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).

```r
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  <2e-16 ***
Temperature       2  12.371   6.186   54.992  1.14e-11 ***
Calcium:Temperature 4  34.801   8.700   77.349  2.2e-16 ***
Residuals         36  4.049   0.112
```
Why should we not trust the results from multiple statistical tests?

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

Go to Lecture 6 and read this guide (reconvene at noon)

Why should we not trust the results from multiple statistical tests?
Guided discussion - t-distribution assuming $H_0$ as true

- Standardized values

- Distribution

- Acceptance area

- Rejection area

- $\alpha/2$

- $-2.063899$ and $2.063899$

- Probability density

- $t$ standardized values
Regarding Calcium, there are 3 possible pairwise tests contrast Growth across levels (3 \times 2/2 = 3).

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

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

<table>
<thead>
<tr>
<th>Condition</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>intermediate:high-high:high</td>
<td>2.15154803</td>
</tr>
<tr>
<td>low:high-high:high</td>
<td>-1.72154916</td>
</tr>
<tr>
<td>high:intermediate-high:high</td>
<td>-2.76050275</td>
</tr>
<tr>
<td>intermediate:intermediate-high:high</td>
<td>-3.86300578</td>
</tr>
<tr>
<td>low:intermediate-high:high</td>
<td>-2.70381093</td>
</tr>
<tr>
<td>high:low-high:high</td>
<td>-4.17303298</td>
</tr>
<tr>
<td>intermediate:low-high:high</td>
<td>-2.80337496</td>
</tr>
<tr>
<td>low:low-high</td>
<td>-3.89620697</td>
</tr>
<tr>
<td>low:high-intermediate:high</td>
<td>-3.87309719</td>
</tr>
<tr>
<td>high:intermediate-intermediate:high</td>
<td>-4.91205078</td>
</tr>
<tr>
<td>intermediate:intermediate-intermediate:high</td>
<td>-6.01455381</td>
</tr>
<tr>
<td>low:intermediate-intermediate:high</td>
<td>-4.85535896</td>
</tr>
<tr>
<td>high:low-intermediate:high</td>
<td>-6.32458101</td>
</tr>
<tr>
<td>intermediate:low-intermediate:high</td>
<td>-4.95492299</td>
</tr>
<tr>
<td>low:low-intermediate:high</td>
<td>-6.04775500</td>
</tr>
<tr>
<td>high:intermediate-low:high</td>
<td>-1.03895359</td>
</tr>
<tr>
<td>intermediate:intermediate-low:high</td>
<td>-2.14145662</td>
</tr>
<tr>
<td>low:intermediate-low:high</td>
<td>-0.98226177</td>
</tr>
<tr>
<td>high:low-low:high</td>
<td>-2.45148382</td>
</tr>
<tr>
<td>intermediate:low-low:high</td>
<td>-1.08182580</td>
</tr>
<tr>
<td>low:low-low</td>
<td>-2.17465781</td>
</tr>
<tr>
<td>intermediate:intermediate-high:intermediate:high</td>
<td>-1.10250303</td>
</tr>
<tr>
<td>low:intermediate-high:intermediate</td>
<td>0.05669182</td>
</tr>
<tr>
<td>high:low-high:intermediate</td>
<td>-1.41253023</td>
</tr>
<tr>
<td>intermediate:low-high:intermediate</td>
<td>-0.04287221</td>
</tr>
<tr>
<td>low:low-high:intermediate</td>
<td>-1.13570422</td>
</tr>
<tr>
<td>low:intermediate-intermediate:intermediate:intermediate:high</td>
<td>1.15919485</td>
</tr>
<tr>
<td>high:low-intermediate:intermediate</td>
<td>-0.31002720</td>
</tr>
<tr>
<td>low:low-intermediate:intermediate</td>
<td>-0.03320119</td>
</tr>
<tr>
<td>high:low-low:intermediate</td>
<td>-1.46922205</td>
</tr>
<tr>
<td>intermediate:low-low:intermediate</td>
<td>-0.09956403</td>
</tr>
<tr>
<td>low:low-low</td>
<td>-1.19239604</td>
</tr>
<tr>
<td>intermediate:low-high:low</td>
<td>1.36965802</td>
</tr>
<tr>
<td>low:high-low</td>
<td>0.27682601</td>
</tr>
<tr>
<td>low:high-low:low</td>
<td>-1.09283201</td>
</tr>
</tbody>
</table>

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.
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.
What happens when we conduct too many statistical tests?

A past classroom demonstration using a survey
Past classroom surveys:
Would you expect odd- and even day born individuals to differ in their preferences?

<table>
<thead>
<tr>
<th></th>
<th>dislike</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Love it</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>1) Do you like soccer?</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2) Do you like playing video games?</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3) Do you like eating out?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4) Do you enjoy writing?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5) Do you like cats?</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6) Do you like to watch movies?</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7) Do you like to read novels?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

......

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

Response: Do you like to listen to the radio.

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birthday</td>
<td>1</td>
<td>13.220</td>
<td>13.2196</td>
<td>12.5081</td>
<td>0.001226 **</td>
</tr>
<tr>
<td>Address</td>
<td>1</td>
<td>7.031</td>
<td>7.0309</td>
<td>6.6525</td>
<td>0.014546 *</td>
</tr>
<tr>
<td>Birthday:Address</td>
<td>1</td>
<td>10.440</td>
<td>10.4397</td>
<td>9.8778</td>
<td>0.003524 **</td>
</tr>
<tr>
<td>Residuals</td>
<td>33</td>
<td>34.877</td>
<td>1.0569</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

class survey: 24 questions
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 ($\alpha$), but conducting 100 tests, there will be a chance of 5 being significant for an $\alpha = 0.05$
Probability of committing 1 type I error (false positive) is the same for 1 or multiple tests ($\alpha$), but conducting 100 tests, there will be a chance of 5 being significant for an $\alpha = 0.05$
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.
ANOVA = 3 tests
pairwise t-tests = 42 tests

3 pairwise 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

36 pairwise tests

ANOVA Table:

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calcium</td>
<td>2</td>
<td>125.190</td>
<td>62.595</td>
<td>556.500</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>Temperature</td>
<td>2</td>
<td>12.371</td>
<td>6.186</td>
<td>54.992</td>
<td>1.137e-11 ***</td>
</tr>
<tr>
<td>Calcium:Temperature</td>
<td>4</td>
<td>34.801</td>
<td>8.700</td>
<td>77.349</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>36</td>
<td>4.049</td>
<td>0.112</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ANOVA for Growth vs Temperature and Calcium interaction:

Graph showing growth vs temperature for different calcium levels.
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:

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

80.6% chance of finding at least 1 significant test when $H_0$ is true!
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. 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), 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.

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

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

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 but modern use credited to Olive Dunn

\[ \alpha_{\text{Bonferroni}} = \frac{\alpha}{m} = \frac{0.05}{32} = 0.0015625 \]

Instead of using the original pre-established (desired) \( \alpha \), use \( \alpha \) 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!

\[
\alpha_{\text{Bonferroni}} = \frac{\alpha}{m}=\frac{0.05}{32}=0.0015625 \\
\text{Total number of tests}
\]

\[
1 - (1-\alpha_{\text{Bonferroni}})^{32} = 1 - (1 - 0.0015625)^{32}=0.04880777 \sim 0.05
\]

\[
P_{\text{Bonferroni}} = m \times P \quad \text{Original P value}
\]

\[
P_{\text{Bonferroni}} \text{ adjusted} \quad \text{Adjusted P value (adjusted P value that can be compared against any alpha)}
\]

Instead of using the original pre-established (desired) \(\alpha\), use \(\alpha\) adjusted instead to guarantee a family-wise (type I) error rate (FWER).
This example - not so many pairwise tests, but still an issue

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Sum of squares</th>
<th>df</th>
<th>Mean square</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>202.5</td>
<td>1</td>
<td>202.5</td>
<td>81</td>
<td>0.0000185</td>
</tr>
<tr>
<td>Within</td>
<td>20</td>
<td>8</td>
<td>2.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>222.5</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ \text{Ho: } \mu_{\text{control}} = \mu_{\text{knee}} = \mu_{\text{eyes}} \]

\[ \text{Ha: at least one } \mu \text{ is different from another } u \text{ or other } u_s; \]
\[ \text{but which pairs?} \]

\[ X_{\text{control}} - X_{\text{knee}} \]
\[ X_{\text{control}} - X_{\text{eyes}} \]
\[ X_{\text{knee}} - X_{\text{eyes}} \]

\( 3 \text{ t-tests necessary} \)

Back to the problem about “The knees who say night”
Bonferroni correction

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

OR

Adjust the P-value as below and contrast with the original $\alpha$ (0.05)

$P_{\text{Bonferroni}} = mP$

Conclude based on these adjusted P-values

<table>
<thead>
<tr>
<th>comparison</th>
<th>uncorrected P (t test)</th>
<th>Bonferroni P (t test)</th>
<th>3 x uncorrected P (t test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>control vs eyes</td>
<td>0.0029</td>
<td>0.0088</td>
<td>3 x 0.0029</td>
</tr>
<tr>
<td>control vs knee</td>
<td>0.9418</td>
<td>1.0000</td>
<td>3 x 0.9418 = 2.8253</td>
</tr>
<tr>
<td>knee vs eyes</td>
<td>0.0044</td>
<td>0.0132</td>
<td>3 x 0.0044</td>
</tr>
</tbody>
</table>

Adjusted $\alpha = 0.0166667$

P-values greater than 1 are set to 1
Bonferroni correction (common table presentation)

<table>
<thead>
<tr>
<th>comparison</th>
<th>uncorrected P (t test)</th>
<th>Bonferroni P (t test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>control vs eyes</td>
<td>0.0029</td>
<td>0.0088</td>
</tr>
<tr>
<td>control vs knee</td>
<td>0.9418</td>
<td>1.0000</td>
</tr>
<tr>
<td>knee vs eyes</td>
<td>0.0044</td>
<td>0.0132</td>
</tr>
</tbody>
</table>

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.
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).
**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.
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
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)**

- Real effect in 10% = 100 tests
- No effect in 90% = 900 tests

Adapted from Colquhoun 2017
False Discovery Rates

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

1000 tests

- **Real effect in 10% = 100 tests**
  - 80% significant (80 true positive)
  - 20% non-significant (20 false negative)

- **No effect in 90% = 900 tests**
  - 95% non-significant (855 true negative)
  - 5% significant (45 false positive)

**Statistical assessment**

\[ \alpha = 0.05 \]

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

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

Adapted from Colquhoun 2017
False Discovery Rates

So, based on an \( \alpha=0.05 \), 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
Based on an \( \alpha = 0.05 \), 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%)

To do that, we need to estimate FDR. How to estimate this value based on real data where we don’t know the truth about false positives and negative as in this example?

\[
\text{FDR} = \frac{45}{45 + 80} = \frac{45}{125} = 0.36 = 36%
\]

\( \alpha = 0.05 \)

- 80% significant (80 true positive)
- 20% non-significant (20 false negative)
- 95% non-significant (855 true negative)
- 5% significant (45 false positive)

Correct knowledge
Incorrect knowledge (type I error)

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 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).

```r
vector.pvalues <- matrix(0,1000)
for (i in 1:100000){
  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")
```
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.
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.

- 495 ($\sim 0.05$)
- 503 ($\sim 0.05$)
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.
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).

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

```r
define 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")
```
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).

\[
\begin{align*}
\mu_1 &= 10 \\
\mu_2 &= 11
\end{align*}
\]

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

![Histogram with annotations](image)

- **5200 p-values smaller than 0.05**
- **550** estimated as the number of false positives.
- **5200 - 550 = 4650** estimated as the number of true positives.

FDR = \( \frac{550}{(4650+550)} = 0.106 = 10.6\% \)

FDR = FP/(FP + TP)
FDR then requires an estimate of the number of true positives!

FDR = FP / (TP + FP)

Estimate of True Positive

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.
Statistical power

Type I error (false positives)

Type II error (false negatives)

\[ \text{FDR} = \frac{\text{FP}}{\text{TP} + \text{FP}} \]
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
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
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.
Step 5: Adjust probabilities based on the FDR principle (NOT CRITICAL TO KNOW)

The largest probability is always the same

Adjusted Probabilities
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

adjusted Probabilities

|   |   |   |   |   |   |   | 0.90 | 0.91 |   |
Step 5: Adjust probabilities based on the FDR principle (NOT CRITICAL TO KNOW)

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
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

adjusted Probabilities

0.10  0.55  0.70  0.77  0.82  0.85  0.87  0.89  0.90  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).
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).
Bonferroni versus FDR (quick contrast)

11000 p-values (tests)

Number of significant tests after adjustment

Bonferroni = 0
FDR = 0

Bonferroni = 2
FDR = 1200
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):

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