

Analysis of Quantitative data

One-Way + Two-Way ANOVA

Anne Segonds-Pichon
v2020-12

Comparison between more than 2 groups

One factor = One predictor

One-Way ANOVA

Signal-to-noise ratio

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

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

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

Analysis of variance: how does it work?

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

= F ratio

One-Way Analysis of variance

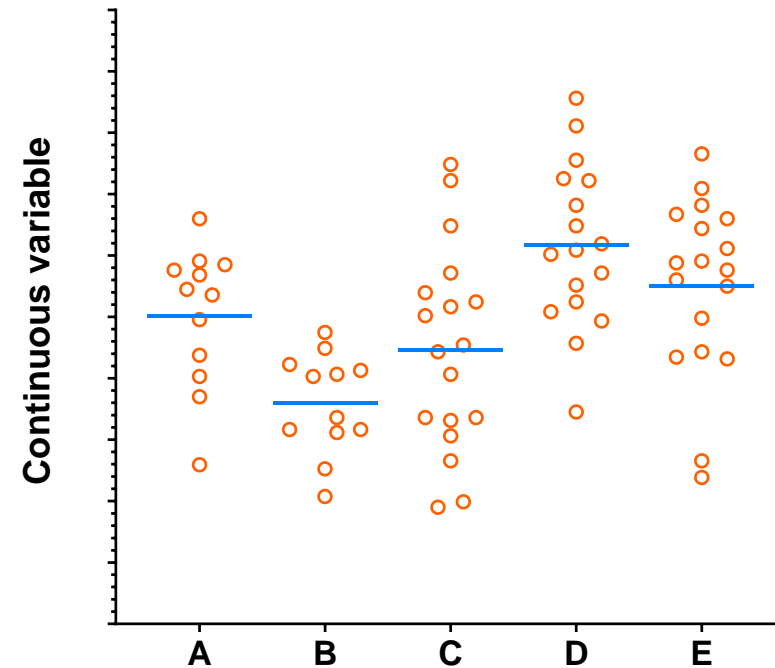
Step 1: Omnibus test

- 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

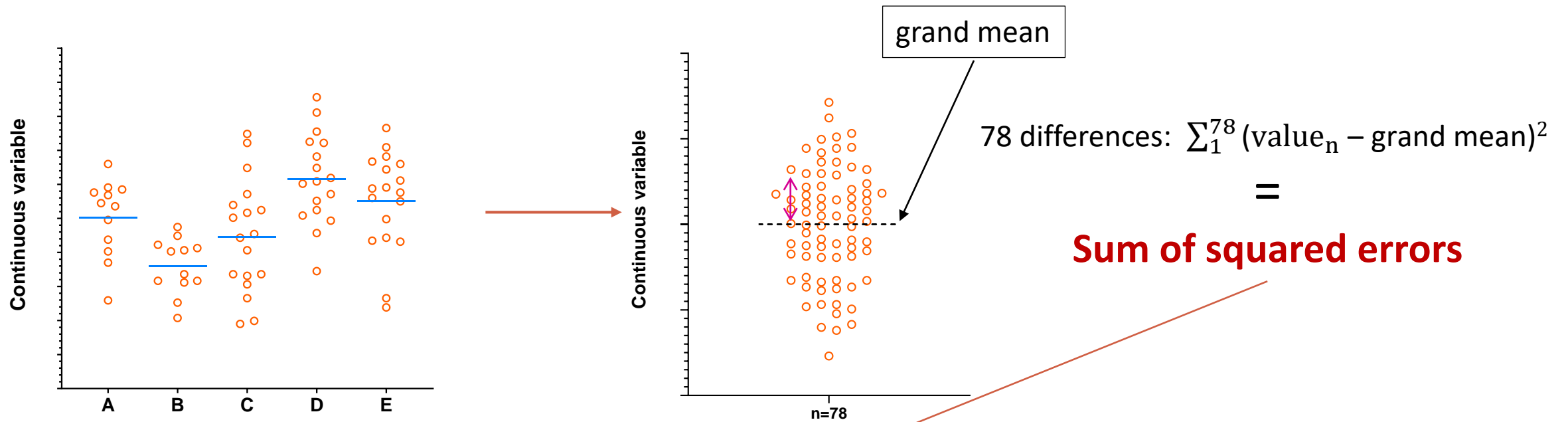
- They tell us if there are differences between the means pairwise.

Analysis of variance: how does it work?



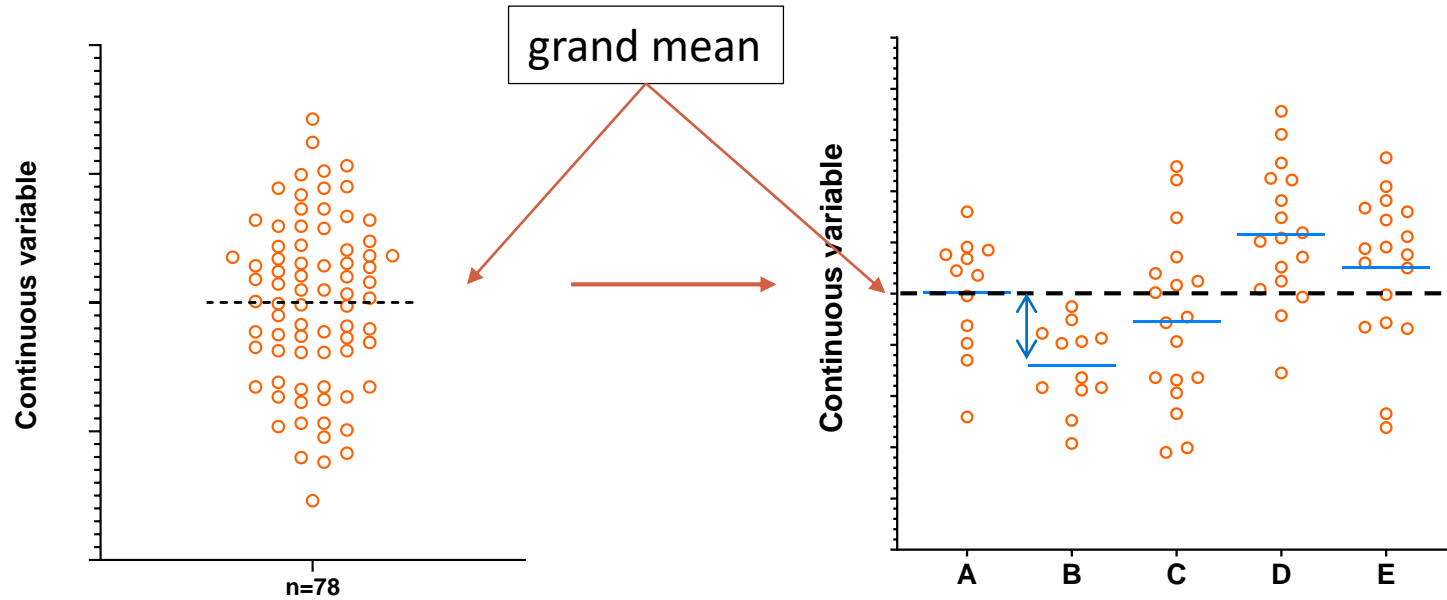
Source of variation	Sum of Squares	df	Mean Square	F	p-value
Between Groups	18.1	4	4.5	6.32	0.0002
Within Groups	51.8	73	0.71		
Total	69.9				

Analysis of variance: how does it work?



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?



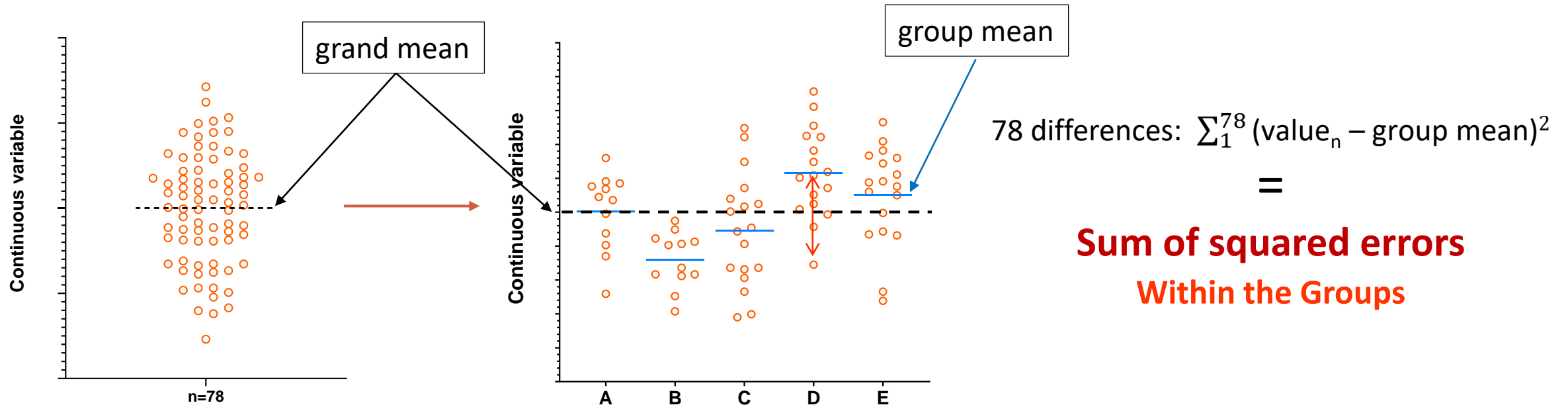
5 differences: $\sum_1^5 (\text{mean}_n - \text{grand mean})^2$

=

Sum of squared errors
Between the groups

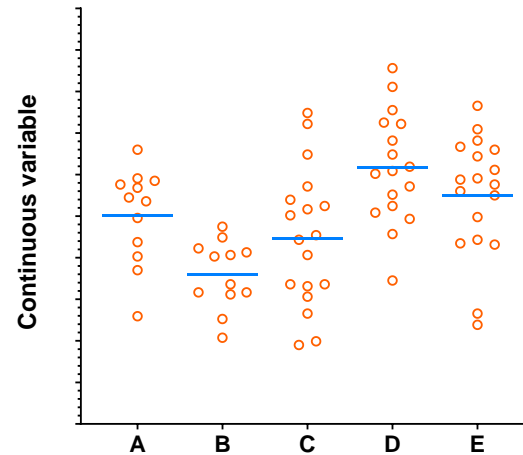
Source of variation	Sum of Squares	df	Mean Square	F	p-value
Between Groups	18.1				
Within Groups					
Total	69.9				

Analysis of variance: how does it work?



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				

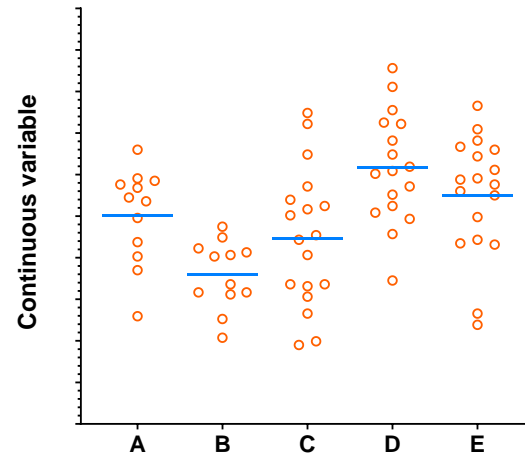
df: degree of freedom with $df = n-1$

n = number of values, k =number of groups

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

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

Analysis of variance: how does it work?



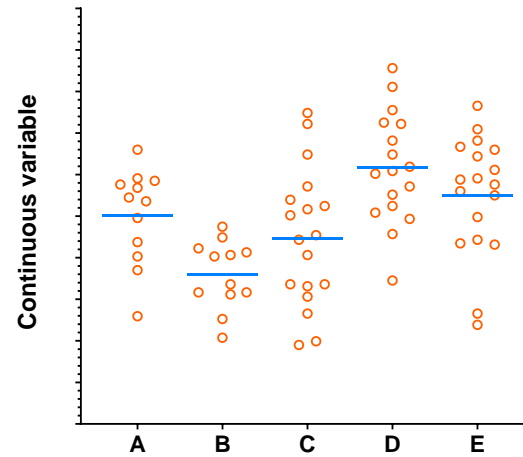
	Source of variation	Sum of Squares	df	Mean Squares	F ratio	p-value
Signal	Between Groups	18.1	4	4.5		
Noise	Within Groups	51.8	73	0.71		
	Total	69.9				

df: degree of freedom with $df = n-1$

$$18.2/4 = 4.5 \quad 51.8/73 = 0.71$$

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	6.34	0.0002
Within Groups	51.8	73	0.71		
Total	69.9				

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

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

Comparison of more than 2 means

- Running multiple tests on the same data increases the **familywise error rate**.
- What is the familywise error rate?
 - The error rate across tests conducted on the same experimental data.
- One of the basic rules ('laws') of probability:
 - The Multiplicative Rule: The probability of the joint occurrence of 2 or more independent events is the product of the individual probabilities.

$$P(A,B) = P(A) \times P(B)$$

For example:

$$P(2 \text{ Heads}) = P(\text{head}) \times P(\text{head}) = 0.5 \times 0.5 = 0.25$$

Familywise error rate

- **Example:** All pairwise comparisons between 3 groups A, B and C:
 - A-B, A-C and B-C
- Probability of making the Type I Error: **5%**
 - The probability of not making the Type I Error is 95% ($=1 - 0.05$)
- Multiplicative Rule:
 - Overall probability of no Type I errors is: $0.95 * 0.95 * 0.95 = 0.857$
- 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**: $\alpha_{\text{adjust}} = 0.05/n$ comparisons e.g. 3 comparisons: $0.05/3=0.016$
 - Problem: very conservative leading to loss of power (lots of false negative)
 - 10 comparisons: threshold for significance: $0.05/10: 0.005$
 - Pairwise comparisons across 20.000 genes ☹️
- **FDR: Benjamini-Hochberg**: the procedure controls the expected proportion of “discoveries” (significant tests) that are false (false positive).
 - Less stringent control of Type I Error than FWER procedures which control the probability of at least one Type I Error
 - More power at the cost of increased numbers of Type I Errors.
- **Difference between FWER and FDR:**
 - a p-value of 0.05 implies that 5% of all tests will result in false positives.
 - a FDR adjusted p-value (or **q-value**) of 0.05 implies that 5% of significant tests will result in false positives.

One-Way Analysis of variance

Step 1: Omnibus test

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

Step 2: Post-hoc tests

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

Example: protein.expression.csv

- **Question:** is there a difference in protein expression between the 5 cell lines?
- **1 Plot the data**
- **2 Check the assumptions for parametric test**

Exercise 6: One-way ANOVA: Data Exploration

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()`

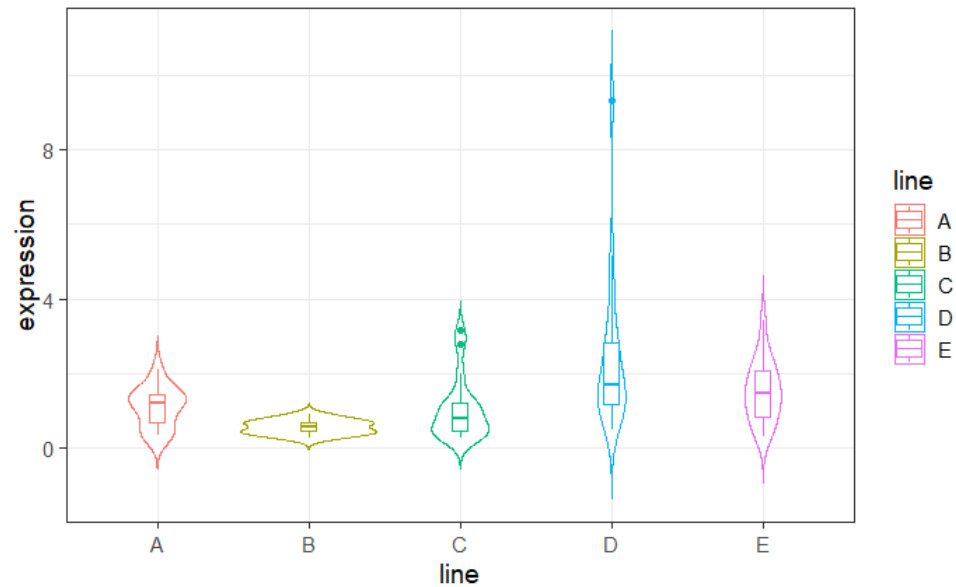
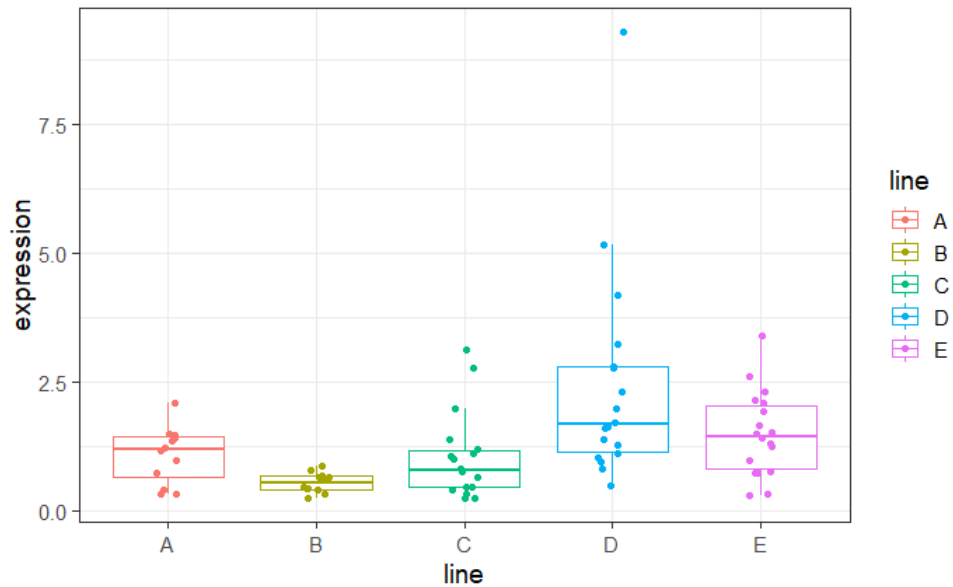
Exercise 6: One-way ANOVA : Data Exploration - Answers

```
protein %>%
```

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

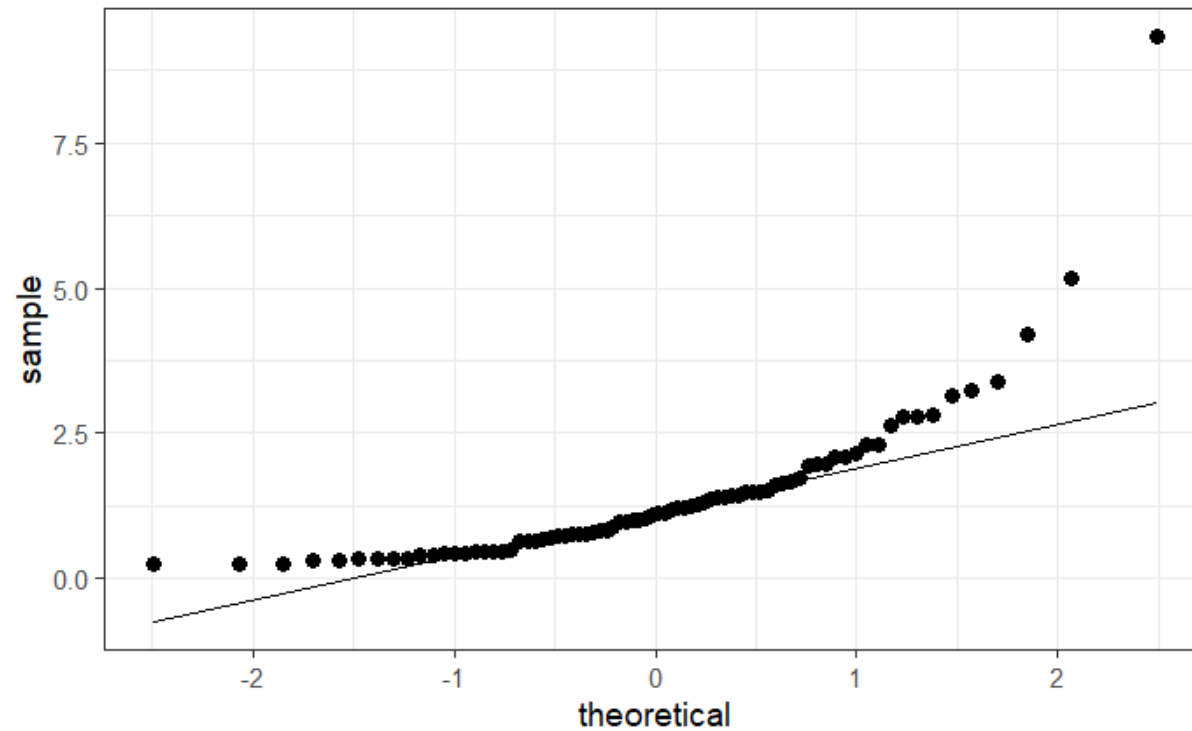
```
protein %>%
```

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



Exercise 6: One-way ANOVA – Answers

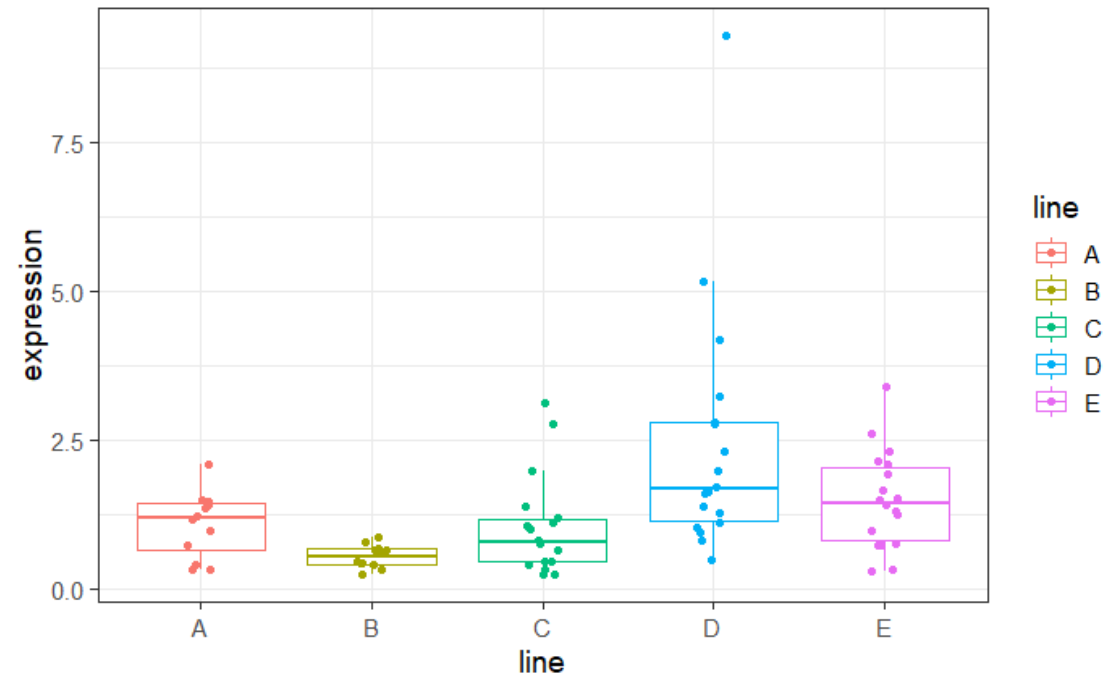
```
protein %>%  
  ggplot(aes(sample = expression))+  
    stat_qq(size=3)+  
    stat_qq_line()
```



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

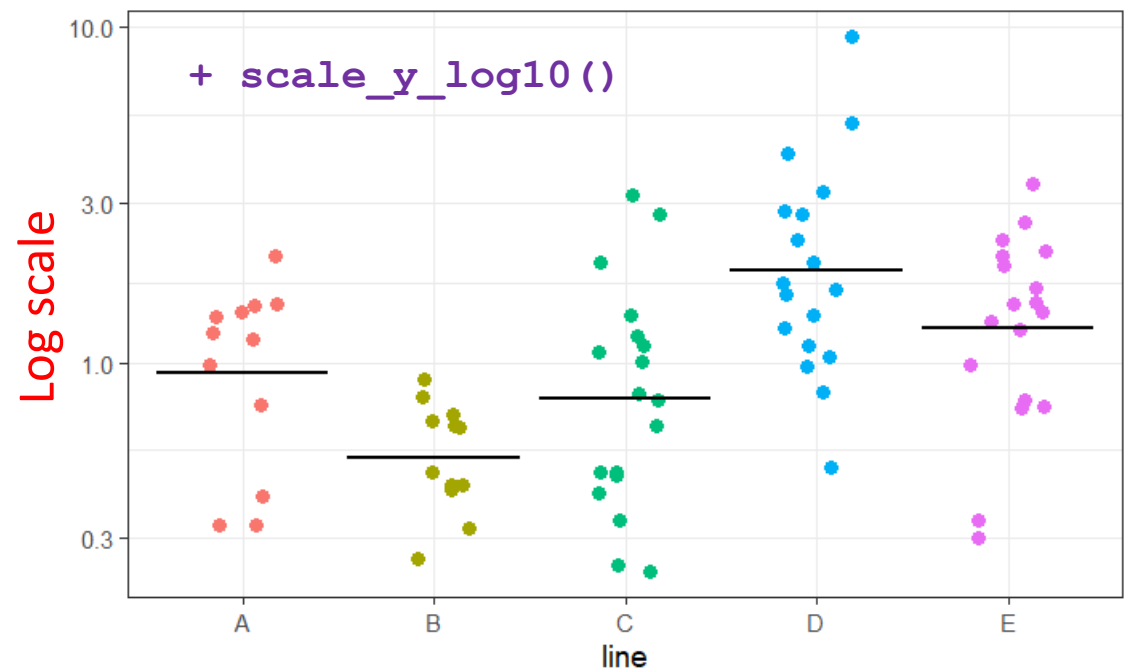
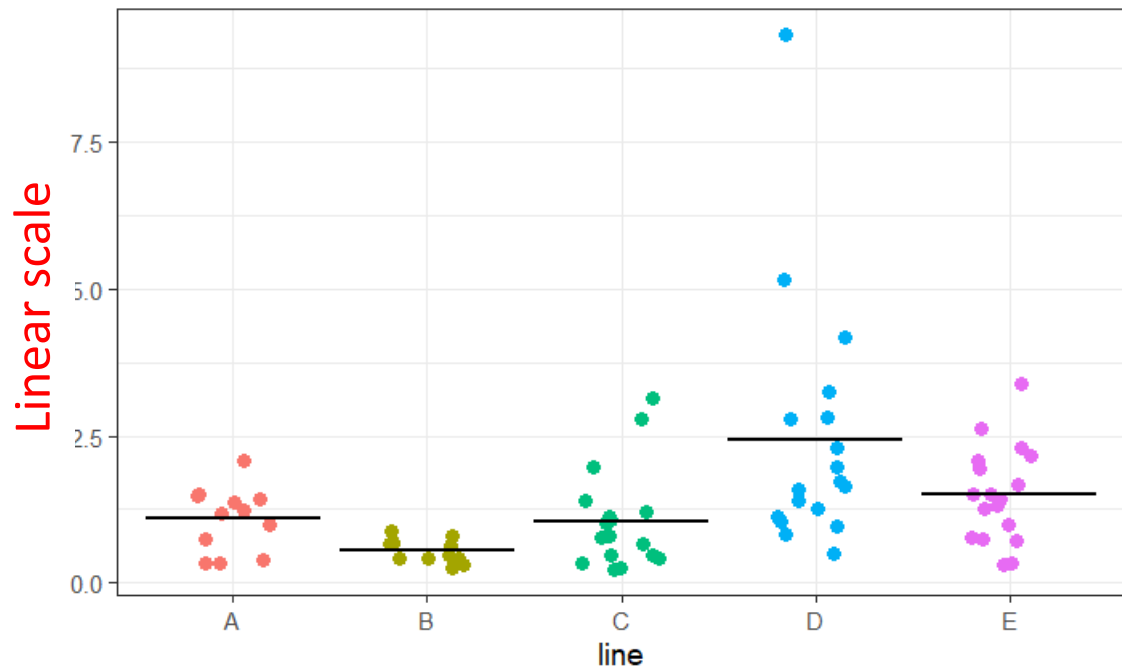
```
protein %>%  
  group_by(line) %>%  
  shapiro_test(expression) %>%  
  ungroup()
```

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)
```



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

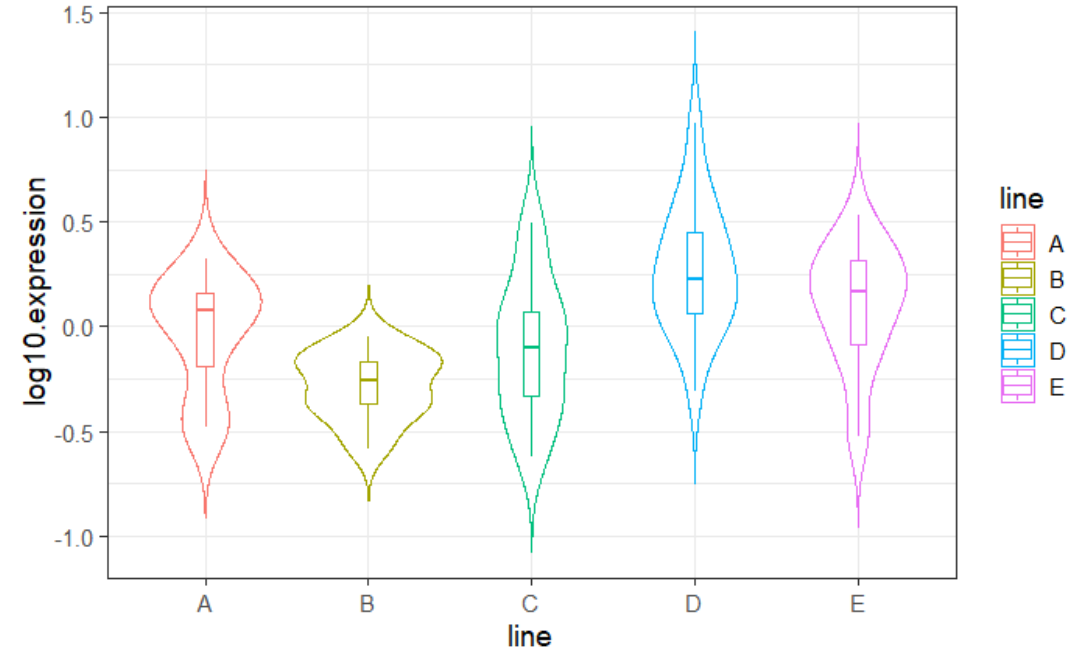
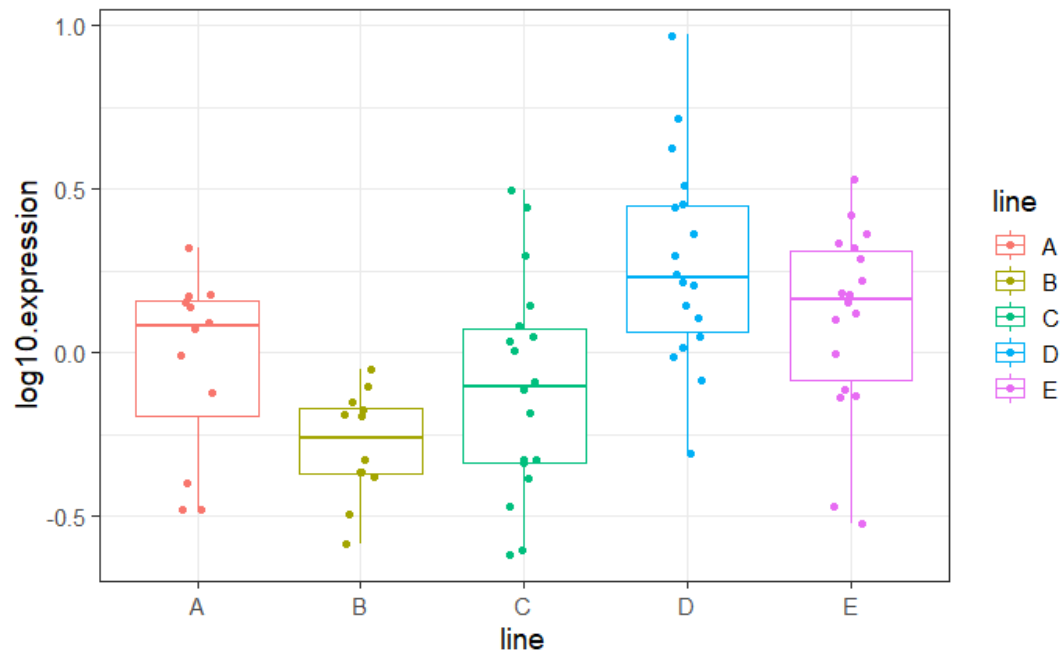
One-way ANOVA: change of scale

```
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)
```

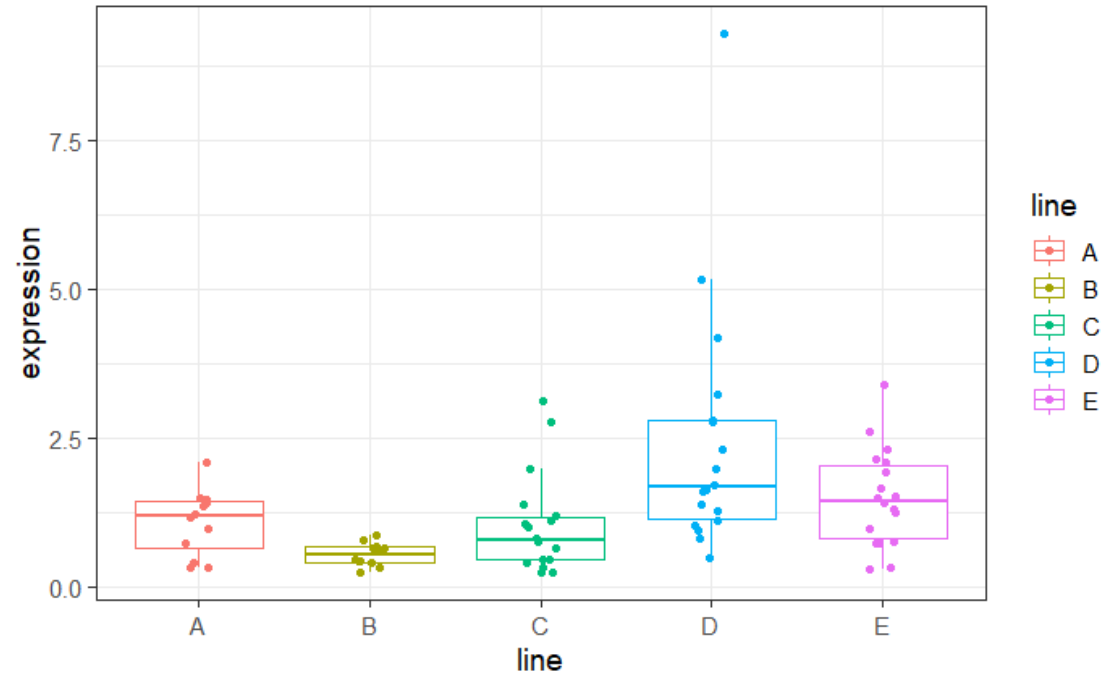


One-way ANOVA – Outliers identification

```
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

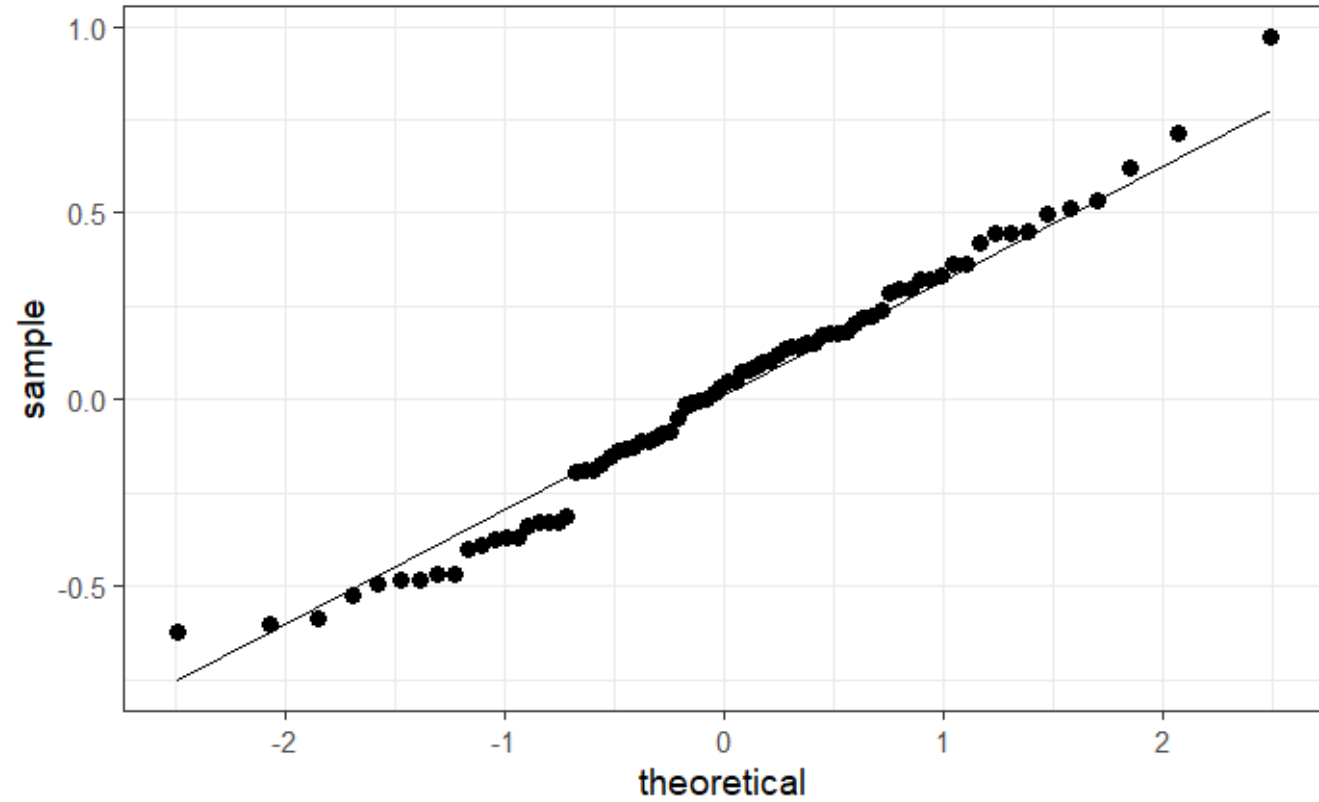
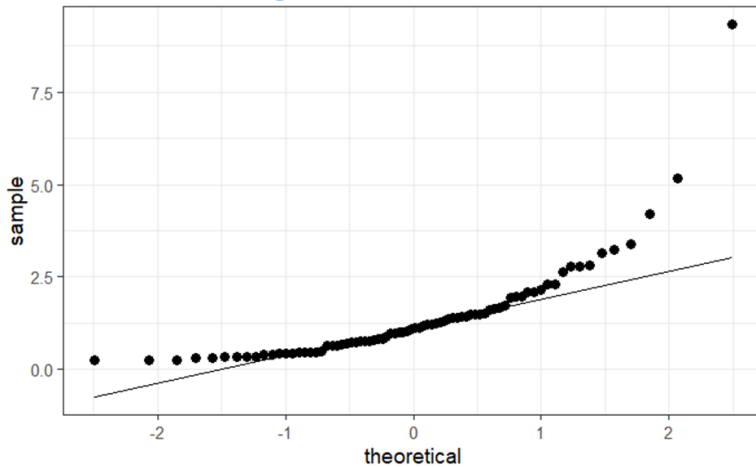
3 rows



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()
```

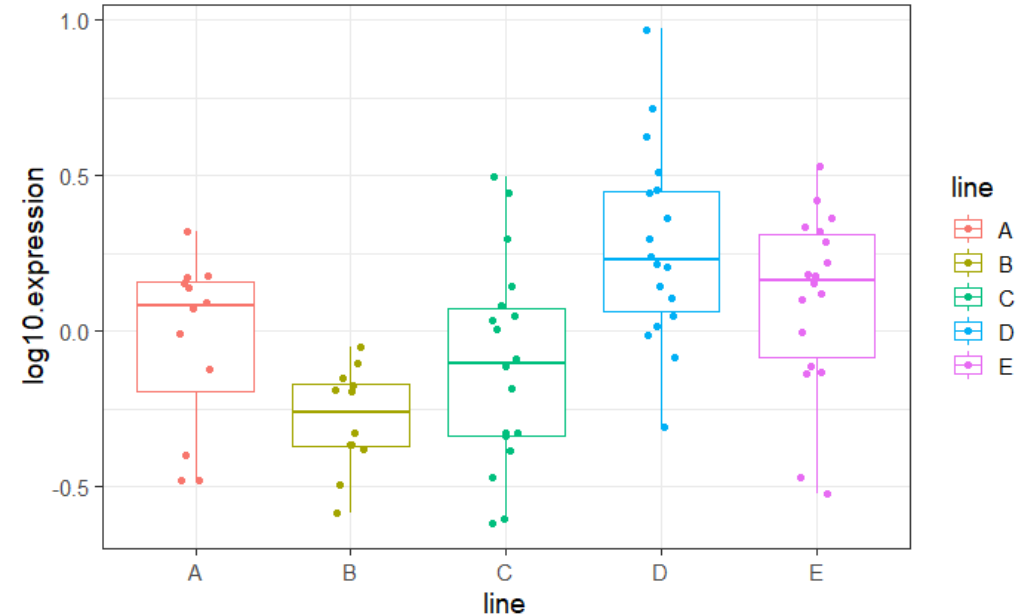
line <chr>	variable <chr>	statistic <dbl>	p <dbl>
A	log10.expression	0.8542464	0.04143953
B	log10.expression	0.9458450	0.57725321
C	log10.expression	0.9657060	0.71417958
D	log10.expression	0.9868425	0.99348831
E	log10.expression	0.9313425	0.20502703

First assumption ✓ish

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

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

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

Default



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	group1	group2	estimate	conf.low	conf.high	p.adj	p.adj.signif
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<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 effect size (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)
```

```
line          Df Sum Sq Mean Sq F value    Pr(>F)
Residuals    73  6.046   0.0828
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

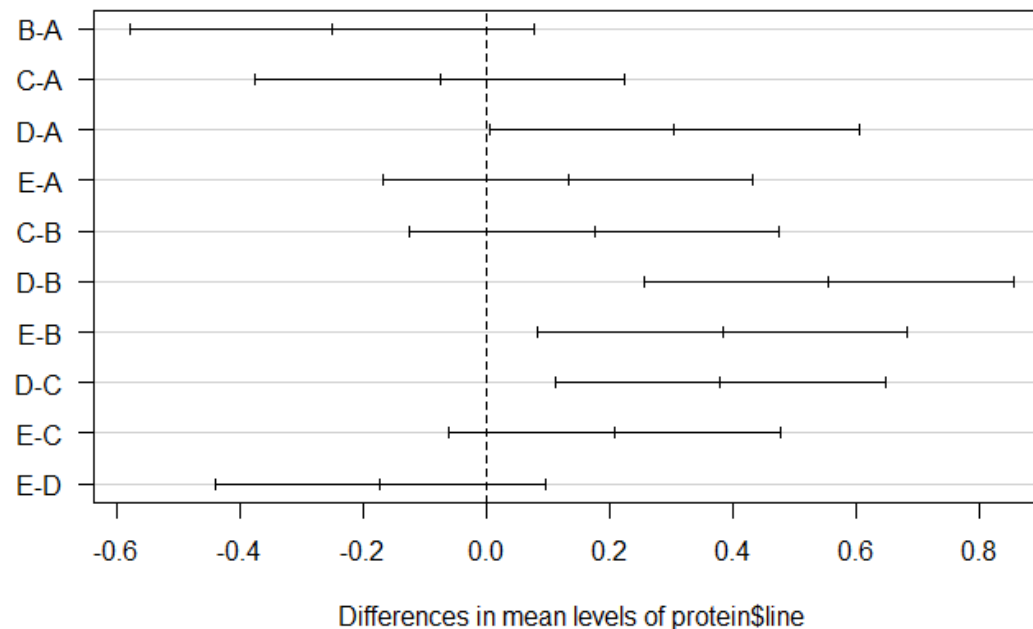
```
TukeyHSD(anova.log.protein,"line")
```

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

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

\$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

95% family-wise confidence level

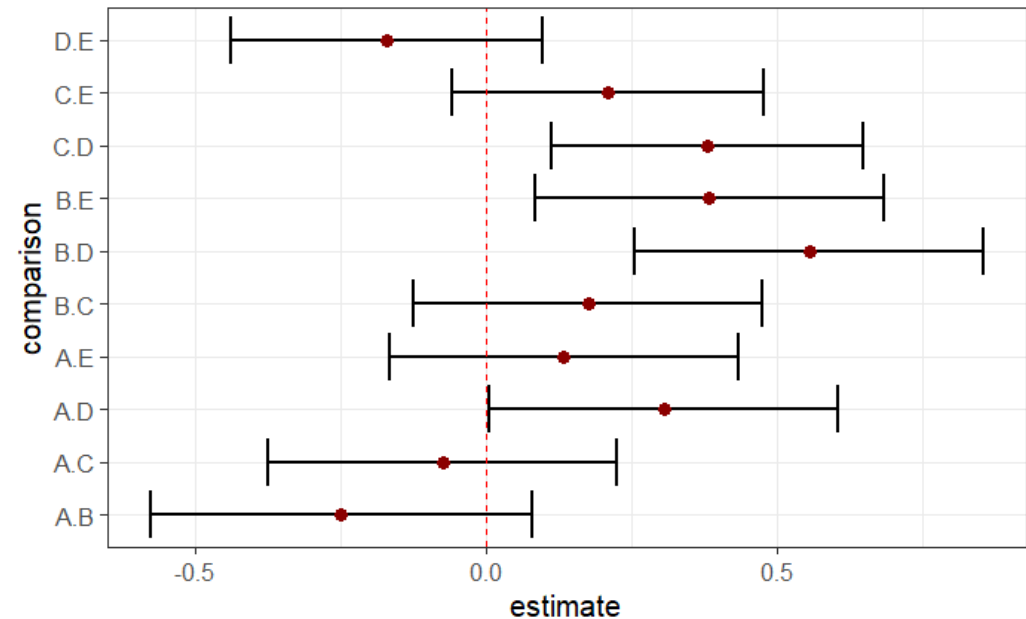


Analysis of variance (tidyverse)

To plot confidence intervals

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

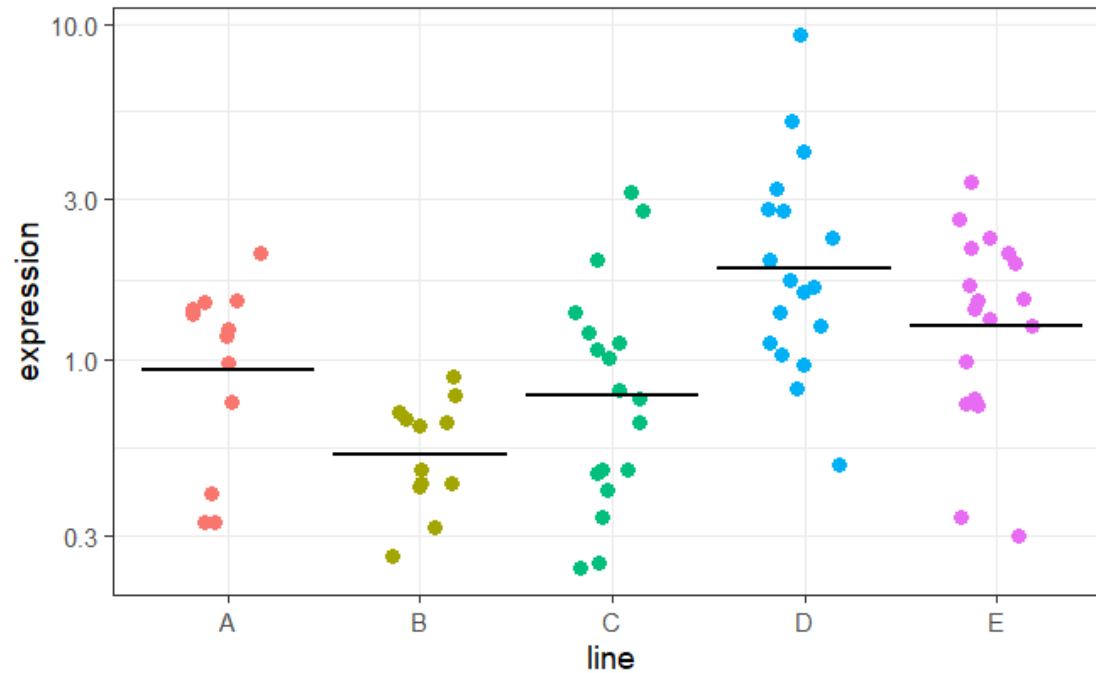
term	group1	group2	null.value	estimate	conf.low	conf.high	p.adj	p.adj.signif	comparison
line	A	B	0	-0.25024832	-0.578882494	0.07838585	2.19e-01	ns	A.B
line	A	C	0	-0.07499724	-0.374997820	0.22500335	9.56e-01	ns	A.C
line	A	D	0	0.30549397	0.005493391	0.60549456	4.39e-02	*	A.D
line	A	E	0	0.13327517	-0.166725416	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

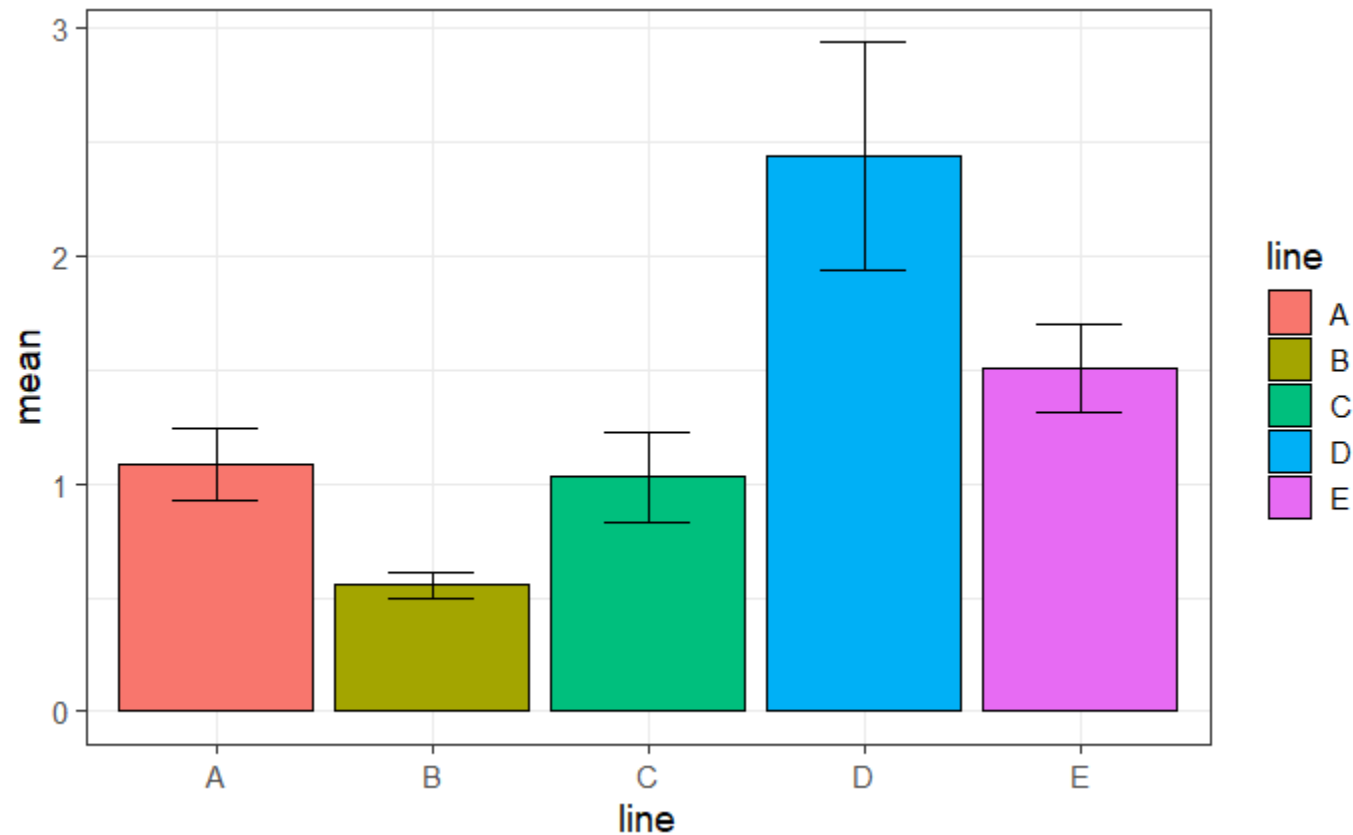
```
protein %>%  
  ggplot(aes(line, expression, colour=line))+  
  geom_jitter(height = 0, width=0.2, size=3, show.legend=FALSE)+  
  stat_summary(geom="errorbar", fun=mean, fun.min=mean, fun.max = mean, colour="black", size=1)+  
  scale_y_log10()
```



Analysis of variance

```
protein %>%
```

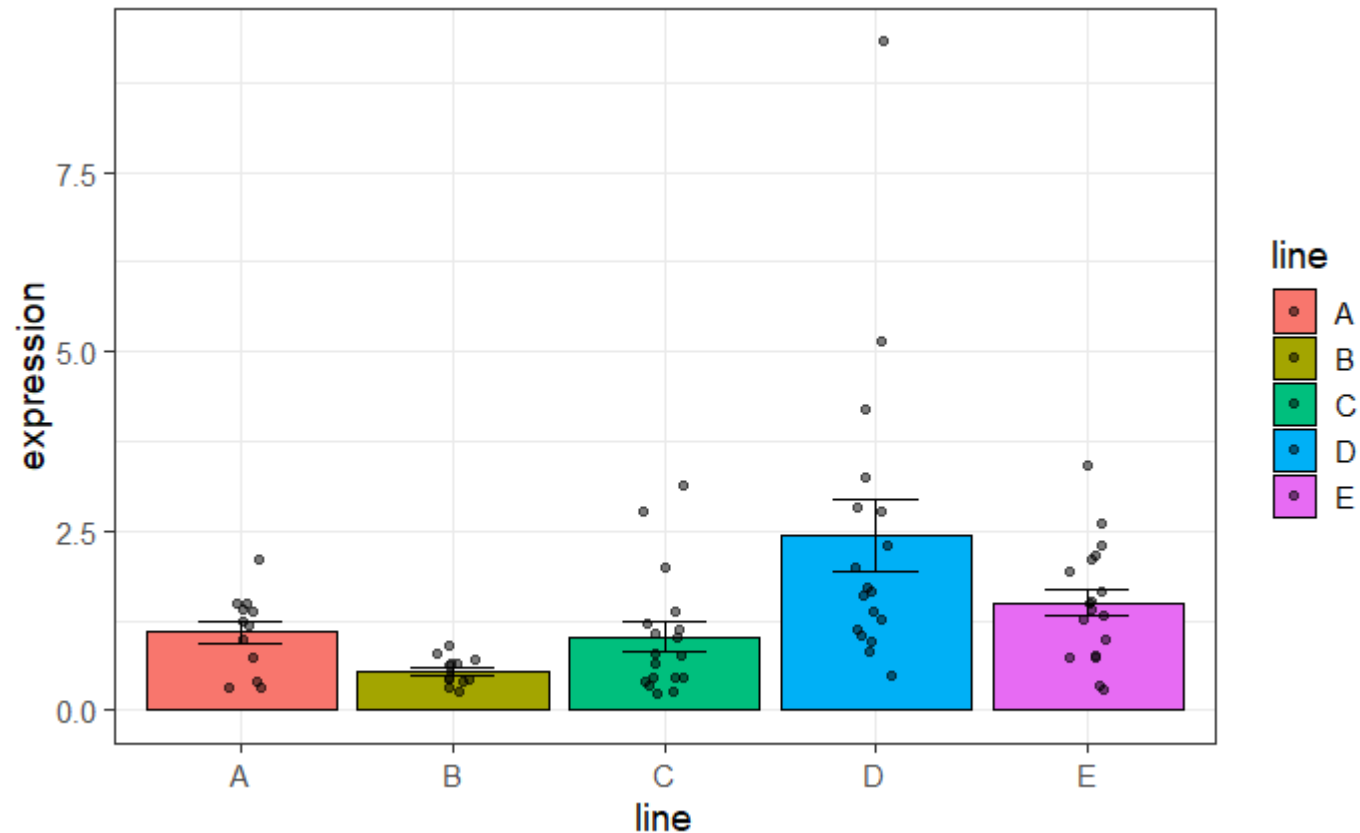
```
  ggplot(aes(x=line, y=expression, fill=line)) +  
    geom_bar(stat = "summary", fun="mean", colour="black")+  
    stat_summary(geom="errorbar", colour="black", width=0.4)
```



Analysis of variance

```
protein %>%
```

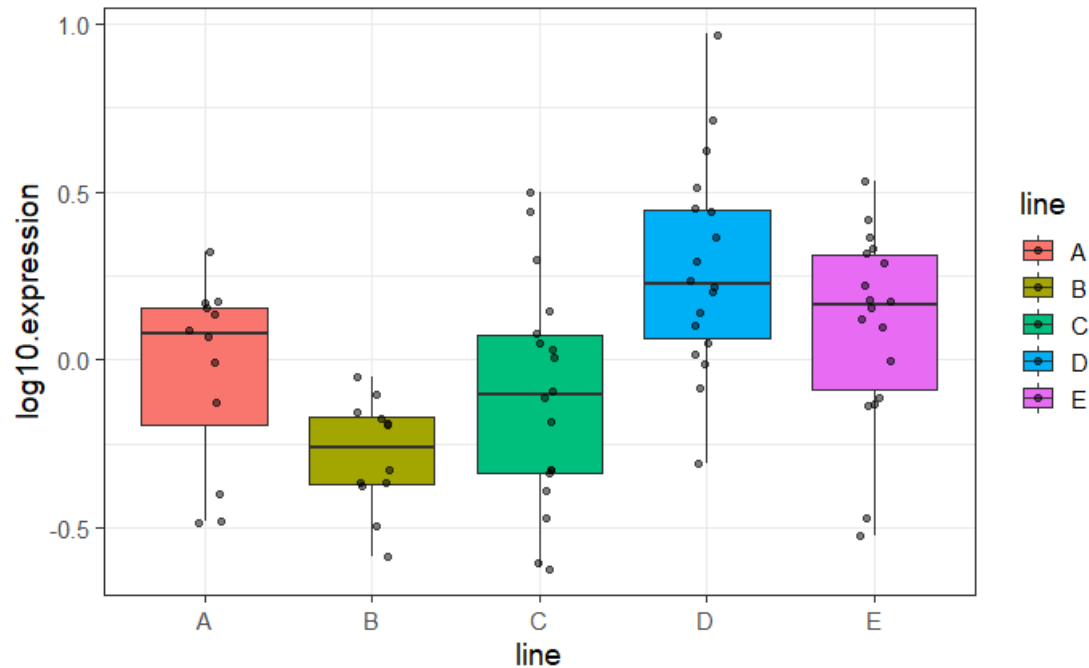
```
  ggplot(aes(x=line, y=expression, fill=line)) +  
    geom_bar(stat="summary", fun="mean", colour="black")+  
    stat_summary(geom="errorbar", colour="black", width=0.4)+  
    geom_jitter(height=0, width=0.1, alpha=0.5)
```



Analysis of variance

```
protein %>%
```

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



Exercise 7: Repeated measures ANOVA

neutrophils.long.csv



- A researcher is looking at the difference between 4 cell groups. He has run the experiment 5 times. Within each experiment, he has neutrophils from a WT (control), a KO, a KO+Treatment 1 and a KO+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

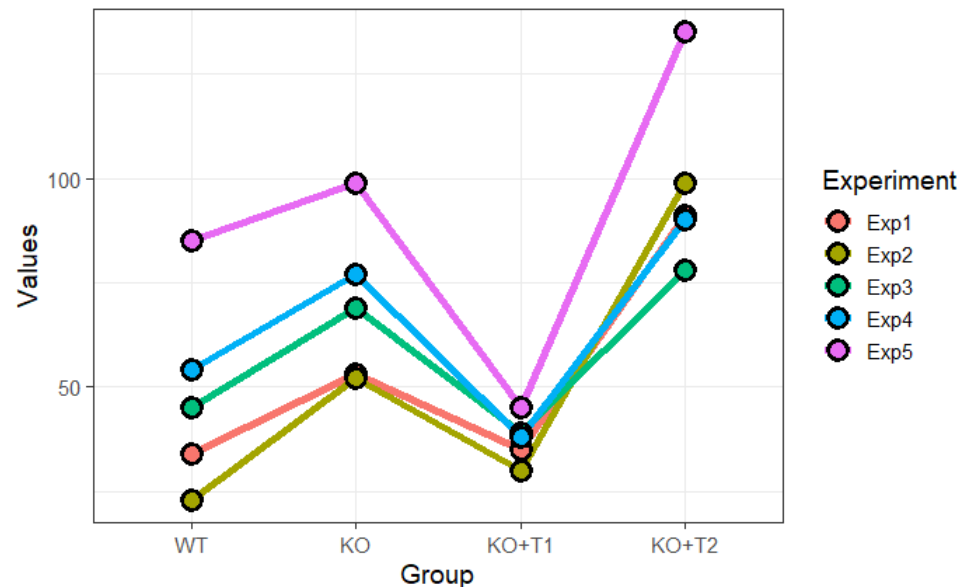
neutrophils.long.csv

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



```
neutrophils.long %>%
```

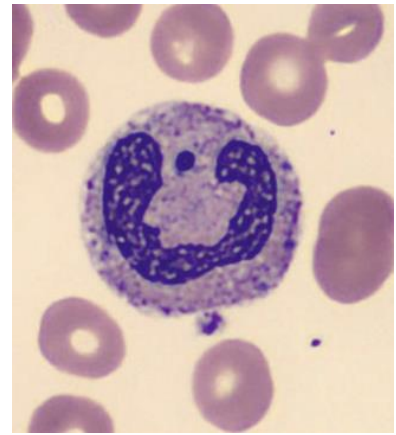
```
  ggplot(aes(Group, Values, group=Experiment, colour=Experiment, fill=Experiment))+  
    geom_line(size=2)+  
    geom_point(size=4, shape = 21, colour= "black", stroke=2)+  
    scale_x_discrete(limits = c("WT", "KO", "KO+T1", "KO+T2"))
```



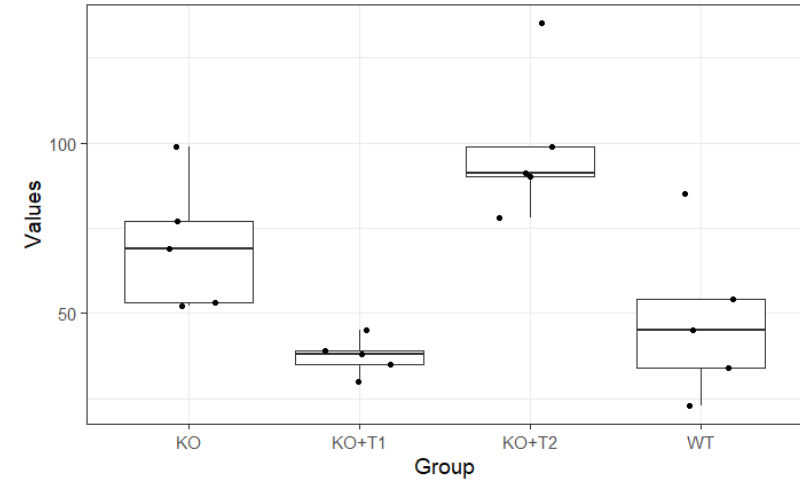
Exercise 7: Repeated measures ANOVA

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)
```

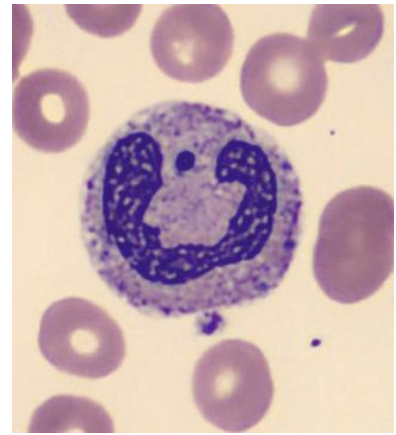


```
neutrophils.long %>%  
  group_by(Group) %>%  
    shapiro_test(Values) %>%  
    ungroup()
```

Group <chr>	variable <chr>	statistic <dbl>	p <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)

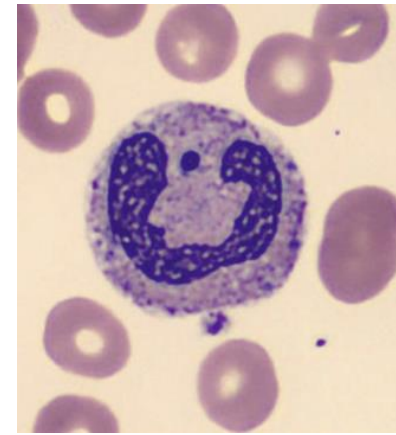
Effect	DFn	DFd	F	p	p<.05	ges
1 Group	3	12	28.575	9.51e-06	*	0.656

```
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. <chr>	group1 <chr>	group2 <chr>	n1 <int>	n2 <int>	statistic <dbl>	df <dbl>	p <dbl>	p.adj <dbl>	p.adj.signif <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	**

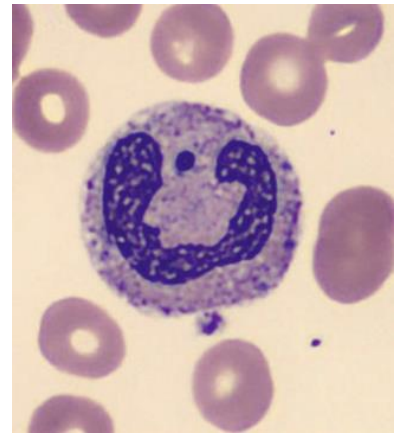
```
neutrophils.long %>%  
  pairwise_t_test(Values~Group, paired=TRUE, ref.group = "WT",  
  p.adjust.method = "holm")
```

	.y. <chr>	group1 <chr>	group2 <chr>	n1 <int>	n2 <int>	statistic <dbl>	df <dbl>	p <dbl>	p.adj <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 <chr>	Group <chr>	Values <dbl>	Difference <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

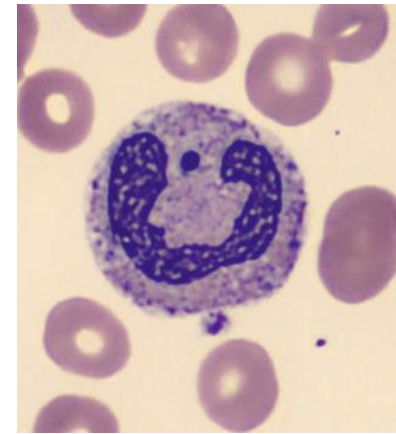
1-10 of 20 rows

Previous

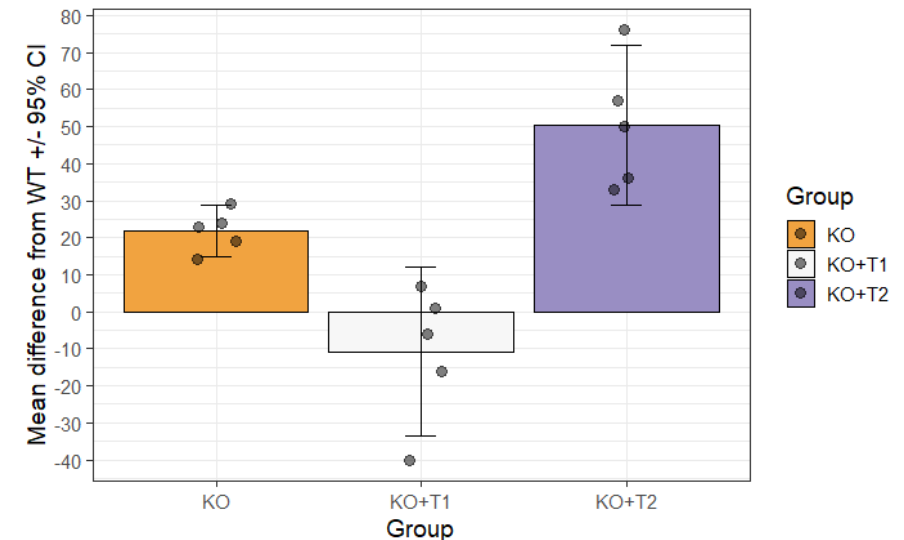
Exercise 7: Repeated measures ANOVA

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

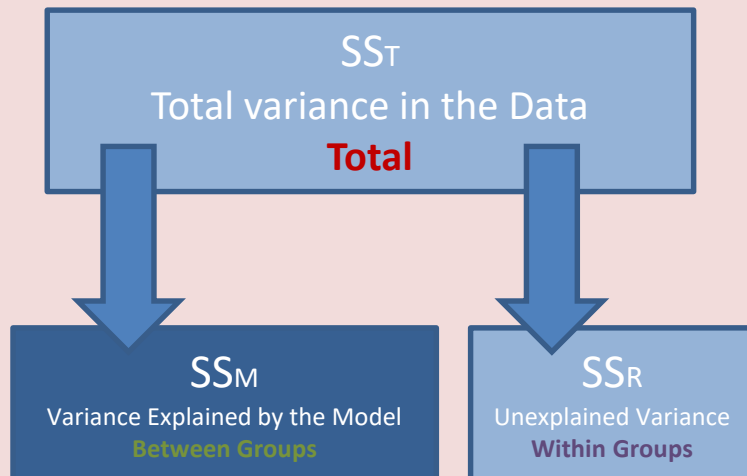
Two-Way ANOVA

Two-way Analysis of Variance (Factorial ANOVA)

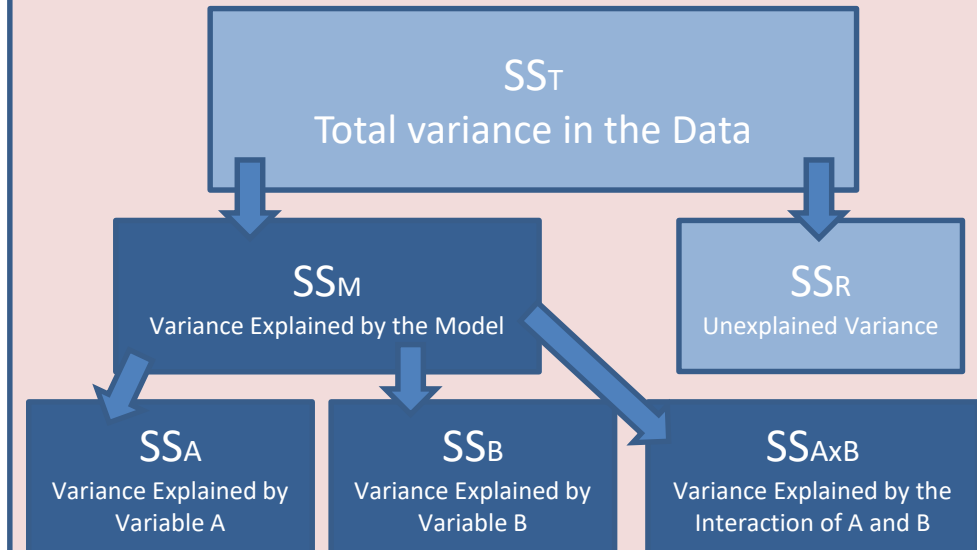
Source of variation	Sum of Squares	Df	Mean Square	F	p-value
Variable A (Between Groups)	2.665	4	0.6663	8.42	<0.0001
Within Groups (Residual)	5.775	73	0.0791		
Total	8.44	77			

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



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



Two-way Analysis of Variance

- Interaction plots: Examples

- Fake dataset:

- 2 factors: **Genotype** (2 levels) and **Condition** (2 levels)

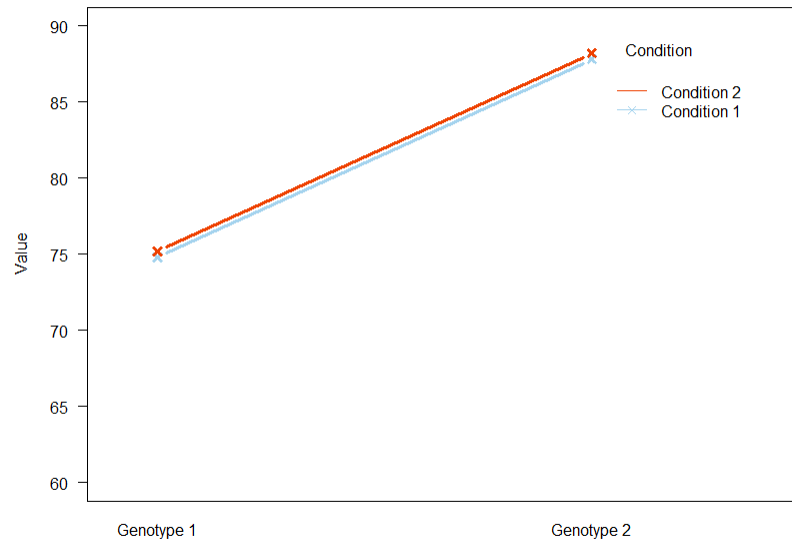
Genotype	Condition	Value
Genotype 1	Condition 1	74.8
Genotype 1	Condition 1	65
Genotype 1	Condition 1	74.8
Genotype 1	Condition 2	75.2
Genotype 1	Condition 2	75
Genotype 1	Condition 2	75.2
Genotype 2	Condition 1	87.8
Genotype 2	Condition 1	65
Genotype 2	Condition 1	74.8
Genotype 2	Condition 2	88.2
Genotype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

Two-way Analysis of Variance

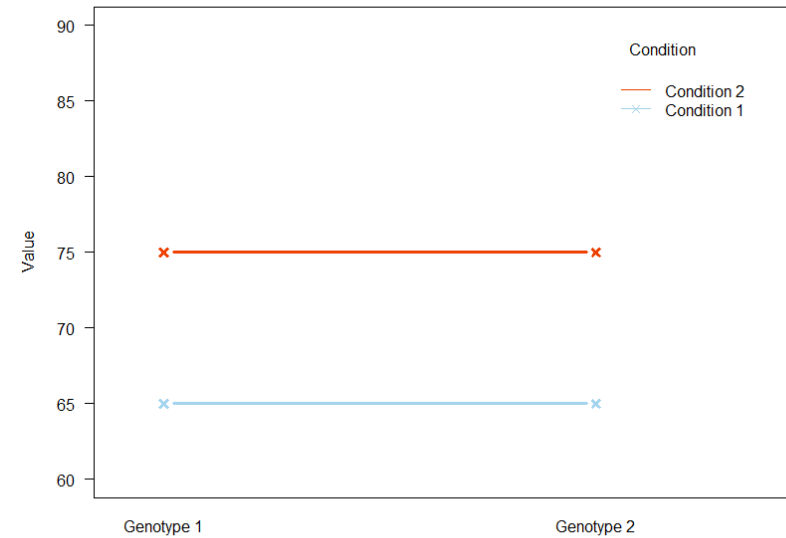
- Interaction plots: Examples

- 2 factors: **Genotype** (2 levels) and **Condition** (2 levels)

Single Effect



Genotype Effect



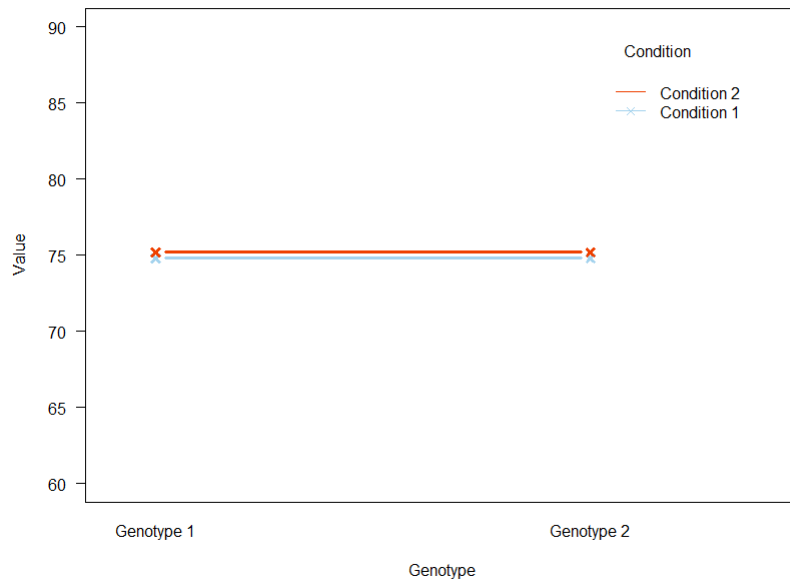
Condition Effect

Two-way Analysis of Variance

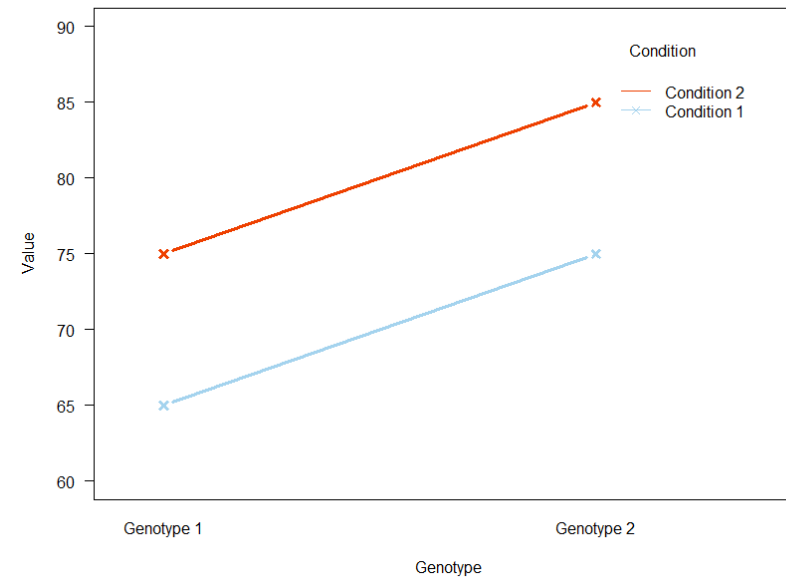
- Interaction plots: Examples

- 2 factors: **Genotype** (2 levels) and **Condition** (2 levels)

Zero or Both Effect



Zero Effect

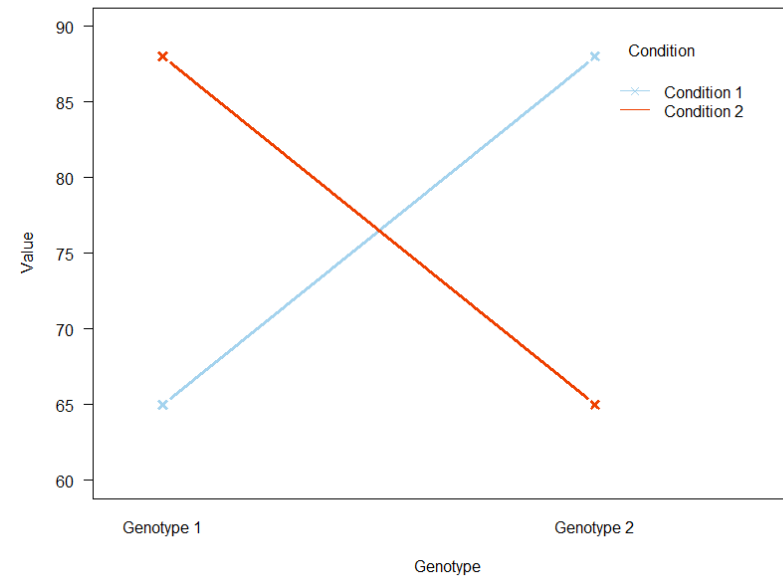
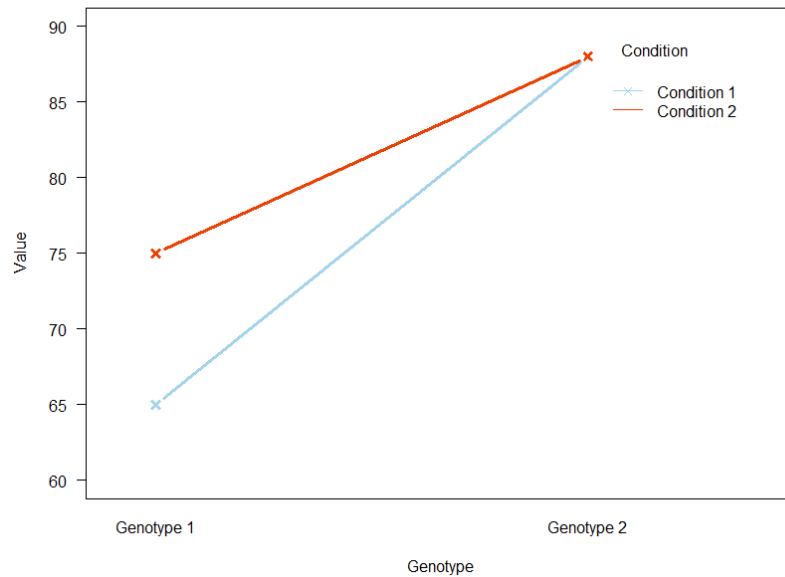


Both Effect

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

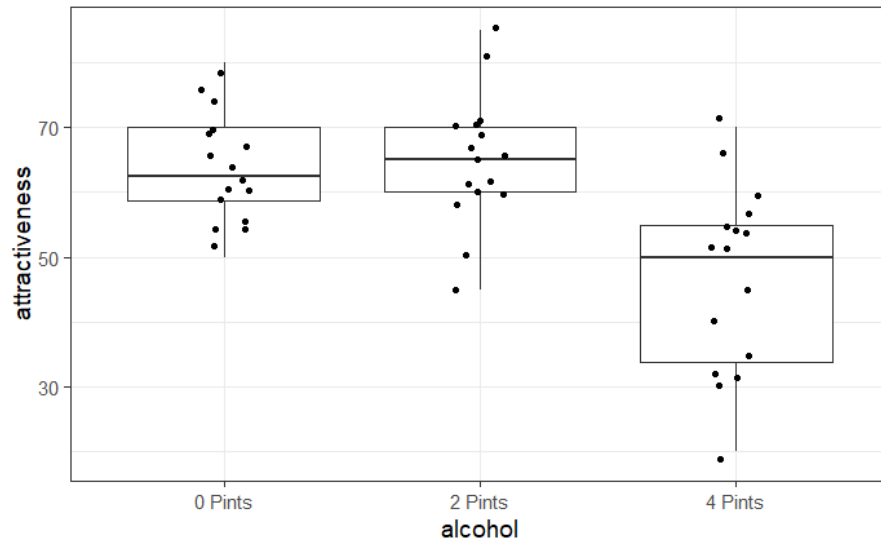
`goggles.csv`

- 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)`

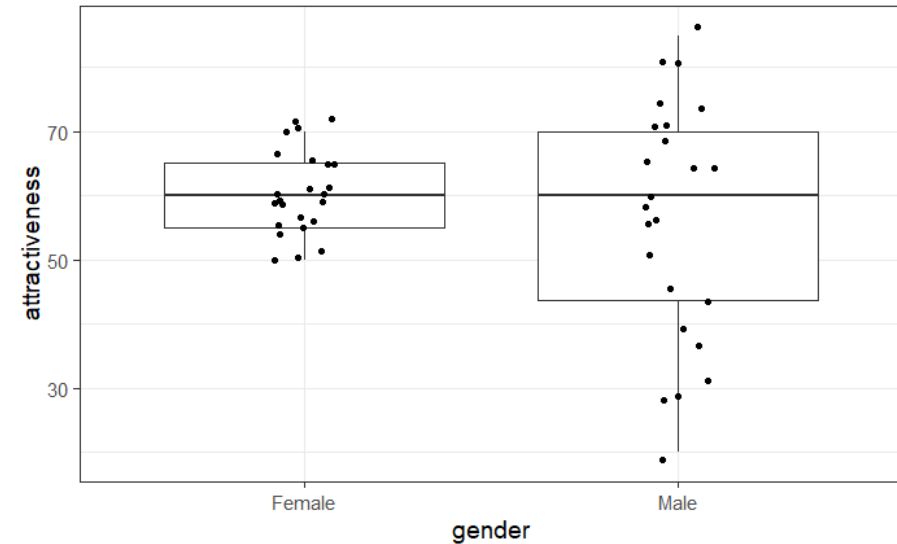
Two-way Analysis of Variance

- 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)
```



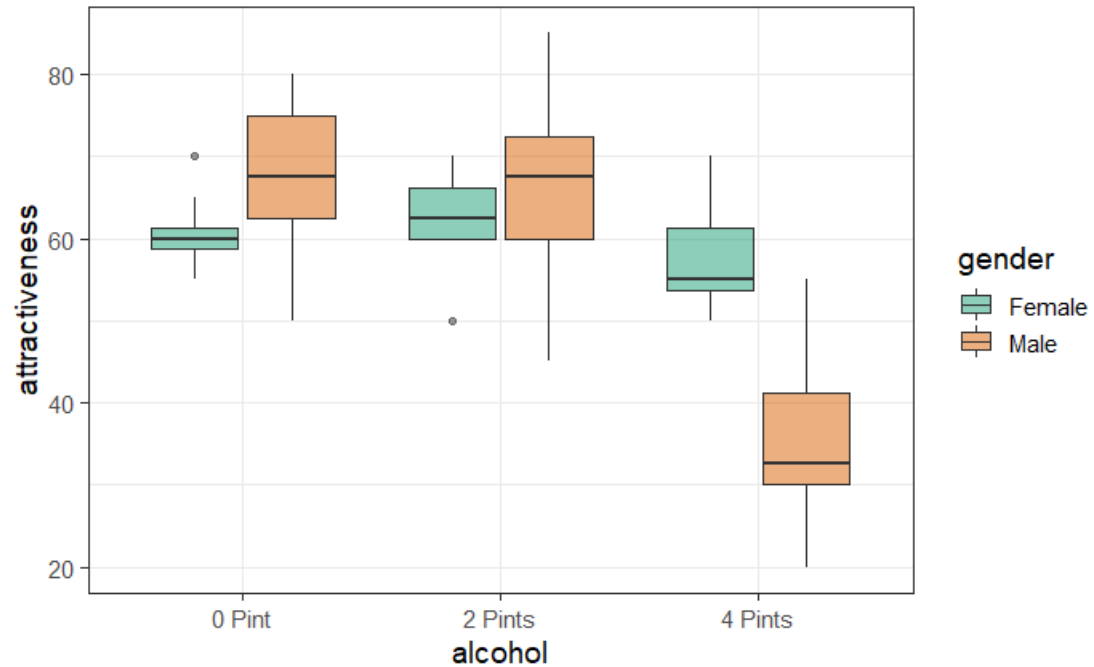
```
goggles %>%  
  ggplot(aes(x=gender, y=attractiveness))+  
  geom_boxplot()+  
  geom_jitter(height=0, width=0.1)
```



Two-way Analysis of Variance

```
goggles %>%
```

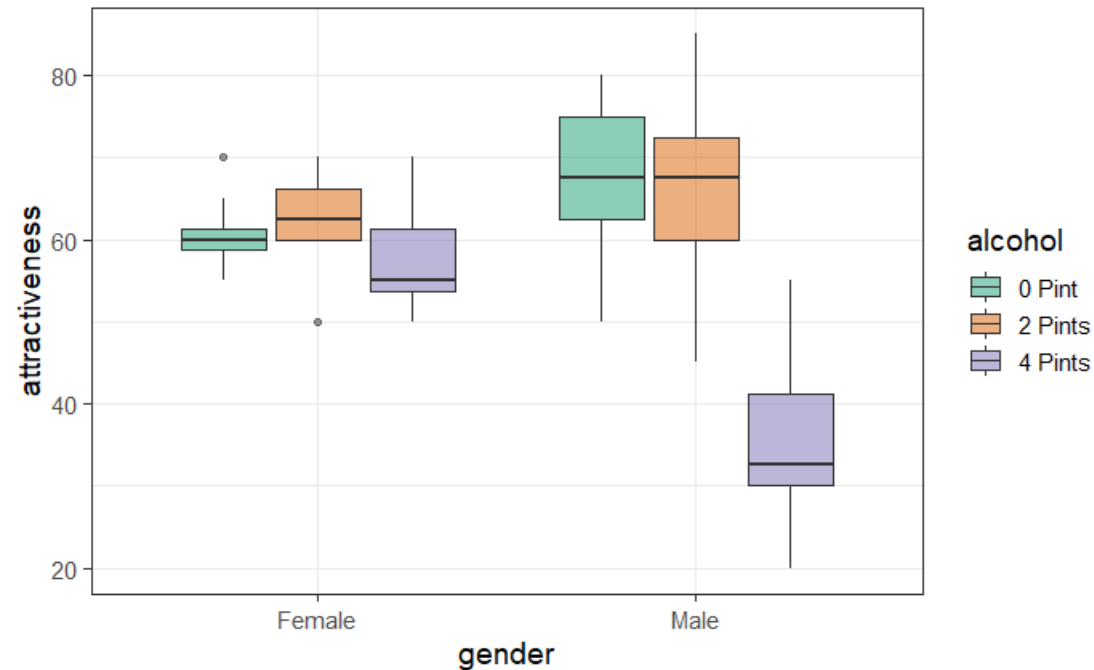
```
  ggplot(aes(beer, attractiveness, fill=gender)) +  
    geom_boxplot(alpha=0.5) +  
    scale_fill_brewer(palette="Dark2")
```



Two-way Analysis of Variance

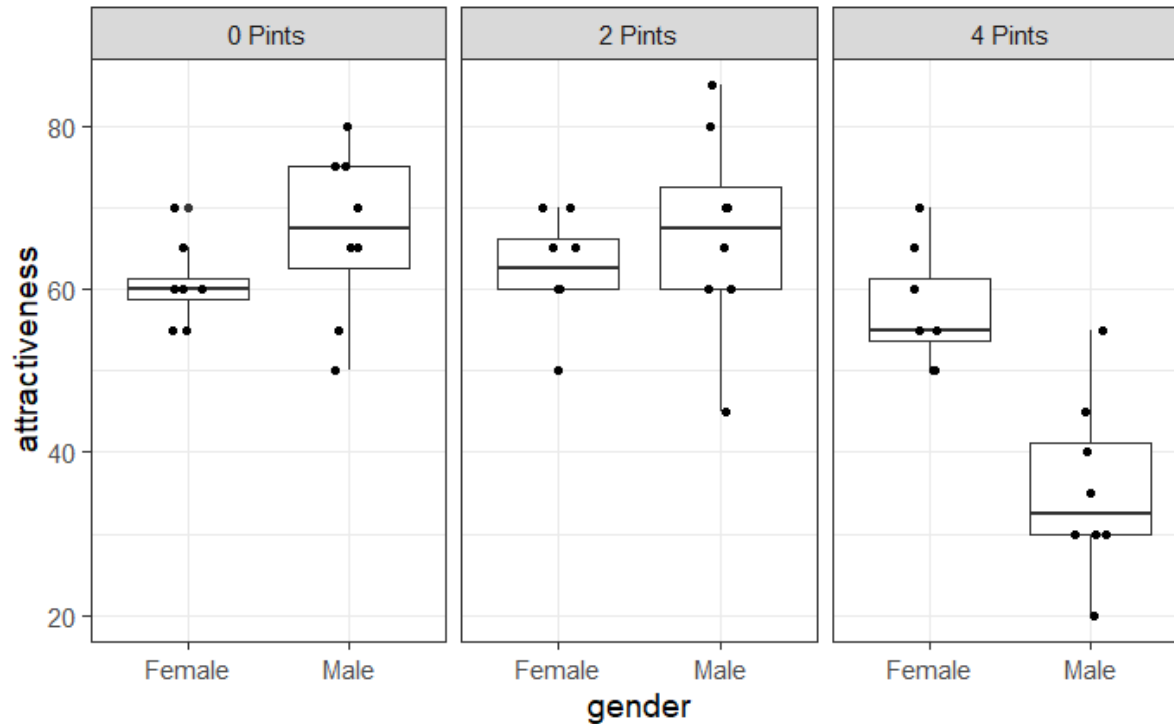
```
goggles %>%
```

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



Two-way Analysis of Variance

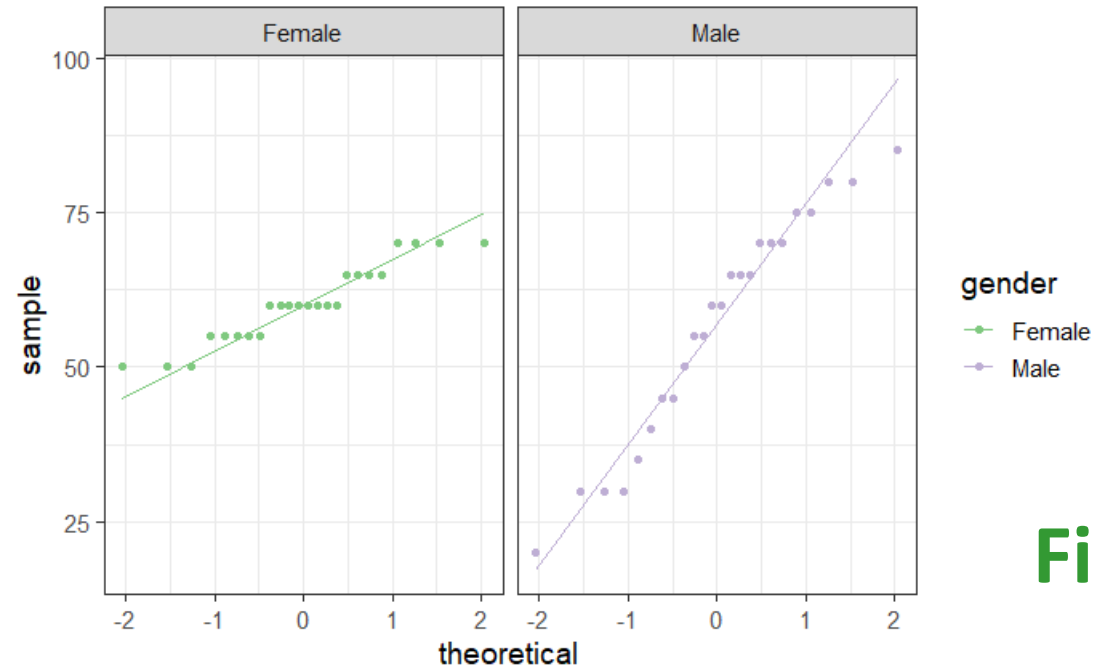
```
goggles %>%  
  ggplot(aes(x=gender, y=attractiveness))+  
  geom_boxplot()+  
  geom_jitter(height=0, width=0.1)+  
  facet_grid(cols=vars(alcohol))
```



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 <chr>	alcohol <chr>	variable <chr>	statistic <dbl>	p <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)
```

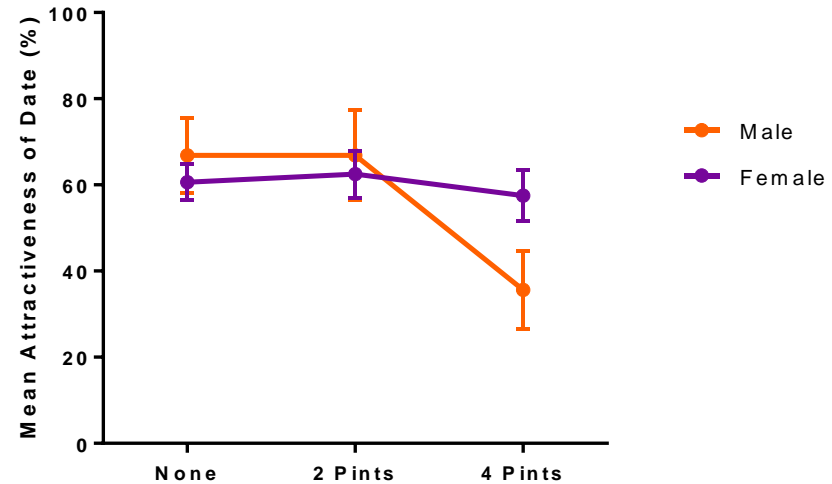
df1 <int>	df2 <int>	statistic <dbl>	p <dbl>
5	42	1.425225	0.2350678

Second assumption ✓

Two-way Analysis of Variance

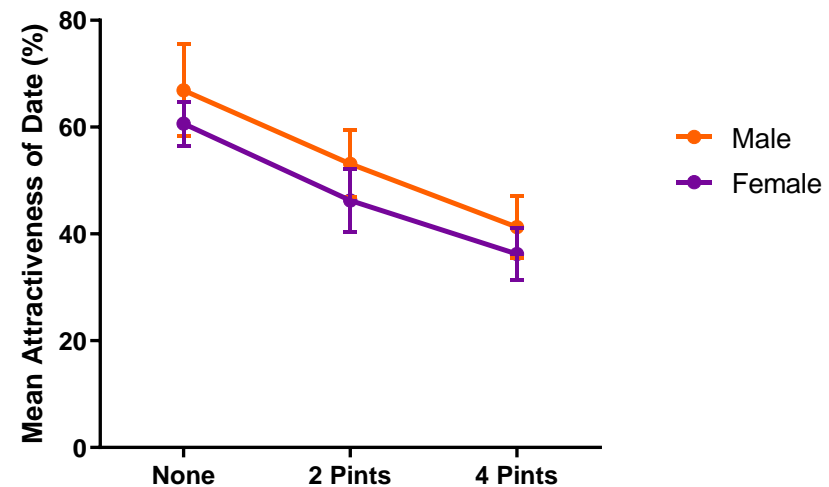
With significant interaction (real data)

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

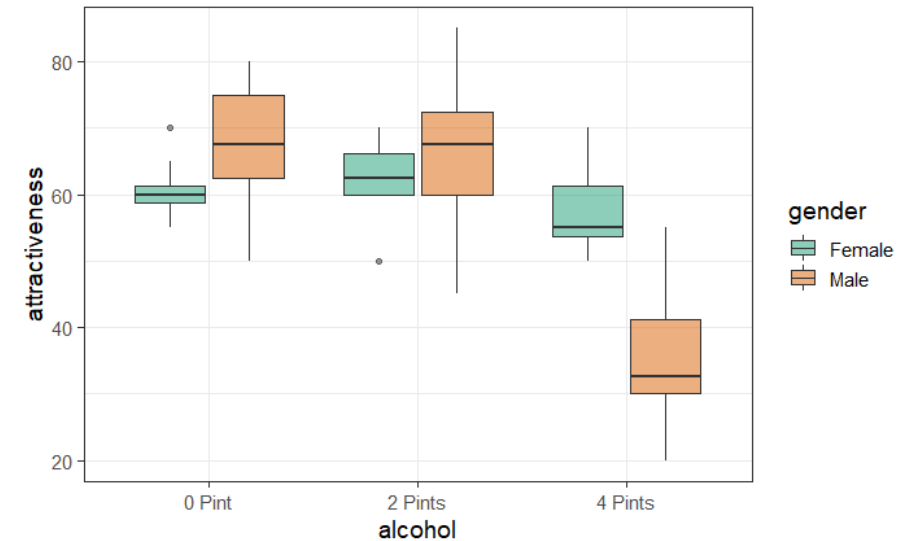
```
goggles %>%
```

```
  group_by(alcohol) %>%
```

```
  tukey_hsd(attractiveness ~ gender) %>%
```

```
  ungroup()
```

alcohol	term	group1	group2	estimate	conf.low	conf.high	p.adj	p.adj.signif
<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<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 ($p=0.0003$)

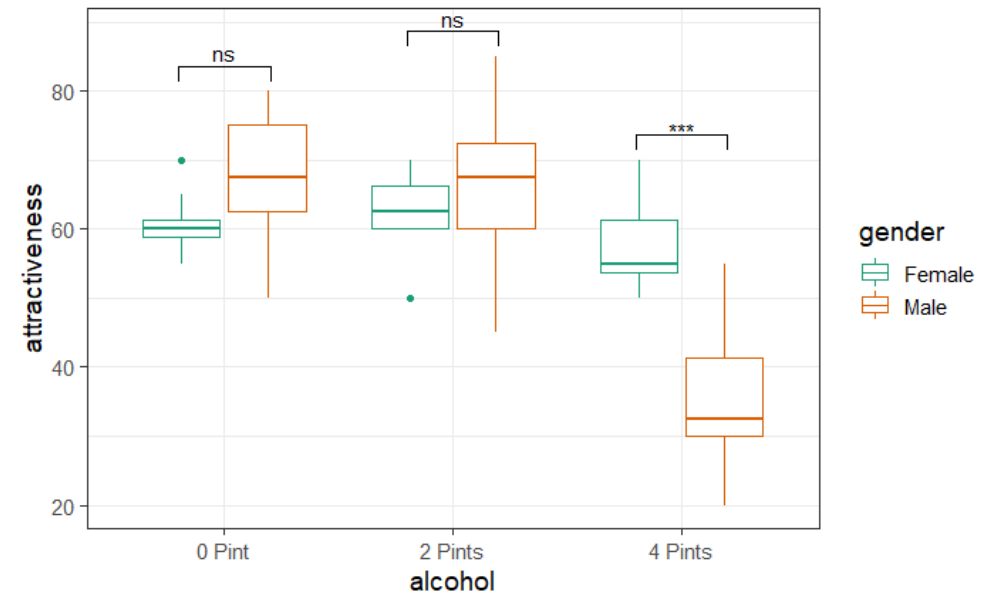
Two-way Analysis of Variance

- *Work in progress* # ggpubr package #

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

beer	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(beer, attractiveness, colour = gender))+  
  geom_boxplot()+  
  stat_pvalue_manual(tukey.results)+  
  scale_colour_brewer(palette = "Dark2")
```



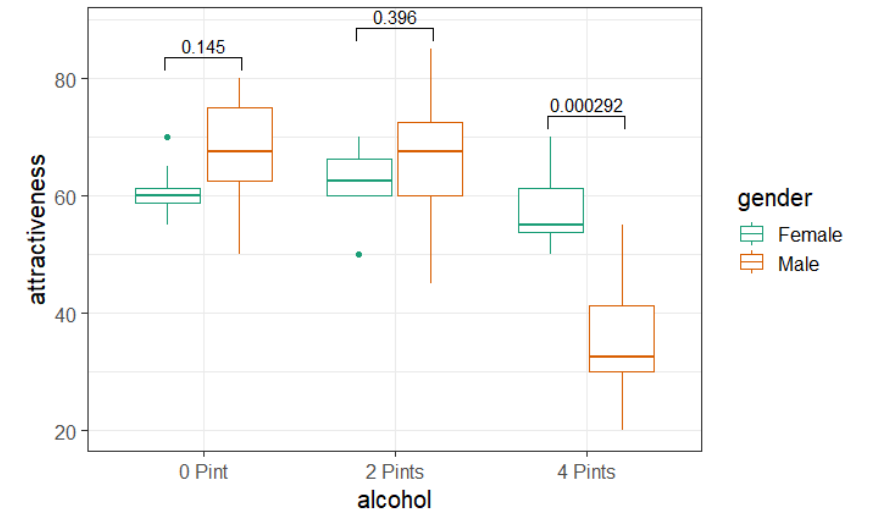
Two-way Analysis of Variance

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

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

beer	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(beer, 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 <chr>	alcohol <chr>	mean <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()
```

