Analysis of Quantitative data
One-Way + Two-Way ANOVA
Anne Segonds-Pichon
v2020-12
Comparison between more than 2 groups
One factor = One predictor
One-Way ANOVA
Signal-to-noise ratio

\[
\frac{\text{Difference}}{\text{Variability}} = \frac{\text{Signal}}{\text{Noise}}
\]

\[
\frac{\text{Signal}}{\text{Noise}} = \text{statistical significance}
\]

\[
\frac{\text{Signal}}{\text{Noise}} = \text{no statistical significance}
\]
Analysis of variance: how does it work?

\[
\frac{\text{Signal}}{\text{Noise}} = \frac{\text{Difference between the means}}{\text{Variability in the groups}} = F \text{ ratio}
\]
Step 1: Omnibus test
• It tells us if there is a difference between the means but not which means are significantly different from which other ones.

Step 2: Post-hoc tests
• They tell us if there are differences between the means pairwise.
Analysis of variance: how does it work?

<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-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>18.1</td>
<td>4</td>
<td>4.5</td>
<td>6.32</td>
<td>0.0002</td>
</tr>
<tr>
<td>Within Groups</td>
<td>51.8</td>
<td>73</td>
<td>0.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>69.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Analysis of variance: how does it work?

<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-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td><strong>69.9</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

78 differences: $\sum_{1}^{78} (\text{value}_n - \text{grand mean})^2$
Analysis of variance: how does it work?

### Source of variation

<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-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td><strong>18.1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>69.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Analysis of variance: how does it work?

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Squares</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>18.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Groups</td>
<td><strong>51.8</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>69.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

78 differences: \[ \sum_{1}^{78} (value_n - \text{group mean})^2 \]

= Sum of squared errors

Within the Groups
Analysis of variance: how does it work?

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Squares</th>
<th>F ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>18.1</td>
<td>k-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Groups</td>
<td>51.8</td>
<td>n-k</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>69.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

df: degree of freedom with df = n-1

n = number of values, k = number of groups

Between groups: df = 4 (k-1)
Within groups: df = 73 (n-k = n_1-1 + ... + n_5-1)
Analysis of variance: how does it work?

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Squares</th>
<th>F ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>18.1</td>
<td>4</td>
<td>4.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Groups</td>
<td>51.8</td>
<td>73</td>
<td>0.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>69.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

df: degree of freedom with df = n-1

\[
\frac{18.2}{4} = 4.5 \quad \frac{51.8}{73} = 0.71
\]

Mean squares = **Sum of Squares / n-1 = Variance!**
Analysis of variance: how does it work?

### ANOVA Table

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Squares</th>
<th>F ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>18.1</td>
<td>4</td>
<td>4.5</td>
<td>6.34</td>
<td>0.0002</td>
</tr>
<tr>
<td>Within Groups</td>
<td>51.8</td>
<td>73</td>
<td>0.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>69.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mean squares = \textbf{Sum of Squares} / n-1 = Variance

\[ F \text{ ratio} = \frac{\text{Variance between the groups}}{\text{Variance within the groups (individual variability)}} = \frac{4.5}{0.71} = 6.34 \]
Comparison of more than 2 means

• Running multiple tests on the same data increases the familywise error rate.

• What is the familywise error rate?
  – The error rate across tests conducted on the same experimental data.

• One of the basic rules (‘laws’) of probability:
  – The Multiplicative Rule: The probability of the joint occurrence of 2 or more independent events is the product of the individual probabilities.

\[ P(A,B) = P(A) \times P(B) \]

For example:

\[ P(2 \text{ Heads}) = P(\text{head}) \times P(\text{head}) = 0.5 \times 0.5 = 0.25 \]
Familywise error rate

- **Example**: All pairwise comparisons between 3 groups A, B and C:
  - A-B, A-C and B-C

- Probability of making the Type I Error: **5%**
  - The probability of not making the Type I Error is **95% (=1 – 0.05)**

- Multiplicative Rule:
  - Overall probability of no Type I errors is: \(0.95 \times 0.95 \times 0.95 = 0.857\)

- So the probability of making at least one Type I Error is \(1-0.857 = 0.143 \text{ or } 14.3\%\)
  - The probability has increased from 5% to 14.3%

- Comparisons between 5 groups instead of 3, the familywise error rate is **40% (=1-(0.95^n))**
Familywise error rate

• **Solution** to the increase of familywise error rate: correction for multiple comparisons
  – **Post-hoc tests**

• Many different ways to correct for multiple comparisons:
  – Different statisticians have designed corrections addressing different issues
    • e.g. unbalanced design, heterogeneity of variance, liberal vs conservative

• However, they all have **one thing in common**:
  – the more tests, the higher the familywise error rate: the more stringent the correction

• Tukey, Bonferroni, Sidak, Benjamini-Hochberg ...
  – Two ways to address the multiple testing problem
    • **Familywise Error Rate** (FWER) vs. **False Discovery Rate** (FDR)
Multiple testing problem

• **FWER: Bonferroni**: \( \alpha_{\text{adjust}} = 0.05/n \) comparisons e.g. 3 comparisons: 0.05/3 = 0.016
  – Problem: very conservative leading to loss of power (lots of false negative)
  – 10 comparisons: threshold for significance: 0.05/10: 0.005
  – Pairwise comparisons across 20,000 genes 😎

• **FDR: Benjamini-Hochberg**: the procedure controls the expected proportion of “discoveries” (significant tests) that are false (false positive).
  – Less stringent control of Type I Error than FWER procedures which control the probability of at least one Type I Error
  – More power at the cost of increased numbers of Type I Errors.

• Difference between FWER and FDR:
  – a p-value of 0.05 implies that 5% of all tests will result in false positives.
  – a FDR adjusted p-value (or q-value) of 0.05 implies that 5% of significant tests will result in false positives.
One-Way Analysis of variance

**Step 1: Omnibus test**
- It tells us if there is (or not) a difference between the means but not which means are significantly different from which other ones.

**Step 2: Post-hoc tests**
- They tell us if there are (or not) differences between the means pairwise.
- A correction for multiple comparisons will be applied on the p-values.
- These post hoc tests should only be used when the ANOVA finds a significant effect.
Example: protein.expression.csv

- **Question**: is there a difference in protein expression between the 5 cell lines?

- 1 Plot the data
- 2 Check the assumptions for parametric test
Exercise 6: One-way ANOVA: Data Exploration
proteína.expression.csv

- **Question**: Difference in protein expression between 5 cell types?

  - Load *protein.expression.csv*

  - Plot the data using at least 2 types of graph
    - `geom_boxplot()`, `geom_jitter()`, `geom_violin()`

  - Draw a QQplot
    - `ggplot(aes(sample =)) + stat_qq() + stat_qq_line()`

  - Check the first assumption (Normality) with a formal test
    - `shapiro_test()`
Exercise 6: One-way ANOVA: Data Exploration - Answers

```r
protein %>%
  ggplot(aes(x=line, y=expression, colour=line)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(height=0, width=0.1)

protein %>%
  ggplot(aes(x=line, y=expression, colour=line)) +
  geom_violin(trim=FALSE) +
  geom_boxplot(width=0.1)
```
Exercise 6: One-way ANOVA – Answers

```r
protein %>%
ggplot(aes(sample = expression)) +
  stat_qq(size=3) +
  stat_qq_line()
```
Exercise 6: One-way ANOVA – Answers. What do we do now?

protein %>%
  group_by(line) %>%
  shapiro_test(expression) %>%
  ungroup()
One-way ANOVA: change of scale

protein %>%
  ggplot(aes(line, expression, colour=line)) +
  geom_jitter(height=0, width=0.2, size=3, show.legend=FALSE) +
  stat_summary(geom = "errorbar", fun=mean, fun.min=mean, fun.max=mean, colour="black", size=1) +
  scale_y_log10()

protein %>%
  mutate(log10.expression=log10(expression)) -> protein
One-way ANOVA: change of scale

protein %>%
  ggplot(aes(x=line, y=log10.expression, colour=line)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(height=0, width=0.1)

protein %>%
  ggplot(aes(x=line, y=log10.expression, colour=line)) +
  geom_violin(trim=FALSE) +
  geom_boxplot(width=0.1)
One-way ANOVA – Outliers identification

protein %>%
  group_by(line) %>%
  identify_outliers(expression) %>%
  ungroup()

<table>
<thead>
<tr>
<th>line</th>
<th>expression</th>
<th>log10.expression</th>
<th>is.outlier</th>
<th>is.extreme</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3.14</td>
<td>0.4969296</td>
<td>TRUE</td>
<td>FALSE</td>
</tr>
<tr>
<td>C</td>
<td>2.78</td>
<td>0.4440448</td>
<td>TRUE</td>
<td>FALSE</td>
</tr>
<tr>
<td>D</td>
<td>9.32</td>
<td>0.9694159</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
</tbody>
</table>

3 rows
One-way ANOVA: change of scale

```r
protein %>%
ggplot(aes(sample=log10.expression)) +
  stat_qq(size=3) +
  stat_qq_line()
```

First assumption ✓

Before log-transformation
Assumptions of Parametric Data

protein %>%
  group_by(line) %>%
  shapiro_test(log10.expression) %>%
  ungroup()

<table>
<thead>
<tr>
<th>line</th>
<th>variable</th>
<th>statistic</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>log10.expression</td>
<td>0.8542464</td>
<td>0.04143953</td>
</tr>
<tr>
<td>B</td>
<td>log10.expression</td>
<td>0.9458450</td>
<td>0.57725211</td>
</tr>
<tr>
<td>C</td>
<td>log10.expression</td>
<td>0.9637060</td>
<td>0.71417958</td>
</tr>
<tr>
<td>D</td>
<td>log10.expression</td>
<td>0.9868425</td>
<td>0.99348831</td>
</tr>
<tr>
<td>E</td>
<td>log10.expression</td>
<td>0.9313426</td>
<td>0.20502703</td>
</tr>
</tbody>
</table>

First assumption ✔ ish

protein %>%
  levene_test(log10.expression ~ line)

<table>
<thead>
<tr>
<th>df1</th>
<th>df2</th>
<th>statistic</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>73</td>
<td>0.982112</td>
<td>0.4227373</td>
</tr>
</tbody>
</table>

Second assumption ✔
Analysis of variance

• Step 1: omnibus test
  ```r
data %>%
anova_test(y~x)
```

• Step 2: post-hoc tests
  - **Tukey correction**
    ```r
data %>%
tukey_hsd(y~x)
```
  - **Bonferroni correction**  # emmeans package#
    ```r
data %>%
emmeans_test(y~x, p.adjust.method="bonferroni")
```

**R way:**
```
aov(y~x, data= ) -> model then summary(model)
pairwise.t.test(y, x, p.adj = "bonf")
TukeyHSD(model)
```

Have a go!
Analysis of variance

```
protein %>%
anova_test(log10.expression ~ line)
```

ANOVA Table (type II tests)

<table>
<thead>
<tr>
<th>Effect</th>
<th>DFn</th>
<th>DFd</th>
<th>F</th>
<th>p</th>
<th>ges</th>
<th>p &lt; .05</th>
</tr>
</thead>
<tbody>
<tr>
<td>line</td>
<td>4</td>
<td>73</td>
<td>8.123</td>
<td>1.78e-05</td>
<td>0.308</td>
<td></td>
</tr>
</tbody>
</table>

generalised effect size (Eta squared $\eta^2$) = $R^2$ ish

```
protein %>%
tukey_hsd(log10.expression ~ line)
```

Tukey correction

<table>
<thead>
<tr>
<th>term</th>
<th>group1</th>
<th>group2</th>
<th>estimate</th>
<th>conf.low</th>
<th>conf.high</th>
<th>p.adj</th>
<th>p.adj.signif</th>
</tr>
</thead>
<tbody>
<tr>
<td>line</td>
<td>A</td>
<td>B</td>
<td>-0.25024832</td>
<td>-0.57882494</td>
<td>0.07838585</td>
<td>2.19e-01</td>
<td>ns</td>
</tr>
<tr>
<td>line</td>
<td>A</td>
<td>C</td>
<td>-0.07499724</td>
<td>-0.37499782</td>
<td>0.22500335</td>
<td>9.56e-01</td>
<td>ns</td>
</tr>
<tr>
<td>line</td>
<td>A</td>
<td>D</td>
<td>0.30549397</td>
<td>0.00549391</td>
<td>0.60549456</td>
<td>4.39e-02</td>
<td>*</td>
</tr>
<tr>
<td>line</td>
<td>A</td>
<td>E</td>
<td>0.13327517</td>
<td>-0.16672541</td>
<td>0.43327575</td>
<td>7.27e-01</td>
<td>ns</td>
</tr>
<tr>
<td>line</td>
<td>B</td>
<td>C</td>
<td>0.1752108</td>
<td>-0.12474949</td>
<td>0.4752167</td>
<td>4.81e-01</td>
<td>ns</td>
</tr>
<tr>
<td>line</td>
<td>B</td>
<td>D</td>
<td>0.55574230</td>
<td>0.25574171</td>
<td>0.85574288</td>
<td>1.83e-04</td>
<td>****</td>
</tr>
<tr>
<td>line</td>
<td>B</td>
<td>E</td>
<td>0.38352349</td>
<td>0.08352290</td>
<td>0.68352407</td>
<td>5.48e-03</td>
<td>**</td>
</tr>
<tr>
<td>line</td>
<td>C</td>
<td>D</td>
<td>0.38049121</td>
<td>0.11216253</td>
<td>0.6481989</td>
<td>1.54e-03</td>
<td>**</td>
</tr>
<tr>
<td>line</td>
<td>C</td>
<td>E</td>
<td>0.20827240</td>
<td>-0.06005627</td>
<td>0.47660108</td>
<td>2.02e-01</td>
<td>ns</td>
</tr>
<tr>
<td>line</td>
<td>D</td>
<td>E</td>
<td>-0.17221881</td>
<td>-0.44054748</td>
<td>0.09610987</td>
<td>3.84e-01</td>
<td>ns</td>
</tr>
</tbody>
</table>
Analysis of variance

\[
\text{protein} \gg \text{anova_test(log10.expression~line)}
\]

ANOVA Table (type II tests)

<table>
<thead>
<tr>
<th>Effect DFn DFd</th>
<th>( p &lt; .05 )</th>
<th>ges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line</td>
<td>4</td>
<td>8</td>
</tr>
</tbody>
</table>

Generalised effect size (Eta squared \( \eta^2 \)) = R^2 ish

\[
\text{protein} \gg \text{emmeans_test(log10.expression ~ line, p.adjust.method = "bonferroni")}
\]

**Bonferroni correction**

<table>
<thead>
<tr>
<th>( y )</th>
<th>group1</th>
<th>group2</th>
<th>df</th>
<th>statistic</th>
<th>( p )</th>
<th>p.adj</th>
<th>p.adj_signif</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \log_{10} \text{expression} )</td>
<td>A</td>
<td>B</td>
<td>73</td>
<td>2.1299578</td>
<td>3.654611e-02</td>
<td>3.654611e-01 ns</td>
<td></td>
</tr>
<tr>
<td>( \log_{10} \text{expression} )</td>
<td>A</td>
<td>C</td>
<td>73</td>
<td>0.6992552</td>
<td>4.866147e-01</td>
<td>1.000000e+00 ns</td>
<td></td>
</tr>
<tr>
<td>( \log_{10} \text{expression} )</td>
<td>A</td>
<td>D</td>
<td>73</td>
<td>-2.8483483</td>
<td>5.705474e-03</td>
<td>5.705474e-02 ns</td>
<td></td>
</tr>
<tr>
<td>( \log_{10} \text{expression} )</td>
<td>A</td>
<td>E</td>
<td>73</td>
<td>-1.2426238</td>
<td>2.179833e-01</td>
<td>1.000000e+00 ns</td>
<td></td>
</tr>
<tr>
<td>( \log_{10} \text{expression} )</td>
<td>B</td>
<td>C</td>
<td>73</td>
<td>-1.6339966</td>
<td>1.065653e-01</td>
<td>1.000000e+00 ns</td>
<td></td>
</tr>
<tr>
<td>( \log_{10} \text{expression} )</td>
<td>B</td>
<td>D</td>
<td>73</td>
<td>-5.1816001</td>
<td>1.882302e-06</td>
<td>1.882302e-05 ****</td>
<td></td>
</tr>
<tr>
<td>( \log_{10} \text{expression} )</td>
<td>B</td>
<td>E</td>
<td>73</td>
<td>-3.5758757</td>
<td>6.238766e-04</td>
<td>6.238766e-03 **</td>
<td></td>
</tr>
<tr>
<td>( \log_{10} \text{expression} )</td>
<td>C</td>
<td>D</td>
<td>73</td>
<td>-3.9663413</td>
<td>1.687079e-04</td>
<td>1.687079e-03 **</td>
<td></td>
</tr>
<tr>
<td>( \log_{10} \text{expression} )</td>
<td>C</td>
<td>E</td>
<td>73</td>
<td>-2.1710868</td>
<td>3.317601e-02</td>
<td>3.317601e-01 ns</td>
<td></td>
</tr>
<tr>
<td>( \log_{10} \text{expression} )</td>
<td>D</td>
<td>E</td>
<td>73</td>
<td>1.7952545</td>
<td>7.675206e-02</td>
<td>7.675206e-01 ns</td>
<td></td>
</tr>
</tbody>
</table>
Analysis of variance (R)
To plot confidence intervals

```r
aov(log10.expression~line, data=protein.stack.clean) -> anova.log.protein
summary(anova.log.protein)
```

```
  df  sum sq  mean sq     F value    Pr(>F)
line       4  2.691  0.6728     8.123   1.78e-05  ***
Residuals  73  6.046  0.0828
---
signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```r
TukeyHSD(anova.log.protein) -> tukey.plot(tukey, las=1)
```

### TukeyHSD (anova.log.protein, "line")

Tukey multiple comparisons of means
95\% family-wise confidence level

Fit: aov(formula = log10.expression ~ line, data = protein.stack.clean)

<table>
<thead>
<tr>
<th>Sline</th>
<th>diff</th>
<th>lwr</th>
<th>upr</th>
<th>p adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>B-A</td>
<td>-0.25024832</td>
<td>-0.578882494</td>
<td>0.07838585</td>
<td>0.2187264</td>
</tr>
<tr>
<td>C-A</td>
<td>-0.07499724</td>
<td>-0.374997820</td>
<td>0.22500335</td>
<td>0.9560187</td>
</tr>
<tr>
<td>D-A</td>
<td>0.30543937</td>
<td>0.005493391</td>
<td>0.60549456</td>
<td>0.0438762</td>
</tr>
<tr>
<td>E-A</td>
<td>0.13327317</td>
<td>-0.166725416</td>
<td>0.43327375</td>
<td>0.7265567</td>
</tr>
<tr>
<td>C-B</td>
<td>0.17525108</td>
<td>-0.124794949</td>
<td>0.47525167</td>
<td>0.4809387</td>
</tr>
<tr>
<td>D-B</td>
<td>0.55574330</td>
<td>0.255741712</td>
<td>0.85574288</td>
<td>0.0600183</td>
</tr>
<tr>
<td>E-B</td>
<td>0.38352349</td>
<td>0.083522904</td>
<td>0.68352407</td>
<td>0.0054767</td>
</tr>
<tr>
<td>C-C</td>
<td>0.38041211</td>
<td>0.121265323</td>
<td>0.64881988</td>
<td>0.0015431</td>
</tr>
<tr>
<td>D-C</td>
<td>0.20827240</td>
<td>-0.060056276</td>
<td>0.47660108</td>
<td>0.2023355</td>
</tr>
<tr>
<td>E-C</td>
<td>-0.17221881</td>
<td>-0.440547487</td>
<td>0.09610987</td>
<td>0.3641389</td>
</tr>
</tbody>
</table>
Analysis of variance (tidyverse)
To plot confidence intervals

protein %>%
tukey_hsd(log10.expression~line) %>%
mutate(comparison = paste(group1, sep=".", group2)) -> tukey.conf

tukey.conf %>%
ggplot(aes(x=comparison, y=estimate, ymin=conf.low, ymax=conf.high)) +
geom_errorbar(colour="black", size=1)+
geom_point(size=3, colour="darkred")+
geom_hline(yintercept=0, linetype="dashed", color = "red")+
coord_flip()
Analysis of variance

```r
protein %>%
  ggplot(aes(line, expression, colour=line)) +
  geom_jitter(height = 0, width=0.2, size=3, show.legend=FALSE) +
  stat_summary(geom="errorbar", fun=mean, fun.min=mean, fun.max = mean, colour="black", size=1) +
  scale_y_log10()
```
protein %>%
ggplot(aes(x=line, y=expression, fill=line)) +
  geom_bar(stat = "summary", fun="mean", colour="black") +
  stat_summary(geom="errorbar", colour="black", width=0.4)
Analysis of variance

```r
protein %>%
ggplot(aes(x=line, y=expression, fill=line)) +
geom_bar(stat="summary", fun="mean", colour="black") +
stat_summary(geom="errorbar", colour="black", width=0.4) +
geom_jitter(heigth=0, width=0.1, alpha=0.5)
```
Analysis of variance

protein %>%
ggplot(aes(x=line, y=log10.expression, fill=line)) +
  geom_bar(stat="summary", fun="mean", colour="black") +
  stat_summary(geom="errorbar", colour="black", width=0.4) +
  geom_jitter(height=0, width=0.1, alpha=0.5)
**Exercise 7: Repeated measures ANOVA**

*neutrophils.long.csv*

- A researcher is looking at the difference between 4 cell groups. He has run the experiment 5 times. Within each experiment, he has neutrophils from a WT (control), a KO, a KO+Treatment 1 and a KO+Treatment2.
- **Question**: Is there a difference between KO with/without treatment and WT?

- Load *neutrophils.long.csv*
- Plot the data so that you have an idea of the consistency of the results between the experiments.
- Check the first assumption
- Run the repeated measures ANOVA and post-hoc tests

```r
anova_test(dv =, wid =, within =) -> res.aov
get_anova_table(res.aov)
pairwise_t_test(p.adjust.method =)
```

- Choose a graphical presentation consistent with the experimental design
Exercise 7: Repeated measures ANOVA
neutrophils.long.csv

- Plot the data so that you have an idea of the consistency of the results between the experiments.

```r
neutrophils.long %>%
ggplot(aes(Group, Values, group=Experiment, colour=Experiment, fill=Experiment)) +
  geom_line(size=2) +
  geom_point(size=4, shape = 21, colour= "black", stroke=2) +
  scale_x_discrete(limits = c("WT", "KO", "KO+T1", "KO+T2"))
```
**Exercise 7**: Repeated measures ANOVA
neutrophils.long.csv

- Check the first assumption

```r
neutrophils.long %>%
  group_by(Group) %>%
  shapiro_test(Values) %>%
  ungroup()
```

```r
ggplot(aes(Group, Values)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(height = 0, width = 0.2)
```
Exercise 7: Repeated measures ANOVA
neutrophils.long.csv

- Run the repeated measures ANOVA and post-hoc tests

```r
neutrophils.long %>%
  anova_test(dv = Values, wid = Experiment, within = Group) -> res.aov
get_anova_table(res.aov)
```

ANOVA Table (type III tests)

<table>
<thead>
<tr>
<th>Effect</th>
<th>DFn</th>
<th>DFd</th>
<th>F</th>
<th>p</th>
<th>p.&lt;.05</th>
<th>ges</th>
<th>* 0.656</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>3</td>
<td>12</td>
<td>28.575</td>
<td>0.51e-06</td>
<td>4</td>
<td>0.00979</td>
<td>0.003 **</td>
</tr>
</tbody>
</table>

```r
neutrophils.long %>%
  pairwise_t_test(Values~Group, paired=TRUE, ref.group = "WT", p.adjust.method = "bonferroni")
```

<table>
<thead>
<tr>
<th>.y.</th>
<th>group1</th>
<th>group2</th>
<th>n1</th>
<th>n2</th>
<th>statistic</th>
<th>df</th>
<th>p</th>
<th>p.adj</th>
<th>p.adj.signif</th>
</tr>
</thead>
<tbody>
<tr>
<td>Values</td>
<td>WT</td>
<td>KO</td>
<td>5</td>
<td>5</td>
<td>-8.657886</td>
<td>4</td>
<td>0.000979</td>
<td>0.003 **</td>
<td></td>
</tr>
<tr>
<td>Values</td>
<td>WT</td>
<td>KO+T1</td>
<td>5</td>
<td>5</td>
<td>1.310271</td>
<td>4</td>
<td>0.260000</td>
<td>ns</td>
<td></td>
</tr>
<tr>
<td>Values</td>
<td>WT</td>
<td>KO+T2</td>
<td>5</td>
<td>5</td>
<td>-6.481813</td>
<td>4</td>
<td>0.003000</td>
<td>0.009 **</td>
<td></td>
</tr>
</tbody>
</table>
Exercise 7: Repeated measures ANOVA
neutrophils.long.csv

• Run the repeated measures ANOVA and post-hoc tests

```r
eutrophils.long %>%
  pairwise_t_test(Values~Group, paired=TRUE, ref.group = "WT",
  p.adjust.method = "bonferroni")
```

<table>
<thead>
<tr>
<th>.y.</th>
<th>group1</th>
<th>group2</th>
<th>n1</th>
<th>n2</th>
<th>statistic</th>
<th>df</th>
<th>p</th>
<th>p.adj</th>
<th>p.adj.signif</th>
</tr>
</thead>
<tbody>
<tr>
<td>Values</td>
<td>WT</td>
<td>KO</td>
<td>5</td>
<td>5</td>
<td>-8.657886</td>
<td>4</td>
<td>0.000979</td>
<td>0.003</td>
<td>**</td>
</tr>
<tr>
<td>Values</td>
<td>WT</td>
<td>KO+T1</td>
<td>5</td>
<td>5</td>
<td>1.310271</td>
<td>4</td>
<td>0.260000</td>
<td>0.780</td>
<td>ns</td>
</tr>
<tr>
<td>Values</td>
<td>WT</td>
<td>KO+T2</td>
<td>5</td>
<td>5</td>
<td>-6.481813</td>
<td>4</td>
<td>0.003000</td>
<td>0.009</td>
<td>**</td>
</tr>
</tbody>
</table>

```r
eutrophils.long %>%
  pairwise_t_test(Values~Group, paired=TRUE, ref.group = "WT",
  p.adjust.method = "holm")
```

<table>
<thead>
<tr>
<th>.y.</th>
<th>group1</th>
<th>group2</th>
<th>n1</th>
<th>n2</th>
<th>statistic</th>
<th>df</th>
<th>p</th>
<th>p.adj</th>
<th>p.adj.signif</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Values</td>
<td>WT</td>
<td>KO</td>
<td>5</td>
<td>5</td>
<td>-8.657886</td>
<td>4</td>
<td>0.000979</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>2 Values</td>
<td>WT</td>
<td>KO+T1</td>
<td>5</td>
<td>5</td>
<td>1.310271</td>
<td>4</td>
<td>0.260000</td>
<td>0.260</td>
<td></td>
</tr>
<tr>
<td>3 Values</td>
<td>WT</td>
<td>KO+T2</td>
<td>5</td>
<td>5</td>
<td>-6.481813</td>
<td>4</td>
<td>0.003000</td>
<td>0.006</td>
<td></td>
</tr>
</tbody>
</table>

Tukey 😞
Exercise 7: Repeated measures ANOVA

neutrophils.long.csv

- Choose a graphical presentation consistent with the experimental design

```r
neutrophils.long %>%
  group_by(Experiment) %>%
  mutate(Difference = Values - Values[Group == "WT"])
  ungroup() -> neutrophils.long
```
Exercise 7: Repeated measures ANOVA
neutrophils.long.csv

- Choose a graphical presentation consistent with the experimental design

```r
eutrophils.long %>%
  filter(Group != "WT") %>%
  ggplot(aes(Group, Difference, fill=Group)) +
  geom_bar(stat = "summary", fun="mean", colour="black") +
  stat_summary(geom="errorbar", fun.data=mean_cl_normal, width=0.15) +
  geom_jitter(height = 0, width=0.1, alpha=0.5, size=3) +
  ylab("Mean difference from WT +/- 95% CI") +
  scale_y_continuous(breaks=seq(from=-40, by=10, to=80)) +
  scale_fill_brewer(palette = "PuOr")
```
Comparison between more than 2 groups
Two factors = Two predictors
Two-Way ANOVA
Two-way Analysis of Variance (Factorial ANOVA)

<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-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable A (Between Groups)</td>
<td>2.665</td>
<td>4</td>
<td>0.6663</td>
<td>8.42</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Within Groups (Residual)</td>
<td>5.775</td>
<td>73</td>
<td>0.0791</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>8.44</td>
<td>77</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

One-way ANOVA= 1 predictor variable

- **SS_T**: Total variance in the Data
- **SS_M**: Variance Explained by the Model
- **SS_R**: Unexplained Variance

Two-way ANOVA= 2 predictor variables: A and B

- **SS_T**: Total variance in the Data
- **SS_M**: Variance Explained by the Model
- **SS_R**: Unexplained Variance
- **SS_A**: Variance Explained by Variable A
- **SS_B**: Variance Explained by Variable B
- **SS_AxB**: Variance Explained by the Interaction of A and B

<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-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable A * Variable B</td>
<td>1978</td>
<td>2</td>
<td>989.1</td>
<td>F (2, 42) = 11.91</td>
<td>P &lt; 0.0001</td>
</tr>
<tr>
<td>Variable B (Between groups)</td>
<td>3332</td>
<td>2</td>
<td>1666</td>
<td>F (2, 42) = 20.07</td>
<td>P &lt; 0.0001</td>
</tr>
<tr>
<td>Variable A (Between groups)</td>
<td>168.8</td>
<td>1</td>
<td>168.8</td>
<td>F (1, 42) = 2.032</td>
<td>P = 0.1614</td>
</tr>
<tr>
<td>Residuals</td>
<td>3488</td>
<td>42</td>
<td>83.04</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Two-way Analysis of Variance

• **Interaction plots: Examples**

  • Fake dataset:
    • 2 factors: **Genotype** (2 levels) and **Condition** (2 levels)

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Condition</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotype 1</td>
<td>Condition 1</td>
<td>74.8</td>
</tr>
<tr>
<td>Genotype 1</td>
<td>Condition 1</td>
<td>65</td>
</tr>
<tr>
<td>Genotype 1</td>
<td>Condition 1</td>
<td>74.8</td>
</tr>
<tr>
<td>Genotype 1</td>
<td>Condition 2</td>
<td>75.2</td>
</tr>
<tr>
<td>Genotype 1</td>
<td>Condition 2</td>
<td>75</td>
</tr>
<tr>
<td>Genotype 1</td>
<td>Condition 2</td>
<td>75.2</td>
</tr>
<tr>
<td>Genotype 2</td>
<td>Condition 1</td>
<td>87.8</td>
</tr>
<tr>
<td>Genotype 2</td>
<td>Condition 1</td>
<td>65</td>
</tr>
<tr>
<td>Genotype 2</td>
<td>Condition 1</td>
<td>74.8</td>
</tr>
<tr>
<td>Genotype 2</td>
<td>Condition 2</td>
<td>88.2</td>
</tr>
<tr>
<td>Genotype 2</td>
<td>Condition 2</td>
<td>75</td>
</tr>
<tr>
<td>Genotype 2</td>
<td>Condition 2</td>
<td>75.2</td>
</tr>
</tbody>
</table>
Two-way Analysis of Variance

- **Interaction plots: Examples**
  - 2 factors: Genotype (2 levels) and Condition (2 levels)

**Single Effect**

**Genotype Effect**

**Condition Effect**
Two-way Analysis of Variance

- **Interaction plots**: Examples
  - 2 factors: Genotype (2 levels) and Condition (2 levels)

Zero or Both Effect

Zero Effect

Both Effect
Two-way Analysis of Variance

- **Interaction plots: Examples**
  - 2 factors: Genotype (2 levels) and Condition (2 levels)
Two-way Analysis of Variance

Example: goggles.csv

– The ‘beer-goggle’ effect

– **Study**: effects of alcohol on mate selection in night-clubs.
– Pool of independent judges scored the levels of attractiveness of the person that the participant was chatting up at the end of the evening.
– **Question**: is subjective perception of physical attractiveness affected by alcohol consumption?
  – Attractiveness on a scale from 0 to 100
Exercise 8: Two-way ANOVA
goggles.csv

- Load goggles.csv

- Graphically explore the data
  - effect of alcohol only
  - effect of gender only
  - effect of both

- Check the assumptions visually (plot+qqplot) and formally (test)
  levene_test(y ~ factor1*factor2)
Two-way Analysis of Variance

- As always, first step: get to know the data

```r
goggles %>%
  ggplot(aes(x=alcohol, y=attractiveness)) +
  geom_boxplot() +
  geom_jitter(height=0, width=0.1)
```

```r
goggles %>%
  ggplot(aes(x=gender, y=attractiveness)) +
  geom_boxplot() +
  geom_jitter(height=0, width=0.1)
```
Two-way Analysis of Variance

goggles %>%
ggplot(aes(alcohol, attractiveness, fill=gender)) +
  geom_boxplot(alpha=0.5) +
  scale_fill_brewer(palette="Dark2")
Two-way Analysis of Variance

goggles %>%
  ggplot(aes(gender, attractiveness, fill=alcohol)) +
  geom_boxplot(alpha=0.5) +
  scale_fill_brewer(palette="Dark2")
Two-way Analysis of Variance

goggles %>%
ggplot(aes(x=gender, y=attractiveness)) +
  geom_boxplot() +
  geom_jitter(height=0, width=0.1) +
  facet_grid(cols=vars(alcohol))
Two-way Analysis of Variance

Checking the assumptions

goggles %>%
ggplot(aes(sample = attractiveness, colour=gender)) +
  stat_qq() +
  stat_qq_line() +
  facet_grid(cols=vars(gender)) +
  scale_colour_brewer(palette = "Accent")

First assumption ✓
Two-way Analysis of Variance

Checking the assumptions

goggles %>%
group_by(gender, alcohol) %>%
  shapiro_test(attractiveness) %>%
  ungroup()

First assumption ✓

<table>
<thead>
<tr>
<th>gender</th>
<th>alcohol</th>
<th>variable</th>
<th>statistic</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>0 Pint</td>
<td>attractiveness</td>
<td>0.8715152</td>
<td>0.1559521</td>
</tr>
<tr>
<td>Female</td>
<td>2 Pints</td>
<td>attractiveness</td>
<td>0.8989639</td>
<td>0.2828089</td>
</tr>
<tr>
<td>Female</td>
<td>4 Pints</td>
<td>attractiveness</td>
<td>0.8972707</td>
<td>0.2729917</td>
</tr>
<tr>
<td>Male</td>
<td>0 Pint</td>
<td>attractiveness</td>
<td>0.9410603</td>
<td>0.6215419</td>
</tr>
<tr>
<td>Male</td>
<td>2 Pints</td>
<td>attractiveness</td>
<td>0.9666411</td>
<td>0.8704264</td>
</tr>
<tr>
<td>Male</td>
<td>4 Pints</td>
<td>attractiveness</td>
<td>0.9508657</td>
<td>0.7199577</td>
</tr>
</tbody>
</table>

Second assumption ✓

goggles %>%
  levene_test(attractiveness ~ gender*alcohol)

<table>
<thead>
<tr>
<th>df1</th>
<th>df2</th>
<th>statistic</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>42</td>
<td>1.425225</td>
<td>0.2350678</td>
</tr>
</tbody>
</table>
## Two-way Analysis of Variance

### With significant interaction (real data)

<table>
<thead>
<tr>
<th>ANOVA table</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F (DFn, DFd)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interaction</td>
<td>1978</td>
<td>2</td>
<td>989.1</td>
<td>F (2, 42) = 11.91</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Alcohol Consumption</td>
<td>3332</td>
<td>2</td>
<td>1666</td>
<td>F (2, 42) = 20.07</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Gender</td>
<td>168.8</td>
<td>1</td>
<td>168.8</td>
<td>F (1, 42) = 2.032</td>
<td>0.1614</td>
</tr>
<tr>
<td>Residual</td>
<td>3488</td>
<td>42</td>
<td>83.04</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Without significant interaction (fake data)

<table>
<thead>
<tr>
<th>ANOVA table</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F (DFn, DFd)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interaction</td>
<td>7.292</td>
<td>2</td>
<td>3.646</td>
<td>F (2, 42) = 0.06872</td>
<td>0.9337</td>
</tr>
<tr>
<td>Alcohol Consumption</td>
<td>5026</td>
<td>2</td>
<td>2513</td>
<td>F (2, 42) = 47.37</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Gender</td>
<td>438.0</td>
<td>1</td>
<td>438.0</td>
<td>F (1, 42) = 8.257</td>
<td>0.0063</td>
</tr>
<tr>
<td>Residual</td>
<td>2228</td>
<td>42</td>
<td>53.05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Two-way Analysis of Variance

goggles %>%
anova_test(attractiveness ~ alcohol + gender + alcohol * gender)

![ANOVA Table (type II tests)](image)

<table>
<thead>
<tr>
<th>Effect</th>
<th>DFn</th>
<th>DFd</th>
<th>F</th>
<th>p</th>
<th>p &lt; .05</th>
<th>ges</th>
</tr>
</thead>
<tbody>
<tr>
<td>alcohol</td>
<td>2</td>
<td>42</td>
<td>7.66e-07</td>
<td>*</td>
<td>0.489</td>
<td></td>
</tr>
<tr>
<td>gender</td>
<td>1</td>
<td>42</td>
<td>2.032e-01</td>
<td>0.046</td>
<td></td>
<td></td>
</tr>
<tr>
<td>alcohol:gender</td>
<td>2</td>
<td>42</td>
<td>11.91e-05</td>
<td>*</td>
<td>0.362</td>
<td></td>
</tr>
</tbody>
</table>

goggles %>%
group_by(alcohol) %>%
tukey_hsd(attractiveness ~ gender) %>%
tukey_hsd(attractiveness ~ gender) %>%
ungroup()

![Box plot](image)

**Answer:** There is a significant effect of alcohol consumption on the way the attractiveness of a date is perceived but it varies significantly between genders ($p=7.99e-05$).

With 2 pints or less, boys seem to be very slightly more picky about their date than girls (but not significantly so) but with 4 pints the difference is reversed and significant ($p=0.0003$).
Two-way Analysis of Variance

- **Work in progress** # ggpubr package #

```r
goggles %>%
group_by(alcohol) %>%
tukey_hsd(attractiveness ~ gender) %>%
add_xy_position(x = "alcohol") %>%
ungroup() -> tukey.results
```

---

### Data Summary

<table>
<thead>
<tr>
<th>Alcohol</th>
<th>Gender</th>
<th>Group1</th>
<th>Group2</th>
<th>Estimate</th>
<th>Cont.low</th>
<th>Cont.high</th>
<th>p.adj</th>
<th>p.adj.signif</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 Pint</td>
<td>Female</td>
<td>Male</td>
<td></td>
<td>6.250</td>
<td>-2.427379</td>
<td>14.93738</td>
<td>0.145000</td>
<td>ns</td>
</tr>
<tr>
<td>2 Pints</td>
<td>Female</td>
<td>Male</td>
<td></td>
<td>4.575</td>
<td>-6.336958</td>
<td>15.08696</td>
<td>0.359000</td>
<td>ns</td>
</tr>
<tr>
<td>4 Pints</td>
<td>Female</td>
<td>Male</td>
<td></td>
<td>-21.675</td>
<td>-31.468394</td>
<td>-12.00661</td>
<td>0.000282</td>
<td>***</td>
</tr>
</tbody>
</table>

---

```r
goggles %>%
ggplot(aes(alcohol, attractiveness, colour = gender)) +
geom_boxplot() +
stat_pvalue_manual(tukey.results) +
scale_colour_brewer(palette = "Dark2")
```
Two-way Analysis of Variance

- **Work in progress** # ggpubr package # Actual p-values rather than NS or *

```r
# Two-way Analysis of Variance

goggles %>%
group_by(alcohol) %>%
tukey_hsd(attractiveness ~ gender) %>%
mutate(p.adj.signif = p.adj) %>%
add_xy_position(x = "alcohol") %>%
ungroup() -> tukey.results

# ggplot code

```r
ggplot(aes(alcohol, attractiveness, colour = gender)) + geom_boxplot() +
stat_pvalue_manual(tukey.results) +
scale_colour_brewer(palette = "Dark2")
```
Two-way Analysis of Variance

- Now a quick way to have a look at the interaction

```r
goggles %>%
group_by(gender, alcohol) %>%
s summarise(mean=mean(attractiveness)) %>%
ungroup() -> goggles.summary
```

---

```r
goggles.summary %>%
ggplot(aes(x=alcohol, y= mean, colour=gender, group=gender)) +
geom_line() +
geom_point()
```