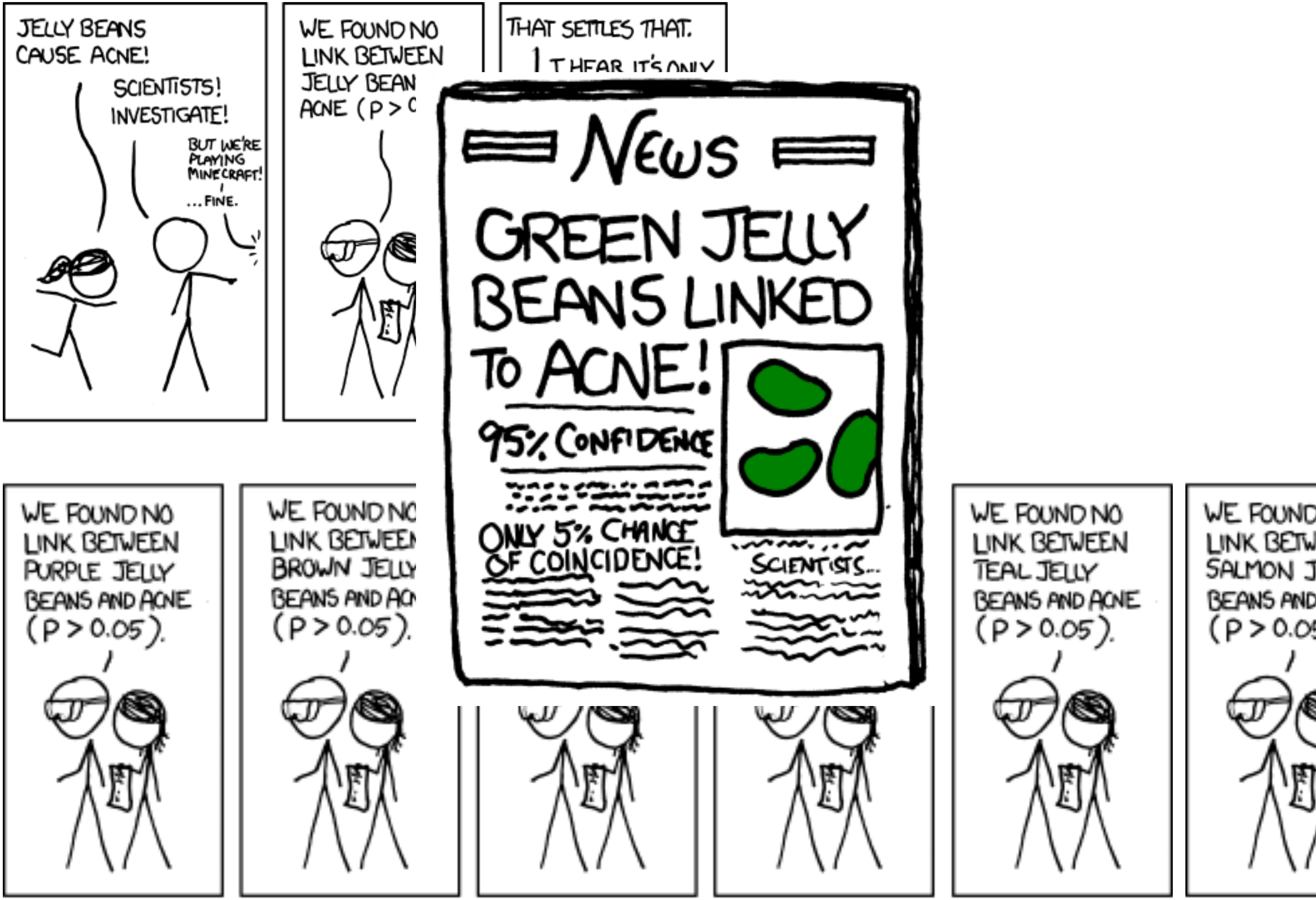
P-values and statistical tests

7. Multiple test corrections

Marek Gierliński
Division of Computational Biology

Hand-outs available at http://is.gd/statlec


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEAN ACNE (P > C
THAT SETTLES THAT.
I HEAR IT'S ONLY
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND AC (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND LINK BETW SALMON J BEANS AND (P > 0.0

Lets perform a test m times

False positives

True positives

	H_0 true	H_0 false	Total
Significant	FP	TP	D
Not significant	TN	FN	$m - D$
Total	m_0	m_1	m

Number of discoveries

Number of tests

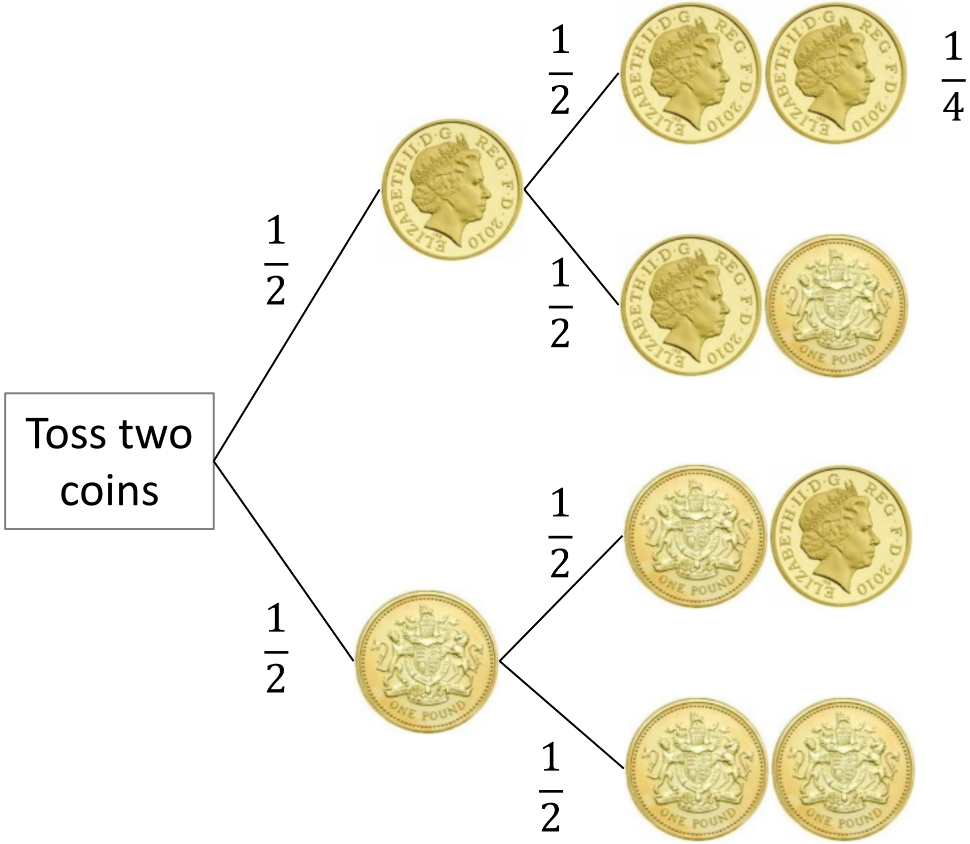
True negatives

False negatives

Family-wise error rate

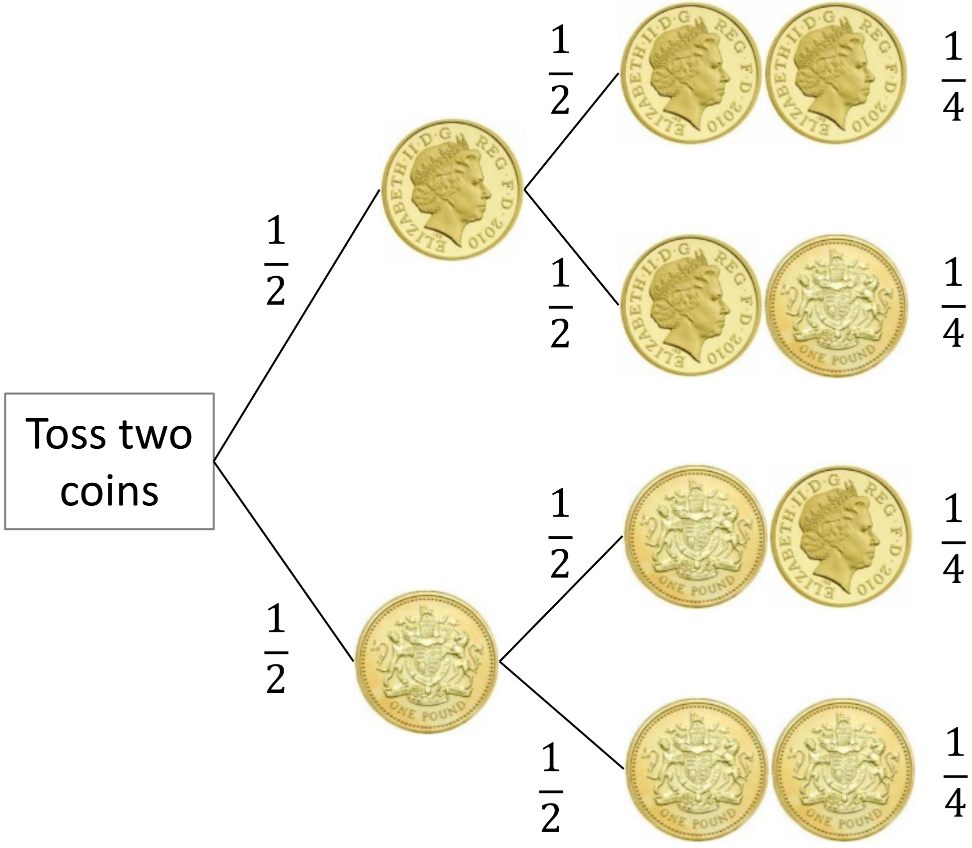
$$FWER = \Pr(FP \geq 1)$$

Probabilities of independent events multiply



$P(H \text{ and } H) = P(H) \times P(H)$

Probabilities of either event is $1 - (1 - p)^2$



$P(H \text{ or } H) = ?$

$P(T) = 1 - P(H)$

$$P(T \text{ and } T) = P(T) \times P(T)$$
$$= \left(1 - P(H)\right)^2$$

$P(H \text{ or } H) = 1 - P(T \text{ and } T)$

$P(H \text{ or } H) = 1 - \left(1 - P(H)\right)^2$

$$P(H \text{ or } H) = 1 - \left(1 - \frac{1}{2}\right)^2 = \frac{3}{4}$$

False positive probability

H_0: no effect
Set $\alpha = 0.05$

One test

Probability of having a false positive
$$P_1 = \alpha$$

Two independent tests

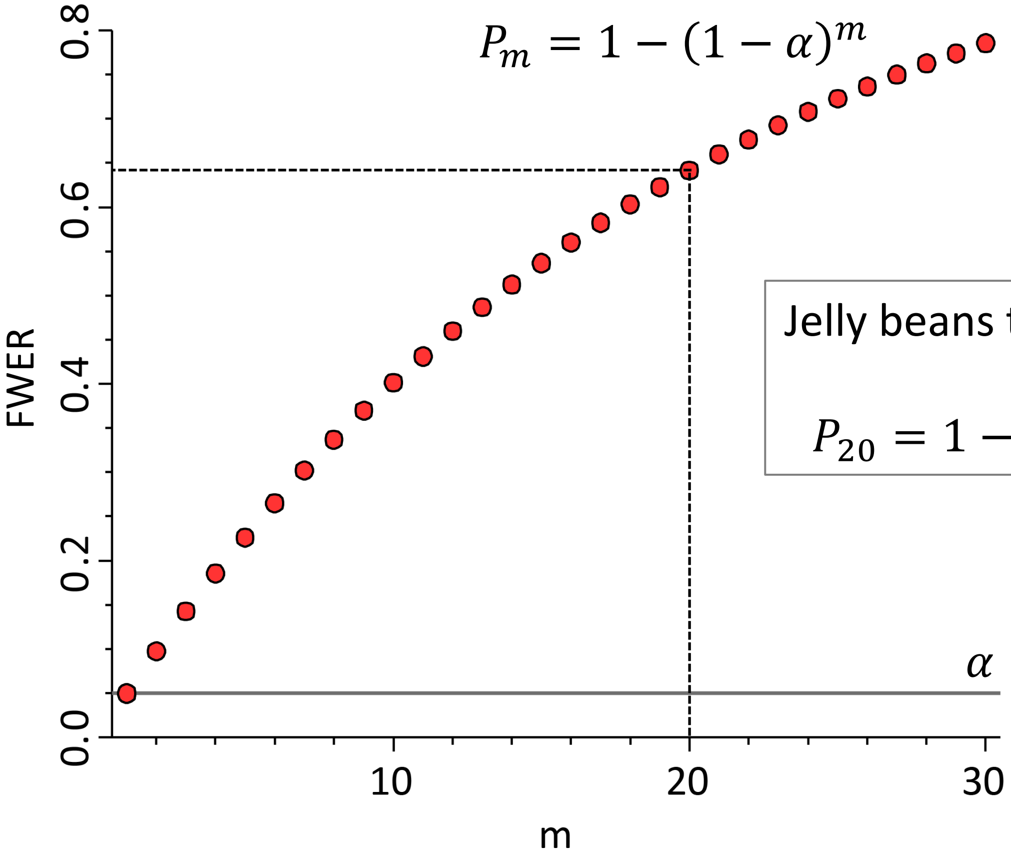
Probability of having at least one false positive in either test
$$P_2 = 1 - (1 - \alpha)^2$$

m* *independent **tests**

Probability of having at least one false positive in any test
$$P_m = 1 - (1 - \alpha)^m$$

Family-wise error rate (FWER)

Probability of having at least one false positive among m tests; $\alpha = 0.05$



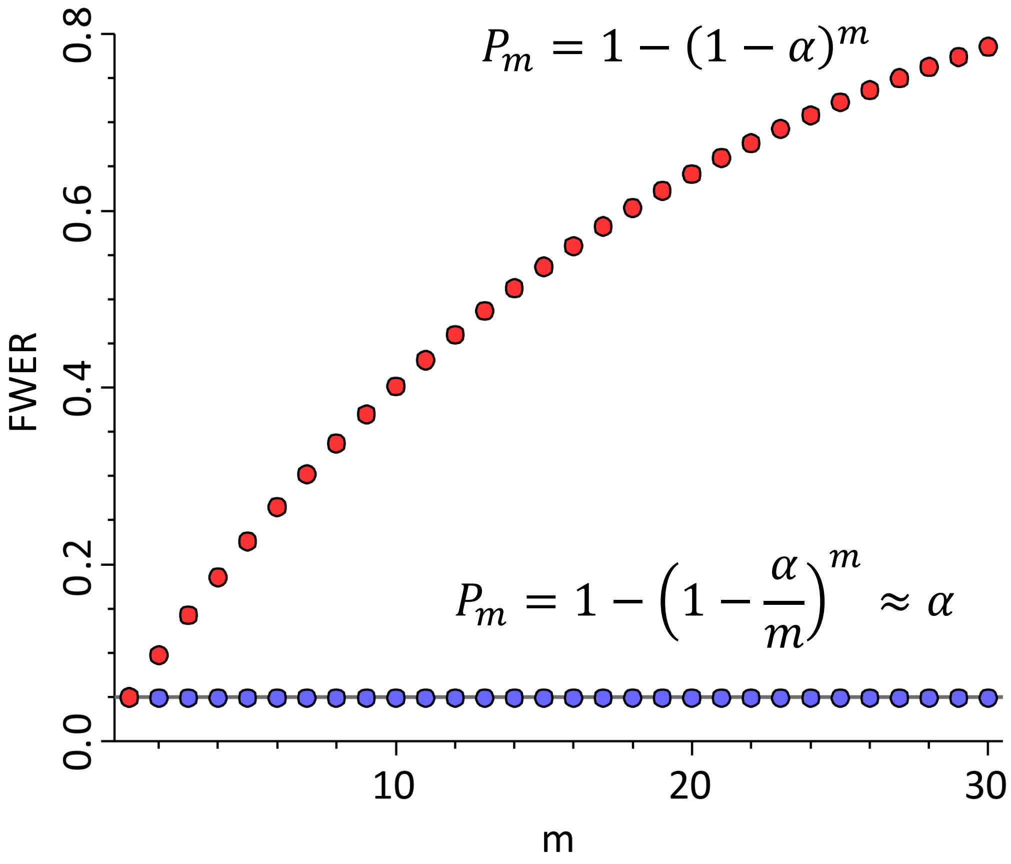
$$P_m = 1 - (1 - \alpha)^m$$

Jelly beans test, $m = 20, \alpha = 0.05$

$$P_{20} = 1 - (1 - 0.05)^{20} = 0.64$$

Bonferroni limit – to control FWER

Probability of having at least one false positive among m tests; $\alpha = 0.05$



Controlling FWER

We want to make sure that

$$FWER \le \alpha'.$$

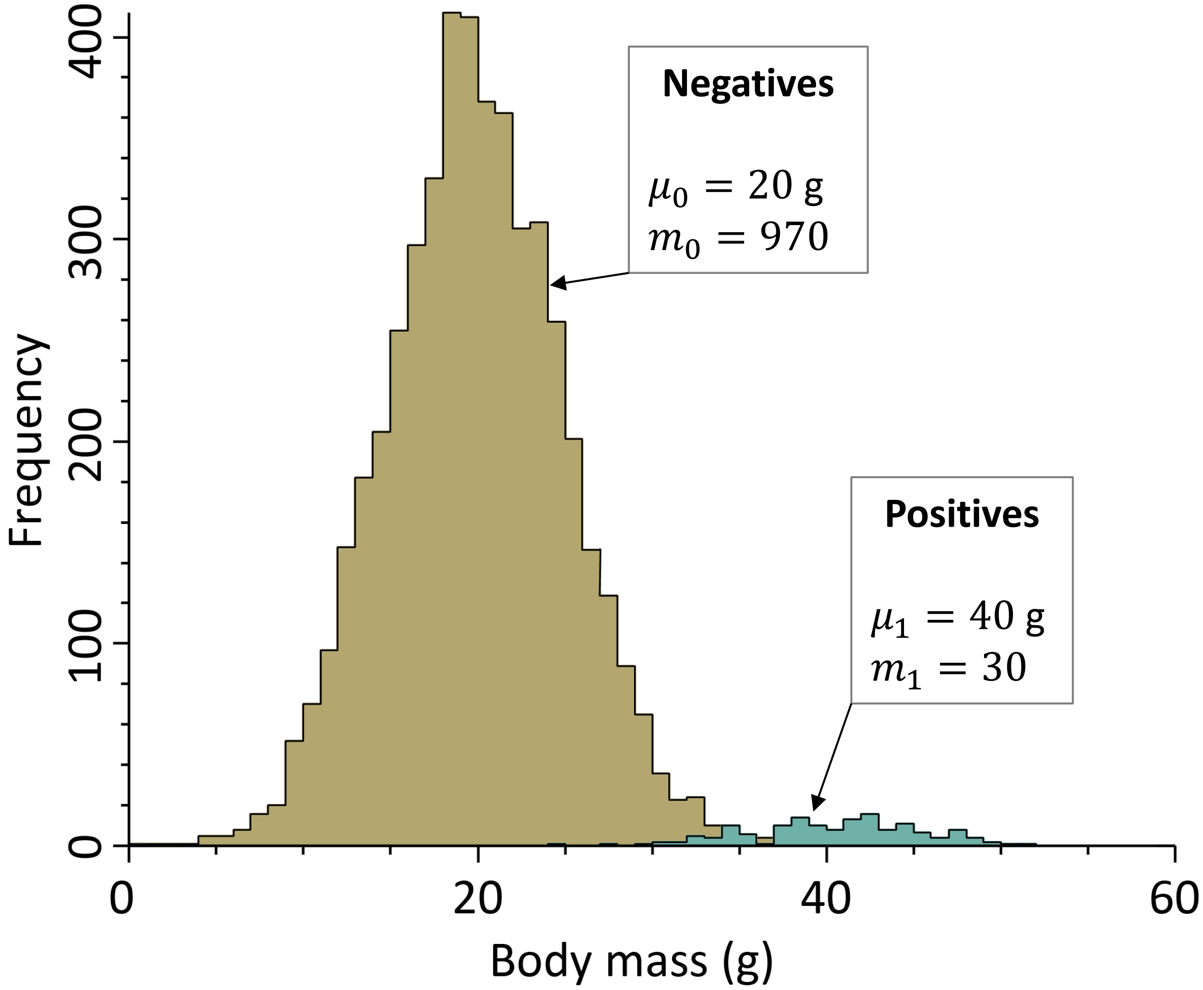
Then, the FWER is controlled at level α'.

Bonferroni limit

$$\alpha' = \frac{\alpha}{m}$$

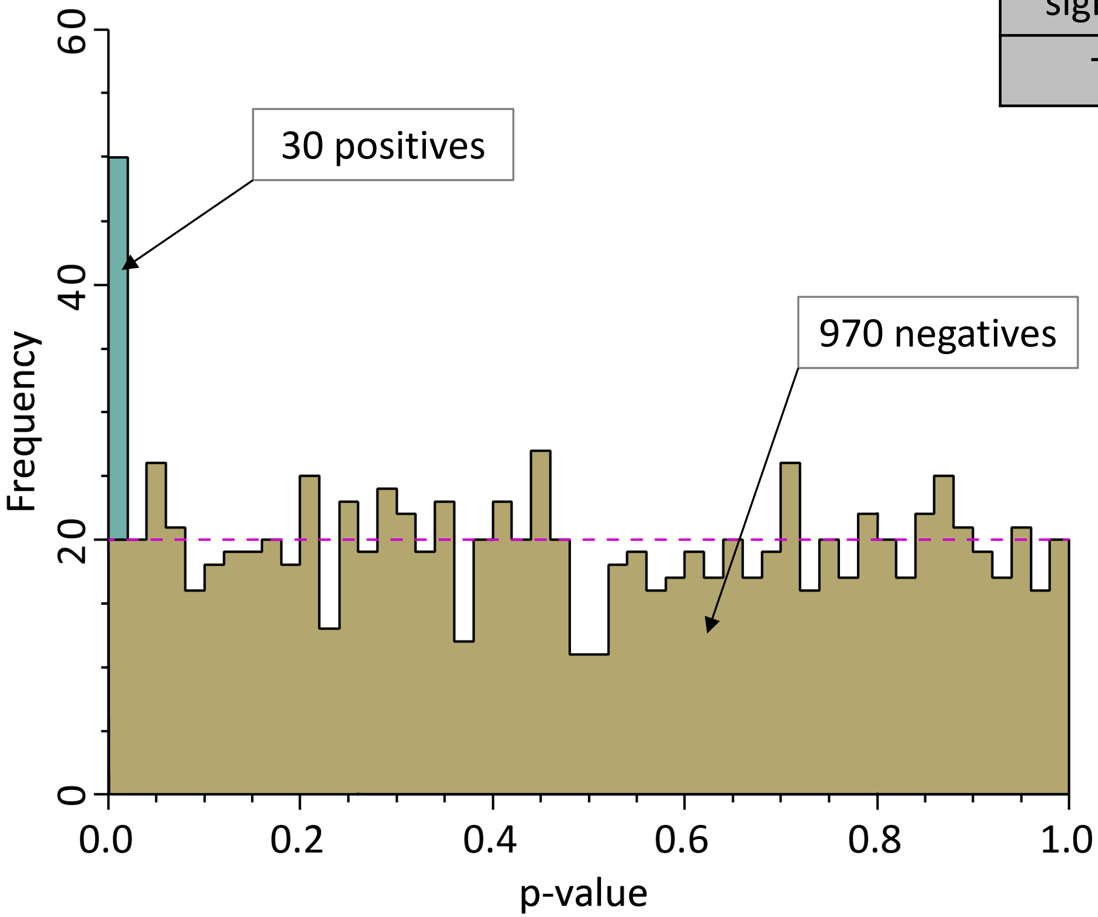
Test data (1000 independent experiments)

Random samples, size $n = 5$, from two normal distributions



One sample t-test, H_o: $\mu = 20$ g

No correction			
	H_0 true	H_0 false	Total
Significant	56	30	86
Not significant	914	0	914
Total	970	30	1000



One sample t-test, H$_o$: $\mu = 20$ g

No correction			
	H$_0$ true	H$_0$ false	Total
Significant	$FP = 56$	$TP = 30$	86
Not significant	$TN = 914$	$FN = 0$	914
Total	970	30	1000

Family-wise error rate $FWER = \Pr(FP \geq 1)$

False positive rate $FPR = \frac{FP}{m_0} = \frac{FP}{FP + TN}$

False negative rate $FNR = \frac{FN}{m_1} = \frac{FN}{FN + TP}$

$$FPR = \frac{56}{56 + 914} = 0.058$$

$$FNR = \frac{0}{0 + 30} = 0$$

Bonferroni limit

	No correction	Bonferroni
α	0.05	5×10^{-5}
FPR	0.058	0
FNR	0	0.87

No correction			
	H_0 true	H_0 false	Total
Significant	56	30	86
Not significant	914	0	914
Total	970	30	1000

Bonferroni			
	H_0 true	H_0 false	Total
Significant	0	4	4
Not significant	970	26	996
Total	970	30	1000

Holm-Bonferroni method

Sort p-values

$$p_{(1)}, p_{(2)}, \dots, p_{(m)}$$

Reject (1) if $p_{(1)} \leq \frac{\alpha}{m}$

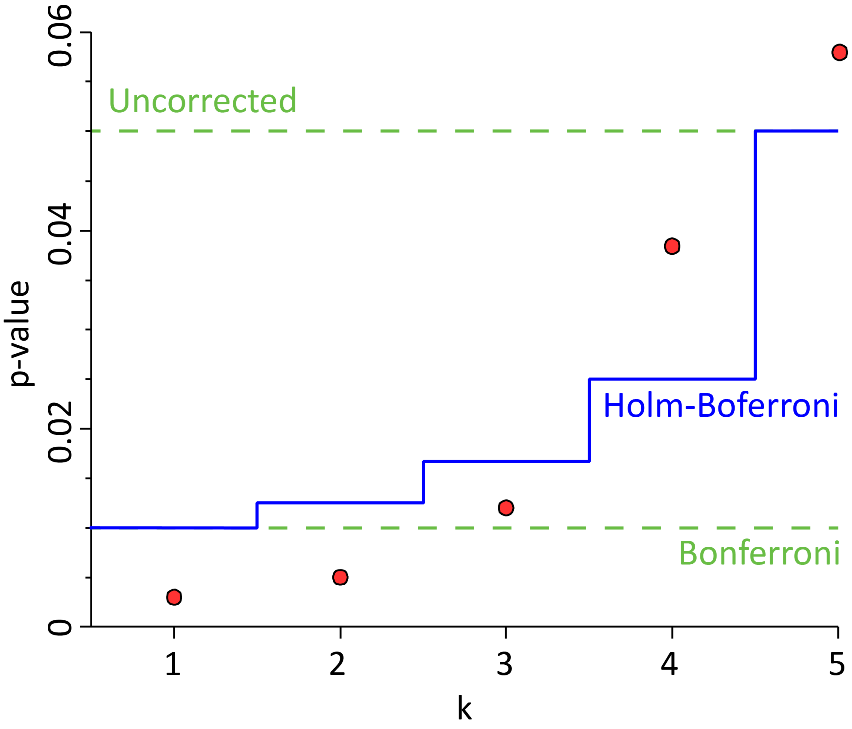
Reject (2) if $p_{(2)} \leq \frac{\alpha}{m-1}$

Reject (3) if $p_{(3)} \leq \frac{\alpha}{m-2}$

…

Stop when $p_{(k)} > \frac{\alpha}{m-k+1}$

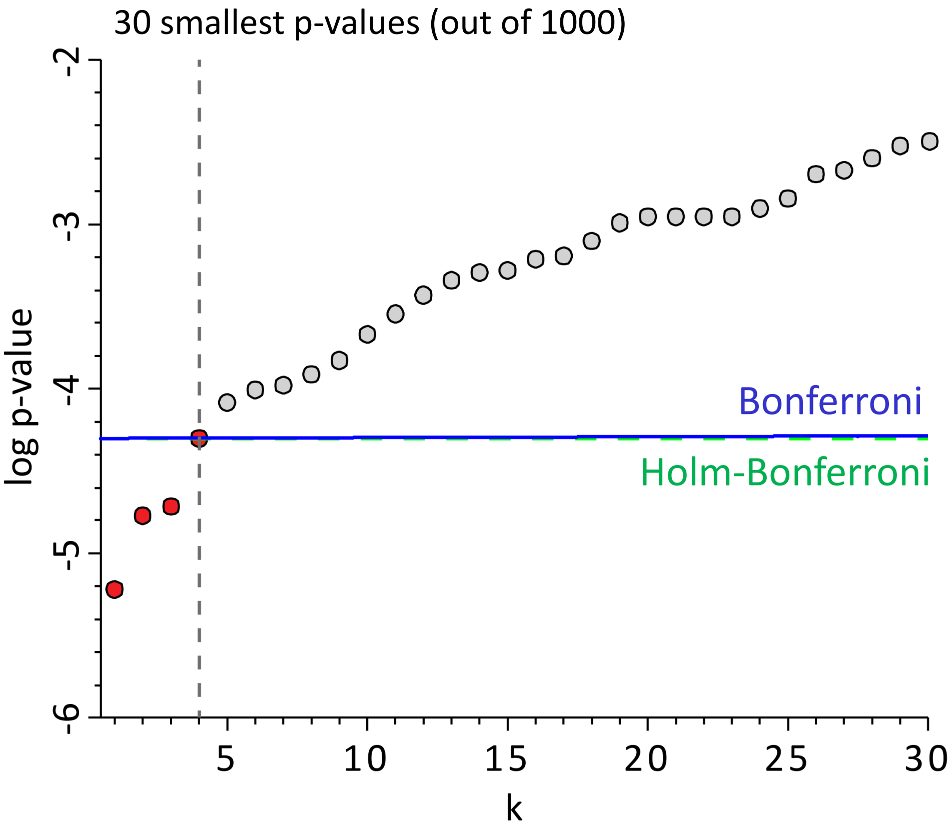
Holm-Bonferroni method controls FWER



k	p	α	$\frac{\alpha}{m}$	$\frac{\alpha}{m-k+1}$
1	0.003	0.05	0.01	0.01
2	0.005	0.05	0.01	0.0125
3	0.012	0.05	0.01	0.017
4	0.04	0.05	0.01	0.025
5	0.058	0.05	0.01	0.05

Holm-Bonferroni method

	No correction	Bonferroni	HB
α	0.05	5×10^{-5}	5×10^{-5}
FPR	0.058	0	0
FNR	0	0.87	0.87



Holm-Bonferroni			
	H_0 true	H_0 false	Total
Significant	0	4	4
Not significant	970	26	996
Total	970	30	1000

False discovery rate

$$FPR = \frac{FP}{D}$$

False discovery rate

False positive rate

$$FPR = \frac{FP}{m_0} = \frac{FP}{FP + TN}$$

The fraction of truly non-significant events we falsely marked as significant

$$FPR = \frac{56}{970} = 0.058$$

False discovery rate

$$FPR = \frac{FP}{D} = \frac{FP}{FP + TP}$$

The fraction of discoveries that are false

$$FDR = \frac{56}{86} = 0.65$$

No correction			
	H_0 true	H_0 false	Total
Significant	$FP = 56$	$TP = 30$	86
Not significant	$TN = 914$	$FN = 0$	914
Total	970	30	1000

Benjamini-Hochberg method

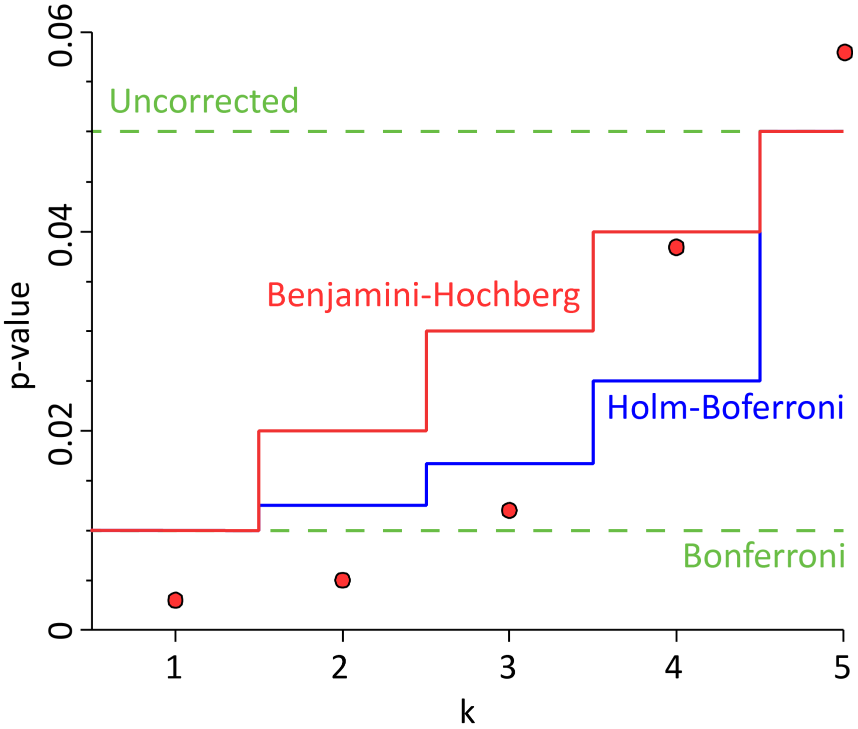
Sort p-values

$$p_{(1)}, p_{(2)}, \dots, p_{(m)}$$

Find the largest k, such that

$$p_{(k)} \leq \frac{k}{m}\alpha$$

Reject all null hypotheses for $i = 1, \dots, k$

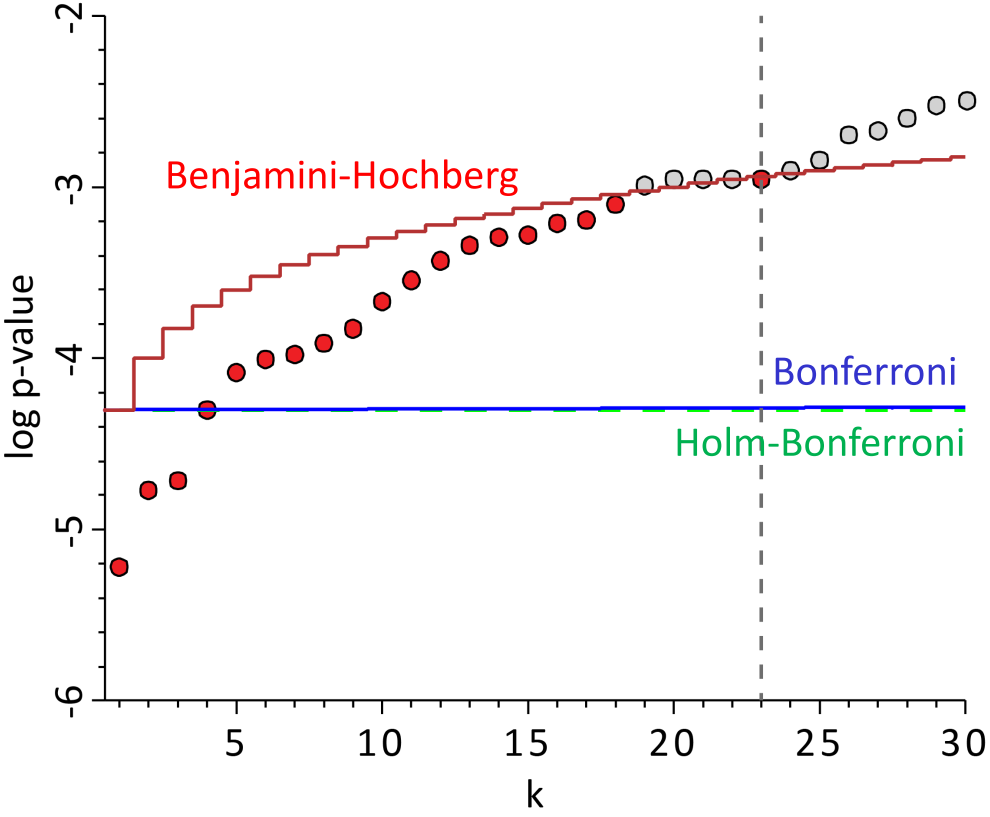


k	p	α	$\frac{\alpha}{m}$	$\frac{\alpha}{m-k+1}$	$\frac{k}{m}\alpha$
1	0.003	0.05	0.01	0.01	0.01
2	0.005	0.05	0.01	0.0125	0.02
3	0.012	0.05	0.01	0.017	0.03
4	0.038	0.05	0.01	0.025	0.04
5	0.058	0.05	0.01	0.05	0.05

Benjamini-Hochberg method controls FDR

Benjamini-Hochberg method

	No correction	Bonferroni	HB	BH
α	0.05	5×10^{-5}	3.7×10^{-5}	0.0011
FPR	0.058	0	0	0.0021
FNR	0	0.87	0.87	0.30
FDR	0.65	0	0	0.087



Benjamini-Hochberg			
	H_0 true	H_0 false	Total
Significant	2	21	23
Not significant	968	9	977
Total	970	30	1000

Controlling FWER and FDR

Holm-Bonferroni

controls FWER

$$FWER = \Pr(FP \geq 1)$$

Controlling FWER - guaranteed

$$FWER \leq \alpha'$$

Benjamini-Hochberg

controls FDR

$$FDR = \frac{FP}{FP + TP}$$

FDR is a random variable

Controlling FDR - guaranteed

$$E[FDR] \leq \alpha$$

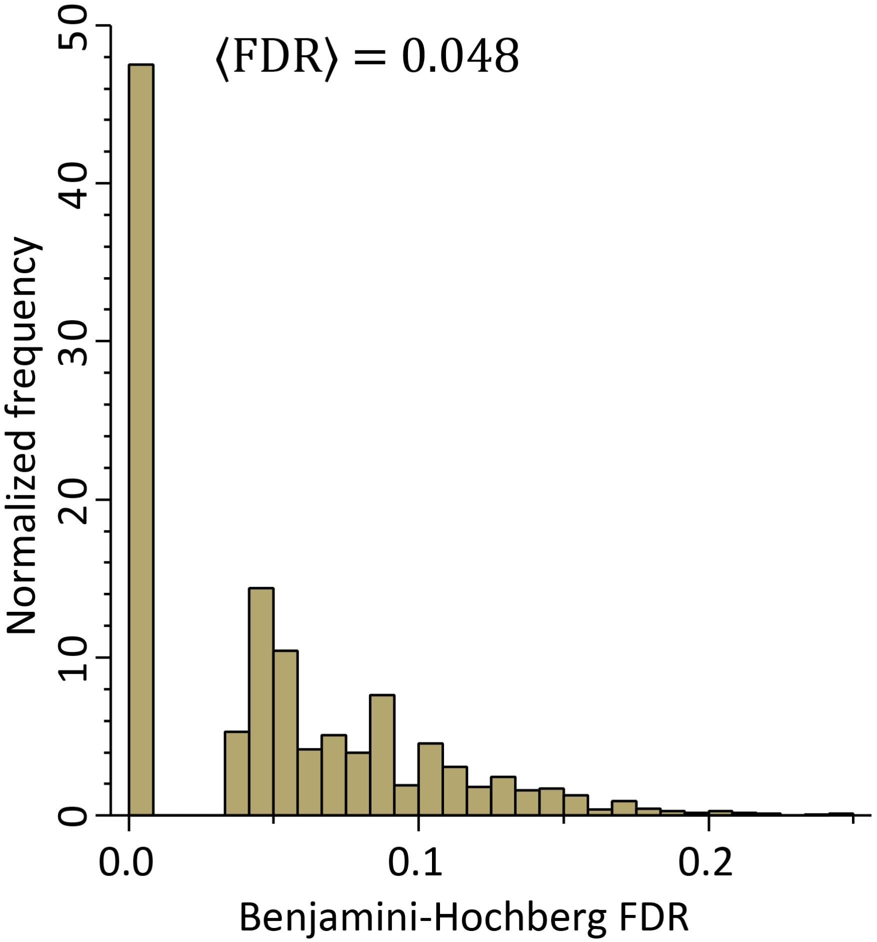
Benjamini-Hochberg procedure controls FDR

Controlling FDR

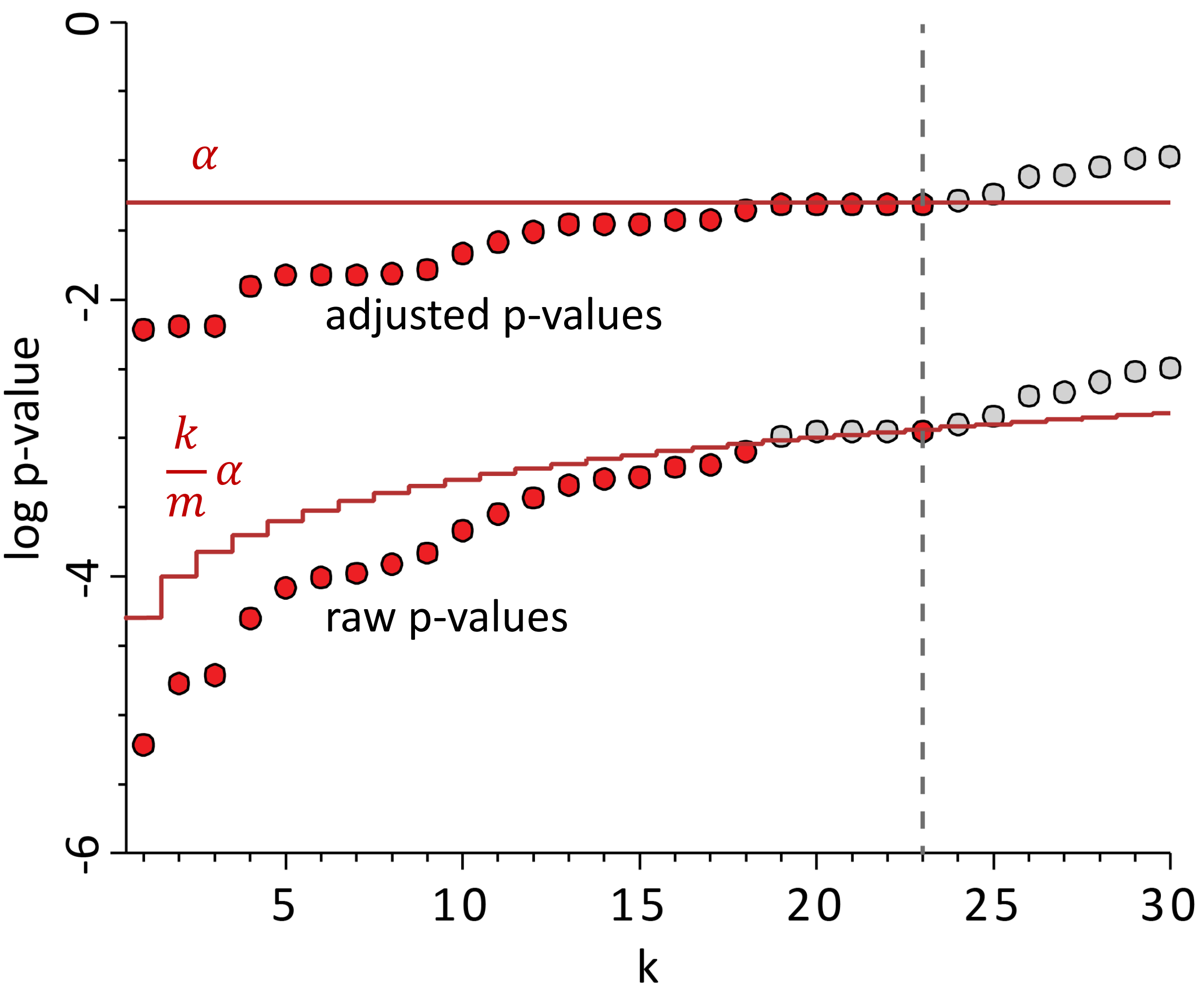
$$E[FDR] \leq \alpha$$

$E[FDR]$ can be approximated by the mean over many experiments

Bootstrap: generate test data 10,000 times, perform 1000 t-tests for each set and find FDR for BH procedure



Adjusted p-values



p-values can be "adjusted", so they compare directly with α, and not $\frac{k}{m}\alpha$

Problem: adjusted p-value does not express any probability

Despite their popularity, I recommend against using adjusted p-values

How to do this in R

```
# Read generated data
> d = read.table("http://tiny.cc/two_hypotheses", header=TRUE)
> p = d$p

# Holm-Bonferroni procedure
> p.adj = p.adjust(p, "holm")
> p[which(p.adj < 0.05)]
[1] 1.476263e-05 2.662440e-05 3.029839e-05

# Benjamini-Hochberg procedure
> p.adj = p.adjust(p, "BH")
> p[which(p.adj < 0.05)]
 [1] 1.038835e-03 6.670798e-04 1.050547e-03 1.476263e-05 5.271367e-04
 [6] 3.503370e-04 9.664789e-04 1.068863e-03 7.995860e-04 5.404476e-04
[11] 9.681321e-04 1.580069e-04 1.732747e-04 3.159954e-04 2.662440e-05
[16] 4.709732e-04 1.517964e-04 2.873971e-04 3.258726e-04 4.087615e-04
[21] 3.029839e-05 9.320438e-04 1.713309e-04 2.863402e-04 4.082322e-04
```

Estimating false discovery rate

Control and estimate

Controlling FDR

1. Fix acceptable FDR limit, α, beforehand
2. Find a thresholding rule, so that

$$E[FDR] \leq \alpha$$

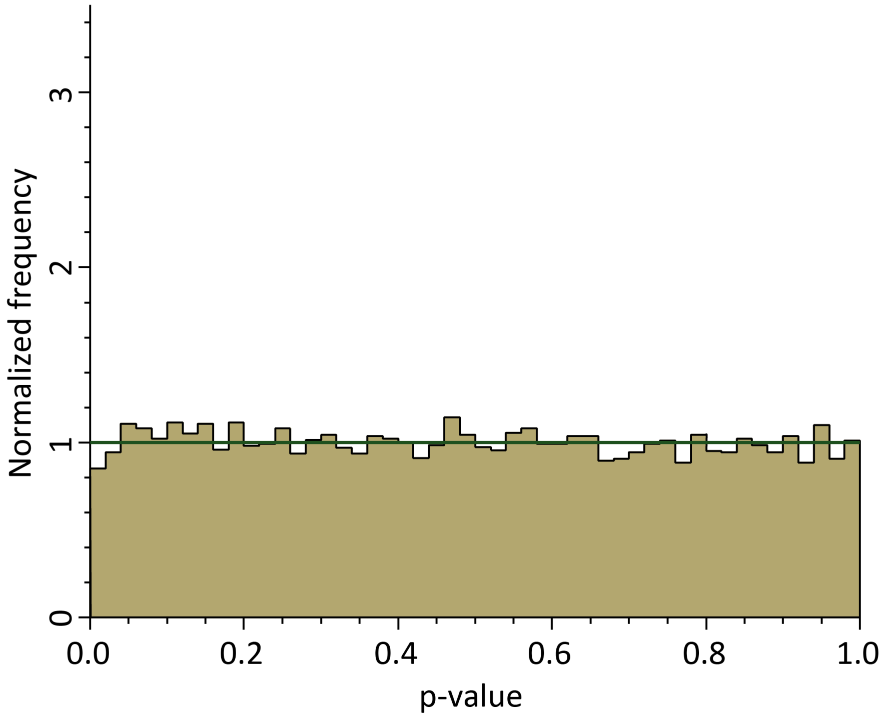
Estimating FDR

For each p-value, p_i, form a point estimate of FDR,

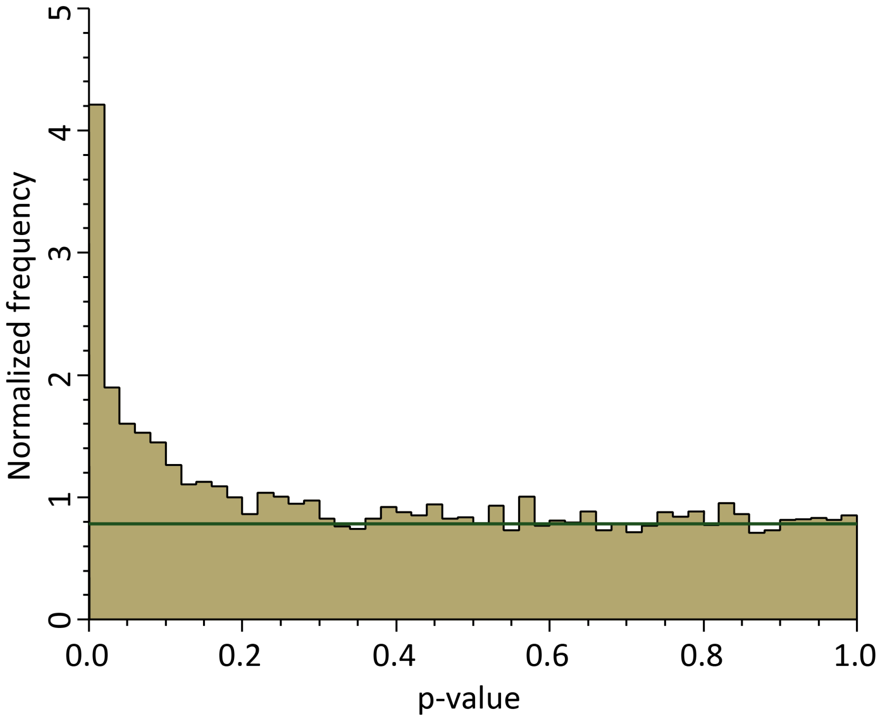
$$\widehat{FDR}(p_i)$$

P-value distribution

100% null

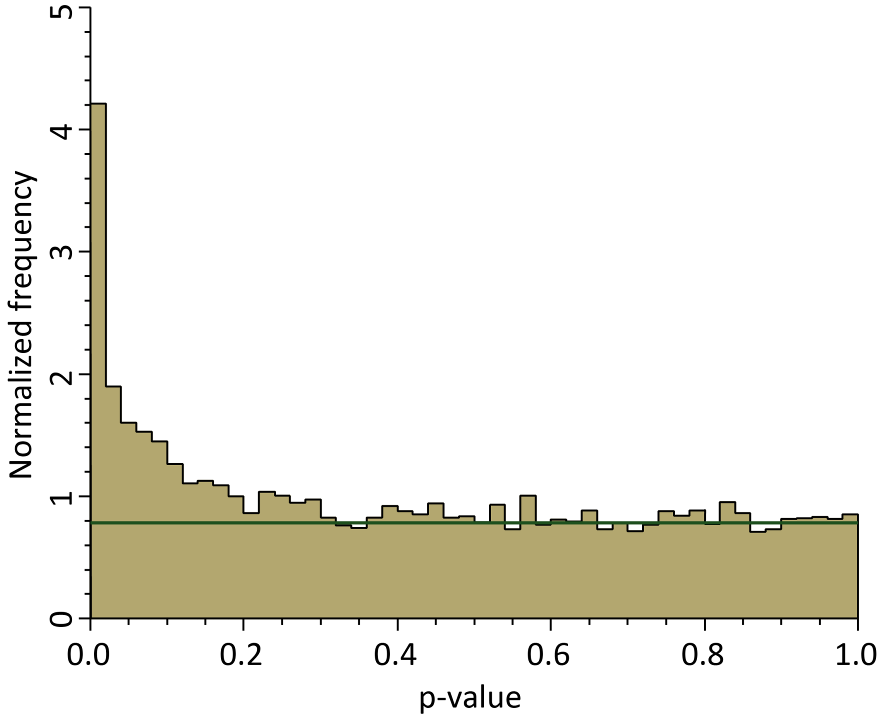


Data set 2
80% null, 20% alternative

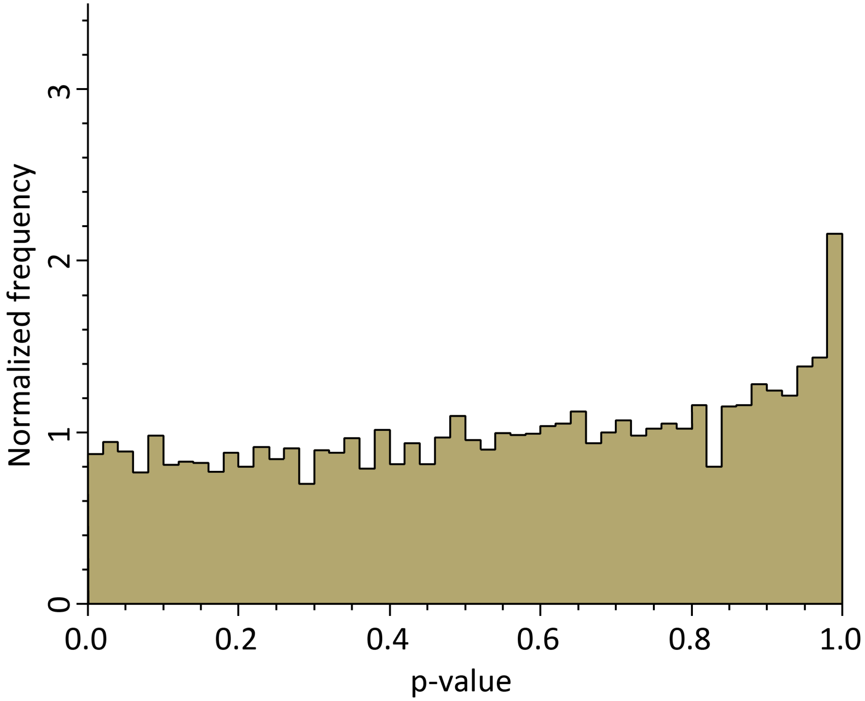


P-value distribution

Good

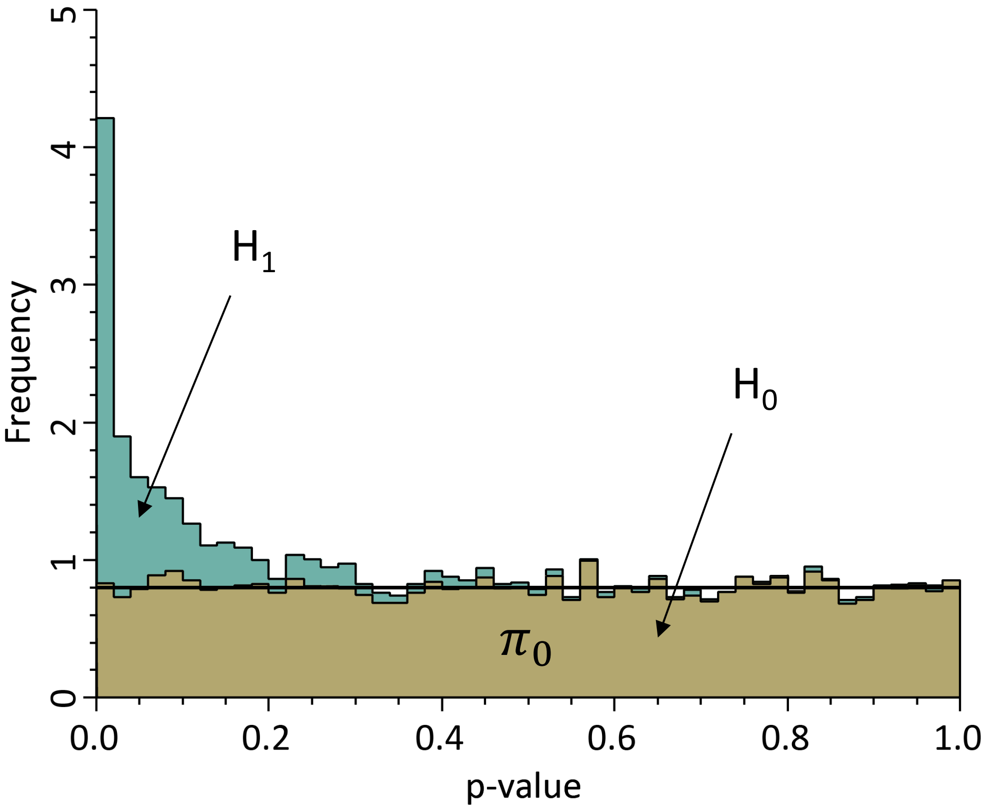


Bad!



Definition of π_0

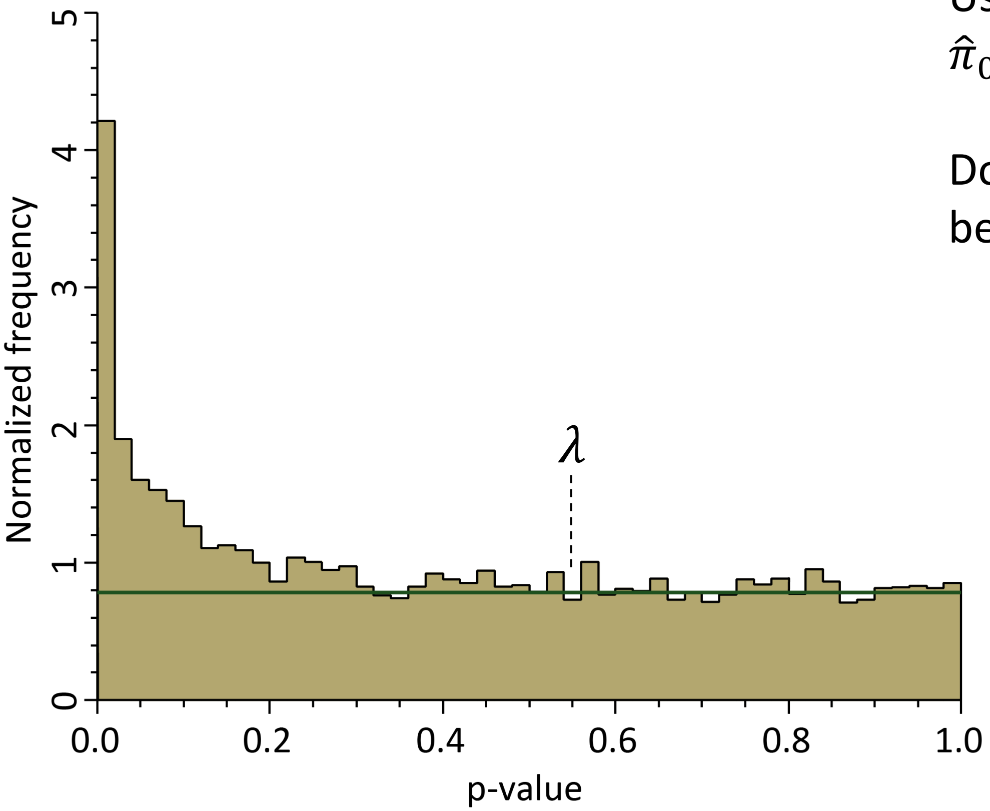
80% null, 20% alternative



Proportion of null tests

$$\pi_0 = \frac{\#\{\text{non-significant tests}\}}{\#\{\text{all tests}\}}$$

Storey method

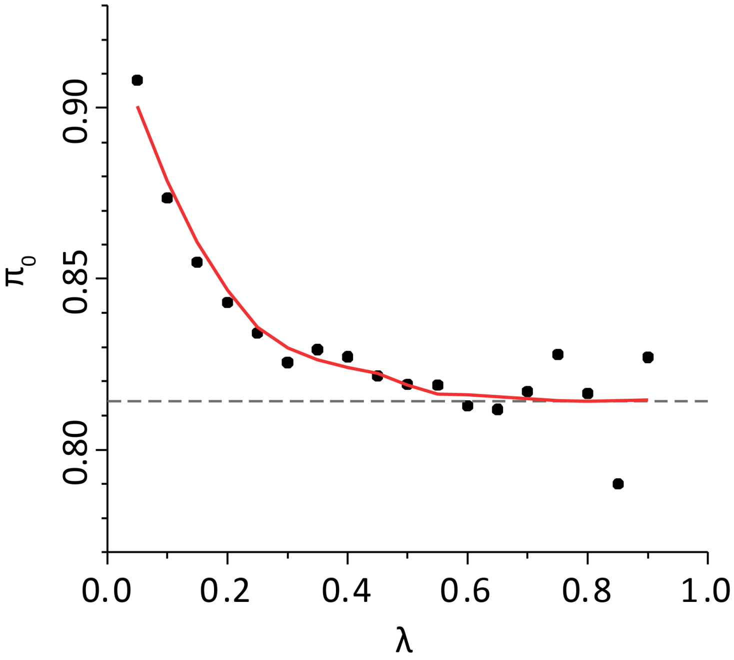


Storey, J.D., 2002, *JR Statist Soc B*, **64**, 479

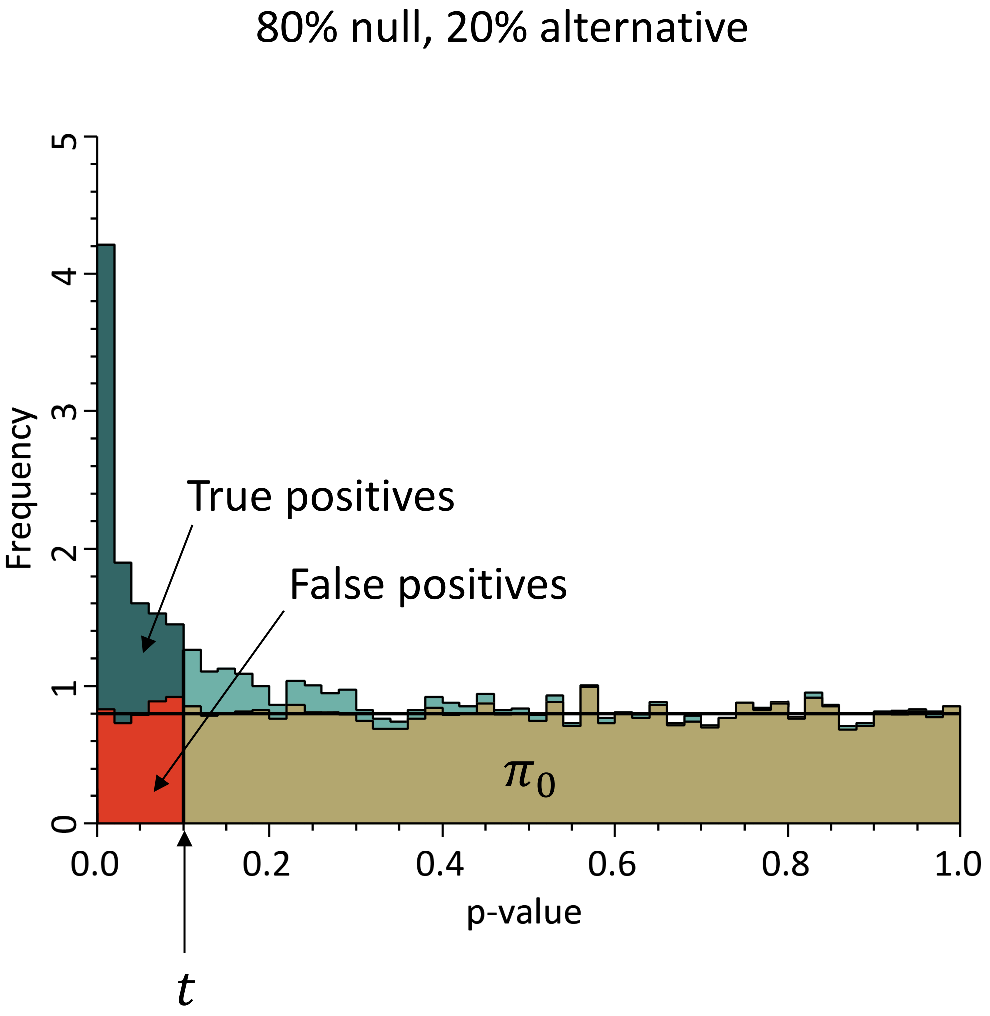
Estimate π_0

Use the histogram for $p > \lambda$ to estimate $\hat{\pi}_0(\lambda)$

Do it for all $0 < \lambda < 1$ and then find the best $\hat{\pi}_0$



Point estimate of FDR



Point estimate, $\boldsymbol{FDR(t)}$

Arbitrary limit t, every $p_i < t$ is significant. No. of significant tests is

$$R(t) = \#\{p_i < t\}$$

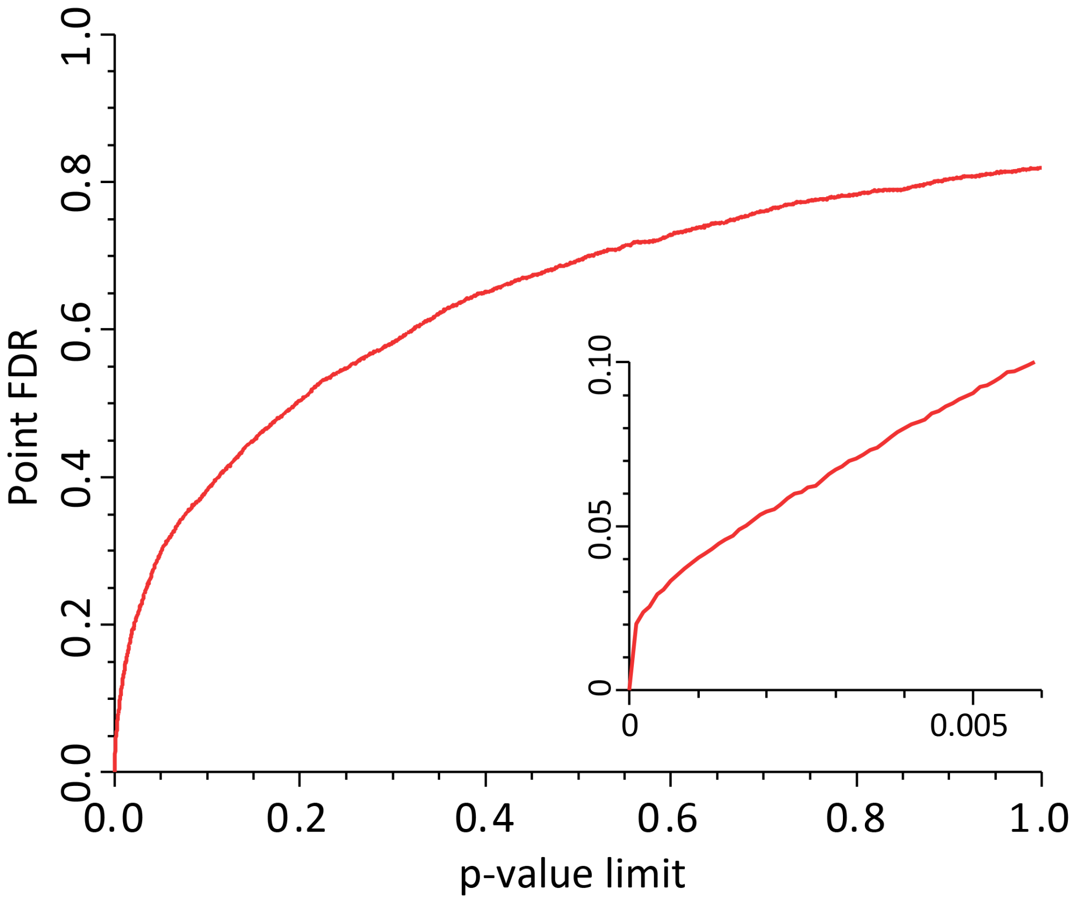
No. of false positives is

$$FP(t) = t\pi_0 m$$

Hence,

$$FDR(t) = \frac{t\pi_0 m}{R(t)}$$

Storey method



Point estimate of FDR

This is the so-called q-value:

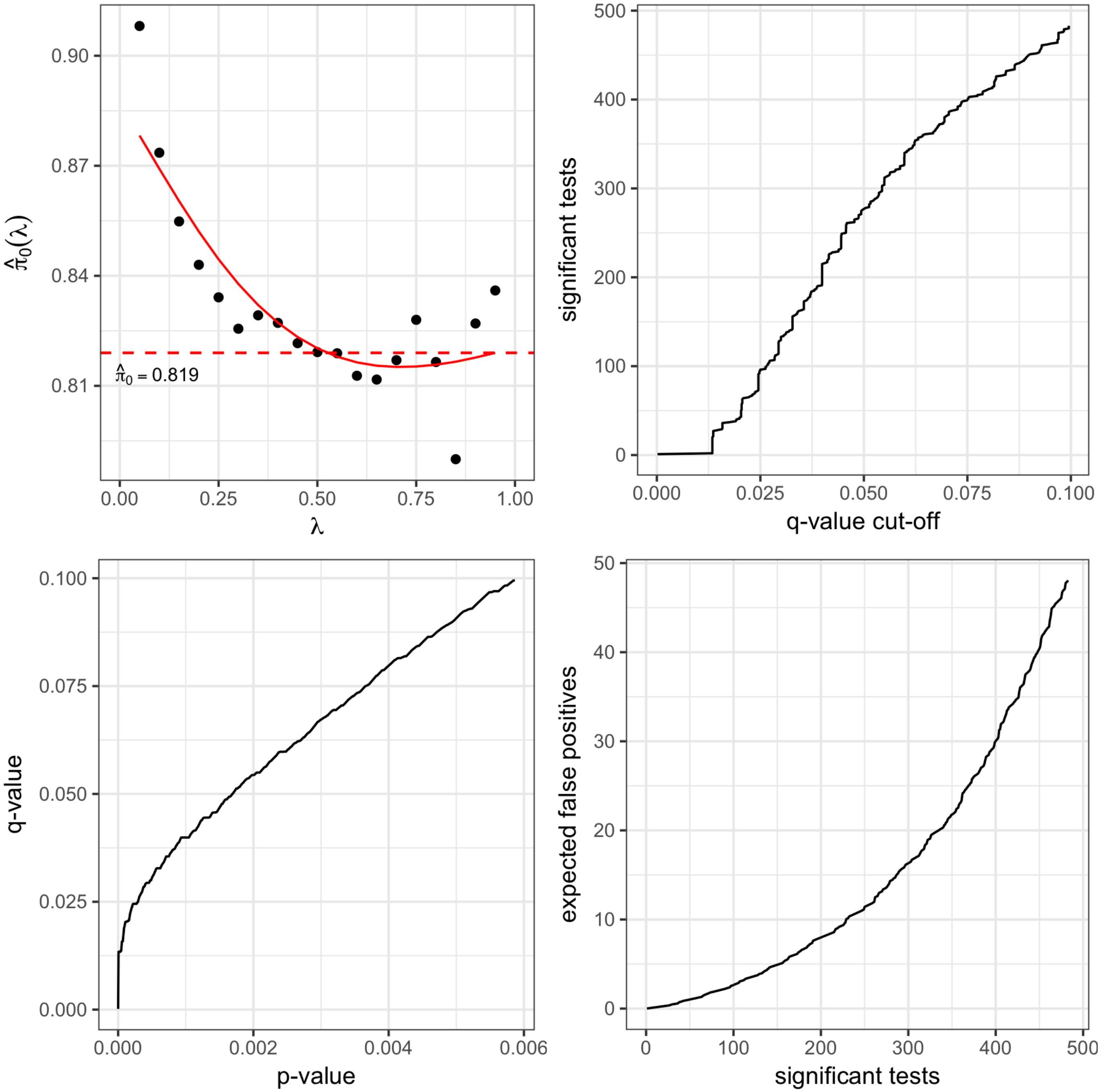
$$q(p_i) = \min_{t \geq p_i} FDR(t)$$

If monotonic

$$q(p_i) = FDR(p_i)$$

How to do this in R

```
> library(qvalue)

# Read data set 1
> pvalues = read.table('http://tiny.cc/multi_FDR', header=TRUE)
> p = pvalues$p

# Benjamini-Hochberg limit
> p.adj = p.adjust(p, method='BH')
> sum(p.adj <= 0.05)
[1] 216

# q-values
> qobj = qvalue(p)
> q = qobj$qv
> summary(qobj)

pi0:	0.8189884

Cumulative number of significant calls:
          <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1    <1
p-value       40    202   611    955  1373 2138 10000
q-value        0      1     1     96   276  483 10000
local FDR      0      1     3     50   141  278  5915

> plot(qobj)
> hist(qobj)
```

0.90
0.87
0.84
0.81
$\hat{\pi}_0(\lambda)$
$\hat{\pi}_0 = 0.819$
0.00
0.25
0.50
0.75
1.00
λ
500
400
300
200
100
0
significant tests
0.000
0.025
0.050
0.075
0.100
q-value cut-off
0.100
0.075
0.050
0.025
0.000
q-value
0.000
0.002
0.004
0.006
p-value
50
40
30
20
10
0
expected false positives
0
100
200
300
400
500
significant tests

Interpretation of q-value

No.	ID	p-value	q-value
...	...	...	...
100	9249	0.000328	0.0266
101	8157	0.000328	0.0266
102	8228	0.000335	0.0269
103	8291	0.000338	0.0269
104	8254	0.000347	0.0272
105	8875	0.000348	0.0272
106	8055	0.000353	0.0273
107	8235	0.000375	0.0284
108	8148	0.000376	0.0284
109	8236	0.000381	0.0284
110	8040	0.000382	0.0284
...	...	...	...

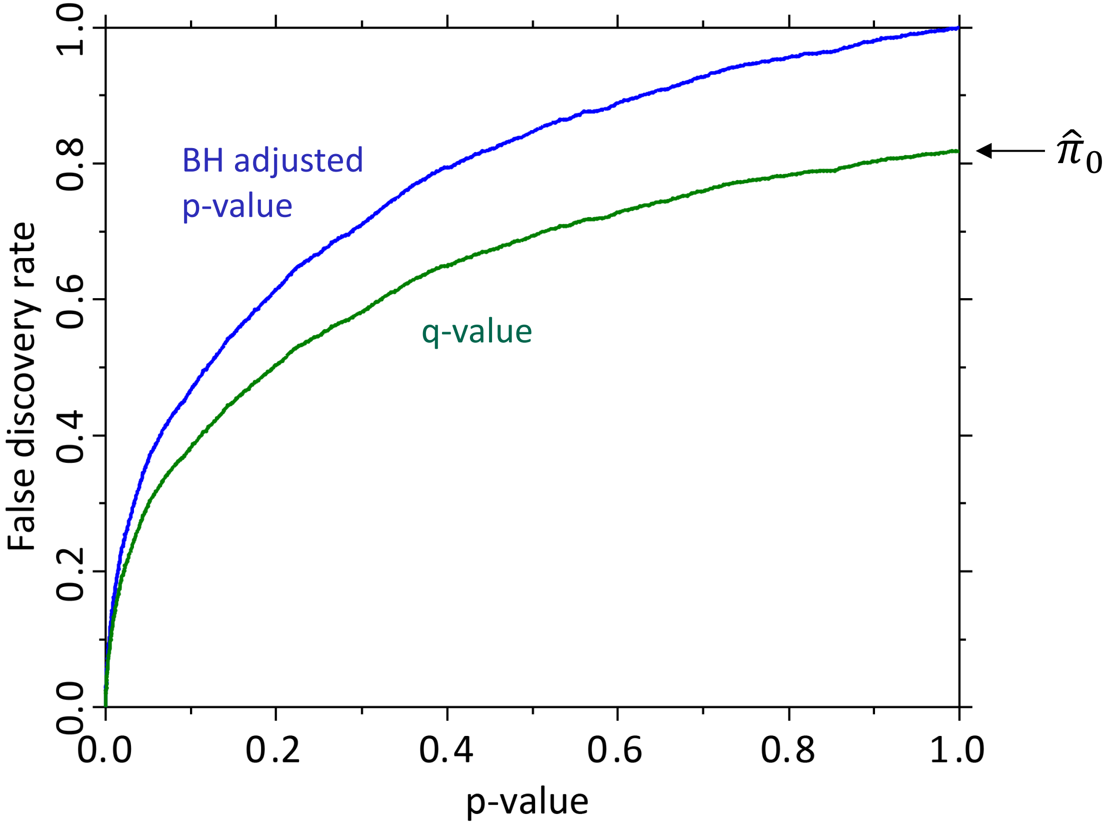
There are 106 tests with $q \leq 0.0273$

Expect 2.73% of false positives among these tests

Expect ~ 3 false positives if you set a limit of $q \leq 0.0273$ or $p \leq 0.00353$

q-value tells you how many false positives you should expect after choosing a significance limit

Q-values vs Benjamini-Hochberg



When $\hat{\pi}_0 = 1$, both methods give the same result.

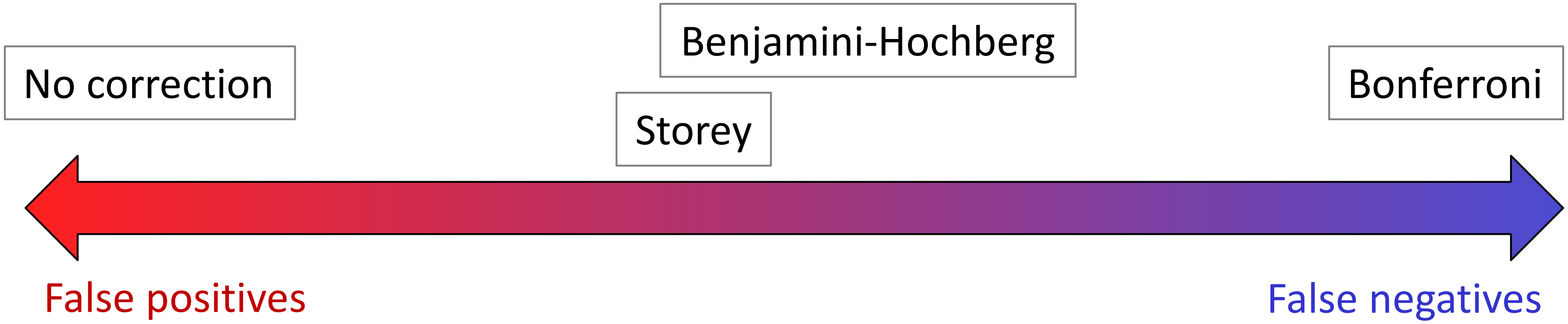
For the same FDR, Storey's method provides more significant p-values.

Hence, it is more powerful, especially for small $\hat{\pi}_0$.

But this depends on how good the estimate of $\hat{\pi}_0$ is.

$\hat{\pi}_0$ - estimate of the proportion of null (non-significant) tests

Which multiple-test correction should I use?



No correction

Storey

Benjamini-Hochberg

Bonferroni

False positives

False negatives

False positive

“Discover” effect where there is no effect

Can be tested in follow-up experiments

Not hugely important in small samples

Impossible to manage in large samples

False negative

Missed discovery

Once you’ve missed it, it’s gone

Multiple test procedures: summary

Method	Controls	Advantages	Disadvantages	Recommendation
No correction	FPR	False negatives not inflated	Can result in $FP \gg TP$	Small samples, when the cost of FN is high
Bonferroni	FWER	None	Lots of false negatives	Do not use
Holm-Bonferroni	FWER	Slightly better than Bonferroni	Lots of false negatives	Appropriate only when you want to guard against any false positives
Benjamini-Hochberg	FDR	Good trade-off between false positives and negatives	On average, α of your positives will be false	Better in large samples
Storey	--	More powerful than BH, in particular for small $\hat{\pi}_0$	Depends on a good estimate of $\hat{\pi}_0$	The best method, gives more insight into FDR

UNIVERSITY OF DUNDEE

THE BARTON GROUP

01000100010000010101010001000001

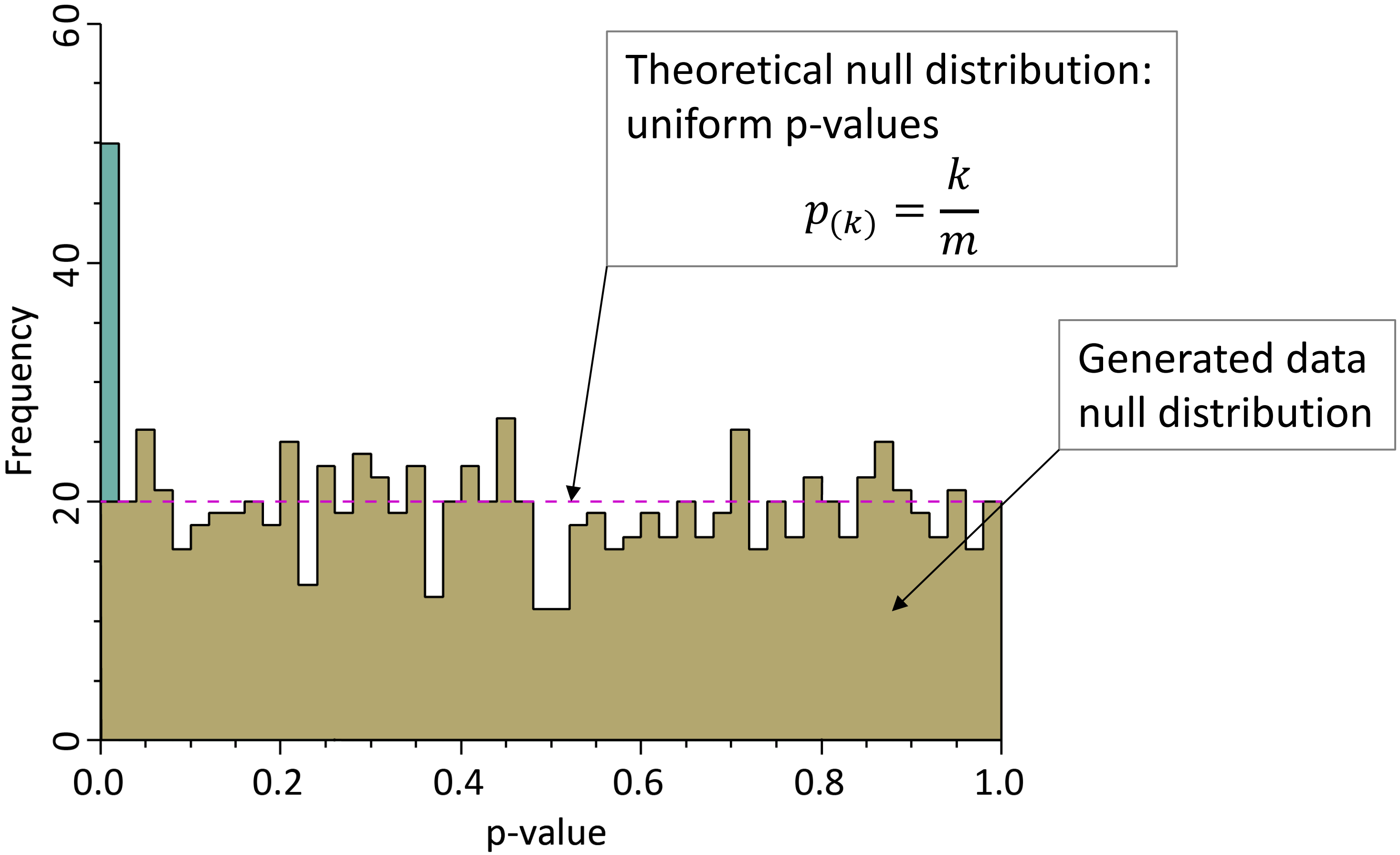
The Data Analysis Group

Hand-outs available at
http://tiny.cc/statlec

Centre for Gene Regulation & Expression

wellcome trust

Benjamini-Hochberg method

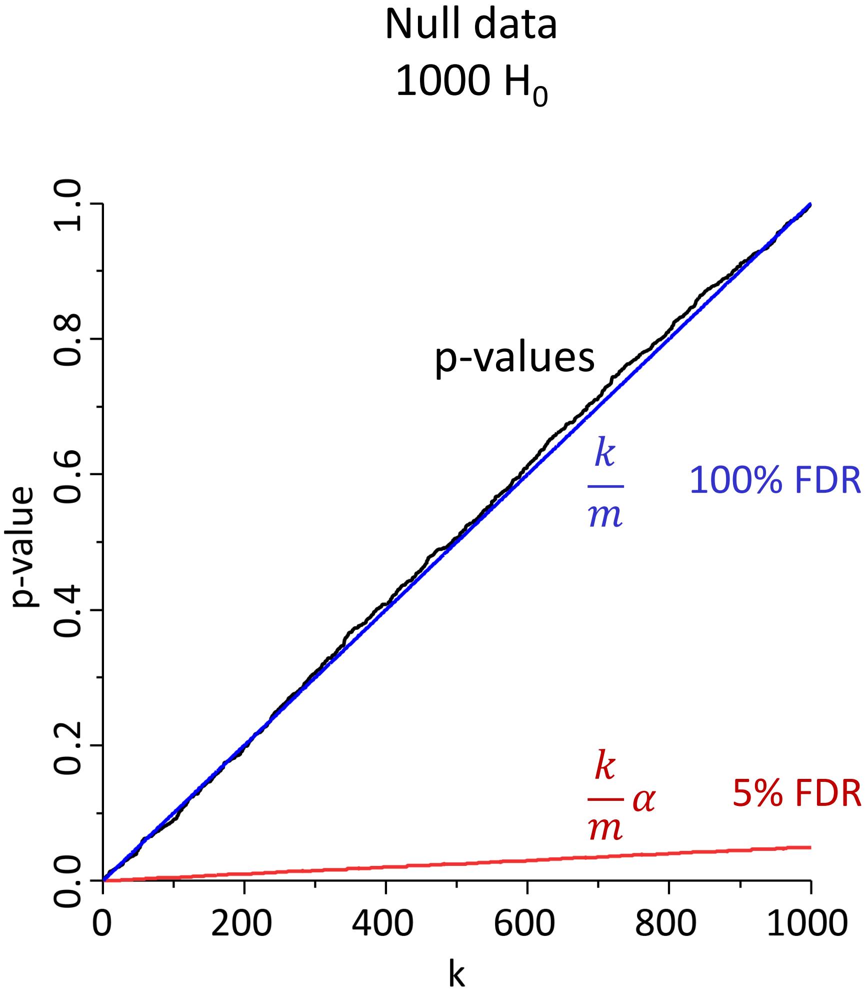


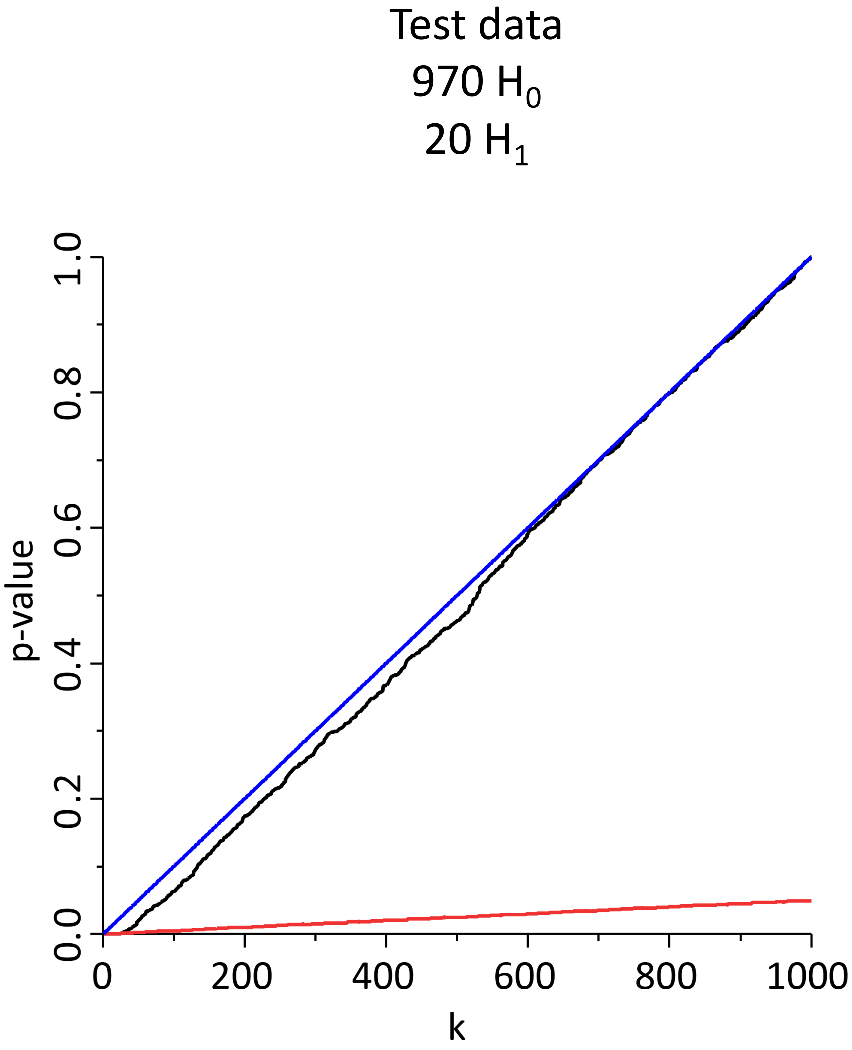
Theoretical null distribution: uniform p-values

$$p_{(k)} = \frac{k}{m}$$

Generated data null distribution

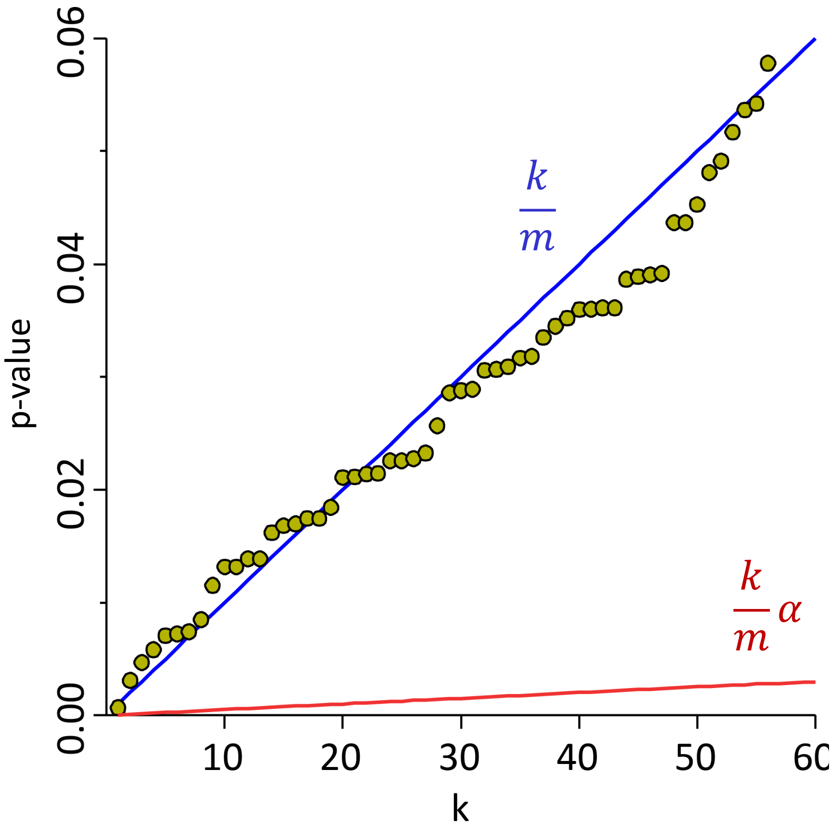
Benjamini-Hochberg method


Null data
1000 H_0

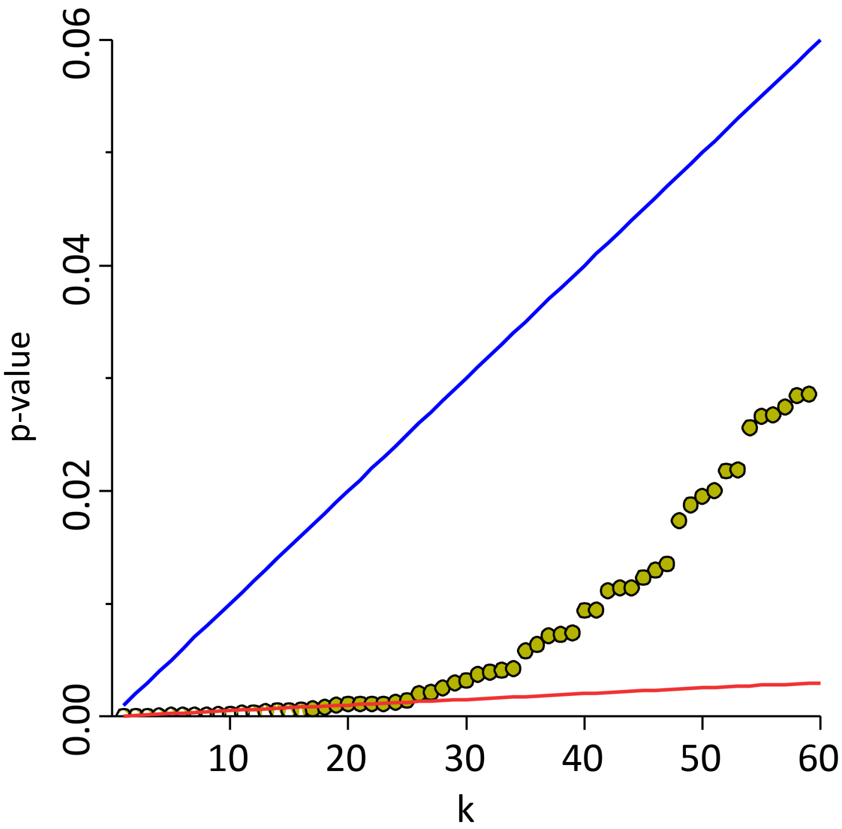

Test data
970 H_0
20 H_1

Benjamini-Hochberg method

Null data
1000 H_0



Test data
970 H_0
20 H_1


p-value
k