Analysis of Quantitative data

Anne Segonds-Pichon

v2020-09
Outline of this section

• Assumptions for parametric data

• Comparing two means: Student’s t-test

• Comparing more than 2 means
  • One factor: One-way ANOVA
  • Two factors: Two-way ANOVA

• Relationship between 2 continuous variables: Correlation
Introduction

• Key concepts to always keep in mind
  – Null hypothesis and error types
  – Statistics inference
  – Signal-to-noise ratio
The null hypothesis and the error types

- **The null hypothesis** ($H_0$): $H_0 = \text{no effect}$
  - e.g. no difference between 2 genotypes
- The aim of a statistical test is to reject or not $H_0$.

<table>
<thead>
<tr>
<th>Statistical decision</th>
<th>True state of $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_0$ True (no effect)</td>
</tr>
<tr>
<td><strong>Reject $H_0$</strong></td>
<td>Type I error $\alpha$</td>
</tr>
<tr>
<td></td>
<td>False Positive</td>
</tr>
<tr>
<td><strong>Do not reject $H_0$</strong></td>
<td>Correct</td>
</tr>
<tr>
<td></td>
<td>True Negative</td>
</tr>
</tbody>
</table>

- Traditionally, a test or a difference is said to be “significant” if the probability of type I error is: $\alpha =< 0.05$
- **High specificity** = low **False Positives** = low **Type I error**
- **High sensitivity** = low **False Negatives** = low **Type II error**
Stats are all about understanding and controlling variation.

If the noise is low then the signal is detectable ...

\[ \text{signal} \] = statistical significance

... but if the noise (i.e. interindividual variation) is large
then the same signal will not be detected

\[ \text{signal} \] = no statistical significance

In a statistical test, the ratio of signal to noise determines the significance.
Analysis of Quantitative Data

• Choose the correct statistical test to answer your question:
  – They are 2 types of statistical tests:
    • **Parametric tests** with 4 assumptions to be met by the data,
    • **Non-parametric tests** with no or few assumptions (e.g. Mann-Whitney test) and/or for qualitative data (e.g. Fisher’s exact and \( \chi^2 \) tests).
Assumptions of Parametric Data

• All parametric tests have 4 basic assumptions that must be met for the test to be accurate.

First assumption: Normally distributed data
– Normal shape, bell shape, Gaussian shape

• Transformations can be made to make data suitable for parametric analysis.
Assumptions of Parametric Data

• Frequent departures from normality:
  - **Skewness**: lack of symmetry of a distribution
    - Skewness < 0 (a) Negatively skewed
    - Skewness = 0 (b) Normal (no skew)
    - Skewness > 0 (c) Positively skewed

  ![Skewness Illustration](image)

  - **Kurtosis**: measure of the degree of ‘peakedness’ in the distribution
    - The two distributions below have the same variance approximately the same skew, but differ markedly in kurtosis.

  ![Kurtosis Illustration](image)

  - More peaked distribution: kurtosis > 0
  - Flatter distribution: kurtosis < 0
  - Flatter distribution: kurtosis < 0
Assumptions of Parametric Data

**Second assumption: Homoscedasticity (Homogeneity in variance)**

- The variance should not change systematically throughout the data

**Third assumption: Interval data (linearity)**

- The distance between points of the scale should be equal at all parts along the scale.

**Fourth assumption: Independence**

- Data from different subjects are independent
  - Values corresponding to one subject do not influence the values corresponding to another subject.
  - Important in repeated measures experiments
Is there a difference between my groups regarding the variable I am measuring?  
- e.g. are the mice in the group A heavier than those in group B?

- Tests with 2 groups:
  - Parametric: Student’s t-test  
  - Non parametric: Mann-Whitney/Wilcoxon rank sum test

- Tests with more than 2 groups:
  - Parametric: Analysis of variance (one-way and two-way ANOVA)  
  - Non parametric: Kruskal Wallis (one-way ANOVA equivalent)

Is there a relationship between my 2 (continuous) variables?  
- e.g. is there a relationship between the daily intake in calories and an increase in body weight?

- Test: Correlation (parametric or non-parametric)
Comparison between 2 groups
Comparison between 2 groups: Student’s $t$-test

• **Basic idea:**
  - When we are looking at the differences between scores for 2 groups, we have to judge the difference between their means relative to the spread or variability of their scores.
  - Eg: comparison of 2 groups: control and treatment
Variability does matter

Scatter plot

Bar chart
Student’s t-test

- **medium variability**
- **high variability**
- **low variability**
Student’s t-test

\[
\frac{\text{signal}}{\text{noise}} = \frac{\bar{x}_T - \bar{x}_C}{\sqrt{\frac{\text{var}_T}{n_T} + \frac{\text{var}_C}{n_C}}}
\]

= t-value
n = 3

Group A

Group B

Values

15
20
25
30

~ 2 x SEM: p~0.05

n = 10

Group A

Group B

Values

15
20
25
30

~ 1 x SEM: p~0.05

~ 2 x SEM: p~0.01

~ 2 x SEM: p~0.01

~ 4.5 x SEM: p~0.01

n = 3

n = 10
Student’s t-test

• **Independent t-test**
  • Difference between 2 means of one variable for **two independent groups**
  • Example: difference in weight between WT and KO mice

• **Paired t-test**
  • Difference between two measures of one variable for **one group**:
    • Example: before-after measurements
      • the second ‘sample’ of values comes from the same subjects (mouse, petri dish ...).
    • Importance of experimental design!

• **One-Sample t-test**
  • Difference between the mean of a single variable and a specified constant.
Example: coyotes

- **Question**: do male and female coyotes differ in size?
- **Sample size**
- **Data exploration**
- **Check the assumptions for parametric test**
- **Statistical analysis: Independent t-test**
Exercise 3: Power analysis

- Example case:

No data from a pilot study but we have found some information in the literature.

In a study run in similar conditions as in the one we intend to run, male coyotes were found to measure: 92cm +/- 7cm (SD).

We expect a 5% difference between genders.

- smallest biologically meaningful difference

```
power.t.test(n = NULL, delta = NULL, sd = 1, sig.level = NULL, power = NULL, type = c("two.sample", "one.sample", "paired"), alternative = c("two.sided", "one.sided"))
```
**Example case:**

We don’t have data from a pilot study but we have found some information in the literature.

In a study run in similar conditions as in the one we intend to run, male coyotes were found to measure: **92cm +/- 7cm (SD)**

We expect a **5% difference** between genders with a similar variability in the female sample.

Mean 1 = 92  
Mean 2 = 87.4 (5% less than 92cm)

delta = 92 – 87.4  
sd = 7

```
power.t.test(delta=92-87.4, sd=7, sig.level=0.05, power=0.8)
```

We need a sample size of **n~76 (2*38)**
Data exploration ≠ plotting data
Exercise 4: Data exploration
coyote.csv

- The file contains individual body length of male and female coyotes.

**Question:** do male and female coyotes differ in size?

- Load *coyote.csv*
- Plot the data as boxplot, histogram, violinplot and stripchart

Data exploration ≠ plotting data
Exercise 4: Data exploration

- Explore data using 4 different representations:
Exercise 4: `facet_grid(rows=vars(row), cols=vars(column))`

`facet_grid(cols=vars(gender))`

2 columns: one per gender

One row
**Exercise 4: geom_jitter()**

- Stripchart
  - Variation of `geom_point()`: `geom_jitter()`

```r
coyote %>%
ggplot(aes(x=gender, y=length)) + geom_point()
```

```r
coyote %>%
ggplot(aes(x=gender, y=length)) + geom_jitter(height=0, width=0.3)
```
Exercise 4: `stat_summary()`

- Stripchart
  - `stat_summary()`
  - What statistical summary: mean: `fun = "mean"`
  - What `geom()`: choice of graphical representation: a line: `geom_errorbar()`

```r
stat_summary(geom="errorbar", fun="mean", fun.min="mean", fun.max="mean")
```

```r
coyote %>%
ggplot(aes(gender, length)) +
  geom_jitter(height=0, width=0.2)+
  stat_summary(geom = "errorbar", fun="mean", fun.min="mean", fun.max="mean")
```
Exercise 4: Data exploration

• Explore data using 4 different representations:

```r
coyote %>%
ggplot(aes(x=gender, y=length))+
geom_boxplot() +
geom_violin() +
geom_histogram() +
geom_jitter() +
stat_summary()
```

Have a go!
**Exercise 4: Exploring data - Stripchart**

coyote %>%
  ggplot(aes(gender, length)) +
  geom_jitter(height=0, width=0.2) +
  stat_summary(geom = "errorbar", fun="mean", fun.min="mean", fun.max="mean")

---

coyote %>%
  ggplot(aes(gender, length, colour=gender)) +
  geom_jitter(height=0, size=4, width=0.2, show.legend = FALSE) +
  ylab("Length (cm)") +
  scale_colour_brewer(palette="Dark2") +
  xlab(NULL) +
  stat_summary(geom="errorbar", fun=mean, fun.min=mean, fun.max=mean, colour="black", size=1.2, width=0.6)
**Exercise 4: Exploring data - Answers**

Boxplots and beanplots

```r
coyote %>%
ggplot(aes(x=gender, y=length)) +
geom_boxplot()
```

```r
coyote %>%
ggplot(aes(x=gender, y=length)) +
geom_violin()
```
Exercise 4: Exploring data - Answers

Boxplots and beanplots

coyote %>%
ggplot(aes(x=gender, y=length, fill=gender)) +
  stat_boxplot(geom="errorbar", width=0.5) +
  geom_boxplot(show.legend=FALSE)+
  ylab("Length (cm)")+xlab(NULL)+
  scale_fill_manual(values = c("orange","purple"))

coyote %>%
ggplot(aes(gender, length, fill=gender)) +
  geom_violin(trim=FALSE, size=1, show.legend=FALSE)+
  ylab("Length (cm)")+xlab(NULL)+
  scale_fill_brewer(palette="Dark2")+
  stat_summary(geom = "point", fun = "median", show.legend=FALSE)
Exercise 4: Exploring data - Answers

Histograms

coyote %>%
  ggplot(aes(length)) +
  geom_histogram(binwidth = 4, colour="black") +
  facet_grid(cols=vars(gender))

also works
facet_wrap(vars(gender))
### Exercise 4: Exploring data - Answers

#### Histograms

```r
coyote %>%
ggplot(aes(length, fill=gender)) +
geom_histogram(binwidth = 4.5, colour="black", show.legend = FALSE) +
scale_fill_brewer(palette="Dark2") +
facet_grid(cols=vars(gender))
```
Exercise 4 extra: Exploring data - Graph combinations

```r
coyote %>%
  ggplot(aes(gender, length)) +
  geom_boxplot(width=0.2) +
  geom_violin()
```

```r
coyote %>%
  ggplot(aes(gender, length, fill=gender)) +
  geom_violin(size=1, trim = FALSE, alpha=0.2, show.legend=FALSE) +
  geom_boxplot(width=0.2, outlier.size=5, outlier.colour = "darkred", show.legend=FALSE) +
  scale_fill_brewer(palette="Dark2") +
  ylab("Length (cm)") +
  xlab(NULL) +
  scale_x_discrete(labels=c("female"="Female", "male"="Male"), limits=c("male", "female"))
```
Exercise 4 extra: Exploring data - Graph combinations

coyote %>%
ggplot(aes(gender, length)) +
geom_boxplot() +
geom_jitter(height=0, width=0.2)

coyote %>%
ggplot(aes(gender, length)) +
geom_boxplot(outlier.shape=NA) +
stat_boxplot(geom="errorbar", width=0.2) +
geom_jitter(height=0, width=0.1, size=2, alpha=0.5, colour="red") +
ylab("Length (cm)")
Checking the assumptions
Normality assumption

QQ plot = Quantile – Quantile plot

Our coyotes

Quantiles:

Mean = 0
SD = 1
Same sample size
Perfectly normal distribution

Normality ✔️ (ish)
Normality assumption: QQ plot

coyote %>%
ggplot(aes(sample = length)) +
  stat_qq()+
  stat_qq_line()

coyote %>%
ggplot(aes(sample = length)) +
  stat_qq(size=2, colour="darkorange3")+
  stat_qq_line()+
  ylab("Body Length (cm)")+
  scale_y_continuous(breaks=seq(from=70, by=5, to=110))+
  scale_x_continuous(breaks=seq(from=-3, by=0.5, to=3))
Assumptions of Parametric Data

- First assumption: Normality
  - Shapiro-Wilk test `shapiro_test()` # rstatix package#
    - It is based on the correlation between the data and the corresponding normal scores.

- Second assumption: Homoscedasticity
  - Levene test `levene_test()`

```
coyote %>%
  group_by(gender) %>%
  shapiro_test(length) %>%
  ungroup()
```

<table>
<thead>
<tr>
<th>gender</th>
<th>variable</th>
<th>statistic</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>female</td>
<td>length</td>
<td>0.9700101</td>
<td>0.3164448</td>
</tr>
<tr>
<td>male</td>
<td>length</td>
<td>0.9844570</td>
<td>0.8189831</td>
</tr>
</tbody>
</table>

Normality ✅

```
coyote %>%
  levene_test(length ~ gender)
```

```
df1  df2  statistic   p
     1   84  0.167929   0.6830022
```

Homogeneity in variance ✅

**Normality**

Other classic: D’Agostino-Pearson test
# fBasic package#
`dagoTest()`

**Homoscedasticity**

More robust: Brown-Forsythe test
# onewaytests package#, `bf()`
Other classic: Bartlett test
`bartlett.test()`
Independent t-test: results (tidyverse)

coyote.csv

coyote %>%
  t_test(length~gender)

<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>
</tr>
</thead>
<tbody>
<tr>
<td>length</td>
<td>female</td>
<td>male</td>
<td>43</td>
<td>43</td>
<td>-1.641109</td>
<td>84</td>
<td>0.105</td>
</tr>
</tbody>
</table>

coyote %>%
  group_by(gender) %>%
  get_summary_stats(length, type = "mean_sd") %>%
  ungroup()

<table>
<thead>
<tr>
<th>gender</th>
<th>variable</th>
<th>n</th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>female</td>
<td>length</td>
<td>43</td>
<td>89.712</td>
<td>6.550</td>
</tr>
<tr>
<td>male</td>
<td>length</td>
<td>43</td>
<td>92.056</td>
<td>6.696</td>
</tr>
</tbody>
</table>

• **Answer:** Males tend to be longer than females but not significantly so (p=0.1045).

• **Power:** How many more coyotes to reach significance?
  • Re-run the power analysis with mean=89.7 for females: n~250
  • **But does it make sense?**
Sample size: the bigger the better?

- It takes huge samples to detect tiny differences but tiny samples to detect huge differences.

- What if the tiny difference is meaningless?
  - Beware of **overpower**
  - Nothing wrong with the stats: it is all about interpretation of the results of the test.

- Remember the important first step of power analysis
  - **What is the effect size of biological interest?**
Independent \( t \)-test: results

The old-fashion way

t = 1.641 < 1.984: not significant

Critical value
Plot ‘coyote.csv’ data: Plotting data

coyote %>%
  ggplot(aes(gender, length, colour=gender)) +
  geom_bar(stat = "summary", fun="mean", width=0.4, alpha=0, colour="black") +
  geom_jitter(height=0, width=0.1)

• Add error bars

coyote %>%
  ggplot(aes(gender, length, colour=gender)) +
  geom_bar(stat = "summary", fun="mean", width=0.4, alpha=0, colour="black") +
  geom_jitter(height=0, width=0.1) +
  stat_summary(geom="errorbar", colour="black", width=0.2)
Prettier version

```r
coyote %>%
  ggplot(aes(gender, length, colour=gender, fill=gender)) +
  geom_bar(stat="summary", fun="mean", width=0.4, alpha=0.2, colour="black", show.legend=FALSE) +
  stat_summary(geom="errorbar", colour="black", width=0.2) +
  geom_jitter(height=0, width=0.1, show.legend=FALSE) +
  scale_colour_brewer(palette="Dark2") +
  scale_fill_brewer(palette="Dark2") +
  theme(legend.position = "none") +
  scale_x_discrete(limits = c("male", "female"), labels = c("male"="Male", "female"="Female")) +
  xlab(NULL) +
  ylab("Length (cm)")
```
Plot ‘coyote.csv’ data: Plotting data

- **Work in progress** # ggsignif package #

```r
coyote %>%
ggplot(aes(gender, length)) +
  stat_boxplot(geom="errorbar", width=0.2) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(height=0, width=0.1, size = 2, alpha = 0.5, colour="red") +
  scale_x_discrete(limits = c("male", "female"), labels = c("male"="Male", "female"="Female")) +
  ylab("Length (cm)") +
  xlab(NULL) +
  geom_signif(comparisons = list(c("female", "male")), map_signif_level=T, test = "t.test")
```

![Graph showing boxplot and jitter plot with significant differences marked]
Exercise 5: Dependent or Paired t-test

working.memory.csv

- A researcher is studying the effects of dopamine depletion on working memory in rhesus monkeys.
  - A group of rhesus monkeys (n=15) performs a task involving memory after having received a placebo. Their performance is graded on a scale from 0 to 100. They are then asked to perform the same task after having received a dopamine depleting agent.

- **Question**: does dopamine affect working memory in rhesus monkeys?
  - Load `working.memory.csv` and check out the structure of the data.
  - Work out the difference: DA.depletion – placebo and assign the difference to a column: `working.memory$difference`
  - Plot the difference as a stripchart with a mean
  - Add **confidence intervals as error bars**
    - Clue: `stat_summary(…, fun.data=mean_cl_normal)` # Hmisc package #
  - Run the paired t-test. `t_text(var ~ 1, mu=0)`
**Exercise 5: Dependent or Paired \( t \)-test - Answers**

```r
working.memory %>%
  mutate(difference = DA.depletion - placebo) -> working.memory
```

# Hmisc package #
working.memory %>%
  ggplot(aes("DA.Depletion", difference)) +
  geom_jitter(height=0, width=0.05, size=4, colour="chartreuse3") +
  stat_summary(geom="errorbar", fun="mean", fun.min="mean", fun.max="mean", width=0.3, size=1) +
  stat_summary(geom="errorbar", fun.data=mean_cl_normal, width=0.15) +
  scale_y_continuous(breaks=-16:0, limits=c(-16, 0)) +
  xlab(NULL) +
  ylab("Mean difference +/- 95% CI")
```
**Exercise 5:** Dependent or Paired *t*-test (tidyverse)

**Question:** does dopamine affect working memory in rhesus monkeys?

```r
working.memory %>%
  shapiro_test(difference)
```

> variable statistic     p
> difference    0.9772671 0.9474075

```r
working.memory %>%
  t_test(difference ~ 1, mu=0)
```

<table>
<thead>
<tr>
<th>.y.</th>
<th>group1</th>
<th>group2</th>
<th>n</th>
<th>statistic</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>difference</td>
<td>1</td>
<td>null model</td>
<td>15</td>
<td>-8.616059</td>
<td>14</td>
<td>5.71e-07</td>
</tr>
</tbody>
</table>

**Answer:** the injection of a dopamine-depleting agent significantly affects working memory in rhesus monkeys (*t*= -8.62, *df*= 14, *p*= 5.71e-07).
Dependent or Paired $t$-test

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

```r
working.memory.long %>%
  t_test(scores ~ treatment, paired = TRUE) -> stat.test

working.memory.long %>%
  ggpaired(x = "treatment", y = "scores", color = "treatment",
           palette = "Dark2", line.color = "gray", line.size = 0.4) +
  scale_y_continuous(breaks = seq(from = 0, by = 5, to = 60),
                    limits = c(0, 60)) +
  stat_pvalue_manual(stat.test, label = "p", y.position = 55)
```
Comparison between more than 2 groups
One factor
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?

Signal
Noise = \frac{\text{Difference between the means}}{\text{Variability in the groups}}

= F \text{ ratio}
One-Way Analysis of variance

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

- 78 differences: \( \sum_{1}^{78} (\text{value}_n - \text{grand mean})^2 \)
- Sum of squared errors

<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>69.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Analysis of variance: how does it work?

Source of variation | Sum of Squares | df | Mean Square | F | p-value
--- | --- | --- | --- | --- | ---
Between Groups | **18.1** | | | | |
Within Groups | | | | | |
Total | 69.9 | | | | |

5 differences: $\sum_{1}^{5} (\text{mean}_n - \text{grand mean})^2 = \text{Sum of squared errors Between the groups}$
### 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 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>

#### Diagram

- **Grand mean**: Average of all data points.
- **Group mean**: Average of data points within each group.

#### Formula

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

**Sum of squared errors**

**Within the Groups**

---

The diagram illustrates the concept of variance analysis, showing how data points are distributed around group means and the grand mean, and how the sum of squared differences is calculated to assess the variance between and within groups.
Analysis of variance: how does it work?

### Signal Noise

<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 $df = n-1$

$n = \text{number of values}, \ k = \text{number of groups}$

**Between groups**: $df = 4 \ (k-1)$

**Within groups**: $df = 73 \ (n-k = n_1-1 + \ldots + n_5-1)$
## Analysis of variance: how does it work?

### Signal vs. Noise

<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?

Mean squares = \( \text{Sum of Squares} / (n-1) = \text{Variance} \)

\[
F \text{ ratio} = \frac{\text{Variance between the groups}}{\text{Variance within the groups (individual variability)}} = \frac{4.5}{0.71} = 6.34
\]
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 \(1 - 0.857 = 0.143\) 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**: $\alpha_{\text{adj}} = 0.05/n$ 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
protein.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 - Answers

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 – QQ plot

```
protein %>%
ggplot(aes(sample = expression)) +
  stat_qq(size=3) +
  stat_qq_line()
```
protein %>%
group_by(line) %>%
  shapiro_test(expression) %>%
  ungroup()

Exercise 6: One-way ANOVA - Answers

<table>
<thead>
<tr>
<th>line</th>
<th>variable</th>
<th>statistic</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>expression</td>
<td>0.9295671</td>
<td>0.3755460156</td>
</tr>
<tr>
<td>B</td>
<td>expression</td>
<td>0.9535144</td>
<td>0.6887867228</td>
</tr>
<tr>
<td>C</td>
<td>expression</td>
<td>0.8196840</td>
<td>0.0029210891</td>
</tr>
<tr>
<td>D</td>
<td>expression</td>
<td>0.7530720</td>
<td>0.0003548725</td>
</tr>
<tr>
<td>E</td>
<td>expression</td>
<td>0.9670693</td>
<td>0.7411280600</td>
</tr>
</tbody>
</table>

protein %>%
  mutate(log10.expression=log10(expression)) -> protein
One-way ANOVA – Outliers identification

```r
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

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)

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: change of scale

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

Before log-transformation

First assumption ✓
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>0.37725321</td>
</tr>
<tr>
<td>C</td>
<td>log10.expression</td>
<td>0.9657060</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>0.9313425</td>
<td>0.20502703</td>
</tr>
</tbody>
</table>

First assumption ✓ ish

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

df1  df2  statistic    p
4     73   0.982112  0.4227373

Second assumption ✓
Analysis of variance

• Step 1: omnibus test

```
data %>%
anova_test(y~x)
```

• Step 2: post-hoc tests

**Tukey correction**
```
data %>%
tukey_hsd(y~x)
```

**Bonferroni correction**  # emmeans package #
```
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)
```
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>p &lt; .05</th>
<th>ges</th>
<th>η²</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>
<td></td>
</tr>
</tbody>
</table>

generalised effect size (Eta squared η²) = R² ish

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

Tukey correction

table

<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>1</td>
<td>line A</td>
<td>B</td>
<td>-0.25024832</td>
<td>-0.578882494</td>
<td>0.07838585</td>
<td>2.19e-01</td>
<td>ns</td>
</tr>
<tr>
<td>2</td>
<td>line A</td>
<td>C</td>
<td>-0.07499724</td>
<td>-0.374997820</td>
<td>0.22500335</td>
<td>9.56e-01</td>
<td>ns</td>
</tr>
<tr>
<td>3</td>
<td>line A</td>
<td>D</td>
<td>0.30549397</td>
<td>0.005493391</td>
<td>0.60549456</td>
<td>4.39e-02</td>
<td>*</td>
</tr>
<tr>
<td>4</td>
<td>line A</td>
<td>E</td>
<td>0.13327517</td>
<td>-0.166725416</td>
<td>0.43327575</td>
<td>7.27e-01</td>
<td>ns</td>
</tr>
<tr>
<td>5</td>
<td>line B</td>
<td>C</td>
<td>0.1752108</td>
<td>-0.124749499</td>
<td>0.4752167</td>
<td>4.81e-01</td>
<td>ns</td>
</tr>
<tr>
<td>6</td>
<td>line B</td>
<td>D</td>
<td>0.55574230</td>
<td>0.255747172</td>
<td>0.85574288</td>
<td>1.83e-05</td>
<td>****</td>
</tr>
<tr>
<td>7</td>
<td>line B</td>
<td>E</td>
<td>0.38352349</td>
<td>0.083522904</td>
<td>0.68352407</td>
<td>5.48e-03</td>
<td>**</td>
</tr>
<tr>
<td>8</td>
<td>line C</td>
<td>D</td>
<td>0.38049121</td>
<td>0.112162532</td>
<td>0.64881989</td>
<td>1.54e-03</td>
<td>***</td>
</tr>
<tr>
<td>9</td>
<td>line C</td>
<td>E</td>
<td>0.20827240</td>
<td>-0.060056276</td>
<td>0.47660108</td>
<td>2.02e-01</td>
<td>ns</td>
</tr>
<tr>
<td>10</td>
<td>line D</td>
<td>E</td>
<td>-0.17221881</td>
<td>-0.440547487</td>
<td>0.09610987</td>
<td>3.84e-01</td>
<td>ns</td>
</tr>
</tbody>
</table>
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>p &lt; .05</th>
<th>ges</th>
<th>( \eta^2 )</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>
<td>R^2 ish</td>
</tr>
</tbody>
</table>

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

```
protein %>%
  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>log10.expression</td>
<td>A</td>
<td>B</td>
<td>73</td>
<td>2.1299578</td>
<td>3.654611e-02</td>
<td>3.654611e-01</td>
<td>ns</td>
</tr>
<tr>
<td>log10.expression</td>
<td>A</td>
<td>C</td>
<td>73</td>
<td>0.6992552</td>
<td>4.866147e-01</td>
<td>1.000000e+00</td>
<td>ns</td>
</tr>
<tr>
<td>log10.expression</td>
<td>A</td>
<td>D</td>
<td>73</td>
<td>-2.8483483</td>
<td>5.705474e-03</td>
<td>5.705474e-02</td>
<td>ns</td>
</tr>
<tr>
<td>log10.expression</td>
<td>A</td>
<td>E</td>
<td>73</td>
<td>-1.2426238</td>
<td>2.179833e-01</td>
<td>1.000000e+00</td>
<td>ns</td>
</tr>
<tr>
<td>log10.expression</td>
<td>B</td>
<td>C</td>
<td>73</td>
<td>-1.6339966</td>
<td>1.065653e-01</td>
<td>1.000000e+00</td>
<td>ns</td>
</tr>
<tr>
<td>log10.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>log10.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>log10.expression</td>
<td>C</td>
<td>E</td>
<td>73</td>
<td>-2.1710868</td>
<td>3.317601e-02</td>
<td>3.317601e-01</td>
<td>ns</td>
</tr>
<tr>
<td>log10.expression</td>
<td>D</td>
<td>E</td>
<td>73</td>
<td>1.7952545</td>
<td>7.675206e-02</td>
<td>7.675206e-01</td>
<td>ns</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)
```

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

```
       diff     lwr     upr     p adj
B-A -0.25024832 -0.378882494 0.07838385 0.2187264
C-A -0.07499724 -0.374997820 0.22500335 0.9560187
D-A  0.30549397  0.005493391 0.60549456 0.0438762
E-A  0.13327317 -0.168725416 0.43327357 0.7265567
C-B  0.17525108 -0.124743499 0.47525167 0.4809387
D-B  0.55742330  0.255747122 0.85742888 0.0001813
E-B  0.48352349  0.083529694 0.88352407 0.0054767
C-D  0.38049121 -0.11265235 0.64818988 0.9915416
E-D  0.20827240 -0.060056276 0.47660108 0.2023355
```

95% family-wise confidence level

```
95% family-wise confidence level

Tukey multiple comparisons of means

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

   Diff   Low  High     Signif
B-A -0.250 -0.378 0.078 *
C-A -0.075 -0.375 0.225 
D-A  0.305  0.005 0.605 **
E-A  0.133 -0.169 0.433 
C-B  0.175 -0.125 0.475 *
D-B  0.557  0.256 0.857 *
E-B  0.484  0.084 0.883 
C-D  0.380 -0.113 0.648 *
E-D  0.208 -0.061 0.477 
```

Differences in mean levels of proteinSline
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()
Analysis of variance

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(height=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+Treatment 2.
• **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
  
  \[ \text{anova_test(dv =, wid =, within =) -> res.aov} \]
  \[ \text{get_anova_table(res.aov)} \]
  \[ \text{pairwise_t_test(p.adjust.method =)} \]

• 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 %>%
  ggplot(aes(Group, Values)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(height = 0, width = 0.2)
```

```r
table <- read_csv('neutrophils.long.csv')

library(shiny)

shapiro_test <- function(data, variable) {
  test_result <- shapiro.test(data[[variable]])
  c(statistic = test_result$statistic, p = test_result$p.value)
}

shapiro_results <- map2_df(table, shapiro_test, variable)

shapiro_results$variable <- as.factor(shapiro_results$variable)

pivot_data <- pivot_longer(shapiro_results, names_to = 'Group', values_to = 'statistic')

pivot_data$variable <- as.factor(pivot_data$variable)

print(pivot_data)
```

<table>
<thead>
<tr>
<th>Group</th>
<th>variable</th>
<th>statistic</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>KO</td>
<td>Values</td>
<td>0.9117498</td>
<td>0.4781767</td>
</tr>
<tr>
<td>KO+T1</td>
<td>Values</td>
<td>0.9865912</td>
<td>0.9664514</td>
</tr>
<tr>
<td>KO+T2</td>
<td>Values</td>
<td>0.8529329</td>
<td>0.2039683</td>
</tr>
<tr>
<td>WT</td>
<td>Values</td>
<td>0.9482754</td>
<td>0.7248636</td>
</tr>
</tbody>
</table>
**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>x 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>* 0.656</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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>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>0.0009</td>
<td>**</td>
</tr>
</tbody>
</table>
```
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"])
```

```
Experiment Group Values Difference
Exp1 WT 34 0
Exp1 KO 53 19
Exp1 KO+T1 35 1
Exp1 KO+T2 91 57
Exp2 WT 23 0
Exp2 KO 52 29
Exp2 KO+T1 30 7
Exp2 KO+T2 9 76
Exp3 WT 45 0
Exp3 KO 69 24
```
Exercise 7: Repeated measures ANOVA
neutrophils.long.csv

- Choose a graphical presentation consistent with the experimental design

```
neutrophils.long %>%
  filter(Grp !="WT") %>%
  ggplot(aes(Grp, Diff, fill=Grp)) +
  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-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 Between Groups
- \( SS_R \) Unexplained Variance Within Groups

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

- \( SS_T \) Total variance in the Data
- \( SS_M \) Variance Explained by the Model
- \( SS_A \) Variance Explained by Variable A
- \( SS_B \) Variance Explained by Variable B
- \( SS_{AxB} \) Variance Explained by the Interaction of A and B
- \( SS_R \) Unexplained Variance

<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)

Interaction
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 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 ✓

Second assumption ✓
# Two-way Analysis of Variance

## With significant interaction (real data)

<table>
<thead>
<tr>
<th>Source</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>Source</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>
goggles %>%
  anova_test(attractiveness ~ alcohol + gender + alcohol*gender)

<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>* 0.489</td>
<td>0.489</td>
<td></td>
</tr>
<tr>
<td>gender</td>
<td>1</td>
<td>42</td>
<td>1.61e-01</td>
<td>0.046</td>
<td>0.046</td>
<td></td>
</tr>
<tr>
<td>alcohol:gender</td>
<td>2</td>
<td>42</td>
<td>7.99e-05</td>
<td>* 0.362</td>
<td>0.362</td>
<td></td>
</tr>
</tbody>
</table>

**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
```

```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

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

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

goggles.summary %>%
ggplot(aes(x=alcohol, y= mean, colour=gender, group=gender)) +
  geom_line() +
  geom_point()
```
Association between 2 continuous variables
One variable X and One variable Y
One predictor
Correlation
Signal-to-noise ratio

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

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

\[
\frac{\text{Signal}}{\text{Noise}} = \text{no statistical significance}
\]
Signal-to-noise ratio and Correlation

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

- Signal is **similarity** of behaviour between variable x and variable y.

- **Coefficient of correlation**: \( r = \frac{\text{similarity}}{\text{variability}} = \frac{\text{Signal}}{\text{Noise}} \)

\[
r = \frac{\text{COV}_{xy}}{SD_x SD_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n - 1) SD_x SD_y}
\]
Correlation

- Most widely-used correlation coefficient:
  - **Pearson product-moment correlation coefficient “r”**
    - The **magnitude** and the **direction** of the relation between 2 variables
    - It is designed to range in value between **-1** and **+1**
    - **-0.6 < r > +0.6 : exciting**

- **Coefficient of determination “r^2”**
  - It gives the proportion of variance in Y that can be explained by X (in percentage).
    - It helps with the interpretation of r
    - It’s basically the **effect size**

<table>
<thead>
<tr>
<th>Coefficient (+ve or –ve)</th>
<th>Strength of the relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0 to 0.2</td>
<td>Negligible</td>
</tr>
<tr>
<td>0.2 to 0.4</td>
<td>Weak</td>
</tr>
<tr>
<td>0.4 to 0.7</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.7 to 0.9</td>
<td>Strong</td>
</tr>
<tr>
<td>0.9 to 1.0</td>
<td>Very strong</td>
</tr>
</tbody>
</table>
Correlation

$p = 0.0002$

$r = -0.34$

$r^2 = 0.12$

$p = 0.04$

$r = -0.83$

$r^2 = 0.68$

Power!!
Correlation
Assumptions

- **Assumptions for correlation**
  - Regression and linear Model (lm)

- **Linearity**: The relationship between X and the mean of Y is linear.

- **Homoscedasticity**: The variance of residual is the same for any value of X.

- **Independence**: Observations are independent of each other.

- **Normality**: For any fixed value of X, Y is normally distributed.
Outliers: the observed value for the point is very different from that predicted by the regression model.

Correlation
Outliers and High leverage points
Correlation
Outliers and High leverage points

• **Leverage points**: A leverage point is defined as an observation that has a value of x that is far away from the mean of x.

• Outliers and leverage points have the potential to be **Influential observations**:
  – Change the slope of the line. Thus, have a large influence on the fit of the model.

• One method to find influential points is to compare the fit of the model **with** and **without** the dodgy observation.
Correlation
Outliers and High leverage points

All good
Correlation
Outliers and High leverage points

Outlier but not influential value
Correlation
Outliers and High leverage points

High leverage but not influential value
Correlation
Outliers and High leverage points

Outlier and High leverage: Influential value
Correlation: Two more things

Thing 1: Pearson correlation is a parametric test

First assumption for parametric test: Normality

Correlation: bivariate Gaussian distribution

Symmetry-ish of the values on either side of the line of best fit.
**Correlation: Two more things**

**Thing 2: Line of best fit comes from a regression**

**Correlation:** nature and strength of the association  
**Regression:** nature and strength of the association and prediction

\[
Y = A + B \times X
\]
Correlation: correlation.csv

- Questions:
  - What is the nature and the strength of the relationship between X and Y?
  - Are there any dodgy points?
**Question:** are there any dodgy points?

```r
read_csv("correlation.csv") -> correlation

correlation %>%
ggplot(aes(variable.x, variable.y, colour=Gender)) +
geom_point(size=3, colour="sienna2")
```
For the lines of best-fit: 3 new functions:

\[
\text{lm}(y \sim x, \text{data} =) \rightarrow \text{fit} \\
\text{coefficients} (\text{fit}) \rightarrow \text{cf.fit} \text{ (vector of 2 values)} \\
\text{geom_abline} (\text{intercept=cf.fit[1]}, \text{ slope=cf.fit[2]})
\]

\[
\text{lm} (\text{variable.y} \sim \text{variable.x}, \text{data=correlation}) \rightarrow \text{fit.correlation} \\
\text{coefficients} (\text{fit.correlation}) \rightarrow \text{cf.correlation} \\
\text{cf.correlation}
\]

\[
(\text{Intercept}) \quad \text{variable.x} \\
8.379803 \quad 3.588814 \\
\text{intercept} \quad \text{slope}
\]
Correlation: correlation.csv

correlation %>%
ggplot(aes(variable.x, variable.y, label = ID)) +
geom_point(size=3, colour="sienna2") +
geom_abline(intercept = coef.correlation[1], slope = coef.correlation[2]) +
geom_text(hjust = 0, nudge_x = 0.15, check_overlap = TRUE)
Correlation: correlation.csv

Assumptions, outliers and influential cases

```r
par(mfrow=c(2,2))
plot(fit.correlation)
```

Linearity, homoscedasticity and outlier

- **Residuals vs Fitted**
- **Normal Q-Q**
- **Scale-Location**
- **Theoretical Quantiles**
- **Residuals vs Leverage**

Normality and outlier

- **Cook's distance**

The **Cook's distance** is a combination of each observation's leverage and residual values; the higher the leverage and residuals, the higher the Cook's distance (influential observation).

- It summarizes how much all the values in the regression model change when the ith observation is removed.
- Consensus: cut-off point =1 (0.5).
Correlation: correlation.csv

correlation %>%
cor_test(variable.x, variable.y)

Y = 8.38 + 3.59*X

summary(fit.correlation)

call:
  lm(formula = variable.y ~ variable.x, data = correlation)

Residuals:
   Min     1Q Median     3Q    Max
-40.034 -3.414  0.867  5.723 17.265

Coefficients:
             Estimate   Std. Error t value Pr(>|t|)  
(Intercept)    8.3798      4.1195  2.034   0.0548 .  
variable.x     3.5888      0.6225  5.765  1.01e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.93 on 21 degrees of freedom
Multiple R-squared:  0.6128,  Adjusted R-squared:  0.5943
F-statistic: 33.23 on 1 and 21 DF,  p-value: 1.01e-05

<table>
<thead>
<tr>
<th>var1</th>
<th>var2</th>
<th>cor</th>
<th>statistic</th>
<th>p</th>
<th>conf.low</th>
<th>conf.high</th>
<th>method</th>
</tr>
</thead>
<tbody>
<tr>
<td>variable.x</td>
<td>variable.y</td>
<td>0.78</td>
<td>5.764871</td>
<td>1.01e-05</td>
<td>0.5471597</td>
<td>0.9034793</td>
<td>Pearson</td>
</tr>
</tbody>
</table>
Correlation: correlation.csv

Have a go: Remove ID 23, then re-run the model and plot the graph again. Then decide what you want to do with ID 21 and 22.
Correlation: correlation.csv

correlation %>%
  filter(ID != 23) -> correlation.23

lm(variable.y ~ variable.x, correlation.23) -> fit.correlation.23
summary(fit.correlation.23)

Call:
  lm(formula = variable.y ~ variable.x, data = correlation.23)

Residuals:
  Min     1Q   Median     3Q    Max
-5.049  -2.784  -1.446  1.679  16.915

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.7103     1.8338   2.023   0.0566 .
variable.x   4.8436     0.2971  16.303   5.13e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.695 on 20 degrees of freedom
Multiple R-squared:  0.93,  Adjusted R-squared:  0.9265
F-statistic: 265.8 on 1 and 20 DF,  p-value: 5.13e-13
Correlation: correlation.csv

correlation.23 %>%
  filter(ID != 21) -> correlation.23.21

lm(variable.y ~ variable.x, correlation.23.21) -> fit.correlation.23.21
summary(fit.correlation.23.21)

Call:
  lm(formula = variable.y ~ variable.x, data = correlation.23.21)

Residuals:
   Min     1Q   Median     3Q    Max
-4.3636 -1.8607  -0.5376  2.2987  5.0434

Coefficients:
                                          Estimate Std. Error   t value     Pr(>|t|)
(Intercept)                   2.4679      1.0757    2.294  0.0333 *
variable.x                    4.9272      0.1719   28.661  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.709 on 19 degrees of freedom
Multiple R-squared:  0.9774
Adjusted R-squared:  0.9762

Method: least.squares
   F-statistic: 821.4 on 1 and 19 DF,  p-value: < 2.2e-16
Extra exercise

Correlation: exam.anxiety.csv

**Question**: Is there a relationship between time spent revising and exam anxiety? And, if yes, are boys and girls different?

- Build a fit for the boys and a fit for the girls
  - `data %>% filter() lm(y~x, data=)`

- Plot the 2 lines of best fit on the same graph
  - `coefficients()` `geom_abline()`

- Check the assumptions visually from the data and with the output for models
  - `par(mfrow=c(2,2)) plot(fit.male)`

- Filter out misbehaving values
  - `rstandard()` `add_column()`

- Plot the final (improved!) model
  - `bind_rows()`
Correlation: exam.anxiety.csv

**Question**: Is there a relationship between time spent revising and exam anxiety? And, if yes, are boys and girls different?

```r
read_csv("exam.anxiety.csv") -> exam.anxiety

exam.anxiety %>%
  ggplot(aes(x=Revise, y=Anxiety, colour=Gender)) + geom_point(size=3)
```
Correlation: exam anxiety.csv

- Is there a relationship between time spent revising and exam anxiety?

```r
exam.anxiety %>%
  filter(Gender=='Female') -> exam.anxiety.female
lm(Angiety~Revise, data=exam.anxiety.female) -> fit.female
coefficients(fit.female) -> cf.fit.female

exam.anxiety %>%
  filter(Gender=='Male') -> exam.anxiety.male
lm(Angiety~Revise, data=exam.anxiety.male) -> fit.male
coefficients(fit.male) -> cf.fit.male
```

Fit for the females

Fit for the males
Correlation: exam anxiety.csv

- Is there a relationship between time spent revising and exam anxiety?

```r
exam.anxiety %>%
  ggplot(aes(x=Revise, y=Anxiety, colour=Gender)) +
  geom_point(size=3) +
  geom_abline(intercept=cf.fit.male[1], slope=cf.fit.male[2]) +
  geom_abline(intercept=cf.fit.female[1], slope=cf.fit.female[2])
```
Correlation: exam anxiety.csv
Assumptions, outliers and influential cases

par(mfrow=c(2,2))
plot(fit.male)
Correlation: exam anxiety.csv
Assumptions, outliers and influential cases

plot(fit.female)
Correlation: exam anxiety.csv

```r
exam.anxiety %>%
group_by(Gender) %>%
cor_test(Revise, Anxiety) %>%
ungroup()
```

**Gender**
- Female: Revise, Anxiety
- Male: Revise, Anxiety

**Correlation Results**

- **Female**
  - Correlation: -0.82, p-value: 1.54e-13
  - Method: Pearson

- **Male**
  - Correlation: -0.60, p-value: 2.94e-06
  - Method: Pearson

**Summary of Models**

- **Male Model**
  - Call: lm(formula = Anxiety ~ Revise, data = exam.anxiety.male)
  - Residuals:
    - Min: -73.124, 1Q: -2.900, Median: 2.221, 3Q: 6.730, Max: 16.600
  - Coefficients:
    - Estimate: 84.1941, Std. Error: 2.6213, t value: 32.119, Pr(>|t|): 3e-16
    - Coef. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  - Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  - Residual standard error: 12.3 on 50 degrees of freedom
  - Multiple R-squared: 0.3568, Adjusted R-squared: 0.344
  - F-statistic: 27.74 on 1 and 50 DF, p-value: 2.937e-06

- **Female Model**
  - Call: lm(formula = Anxiety ~ Revise, data = exam.anxiety.female)
  - Residuals:
    - Min: -22.887, 1Q: -6.263, Median: -1.204, 3Q: 4.197, Max: 38.628
  - Coefficients:
    - Estimate: 91.94181, Std. Error: 2.27858, t value: 40.35, Pr(>|t|): 2e-16
    - Coef. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  - Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  - Residual standard error: 10.42 on 49 degrees of freedom
  - Multiple R-squared: 0.6746, Adjusted R-squared: 0.668
  - F-statistic: 101.6 on 1 and 49 DF, p-value: 1.544e-13

**Equations**

- **Male**
  - Anxiety = 84.19 - 0.53*Revise

- **Female**
  - Anxiety = 91.94 - 0.82*Revise
Correlation: exam.anxiety.csv

Influential outliers: Boys

```r
rstandard(fit.male) -> st.resid.m

eexam.anxiety.male %>%
  add_column(st.resid.m) %>%
  filter(abs(st.resid.m)<3) -> exam.anxiety.male.clean

lm(Anxiety~Revise, data=exam.anxiety.male.clean) -> fit.male2

summary(fit.male2)
```

Call:
lm(formula = Anxiety ~ Revise, data = exam.anxiety.male.clean)

Residuals:
Min  1Q Median  3Q Max
-22.0296 -3.8704  0.5626  6.0786 14.2525

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  86.97461    1.64755  52.790  < 2e-16 ***
Revise       -0.60752    0.06326  -9.603 7.59e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.213 on 49 degrees of freedom
Multiple R-squared:  0.653,  Adjusted R-squared:  0.6459
F-statistic: 92.22 on 1 and 49 DF,  p-value: 7.59e-13

```r
cor_test(Revise, Anxiety)
```

<table>
<thead>
<tr>
<th>var1</th>
<th>var2</th>
<th>cor</th>
<th>statistic</th>
<th>p</th>
<th>conf.low</th>
<th>conf.high</th>
</tr>
</thead>
<tbody>
<tr>
<td>Revise</td>
<td>Anxiety</td>
<td>-0.81</td>
<td>-9.602995</td>
<td>7.59e-13</td>
<td>-0.8863013</td>
<td>-0.6850763</td>
</tr>
</tbody>
</table>
Correlation: exam.anxiety.csv

Influential outliers: Girls

\[
\text{rstandard}(\text{fit.female}) \rightarrow \text{st.resid.f}
\]

\[
\text{exam.anxiety.female} \%>\%
\]
\[
\text{add_column}(\text{st.resid.f}) \%>\%
\]
\[
\text{filter(abs(st.resid.f) < 3)} \rightarrow \text{exam.anxiety.female.clean}
\]

\[
\text{lm(Antiety~Revise, data=exam.anxiety.female.clean)} \rightarrow \text{fit.female2}
\]

\[
\text{summary(fit.female2)}
\]

Call:
\[
\text{lm(formula = Anxiety ~ Revise, data = exam.anxiety.female.clean)}
\]

Residuals:
\[
\begin{array}{c}
\text{Min} \\
\text{1Q} \\
\text{Median} \\
\text{3Q} \\
\text{Max}
\end{array}
\]
\[
\begin{array}{c}
-18.7518 \\
-5.7069 \\
-0.7782 \\
3.2117 \\
18.5538
\end{array}
\]

Coefficients:
\[
\begin{array}{ccc}
\text{Estimate} & \text{Std. Error} & \text{t value} & \text{Pr(>|t|)} \\
(\text{Intercept}) & 92.24536 & 1.93591 & 47.65 & <2e-16 *** \\
\text{Revise} & -0.87504 & 0.07033 & -12.44 & <2e-16 *** \\
\end{array}
\]

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.849 on 48 degrees of freedom
Multiple R-squared: 0.7584, Adjusted R-squared: 0.7584
F-statistic: 154.8 on 1 and 48 DF, p-value: < 2.2e-16

\[
\text{exam.anxiety.female.clean} \%>\%
\]
\[
\text{cor_test(Revise, Anxiety)}
\]
Correlation: exam.anxiety.csv

**Question:** Is there a relationship between time spent revising and exam anxiety? Yes!

```r
bind_rows(exam.anxiety.female.clean, exam.anxiety.male.clean) -> exam.anxiety.clean
coefficients(fit.male2) -> cf.fit.male2
coefficients(fit.female2) -> cf.fit.female2
exam.anxiety.clean %>%
  ggplot(aes(Revise, Anxiety, colour=Gender)) + geom_point(size=3) +
  geom_abline(aes(intercept=cf.fit.male2[1], slope=cf.fit.male2[2]), colour="orange") +
  geom_abline(aes(intercept=cf.fit.female2[1], slope=cf.fit.female2[2]), colour="purple") +
  scale_colour_manual(values = c("purple", "orange"))
```
**Correlation: exam.anxiety**

**Influential outliers: Another check**

<table>
<thead>
<tr>
<th>variable</th>
<th>statistic</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>exam.anxiety.male</td>
<td>0.6992772</td>
<td>5.05199e-09</td>
</tr>
<tr>
<td>exam.anxiety.male.clean</td>
<td>0.9539309</td>
<td>0.04607996</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>variable</th>
<th>statistic</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>exam.anxiety.female</td>
<td>0.9442729</td>
<td>0.01828732</td>
</tr>
<tr>
<td>exam.anxiety.female.clean</td>
<td>0.9767888</td>
<td>0.4258592</td>
</tr>
</tbody>
</table>