Basic Statistics for Business

A Step-by-Step Introduction

Using JASP and Excel

Version: 1.0

September 2018

STATMIND
Management Research & Development

Contents

0 Introduction to Business Statistics with Excel and JASP 4

0.1 Introduction to Business Statistics 4

0.2 JASP Introduction 5

Why JASP? 5

0.3 On This Manual 6

1 The Language of Statistics 10

1.1 Getting Data into JASP 10

1.2 Data Skill Challenge 13

2 Working with Numbers and Graphs 18

2.1 Descriptive Information 18

2.2 Graphs 19

2.2.1 Bar Chart 19

2.2.2 A Bigger Data Set 21

2.3 Data Skill Challenges 25

3 Central Tendency and Variability 30

3.1 Looking at the "Distribution" of Your Data 30

3.2 Data Skill Challenge 1 to Chapter 3 37

4 Probability Distributions 38

4.1 Normal Distributions and Z-scores 38

4.2 Data Skill Challenge 3 43

5 Sampling Distributions 45

5.1 How to Draw Samples? 45

5.2 Data Skill Challenge 45

6 Estimation and Confidence Intervals 48

6.1 Finding Confidence Intervals 48

6.2 Data Skill Challenges 50

7/8 Hypothesis Testing & Z-tests; One-Sample T-tests 52

8.1 The One-sample Z-test 52

8.2 The One-Sample T-test 55

8.3 Data Skill Challenge 56

9 Paired and Independent Samples T-test 60

9.1 Introduction 60

9.2 Paired Samples T-test 61

9.3 Independent Samples T-test 62

9.4 Advanced: Using Regression Analysis for T-test 64

9.5 Data skill challenge 67

10 Analysis of Variance 70

10.1 Introduction 70

10.2 ANOVA 71

10.3 Extra: ANOVA via Regression Analysis 73

10.4 Data Skill Challenge 74

11 Chi-Squared Tests of Fit 77

11.1 Introduction 77

11.2 The Distribution of One Nominal Variable 77

11.3 Testing Relationships Between Two Nominal Variables 79

11.4 Additional Statistics 81

11.5 Data Skill Challenge 82

12 Correlation and Regression 86

12.1 Correlation 86

12.2 Simple Linear Regression 88

12.3 Data Skill Challenge 89

Concluding Remarks 92

References 93

Assignment 94

0 Introduction to Business Statistics with Excel and JASP

Files needed:
JASP-0.8.6-Setup.exe
chapter0.xlsx [Excel file]
chapter0.csv [comma-separated-values text file]

0.1 Introduction to Business Statistics

Modern organizations make use of lots of data. While some have a strong belief in the information contained in data sets, others tend to shy away from numbers and rely on intuition. In the era of Big Data, the divide between these two groups is fading.

In a 2012 article in the Financial Times, Steve Jones referred to Big Data as the fourth factor of production, next to the traditional three factors (land, labor and capital), and in doing so hinted at the end of this divide:

"As the prevalence of Big Data grows, executives are becomingly increasingly wedded to numerical insight. But the beauty of Big Data is that it allows both intuitive and analytical thinkers to excel. More entrepreneurially minded, creative leaders can find unexpected patterns among disparate data sources (which might appeal to their intuitive nature) and ultimately use the information to alter the course of the business."

Steve Jones, Financial Times, December 2012 [1]

Regardless of the extent to which you believe in the power of data, data is a fact of life to managers. In order to make better decisions and communicate clearly, it is essential for managers to be able to extract information from data. Even though extracting information from data is often left to specialists, a basic understanding of statistics is a requirement for managers at all levels.

This module is about basic business statistics. The adjective "basic" does not imply that it's easy! Mastering the basics is a prerequisite for proceeding into more advanced statistics. But here, we will stick to the basics.

The field of statistics can be broken down into two parts, broadly speaking.

The easier part is *descriptive statistics*. Here, the challenge is in summarizing a large amount of data. For example, in your studies you get grades for the various modules that make up the curriculum of your studies. After finishing all modules, you have grades (say, scores between 0 and 100) that you can briefly summarize using an average score; the minimum and the maximum of your scores; the number of scores above the passing level of 60; and so on.

In *inferential* statistics, we assume that the data at hand are sampled from some bigger population. For example, in a survey we can ask a sample of consumers to rate our product. Apart from describing the

[1] See https://www.ft.com/content/5086d700-504a-11e2-9b66-00144feab49a, last accessed on 14 July 2018

mean rating of the sample, our interest is in using the sample results to make a statement about the population of all consumers.

We can go one step further, and raise the question how likely it is that consumers prefer our product (defined by, for example, giving a score of at least 4 on a 5-point scale) given our sample outcome. We are then "testing hypotheses" based on samples, and that's where statistics gets tough!

Our aim is to provide you with an intuitive understanding, just enough to deal with the vast majority of real-life challenges that managers encounter.

In (business) research, our main interest is normally in studying relationships. As a marketing manager, you are interested in the relationship between units sold on the one hand, and decreasing prices or higher marketing budgets on the other. Or, as an HRM manager, you are interested in the impact of training on employee productivity. We will introduce statistics on relationships between two or more "variables", using a technique that is, in a sense, the mother of many more advanced statistical techniques.

Statistics cannot be effectively studied without a tool. The tool that we will use is **JASP**.

0.2 JASP Introduction

JASP provides an environment within which many basic and advanced statistical techniques have been implemented.

Why JASP?

In your daily work you probably use spreadsheet programs like Excel. Although Excel can be used for simple data sets and basic statistics, the program is not ideal for statistical tasks.

In this module, you will make use of both Excel for data entry and data manipulation, and both Excel and **JASP** for carrying out basic statistical tasks.

The advantages of **JASP** are that it is "just enough" and focuses on only the most frequently used techniques; it is very user-friendly, and easy-to-learn; and it is free of cost.

Click here for an introduction to **JASP**. From the introduction, you can go to the download page. Alternatively, you can use the set-up file included in this module.

Introducing JASP 0.9

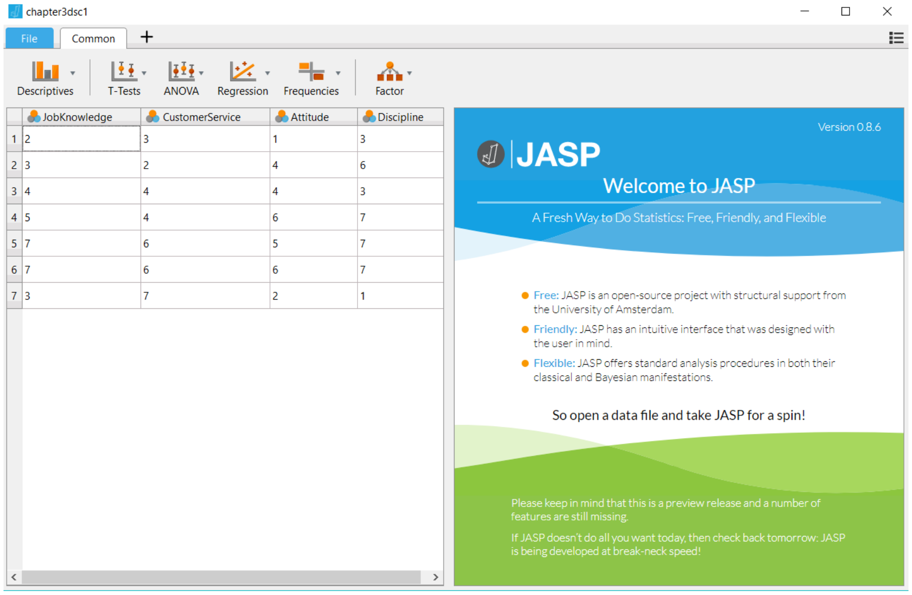
JUNE 20 - 2018

Figure 0.1: Introducing and Downloading JASP

The latest version of **JASP** is 0.9. However, make sure to download version 0.8.6, as, at the time of writing, the latest version contained some bugs. And version 0.8.6 works perfectly for all examples in this manual.

0.3 On This Manual

In this manual we will introduce **JASP** following the recommended textbook by Landers[2]. All examples of Landers will be replicated using **JASP**, and in the process you will become familiar with the program.



	JobKnowledge	CustomerService	Attitude	Discipline
1	2	3	1	3
2	3	2	4	6
3	4	4	4	3
4	5	4	6	7
5	7	6	5	7
6	7	6	6	7
7	3	7	2	1

Version 0.8.6

JASP

Welcome to JASP

A Fresh Way to Do Statistics: Free, Friendly, and Flexible

- Free: JASP is an open-source project with structural support from the University of Amsterdam.
- Friendly: JASP has an intuitive interface that was designed with the user in mind.
- Flexible: JASP offers standard analysis procedures in both their classical and Bayesian manifestations.

So open a data file and take JASP for a spin!

Please keep in mind that this is a preview release and a number of features are still missing.

If JASP doesn't do all you want today, then check back tomorrow: JASP is being developed at break-neck speed!

Figure 0.2: The lay-out of JASP

[2] Landers, R.N. (2013). *A Step-by-Step Introduction to Statistics for Business*. Sage Publications

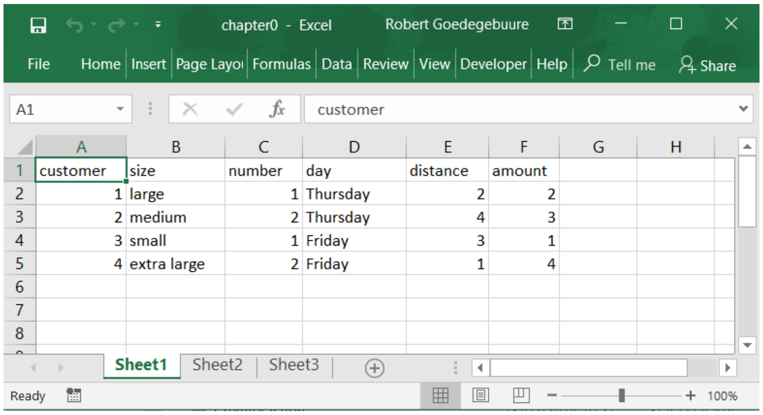
The standard menu consists of just two tabs: **Files**, and **Common**. You can add tabs, but for this module, the standard menu suffices.

Under the **Files** tab, you can open recent files; open new files using the browse option; or use sample files provided by **JASP**.

Data files come in many formats. Software packages may have their own formats, and are able to read several formats including data in Excel.

JASP only reads text files, in *comma separated values* (CSV for short) or *tab delimited* formats. CSV-files are commonly used. CSV-files are plain text files, where the data are separated by commas. These files typically have the extension **csv**. Excel-spreadsheets can be saved in CSV-format.

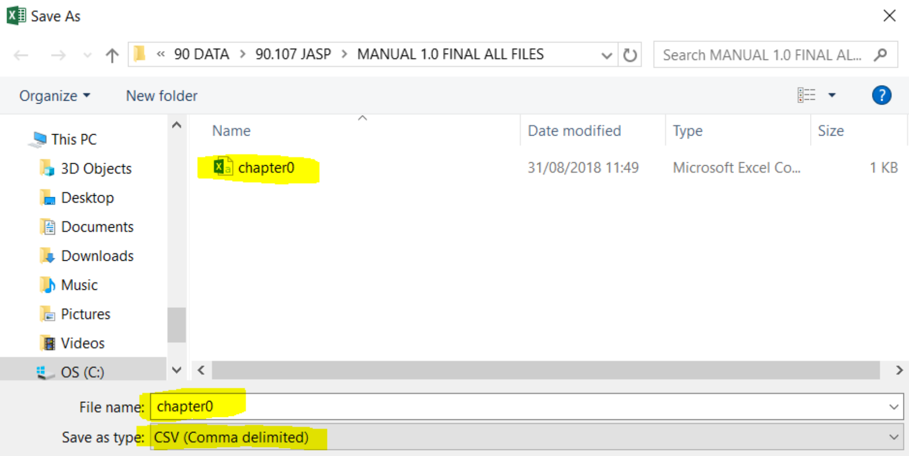
Let's look at a simple example. Suppose you have a small data set on pizza deliveries, stored in Excel. In the first row, you store the names of the variables (or columns). In the next rows, you key in the data of four deliveries. The data indicate that customer 3, who lives three miles away ordered one small pizza on Friday, for 1 €.



	A	B	C	D	E	F	G	H
1	customer	size	number	day	distance	amount		
2	1	large	1	Thursday	2	2		
3	2	medium	2	Thursday	4	3		
4	3	small	1	Friday	3	1		
5	4	extra large	2	Friday	1	4		
6								
7								
8								

Figure 0.3: Sample Data in Excel

Excel files cannot be directly by **JASP**, and first have to be stored in CSV-format. This is easy enough. Just open the Excel file; go to **<File><Save As>**, browse to the folder where you want to store the data, and then use the option CSV, as in the figure below.



File name: chapter0

Save as type: CSV (Comma delimited)

Figure 0.4: Save Data in CSV-Format

Some dos and don’ts when keying in your data in Excel:

- When keying in your data in Excel **DO** use the first row, and no more than one row, to name the variables (or columns);
- For variable names **DO** use short words or codes;
- In variable names, **DO NOT** include spaces, like in “**Number of Pizzas**”. Use **Number** or **NoP** or whatever instead;
- When keying in data in Excel, **DO** make use of numerical codes. Even in the example below, avoid *string variables*, for **size** and **day**; it is easier to use codes from 1 to 4, for small to extra-large pizzas, and 1 to 7 for days of the week.

  The charm of statistical packages is that you can add more telling labels to anyway.

You are probably a very busy researcher, with many projects. It is highly recommended that you create *folders* for each and every project where you store all the files (maybe in subfolders) related to that project.

Once the data are stored in CSV-format, the next step is to open **JASP** and use the browse option under **<File><Open><Computer>** to read the CSV-file.

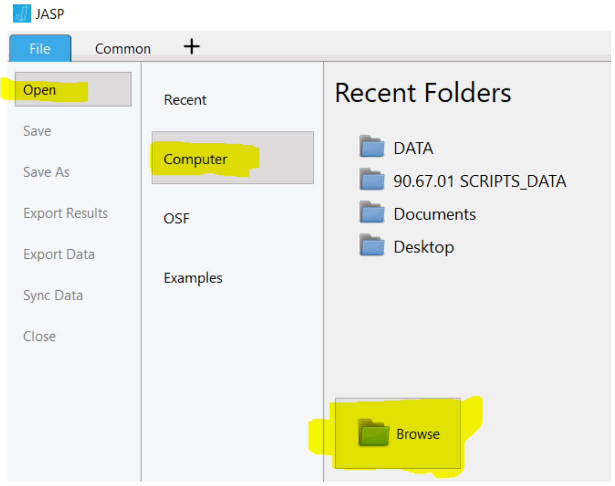
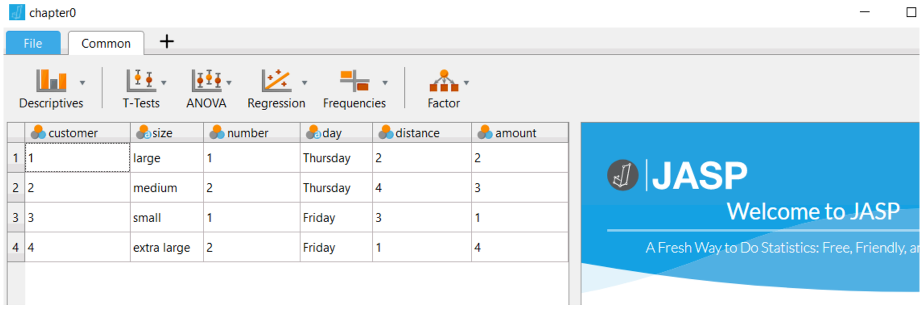


Figure 0.5: Browsing in JASP

In Windows, the file is shown with an icon that looks like:

The “X” in the icon makes you think that the file is an Excel file, but it’s a comma-separated-values text file that can be opened in Excel.

After clicking on the file, it will open. Your screen should look like this.



	customer	size	number	day	distance	amount
1	1	large	1	Thursday	2	2
2	2	medium	2	Thursday	4	3
3	3	small	1	Friday	3	1
4	4	extra large	2	Friday	1	4

Figure 0.6: Open a CSV-file using

Self-test

Make a data set in Excel of 10 of your friends and family members, and record their age, gender, length, and hair color. Save the data set, and then read the data into **JASP**.

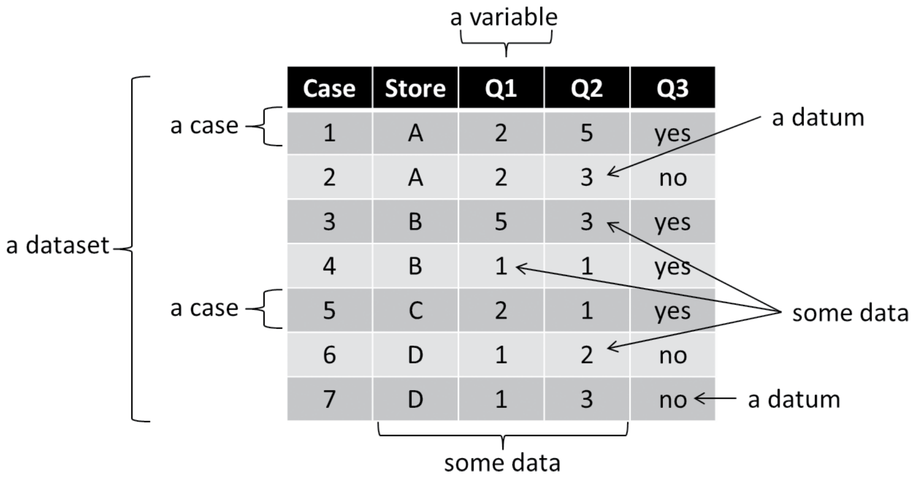
Try to find a way to get the average length!

1 The Language of Statistics

Files needed:
chapter1.xlsx
chapter1.csv
sally.xlsx
sally.csv
sallyForJASP.csv

1.1 Getting Data into JASP

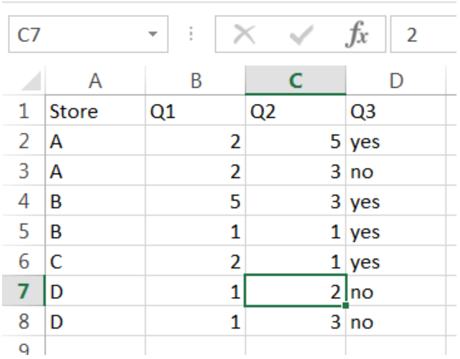
On page 16/17 of Landers there's an example of a data set.



Case	Store	Q1	Q2	Q3
1	A	2	5	yes
2	A	2	3	no
3	B	5	3	yes
4	B	1	1	yes
5	C	2	1	yes
6	D	1	2	no
7	D	1	3	no

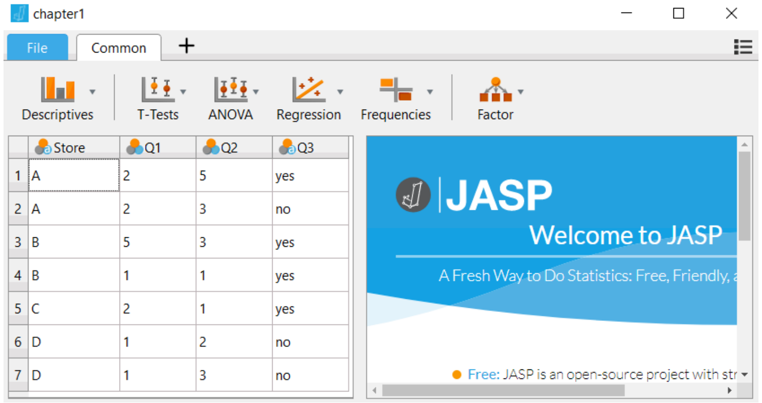
Figure 1.1: Example of a data set (Landers, 2013: 16)

Many researchers use Excel to key in their data. We have done the job for you, in **chapter1.xlsx**. Open the file in Excel, and have a look! As explained, you have to save the file in CSV-format, in order to be able to read the data in **JASP**.



	A	B	C	D
1	Store	Q1	Q2	Q3
2	A	2	5	yes
3	A	2	3	no
4	B	5	3	yes
5	B	1	1	yes
6	C	2	1	yes
7	D	1	2	no
8	D	1	3	no
9				

Figure 1.2: Data for Chapter 1 (in Excel)



	Store	Q1	Q2	Q3
1	A	2	5	yes
2	A	2	3	no
3	B	5	3	yes
4	B	1	1	yes
5	C	2	1	yes
6	D	1	2	no
7	D	1	3	no

JASP

Welcome to JASP

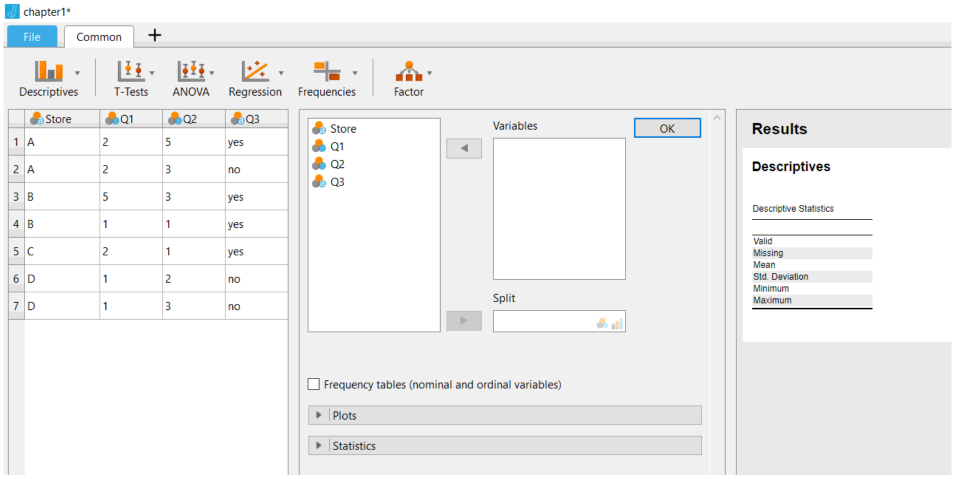
A Fresh Way to Do Statistics: Free, Friendly, a

Free: JASP is an open-source project with str

Figure 1:3: The data imported in JASP

Descriptive information can be obtained using the **<Descriptives>** tab. You can click on the variables that you want to describe, and click them to the **Variables** column. The moment you do so, the **Results** screen on the right will show the number of valid and missing cases; the mean (average) value; the standard deviation; and the minimum and maximum value.

Let’s ask for descriptive information on variables **Q1** to **Q3**.



	Store	Q1	Q2	Q3
1	A	2	5	yes
2	A	2	3	no
3	B	5	3	yes
4	B	1	1	yes
5	C	2	1	yes
6	D	1	2	no
7	D	1	3	no

□ Frequency tables (nominal and ordinal variables)

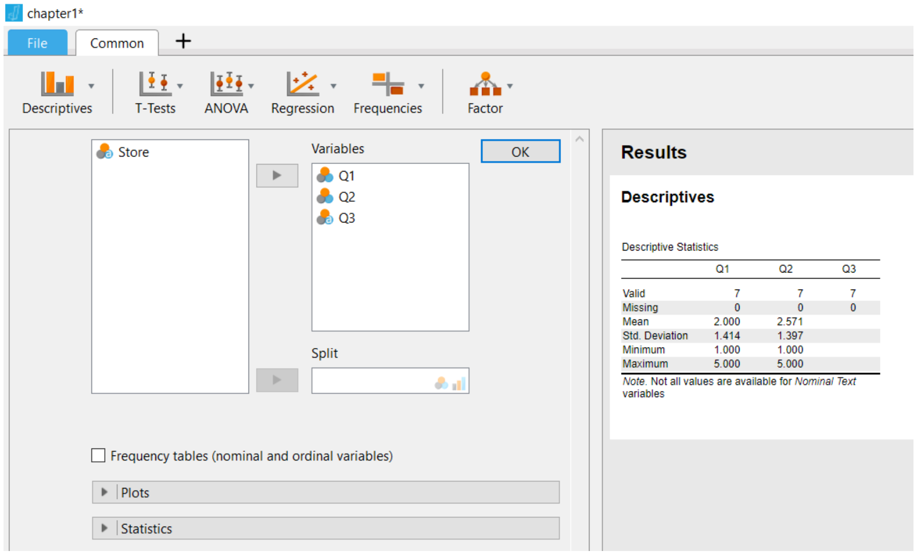
Results

Descriptives

Descriptive Statistics

Valid
Missing
Mean
Std. Deviation
Minimum
Maximum

Figure 1:4: Descriptives in JASP



Variables: Q1, Q2, Q3

□ Frequency tables (nominal and ordinal variables)

▸ Plots

▸ Statistics

Results

Descriptives

Descriptive Statistics

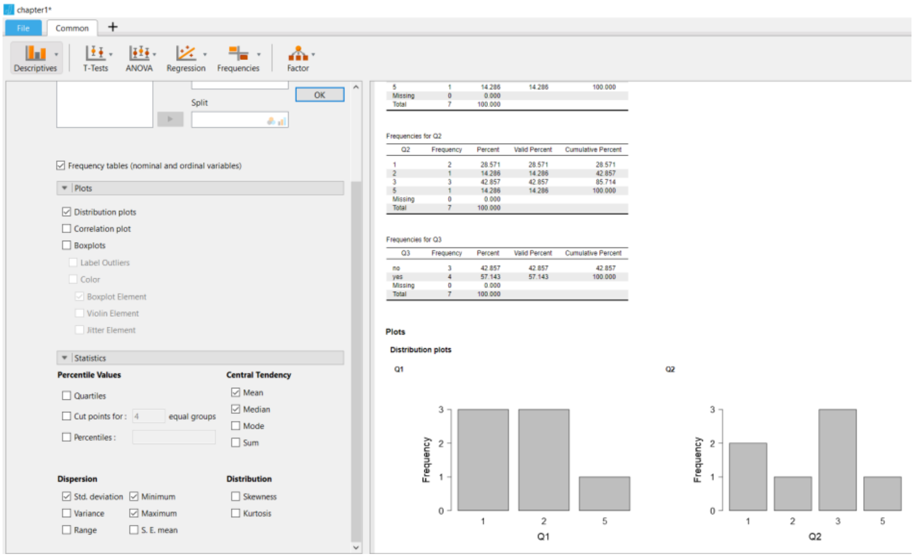
	Q1	Q2	Q3
Valid	7	7	7
Missing	0	0	0
Mean	2.000	2.571	
Std. Deviation	1.414	1.397	
Minimum	1.000	1.000	
Maximum	5.000	5.000	

Note. Not all values are available for *Nominal Text* variables

Figure 1:5: Default Descriptives of Variables Q1-Q3

Variable **Q2**, for example, has a mean value of 2.571, and there are 7 valid (non-missing) values. Since variable **Q3** is a text variable, it is not possible to show its mean value, as explained in the note to the table.

In the options part at the bottom, you can ask for frequency tables, as would be relevant for **Q3**. You can opt for additional descriptive statistics like median and mode, and ask for plots. For example, we tick the option **frequency tables**, and add the median value. Part of the output is shown in figure 1.6. The default options (e.g. Mean) are ticked; you can suppress them by unticking the boxes.



☑ Frequency tables (nominal and ordinal variables)

▾ Plots

☑ Distribution plots
□ Correlation plot
□ Boxplots
□ Label Outliers
□ Color
☑ Boxplot Element
□ Violin Element
□ Jitter Element

▾ Statistics

Percentile Values
□ Quartiles
□ Cut points for : 4 equal groups
□ Percentiles :

Central Tendency
☑ Mean
☑ Median
□ Mode
□ Sum

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

Distribution
□ Skewness
□ Kurtosis

5	1	14.286	14.286	100.000
Missing	0	0.000		
Total	7	100.000		

Frequencies for Q2

Q2	Frequency	Percent	Valid Percent	Cumulative Percent
1	2	28.571	28.571	28.571
2	1	14.286	14.286	42.857
3	3	42.857	42.857	85.714
5	1	14.286	14.286	100.000
Missing	0	0.000		
Total	7	100.000		

Frequencies for Q3

Q3	Frequency	Percent	Valid Percent	Cumulative Percent
no	3	42.857	42.857	42.857
yes	4	57.143	57.143	100.000
Missing	0	0.000		
Total	7	100.000		

Plots

Distribution plots

Figure 1:6: Extra Descriptives of Variables Q1-Q3

A nice feature is that you can copy parts of the results page, and paste them into a Word document (your thesis or report). See the **bar chart** for **Q3** in the figure below.

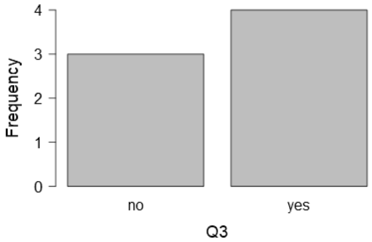


Figure 1:7: Bar Chart for Q3

1.2 Data Skill Challenge

Landers, page 20, Data Skill Challenge 3

Here are Sally's data for her first four pizza deliveries (Customer ID; Size of Pizza Ordered; Number of Pizzas Ordered; Day of the Week; Distance from Store; Time of Day; Total Price):

Customer 1: Large, 1, Thursday, 2 km, evening, £2

Customer 2: Medium, 2, Thursday, 4 km, evening £3

Customer 3: Small, 1, Friday, 3 km, afternoon, £1

Customer 4: Extra Large, Friday, 1 km, evening £4

Task: Enter the data in Excel, and read the data into JASP

Extra 1: Add new variables; change variables

Extra 2: Summarize the data (number of deliveries by day; average amount spent; and so on)

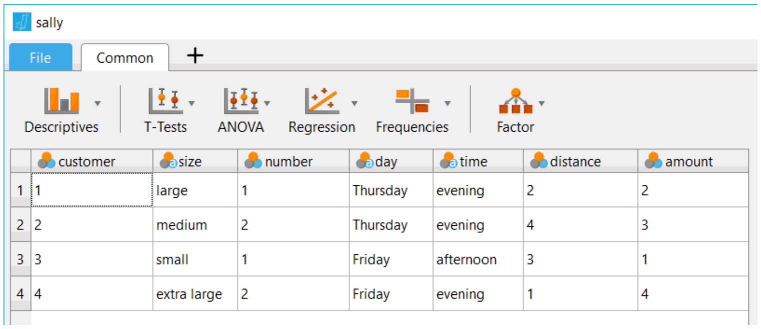
Extra 3: Sorting your data

Solutions

Reading the data

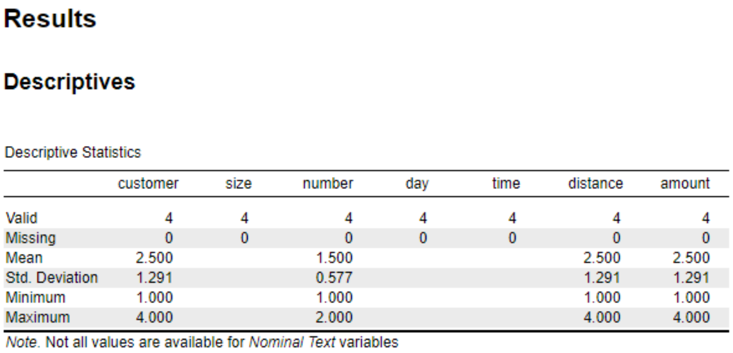
If you don't feel like keying in the data (in Excel, or in **JASP**) then feel free to import the file **sally.csv** in **JASP**. Or, to get some practice, first save **sally.xlsx** in CSV-format and then read the data in **JASP**.

This is what the data looks like.



	customer	size	number	day	time	distance	amount
1	1	large	1	Thursday	evening	2	2
2	2	medium	2	Thursday	evening	4	3
3	3	small	1	Friday	afternoon	3	1
4	4	extra large	2	Friday	evening	1	4

We can summarize information and compute means.



Results

Descriptives

Descriptive Statistics

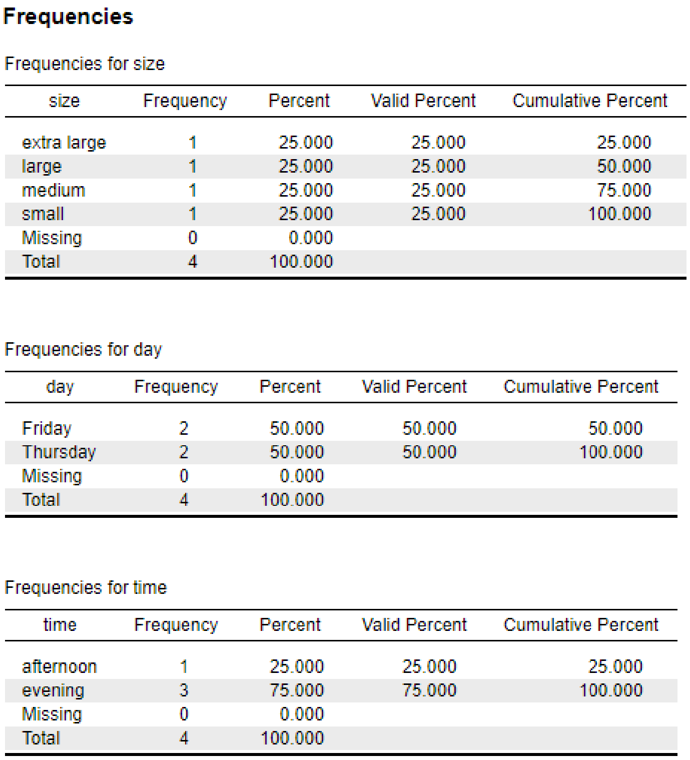
	customer	size	number	day	time	distance	amount
Valid	4	4	4	4	4	4	4
Missing	0	0	0	0	0	0	0
Mean	2.500		1.500			2.500	2.500
Std. Deviation	1.291		0.577			1.291	1.291
Minimum	1.000		1.000			1.000	1.000
Maximum	4.000		2.000			4.000	4.000

Note. Not all values are available for *Nominal Text* variables

The mean distance is 2.5 (kilometers). The distance ranges from a minimum of 1 to a maximum of 4.

The mean size of a pizza we do not know, since it’s coded as a string variable. The customer identifier is coded numerically, but interpreting the mean is pointless.

It would be more elegant to leave out **customer**, **size**, **day** and **time** from the descriptives. For these variables, you can ask for **frequencies** and suppress information on mean, minimum and maximum.



Frequencies

Frequencies for size

size	Frequency	Percent	Valid Percent	Cumulative Percent
extra large	1	25.000	25.000	25.000
large	1	25.000	25.000	50.000
medium	1	25.000	25.000	75.000
small	1	25.000	25.000	100.000
Missing	0	0.000		
Total	4	100.000		

Frequencies for day

day	Frequency	Percent	Valid Percent	Cumulative Percent
Friday	2	50.000	50.000	50.000
Thursday	2	50.000	50.000	100.000
Missing	0	0.000		
Total	4	100.000		

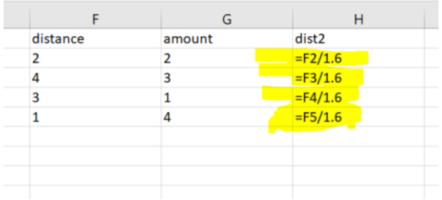
Frequencies for time

time	Frequency	Percent	Valid Percent	Cumulative Percent
afternoon	1	25.000	25.000	25.000
evening	3	75.000	75.000	100.000
Missing	0	0.000		
Total	4	100.000		

Extra 1: Add new records or variables; change variables

JASP does not offer a direct way to edit data, or to generate new variables in your data set. You have to do this in Excel. But luckily, you can use the synchronize (**Sync data**) facility. If you double-click on the data panel, the CSV-file will open in Excel. You have access to all functions in Excel.

As an example, assume that we have the distance in kilometers but we want to have a new variable that gives the distance in miles (1 mile ~ 1.6 km). To create **dist2** we add a column, and use an Excel function to compute the distances in miles.



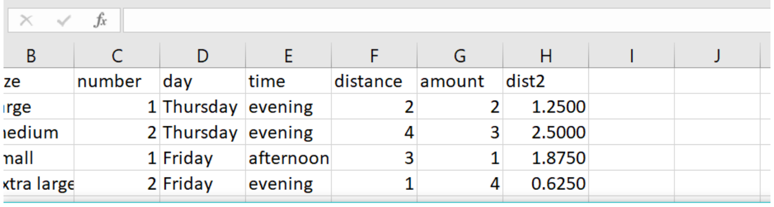
F	G	H
distance	amount	dist2
2	2	=F2/1.6
4	3	=F3/1.6
3	1	=F4/1.6
1	4	=F5/1.6

Upon saving the CSV-file (**sallyForJASP.csv**, in order to keep the original data intact), the new variable is added automatically to the data in **JASP** (after using **Sync Data** under **File**).

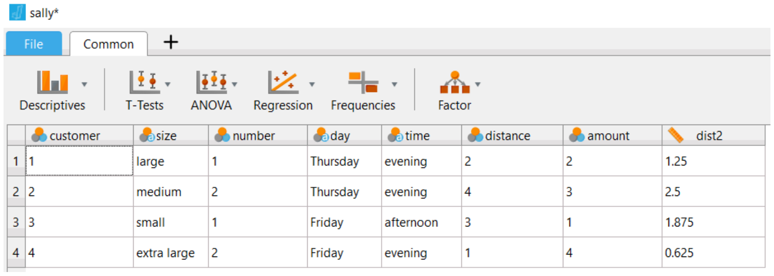
One warning: the CSV-file stores the values rather than formulas, so in order to keep the overview of everything you have done (like the formulas in the table above), you would have to note it down somewhere else.

The best option is to make all changes in the Excel (*.xlsx) file, and add a worksheet with notes, comments, and so on. The Excel-file will keep the formulas.

Next, store the worksheet containing the data as a CSV-file, and read that one in **JASP**.



B	C	D	E	F	G	H	I	J
ze	number	day	time	distance	amount	dist2		
ırge	1	Thursday	evening	2	2	1.2500		
ıedium	2	Thursday	evening	4	3	2.5000		
nall	1	Friday	afternoon	3	1	1.8750		
xtra large	2	Friday	evening	1	4	0.6250		

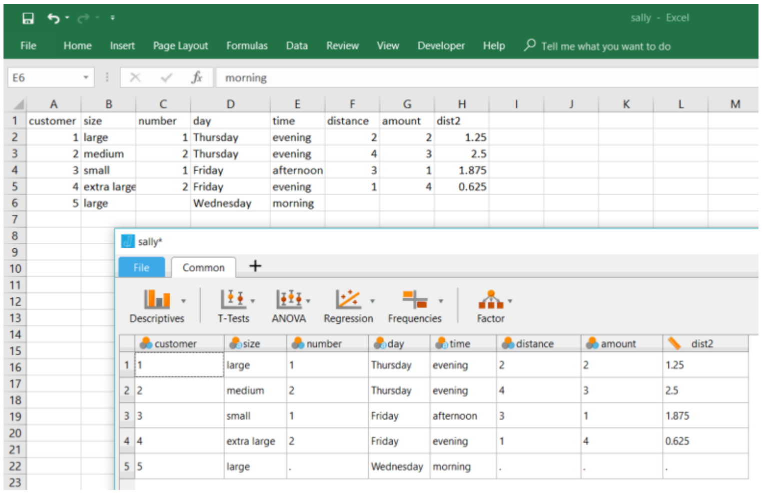


	customer	size	number	day	time	distance	amount	dist2
1	1	large	1	Thursday	evening	2	2	1.25
2	2	medium	2	Thursday	evening	4	3	2.5
3	3	small	1	Friday	afternoon	3	1	1.875
4	4	extra large	2	Friday	evening	1	4	0.625

Self-test:

Suppose that all distances (in kilometers; and converted into miles) were incorrect, since Sally first has to pick up the pizzas from the kitchen which is an extra mile drive. To make that correction, add 1 (mile) to **dist2**.

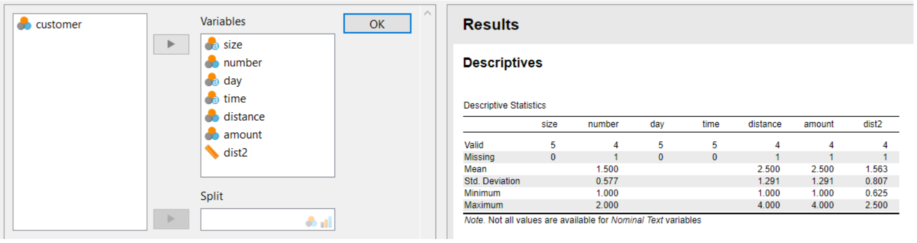
In the same way you can add data for additional respondents. Suppose that a new customer orders one large pizza, on a Wednesday morning; all other data for this delivery are missing.



	A	B	C	D	E	F	G	H
1	customer	size	number	day	time	distance	amount	dist2
2	1	large	1	Thursday	evening	2	2	1.25
3	2	medium	2	Thursday	evening	4	3	2.5
4	3	small	1	Friday	afternoon	3	1	1.875
5	4	extra large	2	Friday	evening	1	4	0.625
6	5	large		Wednesday	morning			

	customer	size	number	day	time	distance	amount	dist2
1	1	large	1	Thursday	evening	2	2	1.25
2	2	medium	2	Thursday	evening	4	3	2.5
3	3	small	1	Friday	afternoon	3	1	1.875
4	4	extra large	2	Friday	evening	1	4	0.625
5	5	large	.	Wednesday	morning	.	.	.

Double-click on the data panel, or go to the CSV-file if it's still open. Add a new line of data, save the file, and the new data are immediately visible in **JASP**. The missing values are shown as a dot, and will be interpreted as missing when describing the data.



Results

Descriptives

Descriptive Statistics

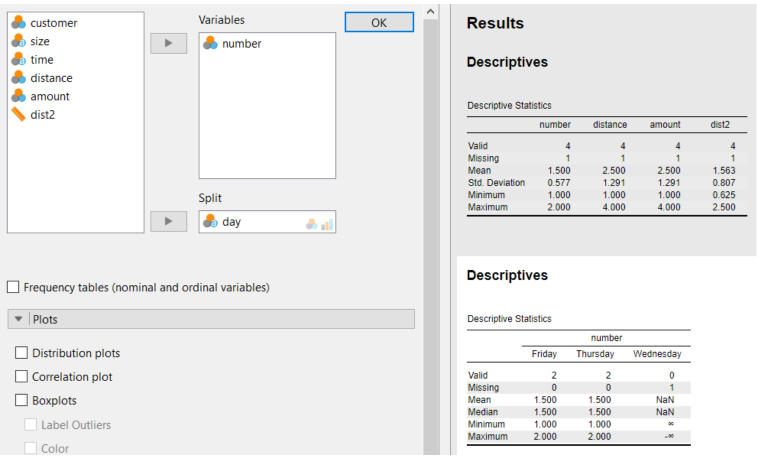
	size	number	day	time	distance	amount	dist2
Valid	5	4	5	5	4	4	4
Missing	0	1	0	0	1	1	1
Mean		1.500			2.500	2.500	1.563
Std. Deviation		0.577			1.291	1.291	0.807
Minimum		1.000			1.000	1.000	0.625
Maximum		2.000			4.000	4.000	2.500

Note. Not all values are available for *Nominal Text* variables

Extra 2: Summarize the data (number of deliveries by day; average amount spent; and so on)

All descriptive statistics are readily obtained under the **Descriptives** tab. Use the default settings first, and start experimenting with the various options. Ticking or unticking options takes immediate effect on the results displayed in the righthand side window.

You can break down the descriptive information for one or more variables, by one variable. For example, you can analyze the number of orders per day of the week. The results below indicate that the average number of pizzas ordered is the same (1.5) on both days for which we have information. We do not have information on the number of pizzas ordered on Wednesday, and a "NaN" (Not a Number) is displayed



□ Frequency tables (nominal and ordinal variables)

□ Distribution plots
□ Correlation plot
□ Boxplots
□ Label Outliers
□ Color

Results

Descriptives

Descriptive Statistics

	number	distance	amount	dist2
Valid	4	4	4	4
Missing	1	1	1	1
Mean	1.500	2.500	2.500	1.563
Std. Deviation	0.577	1.291	1.291	0.807
Minimum	1.000	1.000	1.000	0.625
Maximum	2.000	4.000	4.000	2.500

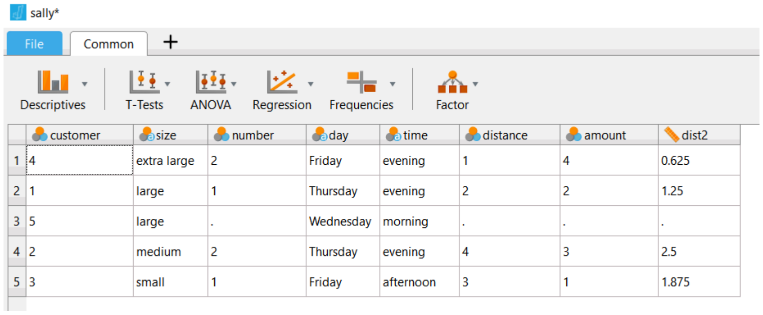
Descriptives

Descriptive Statistics

	number		
	Friday	Thursday	Wednesday
Valid	2	2	0
Missing	0	0	1
Mean	1.500	1.500	NaN
Median	1.500	1.500	NaN
Minimum	1.000	1.000	∞
Maximum	2.000	2.000	-∞

Extra 3: Sorting your data

JASP cannot sort the data for you. Again, you have to open the data in Excel (in CSV-format), and use the Excel-functions to sort the data, and then save the data. The sorted data are now shown in **JASP**. For sorting on the variable **size**:



	customer	size	number	day	time	distance	amount	dist2
1	4	extra large	2	Friday	evening	1	4	0.625
2	1	large	1	Thursday	evening	2	2	1.25
3	5	large	.	Wednesday	morning	.	.	.
4	2	medium	2	Thursday	evening	4	3	2.5
5	3	small	1	Friday	afternoon	3	1	1.875

To summarize: click on **Sync Data** under the **<File>** tab. Then make the changes in the CSV-file, and save the file. The changes will show in **JASP**.

Self-test

Open the CSV-file that you have created in the previous chapter in **JASP**, with data on your 10 friends and family members.

You have recorded their age, in years. Now generate a new variable that contains their age in 10 years from now.

Advanced: if you have recorded length in centimeters, convert it to feet and inches (in new variables)! For example, 180cm is equivalent to 5’11’’ (5 feet and 11 inches, rounded to nearest integers).

Sort the data set in descending order of age.

Summarize age, for male and female persons.

2 Working with Numbers and Graphs

Files needed:

chapter2freq.xls
chapter2freq.csv
chapter2case.xlsx
chapter2case.csv [data, n=1,334]
chapter2dsc.csv
chapter2dscJ.csv
alienation.csv [data]

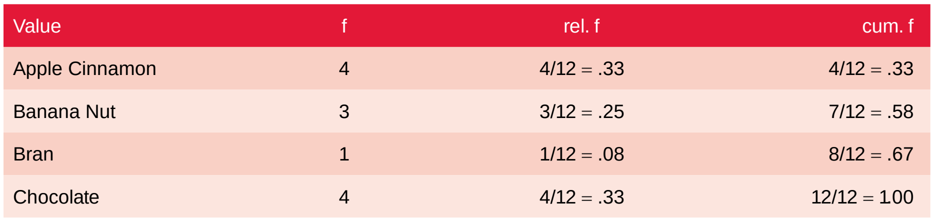
2.1 Descriptive Information

Sometimes you have to key in your own data. But at other times somebody has already done the job for you! On page 27 of his textbook Landers introduces a list of muffin purchases. The list was entered into an SPSS data file by Landers, and converted to a CSV-data file named **chapter2freq.csv**[3].

Muffin Purchases
Chocolate
Chocolate
Banana Nut
Apple Cinnamon
Chocolate
Banana Nut
Apple Cinnamon
Apple Cinnamon
Chocolate
Bran
Apple Cinnamon
Banana Nut

Figure 2.1: The data in chapter2freq.xlsx

On page 27 Landers presents a simple frequency table. It looks like:



Value	f	rel. f	cum. f
Apple Cinnamon	4	4/12 = .33	4/12 = .33
Banana Nut	3	3/12 = .25	7/12 = .58
Bran	1	1/12 = .08	8/12 = .67
Chocolate	4	4/12 = .33	12/12 = 1.00

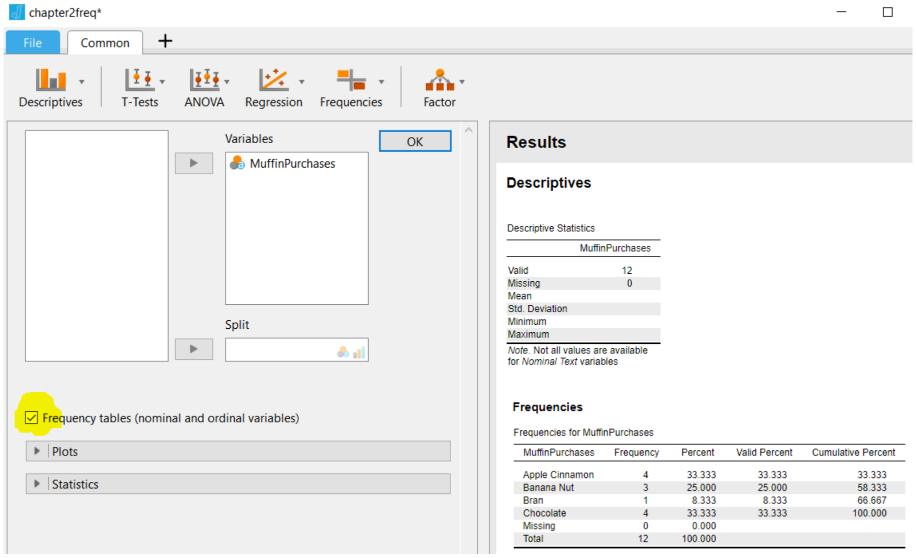
(Source: Landers, 2013)

Figure 2.2: An Example of a Frequency Table

[3] There are several ways to transfer data from one format to the other. A very handy package is **StatTransfer** which has the capability to convert any format into any other format. A (free) alternative is to use the **foreign** package in **R** that enables you to read most formats and convert them into readable formats. Please refer to our module on *Basic Statistics for Business Using **R*** for detailed information.

In the table you can read the frequency of the various flavors sold (4 times Apple Cinnamon); the cumulative frequency (7 times apple cinnamon or banana nut); the relative frequency (0.33 or 33% of the muffins sold are chocolate flavor); and the cumulative relative frequency.

In **JASP** you can obtain the same information using **Descriptives**.



Variables: MuffinPurchases

Split

☑ Frequency tables (nominal and ordinal variables)

Results

Descriptives

Descriptive Statistics

	MuffinPurchases
Valid	12
Missing	0
Mean	
Std. Deviation	
Minimum	
Maximum	

Note. Not all values are available for *Nominal Text* variables

Frequencies

Frequencies for MuffinPurchases

MuffinPurchases	Frequency	Percent	Valid Percent	Cumulative Percent
Apple Cinnamon	4	33.333	33.333	33.333
Banana Nut	3	25.000	25.000	58.333
Bran	1	8.333	8.333	66.667
Chocolate	4	33.333	33.333	100.000
Missing	0	0.000		
Total	12	100.000		

Figure 2.3: Frequency Table in JASP

2.2 Graphs

2.2.1 Bar Chart

We will illustrate how the graphs in Landers can be replicated.

First of all, we want a bar chart like in Landers, page 29.

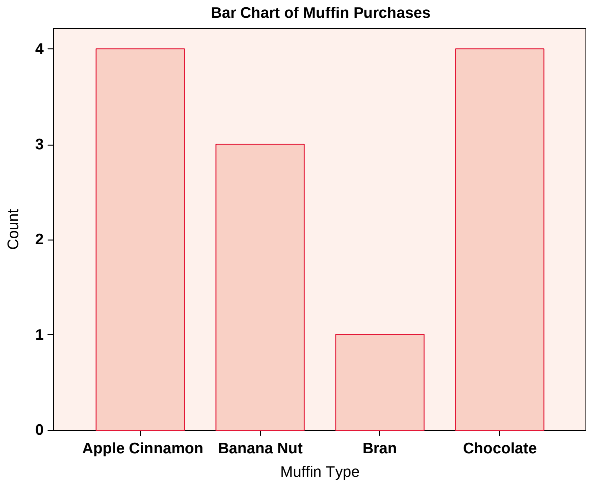
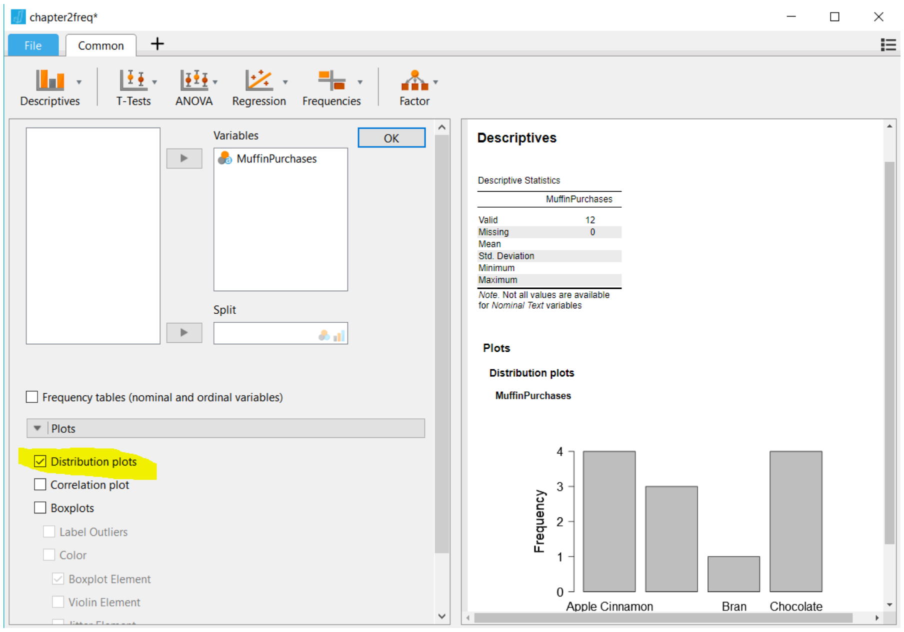


Figure 2.4 Bar Chart (Landers' example)

The bar chart is obtained by opening the **Plots** options, and clicking on **Distribution Plots**.



Variables: MuffinPurchases

Split

□ Frequency tables (nominal and ordinal variables)

☑ Distribution plots
□ Correlation plot
□ Boxplots
□ Label Outliers
□ Color
☑ Boxplot Element
□ Violin Element

Descriptives

Descriptive Statistics

	MuffinPurchases
Valid	12
Missing	0
Mean	
Std. Deviation	
Minimum	
Maximum	

Note. Not all values are available for *Nominal Text* variables

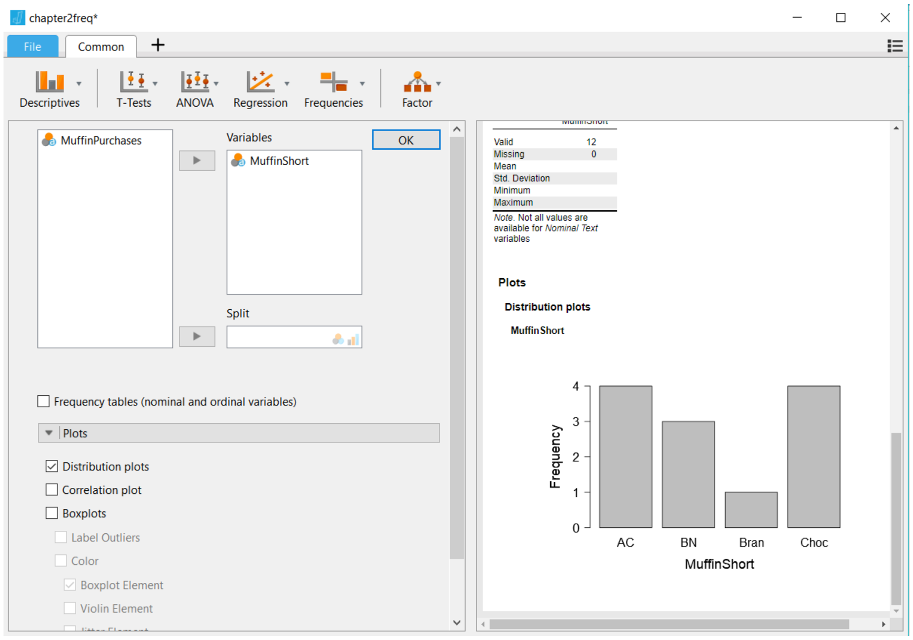
Plots

Distribution plots

MuffinPurchases

Figure 2.5 Bar Chart (Landers' example)

You will notice that the long labels "Apple Cinnamon" and "Banana Nut" do not fit on the horizontal axis, and parts are omitted. That would look bad in your report. For cosmetic reasons, you can add a new column with short labels in your CSV-file, and sync it to **JASP**. For example:



MuffinPurchases

Variables: MuffinShort

Split

□ Frequency tables (nominal and ordinal variables)

☑ Distribution plots
□ Correlation plot
□ Boxplots
□ Label Outliers
□ Color
☑ Boxplot Element
□ Violin Element

	MuffinShort
Valid	12
Missing	0
Mean	
Std. Deviation	
Minimum	
Maximum	

Note. Not all values are available for *Nominal Text* variables

Plots

Distribution plots

MuffinShort

Figure 2.6 Bar Chart – with edited labels

Advanced: recoding data using Excel

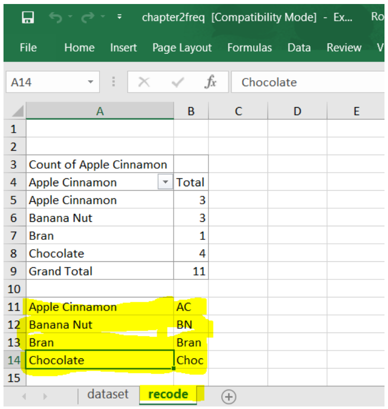
Here, we are "recoding" data. In statistics, we often want to recode our data. In the example above, we recode the existing codes containing long labels one-to-one into shorter labels. In Excel, we can achieve this using the ***vlookup()*** *function.*

We will do it step-by-step.

1. *Open the Excel file* ***chapter2freq.xlsx****.*

2. *The first column contains the long labels. We want to add a second column containing shorter labels using a recoding scheme (in which, for example, "Apple Cinnamon" is recoded into "AC")*

3. *The recoding schemes are in a separate worksheet, **recode**. In the range A11:B14 we add the original codes in column A and the new codes in column B. In cells A3:B9 we have used a pivot table, just to get a sorted list of all original codes in our data.*

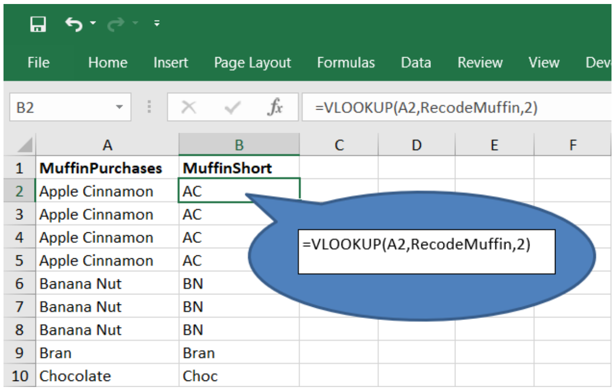


A14 Chocolate

	A	B	C	D	E
1					
2					
3	Count of Apple Cinnamon				
4	Apple Cinnamon ▾	Total			
5	Apple Cinnamon	3			
6	Banana Nut	3			
7	Bran	1			
8	Chocolate	4			
9	Grand Total	11			
10					
11	Apple Cinnamon	AC			
12	Banana Nut	BN			
13	Bran	Bran			
14	Chocolate	Choc			
15					

4. *We name the range A11:B14 in **recode** as **RecodeMuffin***

5. *In the worksheet with data, we add a second column **MuffinShort**, based on a **VLOOKUP()** formula. The formula looks up the value in column A, and then looks up that value in **RecodeMuffin**, the recoding scheme in the **Recodings** worksheet. It returns the value in the second column of the recoding scheme.*



B2 =VLOOKUP(A2,RecodeMuffin,2)

	A	B	C	D	E	F
1	MuffinPurchases	MuffinShort				
2	Apple Cinnamon	AC				
3	Apple Cinnamon	AC				
4	Apple Cinnamon	AC				
5	Apple Cinnamon	AC				
6	Banana Nut	BN				
7	Banana Nut	BN				
8	Banana Nut	BN				
9	Bran	Bran				
10	Chocolate	Choc				

=VLOOKUP(A2,RecodeMuffin,2)

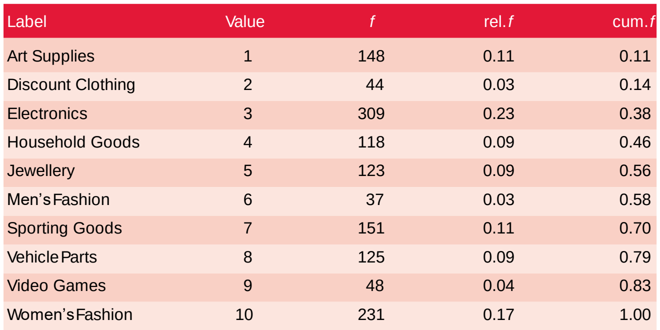
Figure 2.7 Recoding in Excel using VLOOKUP()

2.2.2 A Bigger Data Set

The data set

Normally we have data sets with more observations, and more variables than in the simple examples used so far.

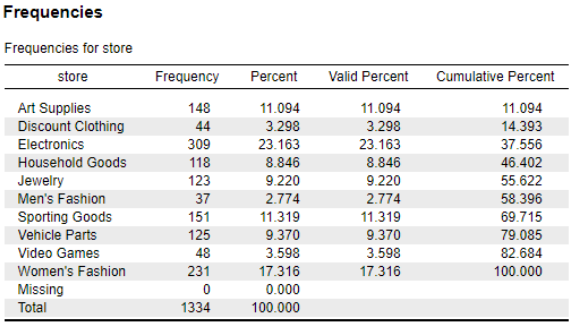
The example of Landers contains no less than 1,334 records.



Label	Value	*f*	rel.*f*	cum.*f*
Art Supplies	1	148	0.11	0.11
Discount Clothing	2	44	0.03	0.14
Electronics	3	309	0.23	0.38
Household Goods	4	118	0.09	0.46
Jewellery	5	123	0.09	0.56
Men's Fashion	6	37	0.03	0.58
Sporting Goods	7	151	0.11	0.70
Vehicle Parts	8	125	0.09	0.79
Video Games	9	48	0.04	0.83
Women's Fashion	10	231	0.17	1.00

Figure 2.8: Frequency Table (Landers' example; page 43)

The data are in **chapter2case.csv**. Read the data in **JASP**, and using **Descriptives**, you get:



Frequencies

Frequencies for store

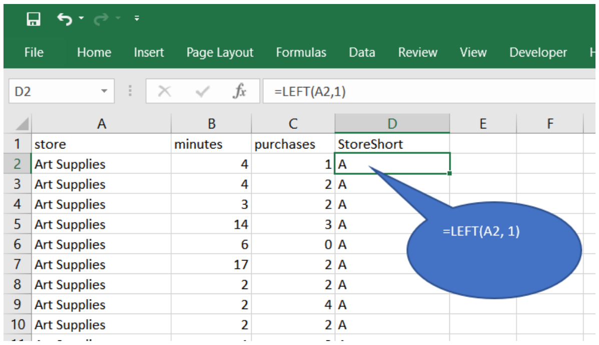
store	Frequency	Percent	Valid Percent	Cumulative Percent
Art Supplies	148	11.094	11.094	11.094
Discount Clothing	44	3.298	3.298	14.393
Electronics	309	23.163	23.163	37.556
Household Goods	118	8.846	8.846	46.402
Jewelry	123	9.220	9.220	55.622
Men's Fashion	37	2.774	2.774	58.396
Sporting Goods	151	11.319	11.319	69.715
Vehicle Parts	125	9.370	9.370	79.085
Video Games	48	3.598	3.598	82.684
Women's Fashion	231	17.316	17.316	100.000
Missing	0	0.000		
Total	1334	100.000		

Figure 2.9: Frequency Table in JASP

Bar chart

Let's make a bar graph of **store**. As expected the full labels do not fit on the x-axis. Using Excel, we have added a column with just the first character of the label.

To accomplish this in Excel, you can use the **LEFT()** function, as shown below, or alternatively use recoding via **VLOOKUP()**. Again, save the data and they are synced to **JASP**.



=LEFT(A2,1)

	A	B	C	D	E	F
1	store	minutes	purchases	StoreShort		
2	Art Supplies	4	1	A		
3	Art Supplies	4	2	A		
4	Art Supplies	3	2	A		
5	Art Supplies	14	3	A		
6	Art Supplies	6	0	A		
7	Art Supplies	17	2	A		
8	Art Supplies	2	2	A		
9	Art Supplies	2	4	A		
10	Art Supplies	2	2	A		

=LEFT(A2, 1)

Figure 2.10: Creating Short Labels

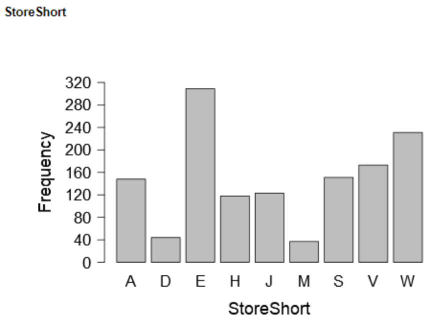


Figure 2.11: Bar chart

Histogram

Minutes spent on websites, is a continuous variable. Rather than a bar chart, we use a histogram to show the frequency or the density of ranges of minutes spent on the websites.

The width of the bins is set by default, and cannot be changed in **JASP**.

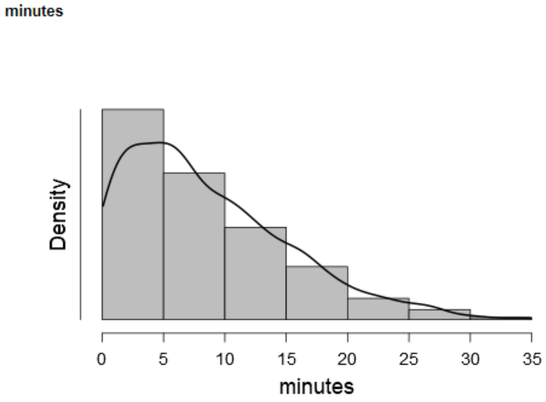


Figure 2.12: Histogram

As an alternative to a histogram for depicting distributions, you can opt for a *box-plot*. We will postpone that to the chapter on ANOVA.

Scattergram

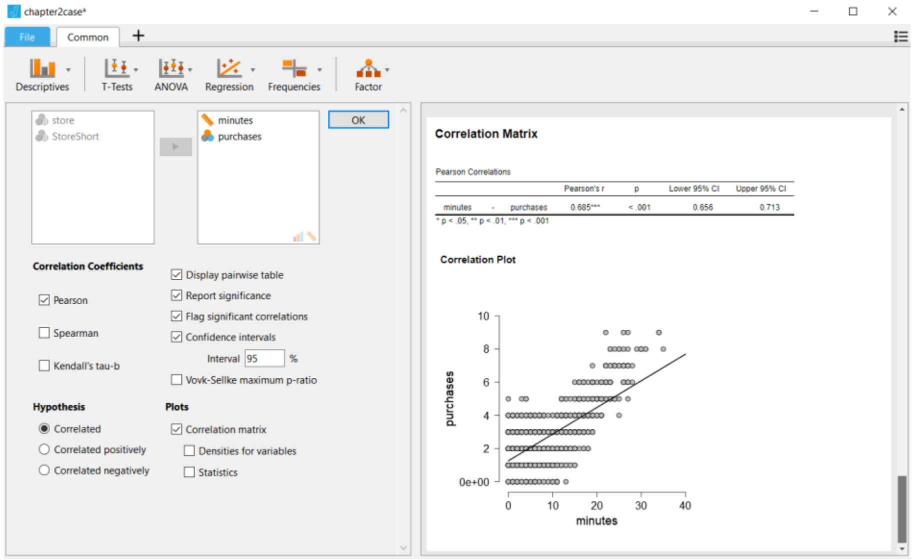
The challenge in statistical research starts when we start analyzing relationships between two or more variables.

In our example, the obvious interesting questions are about the type and strength of the relationship between minutes spent on a website on the one hand, and the number of purchases on the other. We can show the relationship graphically in a scattergram.

The commands below reproduce the scattergram on page 63 of Landers.

Rather than using **Descriptives**, we use **Regression**. Under **Regression**, click on **Correlation Matrix**. Click the two variables of interest, to the right panel. Next, select the options you're interested in. Here, we have selected the plot for the correlation matrix, and added some statistical information.

From the graph it is clear that there is a positive *correlation* between the two variables: the number of purchases goes up with minutes spent on the website. As a measure of strength of the relationship, the *correlation coefficient* is computed. We will come back to that in the chapter on correlation. The correlation is significantly different from zero, indicating that the positive relationship that we have found is not coincidental.



Correlation Coefficients

☑ Pearson

☐ Spearman

☐ Kendall's tau-b

☑ Display pairwise table

☑ Report significance

☑ Flag significant correlations

☑ Confidence intervals

Interval 95 %

☐ Vovk-Sellke maximum p-ratio

Hypothesis

● Correlated

○ Correlated positively

○ Correlated negatively

Plots

☑ Correlation matrix

☐ Densities for variables

☐ Statistics

Correlation Matrix

Pearson Correlations

			Pearson's r	p	Lower 95% CI	Upper 95% CI
minutes	-	purchases	0.685***	< .001	0.656	0.713

* p < .05, ** p < .01, *** p < .001

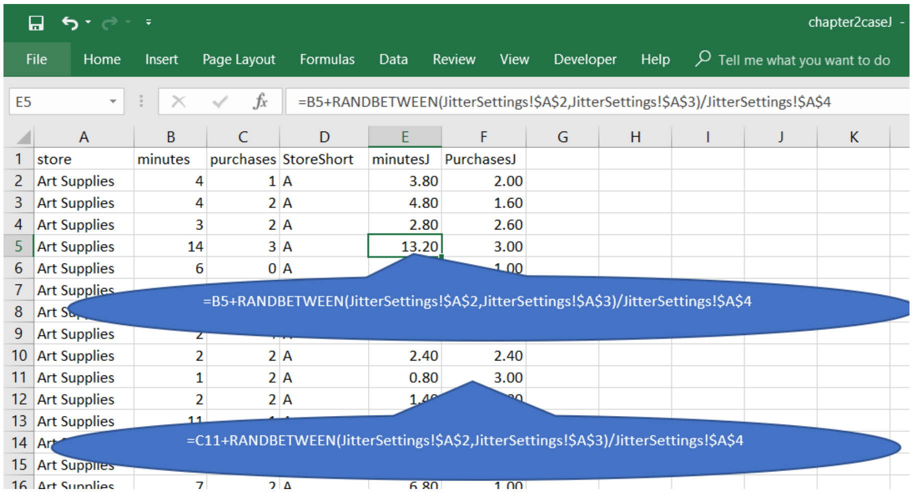
Figure 2.13: Scattergram

The problem with the graph is that, even though we have no less than 1,334 records, you only see a much smaller number of dots. How come? The data are recorded in discrete units: people spend one, two, three, et cetera minutes on websites and make one, two, three et cetera purchases. As a consequence, all occurrences of, say, 3 minutes spent on the website and 2 purchases made are represented by one single dot!

To get a better view of the density, researchers add some "*jitter*": a small random deviation from the actual values in the data set: (3; 2) for example may become (3.05; 1.95). This option is implemented in **STATA**, and in **R** packages.

Advanced: adding jitter

*You can add jitter to the variables, in Excel. We have used the **RANDBETWEEN(bottom,top)** function. The function adds a random number between **bottom** and **top**. Since we want to center the random deviation around the actual values, be pick a negative and a positive number of the same absolute value, for example -5 and +5. Since -5 and +5 would add deviations that are very large relative to the actual numbers we divide the random number by a scale factor (say, 5). In order to experiment with different values, we store these three values (for **bottom**, **top** and **scale**) in a separate worksheet. Be sure to do this in Excel sheet, so that the formulas are stored!*

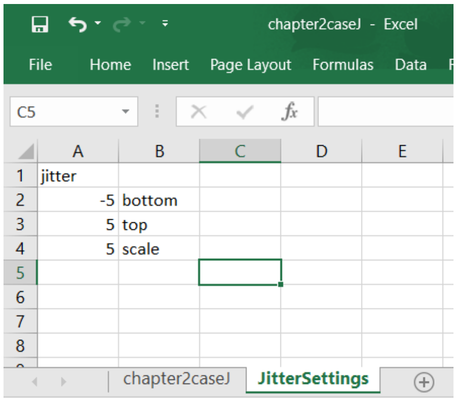


E5 =B5+RANDBETWEEN(JitterSettings!$A$2,JitterSettings!$A$3)/JitterSettings!$A$4

	A	B	C	D	E	F	G	H	I	J	K
1	store	minutes	purchases	StoreShort	minutesJ	PurchasesJ					
2	Art Supplies	4	1	A	3.80	2.00					
3	Art Supplies	4	2	A	4.80	1.60					
4	Art Supplies	3	2	A	2.80	2.60					
5	Art Supplies	14	3	A	13.20	3.00					
6	Art Supplies	6	0	A		1.00					
7	Art Supplies										
8	Art Su										
9	Art Supplies	2									
10	Art Supplies	2	2	A	2.40	2.40					
11	Art Supplies	1	2	A	0.80	3.00					
12	Art Supplies	2	2	A	1.40						
13	Art Supplies	11									
14	Art										
15	Art Supplies										
16	Art Supplies	7	2	A	6.80	1.00					

=B5+RANDBETWEEN(JitterSettings!$A$2,JitterSettings!$A$3)/JitterSettings!$A$4

=C11+RANDBETWEEN(JitterSettings!$A$2,JitterSettings!$A$3)/JitterSettings!$A$4



	A	B
1	jitter	
2	-5	bottom
3	5	top
4	5	scale

Figure 2.14: Adding Jitter to your Data

Remember that adding jitter, is only done for cosmetic reasons: your scattergram will look better! The correlation coefficient to be reported is the one based on data without jitter. If you repeat this exercise, your values will be a little different from mine: the random numbers will be different after each calculation.

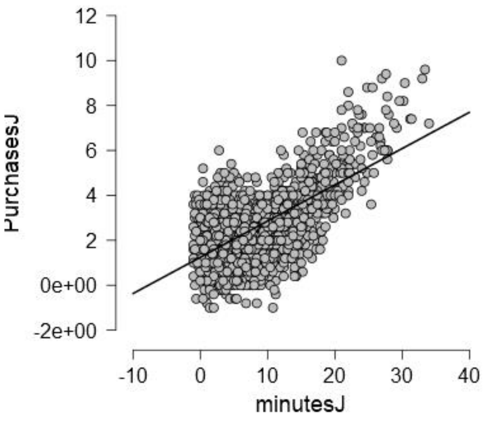


Figure 2.15: Scattergram with Jitter

2.3 Data Skill Challenges

Task 1 (Data skill challenge 3, Landers, page 65)

Consider the following dataset and create a data frame.

Sales in Chinese Yuan: 200000; 125000; 180000; 170000; 210000; 190000; 220000; 180000

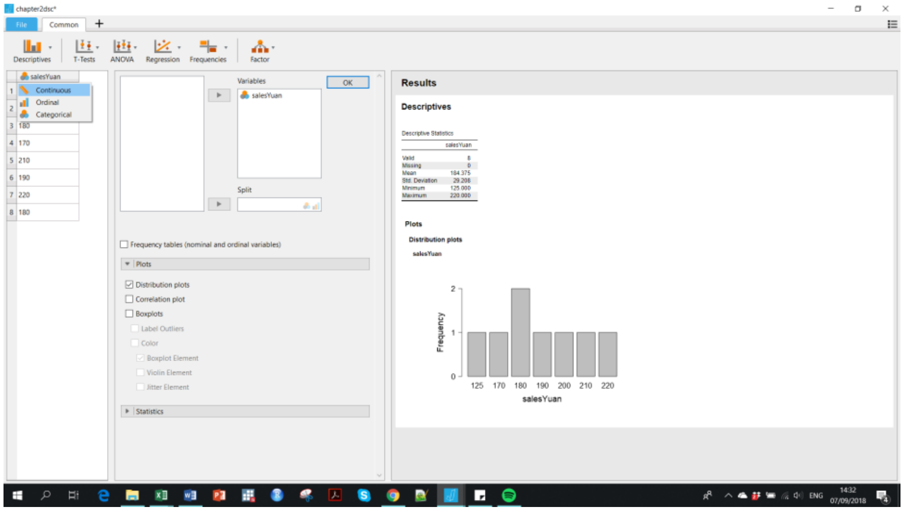
Depict the data in a histogram!

The data are in **chapter2dsc.csv**.

In cases like this, since all figures end with three zeroes we might as well skip them. When drawing a histogram, **JASP** computes defaults for number of bins, and the width of the bins.

Even if we tell **JASP** that the variable is continuous, the limited number of unique values causes the program to produce a bar chart, rather than a histogram. In a histogram, the values on the x-axis can be interpreted as distances. But the distance between 125 and 170, should be larger than the distance between 170 and 180!

You can click on the variable in the data panel, to change the type of variable. You can define the variable as Continuous, Ordinal or Categorical. From its contents, **JASP** would interpret **salesYuan** as categorical (each value is a code for a group), while actually it should be interpreted as a number (sales can be any number in the range from 0 to infinity). As indicated, even changing the type of variable to Continuous does not have an effect on the plot that we get.



	salesYuan
1	
2	
3	180
4	170
5	210
6	190
7	220
8	180

Variables: salesYuan

□ Frequency tables (nominal and ordinal variables)

▾ Plots

☑ Distribution plots

□ Correlation plot

□ Boxplots

□ Label Outliers

□ Color

☑ Boxplot Element

□ Violin Element

□ Jitter Element

▸ Statistics

Results

Descriptives

Descriptive Statistics

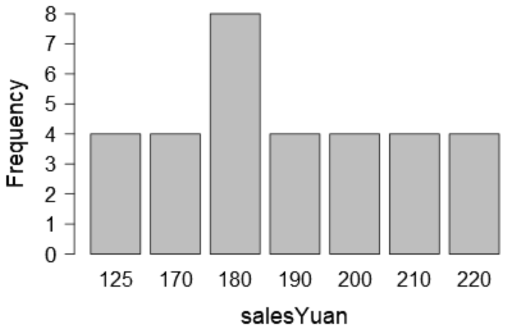
	salesYuan
Valid	8
Missing	0
Mean	184.375
Std. Deviation	29.208
Minimum	125.000
Maximum	220.000

Plots

Distribution plots

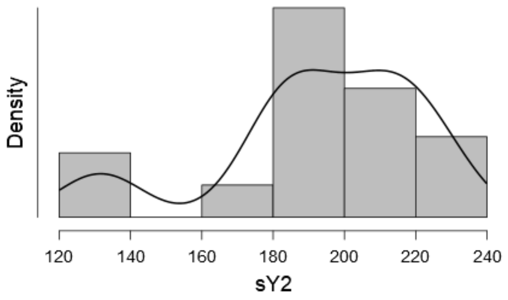
salesYuan

salesYuan



This problem is a minor one. If we create a data set with more (32) unique numbers, then we do get a histogram. In the example below, we have replicated the eight records four times, and added a random number between 0 and 20 to each record. The data can be found in **chapter2dscJ.csv**.

sY2



Task 2

Open the file "alienation.csv"

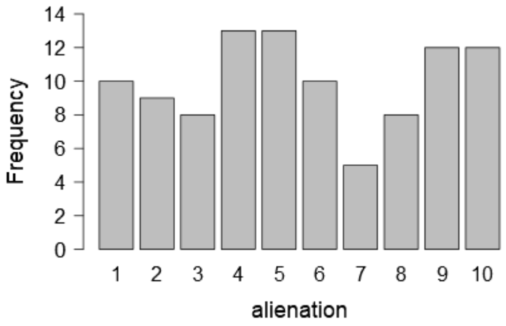
The file contains data on alienation from society, and on income.

- **Summarize the data (how many records do we have? What's the mean of income and alienation? And the minimum and maximum values?)**
- **Display the distribution of the two variables, graphically**
- **Display the relationship between the two variables, graphically, and add a line of best fit!**

Distribution plots

alienation

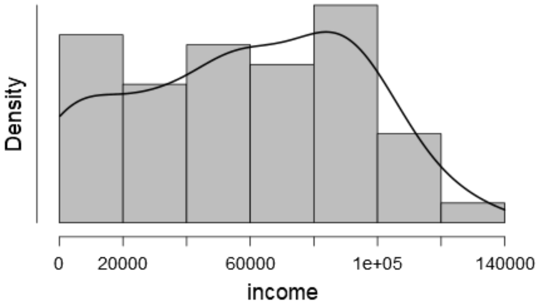
The data can be displayed graphically.



We can build a histogram for **income**.

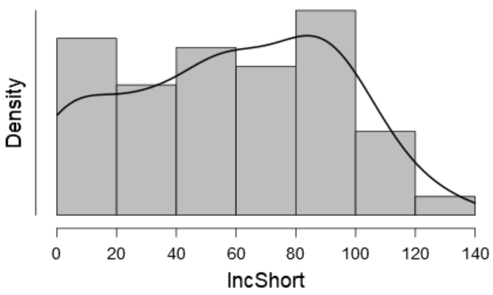
Distribution plots

income



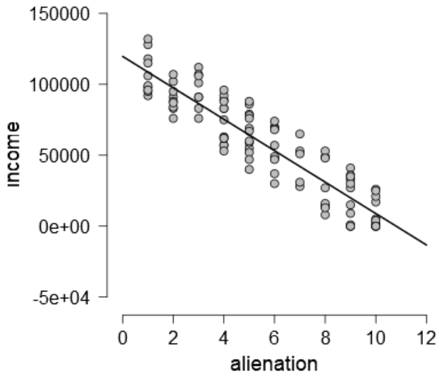
On the axis, the value for 100,000 is displayed as 1e+05 (the so-called scientific notation, which stands for $1*10^5$). In your report, you can avoid this by adding a column for income in thousands of units (dollars, or euros).

IncShort

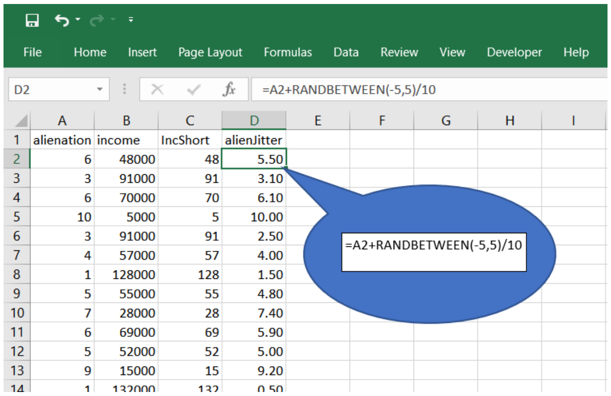


The relationship between the two variables can be shown in a scattergram.

Correlation Plot



Since we have a limited number of observations (100) and at least one of the variables (**income**) is continuous, most of the points are unique, and do not overlap with other points. You can add some jitter yourself, by adding a column (**alienJitter**) in Excel that adds a random positive or negative deviation to the original value of **alienation**, using Excel's **RANDBETWEEN()** function.

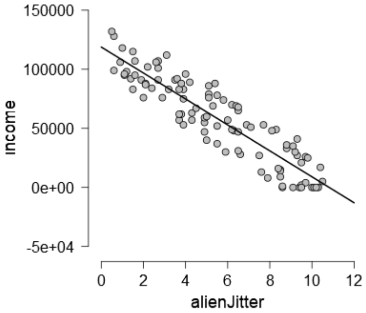


=A2+RANDBETWEEN(-5,5)/10

	A	B	C	D	E	F	G	H	I
1	alienation	income	IncShort	alienJitter					
2	6	48000	48	5.50					
3	3	91000	91	3.10					
4	6	70000	70	6.10					
5	10	5000	5	10.00					
6	3	91000	91	2.50					
7	4	57000	57	4.00					
8	1	128000	128	1.50					
9	5	55000	55	4.80					
10	7	28000	28	7.40					
11	6	69000	69	5.90					
12	5	52000	52	5.00					
13	9	15000	15	9.20					
14	1	132000	132	0.50					

=A2+RANDBETWEEN(-5,5)/10

Correlation Plot



3 Central Tendency and Variability

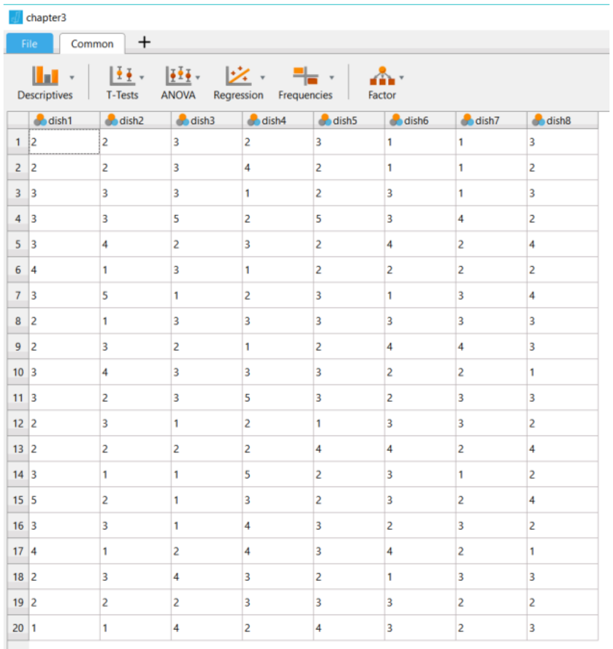
<u>Files needed:</u>
chapter3.xlsx
chapter3.csv
chapter3_dsc1.xlsx

3.1 Looking at the "Distribution" of Your Data

The data for chapter 3 (Landers, page 66-67) represents ratings by 20 restaurant-goers of the eight dishes on the menu of a restaurant.

According to the text, a 7-point scale was used, but since there are no values larger than 5, this is either a mistake in the text or the quality of the food was not that good.

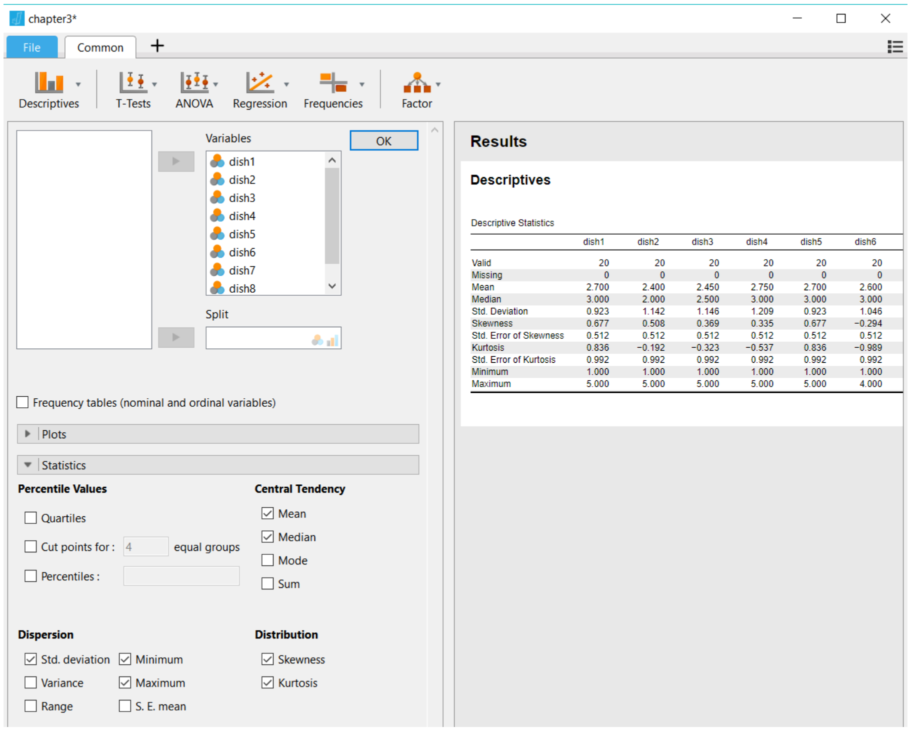
We open the CSV-file (**chapter3.csv**) and show the data.



	dish1	dish2	dish3	dish4	dish5	dish6	dish7	dish8
1	2	2	3	2	3	1	1	3
2	2	2	3	4	2	1	1	2
3	3	3	3	1	2	3	1	3
4	3	3	5	2	5	3	4	2
5	3	4	2	3	2	4	2	4
6	4	1	3	1	2	2	2	2
7	3	5	1	2	3	1	3	4
8	2	1	3	3	3	3	3	3
9	2	3	2	1	2	4	4	3
10	3	4	3	3	3	2	2	1
11	3	2	3	5	3	2	3	3
12	2	3	1	2	1	3	3	2
13	2	2	2	2	4	4	2	4
14	3	1	1	5	2	3	1	2
15	5	2	1	3	2	3	2	4
16	3	3	1	4	3	2	3	2
17	4	1	2	4	3	4	2	1
18	2	3	4	3	2	1	3	3
19	2	2	2	3	3	3	2	2
20	1	1	4	2	4	3	2	3

Figure 3.1: The data for chapter 3

We describe the data, and add some options (*skewness* and *kurtosis*).



Variables

dish1
dish2
dish3
dish4
dish5
dish6
dish7
dish8

Split

□ Frequency tables (nominal and ordinal variables)

Percentile Values

□ Quartiles

□ Cut points for : 4 equal groups

□ Percentiles :

Central Tendency

☑ Mean

☑ Median

□ Mode

□ Sum

Dispersion

☑ Std. deviation ☑ Minimum

□ Variance ☑ Maximum

□ Range □ S. E. mean

Distribution

☑ Skewness

☑ Kurtosis

Results

Descriptives

Descriptive Statistics

	dish1	dish2	dish3	dish4	dish5	dish6
Valid	20	20	20	20	20	20
Missing	0	0	0	0	0	0
Mean	2.700	2.400	2.450	2.750	2.700	2.600
Median	3.000	2.000	2.500	3.000	3.000	3.000
Std. Deviation	0.923	1.142	1.146	1.209	0.923	1.046
Skewness	0.677	0.508	0.369	0.335	0.677	−0.294
Std. Error of Skewness	0.512	0.512	0.512	0.512	0.512	0.512
Kurtosis	0.836	−0.192	−0.323	−0.537	0.836	−0.989
Std. Error of Kurtosis	0.992	0.992	0.992	0.992	0.992	0.992
Minimum	1.000	1.000	1.000	1.000	1.000	1.000
Maximum	5.000	5.000	5.000	5.000	5.000	4.000

Figure 3.2: The data for chapter 3 in JASP

The *mean* of a distribution is simply the sum of values divided by the number of values. The mean (or average) is a widely used measure of the *central tendency* of a distribution.

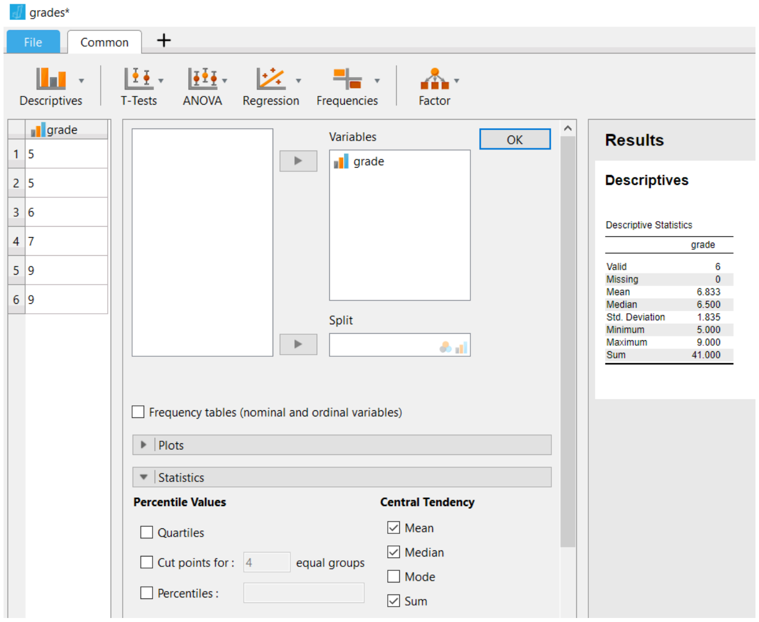
$$\bar{x} = \frac{\sum x_i}{n}$$

The Σ (sigma) is the Greek letter “S”, and is the operator used for computing the sum of values. Here, we compute the sum of all grades (*i* = 1 to 6). The ”n” is the number of observations.

For example, if your grades for 5 modules are: 5, 5, 6, 7, 9, then your mean (or average) grade is the sum of all grades (32) divided by the number of grades (5), or 32/5 = 6.4.

An alternative measure of central tendency is the *median*. Your median grade is the grade that is exactly in the middle of the distribution. To compute the median, you first sort the grades from lowest to highest. The median is the 3rd observed grade in the sorted list, as it has as many observations to its left as to its right; the median is therefore 6. In case of an even number of observations, we pick the mean of observations n/2 and n/2+1. For example, if you would complete your 6th module with a grade of 9, then your sorted list of grades is 5, 5, 6, 7, 9 and 9. The median is in between the 3rd and 4th observation; we take the mean of both, (6+7)/2 = 6.5.

Let’s use **JASP** to do the same. The sum of grades is now 41. The mean is 41/6 = 6.83. The median is, indeed, 6.5.



	grade
1	5
2	5
3	6
4	7
5	9
6	9

Variables: grade

Split

□ Frequency tables (nominal and ordinal variables)

▸ Plots

▾ Statistics

Percentile Values

□ Quartiles

□ Cut points for : 4 equal groups

□ Percentiles :

Central Tendency

☑ Mean

☑ Median

□ Mode

☑ Sum

Results

Descriptives

Descriptive Statistics

	grade
Valid	6
Missing	0
Mean	6.833
Median	6.500
Std. Deviation	1.835
Minimum	5.000
Maximum	9.000
Sum	41.000

Figure 3.3: Mean, median and sum

As a measure for *variability* or *dispersion* in the data, we can use simple methods like minimum and maximum, and range (maximum minus minimum). In our example, the grades are between 5 and 9, and therefore the range is 4.

Widely used measures of dispersion, are the variance, and its square root the standard deviation. These are defined as:

$$Variance = s^2 = \frac{\sum(x - \bar{x})^2}{n}$$

$$Standard\ Deviation = s = \sqrt{s^2}$$

In the formula for variance, the numerator computes the sum of squared deviations from the mean. The closer the grades are to their mean, the lower the variance (dispersion). In the extreme situation of all grades having the same value (say, a grade of 6 for all modules), the mean grade (6) is identical to each and every grade, and the deviations are all zero. The variance and the standard deviation too would be zero.

Since the variance is measured in squared units, the standard deviation, as the square root of the variance, is in the same units as the variable.

Why is the standard deviation so important? Many phenomena in life are “normally distributed”. A normal distribution is characterized by a symmetric, bell-shaped curve, fully determined by its mean and standard deviation. That is, if you assume a distribution to be normal, knowing the mean and the standard deviation makes it possible to make probability statements. If length is normally distributed with a mean of 180cm and standard deviation of 10cm, then we can deduce that 2.5% of the population is taller than 200cm. Referring to the figure below: 95% of the population is in between 180cm minus twice the standard deviation (10cm), and 180cm plus twice the standard deviation, that is between 160 and 200cm; 5% is smaller than 160cm or taller than 200cm. Since the distribution is symmetric, 2.5% is taller than 200cm.

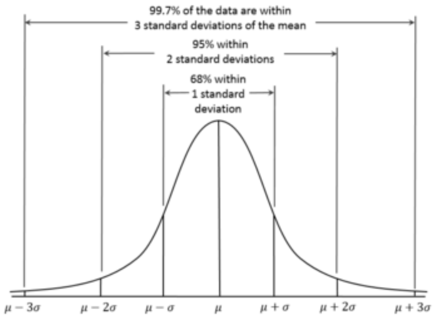


Figure 3.4: The Normal Distribution

This principle is very important, especially in inferential statistics, as we shall see. Often, our empirical distributions are not perfectly normal, and therefore we want to test how likely it is that our sample is drawn from a normally distributed population. One test is based on the shape of the distribution, as revealed by its *skewness* and *kurtosis*.

Skewness is a measure of lack of symmetry. A distribution is symmetric if it looks the same to the left and right of the center point. Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. Distributions with light tails, tend to have a high peak in the center of the distribution; a high kurtosis indicates a high peak.

For a normal (symmetric) distribution the skewness equals zero; the distribution is symmetrical around the mean. Left and right tailed distributions have negative and positive skewness, respectively.

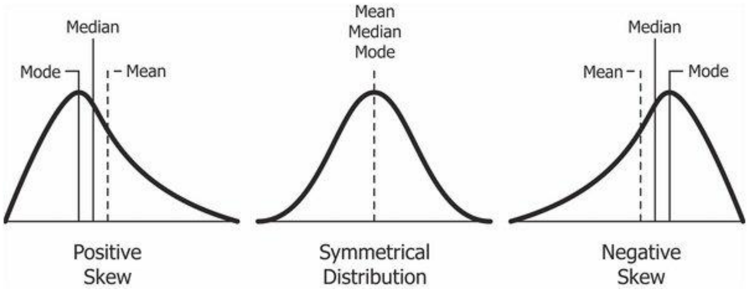


Figure 3.5: Skewed Distributions

The normal distribution has a kurtosis of 3. However, some software packages like **JASP** redefine the kurtosis as centered around zero, simply by subtracting 3. A positive (negative) kurtosis then indicates that the actual distribution is more (less) peaked than expected in a normal distribution.

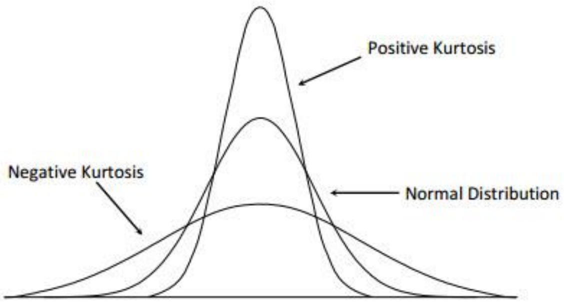


Figure 3.6: Kurtosis

Going back to our data on dishes, the positive (right) skew and strong peak in **dish1** can be illustrated by a bar chart. Check the value for skewness and kurtosis in figure 3.2!

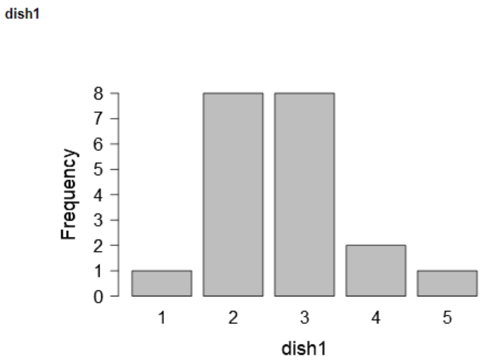
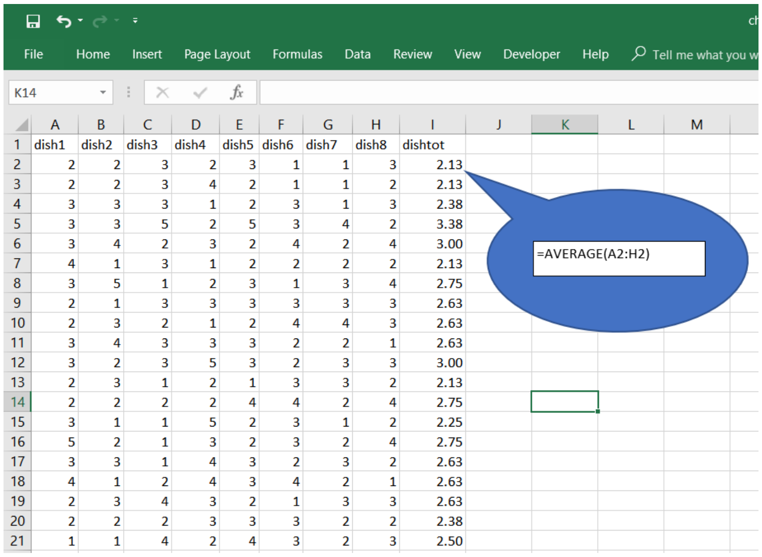


Figure 3.7: Skewed Distribution

We compute (in Excel) a new variable for the average of the eight columns (or variables) **dish1** to **dish8**. Save the file (in CSV-format), and the new variable will be synced to **JASP**.

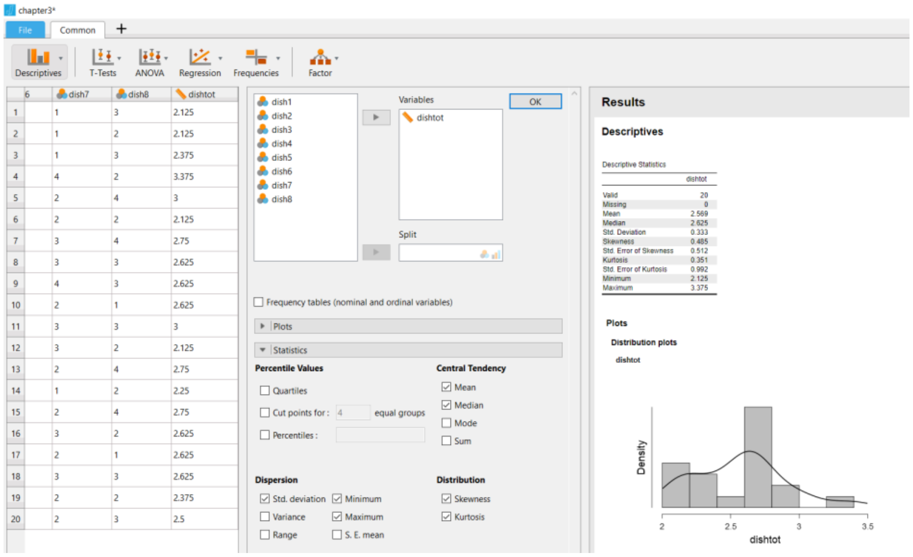


	A	B	C	D	E	F	G	H	I
1	dish1	dish2	dish3	dish4	dish5	dish6	dish7	dish8	dishtot
2	2	2	3	2	3	1	1	3	2.13
3	2	2	3	4	2	1	1	2	2.13
4	3	3	3	1	2	3	1	3	2.38
5	3	3	5	2	5	3	4	2	3.38
6	3	4	2	3	2	4	2	4	3.00
7	4	1	3	1	2	2	2	2	2.13
8	3	5	1	2	3	1	3	4	2.75
9	2	1	3	3	3	3	3	3	2.63
10	2	3	2	1	2	4	4	3	2.63
11	3	4	3	3	3	2	2	1	2.63
12	3	2	3	5	3	2	3	3	3.00
13	2	3	1	2	1	3	3	2	2.13
14	2	2	2	2	4	4	2	4	2.75
15	3	1	1	5	2	3	1	2	2.25
16	5	2	1	3	2	3	2	4	2.75
17	3	3	1	4	3	2	3	2	2.63
18	4	1	2	4	3	4	2	1	2.63
19	2	3	4	3	2	1	3	3	2.63
20	2	2	2	3	3	3	2	2	2.38
21	1	1	4	2	4	3	2	3	2.50

=AVERAGE(A2:H2)

Figure 3.8: Average Scores over 8 Dishes

According to output, **dishtot** is slightly skewed to the right, and the kurtosis is 0.35 (after deducting 3) which is close to that of a normal distribution; the peak in the distribution is somewhat higher than expected under a normal distribution.



	6	dish7	dish8	dishtot
1		1	3	2.125
2		1	2	2.125
3		1	3	2.375
4		4	2	3.375
5		2	4	3
6		2	2	2.125
7		3	4	2.75
8		3	3	2.625
9		4	3	2.625
10		2	1	2.625
11		3	3	3
12		3	2	2.125
13		2	4	2.75
14		1	2	2.25
15		2	4	2.75
16		3	2	2.625
17		2	1	2.625
18		3	3	2.625
19		2	2	2.375
20		2	3	2.5

dish1
dish2
dish3
dish4
dish5
dish6
dish7
dish8

Variables: dishtot

Split:

OK

□ Frequency tables (nominal and ordinal variables)

Plots

Statistics

Percentile Values

□ Quartiles
□ Cut points for : 4 equal groups
□ Percentiles :

Central Tendency

☑ Mean
☑ Median
□ Mode
□ Sum

Dispersion

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

Distribution

☑ Skewness
☑ Kurtosis

Descriptives

Descriptive Statistics

	dishtot
Valid	20
Missing	0
Mean	2.569
Median	2.625
Std. Deviation	0.333
Skewness	0.485
Std. Error of Skewness	0.512
Kurtosis	0.351
Std. Error of Kurtosis	0.992
Minimum	2.125
Maximum	3.375

Plots

Distribution plots

dishtot

Figure 3.9: Distribution of *dishtot*

In the histogram we can add a normal curve when using other packages like **STATA** to visually check how close the actual distribution is to a normal distribution. **JASP** flexibly fits a *kernel* density function.

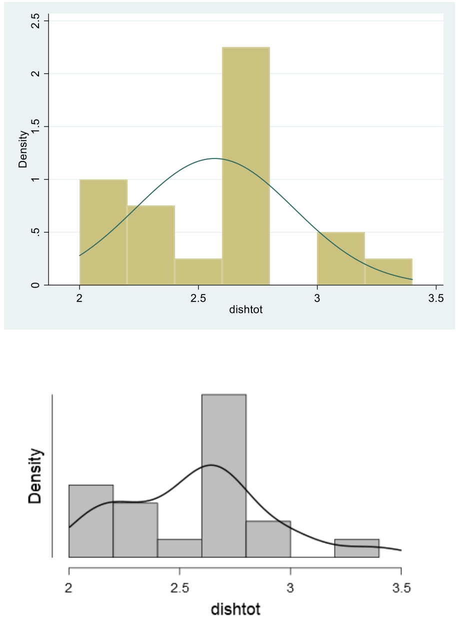
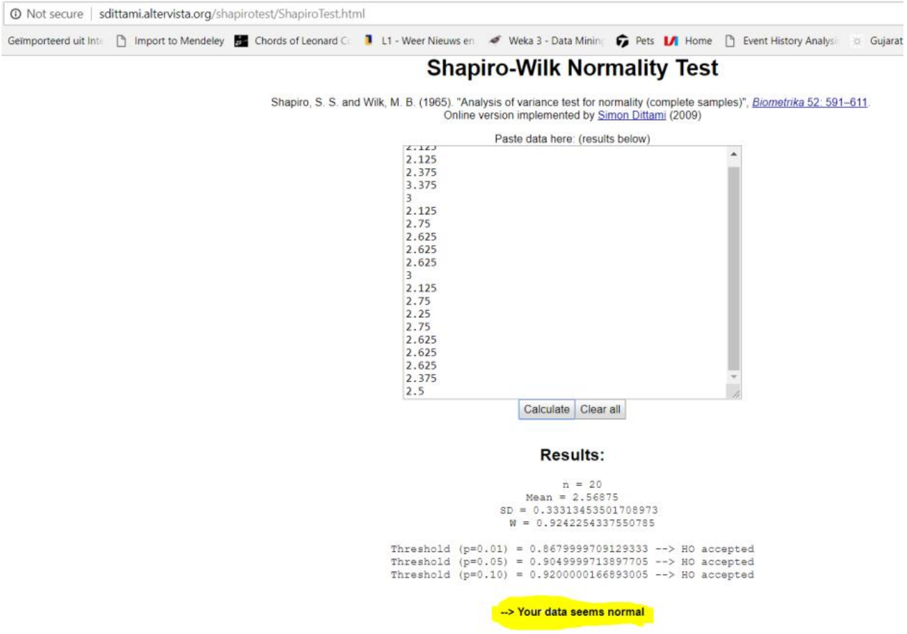


Figure 3.10: Histogram, Fitted Normal Curve, and Kernel Density

The graph indicates that the distribution only remotely resembles a normal distribution. We have too few observations on the left and too many on the right. These deviations may be coincidental. We have a small sample of 20 respondents, and it may not be that unlikely that sampling 20 cases out of a normally distributed population, results in the observed outcome. One test of *normality* is based on skewness and kurtosis.

The test is not implemented in **JASP**, but there are many online tools that can help you out. One such tool can be found here. You can copy your data (from Excel) and paste it in the box, and the results are displayed. The interpretation is that, assuming that the data are sampled from a normal distribution, it is not unlikely to get these results; we don’t have a strong reason to reject the null hypothesis of a normal distribution.



Shapiro-Wilk Normality Test

Shapiro, S. S. and Wilk, M. B. (1965). "Analysis of variance test for normality (complete samples)", *Biometrika 52: 591–611*.
Online version implemented by Simon Dittami (2009)

Paste data here: (results below)

```
2.125
2.375
3.375
3
2.125
2.75
2.625
2.625
2.625
3
2.125
2.75
2.25
2.75
2.625
2.625
2.625
2.375
2.5
```

Results:

```
            n = 20
         Mean = 2.56875
     SD = 0.33313453501708973
      W = 0.9242254337550785

Threshold (p=0.01) = 0.8679999709129333 --> HO accepted
Threshold (p=0.05) = 0.9049999713897705 --> HO accepted
Threshold (p=0.10) = 0.9200000166893005 --> HO accepted
```

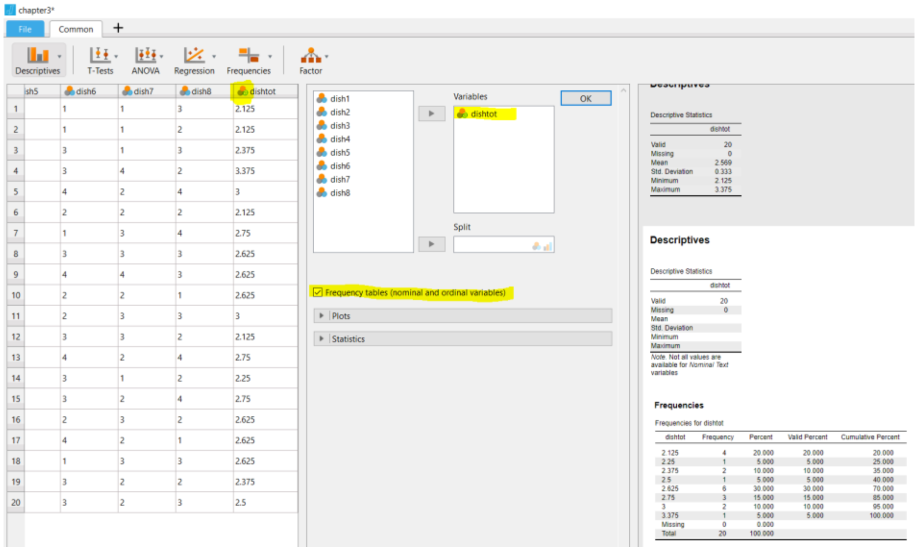
--> Your data seems normal

Figure 3.11: Normality Test Using Online Tools

Landers discusses the **mean**, the **median** and the **mode** of a distribution, as measures of *central tendency*. As we have seen, the mean is the sum of all the values divided by the number of values; the median is the midpoint of the distribution; the mode is the value that occurs most frequently.

The mode is not very popular in social and economic studies. There are two reasons for not using the mode. One important reason is that values may be unique. People’s incomes may be similar, but due to many factors there may be only one person who earns US$ 36,788.26. Maybe there’s not even one. The mode is not useful in case of many unique values – which is typical for continuous variables. Another reason is that a distribution may have two or more modes.

The mode is an option in **JASP**’s Descriptives. The mode for **dishtot** turns out to be 2.625. You can verify that this is the case, by defining the variable as categorical. The program, for good reasons, does not provide frequency tables for continuous variables! The mode here is equal to the median, and close to the mean. However, for continuous variables, this is just coincidence. The mode is only meaningful for categorical and ordinal variables. In one of our previous examples, it makes sense to use the mode for the weekday on which Sally sells most of her pizzas.



	ish5	dish6	dish7	dish8	dishtot
1		1	1	3	2.125
2		1	1	2	2.125
3		3	1	3	2.375
4		3	4	2	3.375
5		4	2	4	3
6		2	2	2	2.125
7		1	3	4	2.75
8		3	3	3	2.625
9		4	4	3	2.625
10		2	2	1	2.625
11		2	3	3	3
12		3	3	2	2.125
13		4	2	4	2.75
14		3	1	2	2.25
15		3	2	4	2.75
16		2	3	2	2.625
17		4	2	1	2.625
18		1	3	3	2.625
19		3	2	2	2.375
20		3	2	3	2.5

dish1
dish2
dish3
dish4
dish5
dish6
dish7
dish8

Variables: dishtot

Split

OK

☑ Frequency tables (nominal and ordinal variables)

▸ Plots

▸ Statistics

Descriptives

Descriptive Statistics

	dishtot
Valid	20
Missing	0
Mean	2.569
Std. Deviation	0.333
Minimum	2.125
Maximum	3.375

Descriptives

Descriptive Statistics

	dishtot
Valid	20
Missing	0
Mean	
Std. Deviation	
Minimum	
Maximum	

Note. Not all values are available for *Nominal Text* variables

Frequencies

Frequencies for dishtot

dishtot	Frequency	Percent	Valid Percent	Cumulative Percent
2.125	4	20.000	20.000	20.000
2.25	1	5.000	5.000	25.000
2.375	2	10.000	10.000	35.000
2.5	1	5.000	5.000	40.000
2.625	6	30.000	30.000	70.000
2.75	3	15.000	15.000	85.000
3	2	10.000	10.000	95.000
3.375	1	5.000	5.000	100.000
Missing	0	0.000		
Total	20	100.000		

Figure 3.12: Checking the Mode of a Distribution

Let's apply what we have learned to one of Landers' data skill challenges.

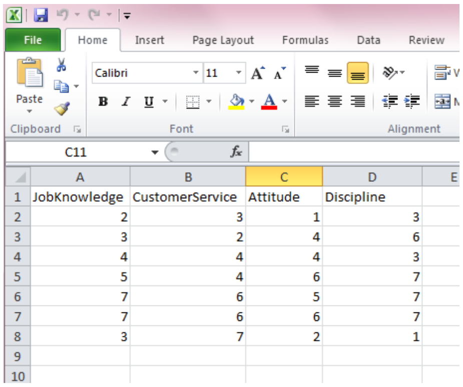
3.2 Data Skill Challenge 1 to Chapter 3

Compute all appropriate central tendency and variability statistics for the data set shown below. These scores are for an employee performance appraisal; responses are given by their supervisors. Each case is a unique employee. Each dimension of employee behavior is assessed on a scale from 1 to 7.

The data are stored in chapter3_dsc1.xlsx

Compute for each employee the average score on all four aspects (use Excel!).

Test if this average score is normally distributed.



	A	B	C	D
1	JobKnowledge	CustomerService	Attitude	Discipline
2	2	3	1	3
3	3	2	4	6
4	4	4	4	3
5	5	4	6	7
6	7	6	5	7
7	7	6	6	7
8	3	7	2	1
9				
10				

Figure 3.13 Data Skill Challenge

By now you should be able to read data entered in Excel-files, into **JASP**.

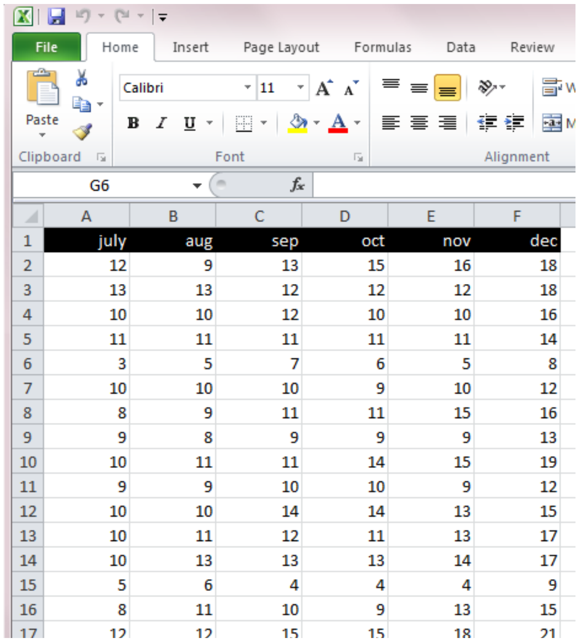
4 Probability Distributions

Files needed:
chapter4.xlsx
chapter4.csv

4.1 Normal Distributions and Z-scores

This chapter focuses on normal distributions and Z-scores.

The data set is described in Landers (page 96). The data set contains data for employee performance (monthly sales) over the last six months.

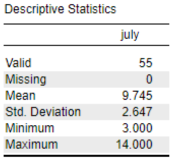


	A	B	C	D	E	F
1	july	aug	sep	oct	nov	dec
2	12	9	13	15	16	18
3	13	13	12	12	12	18
4	10	10	12	10	10	16
5	11	11	11	11	11	14
6	3	5	7	6	5	8
7	10	10	10	9	10	12
8	8	9	11	11	15	16
9	9	8	9	9	9	13
10	10	11	11	14	15	19
11	9	9	10	10	9	12
12	10	10	14	14	13	15
13	10	11	12	11	13	17
14	10	13	13	13	14	17
15	5	6	4	4	4	9
16	8	11	10	9	13	15
17	12	12	15	15	18	21

Figure 4.1 Part of the data used in this chapter

OK, some challenges: how many employees do we have in the data base?

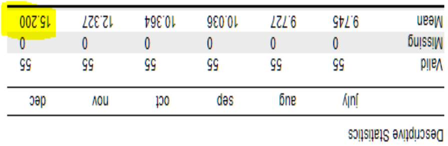
A frequency table of any variable will reveal that the total number of employees is equal to the number of valid plus missing cases, which is 55



Descriptive Statistics

	july
Valid	55
Missing	0
Mean	9.745
Std. Deviation	2.647
Minimum	3.000
Maximum	14.000

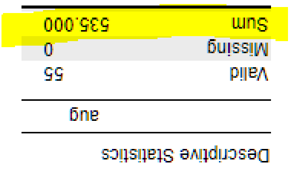
In which month are the average sales per employee at their highest?



Descriptive Statistics

	july	aug	sep	oct	nov	dec
Valid	55	55	55	55	55	55
Missing	0	0	0	0	0	0
Mean	9.745	9.727	10.036	10.364	12.327	15.200

What are the total sales, of all employees in August?



Descriptive Statistics

	aug
Valid	55
Missing	0
Sum	535.000

It's informative to have *standardized* values. Knowing that an employee's sales in August were 16 units is more informative if we know (i) the average sales in August of all employees; (ii) the standard deviation of sales in August; and (iii) the type of distribution.

Economic figures (like sales) tend to be more or less normally distributed.

The **normal distribution** is a symmetric bell-shaped distribution that is characterized by its mean and standard deviation (SD). We obtain the **standard normal (or Z) distribution** with a mean of 0 and an SD of 1, by transforming the original variable X using the formula:

$$Z = \frac{(X - \bar{X})}{SD}$$

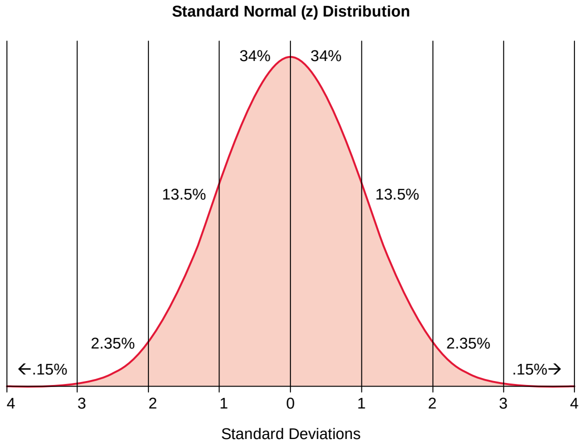
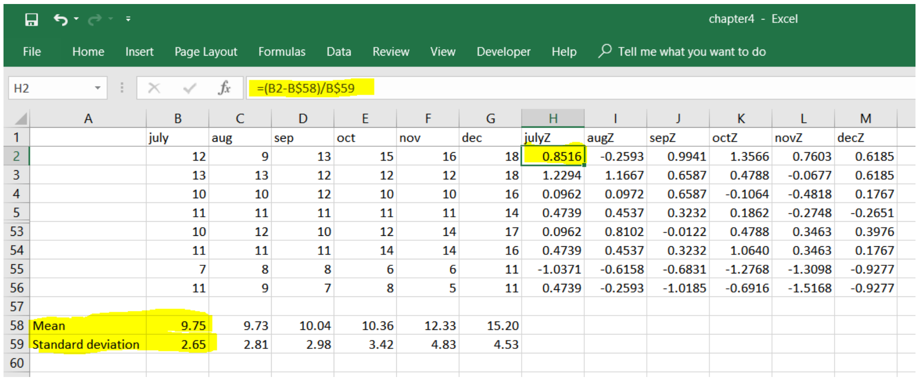


Figure 4.2 The formula for Z; and the Standard Normal Distribution

Standardized values can be easily calculated in Excel, as shown below. The data are in **chapter4.xlsx**.



=(B2-B$58)/B$59

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		july	aug	sep	oct	nov	dec	julyZ	augZ	sepZ	octZ	novZ	decZ
2		12	9	13	15	16	18	0.8516	-0.2593	0.9941	1.3566	0.7603	0.6185
3		13	13	12	12	12	18	1.2294	1.1667	0.6587	0.4788	-0.0677	0.6185
4		10	10	12	10	10	16	0.0962	0.0972	0.6587	-0.1064	-0.4818	0.1767
5		11	11	11	11	11	14	0.4739	0.4537	0.3232	0.1862	-0.2748	-0.2651
53		10	12	10	12	14	17	0.0962	0.8102	-0.0122	0.4788	0.3463	0.3976
54		11	11	11	14	14	16	0.4739	0.4537	0.3232	1.0640	0.3463	0.1767
55		7	8	8	6	6	11	-1.0371	-0.6158	-0.6831	-1.2768	-1.3098	-0.9277
56		11	9	7	8	5	11	0.4739	-0.2593	-1.0185	-0.6916	-1.5168	-0.9277
57													
58	Mean	9.75	9.73	10.04	10.36	12.33	15.20						
59	Standard deviation	2.65	2.81	2.98	3.42	4.83	4.53						
60													

Figure 4.3: Standardized values

The normalized score for employee 1 in July, can be computed as his/her score in July minus the mean score of all employees in July, divided by the standard deviation of scores in July.

$$JulyZ_1 = \frac{(12 - 9.75)}{2.65} = 0.85$$

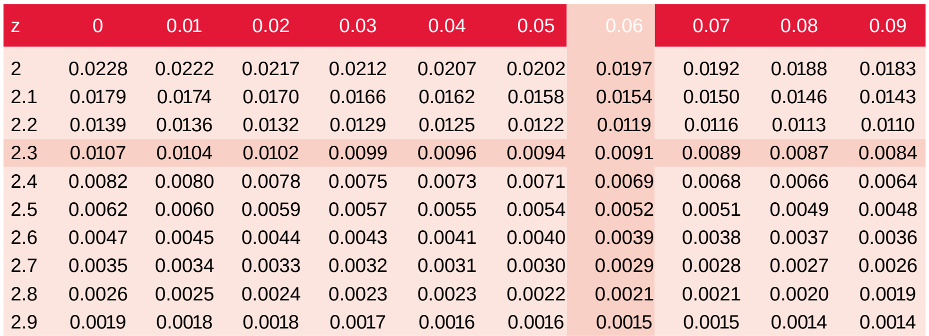
For record=1 we see that the Z-scores are always positive (except for August), signaling that this employee tends to perform above average.

Assuming that sales are normally distributed across employees, in each month, the value of Z informs us about how the position of employee among his peers: Z-scores close to zero are about average, while Z-scores far away from zero indicate very poor or very good performance. Absolute Z-scores of 2 or 3 are quite exceptional.

More precisely, you can calculate the probability of the Z-score *assuming that the distribution is normal*. You can (and should!) test the normality of the distribution (for example using the skewness and kurtosis test discussed in the previous chapter). We will come back to it when discussing *inferential* statistics.

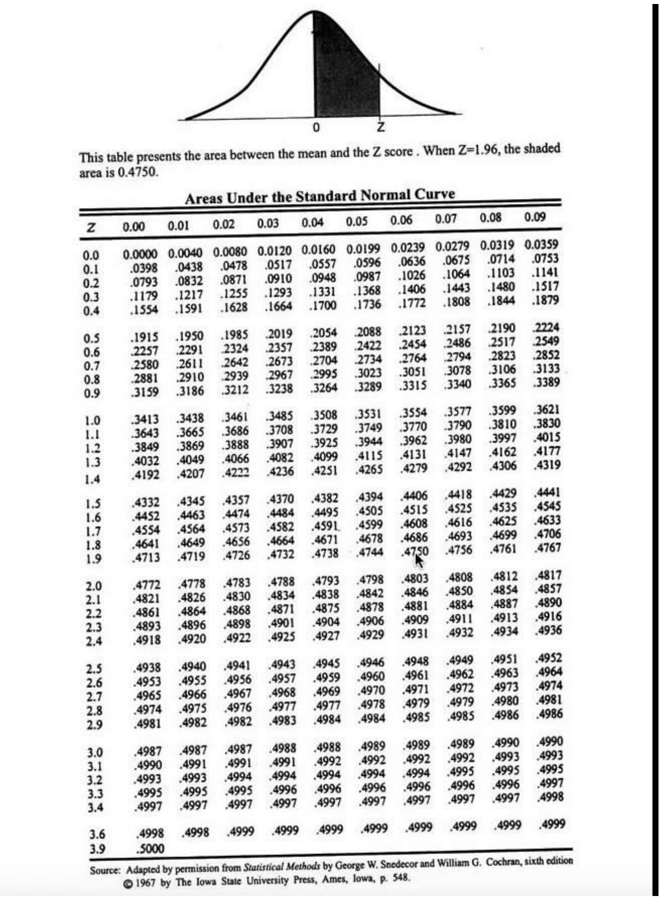
In the good old days, we had to look up the probabilities of Z-scores in tables like the one below. They were (and still are) provided as annexes in statistical texts.

> *Example: Suppose you have just used the formula to calculate that the performance of one of your employees in August, in terms of units sold, is Z=+2.36. From the table below, you conclude that the probability of a Z-score that high, is only 0.0091 (or 0.91%); your employee is among the best performers!*



z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014

Figure 4.4 The Z-table (Standard Normal Distribution)



This table presents the area between the mean and the Z score . When Z=1.96, the shaded area is 0.4750.

Areas Under the Standard Normal Curve

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000									

Source: Adapted by permission from *Statistical Methods* by George W. Snedecor and William G. Cochran, sixth edition © 1967 by The Iowa State University Press, Ames, Iowa, p. 548.

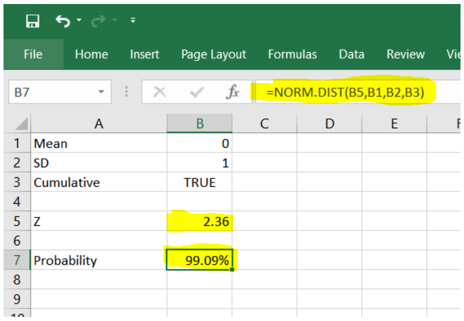
Figure 4.5: Areas under the normal curve and Z-score

The interpretation is that the probability of a Z-score exceeding 2.36 is equal to 0.0091, or 0.91%. You can deduce more information.

- Since probabilities always add up to 1 (100%), the probability of a Z-score less than 2.36 equals 1-0.0091 = 99.09%.
- Since the normal distribution is symmetrical, the probability of a Z-score less than -2.36 is also 0.91%.
- And the probability of an *absolute* Z-score higher than 2.36 (that is, Z<-2.36 or Z>+2.36), equals 2*0.91% = 1.82%.

With modern tools like Excel we don't need those tables. We can ask Excel for probabilities for a Z-score, using the **NORM.DIST()** function.

For Z=2.36, we find:



=NORM.DIST(B5,B1,B2,B3)

	A	B	C	D	E
1	Mean	0			
2	SD	1			
3	Cumulative	TRUE			
4					
5	Z	2.36			
6					
7	Probability	99.09%			
8					
9					

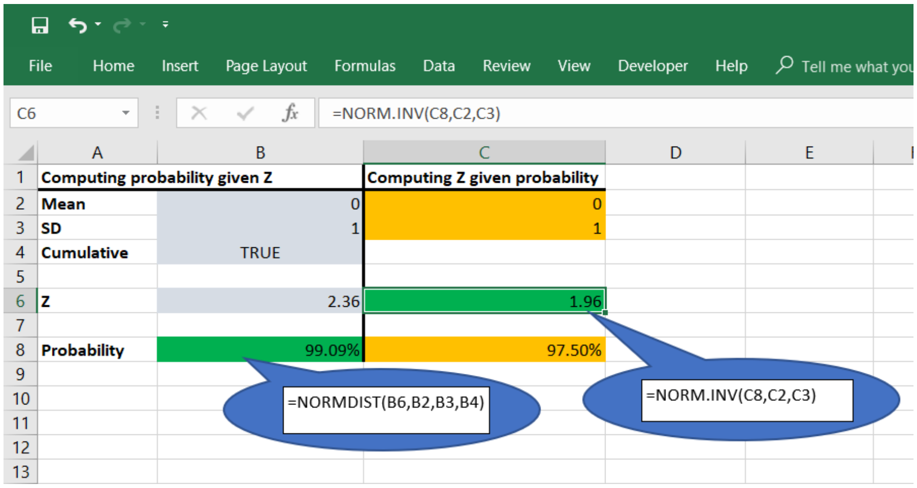
Figure 4.6: Probabilities of Z-scores Using Excel

NORM.DIST() returns the cumulative probability of Z-score. It is easiest to stick to cumulative probabilities (by using the value TRUE, for the last of four parameters between brackets), since – as we have seen earlier – a lot of other information can be simply deduced.

A related function is **NORM.INV()** which returns the Z-score related to a given cumulative probability.

For example, from **NORM.INV()** we can learn that 97.5% of the distribution has a Z-score higher than +1.96.

In an extended Excel worksheet:



C6

=NORM.INV(C8,C2,C3)

	A	B	C	D	E
1	Computing probability given Z		Computing Z given probability		
2	Mean	0	0		
3	SD	1	1		
4	Cumulative	TRUE			
5					
6	Z	2.36	1.96		
7					
8	Probability	99.09%	97.50%		
9					
10					
11					
12					
13					

=NORMDIST(B6,B2,B3,B4)

=NORM.INV(C8,C2,C3)

Figure 4.7: Probabilities and Z-scores Using Excel

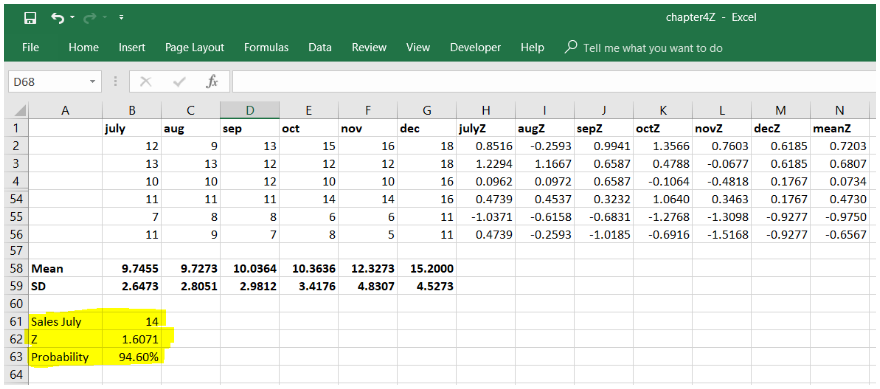
From this we can deduce that 2.5% is on the left of -1.96. And therefore, 5% has a Z-score outside of the range [-1.96; +1.96].

In a normal distribution, 5% of the observations have an absolute value of the Z-score higher than 1.96!

This is all quite handy. If a distribution is (or is assumed to be, or is approximately) normal, and we know the mean and the standard deviation, then we can make statements about the likelihood of an outcome.

For example, you can calculate the probability of an employee having July sales exceeding 14 units. You have to do it step by step.

- First compute the Z-score; for the Z-score you need the mean and the standard deviation for sales in July. We have used Excel to compute these statistics. The Z-score of sales of 14 in July, can be computed as 1.6071.



D68

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		july	aug	sep	oct	nov	dec	julyZ	augZ	sepZ	octZ	novZ	decZ	meanZ
2		12	9	13	15	16	18	0.8516	-0.2593	0.9941	1.3566	0.7603	0.6185	0.7203
3		13	13	12	12	12	18	1.2294	1.1667	0.6587	0.4788	-0.0677	0.6185	0.6807
4		10	10	12	10	10	16	0.0962	0.0972	0.6587	-0.1064	-0.4818	0.1767	0.0734
54		11	11	11	14	14	16	0.4739	0.4537	0.3232	1.0640	0.3463	0.1767	0.4730
55		7	8	8	6	6	11	-1.0371	-0.6158	-0.6831	-1.2768	-1.3098	-0.9277	-0.9750
56		11	9	7	8	5	11	0.4739	-0.2593	-1.0185	-0.6916	-1.5168	-0.9277	-0.6567
57														
58	Mean	9.7455	9.7273	10.0364	10.3636	12.3273	15.2000							
59	SD	2.6473	2.8051	2.9812	3.4176	4.8307	4.5273							
60														
61	Sales July	14												
62	Z	1.6071												
63	Probability	94.60%												
64														

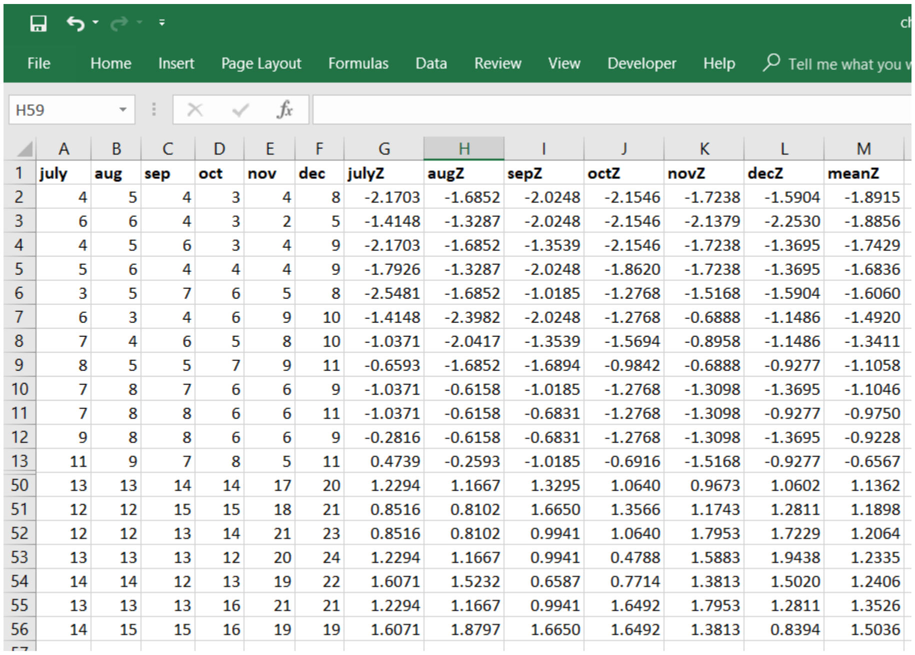
Figure 4.8: Example of Probability of a Z-score

- In conclusion, a value of 14 for sales in July is equivalent to a Z-score of 1.6071. Assuming a normal distribution, 94.60% of all employees would have sales up to 14. Only 5.40% of all employees would have sales of 14 or higher, which means that employees with this level of sales are performing very well.

Self-test

As a challenge, sort the employees is ascending order of their performance in the 6 July to December!

The actual scores and the related Z-scores are in **chapter4.xlsx**. We have saved the relevant data in **chapter4.csv**.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**july**	**aug**	**sep**	**oct**	**nov**	**dec**	**julyZ**	**augZ**	**sepZ**	**octZ**	**novZ**	**decZ**	**meanZ**
2	4	5	4	3	4	8	-2.1703	-1.6852	-2.0248	-2.1546	-1.7238	-1.5904	-1.8915
3	6	6	4	3	2	5	-1.4148	-1.3287	-2.0248	-2.1546	-2.1379	-2.2530	-1.8856
4	4	5	6	3	4	9	-2.1703	-1.6852	-1.3539	-2.1546	-1.7238	-1.3695	-1.7429
5	5	6	4	4	4	9	-1.7926	-1.3287	-2.0248	-1.8620	-1.7238	-1.3695	-1.6836
6	3	5	7	6	5	8	-2.5481	-1.6852	-1.0185	-1.2768	-1.5168	-1.5904	-1.6060
7	6	3	4	6	9	10	-1.4148	-2.3982	-2.0248	-1.2768	-0.6888	-1.1486	-1.4920
8	7	4	6	5	8	10	-1.0371	-2.0417	-1.3539	-1.5694	-0.8958	-1.1486	-1.3411
9	8	5	5	7	9	11	-0.6593	-1.6852	-1.6894	-0.9842	-0.6888	-0.9277	-1.1058
10	7	8	7	6	6	9	-1.0371	-0.6158	-1.0185	-1.2768	-1.3098	-1.3695	-1.1046
11	7	8	8	6	6	11	-1.0371	-0.6158	-0.6831	-1.2768	-1.3098	-0.9277	-0.9750
12	9	8	8	6	6	9	-0.2816	-0.6158	-0.6831	-1.2768	-1.3098	-1.3695	-0.9228
13	11	9	7	8	5	11	0.4739	-0.2593	-1.0185	-0.6916	-1.5168	-0.9277	-0.6567
50	13	13	14	14	17	20	1.2294	1.1667	1.3295	1.0640	0.9673	1.0602	1.1362
51	12	12	15	15	18	21	0.8516	0.8102	1.6650	1.3566	1.1743	1.2811	1.1898
52	12	12	13	14	21	23	0.8516	0.8102	0.9941	1.0640	1.7953	1.7229	1.2064
53	13	13	13	12	20	24	1.2294	1.1667	0.9941	0.4788	1.5883	1.9438	1.2335
54	14	14	12	13	19	22	1.6071	1.5232	0.6587	0.7714	1.3813	1.5020	1.2406
55	13	13	13	16	21	21	1.2294	1.1667	0.9941	1.6492	1.7953	1.2811	1.3526
56	14	15	15	16	19	19	1.6071	1.8797	1.6650	1.6492	1.3813	0.8394	1.5036

Figure 4.9: Sorted Z-scores

The employees are sorted from weak to strong. Without employee names or IDs the information is not helpful. But presuming the names are known, the HR manager can decide to give additional training to the weakest performers. If there's budget to train the weakest 10% and the distribution is approximately normal, then the cut-off point would be at Z-scores of -1.28 (check this yourself!).

4.2 Data Skill Challenge 3

Given this data set: 5, 5, 7, 2, 3, 4, 4

a. Convert these values to Z-scores

b. What proportion of cases would you expect to fall above 4?

c. What score would be at the 75th percentile?

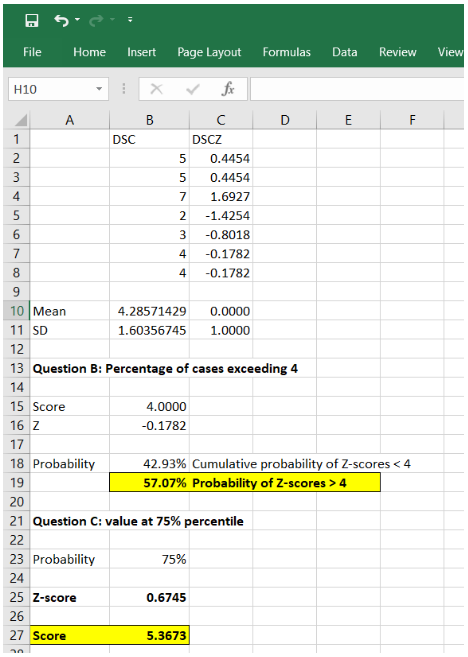
This exercise can be completed in Excel.

First we key in the data, and use Excel functions to compute the mean and the standard deviation. With these statistics, we can add a column with Z-scores. Note that this Z-transformation leads to a variable with mean 0 and standard deviation 1.

For Question B, we compute the Z-score of a score of 4, as -0.1782. We then use the **NORM.DIST()** function to compute the cumulative probability of that Z-score. This probability is the probability that of a Z-score of up to -0.1782. The probability of a Z-score higher than -0.1782 the is the complement: 100% - 42.93% = 57.07%.

For question C we use the **NORM.INV()** function. The probability is now given as 75%, and the related Z-score is 0.6745. We use this Z-score to compute the raw score, as:

$$RawScore = mean + Z * SD = 4.2857 + 0.6745 * 1.6036 = 5.3673$$



	A	B	C	D	E	F
1		DSC	DSCZ			
2		5	0.4454			
3		5	0.4454			
4		7	1.6927			
5		2	-1.4254			
6		3	-0.8018			
7		4	-0.1782			
8		4	-0.1782			
9						
10	Mean	4.28571429	0.0000			
11	SD	1.60356745	1.0000			
12						
13	**Question B: Percentage of cases exceeding 4**					
14						
15	Score	4.0000				
16	Z	-0.1782				
17						
18	Probability	42.93%	Cumulative probability of Z-scores < 4			
19		**57.07%**	**Probability of Z-scores > 4**			
20						
21	**Question C: value at 75% percentile**					
22						
23	Probability	75%				
24						
25	**Z-score**	**0.6745**				
26						
27	**Score**	**5.3673**				

This is not easy, admittedly. But getting acquainted with these fundamentals of statistics will make it easy for you to understand and interpret the statistical output that comes with the techniques that we will discuss!

5 Sampling Distributions

<u>Files needed:</u>
chapter5.xlsx
chapter5_dsc.csv

5.1 How to Draw Samples?

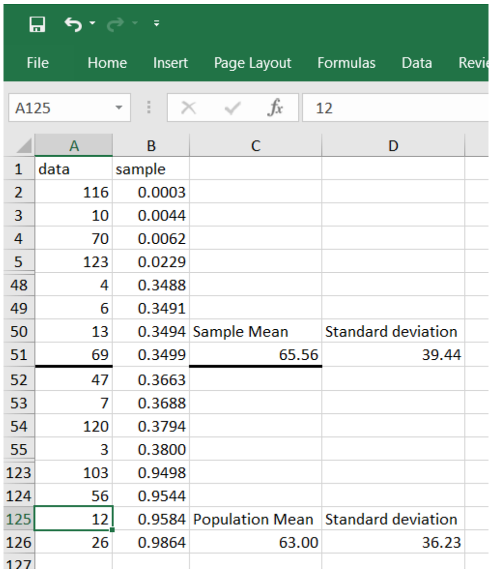
Computers make it an easy task to sample records from your data set. Below we will use a fictitious data set of 125 employees, and take a random sample of 50 of them.

In your data file listing the population or sampling frame, you can add a column with a random number. You can then sort the data on that column (from low to high, or high too low) and use the first 50 records of the sorted file as your sample.

Here, cases 116; 10; 70; etcetera have been sampled.

> *Note that we have hidden some of the rows for brevity; you can unhide them by highlighting rows 5 and 48, and then right-click your mouse to find the unhide option.*

Since the records are numbered from 1 to 125, the mean record number in the population is (1+125)/2 = 63. The sample mean is quite close. Every time you randomly draw a new sample, the sample means will be different, but quite close to the population mean.



A125 | fx | 12

	A	B	C	D
1	data	sample		
2	116	0.0003		
3	10	0.0044		
4	70	0.0062		
5	123	0.0229		
48	4	0.3488		
49	6	0.3491		
50	13	0.3494	Sample Mean	Standard deviation
51	69	0.3499	65.56	39.44
52	47	0.3663		
53	7	0.3688		
54	120	0.3794		
55	3	0.3800		
123	103	0.9498		
124	56	0.9544		
125	12	0.9584	Population Mean	Standard deviation
126	26	0.9864	63.00	36.23
127				

Figure 5.1 Sampling

One reason for this type of exercise would be to check if the sample looks similar to the population from which the sample was drawn, on key characteristics like gender or age.

5.2 Data Skill Challenge

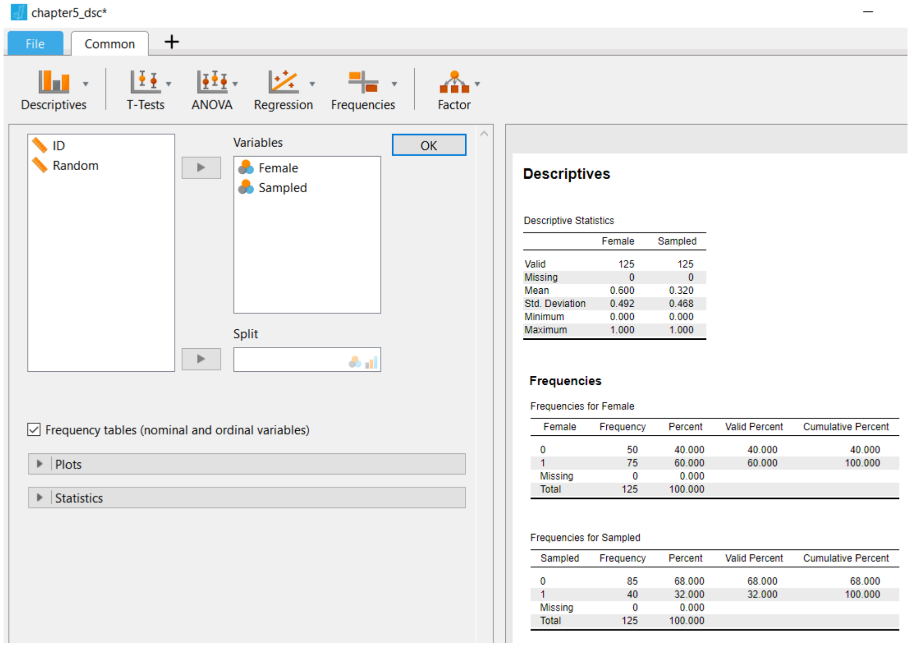
Let's extend the example of this chapter. In addition to the IDs of 125 employees, we now also have information on the employee's gender. In <u>chapter5_dsc.csv</u>, there's a second column for the variable female; this variable is a so-called *dummy* variable with values 0 or 1. It is good practice to name the variable after what code=1 stands for, that is, 1=female (and 0=male).

The third column contains a random number between 1 and 10,000, generated by Excel. We have sorted the data on random number from smallest to largest, and selected the first 40 records for our sample.

We want to check if the sample of 40 is representative in terms of gender. This means that the proportion of females is our sample is the same as (or not significantly different from) the proportion of females in the population.

Try this for yourself, using the tools discussed above.

In the data set we have the variables **female** and **sampled**. The variable **sampled** is coded “y” for the 40 records with the smallest random numbers (in var **sample**). From the frequency table we see that the data set contains 75 female respondents (**female = 1**), and 40 records which are sampled. Our hope and expectation are that the sample is representative of the population, in terms of gender. That is, in the sample we would expect around 6 out of 10 persons to be female.



☑ Frequency tables (nominal and ordinal variables)

Descriptives

Descriptive Statistics

	Female	Sampled
Valid	125	125
Missing	0	0
Mean	0.600	0.320
Std. Deviation	0.492	0.468
Minimum	0.000	0.000
Maximum	1.000	1.000

Frequencies

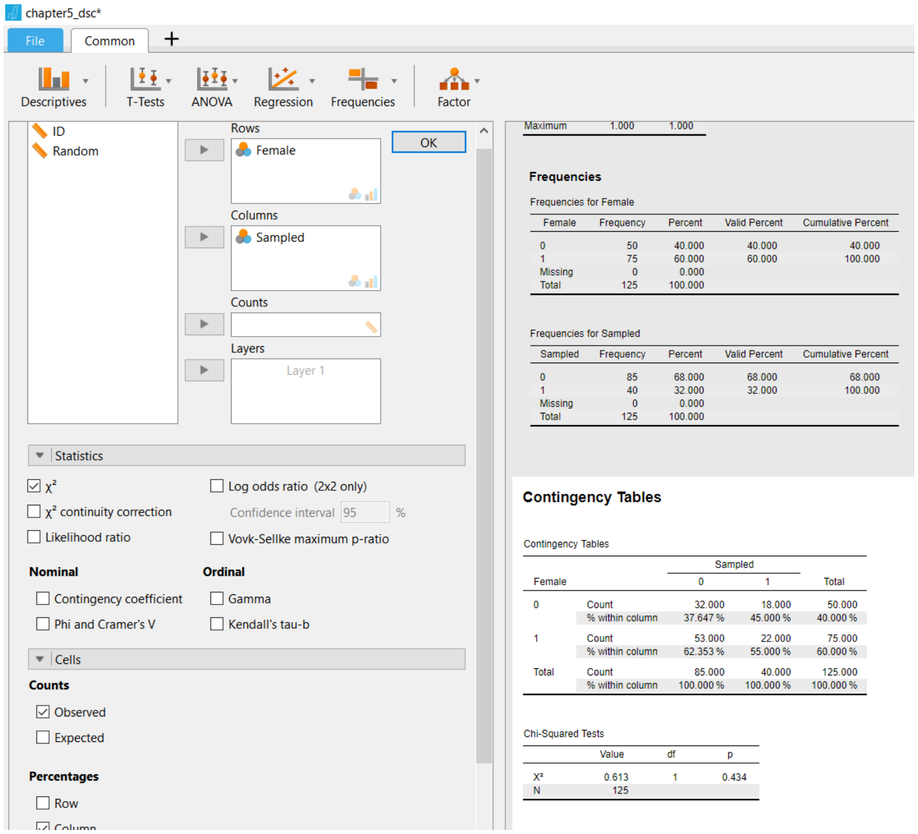
Frequencies for Female

Female	Frequency	Percent	Valid Percent	Cumulative Percent
0	50	40.000	40.000	40.000
1	75	60.000	60.000	100.000
Missing	0	0.000		
Total	125	100.000		

Frequencies for Sampled

Sampled	Frequency	Percent	Valid Percent	Cumulative Percent
0	85	68.000	68.000	68.000
1	40	32.000	32.000	100.000
Missing	0	0.000		
Total	125	100.000		

We can check this using **<Frequencies><Contingency Tables>**. From the output below we learn that 55% of the sampled persons are female, which is slightly less than expected. We can use a statistical test to verify that random sampling has not led to a sample that is significantly different from the population in terms of gender. The formal test is a chi-square test (to be discussed in detail in chapter 11). If the probability of this test statistic is less than 5% then we would start doubting the sampling procedure. Here, however, the probability is 43,4%, implying that repeated sampling of 40 out of 125 using this procedure would produce a test statistic this high in over 40% of the cases. Nothing to worry about!



Rows: Female

Columns: Sampled

Counts

Layers

Statistics

☑ χ²

☐ χ² continuity correction

☐ Likelihood ratio

☐ Log odds ratio (2x2 only)

Confidence interval 95 %

☐ Vovk-Sellke maximum p-ratio

Nominal

☐ Contingency coefficient

☐ Phi and Cramer's V

Ordinal

☐ Gamma

☐ Kendall's tau-b

Cells

Counts

☑ Observed

☐ Expected

Percentages

☐ Row

☑ Column

Maximum	1.000	1.000

Frequencies

Frequencies for Female

Female	Frequency	Percent	Valid Percent	Cumulative Percent
0	50	40.000	40.000	40.000
1	75	60.000	60.000	100.000
Missing	0	0.000		
Total	125	100.000		

Frequencies for Sampled

Sampled	Frequency	Percent	Valid Percent	Cumulative Percent
0	85	68.000	68.000	68.000
1	40	32.000	32.000	100.000
Missing	0	0.000		
Total	125	100.000		

Contingency Tables

Contingency Tables

		Sampled		
Female		0	1	Total
0	Count	32.000	18.000	50.000
	% within column	37.647 %	45.000 %	40.000 %
1	Count	53.000	22.000	75.000
	% within column	62.353 %	55.000 %	60.000 %
Total	Count	85.000	40.000	125.000
	% within column	100.000 %	100.000 %	100.000 %

Chi-Squared Tests

	Value	df	p
Χ²	0.613	1	0.434
N	125		

6 Estimation and Confidence Intervals

Files needed:
chapter6.txt
chapter6_example.xlsx

6.1 Finding Confidence Intervals

In the previous chapter we have seen that samples can provide accurate information. But how accurate? Even in relatively large samples you can be quite unlucky, and find sample statistics, like the mean, that are way off the mark. That's bad. But the point is, we don't know. The reason for taking a sample is to learn about the population mean, rather than to verify what we already know!

It is advisable to report your findings not as simple *point estimates*, but to provide additional information about the error margins of your estimate. We use *confidence intervals* to make statements like: in repeated sampling, in the majority of cases (say, 95% of the cases) the sample statistic will be in this interval. The smaller the interval, the more accurate your estimate. The width of the interval depends on the sample size, and on the variance of the variable in the population.

> *Example: Peter is interested in identifying the confidence interval (CI) surrounding daily production at his 12 plants. He recently received a report from HQ that production worldwide at each company plant varies by 2,000 (standard deviation).*
>
> *This is how you would calculate the CIs by hand (LB and UB are the lower and upper bounds, respectively). The so-called standard error ($\sigma_{\bar{x}}$) is the accuracy of the sample mean, as an estimator of the population mean. The larger the sample size (n), the more accurate the estimation.*

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2000}{\sqrt{12}} = \frac{2000}{3.464102} = 577.350205$$

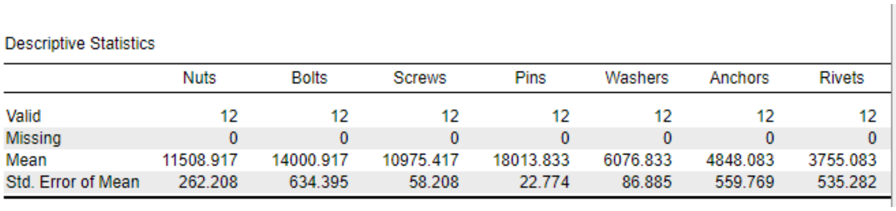
$$\begin{aligned}
CI_{95LB} &= \bar{x} - z\sigma_{\bar{x}} \\
&= 14000 - 1.96(577.350205) \\
&= 14000 - 1131.606402 \\
&= 12868.393598 \\
&= 12868.39
\end{aligned}$$

$$\begin{aligned}
CI_{95UB} &= \bar{x} + z\sigma_{\bar{x}} \\
&= 14000 + 1.96(577.350205) \\
&= 14000 + 1131.606402 \\
&= 15131.606402 \\
&= 15131.61
\end{aligned}$$

For computing the 95% confidence interval (or equivalently α=0.05) the Z-score to be applied is 1.96.

In most cases, the population variance is unknown. In those cases, we have to estimate the variance from the data in our sample. For small samples and unknown population variance, we use the t-distribution rather than the Z-distribution.

Applied to the data in **chapter6.txt** (note that the data are not "comma separated" but "tab delimited"; **JASP** can read that format as well), we want to estimate 95%-confidence intervals for the production of various items. In **Descriptives** we now ask for the standard error of the mean.



Descriptive Statistics

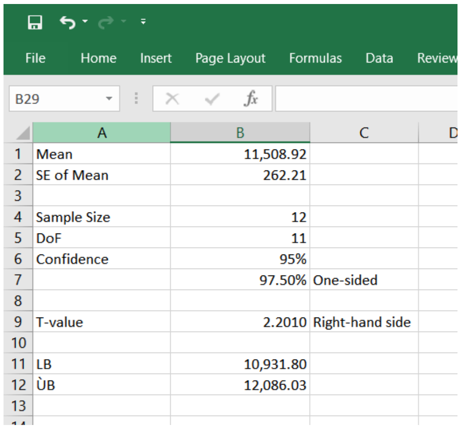
	Nuts	Bolts	Screws	Pins	Washers	Anchors	Rivets
Valid	12	12	12	12	12	12	12
Missing	0	0	0	0	0	0	0
Mean	11508.917	14000.917	10975.417	18013.833	6076.833	4848.083	3755.083
Std. Error of Mean	262.208	634.395	58.208	22.774	86.885	559.769	535.282

Figure 6.1: Standard Error of Mean, in JASP

The confidence interval for the production of nuts, is equal to

$$CI_{nuts} = Mean \pm t_{(95\%;11)} * SE_{Mean}$$

The value of $t_{(95\%;\ 11)}$ can be obtained from the **T.INV()** function in Excel. The 11 is the so-called number of degrees of freedom, which is the sample size minus 1. The t-value for *t* is 2.2010. The data and formulas can be found in **chapter6_example.xlsx**.



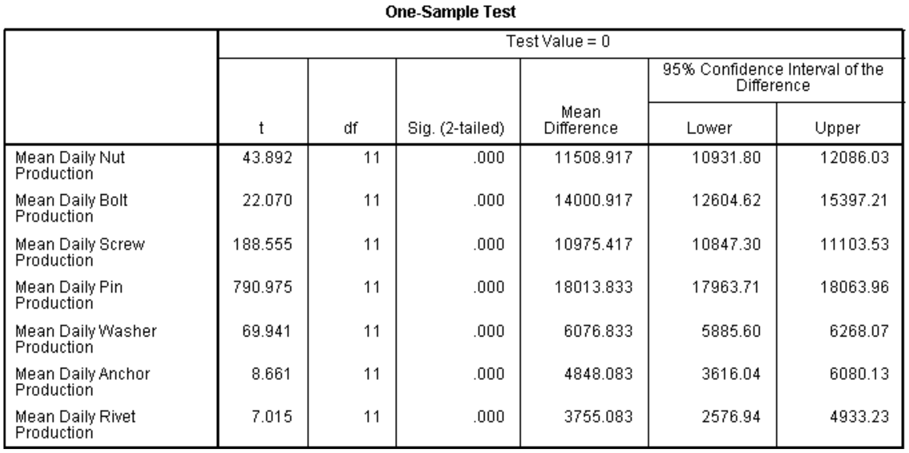
	A	B	C	D
1	Mean	11,508.92		
2	SE of Mean	262.21		
3				
4	Sample Size	12		
5	DoF	11		
6	Confidence	95%		
7		97.50%	One-sided	
8				
9	T-value	2.2010	Right-hand side	
10				
11	LB	10,931.80		
12	ÙB	12,086.03		
13				

Figure 6.2: Obtaining the t-value, in Excel

The confidence interval then can be computed as 11,508.92 ± 2.2010*262.21, or [10,931.80; 12,086.03].

This is equivalent to the results in Landers, who uses SPSS to do the same.

Verify for yourself the confidence intervals for the production of the other items.



One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Mean Daily Nut Production	43.892	11	.000	11508.917	10931.80	12086.03
Mean Daily Bolt Production	22.070	11	.000	14000.917	12604.62	15397.21
Mean Daily Screw Production	188.555	11	.000	10975.417	10847.30	11103.53
Mean Daily Pin Production	790.975	11	.000	18013.833	17963.71	18063.96
Mean Daily Washer Production	69.941	11	.000	6076.833	5885.60	6268.07
Mean Daily Anchor Production	8.661	11	.000	4848.083	3616.04	6080.13
Mean Daily Rivet Production	7.015	11	.000	3755.083	2576.94	4933.23

Figure 6.3: Confidence Intervals (Landers, using SPSS)

6.2 Data Skill Challenges

Test yourself:

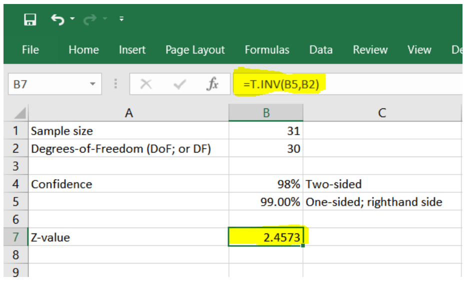
Determine the critical Z-value for the following situation:

(a) 98% confidence; σ unknown; n=31

You could use the statistical tables annexed in many traditional statistical textbooks. But the functions in Excel are easier and more precise. A challenge is to be familiar with these functions!

In case the population standard deviation is not known, we use the sample standard deviation to estimate the population standard deviation and we switch to the t-distribution. We use Excel's **T.INV()** function, and determine the degrees of freedom, which is 30 (31 minus 1). We conclude that the critical t-value is 2.4573 (compare Landers: 392).

Note that in an 98% confidence interval, we are looking for the t-values corresponding to the lower and upper 1% of the distribution. The function (with 99%) finds the t-value for the upper 1%: 1% of the distribution has a t-value beyond 2.4573. Since – like the normal distribution – the t-distribution is symmetrical, we can deduce that another 1% has a t-value lower than -2,4573. In total then, 98% lies in the interval from t=-2.4573 to t=+2.4573.



B7 =T.INV(B5,B2)

	A	B	C
1	Sample size	31	
2	Degrees-of-Freedom (DoF; or DF)	30	
3			
4	Confidence	98%	Two-sided
5		99.00%	One-sided; righthand side
6			
7	Z-value	2.4573	
8			
9			

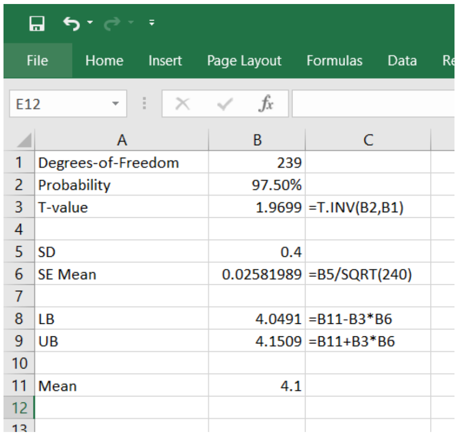
Indeed, this is not easy to remember, if you don't use it a lot. Even experienced researchers have to think anew themselves every time they use these functions.

Let's look at a practical example.

When looking through a report in Forbes you read about a survey of 240 CFOs, finding that on a 5-point scale CFOs report that their organization's financial health is 4.1 with a standard deviation of 0.4.

New/Advanced: What's the margin of error for this mean?

The standard error of the mean, can be computed as the standard deviation divided by the square root of the sample size (see cell B6). The t-value for a 95% confidence interval are computed as before. Don’t forget that the t-value is for the right-tail, so we have to plug in 97.5%; the other 2.5% are at the left tail. Lastly, we can compute the lower and upper bounds of the interval (file **chapter6_example.xlsx**).



	A	B	C
1	Degrees-of-Freedom	239	
2	Probability	97.50%	
3	T-value	1.9699	=T.INV(B2,B1)
4			
5	SD	0.4	
6	SE Mean	0.02581989	=B5/SQRT(240)
7			
8	LB	4.0491	=B11-B3*B6
9	UB	4.1509	=B11+B3*B6
10			
11	Mean	4.1	
12			
13			

The *margin of error* is not a standard statistic, so we have to do some simple additional calculations to get it. The margin of error is defined as half of the width of the confidence interval, divided by the mean. You can use a traditional calculator, or use Excel: 1.9699*0.02582 / 4.1 = 1.24%. Our estimate is pretty accurate.

> *We can now conclude that repeated samples of size 240 from the population of CFOs, in 95 out of 100 cases will produce a sample mean in the interval 4.049 to 4.151. Some researchers would say that the probability that the population mean is in the stated interval, is 95% (.095) which is subtly different from our correct interpretation!*

7/8 Hypothesis Testing & Z-tests; One-Sample T-tests

<u>Files needed:</u>
chapter8_1.xlsx
chapter8_1.csv
chapter4.xlsx

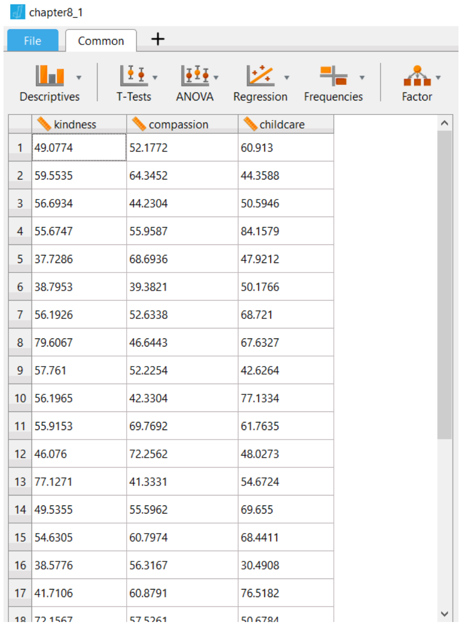
8.1 The One-sample Z-test

We combine chapters 7 and 8 of Landers since they are strongly related. Actually, both chapters are related to chapter 6, on confidence intervals, as well.

When using confidence intervals, the question we ask ourselves is, suppose we keep on sampling repeatedly, in which interval would our sampling statistic fall in 95% of all cases? For example, from an opinion poll on political preferences we may conclude that an estimated 60% of the electorate would vote for candidate A (and 40% for the only alternative, candidate B), with a 95% confidence interval of, say, 55% to 65%.

In hypothesis testing we work the other way around. Here, we test the sample outcome against some *null hypothesis*. If the null hypothesis would be that there is no overall preference for either candidate (both get 50% of the votes), then we would reject that hypothesis at a significance level of 5%, based on our sample data (since the hypothesized value 50% is well outside of our 95% confidence interval).

The data set that we will use for illustrating the concepts, is **chapter8_1.csv**.



	kindness	compassion	childcare
1	49.0774	52.1772	60.913
2	59.5535	64.3452	44.3588
3	56.6934	44.2304	50.5946
4	55.6747	55.9587	84.1579
5	37.7286	68.6936	47.9212
6	38.7953	39.3821	50.1766
7	56.1926	52.6338	68.721
8	79.6067	46.6443	67.6327
9	57.761	52.2254	42.6264
10	56.1965	42.3304	77.1334
11	55.9153	69.7692	61.7635
12	46.076	72.2562	48.0273
13	77.1271	41.3331	54.6724
14	49.5355	55.5962	69.655
15	54.6305	60.7974	68.4411
16	38.5776	56.3167	30.4908
17	41.7106	60.8791	76.5182
18	72.1567	57.5261	50.6784

Figure 7&8.1: The Data for Chapter 7/8

The description of the variables can be found in Landers, pages 187 and 188. In short: the data measure three traits of the 27 tutors in a company that provides tutoring services.

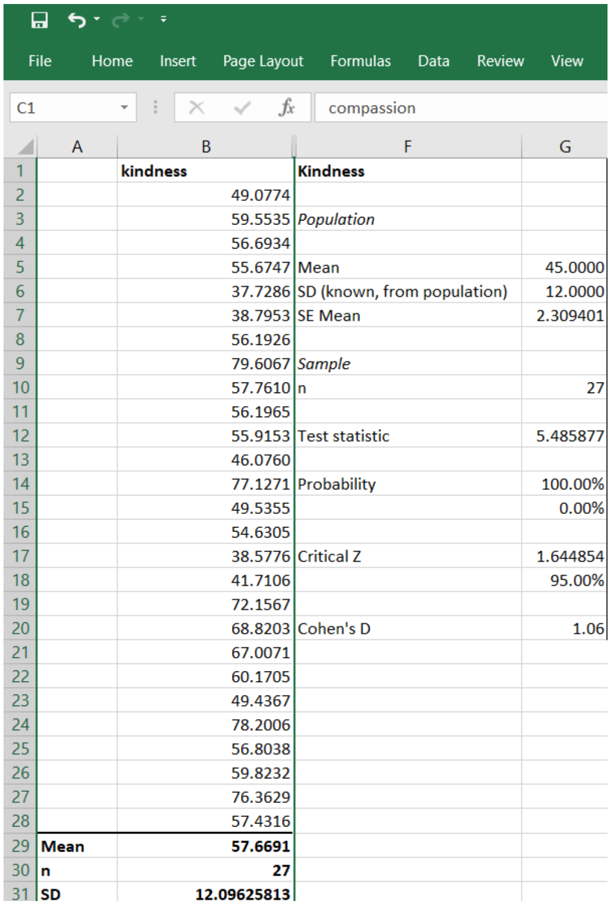
- **Kindness** data are based on the MacMillen Kindness Inventory;
- **Compassion** is measured by the Cincinnati Index of Compassion; and
- **Childcare** reflects the scores of the Child Focus Survey.

All tests produce scores on a scale from 0 to 100.

- The national average for the MacMillen Kindness Inventory is 45, with a standard deviation of 12 (based on a sample of 240,000 people). The standard deviation can be considered the standard deviation in the population;
- The Compassion Index and the Childcare index have averages of 55 and 67, respectively, based on very large samples; the population standard deviations are not known.

We want to test the hypotheses that the sample scores for our 27 tutors are higher than the national (population) averages.

For **kindness** we can use a one sample Z-test, since the standard deviation in the population is known. The best way is to use Excel as a statistical calculator, and follow the steps in Landers (page 195-197).



	A	B	F	G
1		**kindness**	**Kindness**	
2		49.0774		
3		59.5535	*Population*	
4		56.6934		
5		55.6747	Mean	45.0000
6		37.7286	SD (known, from population)	12.0000
7		38.7953	SE Mean	2.309401
8		56.1926		
9		79.6067	*Sample*	
10		57.7610	n	27
11		56.1965		
12		55.9153	Test statistic	5.485877
13		46.0760		
14		77.1271	Probability	100.00%
15		49.5355		0.00%
16		54.6305		
17		38.5776	Critical Z	1.644854
18		41.7106		95.00%
19		72.1567		
20		68.8203	Cohen's D	1.06
21		67.0071		
22		60.1705		
23		49.4367		
24		78.2006		
25		56.8038		
26		59.8232		
27		76.3629		
28		57.4316		
29	**Mean**	**57.6691**		
30	**n**	**27**		
31	**SD**	**12.09625813**		

Figure 7&8.2: Z-test for Kindness

In the above figure we have computed the mean for **kindness** as 57.67 (rounded to two decimals) using the **AVERAGE()** function in Excel. The sample mean of 57.67 is well above the mean of the population (45).

Since the standard deviation of the population is known, we can directly compute the standard error of the mean, as 12 divided by the square root of the sample size (square root of 27), which gives 2.31 in cell G7.

$$SE_{Mean} = \frac{Standard\ Deviation}{\sqrt{Sample\ Size}} = \frac{12}{\sqrt{27}} = 2.3094$$

The Z-value of our finding is therefore (57.67-45)/2.31 = 5.49. The cumulative probability of a Z-value of 5.49 is close to 100%; the probability of Z-value of 5.49 or higher is therefore effectively zero. We reject the hypothesis that our sample mean is equal to the population mean; our sample outcome is significantly higher than the population mean.

Compare these results to page 197 in Landers!

We conclude that the (null) hypothesis that our sample mean is the same as the population mean is firmly rejected. The Z-value is well above the critical value of 1.96 (for two-sided testing); since we are testing whether the scores for our tutors are higher than for the population of tutors, a one-sided test is appropriate.

Our report would read:

H_0: $\mu \leq 45$
H_1: $\mu > 45$
$\alpha = .05$
$Z_{critical} = 1.645$
$Z = 5.49$, $p<.05$

First, we state the null hypothesis and the alternative hypothesis. The "interesting" hypothesis is our alternative hypothesis that we want to test against the null hypothesis. Since we hypothesize that our tutors are better than average, we use one-sided testing. The advantage of one-sided testing is that – by ignoring the left part of the distribution – we have a higher chance of detecting a significant difference!

Our significance level is α=.05 which corresponds to 95% confidence. The alpha (α) level is the probability of rejecting the null hypothesis when in fact the null hypothesis is true. We want this probability to be very small. The value of .05 (5%) is commonly used, but you are free to be more or less strict. If you want to be more certain that your conclusion of our tutors being kinder than average is true, you can raise your confidence level and lower α, as α is equivalent to 1 minus the confidence level.

A shorter version of our conclusion would read: the null hypothesis is rejected (Z=5.49; P<0.05).

Some would report the critical value of Z, that is, the value above which the null hypothesis is rejected. Since we use one-sided testing, we use the **NORM.INV(95%,0,1)** function in Excel, to get the Z-value below which 95% (and beyond which the other 5%) of the distribution lies. 5% of the distribution has a Z-value higher than Z=1.6449; the probability of an even higher Z-score, farther away from the mean, must be smaller than 5%. We therefore reject the null hypothesis.

Since computers have no difficulty in computing the exact probability of Z-score, it has become less common to report the critical value. The exact probability of Z-score of 5.4859 is, in two decimals, 0.00%, and we can report that.

To be complete, we can also report the effect size. For one sample tests, we can compute Cohen's *d*.

$$d = \frac{(\overleftarrow{x}-\mu)}{\sigma} = \frac{(57.669-45)}{12} = 1.06$$

Cohen's *d* can be interpreted as follows:

Absolute value of *d*	**Size of effect**
<0.2	Very small
0.2 – 0.5	Small
0.5 – 0.8	Medium
>.8	Large

Figure 7&8.3: Effect Sizes and Cohen's D

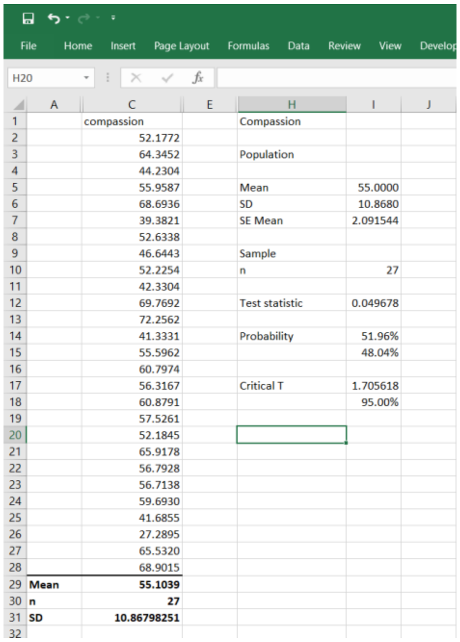
The effect therefore is large.

8.2 The One-Sample T-test

Below we test if the sample means of the variables **compassion** and **childcare** are significantly higher than their population means.

The sample mean (55.1039) for compassion is not much higher than the hypothesized mean of 55. We can therefore suspect that this difference is not significant.

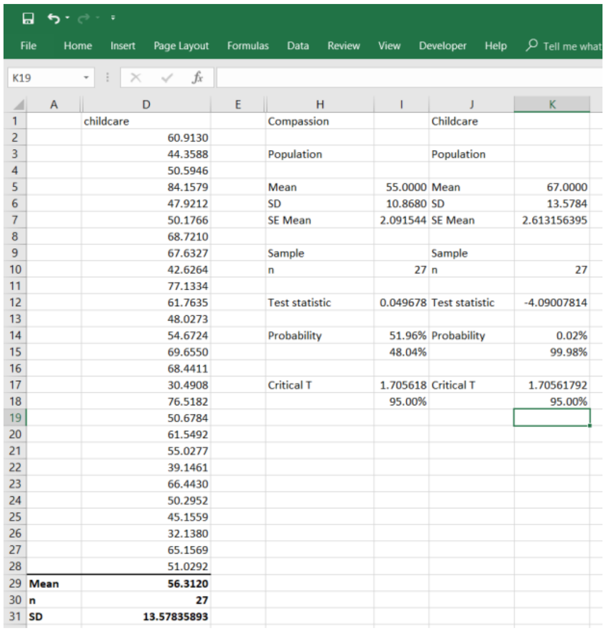
The t-value is .0497, and from the output we can conclude that, when drawing samples of size 27 from a normally distributed population, in no less than 48% of the cases we will have a t-value of .0497 or higher, in other words, it is not that remarkable. Statisticians – at least in social and economic studies – draw the line at 5%. If the probability of a t-value is 5% or less then we would start doubting the null hypothesis. But here, in one-sided testing, the probability is 48%, and therefore we accept the null hypothesis that the score on **compassion** in our example is smaller than or equal to score in the population.



	A	C	E	H	I	J
1		compassion		Compassion		
2		52.1772				
3		64.3452		Population		
4		44.2304				
5		55.9587		Mean	55.0000	
6		68.6936		SD	10.8680	
7		39.3821		SE Mean	2.091544	
8		52.6338				
9		46.6443		Sample		
10		52.2254		n	27	
11		42.3304				
12		69.7692		Test statistic	0.049678	
13		72.2562				
14		41.3331		Probability	51.96%	
15		55.5962			48.04%	
16		60.7974				
17		56.3167		Critical T	1.705618	
18		60.8791			95.00%	
19		57.5261				
20		52.1845				
21		65.9178				
22		56.7928				
23		56.7138				
24		59.6930				
25		41.6855				
26		27.2895				
27		65.5320				
28		68.9015				
29	**Mean**	**55.1039**				
30	**n**	**27**				
31	**SD**	**10.86798251**				
32						

Figure 7&8.4: T-test for Compassion

For **childcare** the procedure is the same.



	A	D	E	H	I	J	K
1		childcare		Compassion		Childcare	
2		60.9130					
3		44.3588		Population		Population	
4		50.5946					
5		84.1579		Mean	55.0000	Mean	67.0000
6		47.9212		SD	10.8680	SD	13.5784
7		50.1766		SE Mean	2.091544	SE Mean	2.613156395
8		68.7210					
9		67.6327		Sample		Sample	
10		42.6264		n	27	n	27
11		77.1334					
12		61.7635		Test statistic	0.049678	Test statistic	-4.09007814
13		48.0273					
14		54.6724		Probability	51.96%	Probability	0.02%
15		69.6550			48.04%		99.98%
16		68.4411					
17		30.4908		Critical T	1.705618	Critical T	1.70561792
18		76.5182			95.00%		95.00%
19		50.6784					
20		61.5492					
21		55.0277					
22		39.1461					
23		66.4430					
24		50.2952					
25		45.1559					
26		32.1380					
27		65.1569					
28		51.0292					
29	**Mean**	**56.3120**					
30	**n**	**27**					
31	**SD**	**13.57835893**					

Figure 7&8.5: T-test for Childcare

Here we have to be careful. The t-value is quite high (in *absolute* terms), namely -4.09. We might conclude that **childcare** in our sample is significantly different from the population, which is probably true.

However, we are testing, one-sidedly, the hypothesis that scores in our sample are higher than in the population. The null hypothesis is (see Landers: 195) that the mean value is smaller than or equal to 67. The alternative hypothesis is that the mean in our sample is higher than the national score. We don't find support for the alternative hypothesis (on the contrary, our sample score is well below the national average!). In almost all samples of this size, the t-value will be higher than -4.09. We accept the null hypothesis.

We can again compute Cohen's *d* for effect size. But this time, we do not know the standard deviation in the population. We can use the sample variance as an estimator of the population variance.

For example, for **compassion** we get the following.

$$d = \frac{(\bar{x} - \mu)}{\sigma} = \frac{(55.1039 - 55)}{10.8680} = 0.01$$

As expected, a very small effect size.

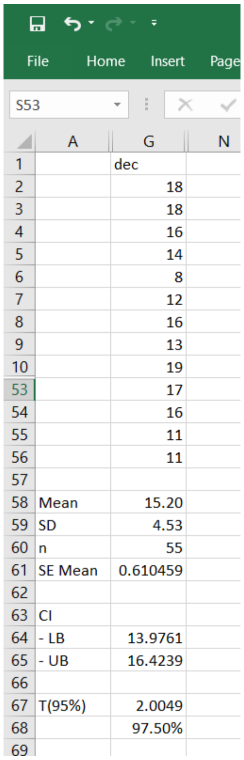
8.3 Data Skill Challenge

Data Skill Challenge 1, page 217

Jill, whose data on used car sales we encountered in chapter 4, has decided to compare her employees' December sales with average (mean) sales in her region, which she read from an online newspaper was 14. Conduct the complete hypothesis testing process with this data set!

Since we do not know the standard deviation in the population we have to use a T-test.

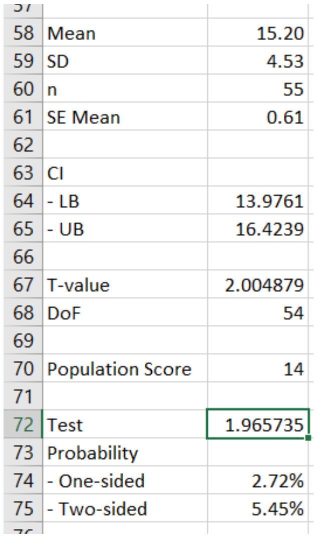
Let's first compute a confidence interval.



	A	G	N
1		dec	
2		18	
3		18	
4		16	
5		14	
6		8	
7		12	
8		16	
9		13	
10		19	
53		17	
54		16	
55		11	
56		11	
57			
58	Mean	15.20	
59	SD	4.53	
60	n	55	
61	SE Mean	0.610459	
62			
63	CI		
64	- LB	13.9761	
65	- UB	16.4239	
66			
67	T(95%)	2.0049	
68		97.50%	
69			

The 95% confidence interval [13.9761; 16.4239] just includes the hypothesized value (14) for December sales, and therefore we can conclude that the observed value is not significantly different from the hypothesized value. All data and formula are in **chapter4.xlsx**.

However, in hypothesis testing we work the other way around, and the formal procedure would use a T-test.



58	Mean	15.20
59	SD	4.53
60	n	55
61	SE Mean	0.61
62		
63	CI	
64	- LB	13.9761
65	- UB	16.4239
66		
67	T-value	2.004879
68	DoF	54
69		
70	Population Score	14
71		
72	Test	1.965735
73	Probability	
74	- One-sided	2.72%
75	- Two-sided	5.45%

Since the hypothesis is formulated in a neutral way (we compare our observations against a hypothesized value), we use a two-sided test. An *absolute* t-value of 1.97 (t<-1.97 or t>1.97) or higher, has a probability of 5.45%. This is just above the cut-off point of 5%, and therefore we accept the null hypothesis of no difference.

If the test was one-sided (the alternative hypothesis states that our December sales are higher than the population average), then we would reject the null hypothesis! In only 2.72% of all samples, will the sample mean be this much higher than the population mean of 14.

Since one-sided testing is more likely to result in a significant difference, many researchers prefer one-sided testing. However, there has to be a theoretical justification (for example, an intervention by Jill, like a training program to her salespersons in order to boost sales). Two-sided testing is preferred by those who are more conservative.

Data Skill Challenge 3, page 217/218

Larry runs a manufacturing company.

The most recent issue of a manufacturing trade magazine published an article stating that the average number of industrial accidents for plastics manufacturers is 15 per year.

Larry wants to know whether his company has fewer accidents than the average for the industry.

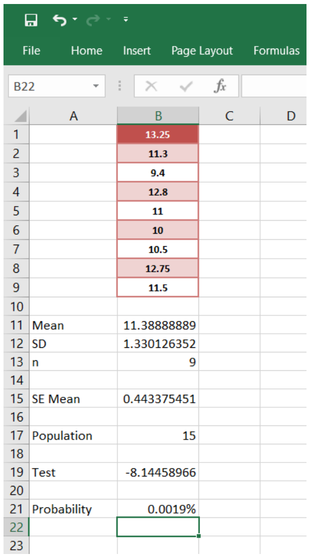
The data for Larry's company for the past nine years are presented in the table below.

Conduct the complete hypothesis testing process with this data set!

Year	# of Accidents
1	13.25
2	11.30
3	9.40
4	12.80
5	11.00
6	10.00
7	10.50
8	12.75
9	11.50

Figure 7&8.6: Data Skill Challenge 3

By now it should be easy to perform this test. Note that the test is left-sided: we test the hypothesis that accidents in Larry's plants are below the industry average of 15.



	A	B	C	D
1		13.25		
2		11.3		
3		9.4		
4		12.8		
5		11		
6		10		
7		10.5		
8		12.75		
9		11.5		
10				
11	Mean	11.38888889		
12	SD	1.330126352		
13	n	9		
14				
15	SE Mean	0.443375451		
16				
17	Population	15		
18				
19	Test	-8.14458966		
20				
21	Probability	0.0019%		
22				
23				

The output shows the sample mean (11.39), which is well below the hypothesized mean of 15.

The standard deviation of the population is not known, and estimated from the sample data. From the estimated standard deviation, we can compute the standard error of the mean, by dividing by the square root of the sample size.

The t-value then is the difference between the sample mean and the hypothesized mean, divided by the standard error of the mean. This gives a t-value of -8.14. The probability of such a highly negative t-value, is very low.

The difference is highly significant, using the left-sided test; the probability of a t-value as low as -8.14 is close to zero. We reject the null hypothesis, and find support for the alternative hypothesis that Larry's plant is doing well in terms of accidents as compared to the industry.

9 Paired and Independent Samples T-test

Files needed:
chapter9_1.csv
chapter9_2.csv
chapter9_dsc.txt

9.1 Introduction

Now that we have learned how to compare a sample result to a hypothesized value, a small next step is to compare groups.

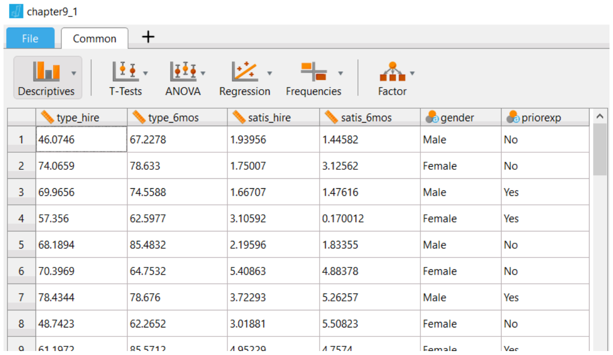
> *For example, we can compare the salaries of male and female employees in our organization and see if there's a statistically significant difference.*
>
> *Another type of comparison would be to compare the performance of the employees in our organization at two points in time, and see if there is a difference.*

For comparing two groups, or comparing one group at two points in time, we use the T-test. Once you understand the principle of the T-test, you can test more complex relationships using related techniques. More complex relationships are, for example, comparisons between three or more groups; over three or more periods in time; group comparisons over time; or add "explanatory" variables to the model.

But let's start with the simpler situation. When making comparisons between two groups, we have to ask ourselves whether the groups are **independent** (e.g. males versus females) or **dependent** (or **matched** or **paired**, like the performance of the same employees at two points in time).

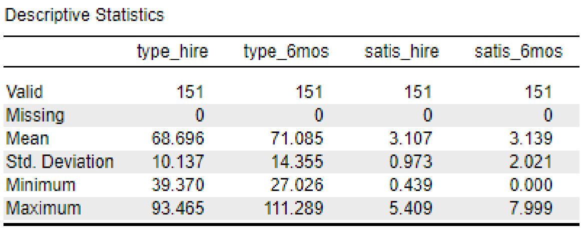
We will use the data set **chapter9_1.csv** to illustrate the T-tests. Below you find summary information on the data.

There are variables for typing skills when hired and after 6 months; and the same for satisfaction. All the information is available for men and women. Verify for yourself that there are 151 records in the data set. The typing skills for the complete sample of 151 men and women have gone up from 68.70 to 71.09 (by 2.39 that is) over the 6 months period.



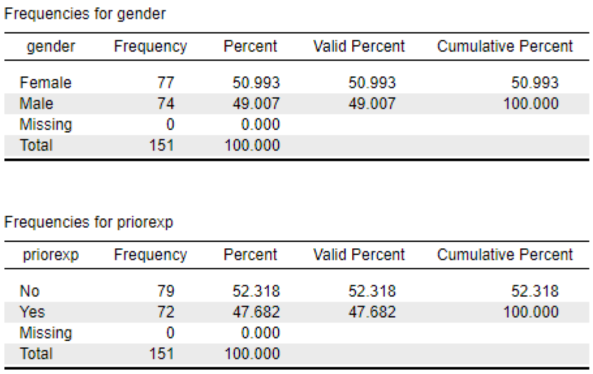
	type_hire	type_6mos	satis_hire	satis_6mos	gender	priorexp
1	46.0746	67.2278	1.93956	1.44582	Male	No
2	74.0659	78.633	1.75007	3.12562	Female	No
3	69.9656	74.5588	1.66707	1.47616	Male	Yes
4	57.356	62.5977	3.10592	0.170012	Female	Yes
5	68.1894	85.4832	2.19596	1.83355	Male	No
6	70.3969	64.7532	5.40863	4.88378	Female	No
7	78.4344	78.676	3.72293	5.26257	Male	Yes
8	48.7423	62.2652	3.01881	5.50823	Female	No
9	61.1972	85.5712	4.95229	4.7574	Female	Yes

Figure 9.1: Data for Chapter 9



Descriptive Statistics

	type_hire	type_6mos	satis_hire	satis_6mos
Valid	151	151	151	151
Missing	0	0	0	0
Mean	68.696	71.085	3.107	3.139
Std. Deviation	10.137	14.355	0.973	2.021
Minimum	39.370	27.026	0.439	0.000
Maximum	93.465	111.289	5.409	7.999



Frequencies for gender

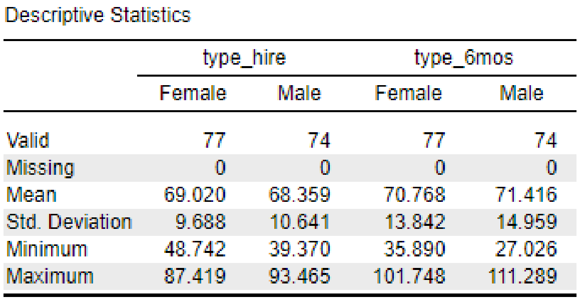
gender	Frequency	Percent	Valid Percent	Cumulative Percent
Female	77	50.993	50.993	50.993
Male	74	49.007	49.007	100.000
Missing	0	0.000		
Total	151	100.000		

Frequencies for priorexp

priorexp	Frequency	Percent	Valid Percent	Cumulative Percent
No	79	52.318	52.318	52.318
Yes	72	47.682	47.682	100.000
Missing	0	0.000		
Total	151	100.000		

Figure 9.2: Descriptive Statistics for Data in Chapter 9

It should become second nature to ask basic questions about your data set, and answer them using the basic commands in **JASP**. For example, you might be interested in the differences in typing skills of male and female employees.



Descriptive Statistics

	type_hire		type_6mos	
	Female	Male	Female	Male
Valid	77	74	77	74
Missing	0	0	0	0
Mean	69.020	68.359	70.768	71.416
Std. Deviation	9.688	10.641	13.842	14.959
Minimum	48.742	39.370	35.890	27.026
Maximum	87.419	93.465	101.748	111.289

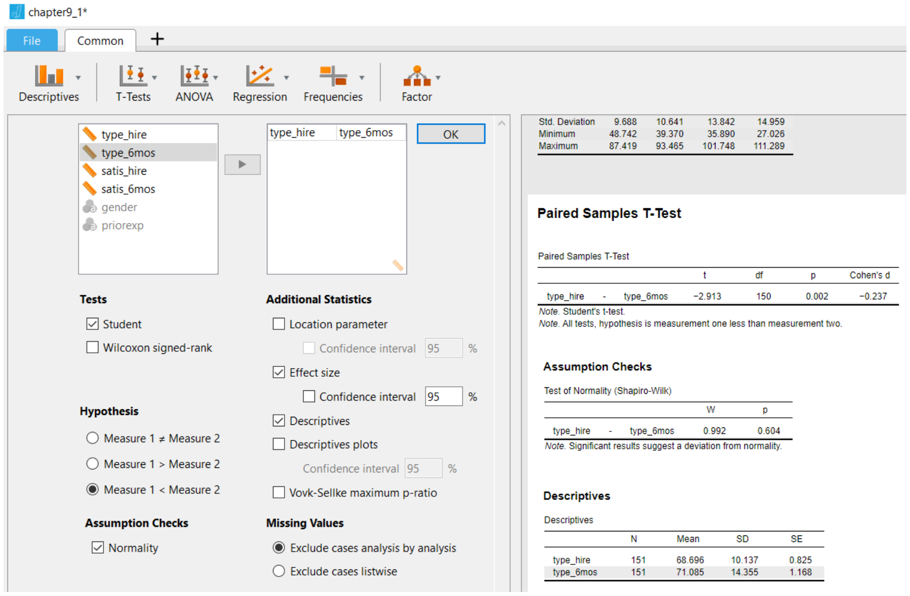
Figure 9.3: Breakdown of Typing Skills by Gender

From this simple overview we learn many things. First of all, we have 151 records for both years. There seem to be no missing data. The typing skills for male respondents were lower, when hired, but have increased faster than for female over the 6 months.

9.2 Paired Samples T-test

The paired samples T-test compares two variables in the data set, in this case **type_hire** and **type_6mos**. The two variables are matched: for each individual, the two variables measure the typing skills of that individual at two points in time.

The results of the test are shown below.



type_hire
type_6mos
satis_hire
satis_6mos
gender
priorexp

type_hire	type_6mos

Tests

☑ Student
☐ Wilcoxon signed-rank

Hypothesis

○ Measure 1 ≠ Measure 2
○ Measure 1 > Measure 2
● Measure 1 < Measure 2

Assumption Checks

☑ Normality

Additional Statistics

☐ Location parameter
☐ Confidence interval 95 %
☑ Effect size
☐ Confidence interval 95 %
☑ Descriptives
☐ Descriptives plots
Confidence interval 95 %
☐ Vovk-Sellke maximum p-ratio

Missing Values

● Exclude cases analysis by analysis
○ Exclude cases listwise

Std. Deviation	9.688	10.641	13.842	14.959
Minimum	48.742	39.370	35.890	27.026
Maximum	87.419	93.465	101.748	111.289

Paired Samples T-Test

Paired Samples T-Test

			t	df	p	Cohen's d
type_hire	-	type_6mos	−2.913	150	0.002	−0.237

Note. Student's t-test.
Note. All tests, hypothesis is measurement one less than measurement two.

Assumption Checks

Test of Normality (Shapiro-Wilk)

			W	p
type_hire	-	type_6mos	0.992	0.604

Note. Significant results suggest a deviation from normality.

Descriptives

Descriptives

	N	Mean	SD	SE
type_hire	151	68.696	10.137	0.825
type_6mos	151	71.085	14.355	1.168

Figure 9.4: Paired T-test of Typing Skills

The results of paired sample test are straightforward. The means are the same as in figure 9.3. The difference of -2.39 is due to the fact that we test the first measurement against the second one. We can do the opposite test, and will come to the very same conclusions – it's just the mirror image. Since skills are bound to improve, we would apply a one-sided test (H_0: typing skills when hired < typing skills after 6 months). The t-value of 2.91 has a probability of .002 (0.2%), and we reject the null hypothesis; skills have improved, looking at all 151 employees).

Self-test

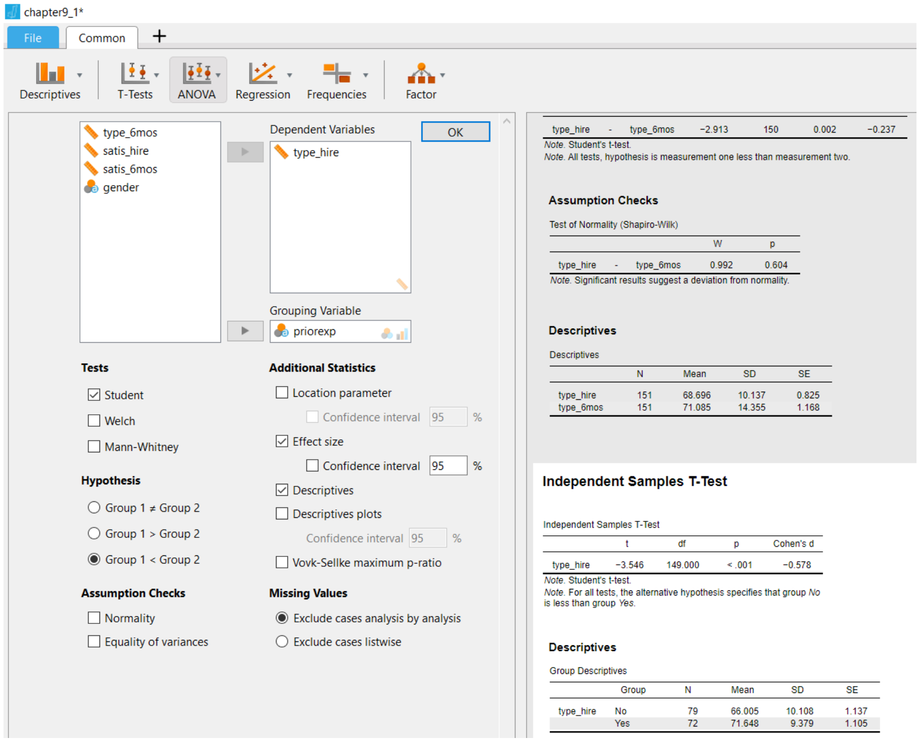
Check for yourself that you get similar results when first entering **type_6mos** followed by **type_hire**. Don't forget to change the hypothesis: Measure 1 is now hypothesized to be larger than Measure 2!

9.3 Independent Samples T-test

Let's move on to testing differences between independent rather than paired samples.

Making use of the same data set, we may be interested in the difference between male and female employees when it comes to typing skills, or the difference between employees with versus without prior experience in typing. For the latter test, it seems reasonable to hypothesize that people with prior experience have better typing skills especially at first ("when hired"); the differences may fade after 6 months of experience.

The results are shown below. Use **<Independent Samples T-Test>**, under the **<T-Tests>** tab.



type_6mos
satis_hire
satis_6mos
gender

Dependent Variables
type_hire

Grouping Variable
priorexp

Tests

☑ Student
☐ Welch
☐ Mann-Whitney

Hypothesis

○ Group 1 ≠ Group 2
○ Group 1 > Group 2
● Group 1 < Group 2

Assumption Checks

☐ Normality
☐ Equality of variances

Additional Statistics

☐ Location parameter
☐ Confidence interval 95 %
☑ Effect size
☐ Confidence interval 95 %
☑ Descriptives
☐ Descriptives plots
Confidence interval 95 %
☐ Vovk-Sellke maximum p-ratio

Missing Values

● Exclude cases analysis by analysis
○ Exclude cases listwise

type_hire	-	type_6mos	−2.913	150	0.002	−0.237

Note. Student's t-test.
Note. All tests, hypothesis is measurement one less than measurement two.

Assumption Checks

Test of Normality (Shapiro-Wilk)

			W	p
type_hire	-	type_6mos	0.992	0.604

Note. Significant results suggest a deviation from normality.

Descriptives

Descriptives

	N	Mean	SD	SE
type_hire	151	68.696	10.137	0.825
type_6mos	151	71.085	14.355	1.168

Independent Samples T-Test

Independent Samples T-Test

	t	df	p	Cohen's d
type_hire	−3.546	149.000	< .001	−0.578

Note. Student's t-test.
Note. For all tests, the alternative hypothesis specifies that group *No* is less than group *Yes*.

Descriptives

Group Descriptives

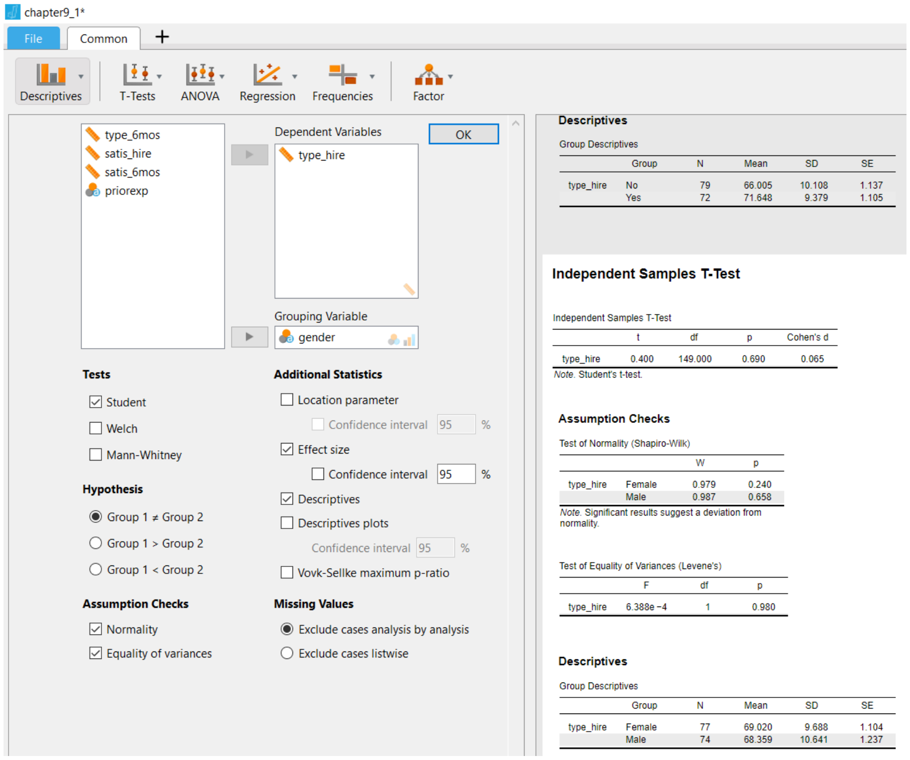
	Group	N	Mean	SD	SE
type_hire	No	79	66.005	10.108	1.137
	Yes	72	71.648	9.379	1.105

Figure 9.5: Independent Samples T-test of Typing Skills

Employees with prior experience have better typing skills at the moment of hiring (71.65 versus 66.01).

The difference in typing skills is highly significant. Since we have opted for one-sided testing, the significance to be reported is *p* < .001. **JASP** does not provide exact probabilities, when the probabilities are very small. But a probability of less than 0.1% is much smaller than our 5% benchmark.

For testing the difference between men and women, we do not have an *a priori* reason to hypothesize that one group is more skilled than the other, so we use a two-sided test. We tick Group 1 ≠ Group 2, as our (alternative) hypothesis; the null hypothesis is that there is no difference.



type_6mos
satis_hire
satis_6mos
priorexp

Dependent Variables: type_hire

Grouping Variable: gender

Tests

☑ Student
☐ Welch
☐ Mann-Whitney

Hypothesis

● Group 1 ≠ Group 2
○ Group 1 > Group 2
○ Group 1 < Group 2

Assumption Checks

☑ Normality
☑ Equality of variances

Additional Statistics

☐ Location parameter
☐ Confidence interval 95 %
☑ Effect size
☐ Confidence interval 95 %
☑ Descriptives
☐ Descriptives plots
Confidence interval 95 %
☐ Vovk-Sellke maximum p-ratio

Missing Values

● Exclude cases analysis by analysis
○ Exclude cases listwise

Descriptives

Group Descriptives

	Group	N	Mean	SD	SE
type_hire	No	79	66.005	10.108	1.137
	Yes	72	71.648	9.379	1.105

Independent Samples T-Test

Independent Samples T-Test

	t	df	p	Cohen's d
type_hire	0.400	149.000	0.690	0.065

Note. Student's t-test.

Assumption Checks

Test of Normality (Shapiro-Wilk)

		W	p
type_hire	Female	0.979	0.240
	Male	0.987	0.658

Note. Significant results suggest a deviation from normality.

Test of Equality of Variances (Levene's)

	F	df	p
type_hire	6.388e −4	1	0.980

Descriptives

Group Descriptives

	Group	N	Mean	SD	SE
type_hire	Female	77	69.020	9.688	1.104
	Male	74	68.359	10.641	1.237

Figure 9.6: Independent Samples T-test of Typing Skills

The difference in typing skills between the genders is quite small (69.020 versus 68.359). This difference translates into a t-value of 0.400, which (at 149 degrees of freedom) is not significant. The p-value is the probability of a t-value of up to 0.400 to occur assuming that there is no difference between the genders. Since the probability>5%, we accept the null hypothesis (and reject the alternative hypothesis). We have not found evidence in support of differences in typing skills between the groups.

Note: in research, avoid statement like "having found proof". You cannot prove things. The data either support your hypothesis or it doesn't, but even in the case of strong support (low probability of the null hypothesis being true), there is always some probability that you draw the wrong conclusion!

Self-test

Repeat the exercise for employee satisfaction. Has employee satisfaction changed over the six months period? Is there a difference in satisfaction between male and female employees (when hired; and after six months)?

Note that the output gives the effect sizes. Interpret the effect size!

9.4 Advanced: Using Regression Analysis for T-test

The T-test is just a special version of regression analysis that we will discuss in chapter 12. To be more precise: the T-test is a special case of Analysis of Variance (ANOVA) which in turn is a special case of regression analysis.

While the T-test is restricted to differences between two groups, ANOVA can be applied in situations of two or more groups. ANOVA is a regression analysis with dummy variables as independent variables. While regression analysis is limited to models with one dependent variable, more complex models with more than one dependent or endogenous variables, can be estimated using structural equation modeling (SEM). The question why statisticians still use the T-test and ANOVA for these special cases rather than use the parent *technique in the hierarchy (SEM), is not easy to answer. One reason is familiarity with the T-test, and the ease of interpretation. A second reason is that ANOVA and regression analysis have been developed in separate disciplines. While ANOVA is popular in psychology where researchers use experiments, regression analysis is more popular among economists who cannot make use of experimental designs. A third reason is that the T-test comes with options that are not implemented in higher order techniques. Still, it is good to realize that the techniques are hierarchically related.*

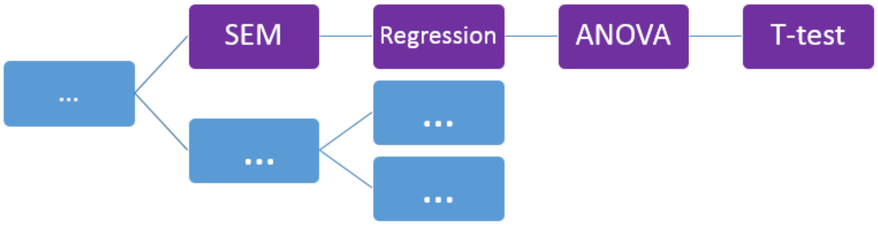


Figure 9.7: A Hierarchy of Techniques

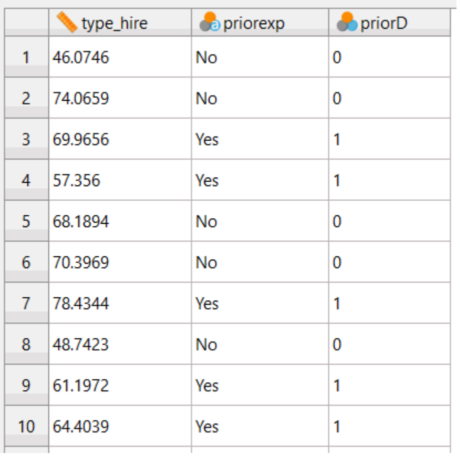
*A regression model is a model in which one or more independent variables (or predictors) are used to explain (or predict) the dependent variable. In the formula (cf. Landers, page 337) typing skills as the dependent variable, would be represented by **y**, while the grouping variable (let's take prior experience) is represented by the independent variable **x**.*

> *Dummy coding: The important point is that we cannot use just any coding scheme for the grouping variable; it has to be 0 and 1! For two groups, that's just enough. We will use the code 0 for no prior experience, and the code 1 for prior experience. We call this "dummy" coding: prior experience is now a dummy (dichotomous, or 0/1) variable. But there's nothing dumb about it, it is widely used by researchers. In addition, it's pretty flexible since we can use it in case of three or more categories too (as we will explain in the next chapter on ANOVA). The rule is that you need one dummy variable in case of two categories, two dummies in case of three categories, or in general, you need **(k-1)** dummies in case of **k** groups.*

The reason to explain a T-test using regression is twofold. First, we want to show you that indeed the T-test and the regression model with a dummy variable do generate the same results. And secondly, regression provides us with a measure called R², the coefficient of determination. The square root of R² is, of course, R which is a correlation coefficient. The correlation coefficient R is often used as a measure of the effect size. Nowadays, in academic publications, you are requested to report effect sizes.

> *Effect size. To understand the importance of effect size, think about the following example. In anticipation of the presidential elections in the US, you ask a small sample of people whom they will vote for. Out of 20 people, 12 say they will vote for the democratic candidate, and 8 for his republican rival. The difference (60% versus 40%) is quite large. The sample however is too small to draw hard conclusions. In short: the effect is large, but the difference is not significant. As the elections are nearing, you decide to increase your sample, from 20 to 10,000. This time the difference is much smaller (52% versus 48%). The "effect" is much smaller but the small difference may very well be significant (it is unlikely to get this result if the preferences in the population are 50/50).*

*Let's use regression analysis, to see if typing skills differ between employees with and without experience (when hired). We use **chapter9_2.csv** which is a subset of **chapter9_1.csv**, but with an added column (**priorD**) which is dummy-coded (0, if **priorexp** is "No"; and 1, if "Yes"). See below.*

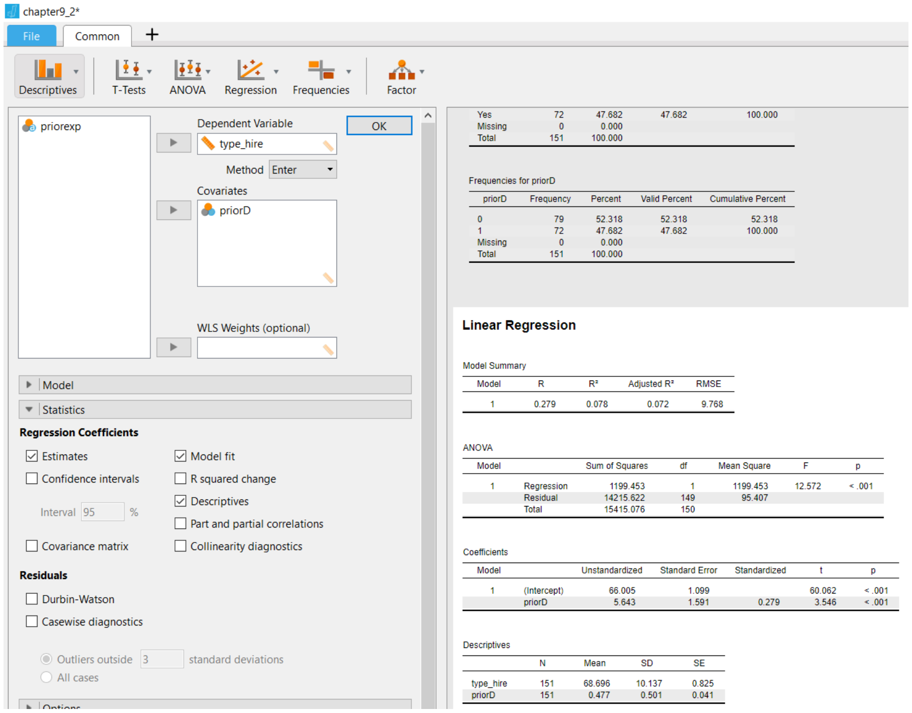


	type_hire	priorexp	priorD
1	46.0746	No	0
2	74.0659	No	0
3	69.9656	Yes	1
4	57.356	Yes	1
5	68.1894	No	0
6	70.3969	No	0
7	78.4344	Yes	1
8	48.7423	No	0
9	61.1972	Yes	1
10	64.4039	Yes	1

Figure 9.8: Data in Chapter9_2, with Dummy Coding

For regression analysis, just click on the ***<Regression>*** *tab, and define the dependent and independent variable. The output provides the same information as the T-test but in a different format. Take some minutes to study the contents of figure 9.9!*

The intercept, is the mean value of the dependent variable if the independent variable equals zero. That is, it gives us the typing skills for the group with no prior experience. The value is 66.005, the same as we have seen before. The "effect" of having prior experience is the (unstandardized) coefficient of ***priorD****. The coefficient of 5.643 is the exact difference between the mean typing skills of the group with prior experience versus the group without. Check this for yourself! The coefficient comes with a t-value of 3.546, identical to the t-value we found when applying the T-test.*



Dependent Variable: type_hire

Method: Enter

Covariates: priorD

WLS Weights (optional)

Model

Statistics

Regression Coefficients

☑ Estimates

☐ Confidence intervals

Interval 95 %

☐ Covariance matrix

☑ Model fit

☐ R squared change

☑ Descriptives

☐ Part and partial correlations

☐ Collinearity diagnostics

Residuals

☐ Durbin-Watson

☐ Casewise diagnostics

● Outliers outside 3 standard deviations

○ All cases

Options

Yes	72	47.682	47.682	100.000
Missing	0	0.000		
Total	151	100.000		

Frequencies for priorD

priorD	Frequency	Percent	Valid Percent	Cumulative Percent
0	79	52.318	52.318	52.318
1	72	47.682	47.682	100.000
Missing	0	0.000		
Total	151	100.000		

Linear Regression

Model Summary

Model	R	R²	Adjusted R²	RMSE
1	0.279	0.078	0.072	9.768

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	1199.453	1	1199.453	12.572	< .001
	Residual	14215.622	149	95.407		
	Total	15415.076	150			

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	66.005	1.099		60.062	< .001
	priorD	5.643	1.591	0.279	3.546	< .001

Descriptives

	N	Mean	SD	SE
type_hire	151	68.696	10.137	0.825
priorD	151	0.477	0.501	0.041

Figure 9.9: T-test Using Regression Analysis

While the T-test computes Cohen's D as a measure of effect size, here we have R (0.279).

The interpretation of R as effect size is that an R of 0.1 is a weak effect; 0.3 is a medium affect; and 0.5 is a strong effect. Here, with R=0.279, we have a (weak to) medium sized effect. Cohen's D of around 0.6 would lead to the same conclusion.

9.5 Data skill challenge

A regional manager implements a policy change for stores in his region (region A) to begin greeting customers whenever they are standing within a 3-meter distance in the store. After the policy has been in place for one month he compares average customer satisfaction for his 10 stores (region A) with the average customer satisfaction in region B. Customer satisfaction is measured on a 1-5 scale with 1 being "very unsatisfied" to 5 being "very satisfied". He expects that his stores (region A) will have higher customer satisfaction ratings compared to Region B.

Below are the data. Read them into a data file!

(a) Calculate the mean customer satisfaction in both regions.

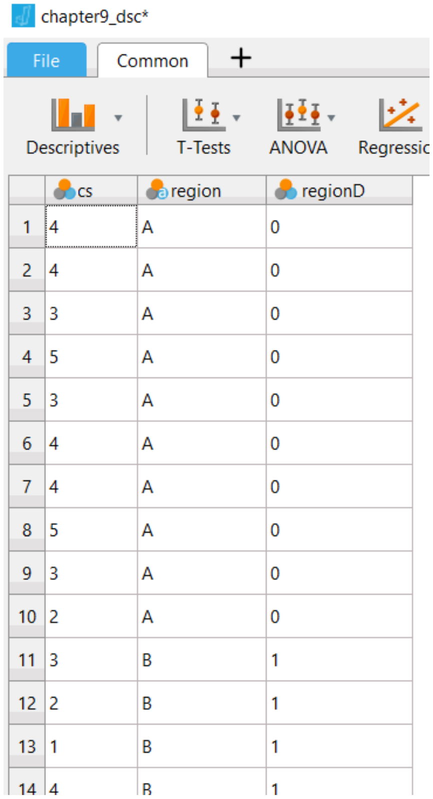
(b) Test whether customer satisfaction differs by region.

(c) Calculate the effect size.

(d) Use both a T-test and regression analysis with a dummy variable for region.

Mean customer satisfaction for stores 1-10 in Region A: 4, 4, 3, 5, 3, 4, 4, 5, 3, 2.

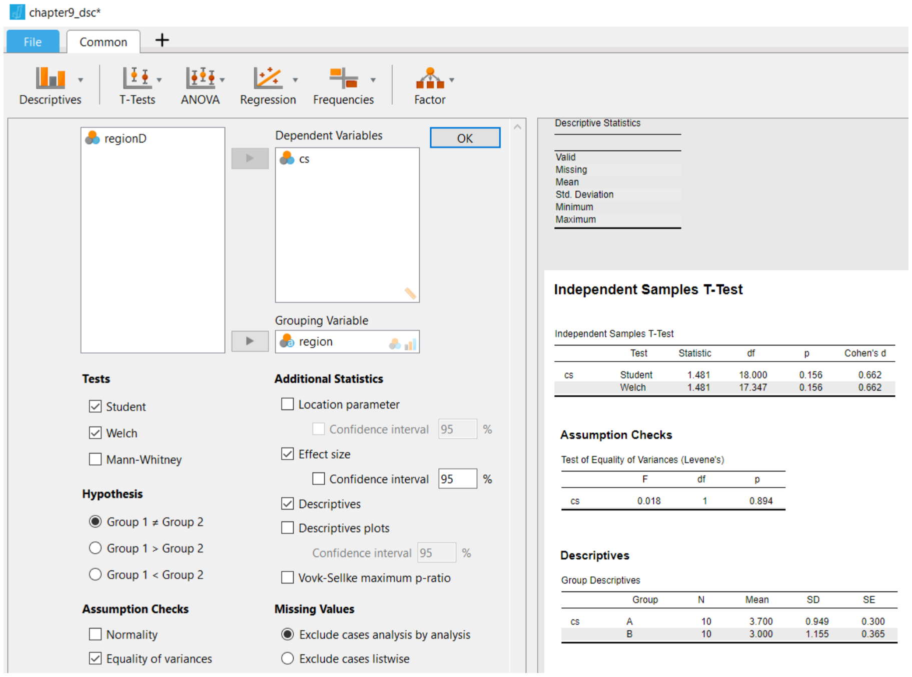
Mean customer satisfaction for stores 1-10 in Region B: 3, 2, 1, 4, 3, 3, 4, 5, 2, 3.



	cs	region	regionD
1	4	A	0
2	4	A	0
3	3	A	0
4	5	A	0
5	3	A	0
6	4	A	0
7	4	A	0
8	5	A	0
9	3	A	0
10	2	A	0
11	3	B	1
12	2	B	1
13	1	B	1
14	4	B	1

Figure 9.10: Data Skill Challenge Data

The T-test is easy enough:



regionD

Dependent Variables
cs

Grouping Variable
region

Tests

☑ Student
☑ Welch
☐ Mann-Whitney

Hypothesis

● Group 1 ≠ Group 2
○ Group 1 > Group 2
○ Group 1 < Group 2

Assumption Checks

☐ Normality
☑ Equality of variances

Additional Statistics

☐ Location parameter
☐ Confidence interval 95 %
☑ Effect size
☐ Confidence interval 95 %
☑ Descriptives
☐ Descriptives plots
Confidence interval 95 %
☐ Vovk-Sellke maximum p-ratio

Missing Values

● Exclude cases analysis by analysis
○ Exclude cases listwise

Descriptive Statistics

Valid
Missing
Mean
Std. Deviation
Minimum
Maximum

Independent Samples T-Test

Independent Samples T-Test

	Test	Statistic	df	p	Cohen's d
cs	Student	1.481	18.000	0.156	0.662
	Welch	1.481	17.347	0.156	0.662

Assumption Checks

Test of Equality of Variances (Levene's)

	F	df	p
cs	0.018	1	0.894

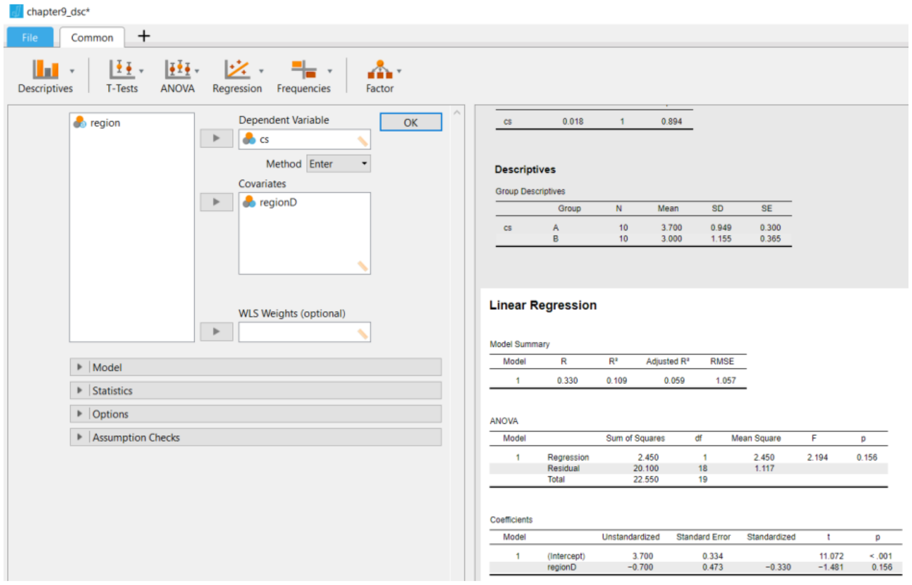
Descriptives

Group Descriptives

	Group	N	Mean	SD	SE
cs	A	10	3.700	0.949	0.300
	B	10	3.000	1.155	0.365

The scores for Region A are higher, probably due to the policy intervention. However, the p-value related to the t-value of 1.48 is above 5% (.0779, or 7.79%) and therefore we retain the null hypothesis.

Let's see how to do the same using regression analysis. We need to use the dummy variable **regionD**.



region

Dependent Variable
cs

Method Enter

Covariates
regionD

WLS Weights (optional)

▸ Model
▸ Statistics
▸ Options
▸ Assumption Checks

cs	0.018	1	0.894

Descriptives

Group Descriptives

	Group	N	Mean	SD	SE
cs	A	10	3.700	0.949	0.300
	B	10	3.000	1.155	0.365

Linear Regression

Model Summary

Model	R	R²	Adjusted R²	RMSE
1	0.330	0.109	0.059	1.057

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	2.450	1	2.450	2.194	0.156
	Residual	20.100	18	1.117		
	Total	22.550	19			

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	3.700	0.334		11.072	< .001
	regionD	−0.700	0.473	−0.330	−1.481	0.156

We leave it to you to find the key statistics, and compare them to the T-test.

In addition to the non-significant difference, we can also report the effect size: The square root of R^2, is .33, which is a medium effect size. You notice that insignificant effects, can still be classified as medium! Effect size is simply a different concept than significance!

One of the reasons to still use the T-test even though it's just a special case in a hierarchy of techniques, are the options that you can use in case assumptions are violated. One of the assumptions is that the variances in the two groups are the same. You can use a so-called *Welch correction* if the assumption of equal variances does not hold. We have added the Welch-correction in the T-test.

The changes are minor due to the fact that the variances are in the same league. The Welch formula makes a correction in the degrees of freedom, which then translates into a different t-value and p-value.

10 Analysis of Variance

<u>Files needed:</u>
chapter10.csv
chapter10_dummy.csv
chapter10_dummyExcel.xlsx
chapter10_dsc.xlsx
chapter10_dsc.csv

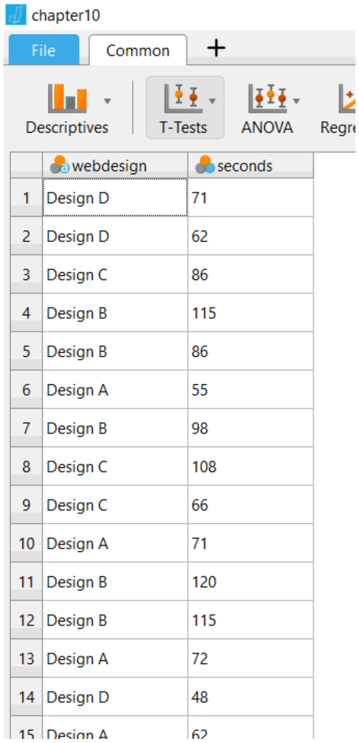
10.1 Introduction

Now that you are familiar with the T-test, and you understand that the T-test is just a special case of regression analysis, ANOVA shouldn't pose a problem. Despite its somewhat confusing name (*analysis of variance*) ANOVA is an extension of the T-test analysis applied to two or more groups.

ANOVA tests whether the means of all the groups are the same.

In the same vein, ANOVA too is related to regression analysis: where we used a regression model with one dummy variable in a T-test for two groups, we use a regression model with two or more dummies in case of three or more groups.

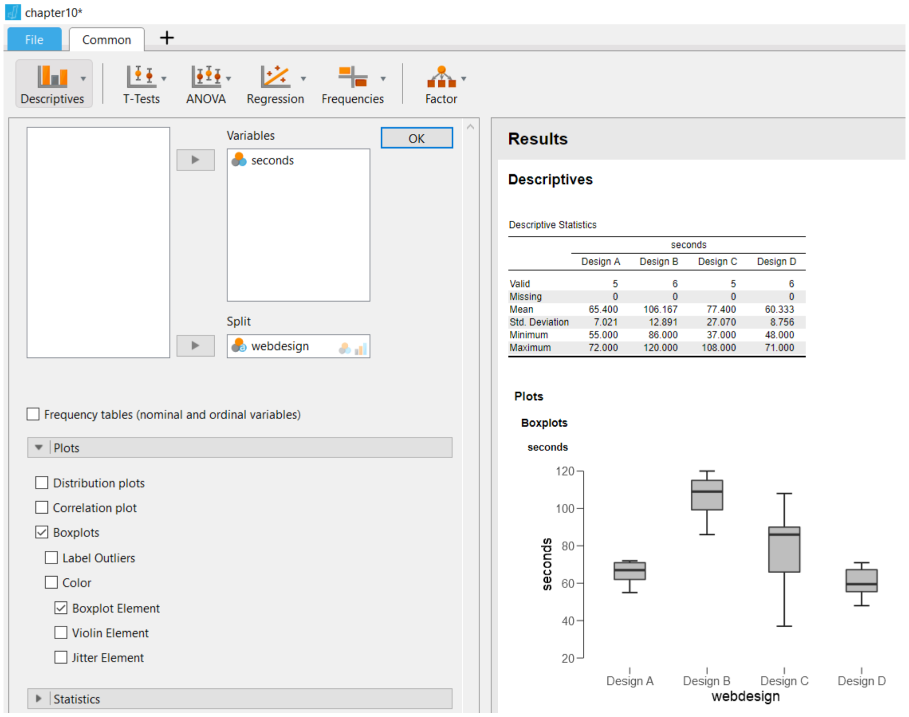
Let's look at the example used by Landers. A company has designed four websites and - using an experimental design - has recorded how many seconds the respondents spent on these websites. The main question is, is there a difference in the time spent on each of these websites?



	webdesign	seconds
1	Design D	71
2	Design D	62
3	Design C	86
4	Design B	115
5	Design B	86
6	Design A	55
7	Design B	98
8	Design C	108
9	Design C	66
10	Design A	71
11	Design B	120
12	Design B	115
13	Design A	72
14	Design D	48
15	Design A	62

Figure 10.1: The Data for ANOVA

A boxplot is a nice way to look at the distributions of the time spent on websites, broken down by **design**.



Variables: seconds

Split: webdesign

☐ Frequency tables (nominal and ordinal variables)

☐ Distribution plots
☐ Correlation plot
☑ Boxplots
☐ Label Outliers
☐ Color
☑ Boxplot Element
☐ Violin Element
☐ Jitter Element

Results

Descriptives

Descriptive Statistics

	seconds			
	Design A	Design B	Design C	Design D
Valid	5	6	5	6
Missing	0	0	0	0
Mean	65.400	106.167	77.400	60.333
Std. Deviation	7.021	12.891	27.070	8.756
Minimum	55.000	86.000	37.000	48.000
Maximum	72.000	120.000	108.000	71.000

Plots

Boxplots

seconds

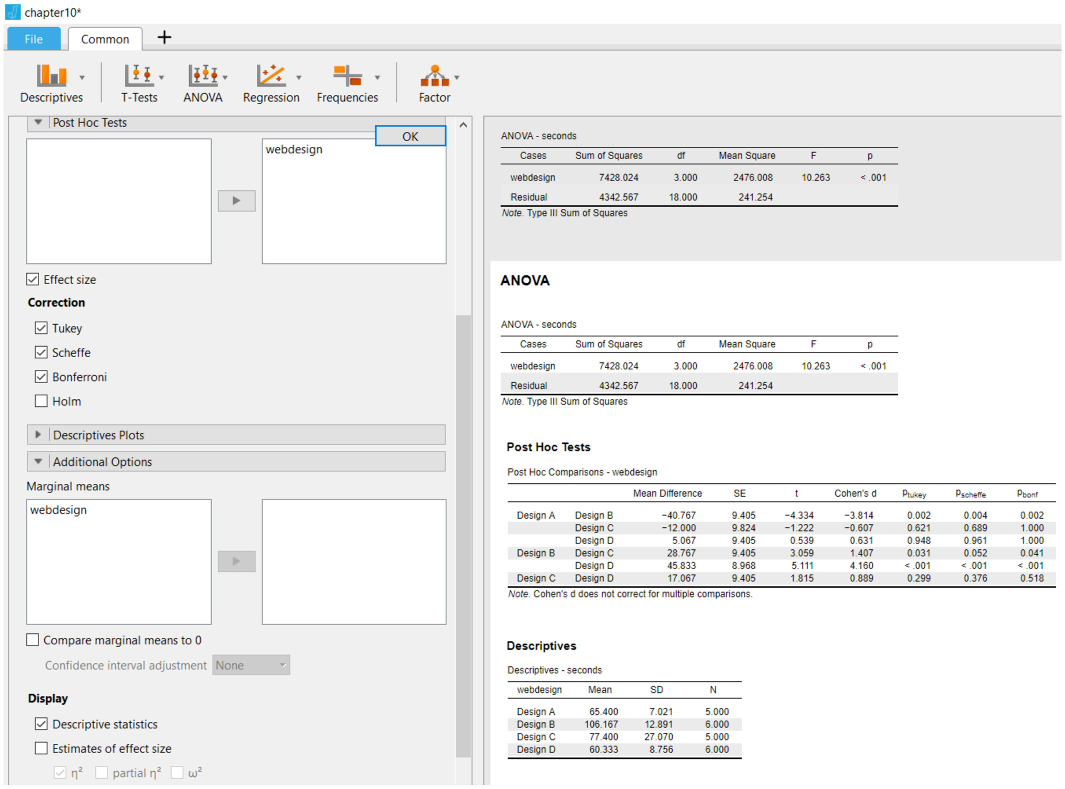
Figure 10.2 Boxplot

The graph and the output show that the mean of design B is, with 106 seconds, well higher than the means for the other designs. The horizontal lines within the four boxes of the boxplot represent the medians (rather than the means). The boxes themselves contain "the middle half" of the observations for each group – which is a bit tedious with groups of 5 respondents. The middle half for design B has no overlap with any of the other middle halves; however, designs A and D do overlap, in terms of time spent.

Now that we have a bird's eye view of the differences, we can use ANOVA to test whether the differences are significant. ANOVA tests whether the time spent on these websites is the same. But it may be that while (like here) the answer seems to be no (B is well higher than the rest), still some of the groups are quite similar. To find out we need an additional test that tells us in "pairwise" comparisons, which groups differ from the others!

10.2 ANOVA

The ANOVA can be obtained by clicking the **<ANOVA>** tab. We have ticked the option *Scheffe*, along with some other options, for multiple comparison tests.



Post Hoc Tests

webdesign

☑ Effect size

Correction

☑ Tukey

☑ Scheffe

☑ Bonferroni

☐ Holm

Descriptives Plots

Additional Options

Marginal means

webdesign

☐ Compare marginal means to 0

Confidence interval adjustment None

Display

☑ Descriptive statistics

☐ Estimates of effect size

☑ η^2 ☐ partial η^2 ☐ ω^2

ANOVA - seconds

Cases	Sum of Squares	df	Mean Square	F	p
webdesign	7428.024	3.000	2476.008	10.263	< .001
Residual	4342.567	18.000	241.254		

Note. Type III Sum of Squares

ANOVA

ANOVA - seconds

Cases	Sum of Squares	df	Mean Square	F	p
webdesign	7428.024	3.000	2476.008	10.263	< .001
Residual	4342.567	18.000	241.254		

Note. Type III Sum of Squares

Post Hoc Tests

Post Hoc Comparisons - webdesign

		Mean Difference	SE	t	Cohen's d	p_{tukey}	$p_{scheffe}$	p_{bonf}
Design A	Design B	−40.767	9.405	−4.334	−3.814	0.002	0.004	0.002
	Design C	−12.000	9.824	−1.222	−0.607	0.621	0.689	1.000
	Design D	5.067	9.405	0.539	0.631	0.948	0.961	1.000
Design B	Design C	28.767	9.405	3.059	1.407	0.031	0.052	0.041
	Design D	45.833	8.968	5.111	4.160	< .001	< .001	< .001
Design C	Design D	17.067	9.405	1.815	0.889	0.299	0.376	0.518

Note. Cohen's d does not correct for multiple comparisons.

Descriptives

Descriptives - seconds

webdesign	Mean	SD	N
Design A	65.400	7.021	5.000
Design B	106.167	12.891	6.000
Design C	77.400	27.070	5.000
Design D	60.333	8.756	6.000

Figure 10.3 ANOVA

In the output, we find the overall test. The F-test, with an F-value of 10.263 (and 3 and 18 degrees of freedom), rejects the hypothesis that all means are the same; the probability is less than 0.1%. We report that F(3, 18) = 10.26; p<5%.

From the descriptives (and our boxplots) we see that design B outperforms all other designs. The hypothesis that all means are the same is firmly rejected.

The output contains pair-wise comparisons. We have just learned that not all means are the same. But the more relevant question is, which means differ from one another?

With 4 websites we can compare 6 pairs: A to B, C and D; that makes three; B to C and D; and C to D, that's six in total. You can compute the number of pairs as ½*k*(k-1), with k representing the number of groups. Here we have ½*4*3 = 6.

The idea behind multiple comparison tests is that when making several pairwise comparisons we are *capitalizing on chance*. We call differences significant if the probability of our test-statistic - under the assumption of no difference in the population - is smaller than 5%. That is, in repeated testing we are bound to be wrong 5% of our trials.

> *Suppose we make 6 comparisons between groups that in reality do not differ from another. Since we are testing with 95% confidence, the probability that we are right is 95% for each of the 6 trials. But with 6 comparisons, the probability that we make at least one mistake equals 1 minus the probability that we are right all the time: 1 – 0.95^6 = 1 – 0.74 = 0.26. The chance that we make at least one mistake, increases with the number of trials!*

Multiple comparison tests correct for that by being stricter. There are several kinds of corrections, one of which *Scheffe*. All corrections have their pros and cons, but the results are quite robust. You can try various corrections to check if the findings are consistent.

In the output there's a matrix in which A (in the first column) is compared to B, C and D; next, B is compared to C and D; and finally, C to D. Although overall the ANOVA found that not all groups (websites, here) are the same, the only significant differences (based on *Scheffe* corrections) are between A and B; and between B and D. The differences between all other combinations of websites are not significantly different from zero, at the 5% level of significance.

10.3 Extra: ANOVA via Regression Analysis

*Since ANOVA is a special case of regression analysis, we can use **regress** with dummy variables to accomplish the same. In the T-test we examined the differences between two groups. In ANOVA we use two or more groups. Applying regression to our example with four groups, we need three dummy variables. The group that serves as the base or reference group will have a code of zero on all three dummies.*

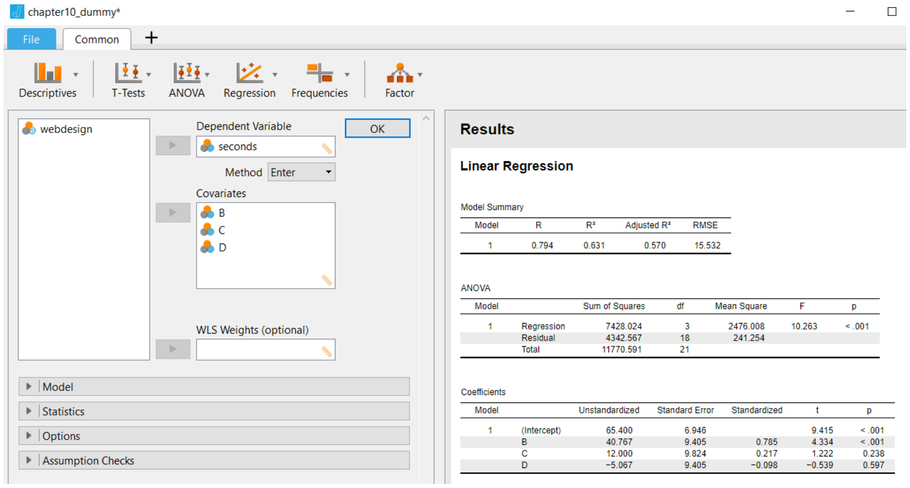
Table 10.1: Scheme for dummy coding, for our four web designs

Design	*Dummy 1 (design B)*	*Dummy 2 (design C)*	*Dummy 3 (design D)*
A (reference group)	*0*	*0*	*0*
B	*1*	*0*	*0*
C	*0*	*1*	*0*
D	*0*	*0*	*1*

*The data are in **chapter10_dummy.csv**.*

*If you recall the use of regression for the T-test, you can interpret the key statistics. The constant term **_cons** is the mean of the reference group (design A) which is 65.4. Seconds spent on design B exceed design A by 40.77 seconds: 65.400 + 40.767 = 106.167. The difference is significant (t=4.33; p<0.001). And so on. The F-statistic for the regression is the same as for ANOVA (F(3, 18)=10.263; p<0.001).*

*The big disadvantage of using regression analysis in **JASP**, is that we don't have all pairwise comparisons.*



Dependent Variable: seconds

Method: Enter

Covariates: B, C, D

WLS Weights (optional)

Results

Linear Regression

Model Summary

Model	R	R²	Adjusted R²	RMSE
1	0.794	0.631	0.570	15.532

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	7428.024	3	2476.008	10.263	< .001
	Residual	4342.567	18	241.254		
	Total	11770.591	21			

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	65.400	6.946		9.415	< .001
	B	40.767	9.405	0.785	4.334	< .001
	C	12.000	9.824	0.217	1.222	0.238
	D	−5.067	9.405	−0.098	−0.539	0.597

Figure 10.4 ANOVA Using Regression Analysis

10.4 Data Skill Challenge

The company in our example has decided to run a second wave of tests, keeping designs B and C but adding designs E, F, G and H. The data are in the table below.

(a) Read the data in a data file (note that the lay-out as below is not appropriate!)

(b) Present overviews using descriptives and boxplots

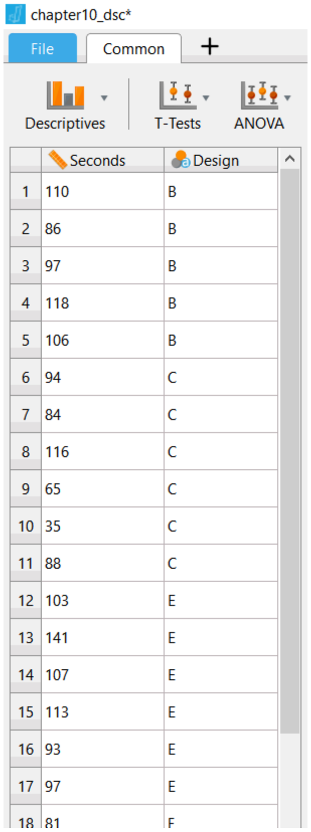
(c) Conduct the full hypothesis testing procedure and draw conclusions.

(d) Which designs are significantly different from other designs?

Design B	Design C	Design E	Design F	Design G	Design H
110	94	103	81	56	140
86	84	141	79	60	115
97	116	107	70	80	130
118	65	113	57	57	146
106	35	93	93	55	109
	88	97			126

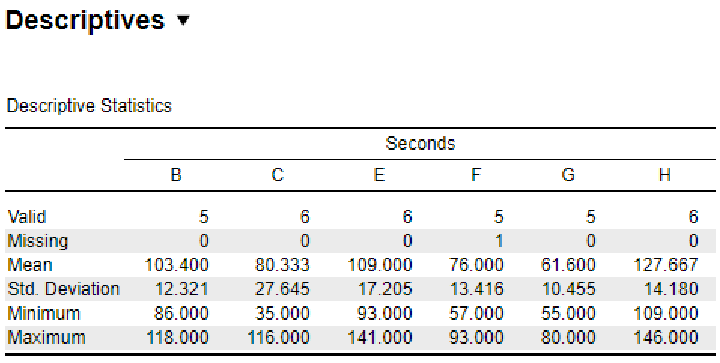
Figure 10.5: Data for Data Skill Challenge

The data, for your convenience, are stored in **chapter10_dsc.csv**. You can copy the data from figure 10.5, and paste them into Excel. For analysis in **JASP**, you have to rearrange the data in two columns, one for **seconds**, and one for **design**. Next, save the file in CSV-format (we have done that for you), and read the data in **JASP**.



	Seconds	Design
1	110	B
2	86	B
3	97	B
4	118	B
5	106	B
6	94	C
7	84	C
8	116	C
9	65	C
10	35	C
11	88	C
12	103	E
13	141	E
14	107	E
15	113	E
16	93	E
17	97	E
18	81	F

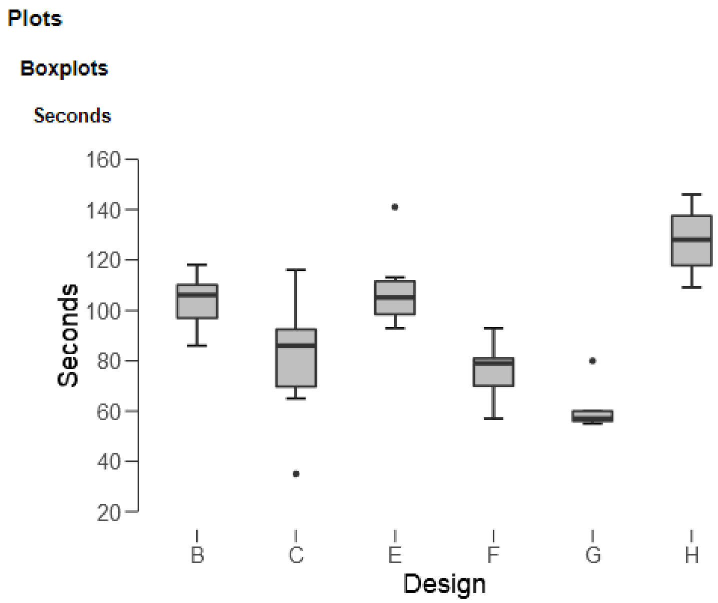
Always describe the data, to get a good feel of the data, before performing advanced analyses like ANOVA. From the descriptives, it seems that design H is doing better than all others. Design G is doing poorly. But are the differences significant, that is, can we draw hard conclusions?



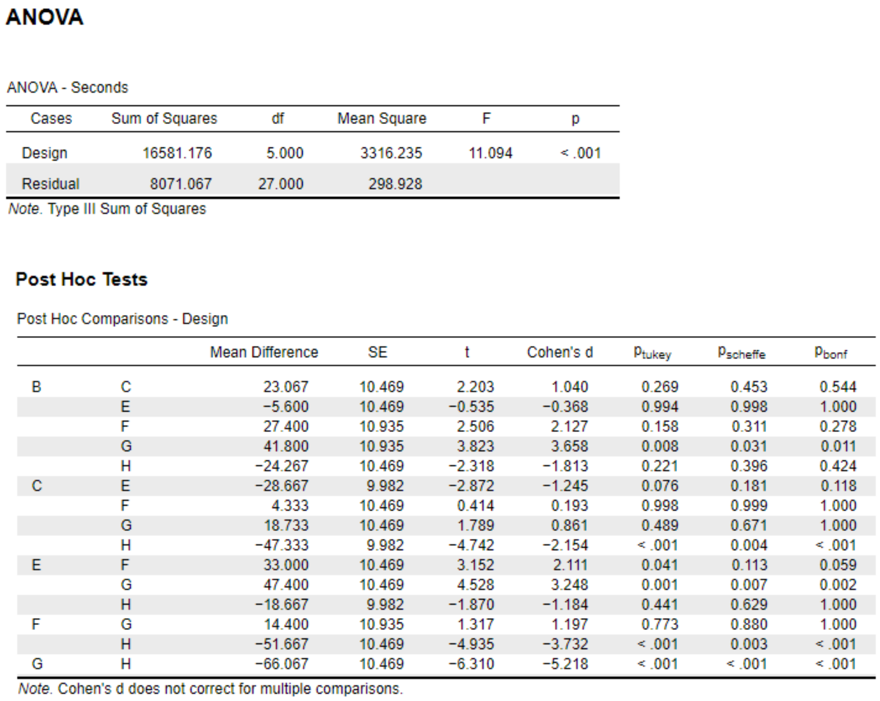
Descriptives ▼

Descriptive Statistics

	Seconds					
	B	C	E	F	G	H
Valid	5	6	6	5	5	6
Missing	0	0	0	1	0	0
Mean	103.400	80.333	109.000	76.000	61.600	127.667
Std. Deviation	12.321	27.645	17.205	13.416	10.455	14.180
Minimum	86.000	35.000	93.000	57.000	55.000	109.000
Maximum	118.000	116.000	141.000	93.000	80.000	146.000



From the *post-hoc* multiple comparison tests, it turns out that H is not significantly better than B and E. Design H is outperforming designs C, F and G. Design G is not significantly worse than design C and F. And so on.



ANOVA

ANOVA - Seconds

Cases	Sum of Squares	df	Mean Square	F	p
Design	16581.176	5.000	3316.235	11.094	< .001
Residual	8071.067	27.000	298.928		

Note. Type III Sum of Squares

Post Hoc Tests

Post Hoc Comparisons - Design

		Mean Difference	SE	t	Cohen's d	p_{tukey}	$p_{scheffe}$	p_{bonf}
B	C	23.067	10.469	2.203	1.040	0.269	0.453	0.544
	E	−5.600	10.469	−0.535	−0.368	0.994	0.998	1.000
	F	27.400	10.935	2.506	2.127	0.158	0.311	0.278
	G	41.800	10.935	3.823	3.658	0.008	0.031	0.011
	H	−24.267	10.469	−2.318	−1.813	0.221	0.396	0.424
C	E	−28.667	9.982	−2.872	−1.245	0.076	0.181	0.118
	F	4.333	10.469	0.414	0.193	0.998	0.999	1.000
	G	18.733	10.469	1.789	0.861	0.489	0.671	1.000
	H	−47.333	9.982	−4.742	−2.154	< .001	0.004	< .001
E	F	33.000	10.469	3.152	2.111	0.041	0.113	0.059
	G	47.400	10.469	4.528	3.248	0.001	0.007	0.002
	H	−18.667	9.982	−1.870	−1.184	0.441	0.629	1.000
F	G	14.400	10.935	1.317	1.197	0.773	0.880	1.000
	H	−51.667	10.469	−4.935	−3.732	< .001	0.003	< .001
G	H	−66.067	10.469	−6.310	−5.218	< .001	< .001	< .001

Note. Cohen's d does not correct for multiple comparisons.

11 Chi-Squared Tests of Fit

Files needed:
chapter11_example.xlsx
chapter11.csv
chapter11_dsc.csv
chapter11_EXTRA.csv

11.1 Introduction

In the previous chapters we have looked at situations in which the dependent variable is measured on an interval or ratio scale.

> ***Interval or ratio scales*** *give rich information. Income, for example, is measured on a ratio scale. A person earning 1,000 € earns twice as much as a person earning 500 €.*
>
> ***Ordinal scaled data*** *contain less information. Suppose we use scores of 1 to 5 to measure overall job satisfaction, an employee with a score of 4 is quite satisfied, and more satisfied than an employee with a score of 2 – but not necessarily twice as satisfied.*
>
> ***In nominal scaled data****, numbers (codes) are meaningless. In previous examples we have use dummy coding for gender, but males and females can be coded with any by itself meaningless number (0 and 1; or 1 and 2; or 11 and 99) as long as we are to distinguish them as groups.*

Dependent (interval or ratio scaled) variables were "explained" by independent variables that were measured as "nominal" variables (groups), like in T-tests and in ANOVA.

In many cases we want to examine relationships between variables that are all measured on a nominal scale, that is, variables that just categorize cases or respondents into *groups*. Examples are:

- Political preferences (you can vote for one out of many parties; one could argue that the parties are on a scale that ranges from the extreme left to the extreme right but in principle the parties are just categories that you can vote for – or against);
- We may be interested in the relationship between gender and wearing glasses. Male and female, and wearing glasses or not, are examples of groupings.

Our interest then is in *count data*: how many people prefer each of the parties? How many men and how many women (as proportions of the total) are wearing glasses? Is gender related to wearing glasses?

11.2 The Distribution of One Nominal Variable

Let's look at the sample. The numbers are the same as in the Landers example on page 301 on chewing gum but we switched to a topic that's harder to chew: politics.

Political parties	Votes (in Sample of 33)	Hypothesized	Previous Election
Liberal Party	10	1/3 (33%)	2/5
Labor Party	14	1/3	2/5
Nationalist Party	9	1/3	1/5

Figure 11.1: Example of Count Data

In figure 11.1 we see that in a **sample** of 33 voters, most interviewees prefer the Labor Party. Statistically, we can determine how likely the sample outcome is, under the null hypothesis that there are no differences in the population. Or we can compare the results to the outcomes of the previous election, and wonder if things have changed.

The *chi-square statistic* is based on the differences between observed (O) and expected (E) counts. The larger the differences, the higher the statistic. In the formula, the differences are squared, which sees to it that the negative and positive differences always become positive and don't cancel out. Since large samples will produce bigger differences, we compensate for that by dividing the squared differences by the expected counts.

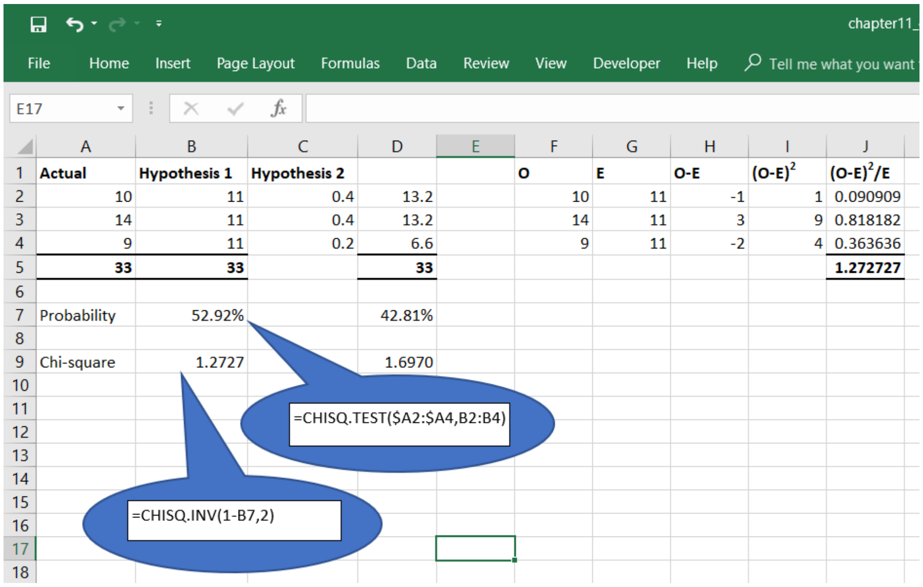
$$Chi\ Square = \sum \frac{(O-E)^2}{E}$$

The chi-square statistic comes with a number of degrees of freedom (DoF). For three groups (like here), the DoF is equal to *k*-1=2; for a *r*c* table, the DoF is equal to the *(r-1)*(c-1)*, where *r* and *c* are the number of rows and columns. For a 2*2 table, the DoF is 1.

In our examples of preferences for political parties, the "expected" counts (assuming no preferences in the population) are 33/3 = 11 for each of the three groups. Once we know the counts for two of the three groups, the count for the third group is fixed: that is, given the sample size, only two counts can vary! In statistical terms, the number of DoF is two.

The test starts from a null hypothesis. For the first question, our null hypothesis would be that the proportions in the sample are 1/3 (≈33%) for each of the parties. If the sample outcome is unlikely to occur under the null hypothesis (p<0.05; the probability is lower than 5%), then we reject the null hypothesis in favor of the alternative hypothesis that the Labor Party is now on top.

Doing the calculations manually based on the data in figure 11.1, would not be that hard. But we don't want to do it manually. We can use functions in Excel.



	A	B	C	D	E	F	G	H	I	J
1	Actual	Hypothesis 1	Hypothesis 2			O	E	O-E	$(O-E)^2$	$(O-E)^2/E$
2	10	11	0.4	13.2		10	11	-1	1	0.090909
3	14	11	0.4	13.2		14	11	3	9	0.818182
4	9	11	0.2	6.6		9	11	-2	4	0.363636
5	33	33		33						1.272727
6										
7	Probability	52.92%		42.81%						
8										
9	Chi-square	1.2727		1.6970						

=CHISQ.TEST($A2:$A4,B2:B4)

=CHISQ.INV(1-B7,2)

Figure 11.2: Data for the Chi-square test (chapter11_example.xlsx)

In figure 11.2, we have entered the data in columns A (observed) and columns B and D (expected). For hypothesis 2 our expectancies are formulated in proportions (in column C), so we compute the expected counts by multiplying the sample size by these proportions. For the sake of illustration, we have computed the chi-square, in columns F to J, step-by-step. The chi-square statistic (1.2727, with DoF=2) can be looked up in chi-square tables; but luckily, Excel has a function to do the job for us.

In cell A7, we have used the **CHISQ.TEST()** function in Excel, to compute the probability of our sample outcome (in column A), given the expected values in column B. In cell D7, we have done the same for comparing the sample outcome to the previous election. The probabilities are well beyond 5%, and therefore we accept the null hypotheses in both cases.

In cells A9 and D9, we have used the **CHISQ.TEST()** function to calculate the chi-square statistic; note that we have to add the DoF (2) as a second argument in this function. The function returns the value of chi-square, associated with the cumulative probability that the chi-square is smaller than or equal to that value. The probability returned by the **CHISQ.TEST()** function, however, is the probability of chi-square exceeding that value. Therefore, in the **CHISQ.TEST()** function, we have to use 1 minus the probability in cells B7 and D7, to get the chi-square!

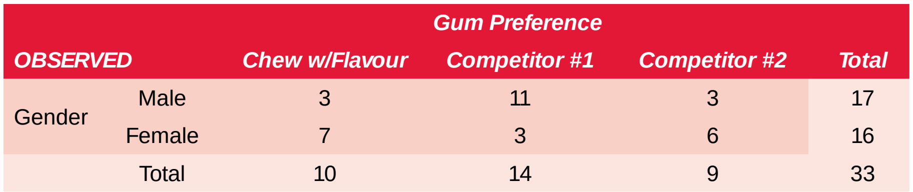
To summarize our test of hypothesis 1:

- Probability of sampling outcome is 52.92%, this is the probability of the chi-square as high (or higher) than the chi-square in this case.
- Since the chi-square is computed behind the scenes, we have to retrieve it!
- Probability of chi-square being less or equal than the chi-square, is 1- 52.92% = 47.08%.
- The chi-square associated with a probability of 47.08%, is 1.2727.
- *Advanced, or extra: the critical value of chi-square (α=0.05; DoF=2) is 5.99 (see 11.4). Since 1.2727 < 5.99, we accept the null hypothesis of no difference; the sampling outcome is not unlikely to occur, under the assumption of no relationship between the variables.*

The outcome shows the chi-square statistic; a value of 1.2727 is not very unlikely (p=0.53) under the null hypothesis of equal probabilities in the population, and therefore we don't find support for our alternative hypothesis. We report that based on our data, we find no evidence to support our alternative hypothesis (Chi-square = 1.27 (2); p=0.53).

11.3 Testing Relationships Between Two Nominal Variables

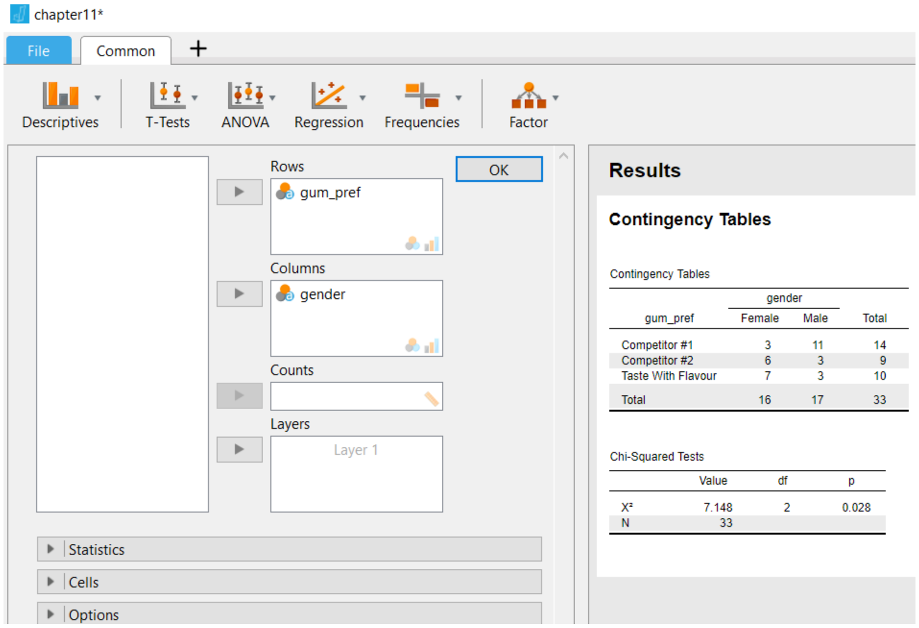
For the example of Landers (page 304) we first read in the data from **JASP** file.



			Gum Preference		
OBSERVED		*Chew w/Flavour*	*Competitor #1*	*Competitor #2*	*Total*
Gender	Male	3	11	3	17
	Female	7	3	6	16
	Total	10	14	9	33

Figure 11.3 Data for the Chi-square test with two variables

It doesn't make a lot of difference which of the variables (gender or male preference) appears in the rows and which in the columns, but – as different from Landers' figure shown above - we would have a slight preference to put gender in the columns. After reading in the data in **chapter11.csv** we use the **<Frequencies><Contingency Tables>** to produce the table along with the chi-square test.



Rows: gum_pref

Columns: gender

Counts

Layers

Results

Contingency Tables

Contingency Tables

	gender		
gum_pref	Female	Male	Total
Competitor #1	3	11	14
Competitor #2	6	3	9
Taste With Flavour	7	3	10
Total	16	17	33

Chi-Squared Tests

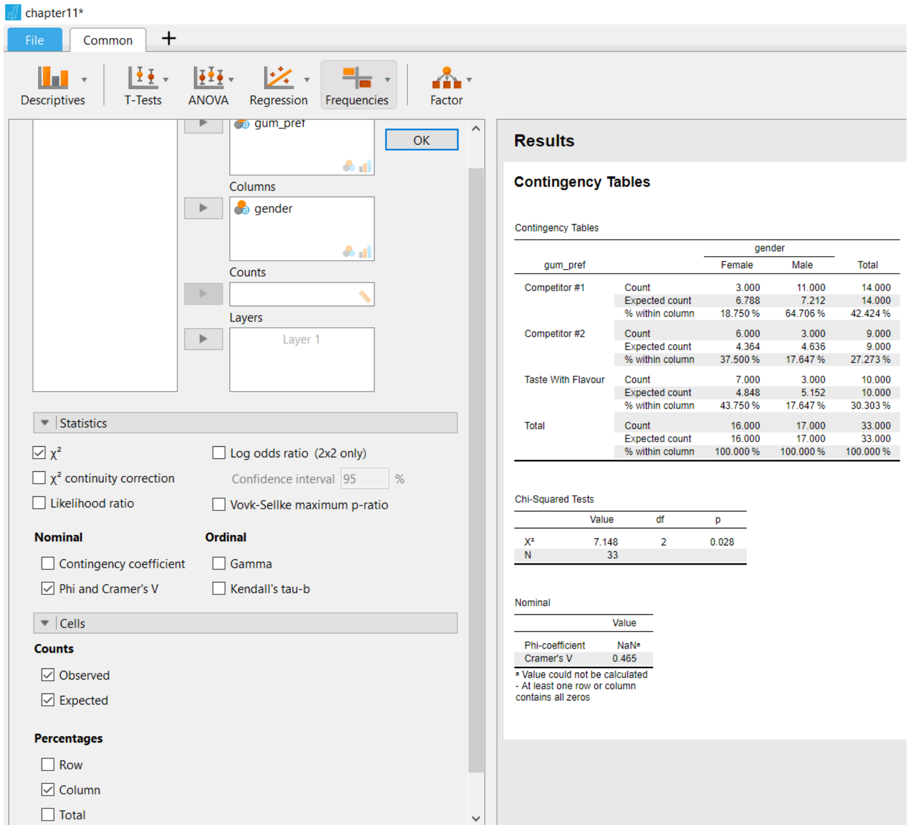
	Value	df	p
X^2	7.148	2	0.028
N	33		

Figure 11.4 Contingency Table with Chi-square Test

We can opt for more information in the cells of the table.

In the figure below, we have added the expected counts the cells. It is easy to compute the expected counts yourself. For example, since we have 16 female persons in our sample, and 14 persons prefer competitor #1, we would expect (under the null hypothesis of no relationship between gender and preference) (16/33) * (14/33) * 33 = 6.788 in this cell. This is equivalent to assuming that the proportion of female persons preferring competitor #1 is the same as the proportion of any person competitor #1; it is also equivalent to assuming that the gender distribution of persons preferring competitor #1, is the same as the overall gender distribution.

For easy interpretation, we have also added column percentages. From the results we see that overall, 42% of the sample prefers competitor #1. However, the percentages vary between female (19%) and male (65%) respondents.



gum_pref

Columns

gender

Counts

Layers

Layer 1

Statistics

☑ χ^2 ☐ Log odds ratio (2x2 only)

☐ χ^2 continuity correction Confidence interval 95 %

☐ Likelihood ratio ☐ Vovk-Sellke maximum p-ratio

Nominal

☐ Contingency coefficient

☑ Phi and Cramer's V

Ordinal

☐ Gamma

☐ Kendall's tau-b

Cells

Counts

☑ Observed

☑ Expected

Percentages

☐ Row

☑ Column

☐ Total

Results

Contingency Tables

Contingency Tables

		gender		
gum_pref		Female	Male	Total
Competitor #1	Count	3.000	11.000	14.000
	Expected count	6.788	7.212	14.000
	% within column	18.750 %	64.706 %	42.424 %
Competitor #2	Count	6.000	3.000	9.000
	Expected count	4.364	4.636	9.000
	% within column	37.500 %	17.647 %	27.273 %
Taste With Flavour	Count	7.000	3.000	10.000
	Expected count	4.848	5.152	10.000
	% within column	43.750 %	17.647 %	30.303 %
Total	Count	16.000	17.000	33.000
	Expected count	16.000	17.000	33.000
	% within column	100.000 %	100.000 %	100.000 %

Chi-Squared Tests

	Value	df	p
Χ²	7.148	2	0.028
N	33		

Nominal

	Value
Phi-coefficient	NaN[a]
Cramer's V	0.465

[a] Value could not be calculated - At least one row or column contains all zeros

Figure 11.5 Contingency Table with Additional Information in Cells

The difference between preferences of male and female respondents is significant (Chi-squared=7.15 (2); p<5%). The number in brackets (2), is the number of degrees of freedom, which is equal to (number of rows minus 1) times (number of columns minus 1). In a 3-by-2 table like we have here, we have 2*1=2 degrees of freedom.

11.4 Additional Statistics

Cramer's V, a measure of effect size that takes a value between 0 and +1, is displayed optionally. Cramer's V is a measure of effect size, with the same rules as the correlation coefficient R. Values of 0.1; 0.3; and 0.5 are interpreted as weak, moderate and strong effects. Here, the effect size is medium to strong.

If you want to report the critical value of chi-square (as in Landers, page 307) then again you do not need to look it up in a table. We can use the Excel function **CHISQ.INV(Probability, DoF)**. Here, the probability to use is 0.95 (when testing at α=0.05), and the DoF is 2.

=CHISQ.INV(0.95,2) = 5.99

Since our test-statistic (7.15) is higher than the critical value (5.99), we reject the null hypothesis. Remember that α is the probability that of falsely rejecting the null hypothesis. We want to keep that error as small as possible. In many disciplines, researchers use an α of 0.05, but it all depends on how serious making this "Type I" error is! In medical studies, it is conceivable that researchers are stricter, when it comes to, say, testing a new medicine.

11.5 Data Skill Challenge

Data Skill Challenge 1

Maria decides to run a focus group, comparing four Chew-with-Flavor's gums to determine which one is preferred. Gums flavors A; B; C; and D are picked 12; 7; 8 and 21 times, respectively.

Complete the full hypothesis testing process, given this data.

The answers to this challenge are in worksheet **DSC1** of **chapter11_example.xlsx**.

Data Skill Challenge 3

Chaitra decides to track the number of accidents at the 12 manufacturing plants she manages. Right now, safety training is conducted by two separate units: one trains the day shift while the other one trains the night shift. Both are generally effective but she is worried that that the night-shift trainers aren't getting out to some of the plants further from the home office.

(a) Complete the full hypothesis testing process given the observed numbers of accidents recorded in the table below, to see if accident counts by shift and plant are related.

(b) Compute the chi-square statistic

(c) Compute the critical value for the chi-square statistic

(d) Compute Cramer's V

(e) What is your main conclusion?

Plant	Day	Night
1	3	4
2	5	7
3	1	0
4	3	1
5	0	1
6	6	7
7	2	0
8	7	6
9	2	9
10	0	13
11	4	6
12	3	12

Figure 11.6: Data for Data Skill Challenge

This is a bit of a challenge.

A solution is to use Excel, and to compute the expected counts. After that, the procedure follows our earlier example.

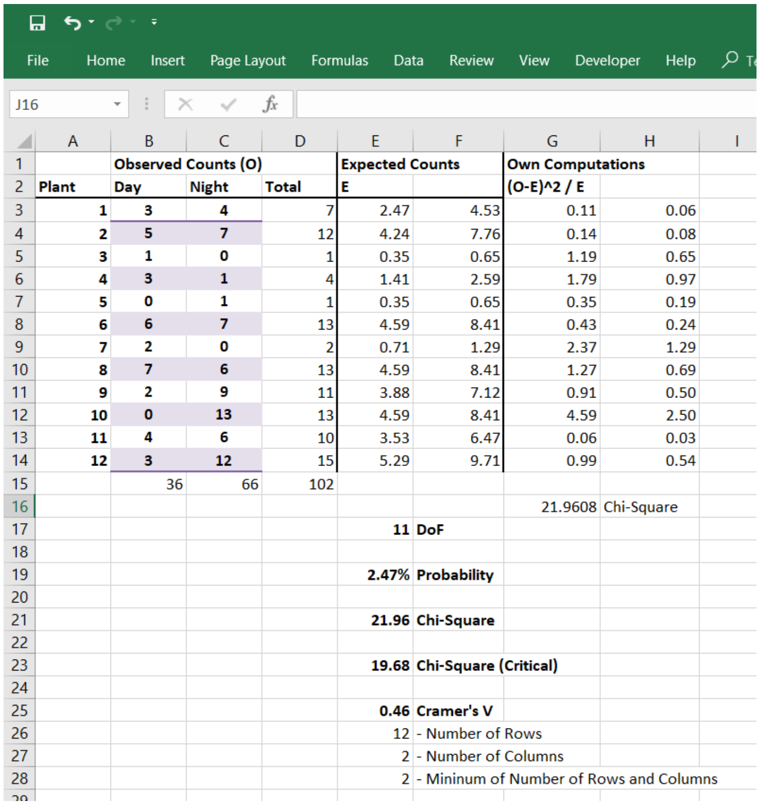
A solution (see below) can be found in worksheet **DSC3**, in **chapter11_example.xlsx**. The expected counts can be computed. For example, for Plant 1 we would expect values of 2.47 and 4.53 for accidents during day and night shifts. Since 7 accidents occurred in this plant, and overall 36 and 66 (in total 102) accidents took place during day and night shifts, we would expect 7*(36/102) and 7*(66/102) to take place during these shifts in plant 1. The observed number (3) is somewhat higher than expected (2.47) – but then again, the observed number has to be an integer (2 or 3).

In the Excel-sheet, we have added Cramer's V. The formula for Cramer's V is:

$$Cramer's\ V = \sqrt{\frac{Chi\ Square}{n * (k - 1)}}$$

In the formula, the chi-square is the value computed using the **CHISQ.INV()** function (21.96); *n* is the sample size (102); and *k* is the minimum of the number of rows and the number of columns. With 12 rows and 2 columns, the minimum is 2. Cramer's V is 0.46, which signifies a medium to strong effect.

For obtaining the chi-square value, remember that the probability (from **CHISQ.TEST()**) is the probability of a chi-square that high, or higher. Here, the probability of obtaining the observed results under the null hypothesis is as small as 2.47%. To display the chi-square, we use the **CHISQ.INV()** function, with – as its first argument - the probability of 1 – 0.0247 = .09753 (97.53%). That is, 97.53% of the distribution has a value of up to 21.96 (and 2.47% a value higher than 21.96).



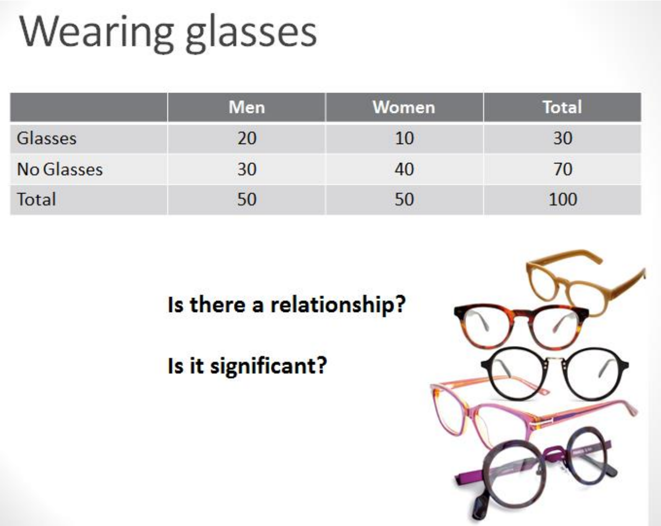
	A	B	C	D	E	F	G	H	I
1		Observed Counts (O)			Expected Counts		Own Computations		
2	Plant	Day	Night	Total	E		(O-E)^2 / E		
3	1	3	4	7	2.47	4.53	0.11	0.06	
4	2	5	7	12	4.24	7.76	0.14	0.08	
5	3	1	0	1	0.35	0.65	1.19	0.65	
6	4	3	1	4	1.41	2.59	1.79	0.97	
7	5	0	1	1	0.35	0.65	0.35	0.19	
8	6	6	7	13	4.59	8.41	0.43	0.24	
9	7	2	0	2	0.71	1.29	2.37	1.29	
10	8	7	6	13	4.59	8.41	1.27	0.69	
11	9	2	9	11	3.88	7.12	0.91	0.50	
12	10	0	13	13	4.59	8.41	4.59	2.50	
13	11	4	6	10	3.53	6.47	0.06	0.03	
14	12	3	12	15	5.29	9.71	0.99	0.54	
15		36	66	102					
16							21.9608	Chi-Square	
17					11	DoF			
18									
19					2.47%	Probability			
20									
21					21.96	Chi-Square			
22									
23					19.68	Chi-Square (Critical)			
24									
25					0.46	Cramer's V			
26					12	- Number of Rows			
27					2	- Number of Columns			
28					2	- Mininum of Number of Rows and Columns			

Figure 11.7: Solution for Data Skill Challenge 3

The significant chi-square statistic indicates that indeed there is a relationship (an association, we would call it) between plants, and day and night shift accidents. The strength of the association is reflected by a Cramer's V of .46. Like for correlation, Cramer's V is interpreted as small (.10), medium (.30) or large (.50) effects. In our case, the effect size of .46 is medium to large.

Data Skill Challenge: Extra

From a sample of 100 respondents, we found the following data on men and women wearing glasses.



Wearing glasses

	Men	Women	Total
Glasses	20	10	30
No Glasses	30	40	70
Total	50	50	100

Is there a relationship?

Is it significant?

Figure 11.8: Gender and Wearing Glasses

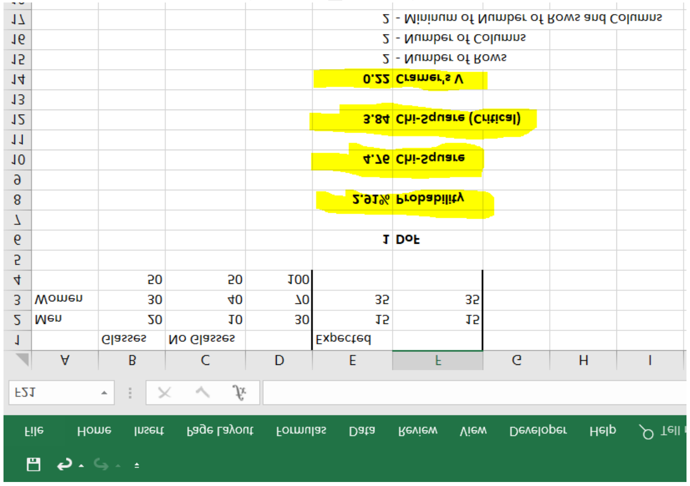
(a) Read the data into Excel

(b) Compute the chi-square statistic

(c) Compute the critical value for the chi-square statistic

(d) Advanced: compute Cramer's V

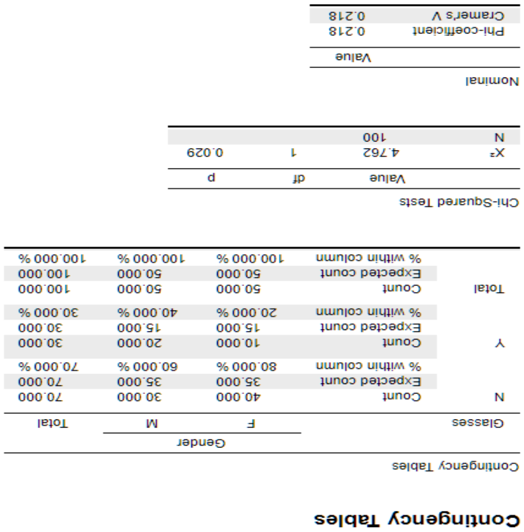
Answer to Extra Data Skills Challenge
(See chapter11_example; worksheet EXTRA)



	A	B	C	D	E	F
1		Glasses	No Glasses		Expected	
2	Men	20	10	30	15	15
3	Women	30	40	70	35	35
4		50	50	100		
5						
6					1	DoF
7						
8					2.91%	Probability
9						
10					4.76	Chi-Square
11						
12					3.84	Chi-Square (Critical)
13						
14					0.22	Cramer's V
15					2	- Number of Rows
16					2	- Number of Columns
17					2	- Minimum of Number of Rows and Columns

Apart from the solution directly applied to the crosstabulation, you can generate a database of 100 records and two columns (variables), as in **chapter11_example; worksheet EXTRA**, and stored in **chapter11_EXTRA.csv**.

Answer to Extra Data Skills Challenge (chapter11_EXTRA.csv)



Contingency Tables

Contingency Tables

		Gender		
Glasses		F	M	Total
N	Count	40.000	30.000	70.000
	Expected count	35.000	35.000	70.000
	% within column	80.000 %	60.000 %	70.000 %
Y	Count	10.000	20.000	30.000
	Expected count	15.000	15.000	30.000
	% within column	20.000 %	40.000 %	30.000 %
Total	Count	50.000	50.000	100.000
	Expected count	50.000	50.000	100.000
	% within column	100.000 %	100.000 %	100.000 %

Chi-Squared Tests

	Value	df	p
X²	4.762	1	0.029
N	100		

Nominal

	Value
Phi-coefficient	0.218
Cramer's V	0.218

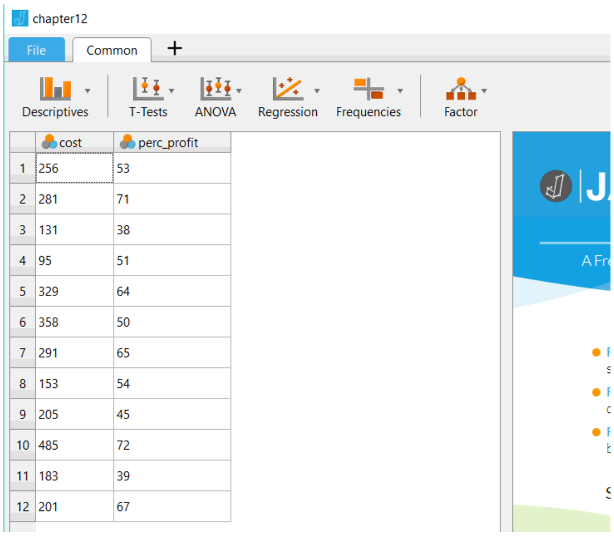
12 Correlation and Regression

Files needed:
chapter12.csv
chapter12_dsc.csv

12.1 Correlation

In the previous chapter we analyzed the relationship between nominal variables (groups, like male/female; or cities). In this chapter we discuss the relationship between two variables measured at an **interval or ratio scale** (e.g. cost; profit; Likert scales). Since the two variables have meaningful values we can use a scattergram to depict our cases (or respondents).

In the example of Landers, we have data on the costs spent on projects, and the profitability of the project.



	cost	perc_profit
1	256	53
2	281	71
3	131	38
4	95	51
5	329	64
6	358	50
7	291	65
8	153	54
9	205	45
10	485	72
11	183	39
12	201	67

Figure 12.1: Data for Chapter 12

We make a scattergram, and fit a regression line.

Self-test questions

What you prefer on the vertical axis: cost or profit? Why?

[It is common to have the explanatory variable on the horizontal (x) axis, and the variable to be explained on the vertical (y) axis!]

Figure out how to switch the axes in **JASP**!

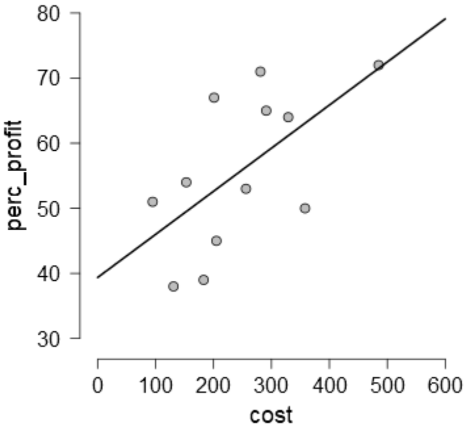
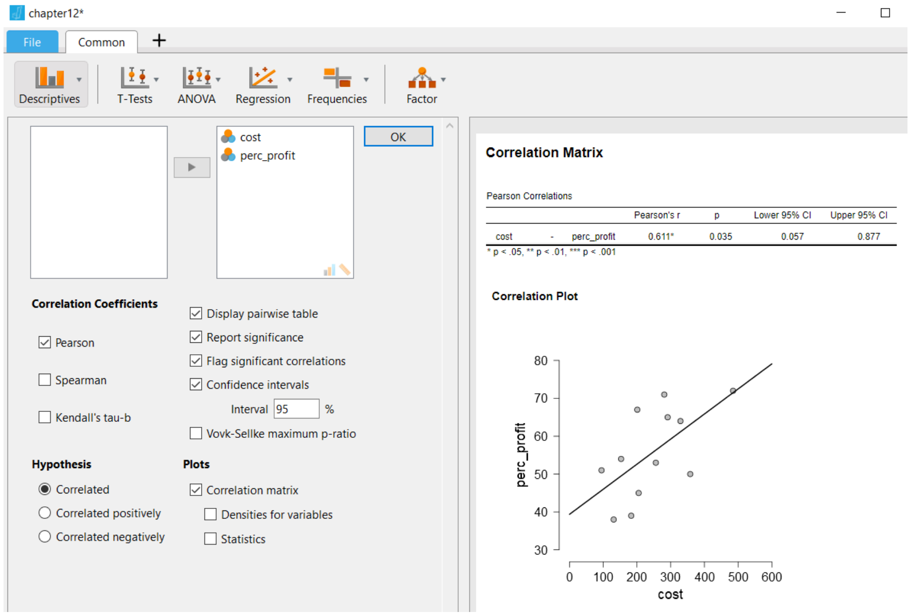


Figure 12.1: Scattergram of the data

The “regression line” slopes upward, indicating a positive correlation between the variables. Some points are quite far from the line-of-best-fit, indicating that the correlation is not perfect. Our statistical interest would be to know if the correlation is significantly different from zero.



Correlation Coefficients

☑ Pearson

☐ Spearman

☐ Kendall's tau-b

☑ Display pairwise table

☑ Report significance

☑ Flag significant correlations

☑ Confidence intervals

Interval 95 %

☐ Vovk-Sellke maximum p-ratio

Hypothesis

● Correlated

○ Correlated positively

○ Correlated negatively

Plots

☑ Correlation matrix

☐ Densities for variables

☐ Statistics

Correlation Matrix

Pearson Correlations

			Pearson's r	p	Lower 95% CI	Upper 95% CI
cost	-	perc_profit	0.611*	0.035	0.057	0.877

* p < .05, ** p < .01, *** p < .001

Correlation Plot

Figure 12.2: Scattergram of the data

The correlation coefficient is 0.611. Since our sample is small, the confidence interval is quite wide, ranging from .057 to .877. In repeated sampling, 95% of the correlations found would be in this interval.

Since a correlation of zero is well outside of this range, we can conclude that the correlation is significantly different from zero.

In a formal hypothesis test, the probability of finding a correlation of 0.611 in a sample of this size under the null hypothesis that there is no correlation in the population from which this sample is drawn, is 3.5% (0.035). When testing at 95% confidence (and hence α = 0.05), we conclude that the correlation between the two variables significantly different from zero. The correlation coefficient itself is a measure of effect size. The effect here is strong.

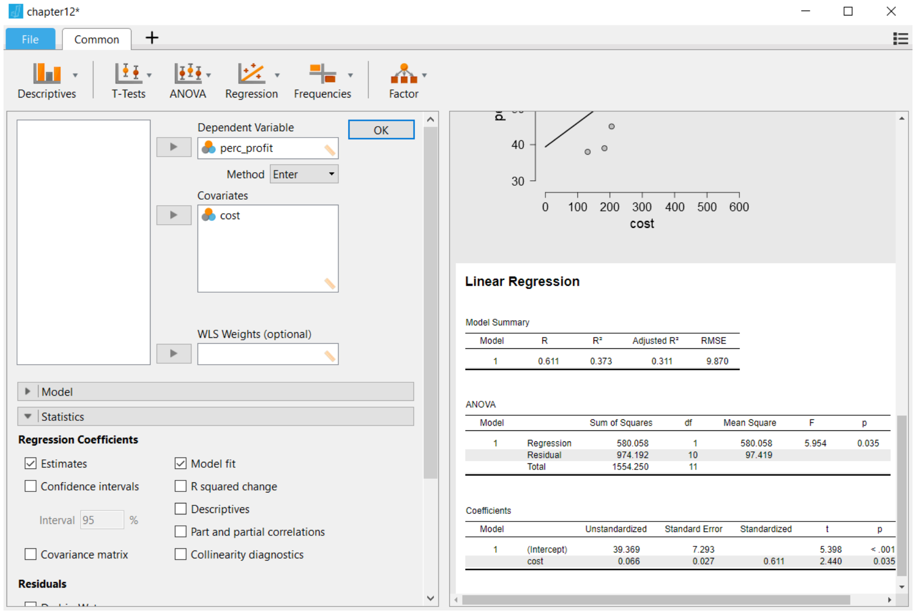
As Landers explains, it's good to compute the coefficient of determination which is the square of the correlation coefficient. The reason is that the coefficient of determination has a clear interpretation: it indicates how much of the variance the two variables have in common. Here, it's $0.61^2 = 0.37$ (or 37%).

12.2 Simple Linear Regression

In regression analysis we make a distinction between the dependent variable, and independent (explanatory) variables. In simple regression analysis we have one dependent and one independent variable, but the idea can be expanded to the situation of one dependent and two or more independent variables.

> *Regression analysis is quite flexible: we can use categorical variables as independent variables (using dummy variables, or factor variables); we can estimate relationships that are nonlinear; in the "family" of regression techniques we can also estimate models in which the dependent variable is a grouping variable. These extensions are beyond the scope of this module.*

We will stick to the fundamentals of regression analysis as discussed in Landers. Regression analysis in **JASP** is found under the **<Regression>** tab. You can click one variable to the dependent variable box, and one or more (non-categorical) variables to the *Covariates* box.



Dependent Variable: perc_profit

Method: Enter

Covariates: cost

WLS Weights (optional)

Regression Coefficients

☑ Estimates

☐ Confidence intervals

Interval 95 %

☐ Covariance matrix

☑ Model fit

☐ R squared change

☐ Descriptives

☐ Part and partial correlations

☐ Collinearity diagnostics

Residuals

Linear Regression

Model Summary

Model	R	R²	Adjusted R²	RMSE
1	0.611	0.373	0.311	9.870

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	580.058	1	580.058	5.954	0.035
	Residual	974.192	10	97.419		
	Total	1554.250	11			

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	39.369	7.293		5.398	< .001
	cost	0.066	0.027	0.611	2.440	0.035

Figure 12.3: Regression Analysis in JASP

The vital statistics are:

- The **coefficients**. The intercept is of less relevance here but "locates" the graph. The coefficient for cost indicates how much the profit percentage increases with a unit increase in cost.
- The **R²** gives us the coefficient of determination (same as the multiple R-squared as we saw earlier on). Again, we conclude that 37% of the variation in profit is explained by variations in cost.
- The **t-statistic** tells us that the coefficient for cost is significant at the 5% level (t=2.44; p=3.5%).
- The **F-statistic** tells us that the overall model is significant; since we have only one independent variable in the model, we knew that already from the t-statistic.

It is not a coincidence that the p-value for the F-statistic is identical to the t-statistic for the coefficient; the F statistic here is the square of the t-statistic. In case of one independent variable it is redundant to report *both! Only for "multiple" regression with two or more independent variables, does the F-statistic add information.*

Now it should be easy to follow Landers' discussion on page 339 and 340, on predicting values from a regression line.

Once the regression model has been estimated you can compute *predicted* values.

The regression function is:

$$Profit = 39.369 + 0.066 * Cost$$

The predicted profit for a project costing 300, is:

$$Profit = 39.369 + 0.066 * 300 = 59.169$$

12.3 Data Skill Challenge

Shane works at an ice cream store. He notices that on every warm day his boss makes sure to have extra supplies in anticipation of having a lot of customers. Shane decided to test whether he can predict the number of customers based on the temperature. If so, how many customers should he expect if the temperature is 38°C? He tracks the temperature and the number of customers for one week. His data are provided below.

(a) Compute the correlation between the number of customers and temperature

(b) Make a scattergram with number of customers (on the vertical axis) and temperature

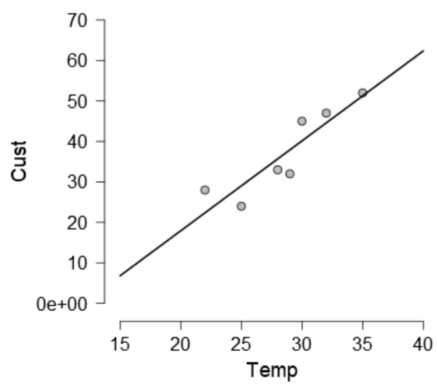
(c) Estimate the regression line; is the coefficient for temperature significant?

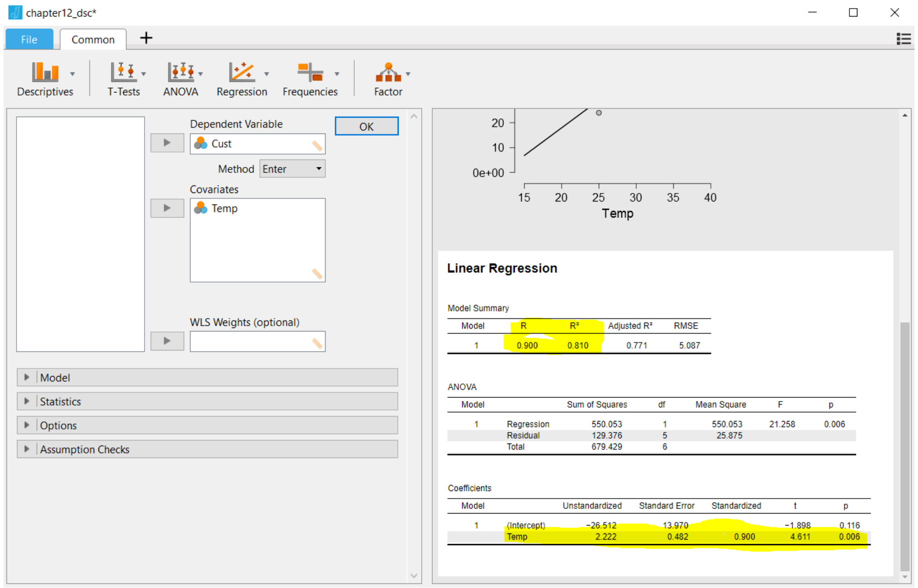
(d) Predict the number of customers if the temperature is 38°C

Temperature (degrees °C)	# of customers
22	28
25	24
29	32
28	33
35	52
32	47
30	45

Figure 12.4: Data for Data Skill Challenge (chaoter12_dsc.csv)

Correlation Plot





Dependent Variable: Cust

Method: Enter

Covariates: Temp

WLS Weights (optional):

Linear Regression

Model Summary

Model	R	R²	Adjusted R²	RMSE
1	0.900	0.810	0.771	5.087

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	550.053	1	550.053	21.258	0.006
	Residual	129.376	5	25.875		
	Total	679.429	6			

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	−26.512	13.970		−1.898	0.116
	Temp	2.222	0.482	0.900	4.611	0.006

From the output we learn that 81% of the variation in the number of customers is explained by the model. The scattergram indicates that, as expected, ice-cream sales go up with temperature.

The coefficient for **Temp** is significantly different from zero as indicated by the high t-value of 4.611. A value that high has a probability of 0.006 (or 0.6%) which is well below the 5% benchmark.

The F-statistic indicates that the regression model as a whole is significant, but with one independent variable the t-value for the one independent variable and the F-statistic are equivalent. You can see that from the identical probabilities; the F-statistic here is the square of the t-statistic. Only in regression models with two or more independent variables the F-statistic would add information.

Since the model seems to make sense we can now write the function for the regression line:

$$Number\ of\ customers\ =\ -26.512 + 2.222 * Temperature$$

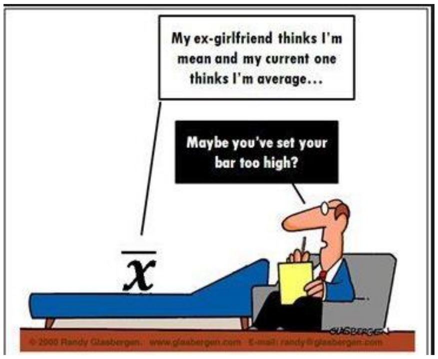
The “intercept” of -26.512 is the number of customers when the temperature drops to zero. This obviously doesn’t make sense. A negative number implies that customers are sending their back ice-creams to the store (luckily, they won’t melt at 0 degrees). The regression line is based on observations in a small range of temperatures; in our sample the temperature is in the 22 to 35°C range and we have to be very careful in making predictions for temperatures outside of that range. Temperatures of 0°C are unlikely to occur anyway, in summer – depending on where you're living of course. In general, the intercept is in most regression models not that relevant; it just "locates" the regression line. The main interest is in the coefficient for **Temp**. It is estimated that we have 2.222 extra customers with every 1 degree increase in temperature.

The predicted number of customers at a 38°C temperature then is: -26.512 + 38*2.222 = 57.924 (57 or 58, if we don't allow broken persons to buy our product).

Concluding Remarks

We hope that this manual has taught you the fundamentals of business statistics, and how to use Excel and **JASP** to perform basic analyses. The book by Landers or any other textbook on basic statistics, and this manual serve as a primer in basic statistics, and enable you to deal with the majority of statistical challenges that you will encounter in your study or profession.

For more advanced quantitative academic research, at masters or doctoral level, you might have to dig deeper. To this end, we have developed many modules to learn about other techniques, with regression analysis as discussed in chapter 12 of this manual as a starting point. For advanced statistical analysis we recommend using STATA or **R**.


My ex-girlfriend thinks I'm mean and my current one thinks I'm average...
Maybe you've set your bar too high?
$\bar{x}$
GLASBERGEN
© 2000 Randy Glasbergen. www.glasbergen.com E-mail: randy@glasbergen.com

References

Books

Field, A., Miles, J. and Field, Z. (2012). *Discovering Statistics Using R.* Sage Publications.

Landers, R. N. (2013). *A Step by Step Introduction to Statistics for Business*. Sage Publications.

Websites

https://JASP-stats.org/, last accessed 16 September 2018

Assignment

1. Here are the sales data for the branch offices of a sales organization
   - Branch 1: West, 20 units sold, € 34,000
   - Branch 2: West, 40 units sold, € 50,000
   - Branch 3: East, 16 units sold, € 35,000
   - Branch 4: East, 93 units sold, € 85,000

   Place the scores in a dataset (preferably Excel), in rows and columns as appropriate. Don't forget to add a column for the branch number. Describe the data, after reading the data into JASP.

2. Read the data of **Assignment_2.csv** in JASP. The dataset contains data on orders by representatives, in several regions of the country.

   Make a plan-of-analysis for this dataset, and report the results (to the sales manager). At a minimum, address the following issues: overall sales; sales by region and representative; distribution of order sizes; and unit price per item.

3. Given the dataset: 2, 3, 3, 1, 4, 2, 3

   a. Convert each score into a Z-score
   b. What percentage of cases would you expect to fall below 3?
   c. What score would be at the 40th percentile?

4. Your organization has conducted a consumer satisfaction survey, finding these overall satisfaction scores: 1, 5, 3, 2, 4, 1, 1, 2, 4, 4, 3, 5, 2, 3, 5, 2, 3, 4, 4, 3, 5, 4

   Describe the information, and find the confidence interval for the mean.

5. The manager of Petra, an employee at a call center, has asked her to determine whether the average number of complaints received by the workers at their branch, is different from the average number of complaints for the company overall. The company receives 24 complaints per day with a standard deviation of 5.25.

   The sample data for Petra's branch, for 5 days, are 23; 28; 34; 26 and 32.

   Conduct the complete hypothesis testing process with this dataset.

6. A regional manager implements a policy change for stores in his region to begin greeting customers whenever they are standing within a 3-meter distance in a store. He compares employee satisfaction for five employees, from pre-change to post-change. Employee satisfaction is measured on a 1-5 scale (very dissatisfied to very satisfied). He expects that his stores will have different ratings before and after the change.

   Mean employee satisfaction of five sampled employees (A-E) are: 4, 4, 3, 5, 4 (before) and 3, 2, 1, 2, 2 (after).

   Conduct the complete hypothesis testing process with this dataset.

7. Lionel is a waiter at a local diner. He notices that he earns less tips when he works the lunch shift compared with the breakfast or dinner shift. He is curious whether customers tip differently depending on the time of the day, and decides to test this by comparing the average amount tipped during each shift for one week.

The data are in the table below.

Breakfast	Lunch	Dinner
3	3	5
2	1	4
3	2	5
6	5	6
1	4	8
4	4	5
5	2	2

Conduct the complete hypothesis testing process with this dataset.

8. Sebastian manages the food services division at an amusement park. He wants to know of there is an interaction between the type of food sold and the color of the food cart. His cart currently sells hot dogs, ice cream and popcorn. He paints half of his carts blue, and the other half red. He records the number of sales for each food cart.

	Hot Dogs	Ice Cream	Popcorn
Red	10	8	15
Blue	12	22	19

Conduct the complete hypothesis testing process with this dataset.

9. Klaus is an office manager at a data entry company. He is interested in finding ways to improve employee productivity. He wonders if the number of hours worked is related to productivity. To test this, he installs software on each of his employees' computers that measures how many cells of data they enter and how long they work. The data are provided below.

# of Minutes Worked	# of Cells Entered
240	500
390	18
495	592
270	340
345	689
525	703
330	440
435	478

Conduct the complete hypothesis testing process with this dataset.

10. Design a quantitative study on your own. Think of an interesting research question, and collect relevant data. Formulate your hypothesis, and test the hypothesis using the data that you have collected!