Multiple Testing

Naomi Altman
The Pennsylvania State University
nsa1@psu.edu
May 2017

May Institute at Northeastern U.

Multiple Testing

Naomi Altman
The Pennsylvania State University
nsa1@psu.edu
May 2017

May Institute at Northeastern U.

The problem with p-values

- Suppose we test 10,000 hypotheses and the null is true for all

- If we reject the null (declare significance) at $p \leq \alpha$

Altman

The problem with p-values

- Suppose we test 10,000 hypotheses and the null is true for all

- If we reject the null (declare significance) at $p \leq \alpha$
- How many rejections do we expect?
- How many of these are *false* rejections?

Altman

The problem with p-values

- Suppose we test 10,000 hypotheses and the null is true for 9000

- If we reject the null (declare significance) at $p \leq \alpha$, then we expect $9{,}000\alpha$ tests to falsely reject
- This can cost us a lot as we follow-up false leads

Altman

The problem with p-values

- Suppose we test 10,000 hypotheses and the alternative is true for 1000 of them

- If we have power β at $p \leq \alpha$

Altman

The problem with p-values

- Suppose we test 10,000 hypotheses and the alternative is true for 1000 of them

- If we have power β at $p \leq \alpha$
- How many rejections do we expect?
- How many of these are *false* rejections?

Altman

The problem with p-values

- Suppose we test 10,000 hypotheses and the alternative is true for 1000 of them

- If we have power 80% at $p \leq .05$

- We expect to reject 9000*.05=450 of the nulls and 1000*.8=800 of the true alternatives

- We expect a False Discovery Rate of 450/1250=36%

Altman

The problem with p-values

- Suppose we test 10,000 hypotheses and the alternative is true for 1000 of them
- If we have power 80% at $p \leq .05$
- We expect to accept 9000*.95=8550 of the nulls and 1000*.2=200 of the true alternatives
- We expect a False NONDiscovery Rate of 200/8750=2.3%

Altman

P-value Histogram

Consider the histogram of p-values from the null tests. What does it look like?

What percentage of tests have $p<0.01$?

$p<0.02$

p between 0.01 and 0.02?

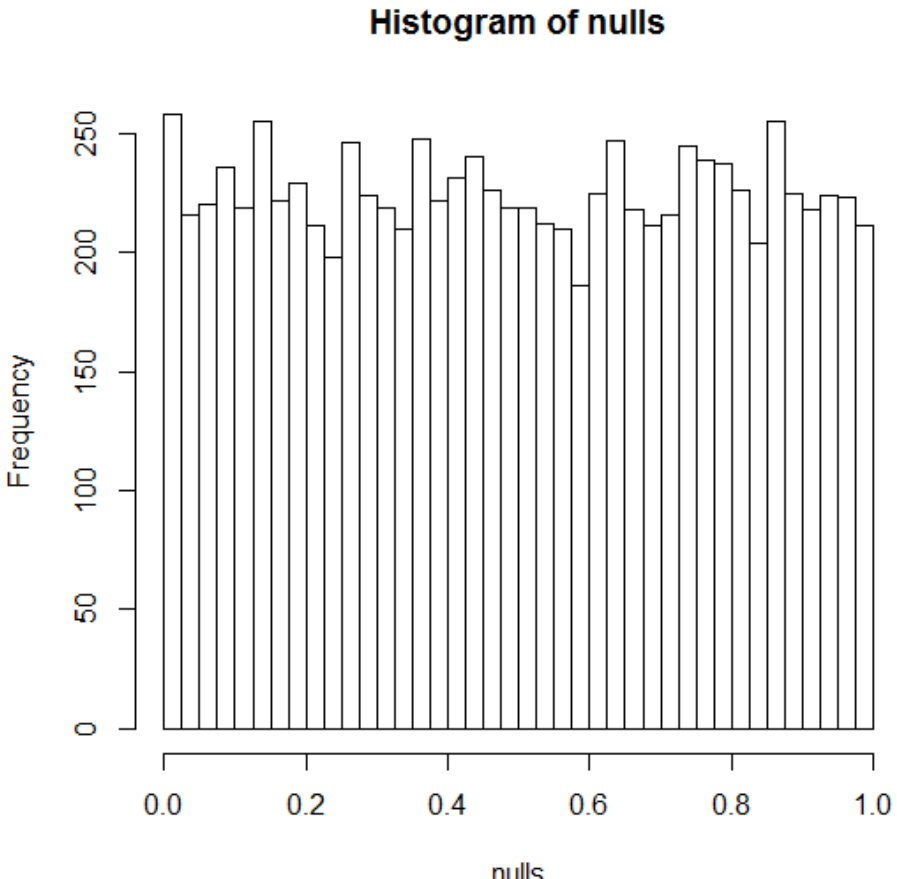
p between 0.02 and 0.03?

p between 0.49 and 0.50?

Altman

P-value Histogram

Consider the histogram of p-values from the null tests. What does it look like?



Altman

P-value Histogram

Consider the histogram of p-values from the NON-null tests. What does it look like?

What percentage of tests have $p<0.01$?

$p<0.02$

p between 0.01 and 0.02?

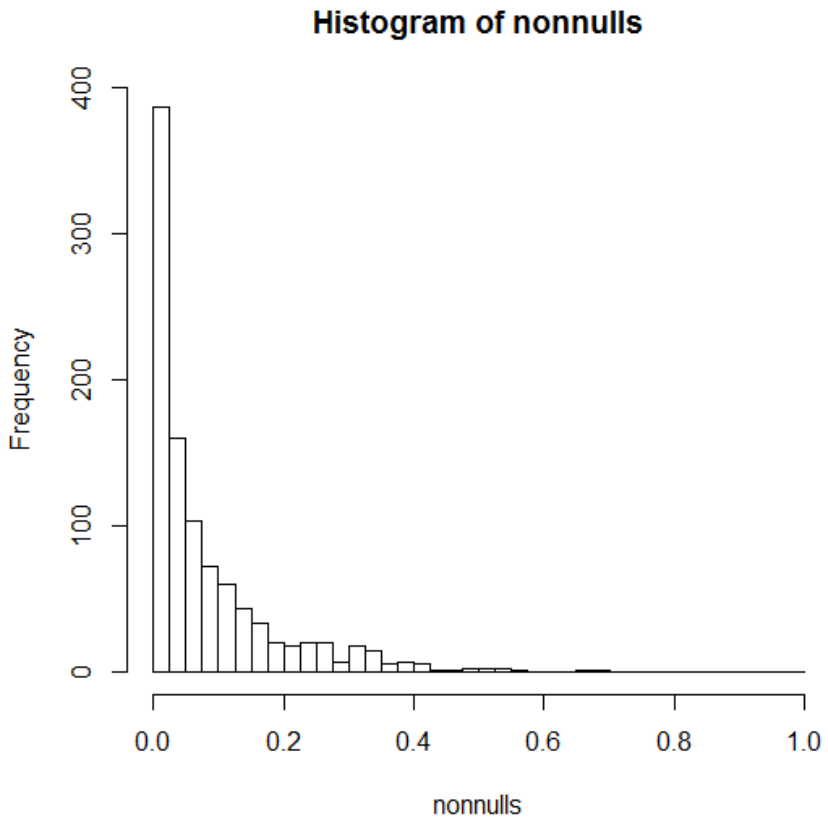
p between 0.02 and 0.03?

p between 0.49 and 0.50?

Altman

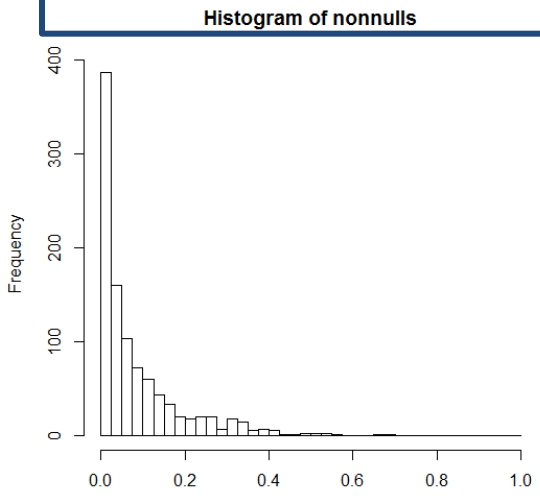
P-value Histogram

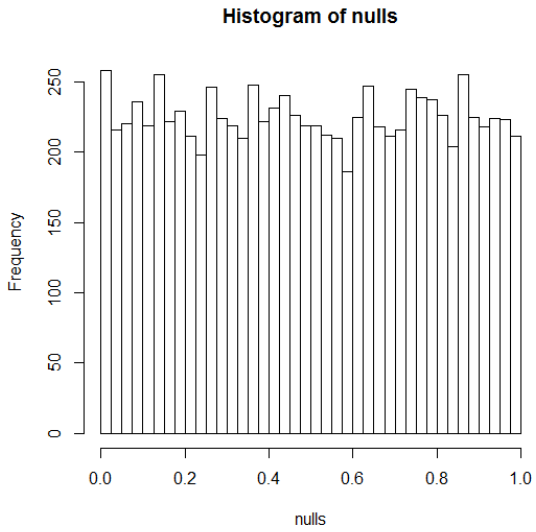
Consider the histogram of p-values from the NON-null tests. What does it look like?



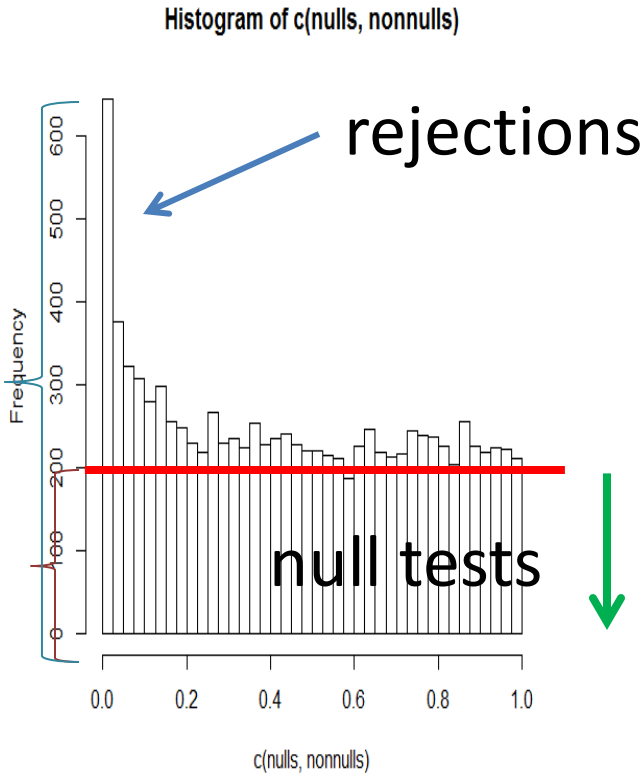
Altman

P-value Histogram



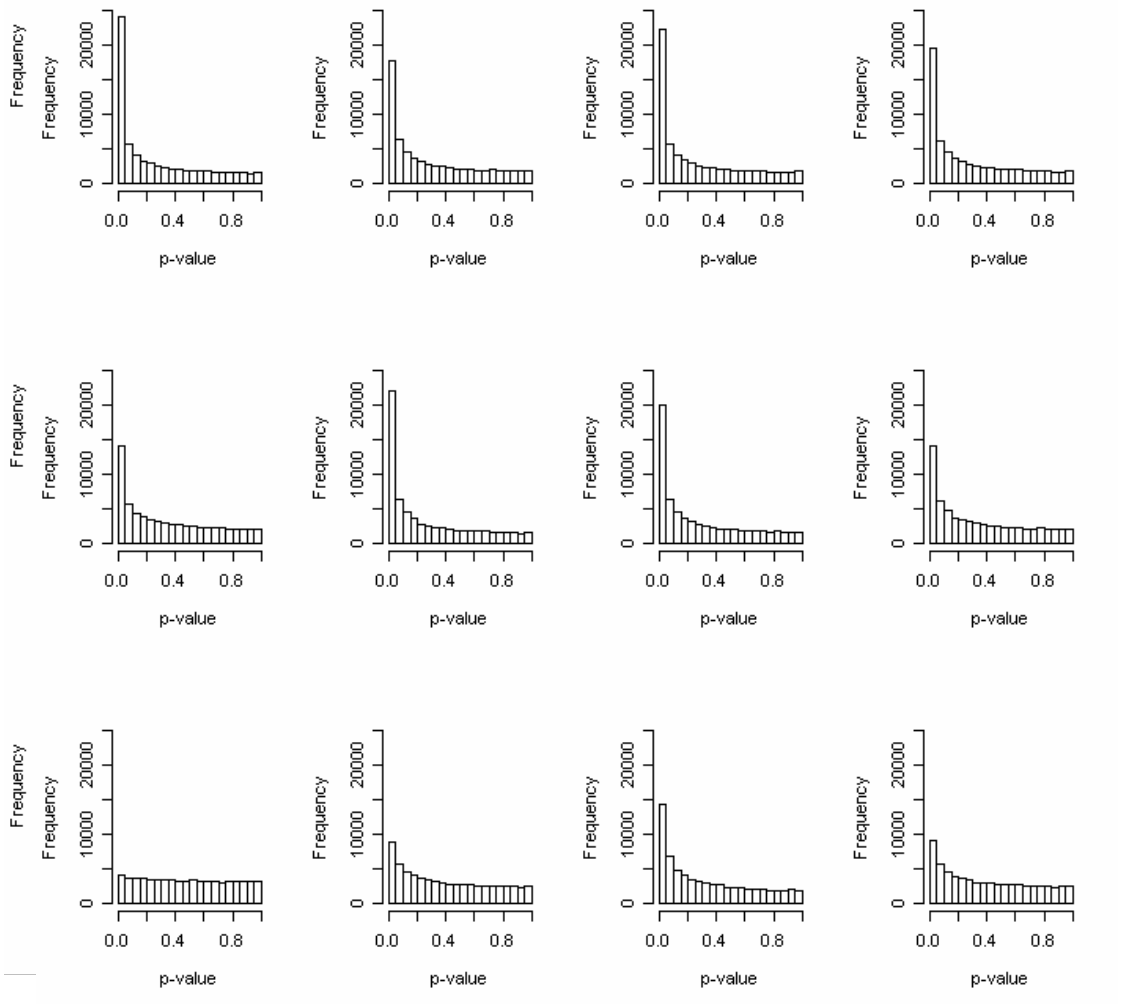


false rejections



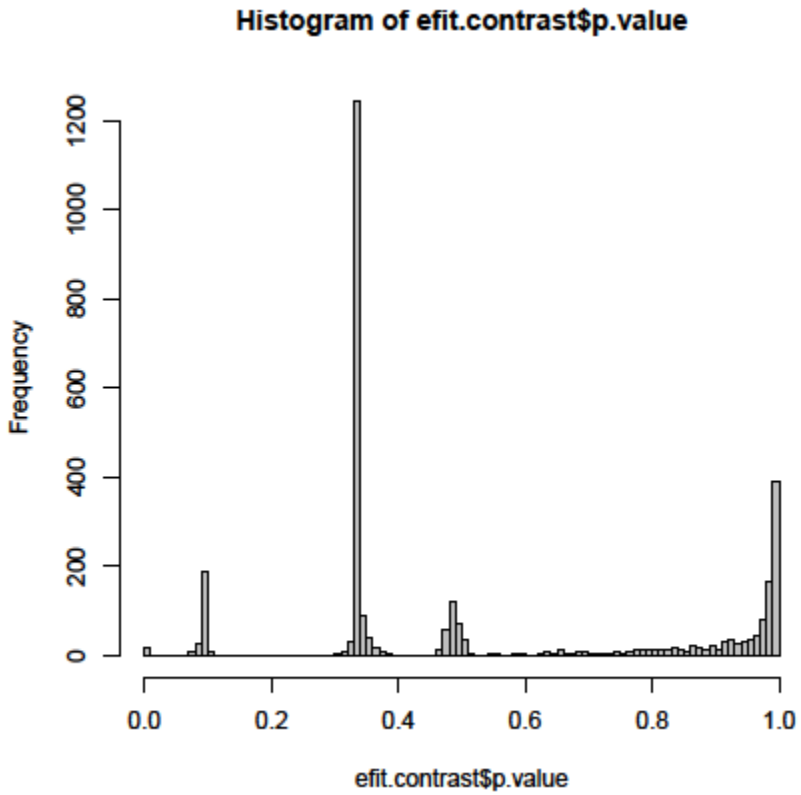
Altman

P-values from a Microarray Study



Altman

P-values from a Proteomic Study



Altman

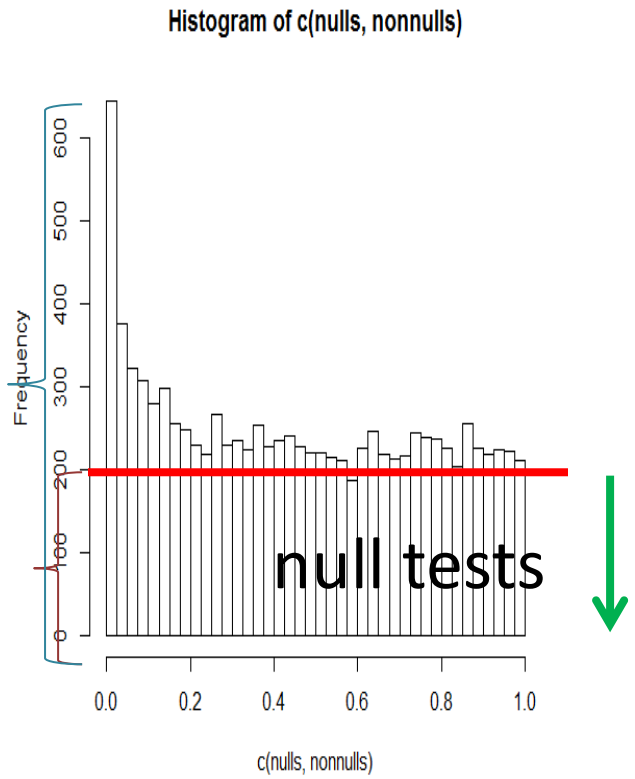
Estimating m_0

Since false discoveries can occur only among the null tests, it is helpful to estimate the number of null tests

m_0

or the proportion of null tests

$\pi_0 = m_0/m$



Altman

Estimating m_0

There are 2 very simple (and quite accurate) methods:

Storey's method: area under flat part of curve

Pounds and Cheng method:

$$m_0 \sim 2*m*\text{average(p-value)}$$

(assumes all the non-nulls have p-value=0).

Note that both may not yield whole numbers

Altman

When performing m tests

	Not Significant	**Significant**	
H_0 true	U	V	m_0
H_0 false	T	S	$m-m_0$
	W	R	m

Total errors: T+V

Altman

The usual approach

Try to control the number of false significant results.

(Why not total errors?)

Before 1995

	Not Significant	**Significant**	
H_0 true	U	V	m_0
H_0 false	T	S	$m-m_0$
	W	R	m

Control Prob(V>0)

Problem? As m gets big ...

Altman

Before 1995

	Not Significant	**Significant**	
H_0 true	U	V	m_0
H_0 false	T	S	$m-m_0$
	W	R	m

Control Prob(V>0)

Problem? As m gets big ...

This is called family-wise error rate (FWER)

The most famous method is the Bonferroni method

Altman

Bonferroni Method

To control FWER at level α reject when

$$p<\alpha/m$$

Expected rejections?

Power?

Adaptive Bonferroni Method

To control FWER at level α reject when

$$\mathbf{p<\alpha/m_0}$$

Slightly better (but not much)

Holm's Method

A more powerful method was devised by Holm:

sort the p-values so that

$p_1 \leq p_2 \ldots \leq p_m$

Starting from the smallest p-value, reject if

$p_i \leq \alpha/(m-i+1)$

So p_1 has to satisfy Bonferroni's criterion, but the larger p-values do not

1995 “Omics” Begin

Altman

1995 “Omics” Begin

	Not Significant	**Significant**	
H_0 true	U	V	m_0
H_0 false	T	S	$m-m_0$
	W	R	m

Benjamini and Hochberg

False Discovery Rate

If m is large, we ought to tolerate a few errors:

Control $q=E(V/R \mid R>0)Prob(R>0)$ the expected percentage of rejections which are false.

Altman

1995 “Omics” Begin

	Not Significant	**Significant**	
H_0 true	U	V	m_0
H_0 false	T	S	$m-m_0$
	W	R	m

Benjamini and Hochberg

False Discovery Rate

BH show that rejecting when $p_i \leq q\ i/m$ (sorted) controls FDR at level q.

Altman

1995 “Omics” Begin

	Not Significant	**Significant**	
H_0 true	U	V	m_0
H_0 false	T	S	$m-m_0$
	W	R	m

Benjamini and Hochberg

False Discovery Rate

BH show that rejecting when $p_i \leq q\ i/m$ (sorted) controls FDR at level $q\ m_0/m$.

BH “adjusted p-value”: $\min(p_i\ m/i, 1)$

Altman

Adaptive BH

	Not Significant	**Significant**	
H_0 true	U	V	m_0
H_0 false	T	S	$m-m_0$
	W	R	m

Benjamini and Hochberg

False Discovery Rate

BH show that rejecting when $p_i \leq q\ i/m_0$ (sorted) controls FDR at level q

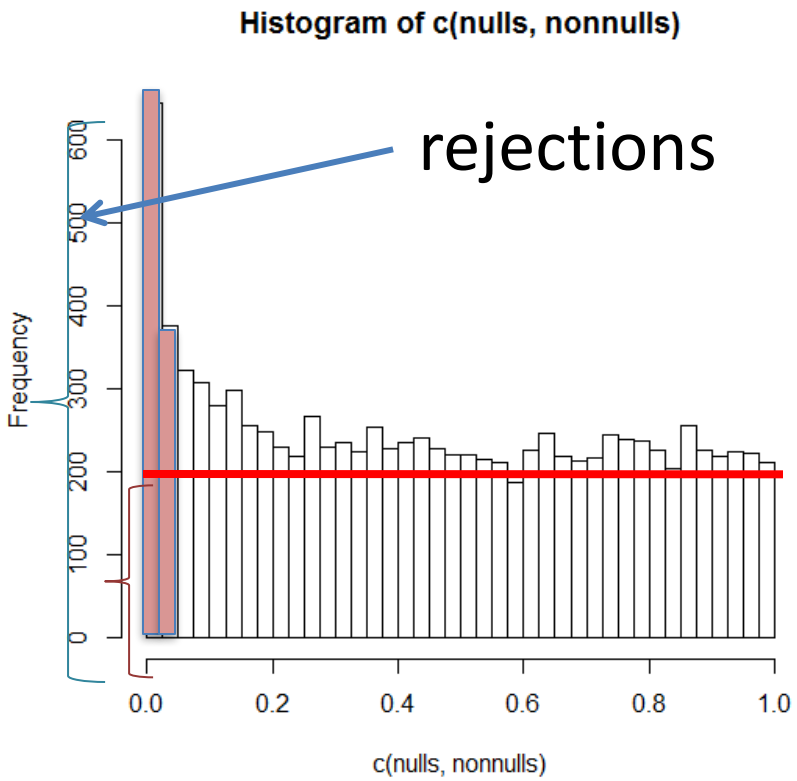
BH "adjusted p-value": $\min(p_i\ m_0/i, 1)$

Altman

Storey

Estimated
q(p)=FDR(p)
=p m_0/R(p)

false
rejections
αm_0



Altman

Our objective

We will simulate:

- 9000 features with no difference in mean intensity
- 1000 features with a difference that gives power 80% if we reject at $p<0.05$

We will look at the estimate of π_0 (m_0 /m)
The number of false discoveries
The number of true discoveries
Power

Bonferroni
BH
q-values

Altman