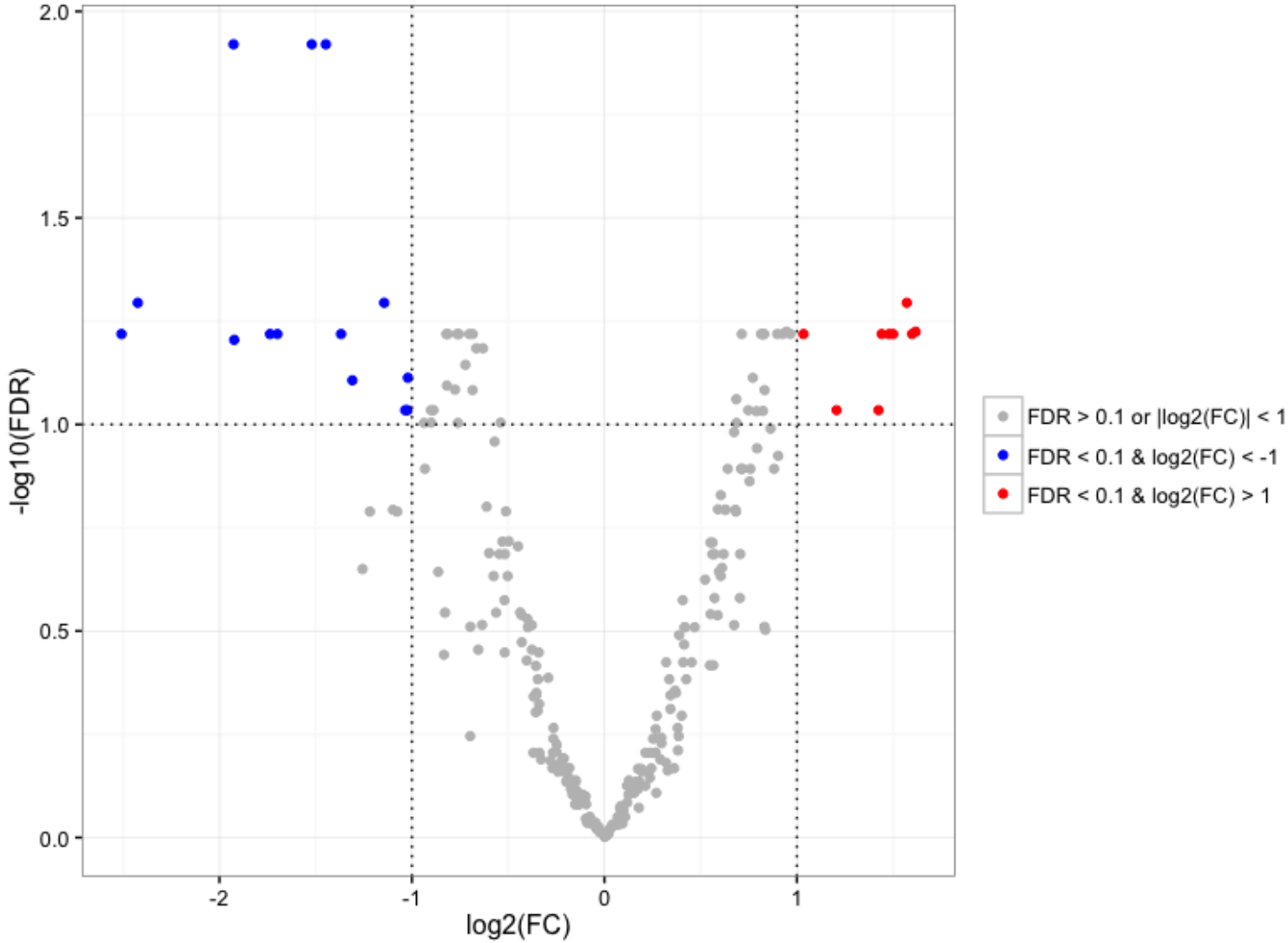
Benjamini-Hochberg q-values

- p.adjust in R computes FDR (q-value).
- In practice, $q < 0.10$ is "significant".
- Example: $q = 0.05$ means we expect 5% of significant tests to result in false positives.

Microbiome Example for FDR

- Goal: study the association of the microbiome with asthma exacerbations.
- $n = 3122$ samples and $N = 268$ taxa.
- Question: which taxa are **differentially expressed** between samples with/without asthma exacerbations?
- 498 exacerbators, 2624 non-exacerbators.
- These are (possibly overdispersed) **count data** so we fit negative binomial regressions for each taxa.

Microbiome Example for FDR



Bayesian Interpretation of FDR

- Each of the N cases is null with prior probability π_0 or non-null with probability $\pi_1 = 1 - \pi_0$.
- Each z statistic has density $f_0(z)$ if null (i.e. $N(0,1)$), $f_1(z)$ if non-null (unknown).
- $F_0(z)$ and $F_1(z)$ are cdf's with "survival curves" $S_0(z) = 1 - F_0(z)$ and $S_1(z) = 1 - F_1(z)$.
- Define $S(z) = \pi_0 S_0(z) + \pi_1 S_1(z)$ and $f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$.

Bayesian Interpretation of FDR

- Suppose $z_j > z_0 = 3$. Then

$$\text{Fdr}(z_0) \equiv P(\text{case } j \text{ is null} | z_j \geq z_0) = \pi_0 S_0(z_0)/S(z_0)$$

- $S_0(z_0)$ usually known (i.e. $1 - \Phi(z_0)$)
- $\hat{S}(z_0) = \#\{z_j \geq z_0\}/N$
- Empirical Bayes estimate:

$$\widehat{\text{Fdr}}(z_0) = \pi_0 S_0(z_0)/\hat{S}(z_0)$$

Bayesian Interpretation of FDR

- $p_{(j)} \leq (j/N)q$ from BH procedure beomes

$$S_0(z_{(j)}) \leq \hat{S}(z_{(j)})q$$

- So then $\widehat{\text{Fdr}}(z_0) \leq \pi_0 q$
- **BH rejects cases for which the empirical Bayes posterior probability of nullness is too small.**

A Note about False-Negative Rates

- The false negative proportion:

$$\mathrm{Fnp} = (N_1 - b)/(N - R)$$

- The expectation of Fnp is a measure of Type II error.
- Let $\mathcal{A}$ be the region in which a null hypothesis is accepted. Then $1 - \widehat{\mathrm{Fdr}}(\mathcal{A})$ estimates the Bayesian false negative rate.

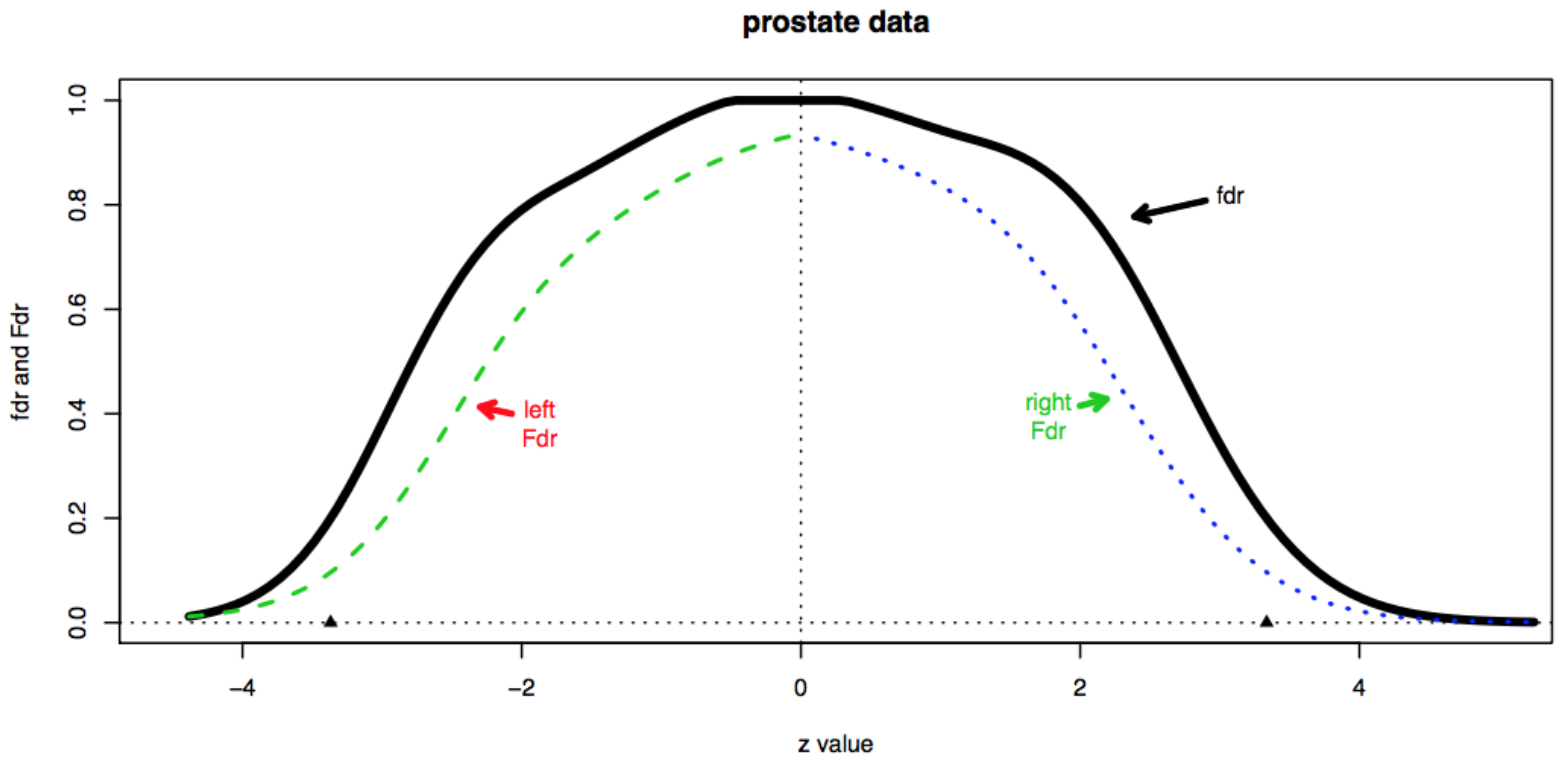
Local FDR

- Instead of a tail-area probability, what about $z_j = z_0$?
- **Local** FDR:

$$\text{fdr}(z_0) = P(\text{case } j \text{ is null} | z_j = z_0) = \pi_0 f_0(z_0)/f(z_0)$$

- π_0 is unknown but can be estimated (Efron 2010) or set to 1 (most cases are null)
- $f(z)$ is unknown but can be estimated

Local FDR


prostate data
fdr and Fdr
1.0
0.8
0.6
0.4
0.2
0.0
−4
−2
0
2
4
z value
fdr
left
Fdr
right
Fdr

Local FDR

- Conventionally "interesting" threshold: $\widehat{\text{fdr}}(z) \leq 0.2$.
- Local and tail-area FDR:

$$\text{Fdr}(z_0) = E[\text{fdr}(z)|z \geq z_0]$$

- Often $\text{Fdr}(z_0) < \text{fdr}(z_0)$.

Local FDR

Computing $\hat{f}(z)$:

- A fourth-degree log polynomial Poisson regression fit to the histogram of z-values.
- See next figure.

Local FDR

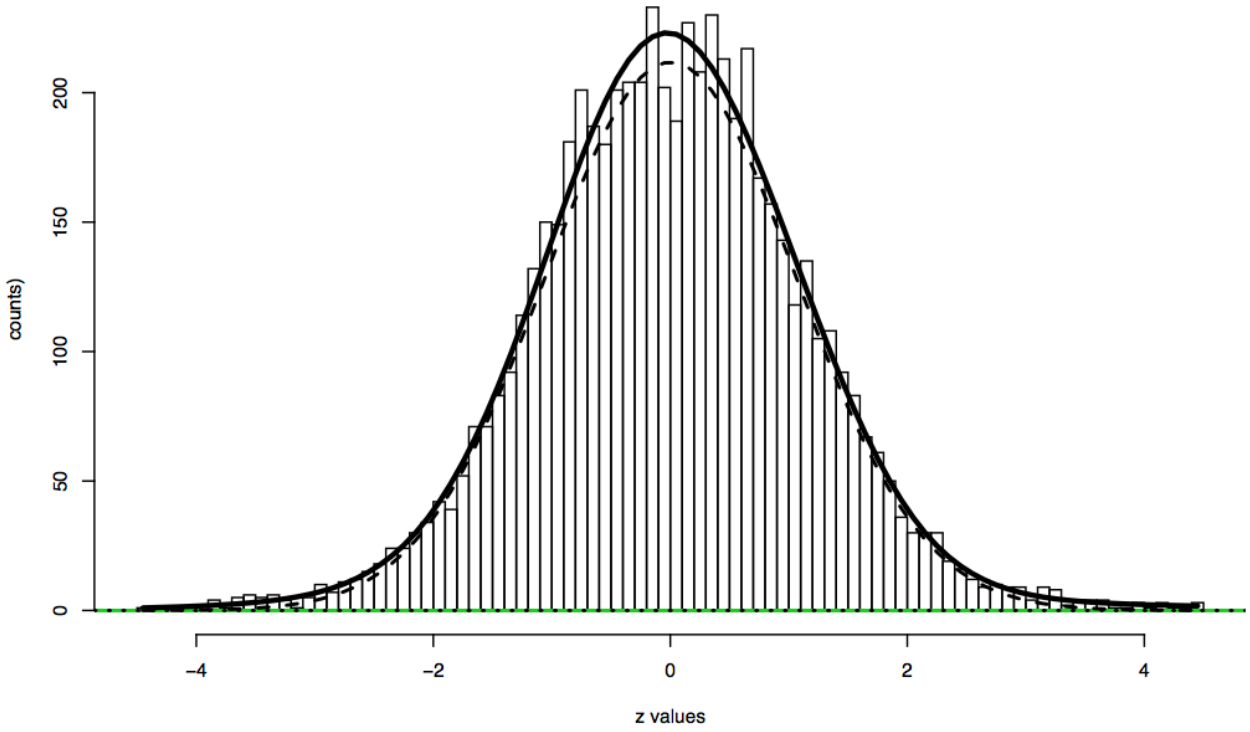


Figure 5.1a z-value histogram for the prostate data, $N = 6033$, Section 2.1. Heavy curve is estimate $\hat{f}(z)$ for mixture density $f(z)$, scaled to match histogram area. Dashed curve is scaled estimate $\hat{\pi}_0 f_0(z)$ where f_0 is the standard normal density (2.9).

Choosing the Null Distribution

- In large-scale testing we can examine hundreds/thousands/millions of z-values.
- The chosen null distribution is inappropriate.
- What if we empirically determine the null distribution?
- Use the R package locfdr (also computes local FDR).

Empirical Null Distribution

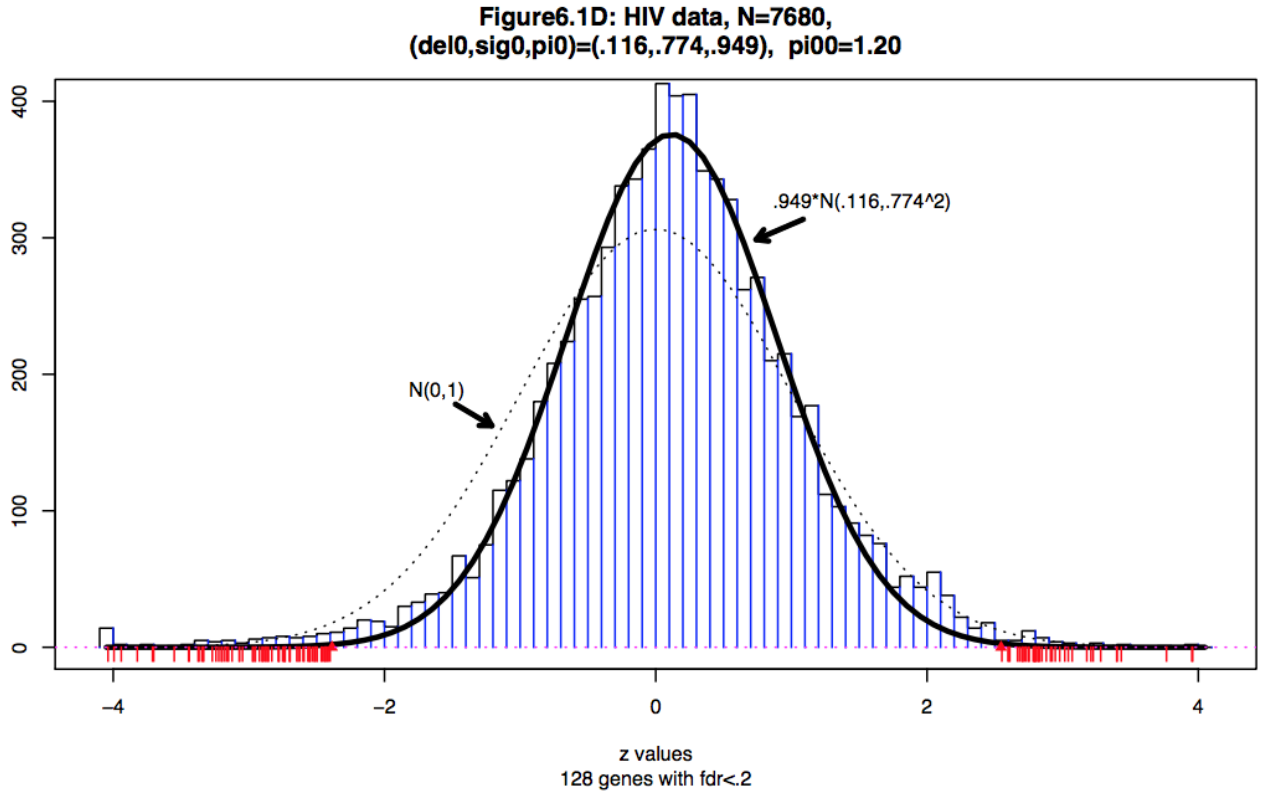


Figure 6.1d *z*-value histogram for HIV data as in Figure 6.1a. Notice reduced scale on x-axis. In this case the theoretical null is too wide.

FDR Summary

- FDR: a more *liberal* criterion than FWER.
- Many practitioners prefer to control FDR; in gene studies it is often more important to discover *interesting* genes.
- The BH procedure for FDR rejects cases for which the empirical Bayes posterior probability of nullness is too small.
- **Local** FDR: investigate more than just the tail-area probability.