Choosing the right test

Choosing the right test

Which analysis should I use?

1. A clearly defined research question
2. What is the dependent variable and what type of variable is it?
3. How many independent variables are there and what data types are they?
4. Are you interested in comparing means or investigating relationships?
5. Do you have repeated measurements of the same variable for each subject?

Research question

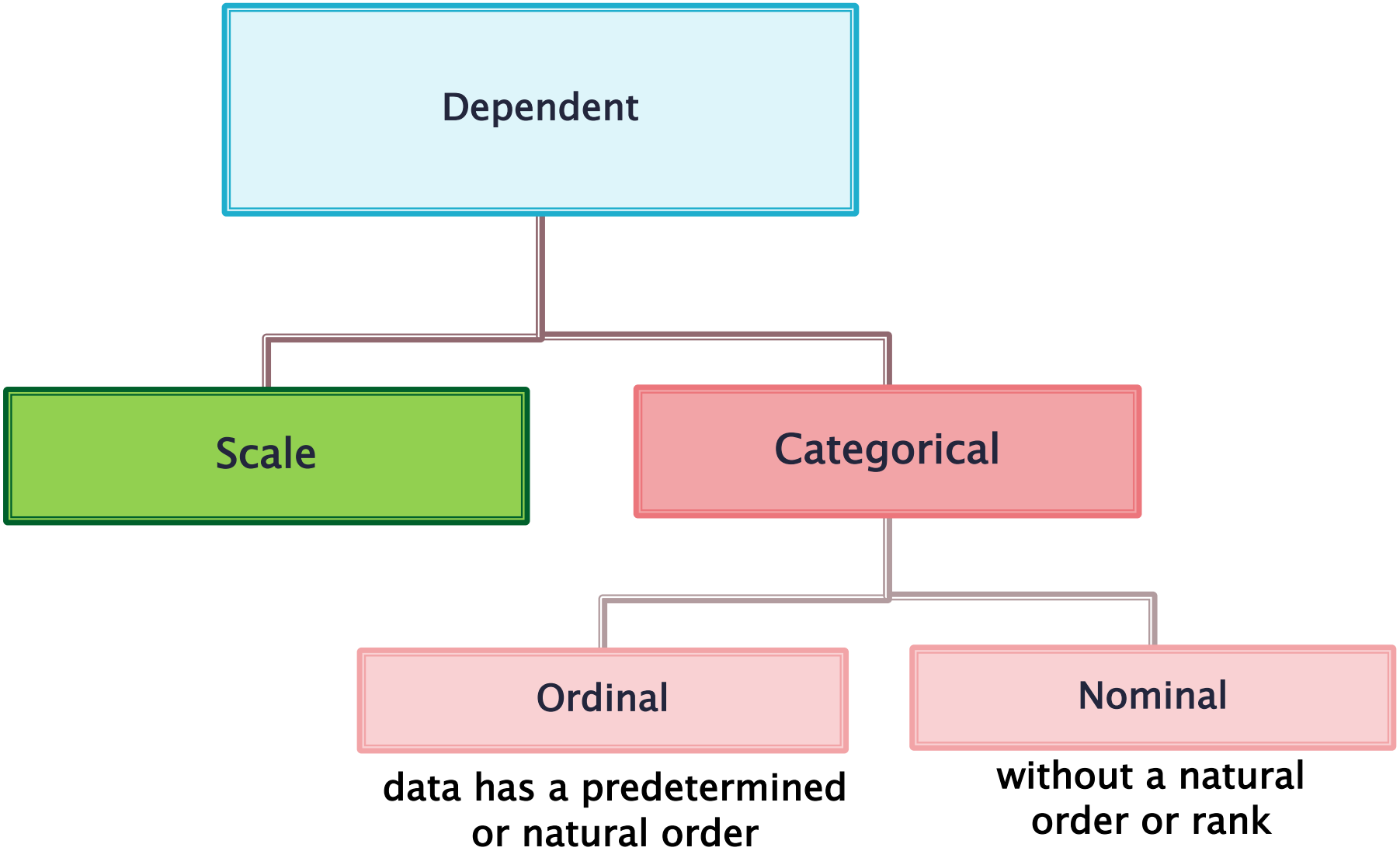
- Clear questions with measurable quantities
- Which variables will help answer these questions
- Think about what test is needed before carrying out a study so that the right type of variables are collected

Dependent variables

Does **attendance** have an association with **exam score**?

Do **women** do more **housework** than **men**?

What variable type is the dependent?



Are boys better at math?

- How can 'better' be measured and what type of variable is it?

Exam score (Scale)

- Do boys think they are better at maths??

I consider myself to be good at maths (ordinal)

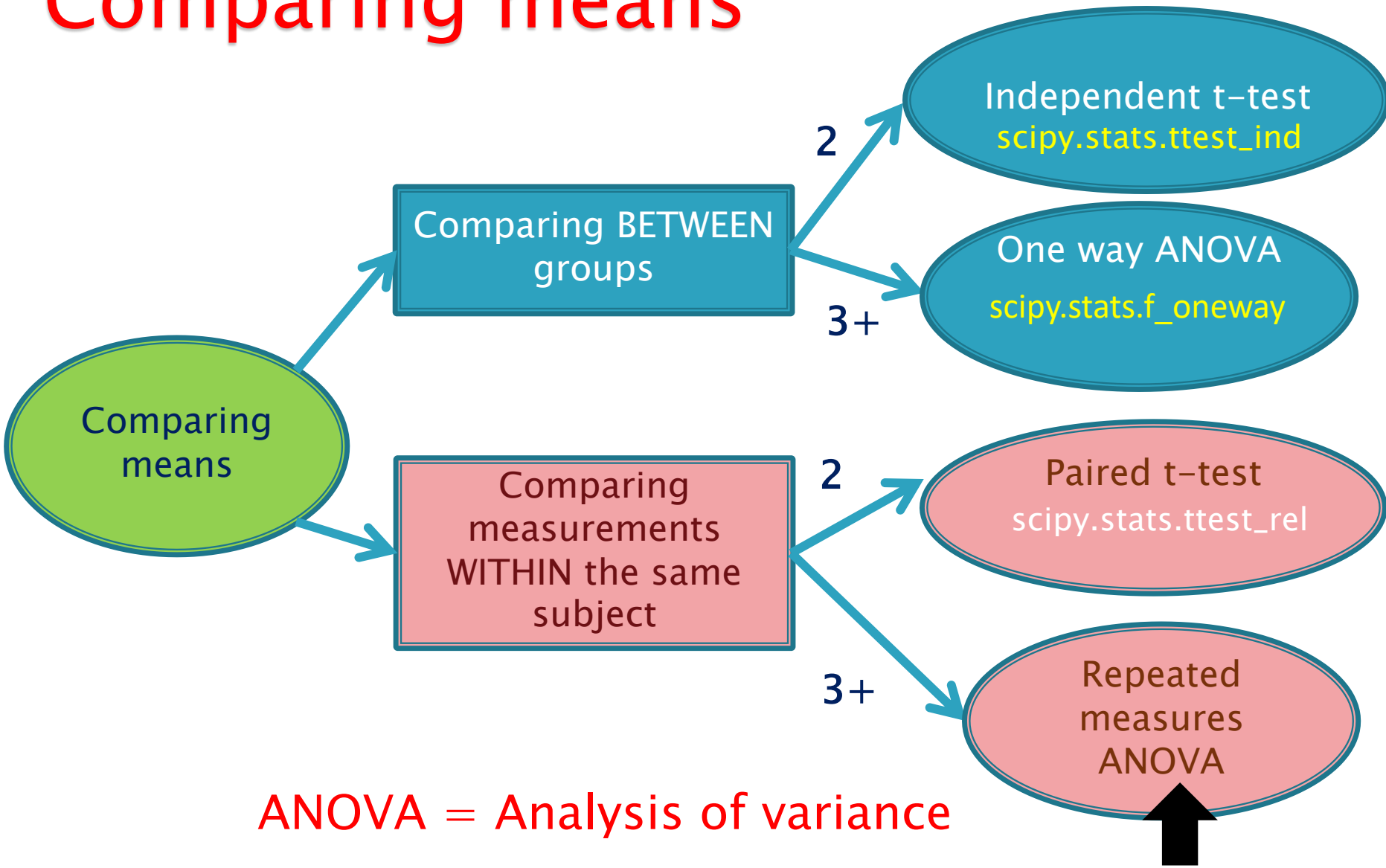
Comparing means

- Dependent = Scale
- Independent = Categorical

- How many means are you comparing?

- Do you have independent groups or repeated measurements on each person?

Comparing means



ANOVA = Analysis of variance

```
from statsmodels.stats.anova import AnovaRM
```

Exercise – Comparing means

Research question	Dependent variable	Independent variable	Test
Do women do more housework than men?	Housework (hrs per week) (Scale)	Gender (Nominal)	
Does Margarine X reduce cholesterol? Everyone has cholesterol measured on 3 occasions	Cholesterol (Scale)	Occasion (Nominal)	
Which of 3 diets is best for losing weight?	Weight lost on diet (Scale)	Diet (Nominal)	

Exercise: Solution

Research question	Dependent variable	Independent variable	Test
Do women do more housework than men?	Housework (hrs per week) (Scale)	Gender (Nominal)	Independent t-test
Does Margarine X reduce cholesterol? Everyone has cholesterol measured on 3 occasions	Cholesterol (Scale)	Occasion (Nominal)	Repeated measures ANOVA
Which of 3 diets is best for losing weight?	Weight lost on diet (Scale)	Diet (Nominal)	One-way ANOVA

Tests investigating relationships

Investigating relationships between	Dependent variable	Independent variable	Test
2 categorical variables	Categorical	Categorical	Chi-squared test
2 Scale variables	Scale	Scale	Pearson's correlation
Predicting the value of an dependent variable from the value of a independent variable	Scale	Scale/binary	Simple Linear Regression
	Binary	Scale/ binary	Logistic regression

Note: Multiple linear regression is when there are several independent variables

Exercise: Relationships

Research question	Dependent variable	Independent variables	Test
Does attendance affect exam score?	Exam score (Scale)	Attendance (Scale)	
Do women do more housework than men?	Housework (hrs per week) (scale)	Gender (Binary) Hours worked (Scale)	
Were Americans more likely to survive on board the Titanic?	Survival (Binary)	Nationality (Nominal)	
	Survival (Binary)	Nationality , Gender, class	

Note: There may be 2 appropriate tests for some questions

Exercise: Solution

Research question	Dependent variable	Independent variables	Test
Does attendance affect exam score?	Exam score (Scale)	Attendance (Scale)	Correlation/ regression
Do women do more housework than men?	Housework (hrs per week) (scale)	Gender (Binary) Hours worked (Scale)	Regression
Were Americans more likely to survive on board the Titanic?	Survival (Binary)	Nationality (Nominal)	Chi-squared
	Survival (Binary)	Nationality , Gender, class	Logistic regression

Note: There may be 2 appropriate tests for some questions

Non-parametric tests

Parametric or non-parametric?

Statistical tests fall into two types:

Parametric tests

Assume data follows a particular distribution e.g. normal

Non-parametric

Nonparametric techniques are usually based on ranks/ signs rather than actual data

Ranking raw data

- Nonparametric techniques are usually based on ranks or signs
- Scale data is ordered and ranked
- Analysis is carried out on the ranks rather than the actual data

Original variable

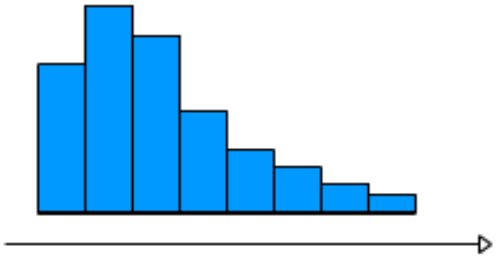
Rank of subject

Group	ReactionTime	Rank
Placebo	.37	1
Placebo	.38	2
Placebo	.61	3
Placebo	.78	4
Placebo	.83	5
Placebo	.86	6
Placebo	.90	7
Placebo	.95	8
Alcohol	.98	9
Alcohol	1.11	10
Alcohol	1.27	11
Alcohol	1.32	12
Alcohol	1.44	13
Alcohol	1.45	14
Alcohol	1.46	15
Placebo	1.63	16
Alcohol	1.76	17
Placebo	1.97	18
Alcohol	2.56	19
Alcohol	3.07	20

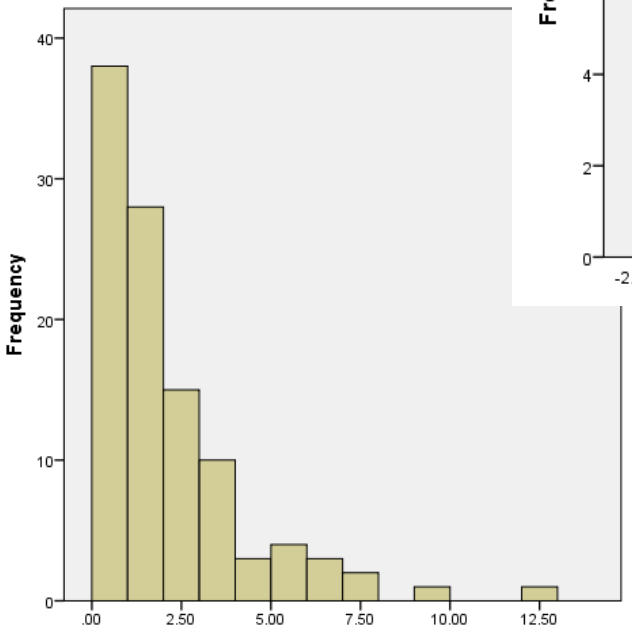
Non-parametric tests

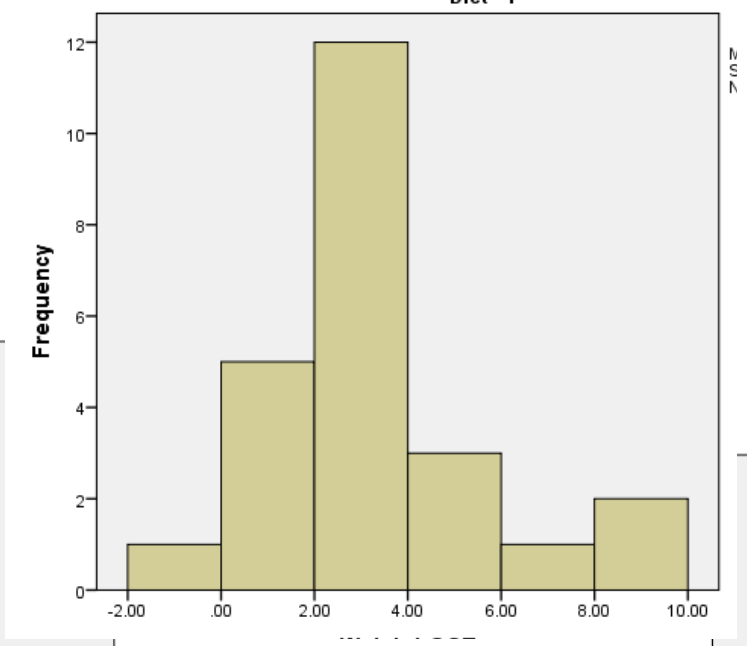
- Non-parametric methods are used when:
  - Data is ordinal
  - Data does not seem to follow any particular shape or distribution (e.g. Normal)
  - Assumptions underlying parametric test not met
  - A plot of the data appears to be very skewed
  - There are potential influential outliers in the dataset
  - Sample size is small

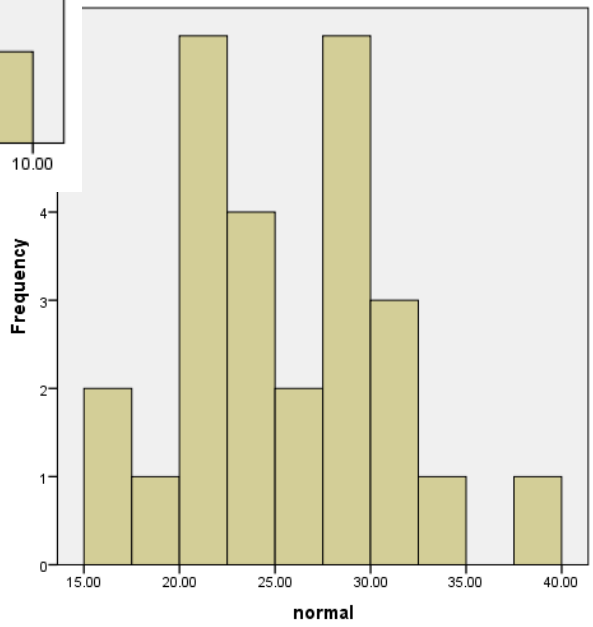
Note: Parametric tests are fairly robust to non-normality. Data has to be very skewed to be a problem



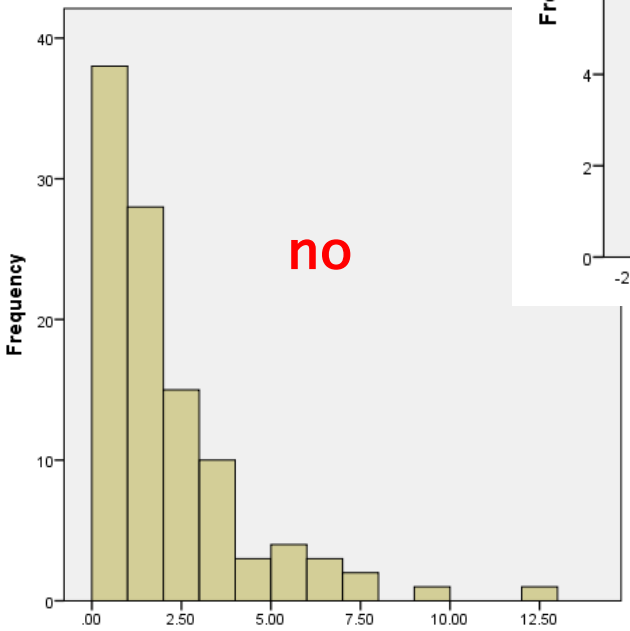
Do these look normally distributed?



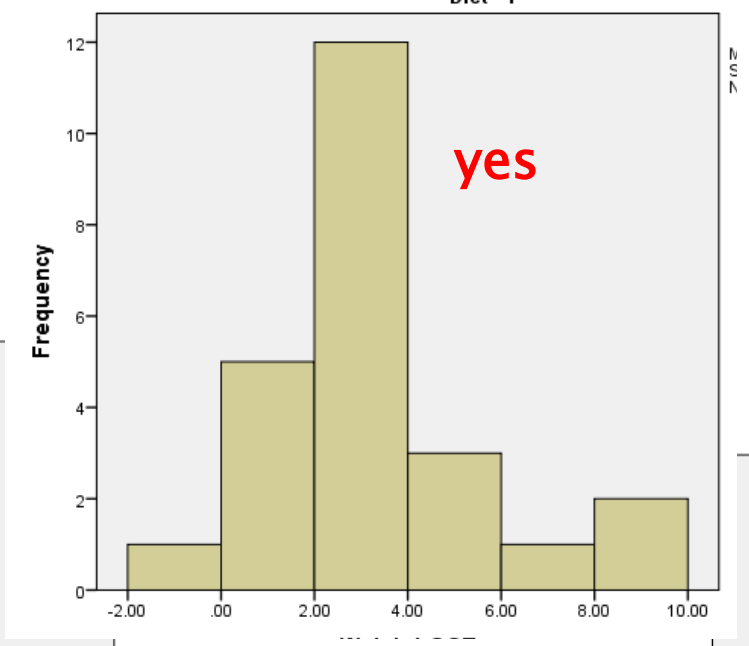




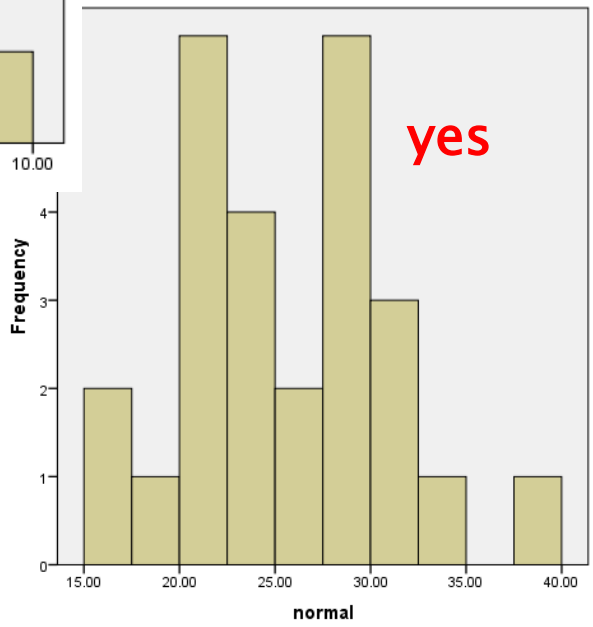
Do these look normally distributed?



no



yes



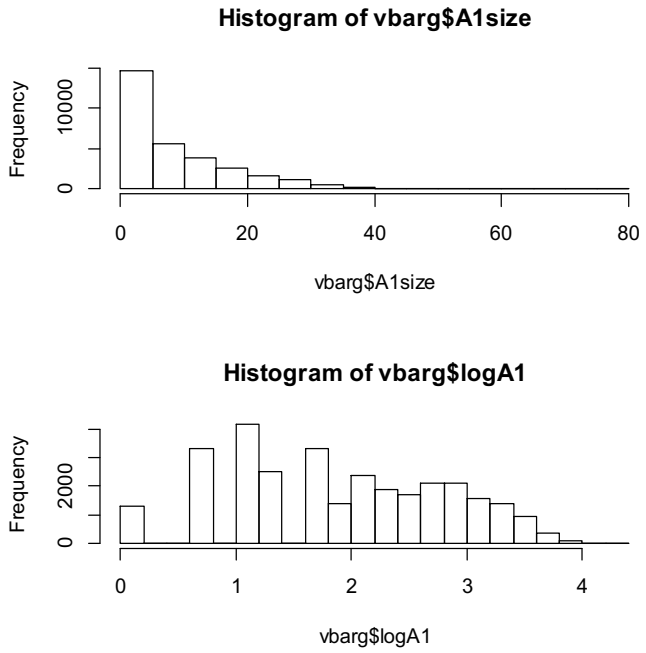
yes

What can be done about non-normality?

If the data are not normally distributed, there are two options:

1. Use a non-parametric test
2. Transform the dependent variable

For positively skewed data, taking the log of the dependent variable often produces normally distributed values



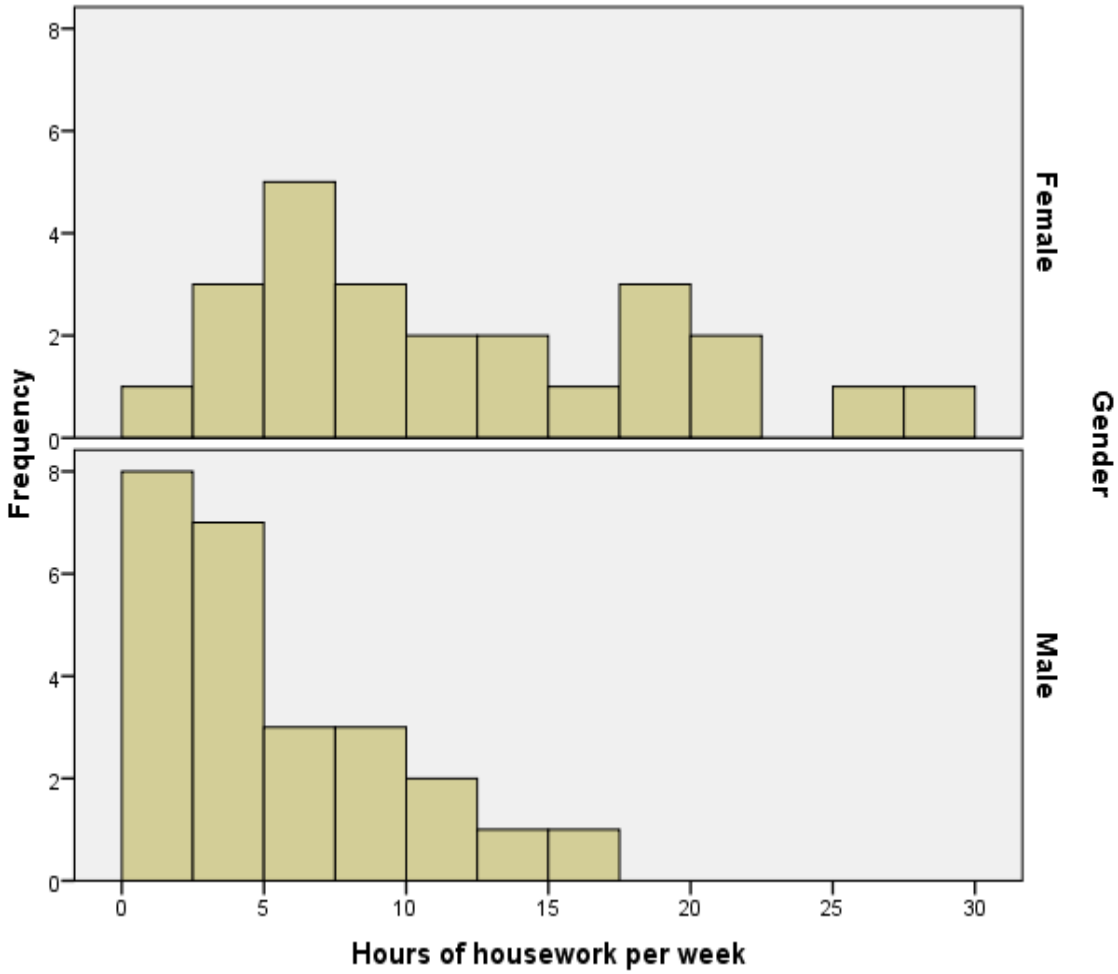
Non-parametric tests

Parametric test	What to check for normality	Non-parametric test
Independent t-test	Dependent variable by group	Mann-Whitney test
Paired t-test	Paired differences	Wilcoxon signed rank test
One-way ANOVA	Residuals/Dependent	Kruskal-Wallis test
Repeated measures ANOVA	Residuals	Friedman test
Pearson's Correlation Co-efficient	At least one of the variables should be normal	Spearman's Correlation Co-efficient
Linear Regression	Residuals	None – transform the data

Notes: The residuals are the differences between the observed and expected values.

Exercise: Comparison of housework

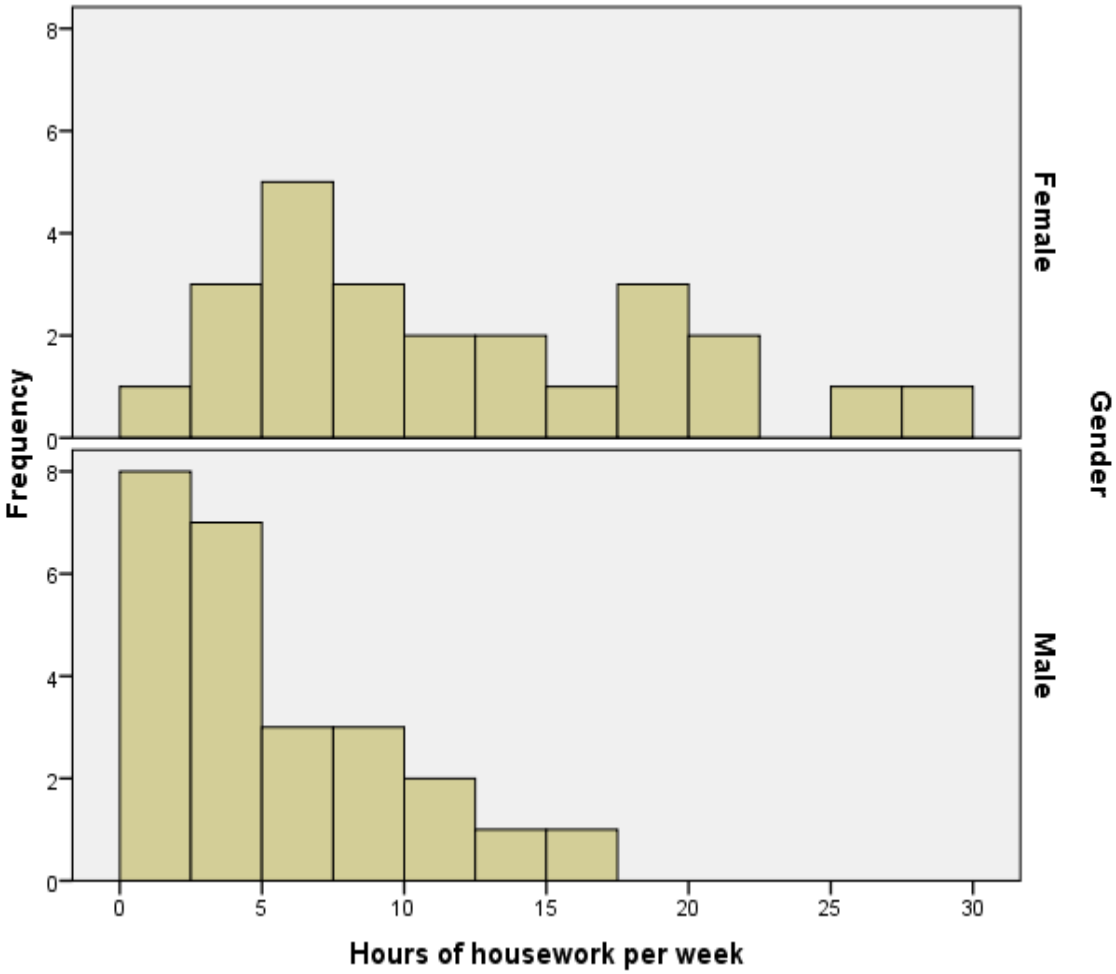
Which test should be carried out to compare the hours of housework for males and females? Look at the histograms of housework by gender to decide.



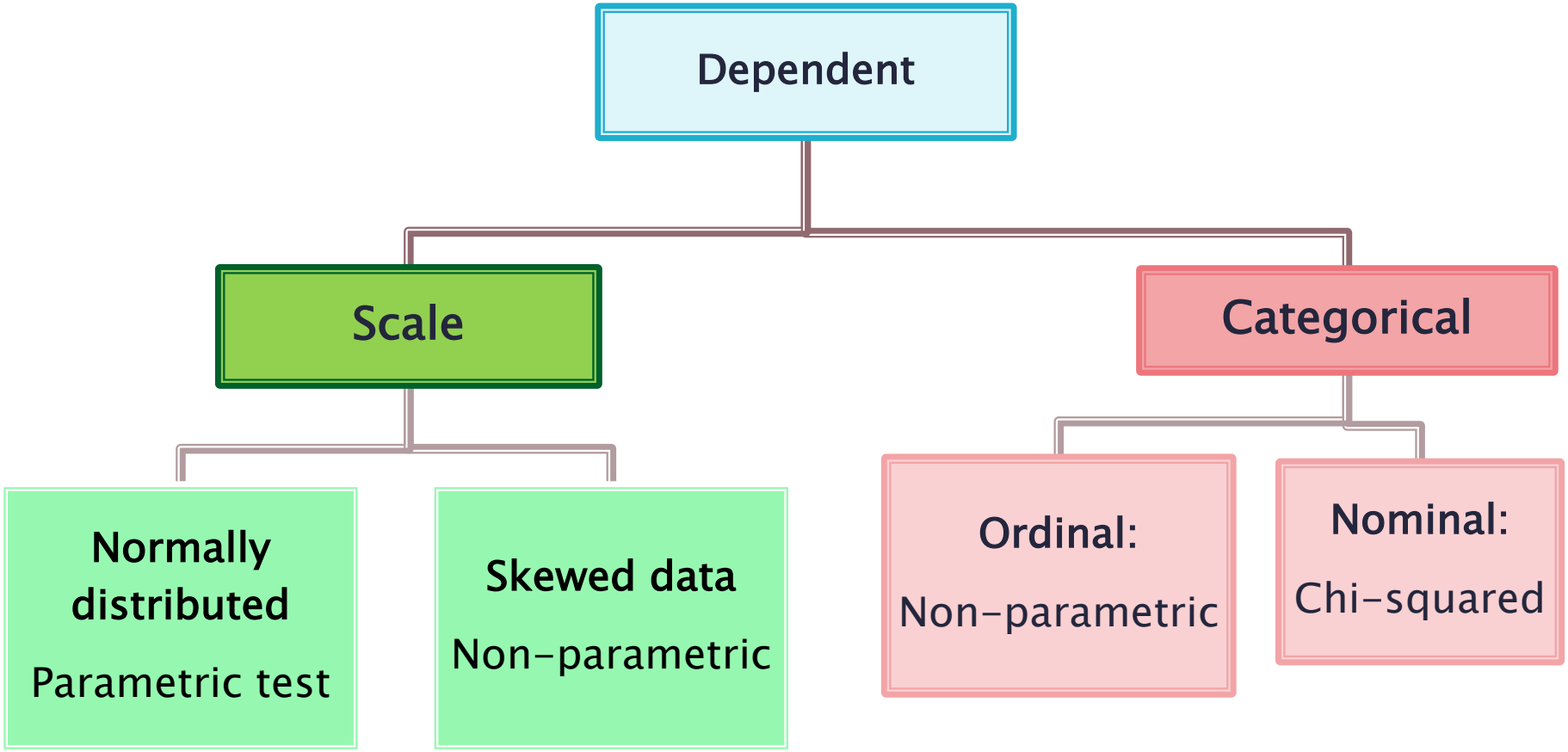
Exercise: Solution

Which test should be carried out to compare the hours of housework for males and females?

The male data is very skewed so use the Mann-Whitney.



Summary


Dependent
Scale
Categorical
Normally distributed
Parametric test
Skewed data
Non-parametric
Ordinal:
Non-parametric
Nominal:
Chi-squared

Mann Whitney Test

Legal drink drive limits

Research question: Does drinking the legal limit for alcohol affect driving reaction times?

- Participants were given either alcoholic or non-alcoholic drinks and their driving reaction times tested on a simulated driving situations
- They did not know whether they had the alcohol or not

Placebo: 0.90 0.37 1.63 0.83 0.95 0.78 0.86 0.61 0.38 1.97
Alcohol: 1.46 1.45 1.76 1.44 1.11 3.07 0.98 1.27 2.56 1.32

Mann–Whitney test

- Nonparametric equivalent to independent t-test
- The data from both groups is ordered and ranked
- The mean rank for the groups is compared

If most of the lowest ranks are in the same group, this suggests a difference between the two groups.

Group	ReactionTime	Rank
Placebo	.37	1
Placebo	.38	2
Placebo	.61	3
Placebo	.78	4
Placebo	.83	5
Placebo	.86	6
Placebo	.90	7
Placebo	.95	8
Alcohol	.98	9
Alcohol	1.11	10
Alcohol	1.27	11
Alcohol	1.32	12
Alcohol	1.44	13
Alcohol	1.45	14
Alcohol	1.46	15
Placebo	1.63	16
Alcohol	1.76	17
Placebo	1.97	18
Alcohol	2.56	19
Alcohol	3.07	20

Formulation:
https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test

Drink driving reactions

H_0: There is no difference between the alcohol and placebo populations on reaction time

H_a: The alcohol population has a different reaction time distribution to the placebo population

Test Statistic = Mann-Whitney U

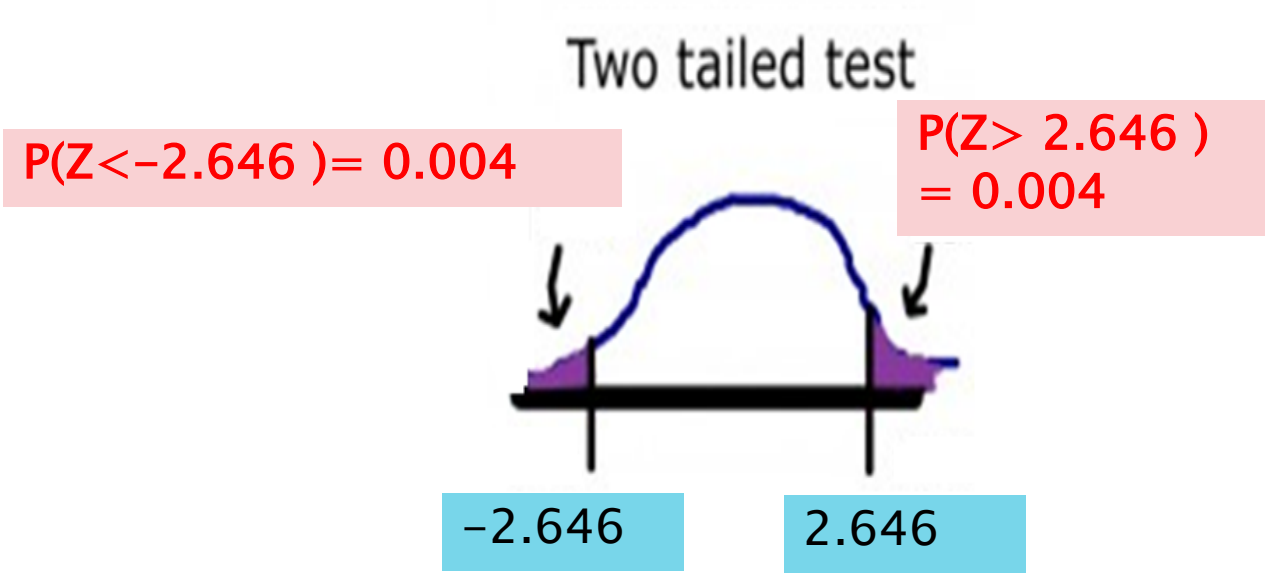
The test statistic U can be approximated to a z score to get a p-value

Mann-Whitney results

- Interpret the results

Test Statistics

	ReactionTime
Mann-Whitney U	15.000
Z	-2.646
p-value (2-tailed)	.008



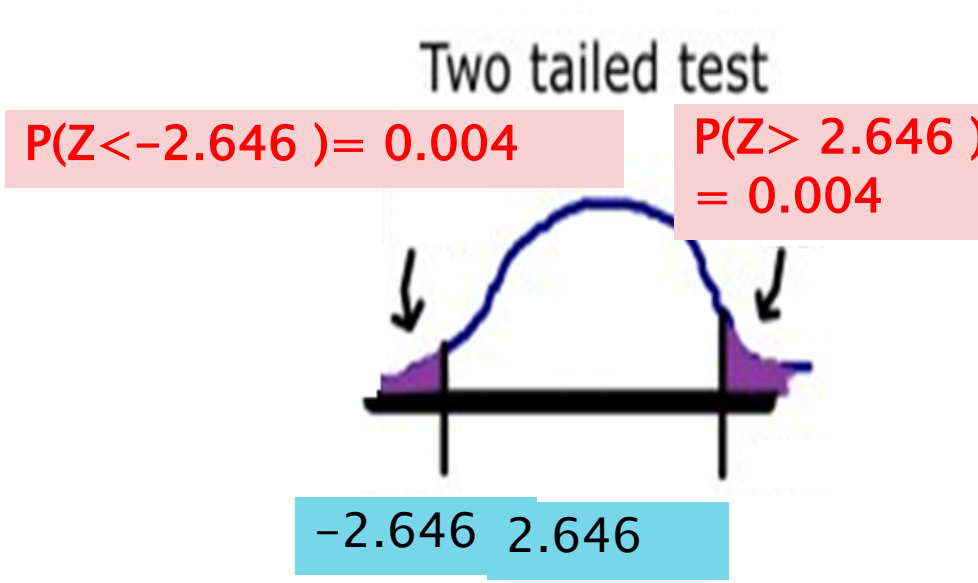
Mann-Whitney results – solution

- Interpret the results

- p =0.008
- Highly significant evidence to suggest a difference in the distributions of reaction times for those in the placebo and alcohol groups

Test Statistics

	ReactionTime
Mann-Whitney U	15.000
Z	-2.646
p-value (2-tailed)	.008



ANOVA

ANOVA

Compares the means of several groups

Which diet is best?
Dependent: Weight lost (Scale)
Independent: Diet 1, 2 or 3 (Nominal)

Null hypothesis: The mean weight lost on diets 1, 2 and 3 is the same $H_0 : \mu_1 = \mu_2 = \mu_3$

Alternative hypothesis: The mean weight lost on diets 1, 2 and 3 are not all the same

Summary statistics

	Overall	Diet 1	Diet 2	Diet 3
Mean	3.85	3.3	3.03	5.15
Standard deviation	2.55	2.24	2.52	2.4
Number in group	78	24	27	27

- Which diet was best?
- Are the standard deviations similar?

How Does ANOVA Work?

- ANOVA = Analysis of variance

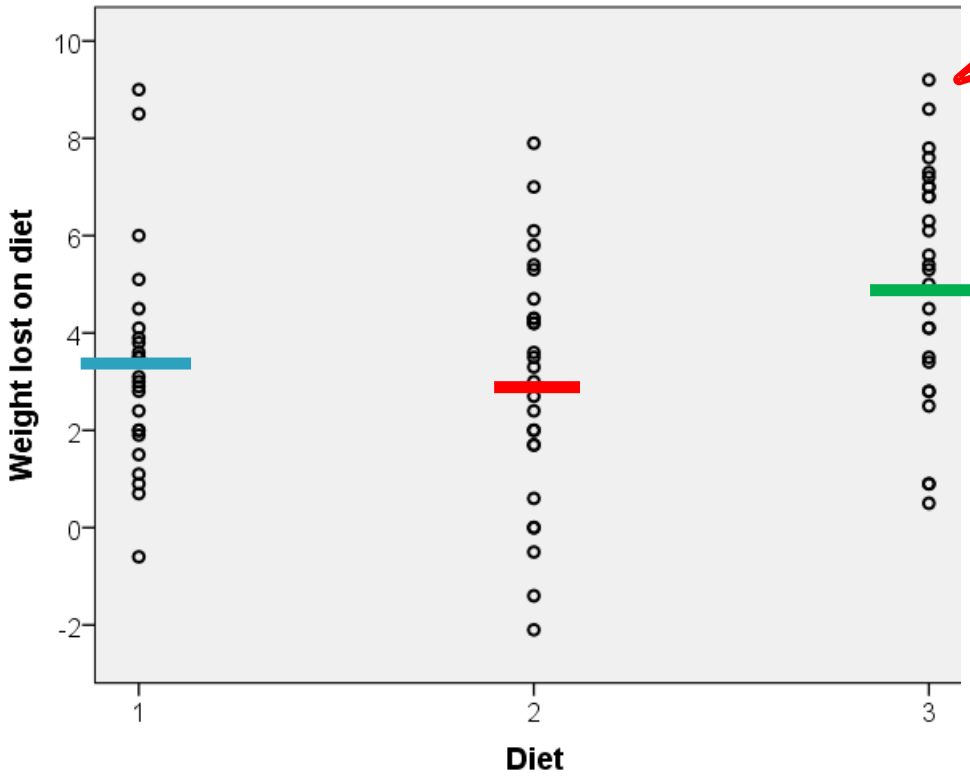
- We compare variation **between** groups relative to variation **within** groups

- Population variance estimated in two ways:
  - One based on variation **between** groups we call the **M**ean **S**quare due to **T**reatments/ **MST**/ $\mathbf{MS_{between}}$
  - Other based on variation **within** groups we call the **M**ean **S**quare due to **E**rror/ **MSE**/ $\mathbf{MS_{within}}$

Within group variation

Residual =difference between an individual and their group mean

SS_{within}=sum of squared residuals

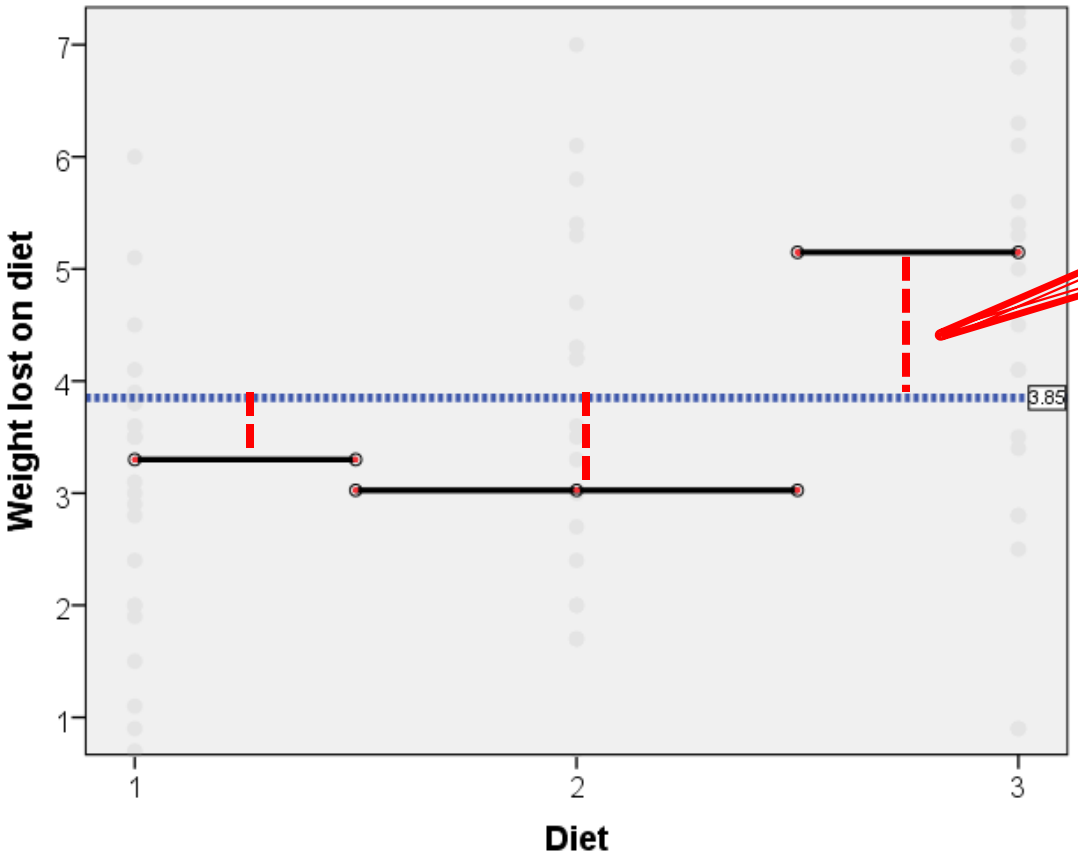


Person lost 9.2kg kg so residual =9.2 – 5.15 = 4.05

Mean weight lost on diet 3 = 5.15kg

Between group variation

Differences between each group mean and the overall mean



Diet 3 difference =5.15 – 3.85 = 1.3

Overall mean = 3.85

Sum of squares calculations

- K = number of groups

$$SS_{within} = \sum_{j=1}^{k}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_j}\right)^2$$

Group means

$$= \sum_{i=1}^{24}\left(x_i - 3.3\right)^2 + \sum_{i=1}^{27}\left(x_i - 3.03\right)^2 + \sum_{i=1}^{27}\left(x_i - 5.15\right)^2 = 430.179$$

$$SS_{Between} = \sum_{j=1}^{k} n_j\left(\overline{x_j} - \overline{x_T}\right)^2$$

$$= 24\left(3.3 - 3.85\right)^2 + 27\left(3.03 - 3.85\right)^2 + 27\left(5.15 - 3.85\right)^2 = 71.094$$

Overall Mean

ANOVA test statistic

Summary ANOVA

Source	Sum of Squares	Degrees of Freedom	Variance Estimate (Mean Square)	*F* Ratio
Between	SS_B	$K-1$	$MS_B = \frac{SS_B}{K-1}$	$\frac{MS_B}{MS_W}$
Within	SS_W	$N-K$	$MS_W = \frac{SS_W}{N-K}$	
Total	$SS_T = SS_B + SS_W$	$N-1$		

Test Statistic (usually reported in papers)

N = total observations in all groups,

K = number of groups

Test Statistic (by hand)

- Filling in the boxes

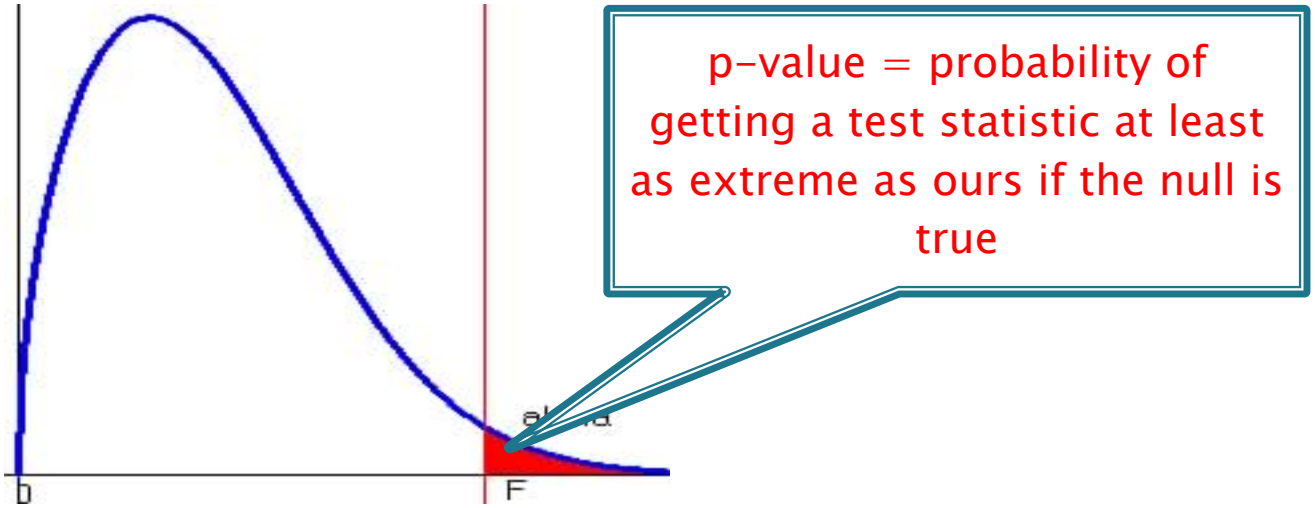
	Sum of squares	Degrees of freedom	Mean square	F-ratio (test statistic)
$SS_{between}$	71.045	2	35.522	**6.193**
SS_{within}	430.180	75	5.736	
SS_{total}	501.275	77		

$$F = \frac{\text{Mean between group sum of squared differences}}{\text{Mean within group sum of squared differences}}$$

If $F > 1$, there is a bigger difference between groups than within groups

P-value

- The p-value for ANOVA is calculated using the F-distribution
- If you repeated the experiment numerous times, you would get a variety of test statistics



One way ANOVA

$$\text{Test Statistic} = \frac{\text{between group variation}}{\text{within group variation}} = \frac{MS_{Diet}}{MS_{Error}} = 6.197$$

Tests of Between-Subjects Effects

Dependent Variable: Weight lost on diet (kg)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	71.094[a]	2	35.547	6.197	.003
Intercept	1137.494	1	1137.494	198.317	.000
Diet	71.094	2	35.547	6.197	.003
Error	430.179	75	5.736		
Total	1654.350	78			
Corrected Total	501.273	77			

a. R Squared = .142 (Adjusted R Squared = .119)

$MS_{between}$
MS_{within}

There was a significant difference in weight lost between the diets (p=0.003)

Post hoc tests

If there is a significant ANOVA result, pairwise comparisons are made

They are t-tests with adjustments to keep the type 1 error to a minimum

- Tukey's and Scheffe's tests are the most commonly used post hoc tests.
- Hochberg's GT2 is better where the sample sizes for the groups are very different.

Post hoc tests

- Which diets are significantly different?

Multiple Comparisons

Dependent Variable: Weight lost on diet (kg)

	(I) Diet	(J) Diet	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	1	2	.2741	.67188	.912	-1.3325	1.8806
		3	-1.8481*	.67188	.020	-3.4547	-.2416
	2	1	-.2741	.67188	.912	-1.8806	1.3325
		3	-2.1222*	.65182	.005	-3.6808	-.5636
	3	1	1.8481*	.67188	.020	.2416	3.4547
		2	2.1222*	.65182	.005	.5636	3.6808

- Write up the results and conclude with which diet is the best.

Pairwise comparisons

- Results

Test	p-value
Diet 1 vs Diet 2	
Diet 1 vs Diet 3	
Diet 2 vs Diet 3	

- Report:

Pairwise comparisons

- Results

Test	p-value
Diet 1 vs Diet 2	P = 0.912
Diet 1 vs Diet 3	P = 0.02
Diet 2 vs Diet 3	P = 0.005

There is no significant difference between Diets 1 and 2 but there is between diet 3 and diet 1 ($p = 0.02$) and diet 2 and diet 3 ($p = 0.005$).

The mean weight lost on Diets 1 (3.3kg) and 2 (3kg) are less than the mean weight lost on diet 3 (5.15kg).

Assumptions for ANOVA

Assumption	How to check	What to do if assumption not met
Normality: The residuals (difference between observed and expected values) should be normally distributed	Histograms/ QQ plots/ normality tests of residuals	Do a Kruskall-Wallis test which is non-parametric (does not assume normality)
Homogeneity of variance (each group should have a similar standard deviation)	Levene's test	Welch test instead of ANOVA and Games-Howell for post hoc or Kruskall-Wallis

Ex: Can equal variances be assumed?

- Null:

Levene's Test of Equality of Error Variances[a]

Dependent Variable: WeightLOST

F	df1	df2	Sig.
.659	2	75	.520

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + Diet

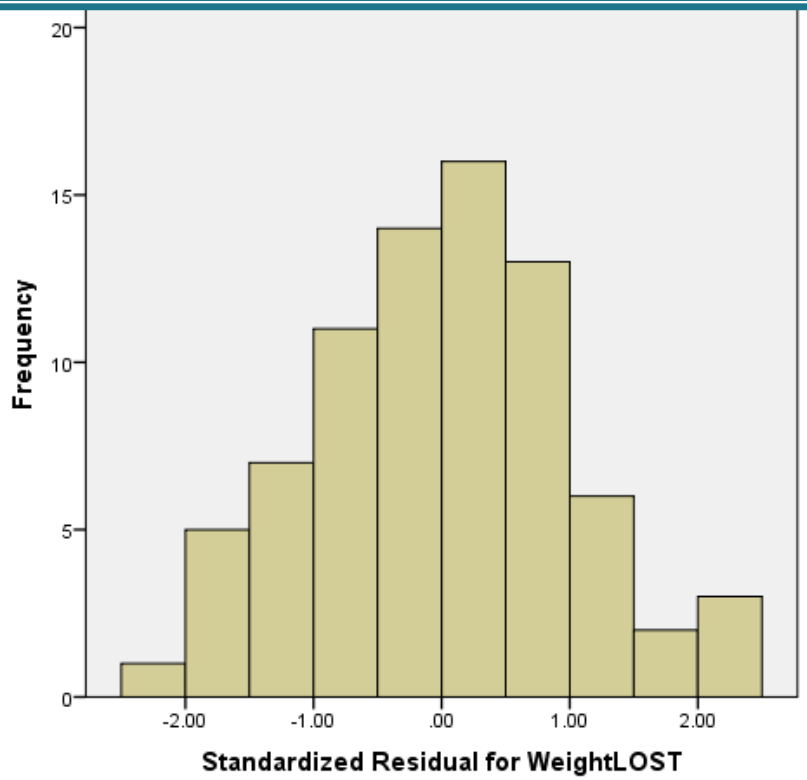
p =

Reject/ do not reject

- Conclusion:

Exercise: Can normality be assumed?

Histogram of standardised residuals



Can normality be assumed?

Should you:

a) Use ANOVA

b) Use Kruskall-Wallis

Ex: Can equal variances be assumed?

- Null:

Levene's Test of Equality of Error Variances[a]

Dependent Variable: WeightLOST

F	df1	df2	Sig.
.659	2	75	.520

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + Diet

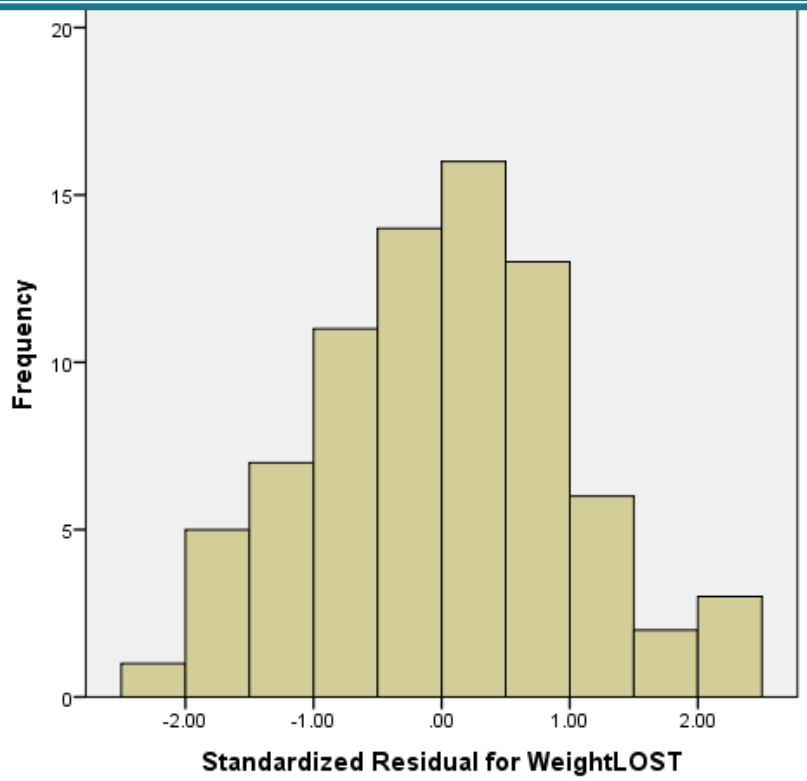
p = 0.52

Do not reject

- Conclusion: Equality of variances can be assumed

Ex: Can normality be assumed?

Histogram of standardised residuals



Can normality be assumed?

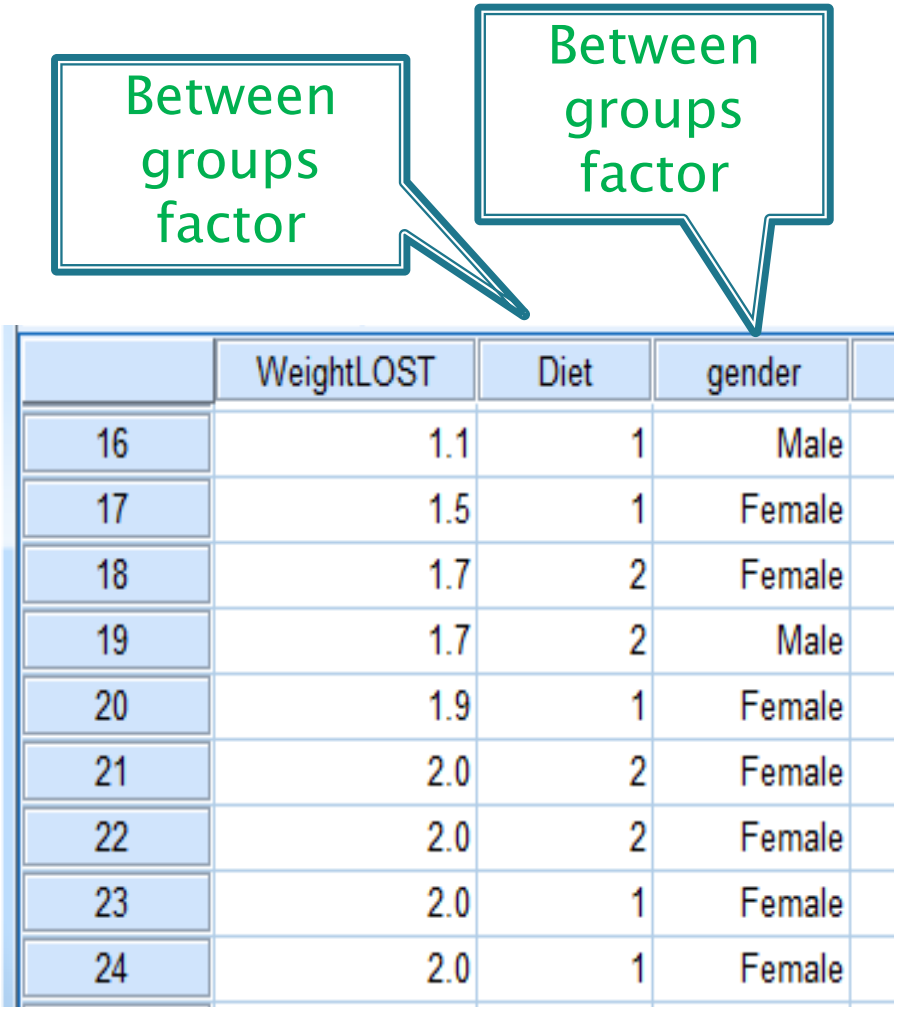
Yes

Use ANOVA

ANOVA

- Two-way ANOVA has 2 categorical independent between groups variables

e.g. Look at the effect of gender on weight lost as well as which diet they were on



	WeightLOST	Diet	gender
16	1.1	1	Male
17	1.5	1	Female
18	1.7	2	Female
19	1.7	2	Male
20	1.9	1	Female
21	2.0	2	Female
22	2.0	2	Female
23	2.0	1	Female
24	2.0	1	Female

Two-way ANOVA

- Dependent = Weight Lost
- Independents: Diet and Gender

- Tests 3 hypotheses:

1. Mean weight loss does not differ by diet
2. Mean weight loss does not differ by gender
3. There is no interaction between diet and gender

What's an interaction?

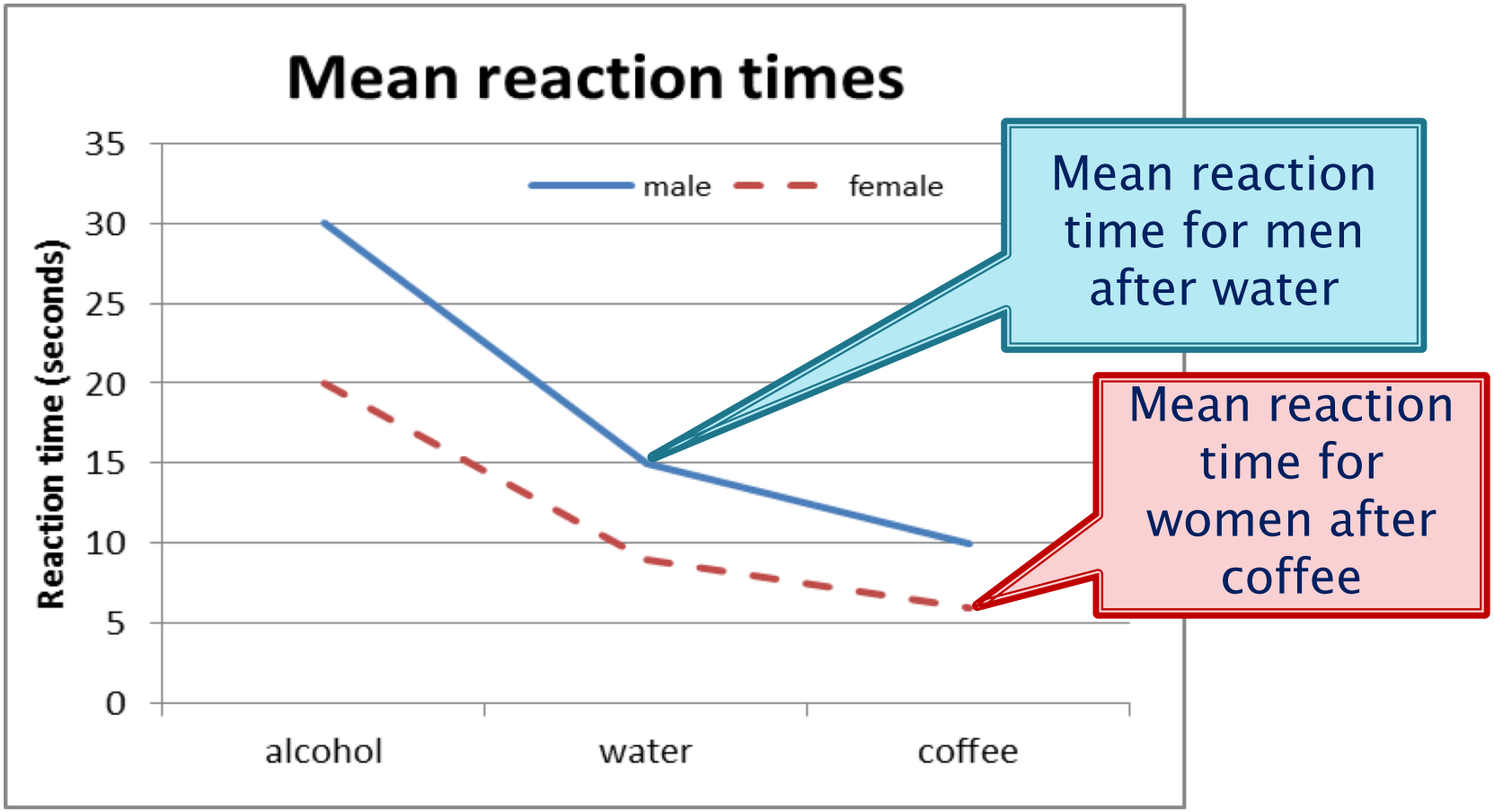
Means plot

- Mean reaction times after consuming coffee, water or beer were taken and the results by drink or gender were compared.

Mean Reaction times	male	female
alcohol	30	20
water	15	9
coffee	10	6

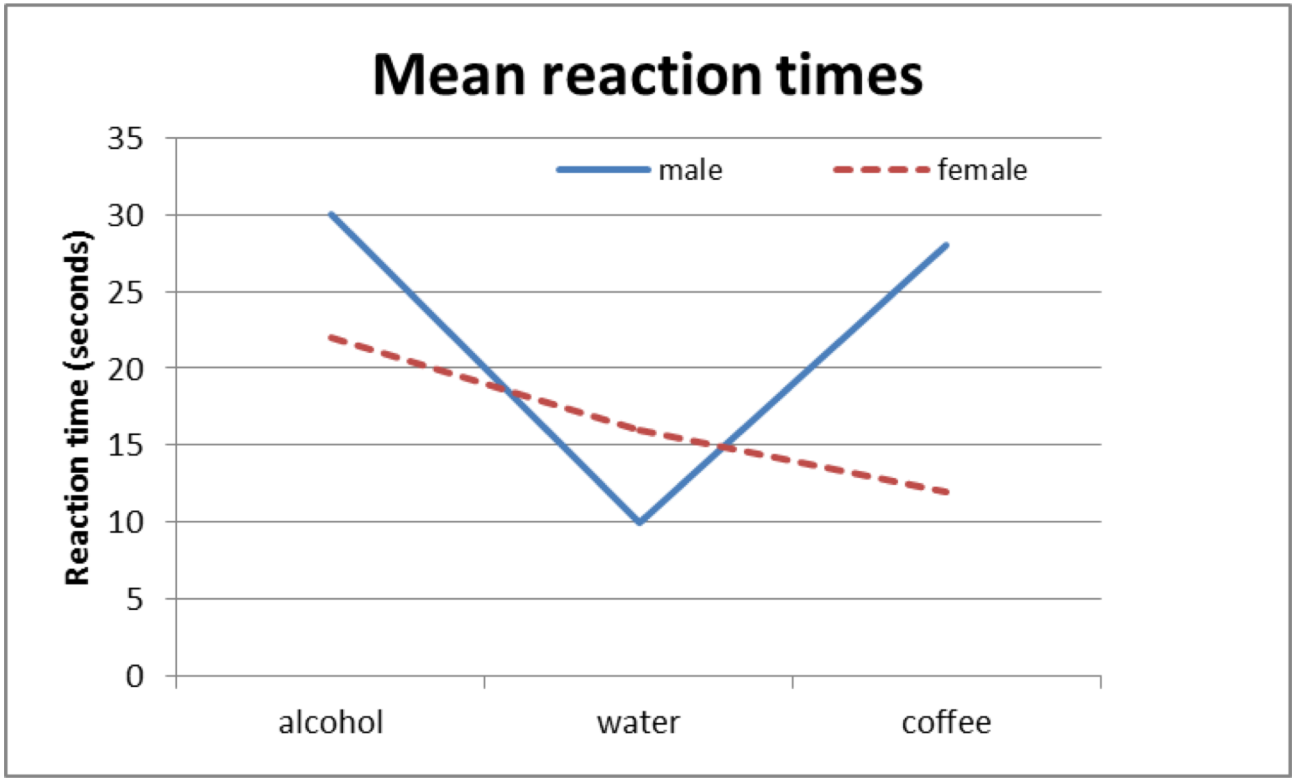
Means/ line/ interaction plot

- No interaction between gender and drink



Means plot

- Interaction between gender and drink



ANOVA

- Mixed between-within ANOVA includes some repeated measures and some between group variables

e.g. give some people margarine B instead of A and look at the change in cholesterol over time

Between groups factor

ID	Cholesterol Before	Cholesterol after 4 weeks	Cholesterol after 8 weeks	Margarine
1	6.42	5.83	5.75	A
2	6.76	6.2	6.13	A
3	6.56	5.83	5.71	A
4	4.8	4.27	4.15	B
5	8.43	7.71	7.67	A
6	7.49	7.12	7.05	B
7	8.05	7.25	7.1	A
8	5.05	4.63	4.67	A