Days 1-3
Analysis of Quantitative data

Anne Segonds-Pichon
v2020-02
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

• Non-parametric tests

• R packages needed:
  • tidyverse, dunn.test, plotrix, rstatix and Hmisc
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 \text{ True (no effect)})</td>
</tr>
<tr>
<td>Reject (H_0)</td>
<td>Type I error (\alpha) False Positive</td>
</tr>
<tr>
<td>Do not reject (H_0)</td>
<td>Correct 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: \(\alpha \leq 0.05\)

- **High specificity** = low *False Positives* = low *Type I error*
- **High sensitivity** = low *False Negatives* = low *Type II error*
Statistical inference

Sample → Difference → Meaningful? → Yes → Real? → Statistical test → Statistic e.g. t, F → Big enough? → Sample + Noise +
Stats are all about understanding and controlling variation.

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

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

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.

1) **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

  ![Normal Distribution](image)

  PARANORMAL DISTRIBUTION

  • **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 Examples](image)

    More peaked distribution: kurtosis > 0
    Flatter distribution: kurtosis < 0

Assumptions of Parametric Data

2) **Homogeneity in variance**
   - The variance should not change systematically throughout the data

3) **Interval data (linearity)**
   - The distance between points of the scale should be equal at all parts along the scale.

4) **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
Analysis of Quantitative Data

• 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

• 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
Student's $t$-test
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
SE gap ~ 2 n=3

SE gap ~ 4.5 n=3

SE gap ~ 1 n>=10

SE gap ~ 2 n>=10

~ 2 x SE: p~0.05

~ 4.5 x SE: p~0.01

~ 1 x SE: p~0.05

~ 2 x SE: p~0.01
CI overlap ~ 1 n=3

CI overlap ~ 0.5 n=3

CI overlap ~ 0.5 n>=10

CI overlap ~ 0 n>=10

~ 1 x CI: p~0.05
~ 0.5 x CI: p~0.01
~ 0.5 x CI: p~0.05
~ 0 x CI: p~0.01
Student’s $t$-test

• **3 types:**
  – **Independent t-test**
    • compares means for two independent groups of cases.
  
  – **Paired t-test**
    • looks at the difference between two variables for a single group:
      – the second ‘sample’ of values comes from the same subjects (mouse, petri dish ...).

  – **One-Sample t-test**
    • tests whether the mean of a single variable differs from 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 with a similar variability in the female sample

  • **smallest biologically meaningful difference**

```r
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: 92 cm +/- 7 cm (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 92 cm)

delta = 92 – 87.4
sd = 7

We need a sample size of n ~ 76 (2*38)
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, beanplot and stripchart
  # beanplot package #

Data exploration ≠ plotting data
Exercise 4: Data exploration

- Explore data using 4 different representations:
Exercise 4: `facet_grid(raw ~ column)`

2 columns: one per gender

One row
Exercise 4: `geom_jitter()`

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

```r
ggplot(coyote, aes(gender, length)) + geom_point()
```

```r
= ggplot(coyote, aes(gender, length)) + geom_jitter(position = "jitter")
```

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

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

\[
\text{stat_summary(} \text{fun.y="mean"}, \text{fun.ymin="mean"}, \text{fun.ymax="mean"}, \text{geom= "errorbar"})
\]

- mean=minimum=max

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

- Explore data using 4 different representations:
  - ```geom_jitter()```
  - ```stat_summary()```
  - ```facet_grid(row~column)```
Exercise 4: Exploring data - Answers
Boxplots and beanplots

```r
ggplot(coyote, aes(gender, length)) + geom_boxplot()
```

```r
ggplot(coyote, aes(gender, length)) + geom_violin()
```
Exercise 4: Exploring data - Answers
Boxplots and beanplots

```
ggplot(coyote, aes(gender, length, fill=gender)) +
ylab("Length (cm)") +
scale_fill_manual(values = c("orange","purple")) +
theme(legend.position = "none") +
stat_boxplot(geom="errorbar", width=0.5) +
theme(axis.title.x = element_blank()) +
geom_boxplot()
```

```
ggplot(coyote, aes(gender, length, fill=gender)) +
geom_violin(trim=FALSE, size=1) +
ylab("Length (cm)") +
scale_fill_manual(values = c("orange","purple")) +
theme(legend.position = "none") +
stat_summary(geom = "point", fun.y = "median")
```
Exercise 4: Exploring data - Answers

Histograms

ggplot(coyote, aes(length)) +
  geom_histogram(binwidth = 4, colour="black") +
  facet_grid(~gender)  facet_wrap() also works
**Exercise 4: Exploring data - Answers**

**Histograms**

```r
ggplot(coyote, aes(length, fill=gender)) + geom_histogram(binwidth = 4.5, colour="black") + scale_fill_manual(values = c("orange","purple")) + theme(legend.position = "none") + facet_grid(~gender)
```
Exercise 4: Exploring data - Stripchart

```r
ggplot(coyote, aes(gender, length)) +
  geom_jitter(height=0, width=0.2) +
  stat_summary(fun.y="mean", fun.ymin="mean", fun.ymax="mean", geom="errorbar", width=0.6)
```

```r
ggplot(coyote, aes(gender, length, colour=gender)) +
  geom_jitter(height=0, size=4, width=0.2) +
  theme(legend.position = "none") +
  ylab("Length (cm)") +
  scale_color_manual(values=c("darkorange","purple")) +
  theme(axis.title.x = element_blank()) +
  stat_summary(fun.y=mean, fun.ymin=mean, fun.ymax=mean, geom="errorbar", colour="black", size=1.2, width=0.6)
```
Exercise 4 extra: Exploring data - Graph combinations

ggplot(coyote, aes(gender, length)) +
geom_violin() +
geom_boxplot(width=0.2)

ggplot(coyote, aes(gender, length)) +
geom_boxplot() +
geom_jitter(height=0, width=0.2)
Exercise 4 extra: Exploring data - Graph combinations

```r
# Plot 1
ggplot(coyote, aes(gender, length)) +
geom_boxplot() +
geom_jitter(height=0, width=0.2)
```

```r
# Plot 2
ggplot(coyote, aes(gender, length)) +
geom_boxplot() +
stat_boxplot(geom="errorbar", width=0.2) +
geom_boxplot(outlier.shape=8, outlier.size = 5) +
geom_jitter(height=0, width=0.1, size = 2, alpha = 0.5, 
colour="red") +
ylab("Length (cm)")
```
Assumptions of Parametric Data

- First assumption: **Normality**
  - Shapiro-Wilk test `shapiro_test()` # rstatix package#

- Second assumption: **Homoscedasticity**
  - Bartlett test `bartlett.test()`

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

<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>

```
bartlett.test(coyote$length ~ coyote$gender)
```

Bartlett test of homogeneity of variances
data:  coyote$length by coyote$gender
Bartlett's K-squared = 0.02021, df = 1, p-value = 0.887
Independent *t*-test: results

**coyote.csv**

t.test(coyote$length~coyote$gender, var.equal=TRUE)

- **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?**
Plot ‘coyote.csv’ data: Plotting data

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

- Add error bars

```
ggplot(coyote, aes(gender, length, colour=gender)) +
  geom_jitter(height=0, width=0.1) +
  geom_bar(stat = "summary", fun.y="mean", width=0.4, alpha=0, colour="black") +
  stat_summary(geom="errorbar", colour="black", width=0.2)
```
Plot ‘coyote.csv’ data: Plotting data

- Prettier version

```r
ggplot(coyote, aes(gender, length, colour=gender)) +
  geom_bar(stat = "summary", fun.y="mean", width=0.4, alpha=0.2, colour="black", fill =c("darkorange","purple")) +
  stat_summary(geom="errorbar", colour="black", width=0.2) +
  geom_jitter(height=0, width=0.1) +
  scale_color_manual(values=c("darkorange","purple")) +
  ylab("Length (cm)") +
  theme(legend.position = "none") +
  theme(axis.title.x = element_blank())
```

Another example of t-test: working.memory.csv
Exercise 5: Dependent or Paired $t$-test

working.memory.csv

- A researcher is studying the effects of dopaminedepletion on working memory in rhesus monkeys.
- **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.
Exercise 5: Dependent or Paired $t$-test - Answers

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

ggplot(working.memory, aes("", difference)) +
  geom_jitter(height=0, width=0.1, size=4, colour="chartreuse3") +
  stat_summary(fun.y="mean", fun.ymin="mean", fun.ymax="mean", geom="errorbar", 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))
Exercise 5: Dependent or Paired t-test - Answers

Question: does dopamine affect working memory in rhesus monkeys?

Answer: the injection of a dopamine-depleting agent significantly affects working memory in rhesus monkeys (t=8.62, df=14, p=5.715e-7).

```r
> shapiro.test(working.memory$Difference)
Shapiro-wilk normality test
data:  working.memory$Difference
W = 0.97727, p-value = 0.9474

> t.test(working.memory$placebo, working.memory$DA.depletion, paired=T)

Paired t-test
data:  working.memory$Placebo and working.memory$DA.depletion
t = 8.6161, df = 14, p-value = 5.715e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
6.308997 10.491003
sample estimates:
mean of the differences
8.4
```
Comparison between more than 2 groups

One factor
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%** ($1 - (0.95)^5$)
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.
Analysis of variance

• Extension of the 2 groups comparison of a $t$-test but with a slightly different logic:

  • $t$-test = mean1 – mean2
  
  Pooled SEM

  • ANOVA = variance between means
  
  Pooled SEM

• ANOVA compares variances:
  – If variance between the several means $>$ variance within the groups (random error) then the means must be more spread out than it would have been by chance.
The statistic for ANOVA is the **F ratio**.

\[
F = \frac{\text{Variance between the groups}}{\text{Variance within the groups (individual variability)}}
\]

\[
F = \frac{\text{Variation explained by the model (= systematic)}}{\text{Variation explained by unsystematic factors (= random variation)}}
\]

If the variance amongst sample means is greater than the error/random variance, then \( F > 1 \)
- In an ANOVA, we test whether \( F \) is significantly higher than 1 or not.
Variance (= SS / N-1) is the mean square
- df: degree of freedom with df = N-1

<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>2.665</td>
<td>4</td>
<td>0.6663</td>
<td>8.423</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Within Groups</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>

In Power Analysis:
Pooled SD=\sqrt{MS(Residual)}
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
- 3 Statistical analysis: ANOVA
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
  - Check the first assumption (Normality) with a formal test (Shapiro-Wilk test)
Exercise 6: One-way ANOVA - Answers

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

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

```r
protein %>%
  group_by(line) %>%
  shapiro_test(expression)
```

<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>0.5295671</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

```r
ggplot(protein, aes(line, log10.expression, colour=line)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(height = 0, width = 0.1)

ggplot(protein, aes(line, log10.expression, colour=line)) +
  geom_violin(trim = FALSE) +
  geom_boxplot(width = 0.1)
```
## Assumptions of Parametric Data

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

<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>0.57725321</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>0.9813425</td>
<td>0.20502703</td>
</tr>
</tbody>
</table>

First assumption ✓

bartlett.test(protein$log10.expression, protein$line)

Bartlett test of homogeneity of variances
data:  protein$log10.expression and protein$line
Bartlett's K-squared = 5.8261, df = 4, p-value = 0.2125

Second assumption ✓
Analysis of variance: Post hoc tests

• The ANOVA is an “omnibus” test (Step 1): it tells you that there is (or not) a difference between your means but not exactly which means are significantly different from which other ones.

  – To find out, you need to apply post hoc tests (Step 2).

  – These post hoc tests should only be used when the ANOVA finds a significant effect.
Analysis of variance

• Step 1: omnibus test:
  • `aov(y~x, data= ) -> model`
  • `summary(model)`

• Step 2: post-hoc tests
  • `pairwise.t.test(y, x, p.adj = "bonf")`
  • `TukeyHSD(model)`

Have a go!
Analysis of variance

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

                     Df  Sum Sq Mean Sq F value  Pr(>F)
line                  4 2.0914  0.5229  8.123 1.78e-03 ***
Residuals             73 6.0460  0.0828

signif. codes:  < 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 1

pairwise.t.test(protein.stack.clean$log10.expression, protein.stack.clean$line, p.adj = "bonf")

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)

```r
$line
diff  lwr   upr     p adj
B-A -0.25024832 -0.57882494  0.07838585 0.218726
C-A -0.07499724 -0.37499720  0.2200333 0.930018
D-A 0.30549397  0.00549339  0.60549565 0.043876
E-A 0.13327517 -0.16672483  0.43327537 0.728957
C-B 0.17525108 -0.12474999  0.47525167 0.480938
D-B 0.55574230  0.25574172  0.85574288 0.000018
E-B 0.38352349  0.08352280  0.68352407 0.005476
D-C 0.38049121  0.11216253  0.6481987 0.001543
E-C 0.20827240 -0.06008276  0.4766108 0.202335
E-D -0.17221881 -0.44054748  0.08610987 0.384198
```
Analysis of variance

TukeyHSD(\texttt{anova.log.protein})

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

$\text{Fit: aov(formula = protein} \log_{10}.\text{expression - protein} \times \text{line})$

$\begin{array}{cccccc}
\text{diff} & \text{lwr} & \text{upr} & p \text{ adj} \\
B-A & -0.23024832 & -0.57882494 & 0.07838585 & 0.2187264 \\
C-A & -0.7499724 & -0.37499782 & 0.22500335 & 0.9580235 \\
D-A & 0.30549397 & 0.00549393 & 0.60549455 & 0.0438763 \\
E-A & 0.13327317 & -0.16672516 & 0.33377575 & 0.7265567 \\
C-B & 0.17552108 & -0.12474949 & 0.47525167 & 0.4809387 \\
D-B & 0.5557423 & 0.25574172 & 0.85574288 & 0.0000183 \\
E-B & 0.38352349 & 0.08352290 & 0.68352407 & 0.0054767 \\
D-C & 0.38049121 & 0.11216252 & 0.64818198 & 0.0015431 \\
E-C & 0.20827240 & -0.06005627 & 0.47660108 & 0.2023355 \\
E-D & -0.17221881 & -0.44054748 & 0.09610387 & 0.3841989 \\
\end{array}$

TukeyHSD(\texttt{anova.log.protein}) \rightarrow \text{tukey plot(\texttt{tukey}, las=1)}
Analysis of variance

ggplot(protein, aes(line, log10.expression, colour=line))+
  geom_jitter(height = 0, width = 0.1, size=2)+
  stat_summary(geom="errorbar", fun.y=mean, fun.ymax = mean, fun.ymin = mean, colour="black", size=1, width=0.5)
Analysis of variance

ggplot(protein, aes(line, expression, fill=line)) +
  geom_bar(stat = "summary", fun.y="mean", colour="black") +
  stat_summary(geom="errorbar", colour="black", width=0.4)
Analysis of variance

ggplot(protein, aes(line,expression, fill=line)) +
  geom_bar(stat = "summary", fun.y="mean",colour="black") +
  stat_summary(geom="errorbar", colour="black", width=0.4) +
  geom_jitter(height=0, width=0.1, alpha=0.5)
Power Analysis for ANOVA

- Different ways to go about it:
  - $\eta^2$: explained proportion variance of the total variance.
    - Can be translated into effect size d.
    - Not very useful: only looking at the omnibus part of the test.
  - Minimum power specification: looks at the difference between the smallest and the biggest means.
    - All means other than the 2 extreme one are equal to the grand mean.
  - Smallest meaningful difference
    - Works like a post-hoc test.
Comparison between more than 2 groups

Two factors
# 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

### Source of variation Summary Table

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Sum of Squares</th>
<th>Df</th>
<th>Mean Square</th>
<th>F (K-n)</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>11.91</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Variable B (Between groups)</td>
<td>3332</td>
<td>2</td>
<td>1666</td>
<td>20.07</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Variable A (Between groups)</td>
<td>168.8</td>
<td>1</td>
<td>168.8</td>
<td>2.032</td>
<td>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

Example: goggles.csv

- The ‘beer-goggle’ effect

<table>
<thead>
<tr>
<th>Alcohol</th>
<th>Gender</th>
<th>None</th>
<th>2 Pints</th>
<th>4 Pints</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>65</td>
<td>50</td>
<td>70</td>
<td>55</td>
<td>45</td>
</tr>
<tr>
<td>70</td>
<td>55</td>
<td>65</td>
<td>65</td>
<td>60</td>
</tr>
<tr>
<td>60</td>
<td>80</td>
<td>60</td>
<td>70</td>
<td>85</td>
</tr>
<tr>
<td>60</td>
<td>65</td>
<td>70</td>
<td>55</td>
<td>65</td>
</tr>
<tr>
<td>60</td>
<td>70</td>
<td>65</td>
<td>55</td>
<td>70</td>
</tr>
<tr>
<td>55</td>
<td>75</td>
<td>60</td>
<td>60</td>
<td>70</td>
</tr>
<tr>
<td>60</td>
<td>75</td>
<td>60</td>
<td>50</td>
<td>80</td>
</tr>
<tr>
<td>55</td>
<td>65</td>
<td>50</td>
<td>50</td>
<td>60</td>
</tr>
</tbody>
</table>

- 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 7: Two-way ANOVA

goggles.csv

• Load goggles.csv

• Graphically explore the data
  • effect of alcohol only
  • effect of gender only
  • effect of both
  • Clue: you can use `facet_grid()` or `facet_wrap()`
Two-way Analysis of Variance

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

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

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

```
ggplot(goggles, aes(alcohol, attractiveness)) +
  geom_boxplot() +
  geom_jitter(height=0, width=0.1) +
  facet_grid(~gender)
```
Two-way Analysis of Variance

```r
ggplot(goggles, aes(alcohol, attractiveness)) +
  geom_boxplot() +
  geom_jitter(height=0, width=0.1) +
  facet_grid(~alcohol)
```
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

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

```r
anova.goggles <- aov(attractiveness ~ alcohol + gender + alcohol * gender, data = goggles)
summary(anova.goggles)

TukeyHSD(anova.goggles)
```

**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

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

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

ggplot(goggles.summary, aes(alcohol, mean, group=gender, colour=gender)) +
  geom_line() +
  geom_point()
```
Two-way Analysis of Variance

Interaction between Alcohol and Gender `interaction.plot()`

goggles$alcohol <- factor(goggles$alcohol, levels = c("None", "2 Pints", "4 Pints"))

`interaction.plot(goggles$alcohol, goggles$gender, goggles$attractiveness)`
Association between 2 continuous variables
Correlation

- A correlation coefficient is an index number that measures:
  - The **magnitude** and the **direction** of the relation between 2 variables
  - It is designed to range in value between -1 and +1
Correlation

• Most widely-used correlation coefficient:
  – Pearson product-moment correlation coefficient “r”

\[ r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \]

• The 2 variables do not have to be measured in the same units but they have to be proportional (meaning linearly related)

– Coefficient of determination:
  • r is the correlation between X and Y
  • \( r^2 \) is the coefficient of determination:
    – It gives you the proportion of variance in Y that can be explained by X, in percentage.
Correlation

• **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.
Correlation

• **Assumptions for correlation**
  – Regression and linear Model (lm)

• **Outliers**: the observed value for the point is very different from that predicted by the regression model.

• **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.

• **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 each observation.

• Bottom line: **influential outliers** are problematic.
Correlation: exam anxiety.csv

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

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

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

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

• For the lines of best-fit: 3 new functions:

```r
lm(y~x, data=) -> fit
coefficients(fit) -> cf.fit (vector of 2 values)
geom_abline(intercept=cf.fit[1], slope=cf.fit[2])
```

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

```r
fit.male
```

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

```r
fit.female
```
• Is there a relationship between time spent revising and exam anxiety?

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

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

Linearity, homoscedasticity and outlier

Normality and outlier

- Homoscedasticity
- Influential cases
Correlation: exam anxiety.csv
Assumptions, outliers and influential cases

plot(fit.female)

Linearity, homoscedasticity and outlier

Homoscedasticity

Normality and outlier

Influential cases
Correlation: exam anxiety.csv

Correlation between Revise and Anxiety for males:

\[
\text{Correlation: exam anxiety.csv}
\]

Correlation between Revise and Anxiety for females:

\[
\text{Correlation: exam anxiety.csv}
\]
Correlation: exam anxiety.2

Influential outliers

```r
exam.anxiety.male %>%
  filter(Code!=78) -> exam.anxiety.male
lm(Anxiety~Revise, data=exam.anxiety.male) -> fit.male
fit.male
summary(fit.male)

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

Coefficients:
(Intercept) Revise
  86.9746  -0.6073

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

Coefficients:
(Intercept) Revise
  92.245  -0.875

exam.anxiety.male %>%
  filter(Code!=78) -> exam.anxiety.male
lm(Anxiety~Revise, data=exam.anxiety.male) -> fit.male
fit.male
summary(fit.male)

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

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: 7.391e-13

exam.anxiety.female %>%
  filter(Code!=78) -> exam.anxiety.female
lm(Anxiety~Revise, data=exam.anxiety.female) -> fit.female
fit.female
summary(fit.female)

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

Residuals:       Min     1Q    Median     3Q    Max
     -18.7518 -5.7069  -0.7782  3.2117 18.5538

Coefficients:     Estimate Std. Error t value Pr(>|t|)
(Intercept)     92.24536    1.93591  47.65  <2e-16 ***
Revise          -0.87504    0.07033  -12.44  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.849 on 48 degrees of freedom
Multiple R-squared: 0.7633, Adjusted R-squared: 0.7584
F-statistic: 154.8 on 1 and 48 DF,  p-value: < 2.2e-16
```
Non-parametric tests
Non-parametric test
Mann-Whitney = Wilcoxon rank test

- Non-parametric equivalent of the t-test.
- **What if the data do not meet the assumptions for parametric tests?**
  - The outcome is a rank or a score with limited amount of possible values: non-parametric approach.

- **How does the Mann-Whitney test work?**

  - **Statistic of the Mann-Whitney test:** \( W (U) \)
    - \( W = \) sum of ranks – mean rank: \( W_1 = 3.5 \) and \( W_2 = 10.5 \)
    - Smallest of the 2 Ws: \( W_1 + \) sample size → **p-value**

  - R: `wilcox.test()`
Non-parametric test: Wilcoxon’s signed-rank

- Non-parametric equivalent of the paired t-test
- How does the test work?

<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>3</td>
<td>-6</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>-3</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>-6</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>-3</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>-6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>-5</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>-5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2.5</td>
</tr>
<tr>
<td>3</td>
<td>2.5</td>
</tr>
<tr>
<td>4</td>
<td>4.5</td>
</tr>
<tr>
<td>5</td>
<td>4.5</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

- Statistic of the Wilcoxon’s signed-rank test: $T (W)$
  - Here: Wilcoxon’s $T = 1$ (smallest of the 2 (absolute value))
  - $N = 9$ (we ignore the 0 difference): $T + N \rightarrow p$-value

- R: `wilcox.test(paired = TRUE)`
Exercise 8: Independent test
smelly.teeshirt.csv

• Hypothesis: Group body odour is less disgusting when associated with an in-group member versus an out-group member. Two groups of Cambridge University students are presented with one of two smelly, worn t-shirts with university logos.

• Question: are Cambridge students more disgusted by worn smelly T-shirts from Oxford or Cambridge? Disgust score: 1 to 7, with 7 the most disgusting

• Load smelly teeshirt.csv
• Explore the data with an appropriate combination of 2 graphs
• Answer the question with a non-parametric approach
Exercise 8: smelly.teeshirt.long.csv

- **Question:** are Cambridge students more disgusted by worn smelly T-shirts from Oxford or Cambridge?
  - Disgust score: 1 to 7, with 7 the most disgusting

```python
read_csv("smelly.teeshirt.long.csv") -> smelly.teeshirt
smelly.teeshirt

ggplot(smelly.teeshirt, aes(university, smell)) +
geom_boxplot() +
geom_jitter(height=0, width=0.1, size=2, colour="red")
```

```r
wilcoxon.test(smelly.teeshirt$smell~smelly.teeshirt$university)
```

**Answer:** T-shirts from Oxford are significantly more disgusting than the ones from Cambridge ($W=5, p=0.0047$).

What do you think of the design??
A group of 9 disabled children with muscle spasticity (or extreme muscle tightness limiting movement) in their right upper limb underwent a course of injections with botulinum toxin to reduce spasticity levels.

- **Question**: do botulinum toxin injections reduce muscle spasticity levels?
  - Score: 1 to 10, with 10 the highest spasticity
  - Load `botulinum.csv`
  - Convert botulinum into a long format
    - Clue: `gather()`
  - Plot the data
  - Answer the question with a non-parametric approach
  - Work out and plot the difference (after – before)
**Exercise 9: Dependent test - botulinum.csv**

```r
botulinum <- read_csv("botulinum.csv")

botulinum %>%
  gather(treatment, score) -> botulinum.long

ggplot(botulinum.long, aes(treatment, score)) +
  geom_boxplot() +
  geom_jitter(height=0, width=0.1)

wilcox.test(botulinum$before, botulinum$after, paired = TRUE)

botulinum$difference <- botulinum$after - botulinum$before

ggplot(botulinum, aes("", difference)) +
  geom_boxplot() +
  geom_jitter(width=0.1) +
  ylim(-7, 0)
```

**Answer:** There was a significant difference pre- and post- treatment in ratings of muscle spasticity (p=0.008).  
*Note: T=V*
Non Parametric approach: Kruskal-Wallis

• Non-parametric equivalent of the one-way ANOVA
• It is a test based on ranks
• `kruskal.wallis()` produces omnibus part of the analysis
• Post-hoc test associated with Kruskal-Wallis: **Dunn test**
• `dunn.test()` gives both Kruskall-Wallis and pairwise comparisons results
• Statistic associated with Kruskal-Wallis is H and it has a Chi$^2$ distribution
• The Dunn test works pretty much like the Mann-Whitney test.
Exercise 10: creatine.csv

- Creatine, a supplement popular among body builders
- Three groups: No creatine; Once a day; and Twice a day.
- **Question**: does the average weight gain depend on the creatine group to which people were assigned?

```r
creatine <- read_csv("creatine.csv")

ggplot(creatine, aes(creatine, gain)) +
  geom_boxplot() +
  geom_jitter(width=0.1)

creatine %>%
  group_by(creatine) %>%
  summarise(sd = sd(gain))
```

<table>
<thead>
<tr>
<th>creatine</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>488.5317</td>
</tr>
<tr>
<td>Once</td>
<td>2005.1585</td>
</tr>
<tr>
<td>Twice</td>
<td>1047.8519</td>
</tr>
</tbody>
</table>
Exercise 10: creatine.csv

dunn.test(creatine$gain, creatine$creatine, kw=TRUE, method="bonferroni", alpha=0.05) ## dunn.test package ##

```
Kruskal-Wallis rank sum test
data: x and group
Kruskal-Wallis chi-squared = 3.8677, df = 2, p-value = 0.14

Comparison of x by group (Bonferroni)

Col Mean | NO | Once
Row Mean |   |   |
---------|---|---|
Once     | -0.160162 | 1.0000
Twice    | -1.784927 | -1.704706
         | 0.1114    | 0.1324
```

**Answer:** this study did not demonstrate any effect from creatine ($\chi^2 = 3.87, p = 0.14$).
Non-Parametric:
Spearman Correlation Coefficient

• Only really useful for ranks (either one or both variables)
• \( \rho \) (rho) is the equivalent of \( r \) and calculated in a similar way

**Example:** dominance.csv

• Six male colobus monkeys ranked for dominance
• Question: is social dominance associated with parasitism?
  • Eggs of *Trichirus* nematode per gram of monkey faeces

```r
read_csv("dominance.csv")
```

```
<table>
<thead>
<tr>
<th>Monkey</th>
<th>Dominance</th>
<th>Eggs per gram</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erroll</td>
<td>1</td>
<td>5777</td>
</tr>
<tr>
<td>Milo</td>
<td>2</td>
<td>4225</td>
</tr>
<tr>
<td>Fraiser</td>
<td>3</td>
<td>2674</td>
</tr>
<tr>
<td>Fergus</td>
<td>4</td>
<td>1249</td>
</tr>
<tr>
<td>Kabul</td>
<td>5</td>
<td>749</td>
</tr>
<tr>
<td>Hope</td>
<td>6</td>
<td>870</td>
</tr>
</tbody>
</table>
```

```r
ggplot(dominance, aes(dominance, eggs.per.gram)) +
geom_col(fill="Magenta", colour="black", size=1) +
scale_x_continuous(breaks=seq(1:6)) +
scale_y_continuous(breaks = seq(0, 6000, 1000))
```
Non-Parametric: Spearman Correlation Coefficient

• **Example**: dominance.csv

```r
cor.test(dominance$dominance, dominance$eggs.per.gram, method = "spearman")
```

```
Spearman's rank correlation rho

data:  dominance$dominance and dominance$eggs.per.gram
S = 68, p-value = 0.01667
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
-0.9428571
```

• **Answer**: the relationship between dominance and parasitism is significant ($\rho = -0.94$, p=0.017) with high ranking males harbouring a heavier burden.