DATA ANALYSIS USING SPSS

Dr. Mark Williamson, PhD
(based on PDF of Andrew Garth, Sheffield Hallam University)

Purpose

- The intent of this presentation is to teach you to explore, analyze, and understand data
- The software used is SPSS (Statistical Package for the Social Sciences)
  - *commonly used in social sciences and health fields*
  - *as opposed to other statistical software such as SAS or R, it requires little to no coding background*
- This presentation is heavily indebted to the work of Andrew Garth (Sheffield Hallam University) and his full document can be found at the link below: https://students.shu.ac.uk/lits/it/documents/pdf/analysing_data_using_spss.pdf
- All the data files used in this presentation can be found at the link below (download the SPSSDATA.zip): http://teaching.shu.ac.uk/hwb/ag/resources/resourceindex.html

Outline

- First, we will look at the Big Picture
- Next, we'll define our terms
- Then, we'll get set up for working in SPSS
- Only then will we get into the meat of things, which will focus on aspects of data analysis
    - *Descriptive Statistics and Graphs (Exploring our Data)*
    - *Inferential Statistics (Analyzing our Data, and Interpreting our Results)*

The Big Question.

- How should I analyze my data?

It depends on the nature of the data and
what questions you want to answer

To answer those questions, you need to explore your data.
and select the proper analysis

Big Picture Steps in Statistical Analysis

1. <u>Explore your data</u>
    1. Look at data
    2. Identify data
    3. Graph/Describe data
    4. Formulate Question (Hypothesis)

2. <u>Analyze your data</u>
    1. Set up hypothesis
    2. Check normality
    3. Select and run appropriate test

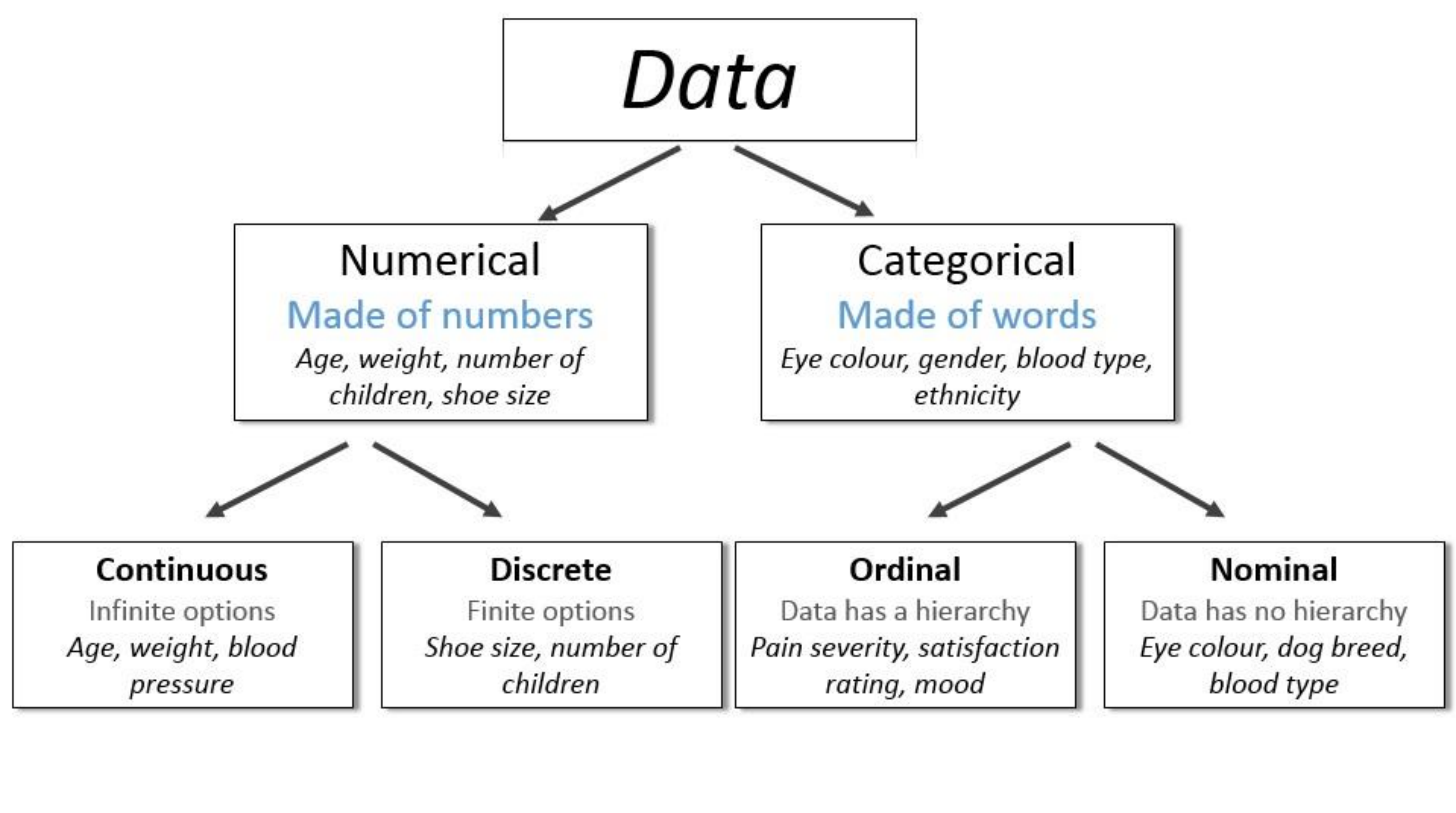
3. <u>Interpret your results</u>
    1. Find the Test Statistic, DF, and P-value
    2. Determine if significant
    3. State if null hypothesis rejected or not
    4. Write result
    5. Present appropriate plot

Before we can start analysis, we need to get set up on the basics

- Defining Terms
- Working in SPSS

Defining Terms

- There are two basic data types, each with two sub-types
  - *Numerical: expressed by numbers*
    - Discrete: numbers take on integer values only (number of children, number of siblings)
    - Continuous: numbers can take on decimal values (height, weight)
  - *Categorical: expressed by categories (also known as factors/groups)*
    - Nominal: no meaningful order between categories (gender, occupation)
    - Ordinal: categories can be put in meaningful order (agreement, level of pain, etc.)
- If data is not used for analysis, it can be labeled as a nuisance or bookkeeping variable



Data

Numerical

Made of numbers

Age, weight, number of children, shoe size

Categorical

Made of words

Eye colour, gender, blood type, ethnicity

Continuous

Infinite options

Age, weight, blood pressure

Discrete

Finite options

Shoe size, number of children

Ordinal

Data has a hierarchy

Pain severity, satisfaction rating, mood

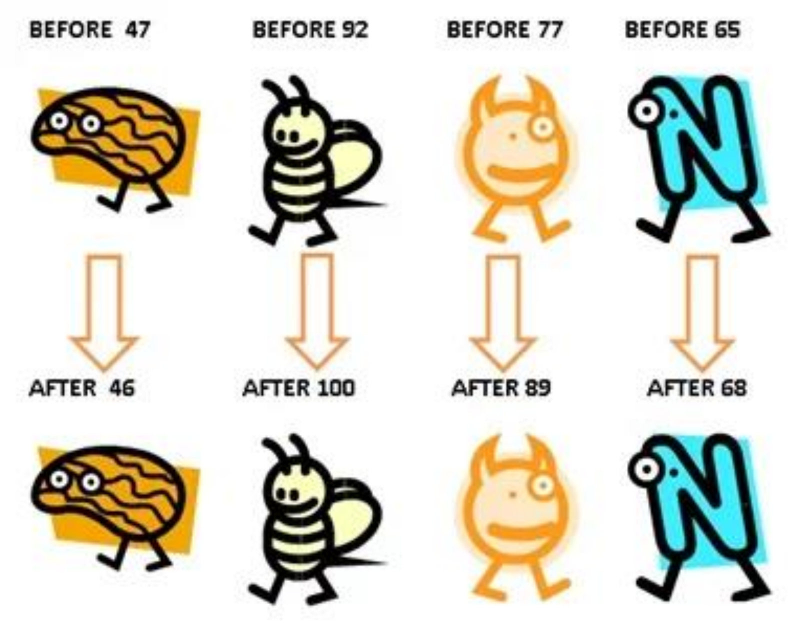
Nominal

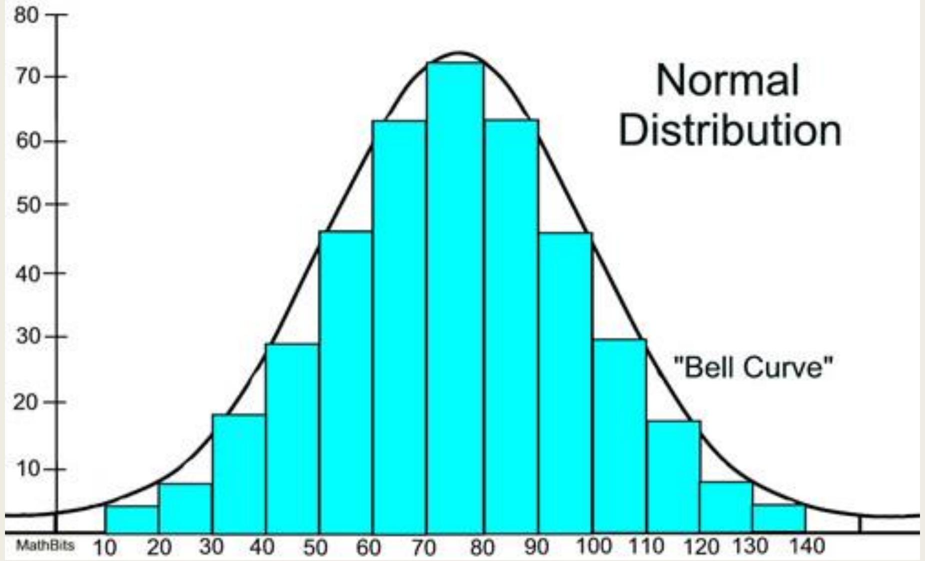
Data has no hierarchy

Eye colour, dog breed, blood type

Defining Terms 2

- Data can also be **paired or unpaired**
  - *Paired: categories are related to one another*
    - Often result of before and after situations (treatments/events)
    - Since each part of the pair is related to each other, this needs to be considered
    - If there are pairs of higher than 2, this is called repeated measures
  - *Unpaired: categories are not related to one another*
- Numerical data can be **parametric or non-parametric**
  - *Simply put, parametric data approximately fits a normal distribution*
    - Data are symmetric around a central point
    - "Bell curve"
    - Also known as normally distributed
  - *Data must be parametric (normally distributed) for many statistical tests*
    - If the data are not parametric, you cannot use the test results
    - If the data are non-parametric (does not fit a normal distribution), there are non-parametric tests for use, but they are weaker





Defining Terms 3

Recap

- Data can be:
  - *Numerical, categorical, or nuisance*
  - *Paired or unpaired*
  - *Parametric or non-parametric (usually must run a test to tell)*

Examples

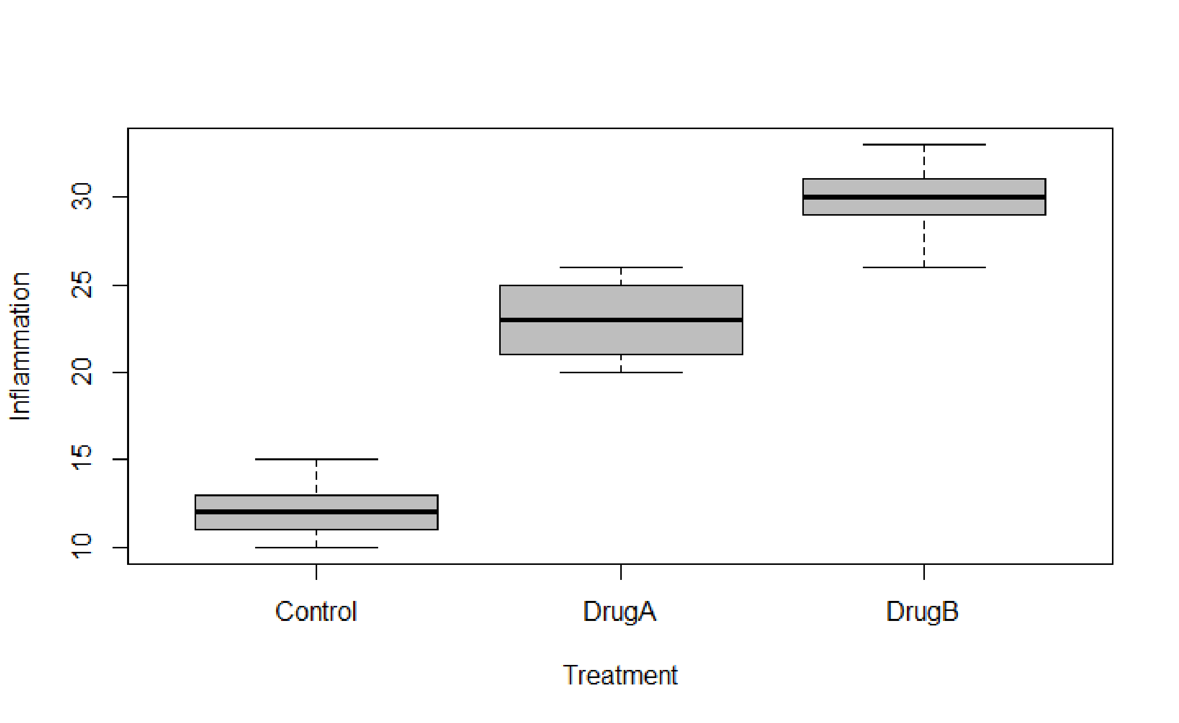
- **Numerical continuous:** height, weight, drug concentration
- **Numerical discrete:** number of siblings, number of drinks in a day, flower petal number
- **Categorical ordinal:** time of day (morning, noon, night), position (assistant professor, associate professor, department chair, dean)
- **Categorical nominal:** flower color, college major, drug treatment (A, B, C)
- **Nuisance:** sample number, subject name, date, id number
- **Paired:** Before, during, and after treatment; pre- and post-disaster

Defining Terms 4

- For statistical tests, we use two types of variables:
  - *Independent Variable- variation does not depend on another variable*
    - Usually denoted as X
    - Typically represents what the researcher set up (treatment, group, etc.)
  - *Dependent Variable – value depends on another variable (the independent one)*
    - Usually denoted as Y
    - Represents the variable that the researcher is interested in
    - Output or outcome
- Almost all statistical tests give three important pieces of information
  - *Test statistic*
    - Variable calculated from sample data and used in hypothesis test
    - Used to determine whether a test was significant or not
  - *Degrees of Freedom*
    - Number of values of quantities that can be assigned to a statistical distribution
    - Should be reported with test results
  - *P-value*
    - Measure of significance for the test statistic
    - Typically 0.05 is the cutoff value

Assessment 1

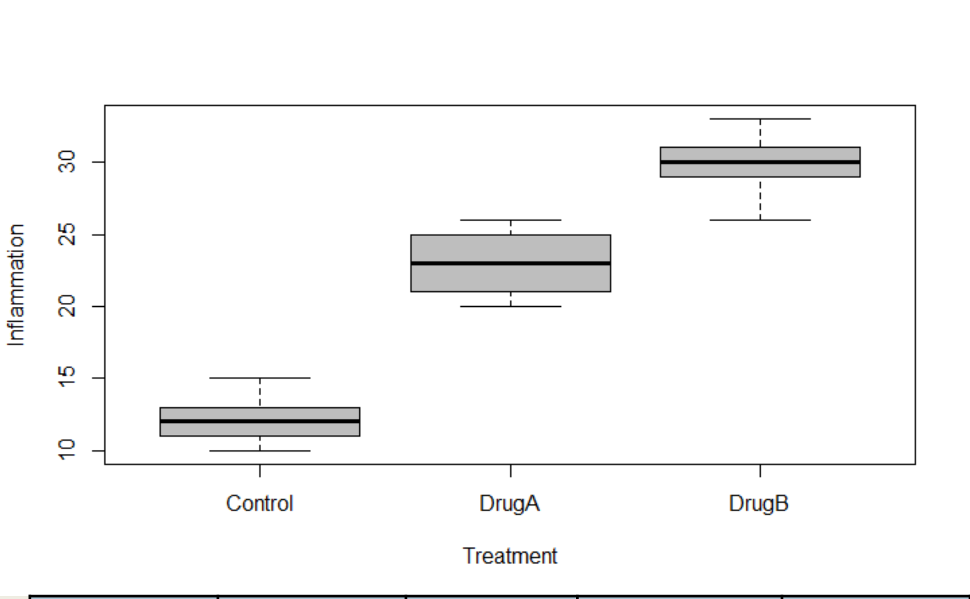
1. What types of data (categorical [nominal, ordinal], numerical [discrete, continuous] are each of the following examples
   a) Number of vaccine shots administered
   b) Highest level of education attained (high school, bachelors, masters, PhD)
   c) Country of origin
   d) Tumor size
2. In the boxplot graph to the right, which axis is the independent variable plotted on? Which axis is the dependent variable plotted on?
3. In the table to the right, label each of the columns as numerical, categorical, or nuisance



Sample #	User ID	Height	Treatment	Group
1	34AF001	162.3	1	A
2	67AF001	159.1	1	B
3	78AF001	160.2	1	C
4	22AF001	165.0	2	A
5	13AF001	157.5	2	B
6	49AF001	155.0	2	C

Assessment 1 Answers

1. What types of data (categorical [nominal, ordinal], numerical [discrete, continuous] are each of the following examples
   a) Number of vaccine shots administered (numerical discrete)
   b) Highest level of education attained (high school, bachelors, masters, PhD) (categorical ordinal)
   c) Country of origin (categorical nominal)
   d) Tumor diameter (numerical continuous)
2. In the graph to the right which axis is the independent variable plotted on? Which axis is the dependent variable plotted on? Independent on X-axis (Treatment), Dependent on Y-axis (Inflammation)
3. In the table to the right, label each of the columns as numerical, categorical, or nuisance (nuisance, nuisance, numerical, categorical, categorical)



Sample #	User ID	Height	Treatment	Group
1	34AF001	162.3	1	A
2	67AF001	159.1	1	B
3	78AF001	160.2	1	C
4	22AF001	165.0	2	A
5	13AF001	157.5	2	B
6	49AF001	155.0	2	C

Starting in SPSS : Access

- You can get access to SPSS using the CitrixWorkspaceApp for UND
- Some UND computers also have it downloaded
- If all else fails, you can try a free trial (https://www.ibm.com/account/reg/us-en/signup?formid=urx-19774)

- From here on out, I will be using the following formats

1. White boxes with green border are instructions in SPSS.
2. These will guide you through how to do the exploration/analysis I show yourself.

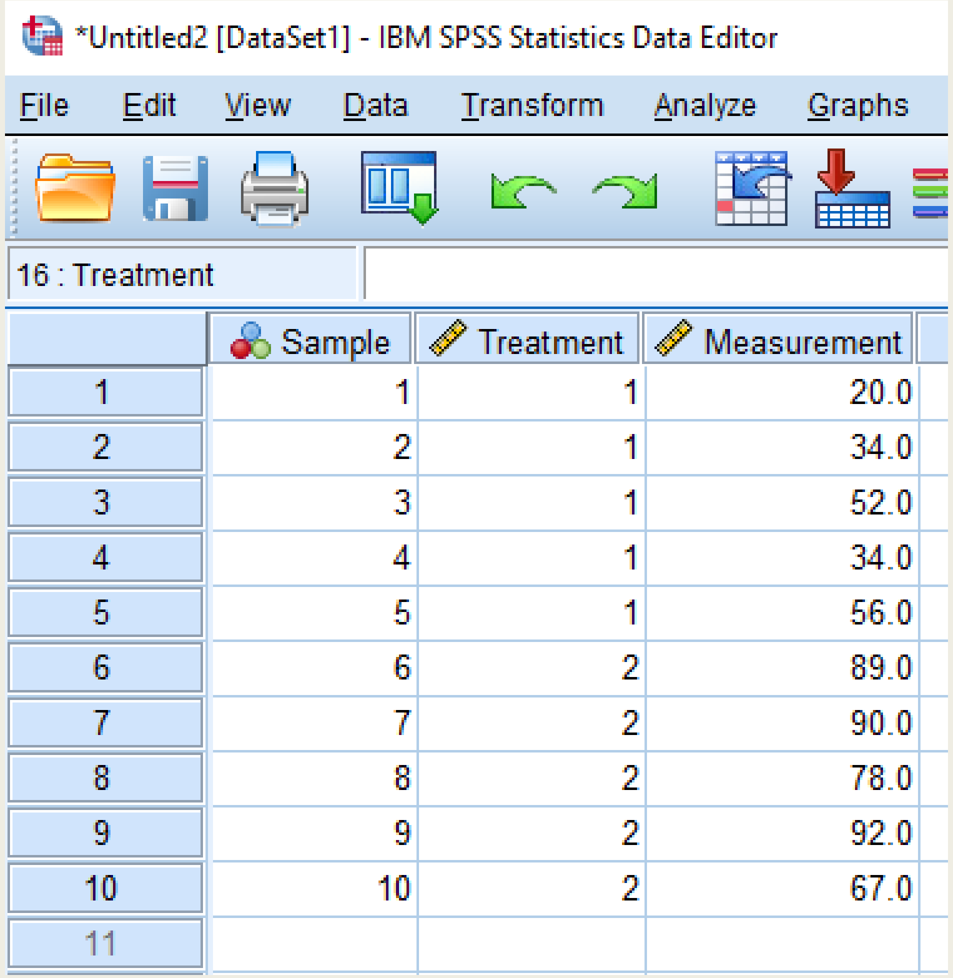
White boxes with purple borders are summaries

1. Orange boxes with red border are general step outlines

Blue boxes are reminders

Starting in SPSS: Data Format

- Specifics of format depends on the kind of data
- Principles that apply in most situations
  1. *Each case goes in its own row*
  2. *Categorical variables are best represented by numbers (even though they are not): can be labeled with Variable Labels option*
  3. *Variable names for the columns are limited in length, so again can be labeled with Variable Labels option*
  4. *Multiple groups of subjects should still be set up with each case having its own row: create a new variable column and give it the group label*



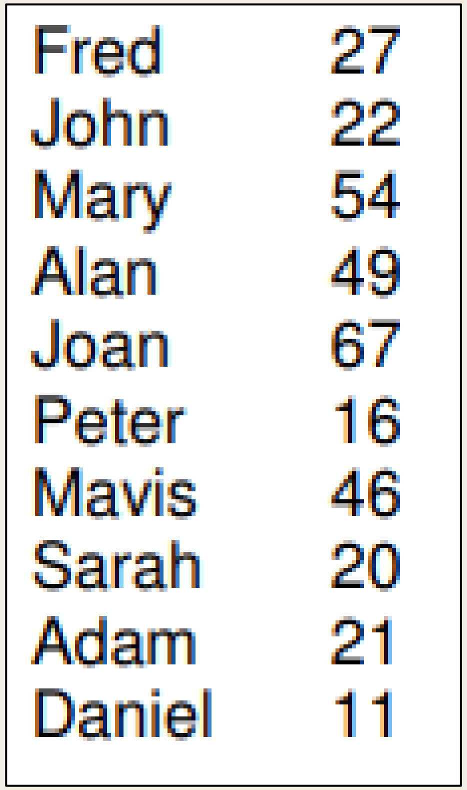
16 : Treatment

	Sample	Treatment	Measurement
1	1	1	20.0
2	2	1	34.0
3	3	1	52.0
4	4	1	34.0
5	5	1	56.0
6	6	2	89.0
7	7	2	90.0
8	8	2	78.0
9	9	2	92.0
10	10	2	67.0
11			

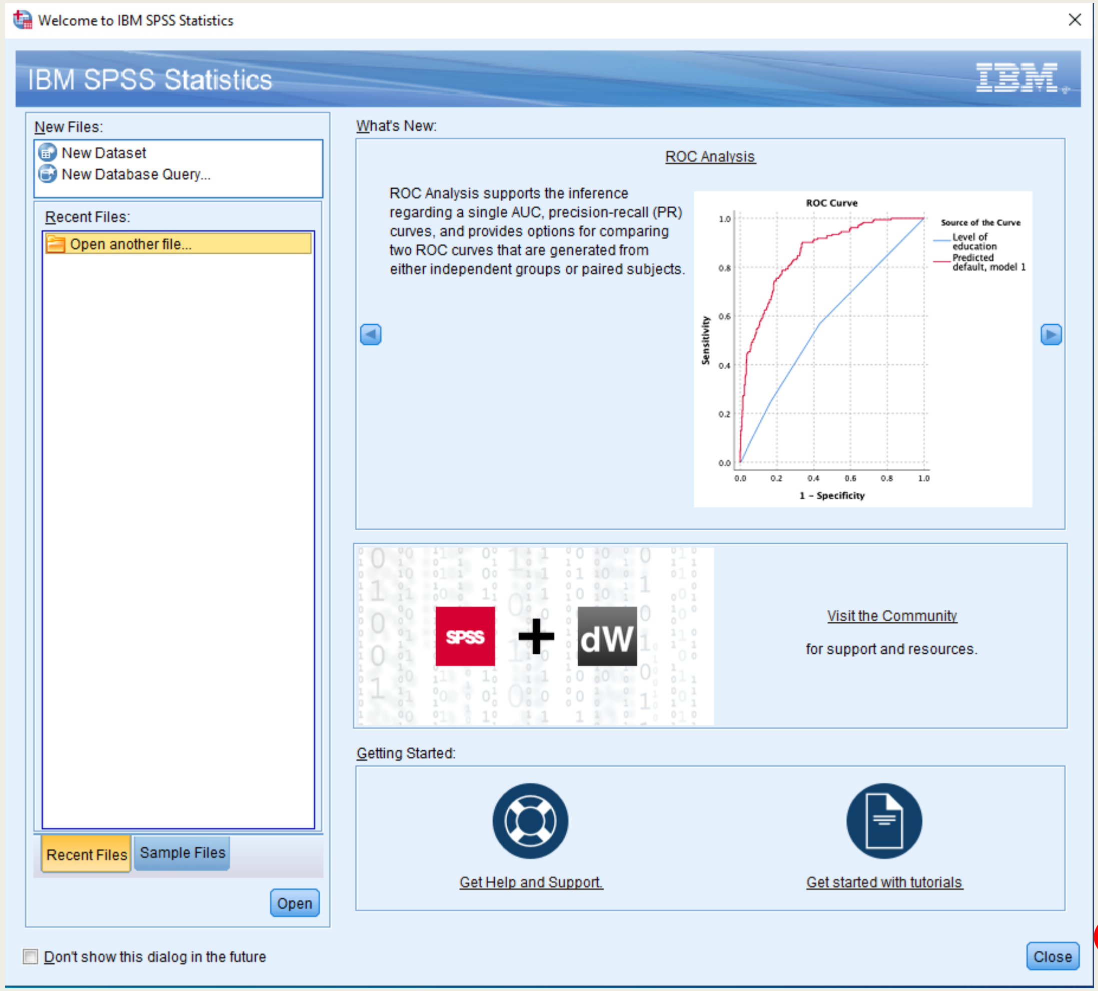
Starting in SPSS: Entering data

- There are two ways to enter data into SPSS
  - *Manually (entering the data by hand)*
  - *Loading in a file (data is saved in some form and can be opened in SPSS)*
- Let's try manual first
- You can look at the data in two ways
  - ***Variable View***
  - ***Data View***
- SPSS gives a lot of information, most which you don't need
  - *Ignore what you don't need*

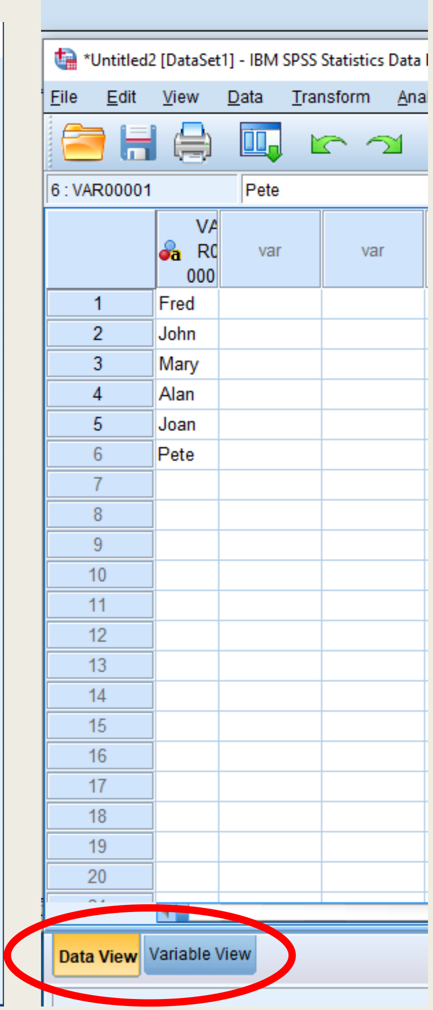
1. Start SPSS from wherever you have it
2. Double click **New Dataset** at the top left
3. In the box on the right there are 10 people's names, type them into the first column
4. You may notice a problem when you get to Peter.
   1. *Peter has 5 letters in his name, unfortunately SPSS has assumed all the cases are similar to the first one and Peter has become Pete.*
   2. *We can alter this by switching to the Variable View (click the tab at the bottom of the SPSS window). You should see a row of information about variable one (var0001), which is where we are storing these names.*
   3. *Change the Width from 4 to 12.*
   4. *Go back to the Data View and type in Peter again.*
   5. *Finish typing the names.*
5. Go back to the Variable View and change the column name (variable) to person rather than var00001.
6. Do the same for var00002, replacing it with the name 'age'.



Fred	27
John	22
Mary	54
Alan	49
Joan	67
Peter	16
Mavis	46
Sarah	20
Adam	21
Daniel	11

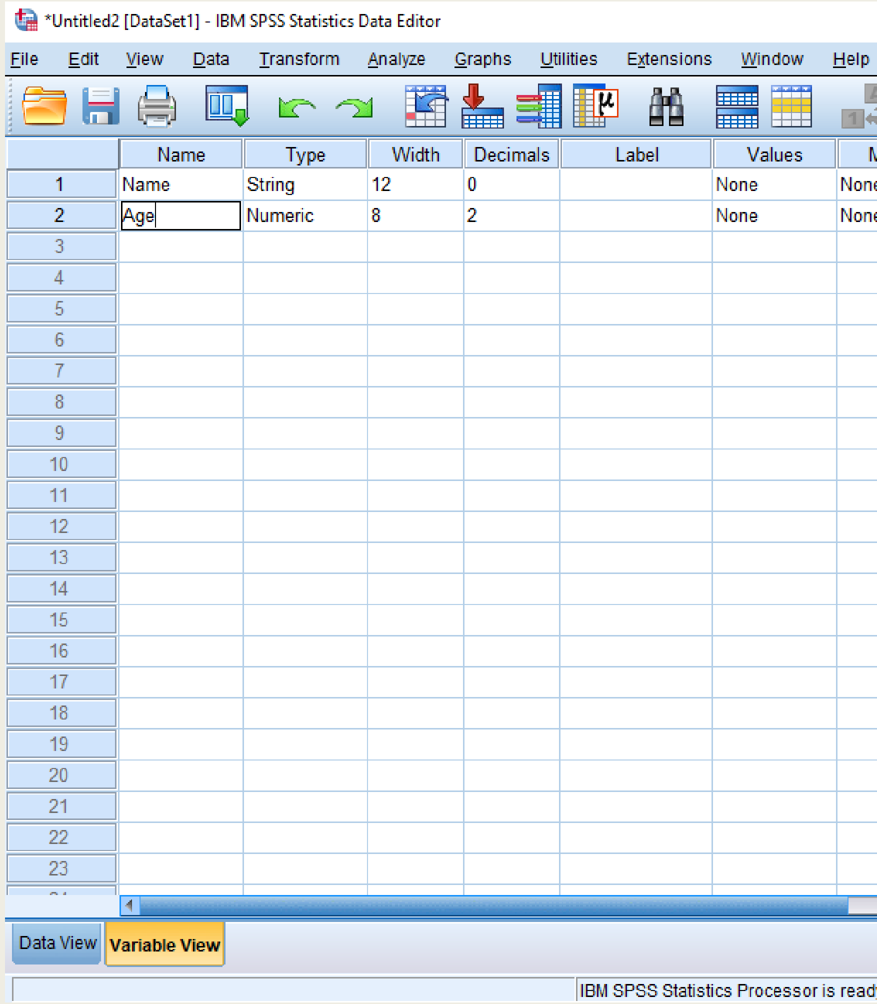


ROC Analysis supports the inference regarding a single AUC, precision-recall (PR) curves, and provides options for comparing two ROC curves that are generated from either independent groups or paired subjects.



6 : VAR00001 Pete

	VAR000	var	var
1	Fred		
2	John		
3	Mary		
4	Alan		
5	Joan		
6	Pete		
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			



	Name	Type	Width	Decimals	Label	Values	M
1	Name	String	12	0		None	None
2	Age	Numeric	8	2		None	None
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							
22							
23							

IBM SPSS Statistics Processor is read

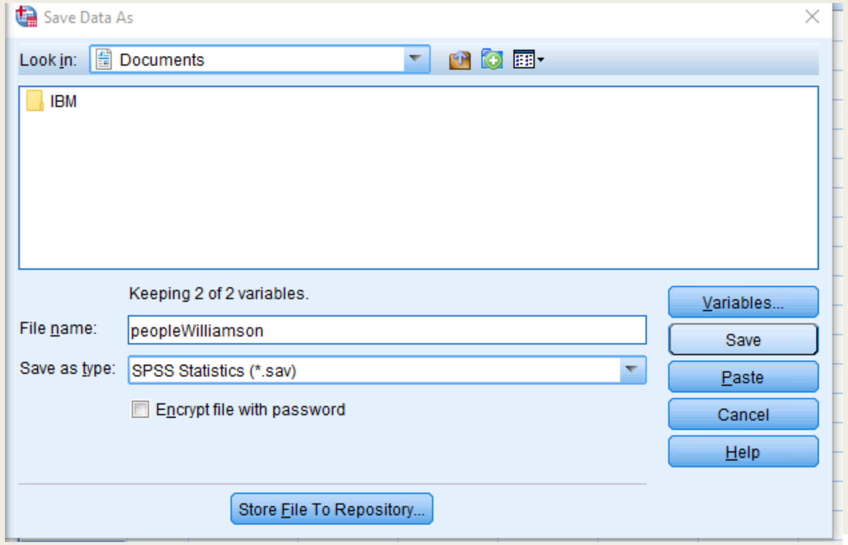
Starting in SPSS: Saving data

- Graphs and analyses will not be saved unless you save them specially
- Save often
- It is good practice to have multiple copies of data (especially when working on original data)

Reminder: the data needed for the tasks to follow are at:

https://teaching.shu.ac.uk/hwb/ag/resources/resourceindex.html

1. To save the names and ages from the previous slide, choose **Save** from the **File** menu. Call it **people** and put your name at the end of the word (ex. **peopleAnderson**).
2. You can save anywhere you want by using the **Look in**: and selecting the appropriate location
3. To save graphs or analyses, we need to do an analysis first
   1. *Click on the Analyze menu and choose Descriptive Statistics, then Descriptives.*
   2. *The button between the two windows let you choose the variables to be analyzed, in our case the choice is simple, just click the center button to move the age variable over to the right then click OK.*
   3. *SPSS should display the results in a separate window, you will see this appear in front of the Data Editor and a new button will appear on the Windows task bar at the bottom of your screen. The new window has a title, have a look in its title bar at the top of its window.*
   4. *Look at the output. If you want to save results like this, you have to save it separately.*



Save Data As

Look in: Documents

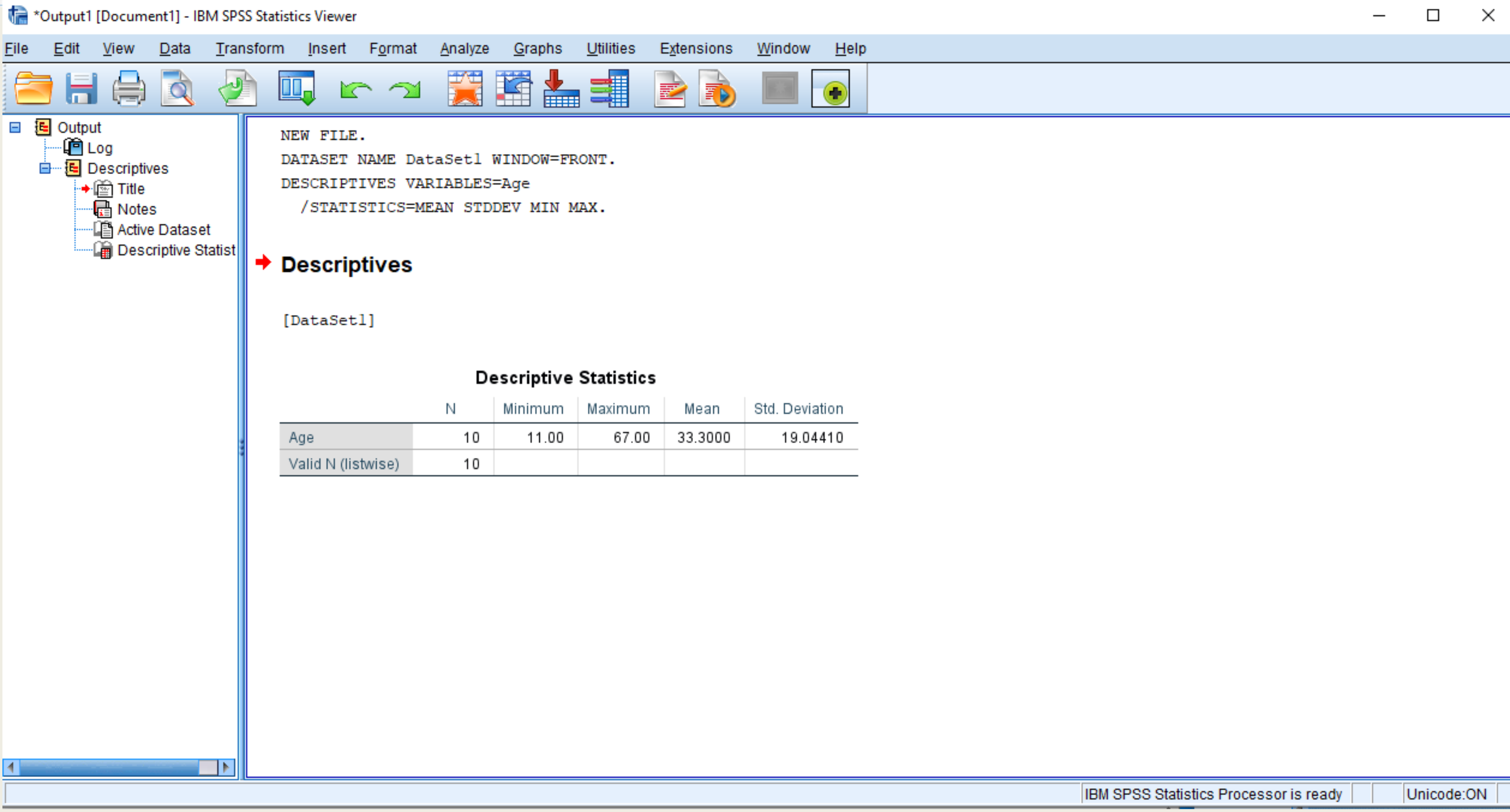
IBM

Keeping 2 of 2 variables.

File name: peopleWilliamson

Save as type: SPSS Statistics (*.sav)

□ Encrypt file with password



*Output1 [Document1] - IBM SPSS Statistics Viewer

```
NEW FILE.
DATASET NAME DataSet1 WINDOW=FRONT.
DESCRIPTIVES VARIABLES=Age
  /STATISTICS=MEAN STDDEV MIN MAX.
```

Descriptives

```
[DataSet1]
```

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age	10	11.00	67.00	33.3000	19.04410
Valid N (listwise)	10				

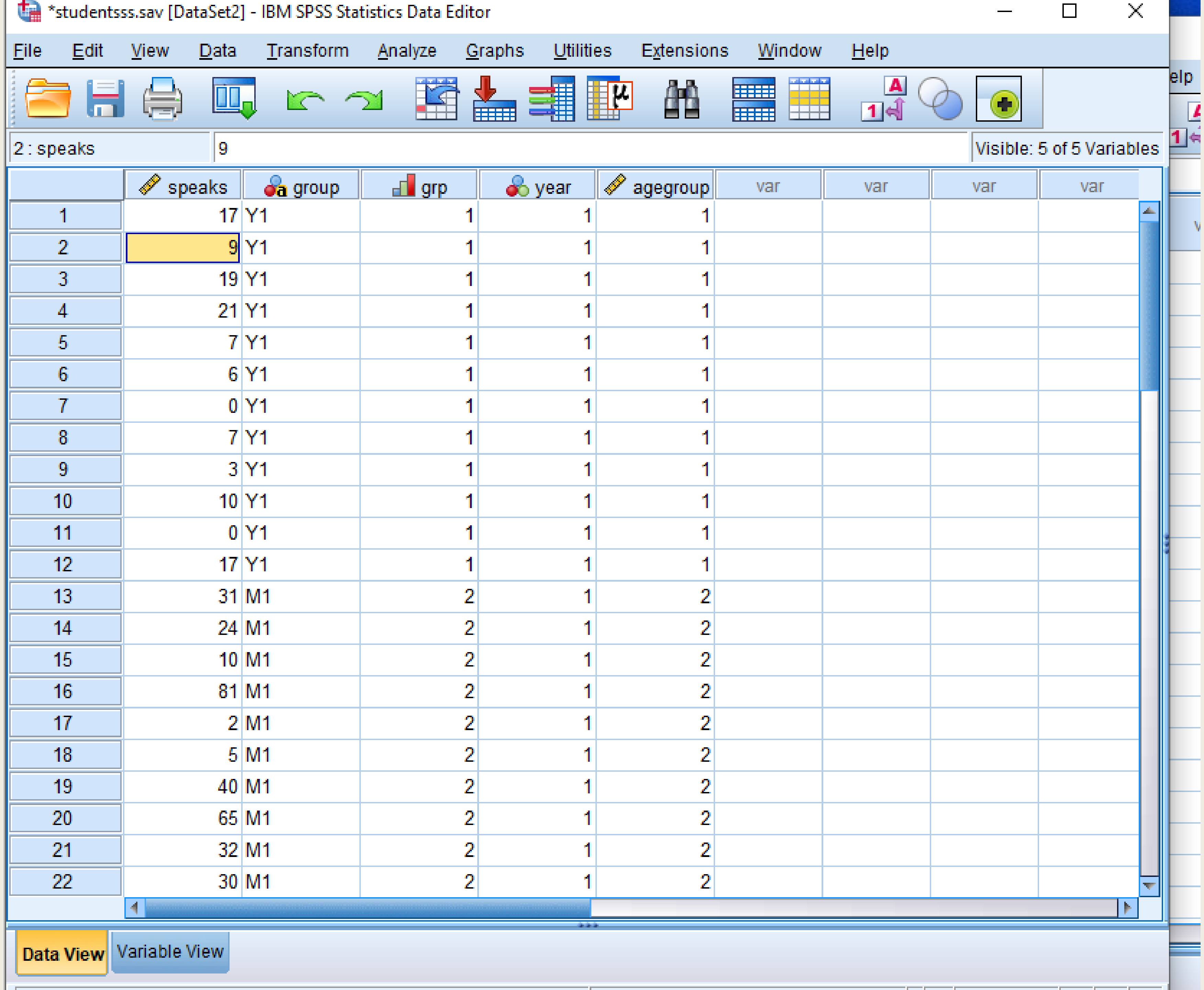
IBM SPSS Statistics Processor is ready

Unicode:ON

Starting in SPSS: Looking at data

- Seeing what data looks like is the first step to data analysis
- It gives a broad-overview in what is going on
- Again, each row is a different sample, while the columns show the value of different variables for that sample
- Looking at the data tells you a lot of big-picture things
  - *How many samples there are*
  - *How many variables there are*
  - *The types of variables and their values*
  - *If there is any missing data*
- **We will examine some data collected by an Occupational Therapy student, looking at how age affected OT students' participation in discussion in class.**
- **She counted how many times each student contributed orally in a period totaling 12 hours of classes. The students were from the 1st and 2nd years of the course and were classed as young if under 21 and mature if 21 or over, making 4 groups altogether.**

1. Open up **Studentss** in SPSS
   1. *choose the File menu and select Open-> Data (will need to search for wherever you downloaded the sample files)*
2. Take a look at the data and answer the following questions.
   1. *What is each column telling you?*
   2. *Which group is which?*
   3. *How many students were in each group?*
   4. *Do older students contribute more frequently in class discussion?*



2 : speaks 9

Visible: 5 of 5 Variables

	speaks	group	grp	year	agegroup	var	var	var	var
1	17	Y1	1	1	1				
2	9	Y1	1	1	1				
3	19	Y1	1	1	1				
4	21	Y1	1	1	1				
5	7	Y1	1	1	1				
6	6	Y1	1	1	1				
7	0	Y1	1	1	1				
8	7	Y1	1	1	1				
9	3	Y1	1	1	1				
10	10	Y1	1	1	1				
11	0	Y1	1	1	1				
12	17	Y1	1	1	1				
13	31	M1	2	1	2				
14	24	M1	2	1	2				
15	10	M1	2	1	2				
16	81	M1	2	1	2				
17	2	M1	2	1	2				
18	5	M1	2	1	2				
19	40	M1	2	1	2				
20	65	M1	2	1	2				
21	32	M1	2	1	2				
22	30	M1	2	1	2				

Starting in SPSS: Exploring the Data

- When analyzing data, it is necessary to know what variable is what
- Dependent variable:
  - *depends on the factor*
  - *Is usually numerical*
  - *In our case, it is ‘speaks’*
- Independent variable (Factor):
  - *Is the groups that the different samples are grouped into*
  - *Is usually categorical*
  - *In our case, it is ‘group’*

1. Click on the Analyze menu->Descriptive Statistics->Explore.
2. Transfer the ***speaks*** variable to the Dependent list and the ***group*** variable to the Factor list and then click OK.
3. Take a look at the results.

Case Processing Summary

	group	Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
speaks	M1	11	100.0%	0	0.0%	11	100.0%
	M2	11	100.0%	0	0.0%	11	100.0%
	Y1	12	100.0%	0	0.0%	12	100.0%
	Y2	10	100.0%	0	0.0%	10	100.0%

Descriptives

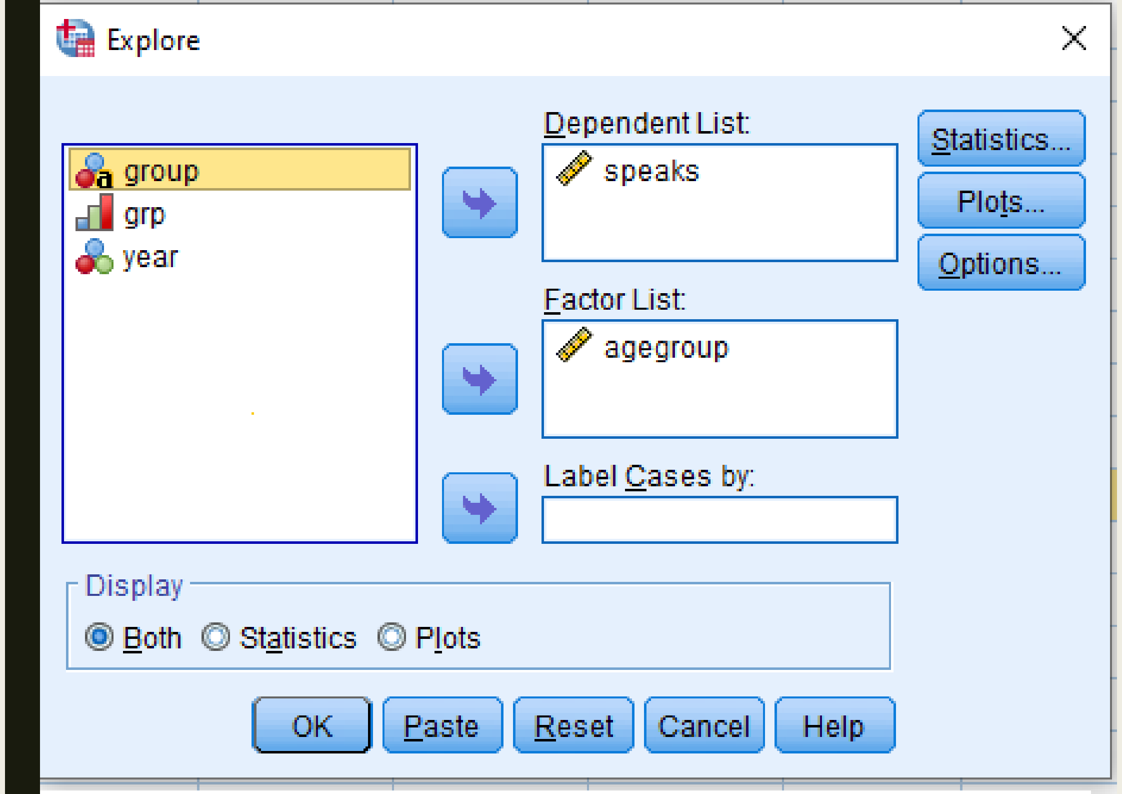
	group			Statistic	Std. Error
speaks	M1	Mean		33.09	7.303
		95% Confidence Interval for Mean	Lower Bound	16.82	
			Upper Bound	49.36	
		5% Trimmed Mean		32.16	
		Median		31.00	
		Variance		586.691	
		Std. Deviation		24.222	
		Minimum		2	
		Maximum		81	
		Range		79	
		Interquartile Range		34	
		Skewness		.677	.661
		Kurtosis		.185	1.279
	M2	Mean		46.91	10.964
		95% Confidence Interval for Mean	Lower Bound	22.48	
			Upper Bound	71.34	
		5% Trimmed Mean		42.84	
		Median		34.00	
		Variance		1322.291	
		Std. Deviation		36.363	
		Minimum		19	
		Maximum		148	
		Range		129	
		Interquartile Range		28	
		Skewness		2.475	.661
		Kurtosis		6.939	1.279
	Y1	Mean		9.67	2.101
		95% Confidence Interval for Mean	Lower Bound	5.04	
			Upper Bound	14.29	
		5% Trimmed Mean		9.57	
		Median		8.00	
		Variance		52.970	
		Std. Deviation		7.278	
		Minimum		0	
		Maximum		21	
		Range		21	
		Interquartile Range		13	
		Skewness		.245	.637
		Kurtosis		-1.248	1.232
	Y2	Mean		16.50	3.845
		95% Confidence Interval for Mean	Lower Bound	7.80	
			Upper Bound	25.20	
		5% Trimmed Mean		15.89	
		Median		12.00	
		Variance		147.833	
		Std. Deviation		12.159	
		Minimum		4	
		Maximum		40	
		Range		36	
		Interquartile Range		16	
		Skewness		1.292	.687
		Kurtosis		.542	1.334

Part A-3d

Using descriptive statistics

- It is hard to read out the various descriptive statistics from graphs
- Instead, we can calculate them and spit out numbers in tables: such as medium, mean, interquartile range, and, standard deviation
- Measures of central tendency, or ‘average:
  - *Mean: all values are summed and divided by the number of values*
  - *Median: middle value*
  - *Mode: the most common value*
- Measures of spread:
  - *Interquartile range*
  - *Standard Deviation*

1. Go back to the **studdentsss** file
2. Got to Analyze menu, select Descriptive Statistics, then Explore. The dependent list refers to the quantity we are measuring, in this case, the number of times people speak. In the factor list we put the factor that we are investigating, in this case "agegroup".
3. From the output find the Mean and Median of each group. The mean and median are both forms of average, do they seem to agree?



Explore

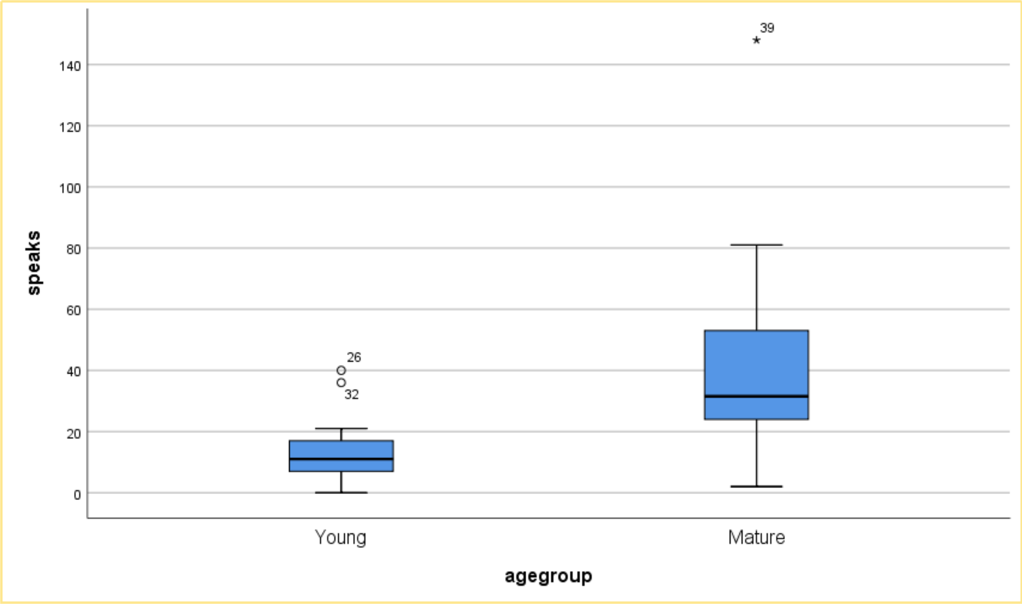
Dependent List: speaks

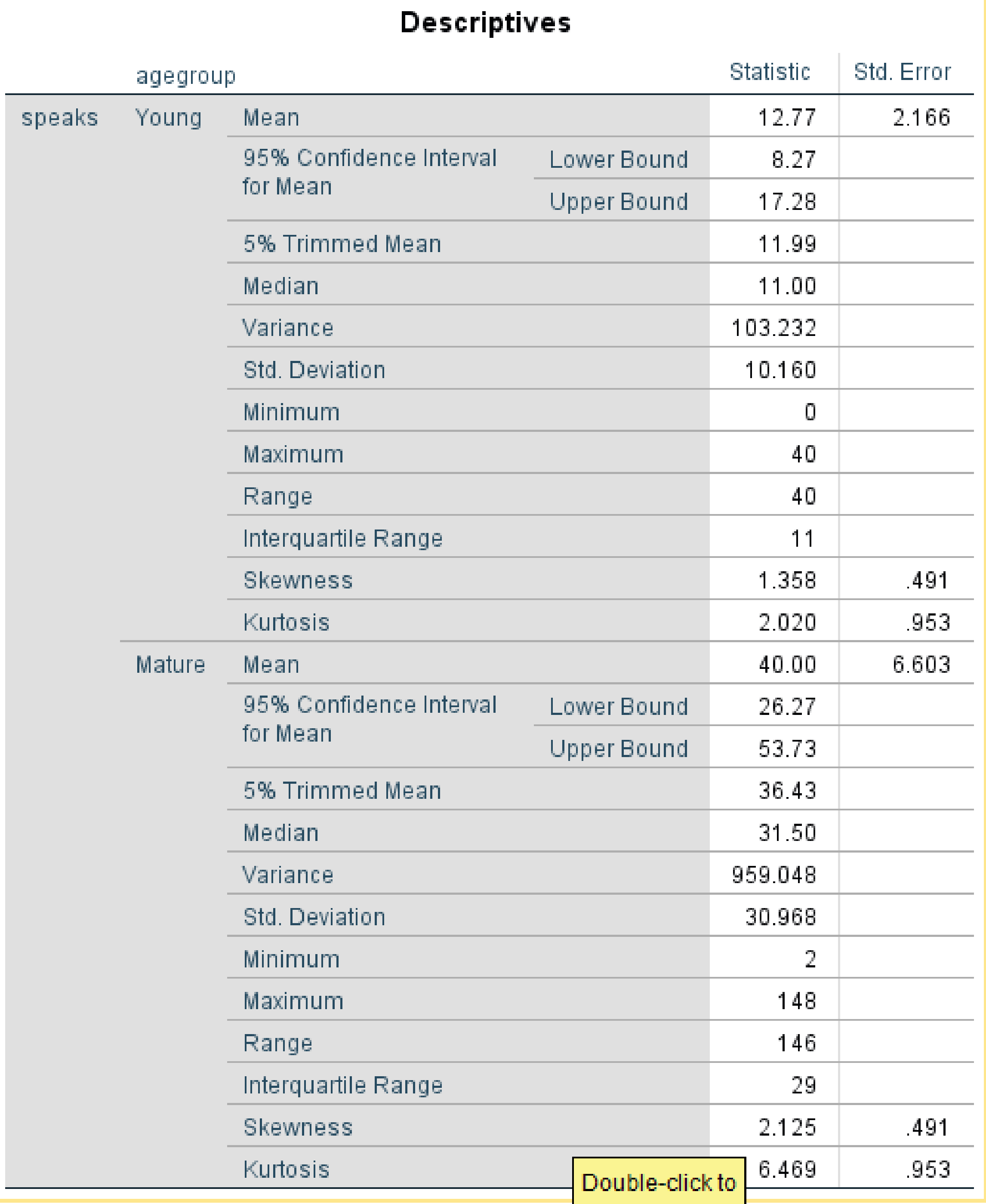
Factor List: agegroup

Label Cases by:

Display

● Both ○ Statistics ○ Plots





Descriptives

	agegroup			Statistic	Std. Error
speaks	Young	Mean		12.77	2.166
		95% Confidence Interval for Mean	Lower Bound	8.27	
			Upper Bound	17.28	
		5% Trimmed Mean		11.99	
		Median		11.00	
		Variance		103.232	
		Std. Deviation		10.160	
		Minimum		0	
		Maximum		40	
		Range		40	
		Interquartile Range		11	
		Skewness		1.358	.491
		Kurtosis		2.020	.953
	Mature	Mean		40.00	6.603
		95% Confidence Interval for Mean	Lower Bound	26.27	
			Upper Bound	53.73	
		5% Trimmed Mean		36.43	
		Median		31.50	
		Variance		959.048	
		Std. Deviation		30.968	
		Minimum		2	
		Maximum		148	
		Range		146	
		Interquartile Range		29	
		Skewness		2.125	.491
		Kurtosis		6.469	.953

Double-click to activate

Assessment 2

1. When formatting data in SPSS, should each sample be put in its own row?
2. Will SPSS automatically save results and graphs?
3. What is the mean, median, and mode of the dataset to the right?

Number of Siblings
2
1
1
2
3
5
10
2
4
1

Assessment 2 Answers

1. When formatting data in SPSS, should each sample be put in its own row? YES
2. Will SPSS automatically save results and graphs? NO
3. What is the mean, median, and mode of the dataset to the right?

   3.33, 2, 2

Number of Siblings
2
1
1
2
3
5
10
2
4

Descriptive Statistics and Graphs (Exploring our Data)

- A large part of data analysis is exploring your data and understanding more about it, both by visually graphing it and generating statistics such as means
- This section will go over a variety of the basic approaches

Rules for Exploring Data

- Discipline
  - If you discipline yourself by doing each of these things each time you look at your data, you will develop the skill to intelligibility see the data
  - This will give you the freedom to analyze the data without struggling to comprehend even the most basic understanding of the data
  - Computers are fast but dumb, so they rely on you to supply the intelligence to make sure the results are useful
- Rules
  1. **Look at data:** open up the file and look at the raw data (or, if the data is too large, a subset)
  2. **Identify data:** for each column determine what type of data it is
     a) If it is numerical, is it continuous or discrete?
     b) If it is categorical, how many categories and is it nominal or ordinal?
     c) Or if it is not useful, call it a nuisance variable?
     d) Are their any variables that may be paired?
  3. **Graph/Describe data:** for each variable or set of variables (comparison), graph and run descriptive statistics
  4. **Write Research Question:** Write out in a clear sentence what each comparison is trying to test

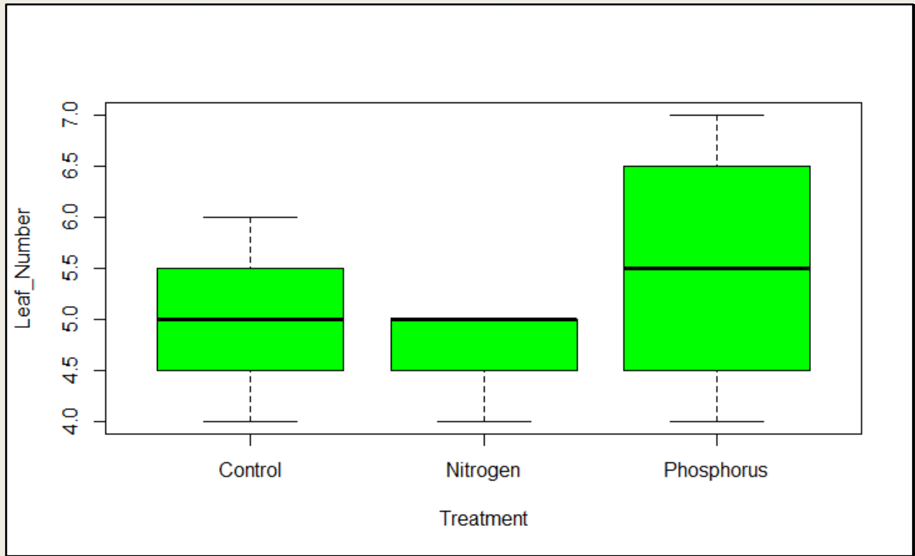
Rules: Example with Plant data

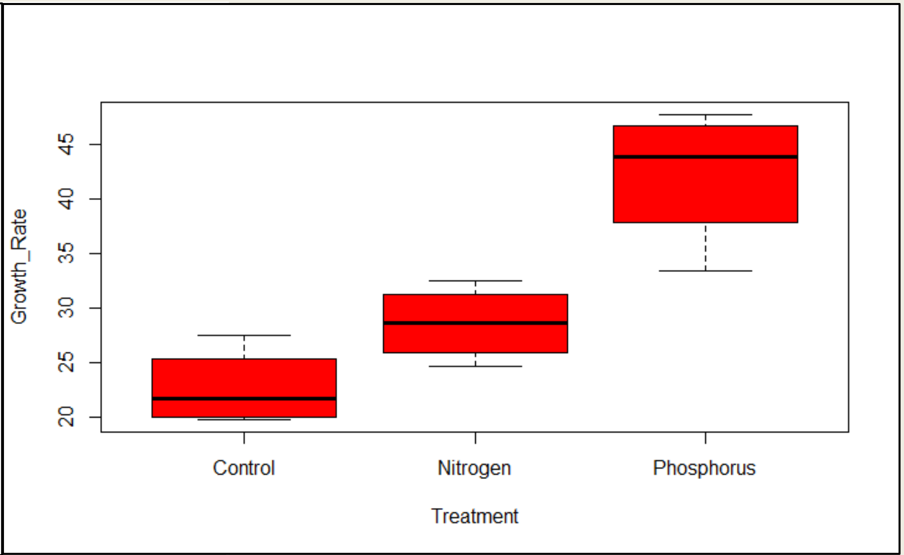
1. Look at the Data

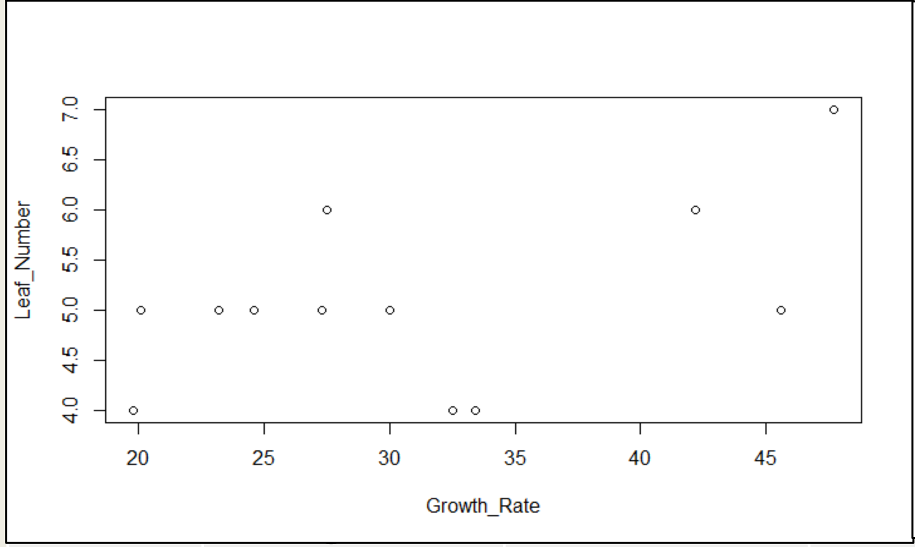
2. Describe Each Variable

3. Graph/Stats each Comparison

4. Write Research Question







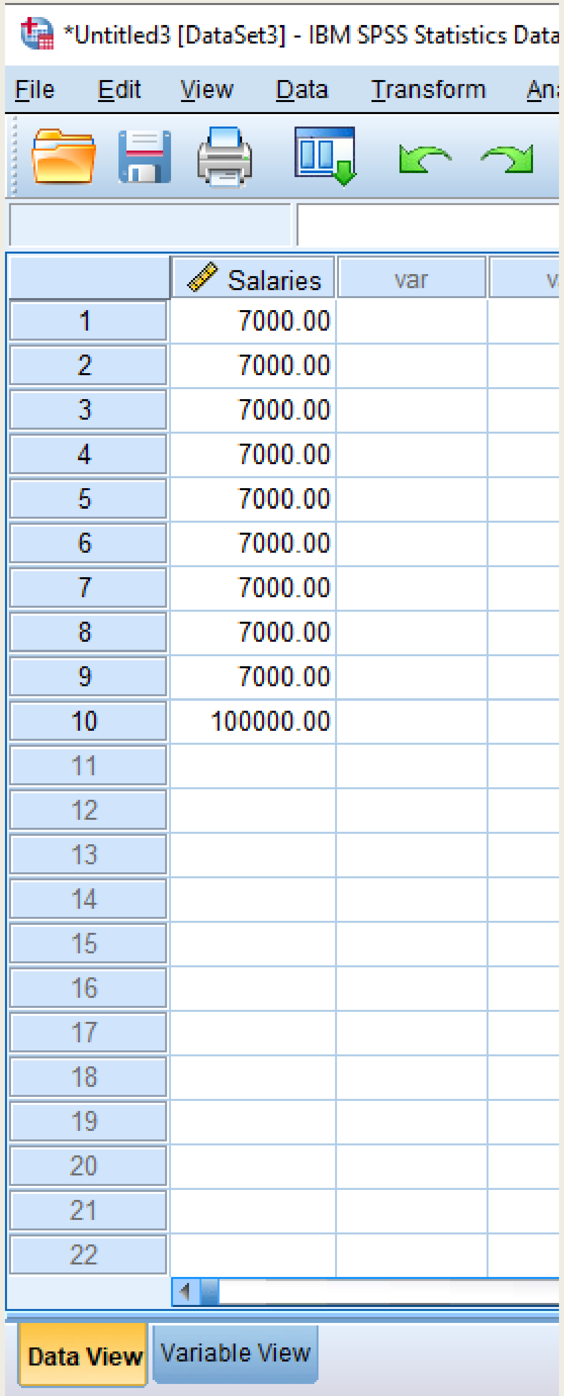
- Is there a difference in [plant] growth rate across nutrient treatments?
- Is there a difference in [plant] leaf number across nutrient treatments?
- Is Leaf number related to growth rate?

Part A-3

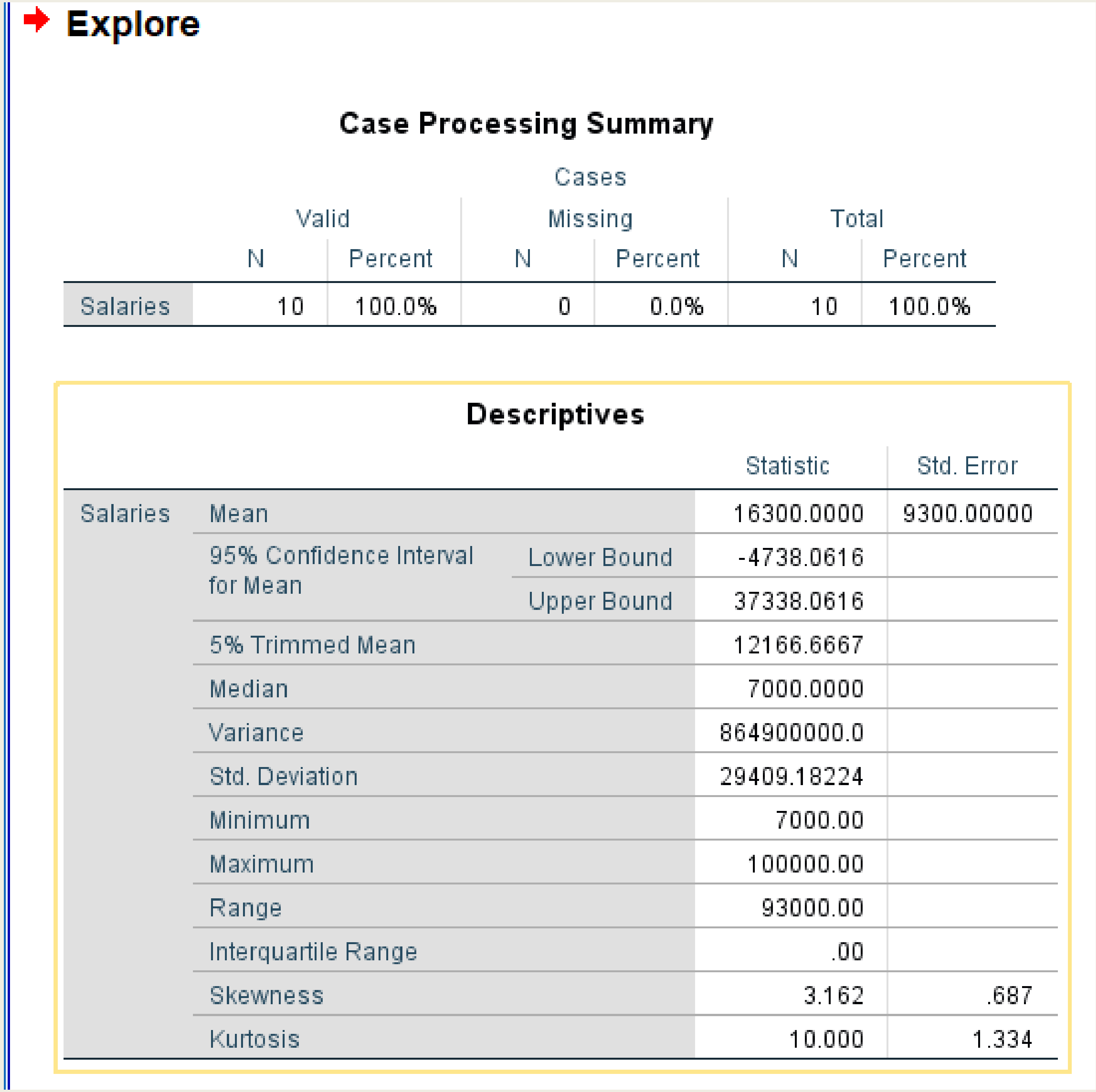
Mean vs. median

Summary: Mean vs. Median - both are types of average. The mean is based on all the data values, however because of this it is prone to being unduly affected by outliers in the data, most noticeably when the sample is small. The median however is largely unaffected by one or two extreme outliers, even in small samples, it is simply the middle value.

1. Open a new file. (File->New->Data) We are going to type in a few figures.
2. Put the following numbers in the first column (7000, 7000, 7000, 7000, 7000, 7000, 7000, 7000, 7000, 100000).
3. Give the column the title 'Salaries' (you need to click onto the Variable View for this
4. Back in Data View you may want to alter the column width by dragging the vertical bar next to the variable name.
5. The numbers represent the annual salaries of the 10 permanent employees of a small (mythical) private clinic. Which is the director's?
6. Run Descriptive Statistics->Explore to find the mean and the median. If you were the union negotiator for the employees of the clinic which of the two average salaries would you quote to the press? If you were the owner of the clinic which might you quote?
7. Find the inter-quartile range and the standard deviation. Can you sketch what the Boxplot would look like? Create the Boxplot on SPSS if you like.



	Salaries	var
1	7000.00	
2	7000.00	
3	7000.00	
4	7000.00	
5	7000.00	
6	7000.00	
7	7000.00	
8	7000.00	
9	7000.00	
10	100000.00	
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		
21		
22		



Explore

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Salaries	10	100.0%	0	0.0%	10	100.0%

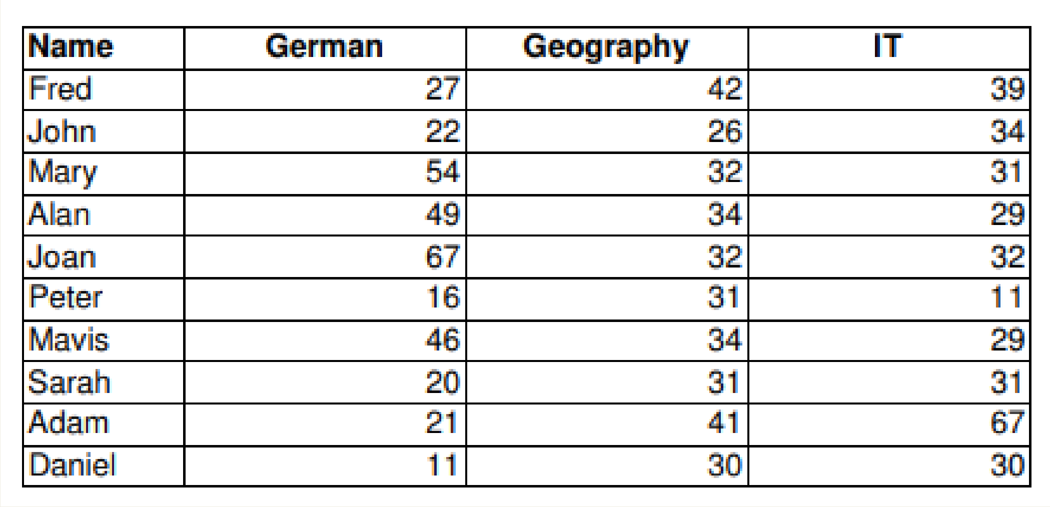
Descriptives

			Statistic	Std. Error
Salaries	Mean		16300.0000	9300.00000
	95% Confidence Interval for Mean	Lower Bound	-4738.0616	
		Upper Bound	37338.0616	
	5% Trimmed Mean		12166.6667	
	Median		7000.0000	
	Variance		864900000.0	
	Std. Deviation		29409.18224	
	Minimum		7000.00	
	Maximum		100000.00	
	Range		93000.00	
	Interquartile Range		.00	
	Skewness		3.162	.687
	Kurtosis		10.000	1.334

Part A-4

Standard deviation

- What is the Standard Deviation (S.D.) really measuring?
- What can it tell us about our data?
- Let’s take a look at some data
  - *The table below shows the German, Geography and IT results of a group of ten*

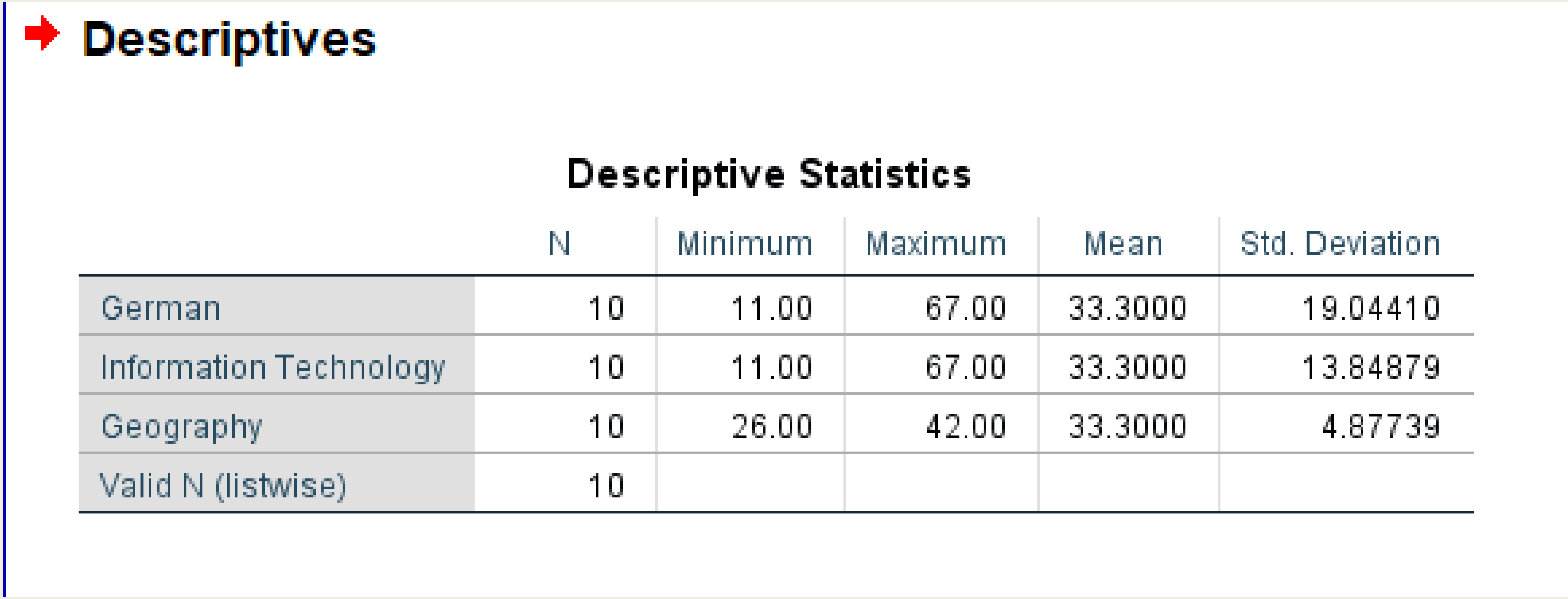


Name	German	Geography	IT
Fred	27	42	39
John	22	26	34
Mary	54	32	31
Alan	49	34	29
Joan	67	32	32
Peter	16	31	11
Mavis	46	34	29
Sarah	20	31	31
Adam	21	41	67
Daniel	11	30	30

1. Open the file **std dev example** in SPSS
2. Use the Descriptive Statistics->Descriptives to fill out the table below

	German	Geography	IT
MEAN			
MAX			
MIN			

3. Which set(s) of figures has the largest range?
4. Which set(s) of figures has the largest number in it?
5. Which set(s) of figures contains the smallest number?
6. Which set of figures has the largest minimum?



Descriptives

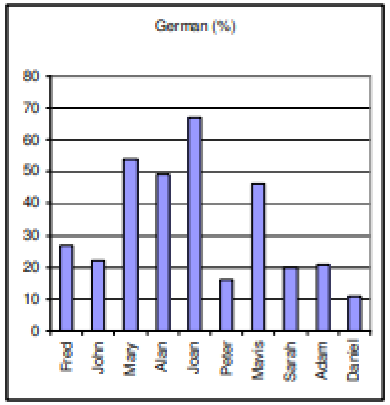
Descriptive Statistics

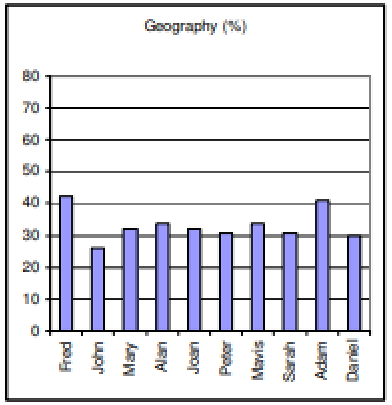
	N	Minimum	Maximum	Mean	Std. Deviation
German	10	11.00	67.00	33.3000	19.04410
Information Technology	10	11.00	67.00	33.3000	13.84879
Geography	10	26.00	42.00	33.3000	4.87739
Valid N (listwise)	10				

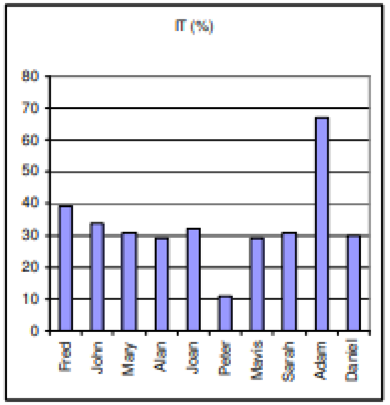
Part A-4

Standard deviation 2

- Given the figures for mean, maximum and minimum it is hard to differentiate between the German and IT figures, the mean, (arithmetic mean) of the figures is the numbers all added together then divided by the number of numbers.
- However it gives no indication of the distribution of the marks within the sets of figures. To do this we could graph the three sets of figures and see if that helps us (later we will create bar charts, for now just look at these).





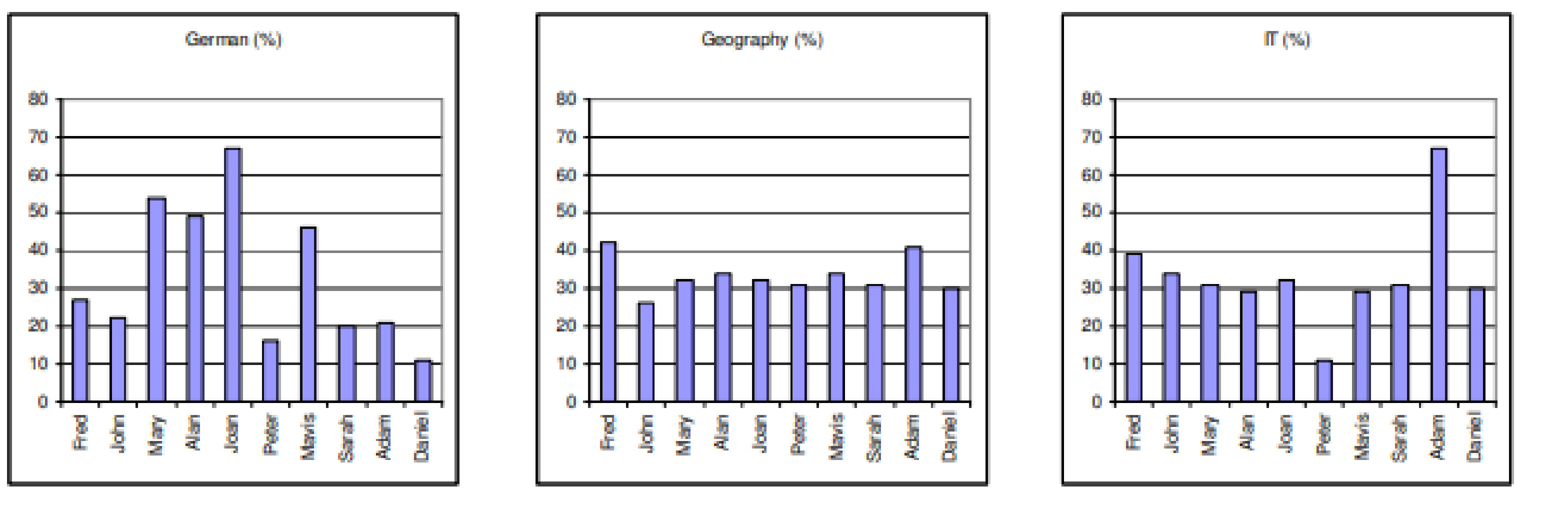


- Look at the three graphs above. Which two do you think are most similar?
- *Possibly Geography and IT but it is rather subjective. They do seem to have less variation in the values than the German results.*

Standard deviation 3

- Question: How can we asses in a fair, unambiguous way, which of three has the least widely deviating set of numbers?

- Answer: Use the Standard Deviation.



- The standard deviation of a set of numbers is a measure of how widely values are dispersed from the mean value. It can be calculated manually, or SPSS can calculate it for you.

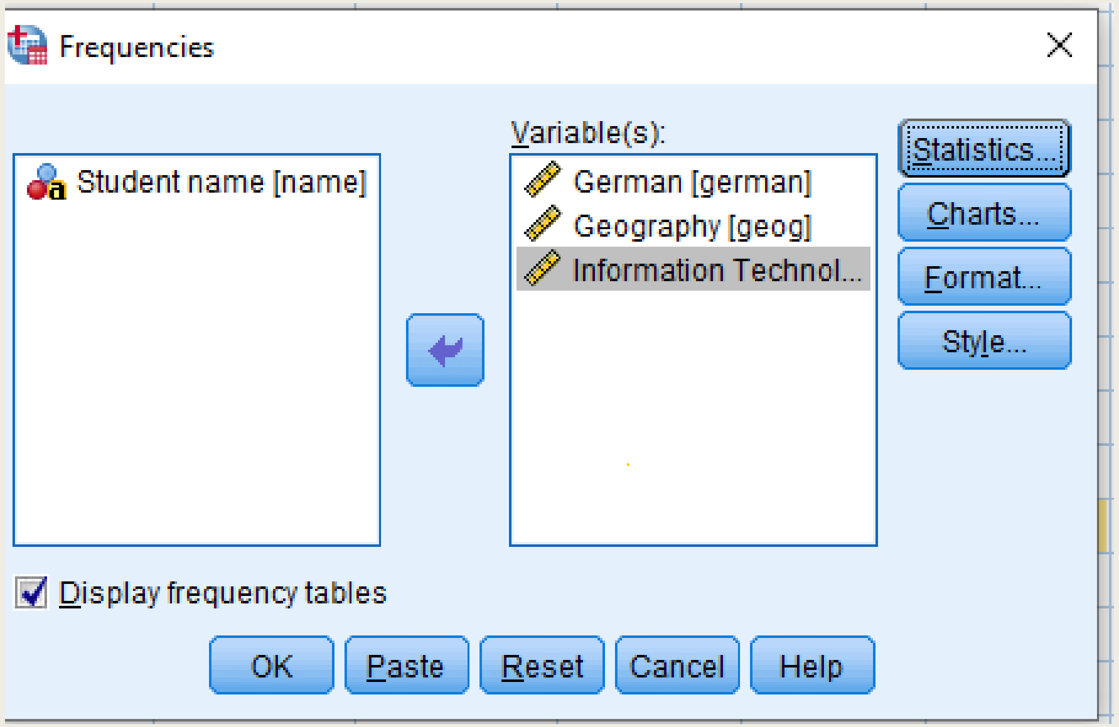
Part A-4

Standard deviation 4

- Let’s work out the standard deviation of the numbers in each column from the **std dev example**
  - *Higher Standard Deviation values indicate a greater spread of values*
  - *Lower Standard Deviation values indicate a tighter spread of values*

Summary: Range, IQR & SD are all measures of spread. Only the SD takes all the data values into account, however this leaves it open to problems similar to the mean, i.e. a tendency to be swayed inordinately by extreme values. The range is extremely sensitive to outliers, since it is based only on the smallest and largest values. The Inter Quartile Range is again based on only two values, the upper and lower quartiles, these are on each end of the middle half of the data, therefore less effected by extremes.

1. Use Descriptive Statistics then Frequencies from the Analyze menu.
2. Select the three variables (get German, Geography and Information Technology (IT) from the left into the right pane).
3. Click the “Statistics” button and select the Standard deviation as well as mean, maximum and minimum, then click “Continue”.
4. Before pressing OK on the Frequencies dialog box, uncheck the option to display frequency tables then click OK.
5. Compared the standard deviations.
   1. *Which set of figures, German, Geography or IT, is the least spread out?*
   2. *Of the two subjects with the same mean, and the same range, which varies least?*
   3. *Which of the three sets of figures, German, Geography or IT varies most?*

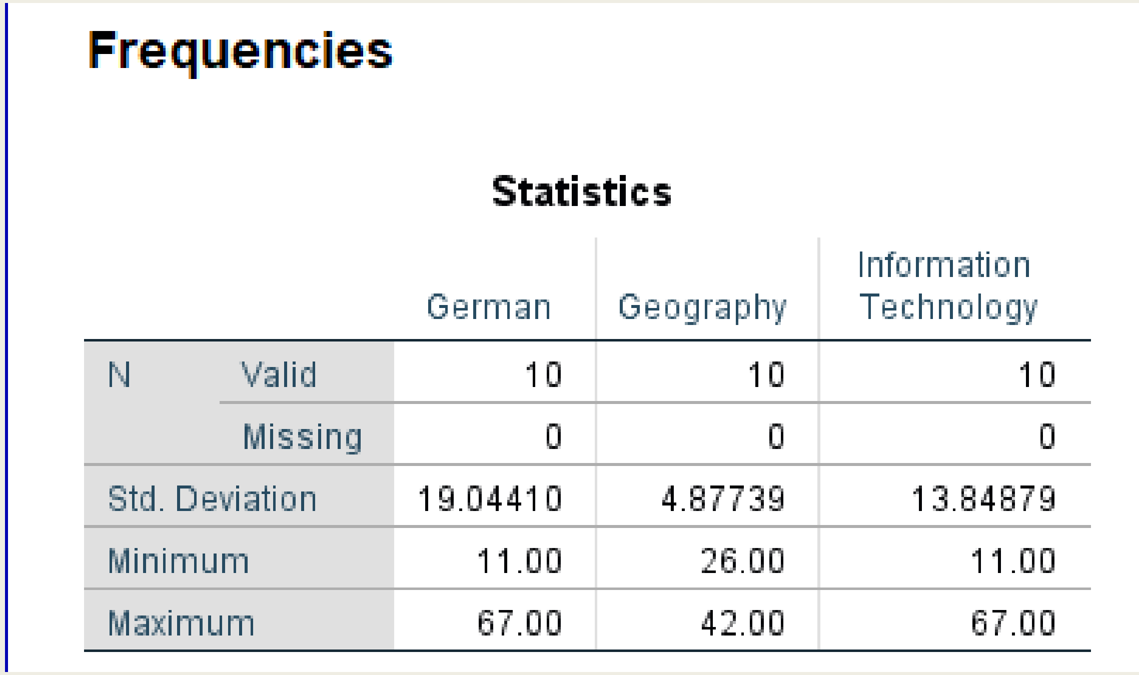


Frequencies

Student name [name]

Variable(s):
- German [german]
- Geography [geog]
- Information Technol...

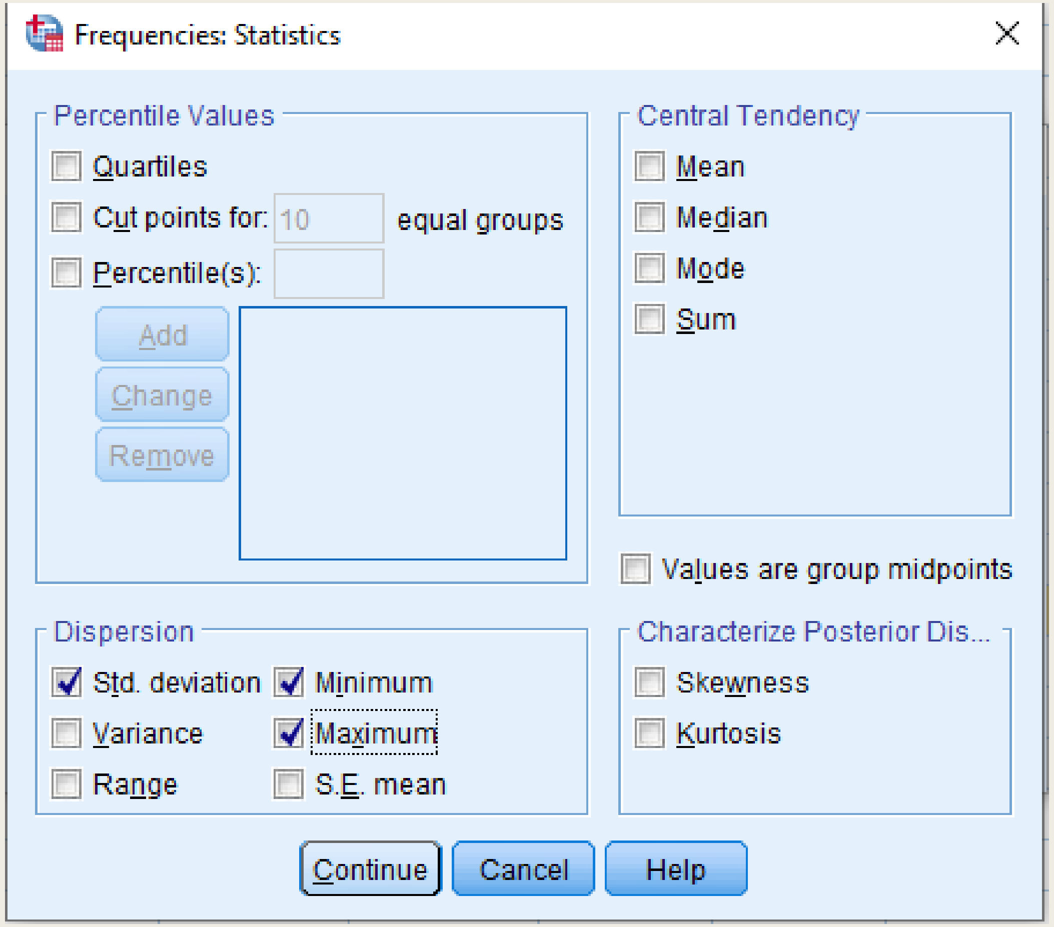
☑ Display frequency tables



Frequencies

Statistics

		German	Geography	Information Technology
N	Valid	10	10	10
	Missing	0	0	0
Std. Deviation		19.04410	4.87739	13.84879
Minimum		11.00	26.00	11.00
Maximum		67.00	42.00	67.00



Frequencies: Statistics

Percentile Values
- □ Quartiles
- □ Cut points for: 10 equal groups
- □ Percentile(s):

Central Tendency
- □ Mean
- □ Median
- □ Mode
- □ Sum

□ Values are group midpoints

Dispersion
- ☑ Std. deviation
- ☑ Minimum
- □ Variance
- ☑ Maximum
- □ Range
- □ S.E. mean

Characterize Posterior Dis...
- □ Skewness
- □ Kurtosis

Assessment 3

1. In the data to the right, which subject had the highest average score?
2. In the data to the right, which subject had the most variation in score? Which had the least?
3. What are the 4 rules for exploring data?

Exam Scores

Subject	N	Mean	Standard Deviation
Art	10	95	3.3
Spelling	10	70	5.8
Math	10	67	3.5
Science	10	84	12.3
Social Studies	10	89	2.1
Physical Education	10	98	1.2

Assessment 3 Answers

1. In the data to the right, which subject had the highest average score?

   Physical Education

2. In the data to the right, which subject had the most variation in score? Which had the least?

   Science, Physical Education

3. What are the 4 rules for exploring data?

   1. Look at the Data
   2. Describe Each Variable
   3. Graph/Stats each Comparison
   4. Write Research Question

Exam Scores			
Subject	N	Mean	Standard Deviation
Art	10	95	3.3
Spelling	10	70	5.8
Math	10	67	3.5
Science	10	84	12.3
Social Studies	10	89	2.1
Physical Education	10	98	1.2

Graphs

- Graphs serve two purposes
  - *Quickly visualize data during data exploration*
  - *Present results of significant statistical analyses*

Types of Graphs to be covered

Type of Graph	Data Type	Usage	Basic Example	Another Example
Histogram	Single numerical variable	Data exploration (determining normality)	*Heights of freshman students*	*Tooth number of apex-predator dinosaurs*
Boxplot	Single numerical variable; single numerical variable + categorical variable	Data exploration, presenting non-parametric t-tests/ANOVA	*Heights of freshman students; Heights of students by grade*	*Weights of apex-predator dinosaurs; Weight of apex-predator dinosaurs by geological period*
Bar Chart	Single numerical variable + categorical variable	Presenting Parametric T-test/ANOVA results	*Heights of students by grade*	*Tooth number of sharks by species*
Scatterplot	Two numerical variables	Data exploration, presenting correlation results	*Heights and weights of students*	*Weights and top swimming speed of sharks*
Line Charts	Two numerical variables (one usually time)	Data exploration	*Heart rate over time*	*Ounces of coffee drank by students over time*
Multiple Line Charts	Three or more numerical variable (one usually time, rest on same scale)	Data exploration	*Various concentrations of nutrients in bloodstream over time*	*Ounces of various caffeinated beverage drank by students over time*
Pie graph	Single numerical variable (proportions) + categorical variable	Data exploration	*Percentage of students across grades*	*Percentage of different caffeinated beverages drank in a month*

Histogram and Normal Distribution

- Histograms can be used to look at the distribution of data
- This is important for determining if the data is parametric or not

Reminder: if data is parametric, it will approximate a normal distribution (bell curve) when viewed as a histogram. Many statistical tests can only be used if the data is parametric

1. Open the file **Reconstructed male heights 1883** in SPSS.
2. This file contains data that is similar to that from which the table you have seen was derived. The file contains 8585 heights, measured in inches.
3. We are going to create a histogram from the values in the variable called **hgtrein**
4. From the menus choose Graph->Chartbuilder.
5. A dialog box will come up, choose OK.
6. In the bottom section Choose **Histogram** and double click the first image
7. Drag the hgtrein (Heights in inches - reconstructed) variable over to the box representing the horizontal (X) axis of the graph.
8. Click OK and wait to see the graph in the output viewer. You should see a normal (bell shaped) pattern to the distribution of the data.
9. To see a normal curve superimposed on the graph go back to the Create Histogram dialog box (from the menus Graph, (Legacy,) Interactive, Histogram) then click on the Histogram tab and tick the "Normal curve" check box, then Click OK.
10. Are these data Discrete or Continuous?

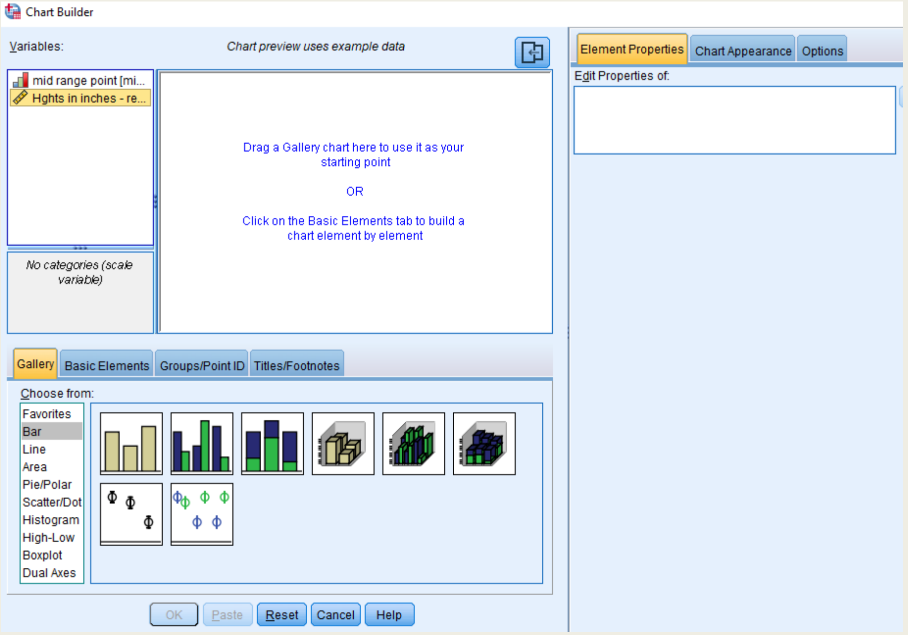


Chart Builder

Variables: mid range point [mi... Hghts in inches - re...

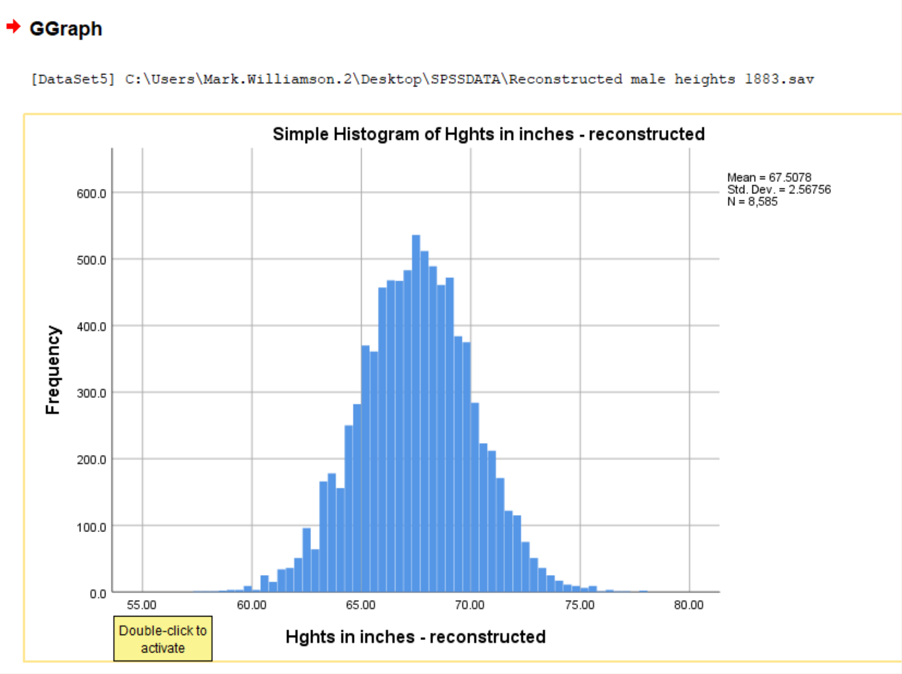
Chart preview uses example data

Drag a Gallery chart here to use it as your starting point

OR

Click on the Basic Elements tab to build a chart element by element

No categories (scale variable)



GGraph

```
[DataSet5] C:\Users\Mark.Williamson.2\Desktop\SPSSDATA\Reconstructed male heights 1883.sav
```

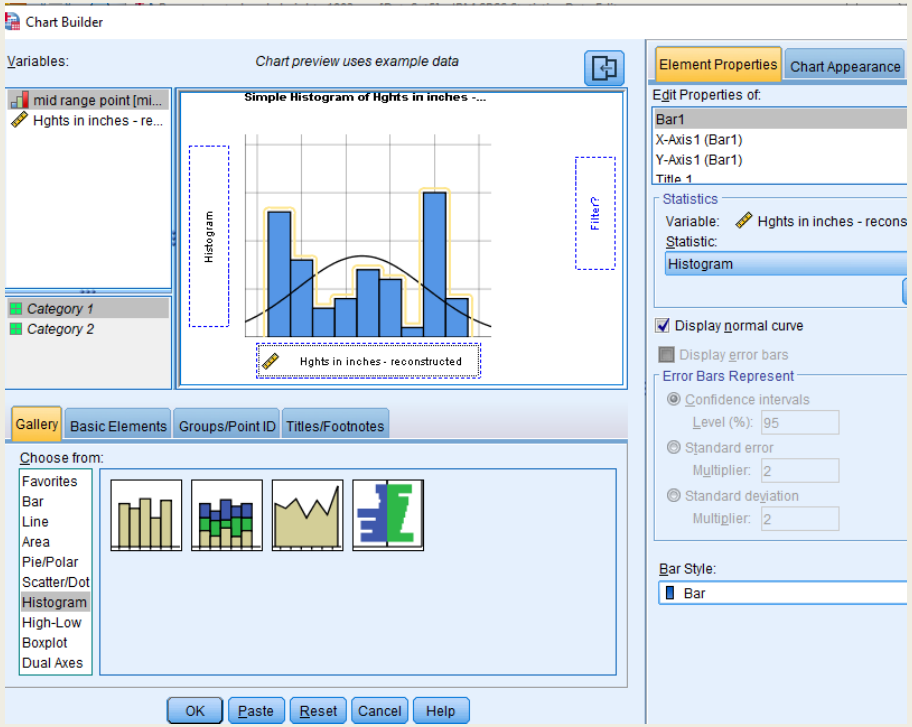


Chart Builder

Variables: mid range point [mi... Hghts in inches - re...

Chart preview uses example data

Edit Properties of: Bar1

Statistics

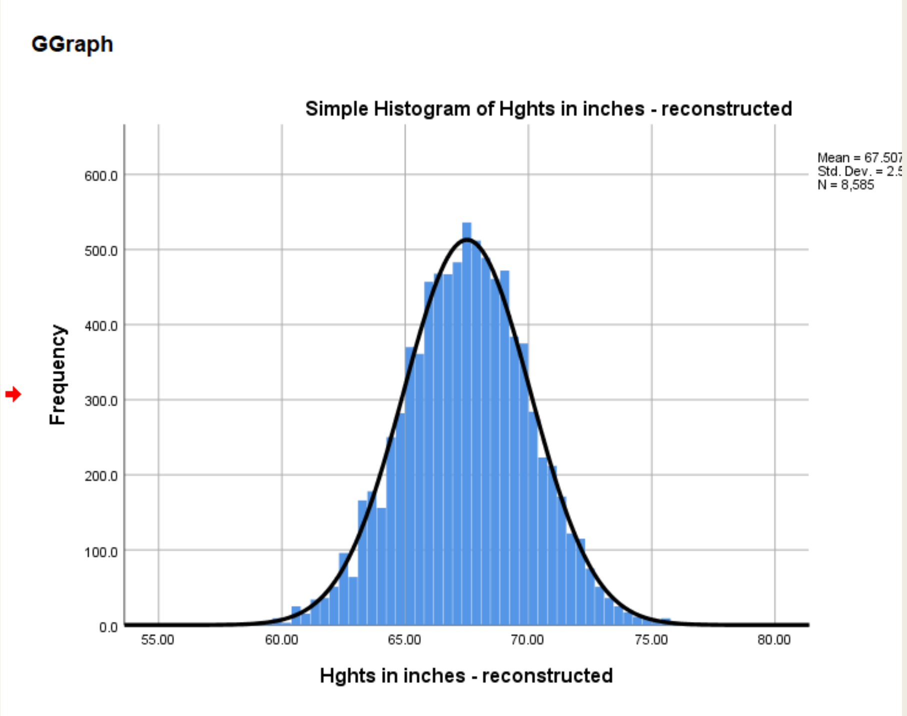
Variable: Hghts in inches - recons

Statistic: Histogram

☑ Display normal curve

□ Display error bars

Bar Style: Bar



GGraph

Histogram 2

Radiologist example:

- The file **Radiologist dose with and without lead combined.sav** contains data gathered to assess the effect of a lead screen to reduce the radiation dose to Radiologists hands while carrying out procedures on patients being irradiated.
- In the trials the lead screen was placed between the patient and the radiologist, the intended effect was to reduce the radiation dose to the radiologist, however there were fears that working through the screen would lengthen the procedure. We want to answer two questions with this data, one about the hand dose and the other about the length of time the examination took.

Summary: Histograms are for displaying continuous data, e.g. height, age etc, the bars touch, signifying the continuous nature of the data. The area of the bars represent the number in each range, the bars are usually of equal widths but this need not always be the case. Histograms should be clearly labelled and the units of measure displayed. The use of Histograms compared to Bar Charts is summarized after the section on Bar Charts.

1. Open **Radiologist dose with and without lead combined** file in SPSS
2. Look at the data, the variable called "**screen**" is the variable that lets you discriminate between procedures carried out with or without the lead screen. If there is a 1 in the screen variable column it means the procedure was carried out with the screen in place, if not the value is 0.
3. We can use this discriminatory variable to create two histograms at once, by using it as a panel variable.
4. The variable we are interested in is the dose to the radiologists' left hand, the left-hand would be nearest the patient so we will concentrate on the left-hand dose variable.
5. Draw histogram using the left-hand dose variable (**lhdose**)
6. Go to the Groups/Point ID tab and click the Rows panel variable
7. Drag the discriminatory variable (**Lead or No Lead**) as the panel variable.
8. What do the histograms show us about the data?
9. If you have time draw a similar histogram using the **extimmin** variable. Does this back up the fears about the increase in examination time?

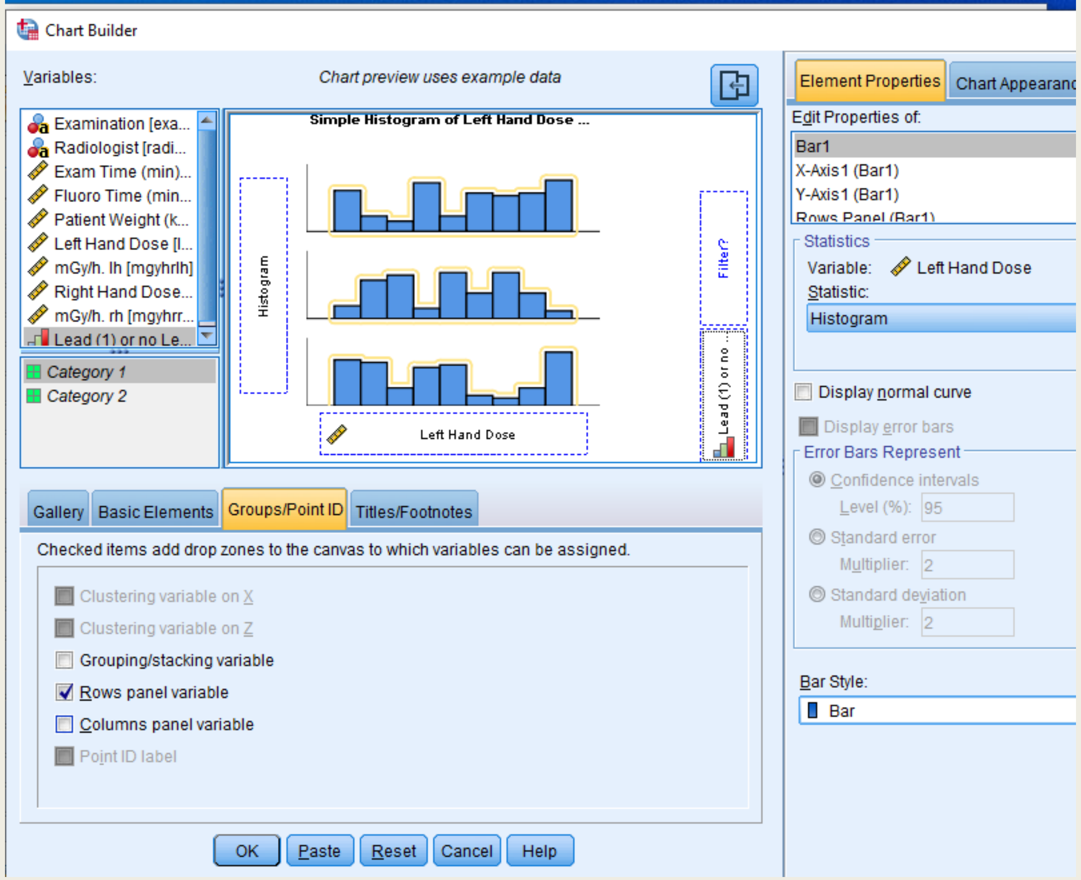


Chart Builder

Variables:

Chart preview uses example data

- Examination [exa...
- Radiologist [radi...
- Exam Time (min)...
- Fluoro Time (min...
- Patient Weight (k...
- Left Hand Dose [l...
- mGy/h. lh [mgyhrlh]
- Right Hand Dose...
- mGy/h. rh [mgyhrr...
- Lead (1) or no Le...

- Category 1
- Category 2

Checked items add drop zones to the canvas to which variables can be assigned.

□ Clustering variable on X

□ Clustering variable on Z

□ Grouping/stacking variable

☑ Rows panel variable

□ Columns panel variable

□ Point ID label

Edit Properties of:

- Bar1
- X-Axis1 (Bar1)
- Y-Axis1 (Bar1)
- Rows Panel (Bar1)

Statistics

Variable: Left Hand Dose

Statistic: Histogram

□ Display normal curve

□ Display error bars

Error Bars Represent

● Confidence intervals

Level (%): 95

○ Standard error

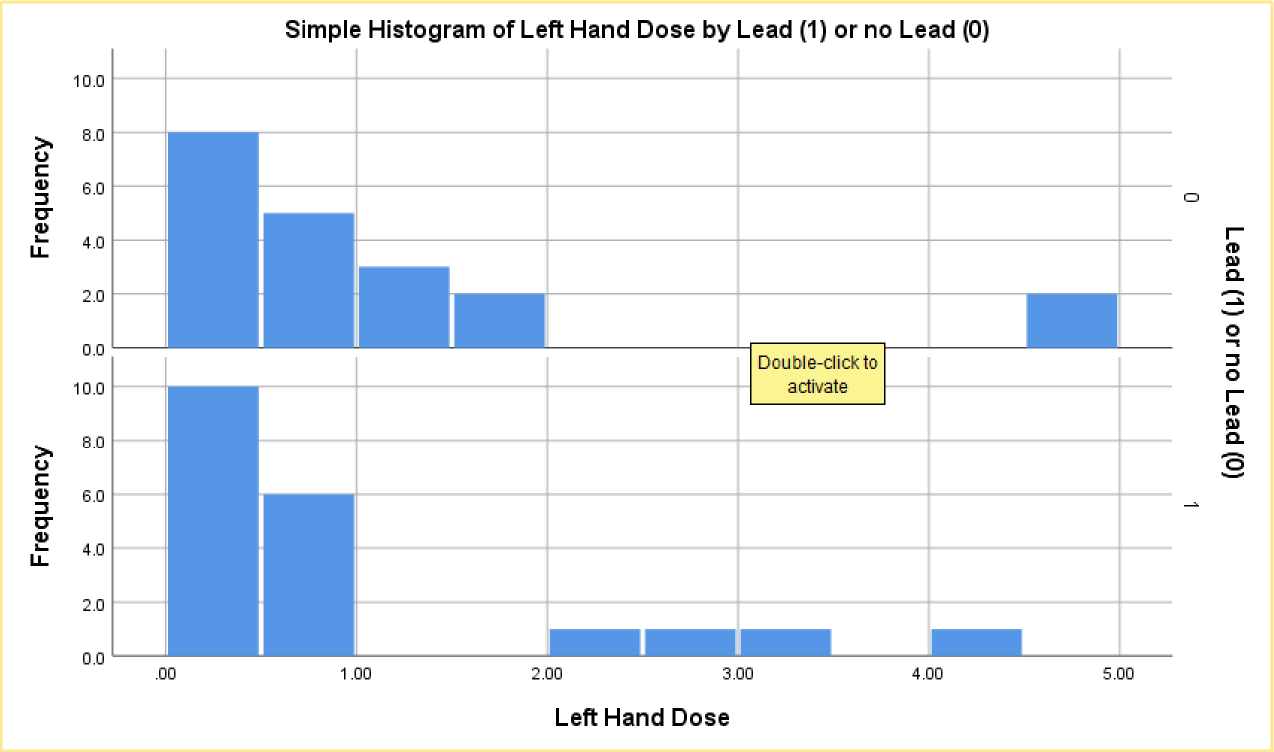
Multiplier: 2

○ Standard deviation

Multiplier: 2

Bar Style: Bar

GGraph



Drawing boxplots

- Boxplots are a great way to visualize data between varies groups
- Requires: numerical dependent variable and a factor with 2 or more groups
- For paired data, you can draw boxplots straight from the graph menu

1. Go back to in **studentsss**
2. Choose Chartbuilder under Graphs
3. In the bottom section Choose **Boxplot** and double click the first image
4. Drag the **speaks** variable to the y-axis and the **year** variable to the x-axis.
5. Look at your boxplots. Can you see an asterisk or circle beyond the whiskers? In SPSS an asterisk represents an extreme outlier (a value more than 3 times the interquartile range from a quartile). A circle is used to mark other outliers with values between 1.5 and 3 box lengths from the upper or lower edge of the box. The box length is the interquartile range.
6. Which number on your data screen does the most extreme outlier correspond to? (SPSS gives a bit of a hint here!) Why is it an extreme outlier?
7. Look at the boxplots, which group has the highest median? What does this tell you about the groups?
8. Look at the boxplots, which group has the highest interquartile range (IQR)? What does this tell you about the groups? Refer to a glossary to review IQR

Summary: Boxplots are good for seeing the range and level of data and highlighting outliers. The box shows the IQR (Inter Quartile Range) and the bar in the box shows the median. Boxplots should be clearly labelled with the units of measure displayed.

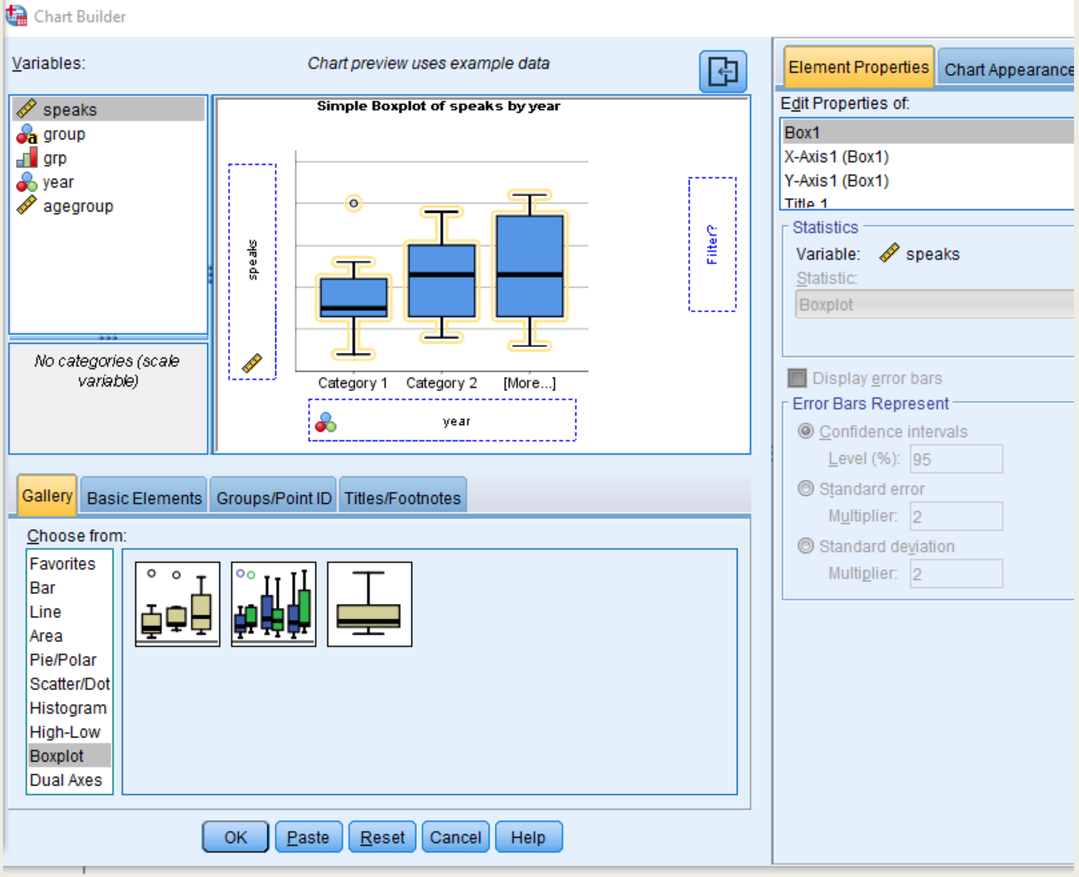
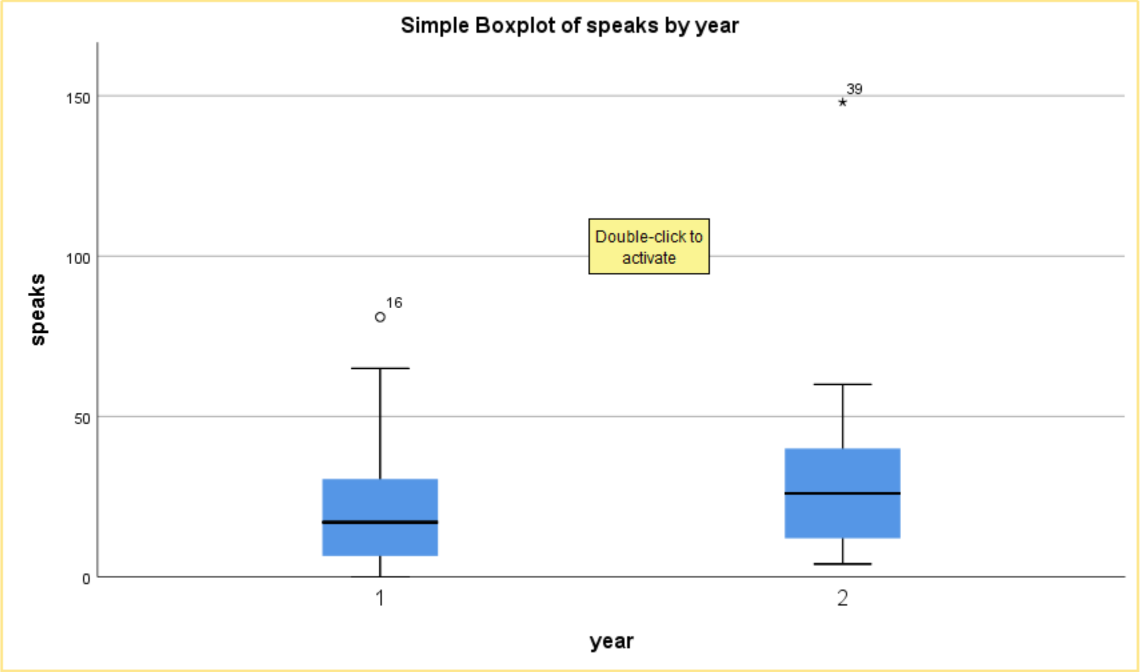


Chart Builder
Variables:
speaks
group
grp
year
agegroup
No categories (scale variable)
Chart preview uses example data
Simple Boxplot of speaks by year
speaks
Category 1
Category 2
[More...]
year
Filter?
Gallery
Basic Elements
Groups/Point ID
Titles/Footnotes
Choose from:
Favorites
Bar
Line
Area
Pie/Polar
Scatter/Dot
Histogram
High-Low
Boxplot
Dual Axes
OK
Paste
Reset
Cancel
Help
Element Properties
Chart Appearance
Edit Properties of:
Box1
X-Axis1 (Box1)
Y-Axis1 (Box1)
Title 1
Statistics
Variable:
speaks
Statistic:
Boxplot
Display error bars
Error Bars Represent
Confidence intervals
Level (%):
95
Standard error
Multiplier:
2
Standard deviation
Multiplier:
2

GGraph



Simple Boxplot of speaks by year
150
100
50
0
speaks
39
16
Double-click to activate
1
2
year

Bar Charts

- Bar charts and histograms look similar at first
- there is however a definite difference in the type of data each is designed to show and this subtle difference is an important one if you are using them in your research.
- Bar charts are for non-continuous data, i.e. data in categories that are not related in any order.
- Histograms are for displaying continuous data
- the graph can be edited after it is drawn, just double click on the graph and then click into the labels you wish to alter

Summary: Bar charts are for non-continuous data e.g. the number of people from each of five towns, the bars do not touch. Bar charts should be clearly labelled and the units of measure displayed. Bar charts and Histograms look similar, however the type of data they should be used on is different. In a Histogram the bars touch each other, this denotes the continuous nature of the data being displayed. Bar charts should be used for discrete data. If you aren't sure about the difference between continuous and discrete data look it

1. open the file **shoetype**s in SPSS
2. this file contains data about the type of shoes worn at the time the data were gathered and number of pairs owned by a sample of 100 people. We can use SPSS to analyze the data by using bar charts among other methods.
3. Graphs->Chartbuilder->Select Bar
4. Drag the **footwaretxt** variable to the x-axis then click OK. The graph above should appear.
5. Try again but this time, include a Rows Panel variable, then drag the **gendertxt** variable over Panel box and see what happens.

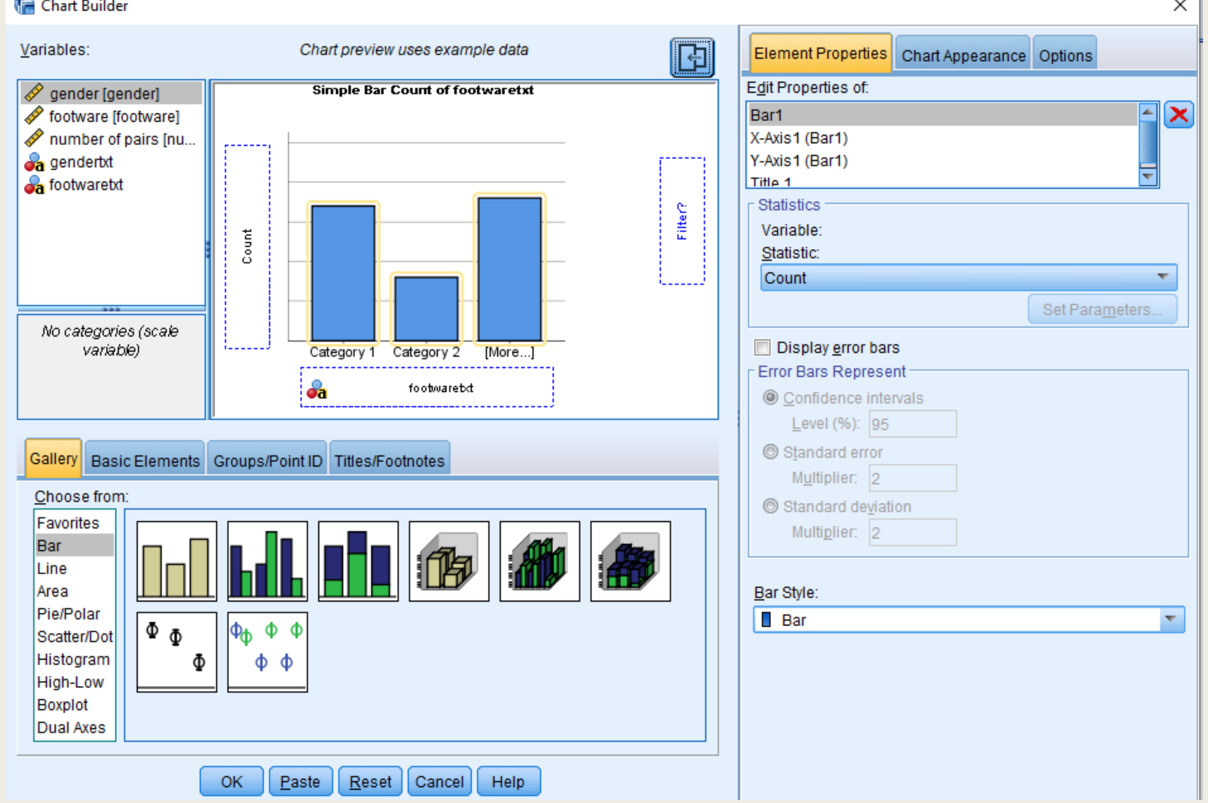


Chart Builder

Variables:

- gender [gender]
- footware [footware]
- number of pairs [nu...
- gendertxt
- footwaretxt

No categories (scale variable)

Chart preview uses example data

Edit Properties of:

- Bar1
- X-Axis1 (Bar1)
- Y-Axis1 (Bar1)
- Title 1

Statistics

Variable:

Statistic: Count

Set Parameters...

□ Display error bars

Error Bars Represent

● Confidence intervals

Level (%): 95

○ Standard error

Multiplier: 2

○ Standard deviation

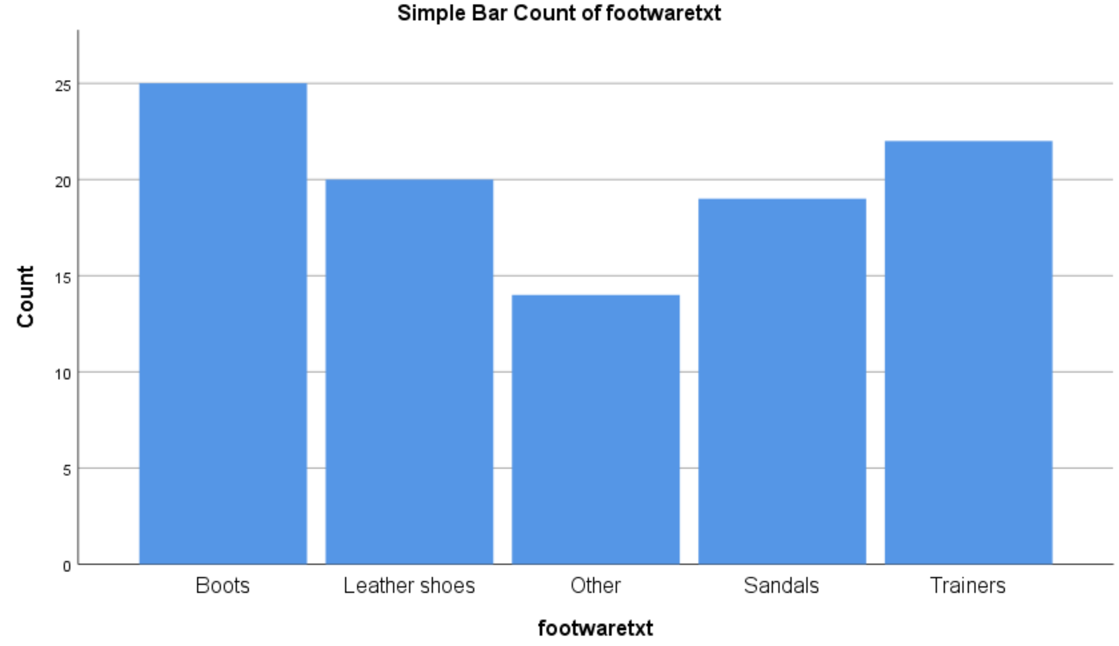
Multiplier: 2

Bar Style: Bar

```
END GPL.
```

GGraph

```
[DataSet8] C:\Users\Mark.Williamson.2\Desktop\SPSSDATA\shoetypes.sav
```



Percentages

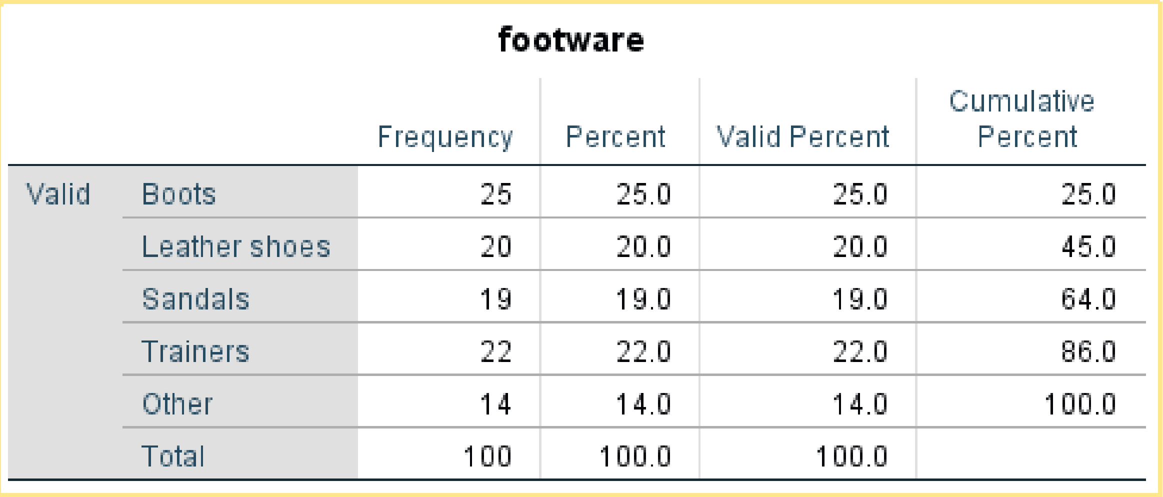
- Percentages are often used in bar charts
- General formula for calculation percentages

  100 × the individual value ÷ the total of the values

- If percentages span across all values, the total needs to sum to 100% across all groups

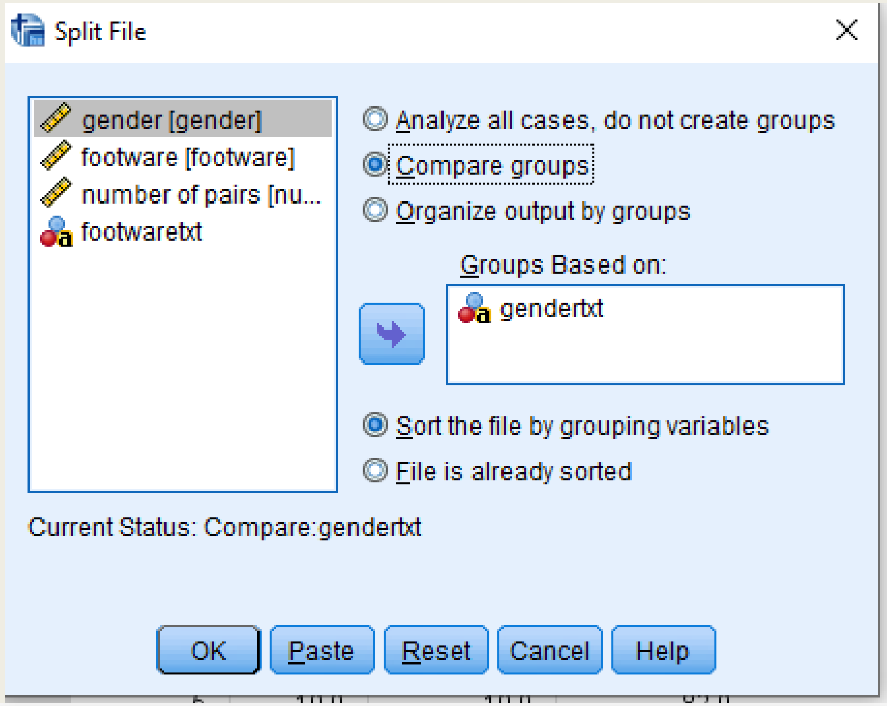
Summary: Percentages show proportions, it should be clear what they are percentages of.

1. You can very quickly create summary percentages using the "frequencies" command, for example in the shoes file,
2. What percentage of subjects were wearing each type of shoe?
3. Analyze->Descriptive Statistics->Frequencies
4. Add **footware** to the variables list
5. Does the percentage of footwear types differ in the different gender grouped?
6. Lets get SPSS to do everything twice, once for males and once for females, we can do this using the split file command. Choose Data->Split file. Now calculate the percentages again as you did before.
7. The output should now be split into two groups, one for Male and one for Female. Tables like this are rarely in the ideal format for inclusion in a dissertation or paper but can be copied and pasted into a word processor and manipulated there.
8. Remove the split once you have done with it. If you leave it on you may get some strange results. Choose Data, Split file. Then select the "Analyse all cases" option, then click OK.
9. Don't forget to switch this feature off when you don't need it!



footware

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Boots	25	25.0	25.0	25.0
	Leather shoes	20	20.0	20.0	45.0
	Sandals	19	19.0	19.0	64.0
	Trainers	22	22.0	22.0	86.0
	Other	14	14.0	14.0	100.0
	Total	100	100.0	100.0	



Split File

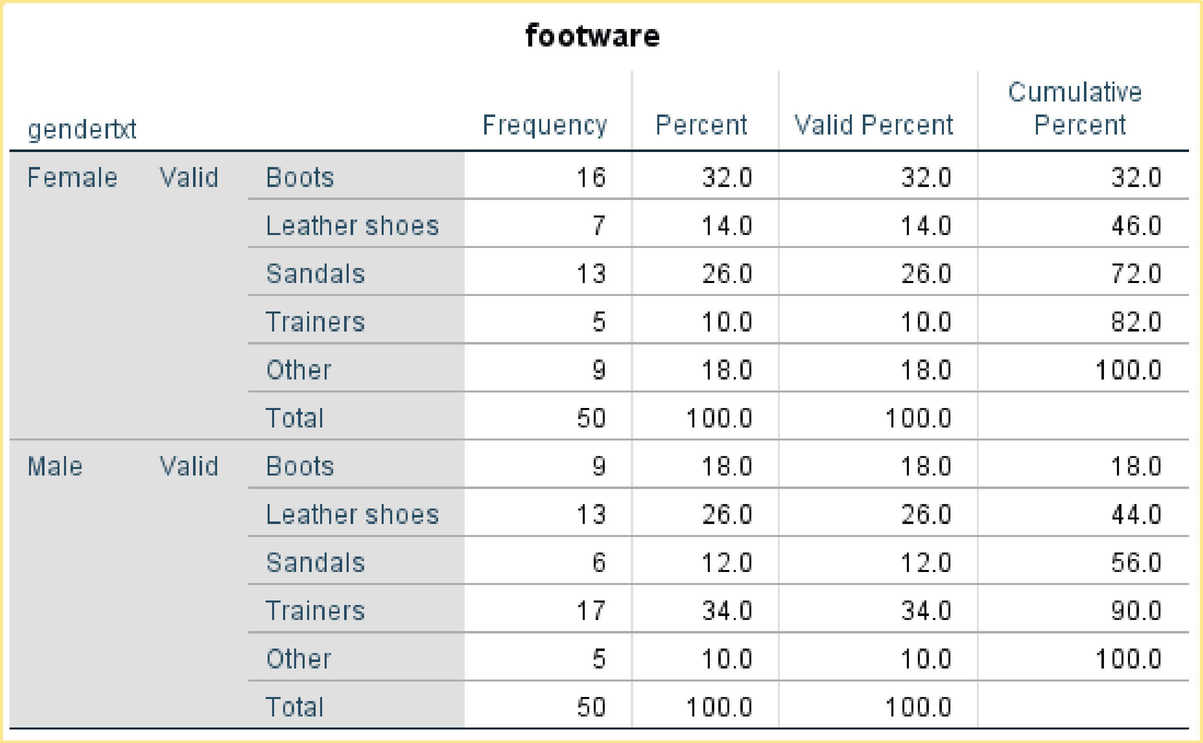
- gender [gender]
- footware [footware]
- number of pairs [nu...
- footwaretxt

○ Analyze all cases, do not create groups
● Compare groups
○ Organize output by groups

Groups Based on:
gendertxt

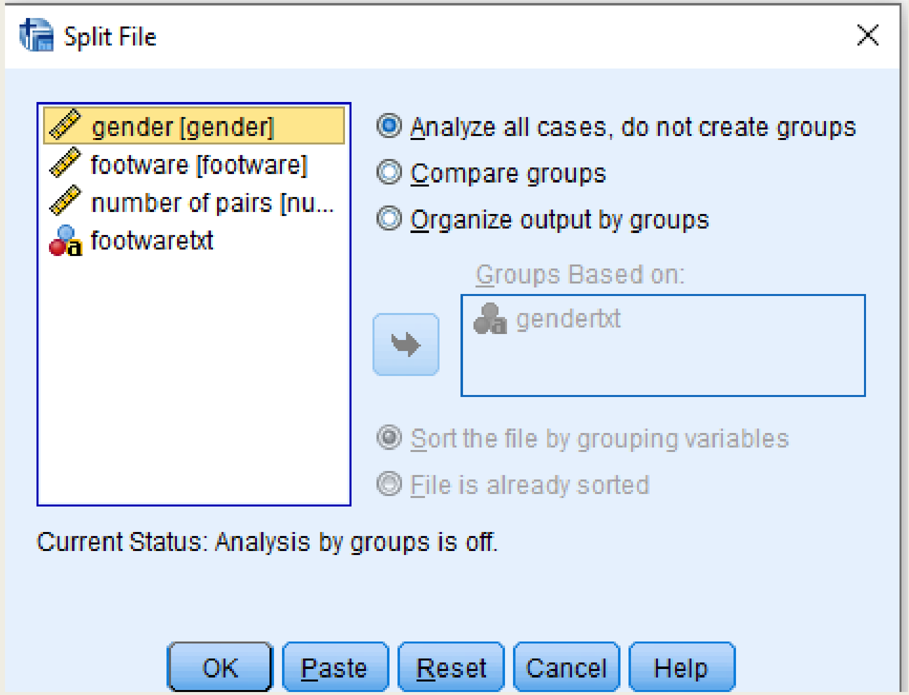
● Sort the file by grouping variables
○ File is already sorted

Current Status: Compare:gendertxt



footware

gendertxt			Frequency	Percent	Valid Percent	Cumulative Percent
Female	Valid	Boots	16	32.0	32.0	32.0
		Leather shoes	7	14.0	14.0	46.0
		Sandals	13	26.0	26.0	72.0
		Trainers	5	10.0	10.0	82.0
		Other	9	18.0	18.0	100.0
		Total	50	100.0	100.0	
Male	Valid	Boots	9	18.0	18.0	18.0
		Leather shoes	13	26.0	26.0	44.0
		Sandals	6	12.0	12.0	56.0
		Trainers	17	34.0	34.0	90.0
		Other	5	10.0	10.0	100.0
		Total	50	100.0	100.0	



Split File

- gender [gender]
- footware [footware]
- number of pairs [nu...
- footwaretxt

● Analyze all cases, do not create groups
○ Compare groups
○ Organize output by groups

Groups Based on:
gendertxt

● Sort the file by grouping variables
○ File is already sorted

Current Status: Analysis by groups is off.

Scatterplots

- Used when data are paired: each point on a diagram represents a pair of numbers
- A better description is that you use scatter plots when comparing two numerical variables (Unlike a numerical and categorical like in box or bar)
- Scatter plots are used to detect correlation
  - *Correlation is not causation*
  - *Strong, weak, or no correlation,*
  - *Positive or negative*

1. Open the file **Step** in SPSS
2. These data come from an experiment to see whether subjects could perform more step exercises in a fixed time in a group or on their own. A physiotherapy student collected them as part of a third year project.
3. Look at the data; you will see that the columns are of equal length, this is another indication that the data are paired.
4. We are going to draw a scatterplot for these two columns with the number of steps done individually on the x-axis.
5. Graphs->Chartbuilder->Scatter/Dot.
6. Drag **individ** to X-axis and **group** to Y-axis.
7. Do the points appear to form a line?
8. If they do is it a clear, quite thin line or more like a cloud?
9. Does it slope up or down from left to right?
10. Look at your answers and decide if there is a strong, weak or no correlation. Is it positive or negative?

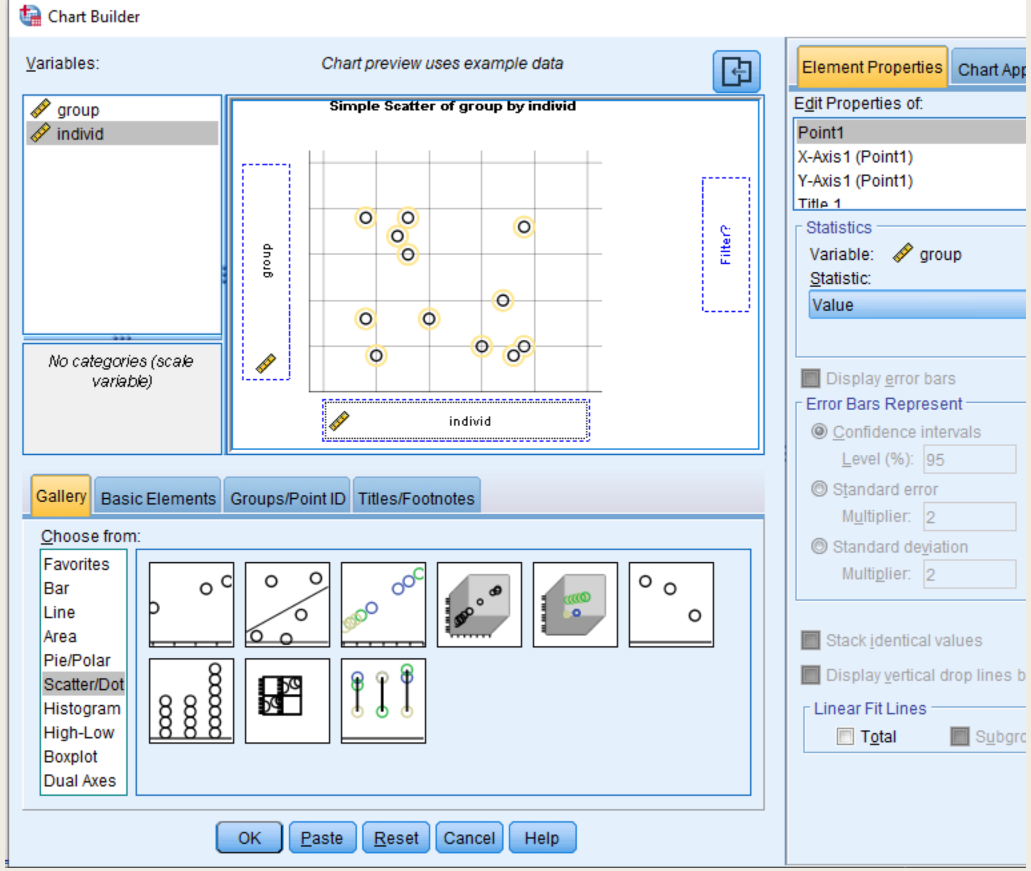


Chart Builder

Variables:

- group
- individ

No categories (scale variable)

Chart preview uses example data

Simple Scatter of group by individ

Edit Properties of:

- Point1
- X-Axis1 (Point1)
- Y-Axis1 (Point1)
- Title 1

Statistics

Variable: group

Statistic: Value

□ Display error bars

Error Bars Represent

○ Confidence intervals

Level (%): 95

○ Standard error

Multiplier: 2

○ Standard deviation

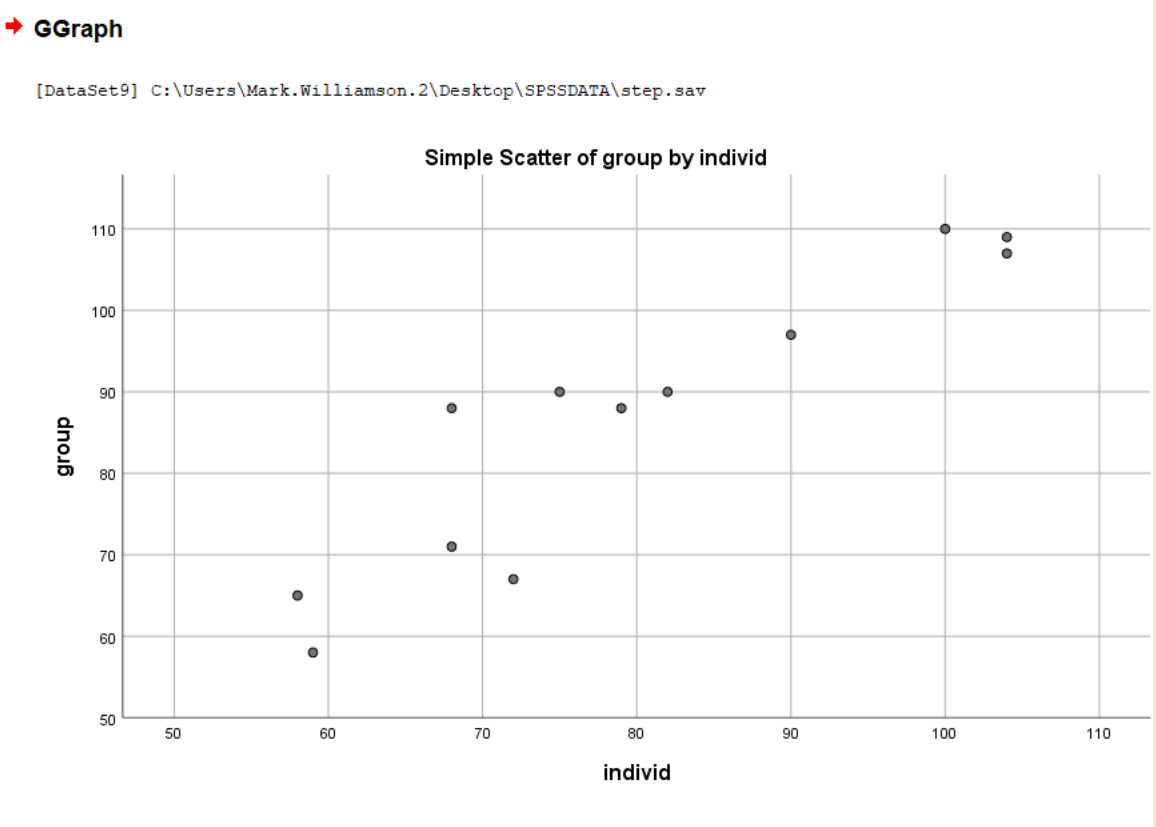
Multiplier: 2

□ Stack identical values

□ Display vertical drop lines b

Linear Fit Lines

□ Total □ Subgro



GGraph

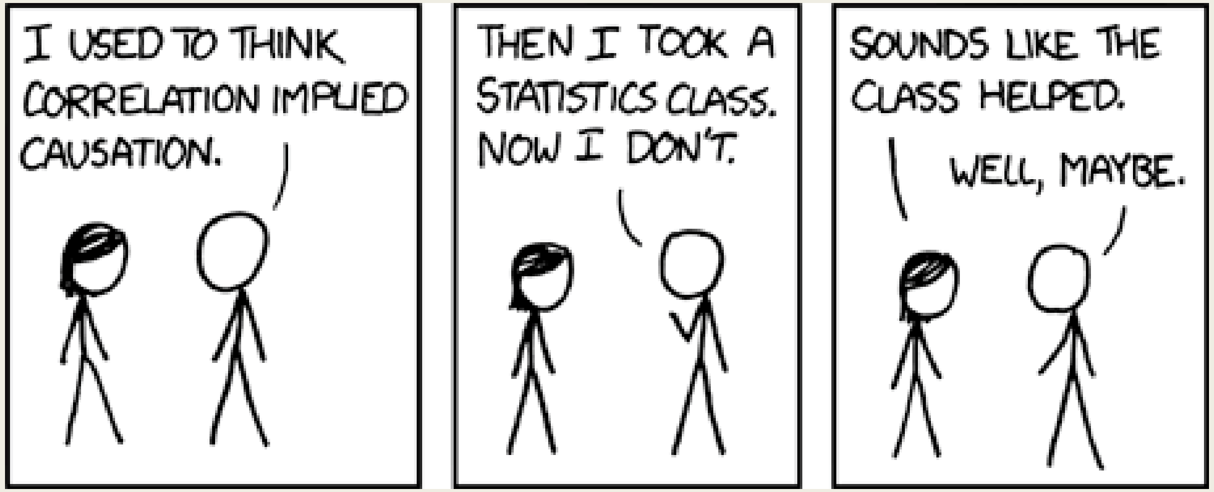
```
[DataSet9] C:\Users\Mark.Williamson.2\Desktop\SPSSDATA\step.sav
```

Simple Scatter of group by individ

Scatterplots 2

Summary: Scatter plots are used to show paired data, where for example one person is tested under two circumstances, each individual will have a pair of readings. In this example a scatter plot can be used to indicate changes between the performance in different circumstances. Scatter plots are also typically used to show correlation. Scatter plots should be clearly labelled and the units of measure displayed.

1. Adding more information to the previous plot
2. Go back to the Chartbuilder.
3. To add a linear fit line, select the Total box under Linear Fit Lines, or select the second scatterplot image.
4. Does the line match what you predicted?


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

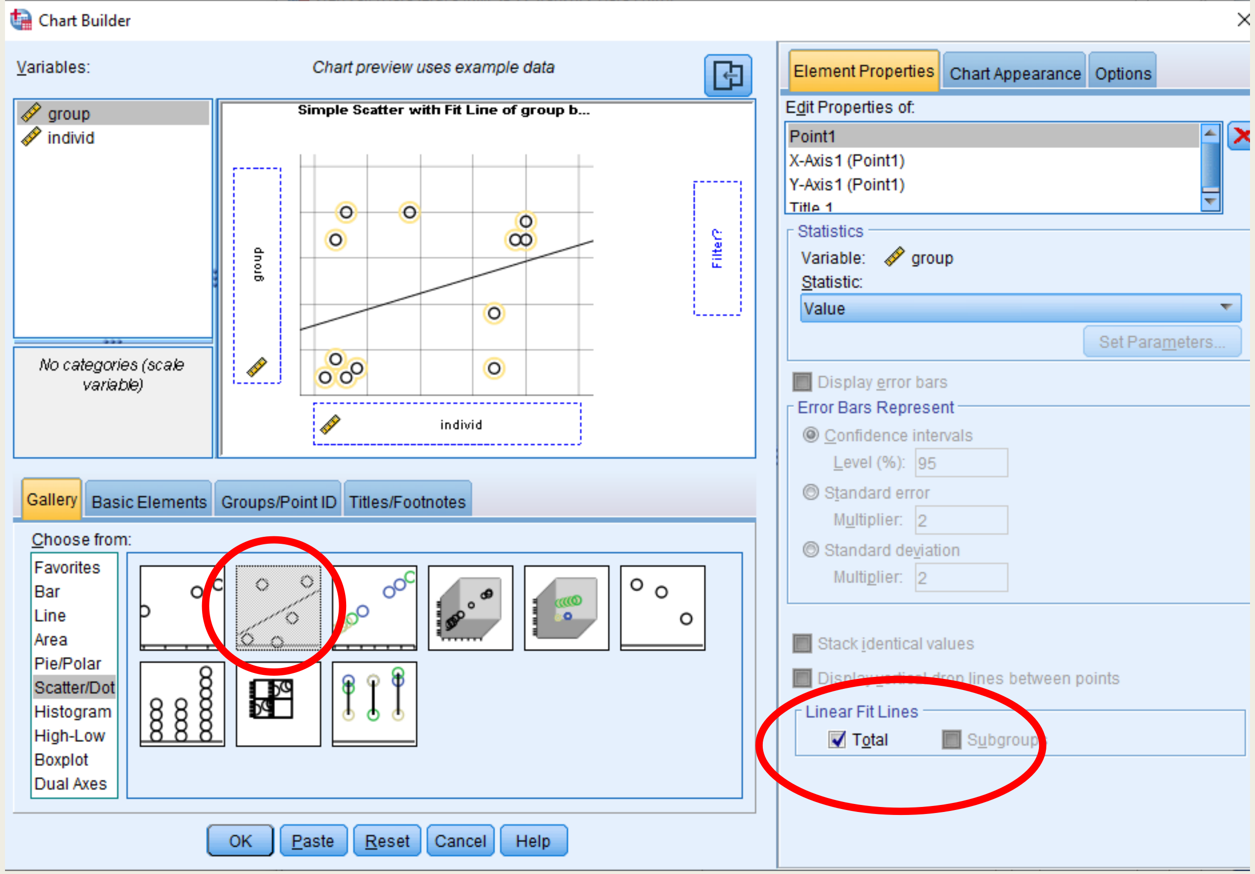


Chart Builder

Variables: group, individ

Chart preview uses example data

Simple Scatter with Fit Line of group b...

No categories (scale variable)

Gallery | Basic Elements | Groups/Point ID | Titles/Footnotes

Choose from: Favorites, Bar, Line, Area, Pie/Polar, Scatter/Dot, Histogram, High-Low, Boxplot, Dual Axes

OK | Paste | Reset | Cancel | Help

Element Properties | Chart Appearance | Options

Edit Properties of: Point1, X-Axis1 (Point1), Y-Axis1 (Point1), Title 1

Statistics

Variable: group

Statistic: Value

Set Parameters...

□ Display error bars

Error Bars Represent

● Confidence intervals

Level (%): 95

○ Standard error

Multiplier: 2

○ Standard deviation

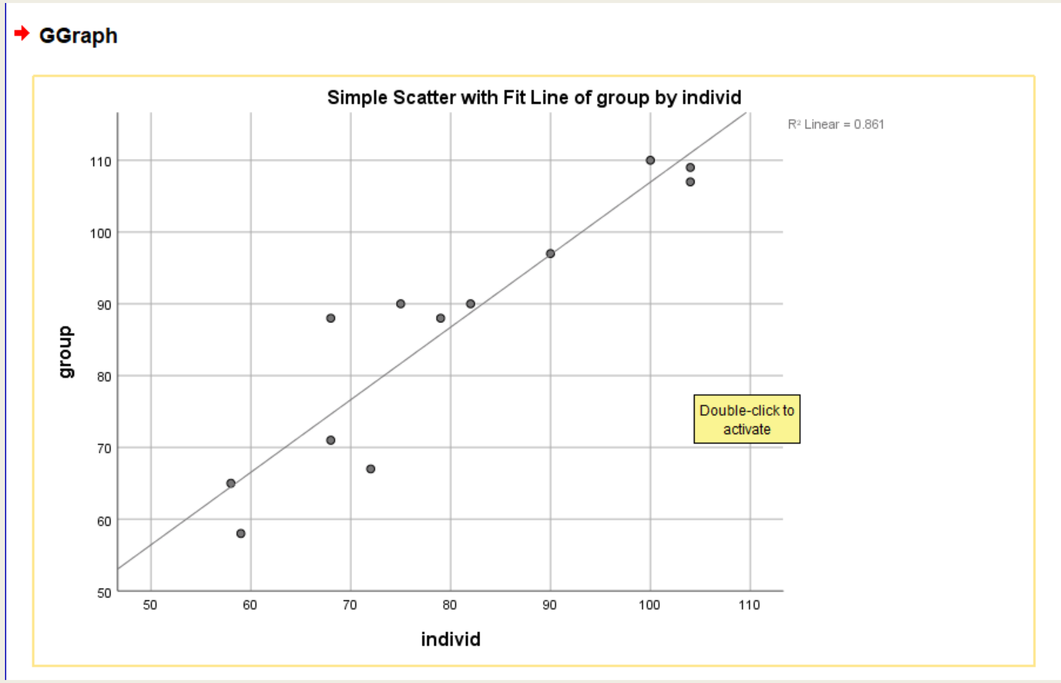
Multiplier: 2

□ Stack identical values

□ Display vertical drop lines between points

Linear Fit Lines

☑ Total □ Subgroup



GGraph

Line graphs

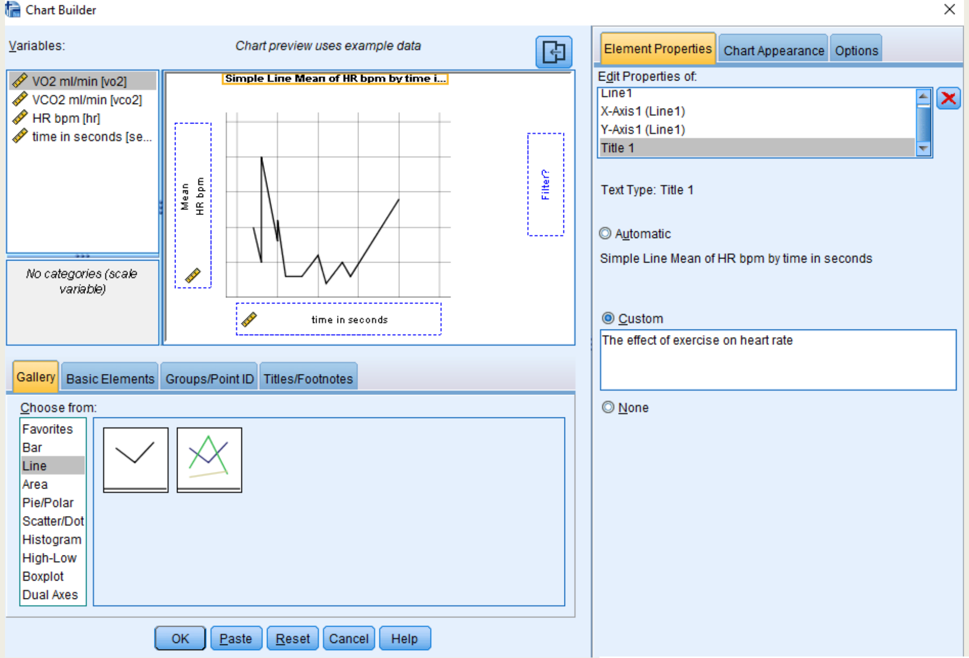
Line graphs are useful in time-based designs

Typically consists of a numerical variable over time

Example: Oxygen used walking description

- The variables in the file are: vo2 Volume of O2 ml/min vco2 Volume of CO2 ml/min hr Heart Rate beats per minute seconds time in seconds from start of procedure
- The protocol employed to take the measurements consisted of:
  - *5 minutes rest, to achieve baseline values for heart rate and enable the subject to get used to the equipment, followed by:*
  - *10 minutes exercise, (walking at a self-selected speed) followed by:*
  - *a second 5 minutes rest, to ensure baseline values return to the norm for the subject. This is important when interpreting the graph we are about to draw.*

1. Open the file **Oxygen used walking** in SPSS
2. The data is just part of a large dataset collected by a student researching the effect of tibial malunion on oxygen expenditure during exercise.
3. For our purposes the data gives us a good example of a variable changing over time. The file contains the data from just one subject.
4. From the menus choose Graphs->Chartbuilder->Line
5. Drag the Heart Rate (**HR bpm**) onto the Y-axis and the Time (**time in seconds**) onto the X-axis.
6. Look at the graph. It is easy to see when the subject started and stopped walking!
7. The increase looks massive, but it is because the graph used a false origin (not set at zero). We'll want to redraw and label better.
8. Go back to Chartbuilder
9. Click on the "Titles" tab and switch the text from Automatic to Custom and add an appropriate name, such as "The effect of exercise on heart rate."
10. Click on the Y-axis. Under the Scale Range, Switch off the Automatic feature. Set the minimum to 0 and the maximum to 100. Press OK.



Edit Properties of:
Line1
X-Axis1 (Line1)
Y-Axis1 (Line1)
Title 1

Text Type: Title 1

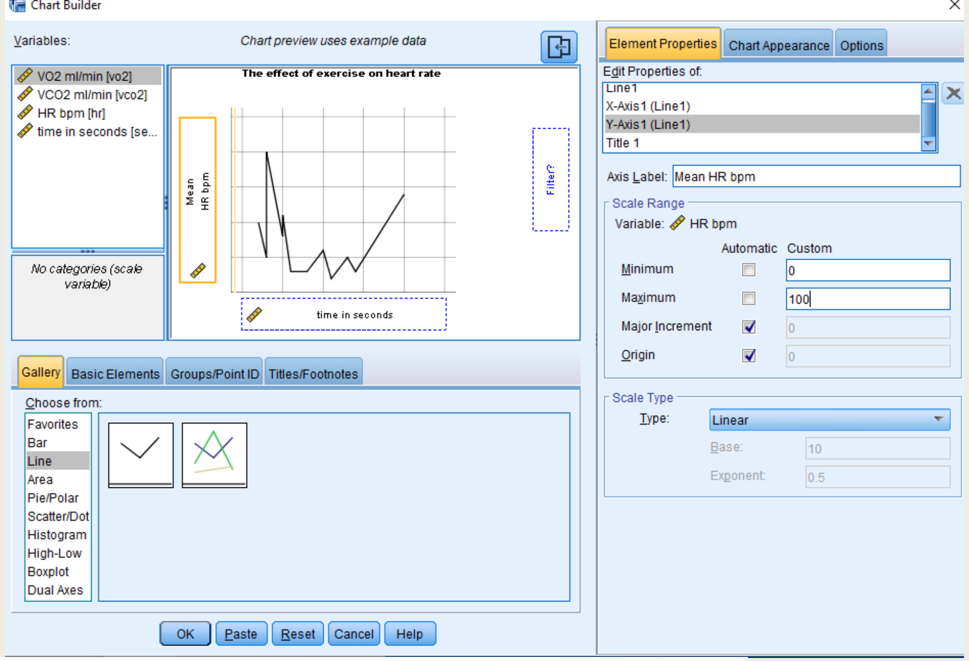
○ Automatic

Simple Line Mean of HR bpm by time in seconds

● Custom

The effect of exercise on heart rate

○ None



Edit Properties of:
Line1
X-Axis1 (Line1)
Y-Axis1 (Line1)
Title 1

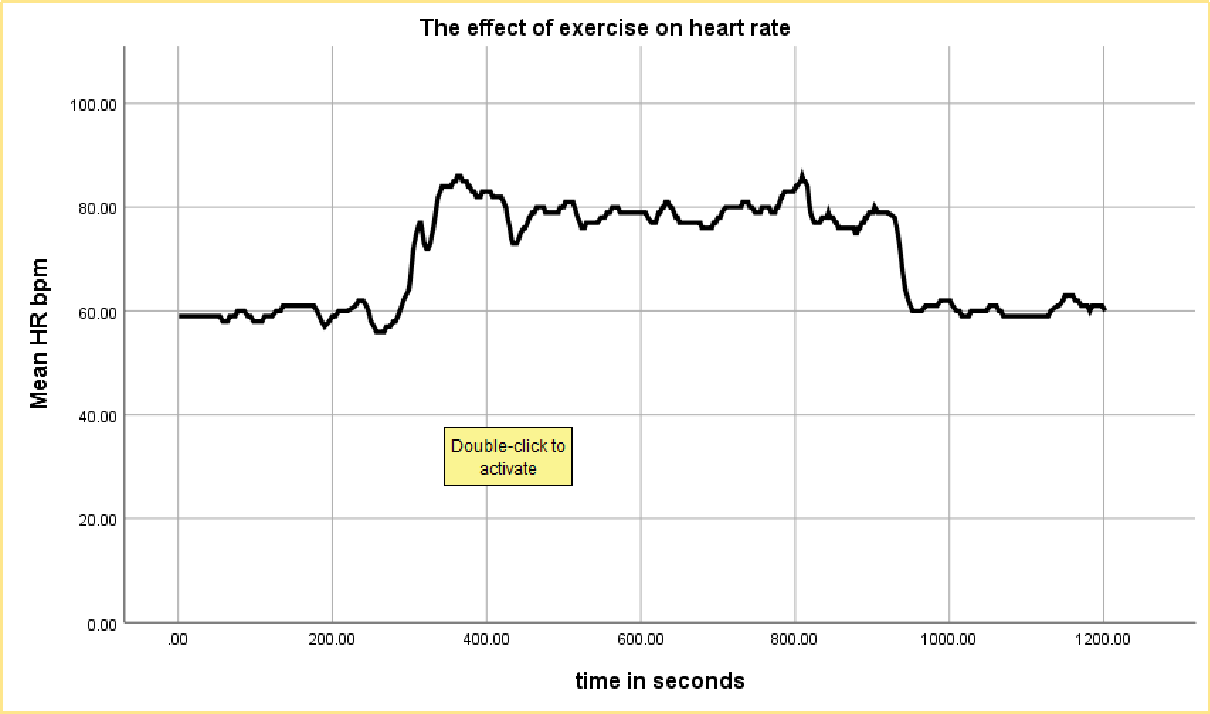
Axis Label: Mean HR bpm

Scale Range
Variable: HR bpm

	Automatic	Custom
Minimum	□	0
Maximum	□	100
Major Increment	☑	0
Origin	☑	0

Scale Type
Type: Linear
Base: 10
Exponent: 0.5

GGraph

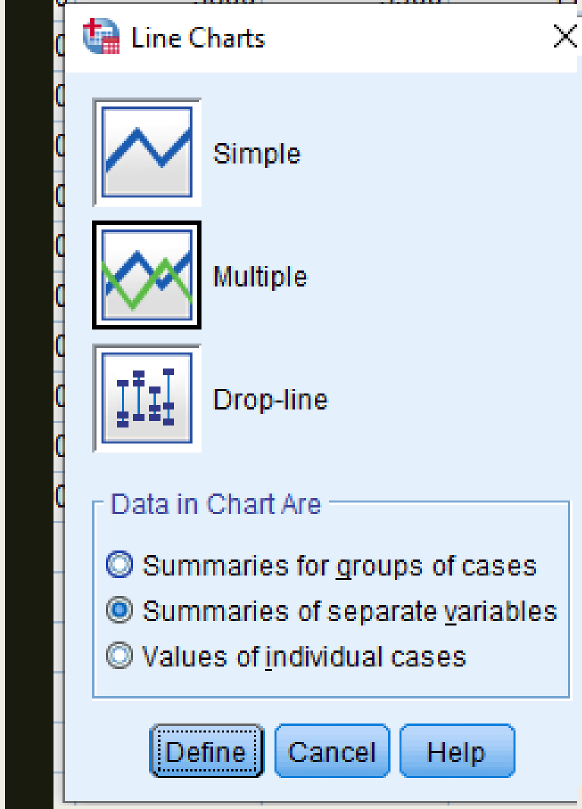


Multiple Line graphs

- More than one line can be plotted at once, as long as the time variable is consistent
- Example: Children looked after
  - *The data are from the Department for Education and Skills*
  - *gives figures for children looked after by Local Authorities in England.*

Summary: Line graphs are ideal for showing the changes in a variable as another alters, e.g. changes over time. The independent variable goes on the x-axis and the dependent variable goes up the y-axis. More than one line is often shown on the chart allowing comparisons. Line graphs should be clearly labelled and the units of measure displayed.

1. We will use the older graphing system and the data in the file called **Children looked after**
2. The variable names may look a bit strange at first, go to the Variable view.
3. Graphs->Legacy Dialogs->Line->Multiple
4. Select the option for Multiple lines and Summaries of separate variables. Then press "Define".
5. Transfer the variables "**Boys 1-4**" and "**Girls 1-4**" to the top box and the "year" to the Category Axis then click OK.
6. The graph that appears should let you answer the following questions;
7. In the 11 years covered by the data do the numbers of girls and boys aged 1 to 4 looked after by Local Authorities in England appear to increase?
8. Are the number of boys and girls in the age group 1 to 4 staying in roughly the same proportion, i.e. do they seem to increase or decrease together?
9. Now plot the data for the 16 and over age group, can you see any difference between the girls and boys?



Line Charts

Simple

Multiple

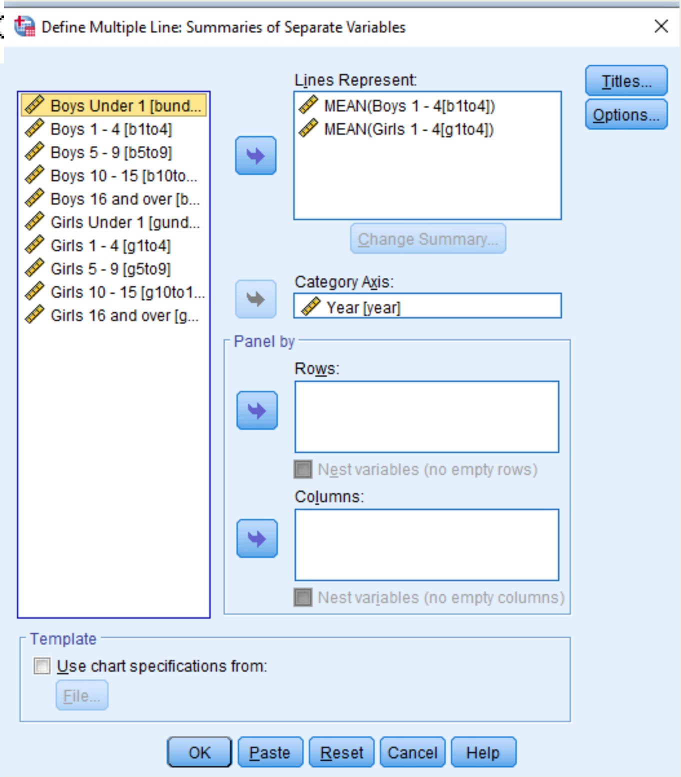
Drop-line

Data in Chart Are

○ Summaries for groups of cases

● Summaries of separate variables

○ Values of individual cases



Define Multiple Line: Summaries of Separate Variables

- Boys Under 1 [bund...
- Boys 1 - 4 [b1to4]
- Boys 5 - 9 [b5to9]
- Boys 10 - 15 [b10to...
- Boys 16 and over [b...
- Girls Under 1 [gund...
- Girls 1 - 4 [g1to4]
- Girls 5 - 9 [g5to9]
- Girls 10 - 15 [g10to1...
- Girls 16 and over [g...

Lines Represent:
MEAN(Boys 1 - 4[b1to4])
MEAN(Girls 1 - 4[g1to4])

Category Axis:
Year [year]

Panel by

Rows:

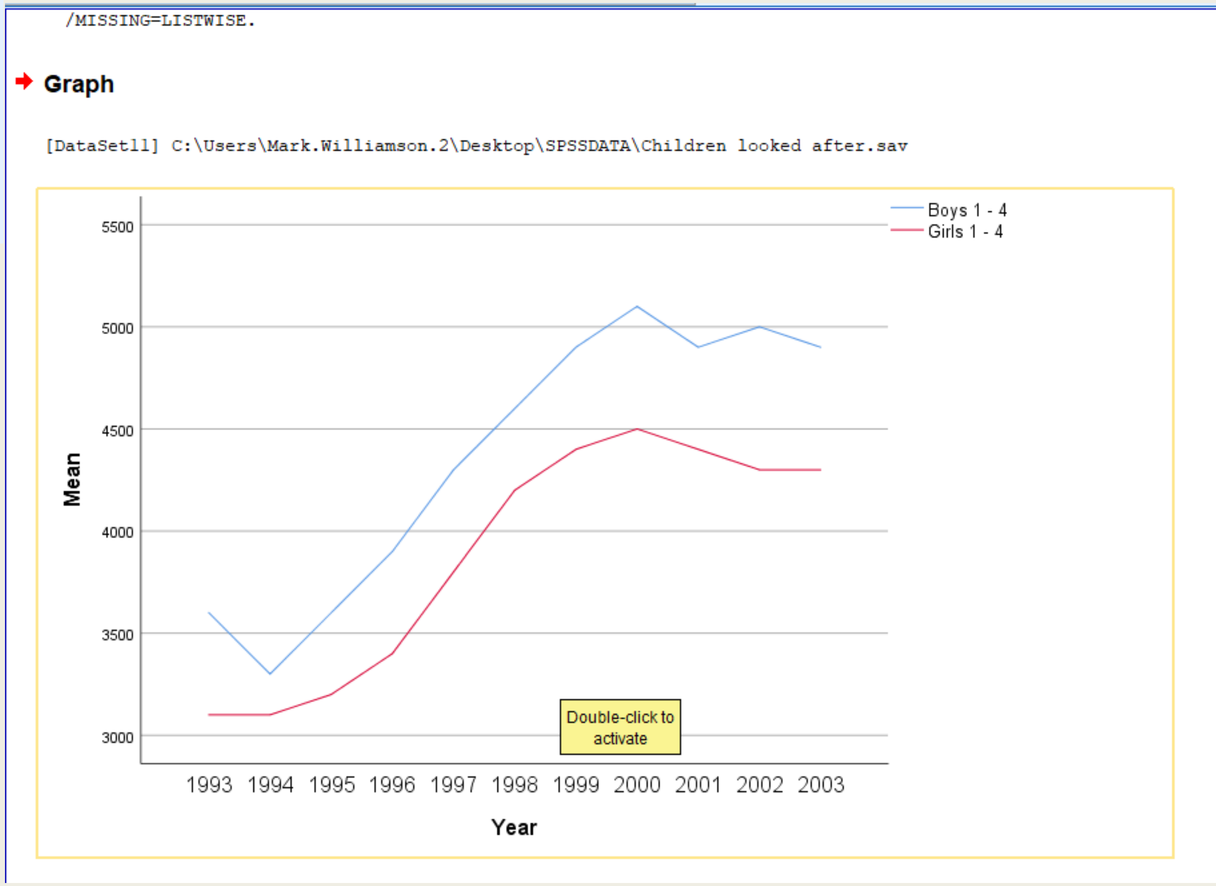
□ Nest variables (no empty rows)

Columns:

□ Nest variables (no empty columns)

Template

□ Use chart specifications from:



```
/MISSING=LISTWISE.
```

Graph

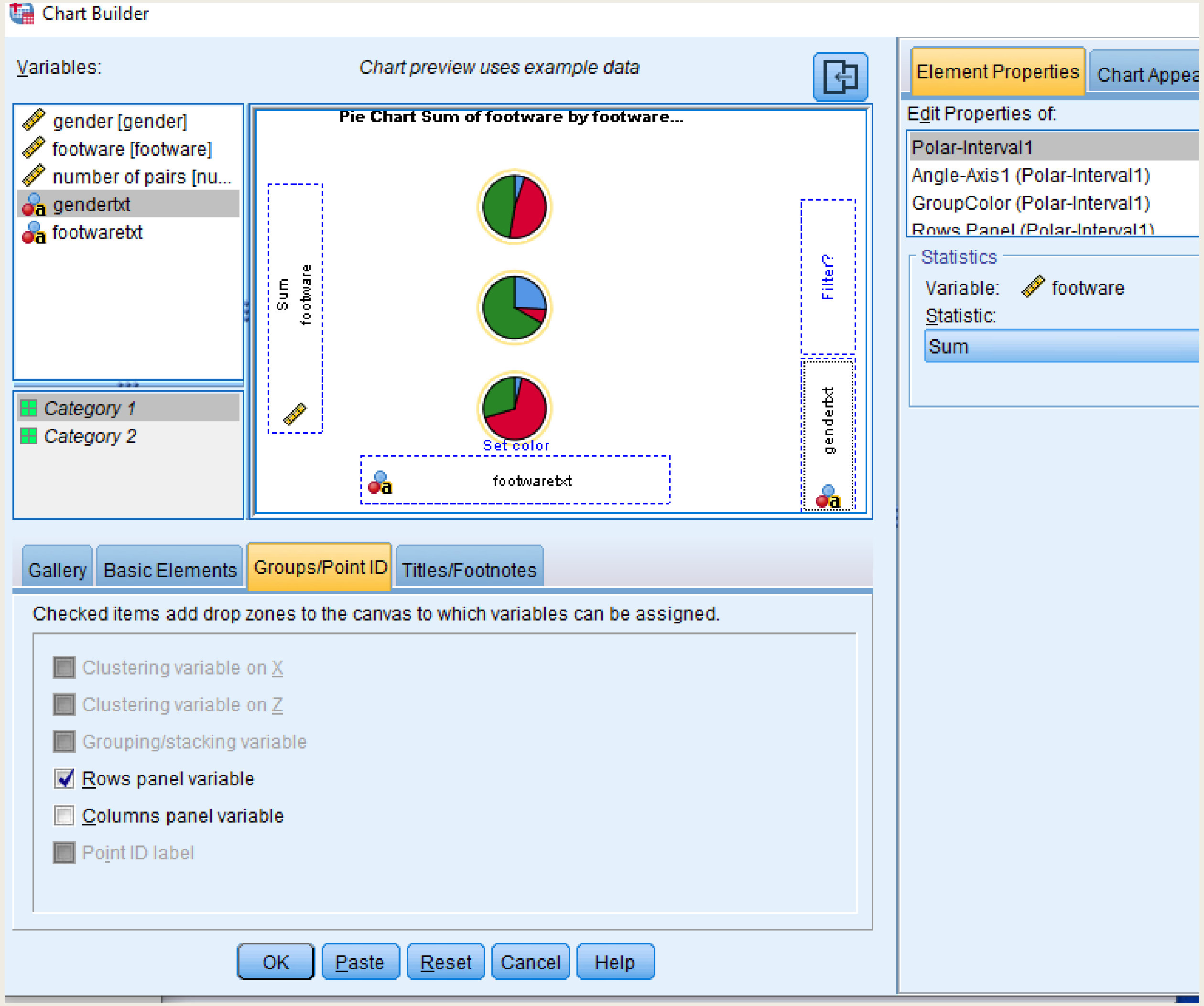
```
[DataSet11] C:\Users\Mark.Williamson.2\Desktop\SPSSDATA\Children looked after.sav
```

Pie charts

- Pie charts are ideal for showing proportions and summarizing data
- They can be made using raw data or pre-aggregated data

Summary: Pie charts, are used to show proportion, e.g. the number of votes cast for each party in an election. The pie should add up to 100% of the observed data. Pie charts should be clearly labelled and the units of measure

1. Open the file **shoetypes** in SPSS again
2. Graphs->Chartbuilder->Pie.
    1. *Put **footware** in the left bar (**footwaretext** should be in the Set color bar on bottom)*
    2. *Click the Rows panel variable, then drag **gendertxt** to the panel bar. You should get two pie charts, one for each gender, this might help identify any differences between the gender groups in their choice of shoes.*
3. Open the file **hip patient numbers** in SPSS.
    1. *This is a simplified version of the NHS hip fracture discharge data for 1997 to 1999 for England for patients aged 65 and over.*
4. Drag the “Trust Cluster” variable to the “Slice By” box, this will tell SPSS to make each slice of the pie represent one type of trust (Small/medium acute, Large acute, Very large acute, Acute teaching, Multiservice).
5. Drag the “Patient 97” variable to the “Count” box and press OK



Checked items add drop zones to the canvas to which variables can be assigned.

- ☐ Clustering variable on X
- ☐ Clustering variable on Z
- ☐ Grouping/stacking variable
- ☑ Rows panel variable
- ☐ Columns panel variable
- ☐ Point ID label

Edit Properties of: Polar-Interval1

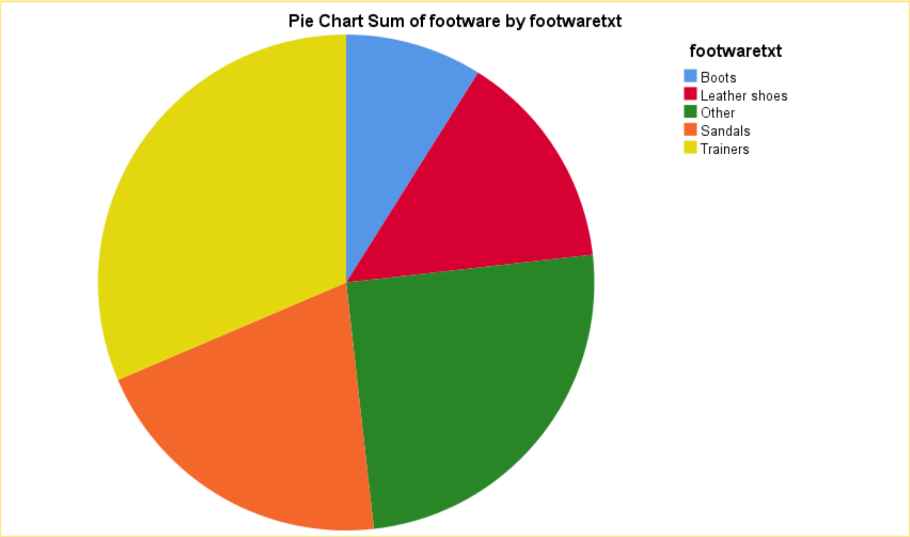
Statistics

Variable: footware

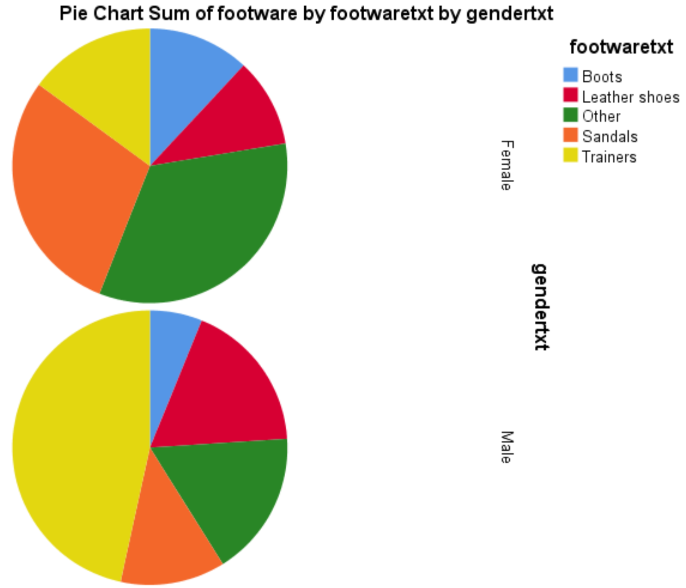
Statistic: Sum

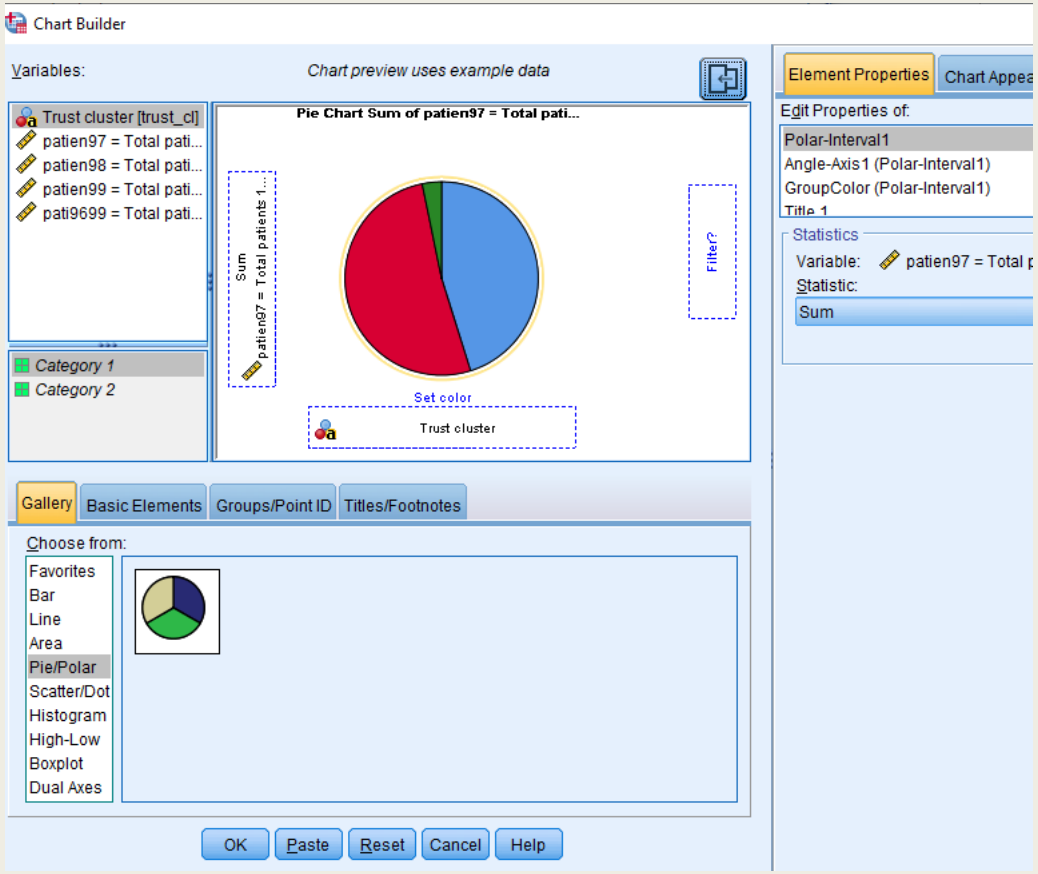
GGraph

```
[DataSet8] C:\Users\Mark.Williamson.2\Desktop\SPSSDATA\shoetypes.sav
```



GGraph

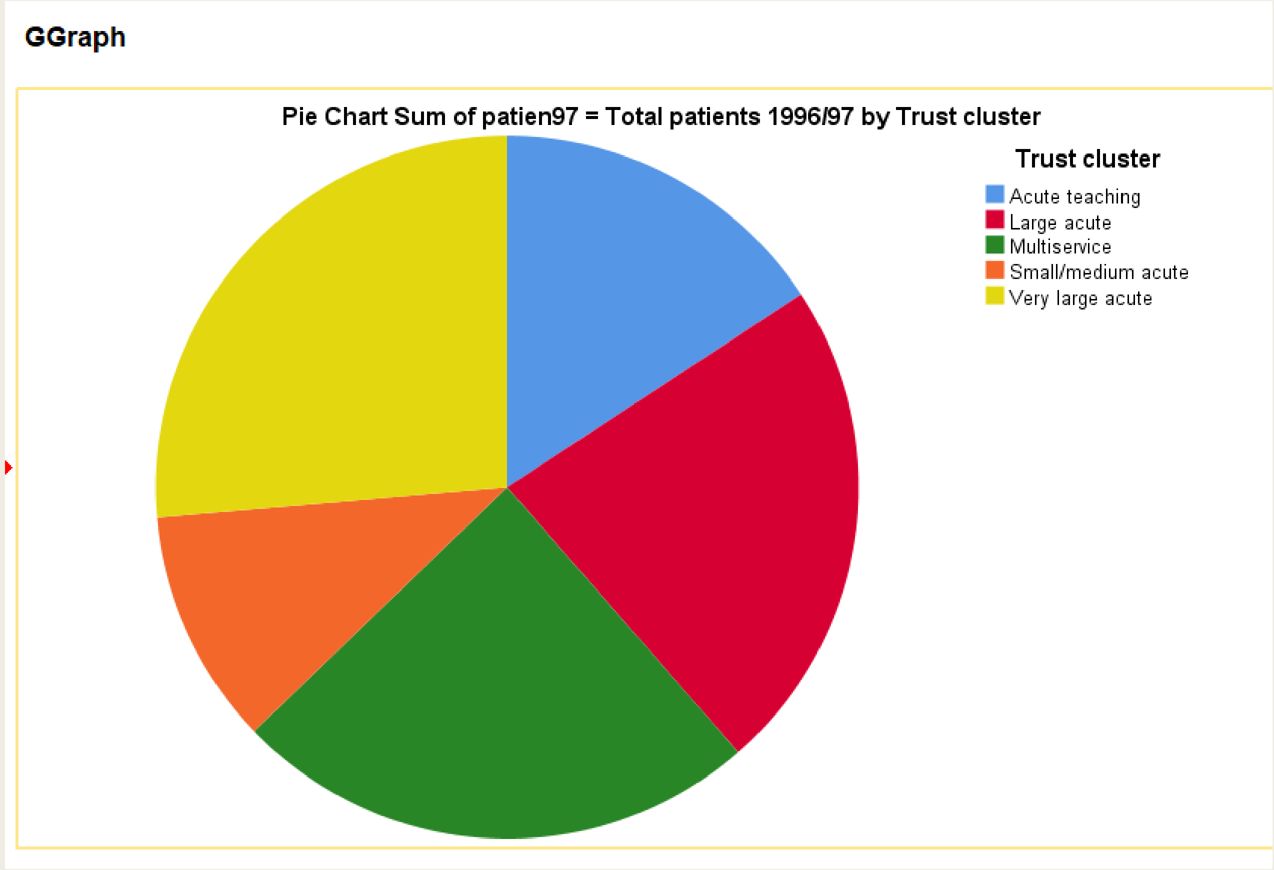




Edit Properties of: Polar-Interval1

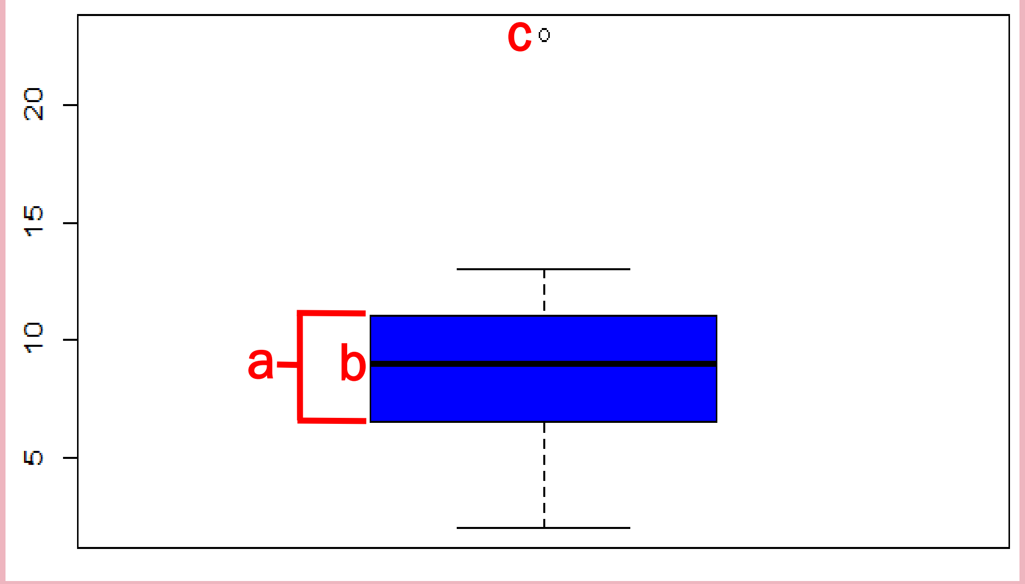
Variable: patien97 = Total p

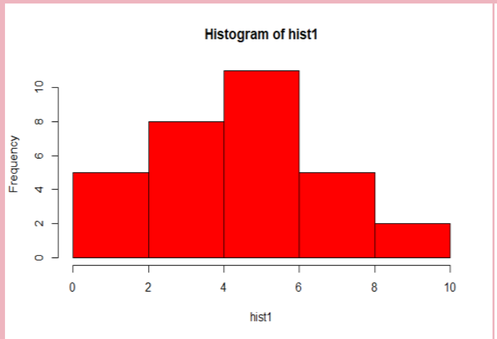
Statistic: Sum

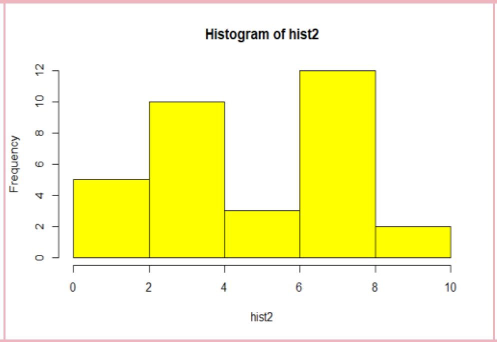


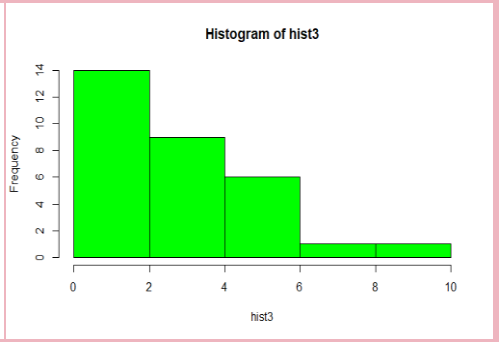
Assessment 4

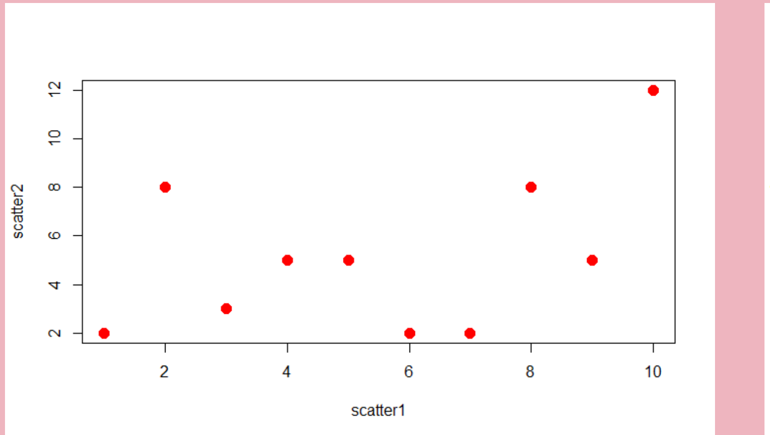
1. In the boxplot to the right, label the letters with the appropriate term
   a)
   b)
   c)
2. For the three histograms to the right, label them as parametric (normally distributed) or non-parametric
3. For the scatterplots to the right, label the correlation as:
   a) Strong, Weak, or None
   b) Positive, Negative, or None

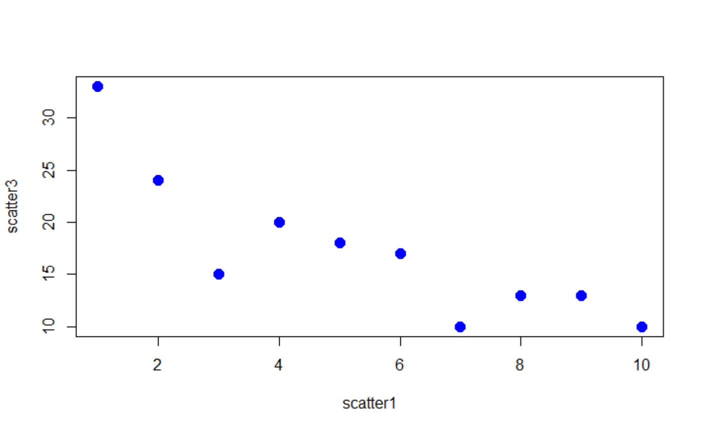












Assessment 4 Answers

1. In the boxplot to the right, label the letters with the appropriate term

   a) Interquartile Range

   b) Median

   c) Extreme Value / Outlier

2. For the three histograms to the right, label them as parametric (normally distributed) or non-parametric

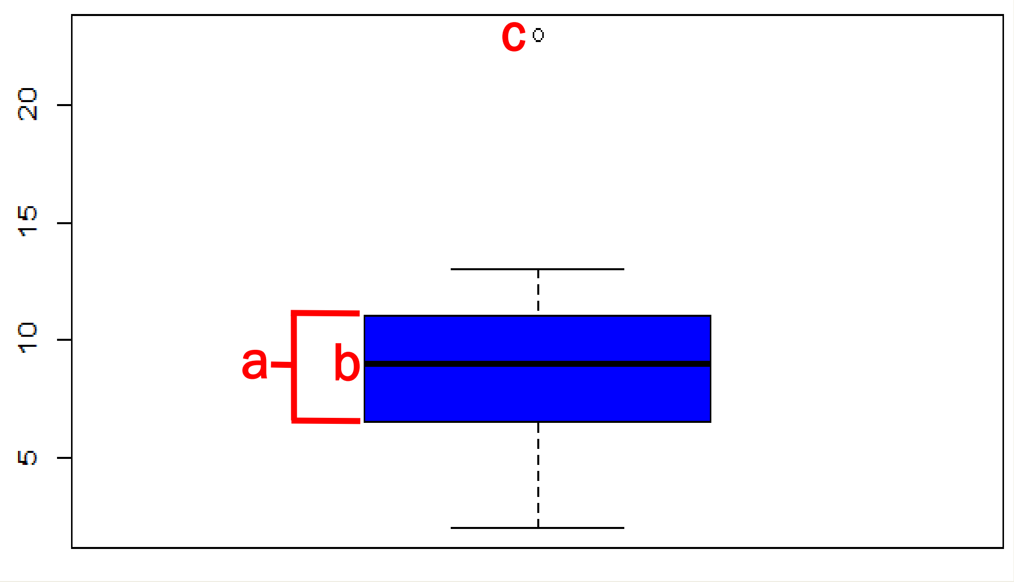
   parametric, non-parametric, non-parametric

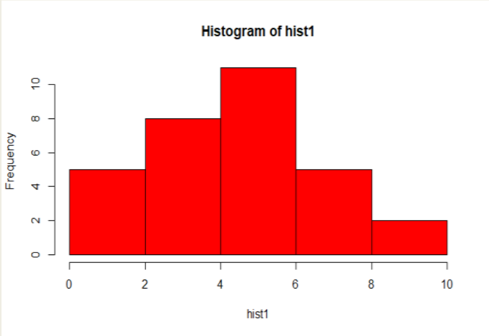
3. For the scatterplots to the right, label the correlation as:

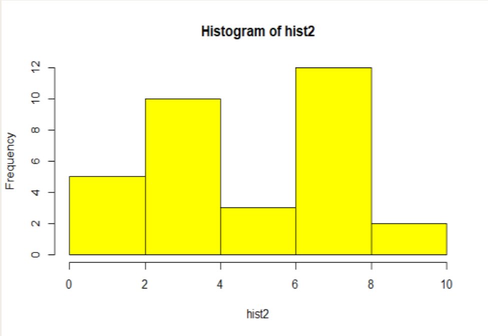
   a) Strong, Weak, or None

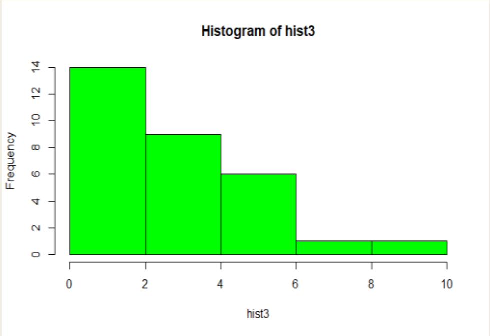
   b) Positive, Negative, or None

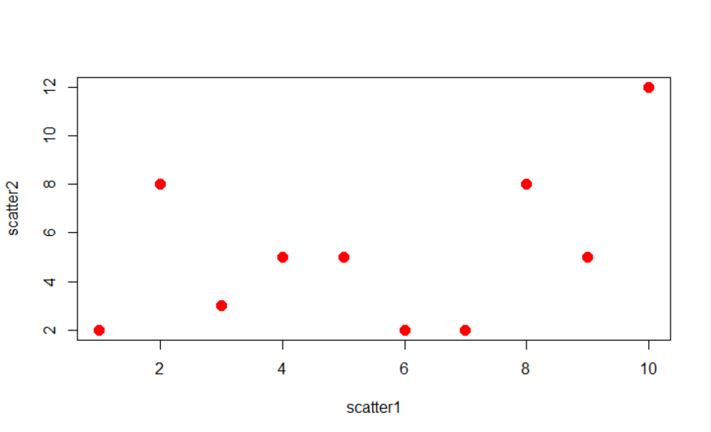
   None-None, Strong-Negative

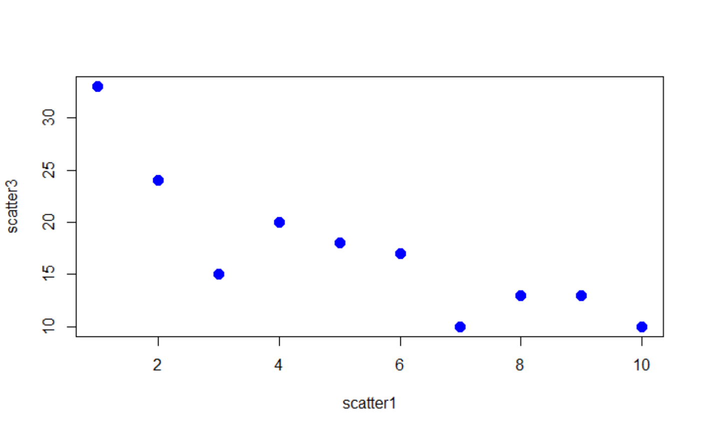












Inferential Statistics (Analyzing our Data)

- If we want to draw conclusions about an entire population from our sample, we enter the realm of inferential statistics
- This section will go over a variety of the basic tests

Part B-2

Guidelines of tests

- You ought to be interested in using statistics to make as accurate mathematical inferences about the complexities of reality to make the world a better place
- The statistics only tell you as much as you put into them, and again, they are only mathematical representations
- It is up to you to be as disciplined as possible in setting up your data and analyzing it in such a way as to best get at the truth of the world
- The following are my strong suggestions of how to go about analyzing data: think of them like football drills: you need to master the basics to be any good at answering questions with statistics

Part B-2

Rules of Analysis

A. Explore your data (outlined in first section)
   1. Look at data
   2. Identify data
   3. Graph/Describe Data
   4. Formulate Question

B. Analyze your data
   1. Set up hypothesis (null and alternative)
   2. Check normality
   3. Select and run appropriate test

C. Interpret your results
   1. Find the Test Statistic, DF, and P-value
   2. Determine if significant
   3. State if null hypothesis rejected or not
   4. Write result
   5. Present appropriate plot

Analyze your data: Set up Hypothesis

- When running a statistical test, there are two hypotheses being tested
  - ***Null Hypothesis:*** *the default, or 'boring' state*
    - Typically 'no change', 'no difference', or 'no relationship'
  - ***Alternative Hypothesis:*** *something else happening*
- Construct the two hypotheses based on your question from the data exploration step
- Example 1
  - *Question: Is there a difference between male and female shark body length?*
  - *Null Hypothesis: There is no difference in shark length by gender.*
  - *Alternative Hypothesis: There is a difference in shark length by gender.*
- Example 2
  - *Question: Is there relationship between the cups of coffee consumed during studying and exam grade?*
  - *Null Hypothesis: There is no relationship between cups of coffee and exam grade.*
  - *Alternative Hypothesis: There is a relationship between cups of coffee and exam grade.*

Analyze your data: Check Normality

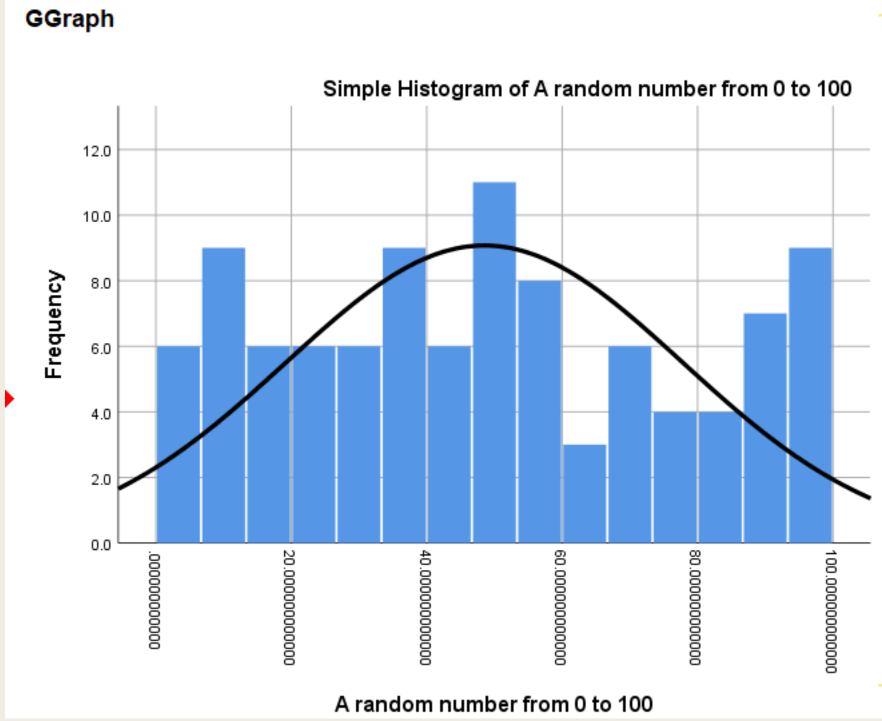
- Many tests can only be run with data that is parametric (normally distributed)
- Check normality by **histogram**, **QQ-plot**, and **test for normality**
- Usually try two or three, as each gives some different information
  - *I prefer using histograms and QQ-plots, as the test for normality is strict and most data isn't neat enough to pass*
- Determining if something is normally distributed via graph inspection is partly an art (you have to get good at looking at the graphs)

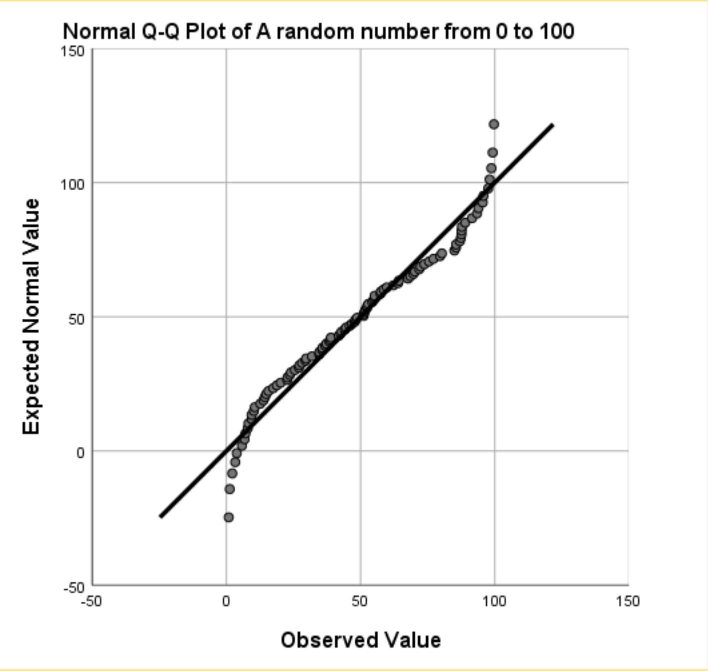
Analyze your data: Tests for Normality

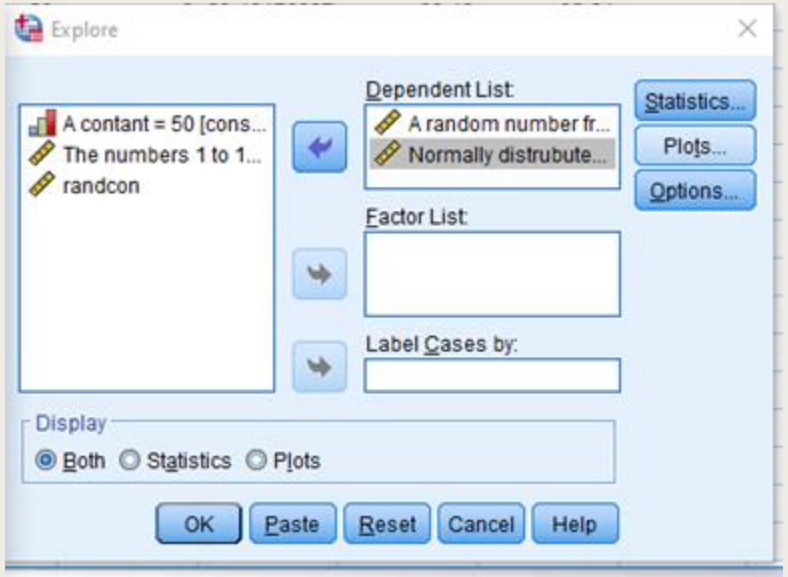
- Histogram
  - *Bars should approximate the bell curve if it is normally distributed*
  - *Doesn't have to be perfect*
- QQ plot
  - *In this plot, the normal distribution is a straight line*
  - *If normally distributed, the points should cluster around the straight line*
  - *Should not have 'tails'*
- Test of normality
  - *Statistical test*
  - *Kolmogorov-Smirnov standard*
  - *Shapiro-Wilk for small sample size*
  - *Sig. column (p-value) interpreted as if* ***more than*** *0.05, from normal distribution*
  - *If less, then* ***not*** *from normal dist.*

1. Open **tests for normality** file in SPSS
2. For Histogram
   1. *Graphs->Chartbuilder->Histogram*
   2. *Variable under investigation to the horizontal (try both Random number and Normally distributed)*
   3. *Select normal curve*
3. Q-Q plot
   1. *Analyze->Descriptive Statistics->Q-Q Plots*
   2. *Can run both variables at the same time*
4. Test for normality
   1. *Analyze->Descriptive Statistics->Explore*
   2. *Put variables to check under the Dependent list box*
   3. *Click to select Normality plots with tests*

Reminder: every test has a test statistic and a p-value. The p-value tells you if the test statistic is significant





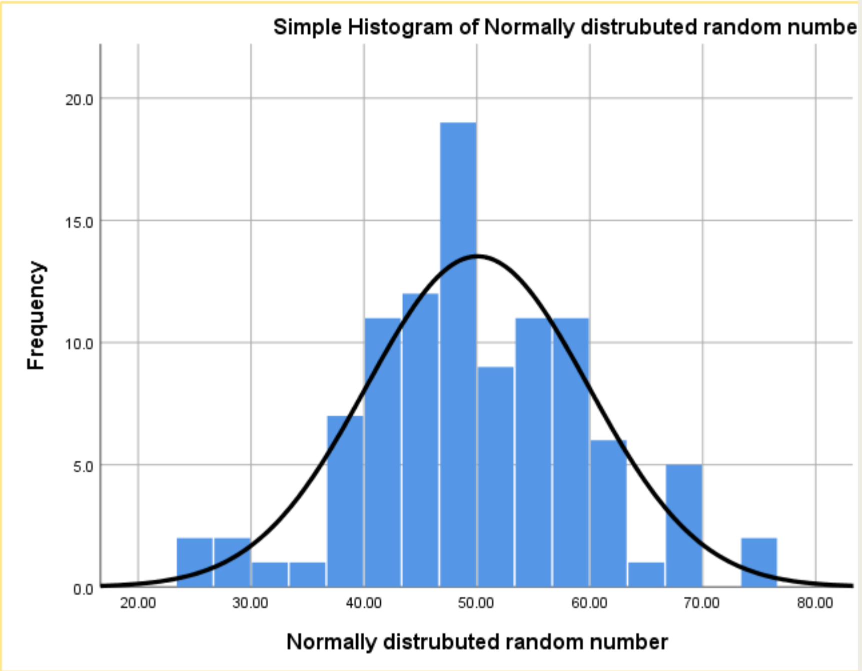


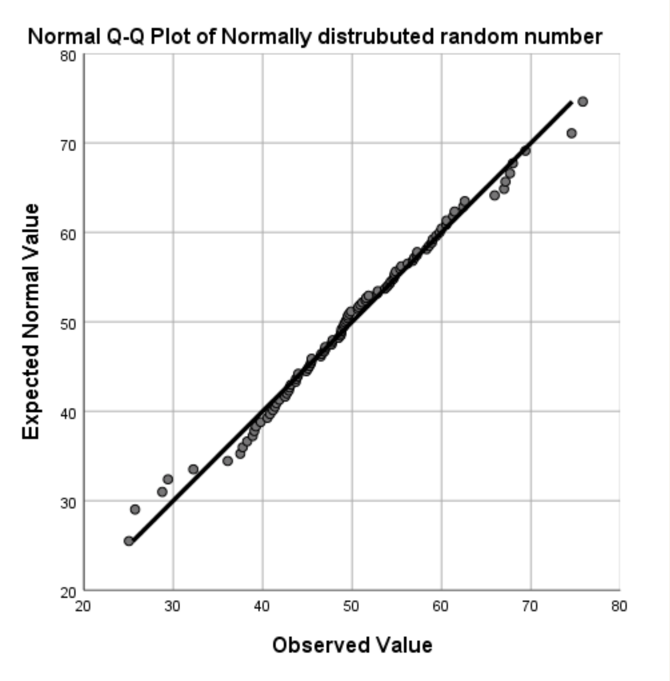
Dependent List: A random number fr... ; Normally distrubute...

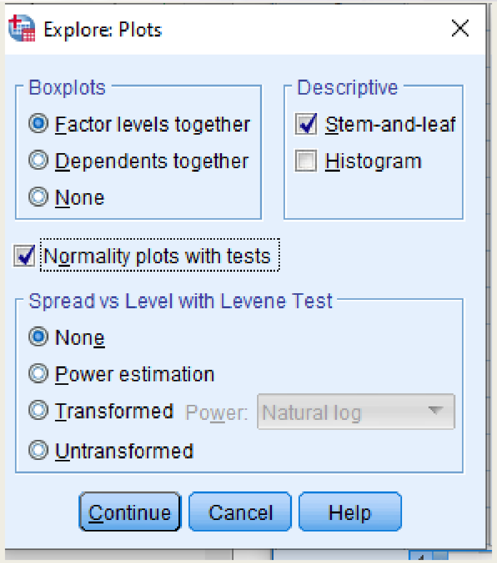
Factor List:

Label Cases by:

Display: ● Both ○ Statistics ○ Plots





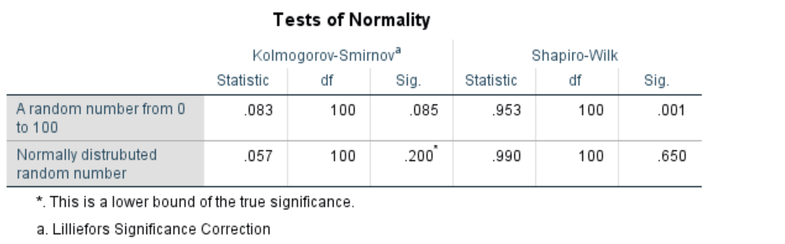


Boxplots: ● Factor levels together ○ Dependents together ○ None

Descriptive: ☑ Stem-and-leaf □ Histogram

☑ Normality plots with tests

Spread vs Level with Levene Test: ● None ○ Power estimation ○ Transformed Power: Natural log ○ Untransformed

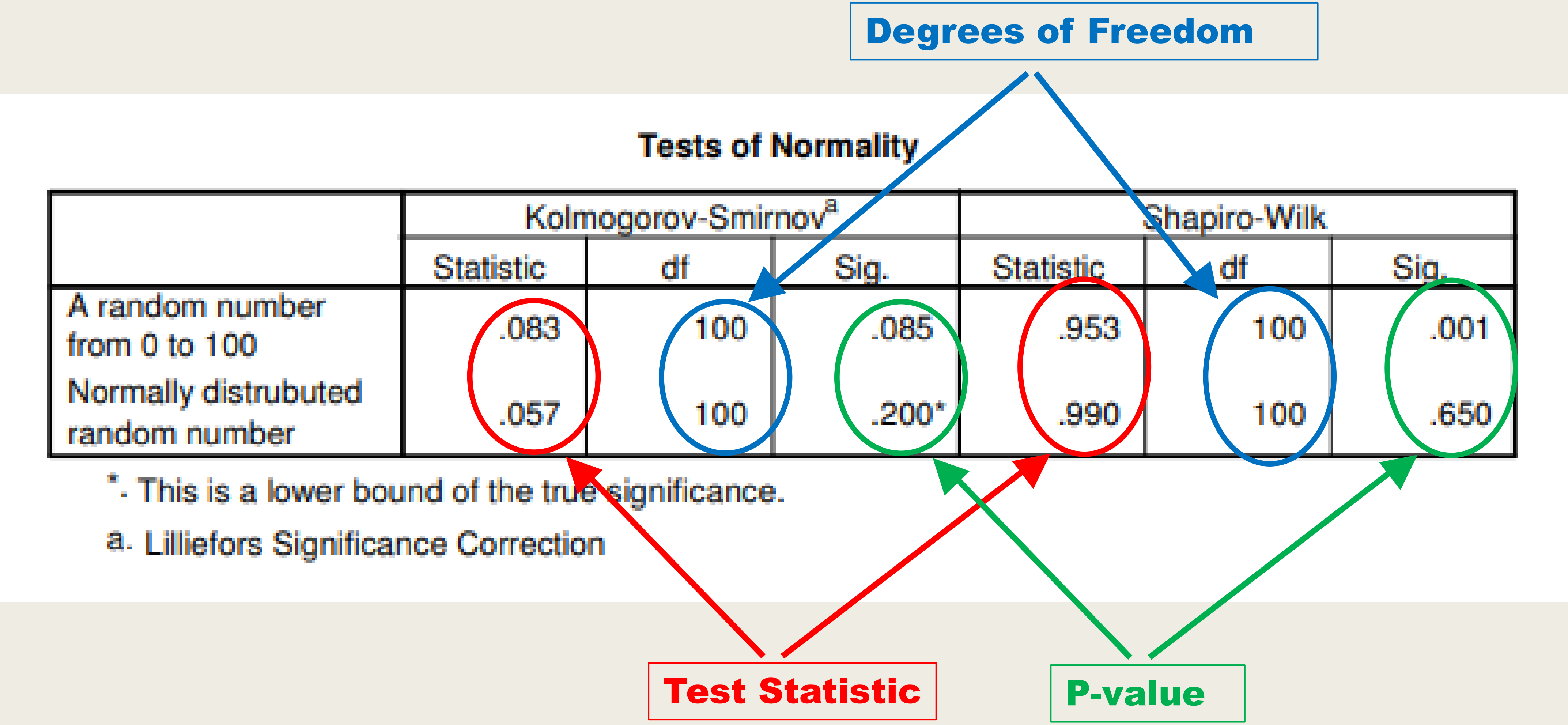


Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
A random number from 0 to 100	.083	100	.085	.953	100	.001
Normally distrubuted random number	.057	100	.200*	.990	100	.650

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
A random number from 0 to 100	.083	100	.085	.953	100	.001
Normally distrubuted random number	.057	100	.200*	.990	100	.650

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Degrees of Freedom

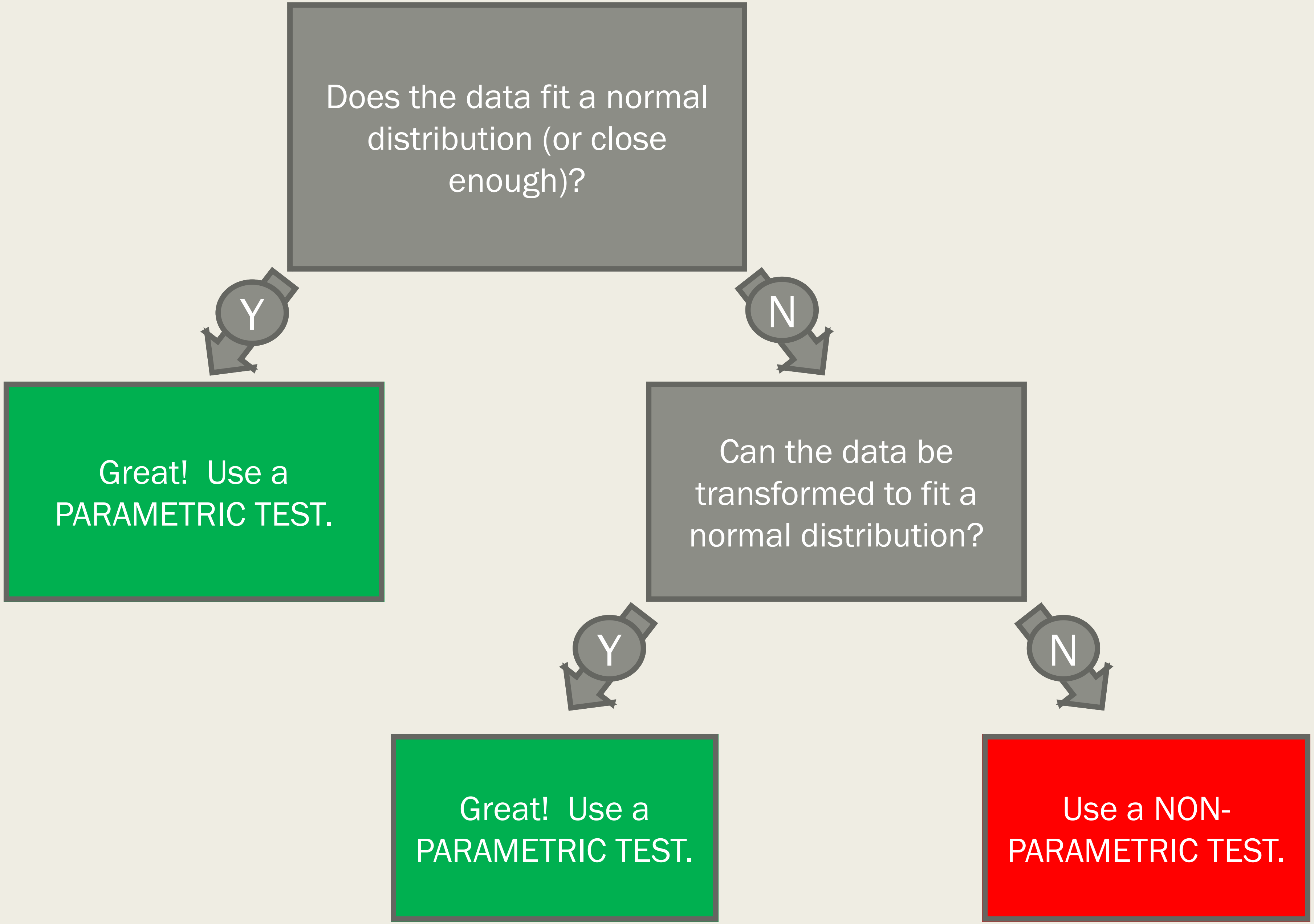
Test Statistic

P-value

Assessing normality

- If the data is not normally distributed, you can try transformations
  - *These only work for numerical data, and some only work for certain kinds of numerical*
  - *Most common:*
    - Log transformation
      - *Can't use on variables that include zero or negative numbers*
    - Square root transformation
      - *Can't use on variables that include negative numbers*
- Doing this in SPSS
  - *Can play around with Transform->Compute Variable*
  - *This can create new columns of variables based on transformations*
  - *(We won't be transforming data in this presentation, but it is useful to know in the future)*
- Then, check for normality (histogram, Q-Q plot, etc.) on the transformed data,
  - *If it looks normally distributed, use the transformed data in the analysis*
  - *If not, try another transformation*
  - *If still not, will have to use non-parametric*

Reminder: Non-parametric tests are weaker, so we only use those tests if we cannot use parametric ones



Does the data fit a normal distribution (or close enough)?

Y

Great! Use a PARAMETRIC TEST.

N

Can the data be transformed to fit a normal distribution?

Y

Great! Use a PARAMETRIC TEST.

N

Use a NON-PARAMETRIC TEST.

Analyze Your Data: Select Appropriate Test

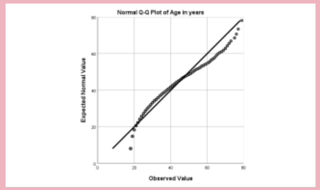
- Depends on the variables and the types of questions you want to answer
- Whether the data is numerical, categorical
- How many categories there are in the categorical variable
- Whether the data is paired or not
- Whether the data is parametric or not

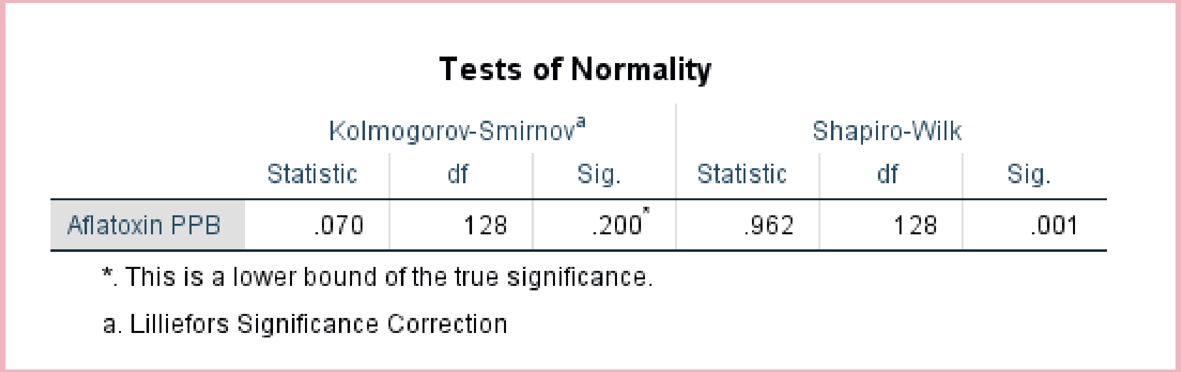
Interpret Your Results

1. Find the Test Statistic, DF, and P-value
   - *Generally picking them out of a results table*
   - *If the test does not give degrees of freedom (DF), use number of samples instead (N)*
2. Determine if significant
   - *If the p-value is below a certain threshold (usually 0.05), it is significant*
   - *If the p-value is above, it is not significant*
3. State if null hypothesis rejected or not
   - *Null hypothesis rejected if significant p-value for the test statistic*
4. Write result
   - *Answer the question stated from the data exploration*
   - *Include the test statistic, p-value, and degrees of freedom*
   - *Also include type of test*
   - *Example 1: Female sharks were significantly larger than males sharks (two-tailed T-test, F=5.67, p-value=0.0024, DF=19)*
   - *Example 2: There was no relationship between the number of cups of coffee drunk and exam score (Pearson Correlation, F=1.23, p-value=0.5863, DF=24)*
5. Present appropriate plot
   - *Don't typically plot anything for a non-significant test*
   - *Simple tests like t-tests can just have the written results; more complex analyses like correlation or ANOVA should get a plot*

Assessment 5

1. Based on the Q-Q plot to the right, would you consider the data normally distributed or not?
2. Based on the normality test to the right, would you consider the data normally distributed or not?
3. A variable in a dataset is assessed for normality and found to not be normally distributed. However, a logarithmic transformation of the data is normally distributed. Can you use a parametric test?





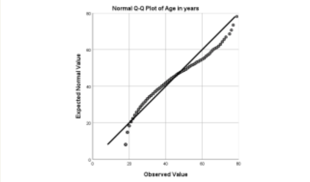
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Aflatoxin PPB	.070	128	.200*	.962	128	.001

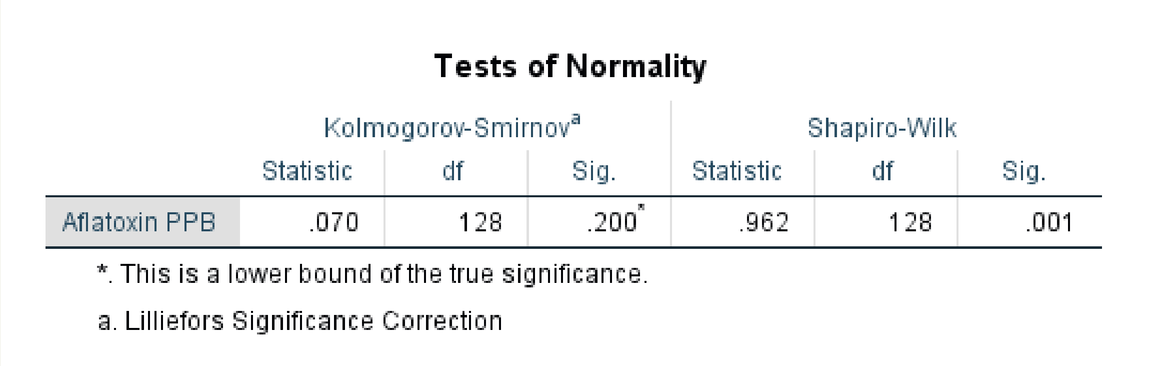
*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Assessment 5 Answers



1. Based on the Q-Q plot to the right, would you consider the data normally distributed or not?
   No, the points don't follow the straight line very well at all
2. Based on the normality test to the right, would you consider the data normally distributed or not?
   No, significant p-value means it likely does not follow a normal distribution
3. A variable in a dataset is assessed for normality and found to not be normally distributed. However, a logarithmic transformation of the data is normally distributed. Can you use a parametric test?
   Yes, on the transformed data



Tests of Normality

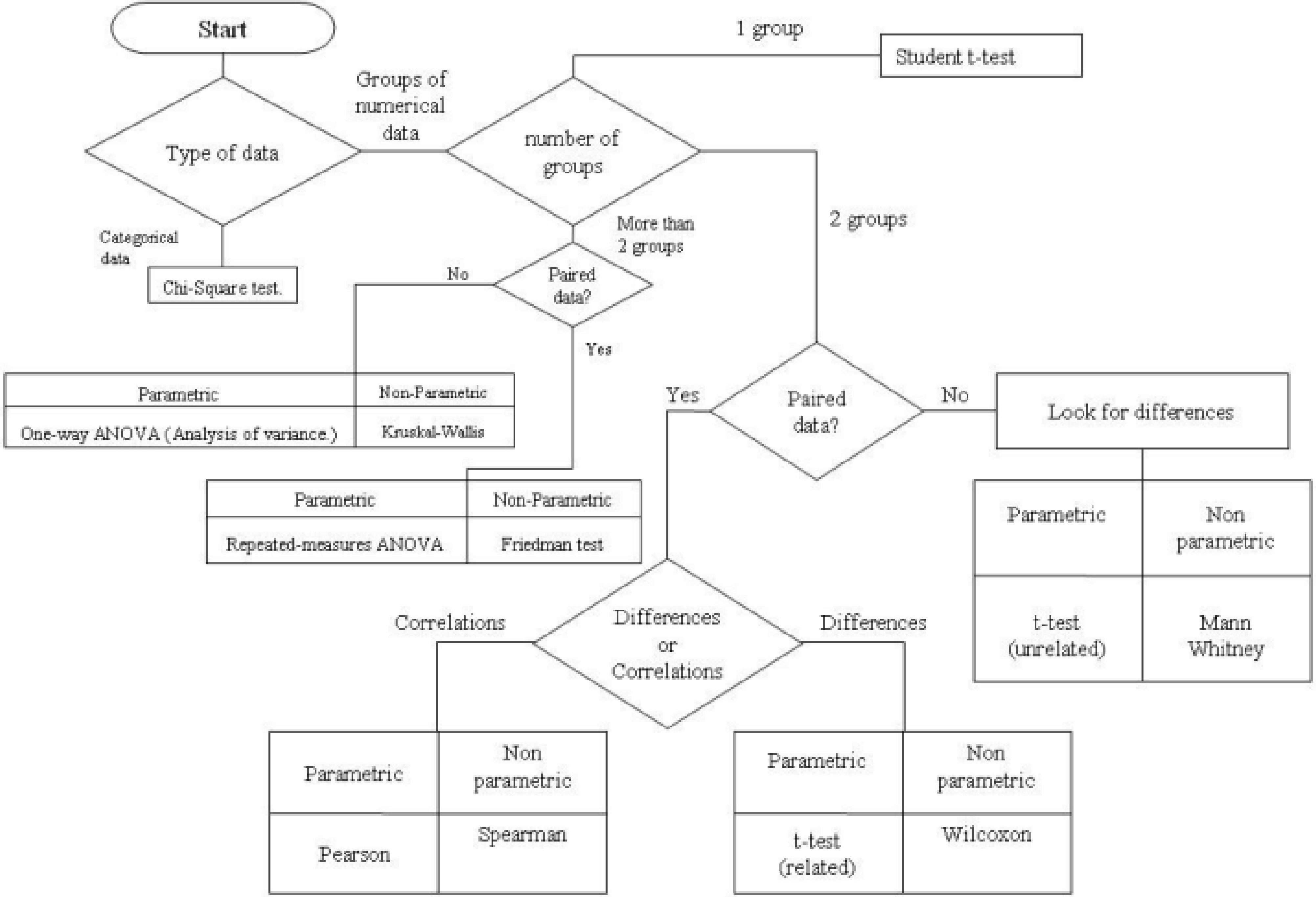
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Aflatoxin PPB	.070	128	.200[*]	.962	128	.001

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Types of tests

- 1 categorical variable + 1 numerical variable
  - Categorical variable is non-paired and group number is one:
    - One Sample T-test
  - Categorical variable is non-paired and group number is two:
    - Parametric: T-test
    - Non-parametric: Mann-Whitney Test
  - Categorical variable is paired and group number is two:
    - Parametric: Paired T-test
    - Non-parametric: Wilcoxon Signed Ranks Test
  - Categorical variable is non-paired and group number is greater than two:
    - Parametric: ANOVA
    - Non-parametric: Kruskal Wallace Test
  - Categorical variable is paired and group number is greater than two:
    - Parametric: Repeated Measures ANOVA
    - Non-parametric: Friedman test
- 2 numerical variables
  - Correlation
    - Parametric: Pearson correlation (usually)
    - Non-parametric: Spearman rank-order correlation
- 2 categorical variables
  - Chi Square test



Start

Type of data

Categorical data

Chi-Square test.

Groups of numerical data

number of groups

1 group

Student t-test

More than 2 groups

Paired data?

No

Parametric	Non-Parametric
One-way ANOVA (Analysis of variance.)	Kruskal-Wallis

Yes

Parametric	Non-Parametric
Repeated-measures ANOVA	Friedman test

2 groups

Paired data?

No

Look for differences

Parametric	Non parametric
t-test (unrelated)	Mann Whitney

Yes

Differences or Correlations

Correlations

Parametric	Non parametric
Pearson	Spearman

Differences

Parametric	Non parametric
t-test (related)	Wilcoxon

One Sample T-test

- 1 categorical variable + 1 numerical variable
  - *Categorical variable is non-paired and group number is one*
- This is when you have a single numerical variable you are interested in and want to know if it is different from some value
  - *Is the average height of basketball players greater than 6.2 feet?*
  - *Is the infant mortality rate in a certain county less than 2 death in 1000?*
  - *Is the effectiveness of treatment of a new drug any different from zero?*

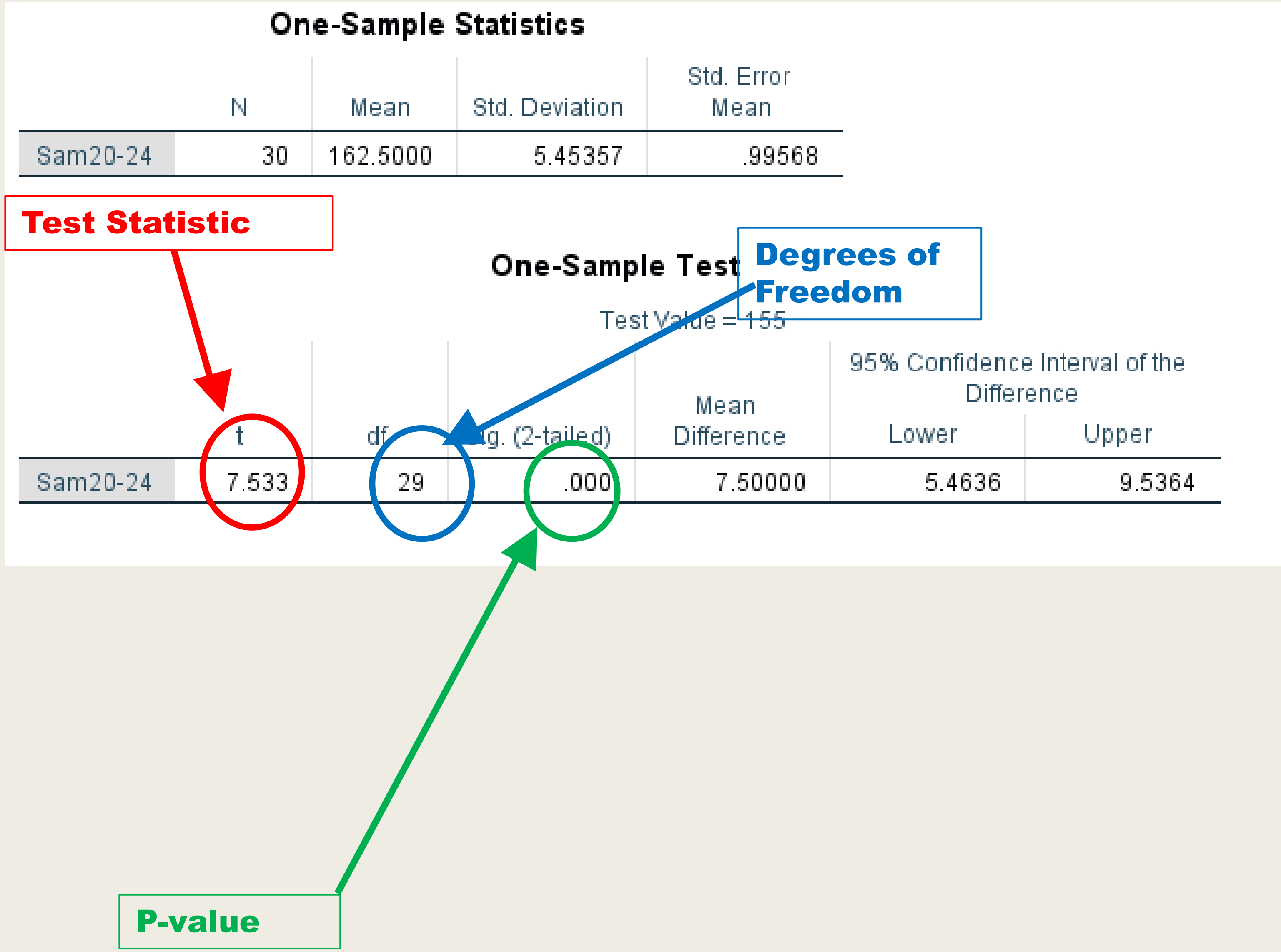
1. Explore your data
   1. *Easy, since all you have is one variable*
   2. *Histogram and maybe boxplot*
2. Check normality
   1. *Histogram, QQ-plot*
3. Set up hypothesis
   1. *Null: the variable is no different from a certain value*
   2. *Alternative: it is different*
4. Select and run appropriate test
   1. *Student's T-test*
   2. *If non-parametric, mumble, mumble Mann-Whitney*
5. Interpret results
   1. *Null rejected or failed to reject?*
   2. *What does it mean for your question*
   3. *Write it out*

Part B-2

One Sample T-test

- Example: Women's Height
  - *data on the heights of women of different ages (women, age, height).*
  - *Focus on just the first column of young women (women from ages 20-24)*
  - ***Question: Is the average height of women ages 20-24 different from 155cm?***

1. Open **waheig2** file in SPSS
2. Explore data and check normality
   1. *We know that it is for this*
3. Define null and alternative hypothesis to question.
   1. *Null= there is no difference*
   2. *Alt= there is a difference (younger women different than 155cm)*
4. Run t-test (2-sample T-test)
   1. *Analyze -> Compare Means -> One-Sample T-test*
   2. *Sam20-24 goes into the Test variable box*
   3. *Set Test value to 155*
5. Interpret Results
   1. *See next page*
   2. *What is the test statistic, degrees of freedom, and p-value?*
   3. *Did you reject or fail to reject the null?*
   4. *What does it mean for the question?*
   5. *Are young women on average different from a height of 155cm? How so?*



One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Sam20-24	30	162.5000	5.45357	.99568

One-Sample Test

	Test Value = 155					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Sam20-24	7.533	29	.000	7.50000	5.4636	9.5364

Test Statistic

Degrees of Freedom

P-value

1. **Find the Test Statistic, DF, and P-value**
   - **t=7.533**
   - **DF=29**
   - **P-value<0.0001**

2. **Determine if significant**
   - **P-value < 0.05**
   - **Significant**

3. **State if null rejected or not**
   - **Reject Null**

4. **Write result**
   - **Young women were significantly taller (mean=162.5) than the value of 155 cm (1-sample t-test, t=7.533, DF=29, p-value<0.0001).**

5. **Present appropriate plot**
   - **N/A**

Part B-2

T-test

- 1 categorical variable + 1 numerical variable
  - *Categorical variable is non-paired and group number is two*
- This is when you have a single numerical variable you are interested in and want to know if it is different between two groups
  - *Are men taller than women?*
  - *Does treatment A reduce mortality more than treatment B?*
  - *Are there more sharks attacks on the East Coast or the West Coast?*

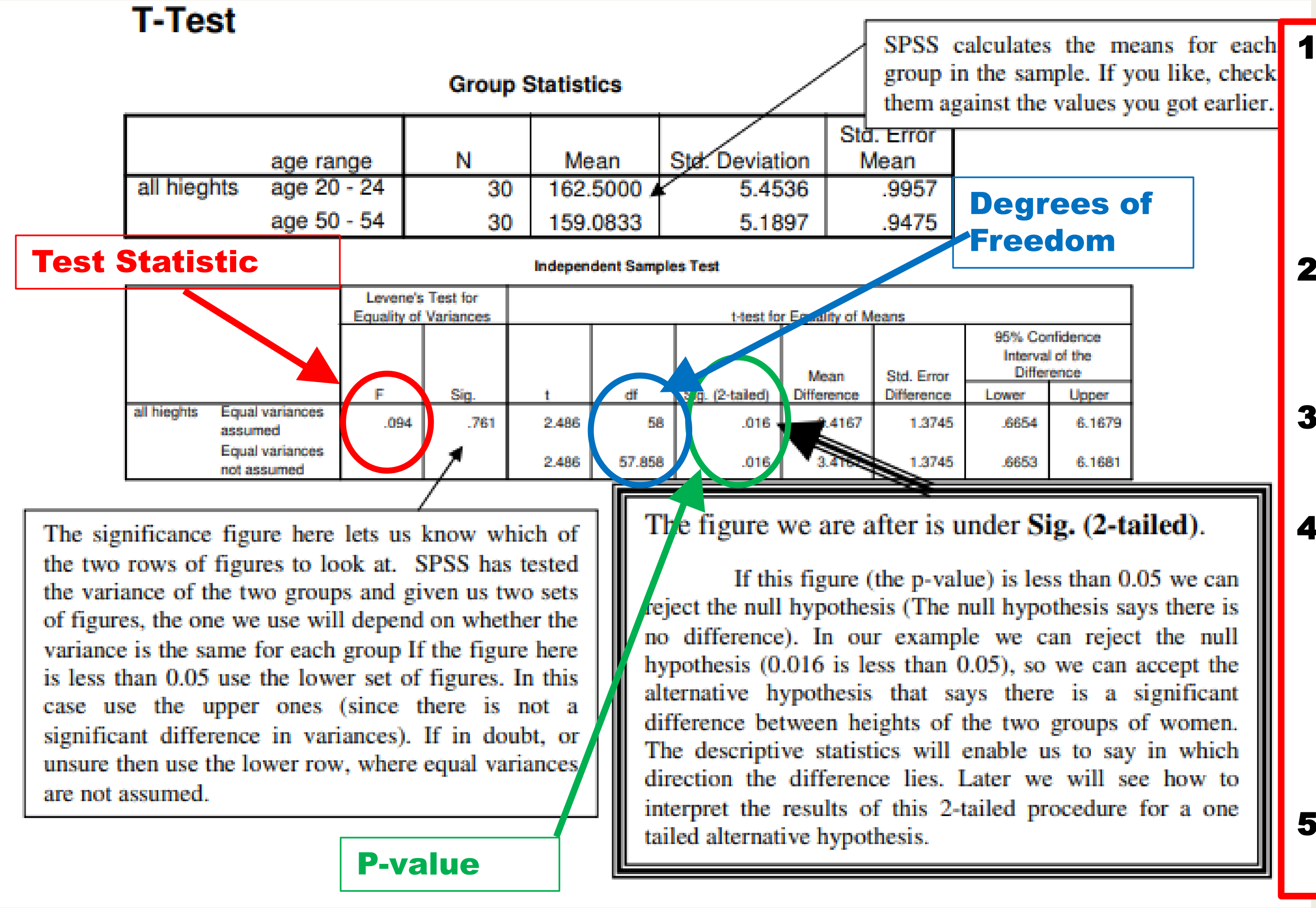
1. **Explore your data**
   1. *Histogram of numerical variable*
   2. *Boxplot of numerical variable grouped by the categorical variable*
2. **Check normality**
   1. *Histogram, QQ-plot*
3. **Set up hypothesis**
   1. *Null: there is no difference between groups*
   2. *Alternative: there is a difference*
4. **Select and run appropriate test**
   1. *Parametric: T-test*
   2. *Non-parametric: Mann-Whitney*
5. **Interpret results**
   1. *Null rejected or failed to reject?*
   2. *What does it mean for your question*
   3. *Write it out*

Part B-2

T-test: Examples

- Parametric Example: Women Height
  - *data on the heights of women of different ages (women, age, height).*
  - *This is not paired data, these are 60 different women not the same 30 measured twice with 30 years between!*
  - ***Question: Is there a difference in height between the two age groups of women?***

1. Open **waheig2S** file in SPSS
2. Explore data and check normality
   1. *We know that it is for this*
3. Define null and alternative hypothesis to question.
   1. *Null= there is no difference*
   2. *Alt= there is a difference (younger women taller)*
4. Run t-test (2-sample T-test)
   1. *Analyze -> Compare Means -> Independent-Samples T-test*
   2. *All heights goes in to Test variables*
   3. *Age range goes in to Grouping variable*
5. Interpret Results
   1. *See next page*
   2. *What is the test statistic, degrees of freedom, and p-value?*
   3. *Did you reject or fail to reject the null?*
   4. *What does it mean for the question?*
   5. *Are the two groups of women different?*
   6. *If so, how are they different (which group is taller?)*



T-Test

Group Statistics

	age range	N	Mean	Std. Deviation	Std. Error Mean
all hieghts	age 20 - 24	30	162.5000	5.4536	.9957
	age 50 - 54	30	159.0833	5.1897	.9475

SPSS calculates the means for each group in the sample. If you like, check them against the values you got earlier.

Degrees of Freedom

Test Statistic

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
all hieghts	Equal variances assumed	.094	.761	2.486	58	.016	3.4167	1.3745	.6654	6.1679
	Equal variances not assumed			2.486	57.858	.016	3.4167	1.3745	.6653	6.1681

The significance figure here lets us know which of the two rows of figures to look at. SPSS has tested the variance of the two groups and given us two sets of figures, the one we use will depend on whether the variance is the same for each group If the figure here is less than 0.05 use the lower set of figures. In this case use the upper ones (since there is not a significant difference in variances). If in doubt, or unsure then use the lower row, where equal variances are not assumed.

The figure we are after is under **Sig. (2-tailed)**.

If this figure (the p-value) is less than 0.05 we can reject the null hypothesis (The null hypothesis says there is no difference). In our example we can reject the null hypothesis (0.016 is less than 0.05), so we can accept the alternative hypothesis that says there is a significant difference between heights of the two groups of women. The descriptive statistics will enable us to say in which direction the difference lies. Later we will see how to interpret the results of this 2-tailed procedure for a one tailed alternative hypothesis.

P-value

1. **Find the Test Statistic, DF, and P-value**
   - **F=0.094**
   - **DF=58**
   - **P-value=0.016**
2. **Determine if significant**
   - **P-value < 0.05**
   - **Significant**
3. **State if null rejected or not**
   - **Reject the null**
4. **Write result**
   - **Younger women, age range of 20-24, are significantly taller than older women of an age range of 50-54 (2-tailed T-test, F=0.094, DF=58, P-value=0.016).**
5. **Present appropriate plot**
   - **N/A**

Interpreting Results

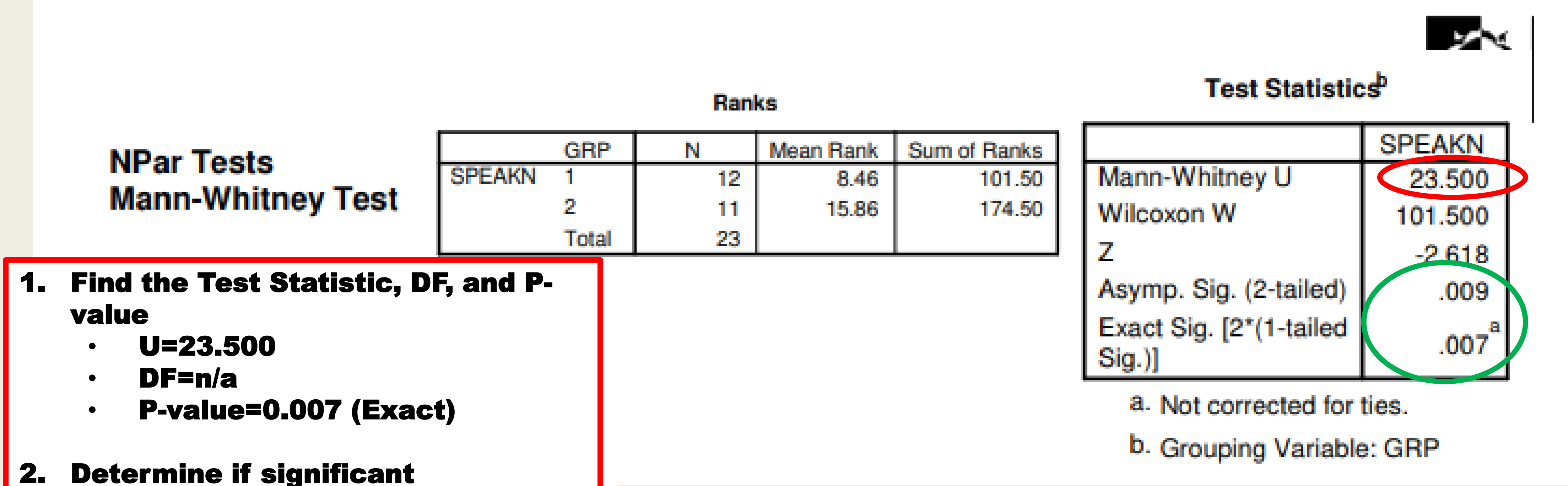
- Look at table
- Find the various values:
  - *Test statistic:*
  - *Degree of freedom or number of samples*
  - *P-value: tells you whether to reject or fail to reject the null hypothesis*
    - Typically, the cutoff value is 0.05
    - So, if value is <0.05, Reject the Null (and Retain the Alternative)
    - If value >0.05, Fail to Reject the Null, so retain the Null over the Alternative
      - *(another way of saying, "**based on our statistical test, there is no evidence that reality is anything other than the null hypothesis**")*

Part B-2

T-test: Examples 2

- Non-Parametric Example: Student Contribution
  - *The file has all the numbers representing the number of times each student contributed in the variable called “speakn” and the age group in the variable called “grp”*
  - *Each row of this data represents a student, the number in the “speakn” column is the amount they contributed and the number in the “grp” column tells us their age and year grouping.*
  - *The middle column is just some text to help you see which group is which, if you go to variable view you will see the “grp” variable labels similar to the ones explained in the previous task*
  - ***Question: Do mature first year students contribute more than young first year students?***

1. Open **studentsss** file in SPSS
2. Explore data and check normality
   1. *Should not be normally distributed*
3. Define null and alternative hypothesis to question.
   1. *Fill out yourself*
4. Run t-test (Mann-Whitney)
   1. *Analyze -> Nonparametric Tests-> Legacy Dialoges->2 independent samples*
   2. ***Speakn** goes in the Test Variable*
   3. ***Age** goes in the Grouping Variable*
      1. Need to define groups (1=Year1 young, 2=Year1 mature)
   4. *Make sure the Mann-Whitney test is ticked (under Test Type)*
5. Interpret Results
   1. *See next page*
   2. *What is the test statistic, degrees of freedom, and p-value?*
   3. *Did you reject or fail to reject the null?*
   4. *What does it mean for the question?*
   5. *Are the two groups of students different? If so, how?*



NPar Tests
Mann-Whitney Test

Ranks

	GRP	N	Mean Rank	Sum of Ranks
SPEAKN	1	12	8.46	101.50
	2	11	15.86	174.50
	Total	23		

Test Statistics[b]

	SPEAKN
Mann-Whitney U	23.500
Wilcoxon W	101.500
Z	-2.618
Asymp. Sig. (2-tailed)	.009
Exact Sig. [2*(1-tailed Sig.)]	.007[a]

a. Not corrected for ties.

b. Grouping Variable: GRP

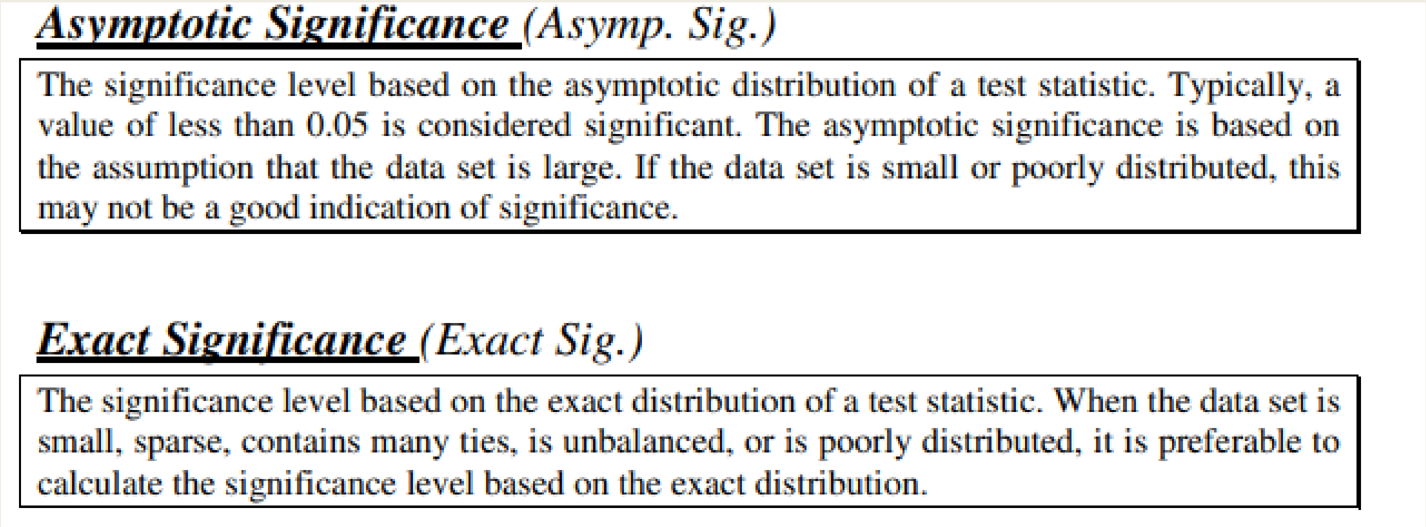
1. **Find the Test Statistic, DF, and P-value**
   - **U=23.500**
   - **DF=n/a**
   - **P-value=0.007 (Exact)**

2. **Determine if significant**
   - **P-value < 0.05**
   - **Significant**

3. **State if null rejected or not**
   - **Reject the null**

4. **Write result**
   - **Mature students (group 2) spoke significantly more than young students (Mann-Whitney Test, U=23.500, N=23, P-value=0.007 with exact significance).**

5. **Present appropriate plot**
   - **N/A**



Asymptotic Significance *(Asymp. Sig.)*

The significance level based on the asymptotic distribution of a test statistic. Typically, a value of less than 0.05 is considered significant. The asymptotic significance is based on the assumption that the data set is large. If the data set is small or poorly distributed, this may not be a good indication of significance.

Exact Significance *(Exact Sig.)*

The significance level based on the exact distribution of a test statistic. When the data set is small, sparse, contains many ties, is unbalanced, or is poorly distributed, it is preferable to calculate the significance level based on the exact distribution.

Part B-2

Paired T-test

- 1 categorical variable + 1 numerical variable
  - *Categorical variable is paired and group number is two*
- This is when you have a single numerical variable you are interested in and want to know if it is different between two groups, but the groups have a integral relationship (they are paired)
- Often 'before' and 'after' type data
  - *Is heartrate different before and after exercise*
  - *Is the number of bear attacks in parks lower after preventative measures have been implemented?*
  - *Does treatment reduce symptoms?*

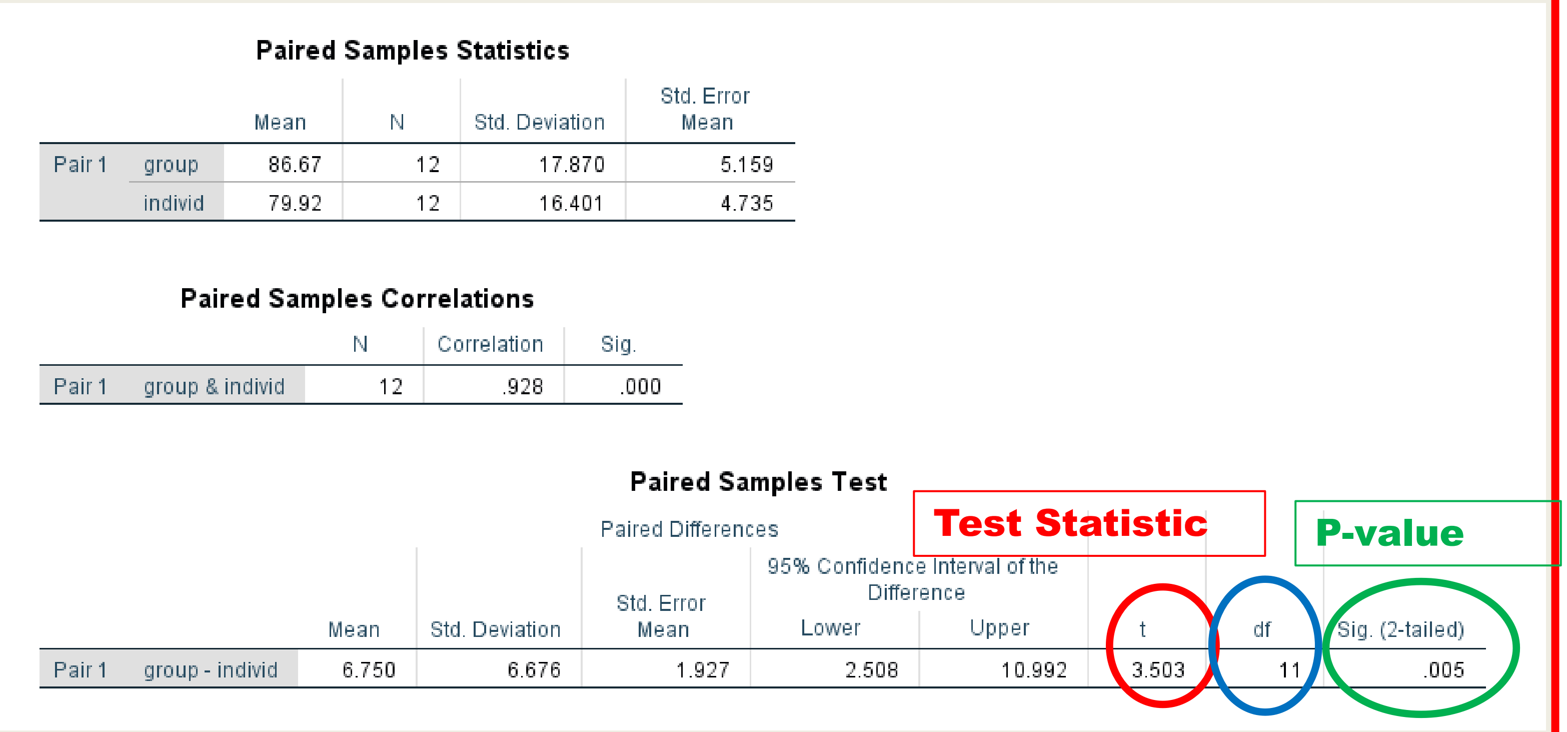
1. Explore your data
   1. *Histogram of numerical variable*
   2. *Boxplot of numerical variable grouped by the categorical variable*
2. Check normality
   1. *Histogram, QQ-plot*
3. Set up hypothesis
   1. *Null: there is no difference between groups*
   2. *Alternative: there is a difference*
4. Select and run appropriate test
   1. *Parametric: Paired T-test*
   2. *Non-parametric: Wilcoxon test*
5. Interpret results
   1. *Null rejected or failed to reject?*
   2. *What does it mean for your question*
   3. *Write it out*

Part B-2

Paired T-test: Examples

- Parametric/Non-par Example: Student Steps
  - *data in this file come from an experiment to see whether subjects could perform more step exercises in a fixed time in a group or on their own*
  - *Paired data often occur in 'before and after' situations. They are also known as 'related samples'. These data are paired, it's the same person doing step exercises under two different conditions.*
  - ***Question: Is there a difference in the number of exercises completed in a fixed time for students alone versus in a group?***

1. Open **Step** file in SPSS
2. Explore data and check normality
   1. *Whether it is normally distributed or not, try both ways.*
3. Define null and alternative hypothesis to question.
4. Run Both Parametric and Non-Parametric
   1. *Parametric: Analyze->Compare Means->Paired Samples T-test*
      1. Both **group** and **individ** goes in Test Pair list
   2. *Non-Parametric: Analyze-> Nonparametric Tests -> 2 related samples*
      1. Both **group** and **individ** goes in Test Pair list
      2. Make sure Wilcoxon is selected in Test Types
      3. Also include Descriptive Statistics
5. Interpret Results
   1. *See next page*
   2. *What is the test statistic, degrees of freedom, and p-value?*
   3. *Did you reject or fail to reject the null?*
   4. *What does it mean for the question?*
   5. *Are the two groups different?*



Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	group	86.67	12	17.870	5.159
	individ	79.92	12	16.401	4.735

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	group & individ	12	.928	.000

Paired Samples Test

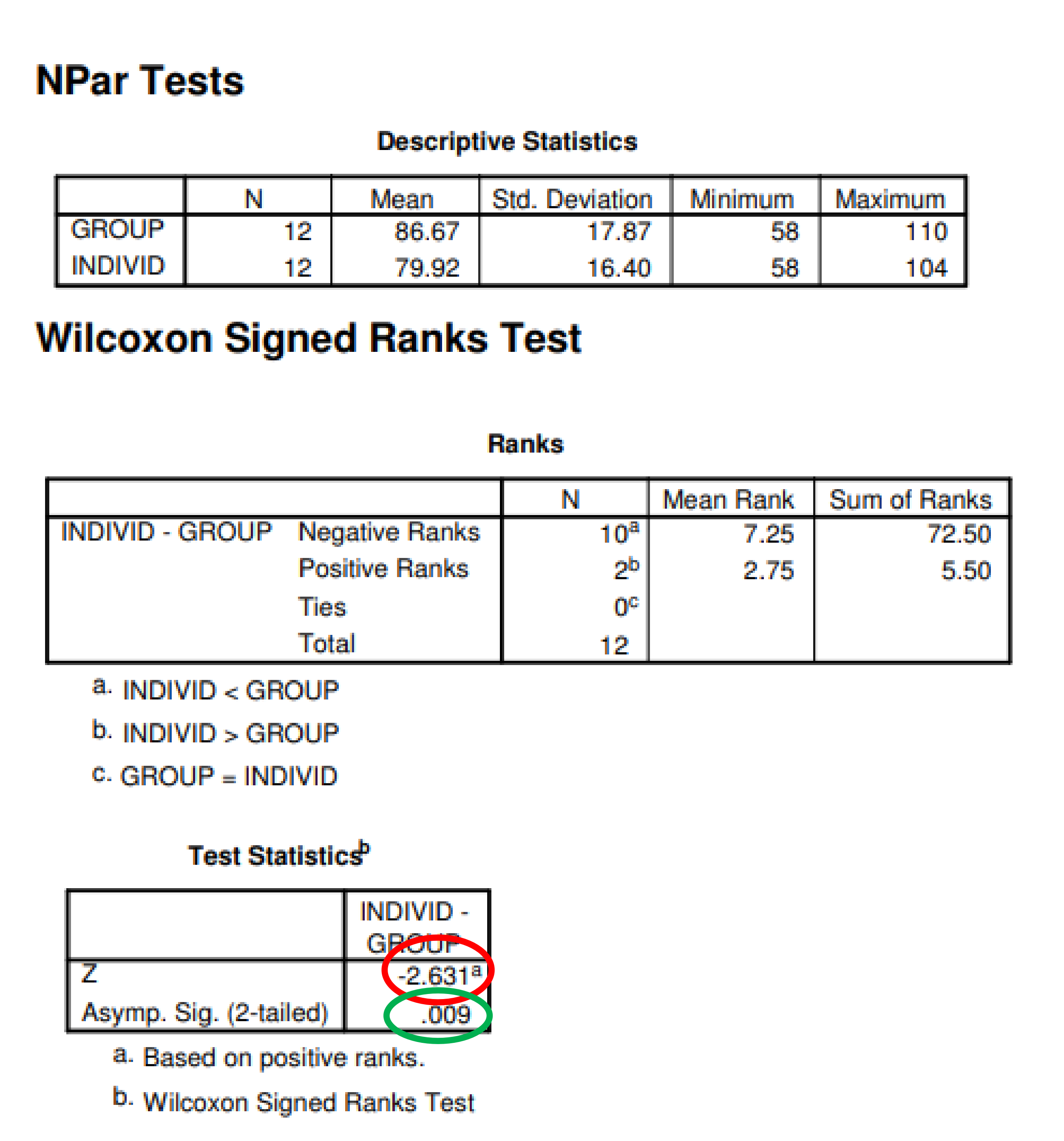
		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	group - individ	6.750	6.676	1.927	2.508	10.992	3.503	11	.005

Test Statistic

P-value

Degrees of Freedom

1. **Find the Test Statistic, DF, and P-value**
   - **t=3.503**
   - **DF=11**
   - **P-value=0.005**
2. **Determine if significant**
   - **P-value < 0.05**
   - **Significant**
3. **State if null rejected or not**
   - **Reject the null**
4. **Write result**
   - **Subjects had a tendency to complete more steps under group conditions than under**
   - **individual conditions. (Paired Samples T-test, t=3.503, DF=11, p-value=0.005).**
5. **Present appropriate plot**
   - **N/A**



NPar Tests

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
GROUP	12	86.67	17.87	58	110
INDIVID	12	79.92	16.40	58	104

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
INDIVID - GROUP	Negative Ranks	10[a]	7.25	72.50
	Positive Ranks	2[b]	2.75	5.50
	Ties	0[c]		
	Total	12		

[a] INDIVID < GROUP

[b] INDIVID > GROUP

[c] GROUP = INDIVID

Test Statistics[b]

	INDIVID - GROUP
Z	-2.631[a]
Asymp. Sig. (2-tailed)	.009

[a] Based on positive ranks.

[b] Wilcoxon Signed Ranks Test

1. **Find the Test Statistic, DF, and P-value**
   - **Z=-2.631**
   - **DF=n/a**
   - **P-value=0.002**

2. **Determine if significant**
   - **P-value < 0.05**
   - **Significant**

3. **State if null rejected or not**
   - **Reject the null**

4. **Write result**
   - **Subjects had a tendency to complete more steps under group conditions than under**
   - **individual conditions. (2-tailed Wilcoxon signed ranks test, Z=-2.631, n=24, p = 0.009).**

5. **Present appropriate plot**
   - **N/A**

Part B-2

Correlation

- 2 numerical variables
- This is when you have two numerical variables and you want to see if there is a relationship between the two (positive, negative)
  - *Is there a relationship between drug concentration and inflammation level?*
  - *Is there a relationship between length and weight in trout?*
  - *Is there a relationship between the number of guns and violent crime rates in a city?*
- Remember, correlation is not causation

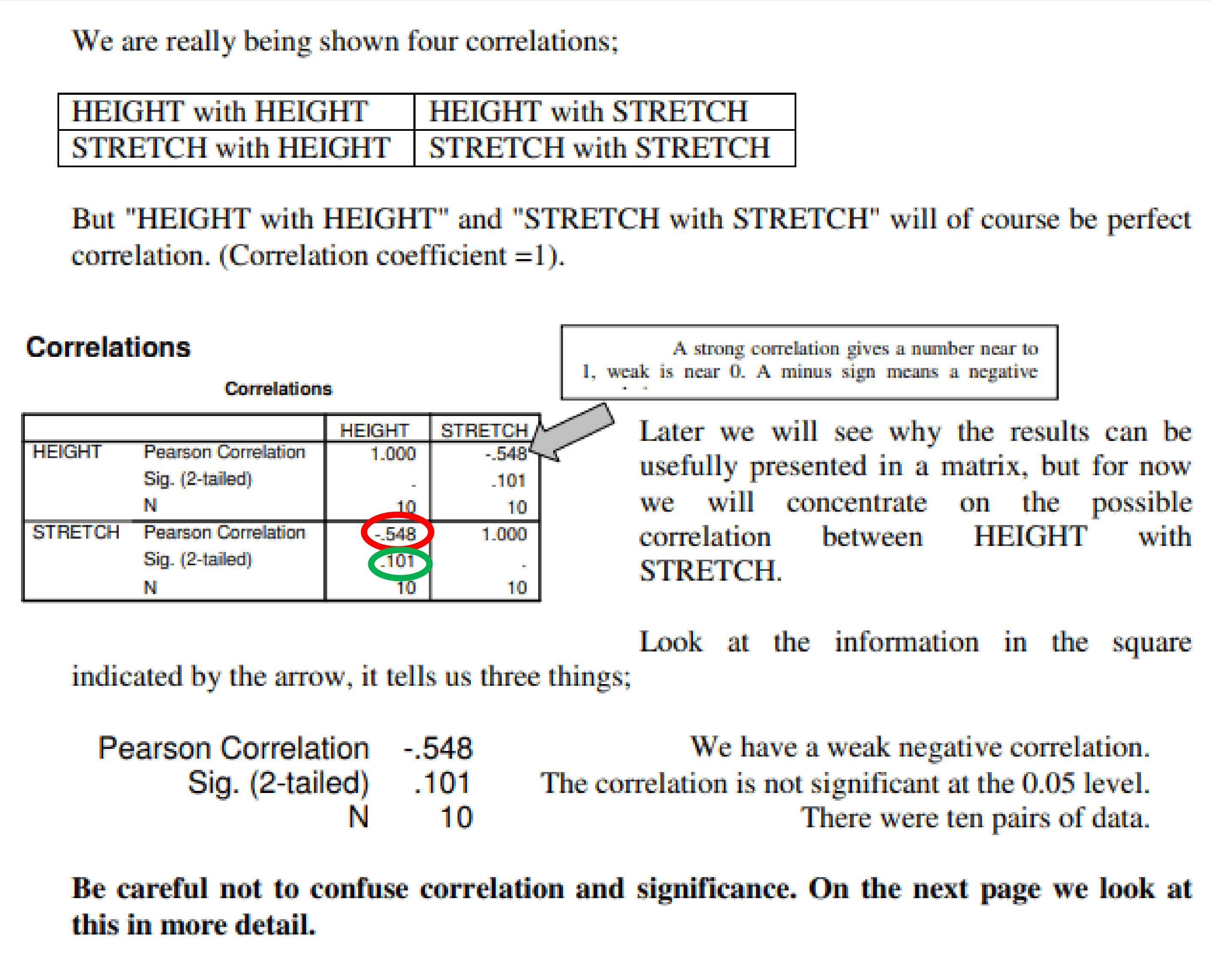
1. Explore your data
   1. *Determine which variable is dependent and which is independent*
   2. *Histogram of dependent variable*
   3. *Scatterplot (independent on x-axis, dependent on y-axis)*
2. Check normality
   1. *Histogram, QQ-plot, of dependent variable*
3. Set up hypothesis
   1. *Null: there is no relationship between the variables*
   2. *Alternative: there is a relationship*
4. Select and run appropriate test
   1. *Parametric: Pearson Correlation*
   2. *Non-parametric: Spearman Rank-correlation*
5. Interpret results
   1. *Null rejected or failed to reject?*
   2. *What does it mean for your question*
   3. *Write it out*

Part B-2

Correlation: Examples

- Parametric Example: Women Height
  - *file contains data from a student project on the effect of heat on hip stretches.*
  - *The first column gives the subject's height, and the second column gives 54 the increase in hip extension after stretching exercises.*
  - *(Other columns relate to the discomfort experienced, and the stretch and discomfort when heat is used; for our purposes those are nuisance variables)*
  - *This is paired data (measurements taken under two conditions)*
  - ***Question: Is there a relationship between the subject's stretch increase and height***

1. Open **Heathip** file in SPSS
2. Explore data and check normality
   1. *Determine which is dependent and which is independent*
   2. *Whether it is normally distributed or not*
   3. *Plot scatterplot (height on x-axis and stretch (without heat) on y-axis): Graphs->Interactive Scatterplot*
3. Define null and alternative hypothesis to question.
4. Run Appropriate test (Try both)
   1. *Parametric*
      1. Analyze->Correlate->Bivariate
      2. **Height** and **stretch** go in Variable
      3. Make sure Pearson is checked under Correlation Coefficients
      4. Also check that Two-Tailed is set up and Flag significant correlations
   2. *Non-Parametric: same thing but check "Spearman" instead of Pearson*
5. Interpret Results
   1. *See next page*
   2. *What is the test statistic, degrees of freedom, and p-value?*
   3. *Did you reject or fail to reject the null?*
   4. *What does it mean for the question?*
   5. *Is there a relationship? If so, what type (positive/negative) and how strong?*



We are really being shown four correlations;

HEIGHT with HEIGHT	HEIGHT with STRETCH
STRETCH with HEIGHT	STRETCH with STRETCH

But "HEIGHT with HEIGHT" and "STRETCH with STRETCH" will of course be perfect correlation. (Correlation coefficient =1).

Correlations

A strong correlation gives a number near to 1, weak is near 0. A minus sign means a negative

Correlations

		HEIGHT	STRETCH
HEIGHT	Pearson Correlation	1.000	-.548
	Sig. (2-tailed)	.	.101
	N	10	10
STRETCH	Pearson Correlation	-.548	1.000
	Sig. (2-tailed)	.101	.
	N	10	10

Later we will see why the results can be usefully presented in a matrix, but for now we will concentrate on the possible correlation between HEIGHT with STRETCH.

Look at the information in the square indicated by the arrow, it tells us three things;

Pearson Correlation	-.548	We have a weak negative correlation.
Sig. (2-tailed)	.101	The correlation is not significant at the 0.05 level.
N	10	There were ten pairs of data.

Be careful not to confuse correlation and significance. On the next page we look at this in more detail.

1. **Find the Test Statistic, DF, and P-value**
   - **Pearson Correlation=-0.548**
   - **N=10**
   - **P-value=0.101**
2. **Determine if significant**
   - **P-value > 0.05**
   - **Not Significant**
3. **State if null rejected or not**
   - **Failed to reject Null**
4. **Write result**
   - **There was no correlation between Height and Stretch Increase in subjects.**
5. **Present appropriate plot**
   - **N/A**

- If a result is not significant, it is no necessary to include the test statistic and p-value
- Should not graph results

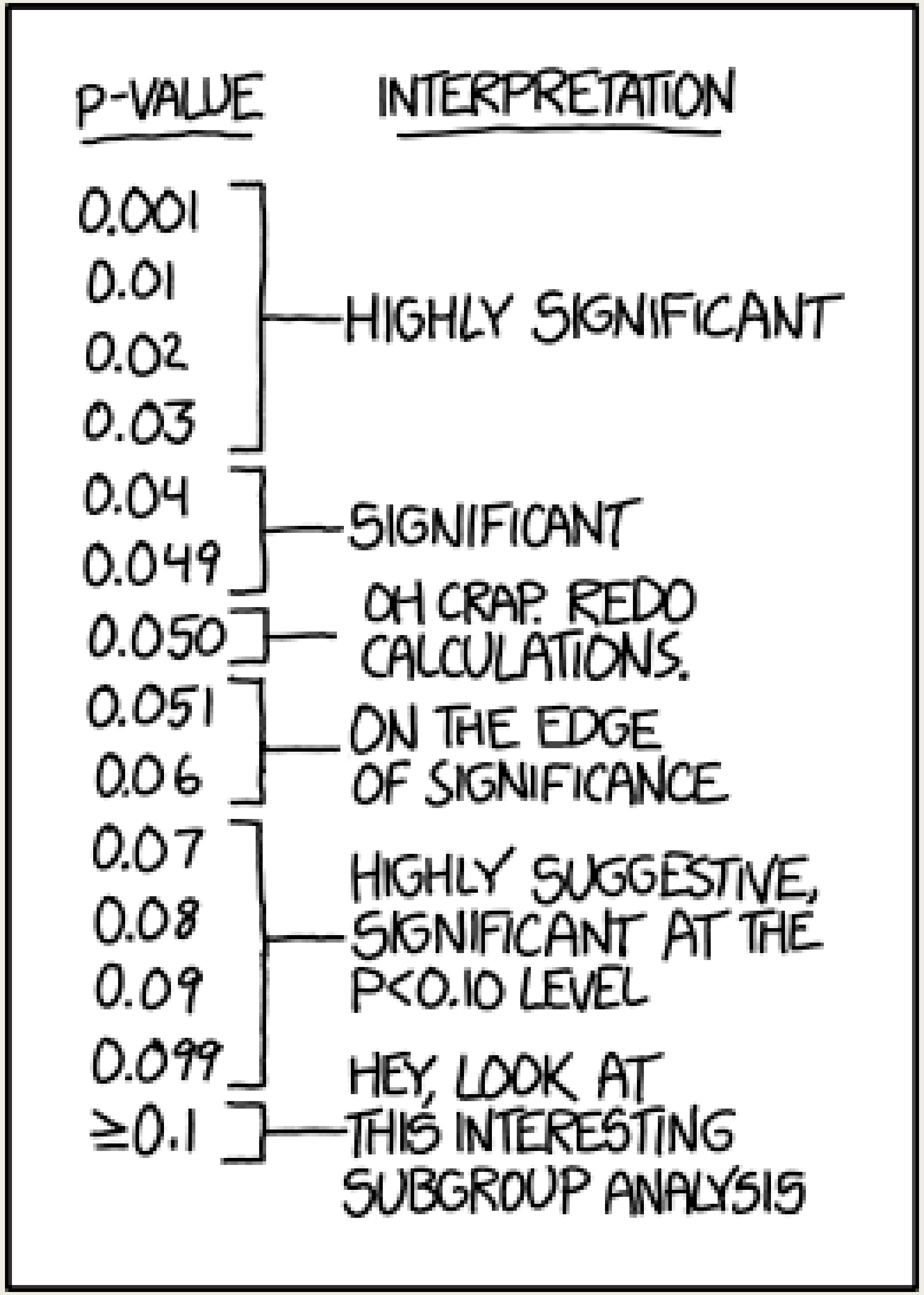
Part B-5

Correlation Notes

- Looking for correlation is different from looking for increases or decreases
- Correlation does not necessarily mean a causal relationship. Just because two values appear to go up and down together does not mean one is causing the other.
- The Pearson's coefficient is designed primarily for looking at linear relationships. Two variables can be related, but if the relationship is not linear, Pearson's correlation coefficient is not an appropriate statistic for measuring their association.
- The number of observations as with other statistics effects the significance.

P-values a summary

- "P-values do not simply provide you with a Yes or No answer, they provide a sense of the strength of the evidence against the null hypothesis.
- The lower the p-value, the stronger the evidence.
- Once you know how to read p-values, you can more critically interpret journal articles, and decide for yourself if you agree with the conclusions of the author. " - TexaSoft, (1996-2001)


P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

Part B-2

Chi-Square

- 2 categorical variables
- This is when you have two categories (simple case is that the categories have two groups in each, but doesn't need to be the case)
- End up getting frequencies of each category class and then generation overall ratios
- Tests for whether the categories are independent or not; can be set up against many null frequencies
- Non-parametric, so there isn't a parametric/non-parametric dichotomy
  - *The data are assumed to be a random sample.*
  - *The expected frequencies for each category should be at least 1.*
  - *No more than 20% of the categories should have expected frequencies of less than 5."*
- Examples:
  - *Is hair color independent of gender?*
  - *Are the ratios of expected genetic crosses of pea plants independent of the observed genetic crosses?*
  - *Is coffee type (caffeinated, decaf) independent of mood (happy, sad)?*

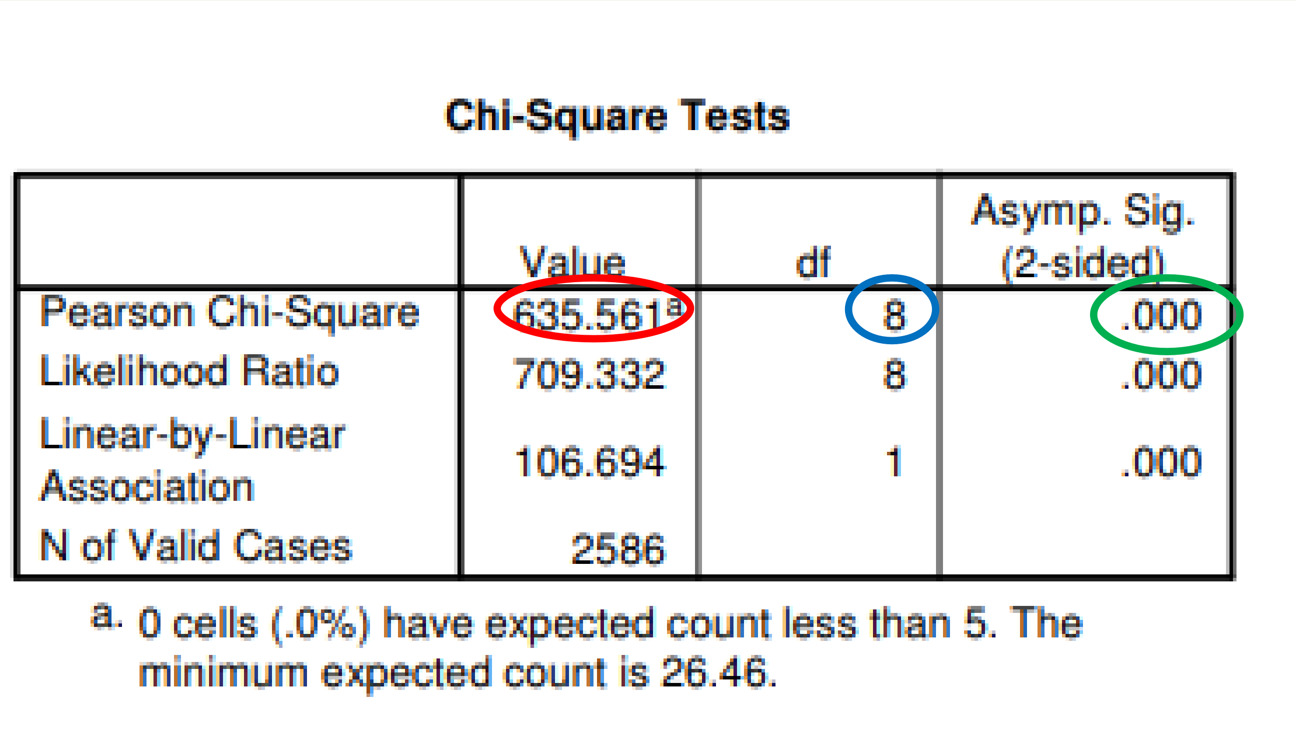
1. Explore your data
2. Check normality
   1. *Not applicable*
3. Set up hypothesis
   1. *Null: categories are independent*
   2. *Alternative: categories are not independent*
4. Select and run appropriate test
   1. *Chi-Square*
5. Interpret results
   1. *Null rejected or failed to reject?*
   2. *What does it mean for your question*
   3. *Write it out*

Part B-2

Chi Square: Examples

- Parametric Example: Male/Female Ratio
  - *Does the ratio of males to females in each school in SHU reflect the overall ratio in the university? (or put another way is there a larger than expected number of one gender in some schools?)*
  - *The data we have available are from a survey of students done in 2001.*
  - *You will see that the data is all numeric. If you want to know what the numbers represent you can look under the Variable View to find out, but this isn't necessary for our purpose. The crosstab system automatically labels the output!*
  - ***Question: Does each school in the SHU have male/female ratio that reflect the overall ratio?***

1. Open **Students data 2001** file in SPSS
2. Explore data
   1. *Analyze->Descriptive Statistics->Crosstabs*
   2. *Put **Gende**r under Row(s)*
   3. *Put **School** under Column(s)*
   4. *Examine Crosstabulation table*
3. Define null and alternative hypothesis to question.
   1. *Get the 'expected' values*
   2. *Go to Crosstabs dialog box*
   3. *Click Cells button then select 'Expected' under the counts section*
4. Run Appropriate test
   1. *Analyze->Descriptive Statistics->Crosstabs*
   2. *Click Statistics button, then select "Chi-Square"*
5. Interpret Results
   1. *See next page*
   2. *What is the test statistic, degrees of freedom, and p-value?*
   3. *Did you reject or fail to reject the null?*
   4. *What does it mean for the question?*



Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	635.561[a]	8	.000
Likelihood Ratio	709.332	8	.000
Linear-by-Linear Association	106.694	1	.000
N of Valid Cases	2586		

[a] 0 cells (.0%) have expected count less than 5. The minimum expected count is 26.46.

1. **Find the Test Statistic, DF, and P-value**
   - **Chi-Square=635.561**
   - **DF=8**
   - **P-value<0.0001**
2. **Determine if significant**
   - **P-value < 0.05**
   - **Significant**
3. **State if null rejected or not**
   - **Reject the null**
4. **Write result**
   - **There is a significant difference in the representation of the sexes across the schools (2-tailed chi square test, chi-sq=635.561, df=8, p-value<0.0001).**
5. **Present appropriate plot**
   - **N/A**

Part B-2

1-way ANOVA

- 1 categorical variable + 1 numerical variable
  - *Categorical variable is un-paired and group number is greater than two*
- This is when you have a single numerical variable you are interested in and want to know if it is different between multiple groups
  - *Is there a difference in grade point average between Freshmen, Sophomores, Juniors, and Seniors?*
  - *Is there a difference in soil moisture retention between 5 types of soil plot designs?*
  - *Is there a size difference between 4 different species of owls?*
- Important thing, the ANOVA test doesn’t tell you which groups are different, only that there is a statistical difference
- Need to do a Post Hoc test to determine what the difference is

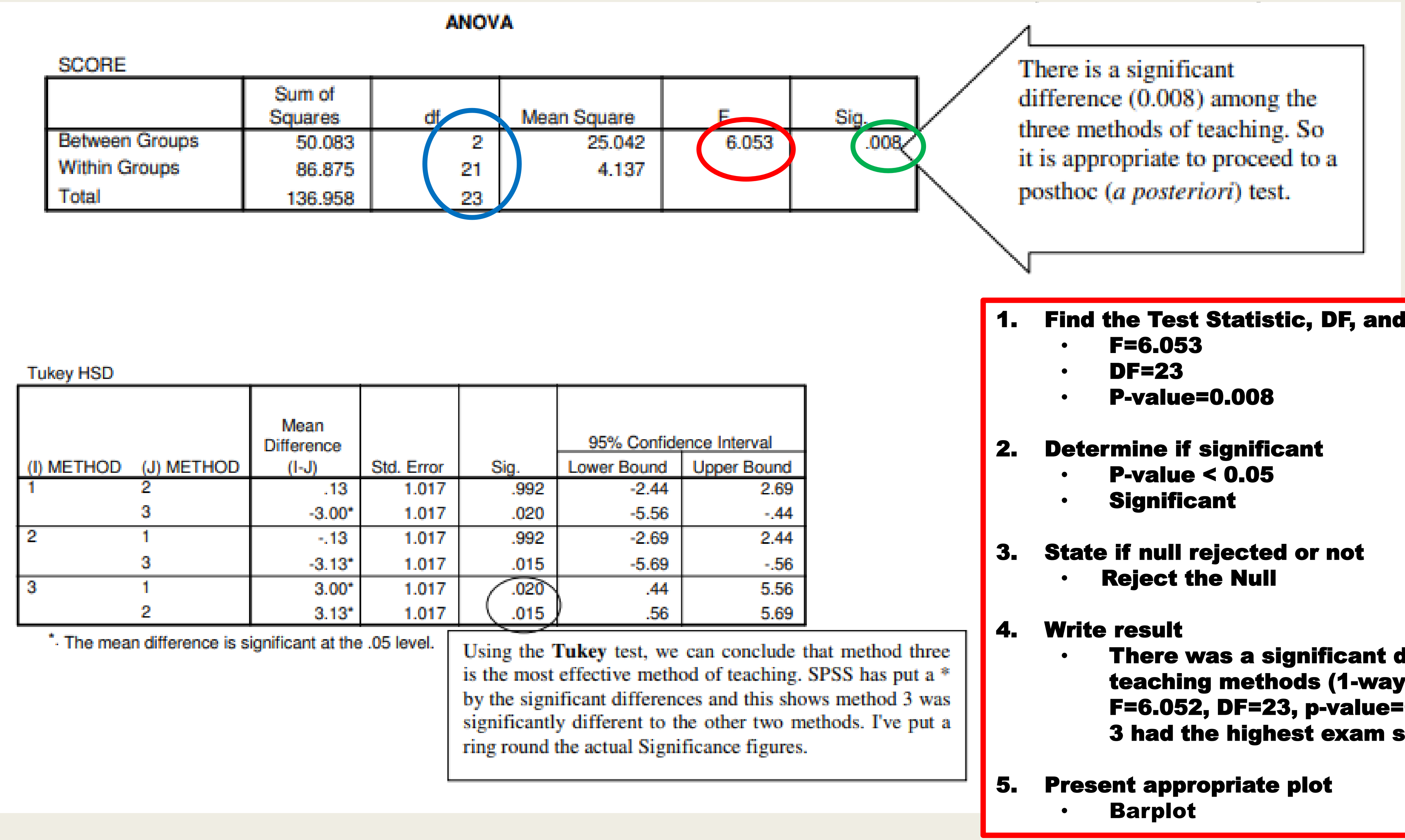
1. Explore your data
   1. *Histogram of numerical variable*
   2. *Boxplot of numerical variable grouped by the categorical variable*
2. Check normality
   1. *Histogram, QQ-plot*
3. Set up hypothesis
   1. *Null: there is no difference between groups*
   2. *Alternative: there is a difference*
4. Select and run appropriate test
   1. *Parametric: One-Way ANOVA*
   2. *Non-parametric: Krustal Wallace Test*
5. Interpret results
   1. *Null rejected or failed to reject?*
   2. ***Post Hoc test***
   3. *What does it mean for your question*
   4. *Write it out*

Part B-2

One Way ANOVA: Examples

- Parametric Example: Teaching Methods
  - *An experimenter is interested in evaluating the effectiveness of three methods of teaching a given course.*
  - *A group of 24 subjects is available to the experimenter*
  - *This group is considered by the experimenter to be the equivalent of a random sample from the population of interest.*
  - *Three subgroups of eight subjects each are formed at random; the subgroups are then taught by one of the three methods. Upon completion of the course, each of the subgroups is given a common test (exam) covering the material in the course*
  - *Note: the Method is set up with numbers (1,2,3) but is actually categorical*
  - ***Question: Is there a difference in scores between the three methods?***

1. Open **anova one way example** file in SPSS
2. Explore data and test for normality
   1. *Histogram of scores*
   2. *Boxplots of scores by method*
3. Define null and alternative hypothesis to question.
   1. *Fill out yourself*
4. Run Appropriate test
   1. *Analyze->Compare Means->One-Way ANOVA*
   2. *Score in Dependent List*
   3. *Method in Factor*
5. Interpret Results
   1. *See next page*
   2. *What is the test statistic, degrees of freedom, and p-value?*
   3. *Did you reject or fail to reject the null?*
   4. *If reject, null, run post hoc and determine the difference.*
      1. Go back to One Way ANOVA dialog box
      2. Choose Post Hoc -> Tukey
   5. *What does it mean for the question?*



ANOVA

SCORE

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	50.083	2	25.042	6.053	.008
Within Groups	86.875	21	4.137		
Total	136.958	23			

There is a significant difference (0.008) among the three methods of teaching. So it is appropriate to proceed to a posthoc (*a posteriori*) test.

Tukey HSD

(I) METHOD	(J) METHOD	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	.13	1.017	.992	-2.44	2.69
	3	-3.00*	1.017	.020	-5.56	-.44
2	1	-.13	1.017	.992	-2.69	2.44
	3	-3.13*	1.017	.015	-5.69	-.56
3	1	3.00*	1.017	.020	.44	5.56
	2	3.13*	1.017	.015	.56	5.69

*. The mean difference is significant at the .05 level.

Using the **Tukey** test, we can conclude that method three is the most effective method of teaching. SPSS has put a * by the significant differences and this shows method 3 was significantly different to the other two methods. I've put a ring round the actual Significance figures.

1. **Find the Test Statistic, DF, and P-value**
   - **F=6.053**
   - **DF=23**
   - **P-value=0.008**

2. **Determine if significant**
   - **P-value < 0.05**
   - **Significant**

3. **State if null rejected or not**
   - **Reject the Null**

4. **Write result**
   - **There was a significant difference in teaching methods (1-way ANOVA, F=6.052, DF=23, p-value=0.008). Method 3 had the highest exam scores**

5. **Present appropriate plot**
   - **Barplot**

Notes on ANOVA

- Basically, ANOVA answers the question “Is there a significant difference between the samples

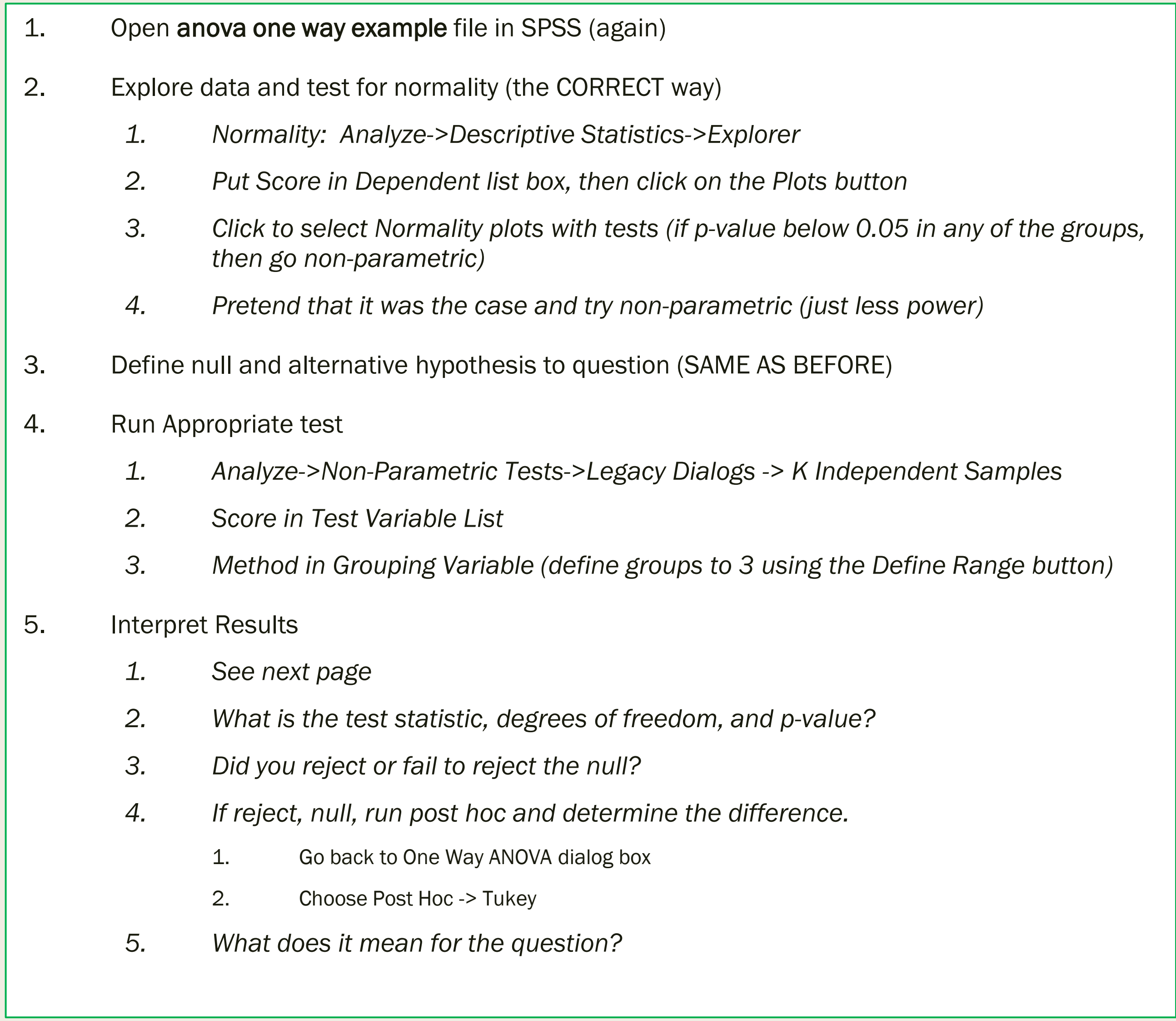
(is any one different from the others)?”

- If there is not (Sig. >0.05) then there is no need to go any further
- If there is then you might want to know which sample(s) is different from each other.
- A supplementary (Post-hoc) test is carried out to investigate differences between the samples.
- Selecting a post test is not simple; generally, to compare groups with each other choose the Tukey test.)

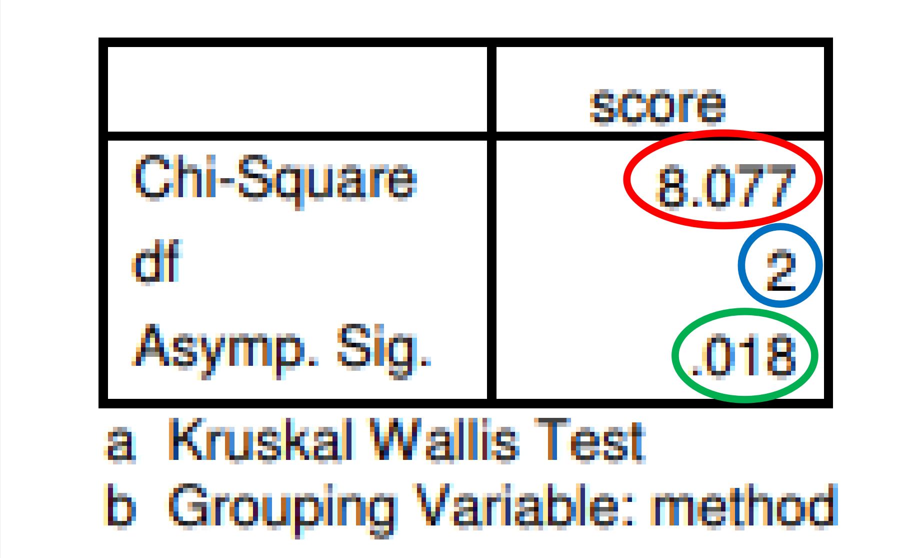
Part B-2

One Way ANOVA: Examples

- Non-Parametric Example: Teaching Methods
  - *The data are really three different sets of scores, one set for each group, so when we test them for normality, we need to remember this, if we treat them as one group then any differences between the groups might lead us to thing that the data aren't normally distributed when the data from each group is*
  - *It is the normality of each group that matters*
  - ***Question: Is there a difference in scores between the three methods?***



1. Open **anova one way example** file in SPSS (again)
2. Explore data and test for normality (the CORRECT way)
   1. *Normality: Analyze->Descriptive Statistics->Explorer*
   2. *Put Score in Dependent list box, then click on the Plots button*
   3. *Click to select Normality plots with tests (if p-value below 0.05 in any of the groups, then go non-parametric)*
   4. *Pretend that it was the case and try non-parametric (just less power)*
3. Define null and alternative hypothesis to question (SAME AS BEFORE)
4. Run Appropriate test
   1. *Analyze->Non-Parametric Tests->Legacy Dialogs -> K Independent Samples*
   2. *Score in Test Variable List*
   3. *Method in Grouping Variable (define groups to 3 using the Define Range button)*
5. Interpret Results
   1. *See next page*
   2. *What is the test statistic, degrees of freedom, and p-value?*
   3. *Did you reject or fail to reject the null?*
   4. *If reject, null, run post hoc and determine the difference.*
      1. Go back to One Way ANOVA dialog box
      2. Choose Post Hoc -> Tukey
   5. *What does it mean for the question?*



	score
Chi-Square	8.077
df	2
Asymp. Sig.	.018

a Kruskal Wallis Test
b Grouping Variable: method

- Notice that the nonparametric test still says that there is a significant difference between the groups (p=0.018) however it isn't quite as well convinced as the more sensitive ANOVA. This is a good illustration of the minor penalty that you pay for the more rugged 75 nonparametric tests, they are less likely to catch a small effect that does exist, i.e. they are less powerful.
- Run Post-Hoc test like before (Tukey)
- So to recap; generally scores would be better treated by nonparametric methods. In this example we did find them to be normally distributed and used them as an example in applying a one way ANOVA and its nonparametric equivalent, the Kruskal-Wallis test. Finally, the two tests agreed but we noticed a slight difference in how certain they were.

1. **Find the Test Statistic, DF, and P-value**
   - **Chi-Square=8.077**
   - **DF=2**
   - **P-value=0.018**
2. **Determine if significant**
   - **P-value < 0.05**
   - **Significant**
3. **State if null rejected or not**
   - **Reject the Null**
4. **Write result**
   - **There was a significant difference in teaching methods (Kruskal Wallis Test, Chi-Square=8.077, DF=2, p-value=0.018). Method 3 had the highest exam scores.**
5. **Present appropriate plot**
   - **Boxplot**

Part B-2

Repeated Measures ANOVA

- 1 categorical variable + 1 numerical variable
  - *Categorical variable is paired and group number is greater than two*
- This is when you have a single numerical variable you are interested in and want to know if it is different between multiple groups that are paired, usually the same subject
  - *Is there a difference in grade point average between students their Freshman, Sophomore, Junior, or Senior year?*
  - *Is there a difference in soil moisture retention 1, 2, 3, 4, 5, or 6 years post treatment?*
  - *Is there a difference in heartrate after 1, 2, and 3 cups of coffee?*
- Repeated Measures ANOVA like extension of the paired t-test, like 1-way ANOVA like extension of the independent samples t-test

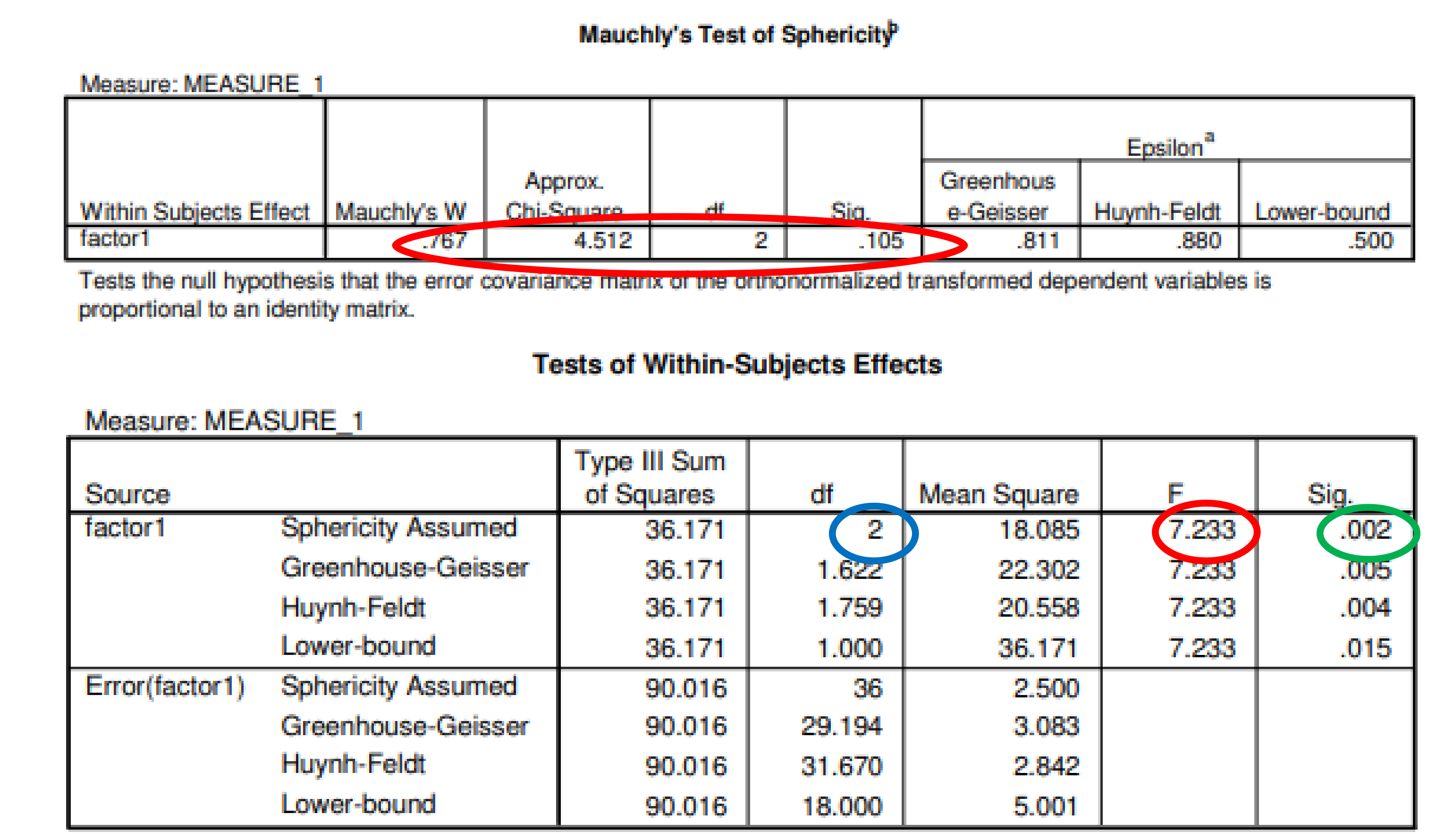
1. Explore your data
   1. *Histogram of numerical variable*
   2. *Boxplot of numerical variable grouped by the categorical variable*
2. Check normality
   1. *Histogram, QQ-plot, Test for Normality*
   2. *Do these by group*
3. Set up hypothesis
   1. *Null: there is no difference between groups*
   2. *Alternative: there is a difference*
4. Select and run appropriate test
   1. *Parametric: Repeated Measures ANOVA*
   2. *Non-parametric: Friedman Test*
5. Interpret results
   1. *Null rejected or failed to reject?*
   2. ***Post Hoc test***
   3. *What does it mean for your question*
   4. *Write it out*

Part B-2

Repeated Measures ANOVA: Examples

- Parametric Example: Jumping
  - *The following data shows the results of an experiment where subjects jumped three times.*
  - *Each subject jumped three times, the height was recorded, the column labelled Jump1 has each subjects first jump in it, the column labelled Jump2 has each subjects second jump in it and so on.*
  - ***Question: Is there a difference in energy between the three jumps?***

1. Open **Three Jumps** file in SPSS
2. Explore data and test for normality (the CORRECT way)
   1. *Normality: Analyze->Descriptive Statistics->Explorer*
   2. *Put the three variables containing the energies (jump 1-3) in Dependent Box*
   3. *Click to select Normality plots with tests (if p-value below 0.05 in any of the groups, then go non-parametric)*
3. Define null and alternative hypothesis to question
   1. *Fill out yourself*
4. Run Appropriate test
   1. *Analyze->General Linear Model->Repeated Measures*
   2. *Repeated Measure; define factors dialog should appear -> put 3 in number of levels, as there were three jumps, then click Add button*
   3. *Click define, highlight the three jump variables and send them into the box with the question marks in and click OK*
5. Interpret Results
   1. *See next page*
   2. *What is the test statistic, degrees of freedom, and p-value?*
   3. *Did you reject or fail to reject the null?*
   4. *If reject, null, run post hoc and determine the difference.*
      1. Analyze->Compare Means->Paired-Samples T Test (no more than three levels)
   5. *What does it mean for the question?*



Mauchly's Test of Sphericity[b]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
factor1	.767	4.512	2	.105	.811	.880	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
factor1	Sphericity Assumed	36.171	2	18.085	7.233	.002
	Greenhouse-Geisser	36.171	1.622	22.302	7.233	.005
	Huynh-Feldt	36.171	1.759	20.558	7.233	.004
	Lower-bound	36.171	1.000	36.171	7.233	.015
Error(factor1)	Sphericity Assumed	90.016	36	2.500		
	Greenhouse-Geisser	90.016	29.194	3.083		
	Huynh-Feldt	90.016	31.670	2.842		
	Lower-bound	90.016	18.000	5.001		

- Lots going on
- Two are of importance; Mauchly's Test of Sphericity and Tests of Within-Subjects Effects.
- The first one, Mauchly's Test of Sphericity, tells us which line of the second one to read (don't worry to much, just whatever) if sig value greater than 0.05, use first line in the Tests of Within-Subjects Effects,
- If not, use Greenhouse-Geisser

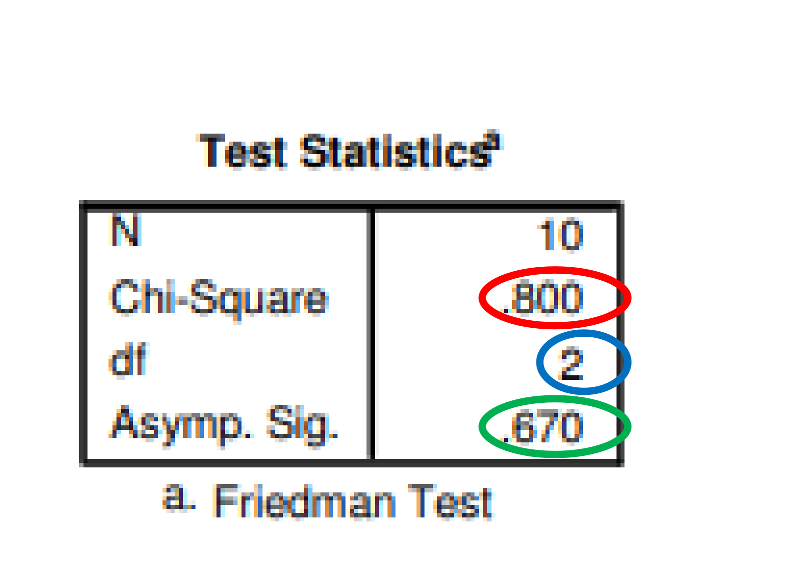
1. **Find the Test Statistic, DF, and P-value**
   - **F=7.233**
   - **DF=2**
   - **P-value=0.002**
2. **Determine if significant**
   - **P-value < 0.05**
   - **Significant**
3. **State if null rejected or not**
   - **Reject the Null**
4. **Write result**
   - **There was a significant in energy used between the three jumps (Repeated Measures ANOVA, Spericity Assumed, F=7.233, DF=2, p-value=0.002).**
5. **Present appropriate plot**
   - **Barplot**

Part B-2

Repeated Measures ANOVA: Examples

- Non-Parametric Example: Walking
  - *One group of subjects walked in three conditions; no crutches, elbow crutches and axillary crutches*
  - *the energy used was measured (indirectly by looking at the oxygen used)*
  - *Notice that there is no discriminatory variable in this data set, this is because there is only one group, each person was measured in all three conditions.*
  - ***Question: Is there a difference in energy used between the three methods?***

1. Open **enerjy and cruches** file in SPSS
2. Explore data and test for normality (the CORRECT way)
   1. *Normality: Analyze->Descriptive Statistics->Explorer*
   2. *Put the three variables containing the energies in Dependent Box*
   3. *Click to select Normality plots with tests (if p-value below 0.05 in any of the groups, then go non-parametric)*
3. Define null and alternative hypothesis to question
   1. *Fill out yourself*
4. Run Appropriate test
   1. *Analyze->Non-Parametric Tests->Legacy Dialogs->K-related samples*
   2. *Put all the three variables in the Test Variable Box*
5. Interpret Results
   1. *See next page*
   2. *What is the test statistic, degrees of freedom, and p-value?*
   3. *Did you reject or fail to reject the null?*
   4. *If reject, null, run post hoc and determine the difference.*
      1. NONE FOR NOW
   5. *What does it mean for the question?*



Test Statistics[a]

N	10
Chi-Square	.800
df	2
Asymp. Sig.	.670

a. Friedman Test

1. **Find the Test Statistic, DF, and P-value**
   - **Chi-Square=7.233**
   - **DF=2**
   - **P-value=0.670**

2. **Determine if significant**
   - **P-value > 0.05**
   - **Not Significant**

3. **State if null rejected or not**
   - **Fail to Reject the Null**

4. **Write result**
   - **There was no significant difference in energy used between the three jumps walking measures.**

5. **Present appropriate plot**
   - **N/A**

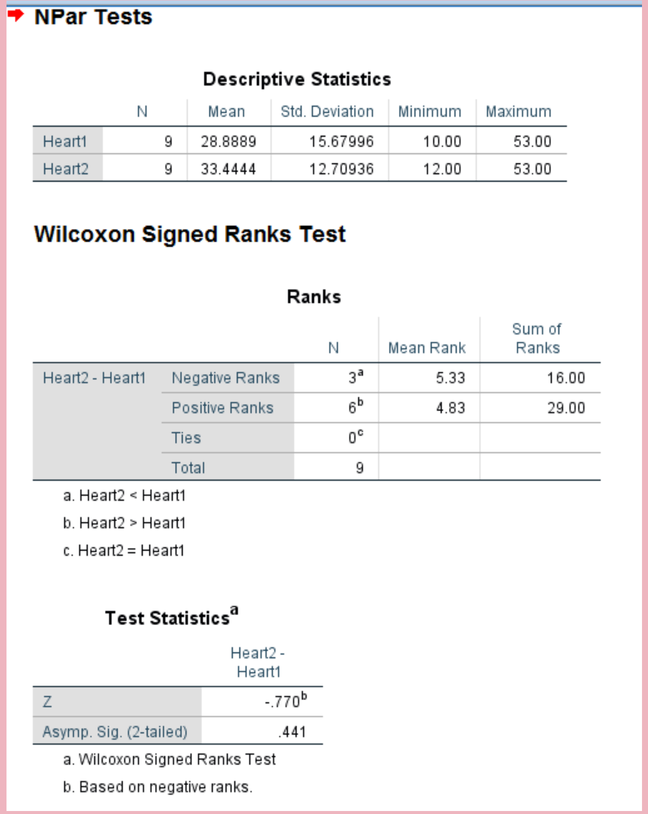
Part C-7

Mixed Designs

- Can combine things for more complicated analysis
- We can think of the repeated measures ANOVA as an extension of the paired t-test, and the One-way ANOVA as an extended version of the independent samples t-test
- A mixed design is when we have for example two groups of subjects measured repeatedly, e.g. two treatments, each treatment group being measured before and a couple of times after treatment
- Beyond the scope of this Presentation
- Professional approach -> just use generalized linear mixed model with the appropriate distribution

Assessment 6

1. A Non-parametric paired test was performed one two different heart rates of the same individual. Based on the results to the right, was there a significant difference between the two heart measurements? If so, how?
2. The number of bears and the number of marshmallows was counted at various campsites. Is there a correlation between bears and marshmallows? If so, what is the strength and direction?
3. Individuals from three different animal groups (Sharks=1, Dinosaurs=2, Bears=3) were rated for coolness. A one-way ANOVA was performed, along with a Post Hoc Test. Was there a significant difference in coolness across the three groups?



NPar Tests

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Heart1	9	28.8889	15.67996	10.00	53.00
Heart2	9	33.4444	12.70936	12.00	53.00

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
Heart2 - Heart1	Negative Ranks	3[a]	5.33	16.00
	Positive Ranks	6[b]	4.83	29.00
	Ties	0[c]		
	Total	9		

a. Heart2 < Heart1

b. Heart2 > Heart1

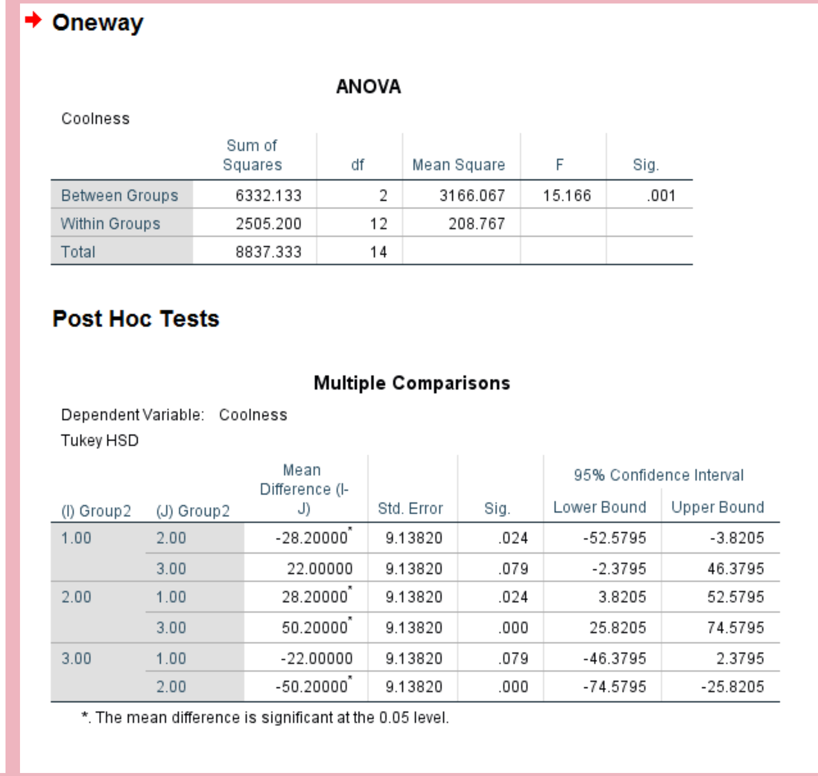
c. Heart2 = Heart1

Test Statistics[a]

	Heart2 - Heart1
Z	-.770[b]
Asymp. Sig. (2-tailed)	.441

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.



Oneway

ANOVA

Coolness

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	6332.133	2	3166.067	15.166	.001
Within Groups	2505.200	12	208.767		
Total	8837.333	14			

Post Hoc Tests

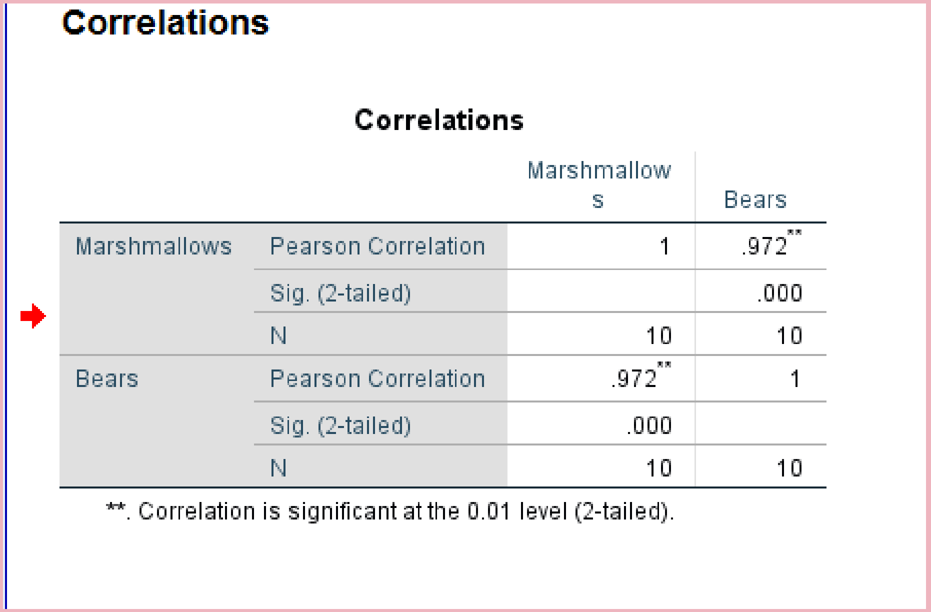
Multiple Comparisons

Dependent Variable: Coolness

Tukey HSD

					95% Confidence Interval	
(I) Group2	(J) Group2	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1.00	2.00	-28.20000*	9.13820	.024	-52.5795	-3.8205
	3.00	22.00000	9.13820	.079	-2.3795	46.3795
2.00	1.00	28.20000*	9.13820	.024	3.8205	52.5795
	3.00	50.20000*	9.13820	.000	25.8205	74.5795
3.00	1.00	-22.00000	9.13820	.079	-46.3795	2.3795
	2.00	-50.20000*	9.13820	.000	-74.5795	-25.8205

*. The mean difference is significant at the 0.05 level.



Correlations

Correlations

		Marshmallows	Bears
Marshmallows	Pearson Correlation	1	.972**
	Sig. (2-tailed)		.000
	N	10	10
Bears	Pearson Correlation	.972**	1
	Sig. (2-tailed)	.000	
	N	10	10

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

Assessment 6 Answers

1. A Non-parametric paired test was performed one two different heart rates of the same individual. Based on the results to the right, was there a significant difference between the two heart measurements? If so, how?

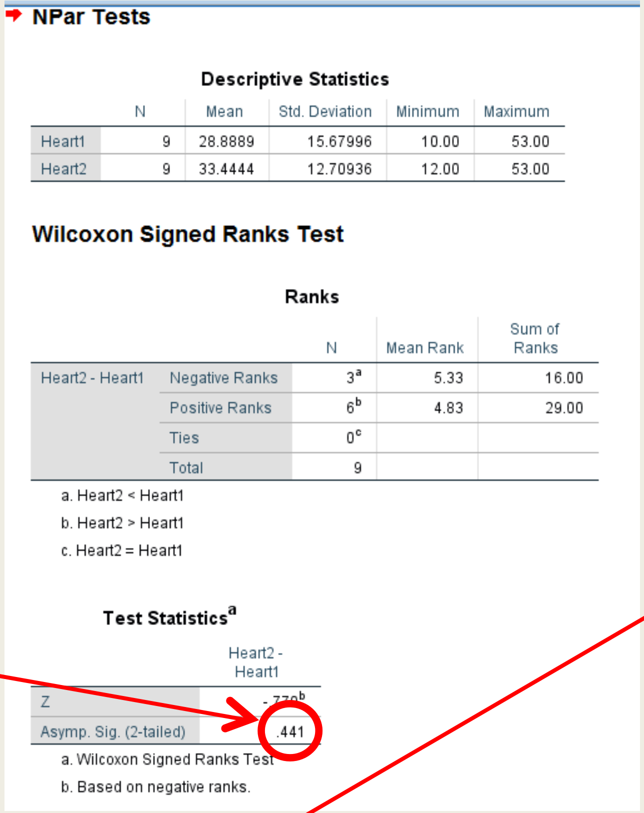
   No significance.

2. The number of bears and the number of marshmallows was counted at various campsites. Is there a correlation between bears and marshmallows? If so, what is the strength and direction?

   Significant correlation. Strong-Positive

3. Individuals from three different animal groups (Sharks=1, Dinosaurs=2, Bears=3) were rated for coolness. A one-way ANOVA was performed, along with a Post Hoc Test. Was there a significant difference in coolness across the three groups?

   Significant difference.



NPar Tests

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Heart1	9	28.8889	15.67996	10.00	53.00
Heart2	9	33.4444	12.70936	12.00	53.00

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
Heart2 - Heart1	Negative Ranks	3[a]	5.33	16.00
	Positive Ranks	6[b]	4.83	29.00
	Ties	0[c]		
	Total	9		

a. Heart2 < Heart1

b. Heart2 > Heart1

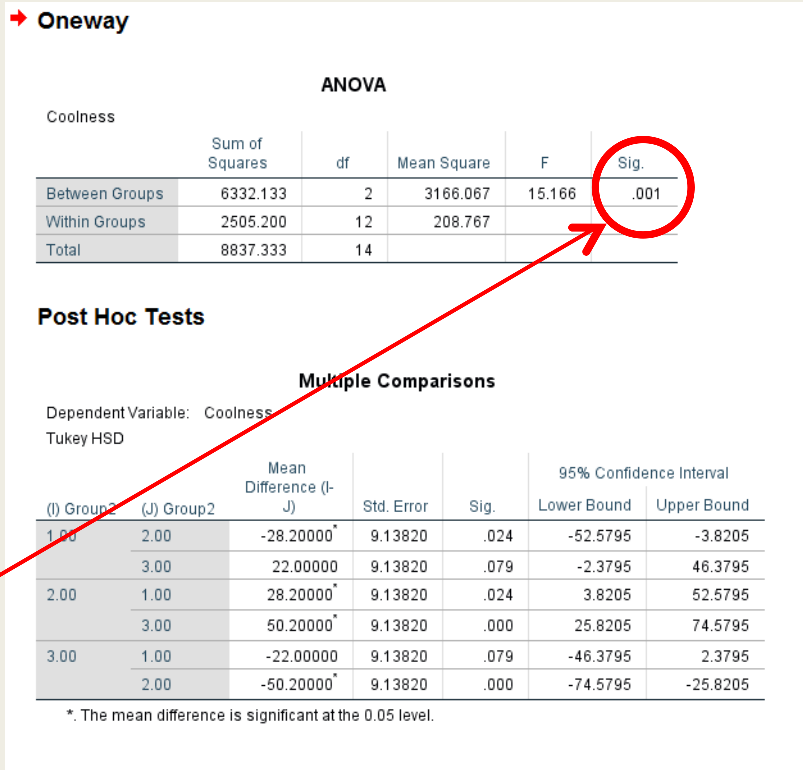
c. Heart2 = Heart1

Test Statistics[a]

	Heart2 - Heart1
Z	-.770[b]
Asymp. Sig. (2-tailed)	.441

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.



Oneway

ANOVA

Coolness

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	6332.133	2	3166.067	15.166	.001
Within Groups	2505.200	12	208.767		
Total	8837.333	14			

Post Hoc Tests

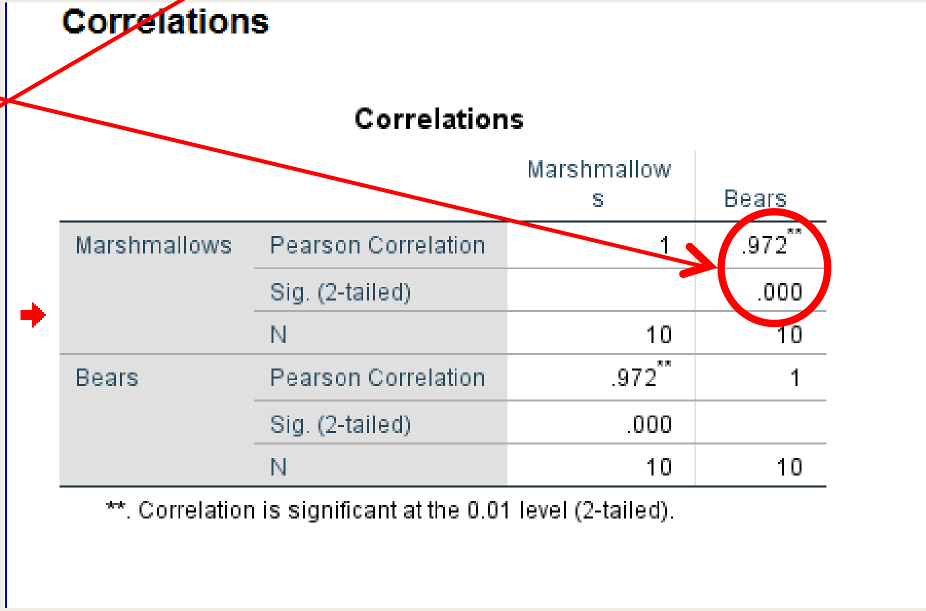
Multiple Comparisons

Dependent Variable: Coolness

Tukey HSD

					95% Confidence Interval	
(I) Group2	(J) Group2	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1.00	2.00	-28.20000*	9.13820	.024	-52.5795	-3.8205
	3.00	22.00000	9.13820	.079	-2.3795	46.3795
2.00	1.00	28.20000*	9.13820	.024	3.8205	52.5795
	3.00	50.20000*	9.13820	.000	25.8205	74.5795
3.00	1.00	-22.00000	9.13820	.079	-46.3795	2.3795
	2.00	-50.20000*	9.13820	.000	-74.5795	-25.8205

*. The mean difference is significant at the 0.05 level.



Correlations

Correlations

		Marshmallows	Bears
Marshmallows	Pearson Correlation	1	.972**
	Sig. (2-tailed)		.000
	N	10	10
Bears	Pearson Correlation	.972**	1
	Sig. (2-tailed)	.000	
	N	10	10

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

Part D: Other Analyses

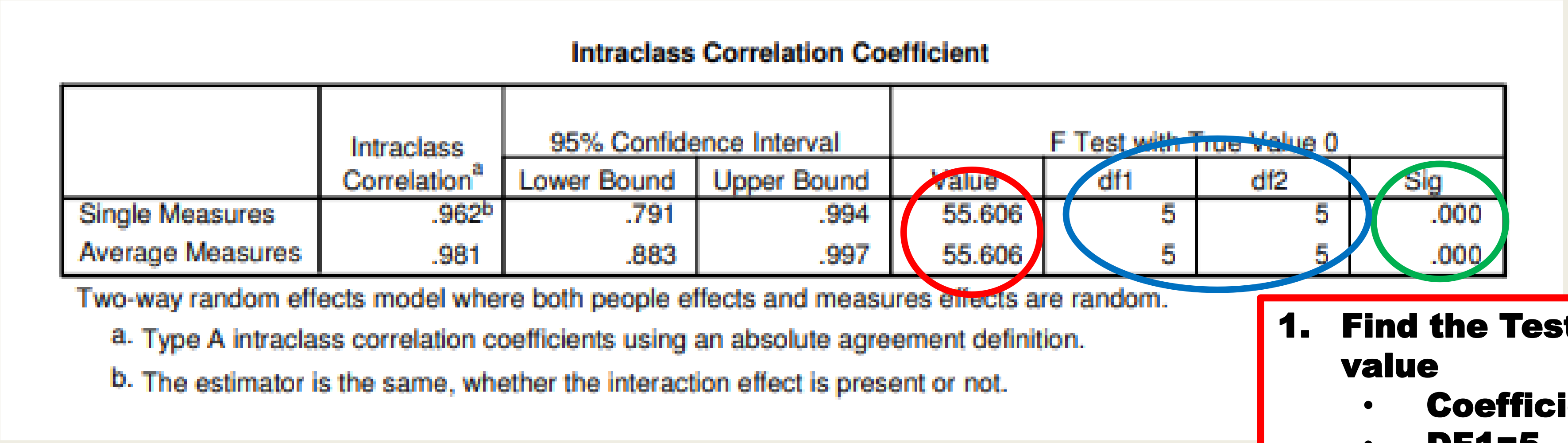
- Reliability and Sensitivity
  - *Reliability: how consistent measurements are*
    - Inter-rater reliability deals with the issue of reliability between different people (raters).
    - Intra-rater reliability deals with whether one rater is consistent, i.e. when they re-look at the same subjects do they rate them in a similar way again.
  - *Sensitivity: how much measurements can detect true effects*

Part B-2

Reliability: Examples

- Imagine that a student wants to find out if a certain exercise can improve performance.
- To measure performance they decide to use a simple measured jump. However to be sure that he can sensibly repeat the measures after the exercise regime has been completed he wants to estimate the reliability of his measurement method.
- To get round the problems of (XYZ), use Intraclass Correlation Coefficient
- The coefficient will tell us how much agreement the two measurements have
- Can also use Cronbach’s Alpha, another measure of reliability (Note that a reliability coefficient of .70 or higher is considered "acceptable" in most Social Science research situations using Cronbach's Alpha)
- Alpha also works for more than 2 measures

1. Open **ICC and Cronbachs alphs** file in SPSS
2. Calculate the ICC
    1. *Analyze->Scale->Reliability Analysis*
    2. *Put Jump 1 and 2 into the Items box and click statistics*
    3. *Tick the Intraclass Correlation Coefficient*
3. Calculate the Alpha
    1. *Done at the same time by the analysis*



Intraclass Correlation Coefficient

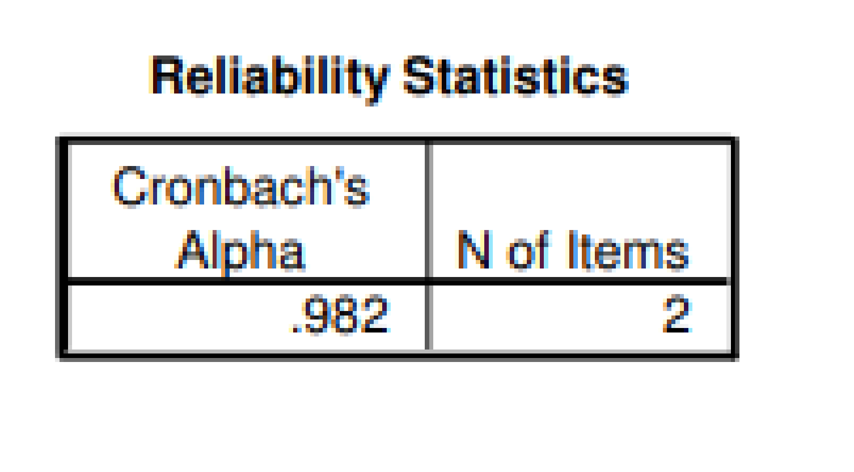
	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.962[b]	.791	.994	55.606	5	5	.000
Average Measures	.981	.883	.997	55.606	5	5	.000

Two-way random effects model where both people effects and measures effects are random.

a. Type A intraclass correlation coefficients using an absolute agreement definition.

b. The estimator is the same, whether the interaction effect is present or not.

- The Intraclass Correlation Coefficient (ICC) in this case is 0.962 we use the single measures because the figures we fed SPSS were raw measurements not an average of several attempts. This value, 0.962 shows a considerable amount of agreement!



Reliability Statistics

Cronbach's Alpha	N of Items
.982	2

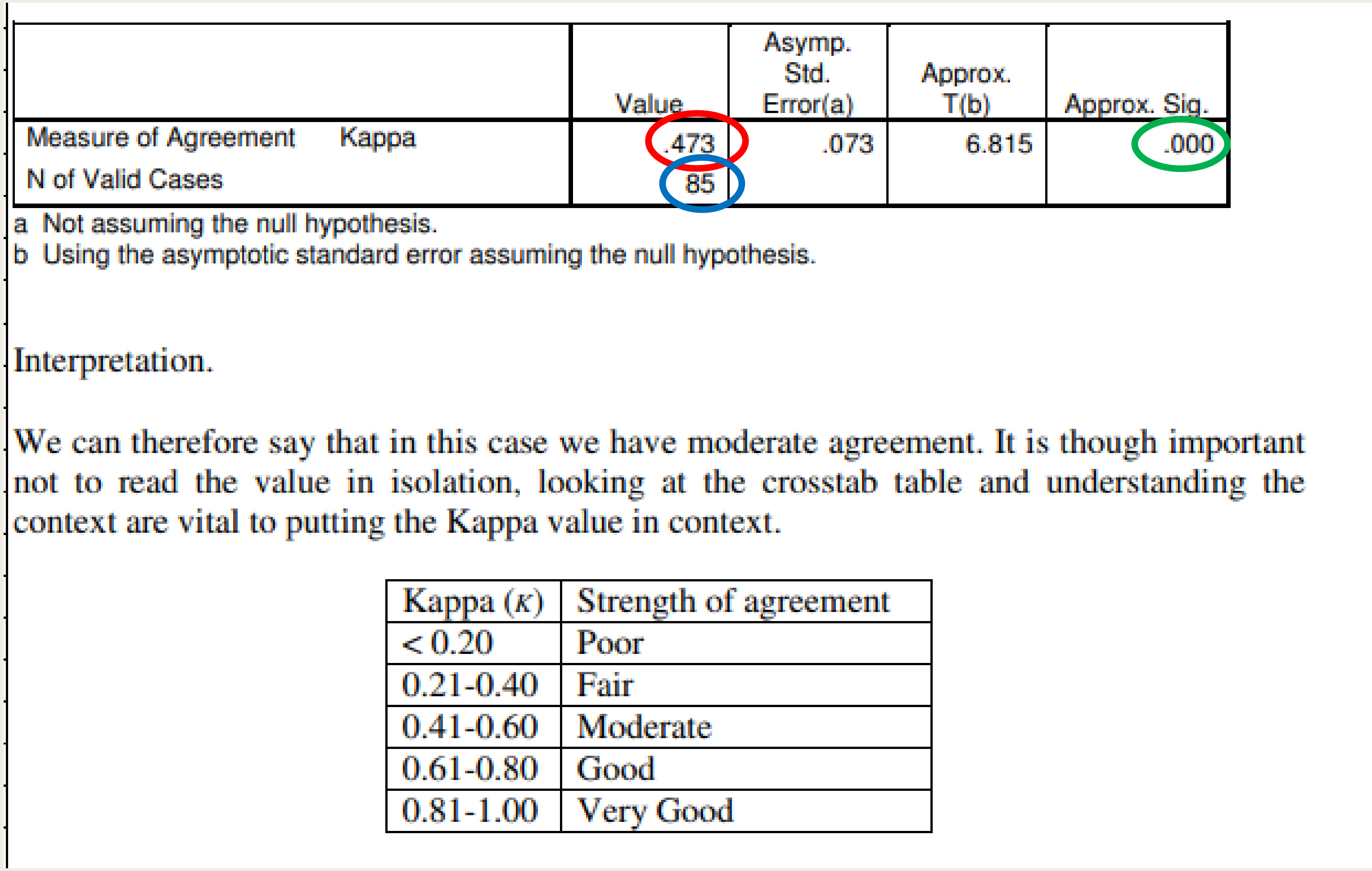
1. **Find the Test Statistic, DF, and P-value**
   - **Coefficient=55.606**
   - **DF1=5**
   - **DF2=5**
   - **P-value<0.0001**
2. **Determine if significant**
   - **P-value < 0.05**
   - **Significant**
3. **State if null rejected or not**
   - **Reject the Null**
4. **Write result**
   - **There was a significant correlation between jump measurements (ICC=55.606, DF=5,5, p-value<0.0001**
5. **Present appropriate plot**
   - **N/A**

Part D-3

Inter-rater agreement using Kappa

- Do two measurement agree, or two measurers (raters)?
- Between any two raters, even if they just guessed, there would have still been a degree of agreement just due to chance
- The Kappa statistic takes this into account

1. Open **radiologist eg from p403** file in SPSS
2. Analyze->Descriptives->Crosstabs
3. Put Radiologist 1 in Rows
4. Put Radiologist 2 in Columns
5. Click statistics and choose Kappa



		Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Measure of Agreement	Kappa	.473	.073	6.815	.000
N of Valid Cases		85			

a Not assuming the null hypothesis.
b Using the asymptotic standard error assuming the null hypothesis.

Interpretation.

We can therefore say that in this case we have moderate agreement. It is though important not to read the value in isolation, looking at the crosstab table and understanding the context are vital to putting the Kappa value in context.

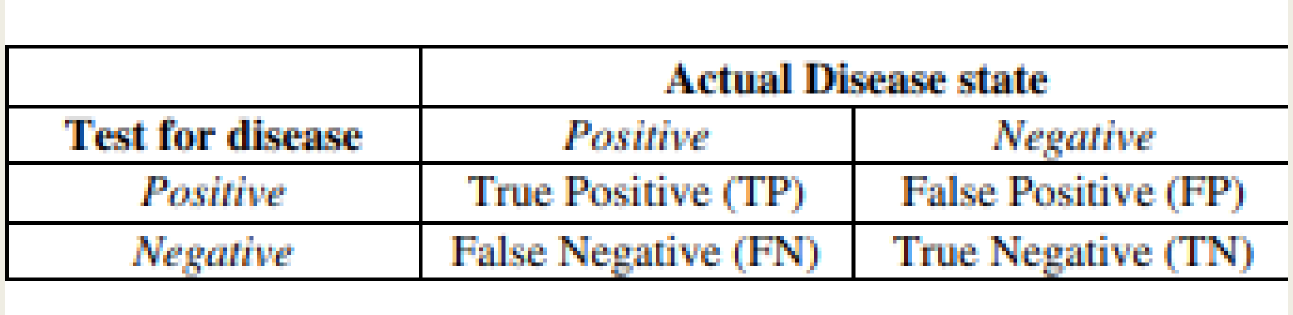
Kappa (κ)	Strength of agreement
< 0.20	Poor
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Good
0.81-1.00	Very Good

1. **Find the Test Statistic, DF, and P-value**
   - **Kappa=0.473**
   - **N=85**
   - **P-value<0.0001**
2. **Determine if significant**
   - **P-value < 0.05**
   - **Significant**
3. **State if null rejected or not**
   - **Reject the Null**
4. **Write result**
   - **There was moderate agreement between raters (Kappa=0.473, N=85, p-value<0.0001).**
5. **Present appropriate plot**
   - **N/A**

Part D-4

Calculating sensitivity and specificity of a diagnostic test

- The table below is a 2x2 cross tabulation (contingency table) representing the findings of a diagnostic test when compared to the actual disease state. I.e. a comparison of what the test indicated and the real facts.
- The four cells TP, FP, FN & TN would have in them the number in each category, they will total the number of cases investigated.
- The crosstabulate command in SPSS or the pivot table feature in MS Excel can calculate the matrix values.

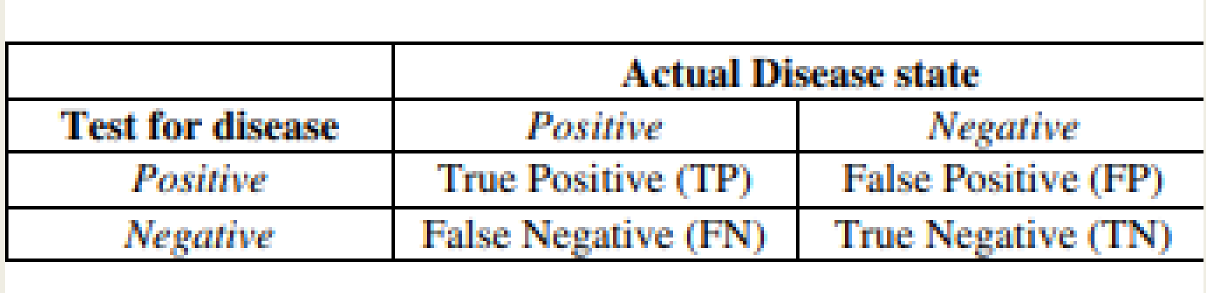


	Actual Disease state	
Test for disease	*Positive*	*Negative*
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

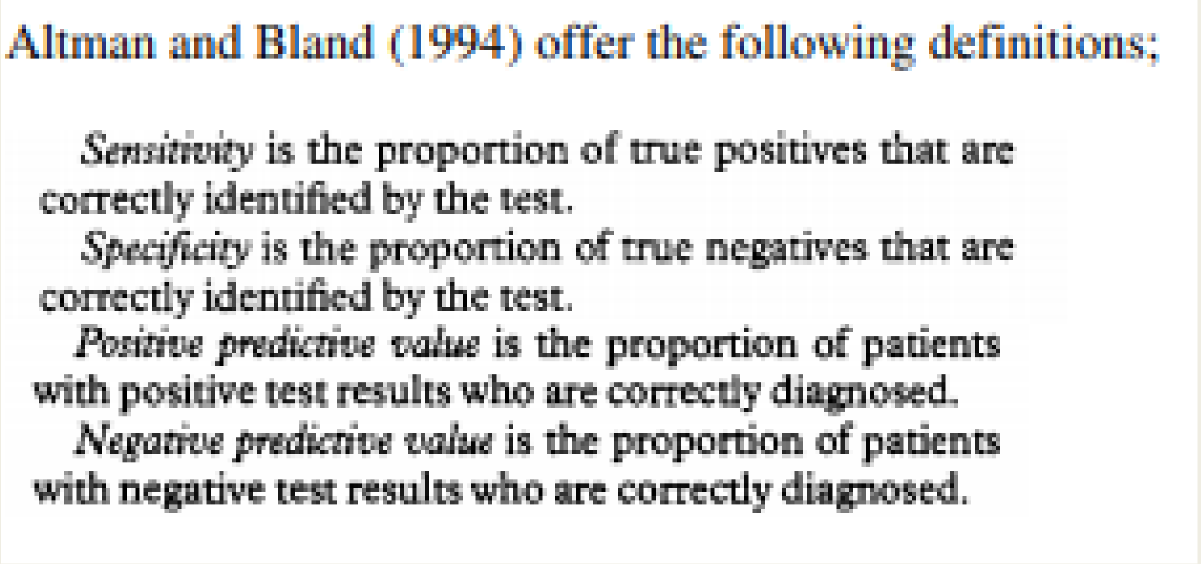
Part D-4

Calculating sensitivity and specificity of a diagnostic test 2

- Sensitivity=TP/(TP+FN)
- Specificity=TN/(FP+TN)
- Prevalence=(TP+FN)/(TP+FN+FP+TN)
- Positive Predictive Value =TP/(TP+FP)
- Negative Predictive Value =TN/(FN+TN)
- Positive Likelihood=SENS/(1-SPEC)
- Negative Likelihood=(1-SENS)/SPEC
- Overall Accuracy = (TP + TN)/(TP + FP + FN + TN)



	Actual Disease state	
Test for disease	*Positive*	*Negative*
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)



Altman and Bland (1994) offer the following definitions;

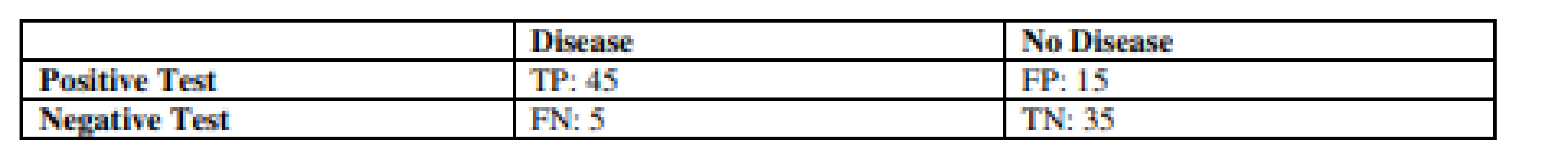
Sensitivity is the proportion of true positives that are correctly identified by the test.

Specificity is the proportion of true negatives that are correctly identified by the test.

Positive predictive value is the proportion of patients with positive test results who are correctly diagnosed.

Negative predictive value is the proportion of patients with negative test results who are correctly diagnosed.

Example



	Disease	No Disease
Positive Test	TP: 45	FP: 15
Negative Test	FN: 5	TN: 35

- Sensitivity = TP/(TP+FN) = 45/(45+5) = 45/50 = 0.9
- Specificity = TN/(FP+TN) = 35/(15+35) = 35/50 = 0.7

Odds Ratio

- *Information and examples based on Explaining Odds Ratio (by Magdalena Szumilas)*
- Odds Ratio (OR): Measure of association between exposure and outcome
- Odds that an outcome will occur given a particular exposure, compared to absence of that exposure
- Uses: odds of disease/disorder given certain exposure (health characteristics, medical history, treatment, environmental factor, etc.)
  - *OR=1 (exposure does no affect odds of outcome)*
  - *OR>1 (exposure associated with higher odds of outcome)*
  - *OR<1 (exposure associated with higher odds of outcome)*

Odds Ratio Calculation

- Odds Ratio (OR) $= \dfrac{\text{(n) exposed cases/(n) unexposed cases}}{\text{(n) exposed non-cases/(n) unexposed non-cases}} = \dfrac{\text{(n) exposed cases} \times \text{(n) unexposed non-cases}}{\text{(n) exposed non-cases} \times \text{(n) unexposed cases}}$

- OR = a/c / b/d = ad /bc

  a=Number of exposed cases

  b=number of exposed non-cases

  c=number of unexposed cases

  d=number of unexposed non-cases

Odds Ratio Example

- Can do by hand
- In the study, 186 of the 263 adolescents previously judged as having experienced a suicidal behavior requiring immediate psychiatric consultation did not exhibit suicidal behavior (non-suicidal, NS) at six months follow-up. Of this group, 86 young people had been assessed as having depression at baseline. Of the 77 young people with persistent suicidal behavior at follow-up (suicidal behavior, SB), 45 had been assessed as having depression at baseline.
- What is the OR of suicidal behavior at six months follow-up given presence of depression at baseline?

- Determine the following numbers

  a: Number of exposed cases (+ +) = ?

  b: Number of exposed non-cases (+ –) = ?

  c: Number of unexposed cases (– +) = ?

  d: Number of unexposed non-cases (– –) = ?

Odds Ratio Example Cont.

Q1: Who are the exposed cases (++ = a)?

A1: Youth with persistent SB assessed as having depression at baseline

a=45

Q2: Who are the exposed non-cases (+ – = b)?

A2: Youth with no SB at follow-up assessed as having depression at baseline **b**=86

Q3: Who are the unexposed cases (– + = c)?

A3: Youth with persistent SB not assessed as having depression at baseline

c: 77(SB) –45(depression) = 32

Q4: Who are the unexposed non-cases (– – = d)?

A4: Youth with no SB at follow-up not assessed as having depression at baseline

d: 186(NS) –86(depression) = 100

Odds Ratio (OR) = a/c / b/d = ad/bc

= 45*100 / 86*32 = 1.63

Odds of persistent suicidal behavior is 1.63, higher given baseline depression diagnosis compared to no baseline depression

Assessment 7

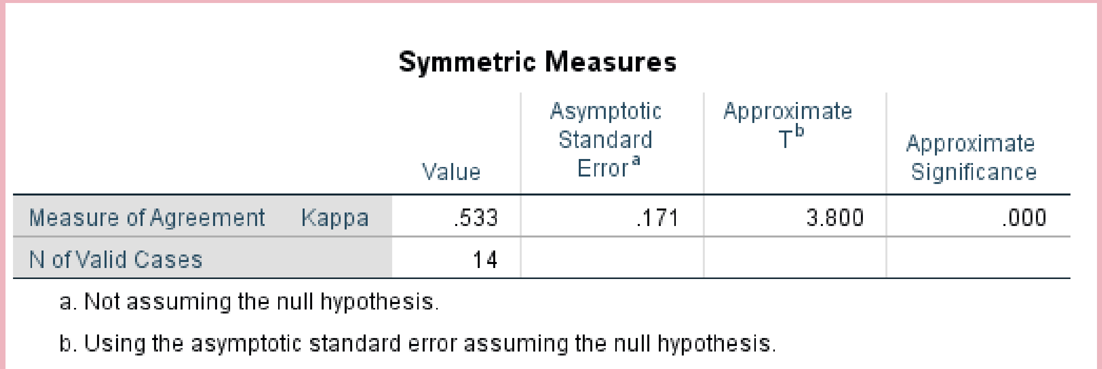
1. Two scientists observed the number of cups coffee I drank each day over a two-week period. A Kappa statistic (table to the right) was calculated to see how well their observations agreed with each other. Was the agreement significant? If so, what was the strength of the agreement.
2. On the right is a contingency table for the test of a certain disease (fear of empty coffee cups). What is the Sensitivity and Specificity of the test?
3. What is the odds ratio for the following information:

   Number of exposed cases= 45

   Number of exposed non-cases= 66

   Number of unexposed cases= 87

   Number of unexposed non-cases= 92

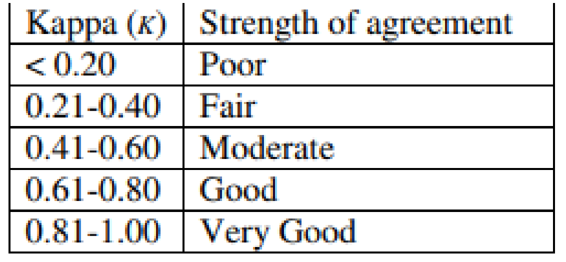


Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.533	.171	3.800	.000
N of Valid Cases		14			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.



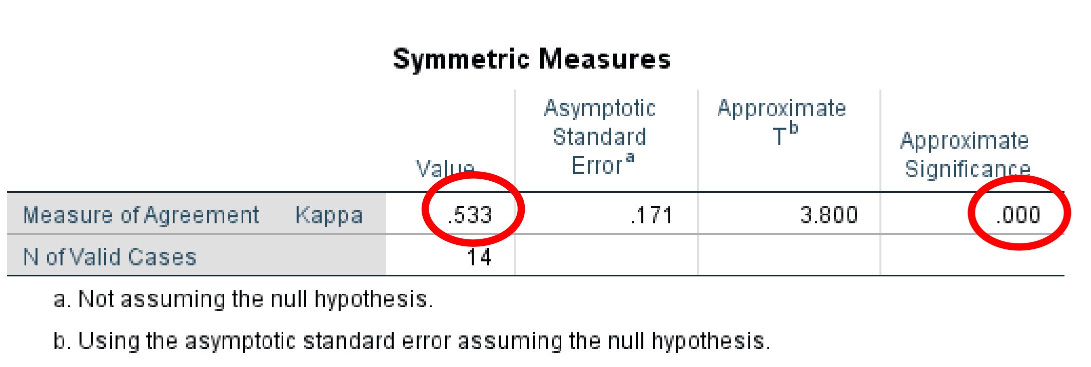
Kappa (κ)	Strength of agreement
< 0.20	Poor
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Good
0.81-1.00	Very Good

	Actual Disease State	
Test for Disease	Disease	No Disease
Positive Test	TP: 37	FP: 3
Negative Test	FN: 13	TN: 18

Assessment 7 Answers

1. Two scientists observed the number of cups coffee I drank each day over a two-week period. A Kappa statistic (table to the right) was calculated to see how well their observations agreed with each other. Was the agreement significant? If so, what was the strength of the agreement.

   Yes, moderate strength

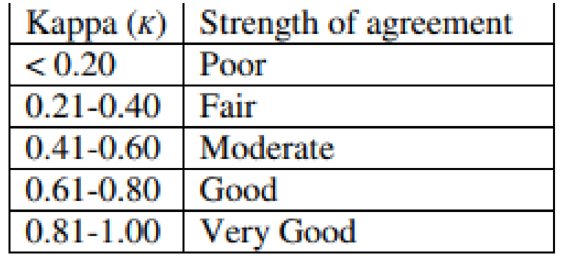


Symmetric Measures

		Value	Asymptotic Standard Error^a	Approximate T^b	Approximate Significance
Measure of Agreement	Kappa	.533	.171	3.800	.000
N of Valid Cases		14			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.



Kappa (κ)	Strength of agreement
< 0.20	Poor
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Good
0.81-1.00	Very Good

2. On the right is a contingency table for the test of a certain disease (fear of empty coffee cups). What is the Sensitivity and Specificity of the test?

   Sensitivity = TP/(TP+FN) = 37/(37+13) = 0.74

   Specificity= TN/(FP+TN) = 18/(3+18) = 0.857

	Actual Disease State	
Test for Disease	Disease	No Disease
Positive Test	TP: 37	FP: 3
Negative Test	FN: 13	TN: 18

1. What is the odds ratio for the following information:

   Number of exposed cases= 45

   Number of exposed non-cases= 66

   Number of unexposed cases= 87

   Number of unexposed non-cases= 92

   OR= a/b / c/d = ad /bc = 45*87 / 66*92 = 0.645

Summary

- SPSS is a very useful tool in analyzing data that does not depend on lots of computer coding background
- Analyzing data helps answer important questions we have about science, medicine, business, society, etc.
- The three major steps are:
  - *Explore Your Data*
  - *Analyze Your Data*
  - *Interpret your results*
- For more information on SPSS, start by checking out the resources below
  - *https://www.ibm.com/products/spss-statistics/resources*
  - *https://www.spss-tutorials.com/*