



# **Introduction to SPSS + Statistical Tests for Quantitative Data**

By Research Comm :)



# Overview

1. Introduction to stats
2. Deeper dive into stats
3. SPSS functions



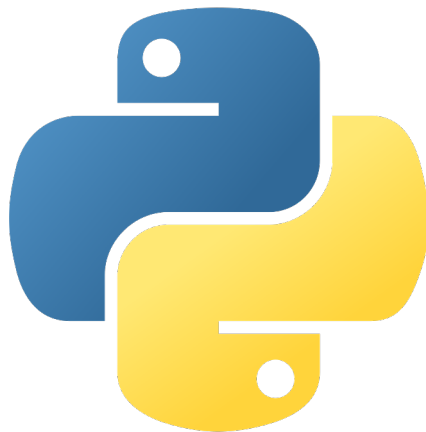


## Why SPSS

- Easy to learn (just need to know how to click buttons)
- For quick analysis of results
- Relevant calculations required for the test are mostly spoon-fed to us :)
- Intuitive interface

## Why not SPSS

- Unable to handle more complex statistical analyses
- \$\$\$ :(

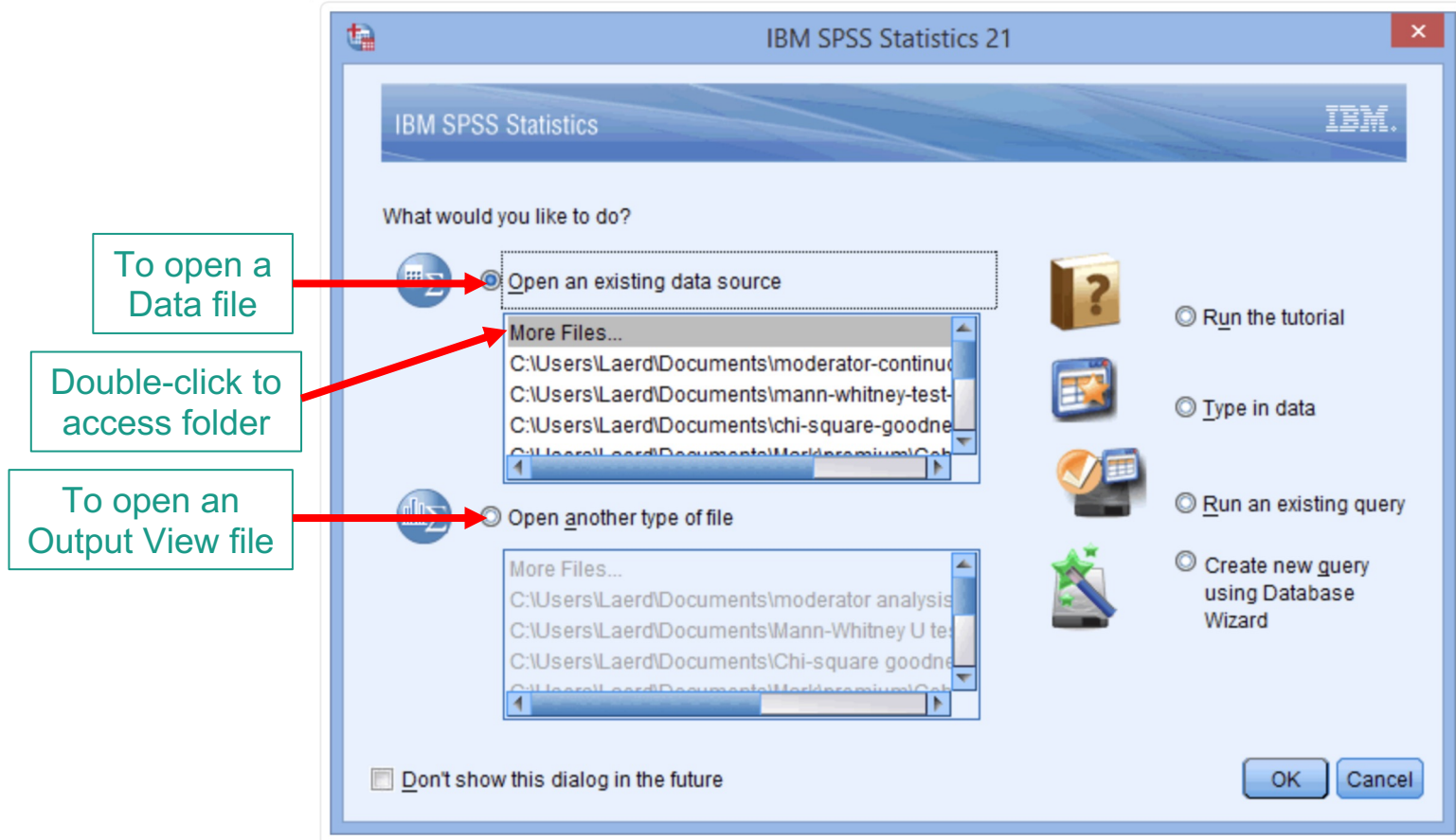




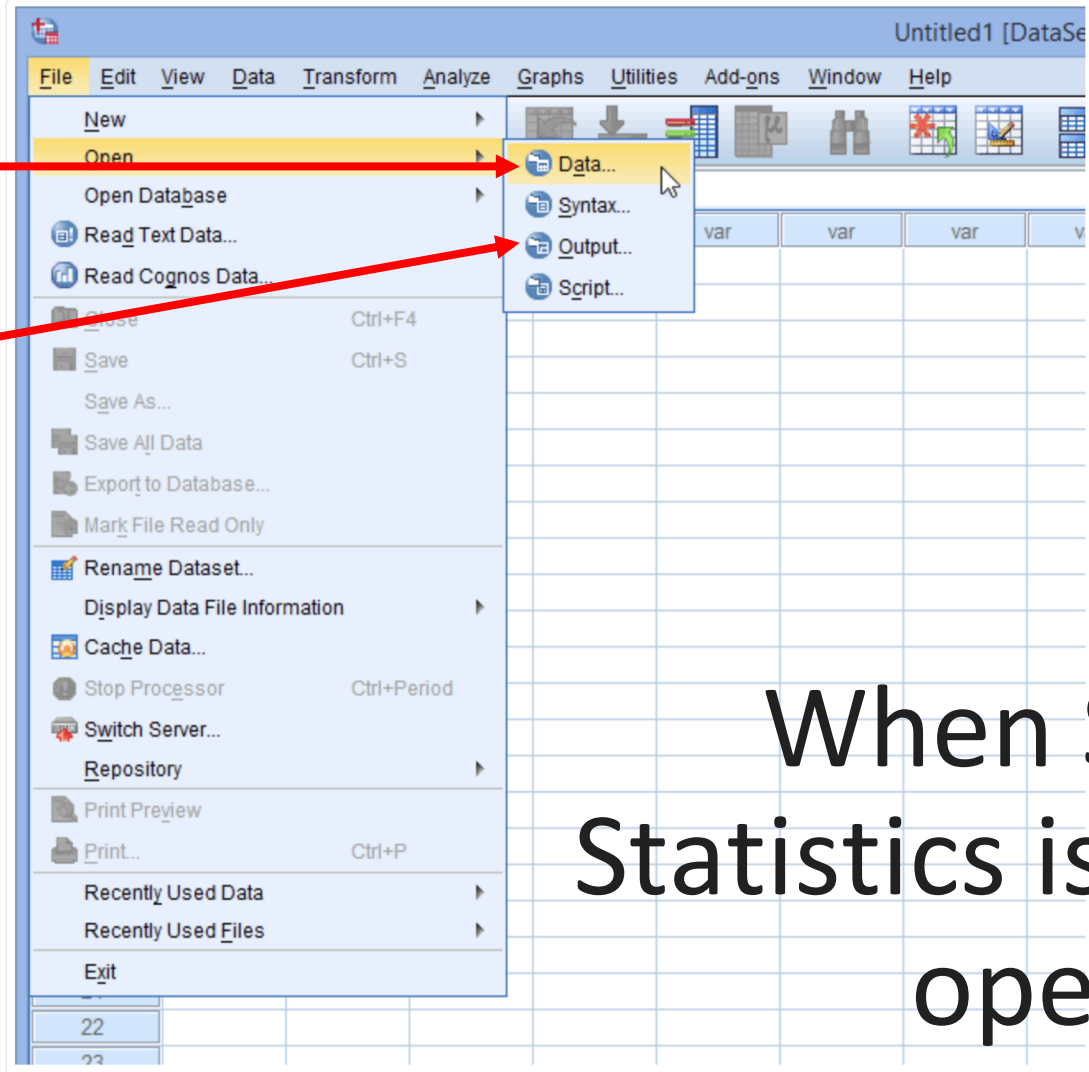


# 1. (re)Introduction to stats

# Opening a File



# When starting SPSS Statistics



To open a Data file

To open an Output View file

When SPSS Statistics is already open

# Data Setup

# Data View

Each column represents a variable

data-setup.sav [DataSet5] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

4 : gender 2

	gender	education	age	qu1	qu2	qu3	var	var
1	Male	School	18	Strongly Agree	Agree	Strongly Agree		
2	Male	College	20	Agree	Neutral	Agree		
3	Male	University	24	Strongly Disagree	Strongly Disagree	Neutral		
4	Female	School	17	Agree	Strongly Disagree	Agree		
5	Female	College	19	Strongly Agree	Strongly Agree	Strongly Agree		
6	Female	University	21	Agree	Neutral	Agree		
7	Male	School	16	Agree	Agree	Neutral		
8	Female	University	23	Neutral	Strongly Disagree	Disagree		

Data View Variable View

IBM SPSS Stat

# Variable View

Each row represents a new variable

Each column represents a property of the variable

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	gender	Numeric	8	0	Gender of participants	{1, Male}...	None	10	Center	Nominal	None
2	education	Numeric	8	0	Education level	{1, School}...	None	10	Center	Ordinal	None
3	age	Numeric	8	0	Age of participants	None	None	10	Center	Scale	None
4	qu1	Numeric	8	0	Question 1	{1, Strongly Agree}...	None	14	Center	Ordinal	None
5	qu2	Numeric	8	0	Question 2	{1, Strongly Agree}...	None	14	Center	Ordinal	None
6	qu3	Numeric	8	0	Question 3	{1, Strongly Agree}...	None	14	Center	Ordinal	None
7											
8											

Data View **Variable View**

IBM SPSS Statis

### TYPE

The variable's data type (e.g. numeric, currency)

### WIDTH

Maximum no. of characters that can be entered

### LABEL

Explanatory label for variable

### COLUMNS

How long each variable name can be (otherwise truncated)

### NAME

The name of the variable must start with a **letter**  
X characters: /, \* or blank space

### DECIMAL

No. of decimal places

### VALUES

Text label for categorical variables

### MISSING

Describe missing values

### ALIGN

Alignment of data in Data View

### MEASURE

Variable's measurement type

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	gender	Numeric	8	0	Gender of participants	{1, Male}...	None	10	Center	Nominal	None
2	education	Numeric	8	0	Education level	{1, School}...	None	10	Center	Ordinal	None
3	age	Numeric	8	0	Age of participants	None	None	10	Center	Scale	None
4	qu1	Numeric	8	0	Question 1	{1, Strongly Agree}...	None	14	Center	Ordinal	None
5	qu2	Numeric	8	0		{1, Strongly Agree}...	None	14	Center	Ordinal	None
6	qu3	Numeric	8	0		{1, Strongly Agree}...	None	14	Center	Ordinal	None
7											

Open variable type dialogue box to select



# Value Labels

Assign a numerical value  
to categorical groups

Key in numerical  
value

Key in category

Click "Add"

Value Labels

Value Labels

Value:

Label:

Add

Change

Remove

1.00 = "Strongly Agree"  
2.00 = "Agree"  
3.00 = "Neutral"  
4.00 = "Disagree"  
5.00 = "Strongly Disagree"

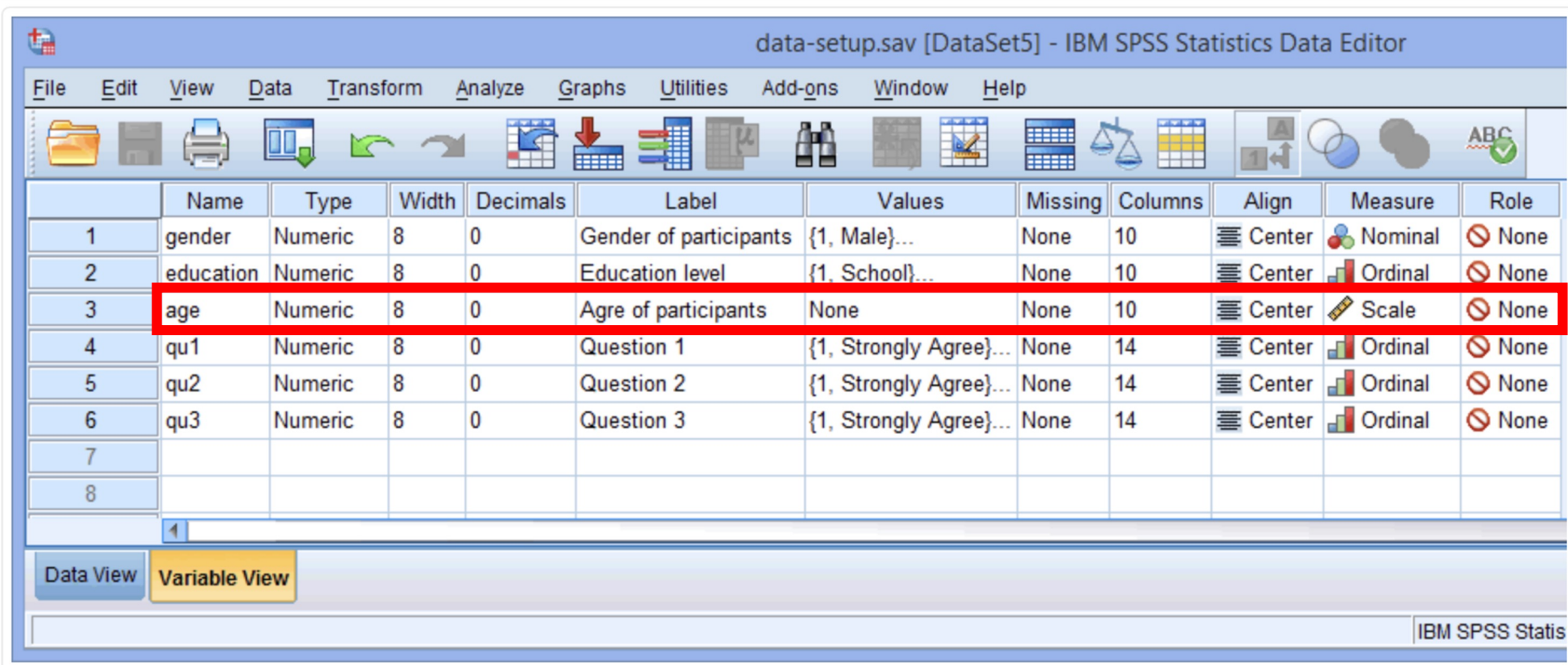
Spelling...

OK Cancel Help

# Entering Continuous Data

data-setup.sav [DataSet5] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	gender	Numeric	8	0	Gender of participants	{1, Male}...	None	10	Center	Nominal	None
2	education	Numeric	8	0	Education level	{1, School}...	None	10	Center	Ordinal	None
3	age	Numeric	8	0	Age of participants	None	None	10	Center	Scale	None
4	qu1	Numeric	8	0	Question 1	{1, Strongly Agree}...	None	14	Center	Ordinal	None
5	qu2	Numeric	8	0	Question 2	{1, Strongly Agree}...	None	14	Center	Ordinal	None
6	qu3	Numeric	8	0	Question 3	{1, Strongly Agree}...	None	14	Center	Ordinal	None
7											
8											

Data View Variable View

IBM SPSS Statis

# Entering Categorical Data

data-setup.sav [DataSet5] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

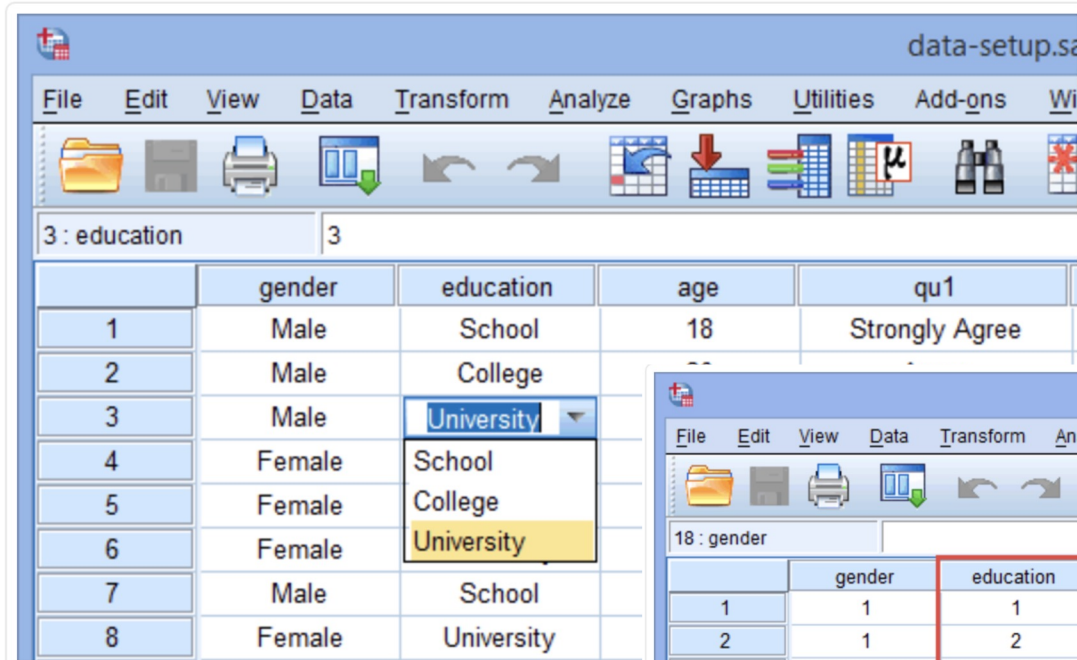
The screenshot shows the IBM SPSS Statistics Data Editor interface. The main window displays a list of variables with their properties. The 'gender' and 'education' rows are highlighted with a red border. The 'gender' row shows a 'Numeric' type with a 'Label' of 'Gender of participants' and 'Values' of '{1, Male}...'. The 'education' row shows a 'Numeric' type with a 'Label' of 'Education level' and 'Values' of '{1, School}...'. The 'Measure' column for 'gender' is 'Nominal' and for 'education' is 'Ordinal'. The 'Role' column for both is 'None'.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	gender	Numeric	8	0	Gender of participants	{1, Male}...	None	10	Center	Nominal	None
2	education	Numeric	8	0	Education level	{1, School}...	None	10	Center	Ordinal	None
3	age	Numeric	8	0	Age of participants	None	None	10	Center	Scale	None
4	qu1	Numeric	8	0	Question 1	{1, Strongly Agree}...	None	14	Center	Ordinal	None
5	qu2	Numeric	8	0	Question 2	{1, Strongly Agree}...	None	14	Center	Ordinal	None
6	qu3	Numeric	8	0	Question 3	{1, Strongly Agree}...	None	14	Center	Ordinal	None
7											
8											

Data View Variable View

IBM SPSS Statis

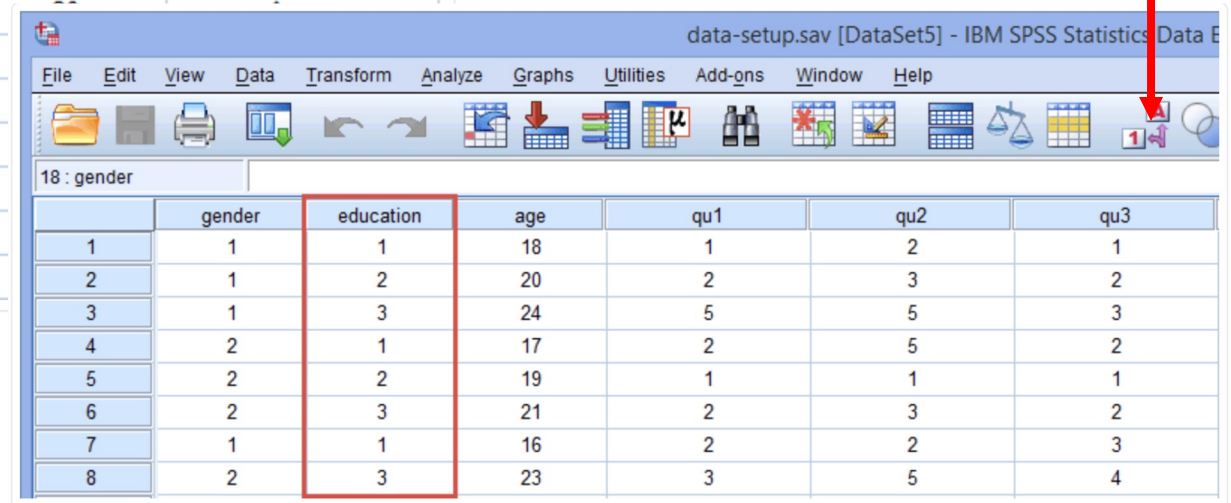
# Entering Categorical Data – Data View



3 : education 3

	gender	education	age	qu1
1	Male	School	18	Strongly Agree
2	Male	College		
3	Male	University		
4	Female	School		
5	Female	College		
6	Female	University		
7	Male	School		
8	Female	University		

Click "Value Labels" on the main toolbar to display underlying codes



18 : gender

	gender	education	age	qu1	qu2	qu3
1	1	1	18	1	2	1
2	1	2	20	2	3	2
3	1	3	24	5	5	3
4	2	1	17	2	5	2
5	2	2	19	1	1	1
6	2	3	21	2	3	2
7	1	1	16	2	2	3
8	2	3	23	3	5	4



## 2. Intro to Stats

- Population vs Sample
- Descriptive vs Inferential Statistics
- Experimental design



# What is Statistics?

Statistics is the study of how to collect, organise, analyse, and interpret numerical information and data

We measure variables from individuals

Eg. measuring CrCl in CKD patients



# Population vs Sample

<b>Population Data</b>	<b>Sample Data</b>
A group of people or objects with a common theme; when every member of that group is considered, it is a population	Small proportion of the population
Census	Sample
Expensive	Cheaper
Time-Consuming	More efficient



# Population vs Sample

Parameters: Measure that describes the entire population

Statistic: A measure that describes only a sample of a population





## Descriptive vs Inferential Statistics

Descriptive Statistics (for both samples and populations): Involve methods of organizing, picturing, and summarizing information from samples AND populations

Inferential Statistics (for samples ONLY): Involves methods of using information from a sample to draw conclusions regarding the populations (HYPOTHESIS TESTING)

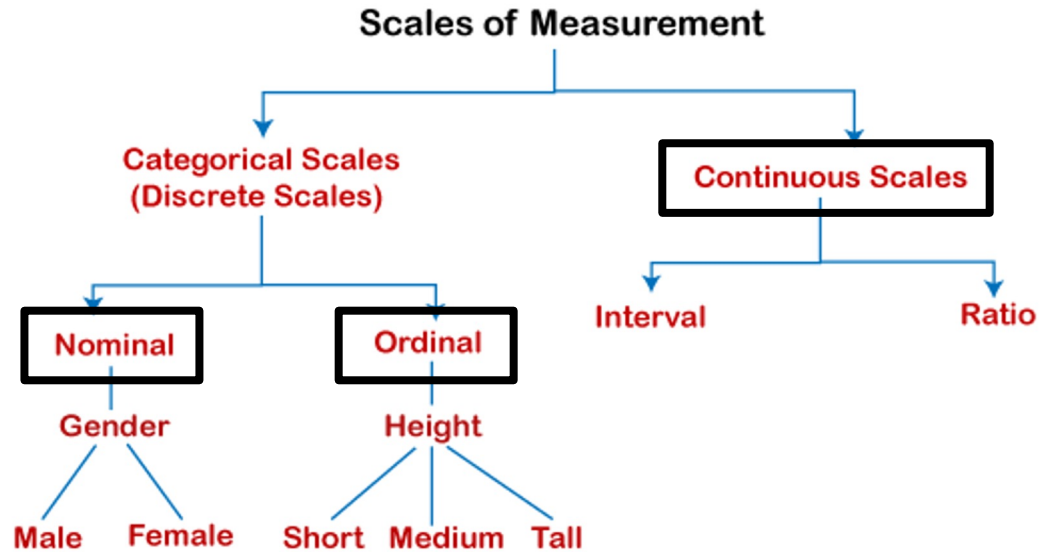
**THIS FORMS THE BASIS OF ALL THAT WE  
DO IN STATISTICS**



# Experimental Design

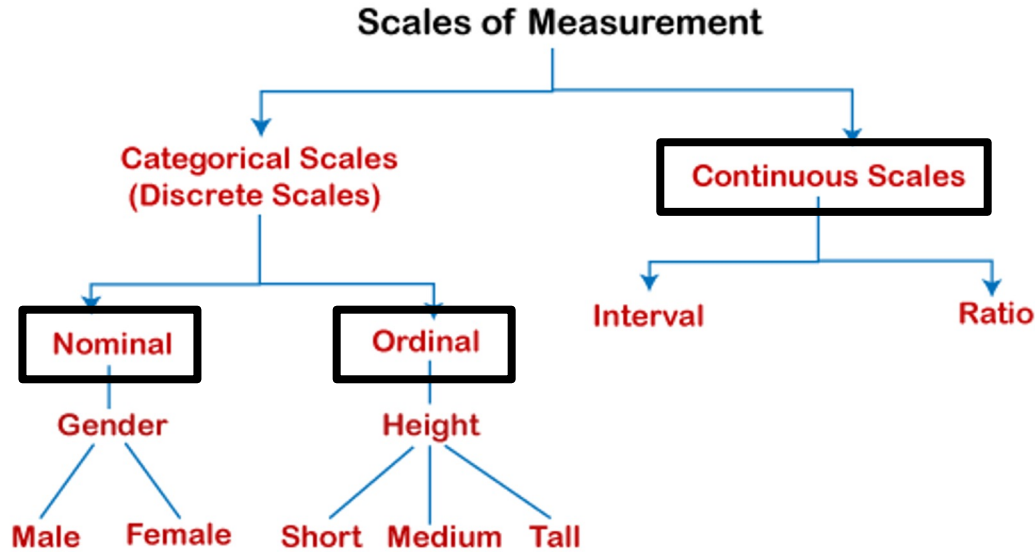
1. State a hypothesis
2. Identify the individuals/population of interest
3. Specify the variables to measure
4. Collect data from your sample
5. Use descriptive or inferential statistics to answer your hypothesis

# Categorical vs Continuous



- You just need to be familiar with these 3 options for SPSS :D

# Categorical vs Continuous





# Experimental Design

EXAMPLE:

Hypothesis: Air pollution causes asthma in children who live in urban settings

Individuals/Population: Children in urban settings

Variables: Degree of air pollution (PSI?), Child is diagnosed with asthma



# Experimental Design

Once we collect data for this study, we want to

a) compute descriptive statistics to describe

1. Associations & Correlations between variables
2. Prediction & Relationships
3. Group Differences

b) compute inferential statistics (aka hypothesis testing) to determine if the statistics in our sample can be generalised to the general population as population parameters



# Experimental Design

The tests that we select are based on

- a) The study design itself (especially the type of variables involved)
  
- a) The data collected



### 3. More on stats

- Data normally distributed?
- Any outliers?
- **Assumptions** met for the specific test?\*
- Reporting results





## To first get an overview of your data...

- Menu bar -> Analyze -> Descriptive Statistics -> Explore...
- Input “Dependent List” and “Factor List”
  - Factor list - e.g. can be used to separate “age” according to “male vs female”
- Select everything else you want to find out in the options buttons at the side

Output:

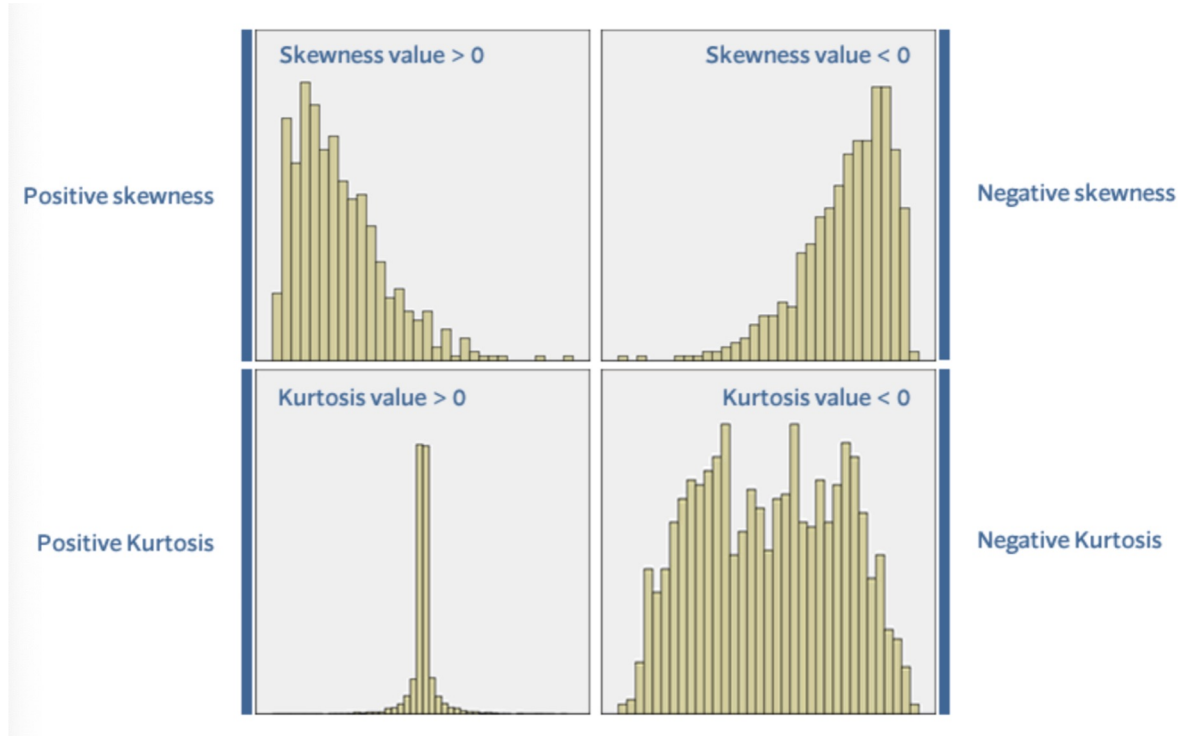
- Summary
- Descriptives
- Outliers
- Histograms
- **\*Checking for normality** (look for table with Shapiro-Wilk)



# Normality Tests

	<b>Numerical Method</b>	<b>Graphical Method</b>
<b>Tests</b>	Skewness and Kurtosis values Shapiro-Wilk test	Normal Q-Q Plot Histogram
<b>Advantages</b>	Objective judgment of normality	Using individual's own judgment to assess normality
<b>Disadvantages</b>	Not sensitive enough at low sample sizes (i.e. not detecting violations of normality)  Overly sensitive to large sample sizes (i.e. very small deviations from normality are detected)	Lack of objectivity

# Skewness and Kurtosis



# Shapiro-Wilk

H0: Data assumes normal distribution

If  $p < 0.05$  -> Reject H0 that data is normally distributed -> AKA Age is not normally distributed in this case

## Tests of Normality

	Gender	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Age	Female	.205	37	<.001	.816	37	<.001
	Male	.125	63	.016	.907	63	<.001

a. Lilliefors Significance Correction

# Histogram and Q-Q Plots

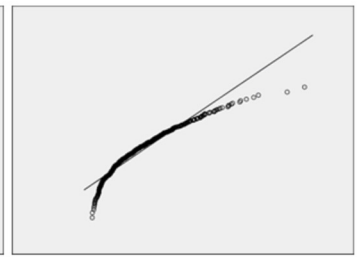
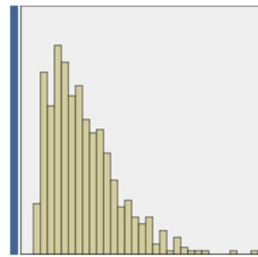
## Histogram

- Look for classic bell-curve shape
- Width of columns (bins) would also affect the shape

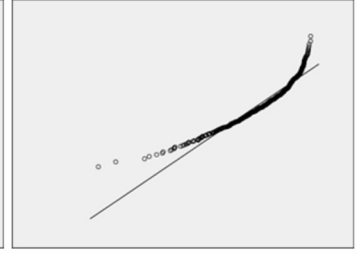
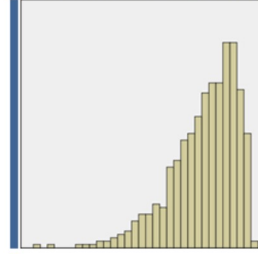
## Q-Q Plots

- Data is normally distributed if it is positioned along the diagonal line

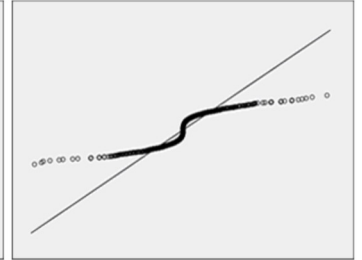
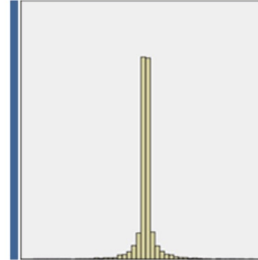
Positive skewness



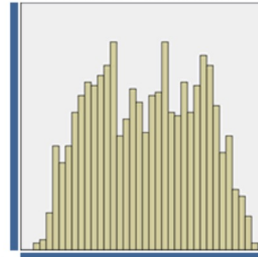
Negative skewness



Positive Kurtosis

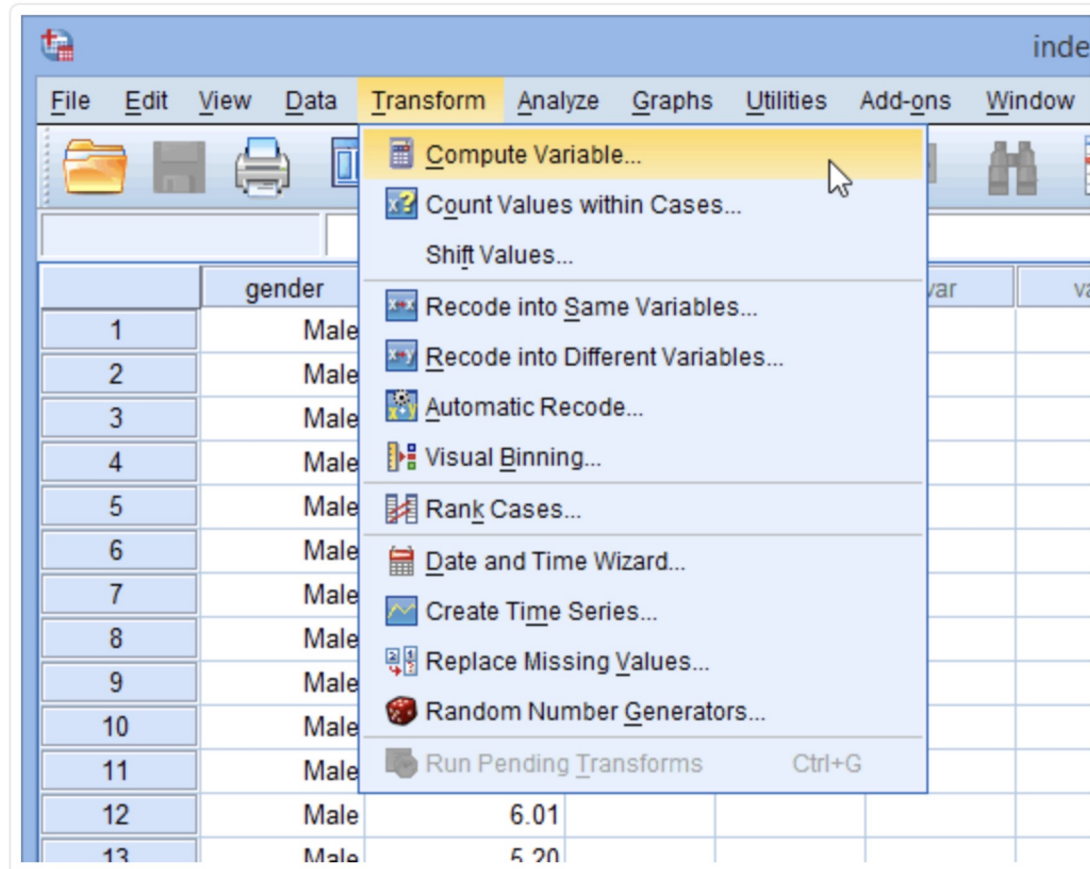


Negative Kurtosis



# Transforming Data

# Performing Transformations



Give a name to the new variable that will be created when you apply a transformation to an existing variable (e.g., you might call it "engagement\_log10" if you want to apply a log10 transformation to the "engagement" variable).

Target Variable:

Type & Label...

- gender
- engagement

Numeric Expression:

Function group:

- All
- Arithmetic
- CDF & Noncentral CDF
- Conversion
- Current Date/Time
- Date Arithmetic
- Date Creation

Functions and Special Variables:

If... (optional case selection condition)

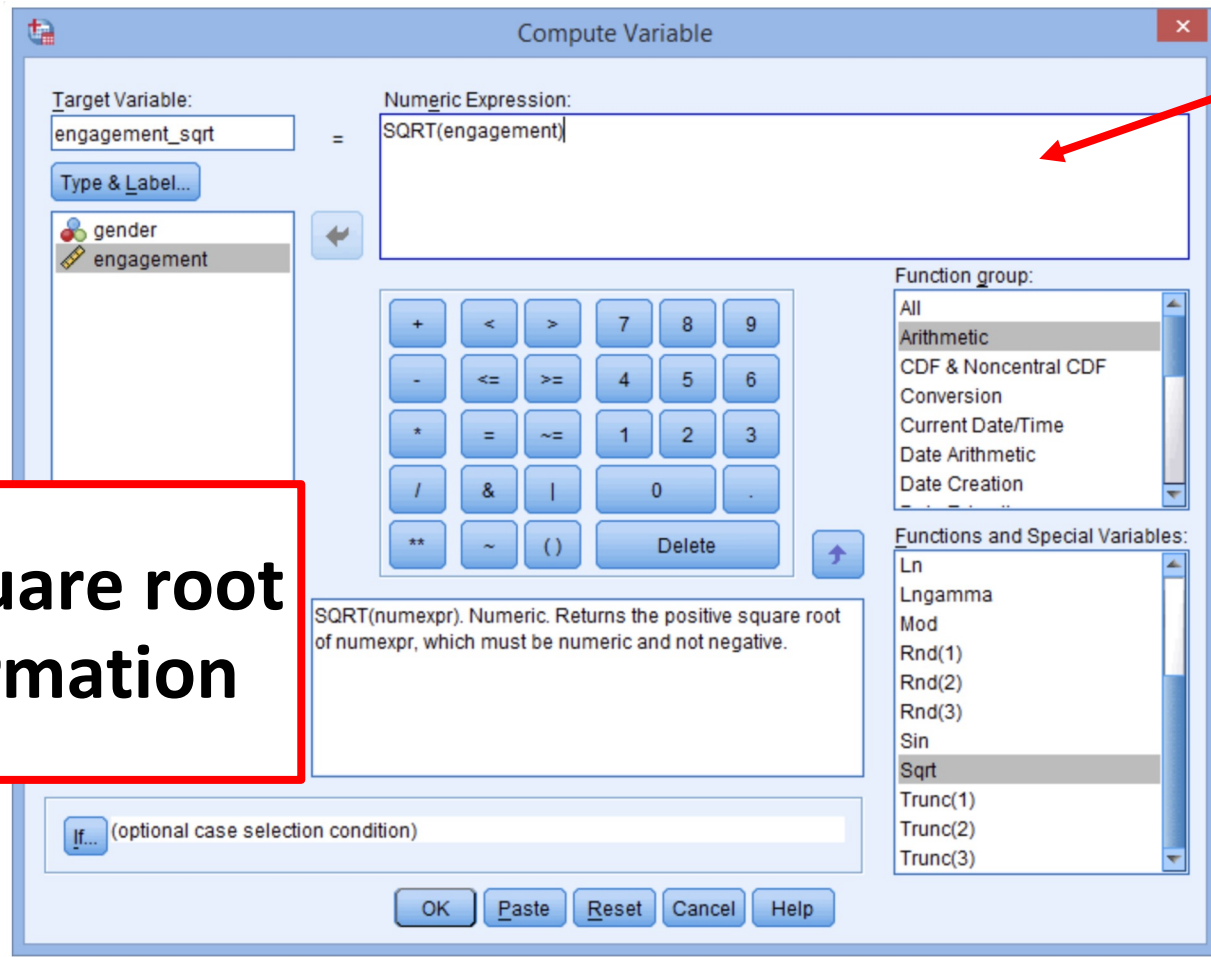
OK Paste Reset Cancel Help

Enter the transformation you want to apply in this box.

A list of predefined functions (e.g., log transformations) that can be selected. All options available will appear in this section below.



# Moderately, positively skewed data



Type:  
SQRT(variable)

# Moderately, negatively skewed data

Target Variable: engagement\_sqrt\_ref

Numeric Expression: SQRT(8.23 - engagement)

Function group: All, Arithmetic, CDF & Noncentral CDF, Conversion, Current Date/Time, Date Arithmetic, Date Creation

Functions and Special Variables: Ln, Lngamma, Mod, Rnd(1), Rnd(2), Rnd(3), Sin, Sqrt, Trunc(1), Trunc(2), Trunc(3)

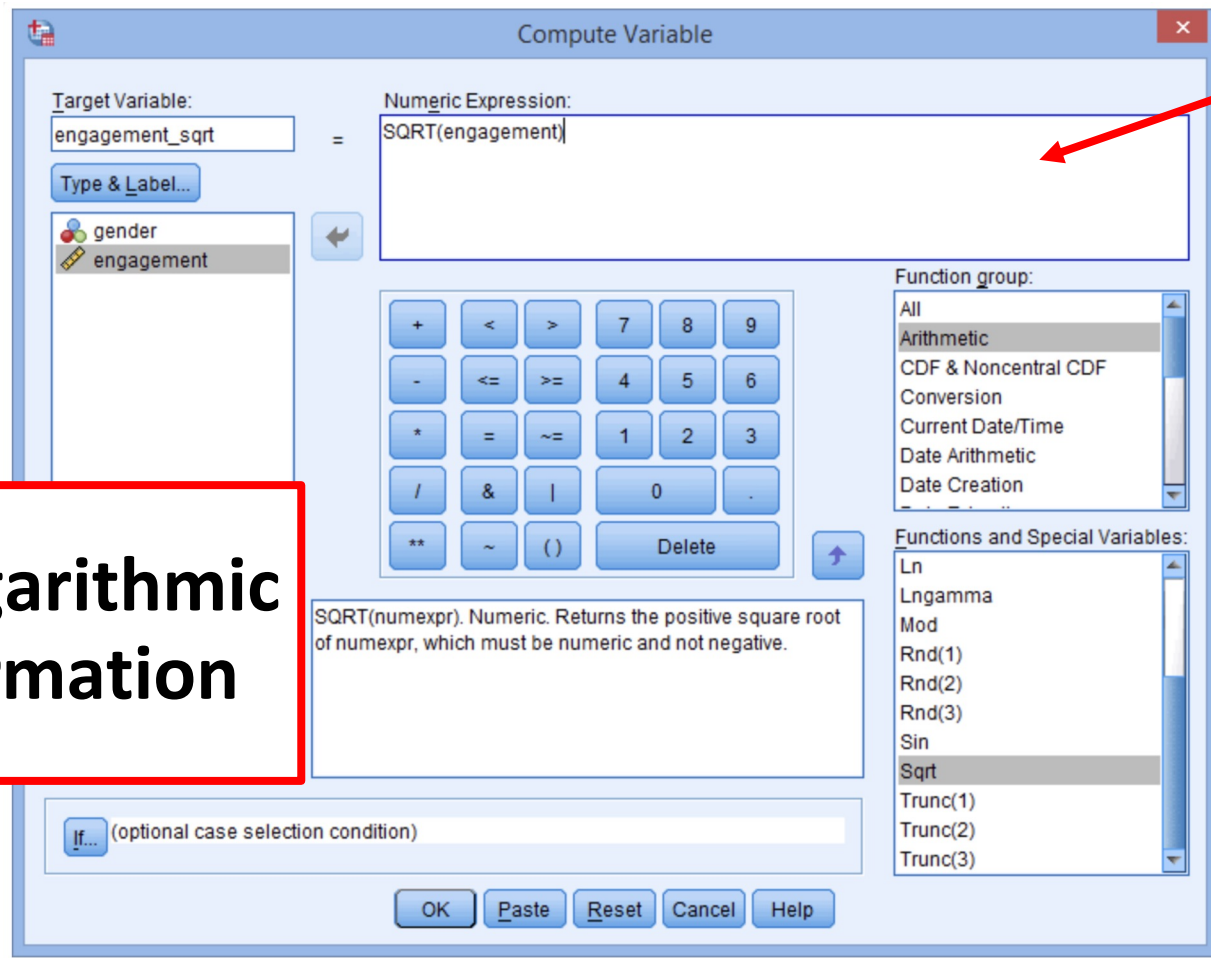
OK Paste Reset Cancel Help

Type: **SQRT**(# - variable)

## Apply reflect and square root transformation

1. Find the largest dependent variable value
2. Add 1 to its value
3. Each dependent value has to be subtracted from the value in (2)
4. Take the square root

# Strongly, positively skewed data



Type:  
**LG10(variable)**

**Apply logarithmic transformation**

# Strongly, negatively skewed data

Target Variable: engagement\_sqrt\_ref

Numeric Expression: SQRT(8.23 - engagement)

Function group: Arithmetic

Functions and Special Variables: Sqrt

SQRT(numexpr). Numeric. Returns the positive square root of numexpr, which must be numeric and not negative.

OK Paste Reset Cancel Help

Type: **LOG10(# - variable)**

## Apply reflect and logarithmic transformation

1. Find the largest dependent variable value
2. Add 1 to its value
3. Each dependent value has to be subtracted from the value in (2)
4. Take the logarithm

# Extremely, positively skewed data

Target Variable: engagement\_sqrt

Numeric Expression: SQRT(engagement)

Type & Label...

gender  
engagement

Function group: Arithmetic

Functions and Special Variables: Sqrt

SQRT(numexpr). Numeric. Returns the positive square root of numexpr, which must be numeric and not negative.

If... (optional case selection condition)

OK Paste Reset Cancel Help

Type: 1/variable

**Apply  
inverse/reciprocal  
transformation**

# Extremely, negatively skewed data

Target Variable: engagement\_sqrt\_ref

Numeric Expression: SQRT(8.23 - engagement)

Function group: All, Arithmetic, CDF & Noncentral CDF, Conversion, Current Date/Time, Date Arithmetic, Date Creation

Functions and Special Variables: Ln, Lngamma, Mod, Rnd(1), Rnd(2), Rnd(3), Sin, Sqrt, Trunc(1), Trunc(2), Trunc(3)

OK Paste Reset Cancel Help

Type: 1/(# - variable)

**Apply reflect and inverse/reciprocal transformation**

1. Find the largest dependent variable value
2. Add 1 to its value
3. Each dependent value has to be subtracted from the value in (2)
4. Take the inverse



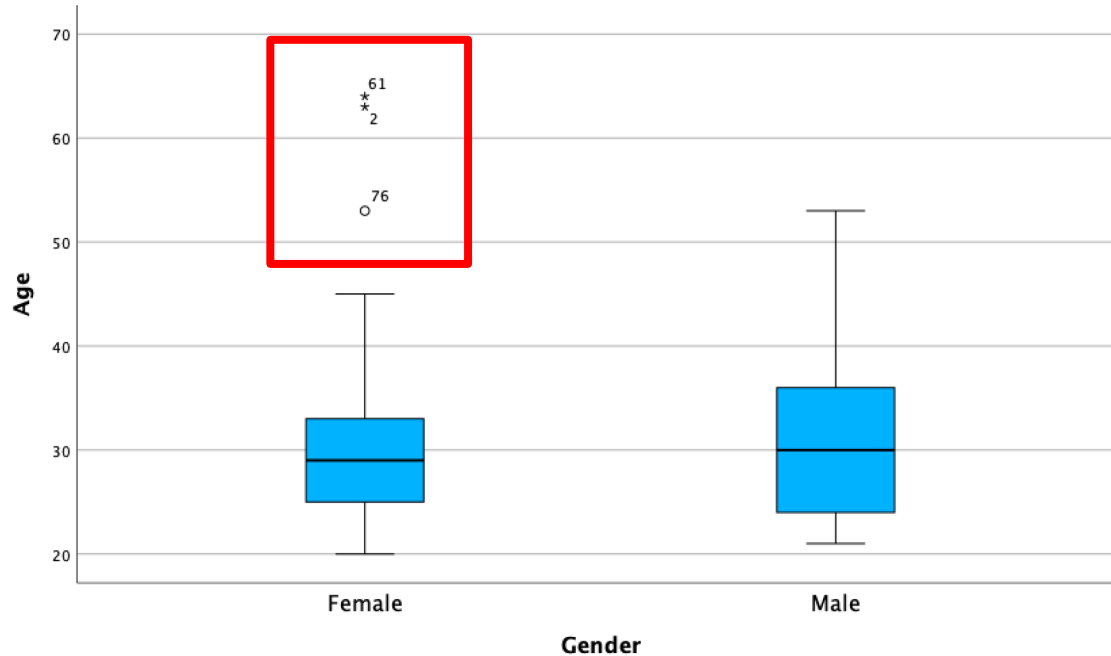
# Data Transformation

<b>Positively Skewed</b>	<b>Negatively Skewed</b>
Square Root	Reflect & Square Root
Log	Reflect & Log
Inverse/Reciprocal	Reflect & Inverse/Reciprocal

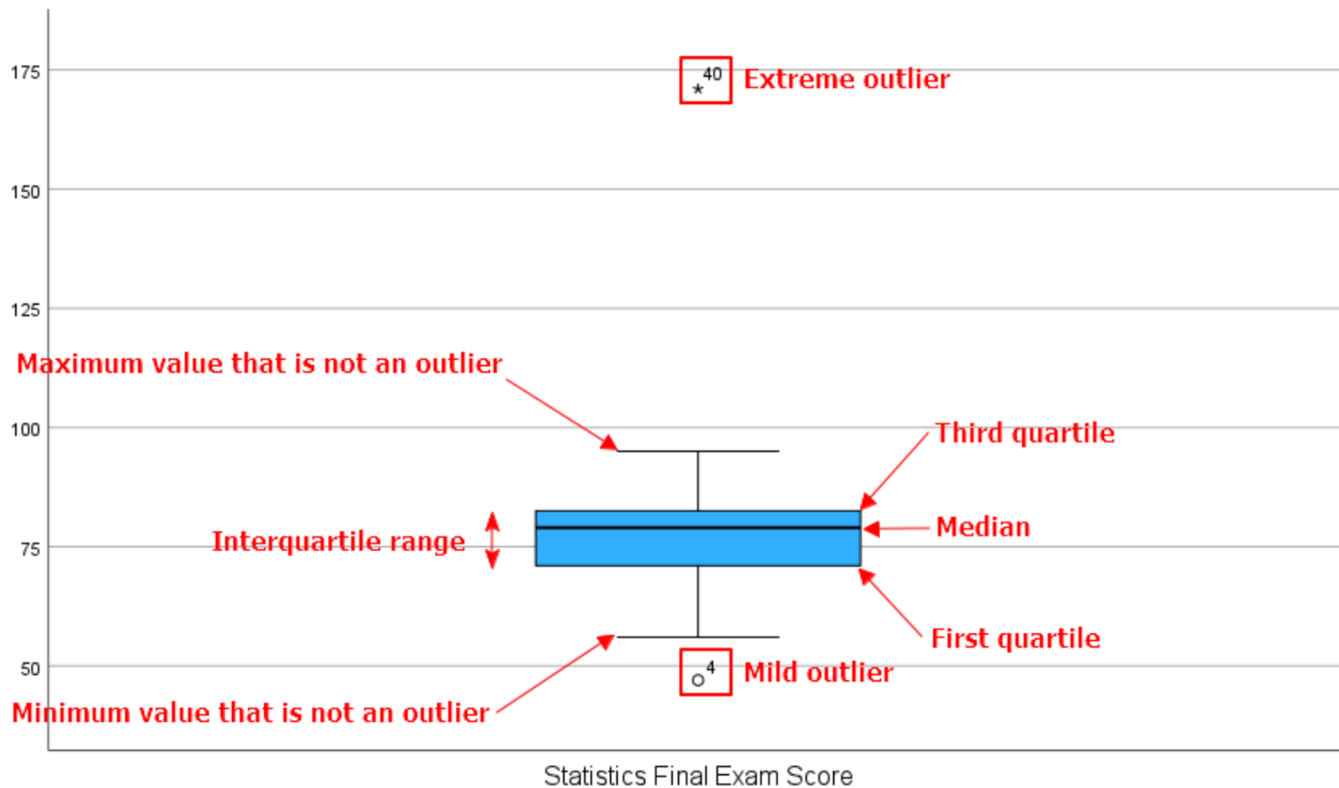
# Outliers

o Mild outliers:  $> 1.5 \times \text{IQR}$  below Q1 or above Q3

\* Extreme outliers:  $> 3.0 \times \text{IQR}$  below Q1 or above Q3









## Dealing with outliers

If keeping outliers:

1. Run **non-parametric** test instead
2. Modify by **replacing with a less extreme value** (e.g. next largest value)
3. **Transform** the dependent variable
4. **Include without change**, if you believe result will not be affected (e.g. similar after running the test with and without the outlier)



## Dealing with outliers

### If removing outliers:

- Generally considered as **last resort**.
  - In your paper: Provide information about the data points removed (e.g. their value and impact on results)

Acceptable in the following example: To investigate the effect of exercise on young males. One participant's cholesterol concentration was particularly high (outlier), indicating considerable risk of heart disease. If the study initially wanted only healthy individuals, and exclude those with risk of heart diseases, then the data can be removed.



## Why is testing for normality important?

Data Setup	Parametric test	Non-Parametric test
1 Variable 2 Categories Between Subjects	independent t-test	Mann-Whitney U test
1 Variable 2 Categories Within-Subjects	paired t-test	Wilcoxon Signed Rank Test
1 Variable >2 Categories Between Subjects	One-way ANOVA	Kruskal Wallis Test
1 Variable >2 Categories Within Subjects	repeated measures ANOVA	Friedman test Mood's median test
1 Variable (Correlation)	Pearson's r	Spearman's $\rho$ (rho)



## After selecting the appropriate tests...

- Pearson's correlation
- Independent t-test
- Mann-Whitney U test (non-parametric independent t-test)
- (One-way) ANOVA
- One-way repeated measures ANOVA (multiple paired-samples t-test)
- Linear regression

**<https://tinyurl.com/spsslkc>**



## Pearson's correlation

Used to determine the strength and direction of a linear relationship between 2 continuous variables

**Pearson correlation coefficient,  $r$ :**

-1 (perfect negative linear relationship)

+1 (perfect positive linear relationship)

0 (no relationship)



## Examples

1. To determine whether there is an association between exam performance and time spent revising
2. To determine whether there is a relationship between "amount of cigarettes reduced" and "withdrawal pain" in participants who failed to quit smoking after 6 months hypnotherapy intervention



Analyze -> Correlate -> Bivariate

The screenshot shows the SPSS software interface. The 'Analyze' menu is open, and the 'Correlate' option is selected. A sub-menu is displayed, showing 'Bivariate...' as the active option. The background shows a data editor window with a search bar containing 'pears' and a data table with columns labeled 'var'.

Power Analysis >  
Meta Analysis >  
Reports >  
Descriptive Statistics >  
Bayesian Statistics >  
Tables >  
Compare Means and Proportions >  
General Linear Model >  
Generalized Linear Models >  
Mixed Models >  
**Correlate >**  
Regression >  
Loglinear >  
Neural Networks >  
Classify >  
Dimension Reduction >  
... >  
Survival >  
Multiple Response >  
Missing Value Analysis >

pearson-correlation.sav [DataSe...]  
pears|  
var var var var  
+ Bivariate with Confidence Intervals...  
**12 Bivariate...**  
123 Partial...  
6 Distances...  
+ Canonical Correlation

The screenshot shows the 'Correlation Coefficients' dialog box. The 'Pearson' option is selected with a checked checkbox, while 'Kendall's tau-b' and 'Spearman' are unselected.

Correlation Coefficients  
 Pearson  Kendall's tau-b  Spearman

#### Missing Values

Exclude cases pairwise

Exclude cases listwise

**Exclude cases pairwise:** Any missing value or variable will only affect the analyses involving that variable

**Exclude cases listwise:** Any missing value or variable will affect all analyses involving that subject



# Assumptions

1. Both variables are continuous (e.g. interval or ratio)
2. Both variables should be paired, each participant has 2 values: 1 for each variable (e.g. a student has 2 variables: revision time and exam performance)
3. Linear relationship between the 2 continuous variables
4. No significant outliers
5. Bivariate normality



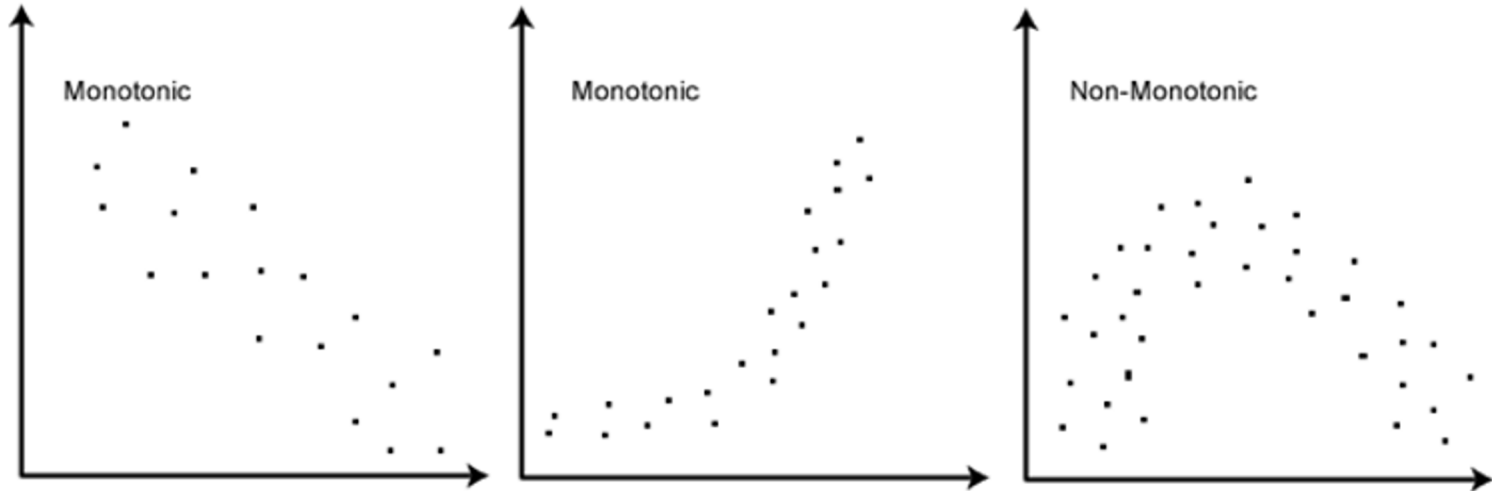
## Checking for linear relationship


Graphs -> Chart builder -> Scatter/Dot (drag and drop) -> Select x- and y-axis (drag and drop) -> Edit properties of axes under "Element Properties" -> Set 0 as minimum



## If not linear?

1. Determine whether **monotonic** or **non-monotonic** relationship





2. If monotonic -> Go to **Spearman's rank-order correlation** OR try to **transform data** into linear relationship if ur a pro...

3. If non-monotonic -> may need to transform one or more variables to get a monotonic relationship



## Interpreting results

	Coefficient value	Strength of association
	$0.1 <  r  < .3$	Small correlation
	$0.3 <  r  < .5$	Medium/moderate correlation
	$ r  > .5$	Large/strong correlation

**Table:** Column meanings for the "**Correlations**" table.



## Coefficient of determination

**Correlations**

		time_tv	cholesterol
time_tv	Pearson Correlation	1	.371**
	Sig. (2-tailed)		.000
	N	100	100
cholesterol	Pearson Correlation	.371**	1
	Sig. (2-tailed)	.000	
	N	100	100

\*\* . Correlation is significant at the 0.01 level (2-tailed).

= square of correlation coefficient ( $r^2$ )

- Proportion of variance in one variable that is "explained" by the other variable

-> e.g. If  $r^2=0.14$ , then daily time spent watching TV statistically explained 14% of the variability in cholesterol concentration



### Correlations

		time_tv	cholesterol
time_tv	Pearson Correlation	1	.371**
	Sig. (2-tailed)		.000
	N	100	100
cholesterol	Pearson Correlation	.371**	1
	Sig. (2-tailed)	.000	
	N	100	100

\*\* . Correlation is significant at the 0.01 level (2-tailed).

Note: Statistical significance here does not determine the strength of the relationship

\*In SPSS, p-value = .000 refers to  $p < 0.0005$

Typically reported as  $p < 0.001$

For p-values  $> 0.001$ , do write the actual values, rather than  $< 0.05$



## Reporting results

### Statistically significant:

A Pearson's product-moment correlation was run to assess the relationship between cholesterol concentration and daily time spent watching TV in males aged 45 to 65 years. One hundred participants were recruited.

Preliminary analyses showed the relationship to be linear with both variables normally distributed, as assessed by Shapiro-Wilk's test ( $p > .05$ ), and there were no outliers.

There was a statistically significant, moderate positive correlation between daily time spent watching TV and cholesterol concentration,  $r(98) = .37$ ,  $p < .0005$ , with time spent watching TV explaining 14% of the variation in cholesterol concentration.



## Reporting results

### Not statistically significant:

A Pearson's product-moment correlation was run to assess the relationship between cholesterol concentration and daily time spent watching TV in males aged 45 to 65 years. One hundred participants were recruited.

Preliminary analyses showed the relationship to be linear with both variables normally distributed, as assessed by Shapiro-Wilk's test ( $p > .05$ ), and there were no outliers.

There was a no statistically significant correlation between daily time spent watching TV and cholesterol concentration,  $r(98) = .28$ ,  $p = .765$ , with time spent watching TV explaining 9% of the variation in cholesterol concentration.

Table 1

*Pearson correlations for main study variables*

	Time watching TV	Cholesterol	CRP
Cholesterol	.371*		
CRP	.341*	.886*	
TAG	.312*	.858*	.981*

**Note.** CRP = C-Reactive Protein, TAG = Triglyceride, \* = statistically significant at  $p < .05$  level.




## Independent samples t-test

Parametric test used to determine whether difference between 2 independent groups is significant



## Examples

1. Whether there is a statistically significant difference in salary (dependent variable) between "under 30y/o" and "above 30y/o" groups
  - Differences between 2 independent groups
2. Whether there is a statistically significant difference in the body fat in mm (dependent variable) between the 2 groups split according to level of physical activity (independent variable) ('none', 'frequent' exercise)
  - Differences between interventions



3. Whether the difference in change in blood glucose concentration for each group was significant in dietary group and control group at the end of the 6 week period (can also look at two-way mixed ANOVA or one-way ANCOVA)  
- Differences in change post intervention



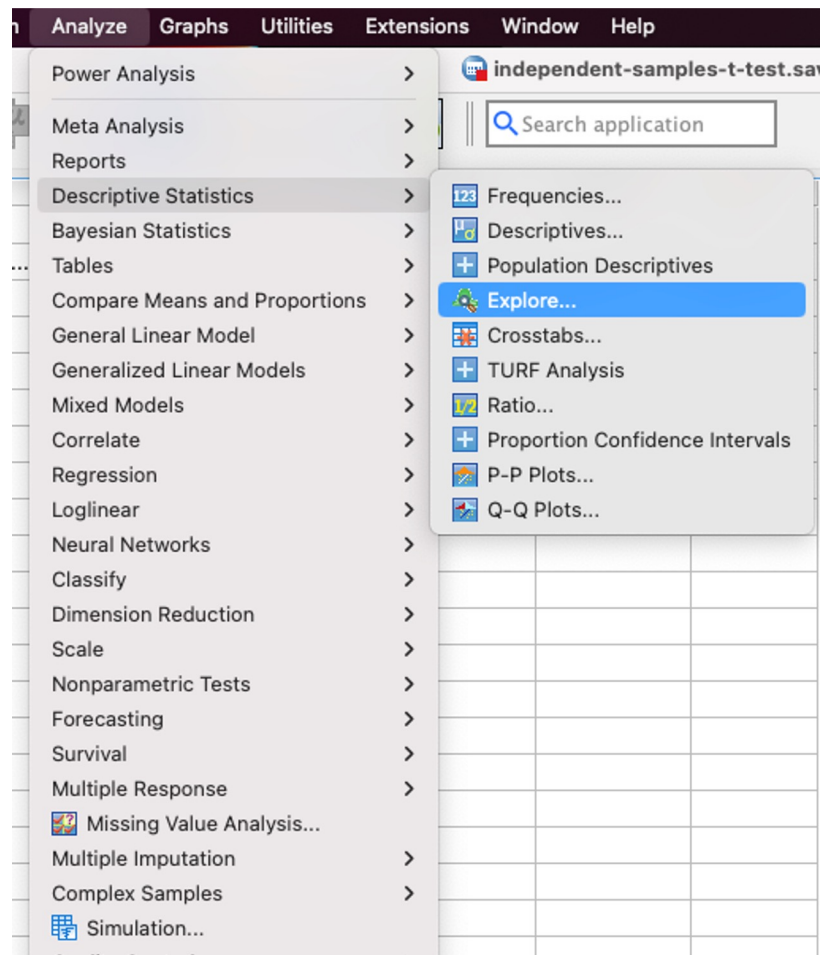
## Assumptions

1. One dependent variable, continuous or ordinal
2. One independent variable, consisting of 2 categorical, independent groups
3. Independence of observations (e.g. same participant cannot be in more than 1 group, otherwise look at **paired-samples t-test**)
4. No significant outliers (dependent variables) within the groups of independent variables
5. Dependent variable approx. normally distributed for each group of independent variable
6. Homogeneity of variances of dependent variables in each group of independent variable





# Explore...

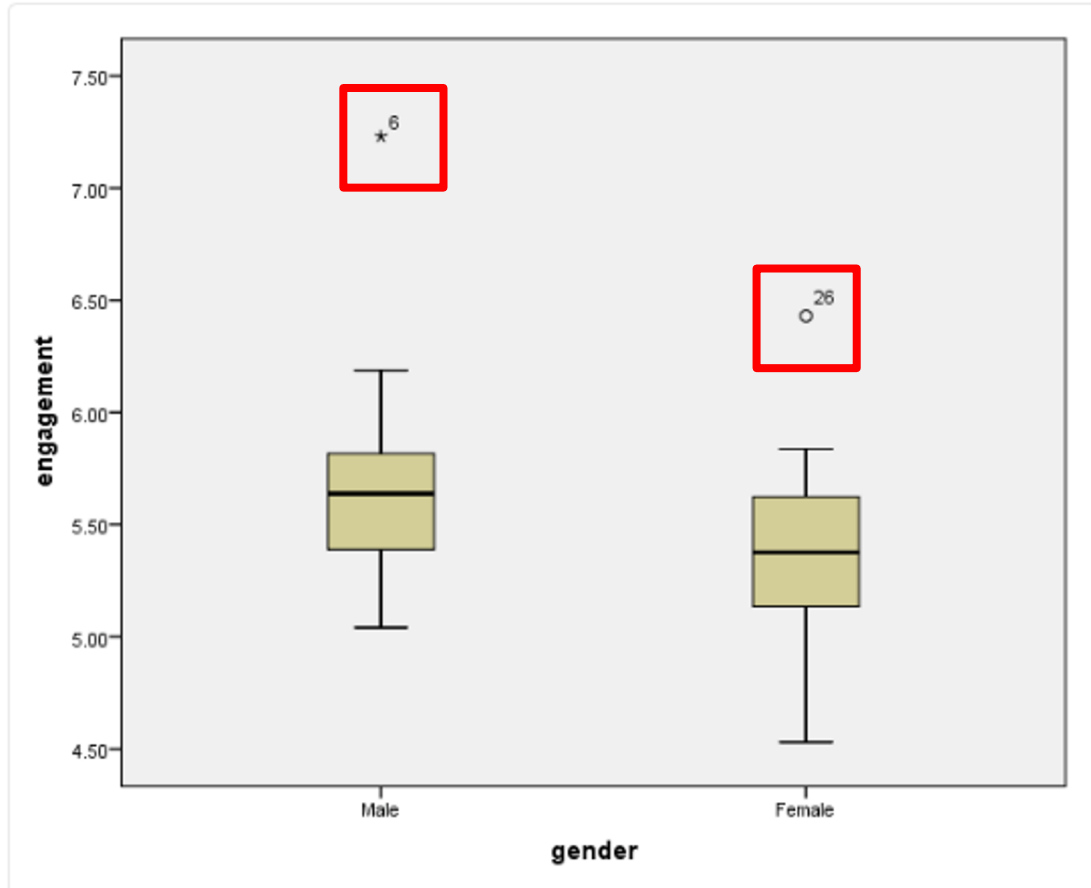


The image shows a screenshot of the SPSS software interface. The 'Analyze' menu is open, displaying a list of statistical analysis options. The 'Explore...' option is highlighted in blue. The background shows a spreadsheet with a search bar and the file name 'independent-samples-t-test.sav'.

independent-samples-t-test.sav

Search application

- Power Analysis >
- Meta Analysis >
- Reports >
- Descriptive Statistics >**
  - Frequencies...
  - Descriptives...
  - Population Descriptives
  - Explore...**
  - Crosstabs...
  - TURF Analysis
  - Ratio...
  - Proportion Confidence Intervals
  - P-P Plots...
  - Q-Q Plots...
- Bayesian Statistics >
- Tables >
- Compare Means and Proportions >
- General Linear Model >
- Generalized Linear Models >
- Mixed Models >
- Correlate >
- Regression >
- Loglinear >
- Neural Networks >
- Classify >
- Dimension Reduction >
- Scale >
- Nonparametric Tests >
- Forecasting >
- Survival >
- Multiple Response >
- Missing Value Analysis...
- Multiple Imputation >
- Complex Samples >
- Simulation...



o **Mild outliers:**  $> 1.5 \times \text{IQR}$   
*below Q1 or above Q3*

\* **Extreme outliers:**  $> 3.0 \times \text{IQR}$   
*below Q1 or above Q3*



## Managing outliers

If keeping outliers:

1. Run non-parametric Mann-Whitney U test instead
2. Modify by replacing with a less extreme value (e.g. next largest value, meaning 2nd largest value = 5.55, the altered value of outlier = 5.56)
3. Transform the dependent variable
4. Include without change, if you believe result will not be affected (e.g. similar after running the test with and without the outlier)



### **If removing outliers:**

Generally considered as last resort. Provide information about the data points removed (e.g. their value and impact on results)

**Acceptable in the following example:** To investigate the effect of exercise on young males. One participant's cholesterol concentration was particularly high (outlier), indicating considerable risk of heart disease. If the study initially wanted only healthy individuals, and exclude those with risk of heart diseases, then the data can be removed.

## Effect size (Cohen, 1988)

\* Journals are increasingly asking for effect sizes to be reported whenever possible :/

	Effect Size	Strength
	.2	small
	.5	medium
	.8	large

Table 4.3: Cohen's d.

### Independent Samples Effect Sizes

	Standardizer <sup>a</sup>	Point Estimate	95% Confidence Interval	
			Lower	Upper
Engagement with TV advert	Cohen's d	.748	.101	1.385
	Hedges' correction	.733	.099	1.358
	Glass's delta	.658	-.003	1.304

- a. The denominator used in estimating the effect sizes.  
Cohen's d uses the pooled standard deviation.  
Hedges' correction uses the pooled standard deviation, plus a correction factor.  
Glass's delta uses the sample standard deviation of the control group.



# Statistical significance vs Effect size

**Statistical significance:** Only tells us whether there is a difference (whether an effect exists)

- **DOES NOT** tell us whether the difference is big, important or helpful in decision making
- If sample is sufficiently large, a stat test will almost always give significant difference
  - E.g. increase in score by 1 point out of 100 points can also be significant...

**Effect size:** Tells us the magnitude of difference between groups



## Statistical significance vs Clinical significance

E.g. Increased knee flexion angles in knee osteoarthritis

- MDC\_90 value for knee flexion contracture: +6 degrees flexion (stratford)
- If results from gait analysis shows statistically significant difference of less than 6 degrees of additional flexion -> **Are these results clinically relevant?**



# Mann-Whitney U Test

Non-parametric test (equivalent of **independent-samples t-test**)

- Can be used when results that require independent-samples t-test do not follow normal distribution (from Shapiro-Wilk test), or have some outliers





## Examples

1. Knee OA group vs Healthy group - comparing KOOS scores (0-100) or pain scores. Healthy group will likely have little to no knee pain and normal knee function. Knee OA group will likely have more pain, and poorer knee function, but these scores generally will not follow a normal distribution if sample sizes are not large enough
2. Working group vs Retired group - comparing stress levels (strongly agree, agree, neutral, disagree, strongly disagree).



## Assumptions

1. One dependent variable, continuous or ordinal
2. One independent variable, consisting of 2 categorical, independent groups (e.g. male and female, employed and unemployed, intervention and control, bus and car)
3. Independence of observations (e.g. same participant cannot take both bus and car), otherwise go to Wilcoxon signed-rank test
4. Look at whether **distribution** of scores for each group of independent variable have **same shape or different shape/variability**



## Additional steps...

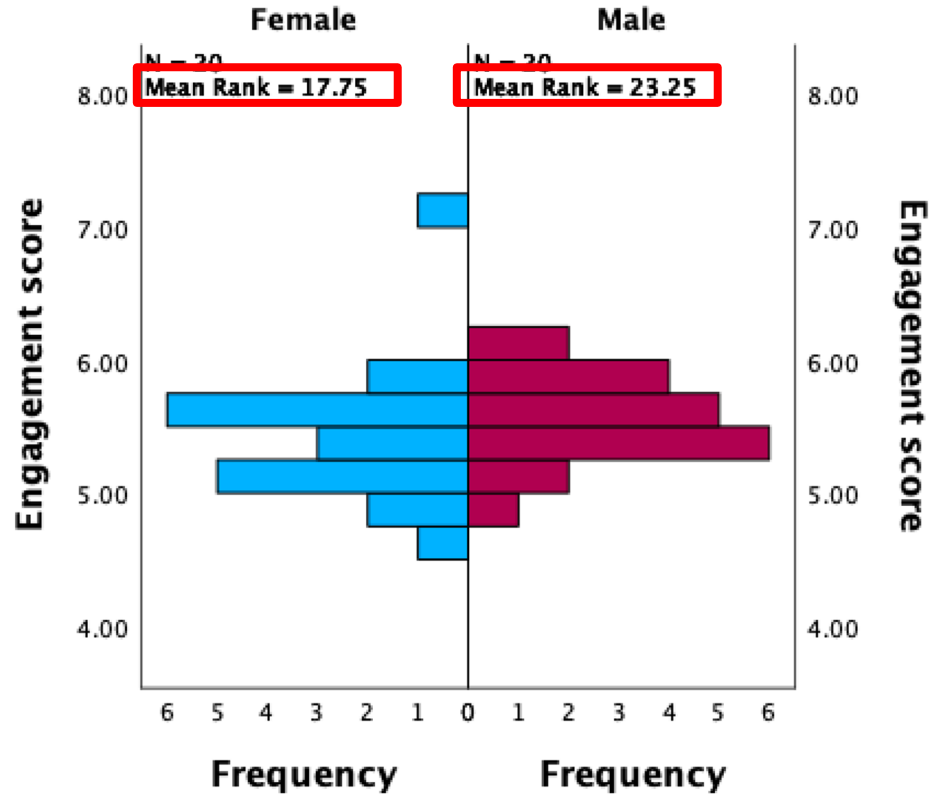
Upon visual inspection of histogram,

If shape of both groups have the same shape -> Test can compare **medians**

If shape of both groups do not have same shape -> Test can compare **mean ranks**

# Independent-Samples Mann-Whitney U Test

## Gender





## Reporting results

For medians,

A Mann-Whitney U test was run to determine if there were differences in engagement score between males and females. Distributions of the engagement scores for males and females were similar, as assessed by visual inspection. **Median** engagement score for males (5.58) and females (5.38) was not statistically significantly different,  $U = 145$ ,  $z = -1.488$ ,  $p = .142$ , using an exact sampling distribution for U (Dineen & Blakesley, 1973).

For ranks,

A Mann-Whitney U test was run to determine if there were differences in engagement score between males and females. **Distributions of the engagement scores for males and females were not similar**, as assessed by visual inspection. Engagement scores for males (**mean rank = 23.25**) were statistically significantly higher than for females (**mean rank = 17.75**),  $U = 218$ ,  $z = -3.422$ ,  $p = .001$ , using an exact sampling distribution for U (Dineen & Blakesley, 1973).



## One-way ANOVA

Parametric test (equivalent of **multiple independent-samples t-test**) used to determine whether difference between **3 or more independent groups**

\*Running multiple t-tests instead would increase Type I error rate. One-way ANOVA would control for the Type I error rate.



## Examples

1. Whether there is a statistically significant difference in the body fat in mm (dependent variable) between the 3 groups split according to level of physical activity (independent variable) ('none', 'moderate', 'frequent' exercise)
2. Whether there is a difference in bone density (dependent variable) between the 3 groups split according to frequency of smoking (independent variable) (non-smoker, occasional, frequent)



## Examples

3. Whether there is a difference in VO<sub>2</sub>max (dependent variable) between swimmers, runners and cyclists
4. Whether there is a difference in smoking cessation (dependent variable) based on treatment type (1 group with nicotine patches, 1 group with hypnotherapy, 1 group with moral support)





# Assumptions

1. One dependent variable, continuous or ordinal
2. One independent variable, consisting of 2 (typically 3) or more categorical, independent groups
3. Independence of observations (e.g. same participant cannot be in more than 1 group)
4. No significant outliers (dependent variables) within the groups of independent variables
5. Dependent variable approx. normally distributed for each group of independent variable
6. **Homogeneity of variances** of dependent variables in each group of independent variable



- Analyze -> Compare means -> One-way ANOVA
- Options -> descriptive, homogeneity of variance test, Welch, means plot
- Post-hoc -> Tukey (homogeneity of variances not violated) + Games-Howell (homogeneity of variances violated)



Assumption of homogeneity of variances in a population:

### **Levene's Test for Equality of Variances**

- If population variance of both groups is equal,  $p > 0.05$ , meeting the assumption of homogeneity of variances
- If population variance of both groups is not equal,  $p < 0.05$ , violating the assumption of homogeneity of variances



### **When homogeneity of variances is met:**

- Refer to ANOVA table for results -> Tells us whether difference exists between any of the groups


Tukey post hoc test: To test all possible group comparisons -> Tells us exactly which groups are different

- Can also use Bonferonni

### **When homogeneity of variances is violated:**

- Refer to "Robust Tests of Equality of Means" table for results of Welch's ANOVA

Games-Howell hoc test: To test all possible group comparisons -> Tells us exactly which groups are different



**Effect size:** omega squared ( $\omega^2$ ) or partial eta squared ( $\eta^2$ )



## Reporting results

One-way ANOVA not statistically significant, but variances were equal:

With test of assumptions: A one-way ANOVA was conducted to determine if the ability to cope with workplace-related stress (CWWS score) was different for groups with different physical activity levels. Participants were classified into four groups: sedentary (n = 7), low (n = 9), moderate (n = 8) and high levels of physical activity (n = 7). There were no outliers, as assessed by boxplot; data was normally distributed for each group, as assessed by Shapiro-Wilk test ( $p > .05$ ); and there was homogeneity of variances, as assessed by Levene's test of homogeneity of variances ( $p = .120$ ). Data is presented as mean  $\pm$  standard deviation. CWWS score increased from the sedentary ( $4.2 \pm 0.8$ ), to low ( $5.9 \pm 1.7$ ), to moderate ( $7.1 \pm 1.6$ ) to high ( $7.5 \pm 1.2$ ) physical activity groups, in that order, but the differences between these physical activity groups was not statistically significant,  $F(3, 27) = 1.116$ ,  $p = .523$ .



## Reporting results

One-way ANOVA was statistically significant, variances were equal and a post hoc test was carried out

Without test of assumptions: A one-way ANOVA was conducted to determine if the ability to cope with workplace-related stress (CWWS score) was different for groups with different physical activity levels. Participants were classified into four groups: sedentary ( $n = 7$ ), low ( $n = 9$ ), moderate ( $n = 8$ ) and high levels of physical activity ( $n = 7$ ). Data is presented as mean  $\pm$  standard deviation. CWWS score was statistically significantly different between different physical activity groups,  $F(3, 27) = 8.316$ ,  $p < .0005$ ,  $\omega^2 = 0.42$ . CWWS score increased from the sedentary ( $4.2 \pm 0.8$ ), to low ( $5.9 \pm 1.7$ ), to moderate ( $7.1 \pm 1.6$ ) to high ( $7.5 \pm 1.2$ ) physical activity groups, in that order. Tukey post hoc analysis revealed that the increase from sedentary to moderate (2.97, 95% CI (0.99 to 4.96)) was statistically significant ( $p = .002$ ), as well as the increase from sedentary to high (3.35, 95% CI (1.30 to 5.40),  $p = .001$ ), but no other group differences were statistically significant.



# One-way repeated measures ANOVA

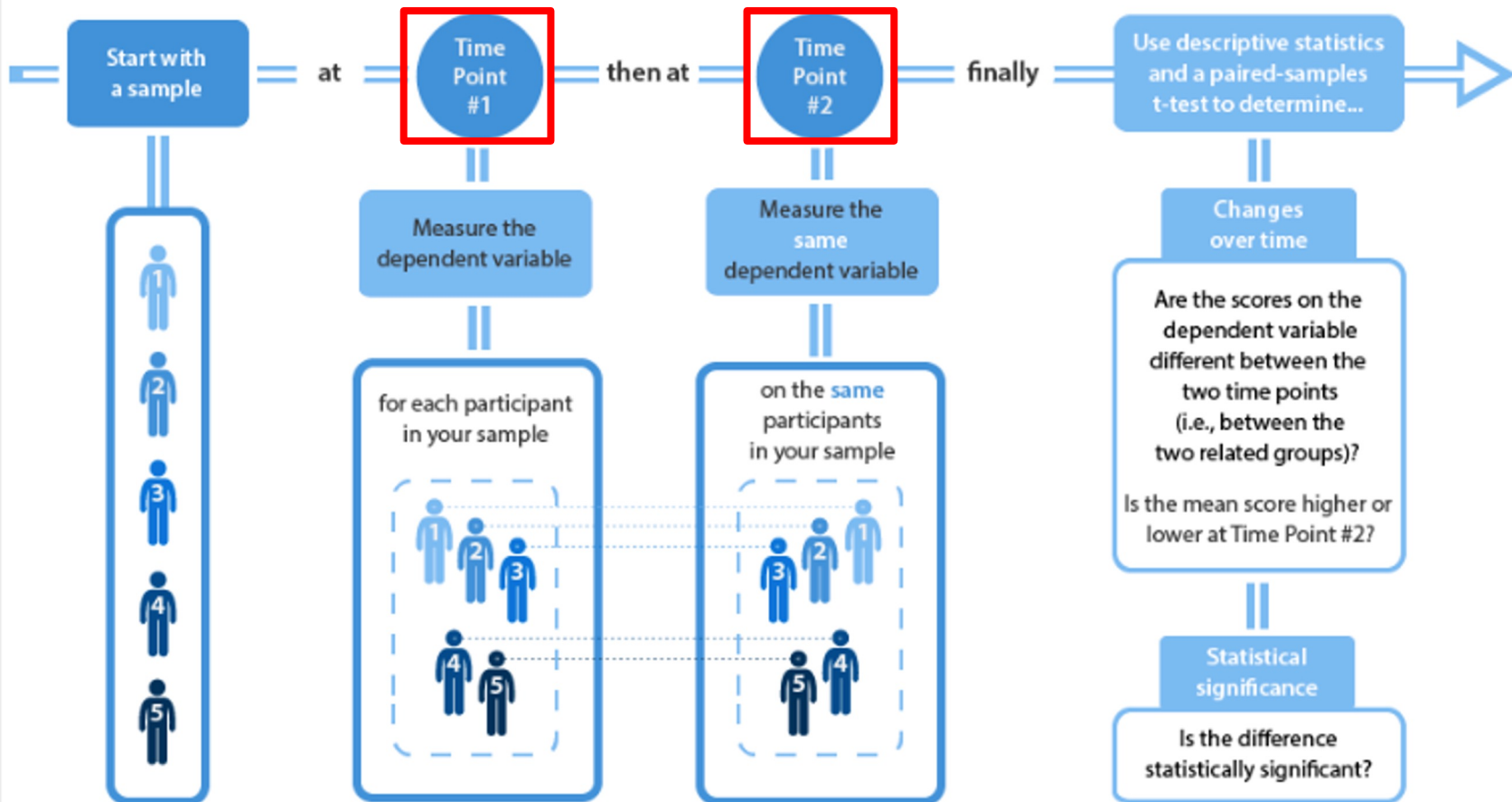
Parametric test (equivalent of multiple **paired-samples t-test**)

Used to determine whether difference between the means of 3 or more levels of a within-subjects factor.

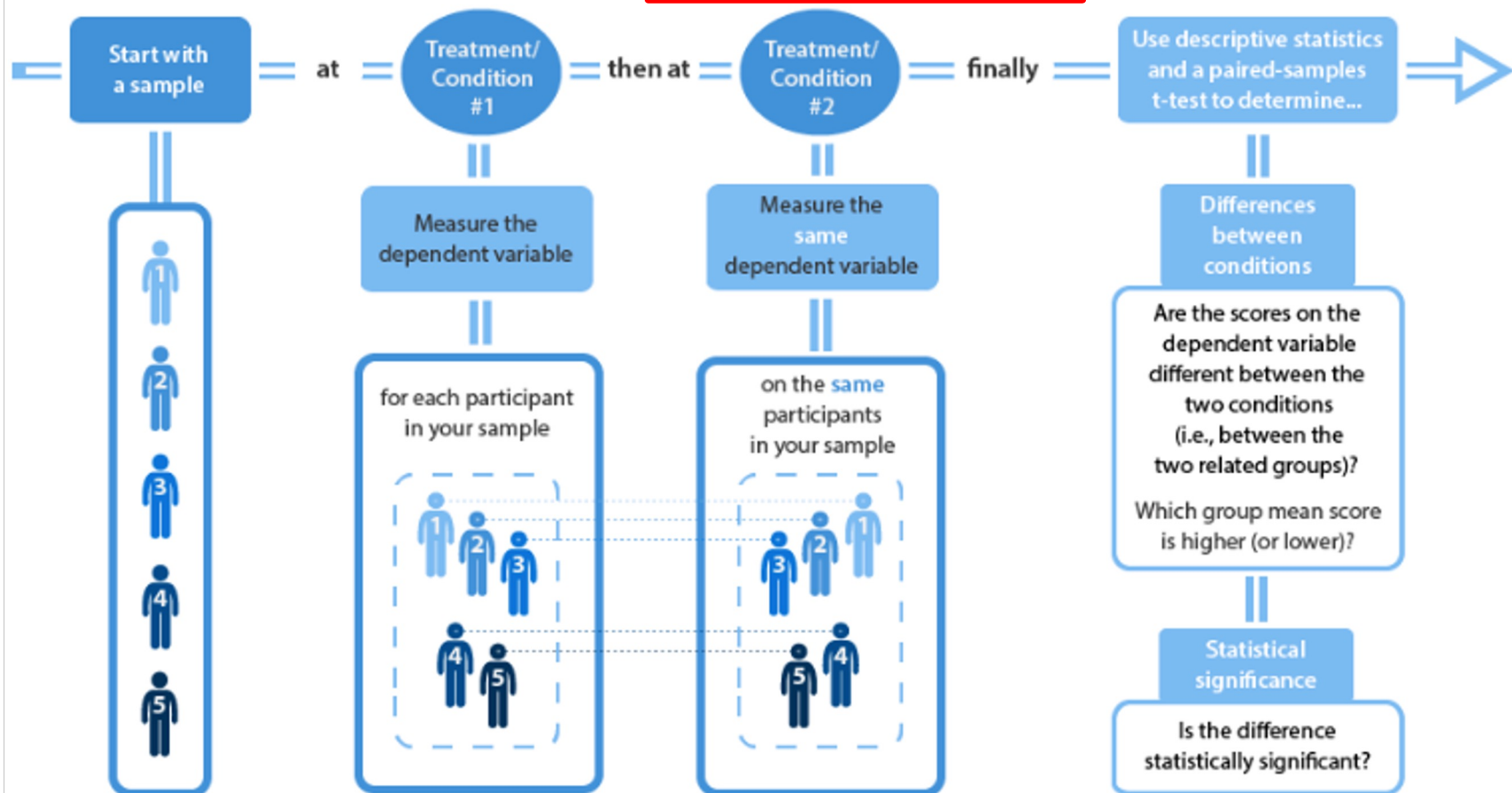
- Participants are the same in the group, tested on 3 or more time durations/scores/treatments on the same dependent variable



## Determine if there are changes over time between two related groups



Determine if there are differences between **two treatments/conditions**

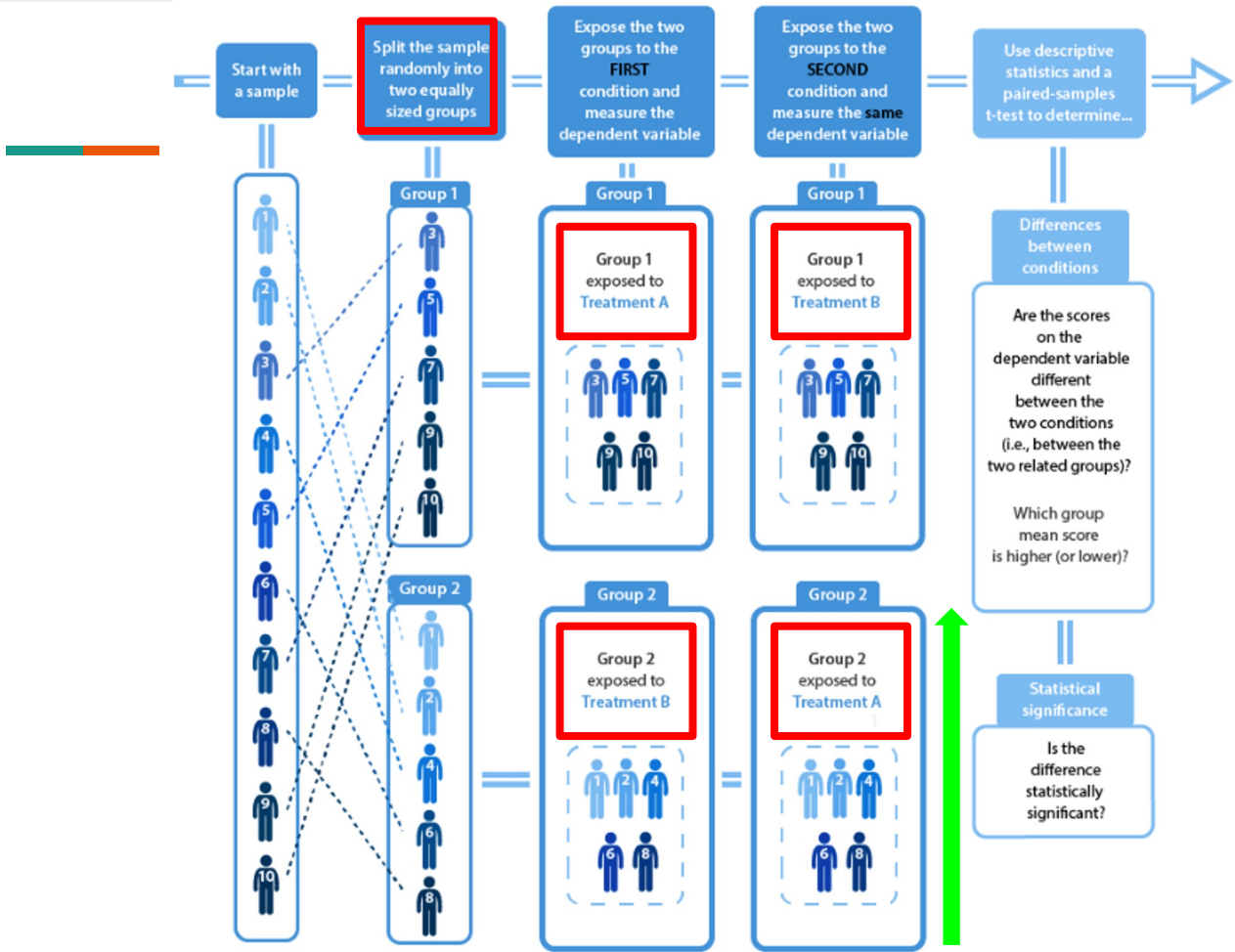




## Cross-over design

- Exposed to 2 conditions in a different order
  - Group 1 - treatment A then B
  - Group 2 - treatment B then A

Reduces possible bias associated with the order in which participants are exposed to a particular condition





## Examples

1. Effects of duration of therapy on cigarette consumption (dependent variable) on 1 group of 30 smokers. Cigarette consumption is measured and compared at 3 different time points (time is the within-subjects factor): before therapy/0 months, 6 months, 12 months mark
  - Determine if there are differences between 3 or more time points
2. Red background vs green background vs blue background on reaction times
  - Determine if there are differences between 3 or more conditions/treatments/interventions

- 
- Determine if there are differences between 3 or more change scores

Same group of 30 participants, undergoing 3 or more different interventions

The same dependent variable (blood glucose concentration) is measured pre and post (1) exercise intervention, (2) dietary intervention and (3) no intervention/control

Change in blood glucose concentrations pre and post interventions is calculated, for all 3 interventions, and are compared using **one-way** repeated measures ANOVA

\*Can also use **two-way** repeated measures ANOVA



# Linear regression

Used to assess the linear relationship between 2 continuous variables to predict the value of a dependent variable based on the value of the independent variable. **Used to determine:**

1. Whether linear regression between the 2 variables is statistically significant
2. **How much variation in the dependent variable is explained by/due to the independent variable**
3. Direction and magnitude of the relationship
4. Predict values of dependent variables based on different values of independent variables

$Y = \beta_0 + \beta_1 X + \epsilon$ , where  $\beta_0$  is the intercept/constant,  $\beta_1$  is the slope coefficient/gradient,  $\epsilon$  is the errors



## Examples

1. Predict the distance athletes can run (dependent, continuous variable) in 30 min based on their VO2 max (independent, continuous variable)
2. Predict how much does the amount of time spent exercising explain cholesterol concentration





# Assumptions

1. One dependent variable, continuous
2. One independent variable, continuous
3. Linear relationship (approx. straight line on scatterplot) present between dependent and independent variables
4. Independence of observations
5. No significant outliers
6. **Variances along the best fit line remain similar throughout**  
(homoscedasticity)
7. **The residuals/errors of regression line are approx. normally distributed**



**Additional steps...**



### 3. Linear relationship?

Visual inspection of scatterplot to determine linear relationship. **If not linear,**

1. Perform a transformation
2. Run a polynomial regression - where 1 or more independent variables is raised to a power of 2 or more ( $Y = \beta_0 + \beta_1X + \varepsilon$  to  $Y = \beta_0 + \beta_1X + \beta_2X^2 + \varepsilon$ ). Possible for curved lines/U-/inverted-U shaped lines (however, may not fit all types of relationships)
3. Run a nonlinear regression



## 4. Independence of observations?

If you suspect that the observations could be related, interpret the Durbin-Watson test in "Model Summary" table.

Durbin-Watson test statistic is between 0-4.

A value closer to 2 shows independence of errors (residuals)



## 5. Any outliers?

To determine outliers, either visually inspect the scatter plot, or perform casewise diagnostics to highlight any cases where the standardised residual  $> 3SD$

### **If keeping outliers:**

1. Transform the dependent variable
2. Include without change, if you believe result will not be affected (e.g. similar after running the linear regression with and without the outlier)

### **If removing outliers:**

Generally considered as last resort. Provide information about the data points removed (e.g. their value and impact on results)



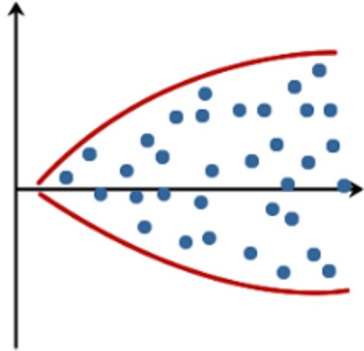
## 6. Are variances along best fit line remains similar throughout?

Visual inspection of scatterplot to test for homoscedasticity

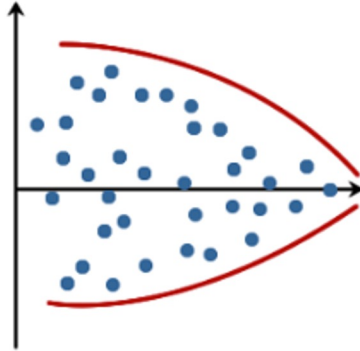
- regression standardized residuals on y-axis
- regression standardised predicted value on x-axis

If there is homoscedasticity, residuals/errors of prediction will be equal across the standardised predicted/fitted values (e.g. plot will show no pattern, and will be approx. constantly spread)

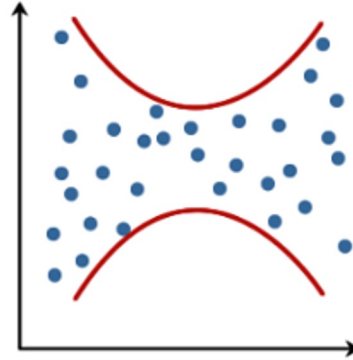
Increasing funnel



Decreasing funnel



Fan shaped



If there is heteroscedasticity, residuals will not be evenly spread (e.g. increasing funnel, decreasing funnel, fan shaped, etc):

1. Run a weighted least squares regression equation
2. Run a regression with robust standard errors
3. Run a robust regression
4. Transform dependent variable



## **7. Residuals approx. normally distributed?**

Check histogram and/or Normal P-P Plot for normality of regression standardised residuals





## To determine how well the model fits:

Percentage/proportion of variance explained (using "Model Summary" table)

- **R**: refers to multiple correlation coefficient = absolute Pearson correlation coefficient between the 1 dependent and 1 independent variable, measuring the strength of association between the 2 variables (not usually of interest in linear regression analysis)
- **R<sup>2</sup>**: refers to proportion of variation explained by the model in the **sample** (e.g. 0.129 means that 12.9% of the variance in the dependent variable can be explained by the independent variable)
- **Adjusted R<sup>2</sup>**: refers an estimate of the proportion of variation explained by the model in the **population** (corrects for positive bias in the sample, hence would be lower than unadjusted). **Adjusted R square is also an estimate of Cohen's effect size**



## Multiple regression

Checking for multicollinearity - occurs when 2 or more independent variables are highly correlated with each other

- + Check that independent variables have correlations  $<0.7$
- + Check that tolerance value is  $>0.1$  or  $VIF < 10$



Statistical significance of the model (refer to "ANOVA" table):

- $p < 0.05$  indicates a statistically significant linear relationship

Interpreting coefficients - B refers to gradient, AKA increase in dependent variable per 1 unit of independent variable.

Predicting dependent variables - the predicted values are the expected and mean values + standard error + 95%CI, using general linear model -> univariate.

Note: CI refers to the mean predicted value. SPSS does not produce predicted value for an individual, although it is possible.



## Reporting results

A linear regression was run to understand the effect of average daily time spent watching TV on cholesterol concentration. To assess linearity a scatterplot of cholesterol concentration against average daily time spent watching TV with superimposed regression line was plotted. Visual inspection of these two plots indicated a linear relationship between the variables. There was homoscedasticity and normality of the residuals. One participant was one outlier with a cholesterol concentration of 7.98 mmol/L. They were removed from the analysis due to not representing the target population.

The prediction equation was: cholesterol concentration =  $-0.94 + 0.03697 \times \text{time}$ . Average daily time spent watching TV statistically significantly predicted cholesterol concentration,  $F(1, 97) = 14.39$ ,  $p < .0005$ , accounting for 12.9% of the variation in cholesterol concentration with adjusted  $R^2 = 12.0\%$ , a medium size effect according to Cohen (1988). An extra minute of daily average time spent watching TV leads to a 0.037 (95% CI, 0.018 to 0.056) mmol/L increase in cholesterol concentration.

+ scatterplot