A Gentle but Critical Introduction to Statistical Inference, Moderation, and Mediation

Wouter de Nooy

2016-09-20 - 2017-12-01

List of Figures

List of Tables

Contents

Introduction and Reader's Guide

In the years that I have been teaching inferential statistics to bachelor students in Communication Science, I have learned two things. First, it is paramount that students thoroughly understand the principles of statistical inference before they can apply statistical inference correctly themselves. Second, formal notation, manual calculation, and estimation details distract rather than help students understand what they are doing. This book offers a non-technical but thorough introduction to statistical inference. It discusses a minimal set of concepts needed to understand both the possibilities and pitfalls of estimation, null hypothesis testing, moderation, and mediation analysis. It uses a minimum of formal notation.

Intended Audience and Setting

This book is written as reading material for a follow-up course in statistics, in the bachelor of Communication Science at the University of Amsterdam. Students enrolled in this course, have had an introductory course in statistics that explained how to change research questions into variables and associations between variables, how to select and execute the correct analysis or test (in SPSS) to answer their research question, and how to interpret the results in a language that is both comprehensible for the average reader and complying with professional standards (APA6 standard for reporting test results). In addition, they have learned the very basics of inferential statistics: How to decide which null hypothesis to reject based on reported p values, and how to interpret confidence intervals.

This book is meant for use in a flipped-classroom setting. Students should read the text, watch embedded videos, and play with the interactive content before they meet in class. Class meetings are used to answer questions raised by the students, do group work to exercise with the concepts and techniques presented in the text, and do little tests to check understanding.

Interactive Content

The interactive content in this book replaces simulations that used to be demonstrated during lectures. I expect that doing simulations yourself rather than watching them being done by someone else enhances understanding. I have tried to break down the simulations into smaller steps, confronting the student several times with essentially the same simulation, but with added complexity. I hope that this approach enhances understanding and remembrance and, at the same time, avoids frustration caused by complex dashboards offering all options at once.

Most interactive content starts with a question regarding the student's expectations of what is going to happen in the simulation. I strongly recommend that students state their expectations before they start the simulations to see where their intuitions are right and where they are wrong.

Disclaimer

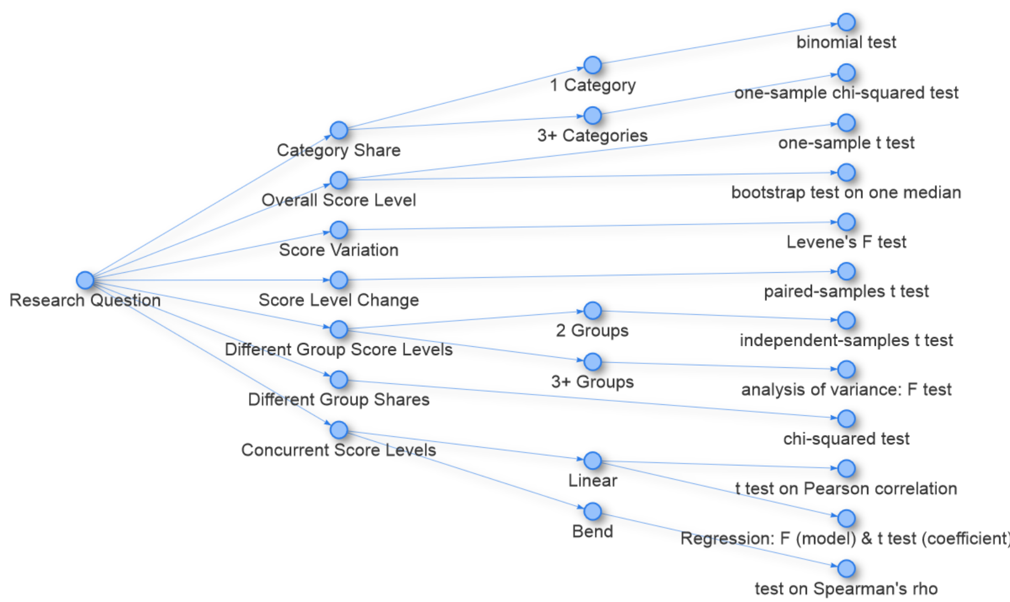
The example data sets have been generated for the purpose of demonstrating statistical techniques. These are not real data and no conclusions should be drawn from the results obtained from the data.

Acknowledgements

Adam Sasiadek developed the more complicated Shiny apps in this book. The College of Communication at the University of Amsterdam generously supported the creation of the apps whereas this university's Grassroots Project for ICT in education refused to support it. Renske van Bronswijk corrected my English. Any remaining errors result from changes and additions that I applied afterwards. My colleague Peter Neijens commented on a draft of this text. As one of the first tutors using this book, Luzia Helfer did many suggestions for improvement of the text.

SPSS Tutorial Videos List

Flow chart



Hint: Hover your mouse pointer over a node, click on a dot, drag with your (left) mouse button, and zoom with your mouse wheel.

Chapter 2: Probability Models: How Do I Get a Sampling Distribution?

Figure 2.5: Bootstrapping in SPSS.

Figure 2.6: Interpreting bootstrap results in SPSS.

Figure 2.7: Performing an exact test in SPSS.

Figure 2.8: Interpreting exact test results in SPSS.

Chapter 3: Estimating a Parameter: Which Population Values Are Plausible?

Chapter 4: Testing a Null Hypothesis: Am I Right or Am I Wrong?

Chapter 5: Which Sample Size Do I Need? Power!

Chapter 7: Moderation with Analysis of Variance

Chapter 8: Moderation with Regression Analysis

Chapter 9: Mediation with Regression Analysis

Chapter 1

Sampling Distribution: How Different Could My Sample Have Been?

Key concepts: inferential statistics, generalization, population, random sample, sample statistic, sampling space, random variable, sampling distribution, probability, probability distribution, discrete probability distribution, expected value/expectation, unbiased estimator, parameter, (downward) biased, representative sample, continuous variable, continuous probability distribution, probability density, (left-hand/right-hand) p value.

Summary

What does our sample tell us about the population from which it was drawn?

Statistical inference is about estimation and null hypothesis testing. We have collected data on a random sample and we want to draw conclusions (make inferences) about the population from which the sample was drawn. From the proportion of yellow candies in our sample bag, for instance, we want to estimate a plausible range of values for the proportion of yellow candies in a factory's stock (confidence interval). Alternatively, we may want to test the null hypothesis that one fifth of the candies in a factory's stock is yellow.

The sample does not offer a perfect miniature image of the population. If we would draw another random sample, it would have different characteristics. For instance, it would contain more or less yellow candies than the previous sample. To make an informed decision on the confidence interval or null hypothesis, we must compare the characteristic of the sample that we have drawn to the characteristics of the samples that we could have drawn.

The characteristics of the samples that we could have drawn is called the sampling distribution.

Sampling distributions are the central element in estimation and null hypothesis testing. In this chapter, we simulate sampling distributions to understand what they are. Here, *simulation* means that we let a computer draw many random samples from a population.

In Communication Science, we usually work with samples of human beings, for instance, users of social media, people looking for health information or entertainment, citizens preparing to cast a political vote, and an organization's stakeholders, or samples of media content such as tweets, tv advertisements, or newspaper articles. In the current and two subsequent chapters, however, we avoid the complexities of these samples.

We focus on a very tangible kind of sample, namely a bag of candies, which helps us understand the basic concepts of statistical inference: sampling distributions (the current chapter), probability distributions (Chapter 2), and estimation Chapter 3). Once we thoroughly understand these concepts, we turn to Communication Science examples.

1.1 Statistical Inference: Making the Most of Your Data

Statistics is a tool for scientific research. It offers a range of techniques to check whether statements about the observable world (empirical reality) are supported by data collected from that world. Scientific theories strive for general statements, that is, statements that apply to many situations. Checking these statements requires lots of data covering all situations addressed by theory.

Collecting data, however, is expensive, so we would like to collect as little data as possible and still be able to draw conclusions about a much larger set. The cost and time involved in collecting large sets of data are also relevant to applied research, such as market research. In this context we also like to collect as little data as necessary.

Inferential statistics offers techniques for making statements about a larger set of observations from data collected for a smaller set of observations. The large set of observations about which we want to make a statement is called the *population*. The smaller set is called a *sample*. We want to *generalize* a statement about the sample to a statement about the population from which the sample was drawn.

Traditionally, statistical inference is generalization from the data collected in a *random sample* to the population from which the sample was drawn. This approach is the focus of the present book because it is currently the most widely used type of statistical inference in the social sciences. We will, however, point out other approaches in Chapter 6.

Statistical inference is conceptually complicated and for that reason quite often used incorrectly. We will therefore spend quite some time on the principles of statistical inference. Good understanding of the principles should help you to recognize and avoid incorrect use of statistical inference. In addition, it should help you to understand the controversies surrounding statistical inference and developments in the practice of applying statistical inference that are taking place. Investing time and energy in fully understanding the principles of statistical inference really pays off later.

1.2 A Discrete Random Variable: How Many Yellow Candies in My Bag?

An obvious but key insight in statistical inference is this: If we draw random samples from the same population, we are likely to obtain different samples. No two random samples from the same population need to be identical, even though they can be identical.

1.2.1 Sample statistic

We are usually interested in a particular characteristic of the sample rather than in the exact nature of each observation within the sample. For instance, I happen to be very fond of yellow candies. If I buy a bag of candies, my first impulse is to tear the bag open and count the number of yellow candies. Am I lucky today? Does my bag contain a lot of yellow candies?

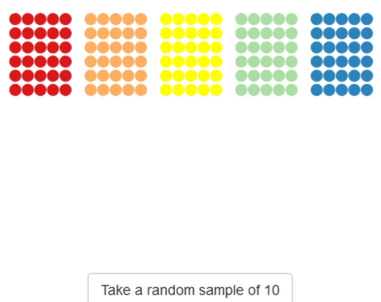


Figure 1.1: How many yellow candies will our sample bag contain?

1. Figure 1.1 shows a population of candies. What do you expect is the number of yellow candies in a random sample of ten candies from this population? Draw several samples and check whether your expectation comes true.

```
* The colours are equally distributed in the population, so one out of each
five candies in the population is yellow. In other words, the proportion of
yellow candies in the population is .2.
* This is the proportion that we would also expect in the sample. A sample
contains ten candies, so two out of these ten are expected to be yellow. If we
draw several samples, we notice that only a minority of our samples contain
exactly two yellow candies.
```

2. What are all the possible outcomes for the number of yellow candies? This collection of all possible outcome scores is called the *sampling space*.

```
* In a sample of ten candies, zero to ten candies can be yellow.
* The numbers 0, 1, 2, ..., 9, 10 constitute all possible outcomes for the
```

```
sample statistic 'Number of yellow candies'. This is called the sampling
space.
```

The number of yellow candies in a bag is an example of a *sample statistic*: a number describing a characteristic of the sample. Each bag, that is, each sample, has one outcome score on the sample statistic. For instance, one bag contains four yellow candies, another bag contains seven, and so on.

The sample statistic is called a *random variable*. It is a variable because it assigns an outcome score to a sample and different samples can have different scores. The value of a random variable may vary from sample to sample. It is a random variable because the score depends on chance, namely the chance that particular elements are drawn during random sampling.

1.2.2 Sampling distribution

Some sample statistic outcomes occur more often than other outcomes. We can see this if we draw very many random samples from a population and collect the frequencies of all outcome scores in a table or chart. We call the distribution of the outcome scores of very many samples a *sampling distribution*.

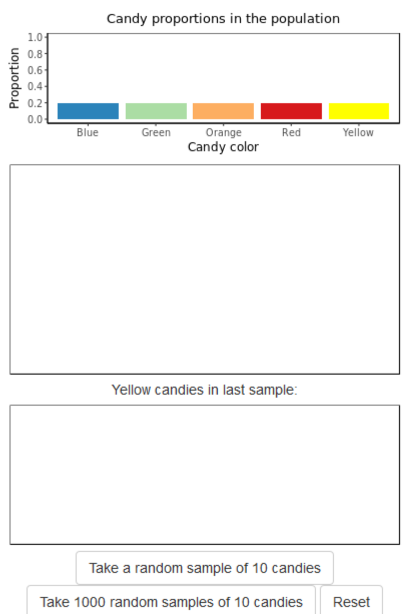


Figure 1.2: What is a sampling distribution?

1. Draw a random sample of ten candies in Figure 1.2. What do the numbers on the horizontal axis of the bottom histograms mean? And what does the vertical axis of

this histogram represent?

```
* The numbers on the horizontal axis specify the sampling space, that is, all
values that the sample statistic "Number of yellow candies" can take.
* The vertical axis shows the number of samples that have been drawn with a
particular value for the sample statistic, that is, with a particular number
of yellow candies in the sample.
```

2. What are the cases (units of analysis) in the three histograms? Hint: There are two different types of cases.

```
* In the upper two graphs, candies are the cases. A candy has a particular
colour, not a (sample) bag of candies.
* In the bottom graph showing the sampling distribution, samples (candy bags)
are the cases. A sample (bag) contains a particular nmber of yellow candies.
```

3. Guess the most likely and most unlikely outcome scores for the number of yellow candies in a sample bag containing ten candies in Figure 1.2. Check your intuitions by drawing 1,000 samples.

```
* If twenty per cent of candies in the population are yellow, we expect about
twenty per cent of candies in our sample to be yellow. Our sample contains ten
candies, so we expect two yellow candies in our sample. Indeed, samples with
two yellow candies are most frequent if we draw 1,000 random samples.
* If we expect two candies, samples are more unlikely if they contain a number
of yellow candies that is further away from two. So we expect the sample
counts to decrease if we move away from two in the sampling distribution. Ten
yellow candies is furthest away from two in our sampling space, so a sample
bag with ten yellow candies is most unlikely.
```

4. After how many samples does the shape of the sampling distribution stop changing?

```
* Theoretically, the sampling distribution represents an infinite number of
samples. In this example, the sampling distribution usually has its final
shape already after drawing 1,000 or 2,000 samples. But you can be unlucky, of
course, and need more samples to arrive at a stable distribution.
```

1.2.3 Probability and probability distribution

What is the probability of buying a bag with exactly five yellow candies? In statistical terminology, what is the probability of drawing a sample with five yellow candies as sample statistic outcome? This probability is the proportion of all possible samples that we could have drawn that happen to contain five yellow candies.

Of course, the probability of a sample bag with exactly five yellow candies depends on the share of yellow candies in the population of all candies. Figure 1.3 displays the probabilities of a sample bag with a particular number of yellow candies if twenty per cent of the candies in the population are yellow. You can adjust the population share of yellow candies to see what happens.

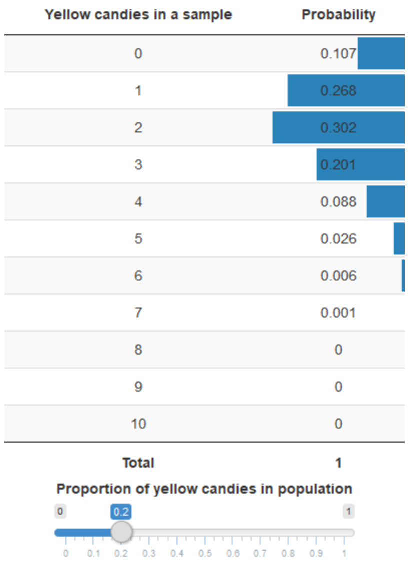


Figure 1.3: How does the probability of drawing a sample bag with two out of ten candies yellow depend on the proportion of yellow candies in the population?

1. In Figure 1.3, what is the sample statistic and what is the sampling space?

```
* The number of yellow candies in the sample (bag) is the sample statistic.
This is the characteristic of the sample (bag) that we are interested in.
* The set of all possible outcomes of the sample statistic is the sampling
space. In this example, the sampling space is the set of (integer) numbers
from 0 to 10.
```

2. Which number of yellow candies is most likely to be found in a sample bag of ten candies? How does this relate to the proportion of candies in the population?

```
* The numbers and horizontal bars in the Probability column represent the
probabilities of outcomes. In the initial situation, the highest probability
is found for a sample bag containing two yellow candies (p = .302). This
amounts to two out of the ten candies in the sample bag, that is, twenty per
cent.
* This percentage is equal to the precentage of yellow candies in the
population. We are most likely to draw a sample with a percentage or
proportion that is equal to the population percentage, here p = .302, even
though the total probability of drawing a sample with another percentage of
yellow candies is higher (p = 1 - .302 = .698).
```

3. What is the probability that a sample bag of ten candies contains not more than three yellow candies if the proportion in the population is .2?

```
* At most three out of ten candies means that we have to sum the probabilities
of zero, one, two, and three yellow candies. This probability equals .107 +
.268 + .302 + .201 = .878. That is a fair chance.
```

4. What do you expect to happen to the probabilities if you increase the proportion of yellow candies in the population (factory stock)? Use the slider to check your answer.

```
* If a larger part of the candies is yellow in the population, we should
expect more yellow candies in our sample bag. The probabilities of small
numbers of yellow candies (low outcome values) should go down whereas the
probabilities of large numbers (high outcome scores) should go up.
* If you move the slider to the right, you will see that the distribution
shifts down in the table.
```

5. What is special about the distribution if the proportion of yellow candies in the population is .5?

```
* If the population proportion is .5, the probability distribution is
symmetric. The probability of a sample bag with four candies is equal to the
probability of a sample bag with six candies. Probabilities are equal for
three and seven yellow candies, two and eight, one and nine, zero and ten
yellow candies.
* The distribution has the classic bell shape of a normal distribution that we
will encounter when we discuss continuous probability distributions.
```

The sampling distribution tells us all possible samples that we could have drawn, that is, if we have drawn very many samples. We can use the distribution of all samples to get the probability of buying a bag with exactly five yellow candies from the sampling distribution: We merely divide the number of samples with five yellow candies by the total number of samples we have drawn.

If we change the (absolute) frequencies in the sampling distribution into proportions (relative frequencies), we obtain the *probability distribution* of the sample statistic: A sampling space with a probability (between 0 and 1) for each outcome of the sample statistic. Because we are usually interested in probabilities, sampling distributions tend to have proportions, that is probabilities, instead of frequencies on the vertical axis. See Figure 1.4 for an example. In this case, the sampling distribution is a probablity distribution.

Figure 1.3 displays the probability distribution of the number of yellow candies per bag of ten candies. This is an example of a *discrete probability distribution* because only a limited number of outcomes are possible. It is feasible to list the probability of each outcome separately.

The sampling distribution as a probability distribution conveys very important information. It tells us which outcomes we can expect, in our example, how many yellow candies we may find in our bag of ten candies. Moreover, it tells us the probability that a particular outcome may occur. If the sample is drawn from a population in which twenty per cent of candies are yellow, we are quite likely to find zero, one, two, three, or four yellow candies in our bag. A bag with five yellow candies would be rare, six or seven candies would be very rare, and a bag with more than seven yellow candies is extremely unlikely even though it is not impossible. If we buy such a bag, we know that we have been extremely lucky.

We may refer to probabilities both as a proportion, that is, a number between 0 and 1, and as a percentage: a number between 0% and 100%. Proportions are commonly considered to be the correct way to express probabilities. When we talk about probabilities, however, we tend to use percentages; we may, for example, say that the probabilities are fifty-fifty.

1.2.4 Expected value or expectation

We haven't yet thought about the value that we are most likely to encounter in the sample that we are going to draw. Intuitively, it must be related to the distribution of colours in the population of candies from which the sample was drawn. In other words, the share of yellow candies in the factory's stock from which the bag was filled or in the machine that produces the candies, seems to be relevant to what we may expect to find in our sample.

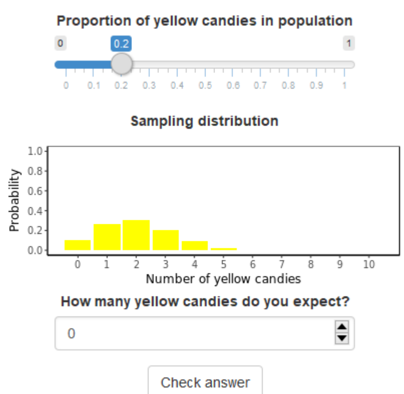


Figure 1.4: What is the expected value of a probability distribution?

1. In Figure 1.4, which number of yellow candies is most likely to occur in a sample bag of ten candies? How does this number change if you change the proportion of yellow candies in the population?

```
* We expect that the proportion of yellow candies in the sample equals the
population proportion, which initially is .2. In a sample of ten candies,
then, the expected number of yellow candies is two.
* Note that two is the outcome with the highest probability in the sampling
distribution.
* The expected number of yellow candies in a sample bag of ten candies is ten
times the population proportion, so the expected number of candies in the
sample changes in accordance with changes in the population proportion.
```

2. How does the mean of the sampling distribution relate to the expected value? Experiment with different values for the population proportion.

```
* The mean of the sampling distribution is equal to the expected value of the
sample statistic. In the initial example, both are two.
* This makes sense: We are equally likely to draw a sample with less yellow
candies than the expected number as a sample with more yellow candies. Samples
with less yellow candies than expected should balance sample with more yellow
candies than expected. The mean represents the "balance point" of a
distribution.
```

3. How does the mean of the sampling distribution relate to the population proportion? Experiment with different values for the population proportion.

```
* The mean of the sampling distribution is equal to the population proportion
if the sample statistic is the sample proportion. If we would have created the
sampling distribution of the proportion of yellow candies in our sample, the
mean of the sampling distribution would be equal to the proportion of yellow
candies in the population.
* Again, we are equally likely to draw a sample with less yellow candies than
the expected proportion as a sample with more yellow candies. Samples with
less yellow candies than expected should balance sample with more yellow
candies than expected. The mean represents the "balance point" of a
distribution.
* Note that the expected value (mean of the sampling distribution) only equals
the population value if the sample statistic is an unbiased estimate of the
population value (parameter). See the next section.
```

If the share of yellow candies in the population is 0.20 (or 20%), we expect one out of each five candies in a bag (sample) to be yellow. In a bag with 10 candies, we would expect two candies to be yellow: one out of each five candies or the population proportion times the total number of candies in the sample = 0.20 * 10 = 2.0. This is the expected value.

The expected value of the proportion of yellow candies in the sample is equal to the proportion of yellow candies in the population. If you carefully inspect a sampling distribution (Figure 1.4), you will see that the expected value also equals the mean of the sampling distribution. This makes sense: Excess yellow candies in some bags must be compensated for by a shortage in other bags.

Thus we arrive at the definition of the *expected value* of a random variable:

> The expected value is the average of the sampling distribution of a random variable.

In our example, the random variable is a sample statistic, more sepcifically, the number of yellow candies in a sample.

The sampling distribution is an example of a probability distribution, so, more generally, the expected value is the average of a probability distribution. The expected value is also called the *expectation* of a probability distribution.

1.2.5 Unbiased estimator

Note that the expected value of the proportion of yellow candies in the bag (sample statistic) equals the true proportion of yellow candies in the candy factory (population statistic). This is always, that is, by definition, true for all sample statistics that are *unbiased estimators* of the population statistic. By the way, we usually refer to the population statistic as a *parameter*.

Most but not all sample statistics are unbiased estimators of the population statistic. Think, for instance, of the actual number of yellow candies in the sample. This is certainly not an unbiased estimator of the number of yellow candies in the population. Because the population is so much larger than the sample, the population must contain many more yellow candies than the sample. If we were to estimate the number in the population (the parameter) from the number in the sample—for instance, we estimate that there are two yellow candies in the population of all candies because we have two in our sample of ten—we are going to vastly underestimate the number in the population. This estimate is *downward biased*: It is too low.

In contrast, the proportion in the sample is an unbiased estimator of the population proportion. That is why we do not use the number of yellow candies to generalize from our sample to the population. Instead, we use the proportion of yellow candies. You probably already did this intuitively.

Sometimes, we have to adjust the way in which we calculate a sample statistic to get an unbiased estimator. For instance, we must calculate the standard deviation and variance in the sample in a special way to obtain an unbiased estimate of the population standard deviation and variance. The exact calculation need not bother us, because our statistical software will take care of that.

1.2.6 Representative sample

Because the share of yellow candies in the population represents the probability of drawing a yellow candy, we also expect 20% of the candies in our bag to be yellow. For the same reason we expect the shares of all other colours in our sample bag to be equal to their shares in the population. As a consequence, we expect a random sample to resemble the population from which it is drawn.

A sample is *representative* of a population if variables in the sample are distributed in the same way as in the population. Of course, we know that a random sample is likely to differ from the population due to chance, so the actual sample that we have drawn is usually not representative of the population.

But we should expect it to be representative, so we say that it is *in principle representative* of the population. We can use probability theory to account for the misrepresentation in the actual sample that we draw. This is what we do when we use statistical inference to construct confidence intervals and test null hypotheses.

1.3 A Continuous Random Variable: Overweight And Underweight.

Let us now look at another variable: the weight of candies in a bag. The weight of candies is perhaps more interesting to the average consumer because it is related to the number of calories that a candy contains.

1.3.1 Continuous variable

Weight is a *continuous variable* because we can always think of a new weight between two other weights. For instance, consider two candy weights: 2.8 and 2.81 gram. It is easy to see that there can be a weight in between these two values, for instance, 2.803 gram. Between 2.8 and 2.803 we can discern an intermediate value such as 2.802. In principle, we could continue doing this endlessly, e.g., find a weight between 2.80195661 and 2.80195662 gram even if our scales may not be sufficiently precise to measure any further differences. It is the principle that counts. If we can always think of a new value in between two values, the variable is continuous.

1.3.2 Continuous sample statistic

We are not interested in the weight of a single candy. If a relatively light candy is compensated for by a relatively heavy candy in the same bag, we still get the calories that we want. We are interested in the average weight of all candies in our sample bag, so average candy weight in our sample bag is our key sample statistic. We want to use this sample statistic to say something about average candy weight in the population of all candies. Can we do that?

The sample mean is an unbiased estimator of the population mean, so the average weight of all candies in the population (at the factory) is the average of the (candy weights in the) sampling distribution. And this is the average weight that we expect in a sample drawn from this population (the expected value or expectation). So far, everything is the same as in the case of the proportion of yellow candies, which was a discrete random variable because it could take only a limited set of values: yellow, blue, green, red, and orange.

1.3.3 Continuous probabilities

When we turn to the probabilities of getting samples with a particular average candy weight, we run into problems with a continuous sample statistic. If we would want to know the probability of drawing a sample bag with an average candy weight of 2.8 gram, we should exclude sample bags with an average candy weight of 2.81 gram, or 2.801 gram, or 2.8000000001 gram, and so on. In fact, we are very unlikely to draw a sample bag with an average candy weight of exactly 2.8 gram, that is, with an infinite number of zeros trailing 2.8. In other words, the probability of such a sample bag is for all practical purposes zero and negligible.

This applies to every average candy weight, so all probabilities are virtually zero. As a consequence, we cannot construct a probability distribution of the sampling space, that is, of all possible outcomes. By the way, note that this is also impossible because we have an infinite number of possible outcomes. After all, we can always find a new weight between two selected weights.

1.3.4 p Values

We can solve this problem by looking at a range of values instead of a single value. We can meaningfully talk about the probability of having a sample bag with an average candy weight of at least 2.8 gram or at most 2.8 gram. We choose a threshold, in this example 2.8 gram, and determine the probability of values above or below this threshold. We can also use two thresholds, for example the probability of an average candy weight between 2.75 and 2.85 gram. This is probably what you were thinking of when I referred to a bag with 2.8 gram as average candy weight.

If we cannot determine the probability of a single value, which we used to depict on the vertical axis, and we have to link probabilities to a range of values on the x axis, for example, average candy weight above/below 2.8 gram, how can we display probabilities? We have to display a probability as an area between the horizontal axis and a curve. This curve is called a *probability density function*, so if there is a label to the vertical axis of a continuous probability distribution, it usually is "Probability density" instead of "Probability".

Figure 1.5 shows an example of a continuous probability distribution for the average weight of candies in a sample bag. This is the familiar normal distribution so we could say that the normal function is the probability density function here. The total area under this curve is set to one, so the area belonging to a range of sample outcomes (average candy weight) is 1 or less, as probabilities should be.

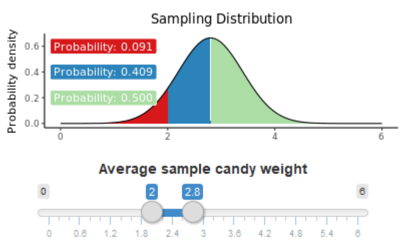


Figure 1.5: How do we display probabilities in a continuous sampling distribution?

1. In Figure 1.5, what is the probability of buying a bag with average candy weight of 2.8 gram or more?

```
* The probability is .5. It is represented by the green surface under the
curve. This is exactly half of the total surface under the curve because 2.8
```

```
gram is the average candy weight in the population and, as a result, the
average value of the sampling distribution of average sample candy weight.
```

2. Is this a left-hand probability, a right-hand probability, or neither?

```
* This is a right-hand probability because it specifies a threshold value
(2.8) and all values that are larger. It concerns the right-hand tail of the
sampling distribution.
```

3. Use the sliders to find the probability of buying a bag with average candy weight between 2.6 and 3.7 gram. Is this a left-hand probability, a right-hand probability, or neither?

```
* If you set the slider handles to 2.6 and 3.7, the blue area represents the
probability that you are looking for. Its value is reported as .564.
* This is neither a left-hand or right-hand probability because it does not
include either the left-hand or right-hand tail of the sampling distribution.
```

4. What is the minimum average weight of the 10% heaviest candy bags?

```
* Drag the rigth-hand slider handle to the maximum value (6) and then adjust
the left-hand slider handle so the (blue) area in the right-tail of the
sampling distribution represent a probability of .100. This area contains the
ten per cent samples with largest average candy weight scores. The value
displayed with the left-hand slider is the minimum average candy weight for
these samples.
```

The probability of values up to (and including) the threshold value or the threshold value and higher are called *p values*. The probability of values up to (and including) the threshold value is known as the *left-hand p value* and the probability of values above (and including) the threshold value is called the *right-hand p value*.

Why did I put *(and including)* between parentheses? It does not really matter whether we add the exact boundary value (2.8 gram) to the probability on the left or on the right because the probability of getting a bag with average candy weight at exactly 2.8 gram (with a very long trail of zero decimals) is negligible.

Are you struggling with the idea of areas instead of heights (values on the vertical axis) as probabilities? Just consider the idea that we could use the area of a bar in a histogram instead of the height as indication of the probability in discrete probability distrbutions, for example, Figure 1.4. After all, the bars in a histogram are all equally wide, so differences between bar areas are proportional to differences in bar height.

1.3.5 Probabilities always sum to 1

While you were playing with Figure 1.5, you may have noticed that displayed probabilities always add up to one. This is true for every probability distribution as you have learned before. In addition, you may have realized that the probabilities can also be interpreted as proportions or percentages. The probability that a sample bag contains candies with average weight over 2.9 gram is equal to the proportion of samples in the sampling distribution with average candy weight over 2.9 gram. Thus, we can use probabilities to find the threshold values that separate the top ten per cent or the bottom five per cent in a distribution.

1.4 Concluding Remarks

A communication scientist wants to know whether children are sufficiently aware of the dangers of media use. On a media literacy scale from one to ten, an average score of 5.5 or higher is assumed to be sufficient.

If we translate this to the simple candy bag example, we realize that the outcome in our sample need not be the true population value. After all, we could very well draw a bag with less or more than twenty per cent yellow candies.

Average media literacy, then, can exceed 5.5 in our sample of children, even if average media literacy is below 5.5 in the population or the other way around. How we decide on this is discussed in later chapters.

1.4.1 Samples characteristics as observations

Perhaps the most confusing aspect of sampling distributions is the fact that samples are our cases (units of analysis) and sample characterstics are our observations. We are accustomed to think of observations as measurements on empirical *things* such as people or candies. We perceive each person or each candy as a case and we observe a characteristic that may change across cases (a variable), for instance the colour of a candy or the weight of a candy.

In a sampling distribution, however, we observe samples (cases) and measure a sample statistic as the (random) variable. Each sample adds one observation to the sampling distribution and its sample statistic value is the value added to the sampling distribution.

1.4.2 Means at three levels

In our first example, the sample statistic is a proportion, namely the proportion of candies that are yellow. The horizontal axis of the sampling distribution represents the proportion of yellow candies. This is fully in line with our research question. If we want to know whether all colours are equally distributed, we are interested in sample proportions, not in properties of individual candies.

Things become a little confusing if we are interested in a sample mean, such as the average weight of candies in a sample bag. Now we have means at three levels: the population, the sampling distribution, and the sample.

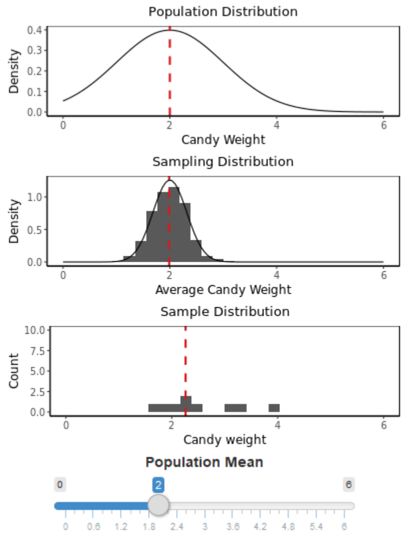


Figure 1.6: What is the relation between the three distributions?

1. In Figure 1.6, explain the meaning of the three means (dotted red lines). Which mean is a mean of means?

```
* The dotted red line in the top graph represents the average of candy weights
in the population.
* The dotted red line in the middle graph represents the average of sampling
means. So it is the average of average candy weight in a sample. It is a mean
of means.
* The dotted red line in the bottom graph represents the average of candy
weights in the sample.
```

2. Is it a coincidence that the mean of the population and sampling distribution are the same? Use a slider to check if these means are the same.

```
* This is not a coincidence because the population mean is the expected value
of the sampling distribution and the expected value of the sampling
distribution is the average of the sampling distribution.
```

3. How does the sample mean relate to the population mean and the mean of the sampling distribution?

```
* The sample mean is one of the sample means included in the sampling
distribution. The sample is, so to speak, one of the possible samples that are
listed in the sampling distribution.
* The sampling distribution depends on the population distribution. For
instance, the mean or centre of the sampling distribution of sample means
equals the population mean (because the sample mean is an unbiased estimator
of the population mean).
```

```
* If the population changes, the sampling distribution changes, so the list of
possible samples (better: the probabilities of sample outcomes) changes, so we
expect a different sample. But this sample need not have the same mean as the
population and the sampling distribution because it is drawn at random. The
sample mean is more likely to be close to the population mean but it can be
quite different from it.
```

The sampling distribution, here, is a distribution of sample means but the sampling distribution itself also has a mean, which is called the expected value or expectation of the sampling distribution. Don't get confused about this. The mean of the sampling distribution is the average of the average weight of candies across all possible samples. This mean of means has the same value as our first mean, namely the average weight of the candies in the population because a sample mean is an unbiased estimator of the population mean.

Think of the three distributions as a hamburger. The top and bottom part of a hamburger are both made of bread. They represent the population and the sample, which consist if the same substance: candies and their weight in our example. The middle part of the hamburger, however, is a completely different type of food. The meat holds the two halves of the bun together. In a similar way, the sampling distribution connects the population to the sample but it is of a very different substance. It consists of samples, for instance, bags of candies instead of single candies.

The sampling distribution sticks to the population because the population statistic (parameter), for example, the average weight of all candies, is equal to the mean of the sampling distribution. The sampling distribution sticks to the sample because it tells us which sample means we will find with what probabilities. The sampling distribution is the vital link connecting the sample to the population. We need it to make statements about the population based on our sample.

1.5 Test Your Understanding

Figure 1.7 simulates drawing random samples from a candy factory's stock of candies. We are interested in the colour of the candies in our sample. The histogram shows the distribution of candies according to colour. Draw some samples and have a look at the number of yellow candies in each sample as well as the average number of yellow candies over all samples.

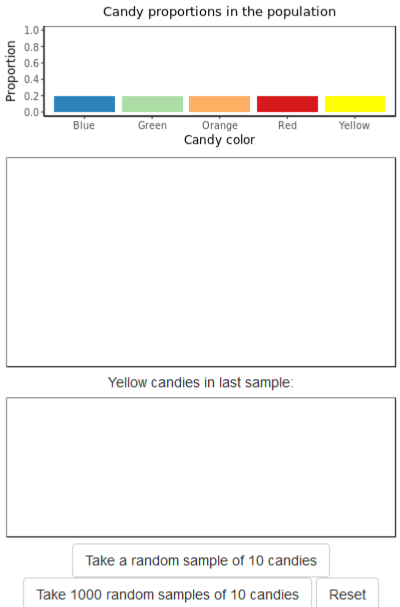


Figure 1.7: A discrete sampling distribution.

1. Figure 1.7 shows a simulated population distribution. What would be the real-world population?

2. Use the button in Figure 1.7 to draw one random sample of ten candies from the population. What do the numbers on the horizontal axis of the bottom histogram represent? What is the statistical name of the variable *Number of yellow candies*? What is the unit of analysis for this characteristic?

3. Which values can the sample characteristic take here and what is the statistical name for this set of values?

4. If you would draw many samples from this population each containing ten candies, what is the number of yellow candies per sample that appears most frequently? Draw 1,000 samples to verify your answer.

5. Is the colour distribution in each sample that you draw representative of the colour distribution in the stock of candies?

6. Why, do you think, is the sample characteristic called a *random variable*?

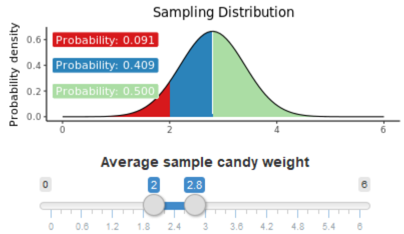


Figure 1.8: A continuous sampling distribution.

7. Use your own words to explain what the sampling distribution in Figure 1.8 represents.
8. What do you think is the average weight of all candies in the population? Justify your answer using the concepts *expected value* and *unbiased estimator*.
9. Use the sliders to find the probability of drawing a sample with average candy weight between 2.0 and 2.9 grams.
10. What, do you expect, is the probability of drawing a sample with average candy weight of exactly 2.9 grams? Use the sliders to check your expectation.
11. Why is this graph an example of a continuous probability distribution?
12. Why is the vertical axis labelled with "Probability density" instead of "Probability"?

1.6 Take-Home Points

- Values of a sample statistic vary across random samples from the same population. But some values are more probable than other values.
- The sampling distribution of a sample statistic tells us the probability of drawing a sample with a particular value of the sample statistic or a particular minimum/maximum value.
- If a sample statistic is an unbiased estimator of a parameter, the parameter value is the average of the sampling distribution, which is called the expected value or expectation.
- For discrete sample statistics, the sampling distribution tells us the probability of individual sample outcomes. For continuous sample statistics, it tells us the p value: the probability of drawing a sample with an outcome that is at least or at most a particular value.

Chapter 2

Probability Models: How Do I Get a Sampling Distribution?

Key concepts: bootstrapping/bootstrap sample, sampling with replacement, exact approach, approximation with a theoretical probability distribution, binomial distribution, (standard) normal distribution, (Student) t distribution, F distribution, chi-squared distribution, condition checks for theoretical probability distributions, sample size, equal population variances, expected values, independent samples, dependent/paired samples.

Summary

How do we get a sampling distribution without drawing many samples ourselves?

In the previous chapter, we drew a large number of samples from a population to obtain the sampling distribution of a sample statistic, for instance, the proportion of yellow candies or average candy weight in the sample. The procedure is quite simple: Draw a sample, calculate the desired sample statistic, add the sample statistic value to the sampling distribution, and repeat this thousands of times.

Although this procedure is simple, it is not practical. In a research project, we would have to draw thousands of samples and administer a survey to each sample or collect data on the sample in some other way. This requires too much time and money to be of any practical value. So how do we create a sampling distribution, if we only collect data for a single sample? This chapter presents three ways of doing this: bootstrapping, exact approaches, and theoretical approximations.

2.1 The Bootstrap Approximation of the Sampling Distribution

The first way to obtain a sampling distribution is still based on the idea of drawing a large number of samples. However, we only draw one sample from the population for which we collect data. As a next step, we draw a large number of samples from our initial sample. The samples drawn in the second step are called *bootstrap samples*. The technique was developed by Bradley Efron (1979; 1987). For each bootstrap sample, we calculate the sample statistic of interest and we collect these as our sampling distribution. We usually want about 5,000 bootstrap samples for our sampling distribution.

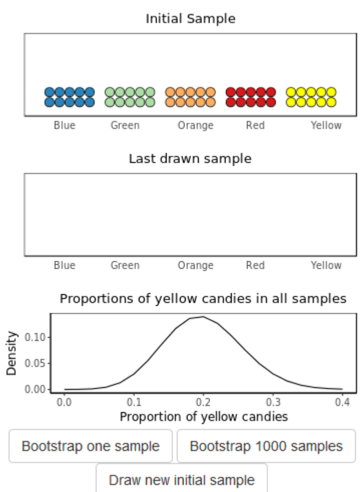


Figure 2.1: How do we create a sampling distribution with bootstrapping?

In Figure 2.1, an initial sample has been drawn from a population containing five candy colours in equal proportions.

1. How large is a bootstrap sample in Figure 2.1? Formulate and motivate your answer before you check it with the *Bootstrap one sample* button.

```
* A bootstrap sample must be just as large as the initial sample. The size of
a sample is very important to the sampling distribution, so we must draw
bootstrap samples with exactly the same number of observations as the initial
sample.
```

2. What element in Figure 2.1 represents the true sampling distribution in this example? If in doubt, see Figure 1.3.

```
* The sampling distribution of a sample proportion is an exact distribution
(named binomial distribution): the probabilities of every number or proportion
of yellow candies in the sample can be calculated. The results are displayed
```

```
as a grey histogram at the bottom of the figure.
```

3. Does the bootstrap sampling distribution resemble the true sampling distribution? Use the “Bootstrap 1000 samples” button and justify your answer.

```
* That depends. If the proportion of yellow candies in the original sample is
close to .2, that is, ten out of fifty candies in the sample are yellow, the
bootstrapped sampling distribution (yellow histogram) is very similar to the
true (exact) sampling distribution (grey histogram).
* If there are much less or many more than ten yellow candies in the sample,
however, the bootstrapped sampling distribution is quite different from the
true sampling distribution. Especially the mean (horizontal location) of the
bootstrapped sampling distribution is different. The shape of the distribution
may still be nearly the same.
```

4. Draw a new initial sample. This sample is probably less representative of the distribution of candy colour in the population. What happens to the bootstrap samples and the bootstrap sampling distribution?

```
* See the answer to Exercise 3.
```

The *bootstrap* concept refers to the story in which Baron von Münchhausen saves himself by pulling himself and his horse by his bootstraps (or hair) out of a swamp. In a similar miraculous way, the bootstrap samples resemble the sampling distribution even though they are drawn from a sample instead of the population. This miracle requires some explanation and it need not work always, as we will discuss in the remainder of this section.

Figure 2.2: Baron von Münchhausen pulls himself and his horse out of a swamp.

2.1.1 Sampling with and without replacement

As we will see in a later chapter, the size of a sample is very important to the shape of the sampling distribution. The sampling distribution of samples with twenty cases can be very different from the sampling distribution of samples with forty cases. To construct a sampling distribution from bootstrap samples, the bootstrap samples must be exactly as large as the original sample.

How can we draw many different bootstrap samples from the original sample if each bootstrap sample must contain the same number of cases as the original sample?

[htbp]

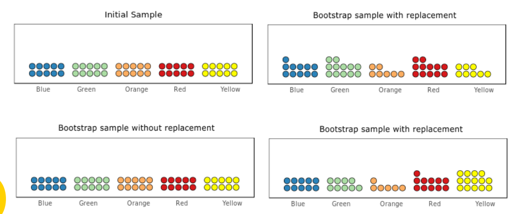


Figure 2.3: Sampling with and without replacement.

1. What are the differences between sampling with and without replacement (Figure 2.3)?

```
* If we draw a sample without replacement from our initial sample of the same
size as the initial sample, the new sample must contain all obervations from
the initial sample. As a result, the new sample is identical to the initial
sample. All samples that we draw are identical.
* Drawing with replacement, an observation can be drawn more than once. This
must have occurred for at least one red candy in the two examples of bootstrap
samples with replacement in the figure. Otherwise, we could never have more red
candies (twelve or eleven) in the bootstrap sample than in the initial sample
(ten red candies).
```

If we allow every case in the original sample to be sampled only once, each bootstrap sample contains all cases of the original sample, so it is an exact copy of the original sample. Thus, we cannot create different bootstrap samples.

By the way, we often use the type of sampling described above, which is called *sampling without replacement*. If a person is (randomly) chosen for our sample, we do not put this person back into the population so she or he can be chosen again. We want our respondents to fill out our questionnaire only once or participate in our experiment only once.

If we do allow the same person to be chosen more than once, we sample *with replacement*. The same person can occur more than once in a sample. Bootstrap samples that are sampled with replacement from the original sample can vary because they need not contain all cases in the original sample. Some cases may not be sampled while other cases are sampled several times. You probably have noticed this in Figure 2.3. Sampling with replacement allows us to obtain different bootstrap samples from the original sample, and still have bootstrap samples of the same size as the original sample.

2.1.2 Calculating probabilities with replacement

You may wonder whether it is OK to sample with replacement when we are bootstrapping. The short answer is: Yes it is. We usually calculate probabilities as if we sampled with replacement. Suppose we want to calculate the probability of picking two yellow candies from a population in which 20% of the candies are yellow. The probability of picking two yellow candies is then calculated as .200 * .200: twice the probability of drawing a yellow candy.

In this calculation, we assume that the probability to draw a yellow candy remains the same while we are sampling. The probability of sampling a yellow candy is assumed to be .200 whether we sample the first or the second candy. We act as if the proportion of yellow candies in the population remains the same, namely 20%. This assumption is very convenient because it simplifies the calculation of probabilities.

2.1.3 Calculating probabilities without replacement

In practice, however, we never want the same respondent to participate twice in our research because this would not yield new information. In actual research, then, we sample *without* replacement. A respondent does not risk being sampled more than once.

How about calculating probabilities if we sample without replacement? If we do not put the sampled yellow candy back in the population, the number of yellow candies in the population is reduced by one after we have drawn the first yellow candy. The probability of drawing a second yellow candy should then be less than 20%.

If the population is large, the decrease in the probability is too small to be in any way relevant. For instance, if we have a population of one million candies and 20% is yellow, the probability of drawing the first yellow candy is 200,000 / 1,000,000 = .200. The probability of drawing the second yellow candy would be 199,999 / 999,999 = 0.1999992 ; the difference between the two probabilities (0,0000008) is negligible.

As you can see, calculating probablities becomes complicated if we sample without replacement. But we can ignore the complications if the population is much larger than the sample because the probabilities hardly differ from the simple probabilities that we have if we sample with replacement.

In practice, then, we always sample *without* replacement but our statistical software calculates probabilities as if we sampled *with* replacement. This is perectly fine as long as the population is much larger than the sample. If the sample contains a large share of the population, however, we should not trust the probabilities that our software reports.

2.1.4 Limitations to bootstrapping

Does the bootstrapped sampling distribution always reflect the true sampling distribution?

Figure 2.4: How is bootstrapping influenced by sample size?

1. When does the bootstrap sampling distribution (yellow histogram) reflect the true sampling distribution (grey histogram) better: at small or large sample sizes? Play with sample size in Figure 2.4 to check your answer.

```
* If you change sample size repeatedly between 15 and 45, you will see that
the bootstrapped sampling distribution (yellow histogram) jumps around the
true sampling distribution (grey histogram).
* For relatively small sample sizes, the bootstrapped sampling distribution is
often quite different from the true sampling distribution.
* At larger sample sizes, say between 120 and 150, the bootstrapped sampling
distribution overlaps the true sampling distribution much more frequently. So
for larger samples, we can trust the bootstrapped sampling distribution more.
But even then, it can sometimes be quite off the mark.
```

2. How does sample size relate to representativeness of the sample? Twenty per cent of the candies in the population are yellow.

```
* The proportion of yellow candies in larger samples is more often close to
the proportion in the population: 0.2. This is the reason that the
bootstrapped sampling distribution usually resembles the true sampling
distribution.
```

3. If you use a very small sample size, it may happen that there is no yellow histogram in the bottom graph. What is the matter if that happens?

```
* If we draw an initial sample without any yellow candies, none of the
bootstrap samples can include yellow candies. As a result, the count of
samples with yellow candies is always zero.
```

```
* The smaller the initial sample, the greater the chance of having a sample
without yellow candies.
```

We can create a sampling distribution by sampling from our original sample with replacement. It is hardly a miracle that we obtain different samples with different sample statistics if we sample with replacement. Much more miraculous, however, is that this bootstrap distribution resembles the true sampling distribution that we would get if we draw lots of samples directly from the population.

Does this miracle always happen? No, it need not happen. First, the original sample that we have drawn from the population must not be too small. We cannot draw many different samples from a small sample. For this reason, the bootstrap distribution cannot resemble the true sampling distribution, in this situation.

Second, the original sample must be more or less representative of the population. The variables of interest in the sample should be distributed more or less the same as in the population. If this is not the case, the sampling distribution may be biased, giving a distorted view of the true sampling distribution.

A sample is more likely to be representative of the population if the sample is drawn in a truly random fashion and if the sample is larger. But we can never be sure. There always is a chance that we have drawn a sample that does not reflect the population well. This is the main problem with the bootstrap approach to sampling distributions.

2.1.5 Any sample statistic can be bootstrapped

The big advantage of the bootstrap approach (*bootstrapping*), however, is that we can get a sampling distribution for any sample statistic that we are interested in. Every statistic that we can calculate for our original sample can also be calculated for each bootstrap sample. The sampling distribution is just the collection of the sample statistics calculated for all bootstrap samples.

Bootstrapping is more or less the only way to get a sampling distribution for the sample median, for instance, the median weight of candies in a sample bag. We may create sampling distributions for the wildest and weirdest sample statistics, for instance the difference between sample mean and sample median squared. I would not know why you would be interested in the squared difference of sample mean and median, but there are very interesting statistics that we can only get at through bootstrapping. A case in point is the strength of an indirect effect in a mediation model (Chapter 9).

2.2 Bootstrapping in SPSS

2.2.1 Instructions

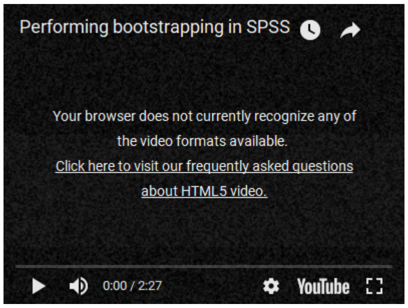


Figure 2.5: Bootstrapping in SPSS.

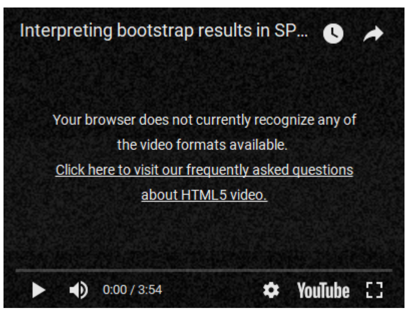


Figure 2.6: Interpreting bootstrap results in SPSS.

In principle, any sample statistic can be bootstrapped. SPSS, however, does not bootstrap all sample statistics. For example, SPSS does not bootstrap the minimum value, maximum value or the range between minimum and maximum value of a variable.

2.2.2 Exercises

1. Download the data set candies.sav and use SPSS to bootstrap the t test on average weight of yellow and red candies (the example above). The test is available in the *Analyze>Compare Means* menu.

```
SPSS syntax:

* Exercise 1: Bootstrap different averages.
* Check data.
```

```
FREQUENCIES VARIABLES=colour weight
  /ORDER=ANALYSIS.
* Execute independent-samples t test with bootstrap.
BOOTSTRAP
  /SAMPLING METHOD=SIMPLE
  /VARIABLES TARGET=weight INPUT=colour
  /CRITERIA CILEVEL=95 CITYPE=BCA  NSAMPLES=5000
  /MISSING USERMISSING=EXCLUDE.
T-TEST GROUPS=colour(4 5)
  /MISSING=ANALYSIS
  /VARIABLES=weight
  /CRITERIA=CI(.95).

Check data:

There are no impossible values on the two variables.

Check assumptions:

The measurement levels of the variables are OK:
colour is a categorical variable and weight is a
numeric variable. There are no additional assumptions
for bootstrapping.

Interpret the results:

The table "Bootstrap for Independent Samples Test" contains
the results that we are interested in.

Levene s test on homegeneity of variances is not statistically
significant, so we may assume that the population variances of
red and yellow candy weight are equal. So we interpret the
top row in table "Bootstrap for Independent Samples Test".

The mean difference between red and yellow candy weight is
-0.04 grams. In our sample, red candies are just a little
lighter than yellow candies.

The bootstrapped 95% confidence interval for this difference
is -0.21 to 0.11. Note that your results can be slightly
different because bootstrapping creates random samples.
In the population, red candies may be heavier of lighter
than yellow candies. We cannot tell which of the two it
is with enough confidence.
```

2. Use the same data set to bootstrap the median of candy weight. Remember that

measures of central tendency can be obtained with the *Frequencies>Statistics* command in the *Analyze>Descriptive Statistics* menu.

```
SPSS syntax:

* Exercise 2: Bootstrap on median candy weight.
* Check data.
FREQUENCIES VARIABLES=weight
  /ORDER=ANALYSIS.
* Bootstrap the median.
BOOTSTRAP
  /SAMPLING METHOD=SIMPLE
  /VARIABLES INPUT=weight
  /CRITERIA CILEVEL=95 CITYPE=BCA  NSAMPLES=5000
  /MISSING USERMISSING=EXCLUDE.
FREQUENCIES VARIABLES=weight
  /FORMAT=NOTABLE
  /STATISTICS=MEDIAN
  /ORDER=ANALYSIS.

Check data:

There are no impossible values on the weight variable.

Check assumptions:

The measurement level of variable weight is OK.

Interpret the results:

Median candy weight in the sample is 2.89 gram. With
95% confidence, we expect median candy weight to be
between 2.77 and 2.92 grams in the population of all candies.
```

2.3 Exact Approaches to the Sampling Distribution

A second approach to constructing a sampling distribution has implicitly been demonstrated in the preceding section on bootstrapping (Section 2.1) and the section on probability distributions (Section 1.2.3). In these sections, we calculated the true sampling distribution of the proportion of yellow candies in a sample from the probabilities of the colours. If we know or think we know the proportion of yellow candies in the population, we can exactly calculate the probability that a sample of ten candies includes one, two, three, or ten yellow candies. See the section on discrete random variables for details (Section 1.2).

1. Explain the meaning of the entries in the column **Combinations** and how they relate

Table 2.1: Number of heads for a toss of three coins.

Outcome	Combination	Probability: Combination	Probability: Outcome
0	tail-tail-tail	1/2 * 1/2 * 1/2 = 1/8 = .125	1/8 = .125
1	tail-tail-head	1/2 * 1/2 * 1/2 = 1/8 = .125	
1	head-tail-tail	1/2 * 1/2 * 1/2 = 1/8 = .125	
1	tail-head-tail	1/2 * 1/2 * 1/2 = 1/8 = .125	3/8 = .375
2	head-head-tail	1/2 * 1/2 * 1/2 = 1/8 = .125	
2	head-tail-head	1/2 * 1/2 * 1/2 = 1/8 = .125	
2	tail-head-head	1/2 * 1/2 * 1/2 = 1/8 = .125	3/8 = .375
3	head-head-head	1/2 * 1/2 * 1/2 = 1/8 = .125	1/8 = .125
Total	8		1.000

to the entries in the **Outcomes** columns.

```
* The column "Combinations" lists all posible outcomes if we toss three coins.
* The sample statistic is the number of heads in a throw of three coins. It
simply counts the number of heads that appear in the combination. This number
can range from zero to three. This is the sampling space.
```

2. Explain how the combinations relate to the probabilities.

```
* There are eight combinations. If the coins are fair, each combination has
the same probability to appear, namely 1/8 = .125. We sum this probability for
all combinations that have the same outcome, namely the same number of heads.
* Thus we arrive at the probability of having no heads in a throw (p = .125),
one head (p = .375), and so on.
```

The calculated probabilities of all possible sample statistic outcomes give us an exact approach to the sampling distribution. Note that I use the word *approach* instead of *approximation* here because the obtained sampling distribution is no longer an approximation, that is, more or less similar to the true sampling distribution. No, it is the true sampling distribution itself.

2.3.1 Exact approaches for categorical data

An exact approach lists and counts all possible combinations. This can only be done if we work with discrete or categorical variables. For an unlimited number of categories, we cannot list all possible combinations.

A proportion is based on frequencies and frequencies are discrete (integer values), so we can use an exact approach to create a sampling distribution for one proportion such as the proportion of yellow candies in the example above. The exact approach uses the binomial probability formula to calculate probabilities. Consult the internet if you want to know this formula; we are not going to use it here.

Exact approaches are also available for the association between two categorical (nominal or ordinal) variables in a contingency table: Do some combinations of values for the two variables occur relatively frequently? For example, are yellow candies more often sticky than red candies? If candies are either sticky or not sticky and they have one out of a limited set of colours, we have two categorical variables. We can create an exact probability distribution for the combination of colour and stickiness. The *Fisher-exact test* is an example of an exact approach to the sampling distribution of the association between two categorical variables.

2.3.2 Computer-intensive

The exact approach can be applied to discrete variables because they have a limited number of values. Discrete variables are usually measured at the nominal or ordinal level. If the number of categories becomes large, a lot of computing time can be needed to calculate the probabilities of all possible sample statistic outcomes. Exact approaches are said to be *computer-intensive*.

It is usually wise to set a limit to the time you allow your computer to work on an exact sampling distribution because otherwise the problem may keep your computer occupied for hours or days.

2.4 Exact Approaches in SPSS

2.4.1 Instructions

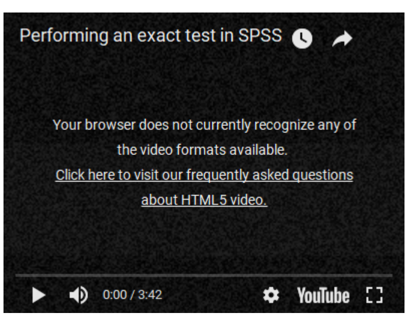


Figure 2.7: Performing an exact test in SPSS.

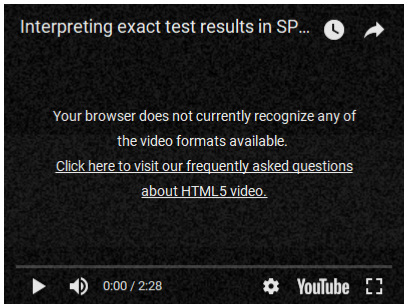


Figure 2.8: Interpreting exact test results in SPSS.

2.4.2 Exercises

1. Download the data set candies.sav and use SPSS to apply a Fisher-exact test to the association between candy colour and candy stickiness.

```
SPSS syntax:

* Exact test on the relation between candy colour
* and candy stickiness.
* Do not forget to deselect bootstrapping.
CROSSTABS
  /TABLES=colour BY sticky
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ PHI
  /CELLS=COUNT COLUMN
  /COUNT ROUND CELL
  /METHOD=EXACT TIMER(5).

Check data:

The contingency table does not show any impossible values for
the two categorical variables.

Check assumptions:

There are no assumptions for a Fisher-exact test.

Interpret the results:

There is a strong association (Phi = .52) between
candy colour and candy stickiness, which is
statistically significant, p = .010 (exact).
Yellow and red candies are less often sticky than
```

```
blue, green, and orange candies.
```

2. With the same data, apply a Fisher-exact test to the association between candy colour and candy spottiness.

```
SPSS syntax:

* Exact test on the relation between candy colour
* and candy spottiness.
* Do not forget to deselect bootstrapping.
CROSSTABS
  /TABLES=colour BY spotted
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ PHI
  /CELLS=COUNT COLUMN
  /COUNT ROUND CELL
  /METHOD=EXACT TIMER(5).

Check data:

The contingency table does not show any impossible values for
the two categorical variables.

Check assumptions:

There are no assumptions for a Fisher-exact test.

Interpret the results:

There is a moderate association (Cramer's V = .27) between
candy colour and candy stickiness, which is
not statistically significant, p = .480 (exact).
Candy colour does not seem relevant to having spots.
```

2.5 Theoretical Approximations of the Sampling Distribution

Because bootstrapping and exact approaches to the sampling distribution require quite a lot of computing power, these methods were not practical in the not so very distant pre-computer age. In those days, mathematicians and statisticians discovered that many sampling distributions look a lot like known mathematical functions. For example, the sampling distribution of the sample mean can be quite similar to the well-known bell-shape of the *normal distribution* or the closely related *(Student) t distribution*. The mathematical functions are called *theoretical probability distributions*.

Figure 2.9: Normal function as theoretical approximation of a sampling distribution.

1. In Figure 2.9, generate a sampling distribution of sample means (the computer draws many random samples from a candy population). Check if the normal function (curve) is a good approximation of the sampling distribution.

```
* The curve fits the histogram of observed sample means quite well.
Discrepancies are mainly due to the jagged layout of the histogram, which
result from binning the data (to create bars) and from the fact that the
number of samples is large but not very large.
* For the sampling distribution of means we know that the normal or (Student)
t distribution represents the sampling distribution very accurately.
```

2. While checking the distribution, pay special attention to the tails because these are used for significance tests (see Chapter 4). The red and green bars represent the 2.5 per cent samples with minimum or maximum average weight. The vertical lines mark the outer 2.5 per cent according to the normal function.

```
* The borders demarcating the lowest and highest 2.5% of sample means in the
theoretical probability distribution (the dotted lines) nicely coincide with
the border between red or green and blue bars in the histogram in most of the
sampling distributions that we generate with this app.
```

3. Generate some new sampling distributions to see if the normal function always yields a good approximation. What changes in the distribution: the mean, the standard deviation, or both?

```
* The mean of the sampling distribution does not change. It is equal to the
population mean (mean candy weight in the population), so the population mean
seems not to be changed when we generate a new sampling distribution.
* The width/peakedness of the sampling distribution changes. This suggests
that the variation (standard devation) in the population changes if we draw a
new sampling distribution. But regardless of the amount of variation, the
normal curve fits the distribution equally well.
```

The normal distribution is a mathematical function linking continuous scores, e.g., a sample statistic such as the average weight in the sample, to p values, that is, to the probability of finding at least, or at most, this score. Such a function is called a *probability density* *function*, Section 1.3.

We like to use a theoretical probability distribution as an approximation of the sampling distribution because it is convenient. The computer can calculate probabilities from the mathematical function very quickly. We also like theoretical probability distributions because they usually offer plausible argumentation about chance and probabilities.

2.5.1 Reasons for a bell-shaped probability distribution

The bell shape of the normal distribution, for instance, makes sense. Our sample of candies is just as likely to be too heavy, as it is too light, so the sampling distribution of the sample mean should be symmetrical. A normal distribution is symmetrical.

In addition, it is more likely that our sample bag has an average weight that is near the true average candy weight in the population than an average weight that is much heavier or much lighter than the true average. Bags with on average extremely heavy or extremely light candies may occur, but they are extremely rare (we are very lucky or very unlucky). From these intuitions we would expect a bell shape for the sampling distribution.

From this argumentation, we conclude that the normal distribution is a reasonable model for the probability distribution of sample means. Actually, it has been proven that the normal distribution exactly represents the sampling distribution in particular cases, for instance the sampling distribution of the mean of a large sample.

As a model, the theoretical probability distribution may actually give a better approximation of the sampling distribution than a sampling distribution created by drawing many samples from the population (as you have done in Figure 2.9), or from the initial sample as in bootstrapping. Sampling is always subject to chance, so we may have accidentally drawn samples that do not cover the sampling distribution well.

2.5.2 Conditions for the use of theoretical probability distributions

Theoretical probability distributions, then, are plausible models for sampling distributions. They are known or likely to have the same shape as the true sampling distributions under particular circumstances or conditions.

If we use a theoretical probability distribution, we must assume that the conditions for its use are met. We have to check the conditions and decide whether they are close enough to the ideal conditions. *Close enough* is of course a matter of judgement. In practice, rules of thumb have been developed to decide if the theoretical probability distribution can be used.

Figure 2.10 shows an example in which the normal distribution is a good approximation for the sampling distribution of a proportion in some situations, but not in all situations.

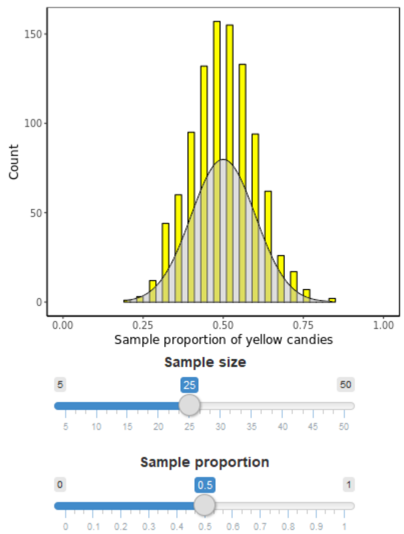


Figure 2.10: How does the shape of the distribution of sample proportions change with sample size and proportion value?

1. How do you expect that sample size affects the shape of the sampling distribution? State your expectation and then check it in the interactive content by changing sample size.

```
* The larger the sample, the more information each sample contains, the more
samples resemble the population, the closer is the sample statistic to the
population value. As a consequence, a larger sample produces a sampling
distribution with a lower spread (standard deviation, variance), so the
sampling distribution is more peaked.
```

2. How do you expect that the value of the proportion in the population affects the shape of the sampling distribution? State your expectation and then check it in the interactive content by changing the population proportion.

```
* The population proportion (parameter value) is equal to the average of the
sampling distribution because the sample proportion is an unbiased estimator
of the population proportion. So if we change the population proportion, the
center of the sampling distribution changes accordingly.
* In addition, the sampling distribution becomes less symetrical/more skewed
if the population proportion approaches zero or one. Because proportions
cannot be less than zero or more than one, the sampling distribution cannot
remain symmetrical if the population proportion is near zero or one.
```

Do theoretical probability distributions fit the true sampling distribution? As you may have noticed from the interactive content, this is not always the case. In general, theoretical probability distributions fit sampling distributions better if the sample is larger. In addition, the value of the parameter may be relevant to the fit of the theoretical probability distribution. The sampling distribution of a sample proportion is more symmetrical, like the normal distribution, if the proportion in the population is nearer .5.

This illustrates that we often have several conditions for a theoretical probability distribution to fit the sampling distribution that we should evaluate at the same time. In the example of proportions, a large sample is less important if the true proportion is closer to .5 but it is more important for true proportions that are more distant from .5.

The rule of thumb for using the normal distribution as the sampling distribution of a sample proportion combines the two aspects by multiplying them and requiring the resulting product to be larger than five. If the probability of drawing a yellow candy is .2 and our sample size is 30, the product is .2 * 30 = 6, which is larger than five. So we may use the normal distribution as approximation of the sampling distribution.

Note that this rule of thumb uses one minus the probability, if the probability is larger than .5. In other words, it uses the smaller of two probabilities: the probability that an observation has the characteristic and the probability that it has not. For example, if we want to test the probability of drawing a candy that is not yellow, the probability is .8 and we use 1 - 0.8 = 0.2, which is then multiplied by the sample size.

Apart from the normal distribution, there are several other theoretical probability distributions. We have the *binomial distribution* for a proportion, the *t distribution* for one or two sample means, regression coefficients, and correlation coefficients, the *F distribution* for comparison of variances and comparing means for three or more groups (analysis of variance, ANOVA), and the *chi-squared distribution* for frequency tables and contingency tables.

For most of these theoretical probability distributions, sample size is important. The larger the sample, the better. There are additional conditions that must be satisfied such as the distribution of the variable in the population. The rules of thumb are summarized in Table 2.2. Bootstrapping and exact tests can be used if conditions for theoretical probability distributions have not been met. Special conditions apply to regression analysis (see Chapter 8, Section 8.1.5).

2.5.3 Checking conditions

Rules of thumb about sample size are easy to check once we have collected our sample. By contrast, rules of thumb that concern the scores in the population cannot be easily checked, because we do not have information on the population. If we already know what we want to know about the population, why would we draw a sample and do the research in the first place?

We can only use the data in our sample to make an educated guess about the distribution of a variable in the population. For example, if the scores in our sample are clearly normally distributed, it is plausible that the scores in the population are normally distributed.

In this situation, we do not *know* that the population distribution is normal but we *assume* it is. If the sample distribution is clearly not normally distributed, we had better not

Table 2.2: Rules of thumb for using theoretical probability distributions.

Distribution	Sample statistic	Minimum sample size	Other requirements
Binomial distribution	proportion	-	-
(Standard) normal distribution	proportion	>= 5 divided by test proportion (<= .5)	-
(Standard) normal distribution	one or two means	> 100	OR variable is normally distributed in the population and population standard deviation is known (for each group)
t distribution	one or two means	> 30	OR variable is normally distributed in each group's population
t distribution	(Pearson) correlation coefficient	-	variables are normally distributed in the population
t distribution	(Spearman) rank correlation coefficient	> 30	-
t distribution	regression coefficient	20+ per predictor variable	See Chapter 8.
F distribution	3+ means	all groups are more or less of equal size	OR all groups have the same population variance
F distribution	two variances	-	no conditions for Levene's F test
chi-squared distribution	row or cell frequencies	expected frequency >= 1 and 80% >= 5	contingency table: 3+ rows or 3+ columns

assume that the population is normally distributed. In short, we sometimes have to make assumptions when we decide on using a theoretical probability distribution.

We could use a histogram of the scores in our sample with a normal distribution curve added to evaluate whether a normal distribution applies. Sometimes, we have statistical tests to draw inferences about the population from a sample that we can use to check the conditions. We discuss these tests in a later chapter.

2.5.4 More complicated sample statistics: differences

Up to this point, we have focused on rather simple sample statistics such as the proportion of yellow candies or the average weight of candies in a sample. Table 2.2, however, contains more complicated sample statistics.

If we compare two groups, for instance, the average weight of yellow and red candies, the sample statistic for which we want to have a sampling distribution must take into account both the average weight of yellow candies and the average weight of red candies. The sample statistic that we are interested in is the difference between the averages of the two samples.

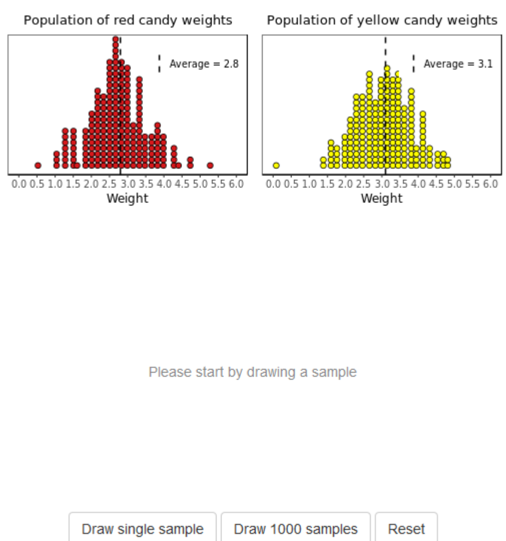


Figure 2.11: How do we obtain a sampling distribution for the mean difference of two independent samples?

1. Click on the button once. Why are these samples called independent?

```
* It is in principle possible to draw a random sample of red candies
separately from a random sample of yellow candies.
```

2. Click on the button several times. What exactly is the sample statistic in the histogram at the bottom of the app?

```
* It is the difference between average weight of red candies and average
weight of yellow candies in a sample.
* This is illustrated by the equation directly above the graph of the sampling
distribution, which subtracts the average weight of yellow candies (in yellow
typeface) from the average weight of red candies (in red typeface). The result
is added to the sampling distribution.
```

3. Click on the button to draw one thousand samples once or more often. Does the sampling distribution look familiar to you?

```
* The sampling distribution has a bell shape like the normal or (Student) t
distribution.
```

4. What, do you expect, is the mean of the sampling distribution?

```
* The true difference in averages in the population is what we expect as the
average difference in the sampling distribution.
* Average weight of red candies in the population is 2.8 grams and the average
weight in the population of yellow candies is 3.1 gram. The average weight
difference in the population is 2.8 - 3.1 = -0.3 gram. This is our
expectation.
* The centre of the sampling distribution is indeed at -0.3 if we draw
thousands of samples.
```

If we draw a sample from both the yellow and the red candies in the population, we may calculate the means for both samples and the difference between the two means. For example, the average weight of yellow candies in the sample bag is 2.76 gram and the average for red candies is 2.82 gram. For this pair of samples, the statistic of interest is 2.76 - 2.82 = -0.06, that is, the difference in average weight. If we repeat this many, many times and collect all differences between means in a distribution, we obtain the sampling distribution that we need.

The sampling distribution of the difference between two means is similar to a t-distribution, so we may use the latter to approximate the former. Of course, the conditions for using the t distribution must be met.

It is important to note that we do not create separate sampling distributions for the average weight of yellow candies and for the average weight of red candies and then look at the difference between the two sampling distributions. Instead, we create *one sampling distribution for the statistic of interest*, namely the difference between means. We cannot combine different sampling distributions into a new sampling distribution. We will see the importance of this when we discuss mediation (Chapter 9).

2.5.5 Independent samples

If we compare two means, there are two fundamentally different situations that are sometimes difficult to distinguish. When comparing the average weight of yellow candies to the average weight of red candies, we are comparing two samples that are *statistically independent* (see Figure 2.11), which means that we could have drawn the samples separately.

In principle, we could distinguish between a population of yellow candies and a population of red candies, and sample yellow candies from the first population and separately sample red candies from the other population. Whether we sampled the colours separately or not does not matter. The fact that we could have done so implies that the sample of red candies is not affected by the sample of yellow candies or the other way around. The samples are statistically independent.

This is important for the way in which probabilities are calculated. Just think of the simple example of flipping two coins. The probability of having heads twice in a row is .5 times .5 that is .25 if the coins are unbiased and the result of the second coin does not depend on the result of the first coin. The second flip is not affected by the first flip.

Imagine that a magnetic field is activated if the first coin lands with heads up and that this magnetic field increases the odds that the second coin will also be heads. Now, the second toss is not independent of the first toss and the probability of getting heads twice is larger than .25.

2.5.6 Dependent samples

The example of a manipulated second toss is applicable to repeated measurements. If we want to know how quickly the yellow colour fades when yellow candies are exposed to sun light, we may draw a sample of yellow candies once and measure the colourfulness of each candy at least twice: at the start and after some time interval. Subsequently, we compare the average colourfulness of the second set of measurements to the average in the first set of measurements.

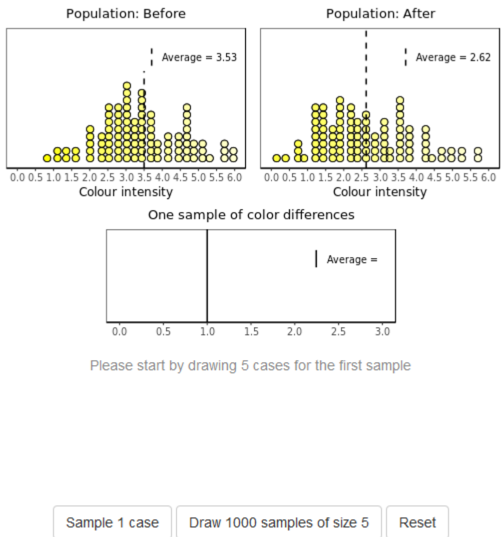


Figure 2.12: Dependent samples.

1. In Figure 2.12, use the **Sample 1 case** button repeatedly to draw a sample of five observations. What is the precise meaning of the numbers on the horizontal axis in the sample histogram?

```
* The numbers on the horizontal axis in the sample histogram represent the
difference in colour intensity for each pair of cases that is drawn.
* In each draw, one case in the before population (red) and the same case in
the after population (orange) is selected. The difference in colour intensity
between the before and after measurement is calculated in the equation below
the population dot plots. The calculated difference for this pair is
represented by a dot in the figure in the middle.
```

2. Why is the sample called dependent or paired?

```
* A case appears both in the before and after population: We have a before and
after measurement of colour intensity for each case (candy). These two
measurements are related or paired because they refer to the same candy.
* As a consequence, if we draw candies for our before measurement, we also draw
the candies for our after measurement. The after sample depends on the before
sample.
```

3. Draw 1,000 samples to obtain a sampling distribution. What is the precise meaning of the numbers on the horizontal axis in the histogram of the sampling distribution?

```
* The numbers on the horizontal axis in the histogram of the sampling
distribution signify the average difference in colour intensity of the candies
in a sample (of five candies).
```

In this example, we are comparing two means, just like the yellow versus red candy weight example, but now the samples for both measurements are the same. It is impossible to draw the sample for the second measurement independently from the sample for the first measurement if we want to compare repeated measurements. Here, the second sample is fixed once we have drawn the first sample. The samples are *statistically dependent*; they create *paired samples*. Even if the second sample depends only partly upon the first sample, the samples are statistically dependent.

With dependent samples, probabilities have to be calculated in a different way, so we need a special sampling distribution. In the interactive content above, you may have noticed a relatively simple solution for two repeated measurements. We just calculate the difference between the two measurements for each candy in the sample and use the mean of this new difference variable as the sample statistic that we are interested in. The t-distribution, again, offers a good approximation of the sampling distribution of dependent samples if the samples are not too small.

For other applications, the actual sampling distributions can become quite complicated but we need not worry about that. If we choose the right technique, our statistical software will take care of this. Of course, we should check whether the conditions are met for approximating the sampling distribution with a theoretical probability distribution.

2.6 SPSS and Theoretical Approximation of the Sampling Distribution

By default, SPSS uses a theoretical probability distribution to approximate the sampling distribution. It chooses the correct theoretical distribution but you yourself should check if the conditions for using this distribution are met. Is the sample size large enough or is it plausible that the variable is normally distributed in the population?

In one case, SPSS automatically selects an exact approach if the conditions for a theoretical approximation are not met: If you do a chi-squared test to a contingency table in SPSS, SPSS will automatically apply Fisher's exact test if the table has two rows and two columns. In all other cases, you have to select a bootstrapping or exact approach yourself if the conditions for a theoretical approximation are not met.

We are not going to practice with theoretical approximations in SPSS, now. Because theoretical approximation is the default approach in SPSS, we will encounter it in the exercises in later chapters.

2.7 Test Your Understanding

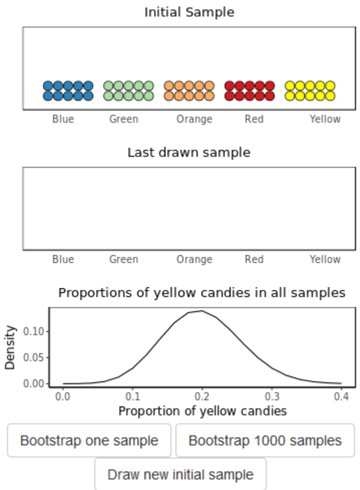


Figure 2.13: How do we bootstrap a sampling distribution?

1. Why does Figure 2.13 not show a population?
2. Which type of sampling is better here: with or without replacement? Justify your answer.
3. Draw a new initial sample in Figure 2.13. Is the bootstrapped sampling distribution going to resemble the true sampling distribution? Note that twenty per cent of the candies in the population are yellow. Motivate your answer. Draw 1,000 bootstrap samples to check your answer.

Table 2.3: Number of heads for a toss of three coins.

Number of heads	Combination
0	tail-tail-tail
1	tail-tail-head
1	tail-head-tail
1	head-tail-tail
2	head-head-tail
2	head-tail-head
2	tail-head-head
3	head-head-head
Total	8

4. Calculate the exact probability distribution of the number of heads in a toss of three fair coins (Table 2.3).

5. In which situations can we use exact probabilities as a sampling distribution?

Figure 2.14: How do we approximate a sampling distribution with a theoretical probability distribution?

6. Generate a sampling distribution of average sample candy weight in Figure 2.14. Try to explain in your own words why the sampling distribution of a sample mean has a bell shape.
7. Which part of the graph in Figure 2.14 represents the theoretical probability distribution?
8. Can we always use this theoretical probability distribution if we are interested in sample means? See if the general outline of the theoretical probability distribution matches the histogram of average sample candy weight observed in a large number of samples. Pay special attention to the lowest (red) and highest (green) 2.5% of observed sample means.

2.8 Take-Home Points

- We may create an exact sampling distribution or simulate a bootstrap sampling distribution in simple situations or if we have a lot of computing power.
- For a bootstrap sampling distribution, we need about 5,000 bootstrap samples from our original sample.
- We can often approximate the sampling distribution of a sample statistic with a known theoretical probability distribution.
- Approximations only work well under conditions, which we have to check.
- Conditions usually involve the size of the sample, sample type (independent vs. dependent/paired), and the shape or variance of the population distribution.
- Samples are independent if, in principle, we can draw a sample for one group without taking into account the sample for another group of cases. Otherwise, the samples are dependent or paired.

Chapter 3

Estimating a Parameter: Which Population Values Are Plausible?

Key concepts: point estimate, interval estimate, confidence (level), precision, standard error, critical value, degrees of freedom, confidence interval, uncertainty.

Summary

Given our sample, what are plausible population values?

In this chapter, we set out to make educated guesses of a population value (parameter, often called "the true value") based on our sample. This type of guessing is called *estimation*. Our first guess will be a single value for the population value. We merely guess that the population value is equal to the value of the sample statistic. This guess is the most precise guess that we can make, but it is most likely to be wrong.

Our second guess uses the sampling distribution to make a statement about the approximate population value. More precisely, we calculate an interval for which we are confident that it includes the population value. The wider the interval, the more confident we are that it contains the true population value but, at the same time, the less precise our guess.

3.1 Point Estimate

If we have to name one value for the population value, our best guess is the value of the sample statistic. For example, if 18% of the candies in our sample bag are yellow, our best guess of the proportion of yellow candies in the population of all candies from which this bag was filled, is .18. What other number can we give if we only have our sample? This type of guess is called a *point estimate* and we use it a lot.

The sample statistic is the best estimate of the population value only if the sample statistic is an unbiased estimator of the population value. As we have learned in Section 1.2.5, the true population value is equal to the mean of the sampling distribution for an unbiased estimator. The mean of the sampling distribution is the expected value for the sample.

In other words, an unbiased estimator neither systematically overestimates the population value, nor does it systematically underestimate the population value. With an unbiased estimator, then, there is no reason to prefer a value higher or lower than the sample value as our estimate of the population value.

Even though the value of the statistic in the sample is our best guess, it is very unlikely that our sample statistic is exactly equal to the population value (parameter). The recurrent theme in our discussion of random samples is that a random sample differs from the population because of chance during the sampling process. The precise population value is highly unlikely to actually appear in our sample.

The sample statistic value is our best point estimate but it is nearly certain to be wrong. It may be slightly or strongly off the mark but it will hardly ever be spot on. For this reason, it is better to estimate a range within which the population value falls. Let us turn to this in the next section.

3.2 Interval Estimate for the Sample Statistic

The sampling distribution of a continuous sample statistic tells us the probability of finding a range of scores for the sample statistic in a random sample. For example, the average weight of candies in a sample bag is a continuous random variable. The sampling distribution tells us the probability of drawing a sample with average candy weight between 2.0 and 3.6 gram. We can use this range as our *interval estimate*.

Note that we are reasoning from sampling distribution to sample now. This is not what we want to do in actual research, where we want to reason from sample to sampling distribution to population. We get to that in Section 3.5. For now, assume that we know the true sampling distribution.

Remember that the average or expected value of a sampling distribution is equal to the population value if the estimator is unbiased. For example, the mean weight of yellow candies averaged over a large number of samples is equal to the mean weight of yellow candies in the population. For an interval estimate, we now select the sample statistic values that are closest to the average of the sampling distribution.

Between which boundaries are the sample statistic values situated that are closest to the population value? Of course, we have to specify what we mean by "closest". Which part of all samples do we want to include? A popular proportion is 95%, so we want to know the boundary values that include 95% of all samples that are closest to the population value.

For example, between which boundaries is the average candy weight situated for 95% of all samples that are closest to the average candy weight in the population?

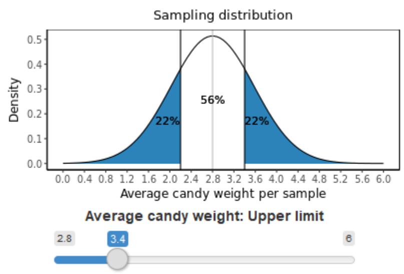


Figure 3.1: Within which interval do we find the sample results that are closest to the population value?

Figure 3.1 shows the sampling distribution of average sample candy weight.

1. What is the average candy weight in the population of candies?

```
* A sample mean is an unbiased estimator of the population mean. As a
consequence, the mean of the sampling distribution is equal to the population.
* The (normal) sampling distribution is symmetrical, so the mean of the
sampling distribution is in its middle, exactly under the top of the sampling
distribution. Here, the value is 2.8 (gram). This is average candy weight in
the population.
```

2. Move the slider until you have found the interval containing 95% of all samples that are closest to the (true) population value. What are the upper and lower limits of the interval that contains these samples?

```
* The exact value of the upper (right) limit changes from application to
application. It usually is somewehere between 3 and 5.
* To obtain the exact lower (left) limit, calculate the difference between the
upper limit and the mean of the distribution (2.8). Then subtract this
difference from the distribution mean (2.8) to get the lower left) limit of
the interval.
* Tip: If the cursor is on the slider handle, you can change the slider value
in minimal steps with the left and right arrow keys on your keyboard.
```

Say, for instance, that 95% of all possible samples in the middle of the sampling distribution have an average candy weight ranging from 1.6 to 4.0 gram. The proportion .95 can be interpreted as a probability. Our sampling distribution tells us that we have 95% probability that the average weight of yellow candies lies between 1.6 and 4.0 gram in a random sample that we draw from this population.

We now have boundary values, that is, a range of sample statistic values, and a probability of drawing a sample with a statistic falling within this range. The probability shows our *confidence* in the estimate. It is called the *confidence level* of an interval estimate.

3.3 Precision, Standard Error, and Sample Size

The width of the estimated interval represents the *precision* of our estimate. The wider the interval, the less precise our estimate. With a less precise interval estimate, we we will have to reckon with a wider variety of outcomes in our sample.

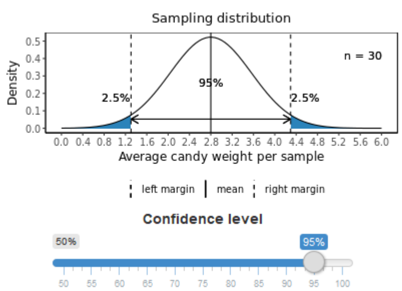


Figure 3.2: How does the confidence level affect the precision of an interval estimate?

1. How, do you think, does the precision (width) of the interval estimate (represented by the double-sided arrow) change if you change the confidence level? First write down what you expect, and why you expect that. Then check what happens if you change the confidence level slider in Figure 3.2.

```
* The more population values we exclude, the larger the probability that the
true population value is NOT in our estimated interval. In other words, our
confidence is smaller if we increase the precision of our estimate or, in other
words, if we reduce the width of the estimated interval.
```

2. What happens if we want to be 100% certain that our interval contains the true population value?

```
* If we want to be 100% sure, we cannot rule out any possible average weight
as the true population value. As a consequence, the interval becomes
infinitely wide. Of course, such an interval is completely useless because it
tells us that any value is possible.
```

If we want to predict something, however, we value precision. We rather conclude that the average weight of candies in the next sample we draw lies between 2.0 and 3.6 gram than between 1.6 and 4.0 gram. If we would be satisfied with a very imprecise estimate, we need not do any research at all. With relatively little knowledge about the candies that we are investigating, we could straightaway predict that the average candy weight is between zero and ten gram. The goal of our research is to find a more precise estimation.

There are several ways to increase the precision of our interval estimate, that is, to obtain a narrower interval for our estimate. The easiest and least useful way is to decrease the probability that our estimate is correct. If we lower the probability that we are right, we can discard a large number of other possible sample statistic outcomes and focus on a narrower range of sample outcomes around the true population value.

This method is not useful because we sacrifice our confidence that the range includes the outcome in the sample that we are going to draw. What is the use of a more precise estimate if we are less certain that it predicts correctly? Therefore, we usually do not change the confidence level and leave it at 95% or thereabouts (90%, 99%). We think it important to be quite sure that our prediction will be right.

3.3.1 Sample size

A less practical but very useful method of narrowing the interval estimate is increasing sample size. If we buy a larger bag containing more candies, we get a better idea of average candy weight in the population.

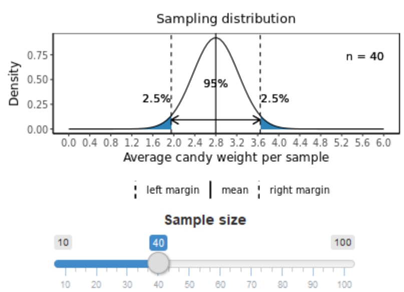


Figure 3.3: How does sample size affect the precision of an interval estimate?

Figure 3.3 shows a sampling distribution of average candy weight in candy sample bags. The horizontal arrow indicates the precision of the interval estimate.

1. How does the precision of the interval estimate change if you change the size of the sample? First write down what you expect, and why you expect that. Then, check what happens if you change the sample size slider.

```
* Larger samples offer more information, so they allow a more precise estimate
of a population value, such as average candy weight. The precision of the
interval will increase, so the interval will become narrower, the arrow will
become shorter.
* The computational reason: A larger sample yields a smaller standard error
(se). See the next section.
```

2. How does the shape of the sampling distribution change if you change sample size? Explain what this means for the values of the sample statistic.

```
* The sampling distribution because more peaked/less flat because more sample
means are found near the (true) population mean.
```

As you may have noticed while playing with Figure 3.3, a larger sample yields a narrower, that is, more precise interval. You may have expected intuitively that larger samples give more precise estimates because they offer more information. This intuition is correct.

In a larger sample, an observation above the mean is more likely to be compensated by an observation below the mean. Just because there are more observations, it is less likely that we sample relatively high scores but no or considerably fewer scores that are relatively low.

In other words, the larger the sample, the more the distribution of scores for a variable in the sample will resemble the distribution of scores for this variable in the population. As a consequence, a sample statistic value will be closer to the population value for this statistic.

Larger samples resemble the population more closely, and therefore large samples drawn from the same population are also closer to one another. The result is that the sample statistic values in the sampling distribution are less varied and more similar. They are more concentrated around the true population value, which is the average of the sampling distribution. The sampling distribution is more peaked, so the middle 95% of all sample statistic values are closer to the centre.

3.3.2 Standard error

The concentration of sample statistic values, such as average candy weight in a sample bag, is expressed by the standard deviation of the sampling distribution. Hitherto, we have only paid attention to the centre of the sampling distribution, its mean, because it is the expected value in a sample and it is equal to the population value if the estimator is unbiased.

Now, we start looking at the standard deviation of the sampling distribution as well, because it tells us how precise our interval estimate is going to be. The sampling distribution's standard deviation is so important that it has received a special name: the *standard error*.

Figure 3.4: The standard error: How wrong are point estimates?

1. Draw a single sample by pressing the button *Take single sample* in Figure 3.4. Explain why we can interpret the red arrow in the sample plot as error.

```
* We use the sample mean (represented by the dotted line in the top graph) as
point estimate for the population mean (represented by the solid line). The
distance between the two shows how wrong our point estimate is. For this
reason, the red arrow in the top chart can be regarded as the size of the error.
```

2. Use the button **Take single sample** several times. What is the meaning of the double-sided red arrow that appears in the sampling distribution?

```
* The double-sided red arrow expresses the standard deviation of the sample
means that we have added to the sampling distribution (according to the
legend). So it is the standard deviation in the empirical sampling
distribution that we create by repeated sampling.
* It gives an impression of how different the means of different random
samples can be. Because those sample means are used as point estimates, the
red arrow expresses the (average) size of the errors that we make if we
generalize the sample mean to the population mean.
```

3. Use the button **Take 100 samples** once or repeatedly. What happens to the length of the double-sided red arrow in the sampling distribution?

```
* The red arrow fits the distance between the dotted lines in the bottom plot.
The dotted lines mark one standard error above or below the mean of the true
sampling distribution (light green). The arrowheads of the red arrows mark one
standard deviation above and below the average of the sample means that we
have drawn.
* The standard error of a sampling distribution, then, is equal to the
standard deviation of the sampling distribution.
```

The word *error* reminds us that the standard error tells us the size of the error that we are likely to make (on average under many repetitions) if we use the value of the sample statistic as a point estimate for the population value.

Let us assume, for instance, that the standard error of the average weight of candies in a sample bag is 0.6. Loosely stated, this means that the average difference between true average candy weight and average candy weight in a sample is 0.6 if we draw a very large number of samples from the same population.

By the way, the standard deviation does not give us the ordinary average difference but it gives us the square root of the average of squared differences. But this detail is irrelevant to how we interpret the standard error.

The smaller the standard error, the more the sample statistic values resemble the true population value, and the more precise our interval estimate with a given confidence level, for instance, 95%. Because we like more precise interval estimates, we prefer small standard errors over high standard errors.

In theory, it is easy to obtain smaller standard errors: just increase sample size. See Figure 3.3: larger samples yield more peaked sampling distributions. In a peaked distribution, values are closer to the mean. In our example, average candy weight in sample bags are closer to the average candy weight in the population. Imagine a horizontal arrow showing the width of the distribution: It becomes smaller for a more peaked distribution, so the standard error is lower.

In practice, however, it is both time-consuming and expensive to draw a very large sample. Usually, we want to settle on the optimal size of the sample, namely a sample that is large enough to have interval estimates at the confidence level and precision that we need but as small as possible to save on time and expenses. We return to this matter in Chapter 5.

3.4 Critical Values

In the preceding section, we learned that the standard error is related to the precision of the interval estimate. A larger standard error yields a less precise estimate and a wider confidence interval.

We are interested in the interval that includes a particular percentage of all samples that can be drawn, usually the 95 per cent of all samples that are closest to the population value. In our current example, the 95 per cent of all samples with average candy weight that is closest to average candy weight in the population (2.8 gram).

In theoretical probability distributions like the normal distribution, the percentage of samples is related to the standard error. If we know the standard error, we know the interval within which we find the 95 per cent of samples that are closest to the population value.

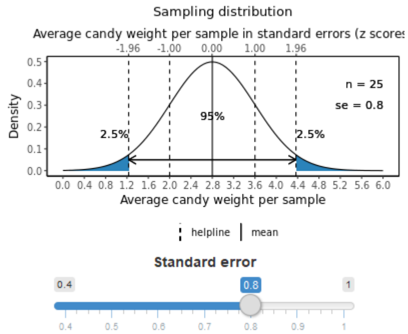


Figure 3.5: How do critical values relate to the standard error in a normal distribution?

Figure 3.5 shows the sampling distribution of average candy weight per sample bag. It contains two horizontal axes, one with average candy weight in grams (bottom) and one with average candy weight in standard errors, also called z scores (top).

1. How do the two horizontal axes tell you the size of the standard error in grams?

```
* The axis at the bottom gives average candy weight in grams and the axis at
the top shows average candy weight in standard errors from the population
mean.
* One standard error equals the distance or difference between z score zero
(0) and z score one (1). If we follow the lines down from these standard error
scores, we obtain the mean average candy weight (2.8 gram) and the average
weight one standard error above the mean (3.6 gram).
* The difference is 0.8 gram, so one standard error represents 0.8 grams (in
average candy weight per sample bag).
```

2. How do you expect the location of the vertical lines on the two horizontal axes to change if you change the size of the standard error? Check your expectation by using the slider.

```
* If we increase the standard error, one standard error coincides with more
grams, so the sampling distribution grows wider; it becomes less peaked. After
all, the variation in sample outcomes (average candy weight) will increase.
* The dotted lines representing plus or minus 1.00 and 1.96 will move away
from the centre but the centre will be fixed at the true average candy weight
in the population, in this example 2.8 gram.
```

In Figure 3.5, we approximate the sampling distribution with a theoretical probability distribution, namely the normal distribution. The theoretical probability distribution links probabilities (areas under the curve) to sample statistic outcome values (scores on the horizontal axis). For example, we have 2.5% probability to draw a sample bag with average candy weight below 1.2 gram or 2.5% probability to draw a sample bag with average candy weight over 4.4 gram.

3.4.1 Standardization and *z* scores

The average candy weights that are associated with 2.5% and 97.5% probabilities in Figure 3.5 depend on the sample that we have drawn. As you will have noticed while playing with Figure 3.3, changing the size of the sample also changes the average candy weights that mark the 2.5% and 97.5% probabilities.

We can simplify the situation if we *standardize* the sampling distribution: Subtract the average of the sampling distribution from each sample mean in this distribution, and divide the result by the standard error. Thus, we transform the sampling distribution into a distribution of standardized scores. The mean of the new standardized variable is always zero.

If we use the normal distribution for standardized scores, which is called the *standard-normal distribution* or *z distribution*, there is a single z value that marks the boundary between the top 2.5% and the bottom 97.5% of any sample. This z value is 1.96. If we combine this value with -1.96, separating the bottom 2.5% of all samples from the rest, we obtain an interval [-1.96, 1.96] containing 95% of all samples that are closest to the mean of the sampling distribution. This is part of the *empirical rule* for the normal distribution.

In a standard-normal or z distribution, 1.96 is called a *critical value*. Together wih its negative (-1.96), it separates the 95% sample statistic outcomes that are closest to the parameter, hence that are most likely to appear, from the 5% that are furthest away and least likely to appear. There are also critical z values for other probabilities, for instance, 1.64 for the middle 90% of all samples and 2.58 for the middle 99%.

3.4.2 Interval estimates from critical values and standard errors

Critical values in a theoretical probability distribution tell us the boundaries, or range, of the interval estimate expressed in standard errors. In a normal distribution, 95% of all sample means are situated no more than 1.96 standard errors from the population mean.

If the standard error is 0.5 and the population mean is 2.8 gram, we have 95% probability that the mean candy weight in a sample that we draw from this population lies between 1.82 gram (this is 1.96 times 0.5 subtracted from 2.8) and 3.78 gram.

Critical values make it easy to calculate an interval estimate if we know the standard error. Just take the population value and add the critical value times the standard error to obtain the upper limit of the interval estimate. Subtract the critical value times the standard error from the population value to obtain the lower limit.

Normal distributions make life easier for us, because there is a fixed critical value for each probability, such as 1.96 for 95% probability, which is well-worth memorizing.

3.5 Confidence Interval for a Parameter

Working through the preceding sections, it may have occurred to you that it is all very well to be able to estimate the value of a statistic in a new sample with a particular precision and probability, but that this is not what we are primarily interested in. Instead, we want to estimate the value of the statistic in the population.

For example, we don't care much about the average weight of candies in our sample bag or in the next sample bag that we may buy. We want to say something about the average weight of candies in the population. How can we do this?

In addition, you may have realized that, if we know the sampling distribution, we also know the precise population value, for instance, average candy weight. After all, the average of the sampling distribution is equal to the population mean for an unbiased estimator. In the preceding paragraphs, we acted as if we knew the sampling distribution. If we know the sampling distribution, and it then follows that we also know the population value, why would we even care about estimating an interval?

Our problem is this: We want to estimate a population value using probabilities. For probabilities we need the sampling distribution but for the sampling distribution, we must know the population. A vicious circle.

In the exact approach to the sampling distribution of the proportion of yellow candies in a sample bag (Figure 3.6), for instance, we must know the proportion of yellow candies in the population. If we know the population proportion, we can exactly calculate the probability of getting a sample bag with a particular proportion of yellow candies. But we don't know the population proportion of yellow candies; we want to estimate it.

A theoretical probability distribution can only be used as an approximation of a sampling distribution if we know some characteristics of the population. We know that the sampling distribution of sample means always has the bell shape of a normal (z) distribution or t distribution. However, knowing the shape is not sufficient for using the theoretical distribution as an approximation of the sampling distribution.

We must also know the population mean because it specifies where the centre of the sampling distribution is located. So, we must know the population mean to use a theoretical probability distribution to estimate the population mean. This sounds like a problem that only Baron von Münchhausen can solve. How can we drag ourselves by the hair out of this swamp?

By the way, we also need the standard error to know how peaked or flat the bell shape is. The standard error can usually be estimated from the data in our sample. But let us not worry about how the standard error is being estimated and focus on estimating the population mean now.

3.5.1 Imaginary population values

How can we find plausible population means? Figure 3.7 shows average candy weight in a random sample (lower scale). Click somewhere under the top axis to select a possible value

Yellow candies in a sample	Probability
0	0.107
1	0.268
2	0.302
3	0.201
4	0.088
5	0.026
6	0.006
7	0.001
8	0
9	0
10	0

Figure 3.6: Probabilities of a sample with a particular number of yellow candies if 20 per cent of the candies are yellow in the population.

for the population mean. The app will then display the interval of most plausible sample means (green if it contains the actual sample means, red otherwise) and the actual sample mean's z value if this would have been the true population value.

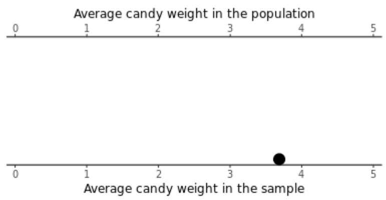


Figure 3.7: For which population means is our sample mean plausible?

1. Click repeatedly on Figure 3.7 to find the highest and lowest value of the population mean for which the sample mean is in the interval of sample means that have 95% probability to occur.

```
* The closer the sample mean is to the population mean, the higher the
probability to draw a sample with more or less this mean. Population means
close to the sample mean will turn up as green dots and their intervals will
include the sample mean.
* The more we move away from the sample mean, the less likely to draw a sample
with a mean that differs at least this much with the population mean. At some
point, the population mean is too far away from the sample mean, so the actual
sample mean is no longer in the range of the 95% most likely sample means to
be drawn from a population with the selected population mean.
```

2. How does the z value of the sample mean help you to minimize the number of clicks you need?

```
* In a normal distribution, z standardizes scores and 95% of all observations
differ no more than 1.96 z scores from the mean. The lowest and highest
population means for which the current sample mean is among the 95% most
plausible samples, then, are reached if the z value of the sample mean is
(almost) 1.96 or -1.96.
* So the z value tells us how close we are to the limit.
```

3. How does the depicted interval estimate help you to minimize the number of clicks you need?

```
* The interval estimate shows the width of the interval between the lowest and
highest population average for which the current sample mean is among the 95%
most plausible samples.
* If the current sample mean is at the left limit of the interval, we reach
the highest population mean for which the sample mean is in the interval.
```

```
Increase the population mean and the interval will no longer include the
current sample mean. In a similar way, the lowest population mean for which
the interval includes the sample mean is found if the sample mean is on the
right limit of the interval.
```

4. What is the most efficient strategy (minimum number of clicks) to determine the lower and upper limits of the population means for which the sample mean is among the 95% most likely samples? Explain why this is the most efficient strategy.

```
* Using the answer to Question 3, we may use the following steps to find the
lower and upper limits.
a. Click in the middle of the sample mean. This selects the population mean
that is equal to the sample mean.
b. Click on the left edge of the interval. This should be (very near to) the
lowest population mean, the interval of which still includes the sample mean.
c. Repeat Step a to display again the interval for the population mean that is
equal to the sample mean.
d. Click the right end of this interval to find the highest population mean
such that the current sample mean is still among the 95% most plausible
samples.
* You may not hit the nail spot-on with every click but the procedure itself
is efficient. You can only do better by gambling the values of the two limits
directly and being very very lucky.
* The important lesson here: Instead of constructing the interval around the
population mean, we can construct it around the sample mean to obtain the
range of population means that are plausible given this sample. We flip the
procedure!
```

How do we solve the Münchhausen problem that we must know the population mean to estimate the population mean? A solution is that we select a lot of imaginary population means. For each imaginary population mean, we calculate the interval within which the sample mean is expected to fall if this imaginary mean would be the true population mean. We use a fixed confidence level, usually a probability of 95 per cent.

As a next step, we check if the mean of the sample that we have actually drawn falls within this interval. If it does, we conclude that this (imaginary) population mean is not at odds with the sample that we have drawn. In contrast, if our sample mean falls outside the interval, we conclude that this population mean is not plausible because our sample is too unlikely to be drawn from a population with this mean.

In this way, we can find all population means that are *consistent* with our sample. If the true population mean is any of these imaginary means, we are sufficiently likely (95% probability) to draw a sample with our actual sample mean.

While playing with Figure 3.7, you may have noticed the z values of the sample mean for the lowest and highest population means for which the sample mean is still within the interval. When you hit the lower bound of the population means, the sample mean has a z value of about 1.96 while it has a z value of about -1.96 for the highest population mean in the range.

It is not a coincidence that we find the critical values of the standard-normal distribution when we reach the minimum and maximum population means that are plausible. We are using the standard-normal distribution to approximate the sampling distribution of the sample mean. The critical z value 1.96 marks the upper limit of the interval containing 95% of all samples with means closest to the population mean and -1.96 marks the lower limit. A distance of 1.96 standard errors, then, is the maximum distance between a population mean and a sample mean that belongs to the 95% sample means closest to the population mean.

As a consequence, we may simply calculated the range of plausible population values by adding and subtracting 1.96 standard errors from the sample mean. This is much more efficient than selecting a lot of imaginary population means!

This can be illustrated with an example: If the average candy weight in our sample is 2.8 gram and the standard error is 0.5, the lower and upper boundary for plausible population means are 1.82 gram (this is 2.8 minus 1.96 times 0.5) and 3.78 gram (2.8 plus 1.96 times 0.5).

Haven't we seen this calculation before? Yes we did, in Section 3.4.2, where we estimated the interval for sample means. We now simply reverse the calculation, using the sample mean to estimate an interval of plausible population means instead of the other way around.

Jerzy Neyman introduced the concept of a confidence interval in 1937:

"In what follows, we shall consider in full detail the problem of estimation by interval. We shall show that it can be solved entirely on the ground of the theory of probability as adopted in this paper, without appealing to any new principles or measures of uncertainty in our judgements". (Jerzy Neyman, 1937: 347)

Figure 3.8: Jerzy Neyman

3.5.2 Confidence interval

The upper and lower bounds for the population mean or, more generally, the parameter that we want to estimate, yield an interval for the parameter. We use this as the interval estimate of the parameter.

This interval is linked to a probability, for instance, 95%. However, it is very important that we understand that this is NOT the probability that the parameter has a particular value, or that it falls within the interval. The parameter is *not* a random variable because it is not affected by the random sample that we draw. In our example, it will be clear that the sample that we draw does not and cannot change the average weight of all candies.

The parameter has one value, which is either within or outside the interval that we have constructed. We just don't know. But we do know that our sample is more likely for population values within the interval.

We use the term *confidence* instead of probability when we use this interval to estimate a parameter. We say that we are 95% confident that the parameter falls within the interval. The interval is called a *confidence interval* and we usually add the confidence level, for instance, the 95% confidence interval (abbreviated: 95%CI) of the average weight of candies in the population ranges from 2.4 to 3.2 gram. An average candy weight between 2.4 and 3.2 gram is plausible given the sample that we have drawn. In our reports, we say that:

> We are 95% confident that the average candy weight in the population is between 2.4 and 3.2 gram.

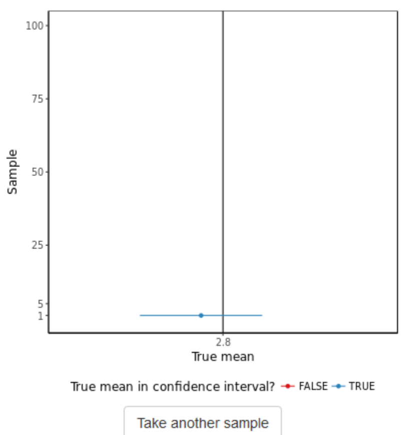


Figure 3.9: How often does a confidence interval include the true population value?

Figure 3.9 shows a 95% confidence interval for average candy weight in the population based on a sample of candies. The vertical line indicates the true average candy weight in the population.

1. Does the (first) confidence interval include the true average candy weight in the population?

```
* The first confidence interval may (blue) or may not (red) include the true
population mean (2.8).
* The sample mean can be close enough to the true population mean for the
confidence interval to include the true population mean. But it can also be too
far from the true population mean.
* A confidence interval, however, is much more likely to include the true
```

```
population mean than not.
```

2. What does the dot in the middle of the confidence interval represent?

```
* The centre of a confidence interval represents the point estimate, that is,
if the standard error is used to calculate the confidence interval. In this
case, the upper and lower boundaries of a confidence interval are situated one
standard error times the critical value below or above the point estimate.
* In the current example, the population mean is estimated, so the point
estimate is the sample mean. A sample mean is an unbiased estimator of the
population mean.
```

3. If you would draw a hundred random samples from the same population and calculate the 95% confidence interval for the mean of each sample, how many confidence intervals do you expect to contain the true population mean? Press the **New Sample** button until there are one hundred confidence intervals. Does your expectation come true? If not, why not?

```
* The confidence level is the probability that our confidence interval includes
the true population value. If the confidence level is 95%, the probability that
a confidence interval includes the true population value is .95, so we expect
that 95 out of 100 confidence intervals include the population mean.
* With probabilities, the expected value is only certain to be found in an
infinitively large number of draws (samples). One hundred samples are quite few
in the light of infinity, so the number of samples including the true
population value may well be smaller or larger than 95.
* This is easier to see if we focus on the number of red confidence intervals
that do not include the true population value. This number is usually
(slightly) larger or smaller than five.
```

The more precise meaning of a confidence interval is rather complicated. If we draw a very large number of samples from the same population, 95% of the sample means would differ less from the population mean than the critical value times the standard error. This is just the definition of critical value.

Because the critical value times the standard error also defines the width of the confidence interval, we can reverse the statement. The population mean would be within the 95% confidence interval of 95% of all sample means. In other words, if we construct the 95% confidence interval for each sample mean, 95% of all samples will have confidence intervals that contain the true population mean.

Our confidence interval is a random variable because it depends on the sample that we draw. After all, we construct the interval around the sample statistic outcome, for instance, average candy weight in our sample (the point estimate), which may change from sample to sample. So we may say that the confidence level is the *probability that our confidence interval includes the true population value*. But we avoid this interpretation because it is easily misread as the probability that the true population value is in the interval. The latter reading is wrong from the perspective that the population value is not a random variable, so it does not have a probability.

Unfortunately, we have no clue whether or not our single sample belongs to the 95 per cent of 'lucky' samples with confidence intervals containing the true population value. We can only hope and be confident that this is the case.

3.5.3 Confidence intervals with bootstrapping

If we approximate the sampling distribution with a theoretical probability distribution such as the normal (z) or t distribution, critical values and the standard error are used to calculate the confidence interval (see Section 3.5.1).

There are theoretical probability distributions that do not work with a standard error, such as the F distribution or chi-squared distribution. If we use those distributions to approximate the sampling distribution of a continuous sample statistic, for instance, the quotient of two variances, we must use bootstrapping to obtain a confidence interval.

As you probably remember from Section 2.1, we simulate an entire sampling distribution if we bootstrap a statistic, for instance the median candy weight in a sample bag. This simulated sampling distribution can be used to estimate the standard error, which is by definition the standard deviation of the sampling distribution. This standard error can then be combined with critical values to calculate the confidence interval.

As an alternative, we can just take the values separating the bottom 2.5% and the top 2.5% of all samples in the bootstrapped sampling distribution as the lower and upper limits of the 95% confidence interval. This is known as the percentile approach.

It is also possible to construct the entire sampling distribution in exact approaches to the sampling distribution. Both the standard error and percentiles can be used to create confidence intervals. This can be very demanding in terms of computer time, so exact approaches to the sampling distribution usually only report p values, not confidence intervals.

3.6 Confidence Intervals in SPSS

3.6.1 Instruction

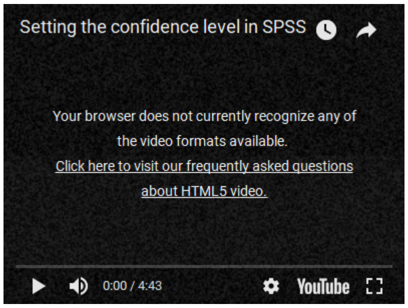


Figure 3.10: Setting the confidence level in SPSS.

3.6.2 Exercises

1. Download the data set candies.sav and use SPSS to calculate the 95% and 99% confidence intervals of average candy weight.

   Hint: Use the *Analyze > Compare Means > One-Sample T Test* command and leave the test value at zero.

   Interpret the results and explain why the 99% confidence interval is wider than the 95% confidence interval.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=weight
  /FORMAT=NOTABLE
  /HISTOGRAM NORMAL
  /ORDER=ANALYSIS.
* 95% CI.
T-TEST
  /TESTVAL=0
  /MISSING=ANALYSIS
  /VARIABLES=weight
  /CRITERIA=CI(.95).
* 99% CI.
T-TEST
  /TESTVAL=0
  /MISSING=ANALYSIS
  /VARIABLES=weight
```

```
  /CRITERIA=CI(.99).

Check data:

There are no impossible values on variable weight.

Check assumptions:

Sample size (N = 50) is well over 30, so we need not care about
the normal distribution of candy weight in the population.

Interpret the results:

Average weight of candies in the population is between 2.77 and
2.90 gram with 95% confidence and between 2.75 and 2.92 gram
with 99% confidence.
If we want to be more confident, we must allow for a broader
range of values, so the confidence interval is wider.
```

2. Let SPSS calculate the 95% confidence interval for median candy weight. Interpret the result. The data are in “candies.sav”.

   Remember the SPSS exercises in Section 2.2.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=weight
  /ORDER=ANALYSIS.
* Bootstrap on median candy weight.
BOOTSTRAP
  /SAMPLING METHOD=SIMPLE
  /VARIABLES INPUT=weight
  /CRITERIA CILEVEL=95 CITYPE=BCA  NSAMPLES=5000
  /MISSING USERMISSING=EXCLUDE.
FREQUENCIES VARIABLES=weight
  /FORMAT=NOTABLE
  /STATISTICS=MEDIAN
  /ORDER=ANALYSIS.

Check data:

There are no impossible values on the weight variable.

Check assumptions:

The measurement level of variable weight is OK.
```

```
Interpret the results:

Median candy weight in the sample is 2.89 gram. With
95% confidence, we expect median candy weight to be
between 2.77 and 2.92 grams in the population of all candies.
```

3. Use SPSS to determine the 95% confidence interval for a paired-samples t test on candy colour fading under sunlight (variables colour_pre and colour_post in "candies.sav"). In your interpretation of the confidence interval, clarify the meaning of the statistic for which the confidence interval was calculated.

   The paired-samples t test is available in SPSS under *Analyze > Compare Means*.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=colour_pre colour_post
  /FORMAT=NOTABLE
  /HISTOGRAM NORMAL
  /ORDER=ANALYSIS.
* Paired-samples t test.
T-TEST PAIRS=colour_pre WITH colour_post (PAIRED)
  /CRITERIA=CI(.9500)
  /MISSING=ANALYSIS.

Check data:

There are no impossible values on variable weight.

Check assumptions:

Sample size (N = 50) is well over 30, so we need not care about
the normal distribution of candy weight in the population.

Interpret the results:

Candy colourfulness fades under sunlight by 1.88 (or between
1.6 to 2.2) points in the data. This change is statistically
significant, t (49) = 12.86, p < 001, 95%CI[1.59; 2.17].
```

4. Use SPSS to determine if candy colourfulness after exposure to sunlight (colour_post) depends on candy weight and candy sweetness. Interpret the 95% confidence intervals for both effects.

   Hint: Use regression analysis, which is available under *Analyze > Regression > Linear* in SPSS.

```
SPSS syntax:

* Check data: Assumption checks in Chapter 8.
* CHeck for impossible values.
FREQUENCIES VARIABLES=weight sweetness colour_post
  /ORDER=ANALYSIS.
* Regression of colour_post on weight and sweetness.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT colour_post
  /METHOD=ENTER weight sweetness.

Check data:

There are no impossible values on the variables.

Check assumptions:

Assumptions are presented later (Chapter 8).

Interpret the results:

We are 95 per cent confident that average candy
colourfulness decreases by 0.23 to 0.52 for
each additional unit of sweetness in the
population.
We are not confident about the direction of the
effect of candy weight on colourfulness. This
effect can be negative (up to a .65 decrease for
each additional candy weight gram) or positive
(up to a 2.05 increase for each additional candy
weight gram).
```

3.7 Test Your Understanding

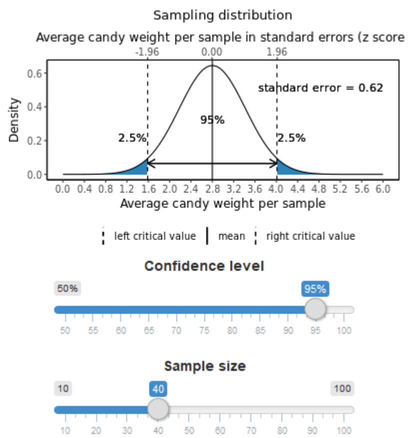


Figure 3.11: Point and interval estimates, confidence intervals.

Figure 3.11 shows the sampling distribution of average candy weight in a sample bag, which is a normal distribution.

1. What is the most likely estimate for average candy weight in the population?
2. The percentage in between the two vertical lines can be interpreted as a probability. A probability of what?
3. The double arrow represents an interval of sample means, in this example, average candy weight. What happens if you change the confidence level? Explain why this makes sense.
4. What happens to the graph if you change sample size?
5. What happens to the standard error if you change sample size? How are sample size and standard error linked? What characteristic of the sampling distribution is expressed by the standard error?
6. The values of the interval limits—average candy weights in this example—on the scale in standard errors are called critical values. What happens to the critical values if you change sample size?
7. What happens to the critical values if you change the confidence level?
8. If 2.8 is average candy weight in the sample but not necessarily true average candy weight in the population, the interval marked by the arrows is called a *confidence interval*. In this example, a confidence interval for what? And what is the point estimate?

3.8 Take-Home Points

- If a sample statistic is an unbiased estimator, we can use it as a point estimate for the value of the statistic in the population.
- A point estimate may come close to the population value but it is most certainly not correct.
- A 95% confidence interval is an interval estimate of the population value: We are 95% confident that the population value lies within this interval. Note that confidence is not a probability!
- A larger sample or a lower confidence level yields a narrower, that is, a more precise confidence interval. The sample statistic is more likely to be closer to the population value. We have less uncertainty in our results.
- A larger sample yields a smaller standard error, which yields a more precise confidence interval because the limits of a 95% confidence interval fall one standard error times the critical value below and above the value of the sample statistic.

Chapter 4

Testing a Null Hypothesis: Am I Right or Am I Wrong?

Key concepts: research hypothesis, statistical null and alternative hypothesis, nil hypothesis, test statistic, p value, significance level (Type I error rate), Type I error, inflated Type I error, capitalization on chance, one-sided and two-sided tests and tests to which this distinction does not apply, rejection region.

Summary

Is my sample probable if the null hypothesis is true?

In the preceding chapter, we have learned that a confidence interval contains the population values that are plausible, given the sample that we have drawn. In the current chapter, we narrow this down to the question whether the expectation of the researcher about the population is plausible.

The expectation is usually called a (research) hypothesis and it must be translated into statistical hypotheses about a population value (parameter): a null hypothesis and an alternative hypothesis.

We test the null hypothesis in the following way. We construct a sampling distribution in one of the ways we have learned in Chapter 2 using the value specified in the null hypothesis as the imaginary population value. In other words, we act as if the null hypothesis is true.

Then, we calculate the probability of drawing a sample such as the one we have drawn or a sample that differs even more from a population for which the null hypothesis is true. If this p value is very low, say, below 5%, we reject the null hypothesis because our sample would be too unlikely if the null hypothesis is true. In this case, the test is statistically significant. The probability threshold that we use is called the significance level of the test.

4.1 A Binary Decision

The overall goal of statistical inference is to increase our knowledge about a population, when we only have a random sample from that population. In Chapter 3, we estimated population values that are plausible considering the sample that we drew. For instance, we looked for all plausible average weights of candies in the population using information about the weight of candies in our sample bag. This is what we do when we estimate a population value.

Estimation is one of two types of statistical inference, the other being null hypothesis testing. When we estimate a population value, we do not use our previous knowledge about the world of candies or whatever other subject we are investigating. We can be completely ignorant about the phenomenon that we are investigating. This approach is not entirely in line with the conceptualisation of scientific progress as an *empirical cycle*, in which scientists develop theories about the empirical world, test these theories against data collected from this world, and improve their theories if they are contradicted by the data.

Hypothesis testing, however, is more in line with this conceptualisation of scientific progress. It requires the researcher to formulate an expectation about the population, usually called a *hypothesis*. If the hypothesis is based on theory and previous research, the scientist uses previous knowledge. As a next step, the researcher tests the hypothesis against data collected for this purpose. If the data contradict the hypothesis, the hypothesis is rejected and the researcher has to improve the theory. If the data does not contradict the hypothesis, it is not rejected and, for the time being, the researcher need not change their theory.

Hypothesis testing, then, amounts to choosing one of two options: reject or not reject the hypothesis. This is a binary decision between believing that the population is as it is described in the null hypothesis, or believing that it is not. This is quite a different approach from estimating a confidence interval as a range of plausible population values. Nevertheless, hypothesis testing and confidence intervals are tightly related as we will see later on in this chapter (Section 4.8).

4.2 Formulating Statistical Hypotheses

A *research hypothesis* is a statement about the empirical world that can be tested against data. Communication scientists, for instance, may hypothesize that:

- a television station reaches half of all households in a country,
- media literacy is below a particular standard (for instance, 5.5 on a 10-point scale) among children,
- opinions about immigrants are not equally polarized among young and old voters,
- the celebrity endorsing a fundraising campaign makes a difference to people's willingness to donate,
- more exposure to brand advertisements increases brand awareness,

- and so on.

As these examples illustrate, research hypotheses seldom refer to statistics such as means, proportions, variances, or correlations. Still, we need such statistics to test a hypothesis. The researcher must translate the hypothesis into a new hypothesis specifying a statistic in the population, for example, the population mean. The new hypothesis is called a *statistical hypothesis*.

Translating the research hypothesis into a statistical hypothesis is perhaps the most creative part of statistical analysis, which is just a fancy way of saying that it is difficult to give general guidelines stating which statistic fits which research hypothesis. All we can do is give some hints.

Research questions usually address shares, score levels, or score variation. If a research question talks about how frequent some characteristic occurs (How many candies are yellow?) or which part has a particular characteristic (Which percentage of all condies are yellow?), we are dealing with one or two categorical variables. Here, we need a binomial, chi-squared, or exact test (see Figure 4.1).

If a research question asks how high a group scores or whether one group scores higher then another group, we are dealing with score levels. The variable of central interest usually is numerical (interval or ratio measurement level) and we are concerned with mean or median scores. There is a range of tests that we can apply, depending on the number of groups that we want to compare (one, two, three or more): t tests or analysis of variance.

Instead of comparing mean scores of groups, a research question about score levels can address associations between numerical variables, for example, Are heavier candies more sticky? Here, the score level on one variable (candy weight) is linked to the score level on another variable (candy stickiness). This is where we use correlations or regression analysis.

Finally, a research question may address the size of the variation of numeric scores, for example, Does the weight of yellow candies vary more strongly than the weight of red candies? Variance is the statistic that we use to measure variation in numeric scores.

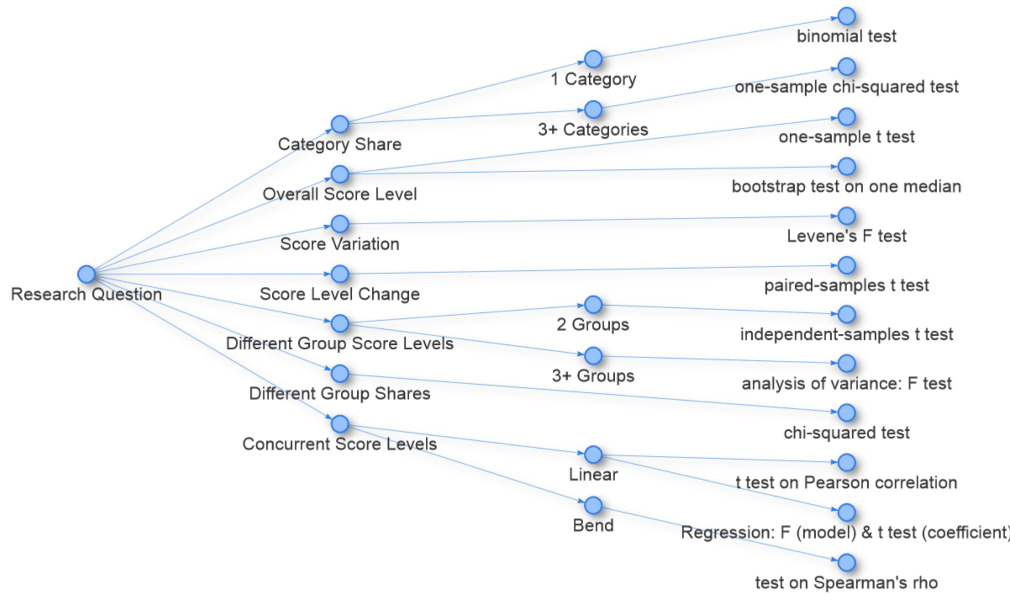


Figure 4.1: Flow chart for selecting the appropriate statistical test.

1. For each of the research hypothesis examples (above), find the appropriate statistical test using Figure 4.1. Hint: Hover your mouse pointer over a node, click on a dot, drag with your (left) mouse button, and zoom with your mouse wheel.

```
* A television station reaches half of all households in a country: Half of
all households refers to a category share. The variable is reaching a
household, which is a dichotomy (yes/no), so we use a binomial test.

* Media literacy is below a particular standard (for instance, 5.5 on a
10-point scale) among children: "Standard" refers to an overall score level.
We are testing a mean (not a median), so we can use a one-sample t test.

* Opinions about immigrants are not equally polarized among young and old
voters: "Polarization" refers to a stronger score variation in opinions. This
brings us to Levene's F test.

* The celebrity endorsing a fundraising campaign makes a difference to
people's willingness to donate: "Difference" probably means that people seeing
one endorser are on average more willing to donate than people seeing another
(or no) celebrity endorser. This refers to score level differences between
groups, which are tested with an indpendent-samples t test or analysis of
variance, depending on the number of groups that we compare.

* More exposure to brand advertisements increases brand awareness: "More ...
```

```
increases ..." refers to concurrent score levels. People with more exposure
are people with more brand awareness. If the association is more or less
linear, we can use a correlation or regression analysis. If the association is
clearly bend, we had better use Spearman's rank correlation.
```

4.2.1 Proportions: shares

This book covers tests for four types of statistics: proportions, means, variance/standard deviations, and associations. A proportion is the statistic best suited to test research hypotheses addressing the share of a category or entity in the population. The hypothesis that a television station reaches half of all households in a country provides an example. All households in the country constitute the population. The share of the television station is the proportion or percentage of all households watching this television station.

If we want to use a statistic, we need to know the variable and cases (units of analysis) for which the statistic must be calculated. In this example, a household does or does not watch the television statement, so our variable is a dichotomy with the two categories ("No, does not watch this station", "Yes, watches this station") usually coded as 0 versus 1 or 1 versus 2.

Each household provides an observation, namely either the score 0 or the score 1 on this variable or no score if there are missing values. To test the research hypothesis that a television station reaches half of all households in a country, we have to formulate a statistical hypothesis about the proportion—of households viewing this television station—in the population—all households in this country. For example, the researcher's statistical hypothesis could be that the proportion in the population is 0.5.

We can also be interested in more than two categories, for instance, does the television station reach the same share of all households in the north, east, south, and west of the country? This translates into a statistical hypothesis containing three or more proportions in the population. If 30% of households in the population are situated in the west, 25 % in the south and east, and 20% in the north, we would expect these proportions in the sample if all regions are equally represented. Our statistical hypothesis is actually a relative frequency distribution, such as, for instance, in Table 4.1.

Table 4.1: Statistical hypothesis about four proportions as a frequency table.

Region	Hypothesized Proportion
North	0.20
East	0.25
South	0.25
West	0.30

A test for this type of statistical hypothesis is called a one-sample chi-squared test. It is up to the researcher to specify the hypothesized proportions for all categories. This is not a simple task: What reasons do you have to expect particular values, say a region's share of thirty per cent of all households instead of twenty-five per cent?

The test is mainly used if the researcher has information on the proportions of the categories in the population. If we draw a sample from all citizens of a country, we usually know the frequency distribution of sex, age, educational level, and so on of all citizens from the national bureau of statistics. With the bureau's information, we can test if the respondents in our sample have the same distribution with respect to sex, age, or educational level as the population; just use the population proportions in the hypothesis. If they do, the sample is *representative* (see Section 1.2.6) of the population with respect to sex, age, or educational level. This is an important check on the representativeness of our sample.

4.2.2 Testing proportions in SPSS

4.2.2.1 Instructions

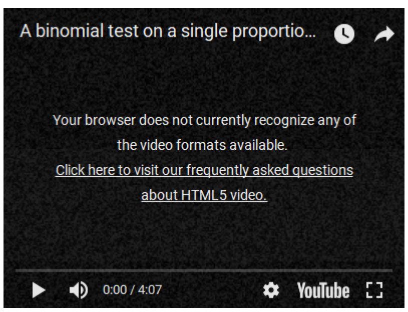


Figure 4.2: A binomial test on a single proportion in SPSS.

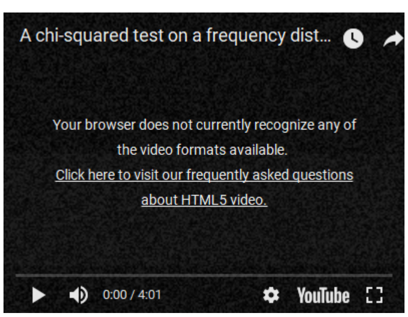


Figure 4.3: A chi-squared test on a frequency distribution in SPSS.

4.2.2.2 Exercises

1. Use the data set households.sav to test the hypothesis that the TV station does not reach 40 per cent of all households in the population.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=tv_reach
  /ORDER=ANALYSIS.
* Binomial test.
* Note: The test is one-sided if the test proprtion is not 0.50.
NPAR TESTS
  /BINOMIAL (0.40)=tv_reach
  /MISSING ANALYSIS.

Check data & assumptions:

Variable tv_reach is a dichotomy as it should be for this test.

Interpret the results:

We have to reject the null hypothesis that the TV station does not reach at
least fourty per cent of all households, p = .039 (one-sided).
```

2. Test the hypothesis that the TV station reaches 55 per cent of all households in the population.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=tv_reach
  /ORDER=ANALYSIS.
* Binomial test.
* Hint: Test the proportion of households not reached because
  this is the first category: 1 - 0.55 = 0.45.
NPAR TESTS
  /BINOMIAL (0.45)=tv_reach
  /MISSING ANALYSIS.

Check data & assumptions:

Variable tv_reach is a dichotomy as it should be for this test.

Interpret the results:

We cannot reject the null hypothesis that the TV station reaches
```

```
at least 55 per cent of all households, p = .260 (one-sided).
```

3. Does half of the households have an income of at most 40,000?

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=income
  /ORDER=ANALYSIS.
* Binomial test.
* Use the cut off option in the binomial test.
NPAR TESTS
  /BINOMIAL (0.50)=income (40000)
  /MISSING ANALYSIS.

Check data:

* There are no apparent impossible income values.

Check assumptions:

There are no assumptions for the binomial test.

Interpret the results:

The proportion of households with an income of at most 40,000
is significantly less than fifty per cent, p = .022. It is
39 per cent in the sample.
```

4. According information from the National Bureau of Statistics, 20 per cent of all households have incomes up to 30,000, 50 per cent have incomes between 30,000 and 50,000, and 30 per cent has incomes over 50,000. Use a test to decide if our sample is representative with respect to income. Hint: recode income first.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=income
  /ORDER=ANALYSIS.
* Recoding income into groups.
RECODE income (Lowest thru 30000=1) (30000  thru 50000=2)
  (50000 thru Highest=3) INTO income_group.
VARIABLE LABELS  income_group 'Grouped income'.
EXECUTE.
* Define Variable Properties.
*income_group.
VALUE LABELS income_group
```

```
  1.00 'low'
  2.00 'medium'
  3.00 'high'.
EXECUTE.
* one-sample chi-squared test.
NPAR TESTS
  /CHISQUARE=income_group
  /EXPECTED=20 50 30
  /MISSING ANALYSIS.

There are no apparent impossible income values.

Check assumptions:

Assumptions for the chi-squared test:

* Expected frequencies never below 1 and max 20% below 5: OK, the SPSS table
note says: "0 cells (0.0%) have expected frequencies less than 5. The minimum
expected cell frequency is 24.0."
* In this case, 0% of the cells have expected frequencies below 5, which is
less than the allowed maximum of 20%. In other words, at least 80% of the
cells have expected frequencies of 5 or more.

Interpret the results:

The distribution of incomes over income groups in the sample
does not differ in a statistically significant way from the
distribution in the population according to the National
Bureau of Statistics, chi-squared (2) = 3.40, p = .182.
In other words, we have no reason to believe that our
sample is not representative of the population with
regards to income.
```

4.2.3 Mean and median: level

Research hypotheses that focus on the level of scores are usually best tested with the mean or another measure of central tendency such as the median value. For example, the hypothesis that media literacy is below a particular standard (e.g., 5.5 on a 10-point scale) among children refers to a level: the level of media literacy scores.

The hypothesis probably does not argue that all children have a media literacy score below 5.5. Instead, it means to say that the overall level is below this standard. The centre of the distribution offers a good indication of the general score level.

For a numeric (interval or ratio measurement level) variable such as the 10-point scale in the example, the mean is a good measure of the distribution's centre. In this example, our statistical hypothesis would be that average media literacy score of all children in the population is (below) 5.5.

4.2.4 Testing one mean in SPSS

4.2.4.1 Instructions

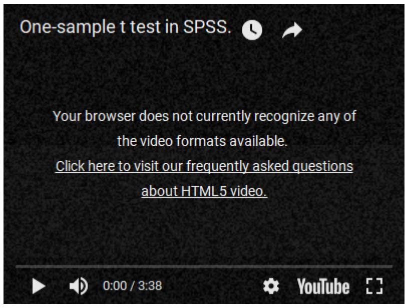


Figure 4.4: A one-sample t test in SPSS.

4.2.4.2 Exercises

1. Use the data set children.sav to test the hypothesis that average parental supervision of the child's media use is 5.5 (on a scale from 1 to 10) in the population.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=supervision
  /ORDER=ANALYSIS.
* Set imposible value (25) to missing.
* Define Variable Properties.
*supervision.
MISSING VALUES supervision(25.00).
EXECUTE.
* One-sample t test.
T-TEST
  /TESTVAL=5.5
  /MISSING=ANALYSIS
  /VARIABLES=supervision
  /CRITERIA=CI(.95).

Check data:

There is one impossible value for parental supervision,
```

```
namely 25. This value must be made missing.

Check assumptions:

Sample size (N = 86) is well over 30, so we need no worry
about the shape of the variable distribution in the
population.

Interpret the results:

We are rather confident that the true average supervision
score in the population is around 5.5, more specifically,
between 4.94 and 5.77, t (85) = -0.68, p = .498, 95%CI[4.94; 5.77].

Note that we have to add the confidence interval limits
to the test value (here: 5.5) to obtain the confidence
interval for the population mean.
```

2. Have a look at the confidence interval reported for Exercise 1. If you would test the hypothesis that average parental supervision in the population is 4.5, would the test be statistically significant? Check your answer by carrying out this test.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=supervision
  /ORDER=ANALYSIS.
* Set imposible value (25) to missing.
* Define Variable Properties.
*supervision.
MISSING VALUES supervision(25.00).
EXECUTE.
* One-sample t test.
T-TEST
  /TESTVAL=4.5
  /MISSING=ANALYSIS
  /VARIABLES=supervision
  /CRITERIA=CI(.95).

Check data:

There is one impossible value for parental supervision,
namely 25. This value must be made missing.

Check assumptions:
```

```
Sample size (N = 86) is well over 30, so we need no worry
about the shape of the variable distribution in the
population.

Interpret the results:

The average parental supervision score in the sample
(M = 5.39, SD = 1.95) makes us doubt that the true
average supervision score in the population is 4.5 or
thereabouts, t (85) = 4.10, p < .001, 95%CI[4.94; 5.77].

Note that we obtain the same confidence interval as in
Exercise 1. The confidence interval does not depend on
the null hypothesis, whereas the significance test does.
```

4.2.5 Variance: (dis)agreement

Although rare, research hypotheses may focus on the variation in scores rather than on score level. The hypothesis about polarization provides an example. Polarization means that we have scores well above the centre and well below the centre rather than all scores concentrated in the middle. If voters' opinions about immigrants are strongly polarized, we have a lot of voters strongly in favour of admitting immigrants as well as many voters strongly opposed to admitting immigrants.

For a numeric variable, the variance or standard deviation—the latter is just the square root of the former—is the appropriate statistic to test a hypothesis about polarization. The research hypothesis concerns the variation of scores in two groups, for instance, young versus old voters. The statistical hypothesis would be that the variance in opinions in the population of young voters is different from the variance in the population of old voters.

4.2.6 Testing two variances in SPSS

4.2.6.1 Instructions

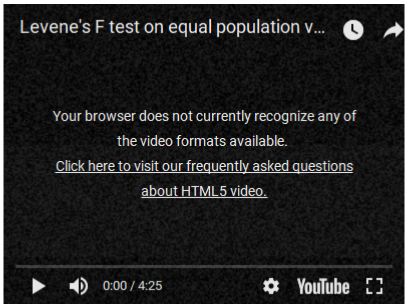


Figure 4.5: Levene's F test on equal population variances in SPSS.

4.2.6.2 Exercises

1. Data set voters.sav contains information about the age and attitude towards immigration among a random sample of voters. Is the attitude towards immigrants equally polarized among young (under 30) and old (30+) voters? Justify your answer with a statistical test.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=age_group immigrant
  /ORDER=ANALYSIS.
* Independent-samples t test with Levene s test.
T-TEST GROUPS=age_group(1 2)
  /MISSING=ANALYSIS
  /VARIABLES=immigrant
  /CRITERIA=CI(.95).

Check data:

There are no impossible values on the variables.

Check assumptions:

There are no assumptions that we have to check for
the Levene test.

Interpret the results:
```

```
The attitude towards immigrants is more polarized among
older voters (SD = 2.06) than among young voters (SD = 1.26).
The difference in variation is statistically significant,
F = 4.99, p = .029.
Note that SPSS does not report the (two) degrees of freedom
of the F test, so we cannot report them either.
SPSS, however, does report the degrees of freedom of Levene's F test in a
one-way analysis of variance. We could have used that approach here as well.
```

2. Use the data of Exercise 1. Create a new variable grouping voter's age with classes 18-35, 36-65, and 66+ years. Is the attitude towards immigrants equally polarized among these three age groups in the population? Justify your answer with a statistical test.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=age immigrant
  /ORDER=ANALYSIS.
* Group age.
RECODE age (Lowest thru 35=1) (36 thru 65=2)
  (66 thru Highest=3) INTO age3.
VARIABLE LABELS  age3 'Voter ages in three groups'.
EXECUTE.
* Define Variable Properties.
*age3.
VALUE LABELS age3
  1.00 '18-35'
  2.00 '36-65'
  3.00 '66+'.
EXECUTE.
* ANOVA with descriptives.
ONEWAY immigrant BY age3
  /STATISTICS DESCRIPTIVES HOMOGENEITY
  /MISSING ANALYSIS.

Check data:

There are no impossible values on the variables.

Check assumptions:

There are no assumptions that we have to check for
the Levene test.

Interpret the results:
```

```
The attitude towards immigrants seems to be more polarized
among middle-aged (SD = 2.17) and aged voters (SD = 2.06)
than among young voters (SD = 1.37).
The difference in variation, however, is not statistically
significant, F (2, 63) = 2.77, p = .070.
This result seems to contradict the statistically significant
test result that we found in Exercise 1, comparing only young
to old voters. There is no contradiction, however. We merely
see that the classification into groups can matter to the
significance of results.
```

4.2.7 Association: relations between characteristics

Finally, research hypotheses may address the relation between two or more variables. Relations between variables are at stake if the research hypothesis states or implies that one (type of) characteristic is related to another (type of) characteristic. The statistical name for a relation between variables is *association*.

Take, for example, an analysis of the effect of a celebrity endorser on the willingness to donate. Here, the endorser to whom a person is exposed (one characteristic) is related to this person's willingness to donate (another characteristic). Another example: If exposure to the campaign increases willingness to donate, a person's willingness to donate is positively related to this person's exposure to the campaign.

4.2.8 Score level differences

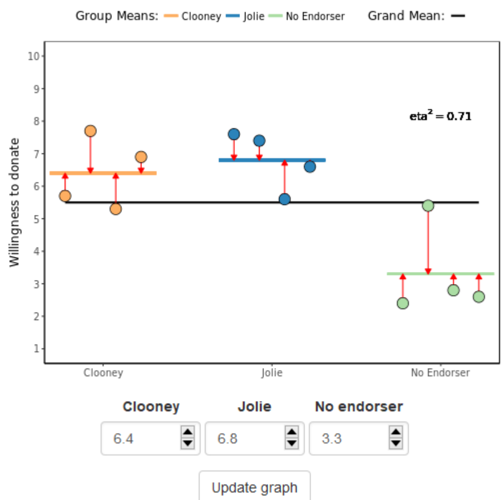


Figure 4.6: How do group level differences express association?

1. Figure 4.6 shows the willingness to donate for twelve respondents and the celebrity they saw endorsing the fund-raising campaign. Do you think that the willingness to donate is associated with the person endorsing the fund-raiser? Motivate your answer.

```
Respondents who saw a campaign without celebrity endorser are on average much
less willing to donate to the fund-raiser than respondents who saw George
Clooney or Angelina Jolie. This suggests that the endorser makes a difference
to willingness to donate, so the two are associated. More specifically, it
seems to make a difference whether the endorser is a celebrity or not.
```

2. Can you create or remove an association between willingness to donate and endorsing celebrity by changing the group averages? Change the group averages to check your expectations.

```
If we increase the differences between the average group scores, the
association becomes stronger. Eta-squared, for example, becomes larger. But if
we decrease the differences between the average group scores, the association
becomes weaker. and eta-squared becomes smaller.
```

Association comes in two related flavours: a difference in score level between groups or the predominance of particular combinations of scores on different variables.

The relation between the endorser's identity and willingness to donate is an example of the first flavour. All people are confronted with one of the celebrities as endorser of the fund-raising campaign. This is captured by a categorical variable: the endorsing celebrity.

The categorical variable clusters people into groups: One group is confronted with Celebrity

A, another group with Celebrity B, and so on. If the celebrity matters to the willingness to donate, the general level of donation willingness should be higher in the group exposed to one celebrity than in the group exposed to another celebrity.

Thus, we return to statistics needed to test research hypotheses about score levels, namely measures of central tendency. If willingness to donate is a numeric variable, we can use group means to test the association between endorsing celebrity (grouping variable) and willingness to donate (score variable). The statistical hypothesis would then be that group means are not equal in the population of all people.

If you closely inspect Figure 4.1, you will see that we prefer to use a t distribution if we compare two different groups (independent-samples t test) or two repeated observations for the same group (paired-samples t test). By contrast, if we have three or more groups, we use analysis of variance with an F distribution.

4.2.9 Comparing means in SPSS

4.2.9.1 Instructions

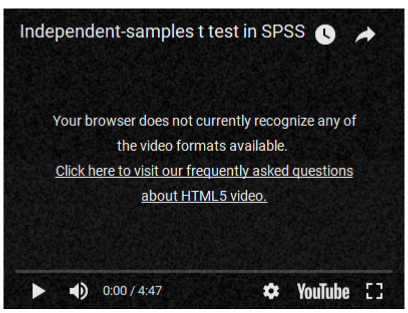


Figure 4.7: An independent-samples t test in SPSS.

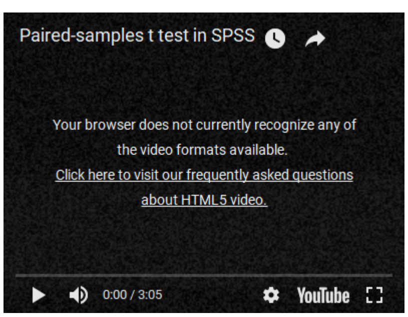


Figure 4.8: A paired-samples t test in SPSS.

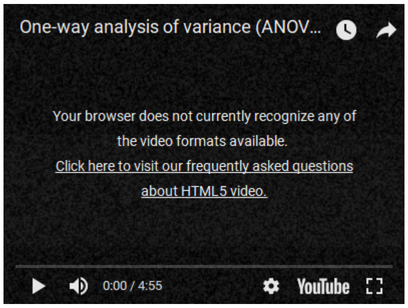


Figure 4.9: One-way analysis of variance (ANOVA) in SPSS.

4.2.9.2 Exercises

1. Use donors.sav to test if people's willingness to donate at the end of the campaign depends on the celebrity endorsing the campaign.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=willing_post endorser
  /ORDER=ANALYSIS.
* One-way analysis of variance.
ONEWAY willing_post BY endorser
  /STATISTICS DESCRIPTIVES HOMOGENEITY
  /PLOT MEANS
  /MISSING ANALYSIS
  /POSTHOC=BONFERRONI ALPHA(0.05).

Check data:

There are no impossible values on the variables.

Check assumptions:

The three groups are more or less of equal sizes:
The largest difference is 4 participants, which is less
than ten per cent of the smallest group (N = 45).
Anyway, we may assume equal population variances,
Levene F (2, 140) = .02, p = .978.

Interpret the results:
```

```
Willingness to donate depends on the endorsing celebrity.
There is a statistically significant difference between average
willingness to donate for the three endorsers, F (2, 140) = 7.44,
p = .001.
People are more willing to donate if they have seen Clooney
(M = 4.99, SD = 1.64) or Jolie (M = 4.95, SD = 1.63) endorse the
fund raiser than people who do not see a celebrity endorser
(M = 3.87, SD = 1.47). The differences between, on the one
hand, no celebrity endorser and, on the other hand, Clooney
(p = .002) or Jolie (p = .004) are statistically significant.

Instead of reporting the F test result in the text, the
ANOVA table can be included.
```

2. Is willingness to donate at the end of the campaign higher for those who remember the campaign than for those who do not remember it?

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=willing_post remember
  /ORDER=ANALYSIS.
* Independent-samples t test.
T-TEST GROUPS=remember(0 1)
  /MISSING=ANALYSIS
  /VARIABLES=willing_post
  /CRITERIA=CI(.95).

Check data:

There are no impossible values on the variables.

Check assumptions:

Sample sizes (N = 66 and N = 77) are of sufficient
size not to worry about the shape of the distribution
of willingness in the population.

Interpret the results:

Willingness to donate at the end of the campaign is significantly
higher for those who remember the campaign (M = 4.94, SD = 1.60)
than for those who do not remember it (M = 4.24, SD = 1.65),
t (141) = -2.57, p = .011, 95%CI[-1.24, -0.16].
Willingness is 0.16 to 1.24 points higher on a 10-point scale for
those who remember the campaign. Considering the range of the
scale, this is quite a small difference.
```

3. Did willingness to donate increase in the population between the start and the end of the campaign?

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=willing_post willing_pre
  /ORDER=ANALYSIS.
* Paired-samples t test.
T-TEST PAIRS=willing_pre WITH willing_post (PAIRED)
  /CRITERIA=CI(.9500)
  /MISSING=ANALYSIS.

Check data:

There are no impossible values on the two variables.

Check assumptions:

Sample size is sufficiently large (N = 143).

Interpret the results:

There is a small but statistically significant increase in
the willingness to donate over the duration of the experiment,
t (142) = 5.74, p < .001, 95%CI[0.08; 0.17].
Average willingness to donate is 0.08 to 0.17 units higher
at the end of the experiment than at the start. This is a
very small difference on a 10-point scale.
```

4.2.10 Combinations of scores

The other flavour of association represents situations in which some combinations of scores on different variables are much more common than other combinations of scores.

Think of the hypothesis that brand awareness is related to exposure to advertisements for that brand. If the hypothesis is true, people with high exposure and high brand awareness should occur much more often than people with high exposure and low brand awareness or low exposure and high brand awareness.

The two variables here are exposure and brand awareness. One combination of scores on the two variables is high exposure combined with high brand awareness. This combination should be more common than high exposure combined with low brand awareness.

Measures of association are statistics that put a number to the pattern of group-score levels or combinations of scores. The exact statistic that we use depends on the measurement level of the variables. For numerical variables, measured at the interval or ratio level, we use Pearson's correlation coefficient or the regression coefficient. For ordinal variables with quite a lot of different scores, we use Spearman's rank correlation.

For categorical variables, measured at the nominal or ordinal level, chi-squared indicates whether variables are statistically associated. The larger chi-squared, the less probable it is that the variables are not associated in the population. If variables are not associated, they are said to be *statistically independent*.

Several measures exist that express the strength of the association between two categorical variables. We use Phi and Cramer's V (two nominal variables, symmetric association), Goodman & Kruskals tau (two nominal variables, asymmetric association), Kendalls tau-b (two categorical ordinal variables, symmetric association), and Somers' d (two categorical ordinal variables, asymmetric association).

4.2.11 Testing associations in SPSS

4.2.11.1 Instructions

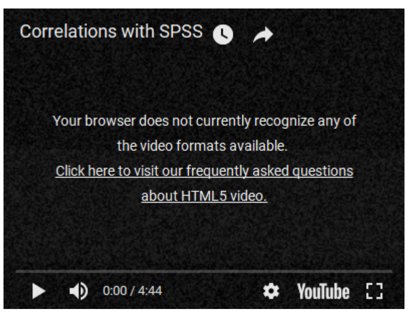


Figure 4.10: Correlations with SPSS.

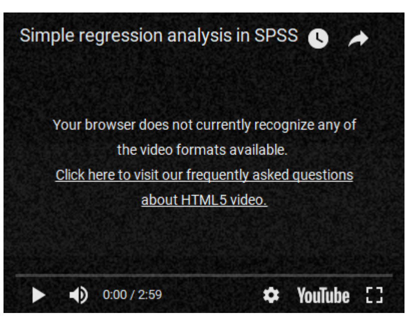


Figure 4.11: Executing and interpreting regression analysis in SPSS.

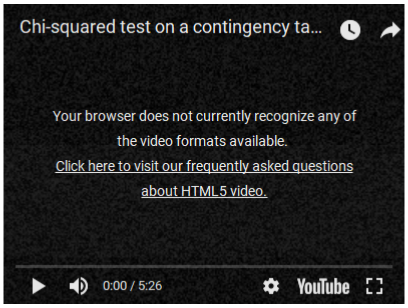


Figure 4.12: Chi-squared test on a contingency table with SPSS.

4.2.11.2 Exercises

1. In the population of all consumers, is brand awareness linked to exposure to advertisements for the brand? Use consumers.sav to answer this question.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=ad_expo brand_aw
  /ORDER=ANALYSIS.
* Check if the association can be linear.
GRAPH
  /SCATTERPLOT(BIVAR)=ad_expo WITH brand_aw
  /MISSING=LISTWISE.
* Correlations.
CORRELATIONS
  /VARIABLES=ad_expo brand_aw
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
NONPAR CORR
  /VARIABLES=ad_expo brand_aw
  /PRINT=SPEARMAN TWOTAIL NOSIG
  /MISSING=PAIRWISE.

Check data:

There are no impossible values that must be changed
into missing values.

Check assumptions:

Can the association be linear? If we check a scatterplot
```

```
of the two variables, the points do not clearly display
a curved shape. But there is one point that may distort
a linear association because it is far away from the
other points, namely a consumer with an exposure score
near one. If the rank correlation is substantially
higher than the Pearson correlation, this single
observation may be responsible.

Interpret the results:

Brand awareness is statistically signficantly associated
with exposure to advertisements for the brand, r = .46,
p < .001. More exposure tends to go together with more
brand awareness.
```

2. How well can we predict brand awareness with ad exposure?

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=ad_expo brand_aw
  /ORDER=ANALYSIS.
* Simple regression.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT brand_aw
  /METHOD=ENTER ad_expo.

Check data:

There are no impossible values.

Check assumptions:

The assumptions will be explained in the chapter on
moderation with regression analysis, so let us not pay
attention to the assumptions yet.

Interpret the results:

Ad exposure predicts about one fifth of the
variation in brand awareness scores, R2 = .21,
F (1, 60) = 15.87, p < .001.
The predictive effect of exposure to brand advertisements
```

```
is moderately strong (b* = 0.46). An additional
unit of exposure increases the predicted brand awareness
with 0.4 points, t = 3.98, p < .001, 95%[0.22; 0.67].
```

3. Does word of mouth involve women rather than men? Interpret the contents, strength, and statistical significance of the association. Have a look at Section 2.4 if you forgot which p value to interpret.

```
SPSS syntax:

* Contingency table with chi-squared test and measure of association.
CROSSTABS
  /TABLES=wom BY gender
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ PHI LAMBDA
  /CELLS=COUNT COLUMN
  /COUNT ROUND CELL
  /BARCHART.

Check data:

There are no impossible values on the two categorical
variables.

Check assumptions:

This is a 2x2 contingency table so we have to use (Fisher)
exact test. This test makes no assumptions.

Interpret the results:

There is no statistically significant association between
word of mouth and sex, p = .161 (Fisher exact), Goodman &
Kruskal tau = .05. We cannot confidently conclude that
either females or males experience word of mouth more
frequently.

Note that a one-sided test is possible too. The p value
of a one-sided exact test is .080 here.
```

4.3 The Null and Alternative Hypothesis

In the preceding section, I referred to statistical hypotheses without further qualification. There are, however, at least two different statistical hypotheses: the null hypothesis (H_0) and the alternative hypothesis (H_1 or H_A).

For reasons that will be explained in Section 4.5.1, the *null hypothesis* must equate the test statistic to a single value. For example, the statistical hypothesis that the proportion of all households in the population reached by a television station is .5 equates the population proportion to .5.

4.3.1 Alternative hypothesis

If the research hypothesis is not the null hypothesis, it is the *alternative hypothesis*. The alternative hypothesis covers all situations not covered by the null hypothesis and the other way round. The null hypothesis stating that the proportion of all households in the population reached by the television station is .5 is linked to the alternative hypothesis that states the proportion is not .5. Thus, we cover all possible outcomes.

The research hypothesis stating that the average media literacy score of all children in the population is below 5.5, is an example of an alternative hypothesis. Most of the research hypothesis examples in the previous section are alternative hypotheses because they do not equate the statistic with a particular value:

- If opinions about immigrants are hypothesized to be different for younger and older voters, the variances are hypothesized to be unequal in the population. But the research hypothesis does not state how unequal.
- If the celebrity endorsing a fundraising campaign makes a difference to the willingness of people to donate, the average willingness is not hypothesized to be equal for all groups but, again, the size of the difference is not specified.
- If we hypothesize that more exposure to brand advertisements increases brand awareness, we expect the correlation or regression coefficient to be positive, that is, larger than zero. But we have not hypothesized a particular value, so the research hypothesis represents the alternative hypothesis.

4.3.2 Research hypotheses tend to be alternative hypotheses

It is not a coincidence that most of the research hypotheses in the examples are alternative hypotheses instead of null hypotheses. This is very common in social research, even though it is not necessarily always the case, as some statistics textbooks would have us believe. Often, our theories tell us to expect differences or changes but not the size of differences or changes.

If the research hypothesis is the alternative hypothesis, we have to formulate the null hypothesis ourselves. This is very important, because it is this hypothesis that is actually tested as we will see in a later section. Some examples:

- A research hypothesis stating that average media literacy of children is below 5.5 is an alternative hypothesis. The associated null hypothesis would be that average media literacy is at least 5.5.

- If the alternative hypothesis states that variances or group means are unequal in the population, the null hypothesis would be that they are equal in the population.
- An alternative hypothesis expecting a correlation between exposure and brand awareness requires the null hypothesis to state that there is no such association in the population.

4.3.3 Nil hypothesis

Null hypotheses are quite often stating that there is no difference or no association in the population. They equate the population statistic to zero. This type of null hypothesis is called a *nil hypothesis* or just plainly *the nil*.

A null hypothesis on association in the population usually is a nil hypothesis, assigning the value zero to the measure of association in the population. For example, Spearman's rho or Pearson's correlation between exposure and brand awareness are hypothesized to be zero in the population. For a measure of association, zero always means that there is no association.

This also applies to the regression coefficient (b or b^* in the regression equation): If it is zero, the predictor variable is useless for predicting the outcome variable. In these cases, the null hypothesis is the nil. If statistical software does not report the null hypothesis that is being tested, you may assume that it equates the parameter of interest to zero.

4.4 One-Sided and Two-Sided Tests

The research hypothesis stating that average media literacy is below 5.5 in the population represents the alternative hypothesis because it does not fix the hypothesized population value to one number. The accompanying null hypothesis must state that the population mean is 5.5 or higher.

This null hypothesis is slightly different from the ones we have encountered so far, which equated the population value to a single value, usually zero. If the null hypothesis equates a parameter to a single value, the null hypothesis can be rejected if the sample statistic is either too high or too low. There are two ways of rejecting the null hypothesis, so this type of hypothesis is called *two-sided* or two-tailed.

By contrast, the null hypothesis stating that the population mean is 5.5 or higher is a *one-sided* or one-tailed hypothesis. It can only be rejected if the sample statistic is at one side of the spectrum: only below (left-sided) or only above (right-sided) a particular value. A test of a one-sided null hypothesis is called a *one-sided test*.

In a left-sided test of the media literacy hypothesis, the researcher is not interested in demonstrating that average media literacy among children can be larger than 5.5. Usually, she is not interested because she rules out the possibility. For instance, she believes that average media literacy among children cannot be substantially higher than 5.5. The researcher brings prior knowledge about the world to bear that has convinced her that average media literacy among children can only be lower than 5.5 on average in the population.

If there is a possibility that the score may also be higher than 5.5 and it is deemed important to note values well over 5.5 and values well below 5.5, the research and null hypotheses should be two-sided. Then, a sample average well above 5.5 would also have resulted in a rejection of the null hypothesis. In a left-sided test, however, a high sample outcome cannot reject the null hypothesis.

4.4.1 Boundary value as hypothesized population value

You may wonder how a one-sided null hypothesis equates the parameter of interest with one value as it should. The special value here is 5.5. If we can reject the null hypothesis stating that the population mean is 5.5 because our sample mean is sufficiently lower than 5.5, we can also reject any hypothesis involving population means higher than 5.5.

In other words, if you want to know if the value is not 5.5 or more, it is enough to find that it is less than 5.5. If it's less than 5.5, then you know it's also less than any number above 5.5. Therefore, we use the boundary value of a one-sided null hypothesis as the value for a one-sided test.

4.4.2 One-sided – two-sided distinction is not always relevant

Note that the difference between one-sided and two-sided tests is only useful if we test a statistic against a particular value or if we test the difference between two groups. In the first situation, we may rule out the possibility that the population value is higher (or lower) than the hypothesized value if we have good reasons to believe that it can only be lower (or higher). In the second situation, we may expect that one group can only score higher than the other group and not the other way around.

In contrast, we cannot meaningfully formulate a one-sided null hypothesis if we are comparing three groups or more. Even if we expect that Group A can only score higher than Group B and Group C, what about the difference between Group B and Group C? If we can't have meaningful one-sided null hypotheses, we cannot meaningfully distinguish between one-sided and two-sided null hypotheses.

4.4.3 Formulate statistical hypotheses in advance

Statistical hypotheses are important because they specify the statistic that we use in our test and they determine whether we do a two-sided or one-sided test. We have to formulate our statistical hypotheses before we have a look at our sample. After all, we want to give the sample data a fair chance to prove our hypothesis wrong. A statistical test is useless if we already know the result in the sample and adjust our hypotheses to this information.

Finally, you should note that scientific papers usually report the research hypothesis but not the statistical hypotheses. The statistical hypotheses of a test are supposed to be known to all researchers. Make sure you know them, as well.

4.5 p Value and Significance Level (α)

The purpose of formulating a null hypothesis is that we can use the value specified in the null hypothesis as a hypothetical population value. This saves us the trouble of looking for plausible population values, which we do if we estimate a confidence interval.

When testing a null hypothesis, we just act as if the null hypothesis is true. We pretend that the value specified in the null hypothesis is the true population value. This allows us to create one sampling distribution that will tell us how plausible our sample is if the null hypothesis would be true. For this reason, the null hypothesis must equate the parameter to one value.

Let us assume that average media literacy is 3.9 in our sample. According to our null hypothesis, the population average is (at least) 5.5. If average media literacy of children in the population would really be 5.5, how plausible is it to draw a sample with 3.9 (or less) as average media literacy? We can use a hypothetical sampling distribution with 5.5 as mean value to answer this question.

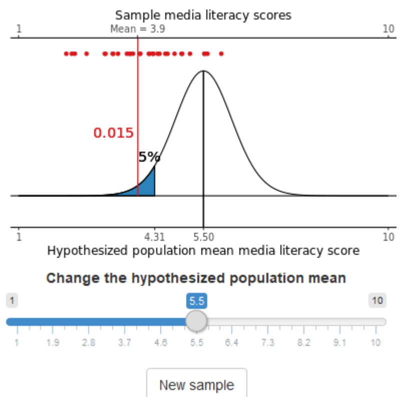


Figure 4.13: Sampling distribution of average media literacy.

1. Figure 4.13 shows the hypothesized population mean, the associated sampling distribution, and the sample scores (red dots) with their mean. What does the red number directly to the left of the sample mean line mean?

```
* It is the probability of drawing a sample with average media literacy cores
below 3.9 from a population with 5.5 as average media literacy scores.
* This is called a (left-sided) p value.
```

2. The significance level is 5% here. Why is it marked by the blue tail in Figure 4.13?

```
* The significance level is the maximum risk that we want to take to reject a
null hypothesis that is true (Type I error). This risk is a probability, so it
is expressed by a surface under the (probability density) curve.
* Here, we are only interested in sample averages that are to low, so the
entire risk is situated in the left tail.
```

3. How low must the sample mean be to have a p value below 5% (a statistically significant test result)?

```
* The limit(s) of the tail that repesent the significance level mark the
rejection region. The limit is called a critical value.
* In the initial example, with average population media literacy hypothesized
to be 5.5, the rejection region contains all sample mean scores below 4.31.
Any sample with average media literacy below 4.31 has a (left-sided) p value
below 5%, so it is statistically significantly different from the hypothesized
population value.
```

4. What happens to the p value of our initial sample mean (3.9) if we change the value of our null hypothesis? Is it always possible to formulate a null hypothesis such that the sample mean is statistically significant? Take some new samples to check your answer.

```
* If the null hypothesis changes, the average, hence the centre of the
sampling distribution moves left or right. And so does the critical value.
* If the sample mean is outside the rejection region, that is, it is larger
than the critical value, choose a higher population mean as your null
hypothesis. If it is sufficiently higher than the sampe mean, the sample mean
ends up in the rejection region under the blue tail. Then, the test is
statistically significant.
```

4.5.1 p Value

If our sample statistic is a continuous variable, for instance, average media literacy, we know that it does not make sense to calculate the probability of finding the exact sample mean that we obtained. In Section 1.3, we have learned that we work with p values in this situation. For example, we use the probability of finding our sample mean or a mean that is even more distant from the hypothesized population mean.

The question now is: How large must the difference be between the value that we expect according to our null hypothesis (the hypothesized population value), and the value that we observe in our sample, before we stop believing that the null hypothesis is true? If our null hypothesis expects the average media literacy of all children to be (at least) 5.5, how small should the sample mean be before we reject the null hypothesis?

The answer is that the difference between hypothesized and sample average media literacy must be so large that it is very improbable that we would draw a sample with the observed mean from a population with the hypothesized mean. In statistical terms, the p value of our sample outcome must be very low before we reject the null hypothesis. With a very low p value, our sample is too improbable to sustain our belief that the null hypothesis is true.

4.5.2 Statistical significance and rejection region

How low should we go? A commonly accepted threshold value is .05 (5%). If the p value is .05 or less, we decide that our sample is too improbable and implausible to be drawn from a population for which the null hypothesis is true. Therefore, we reject the null hypothesis and we say that the test is *statistically significant*.

The decision about the null hypothesis is simple if you have the p value. If the reported p value is lower than the significance level, we reject the null hypothesis. Otherwise, we do not reject the null hypothesis.

> If p is low, the null must go and the test is statistically significant.

This is the golden rule of null hypothesis testing (although some argue that the gold of this rule is fool's gold, see Chapter 6).

The values for the sample statistic for which the test is statistically significant, so that the null hypothesis is rejected, is called the *rejection region*. If the sample statistic is in the rejection region of a test, we reject the null hypothesis, and the test is statistically significant.

However, we usually do not know the rejection region in terms of the sample statistic, so we ordinarily use p values to determine if a test is statistically significant. SPSS reports p values, which are sometimes referred to as *significance* or *Sig.*.

4.5.3 Conditional probability

It is important to remember that the p value that we calculate is a probability **under the assumption that the null hypothesis is true**. Therefore, it is a *conditional probability*.

Compare it to the probability that we throw sixes with a dice. This probability is one out of six under the assumption that the dice is unbiased. Probabilities always rest on assumptions. If the assumptions are violated, we cannot calculate probabilities.

If the dice is biased, we don't know the probability of throwing sixes. In the same way, we have no clue whatsoever of the probability of drawing a sample like the one we have if the null hypothesis is not true in the population. This is why specifying a null hypothesis is so important.

4.5.4 Significance level and Type I error

The threshold value, conventionally .05, is called the *significance level* of the test. If the null hypothesis—for instance, average media literacy is 5.5 in the population—is true, but we accidentally draw a sample with a mean well below 5.5, our p value is smaller than the significance level and we reject the null hypothesis.

We have to reject the null hypothesis in this situation; those are the rules of the game. However, in concluding that the null hypothesis is wrong, we make a mistake. We don't know and don't believe that we make this error, but we still do. This error is called a *Type I error*: rejecting a hypothesis that is actually true.

We cannot entirely avoid this error because samples can be very different from the population from which they are drawn, as we learned in Chapter 1. Thankfully, however, we know the probability that we make this error. This probability is the significance level.

You should understand the exact meaning of probabilities. A significance level of .05 allows five per cent of all possible samples to be so different from the population that we reject the null hypothesis even if it is true.

In other words, if we draw a lot of samples and decide on the null hypothesis for each sample, we would reject a true null hypothesis in five per cent of our decisions. So we have a five per cent chance of making a Type I error.

We decide on that probability when we select the confidence level of the test. We think that .05 is an acceptable probability for making this type of error. However, we do not know whether our sample belongs to the five per cent.

4.5.5 p Values in one-sided tests

In the example of average media literacy, we have only taken into account the left tail of the sampling distribution to calculate the p value. If we expect that average media literacy is below 5.5, we only consider the chance that the sample mean is below 5.5, so we do a one-sided test.

In this case, it makes sense to care only for the left tail of the sampling distribution. The entire probability of rejecting the null hypothesis while it is actually true is located in the left tail of the sampling distribution.

Of course, a one-sided test may also focus solely on the right tail of the sampling distribution. For example, if we hypothesize that exposure to advertisements increases brand awareness among consumers, we expect a positive correlation or regression coefficient. Our null hypothesis specifies that the correlation or regression coefficient is at most zero and we only reject it if the sample value is well above zero.

4.5.6 Significance level in two-sided tests

What is the situation in a two-sided test, for instance, if we hypothesize that average media literacy is not 5.5 in the population? Now, there are two ways in which we may reject the null hypothesis of 5.5 average media literacy: If the sample average media literacy is sufficiently larger than 5.5 or if it is sufficiently smaller. Each way of rejecting the null hypothesis has an associated probability if the null hypothesis is true (Type I error). Their sum is the significance level.

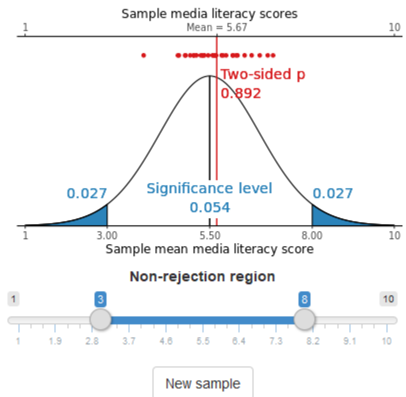


Figure 4.14: How do we obtain two-sided significance levels and p values?

1. In Figure 4.14, use the *Non-rejection region* slider to determine the rejection region for a two-sided test at 5% significance level. Hint: Click a handle and then use the left and right arrow keys on your keyboard to make tiny changes to the handle's value.

```
* A two-sided significance level of .05 has .025 probability in both the left
and right tail of the distribution. Move the left slider handle to 2.95 and the
right slider handle to 8.05 to obtain this situation.
* The two-sided rejection region with 5% significance level unites all sample
mean media literacy scores below 2.95 with those above 8.05.
```

2. What is the significance level of a one-sided test using the rejection area in the right tail that you have just determined? To check your answer, change one of the slider handles such that you obtain a one-sided rejection region and read out the p value.

```
* A one-sided test uses only one of the two tails of the sampling distribution,
the right tail in the current example. The probability of drawing a sample
with average media literacy above 8.05 is 2.5%. So this is the one-sided
significance level.
* If you want to get rid of the left tail in the rejection region, drag the left
slider handle to 1. The left tail probability is zero now, so the right tail
probability equals the total (one-sided) significance level.
* Note that this figure simplifies the situation a little. The tails of a normal
distribution never touch the horizontal axis, so there is a theoretical chance
even for scores below 1. However, media literacy is scored on a scale from 1
to 10, so scores below 1 do not make sense. As a consequence, the probability
of a score below 1 is set to zero. In this respect, the sampling distribution
deviates from the theoretical normal distribution. The latter does not exactly
capture the sampling distribution in this example.
```

3. When does Figure 4.14 indicate that a significance level is not applicable? Why is that so?

```
* The figure reports that a significance level is not applicabe if the two tails
do not contain the same probability but neither of them is zero.
* If neither probability is zero, we have a two-sided test. But in a two-sided
test, the total significance level is equally divided over the left and right
tails. The two probabilities, then, must be equal in a two-sided test.
```

4. When is the two-sided p value largest and when is a one-sided p value largest? Draw several samples and inspect the p values to check your answer.

```
* A two-sided p value is largest if the sample mean is closest to the
hypothesized population mean. The two-sided p value doubles the smallest of
the left-side or right-side p value.
* A one-sided p value is largest if the sample mean is on the end of the
distribution opposite to the tail containing the rejection region. A
right-sided test uses the right-sided p value, a left-sided test uses the
left-sided p value.
```

5. What is special about a significance level of .05?

```
Nothing is special about a significance level of .05. We could have used any
other cutoff value.
```

In a two-sided test, we divide the significance level by two and use half of the probability for the left tail and the other half for the right tail of the sampling distribution. If our significance level is .05, as it usually is, we use .025 as the maximum probability of finding a sample with a statistic that is so low that we would reject a null hypothesis that is true. We use the other .025 probability for drawing a sample with a statistic that is so high that we reject the null hypothesis. Together, the two .025 probabilities constitute the .050 significance level of our two-sided test.

4.5.7 p Values in two-sided tests

Just like the significance level, the p value in a two-sided test has to take into account that we may reject the null hypothesis in two different ways.

A p value gives us the probability of drawing a sample with the value for the sample statistic that we have found in our current sample or a more extreme value. In other words, it is the probability of drawing a sample with a statistic that is at least as distant from the hypothesized population value as our current sample result. In Figure 4.15, the sample mean is 5.15 and we have .042 probability to find a sample mean of 5.15 or less if the null hypothesis is true.

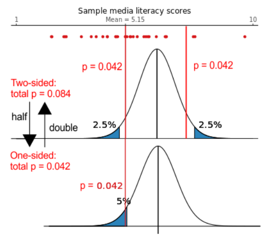
[htbp] 

Figure 4.15: Halve a two-sided p value to obtain a one-sided p value, double a one-sided p value to obtain a two-sided p value.

In a two-sided test, the distance can go in two directions: larger than the hypothesized value or smaller. As a consequence, the p value must cover samples at opposite sides of the sampling distribution. We should not only reckon with sample means that are less than 5.15 but also with sample means that are just as much larger than the hypothesized population value. So we have to take into account a probability of .042 for the right tail of the distribution as well in Figure 4.15. We can double the one-sided p value to obtain the two-sided p value. And we can halve a two-sided p value to obtain the one-sided p value.

You do not need to worry about this if your statistics software reports the type of p value that you need: one-sided or two-sided. If there is no choice between one- and two-sided tests, the reported p value is always correct.

However, if your software reports a one-sided p value but you need a two-sided p value, you have to double the reported p value. In contrast, if the software reports a two-sided p value while you need a one-sided p value, you have to divide the reported p value by two yourself because you do not want to take the probability in the other tail into account. Statistical software usually reports two-sided p values, so this situation may occur.

By the way, the test on the sample mean is statistically significant in the left-sided test in Figure 4.15 (bottom) but not in the two-sided test (top). The sample mean is in the rejection region in the left-sided test because the full 5 per cent significance level is located in the left tail of the distribution.

4.5.8 Unfounded one-sided null hypotheses

Be careful, however. If your left-sided test hypothesizes that average media literacy is below 5.5 but your sample mean is above 5.5, your left-sided test can never be significant. After all, your sample result is fully in line with the null hypothesis.

In this situation, the two-sided p value can be significant but you should never use the two-sided p value if your null hypothesis was one-sided. Changing your null hypothesis during the analysis, even from one-sided to two-sided, is cheating (see Section 4.9) because you must formulate the null hypothesis before you know the sample result. If you test the significance of null hypotheses, you have to live with the fact that you excluded outcomes from your one-sided test that perhaps should not have been excluded.

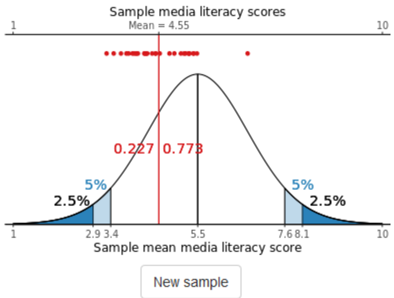


Figure 4.16: Is the test statistically significant?

1. Practice recognizing significant test results in Figure 4.16. Draw some samples and decide if a two-sided, right-sided, and left-sided test is statistically significant at 5% significance level.

```
* If the sample mean is in a dark blue tail, a two-sided test is statistically
significant.
* If the sample mean is in the light or dark blue tail at the right, a
right-sided test is statistically significant but a left-sided test is not.
* If the sample mean is in the light or dark blue tail at the left, a
left-sided test is statistically significant but a right-sided test is not.
```

2. In which situation is a one-sided test not statistically significant whereas a two-sided test is statistically significant at the 5% significance level?

```
* If the sample mean is in the dark blue tail at the right, a two-sided test
is statistically significant but a left-sided test is not.
* If the sample mean is in the dark blue tail at the left, a two-sided test is
statistically significant but a right-sided test is not.
```

4.6 Test Statistic and Degrees of Freedom

A theoretical probability distribution links sample outcomes such as a sample mean to probabilities by means of a *test statistic*. A test statistic is named after the theoretical probability distribution to which it belongs: z for the standard-normal or z distribution, t for the t distribution, F for the F distribution and, you guessed it, chi-squared for the chi-squared distribution.

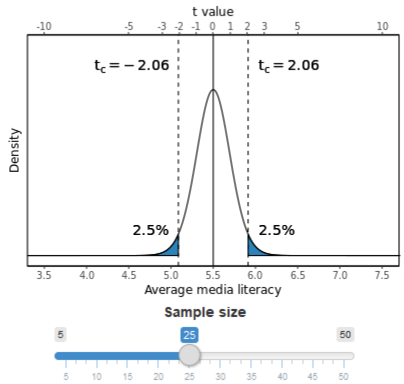


Figure 4.17: Sample size and critical values in a one-sample t test.

Figure 4.17 uses the t distribution to approximate the sampling distribution of average media literacy in a random sample of children.

1. What is the meaning of the coloured tails?

```
* The coloured tails represent the probabilities of drawing a sample with a
(sample) mean that differs a lot from the (true or hypothesized) population
mean. The (true or hypothesized) population mean is represented by the centre
of the distribution.
* In this case, the probabilities of the two coloured tails sum to 5%, so the
tails represent the top five per cent of samples that are most different from
the hypothesized or true population mean.
```

2. What is the meaning of t_c?

```
* t~c~ is the value of the test statistic t that separates the most unlikely
or most extreme samples from the most likely samples.
* It is called the critical value (of the test statistic).
```

3. Why does the distribution become more pointed when sample size increases?

```
* With a larger sample, we have more information, so we have more precise
results. More sample means are close to the true population mean.
* In technical terms, the variation of sample means decreases with larger
sample size; sample means are more alike. The standard deviation of the
sampling distribution is called the standard error. A distribution of larger
samples has a smaller standard error.
```

4. Is there a fixed relation between the t values and the values for average sample media literacy? Change the sample size to find the answer.

```
* No, there is not a fixed relation. A particular average media literacy
score, for example, 4.0 on the bottom scale, does not always correspond to the
same test statistic (t) value on the top scale. Just change sample size to see
```

```
this.
* The t statistic expresses the difference between the (true or hypothesized)
population value and the sample value in standard errors. It translates, for
example, the difference between 5.5 (population average of media literacy) and
4.0 (media literacy in our sample) into standard errors, for instance, 0.5,
which yields a test statstic value of 1.5 / 0.5 = 3.0.
* A larger sample yields a smaller standard error. As a result, we divide the
difference in raw scores, 1.5 in the example, by a smaller number, so the
resulting test statistic is larger.
```

5. What is the relation between sample size and critical t values?

```
* Smaller samples have slightly larger critical t values.
```

A test statistic is calculated from the sample statistic that we want to test, for instance, the sample proportion, mean, variance, or association, but it uses the null hypothesis as well. A test statistic more or less standardizes the difference between the sample statistic and the population value that we expect under the null hypothesis.

The exact formula and calculation of a test statistic is not important to us. Just note that the larger the difference between observation (sample outcome) and expectation (hypothesized population value), the more extreme the value of the test statistic, the less likely (lower p value) it is that we draw a sample with the observed outcome or an outcome even more different from the hypothesized value, and, finally, the more likely we are to reject the null hypothesis.

Actually, we reject the null hypothesis if the test statistic is in the *rejection region*. The value of the test statistic where the rejection region starts, is called the *critical value*. In Section 3.4, we learned that 1.96 is the critical value of z for a two-sided test at 5% significance level in a standard-normal distribution. In a z test, then, a sample z value above 1.96 or below -1.96 indicates a significant test result.

Except for the normal distribution, the probability distributions that we use depend on the degrees of freedom of the test. The degrees of freedom can depend on sample size, the number of groups that we compare, or the number of rows and columns in a contingency table. We don't need to worry about this.

The t distribution depends on the degrees of freedom of the test, for example. The degrees of freedom are determined by sample size: a larger sample yields slightly lower critical values in a t distribution. For samples that are not too small, however, the critical values of t are near 2. You may have noticed this in Figure 4.17.

APA6 requires us to report the degrees of freedom. If SPSS reports the degrees of freedom, usually in a column with the header "df". You should include it between brackets after the name of the test statistic, for instance: t (18) = 0.63. Note that the F test statistic has two degrees of freedom, both of which should be reported (separated by a comma), for example, F (2, 87) = 3.13.

4.7 Test Recipe and Rules for Reporting

Testing a null hypothesis consists of several steps, which are summarized below, much like a recipe in a cookbook.

1. Specify the statistical hypotheses.

In the first step, translate the research hypothesis into a null and alternative hypothesis. This requires choosing the right statistics for testing the research hypothesis (Section 4.2) and choosing between a one-sided or two-sided test if applicable (Section 4.4).

2. Select the significance level of the test.

Before we execute the test, we have to choose the maximum probability of rejecting the null hypothesis if it is actually true. This, as we know now, is the significance level. We almost always select .05 (5%) as the significance level. If we have a very large sample, e.g., several thousands of cases, we may select a lower significance level, for instance, 0.01. See Chapter 5 for more details.

3. Select how the sampling distribution is going to be created.

Are you going to use bootstrapping, an exact approach, or a theoretical probability distribution? Theoretical probability distributions are the most common choice and we focus on these, here. We have to know which theoretical probability distribution can be used for which test. If you are working with statistical software, you automatically select the correct probability distribution by selecting the correct test. For example, a test on the means of two independent samples in SPSS uses the t distribution.

4. Execute the test.

Let your statistical software calculate the p value of the test and/or the value of the test statistic. It is important that this step comes after the first three steps. The first three steps should be made without knowledge of the results in the sample.

5. Decide on the null hypothesis.

Reject the null hypothesis if the p value is lower than the significance level.

6. Report the test results.

The ultimate goal of the test is to increase our knowledge. To this end, we have to communicate our results both to fellow scientists and to the general reader who is interested in the subject of our research.

4.7.1 Reporting to fellow scientists

Fellow scientists need to be able to see the exact statistical test results. According to APA6, we should report the test statistic, the associated degrees of freedom (if any), the value of the test statistic, the p value of the test statistic, and the confidence interval (if any). APA6 requires a particular format for presenting statistical results and it demands that the results are included at the end of a sentence.

The statistical results for a t test on one mean, for example, would be:

t (67) = 2.73, p = .004, 95%CI[4.13, 4.87]

- The degrees of freedom are between parentheses directly after the name of the test statistic. Chi-squared tests add sample size to the degrees of freedom, for instance: chi-squared (12, N = 89) = 23.14, p = .027.
- The value of the test statistic is 2.73 in this example.
- The p value is .004. Note that we report all results with two decimal places except probabilities, which are reported with three decimals. We are usually interested in small probabilities—less than .05—so we need the third decimal here.
- The 95% confidence interval is 4.13 to 4.87, so with 95% confidence we state that the population mean is between 4.13 and 4.87.

Not all tests produce all results reported in the example above. For example, a z test does not have degrees of freedom and F or chi-squared tests do not have confidence intervals. Exact tests or bootstrap tests usually do not have a test statistic. Just report the items that your statistical software produces, and give them in the correct format.

4.7.2 Reporting to the general reader

For fellow scientists and especially for the general reader, it is important to read an interpretation of the results that clarifies both the subject of the test and the test results. Make sure that you tell your reader who or what the test is about:

- What is the population that you investigate?
- What are the variables?
- What are the relevant sample statistics?
- Which comparison(s) do you make?
- Are the results statistically significant and, if so, what are the estimates for the population?
- If the results are significant, how large are the differences or associations?

A test on one proportion, for instance, the proportion of all households reached by a television station, could be reported as follows:

> "The television station reaches significantly and substantially ($p = .61$) more than half of all households in Greece, $z = 4.01$, p < .001."

The interpretation of this test tells us the population ("all households in Greece"), the variable ("reaching a household") and the sample statistic of interest (p for proportion). It tells us that the result is statistically significant, which a fellow scientist can check with the reported p value.

Note that the actual p value is well below .001. If we would round it to three decimals, it would become .000. This suggests that the probability is zero but there is always some probability of rejecting the null hypothesis if it is true. For this reason, APA6 wants us to report p < .001 instead of p = .000.

Finally, the interpretation tells us that the difference from .5 is substantial. Sometimes, we can express the difference in a number, which is called the *effect size*, and give a more precise interpretation (see Chapter 5 for more information).

If you have the value of the test statistic but not the p value, you may report the significance level of the test instead of the p value. In this case, you either report "p < .05" if the test is significant or "n.s." if the test is not significant.

4.8 Relation Between Null-Hypothesis Test and Confidence Interval

The top of Figure 4.18 shows media literacy scores in a random sample of children and their average media literacy score (red). The hypothesized average media literacy in the population of children is shown on the bottom axis. The curve represents the sampling distribution if the null hypothesis is true. The coloured areas under the curve represent 2.5% of the total area each.

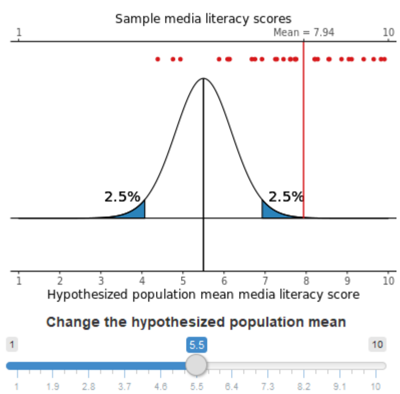


Figure 4.18: How does null hypothesis significance relate to confidence intervals?

1. Would you reject the current null hypothesis with this sample?

```
* If the sample mean (red line) is in the blue tails of
the sampling distribution, the null hypothesis must be
rejected.
* Note that the blue tails extend infinitely away from the
hypothesized value but the density becomes quickly so small
that you can't see the blue surface everywhere.
```

2. What are the lowest and highest sample means for which the null hypothesis is *not* rejected?

```
* If the sample mean (red line) is between the blue
tails, the null hypothesis is not rejected.
* So imagine that the inner boundaries of the tails extend
down to the x axis. At which mean media literacy scores
do these boundaries intersect with the x axis?
* In the initial situation with the hypothesized population
mean at 5.5, the boundary scores are just above four and
just below seven.
```

3. Why does the sampling distribution (curve) move horizontally if you change the hypothesized population mean?

```
* The curve moves because it represents the sampling
distribution if the null hypothesis is true. It assumes
that the true population mean is equal to the hypothesized
population mean.
* The population mean is the expected value of the
sampling distribution, so it is the centre of this
distribution. If we change the hypothesized mean, we
change the centre of the distribution, so it moves
to the left or right.
```

4. Use the slider to find the 95% confidence interval for the population mean given the sample mean.

```
* A confidence interval contains all (hypothetical)
values of the population parameter---here: the
population mean---for which we do not reject the
null hypothesis. So all hypothesized population means
that create a sampling distribution such that the sample
mean is between the critical boundaries of the tails are
part of the confidence interval.
* Change the slider such that the boundary of the right tail
coincides with the red line of the sample mean. This
hypothesized population value is the lower bound of the
95% confidence interval.
* Make the boundary of the left tail meet the sample mean
(red line): this population value is the upper bound
of the confidence interval.
```

Null-hypothesis testing and confidence-interval estimation are related. The long and the short of it is that a 95% confidence interval contains all null hypotheses that would *not* be rejected with the current sample at the 5% significance level, two-sided.

Remember that the probability involved in a 95% confidence interval is not the probability that the population value is within the interval. The population value is a fixed number, not a random variable with a probability distribution (at least, not in the approach to statistical inference that we follow here). For this reason, we say that we are 95% confident that the parameter lies within the 95% confidence interval but not that the population value is included in the interval with 95% probability.

The probability refers to the sample that we have drawn. If the population value would have a value within the interval, our sample result is among the 95% samples that are most plausible to be drawn. However, this is just another way of saying that a null hypothesis with this value for the parameter is sufficiently plausible (at a 5% significance level) considering the sample that we have drawn. So we do *not* reject the null hypothesis if it specifies any of the parameter values included in the 95% confidence interval.

4.8.1 Testing a null hypothesis with a confidence interval

It is easy to execute a null hypothesis test if you know the confidence interval. If the population value specified in the null hypothesis is within the confidence interval, do not reject the null hypothesis. Otherwise, reject the null hypothesis.

Let us imagine, for instance, that the 95% confidence interval for average media literacy among children is 4.11 to 4.87. Our initial null hypothesis states that the average is at least 5.5. An average of 5.5 is clearly outside this 95% confidence interval, so this sample rejects the null hypothesis at the 5% significance level.

We have already encountered this result in a previous section. Note, however, that our original test was one-sided but a confidence interval corresponds always to a two-sided test because it allows the parameter to be both smaller and larger than the sample value.

It is also clear that any null hypothesis specifying a population mean above 5.5 would be rejected. We could already infer that from our one-sided null hypothesis test. What we could not see there, however, is that a hypothesis specifying a population mean of 4.9 or 4.0 would also have been rejected in a two-sided test at the 5% level. A confidence interval gives more information than a single null hypothesis test because it shows us test results for a range of null hypotheses.

4.8.2 Testing a null hypothesis with bootstrapping

Using the confidence interval is the easiest and sometimes the only way of testing a null hypothesis if we create the sampling distribution with bootstrapping. For instance, we may use the median as the preferred measure of central tendency rather than the mean, if the distribution of scores is quite skewed and the sample is not very large. In this situation, a theoretical probability distribution for the sample median is not known, so we resort to bootstrapping.

Bootstrapping creates an empirical sampling distribution: a lot of samples with a median calculated for each sample. A confidence interval can be created from this sampling distribution (see Section 3.5.3). If our null hypothesis about the population median is included in the 95% confidence interval, we do not reject the null hypothesis. Otherwise, we reject it.

4.9 Capitalization on Chance

The relation between null hypothesis testing and confidence intervals may have given the impression that one could test a range of null hypotheses using just one sample and one confidence interval. For instance, we could simultaneously test the null hypotheses that average media literacy among children is 5.5, 4.5, or 3.5. Just check if these values are inside or outside the confidence interval and you are done?

This impression is wrong. The probabilities that we calculate using one sample assume that we only apply one test to the data. If we test the original null hypothesis that average media literacy is 5.5, we run a risk of five per cent to reject the null hypothesis if the null hypothesis is true.

If we apply a second test to the same sample, for example, testing the null hypothesis that average media literacy is 4.5, we again run this risk of five per cent. But the risk of rejecting at least one true null hypothesis increases dramatically if we do two tests. This risk is 9.75 per cent, namely one minus the risk of rejecting no true null hypothesis, which is 1 - .95 * .95 = .0975.

The situation becomes even worse if we do three or more tests on the same sample. The total level of rejecting at least one null hypothesis that is true increases well above the significance level that we want to maintain, namely five per cent.

The phenomenon that we are dealing with probabilities of making Type I errors that are higher (*inflated Type I errors*) than the significance level that we want to use, is referred to as *capitalization on chance*. Applying more than one test to the same data is one way to capitalize on chance. If you do a lot of tests on the same data, it is very rare not to find some statistically significant results even if all null hypotheses are true.

4.9.1 Capitalization on chance in post-hoc tests

This type of capitalization on chance may occur, for example, in an analysis of variance. To test the research hypothesis that the celebrity endorsing a fundraising campaign makes a difference to people's willingness to donate, we may organize an experiment using four versions of a video clip as the treatment, each clip featuring a different celebrity endorsing the campaign. This results in four groups of participants, each group having an average score on their willingness to donate. As a first step, we test the null hypothesis that all groups have equal population means using an F test (analysis of variance).

If this test is statistically significant, we reject the null hypothesis and conclude that at least two groups have different population means. The next question is: Which groups, that is, endorsement by which celebrities, display a different willingness to donate? To answer this question, we must do post-hoc t tests on pairs of groups. With four groups, we have six pairs of groups, so we have six t tests on independent means. The probability of rejecting at least one true null hypothesis of no difference is much higher than five per cent if we use a significance level of five per cent for each single t test.

4.9.2 Correcting for capitalization on chance

We can correct in several ways for this type of capitalization on chance; one such way is by applying the Bonferroni correction. This correction merely divides the significance level that we use for each test by the number of tests that we do. In our example, we do six t tests, so we divide the significance level of five per cent by six. The resulting significance level for each t test is .008. If a t test's p value is below .008, we reject the null hypothesis, but we do not reject it otherwise.

The Bonferroni correction is a rather coarse correction, which is not entirely accurate. However, it has a simple logic that directly links to the problem of capitalization on chance. Therefore, it is a good technique to help understand the problem, which is the main goal we want to attain, here. We will skip better but more complicated alternatives to Bonferroni correction.

Note that we need not apply a correction if we specify a hypothesis beforehand about the two groups that we expect to differ. In the example of celebrity endorsement, we would not have to apply the Bonferroni correction to the t test on the mean difference between participants confronted to Celebrity A and Celebrity C if we had hypothesized that the willingness to donate differs here. Of course, we could have skipped the analysis of variance and gone straight to the t test with such a hypothesis.

4.9.3 Specifying hypotheses afterwards

Capitalization on chance occurs if we apply different tests to the same variables in the same sample. This occurs in exploratory research in which we do not specify hypotheses beforehand but try out different predictors or different outcome variables.

It occurs more strongly if we first have a look at our sample data and then formulate an hypothesis. Knowing the sample outcome, it is easy to specify a null hypothesis that will be rejected. This is plain cheating and it should be avoided at all times.

4.9.4 Advantages of using confidence intervals

If we cannot use the confidence interval to test a range of null hypotheses, what, then, is the added value of a confidence interval over a single null hypothesis test? The confidence interval tells us the plausible values of the parameter rather than whether or not it is one particular value. Putting it simply, we just know more from a confidence interval than from a null hypothesis test.

We could use the additional information, if we were to do a new null hypothesis test on a new sample. We would have a better idea of plausible null hypotheses. We could, for instance, test whether media literacy is 4.49, the middle of the confidence interval in our previous research project. A confidence interval helps us to be more specific in the next study addressing our topic, whether this study is carried out by ourselves or by our colleagues.

4.10 Test Your Understanding

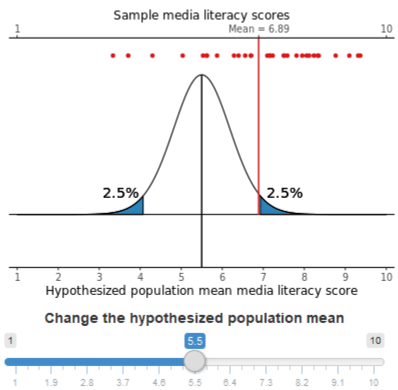


Figure 4.19: Testing null hypotheses.

Figure 4.19 displays a random sample of media literacy scores (red) and a sampling distribution if the null hypothesis is true.

1. What are the null and alternative hypotheses in Figure 4.19? Is the null hypothesis a nil hypothesis here?
2. What represents the p value of the sample mean that we have found in Figure 4.19? Does the p value depend on the type of null hypothesis: one-sided or two-sided?
3. What represents the significance level and rejection region in Figure 4.19?
4. Is the test statistically significant? How do you decide?
5. What happens if you change the hypothesized population mean? Check your answer by using the slider.
6. Is it OK to change your null hypothesis when you know your sample mean? Why is it OK or not OK?

4.11 Take-Home Points

- We use a statistical test if we want to decide on a null hypothesis: reject or not reject? Usually, this boils down to the question: "Is there or is there not an effect (difference, association) in the population?"

- The decision rules should be specified beforehand: decide on the direction of the test (one-sided or two-sided) and the significance level.
- The null and alternative hypotheses always concern a population statistic. Together they cover all possible outcomes for the statistic. The null hypothesis always specifies one (boundary) value for the population statistic.
- We reject the null hypothesis if a test is statistically significant. This means that the probability of drawing a sample with the current or a more extreme outcome (even more inconsistent with the null hypothesis) for the test statistic is below the significance level.
- The 95% confidence interval includes all null hypotheses that would *not* be rejected in a two-sided test at 5% significance level. It contains the population values that are not sufficiently contradicted by the data.
- The calculated p value is only correct if the same data is used for no more than one null hypothesis test and the null hypothesis was formulated beforehand.
- If the same data is used for more null hypotheses tests, the probability of a Type I error increases. We obtain too many significant results, which is called capitalization on chance.

Chapter 5

Which Sample Size Do I Need? Power!

Key concepts: minimum sample size, (unstandardized) effect size, practically significant, standardized effect size, Cohen's d (for means), Type I error, Type II error, test power.

Summary

How large should my sample be?

At the start of a quantitative research project, we are confronted with a seemingly simple practical question: How large should our sample be? In some cases, the statistical test that we plan to use gives us rules of thumb for the minimum size that we need for this test.

This may tell us the minimum sample size but not necessarily the optimal sample size. Even if we can apply the statistical test technically, sample size need not be sufficient for the test to signal the population differences or associations, for short, the effect sizes, that we are interested in.

If we want to know the minimum sample size that we need to signal important effects in our data, things become rather complicated. We have to decide on the size of effects that we deem interesting. We also have to decide on the minimum probability that the statistical test will actually be significant if the true effect in the population is of this size.

This probability is the power of a test: the probability to reject a null hypothesis of no effect if the effect in the population is of a size interesting to us. If we do not reject a false null hypothesis, we make a Type II error.

Thinking about sample size thus confronts us with a problem that we have hitherto neglected, namely the problem of not rejecting a false null hypothesis. This problem is very important if the null hypothesis represents our research hypothesis. If the null hypothesis represents

Table 5.1: Rules of thumb for minimum sample sizes.

Distribution	Sample statistic	Minimum sample size
Binomial distribution	proportion	-
(Standard) normal distribution	proportion	>= 5 divided by test proportion (<= .5)
(Standard) normal distribution	one or two means	$>$ 100
t distribution	one or two means	$>$ 30
t distribution	(Spearman) rank correlation coefficient	$>$ 30
t distribution	regression coefficient	20+ per predictor variable
F distribution	3+ means	all groups are more or less of equal size
chi-squared distribution	row or cell frequencies	expected frequency >= 1 and 80% >= 5

our research hypothesis, our expectations are confirmed if we do *not* succeed in rejecting the null hypothesis.

However, if we do *not* reject the null hypothesis, we cannot make a Type I error, namely rejecting a false null hypothesis. As a consequence, the significance level of our test, which is the maximum probability of making a Type I error, is meaningless. We must know the probability of *not* rejecting a false null hypothesis—the power of the test—to express our confidence that our research hypothesis is true.

This chapter reviews concepts that are central to understanding what it means that a test is statistically significant: sampling distribution, hypotheses, statistical significance, and Type I error. It adds concepts that we need to interpret a test that is *not* statistically significant: effect size, practical significance, Type II error, and test power.

5.1 Sample Size and Test Requirements

Table 2.2 in Chapter 2 shows the conditions that must be satisfied if we want to use a theoretical probability distribution to approximate a sampling distribution. Only if the conditions are met, the theoretical probability distribution resembles the sampling distribution sufficiently for using the theoretical probability distribution.

Conditions often include sample size (Table 5.1 reproduces the sie requirements from Table 2.2). If you plan to do a t-test, either on its own or in post-hoc tests after analysis of variance, each group should contain more than thirty cases. So if you plan on doing t-tests, recruit more than thirty participants for each experimental group or more than thirty respondents for each group in your survey and you are fine. Well, if you have to reckon with non-response, that is, sampled participants or respondents unwilling to participate in your research, you should recruit more participants or respondents to have more than thirty observations in the end.

Chi-squared tests require a minimum of five expected frequencies per category in a frequency distribution or cell in a contingency table. Your sample size should be at least the number of categories or cells times five to come even near this requirement. Regression analysis requires at least 20 cases per predictor variable in the regression model.

The variation of sample size across groups is important to the analysis of variance. If the number of cases is more or less the same across all groups, we need not worry about the variances of the outcome variable in the population for the groups. To be on the safe side, then, it is recommended to design your sampling strategy in such a way that you end up with more or less equal group sizes if you plan to use analysis of variance (ANOVA).

5.2 Effect Size

We have learned that larger samples have smaller standard errors (Section 3.3.1). Smaller standard errors yield (absolutely) larger test statistic values and larger test statistics have smaller p values. In other words, a test on a larger sample is more often statistically significant.

A larger sample offers more precision, so the difference between our sample outcome and the hypothesized value is more often sufficient to reject the null hypothesis. For example, we would reject the null hypothesis that average candy weight is 2.8 gram in the population if the average weight in our sample bag is 2.75 gram and our sample bag is very big. But we may not reject this null hypothesis if we have the same outcome in a small sample bag.

Of course, the size of the difference between our sample outcome and the hypothesized population value matters as well. This difference is called *effect size*. If average candy weight in our sample bag deviates more from the average weight that we expect according to the null hypothesis, we are more likely to reject the null hypothesis. If we think of our statistical test as a security metal detector, our test will pick up larger amounts of metal more easily.

Figure 5.1: Security metal detector

The probability of rejecting a null hypothesis, then, depends both on sample size and effect size, that is, the difference between what we expect null hypothesis) and what we find (sample outcome).

- A larger sample size makes a statistical test more sensitive. The test will pick up (be statistically significant for) smaller effect sizes.

- A larger effect size is more easily picked up by a statistical test. Larger effect sizes yield statistically significant results more easily, so they require smaller samples.

Deciding on our sample size, we should ask ourselves this question: What effect size should produce a significant test result? In the security metal detector example, at what minimum quantity of metal should the alert sound? To answer this question, we should consider the practical aims and context of our research.

5.2.1 Practical significance

Investigating the effects of a new medicine on a person's health, we may require some minimum level of health improvement to make the new medicine worthwhile medically or economically. If a particular level of improvement is clinically important, it is *practically significant*.

If we have decided on a minimum level of improvement that is relevant to us, we want our test to be statistically significant if the average true health improvement in the population is at least of this size. We want to reject the null hypothesis of no improvement in this situation.

For media interventions such as health, political, or advertisement campaigns, one could think of a minimum change of attitude affected by the campaign in relation to campaign costs. A choice between different campaigns could be based on their efficiency in terms of attitudinal change per cost unit.

Note the important difference between practical significance and statistical significance. Practical significance is what we are interested in. If the new medicine is sufficiently effective, we want our statistical test to signal it. In the security metal detector example: If a person carries too much metal, we want the detector to signal it.

Statistical significance is just a tool that we use to signal practically significant effects. Statistical significance is not meaningful in itself. For example, we do not want to have a security detector responding to a minimal quantity of metal in a person's dental filling. Statistical significance is important only if it signals practical significance. We will return to this topic in Chapter 6.

5.2.2 Unstandardized and standardized effect sizes

The difference between our sample outcome and the hypothesized value is the *unstandardized effect size*. If we test a mean, the unstandardized effect size is just the difference between our sample mean and the hypothesized population mean. For example, if we hypothesized that average candy weight in the population is 2.8 gram and we find an average candy weight in our sample bag of 2.75 gram, the unstandardized effect size is -0.05 gram.

Unstandardized effect sizes depend on the scale on which we measure the sample outcome. The unstandardized effect size of average candy weight changes if we measure candy weight in grams, micro grams, kilograms, or ounces. Of course, changing the scale does not affect the meaning of the effect size but the number that we are looking at is very different: 0.05 gram, 50 micro gram, 0.00005 kilo, or 0.00176 ounce. The unstandardized effect size value, then, does not tell us whether the effect size is large or small.

5.2.3 Cohen's d

In scientific research, we rarely have precise norms for differences that are practically significant and differences that are not. Instead, we tend to think of small and large effects as differences that are large or small in comparison to the scores that we usually encounter.

If candy weights vary a lot, we will not be very impressed by a relatively small difference between observed and expected (hypothesized) average candy weight. In contrast, if candy weight is quite constant, a small average difference is important.

For this reason, standardized effect sizes for sample means divide the difference between the sample mean and the hypothesized population mean by the standard deviation in the sample. Thus, we take into account the variation in scores. This standardized effect size for tests on means is known as *Cohen's d*.

The sample outcome can be a single mean, for instance the average weight of candies, but it can also be the difference between two means, for example, the difference between average weight of yellow candies and average weight of red candies. In the latter case, the difference is divided by a combined (*pooled*) standard deviation for yellow and red candy weight.

The direction of an effect is not relevant to effect size. For example, we do not care whether the yellow candies or the red candies are on average heavier. For this reason, Cohen's d is always positive. If you obtain a negative result, just drop the minus sign.

Using an inventory of published results of tests on one or two means, Cohen (1969) proposed rules of thumb for standardized effect sizes:

- 0.2: weak effect,
- 0.5: moderate effect,
- 0.8: strong effect.

Note that Cohen's d can take values above one. These are to be considered strong effects.

5.2.4 How to calculate Cohen's d from SPSS output

5.2.4.1 Instructions

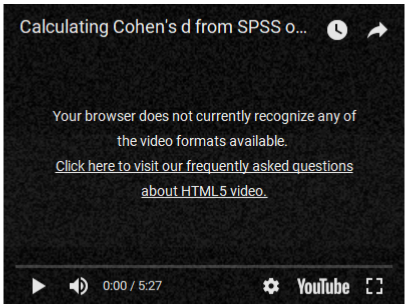


Figure 5.2: Calculating Cohen's d from SPSS output.

5.2.4.2 Exercises

1. Open data set voters.sav that contains information about the age and attitude towards immigration among a random sample of voters. What are the unstandardized and standardized effect sizes if the hypothesized average attitude towards immigrants in the population is 6.0?

```
* Execute a one-sample t test with 6.0 as test value.
* The reported mean difference is (-)0.5; this is the unstandardized effect
size.
* To obtain the standardized effect size (Cohen's d in this case), divide the
unstandardized effect size by the standard deviation of the variable (here:
attitude towards immigration), which is reported to be 1.939 Cohen's d = 0.5 /
1.939 = 0.26. This is a weak effect.
```

2. What are the effect sizes if the null hypothesis states that the average attitude towards immigrants in the population is at least 6.0? And what if it states that average attitude is at most 6.0?

```
* In a one-sided test, we take the boundary value (here: 6.0) as the value
against which we test. It makes sense to use this value also for calculating
effect size. The unstandardized and standardized effect sizes, then, are the
same as in Question 1.
* One could argue, however, that there is no effect, hence zero effect size,
in a right-sided test here. If we assume for whatever reasons that the
population average cannot be below 6.0, a sample average below 6.0 such as 5.5
would have to be due to sampling error. It cannot represent a true effect if
the population average can only be above 6.0.
```

3. What are the unstandardized and standardized effect sizes of a test in which we

compare the attitude towards immigrants of young voters to the attitude of old voters? Again, use data set voters.sav.

```
* Execute an independent-samples t test with groups defined by the variable
age_group.
* The reported mean difference is (-)0.718; this is the unstandardized effect
size.
* We approximate the standardized effect size (Cohen's d in this case) by
dividing twice the t value by the square root of the degrees of freedom. Note
that the F test on equality of population variances is statistically
significant (p = .029), so we do not assume equalx` variances and we use the
bottom row from the results table.
* Cohen's d = 2 * 1.602 / sqrt(30.009) = 3.204 / 5.478 = 0.58. This is a
moderate effect.
* Note that the sample of young voters is too small to conduct a t test. This,
however, is not consequential to Cohen's d because a minimum sample size is
required for a reliable p value but not for a t value.
```

5.2.5 Association as effect size

Measures of association such as Pearson's product-moment correlation coefficient or Spearman's rank correlation coefficient express effect size if the null hypothesis expects no correlation in the population. If zero correlation is expected, a correlation coefficient calculated for the sample expresses the difference between what is observed (sample correlation) and what is expected (zero correlation in the population).

Effect size is also zero according to the standard null hypotheses used for tests on the regression coefficient (b), R^2 for the regression model, and eta^2 for analysis of variance. As a result, we can use the standardized regression coefficient (Beta in SPSS and b^* according to APA6), R^2, and eta^2 as standardized effect sizes.

Because they are standardized, we can interpret their effect sizes using rules of thumb, e.g., an association between 0 and .10 is interpreted as no or a very weak association, between .10 and .30 is weak, between .30 and .50 is moderate, .50 to .80 is strong, and .80 to 1.00 is very strong, while exactly 1.00 is a perfect association (if 1.00 is the maximum value). Note that we ignore the sign (plus or minus) of the effect if we interpret its size.

5.2.6 Effect size and sample size

We can use standardized effect size to express the effects that we are interested in without caring about the precise size of differences. We merely have to choose whether small, moderate, or large effects are of practical interest to us. Preferably, we know from previous research whether small, moderate, or large effects are common in our type of research. If moderate or large effects are rare, we should use a sample size that allows detecting small effects. In contrast, when large effects occur frequently, we can do with a smaller sample that may miss small effects.

If we know the effect size in the sample for which we want statistically significant results, we can figure out the minimum sample size for which the test statistic is statistically significant.

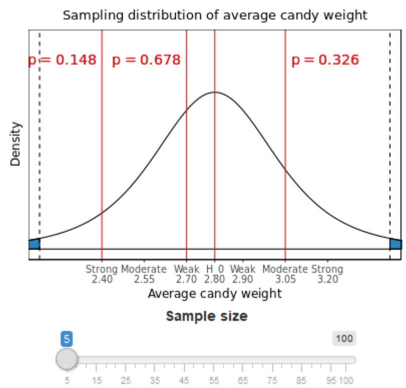


Figure 5.3: What is the minimum sample size required for a significant test result if the sample mean has a particular effect size?

1. Use the slider in Figure 5.3 to find the minimum sample size that we need for a statistically significant test result for each of the three efect sizes represented by red lines.

```
* We need a sample of minimum size 9 to have a statistically significant result
for a sample with average weight 2.40 grams, which represents a strong effect.
With 9 observations in the sample, the red line of a strong effect is in the
2.5% tail of the sampling distribution. The p value is below .05.
* For a moderate effect, we need a sample of at least 18 observations.
* For a weak effect, we need no less than 99 observations.
* By the way, some of these  sample sizes are too small for using the t
distribution as approximation of the sampling distribution. Here, the rule of
thumb (31 cases) has precedence.
```

2. What is the meaning of the p values and why do they decrease if we increase sample size?

```
* The p values give the (two-sided) significance of a sample with average candy
weight equal to the value on the horizontal axis. For example, the leftmost red
line represents a strong effect for a sample with average weight 2.40 grams
under the null hypothesis that average candy weight is 2.8 in the population.
* The larger the sample, the smaller the standard error, the narrower the
sampling distribution, the smaller the probability of drawing a sample with an
average candy weight that differs from the hypothesized value purely by
chance.
```

3. Compare the p values to the blue tails. Is there something wrong in this app?

```
* There is nothing wrong in this app. The p values are two-sided but each tail
represents only half of the significance level. As a consequence, the p value
is twice the surface of the tail to its right or to its left.
```

Effect size as well as test statistics reflect the difference between what we expect according to the null hypothesis and what we observe in our sample. As a consequence, effect size indicators and test statistics are related. In some cases, such as Cohen's d, the relation between effect size and test statistic is very simple.

The test statistic t for a t test on one mean is equal to Cohen's d times the square root of sample size. Here, the only difference between the two is sample size! Sample size influences the test statistic—the larger the sample, the larger the test statistic—but it does not affect effect size. This is one reason why effect size is more interesting than test statistics and their p values.

5.3 Hypothetical World Versus Imaginary True World

In the preceding paragraphs, we determined sample size using the effect size that we expect to find in our sample. We should realize, however, that we are interested in the effect size in the population. The 'true' effect size, so to speak.

The effect of a new medicine or a media campaign in our sample is not important but the effect in the population is. This complicates the calculation of our sample size. Instead of using the effect size in our (future!) sample, we must use the effect size in the population.

5.3.1 Imagining a population with a small effect

Our null hypothesis states that average candy weight in the population is 2.8 grams. Let us decide that a small effect size is practically significant. We can think now of a population that could be the true population if the effect size is small. For example, a population in which average candy weight is 2.9 grams (and the standard deviation is 0.5).

We do not know whether average candy weight is 2.9 grams in the true population. So we may regard this as another hypothesis. Let us call this the alternative hypothesis H_1. Note that this is not an ordinary alternative hypothesis because it does not include all outcomes not covered by the null hypothesis (H_0). Instead, it represents only one value, which is an important value to us because it represents a population with a small but interesting effect size.

Our habit of formulating a null hypothesis and an alternative hypothesis for all situations not covered by the null hypothesis is generally attributed to the statistician R.A. Fisher. This, however, is not entirely correct (see, e.g., Halpin & Stam, 2006). Fisher introduced the concept of a null hypothesis (Ronald Aylmer Fisher, 1935: 18) but not the concept of an alternative hypothesis. The statisticians Jerzy Neyman and Egon Pearson introduced the idea of working with two or more hypotheses. But the two hypotheses do not cover all possible population values and they were usually not called a null and alternative hypothesis. They specify two or more different population values. A statistical test is used to determine which of the hypotheses fits the sample best. (J. Neyman & Pearson, 1933)

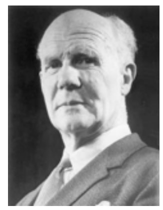


Figure 5.4: Egon Pearson.

Figure 5.5 illustrates this situation. The top graph represents the sampling distribution according to our null hypothesis. This sampling distribution is derived from our hypothetical population in which there is no effect, that is, our null hypothesis is true for this population. In our current example, average candy weight is 2.8 grams in this hypothetical population.

The bottom graph represents the sampling distribution for an imaginary population with a small effect size. Here, the alternative hypothesis is true, for instance, average candy weight is 2.9 grams, which is a bit higher than in the hypothetical population.

By the way, average population candy weights are not depicted in the graphs but you should know by now that the average in a normal or t distribution is situated at the top of the bell shape and that the average of a sample mean sampling distribution is equal to the population average because a sample mean is an unbiased estimator.

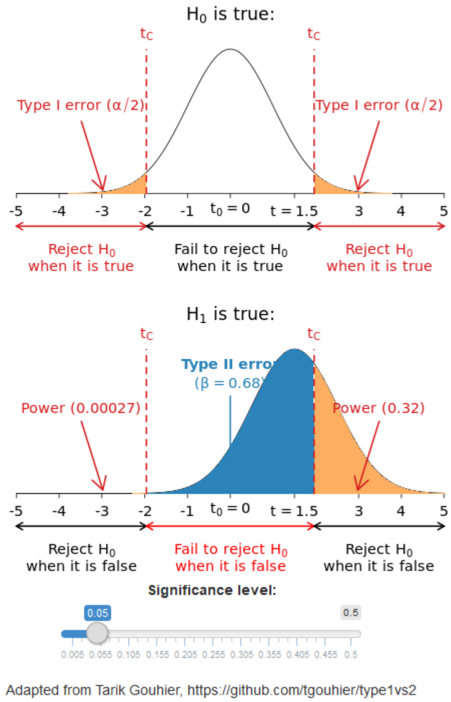


Figure 5.5: Simulation of Type I and Type II error.

Before reading on, try to make sense of the two graphs in Figure 5.5 and how they relate to each other:

1. What is the relation between significance level and Type I error? Formulate your answer very precisely: Details matter now! Check your answer by changing the significance level.

```
* With a higher significance level (alpha goes up), we take a larger risk to
reject the null hypothesis if it is true. The probability of a Type I error,
which we make if we reject a true null hypothesis, goes up.
* The orange tails in the top graph become bigger and the critical values move
towards zero.
```

2. What exactly is Type II error?

```
* We make a Type II error if we do not reject a false null hypothesis.
* The situations in which we make a Type II error are depicted by the red
double-sided arrow in the bottom graph.
```

3. How does the probability of a Type II error relate to the probability of a Type I error? Try to explain the relation in your own words.

```
* If the probability of a Type I error goes up, the probability of a Type II
error goes down, and the other way around.
* With a higher significance level (alpha goes up), we take a larger risk to
reject the null hypothesis if it is true. We have a larger probability of
making a Type I error, which is depicted by the orange tails in the top graph.
* If we take a larger risk to reject a true null hypothesis, we reject the
null hypothesis more easily.
* If we reject the null hypothesis more easily, we are more likely to reject
the null hypothesis if it is not true. The probability to reject a false null
hypothesis is the power of the test. This probability is expressed by the
orange tail in the bottom graph.
* Note that the critical t values move closer to zero both in the top and
bottom graph if we take a larger risk to reject a false null hypothesis. The
critical t values also demarcate the blue area in the bottom graph. Moving
towards zero, they decrease the size of the blue area.
```

5.3.2 Type I error

We have two populations, a hypothetical population and an imaginary true population. Once we have drawn our sample, we only deal with the hypothetical population, for instance, the top graph in Figure 5.5, as we have done in all preceding chapters.

Acting as if the null hypothesis is true, we determine how (un)likely the sample is that we have drawn. If it is very unlikely, we have a p value below the significance level and we reject the null hypothesis. We say: If the null hypothesis was true, our sample would be too unlikely, so we reject the null hypothesis.

We may be wrong. Perhaps the null hypothesis is actually true and we were just very unfortunate to draw a sample that is very different from the population. If so, we make a Type I error (see Section 4.5.4). The probability that we will make this error is equal to the significance level, which is usually set to .05.

5.3.3 The world of the researcher

This is what we are doing once we have the sample. Let us call this the world of the researcher. At present, we have not yet stepped into the world of the researcher because we are still thinking about the size of the sample that we are going to draw.

We can experiment a bit and that is what we do if we ask ourselves: What is going to happen to our statistical test if the true population from which we draw our sample has average candy weight that is a bit higher (small effect) than candy weight according to our null hypothesis?

Table 5.2: Error types and their probabilities.

	Null is true	Null is false
Null is rejected	Type I error, Significance level (alpha)	No error, Power (1 - beta)
Null is not rejected	No error, (1 - alpha)	Type II error, (beta)

5.3.4 The alternative world of a small effect

If we actually sample from this imaginary true population, the bottom graph in Figure 5.5 represents our true sampling distribution. It shows us the true probabilities (areas under the curve) of drawing a sample with a particular minimum or maximum value for the test statistic t. These are the probabilities of our sample if there is a small effect in the population.

Now that we know the true sampling distribution if there is a small effect in the population, we can foresee what is going to happen when we enter the world of the researcher. The researcher is going to use the test values of the top graph to decide on the null hypothesis. If the sample t value is between, say, plus and minus two (the critical values, t_c in Figure 5.5), the researcher is not going to reject the null hypothesis.

5.3.5 Type II error

If there is a (small) effect in the population, the null hypothesis is not true. For example, average candy weight is not 2.8 gram, it is 2.9 gram. If our sample mean is close to 2.8 gram, we may not reject the null hypothesis even if it is not true. This is a *Type II error*: not rejecting a false null hypothesis.

The probability that we make a Type II error if there is a small effect is expressed by the blue section in the bottom graph. It is usually denoted by the Greek letter beta (β). The blue section represents the probability of drawing a sample from this population with a small effect size that has a t value that is NOT in the rejection region, so the null hypothesis is NOT rejected. See the top graph.

Table 5.2 summarizes the four possible situations that may arise if we test a null hypothesis. The null hypothesis may be true or false and we may or may not reject the null hypothesis.

5.3.6 Power of the test

The probability of NOT making a Type II error is called the *power of the test.* It is of course equal to one minus the probability of making a Type II error, that is, 1 - β. The power of the test is represented by the orange sections in the bottom graph. They represent the probability of getting a sample t value that makes the researcher reject the null hypothesis. So a false null hypothesis is rejected and we do not make an error.

Note that we can reject the null hypothesis in two ways: If our sample happens to have a test statistic that is much higher than expected under the null hypothesis or if it is much lower. In the example above, the imagined true population has a mean that is higher than the hypothesized population mean. The bulk of the power of the test therefore is in the right tail of the bottom graph.

However, a little bit of power is situated in the left tail. It is so small that we usually cannot see the orange section in the left tail of the bottom graph but the displayed probability shows that it is there. This is a bit strange because one could say that we reject the null hypothesis for the wrong reason: We think our null hypothesis is too high whereas it actually is too low.

The probability that this happens is very small and usually negligible. In the interactive content, you may encounter power values in the left tail that are so small that they have to be written in scientific notation, e.g., 1.0E-10, which means 1 at the tenth decimal place: 0.0000000001. Anyway, the important thing is that we reject a false null hypothesis even if it is for the 'wrong' reason. Rejecting a false null hypothesis, our conclusion is not erroneous.

5.4 Sample Size, Effect Size, and Power

Finally, after all these sections, we can answer the question that opened this chapter: How large should my sample be? To answer this question, we must consider the significance level, effect size, and test power.

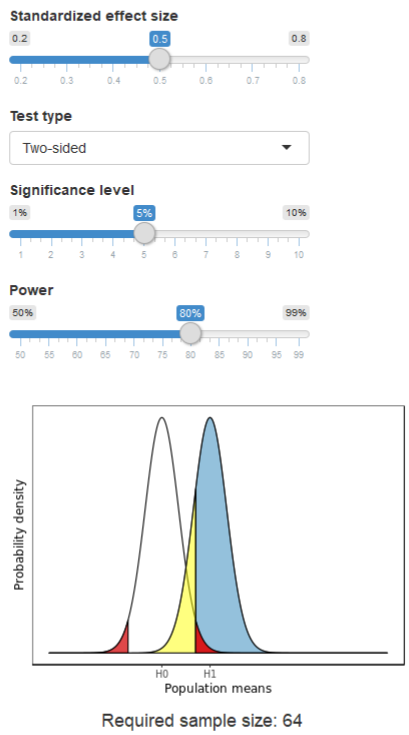


Figure 5.6: How does sample size depend on test power, significance, and effect size?

1. Figure 5.6 shows the sampling distributions of the sample mean under the null hypothesis (H_0, left-hand curve) and under the assumed true value of the population mean (H_1, right-hand curve). Explain the meaning of the red, yellow, and blue surfaces in the graph.

```
* The red areas represent the significance level of the null hypothesis test,
namely the total probability of rejecting the null hypothesis if it is true.
* The yellow area gives the probability that we draw a sample from a population
with true mean H~1~ which does not reject the null hypothesis. With such
samples, we do not reject the null hypothesis, which is false, so we make a
Type II error. The yellow area is the probability of making a Type II error.
* The blue area is the probability that we draw a sample from a population with
true mean H~1~ that rejects the null hypothesis, which is false. It is the
probability that we reject a false null hypothesis, which is the power of the
test.
```

2. What happens to the red, yellow, and blue surfaces if you set power to 50%? And what happens to the minimum sample size? Adjust the power slider to check your answers.

```
* If we set power to 50%, the yellow and blue areas each should occupy half of
the sampling distribution that is centred at H~1~.
* A smaller power implies that we run a larger risk of not rejecting a false
null hypothesis, so we need a smaller sample.
```

3. Do we need a smaller or a larger sample to achieve the specified test power for a larger effect size? Move the effect size slider to check your answer.

```
* With larger effect size, the difference between the true population mean
(H~1~) and the population mean according to the null hypothesis (H~0~) is
larger, and it will be more easily spotted in a smaller sample. In other words,
we have higher power with a larger effect size. We need a smaller sample to
achieve the original test power.
```

4. Do we need a smaller or a larger sample to achieve the specified test power for a one-sided test instead of a two-sided test? Change the test type to check your answer.

```
* With a one-sided test, the total significance level (red area) is situated in
one tail. If this tail is at the right side, namely, at the side of the true
population mean, we are more likely to reject the null hypothesis with a sample
from the true population (the blue area becomes larger), so test power
increases if we keep the same sample size. To achieve the original test power,
we may draw a smaller sample.
```

5. Do we need a smaller or a larger sample to achieve the specified test power for a higher significance level? Move the significance level slider to check your answer.

```
* A higher significance level, for instance, 10% instead 5%, has the same effect
of a change from a two-sided test to a one-sided test. We are more likely to
reject the null hypothesis, so we are more likely to reject a false null
hypothesis, so we have higher test power. To achieve the original test power,
we may draw a smaller sample.
```

6. Why do the sampling distributions become wider if we increase effect size or significance level, or if we decrease test power?

```
* With larger effect size or significance level and with smaller test power,
the (required) sample size decreases. Lower sample sizes produce larger
standard errors, which produce a wider distribution because sample results are
on average further away from the mean of the distribution.
```

Sample size, statistical significance, effect size, and test power are related. To determine the size of your sample, you have three buttons that you should adjust simultaneously. Statistical significance is the easiest button to decide on; we usually leave the significance level at .05. We do not select a smaller value because it will reduce the power of the test (with the same sample and effect size) as you may have noticed in one of the figures in the preceding section.

For effect size, we have to choose among a small, moderate, or large effect. Previous results of research similar to our research project can help us decide whether we have to reckon with small effect sizes (need a larger sample) or not. If we have a concrete number for the (standardized) minimum effect size that is of practical significance, we can use that number.

For power, the conventional rule of thumb is that we like to have at least 80% probability of rejecting a false null hypothesis. You may note that the probability of NOT rejecting a true null hypothesis is higher: .95. After all, it is one minus the probability of rejecting a true null hypothesis (Type I error), which is the significance level.

Power is set to a lower level because the null hypothesis is usually assumed to reflect our best knowledge about the world. From this perspective, we are keener on avoiding the error of falsely rejecting the null hypothesis (our current best knowledge) than falsely accepting it. This approach, however, is not without criticisms as we will discuss in Chapter 6. Anyway, if you want to raise the power to the same level of .95, you can do so; it will require a larger sample.

Unfortunately, test power receives little attention in several software packages for statistical analysis. Using power and effect size to calculate the required sample size is usually not provided in the package. To calculate sample size, we need dedicated software, for example GPower.

5.4.1 So how do we determine sample size?

All in all, using effect size and test power to determine the size of the sample requires several decisions on the part of the researcher. It can be difficult to specify the effect size that we should expect or that is practically relevant. If there is little prior research comparable to our new project, we cannot reasonably specify an effect size and calculate sample size.

Of course, it is important to ensure that our sample meets the requirements of the tests that we want to specify (Section 5.1). In practice, researchers often go well beyond this minimum. They try to collect as large a sample as is feasible just to be on the safe side.

Does this mean that all we have learned about effect size and test power is useless? Certainly not. First of all, we should have learned that effect size is more important than statistical significance because effect size relates to practical significance.

Second, test power and Type II errors are important in situations in which we do not reject the null hypothesis. Then, we should calculate test power to get an impression of our confidence in the result (see the next section). Is our test of sufficient power to yield significant results if there is an effect in the population?

5.5 Research Hypothesis as Null Hypothesis

As noted before (Section 4.3), the research hypothesis usually is the alternative hypothesis. We expect something to change, to be(come) different rather than be or stay the same. We expect an association to be present rather than absent.

In this situation, rejection of the null hypothesis supports our alternative hypothesis, hence our research hypothesis, so we are glad if we reject the null hypothesis. Of course, we know that we can be wrong. Our null hypothesis may still be true even if the probability of drawing a sample like the one we have drawn is so small that we have to reject the null hypothesis. This is a Type I error.

Fortunately, we know the probability of making this error because it is the significance level that we have chosen, five per cent usually. We can live with this probability of making an error if we reject the null hypothesis. So we are doubly glad: We found support for our research hypothesis *and* we know how confident we are about this support.

What if our research hypothesis is our null hypothesis? For example, we have a specific idea of average candy weight in the population from previous research or from specifications by the candy factory. If we want to test whether the candies have the hypothesized average weight, our research hypothesis would specify this average weight. Specifying a particular value, the research hypothesis must be the null hypothesis (Section 4.5.1).

If the research hypothesis is the null hypothesis because it contains a single value for the population parameter, we find support for our research hypothesis if we do *not* reject the null hypothesis. We can be wrong in not rejecting the null hypothesis. If we do not reject a null hypothesis that is actually false, we make a Type II error.

The significance level is irrelevant now because the significance level is the probability of making a Type I error. We do not reject the null hypothesis, so we can never reject a true null hypothesis (Type I error). Instead, the probability of making a Type II error is important, or rather, the probability of not making this error. This is the power of the test.

So if our research hypothesis represents the null hypothesis and our research hypothesis is supported (not rejected), we need test power to know how confident we can be about the support that we have found. Here, test power is key, not statistical significance.

5.6 Test Your Understanding

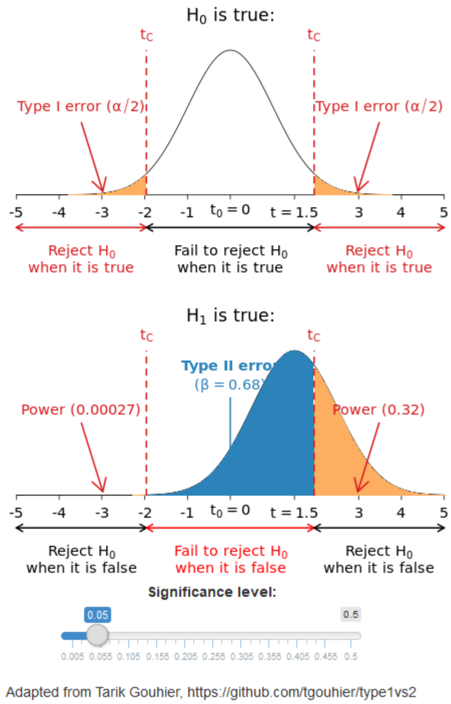


Figure 5.7: Effect size, power, Type I and Type II error.

Figure 5.7 shows sampling distributions for two worlds. Both sampling distributions are approximated with a t distribution. At the top is the hypothetical world of the researcher. In this hypothetical world, the researcher's null hypothesis is true, namely that average candy weight is 2.8 gram in the population. At the bottom is the real world in which average candy weight is 2.9 gram. The standard deviation of candy weights is 0.5.

1. What do the values on the horizontal axes mean?
2. What does t_c mean?
3. Why are the t values and $t_c$values exactly the same on both horizontal axes?
4. What is the unstandardized and standardized effect size if average candy weight is 2.83 gram in our sample?
5. What do the double-sided horizontal red arrows represent in the top graph?
6. Why are the double-sided horizontal arrows red in the top graph and black in the bottom graph and the other way around?

7. Why are the orange sections in the top graph labelled Type I error?
8. What happens to the probability of Type II error (the blue section in the bottom graph) if you change the significance level with the slider?
9. What does *Type II error* mean in the bottom graph?

5.7 Take-Home Points

- Effect size is related to practical significance. Effect sizes are expressed by (standardized) mean differences, regression coefficients, and measures of association such as the correlation coefficient, R^2, and eta^2.
- Statistical significance of a test depends on effect size and sample size.
- Not rejecting a false null hypothesis is a Type II error. A researcher can make this error only if the null hypothesis is not rejected.
- The probability of making a Type II error is commonly denoted with the Greek letter beta (β).
- The probability of *not* making a Type II error is the power of the test.
- The power of a test tells us the probability that we reject the null hypothesis if there is an effect of a particular size in the population. The larger this probability, the more confident we are that we do not overlook this effect when we do not reject the null hypothesis.

Chapter 6

Critical Discussion of Null Hypothesis Significance Testing

Key concepts: problems with null hypothesis significance testing, meta-analysis, replication, frequentist versus Bayesian inference.

Summary

How important is null hypothesis significance testing?

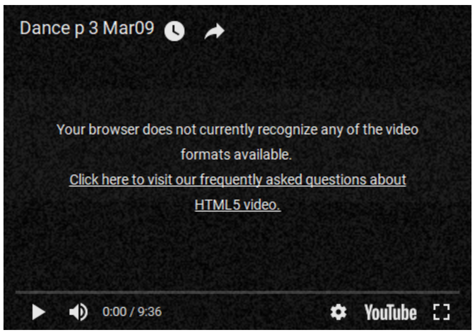
In the preceding chapters, we learned to test null hypotheses. Null hypothesis significance testing is widely used in the social and behavioral sciences. There are, however, problems with null hypothesis significance tests that are increasingly being recognized.

The statistical significance of a null hypothesis test depends strongly on the size of the sample (Chapters 4 and 5), so non-significance may merely mean that the sample is too small. In contrast, irrelevant tiny effects can be statistically significant in a very large sample. Finally, we normally test a null hypothesis that there is no effect whereas we have good reasons to believe that there is an effect in the population. So what does a significant test result really tell us?

Among the alternatives to null hypothesis significance testing, using a confidence interval to estimate effects in the population is easiest to apply. It is closely related to null hypothesis testing, as we have seen in Section 4.8 but it offers us information with which we can draw a more nuanced conclusion about our results.

6.1 Criticisms of Null Hypothesis Significance Testing

In null hypothesis significance testing, we totally rely on the test's p value. If this value is below .05 or another significance level, we reject the null hypothesis and we accept it otherwise. Is this a wise thing to do? Watch the video.



6.1.1 Statistical significance is not a measure of effect size

Perhaps, Chapter 4 on null hypothesis testing should have been titled *Am I Lucky or Unlucky?* instead of *Am I Right or Am I Wrong?* When our sample is small, the power to reject a null hypothesis is rather small, so it occurs often that we retain the null hypothesis even if it is wrong. There is a lot of uncertainty about the population if our sample is small. So we must be lucky to draw a sample that is sufficiently at odds with the null hypothesis to reject it.

If our sample is large or very large, small differences between what we expect according to our null hypothesis can be statistically significant even if the differences are too small to be of any practical value. A statistically significant result need not be practically relevant. In all, statistical significance does not tell us about effect size.

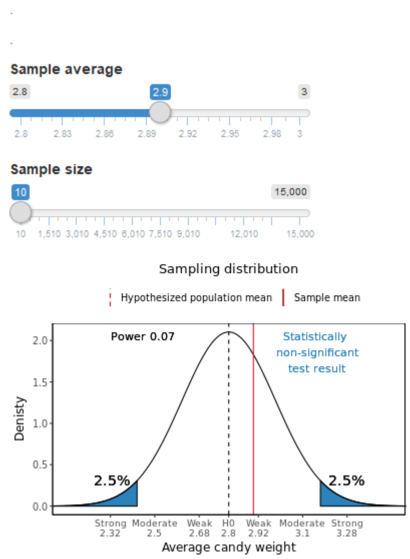


Figure 6.1: Any effect can be statistically significant.

1. What is the null hypothesis in Figure

```
* The null hypothesis is that average candy weight is 2.8 (gram) in the
population.
```

2. In Figure 6.1, what should you do to obtain a statistically significant result for a sample average of 2.9 gram if the null hypothesis states that average candy weight is 2.8?

```
* Increase sample size. Already at a sample size of 140, a sample with average
candy weight of 2.9 gram differs significantly from the hypothesized 2.8 gram.
In other words, a sample of this size with an average of 2. is quite unlikely
to be drawn from the hypothesized population with an average of 2.8.
```

3. Can you get a statistically significant result for the smallest effect size, that is, for the smallest non-zero difference between the observed sample average and the hypothesized population average?

```
* The smallest sample mean value larger than 2.8 that we can select with the
sample average slider is 2.81. The smallest non-zero difference between sample
mean and hypothesized population average, then, is .01 (gram).
* A sample of about 14,000 observations will give a statistically significant
test result.
* So yes, we can get a statistically significant result for the smallest
non-zero effect size.
* A difference of 0.01 gram (less than 0.4% of the hypothesized weight) may
not be practically relevant.
```

4. There is one sample mean for which we can never reject the null hypothesis, no matter how large we make the sample. Which sample mean would that be?

```
* If the sample mean is exactly equal to the hypothesized population mean, in
this example, exactly 2.8 gram, the null hypothesis will never be rejected.
This makes sense because we find exactly what we expect.
* Increasing sample size reduces the width of the interval between the
rejection regions (between the blue tails in this graph). But the hypothesized
value at the centre of this interval will always fall in between the rejection
regions.
```

5. When is a statistically significant result more surprising: with high or low test power? Note: The power calculated in Figure ?? assumes that average candy weight in the population equals average candy weight in the sample.

```
* A statistically significant result is more surprising with low test power.
* With low test power, the difference between hypothesized and true population
values must be quite large to obtain a statistically significant result or we
must be very lucky. In this sense, a statistically significant result is more
surprising with low test power than with high test power.
* With low test power, we are more likely to have an effect that is
practically relevant.
```

It is a common mistake to think that statistical significance is a measure of the strength or practical significance of an effect. In the video (Figure ??, this mistaken interpretation is expressed by the type of sound associated with a p value: the lower the p value of the test, the more joyous the sound.

It is wrong to use statistical significance as a measure of strength or importance. In a large sample, even irrelevant results can be highly significant and in small samples, as demonstrated in the video, results can sometimes by highly significant and sometimes be insignificant. Never forget:

> A statistically significant result ONLY means that the null hypothesis must be rejected.

If we want to say something about the magnitude of an effect in the population, we should use effect size. All we have is the effect size measured in our sample and a statistical test usually telling us whether or not we should reject the null hypothesis that there is no effect in the population.

If the statistical test is significant, we conclude that an effect probably exists in the population. We may use the effect size in the sample as a point estimate of the population effect. This effect size should be at the core of our interpretation. Is it large (strong), small (weak), or perhaps tiny and practically irrelevant?

If the statistical test is not significant, it is tempting to conclude that the null hypothesis is true, namely, that there is no effect in the population and we need not interpret the effect that we find in our sample. But this is not right. Finding insufficient proof for rejecting the null hypothesis does not prove that the null hypothesis is true.

In a two-sided significance test, the null hypothesis specifies one particular value for the sample outcome. If the outcome is continuous, for instance, a mean or regression coefficient, the null hypothesis can hardly ever be true. The true population value is very likely not exactly the same as the hypothesized value. It may be only slightly different but it is different.

Instead of focusing on true versus false, we had better take into account the probability that we reject the null hypothesis, which is test power. If test power is low, as it often is in social scientific research without very large samples, we should realize that there can be a substantive difference between true and hypothesized population values even if the test is not statistically significant.

With low power, we have high probability of not rejecting a false null hypothesis even if the true population value is quite different from the hypothesized value. For example, a small sample of candies drawn from a population with average candy weight of 3.0 gram may not reject the null hypothesis that average candy weight is .8 gram in the population. The statistical test result should not make us conclude that there is no interesting effect. The test may not pick up substantively interesting effects.

In contrast, if our test has very high power, we should expect effects to be statistically significant. Even tiny effects that are totally irrelevant from a substantive point of view. For example, an effect of exposure on attitude of 0.01 on a 10-point scale, is likely to be statistically significant in a very large sample but it is substantively uninteresting.

As noted before (Section 5.3.5), standard statistical software usually does not report the power of a test. For this reason, it is not common practice to evaluate the statistical significance of results in combination with test power.

By now, however, you understand that test power is affected by sample size. You should realize that null hypotheses are easily rejected in large samples but they are more difficult to reject in small samples. Don't let your selection of interesting results be guided predominantly by statistical significance if your sample is not very large.

6.1.2 Knocking down straw men (over and over again)

There is another aspect in the practice of null hypothesis significance testing that is not very satisfactory. Remember that null hypothesis testing was presented as a means for the researcher to use previous knowledge as input to her research. The development of science requires us to expand existing knowledge. Does this really happen in the practice of null hypothesis significance testing?

Imagine that previous research has taught us that one additional unit of exposure to advertisements for a brand increases a person's brand awareness on average by 0.1 unit if we use well-tested standard scales for exposure and brand awareness. If we want to use this knowledge in our own research, we would hypothesize that the regression coefficient of exposure is 0.1 in a regression model predicting brand awareness.

Well, try to test this null hypothesis in your favourite statistics software. Can you actually tell the software that the null hypothesis for the regression coefficient is 0.1? Most likely you can't because the software automatically tests the null hypothesis that the regression coefficient is zero in the population.

This approach is so prevalent that null hypotheses equating the population value of interest to zero have received a special name: the *nil hypothesis* or *the nil* for short (see Section 4.3.3). So you simply cannot include previous knowledge in your test if the software always tests the nil.

The null hypothesis that there is no association between the predictor variable and the outcome variable in the population may be interesting to reject if you really have no clue about the association. But in the example above, previous knowledge makes us expect a positive association of a particular size. Here, it is not interesting to reject the null hypothesis of no association. The null hypothesis of no association is a *straw man* in this example. It is unlikely to stand the test and nobody should applaud if we knock it down.

Rejecting the nil time and again should make us wonder about scientific progress and our contribution to it. Are we knocking down straw men hypotheses over and over again? Is there no way to accumulate our efforts?

6.2 Alternatives for Null Hypothesis Significance Testing

In the social and behavioral sciences, null hypothesis testing is still the dominant type of statistical inference. For this reason, an introductory text like the current one must discuss null hypothesis significance testing. But it should discuss it thoroughly, so the problems and errors that occur with null hypothesis testing become clear and can be avoided.

The problems with null hypothesis significance testing are increasingly being recognized. Alternatives to null hypothesis significance testing have been developed and are becoming more accepted within the field. In this section, some alternatives are briefly sketched.

6.2.1 Estimation instead of hypothesis testing

Following up on a report commissioned by the American Psychological Association APA (Wilkinson, 1999), the 6th edition of the *Publication Manual of the American Psychological Association* recommends reporting and interpreting confidence intervals rather than relying solely on null hypothesis tests.

Estimation is becoming more important: Assessing the precision of our statements about the population rather than deciding pro or con our hypothesis about the population. This is an important step forward and it is easy to accomplish if your statistical software reports confidence intervals.

[htbp]

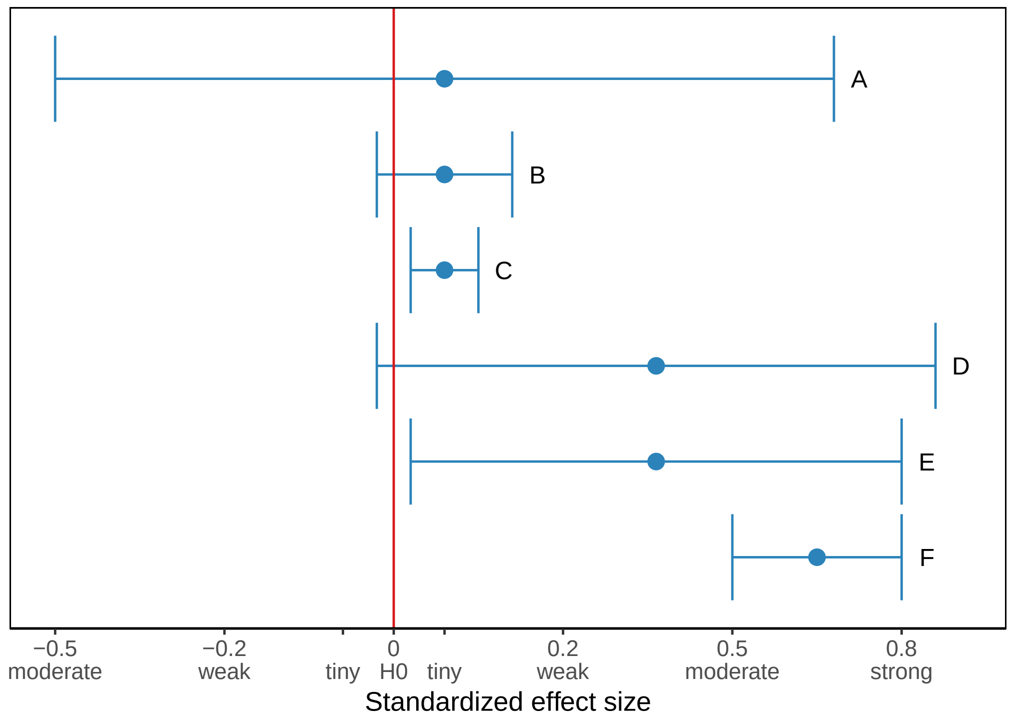


Figure 6.2: What is the most sensible interpretation of the results represented by the confidence interval ?

Figure 6.2 shows six confidence intervals for a population value, for instance, the effect of exposure to advertisements on brand awareness, and the sample result as point estimate (dot). The horizontal axis is labeled by the size of the effect: the difference between the effect in the sample and the absence of an effect according to the null hypothesis.

1. Would you advise the company to use the advertisement based on a null hypothesis significance test?

```
If the 95% confidence interval does not include H0 (the vertical line), we
must reject the null hypothesis at the 5% significance level. That is the rule
of the game called null hypothesis significance testing.
This is the case for confidence intervals C, E, and F.
If we can reject the null hypothesis, we would be confident that there is an
effect of exposure on brand awareness in the population. But this can be a
tiny effect (intervals B and C) or a large effect (interval F). In case of a
tiny effect, would we recommend to use the advertisement?
```

2. Would you advise the company to use the advertisement based on the confidence intervals in Figure 6.2?

```
If less of the confidence interval extends to the left of H0 (negative effect)
and more of it is situated to the right (positive effect), we are more
confident that the effect is positive in the population.
If the range of plausible population values is more strongly positive, we are
more certain that there is a substantial positive effect in the population.
This is the most important reason to recommend using the advertisement.
Interval F offers the most convincing evidence for a substantial positive
effect. Intervals D and E also suggest a positive effect but we are not so
sure about the size of the effect: it may be tiny or perhaps even absent but
it can also be moderate to strong.
```

A confidence interval shows us whether or not our null hypothesis must be rejected (see Section 4.8). The rule is simple: If the value of the null hypothesis is within the confidence interval, the null hypothesis must not be rejected. By the way, note that a confidence interval allows us to test a null hypothesis other than the nil (Section 6.1.2). If we hypothesize that the effect of exposure on brand awareness is 0.1, we reject the hypothesis if the confidence interval of the regression coefficient does not include 0.1.

At the same time, however, confidence intervals allow us to draw a more nuanced conclusion. A confidence interval displays our uncertainty about the result. If the confidence interval is wide, we are quite uncertain about the true population value. If a wide confidence interval includes the null hypothesis near one of its boundaries, we do not reject the null hypothesis but it still is plausible that the population value is substantially larger (or substantially smaller) than the hypothesized value.

We should report that the population value seems to be larger (smaller) than specified in the null hypothesis but that we have inconclusive evidence because the test is not statistically significant. This is better than reporting that there is no difference because the statistical test is not significant.

The fashion of speaking of a null hypothesis as "accepted when false", whenever a test of significance gives us no strong reason for rejecting it, and when in fact it is in some way imperfect, shows real ignorance of the research workers' attitude, by suggesting that in such a case he has come to an irreversible decision.

The worker's real attitude in such a case might be, according to the circumstances:

(a) "The possible deviation from truth of my working hypothesis, to examine which the test is appropriate, seems not to be of sufficient magnitude to warrant any immediate modification."

Or it might be:

(b) "The deviation is in the direction expected for certain influences which seemed to me not improbable, and to this extent my suspicion has been confirmed; but the body of data available so far is not by itself sufficient to demonstrate their reality."

(Ronald Aylmer Fisher, 1955: 73)

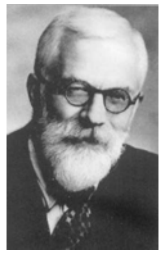


Figure 6.3: Sir Ronald Aylmer Fisher.

In a similar way, a very narrow confidence interval including the null hypothesis and a very narrow confidence interval near the null hypothesis but excluding it should not yield opposite conclusions because the statistical test is significant in the second but not in the first situation. After all, even for the significant situation, we know with high confidence (narrow confidence interval) that the population value is close to the hypothesized value.

Using confidence intervals in this way, we avoid the problem that statistically non-significant effects are not published. Not publishing non-significant results, either because of self-selection by the researcher or selection by journal editors and reviewers, offers a misleading view of research results.

If results are not published, they cannot be used to design new research projects. For example, effect sizes that are not statistically significant are just as helpful to determine sample size as statistically significant effect sizes. A predictor variable without statistically significant effect may have a significant effect in a new research project and should not be discarded if the potential effect size is so substantial that it is practically significant. Moreover, combining results from several research projects helps making more precise estimates of population values, which brings us to meta-analysis.

6.2.2 Meta-analysis

Meta-analysis is a method that capitalizes on previous knowledge. In this method, we collect previous studies on the same topic that use the same or highly similar variables. Combining the results of these studies, we can make statements with higher precision about the population. Basically, we combine the separate samples used for each single study into a large sample, which reduces the uncertainty and allows more precise inferences about the population.

Meta-analysis is a good example of combining research efforts to increase our understanding. It favours estimation over hypothesis testing because the goal is to obtain more precise estimates of population values or effects. Meta-analysis is strongly recommended as a research strategy by Geoff Cumming, who coined the concept *New Statistics*. See Cumming's book (2012), website, or YouTube channel if you are curious to learn more. The video at the start of this chapter is made by Geoff Cumming.

6.2.3 Replication

Another approach that builds upon previous results is *replication*. If we collect new data including the variables that are central in prior research and we execute the same analyses, we *replicate* previous research.

Replication is the surest tool to check results of previous research. Checks do not necessarily serve to expose fraud and mistakes. They tell us whether prior research results still hold at a later time and perhaps in another context. Thus, we can decrease the chance that our previous results derive from an atypical sample. But replication also helps us to develop more general theories and discard theories that apply only to special situations.

6.2.4 Bayesian inference

A more radical way of including previous knowledge in statistical inference is *Bayesian inference*. Bayesian inference regards the sample that we draw as a means to update the knowledge that we already have or think we have on the population. Our previous knowledge is our starting point and we are not going to just discard our previous knowledge if a new sample points in a different direction, as we do when we reject a null hypothesis.

Think of Bayesian inference as a process similar to when we predict the weather. If I try to predict tomorrow's weather, I am using all my weather experience to make a prediction. If my prediction turns out to be more or less correct, I don't change the way I predict the weather. But if my prediction is patently wrong, I try to reconsider the way I predict the weather, for example, paying attention to new indicators of weather change.

Bayesian inference uses a concept of probability that is fundamentally different from the type of inference presented in previous chapters, which is usually called *frequentist inference*. Bayesian inference does not assume that there is a true population value. Instead, it regards the population value as a random variable, that is, as something with a probability.

Again, think of predicting the weather. I am not saying to myself: "Let us assume that tomorrow will be a rainy day. If this is correct, what is the probability that the weather today looks like it does?" Instead, I think of the probability that it will rain tomorrow. Bayesian probabilities are much more in line with our everyday concept of probability than the dice-based probabilities of frequentist inference.

Due to this more intuitive notion of probability, the *credible interval*, which is the Bayesian equivalent of the confidence interval, means what we would like the confidence interval to mean, namely the interval within which the population value is located with the selected probability.

Bayesian inference is intuitively appealing but it has not yet spread widely in the social and behavioral sciences. Therefore, I merely mention this strand of statistical inference and I refrain from giving details. Its popularity, however, is increasing, so you may come in contact with Bayesian inference sooner or later.

6.3 What If I Do Not Have a Random Sample?

In our approach to statistical inference, we have assumed all the time that we could have drawn a very large number of random samples from the same population. The large number of samples constitutes the sampling distribution that tells us about the probability of drawing the one random sample that we have actually drawn.

What if I do not have a random sample? Can I still estimate confidence intervals or test null hypotheses? If you carefully read reports of scientific research, you will encounter examples of statistical inference on non-random samples or data that are not samples at all but rather represent an entire population, for instance, all people visiting a particular web site. Statistical inference is clearly being applied to data that are not sampled at random from an observable population. The fact that it happens, however, is not a guarantee that it is right.

We should note that statistical inference based on a random sample is the most convincing type of inference because we exactly know the nature of the uncertainty in the data, namely the chance introduced by our random sampling. Think of exact methods for creating a sampling distribution. If we know the distribution of candy colours in the population of all candies, we can calculate the exact probability of drawing a sample bag with, for example, 25 per cent of all candies being yellow if we carefully draw the sample at random.

We can calculate the probability because we understand the process of random sampling. For example, we know that each candy has the same probability to be included in the sample. The uncertainty or probabilities arise from the way we designed our data collection, namely as a random sample from a much larger population.

In summary, we work with a concrete population and we know how chance affects our sample if we draw a random sample. We do not have a concrete population and we do not know the workings of chance if we want to apply statistical inference to data that are not collected as a random sample. In this situation, we have to substantiate the claim that our data set can be regarded as a random sample.

6.3.1 Theoretical population

Sometimes, we have data for a population instead of a sample. For example, we have data on all visitors of our website because our website logs visits. If we investigate all people visiting a particular website, what is the wider population?

We may argue that this set of people is representative of a wider set of people visiting similar web sites or of the people visiting this website at different time points. This is called a *theoretical population* because we imagine such a population instead of actually sampling from an observable population.

We have to motivate why we think that our data set (our website visitors) can be regarded as a random sample from the theoretical population. This can be difficult. Is it really just chance that some people visit our website whereas other people visit another (similar) website? Is it really just chance that some visit our website this week but not next week and the other way around? And how about people visiting our website both weeks?

If it is plausible that our data set can be regarded as a random sample from a theoretical population, we may apply inferential statistics to our data set to generalize our results to the theoretical population. Of course, a theoretical population, which is imaginary, is less concrete than an observable population. The added value of statistical inference is more limited.

6.3.2 Data generating process

An alternative approach discards with generalization to a population. Instead, it regards our observed data set as the result of a theoretical *data generating process* (for instance, see Frick, 1998; Hayes, 2013: 50-51). In an experiment, for example, exposure to a celebrity endorsing a fund-raising campaign triggers a process within the participants that results in a particular willingness to donate. Under similar circumstances and personal characteristics, this process yields the same outcomes, that is, generates the same data set.

There is a complication. The circumstances and personal characteristics are very unlikely to be the same every time the process is at work (generates data). A person may pay more or less attention to the stimulus material, she may be more or less susceptible to this type of message, or in a better or worse mood for caring about other people, and so on.

As a consequence, we have variation in the outcome scores for participants who are exposed to the same celebrity and who have the same scores on the personal characteristics that we measured. This variation is supposed to be random, that is, the result of chance. In this approach, then, random variation is not caused by random sampling but by fluctuations in the data generating process.

Compare this to a machine producing candies. Fluctuations in the temperature and humidity within the factory, vibrations due to heavy trucks passing by, and irregularities in the base materials may affect the weight of individual candies. The weights are the data that we are going to analyze and the operation of the machine is the data generating process.

We can use inferential techniques developed for random samples on data with random variation stemming from the data generation process if the probability distributions for sampling distributions apply to random variation in the data generating process. This is the tricky thing about the data generating process approach.

It has been shown that means of random samples have a normal or t distributed sampling distribution (under particular conditions). The normal or t distribution is a correct choice for the sampling distribution here. In contrast, we have no correct criteria for choosing a probability distribution representing chance in the process of generating data that are not a random sample. We have to make up a story about how chance works and to what probability distribution this leads. This is a more contestable choice.

What arguments can a researcher use to justify the choice of a theoretical probability distribution for the sampling distribution? A bell-shaped probability model such as the normal or t distribution is a plausible candidate for capturing the effects of many independent causes on a numeric outcome (see Lyon, 2014 for a critical discussion). If we have many unrelated causes that affect the outcome, for instance, a person's willingness to donate to a charity, particular combinations of causes will push some people to be more willing than the average and other people to be less willing.

So we should give examples of unobserved independent causes that are likely to affect willingness to donate to justify a normal or t distribution. For example, mood differences between participants, fatigue, emotions, prior experiences with the charity, and so on.

This is an example of an argument that can be made to justify the application of t tests in tests on means, correlations, or regression coefficients to data that is not collected as a random sample. The argument can be more or less convincing. The chosen probability distribution can be right or wrong and we will probably never know which of the two it is.

Figure 6.4: Carl Friedrich Gauss.

The normal distribution is usually attributed to Carl Friedrich Gauss (1809). Pierre-Simon Laplace (1812), among others, proved the central limit theorem, which states that under certain conditions the mean of a large number of independent random variables are approximately normally distributed. Based on this theorem, we expect that the overall (average) effect of a large number of independent causes (random variables) produces a variation that is normally distributed.

Figure 6.5: Pierre-Simon Laplace.

6.4 Test Your Understanding

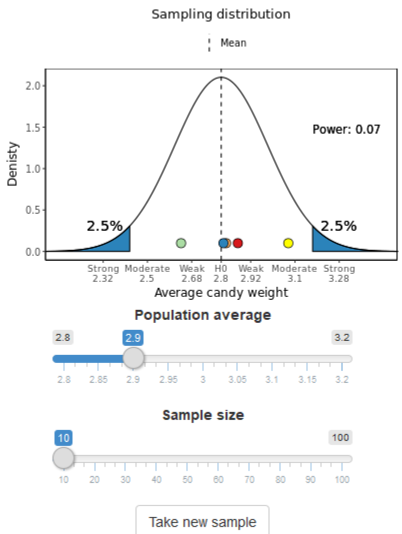


Figure 6.6: How do statistical significance, effect size, sample size, and power relate?

Figure 6.6 displays the sampling distribution for candy weight under the null hypothesis that average candy weight is 2.8 in the population. The horizontal axis shows average candy weight and the standardized effect size (Cohen's d) in a sample: weak, moderate, or strong. Five samples are drawn from a population with the average candy weight specified by the top slider. The samples' average candy weights are represented by coloured dots on the horizontal axis.

1. Which sample means are statistically significant (5% two-sided) and which are not?
2. Is the null hypothesis true for samples with non-significant mean scores?
3. What happens to the statistical significance of the sample means and to test power if you change sample size?

6.5 Take-Home Points

- Null hypothesis significance test results should be interpreted in relation to sample size and, if possible, test power.
- Statistically significant results need not be relevant or important. A small, negligible difference between the sample outcome and the hypothesized population value can be statistically significant in a very large sample with high test power.

- A practically relevant and important difference between the sample outcome and the hypothesized population value need not be statistically significant in a small sample or with a test with low power.
- Give priority to effect size over statistical significance in your interpretation of results.
- A confidence interval shows us how close to and distant from the hypothesized value the true population value is likely to be. It helps us to draw a more nuanced conclusion about the result than a null hypothesis significance test.
- Applying statistical inference to data other than random samples requires justification of either a theoretical population or a data generating process with a particular probability distribution.

Chapter 7

Moderation with Analysis of Variance (ANOVA)

> Key concepts: eta-squared, between groups variance, within groups variance, F test on analysis of variance model, pairwise comparisons, post-hoc tests, one-way analysis of variance, two-way analysis of variance, factorial design, balanced design, main effects, moderation, interaction effect, higher-order interactions.

Summary

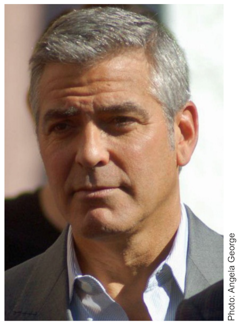
> How do we test score level differences for three or more groups and what if group effects are not the same for all participants?

Imagine an experiment in which participants watch a video promoting a charity. They see George Clooney, Angelina Jolie, or no celebrity endorse the charity's fund-raiser. Afterwards, their willingness to donate to the charity is measured. Which campaign works best, that is, produces highest average willingness to donate? Or does one campaign work better for females, another for males?

In this example, we want to compare the level of outcome scores (average willingness to donate) across more than two groups (participants who saw Clooney, Jolie, or no celebrity). To this end, we use analysis of variance. The null hypothesis tested in analysis of variance states that all groups have the same average outcome score in the population.

This null hypothesis is similar to the one we test in an independent-samples t test for two groups. With three or more groups, we must use the variance of the group means (between-groups variance) to test the null hypothesis. If the between-groups variance is zero, all group means are equal.

In addition to between-groups variance, we have to take into account the variance of outcome scores within groups (within-groups variance). Within-groups variance is related to random



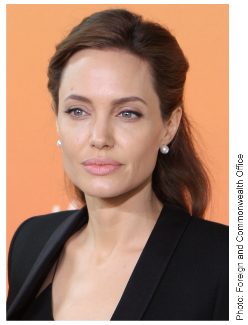


Figure 7.1: George Clooney and Angelina Jolie.

group mean differences that we may expect in random samples. The ratio of between-groups variance over within-groups variance gives us the F test statistic, which has an F distribution.

Differences in average outcome scores for groups on one predictor variable (factor) are called a main effect. A main effect represents an overall or average effect of a factor. If we have only one factor in our model, for instance, the endorser of the fund-raiser, we apply a one-way analysis of variance. With two factors, we have a two-way analysis of variance, and so on.

With two or more factors, we can have interaction effects in addition to main effects. An interaction effect is the joint effect of two or more factors on the outcome variable. An interaction effect is best understood as different effects of one factor across different groups on another factor. For example, Clooney may increase willingness to donate among females but Jolie works best for males.

The phenomenon that a variable can have different effects for different groups on another variable is called moderation. We usually think of one factor as the predictor and the other factor as the moderator. The moderator (e.g., sex) changes the effect of the predictor (e.g., celebrity endorser) on the outcome variable (e.g., willingness to donate).

7.1 Different Score Levels for Three or More Groups

Celebrity endorsement theory states that celebrities who publicly state that they favour a product, candidate, or cause, help to persuade consumers to adopt or support the product, candidate, or cause (for a review, see Erdogan, 1999; for an alternative approach, see McCracken, 1989).

Imagine that we want to test if the celebrity who endorses a fund raiser in a fund-raising campaign makes a difference to people's willingness to donate. We will be using the celebrities George Clooney and Angelina Jolie, and we will compare campaigns with one of them to a campaign without celebrity endorsement.

Let us design an experiment to investigate the effects of celebrity endorsement. We sample a number of people (participants), whom we assign randomly to one of three groups. We show a campaign video with George Clooney to one group, a video with Angelina Jolie to another group, and the third group—the control group—sees a campaign video without celebrity endorsement. So we have three experimental conditions (Clooney, Jolie, no endorser) as our predictor variable.

Our outcome is a numeric scale assessing the participant's willingness to donate to the fund raiser on a scale from 1 (“absolutely certain that I will not donate”) to 10 (“absolutely certain that I will donate”). We will compare the average outcome scores among groups. If groups with Clooney or Jolie as endorser have systematically higher average willingness to donate than the group without celebrity endorsement, we conclude that celebrity endorsement has a positive effect.

In statistical terminology, we have a categorical predictor and a numerical outcome. In experiments, we usually have a very limited set of treatment levels, so our predictor is categorical. For nuanced results, we usually want to have a numeric outcome. Analysis of variance was developed for this kind of data (R. A. Fisher, 1919), so it is widely used in the context of experiments. Note, however, that it can also be used in non-experimental situations as long as the predictor is categorical and the outcome numeric.

7.1.1 Mean differences as effects

Figure 7.2 shows the willingness to donate scores for twelve participants in our experiment. Four participants saw Clooney, four saw Jolie, and four did not see a celebrity endorser in the video that they watched.

Figure 7.2: How do group means relate to effect size?

1. In the sample of (12) participants displayed in Figure 7.2, what do the double-sided vertical arrows represent?

```
* The double-sided vertical arrows represent the differences in average
willingness to donate across the three experimental conditions: exposure to
Clooney as endorser, Jolie as endorser, or no celebrity endorser.
```

2. How do the double-sided vertical arrows relate to effect size (eta^2)? Explain the relation in your own words and change group means to verify your expectations.

```
* The more different the group means, the larger the red arrows, the larger the
between-groups variance, the larger eta-squared.
* For those of you who love the details: Between-groups variance squares the
distances (actually the distances between group means and the grand mean but
that is not relevant here). As a consequence, a long red arrow contributes more
to between-groups variance than two short arrows that are just as long as the
long arrow if they are summed. This is the reason that eta-squared increases if
the group in the middle moves closer to the top group (or the bottom group).
```

Average outcome score for a group represents a group's score level. The group averages in Figure 7.2 tell us for which celebrity the average willingness to donate is higher and for which situation it is lower.

Random assignment of test participants to experimental groups (e.g., which video is shown) creates groups that are in principle equal on all imaginable characteristics except the experimental treatment(s) administered by the researcher. Participants who see Clooney should have more or less the same average age, knowledge, and so on as participants who see Jolie or no celebrity. After all, each experimental group is just a random sample of participants.

If random assignment was done successfully, differences between group score levels can only be caused by the experimental treatment (see Chapter mediation) for a better understanding). Mean differences are said to represent the *effect* of experimental treatment in analysis of variance.

Analysis of variance was developed for the analysis of randomized experiments, where effects can be interpreted as causal effects. Note, however, that analysis of variance can also be applied to non-experimental data. Although mean differences are still called effects in the latter type of analysis, these need not be causal effects.

In analysis of variance, then, we are simply interested in the differences between group means. The conclusion for a sample is easy: Which groups have higher average score on the outcome variable and for which are they lower? A simple means plot, such as Figure 7.3, aids interpretation and helps communicating results to the reader.

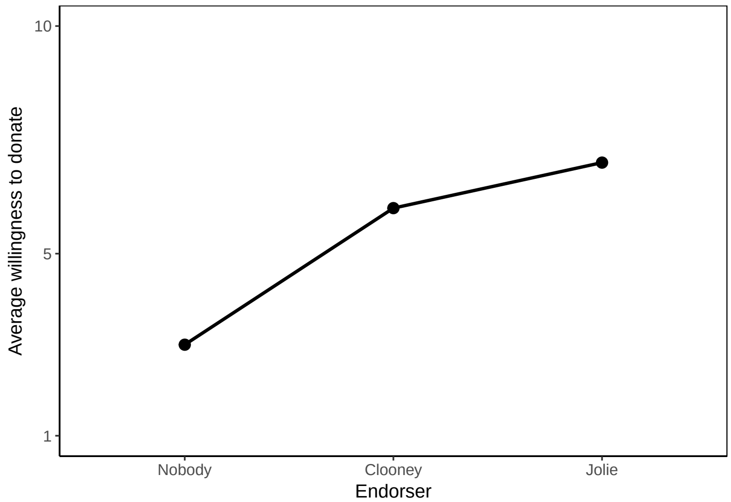


Figure 7.3: A means plot showing that average willingness to donate is higher with a celebrity endorser than without a celebrity endorser.

Effect size in an analysis of variance refers to the overall differences between group means. We use eta^2 as effect size, which gives the proportion of variance in the outcome (willingness to donate) explained or predicted by the group variable (experimental condition).

Rules of thumb for the interpretation of eta^2:

- 0,01 = small or weak effect: 1% explained variance,
- 0,09 = medium-sized or moderate effect: 9% explained variance,
- 0,25 = large or strong effect: 25% explained variance.

7.1.2 Between-groups variance and within-groups variance

For a better understanding of eta^2 and the statistical test of an analysis of variance model, we have to compare the scores to the group averages and to the overall average. Figure 7.4 adds overall average willingness to donate to the plot with participants' scores and average experimental group scores.

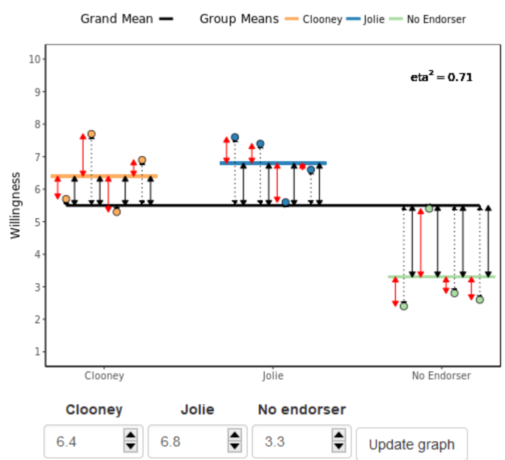


Figure 7.4: Which part of score differences tells us about the differences between groups?

1. In Figure 7.4, what do the solid red arrows represent?

```
* The solid red arrows represent the difference between an individual score
and the average score of the group to which the individual belongs. For
example, the left-most orange dot represents the willingness score of a
participant who was exposed to Clooney as endorser. The orange line represents
the average willingness score of the participants who were exposed to Clooney.
The red arrow is the difference between the individual's willingness score and
the mean score of it's group. Squaring and summing the solid red arrows
(within-groups differences) yields the within-groups variance.
```

2. What do the solid black arrows represent?

```
* The solid black arrows represent the difference between an individual's group
score, for instance, the average willingness score of all participants who were
exposed to Clooney, and the average score of all participants (the grand or
overall mean).
* If we square and sum these differences, we get the variance of group means,
which is called the between-groups variance.
```

3. What do the dotted black arrows in Figure 7.4 represent?

```
* The dotted black arrows represent the difference between individual
willingness scores and the average willingness scores of all participants.
Square and sum them to get the overall variance. Take the square root of the
variance to obtain the overall standard deviation.
```

4. Which arrows relate to effect size eta^2? Change group means (and press the *Update graph* button) to verify your expectation.

```
* The solid black arrows relate to eta-squared. Eta-squared is larger if the
differences between the group means are larger. The solid black arrows express
these differences.
* If you decrease the differences between the group means, eta-squared
decreases and the solid black arrows become smaller. The red arrows (differences
between scores and their group means) remain the same.
* The dotted black arrows change but some become longer and others become
shorter if you decrease the differences between the group means, so the
dotted black arrows are not so clearly related to eta-squared.
```

Let us assume that we have measured the willingness to donate for a sample of 12 participants in our study as depicted in Figure 7.4. Once we have our data, we first have a look at the percentage of variance that is explained, eta^2. What does it mean if we say that a percentage of the variance is explained when we interpret eta^2?

The variance that we want to explain consists of the differences between the scores of the participants on the outcome variable and the overall or grand mean of all outcome scores. Remember that a variance measures deviations from the mean. The dotted black arrows in Figure 7.4 express the distances between outcome scores and the grand average. Squaring and averaging these distances over all observations gives us the total variance in outcome scores.

The goal of our experiment is to explain why some of our participants have a willingness to donate that is far above the grand mean (horizontal black line in Figure 7.4) while others score a lot lower. We hypothesized that participants are influenced by the endorse they have seen. If an endorser is more effective, the overall level of willingness should be higher. In other words, the average willingness should be higher for participants confronted with this endorser.

Group average willingness is what we can predict from the experimental treatment. If we know the group to which a participant belongs—which celebrity she saw endorsing the fundraising campaign—we can use the average outcome score for the group as the predicted outcome for each group member—her willingness to donate due to the endorser she saw. The predicted group scores are represented by the coloured horizontal lines for group means in Figure 7.4.

Now what part of the variance in outcome scores (dotted black arrows in Figure 7.4) is explained by the experimental treatment? If we use the experimental treatment as predictor of willingness to donate, we predict that a participant's willingness equals her group average (horizontal coloured line) instead of the overall average (horizontal black line), which we use if we do not take into account the participant's experimental treatment.

So the difference between the overall average and the group average is what we predict and explain by the experimental treatment. This difference is represented by the solid black arrows in Figure 7.4. The variance of the predicted scores is obtained if we average the squared sizes of the solid black arcs for all participants. This variance is called the *between-groups variance*.

Playing with the group means in Figure 7.4, you may have noticed that eta^2 is high if there are large differences between group means. In this situation we have high between-groups variance—large black arrows—so we can predict a lot of the variation in outcome scores between participants.

In contrast, small differences between group averages allow us to predict only a small part of the variation in outcome scores. If all group means are equal, we can predict none of the variation in outcome scores because the between-groups variance is zero. As we will see in Section 7.1.3, zero between-groups variance is central to the null hypothesis in analysis of variance.

The experimental treatment predicts that a participant's willingness equals the average willingness of the participant's group. It cannot predict or explain that a participant's willingness score is slightly different from her group mean (the red double-sided arrows in Figure 7.4). *Within-groups variance* in outcome scores is what we cannot predict with our experimental treatment; it is prediction error. In some SPSS output, it is therefore labeled as "Error".

7.1.3 F test on the model

Average group scores tell us whether the experimental treatment has effects within the sample (Section 7.1.1). If the group who saw Angelina Jolie as endorser has higher average willingness to donate than the group who did not see an endorser, we conclude that Angelina Jolie makes a difference in the sample. But how about the population?

If we want to test whether the difference that we find in the sample also applies to the population, we use the substantive null hypothesis that all average outcome scores are equal in the population from which the samples were drawn. In our example, the null hypothesis states that people in the population who would see George Clooney as endorser are just as willing to donate as people who would see Angelina Jolie or who would not see a celebrity endorser at all.

We use the variation in group means as the number to express the size of differences between group means. If all groups have the same average outcome score, the between-groups variance is zero. The larger the differences, the larger the between-groups variance (see Section 7.1.2).

We cannot just use the between-groups variance as the test statistic because we have to take into account chance differences between sample means. If we draw samples from the same population, the sample means can still be different because we draw samples at random. These sample mean differences are due to chance, they do not reflect true differences within the population.

We have to correct for chance differences and this is done by taking the ratio of between-groups variance over within-groups variance. This ratio gives us the relative size of observed differences between group means over group mean differences that we expect by chance.

Our test statistic, then, is the ratio of two variances: between-groups variance and within-groups variance. The F distribution approximates the sampling distribution of the ratio of two variances, so we can use this probability distribution to test the significance of the group mean differences we observe in our sample.

Remember that we used the F distribution before for testing a ratio of two variances, namely, in Levene's F test for the null hypothesis that two groups have the same population variance (Section 4.2.6). Now we use it to test if three or more means are equal in the population.

Long story short: We test the substantive null hypothesis that all groups have the same population means in an analysis of variance. But behind the scenes, we actually test between-groups variance against within-groups variance. That is why it is called analysis of variance.

7.1.4 Assumptions for the F test in analysis of variance

There are two important assumptions that we must make if we use the F distribution in analysis of variance: (1) Independent samples and (2) homogeneous population variances.

7.1.4.1 Independent samples

The first assumption is that the groups can be regarded as independent samples. As in an independent-samples t test, it must be possible *in principle* to draw a separate sample for each group in the analysis. Because this is a matter of principle instead of how we actually draw the sample, we have to argue that the assumption is reasonable. We cannot check the assumption against the data.

This is an example of an argument that we can make. In an experiment, we usually draw one sample of participants but we assign participants randomly to one of the experimental conditions. This could have easily been done separately for each experimental group. For example, we first draw a participant for the first condition: seeing George Clooney endorsing the fundraising campaign. Next, we draw a participant for the second condition, e.g., Angelina Jolie. The two draws are independent: whomever we have drawn for the Clooney condition is irrelevant to whom we draw for the Jolie condition. Therefore, draws are independent and the samples can be regarded as independent.

Situations where samples cannot be regarded as independent are the same as in the case of dependent/paired-samples t tests (see Section 2.5.6). For example, samples of first and second observations in a repeated measurement design should not be regarded as independent samples. Some analysis of variance models can handle repeated measurements but we do not discuss them here.

7.1.4.2 Homogeneous population variances

The F test in analysis of variance assumes that the groups are drawn from the same population. This implies that they have the same average score on the outcome variable in the population as well as the same variance of outcome scores. The null hypothesis tests the equality of population means but we must assume that the groups have equal outcome variable variances in the population.

We can use a statistical test to decide whether or not the population variances are equal (homogeneous). This is the same Levene's F test that we have used in combination with independent samples t tests (Section 4.2.9). The test's null hypothesis is that the population variances of the groups are equal. If we do *not* reject the null hypothesis, we decide that the assumption of equal population variances is plausible.

The assumption of equal population variances is less important if group samples are more or less of equal size (a balanced design, see Section 7.3.2). We use a rule of thumb that groups are of equal size if the size of the largest group is less than 10% (of the largest group) larger than the size of the smallest group. If this is the case, we do not care about the assumption of homogeneous population variances.

7.1.5 Which groups have different average scores?

Analysis of variance tests the null hypothesis of equal population means but it does not yield confidence intervals for group means. It does not always tell us which groups score significantly higher or lower.

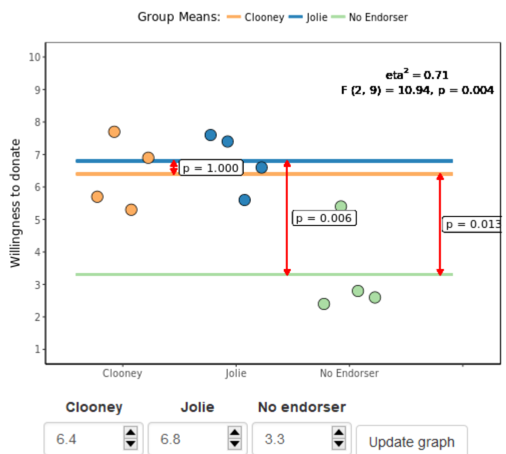


Figure 7.5: Which groups have different average outcome scores in the population?

1. Does the F test in analysis of variance tell us which groups have significantly different average population outcome scores? Can we have the same F test result with different sets of group means? Adjust group means in Figure 7.5 to demonstrate your answer.

```
* If the F test in analysis of variance is statistically significant, we
reject the null hypothesis that all groups have the same mean score in the
population. If we have more than two groups, as in the example, we do not know
which groups score higher and which groups score lower in the population. We
need post-hoc tests to find that out.
```

```
* There is an exception, however. If we have only two groups in our F test,
the statistically significant difference must be between these two groups. A
simple look at the sample means informs us which group has the higher average
score.
* Because the F test does not tell us which group has a higher or lower score,
we can have the same F value for different situations. For example, exchange
the means of two groups. You will get exactly the same F value (in this
balanced design).
```

2. Is it possible that the F test is statistically significant but none of the t tests that compare groups one by one? Can you obtain this situation in Figure 7.5?

```
* Yes, this is possible. For example, set the mean score of the No Endorser
group at 4.5, while you leave the Clooney and Jolie averages at 6.4 and 6.8.
* A t test uses only two of the three groups, so the total number of
observations is lower and, therefore, test power is lower and the null hypothesis is more difficult to
* Note that this situation only occurs if the F test is just significant
(slightly below .05). This illustrates that the .05 threshold is an artificial
boundary.
```

3. Is it okay that we apply both an F test and several t tests to the same group differences?

```
* This is okay provided that we correct for capitalization on chance (see a preceding chapter).
```

If the F test is statistically significant, we reject the null hypothesis that all groups have the same population mean on the outcome variable. In our current example, we reject the null hypothesis that average willingness to donate is equal for people who saw George Clooney, Angelina Jolie, or no endorser for the fund raiser. In other words, we *reject* the null hypothesis that the endorser does *not* matter to willingness to donate.

7.1.5.1 Pairwise comparisons as post-hoc tests

With a statistically significant F test for the analysis of variance model, several questions remain to be answered. Does an endorser increase or decrease the willingness to donate? Are both endorsers equally effective? The F test does not provide answers to these questions. We have to compare groups one by one to see which condition (endorser) is associated with a higher level of willingness to donate.

In a pairwise comparison, we have two groups, for instance, participants confronted with George Clooney and participants who did not see a celebrity endorse the fund raiser, that we want to compare on a numeric outcome, namely their willingness to donate. An independent-samples t test is appropriate here.

With three groups, we can make three pairs: Clooney versus Jolie, Clooney versus nobody, and Jolie versus nobody. We have to execute three t tests on the same data. We already know that there are most likely differences in average scores, so the t tests are executed after the fact, in Latin *post hoc*. Hence the name *post-hoc tests*.

Applying more than one test to the same data increases the probability of finding at least one statistically significant difference even if there are no differences at all in the population. Section 4.9 discussed this phenomenon as capitalization on chance and it offered a way to correct for this problem, namely Bonferroni correction. We ought to apply this correction to the pairwise t tests that we execute if the analysis of variance F test is statistically significant.

7.1.5.2 Two steps in analysis of variance

Analysis of variance, then, consists of two steps. In the first step, we test the general null hypothesis that all groups have equal average outcome scores in the population. If we cannot reject this null hypothesis, we have too little evidence to conclude that there are differences between the groups. Our analysis stops here although it is recommended to report the confidence intervals of the group means to inform the reader. Perhaps our sample was just too small to reject the null hypothesis.

If the F test is statistically significant, we proceed to the second step. Here, we apply independent-samples t tests with Bonferroni correction to each pair of groups to see which groups have significantly different means. In our example, we would compare the Clooney and Jolie groups to the group without celebrity endorser to see if celebrity endorsement increases willingness to donate to the fund raiser. In addition, we would compare the Clooney and Jolie groups to see if one celebrity is more effective than the other.

7.1.5.3 Contradictory results

It may happen that the F test on the model is statistically significant but none of the post-hoc tests is statistically significant. This usually happens when the p value of the F test is near .05. Perhaps the correction for capitalization is too strong; this is known to be the case with the Bonferroni correction. Alternatively, the sample is too small for the post-hoc test. Note that we have fewer observations in a post-hoc test than in the F test because we only look at two of the groups.

This situation illustrates the limitations of null hypothesis significance tests (Chapter 6). Remember that the 5% significance level remains an arbitrary boundary and statistical significance depends a lot on sample size. So do not panic if the F and t tests have contradictory results.

A statistically significant F test tells us that we may be quite confident that at least two group means are different in the population. If none of the post-hoc t tests is statistically significant, we should note that it is difficult to pinpoint the differences. Nevertheless, we should report the sample means of the groups (and their standard deviations) as well as their confidence intervals. The two groups that have most different sample means are most likely to have different population means.

7.2 One-Way Analysis of Variance in SPSS

7.2.1 Instructions

Applying one-way analysis of variance in SPSS and interpreting the results is explained in Section 4.2.9, see Video 4.9.

7.2.2 Exercises

1. How does celebrity endorsement affect the willingness to donate and is one celebrity more effective than the other? Use the data in donors.sav.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=willing_post endorser
  /ORDER=ANALYSIS.
* One-way analysis of variance.
ONEWAY willing_post BY endorser
  /STATISTICS DESCRIPTIVES HOMOGENEITY
  /PLOT MEANS
  /MISSING ANALYSIS
  /POSTHOC=BONFERRONI ALPHA(0.05).

Check data:

There are no impossible values on the variables.

Check assumptions:

The three groups are more or less of equal size: The largest difference is 4
participants, which is less than ten per cent of the smallest group (N = 45).
Anyway, we may assume equal population variances, Levene F (2, 140) = .02, p =
.978.

Interpret the results:

* Willingness to donate depends on the endorsing celebrity. There is a
statistically significant difference between average willingness to donate for
the three endorsers, F (2, 140) = 7.44, p = .001.
* People are more willing to donate if they have seen Clooney (M = 4.99, SD =
1.64) or Jolie (M = 4.95, SD = 1.63) endorse the fund raiser than people who
do not see a celebrity endorser (M = 3.87, SD = 1.47). The differences
between, on the one hand, no celebrity endorser and, on the other hand,
Clooney (p = .002) or Jolie (p = .004) are statistically significant.
```

```
* However, there is not a substantial or statistically significant difference
between Clooney and Jolie with respect to their effect on willingness to
donate (mean difference = 0.04, p = 1.000).

Instead of reporting the F test result in the text, the ANOVA table can be
included.
```

2. The data set smokers.sav contains information on smoking behaviour and attitude towards smoking for a random sample of adults. Does the attitude towards smoking differ among smokers, former smokers, and non-smokers (variable: *status3*)?

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=attitude status3
  /ORDER=ANALYSIS.
* One-way analysis of variance.
ONEWAY attitude BY status3
  /STATISTICS DESCRIPTIVES HOMOGENEITY
  /PLOT MEANS
  /MISSING ANALYSIS
  /POSTHOC=BONFERRONI ALPHA(0.05).

Check data:

There are no impossible scores on the two variables.

Check assumptions:

The Levene test is not statistically significant, F (2, 82) = 2.82, p = .066,
so we may assume that smoking attitude for the three groups have equal
population variances.

Interpret the results:

In the sample, former smokers have a much more negative attitude towards
smoking (M = -1.69, SD = 1.71) than non-smokers (M = 0.64, SD = 1.17) and
smokers (M = 0.80, SD = 1.67). Average attitude towards smoking scores are
  significantly different, F (2, 82) = 18.85, p < .001. We have sufficient
evidence to conclude that one group has a more positive attitude towards
smoking than another group in the population of adults.
```

7.3 Different Score Levels for Two Factors

The participants in the experiment do not only differ because they see different endorsers in the charity video. In addition, there are differences according to sex: female versus male participants. Does participant's sex matter to the effect of the endorser on willingness to donate?

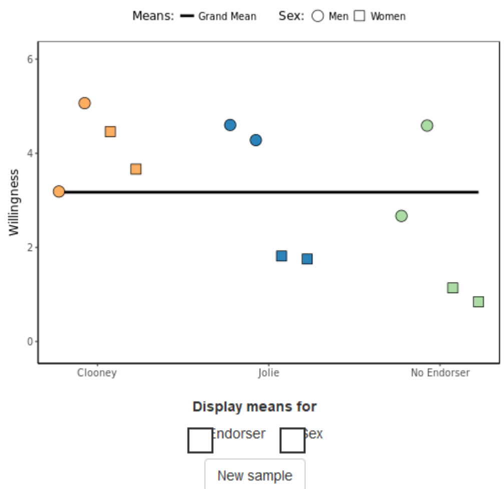


Figure 7.6: How do group means tell us about (main) effects in analysis of variance?

1. How does an analysis of variance test the effect of endorser on willingness to donate with the data displayed in Figure 7.6? Select the endorser factor to check your answer.

```
* Analysis of variance tests the variance (variation, spread) in the mean
scores on the outcome variable among groups. If the group means are more widely
apart, the variance of group means is larger, so the F test value is larger and
further away from what we expect if there are no differences among group means
in the population. The variation among group means is larger if they are more
distant from the overall (or grand) mean.
* To test the endorser effect, analysis of variance looks at the mean scores of
groups defined by the celebrity endorser they have seen. It will check the
distance between the three celebrity group means and the grand mean.
```

2. Which effect on willingness to donate is probably stronger: the effect of endorser or of sex? Motivate your answer, for example, using the grey arrows in the plot. Compare a plot with the endorser factor selected to a plot with the sex factor selected.

```
* The effect of endorser is probably stronger than the effect of sex because
the observations within an endorser group are less equally distributed around
the grand mean (they are more clustered above or below the grand mean) than
observations within a sex group. Both females and males are found nearly
equally above and below the grand mean, so their group means are close to the
```

```
grand mean, and there is little variation in sex group means.
* The grey arrows represent the distance between group mean and grand mean for
each observation. They are much shorter for the difference between females and
males than for differences between endorser groups.
```

3. Where do you expect the group means to show up in the graph if you select both the Endorser and Sex check boxes?

```
* If both boxes are ticked, the scores are grouped by the combination of
endorser and sex. The graph will show the means of subgroups: females exposed to
Clooney, males exposed to Clooney, males exposed to Jolie, and so on.
```

In the preceding section, we have looked at the effect of a single factor on willingness to donate, namely, the endorser to whom participants are exposed. Thus, we take into account two variables: one predictor and one outcome variable. This is an example of bivariate analysis.

Usually, however, we expect an outcome to depend on more than one variable. Willingness to donate does not depend only on the celebrity endorsing a fundraising campaign. It is easy to think of more factors, such as a person's available budget, her personal level of altruism, and so on.

It is straightforward to include more factors in an analysis of variance. These can be additional experimental treatments in the context of an experiment as well as participant characteristics that are not manipulated by the researcher. For example, we may hypothesize that females are generally more charitable than males.

7.3.1 Two-way analysis of variance

If we use one factor, the analysis is called one-way analysis of variance. With two factors, it is called two-way analysis of variance, and with three factors... well, you probably already guessed that name.

A two-way analysis of variance using a factor with three levels, for instance, exposure to three different endorsers, and a second factor with two levels, for example, female versus male, is called a 3x2 (say: three by two) factorial design.

7.3.2 Balanced design

In analysis of variance with two or more factors, it is quite nice if the factors are statistically independent from one another. In other words, it is nice if the scores on one factor are not associated with scores on another factor. This is called a balanced design.

In an experiment, we can ensure that factors are independent if we have the same number of participants in each combination of levels on all factors. In other words, a factorial design is balanced if we have the same number of observations in each subgroup. A subgroup contains

Table 7.1: Number of observations per subgroup in a balanced 3x2 factorial design.

	Female	Male
Clooney	2	2
Jolie	2	2
No endorser	2	2

the participants that have the same level on both factors just like a cell in a contingency table.

Table 7.1 shows an example of a balanced 3x2 factorial design. Each subgroup (cell) contains two participants (cases). If you remember the principles of statistical association and independence in contingency tables (Section 4.2.10), you know that equal distributions of frequencies across columns or across rows indicate statistical independence. In the example, the distributions are the same across columns (and rows), so the factors are statistically independent.

A balanced design is nice but not necessary. Unbalanced designs can be analyzed but estimation is more complicated (a problem for the computer, not for us) and the assumption of equal population variances for all groups (Levene's F test) is more important (a problem for us, not for the computer) because we do not have equal group sizes. Note that the requirement of equal group sizes applies to the subgroups in a two-way analysis of variance. With a balanced design, we ensure that we have the same number of observations in all subgroups, so we are on the safe side.

7.3.3 Main effects in two-way analysis of variance

A two-way analysis of variance tests the effects of both factors on the outcome variable in one go. It tests the null hypothesis that participants exposed to Clooney have the same average willingness to donate in the population as participants exposed to Jolie or those who are not exposed to an endorser. At the same time, it tests the null hypothesis that females and males have the same willingness to donate in the population.

The tested effects are *main effects*, that is, an overall or average difference between the mean scores of the groups on the outcome variable. The main effect of the endorser factor shows the mean differences for endorser groups if we do not distinguish between females and males. Likewise, the main effect for sex shows the average difference in willingness to donate between females and males without taking into account the endorser to whom they were exposed.

We could have used two separate one-way analyses of variance to test the same effects. Moreover, we could have tested the difference between females and males with an independent-samples t test. The results would have been the same (if the design is balanced.) But there is an important advantage to using a two-way analysis of variance, to which we turn in the next section.

7.4 Moderation: Score Level Differences that Depend on Context

In the preceding section, we have analyzed the effects both of endorser and sex on willingness to donate to a fund raiser. The two main effects isolate the influence of endorser on willingness from the effect of sex and the other way around. This assumes that endorser and sex have an effect on their own, a general effect.

We should, however, wonder whether endorser always has the same effect. Even if there is a general effect of endorser on willingness to donate, is this effect the same for females and males? Note that one endorser is a male celebrity who is reputed to be quite attractive to women. The other endorser is a female celebrity with a similar reputation among men. In this situation, shouldn't we expect that one endorser is more effective among female participants and the other among male participants?

If the effect of a factor is different for different groups on another factor, the first factor's effect is *moderated* by the second factor. The phenomenon that effects are moderated is called *moderation*.

With moderation, factors have a combined effect. The context (group score on one factor) affects the effect of the other factor on the outcome variable. The conceptual diagram for moderation expresses the effect of the moderator on the effect of the predictor as an arrow pointing at another arrow. Figure 7.7 shows the conceptual diagram for participant's sex moderating the effect of endorsing celebrity on willingness to donate.

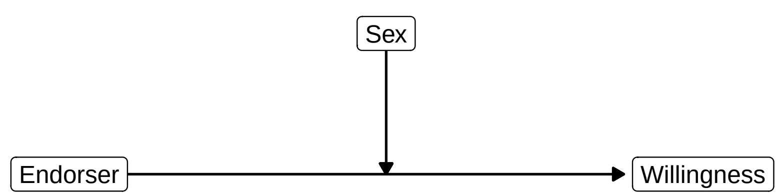
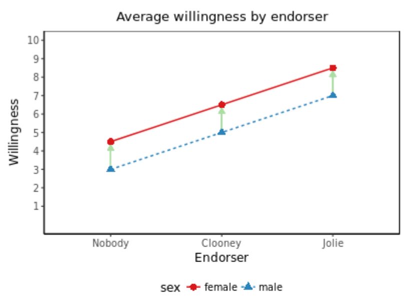


Figure 7.7: Conceptual diagram of moderation.

7.4.1 Types of moderation

Moderation as different effects for different groups is best interpreted using a cross-tabulation of group means, which is visualized as a means plot. In a group means tables, the **Totals** row and column contain the means for each factor separately, for example the means for males and females (factor sex) or the means for the endorsers (factor endorser). These means represent the main effects. The means in the cells of the table are the means of the subgroups, which represent moderation. Draw them in a means plot for easy interpretation.

In a means plot, we use the groups of the predictor on the horizontal axis, for example, the three endorsers. The average score on the outcome variable is used as the vertical axis. Finally, we plot the average scores for every predictor-moderator group, for instance, an endorser-sex combination, and we link the means that belong to the same moderator group, for example, the means for females and the means for males (Figure 7.8).



	Nobody	Clooney	Jolie	Total:
Women				6.5
Men				5
Total:	3.75	5.75	7.75	5.75

Figure 7.8: How can we recognize main effects and moderation in a means plot?

1. Does the plot in Figure 7.8 display a main effect of the factor sex? Motivate your answer.

```
* A main effect exists if the groups on one predictor (factor) have unequal
means.
* In the right-most column of the table, it is easy to see that females have a
higher average willingness score than males, so there is a main effect of sex.
* In the plot, the difference is more difficult to see. You have to guess the
average score of females over all endorsers (the average of the red line) and
the average score of all males (average of the blue line), to see that females
score higher than males on average.
* In this particular plot, the red line is everywhere above the blue line, so
it is relatively easy to see that females score on average higher than males.
But if the lines cross, it will be more difficult to see.
```

2. Is there a main effect of endorser? Again, motivate your answer.

```
* Now we have to look at the average scores for the three endorsers, which are
quite clearly different (bottom row of the table). So yes, there is a main
effect of endorser on willingness.
* In the plot, we have to guess the average score for an endorser, which is
```

```
somewhere in the middle of the green arrow linking the score of females to the
score of males. The average score with Clooney as endorser seems to be higher
than the average score without a celebrity endorser, and the average score for
Angelina Jolie is highest. These differences indicate a main effect for
endorser.
```

3. Explain why participant's sex does not moderate the effect of endorser in Figure 7.8.

```
* In this example, moderation implies that the differences between endorsers
for females are different from the differences between endorsers for males.
Different differences!
* The red line visualizes the differences between endorsers for females. The
blue line shows the differences between endorsers for males. If the two lines
are parallel, the differences are the same for males and females and there is
NO moderation.
* Another way to see this: Focus on the green arrows in the graph. If these arrows
have the same length and the same direction, the differences between males
(arrow origin) and females (arrow end) are the same for all endorsers. This
indicates that the effect of sex (difference between male and female within an
endorser) is the same for all endorsers, so endorser does not moderate the
effect of sex. This also implies the reverse conclusion that sex does not
moderate the effect of endorser because moderation is symmetrical.
```

4. Adjust the means in such a way that sex moderates the effect of endorser on willingness to donate.

```
* Change the means for the males or females such that the lines are no longer
parallel. Now sex moderates the effect of endorser on willingness to donate.
* If the lines are not parallel, one of two situations occurs:
 1. The means vary more (are more different) for females than for males (or
 the other way around). This means that the effect of endorser is stronger for
 females than for males (or the other way around).
 2. The effects are in the opposite direction. Females are more willing to
 donate than males if they see one endorser but males are more willing to
 donate than females if they see another endorser.
```

Moderation happens a lot in communication science for the simple reason that the effects of messages are stronger for people who are more susceptible to the message. If you know more people who have adopted a new product or a healthy/risky lifestyle, you are more likely to be persuaded by media campaigns to also adopt that product or lifestyle. If you are more impressionable in general, media messages are more effective.

7.4.1.1 Effect strength moderation

Moderation refers to contexts that strengthen or diminish the effect of, for instance, a media campaign. Let us refer to this type of moderation as *effect strength moderation*. In our current example, we would hypothesize that the effect of George Clooney as an endorser is

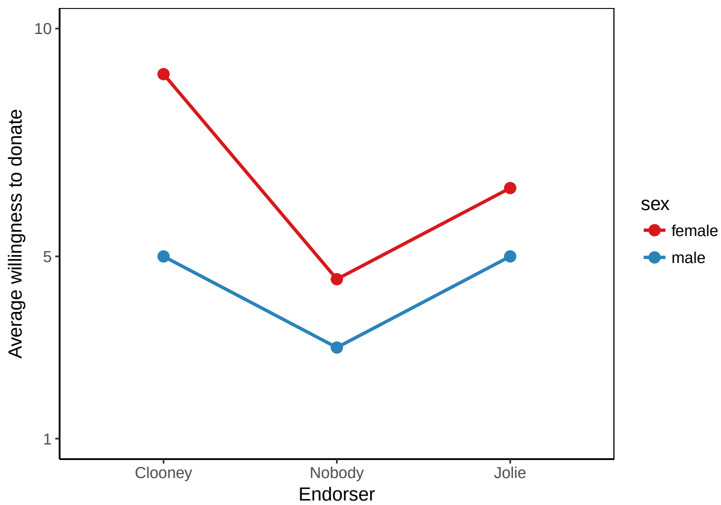


Figure 7.9: Moderation as a stronger effect within a particular context.

stronger for female participants than for male participants.

In analysis of variance, effects are differences between average outcome scores. The effect of Clooney on willingness to donate, for instance, is the difference between the average willingness score of participants exposed to Clooney and the average score of participants who were not exposed to a celebrity endorser.

Different "Clooney effects" for female and male participants imply different differences! The difference in average willingness scores between females exposed to Clooney and females who are not exposed to an endorser is different from the difference in average scores for males. We have four subgroups with average willingness scores that we have to compare. We have six subgroups if we also include endorsement by Angelina Jolie.

A means plot is a very convenient tool to interpret different differences. Connect the means of the subgroups by lines that belong to the same group on the factor you use as moderator. Each line in the plot represents the effect differences within one moderator group. If a line goes up or down, predictor groups have different means, so the predictor has an effect within that moderator group. A flat (horizontal) line tells us that there is no effect at all within that moderator group

The distances between the lines show the difference of the differences. If the lines for females and males are parallel, the difference between endorsers is the same for females and males. Then, the effects are the same and there is *no* moderation. In contrast, if the lines are not parallel but diverge or converge, the differences are different for females and males and there is moderation.

A special case of effect strength moderation is the situation in which the effect is absent (zero) in one context and present in another context. A trivial example would be the effect of an anti-smoking campaign on smoking frequency. For smokers (one context), smoking

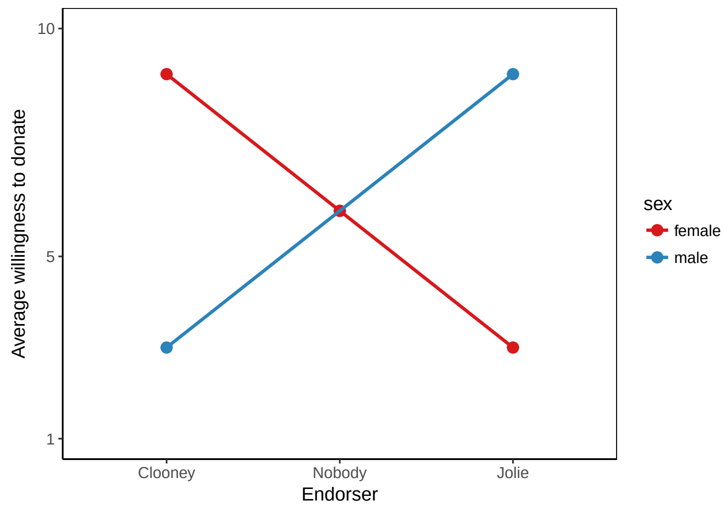


Figure 7.10: Moderation as opposite effects in different contexts.

frequency may go down with campaign exposure and the campaign may have an effect. For non-smokers (another context), smoking frequency cannot go down and the campaign cannot have this effect.

Except for trivial cases such as the effect of anti-smoking campaigns on non-smokers, it does not make sense to distinguish sharply between moderation in which the effect is strengthened and moderation in which the effect is present versus absent. In non-trivial cases, it is very rare that an effect is precisely zero.

7.4.1.2 Effect direction moderation

In the other type of moderation, the effect in one group is the opposite of the effect in another group. For example, Clooney is a more effective endorser among females than males whereas Jolie is more effective among males than females. Let us call this *effect direction moderation*. Females reverse the Jolie effect and males reverse the Clooney effect.

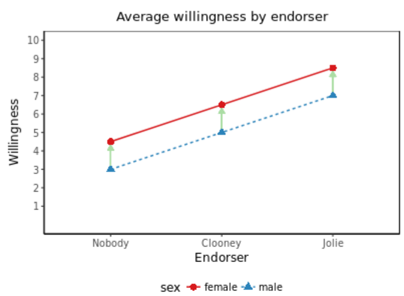
The effect in one group can compensate for the effect in another group if it is about as strong but of the opposite direction. Imagine that George Clooney convinces females to donate but discourages males to donate because his charms backfires on men (pure jealousy, perhaps.) Similarly, Angelina Jolie may have opposite effects on females and males.

In this situation, the main effect of endorser on willingness to donate is (nearly) zero. If we average over females and males, there is no net difference between Clooney, Jolie, and the condition without an endorser. This does not mean that the endorser does not matter. On the contrary, the interaction effects tell us that the endorser is effective for one group but counterproductive for another group. The second part of the conclusion is just as important as the first part. The campaign should avoid to decrease the willingness to donate among particular target groups.

7.4.2 Testing main and interaction effects

The effect of a single factor is called a main effect, as we learned in Section 7.3.3. A main effect reflects the difference between means for groups within one factor. The main effect of sex, for instance, can be that females are on average more willing to donate than males. A two-way analysis of variance includes two main effects, one for each factor (see Section 7.3.1), for example a man effect of sex and a main effect of endorser.

For moderation, however, we compare average scores of subgroups, that is, groups that combine a level on one factor and a level on another factor. In Figure 7.11, for instance, we compare average willingness to donate for combinations of endorser and participant's sex. The effect of differences among subgroups on the outcome variable is called an *interaction effect*. Just like a main effect, an interaction effect is tested with an F test and its effect size is expressed by eta^2.



Endorser effect: F(2, 54) = 28.96, p < .001, eta^2 = 0.47
Sex effect: F(1, 54) = 12.22, p < .001, eta^2 = 0.10
Interaction effect: F(2, 54) = 0.00, p = 1.000, eta^2 = 0.00

	Nobody	Clooney	Jolie	Total:
Women				6.5
Men				5
Total:	3.75	5.75	7.75	5.75

Figure 7.11: How can we recognize main effects and moderation in a means plot?

1. Adjust the means in Figure 7.11 in such a way that the main effect of endorser and the interaction effect of endorser and sex are statistically significant.

```
* If the app loads, there is a main effect of endorser. The effect size
expressed by eta2 is clearly larger than zero and the F test on the endorser
main effect is highly significant (p < .001).
* The main effect indicates that the endorser group averages (bottom row of the
```

```
table) are (clearly) different. We should maintain these differences if we
want to retain the main effect.
* The interaction effect is initially zero and not statistically significant.
This means that the differences between females and males (expressed by the
green arrows) do not vary (sufficiently) between the endorsers. In the initial
situation, the difference between females and males is exactly the same for
all endorsers, namely 1.5.
* Change the difference between males and females within one or more endorser
categories to create a sizable and statistically significant interaction
effect.
* But ensure that the total differences between the endorsers remain. For
example, increase the female scores for one endorser and lower the male scores
for that endorser such that the average for the endorser remains the same.
```

2. Adjust the means in such a way that the main effect of endorser is *not* statistically significant but the interaction effect of endorser and sex is statistically significant.

```
* To obtain a non-significant main effect of endorser, the mean scores for the
endorsers (bottom row of the table) must be more or less the same. So lower
the scores for females and males for endorsers with high average scores and
increase the scores for endorsers with low scores until the means in the
bottom row are equal.
* If you want to have an interaction effect, you must ensure that the
difference between females and males is clearly different for different
endorsers. For example, increase the male scores for one endorser and decrease
the female scores with the same amount.
```

3. Is it possible to have a statistically significant interaction effect but no statistically significant main effects? If so, adjust the scores in Figure 7.11 to prove your case.

```
* Yes, it is possible to have a statistically significant interaction effect
but no statistically significant main effects.
* Adjust the scores in the table such that both the row totals (average per
endorser) and the column totals (average per sex) are equal. To get the same
averages for males and females (rows) while preserving equal averages for
endorsers, you must increase or decrease the female (or male) scores equally
for all endorsers (columns).
* If you start with a table with equal average scores for all subgroups, you
have a situation with no main effects.
* You can obtain an interaction effect in this situation by changing the
average scores for females and males for one endorser, such that the endorser
average score remains the same and applying the opposite change to females and
males for another endorser.
* You will see that the red and blue lines cross.
```

Interpretation of moderation requires some training because we must abstract from main effects. The fact that females score on average higher than males is irrelevant to moderation but it does affect all subgroup mean scores. So the fact that the red line is above the blue line in Figure 7.11 is not relevant to moderation.

Moderation concerns the differences between subgroups that remain if we remove the overall differences between groups, that is, the differences that are captured by the main effects. The remaining differences between subgroup average scores provide us with a between-groups variance. In addition, the variation of outcome scores within subgroups yield a within-groups variance. Note that within-groups variance is not visible in Figure 7.11 because the willingness scores of the participants are not shown.

We can use the between-groups and within-groups variances to execute an F test just like the F test we use for main effects. The null hypothesis of the F test on an interaction effect states that the subgroups have the same population averages if we correct for the main effects. Our statistical software takes care of this correction.

Note that we must include the main effects in the model if we want to correct for them in our test of an interaction effect. If we would exclude main effects, we assume that there are no main effects. Why assume that there are no main effects if we can test their absence?

Moderation between three or more factors is possible. These are called *higher-order interactions*. It is wise to include both main effects and lower-order interactions if we test a higher-order interaction. As a result, our model becomes very complicated and hard to interpret. If a (first-order) interaction between two predictors must be interpreted as different differences, an interaction between three factors must be interpreted as different differences in differences. That's difficult to imagine, so let us avoid them.

7.4.3 Assumptions for two-way analysis of variance

The assumptions for a two-way analysis of variance are the same as for a one-way analysis of variance (Section 7.1.4). Just note that equal group sizes and equal population variances now apply to the subgroups formed by the combination of the two factors.

7.5 Reporting Two-Way Analysis of Variance

The main purpose of reporting a two-way analysis of variance is to show the reader the differences between average outcome scores between groups on the same factor (main effects) and different differences for groups on a second factor (interaction effect). A means plot is very suitable for this purpose. Conventionally, we place the predictor groups on the horizontal axis and we draw different lines for the moderator groups. But you can switch them if it produces a more appealing graph.

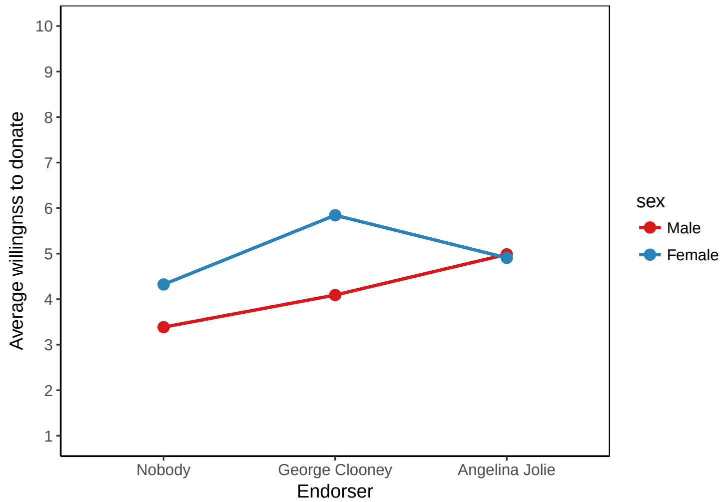


Figure 7.12: An example of a means plot.

For the statistically informed reader, you should include the following information somewhere in your report:

- That you used analysis of variance and the analysis of variance type (one-way or two-way).
- The test result for every effect, consisting of the test name (F), the degrees of freedom, and the significance (p value). APA6 prescribes the following format if you report the test result within your text: F (df1, df2) = F value, p = p value. Note that df1 is the degrees of freedom of the factor and df is the degrees of freedom of the error (within-groups).
- For each effect report eta-squared (eta^2) and interpret it in terms of effect size. If you have to calculate eta-squared by hand, divide the between-groups sum of squares of an effect by the total sum of squares (SPSS: corrected total).
- For each effect worth interpretation, clarify which group or subgroup scores higher. Report the group means and their standard deviations or the mean difference (from the post-hoc tests) for comparisons between groups as well as their p values here.
- As always, don't forget to mention the units (cases) and the meaning of the variables (factors and outcome). They describe the topic of the analysis.
- Report it if the main assumption is violated, that is, if you have (sub)groups of unequal size and the test on homogeneous variances is statistically significant. Report Levene's test just like you report the F test of a main effect (see above). If the assumption is violated, we still report and interpret the results of the analysis of variance but we warn that the results may not be trustworthy.

In a two-way analysis of analysis, the number of numeric results can be large. It is

Table 7.2: An example of a table summarizing results of a two-way analysis of variance.

	Sum of Squares	df	Mean Square	F	p
sex	26.37	1.00	26.37	11.86	0.001
endorser	38.05	2.00	19.03	8.56	< 0.001
endorser*sex	20.60	2.00	10.30	4.63	0.011
error	304.54	137.00	2.22		
Total	142.00	57.92			

recommended to present them as a table (in the text or in an appendix). If you report the table, include the error, the sums of squares and mean squares in the same way that SPSS reports them. Table 7.2 presents an example.

7.6 Two-Way Analysis of Variance in SPSS

7.6.1 Instructions

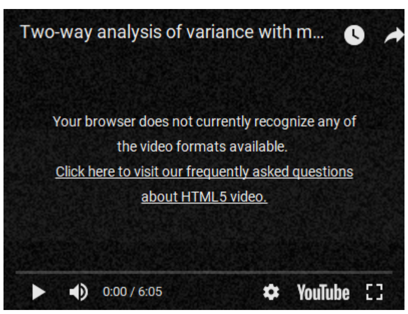


Figure 7.13: Two-way analysis of variance with moderation in SPSS.

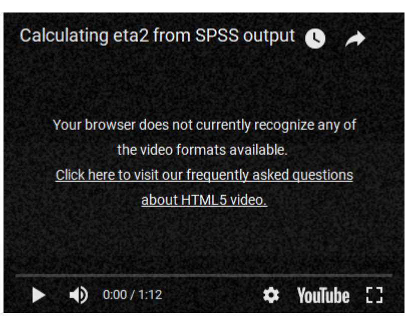


Figure 7.14: Calculating eta^2 from SPSS output.

7.6.2 Exercises

1. Use the data in donors.sav to test if the effect of celebrity endorsement on adults' willingness to donate differs for females and males. Check the assumptions and interpret the results. Create a plot to communicate your results.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=willing_post endorser sex
  /ORDER=ANALYSIS.
* Two-way analysis of variance.
UNIANOVA willing_post BY endorser sex
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /POSTHOC=endorser(BONFERRONI)
  /PLOT=PROFILE(endorser*sex)
  /PRINT=HOMOGENEITY DESCRIPTIVE
  /CRITERIA=ALPHA(.05)
  /DESIGN=endorser sex endorser*sex.

Check data:

There are no impossible values on the three variables.

Check assumptions:

All six subgroups are of nearly equal size and/or the Levene test on
homogeneity of variances is not statistically significant, F (5, 137) = 0.43,
p = .825, so the most important assumptions are met.

Interpret the results:

The question is about moderation, so let us focus our interpretation on the
interaction effect.

* Our two-way analysis of variance shows a statistically significant
interaction effect of endorser and sex of the participant, F (2, 137) = 4.63, p =
.011. We may conclude that the effect of the endorsing celebrity on an
adult's willingness to donate is different for females and males.
* Females who saw George Clooney are more strongly willing to donate (M =
5.84, SD = 1.48) than males (M = 4.09, SD = 1.30). In contrast, there is
hardly any difference in willingness to donate between males (M = 4.99, SD =
1.77) and females (M = 4.91, SD = 1.51) who see Angelina Jolie endorse the
fund raiser.
* This result suggests that females are more sensitive to George Clooney as
endorser than males but the opposite does not apply to Angelina Jolie as
```

```
endorser.
```

2. Use the same data as in Exercise 1 to test if the effect of celebrity endorsement on willingness to donate depends on remembrance of the campaign. Check the assumptions and interpret the results. Again, create a plot to communicate your results.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=willing_post endorser remember
  /ORDER=ANALYSIS.
* Two-way analysis of variance.
UNIANOVA willing_post BY endorser remember
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /POSTHOC=endorser(BONFERRONI)
  /PLOT=PROFILE(endorser*remember)
  /PRINT=HOMOGENEITY DESCRIPTIVE
  /CRITERIA=ALPHA(.05)
  /DESIGN=endorser remember endorser*remember.

Check data:

There are no impossible values on the three variables.

Check assumptions:

The six subggroups are not nearly of equal size but the Levene test on
homogeneity of variances is not statistically significant, F (5, 137) = 0.06,
p = .997, so the most important assumptions are met.

Interpret the results for the interaction effect:

* For adults who do not remember the campaign, the effect of seeing Angelina
Jolie endorse the fund raiser seems to be relatively low in comparison to the
effect for adults who do remember the campaign. This can be seen in the means
plot.
* However, a two-way analysis of variance tells us that the interaction effect
of endorser with campaign remembrance on willingness to donate is very weak
and not statistically significant, F (2, 137) = 0.67, p = .514, eta2 = .01.
* Both the endorsing celebrity and campaign remembrance seem to have
statistically significant effects on willingness to donate, but the effect of
celebrity endorser on willingness to donate is not clearly different for
adults who remember the campaign and those who do not remember the campaign.

Note that eta2 was calculated by hand, dividing the sum of squares of the
interaction effect by the sum of squares of the corrected total.
```

3. Data set smokers.sav contains information about smoking on a random sample of adults. Does the attitude towards smoking depend on the adult's smoking behaviour (smoker, former smoker, or non-smoker) and on exposure to an anti-smoking campaign? Recode exposure scores into three groups: low exposure (scores 3 or less), medium exposure (scores 3 to 7), and high exposure (scores above 7).

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=status3 exposure attitude
  /ORDER=ANALYSIS.
* Group exposure to anti-smoking campaign.
RECODE exposure (Lowest thru 3=1) (3 thru 7 = 2) (ELSE=3) INTO exposure3.
VARIABLE LABELS  exposure3 'Exposure to anti-smoking campaign'.
EXECUTE.
* Define Variable Properties.
*exposure3.
VALUE LABELS exposure3
  1.00 'Low exposure'
  2.00 'Medium exposure'
  3.00 'High exposure'.
EXECUTE.
* Two-way analysis of variance.
UNIANOVA attitude BY status3 exposure3
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /POSTHOC=status3 exposure3(BONFERRONI)
  /PLOT=PROFILE(status3*exposure3)
  /PRINT=HOMOGENEITY DESCRIPTIVE
  /CRITERIA=ALPHA(.05)
  /DESIGN=status3 exposure3 status3*exposure3.

Check data:

All values on the variables seem to be valid.

Check assumptions:

* The six subgroups (defined by combinations of smoking status and exposure)
are definitely not of equal size. The test on homogeneity of variances is
statistically significant, F (8, 76) = 2.17, p = .039.
* We are quite confident that the variance of smoking attitude in the
population is unequal for all subgroups.
* We do not meet the assumptions, so we should report this as a disclaimer
with our results.
```

```
Interpret the results:

Because we are going to interpret all effects, it is recommended to present
the table with between-participants effects instead of reporting all F test
results in the text.

* Exposure clearly matters to smoking attitude according to our two-way
analysis of variance. High exposure is associated with a more negative
attitude towards smoking (M = -0.52, SD = 1.26) than medium exposure (M =
0.18, SD = 1.95, difference: p = .024), which is more negative than the
attitude for adults with low exposure (M = 0.99, SD = 1.34, difference: p =
.006). Exposure accounts for 21 per cent of the variation in attitude towards
smoking.

* Smoking status also matters for attitude towards smoking.
* Former smokers (M = -1.70, SD = 1.71) are significantly more negative
towards smoking then smokers (M = 0.80, SD = 1.67, difference: p < .001) and
non-smokers (M = 0.64, SD = 1.17, difference: p < .001).
* The difference between the latter two groups is not statistically
significant (p = 1.000). Smoking status accounts for 25 per cent of the
differences in smoking attitude.

* There is a weak (eta2 = 0.13) and statistically significant interaction
effect of smoking status with exposure on attitude towards smoking. If we
inspect the means plot,
* we see that the effect of exposure on attitude is moderated by smoking
status. Former smokers with low exposure are much more positive (less
negative) about smoking than we would expect. Or, in other words, medium and
high exposure to the campaign decreases their attitude more than the attitudes
of smokers or non-smokers.

Note that the proportions of explained variance (eta2) have been calculated
manually.
```

7.7 Test Your Understanding

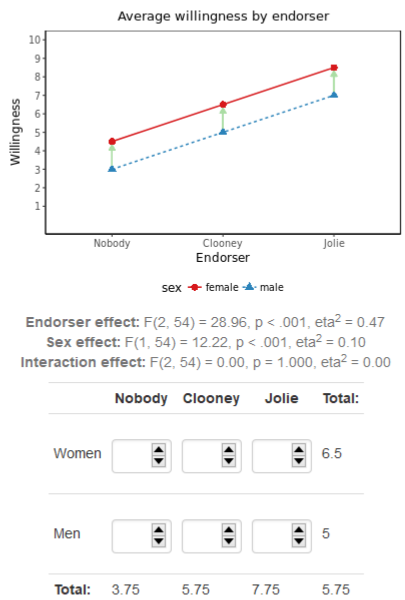


Figure 7.15: How do we recognize main effects and interaction effects in a means plot and in a table of means?

1. What is or are the main effects in Figure 7.15? Explain how you can recognize main effects both in the means plot and in the reported numbers.
2. Change some group means to create substantial and statistically significant moderation of the effect of endorser (Nobody, Clooney, or Jolie) on willingness.
3. What is the null hypothesis of the F tests reported in Figure 7.15?
4. How should we interpret the eta^2 values in Figure 7.15?

7.8 Take-Home Points

- In analysis of variance, we test the substantive null hypothesis that all groups have the same population means. Behind the scenes, we actually test the ratio of between-groups variance to within-groups variance.
- The overall differences in average outcome scores between groups on one factor (predictor) are a main effect in an analysis of variance.

- The differences in average outcome scores between subgroups, that is, groups that combine a level on one factor and a level on another factor, represent an interaction effect. Note that we are dealing with the differences between subgroup scores that remain after the main effects have been removed.
- Moderation is the phenomenon that an effect is different in different contexts. The effect can be stronger or it can have a different direction. In analysis of variance, interaction effects represent moderation.
- Eta-squared measures the size of a main or interaction effect in analysis of variance. It tells us the proportion of variance in the outcome variable that is accounted for by the effect.
- A means plot is very helpful for interpreting and communicating results of an analysis of variance.
- The F tests in analysis of variance do not tell us which groups have different average outcome scores. To this end, we use independent-samples t tests as post-hoc tests with a (Bonferroni) correction for capitalization on chance.
- To apply analysis of variance, we need a numeric outcome variable that has equal population variance in each group of a factor or each subgroup in case of an interaction effect. However, equality of population variances is not important if all groups on a factor or all subgroups in an interaction are more or less of equal size (the largest count is at most 10% of the largest count larger than the smallest count.)

Chapter 8

Moderation with Regression Analysis

> Key concepts: interaction variable, covariate, outcome, regression equation, dummy variables, normally distributed residuals, linearity, homoscedasticity, independent observations, statistical diagram, common support, simple slope, conditional effect, mean-centering.

Summary

> My outcome is numeric but at least one predictor is also numeric, so I cannot apply analysis of variance. How can I investigate moderation with regression analysis?

The linear regression model is a powerful and very popular tool for predicting a numeric outcome variable from one or more predictor variables. In this chapter, we use regression to evaluate the effects of an anti-smoking campaign. We predict attitude towards smoking from exposure to the anti-smoking campaign (numeric), time spent with smokers (numeric), and the respondent's smoking status (categorical).

Regression coefficients, that is, the slopes of regression lines, are the effects in a regression model. They show the predicted difference in the outcome for a one unit difference in the predictor (exposure, time spent with smokers) or the predicted mean difference for two categories (smokers versus non-smokers).

But what if the predictive effect is not the same in all contexts? For example, exposure to an anti-smoking campaign may generally generate a more negative attitude towards smoking. The effect, however, is probably different for people who smoke than for people who do not smoke. In this case, the effect of campaign exposure on attitude towards smoking is moderated by context: Whether or not the person exposed to the campaign is a smoker.

Different effect sizes for different contexts are different regression coefficients for different contexts. We need different regression lines for different groups of people. We can use an interaction variable as a predictor in a regression model to accommodate for moderation as different slopes. An interaction variable is just the product of the predictor and moderator variables.

As a predictor in the model, an interaction variable has a confidence interval and a p value. The confidence interval tells us the plausible values for the size of the interaction effect in the population. The p value tests the null hypothesis that there is no interaction effect at all in the population.

To interpret the interaction effect, we must determine the size of the effect of the predictor on the outcome variable for several interesting values of the moderator. If the moderator is categorical, we want to know the effect (simple slope) within each category of the moderator. For example, the effect of campaign exposure on smoking attitude for smokers and the effect for non-smokers.

The moderator can also be a continuous variable, for example, the time a person spends with people who smoke. In this case, we may look at the effect for the mean value of the moderator (moderate score level) and one standard deviation below (low level) or above (high level) the mean.

An interaction effect in a regression model closely resembles an interaction effect in analysis of variance. In contrast, the effect of a single predictor in a regression model is not a main effect as in analysis of variance. It is a conditional effect, namely the effect for one particular value of the moderator (context). To understand this, we must pay close attention to the regression equation.

8.1 The Regression Equation

In the social sciences, we usually expect that a particular outcome has several causes. Investigating the effects of an anti-smoking campaign, for instance, we would not assume that a person's attitude towards smoking depends only on exposure to a particular anti-smoking campaign. It is easy to think of other and perhaps more influential causes such as personal smoking status, contact with people who do or do not smoke, susceptibility to addiction, and so on.

Figure 8.1 summarizes some hypothesized causes of the attitude towards smoking. A regression model translates this conceptual diagram into a statistical model. The statistical regression model is a mathematical function with the outcome variable (also known as the dependent variable, usually referred to with the letter y) as the sum of a constant (a), the effects (b) of predictors (x), which are *predictive effects*, and an error term (e), which is also called the *residuals*, see Equation (8.1).

$$y = a + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + e \tag{8.1}$$

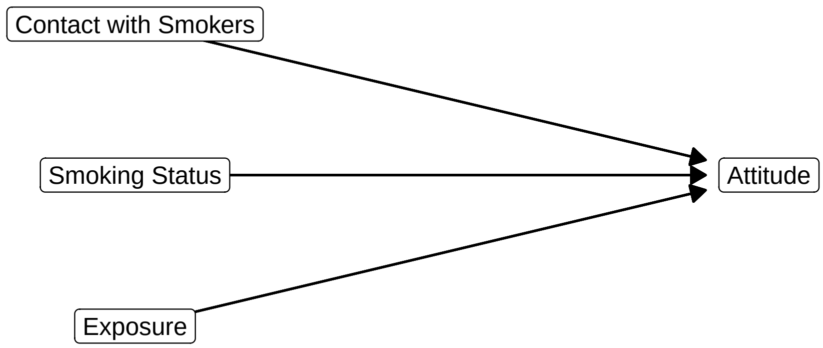


Figure 8.1: A conceptual model with some hypothesized causes of attitude towards smoking.

8.1.1 Interpretation of a regression equation

Let us first have a close look at a simple regression equation, that is, a regression equation with just one predictor (x).

1. How does the plot visualize the constant of the regression equation?

```
* The constant of a regression equation, represented by the symbol a, is the
predicted value of the outcome variable if all predictors are zero.
* Graphically, this is where the regression line cuts the vertical (y) axis.
* In the plot, a is 1.4 according to the equations and the regression line cuts
the vertical axis at 1.4 as it should.
```

2. Explain how the horizontal and vertical lines in the plot help to interpret the unstandardized regression coefficient b.

```
* An unstandardized regression coefficient, denoted by the symbol b, tells us
the predicted difference in the outcome for a difference of one unit in the
predictor variable.
* According to the equations, the predicted attitude decreases by 0.25 for a one
unit difference (4 to 5) in exposure.
* This is the decrease from 0.40 to 0.15 signalled by the horizontal line
segments.
```

Good understanding of the regression equation is necessary for understanding moderation in regression models. So let us have a close look at an example equation (Eq. (8.2)). The outcome variable attitude towards smoking is predicted from a constant and three predictor variables.

$$attitude = constant + b_1 * exposure + b_2 * status + b_3 * contact + e \tag{8.2}$$

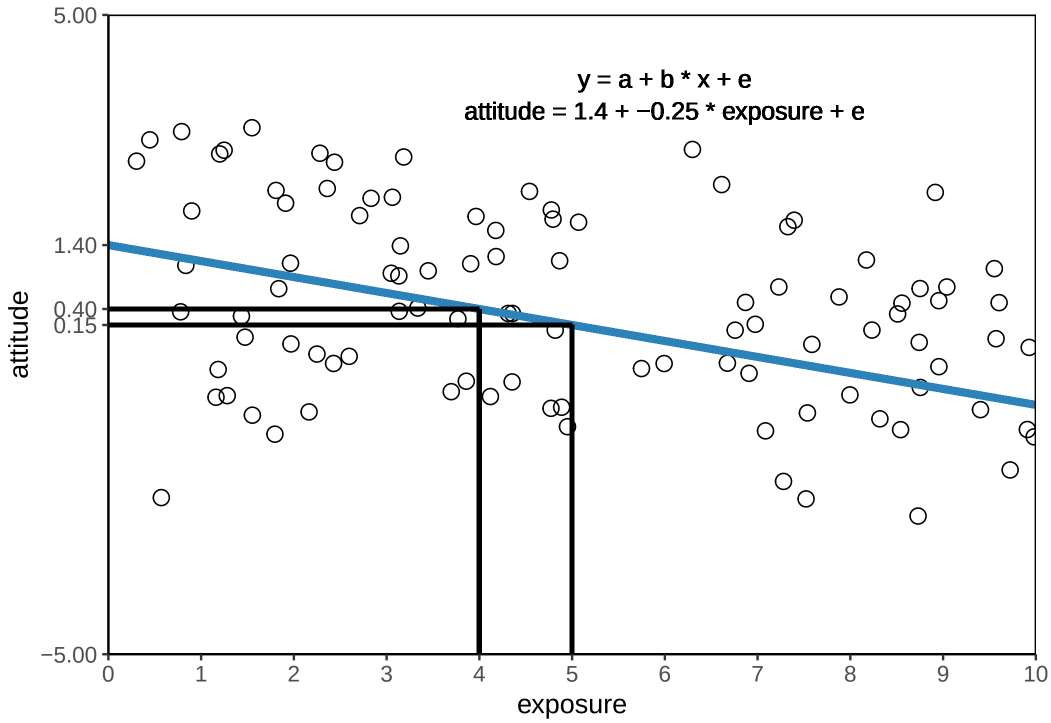


Figure 8.2: What is the meaning of the regression equation?

The constant adds a fixed quantity to the predicted attitude for all participants. It adjusts the overall level of the predicted attitude.

More precisely, the constant is the predicted attitude if a person scores zero on all predictor variables. To see this, plug in zero for all predictors in the equation (Eq. (8.3)) and remember that zero times something is zero. This reduces the equation to the constant and the *error term* e. The error term is the error of our prediction, also known as the *residual*. It does not help to predict the outcome, so the constant is the only remaining predictor.

$$\begin{aligned}
attitude &= constant + b_1 * 0 + b_2 * 0 + b_3 * 0 + e \\
attitude &= constant + 0 + 0 + 0 + e \\
attitude &= constant + e
\end{aligned} \tag{8.3}$$

For all persons scoring zero on exposure, smoking status, and contact with smokers, the predicted attitude equals the value of the regression constant. This interpretation only makes sense if the predictors can be zero. If they include, for instance, scales ranging from one to seven, there are no persons with zero scores on all predictors and the constant has no meaning.

The regression coefficients b represent the predicted difference in the outcome for a difference of one unit in the predictor. For example, plug in the values 5 and 4 for the *status* predictor in the equation. If we take the difference of the two equations, we are left with b_1. All other terms in the two equations cancel out (except, perhaps, the error term e).

$$\begin{array}{rr}
 & attitude = constant + b_1 * 5 + b_2 * status + b_3 * contact + e \\
- & \underline{attitude = constant + b_1 * 4 + b_2 * status + b_3 * contact + e} \\
 & attitude\,difference = b_1 * 5 - b_1 * 4 = b_1 * (5 - 4) = b_1
\end{array} \tag{8.4}$$

We will be plugging in values for predictors in the regression equation a lot in this chapter. It is necessary for understanding and interpreting moderation.

8.1.2 Continuous predictors

In a linear regression, the outcome variable (y) must be numeric and in principle continuous. There are regression models for other types of outcomes, for instance, logistic regression for a dichotomous (0/1) outcome and Poisson regression for a count outcome, but we will not discuss them.

The predictor variables must be either numeric or dichotomous. If exposure is measured as a scale, for instance ranging from zero to ten, the interpretation of the effect of exposure (b_1) is the one that we have encountered in the preceding section: the predicted difference in attitude (outcome variable) for a one unit difference in exposure (predictor variable) while all other predictor values do not change (are held constant).

Whether this predicted difference is small or large depends on the practical context: Is a small decrease in attitude towards smoking worth the effort of the campaign? If we want to apply a rule of thumb for the strength of the effect, we usually look at the standardized regression coefficient (b^* according to APA6, *Beta* in SPSS output). See Section 5.2.5 for some rules of thumb for effect size interpretation.

Note that the regression coefficient is calculated for the predictor values that occur within the data set. For example, if sample exposure scores are within the range three to seven, these values are used to predict attitude towards smoking.

We cannot see this in the regression equation, which allows us to plug in -10, 0, or 100 as exposure values. But the values for attitude that we predict from these exposure values are probably nonsensical because our data do not tell us anything about the relation between exposure and anti-smoking attitude for predictor values outside the three to seven range. We should not pretend to know the effects of exposure levels outside this range. It is good practice to check the actual range of predictor values.

8.1.3 Dichotomous predictors

Instead of a numeric predictor, we can use a dichotomy as a predictor in a regression model. The dichotomy is preferably coded as 1 versus 0, for example, 1 for smokers and 0 for non-smokers among our respondents.

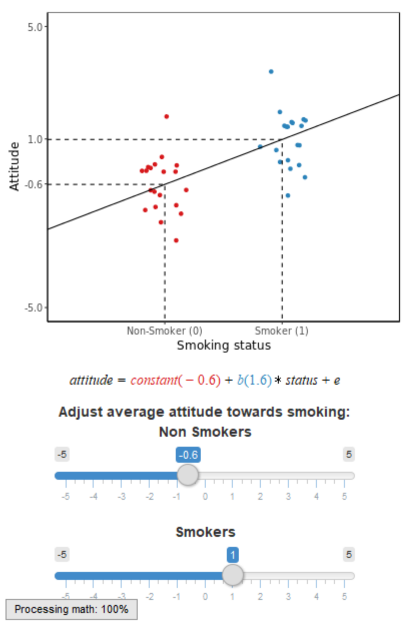


Figure 8.3: What is the difference in attitude between non-smokers and smokers?

1. What is the relation between the constant of the regression line in Figure 8.3 and group

averages? Motivate your answer by changing the average attitude towards smoking for non-smokers in Figure 8.3.

```
* The constant is equal to the average attitude of the non-smokers.
* Use the slider to change the average attitude towards smoking for
non-smokers.
* This will change the value of the constant in the regression equation. Why?
* Non-smokers score 0 on the (smoking) status variable. The regression
equation for non-smokers, then, is:
* attitude = constant + b * status = constant + b * 0 = constant
* Thus, we see that the predicted attitude for non-smokers equals the
constant. In addition, we know that the predicted value for a group
equals the average score of the group. As a result, the constant equals
the average score of non-smokers in this example.
* Note that this is true only if the group is coded zero and if the regression
model contains only one predictor (simple regression model).
```

2. Change the group means to detect the relation between group means and the unstandardized regression coefficient (b). How can we calculate the unstandardized regression coefficient (b) from the group averages? Show that your answer is correct by changing average attitude towards smoking with the slider for smokers.

```
* In this example, the unstandardized regression coefficient (b) is equal to
the average attitude score of smokers minus the average attitude score of
non-smokers.
* This is so because smokers are coded as ones and non-smokers are coded as
zeros. The difference between smokers and non-smokers on the smoking status
variable is one. The regression coefficient (b) tells us the difference in
predicted attitude scores for a difference of one unit on the predictor
variable. As a result, the unstandardized regression coefficient (b) tells us
the difference between the average (= predicted value) for smokers and the
average (= predicted value) for non-smokers.
* With equations:
* Smokers average: attitude = constant + b * status = constant + b * 1 =
  = constant + b
* Non-smokers average: attitude = constant + b * status = constant + b * 0 =
  = constant
* Smokers - Non-smokers = (constant + b) - constant = b
```

Apart from numeric predictors, we can use dichotomous predictors, that is, predictors with only two values, which are preferably coded as 0 and 1 (*dummy coding*). The interpretation of the effect of a dichotomous predictor in a regression model is quite different from the interpretation of a numeric predictor.

For example, let us assume that smoking status is coded as smoker (1) versus non-smoker (0). Because this predictor can only take two values, we effectively have two versions of the regression equation. The first equation (8.5) represents all smokers, so their smoking status score is 1. The smoking status of this group has a fixed contribution to the predicted average attitude, namely b_2.

$$\begin{aligned} attitude &= constant + b_1 * exposure + b_2 * status + b_3 * contact + e \\ attitude &= constant + b_1 * exposure + b_2 * 1 + b_3 * contact + e \\ attitude &= constant + b_1 * exposure + b_2 + b_3 * contact + e \end{aligned} \tag{8.5}$$

Regression equation (8.6) represents all non-smokers. Their smoking status score is 0, so the smoking status effect drops from the model.

$$\begin{aligned} attitude &= constant + b_1 * exposure + b_2 * status + b_3 * contact + e \\ attitude &= constant + b_1 * exposure + b_2 * 0 + b_3 * contact + e \\ attitude &= constant + b_1 * exposure + b_3 * contact + e \end{aligned} \tag{8.6}$$

It does not make sense to interpret the regression coefficient of smoking status (b_2) as predicted difference in attitude for a difference of one 'more' smoking status. After all, the 0 and 1 scores do not mean that there is one unit 'more' smoking. Instead, the coefficient indicates that we are dealing with different groups: smokers versus non-smokers. We have to interpret the effect as a difference between two groups. More specifically, as the mean difference between the attitude of the group represented by the score 1 and the attitude of the *reference group* represented by score 0.

If you compare the final equations for smokers (Eq. (8.5)) and non-smokers (Eq. (8.6)), the only difference is b_2, which is present for smokers but absent for non-smokers. It is the difference between the average outcome score (attitude) for smokers and the average outcome score (attitude) non-smokers. This is exactly the same as in an independent-samples t test!

Imagine that b_2 equals 1.6. This indicates that the average attitude towards smoking among smokers is 1.6 units above the average attitude among non-smokers. Is this a small or large effect? In the case of a dichotomous predictor, we should **not** use the standardized regression coefficient to evaluate effect size. The standardized coefficient depends on the distribution of 1s and 0s, that is, which part of the respondents are smokers. But this should be irrelevant to the size of the effect.

Therefore, it is recommended to interpret only the unstandardized regression coefficient for a dichotomous predictor. Interpret it as the difference in average outcome scores for two groups as we have done in the preceding paragraph.

8.1.4 A categorical predictor and dummy variables

How about a categorical variable, that is, a variable containing three or more groups, for example, the distinction between respondents who smoke, former smokers, and respondents who never smoked? Can we include a categorical variable as a predictor in a regression model? Yes, we can but we need a trick.

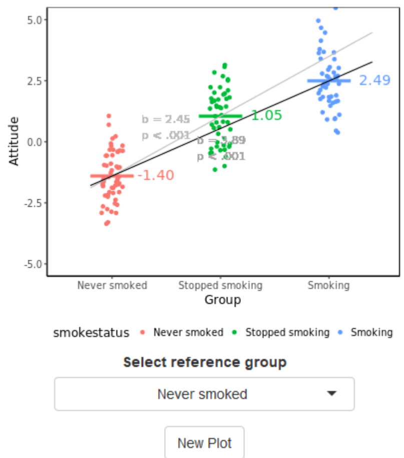


Figure 8.4: What are the predictive effects of smoking status?

1. Interpret the effects of smoking status in Figure 8.4.

```
* The precise answer to this question obviously depends on the group means in
your plot.
* In general, the interpretation focuses on differences between the mean score
of the reference group and the mean scores of the other groups. In this
example, we are talking about average attitude towards smoking for each group
defined by their smoking status. The reference group is selected with the
Select reference drop-down list.
* The unstandardized regression coefficient (b) tells us how much larger
(positive coefficient) or smaller (negative coefficient) the mean score of a
group is in comparison to the reference group.
* The associated p value tells us how uncertain we are that there truly is a
mean difference in the population.
```

2. In the initial state of Figure 8.4, can you tell whether the attitude of smokers is significantly different from the attitude of former smokers? Select a different reference group to motivate your answer.

```
* In the initial state of this figure, people who never smoked are the
reference group. The p values, then, are associated with differences between on
the one hand people who never smoked and on the other hand people who stopped
smoking or are smoking. The comparison between the latter two is not included.
* It is hazardous to derive the p value of the difference between two
non-reference groups from the p values of the differences between these groups
and the reference group. Remember that a p value depends on both effect size
(difference between group means) and on the standard error. The latter depends
on the variance of the outcome variable per group (in the population) and on
sample size per group.
* The solution is to re-estimate the regression model with one of the groups
```

Table 8.1: Dummy variables for a categorical predictor: One dummy variable is superfluous.

Original categorical variable:	neversmoked	smokesnomore	smoking
1 - Never smoked	1	0	0
2 - Former smoker	0	1	0
3 - Smoker	0	0	1

```
that you want to compare as reference group. In this figure, you can do this by
selecting a different group in the drop-down list.
```

3. Select some new plots. For each plot, determine which reference group you think is most convenient for summarizing the results.

```
* There is not one right way of choosing a reference group; think about
arguments.
* 1. Substantive interest: Does your research focus on one particular group? If
so, use this group as the reference group, so it is included in all
comparisons. If, for example, the research is meant to support an anti-smoking
campaign, the group of current smokers is probably of central interest. Make
them your reference group.
* 2. If you expect a particular order in the group means, the group that you
expect to be in the middle is a good choice as reference group. If, for
example, you expect that attitude towards smoking is more positive for smokers
than for former smokers, and the latter are more positive than people who never
smoked, the former smokers are expected to be in the middle. If we use them as
reference group, we can test if they are more positive than people who never
smoked, and more negative than people who are currently smoking.
* 3. If two groups have relatively similar means in comparison to the third
group, you may be interested to know if the relatively small difference is
statistically significant. In this case, one of the two groups that have
similar means is a good reference group.
```

What if smoking status was measured with three categories: (1) never smoked, (2) have smoked, (3) currently smoking? We can only include such a categorical predictor if we change it into a set of dichotomies.

A *categorical variable* contains three or more categories or groups. We can create a new dichotomous variable for each group, indicating whether (score 1) or not (score 0) the respondent is part of this group. In the example, we could create the variables *neversmoked*, *smokesnomore*, and *smoking*. Every respondent would score 1 on one of the three variables and 0 on the other two variables. These variables are called *dummy variables* or *indicator variables*.

If we want to include a categorical predictor in a regression model, we must use all dummy variables as predictors **except one**. In the example, we must include two out of the three dummy variables. We cannot include all three dummy variables because the score on the third dummy variable is determined by the score on the first two dummy variables.

If a respondent scores 1 on either the Never Smoked or the Former Smoker dummy variables, she cannot be a smoker so her score must be 0 on the Smoker dummy variable. Reversely, someone who is not a member of the Never Smoked or Smokes No More Groups, must be a member of the Smoking Group. A person scoring 0 on the first two dummy variables, then, must score 1 on the third.

We cannot include a dummy variable as a predictor that is perfectly predictable from other dummy variables in the regression model. It is like including the same predictor twice: How can the estimation process decide which predictor is responsible for the effect? It can't decide, so the estimation process fails and no regression coefficients are estimated. If this happens, the predictors are said to be perfectly *multicollinear*.

The category or group that is left out of the regression model is the *reference group*. Remember that non-smokers were the reference group in the preceding section because the regression equation did not include a dichotomous predictor on which the non-smokers scored 1. The two groups were represented by only one dichotomy. As you see, we used one dummy less than the number of groups (two).

The interpretation of the effects (regression coefficients) for the included dummies is the same as for a single dichotomous predictor such as smoker versus non-smoker. It is the difference between average outcome score of the group scoring 1 on the dummy variable and the average outcome score of the reference group.

If we exclude the dummy variable for the respondents who never smoked, the regression weight of the dummy variable for the Former Smoker Group gives the average difference between former smokers and non-smokers. If the regression weight is negative, for instance -0.8, former smokers have a more negative attitude towards smoking than non-smokers. If the difference is positive, former smokers have a more positive attitude towards smoking.

Which group should we use as reference category, that is, which dummy should not be used in the regression model? This is hard to say in general. If one group is of greatest interest to us, we could use this as the reference group, so all dummy variable effects express differences with this group. Alternatively, if we expect a particular ranking in average outcome scores, we may pick the group at the highest, lowest or middle rank as the reference group. If you can't decide, run the regression model several times with a different reference group.

8.1.5 Sampling distributions and assumptions

If we are working with a random sample or we have other reasons to believe that our data could have been different due to chance (Section 6.3), we should not just interpret the outcomes for the data set that we collected. We should apply statistical inference—confidence intervals and significance tests—to our results. The confidence interval gives us bounds for the population value of the unstandardized regression coefficient. The p value is used to test the *null hypothesis that the unstandardized regression coefficient is zero in the population.*

Each regression coefficient as well as the constant may vary from sample to sample drawn from the same population, so we should devise a sampling distribution for each of them. These sampling distributions happen to have a t distribution under particular assumptions.

Chapters 3 and 4 have extensively discussed how confidence intervals and p values are constructed and how they must be interpreted. So we may as well focus now on the assumptions under which the t distribution is a good approximation of the sampling distribution of a regression coefficient.

8.1.5.1 Independent observations

The two most important assumptions require that the observations are *independent and identically distributed*. These requirements arise from probability theory. If they are violated, the statistical results should not be trusted.

Each observation, for instance, a measurement on a respondent, must be independent of all other observations. This respondent's outcome variable score is not allowed to depend on outcome scores of other respondents.

It is hardly possible to check that our observations are independent. We usually have to assume that this is the case. But there are situations in which we should not make this assumption. In time series data, for example, the daily amount of political news, we usually have trends, cyclic movements, or issues that affect the amount of news over a period of time. As a consequence, the amount and contents of political news on one day may depend on the amount and contents of political news on the preceding days.

Clustered data should also not be considered as independent observations. Think, for instance, of student evaluations of statistics tutorials. Students in the same tutorial group are likely to give similar evaluations because they had the same tutor and because of group processes: Both enthusiasm and dissatisfaction can be contagious.

8.1.5.2 Identically distributed observations

To check the assumption of identically distributed observations, we inspect the residuals. Remember, the residuals are represented by the error term (e) in the regression equation. They are the difference between the scores that we observe for our respondents and the scores that we predict for them with our regression model.

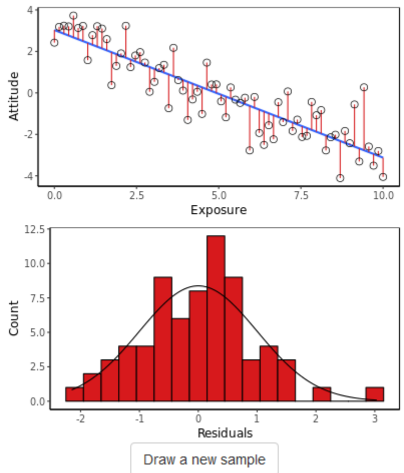


Figure 8.5: What are the residuals and how are they distributed?

1. What do the lines between dots and regression line represent in the scatter plot of Figure 8.5?

```
* A red line depicts the residual or prediction error: The difference between
the actual outcome score (attitude) for a respondent (dot) and the outcome
score predicted by the regression line (blue) for the predictor score
(exposure) of the respondent.
```

2. What is the relation between the scatter plot and the histogram? Can you point out the dot in the scatter plot that belongs to the leftmost bar in the histogram?

```
* The residuals represented by the red lines in the scatterplot are counted in
the histogram.
* Drag the mouse pointer around one or more dots in the scatterplot and you
will see where they are featured in the histogram (blue). The dot that is
furthest below the regression line is counted in the left-most bar of the
histogram.
```

3. Draw some new samples. Are the residuals always normally distributed?

```
* No, sometimes the histogram is clearly skewed (asymmetrical).
```

If we sample from a population where attitude towards smoking depends on exposure, smoking status, and contact with smokers, we will be able to predict attitude from the predictors in our sample. Our predictions will not be perfect, sometimes too high and sometimes to low. The differences between predicted and observed attitude scores are the residuals.

If our sample is truly a random sample with independent and identically distributed observations, our predictions should be equally bad or equally well for each value of the outcome variable, that is, attitude in our example. More specifically, the sizes of our errors (residuals) should be normally distributed for each attitude level (according to the central limit theorem).

So for all possible values of the outcome variable, we must collect the residuals for the observations that have this score on the outcome variable. For example, we should select all respondents who score 4.5 on the attitude towards smoking scale. Then, we select the residuals for these respondents and see whether they are approximately normally distributed.

Usually, we do not have more than one observation (if any) for each single outcome score, so we cannot practically apply this check. Instead, we use a simple and coarse approach: Are all residuals normally distributed?

A histogram with an added normal curve helps us to evaluate the distribution of the residuals. If the curve more or less follows the histogram, we conclude that the assumption of identically distributed observations is plausible. If not, we conclude that the assumption is not plausible and we warn the reader that the results can be biased.

8.1.5.3 Linearity and prediction errors

The regression models that we discuss assume that the association between the predictor and outcome variables is linear. Residuals of the regression model help us to see whether this is the case.

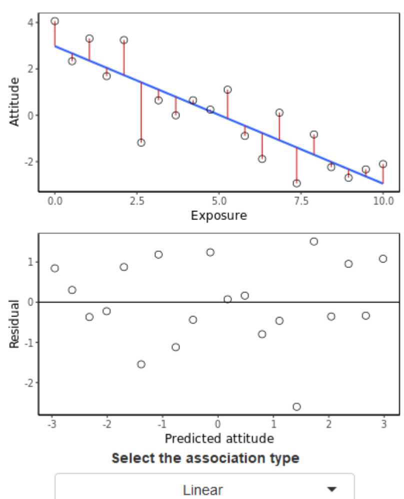


Figure 8.6: How do residuals tell us whether the relation is linear?

1. Which dot in the plot of residuals (Figure 8.6 bottom) corresponds with the left-most observation (dot) in the scatter plot of attitude by exposure (Figure 8.6 top)? Drag your mouse around the left-most dot while pressing the left mouse button to check your choice. Repeat for more dots until you understand the relation between the two plots.

```
* The blue regression line in the top graph represents the predicted values on
the attitude variable from exposure scores. The predicted attitude value for
the left-most observation is the value at which the blue line intersects with
the vertical red line dropping down from that observation.
* The left-most observation is the highest predicted value because the
regression line slopes down to the right. The highest predicted value is at
the far right of the residuals by predicted attitude graph (bottom graph)
because the predicted values are on the horizontal axis of this graph. The
left-most observation in the top graph, then, must correspond with the
right-most observation in the bottom graph.
* Note that this is not always true. If the regression slope is positive, that
is, the regression line goes up from left to right, the left-most observation
in the top graph has the lowest predicted value, so it corresponds to the
left-most observation in the residuals plot.
```

2. Select a U-shaped curve in Figure 8.6. Explain how the plot of residuals tells you that the association is not linear. Do the same for a curved association.

```
* If the association between two variables is U-shaped, the regression line
underestimates the attitude for observations with low exposure, overestimates
the attitude for medium values of exposure, and underestimates the attitude
for high exposure values. As a result, the residuals have a marked pattern
from left to right: a set of positive residuals, followed by a set of negative
residuals, followed by a set of positive residuals.
* The same phenomenon occurs for a curved association.
* In contrast, a linear association yields residuals without a clear pattern.
At all levels of exposure and, hence, at all predicted levels of attitude, we
may encounter both positive and negative residuals. In the residuals by
predicted values plot, we have positive and negative residuals everywhere. We
may underestimate as well as overestimate the outcome variable everywhere.
```

The other two assumptions that we use tell us about problems in our model rather than problems in our statistical inferences. Our linear regression model assumes a linear effect of the predictors on the outcome variable (*linearity*) and it assumes that we can predict the outcome equally well or equally badly for all levels of the outcome variable (*homoscedasticity*).

The regression models that we estimate assume a linear model. This means that an additional unit of the predictor always increases the predicted value by the same amount. If our regression coefficient for the effect of exposure on attitude is -0.25, an exposure score of one predicts a 0.25 more negative attitude towards smoking than zero exposure. Exposure score five predicts the same difference in attitude than score four, exposure score 10 predicts the same difference in comparison to exposure score nine, and so on. As a consequence of the linearity assumption, we can graph a regression model as a straight line.

The relation between a predictor and outcome variable, for example, exposure and attitude towards smoking, need not be linear. It can be curved or have some other fancy shape. Then, the linearity assumption is not met. A straight regression line does not nicely fit the data.

We can see this in a graph showing the (standardized) residuals (vertical axis) against the (standardized) predicted values of the outcome variable (on the horizontal axis). Note that the residuals represent prediction errors. If our regression predictions are systematically too low at some levels of the outcome variable and too high at other levels, the residuals are not nicely distributed around zero for all predicted levels of the outcome variable. This is what you see if the association is curved or U-shaped.

This indicates that our linear model does not fit the data. If it would fit, the average prediction error is zero for all predicted outcome levels. Graphically speaking, our linear model matches the data if positive prediction errors (residuals) are more or less balanced by negative prediction errors everywhere along the regression line.

8.1.5.4 Homoscedasticity and prediction errors

The plot of residuals by predicted values of the outcome variable tells us more than whether a linear model fits the data.

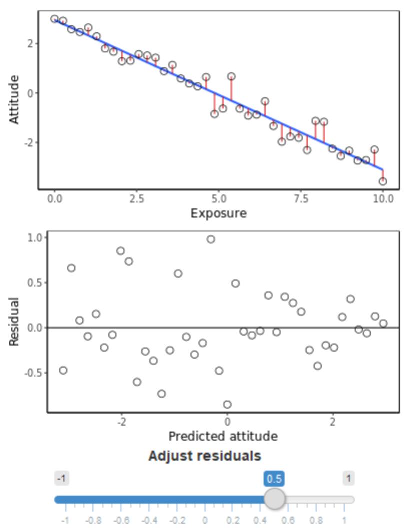


Figure 8.7: How do residuals tell us that we predict all values equally well?

1. What strikes you about the residuals in Figure 8.7?

```
* The residuals are higher for higher values of exposure.
* In the top graph, the residuals tend to become larger from left to right:
The residuals are larger for larger exposure scores (horizontal axis in the
top graph).
* In this particular example, larger exposure scores predict more negative
('lower') attitudes towards smoking (vertical axis in the top graph). We can
predict high attitude levels better (smaller residuals) than low attitude
levels (larger residuals).
```

```
* Attitudes are on the horizontal axis in the bottom graph, so the residuals
tend to become smaller if we go from left (low attitude scores) to right (high
attitude scores).
* The regression model seems to predict attitude better for participants with low
exposure scores than for participants with high exposure scores.
```

2. What happens if you move the slider to the far left?

```
* The pattern reverses. Now, residuals are larger for low values of exposure
or, equivalently in this model with a negative slope, for higher predicted
values of attitude.
* In this situation, we are better at predicting low attitude levels than high
attitude levels. Note that we prefer to predict all attitude levels equally
well.
```

3. At which slider position are all attitude levels predicted equally well or equally badly?

```
* If the slider is positioned at or around zero, the residuals are more or
less equal for low, medium, or high predicted values of attitude.
Here, we can predict all levels of attitude equally well or equally badly.
* This is best seen in the bottom graph: The vertical spread of observations
is more or less the same at the left, middle, and right of the graph. The
vertical diameter of the dot cloud is more or less equal from left to right.
* This is how we like the plot to look.
```

The other assumption states that we can predict the outcome variable equally well at all outcome variable levels. In other words, the prediction errors (residuals) are more or less the same at all levels of the outcome variable. If we have large prediction errors at some levels of the outcome variable, we should also have large prediction errors at other levels. As a result, the vertical width of the residuals by predictions scatter plot should be more or less the same from left to right.

If the prediction errors are not more or less equal for all levels of the predicted outcome, our model is better at predicting some values than other values. For example, low values can be predicted better than high values of the outcome variable. This may signal, among other things, that we need to include moderation in the model.

8.1.6 Visualizing predictions

A line in a scatterplot is a nice visualization of a regression model. The regression line represents the predicted values of the outcome variable for all values of the predictor.

In a simple regression model, we have just one predictor, so we have only one regression line. In a multiple regression model, we can draw a lot of regression lines even if we use the same predictor.

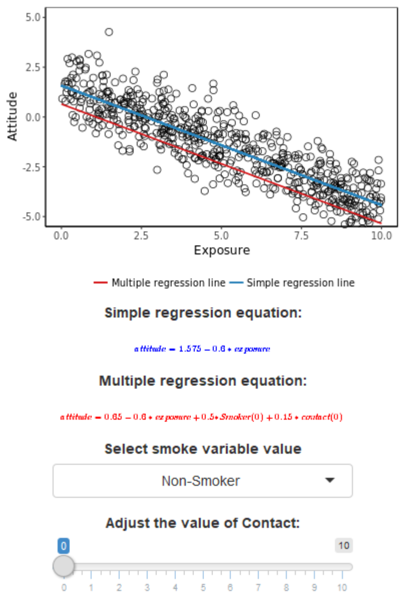


Figure 8.8: How do predictions based on exposure depend on values of smoking status and smoker contact?

1. What happens in Figure 8.8 if you change smoking status or smoking contact score? Can you explain the size of changes?

```
* The (red) multiple regression line moves up or down.
* The change from non-smoker to smoker adds 0.5 to the predicted values. The
red line moves up 0.5 points on the attitude scale.
* An additional unit of contact adds 0.15 to the predicted values. The red line
moves up 0.15 points on the attitude scale.
```

2. The line for a simple regression of attitude on exposure (with just one predictor) is displayed in blue in Figure 8.8. Can you make the multiple regression line equal to the simple regression line? If so, at what covariate values? If not, why not?

```
* It is possible to have the red regression line fully overlap the blue
regression line because they have the same slope, namely -0.6. The multiple and simple reg
the blue line.
* For non-smokers, we have to set contact near six for the red line to overlap
the blue one. The contact value for smokers is near 3.
* If you like, you can calculate the exact values for contact at which the
multiple and simple regression lines are equal. Link the two equations and plug
in the values that you know. For smokers:
* 1.575 - 0.6 * Exposure = 0.65 - 0.6 * Exposure + 0.5 * Smoker + 0.15 * Contact
* Exposure drops from the equation because it occurs both left and right. The
value for smokers is 1 on Smoker.
```

```
* 1.575 = 0.65  + 0.5 * 1 + 0.15 * Contact
* Rearrange:
* 1.575 - 0.65 - 0.5 = 0.15 * Contact
* Divide left and right by 0.15:
* (1.575 - 0.65 - 0.5) / 0.15 = Contact
* 0.425 / 0.15 = Contact
* Contact = 2.833.
```

The regression equation without the error term e predicts the outcome variable scores from the predictor scores. Plug in values for the predictor variables and you can calculate the predicted outcome score. Figure 8.8, for instance, shows a regression equation for the effects of exposure, smoking status, and contact with smokers on attitude towards smoking. If you plug in a score of 4 for exposure, 0 for smoking status (non-smoker), and 6 for contact with smokers, the predicted attitude is 0.65 + -0.6 * 4 + 0.5 * 0 + 0.15 * 6 = -0.85.

If we focus on the relation between one predictor and the outcome variable, the predicted values can be represented by a straight line in a scatter plot. By convention, we display the predictor on the horizontal axis and the outcome variable on the vertical axis of the scatter plot.

In a simple regression, we only have one predictor, so we can only draw one regression line. In multiple regression, however, we have several predictors. We can draw regression lines for each predictor and, importantly, we can draw different regression lines for the same predictor.

If we want to draw the regression line for the predictive effect of one predictor in a multiple regression model, we regard the other predictors as covariates. Let us define a *covariate* as a variable that may predict the outcome but is not our prime interest, so we mainly want to control for its effects. For example, if we focus on the predictive effect of exposure on attitude towards smoking, exposure is our predictor and smoking status and contact with smokers are covariates.

Note that the distinction between predictor and covariates is temporary. As soon as we focus on another variable, that variable becomes the predictor and the other predictors become covariates. The distinction between predictor and covariate is just terminology to show on which variable we focus.

To draw the regression line for the effect of exposure on attitude towards smoking, we must select a value for the covariates. If we would not do so, we have more than one variable that is allowed to vary but we can only display one predictor on the horizontal axis of our scatter plot.

$$
\begin{aligned}
attitude &= 0.65 + -0.6 * exposure + 0.5 * status + 0.15 * contact \\
attitude &= 0.65 + -0.6 * exposure + 0.5 * 0 + 0.15 * 3 \\
attitude &= 0.65 + -0.6 * exposure + 0 + 0.45 \\
attitude &= 1.10 + -0.6 * exposure
\end{aligned}
\tag{8.7}
$$

Let us select non-smokers (*status* equals 0) who score 3 on *contact* in Equation (8.7). If we plug in these values in the regression equation, we obtain a simple regression—just one predictor, namely *exposure*—with a higher constant: 1.10 instead of 0.65. The constant is the intercept of the regression line, that is, the value of the vertical axis where the regression line crosses it.

The slope of the regression line, however, does not change no matter which values we select for the covariates. The regression coefficient of *exposure* remains -0.6. The regression line only moves up or down if we choose different values for the covariates. To visualize the exposure effect, it does not matter which values we chose for the covariates. A popular choice is using their average scores. When we add a moderator, however, the slope also changes as we will see in Section 8.3.

8.2 Regression Analysis in SPSS

8.2.1 Instructions

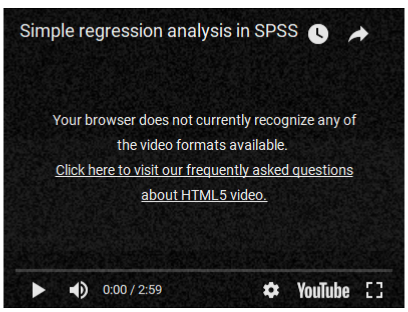


Figure 8.9: Executing and interpreting regression analysis in SPSS.

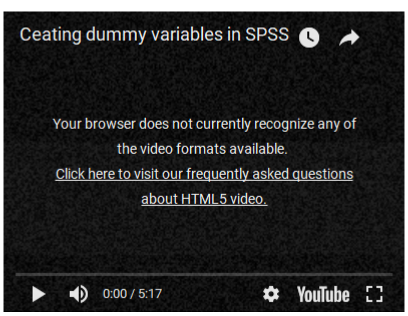


Figure 8.10: Creating dummy variables in SPSS.

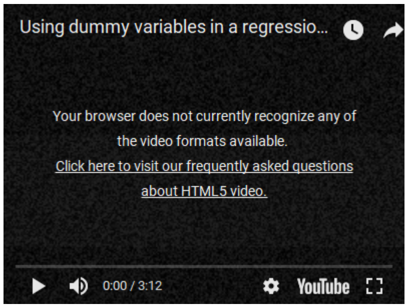


Figure 8.11: Using dummy variables in a regression model in SPSS.

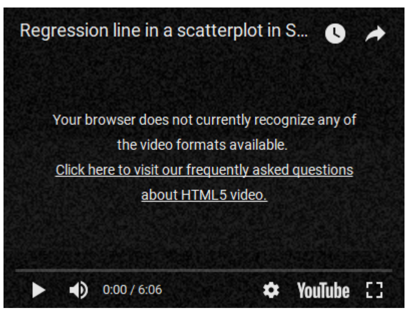


Figure 8.12: Adding a regression line to a scatterplot in SPSS.

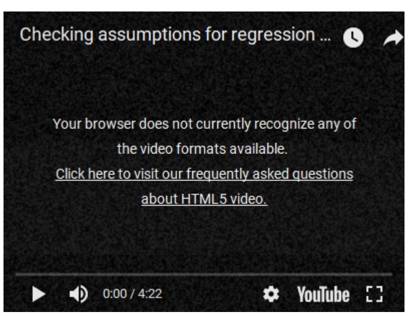


Figure 8.13: Checking assumptions for regression models in SPSS.

8.2.2 Exercises

1. Use the data in smokers.sav to predict the attitude towards smoking from exposure to an anti-smoking campaign. Check the assumptions and interpret the results.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=exposure attitude
  /ORDER=ANALYSIS.
* Simple regression analysis with assumption checks.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT attitude
  /METHOD=ENTER exposure
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID).

Check data:

There are no impossible values on the two variables.

Check assumptions:

* The distribution of the residuals is (left) skewed rather than normal. This
is not good.
* The residuals are nicely centered around zero at all predicted levels of the
outcome variable, so the association seems to be linear.
* The residuals are evenly spread around zero at all predicted levels.
Perhaps, the spread is slightly larger at low predicted values. In all,
however, prediction accuracy seems to be more or less the same at all levels
(homoscedasticity).

Interpret the results:

* Campaign exposure predicts attitude towards smoking among adults reasonably
well, R2 = .17, F (1, 83) = 16.73, p < .001.
* One additional unit of exposure decreases the predicted attitude by 0.12 to
0.34 points, t = -4.09, p < .001, 95%CI[-0.34; -0.12]. This is a moderate to
strong effect (b* = -.41).
```

2. Add smoking status (variable *status3*), and contact with smokers as predictors to the regression model of Exercise 1. Compare the effect of exposure between the two regression models. What is the difference and why is there a difference?

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=exposure status3 contact attitude
  /ORDER=ANALYSIS.

* Create dummy variables for status3.
* With Transform > Create Dummy Variables.
* ENSURE THAT MEASUREMENT LEVEL I SET TO ORDINAL.
* Define Variable Properties.
*status3.
VARIABLE LEVEL  status3(ORDINAL).
EXECUTE.
SPSSINC CREATE DUMMIES VARIABLE=status3
ROOTNAME1=status
/OPTIONS ORDER=A USEVALUELABELS=YES USEML=YES OMITFIRST=NO.

* If your SPSS version does not have this command, user Recode.
RECODE status3 (1=1) (ELSE=0) INTO status_2.
VARIABLE LABELS  status_2 'Former smoker'.
EXECUTE.
RECODE status3 (2=1) (ELSE=0) INTO status_3.
VARIABLE LABELS  status_3 'Smoker'.
EXECUTE.

* Multiple regression analysis with assumption checks.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT attitude
  /METHOD=ENTER exposure status_2 status_3 contact
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID).

Check data:

All values on the variables seem to be valid.

Check assumptions:

* The residuals seem to be a little skewed rather than symmetrical as in a
normal distribution.
* The residuals by predicted values plot is not as it should be. At low
predicted levels, the residuals are below zero, so the average residual is not
```

```
zero. The associations do not seem to be linear.
* In addition, the spread of the residuals is larger at higher predicted
levels (right) than low predicted levels (left). This pattern suggests that
the effects are moderated (see next section).

Interpret the results:

* Include a table with regression coefficients, so we need not report all t
test results in our interpretation.

* The regression model predicts about sixty per cent of the variation in
attitude towards smoking, which is very much for a social scientific model, R2
= .62, F (4, 80) = 32.02, p < .001.
* Exposure to the anti-smoking campaign predicts a more negative attitude
towards smoking, while contact with smokers is associated with a slightly more
positive attitude. Of the two, exposure (b* = -0.44) is a better predictor
than contact with smokers (b* = 0.20).
* Former smokers are on average much (2.85 points) more negative about smoking
than non-smokers and smokers; smokers are on average only 0.20 points below the
average attitude of non-smokers.
* The residuals suggest that the assumptions for using the theoretical
approximation of the sampling distributions may not have been met and/or that
the model should not be linear.
```

3. The data set children.sav contains information about media literacy of children and parental supervision of their media use. Are the two related? Check the assumptions and interpret the results.

```
* The analysis method you choose, depends on your substantive decision on the
direction of the association.
* If you assume no direction, a correlation coefficient is the most
appropriate choice. If you think one variable may depend on another, a
(simple) regression model is the best choice.

SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=medliter supervision
  /ORDER=ANALYSIS.
* Set supervision 25 to missing.
* Define Variable Properties.
*supervision.
MISSING VALUES supervision(25.00).
EXECUTE.
* Undirected: correlation (linear?).
* Check scatterplot.
GRAPH
```

```
  /SCATTERPLOT(BIVAR)=supervision WITH medliter
  /MISSING=LISTWISE.
* Correlations.
CORRELATIONS
  /VARIABLES=medliter supervision
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
* Simple regression: media literacy dependent.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT medliter
  /METHOD=ENTER supervision
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID).
* Simple regression: parental supervision dependent.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT supervision
  /METHOD=ENTER medliter
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID).

Check data:

Score '25' for parental supervision cannot be right because the scale runs to
10. Define this score as a missing value.

Check assumptions:

* In the regression models, the residuals are quite normally distributed,
nicely grouped around zero at all levels of the predicted outcome.
* The variation of residuals is more or less the same at different levels of
the predicted outcome but there are perhaps too few observations for low and
high predicted levels to decide.

Interpret the results:

Parental supervision and child media literacy are weakly correlated (r = .36,
p = .001). More supervision predicts more media literacy, t = 3.54,
```

```
p = .001, 95%CI[0.13; 0.48]. One additional unit of parental supervision
predicts 0.31 additional units of media literacy. One additional unit of media
literacy predicts 0.42 additional units of parental supervision.

Note that the t test on the regression coefficient is the same in a simple
regression model if you use supervision or media literacy as outcome variable.
```

8.3 Different Lines for Different Groups

If we have a categorical independent variable, for instance, people living among smokers versus people living among non-smokers, and we want to determine its effect on a numerical variable, for example, attitude towards smoking, we compare group means. The difference between group means is the main effect of the categorical variable. For example, the average attitude towards smoking is 0.5 points more positive for people living among smokers than for people who do not live among smokers.

In analysis of variance (Chapter 7), the main effect of living among smokers is the average effect for all people regardless of their other characteristics or the contexts that they are in. In other words, a main effect is the overall difference in attitude between people living among smokers and those who are not living among smokers.

What if the effect of living among smokers on attitude may be different in different contexts, e.g., for people who smoke themselves, used to smoke, or never smoked (*smoking status*)? To model this, we added an interaction effect to the main effects in Chapter 7.

The interaction effect tells us whether the attitude difference between, on the one hand, people living among smokers and, on the other hand, people living among non-smokers is different for smokers, former smokers, and non-smokers. In a conceptual diagram, the interaction effect is represented by an arrow pointing at another arrow. The moderator (smoking status) changes the relation between the predictor (living among smokers) and the outcome (attitude towards smoking).

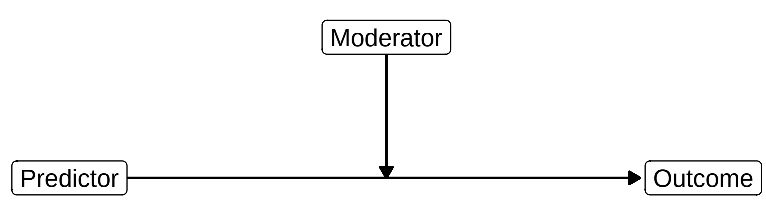


Figure 8.14: Conceptual diagram of moderation.

What if living among smokers is a numerical variable, for example, counting the number of times the respondent is in the company of a smoker? How can we investigate if the effect of this numerical predictor on attitude is moderated by the respondent's smoking status?

Analysis of variance (ANOVA), as discussed in Chapter 7, investigates the effects of categorical variables on a numeric outcome variable. It cannot handle numeric predictors or numeric covariates. Although there are ways to include numeric covariates in analysis of variance, for instance, in the analysis of covariance (ANCOVA), we use regression analysis if we have at least one numerical predictor or covariate and a numerical outcome.

In the current section, we discuss regression models with a numerical predictor and a categorical moderator. A later section (Section 8.5), presents regression models in which both the predictor and moderator are numeric.

8.3.1 A dichotomous moderator and continuous predictor

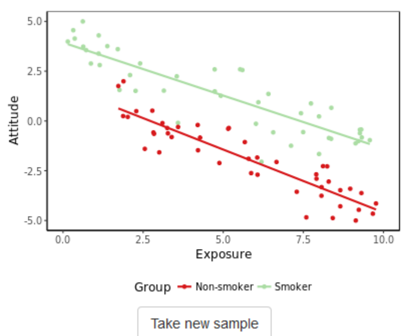


Figure 8.15: Is the effect of exposure on attitude moderated by smoking status?

1. Is the effect of exposure on attitude moderated by the smoking status of respondents (smokers versus non-smokers) in Figure 8.15? Motivate your answer. Press the *Take new sample* button to practice some more with recognizing moderation.

```
* Remember: The slope of the regression line represents the effect (the
regression coefficient) of the predictor. Parallel lines mean equal slopes.
* The answer depends on the sample drawn. If the regression lines for the two
smoking status groups are more or less parallel, the (predictive) effect of
exposure on attitude is more or less the same for both groups.
* In this case, there is no moderation.
* In all other situations, there is moderation.
```

In Section 8.1, we have analyzed the predictive effects of exposure to an anti-smoking campaign and smoking status on a person's attitude towards smoking. We have found a negative effect for exposure and a positive effect for smoking. More exposure predicts a more negative attitude whereas smokers have a more positive attitude towards smoking than non-smokers.

Our current question is: Does exposure to the campaign have the same effect for smokers and non-smokers? We want to compare an effect (exposure on attitude) for different contexts (smokers versus non-smokers), so our current question involves moderation. Is the effect of exposure on attitude moderated by smoking status?

Our moderator (smoker vs. non-smoker) is a dichotomous variable but our predictor (exposure) is numeric, so we cannot use analysis of variance. Instead, we use regression analysis, which allows numeric predictors.

In the context of a regression model, moderation means **different slopes for different groups**. The slope of the regression line is the regression coefficient, which expresses the effect of the predictor on the outcome variable. If we have different effects in different contexts (moderation), we must have different regression coefficients for different groups.

8.3.2 Interaction variable

How do we obtain different regression coefficients and lines for smokers and non-smokers? The statistical trick is quite easy: Include a new predictor in the model that is the product of the predictor (exposure) and the moderator (smoking status). This new predictor is the *interaction variable*. It must be included together with the original predictor and moderator variables, see Equation (8.8). This is also visible in the statistical diagram (Figure 8.16) for moderation in a regression model.

$$\begin{aligned} attitude = {} & constant + b_1 * exposure + b_2 * smoker + b_3 * contact \\ & + b_4 * exposure * smoker + e \end{aligned} \tag{8.8}$$

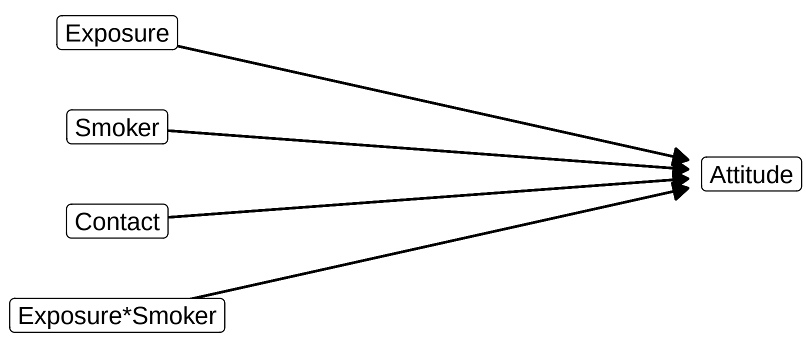


Figure 8.16: Statistical diagram of moderation.

The smoking status variable is coded 1 for smokers and 0 for non-smokers. For clarity, we name this variable smoker with score 1 for Yes and score 0 for No. Remember (Section 8.1.3) that we have two different regression equations, one for each group on the dichotomous predictor *status*. Just plug in the two possible values (1 and 0) for this variable. For non-smokers, the interaction variable drops from the model because multiplying with zero yields zero. For non-smokers, our reference group, b_1 represents the effect of exposure on attitude. It is called the *simple slope* of exposure for non-smokers.

$$
\begin{aligned}
attitude &= constant + b_1 * exposure + b_2 * smoker + b_3 * contact \\
&\quad + b_4 * exposure * smoker + e \\
attitude &= constant + b_1 * exposure + b_2 * 0 + b_3 * contact \\
&\quad + b_4 * exposure * 0 + e \\
attitude &= constant + b_1 * exposure + b_3 * contact + e
\end{aligned} \tag{8.9}
$$

In contrast, the interaction variable remains in the model for smokers, who score 1 on smoking status. Note what happens with the coefficient of the exposure effect if we rearrange the terms a little: The exposure effect equals the effect for the reference group of non-smokers (b_1) plus the effect of the interaction variable (b_4). The simple slope for smokers, then, is $b_1 + b_4$.

$$
\begin{aligned}
attitude &= constant + b_1 * exposure + b_2 * smoker + b_3 * contact \\
&\quad + b_4 * exposure * smoker + e \\
attitude &= constant + b_1 * exposure + b_2 * 1 + b_3 * contact \\
&\quad + b_4 * exposure * 1 + e \\
attitude &= constant + b_1 * exposure + b_4 * exposure + b_2 + b_3 * contact + e \\
attitude &= constant + (b_1 + b_4) * exposure + b_2 + b_3 * contact + e
\end{aligned} \tag{8.10}
$$

The interaction effect (b_4) shows the difference between the simple slope of the exposure effect for smokers ($b_1 + b_4$) and the simple slope for non-smokers (b_1). Let us assume that the unstandardized regression coefficient of the interaction effect is -0.3 (see Section 8.3.4). This means that the effect of exposure on attitude is more strongly negative (or less positive) for smokers than for non-smokers. One additional unit of exposure decreases the predicted attitude for smokers by 0.3 more than for non-smokers.

8.3.3 Conditional effects, not main effects

It is very important to note that the effects of exposure and smoking status in a model with exposure by smoking status interaction are **not** main effects as in analysis of variance. As we have seen in the preceding section (Equation (8.9)), the regression coefficient b_1 for exposure expresses the effect of exposure for the reference group of non-smokers. It is a *conditional effect*, namely the effect for non-smokers only. This is quite different from a main effect, which is an average effect over all groups.

In a similar way, the regression coefficient b_2 for smoking status expresses the effect for persons who score zero on the exposure predictor. Simply plug in the value 0 for exposure in the regression equation (Equation (8.11)).

Table 8.2: Predicting attitude towards smoking: regression analysis results.

	B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
(Constant)	0.900	0.357		2.521	0.014	0.190	1.610
Exposure	-0.162	0.061	-0.162	-2.651	0.010	-0.284	-0.040
Status (smoker)	1.980	0.738	1.980	2.683	0.009	0.512	3.448
Exposure*Status (smoker)	-0.327	0.142	-0.327	-2.311	0.023	-0.609	-0.045

$$
\begin{aligned}
attitude &= constant + b_1 * exposure + b_2 * smoker + b_3 * contact \\
&\quad + b_4 * exposure * smoker + e \\
attitude &= constant + b_1 * 0 + b_2 * smoker + b_3 * contact \\
&\quad + b_4 * 0 * smoker + e \\
attitude &= constant + b_2 * smoker + b_3 * contact + e
\end{aligned}
\tag{8.11}
$$

Smoking status is a dichotomy, so its regression coefficient (b_2) tells us the average difference in attitude between smokers and non-smokers. Due to the inclusion of the interaction variable, it now tells us the difference in average attitude between smokers and non-smokers who have zero exposure to the anti-smoking campaign. Note again that this is a conditional effect, not a main effect.

8.3.4 Interpretation and statistical inference

In Table 8.2, non-smokers are the reference group because they are coded 0 on the *Status* variable. As a consequence, the regression coefficient for exposure gives us the effect of exposure on smoking attitude for non-smokers. It's value is -0.16, so an additional unit of exposure predicts a smoking attitude among non-smokers that is 0.16 points more negative. More exposure to the campaign goes together with a more negative attitude towards smoking for non-smokers. The p value for this effect tests the null hypothesis that the effect is zero in the population. If we reject this null hypothesis, the exposure effect is statistically significant for non-smokers.

The effect of (smoking) status on attitude is conditional on exposure. The regression coefficient for status tells us the difference between smokers and non-smokers who have 0 exposure. So, without exposure to the campaign, smokers are on average 1.98 more positive towards smoking than non-smokers. The p value tests the null hypothesis that the difference is zero for people without exposure to the anti-smoking campaign.

Smokers are coded 1 on the (smoking) status variable, so the regression coefficient for the interaction tells us that the slope of the exposure effect is 0.33 lower for smokers than for non-smokers. In a preceding paragraph, we have seen that the estimated slope of the exposure effect is -0.16 for non-smokers. We can add the regression coefficient of the interaction variable to obtain the estimated slope for smokers, which is -0.49. Now we can compare the two regression lines for the two groups, which gives good insight in the nature of moderation in this example. A graph of the two regression lines is probably the best way to communicate your results.

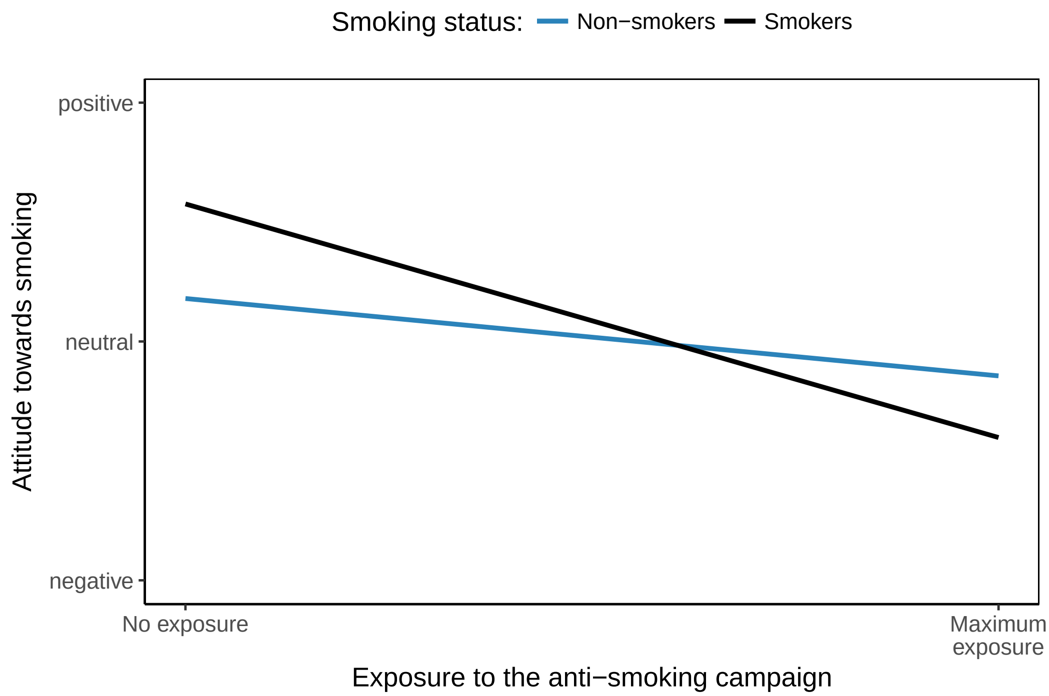


Figure 8.17: The effects of exposure to the anti-smoking campaign on attitude towards smoking among smokers and non-smokers.

The interaction variable is treated as an ordinary predictor in the estimation process, so it receives a confidence interval and a p value. The null hypothesis is that the interaction effect is zero in the population.

Remember that the regression coefficient for the interaction variable expresses the difference between the slope for the indicated group, e.g., smokers, and the slope for the reference group, e.g., non-smokers. If this difference is zero, as stated by the null hypothesis, the two groups have the same slope, so the effect is not moderated by the group variable.

So we know the confidence intervals and p values of the exposure effect for non-smokers (the regression coefficient for exposure) and for the difference between their exposure effect and the exposure effect for smokers (the regression coefficient for the interaction effect). We do not know, however, the confidence interval and statistical significance of the exposure effect for smokers. We cannot add confidence intervals or p values.

If you want to know the confidence interval or p value of the exposure effect for smokers, you have to rerun the regression analysis using a different dummy variable for the moderator. You should create a dichotomous variable that assigns the 1 score to non-smokers and an interaction variable created with this dichotomy. The regression coefficient of the exposure effect now expresses the effect for smokers because smokers are the reference group on the new dummy variable. The associated p value and confidence interval apply to the exposure effect for smokers.

Interaction variables are used just like ordinary predictors, so the general assumptions of regression analysis apply. See Section 8.1.5 for a description of the assumptions and checks.

Let us conclude the interpretation with a warning. The standardized regression coefficients that SPSS reports for interaction effects or effects of predictors that are involved in interaction effects **must not be used**. They are calculated in the wrong way if the regression model includes an interaction variable. As a result, they are meaningless.

8.3.5 A categorical moderator

What if we have three or more groups in our moderator? For example, smoking status measured with three categories: (1) never smoked, (2) have smoked, (3) currently smoking? Does the effect of exposure on attitude vary between people who never smoked, stopped smoking, and are still smoking?

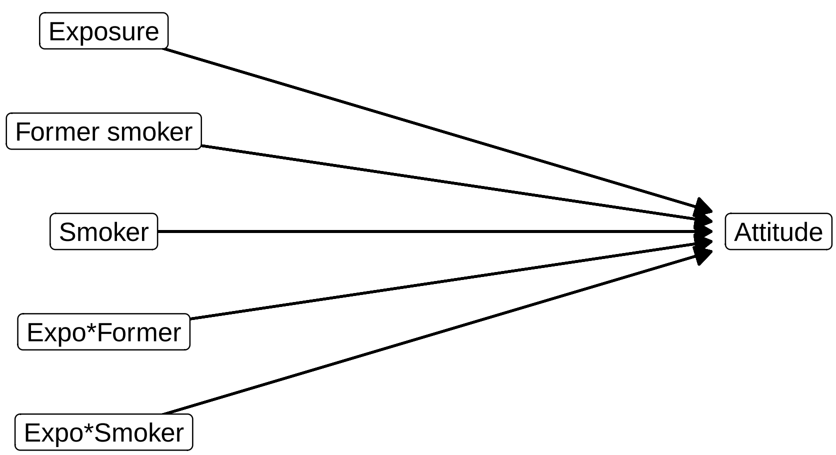


Figure 8.18: When do we have moderation with a categorical moderator?

1. If we have exposure effects within three groups, as in Figure 8.18, when do we have an interaction effect (moderation) and when do we not have an interaction effect? Motivate your answer. Press the *Take new sample* button to practice recognizing moderation.

```
* If the regression lines for all three smoking status groups are more or less
parallel, the (predictive) effect of exposure on attitude is more or less the
same for the three groups. In this case, there is no moderation.
* In all other situations, there is moderation. There is an interaction
```

Table 8.3: Predicting attitude towards smoking for three smoking status groups: regression analysis results.

	B	Std. Error	t	Sig.	Lower Bound	Upper Bound
(Constant)	1.644	0.288	5.717	0.000	1.072	2.216
Exposure	-0.185	0.046	-3.987	0.000	-0.277	-0.093
Former smoker	-1.095	0.544	-2.013	0.048	-2.178	-0.012
Smoker	1.235	0.521	2.372	0.020	0.199	2.272
Exposure*Former smoker	-0.405	0.112	-3.604	0.001	-0.629	-0.181
Exposure*Smoker	-0.304	0.098	-3.116	0.003	-0.498	-0.110

```
effect, for example, if two lines are parallel but the third line is not
parallel to the other two lines.
```

In Section 8.1.4, we learned that we must create dummy variables for all but one groups of a categorical predictor in a regression model. This is what we have to do also for a categorical moderator because we must include a moderator as a predictor in the regression model. If the effect of another predictor, such as exposure, is moderated by a categorical variable, we have to create an interaction variable for each dummy variable in the equation. To create the interaction variables, we multiply the predictor with each of the dummy variable as we have done before.

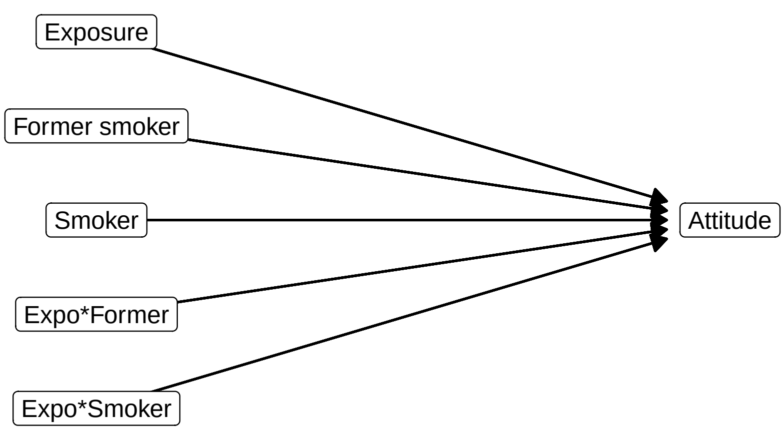


Figure 8.19: Statistical diagram with a moderator consisting of three groups. Non-smokers are the reference group

In the end, we have an interaction variable for all groups but one on the categorical moderator. Figure 8.19 shows the statistical diagram. Estimation of the model yields point estimates (regression coefficients), confidence intervals, and p values for all variables (Table 8.3).

Remember that the effects of predictors that are included in interactions are conditional effects: effects for the reference group or reference value on the other variable involved in the interaction. The p value for Exposure tests the hypothesis that the exposure effect for people who never smoked is zero in the population. For the two dummy variables Former smoker and Smoker, the null hypothesis is tested that they have the same average attitude in the population as the non-smokers (reference group) if they are not exposed to the anti-smoking campaign (zero exposure).

Interaction predictors show effect differences. In Table 8.3, the interaction predictors test the null hypotheses that the effect of exposure is equal for former smokers and non-smokers (Exposure*Former smoker) or for smokers and non-smokers (Exposure*Smoker) in the population.

If we would like to know whether the exposure effect for former smokers is significantly different from zero, we have to rerun the regression model using the people who stopped smoking as reference group. This new model would also tell us whether the exposure effect for people who stopped smoking is significantly different from the exposure effect for people who are still smoking.

8.3.6 Common support

In a regression model with moderation, we have to interpret the effect of a predictor involved in the interaction at a particular value of the moderator (Section 8.3.3). The estimated effect at a particular value of the moderator can only be trusted if there are quite some observations at or near this value of the moderator. In addition, these observations should cover the full range of values on the predictor. After all, the effect that we estimate must tell us whether higher values on the predictor go together with higher (or lower) values on the outcome.

For example, we need quite some observations for smokers to estimate the conditional effect of exposure on attitude for smokers. If there are hardly any smokers in our sample, we cannot estimate the effect of exposure on attitude for them in a reliable way. Even if we have quite some observations for smokers but all smokers have low exposure, we cannot say much about the effect of exposure on attitude for them. If we cannot say much about the effect within this group, we cannot say much about the difference between this effect and effects for other groups. In short, the moderation model is problematic here.

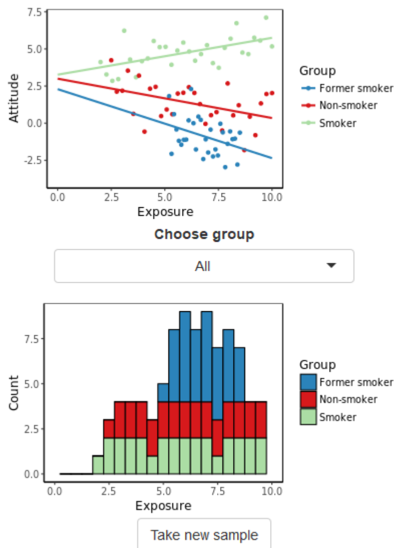


Figure 8.20: How well do the observations cover the predictor within each category of smoking status?

1. What does the histogram represent in Figure 8.20?

```
* The histogram shows the counts of cases with particular exposure values.
* The counts are subdivided (coloured) by their score on the moderator variable
(smoking status).
```

2. Are the exposure values nicely spread for each smoke status group? Inspect each smoke status group separately with the "Show groups" option.

```
* The answer depends on the sample that was drawn.
* A moderator (smoking status) group may have scores for the entire range of
the predictor exposure, that is, from 0 to 10. But scores may also be available
only for part of this range, for example, from 1 to 6 or from 4 to 8.
```

3. What is the problem if exposure scores are not nicely spread over the same range for all smoke status groups?

```
* We may erroneously believe that the moderation effect applies to the entire
range of scores. For example, we may conclude that there is a negative effect
for smokers versus a positive effect for non-smokers. However, the effect for
smokers may be based on a range of exposure values that is different and
perhaps hardly overlapping with the range of exposure values for non-smokers.
In this case, the effect difference may be due to the level of predictor scores
instead of to the moderator.
* In extreme cases, we may have only very few observations for a moderator
group, for example, only a handful of smokers. In this situation, the
regression effect for this group is not to be trusted.
```

The variation of predictor scores for a particular value of the moderator is called *common support* (Hainmueller, Mummolo, & Xu, 2016). If common support for predictors involved in moderation is bad, we should hesitate to draw conclusions from the estimated effects. Guidelines for good common support are hard to give. Common support is usually acceptable if there are observations over the entire range of the predictor.

It is recommended to check the number of observations per value of the moderator. For a categorical moderator, such as smoking status, a scatter plot of outcome (vertical axis) by predictor (horizontal axis) with dots coloured according to the moderator category may do the job. Check that there are observations for more or less all values of the predictor. In Figure 8.21, observations in each moderator category (dot colour) range from low to high predictor values.

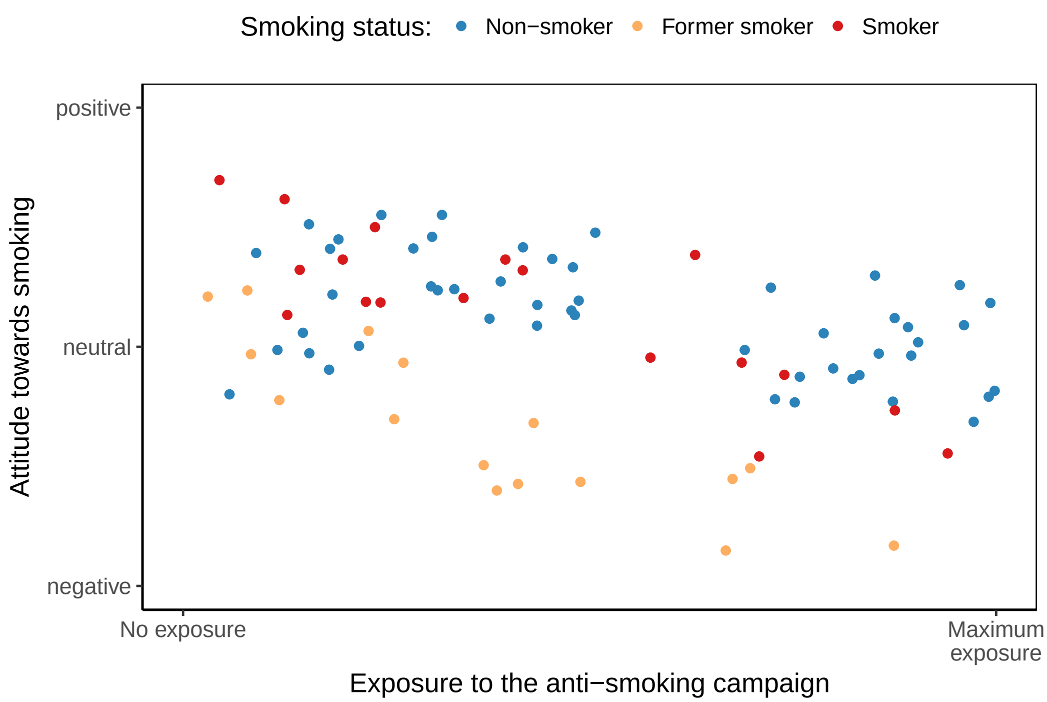


Figure 8.21: The effects of exposure to the anti-smoking campaign on attitude towards smoking among smokers and non-smokers.

8.4 A Dichotomous or Categorical Moderator in SPSS

8.4.1 Instructions

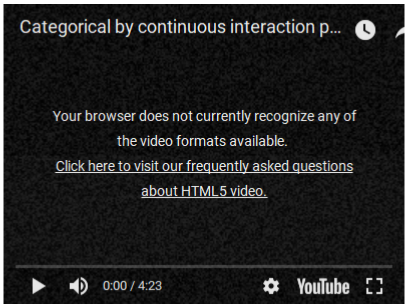


Figure 8.22: Creating categorical by continuous interaction predictors for regression in SPSS.

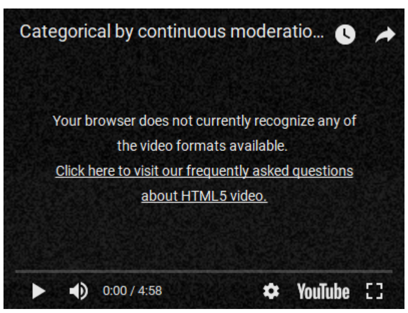


Figure 8.23: Estimating categorical by continuous moderation with regression in SPSS.

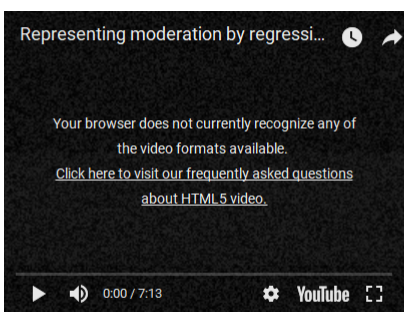


Figure 8.24: Representing moderation by regression lines in a scatterplot in SPSS.

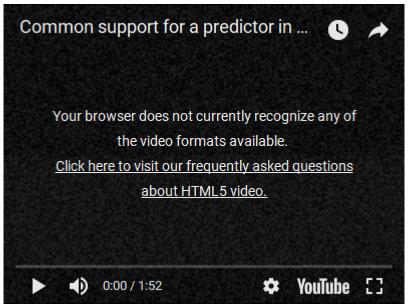


Figure 8.25: Checking common support for a predictor at different moderator values in SPSS.

8.4.2 Exercises

1. Use the data in smokers.sav to predict the attitude towards smoking from exposure moderated by smoking status (variable *status2*). Use contact with smokers as a covariate. Check the assumptions for regression analysis and interpret the results.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=exposure status2 contact attitude
  /ORDER=ANALYSIS.
* Compute interaction variable.
COMPUTE expo_status=exposure * status2.
VARIABLE LABELS  expo_status 'Interaction exposure * smoker'.
EXECUTE.
* Multiple regression.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT attitude
  /METHOD=ENTER exposure status2 expo_status contact
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID).

Check data:

All values seem to be valid.

Check assumptions:

* The residuals are skewed (long tail to the left).
```

```
* The residuals seem to average to zero on most levels of the predicted
outcome, so a linear model seems to fit.
* However, the lower attitude values are predicted worse (more variation) than
the higher levels. The assumptions do not seem to be strongly violated but our
model may not be specified all well.

Interpret the results:

Add table with regression coefficients.

* We can predict smoking attitude for about 26 per cent with the regression
model, R2 = .26, F (4, 80) = 7.10, p < .001.
* The predictive effect of exposure to the anti-smoking campaign on smoking
attitude for non-smokers is more probably negative than positive but it is not
significantly different from zero, b = -0.12, t = -1.79, p = .078,
95%CI[-0.25, 0,01]). More exposure tends to yield a more negative attitude.
* For smokers, the predictive effect of campaign exposure on smoking attitude
is more strongly negative. The moderation of the exposure effect by smoking
status is negative and statistically significant, b = -0.33, t = -2.35, p =
.021, 95%CI[-0.61, -0.05].
* More contact with smokers is associated with a more positive attitude
towards smoking rather than a negative attitude. The predictive effect is weak
(b* = 0.18) but not significantly different from zero, b = 0.15, t = 1.68, p =
.096, 95%CI[-0.03, 0.33].
* Smokers have a more positive attitude than non-smokers if they are not
exposed to the campaign, on average circa 2 (0.5 to 3.4) points more positive
the attitude, and this difference is statistically significant, b = 1.98, t =
2.72, p = .008, 95%CI[0.53, 3.43].

Remember:
* In the presence of an interaction effect, the partial effect of a predictor
is the effect for the reference group or value. It is NOT an overall or
average effect as in analysis of variance.
* Standardized regression coefficients reported by SPSS are not correct for
interaction effects or effects of predictors that are involved in interaction
effects. They can only be used for predictors that are not involved in
interaction effects.
* Contact with smokers is not involved in an interaction in this model, so we
can interpret the standardized regression coefficient for this effect.
```

2. Visualize the moderated effects of exposure on attitude (Exercise 1). Create a scatter plot with two regression lines. Colour the regression lines and the dots (respondents) according to their smoking status category. Interpret how smoking status moderates the effect of exposure on attitude.

```
* First of all, you must write the regression equations for different values
of the moderator. Plug in the estimated values of the regression coefficients
and the means of covariates (here: average contact with smokers).

* SPSS syntax to get the average value of contact with smokers:
FREQUENCIES VARIABLES=contact
 /FORMAT=NOTABLE
 /STATISTICS=MEAN
 /ORDER=ANALYSIS.

attitude = constant + -.118*exposure + 1.982*status + -.329*exposure*status + .152*contact

attitude = -.087 + -.118*exposure + 1.982*status + -.329*exposure*status + .152*5.091

attitude = -.087 + -.118*exposure + 1.982*status + -.329*exposure*status +.774

attitude = .687 + -.118*exposure + 1.982*status + -.329*exposure*status

* Replace status by 0 for non-smokers.

Non-smokers (0):

attitude = .687 + -.118*exposure + 1.982*0 + -.329*exposure*0

attitude = .687 + -.118*exposure

* Replace status by 1 for smokers.

Smokers (1):

attitude = .687 + -.118*exposure + 1.982*1 + -.329*exposure*1

attitude = .687 + 1.982 + -.118*exposure + -.329*exposure

attitude = 2.669 + (-.118 + -.329)*exposure

attitude = 2.669 + -.447*exposure

* Next, create a scatterplot of attitude by exposure, colouring the dots by
smoking status. Use the calculated two equations in the SPSS Chart Editor to
create two lines. Use the icon "Add a reference line from Equation" for each
line. Enter the equation using x instead of exposure as the predictor.
Colour the lines with the colours of the dots in the scatterplot.

SPSS syntax:
```

```
* Scatterplot with dots coloured by smoking status.
GRAPH
  /SCATTERPLOT(BIVAR)=exposure WITH attitude BY status2
  /MISSING=LISTWISE.

Check data:

See Exercise 1.

Check assumptions:

See Exercise 1.

Interpret the results:

* For smokers, the predictive effect of campaign exposure on smoking attitude
is more strongly negative (b = -0.45) than for non-smokers (b = -0.12.
```

3. Check common support of the predictor (exposure) in all groups of the moderator (smoking status). Could you also check common support with the scatter plot you made for Exercise 2?

```
SPSS syntax:

* Histogram of predictor (exposure) for each smoking status.
GRAPH
  /HISTOGRAM=exposure
  /PANEL ROWVAR=status2 ROWOP=CROSS.

Interpret the results:

* Even for smokers, the much smaller group, we have exposure scores over
(almost) the entire range. We have good coverage both for smokers and
non-smokers. Do not mind the gap in scores around 6: we have plenty of
observations around 5 and 7.

* We could have seen this result in the scatterplot because we had green and
blue dots across the entire width of the plot.
```

4. Repeat the analyses of Exercises 1 through 3 but use smoking status with three categories (*status3*).

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=exposure status3 contact attitude
  /ORDER=ANALYSIS.
```

```
* Create dummies and interaction variables.

* With Create Dummy Variables.
* ENSURE THAT MEASUREMENT LEVEL IS SET TO ORDINAL.
* Define Variable Properties.
*status3.
VARIABLE LEVEL  status3(ORDINAL).
EXECUTE.
SPSSINC CREATE DUMMIES VARIABLE=exposure status3
ROOTNAME1=exposure, status ROOTNAME2=expo_status
/OPTIONS ORDER=A USEVALUELABELS=YES USEML=YES OMITFIRST=NO.

* With Recode.
RECODE status3 (1=1) (ELSE=0) INTO status_3.
VARIABLE LABELS  status_3 'Former smoker'.
EXECUTE.
RECODE status3 (2=1) (ELSE=0) INTO status_4.
VARIABLE LABELS  status_4 'Smoker'.
EXECUTE.
* Interaction variables (same name as those given by Create Dummy Variables).
COMPUTE expo_status_2_2=exposure * status_3.
VARIABLE LABELS  expo_status_2_2 'expo * formersmoker'.
EXECUTE.
COMPUTE expo_status_2_3=exposure * status_4.
VARIABLE LABELS  expo_status_2_3 'expo * smoker'.
EXECUTE.

* Multiple regression.
* Statistic Descriptives is added to get the means that we need
* to plug into the regression equation in the moderation plot.
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT attitude
  /METHOD=ENTER exposure status_3 status_4 expo_status_2_2 expo_status_2_3 contact
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID).
* Scatterplot with dots coloured by smoking status.
GRAPH
  /SCATTERPLOT(BIVAR)=exposure WITH attitude BY status3
  /MISSING=LISTWISE.
* Histogram of predictor (exposure) for each smoking status.
GRAPH
```

```
  /HISTOGRAM=exposure
  /PANEL ROWVAR=status3 ROWOP=CROSS.

* Write out the regression equations for all three groups of the moderator.
Plug in the estimated values of the constant and the regression coefficients.

attitude = .550 + -.137*exposure + -1.139*former + 1.223*smoker +
  -.402*exposure*former + -.304*exposure*smoker + .171*contact

Plug in the means of covariates and add to constant:

attitude = .550 + -.137*exposure + -1.139*former + 1.223*smoker +
  -.402*exposure*former + -.304*exposure*smoker + .171*5.091

attitude = 1.421 + -.137*exposure + -1.139*former + 1.223*smoker +
  -.402*exposure*former + -.304*exposure*smoker

Non-smokers (former = 0, smoker = 0):

attitude = 1.421 + -.137*exposure + -1.139*0 + 1.223*0 + -.402*exposure*0 +
  -.304*exposure*0

attitude = 1.421 + -.137*exposure

Former smokers (former = 1, smoker = 0):

attitude = 1.421 + -.137*exposure + -1.139*1 + 1.223*0 + -.402*exposure*1 +
  -.304*exposure*0

attitude = 1.421 + -1.139 + -.137*exposure + -.402*exposure

attitude = 0.282 + (-.137 + -.402)*exposure

attitude = 0.282 + -.539*exposure

Smokers (former = 0, smoker = 1):

attitude = 1.421 + -.137*exposure + -1.139*0 + 1.223*1 + -.402*exposure*0 +
  -.304*exposure*1

attitude = 1.421 + 1.223 + (-.137 + -.304)*exposure

attitude = 2.644 + -.441*exposure

* Create a scatterplot of attitude by exposure, colouring the dots by smoking
```

```
status.
* Use the calculated two equations in the SPSS Chart Editor to create two
lines. Use the icon "Add a reference line from Equation" for each line. Enter
the the equation using x instead of exposure as the predictor. Colour the
lines with the colours of the dots in the scatterplot.

Check data:

All values seem to be valid.

Check assumptions:

* The residuals seem to be skewed a little bit.
* The residuals by predicted values plot gives no reason to doubt the
linearity of the model but the problem of predicting higher values less
accurately than lower values seems to be worse than in Exercise 1. We should
warn the reader that the assumptions seem to be violated.

Interpret the results:

* Again, summarize the tests on the regression coefficients in a table.

* With three smoking status groups, we can predict attitude towards smoking
much better (R2 = 0.70) than with the two groups in Exercise 1 (R2 = .26).

* There is an important difference between non-smokers and former smokers
(they were lumped together in the preceding exercises). On average, former
smokers have a more negative attitude than non-smokers, which may range from a
tiny difference (-0.1) to a large difference (-2 points on the scale from -5 to
+5) if they are not exposed to the campaign.

* In addition, exposure has a stronger negative predictive effect on smoking
attitude among former smokers than among non-smokers. Exposure also has a
stronger negative effect on attitude among smokers than among non-smokers. For
short, exposure to the campaign has less impact on attitude towards smoking
for non-smokers than for former smokers or smokers.

* The coverage of exposure is good for non-smokers and smokers but former
smokers with high exposure are rare.
```

8.5 A Continuous Moderator

With a categorical moderator, it is quite obvious for which values of the moderator we are going to calculate and depict the effect of the predictor on the outcome. If smoking status moderates the effect of exposure on attitude towards smoking, we will inspect a regression line for each smoking status category: smokers, former smokers, and non-smokers. But what if the moderator is a continuous, numeric variable, for example, the intensity of contact with smokers?

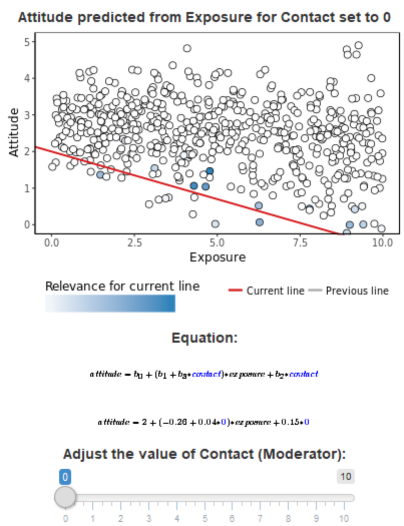


Figure 8.26: How do contact values affect the conditional effect of exposure on attitude?

1. The regression line depicted in Figure 8.26 represents the conditional effect of exposure on attitude for the value of contact with smokers selected with the slider. How many different conditional effects are there?

```
* In principle, there is an unlimited number of conditional effects if the
moderator is a continuous variable.
* In the app, however, the slider allows you to increase or decrease contact
score by 0.5. In the app, there are effectively 21 moderator values that you
can select, so there are 21 different regression lines that can be depicted.
```

2. Is the effect of campaign exposure on attitude towards smoking always negative? Or does more exposure lead to a more positive attitude (higher score) in some cases? If so, in which cases?

```
* Move the slider from left to right to find the moderator value at which the
regression line is horizontal. For higher moderator values, the slope of the
regression line is positive. Here, more exposure leads to a more positive
attitude.
* Or have a close look at the equations. The regression coefficient of the
(simple) slope of the exposure effect is zero if the part between brackets (b1
+ b3*contact) is zero. We know b1 and b3, so we must solve the equation:

> -0.26 + 0.04 * contact = 0
```

```
Your high-school algebra may help you:

> 0.04 * contact = 0.26

> contact = 0.26 / 0.04

> contact = 0.26 / 0.04 = 6.5
```

3. How much does the slope increase if the moderator value is changed from 0 to 1? And how much if it changes from 6 to 7?

```
* Each increment of 1 unit of contact increases the (simple) slope of the
exposure effect on attitude by 0.04, that is, by the value of the interaction
effect.
* This is easy to see in the equation for the effect of exposure:

> (-0.26 + 0.04 * contact) * exposure

* Plug in 0 for contact: The simple slope is -0.26.
* Plug in 1 for contact: The simple slope is -0.26 + 0.04 * 1.
* The difference is 0.04. This difference is the same for every increase of
one unit in contact, so it is also the difference between the slopes at
contact levels six and seven.
```

People hanging around a lot with smokers are likely to have a more positive attitude towards smokers than people who have little contact with smokers. After all, people who really hate smoking will avoid meeting smokers. This is a main effect of contact with smokers on attitude towards smoking.

In addition, the anti-smoking campaign may be less effective for people who spend a lot of time with smokers. Negative perceptions of smoking instilled by the campaign can be compensated by positive experiences of seeing people enjoy smoking. Contact with smokers would decrease the effect of campaign exposure on attitude. The effect of exposure is moderated by contact with smokers.

Our moderator, contact with smokers, is continuous. As a consequence, we can have an endless number of contact levels as groups for which the slope may change. This is the only difference with a categorical moderator. Other than that, we will analyze a continuous moderator in the same way as we analyzed a categorical moderator.

8.5.1 Interaction variable

We need one interaction variable to include a continuous moderator in a regression model. As before, the interaction variable is the product of the predictor and the moderator. Multiply the predictor with the moderator to obtain the interaction variable.

Although we have an endless number of different moderator values or groups, we only need one interaction variable. It represents the gradual (linear) change of the effect of the predictor for higher values of the moderator.

$$\begin{aligned} attitude &= constant + b_1 * exposure + b_2 * contact + b_3 * exposure * contact + e \\ attitude &= constant + (b_1 + b_3 * contact) * exposure + b_2 * contact + e \end{aligned} \tag{8.12}$$

To see this, it is helpful to inspect the regression equation with rearranged terms (Equation (8.12)). Every extra contact with smokers adds b_3 to the slope ($b_1 + b_3 * contact$) of the exposure effect. The addition is gradual—a little bit of additional contact with smokers changes the exposure effect a little bit—and it is linear: A unit increase in contact adds the same amount to the effect whether the effect is at a low or a high level.

We can interpret the regression coefficient of the interaction effect (b_3) here as the predicted change in the exposure effect (slope) for a one unit difference in contact (the moderator). A positive coefficient indicates that the exposure effect is more positive for higher levels of contact with smokers. A negative coefficient indicates that the effect is more negative for people with more contacts with smokers.

Note that positive and negative are used here in there mathematical meaning, not in an appreciative way. A positive effect of exposure implies a more positive attitude towards smoking. Anti-smoking campaigners probably evaluate this as a negative result.

8.5.2 Conditional effect

The regression coefficients for exposure and contact represent conditional effects (see Section 8.3.3), namely, the effects for cases that score zero on the other variable. Plug in zero for the moderator and you will see that all terms with a moderator drop from the equation and only b_1 is left as the effect of exposure.

$$\begin{aligned} attitude &= constant + (b_1 + b_3 * contact) * exposure + b_2 * contact + e \\ attitude &= constant + (b_1 + b_3 * 0) * exposure + b_2 * 0 + e \\ attitude &= constant + b_1 * exposure + e \end{aligned} \tag{8.13}$$

The zero score on the moderator is the *reference value* for the conditional effect of the predictor. Cases that score zero on the moderator are the *reference group* just like cases scoring zero on the dummy variables are the reference group in a model with a categorical moderator (Section 8.1.3).

8.5.3 Mean-centering

Because the effects of predictors involved in an interaction are conditional effects, a zero score on these variables has a special role. It is the reference value for the effect of the other predictor. For example, the effect of exposure on attitude applies to respondents with zero contacts with smokers if the regression model includes an exposure by contact interaction. If zero is so important, we may want to manipulate this value.

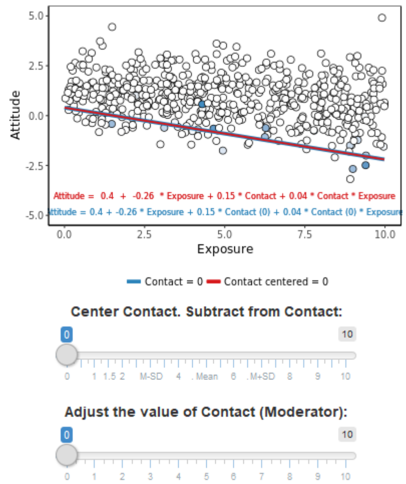


Figure 8.27: What happens if you mean-center the moderator variable?

1. What is the correct interpretation of the estimated regression coefficient of exposure in Figure 8.27?

```
* In the initial situation, the regression line in this figure represents the
predictive effect of exposure on attitude for respondents who score zero on the
moderator (Contact with smokers).
* In this regression model, the regression coefficient of exposure is -0.26,
so an additional unit of exposure decreases the predicted attitude by 0.26 for
respondents who have no (zero) contacts with smokers.
```

2. What happens to the red regression line and the regression equation if you subtract the mean (M) from the respondents' contact with smokers scores? Thus, you mean-center the moderator Contact. Use the slider **Center Contact. Subtract from Contract:** in Figure 8.27 to check your answer.

```
* Mean Contact score is 5. The regression coefficient of exposure changes to
-0.06, namely -0.26 + 0.04 * 5. Accordingly, the red regression line becomes
almost horizontal.
```

3. If Contact is mean-centered, the regression coefficient of exposure represents the effect of exposure on attitude for respondents with a particular score on the original Contact variable. What is this original Contact score? Use the slider **Adjust the value of Contact (Moderator):** to check your answer.

```
* The regression line of the mean-centered Contact moderator coincides with
the blue regression line if we select 5 as value of the original Contact
variable.
```

```
* Apparently, the reference group for the regression coefficient of exposure in
case of a mean-centered moderator, consists of respondents who score 5 on the
original Contact with smokers moderator.
* By mean centering, we assign the score zero to respondents who used to have
a mean score on the moderator Contact with smokers. Now that they score zero
on the new mean-centered variable, they have become the reference group.
```

What if there are no people with zero contact? Then, the interpretation of the regression coefficient b_1 for exposure does not make sense. In this situation, it is better to mean-center the moderator (contact) before you add it to the regression equation and before you calculate the interaction variable.

To *mean-center* a variable, you subtract the variable's mean from all scores on the variable. As a result, a mean score on the original variable becomes a zero score on the mean-centered variable.

$$contactcentered = contact - mean(contact)$$

With mean-centered numerical moderators, a conditional effect in the presence of interaction always makes sense. It is the effect of the predictor for average score on the moderator. An average score always falls within the range of scores that actually occur. If we mean-center the contact with smokers moderator, the regression coefficient b_1 for exposure expresses the effect of exposure on attitude for people with average contacts with smokers. This makes sense.

Remember that the interaction variable is the product of the predictor and moderator (Section 8.3.2). If any or both of these are mean-centered, you should multiply the mean-centered variable(s) to create the interaction variable, see Sections 8.3.3 and 8.5.2.

8.5.4 Symmetry of predictor and moderator

If we want to interpret the conditional effect of contact on attitude (b_2), we must realize that this is the effect for people who score zero on the exposure variable if the exposure by contact interaction is included in the regression model. This is clear if we rearrange the regression equation as in Equation (8.14).

$$\begin{aligned}
attitude &= constant + b_1 * exposure + b_2 * contact + b_3 * exposure * contact + e \\
attitude &= constant + b_1 * exposure + (b_2 + b_3 * exposure) * contact + e \\
attitude &= constant + b_1 * 0 + (b_2 + b_3 * 0) * contact + e \\
attitude &= constant + b_2 * contact + e
\end{aligned} \tag{8.14}$$

But wait a minute, this is what we would do if contact was the predictor and exposure the moderator. That is a completely different situation, is it not? No, technically it does not make a difference which variable is the predictor and which is the moderator. The predictor and moderator are symmetric. The difference is only in our theoretical expectations and in our interpretation.

The conditional effect of the moderator, as stated above, is the effect of the moderator if the predictor is zero. This interpretation makes sense only if there are cases with zero scores on the predictor. In the current example, the scores on exposure range from 0 to 10, so zero exposure is meaningful. But it represents a borderline score with perhaps a very atypical effect of contact on attitude or few observations. For these reasons, it is recommended to *mean-center both the predictor and moderator if they are numeric.*

8.5.5 Visualization of the interaction effect

It can be quite tricky to interpret regression coefficients in a regression model that contains interaction effects. The safest strategy is to draw regression lines for different values of the moderator. But what are interesting values if the moderator is numerical?

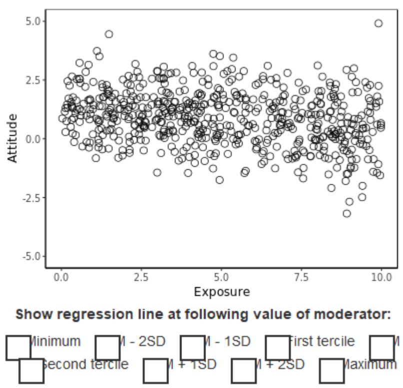


Figure 8.28: Which moderator values are helpful for visualizing moderation?

1. Select one or more options in Figure 8.28 to represent regression lines predicting attitude from exposure at different values of the moderator (contact with smokers). Respondents with moderator values close to the selected value are coloured. Which moderator values would you pick to communicate the results of moderation? Motivate your answer.

```
* The important thing is to understand that a regression line that depends on
few observations is not very trustworthy. The line usually does not represent
the observations well. In addition, these few observations may have different
predictor and outcome values in a new sample, so the regression line may be
quite different at this moderator value in a new sample.
* The minimum and maximum value observed in the current sample represent
regression lines that are quite outside the dot cloud. They are based on very
few observations, so they are not very trustworthy.
* Observations with moderator values at two standard deviations away from the
```

```
mean (options: M - 2SD and M + 2SD), are still quite rare.
* Observations with moderator values at one standard deviations away from the
mean (options: M - 1SD and M + 1SD), are quite common. These regression
lines are nicely embedded in the dot cloud. They are good candidates for
interpreting moderation.
* Of course, moderator values even closer to the mean, such as moderator values
below or above which we find one third of all scores (the first and third
terciles) are also well supported by observations, as is the moderator mean
itself.
```

As we have seen in Section 8.5.1, the regression coefficient of an interaction effect with a continuous moderator can be directly interpreted. It represents the predicted difference in the unstandardized effect size for a one unit increase in the moderator. For example, one more contact with a smoker increases the exposure effect by 0.04.

The size of the interaction effect tells us the moderation trend, for instance, people who are more around smokers tend to be less opposed to smoking if they are exposed to the anti-smoking campaign. But we do not know how much an anti-smoking attitude is fostered by exposure to a campaign and whether exposure to the campaign increases anti-smoking attitude for everyone. Perhaps, people hanging out with smokers a lot may even get a more positive attitude towards smoking from campaign exposure.

We can be more specific about exposure effects at different levels of contact with smokers if we pick some interesting values of the moderator and calculate the conditional effects at these levels.

The minimum or maximum values of the moderator are usually not very interesting. We tend to have few observations for these values, so our confidence in the estimated effect at that level is low. Instead, the values one standard deviation below and above the mean of the moderator are popular values to be picked. One standard deviation below the mean (M - SD) indicates a low value, the mean (M) indicates a central value, and one standard deviation above the mean (M + SD) indicates a high value.

Having picked these values, we can visualize moderation as different regression lines in a plot. We use exactly the same approach as in visualizing moderation by a categorical variable. As a first step, we construct equations for conditional effects of the predictor at different levels of the moderator. Plug the selected value of the moderator into the regression equation. If there are covariates, also plug in a meaningful value for the covariates, usually the average for numeric covariates and zero or one for dichotomous covariates. As a second step, we use the equations to add regression lines to a scatter plot.

If contact (the moderator) is mean-centered, as in the current example, we simply plug in zero for the moderator to obtain the equation for the regression line at the mean of the moderator (contact with smokers). We plug in the value of the standard deviation of contact with smokers to get the regression equation for people who scored one standard deviation above the mean on the moderator. The standard deviation of contact is 2.0 in this example, so Equation (8.15) replaces contact by 2.0 everywhere. Finally, we plug in minus the value of one standard deviation if we want the regression line for the moderator at the mean minus

Table 8.4: Predicting attitude towards smoking: regression analysis results with exposure and contact mean-centered.

	B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Boun
(Constant)	0.169	0.204		0.825	0.412	-0.238	0.57
Exposure (mean-centered)	-0.174	0.063	-0.174	-2.740	0.008	-0.300	-0.04
Contact (mean-centered)	0.159	0.094	0.159	1.685	0.096	-0.029	0.34
Status (smoker)	0.533	0.405	0.533	1.318	0.191	-0.272	1.33
Exposure*Contact (mean-centered)	0.018	0.034	0.018	0.533	0.595	-0.049	0.08

one standard deviation.

$$
\begin{aligned}
attitude &= 3.6 + -0.1 * exposure + 0.1 * status + 0.1 * contact \\
&\quad + 0.03 * contact * exposure \\
attitude &= 3.6 + -0.1 * exposure + 0.1 * (0) + 0.1 * (2.0) + 0.03 * (2.0) * exposure \\
attitude &= 3.4 + -0.04 * exposure
\end{aligned} \tag{8.15}
$$

We also have to plug in a value for each covariate. This example contains one covariate, namely (smoking) status. We plug in the score for non-smokers (0). In the end, our predictor (exposure) should be the only variable in the right hand side of the regression equation (the last line in Equation (8.15)). Note that we do not include the error term (e) in the equation if we predict values; the error term captures prediction errors.

If the moderator is not mean-centered, we have to plug in the value of the mean of the moderator and the value of the mean plus or minus the standard deviation of the moderator. In this example, the mean score of contact with smokers is 5.1, so the moderator mean minus one standard deviation (2.0) equals 3.1 and the mean plus one standard deviation is 7.1.

8.5.6 Statistical inference on conditional effects

The regression model yields a p value and confidence interval for the predictor at the reference value of the moderator. In the model estimated in Table 8.4, for instance, we obtain a p value of 0.008 and a 95% confidence interval of [-0.30, -0.05] for the effect of exposure on attitude. This is the conditional effect of exposure on attitude for cases that score zero on the moderator variable (contact with smokers) as we can verify in Equation (8.16).

$$
\begin{aligned}
attitude &= constant + b_1 * exposure + b_2 * contact + b_3 * status + \\
&\quad + b_4 * exposure * contact + e \\
attitude &= constant + (b_1 + b_3 * contact) * exposure + b_2 * contact + e \\
attitude &= constant + (b_1 + b_4 * 0) * exposure + b_2 * 0 + b_3 * status + e \\
attitude &= constant + b_1 * exposure + b_3 * status + e
\end{aligned} \tag{8.16}
$$

If the variable contact is mean-centered, the p value tests the null hypothesis that the effect of exposure is zero for people who have average contact with smokers. The confidence interval tells us that the effect of exposure on attitude for people with average contacts with smokers ranges between -0.30 and -0.05 with 95% confidence. If the moderator is not mean-centered, the results apply to people who have no contact with smokers.

Note that mean-centering of the moderator changes, so to speak, the regression line that we test from the effect of exposure for people with no smoker contact to the effect for people with average contact with smokers. If we would like to get the p value or confidence interval for the regression line at one standard deviation above or below the mean, we have to center the moderator at those values before we estimate the regression model.

8.5.7 Common support

In Section 8.3.6, we checked the support of the predictor in the data for different groups of the moderator. The basic idea is that we can only sensibly estimate and interpret a conditional effect at a moderator level if we have observations over the entire range of the predictor. For each moderator group, we checked the distribution of the predictor.

With a continuous moderator we can also do this if we group moderator scores. Hainmueller et al. (2016) recommend creating three groups, each containing one third of all observations. These low, medium, and high groups correspond more or less with the minus one standard deviation/mean/plus one standard deviation values that we used for visualizing and testing conditional effects. Create a histogram for the predictor in each of these groups to check common support of moderation in the data.

8.5.8 Assumptions

The general assumptions for regression analysis (Section 8.1.5) also apply to the interaction effect with a continuous moderator. The checks are the same: See if the residuals are more or less normally distributed and check the residuals by predicted values plot.

Note that the linearity assumption also applies to the interaction effect. If the interaction effect is positive, the exposure (predictor) effect must be higher for higher values of contact with smokers (moderator). More precisely, a unit difference on the moderator should result in a fixed increase (or decrease) of the effect of the predictor. You may have noticed this linear change in the effect size in Figure 8.26 at the beginning of this section on continuous moderators.

It is difficult to check this assumption, so let us not pursue this here. Just remember that the interaction effect is assumed to be linear: an ever increasing or decreasing effect of the predictor at higher moderator values.

Table 8.5: Predicting attitude towards smoking. Results in APA6 style. Exposure and contact are mean-centered.

	B	95% CI
Constant	-1.08***	[-1.32, -0.85]
Exposure	-0.18***	[-0.26, -0.10]
Contact	0.21***	[0.12, 0.30]
Former smoker	-1.38***	[-1.74, -1.01]
Smoker	0.10	[-0.41, 0.60]
Exposure * Contact	0.06***	[0.02, 0.09]
Exposure * Former smoker	-0.20**	[-0.33, -0.07]
Exposure * Smoker	-0.02	[-0.17, 0.13]
R^2^	0.59	
F	28.64***	

8.5.9 Higher-order interaction effects

An interaction effect with one moderator, whether continuous or categorical, is called *first-order interaction*. It is possible to have a moderated effect that is moderated itself by a second moderator. For example, the change in the exposure effect due to a person's contact with smokers may be different for smokers than for non-smokers. This is called a *second-order interaction* or *higher-order interaction*. We can include more moderators, yielding even higher higher-order interactions, such as three or four moderators.

An interaction variable that is the product of the predictor and two moderators can be used to include a second-order interaction in a regression model. If you include a second-order interaction, you must also include the effects of the variables involved in the interaction as well as all first-order interactions among these variables in the regression model. All in all, these models become very complicated to interpret, so we do not pay attention to them.

8.6 Reporting Regression Results with Moderation

Note. $N = 150$. CI = confidence interval.

* $p < .05$. ** $p < .01$. *** $p < .001$.

If we report a regression model, we first present the significance test and predictive power of the entire regression model. We may report that the regression model is statistically significant, F (7, 142) = 28.64, p < 0.001, so the regression model very likely helps to predict attitude towards smoking in the population. Retrieve the test information from SPSS; the APA6-style table (Table 8.5) only reports the F value and its significance level.

How well does the regression model predict attitude towards smoking? The effect size of a regression model or its predictive power is summarized by R^2 (*R Square*), which is the proportion of the variation in the outcome variable scores (attitude towards smoking) that can be predicted with the regression model. In this example, R^2 is 0.59, so the regression model predicts 59% of the variance in attitude towards smoking among the respondents. In communication research, R^2 is usually smaller.

R^2 tells us how well the regression model predicts the outcome variable in the sample. Every predictor that we add to the regression model helps to predict results in the sample even if the predictor does not help to predict the outcome in the population. For a better idea of the predictive power of the regression model in the population, we may use *Adjusted R Square*. Adjusted R Square is usually slightly lower than R Square. In the example, Adjusted R Square is 0.56 (not reported in Table 8.5)

As a next step, we discuss the size, statistical significance, and confidence intervals of the regression coefficients. If a predictor is involved in one or more interaction effects, we must be very clear about the reference value and reference group to which the effect applies.

Exposure, in our example, has a negative predictive effect on attitude towards smoking for non-smokers with average contacts with smokers, t = -4.37, p < .001, 95%CI[-0.26, -0.10]. Note that SPSS does not report the degrees of freedom for the t test on regression coefficient, so we cannot report them.

Instead of presenting the numerical results in the text, we may summarize them in an APA6 style table, such as Table 8.5. Note that t and p values are not reported in this table, the focus is on the confidence intervals. The significance level is indicated by stars.

A sizable and statistically significant interaction effect signals that an effect is moderated. In the example reported in Table 8.5, the effect of exposure on attitude seems to be moderated by contact with smokers (b = 0.06, p < .001) and by smoking status (b = -0.20, p = 0.003).

The regression coefficients for interaction effects must be interpreted as effect differences. For a categorical moderator, the coefficient describes the effect size difference between the category represented by the dummy variable and the reference group. The negative effect of exposure is stronger for former smokers than for the reference group non-smokers. The average difference is -0.20.

For a continuous moderator, we can interpret the general pattern reflected by the interaction effect. A positive interaction effect, such as 0.06 for the interaction between exposure and smoker contact, signals that the effect of exposure is more strongly positive or less negative at higher levels of contact with smokers.

This interpretation in terms of effect differences remains difficult to understand. It is recommended to select some interesting values for the moderator and report the size of the effect for each value. For a categorical moderator, each category is of interest. For a continuous moderator, the mean and one standard deviation below and above the mean are usually interesting values. The regression coefficients show whether the effect is positive, negative, or nearly zero at different values of the moderator.

Visualize the regression lines for different values of the moderator rather than presenting the numerical results. If the regression model contains covariates, mention the values that you have used for the covariates. Select one of the categories for a categorical covariate. For numeric covariates, the mean is a good choice. If you are working with mean-centered predictors, be sure to use the mean-centered predictor for the horizontal axis (as in Figure 8.29), not the original predictor.

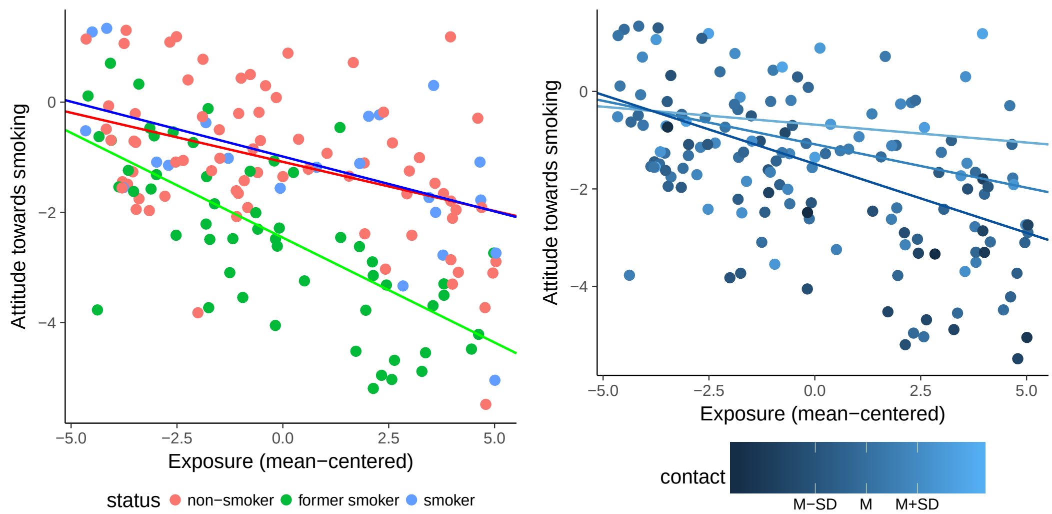


Figure 8.29: The effect of exposure on attitude towards smoking. Left: Effects for groups with different smoking status (at average contact with smokers). Right: Effects at different levels of contact with smokers (effects for non-smokers).

The left panel in Figure 8.29 clearly shows that the effect of exposure on attitude is more or less the same for non-smokers and smokers. The effect is different for former smokers, for whom the exposure effect is more strongly negative. It is difficult to draw this conclusion from the table with regression coefficients.

Check that the predictor has good support at the selected values of the moderator. In the left-hand plot of Figure 8.29, the groups (colours) vary nicely over the entire range of the predictor *exposure*, so that is okay. It is more difficult to see good variation in the right-hand plot.

Do not report that common support is good. If it is bad, warn the reader that we cannot fully trust the estimated moderation because we do not have a nice range of predictor values within each level of the moderator.

Finally, inspect the residual plots but do not include them in the report. Warn the reader if the assumptions of the linear regression model are not met. Do not mention the assumptions if they are met.

8.7 A Continuous Moderator in SPSS

8.7.1 Instructions

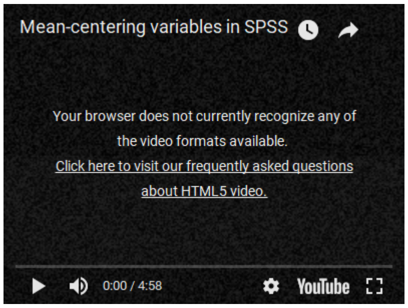


Figure 8.30: Mean-centering variables for regression analysis in SPSS.

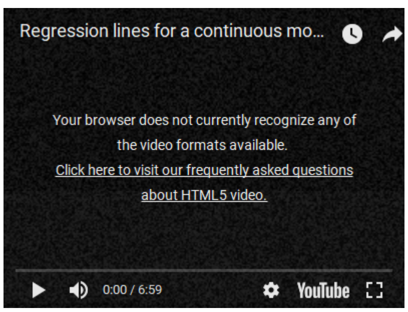


Figure 8.31: Regression lines for a continuous moderator in a scatterplot in SPSS.

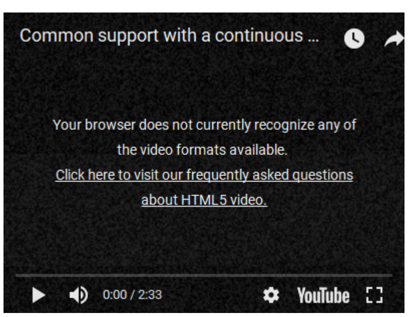


Figure 8.32: Checking common support with a continuous moderator in SPSS.

8.7.2 Exercises

1. With the data in smokers.sav, check if the effect of campaign exposure on attitude towards smoking depends on the contacts that people have with smokers. For now, do not mean-center the variables. Control for the respondent's smoking status (*status2*). Interpret the regression coefficients and check the assumptions of the regression model.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=exposure status2 contact attitude
  /ORDER=ANALYSIS.
* Compute interaction variable.
COMPUTE expo_contact=exposure * contact.
VARIABLE LABELS  expo_contact 'Interaction exposure * contact'.
EXECUTE.
* Multiple regression.
* Statistic Descriptives is added to get the means that we need
* to plug into the regression equation in the moderation plot.
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT attitude
  /METHOD=ENTER exposure contact expo_contact status2
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID).

Check data:

There are no impossible values on the variables.

Check assumptions:

* The residuals are skewed, so the assumption of a normal distribution can be
violated.
* The residuals seem to average to zero at all levels of the predicted
outcome. This supports a linear model. Note that it is not a problem that the
residuals tend to be further from zero if they are below zero.
* Prediction errors seem to be more or less of equal size at different levels
of the outcome variable, so the assumpion of homoscedasticity seems to be met.

Interpret the results:

* The regression model predicts 21 per cent of the variation in the outcome
```

```
variable, F (4, 80) = 5.44, p = .001.
* None of the regression coefficients, however, is statistically significant.
We are not confident that the directions of the estimated effects are the true
directions; they may have the opposite direction in the population.
* But we should realize that the significance tests of the moderated partial
effects apply to the effect at one particular level of the moderator. At this
level of the moderator, the estimated coefficient is not sufficiently
different from zero. At other levels of the moderator, however, the
coefficient can be sufficiently different from zero for the test to be
statistically significance.

* The effects of exposure and contact with smokers must be interpreted with
care due to their interaction effect.
* The estimated effect of exposure applies to adults who score zero on the
variable contact with smokers. In this context, exposure makes the predicted
attitude towards smoking more negative, b = -0.27, t = -1.54, p = .128.
* Contact with smokers makes the attitude more positive for adults who have no
exposure to the anti-smoking campaign, b = 0.07, t = 0.41, p = .682.
* The interaction effect is positive, b = .02, t = 0.53, p = .595, so exposure
seems to be less effective in making the attitude towards smoking more
negative if the adult has more contacts with smokers.
```

2. Visualize the moderating effect of contact with smokers on the exposure effect in a scatter plot with three regression lines. Explain the information conveyed by the plot to your reader

```
SPSS syntax:

* Create scatterplot.
GRAPH
  /SCATTERPLOT(BIVAR)=exposure WITH attitude
  /MISSING=LISTWISE.

Manually add three regression lines:

* Write out the regression equations for different values of the moderator.
* Plug in the estimated values of the regression coefficients, the means of
covariates, and three values for the moderator using its M and SD.

The initial equation for non-smokers (status = 0):

attitude = .648*constant + -.265*exposure + .072*contact +
  .018*exposure*contact + 0.533*status

attitude = .648*1 + -.265*exposure + .072*contact +
  .018*exposure*contact + 0.533*0
```

```
attitude = .648 + (-.265 + .018*contact)*exposure + .072*contact

The equation with Contact = M - SD

attitude = .648 + (-.265 + .018*(5.091 - 1.974))*exposure + .072*(5.091 - 1.974)

attitude = .648 + (-.265 + .018*3.117)*exposure + .072*3.117

attitude = .648 + (-.265 + .056)*exposure + .224

attitude = .872 + -.209*exposure

The equation with Contact = M

attitude = .648 + (-.265 + .018*5.091)*exposure + .072*5.091

attitude = .648 + (-.265 + .092)*exposure + .367

attitude = 1.015 + -.173*exposure

The equation with Contact = M + SD

attitude = .648 + (-.265 + .018*(5.091 + 1.974))*exposure + .072*(5.091 + 1.974)

attitude = .648 + (-.265 + .018*7.065)*exposure + .072*7.065

attitude = .648 + (-.265 + .127)*exposure + .509

attitude = 1.157 + -.138*exposure

Interpret the results:

* The negative predictive effect of exposure on attitude towards smoking is
slightly stronger (more negative) for adults with fewer contacts with smokers.
```

3. Mean-center the predictor and moderator and repeat the regression analysis of Exercise 1. Explain the differences in the results.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=exposure status2 contact attitude
  /ORDER=ANALYSIS.
* Mean-center predictor and moderator.
* Ask for means of predictor and exposure.
FREQUENCIES VARIABLES=exposure contact
```

```
  /FORMAT=NOTABLE
  /STATISTICS=MEAN
  /ORDER=ANALYSIS.
* Subtract mean from variable.
COMPUTE exposure_c=exposure - 4.866.
VARIABLE LABELS  exposure_c 'Exposure (mean-centered)'.
COMPUTE contact_c=contact - 5.091.
VARIABLE LABELS  contact_c 'Contact (mean-centered)'.
EXECUTE.
* Compute new interaction variable.
COMPUTE expo_contact_c=exposure_c * contact_c.
VARIABLE LABELS  expo_contact_c 'Interaction exposure * contact  (mean-centered)'.
EXECUTE.
* Multiple regression.
* Statistic Descriptives is added to get the means that we need
* to plug into the regression equation in the moderation plot.
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT attitude
  /METHOD=ENTER exposure_c contact_c expo_contact_c status2
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID).

Check data: See Exercise 1.

Check assumptions: See Exercise 1.

Interpret the results:

* The size and significance of the interaction effect have not changed at all.
Mean-centering only changes the reference values for the effects of the
predictor and the moderator.
* The coefficient for exposure now expresses the predictive effect of exposure
for adults with average contact with smokers. They have more contact with
smokers than the reference group in Exercise 1, who had no contact with
smokers. The interaction effect tells us that the effect of exposure becomes
less negative at higher levels of contact. This explains that we have a lower
value for the exposure coefficient now. It still is negative, so more exposure
to the anti-smoking campaign predicts a more negative attitude towards
smoking for adults with average contact with smokers.
* In contrast, the positive effect of contact with smokers on attitude is
```

```
stronger now (b = 0.16) than in Exercise 1 (b = 0.07). The interaction effect
tells us that contact has a more positive effect on smoking attitude for
higher levels of campaign exposure. As a result, the effect of contact at
average exposure is stronger than at zero exposure.

* Why do we have a statistical significant result for the effect of exposure
now but not in Exercise 1?
* The size of the unstandardized effect is lower (b = -0.17) now than in
Exercise 1 (b = -0.27). It is closer to zero so we would not expect
statistical significance. However, the standard error is much smaller now: SE
= 0.06 against SE = 0.17. We have quite some observations with about average
contact score (the reference value if we mean-center) but hardly any
observations with minimum (zero) contact score. With fewer observations, we
are less certain about estimates, so we have a larger standard error, and it
is more difficult to be confident that the rergession coefficient is not zero
in the population.
```

4. Check common support of the predictor for the moderator. Divide the moderator into three groups.

```
SPSS syntax:

* Group the moderator.
* Visual Binning.
*contact.
RECODE  contact (MISSING=COPY)
  (LO THRU 4.25076386584132=1)
  (LO THRU 5.83711577142397=2)
  (LO THRU HI=3) (ELSE=SYSMIS) INTO contact_3.
VARIABLE LABELS  contact_3 'Contact with smokers (Binned)'.
FORMATS  contact_3 (F5.0).
VALUE LABELS  contact_3 1 '' 2 '' 3 ''.
VARIABLE LEVEL  contact_3 (ORDINAL).
EXECUTE.
* Histograms of the predictor for each moderator group.
GRAPH
  /HISTOGRAM=exposure
  /PANEL ROWVAR=contact_3 ROWOP=CROSS.

Interpret the results:

* Coverage of the exposure predictor is poor at low contact levels. Especially
low exposure hardly occurs at low contact level. This explains the high
standard error for the exposure effect in Exercise 1 as explained in Exercise
3.
```

5. Let us hypothesize that children's media literacy depends on sex, age, and parental

supervision. Is the effect of parental supervision moderated by the child's age? Use children.sav to answer this research question and apply mean-centering. Report the results as required in this course (APA6), include a moderation plot, and discuss coverage.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=medliter sex age supervision
  /STATISTICS=MEAN
  /ORDER=ANALYSIS.
* Set impossible values to missing.
* Define Variable Properties.
*sex.
MISSING VALUES sex(1).
*supervision.
MISSING VALUES supervision(25.00).
EXECUTE.
* Turn sex into a 0/1 variable.
RECODE sex (2=0) (3=1) INTO girl.
VARIABLE LABELS  girl 'The child is a girl.'.
EXECUTE.
* Mean-center predictor and moderator.
* Ask for means of predictor and exposure.
FREQUENCIES VARIABLES=age supervision
  /FORMAT=NOTABLE
  /STATISTICS=MEAN
  /ORDER=ANALYSIS.
* Subtract mean from variable.
COMPUTE age_c=age - 8.609.
VARIABLE LABELS  age_c 'Age (mean-centered)'.
COMPUTE supervision_c=supervision - 5.358.
VARIABLE LABELS  supervision_c 'Supervision (mean-centered)'.
EXECUTE.
* Check mean centering.
FREQUENCIES VARIABLES=age_c supervision_c
  /FORMAT=NOTABLE
  /STATISTICS=MEAN
  /ORDER=ANALYSIS.
* Compute interaction variable.
COMPUTE age_supervision_c=age_c * supervision_c.
VARIABLE LABELS  age_supervision_c 'Interaction age * supervision (mean-centered)'.
EXECUTE.
* Multiple regression.
* Statistic Descriptives is added to get the means that we need
* to plug into the regression equation in the moderation plot.
REGRESSION
```

```
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT medliter
  /METHOD=ENTER girl age_c supervision_c age_supervision_c
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID).
* Create scatterplot for moderation plot.
* Use the mean-centered variable.
GRAPH
  /SCATTERPLOT(BIVAR)=supervision_c WITH medliter
  /MISSING=LISTWISE.
* Manually add three regression lines.

Check data:

* Score '25' for parental supervision cannot be right because the scale runs
to 10. Define this score as a missing value.
* The sex category '1' cannot be right either.

Check assumptions:

* The residuals are quite normally distributed, as they should.
* The residuals are centered around zero for all levels of the predicted
outcome (linearity) but the variation in residuals seems to be a bit larger at
higher predicted values (the residuals may not be homoscedastic).

Interpret the results:

* The regression model predicts 19 per cent of the differences in media
literacy among children, F (4, 80) = 4.58, p = .002.
* There is no remarkable difference between girls and boys, t = 0.58, p =
.566, 95%CI[-0.47; 0.86].
* Girls may be upto 0.86 more media literate on average than boys but we
cannot rule out that boys have on average more media literacy (up to 0.47).
* Age has a statistically significant positive effect on media literacy for
children at average parental supervision, t = 2.02, p = .047, 95%CI[0.003;
0.35].
* Parental supervision has a positive effect on media literacy, t = 3.36, p =
.001, 95%CI[0.12; 0.47].
* There is no statistically significant interaction effect between age and
parental supervision on media literacy, t = .50, p =.615, 95%CI[-0.07; 0.12].
If there is an interaction effect in the population, it can be negative nearly
```

```
as well as positive.

* Write out the regression equations for different values of the moderator.
* Plug in the estimated values of the regression coefficients, the selected
category of the covariate, and three values for the moderator using its M and
SD.
* Create regression lines for the effect of media literacy at three levels of
parental supervision (M - SD, M, and M + SD) in the scatterplot of media
literacy by mean-centered parental supervision.

Estimated regression equation:

medliter = 4.325 + 0.193 * girl + 0.176 * age_centered + 0.292 * supervision_centered +
  0.025 * age_c * supervision_c

With rearranged terms and sex plugged in for boys:

medliter = 4.325 + 0.193 * 0 + 0.176 * age_centered +
  (0.292  + 0.025 * age_c) * supervision_centered

Age at M - SD (mean-centered so M = 0, SD = 1.937) for boys:

medliter = 4.325 + 0.176*(0 - 1.937) + (0.292 + 0.025*(0 - 1.937))*supervision

medliter = 4.325 + -0.341 + (0.292 + -0.048)*supervision

medliter = 3.984 + 0.244*supervision

Age at M (M = 0) for boys:

medliter = 4.325 + .176*(0) + (0.292 + .025*(0))*supervision

medliter = 4.325 + 0.292*supervision

Age at M + SD (M = 0, SD = 1.937) for boys:

medliter = 4.325 + .176*(0 + 1.937) + (0.292 + 0.025*(0 + 1.937))*supervision

medliter = 4.325 + 0.341 + (0.292 + 0.048)*supervision

medliter = 4.666 + 0.340*supervision

Note: Age was mean-centered for all 87 cases but the regression model only
uses the 85 cases without a missing value on any of the variables. As a
result, the mean of age scores in the regression model is not exactly zero. We
can still use zero here because we merely want to refer to an age value that
```

```
is in the center of the distribution. Both zero and nearly zero are in the
center.

* The regression lines in the moderation plot have quite similar slopes, which
illustrates the absence of a substantial interaction effect.
```

6. Is the effect of parental supervision moderated by sex? Use the data of Exercise 5 to answer this question. You may omit the age predictor from the model. Again, illustrate your answer with a moderation plot.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=medliter sex supervision
  /STATISTICS=MEAN
  /ORDER=ANALYSIS.
* Set impossible values to missing.
* Define Variable Properties.
*sex.
MISSING VALUES sex(1).
*supervision.
MISSING VALUES supervision(25.00).
EXECUTE.
* Turn sex into a 0/1 variable.
RECODE sex (2=0) (3=1) INTO girl.
VARIABLE LABELS  girl 'The child is a girl.'.
EXECUTE.
* Mean-center the predictor.
* Ask for means of parental supervision.
FREQUENCIES VARIABLES=supervision
  /FORMAT=NOTABLE
  /STATISTICS=MEAN
  /ORDER=ANALYSIS.
* Subtract mean from variable.
COMPUTE supervision_c=supervision - 5.358.
VARIABLE LABELS  supervision_c 'Supervision (mean-centered)'.
EXECUTE.
* Compute interaction variable.
COMPUTE girl_supervision_c=girl * supervision_c.
VARIABLE LABELS  girl_supervision_c 'Interaction girl * supervision (mean-centered)'.
EXECUTE.
* Multiple regression.
* Statistic Descriptives is added to get the means that we need
* to plug into the regression equation in the moderation plot.
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
```

```
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT medliter
  /METHOD=ENTER girl supervision_c girl_supervision_c
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID).
* Scatterplot with dots coloured by sex.
* Use the mean-centered predictor.
GRAPH
  /SCATTERPLOT(BIVAR)=supervision_c WITH medliter BY girl
  /MISSING=LISTWISE.
* Note: This model does not contain a covariate, so SPSS can draw the lines.
* Provided that your SPSS installation contains the command:
* Graphs > Regression Variable Plots; Color by: sex..
* With options: Scatterplot Fit Lines: Linear,
*    Grouping: Fit Line for each categorical colour group.
* Use the mean-centered or not mean-centered predictor.
STATS REGRESS PLOT YVARS=medliter XVARS=supervision_c COLOR=sex
/OPTIONS CATEGORICAL=BARS GROUP=1 INDENT=15 YSCALE=75
/FITLINES LINEAR APPLYTO=GROUP.

Check data: See Exercise 5.

Check assumptions:

As with the regression model in Exercise 5.
* The residuals are quite normally distributed and centered around zero for
all levels of the predicted outcome (linearity).
* The variation in residuals seems to be a bit larger at higher predicted
values (the residuals may not be homoscedastic).

Interpret the results:

* The regression model predicts 15 per cent of the variation in outcome
scores, F (3, 81) = 4.76, p = .004.
* There is no remarkable difference between girls and boys, t = 0.37, p =
.711, 95%CI[-0.55; 0.80] for children at average supervision level. Girls may
have up to 0.80 more media literacy on average than boys but we cannot rule
out that boys have on average more media literacy (up to 0.55).
* Parental supervision has a statistically significant positive effect on
media literacy for boys, b = 0.39, t = 2.99, p = .004, 95%CI[0.13; 0.65] that
is weaker than for girls, b = 0.26 but the difference between boys and girls
is not statistically significant, t = -0.76, p = .448, 95%CI[-0.48; 0.22].
```

```
* If you add the regression lines for boys and girls manually, use the
mean-centered supervision variable and the following equations:

Equation for boys (girl = 0):

medliter = 4.374 + 0.125*girl + 0.389*supervision + -0.134*girl*supervision

medliter = 4.374 + 0.125*0 + (0.389 + -0.134*girl)*supervision

medliter = 4.374 + (0.389 + -0.134*0)*supervision

medliter = 4.374 + 0.389*supervision

Equation for girls (girl = 1):

medliter = 4.374 + 0.125*girl + 0.389*supervision + -0.134*girl*supervision

medliter = 4.374 + 0.125*1 + (0.389 + -0.134*girl)*supervision

medliter = 4.499 + (0.389 + -0.134*1)*supervision

medliter = 4.499 + 0.255*supervision
```

8.8 Test Your Understanding

Figure 8.33 shows how much respondents were exposed to an anti-smoking campaign (horizontal axis) and their attitudes towards smoking, ranging from negative (0) to positive (5, vertical axis). A third variable measures the extend to which respondents have daily contact with people who smoke. Does the contact variable moderate the effect of exposure on attitude?

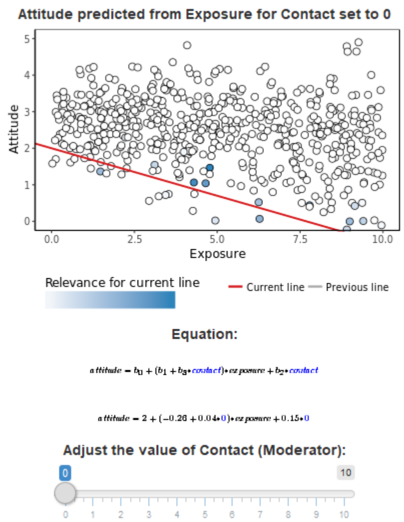


Figure 8.33: How does moderation work in a regression model?

1. What does the red line in Figure 8.33 mean?
2. What happens if you change the position on the slider? Explain you answer.
3. Why does *contact* (with smokers) appear in between brackets together with the regression coefficient for exposure in the regression equation?
4. Which of the regression coefficients represent(s) a partial effect and which a conditional effect? Explain your answer.
5. What is the null hypothesis of a significance test on the interaction effect (b_3)?

8.9 Take-Home Points

- In a regression model, moderation means that there are different slopes (of the predictor) for different groups or contexts (moderator).
- Interaction variables represent moderation in a regression model.
- An interaction variable is the product of the predictor and moderator. If the moderator is categorical, it is represented by one or more dummy variables. There is an interaction variable for each of the moderator's dummy variables.
- Statistical inference for an interaction variable is exactly the same as for "ordinary" regression predictors.
- The effect of the predictor in a model with an interaction variable does *not* represent a main or average effect. It is a conditional effect: The effect for cases that score

zero on the moderator. The same applies to the effect of the moderator, which is the conditional effect for cases scoring zero on the predictor.

- Mean center a numeric moderator and a numeric predictor that are involved in an interaction effect. Observations with a mean score on the moderator are a substantively interesting reference group.
- To interpret moderation, describe the effects (slopes, unstandardized regression coefficients) and preferably visualize the regression lines for different groups. For a numerical variable, select some interesting levels of the moderator, such as the mean and one standard deviation below or above the mean.
- Interpret regression lines for groups or moderator levels only if the predictor scores are nicely distributed for this group or level (common support).
- Don't use the standardized regression coefficients (Beta) for interaction variables in SPSS.

Chapter 9

Mediation with Regression Analysis

Key concepts: partial effect, statistical controlling for effects of other predictors, indirect correlation, confounders, suppression and suppressor, spuriousness and reinforcer, causality, causal and time order, common cause, antecedent variable, direct, indirect, and total effects, causal model, path diagram and path model, parallel and serial mediation, partial mediation, covariate, controlling mediator values.

Summary

If we analyze the effects of two or more predictors on an outcome variable in a regression model, the effect of a predictor is adjusted for the effects of other predictors. Each predictor only predicts the part of the outcome scores that cannot be predicted by the other predictors. If we predict newspaper reading time, for example, from age and interest in politics, age predicts the part of newspaper reading time that interest in politics cannot predict.

Because of adjustment for other predictors, adding new predictors to a regression model may change the effects of all predictors. The effects can become stronger (the new variable was a suppressor), weaker, and even change direction (the new variable was a reinforcer, the effect was partly spurious). For example, adding respondents' news site use to the regression model predicting newspaper reading time from age may change the effect of age on newspaper reading time. For this reason, we cannot be sure that the regression estimates represent the true effects of the predictors.

Indirect correlations play a central role here: The correlation of a predictor (age) with the outcome (newspaper reading time) due to the fact that the predictor (age) is correlated with another predictor (news site use) that is also correlated with the outcome (newspaper reading time). The size of the indirect correlation is simply the product of the correlation

between the two predictors (age and news site use) and the standardized regression effect of the other predictor (news site use) on the outcome variable (newspaper reading time).

If we add a causal order among the predictors of our regression model, we obtain a causal model or path model. The causal model includes an indirect effect: The first predictor affects the scores on the second predictor, which affects the outcome scores. For example, age affects news site use, which affects newspaper reading time. In this path model, the second predictor (news site use) mediates the effect of the first predictor on the outcome variable. The second predictor is called a mediator.

We can estimate a path model as a series of regression models. With additional software, we can also estimate the confidence interval of an indirect effect. The causal order underlying the path model, however, is an assumption that we make. A regression model shows the predictive effects, which need not be causal. We cannot prove that the predictive effects are causal. We can only think of arguments that make a causal effect plausible.

9.1 Controlling for Effects of Other Predictors

In a regression model, we use the variation in scores on predictor variables to predict the variation of scores in the outcome variable: Does a person with a higher score on a predictor also have a higher score or, on the contrary, a lower score on the outcome variable? A simple regression model contains only one predictor but a multiple regression model includes more than one.

For example, European citizens who are older spend more time on reading newspapers and so do citizens who are more interested in politics. We have two predictors (age and interest in politics) to predict the outcome variable (newspaper reading time). The two predictors can be correlated: Older citizens tend to be more interested in politics. How does the regression model decide which predictor is responsible for which part of the variation in outcome scores?

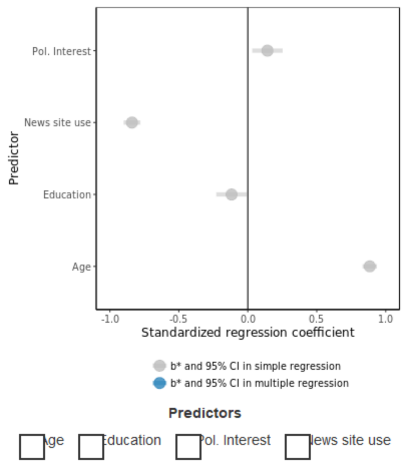


Figure 9.1: How do regression coefficients change if new predictors for reading time are added to the model? The grey dots and lines represent the simple regression coefficients and their 95 per cent confidence intervals in a model predicting newspaper reading time. Blue dots and lines represent results in a regression model including all selected predictors.

1. Select Age and Education in Figure 9.1. Compare the regression coefficients and their confidence intervals for the multiple regression model including the selected predictors (blue dots and lines) to the results for the simple regression models including only one predictor (gray dots and lines).

```
* The effect of age remains more or less the same, around 0.9, if we control
for education.
* The negative effect of education if we do not control for age (gray),
disappears if we control for age (blue). The blue dot for education is near
zero (and positive rather than negative) and the confidence interval includes
zero.
```

2. Try other combinations of Age and a second predictor. When does the regression coefficient of Age change markedly? Can you explain why it changes as it does?

```
* The regression coefficient of age changes most when we add news site use as a
predictor to the model. The effect of age becomes weaker (around 0.7).
* Using internet as news source probably reduces newspaper reading time because
news can be retrieved from the internet. Older people are less likely to use
the internet as news source, so part of the positive effect of age on newspaper
reading time is actually due to the fact that older people use news sites less
instead of their age. That part of the age effect disappears if we add news
site use as predictor to the model.
```

3. Interpret the meaning of the simple regression models and explain why they do not change if predictors are added. In which situation are the light and dark results equal?

```
* A simple regression model contains only one predictor. The regression
coefficient represents the effect of this predictor without controlling for the
effects of other predictors.
* These effects are positive for age and political interest in the example:
Older people and people who are more interested in politics are predicted to
spend more time on reading newspapers. The effects are negative for education
and news site use: People with more (years of) education or more frequently
using news sites are predicted to spend less time on reading newspapers.
* If we select only one confounder, we have a simple regression model, so the
predictor's blue effect must be equal to its gray effect.
```

9.1.1 Partial effect

Conceptually, the regression model first removes the variation in the outcome that is predicted by all other predictors. Then it determines how well the variable predicts the variation that is left (residual variation). This is the variation in outcome scores than can be predicted by this particular variable but not by any of the other predictors in the model.

In this sense, a regression coefficient in a multiple regression model expresses the unique contribution of a variable to the prediction of the outcome. It is the contribution to the prediction of the outcome variable over and above the predictions that we can make with all other predictors in the model. This is called a *partial effect*.

This is what we mean if we say that we are *controlling for all other predictors* in our interpretation of a regression model. We are not controlling as in an experiment where we ensure that participants get a particular value on a predictor (treatment) variable. It is controlling in a statistical sense, using the data that we have collected.

9.1.2 Confounding variables

It is important to note that the effect is only unique in comparison to the other predictors that are included in the model. It may well be that we did not include variables in the model that are actually responsible for part of the effects that are attributed to the predictors in the model. Such left-out variables are called *confounding variables* or, for short, *confounders*.

If we include a confounder as a new predictor in the model, the partial effects of old predictors change. In Figure 9.1, for instance, this happens if you add news site use to a model containing age as a predictor for newspaper reading time. The effects of old predictors are adjusted to a new situation, namely a situation with an added new predictor. The new predictor helps to predict variation in the outcome variable, so the variation left to be explained by old predictors changes. In Section 9.3, we will learn that regression coefficients can go up and down if confounders are included in the model.

With non-experimental data, for example data on newspaper reading habits gathered with a survey, we must include all confounders to ensure that the estimated regression coefficients represent the effects of the predictors and not effects of confounders. In practical applications, we can only include a limited number of variables in our research and we do not always know what are the confounders. We should strive to include the most important confounders in our research project but we should always keep in mind that the predictive effects that we find may be due to variables not included in our regression model.

9.2 Indirect Correlation

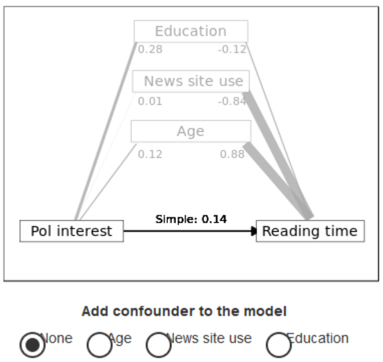


Figure 9.2: What happens to the regression coefficient if we add a confounder to the model? Numbers represent correlations (lines) or standardized regression coefficients (arrow).

1. Which variable confounds the effect of political interest on newspaper reading time most? And why is that so? After formulating your answer, add confounders to the model in Figure 9.2 to check if you are right.

```
* Age is the strongest confounder in this example. If we add age as a
predictor to the model, the partial effect of political interest on reading
time is 0.04. The (simple) effect of this predictor is 0.14 if we do not
control for age, so the effect difference is 0.10.
* The difference between the simple and partial effect of political interest
on reading time is smaller for news site use and education.
* Because age is most strongly related to both political interest and reading
time, it affects the effect of political interest on reading time most
strongly.
```

When is a variable a confounder and when is it a stronger confounder, changing the effect of another predictor a lot? The answer to the first part of this question is easy: A *confounder* is a variable that is correlated with both the predictor and outcome but it is not (yet) included itself in the regression model. Because of the correlations, a confounder establishes an *indirect correlation* between the predictor and outcome variable.

The size of the indirect correlation is equal to the product of the correlation between confounder and predictor and the correlation between confounder and outcome variable. In Figure 9.2, the correlation between age and interest in politics is .12 and the correlation between age and newspaper reading time is .88, the indirect correlation between interest in politics and reading time established by age is .12 * .88 = .11.

9.2.1 Multiplication of correlations

Multiplication makes intuitively sense. If two predictors are perfectly correlated ($r = 1$), the first predictor can exactly predict the second predictor. As a consequence it can predict the outcome via the second predictor (indirect correlation) just as well as the second predictor can predict it itself. For example, if age is perfectly correlated with interest in politics ($r = 1$) and the correlation between interest in politics and newspaper reading time is .88, the latter correlation is exactly the same as the indirect correlation between age and reading time via interest in politics that we get through multiplication (1.00 * .88 = .88). Here, age predicts reading time via interest in politics just as well as interest in politics predicts reading time.

If the correlation between the two predictors is below one (disregarding the sign of the correlation), the first predictor can only predict a proportion of the second predictor. As a consequence, the first predictor can also predict only a proportion of the outcome via the second predictor. This is a proportion of what the second predictor can predict by itself. A correlation below one (and above minus one) is a proportion, so multiplying the second predictor's correlation by the correlation between the two predictors does the job.

9.2.2 Indirect correlation and size of confounding

In Figure 9.2, we start with a simple regression model with political interest as the only predictor of newspaper reading time. Respondent's age, however, is correlated both with political interest ($r = 0.12$) and with newspaper reading time ($r = 0.88$). Age creates a positive indirect correlation between political interest and reading time.

As long as age is not included in the regression model, the model believes that the indirect correlation via age is part of the effect of political interest. It assigns the indirect correlation via age erroneously to the partial effect of political interest, that is, to the regression coefficient of political interest. In this situation, the regression coefficient for political interest expresses both the effect of the political interest itself and the effect of age (the confounder).

Once we add age as a new predictor to the regression model, the indirect correlation via age is removed from the effect of political interest. The effect of age on newspaper reading time is now correctly assigned to age. As a result, the value of the regression coefficient for political interest changes if we add age as a new predictor.

The size of the change usually is closely related to the size of the indirect correlation, so the larger the indirect correlation, the more the regression coefficient of political interest changes if age (the former confounder) is included as a new predictor. This answers the second part of the question with which we started Section 9.2: When is a variable a stronger confounder?

If you love the details: The size of the change is not exactly the same as the size of the indirect correlation. It is equal to the correlation between the confounder (age) and the predictor (political interest) times the partial effect of the confounder (age) on the outcome variable (newspaper reading time) that controls for the effect of the predictor (political interest). This quantity is subtracted from the effect of the predictor (political interest) if we add age as a new predictor to the regression model.

9.2.3 Confounders are not included in the regression model

Finally, it is important to remember again that a confounder, such as age in the present example, is a variable that is not included in our regression model. As long as it is not included, the indirect correlation between predictor (political interest) and outcome (newspaper reading time) via the confounder (age) is not controlled for when the effect of the predictor is estimated. The estimated effect is not correct, it is confounded (confused) with the effect of the confounding variable.

Once the confounder (age) is added to the regression model, however, the estimated effects are controlling for the variable formerly known as a confounder. The effects no longer partly represent the effect of the former confounder. In other words, they are no longer confounded by the effect of that variable. The former confounding variable now is a predictor or, if we are not interested in its effects, a control variable in the regression model.

9.3 Two Types of Confounders

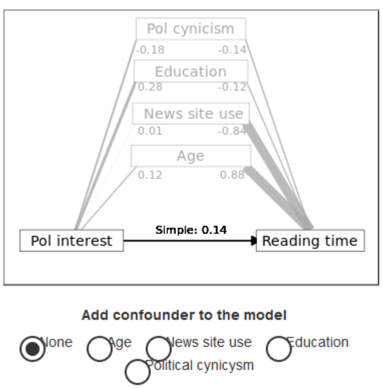


Figure 9.3: When is a regression effect too large and when is it too small due to a confounder?

1. When is a partial regression effect weaker than the simple regression effect and when is it stronger? Formulate a general rule and check it in Figure 9.3.

```
* If the indirect correlation has the opposite sign of the direct effect of
political interest on reading time, which is positive here, the partial effect
is stronger than the simple effect. An indirect correlation is negative if one
of its two correlations is negative and the other is positive.
* If the indirect correlation has the same sign as the direct effect of
```

```
political interest on reading time, the partial effect is weaker than the
simple effect. An indirect correlation is positive if none or both of its two
correlations are negative.
* Note that the simple effect is the regression coefficient if we do not
control for any confounders, so political interest is the only predictor. The
partial effect is the regression coefficient for the effect of political
interest on reading time if we control for (at least) one other predictor.
Then, political interest is not the only predictor in the regression model.
```

2. According to your general rule, what happens to the partial regression effect if we would add interest in sports as a predictor, which is negatively correlated with interest in politics but positively correlated with newspaper reading time?

```
* The indirect correlation between Political interest and Reading time created
by Interest in sports would be negative. It would be the opposite of the sign
of the direct effect, so the partial effect would be stronger.
```

In the preceding section, we learned that a partial effect expressed by a regression coefficient may change if a new predictor is added to the regression model. The partial effect of a predictor changes if the added variable is a confounder: It is correlated both with the predictor and outcome variable. In other words, there is an indirect correlation between the predictor and outcome via the confounding variable.

The partial effect of a predictor can become stronger, weaker, or even change direction if we add a confounder to the regression model. The current section describes the two types of confounders that are responsible for these changes: suppressors and reinforcers.

9.3.1 Suppression

A predictor's effect becomes stronger if we include a confounding variable that is responsible for an indirect correlation that points in the opposite direction of the effect of the predictor. Here, the indirect correlation contradicts the effect of the predictor and as a result, the effect of the predictor is underestimated (suppressed) if the confounder is not included in the model.

There are two situations in which an indirect correlation can have the opposite sign of the effect of a predictor:

1. The indirect correlation is negative but the effect of the predictor is positive.

2. The indirect correlation is positive but the effect of the predictor is negative.

Let us start with the first situation and discuss the second situation later on in this section.

Let us assume that political interest has a positive effect on reading newspapers. People who are more interested in politics tend to spend more time on reading newspapers than people who are less interested in politics. The use of news media confounds this effect if it is correlated with both political interest and newspaper reading time. What happens if people

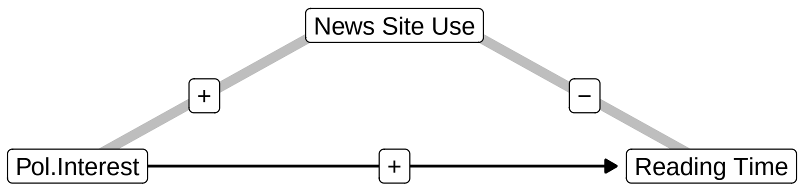


Figure 9.4: News site use as a confounder of the effect of interest in politics on newspaper reading time.

interested in politics use news sites more often (positive correlation) but using news sites decreases newspaper reading time (negative correlation)?

In this situation, the indirect correlation between political interest and newspaper reading time via news site use is negative: Positive times negative yields a negative. The indirect correlation tells us that people interested in politics use news sites more frequently and for that reason read less newspapers. The indirect correlation clearly contradicts the regression effect of political interest on newspaper reading time, which is positive: People who are more interested in politics spend more time on reading newspapers.

If news site use—the confounder—is not included in the regression model, the standardized regression effect of political interest more or less adds the indirect correlation to the effect of political interest. Adding a negative amount (indirect correlation), however, is equal to subtracting this amount from the standardized regression coefficient. The positive effect of political interest on reading time is underestimated. In this example, the effect of political interest is *suppressed* (*masked*) by the confounder news site use. News site use is a *suppressor variable*.

If we include this suppressor variable (news site use) in our regression model, we eliminate its suppression of the effect of political interest on newspaper reading time. The negative effect of news site use on reading time is captured by the regression coefficient for the new news site use predictor. The effect of political interest on newspaper reading time is now controlled for the effect of news site use; it no longer includes the indirect correlation via news site use.

Now, let us have a look at the situation in which the indirect correlation is positive but the regression effect of the predictor is negative. Just reverse the example and make news site use the predictor and political interest the confounder. The regression effect of news site use on newspaper reading time is negative if people tend to use news sites instead of newspapers as sources of information. The indirect correlation via political interest, however, is positive if politically interested people use news sites more and spend more time on reading newspapers. In this scenario, the negative effect of news site use on newspaper reading is estimated too low if we do not control for political interest.

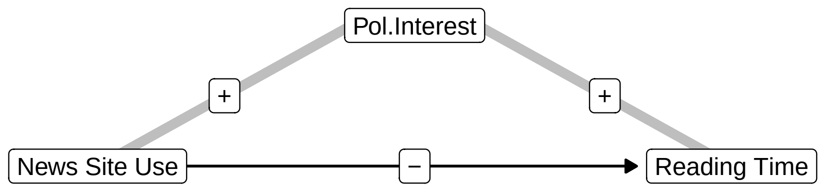


Figure 9.5: Interest in politics as a confounder of the effect of news site use on newspaper reading time.

Suppression can have surprising effects. If the predictor's original effect was close to zero, adding a suppressor variable to the model will strengthen the effect. An effect that we initially believed to be absent may turn out to be substantial and statistically significant. If our regression model tells us that our predictor does not have an effect, we cannot rule out that it does have an effect that is suppressed by a suppressor variable.

Finally, indirect correlations via other predictors can add so much to the original partial effect of a predictor that the standardized regression coefficient becomes larger than 1 or smaller than -1. This clearly illustrates that standardized regression coefficients are not correlations in multiple regression models because correlations can never be higher than 1 or lower than -1. In contrast, the standardized regression coefficient in a simple regression model is equal to the correlation between predictor and outcome. Isn't that confusing?

9.3.2 Reinforcement and spuriousness

Adding a new predictor to a regression model may weaken the effects of other predictors. This happens if the indirect correlation due to a confounder has the same direction (sign) as the regression effect of the predictor. Either both the indirect correlation and regression effect are both positive or they are both negative.

In both situations, regression effects are initially overestimated because the predictors cover part of the effect of an important variable that had not yet been added to the regression model. The part of the effect that is due to the confounding variable is called *spurious*. The confounding variable is called a *reinforcer* because it makes an effect appear stronger than it really is as long as the confounder has not been added to the regression model.

As an example, the effect of political interest on newspaper reading time may include the effect of age on newspaper reading. If older people are more interested in politics and do more newspaper reading, age creates a positive indirect correlation between political interest and newspaper reading.

If age is not included as a predictor in the regression model, the indirect correlation is mistakenly attributed to the effect of interest in politics. The estimated effect is too strong.

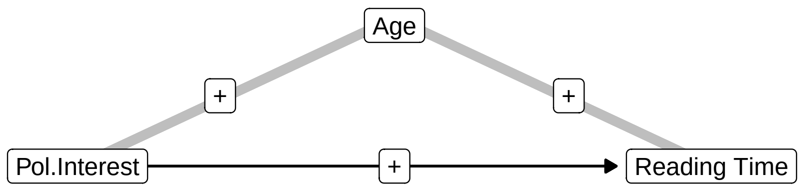


Figure 9.6: Age as a confounder of the effect of interest in politics on newspaper reading time.

Once we include age as a predictor, the effect of political interest is cleansed of the age effect, so the effect size decreases.

In Figure 9.6, age is positively correlated with both political interest and newspaper reading. But a confounder that is negatively correlated with predictor and outcome has the same impact as a confounder that is positively correlated with predictor and outcome. Political cynicism, for instance, can be negatively correlated with both interest in politics and newspaper reading. Similar scenarios are available if the regression effect is negative.

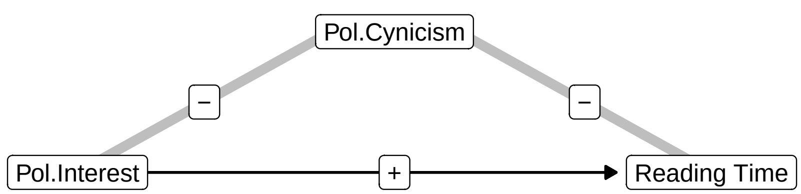


Figure 9.7: Political cynicism as a confounder of the effect of interest in politics on newspaper reading time.

As with suppression, spuriousness can have surprising results. It may happen that the entire estimated effect of a predictor is spurious. Adding a reinforcer variable to the regression model may make the entire effect of a predictor disappear. In other words, an effect that we initially thought was substantial may turn out to be too weak to be of interest.

Actually, the indirect correlation between a predictor and outcome via a confounding variable can be so strong that a positive effect in a model without the confounder changes into a negative effect in a model that includes the variable. The opposite may happen as well: A formerly negative effect may become positive if a strong positive reinforcer variable is added to the model.

The bottom line is simple: We can only trust the results if all important confounders are included in the model.

9.4 Comparing Regression Models in SPSS

9.4.1 Instructions

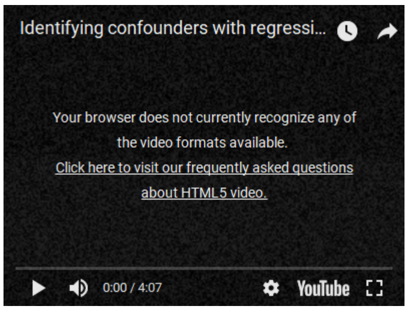


Figure 9.8: Identifying confounders with regression in SPSS.

9.4.2 Exercises

1. The file readers.sav contains the data on newspaper reading time that we have used as example in this chapter. Predict average newspaper reading time by education, interest in politics, news site use, and age. Add the predictors one by one to the regression model in the order specified in the preceding sentence. How do the partial effects change when new predictors are added?

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=age education polinterest newssite readingtime
  /ORDER=ANALYSIS.
* Multiple regression.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT readingtime
  /METHOD=ENTER education
  /METHOD=ENTER polinterest
  /METHOD=ENTER newssite
  /METHOD=ENTER age
```

```
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID).

Check data:

* The variabes do not have impossible values.

Check assumptions:

* The residuals are quite normally distributed.
* They are centered around zero at all levels of the predicted outcome, so a
linear model seems to fit the data.
* The variation in residuals is about the same at all levels of the predicted
outcome, so the outcome is more or less equally well predicted at all levels
of the outcome variable.
* As a conclusion, there are no clear indications that the assumptions for a
linear regression model are violated.

Interpret the results:

* Given the number of regression coefficients that are estimated, it is
helpful to present them as a table instead of reporting all results in the
interpretation.

* Education has a weak negative predictive effect on newspaper reading time
until news site use is introduced as a predictor. Then its effect vanishes.
* Interest in politics has a weak positive predictive effect on newspaper
reading time. This effect becomes weaker when news site use is added as a
predictor and especially when age is added.
* Initially, news site use has a strong negative predictive effect on
newspaper reading time but the effect is much weaker when age is added as a
predictor.
* Age has a strong positive effect on newspaper reading time if we control for
all other predictors in the model.
```

2. In the series of models that you estimated in Exercise 1, for which predictors is age a suppressor or reinforcer? For advanced understanding: Use the correlations between the variables to justify your answers.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=age education polinterest newssite readingtime
  /ORDER=ANALYSIS.
* (Pearson) Correlations.
CORRELATIONS
```

```
  /VARIABLES=age education polinterest newssite readingtime
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.

Interpret the results:

* If the partial effect of a predictor increases (becomes more strongly
positive or more strongly negative) when a new predictor is added to the
regression model, the added predictor was a suppressor while it was not yet
included in the model.
* When age is added as a predictor to the model, none of the partial effects
of other predictors increases, so age did not suppress any of the effects in
the models that did not include age yet.

* If the partial effect of a predictor decreases (becomes less strongly
positive or less strongly negative) or it changes sign when a new predictor is
added to the regression model, the added predictor was a reinforcer while it
was not yet included in the model.
* This happens to the partial effects of education, interest in politics, and
news site use when age is added as a predictor. For these three predictors,
age was a reinforcer. Part of their effects were spurious in the models
without age. They seemed to be effects of education, political interest, and
news site use but they are actually effects of age.

* The signs of the correlations between predictor, confounder, and outcome
variable tells us if a confounder is a suppressor or reinforcer.
* If the sign (direction) of the indirect correlation between predictor and
oucome via the confounder (age) is equal to the sign of the correlation
between predictor and outcome, the confounder is a reinforcer. In this case,
part of the predictive effect is due to the confounder.
* For example, Education is negatively correlated with newspaper reading time.
The indirect correlation via age also is negative because age is negatively
correlated with education but positively with newspaper reading time. Negative
times positive yields a negative.

* For a suppressor, the indirect correlation must be of the opposite sign of
the correlation between predictor and outcome. This situation does not occur
for age as the confounder in the models estimated in Exercise 1.
```

3. Predict average newspaper reading time from education, political cynicism, news site use, and age. Add the predictors one by one to the regression model in the order specified in the preceding sentence. Does suppression occur here? If so, for which predictor and confounder and in which model(s)?

```
SPSS syntax:
```

```
* Check data.
FREQUENCIES VARIABLES=age education polcynic newssite readingtime
  /ORDER=ANALYSIS.
* Multiple regression.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT readingtime
  /METHOD=ENTER education
  /METHOD=ENTER polcynic
  /METHOD=ENTER newssite
  /METHOD=ENTER age
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID).

Check data:

* In addition to the other variables (see Exercise 1), political cynicism does
not have impossible values.

Check assumptions:

* The residuals are similarly distributed as in Exercise 1, so we may assume
that assumptions are met.

Interpret the results:

* When political cynicism is added as a predictor, the negative effect of
education becomes more strong. In the model without political cynicism,
political cynicism suppressed part of the effect of education.
```

9.5 Mediation as Causal Process

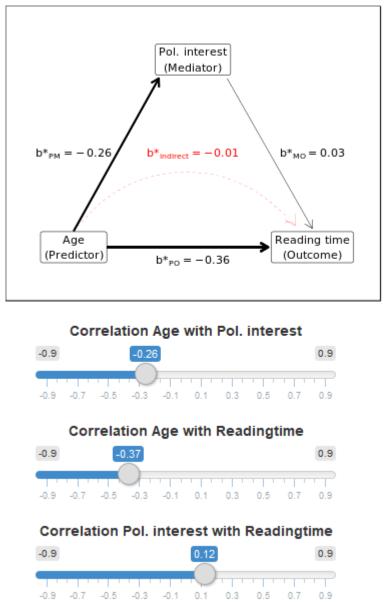


Figure 9.9: How does a common cause affect regression coefficients? The values in this path diagram represent standardized regression coefficients.

1. In Figure 9.9, what does the long, curved arrow represent? Can you motivate your answer with the values that are linked to the arrows?

```
* The curved arrow is the standardized indirect (predictive) effect of age on
reading time.
* It summarizes the assumed causal process that age influences political
interest, which influences reading time. For example, older people are more
interested in politics and because they are more interested in politics, they
spend more time on reading newspapers.
* Technically, the standardized indirect effect represents the predicted
difference (in standard deviations) of reading time that results from a
difference in political interest that results from a difference of one
standard deviation in age.
* An indirect effect is equal to the product of the direct effects. In the
example, the value of the indirect effect is indeed equal to the product of
the effects of age on political interest and political interest on reading
time.
```

2. In a mediation model such as Figure 9.9, which standardized effects are always equal to the correlation between the corresponding variables? Explain why this is so.

```
* A correlation in a mediation model can differ from the standardized
regression coefficient in a mediation model in two ways:

1. Part of the correlation is spurious (suppressed or reinforced) because the
two variables involved in the correlation have a common cause in the model.

2. Part of the correlation is actually an indirect effect, so it is subtracted
from the correlation to obtain the direct effect (standardized regression
coefficient).

* Ad 1: In our mediation model, the predictor is a common cause to the
mediator and outcome. As a result, the correlation between the latter two
variables (political interest and reading time in the example) need not equal
the standardized regression effect.
* Ad 2: Our mediation model can only contain an indirect effect of the
predictor (age) on the outcome (reading time), namely, via the mediator
(political interest). As a result, the regression coefficient of the effect of
age on reading time need not equal their correlation coefficient.
* The predictor-mediator pair (age and political interest) does not have a
common cause in this model nor can there be an indirect effect, so the
standardized regression coefficient for the effect of age on political
interest is always equal to their correlation coefficient in this model.
* Later on, we will see that we can estimate this model with two regression
equations. The first model regresses reading time on both political interest
and age. This is a multiple regression, so effects are controlled for effects
of other predictors. As a consequence, the standardized regression
coefficients need not equal the correlation coefficients. The second
regression model regresses political interest on age. This is a simple
regression model in which we do not control for effects of other predictors.
In such a model, the standardized regression coefficient is always equal to
the correlation.
```

3. Adjust the correlations in such a way that the effect of age on reading time is fully mediated by political interest. How can we see that the effect is fully mediated?

```
* An effect of a predictor on an outcome is fully mediated if the indirect
effect via the mediator is equal to the overall correlation between the
predictor and the outcome. In other words, the correlation only represents the
indirect effect, so there cannot be a direct effect.
* Of course, it makes no sense to say that an effect is fully mediated if
there is no indirect effect (and no direct effect). So we must start with a
non-zero correlation between predictor and outcome.
* As a next step, we must select values for the correlation between predictor
(age) and mediator (political interest) and for the correlation between
mediator and outcome (reading time) such that their product (the indirect
correlation) is equal to the correlation between age and reading time. Use,
for instance, 0.3 for the correlation between age and reading time, 0.5 for the
```

```
correlation between age and political interest, and 0.6 for the correlation
between political interest and reading time. Note that 0.5 * 0.6 = 0.3.
```

9.5.1 Criteria for a causal relation

Researchers are usually interested in causal effects, so let us theorize a causal order between age and reading newspapers. From previous research and personal experience, we strongly suspect that older people spend more time on reading newspapers than young people. In statistical language, we expect a positive correlation between age and newspaper reading time. Can age be a cause of current newspaper reading?

Correlation is the first criterion. A causal relation implies correlation or another type of statistical association. If newspaper reading is not correlated with age, it is hard to imagine that age affects newspaper reading. But correlation does not imply causation, as the saying goes. Correlated variables need not be causally related. We need additional arguments to add plausibility to a causal relation.

The second criterion is the time order between cause and consequence. A consequence must appear after the cause. In our example, a person's age must be fixed before she displays the behaviour that we want to explain, namely reading newspapers. The time order is very plausible here because age stands for the moment a person was born, which must be prior in time to reading newspapers. If there is a causal relation between age and reading newspapers, age must be the cause and newspaper reading the consequence.

A third criterion for causality is that the correlation is not spurious. In Section 9.3.2, we have encountered a spurious effect as an effect that incorrectly includes the effect of a confounder (reinforcer).

In the context of causality, spuriousness is linked to a confounder that is a *common cause* to both the assumed cause (predictor) and consequence (outcome). Age, for instance, can be a common cause both to having (grand)children and reading newspapers. Older people tend to have more (grand)children and they read more newspapers. If we do not control for age when we regress newspaper reading time on the number of (grand)children a person has, we may find a positive effect.

This effect is probably not causal: We do not spend more time reading newspapers because we have more (grand)children. Unless we use newspaper reading to ignore our children and grandchildren when they are around. Most if not all of the effect of (grand)children on newspaper reading is spurious because it results from a common cause, namely age or the habits and opportunities represented by age.

To interpret the effect that we find as causal, then, we must ensure that there are no confounding variables that are common causes to both our predictors and outcome. Including them as controls in the regression model is a way to solve the problem. Unfortunately, we can never be sure that we have included all common causes in our model.

9.5.2 Mediation as indirect effect

A common cause need not remove the entire effect between a predictor and outcome. Even if part of newspaper reading is caused by age, another part can be caused by a variable related to age, for example, interest in politics. During their lifetime, people may gain more experience with politics and, for that reason, become more interested in reading about politics. This may cause them to invest more time in reading newspapers for collecting information.

Not all people become more interested in politics as they grow older and their interest in politics need not increase regularly during all of their lifetime. The relation between age and interest in politics, therefore, will not be perfect. This allows us to technically distinguish between the effect of age and the effect of interest in politics.

If we include both age and interest in politics as predictors in a regression model for newspaper reading time, the partial effect of interest in politics is corrected for the spurious correlation between interest in politics and newspaper reading caused by age as their common cause. The partial effect of political interest can be interpreted as causal if current interest in politics was attained before the newspaper readings that we measure (very plausible) and age is the only common cause of interest in politics and newspaper reading (highly questionable).

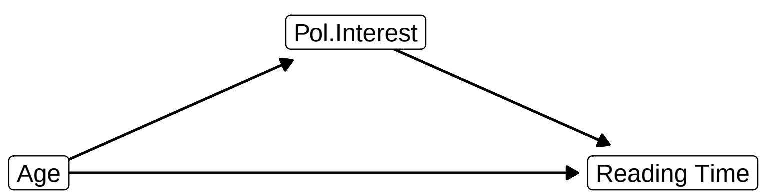


Figure 9.10: Causal diagram for the effects of age and interest in politics on newspaper reading time.

Now let us draw the *causal diagram* for this simple example (Figure 9.10). A causal diagram contains the names of the variables with arrows pointing from causes to consequences. The causal order of variables is represented from left to right. In Figure 9.10, the very first cause (age) is at the left, the final consequence (newspaper reading time) is at the right, and interest in politics is placed in the middle. In this layout, the arrows always point to the right.

In the causal order that we theorize, age is causally prior or *antecedent* to interest in politics, which is antecedent to current newspaper reading time. We have an *indirect effect* of age on newspaper reading by way of interest in politics. When adults grow older, they tend to be more interested in politics and because of this, they tend to spend more time on reading newspapers. We say that interest in politics *mediates* the effect of age on newspaper reading time. Interest in politics is a *mediator*, an *intermediary variable*, or an *intervening variable* in this causal diagram.

A causal diagram like Figure 9.10 is also called a *path diagram*. Each indirect effect is a sequence of direct effects. Each direct effect is a "step" from one variable to another variable, represented by an arrow. An indirect effect, then, can be regarded as a *path* that we can follow to "travel" from one variable to another variable.

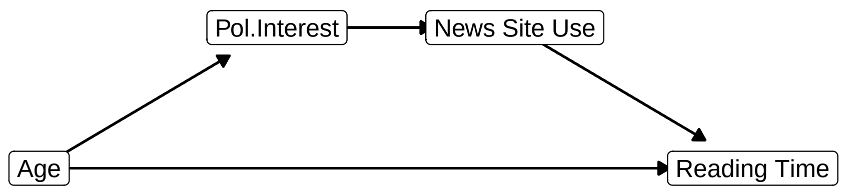


Figure 9.11: Causal diagram for the effect of age on newspaper reading time mediated by interest in politics and news site use.

An indirect effect may contain more than one step or mediator. If we include news site use in the model (Figure 9.11), we would have an indirect effect from age to interest in politics to news site use and finally to newspaper reading time.

9.5.3 Causal process

In our example (Figure 9.11), age has a direct effect on newspaper reading time. What does the direct effect mean? If we start thinking about why older people spend more time on reading newspapers, we soon realize that this is probably not some biological process. It is hard to believe that an ageing human body viewed apart requires more newspaper reading time. The effect is more likely to be social.

In the middle of the 20th century, newspapers were among the most important sources of information. A person who was born and grew up in that period is accustomed to using newspapers as main information source. For later generations, however, news sites on the internet have become important sources of information. Newspapers being less important to them, they are less oriented and accustomed to reading newspapers.

This line of reasoning shows us two things. First, we discover that our common cause may actually represent different things. Age, for instance, refers to life experience in its effect on interest in politics. In contrast, it relates to the period of coming of age in its direct effect on newspaper reading time.

Our second discovery is that we usually look for mediators if we want to understand a direct effect. Date of birth affects exposure to people using newspapers as information sources, which affects the habit of reading newspapers, which finally affects the time spend on reading newspapers later on. Exposure and habit are mediators here. A direct effect of age on newspaper reading merely replaces a causal process that may contain many intermediary steps. Adding mediators to our model is a way of getting more insight in the causal process.

9.6 Path Model with Regression Analysis

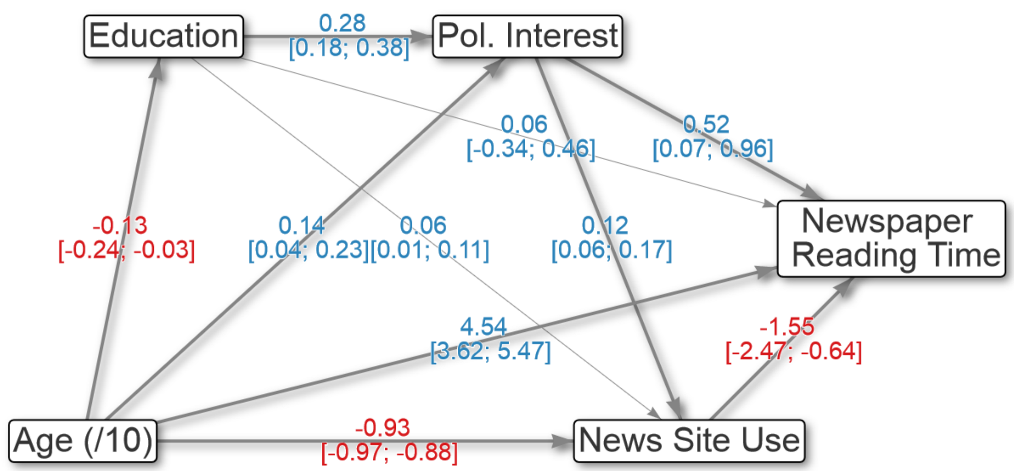


Figure 9.12: Path diagram with unstandardized effect sizes and their 95% confidence intervals.

1. In the causal diagram depicted in Figure 9.12, click on a variable name to highlight the effects in a regression analysis with this variable as the outcome. How many regression analyses does the causal diagram include and what do the results mean?

```
* There are four variables with incoming arrows in this model. Each of them can
be an outcome variable in a regression model, so we can estimate four
regression models.
* The unstandardized and standardized regression coefficients tell us the
(unstandardized or standardized) sizes of the direct effects in the path
model.
* For example, news site use has an unstandardized effect of -1.55 on
newspaper reading time in the diagram. A difference of one unit of news site
use predicts a decrease in newspaper reading time of 1.55 units. We are 95
per cent confident that this decrease is in between 0.6 and 2.5 units in the
population.
```

2. The fat arrows were hypothesized whereas the thin ones were not. Why are the thin ones included in the regression analyses (they have coefficient estimates too)?

```
* If we do not include effects between variables that were not hypothesized in
our regression model, we actually assume that the effects are zero.
* If this assumption is not correct, the results for hypothesized effects are
wrong.
* If we exclude the effect of a predictor on an outcome from the model, we
deny this predictor the opportunity to be a common cause to the outcome and
another predictor. Thus, we fail to signal and filter out the spurious
correlation due to this common cause.
* It is better to include all causally antecedent variables as predictors.
```

3. Which effect sizes with their confidence intervals cannot be obtained with regression analyses?

```
* We cannot get the sizes and confidence intervals of the indirect effects
directly from the regression results.
* Indirect effect sizes are easily calculated. Just multiply the regression
coefficients of all effects that make up the indirect effect.
* Confidence intervals and p values of indirect effects cannot be easily
calculated from the regression results. We use bootstrapping to this end.
```

Mediation or, more generally, path models can be estimated with a series of regression models. Every variable in the path diagram with at least one predictor (or incoming arrow) is an outcome variable, so for each of them, we estimate a regression model. The regression model contains all variables as predictors that may cause changes in the outcome variable. In other words, all variables that are causally antecedent to the outcome are used as predictors. In a well-designed causal diagram, all variables to the left of the outcome are antecedent to it.

In the path diagram displayed in Figure 9.12, we would regress newspaper reading time, the final outcome variable, on all other variables. As a next step, we would predict news site use as outcome from all variables except reading time.

Note that we include education level as predictor of news site use and newspaper reading even if we did not draw a direct arrow. Education level is theorized to be causally prior to news use site, so it can have a direct effect. We hypothesized that it did not have a direct effect on news site use and reading time, so we omitted arrows in our hypothesized model. We must include education as a predictor to check that it does not have these direct effects.

A third regression model would predict political interest from education level and age. The final regression model predicts education level from age .

9.6.1 Requirements

We can estimate mediation and path models with regression analysis if we meet the following requirements:

1. Each variable used as an outcome is numeric. This is a general requirement of a linear regression model. In a path diagram, it means that all mediators and outcome

variables must be numeric.

For detail lovers: Variables with only incoming arrows may be dichotomous but that requires logistic regression, which we do not discuss.

2. Each variable used as a predictor must be a numeric or dichotomous (dummy) variable. Again, this is a general requirement of regression models.

3. There are no causal feedback loops. Causality must work in one direction. It must be impossible to travel from a variable back to it while following the direction of the arrows. Note that it can be difficult to assign a causal order. For example, does political interest cause (low) political cynicism or the other way around? Or are they not causally related?

4. All regression models meet the assumptions for regression analysis. Check if the residuals are normally distributed, centered around zero for all levels of the predicted outcome, and that all outcome levels are predicted equally well (see Section 8.1.5).

9.6.2 Size of indirect effects

The regression results tell us the sizes and statistical significance of all direct effects on the outcome variable. Both unstandardized and standardized regression coefficients can be used to interpret effects in the usual way. But how do we obtain the size, confidence interval, and statistical significance of indirect effects?

The size of an indirect effect is calculated in exactly the same way as the size of indirect correlations (Section 9.2): Just multiply the size of direct effects. This can be done with either the standardized regression coefficients or the unstandardized regression coefficients.

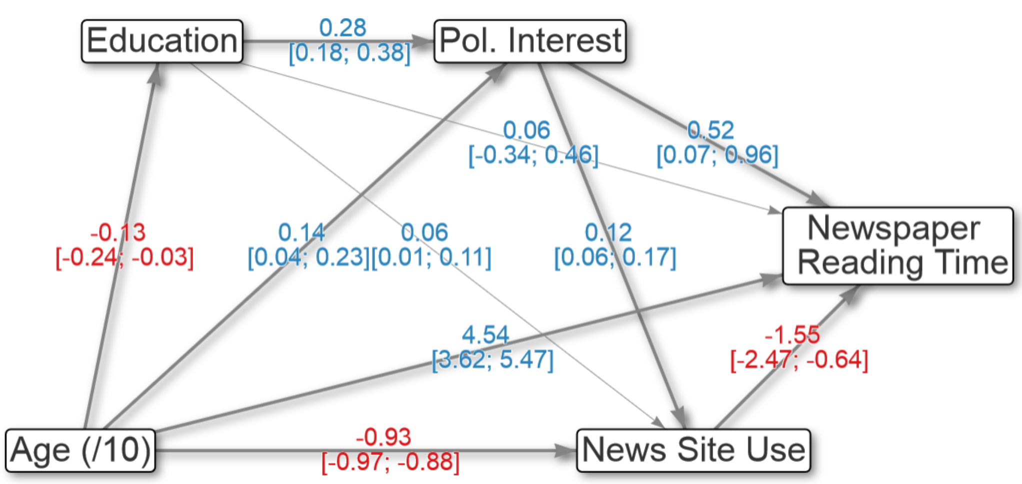


Figure 9.13: Path diagram with unstandardized effect sizes and their 95% confidence intervals.

It may sound weird that we can multiply the unstandardized regression coefficients but it really works. In Figure 9.13, for instance, the unstandardized partial effect of age (measured in tens of years) on interest in politics is 0.14. This means that an additional 10 years of life predict an average increase in interest in politics of 0.14.

In its turn, interest in politics has an unstandardized effect of 0.52 on reading time (in minutes). An additional unit of interest in politics predicts an average increase in reading time of 0.52 minutes.

Ten additional years of life only predict an increase of 0.14 in political interest, not a full unit increase. The predicted increase of 0.14 in political interest predicts 0.14 * 0.52 = 0.07 minutes of additional newspaper reading time. As a result, the additional ten years of life predict 0.14 * 0.52 = 0.07 minutes of additional newspaper reading time.

Note that the indirect effect is interpreted in terms of the measurement units of the initial predictor (age in tens of years) and the final outcome (reading time in minutes): A difference in (tens of) years predicts a difference in reading time in minutes. As a consequence, we can directly compare unstandardized indirect effect sizes of different paths between the same predictor and outcome, as we will see in Section 9.6.4.

9.6.3 Direction of indirect effects

Multiplication of direct effects assigns the right direction (positive or negative) to indirect effects. In the example above, age has a positive effect on interest in politics, which has a positive effect on newspaper reading time. If age goes up, interest in politics goes up and if interest in politics goes up, reading time increases. Thus, higher age is indirectly associated with more reading time through interest in politics: Plus times plus yields a plus.

If people with more interest in politics use news sites more frequently, there is a positive regression effect. If more news site use is associated with less newspaper reading (a negative effect), the indirect effect of interest in politics on reading time via news site use is negative. People with more interest in politics spend less time on reading newspapers because they use news sites more: Positive times negative yields a negative.

9.6.4 Parallel and serial mediation

If the indirect effects of an antecedent variable on an outcome variable contain at most one mediator, we have *single mediation* or *parallel mediation*. Figure 9.14 illustrates single and parallel mediation.

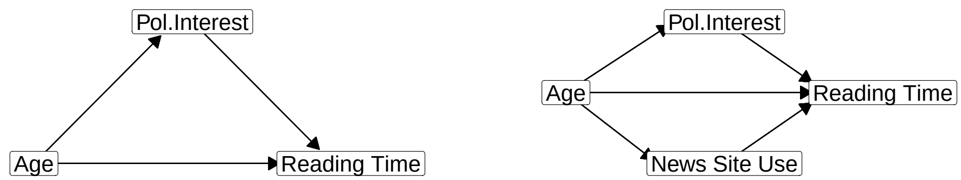


Figure 9.14: Causal diagrams for single (left) and parallel mediation (right).

If any of the indirect effects between an antecedent and outcome variable contains two or more mediators, we are dealing with serial mediation. Figure 9.15 illustrates serial mediation. It contains an indirect effect from age on reading time with two mediators: Age > Political Interest > News Site Use > Reading Time. The distinction between parallel and serial mediation is relevant to the software (PROCESS) that we will use to estimate indirect and total effects (Section 9.9).

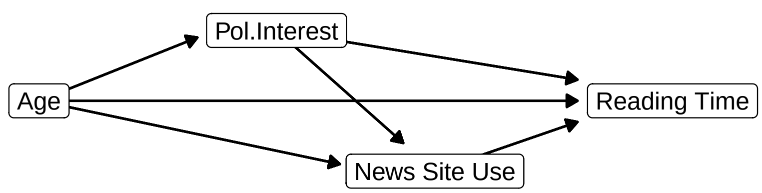


Figure 9.15: Causal diagram for serial mediation.

9.6.5 Partial and full mediation

Table 9.1 lists all direct and indirect effects of age on newspaper reading time in Figure 9.13. We can trace eight different paths from age to newspaper reading time. For each path, we multiply the unstandardized effect sizes.

The unstandardized indirect effects between the same predictor and outcome can be compared directly because they are all expressed in the same measurement units, namely the predicted change in the outcome (reading time in minutes) for a difference of one unit in the predictor (ten additional years of life) (Section 9.6.2). The unstandardized direct effect is also expressed as predicted change in reading time (minutes) for a difference in age (tens of years), so it

Table 9.1: All unstandardized direct and indirect effects of age on newspaper reading time.

	Effect		Effect		Effect		Effect		Total size
1	Age - Reading Time 4.54							=	4.54
2	Age - Education -0.13	*	Education - Reading Time 0.06					=	-0.007
3	Age - Education -0.13	*	Education - Pol. Interest 0.28	*	Pol. Interest - Reading Time 0.52			=	-0.02
4	Age - Education -0.13	*	Education - Pol. Interest 0.28	*	Pol. Interest - News Sites 0.12	*	News Sites - Reading Time -1.55	=	0.01
5	Age - Education -0.13	*	Education - News Sites 0.06	*	News Sites - Reading Time -1.55			=	0.01
6	Age - Pol. Interest 0.14	*	Pol. Interest - Reading Time 0.52					=	0.07
7	Age - Pol. Interest 0.14	*	Pol. Interest - News SItes 0.12	*	News Sites - Reading Time -1.55			=	-0.02
8	Age - News Sites -0.93	*	News Sites - Reading Time -1.55					=	1.44
	Total Effect							=	6.03

can be compared to the indirect effect. In addition, we can sum the unstandardized direct and indirect effects to obtain the total unstandardized effect.

With this in mind, we see that the relation between age and newspaper reading time is dominated by the positive direct effect ($b = 4.54$) and the positive indirect effect via news site use ($b = 1.44$). The remaining indirect effects are relatively small as indirect effects usually are.

Summing all effects, we obtain a *total effect* of age on newspaper reading time around 6 ($b = 6.03$). A person who is ten years older but in other respects the same as another person, is predicted to spend on average 6 additional minutes on reading newspapers per day.

If the direct effect of a predictor on the outcome is zero in a model with mediators, the predictor's effect is *fully mediated*. This clearly is not the case in our example: There still is a substantial direct effect of age on newspaper reading time. This is what we usually encounter; it is called *partial mediation*.

Sometimes, researchers decide that an effect is fully mediated if the direct effect is no longer statistically significant once a mediator is added to the model. This strategy is contestable because a statistically non-significant direct effect does not mean that the effect is absent (zero) in the population. It can be absent but it is much more likely to be present but just too small to be picked up by our significance test (see Chapter 6).

The distinction between full and partial mediation is a little bit problematic. From a substantive point of view, we may argue that direct effects are probably always mediated. As we have seen in Section 9.5.2, a direct effect usually summarizes a causal process that consists of intermediary steps, which is mediation. We may wonder whether it makes theoretical sense to talk about unmediated effects. Do we really believe that age can directly affect newspaper reading time?

If the variables that we entered in the model as mediators do not create any indirect effects, the direct effect is equal to the total effect. We may conclude that the direct effect is not mediated by the mediators that are included in the model. For example, if education, political interest, and news site use do not create indirect effects from age on newspaper reading time (Figure 9.13), we only have the direct effect of age on reading time in our model. However, this effect is very likely to be mediated by other variables that we did not include in the model. We should not conclude that the effect is unmediated because we have not found mediation yet.

9.6.6 Significance of indirect effects

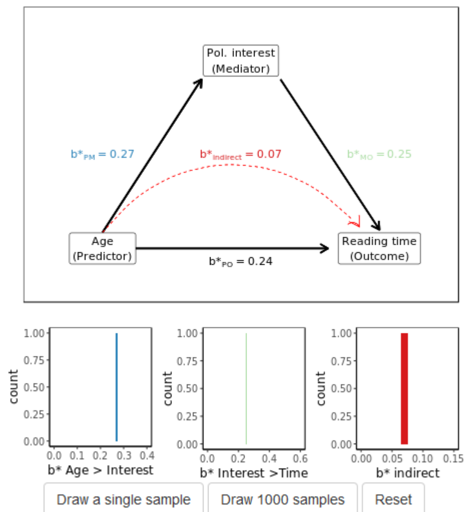


Figure 9.16: Does the sampling distribution of an indirect effect resemble the sampling distributions of its direct effects?

1. How is the sampling distribution of indirect effect size created? Before, simultaneously, or after the sampling distributions of direct effect sizes are created? Check your answer using the **Draw a single sample** button in Figure 9.16 repeatedly.

```
* For every sample, the indirect effect size is calculated as the product of
the effects of the predictor (Age) on the mediator (Political interest) and of
the mediator on the outcome (Reading time). The indirect effect is added to
the sampling distribution.
```

2. Take a few thousand samples. Do all three sampling distributions have the same shape?

```
* If you look carefully, you will see that the sampling distribution of the
regression coefficient of the indirect effect is skewed, asymmetrical. The
left tail is badly populated and the maximum is to the left of the maximum of
the normal curve.
```

SPSS does not calculate the size of indirect effects for us or their confidence intervals and p values. It is easy to calculate the sizes of indirect effects, as we have seen in a preceding section: just take the product of direct effects. In contrast, it is not possible to calculate the confidence interval or p value of an indirect effect in a reliable way from the confidence intervals or p values of the direct effects (see Hayes, 2013: Section 4.4 for a detailed discussion of different approaches).

For statistical inference on the indirect effect, we need the sampling distribution of the size of the indirect effect. This sampling distribution is not the same as the product or

Table 9.2: Bootstrap results for unstandardized indirect effects in a model with two mediators. Effect size, standard error, lower and upper levels of the 95% confidence interval.

	Effect	Boot SE	BootLLCI	BootULCI
Total indirect effect	1.47	.42	.62	2.25
Age - Pol. Interest - Reading Time	.05	.03	.01	.14
Age - Pol. Interest - News Site Use - Reading Time	-.02	.01	-.05	.00
Age - News Site Use - Reading Time	1.44	.42	.60	2.23

some other combination of the sampling distributions of the direct effects that make up the indirect effect. The situation is similar to the sampling distribution of the difference of two sample means (independent-samples t test), which is not equal to the difference between the sampling distribution of one mean and the sampling distribution of the other mean (Section 2.5.4).

We use bootstrapping (Section ??(boot-approx)) to create the sampling distribution of indirect effect size. We have learned the principles and limitations of bootstrapping in Section 2.1, so we need not go into details here. Suffice it to repeat that our original sample must not be too small and it must be quite representative of the population if we apply bootstrapping.

The confidence interval of an indirect effect can be calculated from its bootstrapped sampling distribution. Table 9.2 shows bootstrap results for the indirect effects in a model with age as predictor, newspaper reading time as outcome, and interest in politics and news site use as mediators.

In total, there is a substantive indirect effect of age on newspaper reading time in this model. We are confident that this effect is positive (b = 1.47, 95%CI[0.62, 2.25]). It is easy to see that the indirect effect of age via news site use on reading time is by far the most important indirect effect. On its own, it is responsible for almost the total indirect effect (b = 1.44, 95%CI[0.60, 2.23]).

It may happen that an indirect effect is not statistically significant (the confidence interval includes zero) whereas all direct effects that constitute the indirect effect are statistically significant. In Figure 9.17, for example, both the effect of age on political interest (b = 0.30, 95%CI[0.22, 0.38]) and the effect of political interest on reading time (b = 0.20, 95%CI[0.12, 0.28]) are statistically significant at the .05 level. The indirect effect of age via political interest on reading time, however, is not statistically significant at this level (b = 0.30 * 0.20 = 0.06, 95%CI[-0.02, 0.14]).

This sounds like a paradox but it should not upset you. The unstandardized indirect effect tends to be weaker than the direct effects, that is, closer to zero (for instance, see Table 9.1). With a weaker effect, it is more difficult to reject the null hypothesis that the effect is zero in the population. We need a larger sample to reject null hypotheses for smaller effects (see Chapter 5 on power).

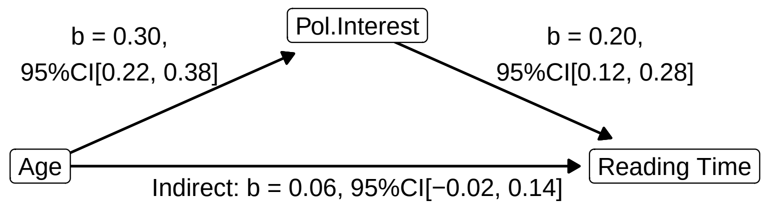


Figure 9.17: Causal diagram for the effects of age and interest in politics on newspaper reading time: unstandardized estimates with 95% confidence intervals.

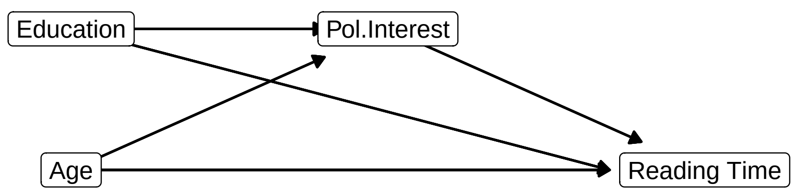


Figure 9.18: Causal diagram for interest in politics as mediator between age and newspaper reading time with education as covariate.

9.7 Controlling for Covariates

We usually have theoretical reasons to expect mediation between one pair of variables, for example, political interest as mediator between age and newspaper reading time. At the same time, we know that our outcome variable and perhaps our mediator may depend on other variables. Newspaper reading time, for instance, may also depend on education. In this situation, we would use the other variables as covariates for which we want to control statistically.

Figure 9.18 presents a model in which education is used as a covariate in a model with political interest mediating the effect of age on newspaper reading time. Education is probably causally antecedent to both political interest and newspaper reading time, so it is allowed to have an effect on both variables. In this way, we control for education and remove spurious correlation between political interest and newspaper reading time due to education as a common cause.

If education is allowed to predict both political interest and newspaper reading time as in Figure 9.18, political interest mediates the effect of education on newspaper reading time.

We are, however, not interested in mediation in the case of a covariate, so we do not estimate or report the indirect effects of education on newspaper reading time. In other words, a *covariate* is a predictor for which we do not investigate if its effect is mediated.

Note that covariates should only be allowed to have an effect on variables that can be caused by the covariate. We should not include effects of a covariate on a variable that is causally antecedent to the covariate. If a control is a consequence rather than a cause of a mediator, it had better be used as another mediator in the model. If, for instance, political cynicism may affect newspaper reading time but it is a consequence of political interest, it should be included as a second (serial) mediator instead of a covariate.

9.8 Reporting Mediation Results

We analyse a path model as a series of regression models, so the general rules for reporting mediation are the same as for reporting regression analyses (see Section 8.6). If you summarize results in a table, make sure that the table includes:

1. The unstandardized regression coefficients for all direct and indirect effects tested in the regression models.
2. The confidence intervals and significance levels of the unstandardized effects.
3. The F test and measure of model fit (R^2) for each regression model.

Table 9.3: Unstandardized effects in a model regressing newspaper reading time on age with one mediator (News Site Use) and two covariates (Education, Political Interest). OLS estimates for direct effects, bootstrap results for indirect effects, using 5,000 bootstraps and a bias-corrected method.

	B		95% CI
Outcome: News Site Use			
constant	6.62	***	[5.92; 7.31]
age	-0.93	***	[-0.97; -0.88]
education	0.06	*	[0.01; 0.11]
pol.interest	0.12	***	[0.06; 0.17]
R^2^	0.86		
F (3, 308)	617.40	***	
Outcome: Newspaper Reading Time			
constant	13.59	**	[5.26; 21.93]
age	4.55	***	[3.62; 5.47]
education	0.06		[-0.34; 0.46]
pol.interest	0.52	*	[0.07; 0.96]
newssite	-1.55	**	[-2.47; -0.64]
R^2^	0.79		
F (4, 307)	290.85	***	
Indirect Effect			
Age > News Site Use > Reading Time	1.44		[0.61; 2.17]

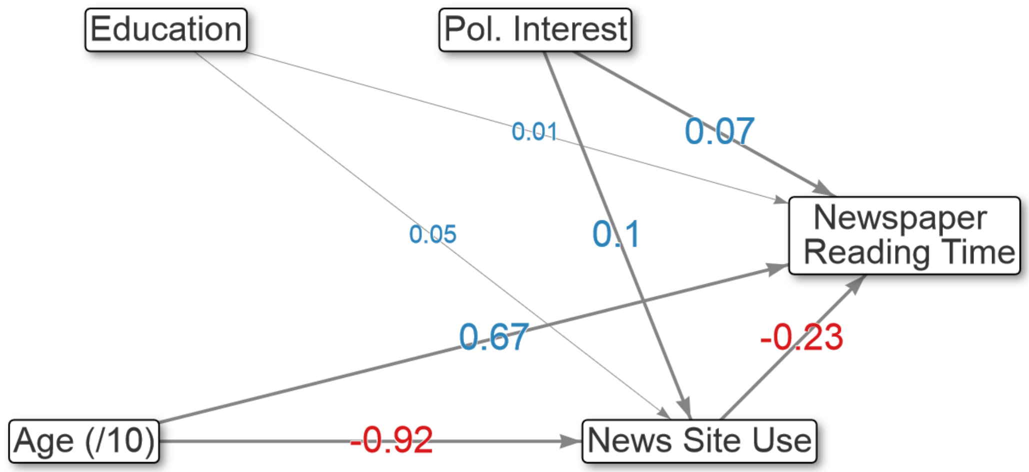


Figure 9.19: Unstandardized direct effects for a path model with one mediator and two covariates.

Note. * p < .050, ** p < .010, *** p < .001.

A path model may yield a lot of direct effects, so it is good practice to present results as a path diagram with the values of the standardized or unstandardized regression coefficients as labels to the arrows. A path model conveniently summarizes the results for the reader (Figure 9.19). Remember that we don't use standardized regression coefficients if the predictor or a covariate is dichotomous or a set of dummy variables (see Section 8.1.3).

If effect mediation is central to your report, focus your presentation and interpretation on the indirect effects and compare them to the direct effects. Report the size and confidence interval of each indirect effect. If possible, add both the direct and indirect effect to a diagram such as Figure 9.19.

Interpret the indirect effect just like any regression effect, namely, as the predicted difference in the outcome for a one unit difference in the predictor. It is usually interesting to compare the sizes of the direct and indirect effects. Is the effect predominantly mediated in the model or is only a minor part of the effect mediated in the model?

Inform the reader that you bootstrapped the indirect effect and report the number of bootstrap samples and the method used for the confidence intervals (see Section 9.9). For a more elaborate discussion of reporting mediation, see (Hayes, 2013: 198-202).

9.9 Mediation with SPSS and PROCESS

9.9.1 Instructions

SPSS cannot apply statistical inference to indirect effects, so we use the PROCESS macro developed for this purpose (Hayes, 2013). If correctly installed (see below), the macro can be used from within the SPSS Regression menu. Please note that you had better not paste the PROCESS commands to the SPSS syntax because it produces a lot of code that is difficult to understand. Instead, run the PROCESS command directly from the menu and manually add a comment to your SPSS syntax file reminding yourself of the model that you estimated with PROCESS.

Download the PROCESS macro and install the SPSS custom dialog file. Check the FAQ at the PROCESS website if installation is not successful. If PROCESS is successfully installed, it can be found in the Analyse > Regression menu.

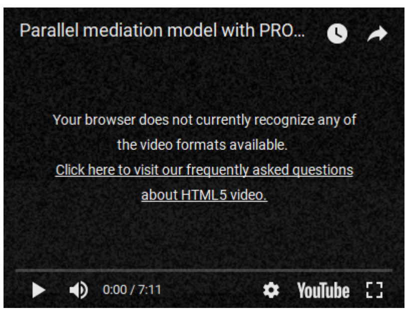


Figure 9.20: Estimating a single or parallel mediation model with PROCESS (Model 4).

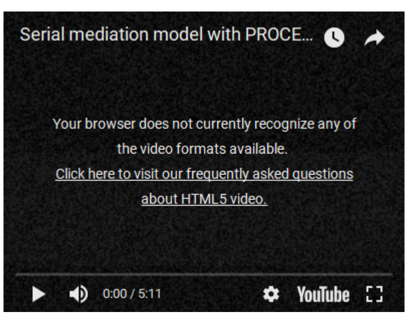


Figure 9.21: Estimating a serial mediation model with PROCESS (Model 6).

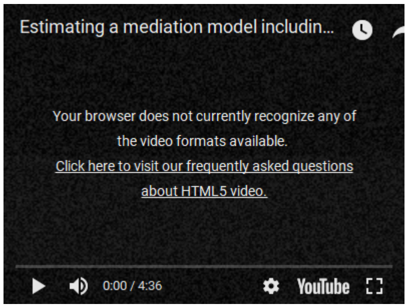


Figure 9.22: Estimating a mediation model including covariates with PROCESS.

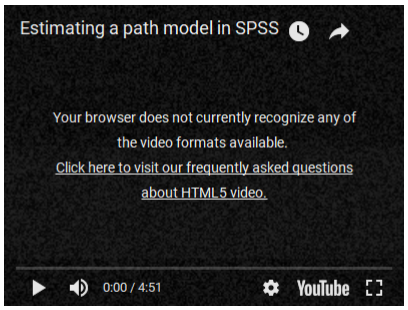


Figure 9.23: Estimating a path model in SPSS.

9.9.2 Exercises

1. Use readers.sav to analyze the causal model depicted in Figure 9.19 with a series of regression models in SPSS. Create a table and draw a path diagram to present the direct effects.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=age education polinterest newssite readingtime
  /ORDER=ANALYSIS.
* Multiple regression for newspaper reading time.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT readingtime
  /METHOD=ENTER age education polinterest newssite
```

```
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID).
* Multiple regression for news site use.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT newssite
  /METHOD=ENTER age education polinterest
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID).

Check data:

The variables do not have impossible values.

Check assumptions:

Regression model with newspaper reading time as outcome variable:

* The residuals are quite normally distributed.
* They are centered around zero at all levels of the predicted outcome, so a
linear model seems to fit the data.
* The variation in residuals is about the same at all levels of the predicted
outcome, so the outcome is more or less equally well predicted at all levels
of the outcome variable.
* As a conclusion, there are no clear indications that the assumptions for a
linear regression model are violated.

Regression model with news site use as outcome variable:

* The residuals are quite normally distributed but they display a pattern of
down-sloping strips. This is the consequence of the fact that the outcome
(news site use) is measured as integer scores, that is, without decimal
places. So this pattern is nothing to worry about.
* Check that the residuals are evenly distributed above and below zero for all
predicted outcome levels and that the variation is more or less equal for all
predicted outcome levels. Because the residuals are organized in strips, the
variation cannot be as large at the extreme right or extreme left as in the
middle. But we may assume that this results from the discrete (integer)
outcome variable instead of from a problem with homoscedasticity.

Results:
```

```
Your table and path diagram should like like the table and path diagrams
reported in the section on reporting mediation results.
```

2. To what extent is the effect of age on newspaper reading time mediated by news site use? Use the data of Exercise 1 and PROCESS to estimate both the unstandardized and standardized indirect effect in a model containing only these three variables. Interpret the results.

```
SPSS syntax:

We should not paste PROCESS commands because they contain a lot of
unintelligble code.

In the PROCESS menu:

* Select readingtime as Outcome Varible (Y).
* Select age as Independent Variable (X),
* Select newssite as M variable(s),
* Select 4 under Model Number (model with only one mediator).
* Under Options: Select Effect size and 3 decimal places in output.
* Under Long names: Select Allow long variable names.

Check data: See Exercise 1.

Check assumptions: Cannot be done with PROCESS, must be done with SPSS. See
Exercise 1.

Interpret the results:

The relevant results in the PROCESS output:

Indirect effect of X on Y

          |  Effect   |  Boot SE  |  BootLLCI  |  BootULCI
newssite  |  1.121    |  .411     |  .307      |  1.923

Completely standardized indirect effect of X on Y

          |  Effect  |  Boot SE  |  BootLLCI  |  BootULCI
newssite  |  .164    |  .060     |  .045      |  .281

Note that these are bootstrap results, so they may change from
one analysis to the next. Your results are most likely to be
not exactly the same as the ones reported here.

* An age difference of ten years (remember that age was measured in decades)
```

```
increases the predicted average newspaper reading time through news site use
with slightly more than one minute. With 95% confidence, the increase is 0.3
and 1.9 additional minutes.
* The indirect effect is weak to moderate, b* = .16, 95%CI[.05; .28].
```

3. Add interest in politics as a covariate to the model of Exercise 2. Are you going to use it as a covariate for the mediator (news site use), the outcome (newspaper reading time), or both? Motivate your choice. Present the unstandardized effects as a path diagram and add the indirect effect to it.

```
SPSS syntax:

* Again, we should not paste PROCESS commands because they contain a lot of
unintelligble code.

In addition to the PROCESS menu selections specified in Exercise 2, you should:

* Select interest in politics under Covariate(s),
* Change the option under Covariate(s) in model(s) of... if you want political
interest to affect either news site use or newspaper reading time (and not
both).

Check data: See Exercise 1.

Check assumptions: Cannot be done with PROCESS, must be done with SPSS. See
Exercise 1.

Interpret the results:

* To decide on the effects of the covariate that we include in the model, we
need a substantive argument about the causal order of political interest, news
site use, and newspaper reading time. The covariate political interest should
have an effect on every variable to which it is antecedent unless we have good
reasons to believe that the effect is zero. We hardly ever have such good
reasons.
* In this answer, political interest is hypothesized to affect both news site
use (the mediator M) and newspaper reading time (the outcome Y), so the option
under Covariate(s) in model(s) of... remains at ...both M and Y.
* If you decide that political interest is consequent to news site use instead
of antecedent, political interest should be included as a (second) mediator
between news site use and reading time. In this case, you should estimate a
model for serial mediation, see Exercise 4.

The relevant results in the PROCESS output:

Indirect effect of X on Y
```

```
          |  Effect  |  Boot SE  |  BootLLCI  |  BootULCI
newssite  |  1.441   |  .416     |  .589      |  2.225

Completely standardized indirect effect of X on Y

          |  Effect  |  Boot SE  |  BootLLCI  |  BootULCI
newssite  |  .212    |  .061     |  .087      |  .325

* The interpretation is very much like the interpretation in Exercise 2. Note
that the indirect effect has become a little stronger now that we control for
interest in politics.
```

4. Add political interest to the model of Exercise 2 such that it mediates the effect of age on news site use: Age > Political Interest > News Site Use > Reading Time. Estimate the model with PROCESS. Report the unstandardized indirect effects as a table and interpret them.

```
SPSS syntax:

* Remember, we should not paste PROCESS commands because they contain a lot of
unintelligble code.

In addition to the PROCESS menu selections specified in Exercise 2, you should:

* Add political interest to the list of M Variable(s) ensuring that it is the first
this list.
* Change Model Number to 6 (Model with two mediators in line).

Check data: See Exercise 1.

Check assumptions: Cannot be done with PROCESS, must be done with SPSS. See
Exercise 1.

Interpret the results:

The main results are reported thus:

Indirect effect(s) of X on Y

        |  Effect  |  Boot SE  |  BootLLCI  |  BootULCI
Total:  |  1.474   |  .418     |  .657      |  2.298
Ind1 :  |  .053    |  .033     |  .007      |  .146
Ind2 :  |  -.021   |  .013     |  -.057     |  -.003
Ind3 :  |  1.441   |  .417     |  .616      |  2.257

Indirect effect key
```

```
* Ind1 :   age -> polinter -> readingt
* Ind2 :   age -> polinter -> newssite -> readingt
* Ind3 :   age -> newssite -> readingt

A table could look like this:

Indirect effect(s) of X on Y

                                                |  Effect | Boot SE | BootLLCI | BootULCI
age->pol. interest->reading time                |  .05    |  .03    |  .01     |  .15
age->pol. interest->news site use->reading time |  -.021  |  .01    |  -.06    |  -.003
age->news site use->reading time                |  1.44   |  .42    |  .62     |  2.26
Total indirect effect                           |  1.47   |  .42    |  .66     |  2.30


* Age has an indirect effect on newspaper reading time. An additional ten
years of life predict that newspaper reading time increases by 1.5 minutes on
average.
* The indirect effect mainly consists of an effect via news site use. Indirect
effects including interest in politics are much weaker.
* The indirect effect through interest in politics is probably positive
whereas the effect through political interest and news site use is probably
negative.
```

5. The data set children.sav contains information about the media literacy of children and parental supervision of their media use. Is the effect of age on media literacy fully or partially mediated by parental supervision? Use PROCESS and SPSS to estimate the model and check the assumptions. Motivate your answer to the question.

```
SPSS syntax:

* Check data.
FREQUENCIES VARIABLES=age supervision medliter
  /ORDER=ANALYSIS.
* Set imposible value (25) to missing.
* Define Variable Properties.
*supervision.
MISSING VALUES supervision(25.00).
EXECUTE.
* Indirect effect test with PROCESS: Model 4.
* Do not paste PROCESS output.
* Regression models for checking assumptions.
* Outcome: media literacy.
REGRESSION
  /MISSING LISTWISE
  /CRITERIA=PIN(.05) POUT(.10)
```

```
  /NOORIGIN
  /DEPENDENT medliter
  /METHOD=ENTER age supervision
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID).
* Outcome: parental supervision.
REGRESSION
  /MISSING LISTWISE
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT supervision
  /METHOD=ENTER age
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID).

We should not paste PROCESS commands because they contain a lot of
unintelligble code.

In the PROCESS menu:

* Select medliter as Outcome Varible (Y).
* Select age as Independent Variable (X).
* Select supervision as M variable(s).
* Select 4 under Model Number (model with only one mediator).
* Under Options: Select Effect size and 3 decimal places in output.
* Under Long names: Select Allow long variable names.

Check data:

* All values are plausible except score 25 on parental supervision because the
value labels tell us that the scale ranges from 1 to 10. Set 25 to missing.

Check assumptions:

* These checks cannot be done in PROCESS, so use linear regressions in SPSS to
produce plots of residuals.

Regression of media literacy:
* The residuals are quite normally distributed and nicely grouped around zero
at all levels of the predicted outcome.
* The variation of residuals is perhaps smaller at low levels of the
predicted outcome than at high levels but there are perhaps too few
observations at low predicted levels to decide.

Regression of parental supervision:
```

```
* The residuals are quite normally distributed, nicely grouped around zero at
all levels of the predicted outcome and with more or less the samen variation
at all levels of the predicted outcome. The assumptions need not be violated.

Interpret the results:

Relevant results from PROCESS:

Direct effect of X on Y

         |  Effect  |  SE    |  t      |  p     |  LLCI  |  ULCI
supervis |  .193    |  .083  |  2.327  |  .022  |  .028  |  .358


Indirect effect of X on Y

         |  Effect  |  Boot SE  |  BootLLCI  |  BootULCI
supervis |  .031    |  .031     |  -.017     |  .110

* The indirect effect of age on media literacy via parental supervision does
not remove the direct effect completely (b = 0.19, t = 2.33, p = .022,
95%CI[0.03; 0.36]), so the effect is only partially mediated if it is mediated
at all.
* Although the largest part of the 95% confidence interval is positive, we
cannot rule out a zero (absent) indirect effect, b = 0.03, 95%CI[-0.02; 0.11].
Perhaps, the effect of age on media literacy is not mediated by parental
supervision.
```

9.10 Criticisms of Mediation

If we think of causality, we usually think of a process in which one thing leads to another thing, which leads to something else, and so on. This is apparent if we want to explain why we think that one phenomenon causes another (see Section 9.5.2). Mediation, however, is difficult to establish with regression analysis and, as some argue, perhaps impossible to establish.

9.10.1 Causal order assumed

It is paramount to note that the regression approach to mediation and path models does not tell us anything about the causal order of the variables. The causal order is purely an assumption that we make. The plausibility of the assumptions depends on how well we can justify the time order of the variables and the absence of common causes for cause-consequence pairs (see 9.5.1).

9.10.2 Time order

To establish the time order of variables, we must think about the time at which the behaviours or opinions that we measure took place. This is what matters, not the time at which we make the measurement. After all, we can collect information on behaviour a long time after the fact, for example by asking respondents when they started using news sites or checking internet use logs.

The more time has passed between the occurrence of the behaviour or opinion that we think is the cause and the occurrence of the one we think is the consequence, the more plausible the causal order. If cause and consequence appear very close in time, it may be difficult to argue that one precedes the other.

9.10.3 Causality or underlying construct?

For causes and consequences that appear nearly simultaneously, we should take into account that the two measurements may measure the same underlying construct. Think of the way we construct a scale from items: We assume that the items measure the same underlying attitude, for instance, interest in politics.

The indicators of a scale are correlated because they have a common cause, namely, the underlying attitude. But it does not make sense to interpret the correlation as a sign of mediation. One item does not trigger another item, and so on. A mediator must be theoretically and conceptually different from both the predictor and outcome. We have to provide arguments that they are really different.

An underlying construct such as an attitude is just one example of a common cause that undermines the conclusion that an effect is mediated. Any common cause of a predictor and mediator or of a mediator and outcome that is not included in the model renders all or part of the correlations spurious, that is, non-causal. If we doubt the causality of any effect in the path constituting the indirect effect, we must doubt the causality of the indirect effect as well. Without causality, there is no mediation. So we should think hard about common causes and include them in our model.

9.10.4 Statistical control is not experimental control

In a strict interpretation, mediation cannot be established with regression models. To understand this, let us look very carefully at the statistical meaning of direct and indirect effects.

A direct effect of age on newspaper reading time refers to changes in reading time that depend on changes in age but not on changes in the mediator, for instance, interest in politics. Strictly speaking, age should change and interest in politics should not change for a person to detect the direct effect of age on reading time. But interest in politics depends on age if it mediates the effect of age, so it tends to change if age changes. That is a problem.

In an experiment, we would have to manipulate the value of the mediator to ensure that it does not change. If we succeed in doing this, which is difficult to imagine, the effect of the predictor on the outcome is really a direct effect, at least within the model. For example, interest in politics is manipulated to be the same for some participants. Then, a change in newspaper reading time due to age change would indeed represent the direct effect for these participants.

Controlling the effect of a predictor on an outcome for the mediator in a regression model, however, is not the same as experimental control as described above. In a regression model, the mediator values vary among persons. The statistical trick is that we only compare people who have the same mediator score when we calculate the direct effect. But even for these people, the effect of the predictor on the mediator may have done it's work: Their actual interest in politics is affected by their age. As a result, the effect of age on newspaper reading time includes an effect of age on interest in politics, so the mediator (interest in politics) is not completely excluded from the direct effect.

9.10.5 Recommendations

All in all, mediation is an intuitively simple and appealing concept. Unfortunately, it is very difficult to substantiate the claim that indirect effects in path models represent mediation. Mediation assumes causal effects and causality is difficult to establish.

If you plan to investigate mediation:

1. Justify that the mediator is theoretically and conceptually different from the predictor and outcome.

2. Motivate the time order of variables in the model.

3. Include variables that can be common causes to predictor, mediator, or outcome in your research project and in the regression models that you are going to estimate.

9.11 Combining Mediation and Moderation

Mediation and moderation (Chapter 8) can occur in the same model. For example, the effect of age on newspaper reading time mediated by interest in politics can be different for females and males. In other words, the indirect effect is different for females and males.

If the indirect effect is different for females and males, at least one of the two direct effects (predictor on mediator or mediator on outcome) must be different for females and males. This direct effect is moderated and as a consequence, indirect effects including this effect are moderated. This is called *moderated mediation*. In the example, sex is the moderator and interest in politics is the mediator of the indirect effect of age on newspaper reading time.

Several models with more than one mediator or with moderated mediation can be estimated with PROCESS. For an overview of the models, see http://www.afhayes.com/public/

templates.pdf or Appendix A in Hayes (2013). The models, however, are quite complex, so we leave them for enthusiasts.

9.12 Test your intuition and understanding

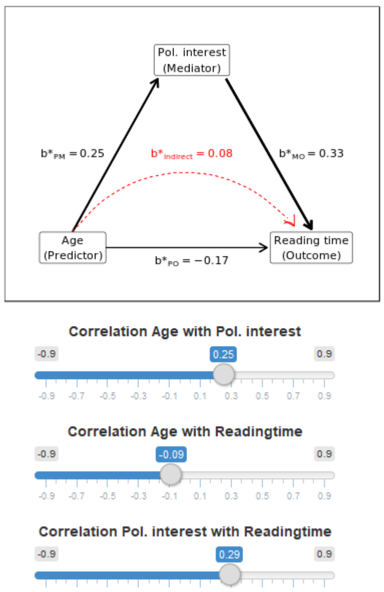


Figure 9.24: How does mediation work and how can we analyze mediation with regression models? The values to the arrows in the diagram are standardized regression coefficients.

1. In Figure 9.24, what does the curved arrow represent? How is its value calculated?
2. How do the numbers tell you that the regression effect of political interest on newspaper reading time is controlled for age? Hint: In a regression model with only one predictor, the standardized regression effect is equal to the correlation between the predictor and outcome variable.
3. If age is excluded from the model as presented in Figure 9.24, would it suppress or reinforce the effect of political interest on newspaper reading time? When would age suppress, reinforce, or not confound the effect of political interest on newspaper reading time? Use the sliders to check your answer.
4. When is the correlation between political interest and reading time completely spurious in this model? Use the sliders to check your answer.
5. Can the standardized effect of political interest on reading time be larger than 1.0 or smaller than -1.0? If so, in which situation? If not, why not? Use the sliders to check your answer.

6. How many path models do we need to estimate the model of Figure 9.24?

9.13 Take-Home Points

- In a multiple regression model, a regression coefficient represents the predictive effect of a variable while controlling for the effects of all other predictors. It is called a *partial effect*: the predictions of the outcome that cannot be predicted by the other predictors.
- If a new predictor is added to a regression model, the regression coefficient of an old predictor changes if the new predictor is correlated with both the old predictor and the outcome. If the old predictor's effect becomes stronger, the new predictor was a suppressor. If it becomes weaker or changes direction (sign), the new predictor was a reinforcer and the old effect was (partially) spurious.
- An estimated regression coefficient only shows the true effect of a predictor if all suppressors and reinforcers are included in the model.
- A causal or path model without causal feedback loops can be estimated as a series of regression models: one regression model for each variable that has at least one predictor in the path model.
- Unstandardized regression coefficients, standardized regression coefficients, and correlations can be multiplied to obtain indirect effects or indirect correlations.
- An indirect effect is a mediated effect. Variables that are at the same time predicted and predictors in an indirect effect are called *mediators*, *intermediary variables*, or *intervening variables*.
- Statistical inference on an indirect effect—its confidence interval and significance level—requires a sampling distribution of the indirect effect's size. This distribution can be bootstrapped with the PROCESS macro (Hayes, 2013).
- Mediation is an intuitively appealing concept but it is difficult to establish. A causal interpretation requires a clear time order between predictor, mediator, and outcome, a clear theoretical and conceptual difference between these three variables, and the inclusion of all common causes of predictor, mediator, and outcome in the regression models.

Read the little but very helpful book on the logic of causal order by James A. Davis (1985) for more information on causality and correlational analysis.

References

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. San Diego, CA: Academic Press.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Davis, J. A. (1985). *The logic of causal order* (Vols. 07-055, p. 72). Beverly Hills, CA: Sage.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann.Statist.*, *7*(1), 1–26. https://doi.org/10.1214/aos/1176344552"

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185. https://doi.org/10.1080/01621459.1987.10478410

Erdogan, B. Z. (1999). Celebrity endorsement: A literature review. *Journal of Marketing Management*, *15*(4), 291–314.

Fisher, R. A. (1919). The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, *52*(2), 399–433. https://doi.org/10.1017/S0080456800012163

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver; Boyd.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society.Series B (Methodological)*, *17*(1), 69–78. Retrieved from http://www.jstor.org/stable/2983785

Frick, R. W. (1998). Interpreting statistical testing: Process and propensity, not population and random sampling. *Behavior Research Methods, Instruments, & Computers*, *30*(3), 527–535. https://doi.org/10.3758/BF03200686

Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium auctore carolo friderico gauss*. sumtibus Frid. Perthes et IH Besser.

Hainmueller, J., Mummolo, J., & Xu, Y. (2016). How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice. https://doi.org/http://dx.doi.org/10.2139/ssrn.2739221

Halpin, P. F., & Stam, H. J. (2006). Inductive inference or inductive behavior: Fisher and neyman: Pearson approaches to statistical testing in psychological research (1940-1960).

The American Journal of Psychology, *119*(4), 625–653. https://doi.org/10.2307/20445367

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press.

Laplace, P. S. de. (1812). *Théorie analytique des probabilités* (Vol. 7). Courcier.

Lyon, A. (2014). Why are normal distributions normal? *The British Journal for the Philosophy of Science*, *65*(3), 621–649.

McCracken, G. (1989). Who is the celebrity endorser? Cultural foundations of the endorsement process. *Journal of Consumer Research*, *16*(3), 310–321.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London.Series A, Mathematical and Physical Sciences*, *236*(767), 333–380.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos Trans R Soc Lond A*, *231*(694-706), 289. Retrieved from http://rsta.royalsocietypublishing.org/content/231/694-706/289.abstract

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594.