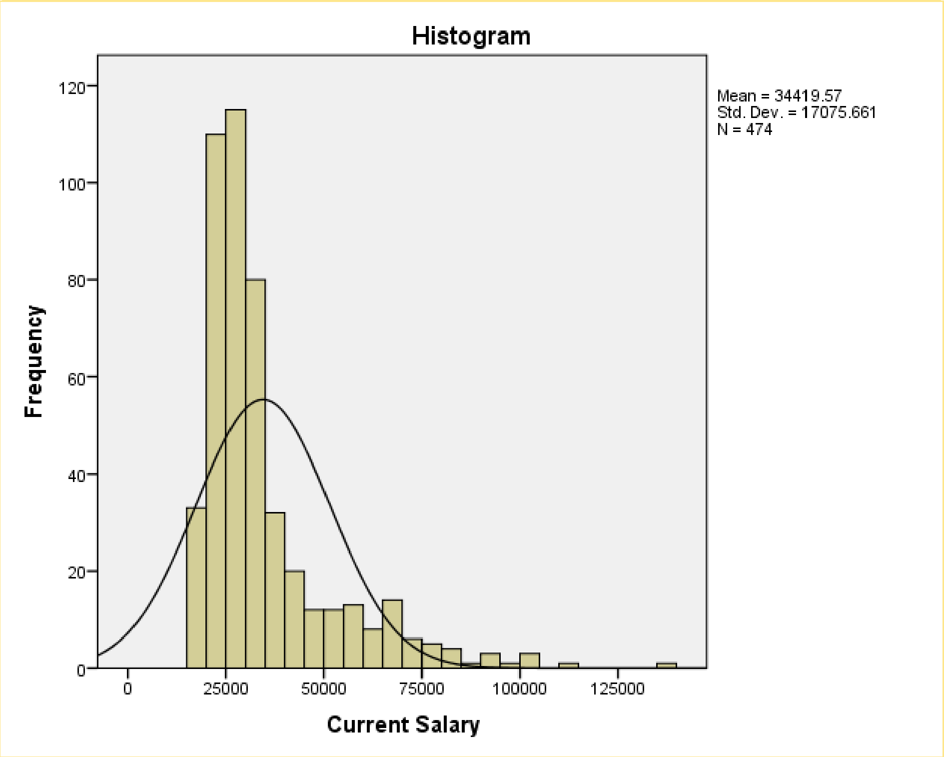
Introduction to Microsoft Excel

Online Workbook

Basic Statistics for Business and Management Students

Using Excel and IBM SPSS Statistics

This online workbook is intended to provide students with an introduction to the IBM SPSS software package.



The document is available online to download for customers who have purchased the textbook Basic Statistics for Business and Management Students - Using Excel and IBM SPSS Statistics.

Created by Glyn Davis & Branko Pecar
Introduction to Microsoft Excel
Thursday, 01 October 2020

Preface

This workbook has been designed using IBM SPSS version 24 and 25, though the latest version 27 is not much different..

The textbook explores the use of SPSS in solving a range of business and management problems identified and solved using SPSS in the textbook.

The chapter topics in this online workbook, include:

1. Introduction to IBM SPSS Statistics.
2. Entering data.
3. Graphing data.
4. Descriptive statistics.
5. Comparing means using the Students' t test.
6. Chi-square and non-parametric tests.
7. Correlation and regression analysis.

IBM SPSS Help Web Site

Access via SPSS Help menu Help.

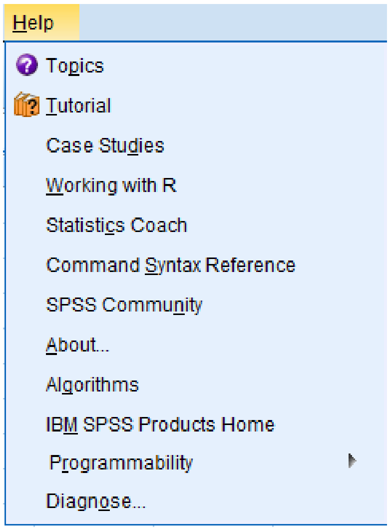


Figure Preface1 SPSS help menu

Contents

Chapter 1 Introduction to IBM SPSS Statistics.

SPSS is a Windows based program that can be used to perform data entry and analysis and to create tables and graphs. SPSS is capable of handling large amounts of data and is a statistical software program that has been designed for both beginners and professional statisticians. SPSS is commonly used in business and the social sciences, so familiarity with this program should serve you well in the future.

Running IBM SPSS

After installing IBM SPSS then you will need to run the program.

Click on Start and Select > IBM SPSS Statistics 23

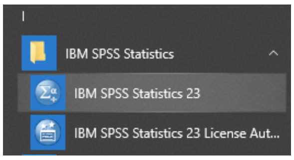


Figure 1.1 Run Windows 10 SPSS

Please note that you may have an earlier version of SPSS or a later version than version 23.

Layout of IBM SPSS

Data View is where you see the data you are using.

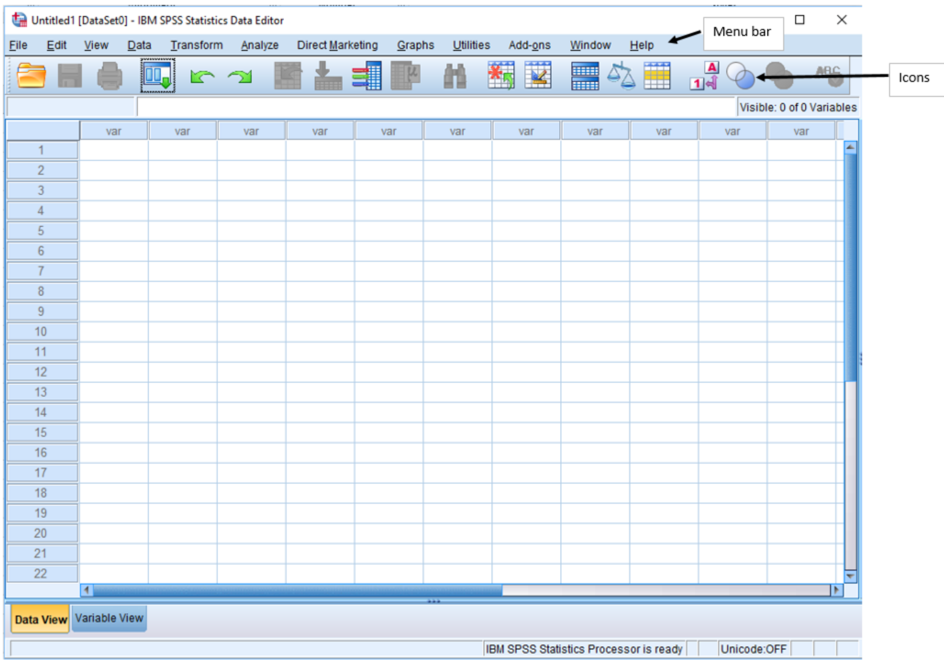


Figure 1.2 SPSS data view

The Data Editor window has two views that can be selected from the lower left-hand side of the screen

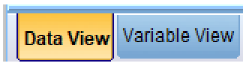


Figure 1.3 SPSS data and variable view

Variable View is where you can specify the format of your data when you are creating a file or where you can check the format of a pre-existing file.

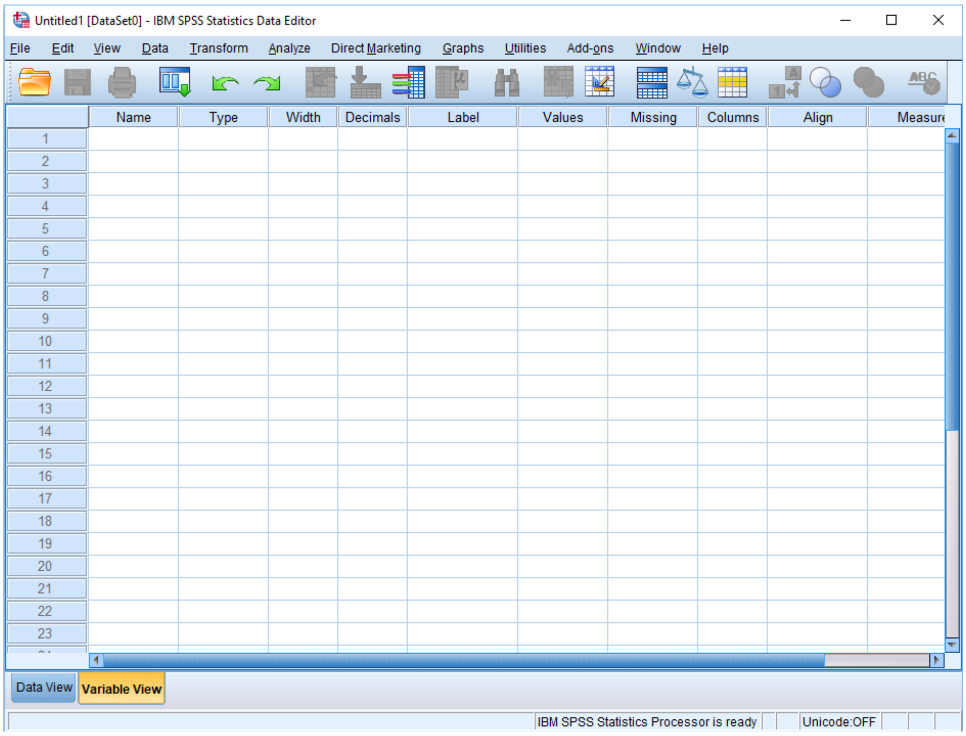


Figure 1.4 SPSS variable view

SPSS filenames – data (filename.sav)

The data in the Data Editor is saved in a file with the filename.sav.

SPSS filenames – output (filename.spv)

The other most commonly used SPSS window is the **SPSS Viewer window** which displays the output from any analyses that have been run and any error messages. When you save an SPSS output file it will add the .spv extension to the filename.

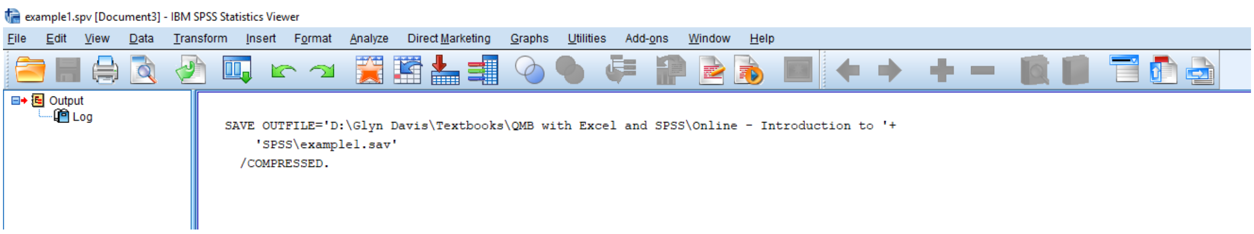


Figure 1.5 SPSS output file

Example 1.1

Enter two columns of numbers into SPSS (X: 3, 6, 7 and Y: 7, 8, 9, 10, 12).

Enter numbers into data view

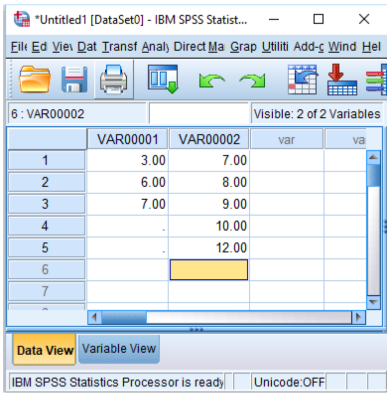


Figure 1.6 Data entered into data view

Observe that the first column is called VAR00001 and the second column VAR00002. Furthermore, the numbers are presented to 2 decimal places. To change this, click on Variable View.

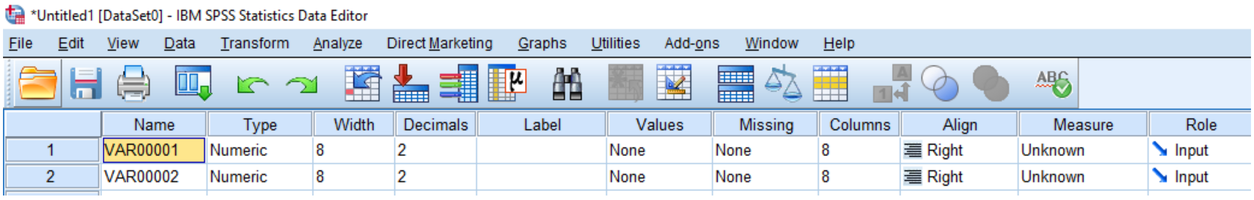


Figure 1.7 Variable view

- Change VAR00001 to X
- Change VAR00002 to Y
- Change for each 2 decimal places to 0.

At this stage, observe that the Measure is defined Unknown. It is important when you have entered the data in data view that you use the variable view to label the column, decide on the general layout, and select the appropriate measure. If your data for variable X is scale (or ratio) then select scale. If it is ordinal (ranked) then select ordinal. Finally, if it is nominal (category) then select nominal

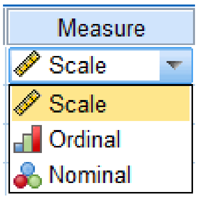


Figure 1.8 Choose the measure

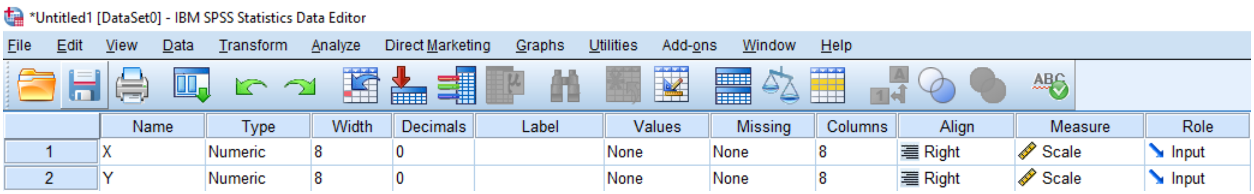


Figure 1.9 Variable view after changes made

If we now click on data view, then we would have

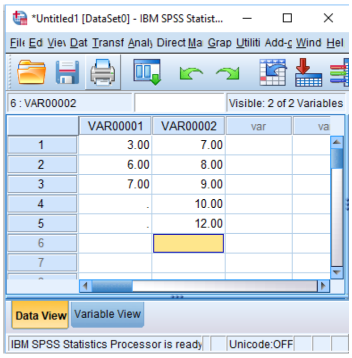


Figure 1.10 Data view

If we have completed the data entry, then we would save the file (say example1). SPSS would save this has **example1.1.sav** and we will save this in the **Online - Introduction to SPSS** folder. Select File > Save.

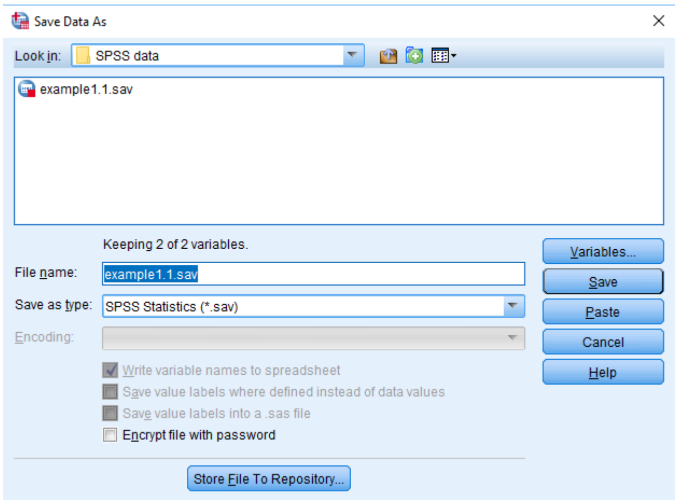


Figure 1.11 Click Save

You should observe that when we saved the data filename changed to **example1.1.sav** and a second menu window appeared. This second window is the SPSS output file. Save this file as **example1.1.spv**.

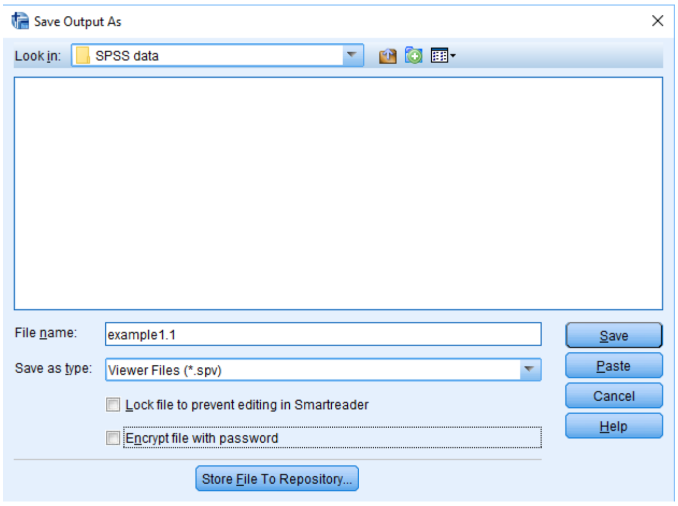


Figure 1.12 Click Save

Finally, if I look at the folder Online – Introduction to SPSS we would observe

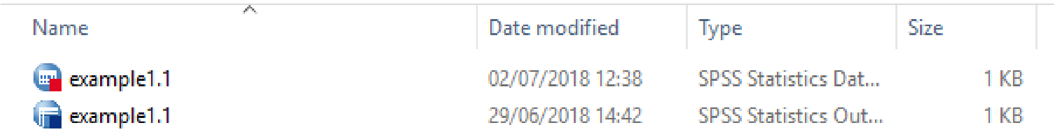


Figure 1.13 View via Windows explorer of file structure

Finally, there is the **Syntax window** which displays the command language used to run various operations. Typically, you will simply use the dialog boxes to set up commands and would not see the Syntax window. The Syntax window would be activated if you pasted the commands from the dialog box to it, or if you wrote you own syntax--something we will not focus on here. Syntax files end in the extension **.sps**.

SPSS Menus and Icons

Now, let's review the menus and icons. Review the options listed under each menu on the Menu Bar by clicking them one at a time.

File

Includes all of the options you typically use in other programs, such as open, save, exit. Notice, that you can open or create new files of multiple types as illustrated to the right.

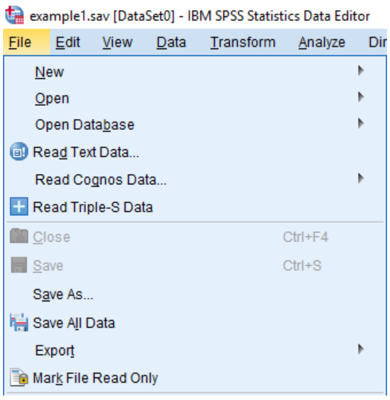


Figure 1.14 File menu (not complete)

Edit

Includes the typical cut, copy, and paste commands, and allows you to specify various options for displaying data and output.

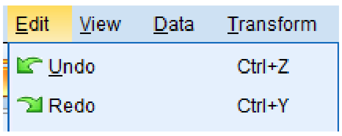


Figure 1.15 Edit menu (not complete)

View

Allows you to select which toolbars you want to show, select font size, add or remove the gridlines that separate each piece of data, and to select whether or not to display your raw data or the data labels.

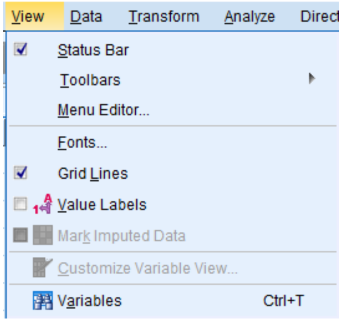


Figure 1.16 View menu

Data

Allows you to select several options ranging from displaying data that is sorted by a specific variable to selecting certain cases for subsequent analyses.

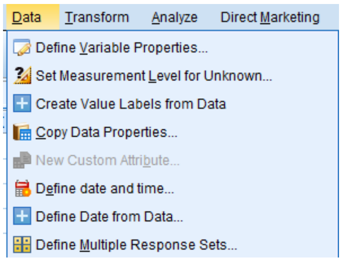


Figure 1.17 Data menu (not complete)

Transform

Includes several options to change current variables. For example, you can change continuous variables to categorical variables, change scores into rank scores, add a constant to variables, etc.

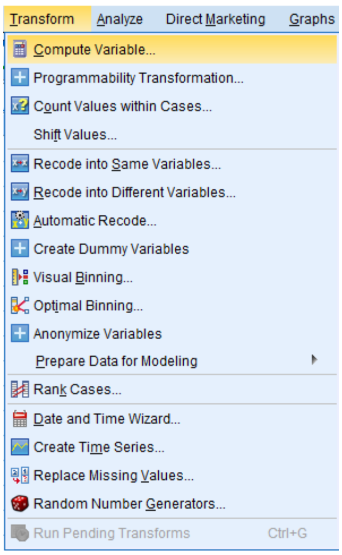


Figure 1.18 Transform menu

Analyze

Includes all of the commands to carry out statistical analyses and to calculate descriptive statistics. Much of this book will focus on using commands located in this menu.

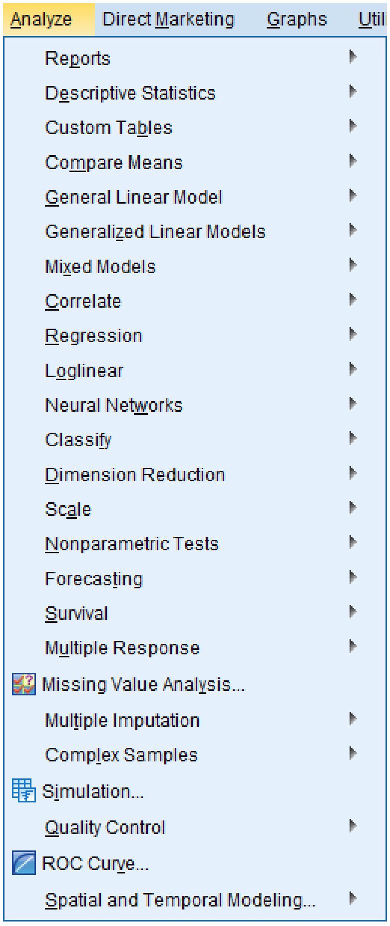


Figure 1.19 Analyze menu

Direct Marketing

The Direct Marketing option provides a set of tools designed to improve the results of direct marketing campaigns by identifying demographic, purchasing, and other characteristics that define various groups of consumers and targeting specific groups to maximize positive response rates.

1. RFM Analysis. This technique identifies existing customers who are most likely to respond to a new offer.
2. Cluster Analysis. This is an exploratory tool designed to reveal natural groupings (or clusters) within your data. For example, it can identify different groups of customers based on various demographic and purchasing characteristics.
3. Prospect Profiles. This technique uses results from a previous or test campaign to create descriptive profiles. You can use the profiles to target specific groups of contacts in future campaigns.
4. Postal Code Response Rates. This technique uses results from a previous campaign to calculate postal code response rates. Those rates can be used to target specific postal codes in future campaigns.

5. Propensity to Purchase. This technique uses results from a test mailing or previous campaign to generate propensity scores. The scores indicate which contacts are most likely to respond.
6. Control Package Test. This technique compares marketing campaigns to see if there is a significant difference in effectiveness for different packages or offers.

The Direct Marketing dialog for selecting a technique also provides a shortcut to the Scoring Wizard, which allows you to score data based on a predictive model. You can build predictive models with Propensity to Purchase and with many procedures available in other add-on modules.

Select Direct Marketing

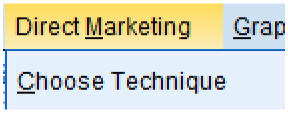


Figure 1.20 Direct marketing menu

Select Choose Technique

Figure 1.21

Please note this option is not part of this textbook but would possibly be useful for students studying digital marketing type courses.

Graphs

Includes the commands to create various types of graphs including box plots, histograms, line graphs, and bar charts.

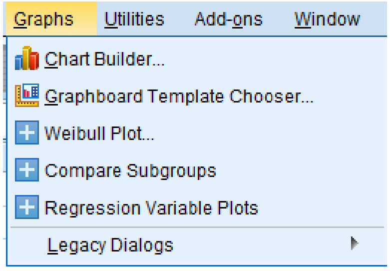


Figure 1.22 Graphs menu

Utilities

Allows you to list file information which is a list of all variables, there labels, values, locations in the data file, and type.

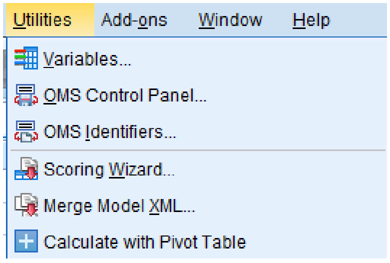


Figure 1.23 Utilities menu

Add-ons

Are programs that can be added to the base SPSS package. You probably do not have access to any of those.

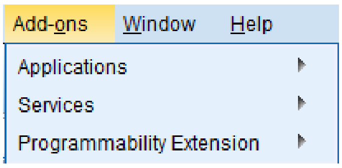


Figure 1.24 Add-ons menu

Window

Can be used to select which window you want to view (i.e., Data Editor, Output Viewer, or Syntax). Since we have a data file and an output file open, let's try this.

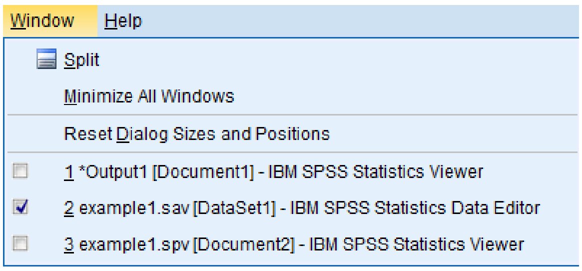


Figure 1.25 Windows menu

Help

Has many useful options including a link to the SPSS homepage, a statistics coach, and a syntax guide. Using topics, you can use the index option to type in any key word and get a list of options, or you can view the categories and subcategories available under contents. This is an excellent tool and can be used to troubleshoot most problems.

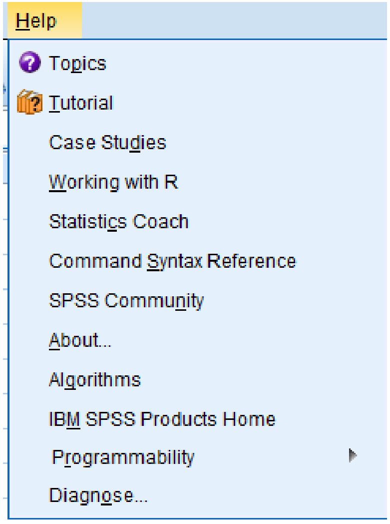


Figure 1.26 Help menu

Menu Icons

The Icons directly under the Menu bar provide shortcuts to many common commands that are available in specific menus.

Place your cursor over the Icons for a few seconds, and a description of the underlying command will appear. For example, this icon  is the shortcut for Save.

Exiting SPSS

To close SPSS, you can either left click on the close button × located on the upper right-hand corner of the screen or select Exit from the File menu. A dialog box like the one below will appear for every open window asking you if you want to save it before exiting.

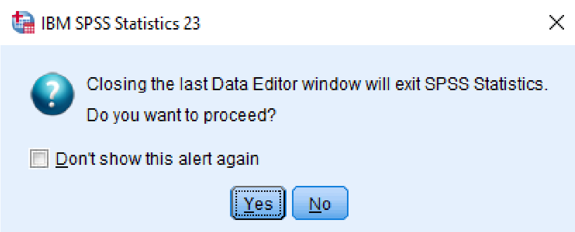


Figure 1.27 SPSS Menu warning before closing file

Important – save your data files (filename.sav) and output files (filename.spv)

You almost always want to save data files and output files BUT make sure you store in the correct folder and use filenames that relate to the problem you are studying for example, we could use for the SPSS data file data_payments.sav and for the SPSS output file data_payments.spv.

Chapter 2 Entering data.

The Logic of Data Files

Each row typically represents the data from 1 case, whether that be a person, animal, or object. Each column represents a different variable. A cell refers to the juncture of a specific row and column. For example, the first empty cell in the right-hand corner would include the data for case 1, variable 1.

Entering Data

Run SPSS and follow along as you read this description.

To enter data, you could simply begin typing information into each cell. If you did so, SPSS would give each column a generic label such as VAR00001. Clearly this is not desirable, unless you have a superior memory, because you would have no way of identifying what VAR00001 meant later on.

Instead, we want to specify names for our variables. To do this, you can double left click on any column head, this will automatically take you to the Variable View. Alternatively, you can simply click on Variable View on the bottom left hand corner of your screen.

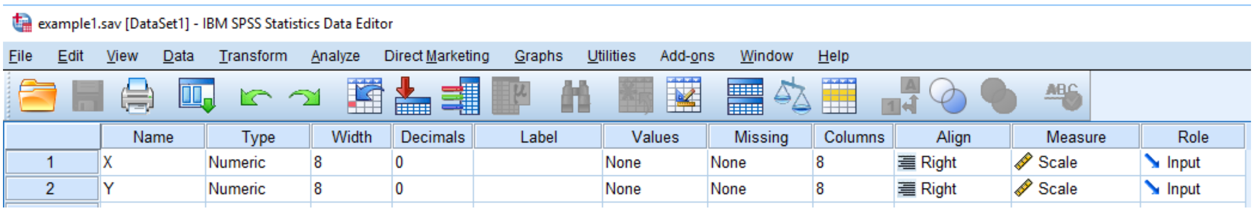


Figure 2.1 Variable view

Column 1 - The first column of variable view is Name.

Column 2 - The second column is Type

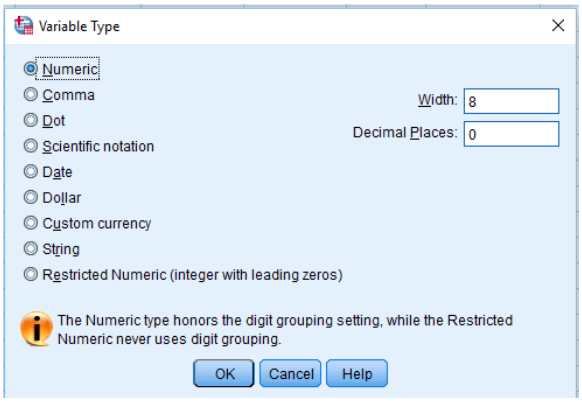


Figure 2.2 Variable type menu

String variables are those that consist of text. For example, you could type Male and Female if gender were a variable of interest. It is important to note that SPSS is case sensitive meaning that “female” and “Female” would not be viewed as the same category. Misspellings are also problematic with string data (e.g., “femal” would not be recognized as the intended “female”). For these reasons, it is advantageous to use numbers to represent common categories, and then supply names for those levels as discussed below.

Column 3 and 4 – Width and Decimals

The next columns are for Width and Decimals. You could have set this while specifying your variable type, or you can specify them in these columns. The default for width is 8 characters and the default for decimals is 2. To change this, left click the cell, and up and down arrows will appear, as illustrated below. Left click the up arrow if you want to increase the number, click the down arrow to decrease the value. Alternatively, you can simply type the desired value in the cell.

Column 5 – Column label

The next column is Label. This is a very nice feature that allows you to provide more information about the variable than you could fit in the 8-character variable name. For example, I could type “time before training started” if variable X represented this variable.

Column 6 - Values

The next column is Values. This allows you to assign variable labels. You will typically use this option for categorical variables. For example, we may want the number 1 to represent males and the number 2 to represent females when we enter data on gender.

Other columns

Of the remaining columns, you are most likely to use Align, which allows you to specify how the data will appear in the cells. Your choices are left justified, right justified, or centered. This is simply a matter of personal preference.

After you have completed specifying your variables, you can click on Data View and begin entering your data. Put your cursor on the cell in which you want to enter data. Type the value. If you hit Enter the cursor will move to the cell under the one you just filled. You can also use the arrow keys to move to the next cell in any given direction. Typically, you will either enter all of the values in one column by going down or you will enter all of the variables in a row going from left to right.

Example 2.1

Consider the data collected by a researcher that is exploring student grades in a statistics examination. The research has collected the following data which requires entered into SPSS.

Student ID	Age	Gender	Average hours of sleep per student	Number of classes missed	Final module grade
1	18	Male	7	0	A
2	18	Female	4	1	C
3	17	Female	6	2	B
4	19	Female	10	5	F
5	20	Male	8	2	B

6	21	Female	7	3	C
7	23	Male	9	1	B
8	22	Male	8	2	A
9	18	Male	6	3	D

Table 2.1

SPSS data file

Data view

	ID	Age	Gender	AvHrsSleep	NumClasses Missed	GradeinCourse
1	1	18	1	7.00	.00	1
2	2	18	2	4.00	1.00	3
3	3	17	2	6.00	2.00	2
4	4	19	2	10.00	5.00	5
5	5	20	1	8.00	2.00	2
6	6	21	2	7.00	3.00	3
7	7	23	1	9.00	1.00	2
8	8	22	1	8.00	2.00	1
9	9	18	1	6.00	3.00	4

Figure 2.3

Gender (1 = Male, 2 = Female)
Grade (1 = A, 2 = B, 3 = C, 4 = D, 5 = F)

Or you could use the following

	VAR00001	VAR00002	VAR00003	VAR00004	VAR00005	VAR00006
1	1.00	18.00	Male	7.00	.00	A
2	2.00	18.00	Female	4.00	1.00	C
3	3.00	17.00	Female	6.00	2.00	B
4	4.00	19.00	Female	10.00	5.00	F
5	5.00	20.00	Male	8.00	2.00	B
6	6.00	21.00	Female	7.00	3.00	C
7	7.00	23.00	Male	9.00	1.00	B
8	8.00	22.00	Male	8.00	2.00	A
9	9.00	18.00	Male	6.00	3.00	D

Figure 2.4

Variable view

Edit variable view to give

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	ID	Numeric	8	0		None	None	12	Center	Ordinal	Input
2	Age	Numeric	8	0		None	None	11	Center	Scale	Input
3	Gender	String	6	0		{1, Male}...	None	11	Center	Nominal	Input
4	AvHrsSleep	Numeric	8	2		None	None	10	Center	Scale	Input
5	NumClassesMissed	Numeric	8	2		None	None	8	Center	Scale	Input
6	GradeinCourse	String	1	0		{1, A}...	None	10	Center	Nominal	Input

Figure 2.5

Where Gender values looks like

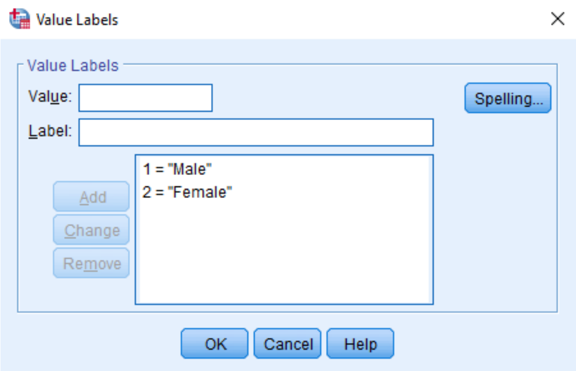


Figure 2.6

Where Grade values looks like

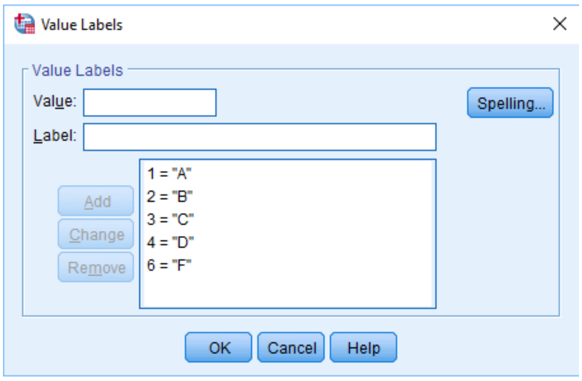


Figure 2.7

Final data view

	ID	Age	Gender	AvHrsSleep	NumClasses Missed	GradeinCourse
1	1	18	1	7.00	.00	1
2	2	18	2	4.00	1.00	3
3	3	17	2	6.00	2.00	2
4	4	19	2	10.00	5.00	5
5	5	20	1	8.00	2.00	2
6	6	21	2	7.00	3.00	3
7	7	23	1	9.00	1.00	2
8	8	22	1	8.00	2.00	1
9	9	18	1	6.00	3.00	4

Figure 2.8

Save SPSS Data file: Example2.1.sav

Inserting a Variable

If you forget to insert a variable, then it is quite easy to add a new variable. In Variable View, highlight the second row and then click Insert Variable on the Edit menu. This will place a new variable before the selected variable.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	ID	Numeric	8	0		None	None	12	Center	Ordinal	Input
2	Age	Numeric	8	0		None	None	11	Center	Scale	Input
3	Gender	String	6	0		{1, Male}...	None	11	Center	Nominal	Input
4	AvHrsSleep	Numeric	8	2		None	None	10	Center	Scale	Input
5	NumClassesMissed	Numeric	8	2		None	None	8	Center	Scale	Input
6	GradeinCourse	String	1	0		{1, A}...	None	10	Center	Nominal	Input

Figure 2.9

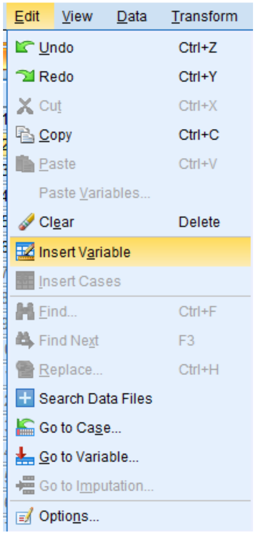


Figure 2.10

In Data View, highlight the second variable column and then click the Insert Variable icon . This will also place a new variable column before the second variable.

	ID	Age	Gender	AvHrsSleep	NumClasses Missed	GradeinCourse
1	1	18	1	7.00	.00	1
2	2	18	2	4.00	1.00	3
3	3	17	2	6.00	2.00	2
4	4	19	2	10.00	5.00	5
5	5	20	1	8.00	2.00	2
6	6	21	2	7.00	3.00	3
7	7	23	1	9.00	1.00	2
8	8	22	1	8.00	2.00	1
9	9	18	1	6.00	3.00	4

Figure 2.11

Re-save SPSS Data file: Example2.1.sav

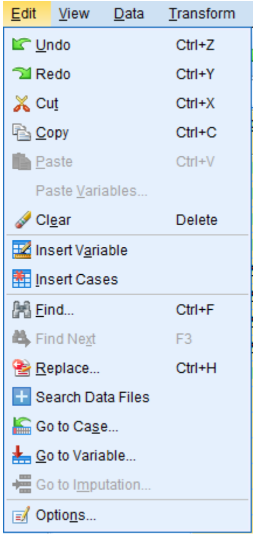


Figure 2.12

Inserting a Case

If you found that a case was missing – let us say that ID 10 was missing, then we can easily add a case.

Highlight the case for ID 9.

	ID	Age	Gender	AvHrsSleep	NumClasses Missed	GradeinCourse
1	1	18	1	7.00	.00	1
2	2	18	2	4.00	1.00	3
3	3	17	2	6.00	2.00	2
4	4	19	2	10.00	5.00	5
5	5	20	1	8.00	2.00	2
6	6	21	2	7.00	3.00	3
7	7	23	1	9.00	1.00	2
8	8	22	1	8.00	2.00	1
9	9	18	1	6.00	3.00	4

Figure 2.13

Click on Insert Case on the Data menu or click on the Insert Case icon . In either case, a blank row will appear before the highlighted case.

	ID	Age	Gender	AvHrsSleep	NumClasses Missed	GradeinCourse
1	1	18	1	7.00	.00	1
2	2	18	2	4.00	1.00	3
3	3	17	2	6.00	2.00	2
4	4	19	2	10.00	5.00	5
5	5	20	1	8.00	2.00	2
6	6	21	2	7.00	3.00	3
7	7	23	1	9.00	1.00	2
8	8	22	1	8.00	2.00	1
9	.	.		.	.	
10	9	18	1	6.00	3.00	4

Figure 2.14

Resave SPSS Data file: Example2.1.sav

Merging Files

Adding Cases.

Sometimes data that are related may be in different files that you would like to combine or merge. For example, in a research methods class, every student may collect and then enter data in their own data file. Then, the instructor might want to put all of their data into one file that includes more cases for data analysis. In this case, each file contains the same variables but different cases. To combine these files, Select Data > Merge Files > Add Cases.

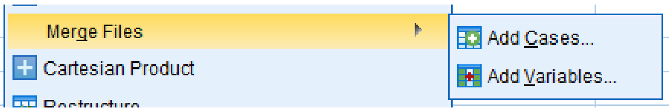


Figure 2.15

Then specify the file from which the new data will come and click Open. A dialog box will appear showing you which variables will appear in the new file. View it, and if all seems in order, click OK. The two files will be merged.

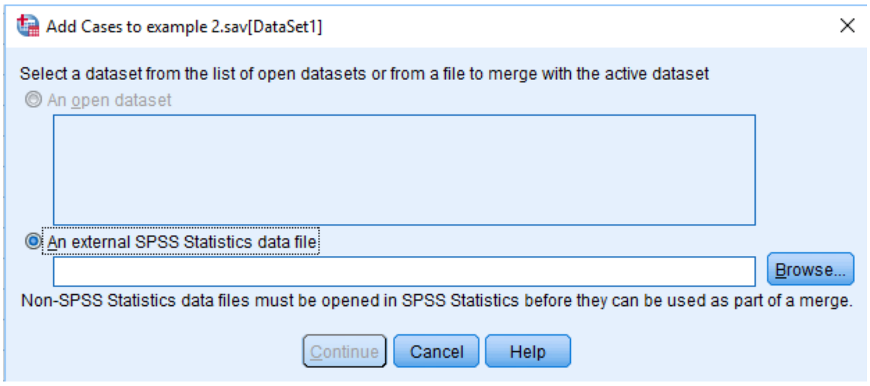


Figure 2.16

Adding Variables.

In other cases, you might have different data on the same cases or participants in different files. I may want to put them together because I'd like to see if demographic variables, like socioeconomic status or gender are related to depression.

In this case, you need to be sure the variables on the same participants end up in the correct row, that is, you want to match the cases. In this case, we will use ID to match cases. SPSS requires that the files you merge be in ascending order by the matching variable. So, in both files, ID must start at 1. You can set this up by sorting cases. Then, make sure one of the files is open.

Reading Data in From Other Sources

SPSS can also recognize data from several other sources. For example, you can open data from Microsoft EXCEL in SPSS, or you can get SPSS to read data entered in a text file.

Chapter 3 Tabulating and graphing data

In this chapter, we will explore the use of SPSS to tabulate data and create graphs.

Example 3.1

Consider the employee data presented in Figure 3.1.

SPSS data view

The first 20 records out of 474 records are presented below

	Staff_id	Gender	Education_level	Job_type	Current_salary	Start_salary	Employ_time	Previous_Emply_time	Minority_classification
1	1	m	15	3	57000	27000	98	144	0
2	2	m	16	1	40200	18750	98	36	0
3	3	f	12	1	21450	12000	98	381	0
4	4	f	8	1	21900	13200	98	190	0
5	5	m	15	1	45000	21000	98	138	0
6	6	m	15	1	32100	13500	98	67	0
7	7	m	15	1	36000	18750	98	114	0
8	8	f	12	1	21900	9750	98	0	0
9	9	f	15	1	27900	12750	98	115	0
10	10	f	12	1	24000	13500	98	244	0
11	11	f	16	1	30300	16500	98	143	0
12	12	m	8	1	28350	12000	98	26	1
13	13	m	15	1	27750	14250	98	34	1
14	14	f	15	1	35100	16800	98	137	1
15	15	m	12	1	27300	13500	97	66	0
16	16	m	12	1	40800	15000	97	24	0
17	17	m	15	1	46000	14250	97	48	0
18	18	m	16	3	103750	27510	97	70	0
19	19	m	12	1	42300	14250	97	103	0
20	20	f	12	1	26250	11550	97	48	0

Figure 3.1 Employee data

SPSS variable view

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Staff_id	Numeric	4	0	Employee Code	None	None	10	Center	Scale	Input
2	Gender	String	1	0	Gender	{f, Female}...	None	7	Center	Nominal	Input
3	Education_level	Numeric	2	0	Educational Level (years)	{0, 0 (Missing)}...	0	11	Center	Ordinal	Input
4	Job_type	Numeric	1	0	Employment Category	{0, 0 (Missing)}...	0	8	Center	Ordinal	Input
5	Current_salary	Numeric	8	0	Current Salary	{0, missing}...	0	10	Center	Scale	Input
6	Start_salary	Numeric	8	0	Beginning Salary	{0, missing}...	0	10	Center	Scale	Input
7	Employ_time	Numeric	2	0	Months since Hire	{0, missing}...	0	9	Center	Scale	Input
8	Previous_Emply_time	Numeric	6	0	Previous Experience (months)	{0, missing}...	None	15	Center	Scale	Input
9	Minority_classification	Numeric	1	0	Minority Classification	{0, No}...	9	14	Center	Ordinal	Input

Figure 3.2

Save SPSS Data file: Example3.1.sav

Frequencies: Counts and Percents

Counts and percents are wonderful statistics because they are easy to explain and quickly grasped. Frequencies also form the very foundation of most explanations of probability. They are an excellent place to begin understanding any data you may work with.

Analyze > Descriptive Statistics > Frequencies

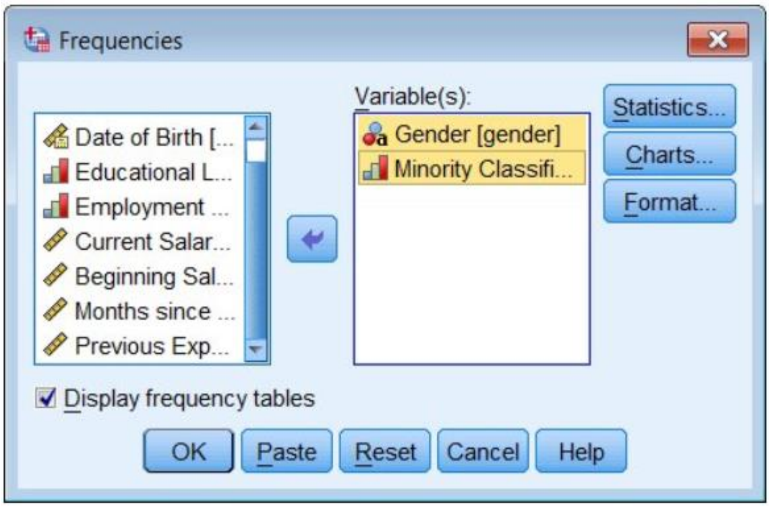


Figure 3.3

Select one or more variables in the selection list on the left and move them into the analysis list on the right by clicking on the arrow in between. Then click OK.

Output

Frequency Table

Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	216	45.6	45.6	45.6
	Male	258	54.4	54.4	100.0
	Total	474	100.0	100.0	

Minority Classification

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	370	78.1	78.1	78.1
	Yes	104	21.9	21.9	100.0
	Total	474	100.0	100.0	

Figure 3.4

Save SPSS Output file: Example3.1.spv

Note that this is one of the few tables were missing values (whether system missing "." or user designated missing) show up in the default output table (however, not in this example).

Crosstabs: Counts by Group

The basic crosstabs command just gives you counts by default. Typically, it is useful to also look at either row-percents or column-percents, which must be specified as options.

Analyze > Descriptive Statistics > Crosstabs

Select one variable as the rows, another variable as the columns. Conventionally you might put an independent variable in the rows and a dependent variable in the columns, although mathematically it doesn't really matter. To get percents in your output, click on the Cells button and specify the kind of percents you want to see.

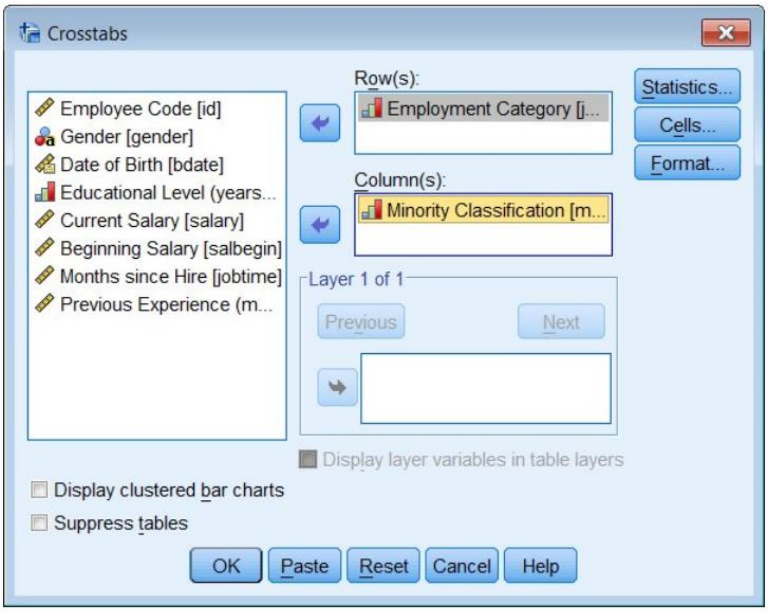


Figure 3.5

Output

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Employment Category * Minority Classification	474	100.0%	0	0.0%	474	100.0%

Employment Category * Minority Classification Crosstabulation

Count

		Minority Classification		Total
		No	Yes	
Employment Category	Clerical	276	87	363
	Custodial	14	13	27
	Manager	80	4	84
Total		370	104	474

Figure 3.6

Resave SPSS Output file: Example3.1.spv

Bar Charts

Like a histogram, the x axis is treated as a categorical variable, and the y axis represents one of a variety of summary statistics: counts (a.k.a. a histogram!), means, sums, etc.

Graphs>- Legacy Dialogs > Bar

This takes you through an initial dialog box, where you choose among several basic schemas for making bar charts

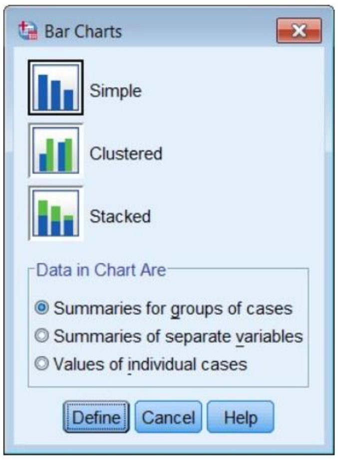


Figure 3.7

Choose Simple

To graph means by groups, select Other statistic for what the bars represent, the variable for which you want to calculate means in the Variable box (means will be the default statistic), and the group in the Category Axis box, e.g. employment category.

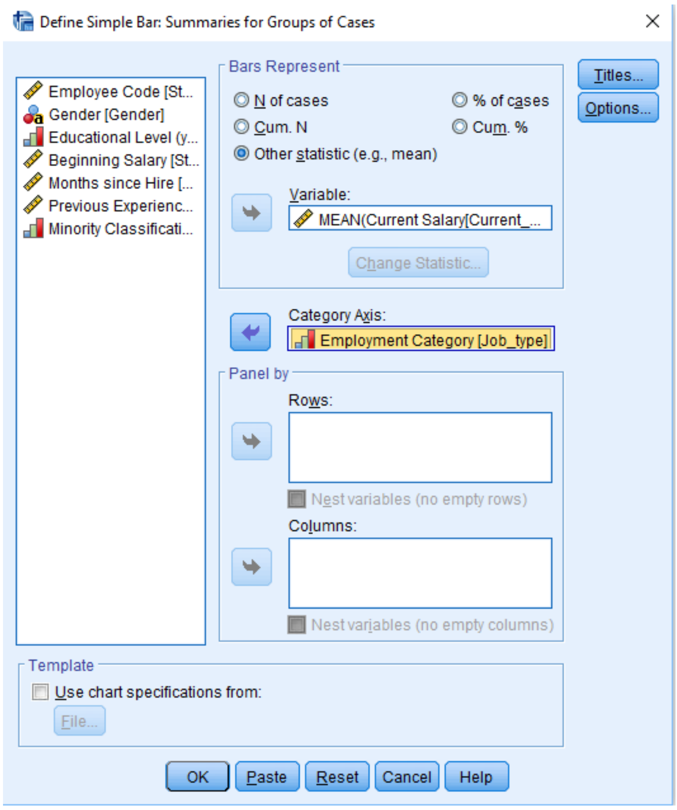


Figure 3.8
Click OK

Output

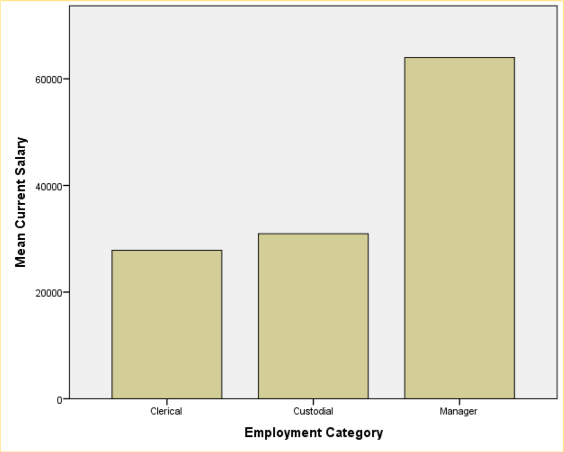


Figure 3.9

Resave SPSS Output file: Example3.1.spv

Boxplots

As with bar charts, you first choose a specific boxplot schema from an initial dialog box

Graphs > Legacy Dialogs > Boxplot

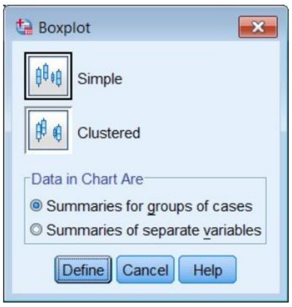


Figure 3.10
Choose Simple
Click Define

and then choose the analytical variable (the one you want to see medians and interquartile ranges for, the y axis), and the categorical variable (the x axis), e.g. employment category.

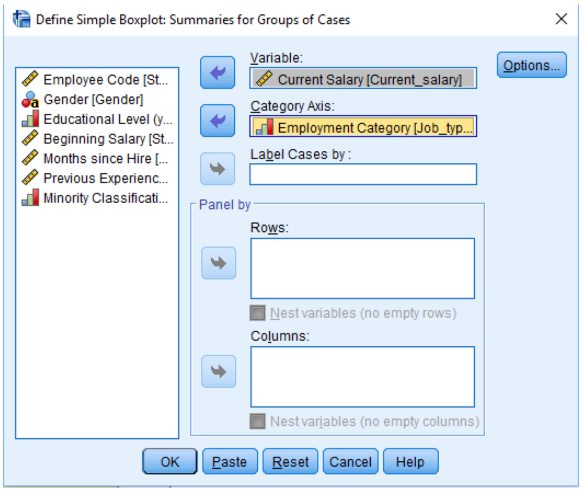


Figure 3.11
Click OK

Output

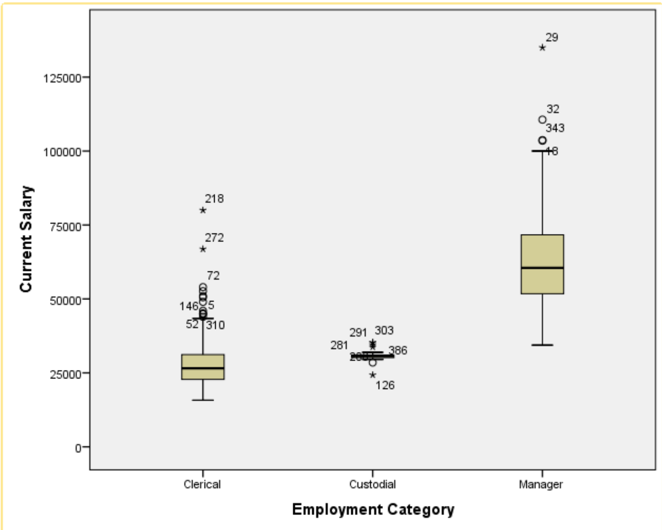


Figure 3.12

Resave SPSS Output file: Example3.1.spv

Histograms

SPSS has three different sets of commands for producing graphs. The easiest to learn and use are the oldest "legacy" graphing commands. They give you graphs with a default visual style (colours used, weight of lines, size of type, etc) that can be customized by hand.

Histograms are vexing because they can be alternately informative or deceptive, depending upon how the bins (the bar boundaries) are chosen. They are useful and popular because they are conceptually very simple, easy to draw and interpret, and if drawn well they can give a good visual representation of the distribution of values of a variable.

Graphs>- Legacy Dialogs > Histogram

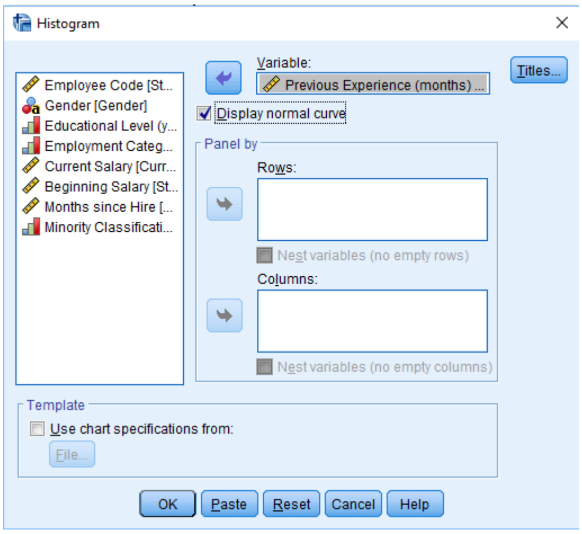


Figure 3.13

The basic histogram command works with one variable at a time, so pick one variable from the selection list on the left and move it into the Variable box. (A useful option if you expect your variable to have a normal distribution is to Display normal curve.)

Output

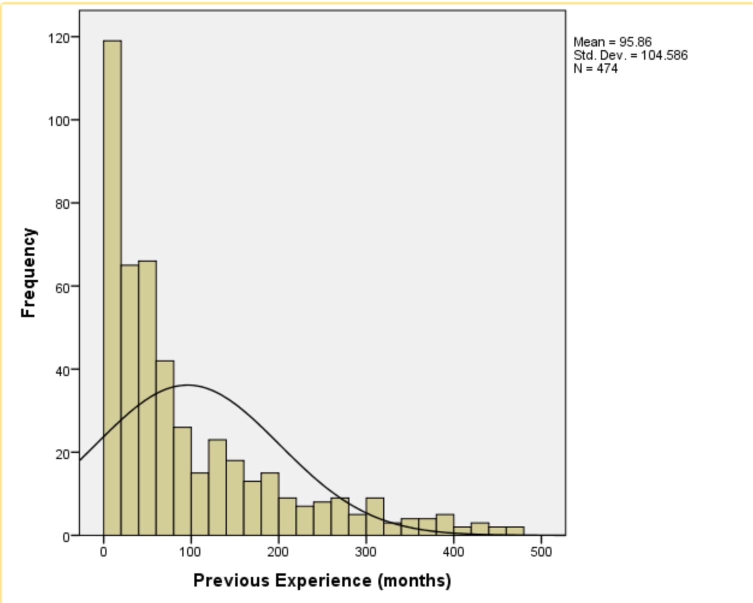


Figure 3.14

Resave SPSS Output file: Example3.1.spv

In this example, the distribution of the data is nothing like a normal distribution! To edit colours, titles, scales, etc. double-click on the graph in the Output Viewer, then double-click on the graph element you want to change.

Scatter Plots

Both simple scatter plots and scatter plot matrixes are pretty easy to produce.

Graphs > Legacy Dialogs > Scatter/Dot

Takes you through two dialog boxes. First you choose the scatter plot schema you want to work with,

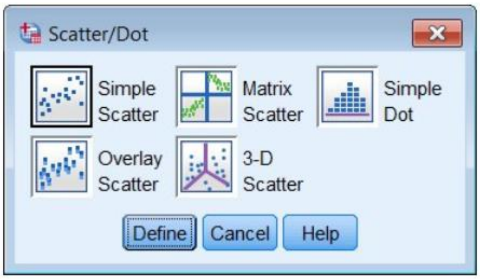


Figure 3.15

Choose Simple Scatter

And then you specify the variables with the x and y coordinates of the points you wish to plot.

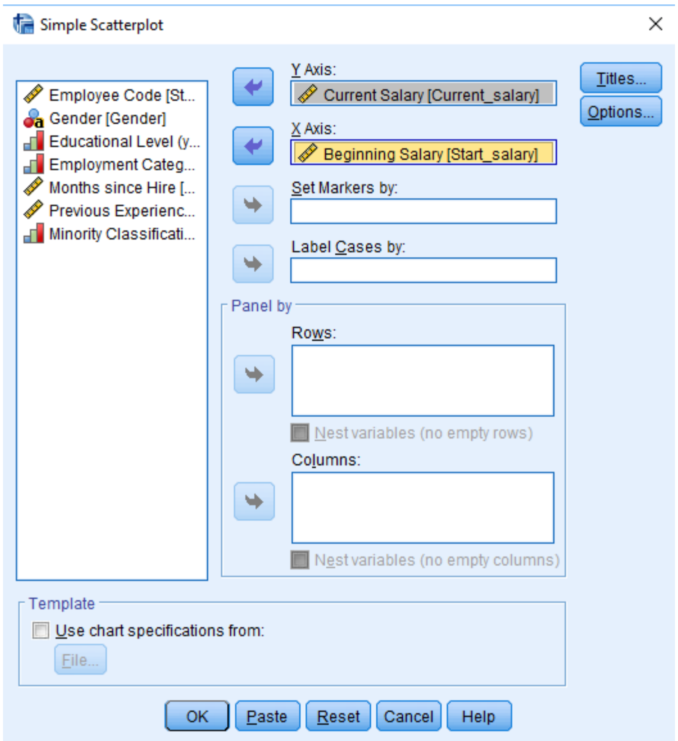


Figure 3.16

Click Ok

Output

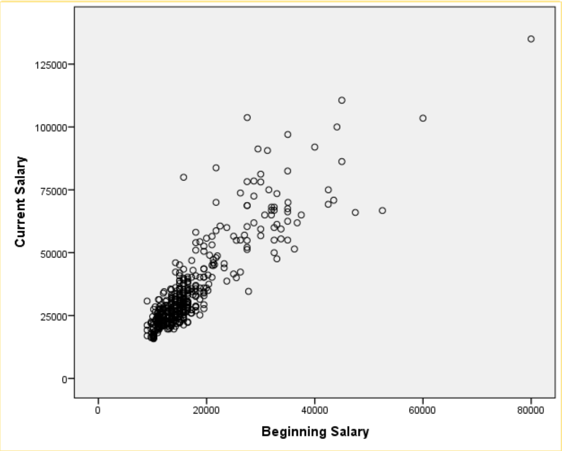


Figure 3.17

Resave SPSS Output file: Example3.1.spv

Chapter 4 Descriptive statistics

To illustrate we shall use the employee data used in Chapter 3 (example3.1.sav).

Example 4.1

Figure 4.1 represents the first 20 out of 474 records illustrated in SPSS Data View.

	Staff_id	Gender	Education_level	Job_type	Current_salary	Start_salary	Employ_time	Previous_Emply_time	Minority_classification
1	1	m	15	3	57000	27000	98	144	0
2	2	m	16	1	40200	18750	98	36	0
3	3	f	12	1	21450	12000	98	381	0
4	4	f	8	1	21900	13200	98	190	0
5	5	m	15	1	45000	21000	98	138	0
6	6	m	15	1	32100	13500	98	67	0
7	7	m	15	1	36000	18750	98	114	0
8	8	f	12	1	21900	9750	98	0	0
9	9	f	15	1	27900	12750	98	115	0
10	10	f	12	1	24000	13500	98	244	0
11	11	f	16	1	30300	16500	98	143	0
12	12	m	8	1	28350	12000	98	26	1
13	13	m	15	1	27750	14250	98	34	1
14	14	f	15	1	35100	16800	98	137	1
15	15	m	12	1	27300	13500	97	66	0
16	16	m	12	1	40800	15000	97	24	0
17	17	m	15	1	46000	14250	97	48	0
18	18	m	16	3	103750	27510	97	70	0
19	19	m	12	1	42300	14250	97	103	0
20	20	f	12	1	26250	11550	97	48	0

Figure 4.1 Employee data

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Staff_id	Numeric	4	0	Employee Code	None	None	10	Center	Scale	Input
2	Gender	String	1	0	Gender	{f, Female}...	None	7	Center	Nominal	Input
3	Education_level	Numeric	2	0	Educational Level (years)	{0, 0 (Missing)}...	0	11	Center	Ordinal	Input
4	Job_type	Numeric	1	0	Employment Category	{0, 0 (Missing)}...	0	8	Center	Ordinal	Input
5	Current_salary	Numeric	8	0	Current Salary	{0, missing}...	0	10	Center	Scale	Input
6	Start_salary	Numeric	8	0	Beginning Salary	{0, missing}...	0	10	Center	Scale	Input
7	Employ_time	Numeric	2	0	Months since Hire	{0, missing}...	0	9	Center	Scale	Input
8	Previous_Emply_time	Numeric	6	0	Previous Experience (months)	{0, missing}...	None	15	Center	Scale	Input
9	Minority_classification	Numeric	1	0	Minority Classification	{0, No}...	9	14	Center	Ordinal	Input

Figure 4.2 Variable view

Save SPSS Data file: Example4.1.sav

Frequencies

Use frequencies menu to calculate a range of descriptive statistics for the current salary variable (salary).

Select Analyze > Descriptives > Frequencies

Transfer salary to the Variable(s) box

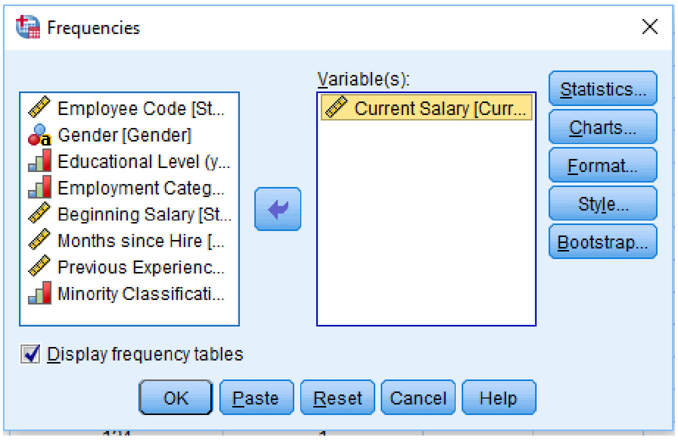


Figure 4.3

Click on Statistics

Choose the required statistics

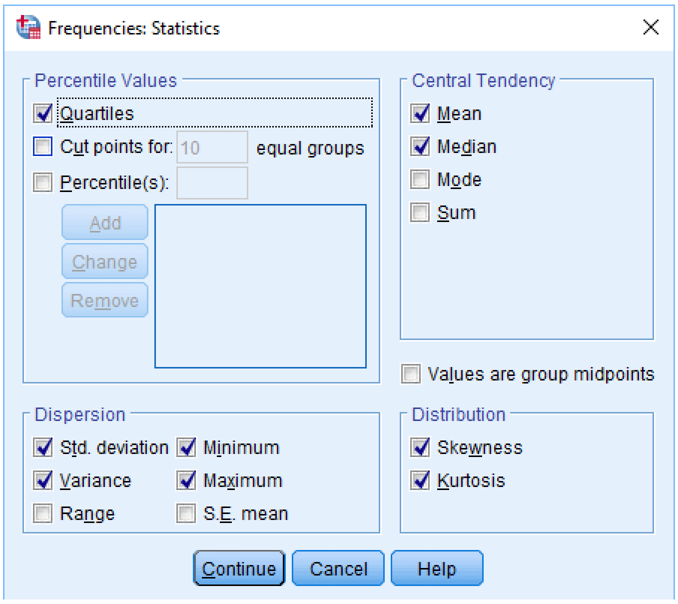


Figure 4.4

Click Continue

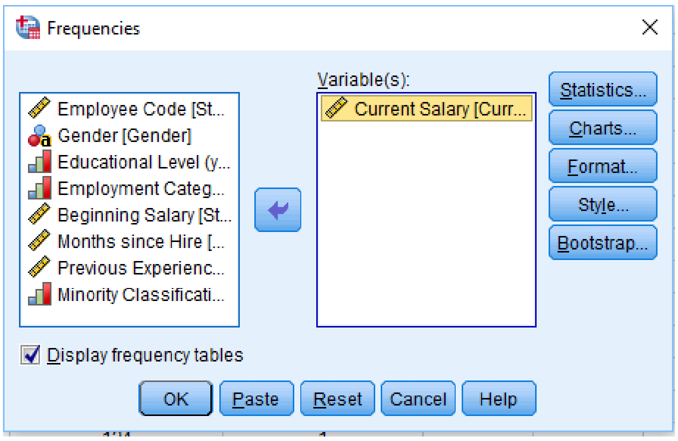


Figure 4.5

Click on Charts

Choose the required statistics

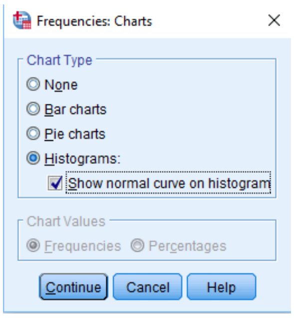


Figure 4.6

Click Continue

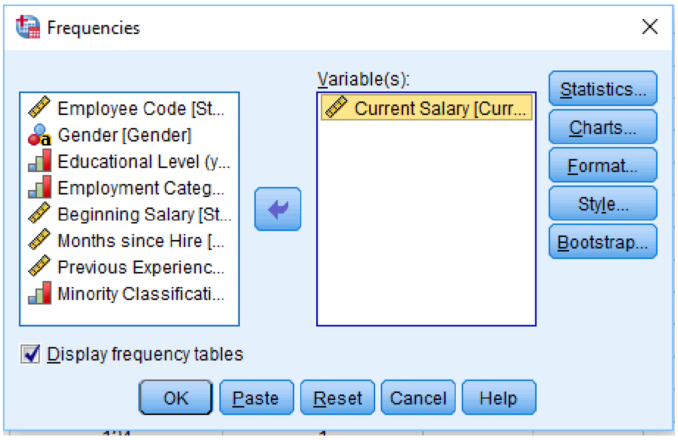


Figure 4.7

Click OK

SPSS output

Statistics

Current Salary

N	Valid	474
	Missing	0
Mean		34419.57
Median		28875.00
Std. Deviation		17075.661
Variance		291578214.5
Skewness		2.125
Std. Error of Skewness		.112
Kurtosis		5.378
Std. Error of Kurtosis		.224
Minimum		15750
Maximum		135000
Percentiles	25	24000.00
	50	28875.00
	75	37162.50

Figure 4.8

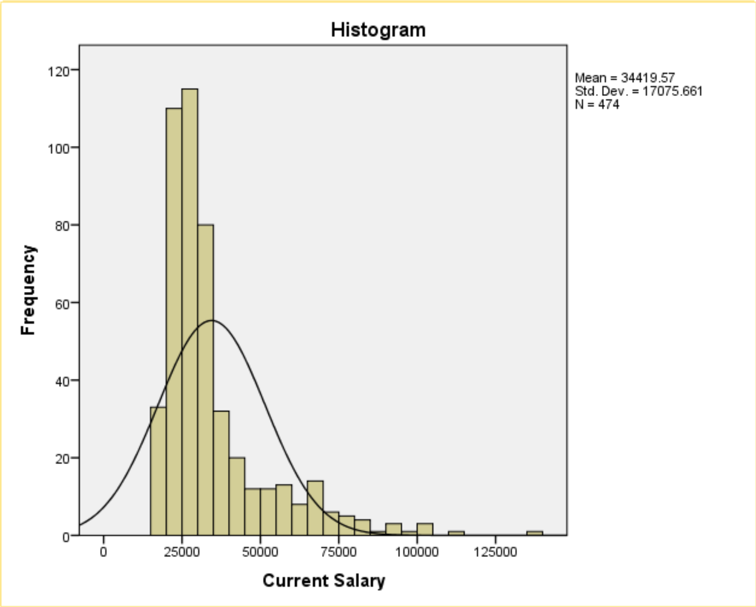


Figure 4.9

Save SPSS Output file: Example4.1.spv

Descriptives

Use descriptives menu to calculate a range of descriptive statistics for the current salary variable (salary).

Select Analyze > Descriptives > Descriptives

Transfer salary to the Variable(s) box

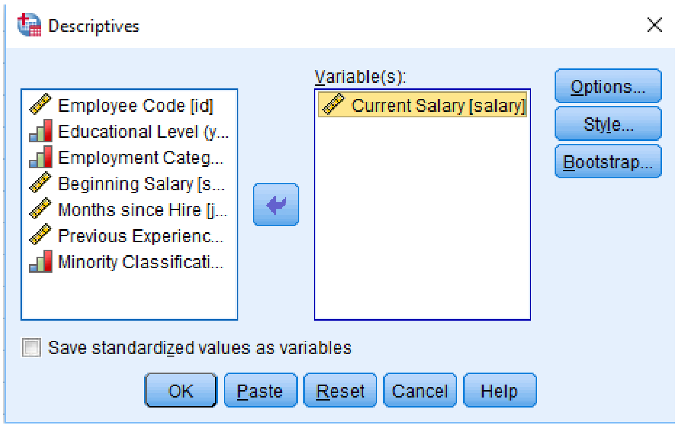


Figure 4.10

Click Options and select the required statistics

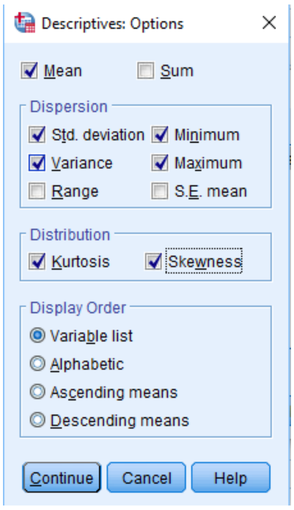


Figure 4.11

Click Continue

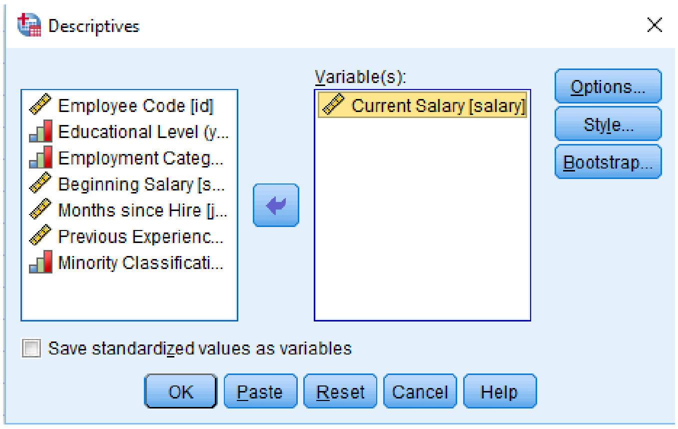


Figure 4.12

Click OK

SPSS output

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Current Salary	474	15750	135000	34419.57	17075.661	291578214.5	2.125	.112	5.378	.224
Valid N (listwise)	474									

Figure 4.13

Save SPSS Output file: Example4.1.spv

From the SPSS solution, we observe the mean salary is 34419.57, standard deviation = 17075.66, etc.

Other SPSS methods to calculate descriptive statistics can be found via the frequencies menu.

Chapter 5 Comparing means using the Students' t test

SPSS contains a large range of parametric statistical testing procedures, including Students' t tests and analysis of variance for three or more samples. In this textbook we limit the discussion to one and two sample t tests. If you are interested, then I have added a document within the online resource which describes factorial experiments and their solution using Microsoft Excel and IBM SPSS.

One sample t test

SPSS one-sample t-test tests if the mean of a single metric variable is equal to some hypothesized population value.

Example 5.1

A local fish shop sells cod to customers with a hypothesised average weight of 400grams. The local trading standards officers have received complaints from customers that the cod are a great deal smaller than the advertised weight provided in the shop window. Trading standards have sampled 40 cod to be tested to confirm if the average cod weight is less than 400 grams?

	Cod_weight_grams
1	415
2	311
3	322
4	352
5	474
6	382
7	288
8	543
9	363
10	381
11	331
12	506
13	430
14	409
15	487
16	294
17	356
18	361
19	388
20	435

Figure 5.1a

	Cod_weight_grams
21	330
22	301
23	301
24	410
25	285
26	331
27	436
28	282
29	244
30	540
31	475
32	355
33	257
34	401
35	251
36	374
37	315
38	293
39	306
40	467

Figure 5.1b

Save SPSS Data file: Example5.1.sav

Quick Data Check

The first part of the analysis is to have a look at your data by using SPSS to create the histogram and to provide summary statistics.

Analyze > Descriptive Statistics > Frequencies

Transfer Body Weight of the Cod to Variable(s) box

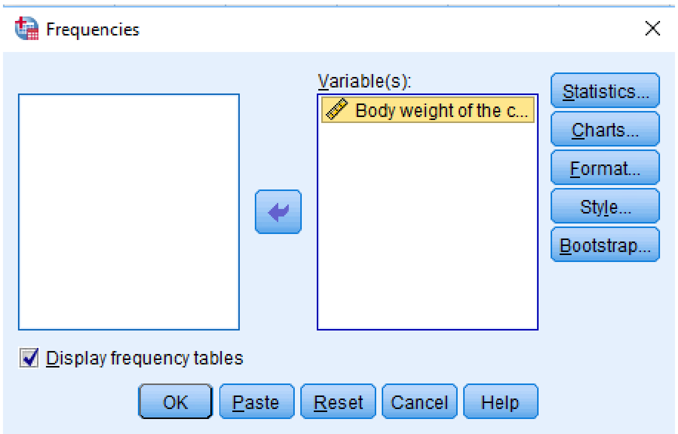


Figure 5.2

Click on Charts

Choose Histograms and Show normal curve on histogram

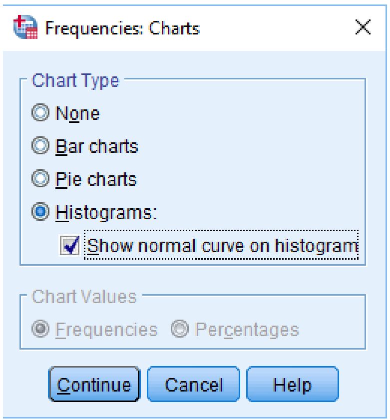


Figure 5.3
Click Continue

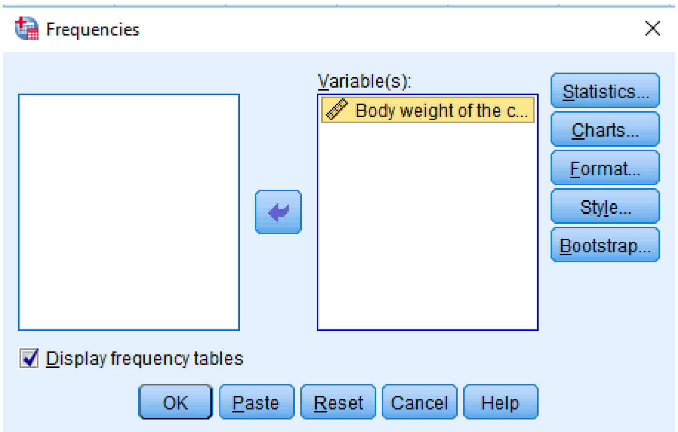


Figure 5.4

Click OK

SPSS Output

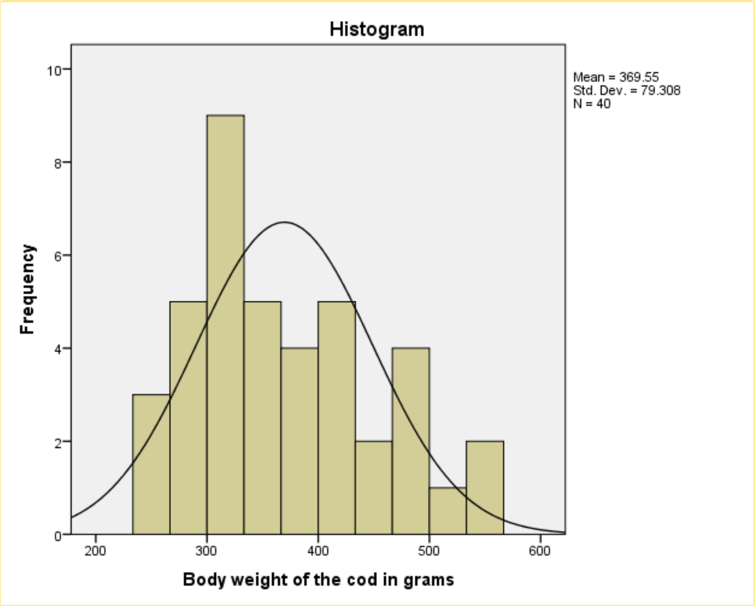


Figure 5.5

Save SPSS Output file: Example5.1.spv

Assumptions One Sample T-Test

Results from statistical procedures can only be taken seriously insofar as relevant assumptions are met. For a one-sample t-test, these are:

1. Independent and identically distributed variables (or, less precisely, “independent observations”).
2. Normality: the test variable is normally distributed in the population. We observe from the histogram above that the data looks approximately normal is shape.

Run SPSS One-Sample T-Test

Analyze > Compare Means > One-Sample T Test

Transfer Body Weight to Test Variable(s) box
Type 400 in Test Value box

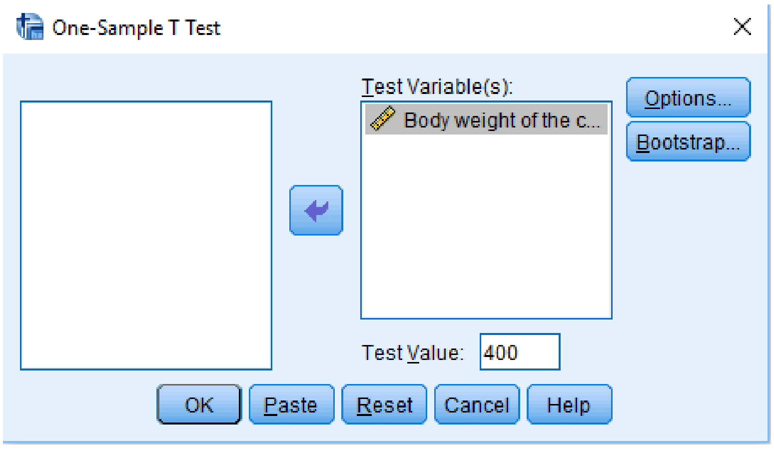


Figure 5.6

Click OK

SPSS output

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Body weight of the cod in grams	40	369.55	79.308	12.540

One-Sample Test

	Test Value = 400					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Body weight of the cod in grams	-2.428	39	.020	-30.450	-55.81	-5.09

Figure 5.7

Resave SPSS Output file: Example5.1.spv

The actual t-test results are found in the One-Sample Test table.

The p value, denoted by “Sig. (2-tailed)” is 0.02; if the population mean is exactly 400 grams, then there's only a 2% chance of finding the result we did. We usually reject the null hypothesis if $p < 0.05$.

We thus conclude that cod do not weigh 400 grams (but probably less than that).

It's important to notice that the p value of 0.02 is 2-tailed. This means that the p value consists of a 1% chance for finding a difference < - 30.45 grams and another 1% chance for finding a difference > 30.45 gram.

The Mean Difference is simply the sample mean minus the hypothesized mean (369.55 - 400 = - 30.45).

Conclusion

Regarding descriptive statistics, the very least we should report, is the mean, standard deviation and N on which these are based. Since these statistics don't say everything about the data, we personally like to include a histogram as well.

We may report the t-test results by writing “we found that, on average, cod weighed less than the 400 grams advertised by the fish shop owner [$t(39) = -2.428$, two-tail p-value = 0.020.”

Two independent sample t-test

SPSS independent samples t-test is a procedure for testing whether the means in two populations on one metric variable are equal. The two populations are identified in the sample by a dichotomous variable. These two groups of cases are considered “independent samples” because none of the cases belong to both groups simultaneously; that is, the samples don't overlap.

Example 5.2

A marketeer wants to know whether women spend the same amount of money on clothes as men. She asks 30 male and 30 female respondents how much many Euros they spend on clothing each month, resulting in example5.5.sav. Do these data contradict the null hypothesis that men and women spend equal amounts of money on clothing?

	Gender	Spent_£_per_month
1	0	168.00
2	0	27.00
3	0	36.00
4	0	68.00
5	0	303.00
6	0	111.00
7	0	12.00
8	0	510.00
9	0	82.00
10	0	109.00
11	0	45.00
12	0	392.00
13	0	201.00
14	0	158.00
15	0	338.00
16	0	16.00
17	0	74.00
18	0	80.00
19	0	121.00
20	0	211.00
21	0	44.00
22	0	20.00
23	0	20.00
24	0	159.00
25	0	11.00
26	0	45.00
27	0	212.00
28	0	9.00
29	0	.00
30	0	500.00

Figure 5.8a

	Gender	Spent_£_per_month
31	1	210.00
32	1	30.00
33	1	3.00
34	1	80.00
35	1	5.00
36	1	48.00
37	1	6.00
38	1	1.00
39	1	3.00
40	1	193.00
41	1	154.00
42	1	18.00
43	1	68.00
44	1	19.00
45	1	40.00
46	1	1.00
47	1	7.00
48	1	393.00
49	1	128.00
50	1	7.00
51	1	94.00
52	1	116.00
53	1	169.00
54	1	66.00
55	1	237.00
56	1	188.00
57	1	41.00
58	1	278.00
59	1	26.00
60	1	12.00

Figure 5.8b

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Gender	Numeric	8	0		{0, Female}...	None	10	Right	Nominal	Input
2	Spent_£_per_month	Numeric	4	2	£'s spent on clothing per month	None	None	14	Right	Scale	Input

Figure 5.9

Save SPSS Data file: Example5.2.sav

Quick Data Check

Before moving on to the actual t-test, we first need to get a basic idea of what the data looks like. We'll take a quick look at the histogram for the amounts spent by running Frequencies.

Analyze > Descriptive Statistics > Frequencies

Transfer £'s spent to the Variables box

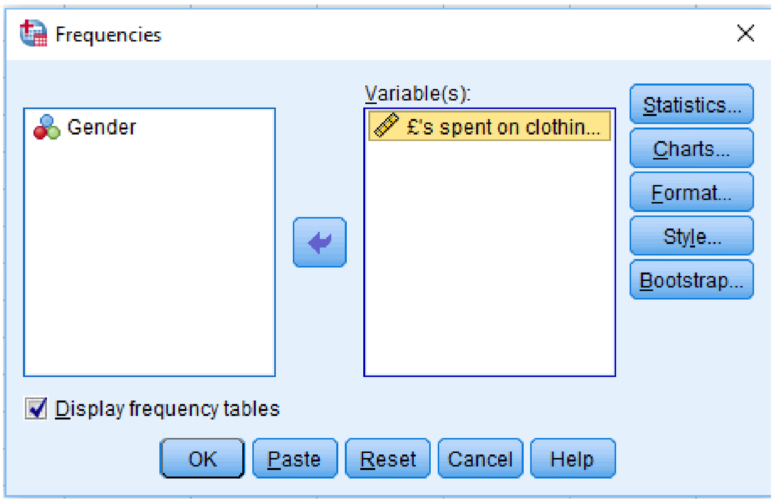


Figure 5.9

Click on Charts

Choose Histograms
Select Show normal curve on histogram

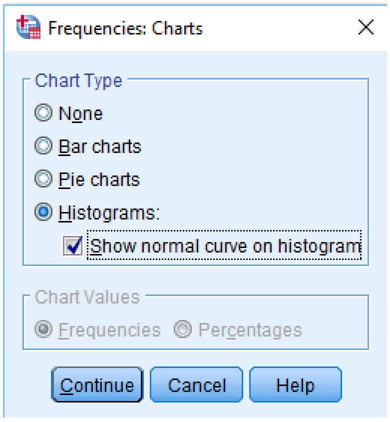


Figure 5.10

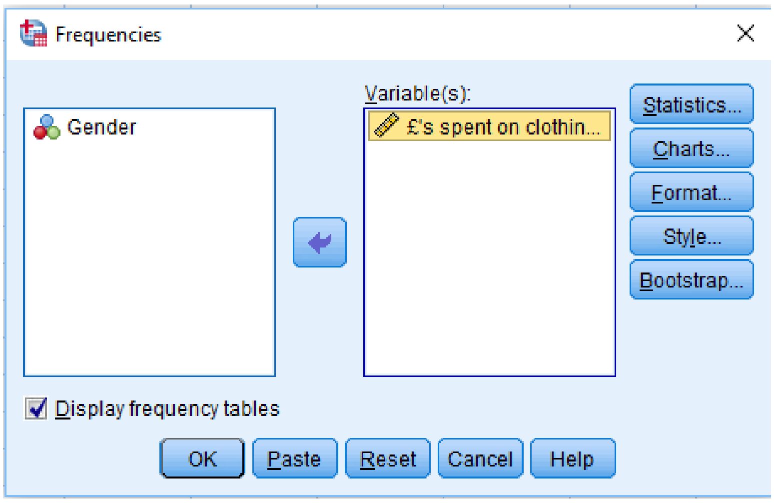


Figure 5.11

Click OK

SPSS Output

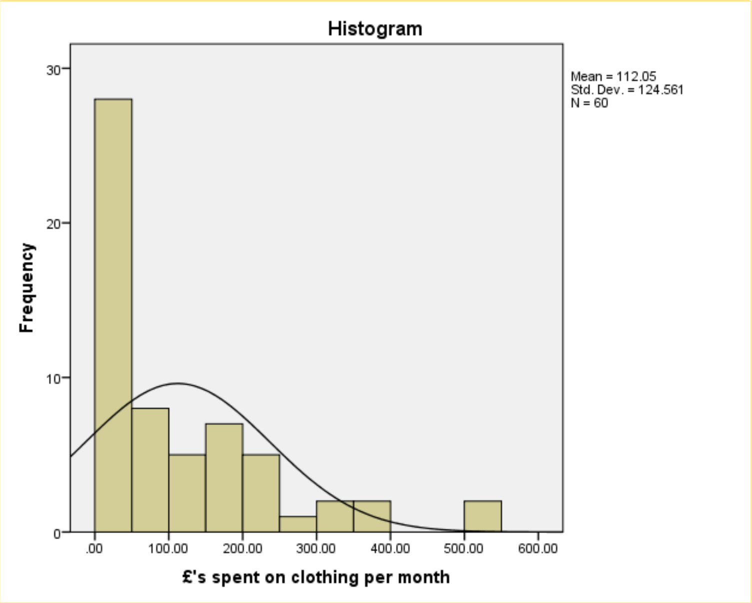


Figure 5.12

Save SPSS Output file: Example5.2.spv

These values look plausible. The maximum monthly amount spent on clothing (around £525) is not unlikely for one or two respondents, the clear majority of whom spend under £100. Also, note that N = 60, which tells us that there are no missing values.

Assumptions Independent Samples T-Test

If we just run our test at this point, SPSS will immediately provide us with relevant test statistics and a p-value. However, such results can only be taken seriously insofar as the independent t-test assumptions have been met. These are:

1. Independent and identically distributed variables (or, less precisely, "independent observations").
2. The dependent variable is normally distributed in both populations.
3. Homoscedasticity: the variances of the populations are equal.

Assumption 1 is mostly theoretical.

Violation of assumption 2 hardly affects test results for reasonable sample sizes (say n >30). If this doesn't hold, perhaps consider a Mann-Whitney test instead of the t-test.

If assumption 3 is violated, test results need to be corrected. For the independent samples t-test, the SPSS output contains the uncorrected as well as the corrected results by default.

Run SPSS Independent Samples T-Test

Analyze > Compare Means > Independent-Samples T Test

Transfer £'s spent to Test Variable(s) box
Transfer Gender to Grouping Variable box
Click on Define box and choose groups as 0 and 1

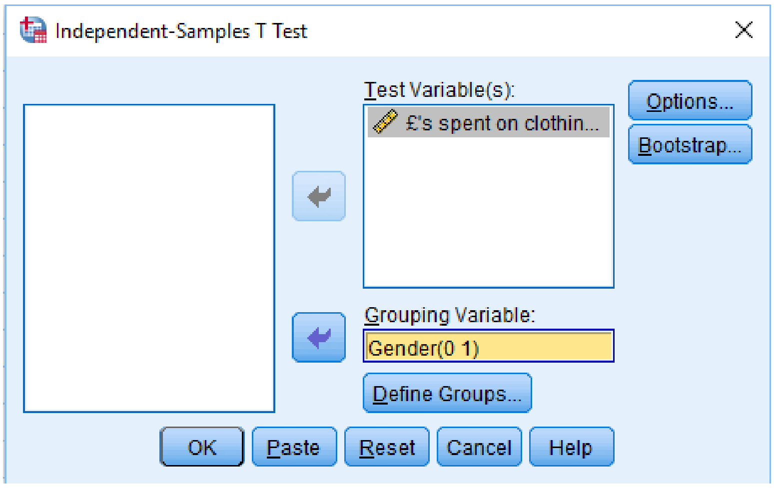


Figure 5.13

Click OK

SPSS Output

From the first table, showing some basic descriptives, we see that 30 female and 30 male respondents are included in the test. Female respondents spent an average of £136 on clothing each month. For male respondents this is only £88. The difference is roughly £48.

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
£'s spent on clothing per month	Female	30	136.0667	143.14303	26.13422
	Male	30	88.0333	99.41223	18.15011

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
£'s spent on clothing per month	Equal variances assumed	2.548	.116	1.510	58	.137	48.03333	31.81861	-15.65853	111.72520
	Equal variances not assumed			1.510	51.695	.137	48.03333	31.81861	-15.82434	111.89101

Figure 5.14

Resave SPSS Output file: Example5.2.spv

As shown in the screenshot, the t-test results are reported twice. The first line ("equal variances assumed") assumes that the assumption of equal variances has been met. If this assumption doesn't hold, the t-test results need to be corrected. These corrected results are presented in the second line ("equal variances not assumed").

Whether the assumption of equal variances holds is evaluated using Levene's test for the equality of variances. As a rule of thumb, if Sig. > 0.05, use the first line of t-test results. Reversely, if its p-value ("Sig.") < 0.05 we reject the null hypothesis of equal variances and thus use the second line of t-test results.

The difference between the amount spent by men and women is around £48. The chance of finding this or a larger absolute difference between the two means is about 14%. Since this is a fair chance, we do not reject the hypothesis that men and women spend equal amounts of money on clothing.

Note that the p-value is two-tailed. This means that the 14% chance consists of a 7% chance of finding a mean difference smaller than £48 and another 7% chance for a difference larger than £48.

Conclusion

When reporting the results of an independent samples t-test, we usually present a table with the sample sizes, means and standard deviations. Regarding the significance test, we'll state that "on average, men did not spend a different amount than women; $t(58) = 1.5$, $p = 0.14$."

Two dependent sample t-test

SPSS paired samples t-test is a procedure for testing whether the means of two metric variables are equal in some population. Both variables have been measured on the same cases. Although "paired samples" suggests that multiple samples are involved, there's really only one sample and two variables.

Example 5.3

A local microbrewery advertises that drinking a pint of their special brew beer will dull the senses and affect reaction times to complete everyday tasks. The microbrewery decides to test this by randomly selecting 30 participants and asking them to perform some tasks before and after having a beer and records their reaction times. For each participant, she calculates the average reaction time over tasks both before and after the beer, resulting in example5.3.sav. Can we conclude from these data that a single beer affects reaction time?

	id	Before_time_average	After_time_average
1	1	992	1452
2	2	1110	1533
3	3	1086	1280
4	4	1442	1504
5	5	927	1093
6	6	1080	1291
7	8	1122	1405
8	10	826	999
9	12	1358	1397
10	14	1016	1137
11	15	1242	1427
12	17	1078	1128
13	18	1144	1272
14	20	1198	1430
15	21	1000	1229
16	22	1039	1180
17	23	1213	1485
18	24	1382	1576
19	25	1416	1578
20	26	900	974
21	27	1259	1321
22	28	1056	1015
23	30	1215	1255
24	31	1455	1315
25	32	1337	1258
26	33	994	974
27	34	985	980
28	36	1110	1144
29	38	1169	1379
30	39	1825	1631

Figure 5.15

Save SPSS Data file: Example5.3.sav

Quick Data Check

Graphs > Legacy Dialogs > Scatter/Dot

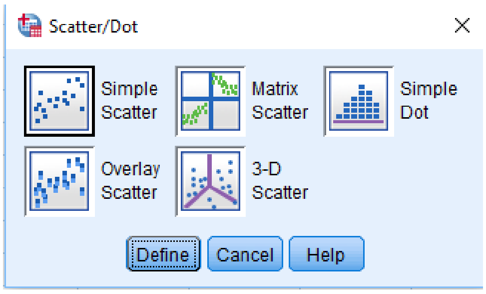


Figure 5.16

Choose Simple Scatter

Click on Define

We then move Before_time_average and After_time_average to X-Axis and Y-Axis.

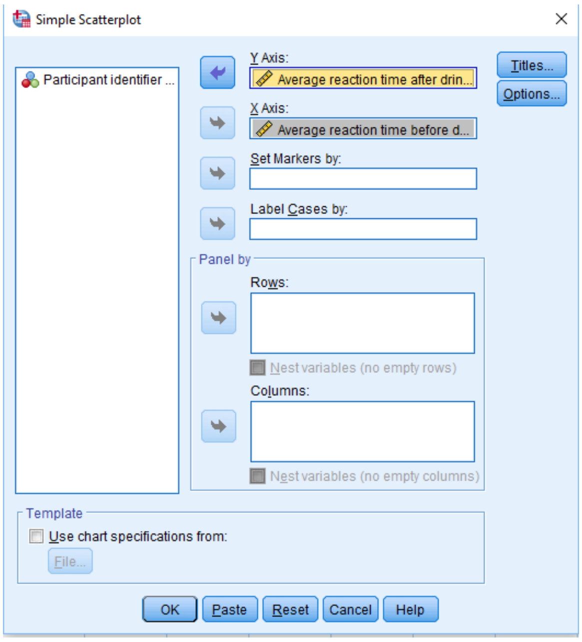


Figure 5.17

Click OK

SPSS Output

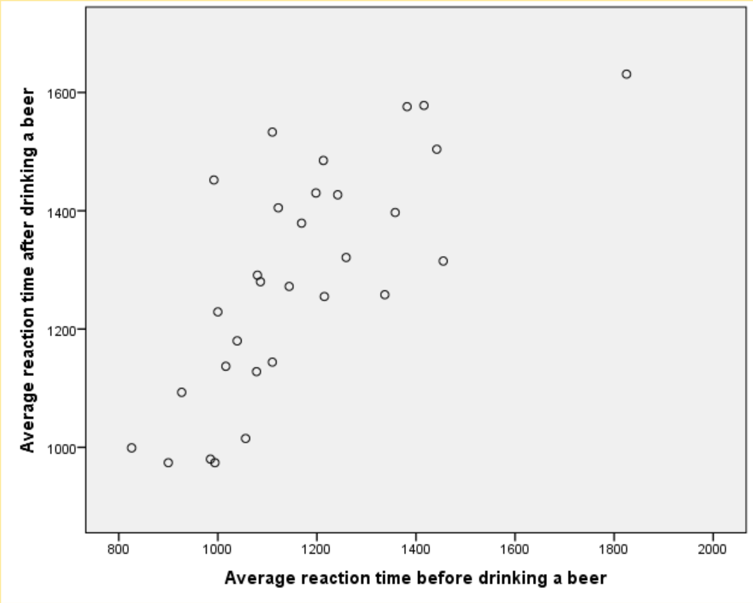


Figure 5.18

Save SPSS Output file: Example5.3.spv

Normal reactions times are between 800 and 1500 ms (= milliseconds). Neither variable has any values that are way out of this normal range, so the data seem plausible. We also see a substantial positive correlation between the variables; respondents who were fast on the first task tend to be fast on the second task as well. The graph seems to suggest that the mean reaction time before a beer is somewhere near 1100 ms (vertical axis) and after a beer perhaps 1300 ms (horizontal axis).

One respondent (right top corner, denoted "outlier") is remarkably slow compared to the others. However, we decide that its scores are not extreme enough to justify removing it from the data.

Assumptions Paired Samples T-Test

SPSS will happily provide us with test results, but we can only take those seriously insofar as the assumptions for our test are met. For the paired samples t-test, these are:

1. Independent observations or, more precisely, independent and identically distributed variables;
2. The difference scores between the two variables must be normally distributed in our population.

The first assumption is often satisfied if each case (row of data values) holds a distinct person or other unit of analysis. The normality assumption is mostly relevant for small sample sizes (say N < 30). If it's violated, consider a Wilcoxon signed-ranks test instead of a t-test. However, our data seems to meet both assumptions, so we'll proceed to the t-test.

Run SPSS Paired Samples T-Test

Analyze > Compare Means > Paired-Samples T Test

Select both variables and move them into the Paired Variables box.

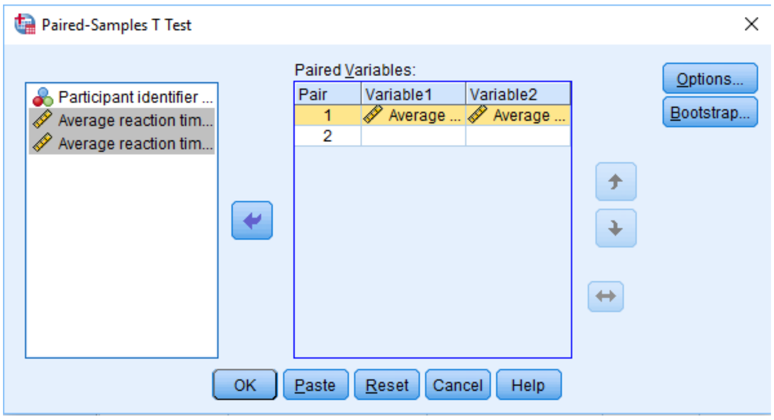


Figure 5.19

Click on Options

Type in 95% into the Confidence Interval Percentage box

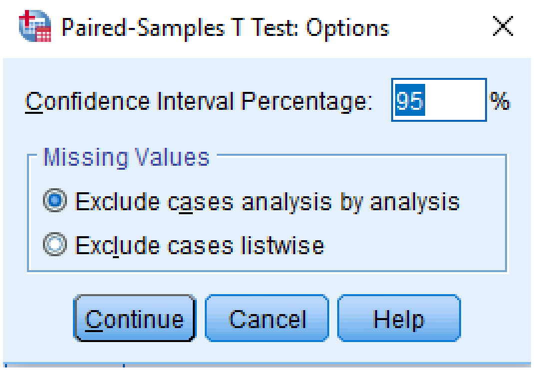


Figure 5.20

Click Continue

Click OK

SPSS Output

The first table ("Paired Samples Statistics") presents the descriptive statistics we'll report. Since N = 30, we don't have any missing values on the test variables and as expected, the mean reaction time before a beer (1166 ms) is lower than after a beer (1288 ms).

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Average reaction time before drinking a beer	1165.87	30	206.274	37.660
	Average reaction time after drinking a beer	1288.07	30	195.374	35.670

Figure 5.21

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Average reaction time before drinking a beer & Average reaction time after drinking a beer	30	.736	.000

Figure 5.22

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Average reaction time before drinking a beer - Average reaction time after drinking a beer	-122.200	146.353	26.720	-176.849	-67.551	-4.573	29	.000

Figure 5.23

Resave SPSS Output file: Example5.3.spv

On average, respondents slow down some 122 ms. We could have calculated this from the first table ourselves. The p-value denoted by “Sig. (2-tailed)” is 0.000 (If we double-click it, we'll see it's precisely 0.000083, meaning a 0.0083 % chance.)

Sig. (2-tailed)
.000

0.000083

Figure 5.24

So, if the population means are equal, there's a 0% chance of finding this result. We therefore reject the null hypothesis. Even a single beer slows people down on the given tasks.

Note that the p-value is two-sided. This means that the p-value consists of a 0.00415% chance of finding a difference < - 122 ms and another 0.00415% chance of finding a difference > 122 ms.

Conclusion

As we mentioned before, we'll always report the descriptive statistics obtained from the paired samples t-test. For the significance test, we may write something like “Participants became slower after drinking a single beer, t(29) = - 4.573, p = 0.000”.

Chapter 6 Chi-square and non-parametric tests

Chi-square test of association

Example 6.1

A sample of 183 year 1 students evaluated several undergraduate business courses at a local university. The data are stored in the SPSS data file example6.1.sav.

Data view

First 20 records out of 183 student records

	First_name	Surname	Gender	Module	q1	q2	q3	q4	q5	q6
1	Thomas	Adams	1	4	4	3	3	4	4	2
2	Skylar	Adams	0	2	4	3	4	4	3	4
3	Hailey	Adams	0	4	4	3	2	3	3	4
4	Amelia	Allen	0	3	3	4	5	4	3	4
5	William	Allen	1	4	3	3	3	4	4	1
6	Brayden	Allen	1	4	4	4	4	4	5	4
7	Lily	Anderson	0	4	3	5	4	3	4	3
8	Lydia	Anderson	0	1	4	5	4	4	4	3
9	Kylie	Anderson	0	1	4	5	5	3	5	5
10	Gabriel	Anderson	1	2	4	4	5	3	4	4
11	Wyatt	Baker	1	2	5	5	4	5	5	5
12	Henry	Baker	1	3	4	4	3	4	4	3
13	Jayden	Baker	1	5	2	3	4	3	2	3
14	Robert	Baker	1	3	5	4	4	2	4	2
15	Easton	Brown	1	4	4	4	5	3	3	3
16	Kevin	Brown	1	4	4	4	5	5	4	5
17	David	Brown	1	4	4	4	5	4	4	5
18	Lucas	Brown	1	5	4	4	4	2	4	3
19	Ellie	Campbell	0	3	2	2	3	3	3	3
20	Dylan	Campbell	1	2	5	5	5	4	5	4

Figure 6.1 Data view

Save SPSS Data file: Example6.1.sav

We'd now like to know:

Is study course associated with gender?

Since course and gender are nominal variables, we'll run a chi-square test to find out. Chi-square independence test can be trusted if two assumptions are met:

1. independent observations. This usually -not always- holds if each case in SPSS holds a unique person or other statistical unit. Since this is that case for our data, we'll assume this has been met.
2. For a 2 by 2 table, all expected frequencies > 5.* For a larger table, no more than 20% of all cells may have an expected frequency < 5 and all expected frequencies > 1.

SPSS will test this assumption for us when we'll run our test.
Select Analyze > Descriptive Statistics > Crosstabs

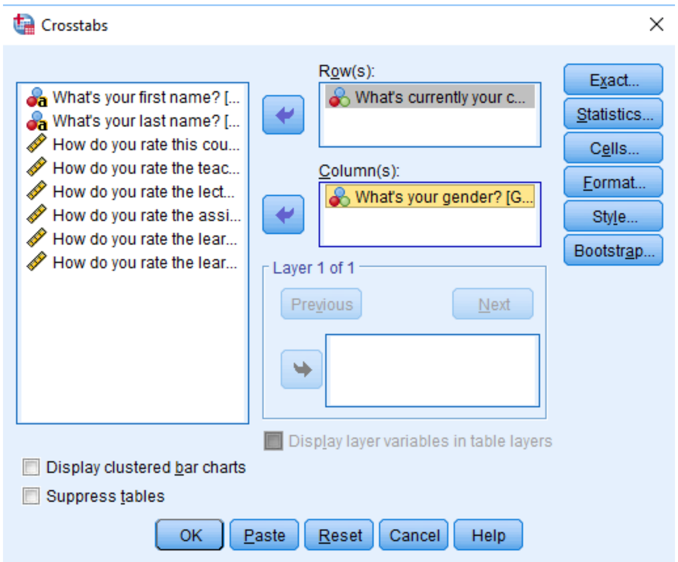


Figure 6.2

Click on Statistics and choose Chi-Square

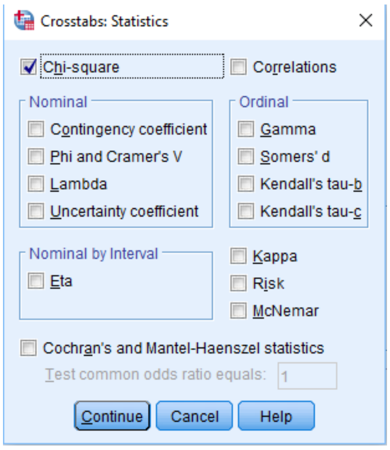


Figure 6.3

Click Continue

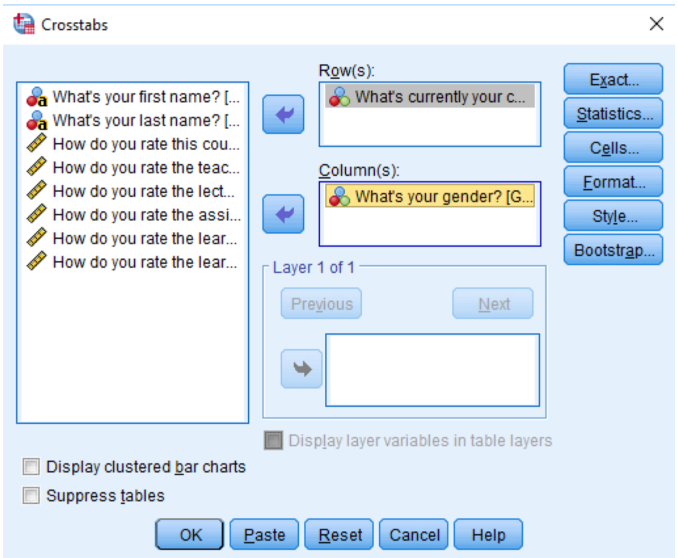


Figure 6.4

Click OK

SPSS output

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
What's currently your course? * What's your gender?	183	100.0%	0	0.0%	183	100.0%

Figure 6.5

First off, we take a quick look at the Case Processing Summary to see if any cases have been excluded due to missing values. In this example, no data missing.

What's currently your course? * What's your gender? Crosstabulation

Count

		What's your gender?		Total
		female	male	
What's currently your course?	e-commerce	54	8	62
	economics	7	28	35
	marketing	12	21	33
	human resources	15	22	37
	Other	4	12	16
Total		92	91	183

Figure 6.6

Next, we inspect our contingency table. Note that its marginal frequencies -the frequencies reported in the margins of our table- show the frequency distributions of either variable separately.
Both distributions look plausible and since there's no “no answer” categories, there's no need to specify any user missing values.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	54.504[a]	4	.000
Likelihood Ratio	59.758	4	.000
Linear-by-Linear Association	25.597	1	.000
N of Valid Cases	183		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 7.96.

Figure 6.7

Save SPSS Output file: Example6.1.spv

First off, our data meet the assumption of all expected frequencies > 5 that we mentioned earlier. Since this assumption holds, we can rely on our significance test for which we use the Pearson Chi-Square test statistic.

Right, we usually say that the association between two variables is statistically significant if Asymptotic Significance (2-sided) < 0.05. Significance is often referred to as "p", short for probability; it is the probability of observing our sample outcome if our variables are independent in the entire population.

Therefore, the asymptotic significance (2 sided) p-value = 0.000 < 0.05.

Conclusion:

we reject the null hypothesis that our variables are independent in the entire population

We report the significance test with something like

An association between gender and study course was observed, $\chi^2(4) = 54.504$, $p = 0.000$

Further, I suggest including our final contingency table (with frequencies and row percentages) in the report as well as it gives a lot of insight into the nature of the association.

Chi-square test of goodness-of-fit

SPSS one-sample chi-square test is used to test whether a single categorical variable follows a hypothesized population distribution.

Example 6.2

A marketer believes that 4 smartphone brands are equally attractive. He asks 44 people which brand they prefer, resulting in example6.2.sav. If the brands are equally attractive, each brand should be chosen by roughly the same number of respondents. In other words, the expected frequencies under the null hypothesis are 11 cases for each brand (44 cases/4brands = 11). The more the observed frequencies differ from these expected frequencies, the less likely it is that the brands really are equally attractive.

SPSS data

	brand_appeal
1	1
2	1
3	1
4	4
5	4
6	1
7	2
8	1
9	2
10	3
11	2

	brand_appeal
12	3
13	4
14	3
15	3
16	1
17	3
18	4
19	3
20	3
21	1
22	1

	brand_appeal
23	1
24	3
25	3
26	3
27	3
28	3
29	4
30	2
31	2
32	3
33	1

	brand_appeal
34	3
35	1
36	4
37	1
38	3
39	3
40	2
41	1
42	1
43	2
44	3

Figure 6.8 a - d

Save SPSS Data file: Example6.2.sav

Quick Data Check

Before running any statistical tests, we always want to have an idea what our data basically look like. In this case we'll inspect a histogram of the preferred brand by running FREQUENCIES.

Analyze > Descriptive Statistics > Frequencies

Transfer Which Brand to Variable(s) box

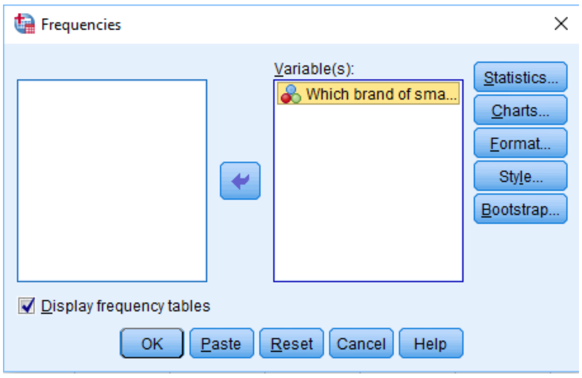


Figure 6.9

Click on Charts

Choose Histograms
Choose Show normal curve on histogram

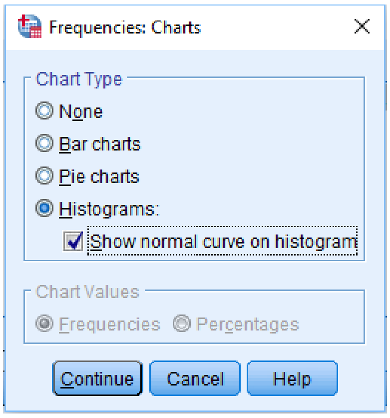


Figure 6.10

Click on Continue

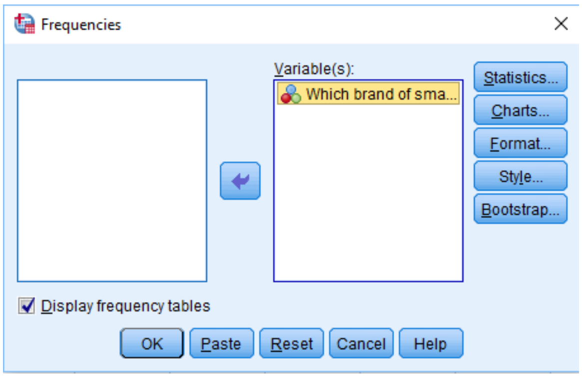


Figure 6.11

Click on OK

SPSS Output

Which brand of smartphone do you prefer?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Samsung	14	31.8	31.8	31.8
	HTC	7	15.9	15.9	47.7
	Apple	17	38.6	38.6	86.4
	Other	6	13.6	13.6	100.0
	Total	44	100.0	100.0	

Figure 6.12

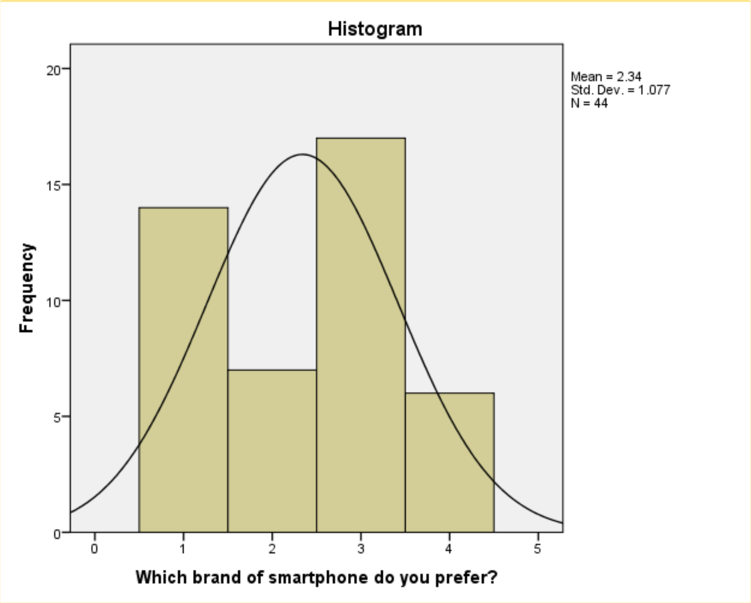


Figure 6.13

Save SPSS Output file: Example6.2.spv

First, N = 43 means that the histogram is based on 43 cases. Since this is our sample size, we conclude that no missing values are present. SPSS also calculates a mean and standard deviation, but these are not meaningful for nominal variables, so we'll just ignore them.

Second, the preferred brands have rather unequal frequencies, casting some doubt upon the null hypothesis of those being equal in the population.

Assumptions One-Sample Chi-Square Test

1. independent and identically distributed variables (or “independent observations”);
2. none of the expected frequencies are < 5;

The first assumption is a design issue and we will presume this assumption has been met.

Whether assumption 2 holds is reported by SPSS whenever we run a one-sample chi-square test.

Run SPSS One Sample Chi-Square Test

Analyze > Nonparametric Tests > Legacy Dialog > Chi-square

Transfer which brand of smartphone to the Test Variable List

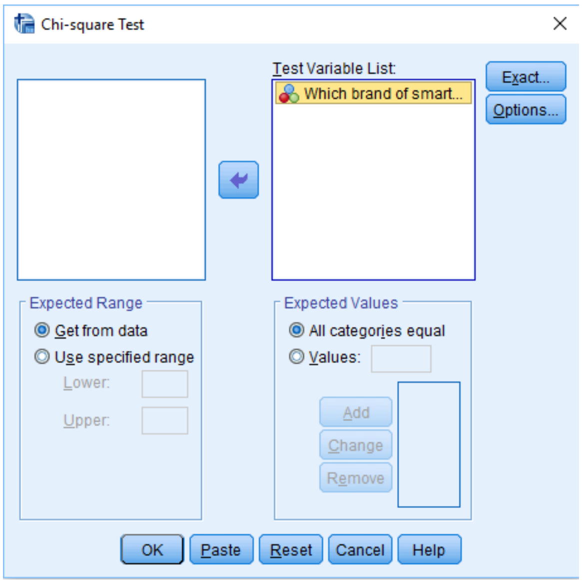


Figure 6.14

Click OK

SPSS Output

Which brand of smartphone do you prefer?

	Observed N	Expected N	Residual
Samsung	14	11.0	3.0
HTC	7	11.0	-4.0
Apple	17	11.0	6.0
Other	6	11.0	-5.0
Total	44		

Figure 6.15

Under Observed N we find the observed frequencies that we saw previously.

Under Expected N we find the theoretically expected frequencies (= 11).

For each frequency the Residual is the difference between the observed and the expected frequency and thus expresses a deviation from the null hypothesis.

Test Statistics

	Which brand of smartphone do you prefer?
Chi-Square	7.818[a]
df	3
Asymp. Sig.	.050

a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 11.0.

Figure 6.16

The Chi-Square test statistic sort of summarizes the residuals and hence indicates the overall difference between the data and the hypothesis.

The larger the chi-square value, the less the data “fit” the null hypothesis.

Degrees of freedom (df) specifies which chi-square distribution applies;

Asymp. Sig. refers to the p value and is 0.050 in this case. If you double click on the table and this bumber value for p (= 0.050) then it will give you the number to a greater number of decimal places (Asym. Sig (p-value) = 0.049922 ≈ 0.05).

Test Statistics

	Which brand of smartphone do you prefer?
Chi-Square	7.818[a]
df	3
Asymp. Sig.	0.049923

a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 11.0.

Figure 6.17

Resave SPSS Output file: Example6.2.spv

If the brands are exactly equally attractive in the population, there's a 5.0% chance of finding our observed frequencies or a larger deviation from the null hypothesis. We usually reject the null hypothesis if $p < 0.05$. Since this is not the case, we conclude that the brands are equally attractive in the population.

Conclusion

When reporting a one-sample chi-square test, we always report the observed frequencies. The expected frequencies usually follow readily from the null hypothesis so reporting them is optional.

Regarding the significance test, we usually write something like “we could not demonstrate that the four brands are not equally attractive; $\chi^2(3) = 7.818$, p-value = 0.05.”

Cochran Q test

SPSS Cochran Q test is a procedure for testing if the proportions of 3 or more dichotomous variables are equal in some population. These outcome variables have been measured on the same people or other statistical units.

Example 6.3

The principal of some university wants to know whether three examinations are equally difficult. Fifteen students took these examinations. The results are as follows:

	id	Result_1	Result_2	Result_3
1	1	0	1	1
2	2	0	1	1
3	3	1	1	1
4	4	1	0	1
5	5	0	0	0
6	6	1	1	1
7	7	1	1	1
8	8	1	0	1
9	9	1	1	1
10	10	0	0	1
11	11	1	1	1
12	12	1	1	1
13	13	1	0	1
14	14	1	0	0
15	15	0	0	1

Figure 6.18

Save SPSS Data file: Example6.3.sav

Quick Data Check

It's always a good idea to take a quick look at what the data look like before proceeding to any statistical tests. We'll open the data and inspect some histograms by running FREQUENCIES.

Analyze > Descriptive Statistics > Frequencies

Transfer the three test variables to the Variables box

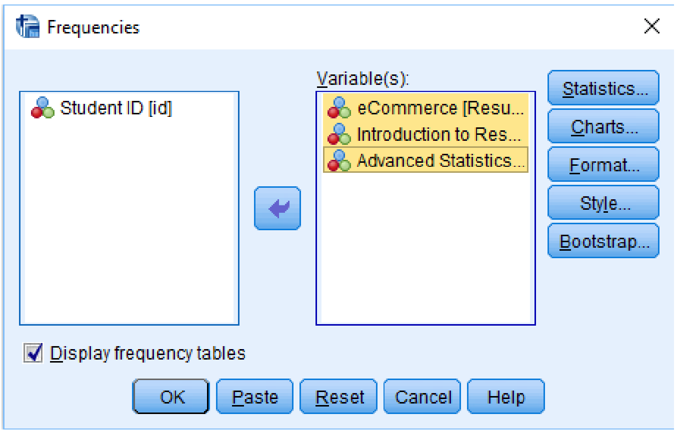


Figure 6.19

Click on Charts
Choose Histograms

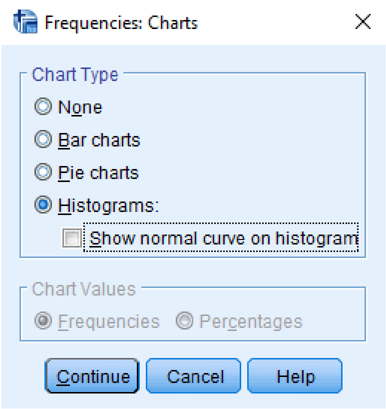


Figure 6.20

Click on Continue

Click on OK

SPSS Output

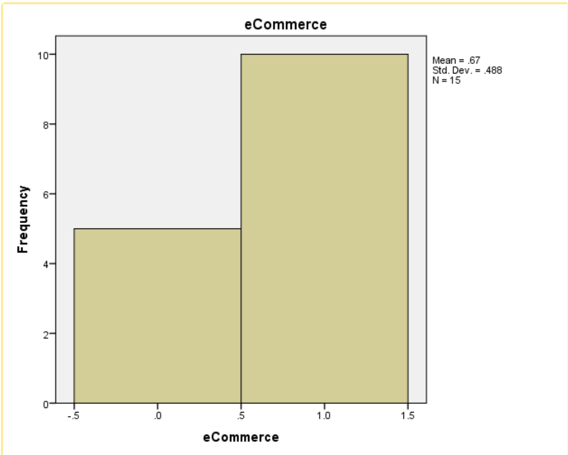


Figure 6.21

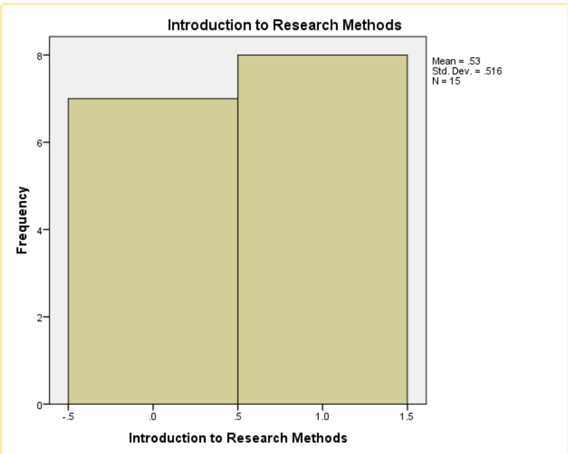


Figure 6.22

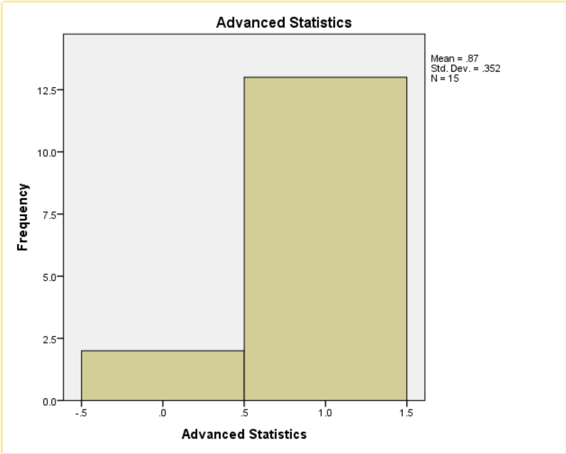


Figure 6.23

The histograms indicate that the three variables are indeed dichotomous (there could have been some “Unknown” answer category but it doesn't occur). Since N = 15 for all variables, we conclude there's no missing values. Values 0 and 1 represent “Failed” and “Passed”.

We therefore readily see that the proportions of students succeeding range from 0.53 to 0.87.

Save SPSS Output file: Example6.3.spv

Assumptions Cochran Q Test

Cochran's Q test requires only one assumption:

Independent observations (or, more precisely, independent and identically distributed variables);

Running SPSS Cochran Q Test

Analyze > Nonparametric Tests > Legacy Dialogs > K Related Samples

Transfer the three test variables to the Variables box
Choose Cochran's Q

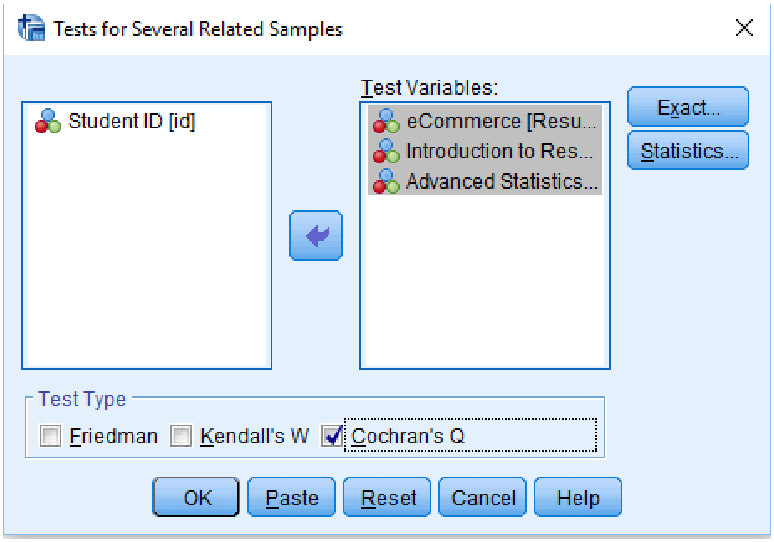


Figure 6.24

Click on Statistics and choose Descriptives

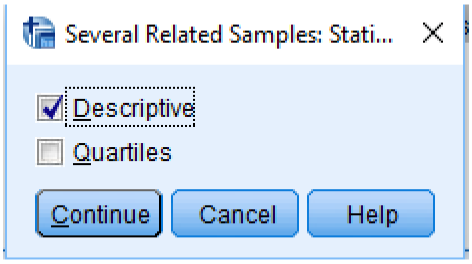


Figure 6.25

Click Continue

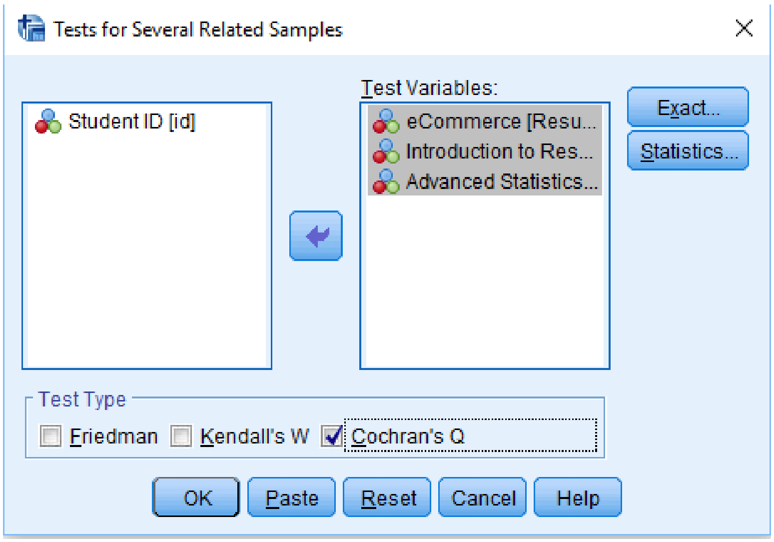


Figure 6.26

Click OK

SPSS Output

The first table (Descriptive Statistics) presents the descriptives we'll report.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
eCommerce	15	.67	.488	0	1
Introduction to Research Methods	15	.53	.516	0	1
Advanced Statistics	15	.87	.352	0	1

Figure 6.27

Since N = 15, the descriptives once again confirm that there are no missing values and the proportions range from 0.53 to 0.87.

Cochran Test

Frequencies

	Value	
	0	1
eCommerce	5	10
Introduction to Research Methods	7	8
Advanced Statistics	2	13

Test Statistics

N	15
Cochran's Q	4.750[a]
df	2
Asymp. Sig.	.093

a. 0 is treated as a success.

Figure 6.28

Resave SPSS Output file: Example6.3.spv

The table Test Statistics presents the result of the significance test. The p-value ("Asymp. Sig.") is 0.093; if the three tests really are equally difficult in the population, there's still a 9.3% chance of finding the differences we observed in this sample. Since this chance is larger than 5%, we do not reject the null hypothesis that the tests are equally difficult.

Conclusion

When reporting the results from Cochran's Q test, we first present the descriptive statistics. Cochran's Q statistic follows a chi-square distribution, so we'll report something like "Cochran's Q test did not indicate any differences among the three proportions, $\chi^2(2) = 4.75$, p-value = 0.093.

Binomial test

SPSS binomial test is used for testing whether a proportion from a single dichotomous variable is equal to a presumed population value.

Example 6.4

A university claims that 75% of the student population is female. A random sample of 15 students are identified and 7 are found to be female. Is there any evidence for the claim to be true?

SPSS data

	id	gender
1	1	1
2	2	1
3	3	1
4	4	0
5	5	0
6	6	0
7	7	1
8	8	1
9	9	0
10	10	1
11	11	1
12	12	0
13	13	1
14	14	0
15	15	0

Figure 6.29

Save SPSS Data file: Example6.4.sav

Data Check

Let's first take a quick look at the FREQUENCIES for gender. Like so, we can inspect whether there are any missing values and whether the variable is really dichotomous. We'll run some FREQUENCIES.

Analyze > Descriptives > Frequencies

Transfer Gender variable to the Variables box

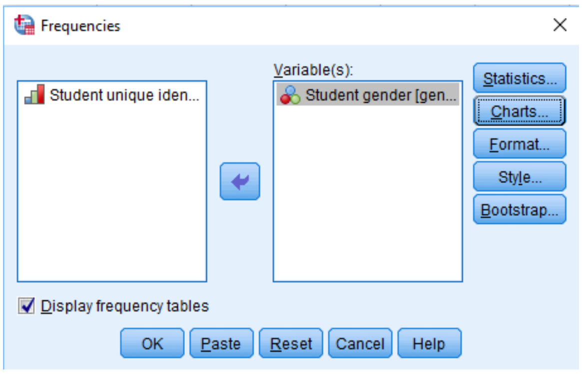


Figure 6.30

Click OK

SPSS Output

Statistics

Student gender

N	Valid	15
	Missing	0

Student gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	7	46.7	46.7	46.7
	Male	8	53.3	53.3	100.0
	Total	15	100.0	100.0	

Figure 6.31

Save SPSS Output file: Example6.4.spv

The output tells us that there are no missing values and the variable is indeed dichotomous. We can proceed our analysis with confidence.

Assumptions Binomial Test

For the binomial test we need just one:

> Independent observations (or, more precisely, independent and identically distributed variables).

Run SPSS Binomial Test

We'd like to test whether the proportion of females differs from 0.75 (our test proportion). Now SPSS Binomial Test has a very odd feature: the test proportion we enter applies to the category that's first encountered in the data.

So, the hypothesis that's tested depends on the order of the cases. Because our test proportion applies to females (rather than males), we need to make sure that our females are at the top of the data file.

Select Data > Sort Cases

> Transfer Student Gender to Sort by box

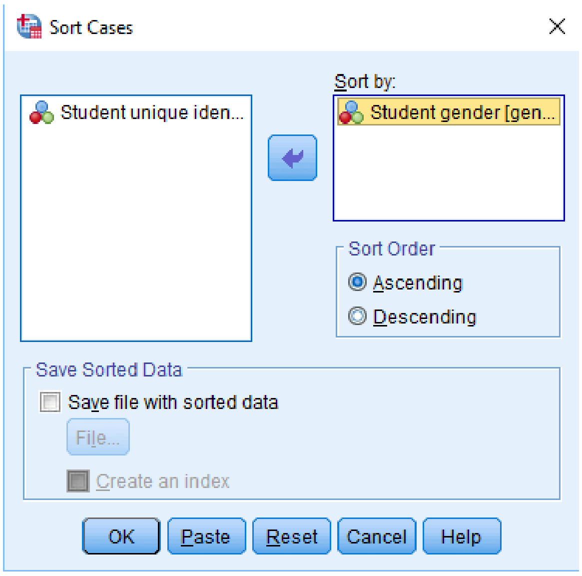


Figure 6.32

Click OK

	id	gender
1	4	0
2	5	0
3	6	0
4	9	0
5	12	0
6	14	0
7	15	0
8	1	1
9	2	1
10	3	1
11	7	1
12	8	1
13	10	1
14	11	1
15	13	1

Figure 6.33

Next, we'll run the actual binomial test.

Analyze > Nonparametric Test > Legacy Dialog > Binomial

Transfer Student Gender to Test Variable List box

Type 0.75 into Test Proportion box

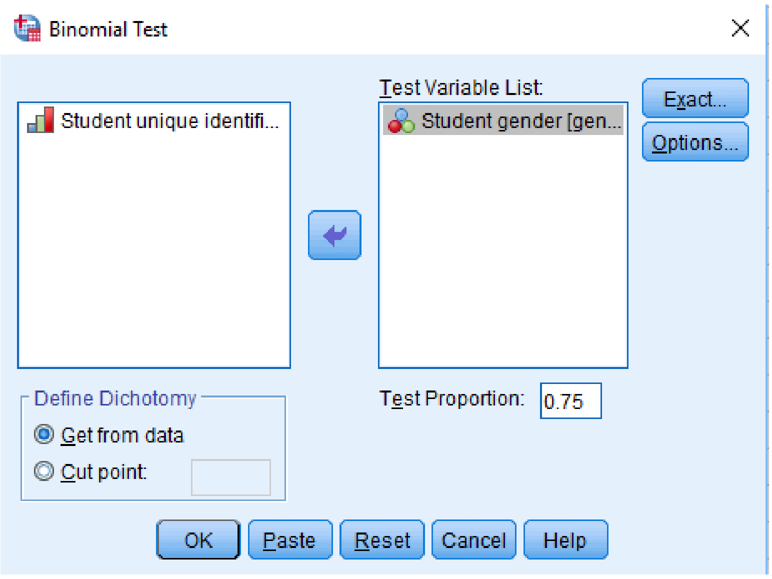


Figure 6.34

Click OK

SPSS Output

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)
Student gender	Group 1	Female	7	.47	.75	.017[a]
	Group 2	Male	8	.53		
	Total		15	1.00		

a. Alternative hypothesis states that the proportion of cases in the first group < .75.

Figure 6.35

Resave SPSS Output file: Example6.4.spv

Since we have 7 females out of 15 observations, the observed proportion is (7 / 15 = 0.47).

Our null hypothesis states that this proportion is 0.75 for the entire population.

The p-value denoted by Exact Sig. (1-tailed) is 0.017. If the proportion of females is exactly 0.75 in the entire population, then there's only a 1.7% chance of finding 7 or fewer female spiders in a sample of N = 15. We often reject the null hypothesis if this chance is smaller than 5% ($p < 0.05$). We conclude that the proportion of females is not 0.75 in the population but probably (much) lower.

Note that the p value is the chance of finding the observed proportion or a “more extreme” outcome. If the observed proportion is smaller than the test proportion, then a more extreme outcome is an even smaller proportion than the one we observe. We ignore the fact that finding very large proportions would also contradict our null hypothesis. This is what's meant by (1-tailed).*

Conclusion

Regarding the significance test, we'll write something like “a binomial test indicated that the proportion of females of 0.47 was lower than the expected 0.75, p = 0.017 (1-sided)”.

McNemar test

Example 6.5

A marketer wants to know whether two products are equally appealing. He asks 20 participants to try out both products and indicate whether they'd consider buying each products (“yes” or “no”). This results in product_appeal.sav. The proportion of respondents answering “yes, I'd consider buying this” indicates the level of appeal for each of the two products.

The null hypothesis is that both percentages are equal in the population.

SPSS data file

	id	product_a	product_b
1	1	1	1
2	2	0	0
3	3	1	1
4	4	0	1
5	5	0	1
6	6	0	0
7	7	0	0
8	8	0	1
9	9	0	0
10	10	1	1
11	11	0	1
12	12	0	1
13	13	0	0
14	14	1	1
15	15	0	0
16	16	1	1
17	17	1	1
18	18	1	1
19	19	0	1
20	20	0	0

Figure 6.36

Save SPSS Data file: Example6.5.sav

Quick data check

Before jumping into statistical procedures, let's first just look at the data. A graph that basically tells the whole story for two dichotomous variables measured on the same respondents is a 3-d bar chart.

Graph > Legacy Dialogs > 3-D Bar

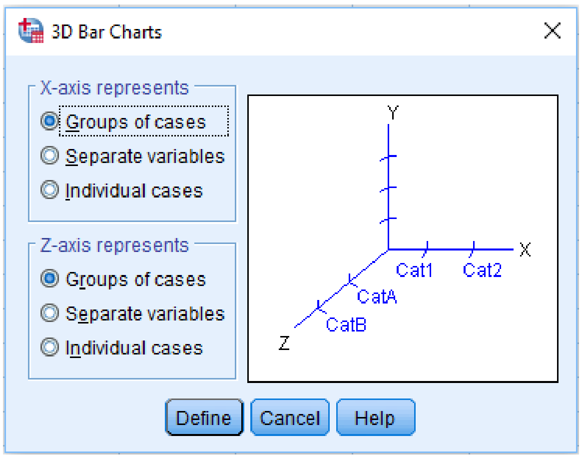


Figure 6.37

Click on Define

Select Number of cases and move product_a (X Category Axis) and product_b (Z Category Axis) into the appropriate boxes.

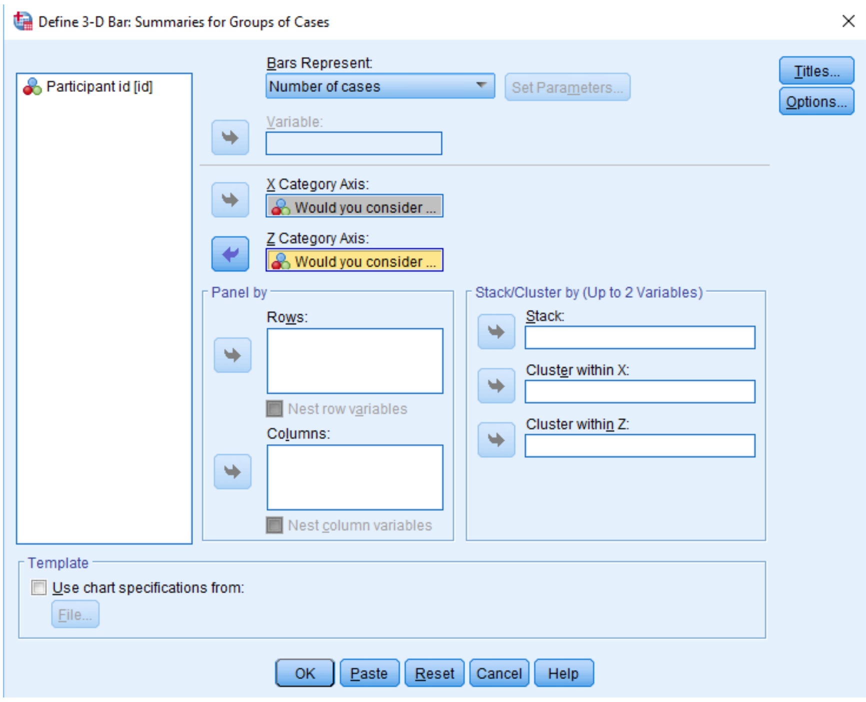


Figure 6.38

Click OK

SPSS output

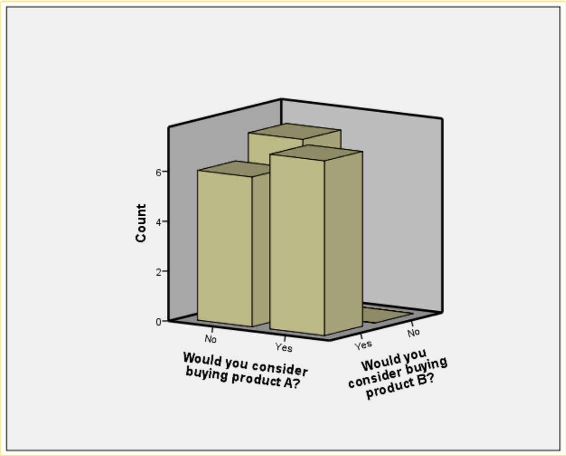


Figure 6.39

Save SPSS Output file: Example6.5.spv

The most important thing we learn from this chart is that both variables are indeed dichotomous.

There could have been some “Don't know/no opinion” answer category but both variables only have “Yes” and “No” answers. There are no system missing values since the bars represent (6 + 7 + 7 + 0 =) 20 valid answers which equals the number of respondents.

Second, product_b is considered by (6 + 7 =) 13 respondents and thus seems more appealing than product_a (considered by 7 respondents).

Third, all of the respondents who consider product_a consider product_b as well but not reversely. This causes the variables to be positively correlated and asks for a close look at the nature of both products.

McNemar test

The results from the McNemar test rely on just one assumption:

> Independent and identically distributed variables (or, less precisely, “independent observations”)

Select Analyze > Nonparametric Tests > Legacy Dialogs > 2 Related Samples

Transfer variables to the Test Pairs box
Choose McNemar test

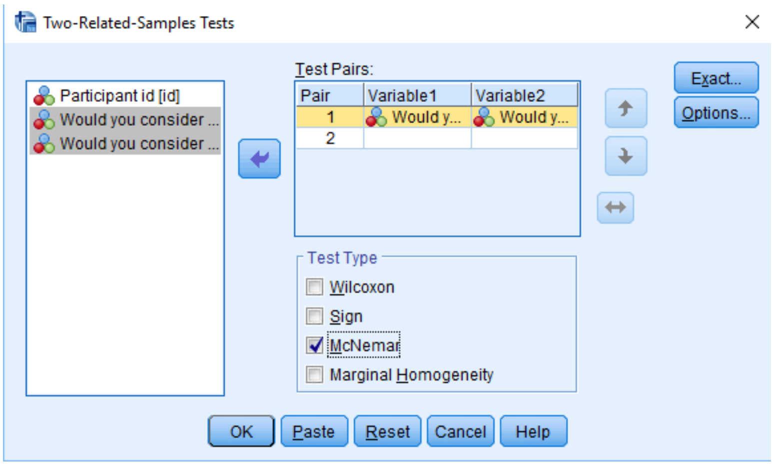


Figure 6.40

Click on Options

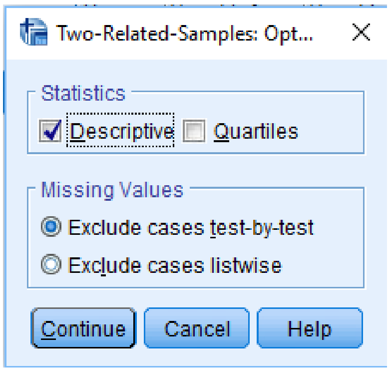


Figure 6.41

Click Continue

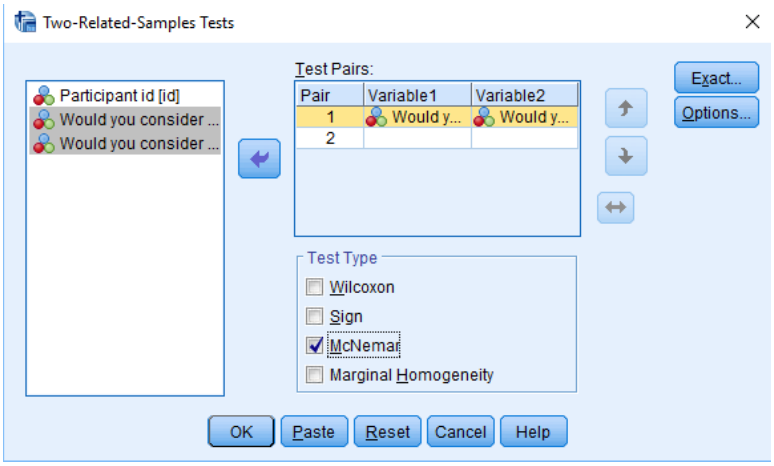


Figure 6.42

Click OK

SPSS McNemar Output

The first table (Descriptive Statistics) confirms that there are no missing values. Note that SPSS reports means rather than proportions. However, if your answer categories are coded 0 (for "absent") and 1 (for "present") the means coincide with the proportions.*

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Would you consider buying product A?	20	.35	.489	0	1
Would you consider buying product B?	20	.65	.489	0	1

Figure 6.43

The proportions are (exactly) 0.35 and 0.65.

The difference is thus - 0.3 where we expected 0 (equal proportions).

Would you consider buying product A? & Would you consider buying product B?

Would you consider buying product A?	Would you consider buying product B? No	Would you consider buying product B? Yes
No	7	6
Yes	0	7

Figure 6.44

The final table (Test Statistics) shows that the 2-tailed p-value is 0.031. If the two proportions are equal in the population, there's only a 3.1% chance of finding the difference we observed in our sample.

Usually, if p-value < 0.05, we reject the null hypothesis.

We therefore conclude that the appeal of both products is not equal.

Test Statistics[a]

	Would you consider buying product A? & Would you consider buying product B?
N	20
Exact Sig. (2-tailed)	.031[b]

a. McNemar Test

b. Binomial distribution used.

Figure 6.45

Note that the p-value is two-sided. It consists of a 0.0155% chance of finding a difference smaller than (or equal to) - 0.3 and another 0.0155% chance of finding a difference larger than (or equal to) 0.3.

Resave SPSS Output file: Example6.5.spv

Sign test for one sample median

A sign test for one median is often used instead of a one sample t-test when the latter's assumptions aren't met by the data. The most common scenario is analyzing a variable which doesn't seem normally distributed with few (say n < 30) observations.

Example 6.6

A car manufacturer had 3 commercials rated on attractiveness by 18 people. They used a percent scale running from 0 (extremely unattractive) through 100 (extremely attractive). A marketer thinks a commercial is good if at least 50% of some target population rate it 80 or higher. Now, the score that divides the 50% lowest from the 50% highest scores is known as the median. In other words, 50% of the population scoring 80 or higher is equivalent to our null hypothesis that

H_0: the population median is at least 79.5 for each commercial

If this is true, then the medians in our sample will be somewhat different due to random sampling fluctuation. However, if we find very different medians in our sample, then our hypothesized 79.5 population median is not credible, and we'll reject our null hypothesis.

SPSS data SPSS data (same data for Examples 6.6 – 6.8)

	id	Gender	Age_group	Edu_level	Family_salary_£	Rating_Family_Car_advert	Rating_Teenager_car_advert	Rating_Eco_Car_advert	Family_car_advert	Teenager_car_advert	Eco_car_advert
1	1	0	3	3	58000	94	31	60	1	0	0
2	2	1	1	3	44500	92	58	67	1	0	0
3	3	0	2	3	67000	100	66	66	1	0	0
4	4	0	1	3	42000	92	49	39	1	0	0
5	5	0	1	3	24000	93	36	100	1	0	1
6	6	1	2	2	44000	49	70	78	0	0	0
7	7	1	1	4	59000	53	50	61	0	0	0
8	8	1	3	4	37000	58	46	83	0	0	1
9	9	0	3	4	63000	95	29	53	1	0	0
10	10	0	2	4	41000	89	75	92	1	0	1
11	11	0	1	3	52000	100	34	47	1	0	0
12	12	1	3	5	63000	84	71	59	1	0	0
13	13	1	2	3	59000	88	53	95	1	0	1
14	14	1	1	4	57000	73	74	63	0	0	0
15	15	1	2	4	52000	78	70	66	0	0	0
16	16	1	3	4	59000	88	76	47	1	0	0
17	17	0	3	4	47000	86	88	31	1	1	0
18	18	0	2	3	49000	90	14	72	1	0	0

Figure 6.46

Save SPSS Data file: Example6.6.sav

Quick Data Check - Histograms

Let's first take a quick look at what our data look like in the first place.

Graphs > Legacy Dialogs > Histogram

Transfer Rating family car advert variable into Variable box

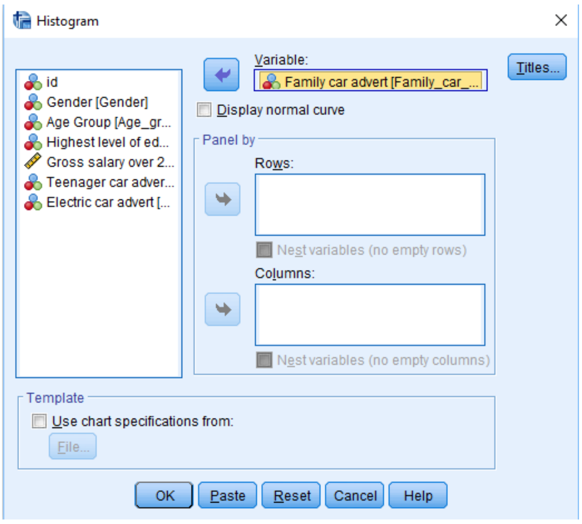


Figure 6.47

Click OK

Repeat for the other two advert ratings.

SPSS output

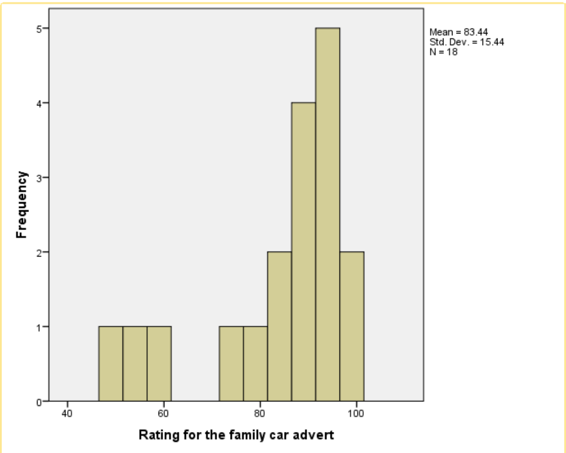


Figure 6.48

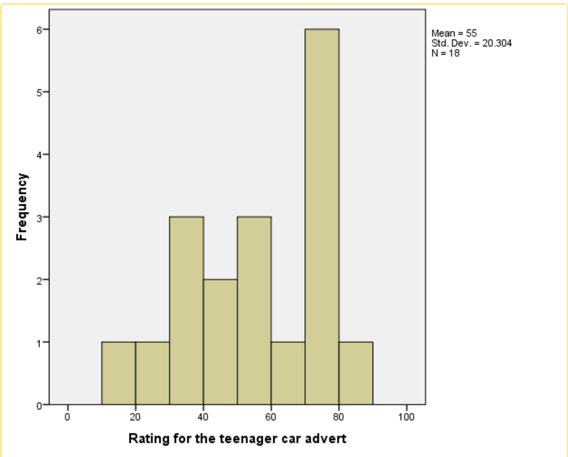


Figure 6.49

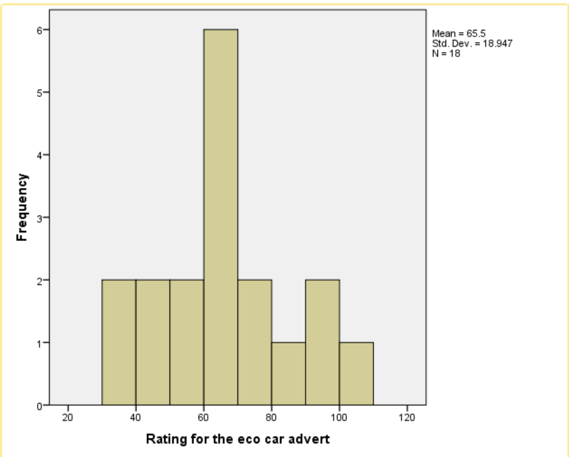


Figure 6.50

First, note that all distributions look plausible. Since n = 18 for each variable, we don't have any missing values. The distributions don't look much like normal distributions. Combined with our small sample sizes, this violates the normality assumption required by t-tests, so we run the non-parametric equivalent "sign test".

Save SPSS Output file: Example6.6.spv

Quick Data Check - Medians

Our histograms included mean scores for our 3 outcome variables but what about their medians? Very oddly, we can't compute medians -which are descriptive statistics- with DESCRIPTIVES. We could use FREQUENCIES but we prefer the table format we get from MEANS as shown below.

Click on Analysis > Descriptives > Frequencies

Transfer the 3 ratings variables into the Variable(s) box

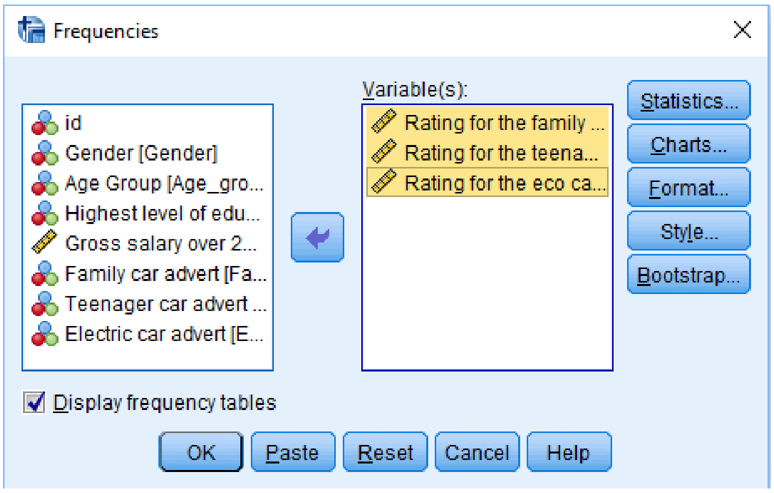


Figure 6.51

Click on Statistics

Choose Mean and Median

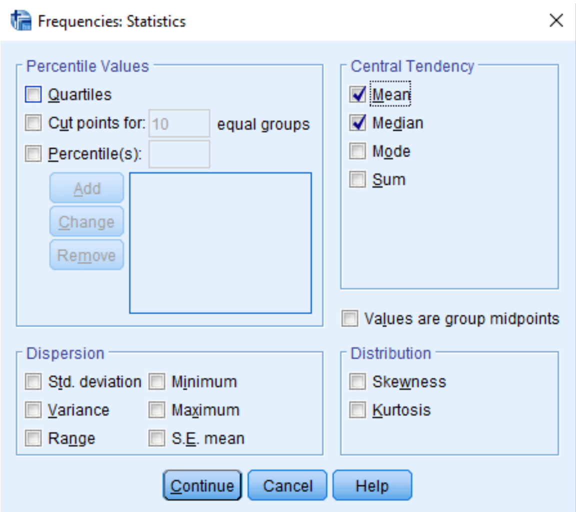


Figure 6.52

Click on Continue

Click on OK

SPSS Output

Statistics

		Rating for the family car advert	Rating for the teenager car advert	Rating for the eco car advert
N	Valid	18	18	18
	Missing	0	0	0
Mean		83.44	55.00	65.50
Median		88.50	55.50	64.50

Figure 6.53

Only our first advertisement ("family car") has a median of 88.50 which is close to 79.5. The other 2 commercials have much lower median values (55.5, 64.5). But are they different enough for rejecting our null hypothesis?

Resave SPSS Output file: Example6.6.spv

SPSS Sign Test - Recoding Data Values
SPSS includes a sign test for two related medians but the sign test for one median is absent.

But remember that our null hypothesis of a 79.5 population median is equivalent to 50% of the population scoring 80 or higher. And SPSS does include a test for a single proportion (a percentage divided by 100) known as the binomial test. We'll therefore just use binomial tests for evaluating if the proportion of respondents rating each commercial 80 or higher is equal to 0.50.

The easy way to go here is to RECODE our data values: values smaller than the hypothesized population median are recoded into a minus (-) sign. Values larger than this median get a plus (+) sign. It's these plus and minus signs that give the sign test its name. Values equal to the median are excluded from analysis so we'll specify them as missing values.

Select Transform > Recode into Different Variables

Transfer the three car rating variables into Numeric Variable -> Output Variable box

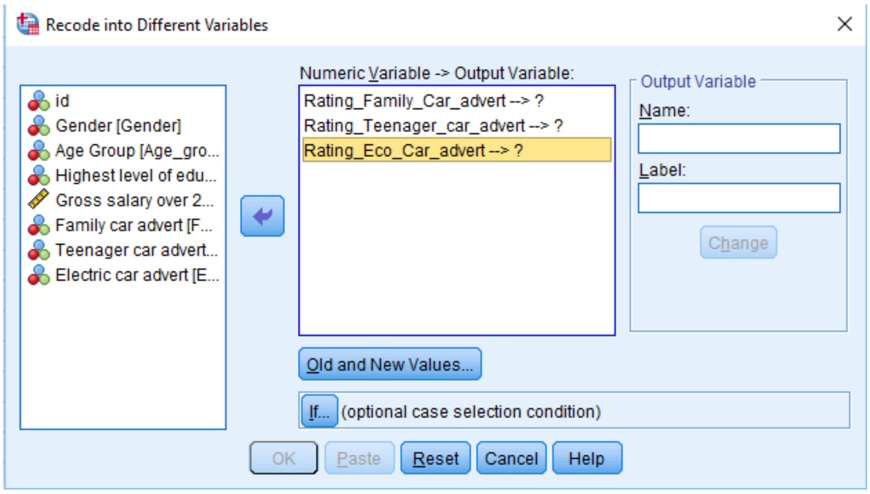


Figure 6.54

Click on the first variable "Rating_Family_Car_advert" and in Name and **L**abel box type "Family" and "Family_Car". Now click on Change.

Repeat for the other two variables.

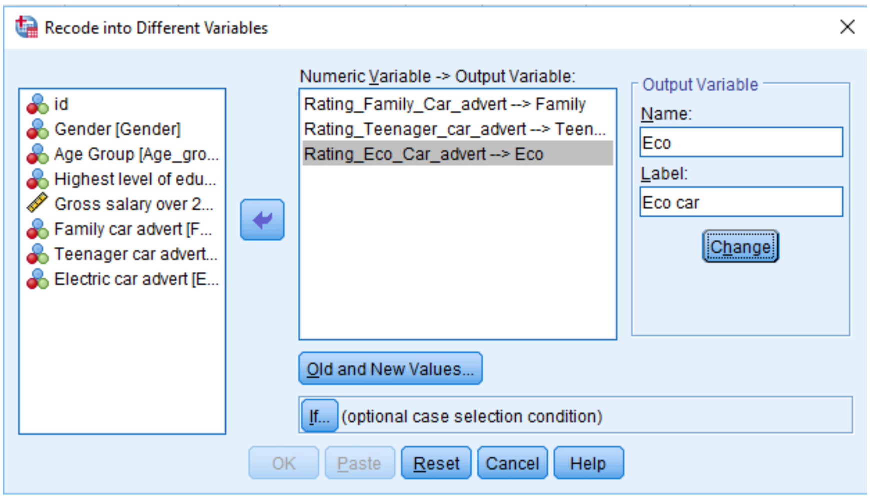


Figure 6.55

Click on “Rating_Family_Car_advert” then click on Old and New Values …

We need to code as follows:

- Lowest thru 79.5 > 0 (represents below median)
- 79.5 thru Highest > 1 (represents above median)

Please note that we should include the possibility that a value of 79.5 exists – remember we do not want to include this possibility in the analysis given we are only interested in values less than or greater than 79.5.

To exclude values = 79.5, then we add an extra recode 79.5 equates to -10000.

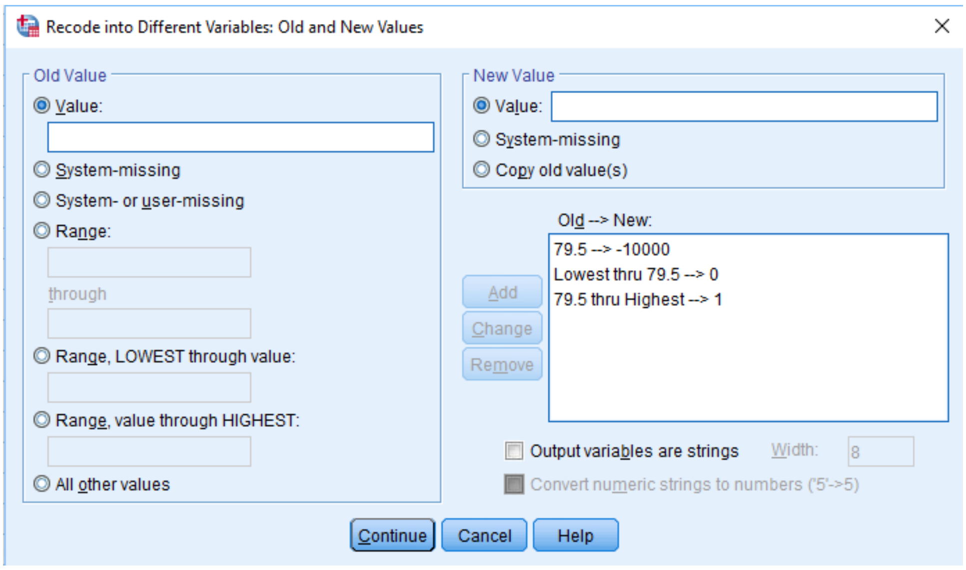


Figure 6.56

Click Continue

Click OK

SPSS output

The new codes will operate for all three variables Family, Teenager, and Eco.

	id	Gender	Age_group	Edu_level	Family_salary_£	Rating_Family_Car_advert	Rating_Teenager_car_advert	Rating_Eco_Car_advert	Family_car_advert	Teenager_car_advert	Eco_car_advert	Family	Teenager	Eco
1	1	0	3	3	58000	94	31	60	1	0	0	1.00	.00	.00
2	2	1	1	3	44500	92	58	67	1	0	0	1.00	.00	.00
3	3	0	2	3	67000	100	66	66	1	0	0	1.00	.00	.00
4	4	0	1	3	42000	92	49	39	1	0	0	1.00	.00	.00
5	5	0	1	3	24000	93	36	100	1	0	1	1.00	.00	1.00
6	6	1	2	2	44000	49	70	78	0	0	0	.00	.00	.00
7	7	1	1	4	59000	53	50	61	0	0	0	.00	.00	.00
8	8	1	3	4	37000	58	46	83	0	0	1	.00	.00	1.00
9	9	0	3	4	63000	95	29	53	1	0	0	1.00	.00	.00
10	10	0	2	4	41000	89	75	92	1	0	1	1.00	.00	1.00
11	11	0	1	3	52000	100	34	47	1	0	0	1.00	.00	.00
12	12	1	3	5	63000	84	71	59	1	0	0	1.00	.00	.00
13	13	1	2	3	59000	88	53	95	1	0	1	1.00	.00	1.00
14	14	1	1	4	57000	73	74	63	0	0	0	.00	.00	.00
15	15	1	2	4	52000	78	70	66	0	0	0	.00	.00	.00
16	16	1	3	4	59000	88	76	47	1	0	0	1.00	.00	.00
17	17	0	3	4	47000	86	88	31	1	1	0	1.00	1.00	.00
18	18	0	2	3	49000	90	14	72	1	0	0	1.00	.00	.00

Figure 6.57

Resave SPSS Data file: Example6.6.sav

SPSS Binomial Test Menu

Minor note: the binomial test is a test for a single proportion, which is a population parameter. So, it's clearly not a nonparametric test. Unfortunately, "nonparametric tests" often refers to both nonparametric and distribution free tests -even though these are completely different things.

Select Analyze > Nonparametric Tests > Legacy Dialogs > Binomial

Transfer the three re-coded variables into the Test Variable List box

Choose Test Proportion = 0.5

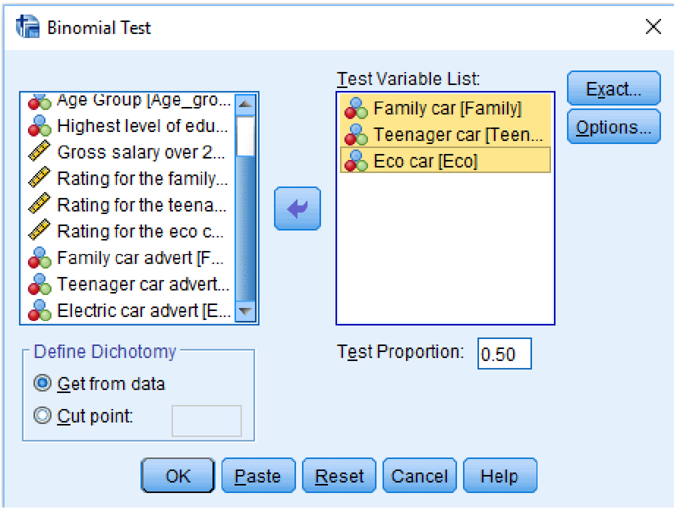


Figure 6.58

Click OK

SPSS Output

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
Family car	Group 1	1.00	13	.72	.50	.096
	Group 2	.00	5	.28		
	Total		18	1.00		
Teenager car	Group 1	.00	17	.94	.50	.000
	Group 2	1.00	1	.06		
	Total		18	1.00		
Eco car	Group 1	.00	14	.78	.50	.031
	Group 2	1.00	4	.22		
	Total		18	1.00		

Binomial Test

Figure 6.59

Resave SPSS Output file: Example6.6.spv

Note: Category 1.00 represents above median (+), and .00 represents below media (-).

For Family car

We saw previously that our first advert ("family car") has a sample median of 89.5. Summary - 5 out of 18 cases score lower than 79.5, and the observed proportion is (5 / 18 =) 0.28 or 28%. The hypothesized test proportion is 0.50; p (denoted as "Exact Significance (2-tailed)") = 0.096: the probability of finding our sample result is roughly 10% if the population proportion really is 50%. We generally reject our null hypothesis if $p < 0.05$, so our binomial test does not refute the hypothesis that our population median is 79.5 given $p = 0.096 > 0.05$.

For Teenager car

We saw previously that our second advert ("teenager car") has a sample median of 55.5. Our p-value of 0.000 means that we've a 0% probability of finding this sample median in a sample of $n = 18$ when the population median is 79.5. Since $p = 0.000 < 0.05$, we reject the null hypothesis: the population median is not 79.5 but -presumably- much lower.

For Eco car

We saw previously that our third advert ("eco car") has a sample median of 64.5. Our p-value of 0.031 means that we've a 3% probability of finding this sample median in a sample of $n = 18$ when the population median is 79.5. Since $p = 0.031 < 0.05$, we reject the null hypothesis: the population median is not 79.5 but -presumably lower.

Sign test for two sample medians

The sign test for two medians evaluates if 2 variables measured on 1 group of cases are likely to have equal population medians. It can be used on either metric variables or ordinal variables. For comparing means rather than medians, the paired samples t-test and Wilcoxon signed-ranks test are better options.

Example 6.7

Re-consider Example 6.6 SPSS data (same data for Examples 6.6 – 6.8)

	id	Gender	Age_group	Edu_level	Family_salary_£	Rating_Family_Car_advert	Rating_Teenager_car_advert	Rating_Eco_Car_advert
1	1	0	3	3	58000	94	31	60
2	2	1	1	3	44500	92	58	67
3	3	0	2	3	67000	100	66	66
4	4	0	1	3	42000	92	49	39
5	5	0	1	3	24000	93	36	100
6	6	1	2	2	44000	49	70	78
7	7	1	1	4	59000	53	50	61
8	8	1	3	4	37000	58	46	83
9	9	0	3	4	63000	95	29	53
10	10	0	2	4	41000	89	75	92
11	11	0	1	3	52000	100	34	47
12	12	1	3	5	63000	84	71	59
13	13	1	2	3	59000	88	53	95
14	14	1	1	4	57000	73	74	63
15	15	1	2	4	52000	78	70	66
16	16	1	3	4	59000	88	76	47
17	17	0	3	4	47000	86	88	31
18	18	0	2	3	49000	90	14	72

Figure 6.60

Save SPSS Data file: Example6.7.sav

Data check

Analyze > Descriptive Statistics > Frequencies

Transfer the rating variables into Variable(s) box

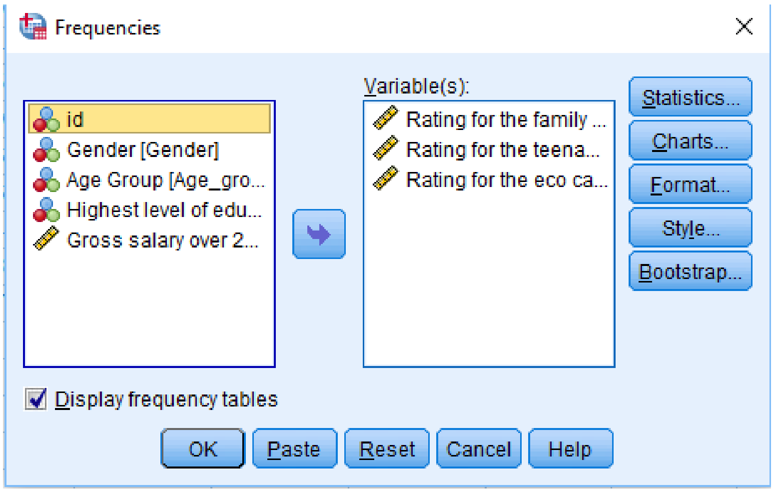


Figure 6.61

Click on Statistics

Choose Mean and Median

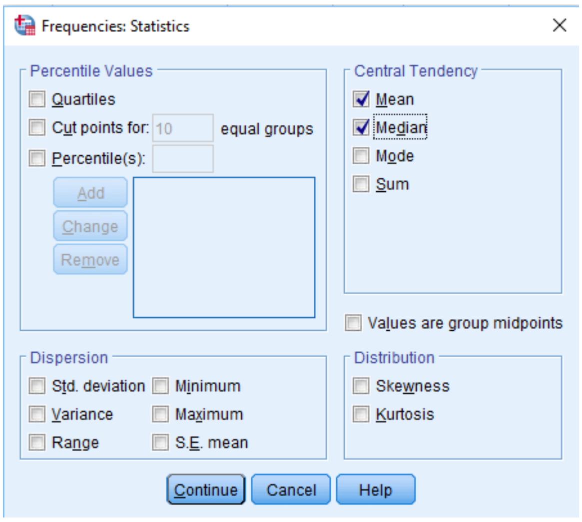


Figure 6.62

Click Continue

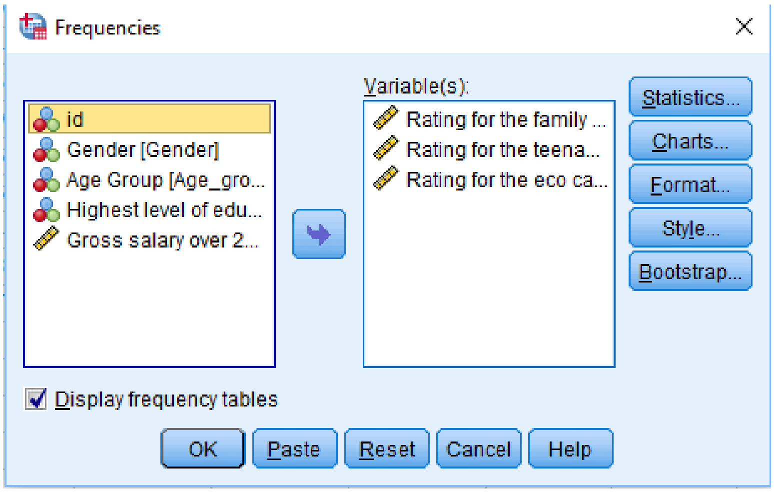


Figure 6.63

Click OK

SPSS Output

Statistics

		Rating for the family car advert	Rating for the teenager car advert	Rating for the eco car advert
N	Valid	18	18	18
	Missing	0	0	0
Mean		83.44	55.00	65.50
Median		88.50	55.50	64.50

Figure 6.64

Save SPSS Output file: Example6.7.spv

The mean and median ratings for the second advert (“Teenager Car”) are very low. We'll therefore exclude this variable from further analysis and restrict our focus to the first and third adverts.

Sign Test - Null Hypothesis

For some reason, our marketing manager is only interested in comparing median ratings, so our null hypothesis is that

the two population medians are equal

for our 2 rating variables.

If our null hypothesis is true, then the plus and minus signs should be roughly distributed 50/50 in our sample. A very different distribution is unlikely under H_0 and therefore argues that the population medians probably weren't equal after all.

Running the Sign Test in SPSS

Analyze > Nonparametric Tests > Legacy Dialogs > 2 Related Samples

Transfer family rating and eco rating into Test Pairs Variable boxes

[We prefer having the best rated variable in the second slot. We'll do so by reversing the variable order]

Choose Sign test

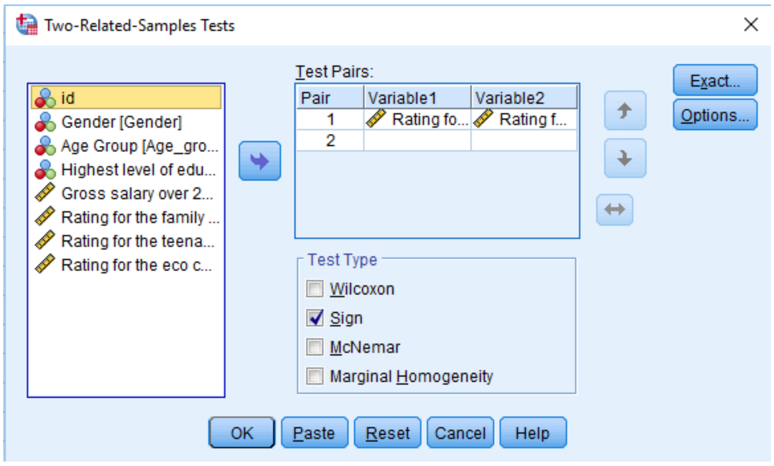


Figure 6.65

Click OK

SPSS output

Frequencies

		N
Rating for the family car advert - Rating for the eco car advert	Negative Differences[a]	6
	Positive Differences[b]	12
	Ties[c]	0
	Total	18

a. Rating for the family car advert < Rating for the eco car advert

b. Rating for the family car advert > Rating for the eco car advert

c. Rating for the family car advert = Rating for the eco car advert

Test Statistics[a]

	Rating for the family car advert - Rating for the eco car advert
Exact Sig. (2-tailed)	.238[b]

a. Sign Test

b. Binomial distribution used.

Figure 6.66

Resave SPSS Output file: Example6.7.spv

We have 18 respondents; our null hypothesis suggests that roughly 9 of them should rate family car advert higher than eco car advert. It turns out that this holds for 12 instead of 9 cases. Can we reasonably expect this difference just by random sampling 18 cases from some large population?

Exact Sig. (2-tailed) refers to our p-value of 0.238. This means there's a 23.8% chance of finding the observed difference if our null hypothesis is true. Our finding doesn't contradict our hypothesis of equal population medians.

In many cases the output will include “Asymp. Sig. (2-tailed)”, an approximate p-value based on the standard normal distribution.* It's not included now because our sample size n <= 25.

Conclusion

“a sign test didn't show any difference between the two medians, exact binomial p-value (2-tailed) = 0.238”

Mann-Whitney test

The Mann-Whitney test is an alternative for the independent samples t test when the assumptions required by the latter aren't met by the data. The most common scenario is testing a non-normally distributed outcome variable in a small sample (say, n < 25).

The Mann-Whitney test is also known as the Wilcoxon test for independent samples -which should not be confused with the Wilcoxon signed-ranks test for related samples.

Example 6.8

Our research question is whether men and women judge our adverts similarly. For each advert separately, our null hypothesis is:

Null hypothesis

H_0: the mean ratings of men and women are equal

Alternative hypothesis

H_1: the mean ratings of men and women are different

SPSS data (same data for Examples 6.6 – 6.8)

	id	Gender	Age_group	Edu_level	Family_salary_£	Rating_Family_Car_advert	Rating_Teenager_car_advert	Rating_Eco_Car_advert
1	1	0	3	3	58000	94	31	60
2	2	1	1	3	44500	92	58	67
3	3	0	2	3	67000	100	66	66
4	4	0	1	3	42000	92	49	39
5	5	0	1	3	24000	93	36	100
6	6	1	2	2	44000	49	70	78
7	7	1	1	4	59000	53	50	61
8	8	1	3	4	37000	58	46	83
9	9	0	3	4	63000	95	29	53
10	10	0	2	4	41000	89	75	92
11	11	0	1	3	52000	100	34	47
12	12	1	3	5	63000	84	71	59
13	13	1	2	3	59000	88	53	95
14	14	1	1	4	57000	73	74	63
15	15	1	2	4	52000	78	70	66
16	16	1	3	4	59000	88	76	47
17	17	0	3	4	47000	86	88	31
18	18	0	2	3	49000	90	14	72

Figure 6.67

Save SPSS Data file: Example6.8.sav

Quick Data Check - Split Histograms

Before running any significance tests, you should look at the data to confirm the data type is appropriate and to see if the variables are approximately normally distributed. Since we're interested in differences between male and female respondents, let's split our histograms by gender.

Select Graphs > Legacy Dialogs > Histogram

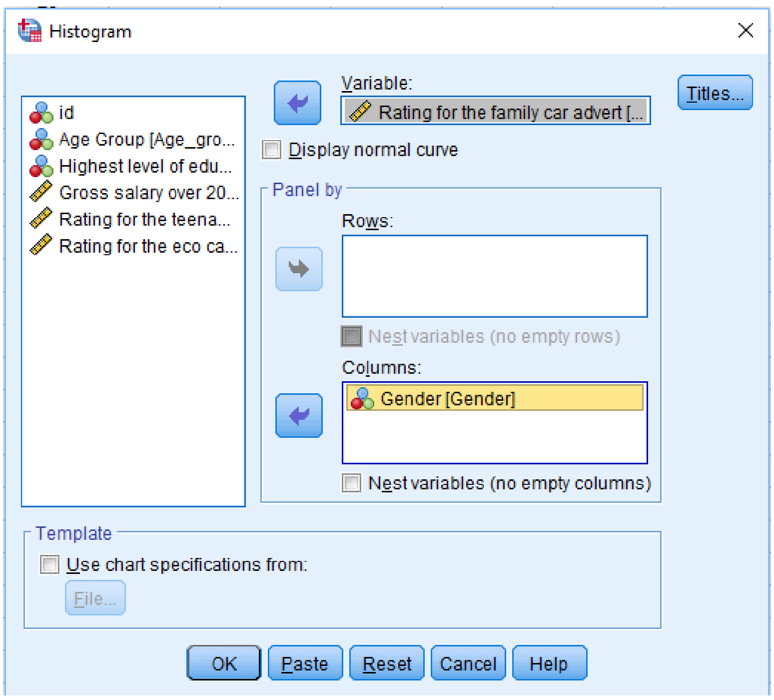


Figure 6.68

Click OK

SPSS output

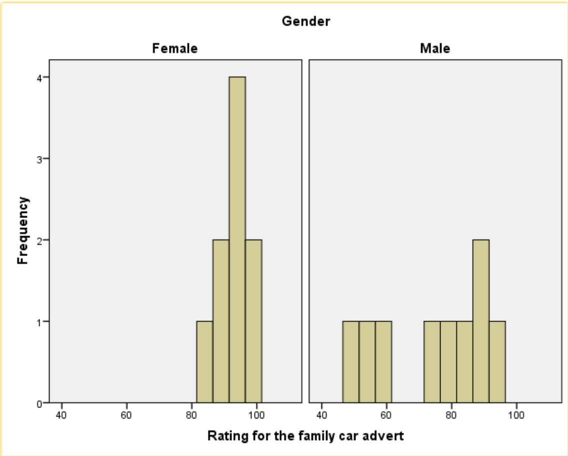


Figure 6.69

Repeat for the other two variables: teenager, eco.

Figure 6.70

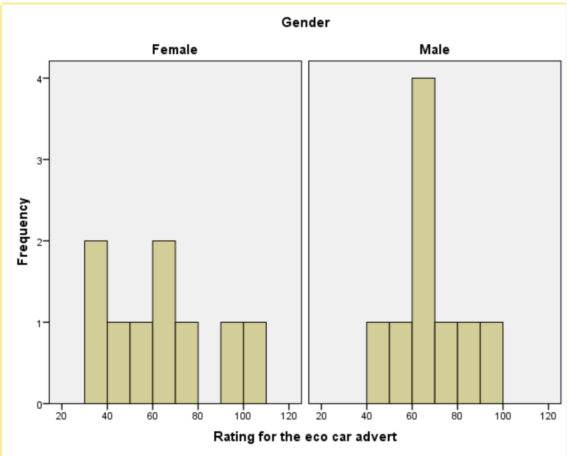


Figure 6.71

Save SPSS Output file: Example6.8.spv

Most importantly, all results look plausible; we don't see any unusual values or patterns. Second, our outcome variables don't seem to be normally distributed and we've a total sample size of only n = 18. This argues against using a t-test for these data.

Finally, by taking a good look at the split histograms, you can already see which adverts are rated more favourably by male versus female respondents. But even if they're rated perfectly similarly by large populations of men and women, we'll still see some differences in small samples. Large sample differences, however, are unlikely if the null hypothesis -equal population means- is really true. We'll now find out if our sample differences are large enough for refuting this hypothesis.

Mann-Whitney Test

Analyze > Nonparametric Tests > Legacy Dialogs > 2 Independent Samples

Transfer the 3 variables into the Test Variable List box
Transfer gender variable into the Grouping Variable box
Click on Define Groups and type “0” for Group 1 and “1” for Group 2 boxes.
Click on Mann-Whitney U

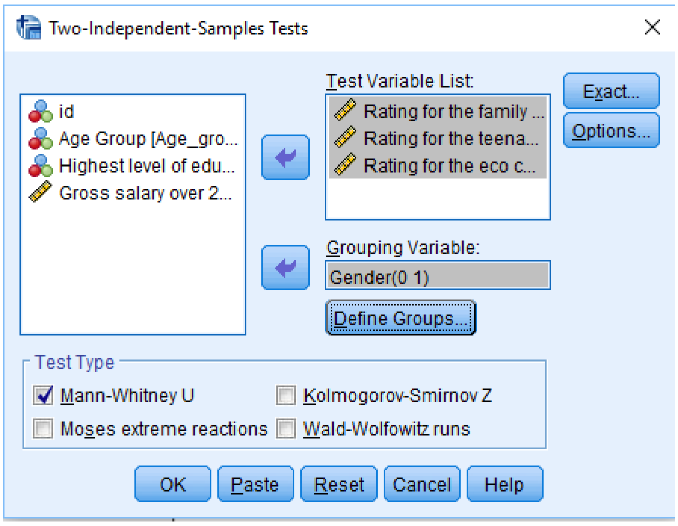


Figure 6.72

Click OK

SPSS Output

The Mann-Whitney test basically replaces all scores with their rank numbers: 1, 2, 3 through 18 for 18 cases. Higher scores get higher rank numbers. If our grouping variable (gender) doesn't affect our ratings, then the mean ranks should be roughly equal for men and women.

Ranks

	Gender	N	Mean Rank	Sum of Ranks
Rating for the family car advert	Female	9	13.39	120.50
	Male	9	5.61	50.50
	Total	18		
Rating for the teenager car advert	Female	9	7.44	67.00
	Male	9	11.56	104.00
	Total	18		
Rating for the eco car advert	Female	9	8.44	76.00
	Male	9	10.56	95.00
	Total	18		

Figure 6.73

Our first advert ("Family car") shows the largest difference in mean ranks between male and female respondents: females seem much more enthusiastic about it. The reverse pattern -but much weaker- is observed for the other two adverts.

Test Statistics[a]

	Rating for the family car advert	Rating for the teenager car advert	Rating for the eco car advert
Mann-Whitney U	5.500	22.000	31.000
Wilcoxon W	50.500	67.000	76.000
Z	-3.095	-1.634	-.840
Asymp. Sig. (2-tailed)	.002	.102	.401
Exact Sig. [2*(1-tailed Sig.)]	.001[b]	.113[b]	.436[b]

a. Grouping Variable: Gender

b. Not corrected for ties.

Figure 6.74

Resave SPSS Output file: Example6.8.spv

From SPSS:

Rating for the family car

Mann-Whitney test statistic U = 5.5
Asymp Sid (2-tailed) p-value = 0.002 < 0.05 [Reject H_0, accept H_1]

Rating for the teenager car

Mann-Whitney test statistic U = 22
Asymp Sid (2-tailed) p-value = 0.102 > 0.05 [Accept H_0, reject H_1]

Rating for the eco car

Mann-Whitney test statistic U = 31
Asymp Sid (2-tailed) p-value = 0.401 > 0.05 [Accept H_0, reject H_1]

Women rated the "Family Car" commercial more favorably than men (p-value = 0.002). The other two commercials didn't show a gender difference (p-values > 0.05).

The p-value of 0.002 indicates a probability of 2 in 1,000: if the populations of men and women rate this advert similarly, then we've a 2 in 1,000 chance of finding the large difference we observe in our sample. Presumably, the populations of men and women don't rate it similarly after all.

Kruskal-Wallis test

The Kruskal-Wallis test is an alternative for a one-way ANOVA if the assumptions of the latter are violated.

Example 6.9

The data in example6.9.sav contains the result of a small experiment regarding a bodybuilding supplement X. These were divided into 3 groups: some didn't take any of X, others took it in the morning, and others took it in the evening. After doing so for a month, their weight gains were measured.

The basic research question is:

> Does the average weight gain depend on the creatine condition to which people were assigned?

That is, we'll test if three means -each calculated on a different group of people- are equal. The most likely test for this scenario is a one-way ANOVA but using it requires some assumptions. Some basic checks will tell us that these assumptions aren't satisfied by our data at hand.

SPSS data

	Group	Weight_gain
1	1	63
2	1	-261
3	1	-153
4	1	-13
5	1	965
6	2	0
7	2	-652
8	2	4724
9	2	-2
10	2	0
11	2	-86
12	3	2239
13	3	171
14	3	40
15	3	1395

Figure 6.75

Save SPSS Data file: Example6.9.sav

Data Check 1 - Histogram

Analyze > Descriptives > Frequencies

Transfer Weight Gain to Variables box

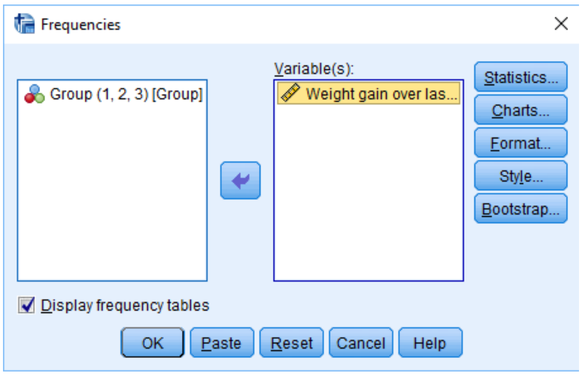


Figure 6.76

Click on Charts
Choose Histogram

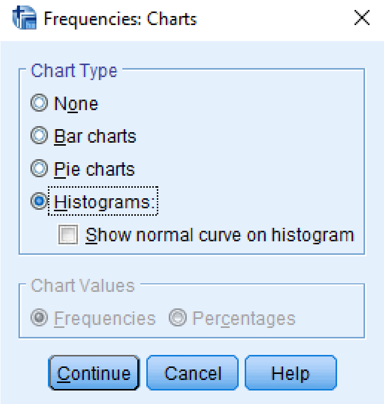


Figure 6.77

Click on Continue

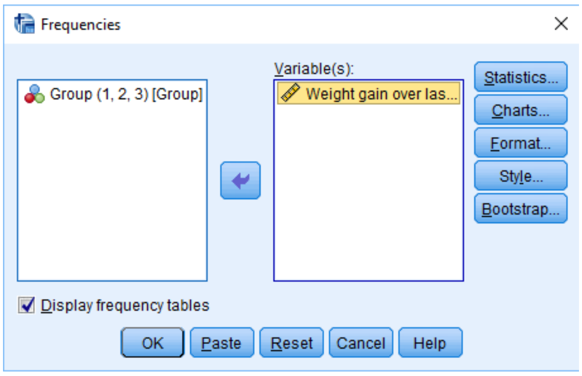


Figure 6.78

Click on OK

SPSS output

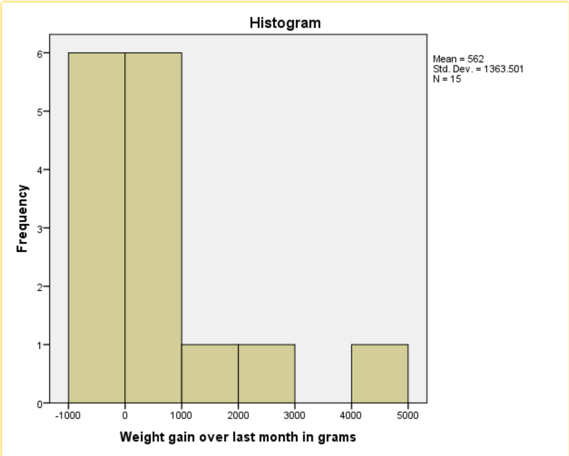


Figure 6.79

First, our histogram looks plausible with all weight gains between -1 and +5 kilos, which are reasonable outcomes over one month. However, our outcome variable is not normally distributed as required for ANOVA. This isn't an issue for larger sample sizes of, say, at least 30 people in each group. However, for our tiny sample at hand, this does pose a real problem.

Save SPSS Output file: Example6.9.spv

Data Check 2 - Descriptives per Group

Right, now after making sure the results for weight gain look credible, let's see if our 3 groups actually have different means. The fastest way to do so is a simple MEANS command as shown below.

Analyze > Compare Means > Means

Transfer Weight gain to the Dependent List box
Transfer Group to the Independent List box

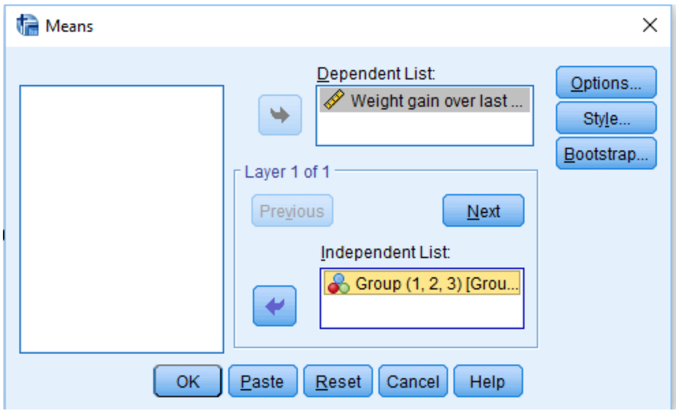


Figure 6.80

Click OK

SPSS Output

Report			
Weight gain over last month in grams			
Group (1, 2, 3)	Mean	N	Std. Deviation
No supplement	120.20	5	488.532
Take supplement in the morning	664.00	6	2005.159
Take supplement in the evening	961.25	4	1047.852
Total	562.00	15	1363.501

Figure 6.81

Resave SPSS Output file: Example6.9.spv

First, note that our evening group (4 participants) gained an average of 961 grams as opposed to 120 grams for no supplement. This suggests that the supplement does make a real difference. But don't overlook the standard deviations for our groups: they are very different, but ANOVA requires them to be equal.* This is a second violation of the ANOVA assumptions.

Run Kruskal-Wallis Test

A test that was designed for precisely this situation is the Kruskal-Wallis test which doesn't require these assumptions. It basically replaces the weight gain scores with their rank numbers and tests whether these are equal over groups.

Analyze > Nonparametric Tests > Legacy Dialogs > K Independent Samples

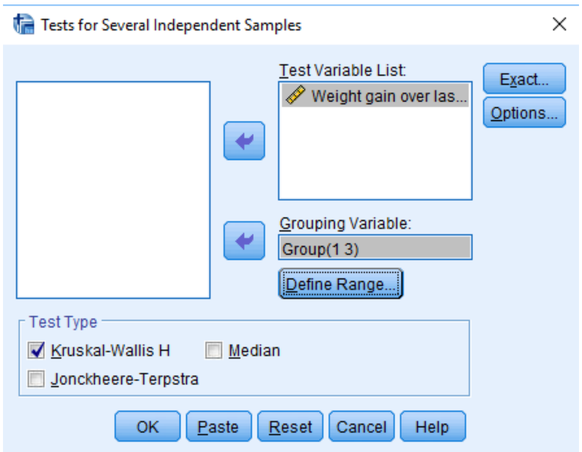


Figure 6.82

Click OK

SPSS Output

Ranks table

Ranks			
	Group (1, 2, 3)	N	Mean Rank
Weight gain over last month in grams	No supplement	5	6.40
	Take supplement in the morning	6	6.83
	Take supplement in the evening	4	11.75
	Total	15	

Figure 6.83

The second table gives the Kruskal-Wallis hypothesis test results.

Test Statistics[a,b]

	Weight gain over last month in grams
Chi-Square	3.868
df	2
Asymp. Sig.	.145

a. Kruskal Wallis Test

b. Grouping Variable: Group (1, 2, 3)

Figure 6.84

Resave SPSS Output file: Example6.9.spv

Our test statistic - incorrectly labelled as "Chi-Square" by SPSS- is known as Kruskal-Wallis H. A larger value indicates larger differences between the groups we're comparing. For our data it is roughly 3.868. We need to know its sampling distribution for evaluating whether this is unusually large.

Asymp. Sig. is the p-value based on our chi-square approximation. The value of 0.145 basically means there's a 14.5% chance of finding our sample results if supplement doesn't have any effect in the population at large. So, if the supplement does nothing whatsoever, we have a fair (14.5%) chance of finding such minor weight gain differences just because of random sampling. If p-value > 0.05, we usually conclude that our differences are not statistically significant.

Conclusion

The official way for reporting our test results includes our chi-square value, df and p-value as in this study did not demonstrate any effect from the supplement, $\chi^2(2) = 3.87$, p-value = 0.15.

Wilcoxon test

For comparing two metric variables measured on one group of cases, our first choice is the paired-samples t-test. This requires the difference scores to be normally distributed in our population. If this assumption is not met, we can use Wilcoxon S-R test instead.

It can also be used on ordinal variables -although ties may be a real issue for Likert items. Don't abbreviate "Wilcoxon S-R test" to simply "Wilcoxon test" like SPSS does: there's a second "Wilcoxon test" which is also known as the Mann-Whitney test for two independent samples.

Example 6.10

A car manufacturer had 18 respondents rate 3 different adverts for one of their cars. They first want to know which advert is rated best by all respondents.

H_0: advert rating the same

SPSS data (example6.10.sav)

	Respondent_ID	Gender	Age_group	Education_level	Salary_£	Advert_1	Advert_2	Advert_3
1	1	0	3	3	25000	94	31	60
2	2	1	1	3	38500	92	58	67
3	3	0	2	3	68500	100	66	66
4	4	0	1	3	42000	92	49	39
5	5	0	1	3	24000	93	36	100
6	6	1	2	2	44000	49	70	78
7	7	1	1	4	59000	53	50	61
8	8	1	3	4	37000	58	46	83
9	9	0	3	4	63000	95	29	53
10	10	0	2	4	41000	89	75	92
11	11	0	1	3	52000	100	34	47
12	12	1	3	5	63000	84	71	59
13	13	1	2	3	59000	88	53	95
14	14	1	1	4	57000	73	74	63
15	15	1	2	4	52000	78	70	66
16	16	1	3	4	59000	88	76	47
17	17	0	3	4	47000	86	88	31
18	18	0	2	3	49000	90	14	72

Figure 6.85

Save SPSS Data file: Example6.10.sav

Quick Data Check

Our current focus is limited to the 3 rating variables, advert_1, advert_2, and advert_3.

Analyze > Descriptive Statistics > Frequencies

Transfer the variable variables into the Variable(s) box

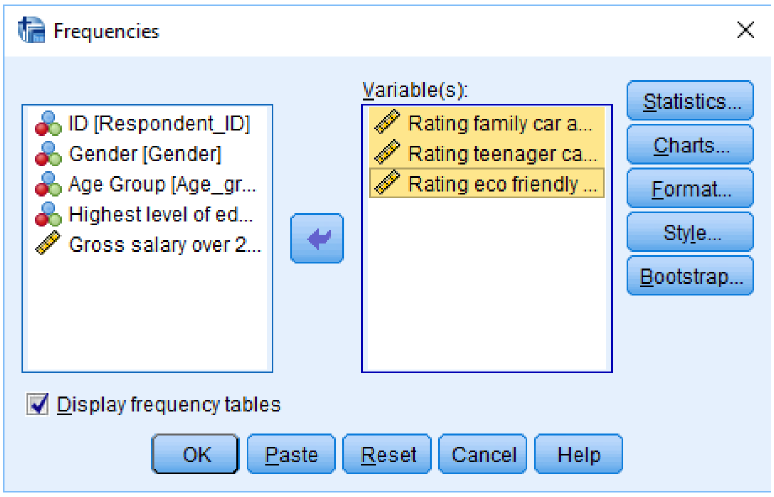


Figure 6.86

Click on Charts

Choose Histograms

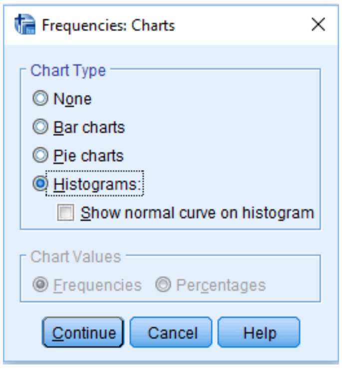


Figure 6.87

Click Continue

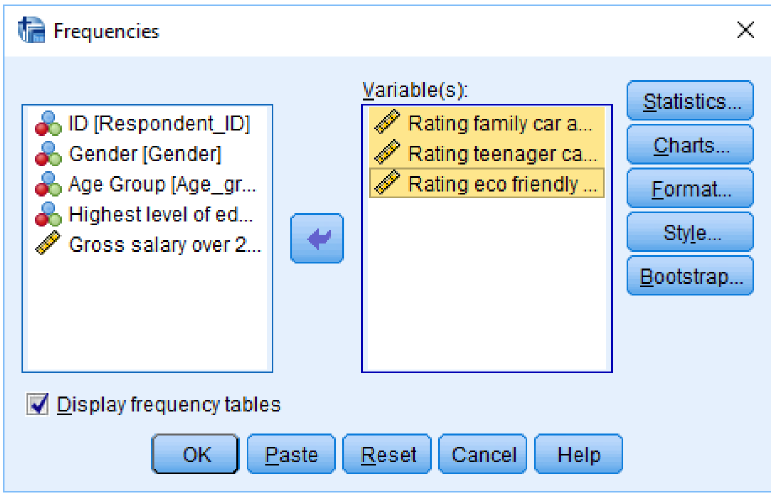


Figure 6.88

Click OK

SPSS output

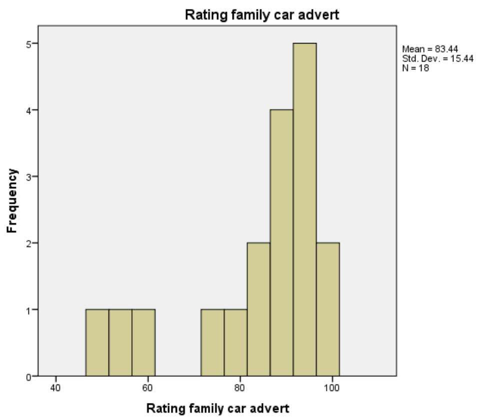


Figure 6.89

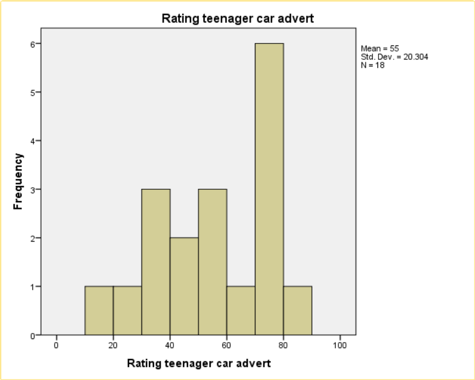


Figure 6.90

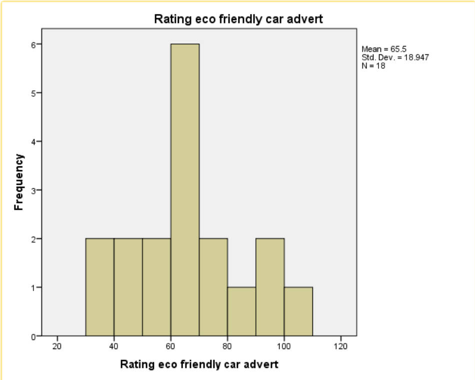


Figure 6.91

Save SPSS Output file: Example6.10.spv

The 3 histograms show that all data values are present and that the sampling distributions do not look normally distributed. From the SPSS output, we observe that advert_2 has a very low average rating of only 55. Therefpe, we decide to test if advert_1 and advert_3 have equal mean ratings.

Difference Scores

Let's now compute and inspect the difference scores between advert_1 and advert_3.

Transform > Compute Variable

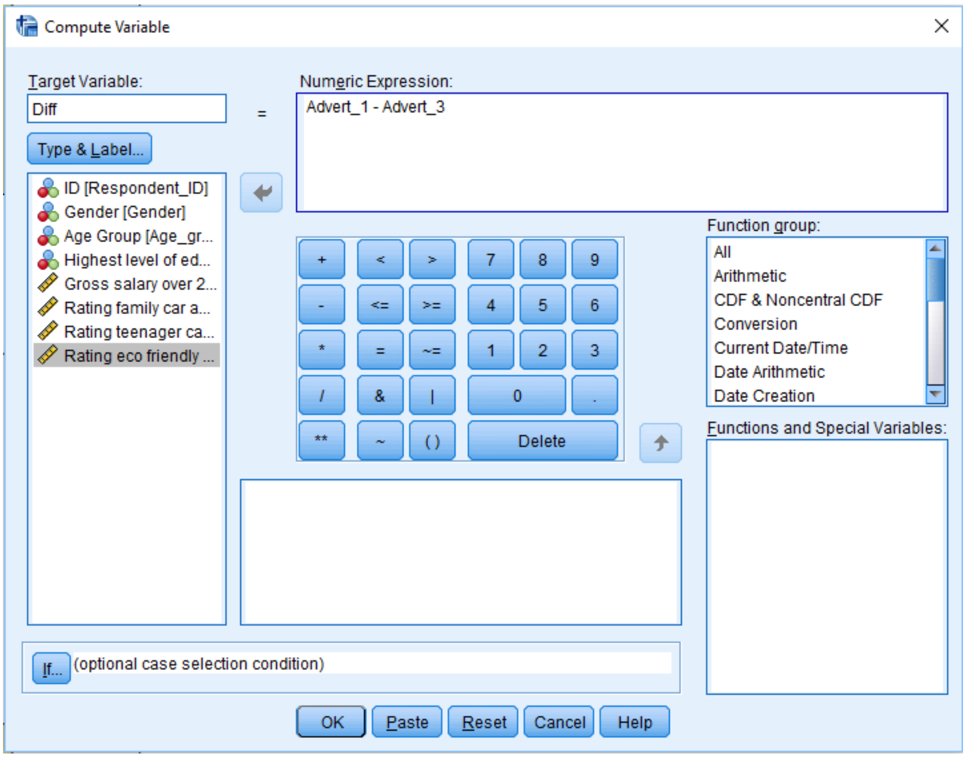


Figure 6.92

Click OK

SPSS output

	Respondent_ID	Gender	Age_group	Education_level	Salary_£	Advert_1	Advert_2	Advert_3	Diff
1	1	0	3	3	25000	94	31	60	34.00
2	2	1	1	3	38500	92	58	67	25.00
3	3	0	2	3	68500	100	66	66	34.00
4	4	0	1	3	42000	92	49	39	53.00
5	5	0	1	3	24000	93	36	100	-7.00
6	6	1	2	2	44000	49	70	78	-29.00
7	7	1	1	4	59000	53	50	61	-8.00
8	8	1	3	4	37000	58	46	83	-25.00
9	9	0	3	4	63000	95	29	53	42.00
10	10	0	2	4	41000	89	75	92	-3.00
11	11	0	1	3	52000	100	34	47	53.00
12	12	1	3	5	63000	84	71	59	25.00
13	13	1	2	3	59000	88	53	95	-7.00
14	14	1	1	4	57000	73	74	63	10.00
15	15	1	2	4	52000	78	70	66	12.00
16	16	1	3	4	59000	88	76	47	41.00
17	17	0	3	4	47000	86	88	31	55.00
18	18	0	2	3	49000	90	14	72	18.00

Figure 6.93

Reave SPSS Data file: Example6.10.sav

Now create a histogram for this new variable (Difference) and ask SPSS to plot the normal curve onto the histogram.

Graphs > Legacy Dialogs > Histograms

Click on Display normal curve

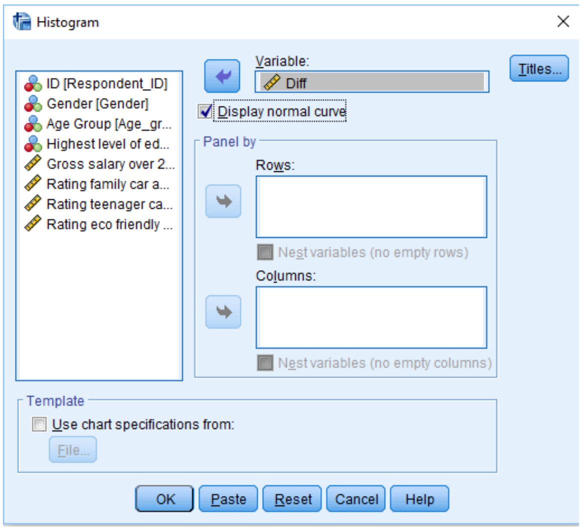


Figure 6.94

Click OK

SPSS output

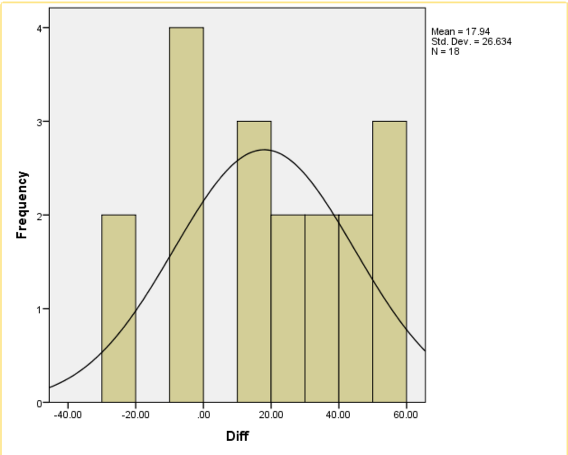


Figure 6.95

Resave SPSS Output file: Example6.10.spv

We could analyse the difference using a paired samples t-test. This requires the difference scores to be normally distributed in our population, but our sample suggests otherwise. This isn't a problem for larger samples sizes (say, n > 25) but we've only 18 respondents in our data.

Fortunately, Wilcoxon S-R test was developed for precisely this scenario: not meeting the assumptions of a paired-samples t-test.

Null hypothesis

H_0: the population distributions for advert_1 and advert_3 are identical

If this is true, then these distributions will be slightly different in a small sample like our data at hand. However, if our sample shows very different distributions, then our hypothesis of equal population distributions will no longer be tenable.

Wilcoxon S-R test in SPSS

Analyze > Nonparametric Tests > Legacy Dialogs > 2 Related Samples

2 Related Samples refers to comparing 2 variables measured on the same respondents. This is similar to “paired samples” or “within-subjects” effects in repeated measures ANOVA.

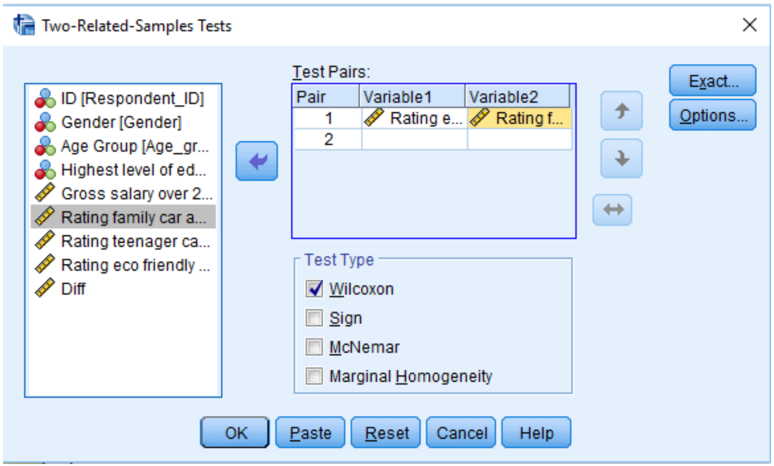


Figure 6.96

Click OK

SPSS Output

The first table compares the summary statistics for the negative and positive ranks.

Ranks

		N	Mean Rank	Sum of Ranks
Rating family car advert - Rating eco friendly car advert	Negative Ranks	6[a]	5.00	30.00
	Positive Ranks	12[b]	11.75	141.00
	Ties	0[c]		
	Total	18		

a. Rating family car advert < Rating eco friendly car advert

b. Rating family car advert > Rating eco friendly car advert

c. Rating family car advert = Rating eco friendly car advert

Figure 6.97

If advert_1 and advert_3 have similar population distributions, then the signs (plus and minus) should be distributed roughly evenly over ranks. This implies that the sum of positive ranks should be close to the sum of negative ranks. This number (141 in our example) is our test statistic and known as Wilcoxon W+. Our table shows a very different pattern: the sum of positive ranks (indicating that the “Family car” was rated better) is way larger than the sum of negative ranks. Can we still believe our 2 commercials are rated similarly?

The second table gives you the Wilcoxon S-R test results

Test Statistics[a]

	Rating family car advert - Rating eco friendly car advert
Z	-2.419[b]
Asymp. Sig. (2-tailed)	.016

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.

Figure 6.98

Resave SPSS Output file: Example6.10.spv

Oddly, our ”Test Statistics“ table includes everything except for our actual test statistic, W+.

Asymp. Sig. (2-tailed) p-value = 0.016. This approximate p-value is based on the standard normal distribution (hence the “Z” right on top of it).

If required you could request an exact value by clicking on the Exact menu and choosing Exact.

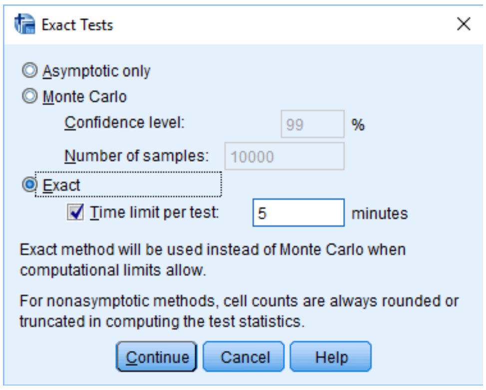


Figure 6.99

If we run this solution, then SPSS output would be:

Test Statistics[a]

	Rating family car advert - Rating eco friendly car advert
Z	-2.419[b]
Asymp. Sig. (2-tailed)	.016
Exact Sig. (2-tailed)	.013
Exact Sig. (1-tailed)	.007
Point Probability	.001

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.

Figure 6.100

Resave SPSS Output file: Example6.10.spv

From Figure 6.100, the Asymp sig (2 sided) p-value = 0.016 and the exact 2-tailed p-value = 0.013. Apparently, the normal approximation is accurate.

Conclusion

A Wilcoxon Signed-Ranks test indicated that the “Family car” advert (mean rank = 11.75) was rated more favourably than the “Eco car” advert (mean rank = 5.0), Z = - 2.419, p = 0.016. Note. If sample sizes are small, then the z approximation may be unnecessary and inaccurate, and the exact p-value is to be preferred.

Friedman test

For testing if 3 or more variables have identical population means, our first option is a repeated measures ANOVA. This requires our data to meet some assumptions - like normally distributed variables. If such assumptions are not met, then our second option is the Friedman test: a nonparametric alternative for a repeated-measures ANOVA. Strictly, the Friedman test can be used on metric or ordinal variables, but ties may be an issue in the latter case.

Example 6.11

The data contain 18 respondents who rated 3 commercials for cars on a percent (0% through 100% attractive) scale.

	Respondent_ID	Gender	Age_group	Education_level	Salary_£	Advert_1	Advert_2	Advert_3
1	1	0	3	3	25000	94	31	60
2	2	1	1	3	38500	92	58	67
3	3	0	2	3	68500	100	66	66
4	4	0	1	3	42000	92	49	39
5	5	0	1	3	24000	93	36	100
6	6	1	2	2	44000	49	70	78
7	7	1	1	4	59000	53	50	61
8	8	1	3	4	37000	58	46	83
9	9	0	3	4	63000	95	29	53
10	10	0	2	4	41000	89	75	92
11	11	0	1	3	52000	100	34	47
12	12	1	3	5	63000	84	71	59
13	13	1	2	3	59000	88	53	95
14	14	1	1	4	57000	73	74	63
15	15	1	2	4	52000	78	70	66
16	16	1	3	4	59000	88	76	47
17	17	0	3	4	47000	86	88	31
18	18	0	2	3	49000	90	14	72

Figure 6.101

Save SPSS Data file: Example6.11.sav

We'd like to know which commercial performs best in the population. So, we'll first see if the mean ratings in our sample are different. If so, the next question is if they're different enough to conclude that the same holds for our population at large. That is, our null hypothesis is that:

the population distributions of our 3 rating variables are identical

Quick Data Check

Inspecting the histograms of our rating variables will give us a lot of insight into our data with minimal effort.

Analyze > Descriptives Statistics > Frequencies

Transfer variables to the Variable(s) box

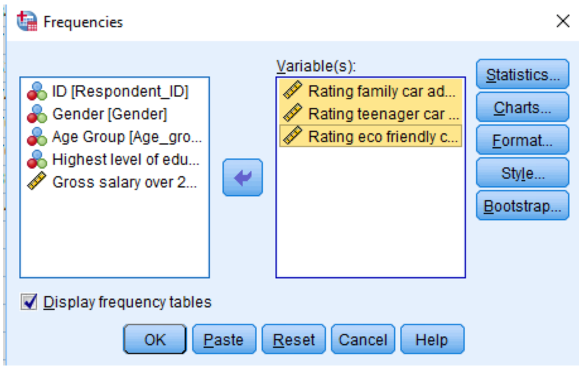


Figure 6.102

Click on Charts
Choose Histograms
Click on Show normal curve on histogram

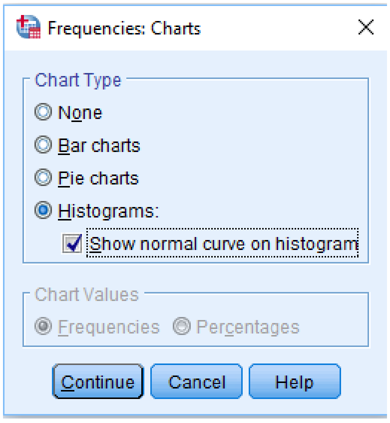
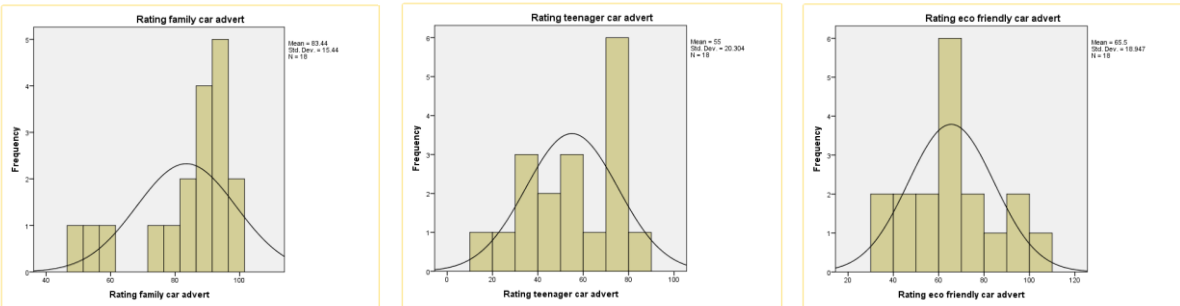


Figure 6.103

Click Continue

Click OK

SPSS Output



Figures 6.104 a - c

Save SPSS Output file: Example6.11.spv

Most importantly, our data look plausible: we don't see any outrageous values or patterns. Note that the mean ratings are different: 83.44, 55 and 65.5. Every histogram is based on all 18 cases so there's no missing values to worry about.

Now, by superimposing normal curves over our histograms, we do see that our variables are not quite normally distributed as required for repeated measures ANOVA. This isn't a serious problem for larger sample sizes (say, n > 25 or so) but we've only 18 cases now. We'll therefore play it safe and use a Friedman test instead.

Running a Friedman Test in SPSS

Analyze > Nonparametric > Legacy Dialogs > K related Samples

Transfer variables into the Test Variables box
Click on Friedman

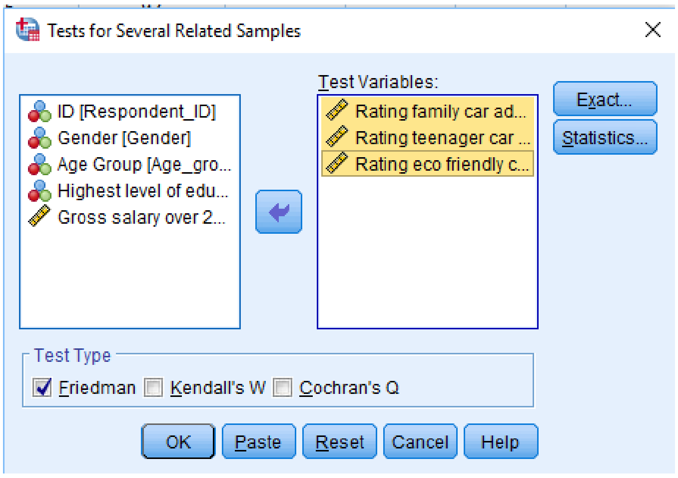


Figure 6.105

Click OK

Click on Exact

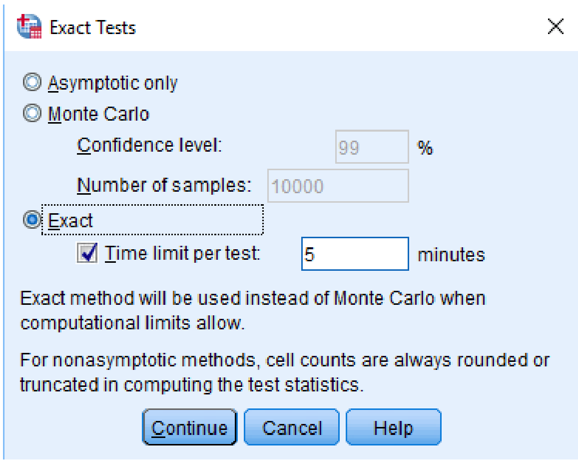


Figure 6.106

Click Continue

Click OK

SPSS Output

Ranks

	Mean Rank
Rating family car advert	2.50
Rating teenager car advert	1.53
Rating eco friendly car advert	1.97

Figure 6.107

Test Statistics[a]

N	18
Chi-Square	8.648
df	2
Asymp. Sig.	.013
Exact Sig.	.012
Point Probability	.001

a. Friedman Test

Figure 6.108

Resave SPSS Output file: Example6.11.spv

First note that the mean ranks differ quite a lot in favour of the first (“Family Car”) advert. Unsurprisingly, the mean ranks have the same order as the means we saw in our histogram.

1. Chi-Square (more correctly referred to as Friedman’s Q) is our test statistic. It basically summarizes how differently our adverts were rated.
2. df are the degrees of freedom associated with our test statistic. It's equal to the number of variables we compare - 1. In our example, 3 variables - 1 = 2 degrees of freedom.
3. Asymp. Sig. is an approximate p-value. Since p-value = 0.013 < 0.05, we reject the null hypothesis of equal population distributions.
4. Exact Sig. is the exact p-value = 0.012. If available, we prefer it over the asymptotic p-value, especially for smaller sample sizes.

Conclusion

We could write something like:

> “a Friedman test indicated that our commercials were rated differently, $\chi2(2)$ = 8.648, p-value = 0.012“

Chapter 7 Regression and correlation analysis

Example 7.1

A company wants to know how job performance relates to IQ, motivation and social support. They collect data on 60 employees, resulting in Example7.1.sav. We'll try to predict job performance from all other variables by means of a multiple regression analysis. Therefore, **job performance = function (IQ, motivation, social support)**.

Data view

	name	perf	iq	mot	soc
1	Henry	85	109	89	73
2	Riley	84	106	84	80
3	Alexis	87	125	59	67
4	Evelyn	69	84	60	58
5	Blake	69	89	60	67
6	Dominic	81	109	62	75
7	Jose	71	121	67	55
8	Tristan	76	102	44	73
9	Kayden	77	111	68	60
10	Makayla	76	106	63	54
11	Ella	90	107	93	75
12	Piper	74	97	52	58
13	Jonathan	74	133	60	50
14	Joshua	65	96	52	74
15	Brooklyn	66	97	65	81
16	Connor	73	116	62	45
17	Sadie	80	108	74	92
18	Zoe	96	102	84	84
19	Cameron	77	94	78	79
20	Jason	73	98	71	68

Figure 7.1

	name	perf	iq	mot	soc
1	Henry	85	109	89	73
2	Riley	84	106	84	80
3	Alexis	87	125	59	67
4	Evelyn	69	84	60	58
5	Blake	69	89	60	67
6	Dominic	81	109	62	75
7	Jose	71	121	67	55
8	Tristan	76	102	44	73
9	Kayden	77	111	68	60
10	Makayla	76	106	63	54
11	Ella	90	107	93	75
12	Piper	74	97	52	58
13	Jonathan	74	133	60	50
14	Joshua	65	96	52	74
15	Brooklyn	66	97	65	81
16	Connor	73	116	62	45
17	Sadie	80	108	74	92
18	Zoe	96	102	84	84
19	Cameron	77	94	78	79
20	Jason	73	98	71	68

Figure 7.2

	name	perf	iq	mot	soc
41	Hannah	85	101	87	65
42	Aubrey	75	94	54	60
43	Eva	81	106	72	55
44	Nora	68	102	32	69
45	Bella	81	98	72	69
46	Jaxson	80	112	72	78
47	Chase	78	87	74	93
48	Caleb	62	73	68	67
49	Madelyn	81	94	67	59
50	London	76	117	66	68
51	Hudson	77	112	58	57
52	Annabelle	74	113	57	76
53	Taylor	69	94	65	53
54	Hunter	68	119	48	44
55	Stella	85	111	91	59
56	Ava	79	104	50	73
57	Samuel	74	99	77	83
58	Angel	81	104	78	83
59	Anna	84	108	58	64
60	Alyssa	92	130	58	75

Figure 7.3

Variable view

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	name	String	9	0	Employee's first name	None	None	8	Left	Nominal	Input
2	perf	Numeric	3	0	Outcome of job performance test	None	None	8	Right	Scale	Input
3	iq	Numeric	3	0	Outcome of IQ test	None	None	8	Right	Scale	Input
4	mot	Numeric	3	0	Outcome of job motivation test	None	None	8	Right	Scale	Input
5	soc	Numeric	3	0	Outcome of social support test	None	None	8	Right	Scale	Input

Figure 7.4

Save SPSS Data file: Example7.1.sav

Quick Data Check

We usually start our analysis with a solid data inspection. Since that's already been done for the data at hand, we'll limit it to a quick check of relevant histograms and correlations.

Histograms

Select Analysis > Descriptive Statistics > Frequencies

Transfer the 4 independent variables into the Variable(s) box

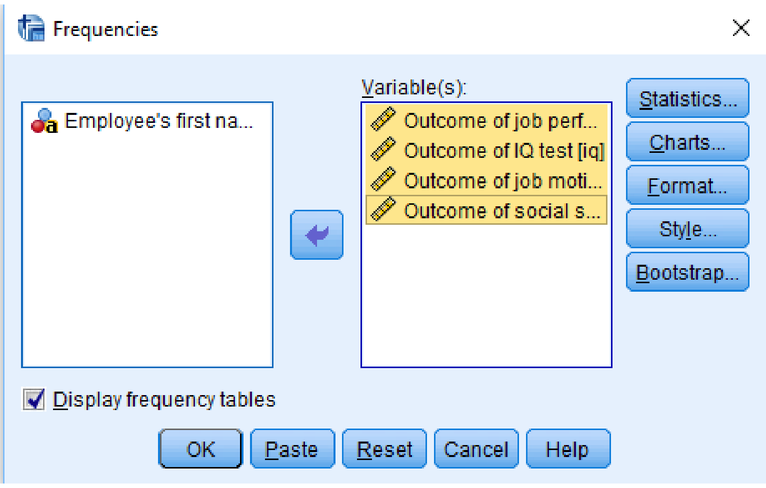


Figure 7.5

Click on Charts
Choose Histograms
Show normal curve on histogram

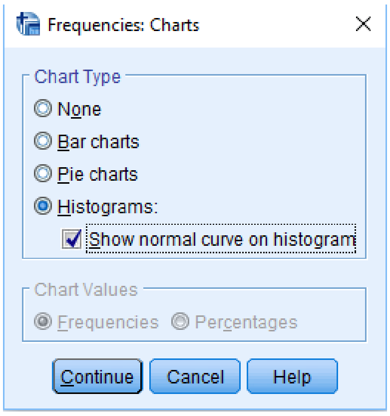


Figure 7.6

Click on Continue

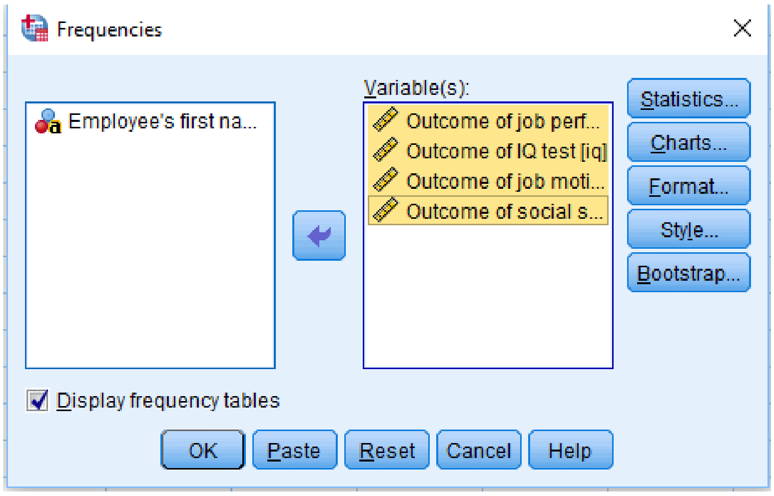


Figure 7.7

Click OK

SPSS output

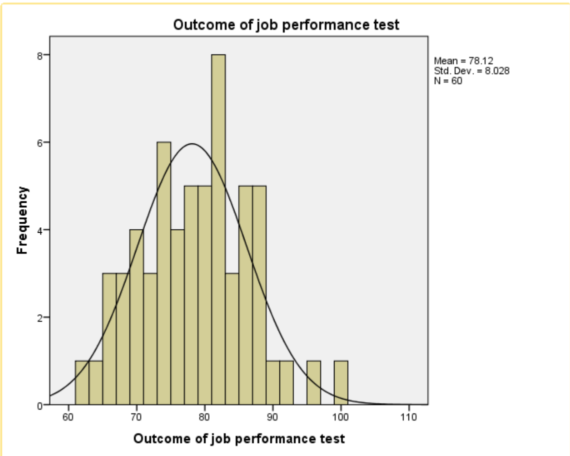


Figure 7.8

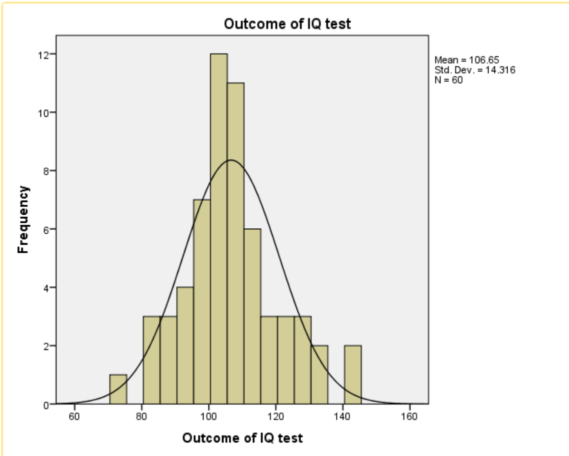


Figure 7.9

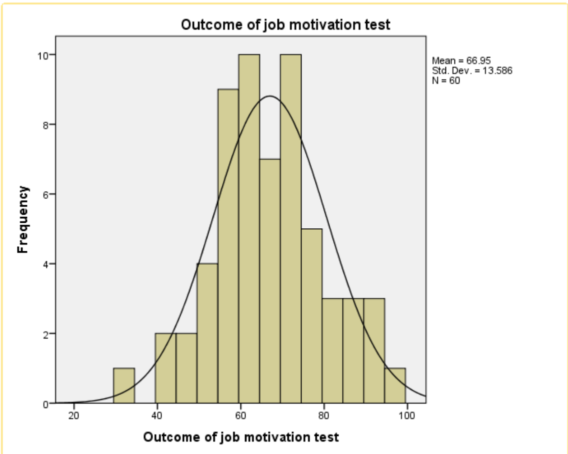


Figure 7.10

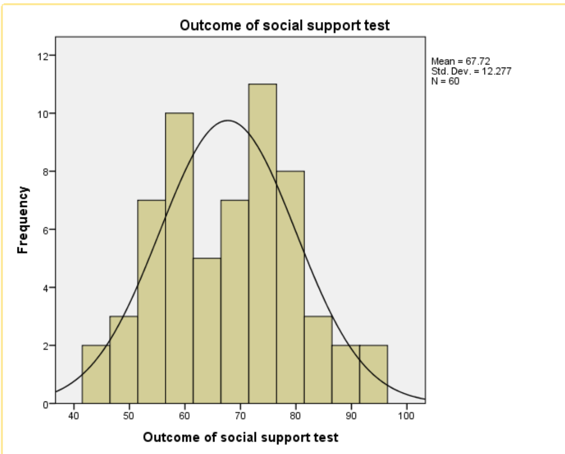


Figure 7.11

Save SPSS Output file: Example7.1.spv

Note that each histogram is based on 60 observations, which corresponds to the number of cases in our data. This means that we don't have any system missing values. Second, note that all histograms look plausible; none of them have weird shapes or extremely high or low values.

Correlations

Next, we'll check whether the correlations among our regression variables make any sense.

Select Analyze > Correlate > Bivariate

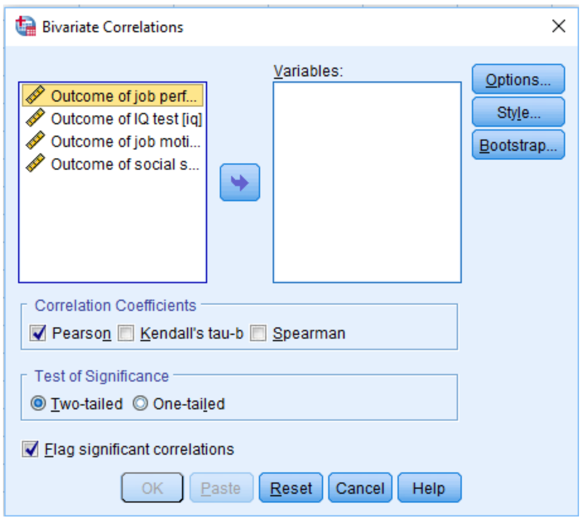


Figure 7.12

Transfer 4 independent variables into the variables box
Given the data is interval/ratio/scale level data then choose Correlation Coefficients: Pearson.

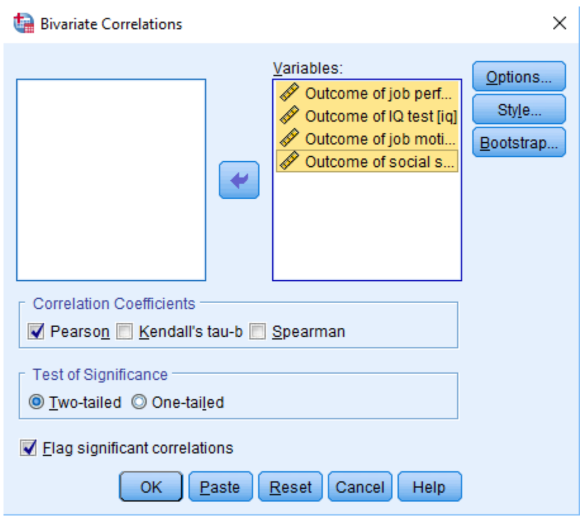


Figure 7.13

Click OK

SPSS output

Correlations

		Outcome of job performance test	Outcome of IQ test	Outcome of job motivation test	Outcome of social support test
Outcome of job performance test	Pearson Correlation	1	.474**	.635**	.397**
	Sig. (2-tailed)		.000	.000	.002
	N	60	60	60	60
Outcome of IQ test	Pearson Correlation	.474**	1	.047	-.092
	Sig. (2-tailed)	.000		.722	.485
	N	60	60	60	60
Outcome of job motivation test	Pearson Correlation	.635**	.047	1	.363**
	Sig. (2-tailed)	.000	.722		.004
	N	60	60	60	60
Outcome of social support test	Pearson Correlation	.397**	-.092	.363**	1
	Sig. (2-tailed)	.002	.485	.004	
	N	60	60	60	60

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

Figure 7.14

Resave SPSS Output file: Example7.1.spv

Most importantly, the correlations are plausible; job performance correlates positively and substantively with all other variables. This makes sense because each variable reflects as positive quality that's likely to contribute to better job performance.

Fit linear regression model

Model to fit: **job performance = function (IQ, motivation, social support)**

Keep in mind that regression does not prove any causal relations from our predictors on job performance. A basic rule of thumb is that we need at least 15 independent observations for each predictor in our model. With three predictors, we need at least (3 x 15 =) 45 respondents. The 60 respondents we have in our data are sufficient for our model.

Select Analyze > Regression > Linear

Transfer job performance to the Dependent box
Transfer the 3 independent variables to the Independent(s) box

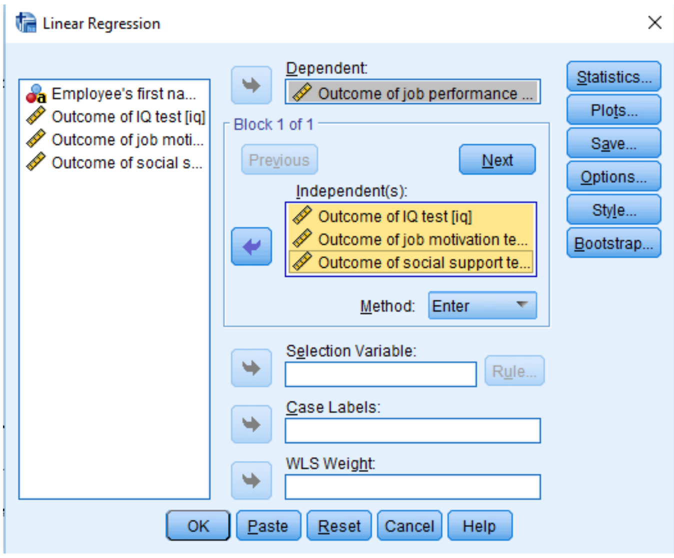


Figure 7.15

Click on Statistics

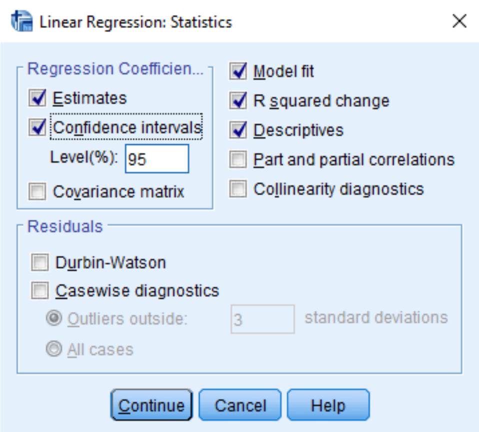


Figure 7.16

Click Continue

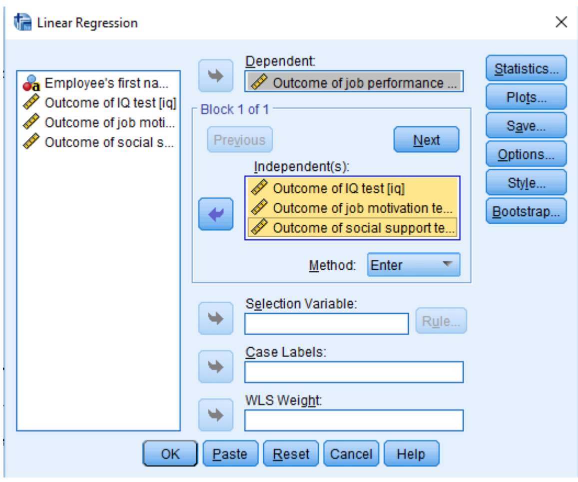


Figure 7.17

Click OK

SPSS output

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	.809[a]	.654	.636	4.844	.654	35.356	3	56	.000

a. Predictors: (Constant), Outcome of social support test, Outcome of IQ test, Outcome of job motivation test

Figure 7.18

R denotes the correlation between predicted and observed job performance. In our case, R = 0.809. Since this is a very high correlation, our model predicts job performance rather precisely. R square is simply the square of R. It indicates the proportion of variance in job performance that can be "explained" by our three predictors. R square = 0.654.

Because regression maximizes R square for our sample (Adjusted $R^2 = 0.636$), it will be somewhat lower for the entire population, a phenomenon known as shrinkage. The adjusted R square estimates the population R square for our model and thus gives a more realistic indication of its predictive power.

The high adjusted R squared tells us that our model does a great job in predicting job performance. On top of that, our b coefficients are all statistically significant and make perfect intuitive sense. We should add, however, that this tutorial illustrates a problem free analysis on problem free data.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2488.395	3	829.465	35.356	.000[b]
	Residual	1313.788	56	23.461		
	Total	3802.183	59			

a. Dependent Variable: Outcome of job performance test

b. Predictors: (Constant), Outcome of social support test, Outcome of IQ test, Outcome of job motivation test

Figure 7.19

The ANOVA table provides a global test to see if the model predictors are a significant contributor to the value of job performance. From SPSS: F test statistic = 35.356, p-value = 0.000 < 0.05, reject H_0 and accept H_1. The model predictors (IQ, motivation, social support) are a significant contributor to the value of job performance.

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)	18.131	6.346		2.857	.006	5.419	30.844
	Outcome of IQ test	.265	.044	.472	5.965	.000	.176	.354
	Outcome of job motivation test	.308	.050	.522	6.163	.000	.208	.408
	Outcome of social support test	.164	.056	.251	2.953	.005	.053	.275

a. Dependent Variable: Outcome of job performance test

Figure 7.20

Resave SPSS Output file: Example7.1.spv

From SPSS, the linear regression model with 3 predictor variables is

Job performance = 18.1 + (0.27 x intelligence) + (0.31 x motivation) +(0.16 x social support)

The b coefficients tell us how many units job performance increases for a single unit increase in each predictor. Therefore, a 1-point increase on the IQ test corresponds to 0.27 points increase on the job performance test. Importantly, note that all b coefficients are positive numbers; higher IQ is associated with higher job performance and so on. B coefficients having the "wrong direction" often indicate a problem with the analysis known as multicollinearity.

The column "Sig." holds the significance levels for our predictors. As a rule of thumb, we say that a b coefficient is statistically significant if its p-value is smaller than 0.05. All of our b coefficients are statistically significant. The beta coefficients allow us to compare the relative strengths of our predictors. These are roughly 2 to 2 to 1 for IQ, motivation and social support.

When applying regression analysis to more difficult data, you may encounter complications such as multicollinearity and heteroscedasticity. These are beyond the scope of this basic regression example.