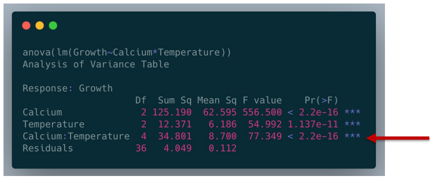
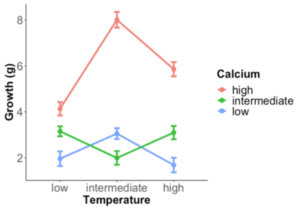


```
anova(lm(Growth~Calcium*Temperature))
Analysis of Variance Table

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
```



Regarding the interaction, there are 3 groups of Calcium and 3 groups of temperature (9 means). There are 36 possible pairwise tests to contrast Growth across levels (9 x 8/2 = 36).

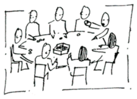
1

Why do we conduct ANOVAs and not simply test pairs of means?

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

A pedagogical guide for understanding the issues underlying

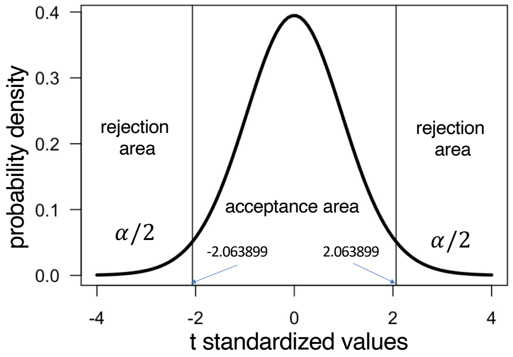
Multiple hypothesis testing



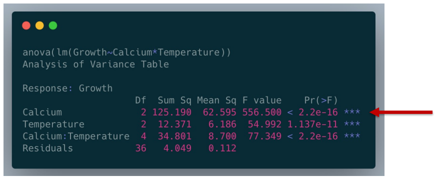
Why should we not trust the results from multiple statistical tests?

2

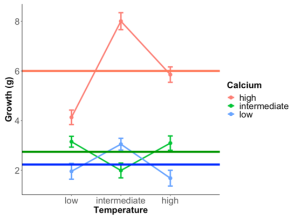
Guided discussion - t-distribution assuming H_0 as true



3



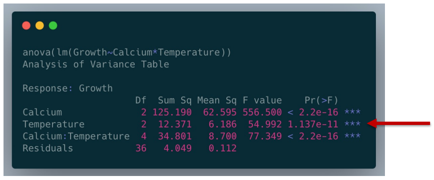
```
anova(lm(Growth~Calcium*Temperature))
Analysis of Variance Table

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
```



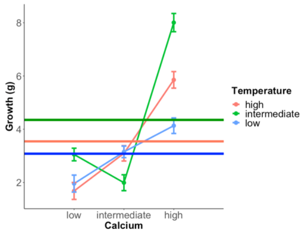
Regarding Calcium, there are 3 possible pairwise tests contrast Growth across levels (3 x 2/2 = 3).

High – intermediate
High – low
Intermediate – low

4



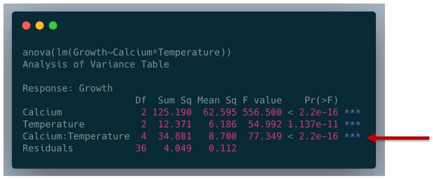
```
anova(lm(Growth~Calcium*Temperature))
Analysis of Variance Table

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
```



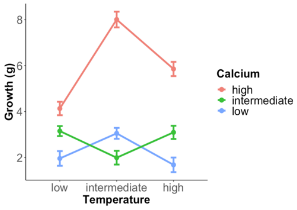
Regarding Temperature, there are 3 possible pairwise tests contrast Growth across levels (3 x 2/2 = 3).

High - intermediate
High - low
Intermediate - low

5

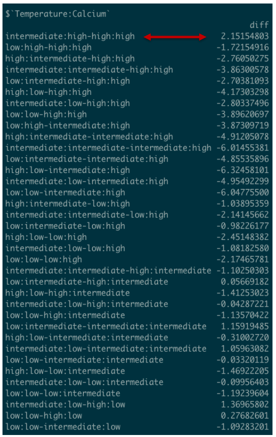


```
anova(lm(Growth~Calcium*Temperature))
Analysis of Variance Table

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
```



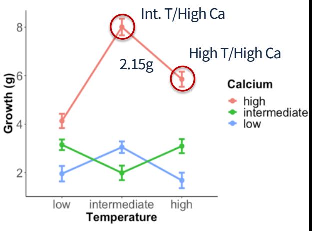
Regarding the interaction, there are 3 groups of Calcium and 3 groups of temperature (9 means). There are 36 possible pairwise tests to contrast Growth across levels (9 x 8/2 = 36).

6



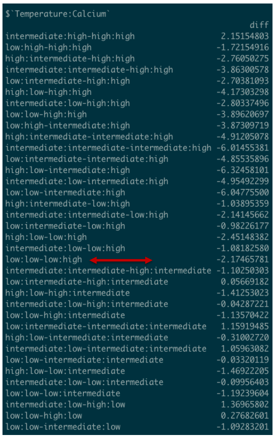
```
$`Temperature:Calcium`
                                                   diff
intermediate:high-high:high                  2.15154803
low:high-high:high                          -1.72154916
high:intermediate-high:high                 -2.76050275
intermediate:intermediate-high:high         -3.86300578
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).



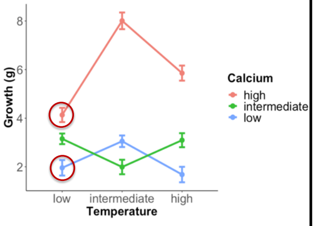
Does the mean growth in intermediate T and high Ca differ significantly from the mean growth in high T and high Ca? Difference = 2.15g.

7



```
$`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).



Does the mean growth in low T and low Ca differ significantly from the mean growth in low T and high Ca? Difference = 2.17g.

8

What happens when we conduct too many statistical tests?

A past classroom demonstration using a survey

9

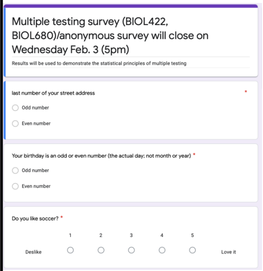
Past classroom surveys:
Would you expect odd- and even day born individuals to differ in their preferences?

	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		

10



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) *

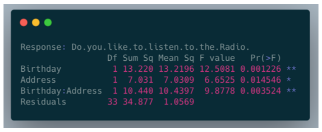
○ Odd number

○ Even number

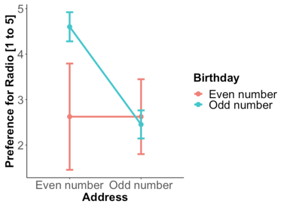
Do you like soccer? *

	1	2	3	4	5	
Deslike	○	○	○	○	○	Love it

Really ????



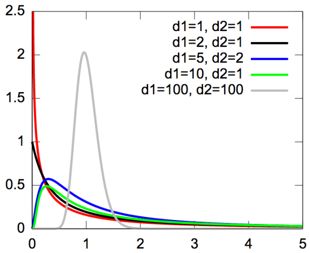
```
Response: Do.you.like.to.listen.to.the.Radio.
                 Df Sum Sq Mean Sq F value   Pr(>F)
Birthday          1 13.220 13.2196 12.5081 0.001226 **
Address           1  7.031  7.0309  6.6525 0.014546 *
Birthday:Address  1 10.440 10.4397  9.8778 0.003524 **
Residuals        33 34.877  1.0569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



class survey:
24 questions

11

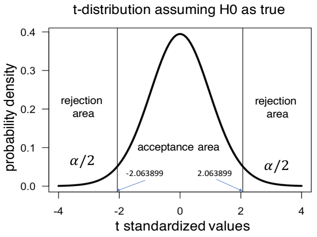
Why when comparing multiple mean values, one should start with an ANOVA and not multiple t-test



Probability of committing 1 type I error (false positive) is the same for 1 or multiple tests (α), but conducting 100 tests, there will be a chance of 5 being significant for an α = 0.05

12

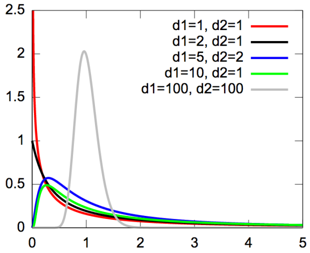
Why when comparing multiple mean values, one should start with an ANOVA and not multiple t-test



Probability of committing 1 type I error (false positive) is the same for 1 or multiple tests (α), but conducting 100 tests, there will be a chance of 5 being significant for an α = 0.05

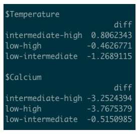
13

Why when comparing multiple mean values, one should start with an ANOVA and not multiple t-test



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.

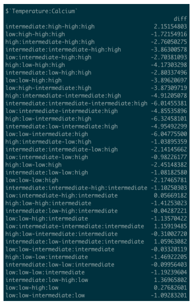
14



```
$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

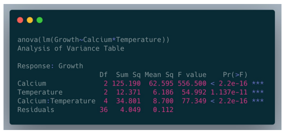
3 pairwise tests



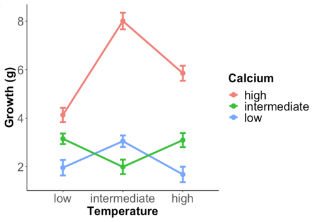
```
$`Temperature:Calcium`
                                                   diff
intermediate:high-high:high                  2.15154803
low:high-high:high                          -1.72154916
high:intermediate-high:high                 -2.76050275
intermediate:intermediate-high:high         -3.86300578
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

ANOVA = 3 tests
pairwise t-tests = 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
```



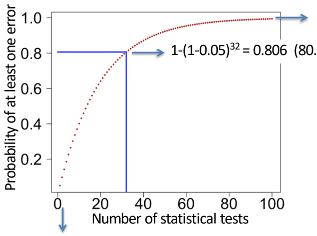
15

16

If we set an alpha of 0.05, i.e., acceptance area of 95% (0.95), then the chance of at least one significant test by chance (i.e., null hypothesis is true) when one should not (i.e., false positive) out of 32 tests is:

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

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



$1-(1-0.05)^{100}=0.9941$ (99.4%)

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

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

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

17

Examples of really huge numbers of multiple tests

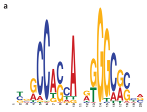
How does multiple testing correction work?

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.



68 million statistical tests

b

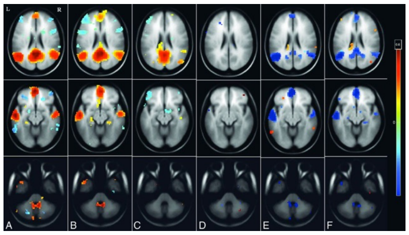
Position	Str	Sequence	Score
19390631	+	TTGACCAGCAGGGGGCGCCG	26.30
32420105	+	CTGGCCAGCAGAGGGCAGCA	26.30
27910537	–	CGGTGCCCCCTGCTGGTCAG	26.18
21968106	+	GTGACCACCAGGGGGCAGCA	25.81
31409358	+	CGGGCCTCCAGGGGGCGCTC	25.56
19129218	–	TGGCGCCACCTGCTGGTCAC	25.44
21854623	+	CTGGCCAGCAGAGGGCAGGG	24.95
12364895	+	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.

18

Examples of really huge numbers of multiple tests

Compare signal changes using t-test (task versus no-task) across thousands of voxels (brain pixels in 3D)



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

19

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.

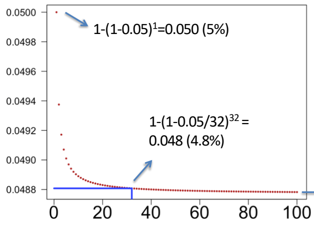
20

Bonferroni correction

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

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

Total number of tests



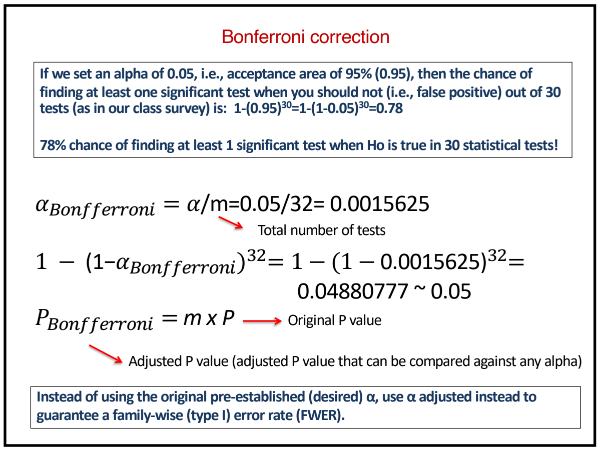
$1-(1-0.05)^1$=0.050 (5%)

$1-(1-0.05/32)^{32}$ = 0.048 (4.8%)

$1-(1-0.05/100)^{100}$= 0.049 (4.9%)

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).

21



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!

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

Total number of tests

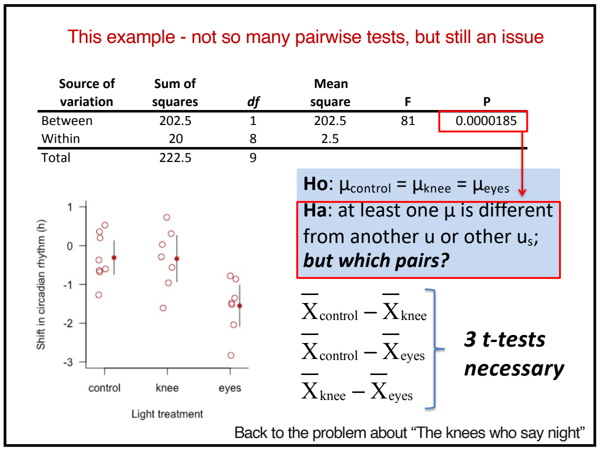
$$1 - (1-\alpha_{Bonfferroni})^{32} = 1 - (1 - 0.0015625)^{32} = 0.04880777 \sim 0.05$$

$P_{Bonfferroni} = m \; x \; P$ → Original P value

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).

22



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

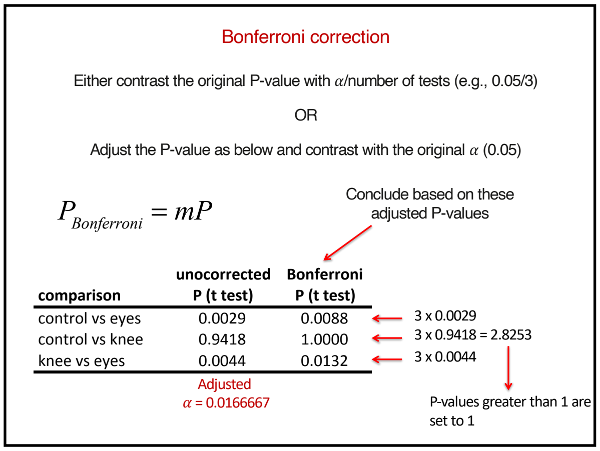
Ha: at least one μ is different from another u or other u_s; ***but which pairs?***

$$\left.\begin{array}{l}\overline{X}_{control} - \overline{X}_{knee}\\ \overline{X}_{control} - \overline{X}_{eyes}\\ \overline{X}_{knee} - \overline{X}_{eyes}\end{array}\right\}$$

3 t-tests necessary

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

23



Bonferroni correction

Either contrast the original P-value with α/number of tests (e.g., 0.05/3)

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

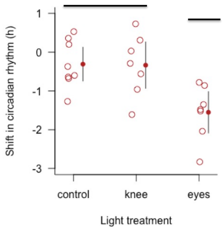
Adjusted
α = 0.0166667

P-values greater than 1 are set to 1

24

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 comparing two-sample means only (i.e., based on t-tests).

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

25

26

False Discovery Rates - FDR (or false positive rate)

How much did you learn that was based on false positives?

Adjustments for multiple tests like the Bonferroni put too much emphasis on controlling for false positives (Type I error) BUT not false negatives (Type II error); thus, they reduce the "power of discovery".

The FDR philosophy: To be "precise", you need to ESTIMATE how often you could be right when you declare a result to be significant (avoid false negatives) and ESTIMATE how often you could be wrong when you declare a result to be significant (avoid false positives).

27

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!!

Gather all tests that lead to a statistically significant result (i.e., all for which $P \leq \alpha$). This subset is called “discoveries”. The FDR estimates the probability that these discoveries are false positives (i.e., Type I error). This improves statistical power as the entire sequence of P-values (and not only individual ones as in the Bonferroni correction procedure) are considered in the adjustment.

28

False Discovery Rates is widely used!

Methods in Ecology and Evolution

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

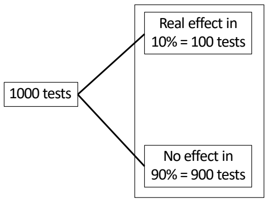
9440–9445 | PNAS | **August 5, 2003** | vol. 100 | no. 16

29

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.

Truth (unknown)



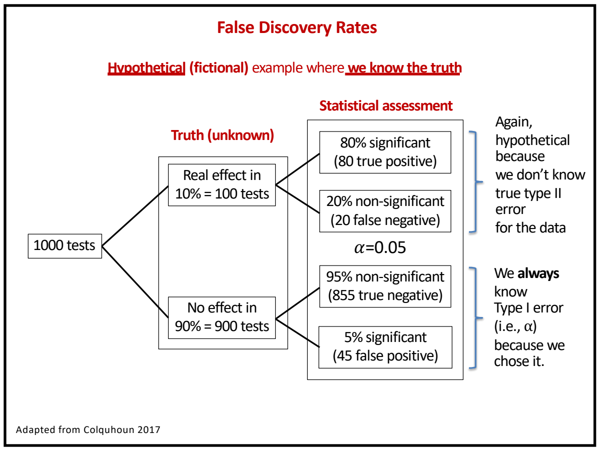
1000 tests

Real effect in 10% = 100 tests

No effect in 90% = 900 tests

Adapted from Colquhoun 2017

30



False Discovery Rates

Hypothetical (fictional) example where we know the truth

Truth (unknown)

Statistical assessment

1000 tests

Real effect in 10% = 100 tests

No effect in 90% = 900 tests

80% significant (80 true positive)

20% non-significant (20 false negative)

α=0.05

95% non-significant (855 true negative)

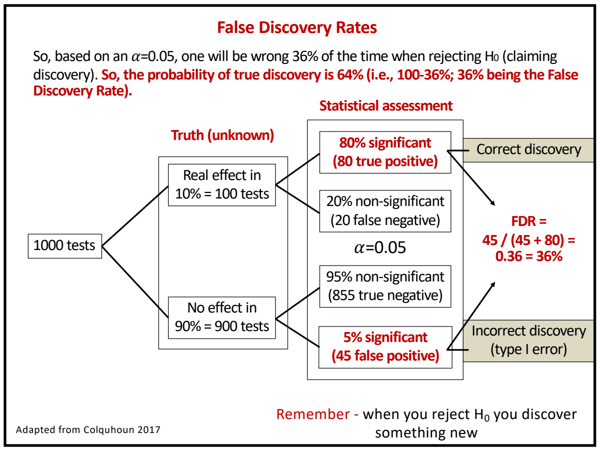
5% significant (45 false positive)

Again, hypothetical because we don't know true type II error for the data

We **always** know Type I error (i.e., α) because we chose it.

Adapted from Colquhoun 2017

31



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).**

Truth (unknown)

Statistical assessment

1000 tests

Real effect in 10% = 100 tests

No effect in 90% = 900 tests

80% significant (80 true positive)

20% non-significant (20 false negative)

α=0.05

95% non-significant (855 true negative)

5% significant (45 false positive)

Correct discovery

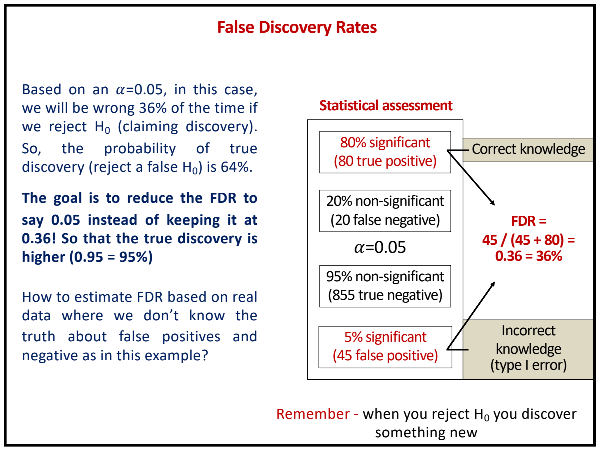
FDR = 45 / (45 + 80) = 0.36 = 36%

Incorrect discovery (type I error)

Remember - when you reject H_0 you discover something new

Adapted from Colquhoun 2017

32



False Discovery Rates

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?

Statistical assessment

80% significant (80 true positive)

20% non-significant (20 false negative)

α=0.05

95% non-significant (855 true negative)

5% significant (45 false positive)

Correct knowledge

FDR = 45 / (45 + 80) = 0.36 = 36%

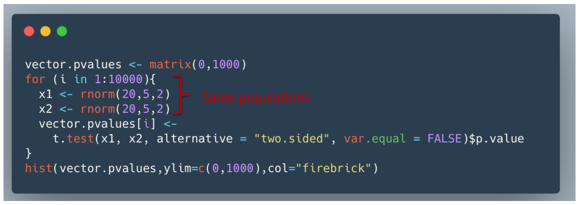
Incorrect knowledge (type I error)

Remember - when you reject H_0 you discover something new

33

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")
```

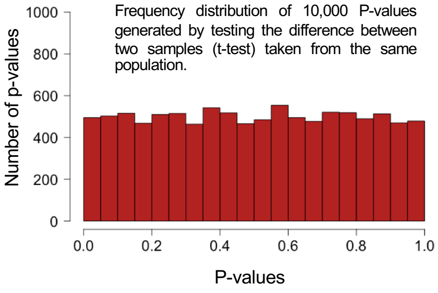
Same populations

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

34

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).

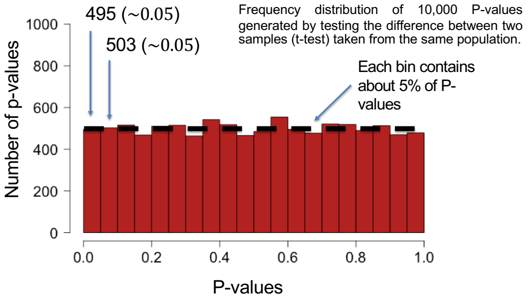


Frequency distribution of 10,000 P-values generated by testing the difference between two samples (t-test) taken from the same population.

35

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

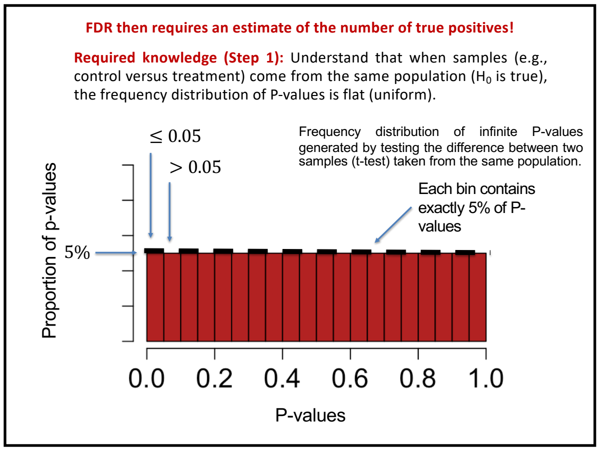
Required knowledge (Step 1): Understand that when samples (e.g., control versus treatment) come from the same population (H_0 is true), the frequency distribution of P-values is flat (uniform).



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

Each bin contains about 5% of P-values

36



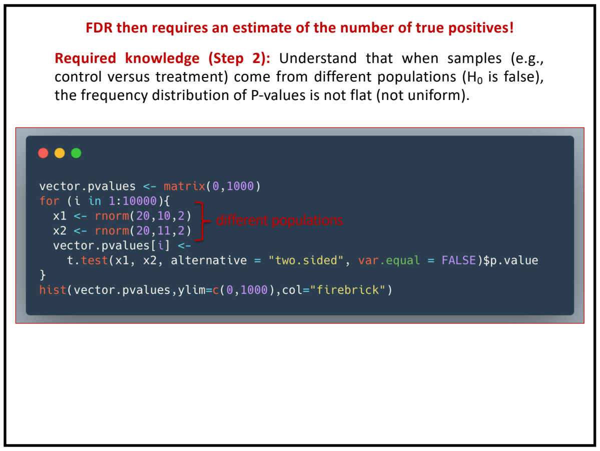
FDR then requires an estimate of the number of true positives!

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).

Frequency distribution of infinite P-values generated by testing the difference between two samples (t-test) taken from the same population.

Each bin contains exactly 5% of P-values

37

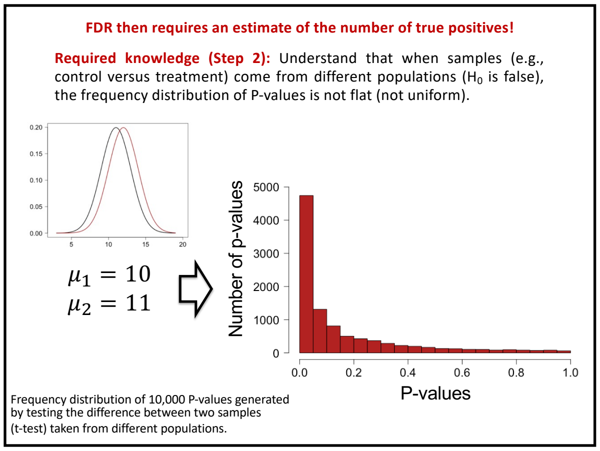


FDR then requires an estimate of the number of true positives!

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")
```

38



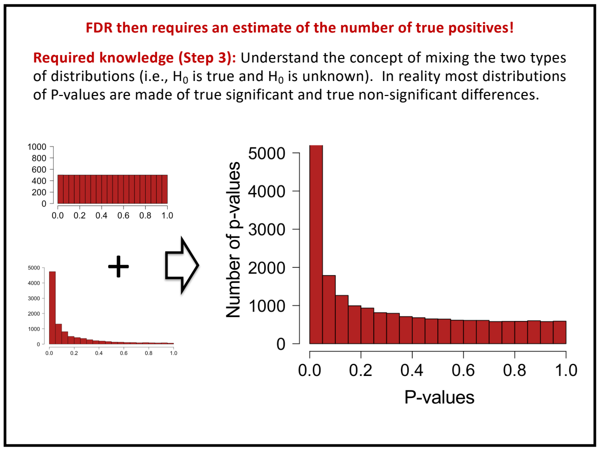
FDR then requires an estimate of the number of true positives!

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).

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

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

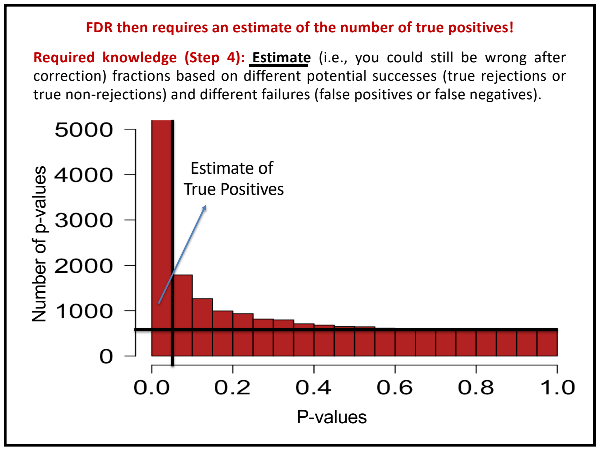
39



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 and H_0 is unknown). In reality most distributions of P-values are made of true significant and true non-significant differences.

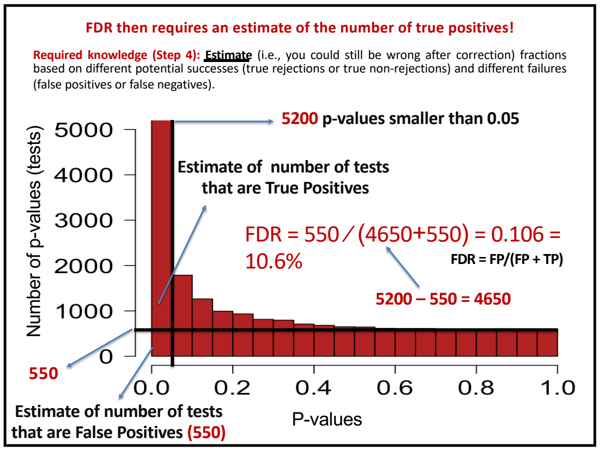
40



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).

41



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).

5200 p-values smaller than 0.05

Estimate of number of tests that are True Positives

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

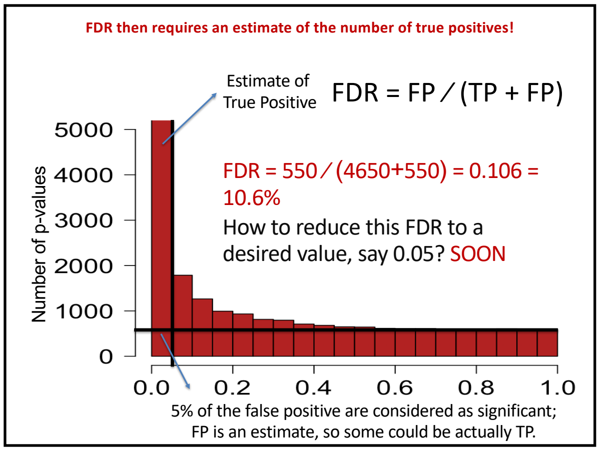
FDR = FP/(FP + TP)

5200 – 550 = 4650

550

Estimate of number of tests that are False Positives (550)

42



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

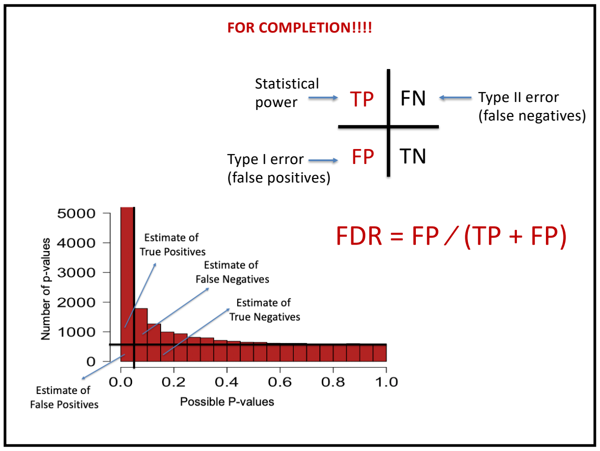
FDR = FP ⁄ (TP + FP)

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

How to reduce this FDR to a desired value, say 0.05? SOON

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

43



FOR COMPLETION!!!!

Statistical power → TP

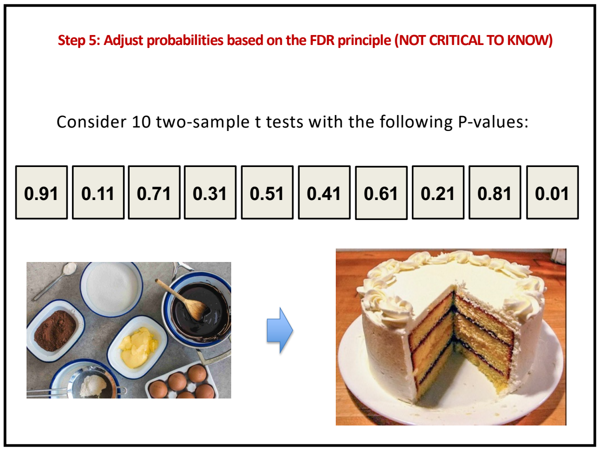
Type II error (false negatives) → FN

Type I error (false positives) → FP

TP	FN
FP	TN

FDR = FP ⁄ (TP + FP)

44



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

Consider 10 two-sample t tests with the following P-values:

0.91	0.11	0.71	0.31	0.51	0.41	0.61	0.21	0.81	0.01

45

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

Consider 10 two-sample t tests with the following P-values:

0.91	0.11	0.71	0.31	0.51	0.41	0.61	0.21	0.81	0.01

0.01	0.11	0.21	0.31	0.41	0.51	0.61	0.71	0.81	0.91

Order P-values

46

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

Consider 10 two-sample t tests with the following P-values:

0.01	0.11	0.21	0.31	0.41	0.51	0.61	0.71	0.81	0.91

Let’s see what happens if this small p-value (significant) when corrected by FDR.

47

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

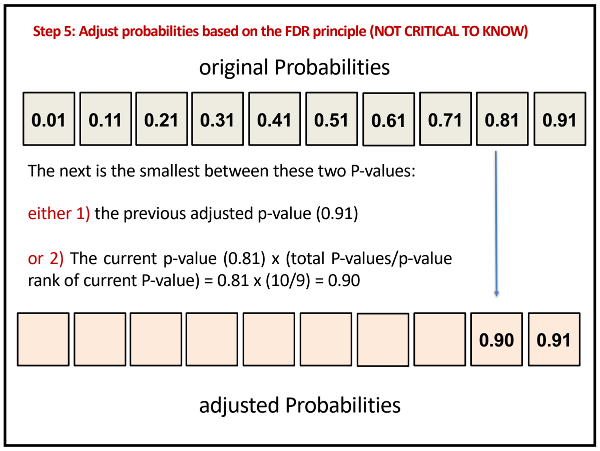
0.01	0.11	0.21	0.31	0.41	0.51	0.61	0.71	0.81	0.91

The largest probability is always the same

									0.91

Adjusted Probabilities

48



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

original Probabilities

0.01	0.11	0.21	0.31	0.41	0.51	0.61	0.71	0.81	0.91

The next is the smallest between these two P-values:

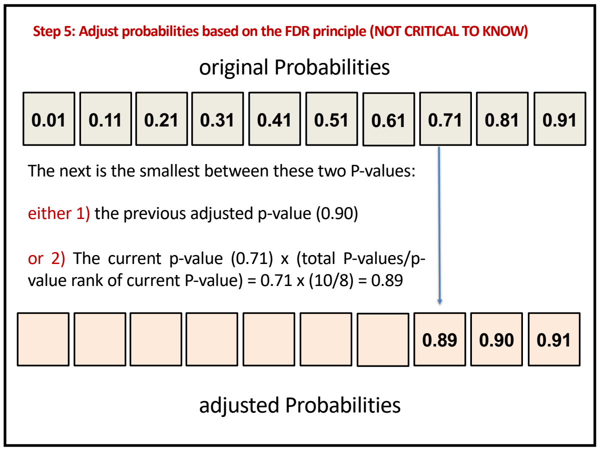
either 1) the previous adjusted p-value (0.91)

or 2) The current p-value (0.81) x (total P-values/p-value rank of current P-value) = 0.81 x (10/9) = 0.90

								0.90	0.91

adjusted Probabilities

49



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

original Probabilities

0.01	0.11	0.21	0.31	0.41	0.51	0.61	0.71	0.81	0.91

The next is the smallest between these two P-values:

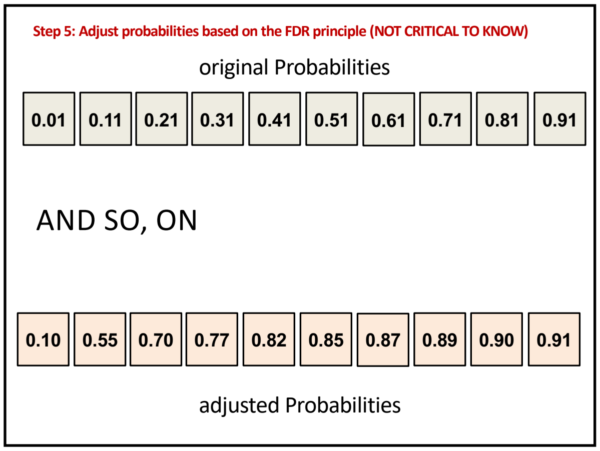
either 1) the previous adjusted p-value (0.90)

or 2) The current p-value (0.71) x (total P-values/p-value rank of current P-value) = 0.71 x (10/8) = 0.89

							0.89	0.90	0.91

adjusted Probabilities

50



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

original Probabilities

0.01	0.11	0.21	0.31	0.41	0.51	0.61	0.71	0.81	0.91

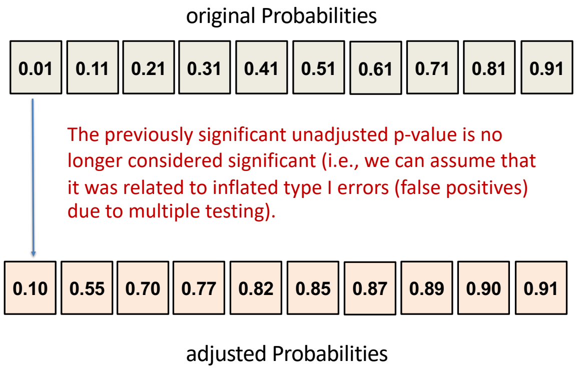
AND SO, ON

0.10	0.55	0.70	0.77	0.82	0.85	0.87	0.89	0.90	0.91

adjusted Probabilities

51

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



original Probabilities

0.01	0.11	0.21	0.31	0.41	0.51	0.61	0.71	0.81	0.91

The previously significant unadjusted p-value is no longer considered significant (i.e., we can assume that it was related to inflated type I errors (false positives) due to multiple testing).

0.10	0.55	0.70	0.77	0.82	0.85	0.87	0.89	0.90	0.91

adjusted Probabilities

52

Should we care about not committing any Type I error?

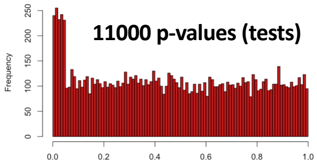
If we want to be protected against any FWER (family-wise error rate), then use Bonferroni like adjustments.

In many cases, we can let go on strict control over FWER, allow some false-positives to gain a lot of statistical power (then use FDR).

53

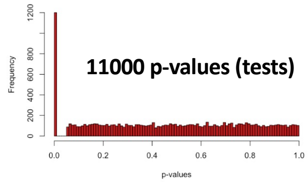
Bonferroni versus FDR (quick contrast)

Number of significant tests after adjustment



11000 p-values (tests)

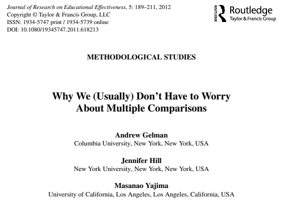
Bonferroni = 0
FDR = 0



11000 p-values (tests)

Bonferroni = 2
FDR = 1200

54



Journal of Research on Educational Effectiveness, 5: 189–211, 2012
Copyright © Taylor & Francis Group, LLC
ISSN: 1934-5747 print / 1934-5739 online
DOI: 10.1080/19345747.2011.618213

Routledge
Taylor & Francis Group

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

Some Bayesian dissent

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!

55

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!

56

Let's reflect on statistical errors and decisions:

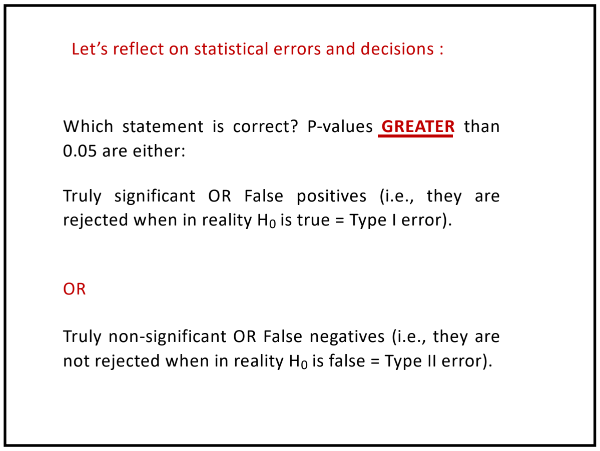
Which statement is correct? P-values **SMALLER** than 0.05 are either:

Truly significant OR False positives (i.e., they are rejected when in reality H_0 is true = Type I error).

OR

Truly non-significant OR False negatives (i.e., they are not rejected when in reality H_0 is false = Type II error).

57



Let’s reflect on statistical errors and decisions :

Which statement is correct? P-values **GREATER** than 0.05 are either:

Truly significant OR False positives (i.e., they are rejected when in reality H_0 is true = Type I error).

OR

Truly non-significant OR False negatives (i.e., they are not rejected when in reality H_0 is false = Type II error).

58