Multiple Testing: Motivation and FWER

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

March 26, 2020

Outline

Outline

Announcements

- Midterm 2: Due March 31 at 5:00pm, email solutions to me
- HW 6: Due March 31 at 5:00pm, email TA Scott Liang at ricestat533@gmail.com
- Today's Lecture
  - Slides
  - Plots produced by R code
  - Slides + R code available on course website
- Lecture Structure
  - Microphones are muted when you enter the class.
  - But please ask questions, remember to unmute / mute
  - Let me know about audio issues
  - You are welcome to try to communicate with other zoom features, although I am somewhat a beginner

Outline for Remainder of Course

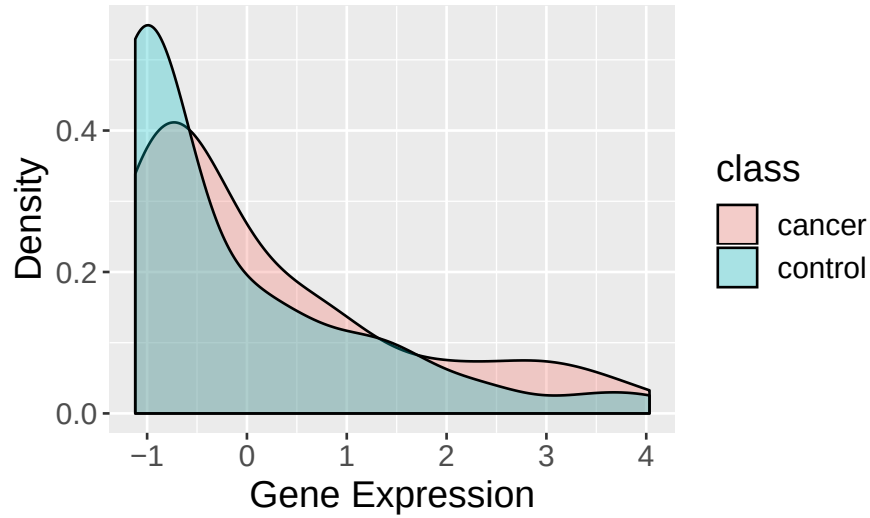
- Textbook: Efron "Large Scale Inference"
  - Available free online, see course website
- Cover parts of chapters 2–5
  - Multiple testing, family wise error rate
  - False discovery rate, local FDR
  - Empirical Bayesian Methods for testing
- Homeworks
  - Questions from Efron and some I write
  - Posted on course website
  - Solutions emailed to TA Scott Liang at ricestat533@gmail.com

Outline

Hypothesis Testing Review

Motivating Example:

- Observe expression of gene for cancer and healthy patients
- Question: Is this gene differentially expressed across 2 groups, e.g. different mean expression in cancer and healthy patients



Mathematical Notation:

- $X_1, \ldots, X_{n_1} \sim N(\mu_1, \sigma^2)$ (healthy controls)
- $X_{n_1+1}, \ldots, X_{n_1+n_2} \sim N(\mu_2, \sigma^2)$ (cancer patients)
- $n = n_1 + n_2$

$$H_0 : \mu_1 = \mu_2 \text{ no difference}$$

$$H_1 : \mu_1 \neq \mu_2 \text{ gene is differentially expressed}$$

Note: Model not particularly accurate, see plot.

Hypothesis Test

One solution: Two sample equal variance t-test

$$T = \frac{\overbrace{\frac{1}{n_2}\sum_{j=n_1+1}^{n} X_j}^{\equiv \bar{X}_2} - \overbrace{\frac{1}{n_1}\sum_{j=1}^{n_1} X_j}^{\equiv \bar{X}_1}}{s}$$

where s is the standard error of the numerator

$$s^2 = \frac{\sum_{j=1}^{n_1}(X_j - \bar{X}_1)^2 + \sum_{j=n_1+1}^{n}(X_j - \bar{X}_2)^2}{n-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

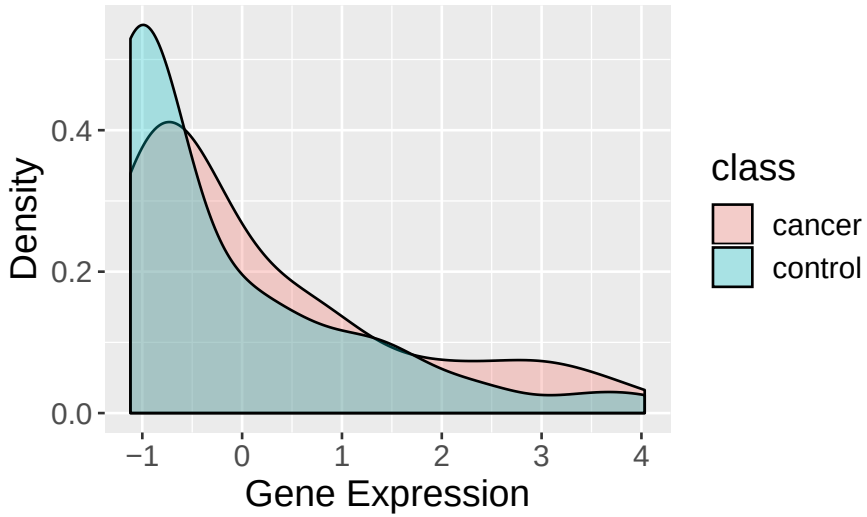
Assuming H_0 is true:

$$T \sim T_{n_1+n_2-2}$$

Conclusions:

- Choose α, reject H_0 if $|T| > T_{\alpha/2,n_1+n_2-2}$.
- Compute p-value, $2P(|T| > |T_{n-2}|)$ where T_{n-2} is t-distributed with $n-2$ dof

Application to Gene



- $\bar{X}_2 \approx 0.21$ (cancer)
- $\bar{X}_1 \approx -0.19$ (healthy)
- $T \approx 1.48$
- p-value ≈ 0.14

Note: T is asymptotically $N(0,1)$ even if X_i not normal. So procedure will be reasonable supposing n is "large."

Modern Hypothesis Testing

Features of Modern Testing:

- Need to test 1000s to millions of hypotheses
- Most null hypotheses are true.

Example: Measure expression of $N \approx 6000$ genes for $n = 102$ patients (50 control, 52 prostate cancer)

- matrix X_{ij} for $i = 1, \ldots, N$ and $j = 1, \ldots, n$
  - rows (i) index gene, N total
  - columns (j) index patient, n total
  - $j = 1, \ldots, n_1$ are controls
  - $j = n_1 + 1, \ldots, n_1 + n_2 = n$ are cancer patients
- $X_{i1}, \ldots, X_{in_1} \sim N(\mu_{i1}, \sigma_i^2)$
- $X_{i,n_1+1}, \ldots, X_{in} \sim N(\mu_{i2}, \sigma_i^2)$

$$H_{0i} : \mu_{i1} = \mu_{i2}$$
$$H_{1i} : \mu_{i1} \neq \mu_{i2}$$

for $i = 1, \ldots, N$.

prostate Data Example

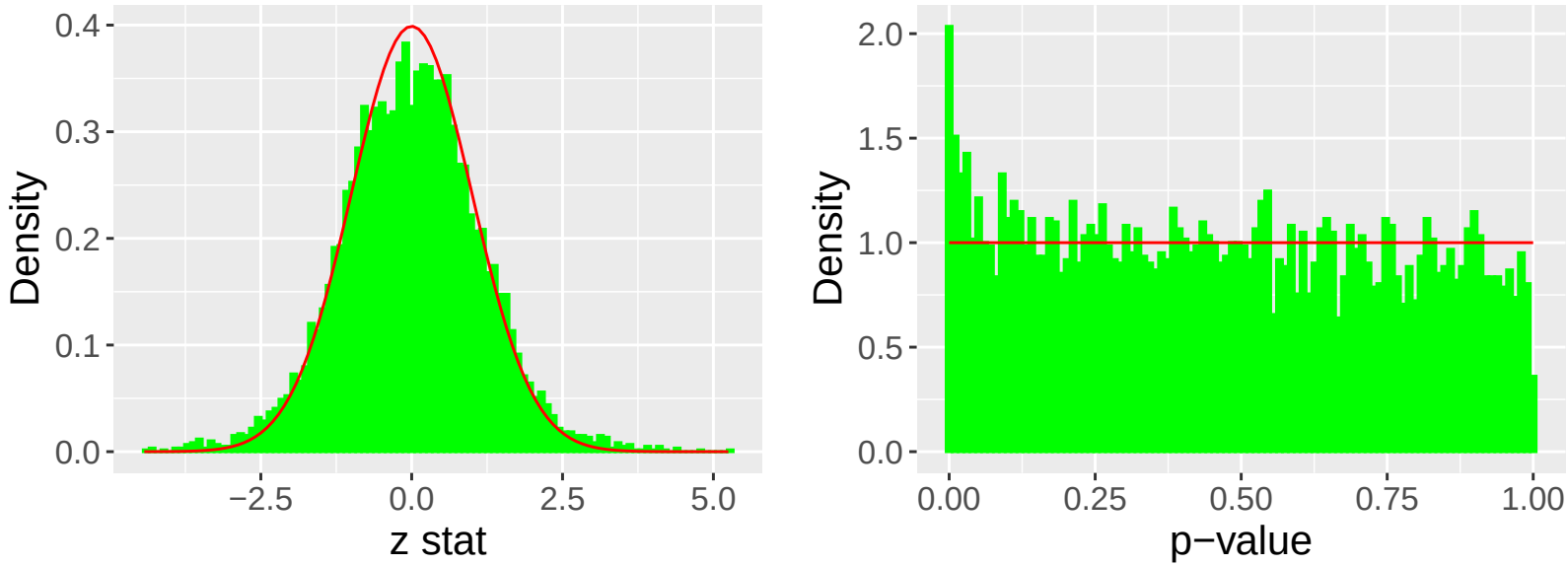
Showed result for $i = 1$. But there are approximately 6000 genes.
So we can compute:

- Test statistics T_i for $i = 1, \ldots, 6000$
- p-values p_i for $i = 1, \ldots, 6000$
- What do we do with this information?
  - How to generalize notations such as Type I Error to many tests?
  - Reasonable thresholds for declaring "significant"

New Opportunities

Plot the distribution of (transformed) test statistics and p-values:

$$\text{z stat} = Z_i = \Phi^{-1}(F_{n-2}(T_i))$$

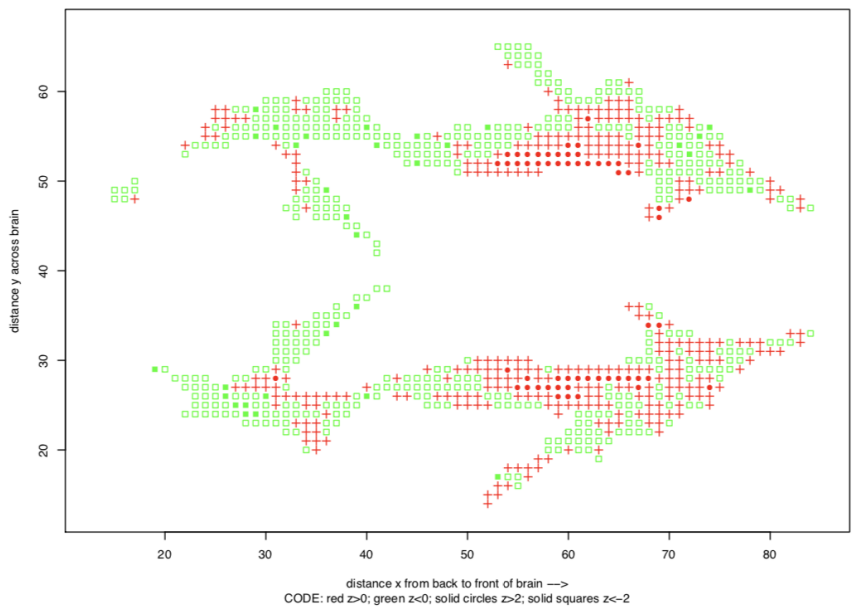


- Excessive large test statistics / small p-values
- Possible Analysis:
  - Classify 26 T statistics greater than 3.5 as discoveries
  - Expect about $2.1 = N(1 - F_{T_{n-2}}(3.5))$ T statistics greater than 3.5 (if all nulls true)
  - So about $2.1/26 < 10\%$ of discoveries are false

This type of analysis is impossible with traditional testing.

Areas of Application

- Genomic data: small number of patients (hundreds or thousands) but large number of variables / patient
  - Gene expression
  - Protein expression
  - Mutation data, e.g. SNPs
- Imaging Data: 1 hypothesis per pixel / voxel



Outline

Family Wise Error Rate Background

- Family wise error rate (FWER) is generalization of Type I Error to multiple testing
- Controlling FWER was popular approach to multiple testing through mid 1990s
- Most appropriate when small number of tests (tens)
- Strongly frequentist

Setup and Notation

- $\mathcal{P}$ denote a model (set of probability distributions).
- R is a (nonrandomized) test function which rejects or does not reject the hypotheses $H_{01}, \ldots, H_{0N}$.

$$R : X \to \mathcal{S}\{1, \ldots, N\}$$

  - $\mathcal{S}$ denotes the power set (all possible subsets of $1, \ldots, N$).
  - $R(X)$ specifies which null hypotheses are rejected.
- For a given $P \in \mathcal{P}$, I_0 specifies which null hypotheses are true.

$$I_0 : P \to \mathcal{S}\{1, \ldots, N\}$$

- The family wise error rate of R is

$$FWER_R = \sup_{P \in \mathcal{P}} P(\cup_{i \in I_0(P)} \{i \in R(X)\})$$

- Family wise error rate (FWER) is the probability that any H_{0i} is falsely rejected. (Equation 3.11 in Efron.)

Bonferroni Correction

- Let p_i be a p-value for hypothesis H_{0i}
  - For any $P \in \mathcal{P}$ where H_{0i} is true $p_i \sim Unif[0,1]$
  - Detail: Actually p_i needs to be stochastically no smaller than $Unif[0,1]$
- The Bonferroni rejection region is

$$R(X) = \{i : i \in \{1, \ldots, N\}, p_i < \alpha/N\}$$

- Bonferroni is stricter than controlling Type I error for single hypothesis at level α

Bonferroni Controls FWER

For any $P \in \mathcal{P}$

$$\begin{aligned} P(\cup_{i \in I_0(P)} \{i \in R(X)\}) &\leq \sum_{i \in I_0(P)} P(\{i \in R(X)\}) \\ &\leq \sum_{i \in I_0(P)} P(p_i < \alpha/N) \\ &\leq \sum_{i \in I_0(P)} \alpha/N \\ &\leq \sum_{i=1}^{N} \alpha/N \\ &= \alpha \end{aligned}$$

Thus

$$FWER_R = \sup_{P \in \mathcal{P}} P(\cup_{i \in I_0(P)} \{i \in R(X)\}) \leq \alpha$$

Adjusted p-values

- R_α for $0 \leq \alpha \leq 1$ is a set of tests such that R_α controls FWER at α
- The adjusted p-value for H_{0i} is

$$\widetilde{p}_i = \inf\{\alpha : i \in R_\alpha(X)\}$$

- If R_α are Bonferroni tests, then

$$\widetilde{p}_i = \min(Np_i, 1)$$

- **Idea:** Rather than specify an α to control FWER and report a set of significant hypotheses, report the adjusted p-values. Reader of results can choose own α.

Bonferroni

- Bonferroni makes no assumptions on dependence structure of p-values (good)
  - Now: Discuss in context of simple normal example
- Bonferroni is very conservative, especially with N large (bad)
  - Discuss later in context of `prostate` data

Sidak's Procedure for Independent Hypotheses

- **Sidak's Procedure:** Reject H_{0i} if

$$p_i \leq 1 - (1 - \alpha)^{1/N}$$

  - Threshold is decreasing in N, so rejecting with many hypotheses becomes more stringent.
  - More liberal than Bonferroni because

$$1 - (1 - \alpha)^{1/N} > \alpha/N$$

- **Theorem:** If p_i are independent, then Sidak's procedure controls FWER at α.
  - Homework question, use facts:

$$P(\cup_i A_i) = 1 - P(\cap_i A_i)$$

$$P(\cap_i A_i) = \prod_i P(A_i) \quad \text{(when } A_i \text{ are independent)}$$

Dependent Tests

Example:

- $X_j \in \mathbb{R}^2$
- $X_j \sim N(\mu, \Sigma)$ for $j = 1, \ldots, n$
- $\mu = (\mu_1, \mu_2)^T$

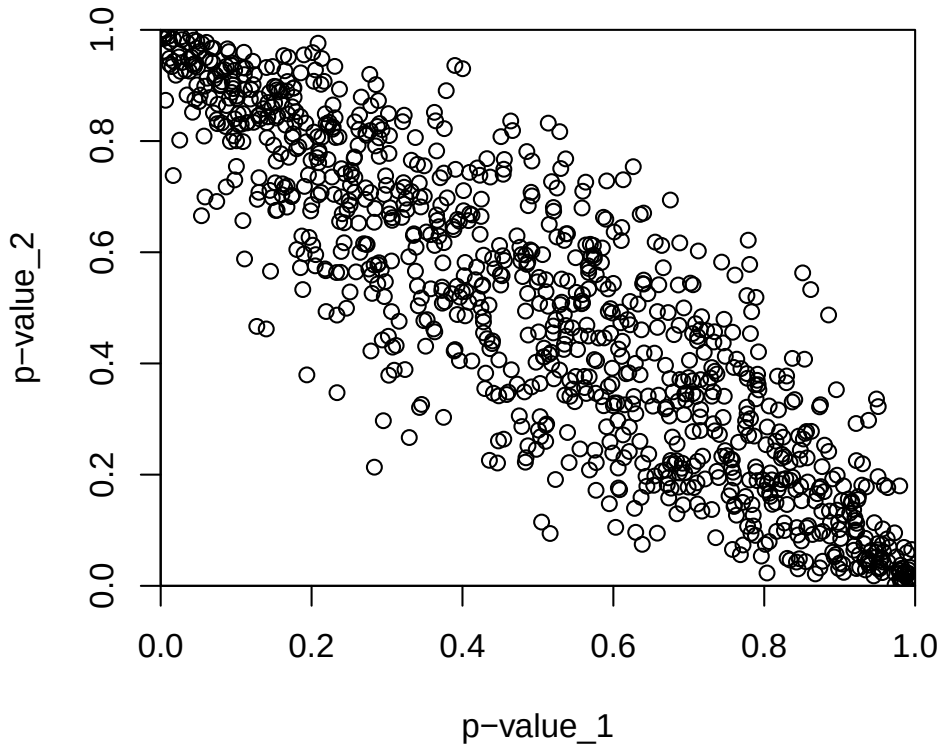
$$\Sigma = \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix}$$

- For $i = 1, 2$ hypotheses are:

$$H_{0i} : \mu_i = 0$$
$$H_{1i} : \mu_i > 0$$

- Simulate under the global null (H_{01} and H_{02} true) $M = 1000$ times
- Compute p-value (1 sided z-test) for each simulation run, each hypothesis
- Result M pairs of p-values

p-value Joint Distribution



- p-values demonstrate strong negative correlation
- Sidak's procedure may not control FWER for such model

Holm's Procedure

- Order the p-values $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(N)}$
- Holm's procedure at α rejects hypothesis for $p_{(i)}$ if

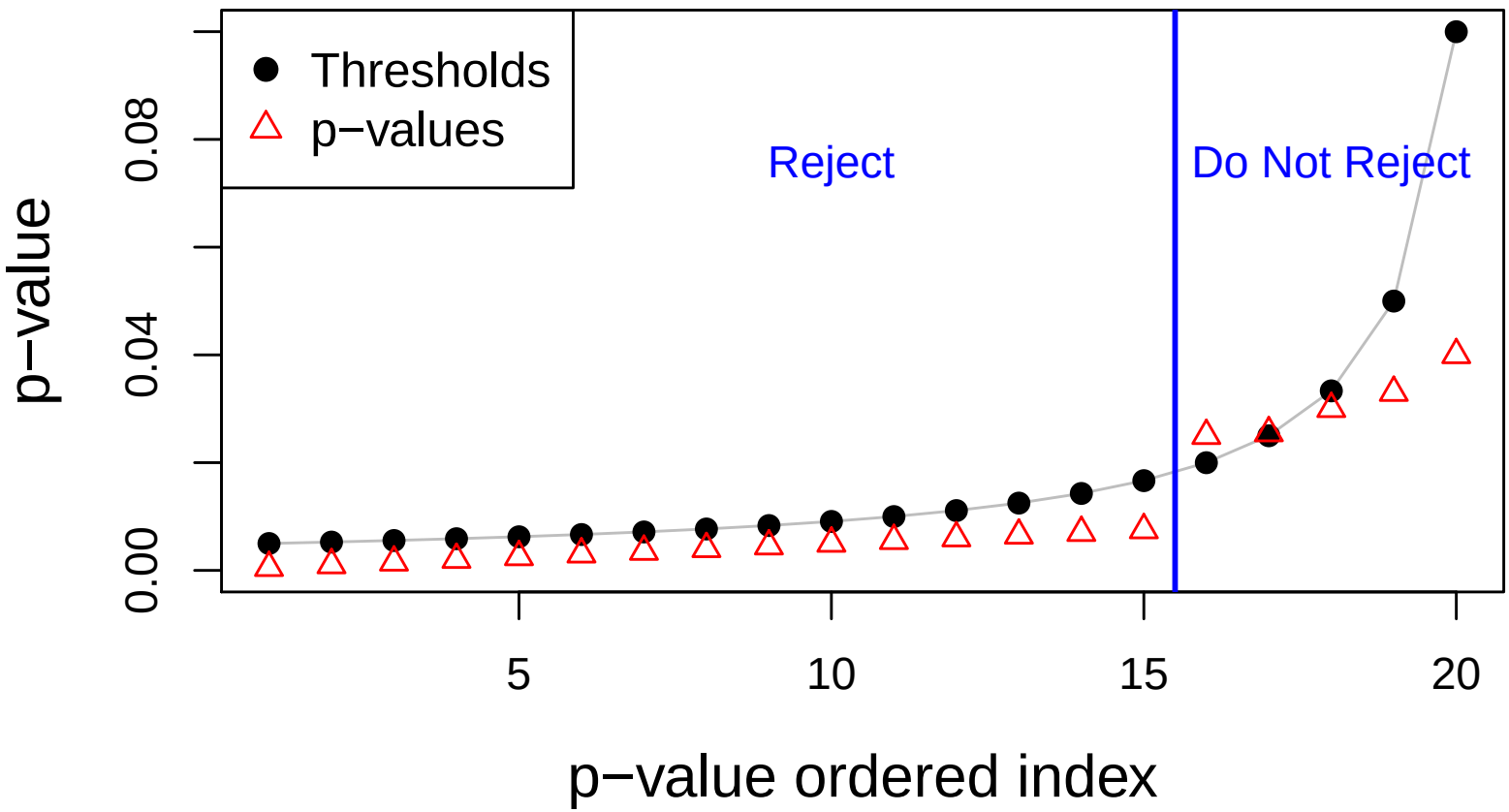
$$p_{(j)} \leq \frac{\alpha}{N - j + 1} \text{ for } j = 1, \ldots, i$$

- Holm has higher power than Bonferroni because

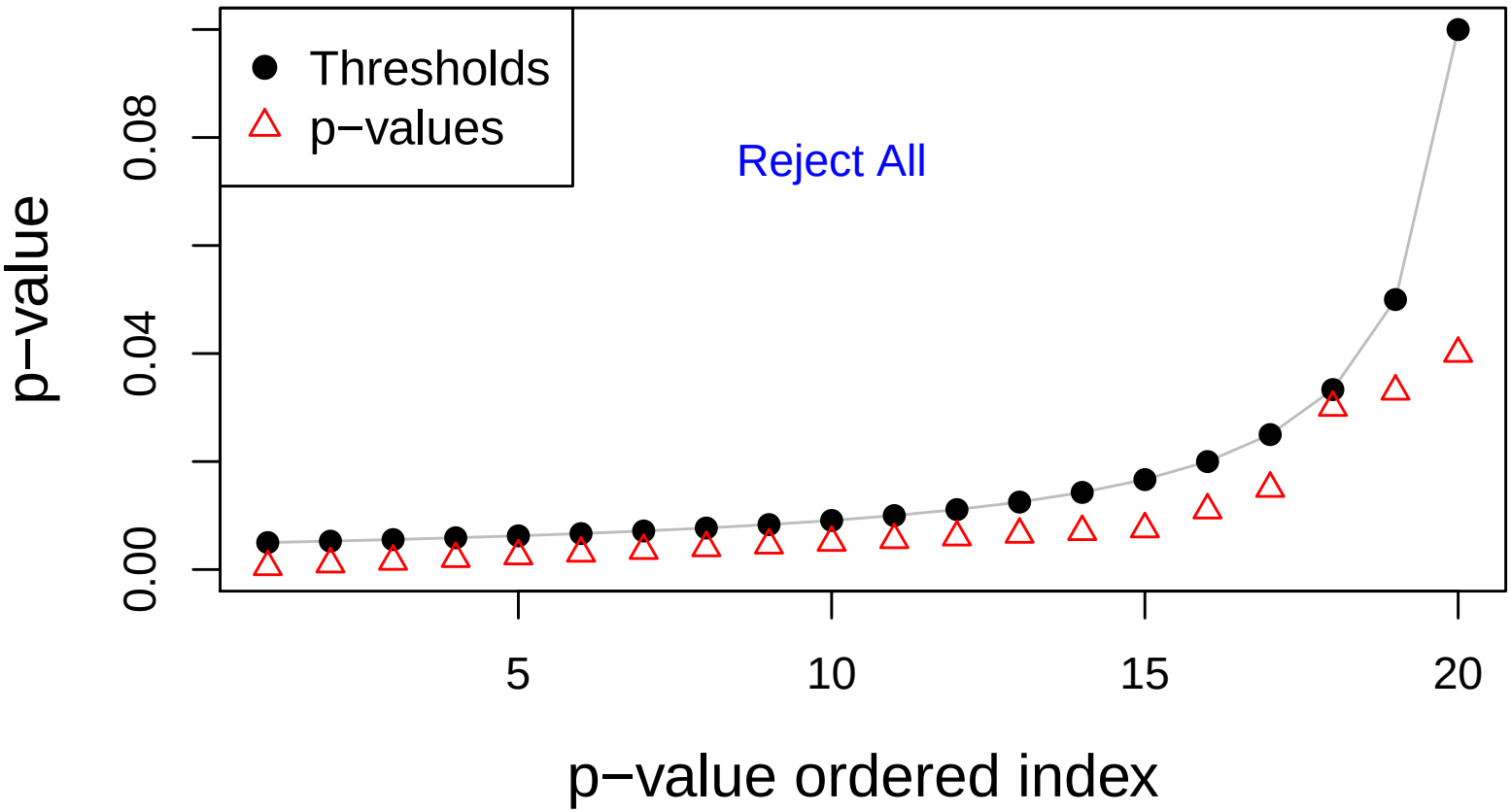
$$\frac{\alpha}{N} \leq \frac{\alpha}{N - j + 1} \text{ for all } j$$

- Holm controls FWER at α

Holm Visual with 20 p-values



Holm Visual with 20 p-values



Message: The decision to reject p_i can change if p_j for $j \neq i$ changes even if i remains the same.

Proof of Holm FWER Control

Step 1:

Define: $N_0 = \#I_0 =$ number of true nulls

Claim:

$$\{\text{Falsely reject a null}\} \subseteq \{\text{true null with p-value } \leq \alpha/N_0\} \quad (1)$$

- $p_{(i)} \equiv$ smallest p-value among true nulls
- $p_{(1)}, \ldots, p_{(i-1)}$ are false nulls
- $i - 1 \leq N - N_0 =$ number of false nulls
- $i \leq N - N_0 + 1$
- If a null is falsely rejected, then $p_{(i)}$ must be rejected

$$p_{(i)} \leq \frac{1}{N - i + 1} \leq \frac{1}{N - (N - N_0 + 1) + 1} = \frac{\alpha}{N_0}$$

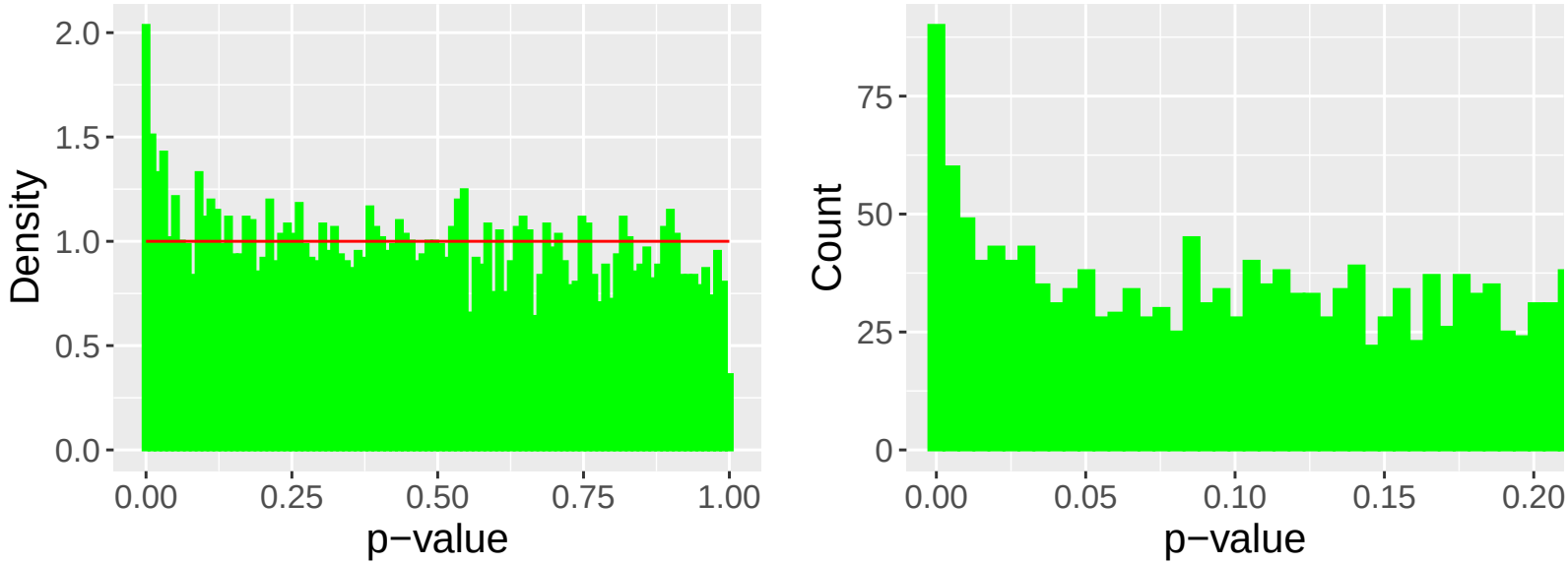
Proof of Holm FWER Control

Step 2: Using Equation (1) and Bonferroni like proof we have

$$\begin{aligned} P(\{\text{Falsely reject a null}\}) &\leq P(\{\text{true null with p-value } \leq \alpha/N_0\}) \\ &\leq P(\cup_{i \in I_0} p_i \leq \alpha/N_0) \\ &\leq \sum_{i \in I_0} P(p_i \leq \alpha/N_0) \\ &\leq \sum_{i \in I_0} \alpha/N_0 \\ &\leq \alpha \end{aligned}$$

Since holds for any $P \in \mathcal{P}$, obtain FWER control.

Bonferroni and prostate Data



- Standard: 477 p-values < 0.05 (too liberal)
- Bonferroni: 2 p-values < 0.05 / N (too conservative)
- Potential solution: Use Bonferroni with larger α
  - Doesn't Work: With $\alpha = 0.5$, Bonferroni bound is $\alpha/N < 0.0001$, satisfied by only 14 hypotheses
  - FWER is too strict a criteria to control in high dimensional test settings

Summary

- Historical development of multiple testing through 1980s focused on controlling FWER
  - Several creative ways to get more power than Bonferroni (e.g. Holms)
  - Among FWER control procedures, Bonferroni remains most popular due to ease of use
  - With additional assumptions (e.g. independence), can obtain additional power
    - Sidak's procedure
    - Hochberg: Did not discuss. Assumes independence, but has similar flavor to Holm
  - Useful with small number of hypotheses (tens)
- Become excessively conservative as N grows large
- Controlling FWER not right criteria for large N
- Desire: Obtain results closer to a Bayesian analysis, i.e. $P(H_{0i} \text{ true}) = 0.02$