Statistics using Microsoft Excel and SPSS

Professor Dr. K. MARIMUTHU
Deputy Vice Chancellor, Academic and International Affairs,
Department of Biotechnology, Faculty of Applied Sciences

AIMST University, Bedong-Semeling, 08100 Semeling, Kedah Darul Aman, Malaysia.
T: +604 - 429 1054 | F: +604 - 429 8102 | HP: +6016 - 4723672
Email: marimuthu@aimst.edu.my | aquamuthu2k@gmail.com

Aim and Objectives

- Aim
  - To illustrate how Excel and SPSS can be used to carry out statistical analysis and tests
- Objectives
  - To show you how to use some of the statistical worksheet functions available in Excel and SPSS
  - To show you how to use some of the tools available in the Analysis ToolPak in Excel

Basic functions and statistical applications in Microsoft Excel

Why use Excel?

- Very popular and widely used
- Software more accessible & user friendly
- Easy to format output
- Better charting facilities than some statistical applications
- Good for exploratory/descriptive statistics
    - graphs and tables

In Life sciences and other fields, we will use Excel to:

1. Store and organize data,
2. Analyze data (descriptive statistics & inferential statistics), *and*
3. Represent data graphically (e.g., in bar graphs, histograms, and scatter plots)

5

Starting MS Excel:

- **Starting MS Excel**: Double click on the Microsoft Excel icon on the desktop
  - **or**
- Click on Start --> Programs --> Microsoft Excel.

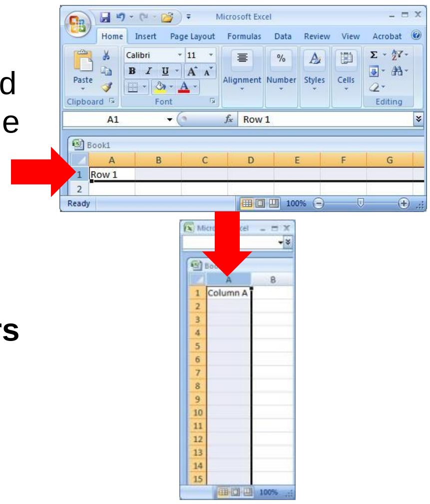
6

Excel Basics

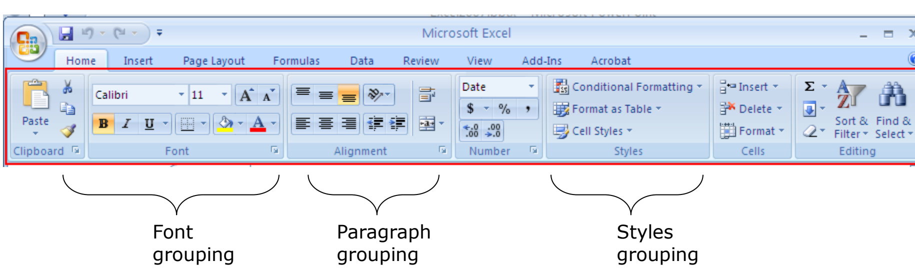
Excel spreadsheets organize information **(text and numbers)** by rows and columns:

This is a **row.**
Rows are represented by **numbers** along the side of the sheet.

This is a **column.**
Columns are represented by **letters** across the top of the sheet.



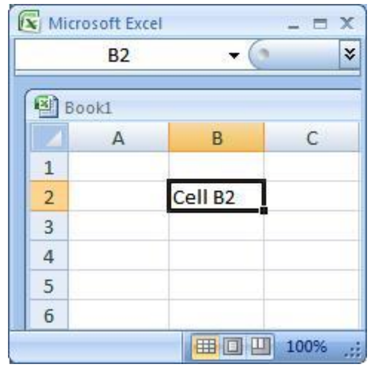
Ribbon



- Home: has the common formatting tools, clipboard, fonts, paragraphs, number, Styles, Cells, and Editing.

Excel Basics

- A **cell** is the intersection between a column and a row.
- Each cell is named for the column letter and row number that intersect to make it.
- For example, **B2** is used to refer to the cell in column B and row 2.
  - **B10:B20** is used to refer to the range of cells in column B and rows 10 through 20.

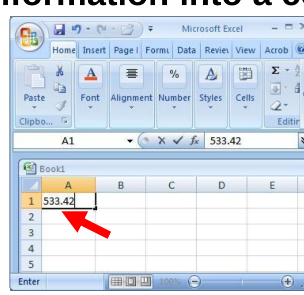


9

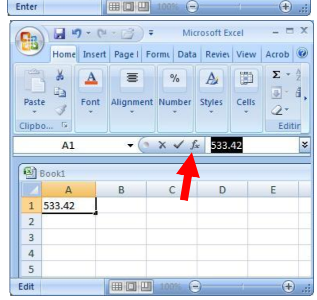
Data Entry

There are <u>two</u> ways to enter information into a cell:

1. Type directly into the cell.
Click on a cell, and type in the data (numbers or text) and press Enter.



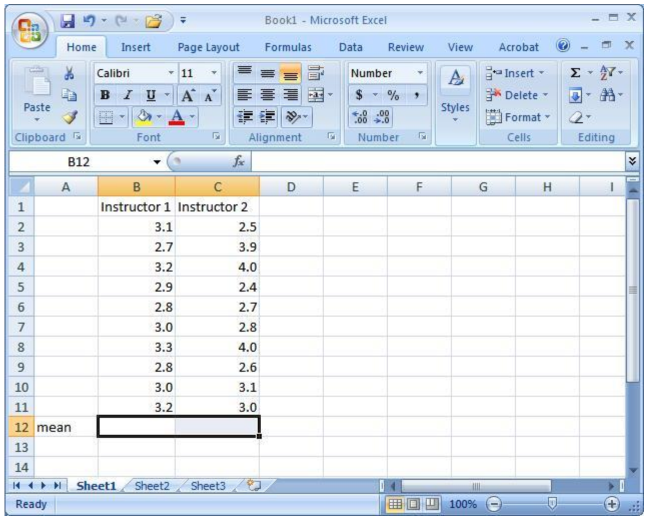
2. Type into the formula bar.
Click on a cell, and then click in the formula bar (the space next to the f_x). Now type the data into the bar and press Enter**.**



10

Data Entry

Enter the following information into your spreadsheet:



Formulas and Functions

- Formulas are equations that perform calculations in your spreadsheet. Formulas always begin with an equals sign (=). When you enter an equals sign into a cell, you are basically telling Excel to "calculate this."

- Functions are Excel-defined formulas. They take data you select and enter, perform calculations on them, and return value (s).

More on Functions

- All functions have a common format – the equals sign followed by the function name followed by the input in parentheses.
- The input for a function can be either:
  - A set of numbers (e.g., "=AVERAGE(2, 3, 4, 5)")
    - This tells Excel to calculate the average of these numbers.
  - A reference to cell (s) (e.g., "=AVERAGE(B1:B18) or "=AVERAGE (B1, B2, B3, B4, B5, B6, B7, B8)"
    - This tells Excel to calculate the average of the data that appear in all the cells from B1 to B8.
    - You can either type these cell references in by hand or by clicking and dragging with your mouse to select the cells.

Functions for Descriptive Statistics

=AVERAGE(first cell:last cell): calculates the mean

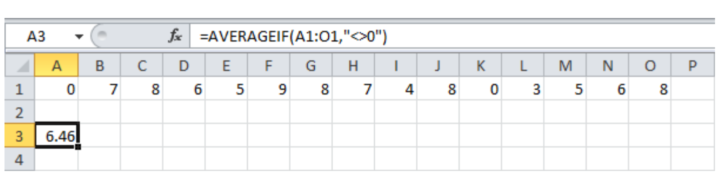
=MEDIAN(first cell:last cell): calculates the median

=MODE(first cell:last cell): calculates the mode

=VAR**P** (first cell:last cell): calculates the variance

=STDEV**P**(first cell:last cell): calculates the standard deviation

- You may directly write the functions for these statistics into cells or the formula bar, OR
- You may use the function wizard (fx in the toolbar)

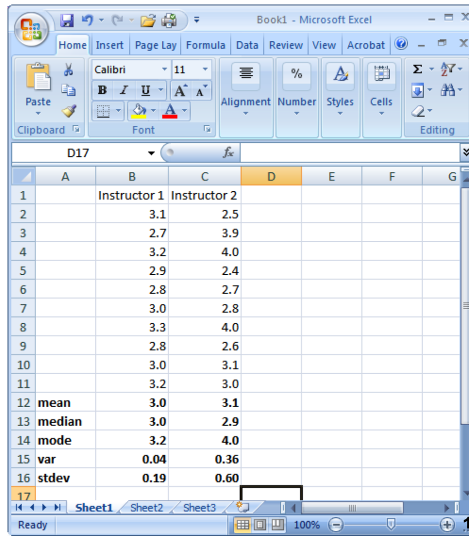


A3 | fx | =AVERAGEIF(A1:O1,"<>0")

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	0	7	8	6	5	9	8	7	4	8	0	3	5	6	8	
2																
3	6.46															
4																

Functions for Descriptive Statistics

- Your Excel spreadsheet should now look like this:



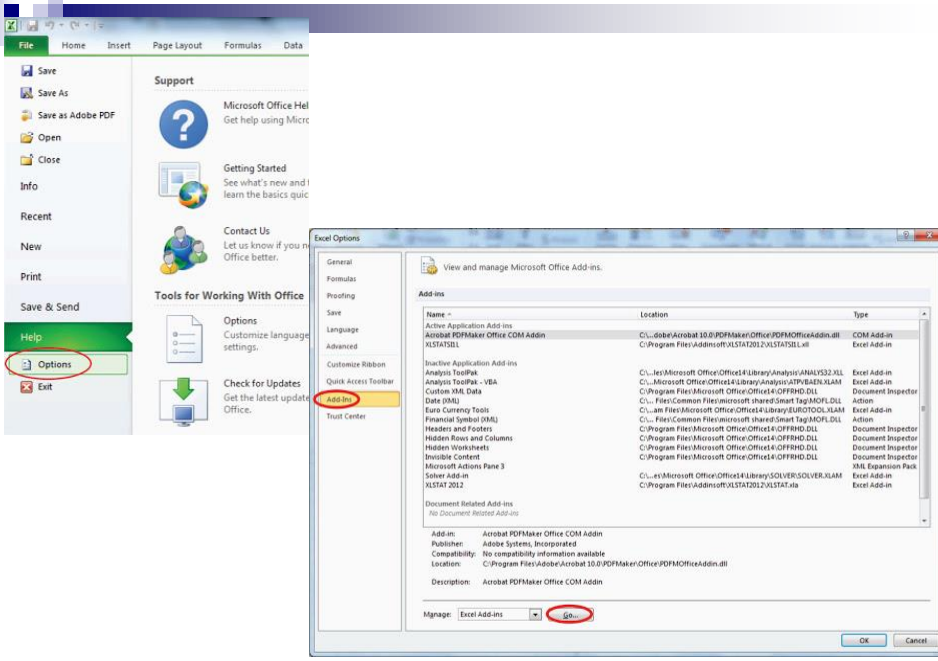
	A	B	C
1		Instructor 1	Instructor 2
2		3.1	2.5
3		2.7	3.9
4		3.2	4.0
5		2.9	2.4
6		2.8	2.7
7		3.0	2.8
8		3.3	4.0
9		2.8	2.6
10		3.0	3.1
11		3.2	3.0
12	**mean**	**3.0**	**3.1**
13	**median**	**3.0**	**2.9**
14	**mode**	**3.2**	**4.0**
15	**var**	**0.04**	**0.36**
16	**stdev**	**0.19**	**0.60**

15

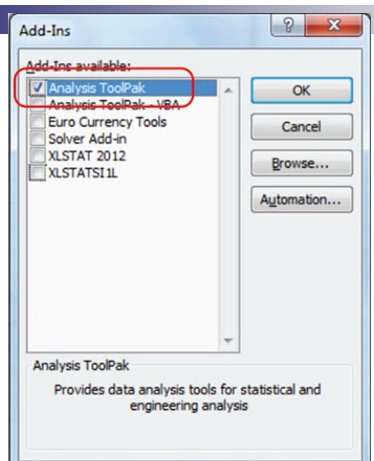
Advanced Microsoft Excel 2007 & 2010

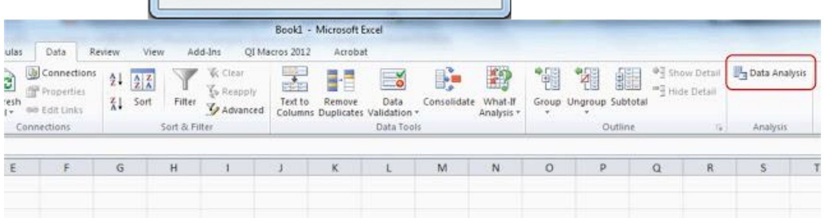
- Click the **Microsoft Office Button** , and then click **Excel Options**.
- Click **Add-Ins**, and then in the **Manage** box, select **Excel Add-ins**.
- Click **Go**.
- In the **Add-Ins available** box, select the **Analysis ToolPak** check box, and then click **OK**.
- If you get prompted that the Analysis ToolPak is not currently installed on your computer, click **Yes** to install it.
- After you load the Analysis ToolPak, the **Data Analysis** command is available in the **Analysis** group on the **Data** tab.

16


File
Home
Insert
Page Layout
Formulas
Data
Save
Save As
Save as Adobe PDF
Open
Close
Info
Recent
New
Print
Save & Send
Help
Options
Exit
Support
Microsoft Office Help
Getting Started
Contact Us
Tools for Working With Office
Options
Check for Updates
Excel Options
General
Formulas
Proofing
Save
Language
Advanced
Customize Ribbon
Quick Access Toolbar
Add-Ins
Trust Center
View and manage Microsoft Office Add-ins.
Add-ins
Manage: Excel Add-ins
Go...
OK
Cancel

17


Add-Ins
Add-Ins available:
Analysis ToolPak
Analysis ToolPak - VBA
Euro Currency Tools
Solver Add-in
XLSTAT 2012
XLSTATSI1L
OK
Cancel
Browse...
Automation...
Analysis ToolPak
Provides data analysis tools for statistical and engineering analysis


Book1 - Microsoft Excel
Data
Review
View
Add-Ins
QI Macros 2012
Acrobat
Connections
Properties
Edit Links
Sort
Filter
Clear
Reapply
Advanced
Text to Columns
Remove Duplicates
Data Validation
Consolidate
What-If Analysis
Group
Ungroup
Subtotal
Show Detail
Hide Detail
Data Analysis
Connections
Sort & Filter
Data Tools
Outline
Analysis

18

Descriptive Statistics and other Statistical methods in Excel

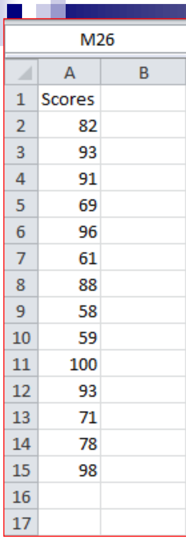
- **Descriptive Statistics and other Statistical methods** : Tools → Data Analysis→ Statistical method.

19

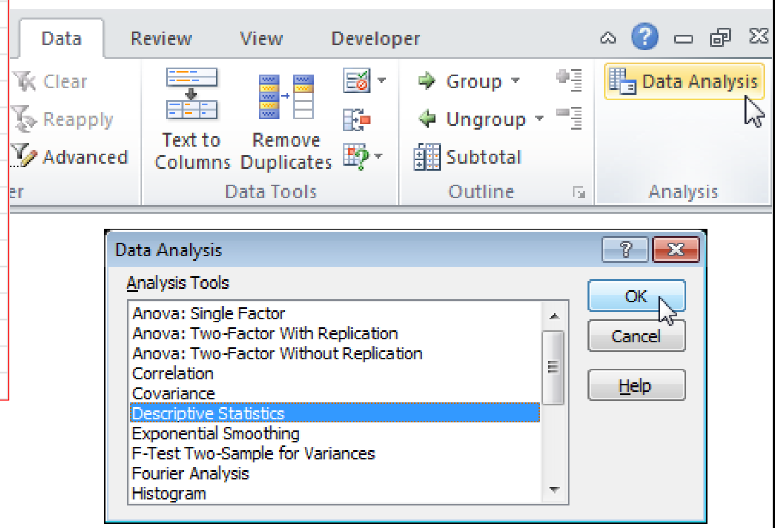
Descriptive Statistics

- Mean, Median, Mode
- Standard Error
- Standard Deviation
- Sample Variance
- Range
- Minimum
- Maximum
- Sum
- Count
- kth Largest
- kth Smallest

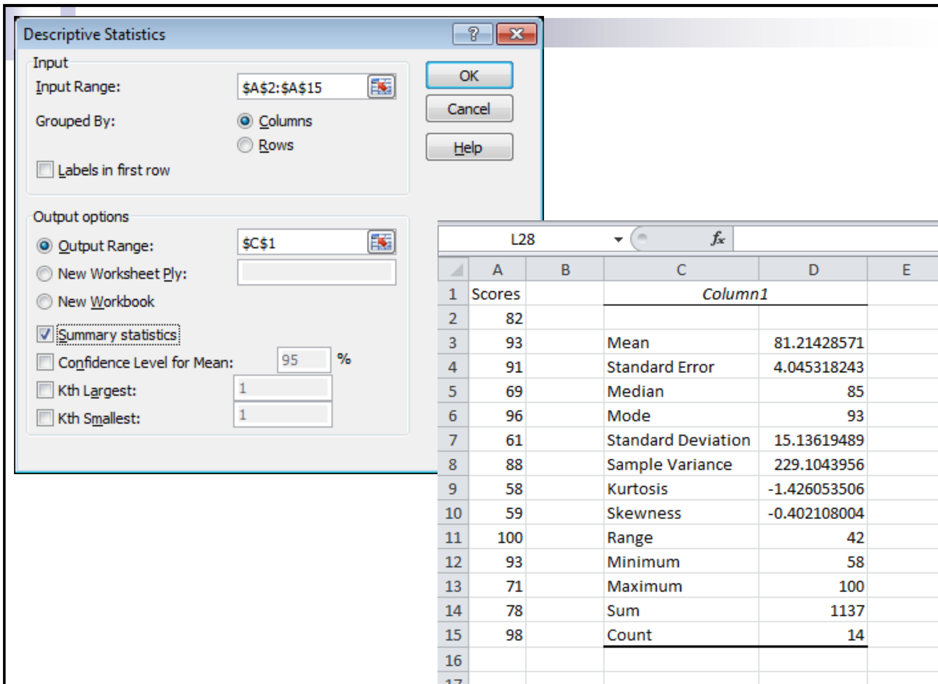
20



Descriptive Statistics and other Statistical methods in Excel 2010



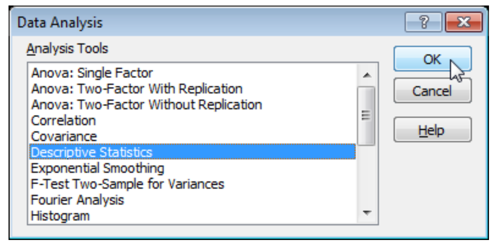
21



22

Analysis ToolPak – in Excel

- Descriptive Statistics
- Correlation
- Linear Regression
- t-Tests
- z-Tests
- ANOVA
- Covariance



Correlation and Regression

- **Correlation** is a measure of the strength of linear association between two variables.

- Every correlation has a *direction* (positive or negative):
  - + correlation: **increase in** one variable are associated with proportional **increase** with another variable.
    - Example: Length weight relation ship. Body height and weight
  - - correlation: **increase in** one variable are associated with ***decrease in*** other variable.
  - Values between -1 and +1
  - Values close to -1 indicate strong negative relationship
  - Values close to +1 indicate strong positive relationship
  - Values close to 0 indicate weak relationship.

- **Linear Regression** is the process of finding a line of best fit through a series of data points.

Calculating Pearson's *r*

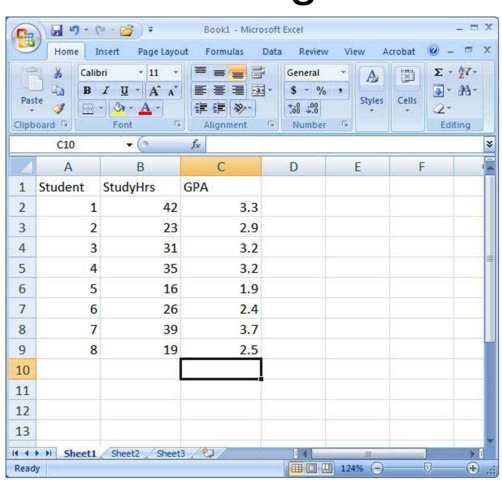
- Correlations are described using the Pearson Product-Moment correlation statistic, or *r* value.

- In Excel, there are many functions that can calculate a correlation statistic, however, we will only use = PEARSON in this class.

Let's say we want to determine if there is a relationship between number of hours spent per week studying for SCBT 33112 Environmental biotechnology and GPA earned in the class at the end of the semester.

To do so, we can calculate Pearson's *r* for our two variables.

Enter the following data into Excel:



Student	StudyHrs	GPA
1	42	3.3
2	23	2.9
3	31	3.2
4	35	3.2
5	16	1.9
6	26	2.4
7	39	3.7
8	19	2.5

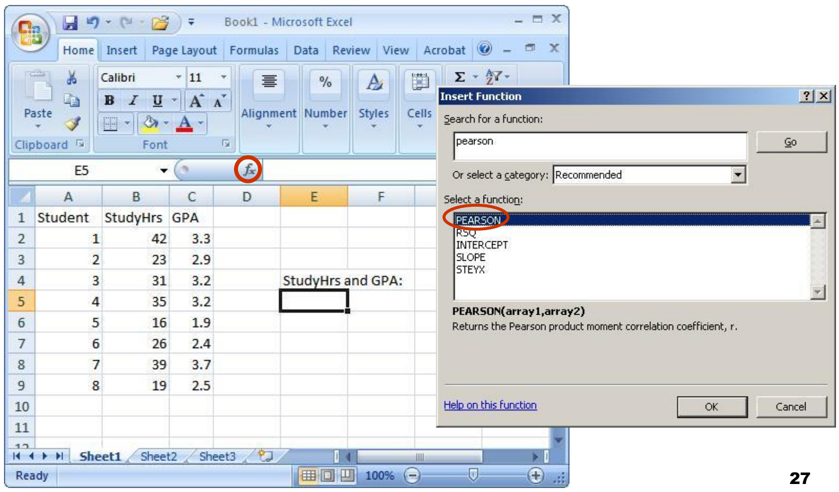
Study Hrs = average number of hours spent per week studying for Environmental biotech

GPA = grade-point average earned in Environmental biotech

Step 1: Select the cell where you want your *r* value to appear (you might want to label it).

Step 2: Click on the function wizard *fx* button.

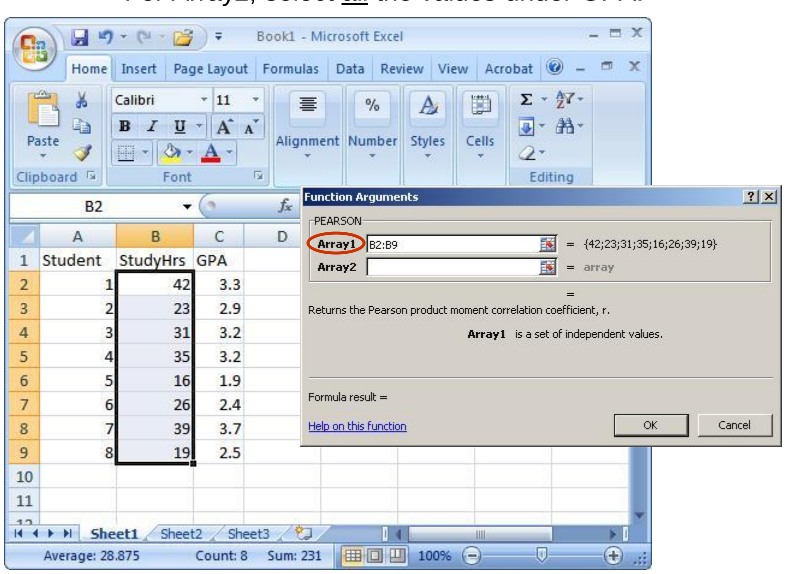
Step 3: Search for and select PEARSON.



27

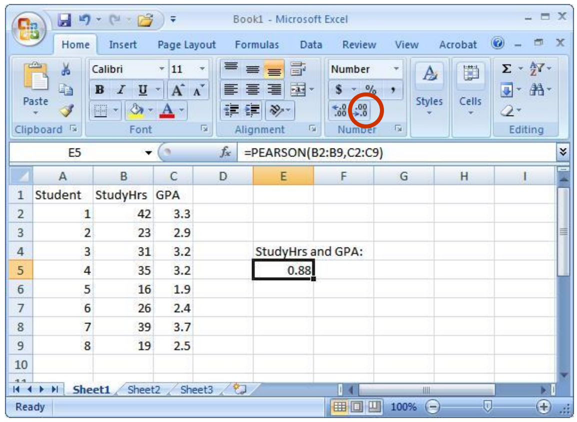
Step 4: For Array1, select <u>all</u> the values under Study Hrs.

For Array2, select <u>all</u> the values under GPA.



28

Step 5: That's it! Once you have your *r* value, don't forget to round to 2 decimal places.



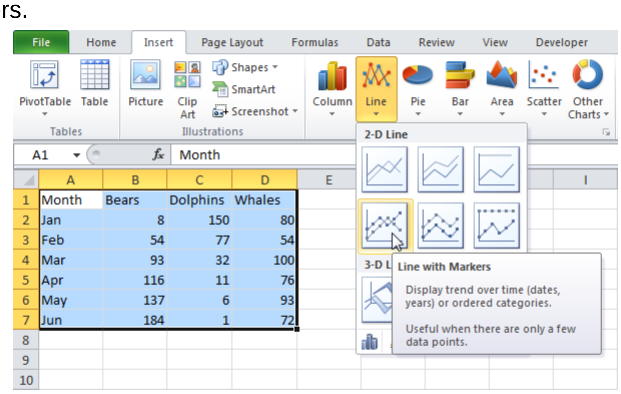
Interpretation: What does the *r* value of 0.88 tell you about the <u>strength</u> and <u>direction</u> of the correlation between Study Hrs and GPA?

Create a Chart in Excel 2010

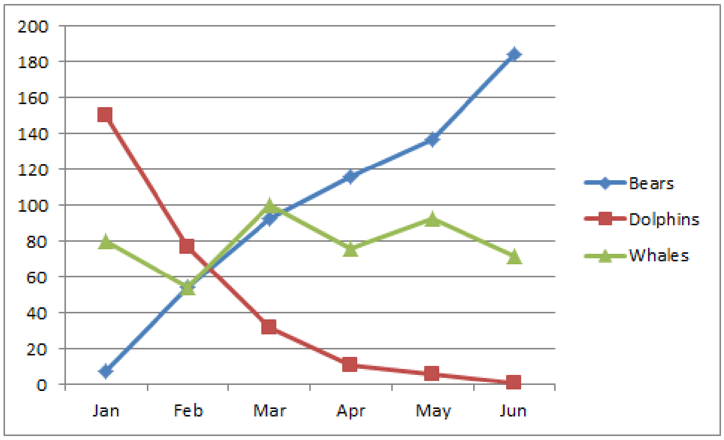
To create a line chart, execute the following steps.

1. Select the range A1:D7.

2. On the **Insert tab**, in the Charts group, **choose Line**, and **select Line** with Markers.



Line graph



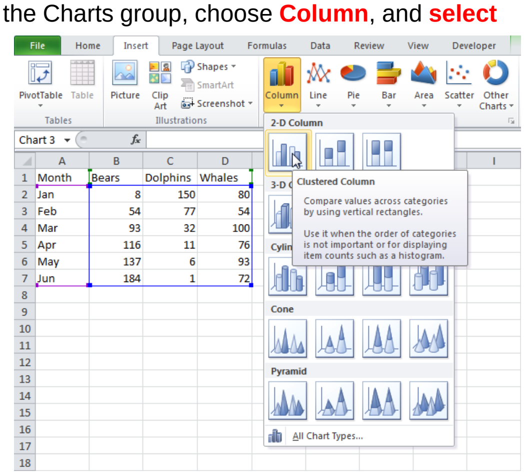
31

Change Chart Type

You can easily change to a different type of chart at any time.

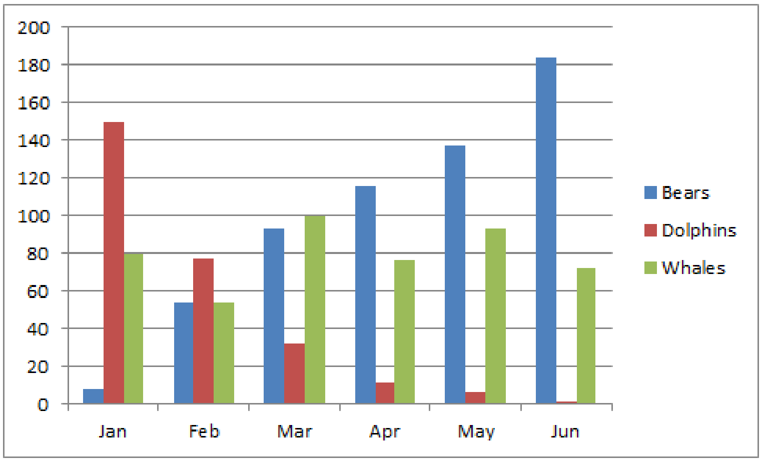
1. Select the chart.

2. On the Insert tab, in the Charts group, choose **Column**, and **select Clustered Column.**



32

Bar Chart - Species density in different months

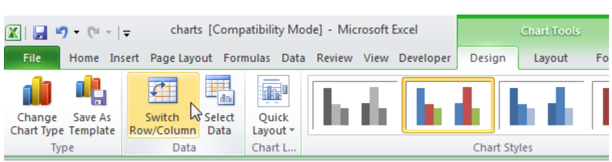


33

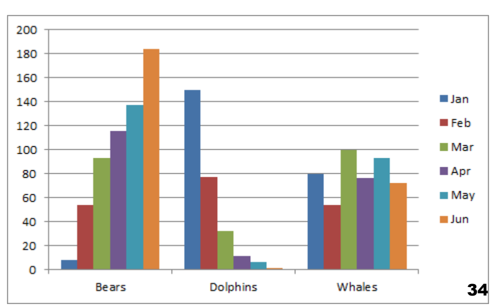
Switch Row/Column

If you want the animals, displayed on the vertical axis, to be displayed on the horizontal axis instead, execute the following steps.

1. Select the chart. The Chart Tools contextual tab activates.
2. On the Design tab, **click Switch Row/Column**.



Bar Chart - Species density in different months



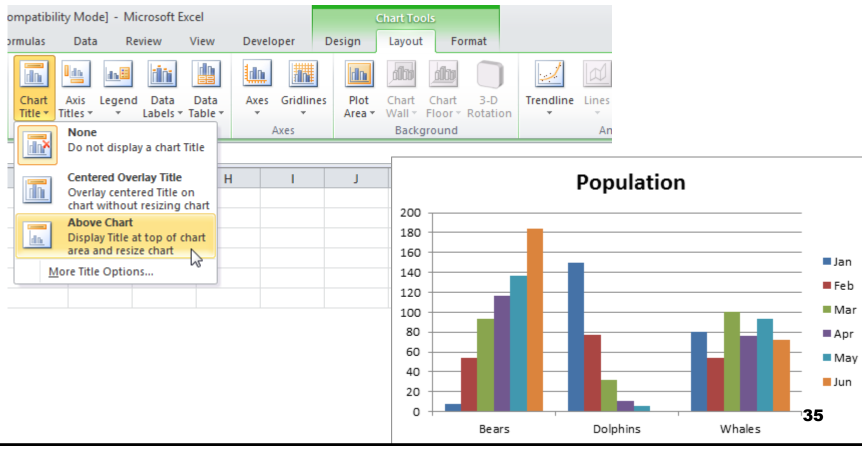
34

Add Chart Title

To add a chart title, execute the following steps.

1. Select the chart. The Chart Tools contextual tab activates.

2. On the Layout tab, click Chart Title, Above Chart.



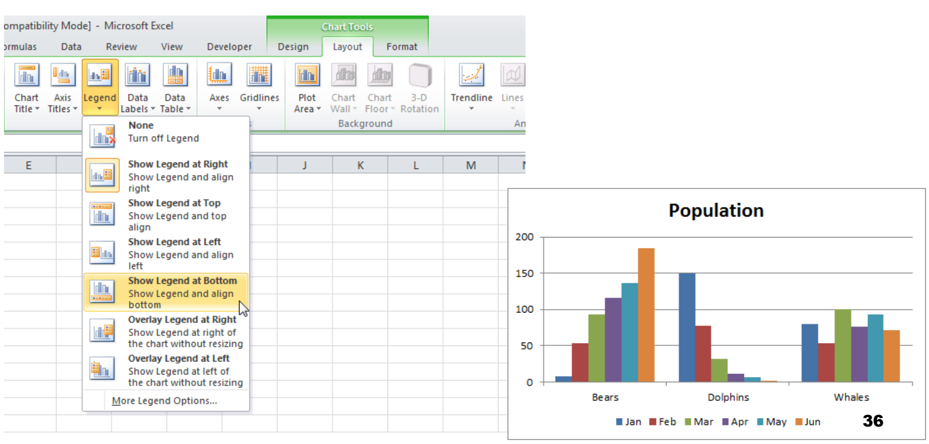
35

Legend Position

By default, the legend appears to the right of the chart. To move the legend to the bottom of the chart,

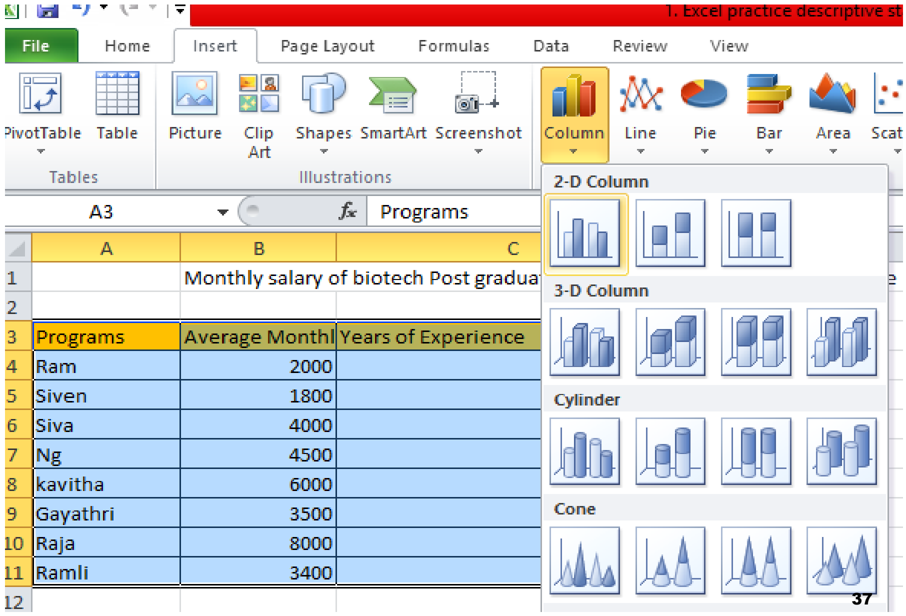
1. Select the chart. The Chart Tools contextual tab activates.

2. **On the Layout tab**, click **Legend**, Show Legend at Bottom.

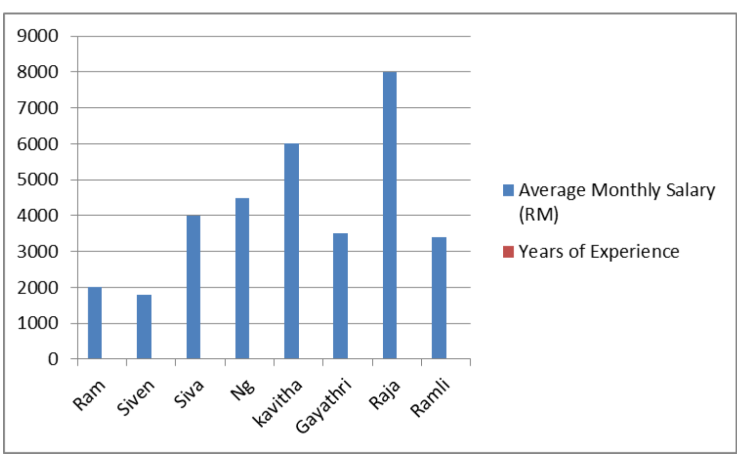


36

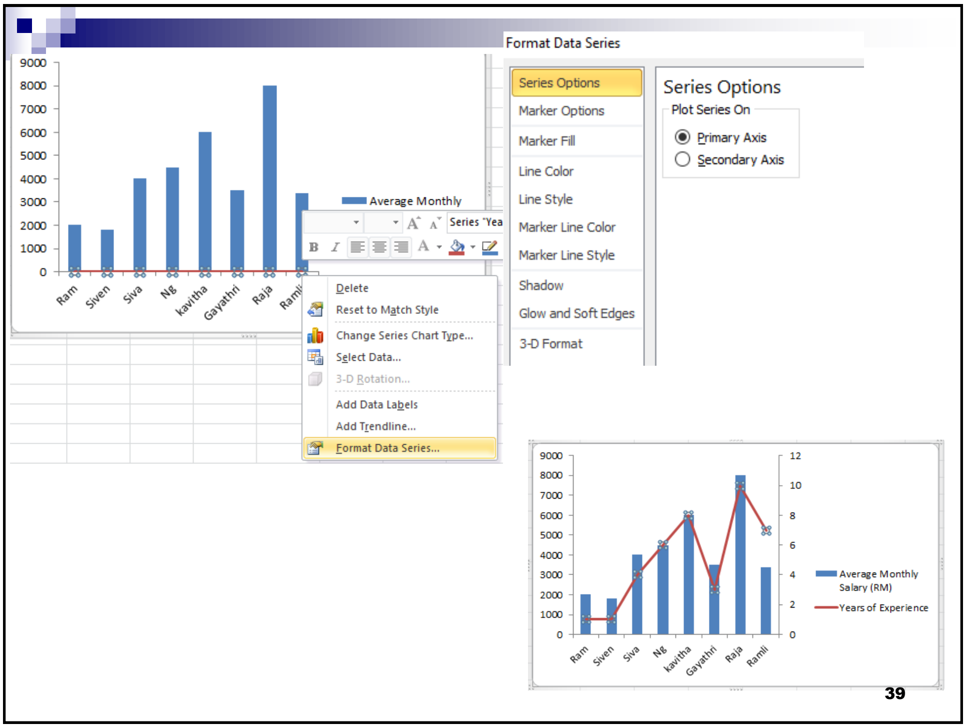
Creating two axis in a graph



37



38


Format Data Series
Series Options
Marker Options
Marker Fill
Line Color
Line Style
Marker Line Color
Marker Line Style
Shadow
Glow and Soft Edges
3-D Format
Series Options
Plot Series On
Primary Axis
Secondary Axis
Average Monthly
Delete
Reset to Match Style
Change Series Chart Type...
Select Data...
3-D Rotation...
Add Data Labels
Add Trendline...
Format Data Series...
Ram
Siven
Siva
Ng
kavitha
Gayathri
Raja
Ramli
Average Monthly Salary (RM)
Years of Experience
39

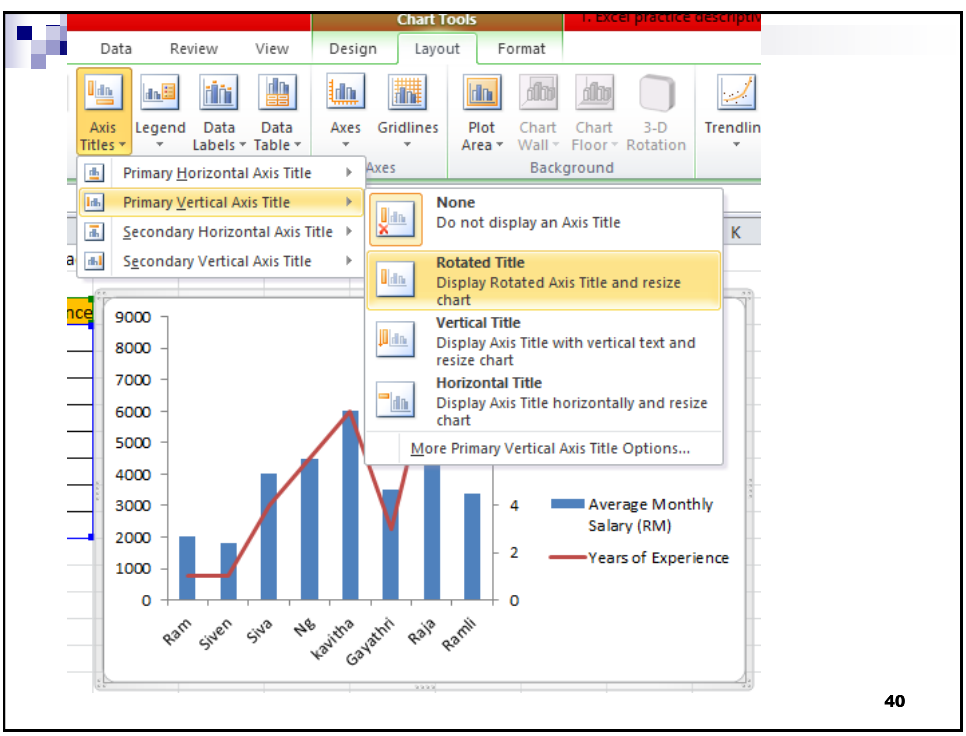
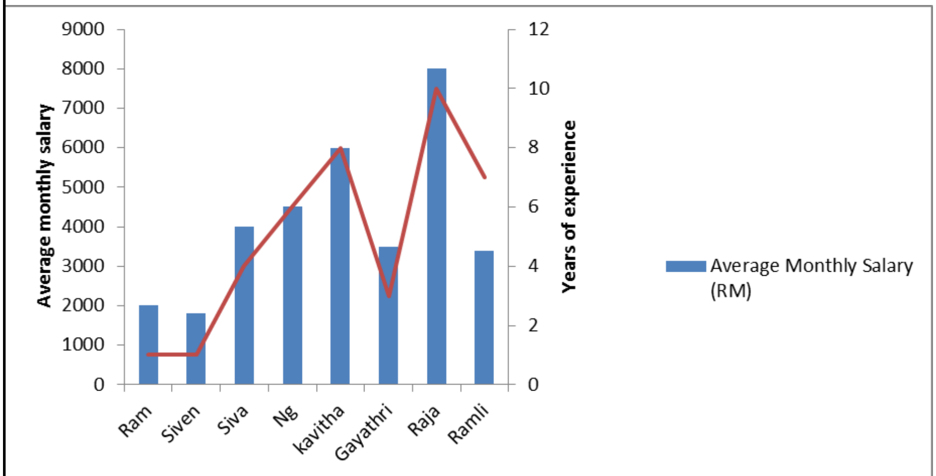


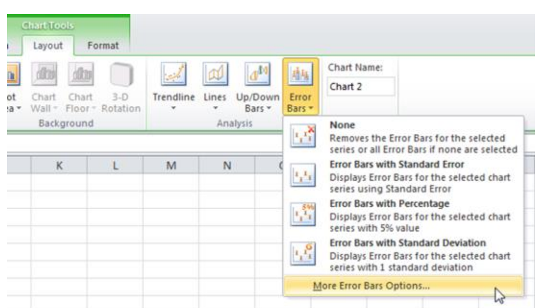
Chart Tools
Data
Review
View
Design
Layout
Format
Axis Titles
Legend
Data Labels
Data Table
Axes
Gridlines
Plot Area
Chart Wall
Chart Floor
3-D Rotation
Trendlin
Axes
Background
Primary Horizontal Axis Title
Primary Vertical Axis Title
Secondary Horizontal Axis Title
Secondary Vertical Axis Title
None
Do not display an Axis Title
Rotated Title
Display Rotated Axis Title and resize chart
Vertical Title
Display Axis Title with vertical text and resize chart
Horizontal Title
Display Axis Title horizontally and resize chart
More Primary Vertical Axis Title Options...
Ram
Siven
Siva
Ng
kavitha
Gayathri
Raja
Ramli
Average Monthly Salary (RM)
Years of Experience
40



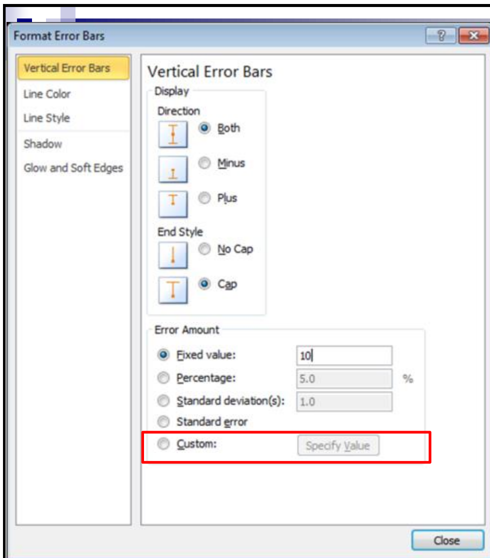
41

Adding Error Bars in Excel graph

1. Select the chart. The Chart Tools contextual tab activates.
2. On the Layout tab, click Error Bars, More Error Bars Options...
3. Choose a Direction. Click Both.
4. Choose an End Style. Click Cap.
5. Click Fixed value and enter the value 10.



42



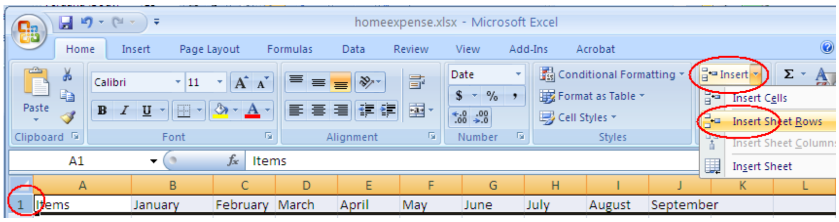
If you have SD values select custom and specify the values by selecting reference cell

43

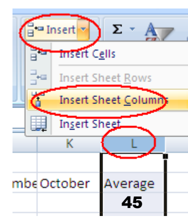
Other formatting functions in Excel

Insert a Row/Column

- Insert a row:
  - Select the row you would like to insert above
    - Clicking on the row number tab.
  - In **Home** tab, go to **Insert** and select **Insert Sheet Rows**.

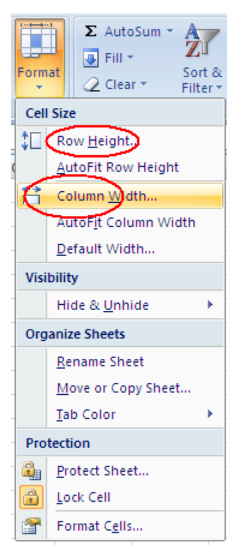


- Insert a column:
  - Select the column you would like to insert next to it
    - Clicking on the column letter tab such as L.
  - In **Home** tab, go to **Insert** and select **Insert Sheet Column**.



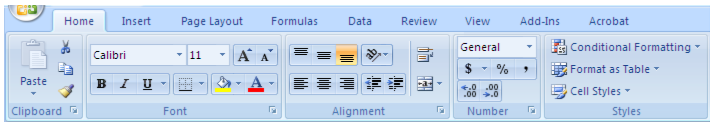
Change Column Width or Row Height

- **Column Width**
  - Drag the border between two columns to adjust a column width.
  - Adjust column width for a group of columns
    - Highlight the columns you want to adjust their width.
    - In **Home** tab, go to **Format** and select **Column Width**...
    - Enter a number of characters for column width. Click on OK.
- **Row Height**
  - Drag the border between two rows to adjust a row width.
  - Adjust row width for a group of rows
    - Highlight the rows you would like to change their height.
    - In **Home** tab, go to **Format** and select **Row Height**.
    - Enter a number of the row height and click on **OK**.
      - One point=.035 cm



46

Format a Worksheet

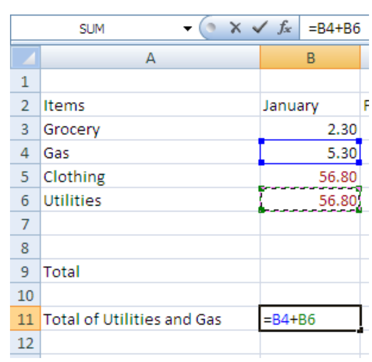


- Change the font size, color, and the background of a cell or group of cells.
- Select the cells you'd like to change. Then select a formatting tool.
- To show cell borders, highlight the cells and select a border.

47

Creating Basic Formula

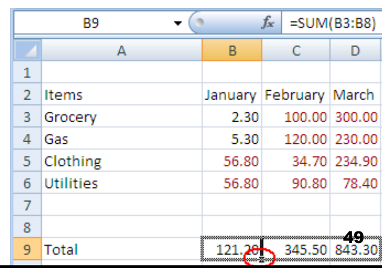
- You conduct a mathematical calculation in Excel by typing a simple formula into a cell. An Excel formula always begins with an equal sign (=).
- Math **operators**
  - Addition: +
  - Subtraction:-
  - Multiplication:*
  - Division:/
- Example: Gas + Utilities
  - Click on the cell that displays the expense of Gas and Utilities.
  - Enter =.
  - Click on the Gas cell for January.
  - Enter +.
  - Click on the Utilities cell for January
  - Hit Enter key.



48

Copy a Formula

- You may copy the same formula onto a series of cells.
  - Example, a total expense in each of all 12 months.
  - Select the total cell for January.
  - Drag the bottom right corner of the cell to expand to the December total cell.
  - The total expense is then calculated for all 12 months.



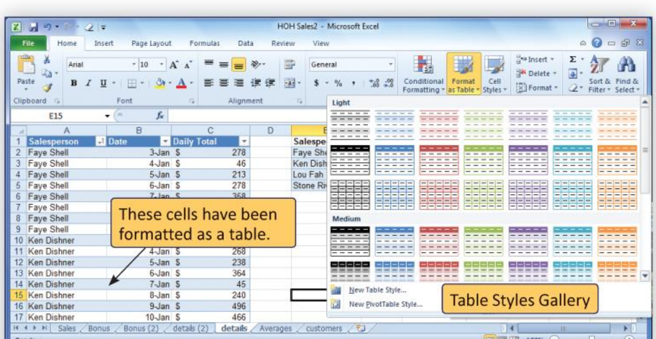
B9 =SUM(B3:B8)

	A	B	C	D
1				
2	Items	January	February	March
3	Grocery	2.30	100.00	300.00
4	Gas	5.30	120.00	230.00
5	Clothing	56.80	34.70	234.90
6	Utilities	56.80	90.80	78.40
7				
8				
9	Total	121.20	345.50	843.30

49

Define Data as a Table

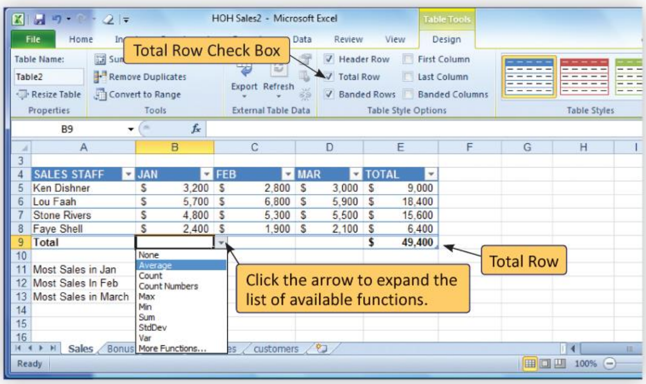
- Select the data for your table.
- On the *Home* tab, in the *Styles* group, click the **Format as Table** button to display the *Table Styles* gallery.
- Click the style you want to use for your table.
- Excel will automatically populate the *Format as Table* dialog box with the selected data range.
- Be sure to check the **My table has headers** check box if appropriate.
- Click **OK** to create the table.



50

To Add a Total Row to a Table

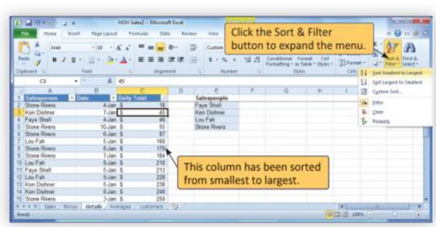
- On the *Table Tools Design* tab, in the *Table Style Options* group, click the **Total Row** check box.
- In the Total row at the bottom of the table, click the column where you want to add a total.
- Click the arrow, and select the function you want to use.



51

To Sort the Data in Your Worksheet

- Click any cell in the column to sort.
- On the *Home* tab, in the *Editing* group, click the **Sort & Filter** button.
- Click the sorting option you want. The sorting options change depending on the type of data in the column you are sorting by.
  - If the numbers in the column are formatted as dates, Excel detects this and offers sorting options **Sort Oldest to Newest** and **Sort Newest to Oldest.**
  - If the column contains text, the sort options are **Sort A to Z** and **Sort Z to A.**
  - If the column contains numbers, the sort options are **Sort Smallest to Largest** and **Sort Largest to Smallest**.



52

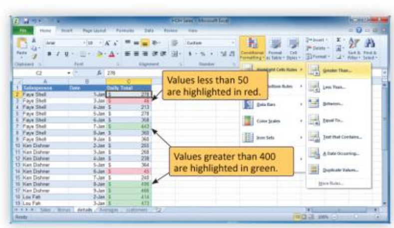
Applying Conditional Formatting with Highlight Cells Rules

- Conditional formatting with **Highlight Cells Rules** allows you to **define formatting for cells that meet specific numerical or text criteria (e.g., greater than a specific value or containing a specific text string).**
- Use this type of conditional formatting when you want to highlight cells based on criteria you define.

53

To Highlight Cells with Conditional Formatting

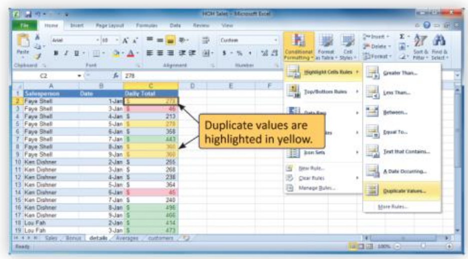
- Select the data you want to apply conditional formatting to.
- On the *Home* tab, in the *Styles* group, click the **Conditional Formatting** button.
- From the menu, point to **Highlight Cells Rules** and click the option you want:
  - **Greater Than . . .**
  - **Less Than . . .**
  - **Between . . .**
  - **Equal To . . .**
  - **Text That Contains . . .**
  - **A Date Occurring . . .**
  - **Duplicate Values . . .**
- Each option opens a dialog box where you can enter the condition to compare selected cells to and the formatting to apply when selected cells match the condition.
- Click **OK** to apply the conditional formatting.



54

To Remove Conditional Formatting

- Select the cells you want to remove the formatting from.
- On the *Home* tab, in the *Styles* group, click the **Conditional Formatting** button.
- Point to **Clear Rules,** and click the option you want from the menu:
  - **Clear Rules from Selected Cells**
  - **Clear Rules from Entire Sheet**
  - **Clear Rules from This Table** (available if the selected cells are part of a table)

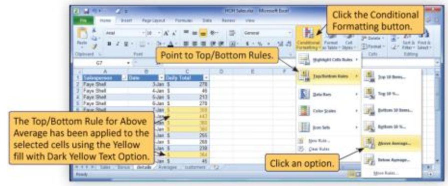


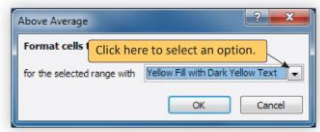
Applying Conditional Formatting with Top/Bottom Rules

- One way to analyze worksheet data is to compare cell values to other cell values.
- To highlight the highest or lowest values or values that are above or below the average, use conditional formatting **Top/Bottom Rules**.
- When you use Top/Bottom Rules conditional formatting, Excel automatically finds the highest, lowest, and average values to compare values to, rather than asking you to enter criteria (as you do when using Highlight Cells Rules).

To Highlight Cells with Conditional Formatting Top/Bottom Rules

1. Select the data you want to apply conditional formatting to.
2. On the *Home* tab, in the *Styles* group, click the **Conditional Formatting** button.
3. From the menu, point to **Top/ Bottom Rules** and click the option:
   - **Top 10 Items . . .**
   - **Top 10% . . .**
   - **Bottom 10 Items . . .**
   - **Bottom 10% . . .**
   - **Above Average . . .**
   - **Below Average . . .**
4. Each option opens a dialog box where you can modify the condition and select formatting to apply when cells match the condition.
5. Click **OK** to apply the conditional formatting.



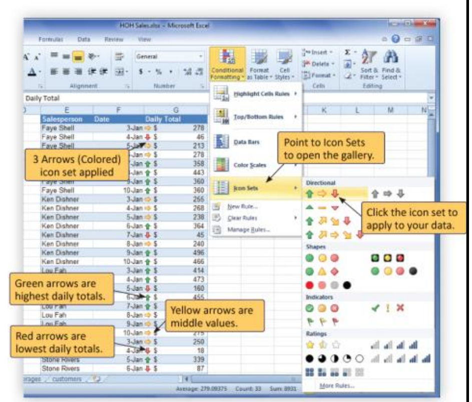


57

Applying Conditional Formatting with Data Bars…

1. Select the data you want to apply conditional formatting to.
2. On the *Home* tab, in the *Styles* group, click the **Conditional Formatting** button.
3. From the menu, point to one of the options, and then click the specific style of formatting you want.

- **Data Bars** —Display a color bar (gradient or solid) representing the cell value in comparison to other values (cells with higher values have longer data bars).
- **Color Scales** —Color the cells according to one of the color scales [e.g., red to green (bad/low to good/high) or blue to red (cold/low to hot/high)].
- **Icon Sets** —Display a graphic in the cell representing the cell value in relation to other values.



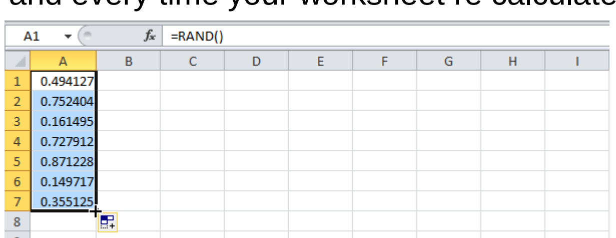
58

Generating Random numbers in Microsoft Excel

Select a Random Decimal Value Between 0 and 1

If you want to generate a random decimal value between 0 and 1, simply use the Excel Rand function in any cell of your worksheet:
=RAND()

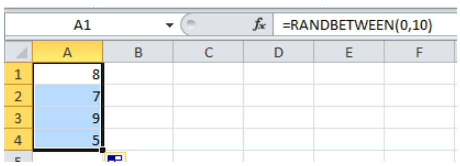
This function will generate a different random decimal, between 0 and 1 and every time your worksheet re calculates.

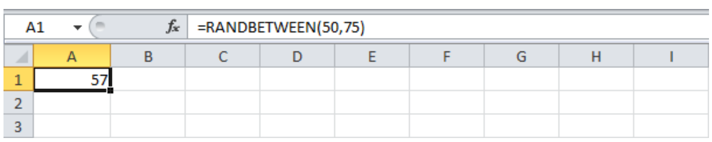


60

Select a Random Integer Between Two Supplied Integers

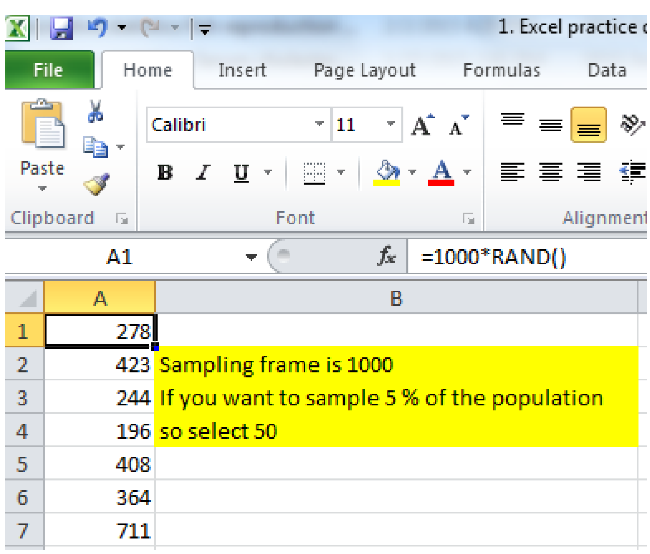
The Excel Randbetween function is used to generate a random integer between two supplied integers. For example:
=RANDBETWEEN(0, 10)
=RANDBETWEEN(50, 75)





61

If you know the sampling frame in advance then you follow this



62

Statistics using SPSS

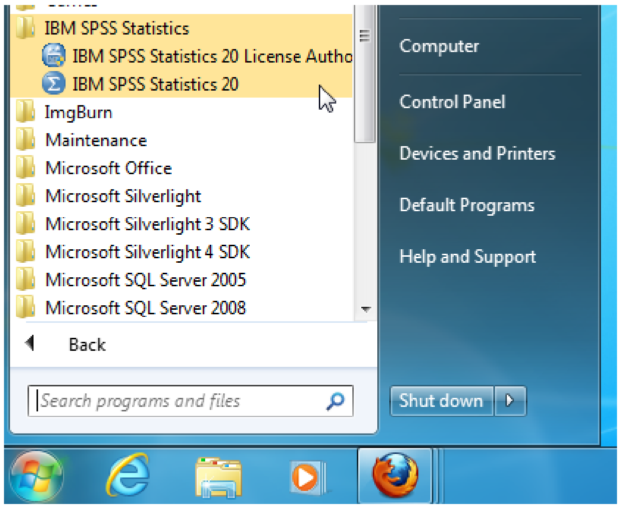
What is SPSS?

- Originally it is an acronym of Statistical Package for the Social Science but now it stands for Statistical Product and Service Solutions
- SPSS is a powerful program which provides many ways to rapidly examine data and test scientific questions.
- One of the most **popular statistical packages** which can perform highly complex data manipulation **and analysis** with simple instructions.
- The program also is capable of producing high-quality graphs and tables

64

Starting SPSS

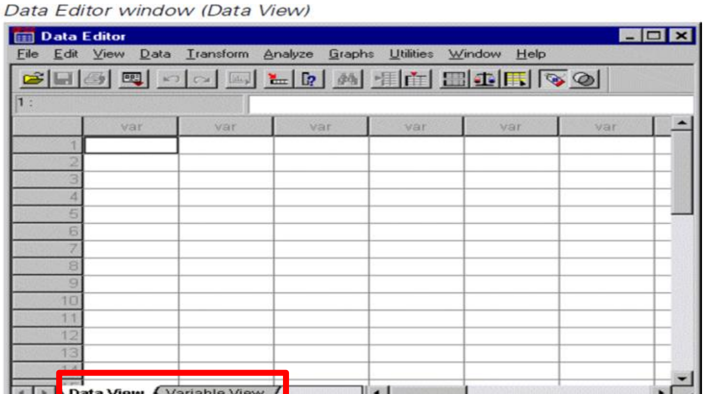
- Start → All Programs → IBM SPSS statistics 20



Data Editor

- The Data Editor window displays the contents of the working dataset.
- It is arranged in a spreadsheet format that contains **variables in columns** and **cases in rows**.
- There are two sheets in the window.
  - The **Data View** is the sheet that is visible when you first open the Data Editor and contains the data.
  - The other one is **Variable view**

Data Editor window (Data View)



Data and Variable View

Data View:

- Allows you to entering the data and examine your actual data.

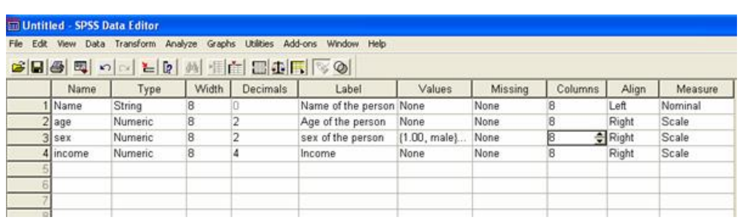
Variable View:

- Variable view is for Naming and define the variables
- Is it a string or numeric variable?
- Are there labels for it? In Variable view, you can add labels to variables so your results will be easier to understand.

67

Variable View window & Defining variable names and their properties in SPSS

- This sheet contains **information about the data set** that is stored with the dataset
- **Name of the variable**
- **Type**
- **Width**
- **Decimal**
- **Label**
- **Values**
- **Missing**
- **column**
- **Align**
- **Measure**

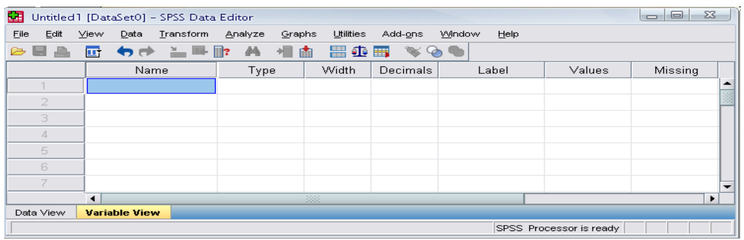


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	Name	String	8	0	Name of the person	None	None	8	Left	Nominal
2	age	Numeric	8	2	Age of the person	None	None	8	Right	Scale
3	sex	Numeric	8	2	sex of the person	{1.00, male}...	None	8	Right	Scale
4	income	Numeric	8	4	Income	None	None	8	Right	Scale
5										
6										
7										

68

Name

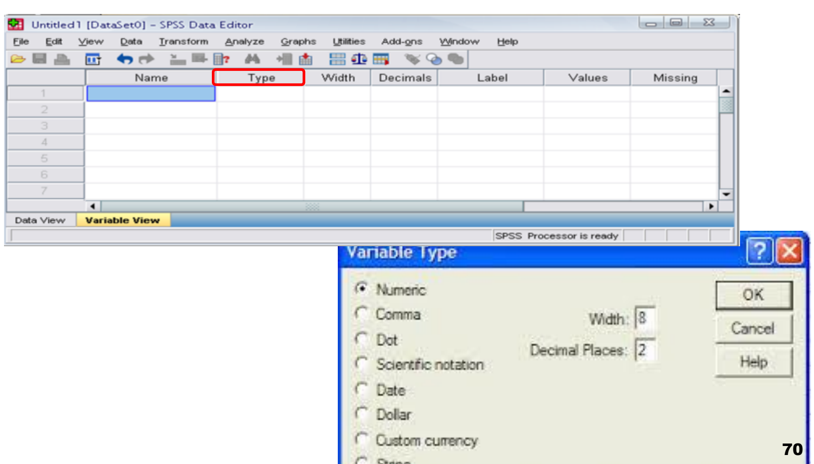
- The first character of the variable name must be alphabetic
  - Variable names must be unique, and have to be less than 64 characters.
  - Spaces are NOT allowed.



69

Variable View window: Type

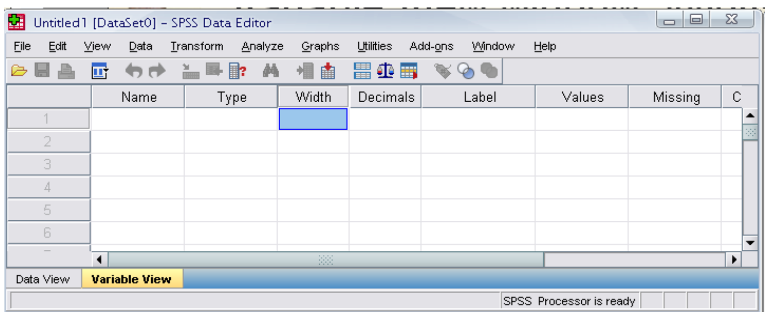
- Click on the 'type' box.
- The two basic types of variables that you will use are **numeric and string**. This column enables you to specify the type of variable.



70

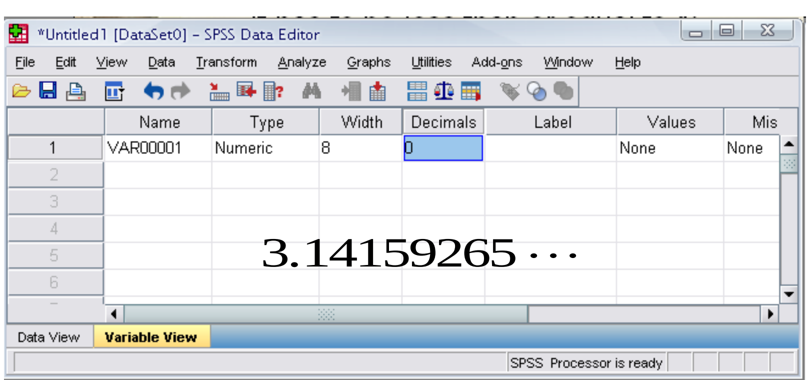
Variable View window: Width

Width allows you to determine the number of characters SPSS will allow to be entered for the variable



Variable View window: Decimals

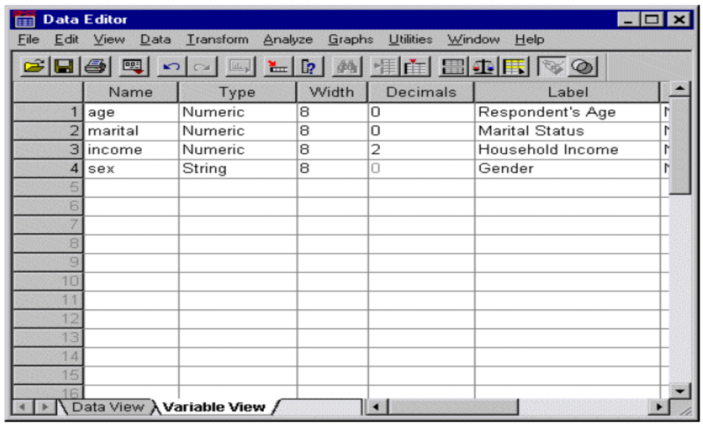
- Number of decimals
- It has to be less than or equal to 16



Variable View window: Label

- You can specify the details of the variable

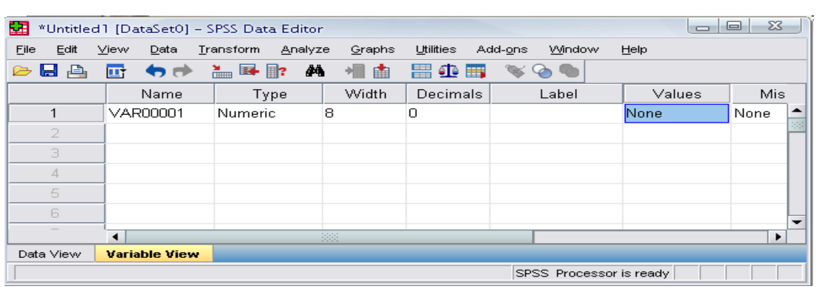
Variable labels entered in Variable View



73

Variable View window: Values

This is used and to suggest which numbers represent which categories when the variable represents a category

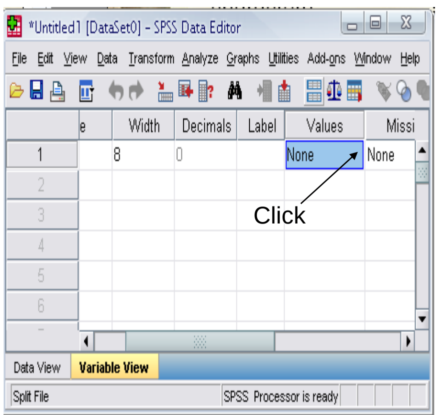


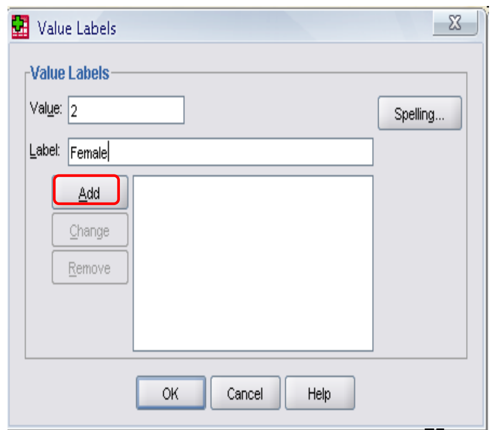
Male 1; Female 2
Treatment group 1; Control group 2

74

Defining the value labels

- Click the cell in the values column as shown below
- Value 1 for male and 2 for female and then click OK





Columns

- The columns property tells SPSS how wide the column should be for each variable.

- Don't confuse this one with width, which indicates how many digits of the number will be displayed. The column size indicates **how much space is allocated**

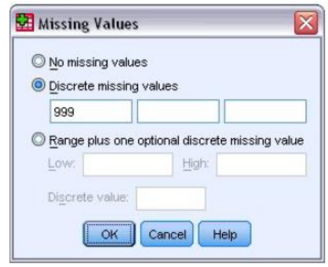
Align

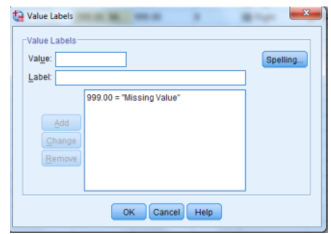
- The alignment property indicates whether the information in the Data View should be left-justified, right-justified, or centered

Missing Values for a Numeric Variable

- Survey respondents may refuse to answer certain questions, may not know the answer.
- If you don't filter or identify these data, your analysis may not provide accurate results.

1. Click the **Variable View** tab at the bottom of the Data Editor window.
2. Click the *Missing* cell in the row, and then click the button on the right side of the cell to open the Missing Values dialog box. In this dialog box, you can specify up to three distinct missing values, or you can specify a range of values plus one additional discrete value.
3. Select **Discrete missing values**.
4. Type 999 in the first text box and leave the other two text boxes empty.
5. Click **OK** to save your changes and return to the Data Editor. Now that the missing data value has been added, a label can be applied to that value.
6. Click the *Values* cell in the *age* row, and then click the button on the right side of the cell to open the Value Labels dialog box.
7. Type 999 in the Value field.
8. Type No Response in the Label field.
9. Click **Add** to add this label to your data file.
10. Click **OK** to save your changes and return to the Data Editor.

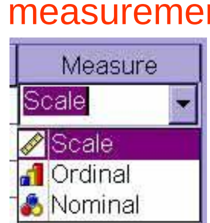




77

Measure

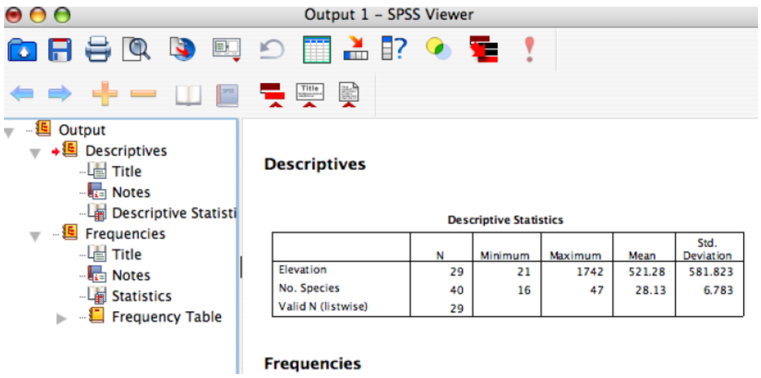
- The Measure property indicates the level of measurement.
- Since SPSS does not differentiate between interval and ratio levels of measurement, both of these quantitative variable types are lumped together as **"scale".**
- Nominal and ordinal levels of measurement, however, **are** differentiated



78

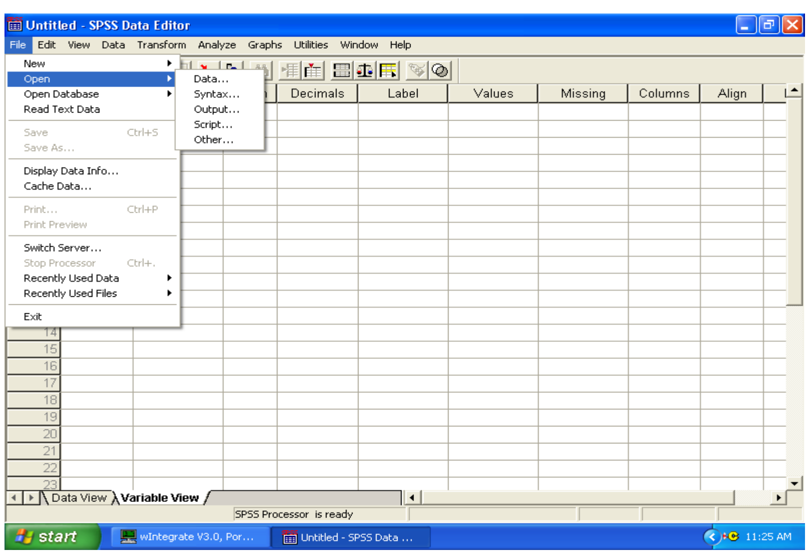
SPSS Viewer

- All output from statistical analyses and graphs is displayed to the SPSS Viewer window.

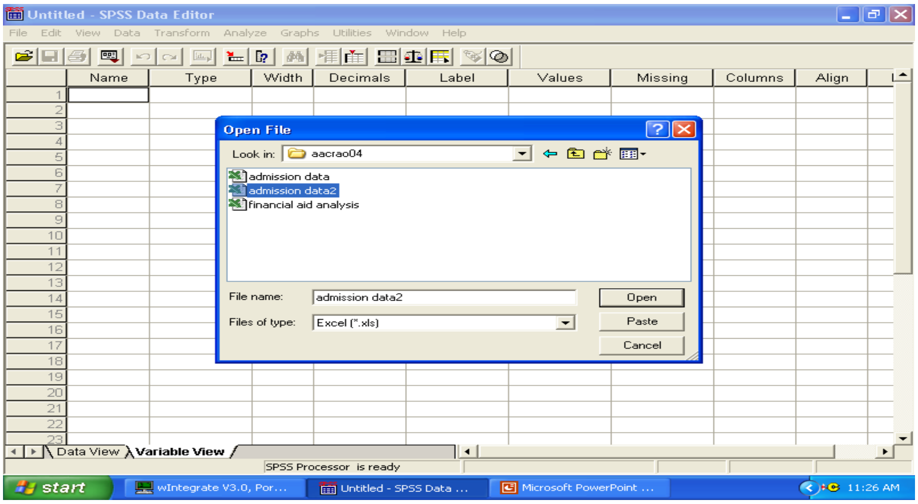


How to access data into SPSS from your computer - **Opening an Excel file**

1. Go to "File" then "Open" and click on "Data"



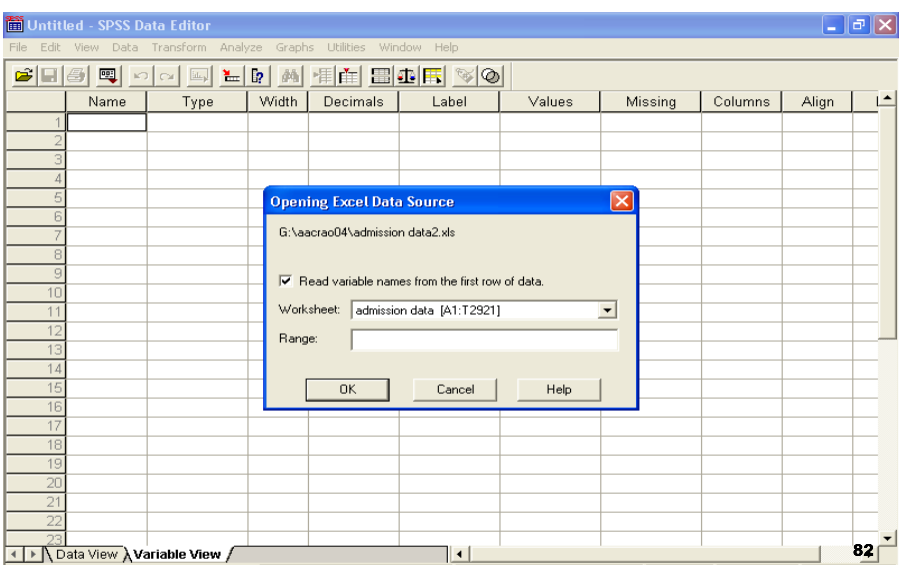
2. Be sure that the “Files of type” is set on Excel (*.xls)



3. Click on the Excel File you want to open.
4. Either Double-Click the file name or click the “Open” button

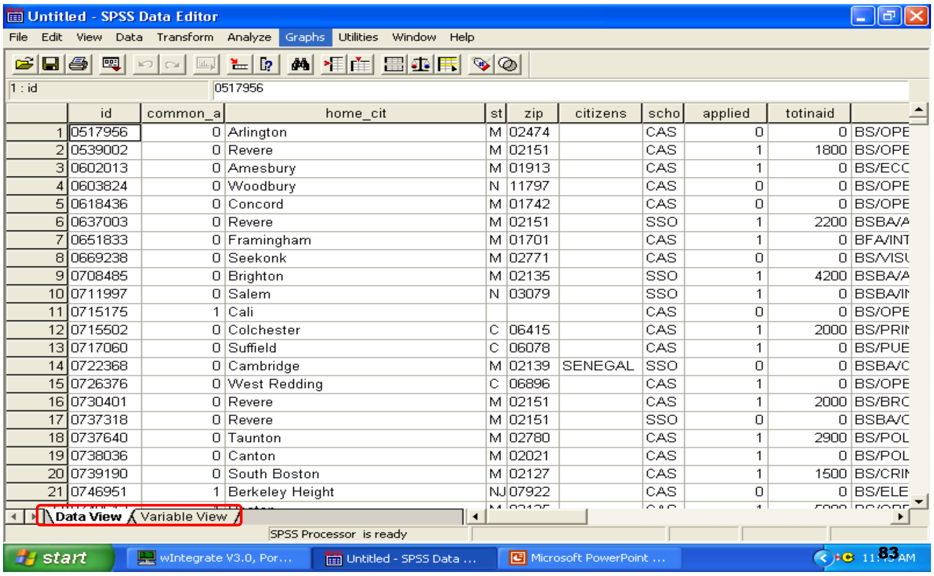
81

5. Choose the name of the worksheet that your data is in. (You can only choose one worksheet at a time)
6. Click “OK”

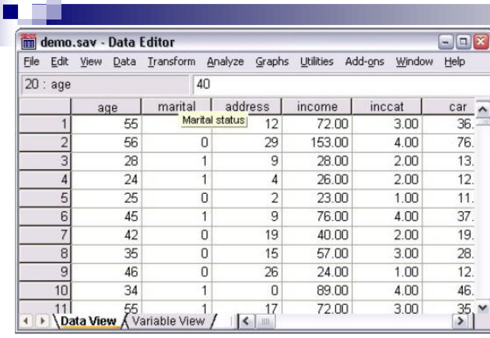


82

If you look at the bottom left, you'll see tabs for **Data View and Variable View**



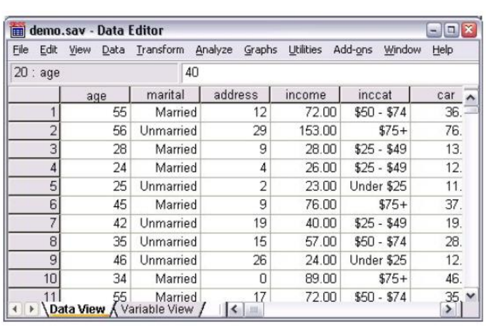
	id	common_a	home_cit	st	zip	citizens	scho	applied	totinaid	
1	0517956	0	Arlington	M	02474		CAS	0	0	BS/OPE
2	0539002	0	Revere	M	02151		CAS	1	1800	BS/OPE
3	0602013	0	Amesbury	M	01913		CAS	1	0	BS/ECC
4	0603824	0	Woodbury	N	11797		CAS	0	0	BS/OPE
5	0618436	0	Concord	M	01742		CAS	0	0	BS/OPE
6	0637003	0	Revere	M	02151		SSO	1	2200	BSBA/A
7	0651833	0	Framingham	M	01701		CAS	1	0	BFA/INT
8	0669238	0	Seekonk	M	02771		CAS	0	0	BS/VISU
9	0708485	0	Brighton	M	02135		SSO	1	4200	BSBA/A
10	0711997	0	Salem	N	03079		SSO	1	0	BSBA/IN
11	0715175	1	Cali				CAS	0	0	BS/OPE
12	0715502	0	Colchester	C	06415		CAS	1	2000	BS/PRIN
13	0717060	0	Suffield	C	06078		CAS	1	0	BS/PUE
14	0722368	0	Cambridge	M	02139	SENEGAL	SSO	0	0	BSBA/C
15	0726376	0	West Redding	C	06896		CAS	1	0	BS/OPE
16	0730401	0	Revere	M	02151		CAS	1	2000	BS/BRC
17	0737318	0	Revere	M	02151		SSO	0	0	BSBA/C
18	0737640	0	Taunton	M	02780		CAS	1	2900	BS/POL
19	0738036	0	Canton	M	02021		CAS	1	0	BS/POL
20	0739190	0	South Boston	M	02127		CAS	1	1500	BS/CRIN
21	0746951	1	Berkeley Height	NJ	07922		CAS	0	0	BS/ELE



	age	marital	address	income	inccat	car
1	55		12	72.00	3.00	36.
2	56	0	29	153.00	4.00	76.
3	28	1	9	28.00	2.00	13.
4	24	1	4	26.00	2.00	12.
5	25	0	2	23.00	1.00	11.
6	45	1	9	76.00	4.00	37.
7	42	0	19	40.00	2.00	19.
8	35	0	15	57.00	3.00	28.
9	46	0	26	24.00	1.00	12.
10	34	1	0	89.00	4.00	46.
11	55	1	17	72.00	3.00	35

To view value Labels

You can use the Value Labels button on the toolbar



	age	marital	address	income	inccat	car
1	55	Married	12	72.00	$50 - $74	36.
2	56	Unmarried	29	153.00	$75+	76.
3	28	Married	9	28.00	$25 - $49	13.
4	24	Married	4	26.00	$25 - $49	12.
5	25	Unmarried	2	23.00	Under $25	11.
6	45	Married	9	76.00	$75+	37.
7	42	Unmarried	19	40.00	$25 - $49	19.
8	35	Unmarried	15	57.00	$50 - $74	28.
9	46	Unmarried	26	24.00	Under $25	12.
10	34	Married	0	89.00	$75+	46.
11	55	Married	17	72.00	$50 - $74	35

Descriptive value labels are now displayed to make it easier to interpret the responses

Running Analysis in SPSS

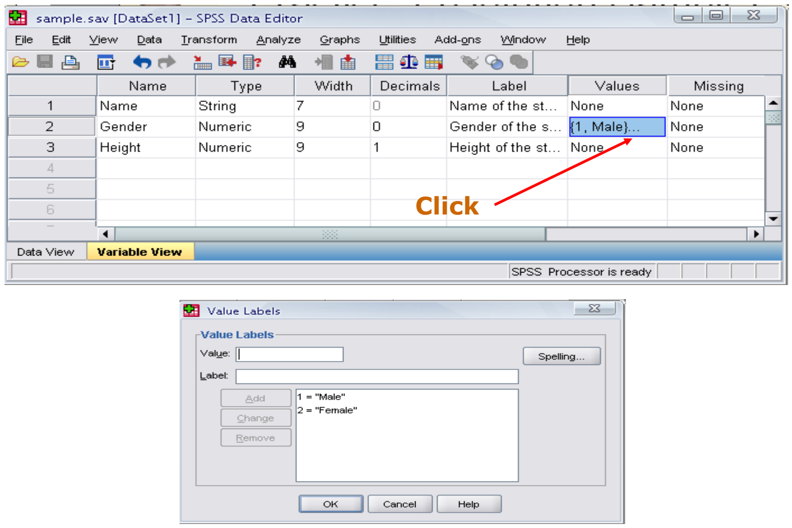
Practice 1

- How would you put the following information into SPSS?

Name	Gender	Height
JAUNITA	2	5.4
SALLY	2	5.3
DONNA	2	5.6
SABRINA	2	5.7
JOHN	1	5.7
MARK	1	6
ERIC	1	6.4
BRUCE	1	5.9

Value = 1 represents Male and Value = 2 represents Female

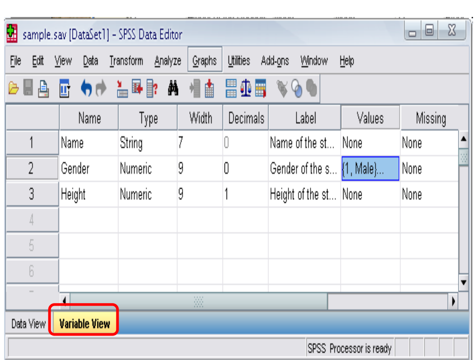
Practice 1 (Solution Sample)

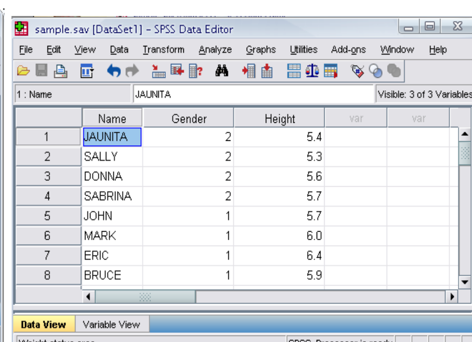


87

Variable View and Data View window

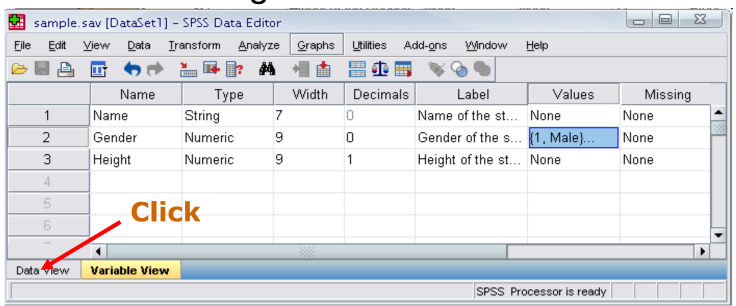
- The Data View window
  This sheet is visible when you first open the Data Editor and this sheet contains the data
- Click on the tab labeled Variable View

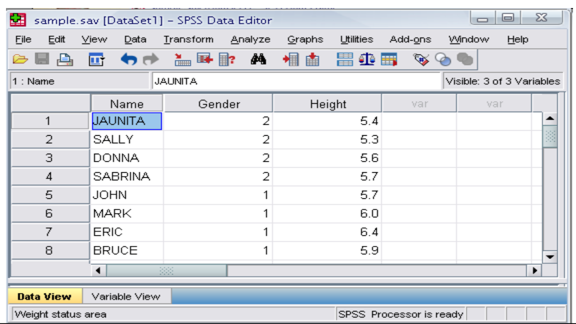




88

Sorting the data

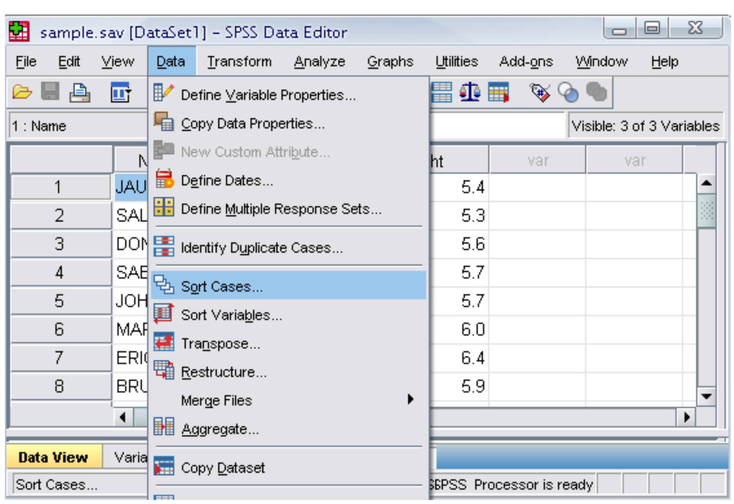




89

Sorting the data

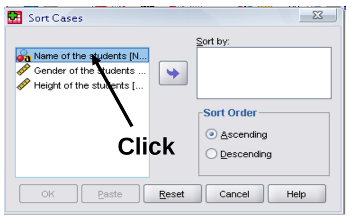
- Click **'Data'** and then click **Sort Cases**

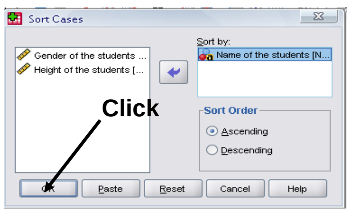


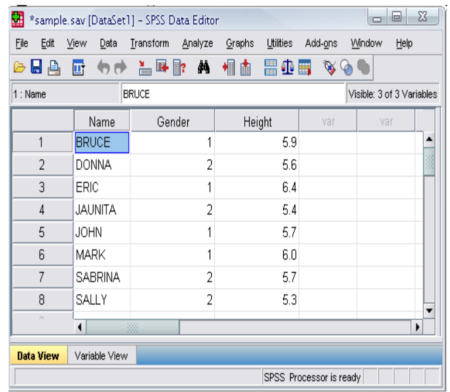
90

Sorting the data

- Double Click ‘Name of the students.’
- Then click ok.



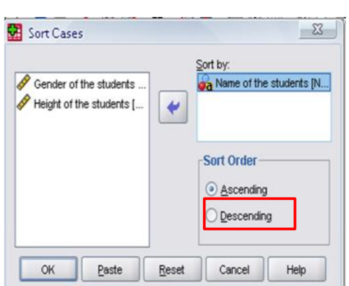


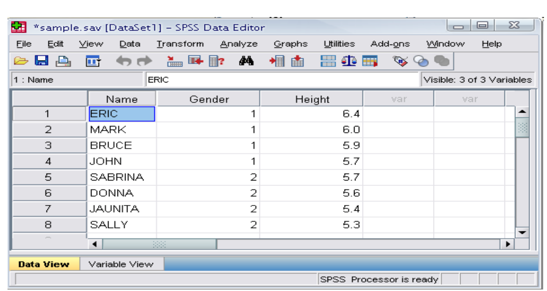


91

Practice 2

- How would you sort the data by the ‘Height’ of students in descending order?
  - Click data, sort cases, double click ‘height of students,’ **click ‘descending**,’ and finally click ok.





92

Recoding Variables in SPSS

Recoding

- You can use recoding to produce different values or codes for a variable. Recoding can be done in one of two ways:
  - Recoding into the same variable
  - Recoding into a different variable

94

Recoding

- There are 3 main types of recoding:
  - Recode single values
  - Recode a given range of values
  - Recode data into two categories

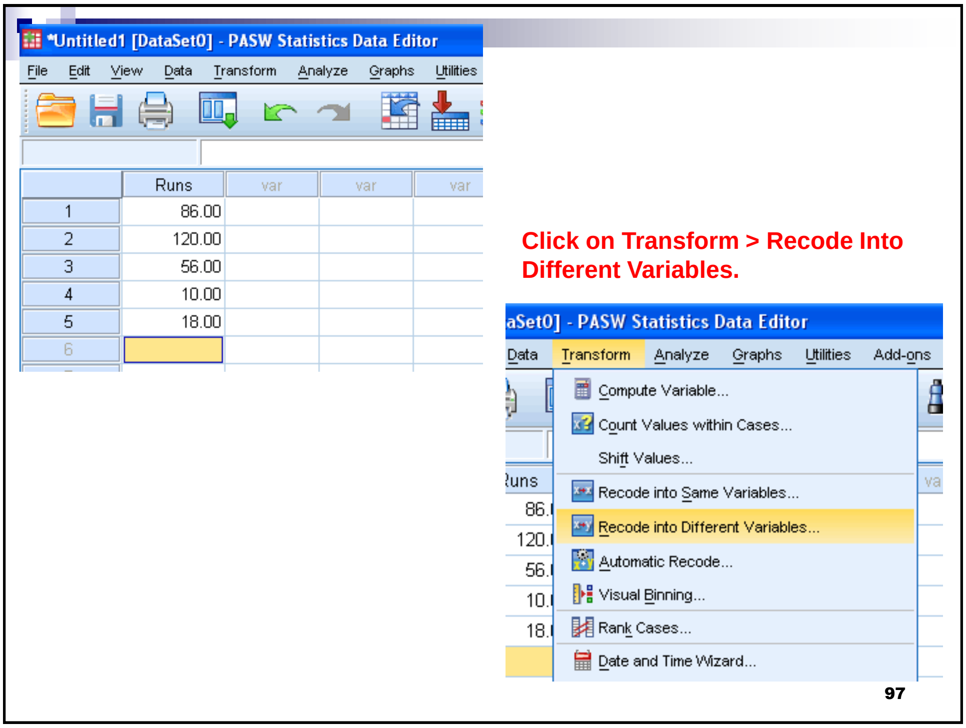
95

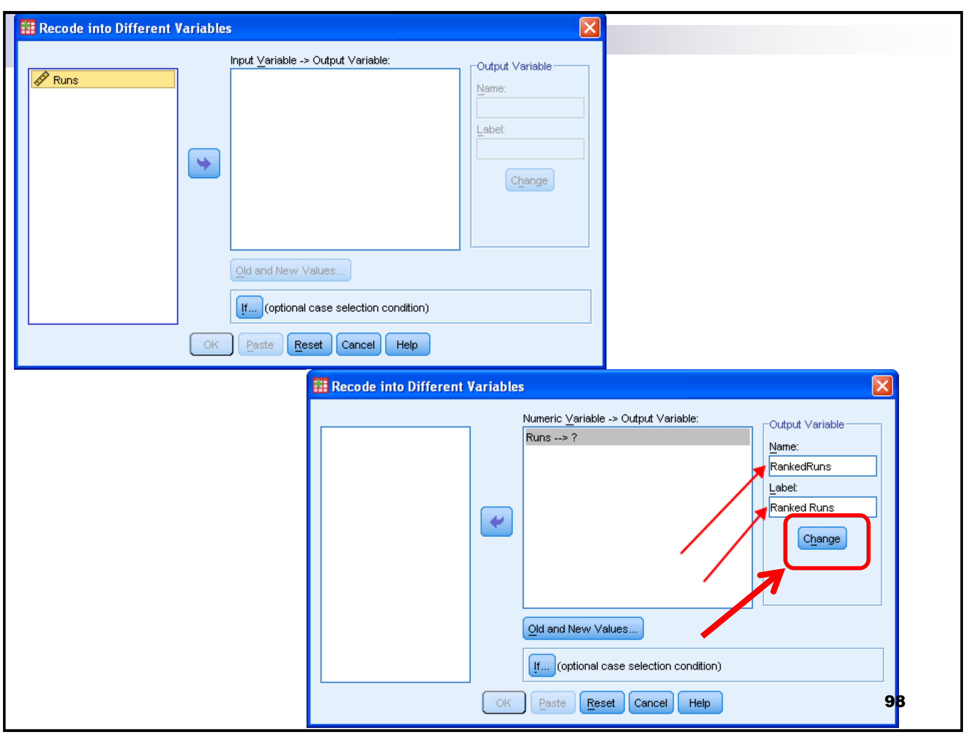
Recode single values in SPSS

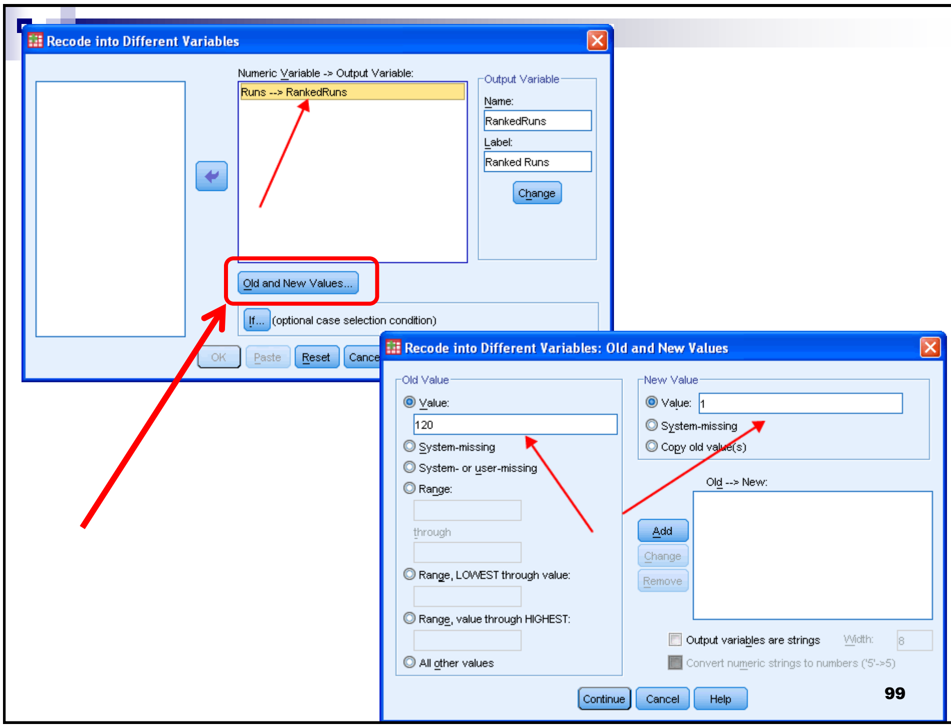
Example: The data given below represents cricket run scored by 5 Batsmen in a match. We can recode the batsman with the highest runs given a code of "1" and the batsman with the lowest runs given a "5".

	Number of runs by batsmen				
Batsmen	1	2	3	4	5
Runs	86	120	56	10	18

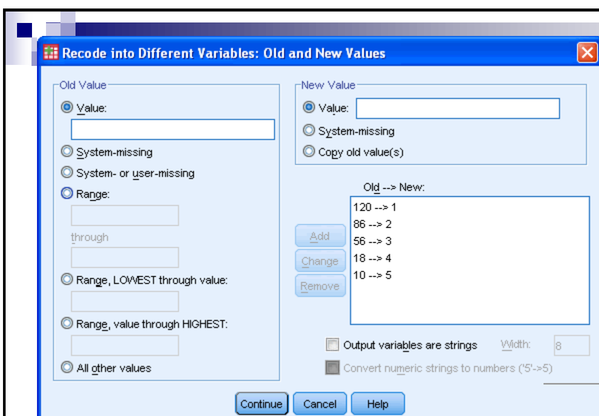
96


*Untitled1 [DataSet0] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities
Runs var var var
1 86.00
2 120.00
3 56.00
4 10.00
5 18.00
Click on Transform > Recode Into Different Variables.
aSet0] - PASW Statistics Data Editor
Data Transform Analyze Graphs Utilities Add-ons
Compute Variable...
Count Values within Cases...
Shift Values...
Recode into Same Variables...
Recode into Different Variables...
Automatic Recode...
Visual Binning...
Rank Cases...
Date and Time Wizard...
97


Recode into Different Variables
Runs
Input Variable -> Output Variable:
Output Variable
Name:
Label:
Change
Old and New Values...
If... (optional case selection condition)
OK Paste Reset Cancel Help
Recode into Different Variables
Numeric Variable -> Output Variable:
Runs --> ?
Output Variable
Name:
RankedRuns
Label:
Ranked Runs
Change
Old and New Values...
If... (optional case selection condition)
OK Paste Reset Cancel Help
98



99



Runs	RankedRuns
86.00	2.00
120.00	1.00
56.00	3.00
10.00	5.00
18.00	4.00

100

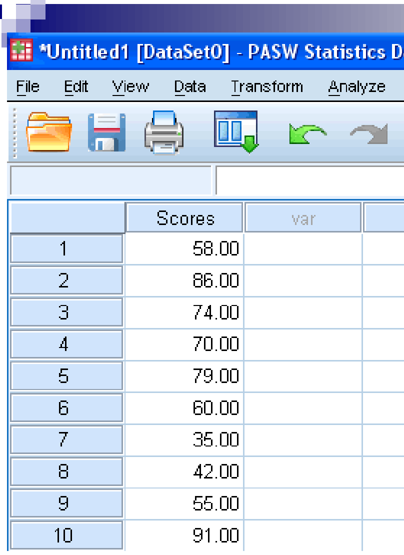
Recode a given range

Example: You are provided with marks of 10 students in a final business statistics examination.

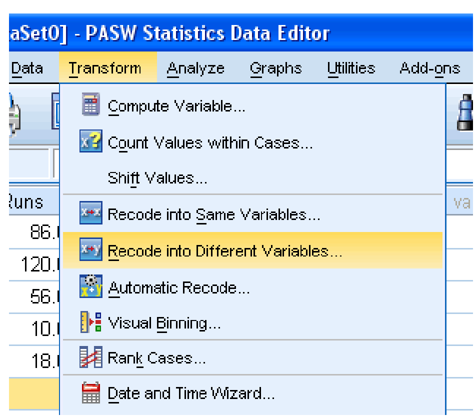
You can recode the data by giving code "1" to scores between 75 - 100, code 2 to scores between 61 - 74, code 3 to scores between 41 - 60 and code 4 to scores between 0 - 40.

Final examination scores of 10 students										
Scores	58	86	74	70	79	60	35	42	55	91

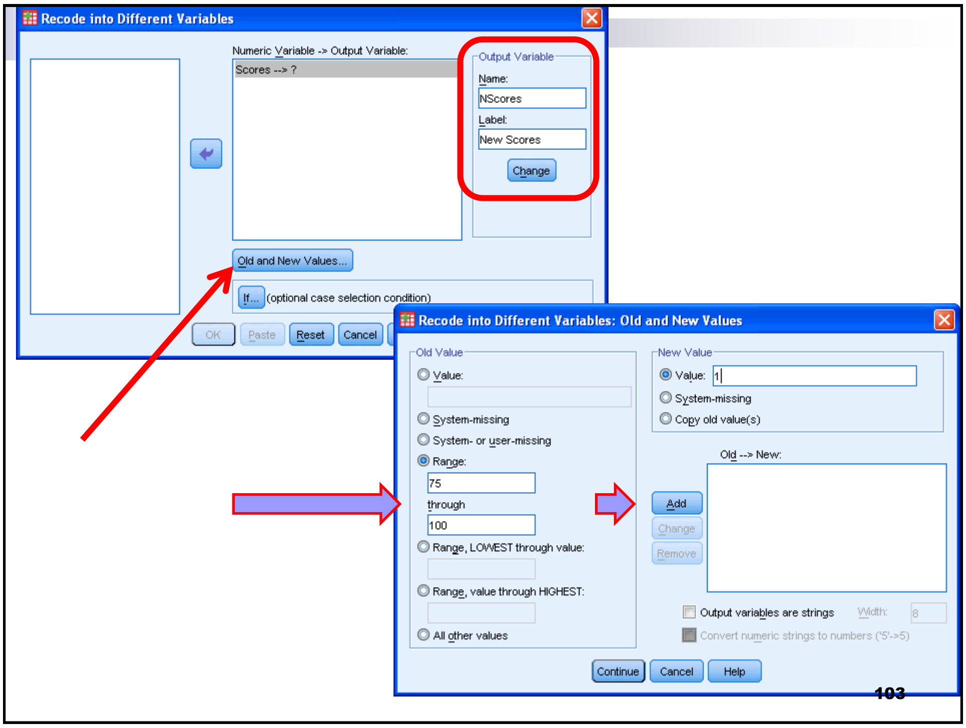
101

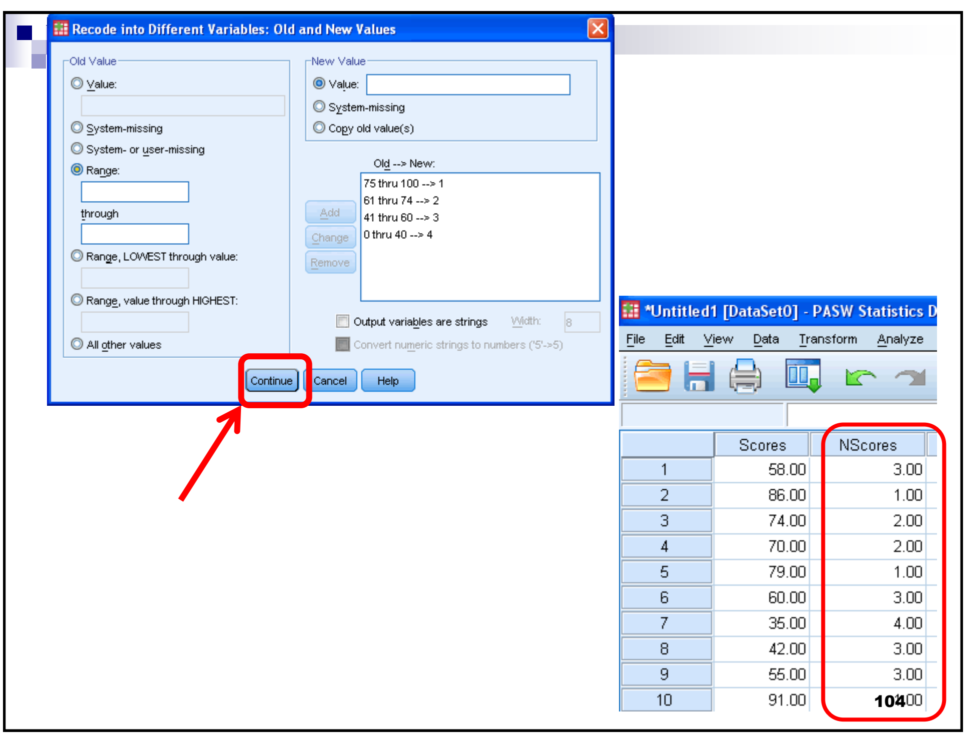


- Click on Transform > Recode Into Different Variables



102


Recode into Different Variables
Numeric Variable -> Output Variable:
Scores --> ?
Output Variable
Name:
NScores
Label:
New Scores
Change
Old and New Values...
If... (optional case selection condition)
OK
Paste
Reset
Cancel
Recode into Different Variables: Old and New Values
Old Value
Value:
System-missing
System- or user-missing
Range:
75
through
100
Range, LOWEST through value:
Range, value through HIGHEST:
All other values
New Value
Value: 1
System-missing
Copy old value(s)
Old --> New:
Add
Change
Remove
Output variables are strings
Width: 8
Convert numeric strings to numbers ('5'->5)
Continue
Cancel
Help
103


Recode into Different Variables: Old and New Values
Old Value
Value:
System-missing
System- or user-missing
Range:
through
Range, LOWEST through value:
Range, value through HIGHEST:
All other values
New Value
Value:
System-missing
Copy old value(s)
Old --> New:
75 thru 100 --> 1
61 thru 74 --> 2
41 thru 60 --> 3
0 thru 40 --> 4
Add
Change
Remove
Output variables are strings
Width: 8
Convert numeric strings to numbers ('5'->5)
Continue
Cancel
Help
*Untitled1 [DataSet0] - PASW Statistics D
File Edit View Data Transform Analyze
Scores
NScores
1 58.00 3.00
2 86.00 1.00
3 74.00 2.00
4 70.00 2.00
5 79.00 1.00
6 60.00 3.00
7 35.00 4.00
8 42.00 3.00
9 55.00 3.00
10 91.00
104

Recoding data into categories

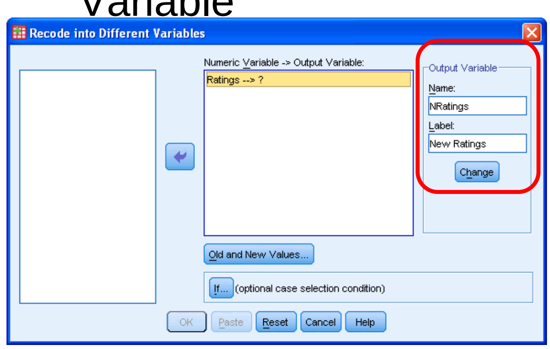
Example: The data represents a customer satisfaction **rating out of 10** for a new service offered by a company.

The company would like to code all those who responded by giving ratings **above 5 a "Satisfactory"** code and those **below 5 a "Dissatisfactory" code.**

Satisfaction scores for a new service												
Scores	3	6	8	9	7	2	10	6	4	8	9	3

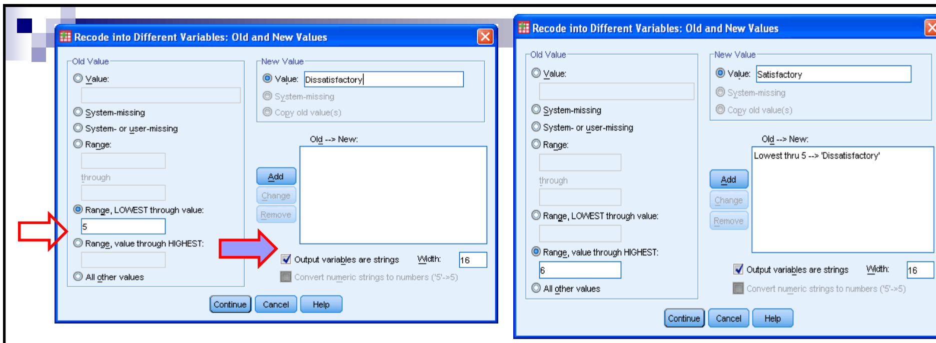
105

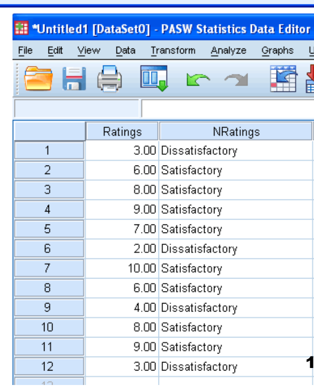
- Click on Transform > Recode Into Different Variable



Name the output variable of "NRatings" and label of "New Ratings"

106





	Ratings	NRatings
1	3.00	Dissatisfactory
2	6.00	Satisfactory
3	8.00	Satisfactory
4	9.00	Satisfactory
5	7.00	Satisfactory
6	2.00	Dissatisfactory
7	10.00	Satisfactory
8	6.00	Satisfactory
9	4.00	Dissatisfactory
10	8.00	Satisfactory
11	9.00	Satisfactory
12	3.00	Dissatisfactory

Enter the value of "5" into the Range, LOWEST through value: in the -Old Value- area, and set the new code to Dissatisfactory into the Value in the -New Value- area.

Click the Output variables are strings

107

Computing New Variables using SPSS

Computing a New Variable

- You may need to compute a new variable based on existing information (from other variables) in your data.
- Example:
  - You may want to convert the units of a variable from feet to meters, or use a subject's height and weight to compute their BMI.

Example:

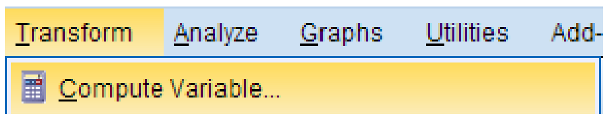
- How to calculate Gonado somatic index (GSI) of fish from a given data of [Ovary weight/body weight*100]

Compute GSI

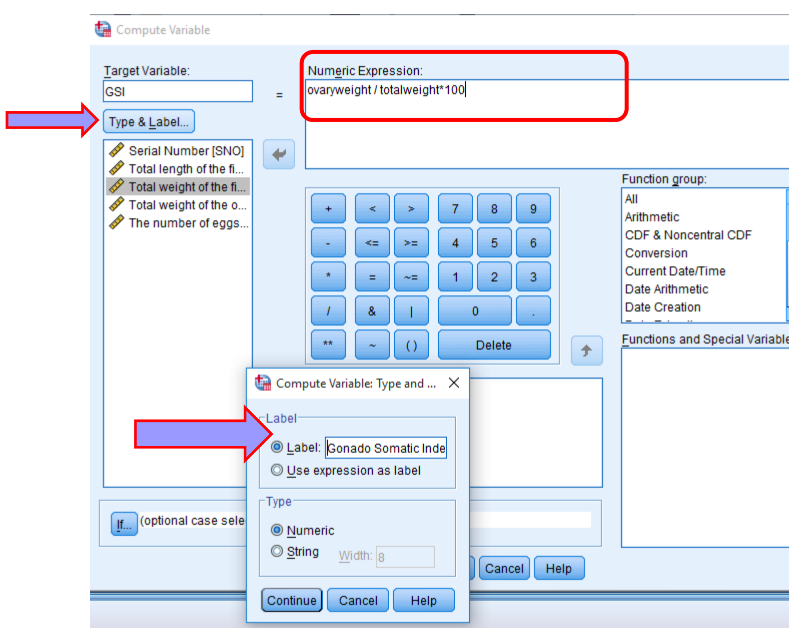
SNO	totallength	totalweight	ovaryweight	fecundity
1	13.20	38.46	5.99	31148
2	12.50	35.71	4.77	24327
3	13.20	40.90	5.41	33001
4	12.40	36.84	5.74	35014
5	13.00	40.73	3.11	19904
6	12.80	38.90	4.35	23925
7	12.70	39.50	3.93	27117
8	13.60	49.56	4.54	25424
9	13.30	50.80	3.52	21120
10	12.90	37.37	4.83	29946
11	13.10	38.89	3.67	17983
12	12.90	34.03	4.56	22800
13	13.00	35.27	3.07	14429
14	13.70	41.55	3.97	19850
15	12.80	36.38	1.27	6477
16	13.00	36.10	2.54	13462
17	12.70	33.22	1.26	6426
18	13.50	46.67	5.47	26803
19	13.00	35.65	1.36	7208
20	12.70	35.54	.60	3120
21	18.30	121.91	5.02	30120
22	18.20	111.93	3.29	19082
23	19.20	137.19	1.76	9504

Computing a New Variable (GSI)

- To compute a new variable, click Transform > Compute Variable.



111



112

SNO	totallength	totalweight	ovaryweight	fecundity	GSI
1	13.20	38.46	5.99	31148	15.57
2	12.50	35.71	4.77	24327	13.36
3	13.20	40.90	5.41	33001	13.23
4	12.40	36.84	5.74	35014	15.58
5	13.00	40.73	3.11	19904	7.64
6	12.80	38.90	4.35	23925	11.18
7	12.70	39.50	3.93	27117	9.95
8	13.60	49.56	4.54	25424	9.16
9	13.30	50.80	3.52	21120	6.93
10	12.90	37.37	4.83	29946	12.92
11	13.10	38.89	3.67	17983	9.44
12	12.90	34.03	4.56	22800	13.40
13	13.00	35.27	3.07	14429	8.70
14	13.70	41.55	3.97	19850	9.55
15	12.80	36.38	1.27	6477	3.49
16	13.00	36.10	2.54	13462	7.04
17	12.70	33.22	1.26	6426	3.79
18	13.50	46.67	5.47	26803	11.72
19	13.00	35.65	1.36	7208	3.81
20	12.70	35.54	.60	3120	1.69
21	18.30	121.91	5.02	30120	4.12
22	18.20	111.93	3.29	19082	2.94
23	19.20	137.19	1.76	9504	1.28

113

Z score compute using SPSS

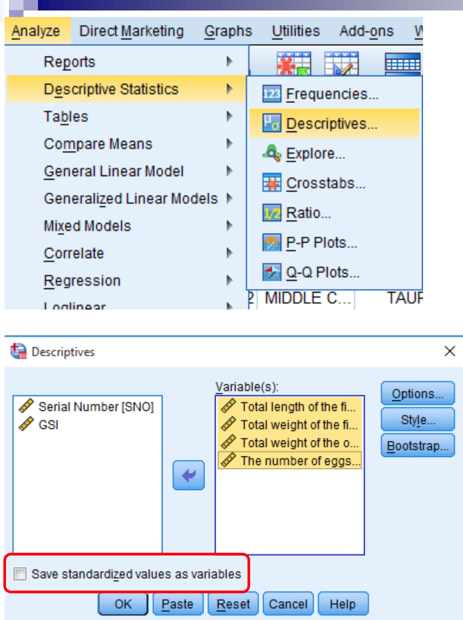
Z score

Z-Scores tell us whether a particular score is equal to the mean, below the mean or above the mean of a bunch of scores. They can also tell us how far a particular score is away from the mean. Is a particular score close to the mean or far away?

If a Z-Score….

- **Has a value of 0, it is equal to the group mean.**
- **Is positive, it is above the group mean.**
- **Is negative, it is below the group mean.**
- **Is equal to +1, it is 1 Standard Deviation above the mean.**
- **Is equal to +2, it is 2 Standard Deviations above the mean.**
- **Is equal to -1, it is 1 Standard Deviation below the mean.**
- **Is equal to -2, it is 2 Standard Deviations below the mean.**

115



In the Data View of the Data Editor window, the z-scores will be added as a new variable (Ztotallength and Ztotalweight) with each individual case having a z-score.

Ztotallength	Ztotalweight
-1.40667	-1.46442
-1.74226	-1.56476
-1.40667	-1.37540
-1.79021	-1.52353
-1.50256	-1.38160
-1.59844	-1.44837
-1.64638	-1.42648
-1.21491	-1.05943
-1.35873	-1.01418
-1.55050	-1.50419
-1.45462	-1.44873
-1.55050	-1.62606
-1.50256	-1.58081
-1.16697	-1.35168
-1.59844	-1.54031
-1.50256	-1.55053
-1.64638	-1.65561
-1.26285	-1.16487
-1.50256	-1.56695
-1.64638	-1.57096
1.03835	1.58035
.99040	1.21622
1.46982	2.13786

116

Descriptive Statistics using SPSS

Descriptive analysis

- Descriptive statistics are commonly used for summarizing data frequency or measures of central tendency (mean, median and mode).
- If your data is categorical, use the frequencies or **crosstabs procedures**.
- If your data is scale level, **use summaries or descriptives.**

118

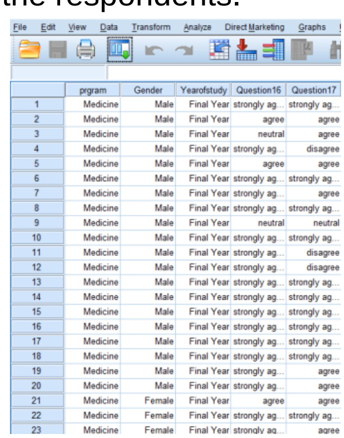
Frequency Analysis using SPSS

Frequency analysis

Frequency analysis is a descriptive statistical method that shows the number of occurrences of each and each category of variables and response chosen by the respondents.

Consider the following variables

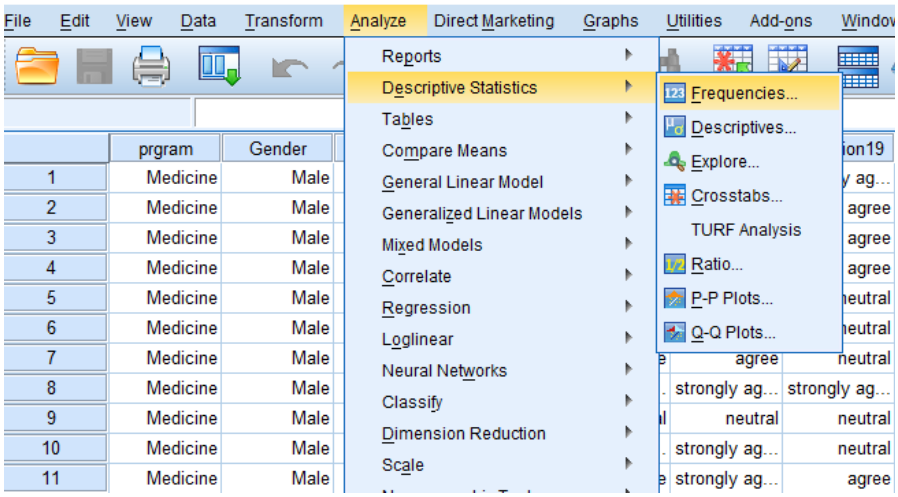
- Program
- Gender
- Year of study



	prgram	Gender	Yearofstudy	Question16	Question17
1	Medicine	Male	Final Year	strongly ag...	strongly ag...
2	Medicine	Male	Final Year	agree	agree
3	Medicine	Male	Final Year	neutral	agree
4	Medicine	Male	Final Year	strongly ag...	disagree
5	Medicine	Male	Final Year	agree	agree
6	Medicine	Male	Final Year	strongly ag...	strongly ag...
7	Medicine	Male	Final Year	strongly ag...	agree
8	Medicine	Male	Final Year	strongly ag...	strongly ag...
9	Medicine	Male	Final Year	neutral	neutral
10	Medicine	Male	Final Year	strongly ag...	strongly ag...
11	Medicine	Male	Final Year	strongly ag...	disagree
12	Medicine	Male	Final Year	strongly ag...	disagree
13	Medicine	Male	Final Year	strongly ag...	strongly ag...
14	Medicine	Male	Final Year	strongly ag...	strongly ag...
15	Medicine	Male	Final Year	strongly ag...	strongly ag...
16	Medicine	Male	Final Year	strongly ag...	strongly ag...
17	Medicine	Male	Final Year	strongly ag...	strongly ag...
18	Medicine	Male	Final Year	strongly ag...	strongly ag...
19	Medicine	Male	Final Year	strongly ag...	agree
20	Medicine	Male	Final Year	strongly ag...	agree
21	Medicine	Female	Final Year	agree	agree
22	Medicine	Female	Final Year	strongly ag...	strongly ag...
23	Medicine	Female	Final Year	strongly ag...	agree

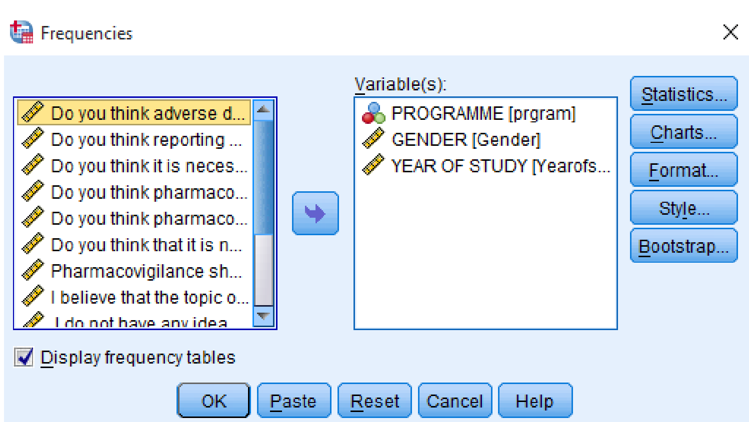
120

From the menus choose:
Analyze > Descriptive Statistics > Frequencies...



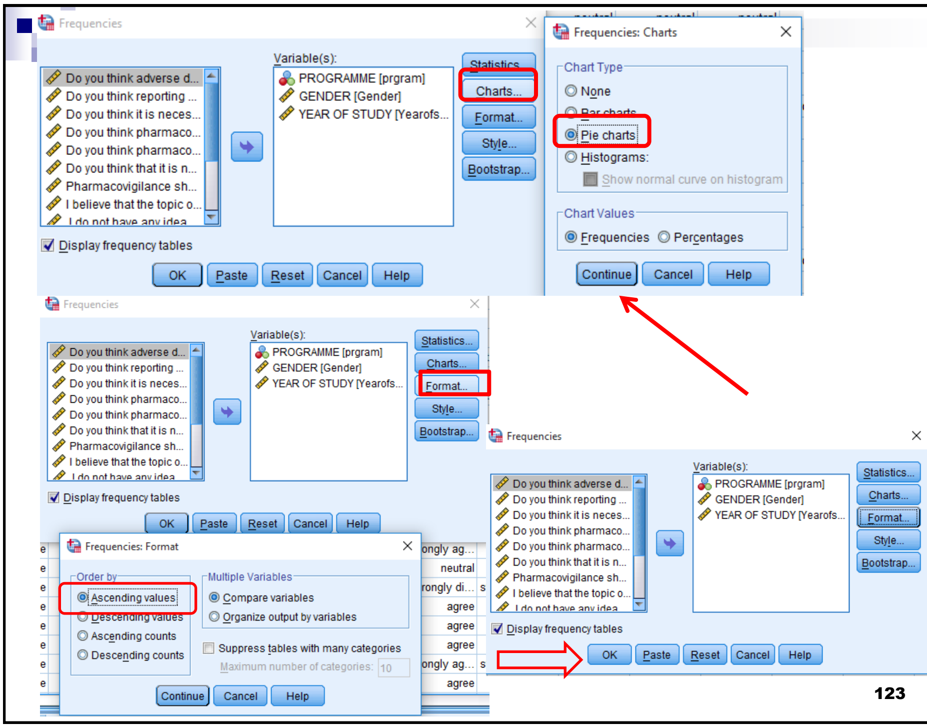
121

Select and place the demographic variables



- **Click Program, Gender, year of study** and drag the variable into the target Variable(s) list.

122



123

Frequencies

Statistics

		PROGRAMME	GENDER	YEAR OF STUDY
N	Valid	364	364	364
	Missing	0	0	0

YEAR OF STUDY

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Final Year	144	39.6	39.6	39.6
	Pre-final Year	220	60.4	60.4	100.0
	Total	364	100.0	100.0	

Double-click to activate

Frequency Table

PROGRAMME

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Medicine	134	36.8	36.8	36.8
	Dentistry	61	16.8	16.8	53.6
	Pharmacy	137	37.6	37.6	91.2
	Nursing	32	8.8	8.8	100.0
	Total	364	100.0	100.0	

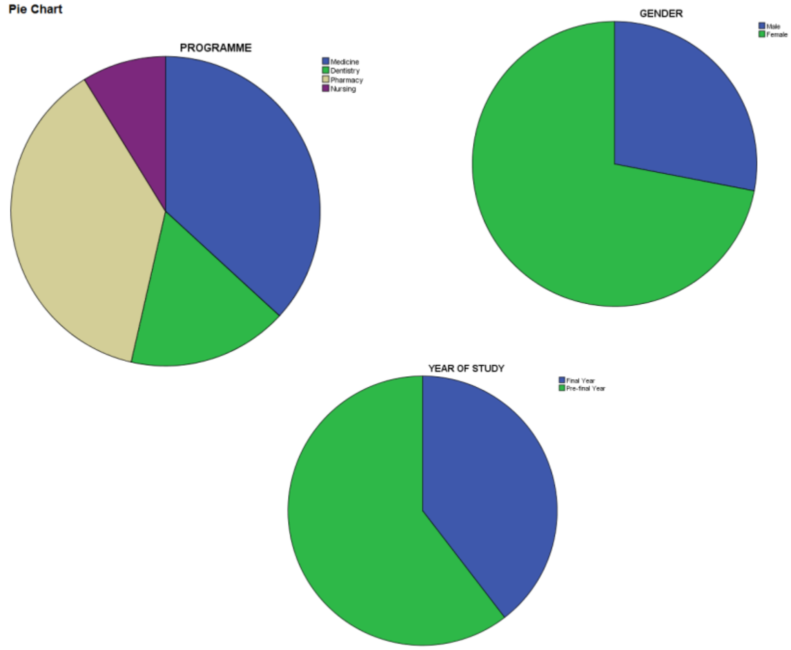
GENDER

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	102	28.0	28.0	28.0
	Female	262	72.0	72.0	100.0
	Total	364	100.0	100.0	

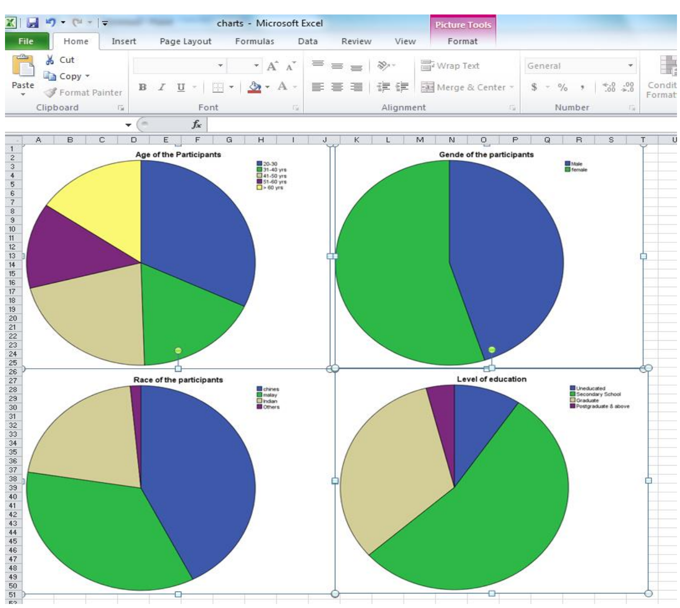
- No missing values in the data set
- Programme, Gender, year of study distribution (frequency and percent)

124

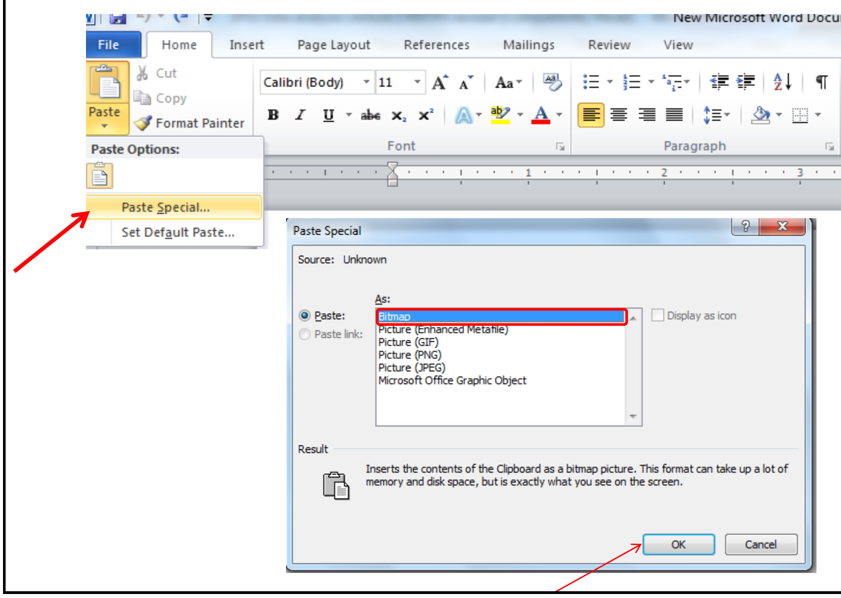
Pie Chart



Copy and paste all graphs to Excel file as follows in two rows



Copy all the chart from excel and paste special and then choose – bitmap in word as follows



127

Give Figure title and Figure number

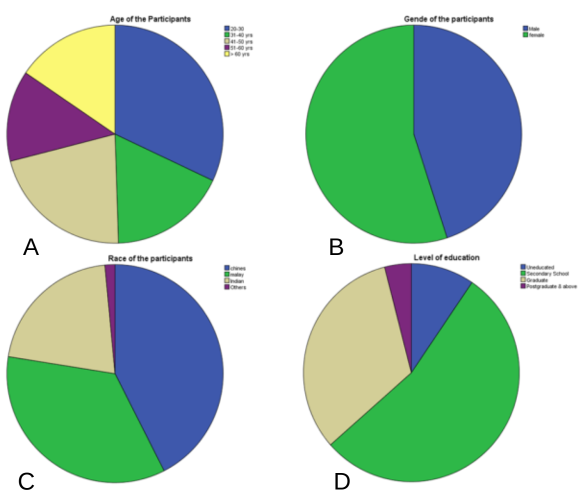
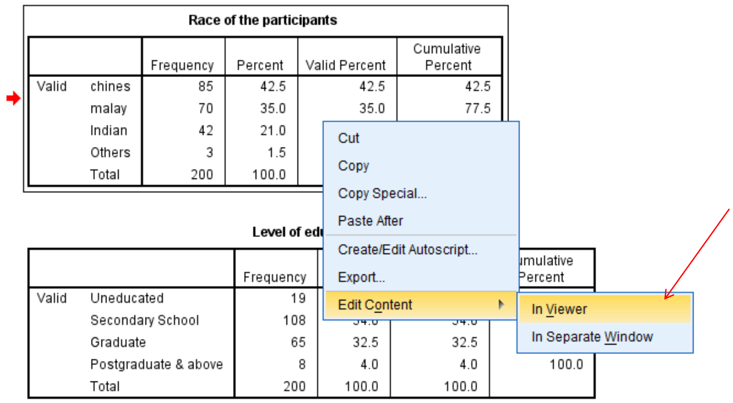


Fig 1. Demographic details of the participants

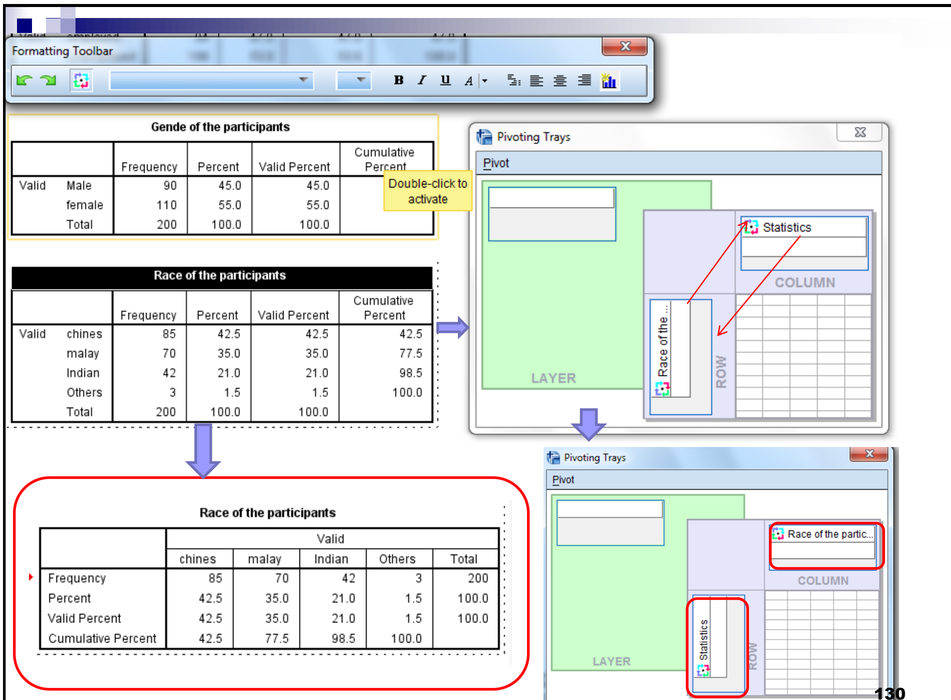
128

If you want to rearrange the column and row items, Try Pivot Table.. Select the table – Right click – Edit Content – In Viewer.



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	chines	85	42.5	42.5	42.5
	malay	70	35.0	35.0	77.5
	Indian	42	21.0		
	Others	3	1.5		
	Total	200	100.0		

129



Gende of the participants

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	90	45.0	45.0	
	female	110	55.0	55.0	
	Total	200	100.0	100.0	

Race of the participants

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	chines	85	42.5	42.5	42.5
	malay	70	35.0	35.0	77.5
	Indian	42	21.0	21.0	98.5
	Others	3	1.5	1.5	100.0
	Total	200	100.0	100.0	

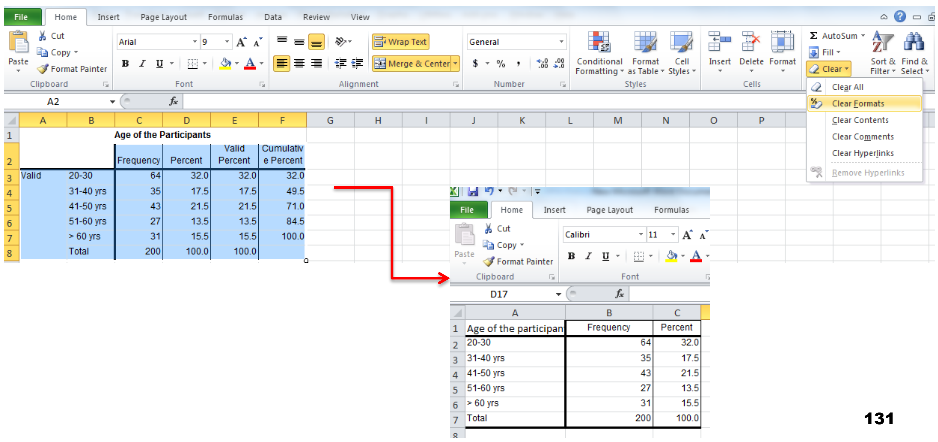
Race of the participants

	Valid				
	chines	malay	Indian	Others	Total
Frequency	85	70	42	3	200
Percent	42.5	35.0	21.0	1.5	100.0
Valid Percent	42.5	35.0	21.0	1.5	100.0
Cumulative Percent	42.5	77.5	98.5	100.0	

130

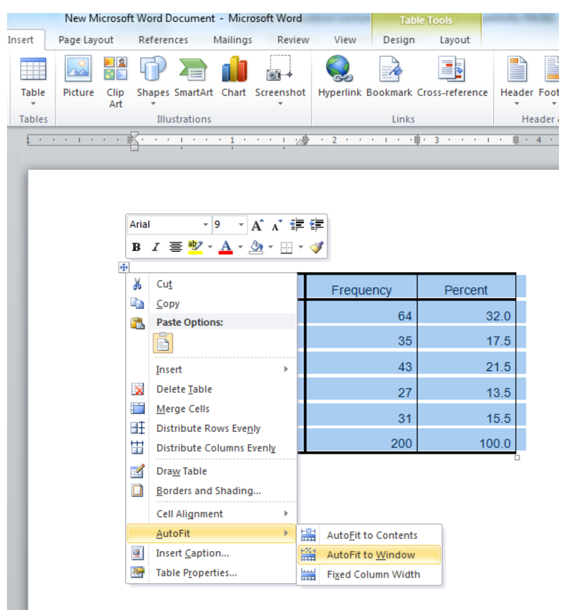
Edit and formatting table

- Copy table from SPSS Output
- Paste in excel file
- Clear formatting as follows
- Delete unwanted columns

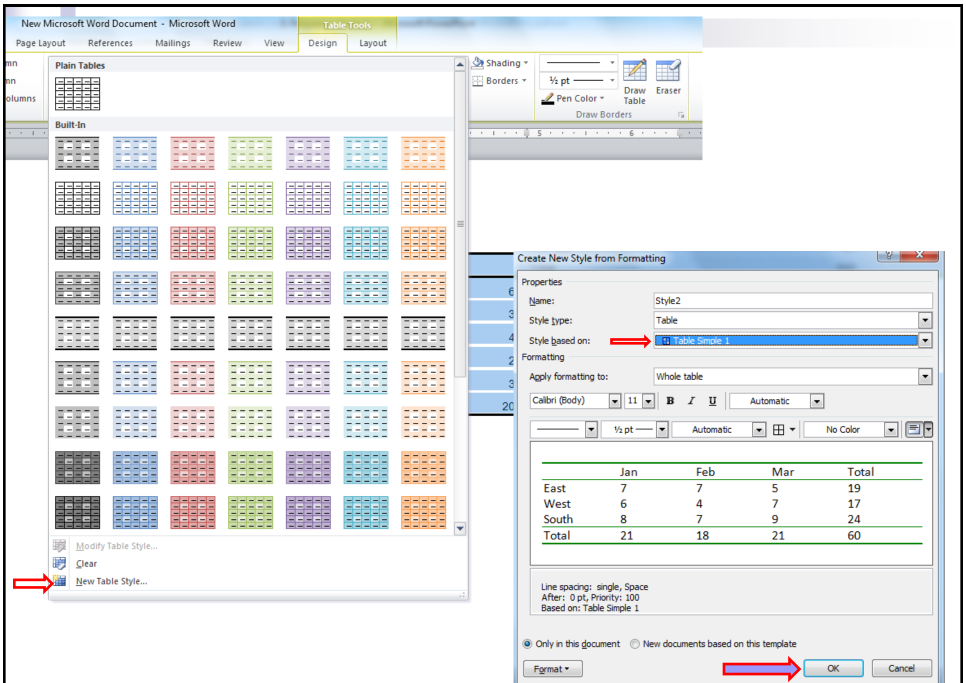


131

Select the table and auto format in word



132



133

APA (*American Psychological Association)* style formatting table

Age of the participants	Frequency	Percent
20-30	64	32.0
31-40 yrs	35	17.5
41-50 yrs	43	21.5
51-60 yrs	27	13.5
> 60 yrs	31	15.5
Total	200	100.0

134

Creating a Bar Chart using SPSS

Bar chart

- A **bar chart** is helpful in graphically describing (visualizing) your data. It will often be used in addition to inferential statistics.
- For example, a bar chart can be appropriate if you are analysing your data using an independent-samples t-test, paired-samples t-test (dependent t-test), one-way ANOVA or repeated measures ANOVA

136

Example

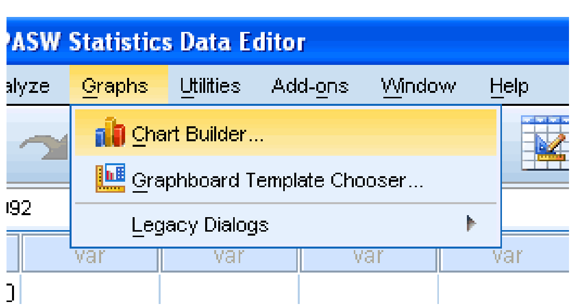
- The concentration of cholesterol (a type of fat) in the blood is associated with the risk of developing heart disease, such that higher concentrations of cholesterol indicate a higher level of risk and lower concentrations indicate a lower level of risk. If you lower the concentration of cholesterol in the blood, your risk for developing heart disease can be reduced. Being overweight and/or physically inactive increases the concentration of cholesterol in your blood**. Both exercise and weight loss can reduce cholesterol concentration**. However, it is not known whether exercise or weight loss is best for lowering blood cholesterol concentration.

A random sample of inactive male individuals that were classified as overweight were recruited to investigate whether an exercise or weight loss intervention is more effective in lowering cholesterol levels. This sample was split into two groups: one group underwent an "exercise training programme" **(labelled "exercise" in the bar chart),** and the other group undertook a "calorie-controlled diet" (lablled "diet" in the bar chart). In order to determine which treatment programme was more effective, the mean cholesterol concentrations were compared between the two groups at the end of the treatment programmes. **The dependent variable was Cholesterol Concentration, and the independent variable, Treatment, which consisted of these two groups: "exercise" and "diet"**

137

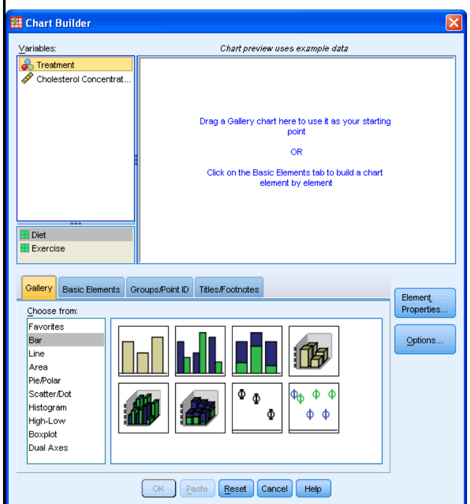
Creating a Bar Chart using SPSS

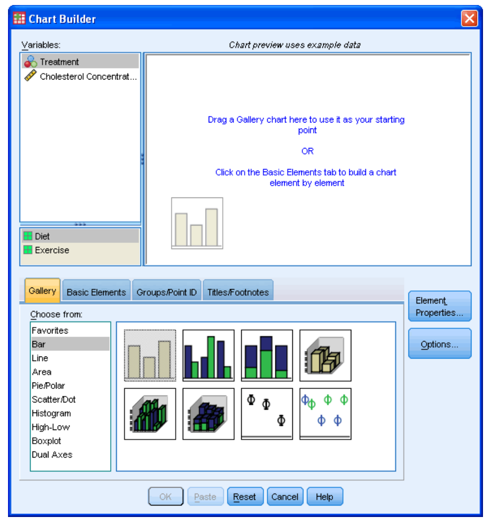
- Click **Graphs > Chart Builder...** on the top menu as shown below:



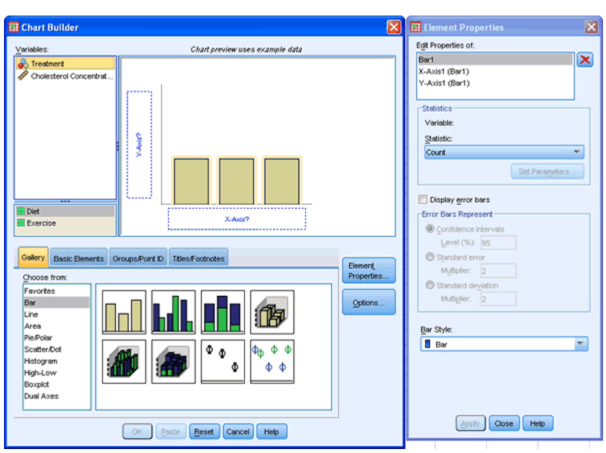
138

Under the Gallery Tab (Gallery), select the **Bar** option and the **simple bar chart** icon (top-left icon).
Drag-and-drop this icon into the **Chart Preview Area**.

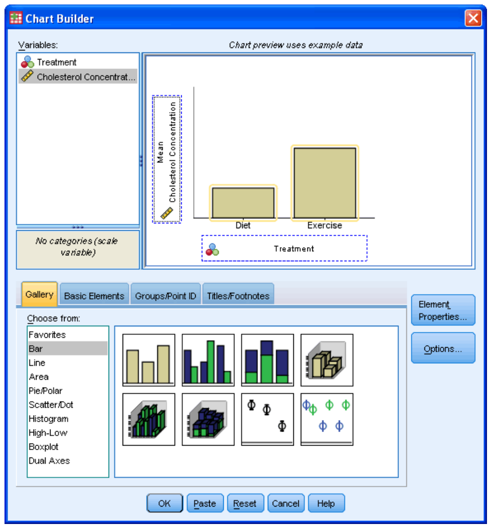




You will be presented with the following dialog boxes: **Chart Builder** and **Element Properties**. You can see, the **Chart Preview Area** with simple bar chart.

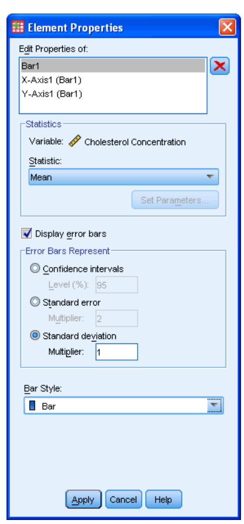


Transfer the independent variable, Treatment, into the "**X-Axis?**" box and the dependent (outcome) variable, Cholesterol Concentration, into the "**Y-Axis?**" box within the **Preview Chart Area** by drag-and-dropping the variables from the Variables: box.



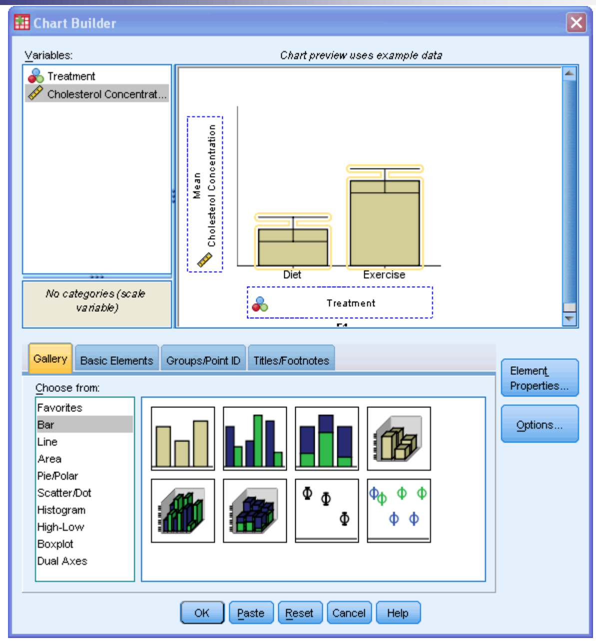
141

Ideally, we want to be able to show a measure of the spread of the data. In this case, we wish to have error bars that represent ± 1 standard deviations. To do this, we tick the Dispay error bars checkbox and then, under the -Error Bars Represent- area, we check the Standard deviation box, and in the Multiplier:, enter "1".



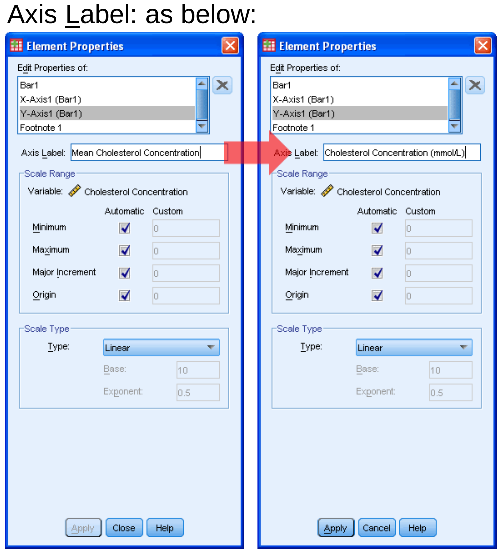
Click apply

142

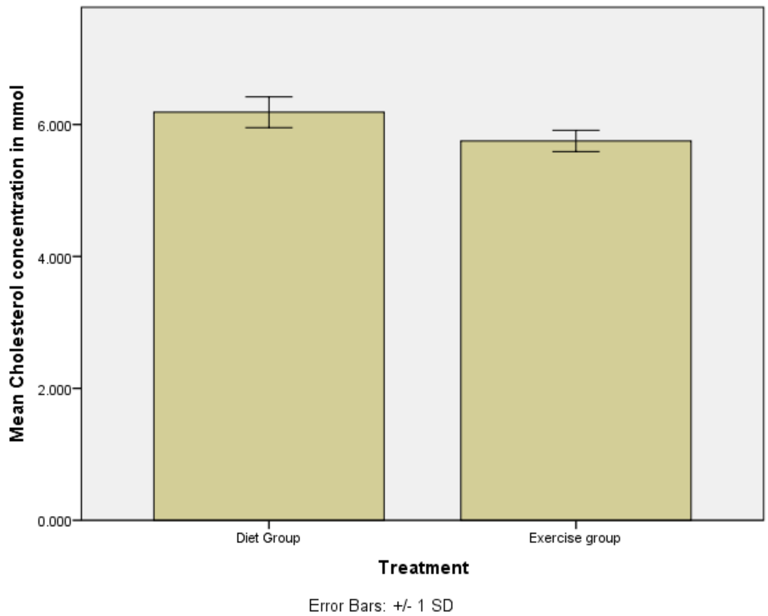


143

- We want to change the y-axis label so that we can remove the "mean" text and add in some units of measurement. We do this by selecting "Y-Axis (Bar1)" in the Edit Properties of: box and then change the Axis Label: as below:



144



145

Summary Measures for Scale Variables using SPSS

Descriptive statistics: Summary Measures

A study was conducted to determine the serum cholesterol (mmol/L) measured on a sample of 86 stroke patients and the results are given in the below table.

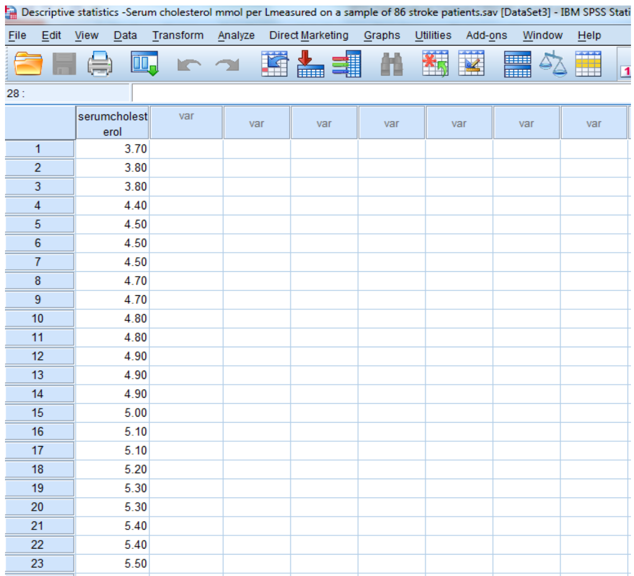
3.7	4.8	5.4	5.6	6.1	6.4	7.0	7.6	8.7
3.8	4.9	5.4	5.6	6.1	6.5	7.0	7.6	8.9
3.8	4.9	5.5	5.7	6.1	6.5	7.1	7.6	9.3
4.4	4.9	5.5	5.7	6.2	6.6	7.1	7.7	9.5
4.5	5.0	5.5	5.7	6.3	6.7	7.2	7.8	10.2
4.5	5.1	5.6	5.8	6.3	6.7	7.3	7.8	10.4
4.5	5.1	5.6	5.8	6.4	6.8	7.4	7.8	
4.7	5.2	5.6	5.9	6.4	6.8	7.4	8.2	
4.7	5.3	5.6	6.0	6.4	7.0	7.5	8.3	
4.8	5.3	5.6	6.1	6.4	7.0	7.5	8.6	

Descriptive Statistics

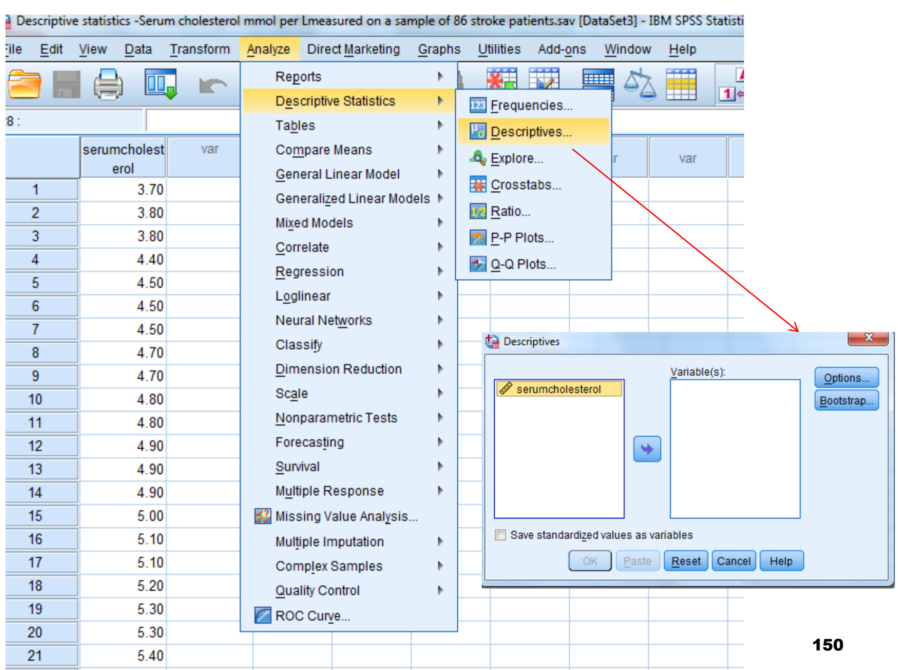
Descriptive measures describe the properties, distribution, dispersion and pattern of data

- Mean – Gives central value
- Median – Gives middle number
- Mode – Gives high frequency number
- Standard deviation with upper and lower limits to mean
- Skewness and Kurtosis – Give the nature of frequency curve
- Maximum and Minimum gives the range of values

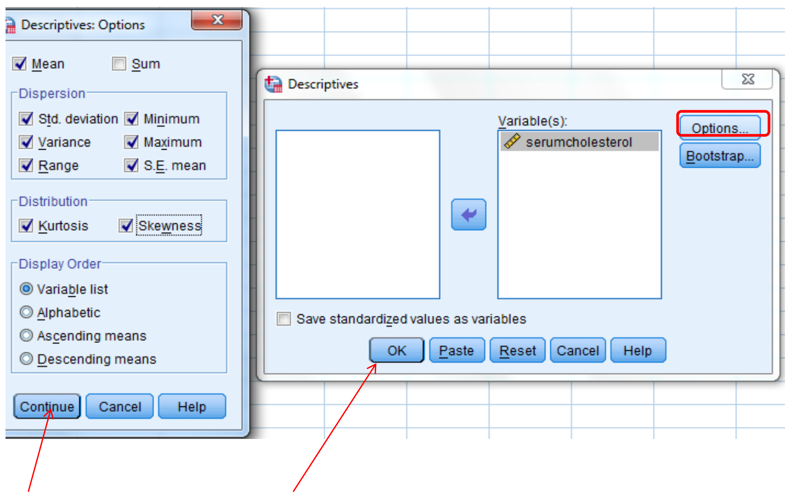
- **Enter the data**
- **Click ANALYZE – Descriptive statistics**



149



150



151

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
serumcholesterol	86	6.70	3.70	10.40	6.3407	1.39978	1.959
Valid N (listwise)	86						

152

Interpretation of descriptive measures

- Mean value is used for comparison
- Mean is not a good measure as it could be affected by extreme values
- Hence median (the middlemost number) in a series is considered instead of mean
- Mode is the very frequently occurring item, like frequent selling item

153

Interpretation of descriptive measures

- Standard deviation (SD) is the distance measure from mean. Normally denoted by σ.
- SD is useful in deciding confidence intervals
- In combination with mean the SD decides the probabilities in normal (Gaussian) distributions
- (1σ to -1σ) cover 68% of area in a normal curve
- (2σ to -2σ) cover 96% of area in a normal curve
- (3σ to -3σ) cover 99.9% of area in a normal curve

154

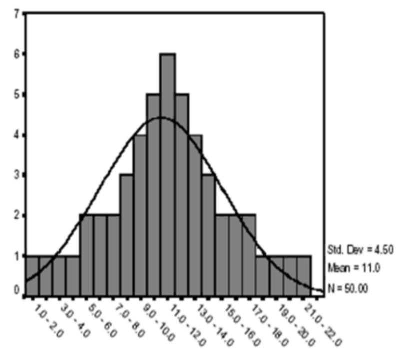
Skewness and Kurtosis results Interpretation

- Skewness value 0 is considered normal
- Skewness between -1 to +1 considered normal and requires no transformation of data
- Beyond -1 to +1 in either direction requires transformation of data to do any further analysis
- Transformation may be in the form of LN (natural log) or square or square root etc
- If the data do not become normal even after transformation, then NON-PARAMETRIC analysis is to be applied.
- Kurtosis between -1 to +1 considered normal and requires no transformation of data
- Beyond -1 to +1 in either direction requires transformation of data to do any further analysis

Testing for Normality using SPSS

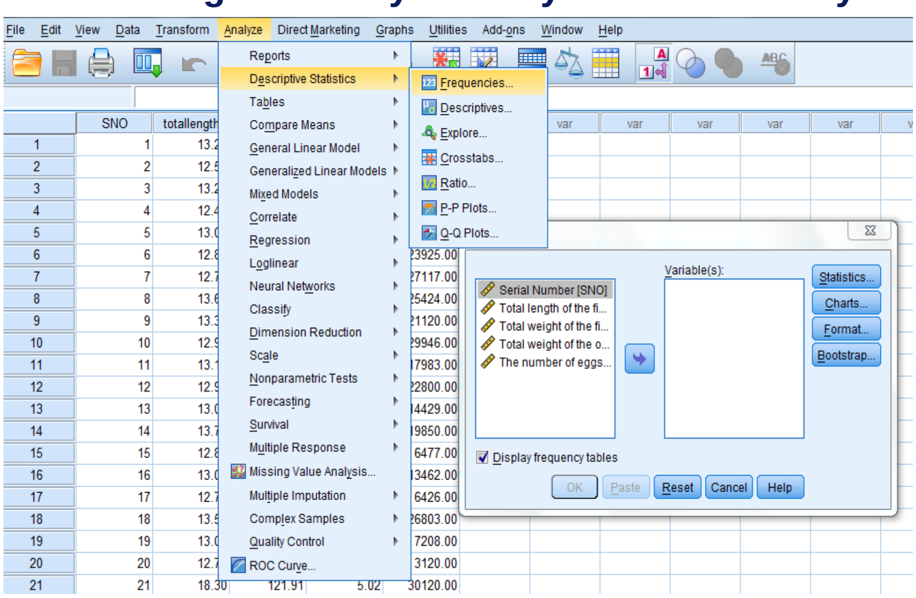
Normality

- The concept of normality is central to statistics. Normality refers to the **'shape'** of the distribution of data. Consider a histogram of values for one variable.
- By drawing a line across the 'tops' of the bars in the histogram, we are able to see the 'shape' of the data.
- When the **'shape' forms a 'bell' shape, we generally call this a normal curve**.
- The figure is approximately normally distributed.
- For data to be normal, they must have the form of a bell curve



157

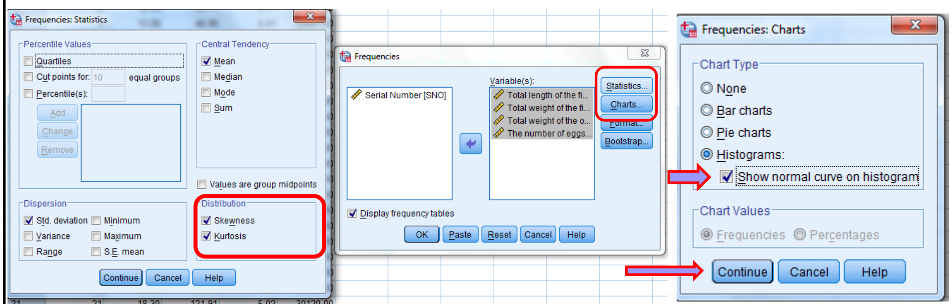
Assessing Normality Visually and Statistically



158

Statistical Approach

Visual Approach



159

Statistical Approach

Statistics

		Total length of the fish (cm)	Total weight of the fish (g)	Total weight of the ovaries (g)	The number of eggs produced each female
N	Valid	70	70	70	70
	Missing	0	0	0	0
Mean		16.1341	78.5963	7.9787	36804.6143
Std. Deviation		2.08587	27.40759	4.26769	19152.37256
Skewness		-.737	-.444	.379	.662
Std. Error of Skewness		.287	.287	.287	.287
Kurtosis		-1.091	-.914	-.660	.064
Std. Error of Kurtosis		.566	.566	.566	.566

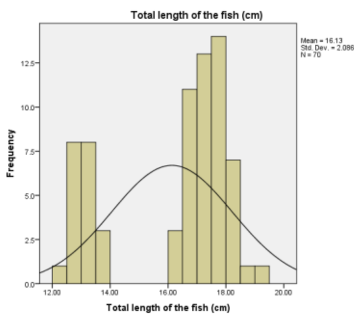
Skewness and kurtosis values between –1.0 and +1.0 is considered normal

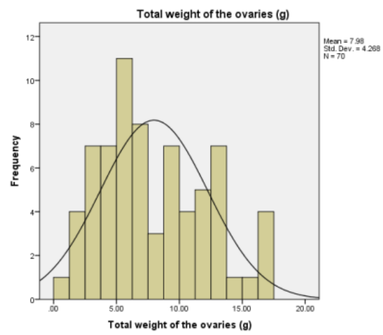
A positive skewness value indicates positive (right) skew; a negative value indicates negative (left) skew.

160

Visual Approach:

Frequency Distributions

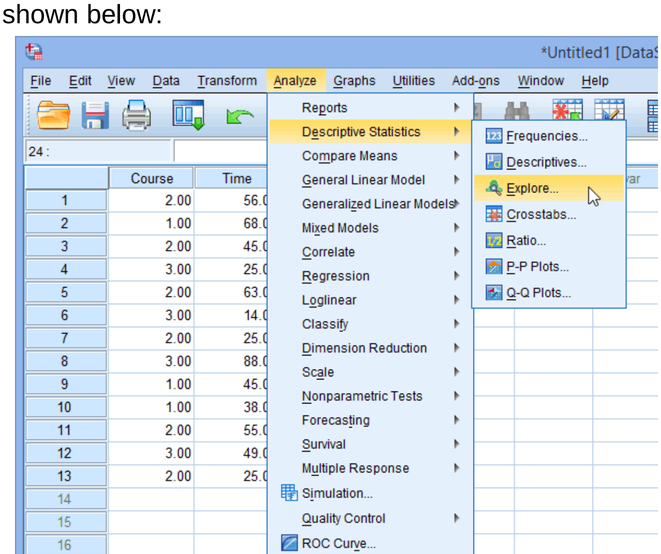




161

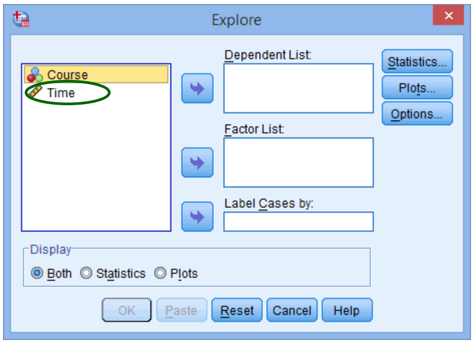
Normality Checking different levels of independent variable

Click Analyze > Descriptive Statistics > Explore... on the top menu, as shown below:



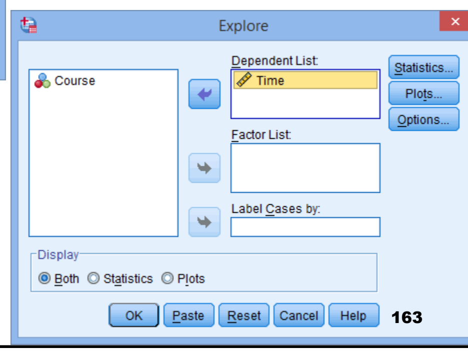
162

You will be presented with the Explore dialogue box, as shown below:



Transfer the variable that needs to be tested for normality into the Dependent List:

Transfer the Time variable into the Dependent List: box.

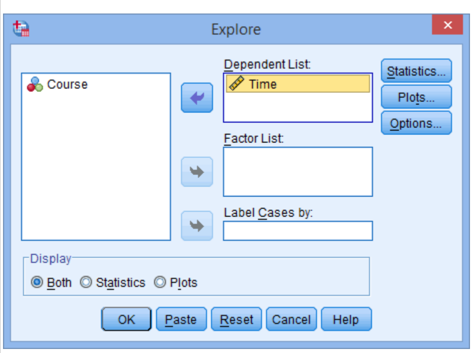


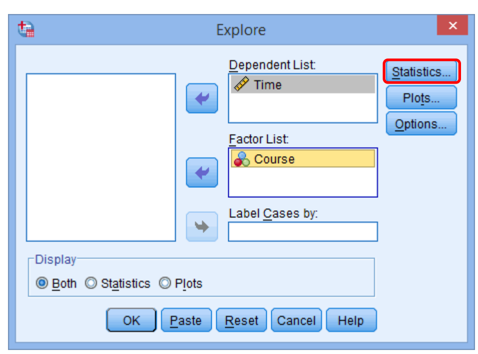
163

[Optional] If you need to establish if your variable is normally distributed for each level of your independent variable, you need to add your independent variable to the Factor List: box

In this example, we transfer the Course variable into the Factor List: box.

You will be presented with the following screen:

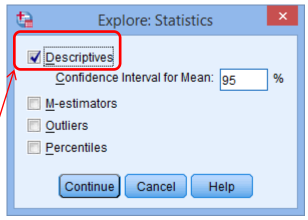




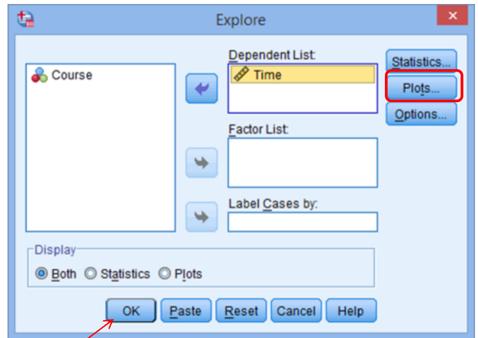
Click the Statistics Button.

164

Click the Statistics Button. You will be presented with the Explore: Statistics dialogue box, as shown below:

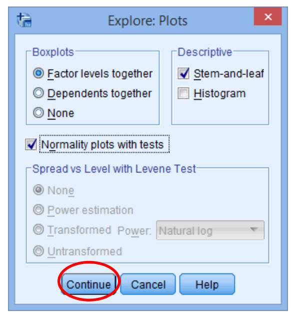


Check descriptives



Click the SPSS Plots Button.

Change the options so that you are presented with this screen:



165

Shapiro-Wilk Test of Normality

Tests of Normality

Course		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Time	Beginner	.177	10	.200*	.964	10	.827
	Intermediate	.166	10	.200*	.969	10	.882
	Advanced	.151	10	.200*	.965	10	.837

a. Lilliefors Significance Correction

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

The above table presents the results from two well-known tests of normality, namely the **Kolmogorov-Smirnov Test and the Shapiro-Wilk Test**.

We can see from the above table that for the **"Beginner", "Intermediate" and "Advanced" Course Group and the dependent variable**, "Time", was normally distributed.

If the Sig. value of the Shapiro-Wilk Test is greater than 0.05, the data is normal. If it is below 0.05, the data significantly deviate from a normal distribution.

166

Kolmogorov-Smirnov D test

- Kolmogorov-Smirnov D test is a test of normality for large samples.
- This test is similar to a chi-square test for goodness-of-fit, testing to see if the observed data fit a normal distribution.
- If the results are significant, then the null hypothesis of no difference between the observed data distribution and a normal distribution is rejected

Shapiro-Wilks W test

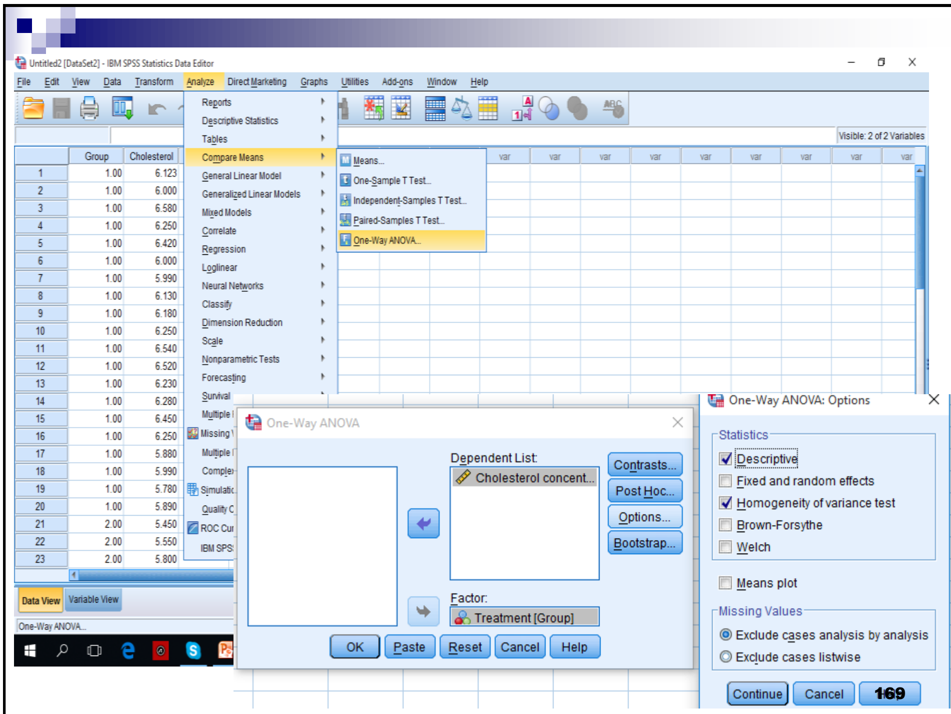
Shapiro-Wilks W test is considered by some authors to be the best test of normality (Zar 1999). Shapiro-Wilks W is limited to "small" data sets up to n = 2000.

167

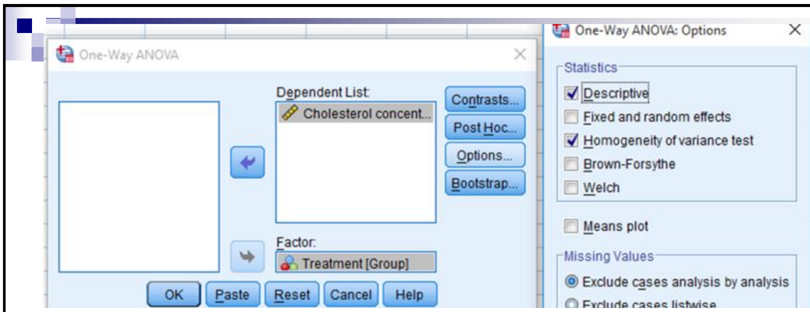
Homogeneity of Variance - **Levene's Test**

For parametric statistics to work optimally, the variance of the data must be the same throughout the data set.
This is known as homogeneity of variance, and the opposite condition is known as heteroscedasticity.

168



169



Descriptives

Cholesterol concentration in mmol

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Diet Group	20	6.18665	.233831	.052286	6.07721	6.29609	5.780	6.580
Exercise group	20	5.75175	.161402	.036091	5.67621	5.82729	5.450	6.000
Total	40	5.96920	.296354	.046858	5.87442	6.06398	5.450	6.580

Test of Homogeneity of Variances

Cholesterol concentration in mmol

Levene Statistic	df1	df2	Sig.
3.012	1	38	.091

Levene's Test "Sig." value is higher than 0.05), the two variances are similar and not rejecting the null hypothesis, and you can proceed to use a parametric test.

170

Student t - tests using SPSS

t test

- **One-sample t-tests:**
  - Used to compare **one sample mean** to a **population mean** or some other known value.
  - Average birth weight of new born baby
  - You hear that the average person sleeps 8 hours a day. You think college students sleep less. You ask 100 college students how long they sleep on an average a day.
  - You get the data and the mean of sleeping hrs is 6.5 hours.

Compare two (or more) sample means to each other

Two general research strategies:

- **Two completely separate (independent) samples**
  - Example: Hemoglobin levels in male and female is same or not?
  - Body fat content in pig fed with two different diets
- **Two related (dependent) samples**
  - measure the size of tumor for cancer patient's before and after a treatment

Reporting Significance

Report p values as being less than .05, .01, or .001.

If a result is not significant, report p as being greater than .05 ($p > .05$)

Here are some examples…

if $p = .017$	report $p < .05$	We conclude that group means are significantly different
if $p = .005$	report $p < .01$	We conclude that group means are significantly different
If $p = .24$	report $p > .05$	We conclude that group means are NOT significantly different

173

One sample t tests using SPSS

One-Sample *t* Test SPSS

- Tests for difference between sample mean and pre-determined population mean.
- Compares the mean score of a sample to a known value. Usually, the known value is a population mean.
- Examples:
  - Comparison of mean dietary intake of a particular group of individuals with the recommended daily intake.
  - Average birth weight of new born baby in Malaysia
  - Blood pressure and glucose is normal in adults
- **Hypotheses:**
  - **Null:** There is no significant difference between the sample mean and the population mean.
    Alternate: There is a significant difference between the sample mean and the population mean.

175

One-Sample *t* Test

Assumption #1: Your **dependent variable** should be measured at the **interval** or **ratio level** (i.e., **continuous**).

Assumption #2: The data are **independent** (i.e., **not correlated/related**), which means that there is no relationship between the observations.

Assumption #3: There should be **no significant outliers**.

Assumption #4: Your **dependent variable** should be **approximately normally distributed**.

176

Example of study

- A researcher is planning a psychological intervention study, but before he proceeds he wants to characterize his participants' depression levels.
- He tests each participant on a particular depression index, where anyone who achieves a **score of 4.0 is deemed to have 'normal' levels of depression**. Lower scores indicate less depression and higher scores indicate greater depression. The study included 40 participants. Depression scores are recorded in the variable dep_score. He wants to know whether his sample is representative of the normal population **(i.e., do they score statistically significantly differently from 4.0).**

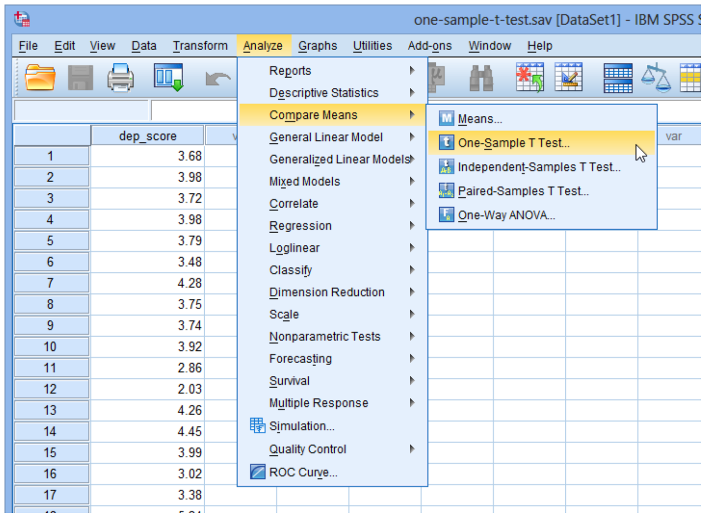
177

One-Sample *t* Test SPSS

- Click "Analyze" → "Compare Means" → "One- Sample T Test…"
  - "Test Value" = Predetermined population mean
  - EXAMPLE:
    - Compared the mean depression score of 4.0 a known population value **of 4**

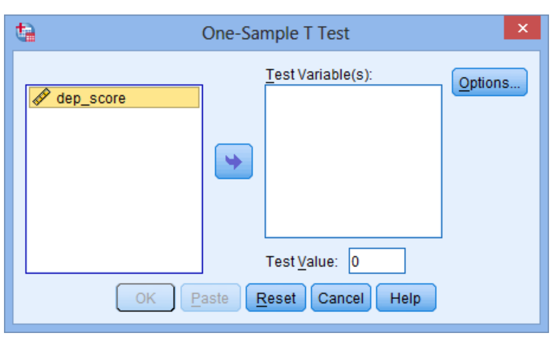
178

Enter the dependent variable, dep_score, (depression score).



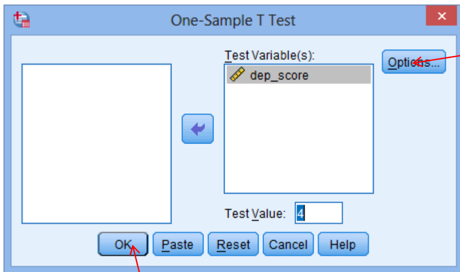
179

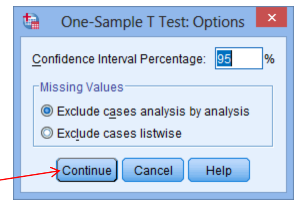
You will be presented with the **One-Sample T Test** dialogue box, as shown below:



180

- Transfer the dependent variable, dep_score, into the Test Variable(s): box
- Enter the population mean you are comparing the sample against in the Test Value: box, by changing the current value of **"0" to "4".**





Click on the options button.
You will be presented with the One-Sample T Test: Options dialogue box

Click the continue button.

Then click the OK to generate the output.

Interpreting the SPSS output of the one-sample t-test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
dep_score	40	3.7225	.73709	.11654

Mean depression score **(3.72 ± 0.74)** was lower than the population 'normal' depression score of **4.0.**

One-sample t-test

One-Sample Test

	Test Value = 4					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
dep_score	-2.381	39	.022	-.27750	-.5132	-.0418

Interpretation:

You are presented with the observed *t*-value ("**t**" column), the degrees of freedom ("**df**"), and the statistical significance (*p*-value) ("**Sig. (2-tailed)**".

In this example, $p < .05$ (it is $p = .022$).

Therefore, it can be concluded that the population means are statistically significantly different.

If *p* > .05, the difference between the sample-estimated population mean and the comparison population mean would not be statistically significantly different.

183

Writing in the manuscript

- A one-sample t-test was run to determine whether depression score in recruited subjects was different to normal, defined as a depression score of 4.0.

- Mean depression score (3.73 ± 0.74) was lower than the normal depression score of 4.0, a statistically significant difference of 0.28 (95% CI, 0.04 to 0.51), $t(39) = -2.831$, $p = .022$.

184

Paired Samples t Test using SPSS

Paired Samples t Test

- Tests if two related samples differ significantly from one another
- Same individuals are studied more than once in different time.
  - Measurement made on the sample people before and after.
  - **Example:** To evaluate the effect of new diet on weight loss.
  - Cholesterol levels before and after drug administration for a group of people

186

Hypothesis:

- **Null:** There is no significant difference between the means of the two variables.

- **Alternate:** There is a significant difference between the means of the two variables.

Example SPSS output:

- Compare the mean test scores before (pre-test) and after (post-test) the subjects completed a test preparation course.
- We want to see if test preparation course improved people's score on the test.

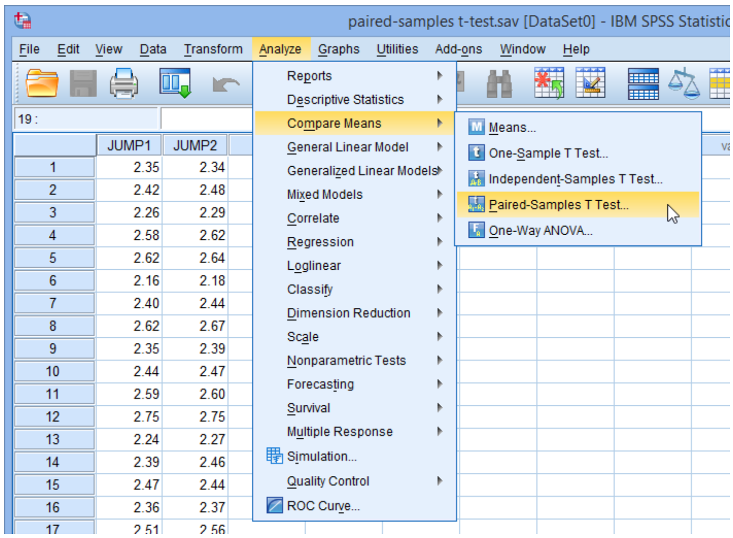
187

Example

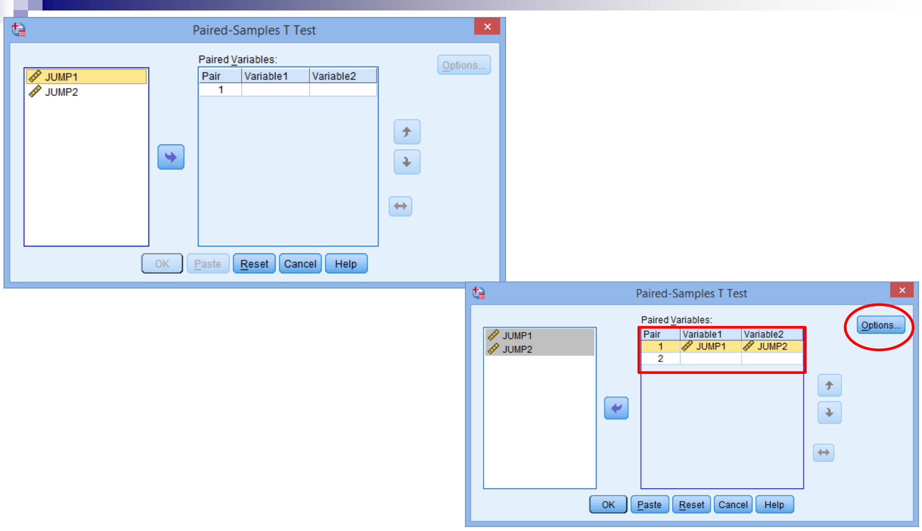
- A group of Sports Science students **(*n* = 20)** are selected from the population to investigate whether a 12-week plyometric-training programme improves their standing long jump performance.
- In order to test whether this training improves performance, the students are tested for their long jump performance before they undertake a plyometric-training programme and then again at the end of the programme (i.e., the dependent variable is "standing long jump performance", and the two related groups are the standing long jump values **"before" and "after"** the 12-week plyometric-training programme).

188

Click **Analyze > Compare Means > Paired-Samples T Test...** on the top menu, as shown below:

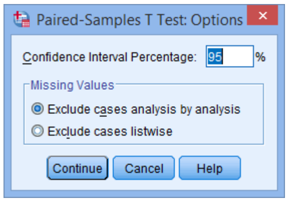


189

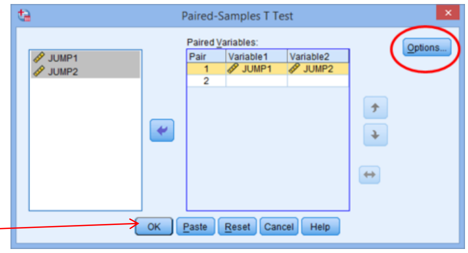


- Transfer the variables JUMP1 and JUMP2 into the Paired Variables: box.
- Click on the options button.

190



Click the continue



- You will be returned to the **Paired-Samples T Test** dialogue box.
- Click OK button.

Output of the Dependent T-Test in SPSS

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	JUMP1	2.4815	20	.16135	.03608
	JUMP2	2.5155	20	.15982	.03574

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	JUMP1 - JUMP2	-.03400	.03185	.00712	-.04891	-.01909	-4.773	19	.000

Reporting the Output of the Dependent T-Test

You might report the statistics in the following format: *t*(degrees of freedom) = *t*-value, *p* = significance level. In our case this would be: *t*(19) = -4.773, $p < 0.0005$. Due to the means of the two jumps and the direction of the *t*-value, we can conclude that there was a statistically significant improvement in jump distance following the plyometric-training programme from 2.48 ± 0.16 m to 2.52 ± 0.16 m ($p < 0.0005$); an improvement of 0.03 ± 0.03 m.

Two Samples independent t Test using SPSS

Independent-Samples *t* Test

Compares the means between two unrelated groups on the same continuous, dependent variable.

- Example: To study the weight gain of fish fed with low and high protein fed groups.
- Example: First year graduate salaries differed based on gender (i.e., dependent variable would be "first year graduate salaries" and independent variable would be "gender", which has two groups: "male" and "female").

Independent Sample Data (Data are time off task)

Experimental (Caff)	Control (No Caffeine)
12	21
14	18
10	14
8	20
16	11
5	19
3	8
9	12
11	13
	15
N_1=9, M_1=9.778, SD_1=4.1164	N_2=10, M_2=15.1, SD_2=4.2805

195

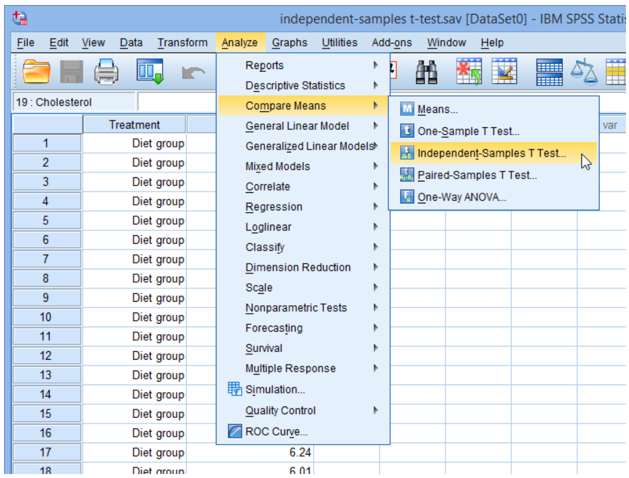
Study Example

- The concentration of cholesterol in the blood is associated with the risk of developing heart disease, such that higher concentrations of cholesterol indicate a higher level of risk, and lower concentrations indicate a lower level of risk.
- If you lower the concentration of cholesterol in the blood, your risk of developing heart disease can be reduced. Being overweight and/or physically inactive increases the concentration of cholesterol in your blood.
- **Both exercise and weight loss can reduce cholesterol concentration**. However, it is not known whether exercise or weight loss is best for lowering cholesterol concentration.
- Investigated whether an exercise or weight loss intervention is more effective in lowering cholesterol levels.
- The researcher recruited a random sample of inactive males that were classified as overweight.
- This sample was then randomly split into two groups:
  - **Group 1 underwent a calorie-controlled diet and**
  - **Group 2 undertook the exercise-training programme.**
- In order to determine which treatment programme was more effective, **the mean cholesterol concentrations were compared between the two groups** at the end of the treatment programmes.

196

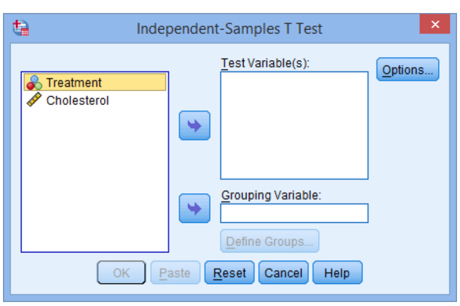
Test Procedure in SPSS

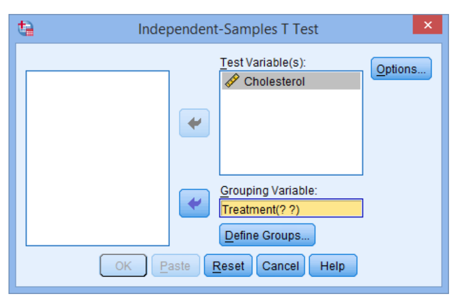
- Click **Analyze > Compare Means > Independent-Samples T Test...** on the top menu, as shown below:



197

You will be presented with the **Independent-Samples T Test** dialogue box, as shown below:

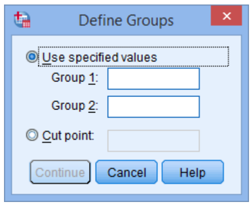


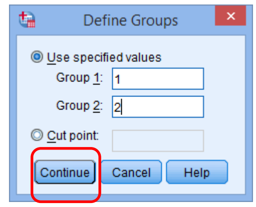


- Transfer the dependent variable, Cholesterol, into the Test Variable(s): box, and transfer the independent variable, Treatment, into the Grouping Variable box

198

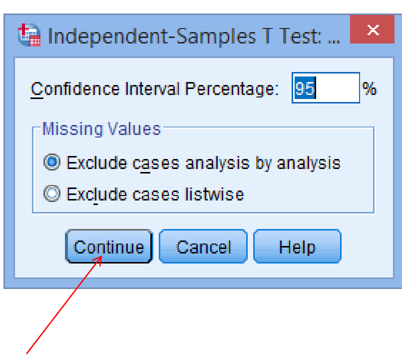
You then need to define the groups (treatments). Click on the **Define Groups** button.
You will be presented with the **Define Groups** dialogue box, as shown below:

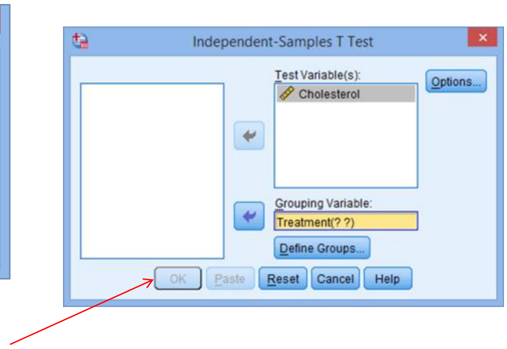




- Enter **1** into the Group 1: box and enter **2** into the Group 2: box.
- Remember that we labelled the **Diet Treatment** group as **1** and the **Exercise Treatment** group as **2**.
- Click the continue

199





- Click the continue button.
- You will be returned to the **Independent-Samples T Test** dialogue box.
- Click the Ok button.

200

Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
Cholesterol Concentration	Diet	20	6.1450	.51959	.11618
	Exercise	20	5.7950	.38179	.08537

Independent Samples Test

			Cholesterol Concentration	
			Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F		.314	
	Sig.		.579	
t-test for Equality of Means	t		2.428	2.428
	df		38	34.886
	Sig. (2-tailed)		.020	.021
	Mean Difference		.35000	.35000
	Std. Error Difference		.14418	.14418
	95% Confidence Interval of the Difference	Lower	.05813	.05727
		Upper	.64187	.64273

Group means are significantly different because the value in the "**Sig. (2-tailed)**" row is less than 0.05.

Interpretation:

This study found that significantly lower cholesterol concentrations **(5.80 ± 0.38 mmol/L)** at the end of an exercise-training programme compared to after a calorie-controlled diet (6.15 ± 0.52 mmol/L), *t*(38) = 2.428, *p* = 0.020.

One-way analysis of variance (ANOVA)using SPSS

One-way analysis of variance (ANOVA)

- The one-way analysis of variance (ANOVA) is used to determine whether there are any significant differences between the means of three or more independent (unrelated) groups.
  - For example, you could use a one-way ANOVA to understand whether **exam performance** differed based on test anxiety levels amongst students, dividing students **into three independent groups (e.g., low, medium and high-stressed students).**

203

Assumptions

- **Assumption #1:** Your **dependent variable** should be measured at the **interval** or **ratio level** (i.e., they are **continuous**).
  - Examples of variables that meet this criterion include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg).
- **Assumption #2:** Your **independent variable** should consist of **two or more categorical**, **independent groups**
- **Assumption #3:** You should have **independence of observations**, which means that there is no relationship between the observations in each group or between the groups themselves.
- **Assumption #4:** There should be **no significant outliers**.
- **Assumption #5:** Your **dependent variable** should be **approximately normally distributed for each category of the independent variable**.

204

Example

- A manager wants to raise the productivity at his company by increasing the speed at which his employees can use a particular spreadsheet program. As he does not have the skills in-house, he employs an external agency which provides training in this spreadsheet program.
- **They offer 3 courses:** a **beginner, intermediate and advanced course**. He is not sure which course is needed for the type of work they do at his company, so he sends **10 employees on the beginner course, 10 on the intermediate and 10 on the advanced course**. When they all return from the training, he gives them a problem to solve using the spreadsheet program, and times how long it takes them to complete the problem. He then compares the three courses **(beginner, intermediate, advanced)** to see if there are any differences in the **average time it took** to complete the problem.

205

SPSS one-way analysis of variance

- SPSS procedure *Analysis, Compare Means, One-Way ANOVA*
- *Dependent List* is for variable for which means are to be calculated, compared.
- *Factor* is for variable used to designate the different samples or groups
- *Options* to specify *Descriptive Statistics*
- *Post Hoc* for multiple comparisons

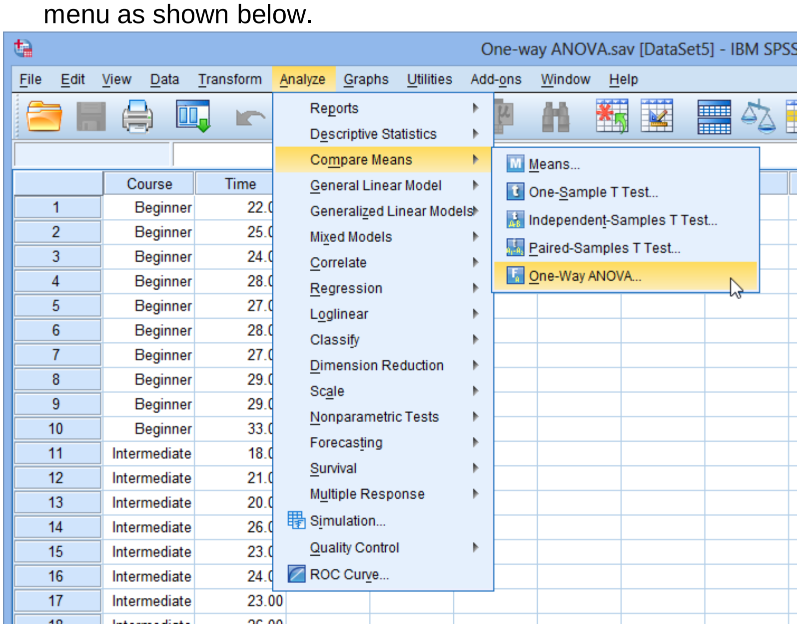
206

Steps

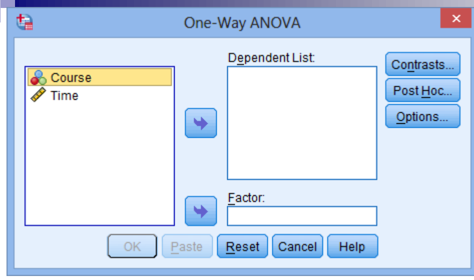
- In SPSS, enter the groups for analysis by creating a grouping variable called Course (i.e., the independent variable), and give the beginners course a value of "1", the intermediate course a value of "2" and the advanced course a value of "3".
- Time to complete the set problem was entered under the variable name Time (i.e., the dependent variable).

207

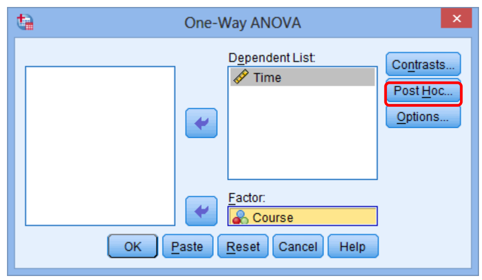
Click **Analyze > Compare Means > One-Way ANOVA...** on the top menu as shown below.



208



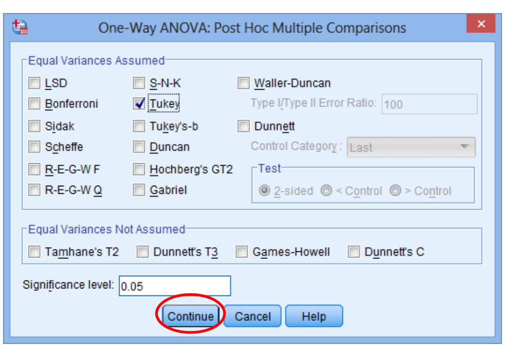
Transfer the dependent variable (Time) into the Dependent List: box and the independent variable (Course) into the Factor:



Click the post hoc button.

209

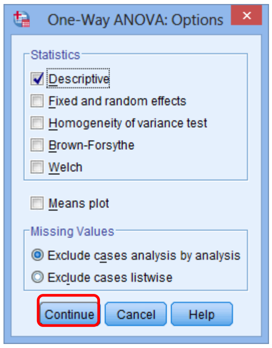
Tick the Tukey checkbox as shown below:

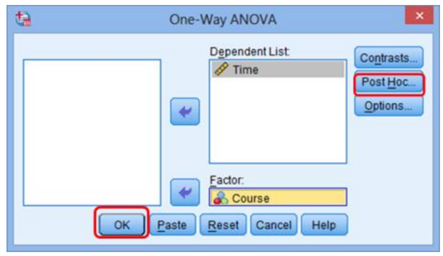


- Click the continue button.

210

Click the options button. Tick the Descriptive checkbox in the –Statistics– area, as shown below:





- Click the continue button.
- Click the OK button.

Descriptives table

Descriptives

Time

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Beginner	10	27.2000	3.04777	.96379	25.0198	29.3802	22.00	33.00
Intermediate	10	23.6000	3.30656	1.04563	21.2346	25.9654	18.00	29.00
Advanced	10	23.4000	3.23866	1.02415	21.0832	25.7168	18.00	29.00
Total	30	24.7333	3.56161	.65026	23.4034	26.0633	18.00	33.00

The descriptives table provides some very useful descriptive statistics, including the mean, standard deviation for the **dependent variable (Time)** **for each separate group (Beginners, Intermediate and Advanced)**.

ANOVA

Time

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	91.467	2	45.733	4.467	.021
Within Groups	276.400	27	10.237		
Total	367.867	29			

- Statistically significant difference between our group means. We can see that the significance level is 0.021 (*p* = .021), which is below 0.05. and, therefore, there is a **statistically significant difference** in the mean length of time to complete the spreadsheet problem between the different courses taken.

213

Multiple Comparisons

Dependent Variable: Time

Tukey HSD

(I) Course	(J) Course	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Beginner	Intermediate	3.60000*	1.43088	.046	.0523	7.1477
	Advanced	3.80000*	1.43088	.034	.2523	7.3477
Intermediate	Beginner	-3.60000*	1.43088	.046	-7.1477	-.0523
	Advanced	.20000	1.43088	.989	-3.3477	3.7477
Advanced	Beginner	-3.80000*	1.43088	.034	-7.3477	-.2523
	Intermediate	-.20000	1.43088	.989	-3.7477	3.3477

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

- Multiple Comparisons, shows that, there is a significant difference in time to complete the problem between the group that took the **beginner course and the intermediate course** **(*p* = 0.046),** as well as between the **beginner course and advanced course** **(*p* = 0.034).**
- **There were no differences between the groups that took the intermediate and advanced course (*p* = 0.989).**

214

Interpretation and reporting

ANOVA

Time

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	91.467	2	45.733	4.467	.021
Within Groups	276.400	27	10.237		
Total	367.867	29			

Multiple Comparisons

Dependent Variable: Time

Tukey HSD

(I) Course	(J) Course	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Beginner	Intermediate	3.60000*	1.43088	.046	.0523	7.1477
	Advanced	3.80000*	1.43088	.034	.2523	7.3477
Intermediate	Beginner	-3.60000*	1.43088	.046	-7.1477	-.0523
	Advanced	.20000	1.43088	.989	-3.3477	3.7477
Advanced	Beginner	-3.80000*	1.43088	.034	-7.3477	-.2523
	Intermediate	-.20000	1.43088	.989	-3.7477	3.3477

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

There was a statistically significant difference between groups as determined by one-way ANOVA ($F(2,27) = 4.467$, $p = .021$). A Tukey post hoc test revealed that the time to complete the problem was statistically significantly lower after taking the intermediate **(23.6 ± 3.3 min, $p = .046$)** and advanced **(23.4 ± 3.2 min, $p = .034$)** course compared to the beginners course **(27.2 ± 3.0 min).** There were no statistically significant differences between the intermediate and advanced groups ($p = .989$).

215

Example

Effect of dietary protein levels on final body weight (g) of catfish fed for 30 days
T1- 30% protein; T2 – 35%, T3 – 40%; T4-45%.

Treatment 1	Treatment 2	Treatment 3	Treatment 4
60 g	50	48	47
67	52	49	67
42	43	50	54
67	67	55	67
56	67	56	68
62	59	61	65
64	67	61	65
59	64	60	56
72	63	59	60
71	65	64	65

216

Two-way analysis of variance (ANOVA)using SPSS

Two-way ANOVA in SPSS

The two-way ANOVA compares the **mean differences between groups that have been split on two independent variables** (called factors).

The primary purpose of a two-way ANOVA is to understand if there is an **interaction between the two independent variables on the dependent variable**.

Example:

Two-way ANOVA is used to understand whether there is an interaction between **gender and educational level on test anxiety amongst university students**, where gender (males/females) and education level (undergraduate/postgraduate) are independent variables, **and test anxiety** is **dependent variable**

218

Example: 2

To determine whether there is an interaction between **physical activity level and gender** on **blood cholesterol concentration** in children, where **physical activity (low/moderate/high)** and **gender (male/female)** are independent variables, and cholesterol concentration is dependent variable.

219

Example

A researcher was interested in whether an individual's interest in politics was influenced by their level of education and gender. They recruited a random sample of participants to their study and asked them about their interest in politics, which they scored from **0 to 100, with higher scores indicating a greater interest in politics**.

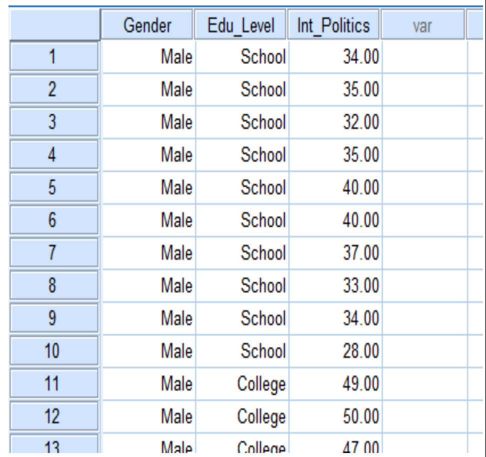
The researcher then divided the participants by **gender (Male/Female) and then again by level of education (School/College/University).**

The **dependent variable** is **"interest in politics",** and the **two independent variables** are **"gender" and "education".**

220

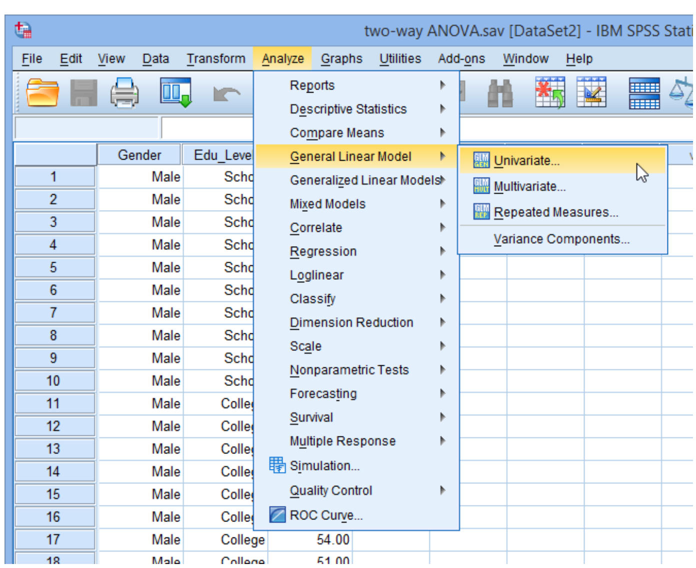
Two-way ANOVA in SPSS

- Enter the two independent variables, and label them Gender and Edu_Level.
- For Gender, we code **"males" as 1 and "females" as 2,** and for **Edu_Level, we code "school" as 1, "college" as 2 and "university" as 3.**
- The participants' interest in politics – the dependent variable – enter under the variable name, **Int_Politics.**

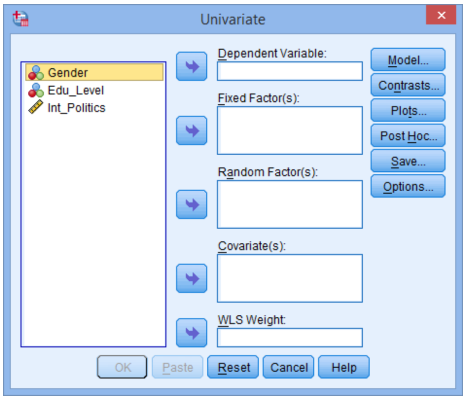


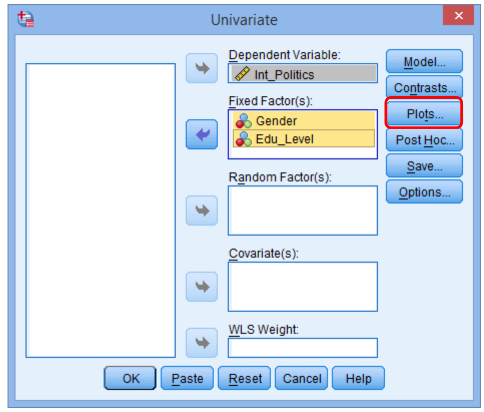
	Gender	Edu_Level	Int_Politics	var
1	Male	School	34.00	
2	Male	School	35.00	
3	Male	School	32.00	
4	Male	School	35.00	
5	Male	School	40.00	
6	Male	School	40.00	
7	Male	School	37.00	
8	Male	School	33.00	
9	Male	School	34.00	
10	Male	School	28.00	
11	Male	College	49.00	
12	Male	College	50.00	
13	Male	College	47.00	

Click **Analyze > General Linear Model > Univariate...**



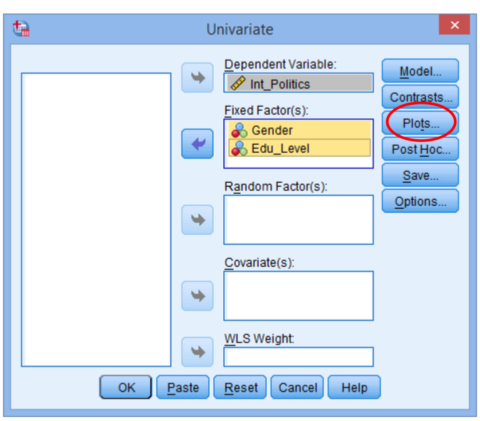
You will be presented with the **Univariate** dialogue box, as shown below:

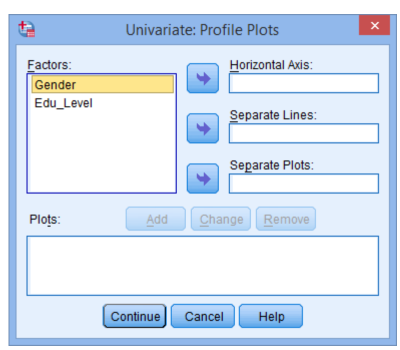




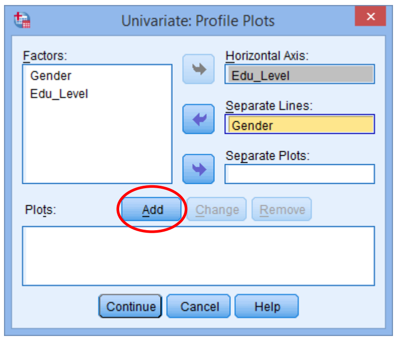
Transfer the dependent variable, Int_Politics, into the Dependent Variable: box, and transfer both independent variables, Gender and Edu_Level, into the Fixed Factor(s): box.

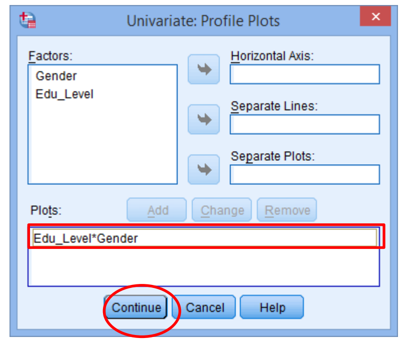
Click on the plots button.
You will be presented with the **Univariate: Profile Plots** dialogue box, as shown below





Transfer the independent variable, Edu_Level, from the Factors: box into the Horizontal Axis: box, and transfer the other independent variable, Gender, into the Separate Lines: box.



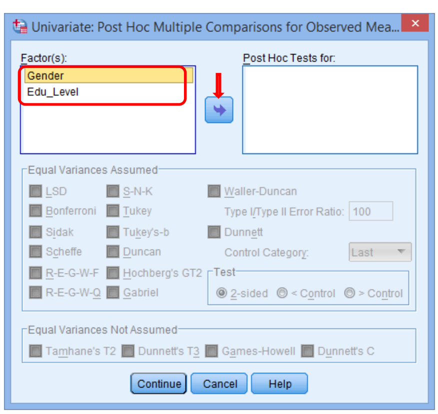


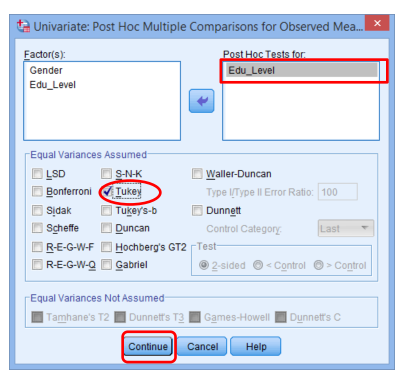
Click the Add Button.
You will see that "Edu_Level*Gender" has been added to the Plots: box, as shown above:

Click the **continue** button.

225

Click the post hoc button. You will be presented with the **Univariate: Post Hoc Multiple Comparisons for Observed Means** dialogue box, as shown below:



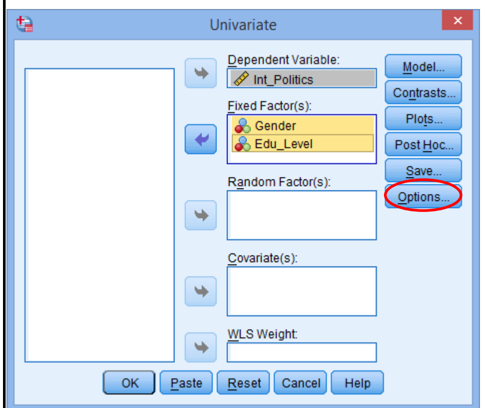


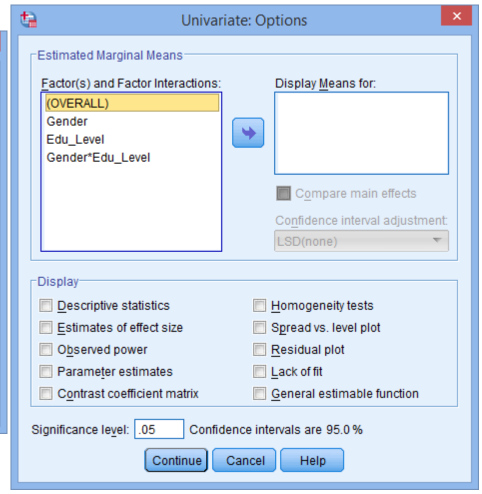
- Transfer Edu_Level from the Factor(s): box to the Post Hoc Tests for: box.
- This will make the –Equal Variances Assumed– area become active and present you with some choices for which post hoc test to use.
- The select Tukey post hoc test.

226

Click the continue button to return to the **Univariate** dialogue box.

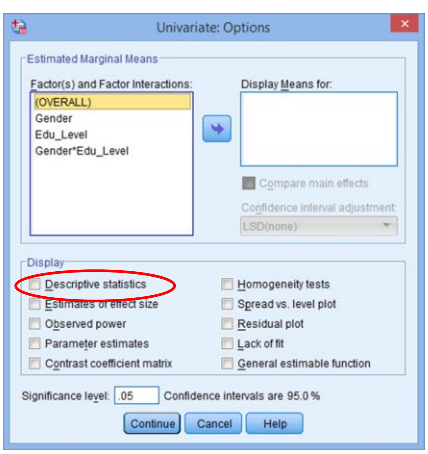
Click the **options** button.
This will present you with the **Univariate: Options** dialogue box, as shown below:

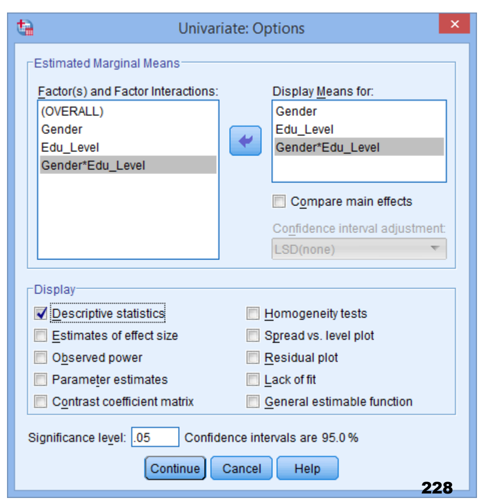




227

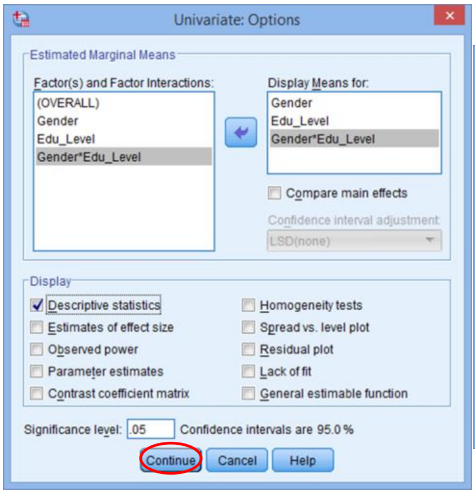
Transfer **Gender**, **Edu_Level** and **Gender*Edu_Level** from the Factor(s) and Factor Interactions: box into the Display Means for: box.
In the –Display– area, tick the Descriptive Statistics option. You will presented with the following screen:

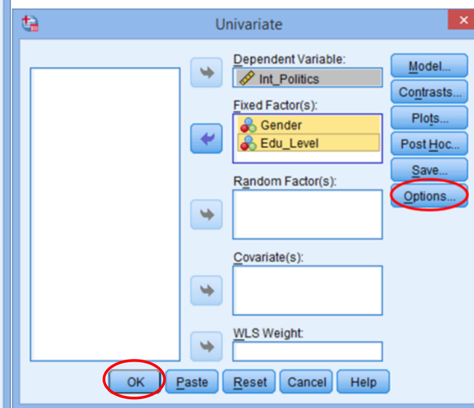




228

Click the continue button to return to the **Univariate** dialogue box.
Click the OK button to generate the output.





229

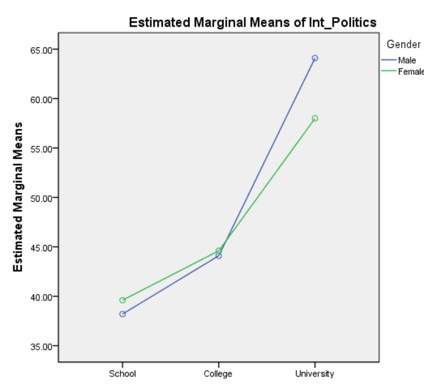
Descriptive Statistics

Dependent Variable:Int_Politics

Gender	Edu_Level	Mean	Std. Deviation	N
Male	School	38.2000	4.18463	10
	College	44.1000	4.26745	10
	University	64.1000	3.07137	10
	Total	48.8000	11.87841	30
Female	School	39.6000	3.27278	10
	College	44.6000	3.27278	10
	University	58.0000	6.46357	10
	Total	47.4000	9.05767	30
Total	School	38.9000	3.72615	20
	College	44.3500	3.71023	20
	University	61.0500	5.83524	20
	Total	48.1000	10.49649	60

Plot of the results

The plot of the mean "interest in politics" score for each combination of groups of "Gender" and "Edu_level" are plotted in a line graph, as shown below:



230

Statistical significance of the two-way ANOVA

Tests of Between-Subjects Effects

Dependent Variable:Int_Politics

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	5525.200[a]	5	1105.040	61.190	.000
Intercept	138816.600	1	138816.600	7686.727	.000
Gender	29.400	1	29.400	1.628	.207
Edu_Level	5328.100	2	2664.050	147.517	.000
Gender * Edu_Level	167.700	2	83.850	4.643	.014
Error	975.200	54	18.059		
Total	145317.000	60			
Corrected Total	6500.400	59			

a. R Squared = .850 (Adjusted R Squared = .836)

- Our independent variables (the "Gender" and "Edu_Level" rows) and their interaction (the "Gender*Edu_Level" row) have a statistically significant effect on the dependent variable, "interest in politics".
- See from the "**Sig.**" column that we have a statistically significant interaction at the *p* = .014 level.
- There was no statistically significant difference in mean interest in politics between males and females (*p* = .207), **but there were statistically significant differences between educational levels (*p* < .0005).**

231

Post hoc tests - Multiple Comparisons Table

Multiple Comparisons

Int_Politics
Tukey HSD

(I) Edu_Level	(J) Edu_Level	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
School	College	-5.4500*	1.34385	.000	-8.6887	-2.2113
	University	-22.1500*	1.34385	.000	-25.3887	-18.9113
College	School	5.4500*	1.34385	.000	2.2113	8.6887
	University	-16.7000*	1.34385	.000	-19.9387	-13.4613
University	School	22.1500*	1.34385	.000	18.9113	25.3887
	College	16.7000*	1.34385	.000	13.4613	19.9387

Based on observed means.
The error term is Mean Square(Error) = 18.059.

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

There is a statistically significant difference between all three different educational levels (p < .0005).

232

Correlation Coefficients: The Pearson *r* & Spearman's *rho*

Correlation Coefficients

Definition: A correlation coefficient is a statistic that indicates the strength & direction of the relationship b/w 2 variables.

- Correlation coefficients provide a single numerical value to represent the relationship b/w the 2 variables
- Correlation coefficients ranges -1 to +1

-1.00 (negative one) a perfect, inverse relationship

+1.00 (positive one) a perfect, direct relationship

0.00 indicates no relationship

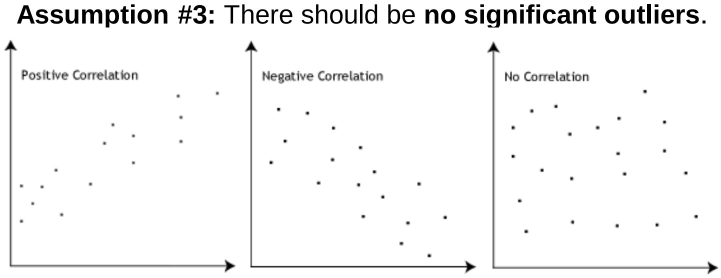
Different indices of correlation coefficient

- **Pearson product moment correlation coefficient (r)** – used to assess the relationship between **2 normally distributed continuous variables** (Parametric)
- **Spearman rank order correlation (rho) coefficient -** used to assess the relationship between **2 continuous variables** and one of which is not normally distributed (Non parametric)
- **Kendall's rank correlation coefficient -** used to assess the relationship between **2 ordinal variables** or **one ordinal and one continuous variable** (Non parametric)

235

Pearson Correlation Assumptions

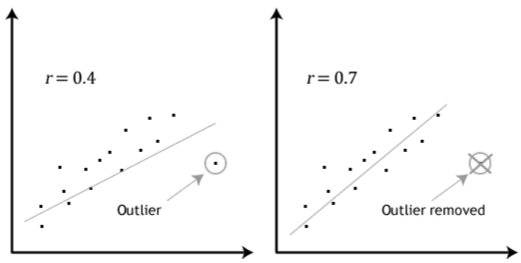
- **Assumption #1:** Your two variables should be measured at the **interval** or **ratio level** (i.e., they are **continuous**).
- **Assumption #2:** There needs to be a **linear relationship** between the two variables. Check the linear relationship exists between your two variables, by scatterplot using SPSS.
- **Assumption #3:** There should be **no significant outliers**.



236

Assumptions

- **Assumption #3:** There should be **no significant outliers**.



- **Assumption #4:** Your variables should be **approximately normally distributed**.

237

Example

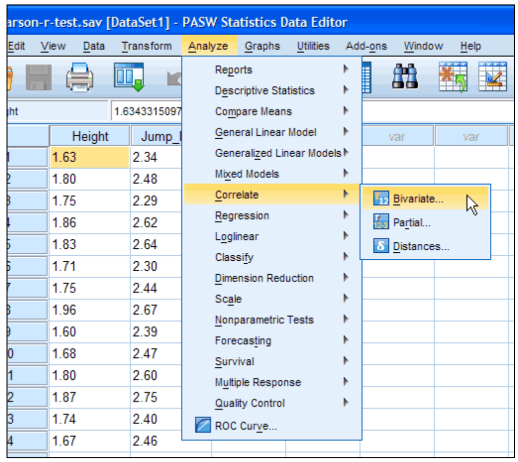
- Person's height is related to how well they perform in a long jump.
- The researcher recruited untrained individuals from the general population, measured their **height** and had them perform a **long jump**.
- **The researcher then investigated whether there is an association between height and long jump performance.**

238

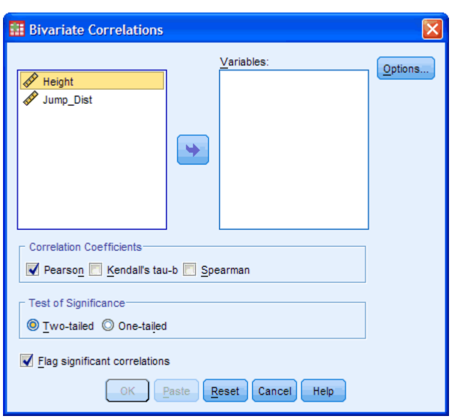
In SPSS, enter the two variables

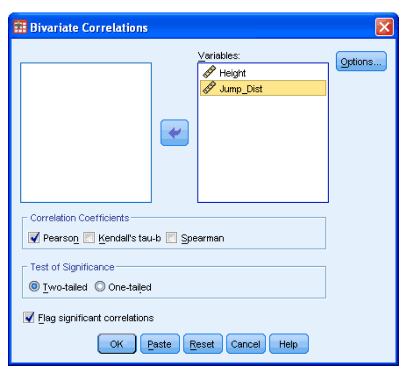
Height (i.e., the person's height) and JumpDist (i.e., long jump distance).

Click **Analyze > Correlate > Bivariate...** on the menu system as shown below



You will be presented with the following screen:

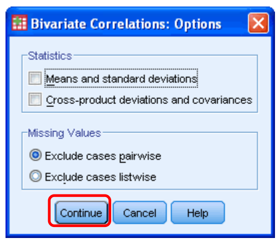


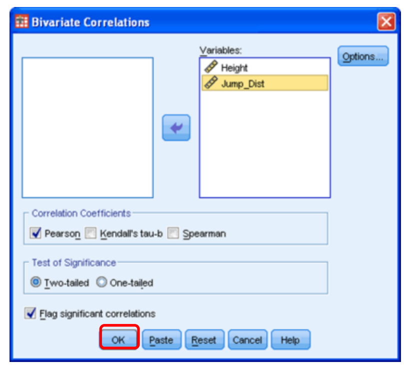


Transfer the variables Height and Jump_Dist into the Variables: box

Make sure that the Pearson tick box is checked under the -Correlation Coefficients- area

- Click the options button. If you wish to generate some descriptives, you can do it here by clicking on the relevant tickbox under the -Statistics- area.





Click the continue.
Then Click the OK.

Correlations table in the output

Correlations

		Height	Jump_Dist
Height	Pearson Correlation	1	.777**
	Sig. (2-tailed)		.000
	N	27	27
Jump_Dist	Pearson Correlation	.777**	1
	Sig. (2-tailed)	.000	
	N	27	27

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

Pearson correlation coefficient, *r*, is 0.777, and that this is statistically significant ($p < 0.0005$).

Reporting the results:

A Pearson product-moment correlation was run to determine the relationship between an **individual's height and their performance in a long jump** (distance jumped). There was a **strong, positive correlation** between height and distance jumped, which was statistically significant ($r = .777$, $n = 27$, $p < .0005$).

Linear Regression Analysis using SPSS

Regression Analysis

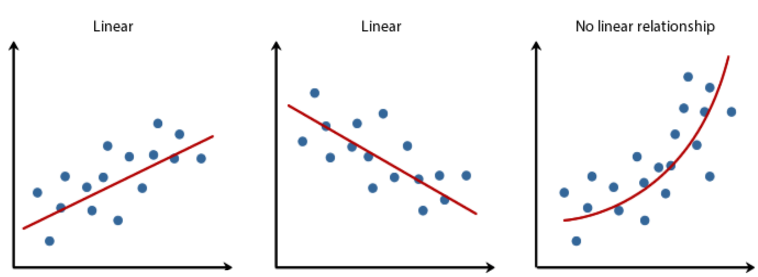
Regression Analysis is the estimation of the linear relationship between a dependent variable and one or more independent variables or covariates.

When to use Correlation and Linear regression

- Correlation and linear regression are used when you have two measurement variables, such as
  - food intake and weight,
  - drug dosage and blood pressure,
  - air temperature and metabolic rate, etc.

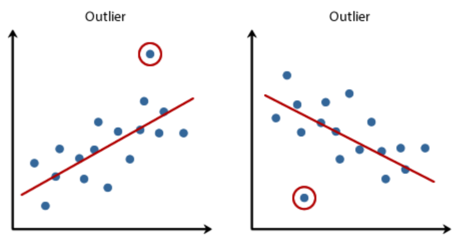
Assumptions

- **Assumption #1:** Your two variables should be measured at the **continuous** level (i.e., they are either **interval** or **ratio** variables). Examples of **continuous variables** include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg).
- **Assumption #2:** There needs to be a **linear relationship** between the two variables.



Assumptions

- **Assumption #3:** There should be **no significant outliers**.



Linear Regression and Correlation

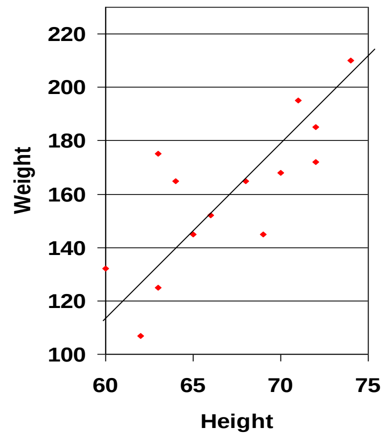
- **In many situations in clinical studies we wish to attempt to answer the question: How is the random variable X related to the random variable Y?**
  - Ex: How is smoking related to lung cancer?
  - Ex: How is age related to development of Alzheimer's Disease?
  - Ex: How is age related to blood pressure?

Such questions are answered statistically using the concepts of *Regression Analysis* which looks for relationships among different variables (either negatively or positively) and *Correlations*, the strengths of the relationships

Linear Regression

Scatterplots

- In order to perform regression analysis visually, it helps to "graph" the points on a scatterplot
- A visual relationship can often be observed when looking at these plots.
- Need to draw the ***line of best fit.***
- ***Best fit means that the sum of the squares of the vertical distances from each point to the line is at minimum.***



249

Is there a relationship between wing length and tail length in songbirds?

Wing length cm	Tail length cm
10.4	7.4
10.8	7.6
11.1	7.9
10.2	7.2
10.3	7.4
10.2	7.1
10.7	7.4
10.5	7.2
10.8	7.8
11.2	7.7
10.6	7.8
11.4	8.3

Is there a relationship between age and systolic blood pressure?

Age (yr)	Systolic blood pressure mm hg
30	108
30	110
30	106
40	125
40	120
40	118
40	119
50	132
50	137
50	134
60	148
60	151
60	146
60	147
60	144
70	162
70	156
70	164
70	158
70	159

250

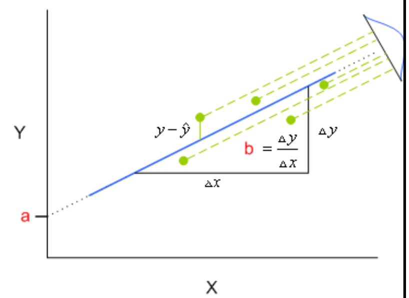
Linear Regression

- The Linear Regression model postulates that two random variables X and Y are related by a straight line as follows:

Y = a + bX

Where

Y is the *dependent variable*
X is the *independent variable*
a is the *intercept*
b is the *slope*



251

LINEAR REGRESSION

- Linear Regression estimates the coefficients of the linear equation, involving one independent variables that best predict the value of the dependent variable.
- Examples:

Birth weight of a baby (independent) and blood pressure (dependent)

Y = a + bX

Blood pressure = a + b (body weight)

252

Example

- A salesperson for a large car brand wants to determine whether there is a relationship between an individual's income and the price they pay for a car.
- The individual's **"income" is the independent variable** and the **"price" they pay for a car is the dependent variable**.

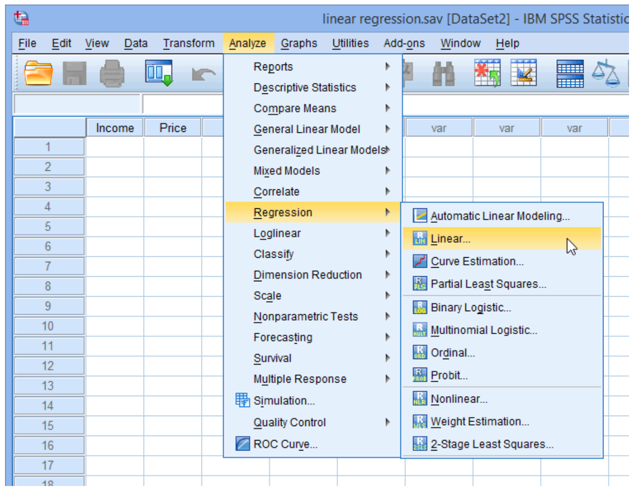
253

SPSS

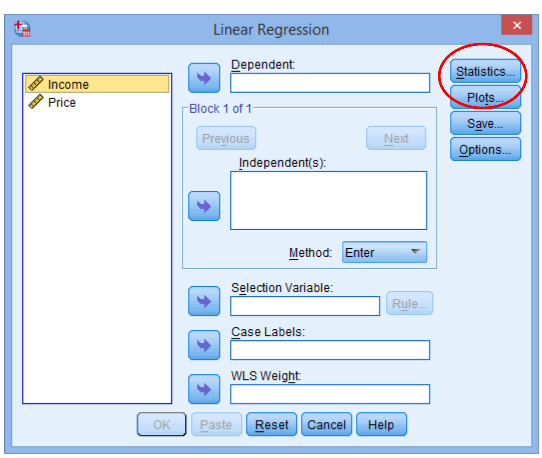
- In SPSS Statistics, enter the two variables
  - **Income (the independent variable), and**
  - **Price (the dependent variable).**

254

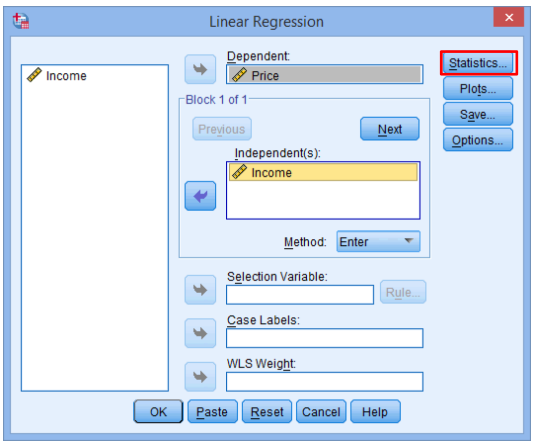
- Click **Analyze > Regression > Linear...** on the top menu, as shown below:

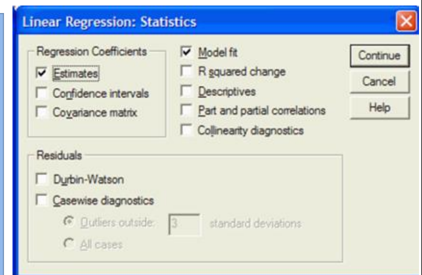


You will be presented with the **Linear Regression** dialogue box



Transfer the independent variable, Income, into the Independent(s): box and the dependent variable, Price, into the Dependent: box.





Click the OK button.

- The first table of interest is the **Model Summary** table, as shown below:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.873[a]	.762	.749	874.779

a. Predictors: (Constant), Income

- This table provides the *R* and R^2 values.
- The *R* value represents the simple correlation and is 0.873 (the "**R**" Column), which indicates a high degree of correlation.
- The R^2 value (the "**R Square**" column) indicates how much of the total variation in the dependent variable, **Price, can be explained by the independent variable, Income**. In this case, 76.2% can be explained, which is very large.

- The next table is the **ANOVA** table, which reports how well the regression equation fits the data (i.e., predicts the dependent variable) and is shown below:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	44182633.37	1	44182633.37	57.737	.000[b]
	Residual	13774291.07	18	765238.393		
	Total	57956924.44	19			

a. Dependent Variable: Price

b. Predictors: (Constant), Income

- This table indicates that the regression model predicts the dependent variable significantly well.
- How do we know this?
- Look at the "**Regression**" row and go to the "**Sig.**" column. This indicates the statistical significance of the regression model that was run. Here, $p < 0.0005$, which is less than 0.05, and indicates that, overall, the regression model statistically significantly predicts the outcome variable (i.e., it is a good fit for the data).

- The **Coefficients** table provides us with the necessary information to predict price from income, as well as determine whether income contributes statistically significantly to the model (by looking at the "**Sig.**" column). Furthermore, we can use the values in the "**B**" column under the "**Unstandardized Coefficients**" column, as shown below:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	8286.786 (a)	1852.256		4.474	.000
	Income	.564 (b)	.074	.873	7.598	.000

a. Dependent Variable: Price

to present the regression equation as:

Y = a+bX

Price = 8287 + 0.564 (Income)

Probit Analysis using SPSS

Probit Analysis

- Probit analysis is a type of regression used to analyze binomial response variables.
- This procedure measures the relationship between the strength of a stimulus and the proportion of cases exhibiting a certain response to the stimulus.
- Probit analysis is still the preferred statistical method in understanding **dose-response relationships**

Example.

- How effective is a new pesticide at killing ants, and what is an appropriate concentration to use?
- You might perform an experiment in which you expose samples of ants to different concentrations of the pesticide and then record the number of ants killed and the number of ants exposed.
- Applying probit analysis to these data, you can determine the strength of the relationship between concentration and killing, and you can determine what the appropriate concentration of pesticide would be if you wanted to be sure to kill, say, 95% of exposed ants.

263

Probit Analysis

- Remember that a binomial response variable refers to a response variable with only two outcomes.
- For example:
  - Flipping a coin: Heads or tails
  - Testing beauty products: Rash/no rash
  - The effectiveness or toxicity of pesticides: Death/no death

264

Application

- Probit analysis is used to analyze many kinds of dose-response or binomial response experiments in a variety of fields.
- Probit Analysis is commonly used in **toxicology** to determine the relative toxicity of chemicals to living organisms.
- This is done by testing the response of an organism under various concentrations of each of the chemicals in question and then comparing the concentrations at which one encounters a response.
- The response is always binomial (e.g. death/no death)

265

Probit Analysis

- Once a regression is run, the researcher can use the output of the probit analysis to compare the amount of chemical required to create the same response in each of the various chemicals.
- There are many endpoints used to compare the differing toxicities of chemicals, but the **LC_{50} (liquids) or LD_{50} (solids)** are the most widely used outcomes of the modern dose-response experiments.
- The LC_{50}/LD_{50} represent the concentration (LC_{50}) or dose (LD_{50}) at which 50% of the population responds.

266

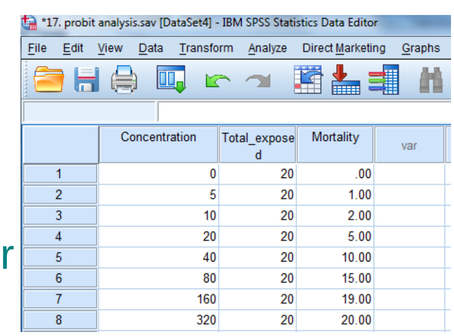
EXAMPLE

- For example, consider comparing the toxicity of two different pesticides to fish, pesticide A and pesticide B.
- If the LC_{50} of pesticide A is 50ug/L and the LC_{50} of pesticide B is 10ug/L, pesticide B is more toxic than A because it only takes 10ug/L to kill 50% of the fish, versus 50ug/L of pesticide B.

267

To run the probit anaylsis in SPSS, follow the following simple steps:

- Simply input a minimum of three columns into the Data Editor
- Number of individuals per container that responded
- Total of individuals per container
- Concentrations

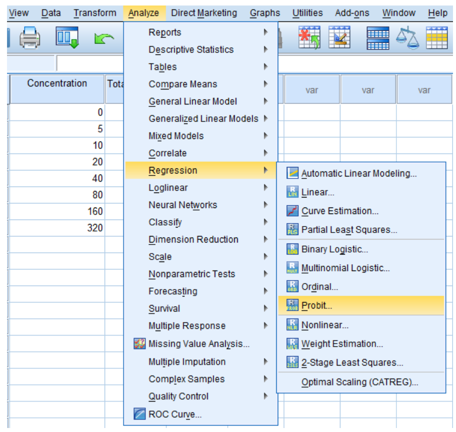


	Concentration	Total_exposed	Mortality	var
1	0	20	.00	
2	5	20	1.00	
3	10	20	2.00	
4	20	20	5.00	
5	40	20	10.00	
6	80	20	15.00	
7	160	20	19.00	
8	320	20	20.00	

268

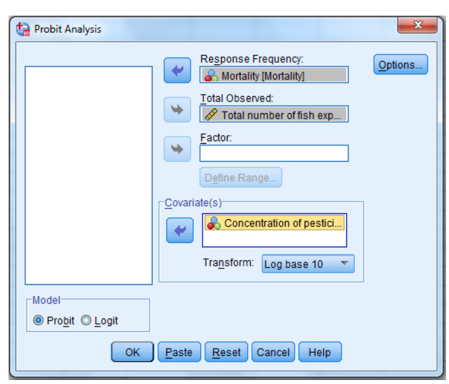
Probit analysis in SPSS

- On the main menu Click Analyze, Regression, Probit, choose the log of your choice to transform:



Probit analysis in SPSS

- Then set your number responded column as the "Response Frequency", the total number per container as the "Total Observed", and the concentrations as the "Covariates".
- Don't forget to select the log base 10 to transform your concentrations.



- Click the OK button in the Probit Analysis dialog box to run the analysis.
- The output will be displayed in a new SPSS Viewer window.

LC_{50} = 36.624 mg/L with 95 % confidence interval of 27.539 – 48.823

Confidence Limits

	Probability	95% Confidence Limits for Concentration of pesticide mg/L			95% Confidence Limits for log(Concentration of pesticide mg/L)[a]		
		Estimate	Lower Bound	Upper Bound	Estimate	Lower Bound	Upper Bound
PROBIT	.010	3.522	1.417	6.107	.547	.151	.786
	.020	4.635	2.047	7.636	.666	.311	.883
	.030	5.516	2.583	8.807	.742	.412	.945
	.040	6.288	3.075	9.810	.798	.488	.992
	.050	6.994	3.541	10.715	.845	.549	1.030
	.060	7.658	3.993	11.555	.884	.601	1.063
	.070	8.292	4.434	12.349	.919	.647	1.092
	.080	8.904	4.869	13.109	.950	.687	1.118
	.090	9.499	5.301	13.845	.978	.724	1.141
	.100	10.082	5.730	14.562	1.004	.758	1.163
	.150	12.904	7.889	17.999	1.111	.897	1.255
	.200	15.699	10.130	21.389	1.196	1.006	1.330
	.250	18.575	12.503	24.901	1.269	1.097	1.396
	.300	21.604	15.043	28.659	1.335	1.177	1.457
	.350	24.850	17.783	32.781	1.395	1.250	1.516
	.400	28.381	20.752	37.400	1.453	1.317	1.573
	.450	32.273	23.989	42.678	1.509	1.380	1.630
	.500	36.624	27.539	48.823	1.564	1.440	1.689
	.550	41.562	31.469	56.112	1.619	1.498	1.749
	.600	47.262	35.871	64.934	1.675	1.555	1.812
	.650	53.977	40.884	75.855	1.732	1.612	1.880
	.700	62.087	46.720	89.752	1.793	1.670	1.953
	.750	72.212	53.724	108.082	1.859	1.730	2.034
	.800	85.442	62.498	133.502	1.932	1.796	2.125
	.850	103.951	74.220	171.530	2.017	1.871	2.234
	.900	133.039	91.682	236.304	2.124	1.962	2.373
	.910	141.208	96.420	255.481	2.150	1.984	2.407
	.920	150.651	101.820	278.147	2.178	2.008	2.444
	.930	161.766	108.080	305.477	2.209	2.034	2.485
	.940	175.151	115.494	339.285	2.243	2.063	2.531
	.950	191.774	124.530	382.562	2.283	2.095	2.583
	.960	213.330	135.999	440.689	2.329	2.134	2.644
	.970	243.180	151.477	524.661	2.386	2.180	2.720
	.980	289.422	174.682	662.056	2.462	2.242	2.821
	.990	380.797	218.382	956.548	2.581	2.339	2.981

a. Logarithm base = 10.

NON PARAMETRIC STATISTICS

Chi-Square Test for Association using SPSS

Chi-square statistic

- Two **non-parametric hypothesis tests** using the chi-square statistic:
  - **the chi-square test for goodness of fit and**
    - Goodness of fit refers to how close the observed data are to those predicted from a hypothesis
  - **the chi-square test for independence.**

Chi-Square Test for Association using SPSS

The chi-square test for independence, also called **Pearson's chi-square test or the chi-square test of association**, is used to discover if there is a **relationship between two categorical variables**.

Assumptions:

Assumption #1: Your **two variables** should be measured at an **ordinal** or **nominal level** (i.e., **categorical** data).

Assumption #2: Your two variable should consist of **two or more categorical**, **independent groups**. Example independent variables that meet this criterion include gender (2 groups: Males and Females), profession (e.g., 5 groups: surgeon, doctor, nurse, dentist, therapist), and so forth.

275

Example:

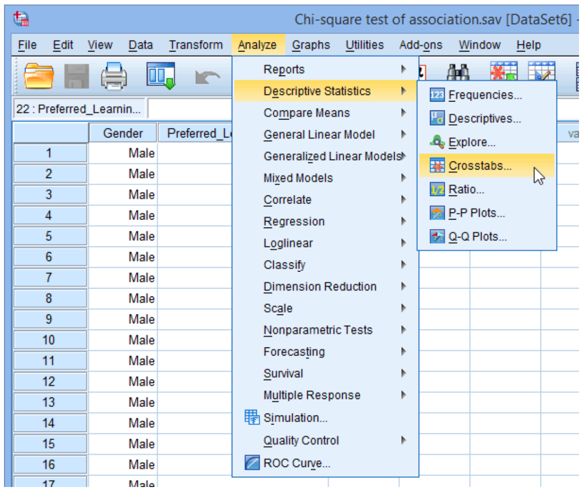
An educator would like to know whether **gender (male/female)** is associated with the preferred type of learning medium (online vs. books). We have two nominal variables: Gender (male/female) and Preferred Learning Medium (online/books).

276

In SPSS enter the two variables gender and learning medium

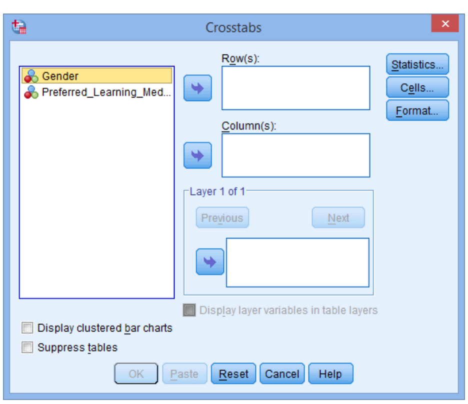
Gender in one column (code for male 1 & female 2) &

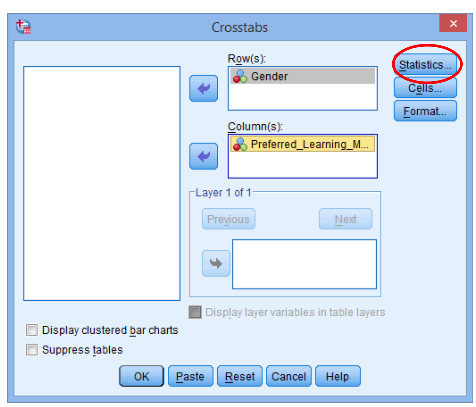
Preferred_Learning_Medium in another column (code for online 1 and books 2)



Click Analyze > Descriptives Statistics > Crosstabs... on the top menu, as shown below:

277



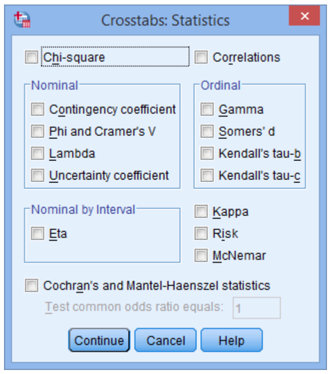


Transfer one of the variables into the Row(s): box and the other variable into the Column(s): box. In our example, we will transfer the Gender variable into the Row(s): box and Preferred_Learning_Medium into the Column(s): box.

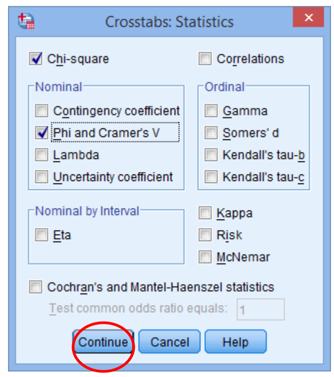
Then Click Statistics Button.

278

Click Statistics Button.



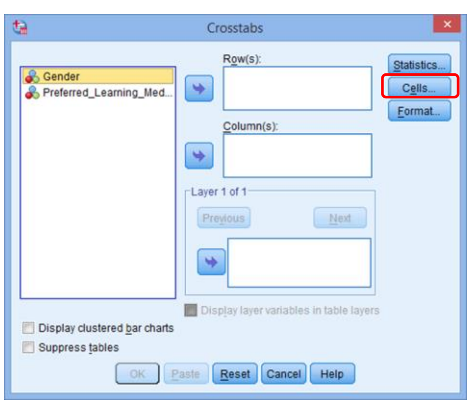
Select the Chi-square and Phi and Cramer's V options, as shown below:

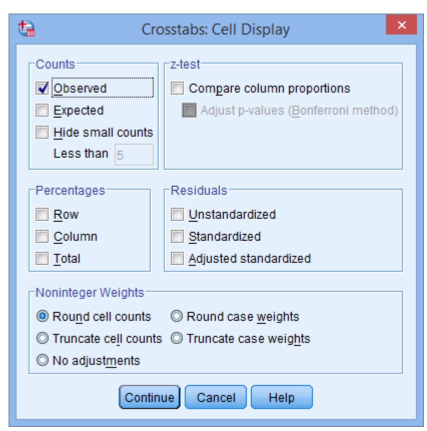


Click the Continue Button.

279

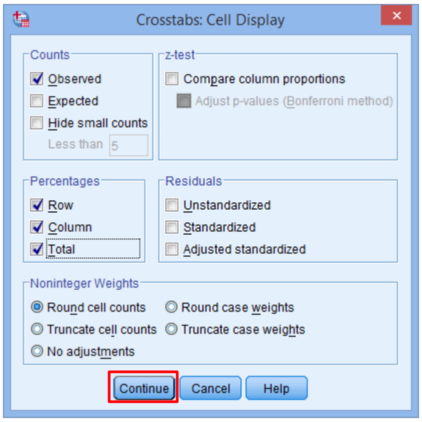
Click the Cells Button .





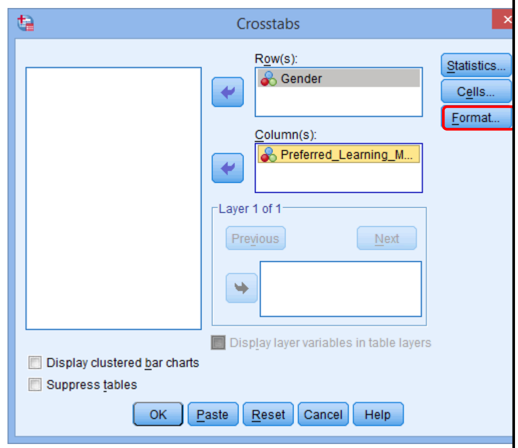
280

Select Observed from the – Counts – area, and Row, Column and Total from the – Percentages– area, as shown below:

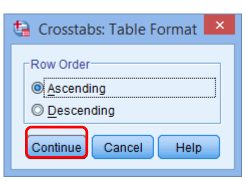


Click the Continue Button.

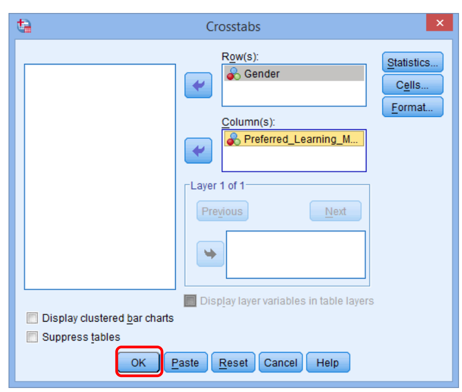
Then Click the Format Button.



281



This option allows you to change the order of the values to either ascending or descending.



Once you have made your choice, click the Continue Button.

Click the OK button to generate your output.

282

The Crosstabulation Table (Gender*Preferred Learning Medium Crosstabulation)

Gender * Preferred Learning Medium Crosstabulation

			Preferred Learning Medium		Total
			Books	Online	
Gender	Male	Count	16	24	40
		% within Gender	40.0%	60.0%	100.0%
		% within Preferred Learning Medium	55.2%	47.1%	50.0%
		% of Total	20.0%	30.0%	50.0%
	Female	Count	13	27	40
		% within Gender	32.5%	67.5%	100.0%
		% within Preferred Learning Medium	44.8%	52.9%	50.0%
		% of Total	16.3%	33.8%	50.0%
Total		Count	29	51	80
		% within Gender	36.3%	63.8%	100.0%
		% within Preferred Learning Medium	100.0%	100.0%	100.0%
		% of Total	36.3%	63.8%	100.0%

This table tells you both males and females prefer to learn using online materials versus books.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.487[a]	1	.485		
Continuity Correction[b]	.216	1	.642		
Likelihood Ratio	.487	1	.485		
Fisher's Exact Test				.642	.321
Linear-by-Linear Association	.481	1	.488		
N of Valid Cases	80				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 14.50.

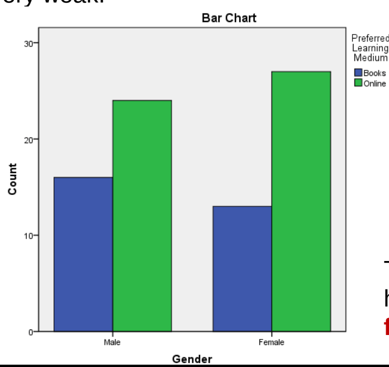
b. Computed only for a 2x2 table

Results of the "Pearson Chi-Square" row. We can see here that $\chi(1) = 0.487$, $p = .485$. This tells that there is no statistically significant association between Gender and Preferred Learning Medium; that is, both Males and Females equally prefer online learning versus books.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.078	.485
	Cramer's V	.078	.485
N of Valid Cases		80	

Phi and Cramer's V are both tests of the strength of association. We can see that the strength of association between the variables is very weak.



The clustered bar chart produced and highlights the **group categories and the frequency of counts** in these groups.

Chi-Square Goodness-of-Fit Test in SPSS

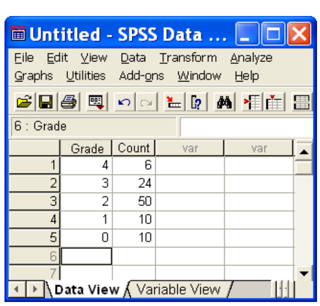
Chi-Square Goodness-of-Fit Test

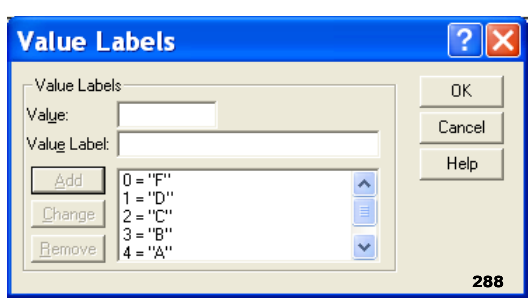
- The chi-square goodness-of-fit test is a single-sample nonparametric test, also referred to as the **one-sample goodness-of-fit test or Pearson's chi-square goodness-of-fit test.**
- It is used to determine whether the distribution of cases (e.g., participants) in a single categorical variable (e.g., "gender", consisting of two groups: "males" and "females") follows a known or hypothesised distribution

287

Chi-Square Goodness-of-Fit Test in SPSS

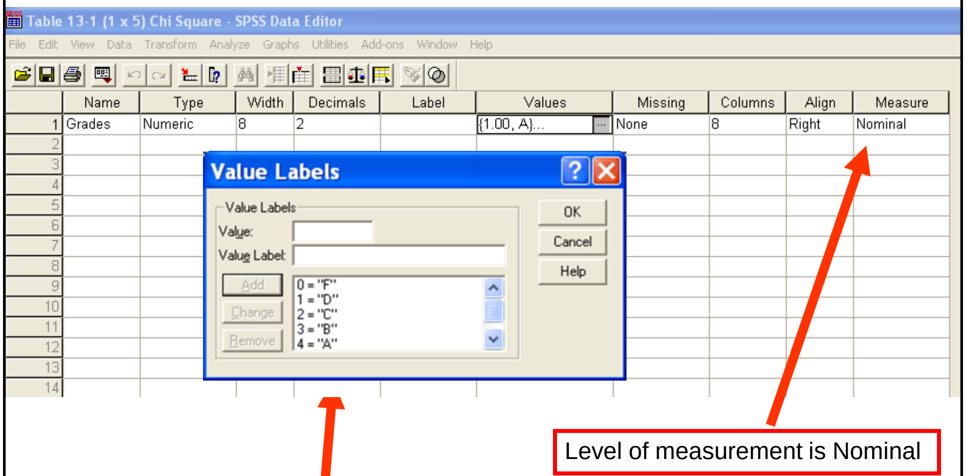
- Suppose we wish to test the null hypothesis that **Dr. Suresh** gives equal numbers of A's, B's, C's, D's, and F's as final grades in his Bioinformatics classes.
- The observed frequencies are: **A: 6, B: 24, C: 50, D: 10, F: 10.**
- The data are entered into SPSS like this:
- H0: Null hypothesis that the counts are uniformly distributed across the categories
- Ha: The counts are not equal





288

Chi Square SPSS Data Input



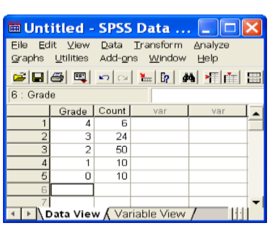
Level of measurement is Nominal

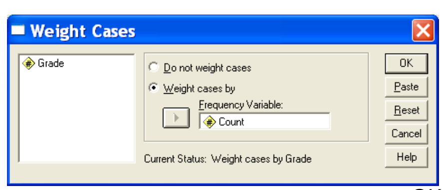
Enter Value Labels

289

Chi-Square Goodness-of-Fit Test in SPSS

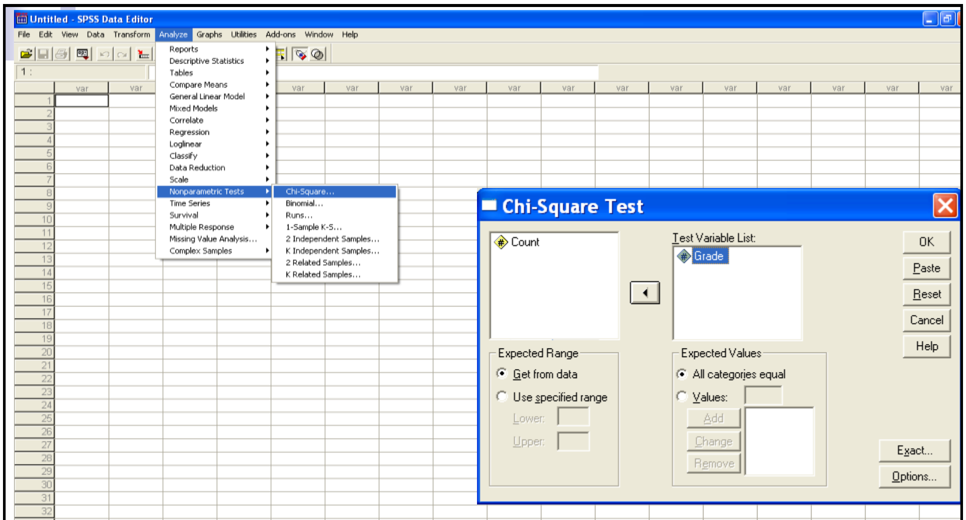
- Now tell SPSS to weight the cases by Count. **Click Data, Weight Cases, Weight Cases By Count.**





OK.

290



Click Analyze, Nonparametric Tests, Chi-square. Move the Grade into the Test Variable List. By default SPSS will use all categories and will test the hypothesis that the counts are, in the population, uniformly distributed across categories.

Click OK

291

Grade

	Observed N	Expected N	Residual
F	10	20.0	-10.0
D	10	20.0	-10.0
C	50	20.0	30.0
B	24	20.0	4.0
A	6	20.0	-14.0
Total	100		

Test Statistics

	Grade
Chi-Square[a]	65.600
df	4
Asymp. Sig.	.000

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 20.0.

We reject the null hypothesis that the counts are uniformly distributed across the categories, $\chi^2(4, N = 100) = 65.60$ $p < .001$.

292

Mann-Whitney Test in SPSS

Mann-Whitney U test

- The Mann-Whitney U test is used to compare differences between two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed.
  - For example, you could use the Mann-Whitney U test to understand whether salaries, measured on a continuous scale, differed based on educational level (i.e., your dependent variable would be "salary" and your independent variable would be "educational level", which has two groups: "high school" and "university").

Assumptions

- **Assumption #1:** Your **dependent variable** should be measured at the **ordinal** or **continuous level**. Examples of **ordinal variables** include Likert items (e.g., a 7-point scale from "strongly agree" through to "strongly disagree"). Examples of **continuous variables** include revision time (measured in hours),exam performance (measured from 0 to 100), weight (measured in kg).
- **Assumption #2:** Your **independent variable** should consist of **two categorical**, **independent groups**. Example gender (2 groups: male or female), employment status (2 groups: employed or unemployed), smoker (2 groups: yes or no), and so forth.
- **Assumption #3:** I**ndependence of observations**, which means that there is no relationship between the observations in each group or between the groups themselves.
  - For example, there must be different participants in each group with no participant being in more than one group.
- **Assumption #4:** A Mann-Whitney U test can be used when your two variables are **not normally distributed**

295

Example

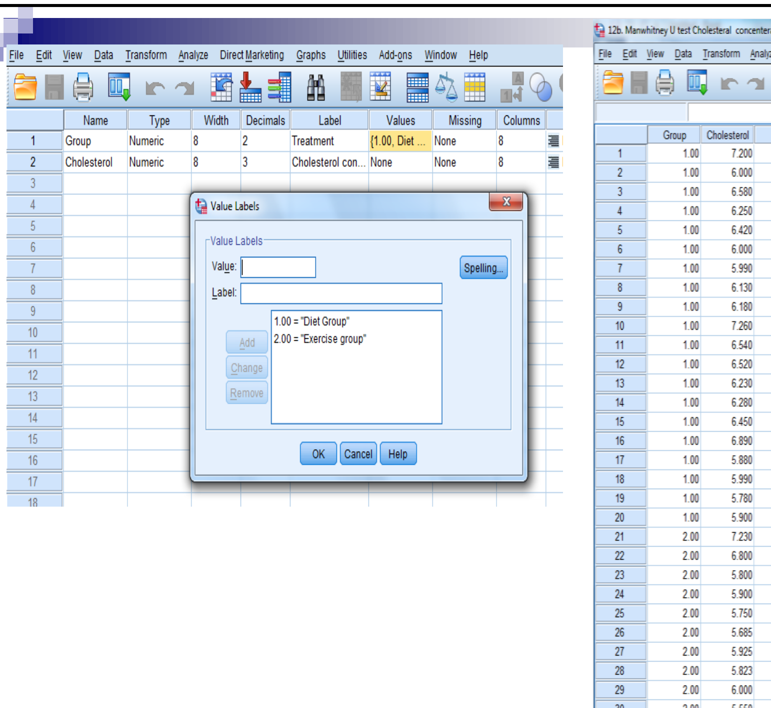
- The concentration of cholesterol in the blood is associated with the risk of developing heart disease, such that higher concentrations of cholesterol indicate a higher level of risk, and lower concentrations indicate a lower level of risk.
- If you lower the concentration of cholesterol in the blood, your risk for developing heart disease can be reduced. Being overweight and/or physically inactive increases the concentration of cholesterol in your blood. Both exercise and weight loss can reduce cholesterol concentration. However, it is not known whether exercise or weight loss is best for lowering cholesterol concentration.
- Therefore, a researcher decided to investigate whether an exercise or weight loss intervention was more effective in lowering cholesterol levels. **The researcher recruited a random sample of inactive males that were classified as overweight.** This sample was then randomly split into **two groups: Group 1 underwent a calorie-controlled diet (i.e., the 'diet' group) and Group 2 undertook an exercise-training programme (i.e., the 'exercise' group).**
- In order to determine which treatment programme was more effective, cholesterol concentrations were compared between the two groups at the end of the treatment programmes.

296

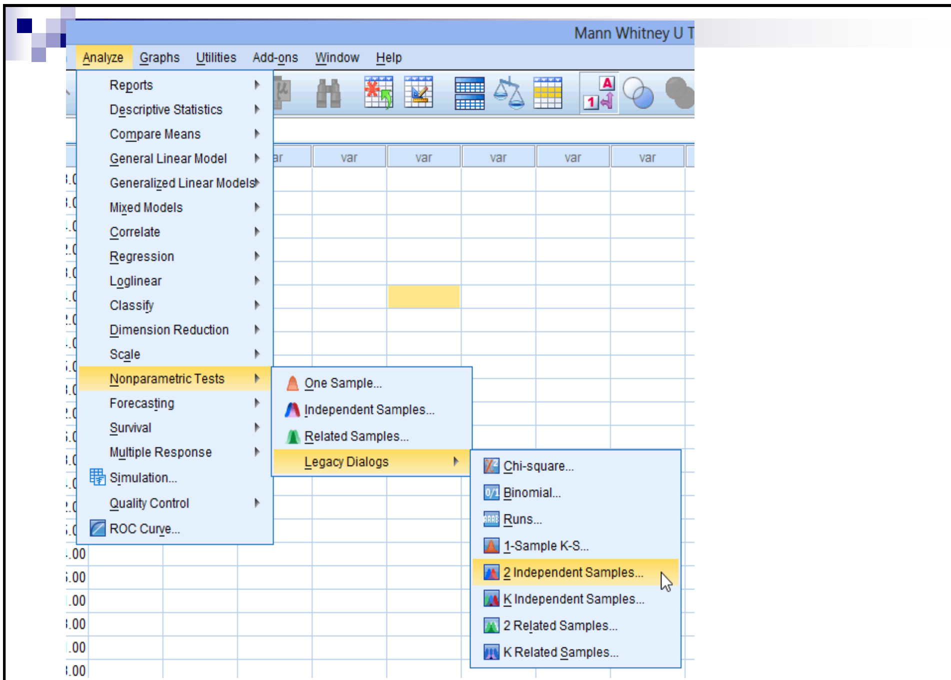
Mann-Whitney U test

- In SPSS Statistics, enter the scores for cholesterol concentration, our dependent variable, under the variable name Cholesterol.
- Next, create a grouping variable, called Group, which represented our independent variable.
- Since our independent variable had two groups - 'diet' and 'exercise' – **Give the diet group a value of "1" and the exercise group a value of "2".**

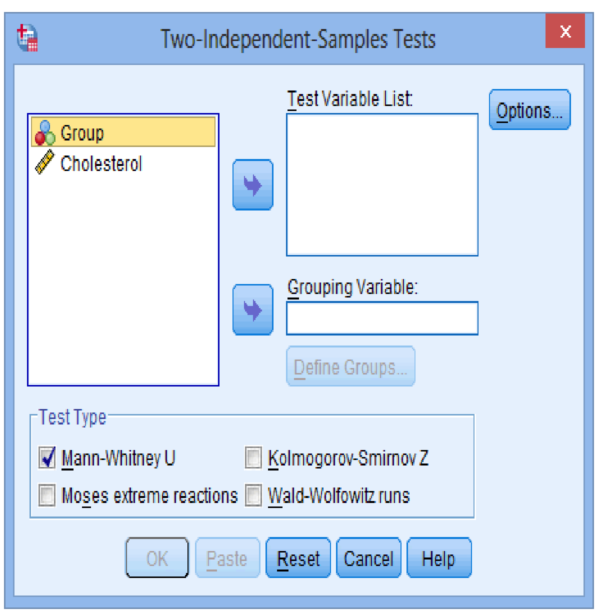
297



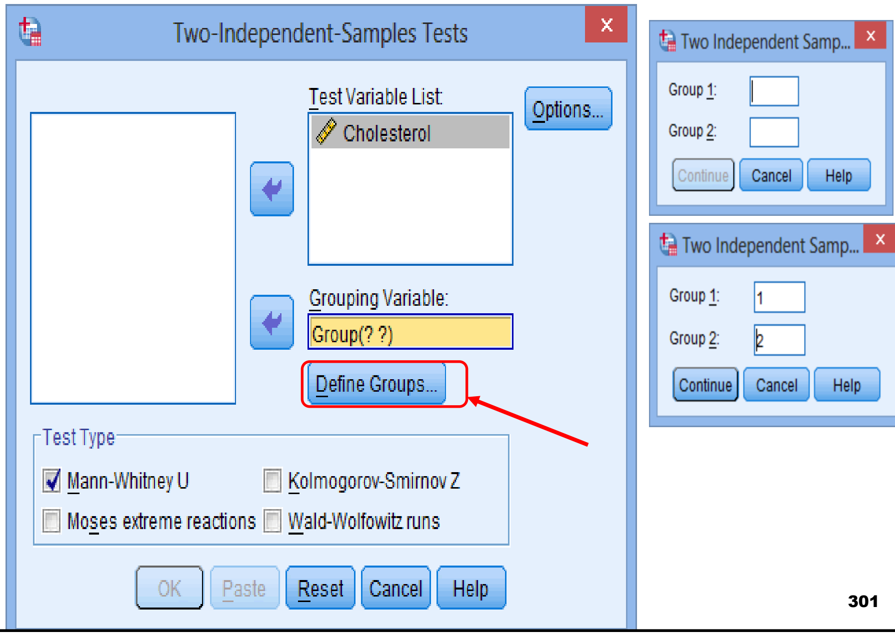
298


Mann Whitney U T
Analyze
Graphs
Utilities
Add-ons
Window
Help
Reports
Descriptive Statistics
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Simulation...
Quality Control
ROC Curve...
One Sample...
Independent Samples...
Related Samples...
Legacy Dialogs
Chi-square...
Binomial...
Runs...
1-Sample K-S...
2 Independent Samples...
K Independent Samples...
2 Related Samples...
K Related Samples...

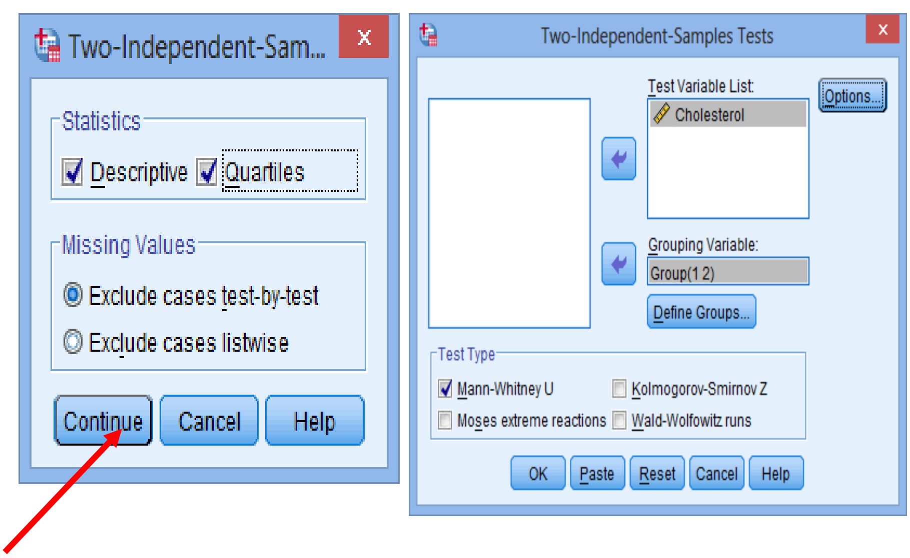
299


Two-Independent-Samples Tests
Group
Cholesterol
Test Variable List:
Options...
Grouping Variable:
Define Groups...
Test Type
Mann-Whitney U
Kolmogorov-Smirnov Z
Moses extreme reactions
Wald-Wolfowitz runs
OK
Paste
Reset
Cancel
Help

300


Two-Independent-Samples Tests
Test Variable List:
Cholesterol
Options...
Grouping Variable:
Group(? ?)
Define Groups...
Test Type
Mann-Whitney U
Kolmogorov-Smirnov Z
Moses extreme reactions
Wald-Wolfowitz runs
OK
Paste
Reset
Cancel
Help
Two Independent Samp...
Group 1:
Group 2:
Continue
Cancel
Help
Two Independent Samp...
Group 1: 1
Group 2: 2
Continue
Cancel
Help

301


Two-Independent-Sam...
Statistics
Descriptive
Quartiles
Missing Values
Exclude cases test-by-test
Exclude cases listwise
Continue
Cancel
Help
Two-Independent-Samples Tests
Test Variable List:
Cholesterol
Options...
Grouping Variable:
Group(1 2)
Define Groups...
Test Type
Mann-Whitney U
Kolmogorov-Smirnov Z
Moses extreme reactions
Wald-Wolfowitz runs
OK
Paste
Reset
Cancel
Help

302

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
Cholesterol Concentration	40	5.9700	.48368	5.20	7.70	5.6250	5.9000	6.2000
Group	40	1.50	.506	1	2	1.00	1.50	2.00

Mann-Whitney Test

Ranks

	Group	N	Mean Rank	Sum of Ranks
Cholesterol Concentration	Diet	20	25.00	500.00
	Exercise	20	16.00	320.00
	Total	40		

303

Test Statistics[b]

	Cholesterol Concentration
Mann-Whitney U	110.000
Wilcoxon W	320.000
Z	-2.446
Asymp. Sig. (2-tailed)	.014
Exact Sig. [2*(1-tailed Sig.)]	.014[a]

a. Not corrected for ties.

b. Grouping Variable: Group

Interpretation:

It can be concluded that cholesterol concentration in the diet group was statistically significantly higher than the exercise group (U = 110, p = .014). Depending on the size of your groups

304

Wilcoxon Signed Rank Test – two related samples

Wilcoxon signed-ranks test

- The Wilcoxon signed-rank test is the nonparametric test equivalent to the dependent t-test (paired sample t test)
- Wilcoxon signed-rank test does not assume normality in the data, it can be used when this assumption has been violated and the use of the dependent t-test is inappropriate.
- It is used to compare two sets of scores that come from the same participants
- **To test difference between paired data**

Wilcoxon signed-ranks test - Assumptions

Assumption #1: D**ependent variable** should be measured at the **ordinal** or **continuous level**.

Assumption #2: I**ndependent variable** should consist of **two categorical**, **"related groups"** or **"matched pairs"**. *"Related groups" indicates that the same subjects are present in both groups. Same subjects in each group is because each subject has been measured on two occasions on the same dependent variable.*

Assumption #3: The **distribution of the differences between the two related groups** needs to be **symmetrical in shape**

307

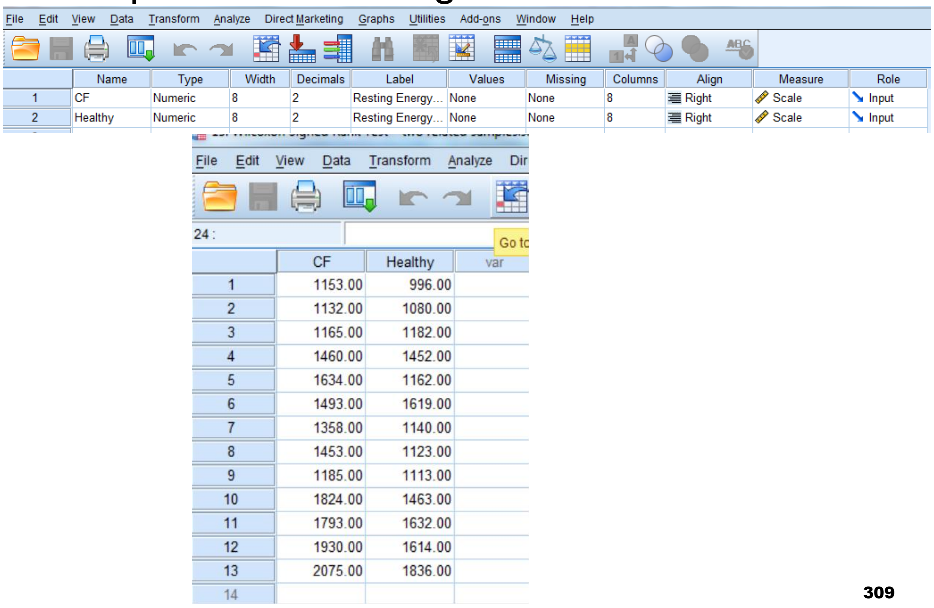
Wilcoxon Signed Rank Test – two related samples – paired samples

Resting Energy Expenditure (REE) for Patient with Cystic Fibrosis

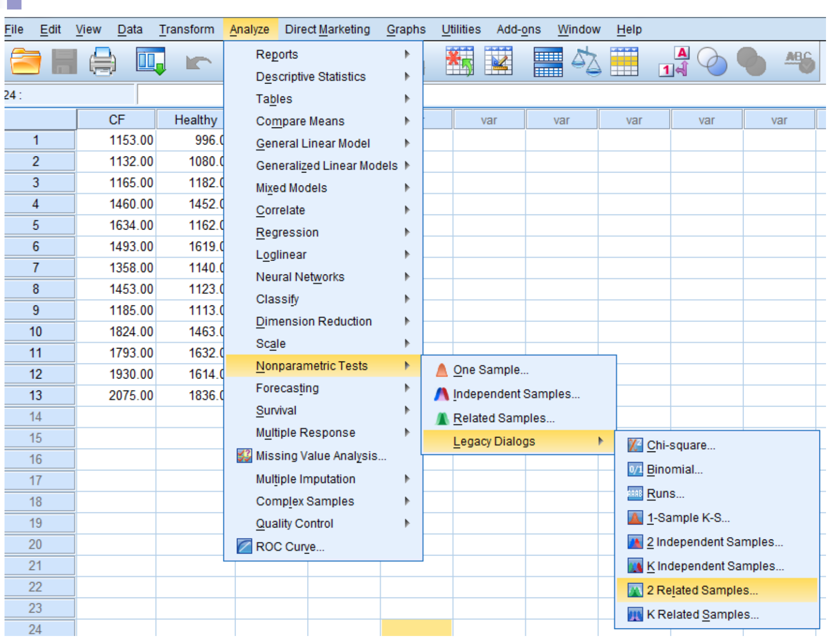
- A researcher believes that patients with cystic fibrosis (CF) expend greater energy during resting than those without CF.
- To obtain a fair comparison matched 13 patients with CF to 13 patients without CF.

308

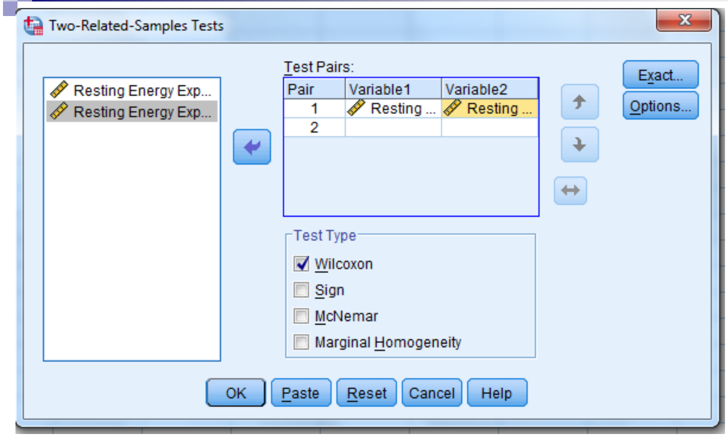
Example: Wilcoxon Signed Rank Test



309



310



311

Ranks

		N	Mean Rank	Sum of Ranks
Resting Energy Expenditure (REE) for NORMAL Patient - Resting Energy Expenditure (REE) for Patient with Cystic Fibrosis	Negative Ranks	11[a]	7.64	84.00
	Positive Ranks	2[b]	3.50	7.00
	Ties	0[c]		
	Total	13		

a. Resting Energy Expenditure (REE) for NORMAL Patient < Resting Energy Expenditure (REE) for Patient with Cystic Fibrosis

b. Resting Energy Expenditure (REE) for NORMAL Patient > Resting Energy Expenditure (REE) for Patient with Cystic Fibrosis

c. Resting Energy Expenditure (REE) for NORMAL Patient = Resting Energy Expenditure (REE) for Patient with Cystic Fibrosis

Test Statistics[a]

	Resting Energy Expenditure (REE) for NORMAL Patient - Resting Energy Expenditure (REE) for Patient with Cystic Fibrosis
Z	-2.691[b]
Asymp. Sig. (2-tailed)	.007

a. Wilcoxon Signed Ranks Test

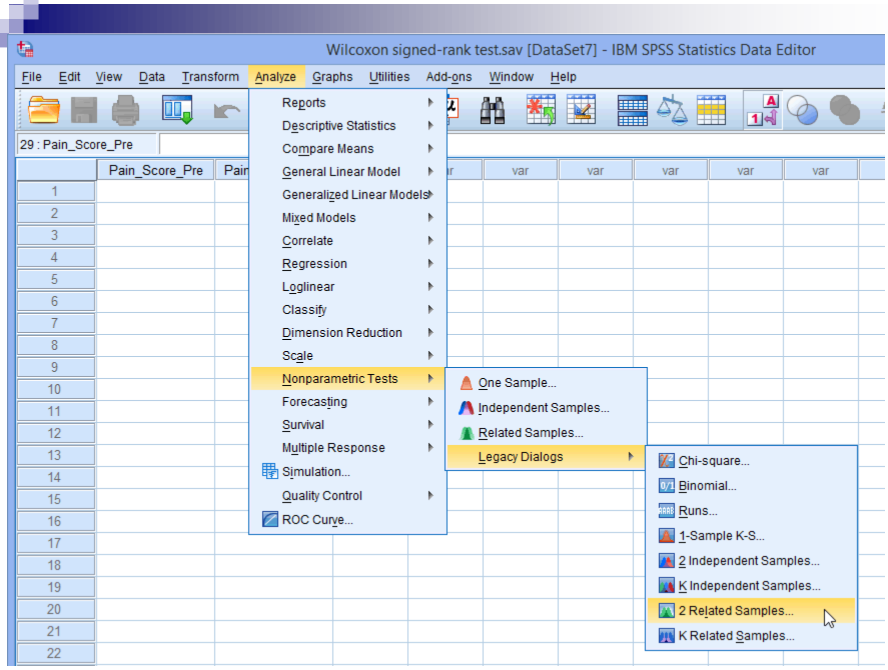
b. Based on positive ranks.

312

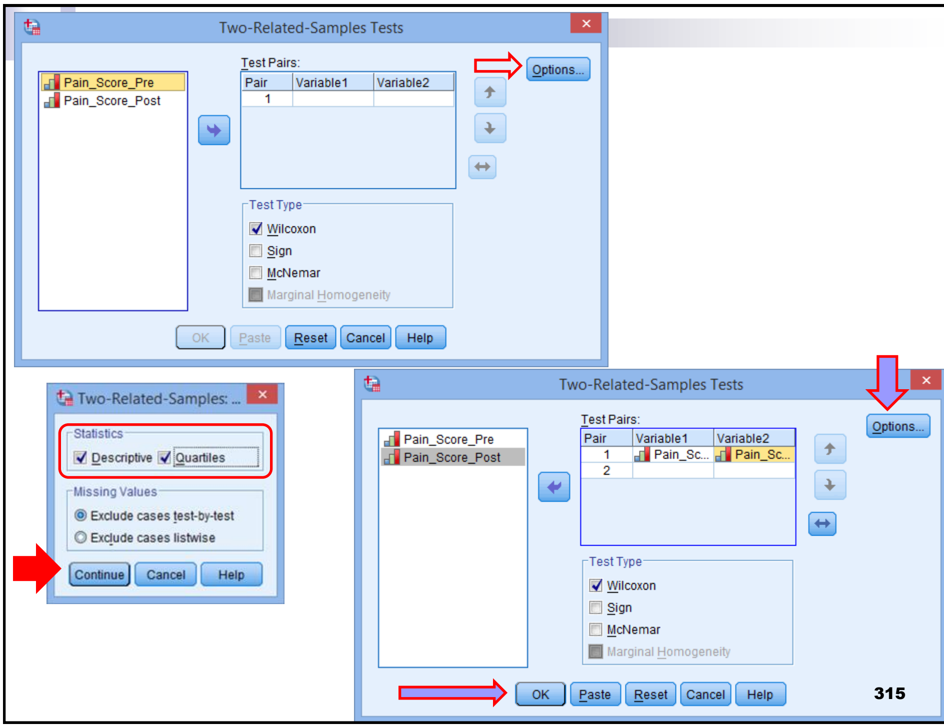
Example

- A researcher is interested in finding methods to reduce lower back pain in individuals without having to use drugs.
- The researcher thinks that having acupuncture in the lower back might reduce back pain.
- To investigate this, the researcher recruits 25 participants to their study. At the beginning of the study, the researcher asks the participants to rate their back pain on a scale of 1 to 10, with 10 indicating the greatest level of pain.
- After 4 weeks of twice weekly acupuncture, the participants are asked again to indicate their level of back pain on a scale of 1 to 10, with 10 indicating the greatest level of pain.
- The researcher wishes to understand whether the participants' pain levels changed after they had undergone the acupuncture, so a Wilcoxon signed-rank test is run.

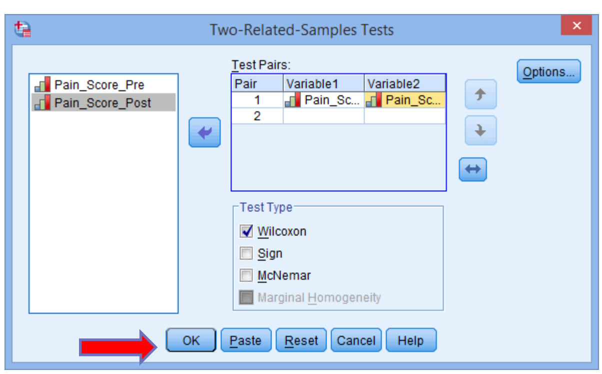
313



314


Two-Related-Samples Tests
Pain_Score_Pre
Pain_Score_Post
Test Pairs:
Pair
Variable1
Variable2
Options...
Test Type
Wilcoxon
Sign
McNemar
Marginal Homogeneity
OK
Paste
Reset
Cancel
Help
Two-Related-Samples: ...
Statistics
Descriptive
Quartiles
Missing Values
Exclude cases test-by-test
Exclude cases listwise
Continue
Cancel
Help

315


Two-Related-Samples Tests
Pain_Score_Pre
Pain_Score_Post
Test Pairs:
Pair
Variable1
Variable2
Pain_Sc...
Pain_Sc...
Options...
Test Type
Wilcoxon
Sign
McNemar
Marginal Homogeneity
OK
Paste
Reset
Cancel
Help

316

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
Pain_Score_Pre	25	5.4400	1.78139	2.00	9.00	4.0000	5.0000	6.5000
Pain_Score_Post	25	5.1600	1.57268	2.00	8.00	4.0000	5.0000	6.0000

Ranks

		N	Mean Rank	Sum of Ranks
Pain_Score_Post - Pain_Score_Pre	Negative Ranks	11[a]	8.00	88.00
	Positive Ranks	4[b]	8.00	32.00
	Ties	10[c]		
	Total	25		

a. Pain_Score_Post < Pain_Score_Pre

b. Pain_Score_Post > Pain_Score_Pre

c. Pain_Score_Post = Pain_Score_Pre

Test Statistics[b]

	Pain_Score_Post - Pain_Score_Pre
Z	-1.807[a]
Asymp. Sig. (2-tailed)	.071

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

The results showed that a 4 week, twice weekly acupuncture treatment course did not reduce a statistically significant change in lower back pain in individuals with existing lower back pain (Z = -1.807, p = 0.071). The, median Pain Score rating is 5.0 in both pre- and post-treatment.

317

McNemar's test using SPSS

McNemar's test using SPSS

- This test is used to determine if there are differences on a **dichotomous dependent variable between two related groups**.
- It can be considered to be similar to the paired-samples t-test, but for a dichotomous rather than a continuous dependent variable

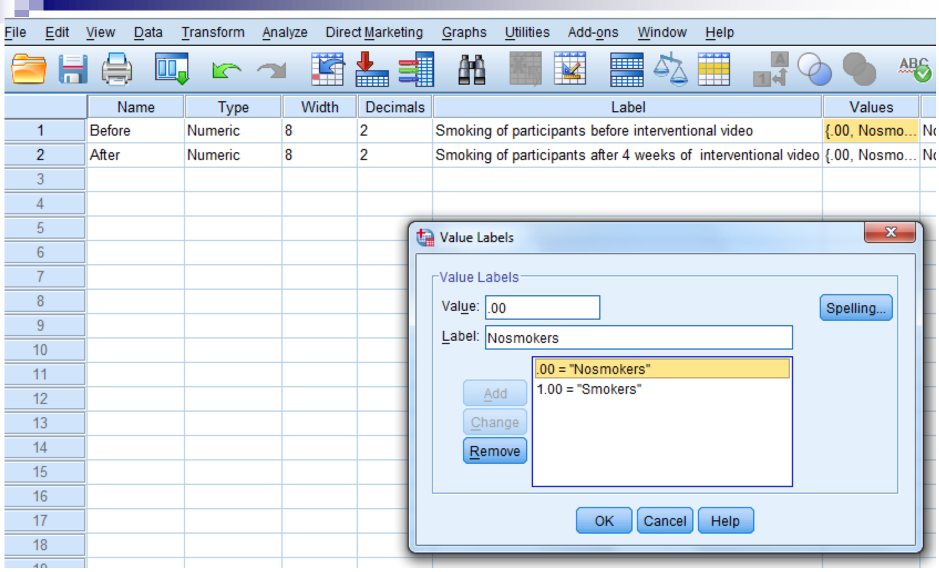
Assumptions

- **Assumption #1:** Will have **one categorical dependent variable with two categories** (i.e., a **dichotomous** variable) and one **categorical independent variable with two related groups**. Examples of dichotomous variables include exam performance (two groups: "pass" and "fail"), preferred choice of cereal brand (two groups: "brand A" and "brand B")
- **Assumption #2:** The two groups of your dependent variable must be **mutually exclusive**. This means that no groups can overlap. In other words, a participant can only be in one of the two groups; they cannot be in both groups at the same time.
- **Assumption #3:** The cases (e.g., participants) are a random sample from the population of interest.

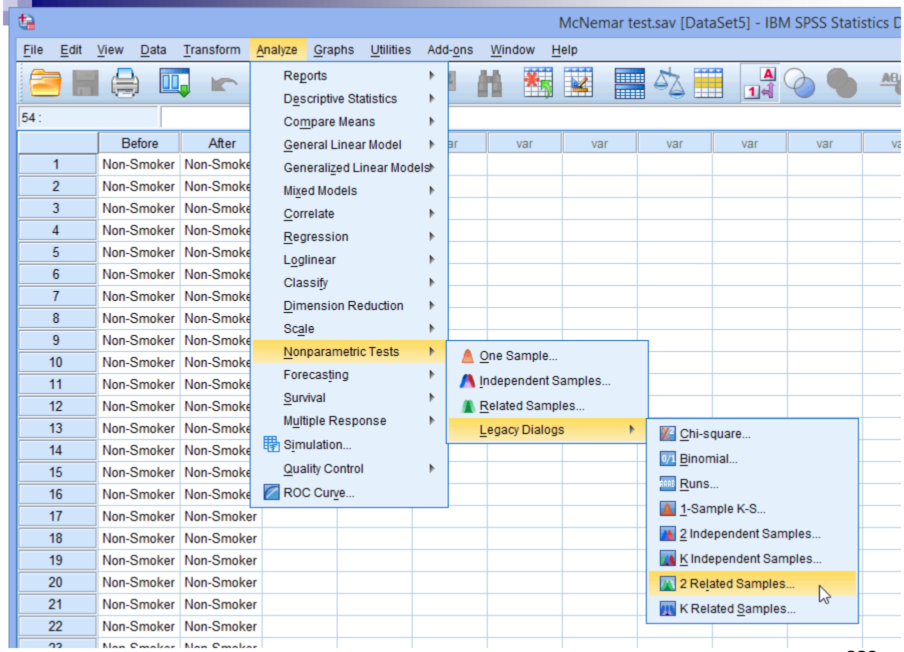
Example

- A researcher wanted to investigate the impact of an intervention on smoking. In this study, 50 participants were recruited to take part, consisting of 25 smokers and 25 non-smokers.
- All participants watched an emotive video showing the impact of health, consequences and deaths from smoking-related cancers had on families. After four weeks after this video intervention, the same participants were asked whether they remained smokers or non-smokers.
- Therefore, participants were categorized as being either smokers or non-smokers before the intervention and then re-assessed as either smokers or non-smokers after the intervention.

321

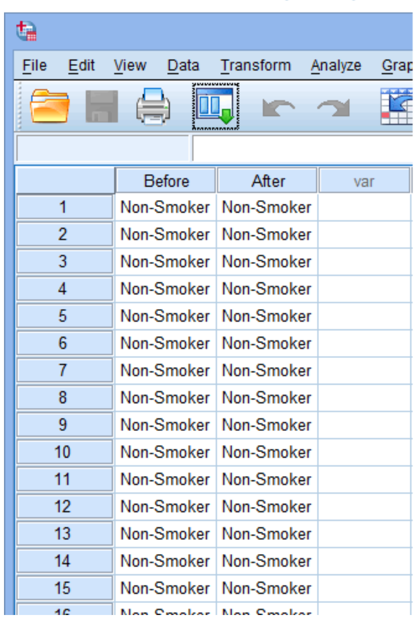


322



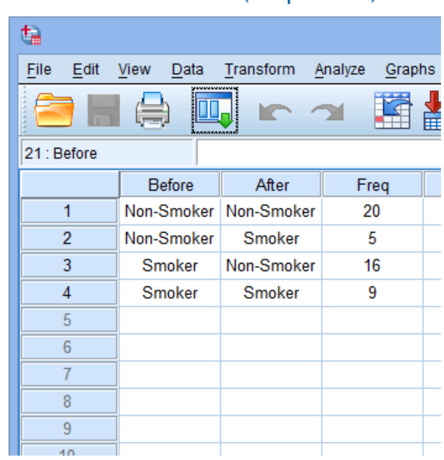
323

Individual scores for each participant



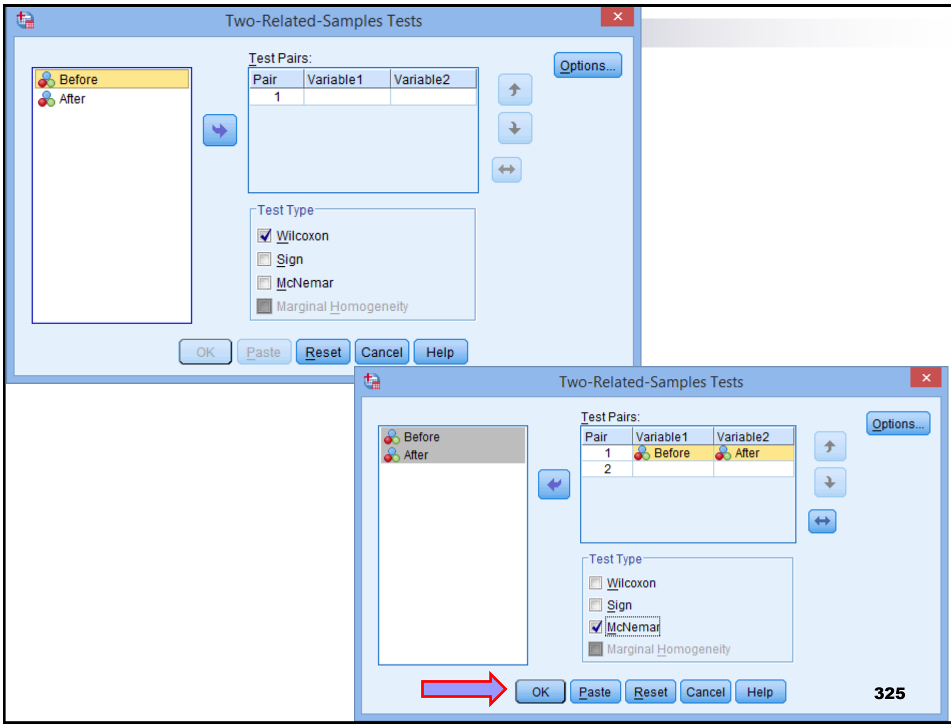
	Before	After
1	Non-Smoker	Non-Smoker
2	Non-Smoker	Non-Smoker
3	Non-Smoker	Non-Smoker
4	Non-Smoker	Non-Smoker
5	Non-Smoker	Non-Smoker
6	Non-Smoker	Non-Smoker
7	Non-Smoker	Non-Smoker
8	Non-Smoker	Non-Smoker
9	Non-Smoker	Non-Smoker
10	Non-Smoker	Non-Smoker
11	Non-Smoker	Non-Smoker
12	Non-Smoker	Non-Smoker
13	Non-Smoker	Non-Smoker
14	Non-Smoker	Non-Smoker
15	Non-Smoker	Non-Smoker

Total count data (frequencies)



	Before	After	Freq
1	Non-Smoker	Non-Smoker	20
2	Non-Smoker	Smoker	5
3	Smoker	Non-Smoker	16
4	Smoker	Smoker	9

324



325

Before & After

Before	After: Non-Smoker	After: Smoker
Non-Smoker	20	5
Smoker	16	9

Test Statistics[a]

	Before & After
N	50
Exact Sig. (2-tailed)	.027[b]

a. McNemar Test

b. Binomial distribution used.

- Fifty participants were recruited to take part in an intervention designed to warn about the dangers of smoking. An exact McNemar's test determined that there was a statistically significant difference in the proportion of non-smokers pre- and post-intervention, *p* = .027.

326

Kruskal-Wallis H Test using SPSS

Kruskal-Wallis H Test using SPSS

- The Kruskal-Wallis H test is a rank-based nonparametric test that can be used to determine if there are statistically significant differences between two or more groups of an independent variable on a **continuous or ordinal dependent variable**.
- It is considered the nonparametric alternative to the **one-way ANOVA,** and an extension of the Mann-Whitney U test to allow the comparison of more than two independent groups.
- Example, to understand whether exam performance, measured on a continuous scale from 0-100, differed based on test anxiety levels (i.e., your dependent variable would be **"exam performance"** and your independent variable would be "test anxiety level", which has three independent groups: students with "low", "medium" and "high" test anxiety levels).

Assumptions

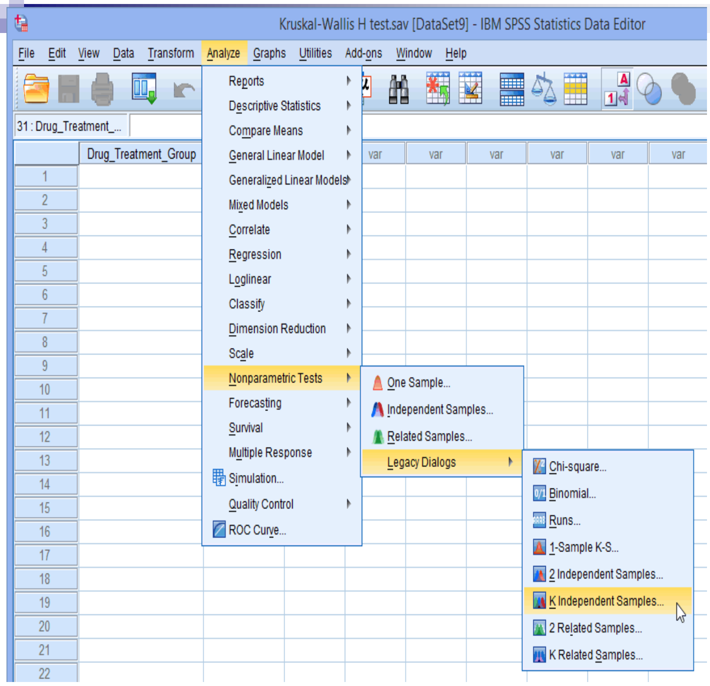
- **Assumption #1:** D**ependent variable** should be measured at the **ordinal** or **continuous level** (i.e., **interval** or **ratio**).
- **Assumption #2:** I**ndependent variable** should consist of **two or more categorical**, **independent groups**. Kruskal-Wallis H test is used when you have **three or more** categorical, independent groups.
  - **Example physical activity level (e.g., four groups: sedentary, low, moderate and high), profession (e.g., five groups: surgeon, doctor, nurse, dentist, therapist),**
- **Assumption #3:** You should have **independence of observations**, which means that there is no relationship between the observations in each group or between the groups themselves.

329

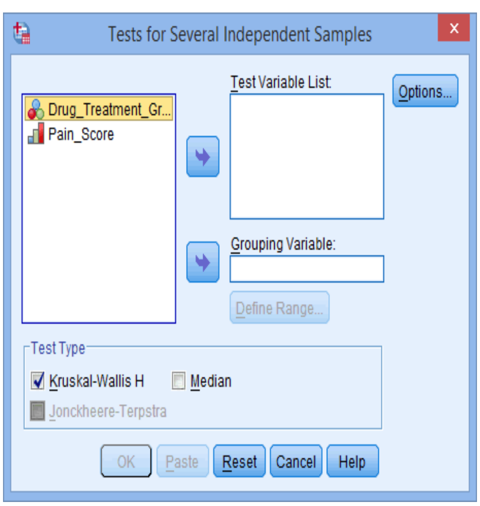
Example

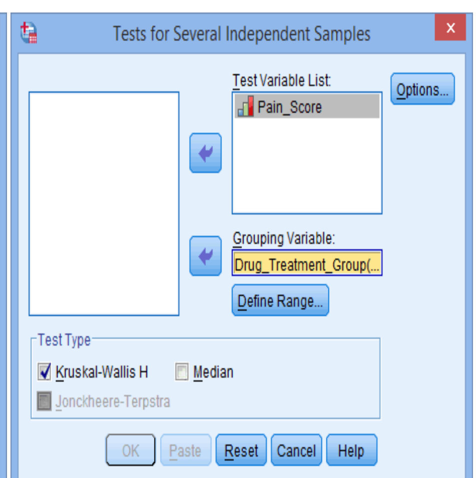
- A medical researcher has heard anecdotal evidence that certain anti-depressive drugs can have the positive side effect of lowering neurological pain in those individuals with chronic, neurological back pain, when administered in doses lower than those prescribed for depression.
- The medical researcher would like to investigate this anecdotal evidence with a study. The researcher identifies **3 well-known, anti-depressive drugs** which might have this positive side effect, and labels them Drug A, Drug B and Drug C.
- The researcher then recruits a group of 60 individuals with a similar level of back pain and randomly assigns them to one of three groups – Drug A, Drug B or Drug C treatment groups – and prescribes the relevant drug for a 4 week period. At the end of the 4 week period, the researcher asks the participants to rate their back pain on a scale of 1 to 10, with 10 indicating the greatest level of pain. **The researcher wants to compare the levels of pain experienced by the different groups at the end of the drug treatment period.**

330

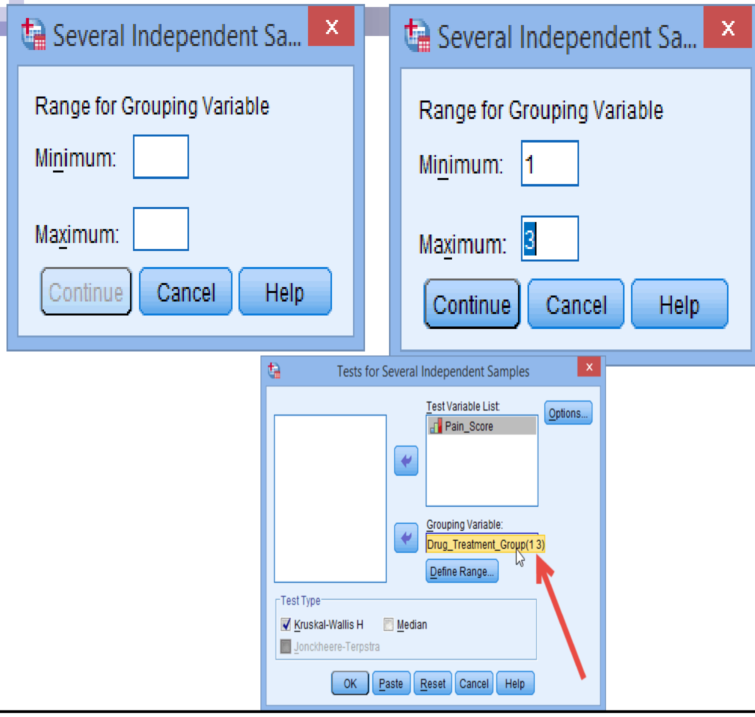

Kruskal-Wallis H test.sav [DataSet9] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
31 : Drug_Treatment_...
Drug_Treatment_Group
var
Reports
Descriptive Statistics
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Simulation...
Quality Control
ROC Curve...
One Sample...
Independent Samples...
Related Samples...
Legacy Dialogs
Chi-square...
Binomial...
Runs...
1-Sample K-S...
2 Independent Samples...
K Independent Samples...
2 Related Samples...
K Related Samples...

331


Tests for Several Independent Samples
Drug_Treatment_Gr...
Pain_Score
Test Variable List:
Options...
Grouping Variable:
Define Range...
Test Type
Kruskal-Wallis H
Median
Jonckheere-Terpstra
OK
Paste
Reset
Cancel
Help

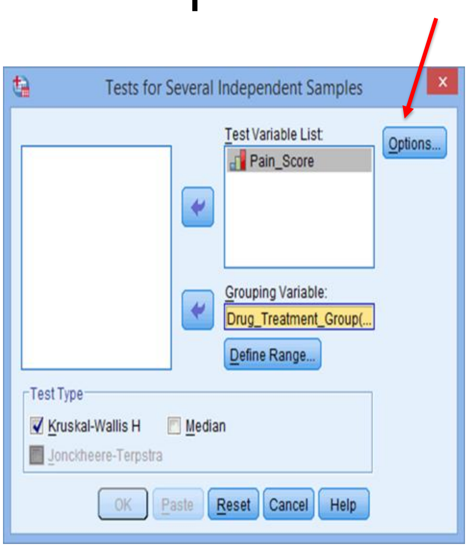

Tests for Several Independent Samples
Test Variable List:
Pain_Score
Options...
Grouping Variable:
Drug_Treatment_Group(...
Define Range...
Test Type
Kruskal-Wallis H
Median
Jonckheere-Terpstra
OK
Paste
Reset
Cancel
Help

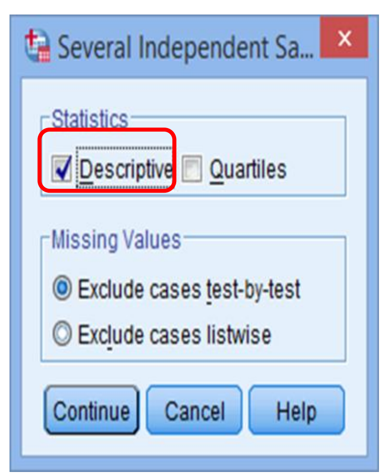
332



333

Click options





334

Kruskal-Wallis Test

Ranks

	Drug Treatment Group	N	Mean Rank
Pain_Score	Drug A	20	35.33
	Drug B	20	34.83
	Drug C	20	21.35
	Total	60	

Test Statistics[a,b]

	Pain_Score
Chi-square	8.520
df	2
Asymp. Sig.	.014

a. Kruskal Wallis Test

b. Grouping Variable: Drug Treatment Group

Interpretation:

A Kruskal-Wallis test showed that there was a statistically significant difference in pain score between the different drug treatments, $\chi^2(2) = 8.520$, $p = 0.014$, with a mean rank pain score of 35.33 for Drug A, 34.83 for Drug B and 21.35 for Drug C.

335

Friedman Test using SPSS

Friedman test

- The Friedman test is the non-parametric alternative to the **one-way ANOVA with repeated measures**.
  - It is used to test for differences between groups when the dependent variable being measured is ordinal.
  - It can also be used for continuous data that has violated the assumptions necessary to run the one-way ANOVA with repeated measures

337

Assumptions

- **Assumption #1: One group** that is measured on **three or more different occasions**.
- **Assumption #2:** Group is a random sample from the population.
- **Assumption #3:** Your **dependent variable** should be measured at the **ordinal** or **continuous level**.
- **Assumption #4:** Samples do **NOT need to be normally distributed**.

338

Example

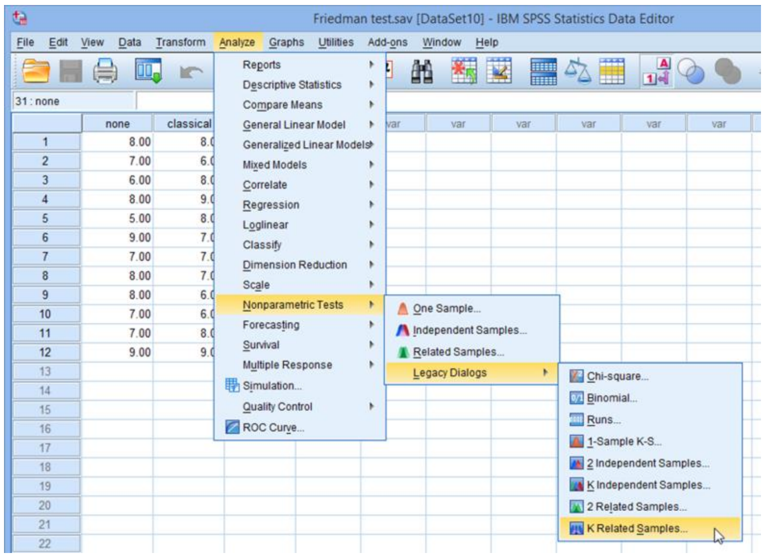
- A researcher wants to examine whether music has an effect on the perceived psychological effort required to perform an exercise session.
- The dependent variable is **"perceived effort to perform exercise"** and the independent variable is **"music type",** which consists of three groups: **"no music", "classical music" and "dance music".**
- To test whether music has an effect on the perceived psychological effort required to perform an exercise session, the researcher recruited 12 runners who each ran three times on a treadmill for 30 minutes.
- For consistency, the treadmill speed was the same for all three runs. In a random order, each subject ran: **(a) listening to no music at all; (b) listening to classical music; and (c) listening to dance music.**
- At the end of each run, subjects were asked to record how hard the running session felt on a scale of 1 to 10, with 1 being easy and 10 extremely hard

339

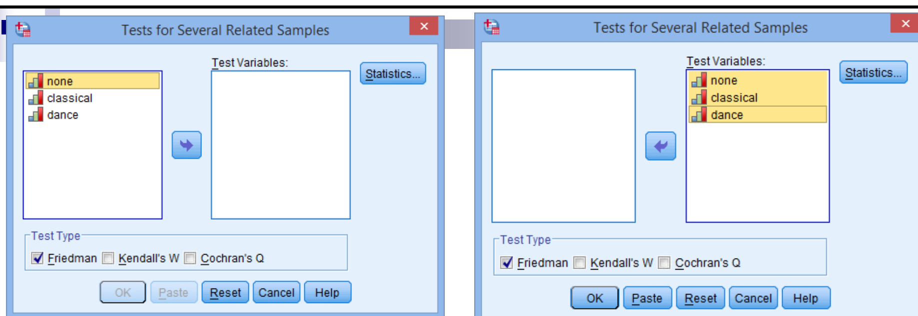
	none	classical	dance
1	8.00	8.00	7.00
2	7.00	6.00	6.00
3	6.00	8.00	6.00
4	8.00	9.00	7.00
5	5.00	8.00	5.00
6	9.00	7.00	7.00
7	7.00	7.00	7.00
8	8.00	7.00	7.00
9	8.00	6.00	8.00
10	7.00	6.00	6.00
11	7.00	8.00	6.00
12	9.00	9.00	6.00

340

Click Analyze > Nonparametric Tests > Legacy Dialogs > K Related Samples

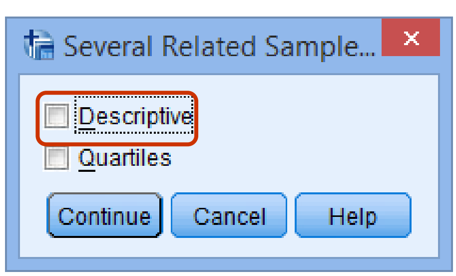


341

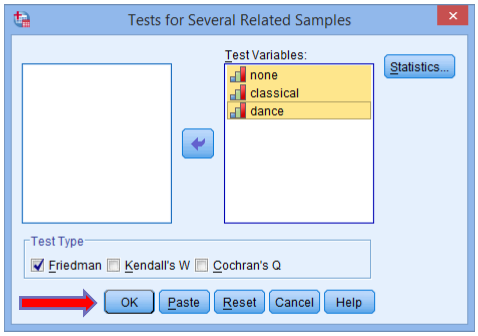


Transfer the dependent variables **none, classical and dance** to the Test Variables:

Click Friedman
Click Statistics



342



Ranks

	Mean Rank
none	2.38
classical	2.17
dance	1.46

Test Statistics[a]

N	12
Chi-Square	7.600
df	2
Asymp. Sig.	.022

a. Friedman Test

We can see that there is an overall statistically significant difference between the mean ranks of the related groups.

Reporting:

There was a statistically significant difference in perceived effort depending on which type of music was listened to whilst running, $\chi^2(2) = 7.600$, $p = 0.022$.

343

Reliability
Cronbach's Alpha (α) using SPSS

Cronbach's Alpha (α) using SPSS

- Cronbach's alpha is the most common measure of internal consistency ("reliability").
- It is most commonly used when you have multiple Likert questions in a survey/questionnaire that form a scale and you wish to determine if the **scale is reliable**.

345

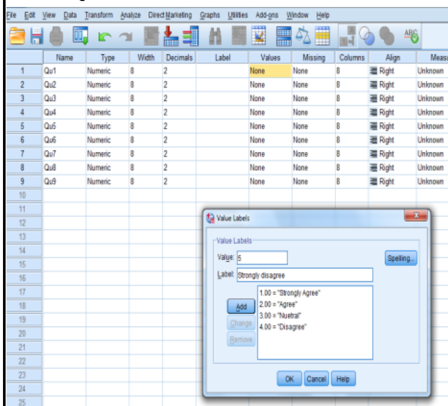
Cronbach's Alpha (α) using SPSS

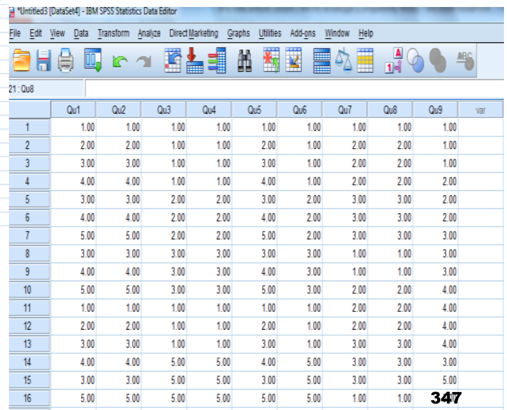
Example

A researcher has devised a nine-question questionnaire to measure how safe people feel at work at an industrial complex.
Each question was a 5-point Likert item **from "strongly disagree" to "strongly agree".**

346

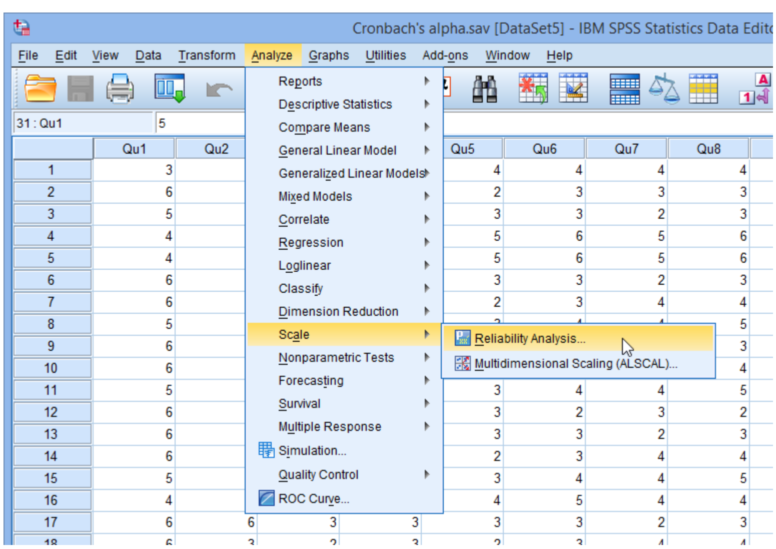
Enter the results of the nine questions. Questions label as Qu1 through to Qu9.





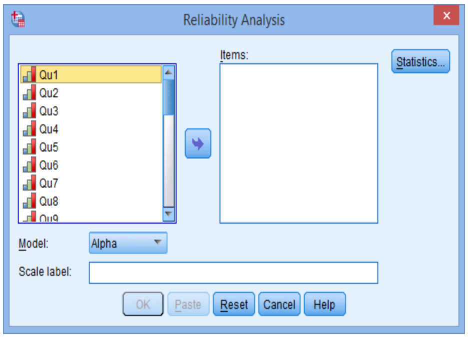
347

Click Analyze > Scale > Reliability Analysis... on the top menu, as shown below:

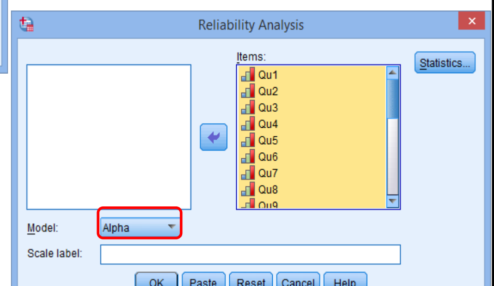


348

You will be presented with the Reliability Analysis dialogue box, as shown below:

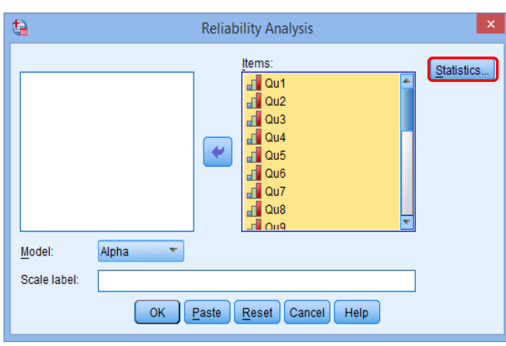


Transfer the variables Qu1 to Qu9 into the Items: box.

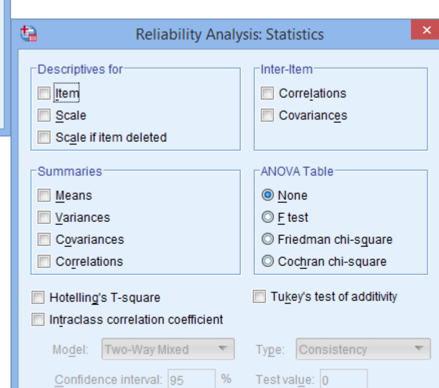


Leave the Model: set as "Alpha", which represents Cronbach's alpha in SPSS.

349

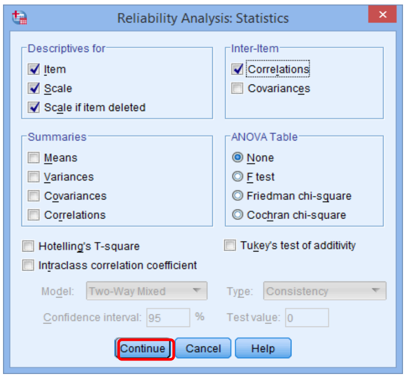


Click on the **Statistics** Button, which will open the Reliability Analysis: Statistics dialogue box

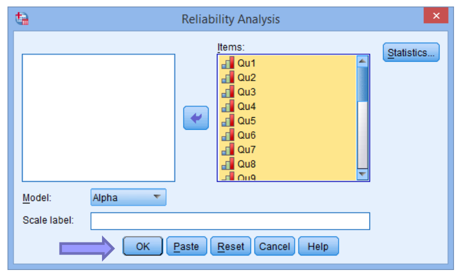


350

Select the Item, Scale and Scale if item deleted options in the –Descriptives for– area, and the Correlations option in the –Inter-Item– area, as shown below:



Click Continue. This will return you to the Reliability Analysis dialogue box.



Click OK to generate the output.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.805	.796	9

Cronbach's alpha is 0.805, which indicates a high level of internal consistency for our scale.

Cronbach's alpha value is 0.70 and above is considered good.

The Item-Total Statistics table presents the "Cronbach's Alpha if Item Delet

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Qu1	24.20	45.029	.633	.588	.767
Qu2	23.93	47.352	.520	.651	.783
Qu3	24.07	46.638	.654	.899	.767
Qu4	23.40	47.114	.551	.823	.779
Qu5	23.60	51.257	.389	.573	.799
Qu6	24.47	50.695	.372	.693	.802
Qu7	24.07	45.210	.615	.777	.770
Qu8	24.20	56.457	.128	.791	.823
Qu9	24.07	45.210	.589	.610	.774

Each column presents the value that Cronbach's alpha would be if that particular item was deleted from the scale. **Except question 8,** would result in a lower Cronbach's alpha.

Multiple Regression Analysis using SPSS

Multiple regression

- Multiple regression is an extension of simple linear regression. It is used when we want to **predict the value of a variable** based on the value of **two or more other variables**.
- The variable we want to predict is called the **dependent variable (or sometimes, the outcome, target or criterion variable**).
- The variables we are using to predict the value of the dependent variable are **called the independent variables (or predictor, explanatory variables).**

Examples:

- The **exam performance** can be predicted based on **revision time, test anxiety, lecture attendance and gender.**
- The **daily cigarette consumption** can be predicted based on **smoking duration, age when started smoking, smoker type, income and gender**.

Multiple regression

- Multiple regression also allows us to determine the overall fit (variance explained) of the model and the **relative contribution of each of the predictors** to the total variance explained.
- For example, you might want to know how much of the variation in exam performance can be explained by **revision time, test anxiety, lecture attendance and gender "as a whole",** but also the "relative contribution" of each independent variable in explaining the variance.

Various types of Regression analysis

- Linear Regression
- Stepwise Regression
- Group-wise Regression
- Hierarchical Regression
- Logistic Regression
- Regression with dummy Variable
- Regression with moderating variable
- Non-Linear Regression

Multiple Linear Regression

- **Linear regression is used for three main purposes**
  1. Used to assess how much each variable contributes in determining the value of the dependent variable
  2. What direction (positive or negative) they contribute
  3. Their relative importance in deciding the value of dependent variable

The R^2 and F Significance determine the reliability and validity of the model

357

Assumptions

- **Assumption #1:** D**ependent variable** should be measured on a continuous scale (i.e., it is either an **interval** or **ratio** variable).
- **Assumption #2:** Should have **two or more independent variables**, which can be either **continuous** (i.e., an **interval** or **ratio** variable) or categorical (i.e., an **ordinal** or **nominal** variable).
- **Assumption #3:** There should be **linear relationship** between (a) the dependent variable and **each** of your independent variables, and (b) the dependent variable and the independent variables **collectively**.
- **Assumption #4:** The data must not show **multicollinearity**, which occurs when we have two or more independent variables that are highly correlated with each other.
- **Assumption #5:** There should be **no significant outliers**, **high leverage points** or **highly influential points**.

358

Example

- A health researcher wants to predict "VO_2max", an indicator of fitness and health. **[VO2 max is a measure of the maximum volume of oxygen that an individaul can use. It is measured in millilitres per kilogramme of body weight per minute (ml/kg/min)].**
- Is it possible to predict an individual's VO_2max based on attributes that can be measured more easily and cheaply.
- The researcher recruited **100 participants to perform a maximum VO_2max test**, but also recorded their **"age", "weight", "heart rate" and "gender".**
- The researcher's goal is to predict **VO_2max** based on these four attributes: **age, weight, heart rate and gender.**

Variable View

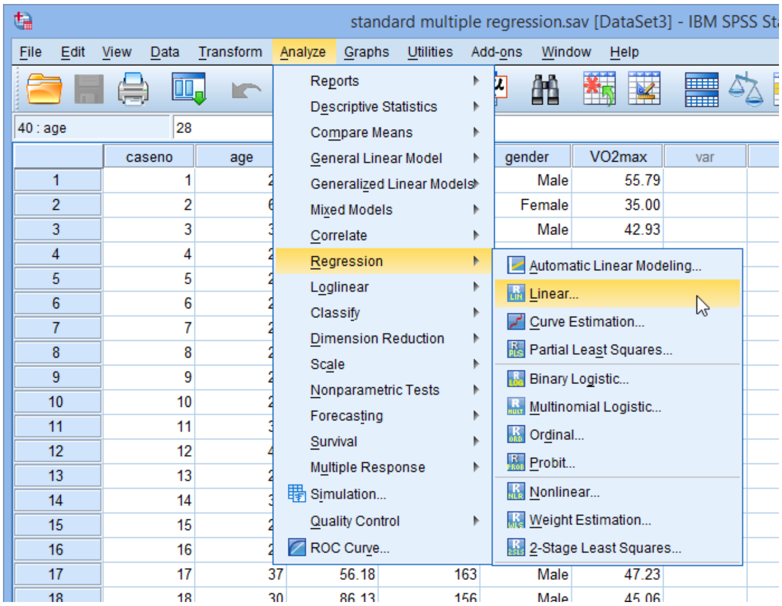
- **In the variable view key in the six variables: (1) VO2max, which is the maximal aerobic capacity; (2) age; (3) weight, which is the participant's weight; (4) heart rate; (5) gender,; and (6) caseno.**

*Untitled2 [DataSet1] - IBM SPSS Statistics Data Editor

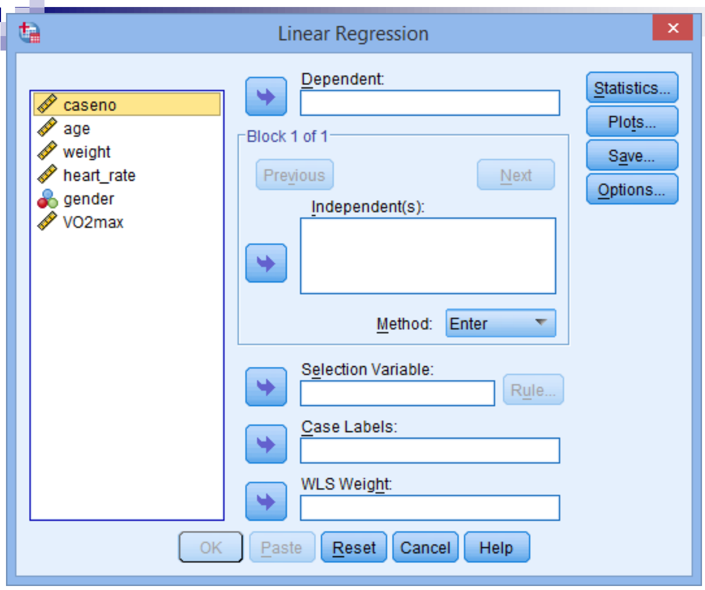
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals	Label	Values	Missing
1	Case_NO	Numeric	8	2	Case Number	None	None
2	Age	Numeric	8	2	Age of the Participants	None	None
3	Weight	Numeric	8	2	Body weight of the participants	None	None
4	Heart_rate	Numeric	8	2	Heart rate of the Participants	None	None
5	Gender	Numeric	8	2	Gender of the participants	None	None
6	VO2Max	Numeric	8	2	VO2max, which is the maximal aer...	None	None
7							

Click **Analyze > Regression > Linear**

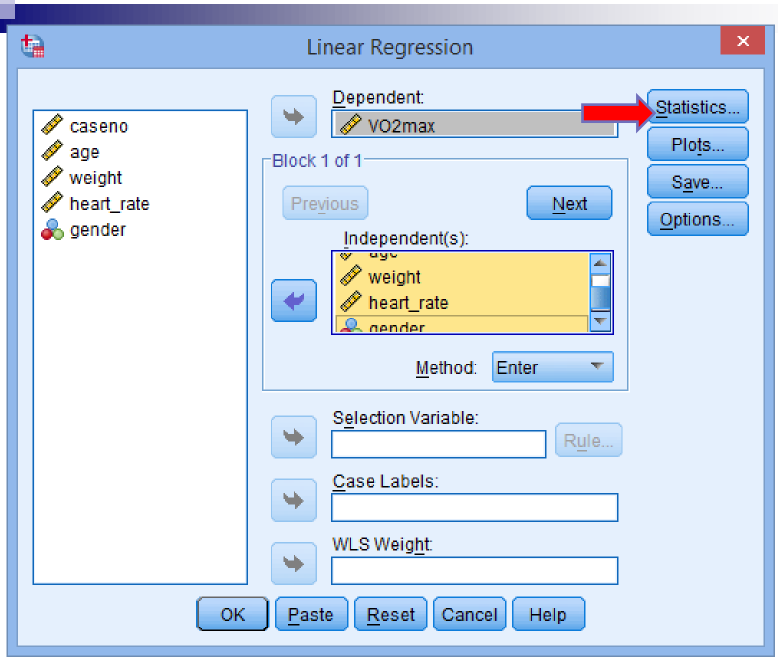


361

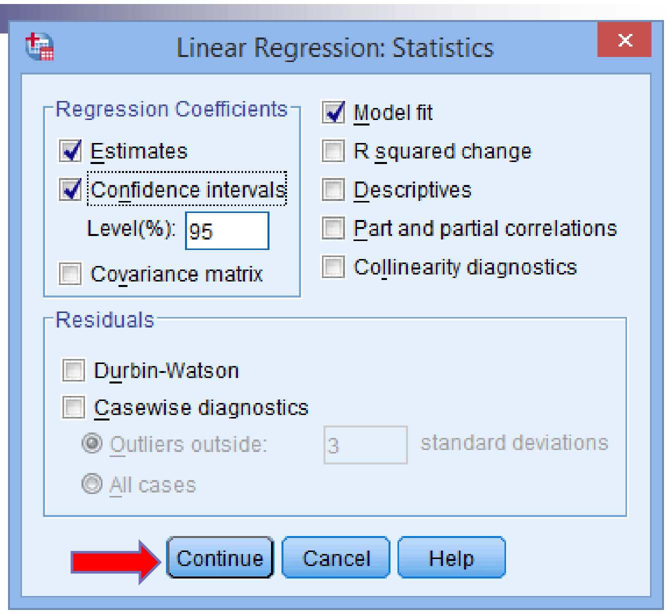


Transfer the dependent variable, VO2max, into the Dependent: and the independent variables, age, weight, heart_rate and gender into the Independent(s)

362

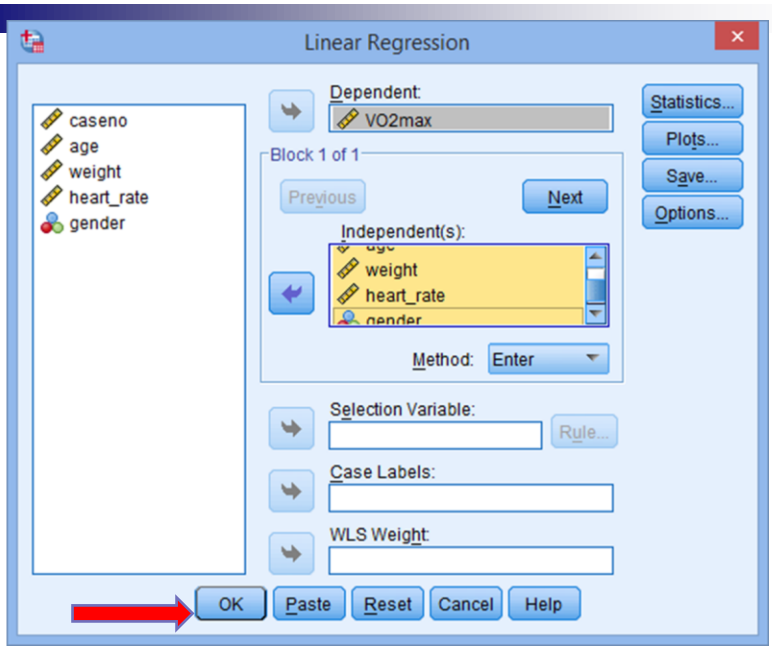


363



Select Confidence intervals in the – Regression Coefficients – area leaving the Level(%): option at "95".

364



365

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.760[a]	.577	.559	5.69097

a. Predictors: (Constant), gender, age, heart_rate, weight

- "R" column represents the value of R, the multiple correlation coefficient.
- R can be used to measure the quality of the prediction of the dependent variable
- The "**R Square**" column represents the R^2 value, which is the proportion of variance in the dependent variable that can be explained by the independent variables.
- Independent variables explain **57.7%** of the variability of our dependent variable,

366

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4196.483	**4**	1049.121	**32.393**	**.000**[b]
	Residual	3076.778	**95**	32.387		
	Total	7273.261	99			

a. Dependent Variable: VO2max

b. Predictors: (Constant), gender, age, heart_rate, weight

The independent variables statistically significantly predict the dependent variable, $F(4, 95) = 32.393$, $p < .0005$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	**87.830**	6.385		13.756	.000	75.155	100.506
	age	**-.165**	.063	-.176	-2.633	.010	-.290	-.041
	weight	**-.385**	.043	-.677	-8.877	.000	-.471	-.299
	heart_rate	**-.118**	.032	-.252	-3.667	.000	-.182	-.054
	gender	**13.208**	1.344	.748	9.824	.000	10.539	15.877

a. Dependent Variable: VO2max

The equation to predict VO2max from age, weight, heart_rate, gender, is:

Predicted VO2max = 87.83 – (0.165 x age) – (0.385 x weight) – (0.118 x heart_rate) + (13.208 x gender)

- Unstandardized coefficients indicate how much the dependent variable varies with an independent variable when all other independent variables are held constant. Consider the effect of age in this example.
- The unstandardized coefficient, for age is equal to **-0.165.**
- Each one year increase in age, there is a decrease in VO2max of **0.165** ml/min/kg.
- **The standardized coefficients indicate which independent variable contributes maximum to dependent variable. They rank the independent variables. Age comes first followed by weight, heart rate and so on.**

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	87.830	6.385		**13.756**	**.000**	75.155	100.506
	age	-.165	.063	-.176	**-2.633**	**.010**	-.290	-.041
	weight	-.385	.043	-.677	**-8.877**	**.000**	-.471	-.299
	heart_rate	-.118	.032	-.252	**-3.667**	**.000**	-.182	-.054
	gender	13.208	1.344	.748	**9.824**	**.000**	10.539	15.877

a. Dependent Variable: VO2max

- This table provides the statistical significance of each of the independent variables.
- This tests whether the unstandardized (or standardized) coefficients are equal to 0 (zero) in the population.
- $p < .05$, you can conclude that the coefficients are statistically significantly different to 0 (zero).

Report
A multiple regression was run to predict VO2max from gender, age, weight and heart rate. These variables statistically significantly predicted VO2max, $F(4, 95) = 32.393$, $p < .0005$, $R2 = .577$. All four variables added statistically significantly to the prediction, $p < .05$.

Stepwise Regression Analysis using SPSS

Stepwise Regression methods

- Researchers should build models with a few important independent variables
- What are those very important variables? How to decide? Stepwise regression addresses this problem
- Some researchers will model their research problem with more than 10 variables or so.
- If all independent variables are analyzed simultaneously the output produced is with all insignificant variables with one R^2 and F values.
- To see clearly the variables which enter the model, the R^2 produced, incremental R^2 , different F values will be an added advantage to the researchers to write clearly the research report.
- The only requirements are that the data is normally distributed (or rather, that the residuals are), and that there is no correlation between the independent variables (known as collinearity).

371

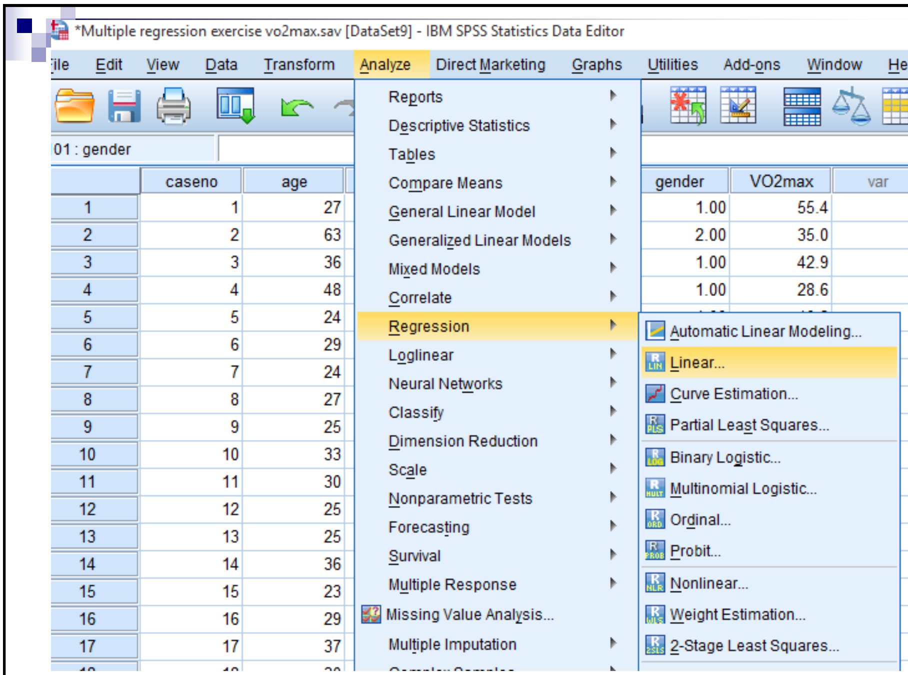
*Multiple regression exercise vo2max.sav [DataSet9] - IBM SPSS Statistics Data Editor

ile Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window

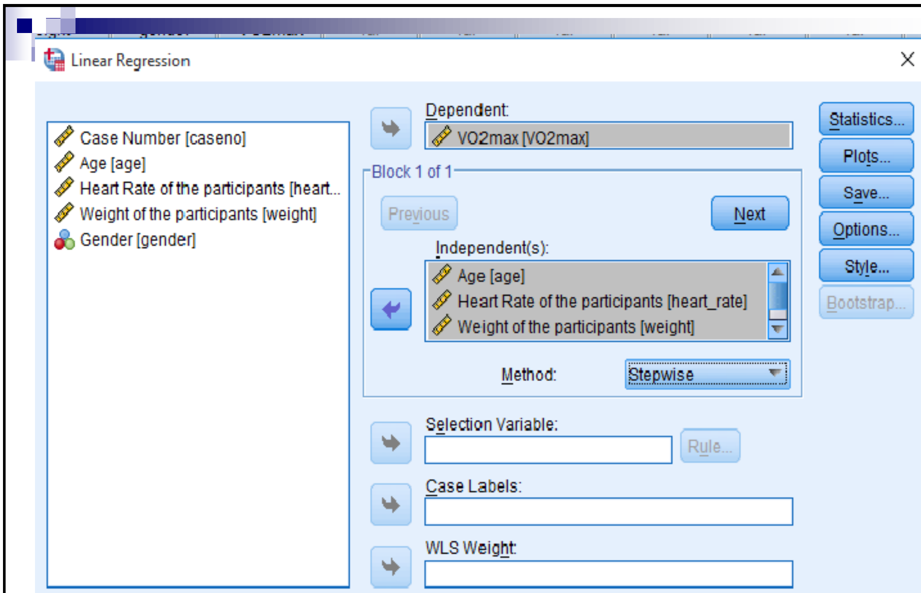
01 : gender

	caseno	age	heart_rate	weight	gender	VO2max
1	1	27	191	67.9	1.00	55.4
2	2	63	115	47.6	2.00	35.0
3	3	36	159	87.4	1.00	42.9
4	4	48	83	86.3	1.00	28.6
5	5	24	133	100.1	1.00	40.2
6	6	29	115	75.3	2.00	33.8
7	7	24	174	78.8	1.00	43.3
8	8	27	111	76.1	2.00	30.2
9	9	25	136	94.3	1.00	40.0
10	10	33	103	77.4	1.00	36.9
11	11	30	157	87.3	1.00	44.1
12	12	25	109	74.4	2.00	38.0
13	13	25	111	75.6	2.00	33.9
14	14	36	142	61.8	2.00	44.7
15	15	23	106	111.8	1.00	31.0
16	16	29	101	83.2	2.00	34.7
17	17	37	176	54.8	1.00	47.9
18	18	30	161	86.9	1.00	45.9

372


*Multiple regression exercise vo2max.sav [DataSet9] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
01 : gender
caseno age gender VO2max var
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Automatic Linear Modeling...
Linear...
Curve Estimation...
Partial Least Squares...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...

373


Linear Regression
Case Number [caseno]
Age [age]
Heart Rate of the participants [heart...
Weight of the participants [weight]
Gender [gender]
Dependent:
VO2max [VO2max]
Block 1 of 1
Previous
Next
Independent(s):
Age [age]
Heart Rate of the participants [heart_rate]
Weight of the participants [weight]
Method:
Stepwise
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
Style...
Bootstrap...

374

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.831[a]	.690	.687	4.7506

a. Predictors: (Constant), Heart Rate of the participants

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4928.968	1	4928.968	218.400	.000[b]
	Residual	2211.712	98	22.568		
	Total	7140.680	99			

a. Dependent Variable: VO2max

b. Predictors: (Constant), Heart Rate of the participants

375

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	11.631	2.216		5.249	.000		
	Heart Rate of the participants	.226	.015	.831	14.778	.000	1.000	1.000

a. Dependent Variable: VO2max

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	Age	-.040[b]	-.691	.491	-.070	.938	1.066	.938
	Weight of the participants	-.023[b]	-.378	.706	-.038	.892	1.121	.892
	Gender	-.068[b]	-1.136	.259	-.115	.883	1.133	.883

a. Dependent Variable: VO2max

b. Predictors in the Model: (Constant), Heart Rate of the participants

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions	
				(Constant)	Heart Rate of the participants
1	1	1.977	1.000	.01	.01
	2	.023	9.221	.99	.99

a. Dependent Variable: VO2max

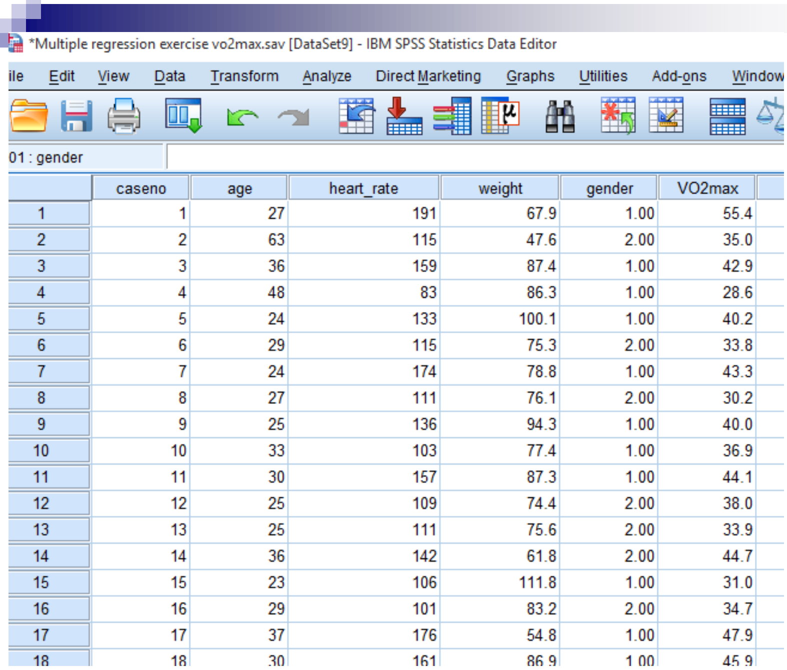
376

Groupwise Regression Analysis using SPSS

Group-wise Regression

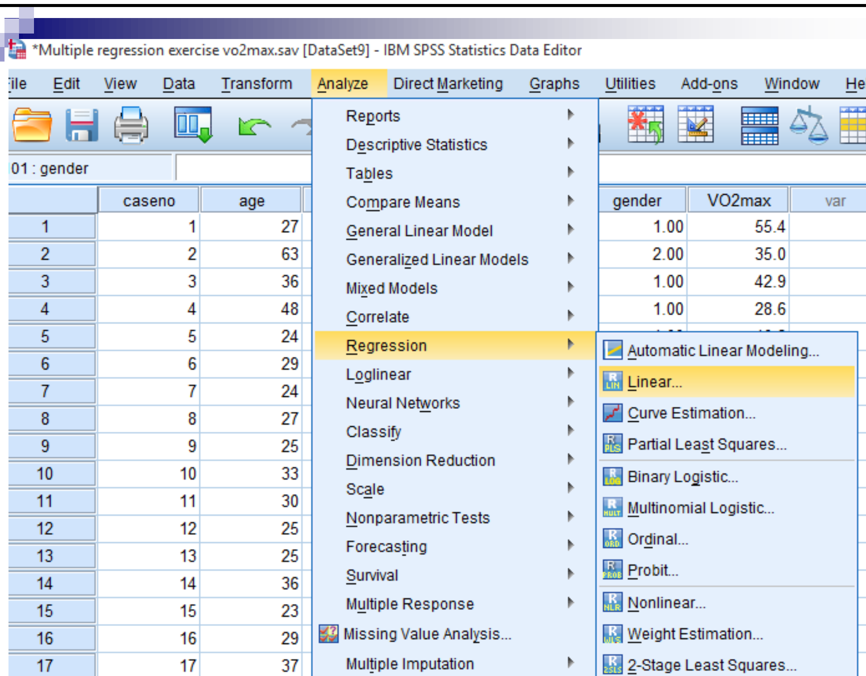
- In Previous example VO2Max we have two groups of participants **male and female**.
- It is interesting to find and present results separately to know more about the independent variables in male and female rather than giving results in general.
- Sometimes we will compare groups separately to get more insight. For example young age group how is it different from other age groups
- We have to run regression two times

378

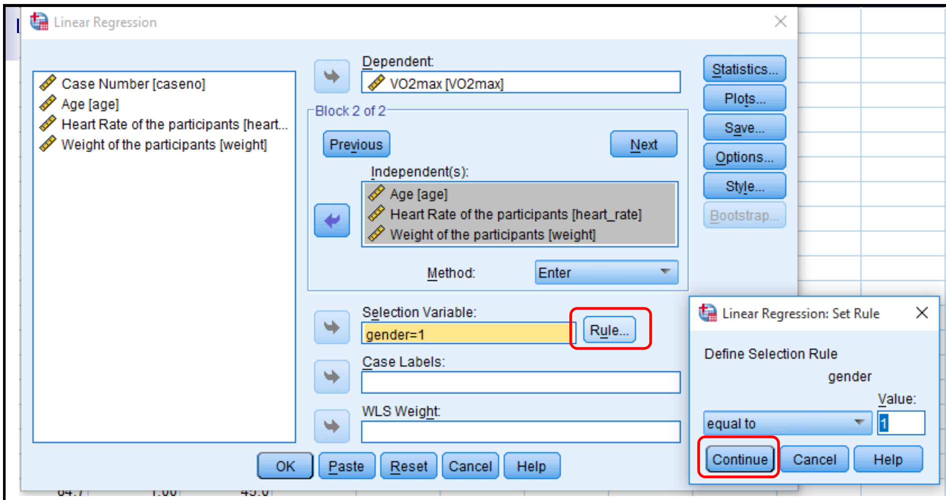


	caseno	age	heart_rate	weight	gender	VO2max
1	1	27	191	67.9	1.00	55.4
2	2	63	115	47.6	2.00	35.0
3	3	36	159	87.4	1.00	42.9
4	4	48	83	86.3	1.00	28.6
5	5	24	133	100.1	1.00	40.2
6	6	29	115	75.3	2.00	33.8
7	7	24	174	78.8	1.00	43.3
8	8	27	111	76.1	2.00	30.2
9	9	25	136	94.3	1.00	40.0
10	10	33	103	77.4	1.00	36.9
11	11	30	157	87.3	1.00	44.1
12	12	25	109	74.4	2.00	38.0
13	13	25	111	75.6	2.00	33.9
14	14	36	142	61.8	2.00	44.7
15	15	23	106	111.8	1.00	31.0
16	16	29	101	83.2	2.00	34.7
17	17	37	176	54.8	1.00	47.9
18	18	30	161	86.9	1.00	45.9

379



380



381

Descriptive Statistics[a]

	Mean	Std. Deviation	N
VO2max	45.898	8.3473	62
Age	30.90	7.796	62
Heart Rate of the participants	149.65	30.822	62
Weight of the participants	83.334	15.0221	62

a. Selecting only cases for which Gender = Male

Descriptive Statistics[a]

	Mean	Std. Deviation	N
VO2max	39.903	7.4314	38
Age	31.92	9.802	38
Heart Rate of the participants	127.76	27.004	38
Weight of the participants	72.295	12.1932	38

a. Selecting only cases for which Gender = Female

Correlations[a]

		VO2max	Age	Heart Rate of the participants	Weight of the participants
Pearson Correlation	VO2max	1.000	-.327	.823	-.587
	Age	-.327	1.000	-.325	.002
	Heart Rate of the participants	.823	-.325	1.000	-.577
	Weight of the participants	-.587	.002	-.577	1.000
Sig. (1-tailed)	VO2max	.	.005	.000	.000
	Age	.005	.	.005	.493
	Heart Rate of the participants	.000	.005	.	.000
	Weight of the participants	.000	.493	.000	.
N	VO2max	62	62	62	62
	Age	62	62	62	62
	Heart Rate of the participants	62	62	62	62
	Weight of the participants	62	62	62	62

a. Selecting only cases for which Gender = Male

Correlations[a]

		VO2max	Age	Heart Rate of the participants	Weight of the participants
Pearson Correlation	VO2max	1.000	-.120	.777	-.224
	Age	-.120	1.000	-.133	-.301
	Heart Rate of the participants	.777	-.133	1.000	-.372
	Weight of the participants	-.224	-.301	-.372	1.000
Sig. (1-tailed)	VO2max	.	.236	.000	.088
	Age	.236	.	.213	.033
	Heart Rate of the participants	.000	.213	.	.011
	Weight of the participants	.088	.033	.011	.
N	VO2max	38	38	38	38
	Age	38	38	38	38
	Heart Rate of the participants	38	38	38	38
	Weight of the participants	38	38	38	38

a. Selecting only cases for which Gender = Female

382

Variables Entered/Removed^{a,b}

Model	Variables Entered	Variables Removed	Method
1	Heart Rate of the participants	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	Age, Weight of the participants[c]	.	Enter

a. Dependent Variable: VO2max

b. Models are based only on cases for which Gender = Male

c. All requested variables entered.

Variables Entered/Removed^{a,b}

Model	Variables Entered	Variables Removed	Method
1	Heart Rate of the participants	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	Age, Weight of the participants[c]	.	Enter

a. Dependent Variable: VO2max

b. Models are based only on cases for which Gender = Female

c. All requested variables entered.

383

Model Summary

Model	R Gender = Male (Selected)	R Square	Adjusted R Square	Std. Error of the Estimate
1	.823[a]	.678	.672	4.7787
2	.840[b]	.706	.691	4.6421

a. Predictors: (Constant), Heart Rate of the participants

b. Predictors: (Constant), Heart Rate of the participants, Age, Weight of the participants

Model Summary

Model	R Gender = Female (Selected)	R Square	Adjusted R Square	Std. Error of the Estimate
1	.777[a]	.603	.592	4.7462
2	.780[b]	.608	.573	4.8535

a. Predictors: (Constant), Heart Rate of the participants

b. Predictors: (Constant), Heart Rate of the participants, Age, Weight of the participants

ANOVA^{a,b}

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3000.505	3	1000.168	46.414	.000[c]
	Residual	1249.844	58	21.549		
	Total	4250.350	61			

a. Dependent Variable: VO2max

b. Selecting only cases for which Gender = Male

c. Predictors: (Constant), Weight of the participants, Age, Heart Rate of the participants

ANOVA^{a,b}

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1232.416	1	1232.416	54.710	.000[c]
	Residual	810.954	36	22.527		
	Total	2043.370	37			
2	Regression	1242.465	3	414.155	17.582	.000[d]
	Residual	800.905	34	23.556		
	Total	2043.370	37			

a. Dependent Variable: VO2max

b. Selecting only cases for which Gender = Female

c. Predictors: (Constant), Heart Rate of the participants

d. Predictors: (Constant), Heart Rate of the participants, Age, Weight of the participants

Coefficients^{a,b}

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	VIF
1	(Constant)	12.537	3.032		4.135	.000		
	Heart Rate of the participants	.223	.020	.823	11.231	.000	1.000	1.000
2	(Constant)	31.217	8.457		3.691	.000		
	Heart Rate of the participants	.183	.026	.675	7.104	.000	.562	1.779
	Age	-.115	.083	-.107	-1.379	.173	.843	1.186
	Weight of the participants	-.109	.050	-.197	-2.194	.032	.628	1.591

a. Dependent Variable: VO2max

b. Selecting only cases for which Gender = Male

Coefficients^{a,b}

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	VIF
1	(Constant)	12.597	3.771		3.340	.002		
	Heart Rate of the participants	.214	.029	.777	7.397	.000	1.000	1.000
2	(Constant)	7.738	9.792		.790	.435		
	Heart Rate of the participants	.222	.033	.808	6.709	.000	.796	1.257
	Age	.009	.089	.011	.097	.924	.839	1.191
	Weight of the participants	.048	.076	.079	.634	.530	.736	1.358

a. Dependent Variable: VO2max

b. Selecting only cases for which Gender = Female

384

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics: Tolerance	VIF	Minimum Tolerance
1	Age	-.066[b]	-.851	.398	-.110	.894	1.118	.894
	Weight of the participants	-.167[b]	-1.904	.062	-.241	.667	1.500	.667

a. Dependent Variable: VO2max

b. Predictors in the Model: (Constant), Heart Rate of the participants

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics: Tolerance	VIF	Minimum Tolerance
1	Age	-.017[b]	-.158	.875	-.027	.982	1.018	.982
	Weight of the participants	.075[b]	.655	.517	.110	.862	1.160	.862

a. Dependent Variable: VO2max

b. Predictors in the Model: (Constant), Heart Rate of the participants

Collinearity Diagnostics[a,b]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions: (Constant)	Heart Rate of the participants	Age	Weight of the participants
1	1	1.980	1.000	.01	.01		
	2	.020	9.891	.99	.99		
2	1	3.884	1.000	.00	.00	.00	.00
	2	.068	7.561	.00	.19	.27	.02
	3	.044	9.369	.00	.03	.40	.25
	4	.004	33.218	1.00	.78	.32	

a. Dependent Variable: VO2max

b. Selecting only cases for which Gender = Male

Collinearity Diagnostics[a,b]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions: (Constant)	Heart Rate of the participants	Age	Weight of the participants
1	1	1.979	1.000	.01	.01		
	2	.021	9.693	.99	.99		
2	1	3.868	1.000	.00	.00	.00	.00
	2	.081	6.928	.00	.05	.64	.03
	3	.047	9.046	.00	.39	.00	.18
	4	.005	28.872	1.00	.56	.35	.78

a. Dependent Variable: VO2max

b. Selecting only cases for which Gender = Female

385

Regression Analysis for dummy variables using SPSS

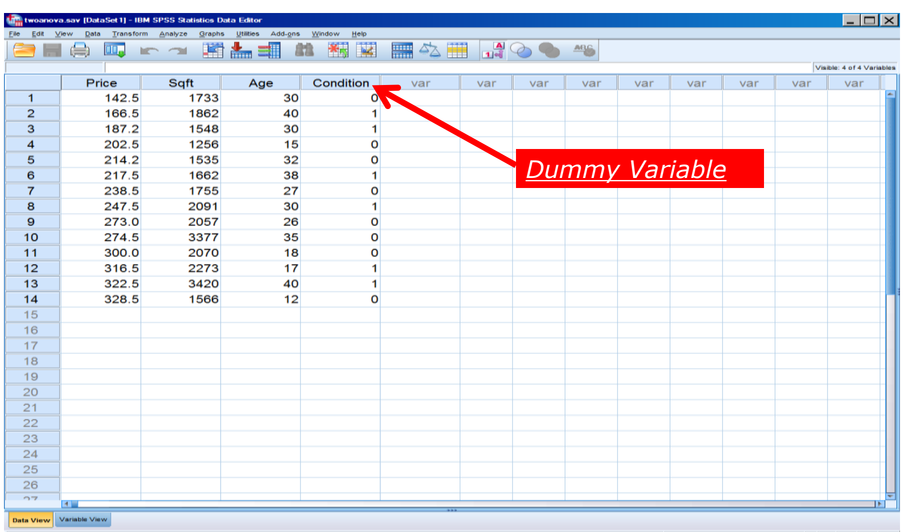
Regression with dummy variable

- Dummy variables are included in the model to capture the effect of dichotomous variables like good or bad (good may be 1 and bad may be zero)
- For example, a researcher wants to study the variables that determine the house prices
- The house price is dependent on **SQ feet, age and locality**. Apart from this present condition of the house is also important.
- If the house is maintained well it will get good price and vice versa.

$$code \begin{cases} 1 = & if\ the\ house\ is\ maintained\ well \\ 0 = & otherwise \end{cases}$$

387

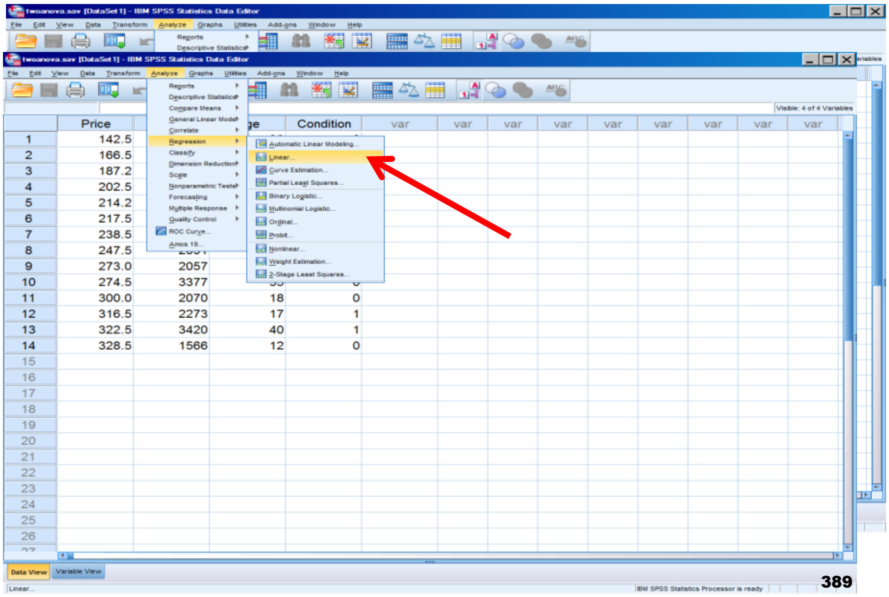
SPSS worksheet and data



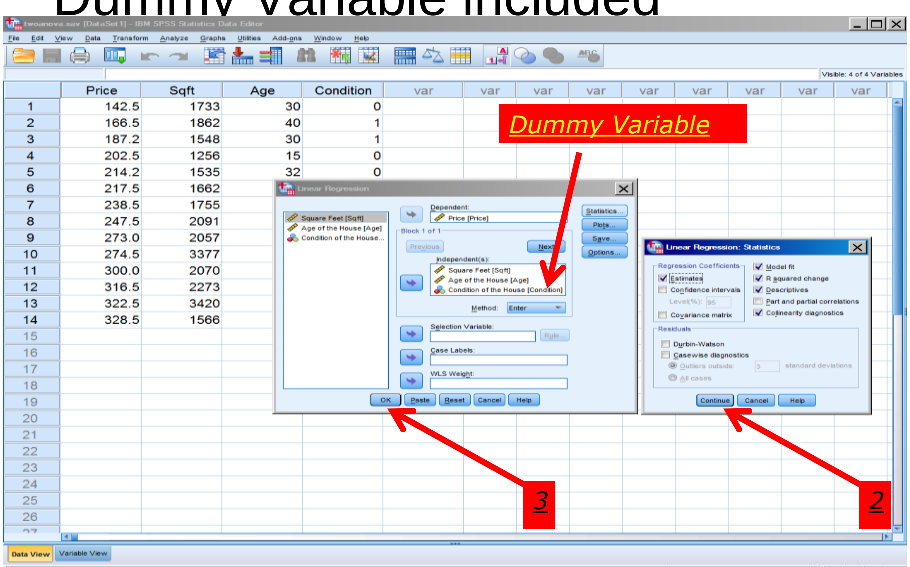
	Price	Sqft	Age	Condition
1	142.5	1733	30	0
2	166.5	1862	40	1
3	187.2	1548	30	1
4	202.5	1256	15	0
5	214.2	1535	32	0
6	217.5	1662	38	1
7	238.5	1755	27	0
8	247.5	2091	30	1
9	273.0	2057	26	0
10	274.5	3377	35	0
11	300.0	2070	18	0
12	316.5	2273	17	1
13	322.5	3420	40	1
14	328.5	1566	12	0

388

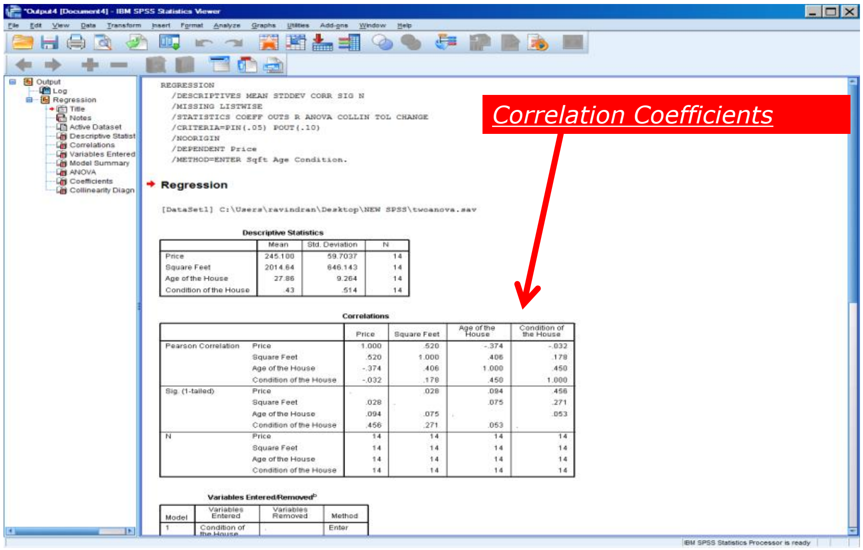
SPSS worksheet and data



Dummy Variable included



Results – Output 1




```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA COLLIN TOL CHANGE
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Price
  /METHOD=ENTER Sqft Age Condition.
```

Regression

[DataSet1] C:\Users\ravindran\Desktop\NEW SPSS\twoanova.sav

Descriptive Statistics

	Mean	Std. Deviation	N
Price	245.100	59.7037	14
Square Feet	2014.64	646.143	14
Age of the House	27.86	9.264	14
Condition of the House	.43	.514	14

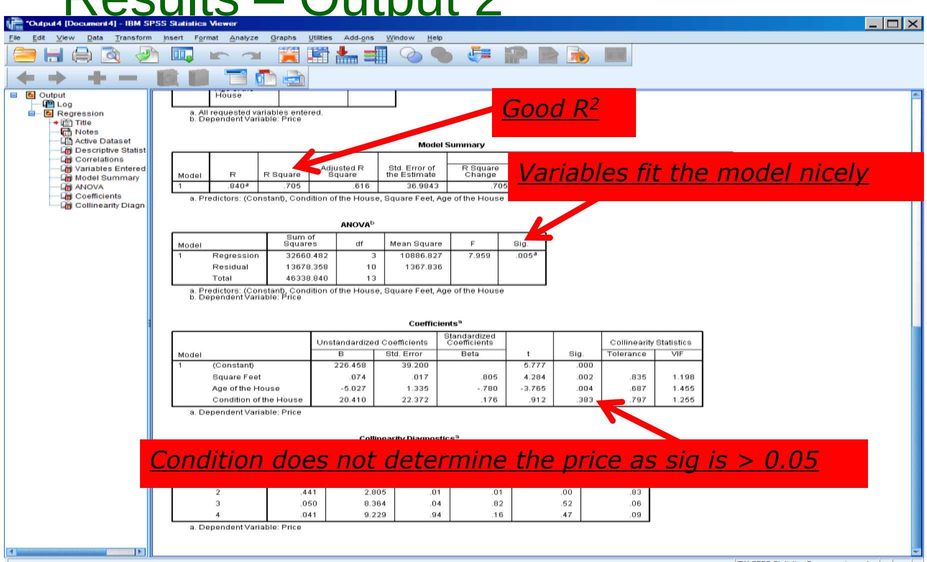
Correlations

		Price	Square Feet	Age of the House	Condition of the House
Pearson Correlation	Price	1.000	.520	-.374	-.032
	Square Feet	.520	1.000	.406	.178
	Age of the House	-.374	.406	1.000	.450
	Condition of the House	-.032	.178	.450	1.000
Sig. (1-tailed)	Price	.	.028	.094	.456
	Square Feet	.028	.	.075	.271
	Age of the House	.094	.075	.	.053
	Condition of the House	.456	.271	.053	.
N	Price	14	14	14	14
	Square Feet	14	14	14	14
	Age of the House	14	14	14	14
	Condition of the House	14	14	14	14

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Condition of the House	.	Enter

Results – Output 2



a. All requested variables entered.
b. Dependent Variable: Price

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change
1	.840[a]	.705	.616	36.9843	.705

a. Predictors: (Constant), Condition of the House, Square Feet, Age of the House

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	32660.482	3	10886.827	7.959	.005[a]
	Residual	13678.358	10	1367.836		
	Total	46338.840	13			

a. Predictors: (Constant), Condition of the House, Square Feet, Age of the House
b. Dependent Variable: Price

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	226.458	39.200		5.777	.000		
	Square Feet	.074	.017	.805	4.284	.002	.835	1.198
	Age of the House	-5.027	1.335	-.780	-3.765	.004	.687	1.455
	Condition of the House	20.410	22.372	.176	.912	.383	.797	1.256

a. Dependent Variable: Price

Collinearity Diagnostics[a]

2	.441	2.805	.01	.01	.00	.83
3	.050	8.364	.04	.82	.52	.06
4	.041	9.229	.94	.16	.47	.09

a. Dependent Variable: Price

Hierarchical Regression analysis using SPSS

Hierarchical regression

- Researcher specifies some order of entry based on theoretical considerations
- They might start with the most important predictor(s) to see what they do on their own before those of lesser theoretical interest are added
- Otherwise, might start with those of lesser interest, and see what they add to the equation (in terms of R^2)
- This test is performed adding main independent variables first and followed by a group of demographic variables in the analysis by hierarchy

Example

- A research hypothesis might state that there are differences between the **average salary** for male employees and female employees, even after we take into account differences between **education levels and prior work experience**.

- In hierarchical regression, the independent variables are entered into the analysis in a sequence of blocks, or groups that may contain one or more variables. In the example above, **education and work experience would be entered in the first block** and **sex would be entered in the second block**.

395

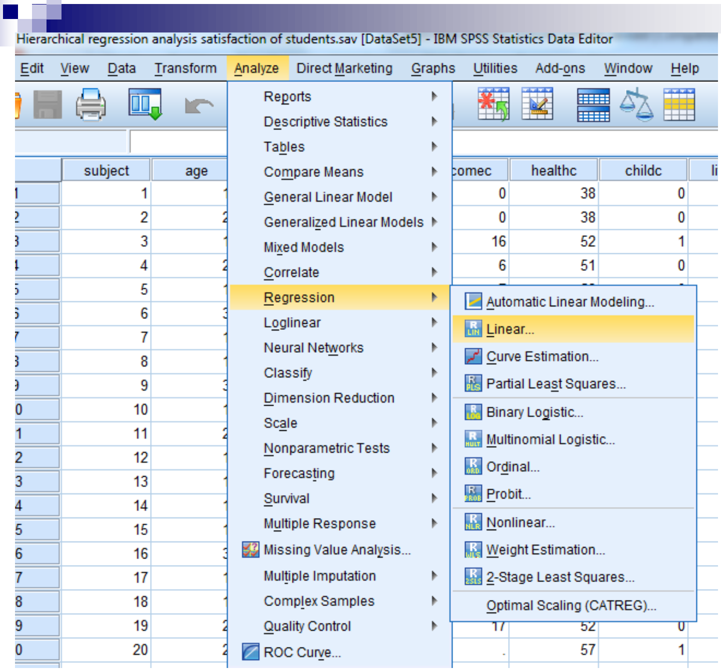
Example

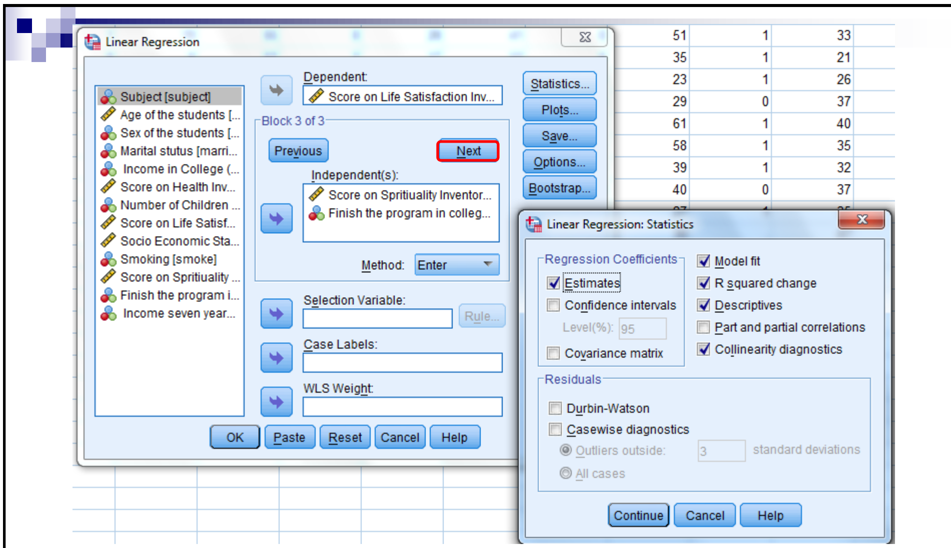
- The prediction of life satisfaction seven years after college from the variables that can be measured while the student is in college.
  - **Age**
  - **Gender** (0=Male, 1=Female)
  - **Married** (0=No, 1=Yes)
  - **IncomeC** Income in College (in thousands)
  - **HealthC** Score on Health Inventory in College
  - **ChildC** Number of Children while in College
  - **LifeSatC** Score on Life Satisfaction Inventory in College
  - **SES** Socio Economic Status of Parents
  - **Smoker** (0=No, 1=Yes)
  - **SpiritC** Score on Sprituality Inventory in College
  - **Finish** Finish the program in college (0=No, 1=Yes)
  - **LifeSat** Score on Life Satisfaction Inventory seven years after College
  - **Income** Income seven years after College (in thousands)

396

	subject	age	gender	married	incomec	healthc	childc	lifesatc	ses	smoke	spiritc	finish	lifesat7	income7
1	1	16	0	0	0	38	0	17	17	1	30	1	22	26
2	2	28	1	0	0	38	0	16	21	1	39	1	20	15
3	3	16	.	1	16	52	1	39	40	0	30	1	42	88
4	4	23	1	0	6	51	0	22	31	0	60	1	48	73
5	5	18	0	1	7	52	0	25	38	0	32	0	.	14
6	6	30	0	1	25	43	2	53	36	1	39	0	33	38
7	7	19	0	1	19	55	0	28	41	0	51	1	33	45
8	8	19	1	0	0	52	2	17	52	0	35	1	21	16
9	9	34	0	0	29	60	2	20	56	0	23	1	26	64
10	10	16	1	0	0	53	0	21	27	0	29	0	37	19
11	11	25	1	0	3	39	0	18	34	1	61	1	40	56
12	12	16	1	1	1	42	0	31	29	1	58	1	35	70
13	13	16	.	0	0	43	0	15	28	1	39	1	32	71
14	14	16	0	1	18	54	1	34	38	0	40	0	37	44
15	15	16	1	0	0	52	0	20	38	0	27	1	35	25
16	16	32	1	1	26	54	1	39	37	0	30	.	47	38
17	17	19	0	0	0	46	0	17	25	0	36	1	26	39
18	18	17	1	1	10	55	2	48	53	0	43	0	42	6
19	19	24	0	0	17	52	0	16	36	0	54	1	38	75
20	20	26	1	1	.	57	1	39	41	0	32	1	42	67

- To examine the prediction of life satisfaction seven years after college in several stages.
- In the first stage, he/she enters **demographic variables that the individual has little or no control over, age, gender, and socio-economic status of parents**.
- In the second block variables are entered that the individual has at least some control, **such as smoking, having children, being married**, etc. The third block consists of the two **attitudinal variables, life satisfaction and spirituality.**





1. **Transfer age, gender, and socio-economic status of parents. Then click next**
2. **Transfer Smoking, having children, being married Then click next**
3. **Life satisfaction and spirituality.**

399

Descriptive Statistics

	Mean	Std. Deviation	N
Score on Life Satisfaction Inventory seven years after College	33.44	8.286	16
Age of the students	21.50	5.785	16
Sex of the students	.56	.512	16
Socio Economic Status of Parents	35.94	11.162	16
Marital stutus	.38	.500	16
Number of Children while in College	.63	.885	16
Smoking	.31	.479	16
Score on Sprituality Inventory in College	41.06	12.272	16
Finish the program in college	.75	.447	16

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Socio Economic Status of Parents, Sex of the students, Age of the students[b]	.	Enter
2	Marital stutus, Smoking, Number of Children while in College[b]	.	Enter
3	Score on Sprituality Inventory in College, Finish the program in college[b]	.	Enter

a. Dependent Variable: Score on Life Satisfaction Inventory seven years after College

b. All requested variables entered.

400

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.346[a]	.120	-.100	8.692	.120	.544	3	12	.661
2	.683[b]	.466	.111	7.814	.347	1.949	3	9	.192
3	.845[c]	.714	.387	6.490	.247	3.024	2	7	.113

a. Predictors: (Constant), Socio Economic Status of Parents, Sex of the students, Age of the students

b. Predictors: (Constant), Socio Economic Status of Parents, Sex of the students, Age of the students, Marital stutus, Smoking, Number of Children while in College

c. Predictors: (Constant), Socio Economic Status of Parents, Sex of the students, Age of the students, Marital stutus, Smoking, Number of Children while in College , Score on Sprituality Inventory in College , Finish the program in college

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	123.398	3	41.133	.544	.661[b]
	Residual	906.539	12	75.545		
	Total	1029.938	15			
2	Regression	480.368	6	80.061	1.311	.343[c]
	Residual	549.569	9	61.063		
	Total	1029.938	15			
3	Regression	735.119	8	91.890	2.182	.160[d]
	Residual	294.818	7	42.117		
	Total	1029.938	15			

a. Dependent Variable: Score on Life Satisfaction Inventory seven years after College

b. Predictors: (Constant), Socio Economic Status of Parents, Sex of the students, Age of the students

c. Predictors: (Constant), Socio Economic Status of Parents, Sex of the students, Age of the students, Marital stutus, Smoking, Number of Children while in College

d. Predictors: (Constant), Socio Economic Status of Parents, Sex of the students, Age of the students, Marital stutus, Smoking, Number of Children while in College , Score on Sprituality Inventory in College , Finish the program in college

401

Coefficients[a]

Demographic Variables coefficients

Main Variables coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	29.134	10.772		2.705	.019		
	Age of the students	-.132	.406	-.092	-.324	.751	.913	1.095
	Sex of the students	4.507	4.456	.279	1.012	.332	.966	1.035
	Socio Economic Status of Parents	.128	.207	.172	.617	.549	.939	1.065
2	(Constant)	20.997	13.773		1.525	.162		
	Age of the students	.365	.427	.255	.855	.415	.668	1.496
	Sex of the students	5.959	4.223	.368	1.411	.192	.870	1.150
	Socio Economic Status of Parents	.080	.387	.108	.208	.840	.219	4.572
	Marital stutus	9.376	4.659	.566	2.013	.075	.750	1.333
	Number of Children while in College	-4.732	4.102	-.505	-1.153	.278	.309	3.238
	Smoking	-7.014	6.152	-.405	-1.140	.284	.469	2.131
3	(Constant)	14.574	11.735		1.242	.254		
	Age of the students	.459	.369	.320	1.243	.254	.616	1.622
	Sex of the students	5.137	3.523	.318	1.458	.188	.862	1.161
	Socio Economic Status of Parents	.114	.392	.154	.291	.779	.146	6.831
	Marital stutus	4.538	4.342	.274	1.045	.331	.596	1.679
	Number of Children while in College	-4.819	4.980	-.515	-.968	.365	.145	6.919
	Smoking	-7.857	5.708	-.454	-1.377	.211	.376	2.659
	Score on Sprituality Inventory in College	.298	.179	.441	1.664	.140	.583	1.716
	Finish the program in college	-8.598	5.297	-.464	-1.623	.149	.500	1.998

a. Dependent Variable: Score on Life Satisfaction Inventory seven years after College

402

Binary Logistics Regression analysis using SPSS

Binary Logistics Regression analysis

- A binomial logistic regression (often referred to simply as logistic regression), predicts the probability that an observation falls into one of two categories of a dichotomous dependent variable based on one or more independent variables that can be either continuous or categorical.

Example

- **Exam performance** can be predicted based on revision time, test anxiety and lecture attendance (i.e., where the dependent variable is **"exam performance", measured on a dichotomous scale – "passed" or "failed"** – and you have three independent variables: **"revision time", "test anxiety" and "lecture attendance").**

- **Drug use** can be predicted based on prior criminal convictions, drug use amongst friends, income, age and gender (i.e., where the dependent variable is **"drug use", measured on a dichotomous scale – "yes" or "no"** – and you have **five independent variables: "prior criminal convictions", "drug use amongst friends", "income", "age" and "gender").**

405

Assumptions

- **Assumption #1:** D**ependent variable** should be measured on a **dichotomous** scale. Examples of **dichotomous variables** include gender (two groups: "males" and "females"), presence of heart disease (two groups: "yes" and "no"), body composition (two groups: "obese" or "not obese")
- **Assumption #2:** O**ne or more independent variables**, which can be either **continuous** (i.e., an **interval** or **ratio** variable) or **categorical** (i.e., an **ordinal** or **nominal** variable).
- **Assumption #3:** I**ndependence of observations** and the dependent variable should have **mutually exclusive and exhaustive categories**.
- **Assumption #4:** There needs to be a **linear relationship between any continuous independent variables and the logit transformation of the dependent variable**.

406

Example

- A health researcher wants to be able to predict whether the "incidence of heart disease" can be predicted based on "age", "weight", "gender" and "VO_2max" (i.e., where VO_2max refers to maximal aerobic capacity, an indicator of fitness and health).
- The researcher recruited 100 participants to perform a maximum **VO_2max test** as well as he recorded **their age, weight and gender**. The participants were also evaluated for the **presence of heart disease** **(Yes or No)**.
- There are six variables:

(1) Heart_disease, which is whether the participant has heart disease: "yes" or "no" (i.e., the dependent variable);

(2) VO2max, which is the maximal aerobic capacity;

(3) age, which is the participant's age;

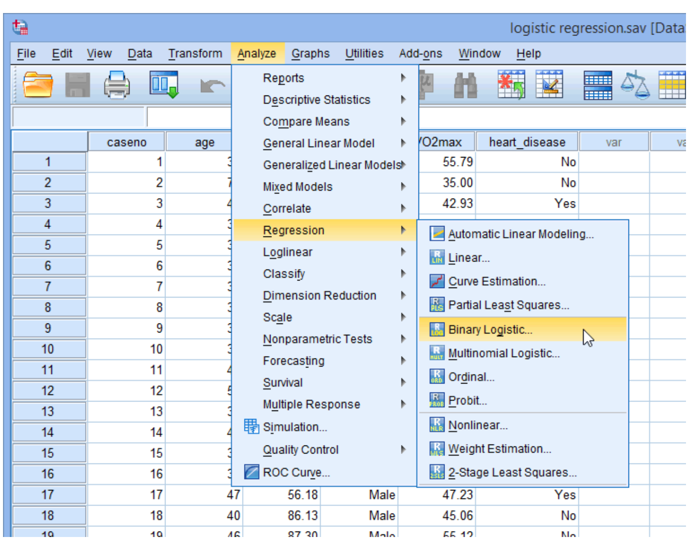
(4) weight, which is the participant's weight (technically, it is their 'mass'); and

(5) gender, which is the participant's gender (i.e., the independent variables); and

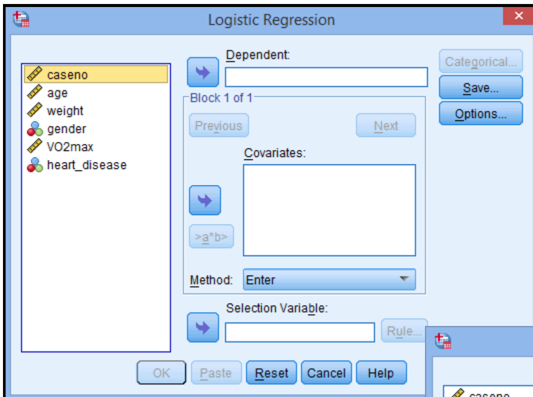
(6) Caseno, which is the case number.

407

Click **Analyze > Regression > Binary Logistic...**

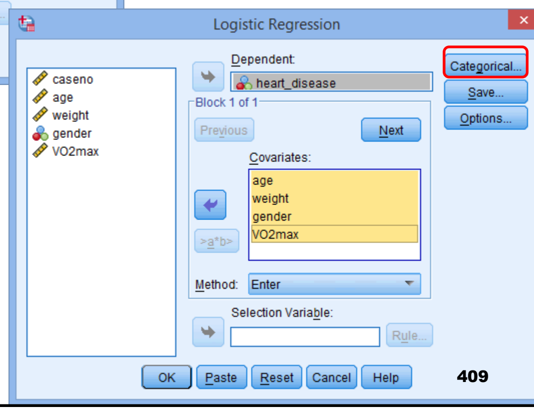


408

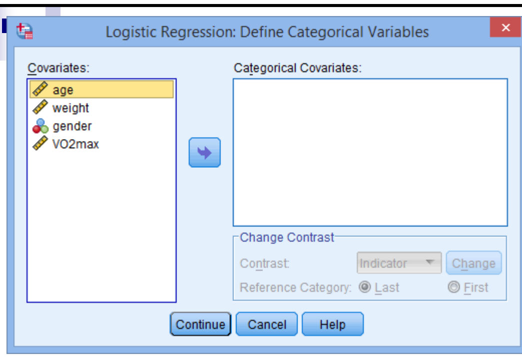


Click the categorical Button

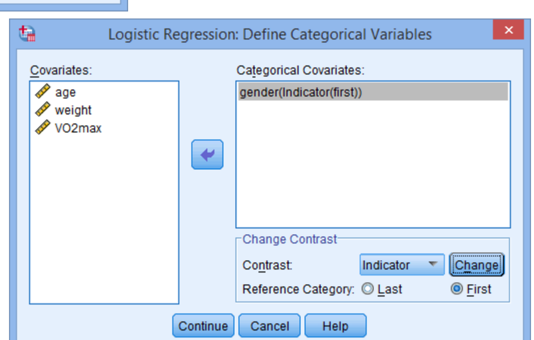
Transfer the dependent variable, heart_disease, into the Dependent: box, and the independent variables, age, weight, gender and VO2max into the Covariates



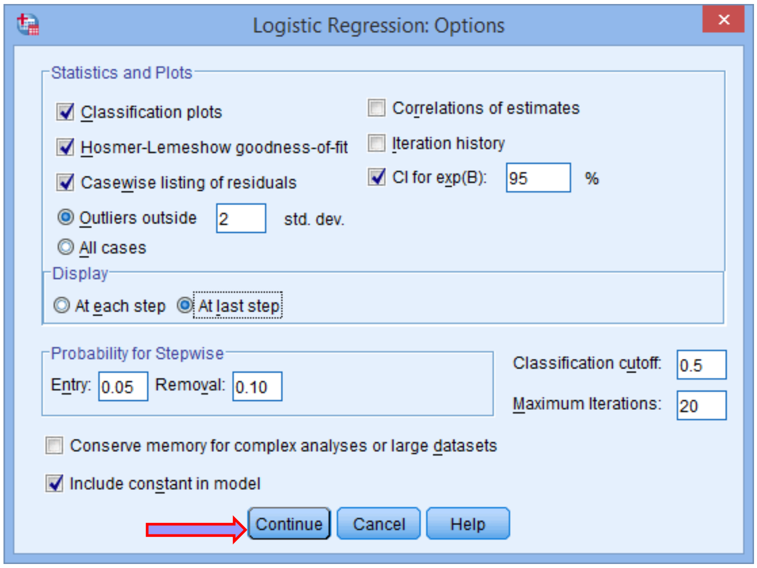
409



In the –Change Contrast– area, change the Reference Category: from the Last option to the First option. Then, click the button, as shown below:



410



411

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	102.088[a]	.240	.330

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

The explained variation in the dependent variable based on our model ranges from 24.0% to 33.0%,

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
Step 1[a]	age	.085	.028	9.132	1	.003	1.089	1.030	1.151
	weight	.006	.022	.065	1	.799	1.006	.962	1.051
	gender(1)	1.950	.842	5.356	1	.021	7.026	1.348	36.625
	VO2max	-.099	.048	4.266	1	.039	.906	.824	.995
	Constant	-1.676	3.336	.253	1	.615	.187		

a. Variable(s) entered on step 1: age, weight, gender, VO2max.

You can see that age ($p = .003$), gender ($p = .021$) and VO2max ($p = .039$) added significantly to the model/prediction, but weight ($p = .799$) did not add significantly to the model. We can use the information in the "Variables in the Equation" table to predict the probability of an event occurring based on a one unit change in an independent variable when all other independent variables are kept constant.

For example, the table shows that the odds of having heart disease ("yes" category) is 7.026 times greater for males as opposed to females.

412

Report

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	age	.085	.028	9.132	1	.003	1.089	1.030	1.151
	weight	.006	.022	.065	1	.799	1.006	.962	1.051
	gender(1)	1.950	.842	5.356	1	.021	7.026	1.348	36.625
	VO2max	-.099	.048	4.266	1	.039	.906	.824	.995
	Constant	-1.676	3.336	.253	1	.615	.187		

a. Variable(s) entered on step 1: age, weight, gender, VO2max.

- A logistic regression was performed to ascertain the effects of age, weight, gender and VO_2max on the likelihood that participants have heart disease.
- The model explained 33.0% (Nagelkerke R^2) of the variance in heart disease. Males were 7.02 times more likely to exhibit heart disease than females.
- Increasing age was associated with an increased likelihood of exhibiting heart disease, but increasing VO_2max was associated with a reduction in the likelihood of exhibiting heart disease.

413

Moderator Analysis using SPSS

Moderator

- **A moderator is a variable that alters the direction or strength of the relationship between a predictor (independent) and an outcome (dependent)**
- Really, it is just an interaction – the effect of one variable depends on the level of another.
- A moderator analysis is used to determine whether the relationship between two variables depends on (is moderated by) the value of a third variable. This relationship is commonly between:
  - (a) a continuous dependent variable and continuous independent variable, which is modified by a dichotomous moderator variable;
  - (b) a continuous dependent variable and continuous independent variable, which is modified by a polytomous moderator variable; or
  - (c) a continuous dependent variable and continuous independent variable, which is modified by a continuous moderator variable

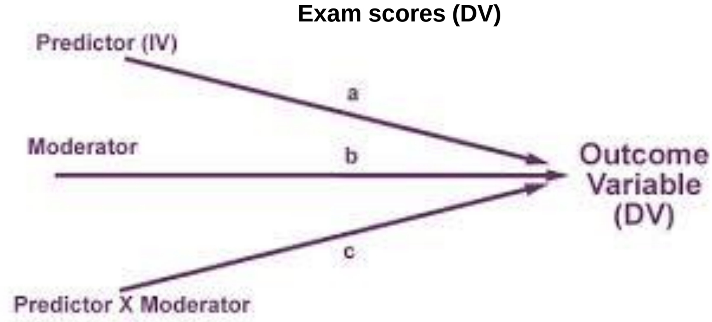
415

- What is the definition of moderation?
  - A moderator analysis is used to determine whether the relationship between two variables depends on (is moderated by) the value of a third variable.
  - IQ (IV)
  - Study Hrs (IV)
  - Exam scores (DV)

416

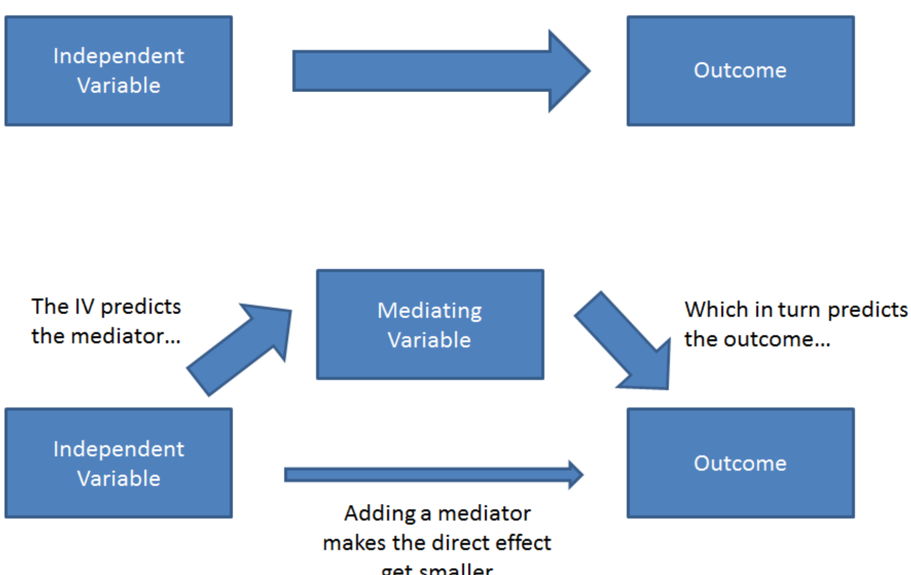
Moderators

IQ (IV)
Study Hrs (IV)
Exam scores (DV)



417

Ordinary Main Effect
(a "direct effect" in mediation terminology)



418

Example

- The relationship between HDL cholesterol and amount of exercise performed per week is different for normal weight and obese participants **(i.e., the continuous dependent variable is "HDL cholesterol", the continuous independent variable is "amount of exercise performed per week"** and the dichotomous moderator variable is **"body composition", consisting of two groups: "normal weight" and "obese**")?

419

Example

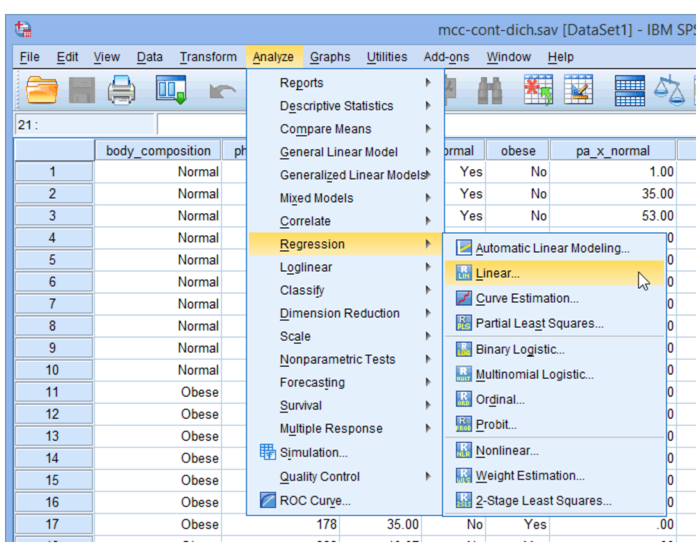
- The relationship between salary and years of education is moderated by gender (i.e., the continuous dependent variable is "salary", the continuous independent variable is "years of education" and the dichotomous moderator variable is "gender", which consists of two groups: "males" and "females").
- If it is, gender (i.e., the dichotomous moderator variable) moderates the relationship between the years of education and salary.

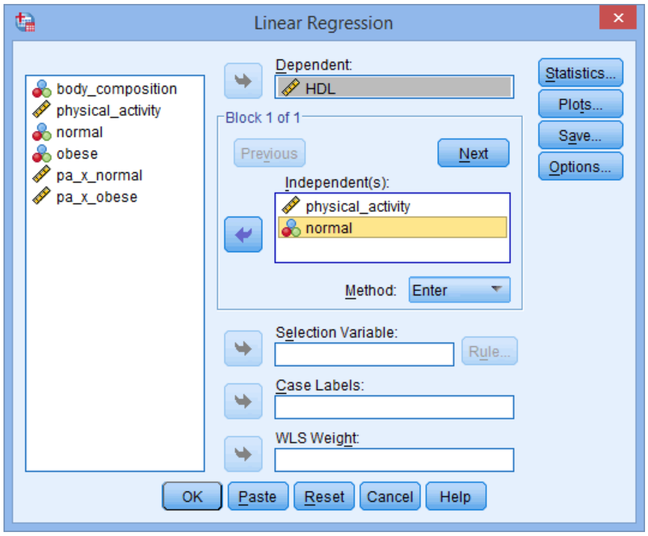
420

Example in SPSS

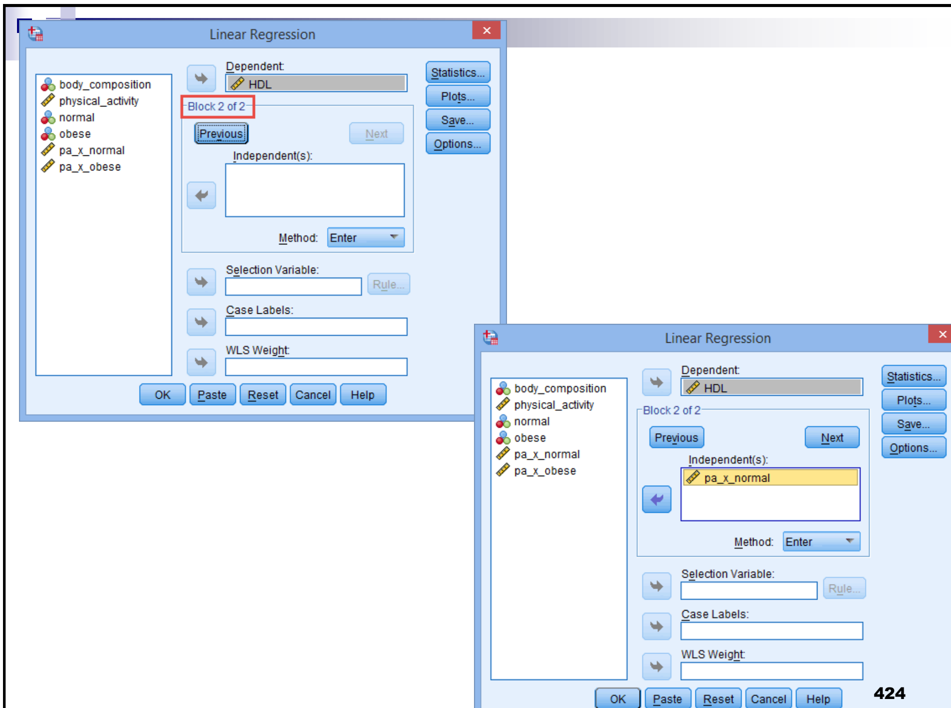
- The high-density lipoprotein cholesterol (HDL cholesterol, for short) is linked to good heart health. The higher the concentration of HDL cholesterol in the blood the better.
- It is known that **exercise** can increase HDL cholesterol concentration. **A researcher wants to understand whether this relationship is similar in normal weight and obese individuals**

Click **Analyze > Regression > Linear.**

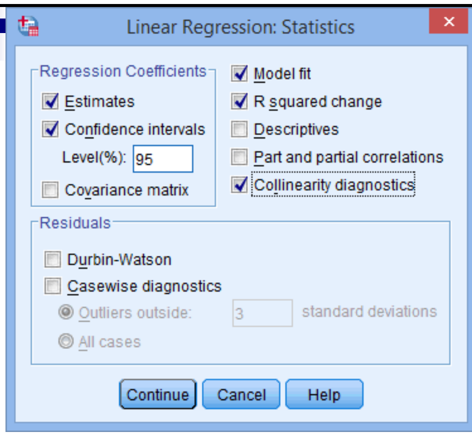


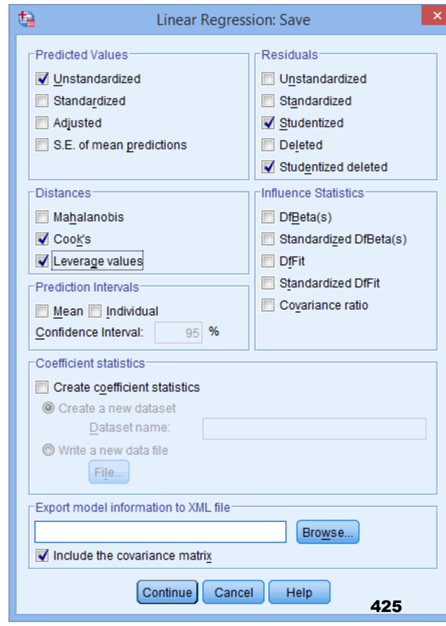

Linear Regression
body_composition
physical_activity
normal
obese
pa_x_normal
pa_x_obese
Dependent:
HDL
Block 1 of 1
Previous
Next
Independent(s):
physical_activity
normal
Method:
Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
OK
Paste
Reset
Cancel
Help

423


Linear Regression
body_composition
physical_activity
normal
obese
pa_x_normal
pa_x_obese
Dependent:
HDL
Block 2 of 2
Previous
Next
Independent(s):
Method:
Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
OK
Paste
Reset
Cancel
Help
Linear Regression
Dependent:
HDL
Block 2 of 2
Previous
Next
Independent(s):
pa_x_normal
Method:
Enter

424





425

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.937[a]	.878	.863	5.35679	.878	60.897	2	17	.000
2	.972[b]	.946	.936	3.67547	**.068**	**20.110**	**1**	**16**	**.000**

a. Predictors: (Constant), normal, physical_activity

b. Predictors: (Constant), normal, physical_activity, pa_x_normal

R^2 change 0.068 is 6.8% (i.e., .068 x 100 = 6.8%), which is the percentage increase in the variation explained by the addition of the interaction term

Coefficients[a]

Model		Unstandardized Coefficients: B	Std. Error	Standardized Coefficients: Beta	t	Sig.	95.0% Confidence Interval for B: Lower Bound	Upper Bound
1	(Constant)	27.155	2.475		10.972	.000	21.933	32.376
	physical_activity	.056	.013	.365	4.298	.000	.028	.083
	normal	24.412	2.396	.865	10.190	.000	19.358	29.467
2	(Constant)	**32.694**	2.100		15.570	.000	28.243	37.146
	physical_activity	**.016**	.013	.104	1.267	.223	-.011	.042
	normal	**13.353**	2.964	.473	4.506	.000	7.070	19.636
	pa_x_normal	**.080**	.018	.537	4.484	.000	.042	.117

a. Dependent Variable: HDL

Using the values obtained above, we can write the regression equation

HDL = 32.694 + (0.016 x physical_activity) + (13.353 x normal) + (0.080 x pa_x_normal)

426

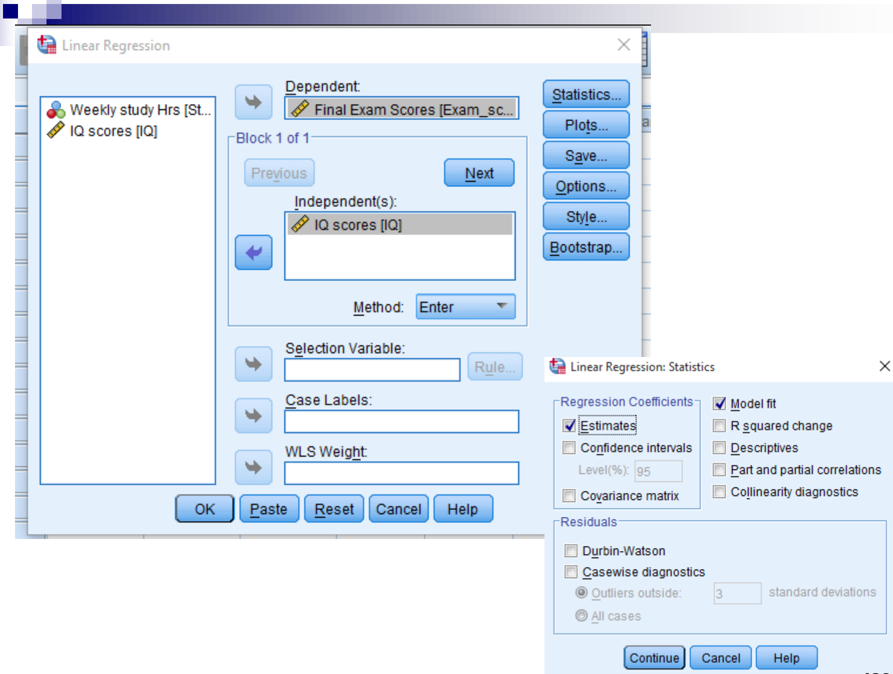
Mediator analysis using SPSS

Mediator

Is there an association of the independent variable with the mediator?

- Exam Score (DV)
- IQ scores (Predictor variable or Independent Variable)
- Weekly study hrs (Mediator)

Study_Hrs	Exam_scor...	IQ
6.00	35.00	122.00
7.00	39.00	124.00
5.00	45.00	118.00
2.00	66.00	117.00
3.00	46.00	119.00
2.00	59.00	112.00
2.00	60.00	109.00
4.00	52.00	104.00
3.00	46.00	120.00
4.00	48.00	107.00
1.00	58.00	105.00
6.00	45.00	124.00
7.00	39.00	116.00
6.00	35.00	121.00
4.00	38.00	125.00



429

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.634[a]	.402	.356	7.80947	.402	8.735	1	13	.011

a. Predictors: (Constant), IQ scores

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	532.758	1	532.758	8.735	.011[b]
	Residual	792.842	13	60.988		
	Total	1325.600	14			

a. Dependent Variable: Final Exam Scores

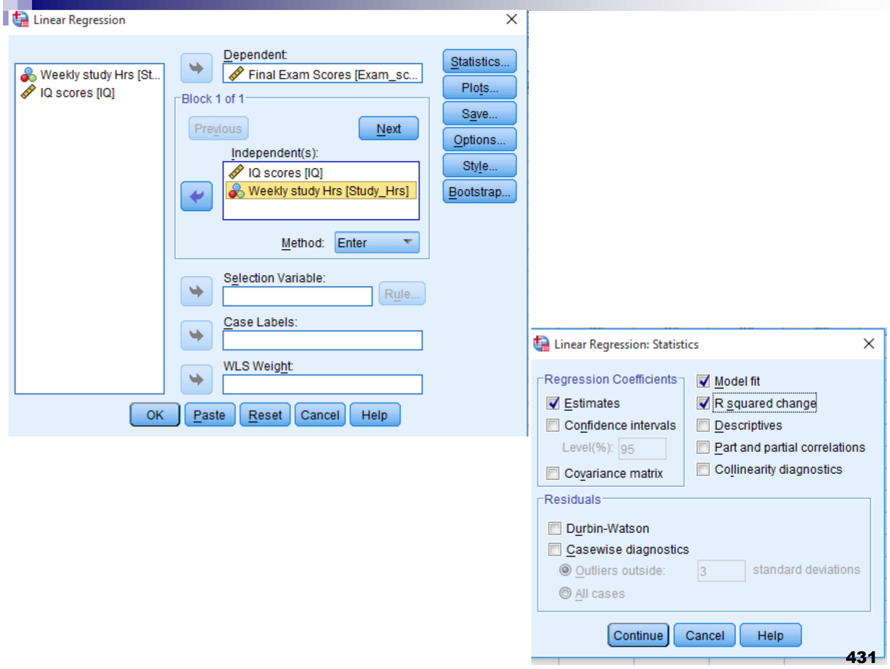
b. Predictors: (Constant), IQ scores

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	148.028	34.106		4.340	.001
	IQ scores	-.866	.293	-.634	-2.956	.011

a. Dependent Variable: Final Exam Scores

430



431

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.855[a]	.732	.687	5.44381	.732	16.365	2	12	.000

a. Predictors: (Constant), Weekly study Hrs, IQ scores

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	969.980	2	484.990	16.365	.000[b]
	Residual	355.620	12	29.635		
	Total	1325.600	14			

a. Dependent Variable: Final Exam Scores

b. Predictors: (Constant), Weekly study Hrs, IQ scores

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	100.498	26.802		3.750	.003
	IQ scores	-.334	.247	-.245	-1.355	.200
	Weekly study Hrs	-3.446	.897	-.694	-3.841	.002

a. Dependent Variable: Final Exam Scores

The study hrs is a mediator variable. When we add mediator variable and the strength or direction of IV and DV decreases

432

MANOVA: Multivariate Analysis of Variance using SPSS

Review of ANOVA: Univariate Analysis of Variance

- An univariate analysis of variance looks for the causal impact of a nominal level independent variable (factor) on a single, and interval level dependent variable.
  - **Analysis of Variance (ANOVA):** Required when there are three or more levels or conditions of the independent variable.
    - **What is the salary (dependent variable)** of different degree levels graduates in Malaysia (**MBBS, BDS, BSc Biotech, Business degree, Physiotherapist (Independent variables)**

434

Review of Factorial ANOVA

- **Two-way ANOVA is applied to a situation in which you have two independent nominal-level variables and one interval or better dependent variable**
- Each of the independent variables may have any number of levels or conditions (e.g., Treatment 1, Treatment 2, Treatment 3…… No Treatment)
- In a two-way ANOVA you will obtain 3 F ratios
  - One of these will tell you if your first independent variable has a significant *main effect* on the DV
  - A second will tell you if your second independent variable has a significant *main effect* on the DV
  - The third will tell you if the *interaction* of the two independent variables has a significant effect on the DV, that is, if the impact of one IV depends on the level of the other

Review: Factorial ANOVA Example

Tests of Between-Subjects Effects

Dependent Variable: TIMENET

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
Corrected Model	284.316[b]	5	56.863	2.380	.049	.170	11.902	.718
Intercept	1104.767	1	1104.767	46.246	.000	.444	46.246	1.000
COLLEGE	80.525	2	40.263	1.685	.194	.055	3.371	.341
MARRY	10.535	1	10.535	.441	.509	.008	.441	.100
COLLEGE * MARRY	171.312	2	85.656	3.586	.034	.110	7.171	.642
Error	1385.544	58	23.889					
Total	2800.500	64						
Corrected Total	1669.859	63						

a. Computed using alpha = .05

b. R Squared = .170 (Adjusted R Squared = .099)

Tests of Hypotheses:

(1) There is no significant main effect for education level (*F*(2, 58) = 1.685, p = .194, partial eta squared = .055) (red dots)

(2) There is no significant main effect for marital status (*F* (1, 58) = .441, p = .509, partial eta squared = .008)(green dots)

(3) There is a significant interaction effect of marital status and education level (*F* (2, 58) = 3.586, p = .034, partial eta squared = .110) (blue dots)

MANOVA: What Kinds of Hypotheses Can it Test?

- A MANOVA or multivariate analysis of variance is a way to test the hypothesis that one or more independent variables, or factors, have an effect on **a set of two or more dependent variables**
  - For example, you might wish to test the hypothesis that **sex and ethnicity** interact to influence a set of job-related outcomes including **attitudes toward co-workers, attitudes toward supervisors, feelings of belonging in the work environment, and identification with the corporate culture**
  - As another example, you might want to test the hypothesis that **three different methods of teaching** writing result in significant differences in **ratings of student creativity, student acquisition of grammar, and assessments of writing quality** by an independent panel of judges

437

MANOVA

- The one-way multivariate analysis of variance (one-way MANOVA) is used to determine whether there are any differences between **independent groups** on **more than one continuous dependent variables**.
- Example:
  - To understand whether there were differences in students' short-term and long-term recall of facts based on three different lengths of lecture (i.e., the two dependent variables are **"short-term memory recall" and "long-term memory recall",** whilst the independent variable is "lecture duration", which has four independent groups: "30 minutes", "60 minutes", "90 minutes" and "120 minutes").

438

Assumptions of MANOVA

1. **Multivariate normality**
   - All of the DVs must be distributed normally
   - Any linear combination of the DVs must be distributed normally
   - All subsets of the variables must have a multivariate normal distribution
2. **Homogeneity of the covariance matrices**
   - In ANOVA we talked about the need for the variances of the dependent variable to be equal across levels of the independent variable
     - In MANOVA, the univariate requirement of equal variances has to hold for each one of the dependent variables
3. Independence of observations
   - Subjects' scores on the dependent measures should not be influenced by or related to scores of other subjects in the condition or level

439

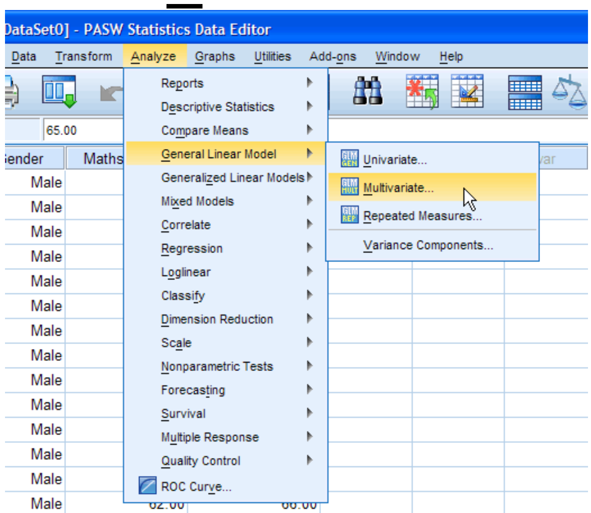
Example – MANOVA in SPSS

The students at a higher secondary school come from three different primary schools. The school Principal wanted to know whether there were academic differences between the students from the three different primary schools.

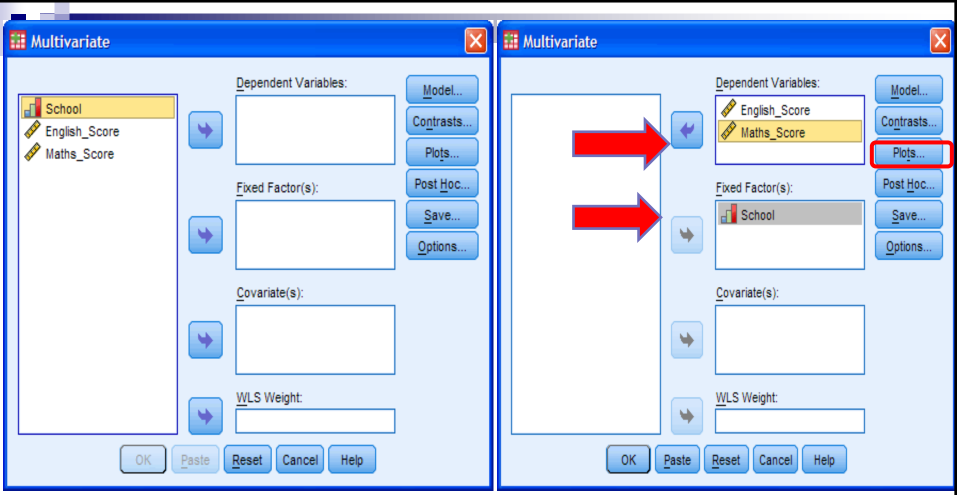
The principal randomly selected 20 students from School A, 20 students from School B and 20 students from School C, and measured their academic performance as assessed by the marks they received in **English and Maths exams**. **We have two dependent variables** were **"English score" and "Maths score",** whilst the **independent variable was "School", which consisted of three categories: "School A", "School B" and "School C".**

440

- Key in the data
- Click **Analyze > General Linear Model > Multivariate.**

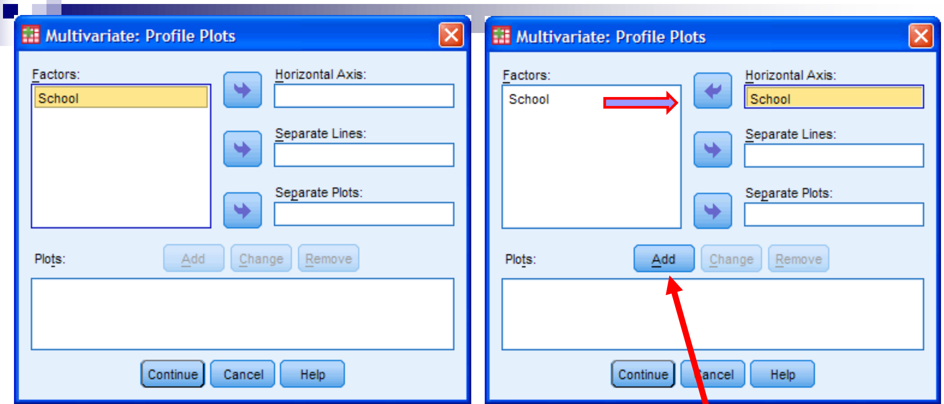


441



Transfer the independent variable, **School, into the Fixed Factor(s);** and transfer the dependent variables, **English_Score and Maths_Score, into the Dependent Variables**:

442

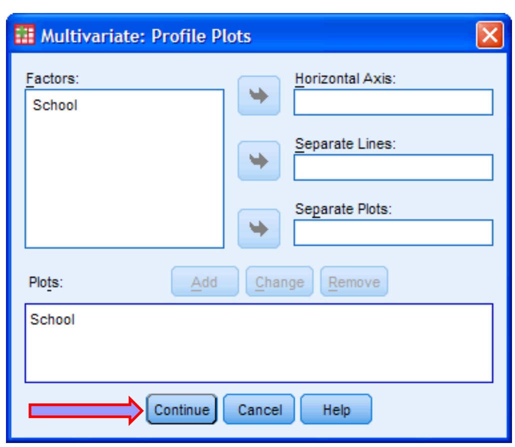


Transfer the independent variable, School, into the Horizontal Axis

Click Add

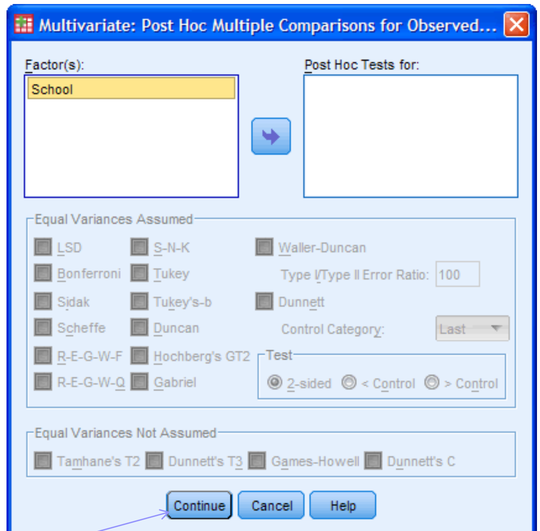
443

Click Add. You will see that "School" has been added to the Plots



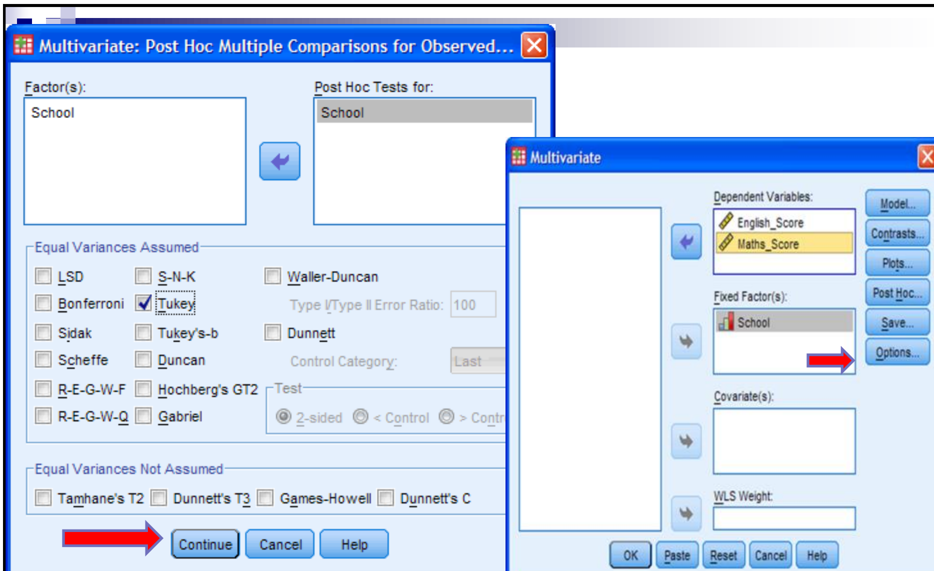
444

Click post hoc

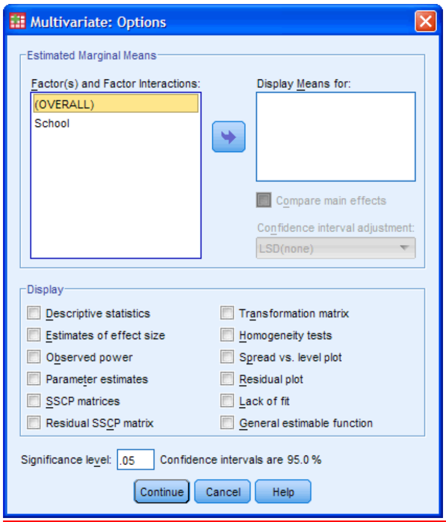


Transfer the independent variable, School, into the Post Hoc Tests for: and select the Tukey checkbox in the -Equal Variances Assumed- area

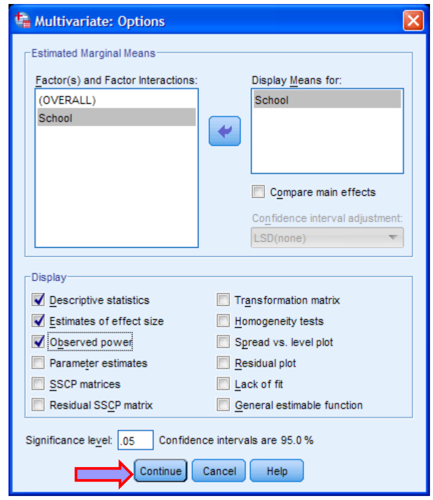
445



446



Transfer the independent variable, "School", from the Factor(s) and Factor Interactions: into the Display Means for:



- Select the Descriptive statistics, Estimates of effect size and Observed power checkboxes in the -Display- area.
- Click continue and click OK

Descriptive Statistics

	School	Mean	Std. Deviation	N
English_Score	School A	75.6000	8.22960	20
	School B	62.7000	9.10234	20
	School C	61.5500	7.14124	20
	Total	66.6167	10.30401	60
Maths_Score	School A	43.9000	8.46603	20
	School B	40.7500	8.16201	20
	School C	30.7500	7.71789	20
	Total	38.4667	9.78145	60

Multivariate Tests[d]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Intercept	Pillai's Trace	.989	2435.089[a]	2.000	56.000	.000	.989	4870.177	1.000
	Wilks' Lambda	.011	2435.089[a]	2.000	56.000	.000	.989	4870.177	1.000
	Hotelling's Trace	86.967	2435.089[a]	2.000	56.000	.000	.989	4870.177	1.000
	Roy's Largest Root	86.967	2435.089[a]	2.000	56.000	.000	.989	4870.177	1.000
School	Pillai's Trace	.616	12.681	4.000	114.000	.000	.308	50.724	1.000
	Wilks' Lambda	.450	13.735[a]	4.000	112.000	.000	.329	54.938	1.000
	Hotelling's Trace	1.075	14.782	4.000	110.000	.000	.350	59.128	1.000
	Roy's Largest Root	.915	26.072[c]	2.000	57.000	.000	.478	52.144	1.000

a. Exact statistic

b. Computed using alpha = .05

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

d. Design: Intercept + School

See the "Sig." value of .000, which means p < .0005. Therefore, we can conclude that this school's students academic performance was significantly dependent on which prior school they had attended (p < .0005).

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	English_Score	2434.233[a]	2	1217.117	18.114	.000	.389	36.228	1.000
	Maths_Score	1885.633[c]	2	942.817	14.295	.000	.334	28.591	.998
Intercept	English_Score	266266.817	1	266266.817	3962.769	.000	.986	3962.769	1.000
	Maths_Score	88781.067	1	88781.067	1346.134	.000	.959	1346.134	1.000
School	English_Score	2434.233	2	1217.117	18.114	.000	.389	36.228	1.000
	Maths_Score	1885.633	2	942.817	14.295	.000	.334	28.591	.998
Error	English_Score	3829.950	57	67.192					
	Maths_Score	3759.300	57	65.953					
Total	English_Score	272531.000	60						
	Maths_Score	94426.000	60						
Corrected Total	English_Score	6264.183	59						
	Maths_Score	5644.933	59						

a. R Squared = .389 (Adjusted R Squared = .367)

b. Computed using alpha = .05

c. R Squared = .334 (Adjusted R Squared = .311)

- The prior schooling has a statistically significant effect on both English ($F(2, 57) = 18.11$; $p < .0005$; partial $\eta 2 = .39$) and Maths scores ($F(2, 57) = 14.30$; $p < .0005$; partial $\eta 2 = .33$).

449

Multiple Comparisons

Tukey HSD

Dependent Variable	(I) School	(J) School	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
English_Score	School A	School B	12.9000*	2.59214	.000	6.6622	19.1378
		School C	14.0500*	2.59214	.000	7.8122	20.2878
	School B	School A	-12.9000*	2.59214	.000	-19.1378	-6.6622
		School C	1.1500	2.59214	.897	-5.0878	7.3878
	School C	School A	-14.0500*	2.59214	.000	-20.2878	-7.8122
		School B	-1.1500	2.59214	.897	-7.3878	5.0878
Maths_Score	School A	School B	3.1500	2.56812	.443	-3.0300	9.3300
		School C	13.1500*	2.56812	.000	6.9700	19.3300
	School B	School A	-3.1500	2.56812	.443	-9.3300	3.0300
		School C	10.0000*	2.56812	.001	3.8200	16.1800
	School C	School A	-13.1500*	2.56812	.000	-19.3300	-6.9700
		School B	-10.0000*	2.56812	.001	-16.1800	-3.8200

Based on observed means.
The error term is Mean Square(Error) = 65.953.

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

- The mean scores for English were statistically significantly different between School A and School B ($p < .0005$), and School A and School C ($p < .0005$), but not between School B and School C ($p = .897$).
- Mean maths scores were statistically significantly different between School A and School C ($p < .0005$), and School B and School C ($p = .001$), but not between School A and School B ($p = .443$).

450

Discriminant analysis using SPSS

Discriminant Analysis

- Discriminant analysis is used to determine which variables discriminate between two or more naturally occurring groups.
- Discriminant analysis characterize the relationship between a set of IVs with a categorical DV with relatively few categories
  - It creates a linear combination of the IVs that best characterizes the differences among the groups
  - Predictive discriminant analysis focus on creating a rule to predict group membership
  - Descriptive DA studies the relationship between the DV and the IVs.

452

Discriminant Analysis

For example, an educational researcher may want to investigate which variables discriminate between high school graduates who decide to

(1) go to college,
(2) attend a trade or professional school, or
(3) seek no further training or education.

For that purpose the researcher could collect data on numerous variables prior to students' graduation. After graduation, most students will naturally fall into one of the three categories. Discriminant Analysis could then be used to determine which variable(s) are the best predictors of students' subsequent educational choice.

Discriminant Analysis

- For example, a medical researcher may record different variables relating to patients' backgrounds in order to learn which variables best predict whether a patient is likely to recover completely (group 1), partially (group 2), or not at all (group 3).
- A biologist could record different characteristics of similar types (groups) of flowers, and then perform a discriminant function analysis to determine the set of characteristics that allows for the best discrimination between the types.

Discriminant Analysis

- Possible applications:
  - Whether a bank should offer a loan to a new customer?
  - Which customer is likely to buy?
  - Identify patients who may be at high risk for problems after surgery

455

Promotional Campaigns

- Identify groups based on their response to promotional campaigns
  - One group purchases a lot on promotion
  - Other does not
- Identify characteristics that distinguish these two groups

456

Discriminant Analysis

- How does it work?
  - Assume the population of interest is composed of distinct populations
  - Assume the IVs follows multivariate normal distribution
  - DS seek a linear combination of the IVs that best separate the populations

457

Discriminat Analysis

- For example
  - A borrower is looking for bank loan
  - There are two groups in customers (Good and Bad) bank will have
  - For good customers bank will give the loan
  - For bad customers extend the loan after collecting good collaterals and guarantee form from a reputed rich person
  - **On what basis will we assign a new comer to the bank to borrow?**

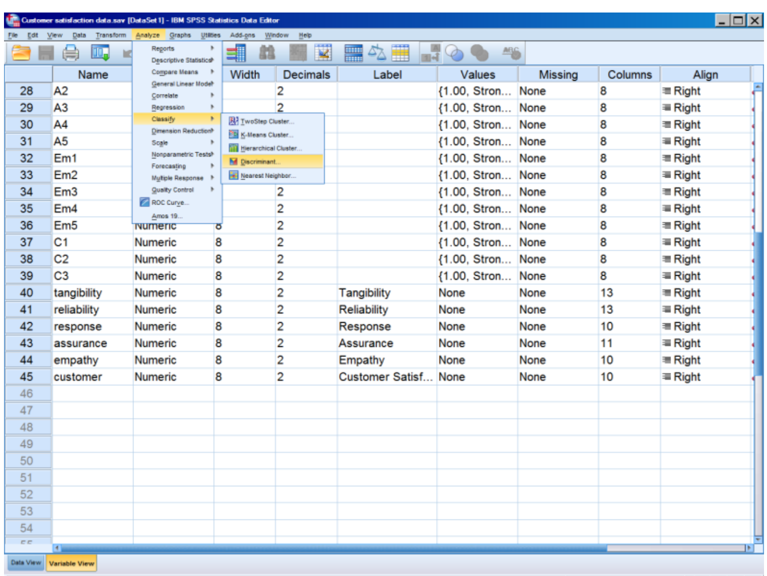
458

Discriminant analysis

- Computing the centroids is the main objective of discriminant analysis
- We will use the customer satisfaction data file
- There are two groups
- If a new customer is to be assigned in either Islamic banking or conventional banking group how well it discriminates
- What are the centroids? What is the function it forms? Let us test.

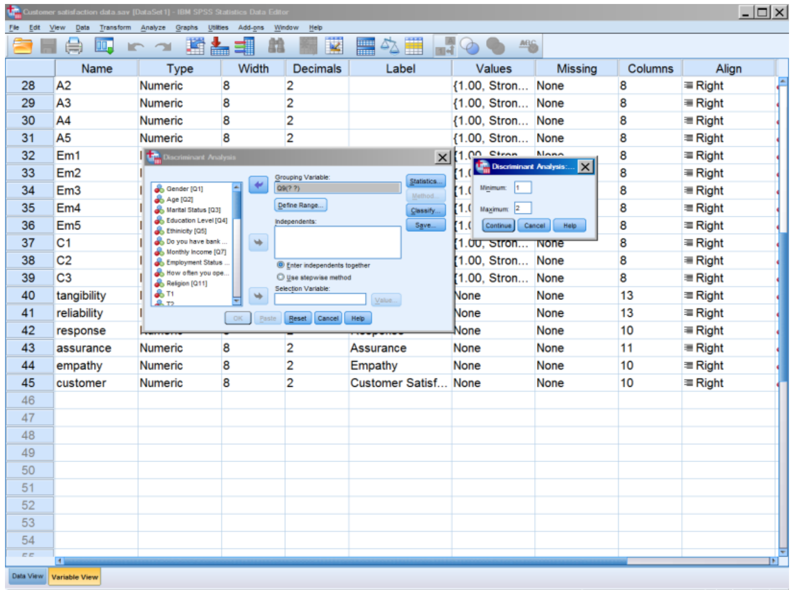
459

Discriminant analysis

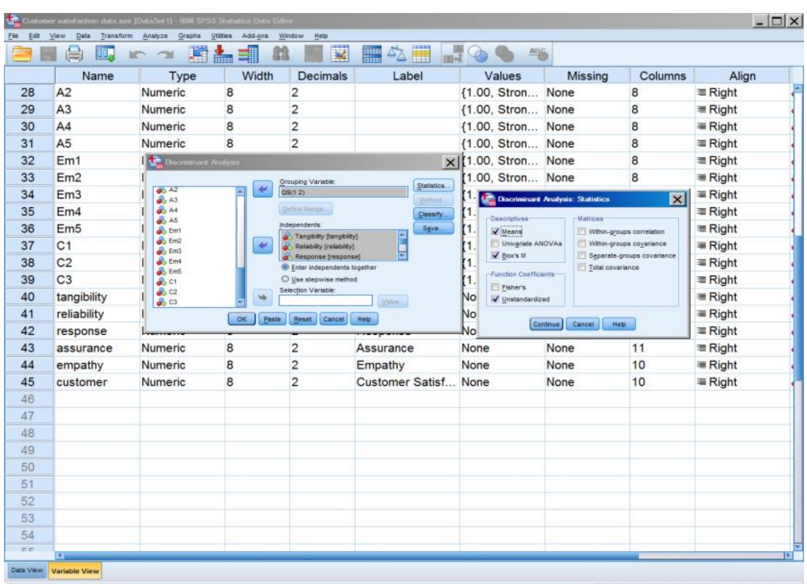


460

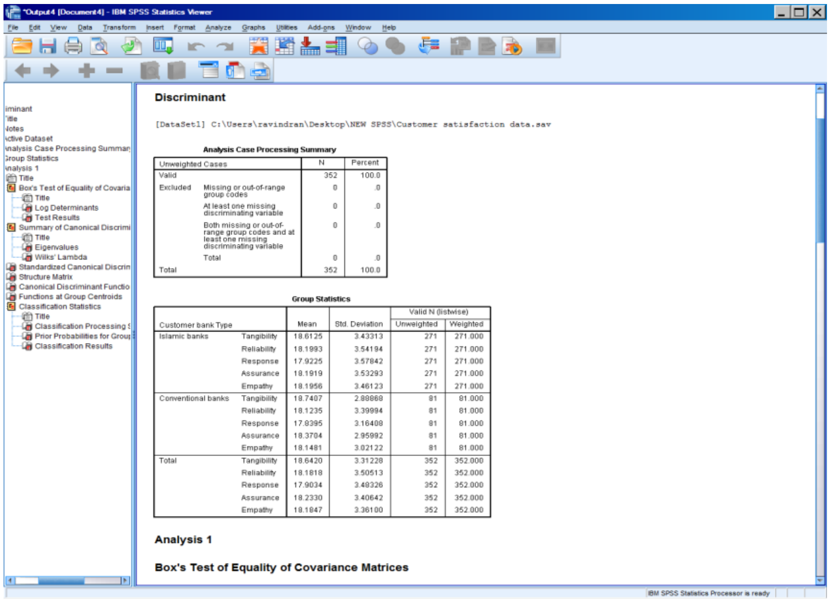
Discriminant analysis



461



462



Discriminant

[DataSet1] C:\Users\ravindran\Desktop\NEW SPSS\Customer satisfaction data.sav

Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		352	100.0
Excluded	Missing or out-of-range group codes	0	.0
	At least one missing discriminating variable	0	.0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	.0
	Total	0	.0
Total		352	100.0

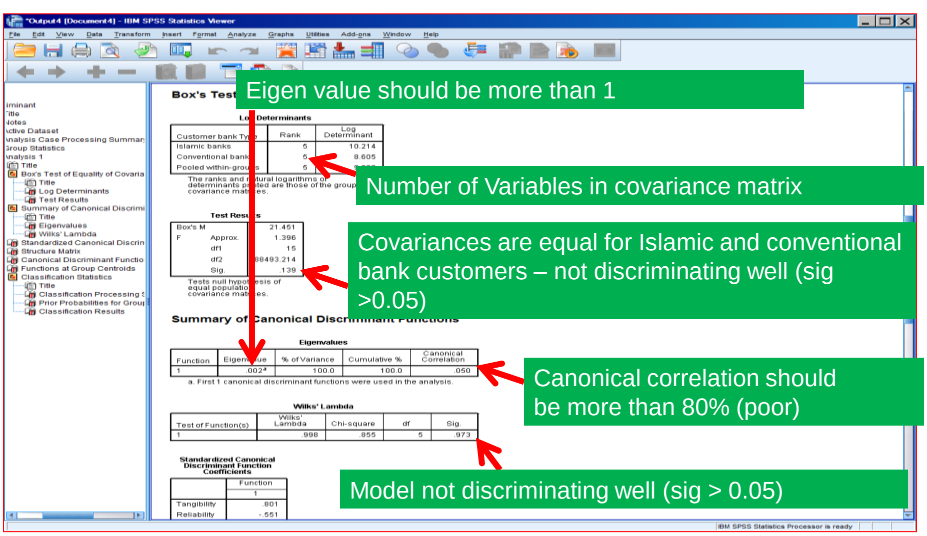
Group Statistics

Customer bank Type		Mean	Std. Deviation	Valid N (listwise) Unweighted	Valid N (listwise) Weighted
Islamic banks	Tangibility	18.6125	3.43313	271	271.000
	Reliability	18.1993	3.54194	271	271.000
	Response	17.9225	3.57842	271	271.000
	Assurance	18.1919	3.53293	271	271.000
	Empathy	18.1956	3.46123	271	271.000
Conventional banks	Tangibility	18.7407	2.88868	81	81.000
	Reliability	18.1235	3.39994	81	81.000
	Response	17.8395	3.16408	81	81.000
	Assurance	18.3704	2.95992	81	81.000
	Empathy	18.1481	3.02122	81	81.000
Total	Tangibility	18.6420	3.31228	352	352.000
	Reliability	18.1818	3.50513	352	352.000
	Response	17.9034	3.48326	352	352.000
	Assurance	18.2330	3.40642	352	352.000
	Empathy	18.1847	3.36100	352	352.000

Analysis 1

Box's Test of Equality of Covariance Matrices

463



Log Determinants

Customer bank Type	Rank	Log Determinant
Islamic banks	5	10.214
Conventional banks	5	8.605
Pooled within-groups	5	[illegible]

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Test Results

Box's M		21.451
F	Approx.	1.396
	df1	15
	df2	88493.214
	Sig.	.139

Tests null hypothesis of equal population covariance matrices.

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.002[a]	100.0	100.0	.050

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

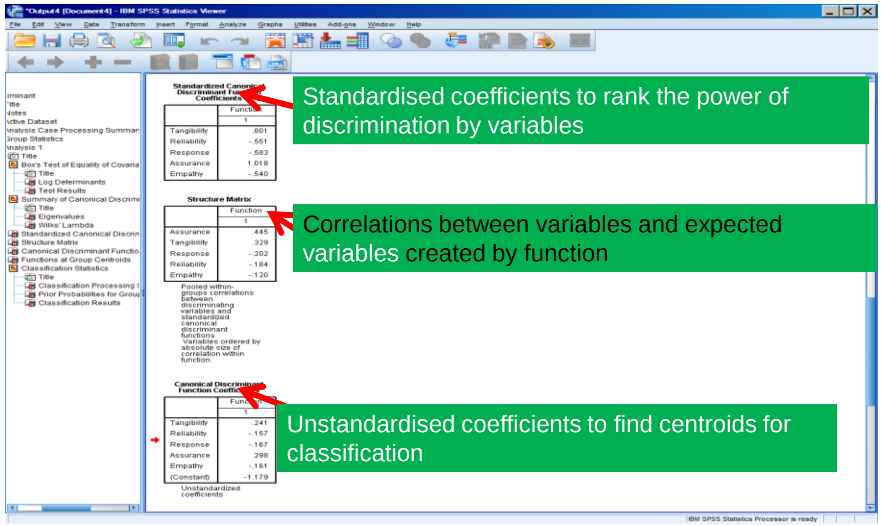
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.998	.855	5	.973

Standardized Canonical Discriminant Function Coefficients

	Function 1
Tangibility	.801
Reliability	-.551

464

Discriminant output - 3

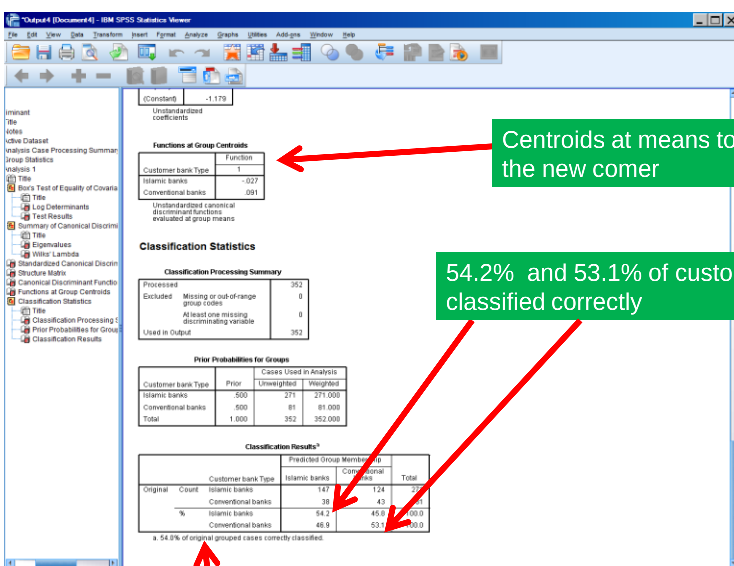


Standardised coefficients to rank the power of discrimination by variables

Correlations between variables and expected variables created by function

Unstandardised coefficients to find centroids for classification

Discriminant output - 4



Centroids at means to classify the new comer

54.2% and 53.1% of customers only classified correctly

54% classification correct - poor (95% and above is good classification)

Discriminant analysis (Interpretation)

•1. Box's M – Compares the covariance matrices Sig (p value) should be less than 0.05 to confirm that covariance matrices are different

•2. The Eigen Value explains the power of discrimination. It should be more than 1

•3. The canonical correlation should be more than 80%

4.Wilk's Lambda is like $1-R^2$. The Chi – square should be significant to prove that the function discriminates well

467

Discriminant analysis (Interpretation)

•5. The Standardized canonical discriminant function coefficients can be used to rank the variables from best discriminator to worst discriminator

•6. The structure matrix gives the correlations between variables and expected variables produced by the discriminant function.

•7. The unstandardized coefficients form the function and also useful to find centroids

•8. Centroids average will help in classifying the new comers

468

Cluster analysis using SPSS

Cluster Analysis

- Cluster analysis is an exploratory data analysis technique design to reveal groups
- How?
  - By distance: close together observations should be in the same group, and observations in the groups should be far apart
- Applications:
  - Plants and animals into ecological groups
  - Companies for product usage

470

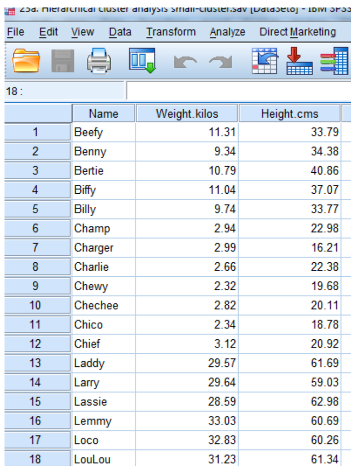
Cluster Analysis

- Two types of method
  - **Hierarchical:** requires observations to remain together once they have joint in a cluster
    - Complete linkage
    - Between groups average linkage
    - Ward's method
    - **It requires the number of clusters to be specified in advance, and the initial number chosen may split natural groupings or combine two or more groups that are rather different from each other.**
  - **Nonhierarchical:** no such requirement
    - Research must pick a number of clusters to **run (K-means algorithm)**

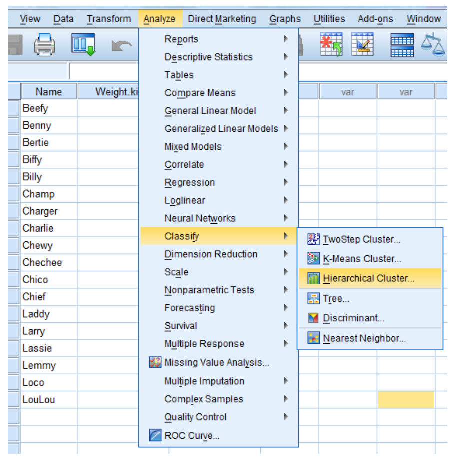
Cluster Analysis

- Recommendations:
  - For relative small samples, use hierarchical (less than a few hundred)
  - For large samples, use K-means

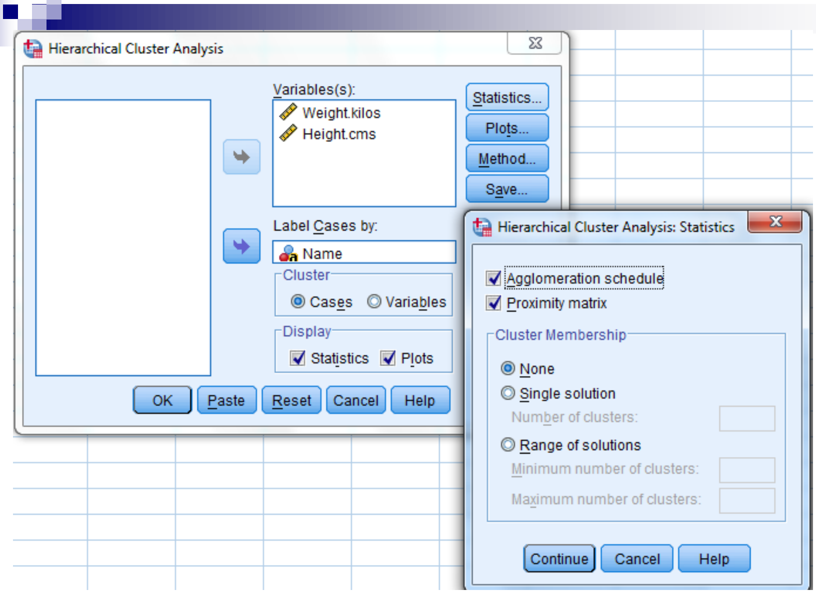
Hierarchical Cluster analysis – Example - Dogs species weight and height



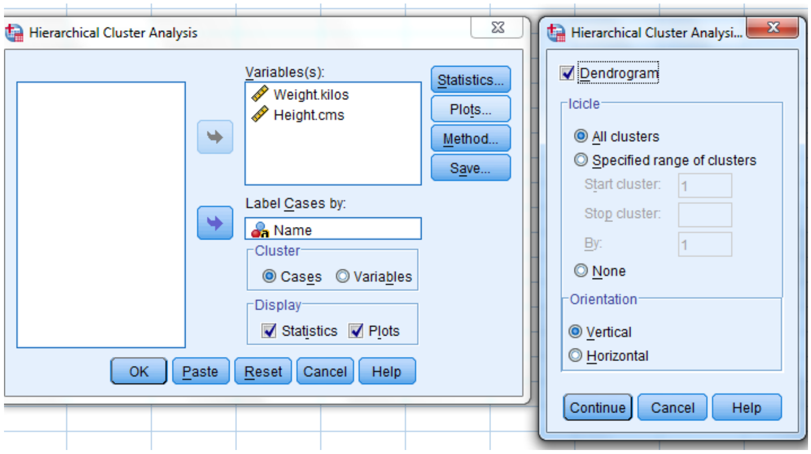
	Name	Weight.kilos	Height.cms
1	Beefy	11.31	33.79
2	Benny	9.34	34.38
3	Bertie	10.79	40.86
4	Biffy	11.04	37.07
5	Billy	9.74	33.77
6	Champ	2.94	22.98
7	Charger	2.99	16.21
8	Charlie	2.66	22.38
9	Chewy	2.32	19.68
10	Chechee	2.82	20.11
11	Chico	2.34	18.78
12	Chief	3.12	20.92
13	Laddy	29.57	61.69
14	Larry	29.64	59.03
15	Lassie	28.59	62.98
16	Lemmy	33.03	60.69
17	Loco	32.83	60.26
18	LouLou	31.23	61.34



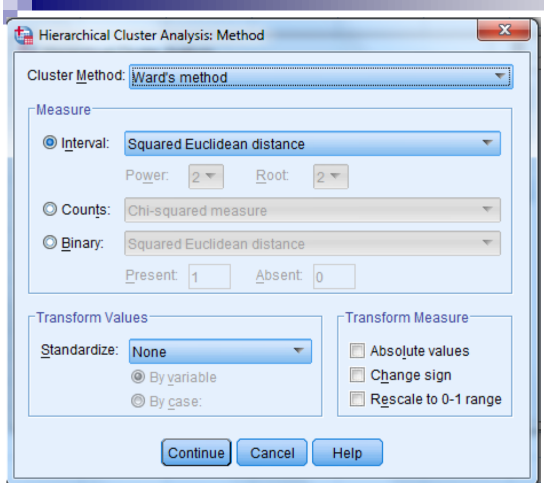
473



474


Hierarchical Cluster Analysis
Variables(s):
Weight.kilos
Height.cms
Statistics...
Plots...
Method...
Save...
Label Cases by:
Name
Cluster
Cases
Variables
Display
Statistics
Plots
OK
Paste
Reset
Cancel
Help
Hierarchical Cluster Analysi...
Dendrogram
Icicle
All clusters
Specified range of clusters
Start cluster:
1
Stop cluster:
By:
1
None
Orientation
Vertical
Horizontal
Continue
Cancel
Help

475


Hierarchical Cluster Analysis: Method
Cluster Method:
Ward's method
Measure
Interval:
Squared Euclidean distance
Power:
2
Root:
2
Counts:
Chi-squared measure
Binary:
Squared Euclidean distance
Present:
1
Absent:
0
Transform Values
Standardize:
None
By variable
By case:
Transform Measure
Absolute values
Change sign
Rescale to 0-1 range
Continue
Cancel
Help

476

Dissimilarity matrix

Proximity Matrix

Case	Squared Euclidean Distance																	
	1:Beefy	2:Benny	3:Bertie	4:Biffy	5:Billy	6:Champ	7:Charger	8:Charlie	9:Chewy	10:Chechee	11:Chico	12:Chief	13:Laddy	14:Larry	15:Lassie	16:Lemmy	17:Loco	18:LouLou
1:Beefy	.000	4.229	50.255	10.831	2.465	186.913	378.279	205.011	279.912	259.223	305.761	232.713	1111.838	973.047	1150.654	1195.368	1163.771	1155.809
2:Benny	4.229	.000	44.093	10.126	.532	170.920	370.471	188.622	265.370	246.143	292.360	219.860	1155.089	1019.713	1188.522	1253.432	1221.554	1206.014
3:Bertie	50.255	44.093	.000	14.427	51.371	381.317	668.462	407.607	520.333	494.083	558.929	456.432	786.577	685.471	806.134	887.847	862.122	837.224
4:Biffy	10.831	10.126	14.427	.000	12.580	264.138	499.942	286.021	378.451	355.210	410.214	323.549	949.505	828.202	979.331	1041.465	1012.580	996.669
5:Billy	2.465	.532	51.371	12.580	.000	162.664	353.916	179.859	253.585	234.482	279.460	208.947	1172.755	1034.078	1208.547	1267.110	1234.868	1221.925
6:Champ	186.913	170.920	381.317	264.138	162.664	.000	45.835	.438	11.274	8.251	18.000	4.276	2207.621	2012.493	2257.922	2327.452	2283.210	2271.814
7:Charger	378.279	370.471	668.462	499.942	353.916	45.835	.000	38.178	12.490	15.239	7.027	22.201	2774.927	2543.775	2842.793	2880.872	2830.828	2834.215
8:Charlie	205.011	188.622	407.607	286.021	179.859	.438	38.178	.000	7.406	5.178	13.062	2.343	2269.424	2071.143	2320.725	2389.993	2345.123	2334.127
9:Chewy	279.912	265.370	520.333	378.451	253.585	11.274	12.490	7.406	.000	.435	.810	2.178	2507.403	2294.805	2565.003	2624.924	2577.596	2571.344
10:Chechee	259.223	246.143	494.083	355.210	234.482	8.251	15.239	5.178	.435	.000	1.999	.746	2444.459	2234.079	2501.930	2559.380	2512.623	2507.041
11:Chico	305.761	292.360	558.929	410.214	279.460	18.000	7.027	13.062	.810	1.999	.000	5.188	2582.741	2365.353	2642.702	2698.324	2650.230	2645.986
12:Chief	232.713	219.860	456.432	323.549	208.947	4.276	22.201	2.343	2.178	.746	5.188	.000	2361.795	2155.683	2417.764	2476.261	2430.320	2423.949
13:Laddy	1111.838	1155.089	786.577	949.505	1172.755	2207.621	2774.927	2269.424	2507.403	2444.459	2582.741	2361.795	.000	7.080	2.624	12.972	12.672	2.878
14:Larry	973.047	1019.713	685.471	828.202	1034.078	2012.493	2543.775	2071.143	2294.805	2234.079	2365.353	2155.683	7.080	.000	16.705	14.248	11.689	7.864
15:Lassie	1150.654	1188.522	806.134	979.331	1208.547	2257.922	2842.793	2320.725	2565.003	2501.930	2642.702	2417.764	2.624	16.705	.000	24.958	25.376	9.659
16:Lemmy	1195.368	1253.432	887.847	1041.465	1267.110	2327.452	2880.872	2389.993	2624.924	2559.380	2698.324	2476.261	12.972	14.248	24.958	.000	.225	3.663
17:Loco	1163.771	1221.554	862.122	1012.580	1234.868	2283.210	2830.828	2345.123	2577.596	2512.623	2650.230	2430.320	12.672	11.689	25.376	.225	.000	3.726
18:LouLou	1155.809	1206.014	837.224	996.669	1221.925	2271.814	2834.215	2334.127	2571.344	2507.041	2645.986	2423.949	2.878	7.864	9.659	3.663	3.726	.000

This is a dissimilarity matrix

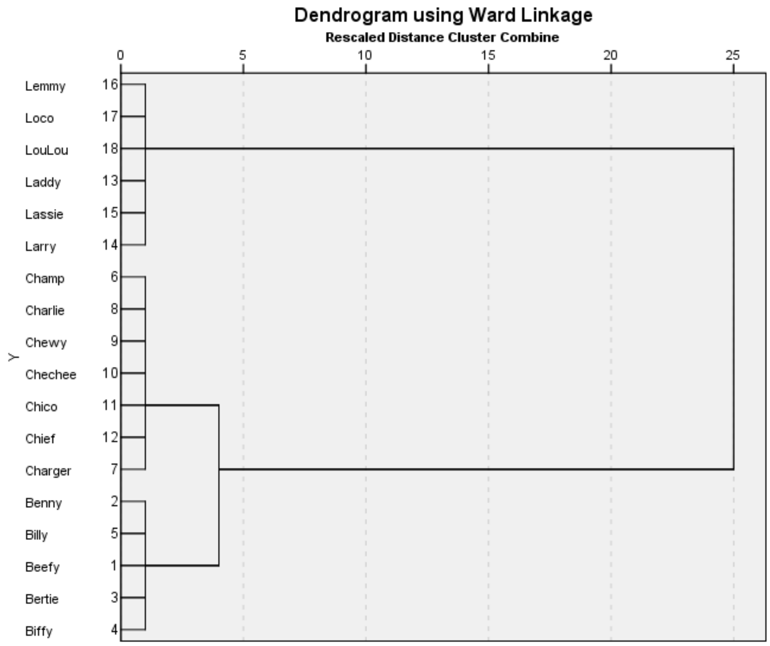
- The more value more the dissimilarty
- If the value is low the organisms are very close to each other.

477

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	16	17	.112	0	0	9
2	9	10	.330	0	0	5
3	6	8	.549	0	0	12
4	2	5	.815	0	0	8
5	9	11	1.679	2	0	7
6	13	15	2.991	0	0	11
7	9	12	4.749	5	0	12
8	1	2	6.892	0	4	15
9	16	18	9.317	1	0	13
10	3	4	16.531	0	0	15
11	13	14	24.022	6	0	13
12	6	9	34.561	3	7	14
13	13	16	49.276	11	9	17
14	6	7	67.473	12	0	16
15	1	3	98.032	8	10	16
16	1	6	1001.275	15	14	17
17	1	13	8167.574	16	13	0

478



479

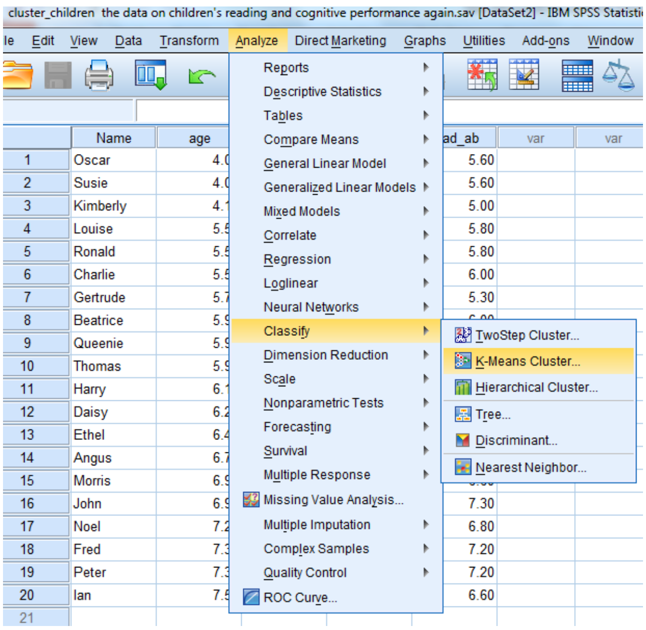
K means cluster analysis

*23 cluster_children.sav [DataSet1] - IBM SPSS Statistics Data Editor

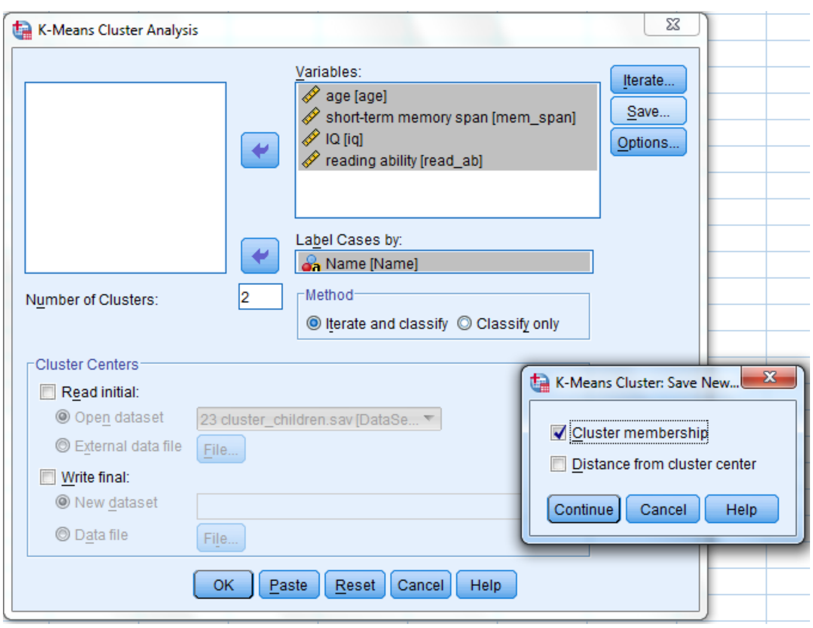
le Edit View Data Transform Analyze Direct Marketing Graphs U

	Name	age	mem_span	iq	read_ab
1	Oscar	4.00	4.20	101.00	5.60
2	Susie	4.00	4.20	101.00	5.60
3	Kimberly	4.10	3.90	108.00	5.00
4	Louise	5.50	4.20	90.00	5.80
5	Ronald	5.50	4.20	90.00	5.80
6	Charlie	5.50	4.10	105.00	6.00
7	Gertrude	5.70	3.60	88.00	5.30
8	Beatrice	5.90	4.00	90.00	6.00
9	Queenie	5.90	4.00	90.00	6.00
10	Thomas	5.90	4.00	90.00	6.00
11	Harry	6.15	5.00	95.00	6.40
12	Daisy	6.20	4.80	98.00	6.60
13	Ethel	6.40	5.00	106.00	7.00
14	Angus	6.70	4.40	95.00	7.20
15	Morris	6.90	4.50	91.00	6.60
16	John	6.90	5.00	104.00	7.30
17	Noel	7.20	5.00	92.00	6.80
18	Fred	7.30	5.50	100.00	7.20
19	Peter	7.30	5.50	100.00	7.20
20	Ian	7.50	5.40	96.00	6.60

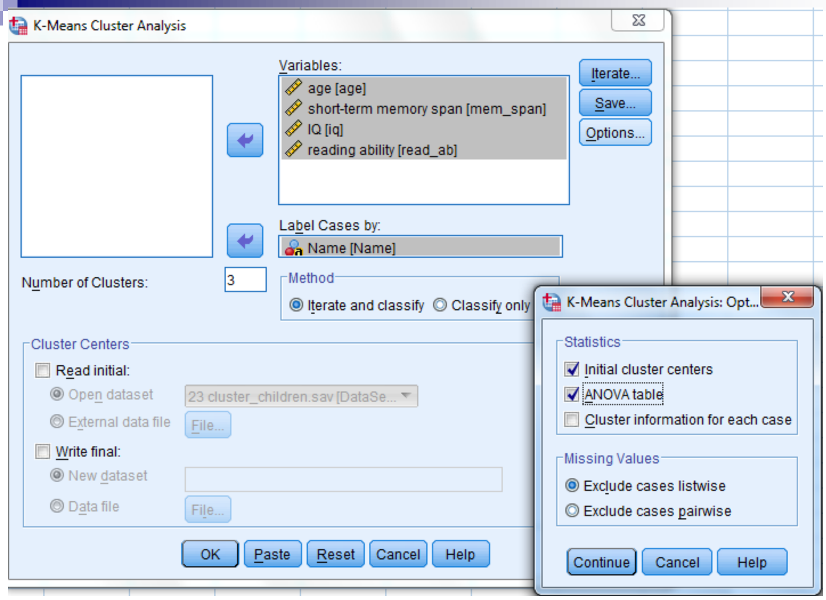
480


cluster_children the data on children's reading and cognitive performance again.sav [DataSet2] - IBM SPSS Statisti
le Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Quality Control
ROC Curve...
TwoStep Cluster...
K-Means Cluster...
Hierarchical Cluster...
Tree...
Discriminant...
Nearest Neighbor...
Name
age
ad_ab
var
var
1 Oscar 4.0 5.60
2 Susie 4.0 5.60
3 Kimberly 4.1 5.00
4 Louise 5.5 5.80
5 Ronald 5.5 5.80
6 Charlie 5.5 6.00
7 Gertrude 5.7 5.30
8 Beatrice 5.9
9 Queenie 5.9
10 Thomas 5.9
11 Harry 6.1
12 Daisy 6.2
13 Ethel 6.4
14 Angus 6.7
15 Morris 6.9
16 John 6.9 7.30
17 Noel 7.2 6.80
18 Fred 7.3 7.20
19 Peter 7.3 7.20
20 Ian 7.5 6.60
21

481


K-Means Cluster Analysis
Variables:
age [age]
short-term memory span [mem_span]
IQ [iq]
reading ability [read_ab]
Iterate...
Save...
Options...
Label Cases by:
Name [Name]
Number of Clusters:
2
Method
Iterate and classify
Classify only
Cluster Centers
Read initial:
Open dataset
23 cluster_children.sav [DataSe...
External data file
File...
Write final:
New dataset
Data file
File...
OK
Paste
Reset
Cancel
Help
K-Means Cluster: Save New...
Cluster membership
Distance from cluster center
Continue
Cancel
Help

482



483

Initial Cluster Centers

	Cluster		
	1	2	3
age	6.20	5.70	4.10
short-term memory span	4.80	3.60	3.90
IQ	98.00	88.00	108.00
reading ability	6.60	5.30	5.00

Iteration History[a]

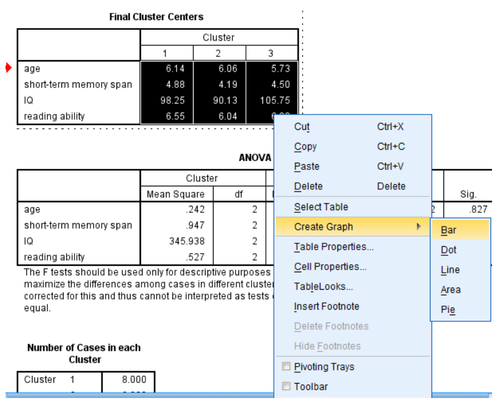
Iteration	Change in Cluster Centers		
	1	2	3
1	.272	2.353	3.133
2	.000	.000	.000

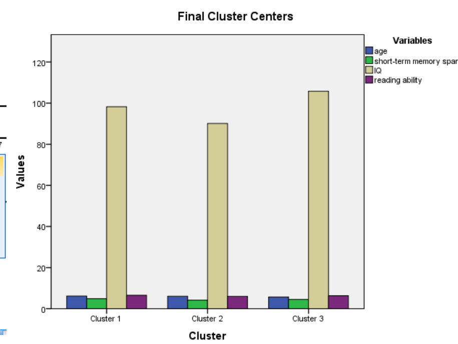
a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is 10.168.

484

Final Cluster Centers

	Cluster		
	1	2	3
age	6.14	6.06	5.73
short-term memory span	4.88	4.19	4.50
IQ	98.25	90.13	105.75
reading ability	6.55	6.04	6.33





485

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
age	.242	2	1.258	17	.192	.827
short-term memory span	.947	2	.261	17	3.622	.049
IQ	345.938	2	3.831	17	90.302	.000
reading ability	.527	2	.466	17	1.130	.346

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Age and reading ability variable is not significant to make the cluster. Short term memory span and IQ is significantly good for having cluster formation.

Number of Cases in each Cluster

Cluster	1	8.000
	2	8.000
	3	4.000
Valid		20.000
Missing		.000

486

*cluster_children the data on children's reading and cognitive performance again.sav [DataSet

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities A

5 :

	Name	age	mem_span	iq	read_ab	QCL_1	QCL_2
1	Oscar	4.00	4.20	101.00	5.60	1	3.67645
2	Susie	4.00	4.20	101.00	5.60	1	3.67645
3	Kimberly	4.10	3.90	108.00	5.00	3	3.13349
4	Louise	5.50	4.20	90.00	5.80	2	.62337
5	Ronald	5.50	4.20	90.00	5.80	2	.62337
6	Charlie	5.50	4.10	105.00	6.00	3	.93742
7	Gertrude	5.70	3.60	88.00	5.30	2	2.35289
8	Beatrice	5.90	4.00	90.00	6.00	2	.28035
9	Queenie	5.90	4.00	90.00	6.00	2	.28035
10	Thomas	5.90	4.00	90.00	6.00	2	.28035
11	Harry	6.15	5.00	95.00	6.40	1	3.25587
12	Daisy	6.20	4.80	98.00	6.60	1	.27164
13	Ethel	6.40	5.00	106.00	7.00	3	1.10623
14	Angus	6.70	4.40	95.00	7.20	1	3.39412
15	Morris	6.90	4.50	91.00	6.60	2	1.37153
16	John	6.90	5.00	104.00	7.30	3	2.37566
17	Noel	7.20	5.00	92.00	6.80	2	2.45990
18	Fred	7.30	5.50	100.00	7.20	1	2.28310
19	Peter	7.30	5.50	100.00	7.20	1	2.28310
20	Ian	7.50	5.40	96.00	6.60	1	2.67956

- Cluster membership
- **How much difference from cluster center**

487

Exploratory factor analysis (Principal Components Analysis) using SPSS

Factor analysis or Principal components analysis

- Principal components analysis is a variable-reduction technique. Data reduction technique
- Its aim is to reduce a larger set of variables into a smaller set of 'artificial' variables, called **'principal components',**
  - Let's say, we have 500 questions on a survey we designed to measure persistence. We want to reduce the number of questions so that it does not take someone 3 hours to complete the survey.
  - It would be appropriate to use PCA to reduce the number of questions by identifying and removing redundant questions. For instance, if question 122 and question 356 are virtually identical (i.e. they ask the exact same thing but in different ways), then one of them is not necessary.
  - The PCA process allows us to reduce the number of questions or variables down to their PRINCIPAL COMPONENTS.

489

Assumptions

- **Assumption #1:** You have **multiple variables** that should be measured at the **continuous level** or **ordinal variables**
- **Assumption #2:** There needs to be a **linear relationship between all variables**. The reason for this assumption is that a PCA is based on **Pearson correlation coefficients**, and as such, there needs to be a linear relationship between the variables.
- **Assumption #3:** You should have **sampling adequacy**, which simply means that for PCA to produce a reliable result, large enough sample sizes are required.
- **Assumption #4:** There should be **no significant outliers**.

490

Example

- A company director wanted to hire another employee for his company and was looking for someone who would **display high levels of motivation, dependability, enthusiasm and commitment**
- In order to select candidates for interview, he prepared a questionnaire consisting of **25 questions** that he believed might answer whether he had the correct candidates. He administered this questionnaire to 315 potential candidates. The questions were phrased such that these qualities should be represented in the questions.
- The director wanted to determine a score for each candidate so that these scores could be used to grade the potential recruits.

491

Example:
Wechsler Intelligence Scale for Children

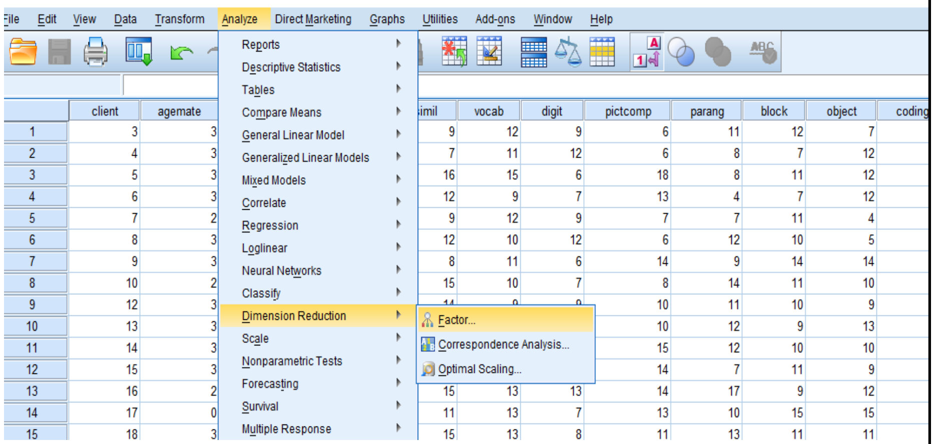
- The most common assessment instrument used by psychologists is the Wechsler Intelligence Scale for Children.
- The test is divided into two sections with each section containing a number of subtests.

Verbal Scale	Performance Scale
Information	Picture Compilation
Similarities	Coding
Arithmetic	Picture Arrangement
Vocabulary	Block Design
Comprehension	Object Assembly

Verbal Scale	Performance Scale
Digit Span	Symbol Search
	Mazes

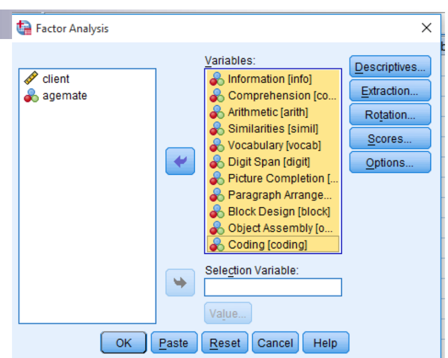
492

Click Analyze > Dimension Reduction > Factor...

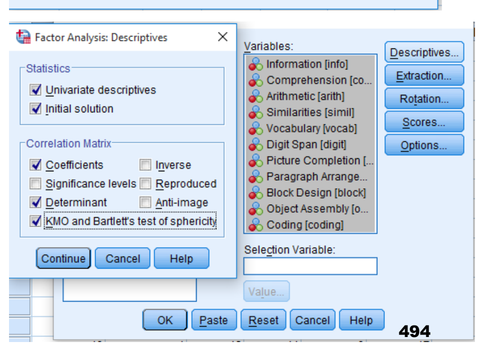


493

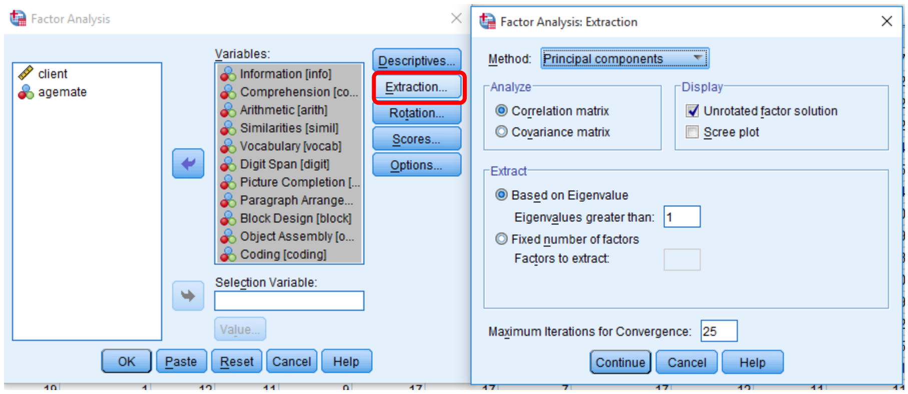
Transfer all the variables you want included in the analysis, into the Variables box.

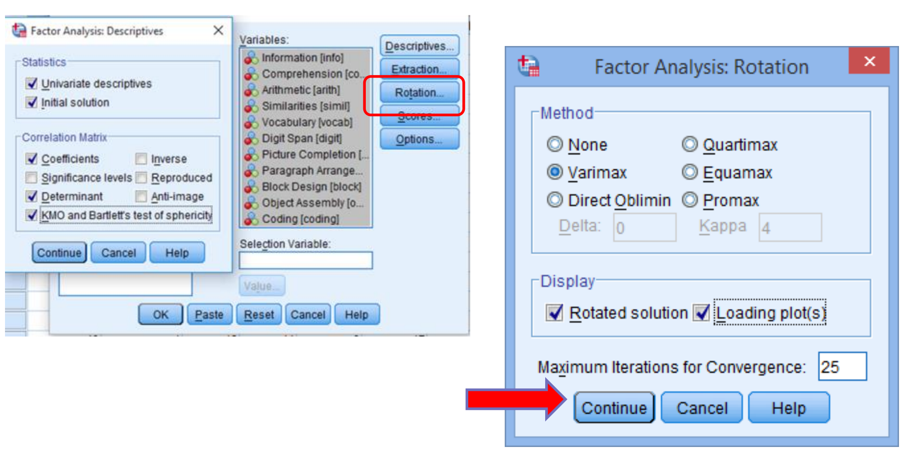


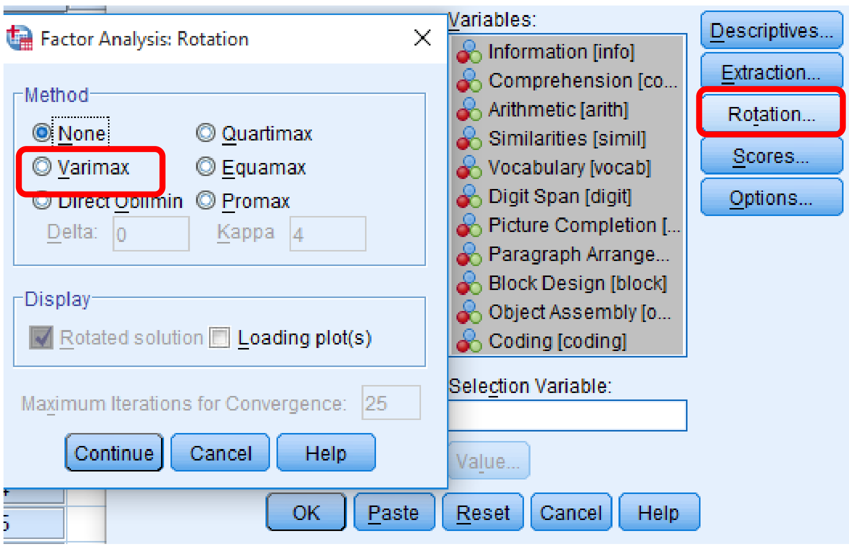
Click Initial solution in the -Statistics- area), also check Coefficients, **KMO and Bartlett's test of sphericity**.



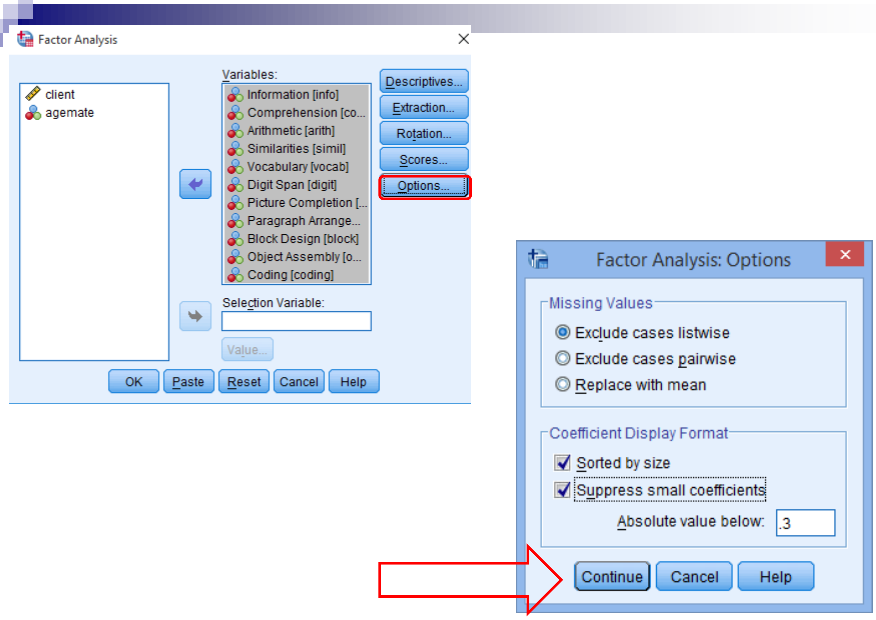
494


Factor Analysis
client
agemate
Variables:
Information [info]
Comprehension [co...
Arithmetic [arith]
Similarities [simil]
Vocabulary [vocab]
Digit Span [digit]
Picture Completion [...
Paragraph Arrange...
Block Design [block]
Object Assembly [o...
Coding [coding]
Selection Variable:
Value...
Descriptives...
Extraction...
Rotation...
Scores...
Options...
OK
Paste
Reset
Cancel
Help
Factor Analysis: Extraction
Method: Principal components
Analyze
Correlation matrix
Covariance matrix
Display
Unrotated factor solution
Scree plot
Extract
Based on Eigenvalue
Eigenvalues greater than: 1
Fixed number of factors
Factors to extract:
Maximum Iterations for Convergence: 25
Continue
Cancel
Help


Factor Analysis: Descriptives
Statistics
Univariate descriptives
Initial solution
Correlation Matrix
Coefficients
Inverse
Significance levels
Reproduced
Determinant
Anti-image
KMO and Bartlett's test of sphericity
Continue
Cancel
Help
Variables:
Information [info]
Comprehension [co...
Arithmetic [arith]
Similarities [simil]
Vocabulary [vocab]
Digit Span [digit]
Picture Completion [...
Paragraph Arrange...
Block Design [block]
Object Assembly [o...
Coding [coding]
Selection Variable:
Value...
Descriptives...
Extraction...
Rotation...
Scores...
Options...
OK
Paste
Reset
Cancel
Help
Factor Analysis: Rotation
Method
None
Quartimax
Varimax
Equamax
Direct Oblimin
Promax
Delta: 0
Kappa 4
Display
Rotated solution
Loading plot(s)
Maximum Iterations for Convergence: 25
Continue
Cancel
Help


Factor Analysis: Rotation
Method
None
Quartimax
Varimax
Equamax
Direct Oblimin
Promax
Delta: 0
Kappa 4
Display
Rotated solution
Loading plot(s)
Maximum Iterations for Convergence: 25
Continue
Cancel
Help
Variables:
Information [info]
Comprehension [co...
Arithmetic [arith]
Similarities [simil]
Vocabulary [vocab]
Digit Span [digit]
Picture Completion [...
Paragraph Arrange...
Block Design [block]
Object Assembly [o...
Coding [coding]
Selection Variable:
Value...
Descriptives...
Extraction...
Rotation...
Scores...
Options...
OK
Paste
Reset
Cancel
Help

497


Factor Analysis
client
agemate
Variables:
Information [info]
Comprehension [co...
Arithmetic [arith]
Similarities [simil]
Vocabulary [vocab]
Digit Span [digit]
Picture Completion [...
Paragraph Arrange...
Block Design [block]
Object Assembly [o...
Coding [coding]
Selection Variable:
Value...
Descriptives...
Extraction...
Rotation...
Scores...
Options...
OK
Paste
Reset
Cancel
Help
Factor Analysis: Options
Missing Values
Exclude cases listwise
Exclude cases pairwise
Replace with mean
Coefficient Display Format
Sorted by size
Suppress small coefficients
Absolute value below: .3
Continue
Cancel
Help

498

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
Information	9.50	2.912	175
Comprehension	10.00	2.965	175
Arithmetic	9.00	2.307	175
Similarities	10.61	3.184	175
Vocabulary	10.70	2.933	175
Digit Span	8.73	2.704	175
Picture Completion	10.68	2.934	175
Paragraph Arrangement	10.37	2.660	175
Block Design	10.31	2.710	175
Object Assembly	10.90	2.844	175
Coding	8.55	2.872	175

The Descriptive Statistics table simply reports the mean, standard deviation, and number of cases for each variable included in the analysis.

499

Correlation Matrix[a]

		Information	Comprehension	Arithmetic	Similarities	Vocabulary	Digit Span	Picture Completion	Paragraph Arrangement	Block Design	Object Assembly	Coding
Correlation	Information	1.000	.467	.494	.513	.625	.345	.230	.202	.229	.185	.007
	Comprehension	.467	1.000	.392	.510	.531	.236	.407	.187	.369	.322	.061
	Arithmetic	.494	.392	1.000	.369	.387	.269	.155	.227	.272	.043	.090
	Similarities	.513	.510	.369	1.000	.538	.260	.369	.298	.261	.269	-.041
	Vocabulary	.625	.531	.387	.538	1.000	.294	.285	.132	.297	.185	.100
	Digit Span	.345	.236	.269	.260	.294	1.000	.075	.148	.073	.035	.173
	Picture Completion	.230	.407	.155	.369	.285	.075	1.000	.249	.382	.363	-.072
	Paragraph Arrangement	.202	.187	.227	.298	.132	.148	.249	1.000	.351	.253	.038
	Block Design	.229	.369	.272	.261	.297	.073	.382	.351	1.000	.399	.107
	Object Assembly	.185	.322	.043	.269	.185	.035	.363	.253	.399	1.000	.053
	Coding	.007	.061	.090	-.041	.100	.173	-.072	.038	.107	.053	1.000

a. Determinant = .051

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.828
Bartlett's Test of Sphericity	Approx. Chi-Square	502.886
	df	55
	Sig.	.000

The Correlation Matrix is the correlation matrix for the variables included. Kaiser-Meyer-Olking (KMO) statistic should be greater than 0.600 and the Bartlett's test should be significant ($p < .05$). KMO is used for assessing sampling adequacy and evaluates the correlations to determine if the data are likely to CORRELATE on components

500

Communalities

	Initial	Extraction
Information	1.000	.693
Comprehension	1.000	.578
Arithmetic	1.000	.487
Similarities	1.000	.616
Vocabulary	1.000	.645
Digit Span	1.000	.474
Picture Completion	1.000	.561
Paragraph Arrangement	1.000	.368
Block Design	1.000	.601
Object Assembly	1.000	.573
Coding	1.000	.792

Extraction Method: Principal Component Analysis.

A communality (h^2) is the sum of the squared component loadings and it represents **the amount of variance in that variable accounted for by all the components**

501

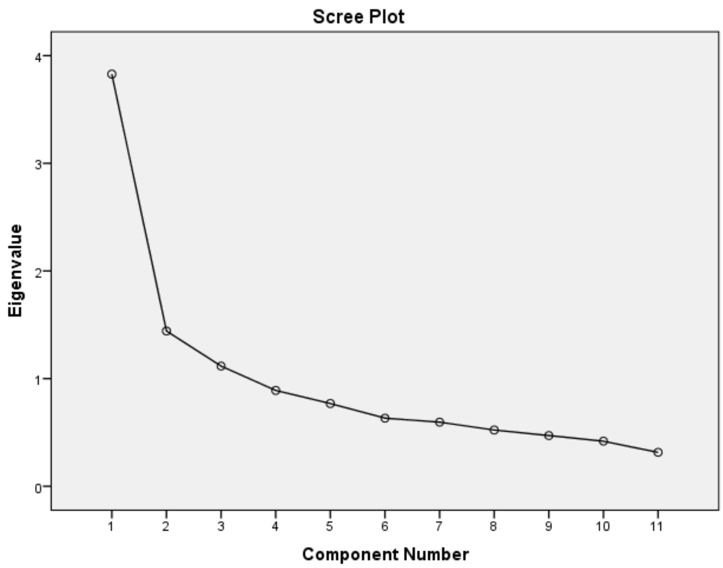
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.829	34.806	34.806	3.829	34.806	34.806	3.023	27.485	27.485
2	1.442	13.109	47.915	1.442	13.109	47.915	2.209	20.084	47.569
3	1.116	10.147	58.062	1.116	10.147	58.062	1.154	10.492	58.062
4	.890	8.092	66.153						
5	.768	6.985	73.138						
6	.633	5.753	78.891						
7	.595	5.412	84.303						
8	.522	4.749	89.051						
9	.471	4.281	93.332						
10	.419	3.806	97.138						
11	.315	2.862	100.000						

Extraction Method: Principal Component Analysis.

- The variance explained by each component as well as the cumulative variance explained by all components. Variance explained means the amount of variance in the total collection of variables/items which is explained by the component(s).
- For instance, component 3 explains 10.492% of the variance in the items.
- We could also say, 58.062% of the variance in our items was explained by the 3 extracted components.

502



eigenvalues greater than 1 was generally accepted

503

Rotated Component Matrix[a]

	Component		
	1	2	3
Information	.826	.107	
Vocabulary	.782	.181	
Similarities	.694	.324	-.172
Arithmetic	.669		.179
Comprehension	.634	.415	
Digit Span	.535		.428
Object Assembly		.756	
Block Design	.173	.742	.141
Picture Completion	.246	.649	-.280
Paragraph Arrangement	.142	.567	.163
Coding		.106	.883

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

Rotated Component Matrix shows you the factor loadings for each variable.
Based on these factor loadings, the factors represent: --**The first 5 subtests loaded strongly on Factor 1, which I'll call "Verbal IQ" --Picture Completion through Object Assembly all loaded strongly on Factor 2, which I'll "Performance IQ" --Coding loaded strongly on Factor 3.**

504

Bootstrap using SPSS

Bootstrapping

- Bootstrapping is a statistical technique that falls under the broader heading of resampling. Bootstrapping can be used in the estimation of nearly any statistic
- One goal of inferential statistics is to determine the value of a parameter of a population.
- It is typically too expensive or even impossible to measure this directly. So we use statistical sampling. We sample a population, measure a statistic of this sample, and then use this statistic to say something about the corresponding parameter of the population.

- For example, in a chocolate factory, we might want to guarantee that candy bars have a particular mean weight. It's not feasible to weigh every candy bar that is produced, so we use sampling techniques to randomly choose 100 candy bars. We calculate the mean of these 100 candy bars, and say that the population mean falls within a margin of error from what the mean of our sample is.

What is a Bootstrap

- A method of Resampling: creating many samples from a single sample
- Generally, resampling is done with replacement
- Used to develop a sampling distribution of statistics such as mean, median, proportion, others.

507

Bootstrap confidence interval

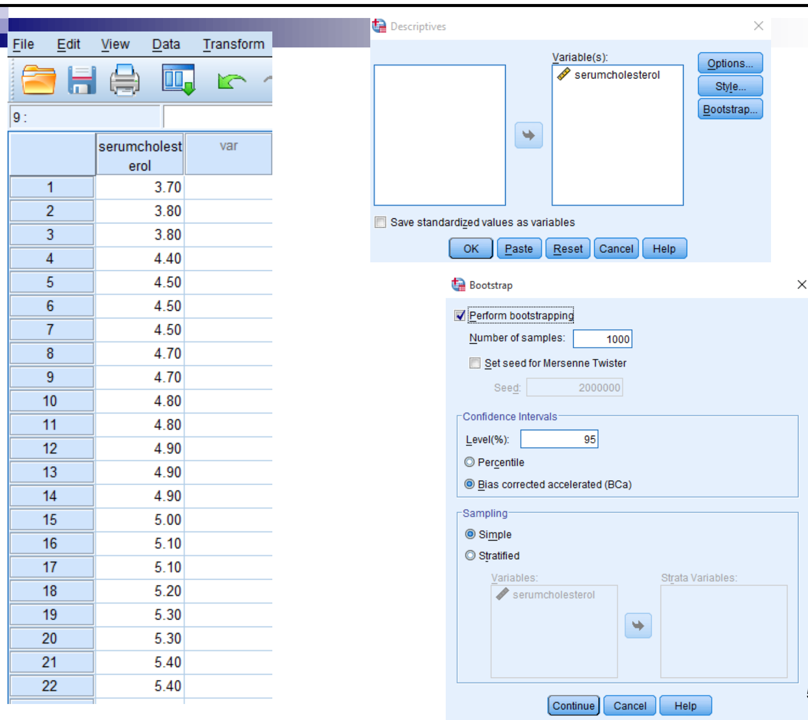
- Bootstrap confidence intervals provide a way of quantifying the uncertainties in the inferences that can be drawn from a sample of data.
- The idea is to use a simulation, based on the actual data, to estimate the likely extent of sampling error.

508

Advantages and Disadvantages

- Advantages:
  - Avoids the costs of taking new samples (Estimate a sampling distribution when only one sample is available)
  - Checking parametric assumptions
  - Used when parametric assumptions cannot be made or are very complicated
  - Estimation of variance in quantiles
- Disadvantages:
  - Relies on a representative sample
  - Variability due to finite replications (Monte Carlo)

509



510

Bootstrap Specifications	
Sampling Method	Simple
Number of Samples	1000
Confidence Interval Level	95.0%
Confidence Interval Type	Bias-corrected and accelerated (BCa)

```
DESCRIPTIVES VARIABLES=serumcholesterol
  /STATISTICS=MEAN STDDEV VARIANCE RANGE MIN MAX KURTOSIS SKEWNESS.
```

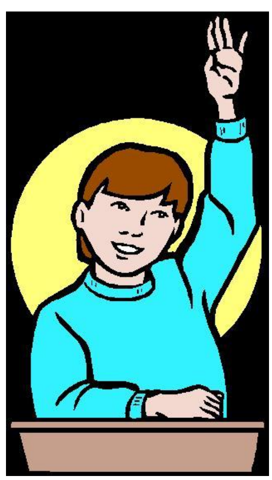
Descriptives

Descriptive Statistics

				Bootstrap[a]			
						BCa 95% Confidence Interval	
		Statistic	Std. Error	Bias	Std. Error	Lower	Upper
serumcholesterol	N	86		0	0	.	.
	Range	6.70					
	Minimum	3.70					
	Maximum	10.40					
	Mean	6.3407		.0005	.1495	6.0649	6.6324
	Std. Deviation	1.39978		-.01128	.11313	1.19786	1.57908
	Variance	1.959		-.019	.315	1.431	2.499
	Skewness	.634	.260	-.030	.228	.220	.973
	Kurtosis	.400	.514	-.045	.517	-.431	1.278
Valid N (listwise)	N	86		0	0	.	.

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

511



THANK YOU

512