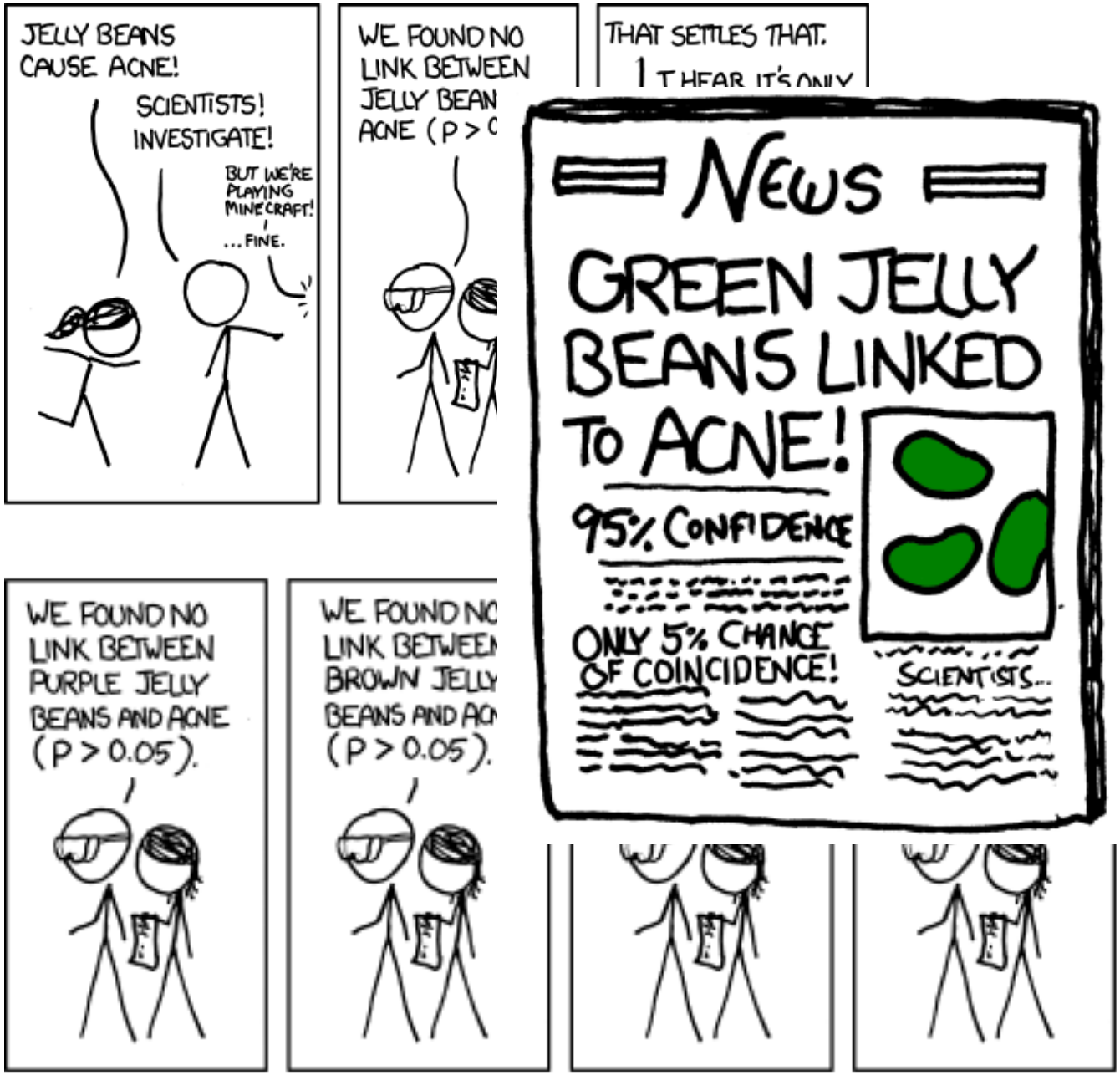
12. Multiple test corrections

"Natural selection is a mechanism for generating an exceedingly high degree of improbability”

Ronald A. Fisher

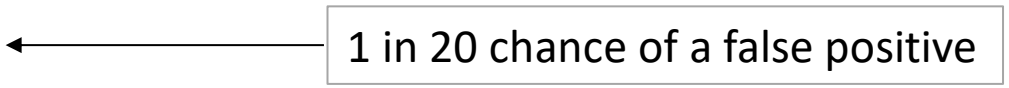

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEAN
ACNE (P >
THAT SETTLES THAT.
I T HEAR IT'S ONLY
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (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 > 0.05

Gene expression experiment

Differential expression: compare gene expression in two conditions (e.g. by t-test)

	gene_id	p.value
1	GENE00001	0.040503700
2	GENE00002	0.086340732
3	GENE00003	0.552768467
4	GENE00004	0.379025917
5	GENE00005	0.990102618
6	GENE00006	0.182729903
7	GENE00007	0.923285031
8	GENE00008	0.938615285
9	GENE00009	0.431912336
10	GENE00010	0.822697032
11	GENE00011	0.004911421
12	GENE00012	0.873463918
13	GENE00013	0.481156679
14	GENE00014	0.442047456
15	GENE00015	0.794117108
16	GENE00016	0.214535451
17	GENE00017	0.231943488
18	GENE00018	0.980911106
19	GENE00019	0.422162464
20	GENE00020	0.915841637

...

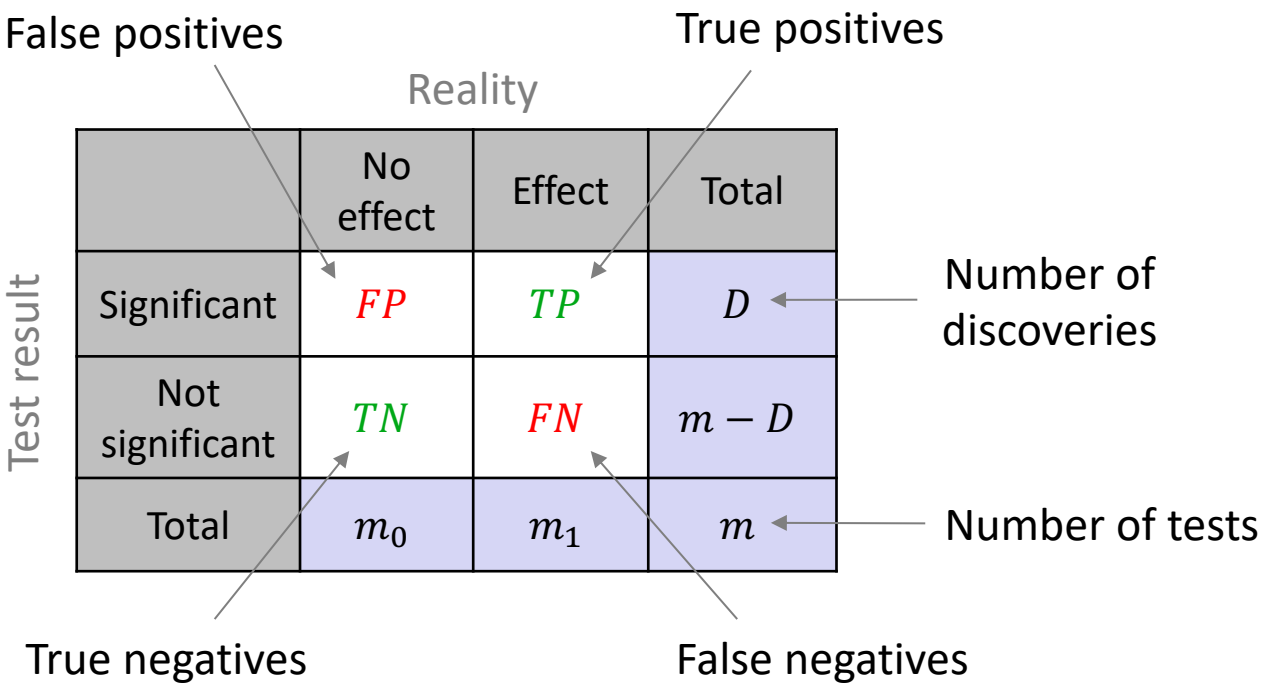


10,000 genes

~500 genes are “significant” by chance

Even unlikely result will eventually happen if you repeat your test many times

Lets perform a test m times



Family-wise error rate

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

Probability of winning at least once

Play lottery
Probability of winning is α

Events	Probability
👍	α
👎	$1-\alpha$

Events	Probability
👍👍	$\alpha \times \alpha$
👍👎	$\alpha \times (1-\alpha)$
👎👍	$(1-\alpha) \times \alpha$
👎👎	$(1-\alpha)^2$

← Probability of not winning at all

Probability of winning at least once
$1-(1-\alpha)^2$

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

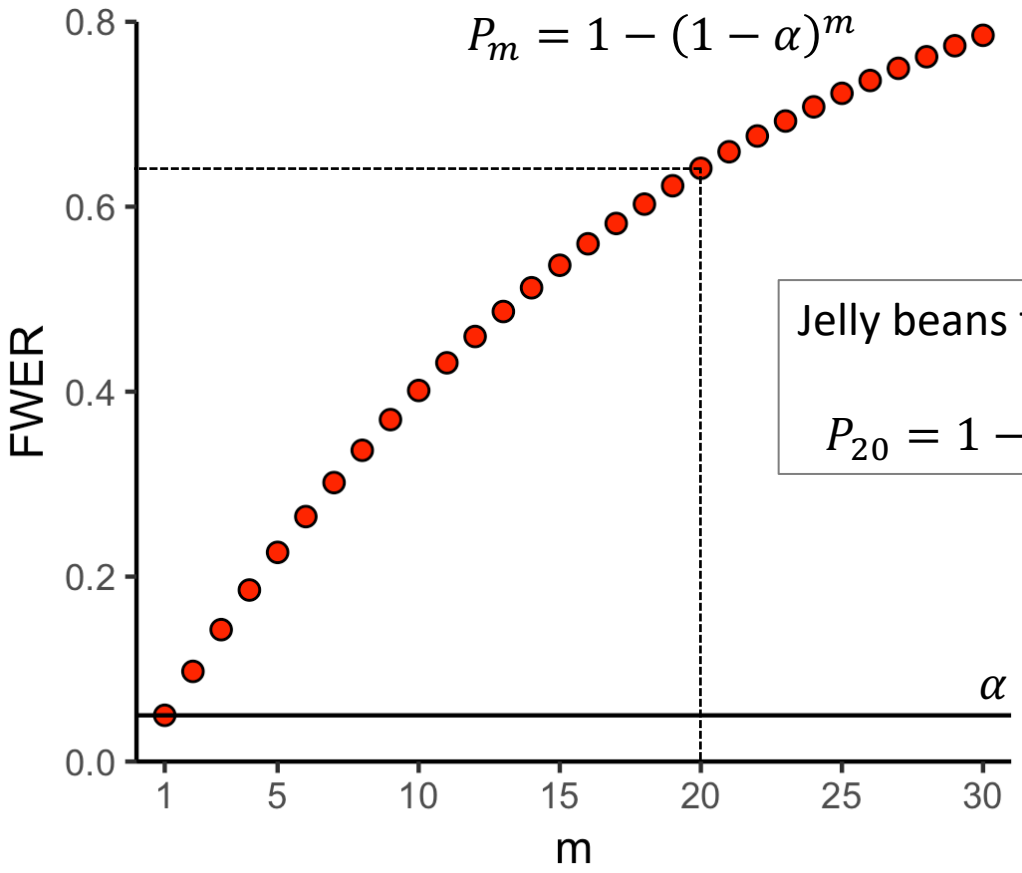
Probability of having at least one false positive in either test
$P_2 = 1 - (1 - \alpha)^2$

***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; $\alpha = 0.05$

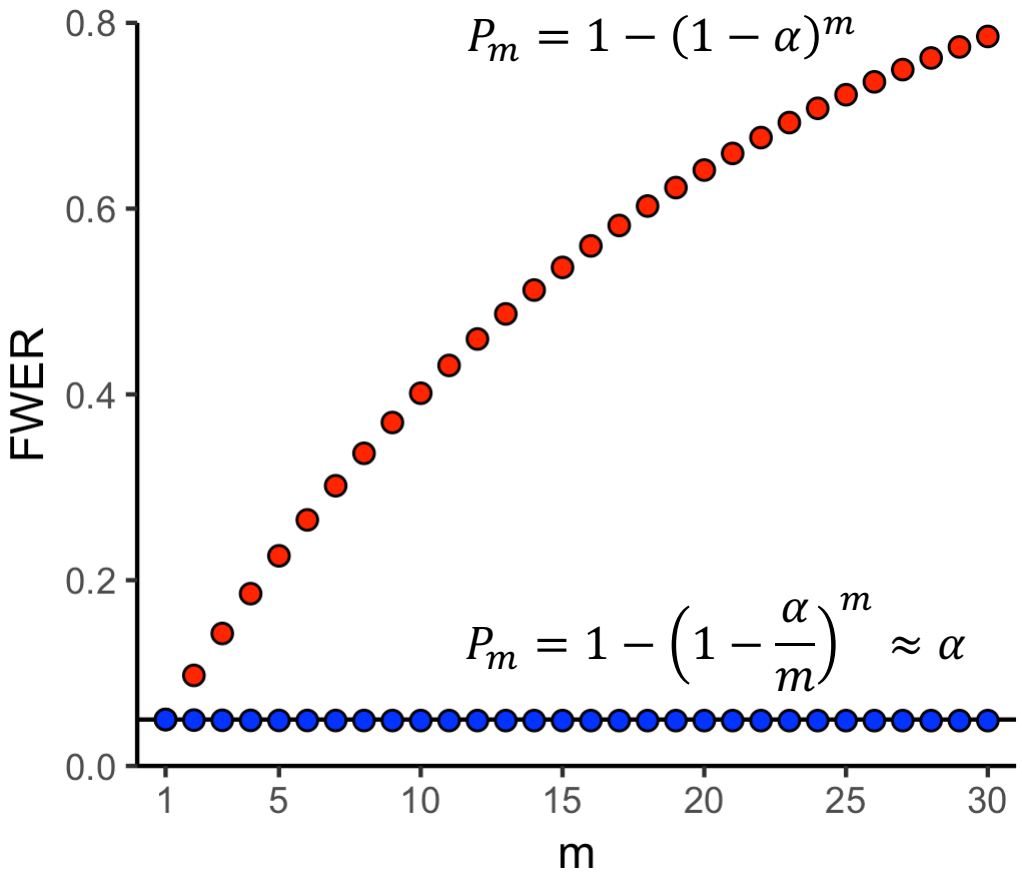


Jelly beans test, $m = 20, \alpha = 0.05$

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

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

$$FWER \leq \alpha.$$

Then, the FWER is controlled at level α.

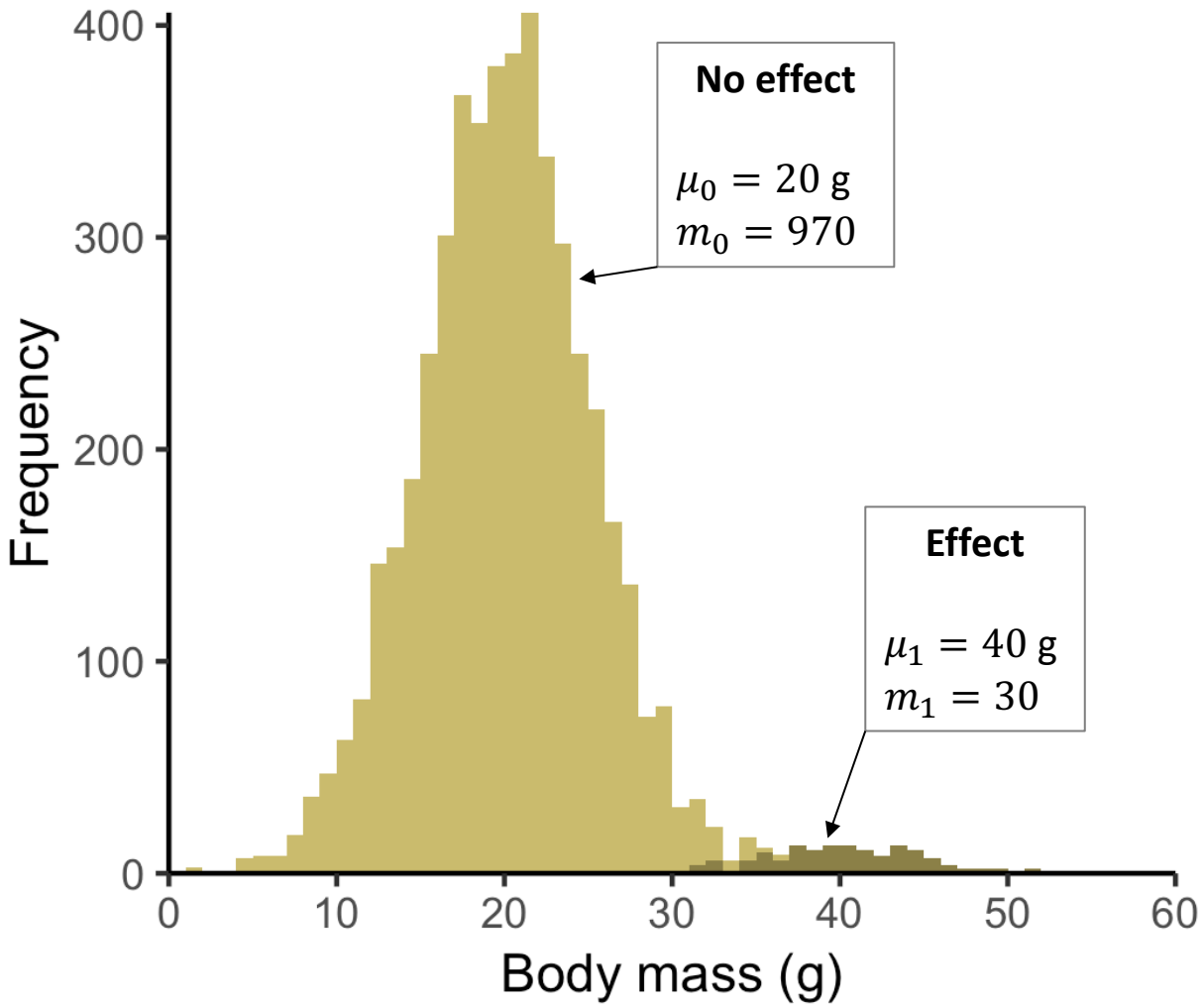
Bonferroni limit

Apply smaller limit per test, α'.

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

Test data (1000 independent experiments)

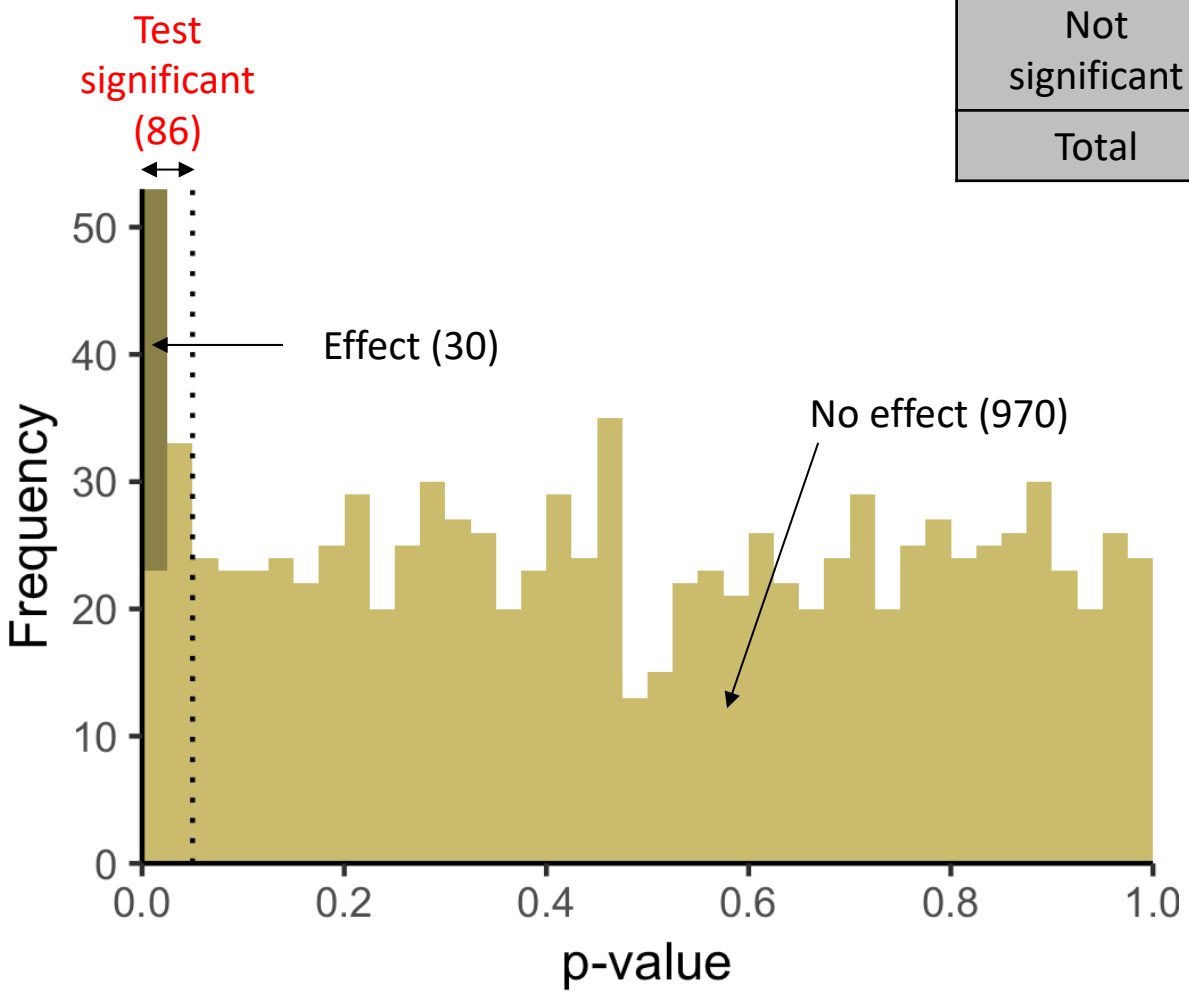
Random samples, size $n = 5$, from two normal distributions



- We have 1000 data sets
- Each set contains 5 values
- We perform one-sample t-test for each sample
- Null hypothesis: $\mu = 20$ g

1000 t-tests, $H_0: \mu = 20$ g

No correction			
	No effect	Effect	Total
Significant	56	30	86
Not significant	914	0	914
Total	970	30	1000



One sample t-test, H_0: $\mu = 20$ g

No correction			
	No effect	Effect	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			
	No effect	Effect	Total
Significant	56	30	86
Not significant	914	0	914
Total	970	30	1000

Bonferroni			
	No effect	Effect	Total
Significant	0	4	4
Not significant	970	26	996
Total	970	30	1000

Holm-Bonferroni method

Sort p-values

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

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

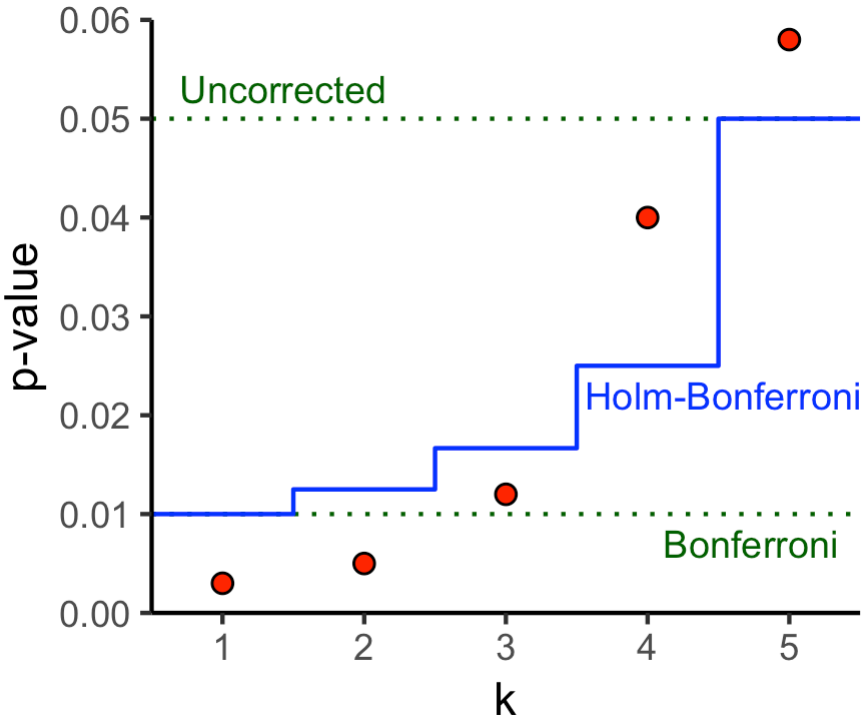
Reject (2) if $p_{(2)} \leq \frac{\alpha}{m-1}$

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

...

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

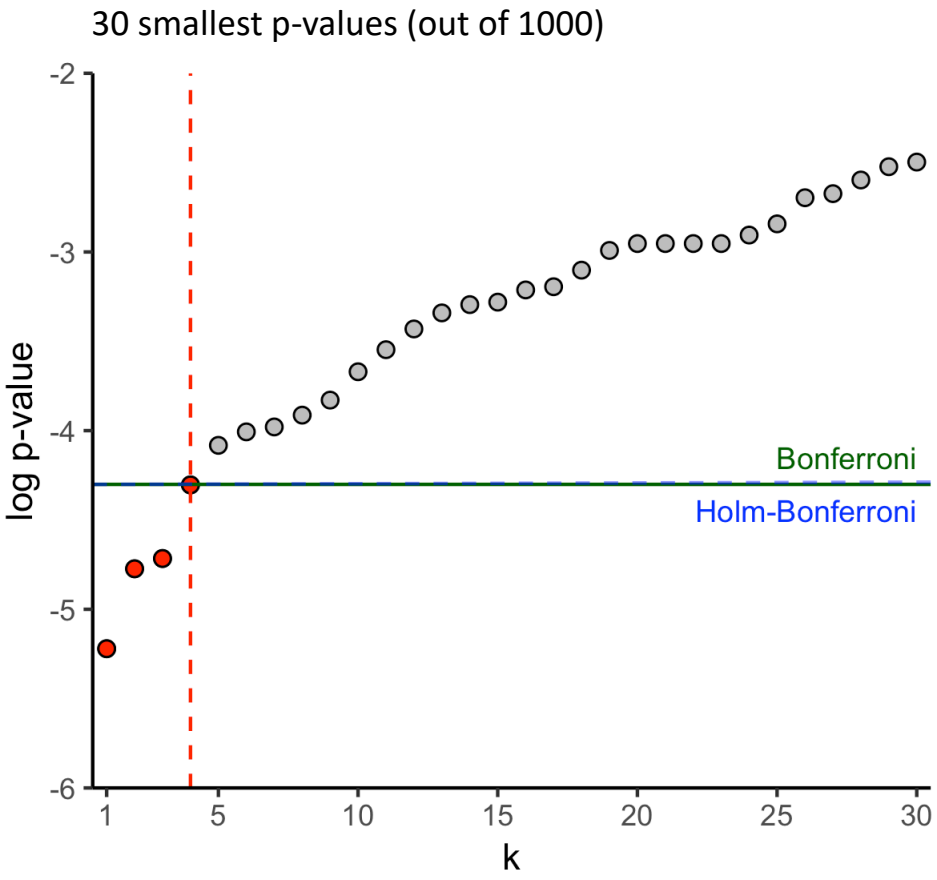
Holm-Bonferroni method controls FWER



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			
	No effect	Effect	Total
Significant	0	4	4
Not significant	970	26	996
Total	970	30	1000

False discovery rate

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

False discovery rate

False positive rate

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

The fraction of events with no effect 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			
	No effect	Effect	Total
Significant	$FP = 56$	$TP = 30$	86
Not significant	$TN = 914$	$FN = 0$	914
Total	970	30	1000

Benjamini-Hochberg method

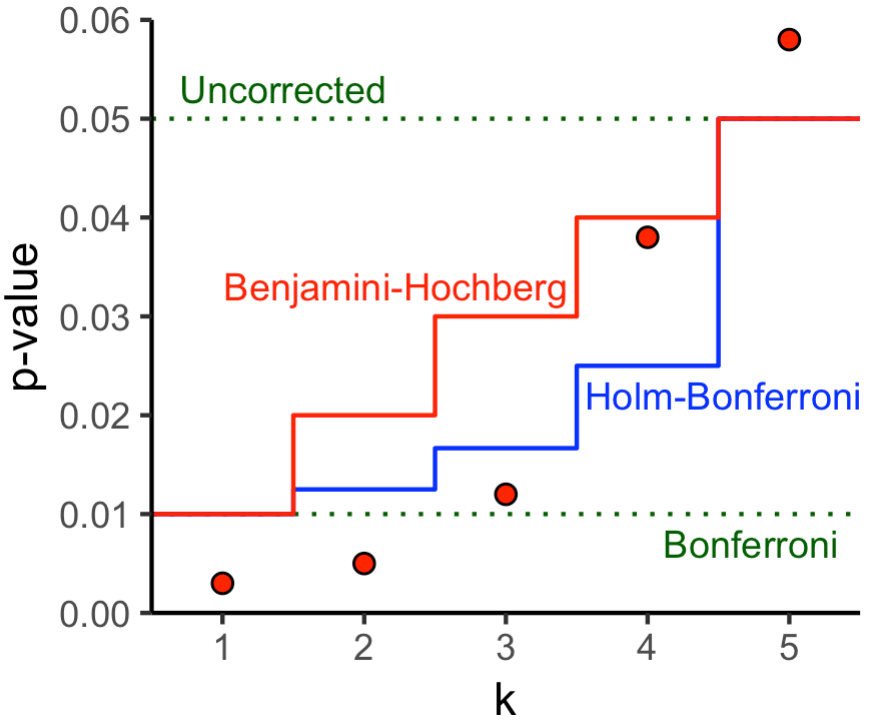
Sort p-values

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

Find the largest k, such that

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

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

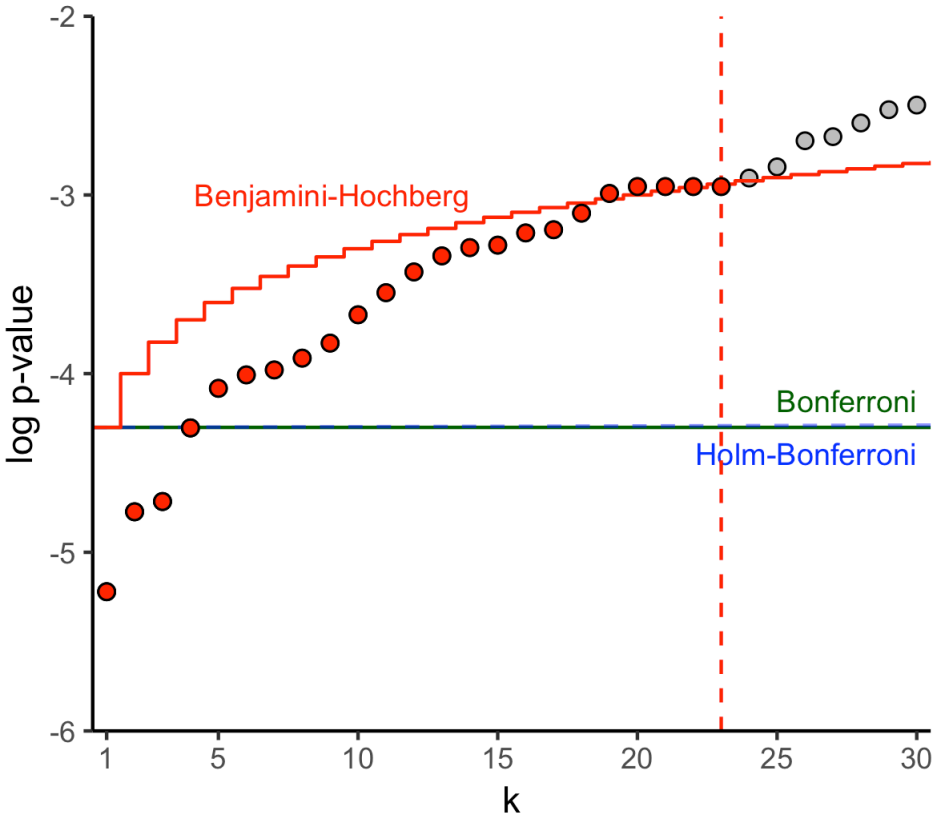


k	p	α	$\frac{\alpha}{m}$	$\frac{\alpha}{m-k+1}$	$\frac{k}{m}\alpha$
1	0.003	0.05	0.01	0.01	0.01
2	0.005	0.05	0.01	0.0125	0.02
3	0.012	0.05	0.01	0.017	0.03
4	0.038	0.05	0.01	0.025	0.04
5	0.058	0.05	0.01	0.05	0.05

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			
	No effect	Effect	Total
Significant	2	21	23
Not significant	968	9	977
Total	970	30	1000

Controlling FWER and FDR

Holm-Bonferroni
controls FWER
(family-wise error rate)

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

Controlling FWER - guaranteed

$$FWER \leq \alpha$$

Benjamini-Hochberg
controls FDR
(false discovery rate)

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

Controlling FDR - guaranteed

$$\overline{FDR} \leq \alpha$$

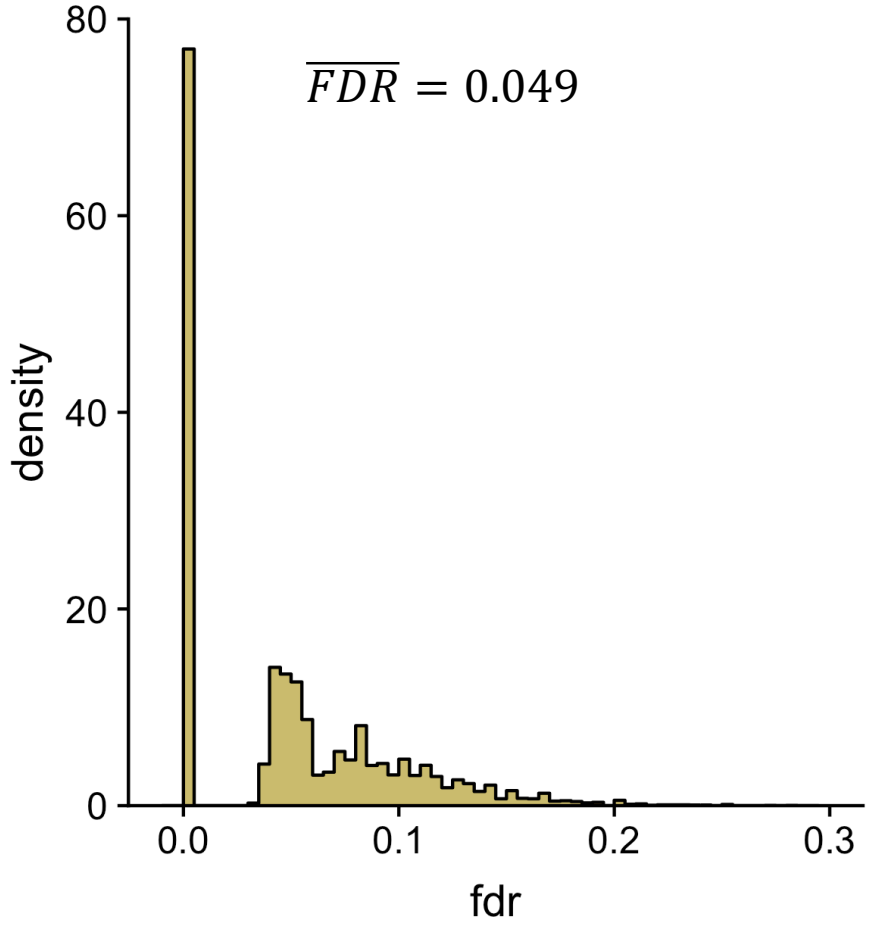
Benjamini-Hochberg procedure controls FDR

Controlling FDR

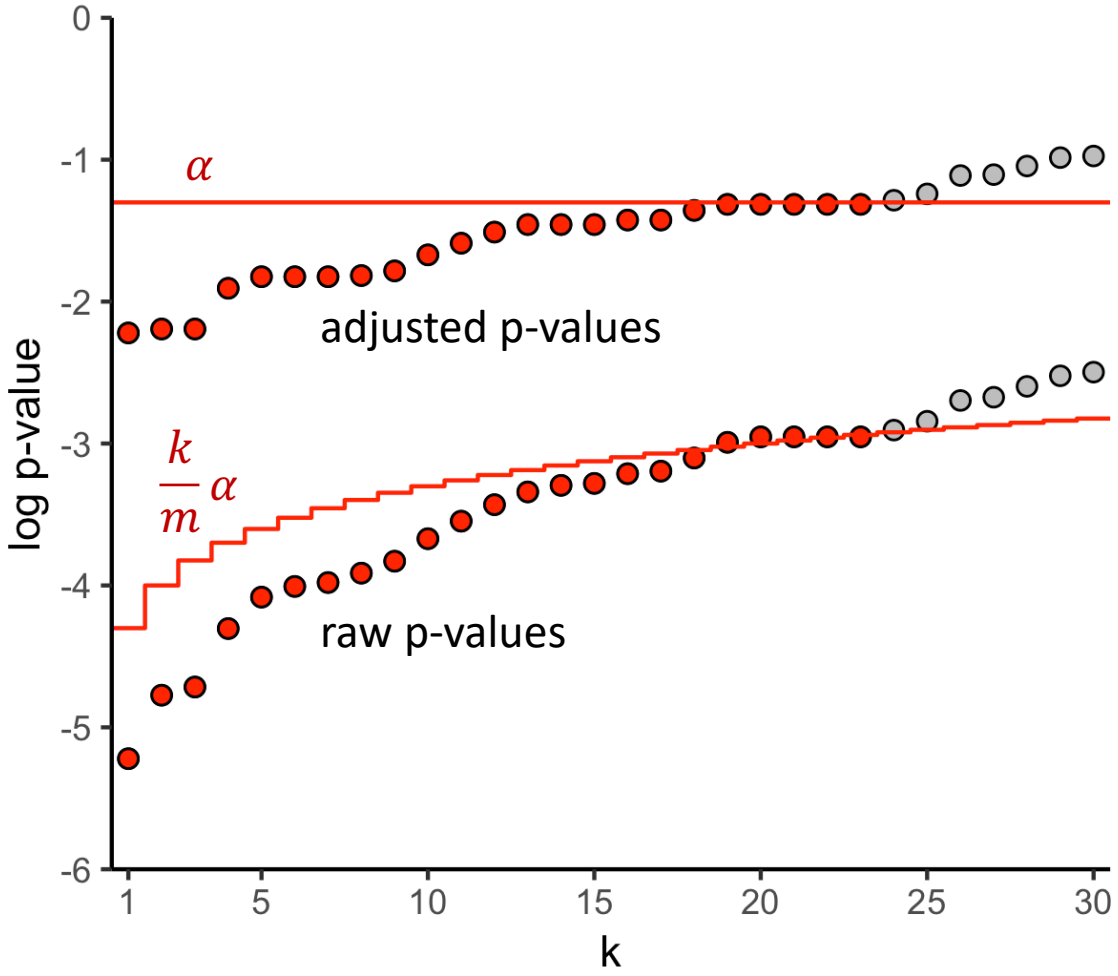
$$\overline{FDR} \leq \alpha$$

$\overline{FDR}$ 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

Useful but mind the interpretation

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[p.adj < 0.05]
[1] 1.476263e-05 2.662440e-05 3.029839e-05

# Benjamini-Hochberg procedure
> p.adj <- p.adjust(p, "BH")
> p[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

$$\overline{FDR} \leq \alpha$$

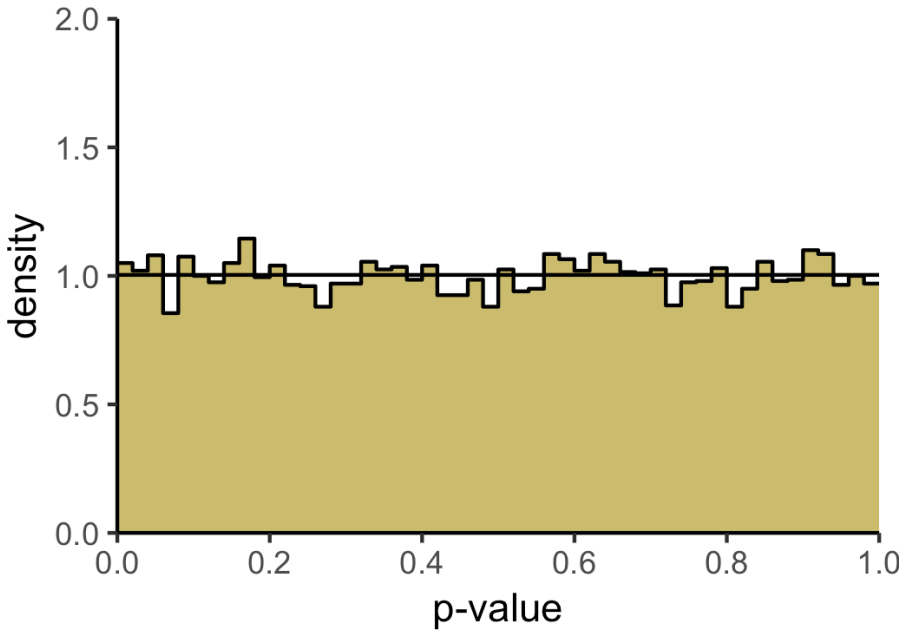
Estimating FDR

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

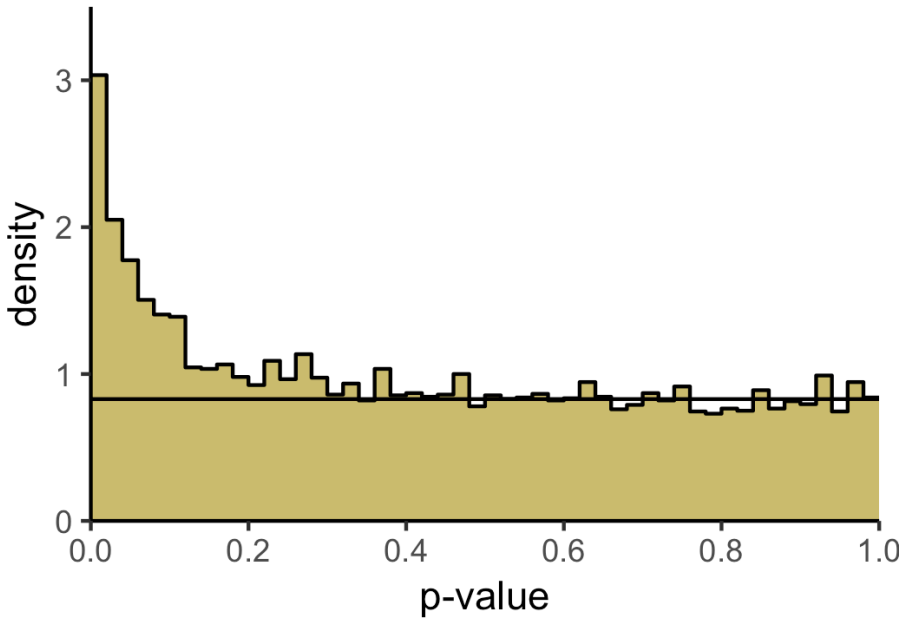
$$FDR(p_i)$$

P-value distribution

Data set 1
100% no effect

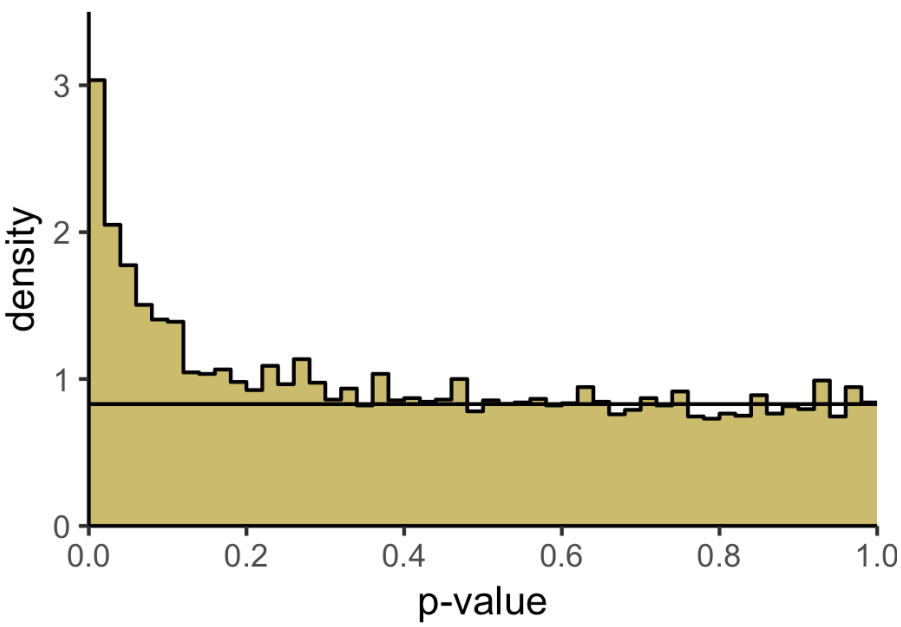


Data set 2
80% no effect,
20% real effect

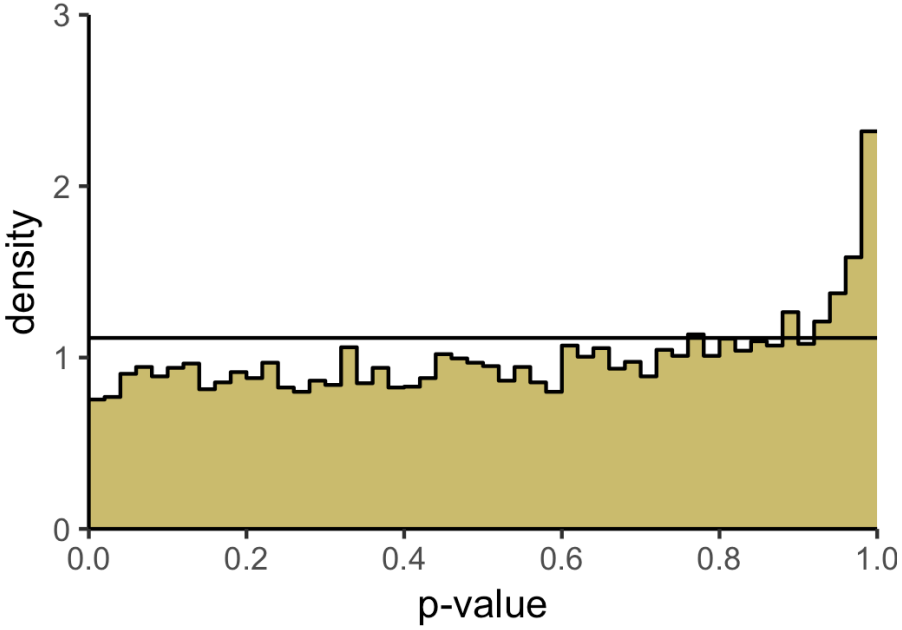


P-value distribution

Good

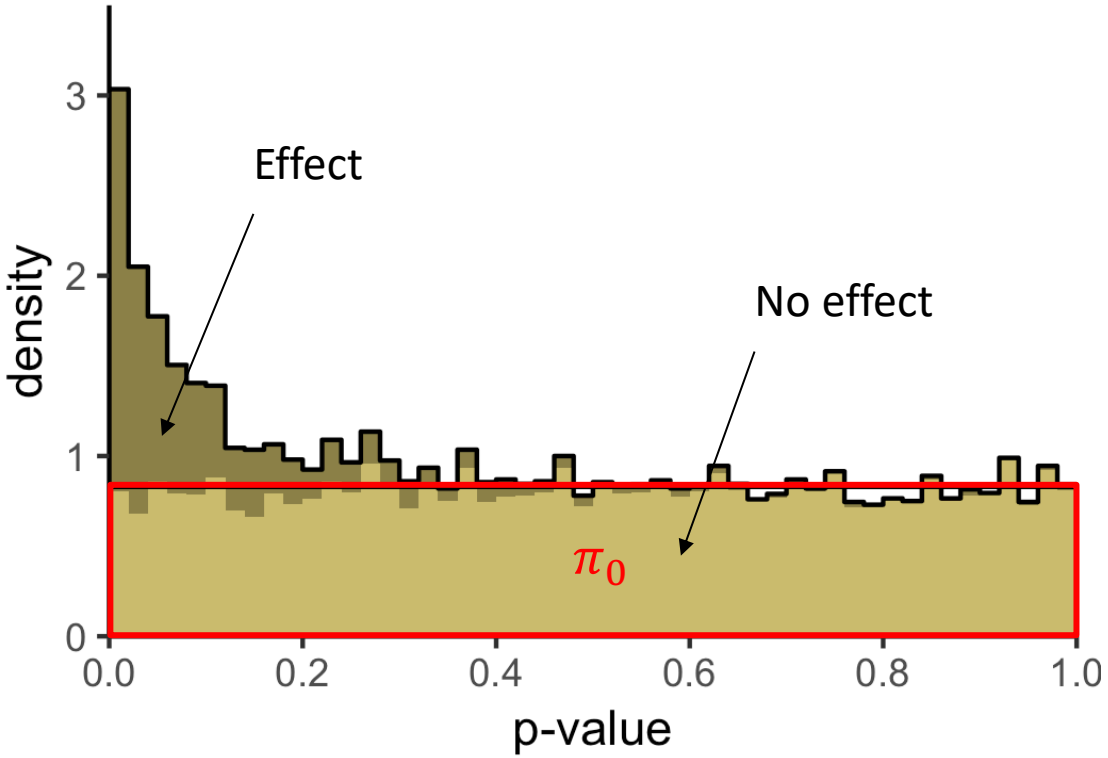


Bad!



Definition of π_0

80% no effect, 20% effect



Proportion of no effects

$$\pi_0 = \frac{\#\{\text{no effect}\}}{\#\{\text{all tests}\}}$$

Total shaded area is 1 (because of normalization)

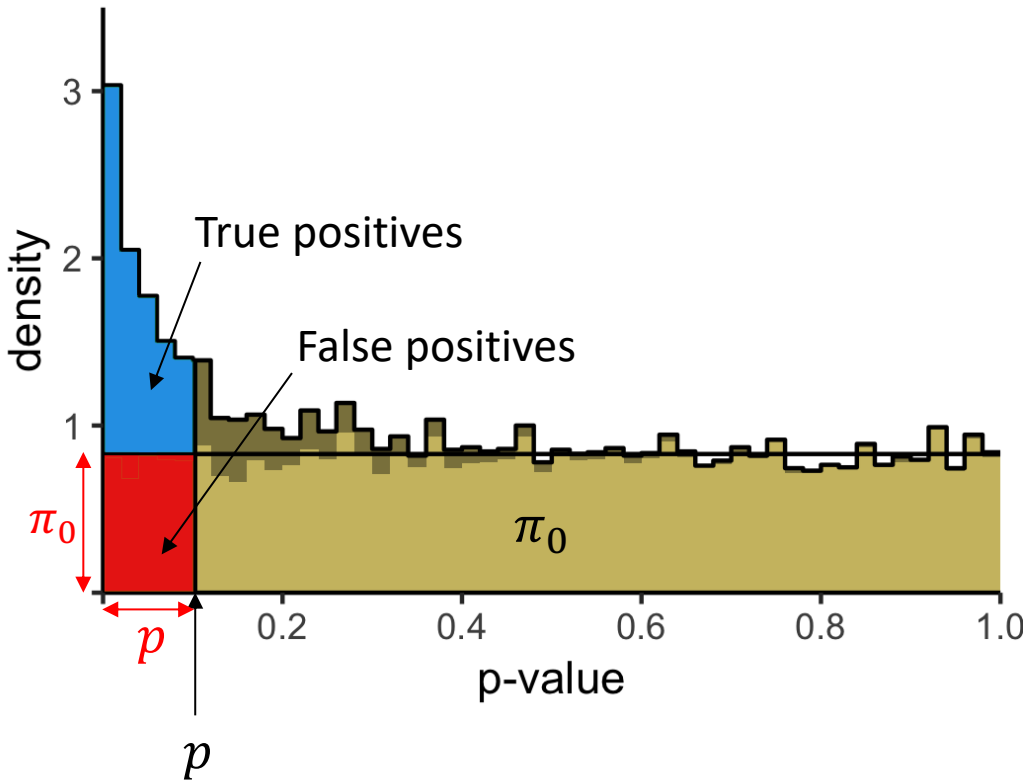
Area of the red rectangle is $\sim\pi_0$

Note

$\#\{\text{set}\}$ denotes number of elements in the set

Storey method: point estimate of FDR

80% no effect, 20% real effect



Point estimate, $FDR(p)$

First, estimate π_0

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

$$D(p) = \#\{p_i < p\}$$

No. of false positives is

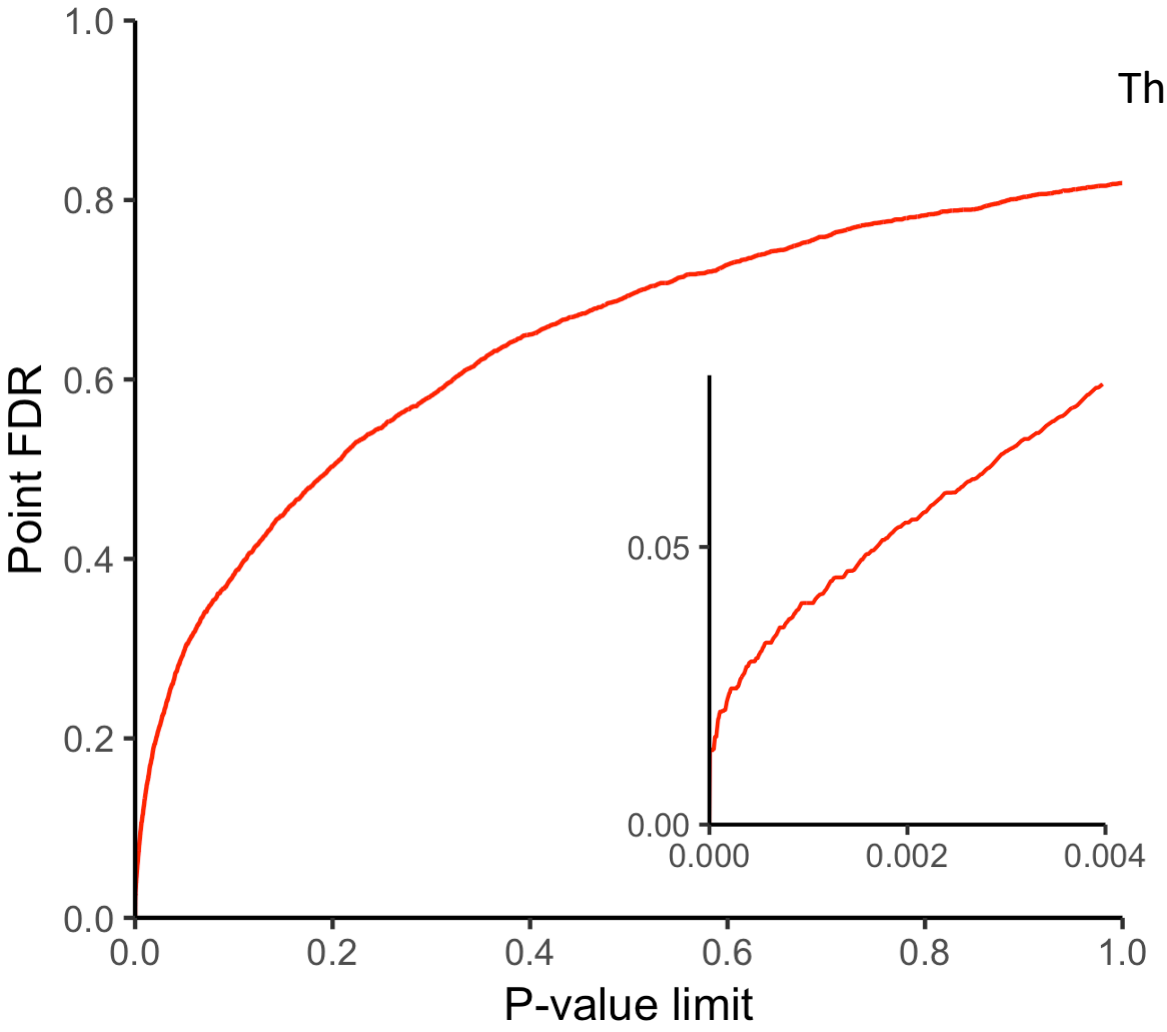
$$FP(p) = p\pi_0 m$$

Hence,

$$FDR(p) = \frac{FP(p)}{D(p)} = \frac{p\pi_0 m}{\#\{p_i < p\}}$$

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

Storey method



Point estimate of FDR

This is the so-called q-value:

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

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
...	...	...	...

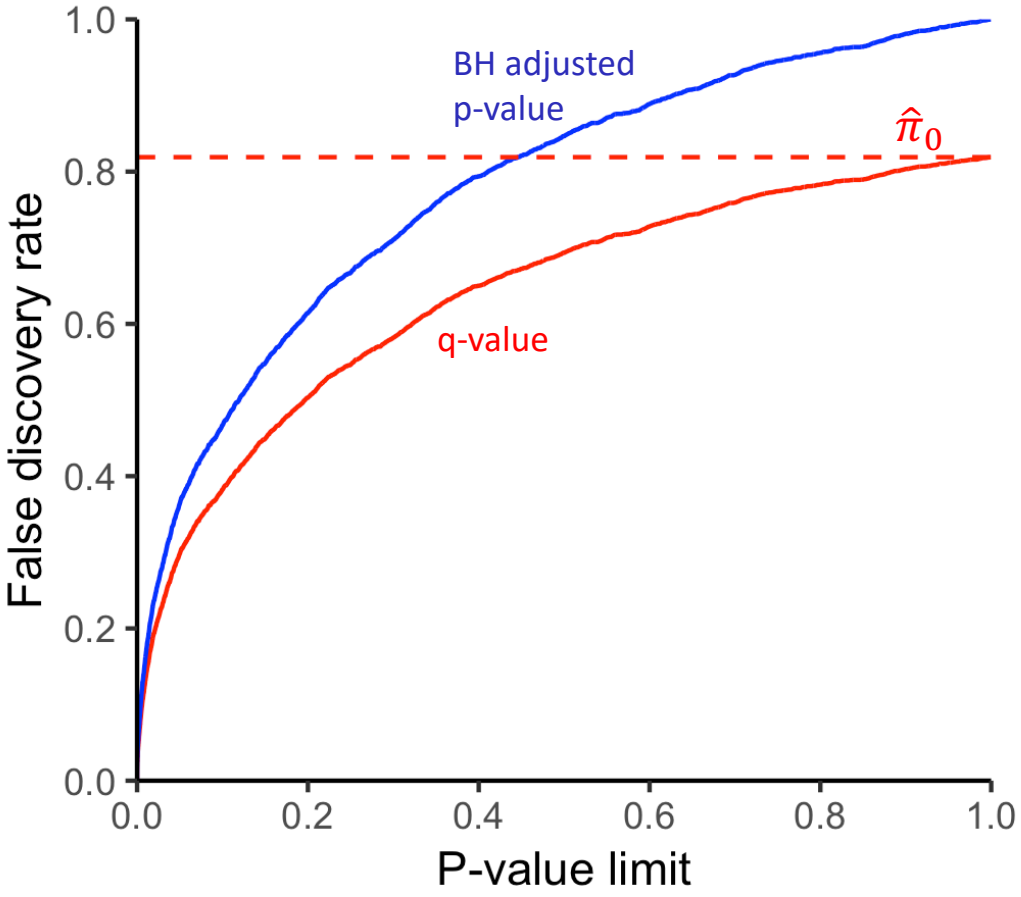
There are 106 tests with $q \leq 0.0273$

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.

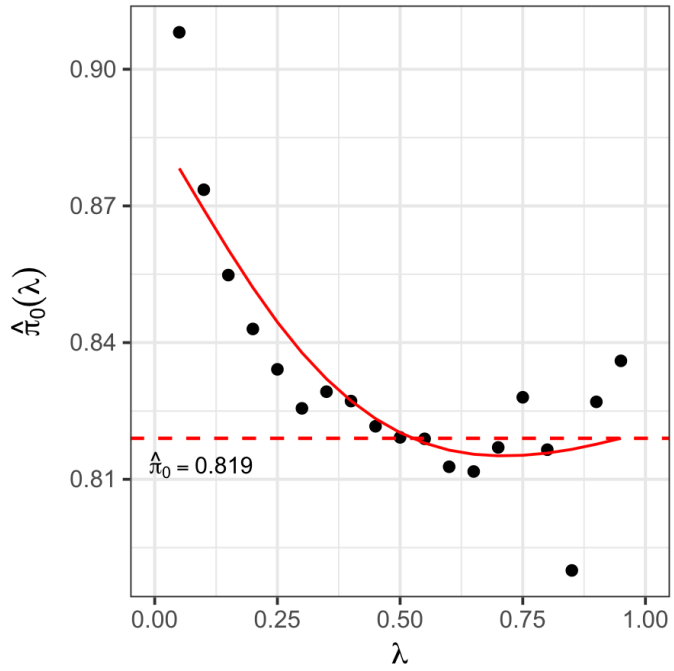
Hence, it is more powerful, especially for small $\hat{\pi}_0$.

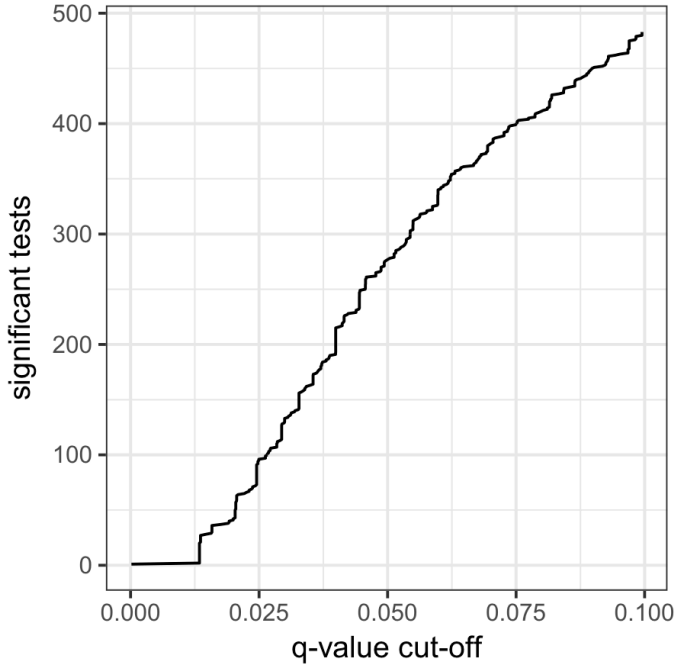
But this depends on how good the estimate of $\hat{\pi}_0$ is.

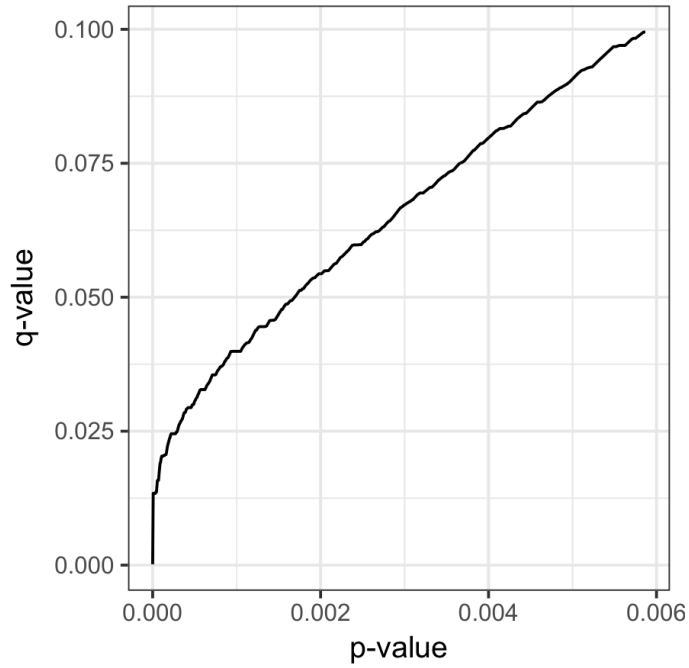
$\hat{\pi}_0$ - estimate of the proportion of null (no effect) tests

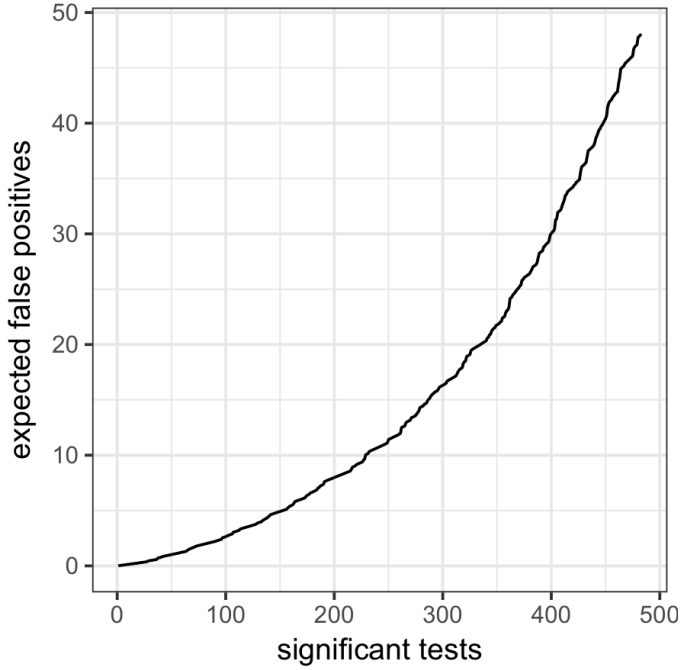
How to do this in R

```
> 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")
> length(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)
```

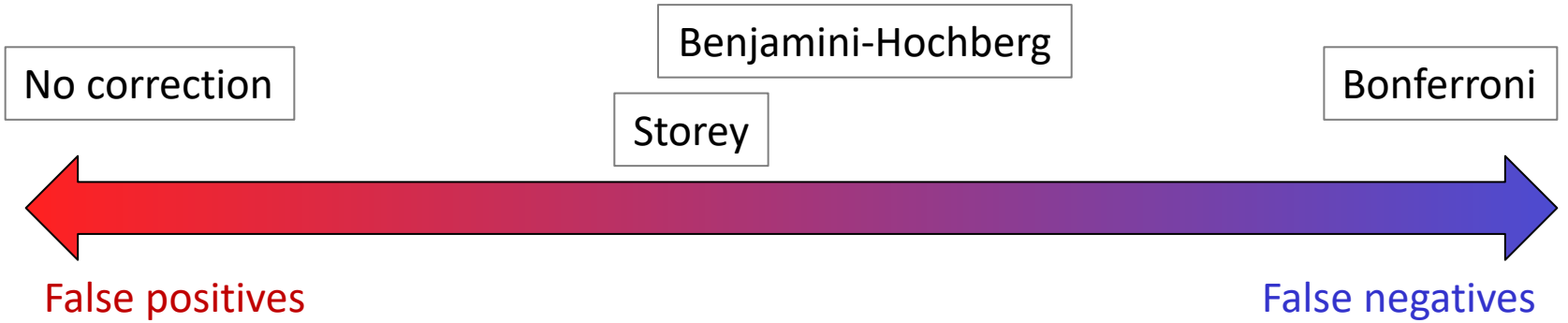

0.90
0.87
0.84
0.81
$\hat{\pi}_0(\lambda)$
$\hat{\pi}_0 = 0.819$
0.00
0.25
0.50
0.75
1.00
λ


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

Which multiple-test correction should I use?



False positive	**False negative**
"Discover" effect where there is no effect	Missed discovery
Can be tested in follow-up experiments	Once you've missed it, it's gone
Not hugely important in small samples	
Impossible to manage in large samples	

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

Acronyms:
FP – false positives; TP – true positives; FN – false negatives; FPR – false positive rate; FWER – family-wise error rate; FDR – false discovery rate; $\hat{\pi}_0$ - estimate of the fraction of non-significant tests

Hand-outs available at
https://dag.compbio.dundee.ac.uk/training/Statistics_lectures.html