

Acceptability Ratings in Linguistics: A Practical Guide to Grammaticality Judgments, Data Collection, and Statistical Analysis

Fabian Bross

Version 1.02

Overview. With this tutorial I give an overview of the techniques of how to conduct acceptability rating studies (or grammaticality judgments, later I will discuss terminology). It is written for people with no prior knowledge of statistics. It contains the following sections: (i) theoretical background on statistics and its relation to empirical research, (ii) designing a questionnaire and conducting your study (the shortest section), (iii) analyzing and visualizing your results (with OpenOffice, R, and JASP). There is also a second part of this tutorial on the use of mixed models to analyze your data (see: www.fabianbross.de/mixedmodels.pdf).

If you have any suggestions, feel free to write me: fabian.bross@ling.uni-stuttgart.de

Please cite as:

Fabian Bross (2019). Acceptability Ratings in Linguistics: A Practical Guide to Grammaticality Judgments, Data Collection, and Statistical Analysis. Version 1.02. Mimeo. Online: www.fabianbross.de/acceptabilityratings.pdf.

Contents

I	Theoretical background on statistics and its relation to empirical research	3
1	The empirical method—a short introduction	3
2	Acceptability or grammaticality judgments? A note on terminology	5
3	Test theory: Creating a construct	8
4	The foundations of acceptability judgments: Measurement theory	15
5	Measures of central tendency	18
6	Measures of dispersion: The standard deviation	20
7	More about populations and samples	25
8	Confidence intervals	26
9	Behind the scenes: parameters (and more about distributions)	29
II	Designing a questionnaire and conducting your study	32
10	How to create the questionnaire: the stimuli	32
11	Use Latin squares for counterbalancing	34
12	The instructions	43
13	Procedure	45
14	Software tips	46
III	Analyzing and visualizing your results	46
15	Analyzing the results	46
16	Do your statistics in different environments: OpenOffice, R/RStudio, JASP	60
17	Multiple comparisons	80
18	Reporting your results	82
19	More on visualizing the results: box plots and beyond	83

Part I

Theoretical background on statistics and its relation to empirical research

There has been much criticism in the way linguists gather their data (e. g., Bard, Robertson & Sorace 1996; Featherston 2007; Gibson & Fedorenko 2010a, 2010b). Given that you are reading this tutorial I assume you know that, at least in some cases, it is not very scientific to rely solely on your own intuitions. It is preferable to collect data in a structured way by carrying out acceptability judgments (see below for terminology and the distinction between acceptability and grammaticality). There are several different ways of collecting such judgments. The most commonly used are magnitude estimation, yes-no choice tasks, forced-choice tasks, and Likert item tasks (if you don't know any of these terms, don't worry). For reasons I will not discuss in detail, I will only concentrate on Likert item tasks in this tutorial, meaning that participants have to rate sentences from unnatural to natural. This is done simply because Likert items are easy to understand for participants, the math that you need for your statistics is fairly easy, and the results are pretty informative (for empirical reasons as to why to use Likert scales and not to rely on other, more complicated, methods in linguistic judgment tasks see Weskott & Fanselow 2011 and Sprouse 2011a).

1. The empirical method—a short introduction

The word empirical is derived from Greek 'empeiría' meaning 'experience'. This means that empirical research collects (and analyzes) data. In this rather vague sense, nearly all academic disciplines are empirical.¹ And indeed, linguistics is always an empirical science since it deals with language data.

There is, however, a difference between linguists who rely on their own intuitions and linguists who collect intuitions in a structured way that follows a strict method to test a previously defined hypothesis. The latter kind of linguistics is what this tutorial is about. The branch of linguistics that does not rely on data collected from participants, but relies on intuitions is sometimes derogatorily called 'armchair linguistics'. Armchair linguistics is not in itself bad as many judgments are rather clear. However, as soon as there is disagreement on data that is hard to judge, linguists should rely on empirical

¹There are, of course, disciplines that are not empirical at all. Prime examples are logic and mathematics. That mathematics is not empirical in the sense we use here is illustrated by Egmont Colerus (2013:25) [1942] in a nice way: "[M]athematics is not in itself an experimental subject like science, which is based on experience. It is purely intellectual, spun out of the brain without any need for experiment. Its results can neither be proved or disproved by experience; proof can only be obtained by ensuring the accuracy of its logical operations."

data collected from many native speakers. Another reason to collect judgments would be a case in which you simply have no access to intuitions because you are not a native speaker of the language under investigation.

Strict methods like the one I will present here follow a philosophy, namely positivism, that I cannot outline here. The short version is: positivism is a version of empiricism:

Empiricism is the philosophical tradition which believes that (a) the world consists of objects (b) these objects have their own characteristics and properties which exist irrespective of what we think they are like, and (c) our knowledge of objects is developed through experience with them. (Buckingham & Saunders 2008:12)

I think what the previous quote says about empiricism is quite uncontroversial: Scientists believe that there are objects in the world that don't change just because we don't look at them. They exist independently of us. We can only gather knowledge about them by collecting and analyzing data. And we can only collect data if we can observe and experience, i. e., measure something. In our case, the object we want to know something about is grammar (think about whether you believe that the grammar of your language exists as an object that has "characteristics and properties which exist irrespective of what" you think they are like).

Positivists are a little bit stricter than empiricists. They especially emphasize the point that you can only know something from observing something (positive data = data you gain through experience). Positivism has its roots in the 19th century (mainly with Auguste Comte) and was criticized by Austrian/British philosopher Karl Popper (1959): Popper introduced the idea that a scientist in general cannot prove something via observation, but can only disprove something. So for him the empirical scientific method starts out with a theoretical generalization and tries to prove it wrong by testing it.² The very short version of this huge debate is that we start out with a hypothesis about how the world works and try to prove this hypothesis wrong via empirical investigation.

Empirical research is often divided into experimental and non-experimental research. Doing acceptability judgments has many features of experimental research, but typically, we would not call a acceptability rating study an experiment, but rather a quasi-experiment (or simply an empirical study). A real experiment is a repeatable empirical study under controlled conditions. In an experiment you test an hypothesis via manipulating one or more variables (for example, one group gets a drug, the other group gets a placebo). The crucial point in an experiment, however, is that there are two or more groups and that the individuals are assigned randomly to these groups. Random assignment plays a crucial role in the definition of experiments (see already Fisher 1925,

²To be 100 percent correct, Popper's view is actually called 'critical rationalism' rather than 'positivism'

1935) and constitutes one case of randomization—a key concept in empirical research: “The term random refers to the equiprobability of events. Random assignment refers to any procedure that assigns subjects to the comparison group on the basis of chance” (Christensen 2012:473). As acceptability judgment tasks in their basic form lack random assignment we do not speak of an experiment, but rather of a quasi-experiment or simply call it an empirical study. However, we will see that we can use Latin squares to create several lists of stimuli which allows us to assign participants to these lists (see Section 11). This will make our judgment tasks more similar to an experiment.

Excursus: Random Assignment, Random Sampling, and Populations

Note that randomly choosing participants from a population is not random assignment. This is called ‘random sampling’. Also note that random assignment always involves the random assignment to groups in order to manipulate a variable. This means that if you compare the speakers of two different dialects, you still don’t have an experiment, since speaking a dialect is not a variable you can manipulate (similar to, for example, gender or age).

Empirical research also involves random sampling. Take an election poll for example. As I’m from Germany, I will illustrate this with the German election system. There are approximately 65 million people eligible to vote in Germany. That’s a lot of people to make a poll! It is, however, enough, to ask 2.000 people, to get an incredibly reliable election prediction (provided that the participants are honest). That 2.000 people is enough for a prediction lies in the fact that they are chosen randomly—in fact, in many cases, we would need much less participants. This leads to a situation in which all voter groups are represented. We call the set of all individuals that are of interest our ‘(statistical) population’ and the subset that we look at our ‘(random) sample’. In this case, the population consist of a finite set of real individuals. A population can also consist of judgments and can be infinite. I will say more about populations and samples later.

2. Acceptability or grammaticality judgments? A note on terminology

“Speakers’ reactions to sentences have traditionally been referred to as grammaticality judgments, but this term is misleading. Since a grammar is a mental construct not accessible to conscious awareness, speakers cannot have any impression about the status of a sentence with respect to that grammar; rather [...] one should say their reactions concern acceptability, that is, the extent to which the sentence sound “good” or “bad” to them.”

– Schütze & Sprouse (2013:27–28)

Linguists often ask themselves whether a particular sentence is grammatical or not. A property of a sentence is that it is a real, observable entity (at least when it is pronounced or written). Although we usually say that we are interested in the grammaticality of a sentence that is not what we are really concerned about. As linguists, we want to know, if a particular grammatical construction, and not a particular sentence, is acceptable or not. To be more precise, the question is whether a particular construction is part of the grammar of a language or not.

As grammar is abstract and not directly observable, we have to stick with individual (concrete) sentences. Things that are not directly observable can nevertheless be measured. Think of intelligence. Intelligence is measured in IQ tests, although it is not a ‘real thing’ that we can see, touch, or measure as we can measure, for example, the height of a person. What you can measure, though, is how an individual is capable of solving a particular exercise. The outcome, i. e., the score, of such an exercise is called an ‘observable variable’ or a ‘manifest variable’. Notice that the term ‘variable’ here simply means that something is measured or counted. It is simply a number. The opposite of a manifest variable is something that cannot directly be observed, like intelligence or grammaticality. We call such abstract concepts ‘latent variables’ or ‘constructs’. To sum up, what we want to measure is an individual’s intelligence (the construct), but we cannot measure it directly. What we can measure directly is the performance of an individual in a particular exercise (a variable). However from the performance of one exercise, we cannot conclude how intelligent an individual is. The individual’s performance on one task can be influenced by many things. Namely, the person could have been distracted by something, it could be that the exercise contains a word the individual does not understand without which s/he would have been able to solve the problem, and many more things.

For linguistics, this means that there is a big difference between a particular sentence and the rules that were used to form or understand this sentence (or that are applied in judging this sentence). What we, as linguists, are interested in is the abstract concept of grammaticality. Thus, grammaticality is a latent variable or a construct that we cannot measure directly. What we can measure is how much individuals like a particular sentence (or a set of particular sentences). The manifest variable (‘How much do you like the sentence from 1 meaning not at all to 7 the sentence is fine’) we can measure is called the acceptability of a sentence (or phrase or word or dialogue). Acceptability is about how much an individual accepts a sentence as being formed according to his or her internal grammar:

- Grammaticality: Not directly observable; an abstract rule that is part of the internal grammar of an individual or of a group of individuals; transferable into a construct (or: latent variable)

- Acceptability: Directly observable through ratings; measurable through concrete entities, namely sentences; a manifest variable (or: observable variable)

Note that I sometimes use the term ‘grammaticality judgments’ while others prefer the term ‘acceptability judgments’ (or ‘acceptability/grammaticality rating’). You will see in the course of this tutorial that both terms are correct in their own way: We use individual acceptability judgments to obtain a grammaticality judgment.

In linguistic terms, following Chomsky (1965), this corresponds to the distinction between performance and competence. As Schütze (2016:20) notes: “Whether a sentence is acceptable is a question about performance.” And: “Whether a sentence is grammatical is a question about competence”. This means that we use the actual behavior of speakers (performance) of a language to get a clue about their grammar (competence). Again: asking a participant whether a sentence is acceptable is something we can measure (a variable) and the underlying grammatical construction or rule is not, although we are interested in it (it is a construct). Actually, we do not even know if the rule we are after in fact is a rule that exists in native speakers’ brains. We don’t know this because we cannot look it up. That’s why it is called a construct as we constructed something we believe exists (or does not exist).

Asking someone for an acceptability judgment raises at least two problems: The first problem is that the internal grammar of the particular person you’re asking could diverge from the grammar of others (you may know this phenomenon from classes or conferences where some people like a sentence and others don’t). This is why you have to ask several informants and not only one (I will come to the question of how many participants you need below).

Excursus: More on variation

That variation between speakers exists cannot be stressed enough because any linguistic theory should be able to model this variation: “It has come to be generally acknowledged that not all speakers of ‘the same language’ might have the same competence, but that does not justify basing the theory only on sentences for which there is universal agreement, and extrapolating by some means to dictate the status of the remainder. In cases where people disagree, that fact cannot be ignored; the theory must be able to describe every speaker’s competence, and thus must allow for variation wherever it occurs” (Schütze 2016:37).

And that variation between speakers will occur in your study is nearly guaranteed! Disagreement between informants is reported since the very beginning of the application of judgment tasks for linguistic purposes (e.g., Hill 1961; Heringer 1970; Labov 1971; for a more comprehensive overview see Cowart 1997:4–5).

The second problem is that an individual might not like a sentence because of an unforeseeable number of reasons. For example: They may be tired or in a bad mood (another reason for asking a large amount of people!) or they may not like its semantics, because the sentence is about dogs and the person is a cat person. Especially, when you are not asking linguists, but laymen, people often don't really understand what you want them to judge and rely on the semantic content or the wording. The solution for this problem is that you do not test one construction using one sentence, but use several sentences that were built from the same rule. This procedure is well-known from psychometrics and classical test theory (e.g., McIver & Carmines 1983), Nunnally & Bernstein (1994), or Oppenheim 1992) so this is what we will take a look at next.

3. Test theory: Creating a construct

“One item a scale doth not make.”

– Carifio & Perla (2007:110)

What I have said so far means: What you want is to measure something that is not directly observable. What you can do, however, is to measure how much people like individual sentences. So you measure acceptability by creating sentences. These sentences are rated by several participants. What you create is a measure of the unobservable (an assumed abstract grammatical rule).

I will now rephrase what I just said very briefly using different terminology: Let's call each sentence you create an item. Assume a very simple case: You want to know something about the linear order of two parts of speech X and Z. There are two possible orders, namely XZ and ZX and your theory predicts XZ, but *ZX. You decide to do a simple judgment task. It is not a good idea to test only two sentences, since the wording of those two sentences could influence your participant's judgments. Additionally, you don't want to test how item a and item b (so, two sentences) are rated, but you want to know how much people, in general, like the construction XZ and ZX: Does their internal grammar allow for both (abstract) rules or only for one?

Now, let's just concentrate on the order XZ. As a friend told you 5 would be a good number, you create 5 items containing the order XZ (and likewise 5 items which only differ from the first 5 items in that they contain the order ZX). The assumption you make is depicted in Figure 1. The figure shows that the construct should determine the judgments of any of the individuals, because it is a representation of the abstract rule XZ you are looking for. Of course, there will be other influences on your items! They will never be judged exactly the same if you ask ten or more people! But: If you constructed your items carefully, it is very likely, that the sentences will be rated very similar. This similarity should show up in form of a correlation. A correlation is a systematic relationship between numbers. The numbers in this case are your ratings. That the ratings correlate means

that all your items measure the same thing. This correlation between all your items is depicted in the lower part of the graphic.

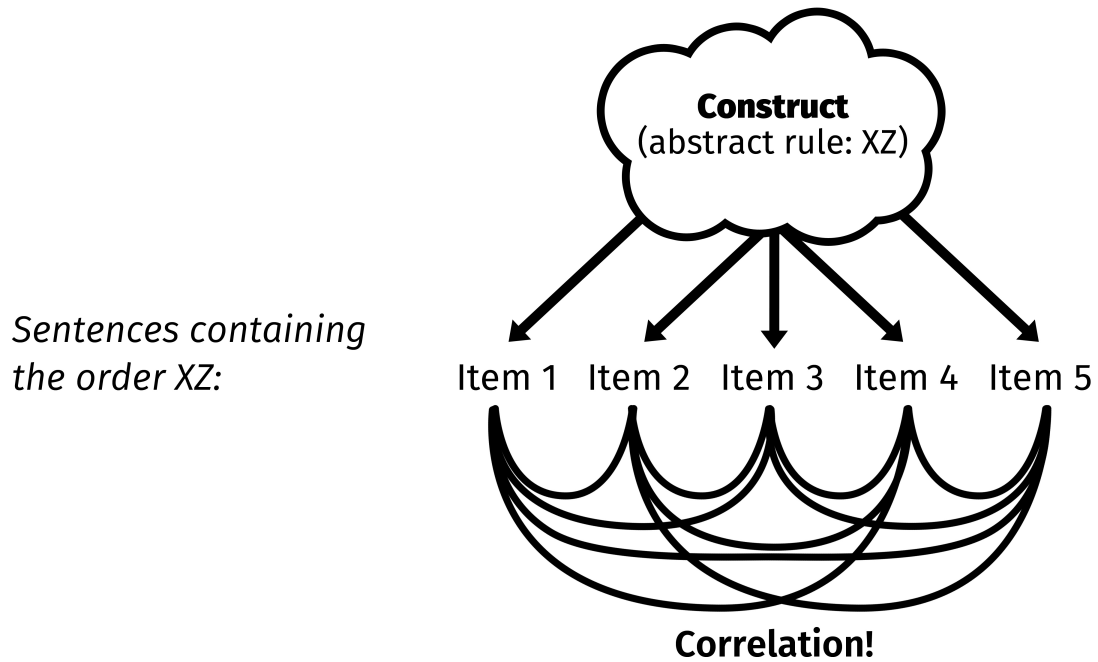


Figure 1: One construct, five items that all correlate

Excursus: Measuring the Correlation

To find out if your items (your sentences) really measure the same thing (i. e. form a scale), you calculate a coefficient called Cronbach's alpha (also called 'tau-equivalent reliability') first described in Cronbach (1951). The formula to do this is:

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum s_i^2}{s_T^2} \right) \quad (1)$$

Where: k = number of items, s_i^2 = variance of the i th item, and s_T^2 = the variance of the sum of all items. As the computer will take care of the math, I will not go into more details here. What you have to know to calculate Cronbach's alpha will become clear by using a very simple example. Assume you have 5 sentences with the same grammatical construction underlying them. Those are your items I1, I2, I3, I4, and I5. As they should all measure the same thing and thus create a scale, their ratings should correlate. You ask 5 people to rate the sentences from 1 to 7 (1 = unnatural, 7 = natural). Let's call the participants P1, P2, P3, P4, and P5.

	I1	I2	I3	I4	I5
P1	3	2	2	3	2
P2	2	2	2	2	3
P3	2	3	3	2	3
P4	2	2	2	3	3
P5	3	3	2	3	3

You can calculate Cronbach's alpha with nearly every common statistics software. You can, for example, use the psych package in R. There are also some very simple to use online resources, for example, https://www.wessa.net/rwasp_cronbach.wasp. On this page, you simply type in the numbers from the table above. The result is an Cronbach's alpha of 0.7576. The highest number you are able to get is a 1 (if you get a negative value, something is wrong). The rule of thumb is that your Cronbach's alpha should be greater than 0.65, so we are doing fine with our value.

Note that the concept of Cronbach's alpha is a little more complex than I presented it here. If you want to find out more I recommend Tavakol & Dennick (2011) as a starter.

The psychological reality of your construct, of course, is not self-evident. Even if 1,000 participants rate all of your 5 items with a 7 (meaning: totally natural), it still could be that there is some hidden factor that you are not aware of. And there is another thing you should be aware of: We now only looked at the order XZ that we predicted to be correct, but in practice you also want to look at the order that should not be well-formed. As we constructed 5 items containing XZ, we also construct 5 items containing ZX, the order that should be ill-formed according to the theory we're working with (i.e., *XZ is the hypothesis). But what does this mean? The answer is: We don't really know! It could mean that there is some abstract rule in the heads of the people that does not allow this order. Or it could mean that there is no rule for this at all! But what have you measured then? What is your construct? In real life, we often simply don't care much about this.

To get an accurate measure of a construct it is useful to understand the basics of classical test theory. As the name suggests, test theory is the theory of how a test works. I will introduce the classical test theory via a simple example: high jump. Let's assume that you are interested in how high your best friend is able to jump. The height your friend is able to jump is a construct. 'Wait? What? But I can measure how high somebody can jump!'. Of course, you can measure the height of a jump. But the height someone is able to jump varies from jump to jump. The 'real' height is not observable.

What you can observe are only instances of jumps.³ What you're interested in is the capability of your friend to jump as high as possible (latent variable/construct), not how high she can jump right now (observable variable).

To measure how high a person is able to jump we use a simple method known from high jump competitions: Your friend will repeatedly jump over a horizontally placed bar. For each height your friend has, let's say, five tries. If your friend is able to jump over the bar at a certain height without dislodging it, the bar is raised to the next level. How high a person can jump depends on several factors: On the mood of that person, on her own height, if the person feels sick, and also on factors that lie outside of that individual, e. g., if she is carrying a backpack if it is super hot, super cold, or even raining. But even without such extreme conditions there will be variation between each jump.

This variation, however, follows a pattern. There will be a few very bad attempts. There will also be a few extraordinary good jumps. But there will be many jumps around one height. Some of them may be a little higher, some a little lower. They will vary around your friend's real capability. You may remember the term 'Gaussian distribution', also called 'normal distribution' from school days (or your statistics class). That means that a bell-shaped curve will emerge, see Figure 2. Put simply, a normal distribution is a symmetrical curve that is shaped like a bell. There will be a few very low and a few very high jumps. This is represented by the left and right. Here, the curve is very low, meaning that only a few instances of really low and really high jumps were measured. There is also one value that occurs very often. This is the value in the middle (the dotted line). Such a curve, of course, will only emerge if you do a lot of jumps, ideally, infinitely many. The true value is simply calculated as the mean of all measurements. However, the true value only emerges if you have an indefinitely large set of measures—a rather unrealistic scenario, of course.

So, to get such a curve, we really need a lot of data. The best thing would be, that we measure your friend not only once on one day, but measure on several days. Perhaps, we do not only want to know how high your friend is able to jump, but how high a person in general is able to jump so we should measure a lot of different people on many different occasions to get the 'real' value of the capability.

³Sometimes a latent variable of this kind, i. e., a latent variable that can be measured directly is called a 'hidden variable'.

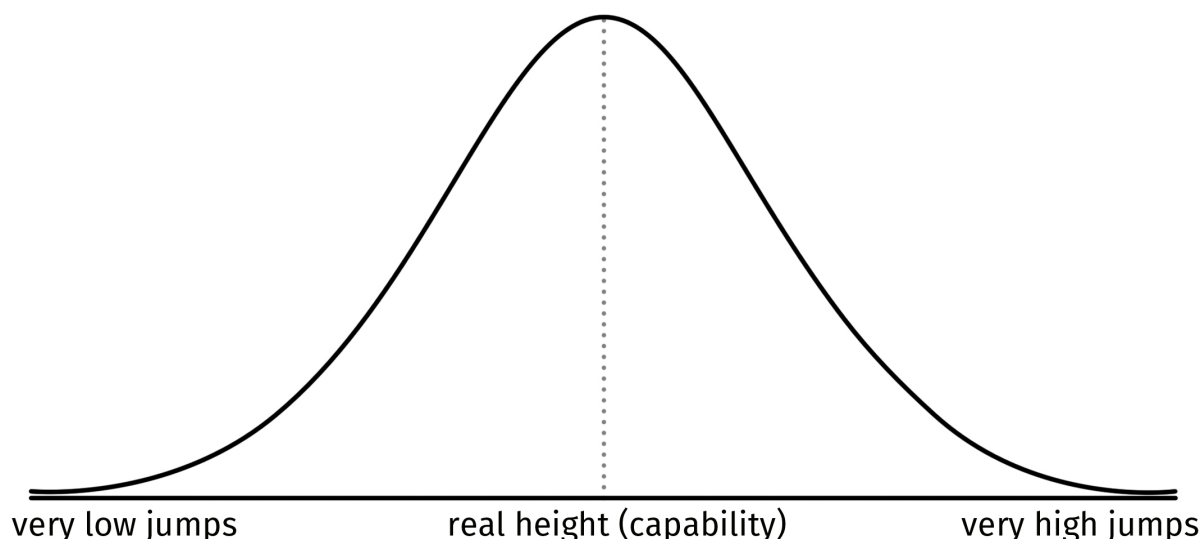


Figure 2: A normal distribution

Excursus: Regression to the Mean

You will find normal distributions everywhere in nature. They come into play when some characteristic is determined by various factors. Think of the length of an average (adult) elephant. The length of an elephant depends on a number of factors. Those factors are, for example, genetics, nutrition, or whether or not the elephant was raised in the wild or in a circus (hopefully not!). If you measure, let's say, a million elephants you will make the following observation: There will be a few very short elephants and a few very long ones. And there will be a lot of elephants that have one particular size (we do measure a little coarse and do not take centimeters into account, or course, every elephant will be of a different size if our measure is infinitivally fine-grained). If you sum up all the lengths of the elephants and divide it by a million (i.e., the number of elephants you measured), the average length will be the one that occurs most often.

This phenomenon is called 'regression to the mean' and was discovered by the English scientist Sir Francis Galton (you may also know him as the 'discoverer' of the fingerprint). Galton measured the heights of people from the same families. What he found was that there was something like a mean height for people in general. Of course, he found some very tall men and some very short ones. But the sons of the very tall men were, on average, a little shorter than their fathers. And the sons of the very short ones were, on average, a little taller. So in the

end, there are extremes, but nature tends to keep them small and around a mean.

At this point, I want to make two recommendations. The first one is a reading and the second one is a watching recommendation. Let's start with my reading tip. If you are a little interested in statistics, I recommend reading Salsburg (2001). About the Galton's discovery of the regression to the mean he writes (p. 13):

Suppose[...] that regression to the mean did not occur. Then, on the average, the sons of tall fathers would be as tall as their fathers. In this case, some of the sons would have to be taller than their fathers (in order to average out the ones who are shorter). The sons of this generation of taller men would then average their heights, so some sons would be even taller. It would go on, generation after generation. Similarly, there would be some sons shorter than their fathers, and some grandsons even shorter, and so on. After not too many generations, the human race would consist of even taller people at one end and ever shorter ones on the other.

Salsburg's book is not an introduction to statistics, but a very well written history of statistics. It's really fun to read!

The other recommendation is that you google the term 'Galton board'. You will find some very interesting videos showing how nature creates normal distributions. This will help you understand the concept better (I will have to say more about normal distributions later).

So far, so good. Now, classical test theory comes into play: this theory takes into account that there is variation in your measurement. We call this variation the 'measurement error', the 'error variation' or simply the 'error', abbreviated with the Greek letter ϵ (epsilon). Let's say, your friend jumps 70 centimeters high on one occasion. We abbreviate this observed score X . Mathematically speaking, this score X is the result of two values: there is the true height your friend is able to jump and there is variation. Thus, we can write the following formula:

$$X = T + \epsilon \quad (2)$$

The formula says, that each score that you will measure is decomposable into a true value T (your friend's capability) and an error ϵ (this error variation can also be negative).⁴

⁴The true score, i. e., the latent variable, is often represented by the Greek letter eta (η). I'll stick with T here.

You may think that we have the problem that we can only observe X (how high did your friend jump/how does an individual rate a sentence), but not T (the ‘real’ height your friend is able to jump/how grammatical is a construction) and ε . However, this is not entirely true. Let’s look at the definition of T :

$$T = M(x) \tag{3}$$

The true value T is defined as the mean of an infinite number of observed scores. Of course, you cannot calculate the mean of an infinite number of scores. What you can calculate, however, is the mean of as many observed values as possible. Indeed, there is not even the need for as many observations/measurements as possible, a reasonable number is enough (we’ll talk about this in a second).

The formulas we have seen so far are axioms (i. e., they serve as a premise, in this case for classical test theory). We can deduce some more formulas from these axioms. From Formula 2, we can get:

$$\varepsilon = X - T \tag{4}$$

What 4 tells us is that the error is the observed value minus the true value. We can also say that, if we make an infinite number of measurements, the sum of all the errors will be zero:

$$\sum_{i=1}^{\infty} \varepsilon = 0 \tag{5}$$

And it doesn’t matter if we talk about measurements from one person or a whole population. Please make sure you understand this! Also look at Figure 2 again. There will be values greater than the mean and values lower than the mean (i. e., some are positive and some are negative). With an infinite number of measurements, the sum of these error will equal zero. Again, we do not have an indefinite set of measurements, but only a limited number. But after collecting a certain number of measurements this does not matter anymore because at a certain point the error will become so low that we simply can ignore it.

When we talk about an error considering the formula in 5, this means that we do not talk about a systematic error, but an unsystematic one. A systematic error (also called ‘bias’) would be, for example, if you would measure your jumping friend only on Mondays after a hard weekend (she would be systematically worse). An unsystematic error arises from a lot of different sources. Sometimes your friend is in a very good mood when she jumps, sometimes she isn’t, sometimes it’s raining and sometimes the sun shines, etc. Sometimes she jumps a little higher and sometimes not.

4. The foundations of acceptability judgments: Measurement theory

“Measurement is never better than the empirical operations by which it is carried out, and operations range from bad to good. Any particular scale, sensory or physical, may be objected to on the grounds of bias, low precision, restricted generality, and other factors, but the objector should remember that these are relative and practical matters and that no scale used by mortals is perfectly free of their taint.”

– Stevens (1946:680)

As far as I have presented grammaticality judgments so far, I have said that participants will rate your items from 1 to 7. This is not the only way to do grammaticality judgments. You could also ask them to rate from 1 to 6. Or just rate if the sentences are acceptable or not. Additionally, there are alternative ways to do acceptability judgments.

There are, however, good reasons to stick with a rating from 1 to 7. First, a rating from 1 to 7 is more informative than just asking if a sentence is acceptable or not. You could, of course also ask participants to rate from 1 to 6 or from 1 to 10. This has the advantage that in such ratings, there is no middle and participants are forced to make a decision in one direction (I think a range from 1 to 10 is too big and people could get confused). See Figure 3. However, there are empirical reasons to stick with a rating from 1 to 7: Using this kind of scale simply has proven to be useful (e. g., Lewis 1993; Finstad 2010; see also Preston & Colman 2000). However, using a rating from 1 to 5 does not do a lot of harm.

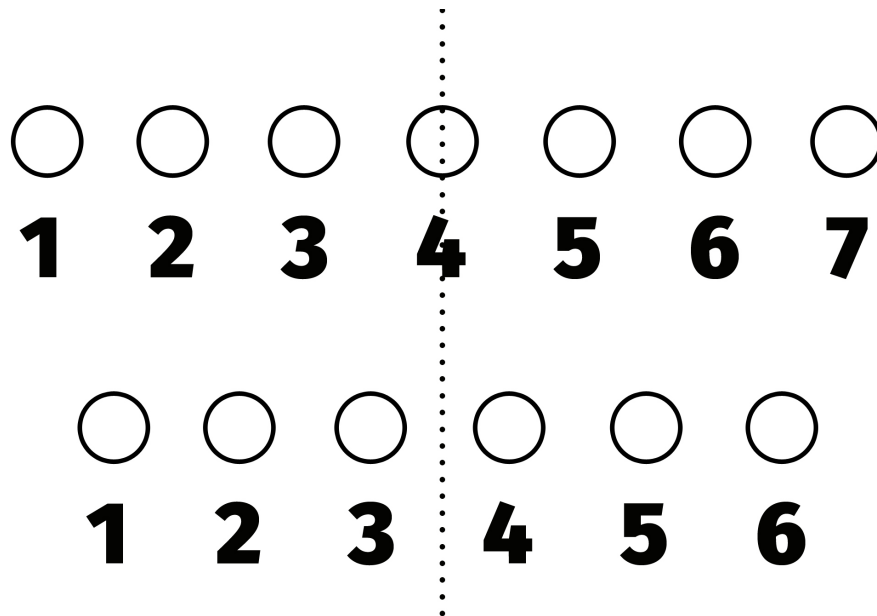


Figure 3: Some rating formats have a middle, i. e., an undecided category, others don't.

Rating formats like the one we use here are called ‘Likert response format’ or ‘Likert-type scale’, named after the inventor of this format (Likert 1932). You surely know the Likert response format. It consists of questions like ‘How many cigarettes do you smoke per day?’ and several options to answer like: ‘1 = 0’, ‘2 = less than 5’, ‘3 = 5 or more’, ‘4 = 10 or more’, ‘5 = 20 or more’. Another example would be: ‘How do you rate the following statement: Beer is important for linguists: 1 = totally agree 2 = agree 3 = neutral 4 = do not agree 5 = do not agree at all’. Many people call such rating scales ‘Likert scales’, but this is not entirely correct and it is useful to understand why: With grammaticality judgments we construct several items for one construct that we want to know something about. The term ‘Likert scale’ refers to the set of items that measures such a construct. In our case this is the set of sentences that are created according to the same linguistic rule. When we talk about how a single sentence is rated there is, of course, also a scale, namely a scale from 1 to 7. We call a particular instance of a sentence to be judged on such a scale a Likert item or a sentence that is judged in a Likert response format. What we need to distinguish is the response format we use for one item (the Likert response format, often from 1 to 7) and a given measurement scale. One item does not constitute a scale. I’ll talk about this issue a little more when I come to the statistical analysis of Likert-type data, but if you want to know more, a short and easy-to-read resource is Carifio & Perla (2007). Their take-home message is “a single item is not a scale in the sense of a measurement scale” (p. 110). This simply means that there is a difference between the response format (e.g., from 1 to 7) and a scale measuring an underlying construct consisting of interrelated items.

It is time now to think a little bit about math. Imagine there are two babies. Peter is 1 year old and Tamara is 2 years old. It is very easy to see that Tamara is older than Peter. Actually, we can be more precise as Tamara is twice as old as Peter. Now look again at the Likert-type scale about tobacco consumption I just mentioned. Someone who answered 2 clearly smokes more than someone who answered 1. But in this case, 2 is not twice as much as 1! What we learn from this is very simple, but very important. There are different kinds of numbers. Actually, there are not different kinds of numbers, but different kinds of data. Depending on the data different mathematical operations are possible. You could, for example, add the ages of the babies up. It makes totally sense to say that they are 3 years old together. But someone who answered 1 when asked how many cigarettes he smokes and someone who answered the same question with a 2 do not smoke 3 cigarettes a day. In fact, you don’t know how much they smoke exactly!

Before you start your study, you have to know what your data will look like. And this is often really important: You have to know what kind of mathematical operations are possible with the kind of data you will get before you even start (as the kind of data, for example, determines how many participants you will need to consult). The reason for this is, of course, that there are different mathematical operations possible with different

kinds of data. What we need is a theory of measurement. There are actually several theories of measurement and there is a huge philosophical debate about the nature of data, about the world, and about mathematics. The theory of measurement we need for our purposes, however, will be quite simple. Nevertheless, it is a good idea to keep in mind that measurement itself presupposes a theory, namely that the thing you are looking at is indeed measurable:

Measurement always presupposes theory: the claim that an attribute is quantitative is, itself, always a theory and that claim is generally embedded within a much wider quantitative theory involving the hypothesis that specific quantitative relationships between attributes obtain. (Michell 1997:359)

The first question we want to ask when we hear a term like ‘theory of x’ is: What is x? So, when talking about a theory of measurement we ask ourselves what measurement is. The classical answer to this question comes from a paper that laid the groundwork for every modern theory of measurement: “... measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules” (Stevens 1946:677). For this purpose, Stanley Smith Stevens developed a classification of different levels of measurement (or: scales of measurement). For our purposes, we need to distinguish three levels:

- Nominal level: Nominal type data is data to which you assign numbers without any empirical relevance. This means that there is no real relation between the numbers and the things you measure. You could say, for example, that all the Spanish speakers in your study will be assigned the number 1 and all Russian speakers the number 2. This does not mean that 2 is more than 1. They are just names (actually, if you know a little Latin the word ‘nominal’ says that). Why should I use the nominal level, you may ask. And that’s a good question! We don’t really need the nominal level here, but it is often useful to have numbers assigned to your data when applying statistical models to your data (e. g., in multiple regression).
- Ordinal level: An ordinal scale represents a rank. This is true for example for a high school diploma, bachelor’s degree, a master’s degree, and a doctor’s degree. They are ranked exactly in this order, so a master’s degree is considered higher than a bachelor’s (since a bachelor’s degree is usually the prerequisite for a master’s degree). You could imagine that you could assign numbers to the degrees and say 1 = high school, 2 = bachelor, 3 = master, 4 = doctorate. As you can imagine, there is only a limited number of mathematical operations you could apply to ordinal data. And you surely have recognized: The way the Likert-type scale asked about tobacco consumption leads to the fact that the data was coded as ordinal

type data. The crucial thing with this kind of data is that the data is ranked, but that the intervals between the steps are either unequal, unclear, or not defined.

- Interval level: Interval type data in fact is very similar to ordinal data, but there is one difference. As ordinal data, interval data is ranked. But additionally, the steps between the scale points are of the same size. A very simple, but very good example is age. The age difference between a 10 year old and a 20 year old is exactly the same as the difference between a 50 year old and a 60 year old. The same was not true for the ordinal level! The difference between a high school diploma and a bachelor is not the same as the difference between a master's and a doctor's degree. In fact nobody can really tell what the difference between two degrees is (in measurable terms of distance).

It is important to note that there is some kind of ranking in the scale levels themselves. Interval scale data represents the highest value in this ranking. This is not only because it provides more information than the lower levels, but because the higher the level, the more mathematical operations are possible. And there is also the following rule: A mathematical operation that can be applied at one level can also be applied to the next higher level. This means, for example, that all operations that are possible for ordinal level data can also be applied to interval level data.

5. Measures of central tendency

Let's look at two very simple mathematical operations we can apply. The first one is called the 'mean' (or 'arithmetic mean') and I'm sure you know what a mean is and how it is calculated. However, if you are not trained using formulas, the equation for the mean might look a little bit scary at first:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (6)$$

There are several things in the formula. Letters, numbers and the sign \sum which is the Greek letter sigma that tells us that we deal with a summation. Let's look at the letters. There are three of them, namely x , n , and i . With n we mean the number of data points we have, i. e., the number of measurements. With x we refer to the individual values we measured. The small \bar{x} with the bar on its head is our mean. Finally, there is a small i . We do not care much about it. It is just an index that tells us that we should start our addition with the first value. The formula just says that we should sum up all the values we measured starting with the first one and divide them by the number of measurements. Assume, we asked 10 people how old they are. As every person has only one age, our n equals 10. What they tell us is the following:

20, 18, 19, 24, 24, 23, 23, 98, 14, 17

Those 10 data points are our 10 x s. If we wanted, we could assign labels to them in the way $x_1 = 20$, $x_2 = 18$, $x_3 = 19$ and so on. Applying the formula in 6 we get:

$$\bar{x} = \frac{20 + 18 + 19 + 24 + 24 + 23 + 23 + 98 + 14 + 17}{10} = \frac{287}{10} = 28.00 \quad (7)$$

This means that the mean age of our participants is 28.00. However, there is one odd thing in our calculation. You surely have noticed that there is one very old person in the group who is 98 years old. Yet, the mean age is still pretty low. But if you compare the mean age with the ages of the participants excluding the 98 year old, the mean age actually is pretty high. This is not good and is actually a misrepresentation of our data. The reason for this is that the mean is a point measure. Point measures are not very informative. For this reason we will see more measures in a second.

When you think back to the scale levels you will notice that the calculation of the mean only makes sense for interval data. It does not make sense to calculate a mean for ordinate level data. If you sum up the numbers you assigned to academic degrees of people and divide it by the number of people you will get some crazy number you cannot make sense of.⁵

Let's get back to the question of what to do with the fact that our point measure is skewed by one very high data point. There is a measure that is not susceptible for outliers as the mean is: the median. The median is defined as the value above which 50 % of the values lie. A logical consequence of this definition is, of course, that the other 50 % of the values lie below the median. Thus, the median is some kind of middle value. Let's look at an example. In the following table you see the grades of 11 people in two subjects (music and math). As they attend a German school, they were graded from 1 (excellent) to 6 (insufficient). We calculate the median and not the mean in this case, because grades are not interval, but ordinal data. This is because the distances between the grades are not defined, i. e., you cannot say that, for example, a 6 is double as bad as a 3.

⁵You can do this of course, the numbers don't care. Actually there is an entertaining paper by Lord (1953) about a similar case. It's called "On the statistical treatment of football numbers".

Pupil	Grade (music)	Grade (math)
Hans	3	1
Daniel	3	1
Eva-Maria	1	4
Theresa	6	1
Lorenz	5	4
Julian	4	2
Torolf	1	3
Gökce	1	4
Katie	3	2
Philip	2	4
Ursula	1	—

To calculate the median grade in music we simply order the grades: 1, 1, 1, 1, 2, 3, 3, 3, 4, 5, 6. The median is the number in the middle, i. e., 3. This is not as easy for the math test, since there is no middle (Ursula was sick). Nevertheless, we order the numbers: 1, 1, 1, 2, 2, 3, 4, 4, 4, 4, 4. To arrive at the median, we take the two numbers that are in the middle (2 and 3) and calculate their mean: $\frac{2+3}{2} = 2.50$.

Remember that the mean we calculated for the ages of 10 peoples was 28.00. If we calculate the median of their ages we get 21.50. This number seems to capture the ages of the individuals more naturally. That the median does not care about outliers can be illustrated by the following fact. There was one person of the age of 98. If this person was not 98, but 969 years old (as Methuselah), the median still would be 21.50, but the mean would be 115.10.

Finally, there is also a measure of central tendency which is called ‘mode’. The mode is simply that value that occurs most often. It can be applied to nominal data. As I said that mathematical operations that can be used on one scale level can also be used on higher levels, the mode can be calculated not only for nominal scales, but also for ordinal and interval data. The median can be calculated for ordinal and interval data and the mean only for interval data. This is depicted in Figure 4.

6. Measures of dispersion: The standard deviation

The point measures we have seen so far are only of limited use as was illustrated by the mean of the ages of 10 people. Another example would be the mean ratings of 10 items. Suppose, there are 10 images and people are requested to rate how much they like the image on a rating scale from 1 (‘I hate it’) to 10 (‘I love it’). Participant 1 rates all the images with a 5 (5, 5, 5, 5, 5, 5, 5, 5, 5, 5), because they all seem to be mediocre to him.

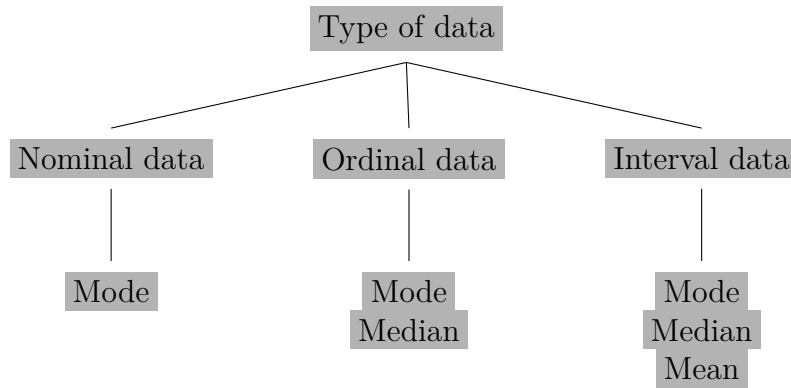


Figure 4: Measures of central tendency and scales of measurements

Participant 2 likes half of the images and dislikes the other half. His ratings look like: 8, 2, 8, 2, 8, 2, 8, 2, 8, 2. In both cases, the mean will be 5 and thus the mean does not tell you pretty much about the dispersion of the values.

You have already heard about the bell curve and about the regression to the mean (the mathematical concept behind this is called ‘central limit theorem’). Bell curves can look very different, but all bell curves have one thing in common: they all have one peak (the mean) and are symmetrically organized around this peak. However, some bell curves are very slim, so the values around the mean do not vary much and other bell curves are wider, so the values spread more around the mean. One of the most popular measures of dispersion is based on how bell curves look: the standard deviation (abbreviated SD or σ). The standard deviation allows you to mathematically describe the shape of a bell curve.

At this point I want to draw your attention to a very important issue I have already talked about, but since it is very important, I will take some space to say a few more words. You have to keep two things apart: On the one hand there is a population that you cannot access and on the other hand there is the data you collected that stems from this population. Your population consists of a probably infinite set of values (as you can ask an incredible amount of participants for judgments of an infinite number of sentences that use the construction you’re interested in). The data you have is just a random sample from this population. You are interested in the population, but you only have your data. What you want is to measure parameters (that mathematically describe your population), but you only have some data points. Those data points, however, stem from the population and you have an assumption about your population, namely that it is a normal distribution, i.e., that it is bell-shaped. What this means is that if you collect enough data you can approximately calculate the underlying normal distribution. Think

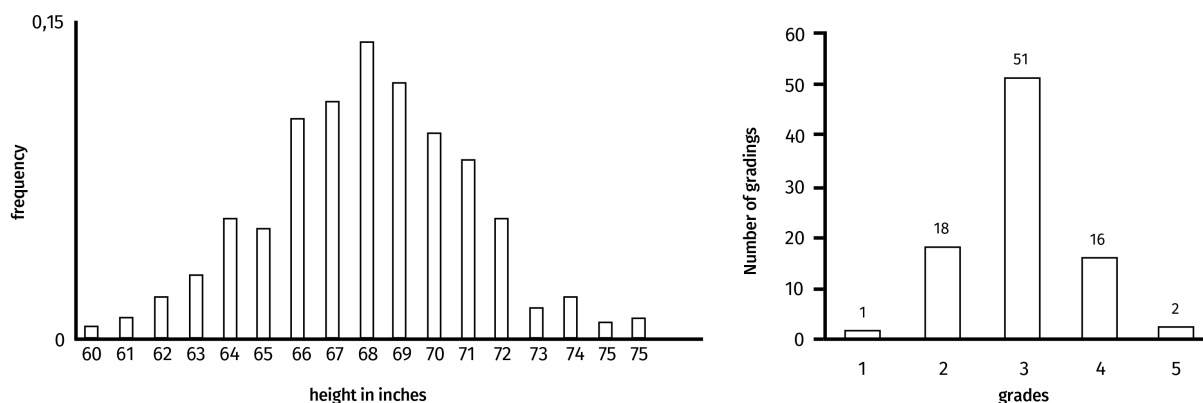


Figure 5: You do not need as much data as in the examples here to be able to get an approximation of the underlying distribution. On the left: the distribution of heights of 500 US citizens (based on Nuñez 2007). On the right: the distribution of grades (again: German grades from 1 to 6) of the same essay given by 88 different elementary school teachers (based on Birkel & Birkel 2002). In both cases a normal distribution arises.

about the elephants again. Assume that there are 1 million elephants in total.⁶ If you measure all of them you would get a normal distribution. However, you do not need to measure all of them, but only a wisely chosen subset of them. All the elephants out there, namely 1 million elephants, are called the ‘population’. The subset you chose for your measures is called ‘sample’. Figure 5 gives you more examples to understand this.

Make sure you have read the caption of Figure 5. The height of the US-citizens depicted on the left is in fact an example that is not very different from the elephant example, except that we are talking about people and height instead of elephants and size. As you can see from the data, the mean height is 68 inches. You can also see that the data is distributed in a way that nearly forms a bell curve. The curve is not perfect, but given that it is only based on measurements of 500 people it is impressive how perfect it looks. If you were to measure more and more people, the curve would look smoother. If you had an infinite number of measurements the curve would be perfectly smooth and symmetric. A perfectly smooth and symmetric curve can be described mathematically. We will call this curve the ‘underlying distribution’. The problem with underlying distributions is that they are rarely achievable, as knowing an underlying distribution would require us to have access to the entire population we are interested in.

An underlying distribution has a mean and a standard deviation. While we have called the mean of some measurements we have obtained \bar{x} (as it was created out of the individual measures x_1, x_2, x_3 etc.) we call the mean of a distribution μ . This value is fixed for each distribution. This is a value we’re interested in, but also a value we don’t

⁶Sadly, there are less elephants in the world in reality.

know. On the other hand we have \bar{x} , the mean we calculated from our data. This value is our best guess of μ . We also have two standard deviations, namely σ and SD (or s). As you can guess, σ is the standard deviation of the underlying distribution that we don't know and s is our best guess of σ , i.e., the standard deviation we calculated from our data. The unknown numbers μ and σ are the parameters that define the shape of the distribution.

As I said, the standard deviation describes the form of a bell curve. When the SD is small, then most values are close to the mean (if the SD is 0 then all values are equal to the mean and there is no variation at all). When the SD is big, the values scatter around the mean to a greater extent. I want to stress that the SD describes the form of the distribution and is not its length. If you only knew the mean and the span of a distribution you still do not know how the bell is shaped and thus, your information is pretty much worthless. This is because there is much more information hidden in the SD than you may think. Before we look at these information, we take a quick look at the way the SD is calculated:

$$SD = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} \quad (8)$$

The formula tells us, that the SD is calculated as the square root of something. This something is a sum divided by the number of data points (n) minus one. We have already seen a sum divided by the number of data points earlier, namely the mean. Thus, what is inside of the square root is very similar to a mean. It is not easy to explain why we divide by $n - 1$ instead of n . The easy explanation is that you actually would divide by n if you would calculate the standard deviation of the population (i.e. σ). As you only have some data points, your SD is only an estimate of σ . And as estimates have an error, you want to somehow try to keep this error as small as possible. This, in short, is what dividing by $n - 1$ does. The bigger your sample is, the smaller the difference between dividing by $n - 1$ and dividing by n will be. In other words: The bigger your sample, the smaller your error will be. In fact, things are a little bit more complicated. If you are interested in this, you should look for the terms 'Bessel's correction' and 'degrees of freedom'. For now, however, we just say that we divide by the number of data points and just subtract 1 to try to correct an estimation error.

In the upper part of the formula you see that a sum is calculated. It is the sum of the quadratic deviation of all your data points from the mean \bar{x} . If you think of a bell curve, this makes complete sense. A bell curve is symmetrical around a mean. If you want to know how the data points scatter around the mean, you want to know some kind of mean

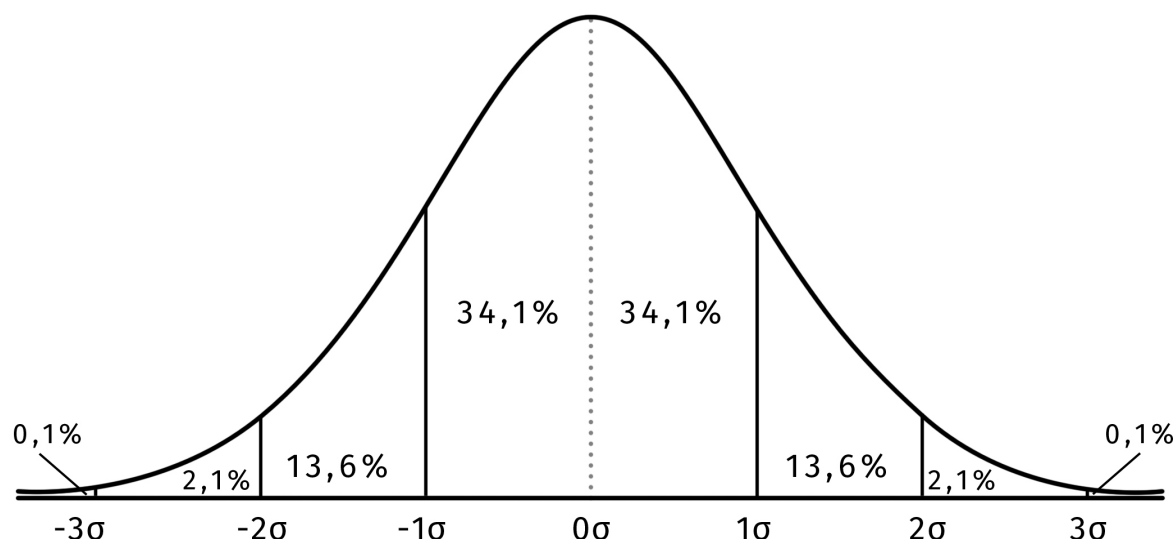


Figure 6: The relation between the standard deviation and the normal distribution (in this case, the standard normal distribution)

value of how all the data points diverge from \bar{x} .⁷ As these data points are symmetrically organized around \bar{x} simply adding up how far away they are from \bar{x} results in 0 because some values are higher than the mean so they are positive and some values are lower than the mean so they are negative. This is fairly understandable: Adding positive and negative values results in zero. This is prevented by the squaring. At the end you get rid of the squaring via the square root. This means that most of the formula is a simple trick that prevents the whole thing from being 0. The most important thing to know is that the *SD* is a measure of the dispersion of your data around the mean. Note that the *SD* always has the same unit as your data.

Now, we come back to the special properties of the *SD*. The good thing is that although these properties are mathematical in nature, we can actually see them in a picture as in Figure 6. In the middle of our distribution we can see our mean. In this case, the mean is 0 (normal distributions with $\mu = 0$ and $\sigma = 1$ are often called ‘standard normal distributions’). The bars represent our standard deviations. In the area of $\pm\sigma$ around the mean are 68.2% of all values and in the area of $\pm 2\sigma$ there are 95.4% of all the values.

There are more measures of dispersion, such as the interquartile range or the standard error of the mean. I will only briefly talk about the standard error of the mean. The standard error of the mean, SEM for short, is calculated as the standard deviation divided by the square root of the number of your data points:

⁷You’ll get this mean value by dividing the whole thing by $n - 1$.

$$SEM = \frac{SD}{\sqrt{n}} \quad (9)$$

As we divide by the number of data points, the SEM gets smaller the more data points we have. You don't need to understand what's really going on in this formula to proceed. We just need the SEM for many calculations in statistics. But note that there is a difference between *SD* and *SEM*. Additionally, there are many more statistical measures of dispersions. Many of them are used as error bars in plots. This means that you always have to indicate which error measure you have used, otherwise your reader cannot interpret your plots. In other words: Error bars without descriptions are useless information.

7. More about populations and samples

I have already stressed several times that there is a strict difference between the population and the sample. You want to know something about the population, but you only have a sample. This conceptual difference is marked in the way we write about populations and samples. Values that refer to the population are written in uppercase and values that refer to your sample in lowercase. Assume we want to know how long a specific endangered species of lizards is. There are only 1.240 lizards (I'm making these numbers up). Our population consists of $N = 1.240$. That's a lot of lizards! You don't have the time (or money) to measure them all. What you need to do is to take a random sample and measure this sample. Let's say, you take 20 of them. We write this down as $n = 20$. We can refer to each of the 20 values by a lowercase letter: $x_1, x_2, x_3, x_4 \dots x_{20}$.

Besides Latin letters, we also use Greek letters. They are used for population parameters that are gained through mathematical operations (this is not so consistently used, however). For example, the sample mean is written as \bar{x} and the population mean is written μ . If you look at Formula 9 again, you directly see that this is the formula to calculate the SEM of a sample. If you wanted to calculate the SEM of a whole population you would have written this instead:

$$SEM = \frac{\sigma}{\sqrt{N}} \quad (10)$$

Let's get back to our 1.240 lizards. At the top of Figure 7 you see the lengths of all 1.240 lizards (each dot represents one length). The distribution is the thing we want to know, but usually the thing we don't know. The mean length of our lizards is 5 centimeters ($\mu = 5$). As you can see, the population of lizard length is normally distributed, meaning that few are very short and few are really long (one of them indeed is veeeeery short and

one of them veeeeery long). Now look at the very bottom of the figure. There are 20 white circles forming a horizontal line. These circles represent a random sample of 20 lizards you caught and measured. Again, some of the lizards you caught are very short and some are very long. Although I'm making this up, the values you see really represent a random sample I've taken these values from the population via computer simulation. Let's get back to our 1.240 lizards. In the top of Figure 7 you see the lengths of all 1.240 lizards. The distribution is the thing we want to know, but usually the thing we don't know. The mean length of our lizards is 5 centimeters ($\mu = 5$). As you can see, the population of lizard length is normally distributed, meaning that few are very short and few are really long (one of them indeed is veeeeery short and one of them veeeeery long). Now look at the very bottom of the figure. There are 20 white circles forming a horizontal line. These circles represent a random sample of 20 lizards you caught and measured. Again, some of the lizards you caught are very short and some are very long. Although I'm making this up, the values you see really do represent a random sample I've taken for these values from the population via computer simulation.

There is another set of circles in the figure, namely a bunch of gray circles. Again, I gained these values by computer simulation. Each gray dot represents a mean of 20 randomly chosen values taken from the population. As you can see, each \bar{x} is surprisingly close to the real mean μ . But you can also see that each estimation, each \bar{x} is different. Some are closer to the real mean, some are farther away. Think about this for a second! Does it ring a bell?

8. Confidence intervals

If you take several samples from a population and calculate the means of your samples, some means are closer to the real mean and some are farther away. Suppose we do this again and again. The means we calculate will be normally distributed (i.e., a bell curve will emerge)! Of course, you will say, our population is also normally distributed. However, even if the population is NOT normally distributed, a large enough sample of means from this population will be normally distributed (see also Pearson 1931; Boneau 1960). It's like magic! And this magic helps us a lot since we do not really need worry if our underlying distribution is normally distributed or not.

My main point, however, is the following: Think about what happens when we take a lot of samples and calculate a lot of means. Suppose we pile up our means and look at them. Again, I ran a computer simulation that is depicted in Figure 8 (see Cummings 2012). The top of the figure shows the lengths of our lizards again. At the bottom you see that I took 254 random samples and calculated their means (so each dot is one mean).

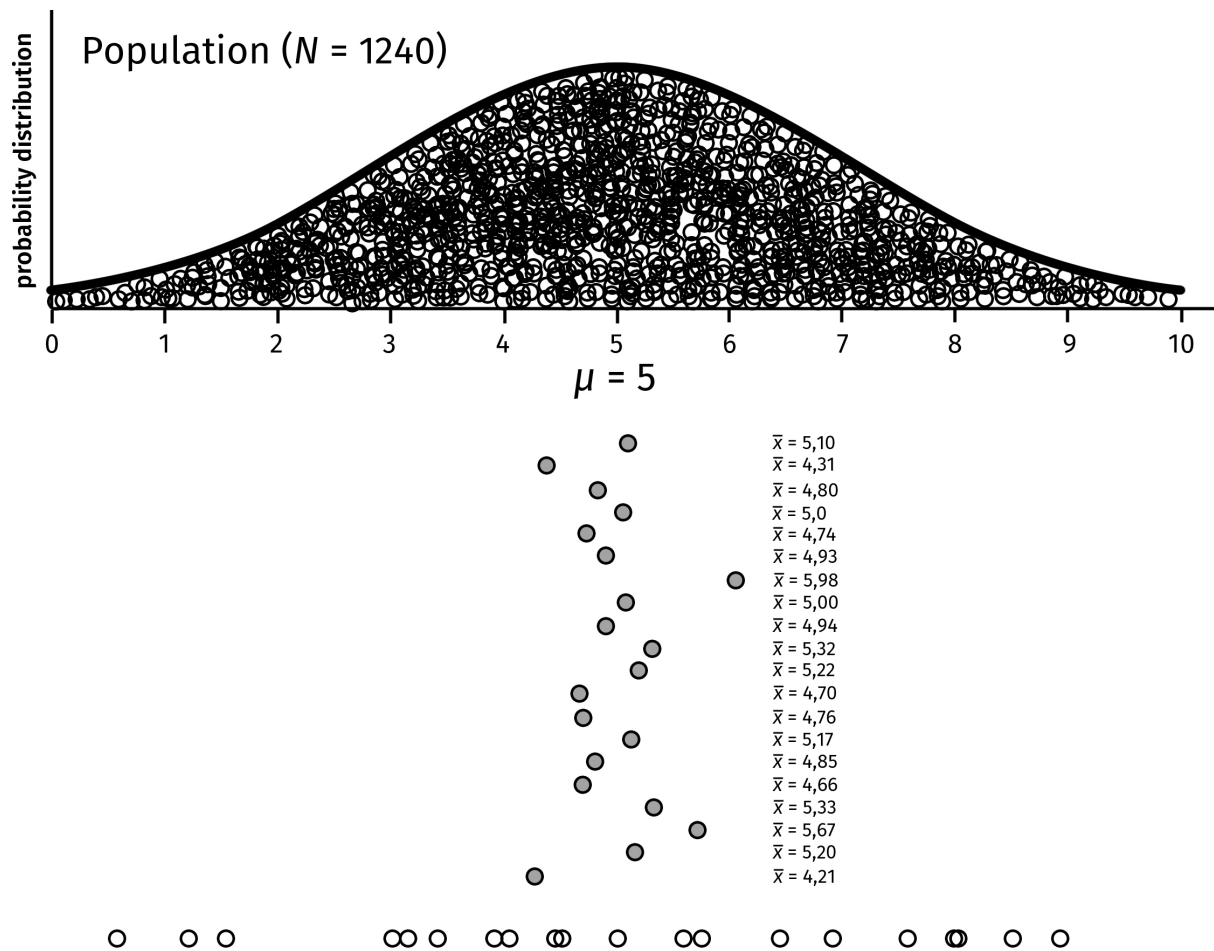


Figure 7: Top: the length of 1.240 lizards; a normally distributed population with a mean of $\mu = 5$. Bottom: look at the white circles first. They represent a random sample of the length of 20 lizards. Some of them are very short, some are very long. Now look at the gray dots. Each of the dots represents the mean of a randomly chosen sample of the length of 20 lizards. They are all pretty close to μ !

Some means are closer to the real mean than others.

When you look at the top half of the figure, you see that there are four lines in the population. Those represent the standard deviations of the population. We know that approximately 96 % of our data lie between ± 2 standard deviations. To be more precise: 95 % of the data lie in the range between $-1.96SD$ and $1.96SD$ (the outer lines). As the SD is something you can calculate for a normal distribution, there is also an SD for the distribution that is formed by the 254 means I calculated from random samples! You can see them at the bottom of the figure. Again, we know that 95 % of the values, that is 95 % of the sample means, lie between ± 1.96 standard deviations (i. e., between the outer lines). Think of it and make sure you understand what happens! This is really hard to process, but once you get it, it makes a lot of sense! Take a break and perhaps re-read parts of the tutorial.

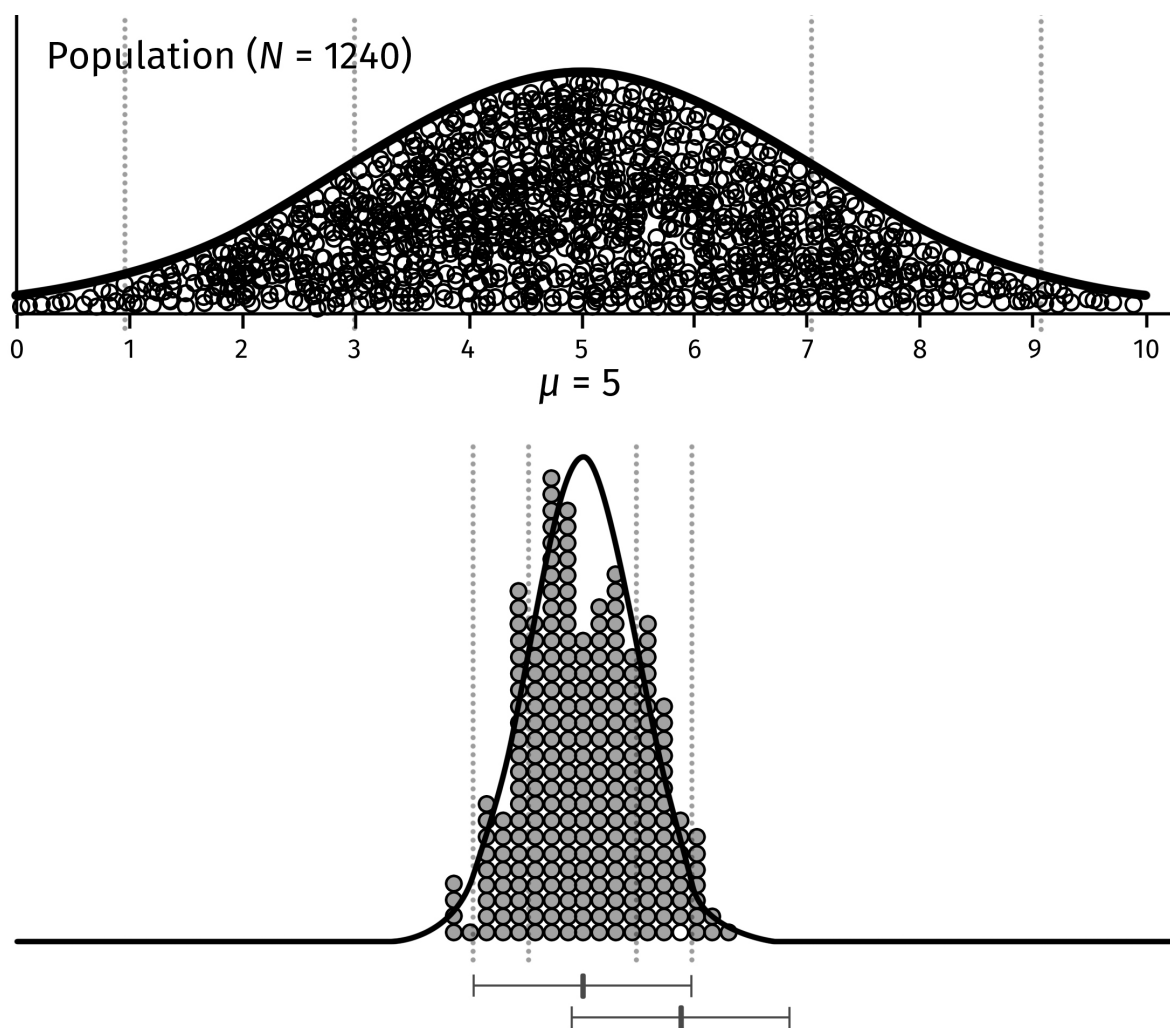


Figure 8: 254 means calculated from equally many random samples from a population

Now imagine what happens to you as a lizard researcher. You only had time to measure 20 lizards. You neither know the real mean μ from the population nor do you have many sample means. You do not know how close your mean \bar{x} is to the real mean μ . But statistical knowledge nevertheless will help you determine how good your estimation \bar{x} is. Look at Figure 8 again. At the bottom half we can see that there are sample means that have the same value as the real mean. At the very bottom of the figure there are two horizontal lines. Look at the centered one. The vertical line in the middle of the horizontal line indicates a sample mean (that is identical or at least very near to the real mean). The two ends of the line represent $-1,96SEM$ of the sample means and $+1,96SEM$ respectively. 95 % of the sample means I calculated are in this area.

Now there is one white dot in the bottom of the figure. Suppose that this is the mean you calculated from your 20 measures. If we look at what happens when we take $-1,96SEM$ and $+1,96SEM$ into account, we get the line at the very bottom of the figure.

Although our mean is rather far away from the real mean, the line still includes the real mean. Do you know how big the chances are that the real mean is inside the bar indicated by the line? It's 95%! We call this a 95% confidence interval since you can be 95% confident that the real mean is within this line.

However, as a real lizard researcher you do not have 254 means, but only one. But from the 20 data points you have, you can still calculate a 95% confidence interval. I will give you the formula for this although in reality, your computer will calculate it for you:

$$\text{95 \% confidence interval: } [\bar{x} - 1,96 \times \frac{s}{\sqrt{n}}, \bar{x} + 1,96 \times \frac{s}{\sqrt{n}}] \quad (11)$$

This formula will lead to two values in the format $[y_1, y_2]$. These two values are the ends of the lines I just showed you in the last figure. Note that this is the formula for a 95% confidence interval for the mean. You can calculate confidence intervals for measures different from the mean, but then you will need to use other formulas. We will see the 95% confidence interval again.

Excursus: Confidence Intervals and the Standard Error

The formula to calculate a 95% confidence interval includes a mean \bar{x} you calculated and $\pm 1.96SD$. As you know from Figure 8 there are two *SDs*: one for the population and one for the sampling distribution of the sample means (actually, the *SD* you really calculate is a third *SD*: the one you get from your 20 samples). I don't want to confuse you, but the standard deviation of this sampling distribution of the sample mean has its own name and you already know it! It's called the standard error (SEM). If you're interested in this I would recommend reading Cummings (2012). He also has a bunch of cool and easy to watch Youtube videos.

9. Behind the scenes: parameters (and more about distributions)

We have already heard about the normal distribution. There are, however, more distributions. We call them probability distributions. Distributions model an entire population. As you will never (or only in very rare cases) have the chance to look at all the data points that exist, you have to choose. In our case, that is in grammaticality judgments, you will have to choose some participants, some sentences, and will get some judgments. There is, however, an infinite set of judgments (the population). This infinite set has a mean

and, as you know, there will be extremes on both sides of the distribution. Take a look at Figure 2 again. A lot of judgments will lie in the middle, and only some at both ends (if we make the assumption that the judgments will follow a normal distribution, but we don't have to worry about the details now). If you ask one person for one judgment of your sentence under consideration, how likely will it be that you will get a judgment that is near the mean compared to a judgment that lies near one of the ends? Do you already see what I'm heading for? Yes, it is a question of probability!

So, besides the normal distribution, there are several other probability distributions. We don't have to care too much about them. What you have to know is what characterizes a distribution like the one in Figure 2. In other words: How can we, mathematically, describe such a curve? What we need was discovered by the English mathematician Karl Pearson. He found that there is a fixed set of numbers that describe a probability distribution:

- The mean
- The standard deviation, a number that describes the dispersion of the values around the mean
- The symmetry of the curve
- The kurtosis, how far the rare values are dispersed around the mean

Again, we don't have to care about the exact mathematical representation of those numbers in detail. What you have to know is that you can describe a probability distribution mathematically. The numbers that are used in this description are called 'parameters' (derive from Greek meaning 'almost measurements'). Pearson's discovery cannot be underestimated. He stated that science should not deal with phenomena that can be observed, but with the things behind those phenomena: You cannot observe a probability distribution. Nevertheless, behind each observable phenomenon, there is such a distribution. Again, I will cite Salsburg (2001) and recommend reading his book:

Pearson proposed that [...] observable phenomena were only random reflections. What was real was the probability distribution. The real "things" of science were not things that we could observe and hold but mathematical functions that described the randomness of what we could observe. The four parameters of a distribution are what we really want to determine in a scientific investigation. (Salsburg 2001:17)

The point is: You, as a linguist, want to know if a certain grammatical construction exists in the minds of the speakers of a language. You cannot access this directly. The same is true for the parameters of a probability distribution. They are not determinable. What you can access are judgments of sentences. You can take some sentences and get some judgments. Those are random data points drawn from a population of infinite judgments, i. e., data points from a probability distribution.

There is another important point that will be relevant later when we learn about the question of how many data points and subjects we need: One of the pioneers in the development of statistics was the English statistician William Sealy Gosset. Gosset started to work for the Guinness Brewing Company in 1899 and developed a method that helped with the measurement of yeast. Guinness did not allow its employees to publish their discoveries, so Gosset decided to publish his insights under a pen name. The pseudonym he used in many influential papers was “Student” (you may already know the Student’s t -test).

You already know this: There is an unobservable population (all the possible data points), i. e. a probability distribution that can be described via its parameters. Those parameters are not accessible. What is accessible is a random sample (your participants’ judgments). Statisticians before Gosset aka Student believed that you would need thousands of data points to calculate the parameters you want to know. Gosset, however, wanted to deal with small samples. One of his great discoveries was that you don’t need large samples. Assuming that your data on the whole follows a normal distribution you will find the approximate parameters of this distribution without knowing the exact numbers for all parameters. Instead you only need a small amount of data points.⁸

Even more important, later researchers found that Gosset’s original assumption about normal distributed data wasn’t even necessary. This means that applying Gossets methods (i. e., Student’s t -test which we will see later on) works for data that may not be normally distributed behind the scenes.

Now you are equipped with some statistical background. The next things we will take a look at is how to create a questionnaire. This is actually not a big deal and thus this section will be rather short. After this section about designing and conducting your judgment study we take a look at how to analyze and visualize the data.

⁸You really don’t need to know this, but for a better understanding you might want to know this: The normal distribution is one type of probability distribution. There are others. As I said, a probability distribution can be described mathematically by a function that is made up of four parameters. The normal distribution, however, can be described by using only two parameters, namely the mean and the standard deviation, because the symmetry and the kurtosis are fixed.

Part II

Designing a questionnaire and conducting your study

10. How to create the questionnaire: the stimuli

The most important part of your study will be your stimuli, of course. In the most simple case you want to compare two constructions A and B. Suppose construction A involves the linear order of two parts of speech XZ and construction B the linear order ZX which you hypothesize to be ill-formed. You would create let's say 5 sentences with the order XZ and 5 sentences with the linear order ZX (we will see in the next section that you actually need to create more sentences). The sentences should be minimal pairs. This means that your stimuli could look like the following:

Construction A	Construction B
Last week, Peter XZ went to the store.	Last week, Peter ZX went to the store.
This morning, Anne XZ bought a beer.	This morning, Anne ZX bought beer.
Yesterday, a man XZ came to my house.	Yesterday, a man ZX came to my house.
Tomorrow, Jun XZ will give me the present.	Tomorrow, Jun ZX will give me the present.
Next year, Laura XZ will graduate.	Next year, Laura ZX will graduate.

The examples show that there is as little variation as possible between the sentences to be judged in the two conditions. How the stimuli will look will, of course, depend on the goal of your study. In fact, there are some things in the example above you might want to avoid: While there is not much variation between the conditions there is much variation inside conditions. There is, for example, variation with respect to the verb structure, concerning definiteness, or concerning tenses in the examples above. You might want to control those things. So better just create one template sentence and just exchange the words:

Construction A	Construction B
Last week, Peter XZ visited a concert.	Last week, Peter ZX visited a concert.
This morning, Anne XZ bought a beer.	This morning, Anne ZX bought a beer.
Yesterday, Rainer XZ saw a movie.	Yesterday, Rainer ZX saw a movie.
Last fall, Jun XZ wrote an article.	Last fall, Jun ZX wrote an article.
Last year, Laura XZ found a watch.	Last year, Laura ZX found a watch.

Now there is as little variation as possible not only between conditions, but also inside the conditions. All sentences now have a definite subject, are in the past tense, contain a transitive verb, and an indefinite object. Depending on the goal of your study these may be aspects you want to control.

Now your test items are prepared (actually, we will see in the next section that we need more example sentences). Do you need more? Yes! Additionally, you want to do one of two things (or both, if the study's design allows):

- You can add some (let's say 5) grammatical sentences and some (again, 5) completely ill-formed sentences in your questionnaire. The data you obtain from those sentences can be used as anchor values against which you can interpret your actual data (and this may be very helpful). Additionally, this method may help you find out if a participant really understood the task or just randomly filled out your questionnaire.
- You can present the participants with some grammatical sentences as examples of natural sentences and some completely ungrammatical sentences as examples of unnatural sentences to your participants at the beginning of your study. This helps your participants to get a feeling of what is meant with 'natural' and 'unnatural' sentences.

Sometimes researchers include unannounced practice items to familiarize participants with the task. Thus, they include several well-formed and ill-formed sentences at the beginning of the questionnaire (Schütze & Sprouse 2013). However, this is only necessary if your participants are not used to the Likert response format (which I guess many people know).

In some cases your questionnaire might consist of many very similar sentences. In this case participants will start to like constructions that they originally didn't like because of the mere exposure effect (see below). In such a case you should include filler sentences, i. e., sentences that do not have anything to do with your study. Filler sentences are often used in experiments to prevent participants from uncovering the true purpose of the experiment (as this might influence the results). Linguistic judgments, however, often are very stable, so you do not have to distract people from the purpose of your study. However, if you are interested in a construction from which you have the feeling that reading it several times may confuse people, use fillers. If you decide to use fillers, be aware that the filler sentences can have an influence on the ratings of the actual sentences to be judged. This means that if you, for example, include only highly acceptable filler sentences, participants may be biased and judge your actual sentences better or the other

way around. I recommend following Cowart's (1997:52) advice: "The best strategy is to include a balanced list of fillers that includes approximately equal numbers of sentences at a wide range of acceptability values." I recommend using as many fillers as you have test items. See also the following helpful quote:

[Filler items] can serve at least three purposes. First, they can reduce the density of the critical comparisons across the whole experiment, reducing the chances that participants will become aware that a particular sentence type is being tested, which could trigger conscious response strategies. Second, they can be used to try to ensure that all the possible responses [...] are used about equally often. This helps to protect against scale bias, which occurs when one participant decides to use the response scale differently from other participants, such as only using one end of the scale (skew), or only using a limited range of responses (compression). (Schütze & Sprouse 2013:39)

11. Use Latin squares for counterbalancing

The procedure presented so far leads to a situation in which one and the same participant sees pairs of sentences using the same lexicalizations only differing in construction type. Especially when your constructions are very similar this can be problematic as the sentences to be rated may become too similar lexically although they differ in grammatical construction. This can lead to carryover effects, meaning that judging the lexicalization of one construction can influence the judgment of the same lexicalization of another construction (thus, on the one hand we want our items to be as similar as possible, but on the other hand, repetition of too similar items is undesirable). This, indeed, is a serious concern. A way out is to use a Latin square design which will ensure that each participant only sees one lexicalization of each condition. Latin squares are a powerful tool for counterbalancing your items.

Excursus: Counterbalancing

Empirical investigations often include unbalanced situations that need counterbalancing. Suppose, you do a forced-choice task. Participants in your study are presented with words and pseudo-words and need to decide if a stimulus they see is a word or not. To do this, they press buttons. There are two buttons in this design, a 'yes button' and a 'no button'. There are not many ways to present the buttons. One way is that the no button is on the left and the yes button is on the right of a keyboard. One problem will be that most people are

right-handed and reactions with the dominant hand are generally faster than reactions of the non-dominant hand. Thus, the design is not balanced. One idea would be to ensure that 50 % of the participants are right-handed and the other 50 % are left-handed. However, finding left-handers is more difficult than finding right-handed participants, so this is impractical.

Another solution would be to only study right-handers. In 50 % of the trials the yes button is on the right and in 50 % of the trials the yes button is on the left. This is a very simple case of counterbalancing. However, in most cases, your variables won't only have two, but more levels and counterbalancing will get more complicated. In this case, a Latin square is a powerful tool for achieving partial counterbalancing.

First, let's look at the basics of how a Latin square works and then let's look at examples. A Latin square is square—what a surprise! It has as many columns as rows. To start with a simple example, we look at a 4×4 Latin square:

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

Each letter in a Latin square represents one condition. In this example, there are 4 conditions (A, B, C, and D). Each condition appears only once in each row and in each column. Thus, condition A in the first row only appears once and condition A in the first column only appears once. Suppose you want to make the following crazy study: You want to know if participants' judgments of the acceptability of sentences get worse when they are exposed to an unpleasant odor. There are 4 levels of odor you want to study. Condition A is neutral (without odor), in condition B participants are exposed to the odor, but only a little bit, condition C is a stinky condition, and condition D is a heavily unpleasant very stinky condition. Thus the odor increases from condition A to D. If you plan to do a within-subject design each subject will judge sentences in each condition. The problem now is that if a participants rate sentences in condition A, condition B, condition C, and finally in condition D, the participant may get used to the odor. Other orders may lead to similar carryover effects, however. Being first presented with sentences in condition D, then in condition C, for example, may have a similar effect. One way out would be to counterbalance the design and study each possible combination of conditions. With 4 conditions, we get 24 possible combinations. If we would test one participant per

combination, we would need 24 participants. If our study had 6 conditions instead, there would be 720 possible combinations, however. Thus, a completely counterbalanced design often would be very effortful. The Latin square is a way out. As you can see from the Latin square above, there are not 24 different combinations of the 4 conditions, but only 4 (each row). Note that this, of course, does not mean that you only need 4 participants.

Now let's look at a linguistic example. We look at how Sprouse, Wagers & Phillips (2013) used a Latin square for studying island effects. Their study is a multifactorial design. As we haven't talked about multifactorial designs yet, let's do that first.

Excursus: Multifactorial designs

The more complex your research question gets the more complex your research design will be. Very often, you do not want to compare 2 constructions, but there are more variables you want to manipulate. If you have 2 constructions A and B there is only one independent variable that you manipulate. This variable is construction type and has 2 levels. However, you could, for example assume that judgements could be influenced by other factors as well, for example, by aspect. This variable could have, to stick with an easy example, also 2 levels, for example, you want to compare sentences in simple present ('Paul sits in the garden') with sentences in present progressive ('Paul is sitting in the garden').

You can now test construction A with and without the aspectual marker (the *-ing* form) and construction B with and without the aspectual marker. Only comparing construction A and B would be called a single-factor design as there is a single factor, namely construction type (in this case with 2 levels). Studies with more than one factor are called multifactorial (or simply factorial) designs. In our simple example we are dealing with a 2×2-design. This means that we have 2 factors with 2 levels each (if you would have 3 factors with 2 levels each we would call it a 2×2×2-design).

Multifactorial designs are designs in which several independent variables are studied at once. Two types of effects could be observed in our example study. It could be, that judgments differ based on construction and/or based on aspectual changes. We would call such effects of an independent variable main effects. Additionally, it could be that we observe effects which come into being through a combination of construction type and aspect. We would call such an effect an interaction effect.

Sprouse, Wagers & Phillips (2013) were interested island effects. A syntactic island is a configuration from which movement is banned. Their starting point is *wh*-movement out of an embedded *whether* clause. This kind of movement is banned in English as *whether* clauses are islands:

- (1) *What_i do you wonder whether John bought *t_i*?

One assumption to draw from this example could be that extraction out of an embedded clause is ill-formed in English. However, this cannot be the case as well-formed examples can be constructed with *that* clauses:

- (2) What_i do you think that John bought *t_i*?

This sounds rather easy, but we already have 2 factors/variables with 2 levels each: Is it the embedding that causes the island effect (\pm embedding) or is it the presence of *whether* (\pm *whether*)? Sprouse, Wagers & Phillips (2013) label the factors ‘structure’ and ‘gap’. The first factor has the two levels ‘island/non-island’ and the second factor has the levels ‘matrix/embedded’. By crossing the factors we get 4 possible combinations:

- | | | | |
|-----|----|---|-----------------------|
| (3) | a. | Who _i <i>t_i</i> thinks that John bought a car? | ‘non-island/matrix’ |
| | b. | What _i do you think that John bought <i>t_i</i> ? | ‘non-island/embedded’ |
| | c. | Who _i <i>t_i</i> wonders whether John bought a car? | ‘island/matrix’ |
| | d. | *What _i do you wonder whether John bought <i>t_i</i> ? | ‘island/embedded’ |

Note that the condition represented in (3a) serves as a baseline condition as it represents the unmarked levels of both factors. What the sentences in (3) illustrate is a basic 2×2-design. Let’s look at how a Latin square could be created. The 2×2-design leads to 4 different conditions. This means that we can create 4 lists using a Latin square. If you want that each participant is presented with each condition 5 times, you would need to create 5 lexicalizations sets for 4 lists (i.e., $4 \times 4 \times 5 = 80$ target sentences). An easy way to do this is using Excel or OpenOffice as described by Simone Gieselman here: <http://idiom.ucsd.edu/~simone/Tutorial.html>. Another way to do this would be to use the Python script ‘turkolizer’ described in Gibson, Piantadosi & Fedorenko (2011).

Let’s briefly look at how to achieve this in Excel or OpenOffice following Simone Gieselman’s tutorial. We basically need 3 columns: A column for the list numbers, a column assigning each example a number (this makes sense if you feed your data in

another program, for example, if you use WebExp), and a column for the actual examples. The results will look like the following (I added a condition column for a better overview):⁹

	A	B	C	D	E	F	G	H	I
1	List	Number	Example	Condition					
2		1 1a	Who thinks that John bought a car?	non-island/matrix					
3		2 1b	What do you think that John bought?	non-island/embedded					
4		3 1c	Who wonders whether John bought a car?	island/matrix					
5		4 1d	What do you wonder whether John bought?	island/embedded					
6		2 2a	Who said that Paula stole a laptop?	non-island/matrix					
7		3 2b	What do you believe Paula stole?	non-island/embedded					
8		4 2c	Who wonders whether John stole a laptop?	island/matrix					
9		1 2d	What do you wonder whether Paula stole?	island/embedded					
10		3 3a	Who knows that Otto sold his house?	non-island/matrix					
11		4 3b	What do you think that Otto sold?	non-island/embedded					
12		1 3c	Who knows whether Otto sold his house?	island/matrix					
13		2 3d	What do you know whether Otto sold?	island/embedded					
14		4 4a	Who thinks that Maria bought a computer?	non-island/matrix					
15		1 4b	What do you think that Maria bought?	non-island/embedded					
16		2 4c	Who wonders whether Maria bought a computer?	island/matrix					
17		3 4d	What do you wonder whether Maria bought?	island/embedded					
18		1 5a	Who said that Gökce smokes cigarettes?	non-island/matrix					
19		2 5b	What do you think that Gökce smokes?	non-island/embedded					
20		3 5c	Who asks whether Gökce smokes cigarettes?	island/matrix					
21		4 5d	What do you wonder whether Gökce smokes?	island/embedded					
22		2 6a	Who thinks that Mica saw a cat?	non-island/matrix					
23		3 6b	What do you think that Mica saw?	non-island/embedded					
24		4 6c	Who wonders whether Mica saw a cat?	island/matrix					
25		1 6d	What do you wonder whether Mica saw?	island/embedded					
26		3 7a	Who said that Gina killed a bug?	non-island/matrix					
27		4 7b	What do you believe Gina killed?	non-island/embedded					
28		1 7c	Who wonders whether Gina killed a bug?	island/matrix					
29		2 7d	What do you wonder whether Gina killed?	island/embedded					
30		4 8a	Who knows that Oskar hit a car?	non-island/matrix					
31		1 8b	What do you think that Oskar hit?	non-island/embedded					
32		2 8c	Who knows whether Oskar hit a car?	island/matrix					
33		3 8d	What do you know whether Otto hit?	island/embedded					
34		1 9a	Who thinks that Algina lost her wallet?	non-island/matrix					
35		2 9b	What do you think that Algina lost?	non-island/embedded					
36		3 9c	Who wonders whether Algina lost her wallet?	island/matrix					
37		4 9d	What do you wonder whether Algina lost?	island/embedded					
38		2 10a	Who said that Alexandra found a watch?	non-island/matrix					
39		3 10b	What do you think that Alexandra found?	non-island/embedded					
40		4 10c	Who asks whether Alexandra found a watch?	island/matrix					
41		1 10d	What do you wonder whether Alexandra found?	island/embedded					

The list column is structured as follows: 1, 2, 3, 4; 2, 3, 4, 1; 3, 4, 1, 2; 4, 1, 2, 3; 1, 2, 3, 4; ... This is the pattern we saw in the Latin square. Now you can select the first column and click on the ‘data’ menu and click ‘sort’. This will sort the items by lists and the result will look like this:

⁹There are not the real examples used in the study, but some quickly made-up examples.

Externe Daten abrufen		Verbindungen		Sortieren und Filtern		Datentools			
C84		fx							
	A	B	C	D	E	F	G	H	I
1	List	Number	Example	Condition					
2		1 1a	Who thinks that John bought a car?	non-island/matrix					
3		1 2d	What do you wonder whether Paula stole?	island/embedded					
4		1 3c	Who knows whether Otto sold his house?	island/matrix					
5		1 4b	What do you think that Maria bought?	non-island/embedded					
6		1 5a	Who said that Gökcé smokes cigarettes?	non-island/matrix					
7		1 6d	What do you wonder whether Mica saw?	island/embedded					
8		1 7c	Who wonders whether Gina killed a bug?	island/matrix					
9		1 8b	What do you think that Oskar hit?	non-island/embedded					
10		1 9a	Who thinks that Algina lost her wallet?	non-island/matrix					
11		1 10d	What do you wonder whether Alexandra found?	island/embedded					
12		1 11c	Who wonders whether Olli bought a skateboard?	island/matrix					
13		1 12b	What do you believe Evanesca wrote?	non-island/embedded					
14		1 13a	Who knows that Uri owns a house?	non-island/matrix					
15		1 14d	What do you wonder whether Maria caught?	island/embedded					
16		1 15c	Who asks whether Ira wrote a novel?	island/matrix					
17		1 16b	What do you think that Lia read?	non-island/embedded					
18		1 17a	Who thinks that Olga sold her car?	non-island/matrix					
19		1 18d	What do you wonder whether Paula saw?	island/embedded					
20		1 19c	Who knows whether Elias touched an elephant?	island/matrix					
21		1 20b	What do you think that Philip built?	non-island/embedded					
22		2 1b	What do you think that John bought?	non-island/embedded					
23		2 2a	Who said that Paula stole a laptop?	non-island/matrix					
24		2 3d	What do you know whether Otto sold?	island/embedded					
25		2 4c	Who wonders whether Maria bought a computer?	island/matrix					
26		2 5b	What do you think that Gökcé smokes?	non-island/embedded					
27		2 6a	Who thinks that Mica saw a cat?	non-island/matrix					
28		2 7d	What do you wonder whether Gina killed?	island/embedded					
29		2 8c	Who knows whether Oskar hit a car?	island/matrix					
30		2 9b	What do you think that Algina lost?	non-island/embedded					
31		2 10a	Who said that Alexandra found a watch?	non-island/matrix					
32		2 11d	What do you wonder whether Olli bought?	island/embedded					
33		2 12c	Who wonders whether Evanesca wrote a letter?	island/matrix					
34		2 13b	What do you think that Uri owns?	non-island/embedded					
35		2 14a	Who thinks that Maria caught a salmon?	non-island/matrix					
36		2 15d	What do you wonder whether Ira wrote?	island/embedded					
37		2 16c	Who wonders whether Lia read the letter?	island/matrix					
38		2 17b	What do you think that Olga sold?	non-island/embedded					
39		2 18a	Who said that Paula saw a lion?	non-island/matrix					
40		2 19d	What do you know whether Elias touched	island/embedded					

For a better overview, it makes sense to open a extra worksheet for each list (you see the worksheets on the bottom of your window). So copy each list in a worksheet:

Access		Web		Text		Quellen		Verbindungen		aktualisieren		Verknüpfungen bearbeiten		Erweitert	
Externe Daten abrufen		Verbindungen		Sortieren und Filtern											
H17		fx													
	A	B	C	D	E	F									
1	List	Number	Example	Condition	Random Number										
2		1 1a	Who thinks that John bought a car?	non-island/matrix											
3		1 2d	What do you wonder whether Paula stole?	island/embedded											
4		1 3c	Who knows whether Otto sold his house?	island/matrix											
5		1 4b	What do you think that Maria bought?	non-island/embedded											
6		1 5a	Who said that Gökce smokes cigarettes?	non-island/matrix											
7		1 6d	What do you wonder whether Mica saw?	island/embedded											
8		1 7c	Who wonders whether Gina killed a bug?	island/matrix											
9		1 8b	What do you think that Oskar hit?	non-island/embedded											
10		1 9a	Who thinks that Algina lost her wallet?	non-island/matrix											
11		1 10d	What do you wonder whether Alexandra found?	island/embedded											
12		1 11c	Who wonders whether Olli bought a skateboard?	island/matrix											
13		1 12b	What do you believe Evanesca wrote?	non-island/embedded											
14		1 13a	Who knows that Uri owns a house?	non-island/matrix											
15		1 14d	What do you wonder whether Maria caught?	island/embedded											
16		1 15c	Who asks whether Ira wrote a novel?	island/matrix											
17		1 16b	What do you think that Lia read?	non-island/embedded											
18		1 17a	Who thinks that Olga sold her car?	non-island/matrix											
19		1 18d	What do you wonder whether Paula saw?	island/embedded											
20		1 19c	Who knows whether Elias touched an elephant?	island/matrix											
21		1 20b	What do you think that Philip built?	non-island/embedded											
22															

Now you can create another worksheet for your fillers and copy the fillers underneath each list (depending on what you will do exactly later on it can be useful to label the fillers to identify them later). If you want to randomize your items right now you can do this, although there are programs that will do the randomization when the stimuli are presented. To randomize the order manually, add a new column labeled 'random number' and type in '=RAND'. This will create a random number. Fill all rows with random numbers next. Schematically, this will look like this:

	A	B	C	D	E
1	List	Number	Example	Condition	Random Number
2	1	1a	Who thinks that John bought a car?	non-island/matrix	0.288482398
3	1	1d	What do you wonder whether Paula stole?	island/embedded	0.818617764
4	1	1c	Who knows whether Otto sold his house?	island/matrix	0.934192415
5	1	14b	What do you think that Maria bought?	non-island/embedded	0.208948144
6	1	15a	Who said that Gökce smokes cigarettes?	non-island/matrix	0.911473266
7	1	16d	What do you wonder whether Mica saw?	island/embedded	0.213942232
8	1	17c	Who wonders whether Gina killed a bug?	island/matrix	0.967341615
9	1	18b	What do you think that Oskar hit?	non-island/embedded	0.952295045
10	1	19a	Who thinks that Algina lost her wallet?	non-island/matrix	0.093027651
11	1	10d	What do you wonder whether Alexandra found?	island/embedded	0.030145354
12	1	11c	Who wonders whether Olli bought a skateboard?	island/matrix	0.305384452
13	1	12b	What do you believe Evanesca wrote?	non-island/embedded	0.090199236
14	1	13a	Who knows that Uri owns a house?	non-island/matrix	0.97173766
15	1	14d	What do you wonder whether Maria caught?	island/embedded	0.513558136
16	1	15c	Who asks whether Ira wrote a novel?	island/matrix	0.017854827
17	1	16b	What do you think that Lia read?	non-island/embedded	0.007180309
18	1	17a	Who thinks that Olga sold her car?	non-island/matrix	0.432826295
19	1	18d	What do you wonder whether Paula saw?	island/embedded	0.716456479
20	1	19c	Who knows whether Elias touched an elephant?	island/matrix	0.370036646
21	1	20b	What do you think that Philip built?	non-island/embedded	0.278633886
22	1	fil1	FILLER1	filler	0.975877774
23	1	fil2	FILLER2	filler	0.603134818
24	1	fil3	FILLER3	filler	0.469548421
25	1	fil4	FILLER4	filler	0.59963916
26	1	fil5	FILLER5	filler	0.170148605
27	1	fil6	FILLER6	filler	0.713191866
28	1	fil7	FILLER7	filler	0.313303425
29	1	fil8	FILLER8	filler	0.487092742
30	1	fil9	FILLER9	filler	0.124442706
31	1	fil10	FILLER10	filler	0.530418216
32	1	fil11	FILLER11	filler	0.97281916
33	1	fil12	FILLER12	filler	0.53536787
34	1	fil13	FILLER13	filler	0.966813311
35	1	fil14	FILLER14	filler	0.975033956
36	1	fil15	FILLER15	filler	0.314082093
37	1	fil16	FILLER16	filler	0.971854024
38	1	fil17	FILLER17	filler	0.00011158
39	1	fil18	FILLER18	filler	0.391344928
40	1	fil19	FILLER19	filler	0.394913433
41	1	fil20	FILLER20	filler	0.882893801

If you now select the ‘random number’ column you can sort the items again. This will result in a random order:

Zwischenablage		Schriftart		Ausrichtung	
E1		fx		Random Number	
	A	B	C	D	E
1	List	Number	Example	Condition	Random Number
2	1	fil17	FILLER17	filler	0.343344544
3	1	16b	What do you think that Lia read?	non-island/embedded	0.381347766
4	1	15c	Who asks whether Ira wrote a novel?	island/matrix	0.766934234
5	1	10d	What do you wonder whether Alexandra found?	island/embedded	0.299537195
6	1	12b	What do you believe Evanesca wrote?	non-island/embedded	0.254408835
7	1	9a	Who thinks that Algina lost her wallet?	non-island/matrix	0.687488902
8	1	fil9	FILLER9	filler	0.512646459
9	1	fil5	FILLER5	filler	0.310737892
10	1	4b	What do you think that Maria bought?	non-island/embedded	0.136055208
11	1	6d	What do you wonder whether Mica saw?	island/embedded	0.420188677
12	1	20b	What do you think that Philip built?	non-island/embedded	0.049727574
13	1	1a	Who thinks that John bought a car?	non-island/matrix	0.46433735
14	1	11c	Who wonders whether Olli bought a skateboard?	island/matrix	0.863210948
15	1	fil7	FILLER7	filler	0.146369799
16	1	fil15	FILLER15	filler	0.198279305
17	1	19c	Who knows whether Elias touched an elephant?	island/matrix	0.655526451
18	1	fil18	FILLER18	filler	0.535562354
19	1	fil19	FILLER19	filler	0.718124583
20	1	17a	Who thinks that Olga sold her car?	non-island/matrix	0.24592897
21	1	fil3	FILLER3	filler	0.743376057
22	1	fil8	FILLER8	filler	0.325198589
23	1	14d	What do you wonder whether Maria caught?	island/embedded	0.527715008
24	1	fil10	FILLER10	filler	0.652944985
25	1	fil12	FILLER12	filler	0.859128309
26	1	fil4	FILLER4	filler	0.698499988
27	1	fil2	FILLER2	filler	0.153327972
28	1	fil6	FILLER6	filler	0.730137713
29	1	18d	What do you wonder whether Paula saw?	island/embedded	0.665061365
30	1	2d	What do you wonder whether Paula stole?	island/embedded	0.615403619
31	1	fil20	FILLER20	filler	0.548838356
32	1	5a	Who said that Gökce smokes cigarettes?	non-island/matrix	0.601075251
33	1	3c	Who knows whether Otto sold his house?	island/matrix	0.138177776
34	1	8b	What do you think that Oskar hit?	non-island/embedded	0.791937521
35	1	fil13	FILLER13	filler	0.660493885
36	1	7c	Who wonders whether Gina killed a bug?	island/matrix	0.852558738
37	1	13a	Who knows that Uri owns a house?	non-island/matrix	0.719547373
38	1	fil16	FILLER16	filler	0.911302886
39	1	fil11	FILLER11	filler	0.017921505

Finally, manual changes can be made if necessary, for example, if you want each list to start with a filler. Note that we now can randomly assign participants to lists which makes our judgment task more similar to a real experiment.

Latin squares can be easily used for more complex tasks. Sprouse, Wagers & Phillips (2013) original study, for example, was more complicated than presented so far: Besides *whether* islands, there are other island types. They tested 4 different island types, namely *whether* islands, Complex NP island, Subject islands, and Adjunct islands. These 4 island types represent the 4 main conditions they used to create a Latin square resulting in 4 lists. The 4 main conditions with the 2×2 manipulation in each condition look like this:

- Condition A: whether islands
 - non-island/matrix
 - non-island/embedded
 - island/matrix
 - island/embedded
- Condition B: Complex NP islands
 - non-island/matrix
 - non-island/embedded
 - island/matrix
 - island/embedded
- Condition C: Subject islands
 - non-island/matrix
 - non-island/embedded
 - island/matrix
 - island/embedded
- Condition D: Adjunct islands
 - non-island/matrix
 - non-island/embedded
 - island/matrix
 - island/embedded

The overview shows that there is a total of 16 critical conditions. I will end this section by a brief summary of how they created their lists:

Four island types (*whether* islands, Complex NP island, Subject islands, and Adjunct islands) were tested, each using a 2×2 manipulation of extraction and structural environment [...], yielding a total of sixteen critical conditions. Eight additional sentence types were included to add some variety to the materials, for a total of twenty-four sentence types. Sixteen lexicalizations of each sentence type were created, and distributed among four lists using a Latin Square procedure. This meant that each list consisted of four tokens

per sentence type, for a total of ninety-six items. Two orders for each of the four lists were created by pseudo-randomizing the items such that related sentence types were never presented successively. This resulted in eight different surveys. (Sprouse, Wagers & Phillips 2013:34)

12. The instructions

You usually do not want to ask linguists to participate in your study. Linguists have theories and hypotheses in mind about a lot of linguistic phenomena—a fact that could have an influence on their judgments (see e.g., Bolinger 1968; Carden 1976; Schütze 2016).¹⁰ Thus, I recommend consulting linguistic laymen (or at least undergraduate students). With this, the next problem arises: You can ask a non-linguist a question like ‘How grammatical is this sentence?’ But what will happen? The term ‘grammaticality’ means something completely different to people who have not studied linguistics: They usually think of prescriptive grammars—and that’s exactly what you don’t want!

In practice, this means that you have to explain the task as exactly as possible. The best option clearly is to avoid terms like ‘grammar’ or ‘grammatical’ completely and just stick with something like ‘natural’. A possible phrasing is: “In the following you will be presented with a number of sentences. We are interested in how people actually use language. You will be asked to rate each sentence. Please rate each sentence from 1 meaning ‘totally unnatural’ to 7 ‘absolutely natural’. A sentence is considered to be natural a) when you would use the sentence in everyday life or b) you would expect your neighbor or a friend to use a sentence like that.” Of course, the exact wording depends on the subject of your study. See also the following helpful paragraph from Schütze & Sprouse (2013:36):

While there is no standard way of wording the instructions for a judgment experiment, there is general agreement that we want to convey to speakers that certain aspects of sentences are not of interest to us and should not factor into their responses. These include violations of prescriptive grammar rules, the likelihood that the sentence would actually be uttered in real life, and the truth or plausibility of its content. [...] We also want to avoid the question of the sentence being understandable, since uncontroversially ungrammatical sentences are often perfectly comprehensible (e.g., What did he wanted?). It is common to instruct participants to imagine that the sentences were being

¹⁰Carden (1976:103) even claims: “The linguist’s own intuitions are plainly untrustworthy.” I also like the way Gibson, Piantadosi & Fedorenko (2013). Their motto is “Expert intuitions are not data” and view judgments by linguists as being predictors and not data.

spoken by a friend, and ask whether the sentences would make them sound like a native speaker of their language.

This is a lot of stuff! In practice, you want to keep your instructions short so it makes sense to work with bullet points like:

- Please do not judge the content or plausibility of the sentences.
- This study is not about rules you find in a grammar book, but about how language is actually used in everyday life.
- Imagine the sentences were used by a neighbor or friend of yours. Would you consider her a native speaker (then the sentence is natural) or does she sound somehow ‘strange’?

Excursus: More on Phrasing

How your actual instructions will look depends on your task. Some phenomena are easier to rate than others. This is especially true for spoken language phenomena. I will illustrate this with two phenomena that are very frequent in spoken German, but only one of them is hard to rate.

There are plenty of so called ‘modal particles’ in German. They seldom occur in writing, but are very frequent in spoken utterances. However, it is no problem for participants to rate the acceptability of a modal particle in a sentence even when the sentence is written. Another phenomenon is a change in word order: As you may know, (surface) word order in German depends on the type of clause. German exhibits SVO in main and SOV in subordinate clauses. The latter word order is, for example, used with subordinate clauses introduced by the complementizer *weil* ‘because’. However, in spoken language, people often use SVO order in this case. People usually are not aware they do this and will—in many cases—reject a sentence, especially a written sentence, as being acceptable.

What I want to say is: Think carefully about your topic. Ask some linguistically trained people for judgments. Try a questionnaire with some laymen and ask them for feedback and look at the results (i. e., run a pilot). Then decide if you have to change your instructions. You could also think of presenting audio instead of written stimuli. Or you could consider giving some audio examples at the beginning of your study.

13. Procedure

One of the most important things you have to do when running your study is to randomize the order of presentation of your stimuli (the sentences to be judged). We have already seen in Section 11 a case of randomization inside the lists created by using a Latin square. But why do randomization? This has to be done mainly because of the fact that the order in which items are presented can influence the way you perceive them (the most famous of these effects may be the serial position effect: the first and last items on a list are remembered better than items in the middle). There is also empirical evidence that order influences acceptability ratings (Greenbaum 1973, 1976; Greenbaum & Quirk 1970).

Additionally, think of what would happen if participants saw five sentences with a construction they normally would not judge acceptable in a row: Chances are not too small that they would become accustomed with the construction and perhaps would like it more when reading the fifth sentence than when reading the first. This is by no means unlikely and this effect even has a name. It is called mere-exposure effect (Zajonc 1968) and there is also evidence that repetition of grammatical constructions alters judgments (e. g., Nagata 1987a, 1987b, 1988). Of course, with randomization you will not get rid of this effect, but it will become less strong. The mere-exposure effect (also called familiarity principle) is, by the way, also the reason why you should not test trained linguists. They might know the phenomenon you're looking at and may be biased!

There are many more reasons to randomize the order of items. People may pay more attention in the beginning of the questionnaire and get tired at the end. People may see patterns in your sentences and be influenced by them (even if the patterns are not really there).

Additional Important Tips for the Procedure

- You need to know a lot of things about your subjects. This mainly depends on the purpose of your study, of course, but standard questions are questions about their gender, about their age, about their native and additional languages that they speak and whether or not they have language impairments. And always keep the data protection laws in your country in mind!
- If you want to know something about spoken language it can, in many cases, be useful to present participants with audio stimuli. They should for reasons of consistency be recorded so every participant hears exactly the same sentence.
- In some cases, especially when you're doing research on a language or dialect you

do not speak fluently, it can be of additional value to ask participants to write down their own versions of the sentences to be judged. The feedback gathered in this way can give you valuable information for future studies. Additionally, you may want to have a native speaker check your stimuli!

14. Software tips

- Google Forms: You can easily use Google Forms for collecting responses. Randomization is easily possible. Responses are saved in a structured spreadsheet you can download.
- Ibex and Ibex Farm: Ibex is a software for self-paced reading tasks and (speeded) acceptability rating studies. Ibex farm provides free hosting so you can run online studies (<http://spellout.net/ibexfarm/>).
- WebExp: a system to perform online experiments. It is also suitable to perform different types of judgment tasks and randomization is, of course, possible.
- MiniJudge: A software for small-scale judgment tasks.
- Amazon Mechanical Turk: To collect large amounts of judgments you can hire people via Mechanical Turk; of course, you have to pay them (see also Sprouse 2011b and for practical matters also Gibson, Piantadosi & Fedorenko 2011 and the website <http://tedlab.mit.edu/software/>).
- PsychoPy: An open source software for psychological experiments written in Python. I really like PsychoPy a lot because it's very flexible. You can easily built a questionnaire with PsychoPy although it's much more powerful. You can also use PsychoPy for online studies (using Pavlovia which is simply a platform for PsychoPy experiments).

Part III

Analyzing and visualizing your results

15. Analyzing the results

“Do not trust any p value.”

15.1 Descriptive statistics and more on Likert scales

The first thing you want to do is to describe your data. A very useful measure to begin with are mean ratings. As you have already learned, however, it is not possible to calculate a mean from our values since we used a Likert item task to get our data. The ‘distance’ between the values of a Likert item is not clearly defined and thus we cannot calculate a mean. This is because what we get from this kind of task are ordinal data.

And now comes the fun part: We can simply ignore this and just calculate the mean. Stevens’ typology of data is often very useful, especially for didactic reasons. The theory of measurement, however, is a purely mathematical theory, what we are doing is empirical research. And empirical research has shown that it is meaningful to calculate a mean out of Likert items. And even Stevens (1951:26) himself noted:

As a matter of fact, most of the scales used widely and effectively by psychologists are ordinal scales. In the strictest propriety the ordinary statistics involving means and standard deviations ought not to be used with these scales [...] On the other hand, [...] there can be invoked a kind of pragmatic sanction: in numerous instances it leads to fruitful results.

And this is not only true in this case, but in others as well:

Experience has shown in a wide range of situations that the application of proscribed statistics to data can yield results that are scientifically meaningful, useful in making decisions, and valuable as a basis for further research. (Velleman & Wilkinson 1993:68)

In fact, this is not only true for the calculation of means, but for a whole set of statistics that are called parametric tests. Parametric tests are tests with a specific set of assumptions. Typical applications of parametric tests (like the *t*-test) involve normally distributed interval (non-ordinal) data. Thus, our Likert data should not be analyzed with parametric tests. However, empirical studies, including simulation studies, have shown repeatedly that parametric tests are so robust against violations of their assumptions that they produce meaningful results for ordinal data. Thus, we will use parametric tests to analyze our data (for more information see, for example, de Winter & Dodou 2010; Norman 2010; Endresen & Janda 2017; see also Pell 2005).

15.2 Testing for differences

Suppose you want to know if there is a difference in the acceptability of two constructions. You tested both constructions and received ratings of five sentences per construction by 25 participants, i. e., each participant rated ten sentences. You will get two means (see Section 16 for details on how to do this). Let's say, again I'm making numbers up, the mean of construction A is 3.8 and the mean of construction B is 4.2. The question you have is if people like construction B better or if the numbers only differ because you drew a random sample.

You will only (really) understand how a statistical test works if you understand the question you have. A mean x you calculate is a measure of a parameter μ and not the parameter itself. It is, given your data, the best guess you have, where the real mean μ of your underlying probability distribution is. Now you have two means x_1 and x_2 . The question you have is the following: Are there two probability distributions with two different μ s? Is this the reason you got x_1 and x_2 ? The more the two values differ the more likely this may become, but you don't know! Another possibility could be that you have, because chances are actually quite high when drawing a random sample, two measures of the same parameter μ . If you want to know if there is a statistical difference between two means, what you really want to know is if they belong to one and the same distribution or to two different distributions. This is what a statistical test does!

We have calculated two means. This is our statistics (the first step of it). We want to know something about the population. The question a statistical test like the one we are interested in now tries to answer is how likely it is that the two means come from one and the same population. Our two means are x_1 and x_2 . Are they just two measures for one μ ? Or are they two measures for two μ s? I have depicted the two possibilities in Figure 9. On the left you see the possibility where your two means are just estimates for the same parameter (i. e., x_1 and x_2 are from the same population). On the right the other possibility is depicted. Here your two means are two estimates. One for μ_1 and one for μ_2 (i. e., x_1 and x_2 are from two different populations).

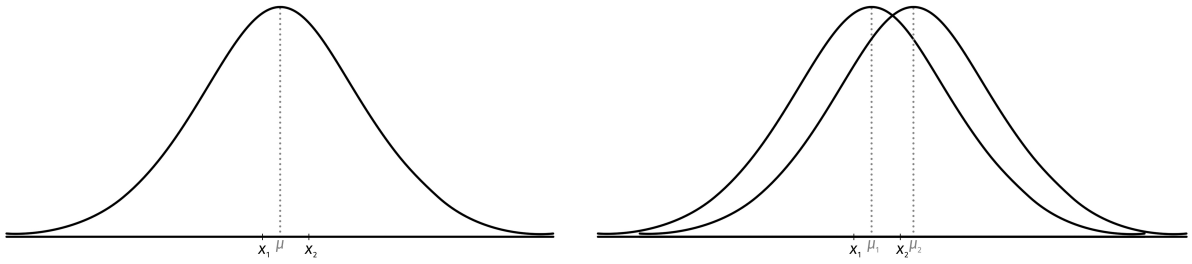


Figure 9: If you have calculated two means, the question arises if there is a statistical difference between the two groups. To decide if this is the case or not a statistical test can help. Based on the data you have such a test can help you to decide how likely it is that the two means x_1 and x_2 are point measures for one μ (the parameter you don't know). The other option would be that x_1 and x_2 do not differ by chance (because you took random samples), but actually are measures for two different μ s that belong to two different populations.

There is one very, very important thing you need to understand before we go on, because what I just told you was very sloppy. I said that a statistical test tries to find out, if x_1 and x_2 stem from the same population or not. But a statistical test cannot really do this as discussed in the introduction. How can the test know the population? It's just impossible! What the test does is that it takes your data as an input and calculates a probability. I will stress this even more later on, but keep in mind that such a test does not calculate how likely it is that x_1 and x_2 stem from one population or not. It takes your data and tells you how likely it is that you will get data like this when there is only one distribution/when there are two populations.

The null hypothesis, also called H_0 , we have is $\mu_1 = \mu_2$. So we assume that the two measures we have are only different because of random sampling and that they really come from the same underlying distribution. We can use different wordings that all mean the same:

- H_0 : $\mu_1 = \mu_2$
- H_0 : The means we calculated from our samples may be different, but this difference only came about by chance.
- H_0 : The means we calculated belong to the same population.

Usually you are interested in showing that your two means are actually different. But in science we cannot really prove things. What we can do instead, is to disprove things.

So we can disprove (or try to disprove) our H_0 . That's why significance testing is also called 'null hypothesis significance testing' or NHST for short. NHST is a kind of yes/no game. You want a definite answer. This answer comes about in form of a p -value (p stands for probability). As this is only a single value and as the world is a really complex thing that cannot be captured in form of yes/no decisions you always should report dispersion measures as well (we will see how this is done later). There are many misconceptions about NHST and about p -values—even among experienced researchers. As it is very important to understand what a p -value tells you (and what it doesn't tell you), I will devote a whole subsection to them.

15.3 Understanding p -values

A p -value is defined as the probability that you will obtain the data you have—or more extreme data—given that your null hypothesis is true. We can rephrase this as the probability that, when the null hypothesis is true, your mean differences are as they are (or of greater magnitude than the results you actually got). This doesn't sound very exciting, of course. But compare this definition to the following, wrong definition (I crossed it out as it is wrong): ~~The p -value is the probability that the null hypothesis is true given your data.~~ This sounds very similar, but it is nonsense! Think why: We really, really want to know the probability of the null hypothesis being true (or false). But that's a thing we will never ever be able to find out! We just cannot see the populations! And actually, the probability of something to be true is either 1 or 0. It's either true or false. There is nothing in between.

The only thing we can look at is our data. What we can calculate from our data is the probability of obtaining the results we obtained (or even more extreme results) given that H_0 is true. So what NHST does is to assume that H_0 is true. This means in our case, that we assume that our measures are estimators of the same population ($\mu_1 = \mu_2$).

What researchers usually want is a p -value as small as possible. The smaller the p -value is, the higher is the probability that we obtained the observed results given that H_0 is true. But when is a p -value small enough? There is no definite answer to this question and there are different traditions of setting a threshold:

- $p < 0.05$, often tagged with a star: *
- $p < 0.01$, often tagged with two stars: **
- $p < 0.001$, often tagged with three stars: ***

The significance level indicates the probability of an observed event to occur given that the null hypothesis is true. This means that a significance level (also called: alpha level/ α level) of 0.05 says that the probability of obtaining the observed results given that H_0 is true is 5 %. An α of 0.05 is very common. In practical terms this means that you obtain two means \bar{x}_1 and \bar{x}_2 and you do a statistical test that gives you a p -value. If you have set your alpha level at 0.05 then you would reject H_0 when the test gives you a p -value that is smaller than 0.05. Then you say that you have a statistically significant result and that it seems to be the case that your values are estimates of two different populations.¹¹ As you reject your null hypothesis, you accept your alternative hypothesis, namely, that there is a difference between your conditions (i. e., your constructs).

This kind of a yes/no decision that is based on alpha levels is called ‘Fisher’s disjunction’ named after the famous statistician Ronald Aylmer Fisher. In fact, many modern statistical methods go back to Fisher. In Fisher’s view, a small p -value meant that “either an exceptionally rare chance has occurred, or the theory [your null hypothesis] is not true” (Fisher 1959:39). Note that Fisher never actually made a statement about how to generally interpret a p -value or to set your alpha level at 0.05. But let’s look at an example before we go on:

Suppose we want to know if a newly developed medication is better than an old one (or if one grammatical construction is accepted more than another one; drugs are simply more illustrative). Suppose additionally, that the new drug in fact is not better than the old one. However, in reality we do not know this. You want to be sure that the new drug really is better before you start advertising it so you hire 20 different research groups that are supposed to compare the new medication to the conventional one. All teams do exactly the same comparison.

If you set an alpha level of 0.05 one team should (on average) reject the null hypothesis and come to the conclusion that one drug is better than the other. The other 19 teams would rightly come to the conclusion that there is no difference, because they obtained a p -value greater than 0.05 when comparing the results between the two groups (old drug vs. new drug). This is because in 5 % of the cases, i. e., in 1 out of 20 cases, you would wrongly reject H_0 and therefore accept H_1 with $\alpha \leq 0.05$. Note that we can only say this if we set our alpha level before we started the study! This means that you should always set your alpha level at the beginning of your study. There are many more things to say. For example, you should always be aware of the fact that, in many cases, you should do several studies to see if your results are just one case out of 20. Or that a very, very small

¹¹ Always try to use a clear language. You do not speak of ~~proving~~ something, because this is something science cannot do. You always obtain probabilities, so you do not write that there actually are two different populations, you still don’t know!

p -value is not better than 0.049 in this case. But the main point of the next subsection will be: Why not just set the alpha level at 0.0000000001 or even smaller?


Excursus: The Neyman-Pearson Tradition

Many statistical methods are based on the work of R. A. Fisher, as I already said. While most statistics textbooks will tell you that you specify a null and an alternative hypothesis, Fisher's system only used one hypothesis:

Under Fisher Hypothesis Testing (FHT), there is only one hypothesis under consideration: the theoretically uninteresting hypothesis called null hypothesis [abbreviated H_0], which for syntax is very often the claim that there is no difference in acceptability between two (or more) sentence types. Statistical tests in FHT assume that H_0 is true, and return the probability of obtaining the observed experimental result, or a result that is more extreme, under this assumption. (Sprouse & Almeida 2017:4)

In Fisher's view, there are three possibilities to interpret a p -value. Either your p -value is very small, and there should be an effect (he considered a p -value smaller than 0.01 as significant), or you obtain a p -value that is between 0.01 and 0.10. Then, in general, more experimentation is needed. The third possibility is that the value is greater than 0.10. Then the chances are really high that there is no effect. That means that Fisher's idea of significance testing was not to get a yes/no answer, but rather to get a measure of how strong the evidence for an effect is. In other words: Fisher's idea was that the smaller the p -value, the stronger the effect.

There is, however, not only one tradition of interpreting p -values, but several. A very influential one is called Neyman-Pearson tradition. It goes back to Jerzy Neyman and Egmont Pearson who also coined the terms 'null hypothesis' and 'alternative hypothesis' and who are the founders of the idea of errors types (see the next subsection). They defended the idea of setting the significance level in advance (i. e., before running an experiment). In their view, p -values only work in the long run. This means that one experiment is simply not enough: You first set your significance level, let's say 0.05. This means that when your result in the end is $p < 0.05$ and there really is an effect, you will miss this effect in only 5 % of the cases. In their view, smaller p -values are not stronger evidence. People following this tradition usually do not report exact p -values, but just state something like "We obtained $p < 0.05$ ".



There is still a big debate about the interpretation of p -values. And although you usually will get very clear results in grammaticality judgments (we talk about this later), you should be aware of the dangers of misinterpreting p -values. I recommend interpreting smaller p -values not as stronger evidence for an effect (i.e. I recommend following the Neyman-Pearson tradition). Nevertheless, it can be useful for your readers to report exact p -values. Following the Neyman-Pearson tradition also means that a large p -value is not evidence that there is no effect. It means that more experimentation is needed.

15.4 Type I and type II errors

There are two types of errors you can make when doing NHST. I will illustrate these errors by means of a story that you might have heard before: There once was a boy that often made fun of the people who lived in his village. To do this, he shouted: “The wolf is coming!,” so people would get afraid and would hide in their homes even though there actually was no wolf coming. Of course, at some point people stopped believing the boy’s warnings and stopped hiding as they realized that there was no wolf. On one day, however, the wolf really did come into the village and the boy saw him. He ran through the streets and shouted: “The wolf is coming! The wolf is coming!,” but nobody believed him.

At the beginning, people believed that there was a wolf, but in fact there was no wolf. Then, people believed there was no wolf, but in fact there was a wolf. These are the two errors of statistics: Believing that there is an effect where in fact there is no effect (type I error) and believing that there is no effect where in fact there is an effect (type II error). A type I error means to reject H_0 when it is actually true. A type II error means to accept H_0 when it is actually false.

Type I errors are also called α errors (alpha errors). The type I error rate is the probability of rejecting H_0 when it is true. This probability is assigned the Greek letter α (also called the significance level). You can set α in advance. If you set your alpha level at 0.05 the probability of committing a type I error is at 5 % (remember the 20 research teams and the new drug). If we call this probability α , the probability of committing a type II error is $1 - \alpha$. If we call the probability of committing a type II error β , then the probability of correctly rejecting the wrong H_0 is $1 - \beta$. Type I and type II errors correlate: the smaller the probability of committing a type I error, the higher the probability of committing a type II error. So it does not make sense to set your alpha level very small to prevent type I errors, because then the chance of committing a type

II error will be really high.

Taken together, there are four possibilities. H_0 may be true, but given your data you reject it (type I error). H_0 may be true and you accept it (perfectly fine). H_0 may be false and you reject it (perfectly fine). Or, H_0 may be false, but you accept it (type II error).¹² I have summarized this in the following table.

	H0 is (in reality)	
	true	false
Decision about H0	reject	type I error (α)
given the data	accept	type II error (β)

One important thing in doing empirical research is the question of how many participants you need. The probability of falsely detecting an effect that is not there is α . The probability of falsely missing an effect that is not there is β . These values depend on each other. If we set our significance level (alpha level) really low, β will get bigger, so the chances that we miss an actual effect rise. At the same time, something different changes, namely the value of $1 - \beta$. This is the probability of correctly rejecting a false H_0 . This value, namely $1 - \beta$ has its own name: power. The power of a test is therefore the probability that the test correctly rejects the null hypothesis.

Suppose that a new drug is developed and that the new drug actually is better than the old one that you usually get for some disease. However, the drug is only slightly better. The measure of the size of an effect is called ‘effect size’ in statistics. If the effect is really small, it will be hard to detect. What you will need is more participants. But how many? The question is how much power does your experiment need. This is something you want to know before you do a study! And to determine your power you need to know which kind of statistical test you want to carry out. Actually, the calculation of the power of a test is so complicated that you will need a special software. You can, for example, use the free tool G*Power (Faul et al. 2007).

However, there is good news for you! Although in many empirical studies, power is far too low (meaning that the chances are high that they miss an actual effect) (see Cohen 1962; Sedlmaier & Gigerenzer 1989), this is usually not a problem of acceptability judgment studies! At least, if the phenomenon you’re looking for is not too subtle. In the next section I will discuss statistical power and the question of how many participants you will need in more detail.

¹²Actually, it would be a little bit more correct to say ‘fail to reject’ instead of ‘accept’.

Before we come to that, it should be noted that power depends on several factors. It depends on the size of the effect as larger effects are easier to detect than smaller ones. It depends on the false positive rate. It depends on the nature of the task. It depends on the statistical test you want to conduct. It depends on how many responses you collect per participant and, crucially, it depends on your sample size: The more participants you have, the higher your power will be.

15.5 How many participants do I need to consult?

While the probability of rejecting the null hypothesis when it is actually true is easy to handle as it is set in advance (usually, $\alpha = 0.05$), the probability of rejecting the null hypothesis when it is actually false, i. e., statistical power, is often overlooked. Of course, the probability of finding an effect does not only depend on your statistics or your research design, but one crucial factor is how big the effect is. A huge effect is easier to detect than a very small effect. How big an effect is is measured by ‘effect sizes’. There is no magic behind effect sizes. An effect size simply is a measure of how big an effect is. For example, the difference between two means is an effect size (the only important point is that p -values are not effect sizes, i. e., a small p -value does not mean that there is a big effect).

There are also standardized effect sizes. A good example of a standardized effect size is ‘Cohen’s d ’ that is used for differences between means.¹³ As you already know, two different means can signify that they come from two different populations. Take a look back at Figure 9 on the right. The bigger the difference between your means, the less overlap there will be between your populations (if there really are two populations). This overlap depends, of course, on the form of the distributions. The form of a distribution can be described mathematically with the standard deviation. This means, that our standardized effect size needs to take the standard deviation into account. That’s exactly what Cohen’s d does.¹⁴ I don’t wanna go into the details, but rather just present how to interpret Cohen’s d by rule of thumb.

Cohen’s d is a number that can be small, medium, or large. A Cohen’s d of 0.2 is generally considered small, a Cohen’s d of 0.5 is considered a medium effect, and a Cohen’s d larger than 0.8 is considered to be a large effect. In a famous article Cohen (1962) analyzed 70 psychological studies and found that the probability of finding medium-sized

¹³Actually, there is some discussion if it is appropriate to use Cohen’s d for repeated measures designs, but there seems to be no conclusion yet.

¹⁴You may have noticed that this means that the standard deviations of your two groups should be similar. Actually, your samples should additionally have the same sizes. If these requirements are not met you should use a measure that is called Glass’ delta.

effects was only 0.48. For small-sized effects the statistical power was only 0.18. Only for large effects was this probability 0.83. Cohen's results were later replicated for more studies (Sedlmeier & Gigerenzer 1989; Maxwell 2004). On the whole it seems that the power of most studies is far too low and that the chance of finding an effect that is really there is actually only 50/50.

A simple way out of this dilemma is to use more participants as power increases with more data. More participants, however, means more work and more costs. What you want to know is how many participants you will need. In other words: You should calculate how many participants you will need for your study in advance. For this, you need to specify the statistical tests you want to carry out. Power calculation is very complex, but fortunately, there is free computer software out there that will help you.

And there is even more good news for you! The first thing is that the statistical tests carried out in grammaticality judgments are fairly easy, so power calculation is not a big deal. The second point is that effect sizes in grammaticality judgments are usually large. So you do not need hundreds of participants—at least when the contrasts you study are not too fine-grained. The third and last point is that there is a study that already compared different effect sizes in grammaticality judgments.

Sprouse & Almeida (2012; 2017) tested many different phenomena with effect sizes ranging between a Cohen's *d* from 0.15 to 1.96 with four different tasks (magnitude estimation, Likert scale, yes-no, and force choice tasks). Their results help us to estimate the number of participants needed for our purpose. I will just present a rough picture here, but note that you will, in many cases, need to calculate how many participants you need (not only for grammaticality judgments, but for every empirical study). See below for more details and software recommendations.

By rule of thumb you want to reach at least 80 % power (Cohen 1988). An empirical study with 80 % power has an 80 % probability of detecting an effect when the effect actually exists. Sprouse & Almeida showed that the lowest number of participants is needed for force-choice tasks to achieve 80 % power and that there is a strong power disadvantage for yes-no tasks. This means that if you want to know something about very fine-grained differences between constructions, i.e. phenomena with small effect sizes, you would like to use a force-choice task. Sprouse & Almeida (2017:13-14) describe the force-choice task they used as follows:

In the (two-alternative) force-choice task (FC), target sentences are presented in vertically arranged pairs, with each sentence in the pair followed by a single radio button. Participants are asked to indicate which of the two sentences is

more acceptable by selecting the radio button next to that sentence. In the current FC experiment, the pairs were lexically matched as to form minimal pairs that varied only by the syntactic property of interest.

As most phenomena in syntax have medium or large effect sizes, I won't go into the details of force-choice tasks in this tutorial, but stick with Likert-scale tasks.

One question you may ask is how would you know if the phenomenon you want to look at has a small, medium, or large effect size. For previously studied phenomena you can simply calculate a Cohen's d by using the means and standard deviations. I recommend reading the paper by Anderson, Kelley & Maxwell (2017) and using G*Power (Faul et al. 2007) and the web-based apps that can be found at <https://designexperiments.com/shiny-r-web-apps>.

If there are no previous study on the phenomenon of interest, and this may be the case, it is a good idea to take a look at Sprouse & Almeida (2012; 2017):

For studies that do not have published means and standard deviations, or for planning a new study where the means and standard deviations are unknown, the situation is a bit more complicated. One possible approach would be to informally compare the phenomena with unknown effect sizes to the known phenomena in this study[...]. Although this method is not precise, it should be possible to arrive at a relatively accurate, albeit coarse, effect size (i.e., small, medium, large, very large) [...]. (Sprouse & Almeida 2012:28-29)

Figure 10 adapted from Sprouse & Almeida (2012:26) shows how much power you get when using Likert item tasks with different sample sizes. As you can see from the Figure, you will need approximately 37 participants when you have a medium sized effect. The graph also shows that Likert item tasks are not suitable for small effect sizes.

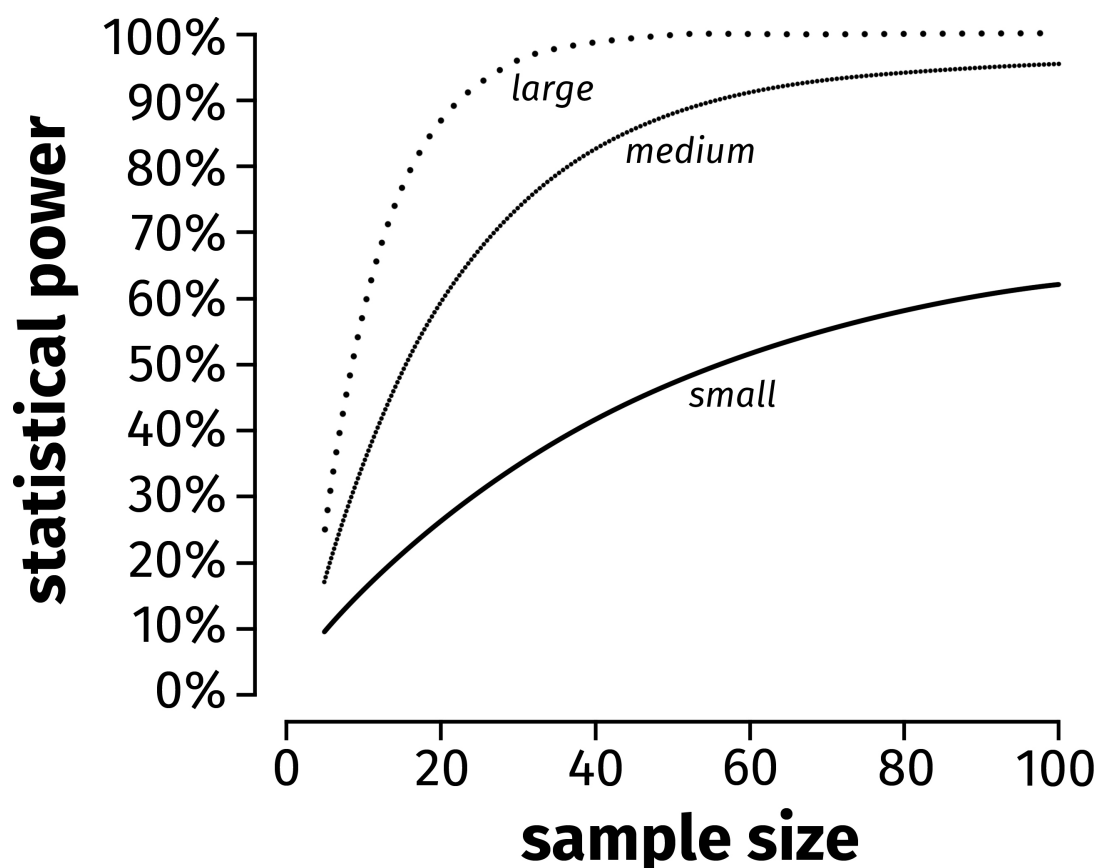


Figure 10: Sample size needed to achieve different percentages of power when using Likert item task in acceptability judgments. Adapted from Sprouse & Almeida (2012:26).

15.6 Back to our statistical test

Let's come back to our statistical test. We have two constructs, a lot of data, and two means. We still ask ourselves if those two means come from the same population or not. You want to run a test, get a p -value, and—based on that p -value—you want to see if you should reject H_0 . Although I said that your H_0 usually is $\mu_1 = \mu_2$, in many cases you have a clear expectation about the outcome of a test. For example, in most cases, you will have an expectation as to which grammatical construction should be worse than another. However, in some cases, you just assume that there will be a difference between the two groups/conditions, but you don't know if one construction will get better ratings, i. e., you assume a difference, but you don't know the direction of this difference.

There are two types of statistical tests: one-sided and two-sided tests (also called one-tailed and two-tailed tests). The choice, however, is simple. If you assume that one construction will be better than the other, you choose a one-sided test, if you don't know, you choose a two-sided test. This decision is made before you take a look at your data,

of course! Note that one-sided tests are more powerful, i. e., while a two-sided test may not show a significant result, a one-sided test may do so using the same data. Only use one-sided tests when you have a reason to do so!

There are more decisions to make before you can choose the right test besides setting your alpha-value (usually 0.05) and specifying if you want a one-sided or two-sided test. One important point is to figure out if you are dealing with paired or unpaired (i. e., independent) data sets. Data is said to be paired if there are two values that belong together. As we have tested two constructions by asking the same participants we are dealing with paired data. Each participant has given the same amount of ratings for the first construction as for the second construction.

To compare the results you obtained you can do a *t*-test. To be more precise, you can run a paired *t*-test, let's say a two-tailed test as you have no clue if construction A will receive higher ratings than construction B. Note that paired tests are also called tests for dependent samples. This should not confuse you. It's just two names for the same thing. The input of your test will be two lists. For every participant you calculate the mean of all judgments of the sentences representing construction A and a mean of all judgments of the sentences of construction B. With this, you receive a list of two means.

Excursus: Should I standardize my data?

Some authors claim that it is useful to normalize the raw data you received by your participants by z-score transforming them. This is done as it is sometimes assumed that not all participants make use of the full range of the Likert items response range (e. g., from 1 to 7). No worries! This is not a complicated thing to understand or to do! Standardization simply means that you transform the data from each participant so that it gets the shape of a normal distribution with a mean equal to 0 and a standard deviation equal to 1. Thus, you receive negative and positive values around a mean of 0. Your computer can do that for you. Schütze & Sprouse (2013:43) describe the procedure:

[E]ach participant's responses are transformed using the z-score *transformation* to eliminate some of the potential scale bias [...]. The z-score transformation allows us to express each participant's responses on a *standardized* scale. It is calculated as follows: For a given participant P, calculate the mean and standard deviation of all of P's judgments. Next, subtract each of P's judgments from the mean. Finally, divide each of these differences by P's standard

deviation.

If you read papers reporting acceptability ratings you will see that some authors transform their data and others do not. While I think that it does no harm to your data (or results) I'm a little bit undecided if you should transform your data without a good reason. In part II of the tutorial we will see cases in which z -transforming your data is helpful.

A short comment before you proceed: You can also fit a mixed model for the analysis of your data. However, a t -test is just fine for simple comparisons. If you are interested in mixed models, take a look at the second part of this tutorial here: www.fabianbross.de/mixedmodels.pdf.

16. Do your statistics in different environments: OpenOffice, R/RStudio, JASP

In this section, I will briefly go through a simple fictional example using three different software tools—all of them are free of charge and can be used on virtually all platforms (Linux, Windows, Mac). I will show you how to do a t -test and do some basic plots in OpenOffice, with R and in JASP. To get the best overview on what is happening, I strongly suggest you to read all three subsections to follow—also because there is more information in the subsections than a pure description of the software.

16.1 OpenOffice

The first thing we need is some data. Again, for reasons of simplicity, I assume that we want to look at two different constructions A and B. For each construction each participant judged 4 sentences. Thus, we have a total of 8 ratings per participant. We got data from 10 participants (for a real study you would need more data, of course). I created an OpenOffice sheet with the judgments the participants gave. This sheet is shown in Figure 11. Note that I simplify matters a lot here. With real data the sheet would look different as by using a Latin square not all participants rated the exact same sentences.

	A	B	C	D	E	F	G	H	I	J	K
1	Participant	Item 1 - Construction A	Item 2 - Construction A	Item 3 - Construction A	Item 4 - Construction A	MEAN A	Item 1 - Construction B	Item 2 - Construction B	Item 3 - Construction B	Item 4 - Construction B	MEAN B
2	1	6	7	6	7	6	1	2	1	1	1
3	2	7	7	5	6	6	1	3	1	3	2
4	3	7	5	6	6	6	1	0	4	3	3
5	4	7	6	5	5	5	3	3	2	1	1
6	5	6	6	5	6	6	1	3	3	2	2
7	6	6	5	5	7	7	1	1	1	3	3
8	7	6	6	7	6	7	2	2	3	2	2
9	8	6	7	7	5	6	1	3	2	2	2
10	9	5	6	5	5	5	1	1	3	2	2
11	10	7	7	5	6	6	3	1	1	2	2

Figure 11: 10 participants rated 8 sentences

It is obvious, that construction A received an extremely high rating (natural) and construction B an extremely low rating (unnatural). Let's nevertheless visualize and analyze the data. The first thing we want is the mean rating for each participant. In OpenOffice, the mean is calculated by a function called "AVERAGE" (the terms 'mean' and 'average' are synonyms). You just write "=AVERAGE()" into the field you want to have your mean value show up in and you can select the cells you want to have your mean from. In Figure 12 I calculated the mean of the ratings participant 1 gave to the sentences making up the scale for construction A. This is labeled 'step 1' in the figure. In step 2 I expanded the mean cell to all the other participants' rating of construction A and in step 3 you can see that I did the exact same thing for construction B (don't be confused by the commas; I'm from Germany, we use commas instead of dots as separators—this doesn't matter).

AVERAGE =AVERAGE(B2:E2)											
	A	B	C	D	E	F	G	H	I	J	K
1	Participant	Item 1 - Construction A	Item 2 - Construction A	Item 3 - Construction A	Item 4 - Construction A	MEAN A	Item 1 - Construction B	Item 2 - Construction B	Item 3 - Construction B	Item 4 - Construction B	MEAN B
2	1	6	7	6	7	=AVERAGE(B2:E2)	1	2	1	1	1
3	2	7	7	5	6		1	3	1	3	2
4	3	7	5	6	6		1	0	4	3	3
5	4	7	6	5	5		3	3	2	1	1
6	5	6	6	5	6		1	3	3	2	2
7	6	6	5	5	7		1	1	1	3	3
8	7	6	6	7	6		2	2	3	2	2
9	8	6	7	7	5		1	3	2	2	2
10	9	5	6	5	5		1	1	3	2	2
11	10	7	7	5	6		3	1	1	2	2

STEP 1

	A	B	C	D	E	F	G	H	I	J	K
1	Participant	Item 1 - Construction A	Item 2 - Construction A	Item 3 - Construction A	Item 4 - Construction A	MEAN A	Item 1 - Construction B	Item 2 - Construction B	Item 3 - Construction B	Item 4 - Construction B	MEAN B
2	1	6	7	6	7	=AVERAGE(B2:E2)	1	2	1	1	1
3	2	7	7	5	6		1	3	1	3	2
4	3	7	5	6	6		1	0	4	3	3
5	4	7	6	5	5		3	3	2	1	1
6	5	6	6	5	6		1	3	3	2	2
7	6	6	5	5	7		1	1	1	3	3
8	7	6	6	7	6		2	2	3	2	2
9	8	6	7	7	5		1	3	2	2	2
10	9	5	6	5	5		1	1	3	2	2
11	10	7	7	5	6		3	1	1	2	2

STEP 2

	A	B	C	D	E	F	G	H	I	J	K
1	Participant	Item 1 - Construction A	Item 2 - Construction A	Item 3 - Construction A	Item 4 - Construction A	MEAN A	Item 1 - Construction B	Item 2 - Construction B	Item 3 - Construction B	Item 4 - Construction B	MEAN B
2	1	6	7	6	7	6.5	1	2	1	1	1.25
3	2	7	7	5	6	6.25	1	3	1	3	2
4	3	7	5	6	6	6	1	0	4	3	2.25
5	4	7	6	5	5	5.75	3	3	2	1	2.25
6	5	6	6	5	6	5.75	1	3	3	2	2.25
7	6	6	5	5	7	5.75	1	1	1	3	3
8	7	6	6	7	6	6.25	2	2	3	2	2.25
9	8	6	7	7	5	6.5	1	3	2	2	2
10	9	5	6	5	5	5.25	1	1	3	2	1.75
11	10	7	7	5	6	6.25	3	1	1	2	1.75

STEP 3

Figure 12: Calculate the means.

What we want now is the mean ratings of the constructions in general. So we calculate the mean of the means. This is shown in Figure 13. As you can see in the figure, construction A was rated to be 6.025 and construction B was rated to be 1.925. These values tell us that, indeed, construction A was rated to be well-formed and construction B was rated

to be ill-formed.

STEP 1

AVERAGE												
=AVERAGE(F2:F11)												
	A	B	C	D	E	F	G	H	I	J	K	
1	Participant	Item 1 - Construction A	Item 2 - Construction A	Item 3 - Construction A	Item 4 - Construction A	MEAN A	Item 1 - Construction B	Item 2 - Construction B	Item 3 - Construction B	Item 4 - Construction B	MEAN B	
2	1	6	7	6	7	6.5	1	2	1	1	1.25	
3	2	7	7	5	6	6.25	1	3	1	3	2	
4	3	7	5	6	6	6	1	1	4	3	2.25	
5	4	7	6	5	5	5.75	3	3	2	1	2.25	
6	5	6	6	5	6	5.75	1	3	3	2	2.25	
7	6	6	5	5	7	5.75	1	1	1	3	1.5	
8	7	6	6	6	7	6.25	2	2	3	2	2.25	
9	8	6	7	7	6	6.5	1	3	2	2	2	
10	9	5	6	5	5	5.25	1	1	3	2	1.75	
11	10	7	7	5	6	6.25	3	1	1	2	1.75	
12												
13												
14			MEAN A	=AVERAGE(F2:F11)								
15			MEAN B									
16												

STEP 2

	A	B	C	D	E	F	G	H	I	J	K
1	Participant	Item 1 - Construction A	Item 2 - Construction A	Item 3 - Construction A	Item 4 - Construction A	MEAN A	Item 1 - Construction B	Item 2 - Construction B	Item 3 - Construction B	Item 4 - Construction B	MEAN B
2	1	6	7	6	7	6.5	1	2	1	1	1.25
3	2	7	7	5	6	6.25	1	3	1	3	2
4	3	7	5	6	6	6	1	1	4	3	2.25
5	4	7	6	5	5	5.75	3	3	2	1	2.25
6	5	6	6	5	6	5.75	1	3	3	2	2.25
7	6	6	5	5	7	5.75	1	1	1	3	1.5
8	7	6	6	6	7	6.25	2	2	3	2	2.25
9	8	6	7	7	6	6.5	1	3	2	2	2
10	9	5	6	5	5	5.25	1	1	3	2	1.75
11	10	7	7	5	6	6.25	3	1	1	2	1.75
12											
13											
14			MEAN A		6.025						
15			MEAN B		1.925						
16											

Figure 13: Mean ratings of the constructions.

However, two means tell us nothing about how the data is distributed. First, we want to know the standard deviation. We get the standard deviation by using the STDEV function. So you type in “=STDEV()” and select the cells from which you want to calculate the standard deviation. This is shown in Figure 14.

STEP 1

AVERAGE

fx

✖

✓

=STDEV(F2:F11)

	A	B	C	D	E	F	G	H	I	J	K
1	Participant	Item 1 - Construction A	Item 2 - Construction A	Item 3 - Construction A	Item 4 - Construction A	MEAN A	Item 1 - Construction B	Item 2 - Construction B	Item 3 - Construction B	Item 4 - Construction B	MEAN B
2	1	6	7	6	7	6.5	1	2	1	1	1.25
3	2	7	7	5	6	6.25	1	3	1	3	2
4	3	7	5	6	6	6	1	1	4	3	2.25
5	4	7	6	5	5	5.75	3	3	2	1	2.25
6	5	6	5	5	6	5.75	1	3	2	2	2.25
7	6	6	5	5	7	5.75	1	1	1	3	1.5
8	7	6	6	6	7	6.25	2	2	3	2	2.25
9	8	6	7	7	6	6.5	1	3	2	2	2
10	9	5	6	5	5	5.25	1	1	3	2	1.75
11	10	7	7	5	6	6.25	3	1	1	2	1.75
12											
13					SD (standard deviation)						
14					6.025						
15			MEAN A								
16			MEAN B		1.925						

STEP 2

	A	B	C	D	E	F	G	H	I	J	K
1	Participant	Item 1 - Construction A	Item 2 - Construction A	Item 3 - Construction A	Item 4 - Construction A	MEAN A	Item 1 - Construction B	Item 2 - Construction B	Item 3 - Construction B	Item 4 - Construction B	MEAN B
2	1	6	7	6	7	6.5	1	2	1	1	1.25
3	2	7	7	5	6	6.25	1	3	1	3	2
4	3	7	5	6	6	6	1	1	4	3	2.25
5	4	7	6	5	5	5.75	3	3	2	1	2.25
6	5	6	5	5	6	5.75	1	3	2	2	2.25
7	6	6	5	5	7	5.75	1	1	1	3	1.5
8	7	6	6	6	7	6.25	2	2	3	2	2.25
9	8	6	7	7	6	6.5	1	3	2	2	2
10	9	5	6	5	5	5.25	1	1	3	2	1.75
11	10	7	7	5	6	6.25	3	1	1	2	1.75
12											
13					SD (standard deviation)						
14					6.025	0.3987828705					
15			MEAN A		1.925	0.1844861711					
16			MEAN B								

Figure 14: Calculating the standard deviation.

As we are math pros now, we also want to calculate the 95%-confidence intervals (CIs). To calculate the CIs, we need three pieces of information: what kind of confidence interval we want to calculate (a 95%-confidence interval; this is expressed via the alpha level which is 0.05 then), the standard deviation (we already have that) and the number of data points we have (this number is 10 as we have 10 values). The CONFIDENCE function

thus needs three arguments and looks like this: “=CONFIDENCE(α ; SD; size)”. Stare at Figure 15 for a moment to see what I did.

STEP 1

STDEV											
=CONFIDENCE(0.05,E14;10)											
	A	B	C	D	E	F	G	H	I	J	K
Participant	Item 1 - Construction A	Item 2 - Construction A	Item 3 - Construction A	Item 4 - Construction A	Item 4 - Construction A	MEAN A	Item 1 - Construction B	Item 2 - Construction B	Item 3 - Construction B	Item 4 - Construction B	MEAN B
1	1	6	7	6	7	6.5	1	2	1	1	1.25
2	2	7	7	5	6	6.25	1	3	1	3	2
3	3	7	5	6	6	6	1	4	3	2.25	3
4	4	7	6	5	5	5.75	3	3	2	1	2.25
5	5	6	6	5	5	5.75	1	3	3	2	2.25
6	6	6	5	5	5	5.75	1	1	1	3	1.5
7	7	6	6	6	6	7	2	2	3	2	2.25
8	8	6	7	7	7	6.5	1	3	2	2	2
9	9	5	6	5	5	5.25	1	1	3	2	1.75
10	10	7	7	5	6	6.25	3	1	1	2	1.75
11											
12											
13											
14											
15											
16											

Figure 15: Confidence intervals.

Now we are able to visualize the results. We make a little plot by selecting our mean values and by clicking “Insert” → “Chart”. This looks like step 1 in Figure 16. By clicking “XY (Scatter)” you will get a simple plot as shown in step 2. Note that you want to adjust the legend later.

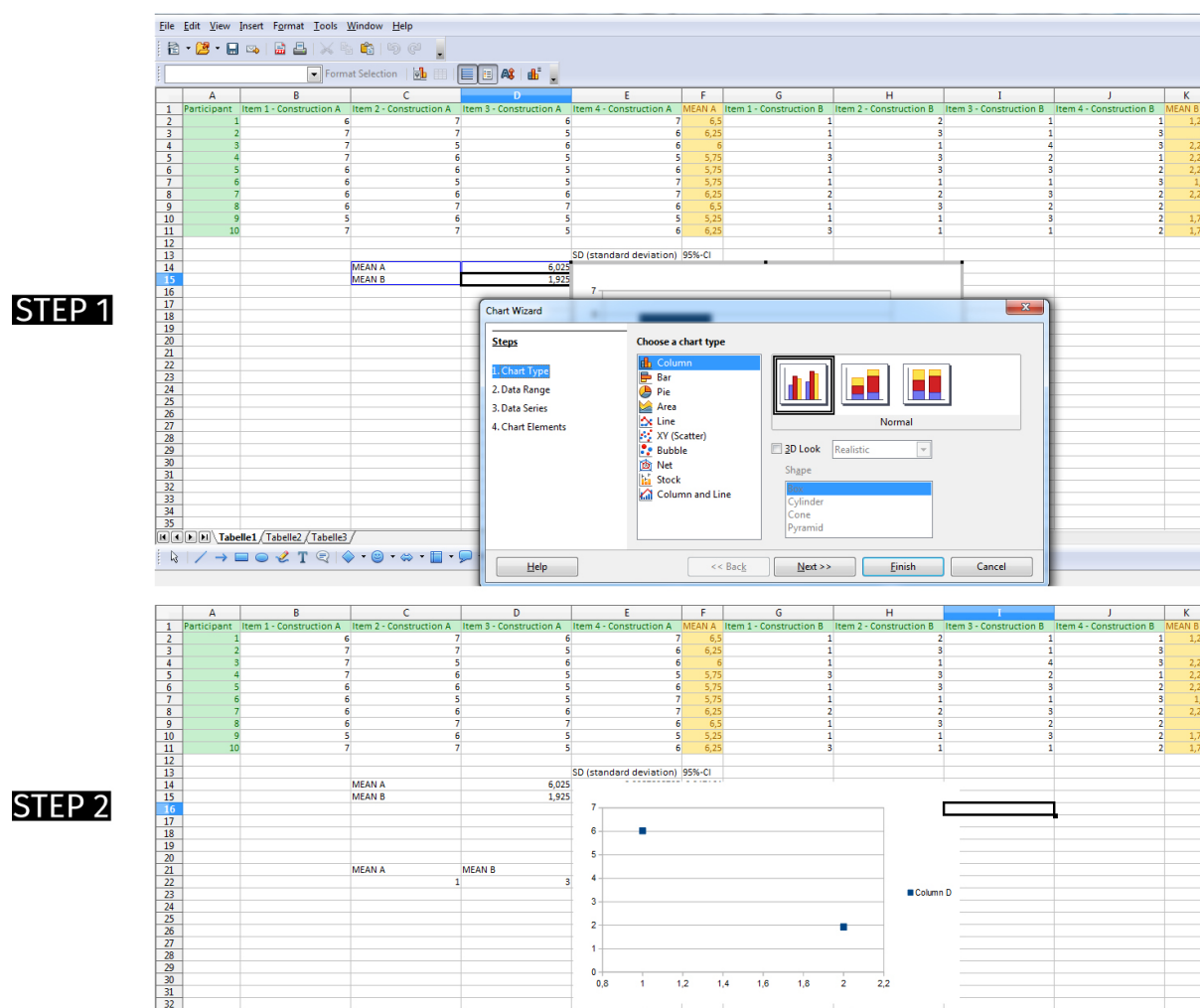


Figure 16: Make a simple plot.

The next step is to indicate the confidence intervals by adding error bars. To do this, double click your plot. Then click on Insert → Y Error Bars as shown in step 1 in Figure 17. Now check the box “Same value for both” and click on “Cell range”. You can now select the two 95%-CI values we calculated (both of them) by clicking on “Positive (+/–)”. See Figure 17. Note that error bars you see in publications do not always represent the 95%-confidence intervals. Sometimes they show the standard deviation, the standard error of the mean, or whatnot. This means for you, that you have to say that the error bars show the 95%-confidence intervals of the mean. You can specify this in your captions. If you do not do this nobody can interpret your plot.

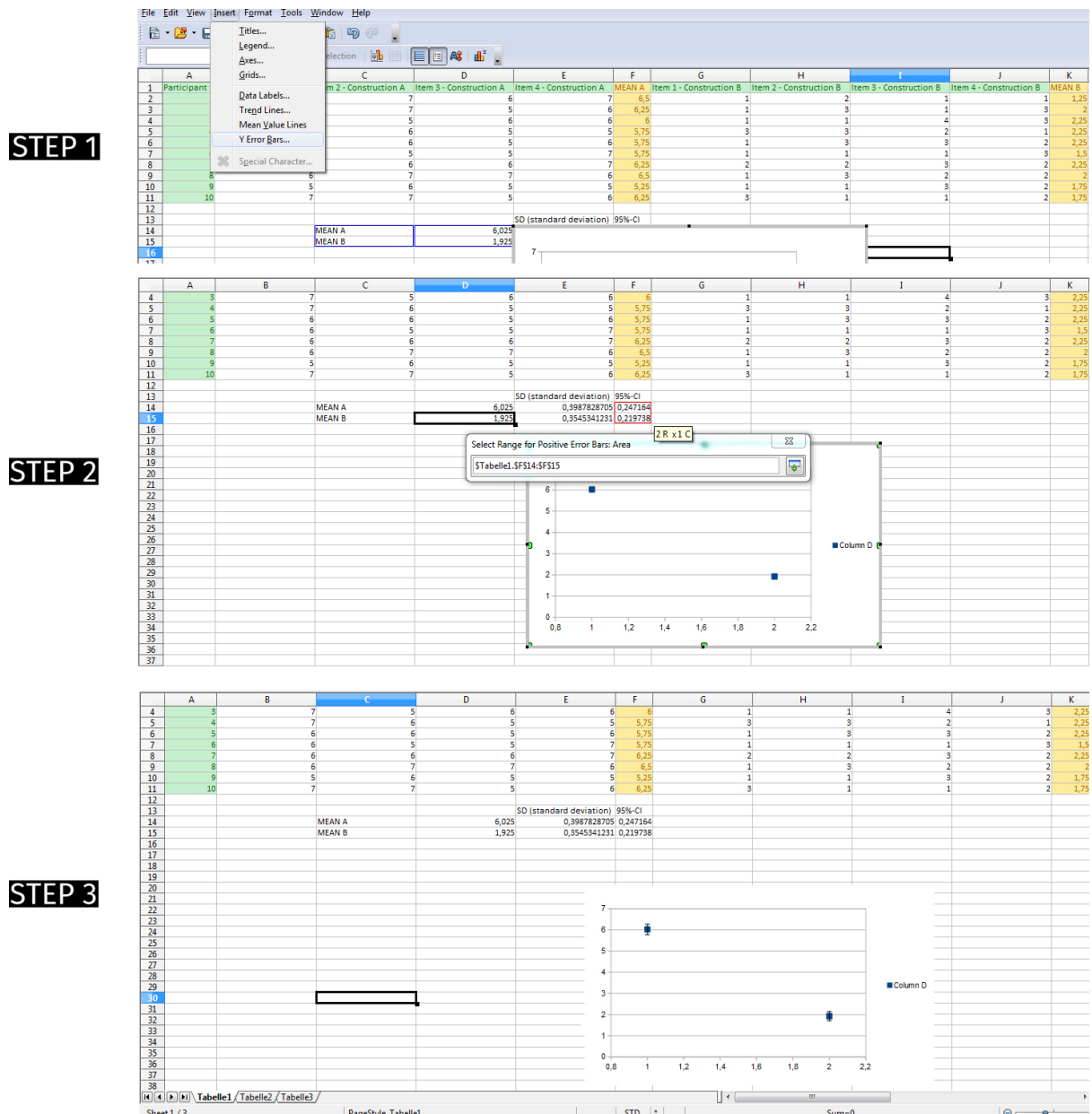


Figure 17: Adding error bars.

The figure we just created may not look 100% professional, but it is already very informative! Before we apply a *t*-test we briefly look at two more examples. First, look at Figure 18. I changed the ratings participants gave to the sentences representing construction A (but left construction B as is). I added some variation. As you can see, the result is that the standard deviation got bigger as well as the confidence interval.

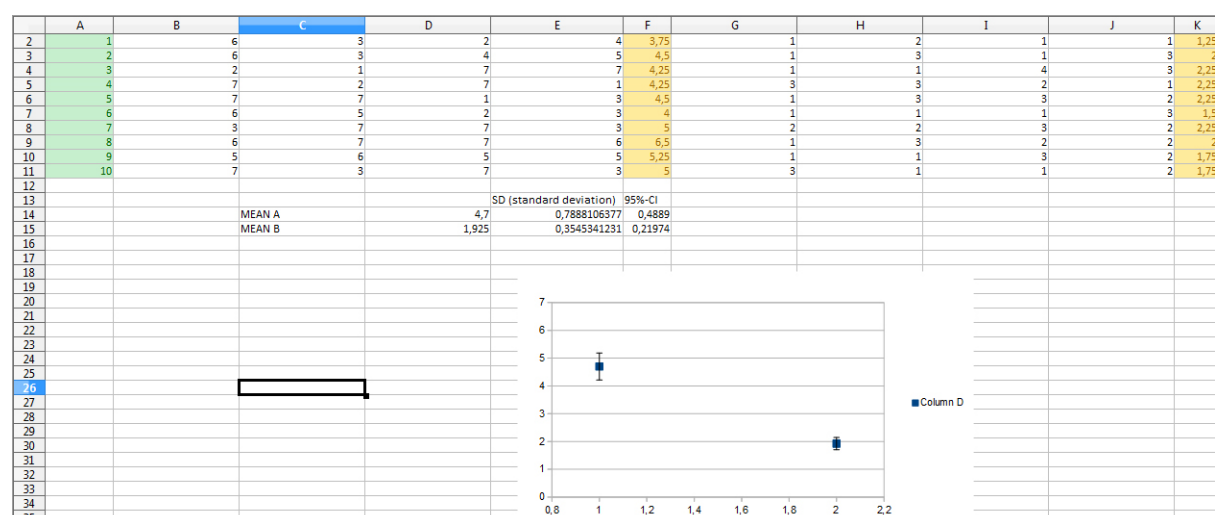


Figure 18: More variation in construction A.

Now look at figure 19. I now added more variation to the judgments of the sentences for construction B. The two means are now very close together (4.575 versus 4). Again, the SD and the 95%-CI got bigger, too.

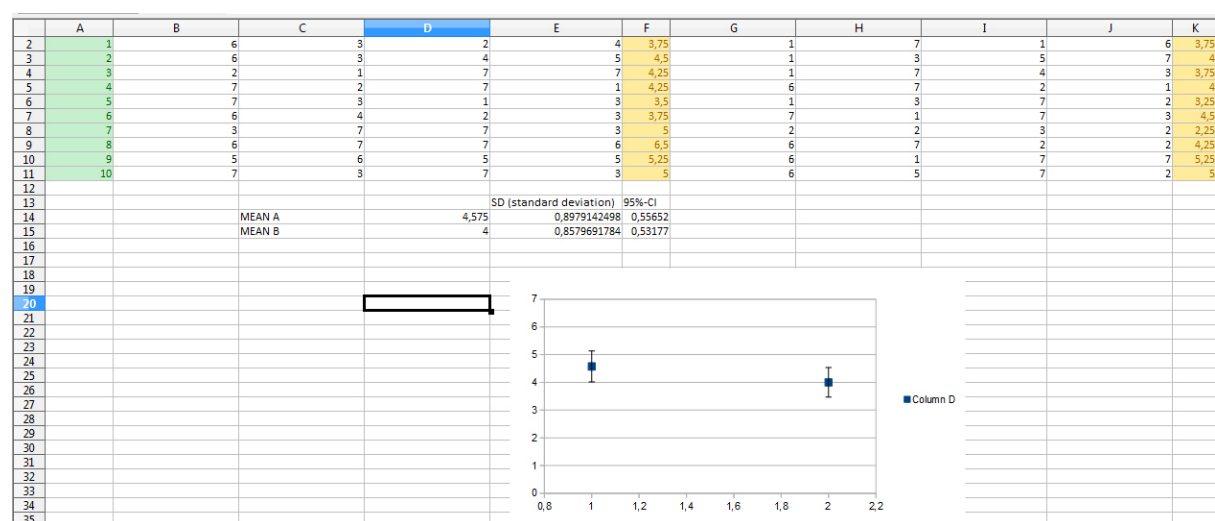


Figure 19: More variation in construction A and B.

Finally, let's apply a *t*-test to compare the ratings the two constructions received. We will do this for all three examples. I have summarized the three examples in Figure 20. The means in example 1 are very far away from each other and the length of the error bars is very short. In example two, the means are also rather far away from each other but one confidence interval is a little bit bigger. In the last example, the means are rather close

together. Although we might say “Hey! There is nearly no difference!” we still need a statistical test producing numbers we can rely on.

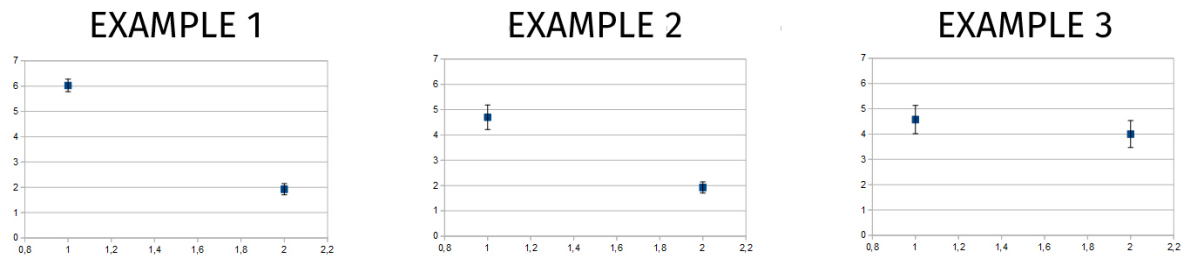


Figure 20: Our three examples.

Let's apply the t -test. We do this by using the TTEST function. This function has the following syntax: “TTEST(data1; data2; mode; type)”. Of course, “data 1” and “data 2” are our mean ratings. With “mode” you can choose between a one-tailed and a two-tailed test. Just type 1 for one-tailed and 2 for two-tailed. We made no predictions about differences between our constructions so we choose a two-tailed test. Finally, we need to specify “type”. We choose 1 which stands for paired data. I did this as shown in Figure 21. As you can see, we get the value 2.80623010563253E-009. This is our p -value. The E just says that there are nine zeros before our number. Thus we calculated a p -value of 0.00000000280623010563253. This value is smaller than 0.05 and our result is statistically significant. Of course, we expected that as the means are very far away from each other. It is very unfortunate that you do not have many options doing a t -test in OpenOffice. Meaning that you do not get much information about the statistics, except a p -value. You can also calculate a t -test online. A trustworthy website is www.socscistatistics.com. Here you also can specify the alpha level and get more numbers that you may need for reporting your statistics.

B20 =TTEST(F2:F11;K2:K11;2;1)											
1	A	B	C	D	E	F	G	H	I	J	K
2	Participant	Item 1 - Construction A	Item 2 - Construction A	Item 3 - Construction A	Item 4 - Construction A	MEAN A	Item 1 - Construction B	Item 2 - Construction B	Item 3 - Construction B	Item 4 - Construction B	MEAN B
3	1	6	7	6	7	6,5	1	2	1	1	1,25
4	2	7	7	5	6	6,25	1	3	1	3	2
5	3	7	5	6	6	6	1	1	4	3	2,25
6	4	7	6	5	5	5,75	3	3	2	1	2,25
7	5	6	6	5	5	5,75	1	3	3	2	2,25
8	6	6	5	5	7	5,75	1	1	1	3	1,5
9	7	6	6	6	7	6,25	2	2	3	2	2,25
10	8	6	7	7	6	6,5	1	3	2	2	2
11	9	5	6	5	5	5,25	1	1	3	2	1,75
12	10	7	7	5	6	6,25	3	1	1	2	1,75
13						SD (standard deviation)					
14						95%-CI					
15			MEAN A	6,025		0,3987828705					
16			MEAN B	1,925		0,3545341231					
17											
18											
19											
20											
21											

Figure 21: Calculating a p -value.

For example 2 we get a p -value of 0.00000682979925227733 (6.82979925227733E-006) and for example 3 a p -value of 0.1270640781. This is also what we expected. In example 2 the means were still very far away from each other, so we get a statistically significant p -value. In example 3, the means are close together and the result is not significant with p being greater than 0.05. However, the fact that two means are close together tells you nothing. There still could be a significant difference! The reason for this is that the means are point measures and they tell you nothing about the spread of the data.

The 95%-confidence intervals, i. e., the error bars, however, may help you to interpret the results by rule of thumb. If the error bars do not overlap, chances are high that there is a very low p -value and a significant difference. If they overlap a little, p is often below 0.05 and if they overlap more, p might be high. However, you do not know for sure and have to make a test in any case, but there is a correlation between p -values and confidence intervals (if you want to know more about this, take a look at Cumming & Finch 2005 for a great overview on visual interpretation of error bars in different circumstances).

16.2 R and RStudio

R is a programming language as well as a software for statistical purposes. It's very powerful and free to use. Some people are scared of R, but I really recommend using it! Additionally, you may use RStudio which is an integrated development environment (IDE). This simply means that RStudio is a software with some extras you can use R in. You can think of R as a programming language and of RStudio as your user interface. In the following, I will go through example 1 from the previous subsection. However, I will simplify matters in a way and assume that you have already calculated the means of the sentences representing the two constructions for each participants.

The first thing we need is to load these means into R. Under normal circumstances I would assume that you have your data in some table. R prefers .csv files. CSV stands for 'comma-separated values' and is a very simple file format to store tables. With RStudio you can easily import .csv files with the "Import Dataset" function. Here, we will simply tell R our numbers by hand. First, you need to install R and RStudio, of course (you'll find out how!). Then we open a new file. We now create two lists. One list with the mean ratings for construction A and one list for the mean ratings of construction B. This is achieved by the "c function". This function simply combines values into a list:

```
A <- c(6.5, 6.25, 6, 5.75, 5.75, 5.75, 6.25, 6.5, 5.25, 6.25)
B <- c(1.25, 2, 2.25, 2.25, 2.25, 1.5, 2.25, 2, 1.75, 1.75)
```

The code above creates two variables, A and B. We now have two lists, a list called ‘A’ containing the mean ratings for construction A and a list called ‘B’ for the mean ratings for construction B. The little arrow simply is an assignment operator: It tells R that I want to give the list a name.

If you type in the code above into your file in RStudio you first need to compile it. This is done by selecting the code and hitting `ctrl` + `↵` (alternatively, you can click on “code” in the menu bar on the top of your screen and then choose “Run selected line(s)”). Figure 22 shows what happens in RStudio when you do this.

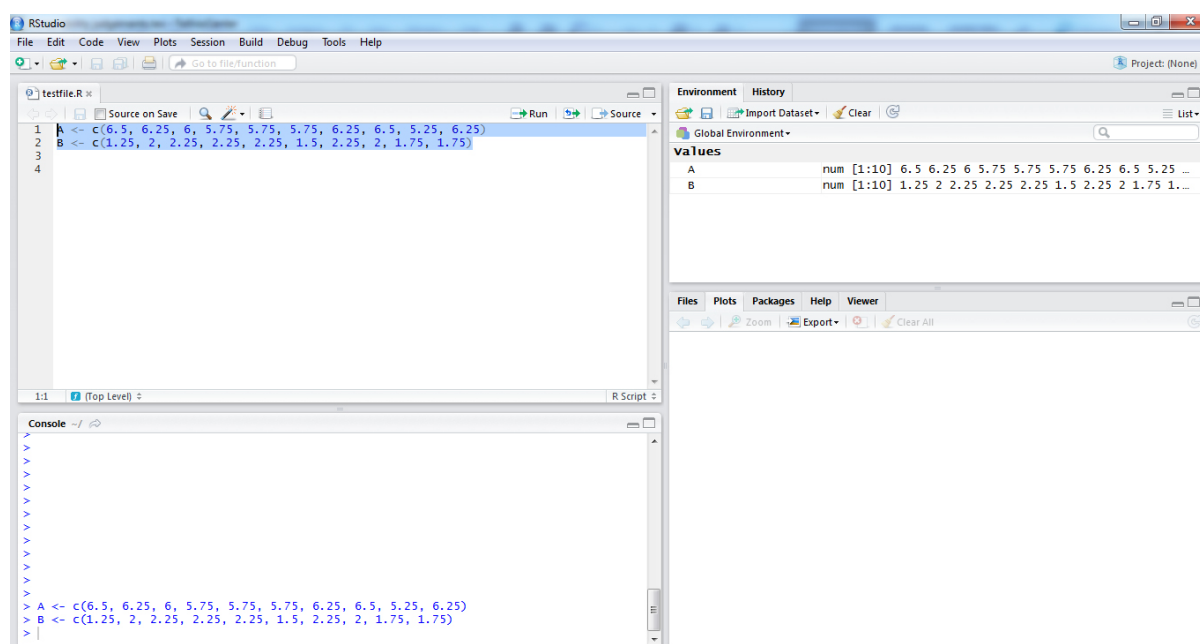


Figure 22: Creating two lists in RStudio.

As you can see from the figure, I wrote my code in the window in the upper left corner. When compiling the code you see the results in the window at the bottom on the left. For now, we will ignore the other two windows on the right. Now we calculate the means of the means and assign each of them a variable. I will call the variables “meanA” and “meanB”:

```
meanA <- mean(A)
meanB <- mean(B)
```

To see the results, you just type in “meanA” and “meanB” and compile the results by selecting “meanA” and “meanB” with the cursor and hitting `ctrl` + `↵`. The result will

look like this:

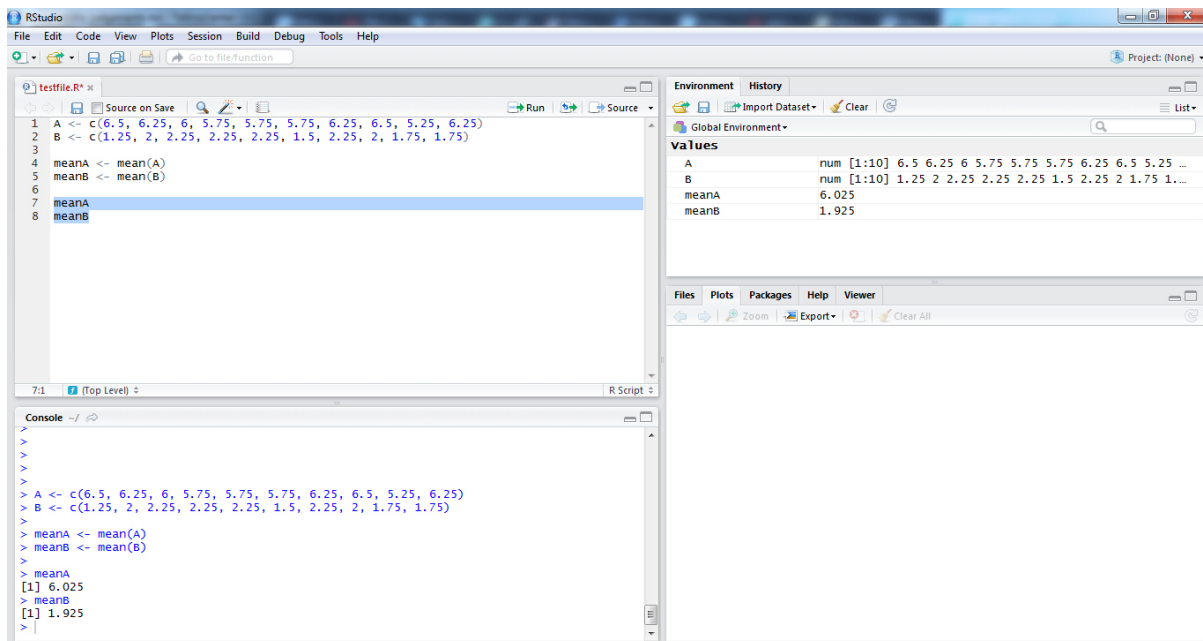


Figure 23: Calculating the means.

We can calculate the standard deviation in a similar way by using the sd function:

```
sdA <- sd(A)
sdB <- sd(B)
```

In the OpenOffice example, we calculated the 95%-confidence interval. We want to do this in R too. First, we need a better format of our data. We will build a data frame. To do this, we first concatenate our two lists into one:

```
numbers <- c(A, B)
```

The result is simply a list of all our values. We named this list “numbers”. The first ten values in this new list are the means of construction A and the other ten values are the means of construction B. Now we want a new column which contains exactly this information. We can do this by creating another list:

```
conditions <- c("contructionA", "contructionA", "contructionA",  
  "contructionA", "contructionA", "contructionA",  
  "contructionA", "contructionA", "contructionA",  
  "contructionA", "contructionB", "contructionB",  
  "contructionB", "contructionB", "contructionB",  
  "contructionB", "contructionB", "contructionB",  
  "contructionB", "contructionB")
```

A more simple way to write this would be to use the rep functions which repeats the things you want as often as you want:

```
conditions <- c(rep("contructionA", 10), rep("contructionB",  
  10))
```

The code above does exactly the same thing: It creates a list called “conditions” which contains ten times the word “contructionA” and ten times the word “contructionB”. Now we create a data frame, i.e., a table. The first column in this table contains our mean values and the second column our means. We call this data frame “df”.

```
df <- data.frame(numbers, conditions)
```

The result will look like this:

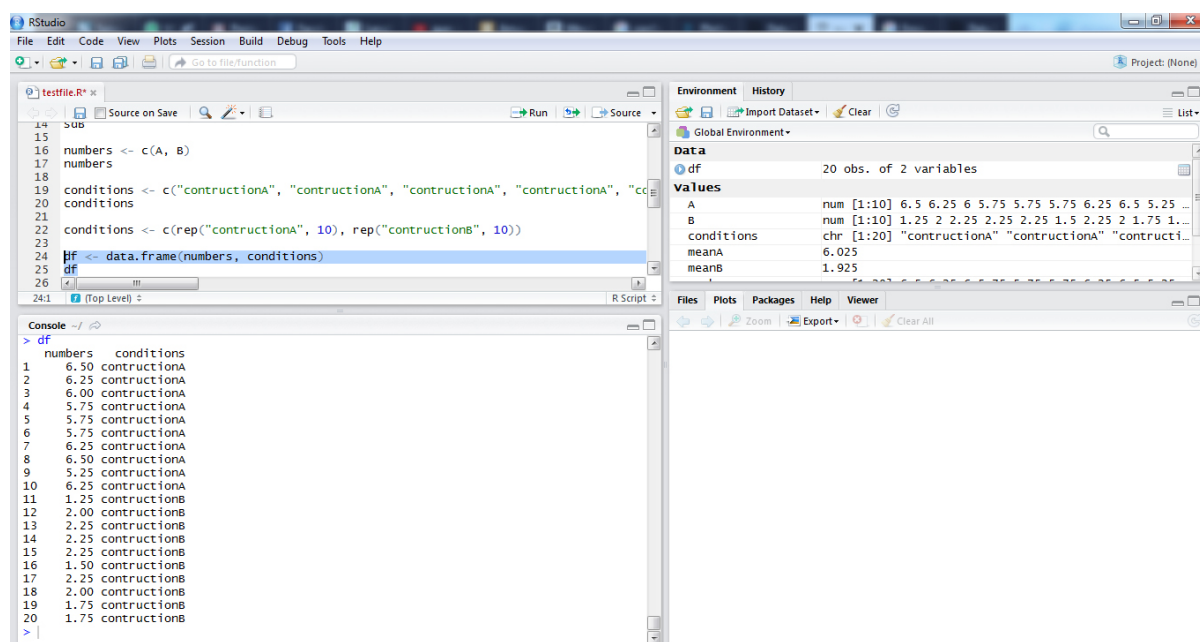


Figure 24: Our first data frame.

Our two columns have names, as you can see. The first column is named “numbers” and the second one “conditions”. In many cases you may want a third column specifying which value comes from which participant. This is often needed because we have two values from each subject. The first and the eleventh value in our first column, for example, belong together. This is an important piece of information that is missing in our table. This information is important for the statistics as we are dealing with a repeated measures design (which produces paired data). Although we do not need this right now, we will give each participant a number. As there are ten participants we need numbers from 1 to 10. We need this twice. Either you write:

```

participants <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1, 2, 3, 4, 5,
  6, 7, 8, 9, 10)

```

Or you keep it short:

```

participants <- c(1:10, 1:10)

```

I think it’s obvious what the code above does: It concatenates the numbers from 1 to 10

twice. Now, we create a new column:

```
df <- data.frame(numbers, conditions, participants)
```

The result looks like in Figure 25:

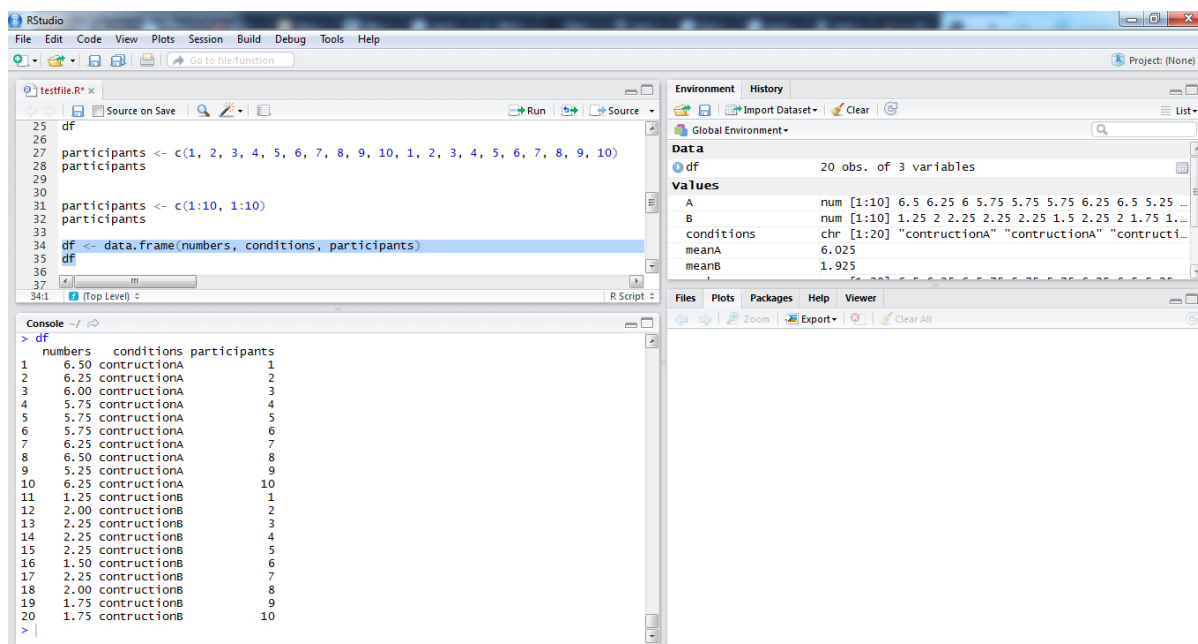


Figure 25: Our second data frame.

We will now use a function called “summarySE”. To use the function, we need an additional package that you need to install. After we have installed the package we need to tell R that we want to use it. We do this with the following code (notice that there are quotation marks in the first command but not in the second):

```
install.packages("Rmisc")
library(Rmisc)
```

We are now ready to get our confidence intervals. Let’s first look at how we get them and then talk about what we did:

```
summarySE(data=df, measurevar="numbers",
          groupvars="conditions", conf.interval=.95)
```

The `summarySE` function needs the following information: First, it needs to know with which table we are working (it's "df"). Then it needs to know our measurement ("measurevar="numbers"). Then we specify our conditions ("withinvars="conditions"). Finally, we specify the confidence level ("conf.interval=.95"). The results are pretty useful. We get the mean, the standard deviation, the standard error, and the confidence intervals:

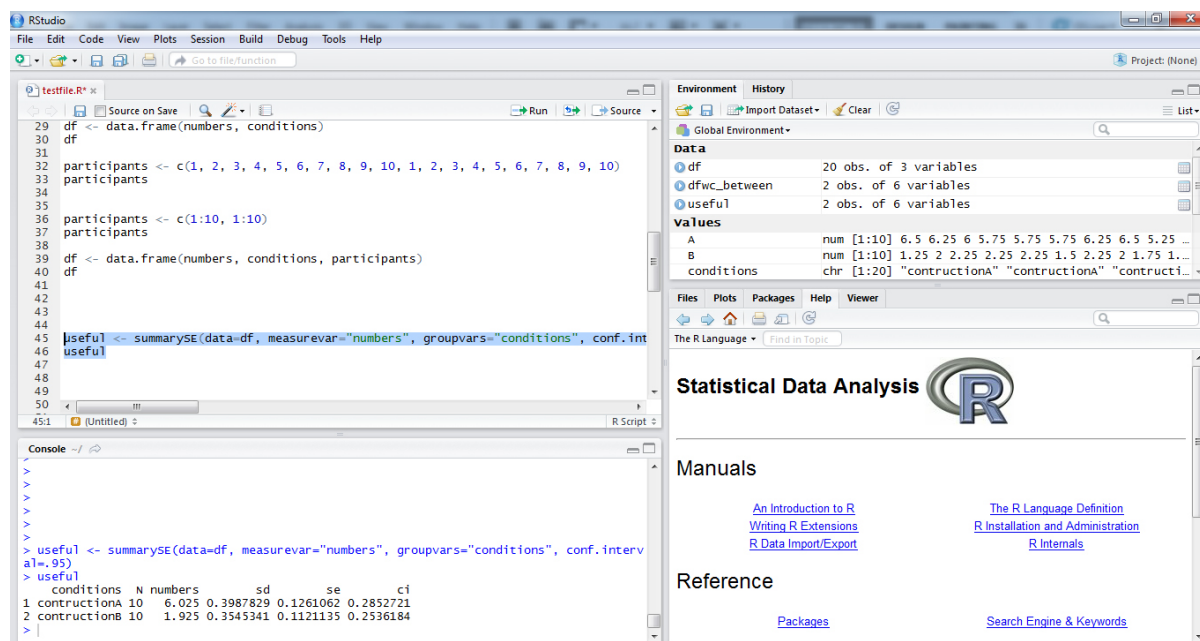


Figure 26: The `summarySE` function.

Actually, to make a nice plot of our data it is useful to give our `summarySE` function a name. Let's call it "useful":

```

useful <- summarySE(data=df, measurevar="numbers",
  groupvars="conditions", conf.interval=.95)

```

To plot this, we will need a package called "ggplot2". Again, we install the package and tell R to use it:

```

install.packages("ggplot2")
library(ggplot2)

```

We can now use a function called "ggplot". It wants to know the data frame we are

working with (it's "df" again) and which column contains the values for the x-axis and which column contains the values for the y-axis. Then we tell R that the error bars will be our confidence intervals. Additionally, we can specify the size and shape of the plot and the limits of the axis:

```
ggplot(useful, aes(x=conditions, y=numbers, group=1)) +
  geom_errorbar(width=.1, aes(ymin=numbers-ci, ymax=numbers+ci))
  +
  geom_point(shape=21, size=3, fill="white") +
  ylim(1,7)
```

I won't go into the details of what this code does. Play around with it! The resulting plot looks pretty good (you can export it in different sizes and formats by clicking on "export" right above the plot):

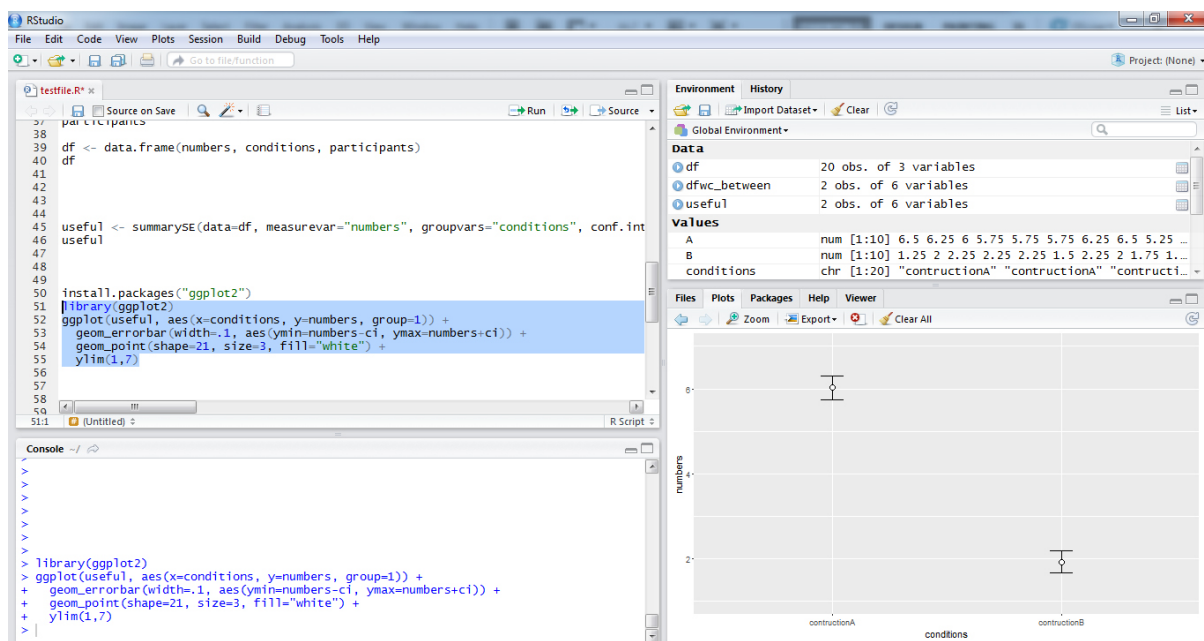


Figure 27: A plot with ggplot.

You can start playing around with ggplot. For example, you could also make a bar plot:

```
ggplot(data=useful, aes(x=conditions, y=numbers)) +
  geom_bar(stat="identity", position=position_dodge(),
  fill="steelblue")+
  >
```

```
geom_errorbar(width=.1, aes(ymin=numbers-ci, ymax=numbers+ci))
+
coord_cartesian(ylim = c(1,7)) +
theme_minimal()
```

This will produce the following output:

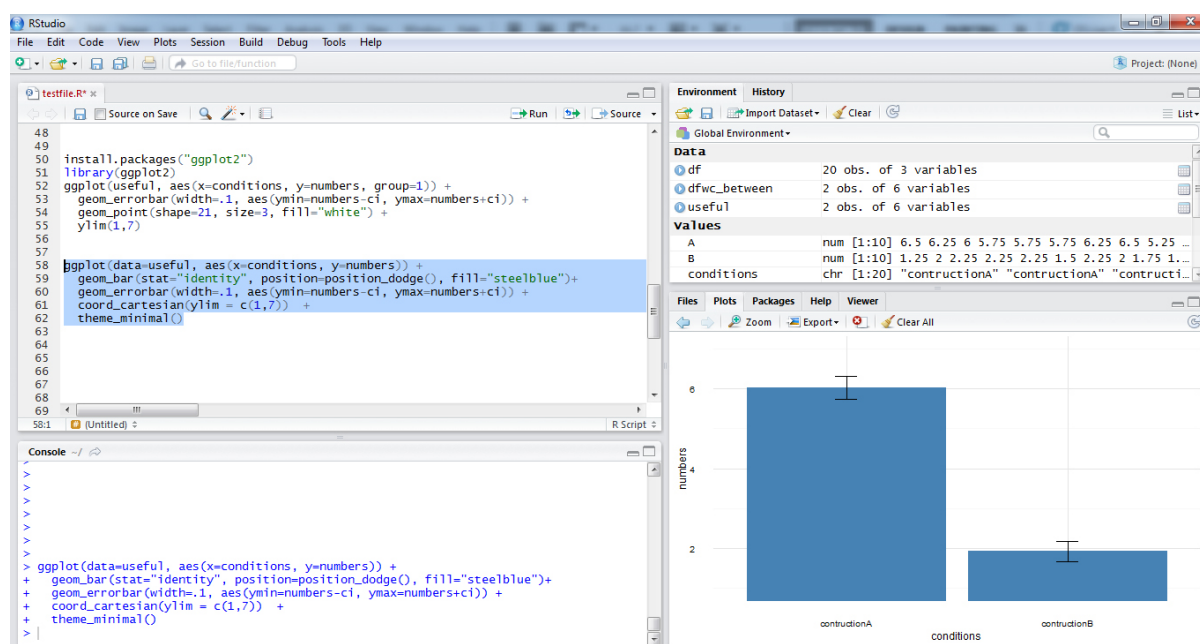


Figure 28: A bar plot with ggplot.

Now, we perform a t -test. To do a paired t -test we type in:

```
t.test(A,B,paired=TRUE)
```

The results will look like this:

Paired t -test

data: A and B

$t = 22.8405$, $df = 9$, $p\text{-value} = 2.806e-09$

alternative hypothesis: **true** difference in means is not equal
to 0


```
95 percent confidence interval:
```

```
3.69393 4.50607
```

```
sample estimates:
```

```
mean of the differences
```

```
4.1
```

As you can see, R tells you several things. First, it reminds you, what you did (a paired t -test) and what the input data was, namely the lists A and B. Then it tells you the statistics. Again, we get a p -value of 2.806e-09 as we did with OpenOffice. As you see, you also got the 95%-CIs as well as the difference between the two means.

The test we just ran was a two-sided t -test. If you want to run a one-sided test you can do this by specifying if you expect the difference to be bigger or smaller. So you either use

```
t.test(A,B,paired=TRUE,alternative="greater")
```

or

```
t.test(A,B,paired=TRUE,alternative="less")
```

That's all! R is a great tool! It just requires some work. The most useful tip is: Google is your friend! And: If you use R or some packages make sure to cite them properly.

16.3 JASP

While R is a programming language, JASP is a computer program with a graphical user interface. It was developed by the JASP Team at the Department of Psychological Methods at the University of Amsterdam and is free to use. You can download it at <https://jasp-stats.org>.

Before we take a look at JASP, we need some data. Again, I will use example 1. I created a simple table that looks like the one in Figure 29. I saved the table in the .csv format.

	A	B	C	D	E	F	G	H	I	J
1	MEAN A	MEAN B								
2	6,5	1,25								
3	6,25	2								
4	6	2,25								
5	5,75	2,25								
6	5,75	2,25								
7	5,75	1,5								
8	6,25	2,25								
9	6,5	2								
10	5,25	1,75								
11	6,25	1,75								
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										

Figure 29: A simple table.

Now you can open JASP and open this table. This will look like in Figure 30.

	MEAN A	MEAN B
1	6,5	1,25
2	6,25	2
3	6	2,25
4	5,75	2,25
5	5,75	2,25
6	5,75	1,5
7	6,25	2,25
8	6,5	2
9	5,25	1,75
10	6,25	1,75

Figure 30: Open the table in JASP.

The use of JASP is pretty straight forward! Above the table you can see different things you can do. If you click on “Descriptives” you will see the names of the two columns “MEAN A” and “MEAN B”. If you select them and click on the arrow they will appear

in the small white window that says “Variables”. On the right you will immediately see some interesting information like the mean or the standard deviation in a table. The cool thing is that you can copy the code of this table and paste it either into Word, OpenOffice, or LaTeX (under “Copy special” you’ll find the LaTeX code). I copied the LaTeX code and it looks like this:

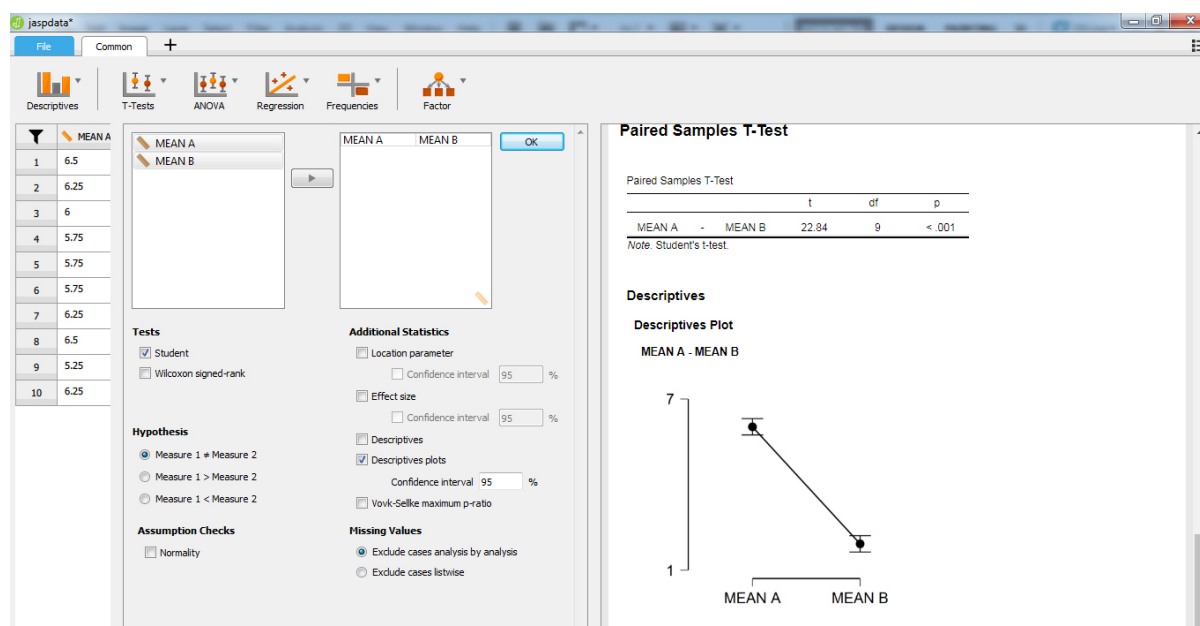
Table 1: Descriptive Statistics		
	MEAN A	MEAN B
Valid	10	10
Missing	0	0
Mean	6.025	1.925
Std. Deviation	0.3988	0.3545
Minimum	5.250	1.250
Maximum	6.500	2.250

That’s not bad! If you click on “Plots” below the variables you can select a plot you like and it will appear immediately in the window on the right. You can easily save the plots in the PNG, PDF, EPS, or TIFF format.

Now, let’s do a t -test. To do this, simply click on the “T-Test” button and choose “Paired Samples T-Test” from the drop-down menu. Again, select “MEAN A” and “MEAN B” and press the button with the arrow. There are a number of settings you can change. You can leave everything as it is, but an interesting feature is “Descriptive plots” which generates a plot with 95%-confidence intervals. Again, the results of the test occur in the window on the right. The table looks like this:

Table 2: Paired Samples T-Test					
		t	df	p	
MEAN A	- MEAN B	22.84	9	< .001	

The results look like in Figure 31. Unfortunately, JASP does not give you an exact p -value. However, That p is smaller than 0.001 might be enough information for many journals.

Figure 31: A t -test in JASP.

If you use JASP, make sure that you cite the software.

17. Multiple comparisons

So far, we have looked at very, very simple cases and only compared two constructions. Depending on your design it may well be that you want to compare more than just one construction. However, there is a problem called ‘alpha inflation’. We already know that with the alpha-level set to 0.05 we have a 5% chance to get a statistically significant result although there is no effect (type I error). In other words: Suppose there is no effect and we sample from a population 20 times and do a t -test with an alpha-level of 0.05, we will get (statistically) one significant result just due to chance. If you do multiple comparisons, the chances to conduct a type I error increases. With each additional test you do, the chances increase. With 20 comparisons, the chance to observe a significant result is already 64%! Of course, that’s very bad! So what to do?

The answer to this question is actually a very tough one and depends on your philosophy. Let’s start the discussion with noticing that there can be two different reasons for conducting multiple comparisons. The first reason would be that you are interested in comparing different conditions (or: constructions) because you have several hypotheses in mind. The second reason could be that you tested a lot of different conditions (or: constructions) and now you just want to look if there are some differences (and compare everything with everything). This is also called ‘fishing for effects’.

If the first type of situation applies to you, many statisticians would recommend not to worry about alpha-inflation and just carry out your tests. That there is a chance of making a type I error does not only apply to your tests, but to tests in general. Thus, if you conduct another study the problem of type I errors will be there again. The same is true if another researcher tries to replicate your study. In this case, the researcher will not correct the alpha level.

Again, let's make ourselves clear what is happening when you conduct multiple comparisons. Feise (2002) summarizes this nicely:

If a null hypothesis is true, a significant difference may still be observed by chance. Rarely can you have absolute proof as to which of the two hypotheses (null or alternative) is true, because you are only looking at a sample, not the whole population. Thus, you must estimate the sampling error. The chance to incorrectly declare an effect because of random error in the sample is called type I error. Standard scientific practice, which is entirely arbitrary, commonly establishes a cutoff point to distinguish statistical significance from non-significance at 0.05. By definition, this means that one test in 20 will appear to be significant when it is really coincidental. When more than one test is used, the chance of finding at least one test statistically significant due to chance and incorrectly declaring a difference increases. When 10 statistically independent tests are performed, the chance of at least one test being significant is no longer 0.05, but 0.40. To accommodate for this, the p -value of each individual test is adjusted upward to ensure that the overall risk or family-wise error rate for all tests remains 0.05. Thus, even if more than one test is done, the risk of finding a difference incorrectly significant continues to be 0.05, or one in twenty.

Now the problem with this logic is that a researcher performs many different tests in his life. Similarly, there will be a lot of tests reported in a journal. Feise (2002) remarks that, following the logic from above, a researcher should adjust his p -values in the course of his lifetime. Similarly, should a journal adjust p -values for each issue? For each year? Or for another period? These questions are hard to answer as it turns out. Additionally, decreasing the chance for type I errors will increase the chance of type II errors. That's something we don't want!

The conclusion from this discussion is the following recommendation: If you have several hypotheses in mind that you want to test and if you came up with these hypotheses before you conducted your study you should simply not care about p -value adjustment. If you are fishing for effects, however you actually should care about alpha-inflation.

When you fish for effects, you first test all the conditions (or: constructions) at once by conducting a one-way repeated measures ANOVA. If this test reveals a significant result you can adjust your p -values by using the Benjamini-Hochberg procedure (or alternatively you do a Bonferroni correction).

18. Reporting your results

There is a very simple method to report what you did called IMRAD which is the abbreviation for ‘Introduction, Methods, Results, and Discussion’. The IMRAD format is a very common structure of papers reporting empirical work. I will briefly describe how it looks but it makes sense to read some articles to familiarize yourself with this format.

In the introduction you explain what the question of your study was and what motivated it. Additionally you state your hypothesis or your hypotheses in case you have several.

In the methods section you describe the methods you used. Often times the methods section has subsections labeled ‘Materials’, ‘Procedure’, and ‘Participants’. Here you describe what materials you used, i.e., the sentences and fillers, the software you used, and what the Likert items looked like and how many lists you created with the Latin square procedure, what your instructions for the participants looked like etc. You also report who the participants were. For example, you can write: “48 native speakers of German with a mean age of 23.05 ($SD = 1.03$) participated in the study. 23 of them were female. None of them reported any language impairments. Each participant was paid 5 Euros for participation.”

In the results section you present your results. This section is not about interpretation, but here you only report your numbers. For example, you can write: “The mean rating of construction A was 6.025 ($SD = 0.3988$), the mean rating of Construction B was 1.925 ($SD = 0.3545$). A two-tailed paired t -test comparing the ratings of the two constructions revealed a statistically significant p -value of 0.000000002806; $t(9) = 22.8405$ ”. Depending on your philosophy you can also abbreviate the numbers and write “... the results were statistically significant with $p < 0.001$; $t(9) = 22.84$ ”. The numbers following the p -value is the t -statistics including the degrees of freedom. You may have wondered what the “ t ” in t -test stands for. It simply means “test”. It compares your results to a specific probability distribution called t -distribution. The general format is “ $t(\text{degrees of freedom}) = t\text{-value}$ ”. You find the degrees of freedom (often abbreviated “df” and the t -value in the output of R or JASP.

Finally, in the discussion section you discuss what your numbers mean. Here, your

interpretation comes into play. The question you now want to answer is: Do the numbers and your statistics support your hypothesis or not?

19. More on visualizing the results: box plots and beyond

We have already learned that the data gained from Likert items are, in a strict mathematical sense, ordinal data. This means, for example, that it is possible to calculate a median, but not a mean from this kind of data. However, conceptually, it has been shown that it does no harm to treat the data as being interval as discussed above. Nevertheless, there is a method to visualize ordinal data which I will discuss here in some detail as it represents a good way of visually presenting your results, namely box plots and its relatives.

You can easily create box plots with R, for example, with the `ggplot2` package we have already used in Section 16.2. You can also create box plots with Microsoft Excel, with OpenOffice, or with JASP, however, there are some limitations. But even if you are not familiar with R, it is easily possible to create professional box plots based on R with the great online tool BoxPlotR (Spitzer et al. 2014). You'll find it by browsing to <http://shiny.chemgrid.org/boxplotr/>. On the website you can simply paste in your data and view and download your box plots (in the .pdf, .eps, or .svg format).

Let's look at box plots and its relatives by using the data from example 3 above. For matters of simplicity, I use the means of the construction. In real life, however, you would use all the original ratings participants gave. I simply copied the data from a table and pasted it into the website. This is how the result looks:

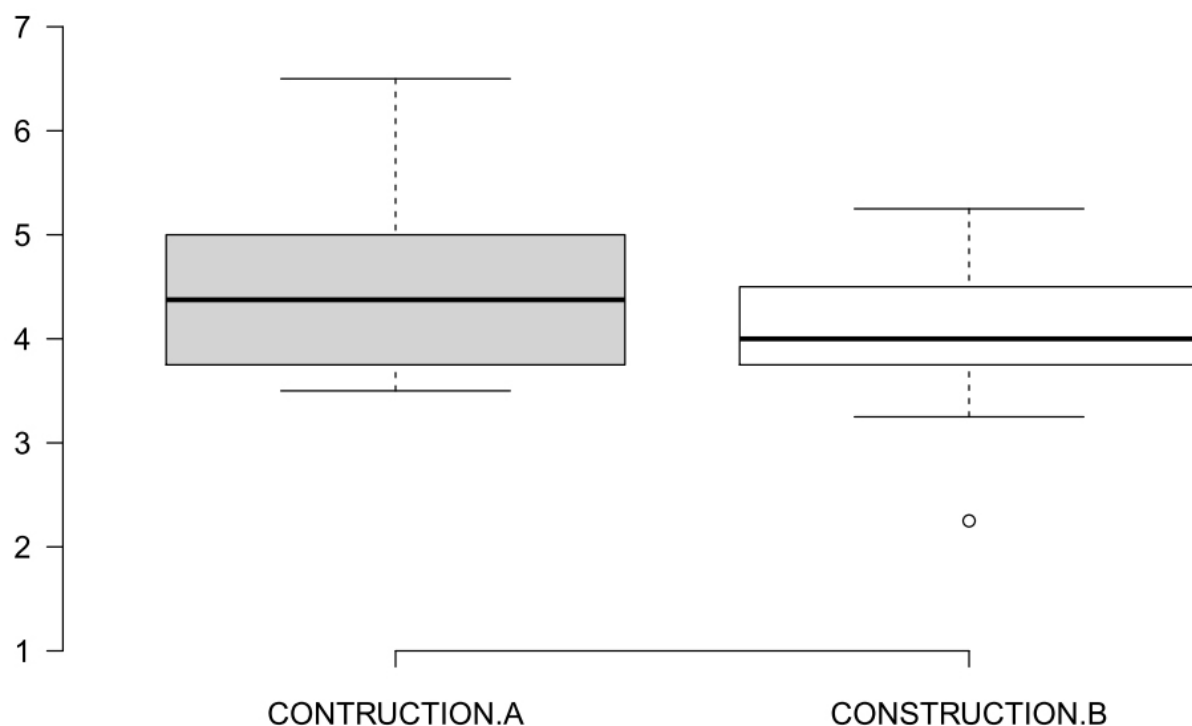


Figure 32: A simple box plot created with BoxPlotR.

Remember that the mean of construction A in example 3 was 4.575 and the mean of construction B was 4. Box plots, however, do not use means, but rather the median. Remember that the median is the middle value of a distribution, i. e., above the median there are 50 % of the data and below the median there are, of course, the other 50 % of the values. It thus cuts the data in half. The median is represented by the bold horizontal lines in Figure 32 (the median of construction A is 4.38 and the median of construction B is 4). If you are not used to box plots you have to concentrate a little: Around the median, there is a box. The upper limit of the box is called the upper quartile. Above the upper quartile, there are 25% of the data, below this line, there are 75 % of the data. The lower limit of the box is called the lower quartile. Below the lower quartile there are 25 % of the data, above the lower quartile there are 75 % of the values. From this, it follows, that the bold line in the middle (representing the median) does not only divide the data in a way that 50 % of the values lie above and 50 % lie below it but this also means that 50 % of the values lie inside of the box. This part is called the interquartile range, or IQR, for short. Stare at Figure 33 for a few seconds to understand this (while ignoring the parts I have not talked about so far).

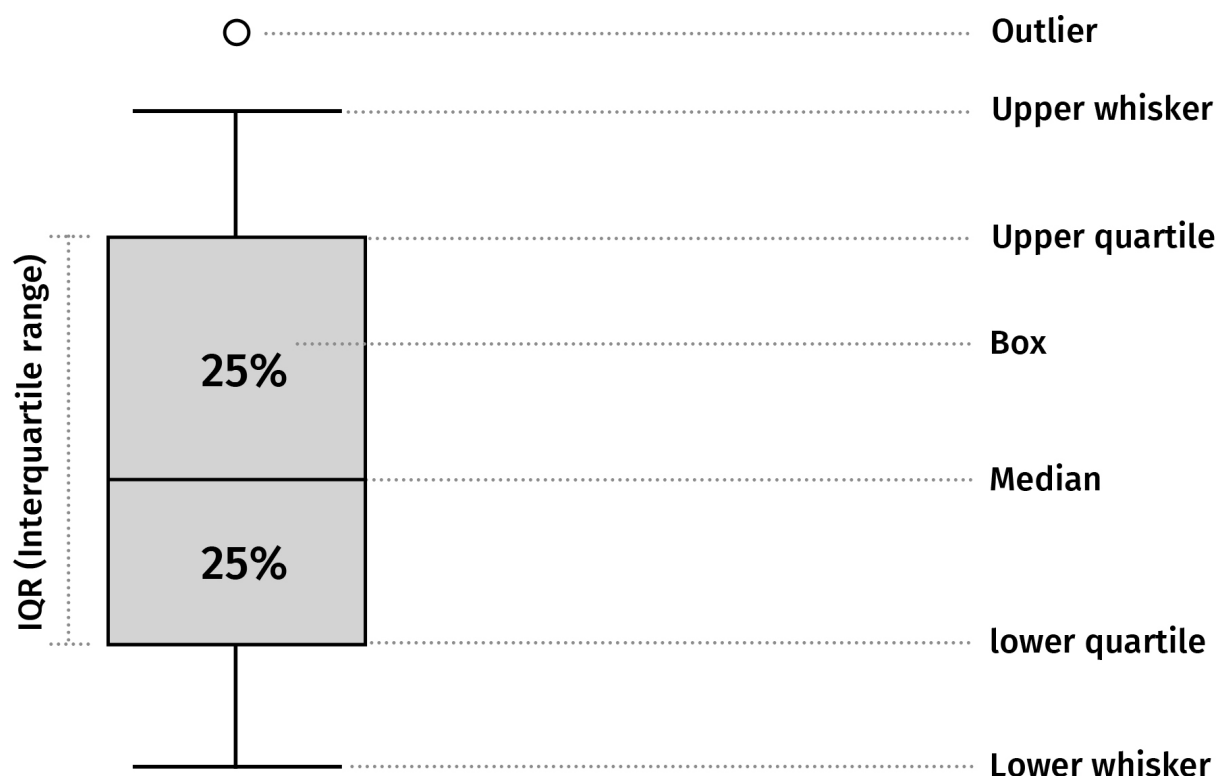


Figure 33: The basic make-up of a boxplot.

A box plot is a great way to visualize the spread of data. The smaller the box, the closer the values spread around the median. Additionally, the box has whiskers (the antennae around the box). As with error bars, there are different definitions of the whiskers, so you have to describe what the whiskers show in the caption. One definition is that the whiskers show the maximal values. This kind of whiskers is called “Spear type whiskers”. However, this is not so common. This is also not the definition used in the examples above. You can tell this from the fact that both examples include outliers which are extreme values that are not captured by the rest of the graphic. Outliers are extreme and rare values. Instead, the whiskers often extend up to 1.5 times the interquartile range away from lower and upper quartile. This kind of whiskers is called “Tukey style whiskers”. Finally, for data sets that are greater than 40 (i. e., $n > 40$), whiskers sometimes spread from the 5th to the 95th percentile. This kind of whiskers is called “Altman style whiskers”. Only with Tukey style and Altman style are there outliers. Outliers are simply values that lie outside of the whisker extent. Take a look at Figure 34 and read the definitions again to understand what is going on. With BoxPlotR you can easily change the definition of the whisker extends.

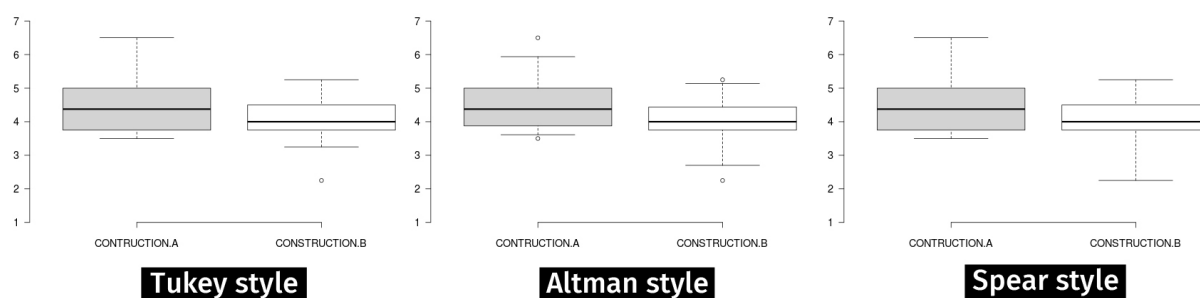


Figure 34: The three basic definitions of whisker extends.

The cool thing with BoxPlotR is that you can just click on an option and see the results immediately. In Figure 35 you see my favorite way of using box plots. There are three changes as opposed to a regular box plot. First, there is a black cross in each box. The crosses represent the means. The gray boxes around the crosses represent the 95 %-confidence intervals of the means. Additionally, there are notches in the box plots. The notches also represent 95 %-confidence intervals, however, not the confidence intervals of the means, but the confidence intervals of the medians. When the notches do not overlap, there is a very great chance that there actually is a difference between your groups.

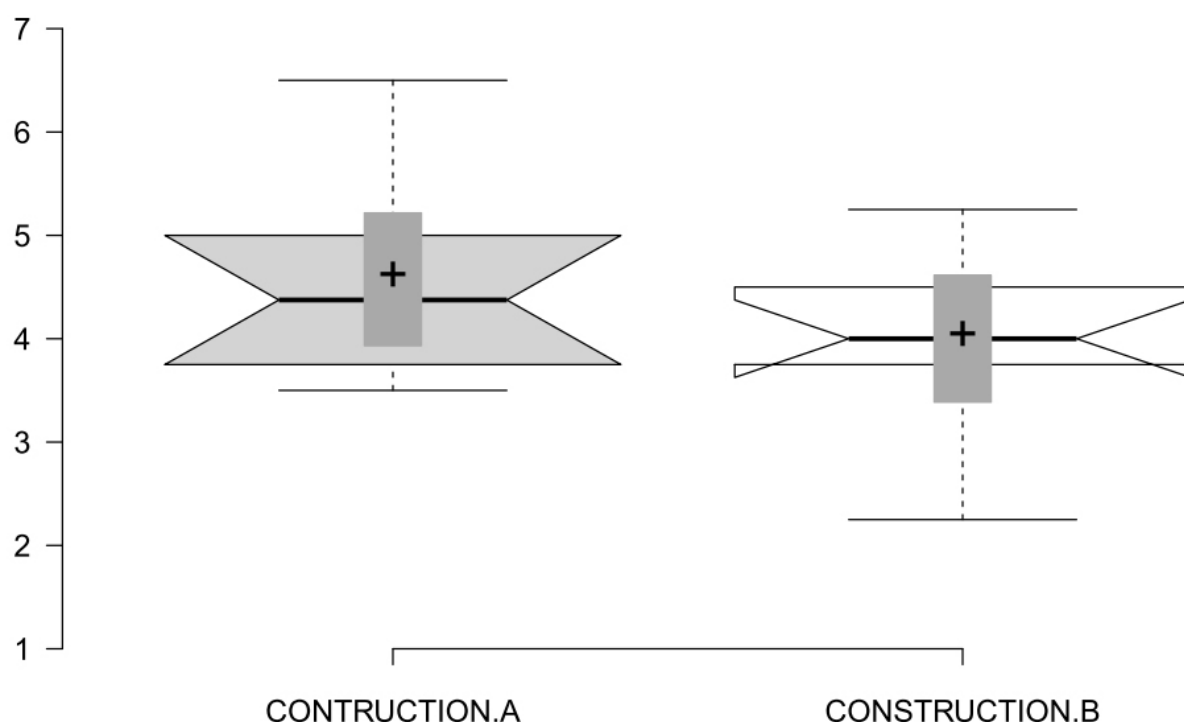


Figure 35: The black crosses show the means, the gray boxes around the means represent the 95 %-confidence interval of the means. The notches represent the 95 %-confidence interval of the medians.

I will briefly show you two other ways of representing your data which are similar to box plots. Their strength is that they visually represent the spread of your data in an very intuitive way. They are a good choice if there are two groups of people who react differently to a construction. For example, if there are people who like your construction and give it a good rating and people who do not like your construction, you can use violin plots or bean plots. To illustrate the advantages of violin and bean plots, I invented some ratings for a construction C. Some people liked the construction, others disliked it. The (mean) ratings participants gave are: 2, 2, 3, 2, 1, 7, 6, 7, 7, 7. Now take a look at Figure 36. As you can see, a box plot shows you that the data is spread out more, but the internal distribution is hidden inside the box. Violin plots show you a little bit more, but the bean plot directly tells you that there are two groups differing in their rating behavior.

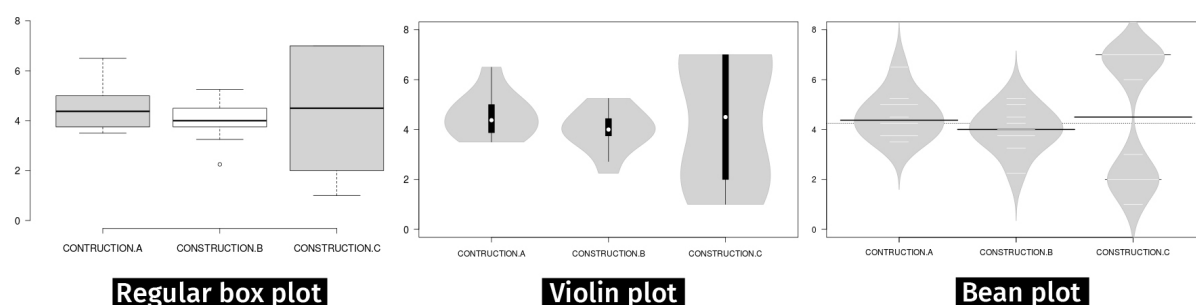


Figure 36: Regular box plot, violin plot, and bean plot.

By the way, the bold black line inside the violin plot is the interquartile range and the white dot indicates the median. The median in the bean plot is indicated by the black horizontal line. If you want to learn more about visualizing your results using box plots I recommend reading Wickham & Stryjewski (2011) and Krzywinski & Altman (2014). The latter paper is, by the way, part of a cool series called “Points of Significance” which gives short introductory overviews of important statistical concepts. I really recommend reading the whole series!

That’s it!

References

- Anderson, F. S., Kelly, K. & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, 1–16.
- Bard, E. G., Robertson, D. & Sorace, A. (1996). Magnitude estimation of linguistic

- acceptability. *Language*, 72, 32–68.
- Birkel, P. & Birkel, C. (2002). Wie einzig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss. *Psychologie in Erziehung und Unterricht*, 49, 219–224.
- Bolinger, D. L. (1968). Judgments of grammaticality. *Lingua*, 21, 34–40.
- Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin*, 57(1), 49–64.
- Buckingham, A. & Saunders, P. (2008). *The Survey Methods Workbook*. Malden: Polity.
- Carden, G. (1976). Syntactic and semantic data. Replication results. *Language in Society*, 5(1), 99–104.
- Carifio, J. & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistens myths and urban legends about Likert scales, Likert response format and their antidotes. *Journal of Social Sciences*, 3(3), 106–116.
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge: MIT Press.
- Christensen, L. (2012). Types of Designs Using Random Assignment. In: Cooper, H. (ed.): *APA handbook of research methods in psychology*. Washington, D.C.: APA Press. 469–488.
- Cohen, J. (1962). The statistical power of abnormal social psychological ressearch: A review. *Journal of Abnormal and Social Psychology*, 65, 145–153.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2nd edition. Hillsdale: Erlbaum.
- Colerus, E. (2013). *Mathematics for Everyman. From Simple Numbers to the Calculus*. Mineola, New York: Dover. Originally published as: Colerus, E. (1942): *Vom Einmaleins zum Integral. Mathematik für jedermann*. Vienna: Bischoff. Note: This is a popular science book, but it is a classic in the German speaking world. If you are interested in mathematics, this is an easy to read choice.
- Cowart, W. (1997). *Experimental Syntax. Applying Objective Methods to Sentence Judgments*. Thousand Oaks, London & New Delhi: Sage.
- Cumming, G. (2012). *Understanding The New Statistics. Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York & London: Routledge.
- Cumming, G. (2013). The new statistics. Why and how. *Psychological Science*, 25(1), 7–29.
- Cumming, G., & Finch, S. (2005). Inference by eye. Confidence intervals and how to read pictures of data. *American Psychologist*, 60(2), 170–180.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 93–96.
- de Winter, J. C. F. & Dodou, D. (2010). Five-Point Likert Items: *t* test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research and Evaluation*, 15(11), 1–16.
- Endresen, A. & Janda, L. A. (2017). Five statistical models for Likert-type experimen-

- tal data on acceptability judgments. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 3(2), 217–250.
- Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191.
- Featherston, S. (2007). Data in generative grammar. The stick and the carrot. *Theoretical Linguistics*, 33, 269–318.
- Feise, R. J. (2002). Do multiple outcome measures require p-value adjustment? *BMC Medical Research Methodology*, 2(1), 8.
- Finstad, K. (2010). Response Interpolation and Scale Sensitivity: Evidence Against 5-Point Scales. *Journal of Usability Studies*, 5(3), 104–110.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.
- Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver & Boyd.
- Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist*, 8, 750–751.
- Gibson, E. & Fedorenko, E. (2010a). Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences*, 14, 233–234.
- Gibson, E. & Fedorenko, E. (2010b). The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, 28(1–2), 88–124.
- Gibson, E., Piantadosi, S. & Fedorenko, K. (2011). Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass*, 5(8), 509–524.
- Gibson, E., Piantadosi, S. T. & Fedorenko, E. (2013). Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida (2013). *Language and Cognitive Processes*, 28(3), 229–240.
- Greenbaum, S. (1973). Informant elicitation of data on syntactic variation. *Lingua*, 31, 201–212.
- Greenbaum, S. (1976). Contextual influence on acceptability judgments. *Linguistics*, 187, 5–11.
- Greenbaum, S. & Quirk, R. (1970). *Elicitation Experiments in English*. Linguistic Studies in Use and Attitude. Coral Gables: University of Miami Press.
- Heringer, J. T. (1970). Research on quantifier-negative idiolects. In: *Papers From the Sixth Regional Meeting, Chicago Linguistic Society*, 287–295.
- Hill, A. A. (1961). Grammaticality. *Word*, 17, 61–73.
- JASP Team (2018). JASP (Version 0.9)[Computer software]. Online: <https://jasp-stats.org/>.
- Krzywinski, M. & Altman, N. (2014). Visualizing samples with box plots. *Nature Methods*, 11, 119–120.

- Labov, W. (1972). Some principles of linguistic methodology. *Language in Society*, 1, 97–120.
- Lewis, J. R. (1993). Multipoint scales: Mean and median differences and observed significance levels. *International Journal of Human-Computer Interaction*, 5(4), 383–392.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 55.
- Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research. Causes, consequences, and remedies. *Psychological Methods*, 9, 147–163.
- McIver, J. P. & Carmines, E. G. (1983). *Unidimensional scaling*. Beverly Hills: Sage.
- Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *The British Journal of Psychology*, 88, 355–383.
- Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Sciences Education*, 15, 625–632.
- Núñez, R. (2007). Inferential Statistics in the Context of Empirical Cognitive Linguistics. In: González-Márquez, M., Mittelberg, I., Coulson, S. & Spivey, M. (eds.): *Methods in Cognitive Linguistics*. Philadelphia: John Benjamins. 87–118.
- Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.
- Oppenheim, A. N. (1992). *Questionnaire design, interviewing, and attitude measurement*. New York: Printer.
- Pearson, E. S. (1931). The analysis of variance in the case of non-normal variation. *Biometrika*, 23(1/2), 114–133.
- Pell, G. (2005). Use and misuse of Likert scales. *Medical Education*, 39(9), 970.
- Popper, K. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.
- Preston, C. C. & Colman, A. M. (2000). Optimal number of response categories in rating scales. Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104(1), 1–15.
- Salsburg, D. (2001). *The Lady Tasting Tea. How Statistics Revolutionized Science in the Twentieth Century*. New York: Holt.
- Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309–316.
- Schütze, C. T. (2016). *The Empirical Base of Linguistics. Grammaticality Judgments and Linguistic Methodology*. Berlin: Language Science Press.
- Schütze, C. D. & Sprouse, J. (2013). Judgment data. In: Podeswa, R. J. & Sharma, D. (eds.): *Research Methods in Linguistics*. Cambridge: Cambridge University Press.
- Spitzer, M., Wildenhain, J., Rappsilber, J. & Tyers, M. (2014). BoxPlotR: a web tool for generation of box plots. *Nature Methods*, 11, 121–122.
- Sprouse, J. (2011a). A test of the cognitive assumptions of magnitude estimation. Commutativity does not hold for acceptability judgments. *Language*, 87, 274–288.
- Sprouse, J. (2011b). A validation of Amazon Mechanical Turk for the collection of ac-

- ceptability judgments in linguistic theory. *Behavior Research Methods*, 43, 155–167.
- Sprouse, J. & Almeida, D. (2012). Power in acceptability judgment experiments and the reliability of data in syntax (unpublished manuscript). University of California, Irvine & Michigan State University.
- Sprouse, J. & Almeida, D. (2017). Design sensitivity and statistical power in acceptability judgment experiments. *Glossa*, 2(1), 1–32.
- Sprouse, J., Wagers, M. W. & Phillips, C. (2013). Deriving competing predictions from grammatical approaches and reductionist approaches to island effects. In: Sprouse, J. & Hornstein, N. (eds.): *Experimental Syntax and Island Effects*. Cambridge: Cambridge University Press, 21–41.
- Stevens, S. S. (1946): On the Theory of Measurement. *Science*, 103, 677–680.
- Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In: Stevens, S. S. (ed.): *Handbook of experimental psychology*. New York: John Wiley.
- Tavakol, M. & Dennick, R. (2011). Making sense of Cronbach’s alpha. *International Journal of Medical Education*, 2, 53–55.
- Velleman, P. F. & Wilkinson, L. (1993): Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading. *The American Statistician*, 47(1), 65–72.
- Weskott, T. & Fanselow, G. (2011). On the informativity of different measures of linguist acceptability. *Language*, 87(2), 249–273.
- Wickham, H. & Stryjewski, L. (2011). 40 years of box plots. Online: <http://vita.had.co.nz/papers/boxplots.pdf>.
- Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, 9(2, 2), 1–27.