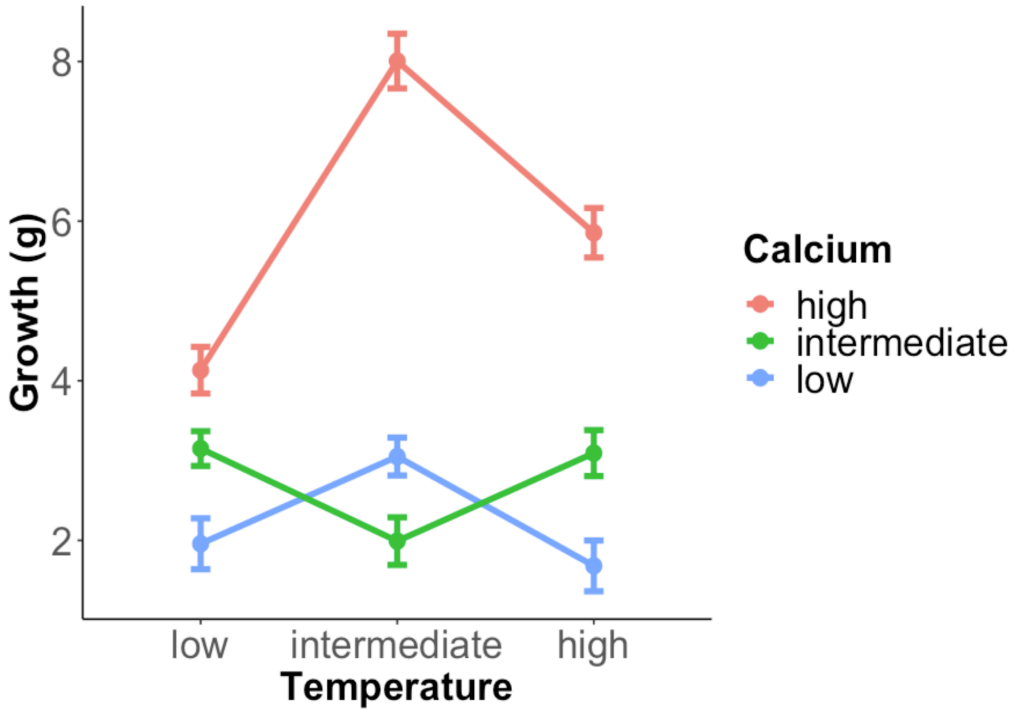
```
anova(lm(Growth~Calcium*Temperature))
Analysis of Variance Table

Response: Growth
                   Df  Sum Sq Mean Sq F value    Pr(>F)
Calcium             2 125.190  62.595 556.500 < 2.2e-16 ***
Temperature         2  12.371   6.186  54.992 1.137e-11 ***
Calcium:Temperature 4  34.801   8.700  77.349 < 2.2e-16 ***
Residuals          36   4.049   0.112
```



For the interaction, there are 3 levels of Calcium and 3 levels of Temperature, yielding 9 group means.

Comparing growth across all groups would require 36 pairwise contrasts (9 \times 8)/2 = 36.

Why do we conduct ANOVA rather than perform multiple pairwise tests of means?

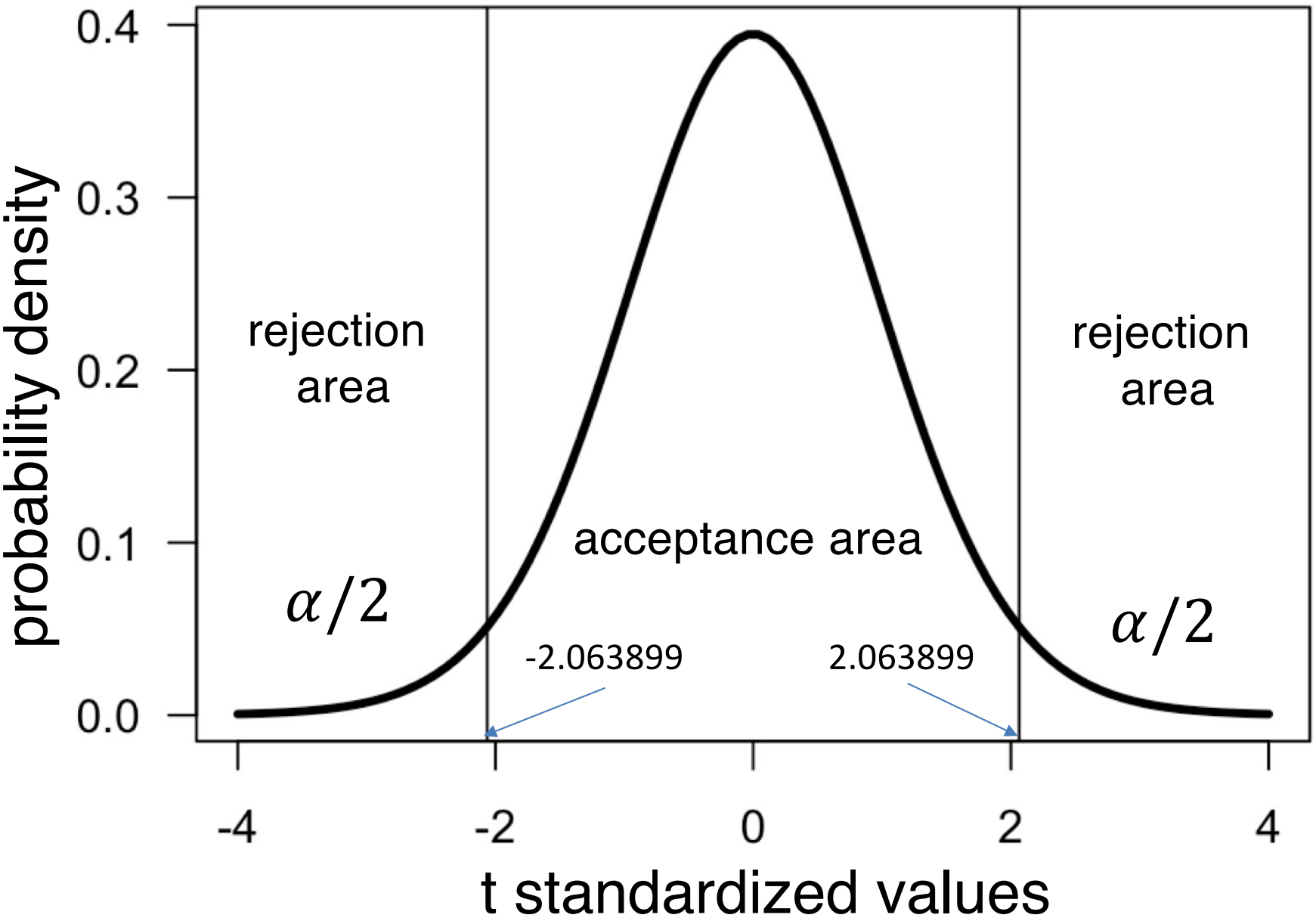
BIOL 422 & 680, Pedro Peres-Neto, Biology, Concordia University

A pedagogical guide for understanding the issues underlying

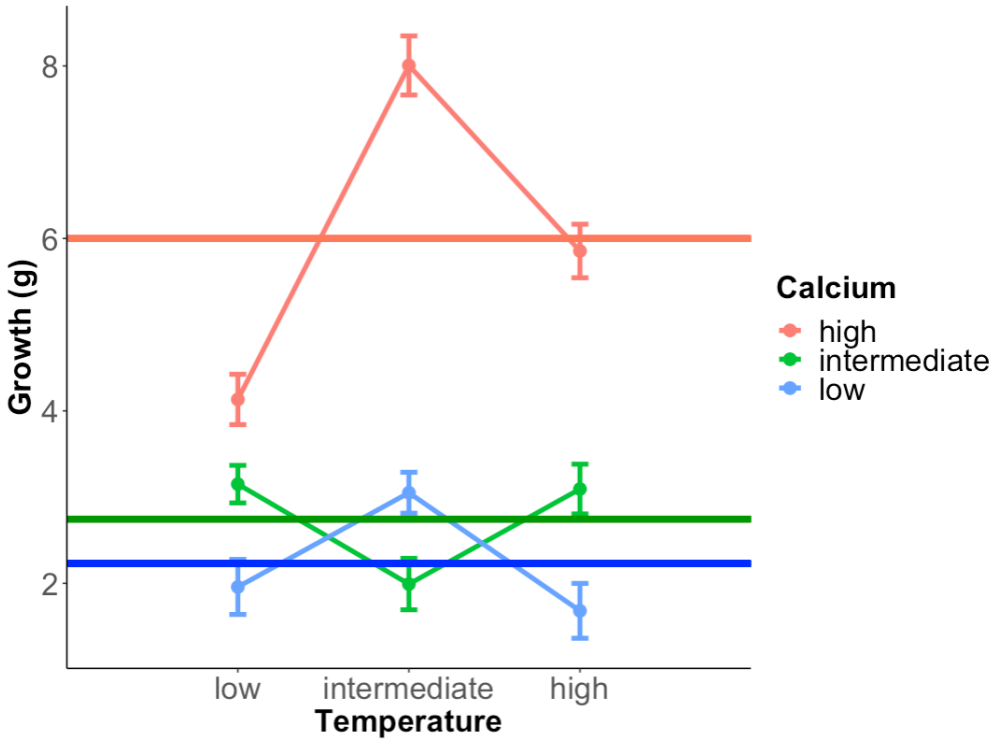
Multiple hypothesis testing

Why should we not trust results obtained from multiple statistical tests?

t-distribution assuming H_0 as true



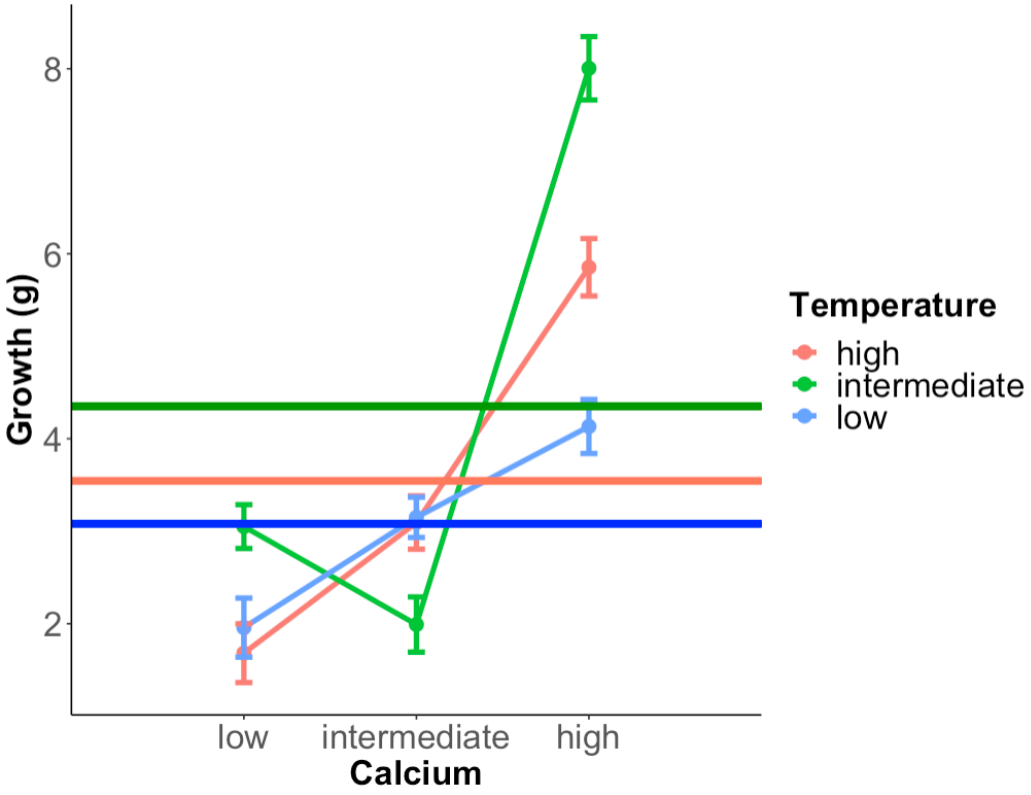
```
anova(lm(Growth~Calcium*Temperature))
Analysis of Variance Table

Response: Growth
                   Df  Sum Sq Mean Sq F value    Pr(>F)
Calcium             2 125.190  62.595 556.500 < 2.2e-16 ***
Temperature         2  12.371   6.186  54.992 1.137e-11 ***
Calcium:Temperature 4  34.801   8.700  77.349 < 2.2e-16 ***
Residuals          36   4.049   0.112
```



Comparing growth across all groups would require 36 pairwise contrasts (9 \times 8)/2 = 36.

High – intermediate
High – low
Intermediate – low

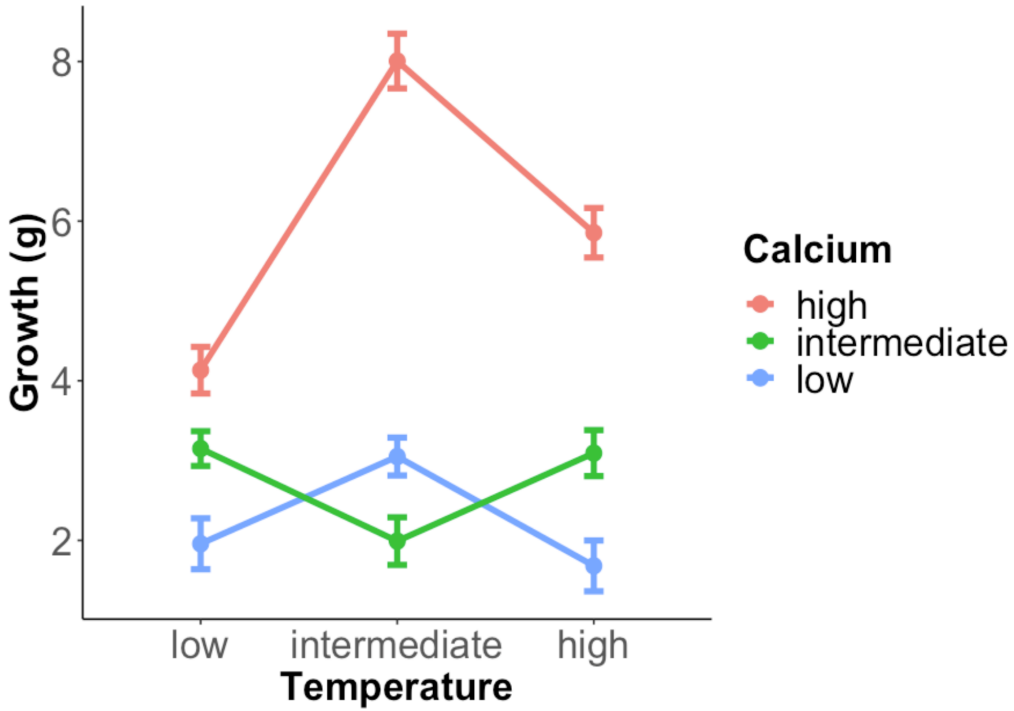
```
anova(lm(Growth~Calcium*Temperature))
Analysis of Variance Table

Response: Growth
                   Df  Sum Sq Mean Sq F value    Pr(>F)
Calcium             2 125.190  62.595 556.500 < 2.2e-16 ***
Temperature         2  12.371   6.186  54.992 1.137e-11 ***
Calcium:Temperature 4  34.801   8.700  77.349 < 2.2e-16 ***
Residuals          36   4.049   0.112
```



Comparing growth across all groups would require 36 pairwise contrasts (9 \times 8)/2 = 36.

High - intermediate
High - low
Intermediate - low

```
anova(lm(Growth~Calcium*Temperature))
Analysis of Variance Table

Response: Growth
                   Df  Sum Sq Mean Sq F value    Pr(>F)
Calcium             2 125.190  62.595 556.500 < 2.2e-16 ***
Temperature         2  12.371   6.186  54.992 1.137e-11 ***
Calcium:Temperature 4  34.801   8.700  77.349 < 2.2e-16 ***
Residuals          36   4.049   0.112
```

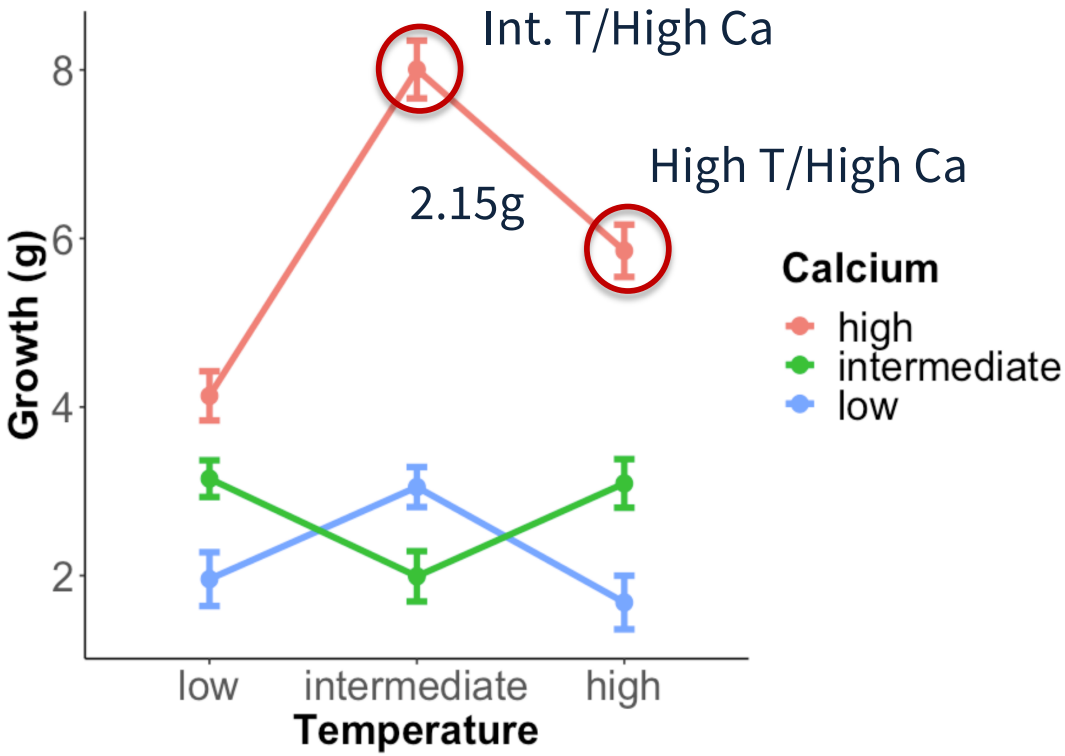


For the interaction, there are 3 levels of Calcium and 3 levels of Temperature, yielding 9 group means.

Comparing growth across all groups would require 36 pairwise contrasts (9 \times 8)/2 = 36.

```
$`Temperature:Calcium`
                                                  diff
intermediate:high-high:high                   2.15154803
low:high-high:high                           -1.72154916
high:intermediate-high:high                  -2.76050275
intermediate:intermediate-high:high          -3.86300578
low:intermediate-high:high                   -2.70381093
high:low-high:high                           -4.17303298
intermediate:low-high:high                   -2.80337496
low:low-high:high                            -3.89620697
low:high-intermediate:high                   -3.87309719
high:intermediate-intermediate:high          -4.91205078
intermediate:intermediate-intermediate:high -6.01455381
low:intermediate-intermediate:high           -4.85535896
high:low-intermediate:high                   -6.32458101
intermediate:low-intermediate:high           -4.95492299
low:low-intermediate:high                    -6.04775500
high:intermediate-low:high                   -1.03895359
intermediate:intermediate-low:high           -2.14145662
low:intermediate-low:high                    -0.98226177
high:low-low:high                            -2.45148382
intermediate:low-low:high                    -1.08182580
low:low-low:high                             -2.17465781
intermediate:intermediate-high:intermediate -1.10250303
low:intermediate-high:intermediate            0.05669182
high:low-high:intermediate                   -1.41253023
intermediate:low-high:intermediate           -0.04287221
low:low-high:intermediate                    -1.13570422
low:intermediate-intermediate:intermediate    1.15919485
high:low-intermediate:intermediate           -0.31002720
intermediate:low-intermediate:intermediate    1.05963082
low:low-intermediate:intermediate            -0.03320119
high:low-low:intermediate                    -1.46922205
intermediate:low-low:intermediate            -0.09956403
low:low-low:intermediate                     -1.19239604
intermediate:low-high:low                     1.36965802
low:low-high:low                              0.27682601
low:low-intermediate:low                     -1.09283201
```

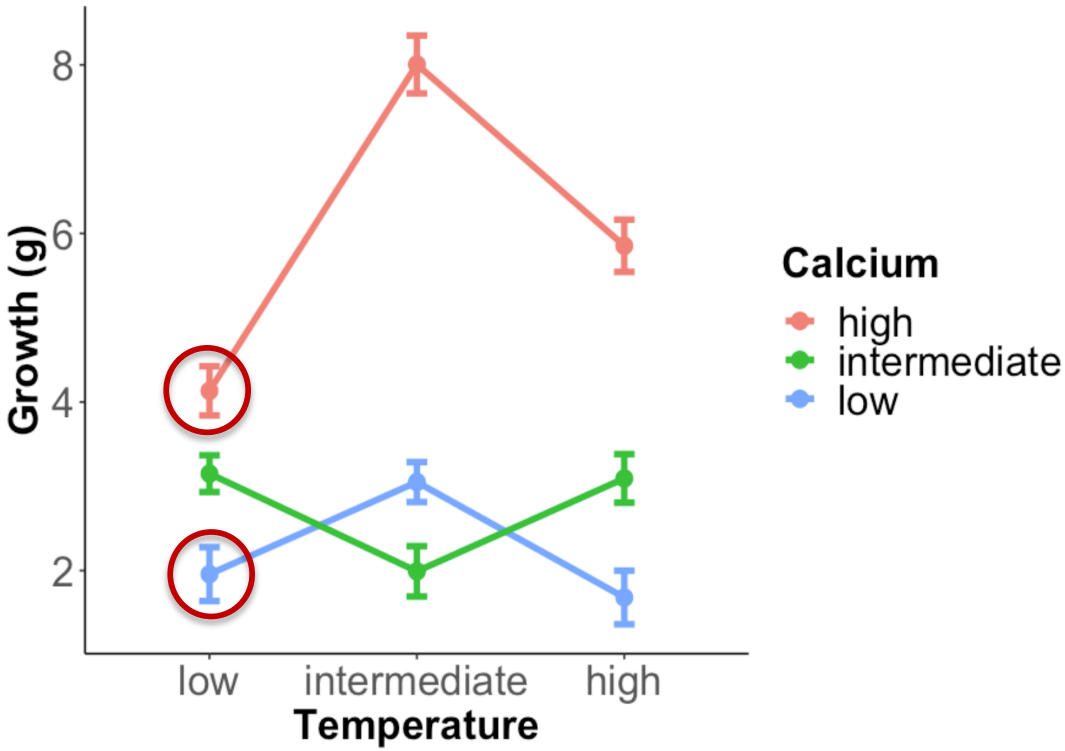
There are 36 possible pairwise tests to contrast Growth across 36 levels (9 x 8/2 = 36).



One possible contrast: Does the mean growth at intermediate temperature and high calcium differ significantly from the mean growth at high temperature and high calcium? (Observed difference = 2.15 g)

```
$`Temperature:Calcium`
                                                    diff
intermediate:high-high:high                     2.15154803
low:high-high:high                             -1.72154916
high:intermediate-high:high                    -2.76050275
intermediate:intermediate-high:high            -3.86300578
low:intermediate-high:high                     -2.70381093
high:low-high:high                             -4.17303298
intermediate:low-high:high                     -2.80337496
low:low-high:high                              -3.89620697
low:high-intermediate:high                     -3.87309719
high:intermediate-intermediate:high            -4.91205078
intermediate:intermediate-intermediate:high    -6.01455381
low:intermediate-intermediate:high             -4.85535896
high:low-intermediate:high                     -6.32458101
intermediate:low-intermediate:high             -4.95492299
low:low-intermediate:high                      -6.04775500
high:intermediate-low:high                     -1.03895359
intermediate:intermediate-low:high             -2.14145662
low:intermediate-low:high                      -0.98226177
high:low-low:high                              -2.45148382
intermediate:low-low:high                      -1.08182580
low:low-low:high                               -2.17465781
intermediate:intermediate-high:intermediate    -1.10250303
low:intermediate-high:intermediate              0.05669182
high:low-high:intermediate                     -1.41253023
intermediate:low-high:intermediate             -0.04287221
low:low-high:intermediate                      -1.13570422
low:intermediate-intermediate:intermediate      1.15919485
high:low-intermediate:intermediate             -0.31002720
intermediate:low-intermediate:intermediate      1.05963082
low:low-intermediate:intermediate              -0.03320119
high:low-low:intermediate                      -1.46922205
intermediate:low-low:intermediate              -0.09956403
low:low-low:intermediate                       -1.19239604
intermediate:low-high:low                       1.36965802
low:low-high:low                                0.27682601
low:low-intermediate:low                       -1.09283201
```

There are 36 possible pairwise tests to contrast Growth across levels (9 x 8/2 = 36).



Another possible contrast: Does the mean growth at low temperature and low calcium differ significantly from the mean growth at low temperature and high calcium? (Observed difference = 2.17 g).

What happens when we conduct too many statistical tests?

A past classroom demonstration using a survey

Past classroom surveys:

Would you expect individuals born on odd versus even days to differ in their preferences?

	dislike				Love it
	1	2	3	4	5
			X		
1) Do you like soccer?	X				
2) Do you like playing video games?			X		
3) Do you like eating out?					
4) Do you enjoy writting?					
5) Do you like cats?			X		
6) Do you like to watch movies?					X
7) Do you like to read novels?					

.....

21) Do you like science fiction?	X				
22) Do you like pizza?		X			
23) Do you like to listen to the radio?				X	
24) Do you like museums?			X		

Multiple testing survey (BIOL422, BIOL680)/anonymous survey will close on Wednesday Feb. 3 (5pm)

Results will be used to demonstrate the statistical principles of multiple testing

last number of your street address *

- Odd number
- Even number

Your birthday is an odd or even number (the actual day; not month or year) *

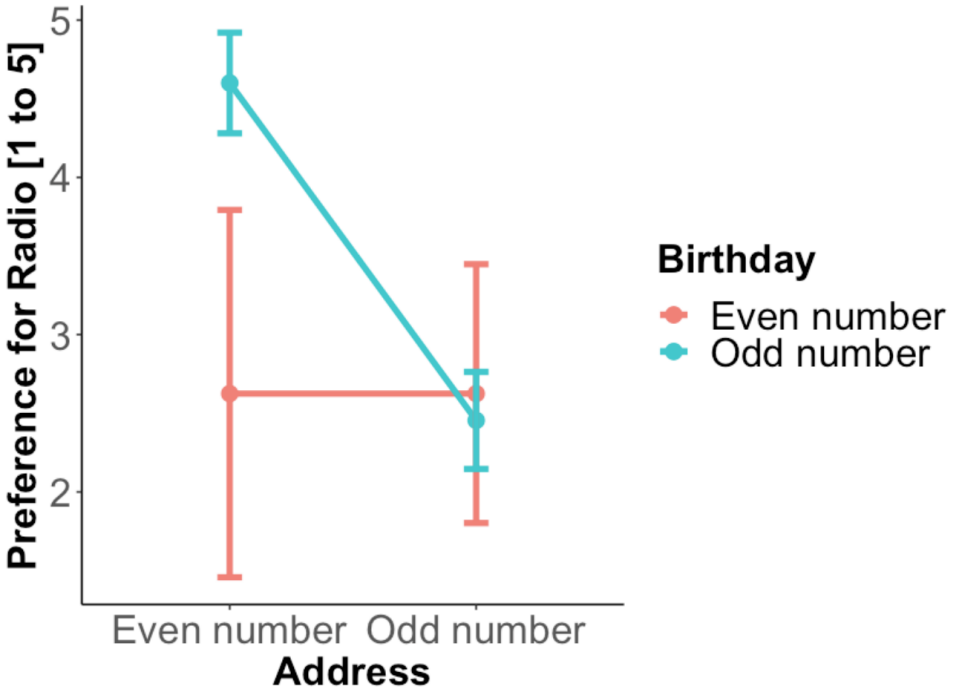
- Odd number
- Even number

Do you like soccer? *

	1	2	3	4	5	
Deslike	○	○	○	○	○	Love it

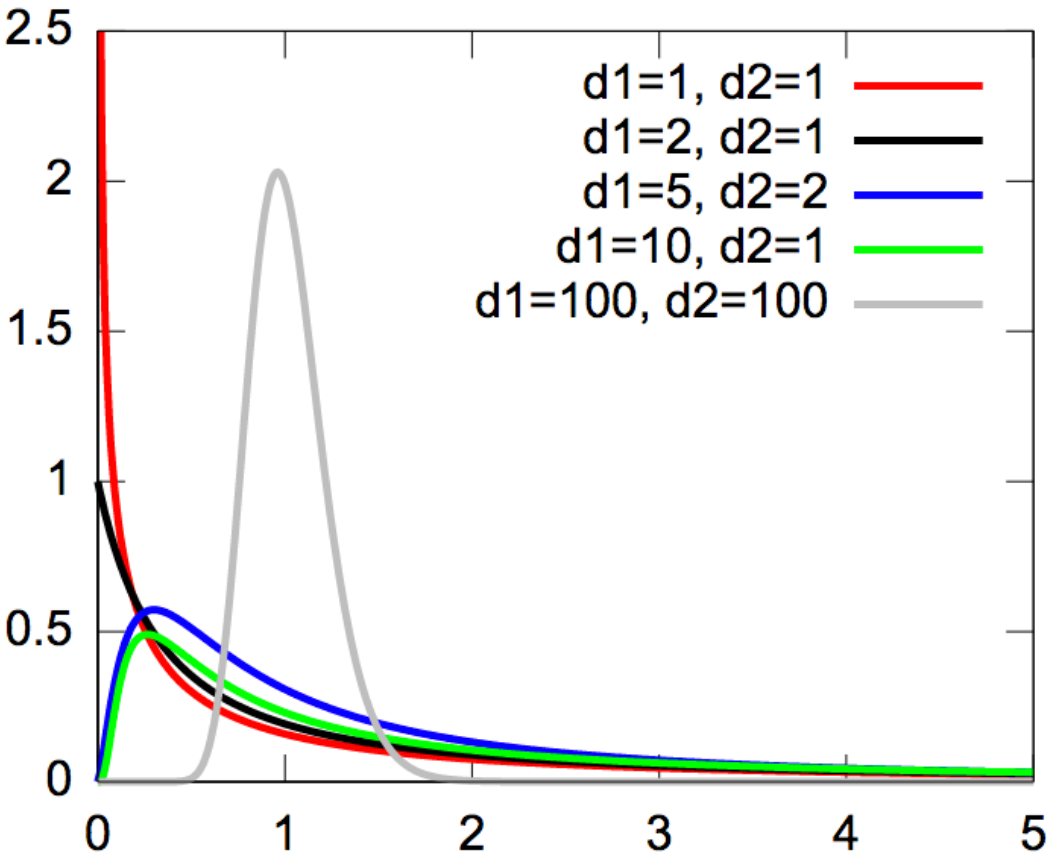
Really ????

```
Response: Do.you.like.to.listen.to.the.Radio.
                Df Sum Sq Mean Sq F value   Pr(>F)
Birthday         1 13.220 13.2196 12.5081 0.001226 **
Address          1  7.031  7.0309  6.6525 0.014546 *
Birthday:Address  1 10.440 10.4397  9.8778 0.003524 **
Residuals       33 34.877  1.0569
```



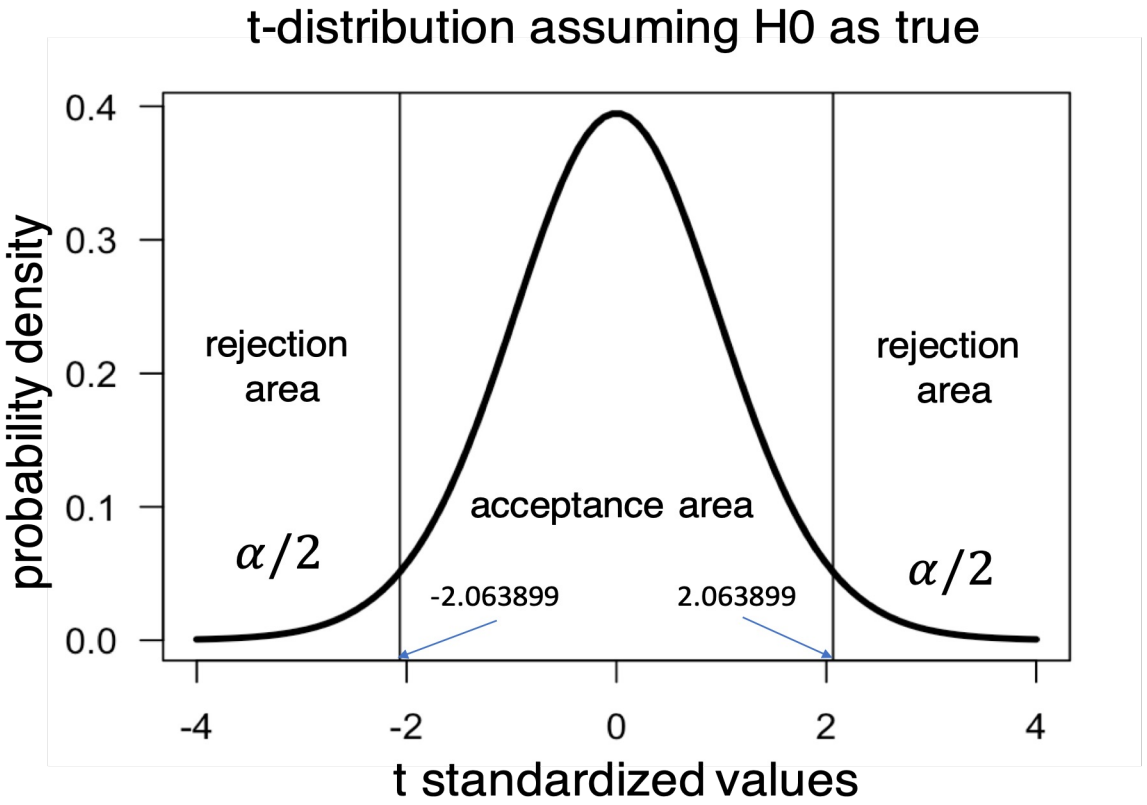
class survey:
24 questions = 24
ANOVAS x 3 p-values each = 72 p-values
24 ANOVAs x 6 p-values each =
144 pairwise t-tests

Why do we conduct ANOVA rather than perform multiple pairwise tests of means?



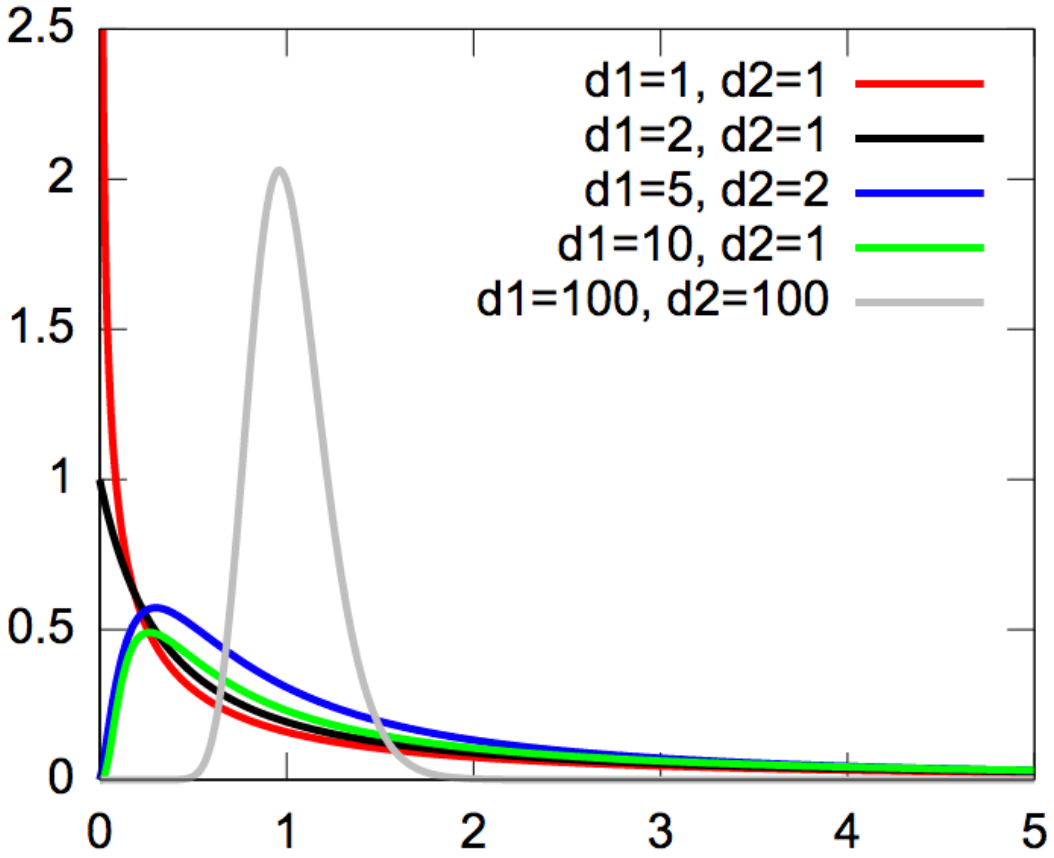
The probability of committing a Type I error for any single test is α. However, when many tests are conducted, the chance of obtaining false positives increases. For example, if 100 independent tests are performed with α = 0.05, we expect about 5 tests to be significant purely by chance.

Why do we conduct ANOVA rather than perform multiple pairwise tests of means?



The probability of committing a Type I error for any single test is α. However, when many tests are conducted, the chance of obtaining false positives increases. For example, if 100 independent tests are performed with α = 0.05, we expect about 5 tests to be significant purely by chance.

Why do we conduct ANOVA rather than perform multiple pairwise tests of means?



class survey:
24 questions = 24
ANOVAS x 3 p-values each = 72 p-values

24 ANOVAs x 6 p-values each = 144 pairwise t-tests (p-values)

Even though multiple ANOVAs will inflate the number of false positives (i.e., type I error), it still generates a much smaller number of tests than pairwise tests (i.e., 72 versus 144 tests).

```
$Temperature
                       diff
intermediate-high  0.8062343
low-high          -0.4626771
low-intermediate  -1.2689115

$Calcium
                       diff
intermediate-high -3.2524394
low-high          -3.7675379
low-intermediate  -0.5150985
```

3 pairwise tests

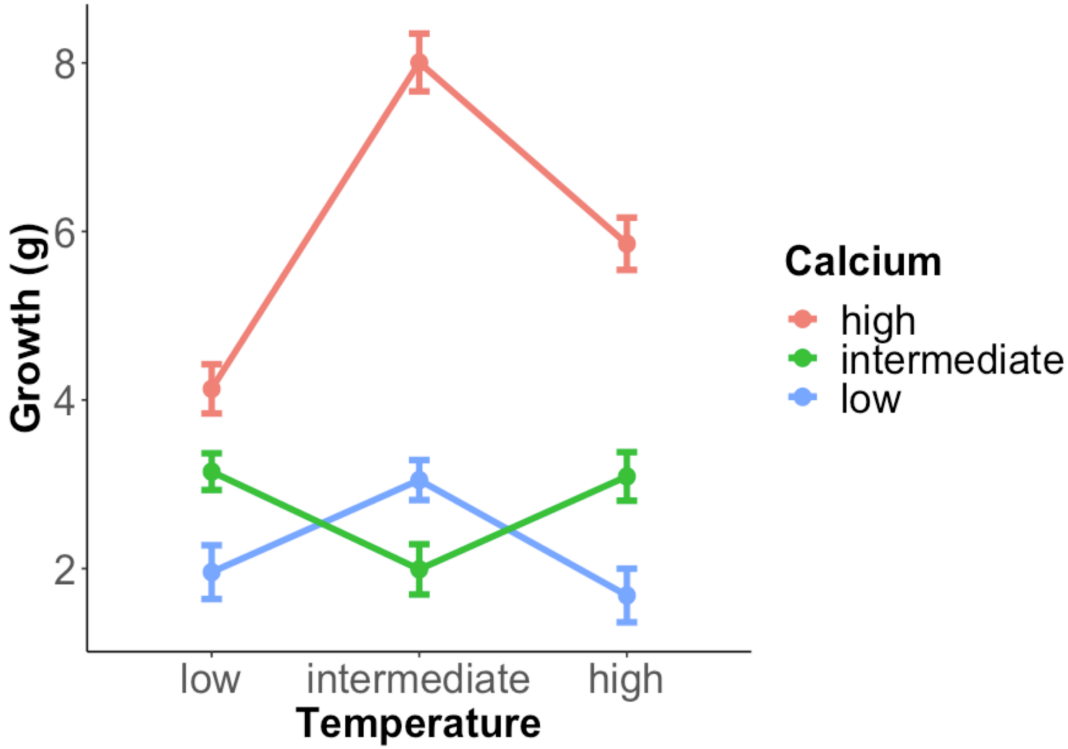
3 pairwise tests

```
$`Temperature:Calcium`
                                                    diff
intermediate:high-high:high                      2.15154803
low:high-high:high                              -1.72154916
high:intermediate-high:high                     -2.76050275
intermediate:intermediate-high:high             -3.86300578
low:intermediate-high:high                      -2.70381093
high:low-high:high                              -4.17303298
intermediate:low-high:high                      -2.80337496
low:low-high:high                               -3.89620697
low:high-intermediate:high                      -3.87309719
high:intermediate-intermediate:high             -4.91205078
intermediate:intermediate-intermediate:high     -6.01455381
low:intermediate-intermediate:high              -4.85535896
high:low-intermediate:high                      -6.32458101
intermediate:low-intermediate:high              -4.95492299
low:low-intermediate:high                       -6.04775500
high:intermediate-low:high                      -1.03895359
intermediate:intermediate-low:high              -2.14145662
low:intermediate-low:high                       -0.98226177
high:low-low:high                               -2.45148382
intermediate:low-low:high                       -1.08182580
low:low-low:high                                -2.17465781
intermediate:intermediate-high:intermediate     -1.10250303
low:intermediate-high:intermediate               0.05669182
high:low-high:intermediate                      -1.41253023
intermediate:low-high:intermediate              -0.04287221
low:low-high:intermediate                       -1.13570422
low:intermediate-intermediate:intermediate       1.15919485
high:low-intermediate:intermediate              -0.31002720
intermediate:low-intermediate:intermediate       1.05963082
low:low-intermediate:intermediate               -0.03320119
high:low-low:intermediate                       -1.46922205
intermediate:low-low:intermediate               -0.09956403
low:low-low:intermediate                        -1.19239604
intermediate:low-high:low                        1.36965802
low:low-high:low                                 0.27682601
low:low-intermediate:low                        -1.09283201
```

36 pairwise tests

6 main effect pairwise t-tests
36 interaction pairwise t-tests
Total = 42 tests

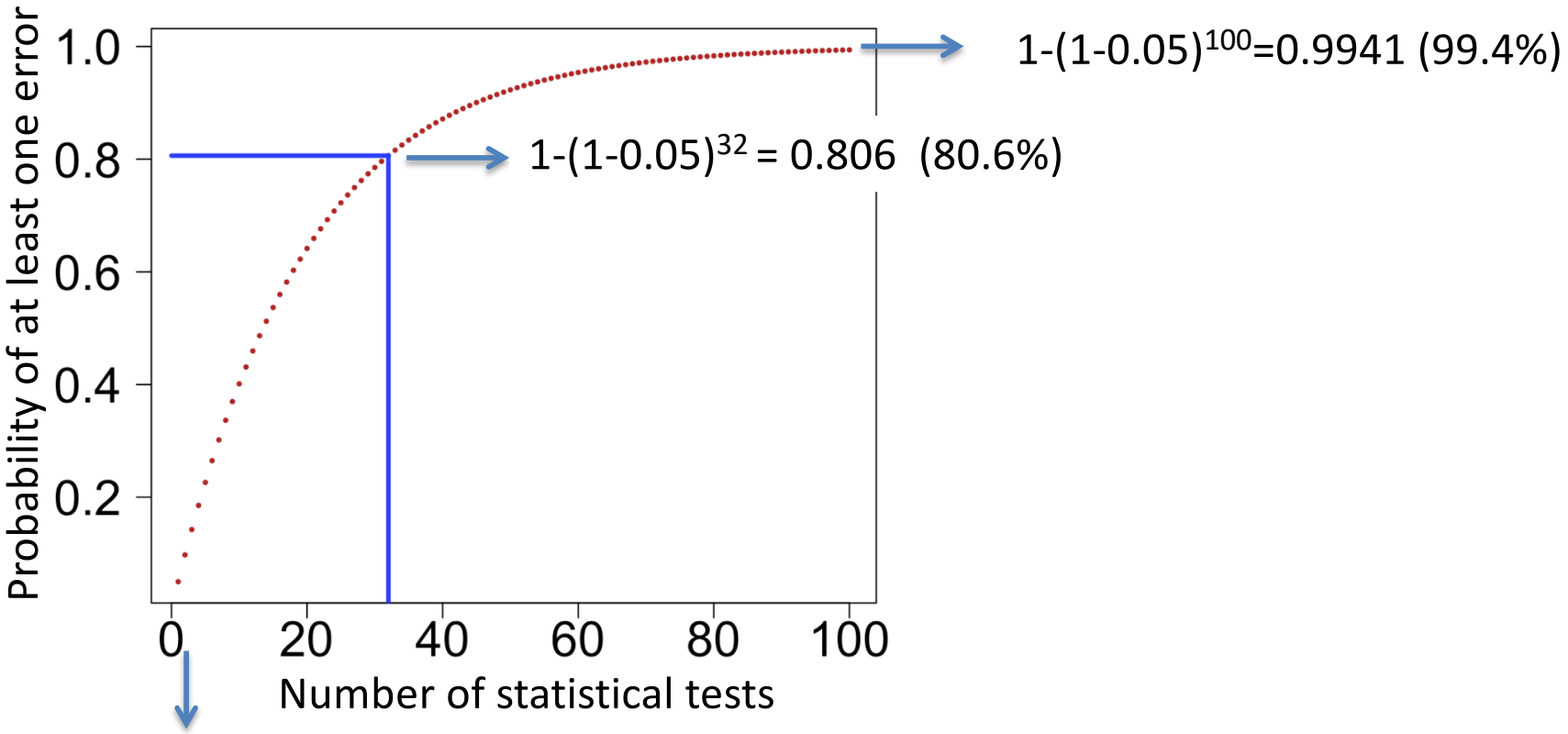
```
anova(lm(Growth~Calcium*Temperature))
Analysis of Variance Table

Response: Growth
                  Df  Sum Sq Mean Sq F value    Pr(>F)
Calcium            2 125.190  62.595 556.500 < 2.2e-16 ***
Temperature        2  12.371   6.186  54.992 1.137e-11 ***
Calcium:Temperature  4  34.801   8.700  77.349 < 2.2e-16 ***
Residuals         36   4.049   0.112
```



If we set alpha = 0.05 (i.e., a 95% acceptance region), what is the probability of obtaining at least one significant result by chance—that is, a false positive - across 32 tests when all null hypotheses are true?

$1-(1-alpha)^{32} = 1-(1-0.05)^{32} = 0.806$ (80.6%)

80.6% chance of finding at least 1 significant test when H_0 is true!



$1-(1-0.05)^{1}=0.050$ (5%)

[1 test leads to the expected alpha (prob. of committing a type I error)

Examples of really huge numbers of multiple tests

How does multiple testing correction work?

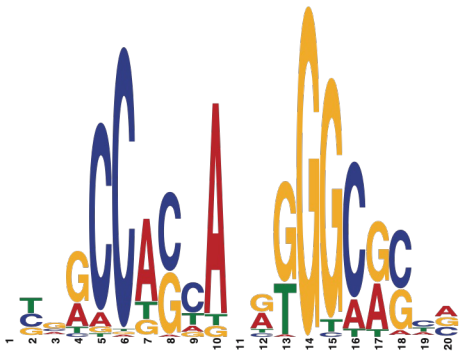
William S Noble

NATURE BIOTECHNOLOGY VOLUME 27 NUMBER 12 DECEMBER 2009

When prioritizing hits from a high-throughput experiment, it is important to correct for random events that falsely appear significant. How is this done and what methods should be used?

As a motivating example, suppose that you are studying CTCF, a highly conserved zinc-finger DNA-binding protein that exhibits diverse regulatory functions and that may play a major role in the global organization of the chromatin architecture of the human genome[1]. To better understand this protein, you want to identify candidate CTCF binding sites in human chromosome 21. Using a previously published model of the CTCF binding motif (**Fig. 1a**)[2], each 20 nucleotide (nt) sub-sequence of chromosome 21 can be scored for its similarity to the CTCF motif. Considering both DNA strands, there are 68 million such subsequences. **Figure 1b** lists the top 20 scores from such a search.

a



68 million statistical tests

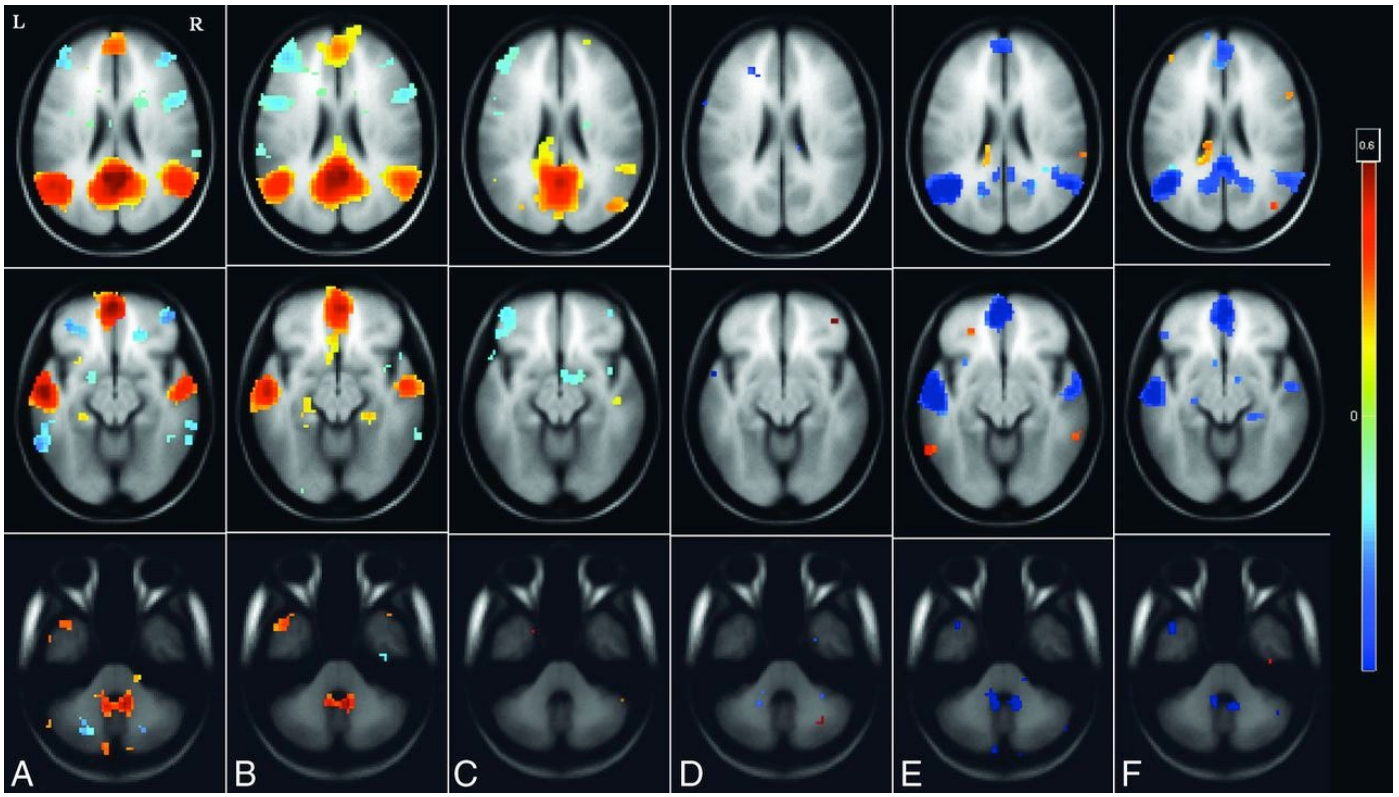
b

Position	Str	Sequence	Score
19390631	+	TTGACCAGCAGGGGGCGCCG	26.30
32420105	+	CTGGCCAGCAGAGGGCAGCA	26.30
27910537	–	CGGTGCCCCCTGCTGGTCAG	26.18
21968106	+	GTGACCACCAGGGGGCAGCA	25.81
31409358	+	CGGGCCTCCAGGGGGCGCTC	25.56
19129218	–	TGGCGCCACCTGCTGGTCAC	25.44
21854623	+	CTGGCCAGCAGAGGGCAGGG	24.95
12364895	+	CCCGCCAGCAGAGGGAGCCG	24.71
13406383	+	CTAGCCACCAGGTGGCGGTG	24.71
18613020	+	CCCGCCAGCAGAGGGAGCCG	24.71
31980801	+	ACGCCCAGCAGGGGGCGCCG	24.71
32909754	–	TGGCTCCCCCTGGCGGCCGG	24.71
25683654	+	TCGGCCACTAGGGGGCACTA	24.58
31116990	–	GGCCGCCACCTTGTGGCCAG	24.58
29615421	–	CTCTGCCCTCTGGTGGCTGC	24.46
6024389	+	GTTGCCACCAGAGGGCACTA	24.46
26610753	–	CACTGCCCTCTGCTGGCCCA	24.34
26912791	–	GGGCGCCACCTGGCGGTCAC	24.34
20446267	+	CTGCCCACCAGGGGGCAGCG	24.22
21872506	–	TGGCGCCACCTGGCGGCAGC	24.22

Wikipedia: High-throughput screening (HTS) is a method for scientific experimentation especially used in drug discovery and relevant to the fields of biology and chemistry. Using robotics, data processing and control software, liquid handling devices, and sensitive detectors, High-throughput screening allows a researcher to quickly conduct millions of chemical, genetic, or pharmacological tests.

Examples of really huge numbers of multiple tests

Compare signal changes between task and no-task conditions using a t-test across more than 250,000 voxels (3D brain pixels).



AJNR AMERICAN JOURNAL OF NEURORADIOLOGY

Seizure Frequency Can Alter Brain Connectivity: Evidence from Resting-State fMRI

R.D. Bharath, S. Sinha, R. Panda, K. Raghavendra, L. George, G. Chaitanya, A. Gupta, and P. Satishchandra

How to avoid inflated false positives (type I errors) due to multiple testing? Or the so-called family-wise error rate (FWER)

There is a large number of specific (e.g., Tukey-test for comparing two the difference between two means) and general procedures; the latter applying to any statistical test as they are used to control for multiple tests by correcting P-values.

There are many commonly used procedures to correct for FWER; here we will review two (very commonly-used) general procedures:

1) Bonferroni correction (simplest): it controls the family Type I error.

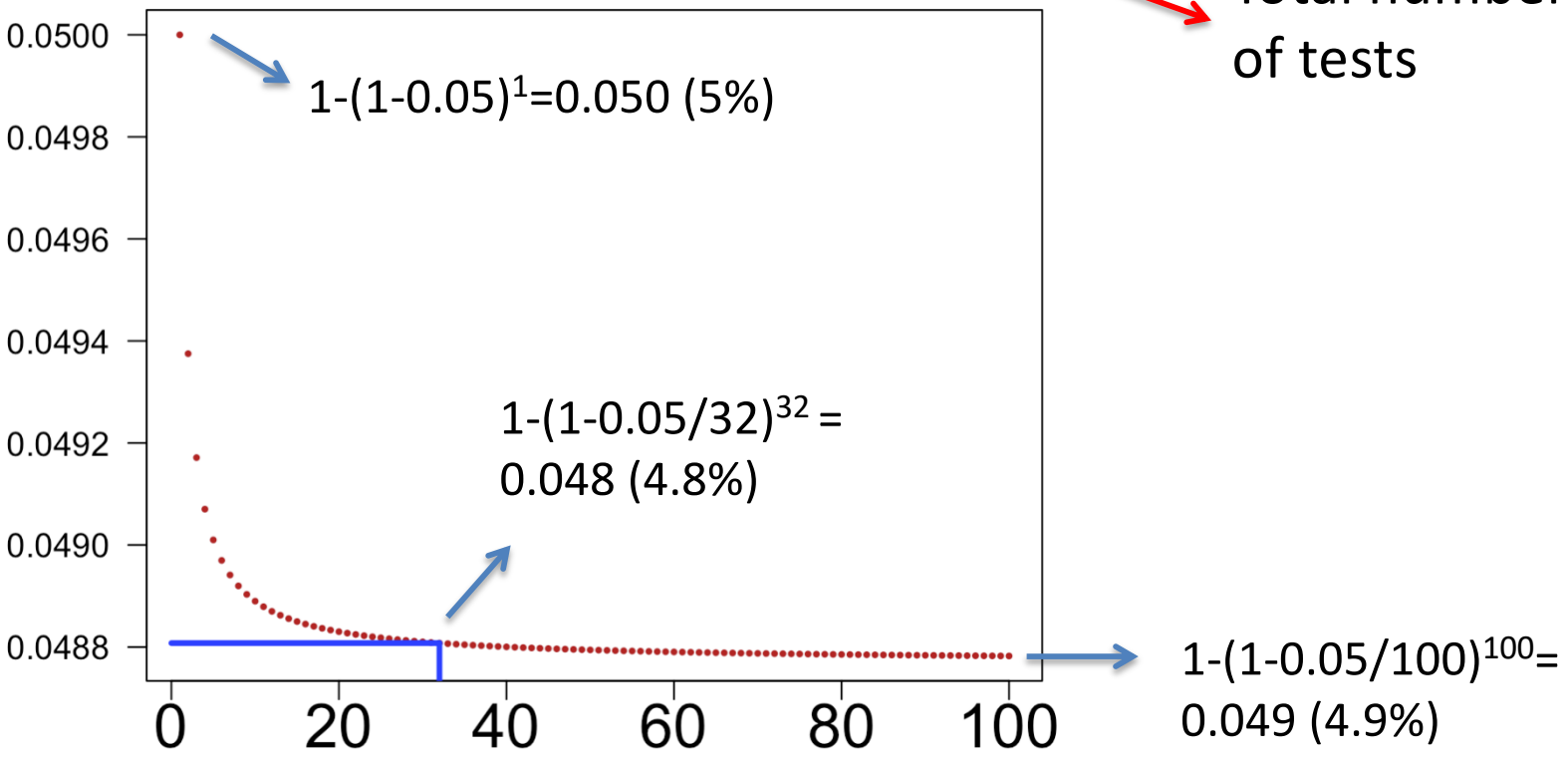
2) False Discovery Rate (FDR; very much used these days): it controls the false discovery rate.

Bonferroni correction

Carlo Emilio Bonferroni developed the correction; modern use credited to Olive Dunn

$$\alpha_{Bonfferroni} = \alpha/\text{m}=0.05/32= 0.0015625$$

Total number of tests



Instead of using the original pre-established (desired) α, use α adjusted by the number of test instead to assure a family-wise (type I) error rate (FWER).

Bonferroni correction

If we set an alpha of 0.05, i.e., acceptance area of 95% (0.95), then the chance of finding at least one significant test when you should not (i.e., false positive) out of 30 tests (as in our class survey) is: $1-(0.95)^{30}=1-(1-0.05)^{30}=0.78$

78% chance of finding at least 1 significant test when Ho is true in 30 statistical tests!

$$\alpha_{Bonfferroni} = \alpha/\mathrm{m}=0.05/32= 0.0015625$$

m: Total number of tests

$$1 - (1-\alpha_{Bonfferroni})^{32}= 1 - (1 - 0.0015625)^{32}= 0.04880777 \sim 0.05$$

$$P_{Bonfferroni} = m \times P$$

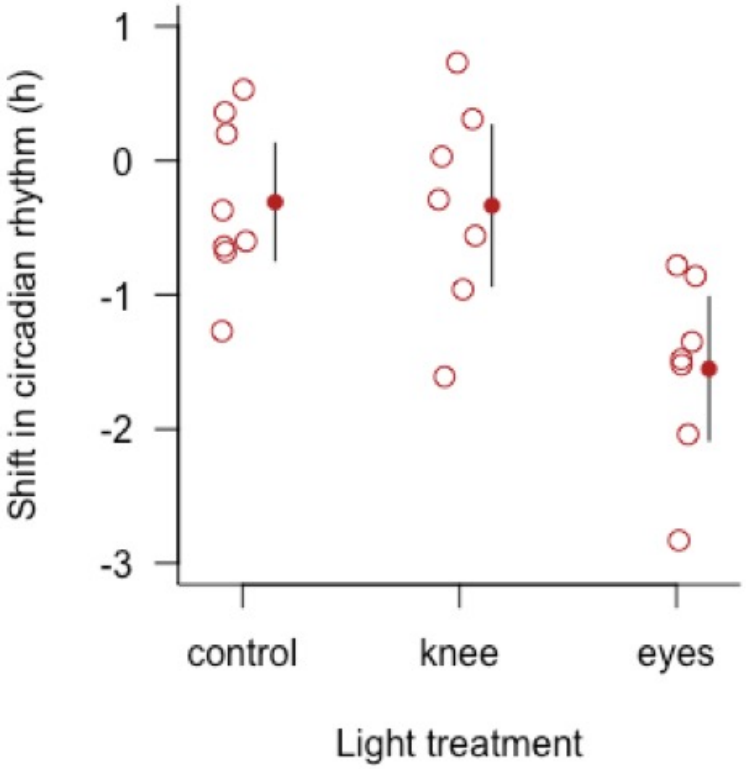
P: Original P value

$P_{Bonfferroni}$: Adjusted P value (adjusted P value that can be compared against any alpha)

Instead of using the original pre-established (desired) α, use α adjusted instead to guarantee a family-wise (type I) error rate (FWER).

This example - not so many pairwise tests, but still an issue

Source of variation	Sum of squares	*df*	Mean square	F	P
Between	202.5	1	202.5	81	0.0000185
Within	20	8	2.5		
Total	222.5	9			



Ho: $\mu_{control} = \mu_{knee} = \mu_{eyes}$

Ha: at least one μ is different from another u or other u_s; ***but which pairs?***

$$\overline{X}_{control} - \overline{X}_{knee}$$
$$\overline{X}_{control} - \overline{X}_{eyes}$$
$$\overline{X}_{knee} - \overline{X}_{eyes}$$

3 pairwise t-tests

Back to the problem about "The knees who say night"

Bonferroni correction

Either contrast the original P-value with α / number of tests
(below: 0.05 / 3 = 0.01667)

OR

Adjust the P-value as below and contrast with the original α (0.05)

$$P_{Bonferroni} = mP$$

Conclude based on these adjusted P-values

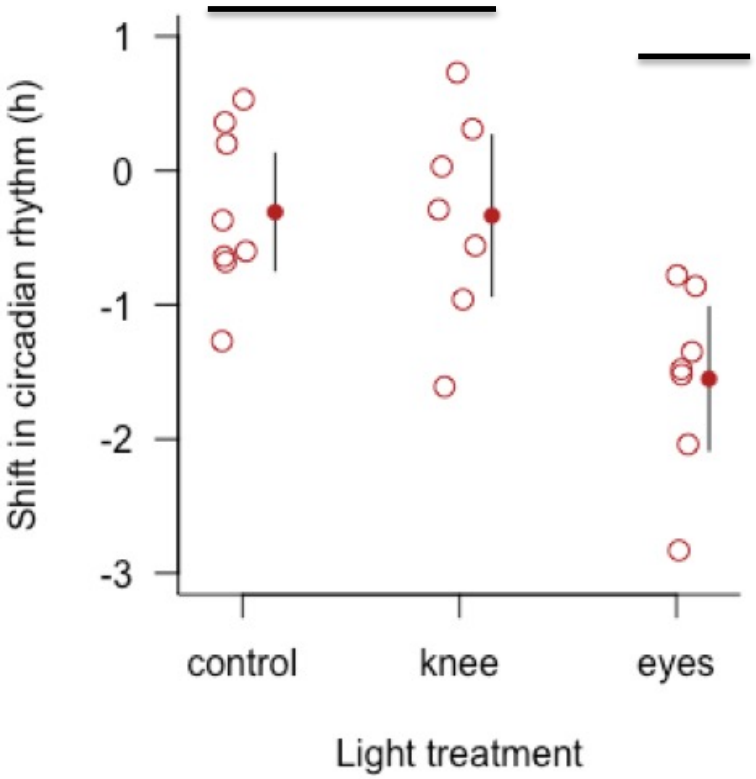
comparison	unocorrected P (t test)	Bonferroni P (t test)	
control vs eyes	0.0029	0.0088	← 3 x 0.0029
control vs knee	0.9418	1.0000	← 3 x 0.9418 = 2.8253
knee vs eyes	0.0044	0.0132	← 3 x 0.0044

Adjusted
α = 0.0166667

P-values greater than 1 are set to 1

Bonferroni correction (common table presentation)

comparison	unocorrected P (t test)	Bonferroni P (t test)
control vs eyes	0.0029	0.0088
control vs knee	0.9418	1.0000
knee vs eyes	0.0044	0.0132



The Tukey test or Tukey's HSD (honest significant difference) usually taught in Intro stats

1) is a solution to correct for single two-sample t-tests.

2) It works well for small number of pairwise comparisons but not large.

When you claim discovery,
how often are you wrong?

To learn from many signals (p-values), it's like medical diagnosis: stricter criteria reduce false alarms but cause more real cases to be missed.

False Discovery Rates - FDR (or false positive rate)
When you claim discovery, how often are you wrong?

A strict control on Type I error comes at a cost (a *trade-off*): as we reduce the chance of false positives, we dramatically increase the chance of false negatives (Type II errors).

In other words, we protect ourselves from being wrong, but we also prevent ourselves from discovering real effects. This is why Bonferroni-type corrections are often said to reduce the power of discovery.

This situation is called a *trade-off* because improving one goal necessarily worsens another.

When we tighten our criteria to reduce false positives, we make it harder for results to be declared significant. This protects us from being wrong, but it also causes real effects to be missed more often.

We cannot minimize both errors at the same time, so reducing one inevitably increases the other.

False Discovery Rates - FDR (or false positive rate)
When you claim discovery, how often are you wrong?

To learn from many signals (p-values), it's like medical diagnosis: stricter criteria reduce false alarms but cause more real cases to be missed.

Bonferroni asks a stricter question than FDR: "What is the probability of making at least one false positive?"

FDR asks a more pragmatic and often more relevant question: "Among the results I am declaring as significant (my discoveries), what proportion are likely to be false positives (Type I error) do to multiple testing?"

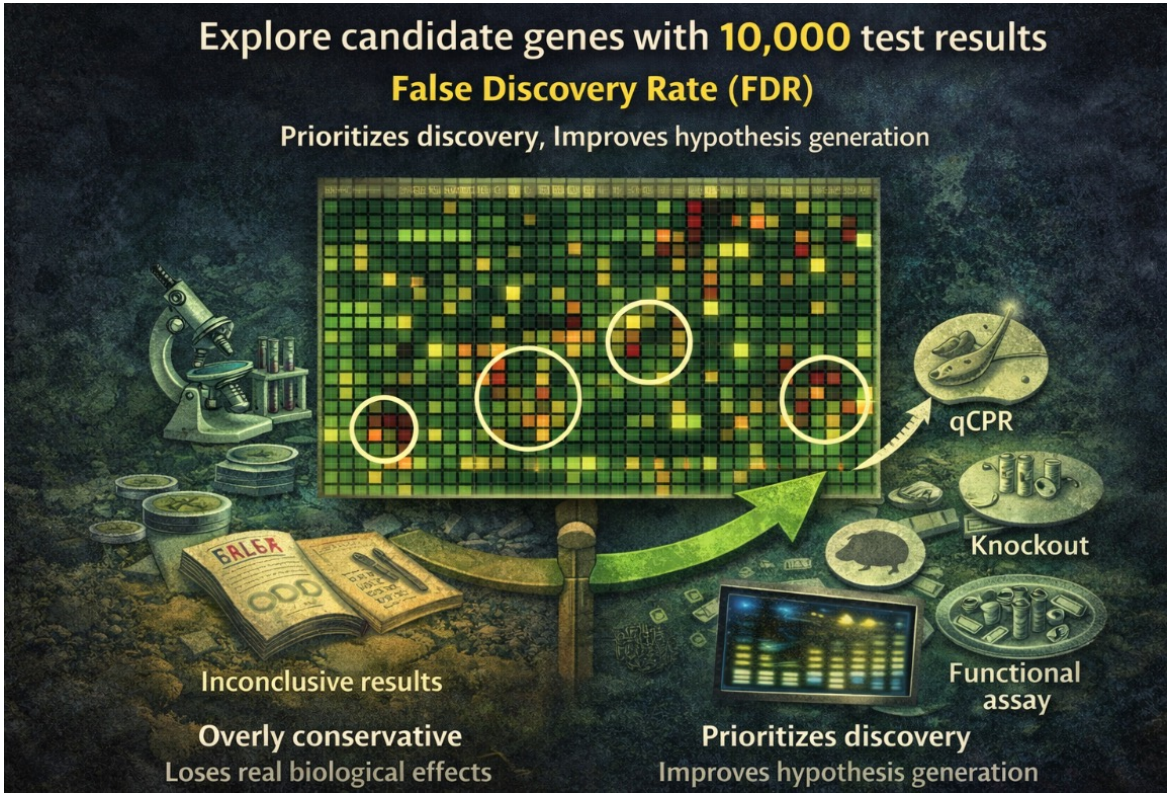
The philosophy behind Bonferroni-type corrections is different and much stricter.

Rather than evaluating tests one by one, FDR asks whether the collection of reported discoveries is mostly correct or substantially contaminated by false positives, by estimating the proportion of false positives among them.

False Discovery Rates - FDR (or false positive rate) versus Bonferroni (strict rules)

False Discovery Rates - FDR (or false positive rate) versus Bonferroni (strict rules)

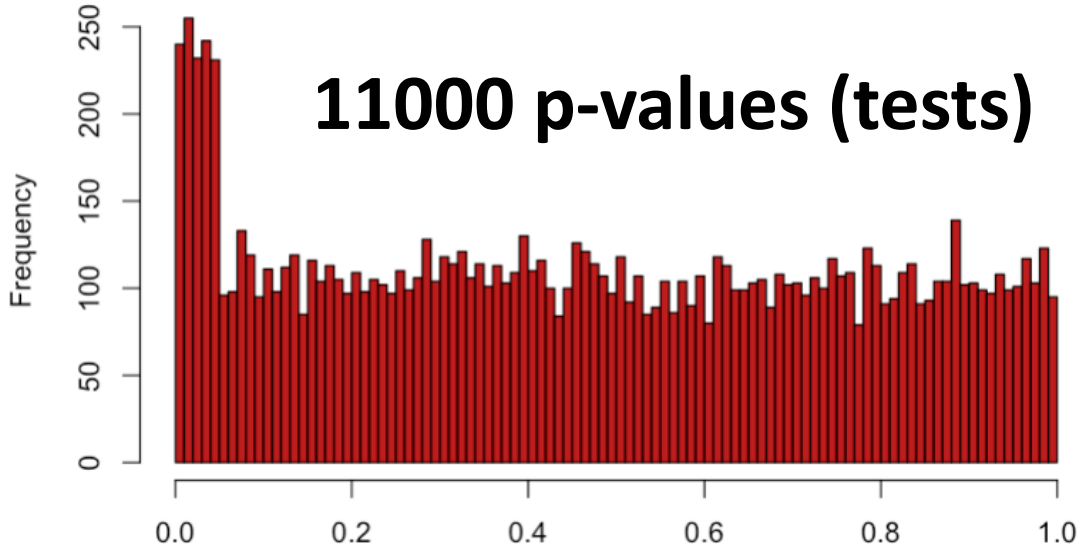
To learn from many signals (p-values), it's like medical diagnosis: stricter criteria reduce false alarms but cause more real cases to be missed.



Goal: identify candidate genes potentially involved in a biological response, which will later be validated using independent experiments

Rather than evaluating tests one by one, FDR asks whether the collection of reported discoveries is mostly correct or substantially contaminated by false positives, by estimating the proportion of false positives among them.

Bonferroni versus FDR (quick contrast)

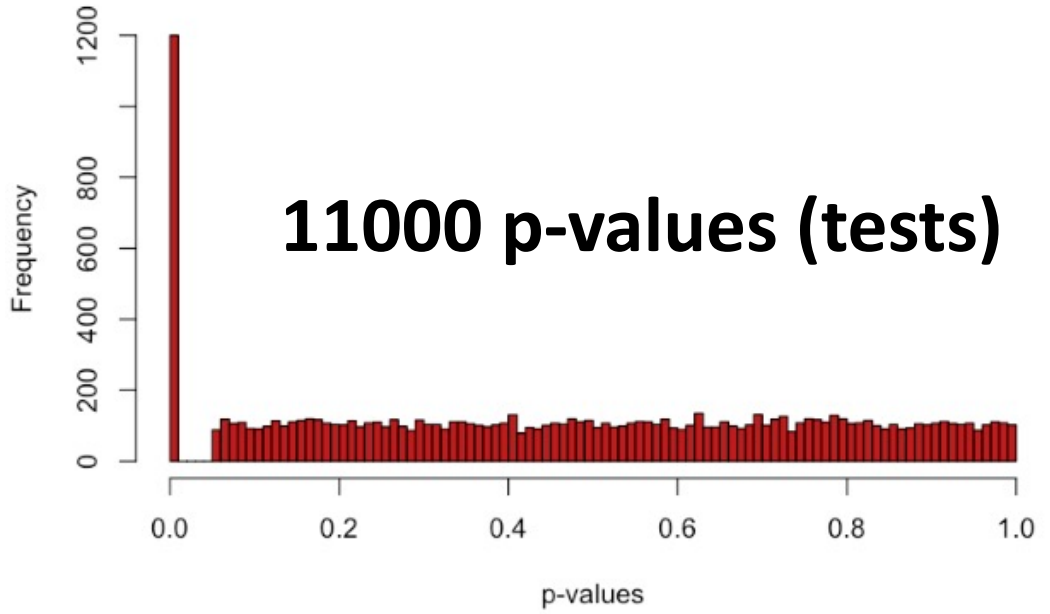


Number of significant tests after adjustment

Bonferroni = 0

FDR = 0

FDR logic: controlling the False Discovery Rate does not guarantee findings; it guarantees that any findings we choose to report are not expected to be heavily contaminated by false positives.



Bonferroni = 2

FDR = 1200

False Discovery Rates is widely used!

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2011, **2**, 278–282 doi: 10.1111/j.2041-210X.2010.00061.x

Using false discovery rates for multiple comparisons in ecology and evolution

Nathan Pike*

Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

Statistical significance for genomewide studies

John D. Storey*† and Robert Tibshirani‡

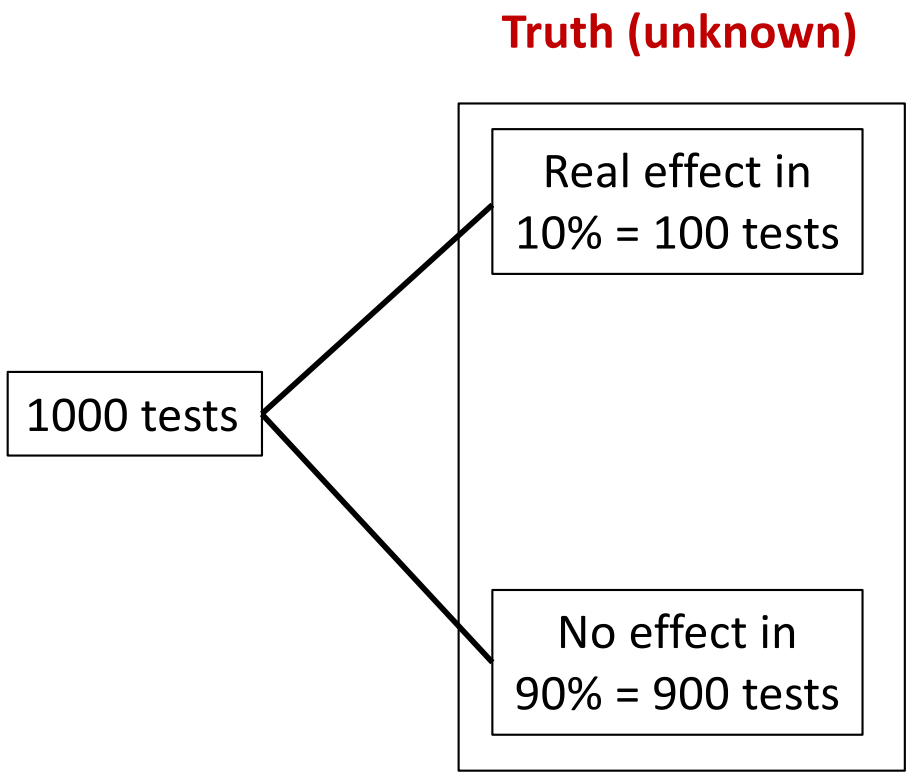
*Department of Biostatistics, University of Washington, Seattle, WA 98195; and ‡Departments of Health Research and Policy and Statistics, Stanford University, Stanford, CA 94305

9440–9445 | PNAS | **August 5, 2003** | vol. 100 | no. 16

PNAS PNAS

False Discovery Rates

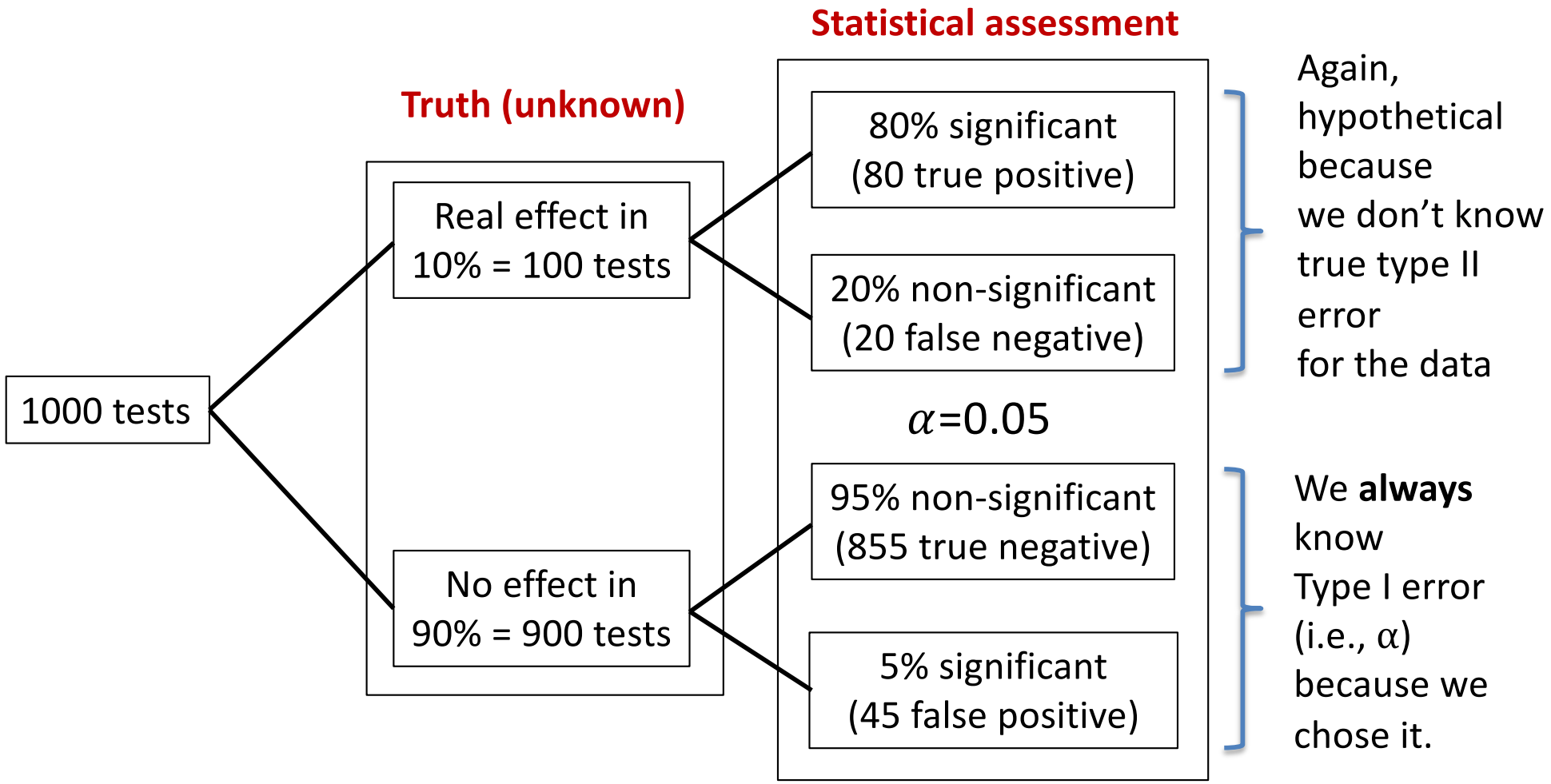
Let's assume a hypothetical (fictional) example where we know the truth about which outcomes are significant and non-significant so that we can better understand the logic behind FDR.



Adapted from Colquhoun 2017

False Discovery Rates

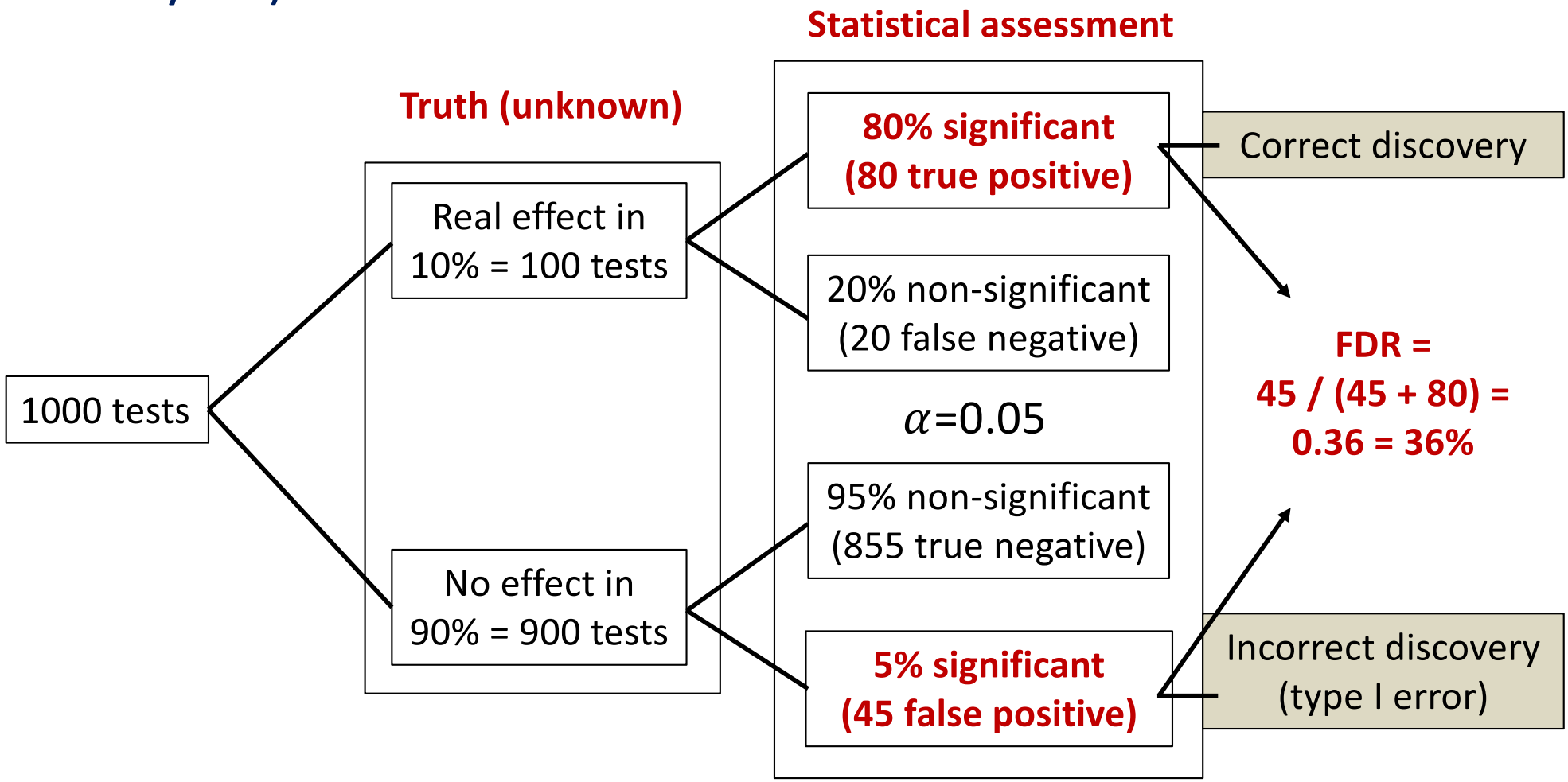
Hypothetical (fictional) example where **we know the truth**



Adapted from Colquhoun 2017

False Discovery Rates

So, based on an α=0.05, one will be wrong 36% of the time when rejecting H_0 (claiming discovery). **So, the probability of true discovery is 64% (i.e., 100-36%; 36% being the False Discovery Rate).**



Remember - when you reject H_0 you discover something new

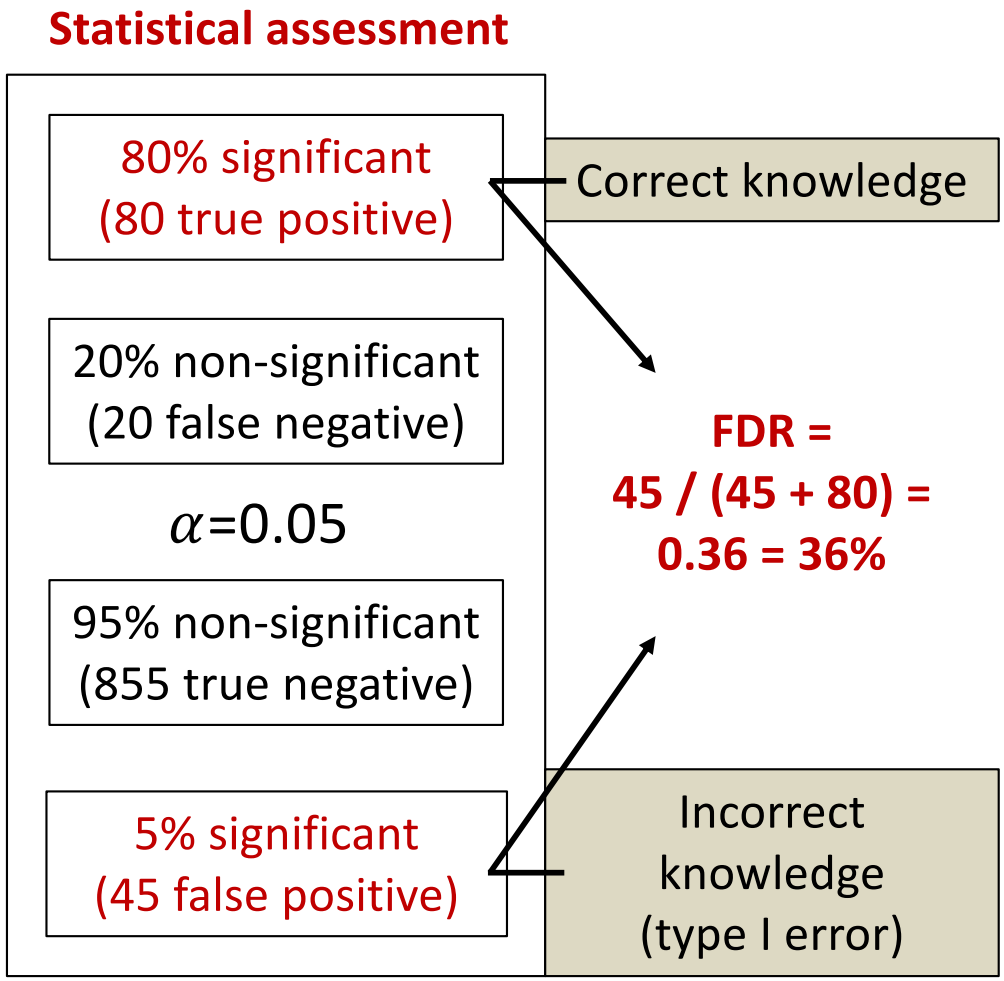
Adapted from Colquhoun 2017

False Discovery Rates.- we are interested in positives (i.e., significant P-values) because those are "discoveries" – some likely wrong and some correct

Based on an α=0.05, in this case, we will be wrong 36% of the time if we reject H_0 (claiming discovery). So, the probability of true discovery (reject a false H_0) is 64%.

The goal is to reduce the FDR to say 0.05 instead of keeping it at 0.36! So that the true discovery is higher (0.95 = 95%)

How to estimate FDR based on real data where we don't know the truth about false positives and negative as in this example?

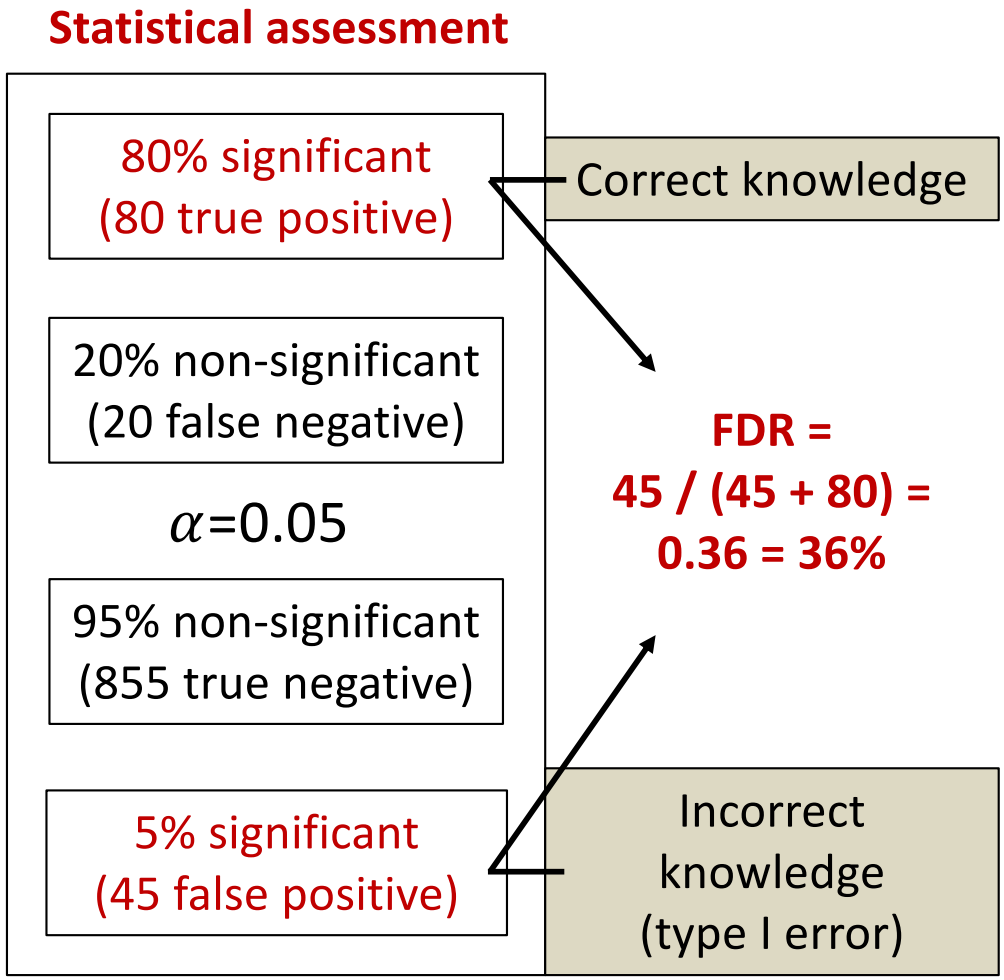


Remember - when you reject H_0 you discover something new

False Discovery Rates.- we are interested in positives (i.e., significant P-values) because those are "discoveries" – some likely wrong and some correct

Interpretation: 36% is **not** the probability that any single result is wrong.

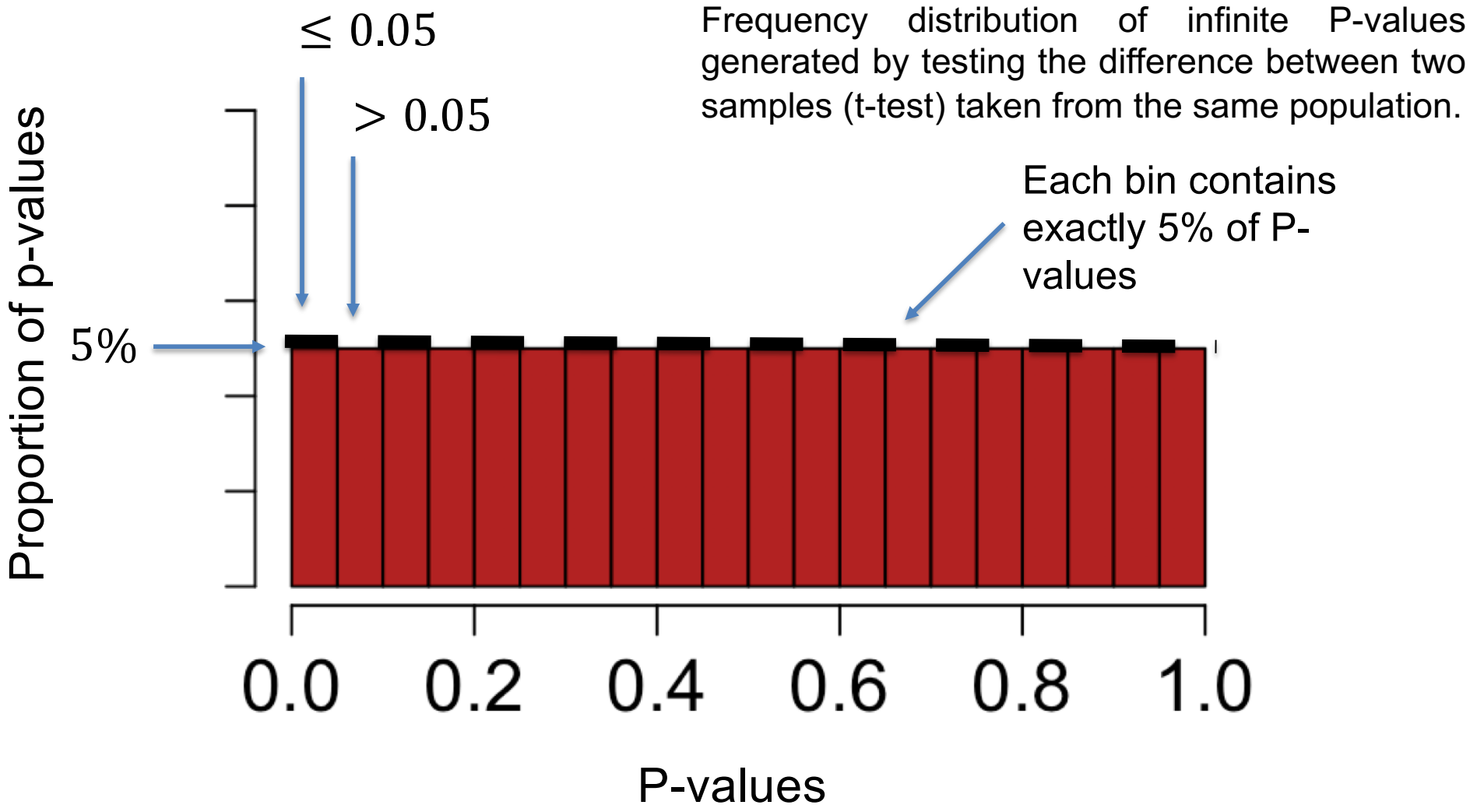
It is a **rate describing the whole set of discoveries**:
if you were to repeat the study many times, about **36% (in average) of the results you call significant would be a false positive, on average**.



Remember - when you reject H_0 you discover something new

FDR then requires an estimate of the number of true positives!

Required knowledge (Step 1): Understand that when samples (e.g., control versus treatment) come from the same population (H_0 is true), the frequency distribution of P-values is flat (uniform).



FDR then requires an estimate of the number of true positives!

Required knowledge (Step 1): Understand that when samples or groups (e.g., control versus treatment) come from the same population (i.e., H_0 is true), the frequency distribution of P-values is flat (uniform).

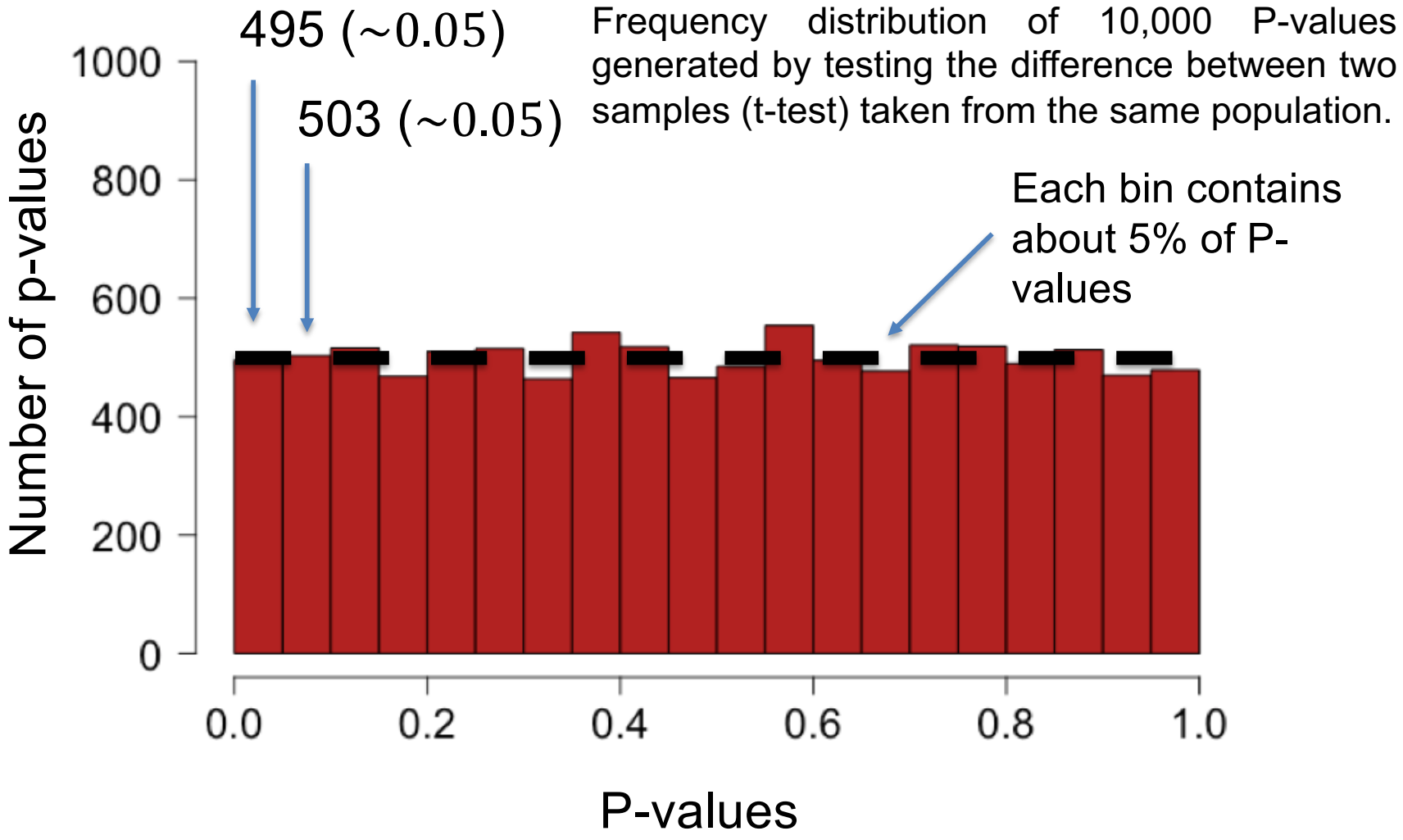
```
vector.pvalues <- matrix(0,1000)
for (i in 1:10000){
  x1 <- rnorm(20,5,2)
  x2 <- rnorm(20,5,2)
  vector.pvalues[i] <-
    t.test(x1, x2, alternative = "two.sided", var.equal = FALSE)$p.value
}
hist(vector.pvalues,ylim=c(0,1000),col="firebrick")
```

Same populations

How to estimate FDR based on real data where we don't know the truth about false positives and negative as in this example?

FDR then requires an estimate of the number of true positives!

A simulation to show that when samples (e.g., control versus treatment) come from the same population (H_0 is true), the frequency distribution of P-values is flat (uniform).



FDR then requires an estimate of the number of true positives!

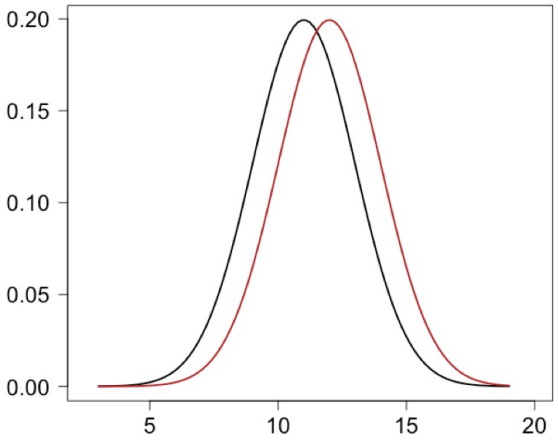
Required knowledge (Step 2): Understand that when samples (e.g., control versus treatment) come from different populations (H_0 is false), the frequency distribution of P-values is not flat (not uniform).

```
vector.pvalues <- matrix(0,1000)
for (i in 1:10000){
  x1 <- rnorm(20,10,2)
  x2 <- rnorm(20,11,2)
  vector.pvalues[i] <-
    t.test(x1, x2, alternative = "two.sided", var.equal = FALSE)$p.value
}
hist(vector.pvalues,ylim=c(0,1000),col="firebrick")
```

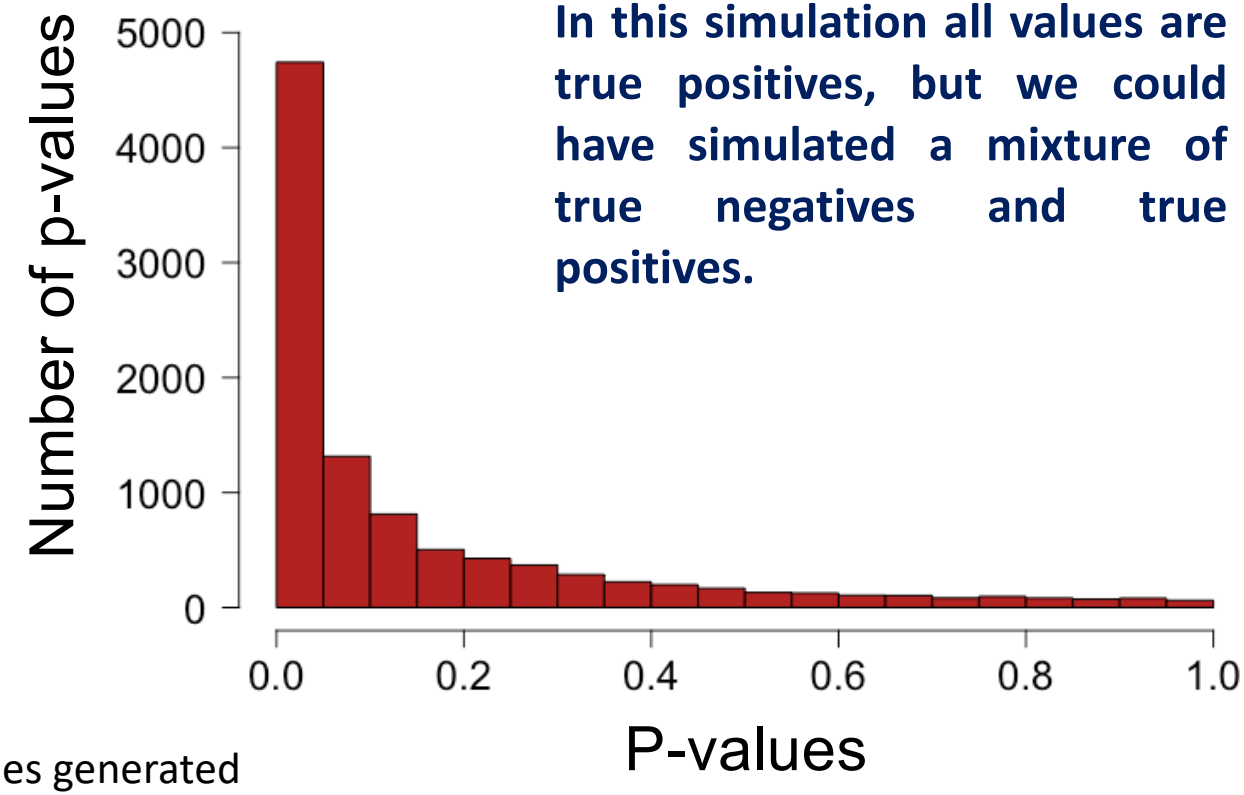
different populations

FDR then requires an estimate of the number of true positives!

Required knowledge (Step 2): Understand that when samples (e.g., control versus treatment) come from different populations (H_0 is false), the frequency distribution of P-values is not flat (not uniform).



$$\mu_1 = 10$$
$$\mu_2 = 11$$

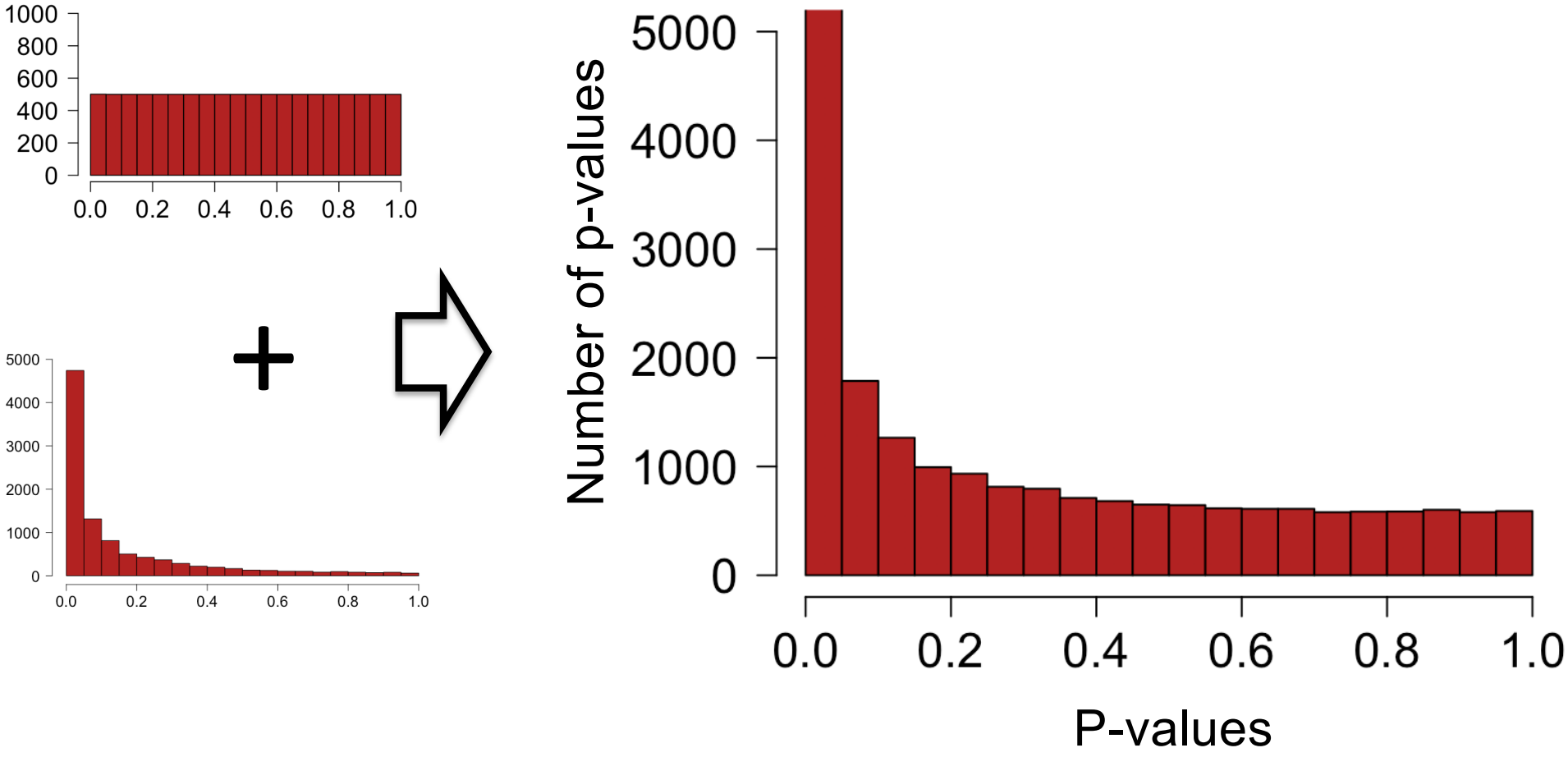


In this simulation all values are true positives, but we could have simulated a mixture of true negatives and true positives.

Frequency distribution of 10,000 P-values generated by testing the difference between two samples (t-test) taken from different populations.

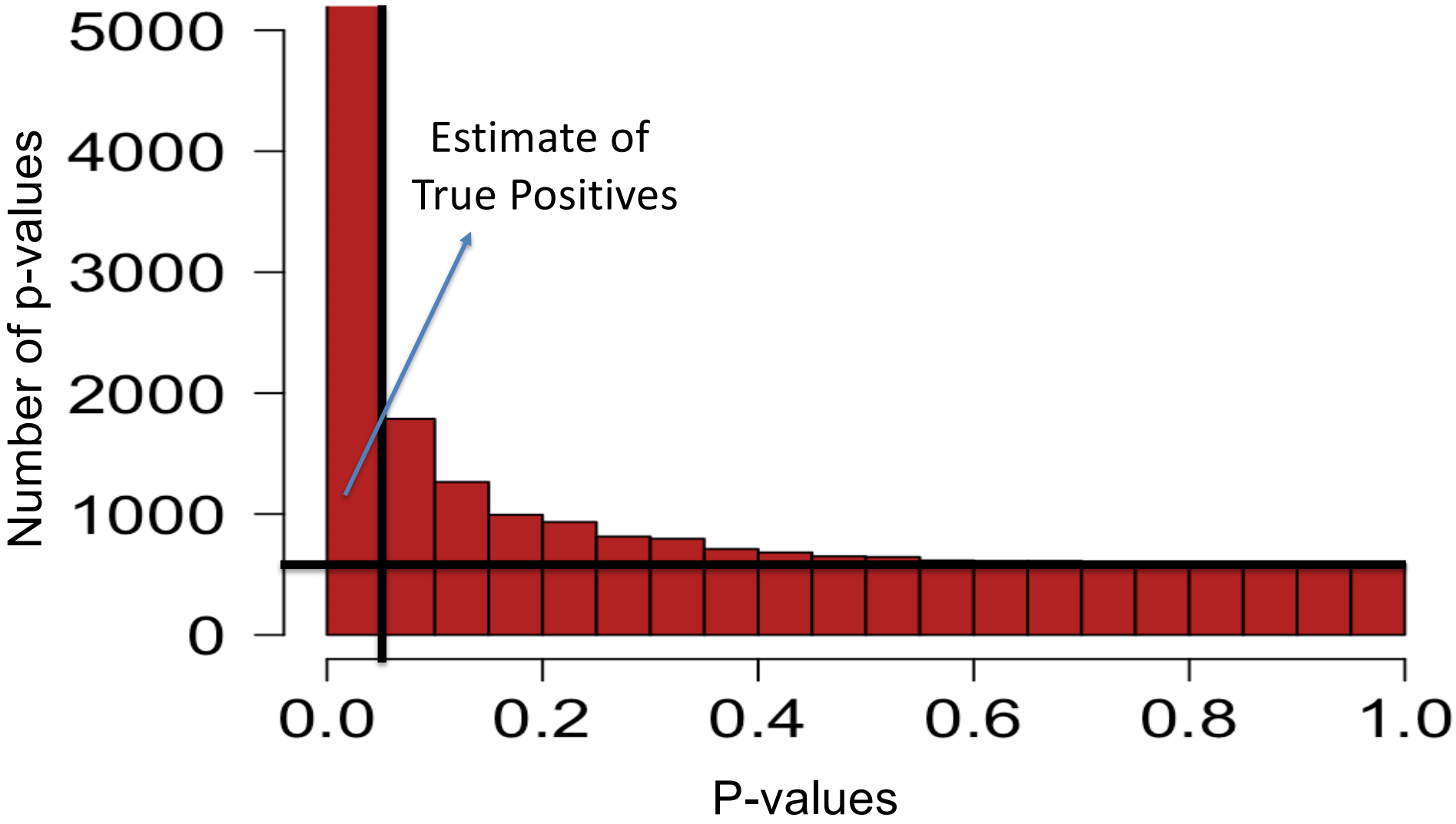
FDR then requires an estimate of the number of true positives!

Required knowledge (Step 3): Understand the concept of mixing the two types of distributions (i.e., H_0 is true (uniform) and H_0 is unknown). In reality, the distribution of P-values in most studies reflects a mixture of tests for which there is a real effect (H_0 is false) and tests for which there is no real effect (H_0 is true).



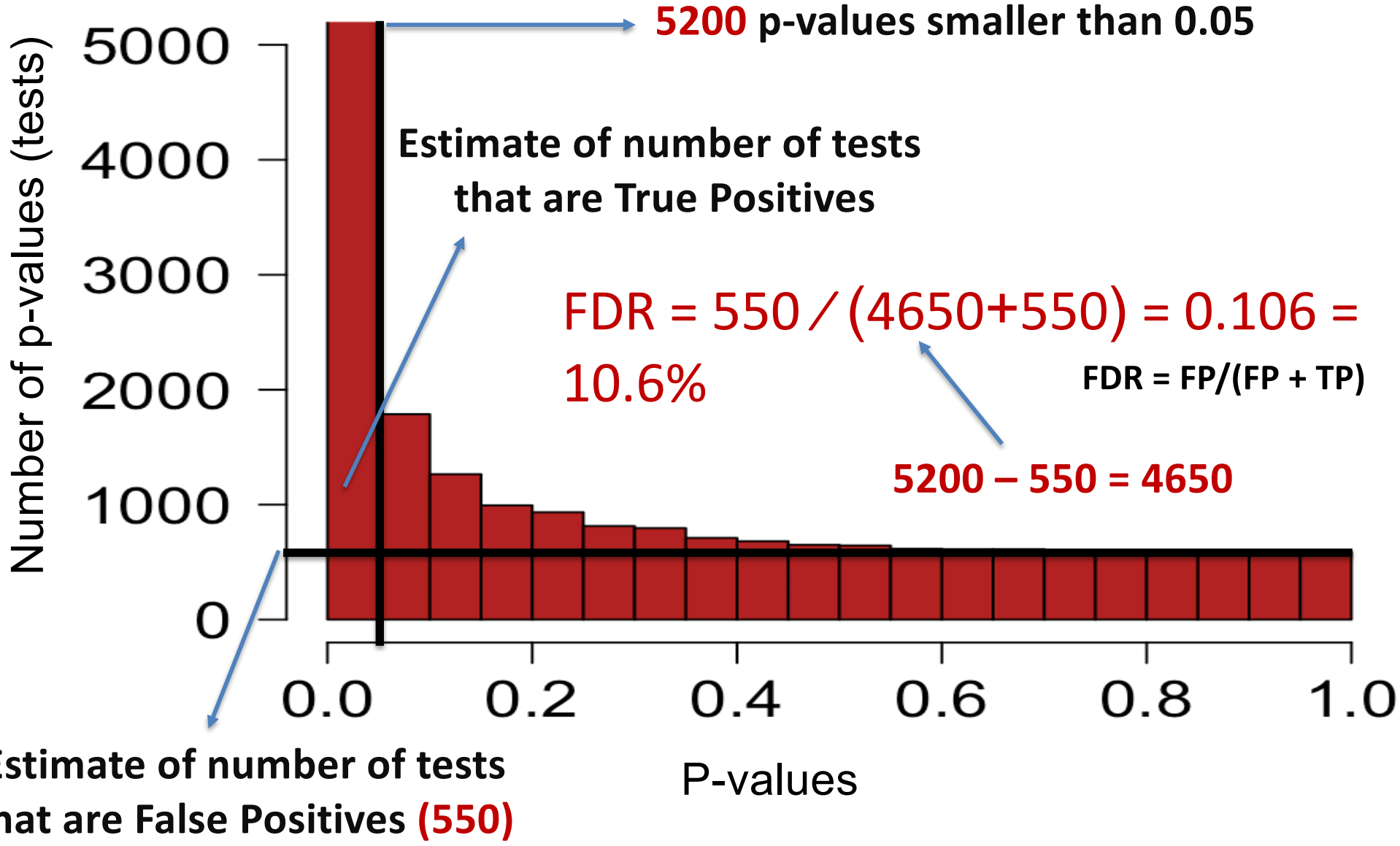
FDR then requires an estimate of the number of true positives!

Required knowledge (Step 4): FDR estimates how often our decisions are right or wrong, while accepting that uncertainty remains even after correction.

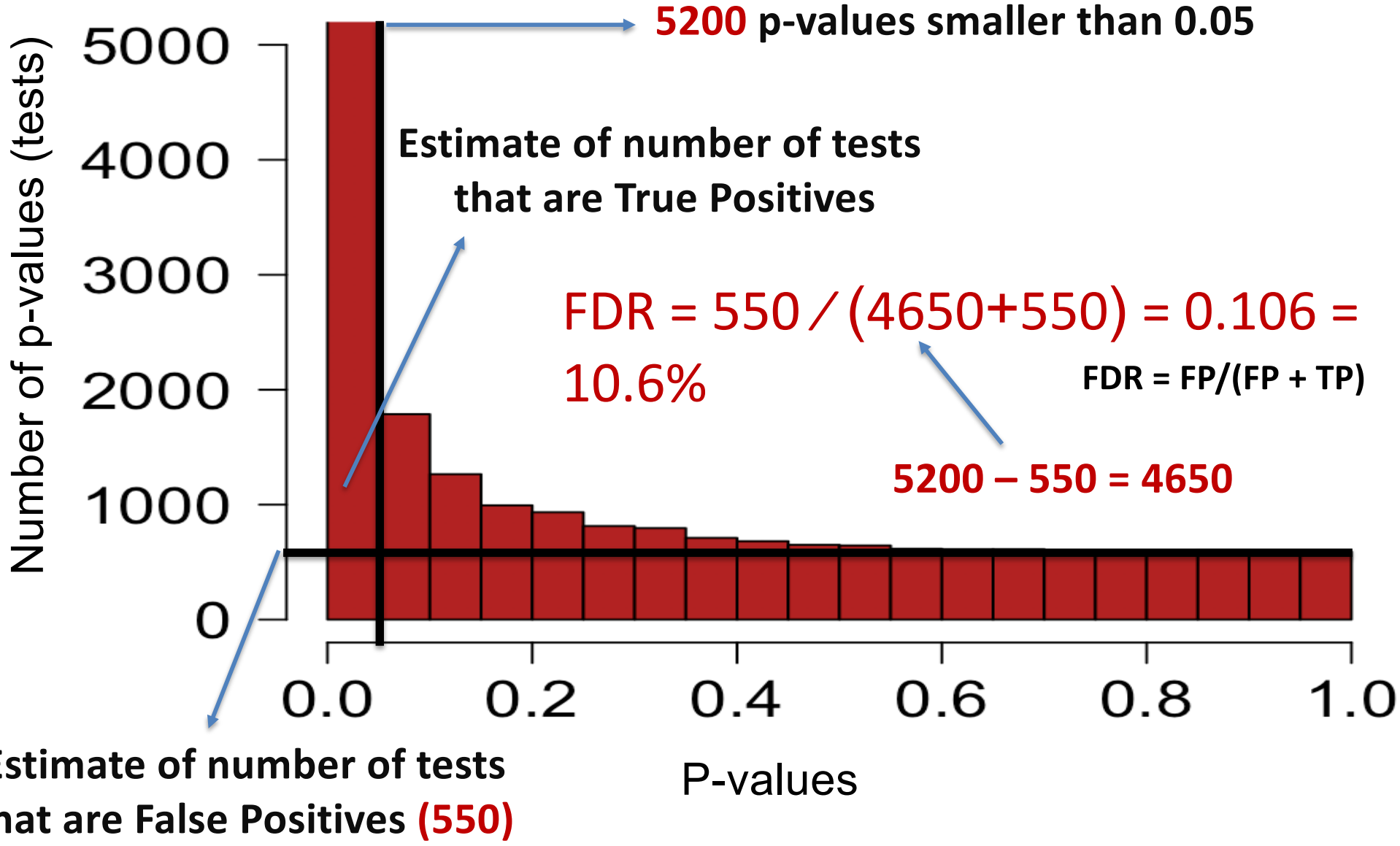


FDR then requires an estimate of the number of true positives!

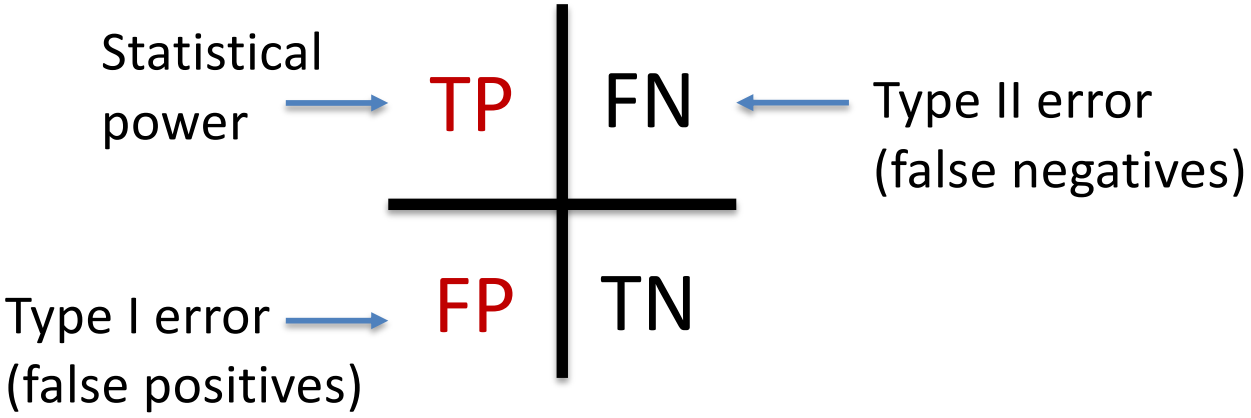
Required knowledge (Step 4): **Estimate** (i.e., you could still be wrong after correction) fractions based on different potential successes (true rejections or true non-rejections) and different failures (false positives or false negatives).

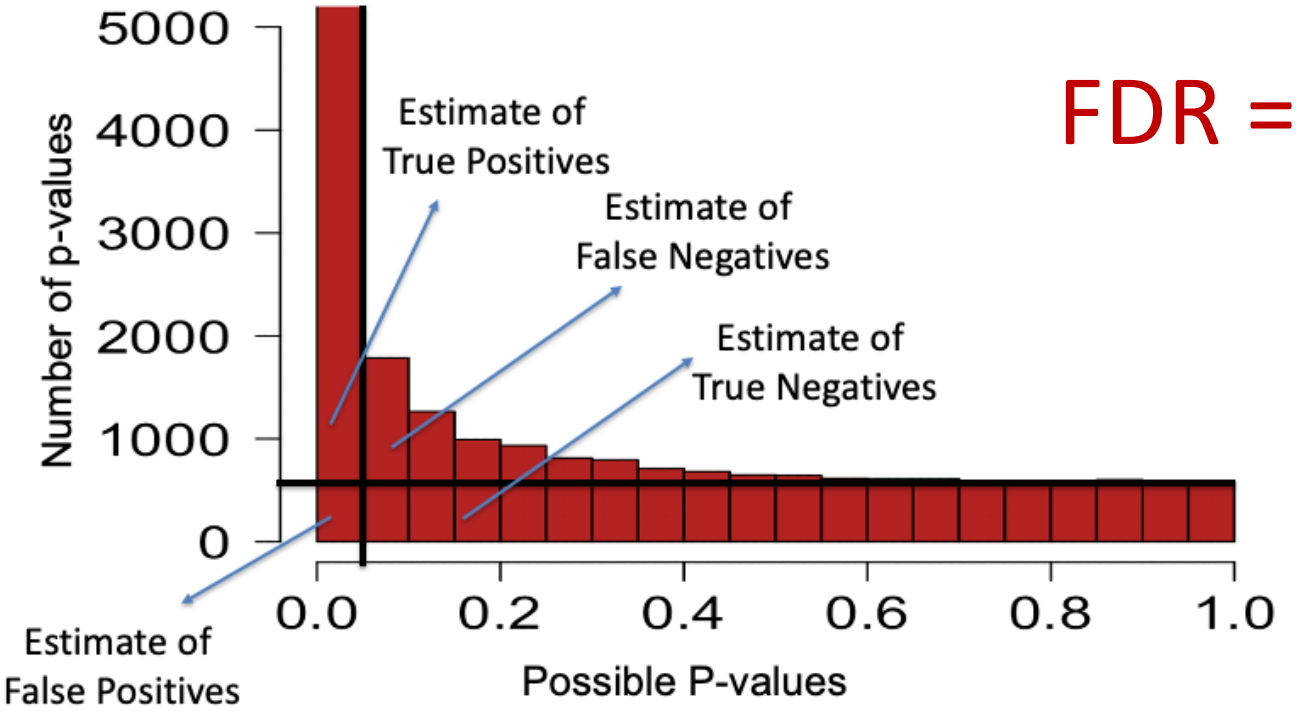


Interpretation: 10.6% is **not** the probability that any single result is wrong. It is a **rate describing the whole set of discoveries**: if you were to repeat this type of analysis many times, about **1 out of every 10 results (~10%) you call significant would be a false positive, on average**.



FOR COMPLETION!!!!



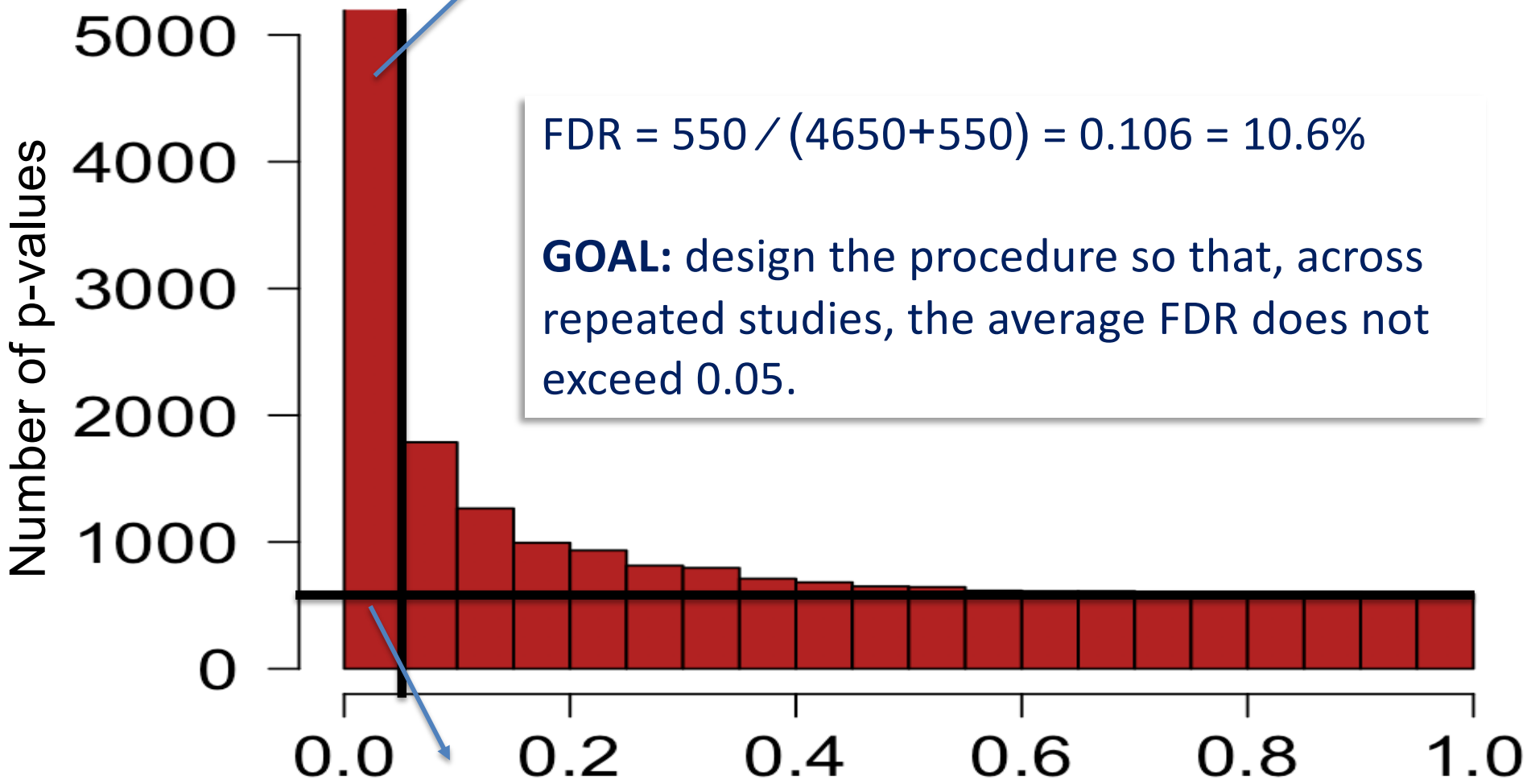


$$FDR = FP / (TP + FP)$$

FDR then requires an estimate of the number of true positives!

Estimate of True Positive

$$FDR = FP / (TP + FP)$$



FDR = 550 / (4650+550) = 0.106 = 10.6%

GOAL: design the procedure so that, across repeated studies, the average FDR does not exceed 0.05.

5% of the false positive are considered as significant;
FP is an estimate, so some could be actually TP.

False Discovery Rates - FDR (or false positive rate)
How much did you learn that was false positive?

The are different types of FDR procedures and the one by Benjamini-Hochberg is likely the most commonly used! To correct the P-values based on the BH-FDR procedure, the calculation is conditional on previous P-values. R does it for you!!

Whatever the FDR is, the goal is to design the procedure so that, across repeated studies, the average FDR does not exceed 0.05.

Step 5: Adjust probabilities based on the FDR principle (NOT CRITICAL TO KNOW)

Rank (i)	Original P-value	BH multiplier (m / i)	BH-adjusted P-value
1	0.01	10.00	0.10
2	0.11	5.00	0.55
3	0.21	3.33	0.70
4	0.31	2.50	0.78
5	0.41	2.00	0.82
6	0.51	1.67	0.85
7	0.61	1.43	0.87
8	0.71	1.25	0.89
9	0.81	1.11	0.90
10	0.91	1.00	0.91

No significant p-value based on the FDR logic: controlling the False Discovery Rate does not guarantee findings; it guarantees that any findings we choose to report are not expected to be heavily contaminated by false positives.

Journal of Research on Educational Effectiveness, 5: 189–211, 2012

ISSN: 1934-5747 print / 1934-5739 online
DOI: 10.1080/19345747.2011.618213

Routledge
Taylor & Francis Group

Some Bayesian dissent

METHODOLOGICAL STUDIES

Why We (Usually) Don't Have to Worry About Multiple Comparisons

Andrew Gelman
Columbia University, New York, New York, USA

Jennifer Hill
New York University, New York, New York, USA

Masanao Yajima
University of California, Los Angeles, Los Angeles, California, USA

Main issues from a Bayesian perspective (my summary):

1) FWER (family wise error, e.g., Bonferroni) is the general goal and this is an issue because it puts sole emphasis on Type I error (even FDR in many ways);
2) issues with dependent tests;
3) FDR good for very large number of tests but Bayesians may not recommend it for small numbers.

Bottom line: journals will request multiple testing and routine procedures are easier to implement and "articulate" than Bayesian ones. So…for the majority of scientists, Type I error is a really BIG ISSUE and needs to be dealt with using appropriate adjustments!

What should be corrected for?

- Variance and multiple t tests?
- All tests in a paper?
- All tests across all papers within a journal issue?
- All test across all papers within a year
- The world is the limit!

Look into this blog (*Why you don't need to adjust your alpha level for all tests you'll do in your lifetime*):
http://daniellakens.blogspot.com/2016/02/why-you-dont-need-to-adjust-you-alpha.html

I don’t necessarily agree with everything in there, but good food for thought!