Babraham Institute

Analysis of Quantitative data

Anne Segonds-Pichon

v2020-09

Babraham Bioinformatics

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

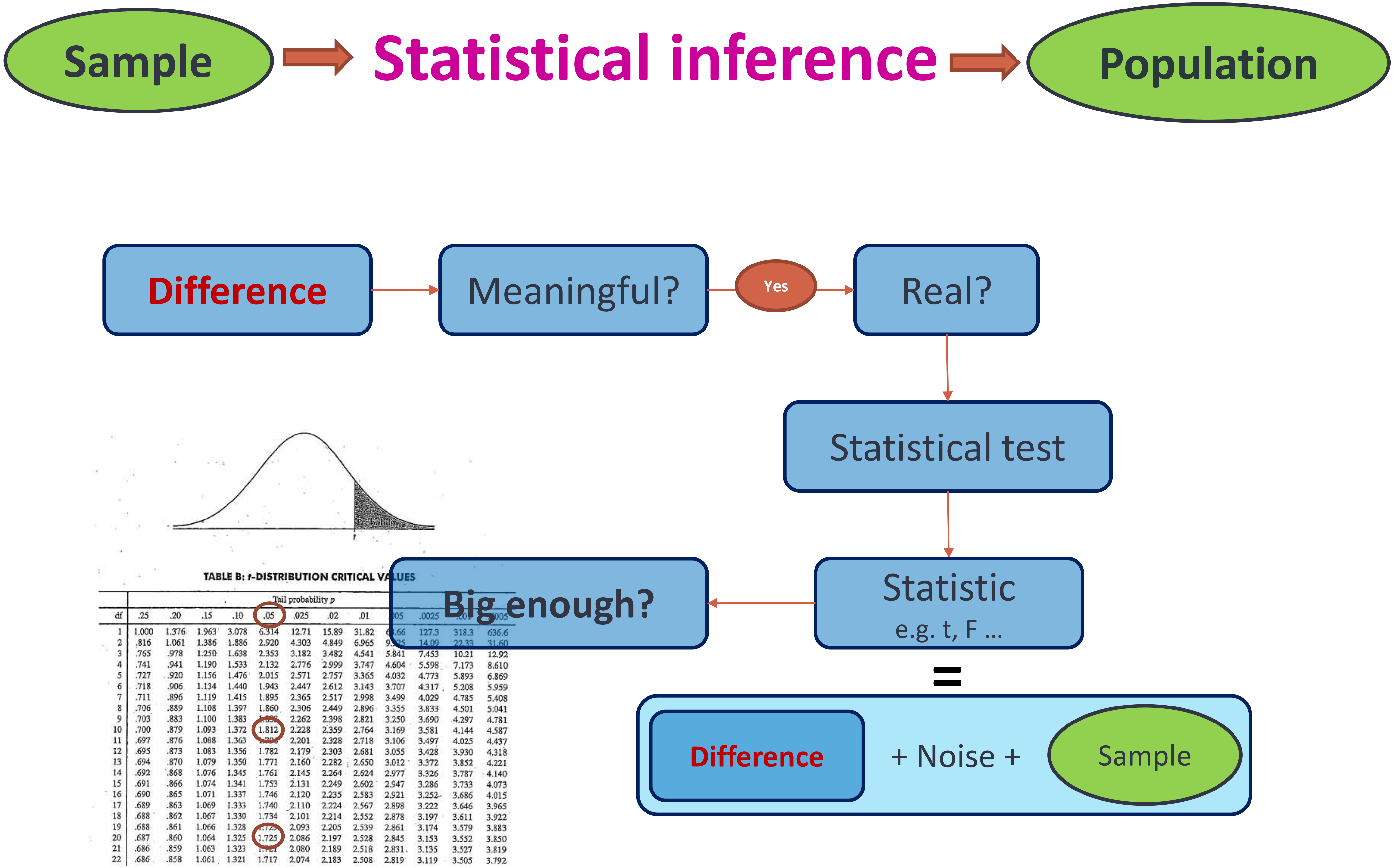
- **Key concepts to always keep in mind**
  - Null hypothesis and error types
  - Statistics inference
  - Signal-to-noise ratio

The null hypothesis and the error types

- The null hypothesis (H_0): H_0 = no effect
  - e.g. no difference between 2 genotypes
- The aim of a statistical test is to reject or not $H_{0.}$

Statistical decision	True state of H_0	
	H_0 True (no effect)	**H_0 False (effect)**
Reject H_0	**Type I error α** **False Positive**	Correct **True Positive**
Do not reject H_0	Correct **True Negative**	**Type II error β** **False Negative**

- Traditionally, a test or a difference is said to be "**significant**" if the probability of type I error is: **α =< 0.05**
- **High specificity** = low **False Positives** = low **Type I error**
- **High sensitivity** = low **False Negatives** = low **Type II error**



Sample → Statistical inference → Population

Difference → Meaningful? → Yes → Real? → Statistical test → Statistic e.g. t, F … → Big enough?

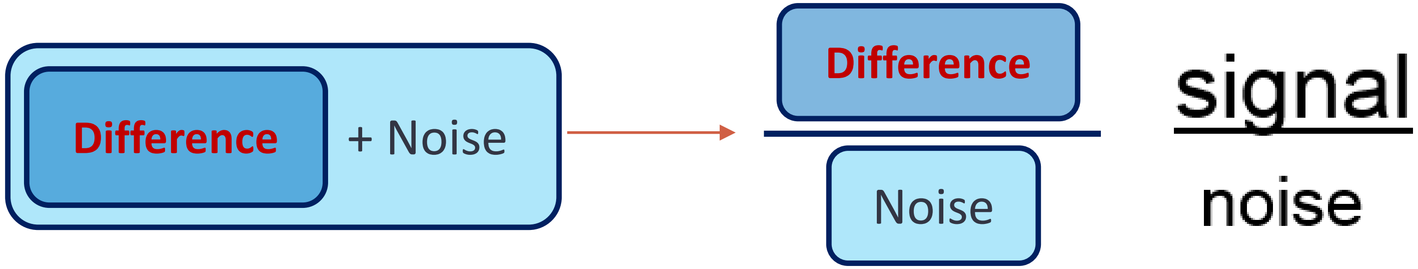
Statistic = Difference + Noise + Sample

TABLE B: *t*-DISTRIBUTION CRITICAL VALUES

	Tail probability *p*											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792

Signal-to-noise ratio

- Stats are all about understanding and controlling variation.



$$\frac{\text{signal}}{\text{noise}}$$ If the **noise is low** then the **signal is detectable** …

= statistical significance

$$\frac{\text{signal}}{\textbf{noise}}$$ … but if the **noise** (i.e. interindividual variation) **is large** then the **same signal will not be detected**

= no statistical significance

- In a statistical test, the ratio of signal to noise determines the significance.

Analysis of Quantitative Data

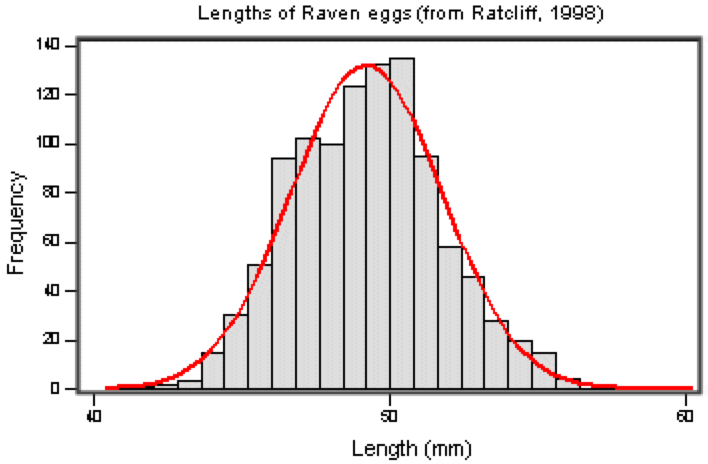
- Choose the correct statistical test to answer your question:
  - They are 2 types of statistical tests:
    - **Parametric tests** with 4 assumptions to be met by the data,
    - **Non-parametric tests** with no or few assumptions (e.g. Mann-Whitney test) and/or for qualitative data (e.g. Fisher’s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

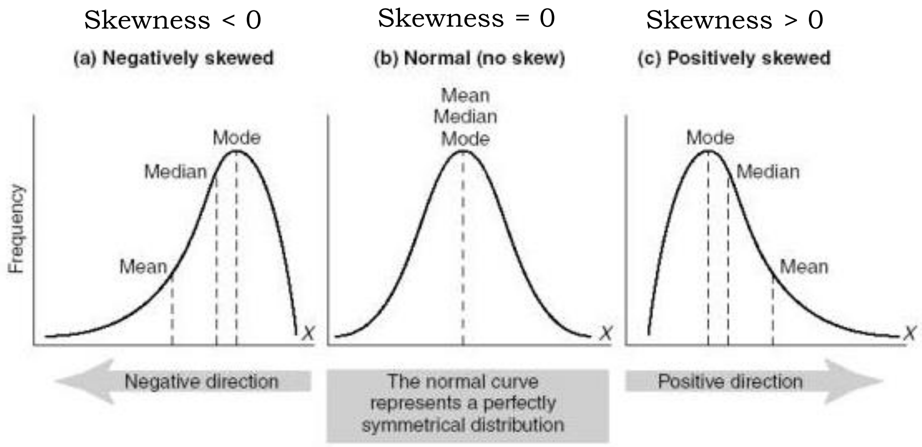
– Normal shape, bell shape, Gaussian shape

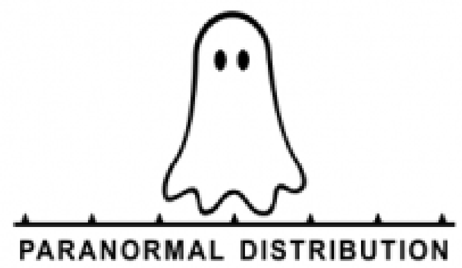

Lengths of Raven eggs (from Ratcliff, 1998)

- Transformations can be made to make data suitable for parametric analysis.

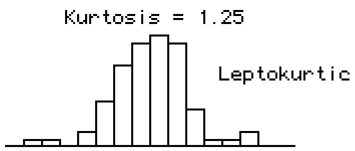
Assumptions of Parametric Data

- Frequent departures from normality:
  - Skewness: lack of symmetry of a distribution

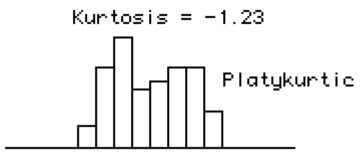




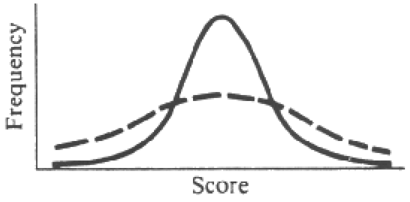
  - Kurtosis: measure of the degree of 'peakedness' in the distribution
    - The two distributions below have the same variance approximately the same skew, but differ markedly in kurtosis.



More peaked distribution: kurtosis > 0



Flatter distribution: kurtosis < 0



(e) Platykurtic and leptokurtic

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

Analysis of Quantitative Data

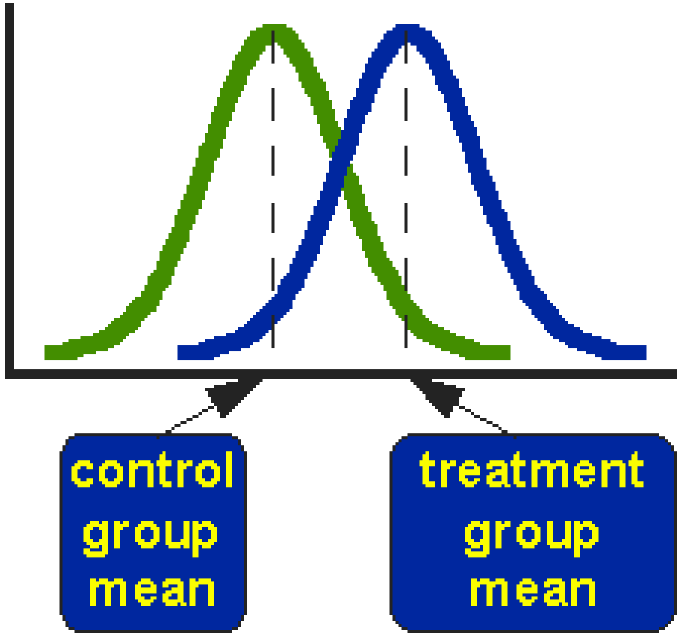
- **Is there a difference between my groups regarding the variable I am measuring?**
  - e.g. are the mice in the group A heavier than those in group B?
    - Tests with 2 groups:
      - Parametric: **Student's *t*-test**
      - Non parametric: **Mann-Whitney/Wilcoxon rank sum test**
    - Tests with more than 2 groups:
      - Parametric: **Analysis of variance (one-way and two-way ANOVA)**
      - Non parametric: **Kruskal Wallis (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

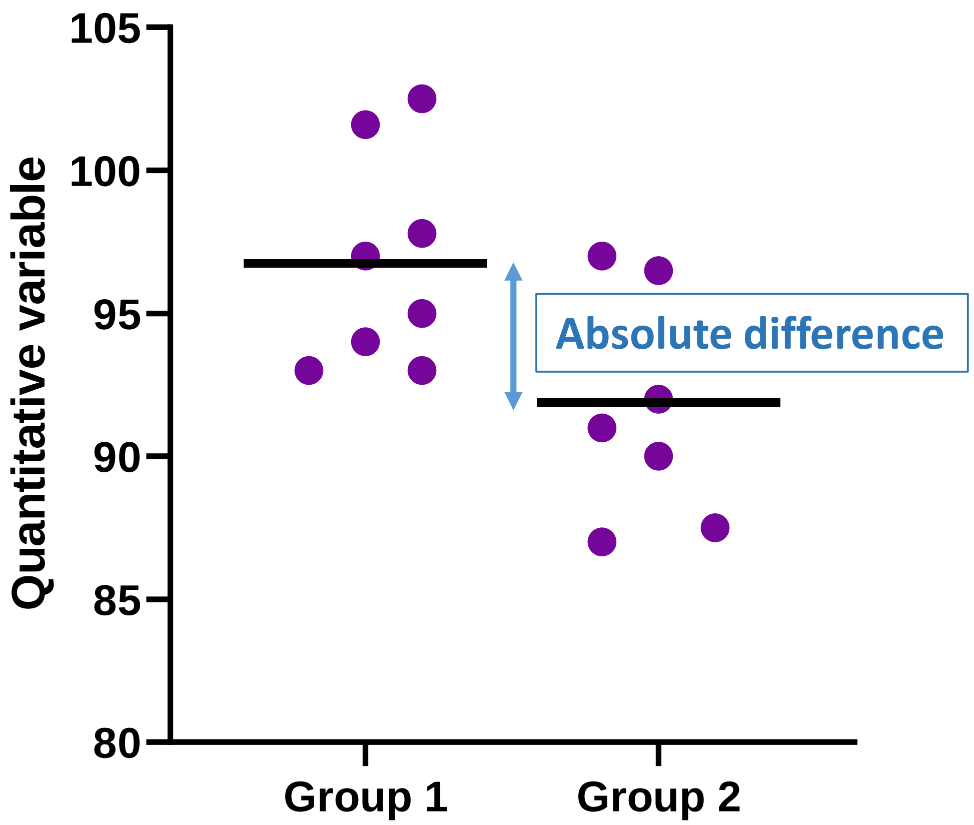
Comparison between 2 groups:

Student's *t*-test

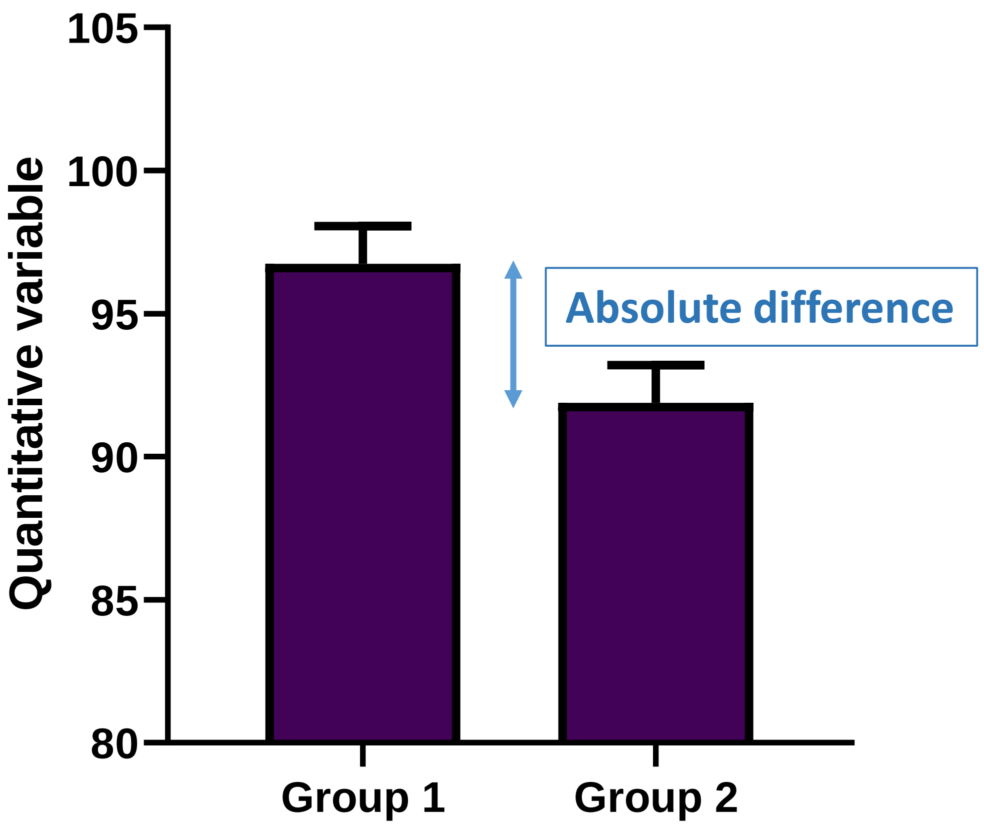
- **Basic idea**:
  - When we are looking at the differences between scores for 2 groups, we have to judge the difference between their means relative to the spread or variability of their scores.
    - Eg: comparison of 2 groups: control and treatment



Variability does matter

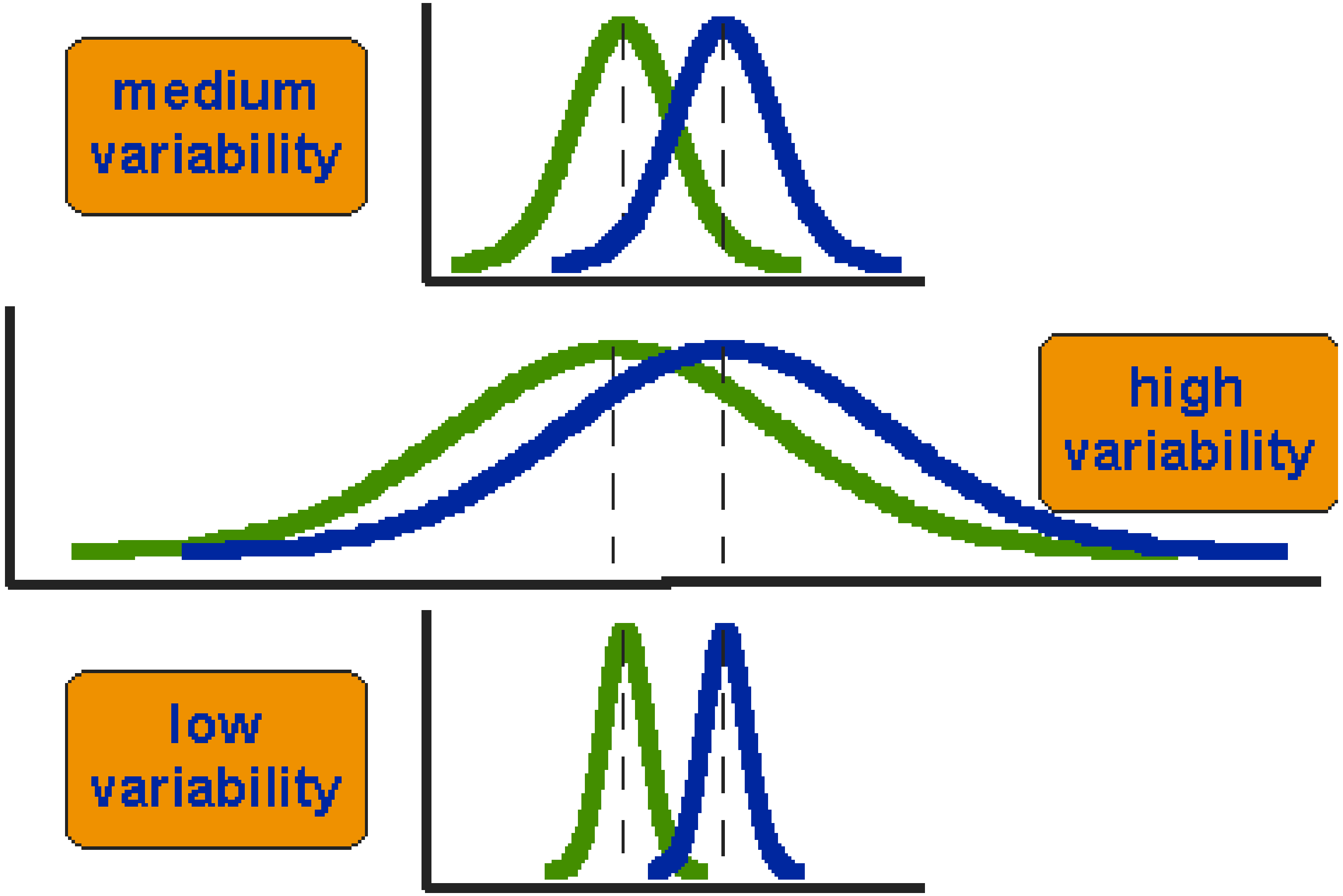


Scatter plot 🙂



Bar chart ☹

Student's *t*-test

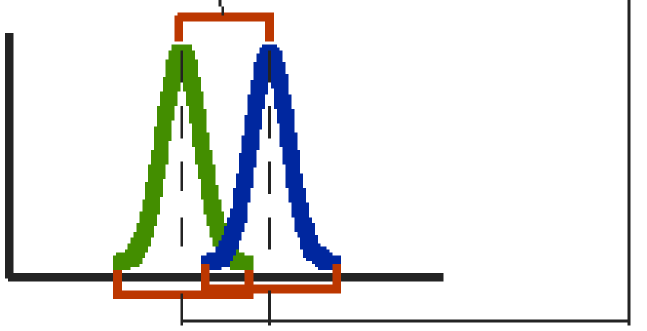


Student's *t*-test

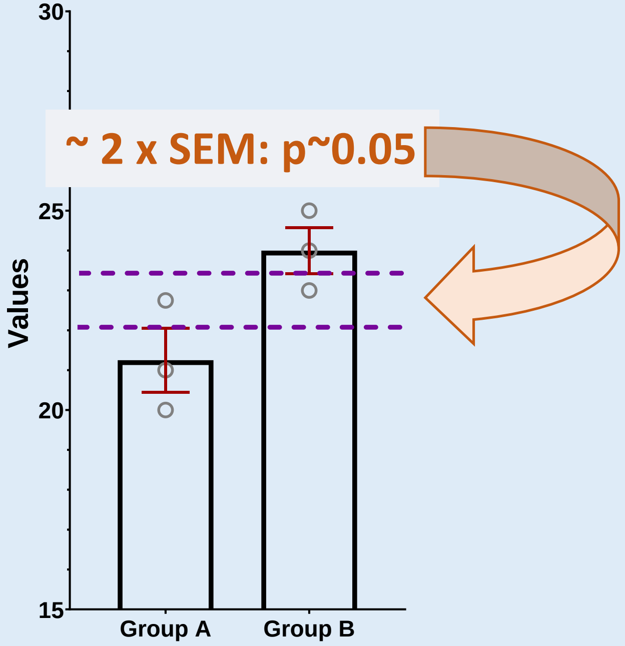
$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}}$$

$$= \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{\mathrm{var}_T}{n_T} + \frac{\mathrm{var}_C}{n_C}}}$$

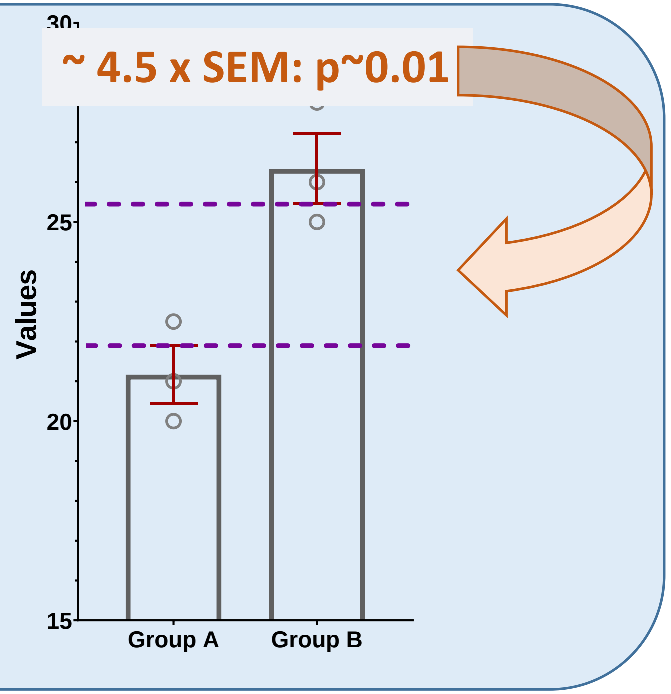
$$= \text{t-value}$$



n = 3

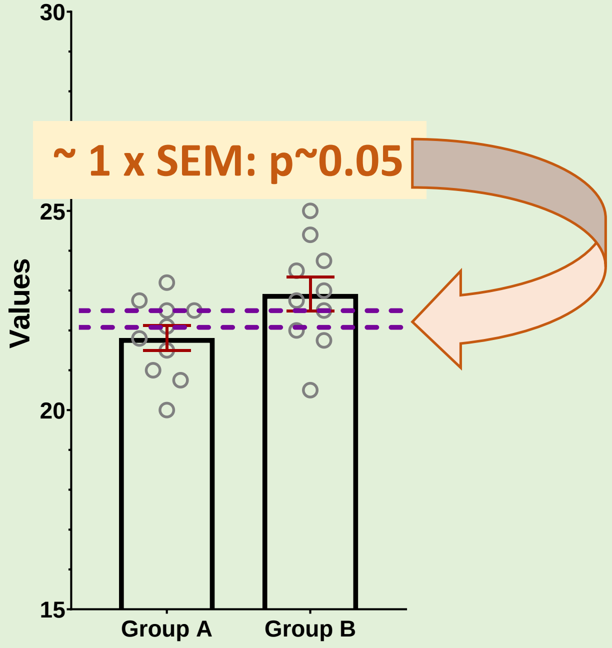


~ 2 x SEM: p~0.05

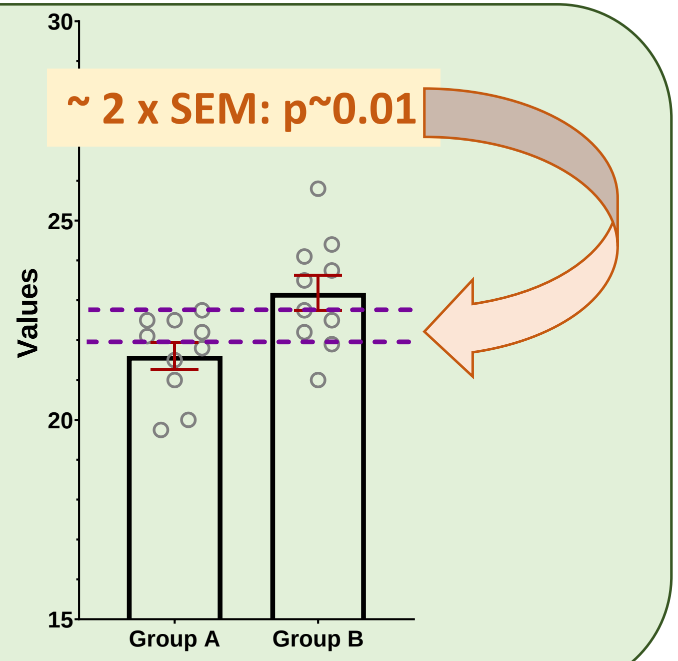


~ 4.5 x SEM: p~0.01

n = 10



~ 1 x SEM: p~0.05



~ 2 x SEM: p~0.01

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"))
```

Exercise 3: Power analysis - *Answers*

Example case:

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

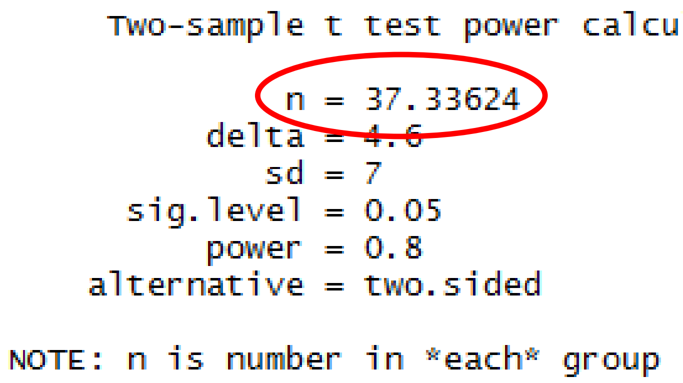
In a study run in similar conditions as in the one we intend to run, **male coyotes** were found to measure: **92cm+/- 7cm (SD)**

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)
```



```
     Two-sample t test power calculation

              n = 37.33624
          delta = 4.6
             sd = 7
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

We need a sample size of **n~76 (2*38)**

Data exploration $\neq$ plotting data

Exercise 4: Data exploration

coyote.csv

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

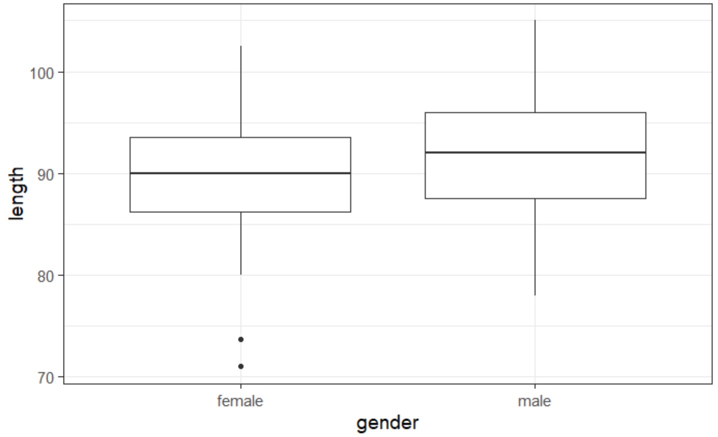
Question: do male and female coyotes differ in size?

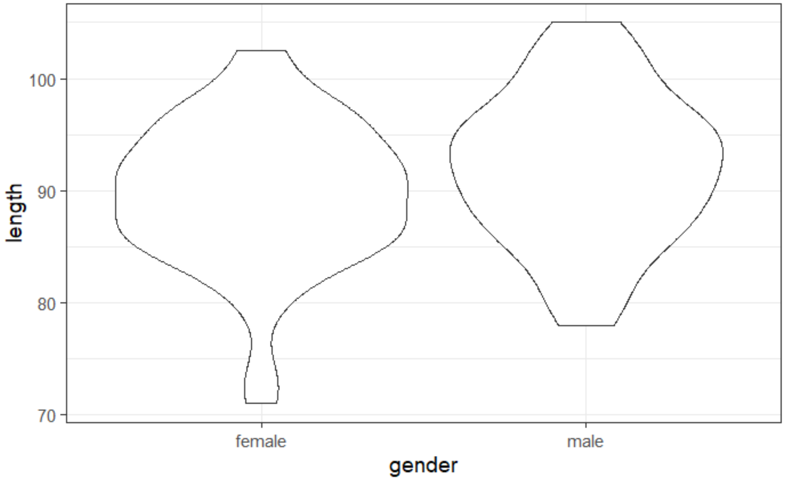
  - Load **coyote.csv**
  - Plot the data as boxplot, histogram, violinplot and stripchart

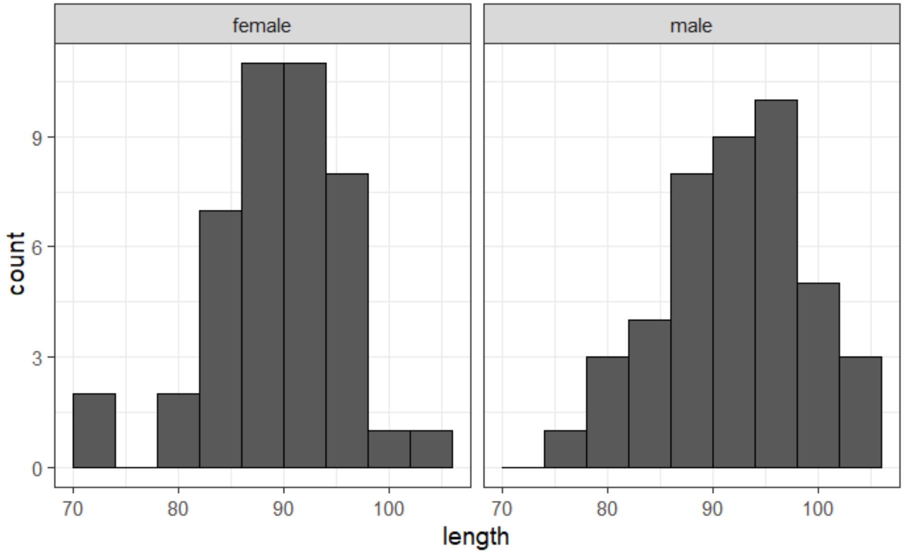
Data exploration ≠ plotting data

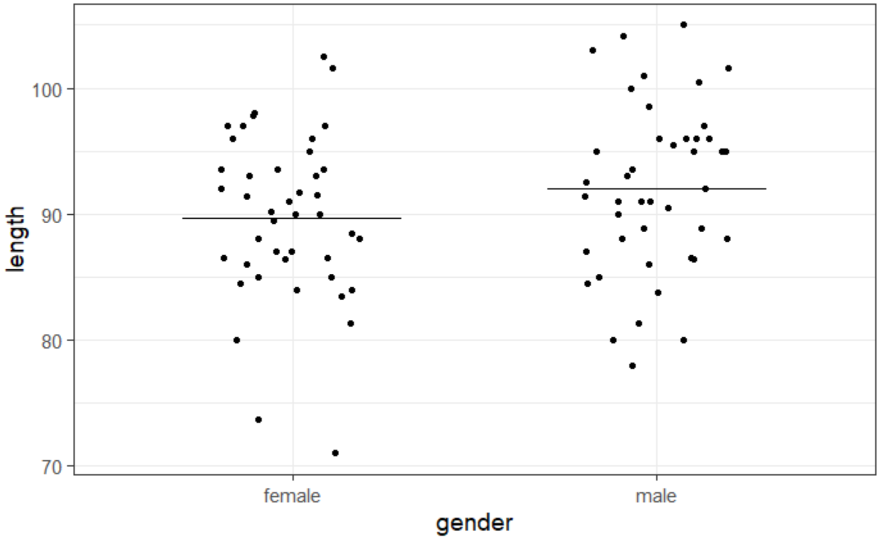
Exercise 4: Data exploration

- Explore data using 4 different representations:







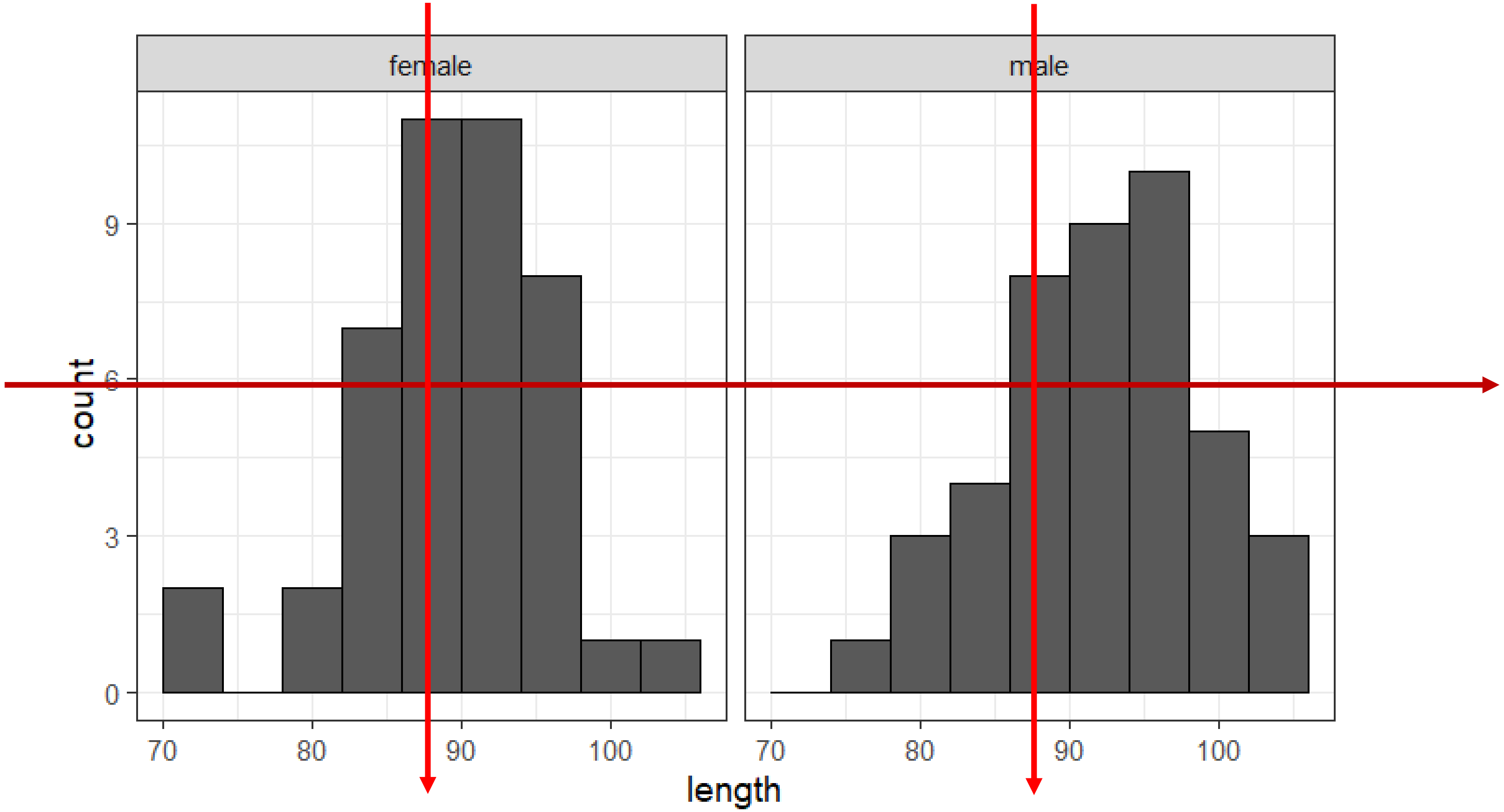


Exercise 4: `facet_grid(rows=vars(row),cols=vars(column))`

`facet_grid(cols=vars(gender))`

2 columns: one per gender

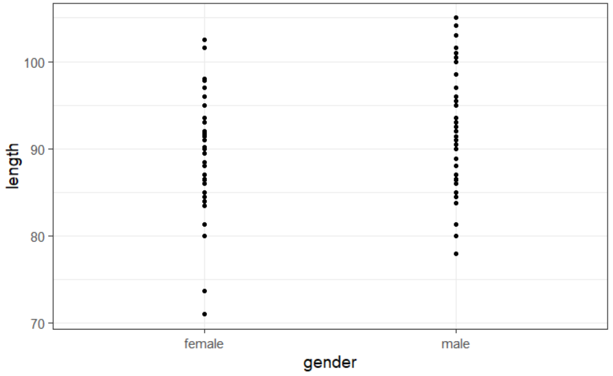
One row



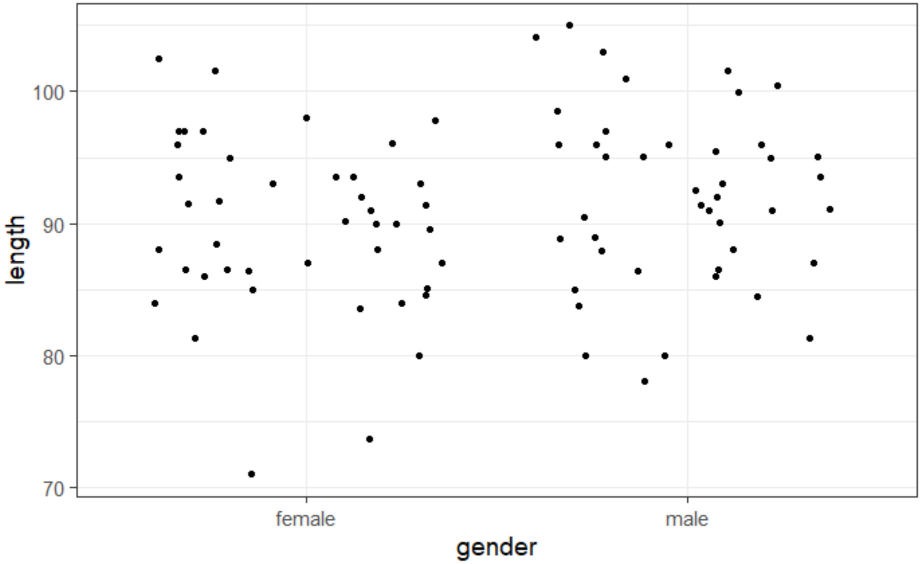
Exercise 4: geom_jitter()

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

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

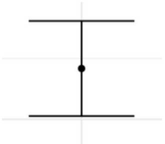


```
coyote %>%
  ggplot(aes(x=gender,y=length)) +
  geom_jitter(height=0, width=0.3)
```



Exercise 4: `stat_summary()`

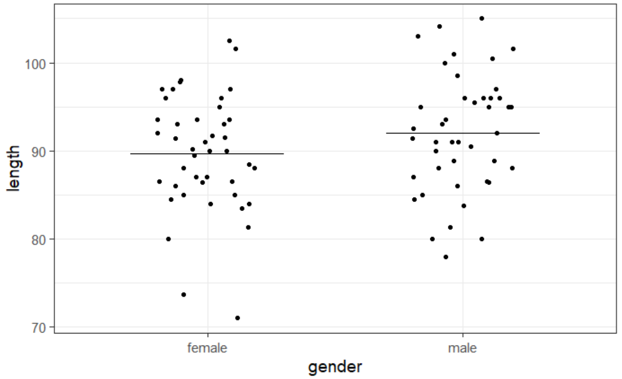
- Stripchart
  - `stat_summary()`
    - What statistical summary: mean: `fun = "mean"`
    - What `geom():` choice of graphical representation: a line: `geom_errorbar()`



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

mean=minimum=max

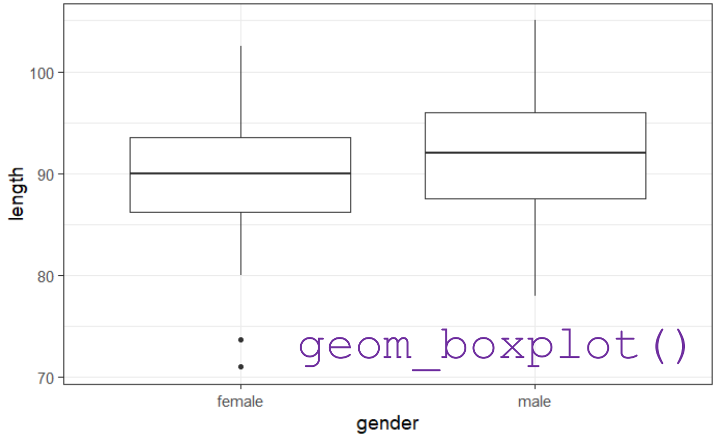
```
coyote %>%
  ggplot(aes(gender,length)) +
    geom_jitter(height=0, width=0.2)+
    stat_summary(geom= "errorbar", fun="mean", fun.min="mean", fun.max="mean")
```



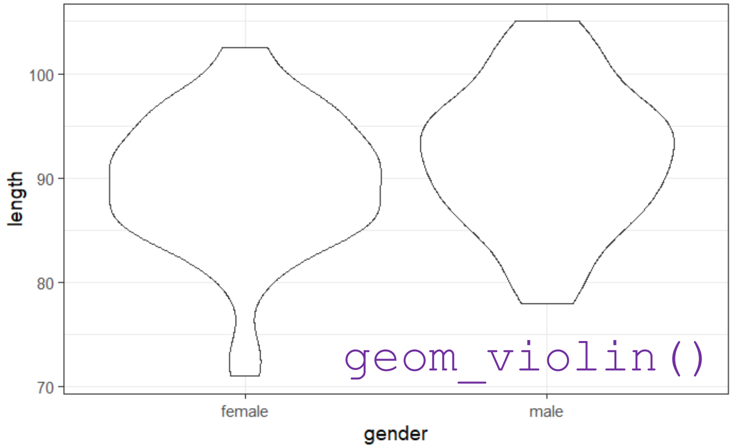
Exercise 4: Data exploration

```
coyote %>%
  ggplot(aes(x=gender, y=length))+
  geom_...()
```

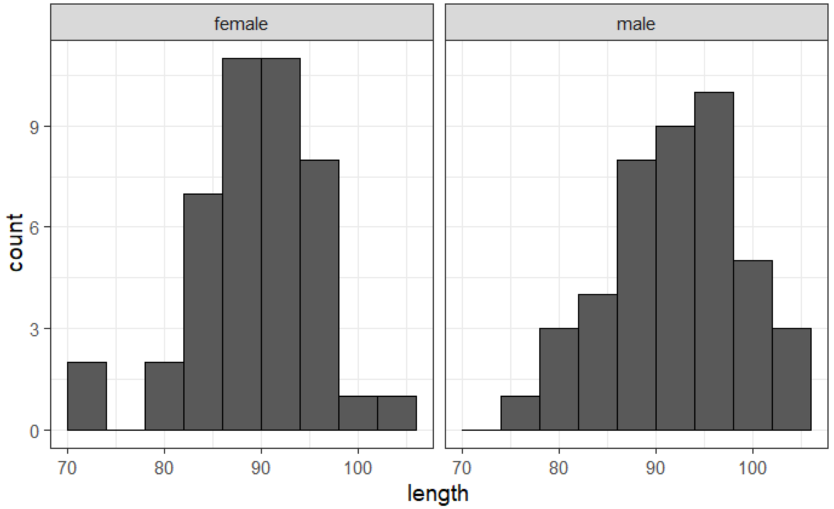
- Explore data using 4 different representations:

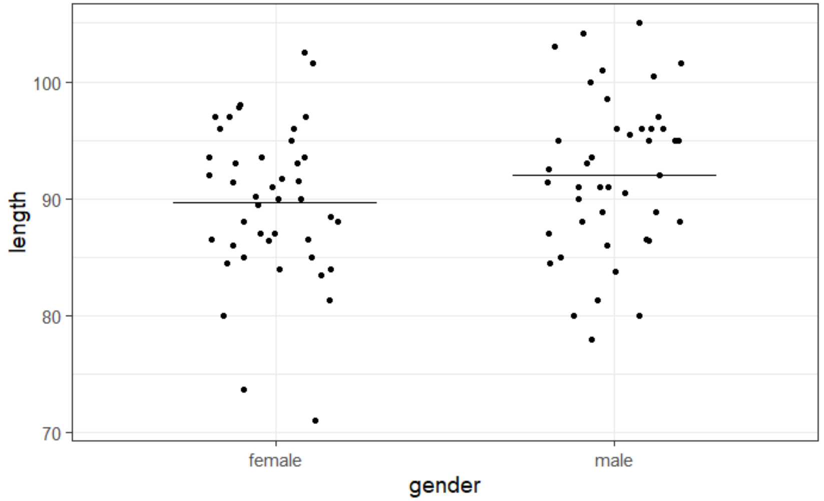


`geom_boxplot()`



`geom_violin()`





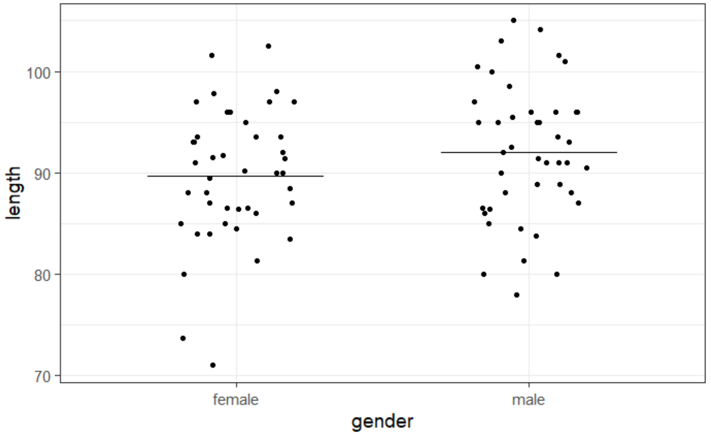
```
facet_grid(rows=vars(row),cols=vars(column))
geom_histogram
```

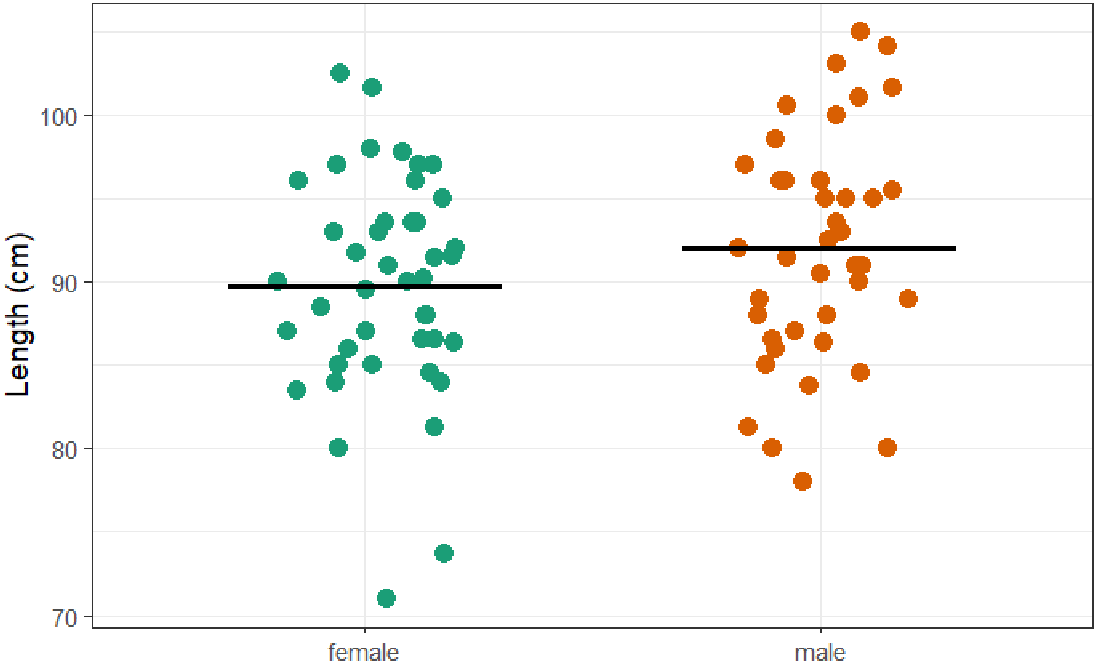
```
geom_jitter()
stat_summary()
```

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")
```

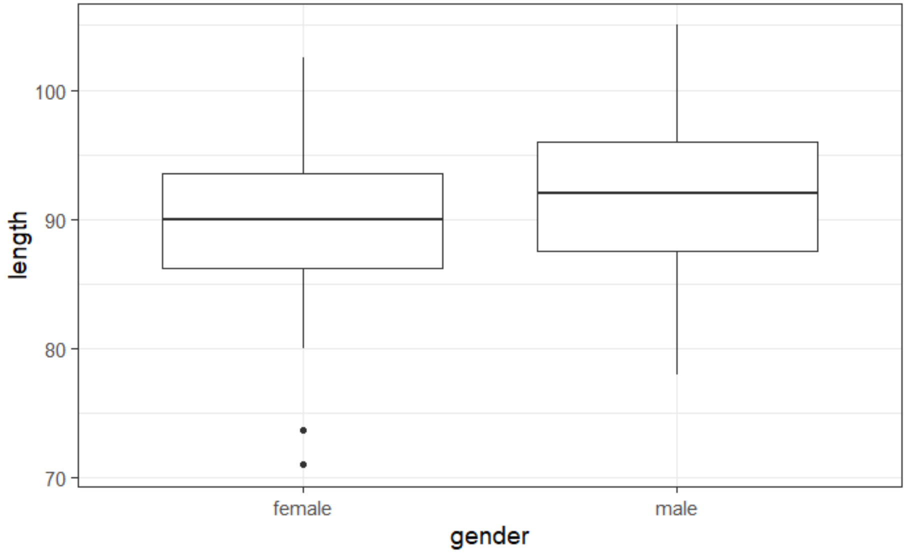


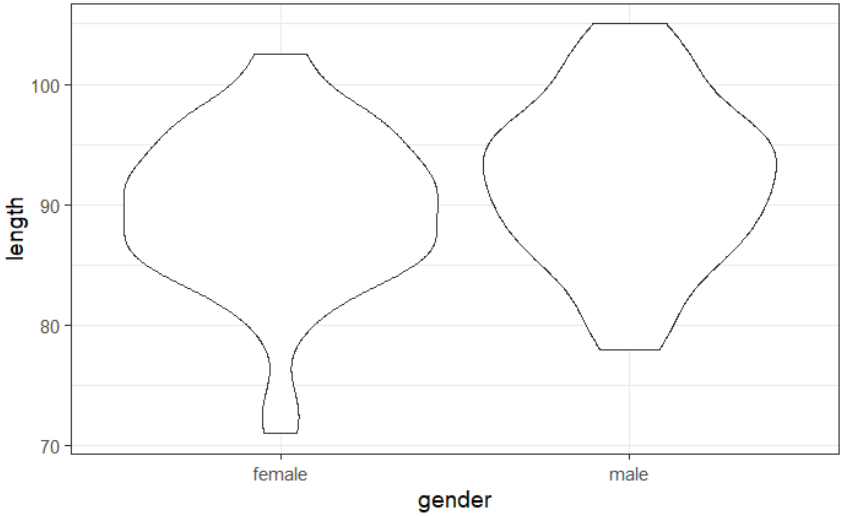


```
coyote %>%
  ggplot(aes(gender,length, colour=gender)) +
    geom_jitter(height=0, size=4, width=0.2, show.legend = FALSE) +
    ylab("Length (cm)")+
    scale_colour_brewer(palette="Dark2")+
    xlab(NULL)+
    stat_summary(geom="errorbar", fun=mean, fun.min=mean, fun.max=mean, colour="black", size=1.2,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

Exercise 4: Exploring data - Boxplots and beanplots

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

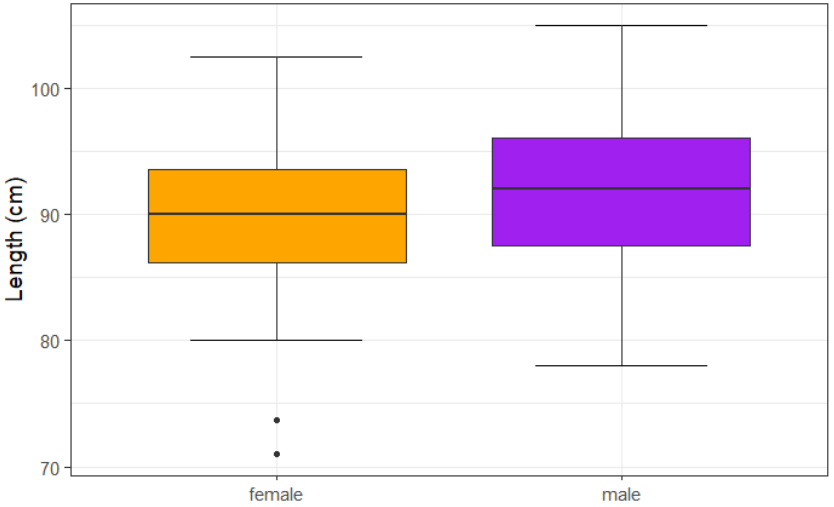


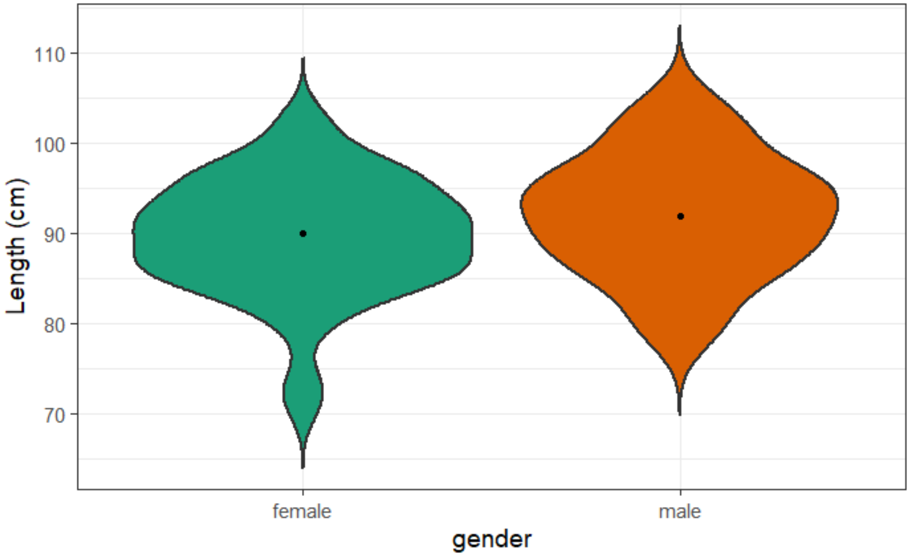


```
coyote %>%
  ggplot(aes(x=gender, y=length)) +
  geom_violin()
```

Exercise 4: Exploring data - 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"))
```





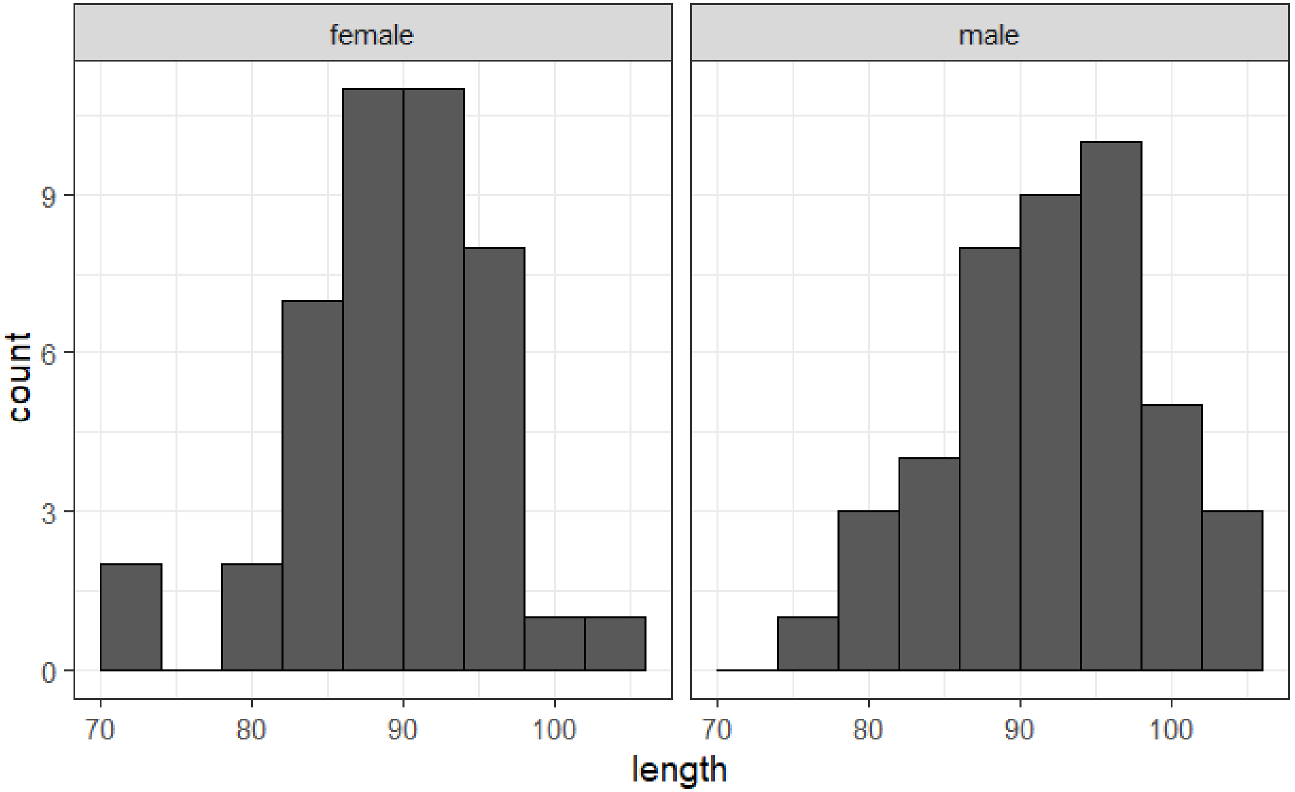
```
coyote %>%
  ggplot(aes(gender, length, fill=gender)) +
    geom_violin(trim=FALSE, size=1, show.legend=FALSE)+
    ylab("Length (cm)")+
    scale_fill_brewer(palette="Dark2")+
    stat_summary(geom = "point", fun = "median",show.legend=FALSE)
```

Exercise 4: Exploring data - Histograms

also works

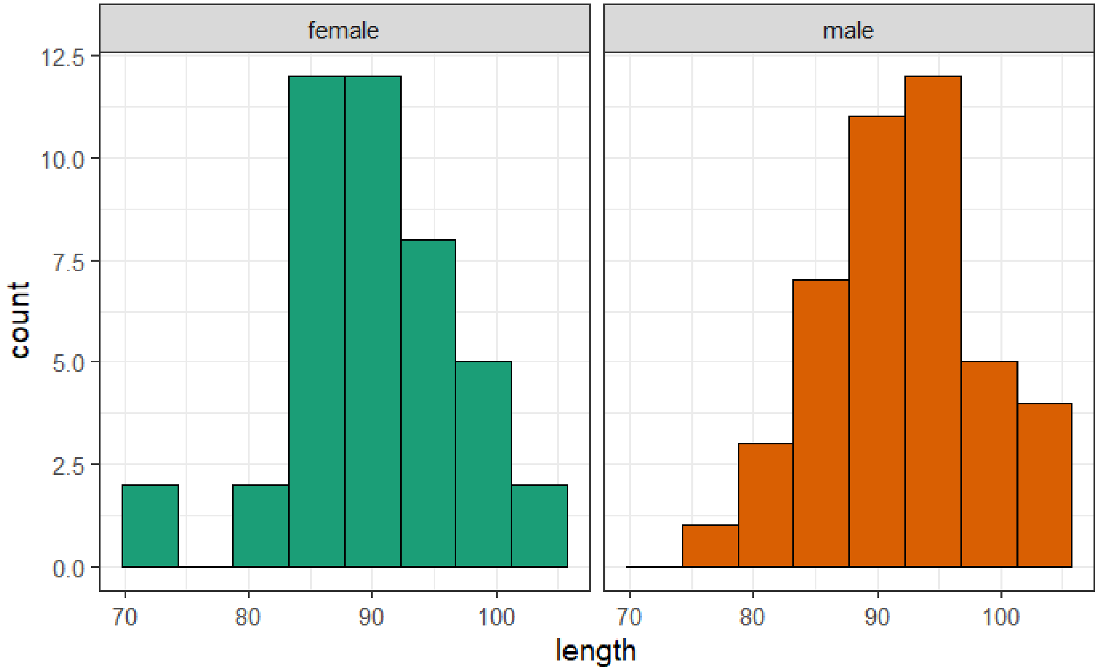
```
facet_wrap(vars(gender))
```

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



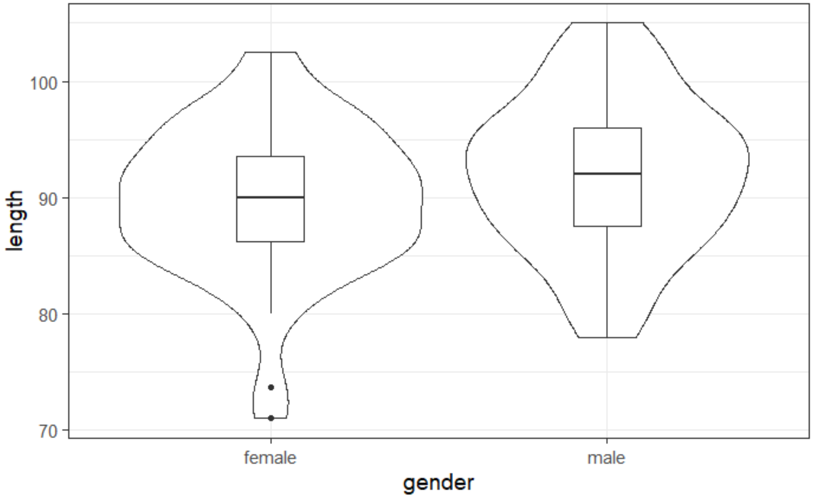
Exercise 4: Exploring data - Histograms

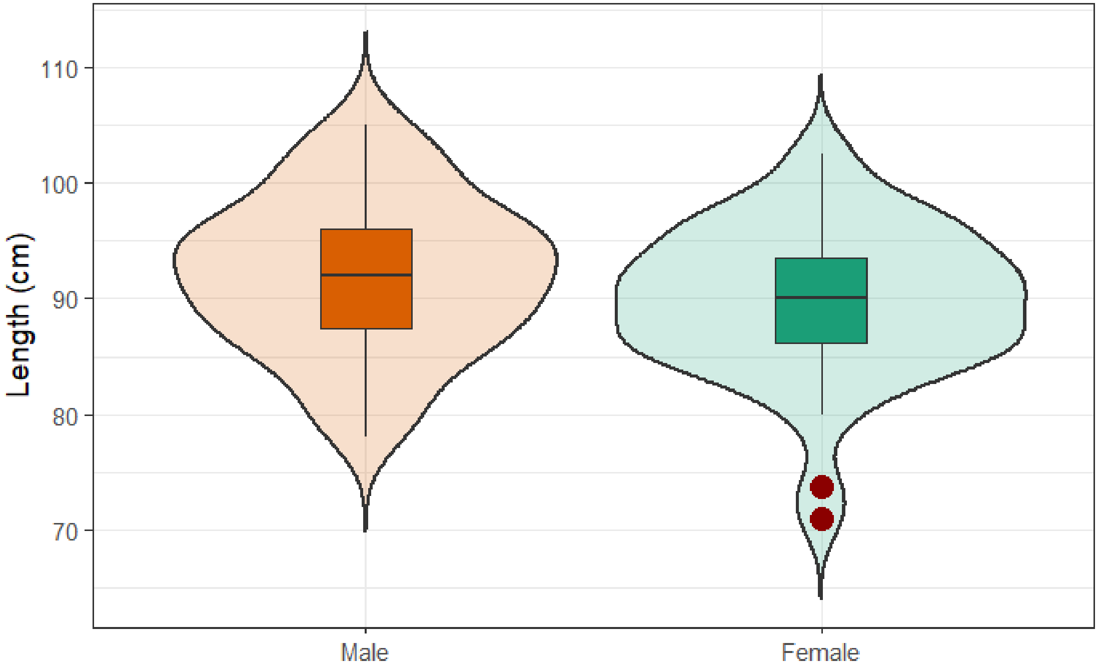
```
coyote %>%
  ggplot(aes(length, fill=gender))+
    geom_histogram(binwidth = 4.5, colour="black", show.legend = FALSE) +
    scale_fill_brewer(palette="Dark2")+
    facet_grid(cols=vars (gender))
```



Exercise 4 extra: Exploring data - Graph combinations

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

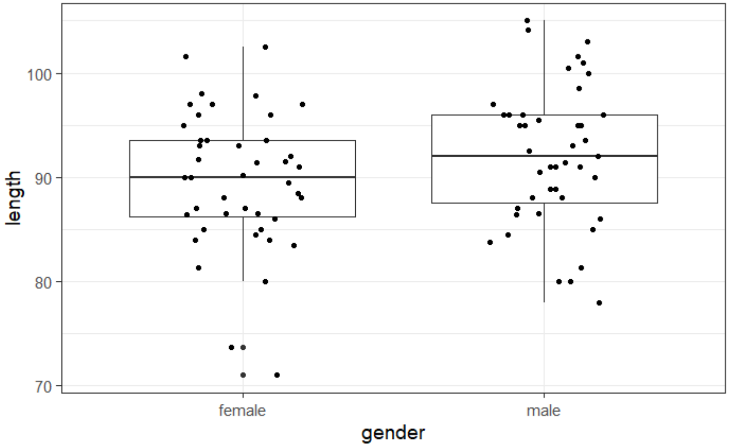


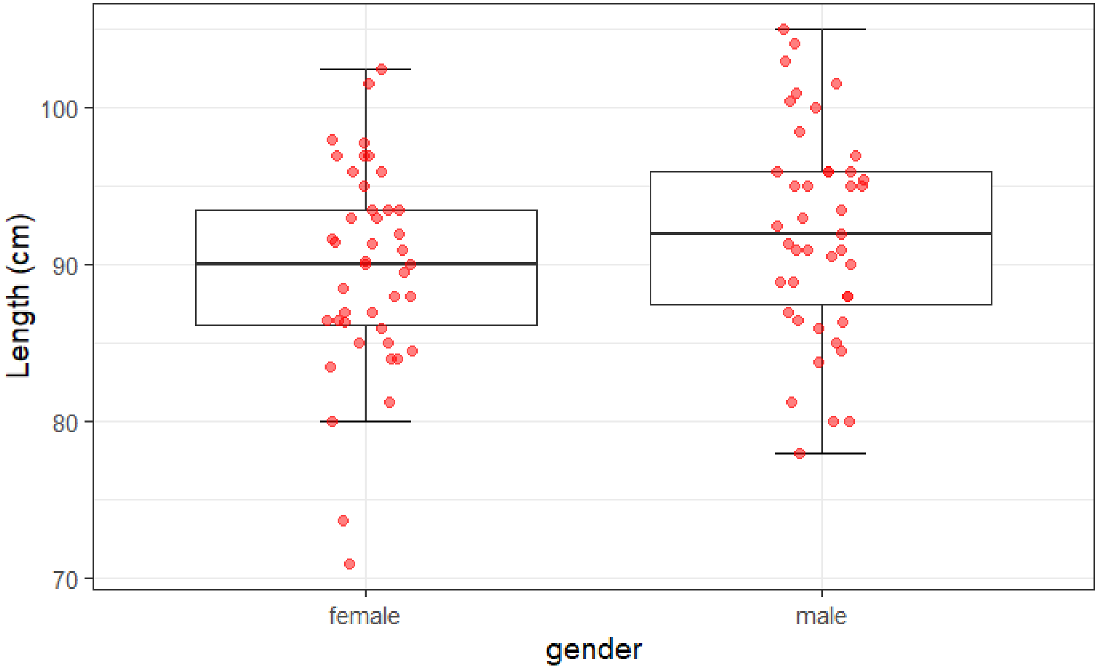


```
coyote %>%
  ggplot(aes(gender,length, fill=gender)) +
       geom_violin(size=1, trim = FALSE, alpha=0.2, show.legend=FALSE) +
       geom_boxplot(width=0.2, outlier.size=5, outlier.colour = "darkred", show.legend=FALSE)+
       scale_fill_brewer(palette="Dark2")+
       ylab("Length (cm)")+
       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)
```

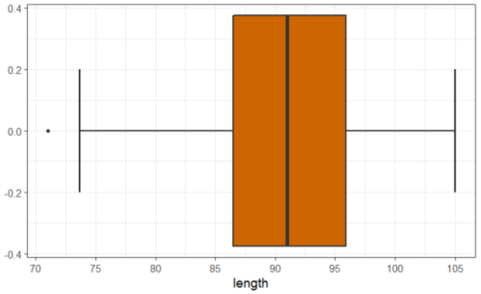




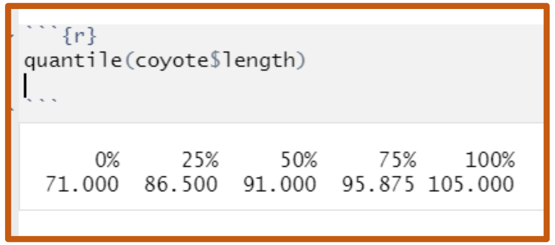
```
coyote %>%
  ggplot(aes(gender, length)) +
       geom_boxplot(outlier.shape=NA)+
       stat_boxplot(geom="errorbar", width=0.2)+
       geom_jitter(height=0, width=0.1, size=2, alpha=0.5, colour="red")+
       ylab("Length (cm)")
```

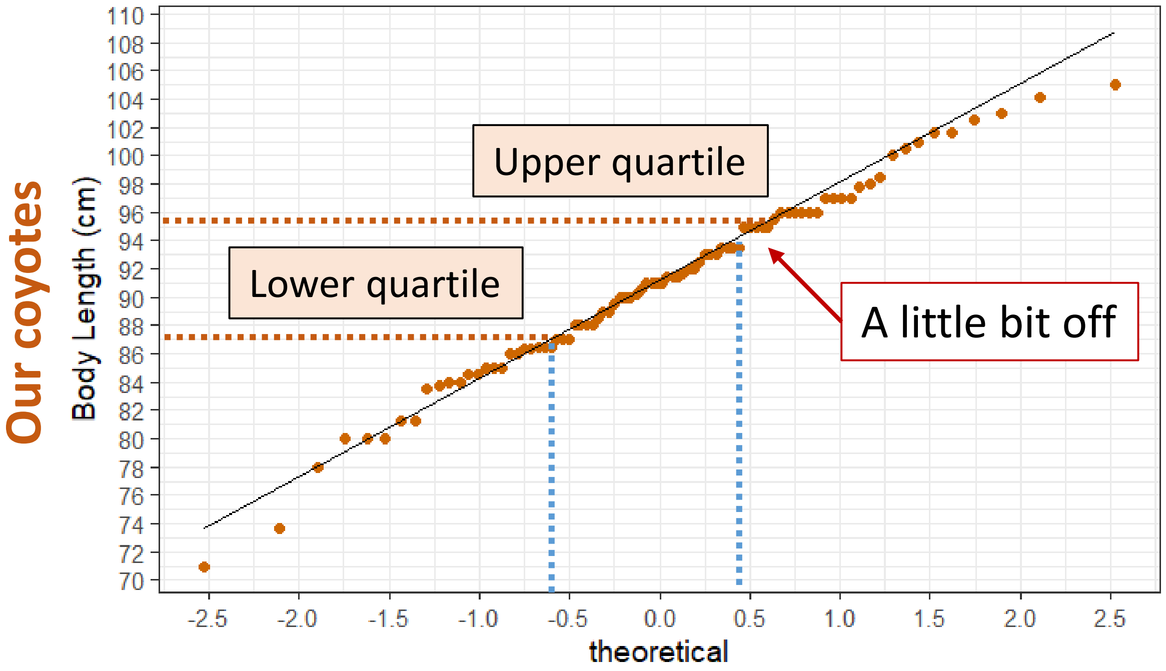
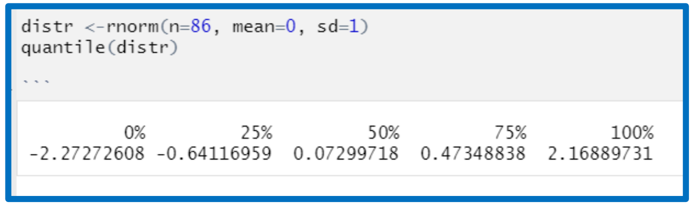
Checking the assumptions

Normality assumption: QQ Plot



Quantiles:

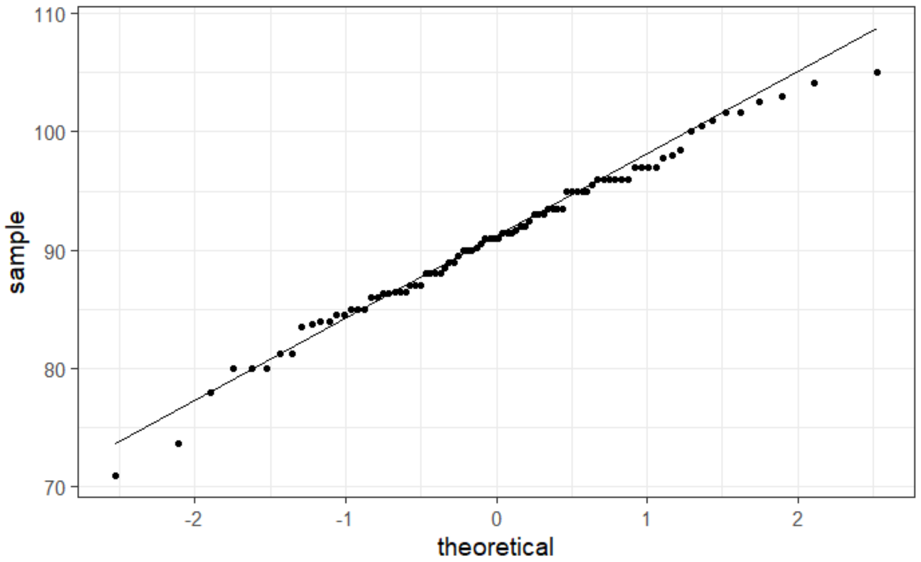


~~~
```{r}
quantile(coyote$length)
~~~
```

```
    0%     25%     50%     75%    100% 
 71.000  86.500  91.000  95.875 105.000 
```

QQ plot= Quantile – Quantile plot



Our coyotes
Body Length (cm)
110
108
106
104
102
100
98
96
94
92
90
88
86
84
82
80
78
76
74
72
70
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
theoretical

Upper quartile

Lower quartile

A little bit off

Normality ☑ (ish)

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

Quantiles:



```
distr <-rnorm(n=86, mean=0, sd=1)
quantile(distr)
```
```

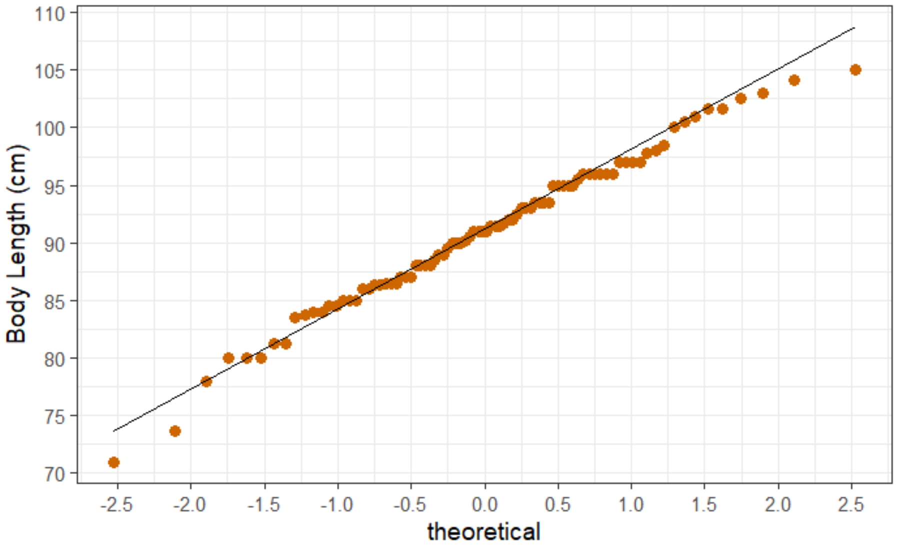
```
         0%         25%         50%         75%        100% 
-2.27272608 -0.64116959  0.07299718  0.47348838  2.16889731 
```

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))
```

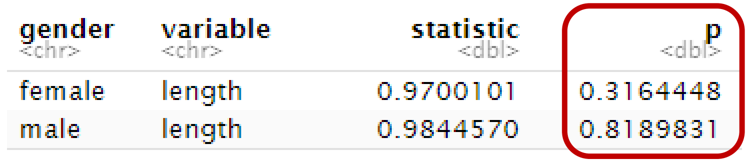


Assumptions of Parametric Data

- 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()`

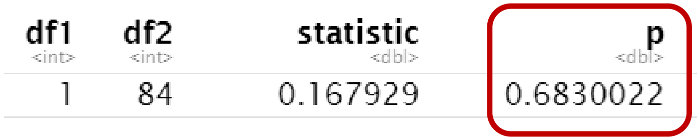
```
coyote %>%
  group_by(gender) %>%
   shapiro_test(length) %>%
       ungroup()
```



```
gender   variable   statistic          p
<chr>    <chr>          <dbl>      <dbl>
female   length     0.9700101  0.3164448
male     length     0.9844570  0.8189831
```

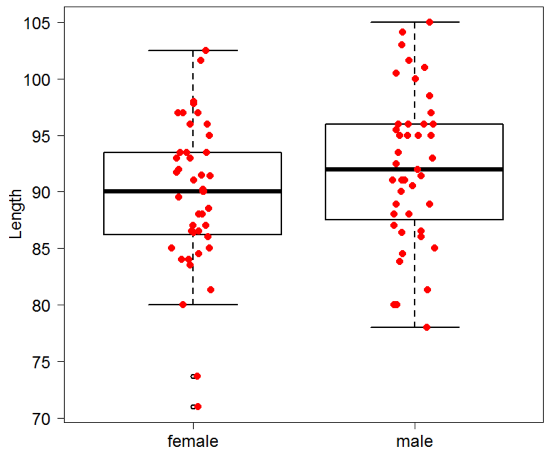
Normality ☑

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



```
df1   df2   statistic          p
<int> <int>     <dbl>      <dbl>
  1    84    0.167929  0.6830022
```

Homogeneity in variance ☑



Normality

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

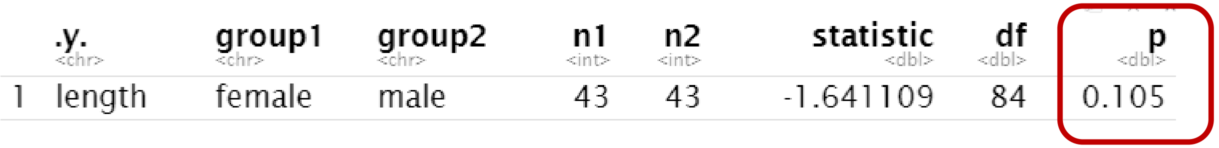
Homoscedasticity

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

Independent *t*-test: results (tidyverse)

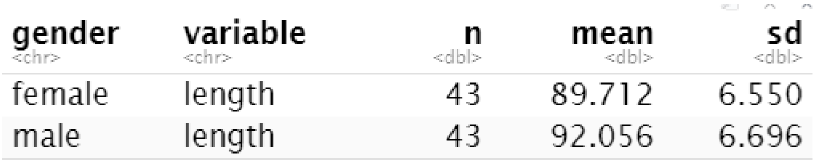
coyote.csv

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



```
  .y.      group1   group2     n1    n2     statistic    df       p
  <chr>    <chr>    <chr>    <int> <int>      <dbl>    <dbl>   <dbl>
1 length   female   male       43    43    -1.641109     84   0.105
```

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



```
gender    variable        n      mean      sd
<chr>     <chr>       <dbl>     <dbl>   <dbl>
female    length         43    89.712   6.550
male      length         43    92.056   6.696
```

- **<u>Answer</u>: Males tend to be longer than females but not significantly so (p=0.1045).**

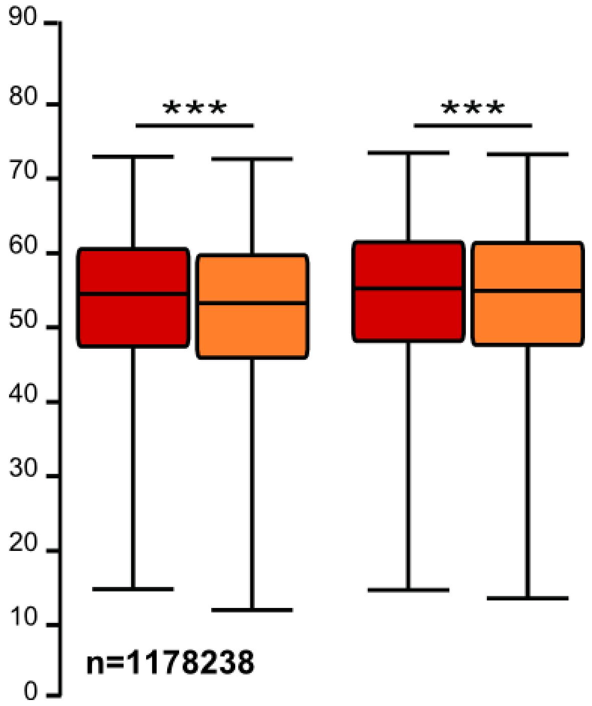
- Power : How many more coyotes to reach significance?
  - Re-run the power analysis with mean=89.7 for females: n~250
    - **But does it make sense?**

Sample size: the bigger the better?

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

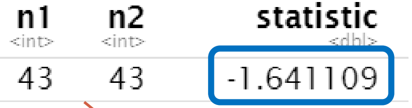
- What if the tiny difference is meaningless?
  - Beware of **overpower**
  - Nothing wrong with the stats: it is all about interpretation of the results of the test.

- Remember the important first step of power analysis
  - **What is the effect size of biological interest?**

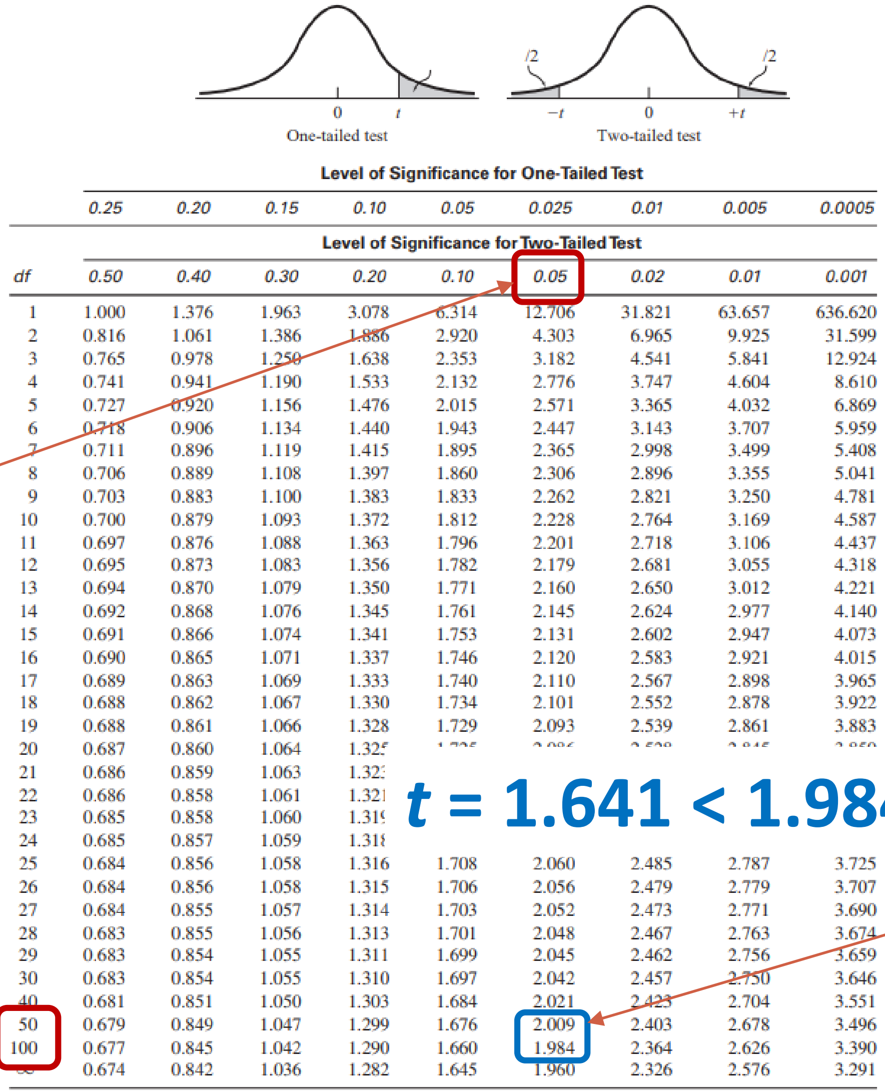


Independent *t*-test: results

The old-fashion way



```
   n1    n2     statistic
<int> <int>         <dbl>
   43    43     -1.641109
```



/2
/2
0
t
−t
0
+t
One-tailed test
Two-tailed test

Level of Significance for One-Tailed Test

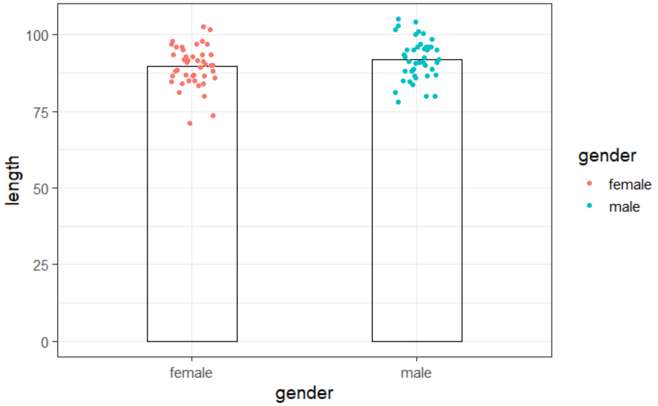
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.0005
	Level of Significance for Two-Tailed Test								
df	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.620
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
21	0.686	0.859	1.063	1.323	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
22	0.686	0.858	1.061	1.321	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
23	0.685	0.858	1.060	1.319	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
24	0.685	0.857	1.059	1.318	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	3.496
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.390
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

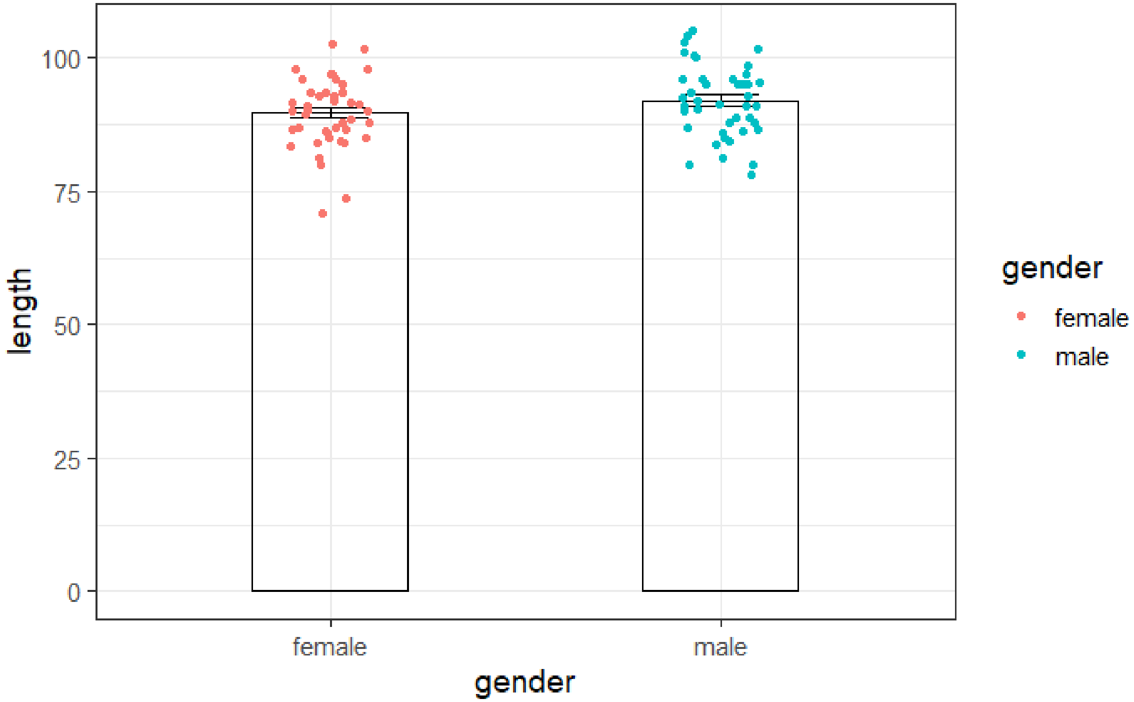
t **= 1.641 < 1.984: not significant**

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)
```



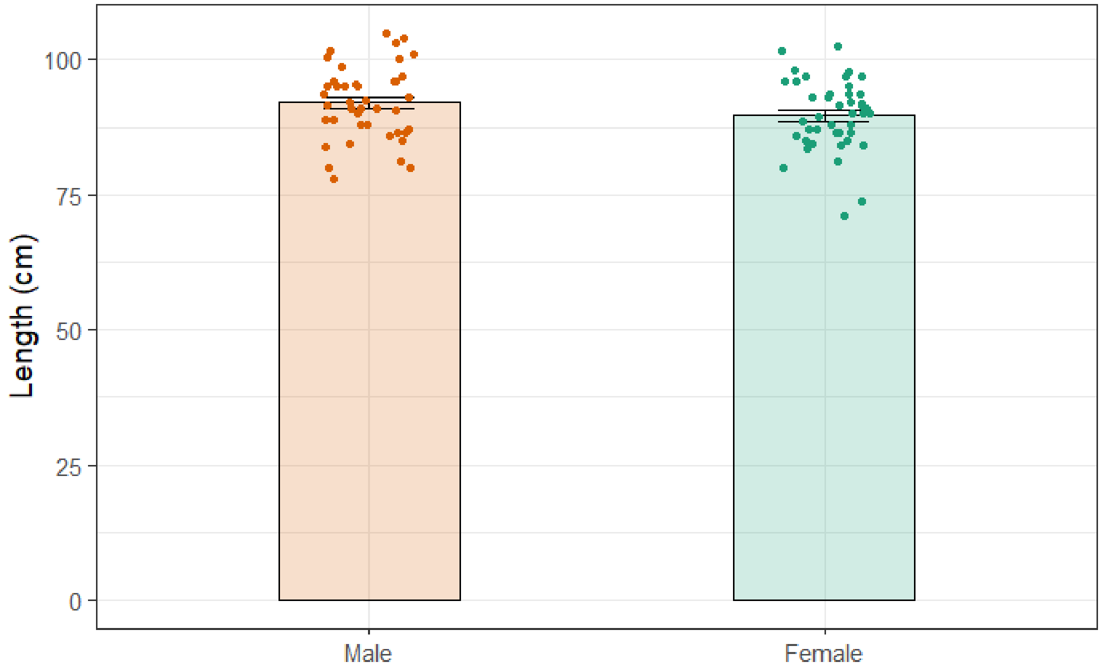


- Add error bars

```
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)
```

Plot ‘coyote.csv’ data: Plotting data

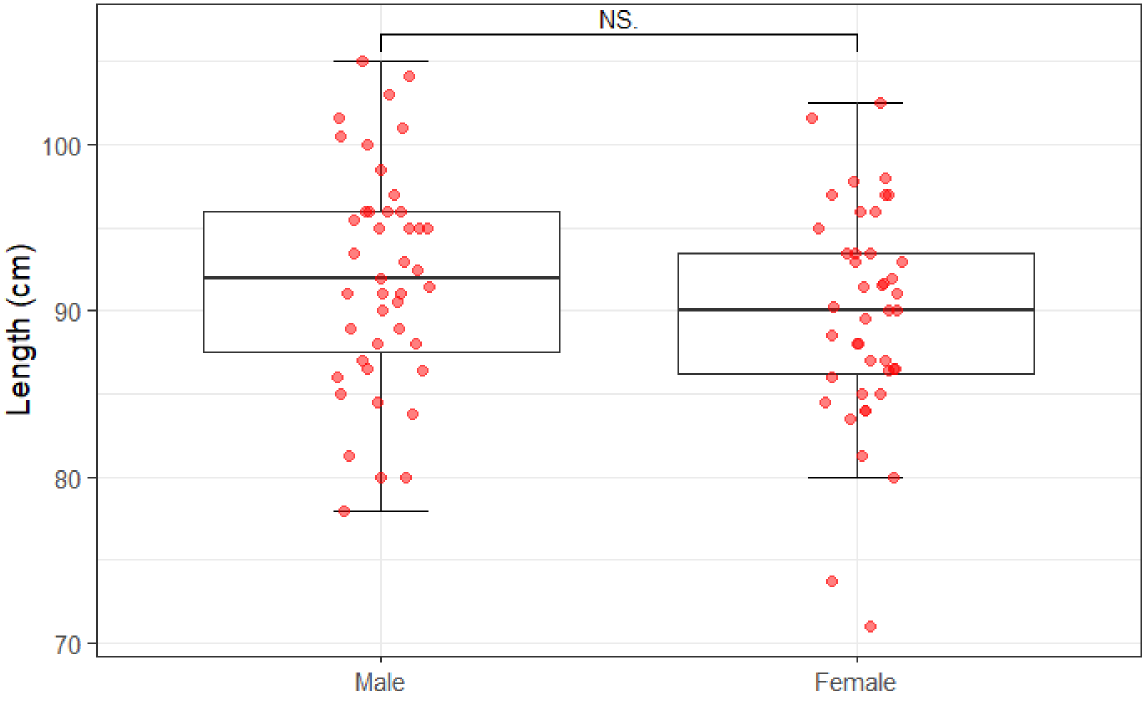
- Prettier version



```
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 #



```
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")
```

Exercise 5: Dependent or Paired *t*-test

working.memory.csv

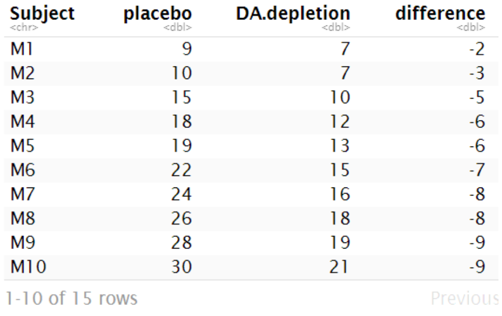
- A researcher is studying the effects of dopamine depletion on working memory in rhesus monkeys.
  - A group of rhesus monkeys (n=15) performs a task involving memory after having received a placebo. Their performance is graded on a scale from 0 to 100. They are then asked to perform the same task after 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_test(var ~ 1, mu=0)`

Exercise 5: Dependent or Paired *t*-test - *Answers*

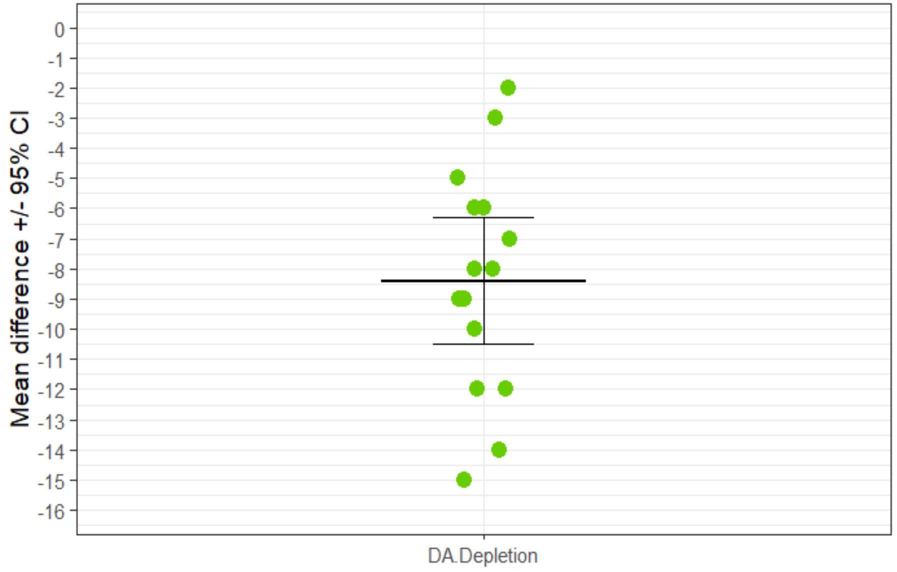
```
working.memory %>%
  mutate(difference = DA.depletion - placebo) -> working.memory
```



Subject	placebo	DA.depletion	difference
<chr>	<dbl>	<dbl>	<dbl>
M1	9	7	-2
M2	10	7	-3
M3	15	10	-5
M4	18	12	-6
M5	19	13	-6
M6	22	15	-7
M7	24	16	-8
M8	26	18	-8
M9	28	19	-9
M10	30	21	-9

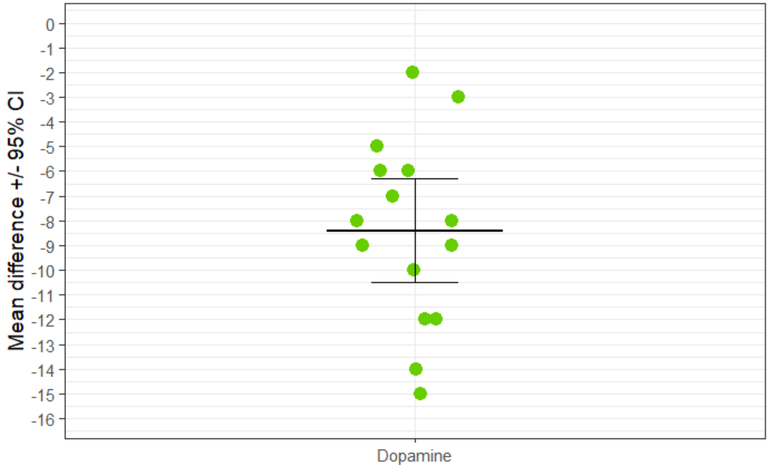
1-10 of 15 rows Previous

```
# 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")
```

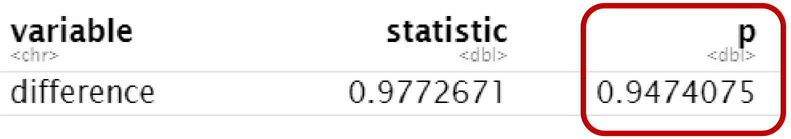


Exercise 5: Dependent or Paired *t*-test (tidyverse)

Question: does dopamine affect working memory in rhesus monkeys?

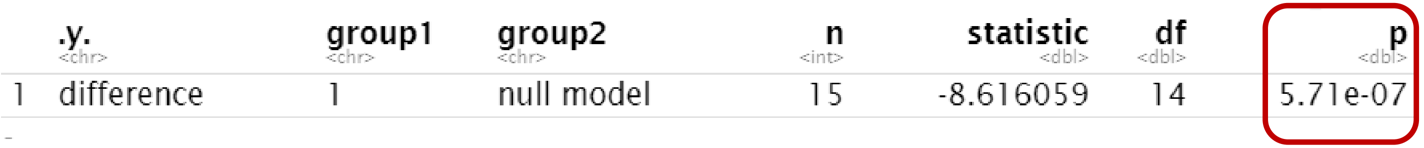


```
working.memory %>%
  shapiro_test(difference)
```



```
variable           statistic          p
<chr>                  <dbl>      <dbl>
difference         0.9772671  0.9474075
```

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



```
  .y.          group1  group2       n   statistic  df          p
  <chr>        <chr>   <chr>    <int>       <dbl> <dbl>    <dbl>
1 difference   1       null model  15   -8.616059  14   5.71e-07
```

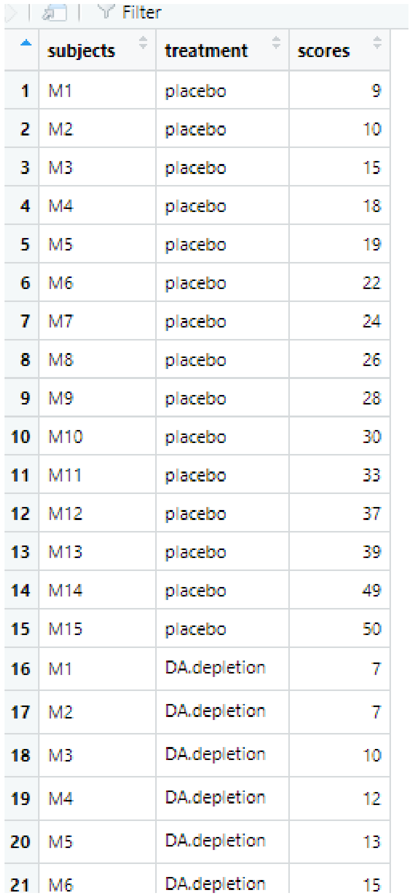
Answer: the injection of a dopamine-depleting agent significantly affects working memory in rhesus monkeys (t=-8.62, df=14, p=5.715e-7).

Dependent or Paired *t*-test

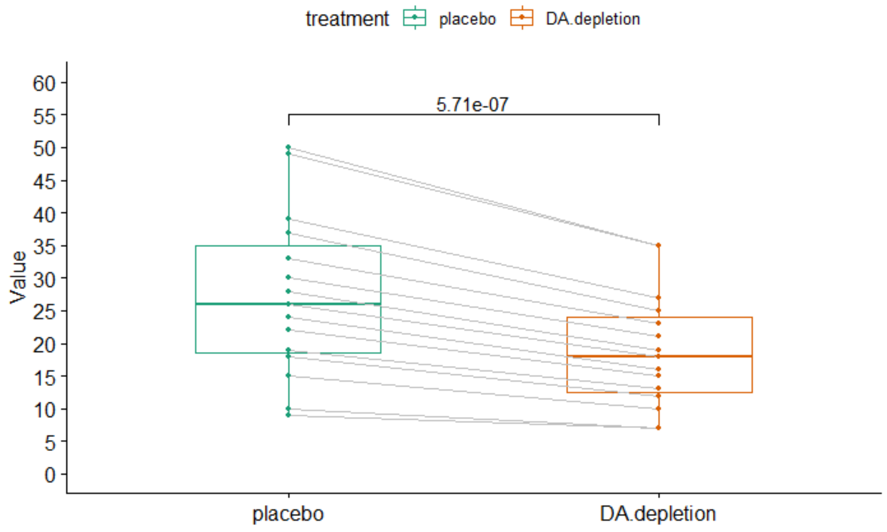
- ***Work in progress*** # ggpubr package #

```
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)
```

working.memory.long



	subjects	treatment	scores
1	M1	placebo	9
2	M2	placebo	10
3	M3	placebo	15
4	M4	placebo	18
5	M5	placebo	19
6	M6	placebo	22
7	M7	placebo	24
8	M8	placebo	26
9	M9	placebo	28
10	M10	placebo	30
11	M11	placebo	33
12	M12	placebo	37
13	M13	placebo	39
14	M14	placebo	49
15	M15	placebo	50
16	M1	DA.depletion	7
17	M2	DA.depletion	7
18	M3	DA.depletion	10
19	M4	DA.depletion	12
20	M5	DA.depletion	13
21	M6	DA.depletion	15



Comparison between more than 2 groups

One factor = One predictor

One-Way ANOVA

Signal-to-noise ratio

$$\frac{\text{Difference}}{\text{Variability}} = \frac{\text{Signal}}{\text{Noise}}$$

$$\frac{\text{Signal}}{\scriptsize\text{Noise}} = \text{statistical significance}$$

$$\frac{\text{Signal}}{\Large\text{Noise}} = \text{no statistical significance}$$

Analysis of variance: how does it work?

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

$$= \text{F ratio}$$

One-Way Analysis of variance

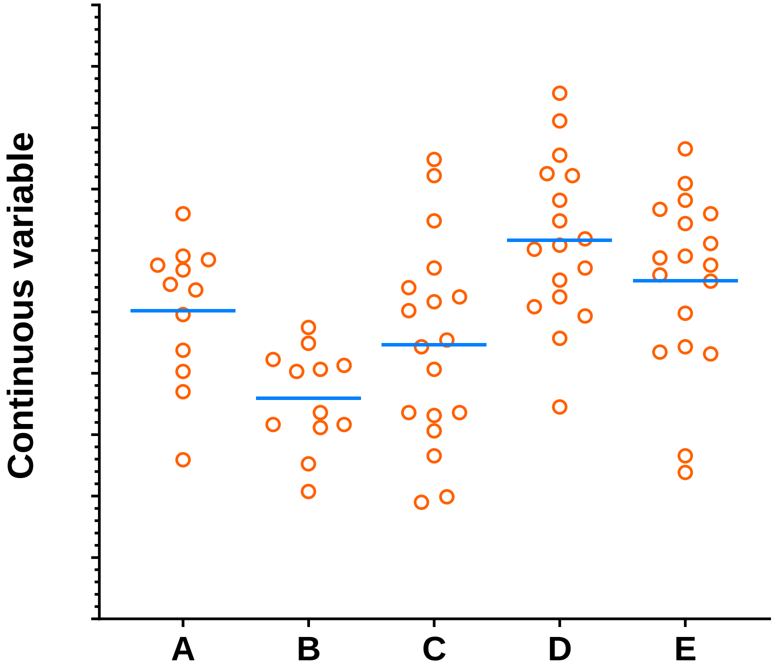
Step 1: Omnibus test

- 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

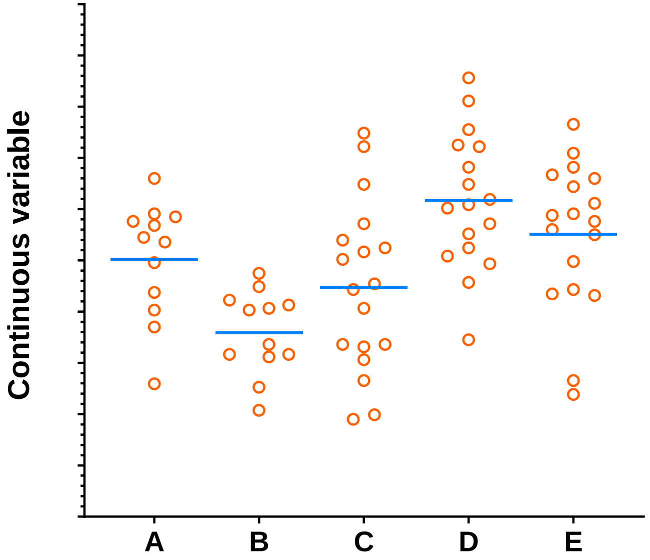
- They tell us if there are differences between the means pairwise.

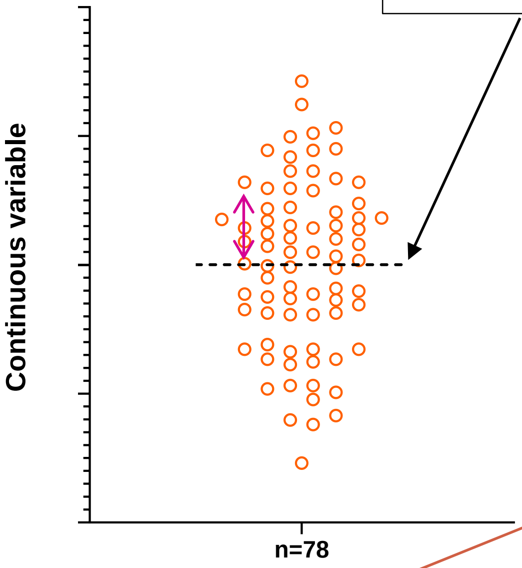
Analysis of variance: how does it work?



Source of variation	Sum of Squares	df	Mean Square	F	p-value
Between Groups	18.1	4	4.5	6.32	0.0002
Within Groups	51.8	73	0.71		
Total	69.9				

Analysis of variance: how does it work?





grand mean

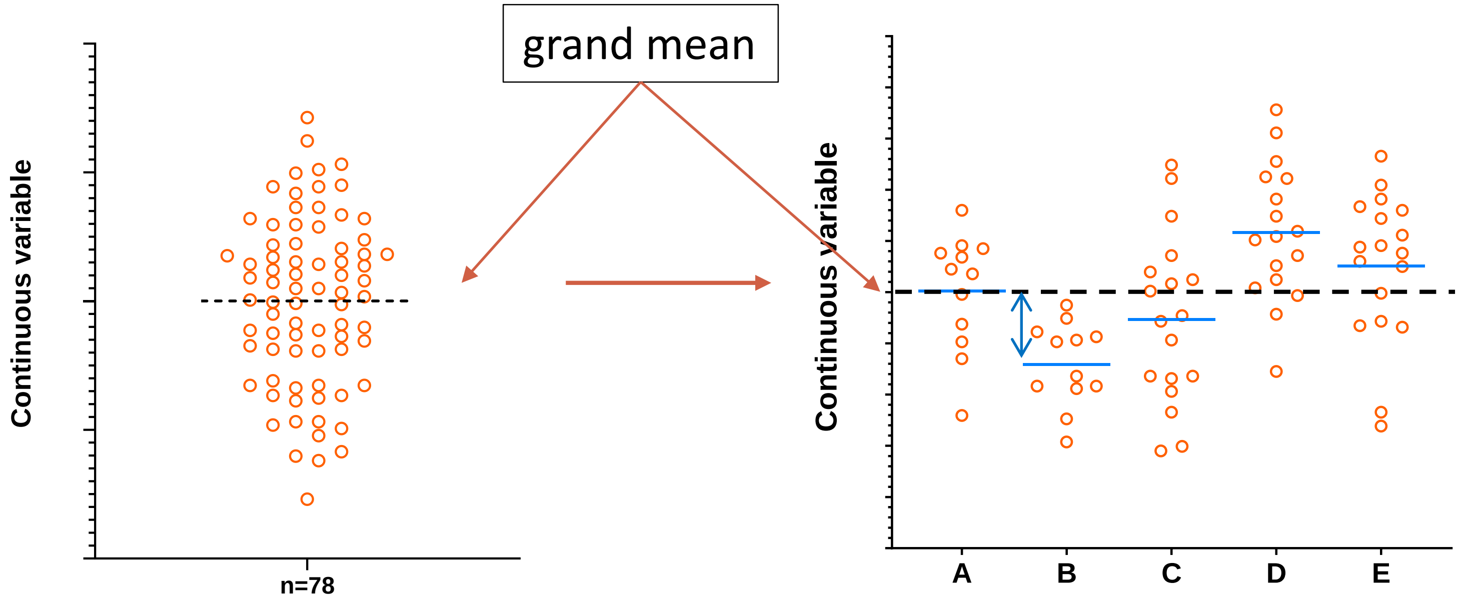
78 differences: $\sum_1^{78}(\text{value}_\text{n} - \text{grand mean})^2$

=

Sum of squared errors

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

Analysis of variance: how does it work?



5 differences: $\sum_1^5 (\text{mean}_\text{n} - \text{grand mean})^2$

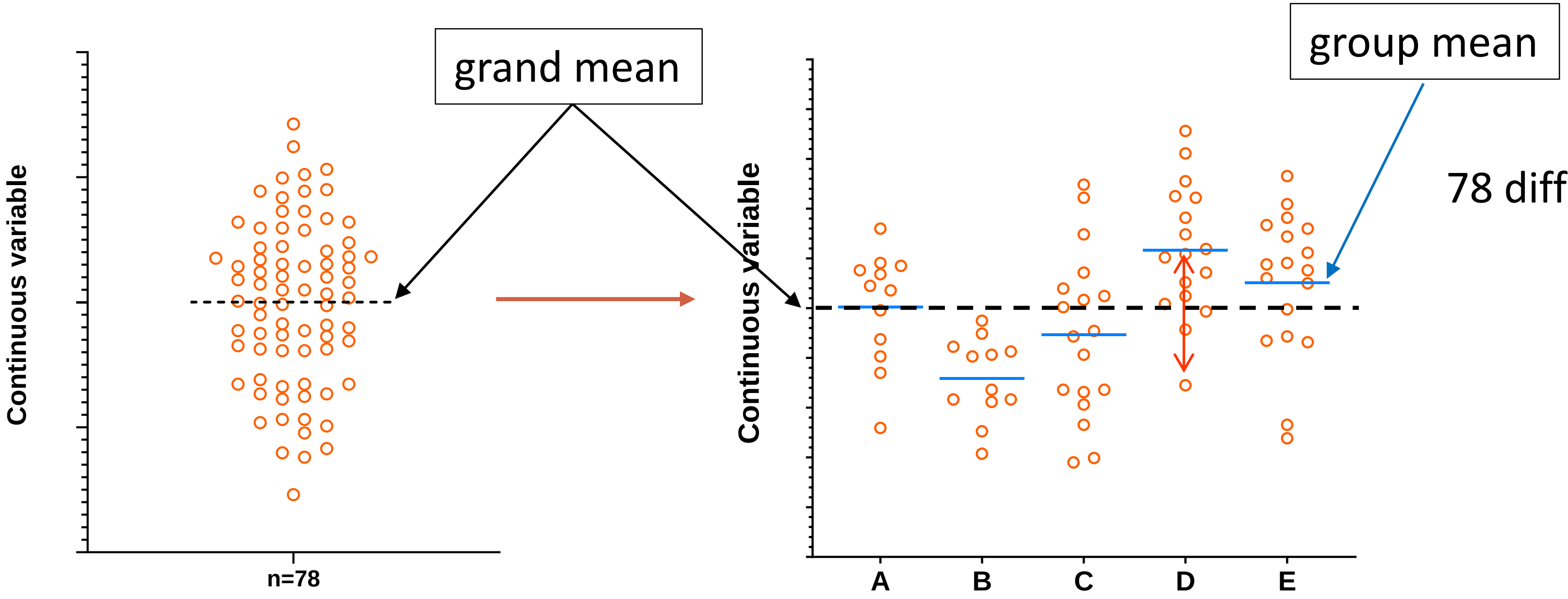
=

Sum of squared errors

Between the groups

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

Analysis of variance: how does it work?



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

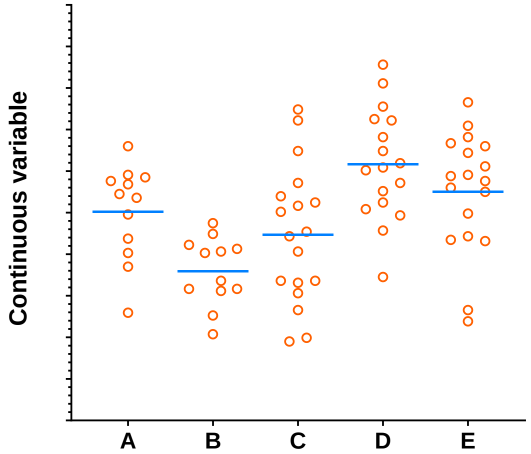
=

Sum of squared errors

Within the Groups

Source of variation	Sum of Squares	df	Mean Squares	F	p-value
Between Groups	18.1				
Within Groups	51.8				
Total	69.9				

Analysis of variance: how does it work?



	Source of variation	Sum of Squares	df	Mean Squares	F ratio	p-value
Signal	Between Groups	18.1	k-1			
Noise	Within Groups	51.8	n-k			
	Total	69.9				

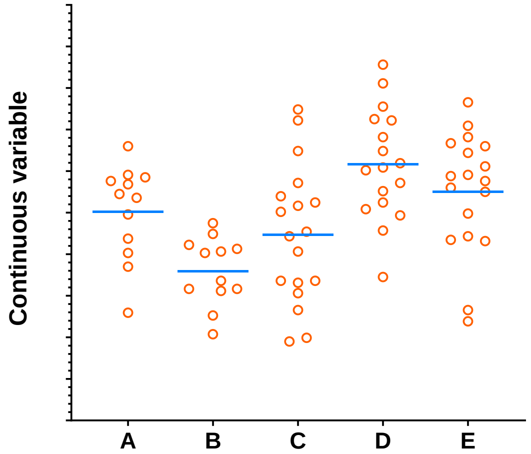
df: degree of freedom with df = n-1

n = number of values, k=number of groups

Between groups: df = 4 (k-1)

Within groups: df = 73 ($n-k = n_1-1 + \ldots + n_5-1$)

Analysis of variance: how does it work?



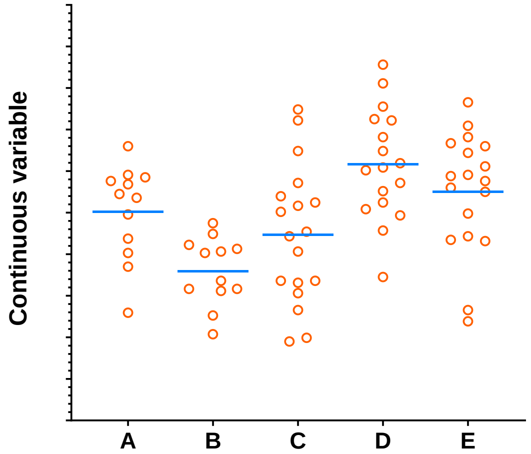
	Source of variation	Sum of Squares	df	Mean Squares	F ratio	p-value
Signal	Between Groups	18.1	4	4.5		
Noise	Within Groups	51.8	73	0.71		
	Total	69.9				

df: degree of freedom with df = n-1

18.2/4 = 4.5 **51.8/73 = 0.71**

Mean squares = **Sum of Squares / n-1 = Variance!**

Analysis of variance: how does it work?



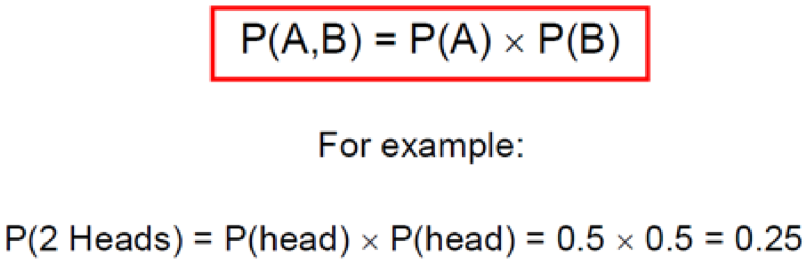
Source of variation	Sum of Squares	df	Mean Squares	F ratio	p-value
Between Groups	18.1	4	4.5	$\boxed{6.34}$	0.0002
Within Groups	51.8	73	0.71		
Total	69.9				

Mean squares = **Sum of Squares / n-1 = Variance**

$$\text{F ratio} = \frac{\text{Variance between the groups}}{\text{Variance within the groups (individual variability)}} = \frac{4.5}{0.71} = \mathbf{6.34}$$

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.



$$\boxed{P(A,B) = P(A) \times P(B)}$$

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 * 0.95 * 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**: α_{adjust} = 0.05/n comparisons e.g. 3 comparisons: 0.05/3=0.016
  - Problem: very conservative leading to loss of power (lots of false negative)
  - 10 comparisons: threshold for significance: 0.05/10: 0.005
  - Pairwise comparisons across 20.000 genes ☹

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

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

One-Way Analysis of variance

<u>Step 1</u>: 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.

<u>Step 2</u>: Post-hoc tests

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

Example: protein.expression.csv

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

- **1 Plot the data**

- **2 Check the assumptions for parametric test**

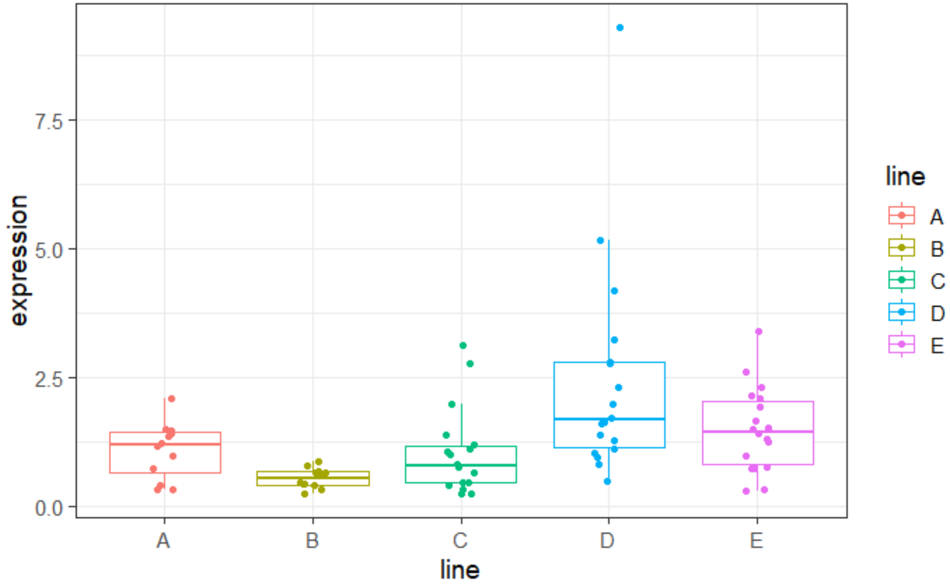
Exercise 6: One-way ANOVA: Data Exploration

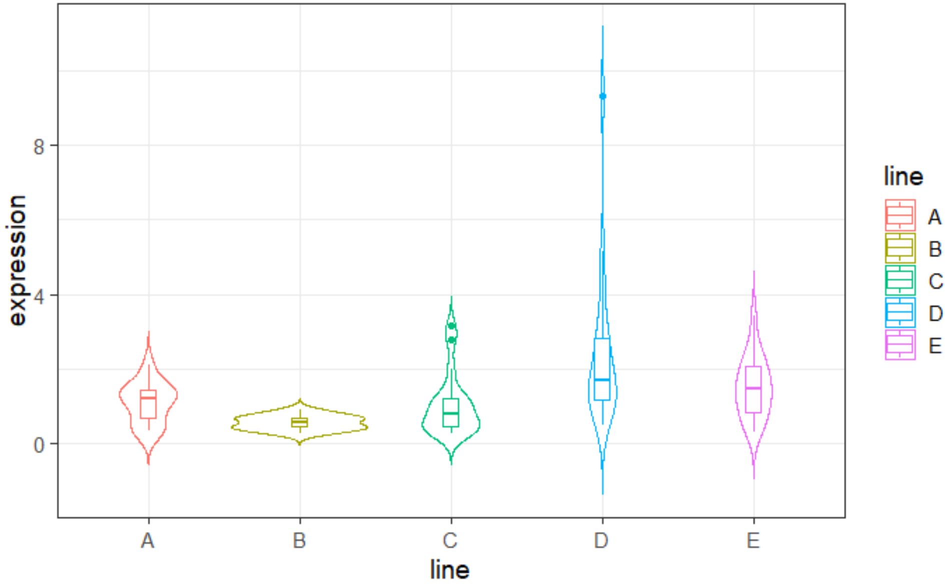
protein.expression.csv

- **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 : Data Exploration - *Answers*

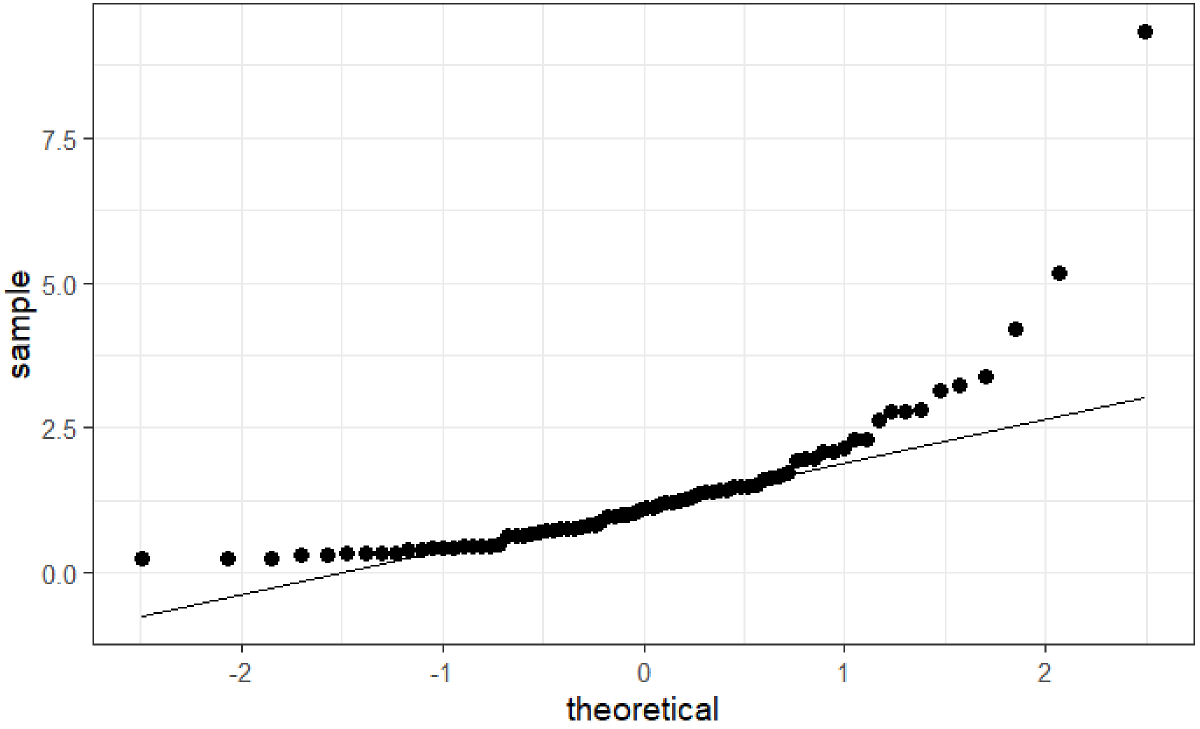
```
protein %>%
  ggplot(aes(x=line, y=expression, colour=line))+
  geom_boxplot(outlier.shape = NA)+
  geom_jitter(height=0, width=0.1)

protein %>%
  ggplot(aes(x=line, y=expression, colour=line))+
  geom_violin(trim=FALSE)+
  geom_boxplot(width=0.1)
```





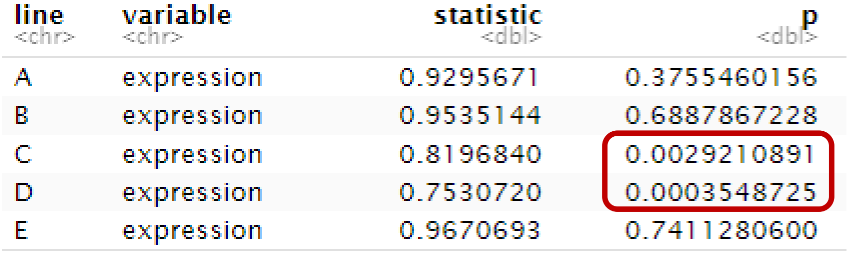
Exercise 6: One-way ANOVA – *Answers*

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

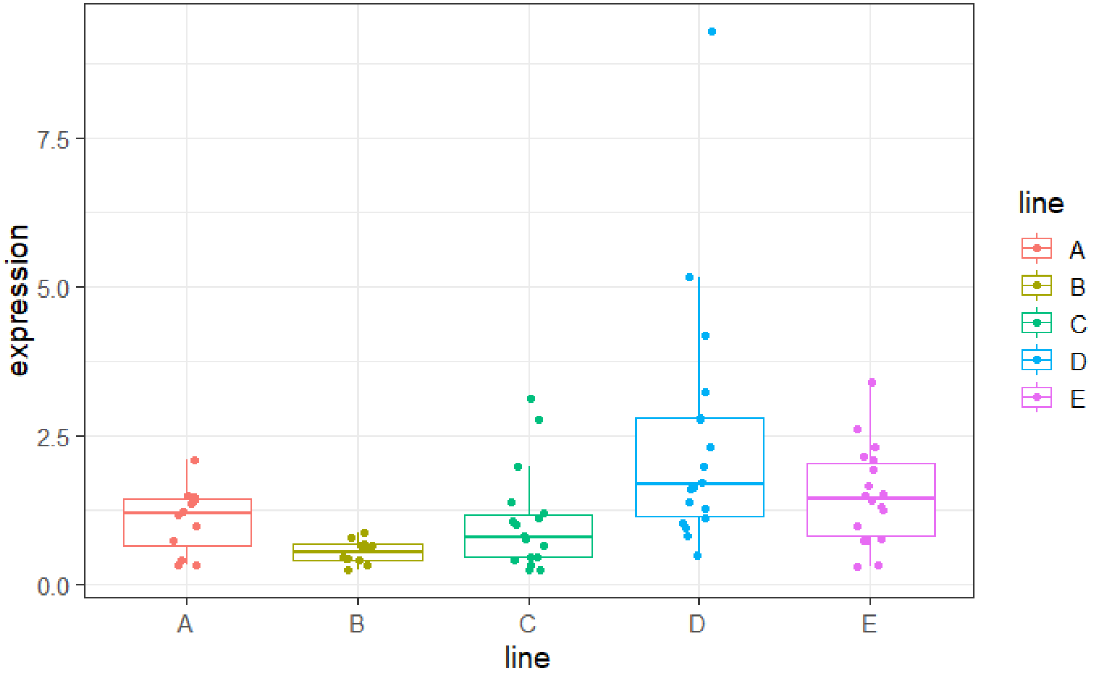


Exercise 6: One-way ANOVA – *Answers. What do we do now?*

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

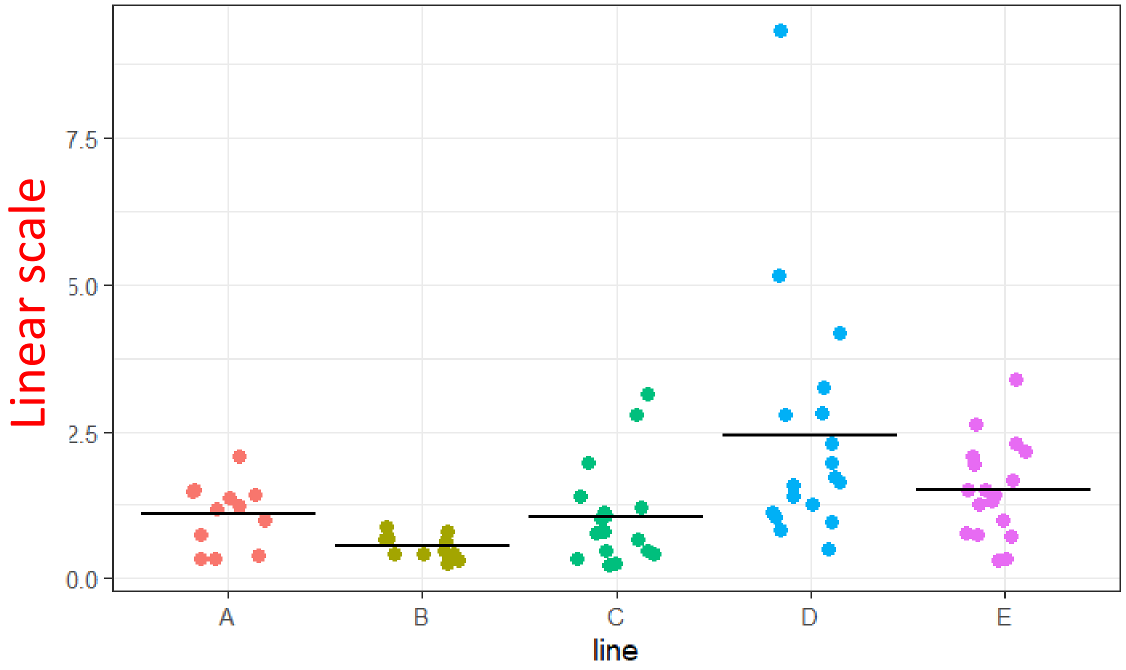


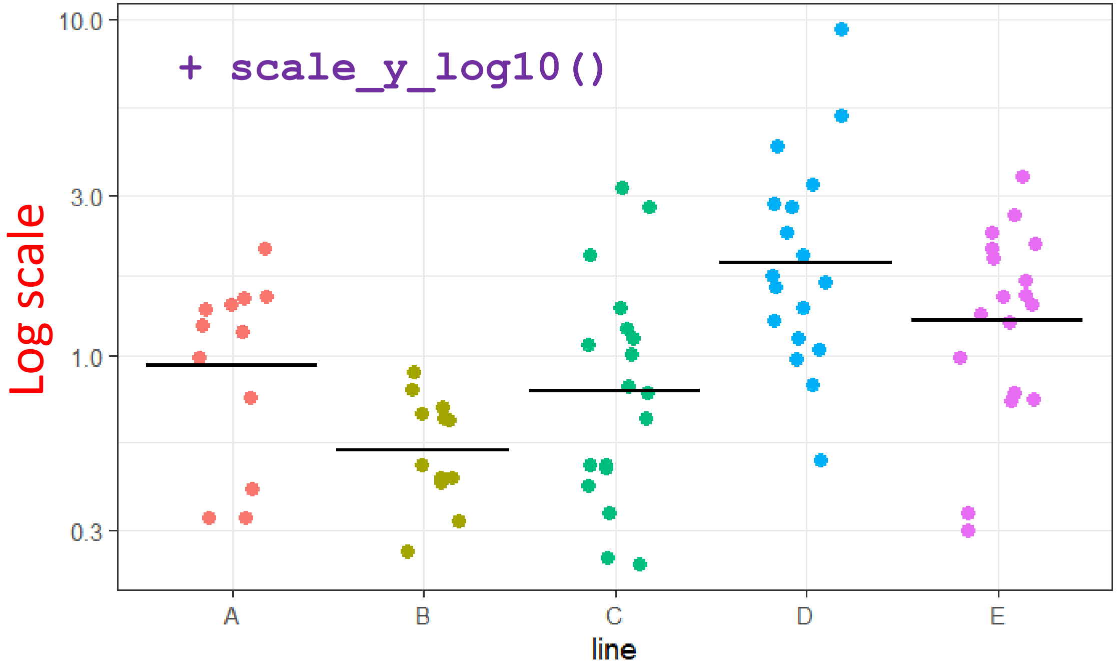
line <chr>	variable <chr>	statistic <dbl>	p <dbl>
A	expression	0.9295671	0.3755460156
B	expression	0.9535144	0.6887867228
C	expression	0.8196840	0.0029210891
D	expression	0.7530720	0.0003548725
E	expression	0.9670693	0.7411280600



One-way ANOVA: change of scale

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



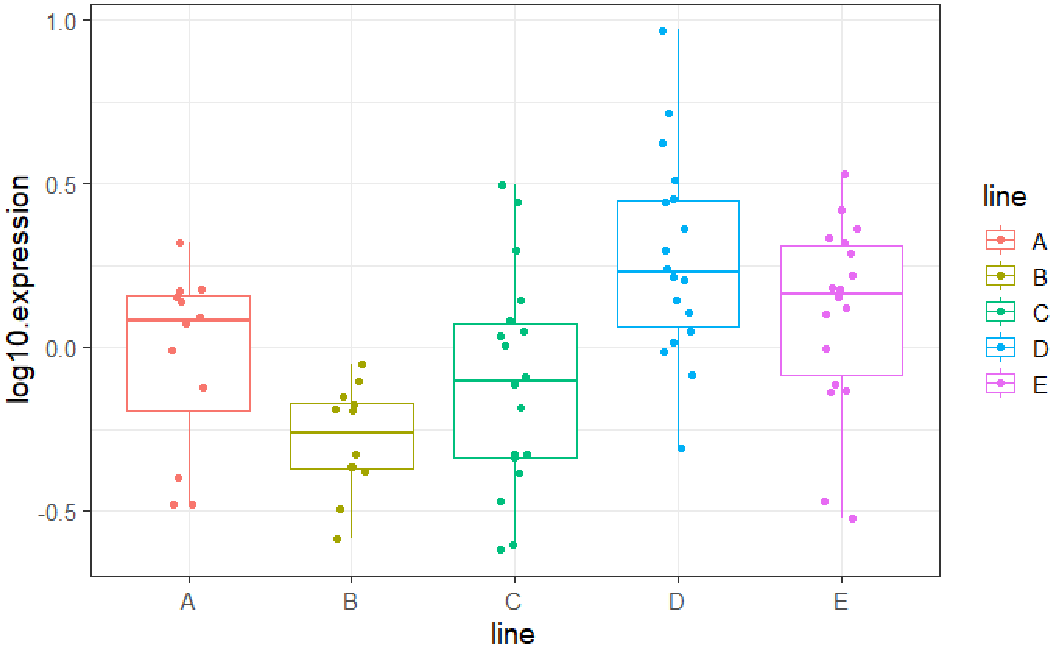


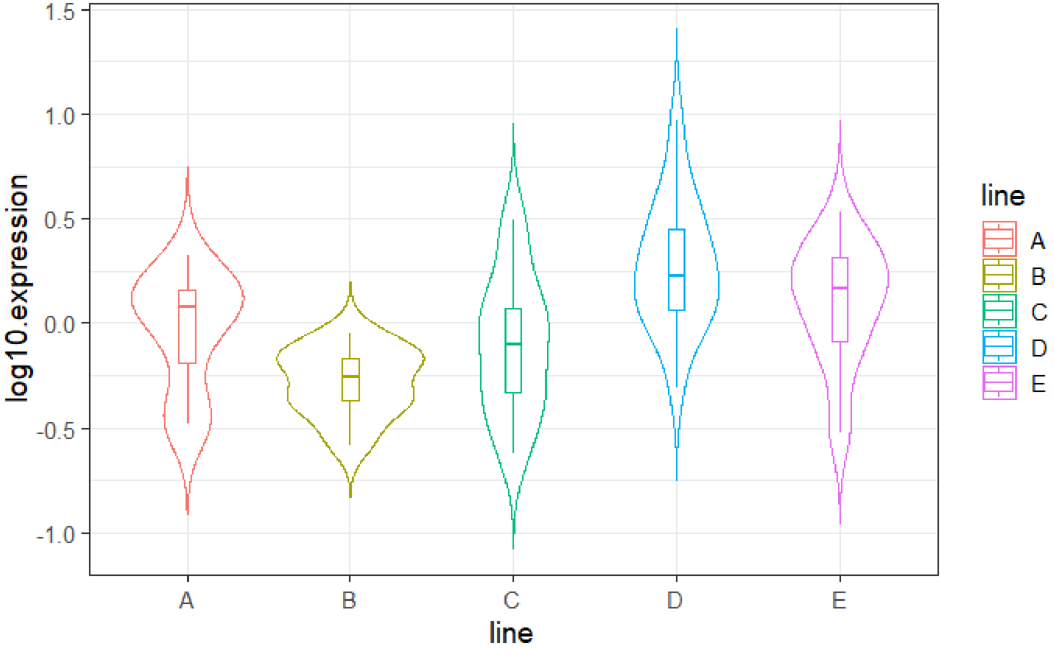
```
+ scale_y_log10()
```

```
protein %>%
  mutate(log10.expression=log10(expression)) -> protein
```

One-way ANOVA: change of scale

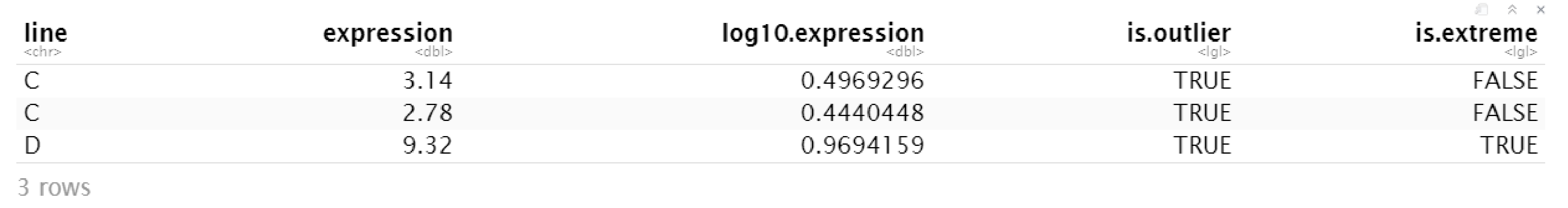
```
protein %>%
  ggplot(aes(x=line, y=log10.expression, colour=line))+
       geom_boxplot(outlier.shape = NA)+
       geom_jitter(height=0, width=0.1)

protein %>%
  ggplot(aes(x=line, y=log10.expression, colour=line))+
        geom_violin(trim=FALSE)+
        geom_boxplot(width=0.1)
```





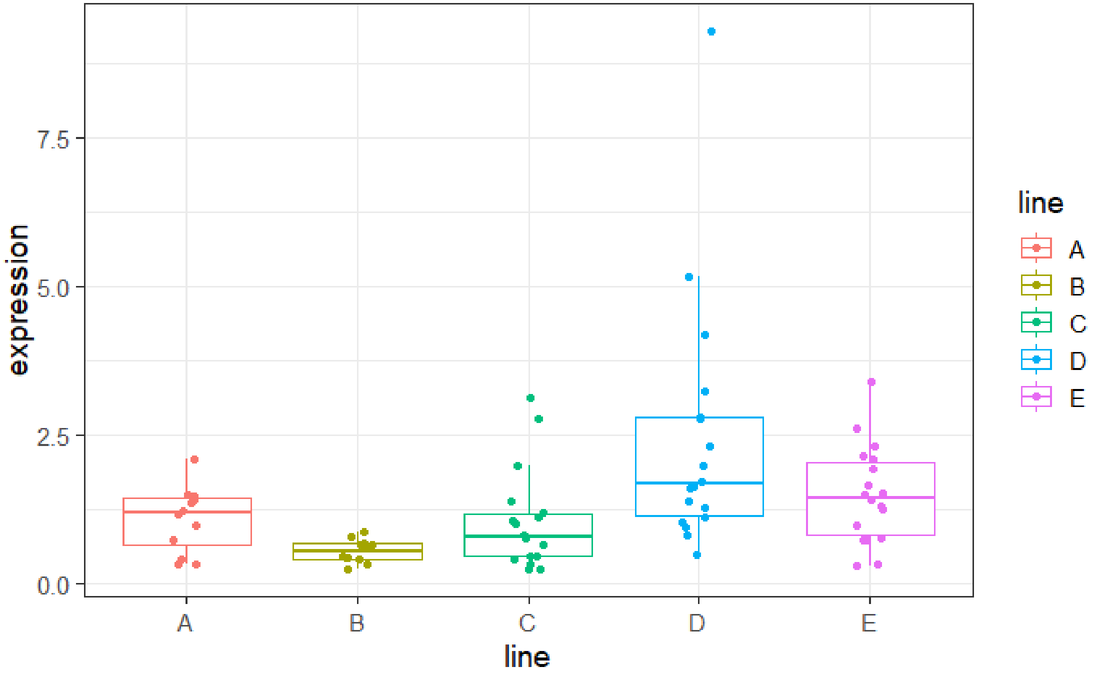
One-way ANOVA – *Outliers identification*

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



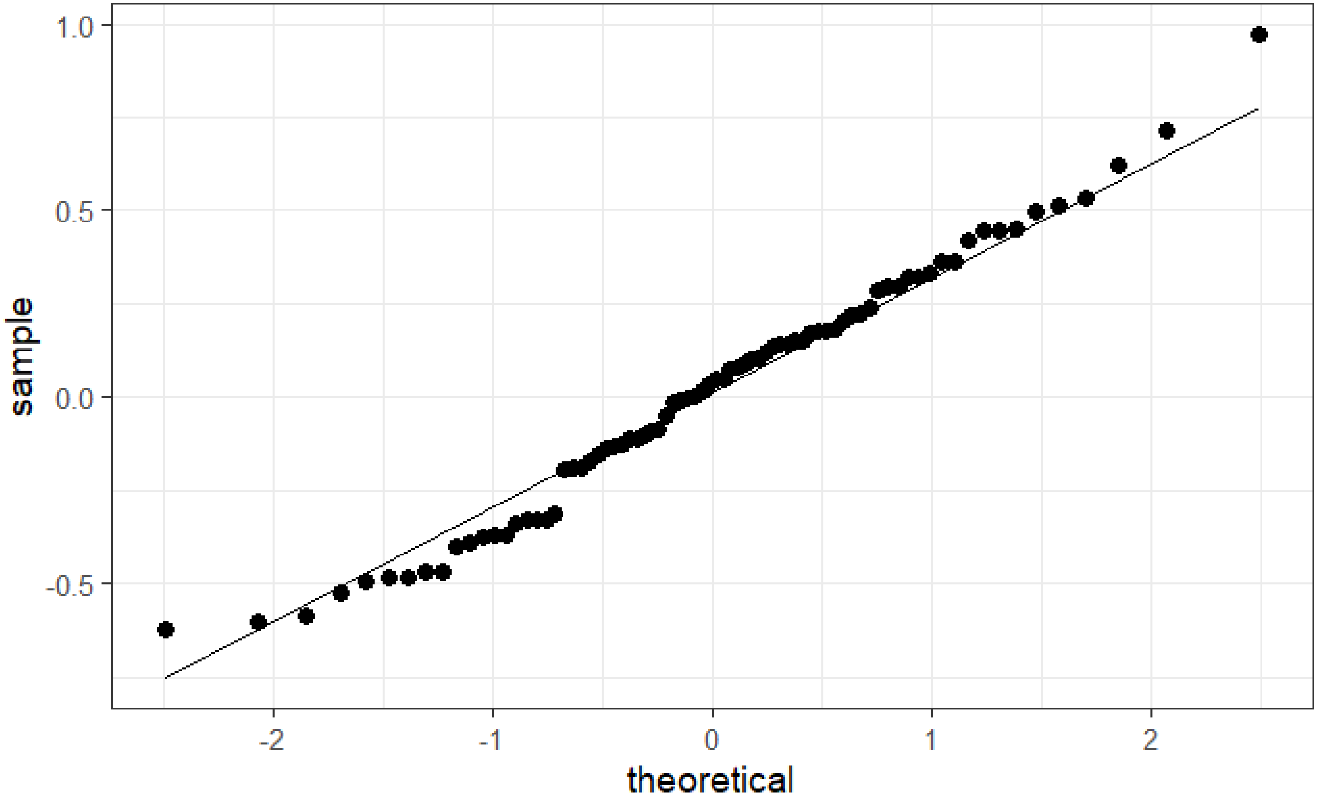
line <chr>	expression <dbl>	log10.expression <dbl>	is.outlier <lgl>	is.extreme <lgl>
C	3.14	0.4969296	TRUE	FALSE
C	2.78	0.4440448	TRUE	FALSE
D	9.32	0.9694159	TRUE	TRUE

3 rows

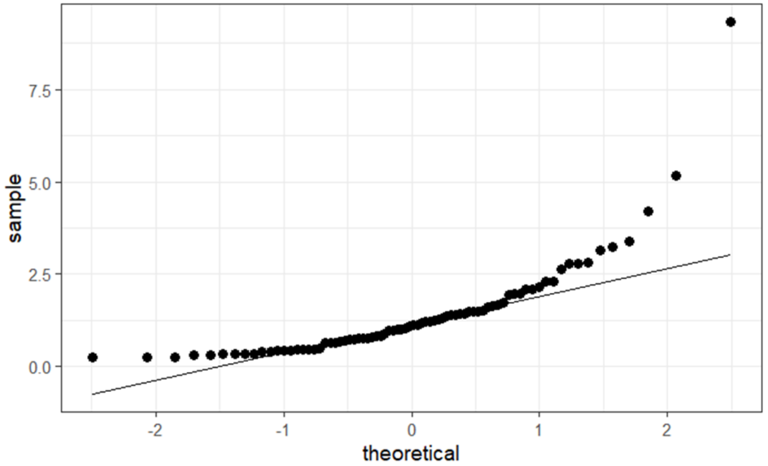


One-way ANOVA: change of scale

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



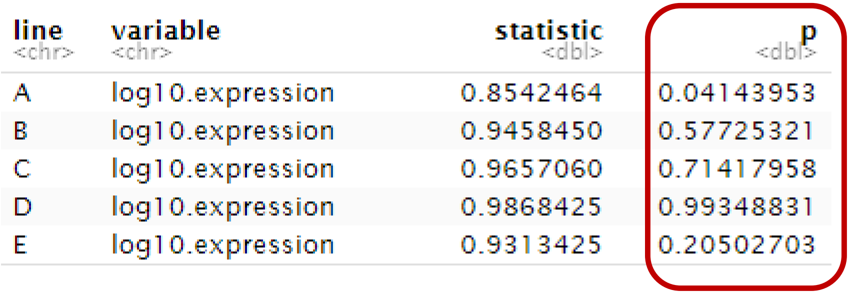
Before log-transformation



First assumption ✓

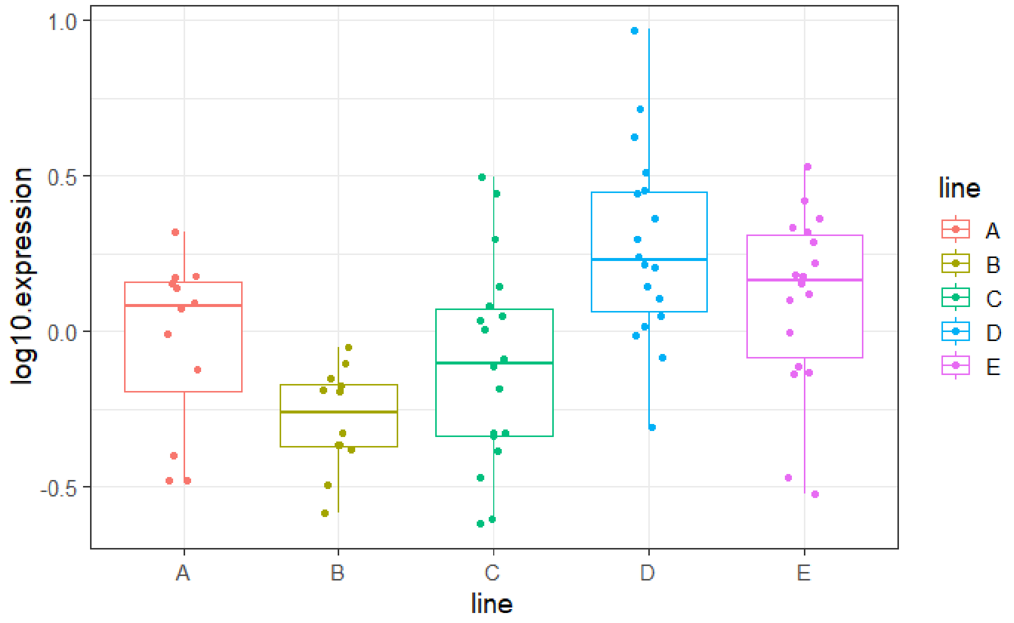
Assumptions of Parametric Data

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

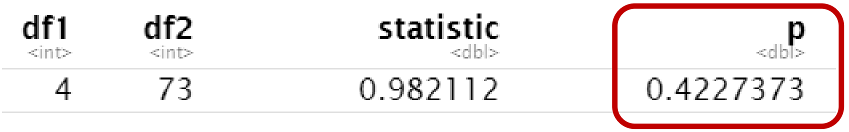


```
line   variable          statistic           p
<chr>  <chr>                 <dbl>       <dbl>
A      log10.expression  0.8542464  0.04143953
B      log10.expression  0.9458450  0.57725321
C      log10.expression  0.9657060  0.71417958
D      log10.expression  0.9868425  0.99348831
E      log10.expression  0.9313425  0.20502703
```

First assumption ✓ish



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



```
df1   df2   statistic          p
<int> <int>     <dbl>      <dbl>
  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 #

Default

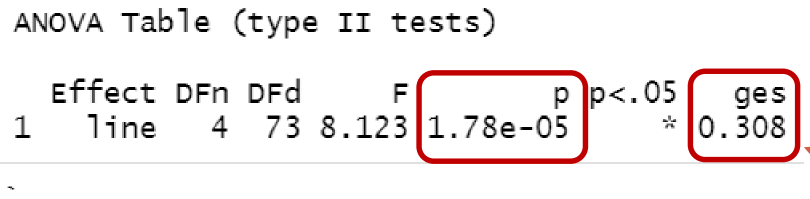
```
data %>%
 emmeans_test(y~x, p.adjust.method="bonferroni")
```

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

Have a go!

Analysis of variance

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

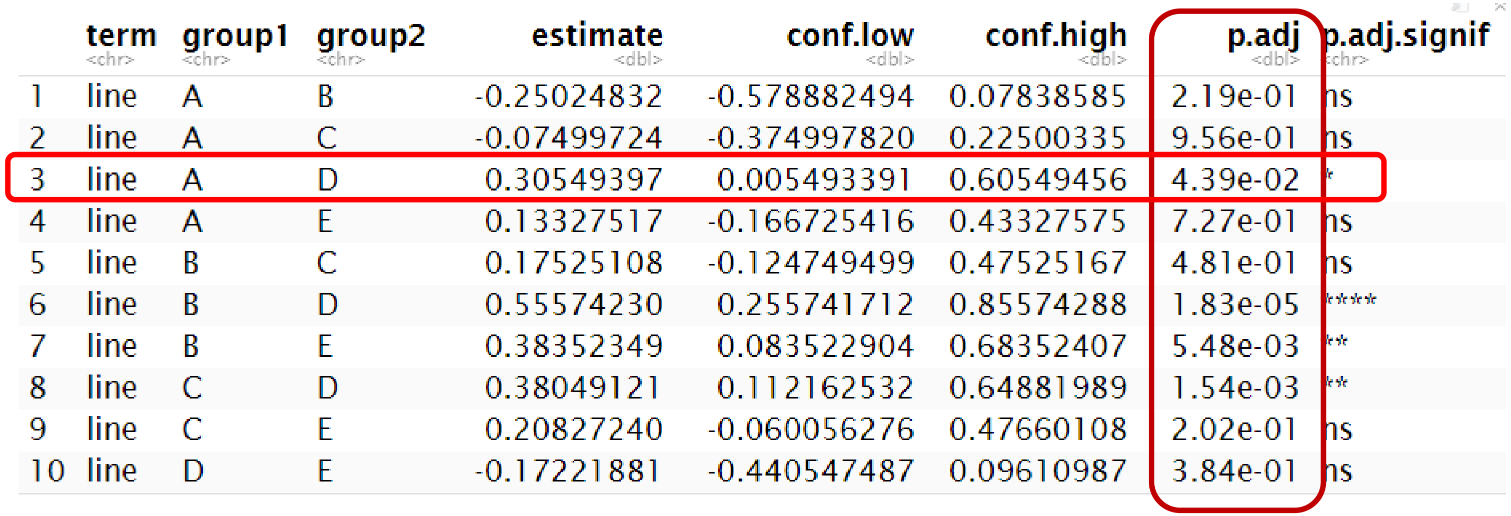


```
ANOVA Table (type II tests)

  Effect DFn DFd     F        p p<.05   ges
1   line   4  73 8.123 1.78e-05     * 0.308
```

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

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

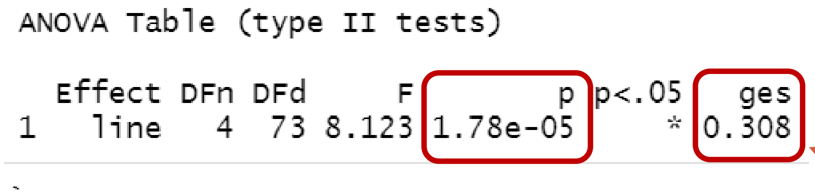
Tukey correction



	term <chr>	group1 <chr>	group2 <chr>	estimate <dbl>	conf.low <dbl>	conf.high <dbl>	p.adj <dbl>	p.adj.signif <chr>
1	line	A	B	-0.25024832	-0.578882494	0.07838585	2.19e-01	ns
2	line	A	C	-0.07499724	-0.374997820	0.22500335	9.56e-01	ns
3	line	A	D	0.30549397	0.005493391	0.60549456	4.39e-02	*
4	line	A	E	0.13327517	-0.166725416	0.43327575	7.27e-01	ns
5	line	B	C	0.17525108	-0.124749499	0.47525167	4.81e-01	ns
6	line	B	D	0.55574230	0.255741712	0.85574288	1.83e-05	****
7	line	B	E	0.38352349	0.083522904	0.68352407	5.48e-03	**
8	line	C	D	0.38049121	0.112162532	0.64881989	1.54e-03	**
9	line	C	E	0.20827240	-0.060056276	0.47660108	2.02e-01	ns
10	line	D	E	-0.17221881	-0.440547487	0.09610987	3.84e-01	ns

Analysis of variance

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

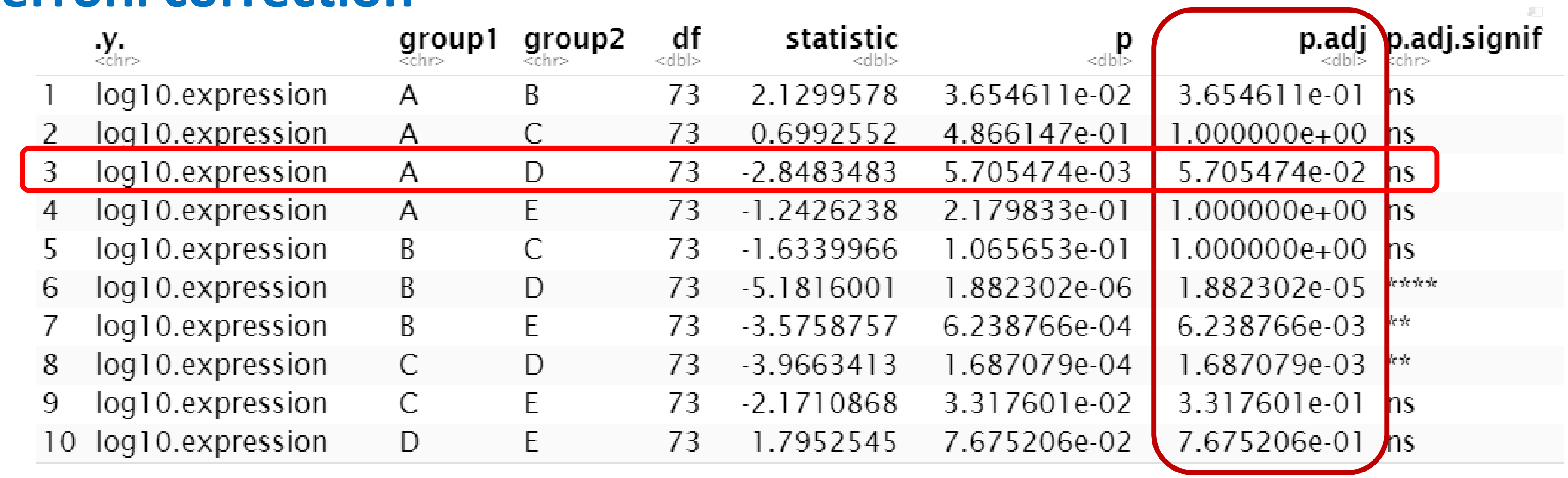


```
ANOVA Table (type II tests)

  Effect DFn DFd     F        p p<.05   ges
1   line   4  73 8.123 1.78e-05     * 0.308
```

generalised **e**ffect **s**ize (Eta squared η^2) = R^2 ish

```
protein %>%
  emmeans_test(log10.expression ~ line, p.adjust.method = "bonferroni")
```

Bonferroni correction



	.y. <chr>	group1 <chr>	group2 <chr>	df <dbl>	statistic <dbl>	p <dbl>	p.adj <dbl>	p.adj.signif <chr>
1	log10.expression	A	B	73	2.1299578	3.654611e-02	3.654611e-01	ns
2	log10.expression	A	C	73	0.6992552	4.866147e-01	1.000000e+00	ns
3	log10.expression	A	D	73	-2.8483483	5.705474e-03	5.705474e-02	ns
4	log10.expression	A	E	73	-1.2426238	2.179833e-01	1.000000e+00	ns
5	log10.expression	B	C	73	-1.6339966	1.065653e-01	1.000000e+00	ns
6	log10.expression	B	D	73	-5.1816001	1.882302e-06	1.882302e-05	****
7	log10.expression	B	E	73	-3.5758757	6.238766e-04	6.238766e-03	**
8	log10.expression	C	D	73	-3.9663413	1.687079e-04	1.687079e-03	**
9	log10.expression	C	E	73	-2.1710868	3.317601e-02	3.317601e-01	ns
10	log10.expression	D	E	73	1.7952545	7.675206e-02	7.675206e-01	ns

Analysis of variance (R)

To plot confidence intervals

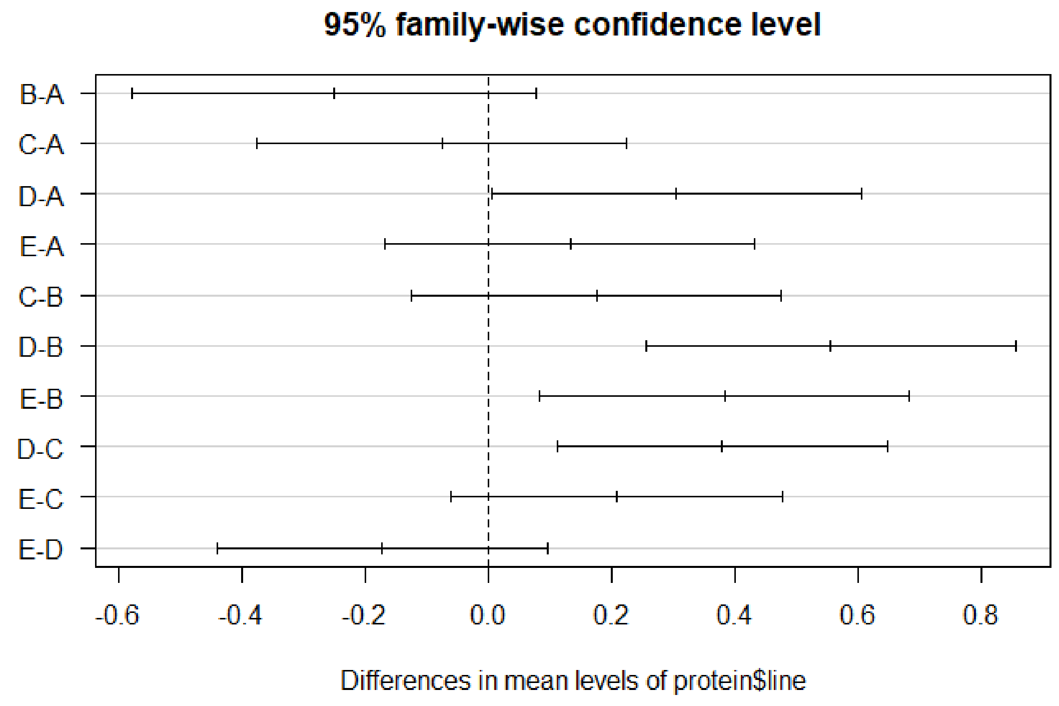
```
aov(log10.expression~line,data=protein.stack.clean) -> anova.log.protein
summary(anova.log.protein)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)    
line         4  2.691  0.6728   8.123 1.78e-05 ***
Residuals   73  6.046  0.0828                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
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)

$line
           diff          lwr        upr     p adj
B-A -0.25024832 -0.578882494 0.07838585 0.2187264
C-A -0.07499724 -0.374997820 0.22500335 0.9560187
D-A  0.30549397  0.005493391 0.60549456 0.0438762
E-A  0.13327517 -0.166725416 0.43327575 0.7265567
C-B  0.17525108 -0.124749499 0.47525167 0.4809387
D-B  0.55574230  0.255741712 0.85574288 0.0000183
E-B  0.38352349  0.083522904 0.68352407 0.0054767
D-C  0.38049121  0.112162532 0.64881989 0.0015431
E-C  0.20827240 -0.060056276 0.47660108 0.2023355
E-D -0.17221881 -0.440547487 0.09610987 0.3841989
```

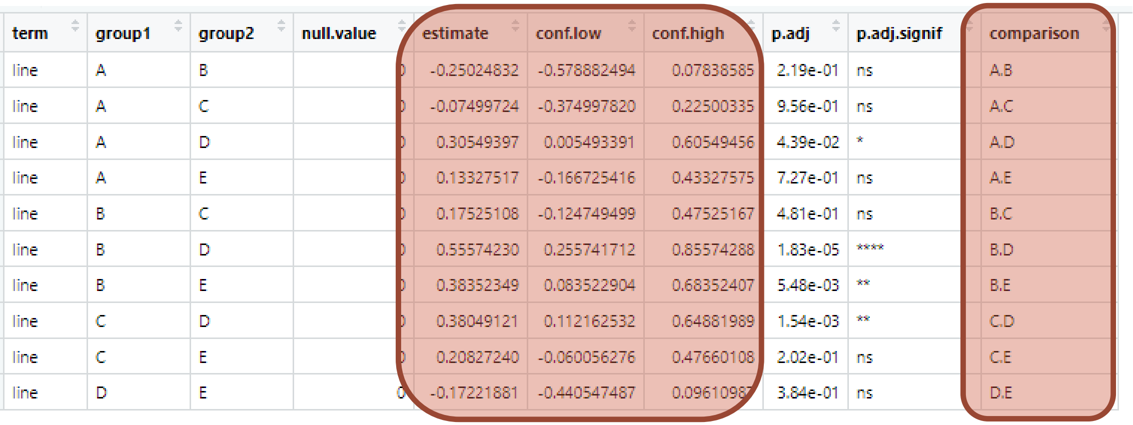
```
TukeyHSD(anova.log.protein)->tukey
plot(tukey, las=1)
```



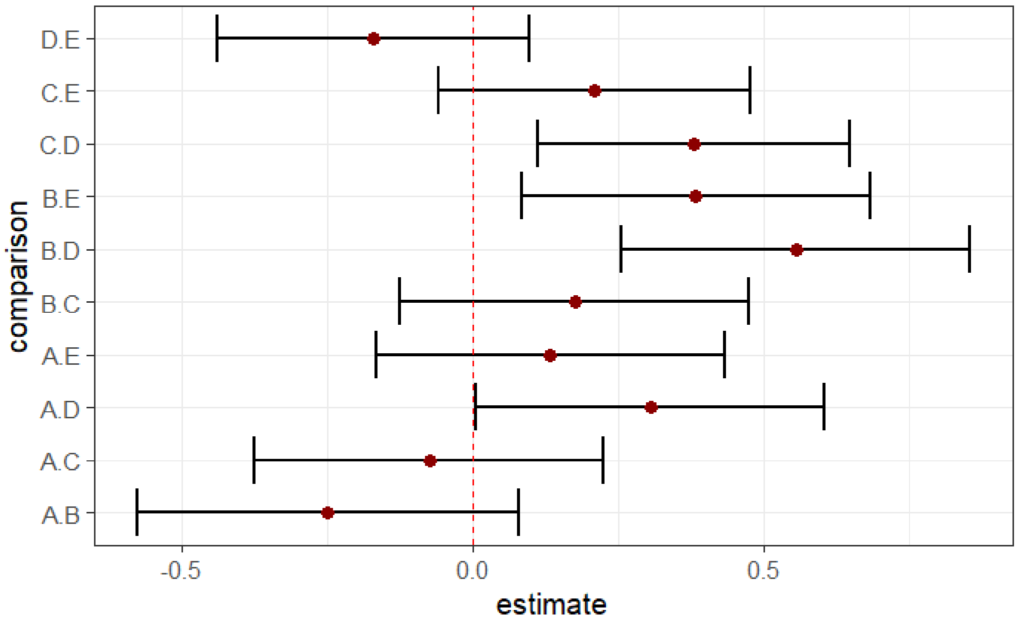
Analysis of variance (tidyverse)

To plot confidence intervals

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



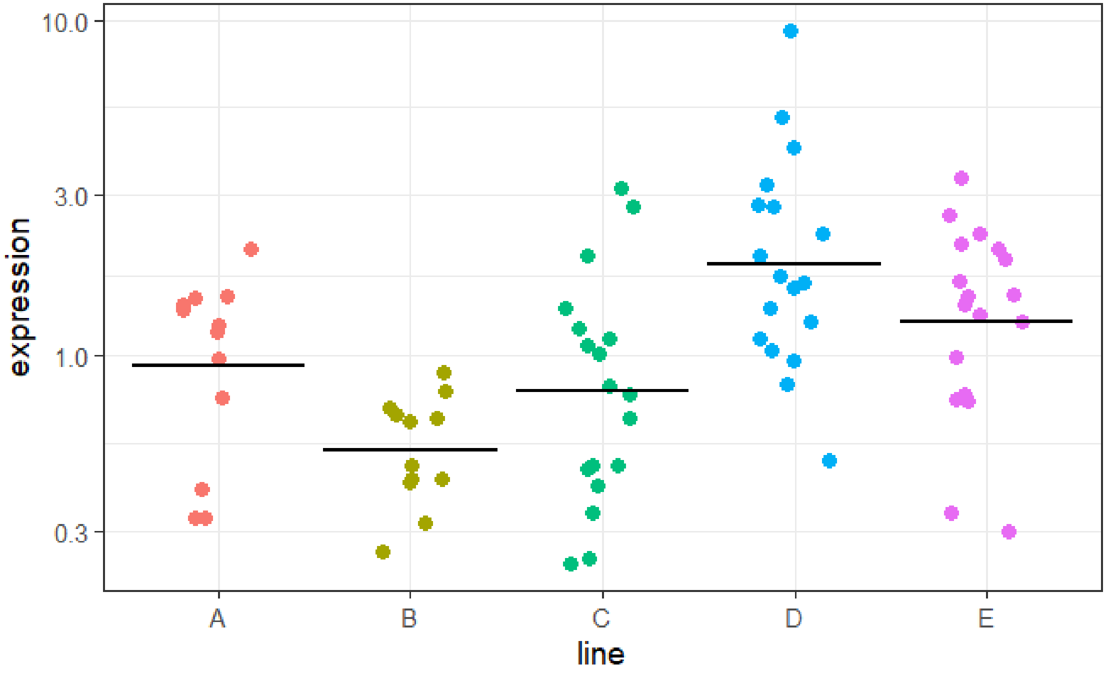
term	group1	group2	null.value	estimate	conf.low	conf.high	p.adj	p.adj.signif	comparison
line	A	B	0	-0.25024832	-0.578882494	0.07838585	2.19e-01	ns	A.B
line	A	C	0	-0.07499724	-0.374997820	0.22500335	9.56e-01	ns	A.C
line	A	D	0	0.30549397	0.005493391	0.60549456	4.39e-02	*	A.D
line	A	E	0	0.13327517	-0.166725416	0.43327575	7.27e-01	ns	A.E
line	B	C	0	0.17525108	-0.124749499	0.47525167	4.81e-01	ns	B.C
line	B	D	0	0.55574230	0.255741712	0.85574288	1.83e-05	****	B.D
line	B	E	0	0.38352349	0.083522904	0.68352407	5.48e-03	**	B.E
line	C	D	0	0.38049121	0.112162532	0.64881989	1.54e-03	**	C.D
line	C	E	0	0.20827240	-0.060056276	0.47660108	2.02e-01	ns	C.E
line	D	E	0	-0.17221881	-0.440547487	0.09610987	3.84e-01	ns	D.E



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

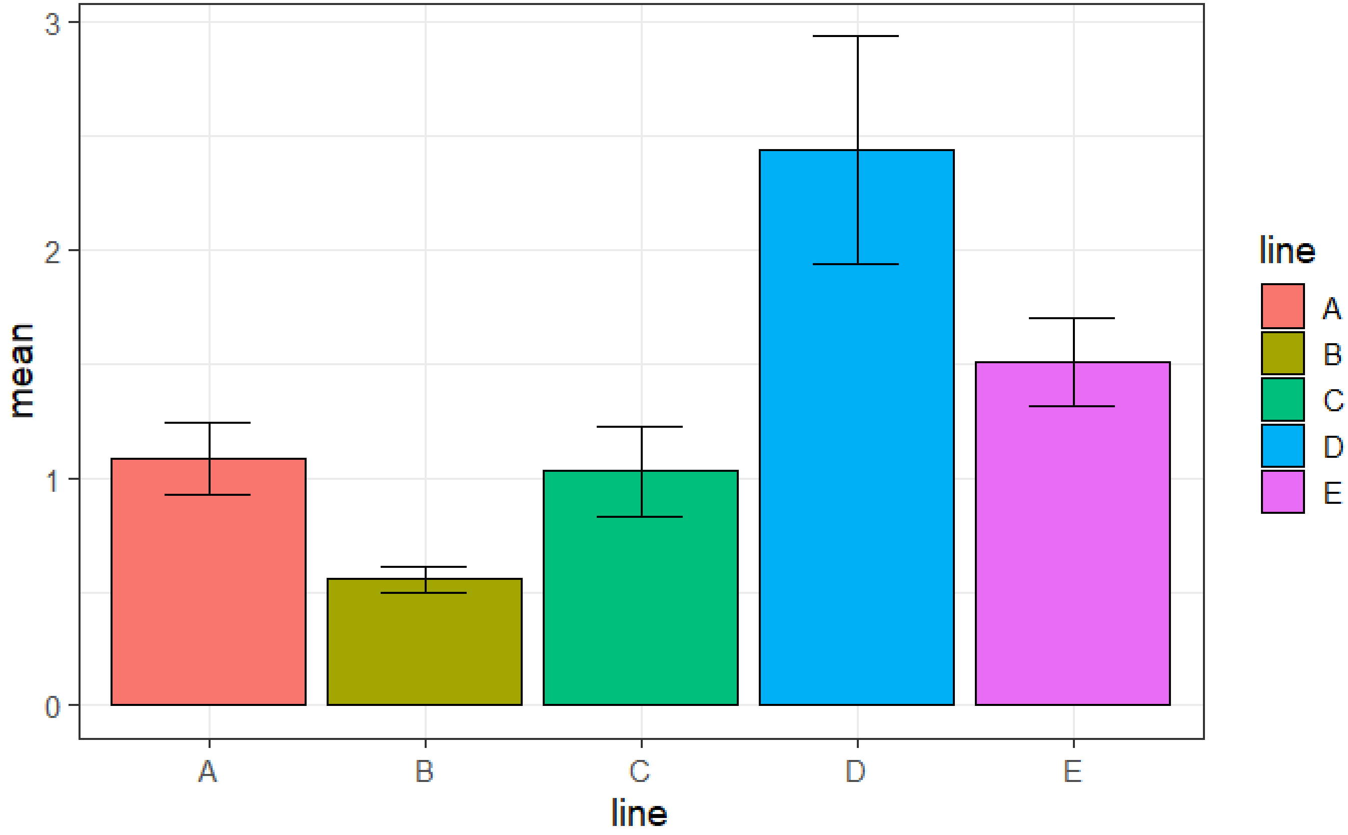
Analysis of variance

```
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()
```



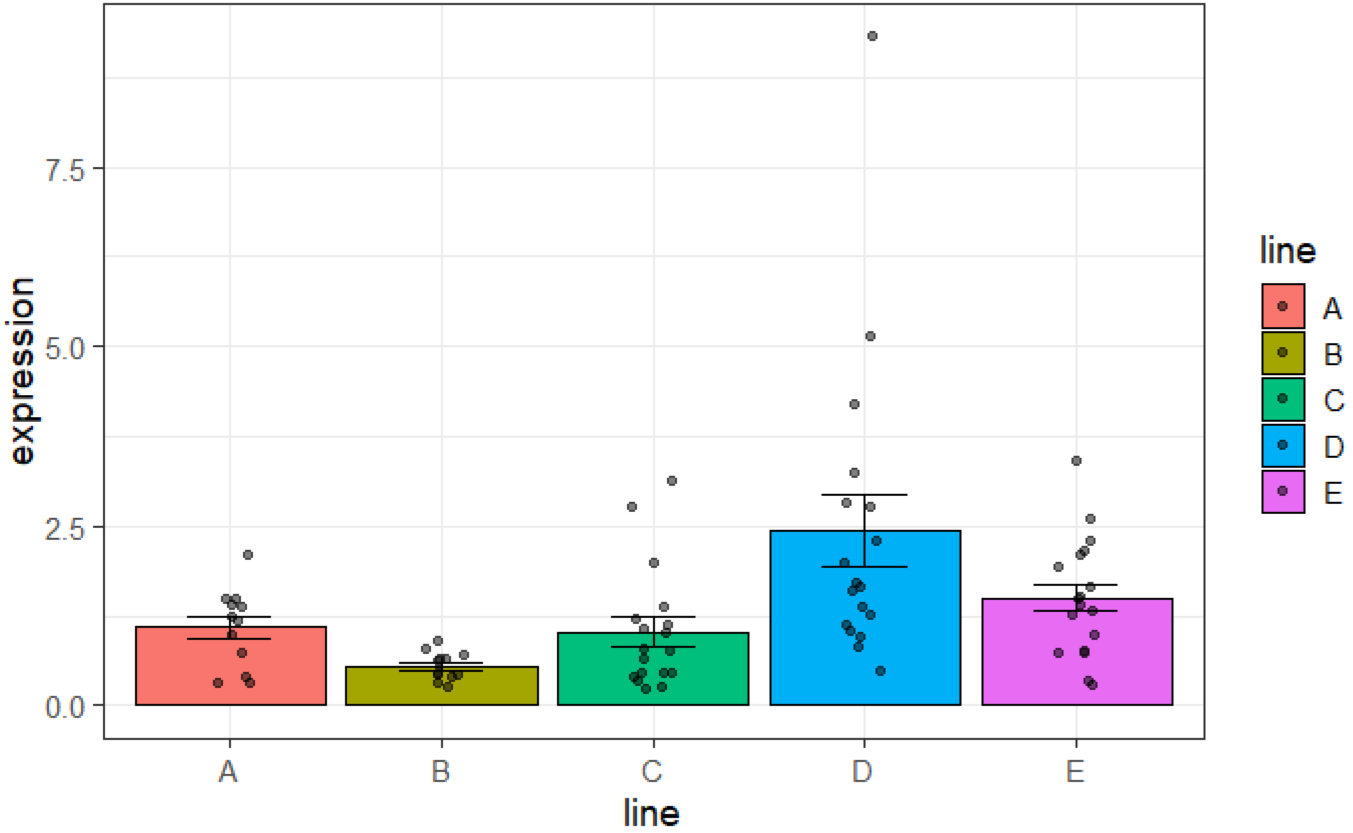
Analysis of variance

```
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)
```



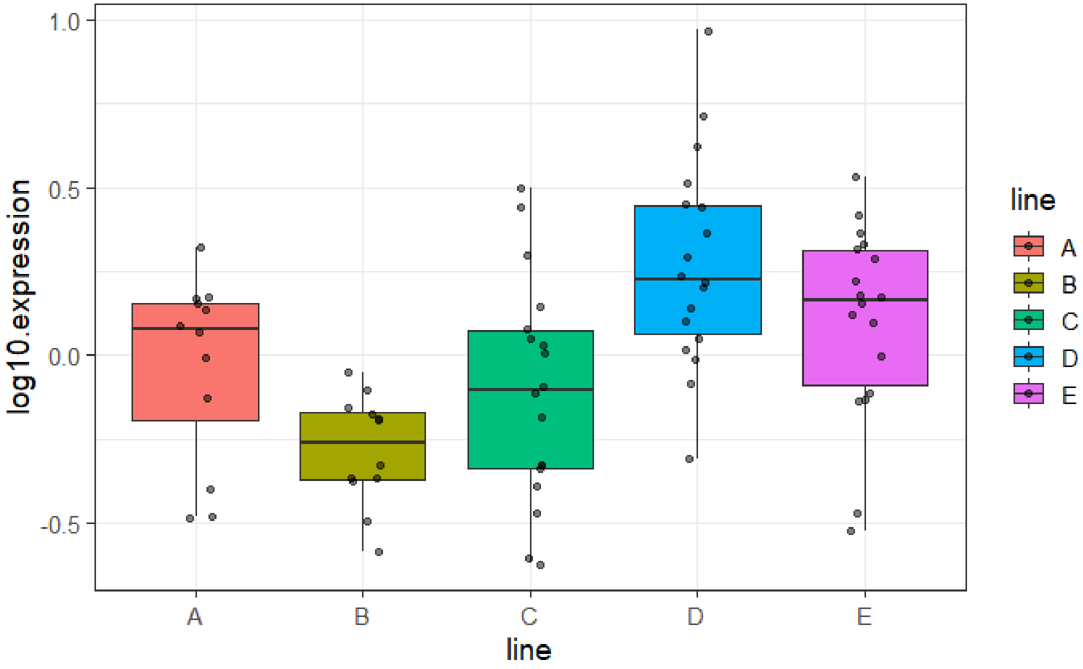
Analysis of variance

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



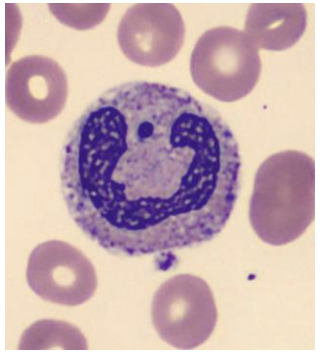
Analysis of variance

```
protein %>%
  ggplot(aes(x=line, y=log10.expression, fill=line)) +
      geom_bar(stat="summary", fun="mean", colour="black")+
      stat_summary(geom="errorbar", colour="black", width=0.4)+
      geom_jitter(heigth=0, width=0.1, alpha=0.5)
```



Exercise 7: Repeated measures ANOVA

neutrophils.long.csv



- A researcher is looking at the difference between 4 cell groups.

He has run the experiment 5 times. Within each experiment, he has neutrophils from a WT (control), a KO, a KO+Treatment 1 and a KO+Treatment2.

- **Question**: Is there a difference between KO with/without treatment and WT?

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

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

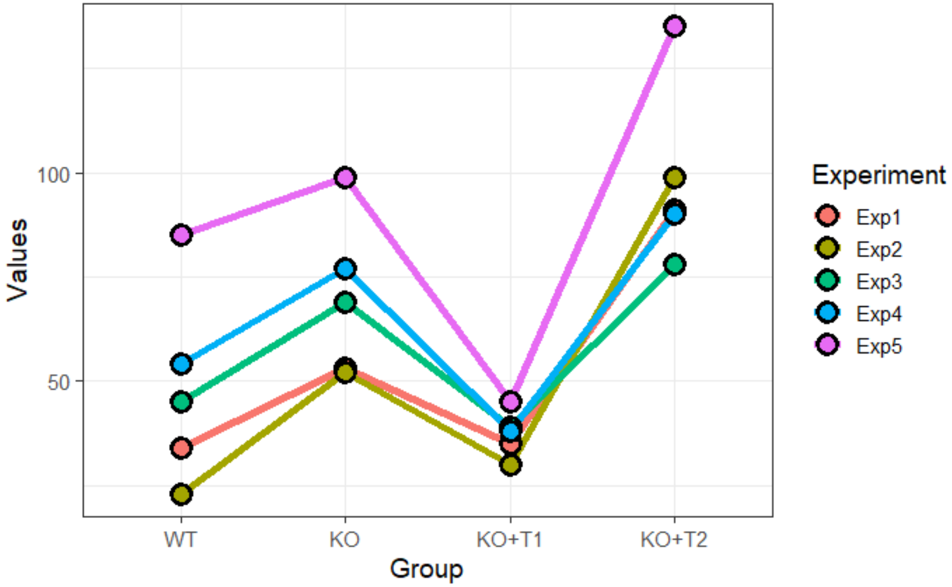
- Choose a graphical presentation consistent with the experimental design

Exercise 7: Repeated measures ANOVA

neutrophils.long.csv

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

```
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"))
```

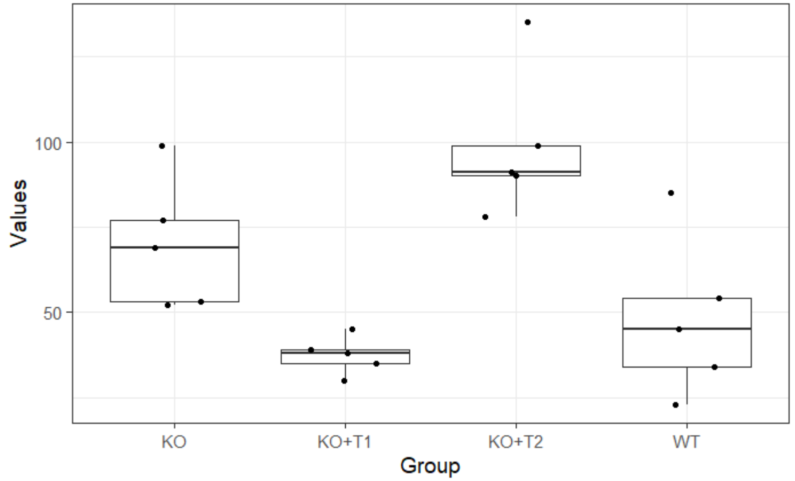


Exercise 7: Repeated measures ANOVA

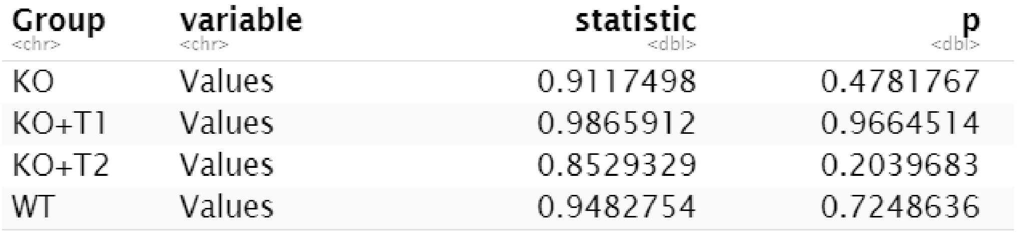
neutrophils.long.csv

- Check the first assumption

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



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



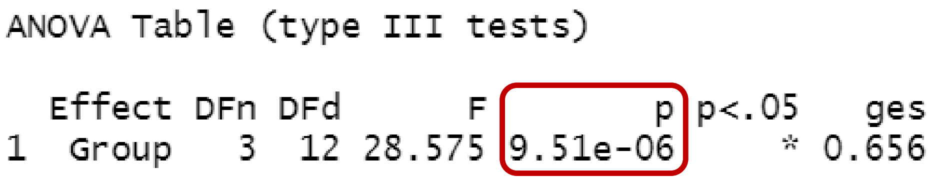
```
Group    variable                 statistic            p
<chr>    <chr>                        <dbl>        <dbl>
KO       Values                   0.9117498    0.4781767
KO+T1    Values                   0.9865912    0.9664514
KO+T2    Values                   0.8529329    0.2039683
WT       Values                   0.9482754    0.7248636
```

Exercise 7: Repeated measures ANOVA

neutrophils.long.csv

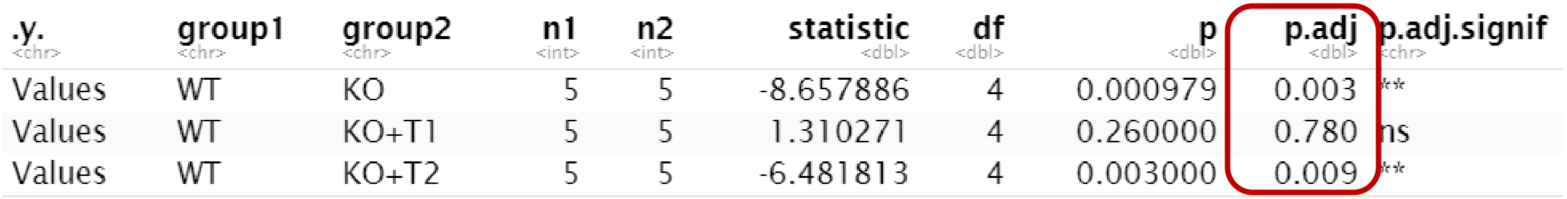
- Run the repeated measures ANOVA and post-hoc tests

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



```
ANOVA Table (type III tests)

  Effect DFn DFd      F        p p<.05   ges
1  Group   3  12 28.575 9.51e-06     * 0.656
```

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



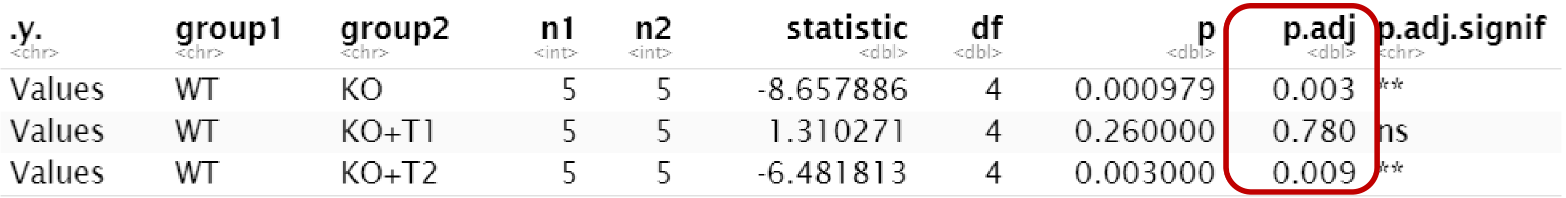
.y.	group1	group2	n1	n2	statistic	df	p	p.adj	p.adj.signif
<chr>	<chr>	<chr>	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
Values	WT	KO	5	5	-8.657886	4	0.000979	0.003	**
Values	WT	KO+T1	5	5	1.310271	4	0.260000	0.780	ns
Values	WT	KO+T2	5	5	-6.481813	4	0.003000	0.009	**

Exercise 7: Repeated measures ANOVA

neutrophils.long.csv

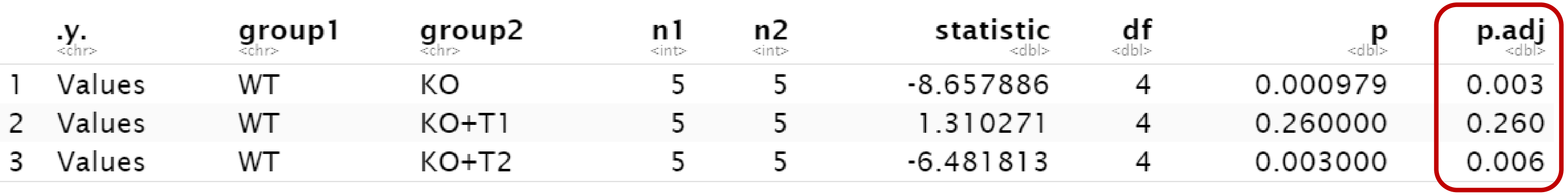
- Run the repeated measures ANOVA and post-hoc tests

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



```
.y.     group1  group2  n1  n2   statistic  df         p   p.adj  p.adj.signif
<chr>   <chr>   <chr>  <int> <int>  <dbl>  <dbl>    <dbl>   <dbl>  <chr>
Values  WT      KO       5   5   -8.657886   4  0.000979   0.003  **
Values  WT      KO+T1    5   5    1.310271   4  0.260000   0.780  ns
Values  WT      KO+T2    5   5   -6.481813   4  0.003000   0.009  **
```

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



```
   .y.     group1  group2  n1  n2   statistic  df         p   p.adj
   <chr>   <chr>   <chr>  <int> <int>  <dbl>  <dbl>    <dbl>   <dbl>
1  Values  WT      KO       5   5   -8.657886   4  0.000979   0.003
2  Values  WT      KO+T1    5   5    1.310271   4  0.260000   0.260
3  Values  WT      KO+T2    5   5   -6.481813   4  0.003000   0.006
```

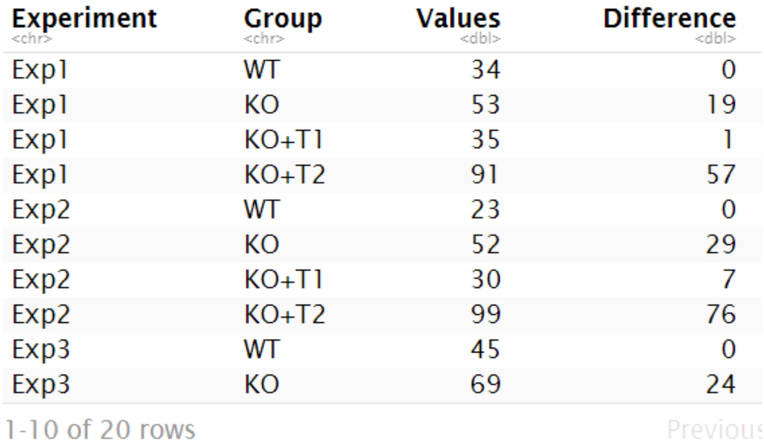
Tukey ☹

Exercise 7: Repeated measures ANOVA

neutrophils.long.csv

- Choose a graphical presentation consistent with the experimental design

```
neutrophils.long %>%
  group_by(Experiment) %>%
    mutate(Difference=Values-Values[Group=="WT"]) %>%
      ungroup() -> neutrophils.long
```



Experiment	Group	Values	Difference
<chr>	<chr>	<dbl>	<dbl>
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	99	76
Exp3	WT	45	0
Exp3	KO	69	24

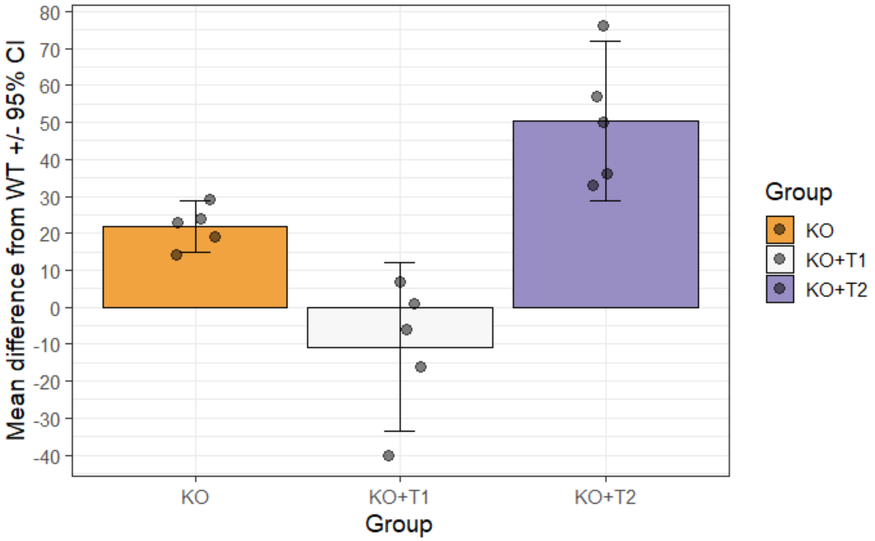
1-10 of 20 rows Previous

Exercise 7: Repeated measures ANOVA

neutrophils.long.csv

- Choose a graphical presentation consistent with the experimental design

```
neutrophils.long %>%
  filter(Group !="WT") %>%
    ggplot(aes(Group, Difference, fill=Group)) +
      geom_bar(stat = "summary", fun="mean", colour="black")+
      stat_summary(geom="errorbar", fun.data=mean_cl_normal, width=0.15)+
      geom_jitter(height = 0, width=0.1, alpha=0.5, size=3)+
      ylab("Mean difference from WT +/- 95% CI")+
      scale_y_continuous(breaks=seq(from=-40, by=10, to=80))+
      scale_fill_brewer(palette = "PuOr")
```



Comparison between more than 2 groups

Two factors = Two predictors

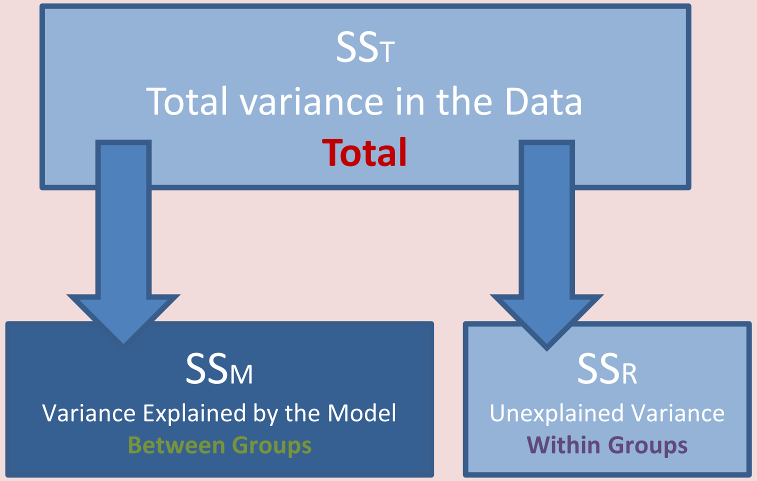
Two-Way ANOVA

Two-way Analysis of Variance (Factorial ANOVA)

Source of variation	Sum of Squares	Df	Mean Square	F	p-value
Variable A (Between Groups)	2.665	4	0.6663	8.42	<0.0001
Within Groups (Residual)	5.775	73	0.0791		
Total	8.44	77			

Source of variation	Sum of Squares	Df	Mean Square	F	p-value
Variable A * Variable B	1978	2	989.1	F (2, 42) = 11.91	P < 0.0001
Variable B (Between groups)	3332	2	1666	F (2, 42) = 20.07	P < 0.0001
Variable A (Between groups)	168.8	1	168.8	F (1, 42) = 2.032	P = 0.1614
Residuals	3488	42	83.04		

One-way ANOVA= 1 predictor variable

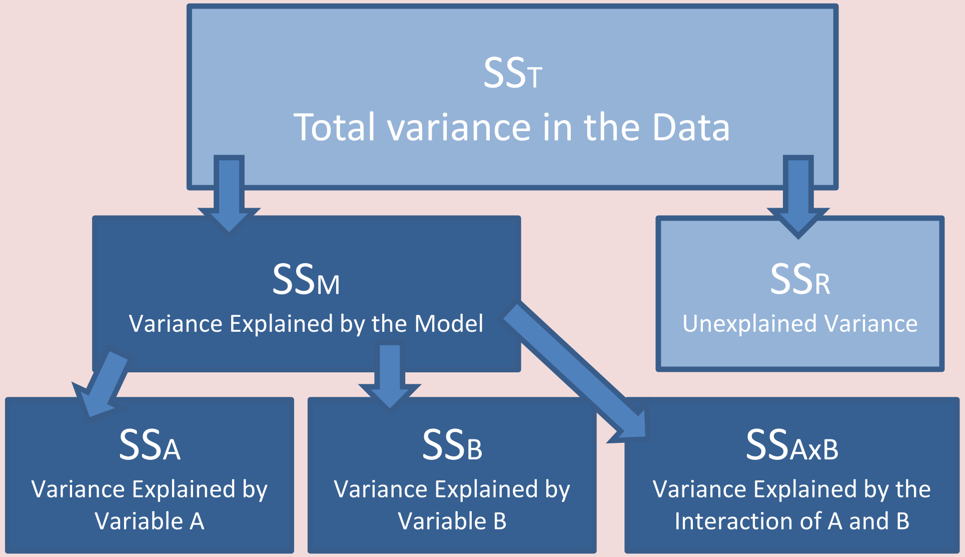


SS_T
Total variance in the Data
Total

SS_M
Variance Explained by the Model
Between Groups

SS_R
Unexplained Variance
Within Groups

2-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

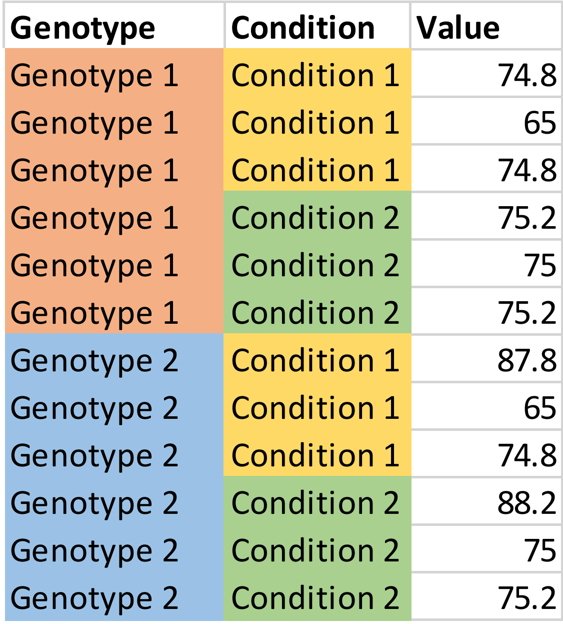
SS_A
Variance Explained by Variable A

SS_B
Variance Explained by Variable B

SS_{AxB}
Variance Explained by the Interaction of A and B

Two-way Analysis of Variance

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

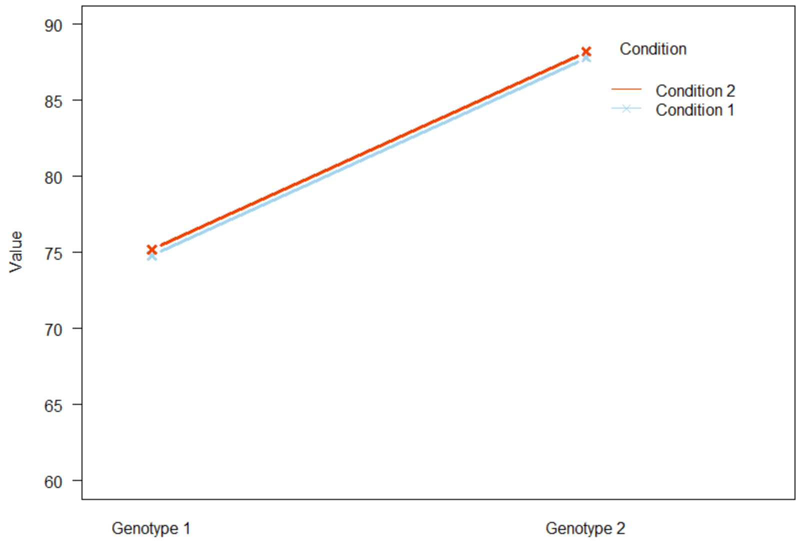


Genotype	Condition	Value
Genotype 1	Condition 1	74.8
Genotype 1	Condition 1	65
Genotype 1	Condition 1	74.8
Genotype 1	Condition 2	75.2
Genotype 1	Condition 2	75
Genotype 1	Condition 2	75.2
Genotype 2	Condition 1	87.8
Genotype 2	Condition 1	65
Genotype 2	Condition 1	74.8
Genotype 2	Condition 2	88.2
Genotype 2	Condition 2	75
Genotype 2	Condition 2	75.2

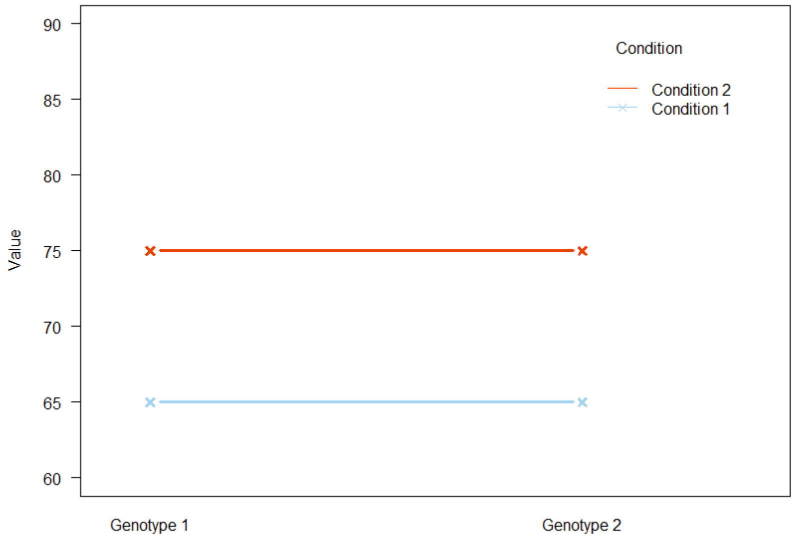
Two-way Analysis of Variance

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

Single Effect



Genotype Effect

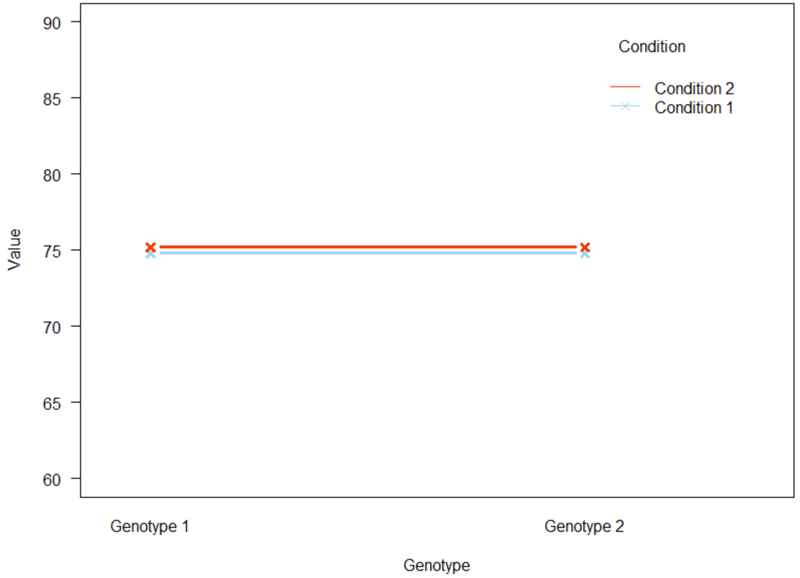


Condition Effect

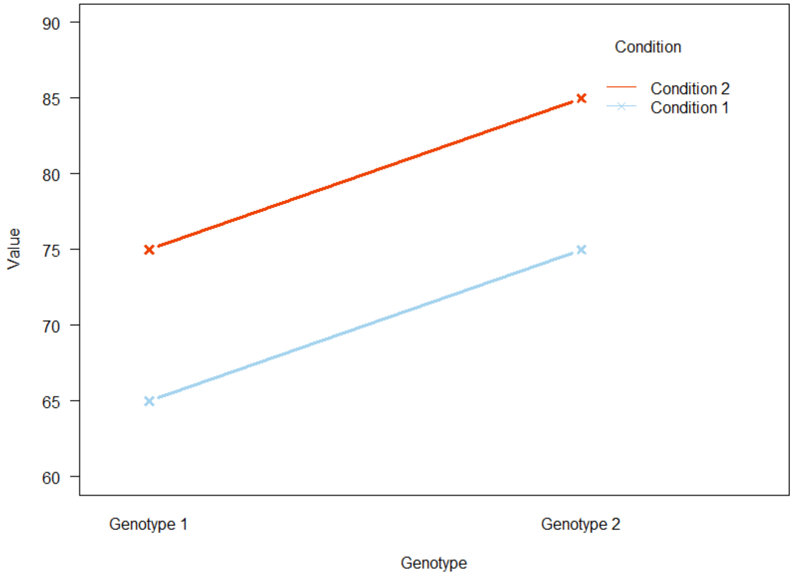
Two-way Analysis of Variance

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

Zero or Both Effect



Zero Effect

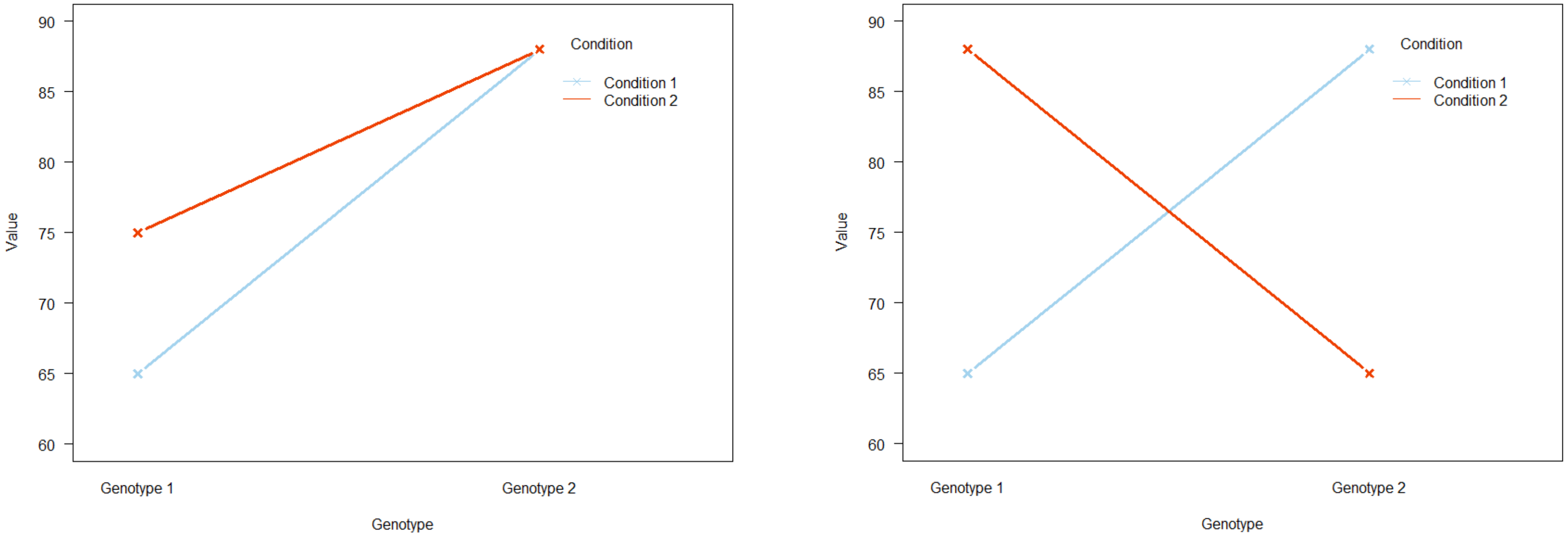


Both Effect

Two-way Analysis of Variance

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

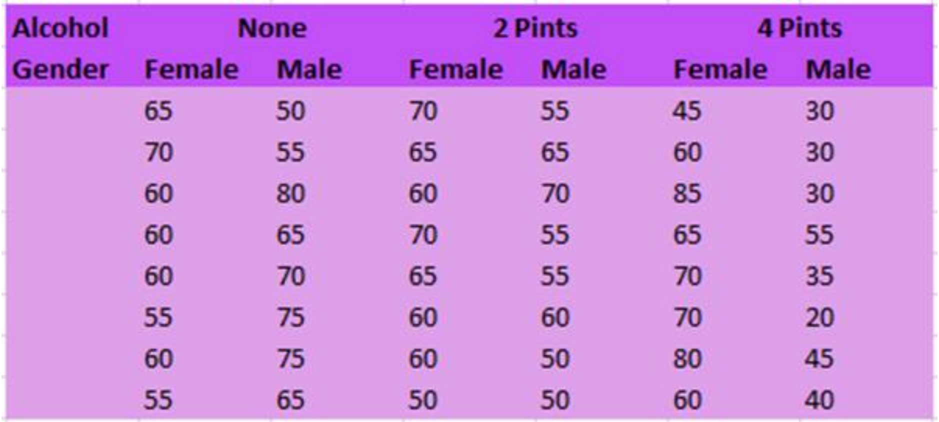
Interaction



Two-way Analysis of Variance

Example: goggles.csv

- The ‘beer-goggle’ effect



Alcohol	None		2 Pints		4 Pints	
Gender	Female	Male	Female	Male	Female	Male
	65	50	70	55	45	30
	70	55	65	65	60	30
	60	80	60	70	85	30
	60	65	70	55	65	55
	60	70	65	55	70	35
	55	75	60	60	70	20
	60	75	60	50	80	45
	55	65	50	50	60	40

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

Exercise 8: Two-way ANOVA

goggles.csv

- Load **goggles.csv**

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

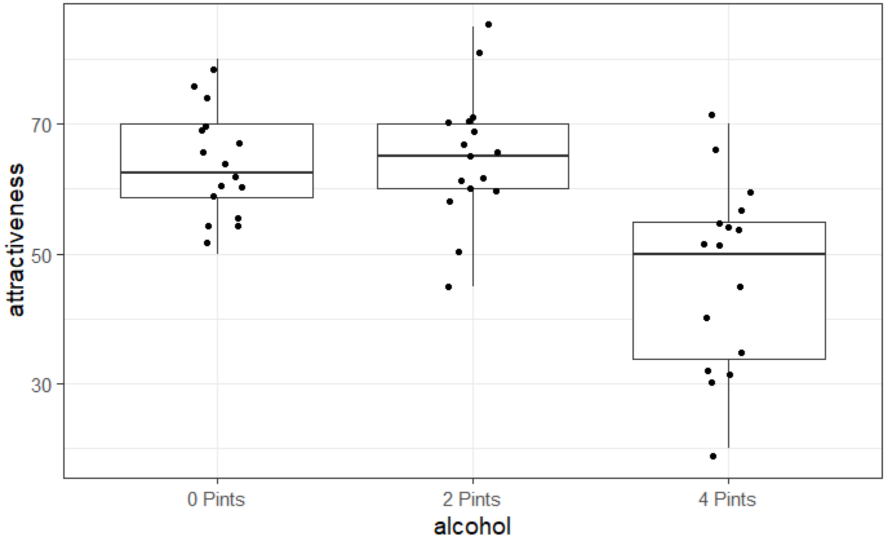
- Check the assumptions visually (plot+qqplot) and formally (test)

```
levene_test(y ~ factor1*factor2)
```

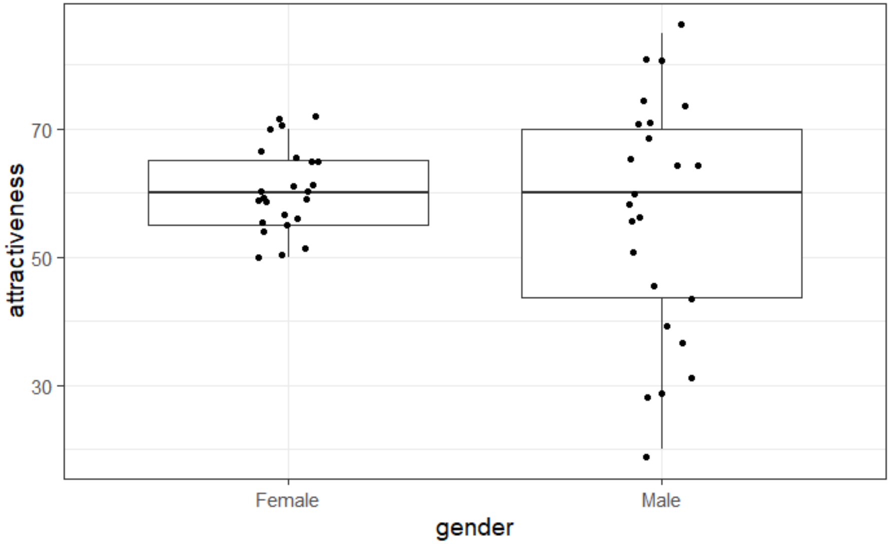
Two-way Analysis of Variance

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

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

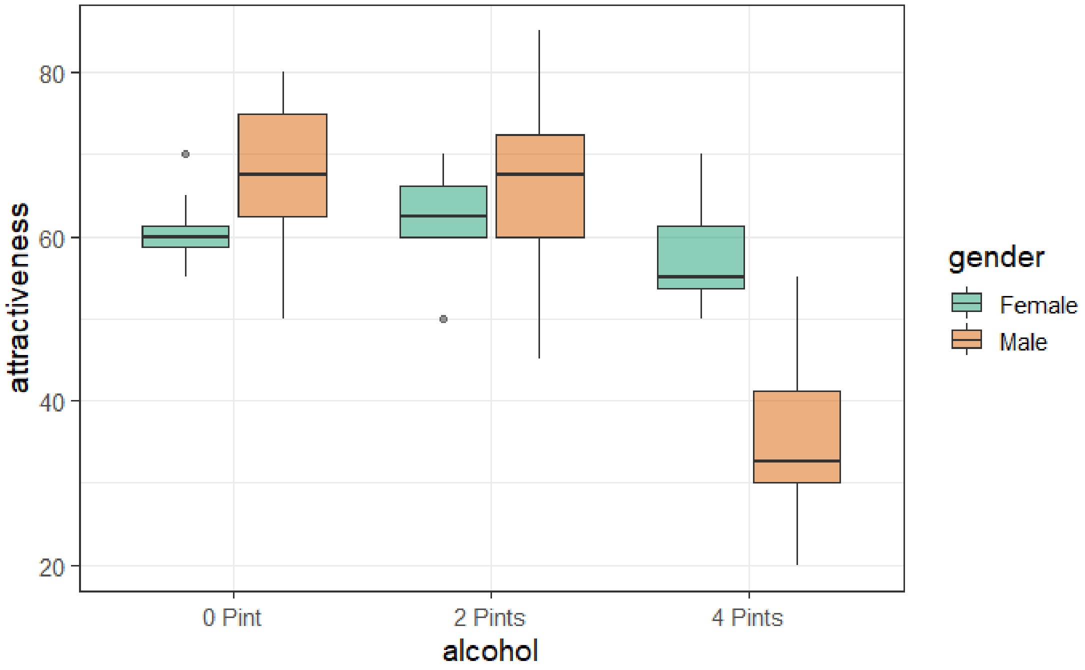


```
goggles %>%
  ggplot(aes(x=gender, y=attractiveness))+
       geom_boxplot()+
       geom_jitter(height=0, width=0.1)
```



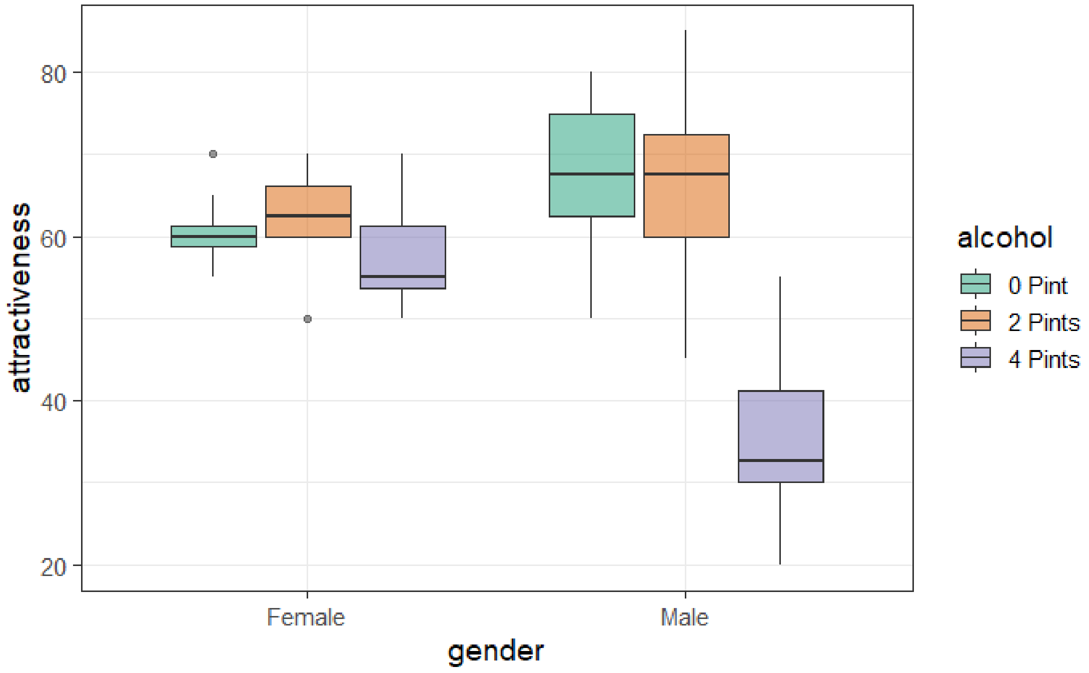
Two-way Analysis of Variance

```
goggles %>%
  ggplot(aes(alcohol, attractiveness, fill=gender))+
      geom_boxplot(alpha=0.5)+
      scale_fill_brewer(palette="Dark2")
```



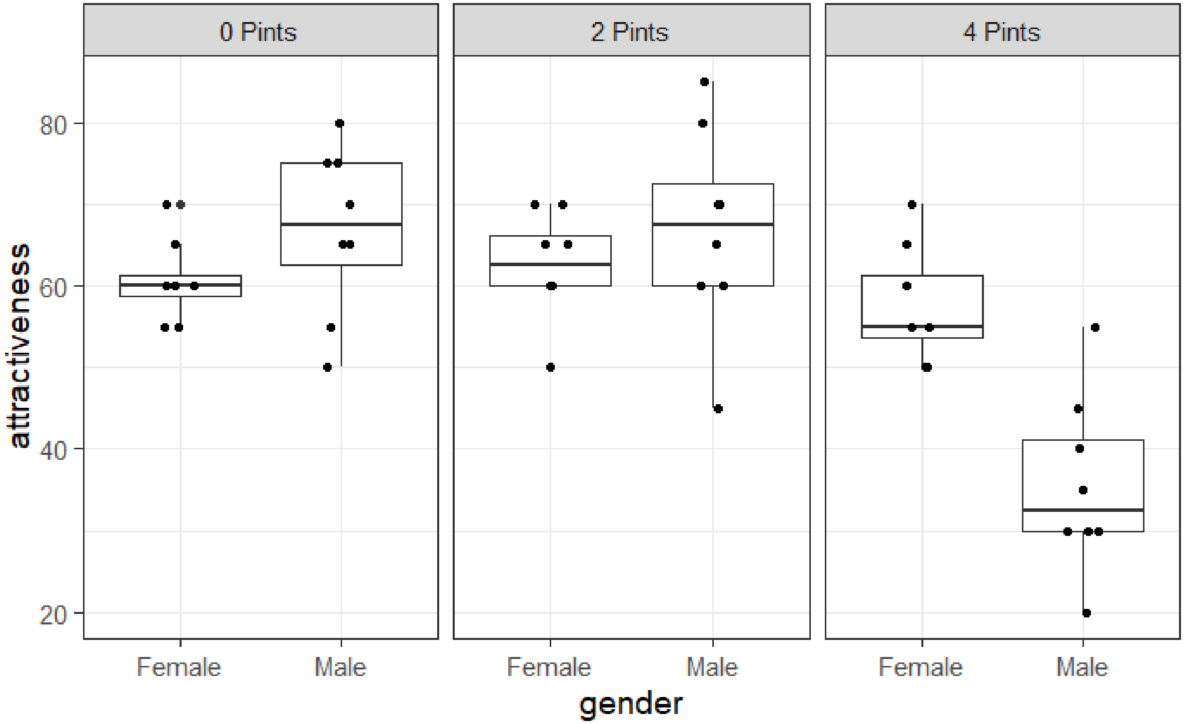
Two-way Analysis of Variance

```
goggles %>%
  ggplot(aes(gender, attractiveness, fill=alcohol))+
  geom_boxplot(alpha=0.5)+
  scale_fill_brewer(palette="Dark2")
```



Two-way Analysis of Variance

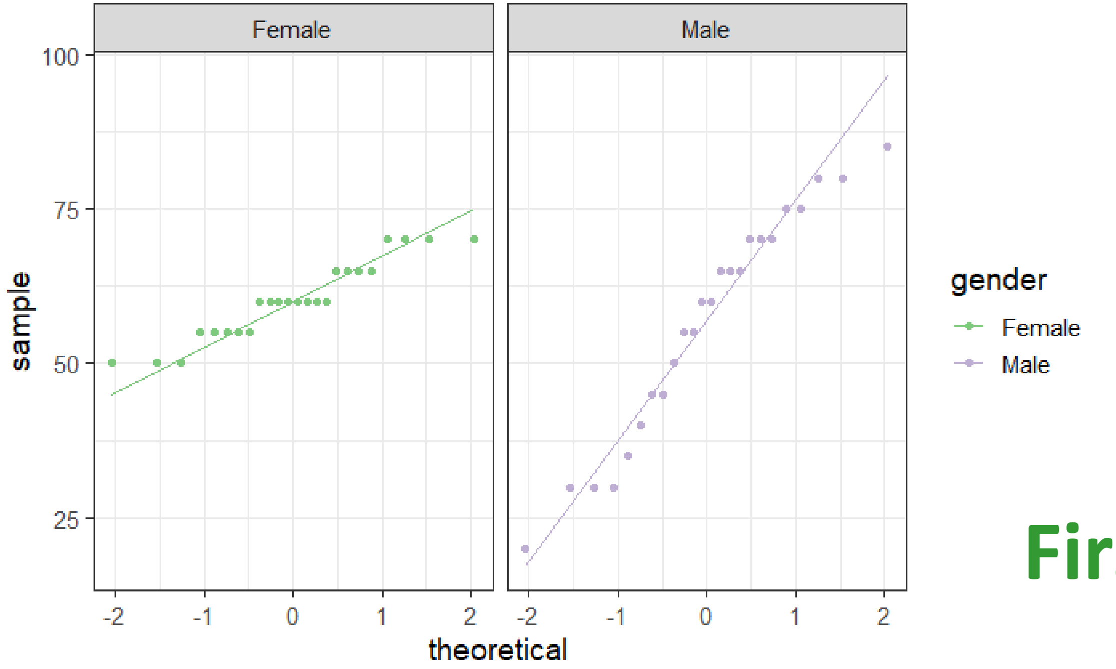
```
goggles %>%
  ggplot(aes(x=gender, y=attractiveness))+
  geom_boxplot()+
  geom_jitter(height=0, width=0.1)+
  facet_grid(cols=vars(alcohol))
```



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")
```

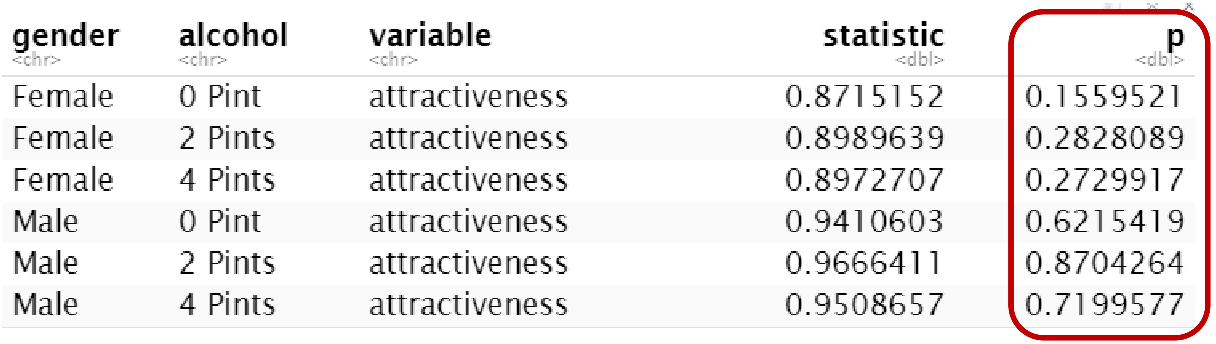


First assumption ✓

Two-way Analysis of Variance

Checking the assumptions

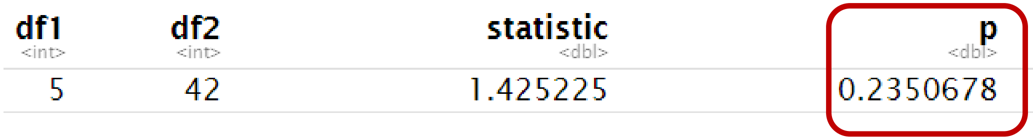
```
goggles %>%
  group_by(gender, alcohol) %>%
    shapiro_test(attractiveness) %>%
      ungroup()
```



```
gender  alcohol  variable          statistic          p
<chr>   <chr>    <chr>                 <dbl>      <dbl>
Female  0 Pint   attractiveness    0.8715152  0.1559521
Female  2 Pints  attractiveness    0.8989639  0.2828089
Female  4 Pints  attractiveness    0.8972707  0.2729917
Male    0 Pint   attractiveness    0.9410603  0.6215419
Male    2 Pints  attractiveness    0.9666411  0.8704264
Male    4 Pints  attractiveness    0.9508657  0.7199577
```

First assumption ✓

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



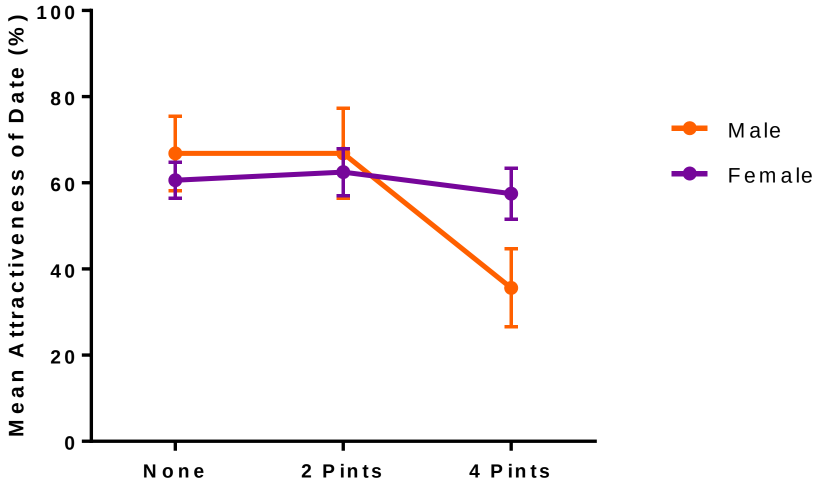
```
df1    df2         statistic          p
<int>  <int>           <dbl>      <dbl>
  5     42          1.425225  0.2350678
```

Second assumption ✓

Two-way Analysis of Variance

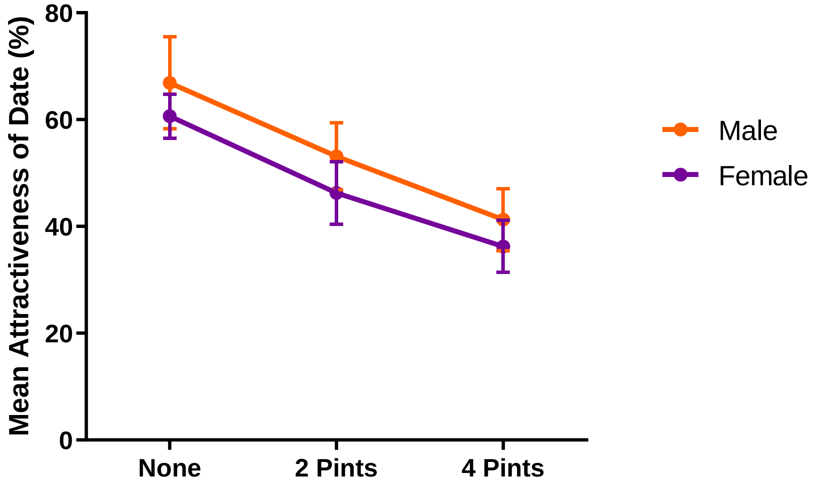
With significant interaction (real data)

ANOVA table	SS	DF	MS	F (DFn, DFd)	P value
Interaction	**1978**	**2**	**989.1**	**F (2, 42) = 11.91**	**< 0.0001**
Alcohol Consumption	3332	2	1666	F (2, 42) = 20.07	< 0.0001
Gender	168.8	1	168.8	F (1, 42) = 2.032	0.1614
Residual	3488	42	83.04		



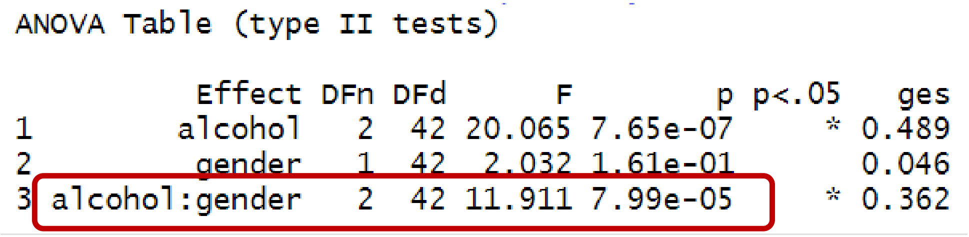
Without significant interaction (fake data)

ANOVA table	SS	DF	MS	F (DFn, DFd)	P value
Interaction	7.292	2	3.646	F (2, 42) = 0.06872	0.9337
Alcohol Consumption	**5026**	**2**	**2513**	**F (2, 42) = 47.37**	**< 0.0001**
Gender	**438.0**	**1**	**438.0**	**F (1, 42) = 8.257**	**0.0063**
Residual	2228	42	53.05		



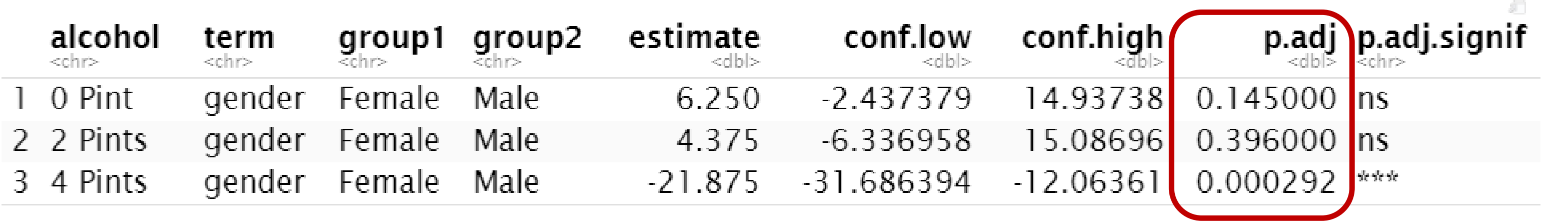
Two-way Analysis of Variance

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

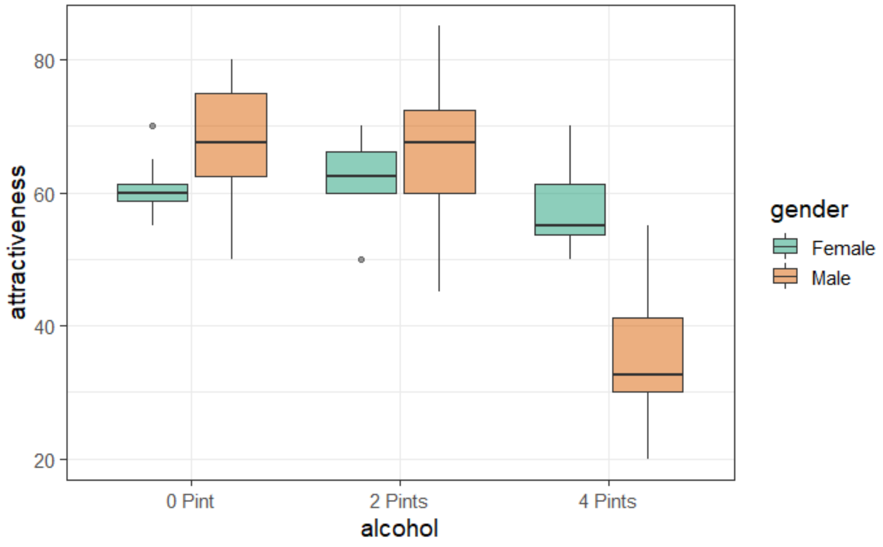


```
ANOVA Table (type II tests)

          Effect DFn DFd      F        p p<.05   ges
1        alcohol   2  42 20.065 7.65e-07     * 0.489
2         gender   1  42  2.032 1.61e-01       0.046
3 alcohol:gender   2  42 11.911 7.99e-05     * 0.362
```

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



	alcohol <chr>	term <chr>	group1 <chr>	group2 <chr>	estimate <dbl>	conf.low <dbl>	conf.high <dbl>	p.adj <dbl>	p.adj.signif <chr>
1	0 Pint	gender	Female	Male	6.250	-2.437379	14.93738	0.145000	ns
2	2 Pints	gender	Female	Male	4.375	-6.336958	15.08696	0.396000	ns
3	4 Pints	gender	Female	Male	-21.875	-31.686394	-12.06361	0.000292	***



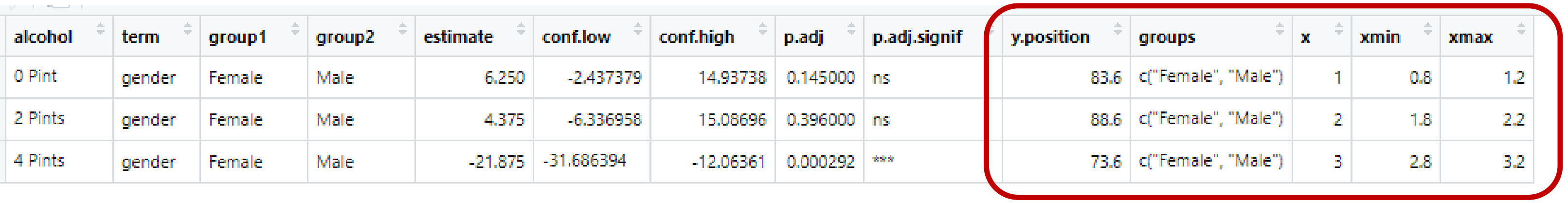
Answer: there is a significant effect of alcohol consumption on the way the attractiveness of a date is perceived but it varies significantly between genders (p=7.99e-05).

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

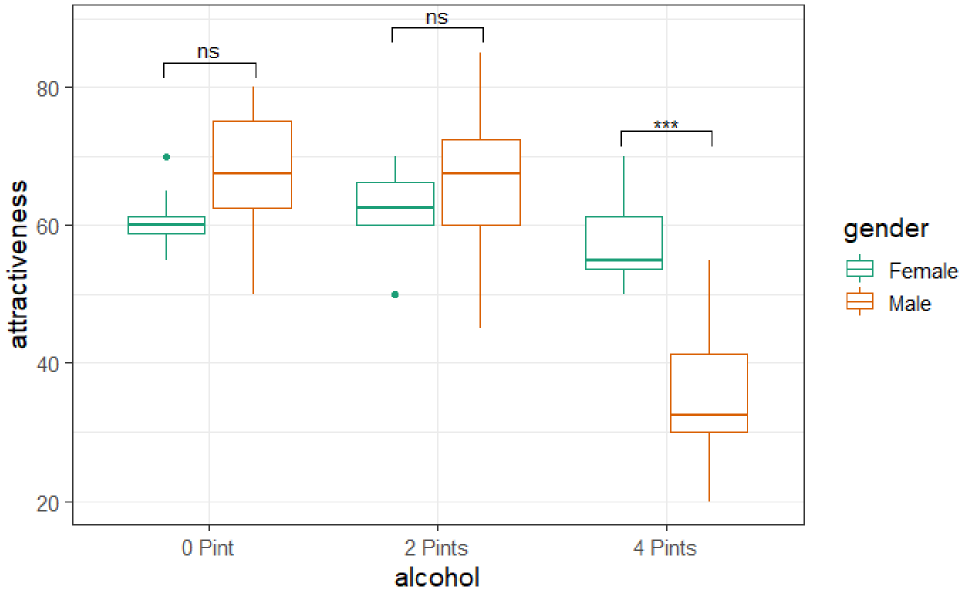
- *Work in progress* # ggpubr package #

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



alcohol	term	group1	group2	estimate	conf.low	conf.high	p.adj	p.adj.signif	y.position	groups	x	xmin	xmax
0 Pint	gender	Female	Male	6.250	-2.437379	14.93738	0.145000	ns	83.6	c("Female", "Male")	1	0.8	1.2
2 Pints	gender	Female	Male	4.375	-6.336958	15.08696	0.396000	ns	88.6	c("Female", "Male")	2	1.8	2.2
4 Pints	gender	Female	Male	-21.875	-31.686394	-12.06361	0.000292	***	73.6	c("Female", "Male")	3	2.8	3.2

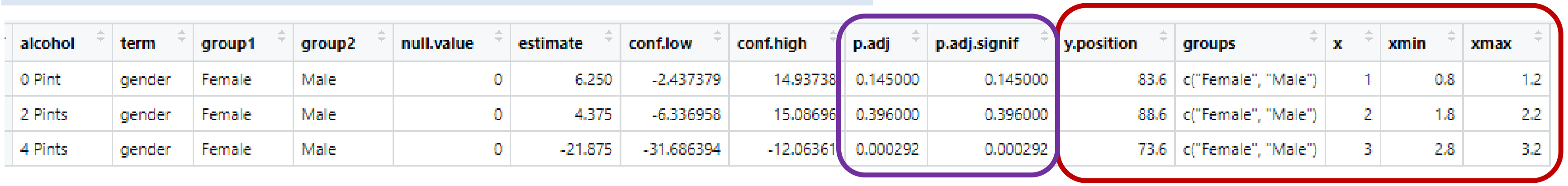
```
goggles %>%
  ggplot(aes(alcohol, attractiveness, colour = gender))+
  geom_boxplot()+
  stat_pvalue_manual(tukey.results)+
  scale_colour_brewer(palette = "Dark2")
```



Two-way Analysis of Variance

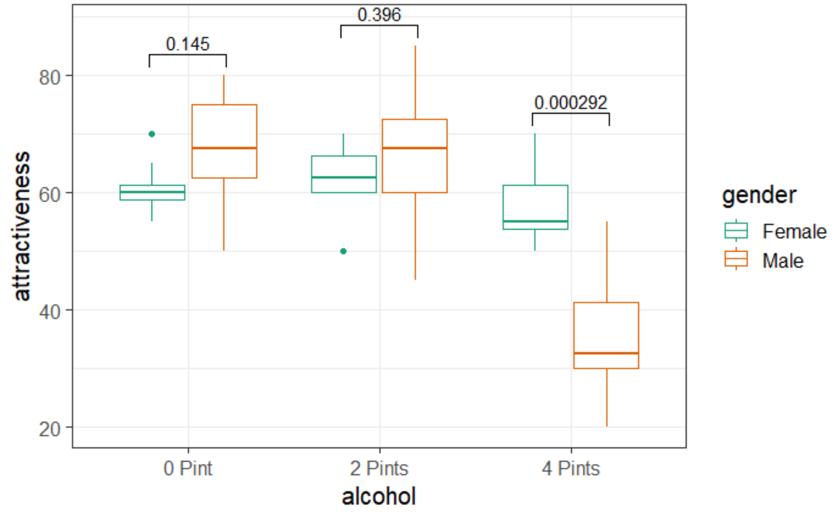
- ***Work in progress*** # ggpubr package # **Actual p-values rather than NS or ***

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



alcohol	term	group1	group2	null.value	estimate	conf.low	conf.high	p.adj	p.adj.signif	y.position	groups	x	xmin	xmax
0 Pint	gender	Female	Male	0	6.250	-2.437379	14.93738	0.145000	0.145000	83.6	c("Female", "Male")	1	0.8	1.2
2 Pints	gender	Female	Male	0	4.375	-6.336958	15.08696	0.396000	0.396000	88.6	c("Female", "Male")	2	1.8	2.2
4 Pints	gender	Female	Male	0	-21.875	-31.686394	-12.06361	0.000292	0.000292	73.6	c("Female", "Male")	3	2.8	3.2

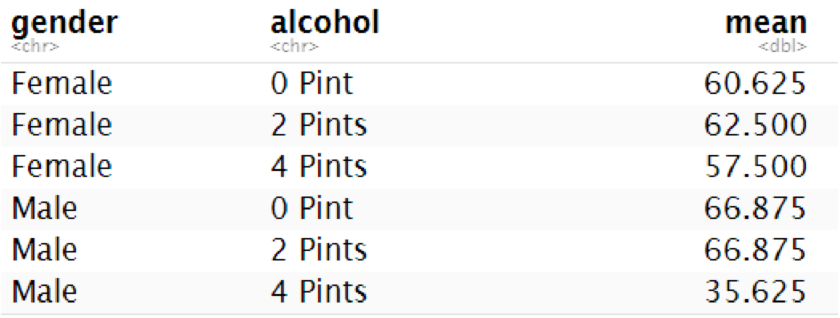
```
goggles %>%
  ggplot(aes(alcohol, attractiveness, colour = gender))+
  geom_boxplot()+
  stat_pvalue_manual(tukey.results)+
  scale_colour_brewer(palette = "Dark2")
```



Two-way Analysis of Variance

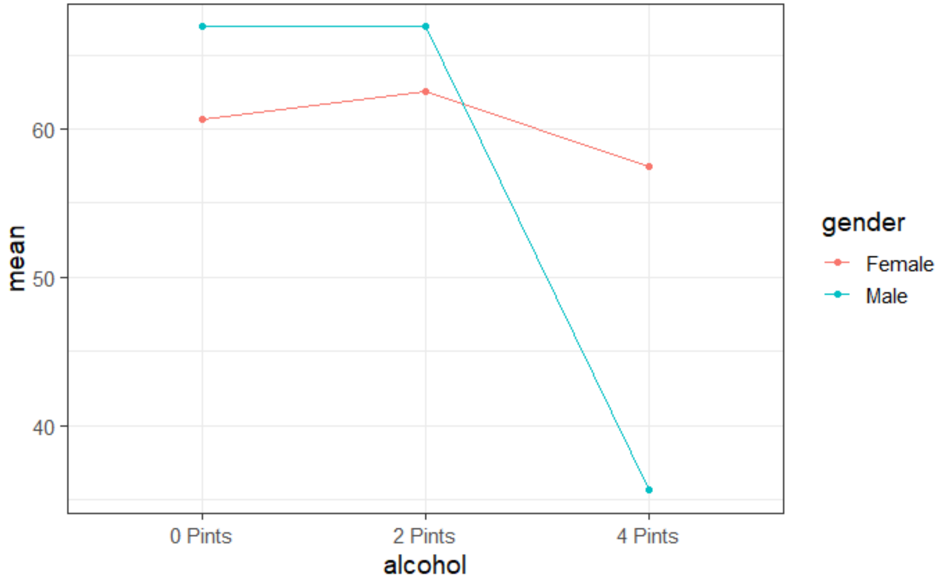
- Now a quick way to have a look at the interaction

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



gender	alcohol	mean
<chr>	<chr>	<dbl>
Female	0 Pint	60.625
Female	2 Pints	62.500
Female	4 Pints	57.500
Male	0 Pint	66.875
Male	2 Pints	66.875
Male	4 Pints	35.625

```
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

$$\frac{\boxed{\text{Similarity}}}{\boxed{\text{Variability}}} = \frac{\text{Signal}}{\text{Noise}}$$

$$\frac{\text{\Large Signal}}{\text{\small Noise}} = \textbf{statistical significance}$$

$$\frac{\text{\Large Signal}}{\text{\Huge Noise}} = \textbf{no statistical significance}$$

Signal-to-noise ratio and Correlation

$$\frac{\text{Similarity}}{\text{Variability}} = \frac{\text{Signal}}{\text{Noise}}$$

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

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

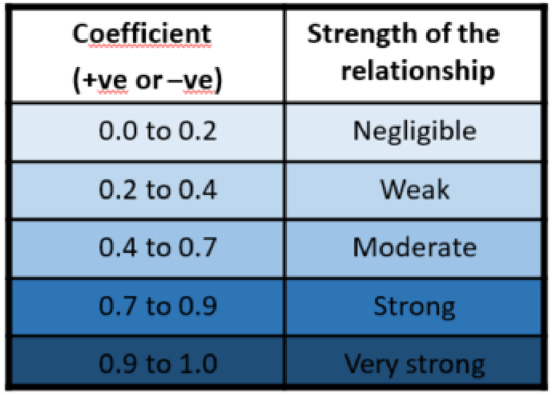
covariance

$$\mathbf{r} = \frac{\text{similarity}}{\text{variability}} = \frac{COV_{xy}}{SD_x SD_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\, SD_x SD_y}$$

Standard Deviation

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 (+ve or –ve)	Strength of the relationship
0.0 to 0.2	Negligible
0.2 to 0.4	Weak
0.4 to 0.7	Moderate
0.7 to 0.9	Strong
0.9 to 1.0	Very strong

- 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**

Correlation

$p = 0.0002$ 😄

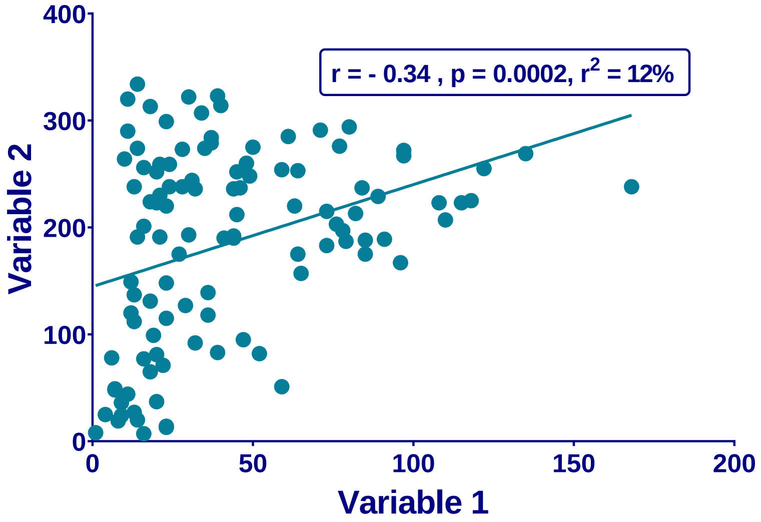
$r = -0.34$ 😐

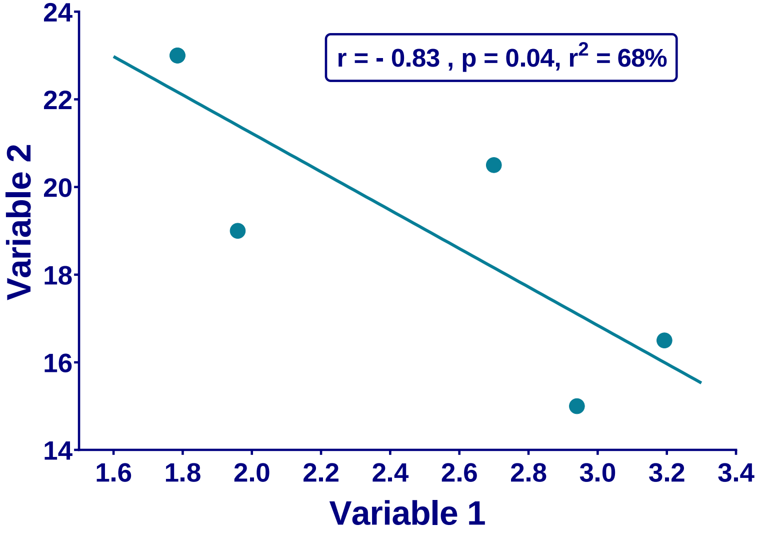
$r^2 = 0.12$ 😐

$p = 0.04$ 😐

$r = -0.83$ 🙂

$r^2 = 0.68$ 😄





Power!!

Correlation

Assumptions

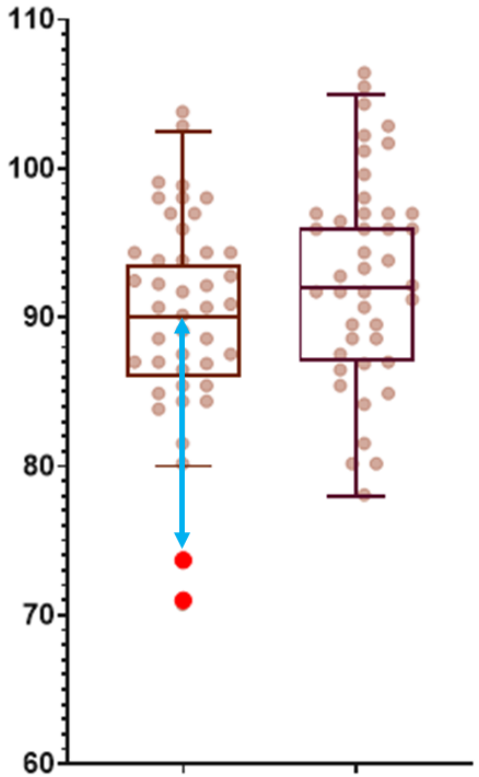
- Assumptions for correlation
  - Regression and linear Model (lm)

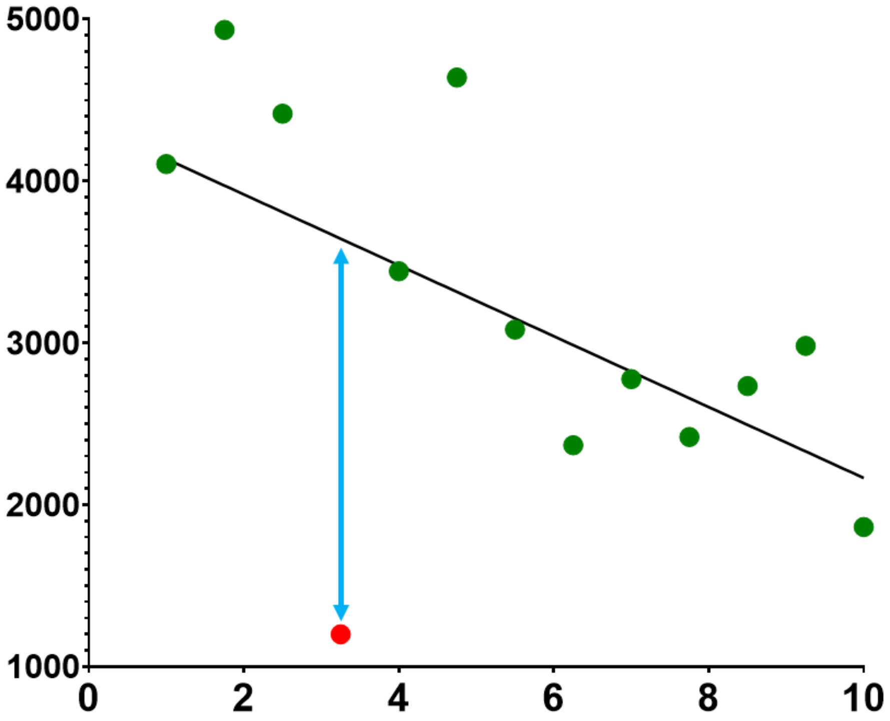
- **Linearity**: The relationship between X and the mean of Y is linear.
- **Homoscedasticity**: The variance of residual is the same for any value of X.
- **Independence:** Observations are independent of each other.
- **Normality:** For any fixed value of X, Y is normally distributed.

Correlation

Outliers and High leverage points

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





Correlation

Outliers and High leverage points

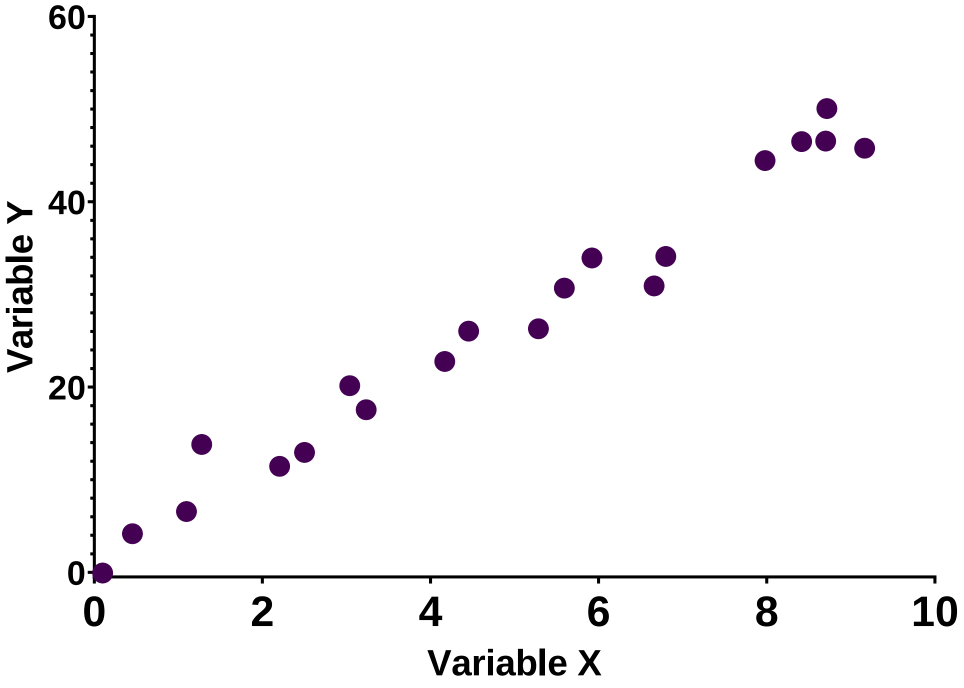
- **Leverage points**: A leverage point is defined as an observation that has a value of x that is far away from the mean of x.

- 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

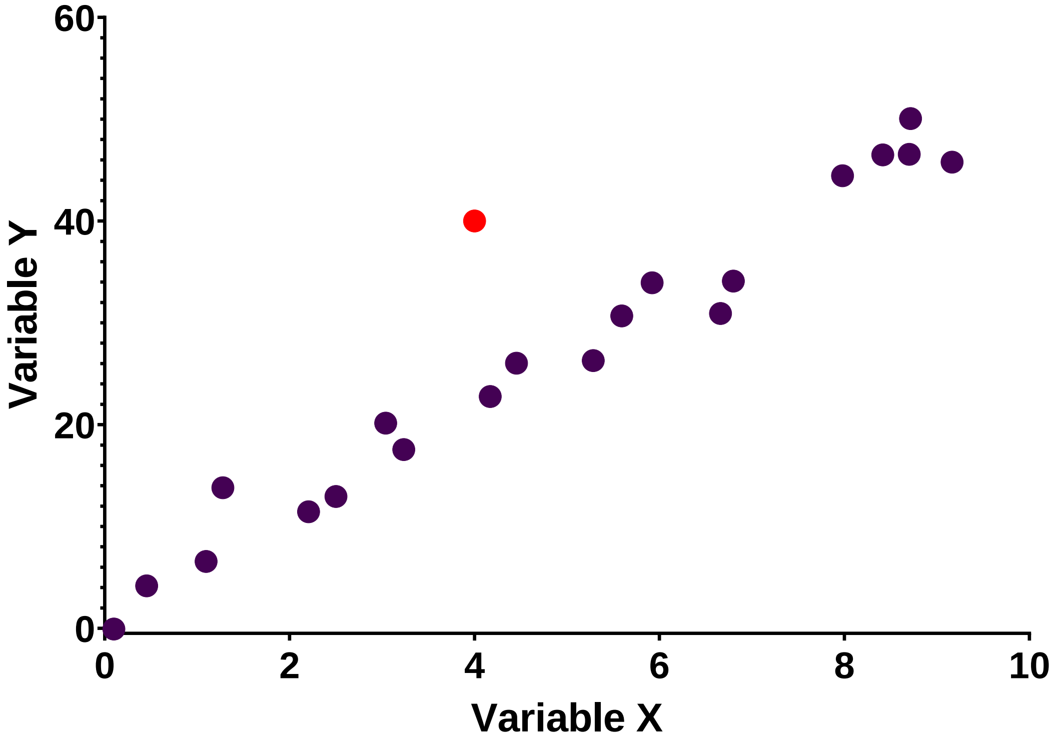
Outliers and High leverage points

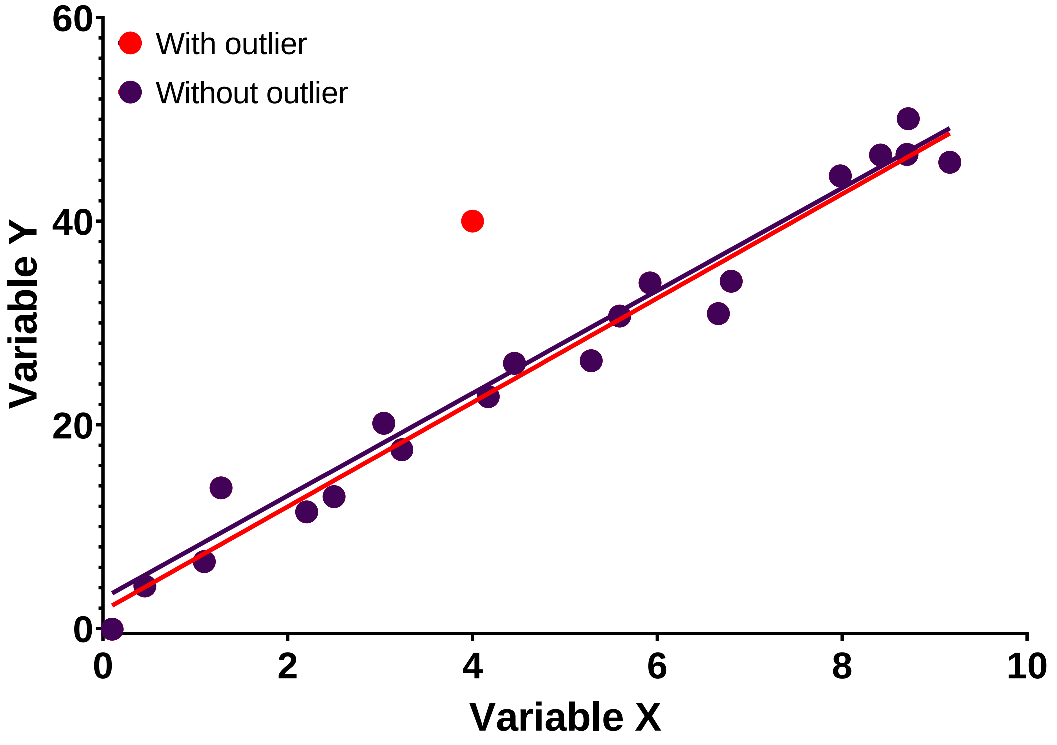


All good

Correlation

Outliers and High leverage points

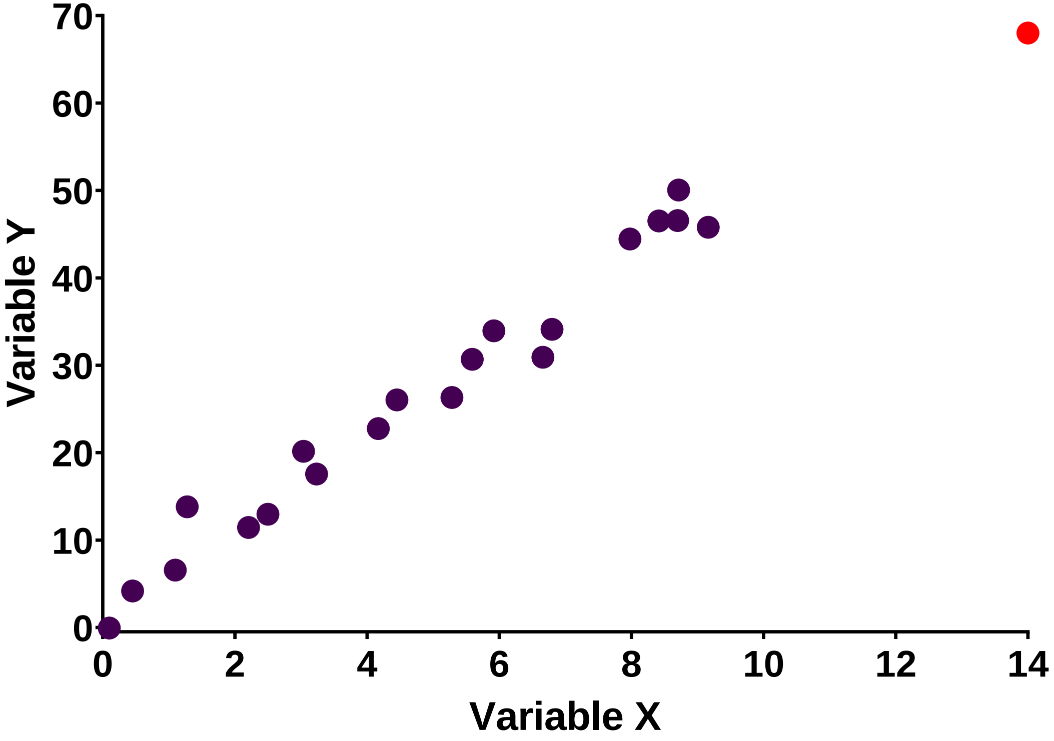


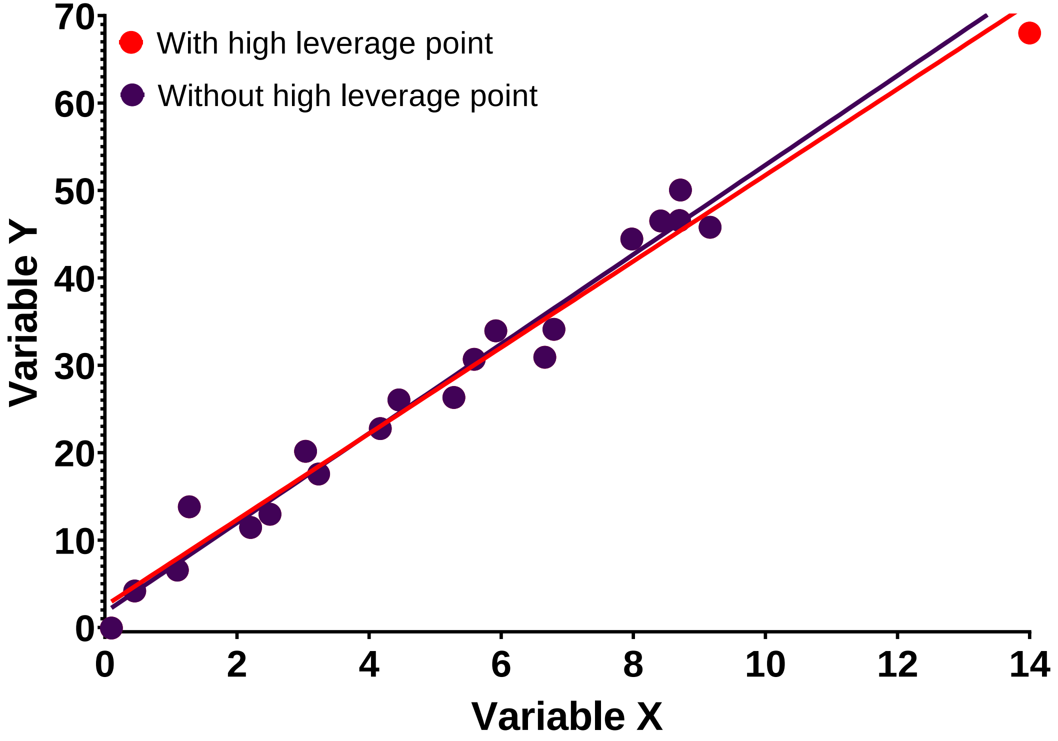


Outlier but not influential value

Correlation

Outliers and High leverage points

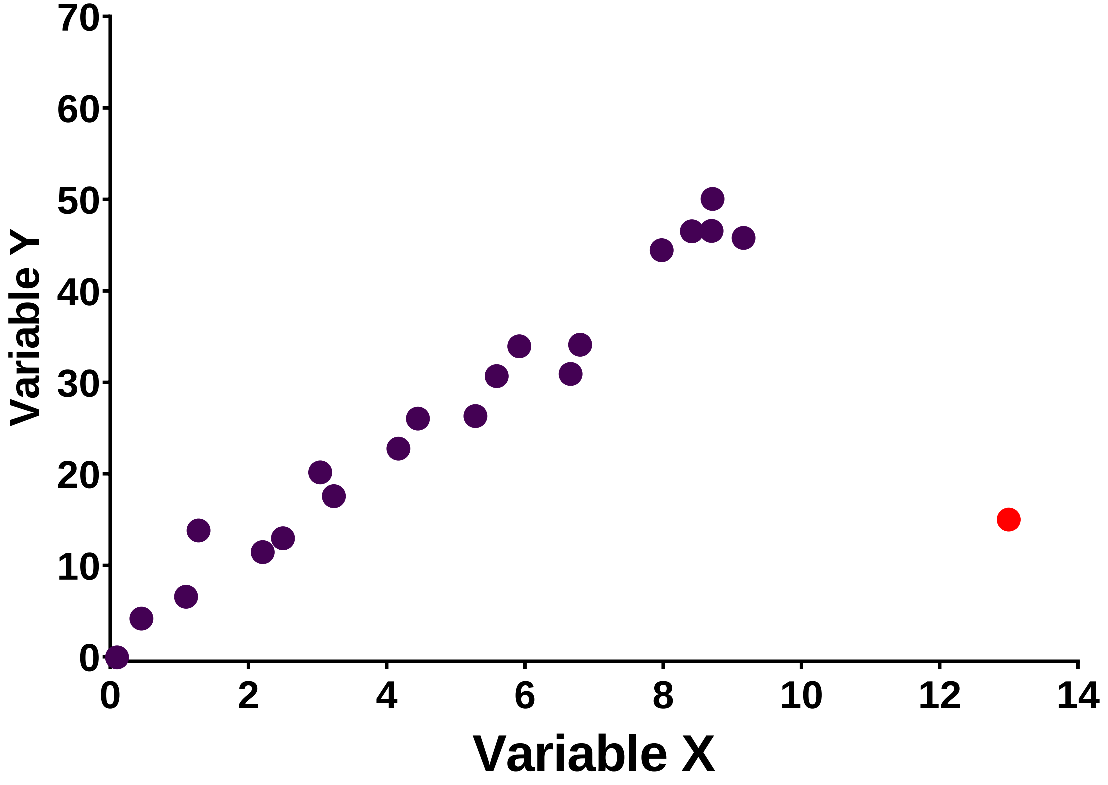


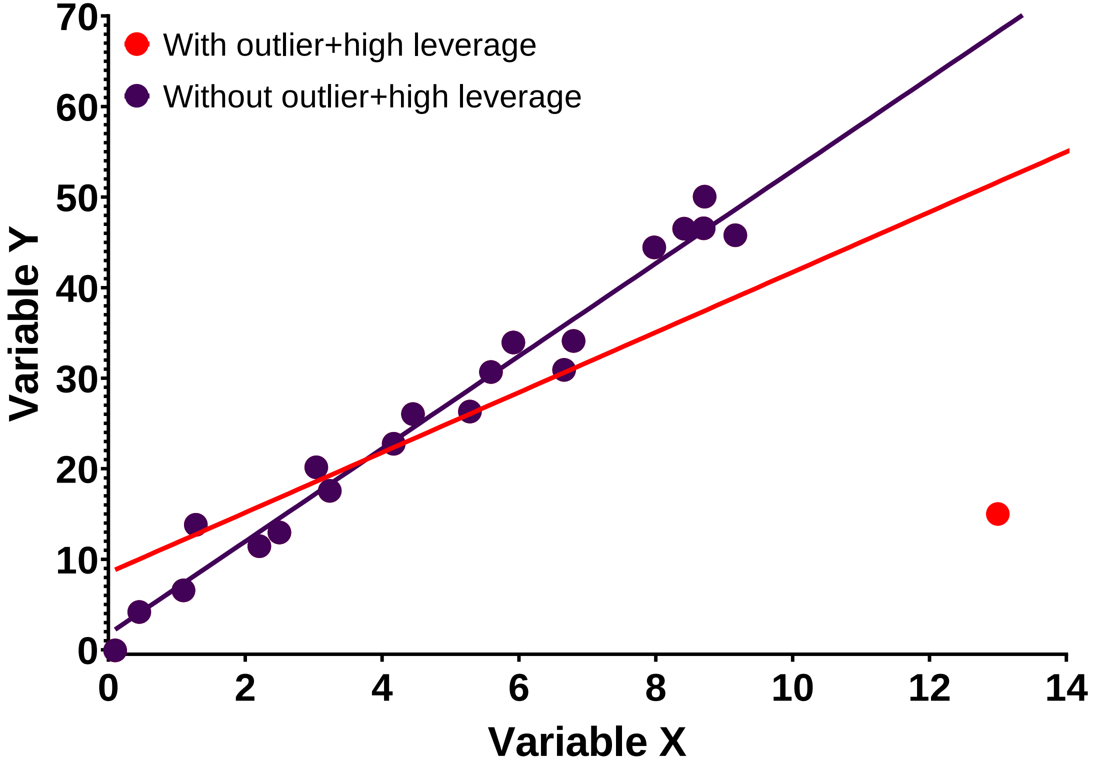


High leverage but not influential value

Correlation

Outliers and High leverage points





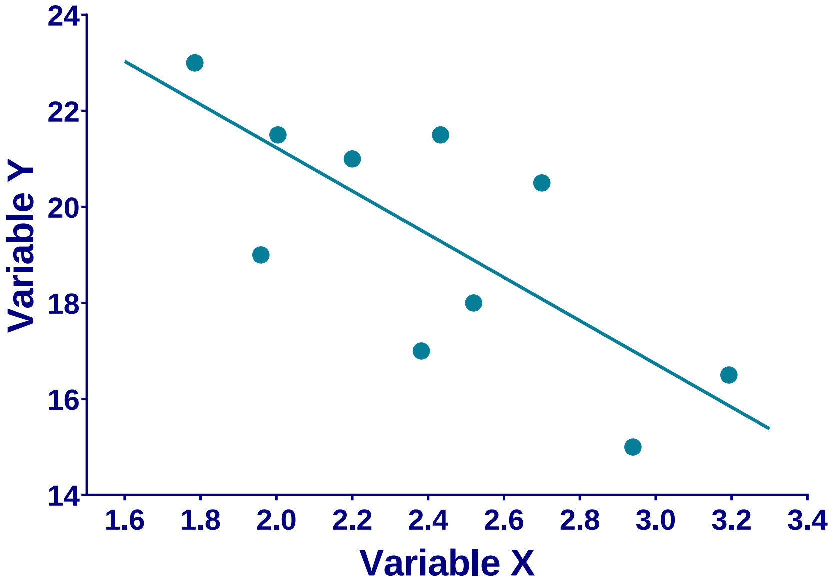
Outlier and High leverage: Influential value

Correlation: Two more things

Thing 1: Pearson correlation is a parametric test

First assumption for parametric test: Normality

Correlation: bivariate Gaussian distribution



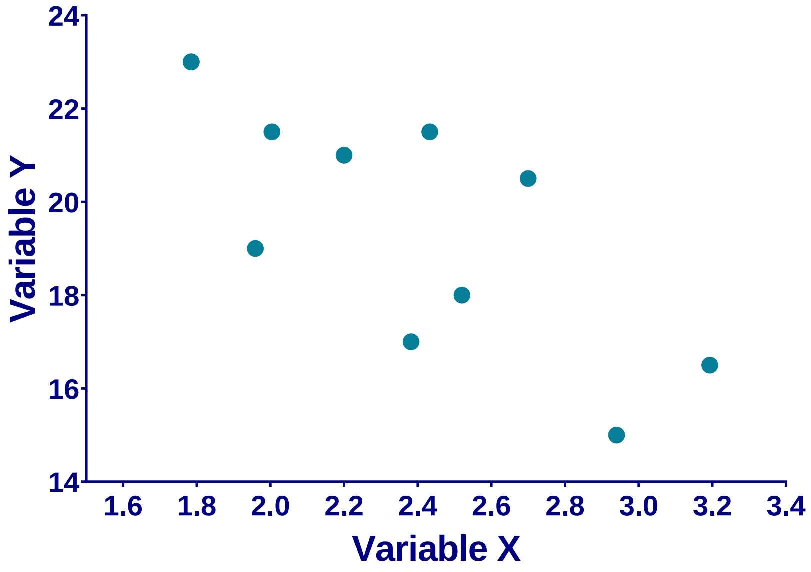
Symmetry-ish of the values on either side of the line of best fit.

Correlation: Two more things

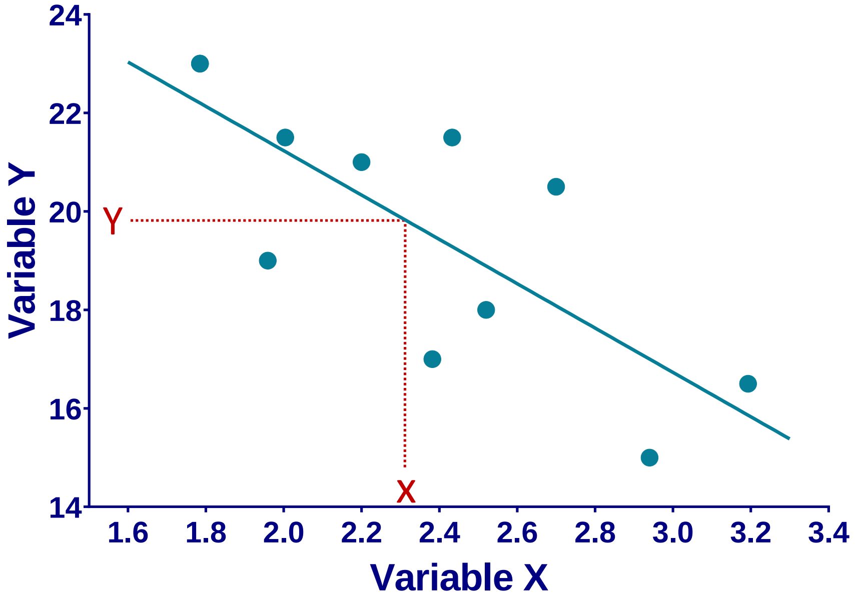
Thing 2: Line of best fit comes from a regression

Correlation: nature and strength of the association

Regression: nature and strength of the association <u>and</u> **prediction**



Correlation = Association

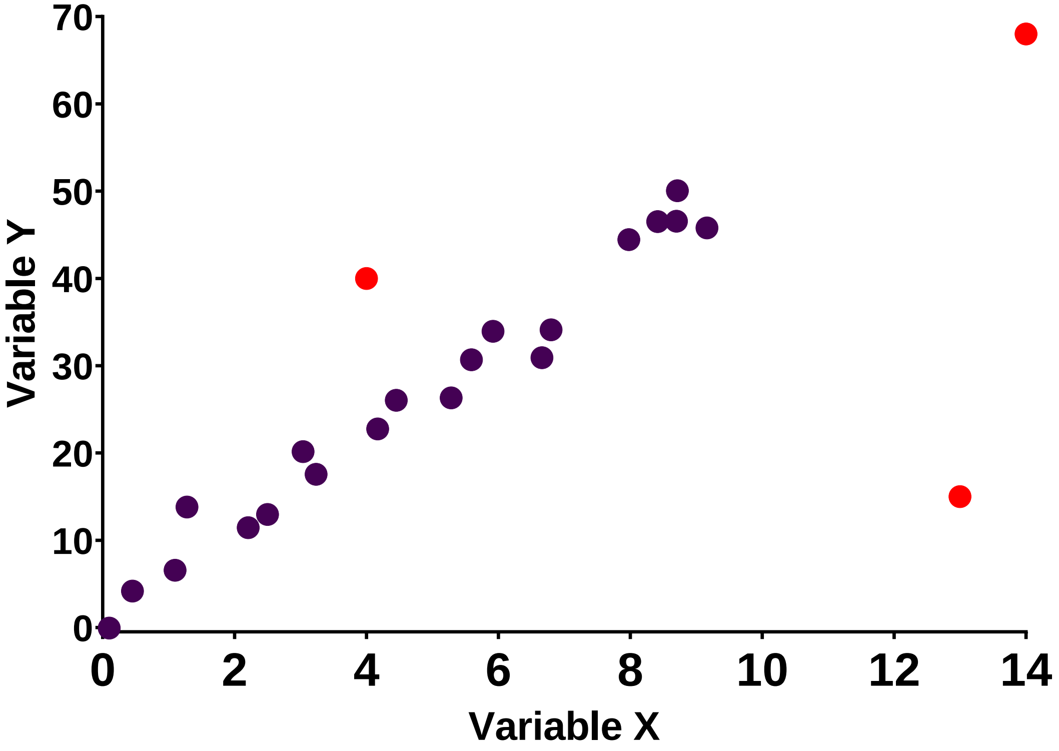


Regression = Prediction

Y = A + B*X

Correlation: correlation.csv

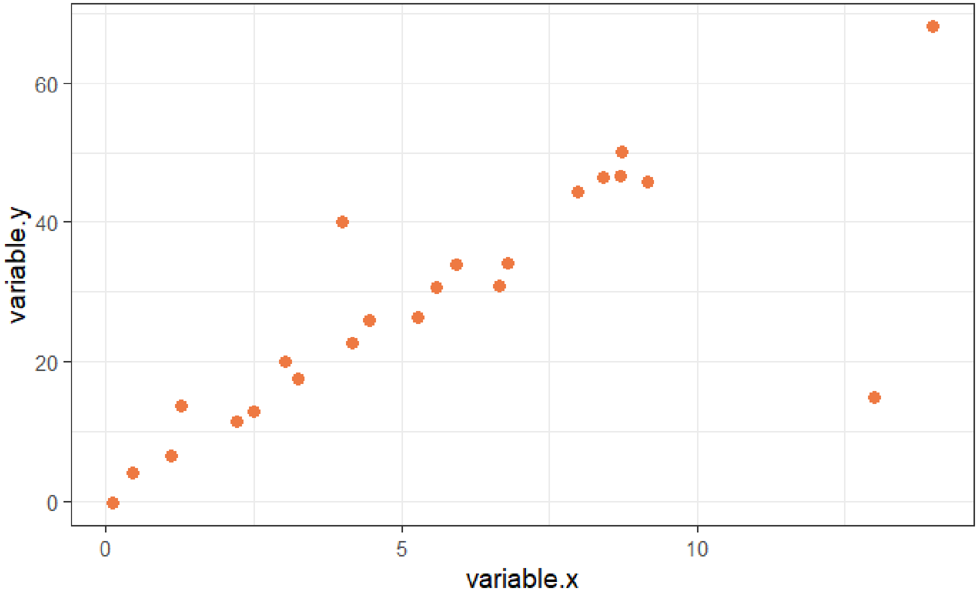
- **Questions**:
  - What is the nature and the strength of the relationship between X and Y?
  - Are there any dodgy points?

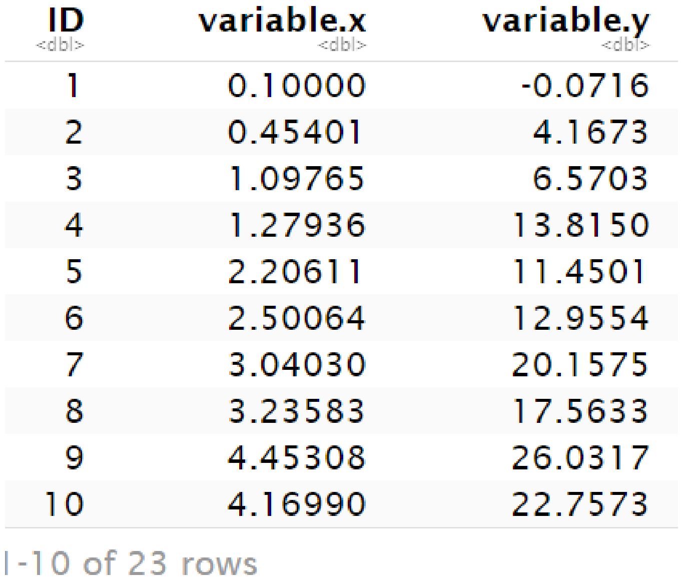


Correlation: correlation.csv

- **Question**: are there any dodgy points?

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

correlation %>%
  ggplot(aes(variable.x, variable.y, colour=Gender)) +
  geom_point(size=3, colour="sienna2")
```





ID <dbl>	variable.x <dbl>	variable.y <dbl>
1	0.10000	-0.0716
2	0.45401	4.1673
3	1.09765	6.5703
4	1.27936	13.8150
5	2.20611	11.4501
6	2.50064	12.9554
7	3.04030	20.1575
8	3.23583	17.5633
9	4.45308	26.0317
10	4.16990	22.7573

1-10 of 23 rows

Correlation: correlation.csv

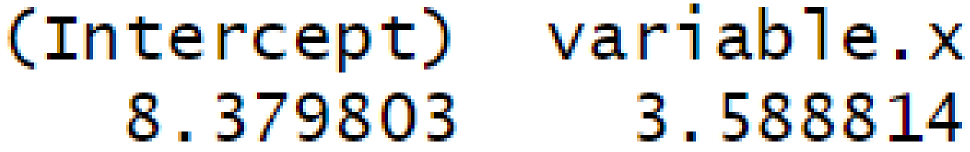
- For the lines of best-fit: 3 new functions:

```
lm(y~x, data=) -> fit
coefficients(fit) -> cf.fit
```

(vector of 2 values)

```
geom_abline(intercept=cf.fit[1], slope=cf.fit[2])
```

```
lm(variable.y ~ variable.x, data=correlation)-> fit.correlation
coefficients(fit.correlation) -> coef.correlation
coef.correlation
```

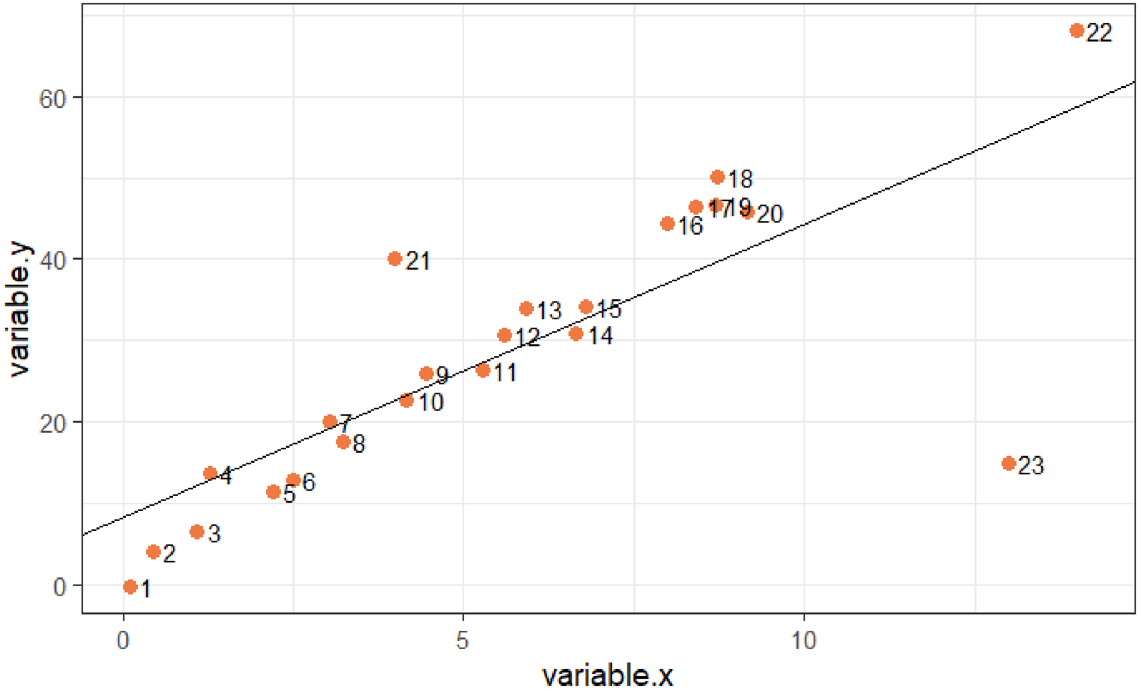


```
(Intercept)  variable.x 
   8.379803    3.588814 
```

intercept   slope

Correlation: correlation.csv

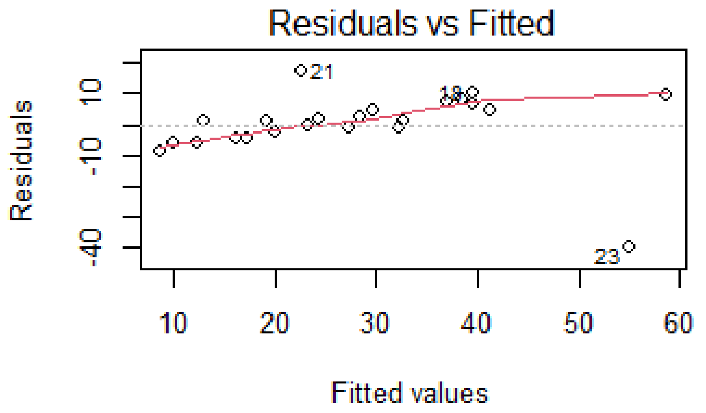
```
correlation %>%
  ggplot(aes(variable.x, variable.y, label = ID)) +
  geom_point(size=3, colour="sienna2") +
  geom_abline(intercept = coef.correlation[1], slope = coef.correlation[2])+
  geom_text(hjust = 0, nudge_x = 0.15)
```



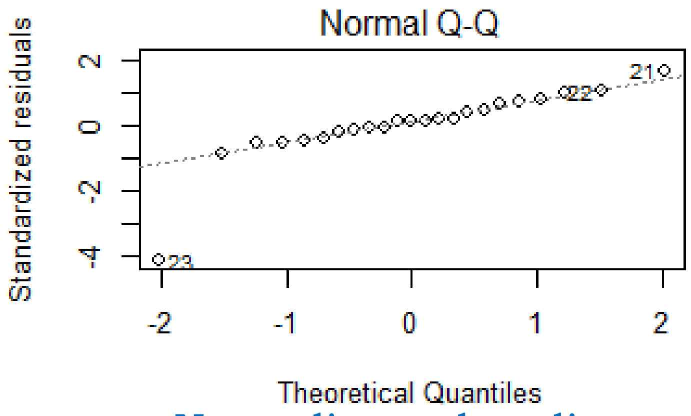
Correlation: correlation.csv

Assumptions, outliers and influential cases

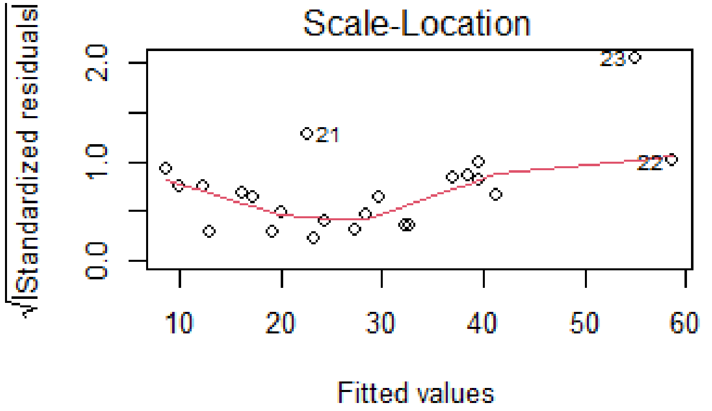
```
par(mfrow=c(2,2))
plot(fit.correlation)
```



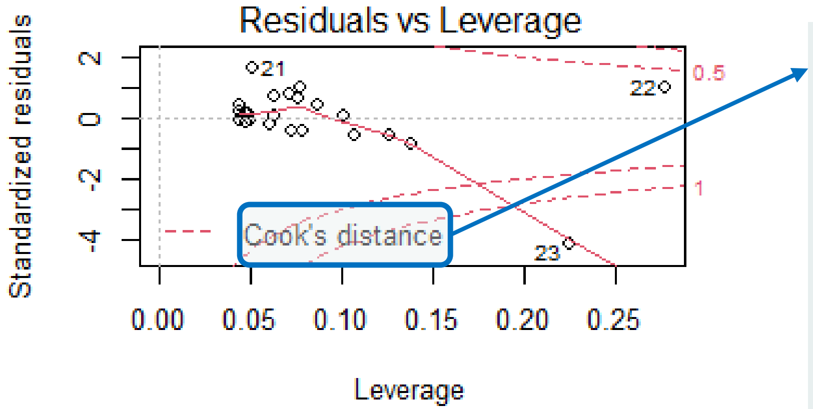
Linearity, homoscedasticity and outlier



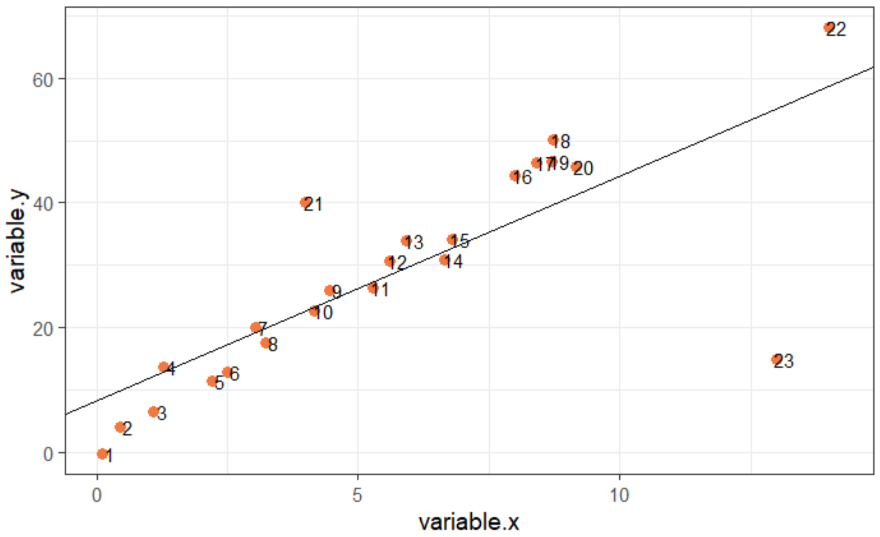
Normality and outlier



Homoscedasticity



Influential cases

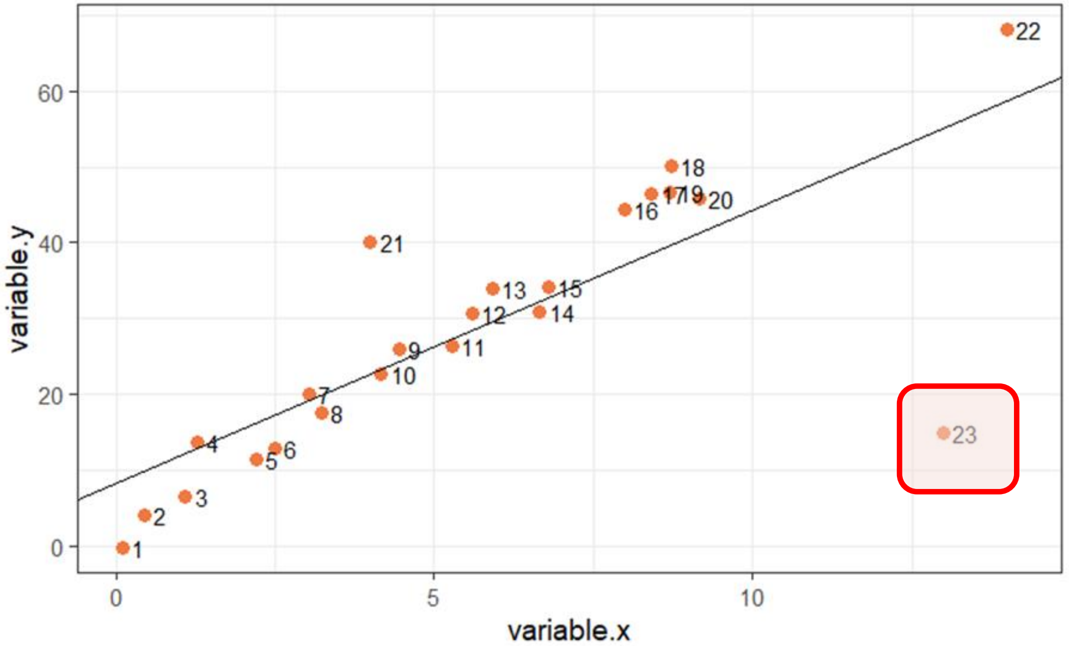


`cooks.distance()`

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).

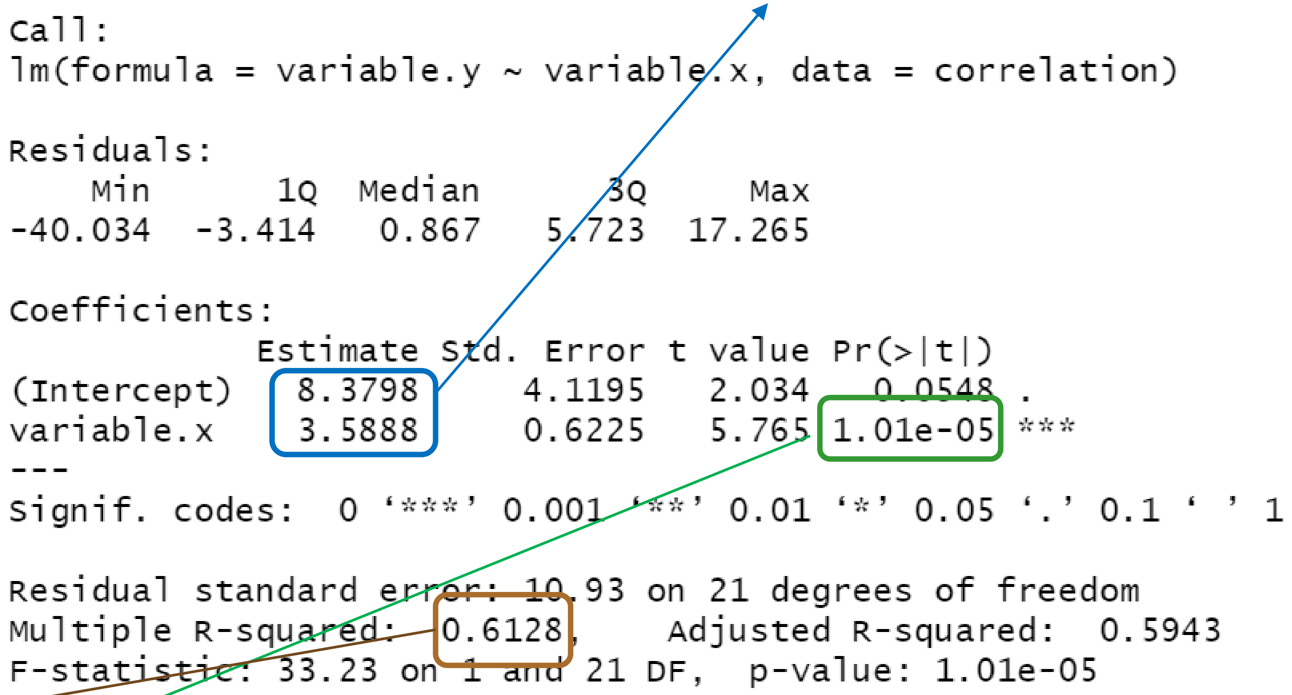
- It summarizes how much all the values in the regression model change when the ith observation is removed.
- Consensus: cut-off point =1 (0.5).

Correlation: correlation.csv



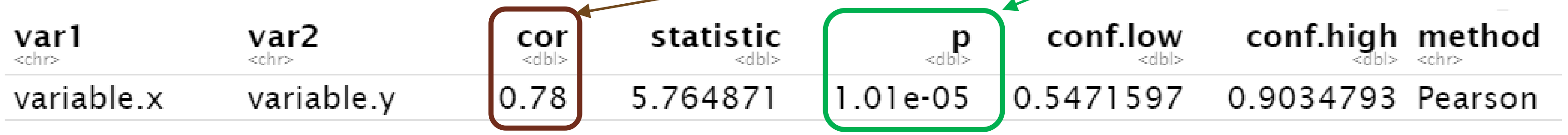
```
summary(fit.correlation)
```

Line of best fit: Y=8.38 + 3.59*X



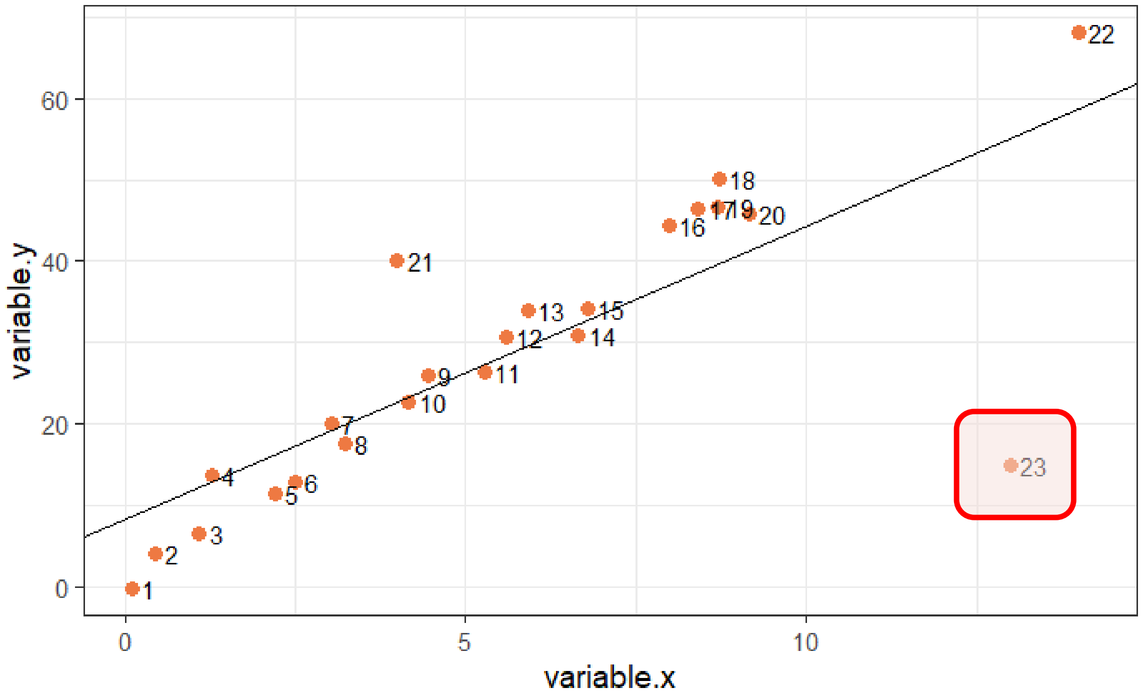
```
Call:
lm(formula = variable.y ~ variable.x, data = correlation)

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
```

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



var1	var2	cor	statistic	p	conf.low	conf.high	method
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
variable.x	variable.y	0.78	5.764871	1.01e-05	0.5471597	0.9034793	Pearson

Correlation: correlation.csv

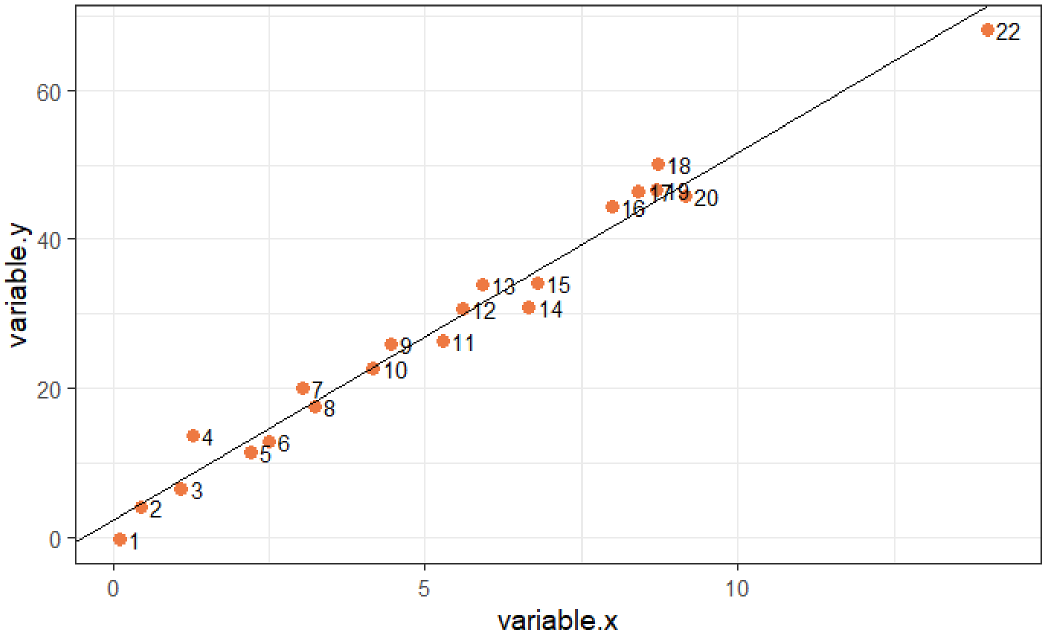


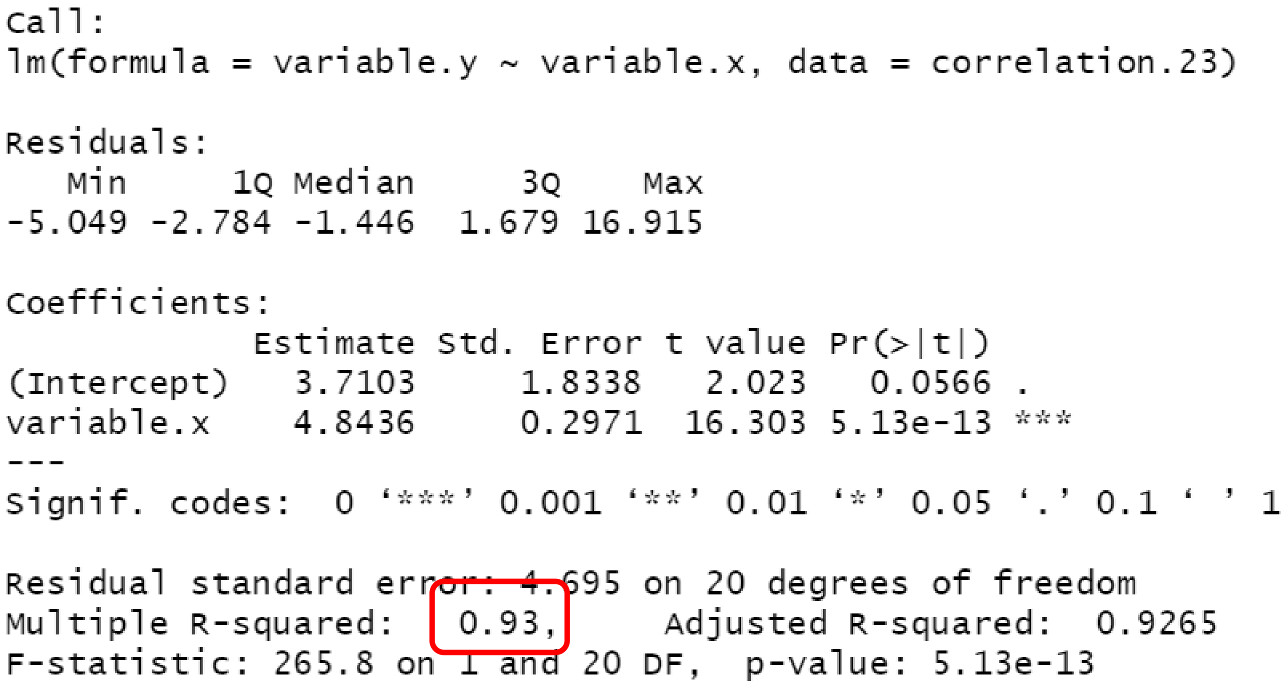
Have a go: Remove ID 23, then re-run the model and plot the graph again. Then decide what you want to do with ID 21 and 22.

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

Correlation: correlation.csv

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

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

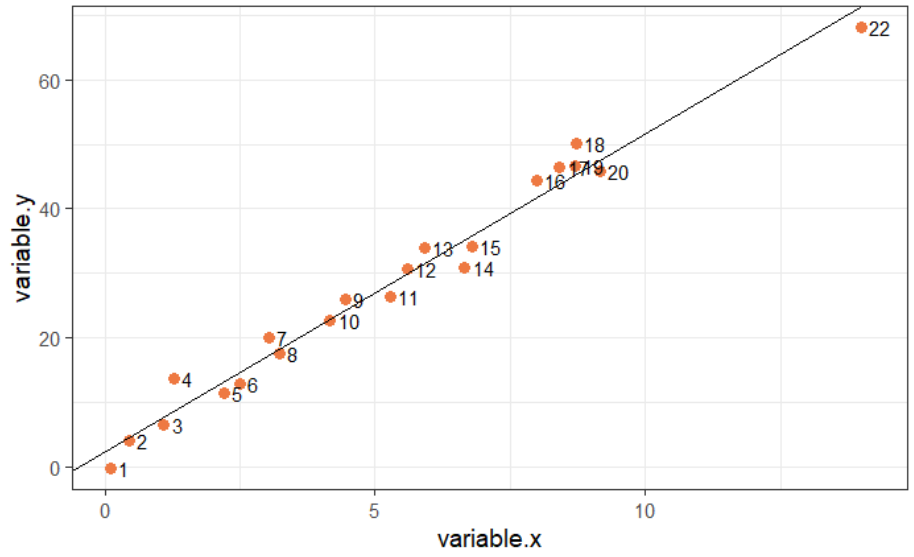


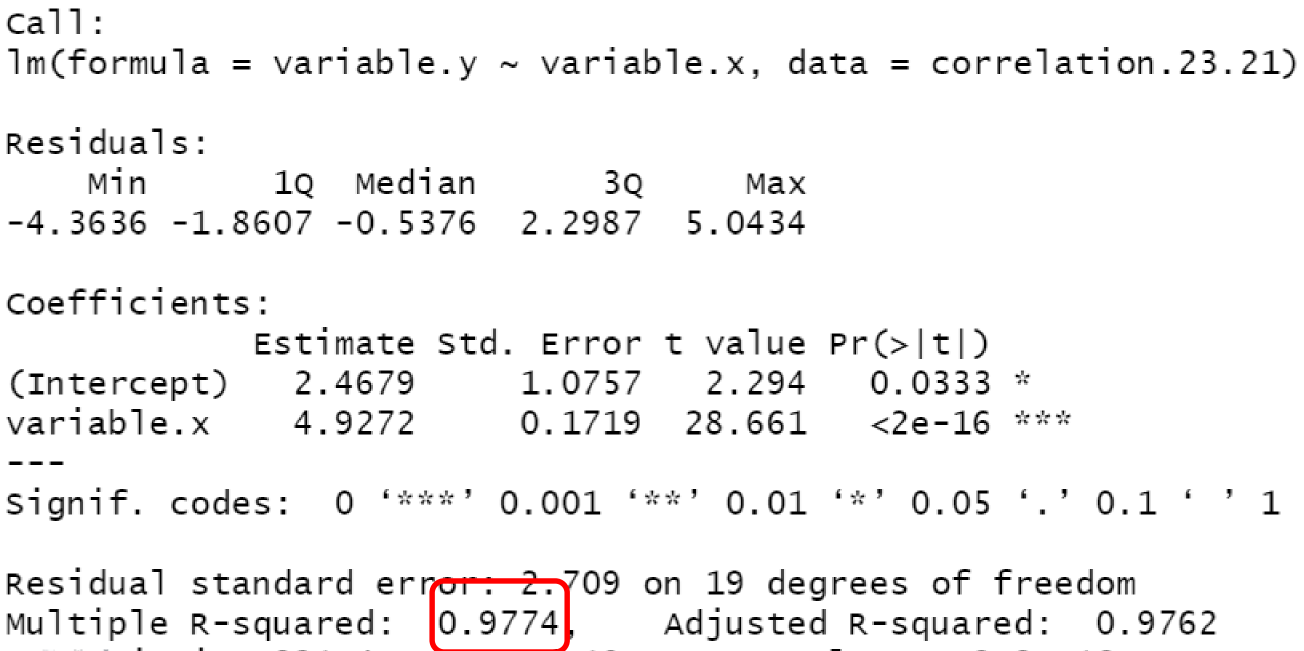


```
Call:
lm(formula = variable.y ~ variable.x, data = correlation.23)

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

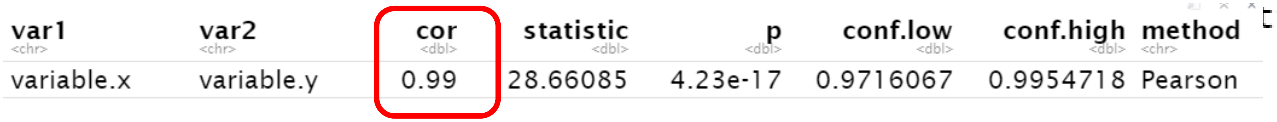
```
correlation.23 %>%
  filter(ID != 21) -> correlation.23.21

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 
tistic: 821.4 on 1 and 19 DF,  p-value: < 2.2e-16
```

```
Correlation.23.21 %>%
  cor_test(variable.x, variable.y)
```



var1 <chr>	var2 <chr>	cor <dbl>	statistic <dbl>	p <dbl>	conf.low <dbl>	conf.high <dbl>	method <chr>
variable.x	variable.y	0.99	28.66085	4.23e-17	0.9716067	0.9954718	Pearson

Extra exercise

Correlation: exam.anxiety.csv

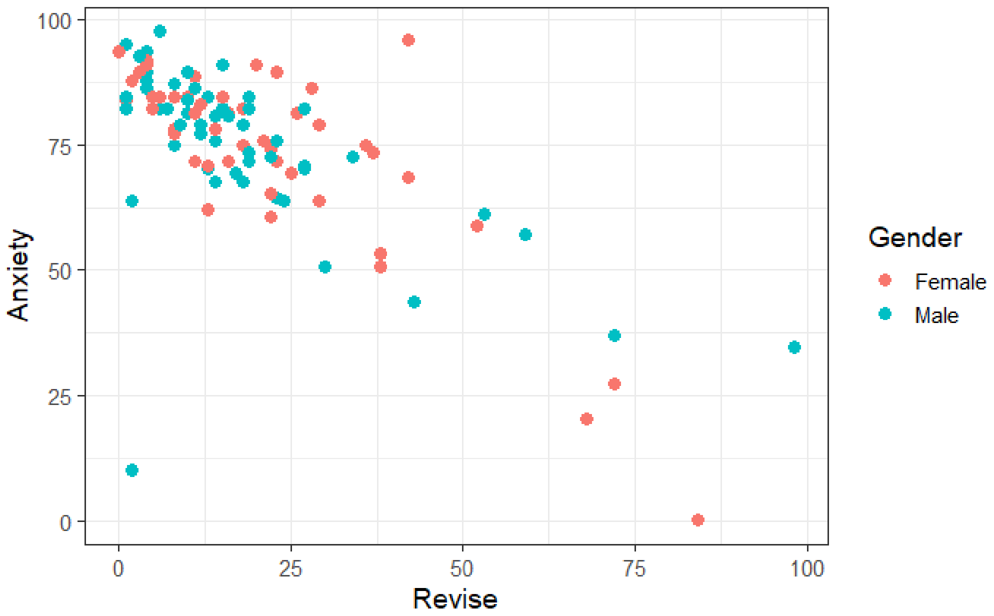
- **Question**: Is there a relationship between time spent revising and exam anxiety? And, if yes, are boys and girls different?

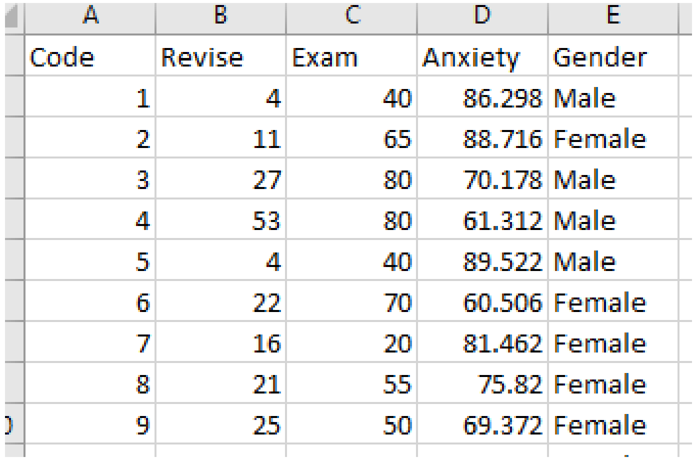
- Build a fit for the boys and a fit for the girls
  - `data %>% filter() lm(y~x, data=)`
- Plot the 2 lines of best fit on the same graph
  - `coefficients()  geom_abline()`
- Check the assumptions visually from the data and with the output for models
  - `par(mfrow=c(2,2))  plot(fit.male)`
- Filter out misbehaving values based on the standardised residuals
  - `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?

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

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





A	B	C	D	E
Code	Revise	Exam	Anxiety	Gender
1	4	40	86.298	Male
2	11	65	88.716	Female
3	27	80	70.178	Male
4	53	80	61.312	Male
5	4	40	89.522	Male
6	22	70	60.506	Female
7	16	20	81.462	Female
8	21	55	75.82	Female
9	25	50	69.372	Female

Correlation: exam anxiety.csv

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

```
exam.anxiety %>%
  filter(Gender=="Female") -> exam.anxiety.female

lm(Anxiety~Revise, data=exam.anxiety.female) -> fit.female

coefficients(fit.female) -> cf.fit.female
```

Fit for the females

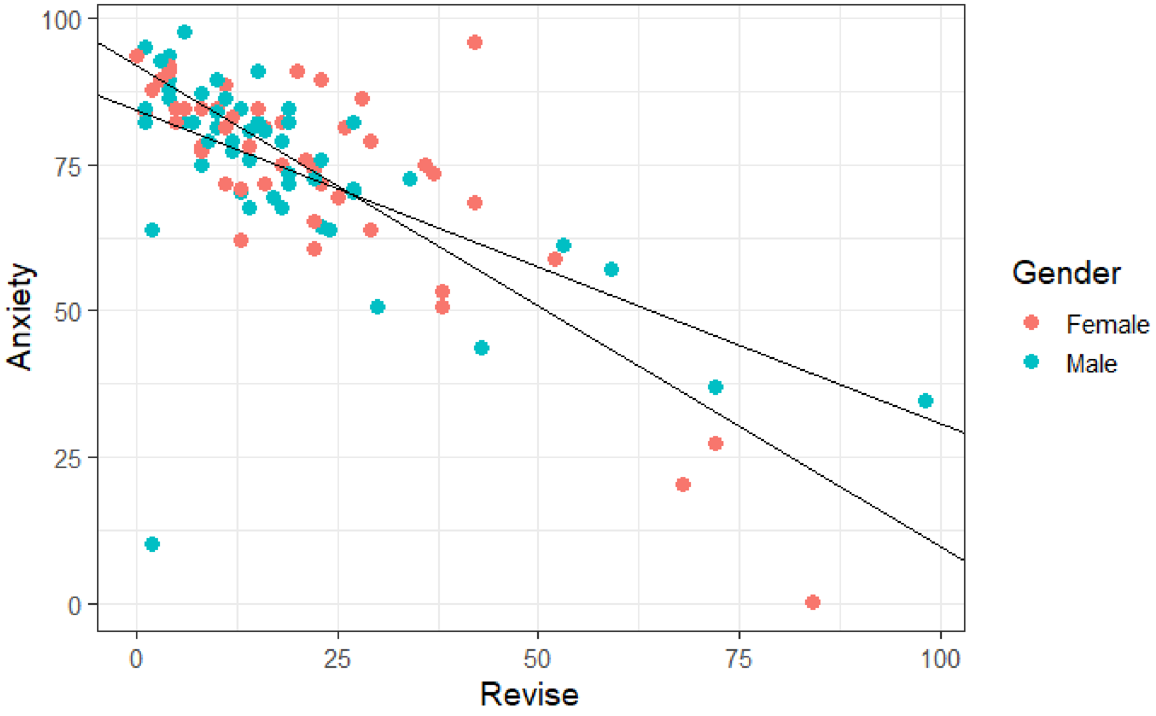
```
exam.anxiety %>%
  filter(Gender=="Male") -> exam.anxiety.male

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

coefficients(fit.male) -> cf.fit.male
```

Fit for the males

Correlation: exam anxiety.csv

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

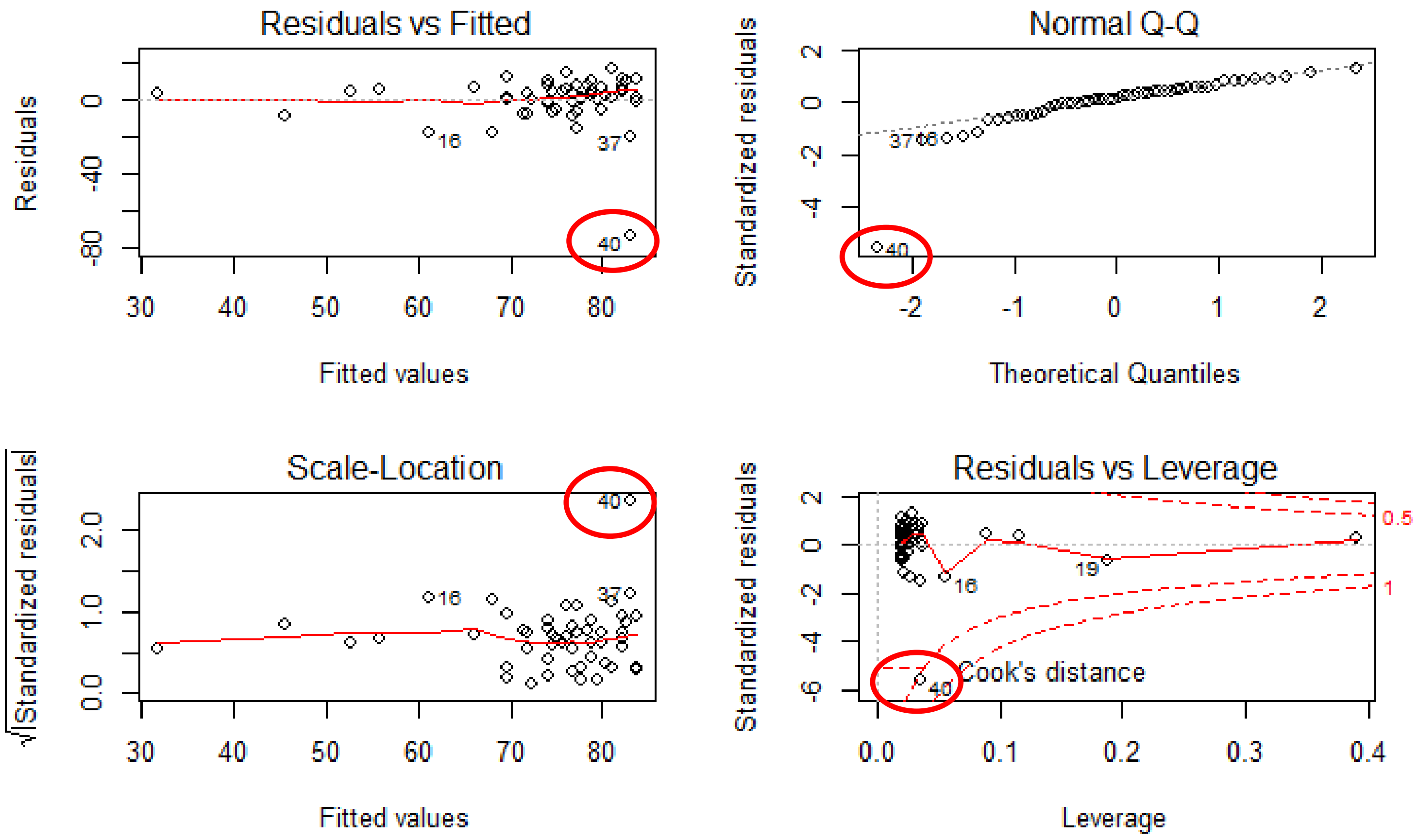
```
exam.anxiety %>%
  ggplot(aes(x=Revise, y=Anxiety, colour=Gender))+
  geom_point(size=3)+
  geom_abline(intercept=cf.fit.male[1], slope=cf.fit.male[2])+
  geom_abline(intercept=cf.fit.female[1], slope=cf.fit.female[2])
```



Correlation: exam anxiety.csv

Assumptions, outliers and influential cases

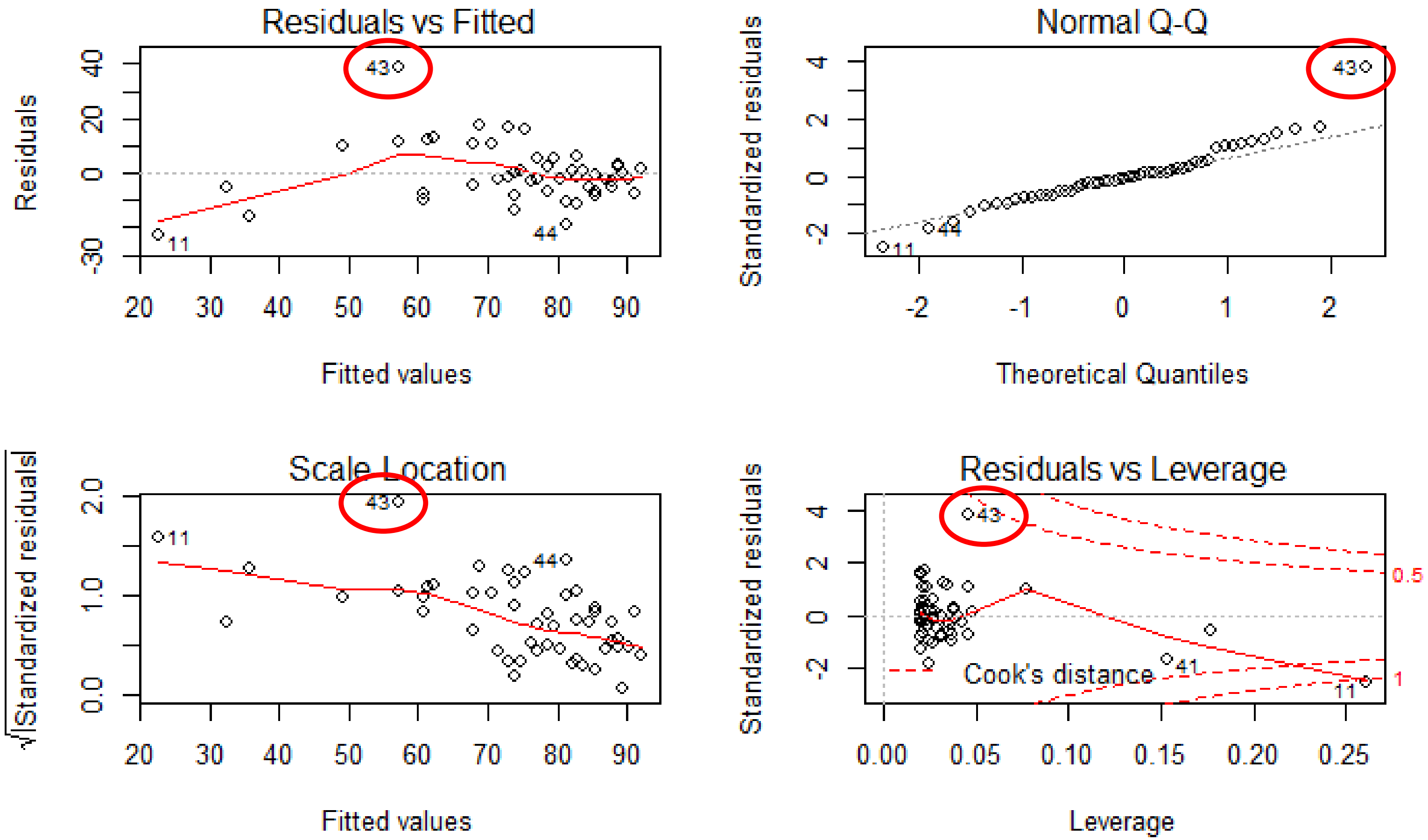
```
par(mfrow=c(2,2))
plot(fit.male)
```



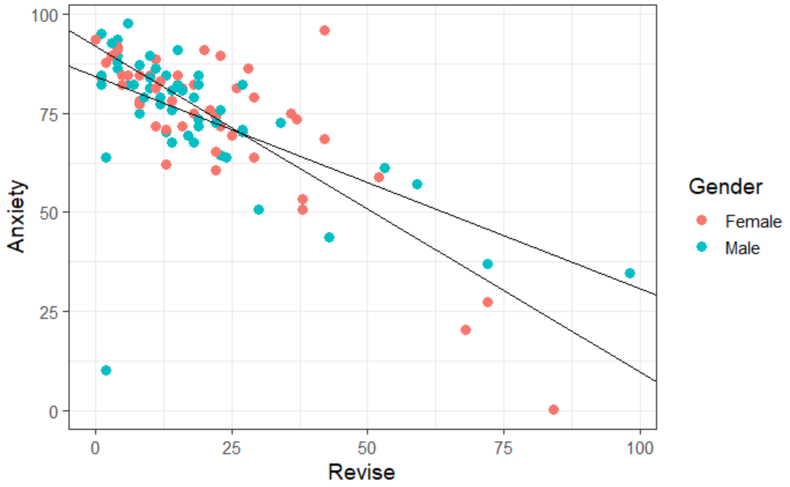
Correlation: exam anxiety.csv

Assumptions, outliers and influential cases

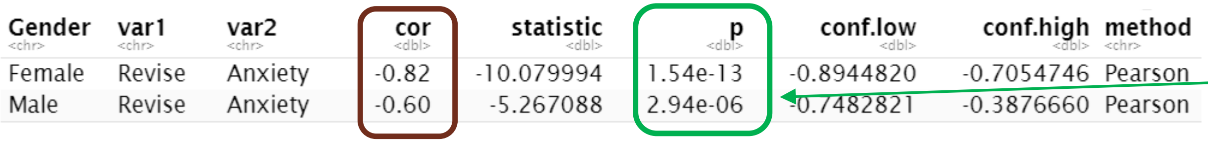
```
plot(fit.female)
```



Correlation: exam anxiety.csv



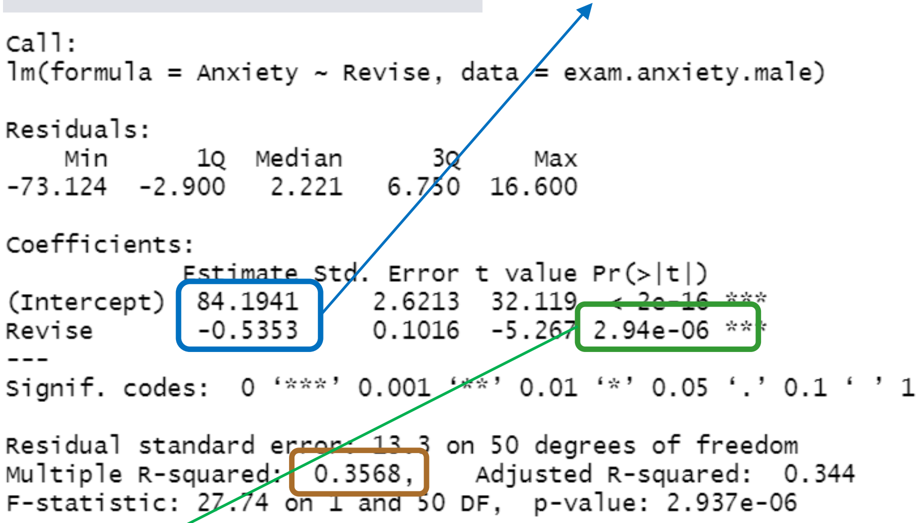
```
exam.anxiety %>%
  group_by(Gender) %>%
    cor_test(Revise, Anxiety) %>%
      ungroup()
```



Gender	var1	var2	cor	statistic	p	conf.low	conf.high	method
<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
Female	Revise	Anxiety	-0.82	-10.079994	1.54e-13	-0.8944820	-0.7054746	Pearson
Male	Revise	Anxiety	-0.60	-5.267088	2.94e-06	-0.7482821	-0.3876660	Pearson

```
summary(fit.male)
```

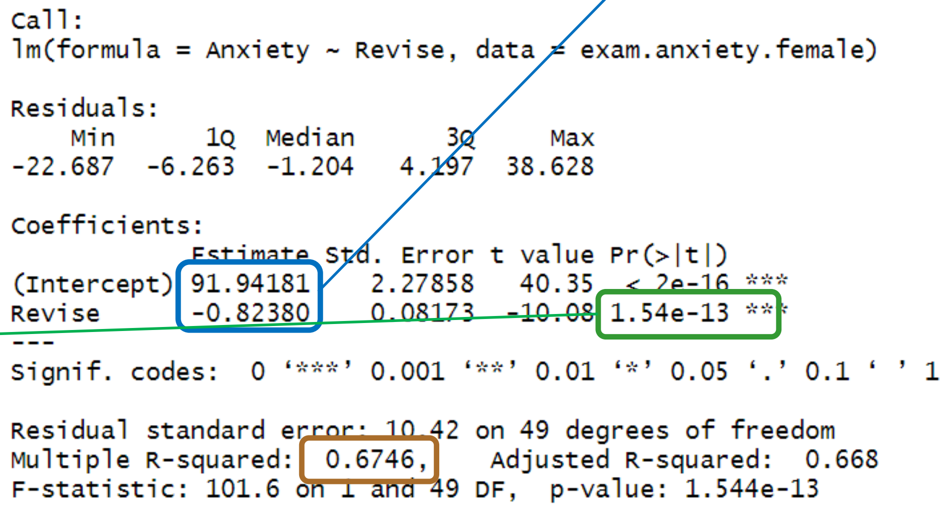
Anxiety=84.19 - 0.53*Revise



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

Residuals:
    Min      1Q  Median      3Q     Max 
-73.124  -2.900   2.221   6.750  16.600 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  84.1941     2.6213  32.119  < 2e-16 ***
Revise       -0.5353     0.1016  -5.267 2.94e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.3 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
```

```
summary(fit.female)
```

Anxiety=91.94 - 0.82*Revise

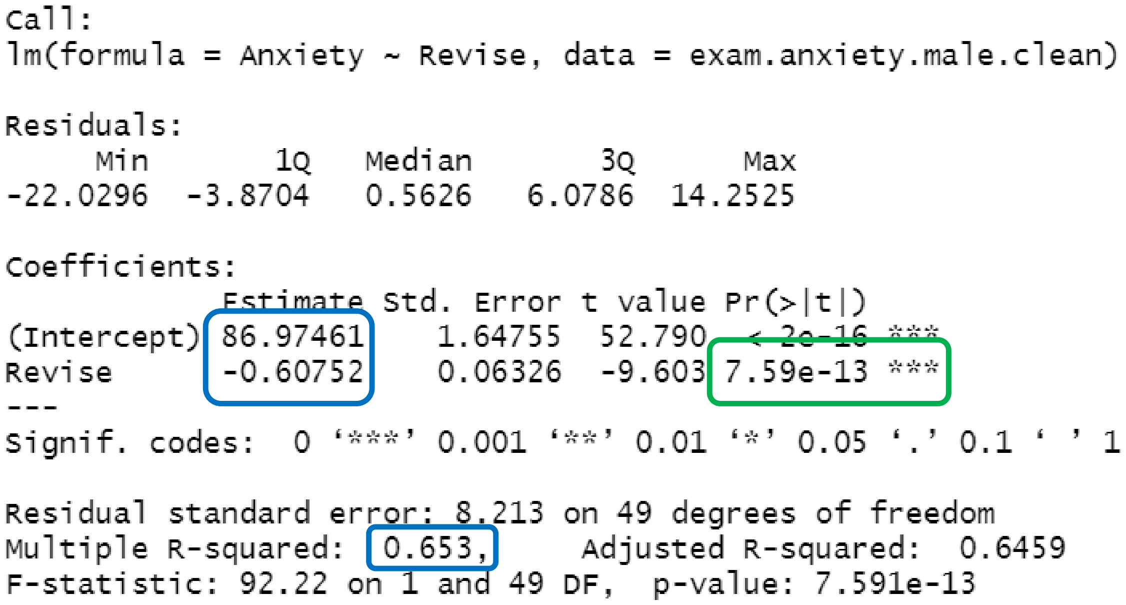


```
Call:
lm(formula = Anxiety ~ Revise, data = exam.anxiety.female)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.687  -6.263  -1.204   4.197  38.628 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 91.94181    2.27858   40.35  < 2e-16 ***
Revise      -0.82380    0.08173  -10.08 1.54e-13 ***
---
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
```

Correlation: exam.anxiety.csv

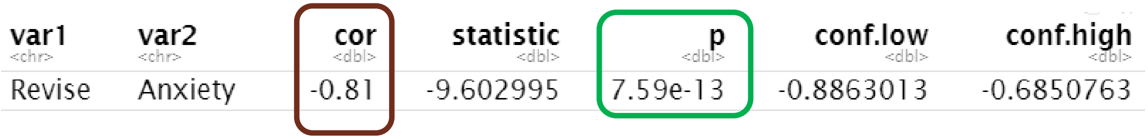
Influential outliers: Boys

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

exam.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: 7.591e-13
```

```
exam.anxiety.male.clean %>%
  cor_test(Revise, Anxiety)
```

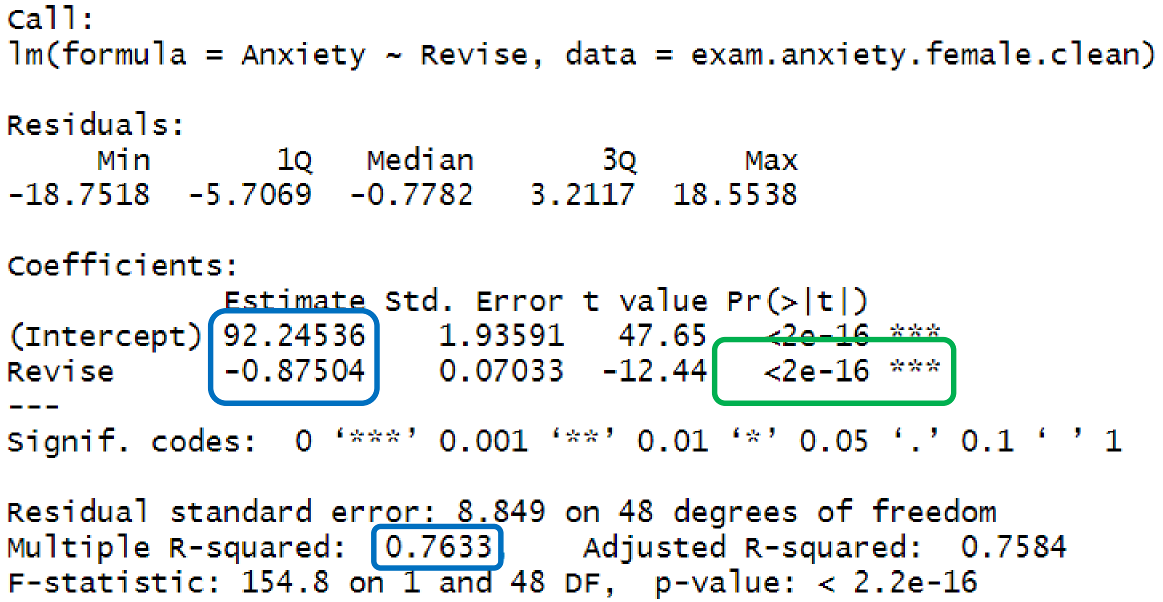


var1 <chr>	var2 <chr>	cor <dbl>	statistic <dbl>	p <dbl>	conf.low <dbl>	conf.high <dbl>
Revise	Anxiety	-0.81	-9.602995	7.59e-13	-0.8863013	-0.6850763

Correlation: exam.anxiety.csv

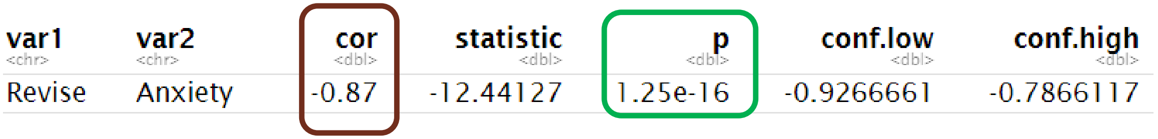
Influential outliers: Girls

```
rstandard(fit.female) -> st.resid.f

exam.anxiety.female %>%
  add_column(st.resid.f) %>%
  filter(abs(st.resid.f) < 3) -> exam.anxiety.female.clean

lm(Anxiety~Revise, data=exam.anxiety.female.clean) -> fit.female2

summary(fit.female2)
```



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

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: 8.849 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
```

```
exam.anxiety.female.clean %>%
  cor_test(Revise, Anxiety)
```

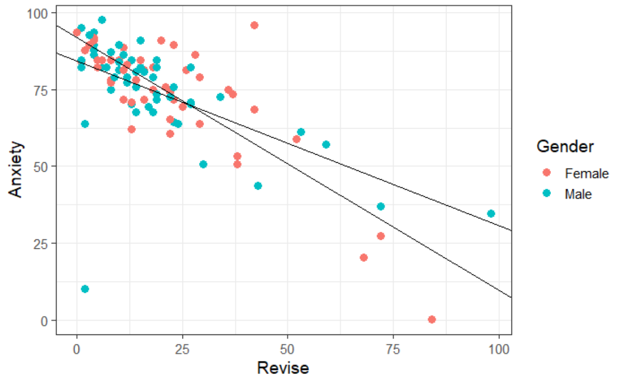


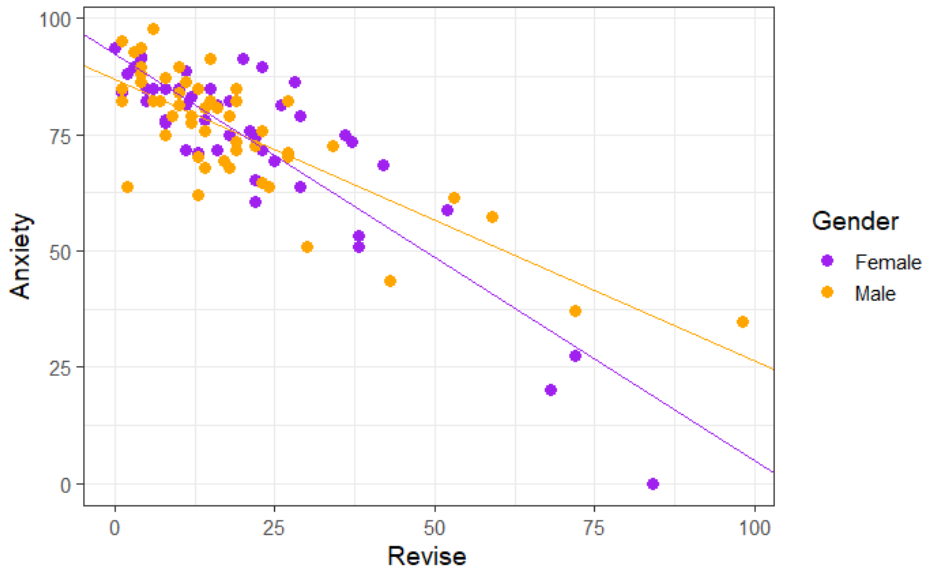
var1	var2	cor	statistic	p	conf.low	conf.high
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Revise	Anxiety	-0.87	-12.44127	1.25e-16	-0.9266661	-0.7866117

Correlation: exam.anxiety.csv

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

```
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"))
```

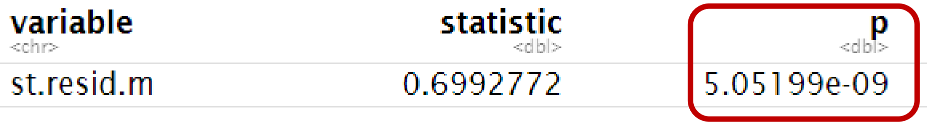




Correlation: exam.anxiety

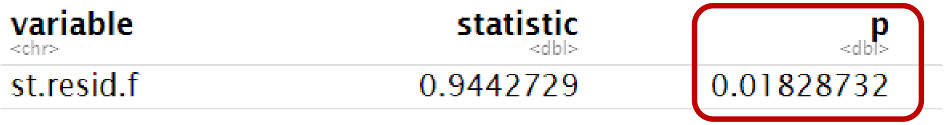
Influential outliers: Another check

```
exam.anxiety.male %>%
  shapiro_test(st.resid.m)
```



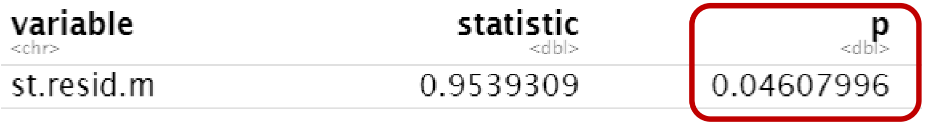
variable	statistic	p
<chr>	<dbl>	<dbl>
st.resid.m	0.6992772	5.05199e-09

```
exam.anxiety.female %>%
  shapiro_test(st.resid.f)
```



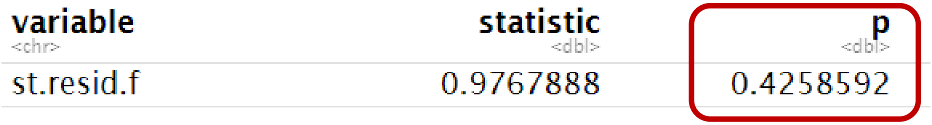
variable	statistic	p
<chr>	<dbl>	<dbl>
st.resid.f	0.9442729	0.01828732

```
exam.anxiety.male.clean %>%
  shapiro_test(st.resid.m)
```



variable	statistic	p
<chr>	<dbl>	<dbl>
st.resid.m	0.9539309	0.04607996

```
exam.anxiety.female.clean %>%
  shapiro_test(st.resid.f)
```

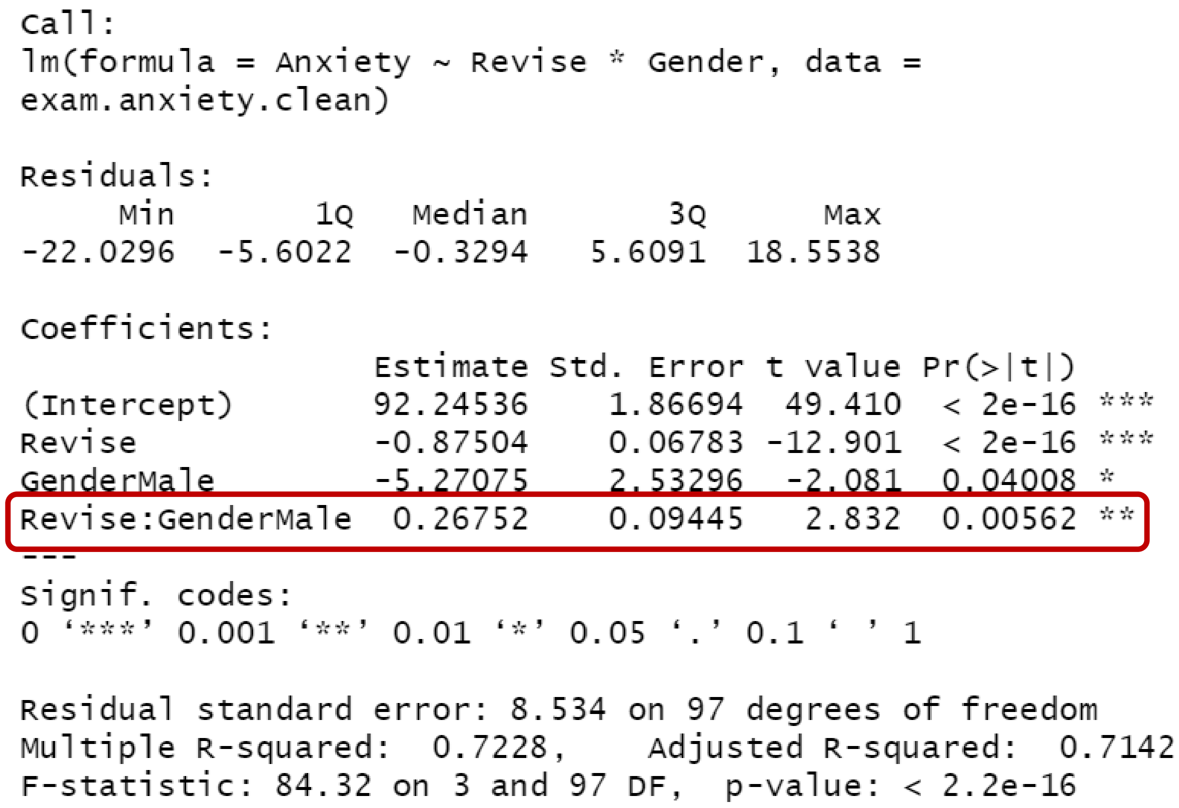


variable	statistic	p
<chr>	<dbl>	<dbl>
st.resid.f	0.9767888	0.4258592

Correlation: exam anxiety.csv

- Difference between boys and girls?

```
lm(Anxiety~Revise*Gender, data=exam.anxiety.clean) -> fit.genders

summary(fit.genders)
```



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

Residuals:
     Min       1Q   Median       3Q      Max 
-22.0296  -5.6022  -0.3294   5.6091  18.5538 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      92.24536    1.86694  49.410  < 2e-16 ***
Revise           -0.87504    0.06783 -12.901  < 2e-16 ***
GenderMale       -5.27075    2.53296  -2.081  0.04008 *  
Revise:GenderMale  0.26752    0.09445   2.832  0.00562 ** 
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.534 on 97 degrees of freedom
Multiple R-squared:  0.7228,	Adjusted R-squared:  0.7142 
F-statistic: 84.32 on 3 and 97 DF,  p-value: < 2.2e-16
```