Research Project

Master Science Education & Communication

”Girls Perform Better at School”

Statistical Facts, Figures and Fallacies in Media Claims

Author:
Eline Pasch (s4102509)

First Supervisor:
Marco Grzegorczyk
Second Supervisor:
Wim Krijnen

March 26, 2020

Preface and Reflection

As a student of the master science education and communication, with a background in mathematics, I have been interested in how numbers are perceived by people with different backgrounds. I noted that people around me often perceive facts based on numbers as the absolute truth. While I know that is is really hard to summarize data in a way that gives a good overview.
Since the media often presents numbers, without enough context, I wanted to understand how the media represents data and draws conclusions from them. But since the media is really broad, I wanted to focus on one claim in general, namely "Girls perform better at school".

This claim has always fascinated me, since it is very vague and did not line up with my view of education. I wondered if this was because I have been studying at a University of Technology, and this made me biased. Or if there are mistakes made in drawing this conclusion.

The research is interesting for journalist who want to find out how they could use statistics in their own articles. It could prevent them from misinterpreting data. Also teachers who want to make sure no unfair division between girls and boys is being made in their classes will be able to read the prove or disprove of some "facts". More general, the research could be interesting for everybody who likes to learn mathematics based on real life examples.

Since the text is not aimed at mathematicians, I had an extra challenge to make the text understandable for a broad audience. I do expects some pre-knowledge from the reader, they should be able to understand, use and interpret graphs and formulas. The rest of the mathematics used will be clarified using a lot of examples that are easy to follow and true to real life situations. These examples could also be nice tools to use for teachers who want to cover statistics in a fun and easy understanding way.

The creation and explanation of examples challenged me to dive deeper in statistical testing. I had to find possible mistakes and search for them in real life data. Giving clear examples, without losing the mathematical concept, was often challenging. Since real data often does not perfectly follow the theory.

The research also helped me to learn more about statistical testing and the underlying principles. For my bachelor, I studied applied mathematics at the University of Technology in Eindhoven. Even tough I followed a few courses in the field of statistics, I did not specifically focus on this. Therefore, a lot of thing discussed were also new for me. Which made writing this thesis it a challenging but highly educational process for me.

Facts are stubborn things, but statistics are pliable.
-Mark Twain

Contents

1 Introduction **4**
- 1.1 Motivation . . . 4
  - 1.1.1 The Power of the Media . . . 4
  - 1.1.2 Mathematical Mistakes in the Media . . . 4
  - 1.1.3 Influence of Mathematics in Media . . . 5
- 1.2 Goal . . . 5
- 1.3 Assumptions and decisions . . . 5
- 1.4 Outline . . . 5
  - 1.4.1 Layout . . . 6

2 Mathematical Background **7**
- 2.1 Averages and Percentages . . . 7
  - 2.1.1 Average . . . 7
  - 2.1.2 Percent . . . 7
  - 2.1.3 Conclusion . . . 8
- 2.2 Outline Statistical Testing . . . 8
- 2.3 Data . . . 8
  - 2.3.1 Collecting Data . . . 9
  - 2.3.2 Continuous and Discrete Data . . . 9
- 2.4 Continuous, Discrete, Nominal and Ordinal Data . . . 10
  - 2.4.1 Paired Data . . . 11
  - 2.4.2 Conclusion . . . 11
- 2.5 Graphical Representation . . . 12
  - 2.5.1 Box Plot . . . 12
  - 2.5.2 Histogram . . . 13
  - 2.5.3 Conclusion . . . 14
- 2.6 Proportion and Underlying Distribution . . . 14
- 2.7 Normal Distribution . . . 16
  - 2.7.1 Probability Calculations . . . 18
  - 2.7.2 Rule of Thumb . . . 20
  - 2.7.3 Checking for Normality . . . 22
  - 2.7.4 Q-Q plots . . . 22
  - 2.7.5 Parametric and Nonparametric Testing . . . 25
- 2.8 Testing . . . 25
  - 2.8.1 Comparing Variables . . . 25
  - 2.8.2 Performing Test . . . 25
  - 2.8.3 Type I and Type II Error . . . 25
- 2.9 Possible Tests . . . 27
- 2.10 Z-test . . . 27
  - 2.10.1 Standardization . . . 27
  - 2.10.2 Two sample z-distribution . . . 28
  - 2.10.3 Assume Null Hypothesis . . . 29
  - 2.10.4 Interval . . . 30
  - 2.10.5 P-value . . . 31
  - 2.10.6 Conclusion . . . 32
- 2.11 T-test . . . 32
  - 2.11.1 T-distribution . . . 32
  - 2.11.2 Interval . . . 33
  - 2.11.3 P-value . . . 33
  - 2.11.4 Conclusion . . . 34
- 2.12 Other Tests . . . 34
- 2.13 Conclusion Statistical Testing . . . 35

3 **Fact checking media** **36**
3.1 Subject and Article Choice . . . 36
3.2 Approach . . . 37
3.3 Assumptions and Notes . . . 37
3.4 Media versus Research . . . 37
3.4.1 Defining a Research Question . . . 38
3.4.2 Hypothesis . . . 38
3.4.3 Analysing Data . . . 38
3.4.4 Testing Hypothesis . . . 40
3.4.5 Conclusion . . . 42
3.5 Different Subject . . . 42
3.5.1 Testing . . . 43
3.5.2 Conclusion . . . 44
3.6 Conclusion . . . 44

4 **Conclusion and Discussion** **45**
4.1 Conclusion . . . 45
4.1.1 Handle Data Incorrectly . . . 45
4.1.2 Statistical and Practical Significance . . . 45
4.1.3 Generalize Research . . . 45
4.1.4 Overall Conclusion . . . 45
4.2 Discussion . . . 45
4.3 Possible Further Research . . . 45
4.3.1 Discussing More Test . . . 45
4.3.2 Checking Media Claims on the Difference Between Boys and Girls in Education 46
4.3.3 Checking General Media Claims . . . 46

5 **Appendix** **47**

1 Introduction

First, we establish the motivation to write this research, which goals are set and which decisions were made in order to fulfill these goals.

1.1 Motivation

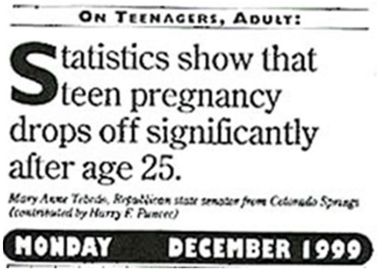
The motivation to write this, is formed by my concern that people will perceive statement based on statistics as the absolute truth. In the media, this could lead to, purposely or not, misinforming a big audience.

1.1.1 The Power of the Media

The fact is, people will learn about current event mostly from media. Therefore it has a very big influence on how people make decisions about social, political or economic issues. The main purpose of the media should therefore be to present the public with information that is as close to reality as possible. Since misinformation could simply lead to decisions that would not have made if they were informed differently.(Başlar, 2011)
This also means that media owners could purposely mislead people. They could use their platform to distribute too little or even wrong information to promote their own interests. (Başlar, 2011)

1.1.2 Mathematical Mistakes in the Media

It is therefore important to critically think before following something that is presented by the media. However, people might think that if a claim is supported by numbers, it must be factual, and therefore, the conclusion in the article is right. But numbers could be interpret incorrectly. Let's take a look at a mathematical mistakes in the media.



ON TEENAGERS, ADULT:

Statistics show that teen pregnancy drops off significantly after age 25.

Mary Anne Tebedo, Republican state senator from Colorado Springs (contributed by Harry F. Puncec)

MONDAY DECEMBER 1999

Figure 1: Example of a mistake in the media (Conquermaths, 2013)

This is a clear example of a mistake that is quickly noticeable. A reader of such a headline will not need a profound mathematical understanding to understand that teen pregnancy only occurs among teens.

But even if we interpret the statement mathematically, it is really dubious. Indeed, there are no teen pregnancies after 25. But the word "drops" claims that until the age of 25, there are teen pregnancies. This would leave me with some questions. For example, why would they choose the age of 25? Why would they say drop? Which statistics did they use to come up with this claim?

Surely, this is an article headline that is just very wrong and probably will not influence the way you think about teen pregnancy. But the misuse of numbers does not only happen in such clear examples.

Research done by Maier (2002) focused on the use of mathematics from a daily newspaper. He concluded that nearly half of local news stories involved mathematical calculations. In his report he talks about eleven different categories of numerical errors among them. From this he concluded that journalists fail to apply the attention and skepticism to numbers overall.

1.1.3 Influence of Mathematics in Media

In conclusion, the media has a influence on somebodies view on a certain topic and journalist make mathematical mistakes. Some clarity should be provided on which numbers are correct and which are not. Therefore, I decided my research on mathematical facts, figures and fallacies in media claims.

1.2 Goal

The goal of this project is to correctly handle the data. It in order to fulfill this, the following subquestions will be looked at.

- How do journalist handle statistics and data now? What are mistakes that are often made?
- What is a right way to handle data? What statistics could be used to draw correct conclusions? What real life situations could link the mathematics to the real world?
- Are the conclusions drawn in previous research right? How is the data collected handled? What statistical tools are applied in order to draw conclusions? What could have been done differently?
- How could somebody with no programming experience work with these statistical principles?

1.3 Assumptions and decisions

In order to successfully provide answers to these goals, the following assumptions and decisions were made:

- Assumed will be that journalist want to provide right conclusions. They will not let their own bias about a subject determine what part of a research will be published.
- A journalist knows how to provide the data within an article.The research will not focus on how to make an article appealing or how to present data in a entertaining way. It will only focus on drawing the right conclusions.
- Since the field of statistics is really broad, only a part of this can be covered. Determined is that the research will focus on descriptive statistics, which means that the goal will be to make sure journalists are able to correctly interpreted and show data, but not to do any predictions based on them. More specific, the comparison of two groups will be covered. Since it is often interesting to compare, for example, different countries, sexes or political preferences.

1.4 Outline

This research could give interesting insights for journalist, educators and others who are interested in using statistics for their own research. As a reader, you can determine to therefore skip to certain topics.

Chapter 2 provides a mathematical background starts by explaining the (mis)use of statistics in the media and continuous to explain a right way to approach numbers. This will be done by explaining different types of data and what the correct way is to approach these data. Some basic statistical testing will be introduced and explained.

Chapter 3 will focus on applying the theory to fact check media statements on whether girls perform better at schools than boys do. This will be done by studying research based on this topic.

In chapter 4, an introduction in R, a statistical tool, will be provided. And finally, in chapter 5, a conclusion and discussion can be found.

1.4.1 Layout

Throughout the research the following layout will be used:

Example

In the green text boxes, examples that will clarify the mathematical principle will be discussed. When you are interested in explaining a concept to others, for example students, these examples will be extra useful.

Concept

In the blue color boxes, a mathematical term, concepts or formula is explained.

Code R

The yellow boxes consist of a piece of code in the statistical tool R. When you read this without programming experience, it is no problem to skip the code. The conclusions will be able to understand, without understanding the code. If you are interested in doing such a calculation in R yourself, you could read about it in chapter **??**.

glossary
When you find a word in italic, it is a mathematical therm or concept that might need some more explanation. The definition of such a word could therefore be found in the *glossary*. If you are reading this on a computer, you could click on the word, and it will guide you to the *glossary*.

5
When the number 5 is found after a picture or number, the R Code written to create this picture can be found in the appendix (chapter 5). Clicking this number will guide you to the code when a computer version is viewed.

2 Mathematical Background

This chapter will start of by showing why the measures used in media nowadays do not portray enough to base a rightful conclusion on. It will show what possibles ways there are to provide more information about data by introducing the concept of statistical testing.

2.1 Averages and Percentages

Journalist are advised to work with averages and percentages. (Kille, 2014) Those are not wrong per definition, but these numbers could easily be manipulated to proof the point of the author.

2.1.1 Average

For example, when talking about average; *mean*, *median* and *mode* could get interchanged. This example will show the difference.

Example 1.1

Let's say you are a party planner and you receive three tasks:

- The first one is that you have to plan a trip for a group of women who have a mean of 33 years old. You think that planning a wine tasting would be fun.
- Secondly, you have to plan a trip for a group that has a median of 16 years old and plan for them to go see a horror movie in the cinema.
- Finally, you have to plan a trip for a group where the mode is 13, for them you think about doing something active, like a hike.

Now, let's say that all this information was about the same group, namely a group of grandmas and granddaughters. There is one grandma of 66 with twin granddaughter of 13. There is a grandma of 69 with a granddaughter of 15 and a grandma of 64 with a granddaughter of 17 and 7.

All of a sudden, none of the activities decided on seem appropriate anymore. So, naturally, we find that looking at averages in all shapes might not be sufficient in order to draw a conclusion about a group.

2.1.2 Percent

Also the term *percent* could give some difficulty. Mainly because it, just like average, does not include anything about the size or distribution of the group.

Example 2

This example shows that a *percentage increase* might lead to a wrong conclusion.

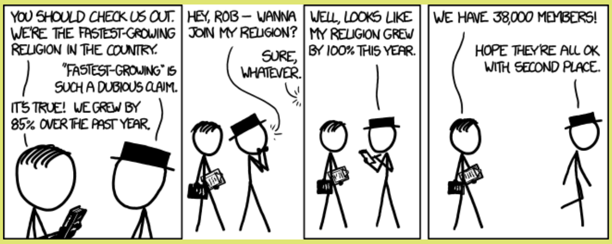

YOU SHOULD CHECK US OUT. WE'RE THE FASTEST-GROWING RELIGION IN THE COUNTRY.
"FASTEST-GROWING" IS SUCH A DUBIOUS CLAIM.
IT'S TRUE! WE GREW BY 85% OVER THE PAST YEAR.
HEY, ROB — WANNA JOIN MY RELIGION?
SURE, WHATEVER.
WELL, LOOKS LIKE MY RELIGION GREW BY 100% THIS YEAR.
WE HAVE 38,000 MEMBERS!
HOPE THEY'RE ALL OK WITH SECOND PLACE.

Figure 2: Fastest Growing Religion. (Xkcd, n.d.)

2.1.3 Conclusion

With what is advised to journalist at the moment the main problem is that they do not depict anything about the group size and the distribution of values within the group. This shortcomings should be solved in order to make sure no wrong conclusions are drawn.

2.2 Outline Statistical Testing

So, we found that only using percentages and averages is not always enough to provide a good image of the entire situation. Different statistical approaches could be used to draw some more interesting conclusions. Let's take a look at the process of *hypothesis testing*, which is very common in research nowadays.

First, a researcher will ask what should be tested. This researcher will already have a possible answer and he/she will use that to formulate the *null hypothesis* and *alternative hypothesis*. The null hypothesis is then a claim that should be tested and the alternative hypothesis is the opposite of this statement. (Diez, 2019)

Then, the researcher will collect data and try to find how likely it is that he found this data, when the null hypothesis is indeed true. In other words he is going to test if his ideas are supported by the real world.

Example 3.1

Imagine, you teach mathematics to a class of second graders, you wonder if the girls performed better at the last test than the boys did. So you want to answer the **research question**: "Did the girls perform better at the last test than the boys did?". You expect that the the girls did do better, and therefore you set the following **null hypothesis:**, "The girls of the second grade scored better at the last test than the boys of that class did". Logically, you might think that the alternative hypothesis claims that the boys did perform better, but there is also a possibility that they both performed equally. Therefore, the **alternative hypothesis:** is "The girls did not perform better".
In order to find the answers you decide to take a look at the grades and split them up between the boys and girls, this will then be your **data**. All girls got 9 out of 10, while all boys scored 6 out of 10. In this case you would find that the data provided supports your null hypothesis. Now, you are able to answer your question; the girls did perform better at the last test.

In example above, it is clear that the null hypothesis cannot be rejected, but sometimes, the results are not as convincing.

Example 3.2

Look at the situation from the last example, but now, you would find that there are only five girls in the class, they on average scored a 6.4, while the 23 boys got an average of 6.3. Then, you might argue that the girls scored better, but there were less of them, so more boys got a high score. So what should you conclude now? And will this conclusion change if you know that no girls scored insufficient, but some boys did?

This last situation is an example of a situation where statistical testing could give an answer. In other words, statistical testing is used to conclude if a certain claim could be true based on the data found. In order to determine what statistical test is the best for the data set provided, the different types of data are important to distinguish.

2.3 Data

All information coming from field notes, surveys and experiments belonging to a scientific experiment could be considered data. (Diez, 2019) This is a very broad definition. Which also means that not all data is automatically useful. So, let's dive deeper in collecting data.

2.3.1 Collecting Data

When reading or writing an article, it is important to check if the conclusions drawn, are done for the correct group. Such a group that is investigated is called a *population*. It is often not possible to collect data from everybody in this population. Instead, data will be collected from a small group of the population; the sample group. It is important that this sample group represent the view the population.

Example 4

For example, you want to publish an article because you worry that girls are privileged in the Dutch schoolsystem. You find a research done in India that investigated the difference between boys and girls in school. It might be tempting to use information from this article, but it is perfectly possible that the Indian school system approaches girls in a different way than the Dutch system. So the conclusions drawn on the population of Indian school children does not represent the Dutch school children.

So in this case, the entire sample group does not represent the population. But it is also possible to sample a group from a population that does not represent the view of the entire population, because there was some bias in picking the sample.

Example 5

For example, you are interested in the favorite sport of people in the Netherlands. You decide to ask all people that arrive at a soccer club on Monday between 10am and 12am.

Very likely all people visiting a soccer club will have soccer as one of their favorite sports. Also, on Monday morning, a soccer club will not be highly visited. Therefore, the test group will be really small and have a bias. Both things that will make sure that the results are not representative.

Since journalist often do not have to collect your own data, we will not dive deeper into this topic. But if you do want to collect your own data, it is also important to check that the sample group represent the population. Welkowitz (2011) will go into more depth on the collection of data.

2.3.2 Continuous and Discrete Data

Since data is a very broad term, some specification is needed.

Example 6

You are being asked to calculate:

The *mean* ..

- .. length of people in 5th grade.
- .. amount of people in a class of all primary school classes.
- .. happiness of the students at a certain school.
- .. gender of all students.

.

You probably feel that you could do the first and second calculation, when the lengths and amounts are provided by adding all numbers and dividing it by the total number of items in the data set. But the third and fourth item will be harder. There is no fixed number that links to happiness.

And there is no such thing as an average gender.

This short example already show that there are different types of data, based on whether you could calculate, for example, an average. Simply said, data where calculations could be based on are called *numerical data*, while non-numerical data is considered *categorical data.* (Diez, 2019)

2.4 Continuous, Discrete, Nominal and Ordinal Data

A more specific distinction could be made. The following chart gives an overview.

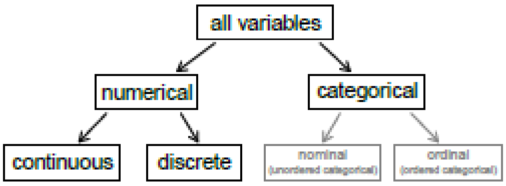


Figure 3: Distinguish Different Kinds of Variables of Data. (Diez, 2019)

Let's take a look at examples of these different types.

Continuous Data

Continuous data could take all values. It is hard to investigate true continuous data, since all measuring tools do have the restriction of having a certain amount of decimals they work with. For example, the length of a person, is considered continuous, while, you will always round up the length up to two decimals.

Discrete Data

Discrete data could only take values with jumps. For example the amount of people, because you will never have half a person.

In reality, the difference between continuous and discrete data might be hard to determine.

Oridinal Data

Ordinal data has a natural ordering, but is not numerical. For example, performance tools like the one below. Since you could give a clear distinction of the order of the smileys. You could say that one is better than the other, but you cannot tell how much better it is exactly.

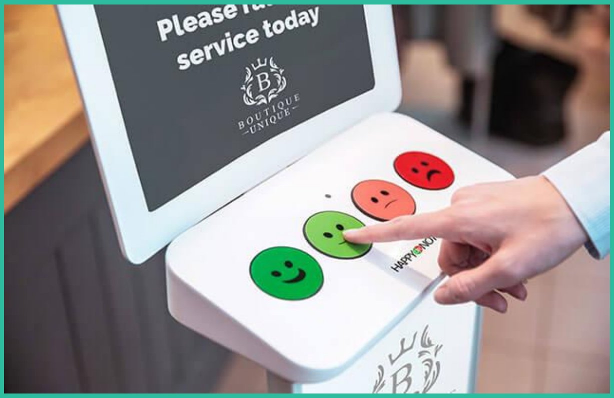


Figure 4: An Example of the Collection of Ordinal Data. (Happy or Not Ltd, 2019)

Nominal Data

Nominal data are data that do not have a natural ordering. For example "men" versus "women". Note that also postal codes could be considered nominal, even though they do exist of numbers. But is is not possible to work with these number, in the sense of adding, subtracting or ordering. So, the fact that the variable exist of numbers, does not automatically make it numerical.

A data set could consist of more than one type of data. For example, listing the gender and grades of a school class will be a collection of *Continuous Data* and *Nominal Data* and could perfectly be used to do some interesting calculations.

The main focus for statistical testing will be on data sets that contain *numerical data*. Simply, because of the calculations that can not be done with *categorical data*. This does not necessarily mean that you could not draw any interesting conclusions from purely *categorical data*. If you for example collect ordinal data, with the smiley system presented before, and you find that a lot of people press the red, unhappy smileys, then you can conclude that something is not going well.

2.4.1 Paired Data

Now, let's look at *paired data*. Whether you need to collect *paired data* or not really depends on the research question.

Example 7

For example, you want to do an experiment on how far children can throw a ball.

If you want to answer the research question "What is the average length a child can throw?". Then you could randomly invite children. The only thing you need to write down is how far they threw. This is not paired to anything. You can now only draw conclusions from the entire data set.

If you want to answer the research question "Do boys throw further than girls?". Then you could randomly invite boys and girls to throw a ball. The data set you would get then does consist of the gender and length, there is only one observation done per person, so the data is not paired. You could then compare the entire group of girls, to the entire group of boys.

If you want to answer the question is "Do boys throw further than their sister?", you do need to make pairs. Then you namely have two different observations from a single family. This is an example of paired data.

So, paired data does consist in different forms. Even similar looking experiments, might need different data formats, therefore it is important to always first formulate your question, before randomly collecting data. When you want to compare two observations from a single source, you will need paired data.

In this research, we will not propose any statistical testing based on paired data. So hence, you might assume that all data discussed is unpaired.

2.4.2 Conclusion

Data could occur in a lot of different forms. Which form of data you need depends on the research question. Therefore, it is important to think about statistical testing before performing an experiment.

2.5 Graphical Representation

Now, some broader explanation about data has been done, we can start with interpreting this data. In the example of the grandmas and granddaughters discussed in the beginning of this chapter, knowing how the data was distributed, could have prevented sending a 7 year old to a wine tasting. There are some graphical tools that could help us finding this distribution.

2.5.1 Box Plot

A common way to visually represent data is a box plot. It is a tool that is useful to compare the distribution the data of two groups. (Lane et al., 2019) In order to draw a box plot, all data will be ordered from low to high. Then the data has to be divided in 4 groups, were all groups consist of the same amount of data. So, between the lowest line, and the second line, the lowest 25 percent of all data is represented. This is then called the first quartile. Then, the middle line represents the value where 50 percent is lower and 50 percent is higher than that value; this value is the *median.*

If the distance between two lines is small, that means that there are a lot of measurements in that group. In other words, the values in that part are often occurring in the data set.

When two boxplots are drawn next to each other,it is easy to determine if the way the data is distributed in a similar way. This could be done by looking at the minimal and maximal value, but also by comparing the size of the boxes.

Example 1.2

This example will continue on the group of grandmas and granddaughters, which had the ages 66, 13, 13, 69, 15, 64, 17, 7. We will use this to draw a boxplot.

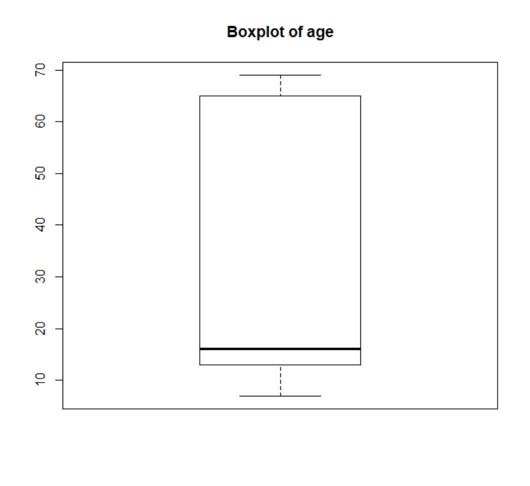


Figure 5: A Boxplot of the Age of Grandmothers and Granddaughters. 5

So,from this, we could conclude that half of the people in the group are between 13 and 17. Something that could not have been found from the averages provided before. A disadvantage of this representation is that we cannot find that is only about 8 people. Also, it does not represent the distribution within a single quartile; we are not able to tell that there are no people between the age of 17 and 64. Therefore, it is a simple way to represent data, but there are also some shortcomings.

2.5.2 Histogram

A *histogram* is an other visual interpretation of data, which just represent the number of data within a certain interval. The peaks are the most important to investigate, since they tell something about where the most observations can be found. The histogram could show one, two or more prominent peaks. The distribution could be described as *unimodal*, *bimodal* or *multimodal* with respect to the number of peaks respectively.
The location of the peak is important as well. The distribution could be *symmetric*, *right-skewed* or *left-skewed*, which describes where the location of the tail of the distribution can be found, in other words, it describes which half has the biggest part of the data.

Example 8.1

This is an example of a*right-skewed bimodal* distribution based on the magnitudes of earthquakes from November 2019 (USGS, 2019).

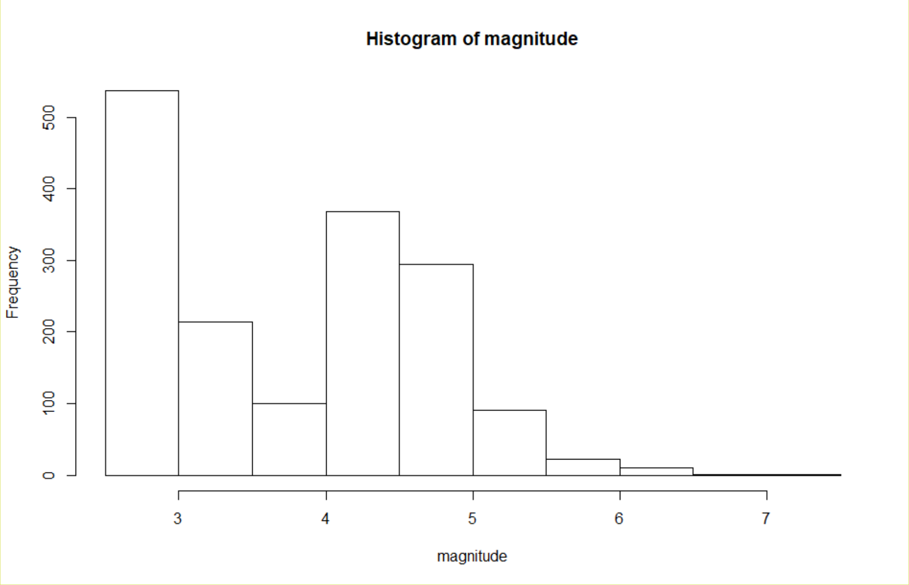


Figure 6: Histogram of the Magnitude of Earthquakes Worldwide in November 2019. 5

Note that this does not mean that there are no shocks with lower magnitude, but earthquakes are not felt then.

You might expect that there would be an *right-skewed unimodal* distribution. But not all seismographs are able to record a event with magnitude lower than 4.5. (USGS, 2016) This shows that a histogram could also quickly show that there might be missing data that gives us a false feeling that there are less earthquakes with a magnitude between 3.5 and 4, than between 4 and 4.5.

Example 1.3

Let's take a look at the ages of the ladies from the examples used before.

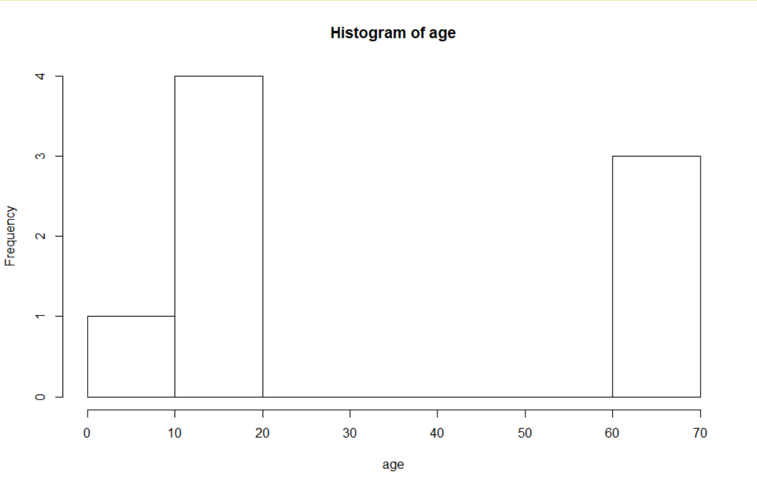


Figure 7: Histogram of the Age of Grandmothers and Granddaughters. 5

Granted, this group is only 8 people, so portraying them in a histogram might not be easier than just giving a list of the ages. But when there is a big amount of data, this will be a neat and easy was to portray how the data is distributed. In this case, we mainly understand that the ages of the ladies are not perfectly centered around an average. Something that might be misleading about the term average.

So overall, histograms are a simple tool that show something about the distribution of data. If the distribution does not seem logical, it is good to question if the data is correctly collected.

2.5.3 Conclusion

Box plots and histograms could both be used to show how data is distributed. There are more ways to represent numerical data, for example frequency polygons or line graphs. Chapter 2 of the book written by Lane et al. (2019) goes into depth into these graphical distributions. It also provides options on how to provide graphic representations of categorical data. All graphical representations could be useful to search for underlying distributions but could also be helpful to represent data to the public.

2.6 Proportion and Underlying Distribution

Histograms and box plots are a nice way to represent the distribution of the data. Now, we have to find a way to describe this distribution.

Example 9

You provide a course to 1000 students, after the course, you ask 100 random students to rate the course. 10 students rate the course with a 9 out of 10. Then you expect that of all students, also about 10 percent would rate the course a 9. Since all students where selected at random.

Reasoning like this will take us from a exact number to a proportion. This number tells us something about the distribution from the collected data.

Now, we can apply this way of reasoning with proportions, to create a histogram of the earthquake data.

Example 8.2

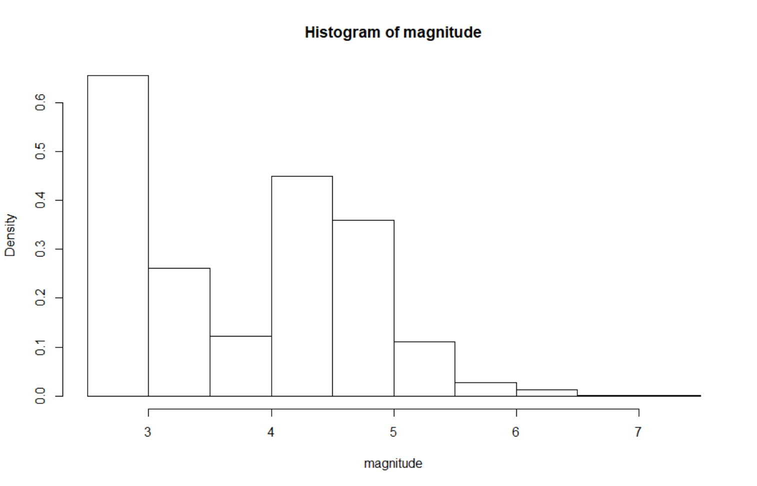


Figure 8: Histogram of the Magnitude of Earthquakes Worldwide until November 2019. (USGS, 2019) 5

The shape is still the same as the histogram found with the numbers, only the values along the y-axis changed by dividing the frequency by the total number of observations. If the data would have been correct (which we already concluded was not possible in the previous chapter) this histogram might even help us conclude that an earthquake has a magnitude between 2 and 3 with a probability of some more than 0.6.
Unfortunately, such an exact probability can only be precisely calculated when the number of experiments done is infinite. Right now, it might just be a coincide that we found these proportions. Let's look at this example.

Example 10

When you would role a fair dices 600 times, you would expect that all numbers are thrown 100 times. We performed such an experiment 2 times and got the following results:

	1	2	3	4	5	6
Experiment 1	108	107	91	106	89	99
Proportion	0.180	0.178	0.152	0.177	0.148	0.165
Experiment 2	107	95	90	94	102	112
Proportion	0.178	0.158	0.150	0.157	0.170	0.187

We know that both experiments have been done with the same dice, so it is impossible that the underlying probability distribution could have changed between experiments. But we do notice that we would get different conclusions about distribution of the proportion for the same dice, when it is based on an other experiment. Since we know the distribution of a dice, it it easy to notice that we just do not get the 100 for every number time, because the dice does not count how often it has been showing one of the numbers.
In real life, we often do not know the underlying distribution, and we are looking for this.

If you repeat an experiment more often, the distribution found from the data will tend to go to the real distribution. This is called the *law of large numbers.* In order to find the real probabilities that make up an distribution, the experiment should be repeated an infinite number of times. Which is not really possible. So there is most possibly a difference between the probability distribution.

But also, there is a difference every time you tend to sample from a distribution, this difference is called a *sampling error.*
This is also why you will have two different notations for *mean.* $\bar{x}$ is the mean of the sample and μ is the mean of the populations, so the mean of the underlying distribution. A list of more variables in the appendix (5) will help you if you get lost in the notations.

So, just because the results are not the same, the underlying distribution of the population can be. This is something that is really important to understand in order to understand why statistical testing would even be used.

2.7 Normal Distribution

When the distribution found from the histogram is symmetric, *unimodal* and bell-shaped, we can try to fit a normal distribution to the data. The formula for the normal distribution is as following:

$$N(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Where μ represent the *mean* and σ the *standard deviation* of the population. (Lane et al., 2019)

Many distributions found in data are nearly normal, but exact normality will not occur in real data. Mainly because a normal distribution is asymptotic, which would mean that even for values very far from the mean, a small density of the population can be found. While real life processes often have a natural occurring limit (Diez, 2019). For example, a person can not have a negative length.

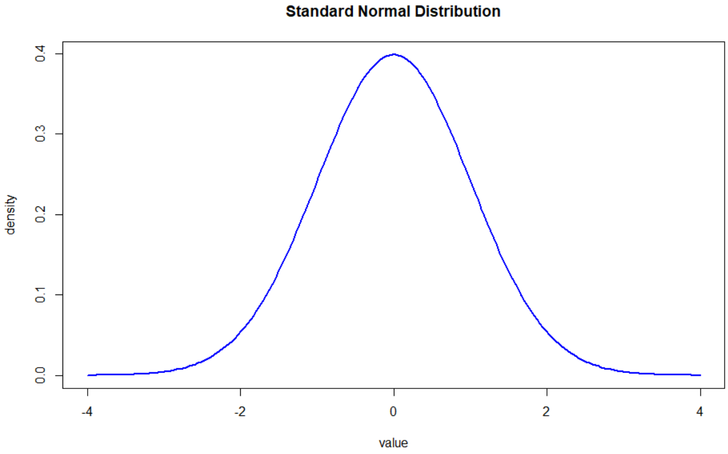


Figure 9: The standard normal distribution

This standard normal distribution has a *mean* of 0 and a *standard deviation* of 1. By changing mean and standard deviation, this standard distribution can be altered to fit other data. The overall shape will still stay symmetric, unimodal and bell-shaped. Changing the mean shifts the bell to left or right, changing the standard deviation stretches or constricts the curve. (Diez, 2019)

The fitting of the normal distribution will be discussed using an example.

Example 11.1

Let's take a look at the following headline.

Iceland's First Baby Of 2020 Is One Of The Biggest Ever

Figure 10: News Article Headline About a Very Big Baby. (Pereira, 2020)

When reading into the article, we discover that the baby mentioned is a shocking 5,9 kilograms. But how unlikely is this really? We can link this information to what we know about the normal distribution.

The normal distribution will be used to approach birth weight. Which has been measured by O'Cathain et al. (2002). A histogram of the events is shown in green in the figure below.

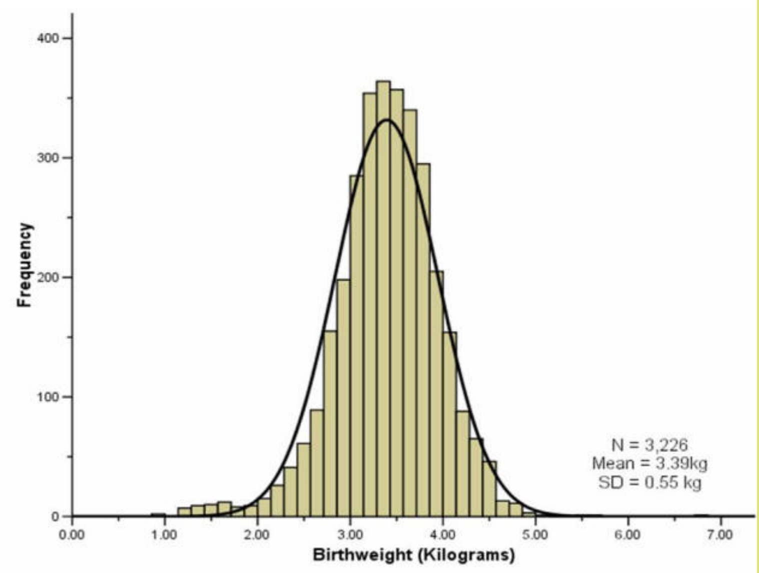


Figure 11: Normal Distribution of Birth Weight of Human. (Campell & Shantikumar, 2016)

If we try to explain this figure we see that the average child in the experiment weighted about 3.39 kilograms. A lot of babies will weigh around this weight. How further away from the 3.39 kilo we look, how less likely it will become to find a baby of that weight.

In this case the normal distribution created to fit the data has a *mean* of 3.39 and a *standard deviation* of 0.55. Both these numbers were calculated from the data, and then filled in into the formula for the normal distribution and are shown by the black line in the figure. Note that this indeed forms a transformation of the standard normal distribution; the shape is still similar, while the location and width are different.

We already feel from the picture that a baby of 5.9 kilograms is indeed very unlikely, but since we assumed an underlying normal distribution, we could even calculate how unlikely this event really is.

2.7.1 Probability Calculations

We concluded that a baby of 5.9 kilograms is "very unlikely", but that does not seem very mathematical yet. Therefore we want to express this unlikeliness with a number; a *probability.* Which in this case is the proportion of a baby of 5.9 kilogram among the population.

probability

Let's take a look at the probability notation.

An example of a probability notation is $\mathbb{P}(X = a)$
Simply put, X is a certain event and a is a possible outcome of this event. So we are interested in the probability that the outcome of something is a.

Example 11

Let's take a look at a fair, normal die. X is the number thrown by the dice. Then $\mathbb{P}(X = 1)$ is the probability that we throw a 1 with the dice. Since there 6 numbers on the dice, the probability of throwing a 1 is 1 in 6, so $\mathbb{P}(X = 1) = \frac{1}{6}$

An other thing we could calculate is $\mathbb{P}(X < a)$
Were we could again see X is a certain event, and a as a possible outcome of this event. Now, we are interested in an outcome lower than a.

Example 11

Let's take a look at a fair, normal dice. X is the number thrown by the dice. Then $\mathbb{P}(X < 3)$ is the probability that we throw a number smaller than 3, so the probability that we throw either 1 or 2 with the dice. Since there 6 numbers on the dice, the probability of throwing a 1 or 2 is 2 in 6, so $\mathbb{P}(X < 3) = \frac{2}{6} = \frac{1}{3}$

Note that the number thrown by a dice is discrete, this means that we would have calculated $\mathbb{P}(X \leq 2)$*, we would also calculate the probability of throwing 1 or 2. In continuous cases, this does not work.*

The normal distribution is not discrete, this does not mean that the data underlying of this distribution does not consist of any rounding up, but it means that theoretically, all values could exist. That is why we talk about a *continuous probability distribution.* In general for a *continuous probability distribution* the following hold.

$$\mathbb{P}(a < X < b) = \int_a^b f(x)dx$$

Were $f(x)$ is a *probability density function.*

This means that in order to find the probability that a certain event is between a and b, the integral over the *probability density function* should be taken. Which basically means that we are interested in the area under the
textitprobability density functionin order to find the probability. If we apply this on the normal distribution, we find:

$$\mathbb{P}(a < X < b) = \int_a^b f(x)dx = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Which means that we calculate the brown area in the following figure to find $\mathbb{P}(a < X < b)$. Which is indeed the proportion of the distribution belonging to that area.

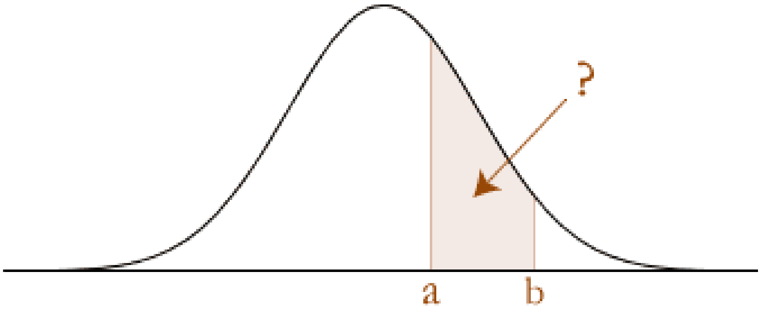


Figure 12: Possible Area Between a and b of a Normal Distribution. (Lund Research Ltd, 2018)

An other notation often used in order to show that you want to find the probability in a normal distribution is $\Phi(a)$. Which stands for:

$$\Phi(a) = \mathbb{P}(X < a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

Which equals the blue area in the following figure.

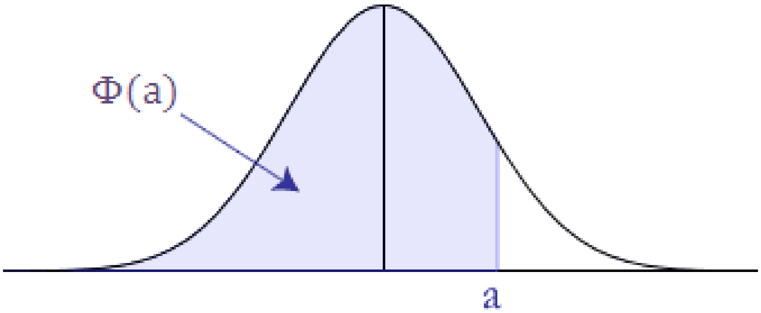


Figure 13: Visual Interpretation of $\Phi(a)$. (Lund Research Ltd, 2018)

Since we are looking at the area under the , some interesting probabilities follow.

Probablity ules

There are some rules that follow if we look at this area intepration.

The total area under the normal distribution is 1, so

$$\mathbb{P}(-\infty < X < \infty) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1$$

The area in a point does not exist, so

$$\mathbb{P}(X = a) = 0$$

and therefore

$$\mathbb{P}(X < a) = \mathbb{P}(X \leq a)$$

From all this the following follows:

$$\mathbb{P}(X > a) = 1 - \mathbb{P}(X < a)$$

Note that these rules apply for all continuous probability functions.

So, now we found some rules and a difficult looking integral. And it might seem impossible to calculate a probability. Well, doing it by hand is not something you want to try. Using a calculator, online tool or programming tool is way easier. Chapter 4 will continue on the possibility of using a programming tool to do such a calculation. But it is quite good to understand the underlying principle in order to later on understand the tests based on it.

Example 11.2

So, theoretically, the probability of a child being born at 5.9 kilograms exactly is 0, $\mathbb{P}(X = 5.9) = 0$. Therefore, this will not be something that is interesting to calculate. Instead, we are going to calculate the probability that a baby is born at 5.9 kilo's or more. So we assume a normal distribution with mean 3.39 and standard deviation of 0.55, based on the experiment done. Looking at it mathematically, we should calculate the following integral

$$\mathbb{P}(X \geq 5.9) = \mathbb{P}(5.9 \leq X < \infty) = \int_{5.9}^{\infty} \frac{1}{\sqrt{2\pi \cdot 0.55^2}} e^{-\frac{(x-3.39)^2}{2\cdot(0.55)^2}} dx.$$

We are definitely not going to solve this by hand. Instead, we will use the rule we found

$$\mathbb{P}(X \geq 5.9) = 1 - \mathbb{P}(X < 5.9).$$

Since most calculation tools can be easily used to calculate $\mathbb{P}(X < 5.9)$

Code R

```
> pnorm(5.9, mean = 3.39, sd = 0.55, lower.tail = TRUE, log.p = FALSE)
 0.9999975

> 1-pnorm(5.9, mean = 3.39, sd = 0.55, lower.tail = TRUE, log.p = FALSE)
 2.513756e-06
```

We find:
$\mathbb{P}(X < 5.9) = 0.9999975$
$\mathbb{P}(X \geq 5.9) = 2.513756e^{-06}$

So we find that the Icelandic baby belonged to the top 0.0002513756 percent heaviest babies, which makes him indeed very exceptional and newsworthy.

As a journalist calculating numbers like this might be adding to your story, since "one of the biggest" might be less convincing than "it belongs to the heaviest 0.0002513756 percent". This is of course only possible when something about the underlying distribution is known.

2.7.2 Rule of Thumb

The normal distribution can be picked as underlying distribution,to grade natural occurring event. Let's take a look of the example of the IQ test.

Example 12

IQ is viewed as a measure of intelligence, but mainly represents how well you reason and solve problems compared to other people of your age. There are different test that could produce such an IQ score, for example, the Wechsler Adult Intelligence Scale and the Stanford-Binet test. Both tests results are produced by fitting the test results of a representative sample and then match the *mean* score to 100 and *standard deviation* to 15. Which will eventually lead to test that use the following IQ score system. (Cherry, 2019)

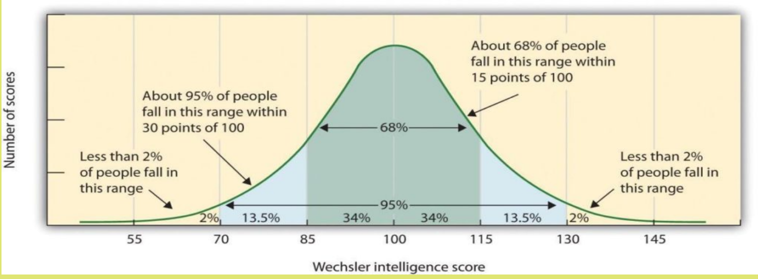


Figure 14: Distribution of IQ Scores. (Lumen Learning, 2019)

So the average IQ score on all tests is about 100. But there is more that could be noticed from the figure, namely that 68 percent of the scores lies somewhere between 85 and 115. 85 is 100-15, so the mean minus the standard deviation. 115 is 100+15, so the mean plus the standard deviation.

95% of the test scores will fall between 70, $100 - 2 \cdot 15$, and 130, $100 + 2 \cdot 15$. And 99.7% falls between 55, $100 - 3 \cdot 15$ and 145, $100 + 2 \cdot 15$.

Rule of Thumb

This connection between these percentages and the amount of times the standard deviation from the mean is not specific for IQ. Instead, it is something that holds for all normal distributions.

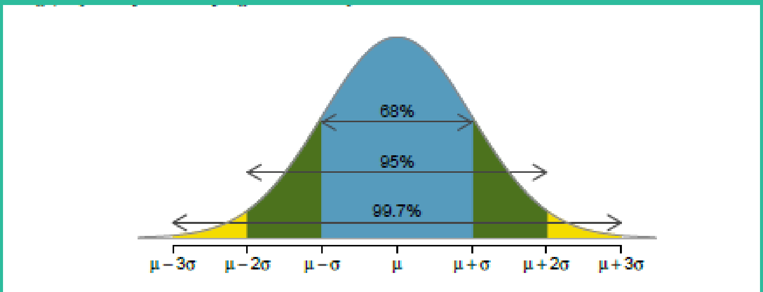


Figure 15: Probabilities and Standard Deviation. (Diez, 2019)

So, we found that the normal distribution is a distribution with a mean. Around this mean, the most data from the data set can be located, how further away from this mean, how less often we find a measuring. We even know that theoretically almost everything (99,7 %) could be found within a difference of 3 standard deviations from the mean.

2.7.3 Checking for Normality

The normal distribution could often be fitted to processes occurring in nature. So some statistical test are designed with the normal distribution as the underlying distribution, this will be discussed later on. So, we concluded that there is no real normality occurring in the real world, but that a lot of samples will somewhat follow this normal distribution. This "somewhat normal distribution" is of course very vague and not mathematical. Therefore, we want to specify this.

Assuming normality is the solution. This basically says that we claim that there is a normal distribution with the mean and standard deviation of the sample and that we believe that this is true, unless we find any strong evidence that this distribution is not appropriate.

Note: The article written by Mordkoff (2000) explains why this assumption is often a good guess by introducing some mathematical theorems, like the Central Limit Theorem. It might be interesting to read when you are interested in a broader mathematical background.

This evidence to support the normality claim could either be based on statistical tests like the Kolmogorov-Smirnov or the Shapiro-Wilk test. Which will not be further discussed. Instead, we will use an approach based on *descriptive statistics.* (Mordkoff, 2000)

The method we will use is called *quantile-quantile plots*, but often the short form, q-q plots, is used. This method is not only used to check for normal distributions, but could also be used in order to check for other underlying distributions. (Lane et al., 2019) We will take a look at the application of them in the case of normal distribution specifically.

2.7.4 Q-Q plots

In general, the basic idea is to compare the theoretically expected quantile from a normal distribution with the quantile found in a data set. Quantiles are subgroups with equal probability. (Lane et al., 2019). So if we have to divide a group into two quantiles, each group will consist of 50 percent of the observations. If we have to divide it into four quantiles, each group will have 25 percent of the observations. Etc.

Let's first determine the theoretical quantile. To get the theoretical quantile we should take a look into the normal distribution.

$$q = \frac{i - 0.5}{N}$$

Where i is a number from 1 up till N and N is the total number of observations. This formula is build up in such a way that the percentage of the underlying distribution between any two up following observations is similar.

From the quantile, we want to know what part of the normal distribution belongs to this. We will do this with the inverse of the phi-function. Since this will bring us from the area to the value of the normal distribution.

$$\Phi^{-1}(q)$$

Example 13.1

We did an experiment with 9 observations in total. We want to know the score that belongs to the 5^{th} observation.

$$q = \frac{5 - 0.5}{9} = 0.5$$

Logically, we expect half of our observations to be smaller than the 5^{th} observation, and half of the observations to be bigger than it.
Now, we are interested in the value that belongs to this using a standard normal distribution

$$\Phi^{-1}(q) = \Phi^{-1}(\frac{5 - 0.5}{9}) = \Phi^{-1}(0.5)$$

Code R

```
> qnorm(0.5)
[1] 0
```

$$\Phi^{-1}(0.5) = 0$$

We find that 0 is the middle value of the standard normal distribution. Something that we already knew.

Note that these steps could be taken for in a similar way when other values of i or n are needed. This is just and example.

Now, we will determine the sample quantile. We collect the data from something that we assume is a normal distribution. We call the number of observations in this data set n. All of these N observations from our collected data are put from lowest to highest. Then we number them. $z_{(1)}$ is the lowest. $z_{(2)}$ is the then lowest, up to z_n, which is the highest value.

So, mathematically,

$$z_{(1)} < z_{(2)} < z_{(3)} < ... < z_{(n)}.$$

Example 13.2

Now, we ask our computer for 9 random observations from a standard normal distribution.

Code R

```
> x<-rnorm(9)
> sort(x)
[1] -1.7095202 -0.5148740 -0.4701114  0.4305446  0.5309747
0.6021393  0.8843518  1.4742833  1.6030414
```

If we take a look at the sample of 9 ordered values from a standard normal distribution, we find that our fifth value is 0.5309747, so $z_{(5)} = 0.5309747$.

So, all coordinates of the Q-Q plot are of the following form:

$$(\Phi^{-1}(\frac{i - 0.5}{n}), z_{(i)})$$

Where i is a number from 1 till n.

Example 13.3

In the example were we checked the 5^th observation of 9 observations in total. This gave us:

$$(\Phi^{-1}(\frac{5-0.5}{9}), z_{(5)}) = (0, 0.5309747).$$

Which is the red coloured dot in the following Q-Q plot. This reasoning could be followed for i from 1 till 9, this would give us entire plot

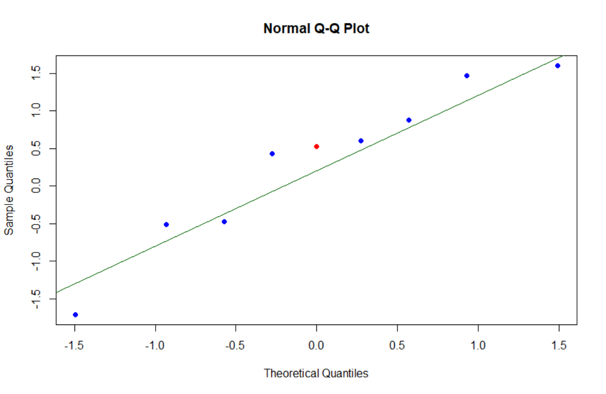


Figure 16: Q-Q plot of 9 Data Points of a Standard Normal Distribution

So, on the x-axis we find the value that we expect to find when a normal distribution is the underlying distribution and on the y-axis we find the value we actually found from the data. The green line shows where the values would be if the expected value and the true value would be similar. So, if all points are on this line, we are pretty sure that the underlying distribution of the data is indeed normal.

Note that we now produced nine random numbers from a standard normal distribution, and the numbers do not fall on the green line. Normally you do not know if the numbers are from a normal distribution, so from a small sample, it is difficult to determine anything about the underlying distribution. Let's take a look at a bigger sample size.

Example 14

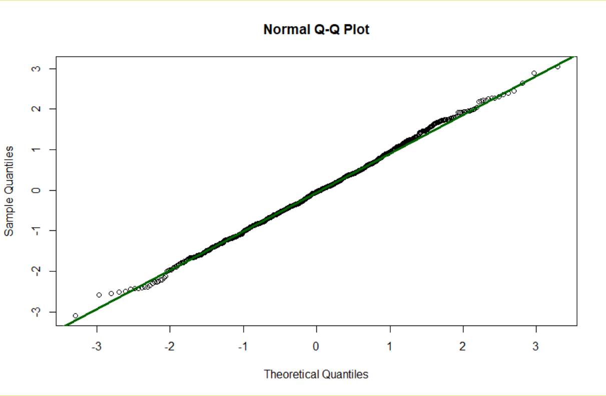


Figure 17: Q-Q plot of 1000 Data Points of a Standard Normal Distribution

Now, the data falls on the line. This verifies our assumption of normality.

2.7.5 Parametric and Nonparametric Testing

So, we now know what the normal distribution is and how we could check for it. But why do we want to know if our data has an underlying normal distribution? This is needed know if you should use a *parametric test* or a *nonparametric test.* If we know that a certain distribution is underlying of the data, a *parametric test* is often more powerful than a similar *nonparametric test.* So checking for an underlying distribution helps in order to find the test that will most likely find a good answer on your question. (Colquhoun, 1971)

Note: we talk about 'a distribution', this is in real life often the normal distribution, but there are other distributions that could be underlying is well. Best is to check from the histogram of this data, and search for an underlying distribution that fits this. If the data seems irregular or does not satisfy the assumptions of a particular distribution, a nonparametric test should be used. Otherwise, search for a parametric test.

2.8 Testing

The focus will lay on testing the difference between two groups. Since this will help us in sufficing the goal of determining the media claims between two groups. Most often those claims will be along the lines of "group 1 is different at something than group 2". So let's first specify what "different" means in the eyes of a mathematician.

2.8.1 Comparing Variables

When talking about a statistical difference, often a difference in means in meant. We want to know if the difference found in the data is significant. In other words, we want to know if the difference found could be found because of a sampling error, or if there is a real difference. This will set us with the following hypotheses.

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

It is also able to compare other variables, but obviously, the setting of the hypothesis depends on the research question. And in the tests that we are going to study the hypotheses mentioned previously will be used.

2.8.2 Performing Test

So, when the hypothesis is set, we want to know how likely it is that with the data found, the null hypothesis could be true. We have to introduce something called the *p-value.* Which is a probability that quantifies the strength of the evidence against the *null hypothesis* and in favor of the *alternative hypothesis.* We will later on explain how you should go from the data to a p-value. (Diez, 2019)

So, when a low *p-value* is found, we find that the probability that the *null hypothesis* is true is low. So we will reject this *null hypothesis*; the evidence does not support the *null hypothesis.* When the *p-value* we fail to reject the *null hypothesis.*(Diez, 2019). In order to understand when this value is low, we have to understand something about the possible errors.

2.8.3 Type I and Type II Error

When making a decision about the hypothesis, we could distinguish four cases:

1. The decision not to reject the null hypothesis could be correct.
2. The decision not to reject the null hypothesis could be incorrect. (*Type II error*)
3. The decision to reject the null hypothesis could be correct.
4. The decision to reject the null hypothesis could be incorrect.(*Type I error*)

(Privitera, 2015)

We note two different types mistakes that could be made. The incorrect decision not to reject a false null hypothesis, which is referred to as a Type II error. Or the incorrect decision to reject a true null hypothesis. This decision is an example of a Type I error. (Privitera, 2015)
Of course, we want to make sure that the amount of wrong judgements is minimized. In order to do so, we should understand how both types of errors influence each other.

Example 15

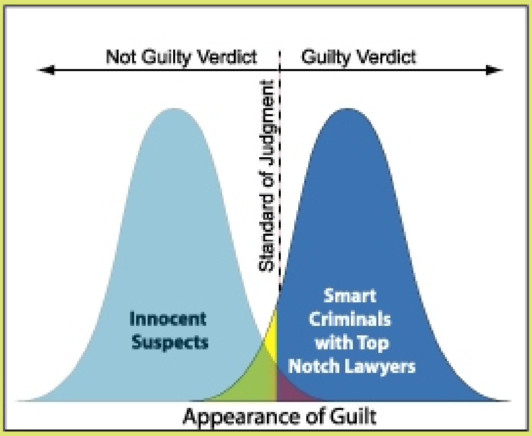


Figure 18: Distribution of Evidence. (Chl, 2013)

In figure 18 we look at two normally distributed amount of evidence of suspect in court. The light blue group represents the evidence found in a case with an innocent suspects, while the dark blue group represents the criminals. There is an overlap between the groups, since it could happen that there is almost no evidence for a guilty suspect, but it could also happen that there is a lot of evidence against an innocent suspect. Still, overall there would be more evidence for guilty suspects, therefore, the two distributions fall together only partly. The following hypotheses are set in this example:

H_0: The suspect is guilty.
H_1: The suspect is not guilty.

For a certain amount of evidence we should draw the conclusion of guilt or innocence, this line is represented by the "standard of judgements". So if we find more evidence we find that the suspect will be found guilty, otherwise H_0 will be rejected. Due to the overlap of these two distributions, it is impossible to find a standard of judgement that makes sure that no innocent people are incarcerated, or that no guilty people will get free. Shifting the "standard of judgement" to the left, would make sure that almost all criminals would get punished, but it will also increase the purple plane; more innocent people would get incarcerated.
A similar reasoning could be followed to conclude that increasing the value of the standard of judgement would make sure almost no innocent suspects would get punished, but quite a lot of guilty suspect would not get free. (Chl, 2013)

This standard of judgement is in statistical testing represented with a α-value and it is compared to the *p-value*, we mentioned before. If p is lower than α we will reject our H_0 and otherwise we will fail to reject H_0.
So if we chose the value of α-value too low, we will have the problem that too much type I errors are made, but if we chose the α-value to high, too much type II errors are made. In most statistical

tests the α-value is chosen to be 0.05, to minimize the total amount of errors. (Chl, 2013)

2.9 Possible Tests

Now, we do understand why we need statistical testing and how to approach it, the only thing left is to determine how we need to get a p-value to draw rightful conclusions from. Unfortunately,there is no simple way to find a p-value, there are very much possible statistical test to get a p-value. In order to select the right one, we have to take a good look into the possibilities and differences.

Since our main goal will be to draw conclusions on the difference between two different groups, we will discuss test that are designed to do such a comparison between the means of two different data sets.

2.10 Z-test

In order to understand the Z-test, we will introduce z-scores. Note that we will talk about the standardization of the normal distribution, so the conclusion drawn are based on this specific case. Z-scores could also be calculated based on other distributions, but we will not go into that.

2.10.1 Standardization

Example 16.1

Two friends from different schools are arguing about who scored better on their mathematics test. The first boy got a 8 out of 10 on his test, while the second scored a 7 out of 10.

The first boy argues that an 8 is higher than a 7, therefore he did better. The second child argues that a lot of his classmates got really low scores, so he was one of the smartest of his class, therefore, his 7 is worth more than the 8 of his friend.

Now, assume that the scores of both math test were normal distributed, than we can conclude who scored better respectively by introducing z-scores.

We know that all normal distributions are described by their mean and standard deviation. If we want to translate any normal distribution into a standard normal distribution ("standardize"), we should therefore change the data in such a way that the mean is 0 and standard deviation is 1.

Look at a normal distribution X with mean μ and standard deviation σ. We know that the values center around the mean, so, if we want them to center them around 0, we should simply deduct μ from all values of the distribution. So now $X - \mu$ is a normal distribution with mean 0 and standard deviation σ. All that is left to do, is to scale these distribution, this could be done by dividing all values by σ.

So $\frac{(X-\mu)}{\sigma}$ would be a the standardization of a normal distribution with mean μ and standard distribution σ.

This figure visualizes a standardized normal distribution:

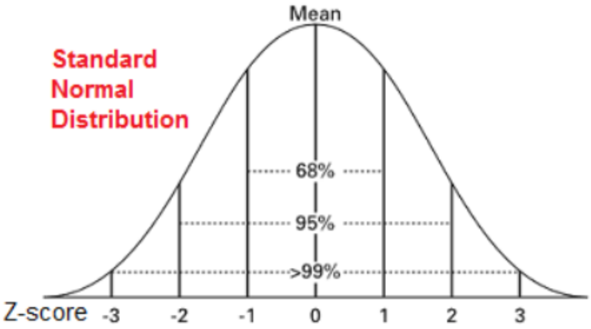


Figure 19: Z-score Distribution. (Mia, 2019)

So, a z-score is simply the standardized form of a value from a normal distribution and can be calculated with.

$$Z = \frac{(x - \mu)}{\sigma}$$

s a result a z-score is the number of times the standard deviation fits between the mean and the actual score and therefore is a measure of the difference between the mean and the actual score. (Lane et al., 2019)

Example 16.2

With this it would be be possible to determine the z-scores of the boys. The first one finds a z-score of 0, which means that he scored the average score. The second child finds that he has a z-score of 2, which means that he scored 2 times standard deviation above the average. If we interpret this with the rule of thumb, we find only 2,5 percent of the class scored better than him. While 50 percent scored better than the first child. We can now decide that the second child indeed did better, concerning the fact that he was more on top of his class.

2.10.2 Two sample z-distribution

We are interested in if the sample mean of two different groups are equal. First, we have to understand why we could use an underlying z-distribution in this case as well.

We have two data sets. The first one consist of data from population 1, and it is assumed to be normally distributed with mean μ_1 and standard deviation σ_1. From the data available we could calculate a mean $\bar{x}_1$ and s_1.
A similar notation is used for our data from population 2.

The sampling distribution of the difference between means can be thought of as the distribution that would result if we repeated the following three steps over and over again:

- Sample n_1 scores from Population 1 and n_2 scores from Population 2
- Compute the means of the two samples
- Compute the difference between means

If you would this experiment enough times, you would find that certain differences in means will occur more often than others. More precisely, there will be a certain value that occurs most often, and how further away from this value you go, how less often this value occurs; the difference between the means are normally distributed. (Lane et al., 2019).

So now, in a similar way as the one sample case, we could calculate the z-score of a specific difference in means.

How are we going to calculate these z-scores? Let's take the formula on how to calculate z-scores and translate it to apply it to the distribution of the difference of means.

First, take a look at the part $\bar{x} - \mu$. So if we would apply this to sample the distribution of the difference of the means, we would find $(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)$, since the mean of the sampling distribution is the difference in means.

Now, all that is left, is dividing it by the standard distribution of the distribution of the difference of means. In order to really understand these sampling distribution you have to understand some rules. Really understanding the mathematical background of the formulas is difficult. For now, it will suffice to just give, explain and apply the formulas.

Formulas

Notation:
Were use a subscript M to show that we talk about the distribution of means. So σ_{M_1} is the standard deviation of the distribution of the sampling means of population 1.

Variance sum law:
$\sigma^2_{M_1-M_2} = \sigma^2_{M_1} + \sigma^2_{M_1}$.
The variance of the distribution combining two independent variables, is the sum of the variance of both independent variables. (Lane et al., 2019)

Variance of the sampling distribution of the mean:
$\sigma^2_M = \frac{\sigma^2}{N}$
The variance of the sampling distribution of the mean is the population variance divided by N, the sample size (the number of scores used to compute a mean). Thus, the larger the sample size, the smaller the variance of the sampling distribution of the mean. (Lane et al., 2019)

Combining these rules would give

$$\sigma^2_{M_1-M_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Taking the root of both sides would then give us

$$\sigma_{M_1-M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

So, the standard deviation of the sampling distributions of the difference in means is
$\sigma_{M_1-M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

Now, we have enough information to calculate the z-score that belongs to the distribution of the difference in sampling means.
Compbining this would give us:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

2.10.3 Assume Null Hypothesis

Now, we could use this z-score to check a hypothesis, for example, the same mean.

We would set the following hypotheses:
$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$.
So, we would use the formula

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

to find the z-scores that belong to the sampling difference in two means, when expected that those means are equal.

After calculating this z score, we should determine if we can indeed conclude that there is no difference in means found from the data, or if there is some difference. If there is no difference, the z-score we calculate will be 0, but when is the z-score too far from 0 to say that it is not likely that the means are indeed similar?

We could distinguish between two cases when looking at the *alternative hypothesis*, namely μ_1 is bigger than μ_2 or μ_1 is smaller than μ_2. This is why there are two sides of the distribution should be tested; z-scores that are really low let's us decide that H_0 probably is not true, just like z-scores that are really high.

Note: There is also a possibility to test if a certain mean is bigger than an other mean. This is done by one-tailed testing, the set up is quite similar till this point. From here one, only one area would be investigated.

There 2 ways to conclude what "too high" and "too low" are in this case. You could either calculate a probability from the z-score, or you could compare the z-score to the interval of possible z scores. We will discuss both methods, but they will, if done correctly, always lead to the same conclusion.

2.10.4 Interval

We construct an interval that all z-scores from data that fail to reject H_0 fall into.

We use the α-value of 5 %. Which means that the H_0 will be rejected 5% of the cases.
Since we have to check two sides, we have to reject all z-scores in the lowest and highest 2.5%. So, let's calculate the z-scores that fall between 2,5% and 97,5% of the data.

Code R

```
> qnorm(0.025)
[1] -1.959964
> qnorm(0.975)
[1] 1.959964
```

This would mean that all z-scores lower than -1.96 and higher than 1.96 will represent the 5% of the data that will lead to a rejection of the H_0. This is visually represented by the following figure.

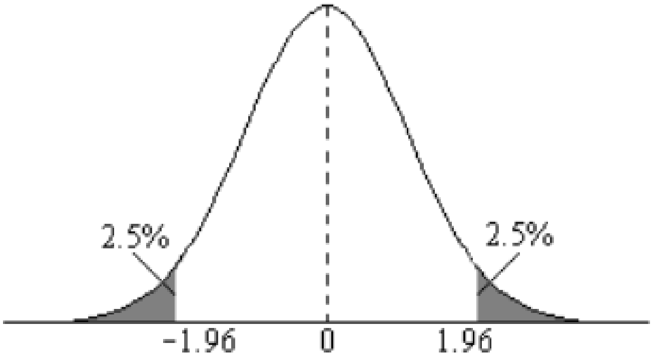


Figure 20: Two-Tailed Testing (Lewinson, 2019)

So, in conclusion from two different data sets, the mean and size could be determined. If we then also know the standard deviation of the distribution of each population, we could determine a z-score. All z-scores bellow -1.96 and above 1.96 will determine that it is very unlikely that the means of the two populations are the same, based on the data found.

2.10.5 P-value

An other option would be to translate a z-score to a p-value. The calculation of the z-score will still be done using the formula that assumes the H_0.

Example 17.1

If we, for example, find a z-score of -2, we would calculate:

Code R

```
> pnorm(-2)
[1] 0.02275013
> 2*pnorm(-2)
[1] 0.04550026
```

This means that or difference in means found belong to the lowest 2.3% of the standard normal distribution. Since we are doing a two-sided test, the value has to be doubled before comparing it to the α-value.

If we would have a z-score that is higher than 0, we want to know if it is within the highest 2.5%. Therefore, we should not just double the probability, because this would provide us with a value higher than 1. Which could obviously never be a true probability.

To solve this, we should deduct the probability we find from 1. This will make sure that we will look at the highest 2.5%.

Example 17.2

So, we say we found a z-score of 2.

Code R

```
> pnorm(2)
[1] 0.9772499
> 1-pnorm(2)
[1] 0.02275013
> 2*(1-pnorm(2))
[1] 0.04550026
```

The p-vale of 0.04550026 means that it is highly improbable that the mean of population 1 and 2 are equal. So, since $0.04550026 < 0.05$, we will reject H_0.

Note that because of the symmetrical character of the normal distribution, we find the same p-value for -2 and 2. This will be the case for all pairs of the negative values and their positive counterpart.

2.10.6 Conclusion

Overall, doing these calculations by hand is time consuming, but understanding the underlying mechanism is really nice in order to understand the conclusions you draw. There are some statistical tools, like R, that could do these calculations for you.

In practice, the two-sample z-test is not used often, because the two population standard deviations σ_1 and σ_2 are usually unknown. A t-distribution will solve this.

2.11 T-test

Let's change the Z-test so we could apply it in the case that we still have two underlying normal distributions, but we do not know the standard deviations of these normal distributions.

In our test statistic from the Z-test, we could replace the standard deviation of the population (σ) with the standard deviation of the samples (s). This would give the following test statistic:

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

Where, again, μ_1 and μ_2 are the hypothesised means. So if we assume $\mu_1 = \mu_2$ we would get the following test statistic:

$$t = \frac{(\bar{x_1} - \bar{x_2})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

We cannot compare this t-value to the z-distribution, since this distribution also assumes a known standard deviation. Instead, we will introduce the t-distribution.

2.11.1 T-distribution

Understanding the formula behind the t-distribution is difficult. Therefore, we will not going discuss that. Instead we are going to take a look at why this t-distribution varies from a z-distribution.

Since we do not know σ, we want to make the tails of the distribution more important, compared to the z-distribution. This will distribute the weight some more, and will not depend on the estimated s as much. Therefore, the t-distribution has a similar shape as the normal distribution (*symmetric*, *unimodal*), but is is more flat out, to distribute the weight more to the ends.

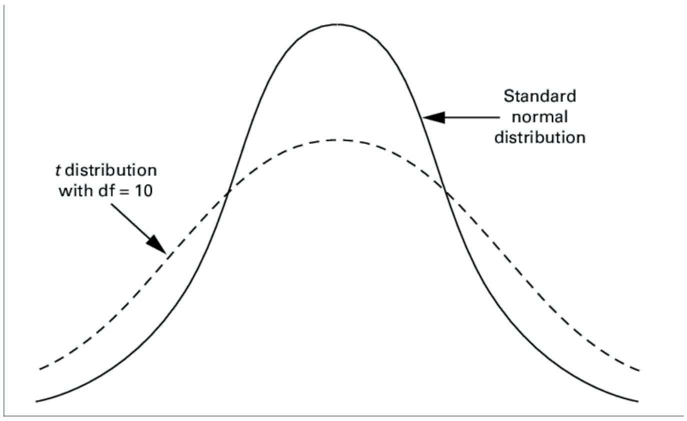


Figure 21: T-distribution and Normal Distribution. (Driscoll, 2001)

So, we see that the T distribution is a flatter version of the standard normal distributions. The *probability density function* of the t-distribution is:

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1+\frac{t^2}{\nu})^{-\frac{\nu+1}{2}}.$$

Which is really difficult to understand, so not try to do that. But what is important is that the formula is based on the variable ν, the degrees of freedom. Which is calculated by n-1: the total number of data, minus 1.
When you are to compare two groups, the degrees of freedom you will use is the smaller of $n_1 - 1$ and $n_2 - 1$.
The t-distribution approaches the standard normal distribution when the degrees of freedom increases. Since, more data will make sure that the sample distribution of the data will approach the underlying distribution. (Lane et al., 2019)

2.11.2 Interval

In a similar way as for the z test, we could provide an interval were all t-scores that do not reject the null hypothesis fall into. Again, we will use $\alpha = 0.05$ and a two-tailed test.

Code R

```
> qt(0.025, 10)
[1] -2.228139
> qt(0.975, 10)
[1] 2.228139
>
> qt(0.025, 1000)
[1] -1.962339
> qt(0.975, 1000)
[1] 1.962339
```

We see that when the degree of freedom increases, the interval gets smaller and eventually will produce the same bounds we found from the z-distribution.

2.11.3 P-value

Also, the p-value could be determined in a similar way.

Code R

```
> pt(2, 10)
[1] 0.963306
> 2*(1-pt(2, 10))
[1] 0.07338803
> pt(2, 1000)
[1] 0.9771148
> 2*(1-pt(2, 1000))
[1] 0.04577035
```

So, with the same t value, you could get a different conclusion on whether to reject H_0 depending on the degree of freedom.

2.11.4 Conclusion

A t-test is needed when the σ values of the population is not known. The test is dependent on the degree of freedom of the data sets. When the data set is big enough it will tend to fall together with the z-test.

Note: We discussed a specific case of t-test, namely one were we compare two means of non-paired data sets. In fact, all test based on the t-distribution are considered t-test, so it is important to carefully look at the t-test you are applying.

2.12 Other Tests

So, overall, we only focused on this very specific case, where we wanted to compare the means of two data sets of which the data is:

- numerical
- non paired
- satisfying the normal assumption

When all else still holds, but he assumption of normality is not met, the Mann-Whitney U Test could be used. Page 626 of Lane et al. (2019) will give an explanation on how to use this test and the underlying principles.

If any other of this points does not hold, the world of statistical testing might still provide you with answers. And you now know a lot about the terms that are important. The following list will guide you to aks the right questions in your search for the right test in an other situation.

- I want to compare ... (number) data sets.
- The data I want to work with is:
  - paired/non paired
  - numerical/categorical
  - parametric/non parametric
- I want to compare:
  - means
  - standard deviation
  - ..

Statistical books like ”Introduction to Statistics” by Lane et al. (2019) will provide more information on statistical testing.

2.13 Conclusion Statistical Testing

In conclusion, the mathematics used in media nowadays mainly consist of percentages and averages. These therms could be interpret incorrectly. They do not provide a good overview of the data, since they leave out information on the sample size and distribution.

We discussed different types of data, distributions and statistical testing and showed how they could be combined to provide a conclusion that takes sample size and distribution into account. We mainly focused on finding if there is a statistical difference between the means of two data sets with assumed normally.
A journalist will now be able to take a critical look at conclusions drawn about a difference in certain groups. He will be aware that a difference in numbers could also be explained by sampling errors. He will understand that it is difficult to summarize numbers and draw conclusions from them. Also, he will have gained some understanding in mathematical terms, concepts and approaches, which will help him understand research.
Overall, hopefully he will be more critical when publishing numbers in the future.

This theoretical approach is now to be tested in the following chapter. Which will show how to apply statistical testing to proof or falsify media claims.

3 Fact checking media

To further investigate how the media uses numbers in order to proof a point, a look will be taken at a article with the remarkable title ”Girls may perform better at school than boys – but their experience is much less happy”. (Smith, 2016) This article is written about the school system in Wales. It uses sources to provide claims in order to prove this main point. This sources are all (based on) researches done. The goal is to investigate a claim done in the article, to understand how it is supported by the research that is referred to. Furthermore, we want to check the statistics used in the research.

3.1 Subject and Article Choice

I chose this article since I wanted to focus on the difference between boys and girls in the school system. I read claims that the Dutch school system fits better with girls. For example, articles with the following headlines:

Example 11

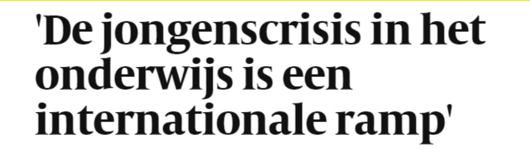

'De jongenscrisis in het onderwijs is een internationale ramp'

Figure 22: ”The Boy Crisis in Education Is an International Disaster” (Haan, 2011)

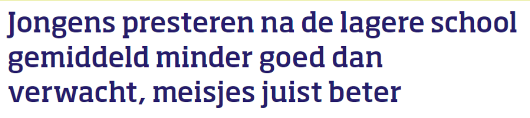

Jongens presteren na de lagere school gemiddeld minder goed dan verwacht, meisjes juist beter

Figure 23: ”Boys Perform Not as Good as Expected After Primary School, While Girls Perform Better Than Expected” (Centraal bureau voor de statistiek, 2014)

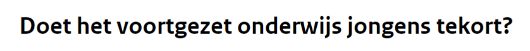

Doet het voortgezet onderwijs jongens tekort?

Figure 24: ”Does Secondary Education Let Boys Down?” (Inspectie van het onderwijs, 2019)

I always found this a strange claim, since it did not fit with my view of the Dutch school system. As a Dutch girl, I felt that the boys in my class got it easier to get a high grade, without stress.

The articles I read were in Dutch and somewhat outdated. Therefore, I decided not to work with any of these. Instead, I searched for an article with a similar claim, but a broader audience.

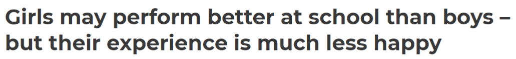

Girls may perform better at school than boys – but their experience is much less happy

Figure 25: Headline of the Article that Grabbed my Attention. (Smith, 2016)

When I found this Welsh article, I was really interested, since it did not only focus on the difference in school results, but also on the effect these results have on the well being of both groups. Which I think is very important, especially because girls having more stress in order to get high results was something that does fit my experience.

Also, the article chosen provided links to scientific research, with data available, so it was a good collection of interesting claims which could be checked with statistical testing.

3.2 Approach

Even tough the happiness of students is interesting, the focus will be on the grades for now, since this is easier to work with factual, numerical data. This is why we chose to check the following claim done by the article:"Girls perform better in most (or all) school subjects than boys, and that this trend has manifested in multiple countries since the early 20^{th} century". After this, the research and data of UCAS Analysis and resarch (2015) that the claim is based on will be investigated. After this, we could draw a conclusion on if the media used the numbers correctly.

In order to check the claim and research, an addition own research will be done. This will be based on data collected about the Dutch high school subjects.

3.3 Assumptions and Notes

The data used is collected by big organisations, therefore assumed is that the data itself correct.

The data will focus on the difference between boys and girls. There is discussion possible in this field. The following assumptions will be made:

- The words "gender" and "sex" can be used as synonyms. Since the numerical difference between these to groups is not that big.
- The sex of the person is determined by what they filled in themselves. This will be done, because, this research is mainly based on other research and this is how these approached it.
- Also, assumed is that all persons are either male or female.

3.4 Media versus Research

First, we will compare the claims in the media, with the claims the initial research makes.

The article by Smith (2016) claims the following: Girls perform better in most (or all) school subjects than boys, and that this trend has manifested in multiple countries since the early 20^{th} century.

The research by UCAS Analysis and resarch (2015) claims that 18 year old women were a third more likely to enter higher education than men in 2015.

It is obvious that those statements do not coincide.

- The research focused on how likely it was for women to get into higher education, compared to men getting in. They found that it was more likely for women to get. Asked should be whether this indeed supports the claim that girls perform better in all school subject. Was it possible for woman to get in because they scored higher on some, but not all subjects?
- The study provides data for 18 year olds, and suggest that more of the female population enter higher education at that age. Where the boys maybe younger when they entered higher education?
- The media article claims that there was a trend in multiple countries since the 20^{th} century. The article is only about 2015. Do they mean the 21^{th} century? Because the school system in the beginning of the 20^{th} century is not representative for the school system right now.
- The study focuses on the England. The claim in the article about multiple countries. About which countries are they speaking? Is Whales part of the countries that they are speaking about?

Overall, it is obvious that the article was not able to state the conclusion they did based on the conclusions of the research.That is why we decide to take a closer look at the data the research is based on. Maybe, this data does support the claim of girls getting higher grades than boys and the research just did not formulate that specific conclusion.

3.4.1 Defining a Research Question

Fortunately, the data belonging to the research was available. It consist of numbers of applicants to higher educations per A level point score and per sex. This A level score is determined by looking at the highest three A level grades achieved by the applicant. The following points per grade are used in the calculation: A* = 6, A = 5, B = 4, C = 3, D = 2, E = 1.. This means that all scores are between 3 and 18 and only consists of entire numbers. (UCAS Analysis and resarch, 2015).
When applying for higher education, a minimum of three A-level scores are needed. Universities like Oxford even advise against taking more A-level courses if this would endanger the results in the first three subjects. (Oxford Royale Academy, 2017)

Therefore we will now assume that taking a look at the results of the three highest A-level scores will indeed give a good overview of the capability of a student. But still, there is not distinction between different school subjects. So, we could only ask ourselves a general question. We will formulate the following research question "Do girls have a different mean A-level score than boys have?"

3.4.2 Hypothesis

From the data we find $\bar{x}_{women} = 11.83$ and $\bar{x}_{men} = 11.89$. So, the mean of the score of men is actually a bit higher than that of the women. We will stat by assuming that there is no difference to be found in the A level scores between boys and girls.

- $H_0 : \mu_{men} = \mu_{women}$
- $H_1 : \mu_{men} \neq \mu_{women}$

3.4.3 Analysing Data

In order to find the right test to answer the question, we will investigate the data.

Since the data only consist of entire numbers, the data itself is discrete. But the underlying distribution is not, since intelligence does not occur only in discrete numbers. These test scores are rounded up.

Also, the data that we are working with is unpaired. Since the scores of boys and girls are not linked.

Finally, we want to know if the assumption of normality is met. We will use descriptive statistics to find if this would be a good assumption.

First, we will produce boxplots.

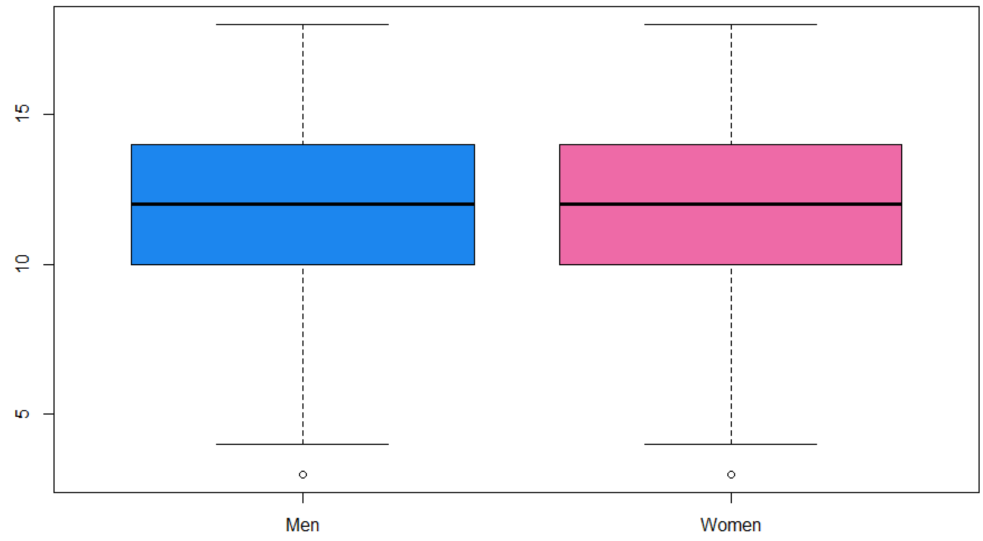


Figure 26: Boxplots of A level scores. 5

Then, we produce histograms and Q-Q plots.

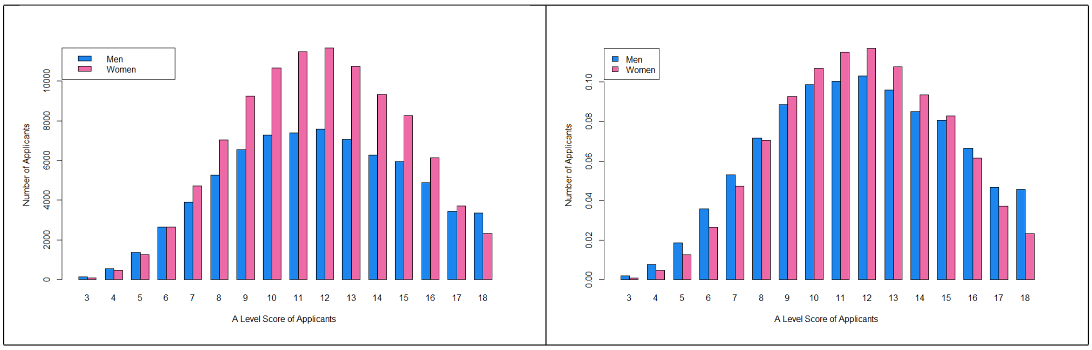


Table 1: Histograms for the A Level Scores of Men and Women

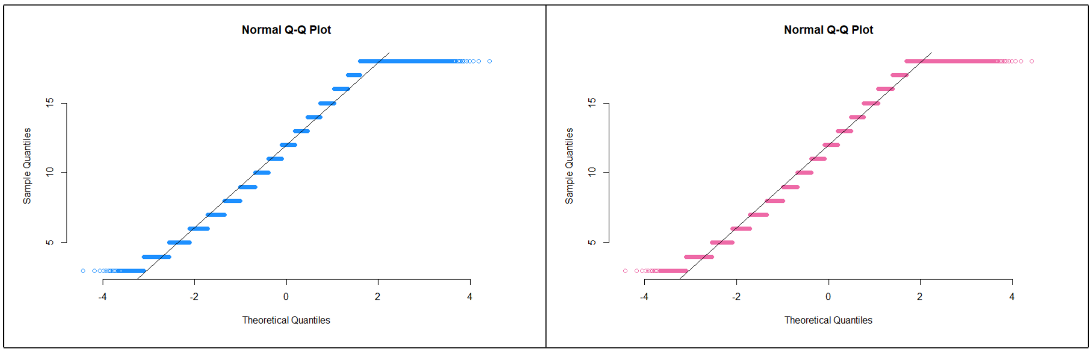


Table 2: Q-Q plot for the A level scores of Men and Women

A bigger version of the pictures can be found in the appendix, chapter 5.

So, there are a few things that could be noted from these plots.

- The underlying distribution of both groups seems very similar.
- The data collected is not continuous, so we find stripes occurring in the plot. Since the theoretical values are continuous. We already noted that, even tough the test scores are not continuous, assuming this for the underlying distribution could work, since the A level scores are rounded up scores of rounded up test scores.
- There is quite a big group with high scores, compared to the group with low scores, that is because the collected data is from students who applied to higher education. And of course, a student with a high score is more likely to apply for higher education. Again,this probably means that the entire underlying distribution of A-level scores is normal, but there are scores missing because of the application process.

3.4.4 Testing Hypothesis

In the previous chapter we found that the t test should provide us with the right conclusion.

Output R

```
t.test(data2015Men, data2015Women)

Welch Two Sample t-test

data:  data2015Men and data2015Women
t = 4.443, df = 205221, p-value = 8.875e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.03523492 0.09085951
sample estimates:
mean of x mean of y
 11.89936  11.83631
```

The p-value provided is really low, namely 8.875e-06, this means that we will reject our null hypothesis, and we find that it is really unlikely that the on average, boys and girls did perform similar on the test. Since we already found that boys scored a higher average, we will use a one sided t-test to find if boys scored significantly higher on average.

So we set:

- $H_0 : \mu_{boys} = \mu_{girls}$
- $H_1 : \mu_{boys} > \mu_{girls}$

Output R

```
t.test(data2015Men, data2015Women, alternative = "less")

Welch Two Sample t-test

data:  data2015Men and data2015Women
t = 4.443, df = 205221, p-value = 1
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 0.086388
sample estimates:
mean of x mean of y
 11.89936  11.83631
```

The p-value is 1, so really no reason so suspect that the A level score of the boys is not higher than that of girls.

So, we found a significant difference, but the difference was between a mean A level score of 11.89 and 11.83, which might in practice not seem like a really important difference in the context. Since it would both be test scores of almost 12. And since this number has such a small practical difference, the number of students that are entering higher education might not necessarily be based by their gender.

So why do we find such a significance difference? This is mainly due to the size of this data set. If we take a look at the test statistic for the t-test.

$$t = \frac{(\bar{x_1} - \bar{x_2})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

If we find that n_1 and n_2 getting big, $\frac{s_1^2}{n_1}$ and $\frac{s_2^2}{n_2}$ will get small. So $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ will get small, and therefore the denominator of the test statistics is quite small, so automatically the t-value will get big, when n is big. So this will more likely result in a rejection of H_0.

Logically, when a lot of data is collected, the distribution of the data will tend to go to the underlying distribution. So, only if the mean is equal, we will find no significant different.

So, based on the data we found something called *statistical significance*, but this is different from the *practical significance.* (Lane et al., 2019)

A more practical situation would be that there are 100 people of each sex that applied to a certain university. A practical question could be if the mean score of the boys that apply varies significantly from the means score of the girls that apply.

Output R

```
> t.test(meansM, meansW)

Welch Two Sample t-test

data:  meansM and meansW
t = 0.0041987, df = 1997.9, p-value = 0.9967
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.02796531  0.02808531
sample estimates:
mean of x mean of y
 11.82273  11.82267
```

So, for a random group of 100 people, no significant difference can be found.

Now we take a look at the situation of 1000 random people of each sex applying to this university. And again check if there is a difference in the means of the applicants.

Output R

```
> t.test(meansM, meansW)

Welch Two Sample t-test

data:  meansM and meansW
t = -1.2098, df = 1997.5, p-value = 0.2265
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.014410692  0.003414692
sample estimates:
mean of x mean of y
 11.83097  11.83647
```

Again, in this experiment, there is no significant difference in the means found.

This shows that there is a difference between statistical and practical significance. So a statistical difference does not mean that the difference found in important. In fact, a small effect can be highly significant if the sample size is large enough. This makes it important to not only use the right test, but also to apply this in a way that provided useful information.

3.4.5 Conclusion

So, let's look back at the claim done by the article, which was "Girls perform better in most (or all) school subjects than boys,and that this trend has manifested in multiple countries since the early 20^{th} century.

Talking about an change that can be seen in multiple countries since the early 20^{th} century and only referring to an article about England in 2015 is already quite strange. Even weirder is that this article did not even compare school subjects. The research claims that boys are indeed having a disadvantage in the current school system, but the data provided with it does not necessarily support this. On average, boys still seem to score significantly higher, based on the data. The problem is that a lot less boys apply.

Still a lot of questions remain, that should be answered in order to draw a rightful conclusion. Since from this data no real conclusion could be drawn on how to solve the problem of less boys getting into higher education. Determined should be if the problem is the motivation of the boys to continue their education, or if they are not able to the high grades to get into the studies they want to.

- What were the total amount of students graduating that year? If this is known, determined could be whether boys are already falling behind at high school, or whether they just do not want to into higher education
- What are the scores of the students who did not apply? If they did not apply because of low grades the way the boys should be approached is different,from when they would have good grades, but just did not feel the urge to apply to higher education.
- What where the other grades of each person? Since the A level scores where based on just three subjects, it is strange that the conclusion in the article is that women perform better in all subjects, which definitely cannot be found from these A level scores.

3.5 Different Subject

Girls perform better than boys implies that girls might perform better at all school subjects. Therefore, we want to check how boys and girls very in different school subjects. In order to do this, numbers from the Dutch Central Institute for Test Development (CITO) are used. They produce the exams that have to be passed in order to get a Dutch high school diploma.

They do not only create the exams, they also collect the results of all high school students each year. I have contacted them and they provided me with some explanations for the numbers on their site. (Cito, 2019a).

Even tough the data represents the Dutch society, I think it is a broad enough data set to investigate if there is a difference in the academic achievements of boys and girls among different school subjects. We might not be able to conclude which school subjects boys or girls excel in overall, but we can conclude if girls and boys perform differently after having similar lessons.

3.5.1 Testing

Cito strives for a normal distribution for their scores, and therefore we will have enough information from the mean and standard deviation the provide. Since the data sets are big enough to represent the entire populations, we will use z scores. We will set the following hypothesis:

We will start with the grades from the school subject "Dutch". The mean and standard deviation used is from the VWO exam of 2019. The *mean* of the boys was 40.24 with a *standard deviation* of 7.47. The *mean* of the girls was 40.83 with a *standard deviation* of 7.29. First, we set:

$$H_0 : \mu_g = \mu_b$$

$$H_1 : \mu_g \neq \mu_b.$$

We find the following z score, find the entire code in the appendix. 5.

Output R

```
> z
[1] 6.978352
> 2*(1-pnorm(z))
[1] 2.986722e-12
```

So again, a small difference in the mean of the test scores will lead to conclude the rejection of H_0. We know that the girls had a higher average, we use a one sided test to determine if they scored significantly higher. So, we set:

$$H_0 : \mu_g = \mu_b$$

$$H_1 : \mu_g > \mu_b.$$

Output R

```
> (1-pnorm(z))
[1] 1.493361e-12
```

We again reject H_0, so we conclude that girls scored significantly higher on the VWO high school exam of 2019. It is very questionable if this statistical significance was also a practical one. Since there are a lot of grades that determine if you get your diploma. Also, there is about 0.5 point of difference between the means. Which comes down to a difference of about 0.1 on the grade scale. (Cito, 2019b)

But this obviously does not conclude that girls scored better on all school subjects. Let's take a look at the scores of the mathematics exam of VWO of the same year. We find a *mean* for the girls of 45.76 points and a *standard deviation* of 13.01. The *mean* score of the boys is 47.24 and the *standard deviation* is 13.64. First, we will perform a two sided test. So:

$$H_0 : \mu_g = \mu_b$$

$$H_1 : \mu_g \neq \mu_b.$$

We find:

Output R

```
> z
[1] -7.595584
> 2*(pnorm(z))
[1] 3.06408e-14
```

So, we reject H_0 and we will perform a one sided test. Since the mean score of the boys is higher, we will set:

$$H_0 : \mu_g = \mu_b$$

$$H_1 : \mu_g < \mu_b.$$

Output R

```
> pnorm(z)
[1] 1.53204e-14
```

So boys score significantly higher on the subject of mathematics.

3.5.2 Conclusion

Whereas girls seem to score significantly higher on Dutch, boys score significantly higher on Mathematics. The practical difference between the two groups is not big. So we can conclude that there is a difference of scores between the genders. But this difference is not always in the advantage of the girls.

We can safely conclude that the statement made in the article of (Smith, 2016) is not supported by the data we checked about the Dutch school system.

In order to find a true answer on how girls and boys score in different school subjects some more data should be checked.

3.6 Conclusion

In conclusion, there are some different things to be noted from the difference in the real data and the statements the media makes based on them.

First, it is possible that the claims done in the media does not coincide with the claim in the research. More specifically, a claim in the media could include periods of time or countries that the research did not even investigate. In my opinion, when there is no reason to assume that the data also holds for any other case, a journalist should not use it to make claims.

Also, in our case, we found that the claim in the media was too general. It might be possible to find data that support the claim. But we mainly found data that supported that boys scored higher grades.

Noted should be that we focused on a specific claim in a specific article. So the conclusion drawn might not be a general one. Obviously not all data based statements in the media are wrong.

Overall, we could conclude that it is important the the media should specify before making a general claim, like "girls perform better at school". Since this is way too broad and therefore is open for misinterpretation.

4 Conclusion and Discussion

4.1 Conclusion

The main goal of the thesis was to provide insight in how to handle data to somebody who did not work with this before. We conclude the following topics a journalist should keep in mind before drawing conclusions on data.

4.1.1 Handle Data Incorrectly

In order to draw a right conclusion, some knowledge on how to handle data is needed. Data should not only be represented by its mean and some random percentages. Since this could leave out information the sample size and distribution. Statistical testing could help getting some more insight in data.

4.1.2 Statistical and Practical Significance

Even when data is handled correctly, it is important to make sure that the difference found is not just based on the fact that the data set is big. Because in that case, a small difference could be statistical significant, without having any practical meaning.

4.1.3 Generalize Research

It might be tempting to use research to support your claim, even if the research did not investigate your specific case. This will provide the audience with a wrong view of the subject. So you should be specific in the claims you make.

4.1.4 Overall Conclusion

Overall, the best advice I could give is to always be very specific about the numbers you use and the conclusions you draw from them. Otherwise a reader might be wrongfully informed about a subject.

4.2 Discussion

I have based my view of media and journalist on literature. It is perfectly possible that my view does not portray the way journalist handle literature correctly. They might have courses on this in their studies that I am not aware of.

Also, we found a piece of media that did not handle data correctly and we could therefore formulate some possible mistakes. It is possible that there are other mistakes that did not come to light in this research.

4.3 Possible Further Research

The following propositions for further research will provide more depth into the topic.

4.3.1 Discussing More Test

In this research, the focus was on statistical testing in the mean of two groups that are unpaired and have a normal distribution as underlying distribution. As discussed, there are a lot of other aspect in which statistical testing could be applied. Discussing these test in a similar way as the unpaired t-test and the z-test for comparing means would provide people with a non-mathematical background with the opportunity to get into statistics.

This similar way would consist of true to life examples that would provide insight in stati

4.3.2 Checking Media Claims on the Difference Between Boys and Girls in Education

It would be interesting to check other media claims, that go further than the actual difference in grades between boys and girls.
The main article about the difference of boys and girls in school of Smith (2016) is based on different research. He drew the following conclusions from this research:

- A teacher's gender has no measurable impact on pupils' academic achievement.
- Girls seem to do better because they have positive perceptions of education than boys.
- Girls seem to do better because they read more than boys.
- Girls seem to do better because they study more than boys.
- Girls seem to do better because they behave better than boys.

All of this claims and their underlying research would be interesting to analyse, because if some of his conclusions are true, then a change in the school system might be needed.

4.3.3 Checking General Media Claims

More general, checking media claims on different subjects is interesting.I think it is important to keep communicating the real meaning of numbers in a certain case to the general public. A further research might not be the solution to solve this problem. A better solution would be to create a website that people could visit to find more information on numbers in the media.

The site could discuss articles and television programs and justify or falsify the numbers that are used. Overall, it would make sure that readers of articles have a better insight in a certain topic before drawing conclusions.

5 Appendix

List of Variables and Formulas

The following notation is used in the research to indicate variables. Overall, it is important to understand that there is a difference variables that belong to a population and variables that belong to a data sample.

μ	mean of a population
$\bar{x}$	mean of a sample
σ	standard deviation of a population
s	standard deviation of a sample
σ^2	variance of a population
s^2	variance of a sample
N	population size
n	sample size
x_i	value of the i^th observation of a data set

The following formals hold:

$$\mu = \frac{\sum_i^N x_i}{N}$$

$$\bar{x} = \frac{\sum_i^n x_i}{n}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

Glossary

alternative hypothesis The opposite of the null hypothesis.(Livingston & Voakes, 2005). 8, 25, 30

bimodal Any distribution with precisely 2 prominent peaks (Diez, 2019). 13

categorical data A value that is categories.(Diez, 2019). 10, 11

Continuous Data A numerical variable that can take all values.(Diez, 2019). 10, 11

continuous probability distribution A probability distribution were were X is is continuous. (Diez, 2019). 18

descriptive statistics The branch of statistics concerned with describing and summarizing data. (Lane et al., 2019). 22, 38

Discrete Data A numerical variable that can only take numerical values with jumps.Diez (2019). 10

glossary List of descriptions.. 6

histogram A tool to summarize discrete or continuous data. It provides a visual interpretation of numerical data by showing the number of data points that fall within a specified range of values. It is similar to a vertical bar graph. (finance insitute, 2019). 13

hypothesis testing Find if data supports a claim.. 8

law of large numbers The proportions in the underlying distribution are noted with p, while the proportions found in the data are called $\hat{p}$. If more data is inspected, $\hat{p}$ will get closer to p. (Diez, 2019). 15

left-skewed A distribution with a long right tail (Diez, 2019). 13

mean The sum of the observed values divided by the number of observations. (Diez, 2019). 7, 9, 16, 17, 21, 43

median If the data are ordered from smallest to largest, the median is the observation right in the middle. If there are an even number of observations, there will be two values in the middle, and the median is taken as their average.(Diez, 2019). 7, 12

mode The value with the most occurrences in the data set.(Diez, 2019). 7

multimodal Any distribution with more than 2 prominent peaks (Diez, 2019). 13

Nominal Data A categorical variable without a special ordering.(Diez, 2019). 11

nonparametric test Tests that, although they involve some assumptions, do not assume a particular distribution. (Colquhoun, 1971). 25

null hypothesis The hypothesis that you will stand by unless the statistical evidence is very strong in the other direction. (Livingston & Voakes, 2005). 8, 25

numerical data A value that can take a wide range of numerical values. It is sensible to add, subtract, or take averages with those values.(Diez, 2019). 10, 11

Oridinal Data A categorical variable were the levels have a natural ordering.Diez (2019). 10

p-value The probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true. We typically use a summary statistic of the data, in this section the sample proportion, to help compute the p-value and evaluate the hypotheses. (Diez, 2019). 25, 26

paired data Two sets of observations are paired if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.(Diez, 2019). 11

parametric test Test which are based on an assumed form of distribution, usually the normal distribution, for the population from which the experimental samples are drawn.(Colquhoun, 1971). 25

percent Parts per 100.. 7

percentage increase The difference between a final value and begin value expressed as a percentage of the begin value. (Study.com, 2019). 7

population A large group of people, animals, objects, or responses that are alike in at least one respect.(Welkowitz, 2011). 9

practical significance The Finding that an effect is large or important. (Lane et al., 2019). 41

probability The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.(Diez, 2019). 18

probability density function A smooth curve that outlines a continuous probability distribution (Diez, 2019). 18, 33

quantile-quantile plots An exploratory graphical device used to check the validity of a distributional assumption for a data set. (Lane et al., 2019). 22

right-skewed A distribution with a long right tail (Diez, 2019). 13

sampling error A number that described how much an estimate will tend to vary from one sample to the next (Diez, 2019). 16

standard deviation A number that describes how far away the typical observation is from the mean.
The square root of the variance.
. 16, 17, 21, 43

statistical significance The null hypothesis of exactly no effect is rejected. (Lane et al., 2019). 41

symmetric A distribution with roughly equal tailing left and right (Diez, 2019). 13, 33

Type I error Failing to reject the null hypothesis when the alternative is actually true. (Diez, 2019). 25

Type II error Rejecting the null hypothesis when H_0 is actually true. (Diez, 2019). 25

unimodal A distribution with one prominent peak.. 13, 16, 33

variance The variance is a widely used measure of variability. It is defined as the mean squared deviation of scores from the mean.(Lane et al., 2019). 49

Figures

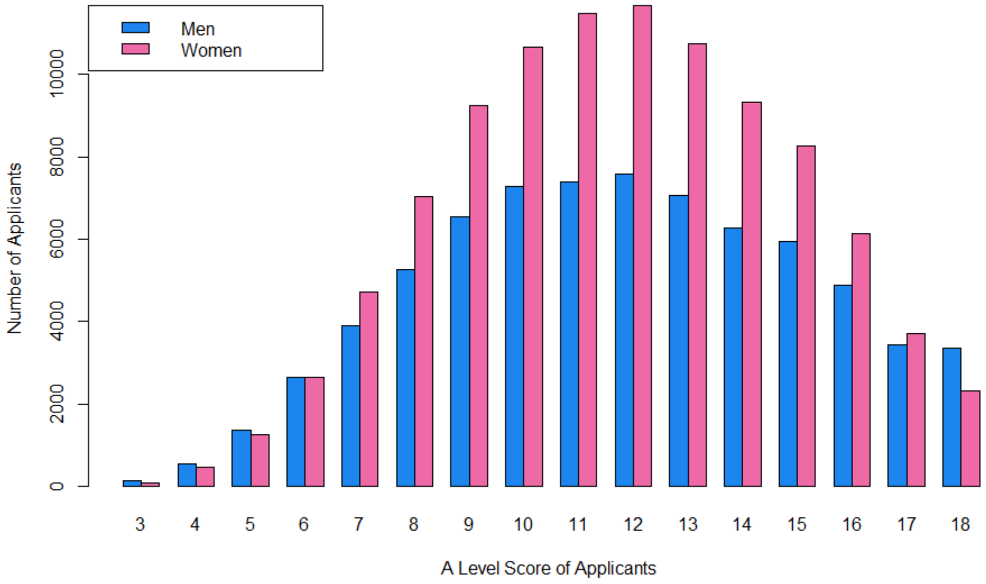


Figure 27: Boxplot

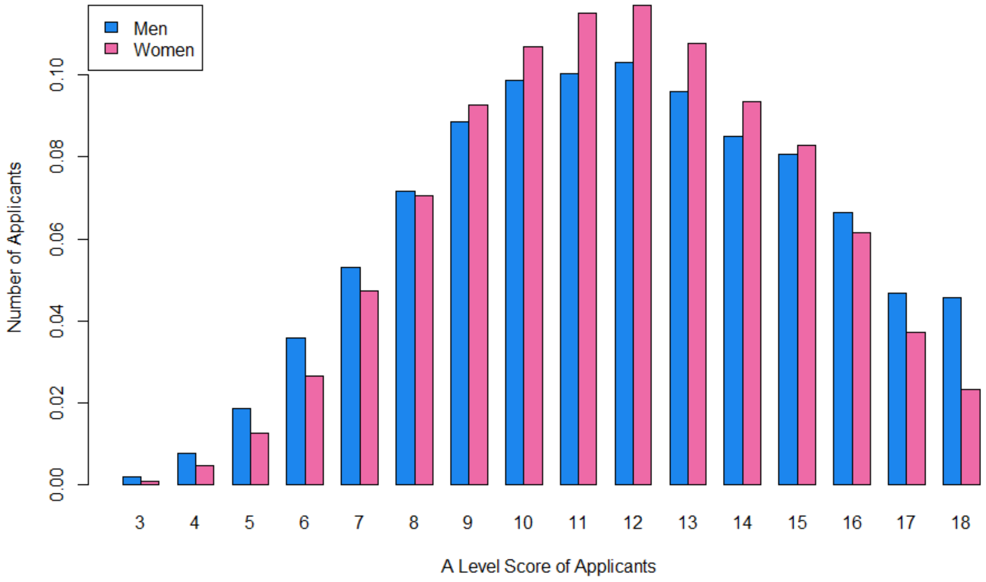


Figure 28: Boxplot

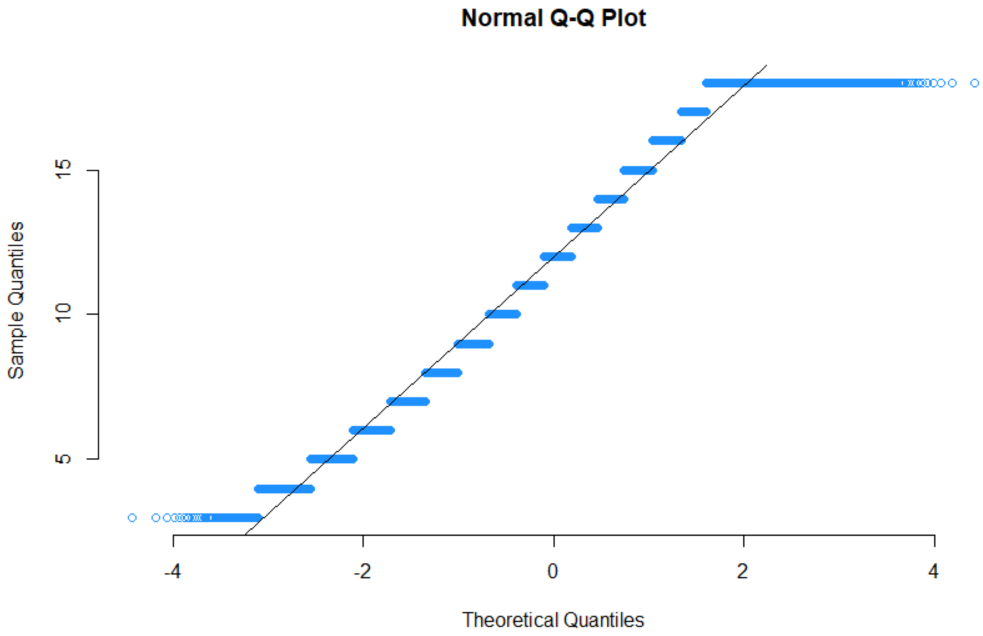


Figure 29: Boxplot

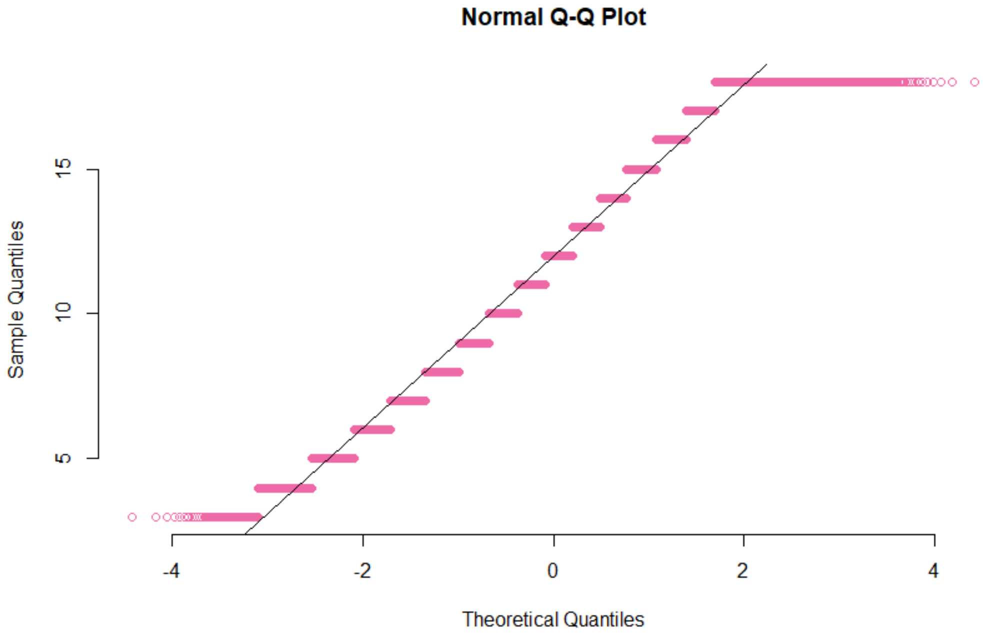


Figure 30: Boxplot

R codes

The following programs are written in R and used to provide insightful examples and do some statistical testing. I am far from a programming expert. So this might not be the most efficient or neat way to program these problems.

Grandmas and Granddaughters

```
age<-c(66, 13, 13, 69, 15, 64, 17, 7)

boxplot(age, main="Boxplot of age")
hist(age, breaks = 8)
```

Histogram Earthquake Magnitudes

Data retrieved from USGS (2019).

```
magnitude<-as.numeric(t(magnitude))
hist(magnitude)
hist(magnitude, freq = FALSE)
boxplot(magnitude)
```

Experiment of Throwing dice

```
#simulate 600 dies throws
y<-table(sample(x = 1:6, size = 600, replace = TRUE))
y

#find the proportion of each of the throws
y/600
```

Normal Distribution

```
value=seq(-4,4,length=200)
density=1/sqrt(2*pi)*exp(-x^2/2)
plot(value,density,type="l",lwd=2,col="blue", main = "Standard Normal Distribution")
```

QQ plot

```
#calculate the theoretical value from the normal distribution
qnorm(0.5)

#produce 9 values from a standard normal distribution
x<-rnorm(9)
#order these 9 values from low to high
sort(x)

#plot a qq plot from the values, color the 5th value red.
qqnorm(sort(x),col=c("blue","blue","blue","blue","red","blue","blue","blue","blue"),pch=16)
#produc a green line that shows were the values are expected
qqline(sort(x),col="darkgreen")

#produce a qq plot with 1000 values from a standard normal distribution
x<-rnorm(1000)
qqnorm(x)
qqline(x,col="darkgreen", lwd=3)
```

Girls Outperforming Boys

Data retrieved from UCAS Analysis and resarch (2015).

```
#distuinguish data between Men and Women
score2015Men<-c(rep(0,16))
score2015Women<-c(rep(0,16))
level<-c(3:18)
for(i in 1:32){
  if(data2015[i,2]=="Men"){
    for(j in 3:18){
      if(data2015[i,1]==j){
        score2015Men[j-2]<-score2015Men[j-2]+ as.numeric(data2015[i,4])
      }
      else{
        score2015Men <- score2015Men
      }
    }
  }
  else{
    for(j in 3:18){
      if(data2015[i,1]==j){
        score2015Women[j-2]<-score2015Women[j-2]+as.numeric(data2015[i,4])
      }
      else{
        score2015Women <- score2015Women
      }
    }
  }
}
score2015Women
score2015Men

#find the distrubition
#produce frequency table
x<-c(score2015Men,score2015Women)
score<-matrix(x, 16,2)
colnames(score)<-c("Men", "Women")
rownames(score)<-c(3:18)


#plot frequency table
barplot(t(score), col=c("dodgerblue2", "hotpink2"), beside = TRUE,
ylab = "Number of Applicants", xlab ="A Level Score of Applicants ")
legend("topleft", legend = c("Men", "Women"), fill=c("dodgerblue2", "hotpink2"))

y<-c(score2015Men/sum(score2015Men), score2015Women/sum(score2015Women))
dens<-matrix(y, 16,2)
colnames(dens)<-c("Men", "Women")
rownames(dens)<-c(3:18)


#plot desity table
barplot(t(dens), col=c("dodgerblue2", "hotpink2"), beside = TRUE,
ylab = "Number of Applicants", xlab ="A Level Score of Applicants ")
legend("topleft", legend = c("Men", "Women"), fill=c("dodgerblue2", "hotpink2"))
```

```
#create numeric dataset with all values to perform more test
#dataset for scores women
data2015Women<-c(rep(3, score2015Women[1]))#create vector with score 3
for(i in 2:16){#fill with other scores
data2015Women <- c(a, rep((2+i), score2015Women[i]))
}
#dataset for scores men
data2015Men<-c(rep(3, score2015Men[1]))
for(i in 2:16){
  data2015Men <- c(a, rep((2+i), score2015Men[i]))
}

boxplot(data2015Men, data2015Women, names=c("Men","Women"), col=c("dodgerblue2", "hotpink2"))

qqnorm(data2015Women)
qqnorm(data2015Women,pch =1,frame= FALSE)

#perform t.test
t.test(data2015Men, data2015Women)
t.test(data2015Men, data2015Women, alternative = "less")

#ttest for a certain group
group<-1000
epr<-1000
meansW<-c(rep(0,epr))
for(j in 1:epr){
part<-sample(1:102085, group, replace=F)
numW<-c(rep(0,group))
for (i in 1:group) {
  numW[i]<-data2015Women[part[i]]
}
mean(numW)
meansW[j]<-mean(numW)
j<-j+1
}

meansM<-c(rep(0,epr))
for(j in 1:epr){
  part<-sample(1:102085, group, replace=F)
  numM<-c(rep(0,group))
  for (i in 1:group) {
    numM[i]<-data2015Women[part[i]]
  }
  mean(numM)
  meansM[j]<-mean(numM)
  j<-j+1
}

t.test(meansM, meansW)
```

Testing Cito Exam Results

Data retrieved from Cito (2019c).

```
#testresultaten Nederlands VWO 2019
#assumpted normality
meanW<-40.83
```

```
sdW<-7.29
nrW<-16650
varW<-(sdW^2)/nrW

meanM<-40.24
sdM<- 7.47
varM<-sdM^2
nrM<-14104
varM<-(sdM^2)/nrM

#calculate z score an p value
z<- (meanW-meanM)/sqrt(varW+varM)
z
(1-pnorm(z))
2*(1-pnorm(z))

#testresultaten wisB
#assumpted normality
#collect data
meanW<-45.76
sdW<-13.01
nrW<-8806
varW<-(sdW^2)/nrW

meanM<-47.24
sdM<- 13.64
varM<-sdM^2
nrM<-9925
varM<-(sdM^2)/nrM

#calculate z score an p value
z<- (meanW-meanM)/sqrt(varW+varM)
z
pnorm(z)
2*(pnorm(z)
```

References

Başlar, G. (2011). *The Influence of Media on the Reconstruction of Social Reality Through Asymmetric Information* (Tech. Rep.).

Campell, M., & Shantikumar, S. (2016). *Standard Statistical Distributions (e.g. Normal, Poisson, Binomial) and their uses.* Retrieved 2020-01-10, from `https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1b-statistical-methods/statistical-distributions`

Centraal bureau voor de statistiek. (2014). *Jongens presteren na de lagere school gemiddeld minder goed dan verwacht, meisjes juist beter.* Retrieved 2019-12-14, from `https://www.cbs.nl/nl-nl/nieuws/2014/39/jongens-presteren-na-de-lagere-school-gemiddeld-minder-goed-dan-verwacht-meisjes-juist-beter`

Cherry, K. (2019). *What the Average IQ Is and What it Means.* Retrieved 2020-01-11, from `https://www.verywellmind.com/what-is-the-average-iq-2795284`

Chl. (2013). *Type I and Type II Errors.* Retrieved 2019-07-01, from `https://stats.stackexchange.com/posts/3568/revisions`

Cito. (2019a). *Informatie centrale examens 2019-2018-2017 vwo.* Retrieved 2020-03-15, from `https://www.cito.nl/onderwijs/voortgezet-onderwijs/centrale-examens-voortgezet-onderwijs/examenvoorbereiding/examenmateriaal/vwo-2019-2018-2017`

Cito. (2019b). *Omzettingstabel normering, Nederlands vwo, CSE 1e tijdvak, 2019 - Examenblad.* Retrieved 2020-03-26, from `https://www.examenblad.nl/examendocument/2019/cse-1/nederlands-vwo/omzettingstabel-normering/2019/vwo`

Cito. (2019c). *Opgaven, uitwerkingen en meer voor centrale examens 2019 vwo.* Retrieved 2020-03-26, from `https://www.cito.nl/onderwijs/voortgezet-onderwijs/centrale-examens-voortgezet-onderwijs/examenvoorbereiding/examenmateriaal/vwo-2019-2018-2017/vwo-2019-tv1`

Colquhoun, D. (1971). *Lectures on Biostatistics* (Tech. Rep.). Retrieved from `http://www.dcscience.net/Lectures{_}on{_}biostatistics-ocr4.pdf`

Conquermaths. (2013). *Maths Fails in the Media.* Retrieved 2020-01-30, from `https://www.conquermaths.com/news/post/index/49/Maths-Fails-in-the-Media`

Diez, D. (2019). *OpenIntro Statistics* (Fourth Edi ed.). Retrieved from `openintro.org/os`

Driscoll, P. (2001, may). Article 7. An introduction to hypothesis testing. Parametric comparison of two groups–2. *Emergency Medicine Journal*, *18*(3), 214–221. Retrieved from `http://emj.bmj.com/cgi/doi/10.1136/emj.18.3.214` doi: 10.1136/emj.18.3.214

finance insitute, C. (2019). *Histogram - Examples, Types, and How to Make Histograms.* Retrieved 2019-12-03, from `https://corporatefinanceinstitute.com/resources/excel/study/histogram/`

Haan, F. (2011). *De jongenscrisis in het onderwijs is een internationale ramp.* Retrieved 2019-12-14, from `https://www.volkskrant.nl/nieuws-achtergrond/de-jongenscrisis-in-het-onderwijs-is-een-internationale-ramp{~}bca4946f/`

Happy or Not Ltd. (2019). *Smiley Terminal^{TM} for Customer and Employee Satisfaction.* Retrieved 2019-11-18, from `https://www.happy-or-not.com/en/smiley-terminal/`

Inspectie van het onderwijs. (2019). *Doet het voortgezet onderwijs jongens tekort? — De Staat van het Onderwijs — Inspectie van het onderwijs.* Retrieved 2019-12-14, from `https://www.onderwijsinspectie.nl/onderwerpen/staat-van-het-onderwijs/trends-in-het-onderwijs/voortgezet-onderwijs/doet-het-voortgezet-onderwijs-jongens-tekort`

Kille, L. W. (2014). *Math basics for journalists: Working with averages and percentages.* Retrieved 2019-11-10, from https://journalistsresource.org/tip-sheets/foundations/math-for-journalists/

Lane, D. M., Scott, D., Hebl, M., Guerra, R., Osherson, D., & Zimmer, H. (2019). *Introduction to Statistics* (Tech. Rep.). Retrieved from http://onlinestatbook.com/Online{_}Statistics{_}Education.pdf

Lewinson, E. (2019). *One-tailed or two-tailed test, that is the question.* Retrieved 2020-03-03, from https://towardsdatascience.com/one-tailed-or-two-tailed-test-that-is-the-question-1283387f631c

Livingston, C., & Voakes, P. (2005). *Workina with Numbers and Statistics. A Handbook for Journalists.* New Jersey: Lawrence Erlbaum Associates. Retrieved from www.erJbaum.com.

Lumen Learning. (2019). *Extremes of Intelligence: Intellectual Disability and Giftedness — Lifespan Development.* Retrieved 2020-01-10, from https://courses.lumenlearning.com/suny-lifespandevelopment/chapter/extremes-of-intelligence-intellectual-disability-and-giftedness/

Lund Research Ltd. (2018). *How to do Normal Distributions Calculations.* Retrieved 2020-02-01, from https://statistics.laerd.com/statistical-guides/normal-distribution-calculations.php

Maier, S. R. (2002). Numbers in the News: A mathematics audit of a daily newspaper. *Journalism Studies*, *3*(4), 507–519. doi: 10.1080/1461670022000019191

Mia, I. (2019). *What is the difference between the z-distribution and the Normal distribution?* Retrieved 2020-02-29, from https://www.quora.com/What-is-the-difference-between-the-z-distribution-and-the-Normal-distribution

Mordkoff, T. J. (2000). *The Assumption(s) of Normality* (Tech. Rep.). Retrieved from http://www2.psychology.uiowa.edu/faculty/mordkoff/GradStats/part1/I.07normal.pdf

O'Cathain, A., Walters, S. J., Nicholl, J. P., Thomas, K. J., & Kirkham, M. (2002, mar). Use of evidence based leaflets to promote informed choice in maternity care: Randomised controlled trial in everyday practice. *British Medical Journal*, *324*(7338), 643–646. doi: 10.1136/bmj.324.7338.643

Oxford Royale Academy. (2017). *How to Choose the Right A-levels: a Guide for GCSE Students.* Retrieved 2020-01-08, from https://www.oxford-royale.com/articles/how-to-choose-a-levels.html{#}aId=55ae4eb1-10a6-4247-98fc-0bd931ef9a92

Pereira, I. (2020). *Iceland's First Baby Of 2020 Is One Of The Biggest Ever.* Retrieved 2020-01-14, from https://grapevine.is/news/2020/01/09/icelands-first-baby-of-2020-is-one-of-the-biggest-ever/

Privitera, G. J. (2015). *Statistics for the Behavioral Sciences.* SAGE Publications, Inc. Retrieved from https://www.sagepub.com/sites/default/files/upm-binaries/40007{_}Chapter8.pdf

Smith, K. (2016, aug). *Girls may perform better at school than boys – but their experience is much less happy.* Retrieved from https://theconversation.com/girls-may-perform-better-at-school-than-boys-but-their-experience-is-much-less-happy-63161

Study.com. (2019). *Percent Increase: Definition & Formula.* Retrieved 2019-12-04, from https://study.com/academy/lesson/percent-increase-definition-formula.html

UCAS Analysis and resarch. (2015). *End of Cycle report* (Tech. Rep.).

USGS. (2016). *The Severity of an Earthquake.* Retrieved 2020-01-11, from https://pubs.usgs.gov/gip/earthq4/severitygip.html

USGS. (2019). *New Earthquake Hazards Program.* Retrieved 2019-12-03, from `https://www.usgs.gov/natural-hazards/earthquake-hazards/lists-maps-and-statistics`

Welkowitz, J. (2011). *Introductory Statistics for the Behavioral Sciences.* Retrieved 2019-11-19, from `https://ebookcentral.proquest.com/lib/rug/reader.action?docID=826865`

Xkcd. (n.d.). *Fastest-Growing.* Retrieved 2019-11-10, from `https://xkcd.com/1102/`