Statistical Inference for Everybody (and a Linguist)

Roland Schäfer, ...

Textbooks in Language Sciences

language science press

Textbooks in Language Sciences

In this series:

1. Müller, Stefan. Grammatical theory: From transformational grammar to constraint-based approaches.
2. Schäfer, Roland. Einführung in die grammatische Beschreibung des Deutschen.
3. Freitas, Maria João & Ana Lúcia Santos (eds.). Aquisição de língua materna e não materna: Questões gerais e dados do português.
4. Roussarie, Laurent. Sémantique formelle: Introduction à la grammaire de Montague.
5. Kroeger, Paul. Analyzing meaning: An introduction to semantics and pragmatics.
6. Ferreira, Marcelo. Curso de semântica formal.
7. Stefanowitsch, Anatol. Corpus linguistics: A guide to the methodology.
8. Müller, Stefan. Chinese fonts for TBLS 8 not loaded! Please set the option tblseight in main.tex for final production.
9. Kahane, Sylvain & Kim Gerdes. Syntaxe théorique et formelle. Vol. 1: Modélisation, unités, structures.

ISSN: 2364-6209

Statistical Inference for Everybody (and a Linguist)

Roland Schäfer, ...

language science press

… Roland Schäfer. 2025. *Statistical Inference for Everybody (and a Linguist)* (Textbooks in Language Sciences). Berlin: Language Science Press.

This title can be downloaded at:
http://langsci-press.org/catalog/book/

ISBN: (Digital)
(Hardcover)
(Softcover)
ISSN: 2364-6209
DOI:
Source code available from www.github.com/langsci/
Errata: paperhive.org/documents/remote?type=langsci&id=

Cover and concept of design: Ulrike Harbort
Fonts: Libertinus, Arimo, DejaVu Sans Mono
Typesetting software: XeLaTeX

Language Science Press
Grünberger Str. 16
10243 Berlin, Germany
http://langsci-press.org

Storage and cataloguing done by FU Berlin

Roland Schäfer in a picture that means that readers will have to calculate everything by hand.

Contents

1 Scientific Inference and Error

population
sample
inference
Texas Marksman
validity
study design
pre-registration
meta-analyses
replication
open science

population
sample
inference
Texas Marksman
validity
study design
pre-registration
meta-analyses
replication
open science

2 Inference: Successes and Failures

Overview

In this chapter, we introduce the notion of frequentist probability. It's the probability of a certain event under the assumption of pure randomness. Pure randomness is not a very precise technical term, but we hope it gives the right impression. For example, frequentist probability of rolling six pips with a fair die is $1 \div 6$ (or roughly 0.17) before you roll the die. This also implies that the proportion of rolls of a fair die showing six pips in a very long or endless sequence of rolls will be $1 \div 6 = 0.17$ (or 17%). Notice that after rolling the die, it either shows six pips or it doesn't show six pips, and there is no longer a probability attached to the result. It becomes a fact. We apply reasoning around frequentist probability to lotteries, parlour games with self-proclaimed psychics, tests of alleged tea gourmetism, and corpus studies. It is shown that unexpected outcomes under the assumption of pure randomness are those that have a low frequentist probability. In other words, those outcomes would be rare if we repeated the experiment over and over again, and if only pure randomness were in control of the outcomes (rather than some influencing factor, such as the manipulation of a die). To quantify these notions and make some careful and limited scientific inferences based on this quantification, we introduce several mathematical concepts. The binominal coefficient is a way of calculating the number of ways of choosing a specific number of elements from a set of elements regardless of their order and without replacement. The p-value calculated in Fisher's Exact Test is a metric that reconstructs the pre-experiment frequentist probability of obtaining a certain result or a more extreme result in simple experiments involving two two-valued variables and counts of their instances. The effect strength for Fisher's Exact Test is a measure of how strongly two variables interact. Finally, probabilty density functions and cumulative distribution functions are general functions to calculate (and plot) such probabilities.

Readers should be warned that this chapter is a tad narrative, slow, and maybe even repetitive, which is necessary to give readers an insight into the basic ideas of frequentist inference. We consider this vital in order to remedy problems with frequentist practice (not with the frequentist philosophy). Take your time and

rest assured that we will up the pace in later chapters. Also, we promise that we will never talk about tea again after this chapter.

Problem Statement: If there's Nothing Going On …

Let's consider three rather simple questions: (i) You know that you aren't prescient, but you decide to play the lottery anyway. How surprised would you be if you won the big prize? (ii) You don't believe that your friend, who claims to be a psychic, actually has psychic abilities. Nevertheless, you give them a chance and invite them to a party where they have to guess the phone numbers of all other guests. How surprised would you be if they guessed the phone numbers of all your other guests correctly? (iii) Given that most grammatical theories (which have something to say about passives) claim that the verb *sleep* cannot be passivised at all or only under very marginal circumstances, how surprised would you be to find ten passives of *sleep* in a corpus of English? Please think thoroughly about your answers to these questions before continuing on.

2.1 Unexpected Outcomes

Did you think about the questions from the Problem Statement? Really? Good. Here's one possible discussion. Most importantly, the questions from the Problem Statement cannot be answered properly in their given form, simply because there are significant data missing. As for (i), the question doesn't specify what kind of lottery we're considering. Is it a simple urn at a funfair from which you get to draw one out of a thousand lots, and only one of the thousand lots is a win? Or is it the Eurojackpot, where you have to guess five numbers out of fifty (plus some additional single numbers) correctly to win the big prize? Most likely, you either decided that you can't answer the question, or you answered it with respect to some specific type of lottery by way of example. Maybe you even wondered whether the lottery is supposed to be fair or not. However, when presented with this specific example, most people typically don't worry too much about how the lottery was conducted and whether it was fair. At least with big national lotteries, they tend to put trust in there being sufficient oversight and the draw being—here it comes—properly random. Most importantly (and disappointingly),

they see no way to rigging the lottery in their own favour. Considering the urn at the funfair, people likely assume that it's rigged anyway, but they don't care (at least in the Free World).[1]

The scenario in (ii) is similar, and there's also relevant information missing. You probably decided whether your degree of surprise would critically depend on the number of guests at the party and the number of digits phone numbers have. In my youth, smaller German villages (like Twiste, located in the Twistetal district) still had three-digit phone numbers, for example. If the local Twiste psychic only had to guess one such phone number, guessing that number correctly even without psychic powers would be much less awe-inspiring than guessing the ten-digit U.S. phone numbers of 28 guests at a party in NYC with the same accuracy, for example. Furthermore—and most likely because it involves a psychic—this scenario makes people much more suspicious of whether and how it was ensured that the psychic didn't cheat. Maybe they have a secret app that exploits a vulnerability in close-by mobile phones, and they simply read the numbers off of peoples' phones. Maybe the party was announced in a group chat on some messenger app, and they tracked all the guests' numbers down in the app before the party. Maybe the host or some other guest conspired with the psychic and gave them all the numbers, either as a practical joke or even because they want to get people to pay for the psychic's services (tracking down dead ancestors who lived as maids and servants at the court of Louis XIV of France).

Example (iii) is much more intricate and, in a way, boring, which is why it only intrigues linguists. Some linguists would smirk at you and claim that they don't care about corpus examples because it was determined once and for all by a cherubic figure (who never laughs) that examples from corpora don't count for anything. Some linguists, on the other hand, would take the ten sentences as conclusive evidence that whatever random modification to their theory they came up with is provably correct, or that somebody else's theory is provably incorrect.[2] What were your thoughts? We certainly hope that you don't belong to either of the aforementioned tribes of linguists and that you saw the parallels to the first two scenarios. Above all, quantitative considerations play a role, among others: How large is the corpus? How often does *sleep* occur in the corpus, regardless of its voice? How many active and passive verbs occur in the corpus? Also, the question of whether it was a fair draw is vastly more complicated than

[1] *It doesn't get more American than this, my friend. Fatty foods, ugly decadence, rigged games.* (Murray Bauman, Episode 7 of Stranger Things 3)

[2] *Whenever I find even one example that contradicts a claim, I consider that claim refuted.* (an unnamed linguist, p. c.)

in the case of a lottery. For example, is it a corpus of language produced by native speakers, children, L2 learners of English, frontier large language models, or even some cute language bot from 1998? Does the composition matter considering your research question? What exactly is your research question? Finally, the underlying theory from which it allegedly follows that *sleep* cannot be passivised needs further inspection. Does it also exclude the figura etymologica for such unaccusative verbs? After all, maybe all ten sentences are instances of silliness such as (1). Would the result still count as unexpected, regardless of any quantitative evaluation?

(1) The sleep of Evil has been slept by many a demon.

It's a muddle! Therefore, we'll use a simple non-linguistic example in Section 2.2 to introduce some important statistical concepts that concern the numerical side of this muddle. The example is about tea, and it's extremely famous, so anyone interested in statistical inference should be aware of it, even if they're not in Tea Studies. Then, we'll return to the questions from the problem statement (except question [ii], which is deferred to the exercises).

2.2 Tea and Milk

2.2.1 An Introduction to Unexpected Outcomes

What unites the examples in the Problem Statement is that they describe a confrontation with chance. Given this confrontation, you're asked what kind of a result would be unexpected under the assumption that there is nothing going on: (i) you're not prescient, or (ii) the psychic isn't actually a psychic. Matters are more complex for the corpus example (iii), and we'll return to it later (for example in Section 2.3.2). In this section, we formalise the notion of *unexpected outcome* in relation to such situations and experiments.

unexpected outcome

First of all, an *unexpected outcome* cannot be one which is deemed totally impossible. If it were impossible to win the lottery, you wouldn't play it. If it had already been established (for example at previous parties) that your psychic friend couldn't guess phone numbers, you wouldn't humiliate them and ask them to guess numbers at your party. Finally, if it were established beyond reasonable doubt that *sleep* cannot be passivised, you wouldn't bother to do a corpus search for passivised forms of that verb. Clearly, unexpected outcomes are not miracles where everything we know about the world is up for debate.

What we usually mean when we deem an outcome *unexpected* is that it had a very slim chance—a low probability—of occurring before we made it occur.

Mathematically, the most straightforward case is the one with the urn at the funfair. If there are a thousand lots in the urn, one of them is a win, all the others are duds, and you draw exactly one, most people have an intuitive understanding that you have a *chance of one in a thousand* (or 1:1000) to win. Usually, it is also understood that this means that if you played this game over and over again, you would end up winning in one of a thousand rounds on average. (Playing the game over and over again, each time with a fresh urn of one thousand lots, not gradually emptying one and the same urn, of course.) That's why playing it once and winning is unexpected or surprising: Winning is a rare event given the way the urn was set up (one winning lot and 999 duds). The maths is slightly more complex for the Eurojackpot because you have to choose five numbers out of fifty and not one lot out of a thousand, but it essentially follows the same logic. For the psychic guessing phone numbers, the idea is also the same once the number of phone numbers and the number of the digits per phone number has been determined. We will return to the third scenario (the corpus study) later, but we can apply a similar logic even to that example.

2.2.2 The Total Number of Possible Outcomes

In each of the scenarios, we need to know the number of potential outcomes in order to quantify how unexpected a single specific outcome is. The higher the number of overall possible outcomes, the more unexpected a specific outcome is. A seminal application of this idea to scientific reasoning is reported in Fisher (1935), and we'll introduce it here before applying the same reasoning to the scenarios from the Problem Statement. In that book, Ronald A. Fisher reports an event where Muriel Bristow, herself a scientist, claimed that she could taste whether the milk or the tea was poured into a cup first. While it is not impossible that some physical properties of the mixed liquids differ depending on their order of being poured into the cup, some doubt was in order. Therefore, Fisher devised an experiment to shed light on the substance of Bristow's claim. She was presented with eight cups, four tea-first cups and four milk-first cups. Otherwise, the cups were identical. Her task in the experiment was to find the four tea-first cups merely by tasting. Very much like winning a lottery after buying just a single lot, some outcomes of this experiment might surprise us by being relatively unexpected if Bristow didn't have the ability she claims to have. We still wouldn't consider it proven beyond all doubt that she does indeed have the ability if that happened. However, we'd at least not consider her claims of being a tea expert refuted if she guessed a surprising number of cups correctly. The question is:

What's a surprising number? How many cups does she have to classify correctly for us to call it an unexpected outcome?

Statistics doesn't offer a final answer to this question. However, it provides the maths upon which we need to base our answer. Remember that Muriel Bristow has to choose four cups out of eight, and we first need to calculate how many distinct sets of fours cups out of eight she could potentially choose, without even considering whether she chose the right ones. Let's do it. In Figure 2.1, we illustrate the 8 cups.[3] While they would all look exactly the same in the real experiment, we've made it easier to follow the argument by showing the tea-first cups with steam and the milk-first ones without steam. Furthermore, we've coloured the cups to make them identifiable individually. There is one grey, one red, one blue, and one green cup for each of the conditions (milk-first or tea-first). Again, in the real experiment, cups should have the exact same physical properties.

Figure 2.1: Four tea-first cups (steaming) and four milk-first cups (not steaming) for Muriel Bristow to choose from; in the actual experiment, they'd all look exactly the same (without colours).

Before choosing her first cup, Ms Bristow obviously has 8 choices. She could pick the red steaming cup, the grey steaming cup, the grey non-steaming cup, etc. Figure 2.2 shows the situation after an arbitrary first cup was chosen.

Figure 2.2: That's the first cup chosen! Choices left: 7.

Before she continues on and picks the second cup, only 7 choices are still available. Notably, for each of the 8 distinct choices she had in the beginning, she now has 7 distinct subsequent choices. In the illustration, she chose the blue steaming

[3] We switch to writing numerals using arabic numbers for clarity, even if that violates one or two style guides.

cup and has the other 7 cups still available. Had she chosen the red steaming cup instead in the first step, she'd now be confronted with a different set of 7 options. That means that after picking another cup (Figure 2.3), she has already decided on one specific choice from among $8 \cdot 7 = 56$ possible choices.[4] Put differently, she has taken 1 out of 56 possible decision paths to choose 2 out of 8 cups.

Figure 2.3: After another cup was chosen, there are now 6 choices left.

The story goes on in a similar vein. Let's assume she chooses the red steaming cup as her third pick, as in Figure 2.4. Now, she has made 1 out of $8 \cdot 7 \cdot 6 = 336$ possible choices, since for each of the $8 \cdot 7$ options left over after her previous decision, she had 6 distinct choices available.

Figure 2.4: As Ms Bristow picks another cup, we're down to 5 choices.

To make things more interesting for later calculations, she now makes her first incorrect guess. She picks the green non-steaming one, which is a milk-first cup. She has now decided on 1 specific configuration from $8 \cdot 7 \cdot 6 \cdot 5 = 1680$ possible configurations. Or has she? As we mentioned in Footnote 4, there's a catch.

Figure 2.5: Ms Bristow has picked her last cup.

[4] This is not exactly true. There's a catch to which we'll return presently. Do you remember from secondary school maths what it is?

First, she chose the blue steaming cup, then the grey steaming cup, then the red steaming cup, and finally the the green non-steaming cup. But is this really the only way to arrive at the same result? In the first step, she had 8 options, and she chose the blue steaming cup, leaving her with 7 choices, etc. She could have chosen the red steaming one and still arrived at the same end result via a different path. For example, she could have chosen the red steaming cup first, then the grey steaming cup, followed by the blue steaming one, and finally the green non-steaming one. Put in quantitative terms, there are groups within the set of 1680 decision chains that yield identical results, at least if the order in which the cups were selected is irrelevant. And for the purpose of this experiment, the order is indeed irrelevant.

How do we know how many of the 1680 decision paths lead to identical results? Well, how many different ways of ordering 4 cups are there? Imagine you had to put 4 cups on a table from left to right one by one. In the first step, you can choose from among the 4 cups. For each of these 4 distinct choices, there are 3 distinct subsequent choices because you'll have 3 cups left. Then there are 2 choices each, then just 1. Hence, the groups of identical outcomes should each have a size of $4 \cdot 3 \cdot 2 \cdot 1 = 24$. There are thus

$$\frac{8 \cdot 7 \cdot 6 \cdot 5}{4 \cdot 3 \cdot 2 \cdot 1} = \frac{1680}{24} = 70$$

truly distinct sets of 4 cups to be chosen from among a set of 8 cups.[5]

This just gives us the number of distinct sets of four cups we can choose from 8 cups. So, how unexpected is her performance (3 cups detected correctly) given that there are 70 different ways of choosing 4 cups out of 8? This would have been easier to quantify if she had guessed all 4 cups correctly. Clearly, there is only 1 such immaculate result. Had Ms Bristow chosen the green steaming cup instead of the green non-steaming cup, it would have been the immaculate guess. Obviously, by merely guessing (without any special sensory ability), Muriel Bristow would produce such a perfect result in 1 out of 70 runs if the experiment were repeated over and over again. Put differently, there is a 1:70 chance of guessing the four cups by mere luck. In technical terms, the *frequentist probability* of hitting the tea jackpot by uninformed guessing is $1 \div 70 \approx 0.014$. This probability is sometimes converted to a percentage, in this case 1.4%.[6] Would this be a highly

frequentist probability

[5] And for those who are beginning to remember their elementary stochastics: We get different results if the order is not irrelevant and if a cup can be chosen more than once (replacement).

[6] From our perspective, this conversion to a percentage is not at all wrong inasmuch as 1.4% of an endless sequence of tries would result in an immaculate result, even if the taster is really just guessing. However, in scientific contexts, probabilities are expressed properly as numbers between 0 and 1, not as percentages.

unexpected result? So unexpected maybe that you'd doubt that Muriel Bristow merely got lucky? Well, you tell us!

2.2.3 In-Depth Probability

Before we continue on to calculate the probability of outcomes that are not perfect—such as 3 correctly classified cups—let's stop and briefly revise our notion of *probability* and informally introduce some terminology and some notation. If this section is too technical for some readers, we encourage them to skip to Section 2.2.4 and maybe return later. Probability Theory is a complex field comprising formal aspects (the axiomatisation of the maths) as well as philosophical aspects (the interpretation of the maths). There is neither just one unique axiomatisation nor one unique interpretation of probability. We present one view (in a highly simplified form) that is useful for practitioners who want to use statistical inference in the analysis of empirical data.

We only contemplate situations that are *experiments* or resemble experiments (such as a lottery). A real and well-designed experiment is characterised by the fact that we know in advance what might happen. For example, we just calculated in Section 2.2.2 the number of possible outcomes in a classical Tea Tasting Experiment (which was 70). That was possible because the protocol of the experiment was defined exactly, such that we could even exhaustively enumerate all 70 possible outcomes on one or two pages of paper. This set of potential outcomes is called the *sample space*, and it's an important property of a proper random experiment that the sample space must be defined (although not necessarily enumerable). Thus, the sample space is a theoretical construct that is fixed before the experiment is conducted.

experiments

sample space

In a proper Tea Tasting Experiment, exactly 4 cups have to be chosen. If more or less cups are chosen, the experiment wasn't actually conducted properly—it failed. For the statistical analysis, selecting 4 cups counts as one single event or outcome. Each of the 70 distinct events in the sample space $\{E_1, \ldots, E_{70}\}$ is assigned a *probability*. Under the assumption that we're just guessing, none of these 70 outcomes is preferred, and we split the probability between them evenly, hence for each E_i, we get $Pr(E_i) = 1 \div 70 \approx 0.014$, where $Pr(\cdot)$ is the *probability function* that maps events onto their probability. What this means is that if we repeated the experiment over and over again by guessing, each possible configuration of 4 cups would be selected equally often. Hence, the notion of probability advocated in this book draws its interpretation from the concept of a hypothetically infinite sequence of repeated events. While it's hypothetical, it's still quite useful and related to everyday experience. It's the reason why you can't rely on a

probability

probability function

fair die to win any money in the long run. There is no way to predict the number of pips because each number has the same probability. No matter which number you bet on, you'll lose in $5 \div 6 = 0.83$ or 83% of all rolls.

event space

Continuing on a slightly more technical note, the so-called *event space* is the set of all possible subsets of the sample space. It contains singleton sets $\{E_i\}$ but also any other subset such as $\{E_i, E_j\}$ or $\{E_i, E_j, E_k\}$ (for $i, j, k \in \{1..70\}$). Crucially, the probability for each of these subsets must be determinable before the experiment. For a singleton set such as $\{E_i\}$, the probability is just the probability of its single member. In the Tea Tasting Experiment, it's $Pr(E_i) = 0.014$. The probability of a set such as $\{E_i, E_j\}$ is easy to determine. What is the probability that *either* E_i *or* E_j will occur? (Remember that E_i and E_j stand for one of 70 unique selections of 4 cups in the example.) It's the sum of the individual probabilities, just as your chances of winning a lottery double if you buy 2 lots instead of 1. Expressing *or* for events as $\cup$ (the set intersection symbol), we get:

$$Pr(E_i \cup E_j) = Pr(E_i) + Pr(E_j)$$

It's 0.028 for any two outcomes of the Tea Tasting Experiment. An essential task in frequentist statistics is to determine the probability of a specific set from the event space. Immediately below, in Section 2.2.4, we'll determine the probability of the set containing all outcomes with three or more correctly chosen cups in order to determine how impressive Ms Bristow's performance were if she classified 3 cups correctly.

The sample space (often denoted Ω), the event space (often denoted $\mathcal{F}$), and the probabilities assigned by the probability function Pr to the sets contained in the event space are key ingredients of probability theory. Furthermore, we introduced the notion of the probability of one *or* ($\cup$) more events. There is one final notion we need to introduce, namely the probability of two independent events to both occur. The simplest example (once again) is a fair die. If you roll it once, one of six possible events/outcomes will occur, and each outcome has the same probability. For example, $Pr(\boxdot) = 1 \div 6 \approx 0.167$ (using the symbol $\boxdot$ instead of some less intuitive E_i). If you roll the die again, the first roll doesn't influence the second roll in any way, and the probability is once again $Pr(\boxdot) = 1 \div 6 \approx 0.167$. However, what's the probability of any sequence of outcomes, such as *4 pips* followed by *3 pips*? It's the probability of one event *and* (written with the set union symbol $\cap$) another independent event: $\boxdot \cap \boxdot$, or more generally $E_i \cap E_j$. It's calculated by multiplication:

$$Pr(E_i \cap E_j) = Pr(E_i) \cdot Pr(E_j)$$

Hence:[7]

$$Pr(\boxdot \cap \boxdot) = Pr(\boxdot) \cdot Pr(\boxdot) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} = 0.028$$

Above, we pointed out that from the view of probability theory and experiment design, the whole experiment is regarded as one event. However, this event can often be divided up into a series of sub-events. In the example, each act of choosing a cup is one such sub-event, which is how we introduced it in Section 2.2.2. The maths for calculating probabilities of such sequences of events is exactly the same as for sequences of experiments as long as the sub-events are *independent events*. Independent events are those which don't influence each other's probability. If you roll a die twice, the number of pips you got in the first roll does not influence the second roll. But a speaker of any natural language uses a certain lexeme in a sentence, the probability of the same lexeme being used in the next sentence increases sharply. For example, a word such as *antidisestablishmentarianism* has a shockingly low probability of ever occurring, even in British English. Once it has been used in a sequence of sentences, however, the probability that it's used again in the next sentence might rise to the levels of words like *technicality* or *opposition*. Clearly, the events are not independent as the first one influences the probabilities of the second one. Please keep this in mind throughout the book (and your career in linguistics).

independent events

By the way, in all cases discussed so far the probabilities were evenly distributed among the possible outcomes, which is called a *uniform random distribution*. A fair die has a uniform random distribution of the number of pips, whereas a manipulated die doesn't. A uniform random distribution is by no means a necessity nor particularly typical, and we'll encounter many cases where other probability distributions are applicable later in this book. This concludes our very short and informal introduction to some rudimentary elements of probability theory. When we add and multiply probabilities in the next section, you can verify that we're doing it according to the foundations laid out in this section.

uniform random distribution

2.2.4 The Number of Some Less Than Perfect Outcomes

Before proceeding to delicate matters of scientific inference, let's calculate the probability of guessing 3 cups correctly and getting 1 wrong, as in our example from Section 2.2.2. To do that, we first introduce a convenient general notation

[7] It's by pure accident that we ended up with two different results in this section which are numerically identical at 0.028.

for the maths of choosing k items out of n. First, notice that in the calculations above we often multiplied a natural number with the next smaller natural number, then the next smaller number, and so on. For example, we calculated $4 \cdot 3 \cdot 2 \cdot 1$. Such an operation, where we multiply a natural number repeatedly with its next smaller neighbour until we reach 1, is called a *factorial*, and it is expressed as $n!$ such that $4! = 4 \cdot 3 \cdot 2 \cdot 1$ if $n = 4$. To calculate the number of possible decision chains for 4 cups out of 8 cups, we calculated $8 \cdot 7 \cdot 6 \cdot 5$, but then we didn't go down all the way to 1. For two cups out of eight, we'd have calculated $8 \cdot 7$ (two choices, then stop), etc. In general and using n as the variable encoding the number of items and k as the variable encoding the number of items to choose, this can be expressed as

factorial

$$\frac{n!}{(n-k)!}$$

To illustrate, we insert the concrete numbers from our example above:

$$\frac{8!}{(8-4)!} = \frac{8!}{4!} = \frac{8 \cdot 7 \cdot 6 \cdot 5 \cdot \cancel{4} \cdot \cancel{3} \cdot \cancel{2} \cdot \cancel{1}}{\cancel{4} \cdot \cancel{3} \cdot \cancel{2} \cdot \cancel{1}} = 8 \cdot 7 \cdot 6 \cdot 5$$

To account for the decision paths that led to identical results, we divided this further by 4! because $k = 4$ items can be arranged in $4 \cdot 3 \cdot 2 \cdot 1$ ways. Generally, we need to divide by $k!$ This gives us the *binomial coefficient* used to calculate the number of distinct sets of k items from n items without replacement and irrespective of their order:

binomial coefficient

$$\binom{n}{k} := \frac{n!}{(n-k)!k!}$$

The binomial coefficient is usually read *n choose k*, but for us non-mathematicians it's okay to read it as *k chosen from n*. Since the factorial results in very large numbers even for relatively low input numbers, it is often practically infeasible to calculate the binomial coefficient using the above formula, and several alternative methods for calculating it are available. They can be found on Wikipedia or in any text book teaching applied maths.

With this, we can now finally calculate how many ways there are of choosing 3 tea-first cups and 1 milk-first cup out of 8 cups in total where there are 4 tea-first and 4 milk-first cups. This is easy if we regard the 8 cups as two sets of 4 cups (milk-first and tea-first). Muriel Bristow thus chose 3 cups out of 4 correctly and 1 cup out of 4 incorrectly, hence:

$$\binom{4}{3}\binom{4}{1} = \frac{4!}{1!3!}\frac{4!}{3!1!} = \frac{24}{6}\cdot\frac{24}{6} = 4\cdot 4 = 16 \tag{2.2}$$

There are thus 16 distinct sets of 3 correct cups and 1 incorrect cup. We already know that there are 70 ways of choosing any 4 cups out of 8, and hence the frequentist probability of achieving such a result by chance is:

$$\frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = \frac{16}{70} \approx 0.23 \tag{2.3}$$

In the long run (repeating the experiment over and over again), even a random guesser would classify 3 cups correctly with a probability of 0.23, which corresponds to 23% of all experiments. So, are 3 correct guesses a highly unexpected result? Not exactly, because in about a quarter (23%) of an endless sequence of runs of such an experiment, anyone would reach this level of accuracy just by guessing.

There is one final amendment that we should make. In order to evaluate how unexpected a result with three correctly chosen cups is, it is more informative to ask how often such a result or an even better result would be when someone's just guessing. Hence, we should add the number of possible configurations with four correct cups, which is 1:

$$\binom{4}{3}\binom{4}{1} + \binom{4}{4}\binom{4}{0} = 4\cdot 4 + 1\cdot 1 = 16 + 1 = 17$$

The probability of obtaining 3 or 4 correct cups by chance is thus:

$$\frac{\binom{4}{3}\binom{4}{1} + \binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = \frac{17}{70} \approx 0.24 \tag{2.4}$$

But still, in light of this number, choosing 3 cups correctly wouldn't be an unexpected result at all. It would provide no grounds whatsoever for rejecting the hypothesis that Muriel Bristow is just guessing. Incidentally, she allegedly chose all four cups correctly when the experiment was actually conducted. Do you find this impressive?

2.2.5 Towards Making an Inference

What does this statistical anecdote show? We hope to have illustrated that there is a way of calculating how unexpected specific results of certain experiments are before the experiment is conducted and under the assumption that the experiment works essentially like a lottery: The person does not have the ability in question, the effect predicted by a theory or a specific hypothesis doesn't exist, the results are completely random. Then, if we run the experiment and the outcome happens to be one of the highly unexpected ones, our surprise might lead us to believe that the experiment wasn't just a lottery, but that there is something going on: The person does indeed have the ability in question, the effect predicted by a theory or a specific hypothesis does indeed exist, etc. In general, we'd assume that there is something going on. We would be making a *statistical inference* based on the outcome of the experiment. Since inference is a process fraught with peril, much more will have to be said about it, for example in Section 2.3.

statistical inference

However, a typical type of misunderstandings that comes with ill ways of phrasing the mathematical results of frequentist statistics must be mentioned immediately. This is the perfect time to mention it because it sounds absurd to most people at this stage. But once they've been taught statistics incorrectly, many people can't get this misunderstanding out of their brains again. The wrong story goes like this:

> *Because guessing three or more cups correctly has a probability of only 0.24, the probability that Muriel Bristow has the ability to detect tea-first cups is* $1 - 0.24 = 0.76$. *We can be 76% certain that she has the ability in question.*

This is all completely wrong. You should find such interpretations surprisingly unwarranted if you followed our argument. First, consider statements such as *the probability that she has the ability in question is 0.76* and how nonsensical they are regardless of the specific number. In reality, it's a fact that she either has the ability or she doesn't have it, and there's no probability attached to a fact. Probability only plays a role when there is a confrontation with chance, and before the proverbial dice are rolled. The probability we calculated above was the probability of obtaining the result which we then actually obtained before we conducted the experiment and under the assumption that there was no ability, effect, etc. at work but just pure randomness. We frequently call this hypothetical absence of an effect the assumption that there's nothing going on.

Second, we'd like to ask you what it really means to be *so-and-so percent certain of something*. We consider this notion (sometimes also called a *subjective*

probability or similar) to be undefined and above all irrelevant to science. Science is about how things really are and not about our subjective beliefs of how they are. However, we suggest you also give those people a chance who think the complete opposite is true. Google *subjective Bayesianism*, read a few blogs, and then read Senn (2011). However, everybody (including—we think—Bayesians of all kinds) agrees that the frequentist probabilities calculated in this book have nothing to do with *being certain* of anything to a quantifiable degree. They serve to test scientific claims in order to slowly establish them as unrefuted in a piecemeal fashion. Please keep the following Big Point in mind when reading the next section, which is about inference.

Big Point: Unexpected Outcomes

The outcome of an experiment is unexpected if it had a low probability before the experiment was conducted under the assumption that there's nothing going on. After that, the outcome is a fact and doesn't have a meaningful frequentist probability attached to it. A low probability of a specific outcome merely means that it would be rare if the experiment were conducted very often.

2.3 Inference With Fisher's Exact Test

2.3.1 p-Values and sig-Levels

The frequentist probability that a certain experimental result had before it was actually obtained (when calculated after the experiment) is often called a *p-value*. In a Tea Tasting Experiment with 8 cups, 4 tea-first cups, and 3 correctly detected tea-first cups, we can calculate $p = 0.24$ with Equation 2.4. The p-value is often used to make automatic inferences along the following lines:

p-value

> *There were 8 cups in total, 4 of them tea-first cups. Ms Bristow agreed to choose 4 cups to demonstrate that she can successfully detect tea-first cups. She detected 3 tea-first cups correctly, hence $p = 0.24$. Since this p-value is higher than the accepted significance threshold of 0.05, we conclude that the experi-*

significance threshold

ment has produced no evidence that Ms Bristow has above-chance abilities to detect tea-first cups.

The purpose of this section is to reconstruct and critique such inferences.[8] However, please keep in mind that *p-value* is just another name for what we calculated in Section 2.2.4 with Equation 2.4. The maths has been established, and this section is about its interpretation.

There are two important aspects with respect to making inferences from unexpected outcomes. Can we be *sure* that Muriel Bristow had the ability of discerning tea-first cups from milk-first cups just because in the real experiment roughly 100 years ago, she pointed out 4 correct cups? The answer is clearly negative because, well, $p \approx 0.014$ and not $p = 0$. To illustrate, consider my grandfather on my father's side, Karl. Karl used to play the German national lottery with a bunch of friends, and in 1962, they won the big prize. (True story!) They guessed 6 numbers out of 49 correctly. Would you take their win as evidence that they were prescient and could foresee the number that would be drawn? Even better, would you take it as evidence that extra-sensory perception (ESP) exists? After all, the p-value is bloody low:

$$p = \frac{1}{\binom{49}{6}} = \frac{1}{13,983,816} \approx 0.0000000715$$

Clearly, most readers would not make such an inference, regardless of how low the p-value is. The interpretation of a statistical result needs to be well-informed about the mechanism that brought about the result, and it must be made with great care. First of all, no known physical and/or cognitive mechanism that we know of could account for ESP, which is why most people don't even bother to run experiments investigating ESP. Good hypotheses about why and how ESP should be real simply don't exist. Also, notice that we did not, in fact, conduct an experiment about my grandfather's ESP abilities, but I merely told a story about him winning the lottery. At some point I started calling it an experiment, probably unnoticed by most readers. I could have told a similar true story about any random person I knew (a friend's girlfriend's uncle or whomever) if it so happened that that person won the lottery at some arbitrary point in the past instead of my grandfather. If you allow yourself to look anywhere for evidence of

[8] By the way, if you're wondering why the accepted threshold should be 0.05, we cannot help you. It was a vague and loose suggestion by Ronald A. Fisher, and it developed a life of its own. We don't recommend using it.

something, you're bound to find it somewhere, and a low p-value becomes utterly meaningless. Interestingly, Karl and his friends repeated the pseudo-experiment over and over again, playing the national lottery every week for roughly ten years in total (over five hundred draws) without ever winning any considerable amount of money again. This really makes it look like they were just guessing numbers and got lucky exactly once. Obviously, repeating experiments (so-called *replication*) is a very good way of further testing inferences made based on unexpected outcomes.

replication

Astonishingly, researchers in soft sciences are often satisfied if $p < 0.05$ in order to proceed to a substantive inference from an experiment. This is the ominous significance threshold mentioned at the beginning of this section. Such thresholds are often called the α level, although we prefer *sig-level* (for *significance level*). Setting $sig := 0.05$ means that researchers are satisfied to make an inference if the outcome of an experiment would only be expected in 1 out of 20 experiments under the assumption that there is nothing going on, i. e., that there really isn't any effect. While this may be justified in some cases, automatically assuming such a sig-level is ill-advised and outright insane. We'll come back to this point again and again, but to show you that a chance of 1 out of 20 usually wouldn't give you much confidence when there is any important matter at stake, think about the old game show *Let's Make a Deal*. In that show, contestants regularly had to choose one door out of three, and there was a big prize behind one of the doors and duds called *Zonk* (or at least much less impressive prizes) behind the other two. Let's modify the rules slightly: In the soft-sciences version of *Let's Make a Deal* there are 20 doors. Behind 19 of them, there are prizes worth a significant fortune (grant money, which means money plus prestige), but one of the doors hides an automated gun turret which instantly kills the contestant if they choose that door. Would you expect any sane person to participate in such a game show? Of course you wouldn't. People who—given a choice—wouldn't take any substantial risk in real life with a 1 in 20 chance are happy to bet the future of linguistics or social psychology on such a chance by setting $sig := 0.05$ (while not doing replication). On the other hand, we have seen that even $p \approx 0.0000000715$ might be meaningless. Clearly, doing maths is easy, but making good inferences isn't. Therefore, two of the major themes in this book are (i) that you shouldn't take unnecessarily high risks in making scientific inferences from data and (ii) that there is no recipe-like procedure that leads to good inferences.

sig-level

In the next section, we'll return to the corpus example from the Problem Statement and formalise the procedure described above in the form of Fisher's Exact Test or just Fisher Test. You'll see that the logic of the Tea Tasting Experiment lies behind the omnipresent 2x2 tables that often pop up in the corpus literature,

especially in research on collocations and collostructions (Stefanowitsch & Gries 2003, Evert 2008).

2.3.2 Fisher's Exact Test and Null Hypotheses

Our examples from the Problem Statement and the Tea Tasting Experiment are all about comparing counts of events. How many cups of tea were chosen correctly relative to the number of possible events of choosing 4 cups of tea? The comparison of these numbers allowed us to calculate the probability of an outcome as extreme as or more extreme than the actual outcome under the assumption that the process generating the guesses (for example, Muriel Bristow) is completely random. Such counts are customarily summarised in *contingency tables*, see Table 2.1.

contingency tables

Table 2.1: The contingency table for the outcome with 3 correctly classified cups

	Reality		
	Tea-First	**Milk-First**	**Sum**
Bristow			
Tea-First	3	1	**4**
Milk-First	1	3	**4**
Sum	**4**	**4**	**8**

A contingency table tabulates counts of items, acts, or events characterised by two *variables*, each having two or more discrete possible values. Here, one variable (called *Reality* in Table 2.1) characterises cups as being a real tea-first or a milk-first cup, and the other variable (called *Bristow* in Table 2.1) characterises cups as designated by Ms Bristow as a tea-first or a milk-first cup. One variable is shown in columns, the other one in rows, and the table shows us how often which values of the two variables co-occur. In row 1, we put the 4 tea-first cups according to Muriel Bristow. In row 2, we put the 4 milk-first cups as classified by her. The row sums in the last column reflect the fact that there were indeed 4 cups for each condition. In the two columns, we count the real tea-first and milk-first cups. As you can see, of the 4 real tea-first cups, Muriel Bristow classified 3 as tea-first (cell 1,1) and 1 as milk-first (cell 2,1).[9] The opposite is true for the real

variables

[9] In matrices and tables, it is customary to index cells first by rows, then by columns. Hence, cell 1,1 is the upper-left cell. Cell 2,1 is the lower-left cell. The upper-right and lower-right cells are indexed 1,2 and 2,2, respectively.

milk-first cups.

To calculate the p-value for Fisher's Exact Test—which is exactly what we've been introducing in this chapter—we only need to consider 6 of the 9 cells of the contingency table: the shaded cells in Table 2.2. With those counts, we can calculate the probability of drawing 3 cups from a set of 4 and independently 1 cup from a different set of 4, which corresponds exactly to the binomial coefficient calculated in Equation 2.2.

Table 2.2: The contingency table for the outcome with 3 correctly classified cups; the relevant cells are highlighted

	Reality		
	Tea-First	**Milk-First**	**Sum**
Bristow			
Tea-First	3	1	**4**
Milk-First	1	3	**4**
Sum	**4**	**4**	**8**

As was shown above, we need to add the probability of obtaining a more extreme result, which is illustrated for completeness in Table 2.3.

Table 2.3: The contingency table for the *even more extreme result*

	Reality		
	Tea-First	**Milk-First**	**Sum**
Bristow			
Tea-First	4	0	**4**
Milk-First	0	4	**4**
Sum	**4**	**4**	**8**

Finally, we turn to the corpus example from the Problem Statement. As we pointed out in Section 2.1, the Problem Statement mentions 10 passives of the verb *sleep*, but we need more information. Let's introduce that information and the argument that comes with it, roughly following the logic behind collostructional analysis.[10] Assume that we drew 90 active sentences containing *sleep* in addition

[10] Paradoxically, we consider collostructional analysis to be suboptimal as an illustration of Fisherian logic of inference. However, as it used to be the most prominent use case of Fisher's Exact Test in linguistics, we use it to introduce the test and critique the specific use of it in the process.

to the 10 passives. Furthermore, assume that the corpus contains 1100 sentences in total, 890 of them active sentences, 210 passive sentences. The corresponding contingency table is shown in Table 2.4.

Table 2.4: A contingency table approximately as found in collostructional analysis

	Voice		
	Active	**Passive**	**Sum**
Verb			
Sleep	90	10	**100**
Other	800	200	**1000**
Sum	**890**	**210**	**1100**

Descriptively, it is the case that *sleep* occurs less frequently in the passive than all other verbs. Only 10% of all occurrences of *sleep* are passives, but 20% of all other verbs are passives. Using the maths introduced in this chapter, we can attempt to quantify how unexpected this result is under the assumption that there's nothing going on, and the story goes as follows. If we drew 100 sentences randomly from this corpus, what would be the frequentist probability of drawing exactly 90 active sentences and 10 passive sentences, given the overall distribution of voice in the corpus? Clearly, it would be:

$$\frac{\binom{890}{90}\binom{210}{10}}{\binom{1100}{100}} \approx \frac{6.573 \cdot 10^{141}}{1.423 \cdot 10^{144}} \approx 0.005$$

This probability is intersting because we obtained the result by querying for all sentences containing *sleep*, not by drawing random sentences. Based on this, we can now inch our way towards making an inference about *sleep*. Our theory states that there should be no or at least very few passives of verbs like *sleep* (unaccusatives) compared to other verbs, many of which can be readily passivised (transitives, unergatives). If *sleep* behaved like the average verb, a sample of sentences containing it would be expected to resemble a random sample from the corpus with respect to the number of actives and passives. The more extreme the distribution of verbal voice in the *sleep* sample is, the more we are inclined to assume that *sleep* does not behave like the rest of the verbs with respect to passivisation. This is actually a similar logic as in the Tea Tasting Experiment. An

unexpected result under the assumption of mere guessing on Muriel Bristow's part is conceptually very similar to an unexpected result regarding the distribution of verbal voice in a corpus sample under the assumption that the sample is not in some way different from the rest of the corpus. The Fisherian type of the frequentist logic of inference is based on this kind of argument: Since it is very difficult to come up with quantitative evidence in favour of research hypotheses (see Chapter 1), a *Null Hypothesis* (or simply *Null*, symbolised H_0) is constructed. The Null states in some way that the effect predicted by the theory is absent. If the result obtained in the experiment has a very low pre-experiment probability under the assumption that the Null is true (i. e., it's an unexpected result), the experiment is assumed to lend some limited support for the theory (or rather for a minor hypothesis which is part of the theory). The usual phrase is that *the Null was rejected*. Unexpected results are not in any way taken as proof of the theory or a part of the theory, and it should be obvious why. First, we never know whether a rare (unexpected) event has occurred by chance, regardless of how unexpected it was before we conducted the experiment. Our calculations are based on the realisation that it is not at all impossible to obtain unexpected results by chance. On the contrary, the p-value quantifies the probability of the actual result under a given Null (i. e., by chance) with perfect precision. Second, take the aforementioned pseudo-experiment regarding my grandfather's ESP abilities with $p = 0.0000000715$. It's practically irrelevant how low the p-value is, most people will not take it as evidence that my grandfather or one of his friends could foresee the numbers drawn in next week's national lottery. Thus, the strength of the evidence depends on many factors, among them the design of the experiment and the p-value.

Null Hypothesis

Let's keep this in mind and complete the maths. The value of $p = 0.005$ calculated above is just the probability of obtaining exactly 10 passives and 90 actives under the informally stated H_0: The verb *sleep* is passivised as often as all other verbs. Since any more extreme (i. e., even lower) number of passives would be at least as good evidence against the Null, we should include them. Hence:

$$p = \frac{\binom{890}{90}\binom{210}{10}}{\binom{1100}{100}} + \frac{\binom{890}{91}\binom{210}{9}}{\binom{1100}{100}} + \cdots + \frac{\binom{890}{100}\binom{210}{0}}{\binom{1100}{100}} \approx 0.008$$

This is indeed the p-value as calculated in Fisher's Exact Test. It's not the only possible p-value, as we'll show immediately.

2.3.2.1 The Direction of the Null Hypothesis

Let's take stock. We've shown that the Tea Tasting Experiment and other confrontations with chance can be analysed statistically with a very simple logic. If Ms Bristow has no special tea-tasting abilities, she would still guess 3 or 4 cups correctly in 24% of an infinite or at least very long sequence of trials. If *sleep* did indeed behave like the totality of all other verbs with respect to passivisation in the fictitious corpus described above, 100 instances of *sleep*, randomly sampled from a corpus would still lead to just 10 or less passives in 0.8% of an infinite or at least very long sequence of trials. If my grandfather and his friends weren't prescient, they'd still have won the big prize in 0.00000715% of an infinite or at least very long sequence of draws.

In the case of the corpus study, we had in mind a specific substantive hypothesis generated by many theories of syntax and the lexicon, namely that unaccusative verbs cannot be passivised or at least are very rarely passivised. Hence, we calculated the probability of obtaining a sample with as few or fewer passives of *sleep* in order to weaken the Null and mildly support the family of hypotheses to which the substantive hypothesis belongs.[11] This means, however, that the Null couldn't have been: *The verb 'sleep' passivises as often as the totality of all other verbs.* If this were the Null, then any strong deviation (of *sleep*) from the overall distribution of passives would be sufficient to weaken it. We'd have to take a result with, for example, 90 passives and 10 actives of *sleep* (i. e., a strong deviation in the other direction) as equally good evidence against the Null as the result that was actually obtained. The test would be a *two-sided test* because it would test for deviations to both sides at the same time.

two-sided test

Based on our substantive hypothesis, however, we were only interested in deviations into one direction, namely a lower relative number (proportion) of occurrences of *sleep* in the passive voice compared to the overall proportion of active and passive voice in the corpus. Hence, the actual Null must have something like: *The verb 'sleep' passivises as often or less often than the totality of all other verbs verb.* The test was a *single-sided test*, in this case a right-sided test (see Section 2.6). If $q_{\text{PASSIVE}(sleep)}$ denotes the proportion of passives of *sleep* and $q_{\text{PASSIVE}(\neg sleep)}$ denotes the proportion of passives of all the other verbs, then the formalisation of the Null is:

single-sided test

$$H_0 \ : \ q_{\text{PASSIVE}(sleep)} \geq q_{\text{PASSIVE}(\neg sleep)}$$

There might be other verbs for which the same theory predicts the exact opposite: that their passive forms occur more often than those of the average verb.

[11] Please review Chapter 1 if the exact wording of this sentence is less than crystal clear to you.

For example, *arrest* could be such a verb, as it is not only transitive but also often used in reports of arrests with phrases like: *A man was arrested by the police.* In order to test a Null based on this substantive hypothesis, we have to reverse the relation between the proportions:

$$H_0 : q_{\text{PASSIVE}(arrest)} \leq q_{\text{PASSIVE}(\neg arrest)}$$

Table 2.5: The fictitious result for the voice distribution of *arrest*

	Voice		
	Active	**Passive**	**Sum**
Verb			
Arrest	69	31	**100**
Other	800	200	**1000**
Sum	**869**	**231**	**1100**

In this case, a result as convincing as the result for *sleep* is described numerically in Table 2.5, assuming there were also exactly 100 instances of *arrest* overall. The p-value for Fisher's Exact Test can be calculated following the know well-established method as follows:

$$p = \frac{\binom{869}{69}\binom{231}{31}}{\binom{1100}{100}} + \frac{\binom{869}{68}\binom{231}{32}}{\binom{1100}{100}} + \cdots + \frac{\binom{869}{0}\binom{231}{100}}{\binom{1100}{100}} \approx 0.009$$

Since we are limited by the corpus size of 1100 and the fixed sample size of 100, we cannot reach exactly 0.008. However, 0.009 is close enough. Such a result could be interpreted as weakening the Null which states that *arrest* is passivised as often or more often than all other verbs. While we can't argue that a precise substantive hypothesis was directly supported, at least a global negation of such a hypothesis failed the test. The directionality involved in formulating the Null is obviously informed by the substantive hypothesis. Hence, it wouldn't be correct to think that Fisher's testing philosophy completely ignores substantive hypotheses. In Chapter 7, we'll turn to methods of directly supporting hypotheses within the Neyman-Pearson philosophy and Deborah Mayo's Severity approach to frequentist inference.

2.4 Sample Size and Effect Strength

When we obtain a low p-value, we can assume that the Null is false or a rare event has occurred. The Null usually states that some effect is absent, that there is nothing going on beyond random chance. We have already illustrated (using the soft-sciences version of *Let's Make a Deal* and my grandfather's lottery win) that there cannot be a fixed threshold for what counts as rare. In this section, we examine how the size of the sample and the p-value are related. Then, we informally introduce the notion of the effect strength, to which we'll return repeatedly in later chapters, very crucially in Chapter 7.

Assume that Fisher and Ms Bristow, after fiercely discussing (as one should) the theoretical foundations of her potential tea-tasting abilities that might have shown in the first experiment, decided to repeat the experiment but with a much larger sample. This time, she agrees to taste 80 cups of tea, 40 of which are tea-first. After the Tea Tasting Marathon, Fisher counts the successes and failures and aggregates the numbers in Table 2.6.

Table 2.6: Contingency table of the Tea Tasting Marathon

	Reality		
	Tea-First	**Milk-First**	**Sum**
Bristow			
Tea-First	30	10	**40**
Milk-First	10	30	**40**
Sum	**40**	**40**	**80**

In this second experiment, the results are similar to those reported in Table 2.2 inasmuch as she guessed 75% of the tea-first cups correctly. However, if we calculate the pre-experiment frequentist probability of obtaining this result or an even better result without tea-tasting abilities by chance (i. e., the p-value), we get:

$$p = \frac{\binom{40}{30}\binom{40}{10}}{\binom{80}{40}} + \frac{\binom{40}{31}\binom{40}{9}}{\binom{80}{40}} + \cdots + \frac{\binom{40}{40}\binom{40}{0}}{\binom{80}{40}} \approx 7.44 \cdot 10^{-6} \tag{2.5}$$

Do you recognise the exponential notation $7.44 \cdot 10^{-6}$? With a negative exponent of -6, it means that you have to move the point 6 digits to the left in the significand (the part before the multiplication sign), hence $7.44 \cdot 10^{-6} = 0.00000744$.

Although the rate of accuracy is the same, the p-value is much lower than the p-value corresponding to Table 2.2, which was 0.24. This should be an intuitively desirable result. If she has the ability in question, this will show much better in a larger number of trials. Once again, an urn serves as a good and even simpler example of this effect. Let's consider a situation where there the suspicion arises that an urn at a funfair might be unfair inasmuch as it does not contain the advertised number of winning lots. The showman claims there are 1000 lots in the urn, 300 of them winners, but you suspect that there are considerably less winners. To test his claim that the urn contains 30% winning lots, you agree to draw a sample of 10 lots. Even if you ended up with 1 winning lot and 9 duds, it wouldn't be (again, intuitively) much evidence of anything because such a small sample is generally not very trustworthy. If you drew a sample of 100 lots, however, and you ended up with 10 winners and 90 duds, the showman would be in trouble.[12] In the same vein, the Tea Tasting Marathon can be compared with the small-sample Tea Tasting Experiment. Since the p-value is interpreted as providing stronger evidence the smaller it is, it behaves as we'd intuitively think it should.

One cautionary note is in order. An experiment is an attempt to find out something about the real world, to detect whether an effect is real. For example, we're curious whether Ms Bristow really has the ability in question or whether the corpus frequencies of certain verbs' voices are in line with our theories (and hence whether our theories are in line with reality). If we draw a larger sample, nothing changes about the real world. We simply increase our chances of detecting an effect (indirectly, by rejecting the Null) if it is there. We have to keep in mind that the p-value obtained in the experiment is just a particular result of us conducting the experiment, which is in essence a confrontation with chance. If there is no effect in the real world, each p-value has exactly the same chance of emerging as the result of our experiment before we actually conduct the experiment, regardless of the sample size.[13] If this sounds weird and kind of wrong to you, please read on until Section 2.5.

Alas, there is another catch, and it's one which becomes ever more important when moving on to realistic experiments from made up and toy scenarios like the Tea Tasting Experiment. Usually, we don't expect humans to be deterministic, and we don't expect human behaviour to be discrete. Even if Ms Bristow could detect tea-first cups, she probably had a non-zero error rate. In fact, this is why we need the type of analysis provided by Fisher's Exact Test in the first

[12] The overall topic of sample size and sample accuracy will be discussed in Chapter 4.

[13] For this to be provably true, certain conditions need to hold, as we will discuss in Chapter 4.

place. If she claimed she had infallible perfect taste, we'd probably just test her with a single cup from time to time and wait for her first error, which would prove her claim wrong beyond doubt. Anyone who misclassifies a single cup does not have truly perfect taste. Clearly, such proof would be much more satisfying than merely rejecting a Null (thus providing weak evidence in favour of a hypothesis, etc.). To everybody's chagrin, the world usually isn't as simple as that. Similarly, even if *sleep* is an unaccusative verb, we expect it to pop up in the passive voice from time to time, either because we consider grammar to be probabilistic (as in many usage-based approaches), or because we attribute such occurrences to performance effects (as in generative and formal approaches with a psycholinguistic interest in processing). We don't take sides here because we don't have to. In any case, both corpora and controlled experiments are based on human behaviour such as linguistic output or reactions to linguistic input. This behaviour is never deterministic, and we cannot expect perfect results from humans. If we could, and if unaccusativity blocked passives strictly, we could just look for one example of any unaccusative verb in the passive voice in order to prove the underlying theory wrong (see Footnote 2).

Why is this important? For the penultimate time, think about Ms Bristow. If she could detect tea-first cups with 75% accuracy, she had the ability of interest, except it wasn't perfect. Also notice that the fictitious corpus study about passives was introduced from the beginning with caveats regarding marginal contexts where even unaccusative verbs could be passivised, etc. We never assumed the results would be perfect in the sense that we'd never encounter *sleep* in the passive voice. Under many probabilistic/usage-based approaches, verbs are even regarded as unaccusative only to a certain degree, such that *sleep* might less unaccusative than *burst*, for example. Nobody expects to never ever find any unaccusative verb in the passive voice among the sentences produced by real speakers, and nobody expects to get unanimous and perfect rejection of all unaccusative verbs in the passive voice in a rating experiment. Therefore, the effect (e. g., unaccusativity leads to reduced passivisation) has a certain *effect strength*. The effect strength is a property of the real-world phenomenon we're trying to capture with our experiment. It is crucially different from the sample size, which is merely a property of our experiment. The sample size only affects our chances of measuring an effect that is real, but it doesn't make an effect real. In other words, it doesn't change the distribution of the p-values. Different effect strengths, on the other hand, correlate with different probabilities for the p-values. Again, Section 2.5 will shed some more light on this, but it's highly important to understand the difference right from the beginning. A large effect is easier to detect per se. The better Ms Bristow's tea-first detection, the easier it is

effect strength

to detect in any experiment because a higher effect strength increases the probability of low p-values. The more prototypical a verb's unaccusativity, the easier it is to recognise as an unaccusative verb in an experiment because a higher effect strength increases the probability of low p-values. A larger sample just makes it easier to detect an effect if it is real.

Did we repeat the same statements several times in the previous paragraph? That was intentional, and we're going to repeat those statements in different form in subsequent chapters. Many practical introductions to statistics blur the distinction by simply stating that *both a large sample size and a high effect strength make it easier to obtain a low p-value.* Among other quite negative effects, this often gives the impression that it's legitimate to increase the sample size until we finally catch that pesky effect. We agree with even the most fanatic Bayesian that this is beyond evil.[14] Only a true understanding of frequentist inference ensures that is applied correctly in everyday research. Therefore, we now dive deeper into the nature of p-values.

2.5 The Distribution of p-Values (With Simulations)

Did you notice that we talked about the probability of obtaining a low p-value? We agree that it might sound confusing to speak of the probability of obtaining a p-value, because the p-value is something like a probability itself. However, the p-value is merely a metric that tells us what the probability of obtaining the result (that was actually obtained) was before we actually obtained it under the assumption that Null is true. Argh! This statistics thing is harder than expected. A p-value clearly isn't an innocent probability itself. If anything, it was a hypothetical probability before we rolled the proverbial die.

To clarify this admittedly twisted notion, we resort to simulations. There are introductions based on and devoted to simulations (see Vasishth & Broe 2011 or the more advanced Carsey & Harden 2014), but we'll keep the fundamentals short since we're not trying to teach you how to do simulations yourself. Luckily, modern computers are powerful machines, and they have very good random number generators.[15] Therefore, we can use them to simulate processes that generate data, where a *process* can be anything from Ms Bristow labelling cups of

[14] If you haven't guessed already, Bayesian statistics is a different approach to statistical inference. We praise many Bayesians for mercilessly exposing bad frequentist practice, and we definitely encourage you to look at Bayesianism. At the same time, we aren't convinced that it's the final solution to all problems of statistical inference.

[15] Here, *good* means that they approach true randomness very well without reconstructible patterns in the sequence of generated numbers.

tea to verbs occurring in the active and passive voice. In the simulation, we have the luxury of controlling all numerical properties of the simulated real world. For example, we could set up a simulation that behaves like a tea-first detecting lady with a 75% accuracy rate. Or we could set up a simulation of a corpus where all verbs have the same tendency to occur in the passive, or a corpus where *sleep* has a tendency much below average to occur in the passive, but this tendency is even stronger for *burst*, etc. The simulation can spit out any amount of data generated according to our specifications. By subsequently analysing those data with our statistical tests, it is possible to see how the tests perform in uncovering the true nature of the process generating the data, i. e., the true nature of the simulated reality. While simulations cannot be used to find out much about the actual real world, there are two advantages to this approach in the evaluation of our statistical tools. First, the data are virtually free. Anyone who has ever conducted an experiment or a corpus study knows that empirical work is tedious and expensive. It's much easier to generate simulated data with a few lines of code. Second, we never know what reality is like in doing real empirical work. On the contrary, we rely on our statistical methods and our knowledge of them when doing empirical work in order to make valid inferences and uncover what reality is like. Therefore, real experiments are mostly useless in the testing and demonstration of statistical methods. Instead, we can use simulations to test the properties of our statistical tools, and we can use them to demonstrate these properties in teaching statistics. In this book, they are used exclusively for demonstration.

Let's simulate a completely untalented tea-taster who really just guesses tea-first and milk-first with equal probability.[16] Based on this uninformed guesser, we simulate a classic Tea Tasting Experiment with four tea-first and four milk-first cups. Table 2.7 shows the raw results, and Table 2.8 shows the contingency table.

Table 2.7: The simulated outcome of a Tea Tasting Experiment

	Cup 1	**Cup 2**	**Cup 3**	**Cup 4**	**Cup 5**	**Cup 6**	**Cup 7**	**Cup 8**
Reality	Milk-First	Tea-First	Milk-First	Tea-First	Milk-First	Milk-First	Tea-First	Tea-First
Guess	Milk-First	Milk-First	Tea-First	Milk-First	Tea-First	Tea-First	Milk-First	Tea-First

[16] When we present a simulation, all data are always generated by code with a random generator. They are not hand-picked or manipulated in any way. Random seed settings are used to make them reproducible. The open-source code for the simulations is written in R, and it can be consulted as part of the Knitr sources for this book.

Table 2.8: Contingency table for the above

	Reality		
	Tea-first	**Milk-first**	**Sum**
Guess			
Tea-first	1	3	**4**
Milk-first	3	1	**4**
Sum	**4**	**4**	**8**

In this single simulated experiment, the random guesser guessed 2 cups correctly and 6 incorrectly. Since the appropriate test is single-sided, the p-value is different from the one corresponding to Table 2.2 although the results looks superficially similar. It's 0.986. Intuitively, many practitioners see such a result and think something like this:[17]

> *Yay! I totally get this result! After all, the person is just guessing, which means that there is no effect. Hence, the p-value should be high, and we can't reject the Null—which is correct. I've finally wrapped my mind around this testing thing. I'm going to finish my thesis now and do some serious empirical work!*

Unfortunately, this is fundamentally incorrect, as we will demonstrate using simulations now.[18]

The true purpose of running simulations is not to run them once but to run them very often. By looking at the results of a large number of so-called replications, we can check whether the claims made by the frequentist theory are true. Let's run 1000 simulations of the same experiment where the person is just guessing. To make the properties of the test pop out more clearly, we also change the number of cups per experiment from 8 to 1000. Since we cannot inspect the results for 1000 times 1000 cups individually, we aggregate them in a certain type of plot—a *histogram*—in Figure 2.6.

histogram

A histogram shows how often certain values occur in a sample or a population. In this case, the values are p-values (on the x-axis). The height of the bars (y-axis)

[17] Please don't ask how we know this.

[18] If you're writing a thesis (or you're about to write one) involving statistical analyses of empirical data, and you don't immediately see why this is false, don't do it. You're not ready yet. Much worse, if you're advising such theses, and you still don't see the problem, it's time to stop and think.

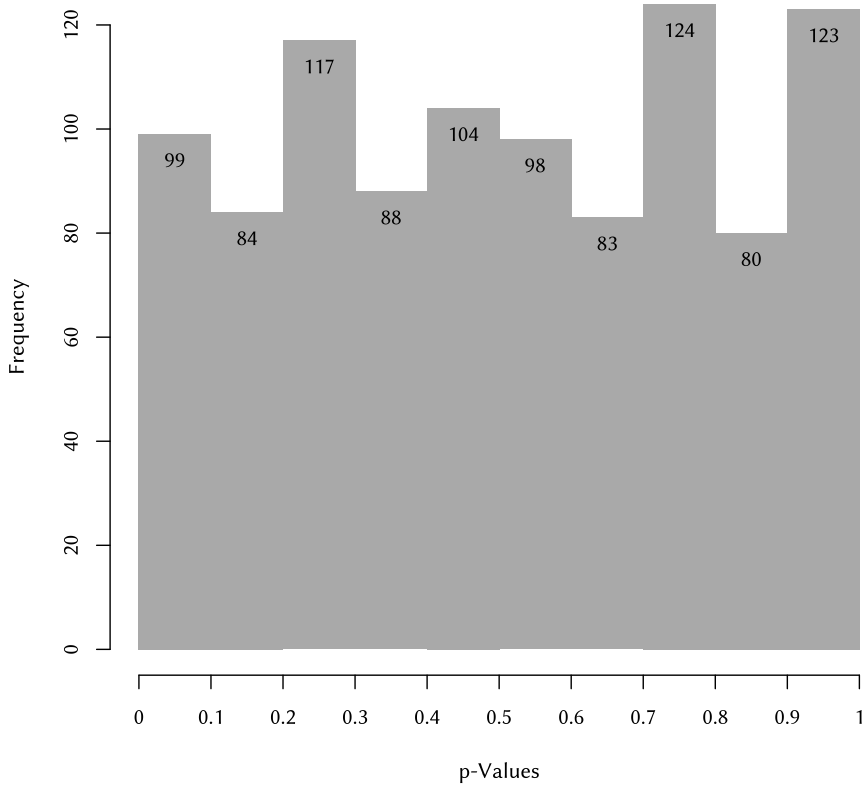

Figure 2.6: Histogram of p-values of 1000 simulated experiments with 1000 cups where the participant is merely guessing

corresponds linearly to the number of p-values that lie in the interval delimited by the respective bar. For example, p-values between 0 and 0.1 were obtained in 99 simulated experiments. p-Values between 0.1 and 0.2 were obtained in 84 simulated experiments, and so on. Some fluctuations aside, it looks like all p-values have the same probability if there is no effect.[19]

[19]Due to the nature of Fisher's Exact Test and the symmetry of the experiment design, there is some patterning in the p-values, which is hidden by the histogram in Figure 2.6. It's really not relevant at the moment, and our point is still absolutely valid. In Chapter 5, even better examples will be shown.

This is a perfectly expected result. Under the Null, each p-value has the same chance of occurring as the result of an experiment. Please revise the calculations in Section 2.2. The whole point was to count the possible outcomes of the experiment. (Remember how we got to the 70 different possible outcomes of Fisher's original experiment?) If there's nothing going on, each of these outcomes has the same chance of occurring, which also means that each p-value gets the same chance. It's explicitly not the case that we can expect high p-values if there is no effect. This point is mostly neglected in introductions for practitioners, and it's one of the most fatally misunderstood points in applied frequentist statistics. At least some texts state that *high p-values should not be interpreted as evidence against the substantive hypothesis*—or—*high p-values should not be interpreted as evidence for the Null*. That's absolutely correct, but now you know why. If the substantive hypothesis is false and the Null describes reality, then high, low, and medium p-values all have the same probability. That's why only low p-values (if any) have an inferential interpretation.

Big Point: Distribution of p-Values under the Null

The null hypothesis states in some way, shape, or form that there is no effect. If it is true, all p-values from 0 to 1 have the same probability. This is the reason why high p-values have no inferential interpretation and must not be taken as evidence for the Null and/or against the substantive hypothesis.

Let's see what happens with the p-values if the participant has the ability to guess tea-first cups with slightly better-than-chance accuracy. Figure 2.7 shows the distributions of p-values for 1000 experiments with 1000 cups. The different panels plot the p-values for simulated participants with a varying tea-first detection accuray. In the upper-left panel, it's 51%, in the upper-right 52%, in the lower-left 55%, and in the lower-right 70%.

This is an expected result. If the Null is not true, low p-values are obtained with a higher chance than high p-values. However, we see that it's not irrelevant how well the participant performs the task, i. e., how large the effect is (see Section 2.4). Even at an accuracy of 1% above chance (51%), 154 of 1000 p-values are already

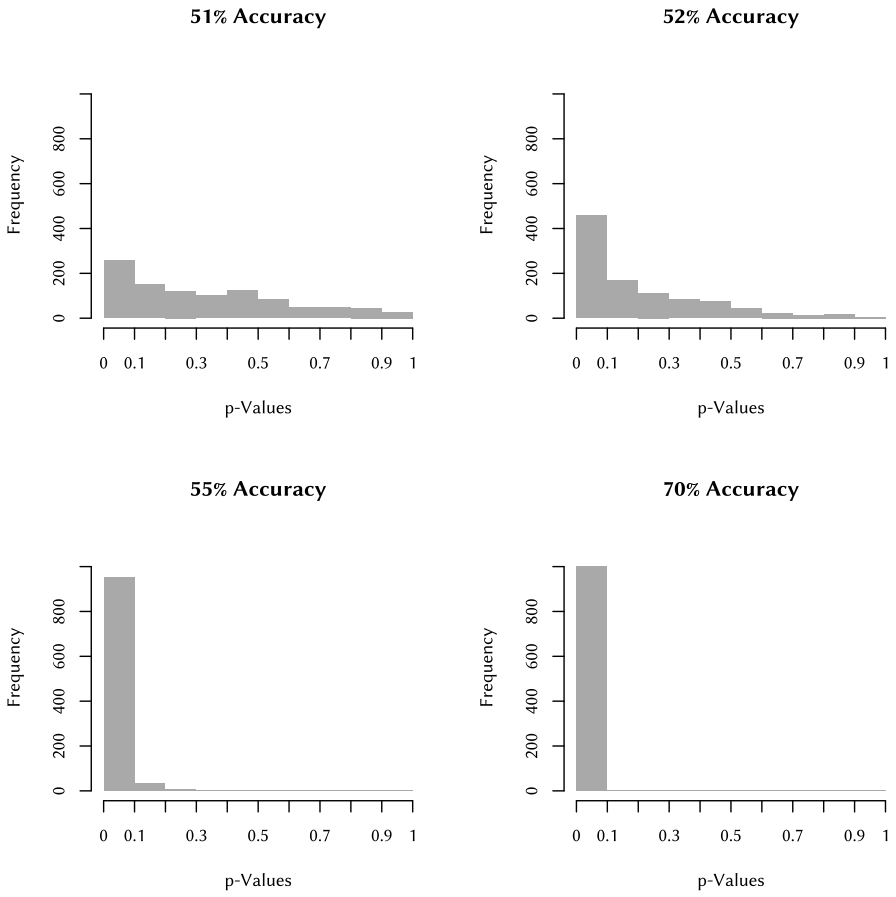

Figure 2.7: Histograms of p-values of 1000 simulated experiments with 1000 cups where the participant has increasing prediction accuracy (increasing from upper-left to lower-right); all y-axes are scaled to 1000 for better comparability

lower than 0.05. At an accuracy of 2% above chance (52%), it's already 329 of 1000 p-values, and at 5% above chance (55%) it's an astonishing 920. If we applied sig $:= 0.05$ in order to test whether there is an effect, we'd almost be guaranteed to *reach significance* (which is what people call it when they reach their pre-set sig-level) at 55% accuracy, albeit the tea-tasting abilities of the participant are middling (to say the least). It's highly doubtful whether this would be good for anything except claiming victory or getting a paper published. It almost certainly wouldn't advance our knowledge of the real world.

What we've just shown is that the effect strength tilts the distribution of p-values towards 0. The stronger the effect is, the more pronounced the tilt. Combined with a large sample, the distribution of the p-values collapses around 0 very easily, which means that mechanical inferences based on pre-set sig-levels are almost always successful, even if the effect is weak and barely above chance level. However, don't misunderstand this logic. Figure 2.6 showed that a large sample alone doesn't mean that we reach significance. Then, Figure 2.7 showed that as soon as there is even the tiniest effect, a large sample has a very high chance of detecting it by allowing a rejection of the Null.[20] This is because Fisher's Exact Test was designed for small samples such as 8 cups, 5 different types of barley, etc. In analyses of large data sets like corpus data with $n = 1000$, $n = 100000$ (or even $n = 200$) Fisher's Exact Test (and virtually all other frequentist tests applicable for the analysis of such data) have a much too high *sensitivity*. A test's sensitivity is its capability of rejecting the Null when there is an effect, and it will be a major topic in Chapter 7. Also related is the introduction of the odds ratio in Chapter 3, which serves as a measure of effect strength with Fisher's Exact Test.

sensitivity

Why go through all the trouble to learn about frequentist inference when it's no good after all? (... we hear you scream.) Well, we don't know why you came here. Only you know what your research questions are, which sources of data you have available, what your previous expertise in data analysis an statistics is. If you've just found out that the methods discussed in this book aren't for you and that you're wasting your time, more power to you! However, please give us chance to convince you to read on. First, we've never said that the methods aren't any good per se. If you have to work with small samples out of necessity and you've come up with a good design for your study and you apply frequentist tests wisely, you will benefit from this book. Second, frequentist tests are used everywhere, and your ability to understand and critique published research depends crucially on your deep knowledge of frequentist methods. Unless you're one of very few researchers who have invested a lot of time in understanding statistics, this book might help you in ironing out some misunderstandings and sharpen your awareness of the omnipresent misuse of frequentist statistics. Third, Section 2.4 and the present section make things look slightly worse than they are. We had to crank up the sample size quite aggressively and hide some quirks of the test (see Footnote 19) in order to make our point. This is because Fisher's Exact Test is extremely well-suited to introduce the logic of frequentist

[20] The usual caveats apply when we use phrases like *rejecting the Null* or *detecting an effect*. Fisher's frequentist logic never delivers proof nor decisive evidence nor even straightforward support for the substantive hypothesis.

inference (the primary goal of this chapter), but it is not the best example when talking about the distribution of p-values and related issues. Fourth, we will formalise notions like a test's sensitivity when introducing the Neyman-Pearson philosophy of frequentist inference in Chapter 7. This will enable you to use frequentist tests much more wisely. Fifth, we will always point out where tests fail or are unnecessary, and what you can do instead.

Our final remark in this section concerns such a case, namely a case where it's much better to use something other than a test or a p-value. If you're not a corpus linguist or if you have no interest in collostructional analyses, you can skip this paragraph. In Section 2.3.2 we claimed that our example application of Fisher's Exact Test was similar to its use in research on collostructions. That's not exactly true, and we urge you to consult papers like Gries (2014, 2015, 2022), maybe also Schmid & Küchenhoff (2013), Küchenhoff & Schmid (2015). You will discover that the design of such analyses is different from our toy example. Most importantly, it was never the goal of collostructional analyses to arrive at scientific inferences via Fisher's logic of statistical inference. It was an exploratory method from the beginning, never an inferential method. The p-values were primarily used to rank lexemes, not to infer substantive hypotheses about individual lexemes.[21] While it was a bad decision to use p-values from Fisher's Exact Test in collostructional analyses, it is now widely recognised that measures of effect strength are much more appropriate for the purpose. Also, possible discussions about the applicability of Fisher's Exact Test in situations where the marginal sums are not fixed (such as most corpus studies) are mostly moot. Virtually all alternative tests are even more sensitive, and we've seen that Fisher's Exact Test is oversensitive, if anything. With that clarification, we proceed to the very last section in this chapter, introducing the concept of probability distributions and their defining functions.

2.6 In-Depth The Hypergeometric Distribution

In this chapter, we tried to introduce the notions of probability and inferences based on probabilities intuitively and by building the necessary maths step by step. In probability theory, knowledge about random phenomena is customarily formulated in the form of probability density functions and their relatives. Here is what they are: In statistics, a variable is a measurement of some kind of

[21] There was some erroneous talk of *significant co-lexemes* and similar things in the early papers. We do not consider such quirks relevant as of today, especially as those papers are now twenty years old.

property of events. In the experiments described in this chapter, the variable was always a count of a two-valued underlying variable like success and failure or actives and passives. Other variables we'll encounter measure different things such as lengths or reaction times. The whole purpose of statistics is to quantify the frequentist probabilities of the values of those variables. For example: What's the probability of guessing k cups correctly? What's the probability of encountering a sentence that is 6 words long? What's the probability of measuring a reaction time of 350 ms or higher?

Section 2.2 was devoted to deriving the maths that specify the probability of obtaining $k = 3$ successes (correctly detected tea-first cups) when there are $K = 4$ potential successes (tea-first cups on the table) and we have $n = 4$ choices to make (cups to choose) from $N = 8$ potential choices in total (the overall number of cups). This effort culminated in Equation 2.4 on page 15, repeated here for convenience:

$$\frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = \frac{16}{70} \approx 0.23$$

The general form of this equation is (as you should verify):

$$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{2.6}$$

Now, think about the design of experiments like the Tea Tasting Experiment: The three parameters N, K, and n are fixed by the design, only the number of successes k varies depending on the performance of the taster. If we plot the probabilities (y-axis) of all possible outcomes of k successes (x-axis) while keeping K, N, and n fixed at certain values, we get the graph in Figure 2.8 (left panel). We also provide the plot for the Tea Tasting Marathon from Section 2.4 (right panel). Notice that the grey lines are merely visual helper lines. As the distribution is discrete (see below), it does not define a curve.

The two graphs in Figure 2.8 are examples of the *Hypergeometric Distribution*, the probability distribution that's the basis of Fisher's Exact Test. Since the outcomes k are discrete counts, there is a defined probability for each outcome that can be calculated with arbitrary precision using the discrete *density function* in

Hypergeometric Distribution

density function

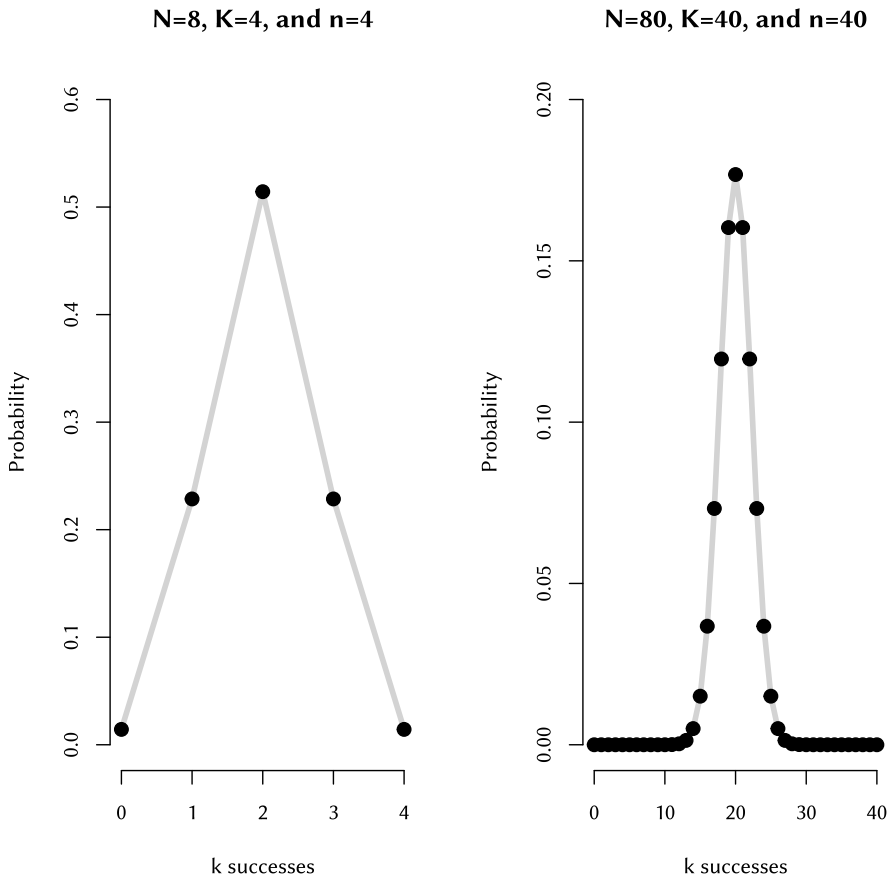

Figure 2.8: Probability density for k successes in Tea Tasting Experiments; the density function of the Hypergeometric Distribution

Equation 2.6.[22]

parameter

Some terminological clarifications are in order. We used the term *parameter* above to describe N, K, and n. Informally speaking, there is not one Hypergeometric Distribution but a large (infinite) class of distributions, one for each fixed permutation of N, K, and n. Thus, the parameters define the concrete function and determine the exact shape of the curve for the admissible values of k,

[22]The term *density function* is a bit unfortunate for thedistributions like the Hypergeometric Distribution as it applies more accurately to continuous distributions (see Chapter 5). Hence, the adjective *discrete* should not be omitted.

which is not itself a parameter. Furthermore, this distribution is *discrete* because it give probabilities for discrete outcomes. After all, there cannot be an experiment where someone guesses 3.5 cups or anything like that correctly. We'll encounter other very much different distributions later.

discrete

A necessary property of such functions is that the sum of the probabilities for all k given a specific set of parameters is 1. Intuitively, this should indeed be a necessary property. For each outcome (a number of successes), the function specifies a probability, and we can be absolutely certain that exactly one outcome will occur if the experiment is conducted. For example, we can only achieve between 0 and 4 correctly classified cups in a classical Tea Tasting Experiment if the experiment is terminated properly (and neither 5 nor 100 nor -1, etc.).[23] The exact maths aside, as exactly one outcome has to occur, the probability of the union of the outcomes must be 1 (see Section 2.2.3).

Finally, for each (discrete or non-discrete) density function there is a *cumulative distribution function*. It gives for each outcome k the summed probabilities of all outcomes to its left and itself. Figure 2.9 shows the two distributions corresponding to the densities in Figure 2.8.[24] The distribution function is highly useful in determining the probability of an outcome as extreme as or more extreme for some k. With Equation 2.4, we calculated the probability of achieving 3 correctly classified cups and 1 incorrectly classified cup, and the result (0.24) is the value of the hypergeometric cumulative distribution function with parameters $N = 8$, $K = 4$, and $n = 4$ at $k = 1$. Similarly, we calculated the probability of achieving 30 correctly and 10 incorrectly classified cups in the Tea Tasting Marathon with Equation 2.5 as $7.44 \cdot 10^{-6}$. This is the value of the Hypergeometric cumulative distribution function with parameters $N = 80$, $K = 40$, and $n = 40$ at $k = 10$.[25]

cumulative distribution function

We will demonstrate several such functions of probability distributions throughout the book. While knowledge of these function is not strictly necessary in order to understand applied frequentist statistics, it might come in handy. First, the proofs underlying frequentism (and Bayesianism, for that matter) make excessive reference to probability distributions. It's therefore helpful for anyone

[23] Notice that the distribution for given parameters N, K, and n only describes properly terminated experiments set up according to those parameters. If the cat jumps onto the table in the middle of the experiment and breaks all cups, the Hypergeometric Distribution is no longer applicable.

[24] Unfortunately, the maths of cumulative distribution functions are often much more complicated than that of the density function, and we do not provide their mathematical formulations. They can be found easily on Wikipedia.

[25] Do you see why it is sufficient to look up the value at $k = 1$ and $k = 10$, respectively? Hint: Contemplate the graphs in Figure 2.9 and think about single-sided tests.

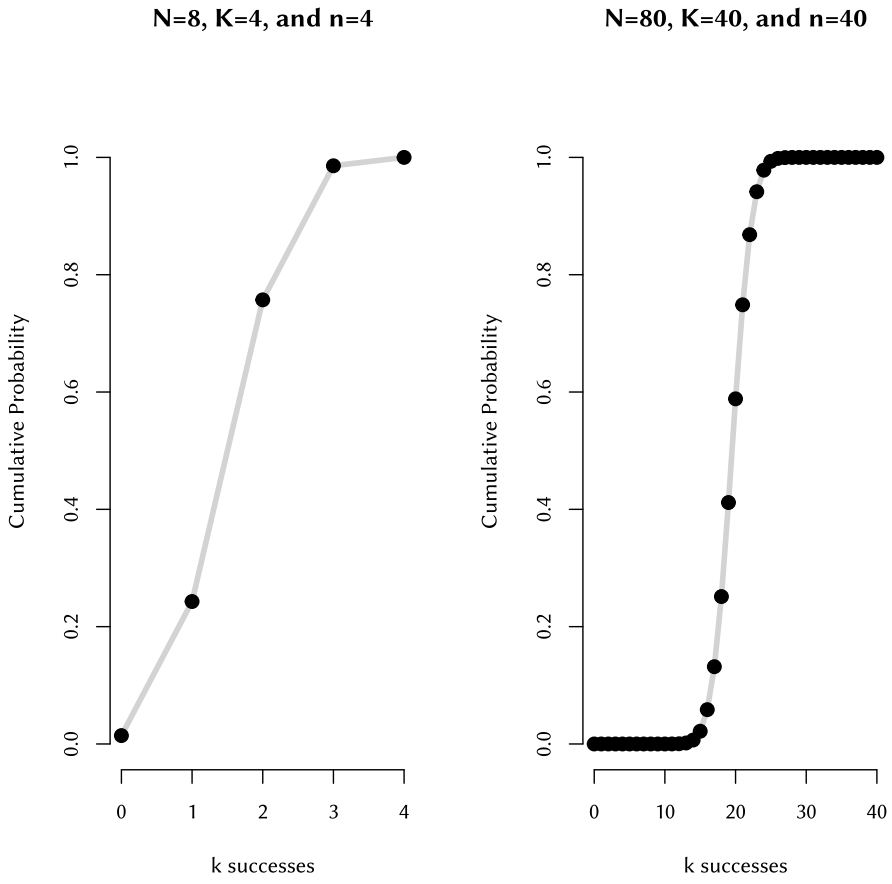

Figure 2.9: Cumulative distribution function for k successes in Tea Tasting Experiments; the distribution function of the Hypergeometric Distribution

who wants to dig deeper and study more sophisticated text books after this one. Second, statistics programs such as R give us access to these functions, and it's much easier to use them with a minimal understanding of their workings.

Exercises for Chapter 2

(1) One Two Three

3 Data: Central Tendency and Variance

Overview

Chapter 2 introduced the logic of frequentist inference using the relatively simple Fisher Exact Test as an example. In order to extend the same logic to other use cases and develop more advanced versions of the same logic, we need to take a step back in this Chapter. Many other tests will require knowledge of measures of central tendency like the mean (or average) and the variance. Variance is a statistic that quantifies how strongly single values deviate from the mean. It helps us to deal with a central reason we need inferential statistics in the first place: Not all measured values are the same. People don't all have the same weight, sentences don't all have the same length, etc. Hence, variance is one of the most crucial concepts in statistics, and it's one of the reasons that making inferences about virtually infinite populations based on finite or even small samples is hard.

While many statistics text books make a huge brouhaha about so-called *descriptive statistics* and *levels of measurement*, we focus those two core concepts (the central tendency and the variance of measurements). We begin by characterising different types of measurements (which are often called levels of measurement). We discuss ways of finding the most characteristic measurements in a sample or a population by finding the mode, median, quartiles and percentiles, and the mean for different types of measurements. Finally, we quantify how much single measurements deviate from the most characteristic measurements in the form of the variance and the standard deviation. Along the way, some types of plots are introduced which provide insight into the structure of data. These are raw data plots, histograms, box plots, density plots, and violinplots.

Problem Statement: Show Me Your Sample!

Imagine you've collected some data. It might be counts of words and constructions from a corpus, acceptability judgements for relative clause constructions on a five-point-scale, word or sentence lengths, reaction times in milliseconds. Ultimately, you want to make an inference, but first you

want to get an overview of your data set and explore it in order to see whether it roughly meets your expectations. Furthermore, you want to publish the results and provide your readers with an overview of the results, as most of them will not look at your raw data. Besides all kinds of plots that aggregate your data visually, you might want to characterise your sample numerically. What's the best numerical characterisation for your kind of data.

3.1 Central Tendency and Typical Values

3.1.1 Binary Measurements

If you go through a corpus of English and look at each noun, it could be a singular or a plural form. We use a paragraph from the Wikipedia article on Critical Rationalism as an example.[1]

> *Critical rationalists hold that scientific theories and any other claims to knowledge can and should be rationally criticized, and (if they have empirical content) can and should be subjected to tests which may falsify them. Thus claims to knowledge may be contrastingly and normatively evaluated. They are either falsifiable and thus empirical (in a very broad sense), or not falsifiable and thus non-empirical.*

In such a case, the measurement consists of a sequence of two possible measurements: singular or plural. The sequence for the highlighted nouns in the sample is:

> *plural, plural, plural, singular, singular, plural, plural, singular, singular*

A main point of this chapter is to introduce the lingo used to describe the properties of such observations in experimental design and statistics. A measurement in an experiment is always characterised by a well-defined *variable*. A variable doesn't always take the form of a numerical measurement (although it can). In any case, it's a description of some property of events. Turning to the example, if a noun occurs, the noun's grammatical number can be characterised as either singular or plural, the two possible values or *levels* of the variable *Number*. In other

variable

levels

[1] Sourced from https://en.wikipedia.org/wiki/Critical_rationalism on 19 February 2025.

words, all events of noun occurrence in English fall into one of two categories: singular or plural. We need some type of formal representation for a series of measurements of such values. We name it **x** in boldprint and regard it as a *tuple*, customarily specified in angled brackets ⟨ ⟩:

tuple

$$\mathbf{x} = \langle plural, plural, plural, singular, singular, plural, plural, singular, singular \rangle$$

A tuple is a set-theoretic structure that (i) has an order, (ii) allows for the same element to occur more than once. Think of it as a list of items where the same item can appear as many times as necessary. It would be perfectly okay to disregard the order of the events because we (albeit often incorrectly, see page 13) assume that the events are independent. From that perspective, we could use a *set* (which is unordered by definition) instead of a tuple, symbolised by curly brackets { } instead of angled brackets ⟨ ⟩. By definition, however, sets don't allow the same element to be added to them twice, which would clearly be inadequate for series of data points such as **x**. We call the individual measurements in **x** x_i, where i is an *index variable* that takes on integer values. The first element in **x** is denoted x_1, the second element x_2, and the last element is always x_n, where n is the sample size. In our example, $n = 9$ and hence x_9 is the last element in **x**. Without digging any deeper into the murky waters of set theory and adjacent areas, we leave it at that. Tuples are called *lists* in Python, *vectors* in R, *arrays* in more substantive programming languages like C++ or Ada, etc.

set

index variable

How are the corresponding results best summarised? Mathematically, one of the few things we can do with these occurrences is to count them, which is exactly what we did with the variables *Bristow* and *Reality* in Chapter 2, both having the possible values *tea-first* and *milk-first* (e. g., Table 2.6). For completeness, we show the very obvious summarisation of these counts in tabular form (Table 3.1). However, we have already shown the more complex type of contingency tables, where the co-occurrences of the levels of two two-level variables per event are tabulated. Hence, Table 3.1 should seem quite unsophisticated to most readers.

Table 3.1: A tabulation of counts of a binary variable

	Singular	**Plural**
Count	4	5

Obviously, the variable *Number* is two-level for languages without duals, trials, etc. such as English. It is called a *binary variable* or *dichotomous variable*.

binary variable

In order to summarise binary data in a single number, we can calculate a proportion, an operation which we assume is well-known (but see below for the general formula). The proportion of singular noun forms in the above sample is $4 \div (4 + 5) = 4 \div 9 \approx 0.4444$. Note that proportions always lie between 0 and 1.

percentage

A *percentage* is simply a proportion times 100, hence the percentage of singular forms is $0.4444 \cdot 100 = 44.44\%$. Consequently, a percentage always lies between 0 and 100. As we'll demonstrate very clearly in Chapter 4, proportions and percentages are fine, but they can be meaningless if the *sample size* is unknown. For binary variables, the sample size n is simply the number of events, hence $n = 9$ in the example. Using the symbol q for proportions (as p is already taken and proportions are always a kind of *quotient*), it should be obvious that for any count c_l of a level l of a binary variable (such as $c_{\text{singular}} = 4$ for *singular* in the example) and it's corresponding proportion q_l (such as $q_{\text{singular}} \approx 0.44$):

sample size

$$c_l = q_l \cdot n$$

proportions

This is because the general formula for calculating *proportions* is:

$$q_l = \frac{c_l}{n} \tag{3.1}$$

For the count of the other level c_m and its corresponding proportion q_m ($c_{\text{plural}} = 5$ and $q_{\text{plural}} \approx 0.56$ in the example), then, necessarily $c_m = (1-q_l) \cdot n$. Also, $q_m = (n - c_l) \div n$. This means that counts of a binary variable are exhaustively summarised by two numbers. Neither can we adequately summarise it by one number (just one count, just one proportion, or just the sample size), nor do we absolutely need to provide more than two numbers (provided it's two well-chosen numbers).[2] Either we specify the sample size n and at least one proportion or one count. Or we specify the raw counts for both levels of the variable. Furthermore, not much is gained by converting counts to proportions, except that some people find proportions more intuitive. We point this out because aggregated numbers for other measurements (to be discussed below) serve a much stronger purpose and are much more informative. For counts of binary variables, any aggregation serves no or just a very limited purpose.

central tendency

Finally, we can ask what is the *central tendency* of a binary variable. The purpose of finding the central tendency is to find a typical value in some sense. The

[2] However, we would like to point out that providing the absolute minimum information is not always the nicest thing to do to your readers. Usually, the more relevant information you provide in your publications, the bettr it is for the reader.

term *central* derives from more complex measurements (see Section 3.1.3 and below), and it's not very intuitive for binary measurements. It will become clearer as we progress through this chapter. In the binary case, the only measure of central tendency is its modal category or just *mode*, which is the level that occurs more often than the other. If $c_l > c_m$, then c_l is the modal category and vice versa. In general, the central tendency is important as the *expected value* of a variable, and the mode of a binary variable is no exception. Without any further information, it's the value that we would predict for any event. If all you knew about English were the short paragraph from Wikipedia quoted above (annotated only for part-of-speech and number, with no information about the syntax, the meaning, the register, etc.), and you were asked to predict the number of the next noun in the text, any sane person would predict *plural* because it appears to be the modal (i. e., more frequent) category judging from the sample. Having much more information about the English language, we are aware that this would probably not lead to a prediction accuracy of 56% in real life. There are two reasons for this: First, producing English nouns is not a mere lottery, and our analysis should reflect this. Second, the sample is very small ($n = 9$) and from a very specific type of text, problems we'll deal with in Chapter 4 and throughout the book. However, if we knew for sure that among the totality of English noun tokens, 5 out of 9 were in the plural, and you had to predict the number of a noun without information about the lexeme, syntactic context, or register of the text, we would always predict plural and be right in 56% of all cases in the long run. In this sense, it would be the *expected value* or *expectation*.[3]

mode

expected value

Finally, Figure 3.1 shows a histogram of the sample of the *Number* variable from the illustrative example discussed above. A histogram for a variable with discrete levels is usually displayed with spaces between the bars (compare Figure 2.6, where there isn't any space between the bars), and it's often called a *bar plot* instead of a histogram. It's still just a special type of histogram. For the benefit of the reader, we have highlighted the mode in colour. This is the only reasonable way to plot such results. We strictly recommend to label the individual bars with the raw counts, not proportions or percentages. Since the bars already provide a good visual impression of the proportions, the relevant numerical information are the counts. At the same time, we can't think of a situation where such a plot by itself isn't an insult to the reader. Looking at just two counts (except maybe very high counts in the millions and above), humans are able to grasp the relative magnitude of those counts easily without visualisation.

bar plot

[3] The difference between an expected value estimated from a sample and one that's true for the population is the subject of Chapter 4.

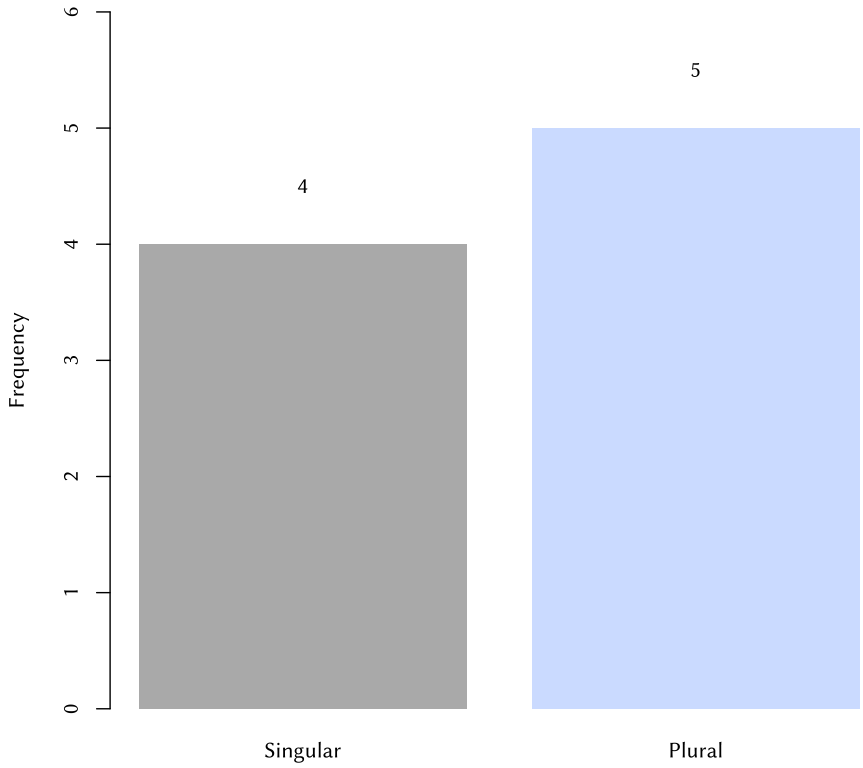

Figure 3.1: Histogram (or bar plot) of counts of a binary variable with the mode highlighted in colour

3.1.2 Multi-Valued Measurements

In this chapter, we continually progress from simple to more complex variables (i. e., measurements). After binary variables, the next more complex thing is simply a more general variant of a binary variable. Let's have a look at a short paragraph from the German Wikipedia article about Critical Rationalism and annotate all lexical nouns with their grammatical case:[4]

[4]Sourced from https://de.wikipedia.org/wiki/Kritischer_Rationalismus on 20 February 2025.

> *Der Realismus$_{Nom}$ ist die dem subjektiven Idealismus$_{Dat}$ widersprechende metaphysische Theorie$_{Nom}$, dass eine vom Menschen$_{Dat}$ unabhängige Wirklichkeit$_{Nom}$ existiert. Während der naive Realismus$_{Nom}$ davon ausgeht, dass die Welt$_{Nom}$ so ist, wie der Mensch$_{Nom}$ sie wahrnimmt, vertritt der kritische Realismus$_{Nom}$ die Auffassung, dass Vorstellungen$_{Nom}$ von ihr durch subjektive Elemente$_{Acc}$, die in der Wahrnehmung$_{Dat}$ und im Denken$_{Dat}$ liegen, mehr oder weniger stark beeinflusst werden. Weil die Sinne$_{Nom}$ und die Verarbeitungsprozesse$_{Nom}$ im Gehirn der angenommenen Außenwelt$_{Dat}$ und der Vorstellung$_{Dat}$ zwischengeschaltet sind, kann man auch vom indirekten Realismus$_{Dat}$ sprechen. Dieser Vermittlungsvorgang$_{Nom}$ schließt eine reine Wahrnehmung$_{Nom}$ aus, denn es kann sich um Täuschungen$_{Nom}$ handeln.*

There are four grammatical cases in German: nominative, accusative, dative, and genitive. They occur at 13, 1, 7, and 0 noun events in the sample, respectively. The sample of measurements looks as follows:[5]

> $\mathbf{x} = \langle$*nominative, dative, nominative, dative, nominative, nominative, nominative, nominative, nominative, nominative, accusative, dative, dative, nominative, nominative, dative, dative, dative, nominative, nominative, nominative*$\rangle$

The variable *Case* measuring grammatical case in German has four levels which have no empirically significant order. Instead of the order nominative, accusative, dative, genitive, many grammars and text books order the four cases according to the Latin tradition: nominative, genitive, dative, accusative, which is perfectly fine, too. At this point, some linguists might object and point out that we should order them by increasing obliqueness or in a way such that syncretisms are analysed better (e. g., Eisenberg 2020). However, such ordering preferences are not empirical observables. They are theoretical results rather than measurements. We specifically chose this example in order to point out this difference. A variable which has more than two discrete levels with no measurable ordering is called a *nominal variable*.

nominal variable

Table 3.2: A tabulation of counts of a nominal (four-level) variable

	Nominative	**Accusative**	**Dative**	**Genitive**
Count	13	1	7	0

[5] We reuse the symbol **x** for any tuple containing measurements from an experiment. If we need to talk about more than one such tuple, we call the others **y** and **z**.

Binary variables are really just the minimal case of a nominal variable. Not surprisingly, we can treat them essentially the same, for example by tabulating the counts as in Table 3.2. The only reasonable aggregations are conversion of raw counts to proportions as in Equation 3.1. The *mode* is just the most frequent value, much like in the binary case, and we consider it trivial to find it for smaller samples. We leave it to the reader to find the mode and calculate the proportions and percentages for each of the four grammatical cases in the sample. Two histograms (more specifically bar plots) for the example are shown in Figure 3.2. Since there is no measurable order in grammatical cases, both plots are equivalent, and you could choose the one that you think is more adequate for your purpose.

mode

3.1.3 Ordered but Discrete Measurements

Moving on to the next more complex type of measurement and variable, we use acceptability ratings as an example. Such ratings are often made on a scale with 5 or 7 points, for example *very high*, *high*, *medium*, *low*, *very low*. For illustration purposes, assume that a rating experiment was conducted where participants were asked to rate the sentence: *It's a mission to boldly go where no man has gone before*. The study had 10 participants, hence $n = 10$. The raw data are, using $\mathbf{x}$ again to denote the tuple of measurements:

> $\mathbf{x} = \langle$*low, medium, very low, medium, very high, high, medium, medium, very low, medium*$\rangle$

We hope that the difference between such a variable—let's call this one *Rating*—and a nominal variable like *Case* from Section 3.1.2 is obvious. The ratings have an intrinsic order, and each level of the variable has a rank: *very high* is higher than *high*, *high* is higher than *medium*, and so forth. A variable like *Rating* is called an *ordinal variable*. All possible ways of aggregating nominal variables (including binary ones) apply to ordinal variables as well: counting and tabulation, calculation of the proportions or percentages of the different levels, the determination of the mode, and bar plots (histograms). The only difference is that tables and plots should respect the intrinsic order of the ordinal variable, as can be observed in Table 3.3 and Figure 3.3.

ordinal variable

Thanks to the higher complexity of this variable (the intrinsic order of its levels), we have an additional measure of central tendency available. To find this measure called the *median*, we first need to sort the raw measurements (not the

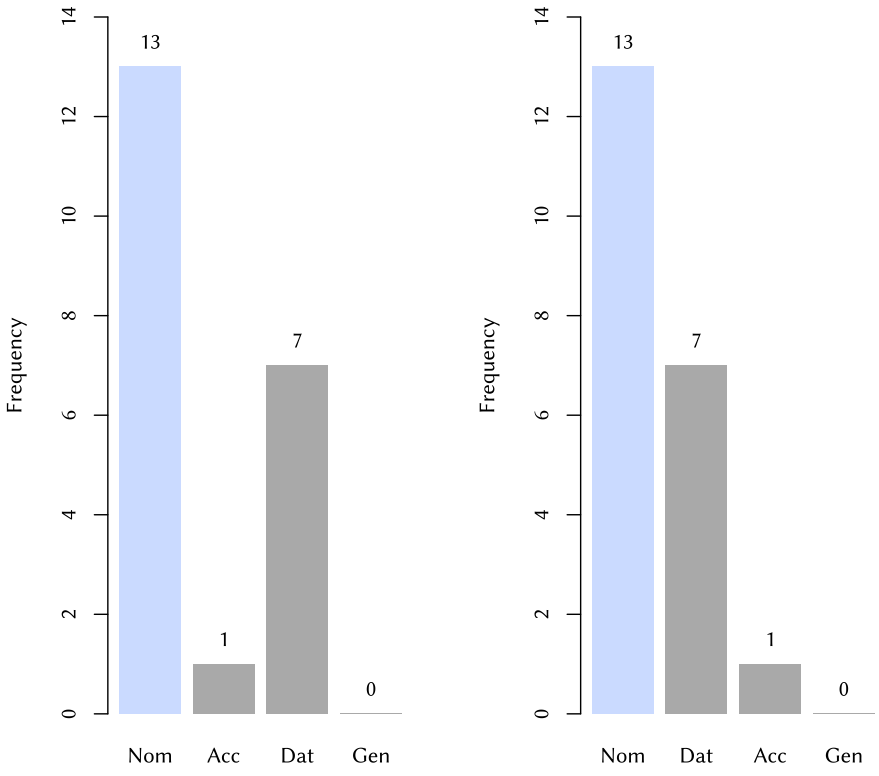

Figure 3.2: Two equivalent histograms of counts of a nominal four-level variable (grammatical case in German) with the mode highlighted in colour

tabulation of the counts) from the lowest rank to the highest rank, which is something we couldn't do with a nominal variable due to the lack of an intrinsic order of its levels. The sorted sample $\mathbf{x}'$ looks like this:

$$\mathbf{x}' = \langle \textit{very low, very low, low, medium, medium, medium, medium, medium, high, very high} \rangle$$

Plotting these single observations from left to right on the x-axis and putting dots on the y-dimension corresponding to the respective ratings, we get Figure 3.4, a *raw data plot*. The grey line is just a helper line for better visual ori- raw data plot

Table 3.3: A tabulation of counts of an ordinal variable

	Very Low	Low	Medium	High	Very High
Count	2	1	5	1	1

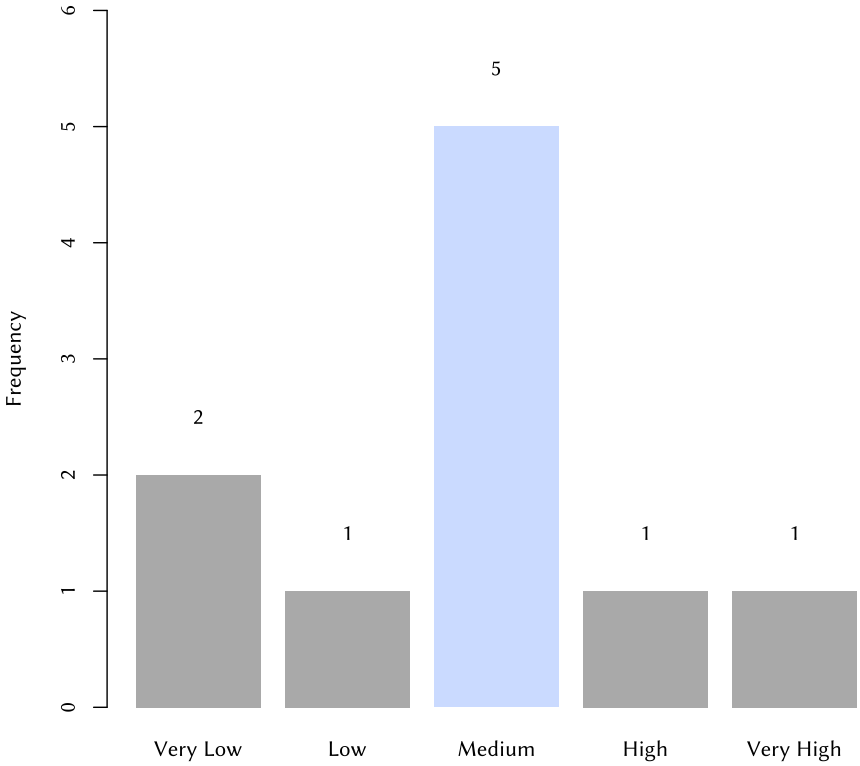

Figure 3.3: Histograms of counts of an ordinal five-level variable (acceptability ratings) with the mode highlighted in colour

entation. The coloured line marks the middle point of the ordered sample in the sense that half the ordered sample lies to its left and the other half to its right.

median

The value that lies in the middle of the ordered sample (marked by the coloured line in Figure 3.4) is the *median*. In the example, the median technically lies be-

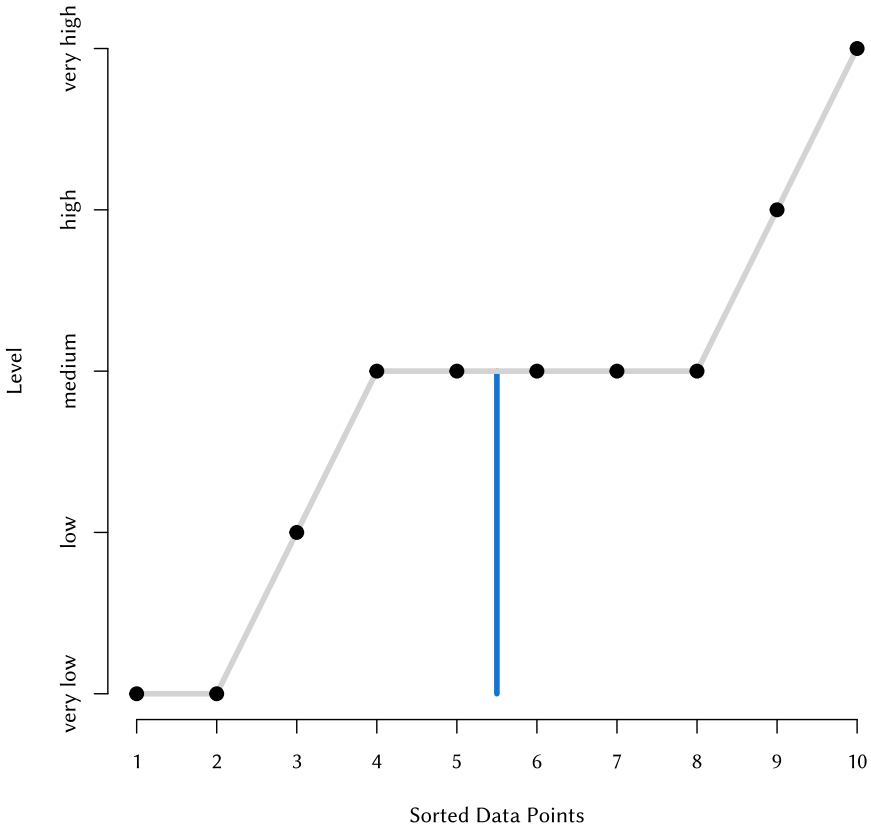

Figure 3.4: Raw data plot of a sorted sample of an ordinal value; the coloured line marks the position of the median

tween two data points (the 5^{th} and the 6^{th} one) as $n = 10$ is an even number. Fortunately, the data points to left and right of the median point have the same rating value, so we can say without further ado that $\tilde{\mathbf{x}} = \textit{medium}$, where we denote the median of $\mathbf{x}$ by $\tilde{\mathbf{x}}$. If, on the other hand, the median point lied between two data points with different values (such as *low* and *medium*), there is no clean way to name the median. One solution would be to say truthfully that the median lies between *low* and *medium* or that it lies at *low*+ (as in the American educational grading system with A−, B+, etc.). Another solution would be to convert the levels from *very low* to *very high* to the numbers between 1 and 5

and state that $\tilde{\mathbf{x}} = 2.5$. For ordinal scales, we strongly advise against the latter solution (conversion to numbers) as it creates the false impression that the ranks correspond to numerical values (see Section 3.1.4).

Starting with the median, we can get a clearer idea of why we speak of measure of central tendency. In a well-defined way, the median position is at the *centre* of the sample. As the expected value of an ordinal variable, however, the mode is probably still the best metric. If *medium* actually is the rating assigned most often, then guessing *medium* would lead to the highest possible success rate. However, median is of course an informative measure, and we'll see in the next section that it's very handy if combined with even more informative measures like the mean.

Likert scales

We've got one final word of warning before moving on to numeric measurements in the next section. Rating scales with 5, 7, 12, or any number of levels are often called *Likert scales* in research papers, text books, and so on. Authors often use the opportunity to point out rather smugly that the correct pronunciation is ['lɪkət] (or some transatlantic variant) and not ['l͡aɪkət]. For example, the Wikipedia article on ordinal data states incorrectly (after providing the correct IPA transcription of the name):[6]

> *A well-known example of ordinal data is the Likert scale.* […] *Examples of ordinal data are often found in questionnaires: for example, the survey question "Is your general health poor, reasonable, good, or excellent?" may have those answers coded respectively as 1, 2, 3, and 4.*

It is not true that any 5-point-scale is a Likert scale. A Likert scale is a complex psychometric construct invented by a person called Likert and operationalised through a set of responses to Likert items. The responses are measured on a 5-point agreement scale, but they are aggregated into the underlying Likert scale. Since the Likert scale cannot be measured but is merely reconstructed from measured responses, we recommend to avoid the phrase *Likert scale* entirely and just call 5-point scales *5-point scales*, etc.

3.1.4 Numeric Measurements

Finally, we turn to numeric measurements. As an example, we use fictitious EEG measurements from an impressive cap-and-cable experiment. We made peoples brains react to split infinitives and measured some negative or positive change

[6] Sourced from https://en.wikipedia.org/wiki/Ordinal_data on 22 February 2025.

in some potential in milliseconds. The 10 measurements ($n = 10$) in our now familiar tuple $\mathbf{x}$ are as follows:[7]

$$\mathbf{x} = \langle 177, 187, 226, 248, 250, 312, 339, 339, 351, 382 \rangle$$

Such numeric measurements have a very clear natural order as they are measured in floating point numbers or integers. It is mathematically and conceptually absolutely clear that 153 is a shorter reaction time than 172, etc. This is why we took the liberty of pre-sorting $\mathbf{x}$ from lowest to highest. While such a numeric measurement defines a ranking like an ordinal measurement, more information is contained in numbers than merely a ranking. Most importantly, the distance between two measurements can be quantified beyond mere ranks. We could say that *high* is 2 ranks above *low* in the example from Section 3.1.3. However, saying that 187 ms is 10 ms longer than 177 ms is much more informative. It cannot (and need not) be determined whether *good* is twice or three times as good as *bad*, but 2 ms are definitely twice as long as 1 ms, and so forth.

numeric variables

Mathematically speaking, *numeric variables* allow us to use standard arithmetic (adding, subtracting, multiplying, dividing), which we cannot do with any of the less complex variables. Sometimes, a distinction is made within the numeric measurements: Variables on a *ratio scale* are those with a defined zero measurement where no measurement below 0 exists. Reaction times or word lengths are good examples as it's impossible to measure negative reaction times or words that are -1 phonemes or graphemes long. The other type of numeric variable is measured on a so-called *interval scale*, which does not have a zero floor and allows measurements towards negative infinity. As such scales never occur in linguistics (and, we think, most empirical sciences) and the distinctions between ratio and interval scales can otherwise be safely neglected, we simply speak of numeric variables.

ratio scale

interval scale

Regarding the central tendency of numeric measurements and the applicable plots, numeric measurements offer many more options and introduce some complications. Let's begin with a raw data plot of the sorted sample in Figure 3.5.

The mode is difficult to determine because we defined it as the most frequent value. However, each measurement is unique, which is intuitively a likely outcome for an experiment measuring anything in millisecods. We'll return to the concept of the mode later. The *median* point is between the the 5^{th} and the 6^{th} data point. The respective measurements are $x_5 = 250$ ms and $x_6 = 312$ ms. It is customary to calculate the arithmetic middle of the two values and declare it

median

[7] Yes, we reuse the symbol $\mathbf{x}$ every time. In real data analysis, one should always use fresh and informative names, of course. Never call real data $\mathbf{x}$, $\mathbf{y}$, etc.

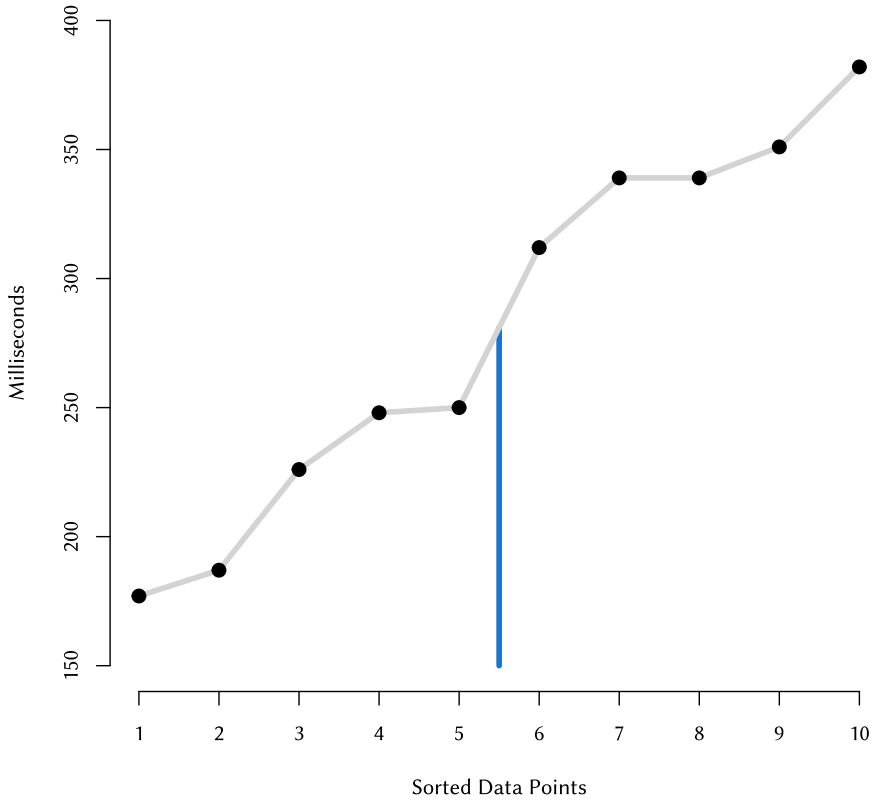

Figure 3.5: Raw data plot of a sorted sample of a numeric variable; the coloured line marks the position of the median

the median: $\tilde{\mathbf{x}} = (250 + 312) \div 2 = 281$ ms. Hence, half of the participants had a faster reaction than 281 ms, and half of the participants had a slower reaction.

While the mean is informative for numeric variables, most readers probably already know that there is one additional operation we can perform to find the central tendency: the *mean* (colloquially called the *average*). It is obtained by adding all measurements and dividing them by the number of measurements (i. e., the sample size):

mean

$$\frac{177 + 187 + 226 + 248 + 250 + 312 + 339 + 339 + 351 + 382}{10} = \frac{2811}{10} \approx 281$$

In this example, the mean is virtually identical to the median. This is typical of symmetric samples. If you look at Figure 3.5, the data points are arranged almost along a straight line from the lower left to the upper right, and the sample is symmetric in this sense. We'll encounter cases where the mean and the median are not the same.

There is a convenient notation used in the general definition of the mean. It's the *sum operator* Σ, which is used like this: sum operator

$$\sum_{i=1}^{n} x_i$$

We use n consistently to denote the sample size ($n = 10$ in the example). The symbol x is used to denote individual values from the tuple $\mathbf{x}$, i. e., the measurements from the experiment. The index i is a counter and tells us which concrete x_i we pull from $\mathbf{x}$. The limits of the sum operator use these symbols to go through the whole tuple $\mathbf{x}$. Starting from $i = 1$ (the lower limit) and counting all the way up to n (the upper limit), it sums up all x_i. Since x_1 is the first element of the tuple and x_n (x_10 in the example), all values from $\mathbf{x}$ are added up. Writing the general form of the sum operator more explicitly, we get:

$$\sum_{i=1}^{n} x_i = x_1 + \cdots + x_n$$

We'll encounter sum operators with more complex expressions than x_i, but the logic remains the same. It's really just the instruction to add some numbers. The general equation to calculate the mean $\bar{\mathbf{x}}$ of a tuple $\mathbf{x}$ is thus:

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.2}$$

As abbreviated variants of the sum operator, we allow the following to denote the sum of all values from $\mathbf{x}$. Either the upper limit n is implied (second variant), or we just pass the symbol for the whole tuple to the sum operator (third variant), in which case no index variable (i) is required:

$$\sum_{i=1}^{n} x_i = \sum_{i} x_i = \sum \mathbf{x}$$

Therefore, alternative notations for the mean are:

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i} x_i}{n} = \frac{\sum \mathbf{x}}{n}$$

The mean and the median are quite informative measures of the central tendency. Finally, we'll now generalise the concept of the median, and we'll reintroduce the mode for numeric variables (which we glossed over at the beginning of this section). Figure 3.6 repeats Figure 3.5 with some additional vertical lines.

The lines mark the aforementioned generalisations of the median. Going outward from the median, the lines at data point 3 and 8 define the lower and upper *quartile*. With the median, they split the sorted sample in four parts called the first (lower), second, third, and fourth (upper) quartile. For numeric samples, it's also informative to inform readers about the minimal and maximal value, which are also marked in the plot. Contrary to an ordinal variable, a numeric variable does not have a pre-defined minimal and maximal value such as *very low* and *very high*. Table 3.4 shows all the information discussed so far that can be used to summarise a numeric variable.[8]

quartile

Table 3.4: Summary of a numeric variable with n=10

	Value
Minimum	177
First Quartile	226
Median	281
Mean	281
Third Quartile	339
Maximum	382
Sample size	10

For a sample with 10 data points, providing a summary in 7 values might not appear to be much of a summary. When samples get bigger, however, this changes very quickly. Table 3.5 shows the same for a very similar sample but with $n = 100$.

The summary characterised a sample of 100 values in 7 values, which is much more economical than just showing all 100 values. Also, most humans aren't very

[8]Notice that statistics software packages don't always calculate the quartiles in the same naïve way we did. There is a number of algorithms available that adjusts the quartiles based on the overall distribution of values in the sample. Please consider the documentation. It's particularly embarrassing if you first notice the resulting discrepancies while teaching with R.

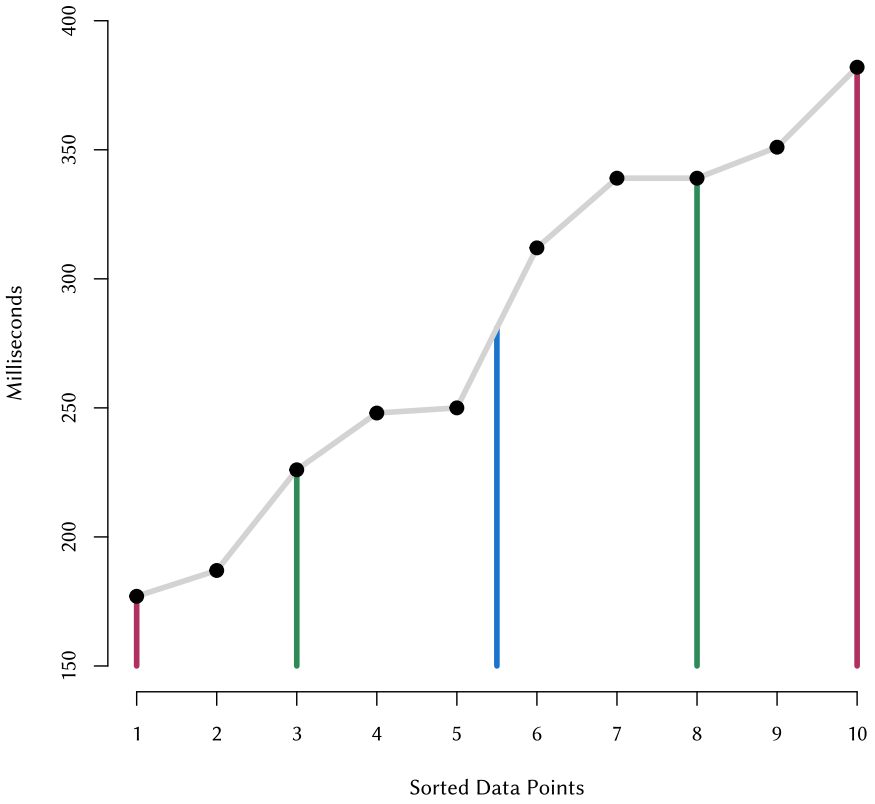

Figure 3.6: Raw data plot of a sorted sample of a numeric variable; the coloured lines mark the position of the median, the lower and upper quartile, as well as the minimum and the maximum

good at processing 100 raw values and get an idea of what the tendencies and trends in the data are. A histogram as in Figure 3.7 (left panel) is a helpful companion to the table. Like in Chapter 2, it's a histogram of a continuous variable, and the individual bars represent counts of occurrences of values from a certain range:

the first bar counts the values that lie between 140 ms and 160 ms, and so on. Histograms of continuous variables are plotted without spaces between the bars, unlike the bar plots used for discrete variables.

Table 3.5: Summary of a numeric variable with n=100

	Value
Minimum	149
First Quartile	227
Median	250
Mean	250
Third Quartile	273
Maximum	348
Sample size	100

Another plot with a similar informative value is shown in the right panel of Figure 3.7. It shows an estimated density function for the sample. Probability density functions were introduced in Section 2.6 for a known *theoretical probability distribution*, namely the Hypergeometric Distribution. While the plots look conceptually very similar, there are major differences between the plots of the Hypergeometric Distribution and Figure 3.7. First, the Hypergeometric Distribution is the distribution of a discrete variable (k successes), and the estimate in Figure 3.7 is for a continuous variable. While k in *k successes* is always an integer and there is no such thing as 3.84 successes, time can in principle be measured with arbitrary precision. Second, Figure 3.7 shows an *empirical distribution*. In other words, it's just a curve guessed from a set of measurements, while the Hypergeometric Distribution is a mathematically well-defined known distribution.[9]

theoretical probability distribution

empirical distribution

However, Figure 3.7 is still a plot of a probability density function. Instead of the individual values adding up to 1 (as with discrete distributions), the area under the curve (the integral of the density function on the interval from $-\infty$ to $+\infty$) is 1. When a statistics software estimates such a curve, it applies a process of inter- and extrapolation from the measured values in order to arrive at a smooth curve, which is also rescaled in order to fulfil the requirement that the area under it is 1. Once such a function has been determined, it is possible to look up the hypothetical proportion (equivalent to the estimated probability) of some value, even if that precise value was not in the original sample. For example, a measurement of 155.34 ms occurs in a hypothetical proportion of 0.00044 of all cases (i. e., has the estimated probability of 0.00044). Finally, the density estimate allows us to define the *mode* of a numeric variable. It's simply the maximum of the curve (or an interval around that point). In the example, the highest probability density

mode

[9] Chapter 4 deals with this difference in some detail.

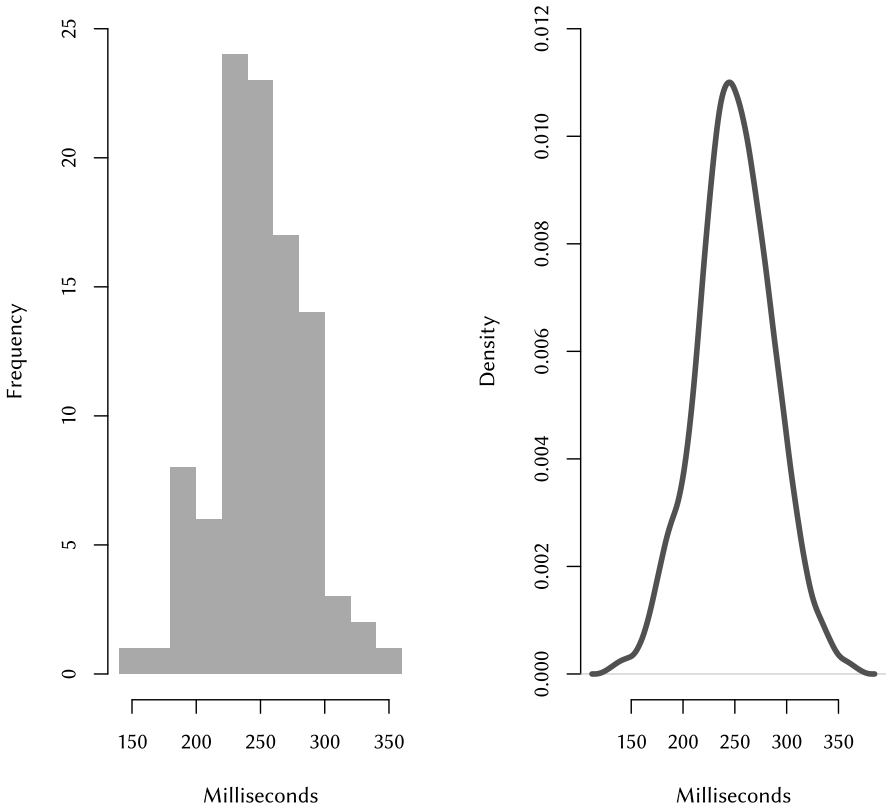

Figure 3.7: Histogram of a sample of a numeric variable with n=100 and a corresponding density plot based on an estimated probability density function

is around 244, and it's 0.011. In this almost perfectly symmetric sample, the mean, the median, and the mode are apparently very close to each other.

Whether a plot of the density estimate is more informative than a histogram depends on the type of data and the research question. Having the two plots side by side in Figure 3.7 should reveal that they contain virtually the same information. However, before using density plots one should do some research as to which specific method the chosen piece of software uses in order to check

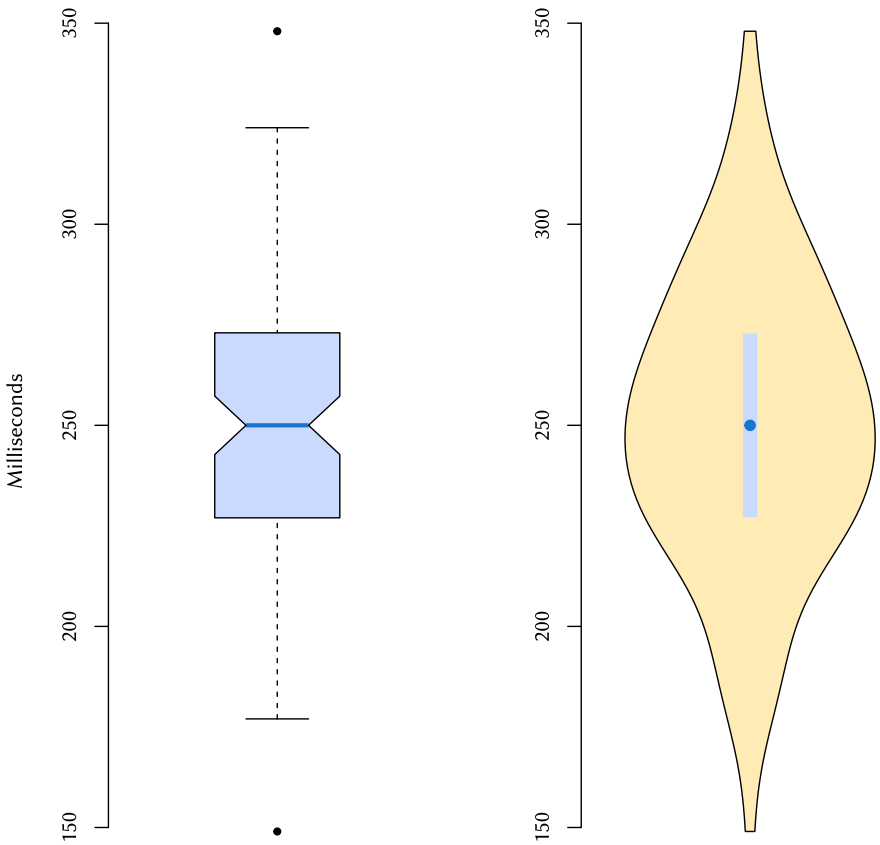

Figure 3.8: A box plot and a violin plot for a numeric sample with n=100

whether it delievers the desired result.[10]

We need the above information about density estimates to understand another type of useful plot. Figure 3.8 shows a box plot (left) and a violin plot (right) of the 100 measurements of milliseconds. A *box plot*—more specifically a box-and-whiskers plot—summarises the distribution of values in sample vertically, and it contains information roughly equivalent to Table 3.5. The line in the middle is at the median position (here 250), and the box around it spans the range from the first quartile (here 227) to the third quartile (here 273). While the box shows in

box plot

[10]Look under *smoother*, *smoothing*, or *kernel*. We use a cosine kernel by default.

which range the central half of the sample values lie, the whiskers are supposed to show the range in which most of the sample values lie. It's usually based on the *inter-quartile range* or IQR. The IQR is the distance between the third and the first quartile (here $273 - 227 = 46$). The IQR is multiplied by 1.5 and subtracted from the first quartile (hence $227 - (273 - 227) \cdot 1.5 = 158$) as well as added to the third quartile (hence $273 + (273 - 227) \cdot 1.5 = 342$) to define the lower and upper bound of the whiskers interval. The concrete length of the whiskers goes up to the closest data points within this interval. Finally, the dots show individual data points that lie outside of the whiskers interval. The idea behind a box plot is to give a straightforward visual impression of where the centre of the sample, the middle 50%, and the majority of the data points lie, and how many data points are so-called *outliers* far away from the majority range.

inter-quartile range

outliers

The violin plot is similar but offers additional information. The dot denotes the median, the box the middle two quartiles, and the body adds a density estimate around those. Arguably, this provides more detailed information than a box-and-whiskers plot.[11] It is definitely useful when the sample is very irregular and has a bumpy distribution. Under such circumstances, the quartiles and the mean tend to create an impression of homogeneity and cleanliness that is unwarranted.

However, both box plots and violin plots are not *innocent* in the sense that there isn't one natural way to plot them. When you use these plots, you're already doing data analysis. Even quartiles—as we mentioned above—can be calculated in different ways, and the decision of what counts as an outlier definitely isn't an automatic one. Some box plots use larger or smaller intervals than $\pm 1.5 \cdot IQR$ for the whiskers, which increases or decreases the number of data points plotted as outliers. Maybe that's warranted for some data sets, but maybe it hides properties of the data set that you and your readers should be aware of. In violin plots, different algorithms used for the density estimate can change the appearance of the plot significantly. Furthermore, symmetric violin plots could be criticised for visually doubling the area under the curve. This can lead readers to overestimate the central areas and underestimate the marginal areas. While many people seem to be enthusiastic about violin plots, at least some of us have very lukewarm feelings about them.

All in all, we advise anyone to use such plots carefully and wisely. Always do the appropriate research to find out what your statistics package actually does when creating the plots. When reading and critiquing published research, always ask whether you understand what was plotted and whether the authors are sufficiently transparent about their plotting methods. Above all, next time somebody

[11] This is the most basic version of a violin plot. There are variants which also add whiskers, outliers, and other information.

shows you a plot that aggregates data in any way, ask persistent questions, especially if you get the feeling there is relevant information missing regarding the genesis of the plots. To our embarrassment, we have (in the distant past) also fallen into the trap of presenting graphics that we didn't fully understand, and we now know that these graphics were suboptimal and potentially misleading. Strangely enough, nobody ever asked us any persistent questions.

3.2 Populations, Samples, and Variance

3.2.1 Populations and Samples

In Section 3.1 we introduced statistics for the central tendency in a sample. Before we move on to discussing the spread of a sample and it variance, we need to mention an important point regarding samples and their populations. As we explained in Chapter 1, the goal of scientific statistical inference is to find out something about a population based on a small sample from that population. The population is the totality of the objects of the type in which we've got some scientific interest. It can be anything from all galaxies in the Laniakea Supercluster to all adult Tories in Buckinghamshire to all sentences of contemporary German. Often, the population is itself a proxy for some bigger research question. Observations of galaxies in the Laniakea Supercluster might be used to test a theory of gravity, a study of Buckinghamshire Tories could be used to test a theory from Social Psychology, and maybe a close examination of all sentences of German will reveal how the brain processes Split Mood Phrases (SplitMP).

While we illustrate statistics such as the mean using small samples in this book, they are all defined for populations, too. At least conceptually, for each variable that describes a property of all objects in the population we can calculate the applicable central tendency, the variance (see immediately below), and so forth. The galaxies in Laniakea have a mean total mass, Buckinghamshire Tories have a mean income, and German sentences have a mean length in phonemes, words, or what not. Why is everything so sample-centric in this chapter, then? First, it's much easier to do exercises where you have to calculate the mean of 10 values rather than the mean of infinitely many (or even hundreds of thousands) values. Second, we don't know all values. The whole point of statistical inference is to generalise from a relatively small sample to a population, simply because we can't look at every single object in the population.

The story should go like this: Our theory makes a prediction about some parameter of a population (say, German sentences in a certain register have a mean length o 6.8 words), and we try to test this prediction with a sample. However,

for this logic to apply, population values must have that parameter (such as a mean) in the first place. Keep this in mind whenever we introduce any statistic for a sample. The same statistics should always be defined (as a parameter) for the population as well.

3.2.2 Spread and Shape of a Distribution

So far, we've discussed how to summarise the central tendency of a sample and how to plot it. We've also argued that the same measures that we use to describe samples also describe populations. As we'll discover throughout the book, it's equally important to quantify how spread out the values from a sample or a population are. Let's illustrate this using two samples. One is the larger sample of reaction times from Section 3.1.4 (called **x** here), the other one is a sample that is superficially similar (called **y**). Compare the summary of both samples in Table 3.6.

Table 3.6: Summary of two numeric samples with n=100

	x	**y**
Minimum	149	95
First Quartile	227	201
Median	250	244
Mean	250	252
Third Quartile	273	304
Maximum	348	398
Sample size	100	100

The two samples look similar in some ways. The means are almost identical (250 and 252) and the medians are also close (250 and 244). However, the minima and maxima as well as the first and third quartiles paint an interesting picture. The extremes appear to be spreading out much more in sample **y** (ranging from 95 to 398) compared to sample **x** (ranging from 149 to 348). This tendency is also visible with the first and third quartiles, which are much further apart in sample **x** (ranging from 201 to 304) than in sample **x** (ranging from 227 to 273). These numbers indicate that the values of the sample are on average somewhat farther away from the mean in **y**. The *spread* of **y** is larger than the spread of **x**. Both a histogram and a density plot should be used to corroborate this and to find out what the shapes of the distributions are. Figure 3.9 shows both types of plots for both samples.

spread

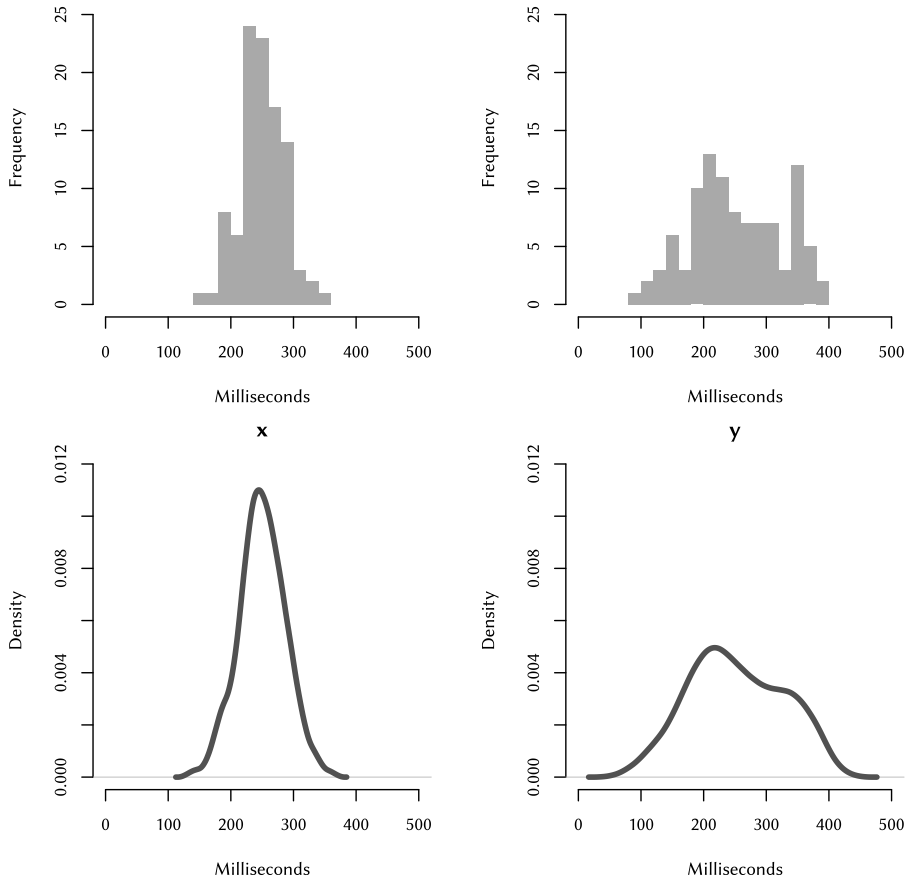

Figure 3.9: Histograms and density plots of two different samples of a numeric variable with n=100

While the distribution of **x** quickly rises and falls in on neat spike, the distribution of **y** rises more slowly, then drops again, only to rise again. By the way, notice the effect of the density estimator: It averages over larger intervals. The second spike around 350, which is clearly visible in the histogram, is so narrow that the averaging going on in the density estimate caps it significantly. However, in both types of plots it looks like **y** has two local maxima, which accounts for the larger spread. Handling data that are distributed like this requires much more care compared to data that looks like **x**, and we'll return to this later in the book.

As we've already pointed out, the data points in **y** are on average farther away from the mean than the ones in **x**, although the means of both samples is virtually the same. While the specific shapes of the distributions can't be summarised in one handy statistic, we have one statistic for the overall spread of a sample: the *variance* (and the *standard deviation* derived from it).

3.2.3 Quantifiying Variance

In order to illustrate the calculation of the variance, we return to the smaller sample from Section 3.1.4 and call it **x**. We repeat the raw data here for convenience:

$$\mathbf{x} = \langle 177, 187, 226, 248, 250, 312, 339, 339, 351, 382 \rangle$$

Its mean is 281. As we've argued that the variance is related to the distances of the individual points around the mean, we begin by plotting the data points (x-axis) and their measured reaction times (y-axis) in Figure 3.10. We also add the mean as a horizontal line plus lines measuring the distance from the mean to the individual data points (using colour to differentiate between negative and positive distances).

Obviously, 5 data points deviate negatively from the mean, and 5 deviate in the positive direction. The way to calculate those distances is $x_i - \mathbf{x}$ for each point x_i Table 3.7 show the results.[12]

Table 3.7: Distances of individual data points to the mean (rounded to integers)

	Measurement	Distance
Data point 1	177	$177 - 281 = -104$
Data point 2	187	$187 - 281 = -94$
Data point 3	226	$226 - 281 = -55$
Data point 4	248	$248 - 281 = -33$
Data point 5	250	$250 - 281 = -31$
Data point 6	312	$312 - 281 = 31$
Data point 7	339	$339 - 281 = 58$
Data point 8	339	$339 - 281 = 58$
Data point 9	351	$351 - 281 = 70$
Data point 10	382	$382 - 281 = 101$

[12]The mean is actually $\bar{\mathbf{x}} = 281.1$. We took the liberty of rounding it to an integer for clarity.

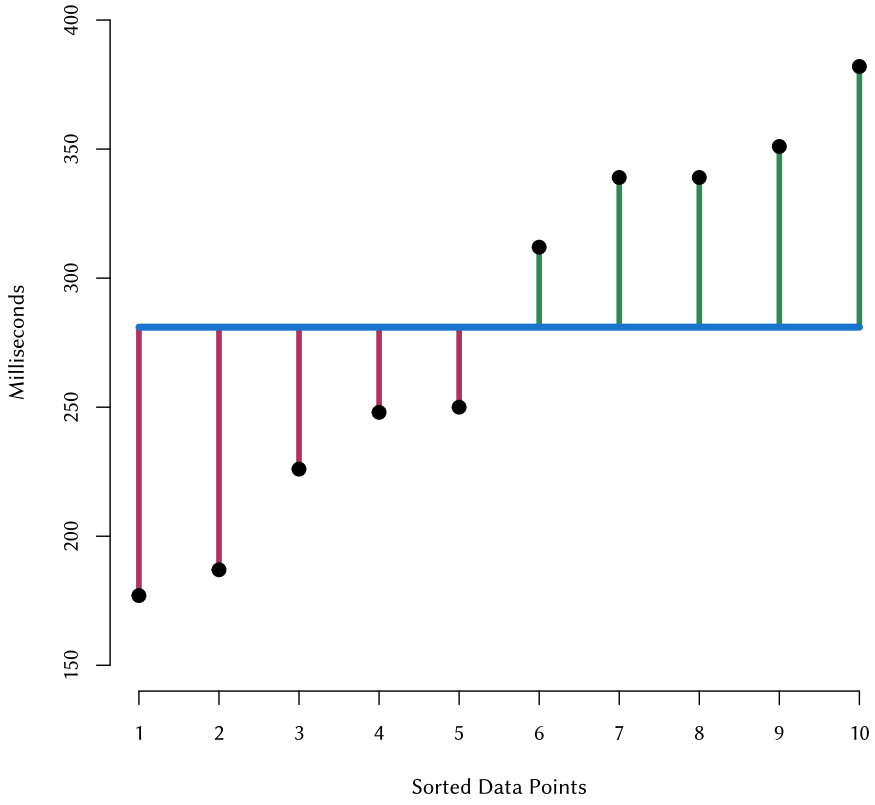

Figure 3.10: Raw data plot of a sample with n=10; distances to mean plotted as vertical lines

Simply adding those distances wouldn't be very helpful. Do you see why? The problem is that negative and positive deviations cancel each other out:

$$\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}}) = \sum_{i=1}^{n}(x_i - \frac{\sum_{j=1}^{n} x_j}{n}) = 0$$

absolute

It would be best to make all values positive. One option would be to take the *absolute* value, which is achieved by discarding the minus sign. While this is an

option, another one is more appropriate for many further calculations in statistics: taking the square of the distance. This makes their absolute value much bigger, but it also means that negative signs cancel out. If we then sum up the squared distances, we get the squared total deviation of the data points from the mean. It's called the *sum of squares* of a sample $\mathbf{x}$, and we abbreviate it as $SQ_\mathbf{x}$, or simply SQ if there is just one sample.[13] sum of squares

$$SQ_\mathbf{x} = \sum_{i=1}^{n} (x_i - \bar{\mathbf{x}})^2 \tag{3.3}$$

While the SQ is the *total* (summed) squared deviation from the mean, we're rather looking for the *average* deviation per data point. We can get there by dividing SQ by n, which gives us the so-called sample *variance* of $\mathbf{x}$ denoted by $s^2_\mathbf{x}$, or simply s^2 if it's clear that we're talking about one sample only. There is a minor catch: We're going to use the variance from a sample as an estimate of the variance in the corresponding population. Statisticians have found out that small samples tend to underestimate the variance of the population and that the best practical remedy is to divide by $n - 1$ instead of n. The larger n, the smaller the effect of this correction, and for populations, the correction is by definition unnecessary. Thus, the sample variance is customarily computed as according to Equation 3.4 and the population variance (denoted σ^2 with a population size of N and a mean of μ) is understood to be given by Equation eq:variancepop. variance

$$s^2_\mathbf{x} = \frac{\sum_{i=1}^{n} (x_i - \bar{\mathbf{x}})^2}{n-1} = \frac{SQ_\mathbf{x}}{n-1} \tag{3.4}$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \tag{3.5}$$

In order to arrive at the *average deviation of the data points from the mean* instead of the *average squared deviation of the data points from the mean*, we need to reverse the upscaling effect of squaring the distances when calculating SQ. We can do this by taking the square root of the variance s^2. The result is called the *standard deviation*, denoted $s_\mathbf{x}$, or simply s if it's clear that there is only standard deviation

[13] This is the abbreviation of German *Summe der Quadrate*, but you can also read it as *Sum of sQuares*. We feel more comfortable with this abbreviation than with *SS*, which is often used in English.

one sample under discussion:

$$s_{\mathbf{x}} = \sqrt{s^2} \tag{3.6}$$

It is customary to switch between variance and standard deviation as the information conveyed by them is equivalent. Most people will probably call the overall phenomenon *variance*, but *standard deviation* is the measure more frequently used in calculations. Let's do it for the small sample $\mathbf{x}$, based on the distances to the mean from Table 3.7. First, we calculate $SQ_{\mathbf{x}}$:

$$\begin{aligned} SQ_{\mathbf{x}} = (-104)^2 + (-94)^2 + (-55)^2 + (-33)^2 + (-31)^2 + (31)^2 + (58)^2 + (58)^2 + (70)^2 + (101)^2 = \\ 10816 + 8836 + 3025 + 1089 + 961 + 961 + 3364 + 3364 + 4900 + 10201 = 47517 \end{aligned}$$

Then, we need to divide SQ by $n - 1$ to calculate the variance s^2:

$$s^2_{\mathbf{x}} = \frac{47517}{9} = 5279.7$$

And finally, we take the square root to calculate the standard deviation:

$$s_{\mathbf{x}} = \sqrt{5279.7} = 72.7$$

The same could be done for the large $\mathbf{x}$ from introduced with Table 3.5, but manually calculating the variance for any sample larger than 10 data points is really tedious (as you'll experience when working on the exercises). Hence, we did it using software, and Table 3.8 summarises the larger $\mathbf{x}$ and $\mathbf{y}$ for you, including variance and standard deviation. Finally, Table 3.9 provides an overview of the statistics that we've introduced in this chapter and to which types of measurements they apply. In the next chapter, we build upon these statistics and discuss the relation between populations and samples in more detail. The ultimate goal is a flexible and general approach to statistical inference based on samples in later chapters.

Table 3.8: Summary of two numeric samples with n=100 including variance and standard deviation

	x	y
Minimum	149	95
First Quartile	227	201
Median	250	244
Mean	250	252
Third Quartile	273	304
Maximum	348	398
Variance	1306.6	5345.9
Standard Deviation	36.1	73.1
Sample size	100	100

Table 3.9: Adequate statistics and plots for summarising types of variables

		Discrete			Continuous
	Symbol	Binary	Nominal	Ordinal	Numeric
Sample Size	n	number of data points			
Median	$\tilde{\mathrm{x}}$	middle value of sorted sample			
Mode	$\dot{\mathrm{x}}$	most frequent value			density maximum
Mean	$\bar{\mathrm{x}}$	—			Equation 3.2
Variance	s^2	—			Equation 3.4
Standard Deviation	s	—			Equation 3.6
Plot		bar plot			histogram, density plot

Exercises for Chapter 3

(1) One Two Three

4 Estimation: Means and Proportions

Overview

In this chapter, we introduce the error intervals (usually called confidence intervals, which is slightly misleading). They're related to frequentist tests which we'll introduce in later chapters, but they serve a different purpose than tests. We almost always measure means of numeric variables or proportions of nominal variables in samples. However, we're almost always interested in the population or in properties of the process that generates the data. Therefore, error intervals are a way of controlling the *safety* and the *precision* of estimating the population parameter from the sample value.

By using error intervals we acknowledge the fact that the sample mean or proportion is virtually never identical to the true parameter. Therefore, we construct an interval around the value calculated from the sample. By adjusting the width of the interval (using special maths) we can adjust the safety and the precision of the estimation. A wider interval has a higher frequentist probability of actually containing the true value (high safety), but it's also less precise because it contains more possible values. The only way to increase the precision while also increasing the safety of the estimation is to increase the sample size. In all of this, it is vital to keep in mind what a frequentist probability is. With error intervals, its primary (to some, its only) purpose is to control the long-run error rate of the estimation.

We begin by showing how the accuracy of sample estimates increases with the sample size and decreases with the variance in the data. The related statistic is the standard error, and we introduce the standard error for sample means and sample proportions. The maths of error intervals is then shown to be based on standard errors. We close with a short section on the density function of the Normal (or Gaussian) Distribution.

Problem Statement: How Reliable Is Your Sample?

Have you ever tried Haribo Berries? They come in packs containing two delicious flavours: raspberry and blackberry. The two flavours are clearly distinct, and you should really prefer blackberry flavour. However, many people feel that there are on average less blackberries than raspberries per pack. Before complaining to Haribo, you decide to take a sample and buy a pack containing 100g. You count them and find that there are 18 raspberries and 12 blackberries. Hm. Is this a precise estimate of the distribution of berries in the average pack? You consider buying a 3kg pack and counting the 900 berries in it. But how much better would the resulting estimate be compared to the 100g bag? Would it be 30 times better? What does this even mean?

Later that day (feeling a tad queasy because you've destroyed the evidence), you go back to "the lab" to do a self-paced reading experiment in order to find out the average per-word reading speed of adult Japanese speakers in a pre-study. It's an exploratory study, and you don't have a hypothesis for the average. Of the 30 invited participants, 4 show up, each reading 10 words. Adjusted for word-length, it took them a mean 100ms per word. A colleague from the theory department pops over for a coffee, looks at the results, and recommends increasing your sample size to get a better estimate. She says you should have at least 100 participants read 1000 words each. How much better would your estimate be? Would it be 2500 times better? What does this even mean?

4.1 Sampling From a Population

4.1.1 Data Generating Processes

Before we discuss sampling accuracy, we need to clarify what we mean by a population. In Section 3.2.1 we colloquially compared three populations: (i) all galaxies in the Laniakea Supercluster, (ii) all adult Tories in Buckinghamshire, and (iii) all sentences of contemporary German. Clearly, these are very different in type and count. There are (according to Wikipedia) roughly 100000 galaxies in Laniakea. The population in Buckinghamshire is roughly 850000, and it's a very conservative county. Let's estimate 70% of the population are Tory support-

ers.[1] The population of Tory supporters in Buckinghamshire probably consists of some 400000 to 450000 humans (having subtracted 20% of the total population to account for minors and apolitical people). How many *sentences of contemporary German* are there? That's a much more difficult question. What counts as German? What counts as a sentence? Do sentences spoken and written by L2 learners count? How advanced do these L2 learners have to be? Is CEFR level B2 good enough, or do we require C1? These questions can't be answered in a statistics text book.[2] What matters is that none of these populations is static. Over the lifespan of the universe, the number of galaxies in Laniakea went from 0 to 100000, and it will return to 0. Galaxies form or are drawn into superclusters all the time, and galaxies are torn apart and will eventually fall apart (in layman's terms) in the distant future. People in Buckinghamshire are born and die, and people change their political affiliation all the time. German sentences are produced at a breathtaking rate, and even the population of speakers of German changes every minute. If the population really mattered as a fixed construct of a well-defined (even if unknown) size, we'd have to repeat all empirical work every day, second, or even millisecond.

While it's sometimes stressed that the population needs to be conceptuall infinite or at least significantly larger than the sample (for mathematical reasons), we see this rarely ever playing an important role in empirical science. What's more, the whole idea of a fixed and huge population from which we draw a sample is not very helpful. Why does any linguist draw a sample of German sentences? In the post-structuralist era, it's not because we want to find out the properties of a massive collection of sentences but because we're interested in the mechanism that generates such sentences. It doesn't matter which model of grammar a linguist believes in: When they look at a sample of sentences they ask what is the grammar like that produces such sentences, be it a formal or a cognitive type of grammar. Hence, we consider it much more appropriate and intuitive to speak of the *data-generating process* (DGP) rather than the population, although we

data-generating process

[1] According to the Buckinghamshire Council website, there are 105 conservative councillors, 15 are independent, 15 are liberal democrats, 6 belong to Labour, and 5 have other affiliations as of 25 February 2025. That's approximately 70% Tories. However, given the British voting system, the council is a bad estimator of the composition of the electorate in terms of political affiliation. (https://buckinghamshire.moderngov.co.uk/mgMemberIndex.aspx)

[2] They should have been answered by linguists (at least those who rely on empirical data) before they started to do research on German per se (as opposed to research on the linguistic behaviour of samples of German-speaking subjects). If you're a corpus linguist who believes that your corpus of choice *represents* a population, you should have a very good answer to these questions. In this respect, the most important property of works such as Biber (1993) is that they're over 30 years old.

conveniently drop back to talking about a *population* because it sometimes just sounds better in a sentence. The data-generating process can be conceived of as cognitive, social, or even purely formal. In fact, our simulations (see Section 2.5) simulate exactly such processes. Since any scientifically interesting process can, in principle, generate ever new data points, the population of those data points is conceptually infinite.

In this chapter, we ask a very crucial question. How reliable can we expect a sample to be in representing the DGP (or, in old-school terms, the population)? Imagine English speakers were programmed (through cognitive constraints) to produce only 2% passives of *sleep*. In this case, the true parameter with which the DGP was set up is 2% (a proportion of 0.02). How well can you expect to approximate this percentage/proportion with a sample of size $n = 1$, $n = 10$, $n = 100$, and so on? Do you see how this question is related to the logic of testing introduced in Chapter 2? However, this chapter is not about *testing* but rather about the *estimation* a parameter of the population from a sample. In order to get there, we need to talk about the distribution of results in repeated experiments in relation to true values of the DGP. We'll use the results from this chapter to develop a general framework for inferential testing in Chapter 5.

estimation

4.1.2 Sampling Berries: One, Two, Three, Many

We begin with a simulation of a known situation. As explained in Section 2.5, we can never do this in actual empirical work. We do empirical research to find out what reality is like, precisely because we don't know what it's like. However, simulating a known (even fictitious) reality allows us to explore and illustrate what happens in actual empirical work. Hence, we now simulate a berry-generating Haribo process that produces blackberries at a rate of 0.3 (or 30%). In the long run, bags of Haribo Berries produced by this process will contain a proportion of $q_{\text{blackberry}} = 0.3$ on average. That does not mean, however, that *each* bag will contain *exactly* 30% blackberries. Also, we expect larger samples to better approximate the true proportion of blackberries. We've been re-iterating this point and variations of it throughout the previous chapters.

The simulated process allows us to pretend that we have a machine that randomly fills bags with berries from a production line that produces 30% blackberries and 70% raspberries. We call each simulated filling of such bag a *simulation run* or a *replication*. In Figure 4.1, we plot the results for 100 replications at a sample size of $n = 1$. Each dot represents one bag containing a single berry filled by a machine that produces 30% blackberries by design. We plot the proportion of blackberries in the sample. As a sample of 1 berry contains either a blackberry

replication

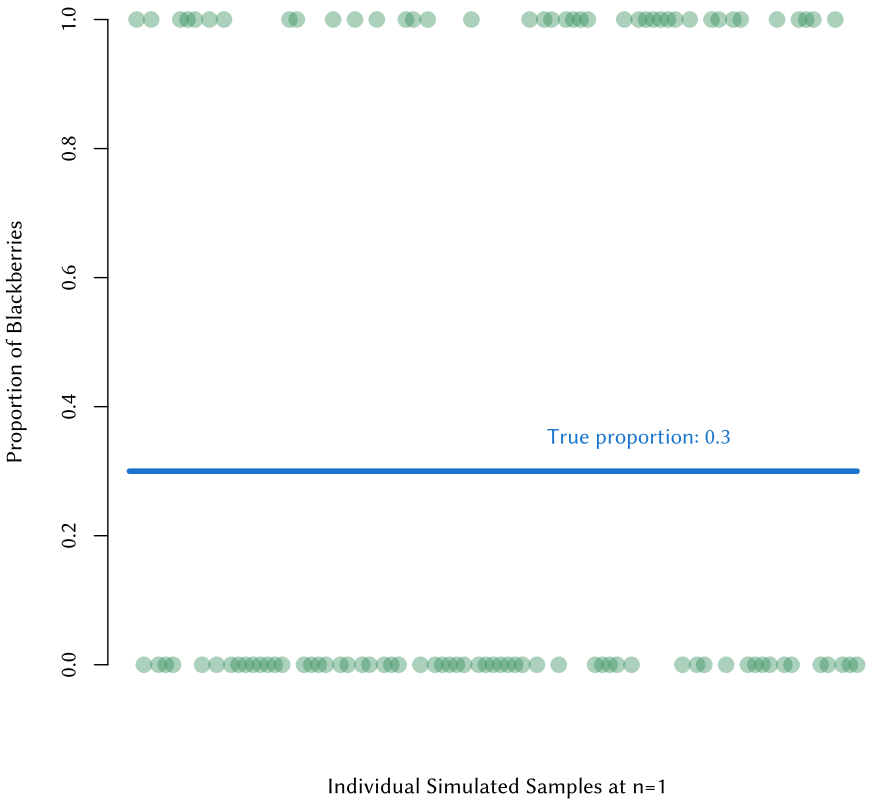

Figure 4.1: Proportions measured in 100 replications at n=1 with a true proportion of 0.3

or a raspberry, and there isn't an in-between option. Hence, the proportion of blackberries can only be either 0 or 1 at $n = 1$.

We got 60 samples of size $n = 1$ containing no blackberry (proportion $q = 0$) and 40 containing one blackberry (proportion $q = 1$). The individual samples do the best they can to approximate the true proportion of 0.3, but at $n = 1$, the options are severely limited. In this extreme case, however, the *proportion of samples* where there was a blackberry (0.4) approximates the true value reasonably well.

We now increase the sample size step by step. At each increment, we simu-

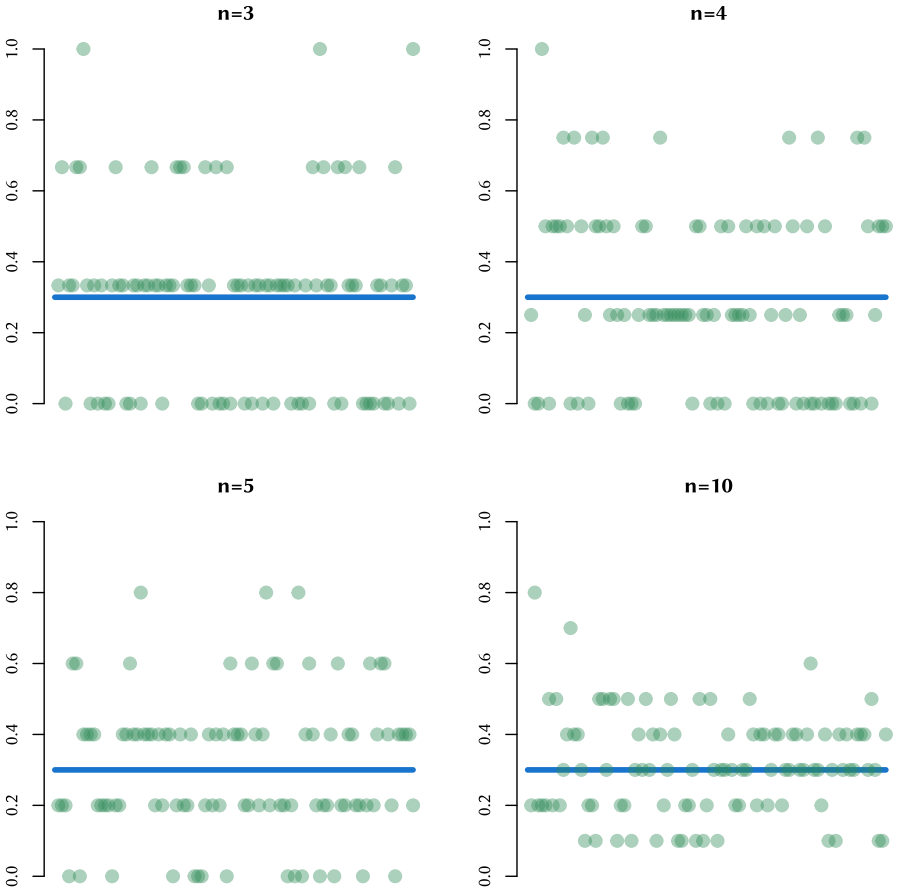

Figure 4.2: Proportions measured in 100 replications at n=3, n=4, n=5, and n=10 with a true proportion of 0.3

late 100 replications and plot the proportion of blackberries per replication in Figure 4.2.

We don't need to analyse these results numerically. It should be immediately obvious that the samples approximate the true value more reliably. With very small samples (up to $n = 5$ in this example), results are still very limited in their possible outcomes, and it's impossible to hit the true proportion of 0.3 precisely. But most of the samples drift towards the best approximation possible. (Keep in mind that each dot represents one sample of the respective sample size.) At $n = 10$, the sample proportions start to form a clearly distinguishable cloud around

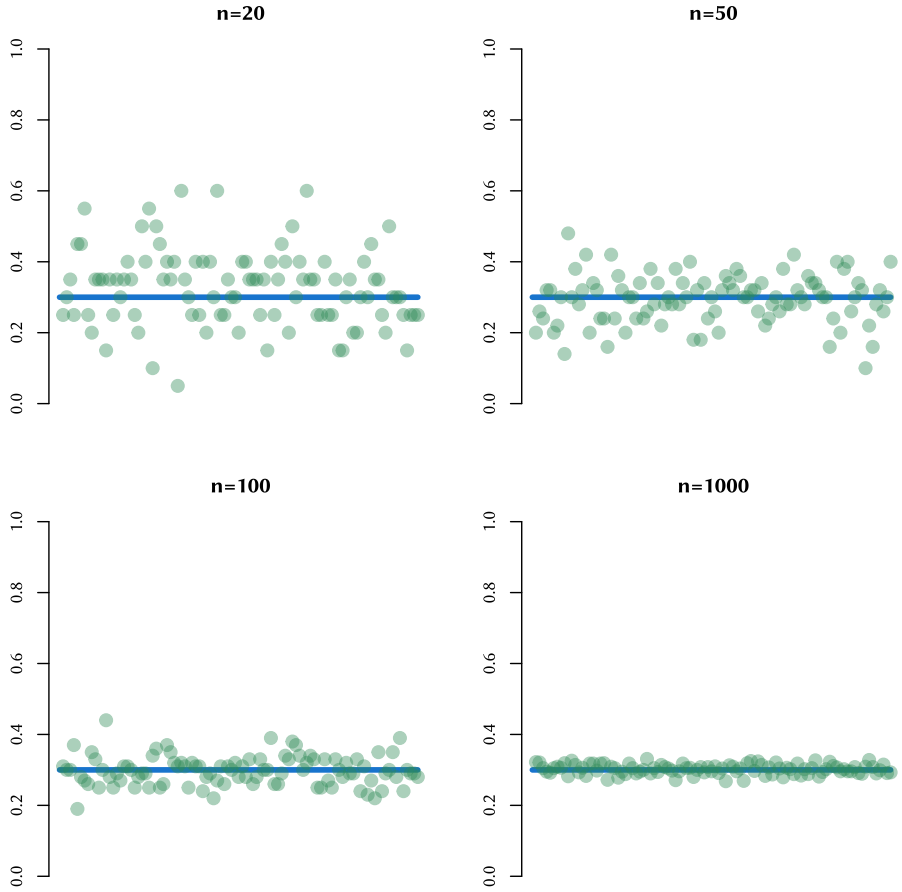

Figure 4.3: Proportions measured in 100 replications at n=20, n=50, n=100, and n=1000 with a true proportion of 0.3

the true value. Let's see what happens with samples of more substantial sizes in Figure 4.3.

While it's not at all impossible to sample 0 or 1000 blackberries if the true proportion of blackberries is 0.3, such a result becomes extremely rare event at this sample size. We see in the lower-right panel of Figure 4.3 that even results lower than 0.2 or higher than 0.4 are so rare that none occurred in the 100 replications. For a binary (or nominal or ordinal) variable, larger samples approximate the true proportion more reliably. In the next section, we try the same with a numeric variable.

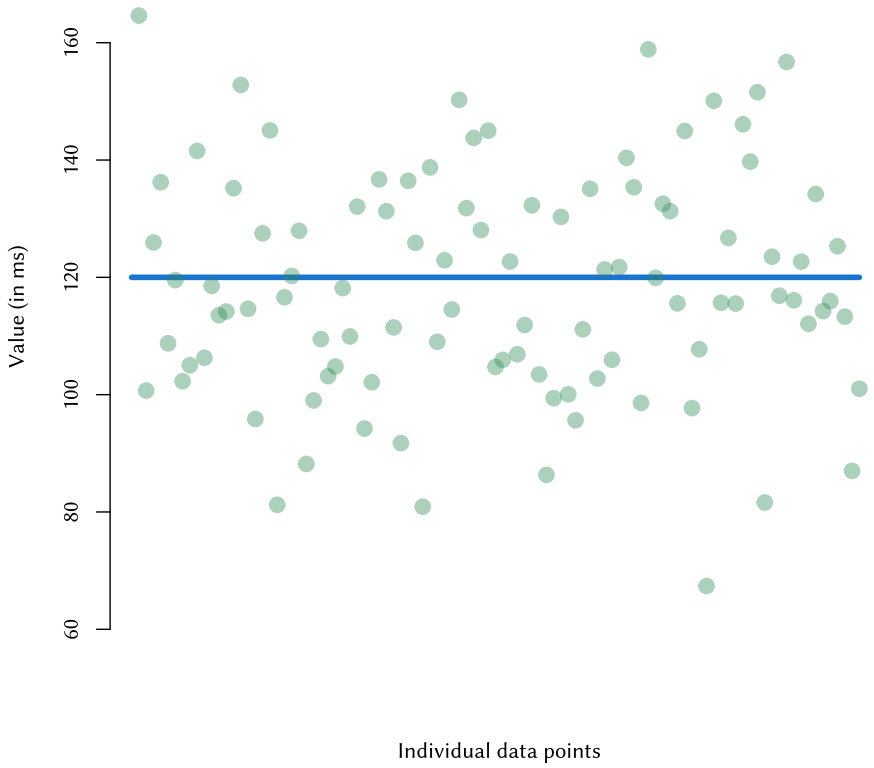

Figure 4.4: Raw data plot of a single random sample with n=100 from a DGP that has μ=120 and σ=20

4.1.3 Sampling Milliseconds: One, Two, Three, Many

We now turn to the per-word reading speed from the Problem Statement. Before we run the simulation, we have to set the parameters that define the simulated reality. Let's assume that the real word-length-adjusted reading speed is at $\mu = 120$ (in milliseconds) per word, and the standard deviation is $\sigma = 20$. With the simulations, we can now ask what will the outcomes be if we take many samples of size $n = 1$, $n = 2$, etc. First take a look at Figure 4.4.

It shows a single sample with $n = 100$ generated by a DGP that is set to $\mu = 120$ and $\sigma = 20$. The dots are single data points (i. e., one person reading one word).

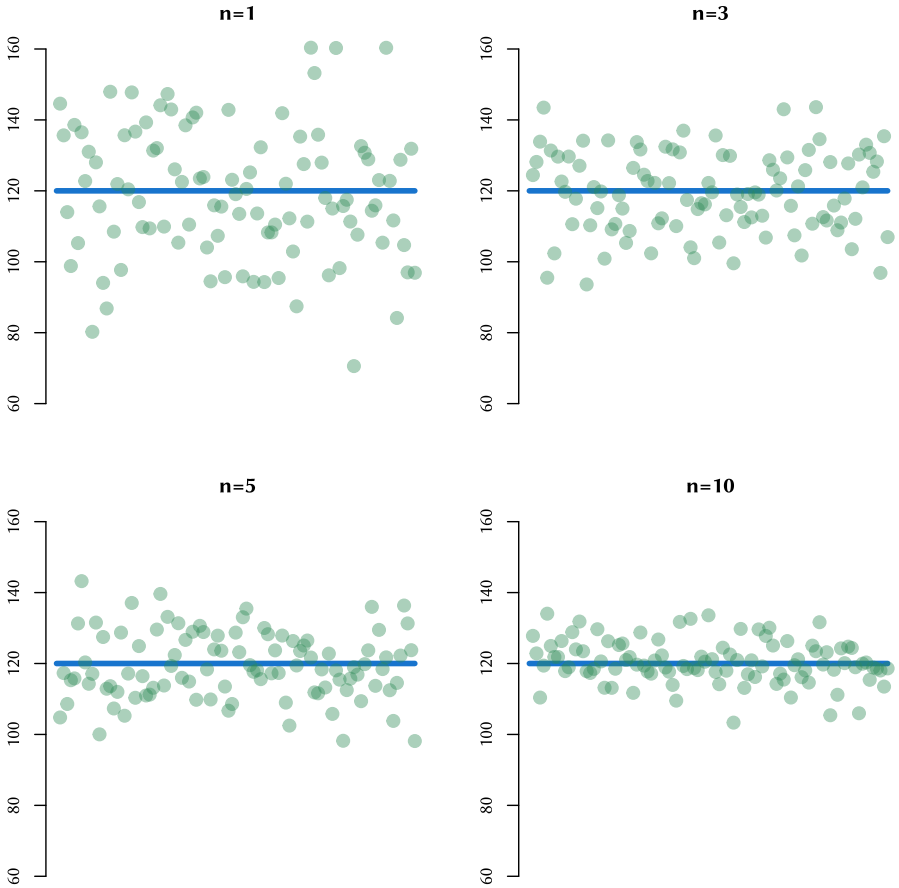

Figure 4.5: Means measured in 100 replications at n=1, n=3, n=5, and n=10 with a true mean of σ=120 and a true standard deviation of μ=20

Values close to $\mu = 120$ have the highest probability (and hence the highest frequency in repeated sampling). The average deviation of the values is at $\mu - \sigma = 100$ and at $\mu + \sigma = 140$. Since this is just the average deviation, and we do not expect there to be a higher number of measurements around these two points, of course. They're just the central tendency of the negative and positive deviations, respectively. Figure 4.5 shows what happens in 100 simulated samples at sizes of $n = 1$, $n = 3$, $n = 5$, and $n = 10$.

We see a very similar effect as in the case of the binary variable. At a sample size of $n = 1$, the mean is virtually identical to a single sampled value. Therefore,

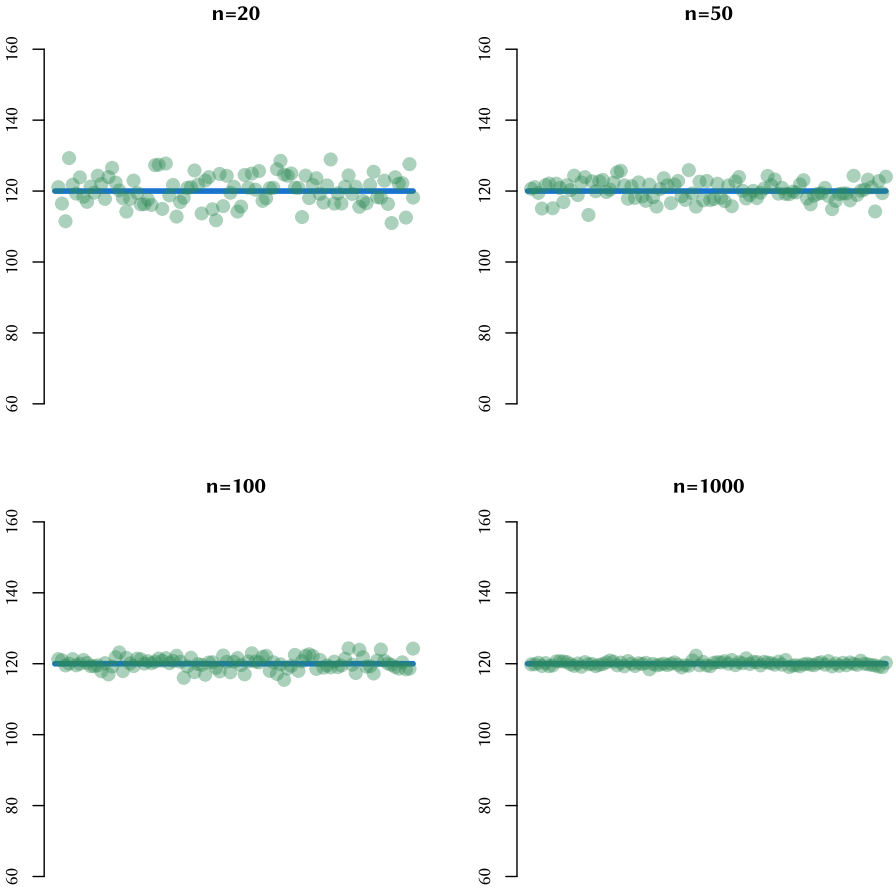

Figure 4.6: Means measured in 100 replications at n=20, n=50, n=100, and n=1000 with a true mean of σ=120 and a true standard deviation of μ=20

the distribution of sample means in the upper left panel of Figure 4.5 looks like a distribution of individual data points in Figure 4.4.[3] The larger the sample, the closer the sample means move (on average) towards the real value as the variance between single data points plays an ever smaller role. This trend continues with much larger samples, see Figure 4.6.

[3]Keep in mind: The dots in Figure 4.4 represent individual data points. The dots in Figure 4.5 represent individual sample means calculated from samples (of different sizes per panel).

Why did we begin with samples of size $n = 1$? After all, it's a ridiculous sample size. We did it because it allowed us to illustrate how a sample approximates parameters like the mean increasingly better than a single measurement. A single value (equivalently a sample with $n = 1$) approximates a parameter of the DGP with a certain reliability, and the reliability is higher the lower the variance in the DGP is. The variance determines how far single data points bounce around the actual parameter, so with larger variance comes higher uncertainty. However, the more such single data points you have in your sample, the more the bouncing around averages out. It's still possible to draw a very extreme sample, but the probability of drawing such a sample gets lower and lower with larger samples.

We hope that our illustration was intuitive enough and laid the foundations for the statistics introduced in Section 4.2 and 4.3. The way we introduced the idea of sampling variation—we hope—has made it clear that it's a frequentist concept. If we took a lot of samples (100 in the examples above), then we'd see the patterns that we saw in the plots. However, nothing can be known for certain or with a quantifiable reliability about the DGB from a single sample. Please keep this in mind. The Statistics Wars were fought about misunderstandings of this and other simple facts (see Chapter 7).

Big Point: Sample Size in Parameter Estimation

When we use a sample to estimate a population parameter, we can rely on some simple facts about variance and sample size. A larger variance in the data-generating process (DGP) makes estimates of their mean less accurate. On the other hand, larger samples average over the variance that makes the individual data points bounce around the true parameter. Hence, parameter estimates from larger samples have a better chance of estimating the true parameter more accurately. However, no amount of quantifiable certainty about the true parameter can ever be gained from a sample!

4.2 Error Intervals for Normal Numeric Measurements

4.2.1 The Variance of Sample Means

Now we know that each increment in sample size counteracts the variance that's inherent in the data points. In other words, repeated (not individual!) samples have the same or less variance than the raw data points. They only have the same variance (under repeated sampling) if $n = 1$. Even with $n = 2$, the variance already decreases (on average). As it happens, we can express this mathematically and at the same time very intuitively for means like so:

$$\text{variance of sample means at sample size } n = \frac{\sigma^2}{n}$$

The variance of the sample means is the variance of the individual measurements σ^2 divided by the sample size n. Hence, when the variance among the individual measurements goes up, the variance among the sample means goes up. On the other hand, when the sample size goes up, the variance among the sample means goes down.[4] In order to get the *average deviation of sample means from the true mean at sample size n*, we need to take the square root. In Section 3.2.3, we did the same to calculate the standard deviation from the variance. The result is called the *standard error* of the mean SE_μ. It's simply the standard deviation of sample means at a given sample size n and given the variance among the individual data points σ^2. The formula in Equation 4.1 is often expressed by the equivalent Equation 4.2, but we consider the first variant more transparent and easier to memorise.

standard error

$$SE_\mu = \sqrt{\frac{\sigma^2}{n}} \tag{4.1}$$

$$SE_\mu = \frac{\sigma}{\sqrt{n}} \tag{4.2}$$

For example, the fictitious distribution of reading times with $\mu = 120$ and $\sigma = 20$ has the following standard error SE_μ for samples of size $n = 16$:

$$SE_\mu = \frac{20}{\sqrt{16}} = \frac{20}{4} = 5$$

We have omitted one important detail. The above is only perfectly true for data that follow the *Normal Distribution* or *Gaussian Distribution*. A normally

Normal Distribution

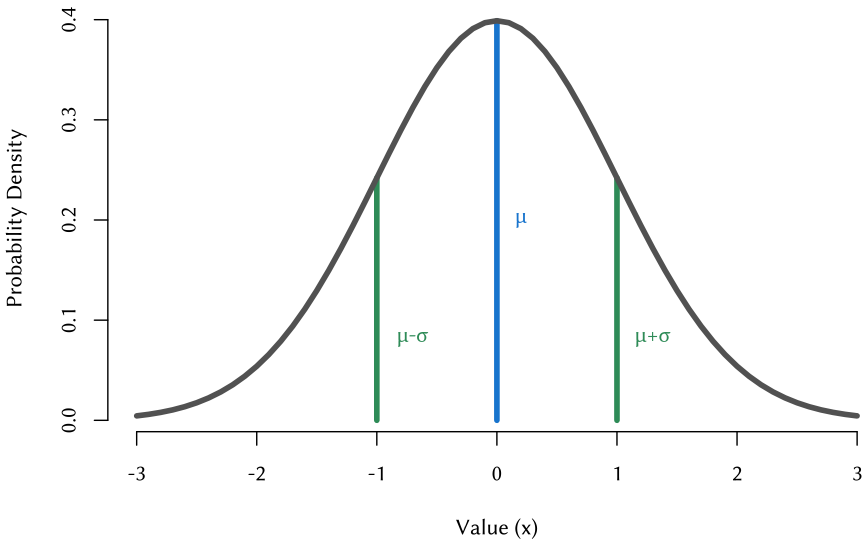

Figure 4.7: Theoretical population distribution for a normal distribution with μ=0 and σ=1

Gaussian Distribution

distributed random variable has a characteristic probability density function. It has two parameters: the mean μ and the standard deviation σ. Figure 4.7 shows a plot of the normal probability density with $\mu = 0$ and $\sigma = 1$.

You've probably seen the Normal Distribution before with the characteristic bell-shaped curve of its density function.[5] Not only do many numeric measurements in nature follow this distribution, it also pops up in an almost creepy way in a family of fundamental proofs in statistics, the so-called *Central Limit Theorem* (which has a lot to do with what we're in the middle of introducing at the moment). Be that as it may, for samples from normally distributed data, the simple formula for the standard error of the mean given above is provably correct. Notice that the density curve is fully defined by the mean and the standard deviation (see also the in-depth Section 4.4). This means that we can calculate the

[4]Notice that, from this angle, we can answer *yes* to the questions from the Problem Statement asking something like: *Would a sample 10 times larger be 10 times more accurate?*

[5]See Section 4.4 for a slightly more in-depth look at the Normal Distribution.

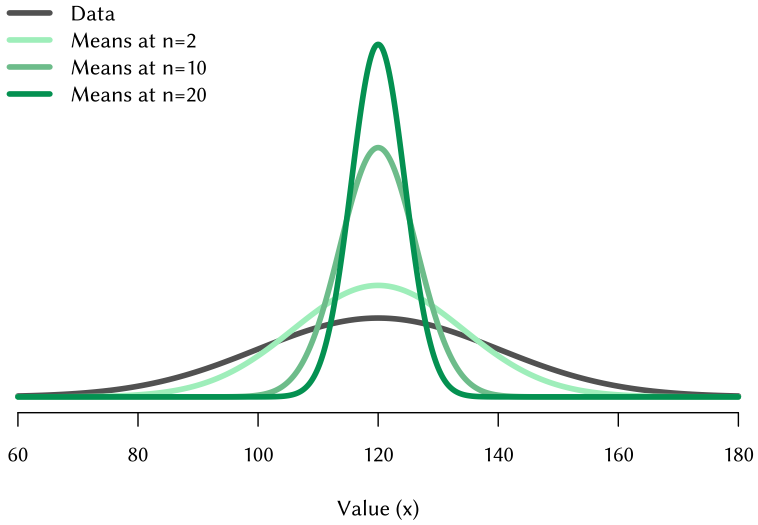

Figure 4.8: Density of a normal distribution with μ=120 and σ=20 (black); long-run sample means at different sample sizes from that normal distribution

probability for each value of x or each interval of x. For example, we can calculate the probability of $x \geq 1$, written: $Pr(x \geq 1)$. As the area under any probability density curve is 1, $Pr(x \geq 1)$ is the integral of the normal density from 1 to infinity. Piece of cake. Let's have that piece of cake now (step by step) and call it the Error Interval.

First, Figure 4.8 shows the theoretical normal density plot of the reading times in milliseconds as introduced in Section 4.1.3.[6] We've added three density functions for the distributions of means at sample sizes 2, 10, and 20. With larger samples, the probability mass piles up around the mean very quickly, and the normal curve narrows sharply. This graph illustrates again what we've shown before: The sample means are centred around the true mean, and they become better and better estimators of the mean with increasing sample size.

Let's focus on one of those curves, the density function of the means at $n = 10$. It's shown in Figure 4.9. In addition, the centre area under the curve that

[6]In density plots, we omit y axis labels whenever possible as they are rarely interpreted directly.

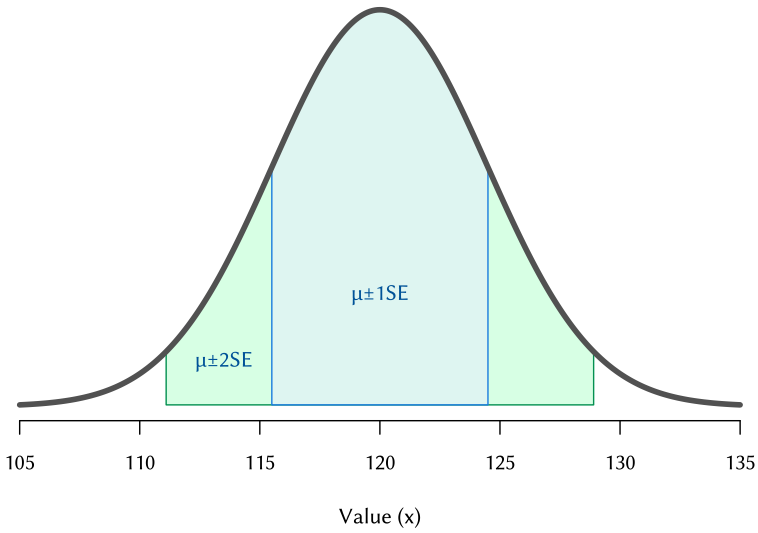

Figure 4.9: Density of the distribution of the means at n=20 with μ=120 and σ=20; are at μ±σ is highlighted

spans from the mean $1SE$ to the left and to the right ($\mu \pm 1SE$) is shaded in blue. The are covering $\mu \pm 2SE$ is shaded in green. As the Normal Distribution has known mathematical properties and we know the real mean and the real standard deviation, we can calculate the size of the shaded areas. Well, *we* here stands for *fully qualified statisticians*, and for the purpose of this book, we just tell you some useful values (see also Tables below). The blue area is approximately 0.68, i. e., 68% of the whole area under the curve. The green area (plus the blue area) is approximately 0.96, i. e., 96%.

How is this useful? Think about what it means in terms of probability. It means that (i) if the true mean of a variable is μ, (ii) its true standard deviation is σ, and (iii) we draw a sample with n, then the probability that the mean of any sample lies in the interval $\mu \pm 1SE$ is 0.68. Similarly, the probability that the mean of any sample lies in the interval $\mu \pm 2SE$ is 0.96. The frequentist interpretation consistently equates the probability of an event occurring with its long-run proportion among repeated repetitions. In our notation for probabilities from Section 2.2.3, for any sample mean $\bar{x}$:

$$Pr(\mu - SE_\mu \leq \bar{\mathbf{x}} \leq \mu + SE_\mu) \approx 0.68$$

$$Pr(\mu - 2SE_\mu \leq \bar{\mathbf{x}} \leq \mu + 2SE_\mu) \approx 0.96$$

The standard error is sometimes also just called σ as it is a kind of standard deviation. We find it best to distinguish clearly between the standard deviation of primary data points σ and the standard deviation of sample means, the standard error of the mean SE_μ.

4.2.2 Theoretical Error Intervals

Given that we can calculate the probability of outcomes that lie within a certain interval, we can also ask the inverse question: *How many standard errors into each direction (from the mean) define the central 90%, 95%, 99%, etc. sample means under long-run repetition?* That number of standard errors is often called the *z-value* or just z. One can use the a function (called the *quantile function* of the Normal Distribution) for calculating the z-value, but pre-calculated tables usually suffice. Table 4.1 is a very sparse example of such a table.

z-value

Table 4.1: z-Values for some Probabilities

Pr	Z
0.50	0.67
0.80	1.28
0.90	1.64
0.95	1.96
0.99	2.58

Let's go through this again. Table 4.1 tells us that 50% of samples of a normally distributed variable lie within a range around the mean that is 0.67 standard errors wide on both sides of the mean. Remember that the standard error itself depends only on the variance in the data and the sample size. Furthermore, 80% of all samples lie between $\mu - 1.28$ standard errors and $\mu + 1.28$ standard errors, and so on. This is formalised in the notion of the *error interval* for the mean EI_μ, where z_a stands for the z-value for a given probability/proportion a as illustrated in Table 4.1:[7]

error interval

$$EI_\mu = \mu \pm z_a \cdot SE_\mu \tag{4.3}$$

[7] We use a because all other variables are taken.

Continuing with the reading time example, we can take the SE_μ calculated above and the mean of $\mu = 120$ to find the error interval with $a = 0.95$:

$$SE_\mu = 120 \pm 1.96 \cdot 5 = 120 \pm 9.8 = \langle 110.2, 129.8 \rangle$$

4.2.3 Empirical Error Intervals

But wait! In normal empirical work, we don't know the true mean and the true standard deviation. Some readers might have noticed this very early on and gotten frustrated because it appeared as if error intervals were a purely theoretical thing without practical use. This is not the case, and the argument will finally bring us back to the frequentist logic of inference. What we do in real life is this: We calculate an error interval based on the statistics of our concrete empirical sample. That is, we use $\bar{\mathbf{x}}$ instead of μ and $s_\mathbf{x}$ instead of σ. After all, μ and σ are unknown. The calculations are the same, only the symbols change:

$$SE_{\bar{\mathbf{x}}} = \sqrt{\frac{s^2_\mathbf{x}}{n}}$$

$$EI_{\bar{\mathbf{x}}} = \bar{\mathbf{x}} \pm z_a \cdot SE_{\bar{\mathbf{x}}}$$

Then, as the inverse of the theoretical error interval, the empirical error interval will contain the true value in a of all cases (with z_a chosen appropriately for a). Hence, we don't already have to know the truth but can still use error intervals. To make it absolutely clear what an error interval is, we have simulated it for you. We simulated 1000 samples with sample size $n = 20$ from the distribution of reading times introduced above. Then, we used the samples to calculate the empirical error intervals exclusively based on each sample. Think about it: We used the simulations of a known fictitious population to generate samples, then we took the samples and calculated the error interval *pretending we didn't know the population parameters* in order to check whether the error interval works as we think it should.[8] It's not as good as a mathematical proof, but it really helps in understanding what statistics is about. Figure 4.10 shows the true mean and all those empirical error intervals. The ones that don't include the true mean are coloured red.

The number of empirical error intervals in these 1000 replications that did include the true mean was 946, i. e., 94.6%. This is called the *coverage*, and it's

coverage

[8] Yes, in a way we're playing god. But as opposed to real life, playing god is a very good idea in statistics (if you have a good random number generator).

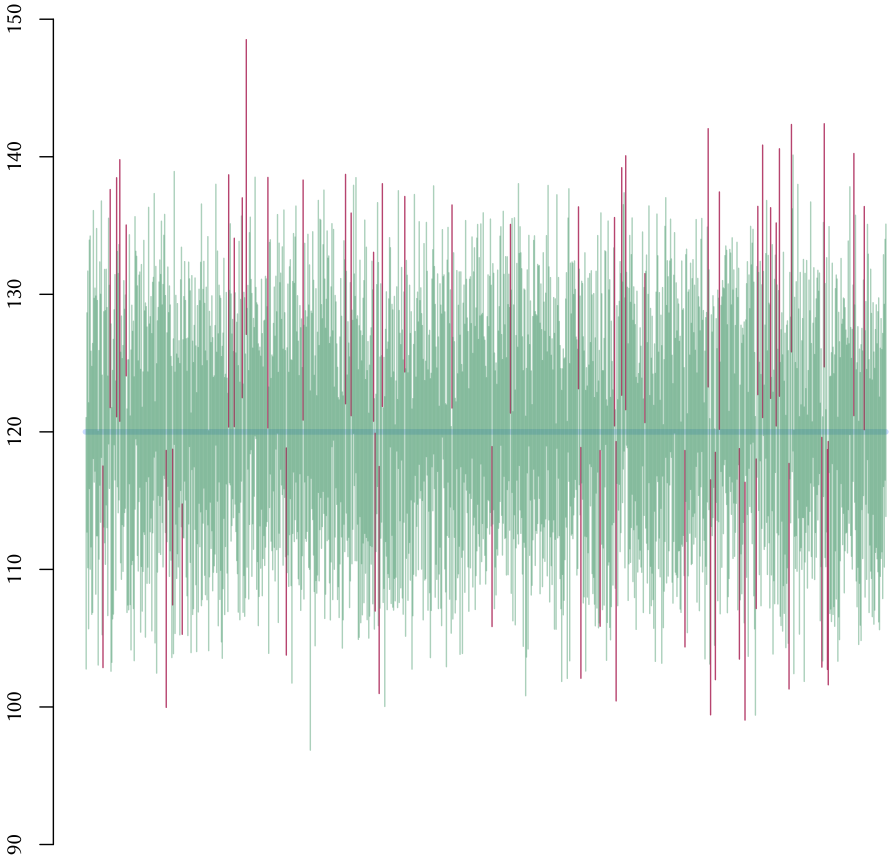

Figure 4.10: Empirical error intervals from 1000 simulated samples; ones that don't contain the true mean marked in red; μ=120, σ=20, z=1.96, n=20

very close to the expected—or rather *requested*—value of 95%. We say *requested* because in deciding to calculate the intervals for a proportion of $a = 0.95$ and consequently choosing $z = 1.96$, we required that in the long run 95% of all empirical error intervals should contain the true mean.[9] We've now called it a

[9] Actually, by calculating error intervals with $a = 0.95$ for any parameter, our long-run error is 5%. It doesn't matter whether we keep taking samples from the same population over and over again or whether we first sample this, then that, and then the other. It's of no big practical consequence, but it's a good thing to keep in mind in order not to misinterpret error intervals.

proportion, but think about it again in terms of *probability*. It's the frequentist probability of drawing a sample whose empirical error interval contains the true mean before we actually draw the sample. Once it has been drawn, the empirical error interval either contains the true mean or it doesn't, but that's a fact and has no probability attached to it.

The term *error interval* is our invention. It's usually called the *confidence interval*. However, there are so many misinterpretations attached to this term that we decided to avoid it. An error interval with $a = 0.95$ is called the *95% confidence interval* (and similar for other values of a). A very wrong way of interpreting an empirical error interval then goes like this:

confidence interval

> STOP *The 95% confidence interval for the mean contains the true mean with 95% certainty. (Meaning: We can be 95% sure that it contains the true mean.)*

What's this even supposed to mean? Next time you hear someone say something like this about a frequentist error interval, ask them what it means to be *95% certain of something*. What it really means is that before taking the sample, there was a probability of 0.95 to draw a sample that contains the true mean (given the chosen n and $z = 1.96$). Now, all we know is that the interval either contains the true mean or a rare event occurred. Please go back to Section 2.3.1 to convince yourself that an event with a probability of 0.05 is not that rare after all.

4.2.4 Varying z and n

We use error intervals to get an estimate of the true mean with a known probability of being wrong. We call this the *safety of the estimate*: The lower the error probability, the safer the estimate. If we set $z = 1.96$, we will estimate 95% of all intervals correctly in the long run. Setting z higher (for example, $z = 2.58$) gives us even better long-run chances of getting our estimates right. So why don't we always set $z = 5$ or something really high? Look at Figure 4.10. If we increased z, the samples themselves wouldn't change. However, we'd require that more intervals should include the true mean, and the only way to do that is to make the intervals wider. This leads to a higher rate of correct estimates, but the *precision of the estimate* decreases. We call precision the narrowness of the error interval: The narrower the interval, the higher the precision. But be aware that the precision is a value set by us. Once again, it does not change the sample itself, and hence it does not change the fact that the sample represents the population (parameter) either well or not.

safety of the estimate

precision of the estimate

First, let's calculate an empirical error interval with $z = 1.96$ and $z = 2.58$ before looking at some more illustrative plots. A single example replications of the simulation led to the following statistics: $\bar{\mathbf{x}} = 119.63$ and $s^2_{\mathbf{x}} = 442.47$. The SE_μ is calculated as follows:

$$SE_{\bar{\mathbf{x}}} = \sqrt{\frac{442.47}{20}} = 4.7$$

The 95% error interval (i. e., $z = 1.96$) and the 99% error interval (i. e., $z = 2.58$) are:

$$EI_{\bar{\mathbf{x}},0.95} = 119.63 \pm 1.96 \cdot 4.7 = \langle 110.42, 128.84 \rangle$$
$$EI_{\bar{\mathbf{x}},0.99} = 119.63 \pm 2.58 \cdot 4.7 = \langle 107.5, 131.76 \rangle$$

The interval is roughly 6ms larger for the added 4% success rate. For a more comprehensive impression of the effect of manipulating z and the effect of the sample size n, we provide Figure 4.11.

As expected, a larger n makes the intervals smaller (increased precision) because it makes the standard error smaller (see Section 4.1.3). A larger z, on the other hand, make the intervals larger because we require that the probability of the interval including the true mean be higher (increased safety). By the way, the fact that the coverage does not always land exactly at 95% or 99% is not surprising. A perfect match between the theoretical expectation and the empirical reality is guaranteed only in the limit, i. e., if we take infinitely many samples. In the remainder of the chapter, we will introduce error intervals for proportions before returning briefly to the Normal Distribution in the in-depth Section 4.4).

Big Point: Interpretation of Error Intervals

An error interval for the mean is an attempt to estimate the true mean. It provides an interval which covers the true mean with a specified frequentist pre-sample probability. For example, a 95% error interval ($a = 0.95$) means that before the sample is drawn, it has a probability of 0.95 of actually including the true mean. Once it is drawn, it either includes the true mean, or a (relatively) rare event has occurred. In the long run, it is guaranteed that a proportion of a intervals thus calculated ($a \cdot 100\%$) include the true mean. An error interval does not quantify any justifiable degree

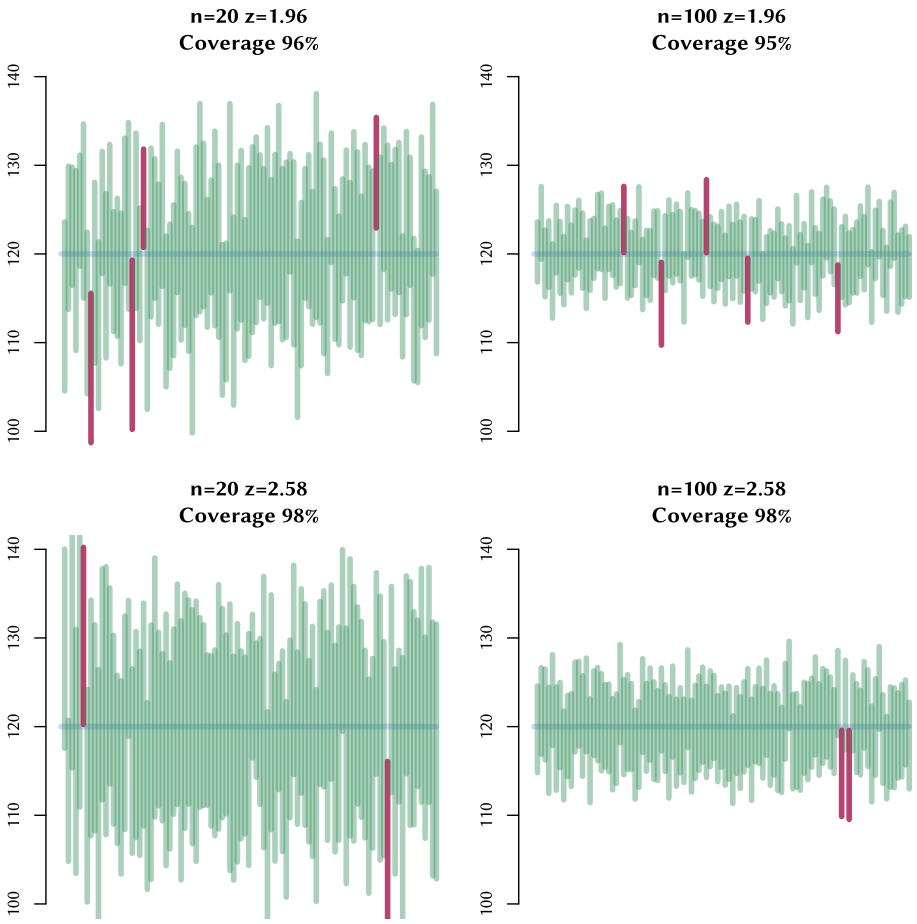

Figure 4.11: Simulations of empirical error intervals for different z and n with 100 replications each

of *confidence* we should have in the estimate, which is why we don't call it *confidence interval*. The term *error interval* reflects the fact that we merely control the error rate of the estimation.

If safe estimates are of the essence, one should increase q (for example, $q = 0.99$ for the 99% error interval). This lowers the error rate in the long run and makes the estimation *safer*. A safer estimates increase the size of the intervals and makes it less *precise* (larger intervals). Changing the

sample size does not change the safety of the estimate. Hence, if *both a precise and a safe* estimation is required, one should increase q and also increase the sample size n. Finally, error intervals (also called confidence intervals) have nothing to do with specifying the degree of certainty that the true mean falls within the interval.

4.3 Error Intervals for Binary Measurements

The idea behind error intervals for proportions is essentially the same as for means. Look back at Figures 4.1, 4.2, and 4.3. The distributions of the proportions from samples with very low sample size show obvious banding and look irregular. However, at least starting with $n = 20$, the plots look very similar to the plots of means in Figure 4.5 and Figure 4.6. Compare the density plots in Figure 4.12. Keep in mind: These are plots of the distribution of proportions and means calculated from a lot of (simulated) samples, not plots of raw individual measurements.

The distributions still look bumpy at $n = 20$, but that is expected. Most importantly, however, there is virtually no qualitative distinction anymore between the distributions of proportions and menas at $n = 20$, even though proportions reflect counts of a discrete variable and means derive from a continuous numerical variable. It was also shown beyond such impressionistic comparisons that the sampling distributions of proportions is reasonably well approximated by a normal distribution. Hence, we can use the normal error interval we introduced for the mean μ for proportions as well. The theoretical error interval for the population proportion ρ is:

$$EI_\rho = \rho \pm z \cdot SE_\rho \tag{4.4}$$

Again, the empirical error interval based on a single sample is the same, except for a single sample proportion q instead of the population proportion ρ:

$$EI_q = q \pm z \cdot SE_q \tag{4.5}$$

But what are SE_ρ and SE_p? The raw data points do not have a normal variance associated with them. They are binary, either and or, 0 or 1. We need a measure for the sample proportions that is not directly derived from variance in the data points.[10] We expect the standard error for proportions to be smaller for

[10] Of course, real statisticians have conducted proofs and done all the proper maths. We skip all

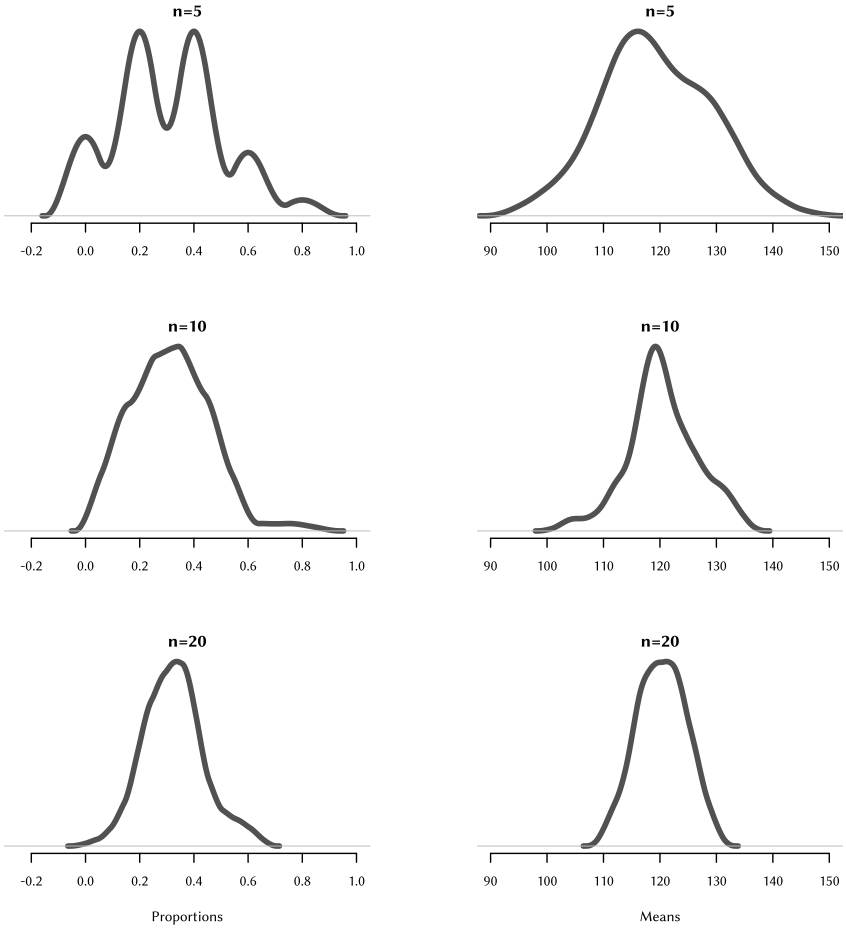

Figure 4.12: Histograms of 1000 sample proportions (left) and 1000 sample means (right) at the same sample sizes

larger samples (see Figure 4.12). Hence, we divide by n. Additionally, consider that proportions range from 0 to 1 and consider at which values between 0 and 1 the variance in proportions might be higher. Again, we use lots of simulated samples for illustration, see Figure 4.13.

We hope you noticed what's going on there. At $\rho = 0.5$, the sample proportion deviates from the true proportion into the negative and the positive direction. The farther towards 0 or 1 the true proportion is moved, the more restricted the

of this and try to make things easy to grasp intuitively.

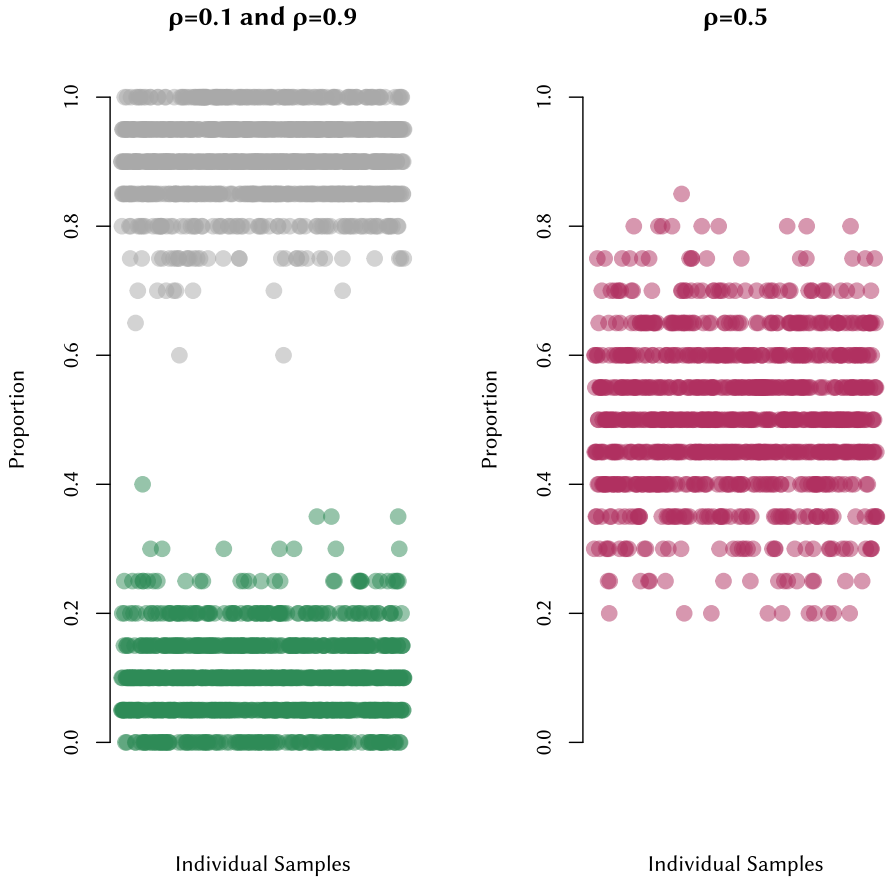

Figure 4.13: Proportions from samples with n=20 with different true population proportions ρ

possible variation is in one of the two directions. As proportions can't be negative or larger than 1, the variance in the sample proportions hits a ceiling of 1 and a floor of 0. Hence, the variance is largest at $\rho = 0.5$ and smallest at $\rho = 0$ and $\rho = 1$.

The standard error for proportions SE_ρ (theoretical) and SE_q (empirical) reflects this:

$$SE_\rho = \sqrt{\frac{\rho \cdot (1 - \rho)}{n}} \tag{4.6}$$

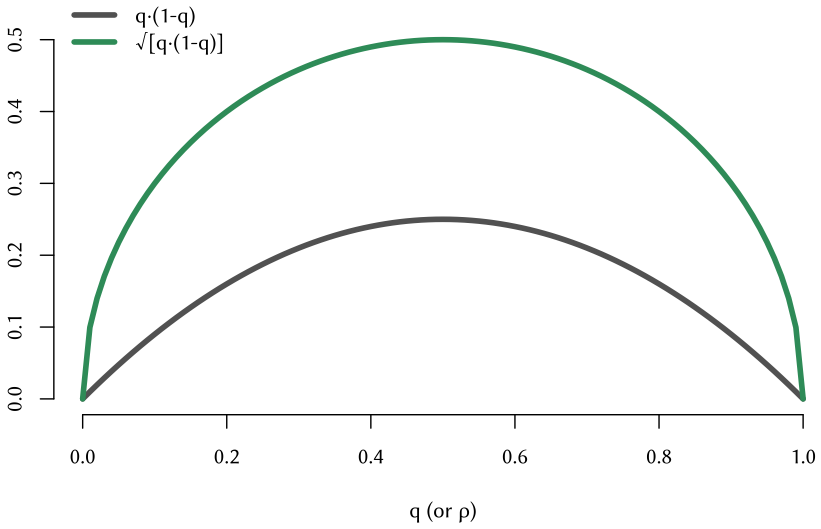

Figure 4.14: Proportions q (x-axis) plotted against q·(1-q) (y-axis) and its square root

$$SE_q = \sqrt{\frac{q \cdot (1-q)}{n}} \tag{4.7}$$

We multiply the proportion ρ by $1-\rho$ (or q by $q-1$). Do you see why that is appropriate (at least what's below the square root)? The simplest illustration is the comparison between, for example, $0.9 \cdot 0.1 = 0.09$ compared to the much larger $0.5 \cdot 0.5 = 0.25$.

Figure 4.14 illustrates it for all values of ρ or q. Multiplying ρ by $1-\rho$ (or q by $q-1$) gives us exactly what we need: a function that starts at the function value 0 for the input 0, then increases as we move towards 0.5 (with a function value of 0.25). Then, the function value returns to 0 as we move towards the input value 1. The square root conserves this property. While it's good that mathematical proofs have led to the development of the maths, we can also make sense of the formulas and understand them from the viewpoint of a mere practitioner.

Therefore, the width of the confidence intervals has the characteristics observable in Figure 4.15. It's largest for centre values and very narrow for extreme pro-

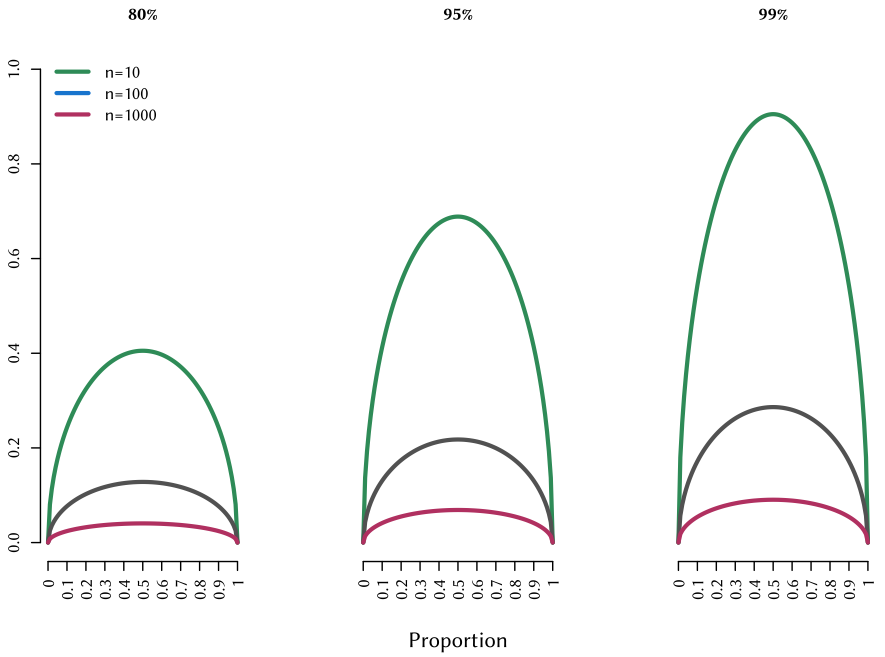

Figure 4.15: Width of error intervals for proportions at three sample sizes, three safety levels, and for proportions from 0 to 1

portions. This concludes our discussion of estimation and sample variance. There is one final word of warning: The error interval discussed here is the so-called *Wald interval* (named after Abraham Wald). It's very simple and it has some sub-optimal properties due to its relying on the distribution of sample proportions being *approximately normal*. For example, close to the extremes, the intervals can extend to minimally below 0 and above 1, which is nonsensical for proportions.[11] In general, it's not very exact. There are other frequentist alternatives available which are more exact but less suitable for a conceptual introduction like ours. If you ever have to calculate error intervals for proportions, look for the *Wilson score interval* or our favourite, the *Clopper-Pearson interval*. The idea behind all of them is the same, the alternatives just use different maths.

Wald interval

Wilson score interval

Clopper-Pearson interval

[11] However, anyone who thinks that this (of all things) is a strong argument against frequentism has no idea what they're talking about.

4.4 In-Depth The Normal Distribution

In this short section, we show the density function of the Normal Distribution as an added bonus for the curious. The purpose isn't to turn you into mathematicians or to prove anything. We just would like to show you that sometimes intimidating functions and equations can be understood almost intuitively with minimal effort in taking them apart. This is the general density function of the Normal Distribution:

$$\frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{4.8}$$

Remember that μ and σ are parameters. They define the precise shape of the distribution, but they do not vary with the actual input values x. Imagine a parametrisation for a given sampling distribution of means—which, as we've shown follows a Normal Distribution—, where $\mu = 120$ and $\sigma = SE_\mu = 20$ as in our example above. These parameters determine the centre x-value (mean) and slope (standard deviation) of the distribution, but the input values correspond to means for which the density function gives us the probability. As x does not occur in the first multiplicand, it is a constant in the parametrised function, and we ignore it for the time being. We're left with:

$$e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

This is an exponential function, a function with the Euler number ($e \approx 2.72$) as its base and *something* in the exponent. In this case, this *something* is first and foremost x^2, which gives us a simple parabola, see Figure 4.16 panel (1):

$$(1)\ x^2$$

The exponential function changes its shape (making it flatter around its minimum), and e^{-x^2} is just the reciprocal of e^{x^2}, which means it inverts the curve, which will then also be squashed between 0 and 1. We get something the shape of which already looks exactly like a normal density function, see Figure 4.16 panel (2).

$$(2)\ e^{-x^2}$$

Dividing the exponent by $2\sigma^2$, which is twice the variance, simply scales the curve horizontally. It spreads it out with increasing variance. Figure 4.16 panel (3) is exactly the same as Figure 4.16 panel (2), except the x-axis is scaled up:

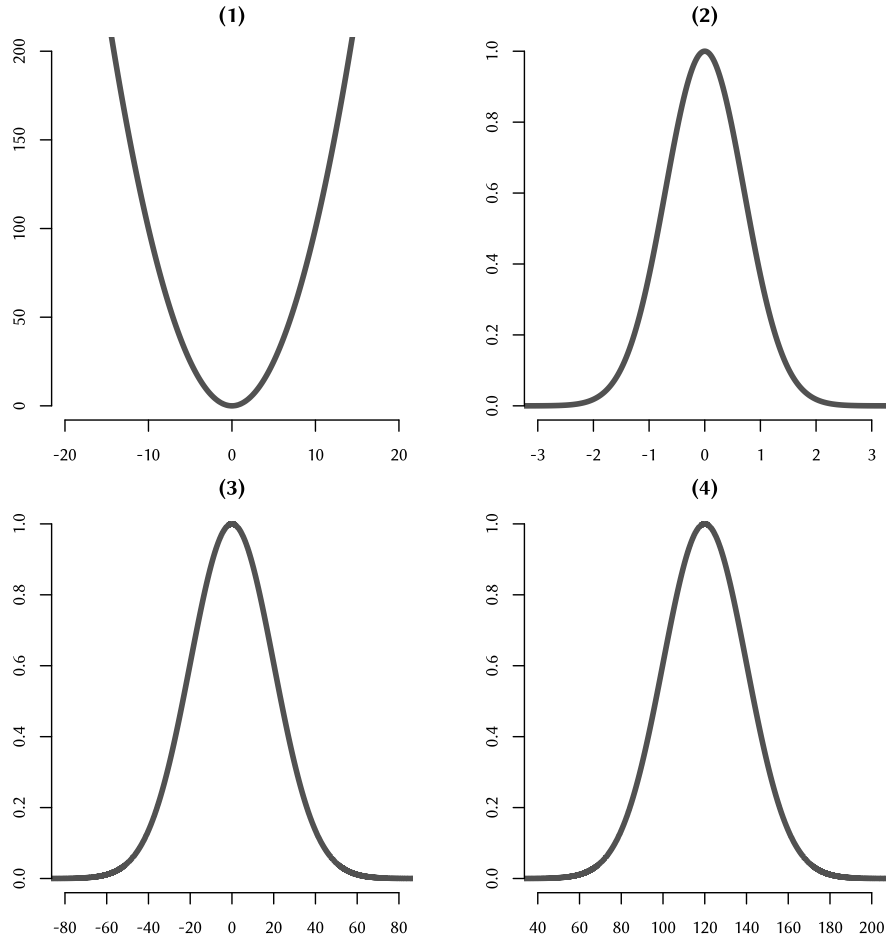

Figure 4.16: Dissecting the density function of the Normal Distribution

$$(3)\ e^{-\frac{x^2}{2\sigma^2}}$$

Subtracting μ from the numerator in the exponent just centres it around the mean μ. This happens independently of the exponential function. Try removing the exponential, and it will still move the curve (which would still look like a scaled parabola by itself). Figure 4.16 panel (4) corresponds to:

$$(4)\ e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

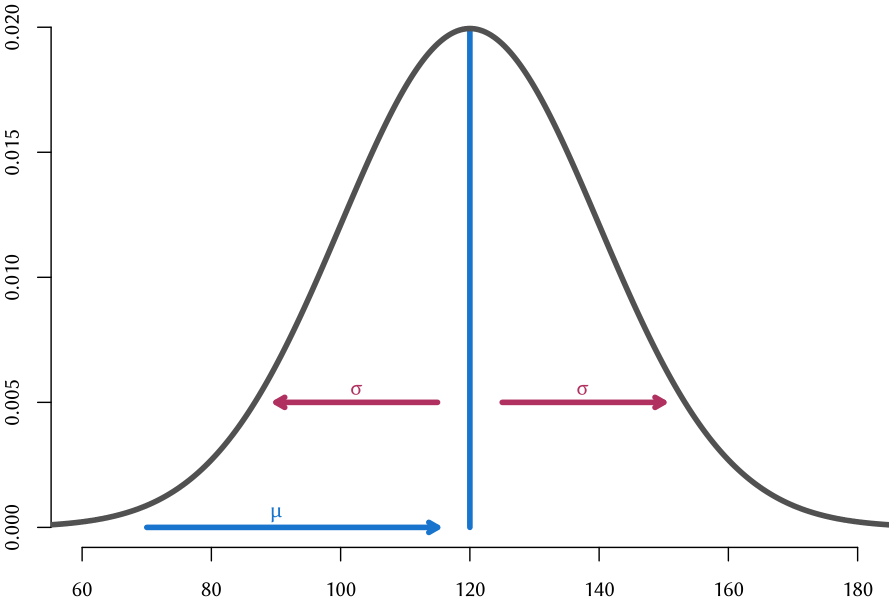

Figure 4.17: Dissecting the density function of the Normal Distribution; final version

It's still exactly the same as Figure 4.16 panel (2), just scaled on both axes and moved horizontally. Finally, the denominator of the constant term is equivalent to the area under the curve plotted in Figure 4.16 panel (4):

$$\frac{1}{\sqrt{2\pi\sigma^2}}$$

Hence, its reciprocal as a constant factor scales the function again to ensure that the area under its curve is 1, which is a requirement for any probability density function, see Figure 4.17. That's it. It's not rocket science to wrap one's mind around why the function looks the way it does and how the parameters μ and σ shape the curve. It's indeed rocket science to come up with (i. e., invent) useful probability distributions, work out their properties, and show why they have something to do with how the world works. To dig at least a bit deeper while still being entertained, two videos by Grant Sanderson (3Blue1Brown) are highly recommended: https://youtu.be/zeJD6dqJ5lo and https://youtu.be/cy8r7WSuT1I. For non-mathematicians, that's already quite sophisticated material.

Exercises for Chapter 4

(1) One Two Three

5 Inference: Mean Differences

Overview

In this chapter, we introduce the z-test and the t-test. We have introduced the logic of frequentist testing (Chapter 2), shown how samples can be summarised through measures of central tendency (Chapter 3), and we've shown how we can adjust the safety and the precision of means and proportions estimated from samples (Chapter 4). In this chapter, we return to inferential tests using the results from the chapters on descriptive statistics and estimation.

The tests introduced in this chapter compare means of samples. The z-test compares the mean of a sample with the mean of a population for which the true mean and the true variance are known for sure. The t-test for a single sample does the same but for cases where the true variance isn't known. Finally, the t-test for two samples compares means of two independent samples.

The substantive hypothesis in such tests is usually that the means differ. Hence, the Null is that they are the same. We calculate the frequentist probability of measuring the difference between the means that was actually measured (or a larger difference) if the Null is true. The frequentist probability is not calculated directly as in our presentation of Fisher's Exact Test. Instead, we calculate a test statistic (z or t), for which the probability density is known (in the form of the Normal Distribution and the very similar t Distribution) and we can use it to look up p-values.

Problem Statement: Inferences About Means

Let's assume you know the mean reaction time for a critical region when native speakers process a certain type of relative clause. This mean reaction time and the corresponding variance in measurements are extremely well established parameters. They were predicted by a robust theory of syntactic processing, and this prediction has been corroborated by a large number of diverse experiments. For an emergent subtype of this kind of

relative clause, the theory predicts considerably higher precessing effort and thus longer reaction times. You conduct an experiment and measure reaction times in the critical region. Which outcomes of the experiment would you interpret as indidcating that reaction times are indeed longer for the emergent type of relative clause?

5.1 Population Means and Sample Means

5.1.1 Introducing the Logic

The Problem Statement exemplifies a simple question: Given a known population mean, do means measured under a specific condition diverge from this known mean? In this section, we show through simple frequentist reasoning how measurements from experiments can provide evidence to answer such questions.[1]

z-Test

The simplest test for such tasks is the *z-Test*. Notice that for the z-Test to be applicable, the given population mean (and the corresponding variance) must be *known*! The test does not take into account any uncertainty in the value for the known population mean, and if you disregard that fact, you will end up in inference hell. This is why the Problem Statement stresses that the mean was predicted by a robust theory and that the prediction was tested in a long series of experiments. If these conditions are not met, other tests (such as the t-test for two samples) might apply, and we're going to introduce such tests as we go along.

For the sake of illustration, let's assume that the population mean is $\mu = 120$ (for example milliseconds) and the population variance is $\sigma^2 = 16$, which corresponds to a standard deviation of $\sigma = 4$. If the population values are generated according to a normal distribution, values are distributed according to the now well-known *Probability Density Function* (PDF) in Figure 5.1.

Probability Density Function

To recapitulate, the PDF gives for each measurement (x-axis) the probability with which it occurs (y-axis). Informally speaking, the curve shows that if we measure random values from this population, the probability of measuring a value close to μ is highest, and measurements deviate on average by the standard deviation σ from μ. The blue line shows the mean, and the green lines show one standard deviation in each direction from the mean. As a result, a very much

[1] Again, we caution readers that we can neither (by no means!) *decide* the questions nor *provide hard evidence* for any possible answer to them, etc.

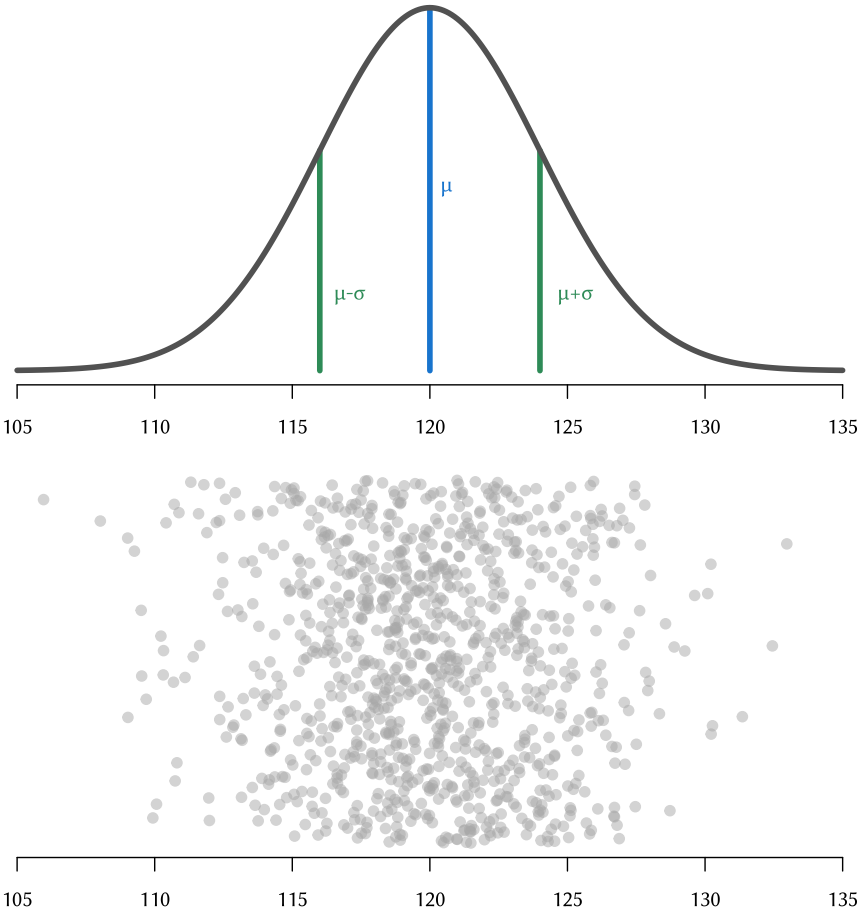

Figure 5.1: Theoretical population distribution for a normal distribution with μ=120 and σ=4 and a simulated random sample of n=1000 measurements from the population

expected sample of $n = 1000$ measurements is shown in the form of the point cloud below the curve (90° rotated compared to the raw data plots we've shown before). It's a simulated sample, and the simulation was set up according to the known facts about the overall population ($\mu = 120$and $\sigma^2 = 16$). The measurements are indeed centred around the mean, and they seem to follow the normal distribution.

If, however, we draw a sample from a different population where the true mean is slightly higher (for example because we're measuring reaction times un-

der a condition that incurs a processing penalty) we expect samples to turn out differently on average and have a higher sample mean compared to the known population mean. However, this on-average expectation can be treacherous because individual samples are not in any way *guaranteed* to represent their population well, as we have shown in Chapter 4. Very similar to Ronald A. Fisher in his experiment with Muriel Bristow (see Chapter 2), we need to ask whether the actual sample warrants any inference regarding the underlying mechanism. It does that if it's a sufficiently unexpected result under the assumption that the desired inference is not correct, i. e., under the Null. It's sufficiently unexpected if it had a low pre-experiment probability of occurring.

We'll show how this works out for the given example. In the case of the reaction times described in the Problem Statement, substantive hypothesis is that reaction times are higher with the emergent subtype of relative clauses because of assumed processing penalties. However—especially if our sample is small—inferring anything from a specific result is tricky. Figure 5.2 shows a possible outcome with $n = 16$ in red, and it should be obvious why it's tricky to infer anything from it. The mean from the sample is higher than the theoretically known mean, but this might very well be just a random deviation. In frequentist terms, such a sample is expected under the Null as well. We need to know the probability of a sample that deviates from the known mean to such a degree (or a stronger degree) in order to proceed to an inference.

Let's call the sample plotted in Figure 5.2 $\mathbf{x}$, a tuple of 16 measurements x_1 through x_{16}. The mean of $\mathbf{x}$ is $\bar{\mathbf{x}} = 122.1$. As we've shown, inferences in frequentist logic (see Chapter 2) are always made by taking into account what the outcome of an experiment could have been under one or several possibly correct hypotheses. In the case at hand, we're interested in the hypothesis that the true mean under the experimental condition (call it μ_1) is larger than the known mean μ. In other words, we would like to gather evidence in support of a substantive hypothesis: $\mu_1 > \mu$. We use the symbol μ_1 for the hypothesised larger mean as it is the mean of a slightly different theoretical population, and μ is the symbol reserved for population means.

For several reasons, we cannot gather evidence that supports this hypothesis directly. First, as a population mean μ_1 is obviously not directly observable. We can virtually never observe whole populations, all we've got are sample. Rather, μ_1 is a hypothesised mean in a population that exists as separate from the known population if the substantive hypothesis is correct. However, if that population is not substantially different from the known one, then we have $\mu_1 = \mu$. While we certainly hope that $\bar{\mathbf{x}}$ is a good indicator of the true value μ_1, we have no guarantees whatsoever that it actually is. Second, as Fisherians we have no

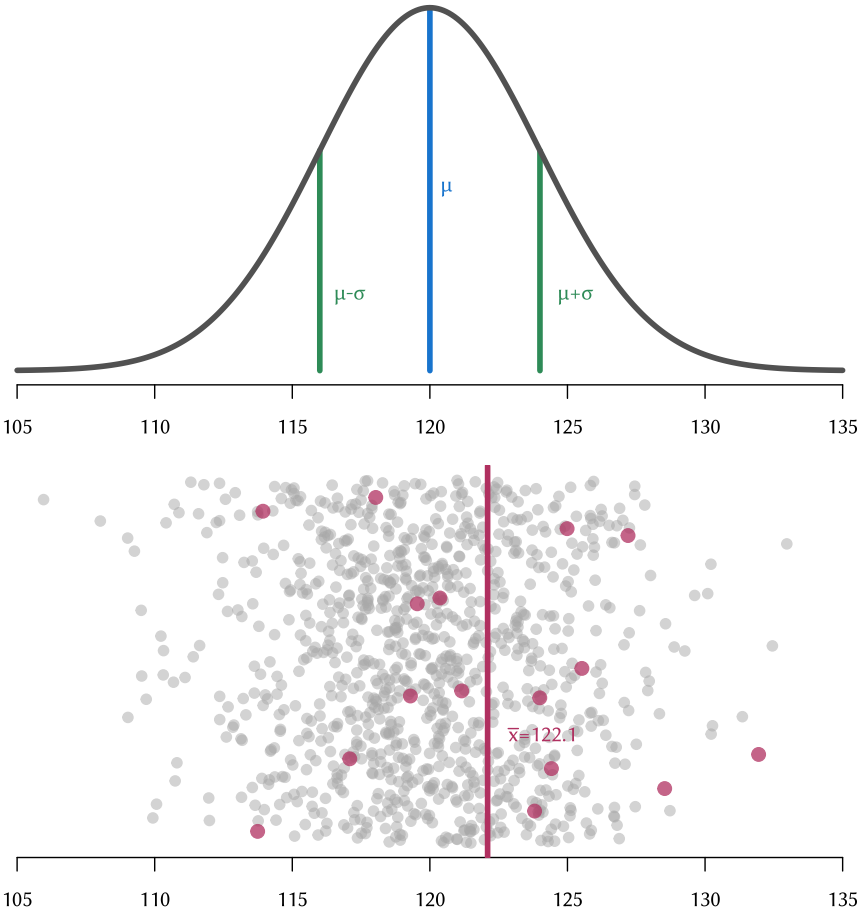

Figure 5.2: Theoretical population distribution for a normal distribution with μ=120 and σ=4 and a simulated random sample of n=16 measurements from some population

formal method of gathering positive evidence.[2] Therefore, we can only check whether the data $\bar{\mathbf{x}}$ are in accord with a Null $H_0 : \mu_1 \not> \mu$, which is equivalent to $H_0 : \mu_1 \leq \mu$.

Let's recapitulate the frequentist logic of inference and adapt it to the case at hand: To test this hypothesis, frequentism assigns a certain well-defined kind of

[2]Only Neyman-Pearson philosophy and the Severity approach will give us this power (pun intended) later in Chapter 7, but with some important caveats.

probability to (obtaining) the data $\mathbf{x}$ given that the Null is correct. This probability can be used to assess whether the data $\mathbf{x}$ are in accord with the Null. First, what do we infer from the data $\mathbf{x}$ (in other words, from our experiment) if they are compatible with the Null? You're right, absolutely nothing! If the data are in accord with the Null, we haven't found any evidence that it is not the case that μ_1 is not greater than μ, and that's the end of it. If that sounds uselessly messed up and disappointing, that's because it is.[3] If you infer anything from such a result, you're not only wrong, but also Bayesians (many of them with Dutch names) will come to haunt you (or at least unfollow you on the messaging platform of your choice)—and they'd be somewhat right to do so. Second, what do we infer from the data $\mathbf{x}$ if they are not compatible with the Null? This is the much more interesting case, but it's difficult to define the admissible inferences without creating false ideas. Let's say rather informally that in such a case we've found some evidence in support of the substantive hypothesis because it and the Null partition the range of possible values of μ_1: either it's greater than μ or it is not greater than (i. e., smaller than or equal to) μ (H_0). Finding no accordance with the Null despite serious attempts to do so (see Chapter 7) provides at least some indication that the substantive hypothesis might be correct. However, if you're looking for a proof of anything, we recommend that you stick to pure theory, logic, theoretical maths, or pseudoscience. There is no proof to be found in (non-trivial) experiments, and statistical inferences are weak and fragile.

5.1.2 Extreme Means Under the Null

We have argued that in Fisherian inference, we have to assess whether $\mathbf{x}$ and its mean $\bar{\mathbf{x}}$ are compatible with the Null. But how do we do this? Because we're comparing means (i. e., numeric values) and not merely counting occurrences of correct and incorrect tea-first detections, it seems difficult to compute the number of all possible outcomes and the number of outcomes as extreme as or more extreme than the one we actually observed. After all, we're dealing with numeric measurements (real values), and there's always a result in between two results, e. g., between $\bar{\mathbf{x}} = 123.99999$ and $\bar{\mathbf{x}} = 124$, there are infinitely many other possible results such as $\bar{\mathbf{x}} = 123.999991$, $\bar{\mathbf{x}} = 123.9999911$, etc. Well, it's not totally impossible to calculate exact probabilities, but there's a convenient shortcut. We'll go through it step by step.

The most naive but not at all wrong thing to do is to calculate the difference between the known population mean μ and the mean of the obtained sample $\bar{\mathbf{x}}$.

[3] Whether you're a linguist or not, please consider that *finding no evidence that A is true* is not the same as *finding evidence that A is false.*

In our case, this is $\bar{x} - \mu = 2.1$. Clearly, a minimal requirement for any further calculations is that this difference is positive. If it were negative, the sample could hardly be interpreted as evidence against N: $\mu_1 \leq \mu$.[4] While this is something one should always do first, it's not suitable for serious inference due to one main reason: It doesn't take into account how large the sample was.

We now follow a very similar logic as in our introduction to Fisher's Exact Test (Chapter 2). The question is: How often would we expect to see a sample of 16 measurements with a sample mean of 122.1 or larger if the true mean is that specified by the Null, which is $\mu = 120$? Luckily, we have already introduced the tool that we need: the *standard error* of the mean SE_μ. The standard error for $n = 16$ and the known variance $\sigma = 4$ (see p. 104) tells us how strongly samples of this size deviate from the true mean (on average) in each direction. The standard error of the mean in this case is (revise Equation 4.1 if necessary):

standard error

$$SE_\mu = \frac{4}{\sqrt{16}} = \sqrt{\frac{4^2}{16}} = 1$$

Remember what the standard error is all about (Chapter 4). If the mean in a population is μ and the standard deviation is σ, then the sample means of samples of size n are themselves normally distributed, and the standard error SE_μ is the standard deviation of that normal distribution. Furthermore, keep in mind that we're talking about the distribution of sample means drawn from a known population. Under the Null, it is also the distribution from which our small sample was drawn. Figure 5.3 contrasts the density of the distribution of individual data points (in our example: individual reaction times) with the much narrower distribution of sample means. Mathematically, it is narrower because the standard error is always smaller or equal to the standard deviation. Intuitively, it should be narrower because on average sample means from samples with $n > 1$ approximate the true mean better than single measurements.[5]

Remember from Chapter 4 that a Normal Distribution is exhaustively defined by the parameters μ and σ. Since (i) the distribution of sample means is Normal, (ii) we know its mean μ, (iii) we know its variance σ^2 and standard deviation σ,

[4] This way of putting it is slightly sloppy and informal. We will return to this notion and make it more precise in Chapter 7. However, in practice it is blatantly obvious that we would never take an experiment that showed lower reaction times as evidence for higher reaction times, etc.

[5] Understanding this argument is crucial. If you're not following it, you should go back to Chapter 4 for an introduction to the distribution of sample means, especially the argument concerning samples of increasing sizes in Sections 4.1.2 and 4.1.3.

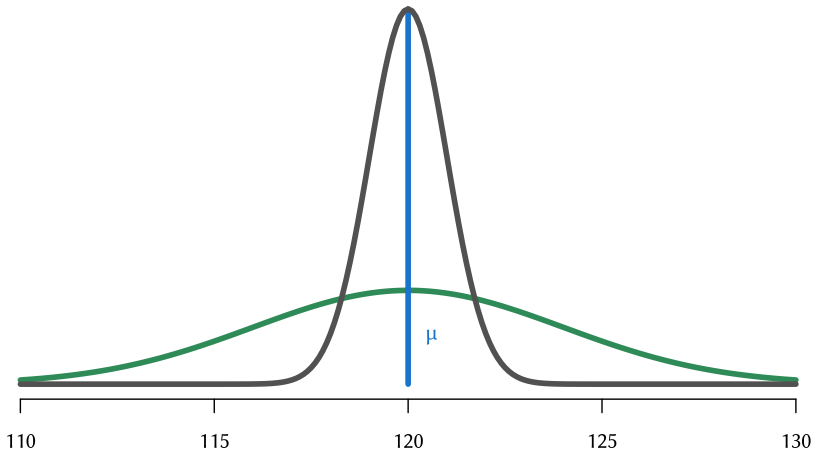

Figure 5.3: Theoretical population distribution for a normal distribution with μ=120 and σ=4 (green) the distribution of sample means for samples from this distribution with n=16 (black)

we can calculate how many samples (in the long run) drawn from the known population would have a mean of 122.1 or larger. In other words, we can calculate how many sample means (from samples of size $n = 16$) would deviate by 2.1 from the population mean anyway due to expected sampling error. Instead of calculating the numbers and probabilities of individual events as in the Tea-Tasting Experiment, we use the known functions of the Normal Distribution to look up those values. This gives us a very precise and well-defined measure of how unexpected the obtained result would be if the Null were true.

cumulative distribution function

The left panel of Figure 5.4 show the PDF of the distribution of sample means, and the highlightes area under the curve corresponds to results as extreme or more extreme than $\bar{x}$. The corresponding *cumulative distribution function* (CDF) can be used to calculate the statistics of interest. The right panel of Figure 5.4 shows the Normal CDF for $\mu = 120$ and $\sigma = 1$ (which happens to be SE_μ). For the observed sample mean $\bar{x} = 122.1$, the CDF has the value 0.98, which means that the probability of obtaining this or a more extreme result is $1 - 0.98 = 0.02$. Think of the value 0.98 as indicating that 98% of the probability mass lie to the right of $\bar{x} = 122.1$. If you have access to a software that has built-in functions for such CDFs, this is the most straightforward way to arrive at a p-value for a sample mean and a known Normal Distribution. In this case, you don't even need to calculate z to perform a z-test.

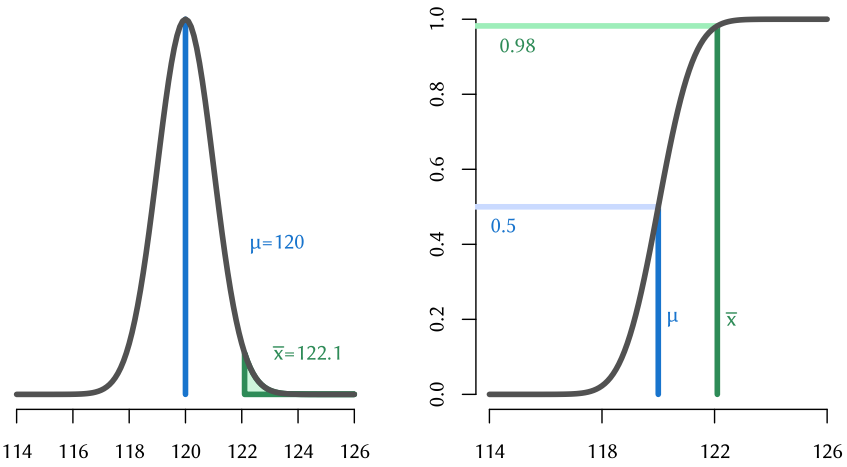

Figure 5.4: Left: PDF of the distribution of sample means for mean μ and standard error SE with obtained sample mean $\bar{\mathbf{x}}$ and corresponding z-value; area under the curve for results equal to or greater than $\bar{\mathbf{x}}$ is shaded; Right: CDF for the same distribution with obtained sample mean $\bar{\mathbf{x}}$

Otherwise, you can calculate the so-called z-score and look up the p-value in a table. To to this, we simply count the distance between the distribution mean and the sample mean in multiples of the standard error, which gives us the *z-score*. It's the same z we introduced in Chapter 4, except we're using it the other way round. For estimation, we looked up z-scores corresponding to probabilities, now we're calculating z-scores in order to lookup probabilities. In this case, it's particularly easy because $SE_\mu = 1$:

z-score

$$z = \frac{\bar{\mathbf{x}} - \mu}{SE_\mu} = \frac{122.1-120}{\sqrt{\frac{4^2}{16}}} = \frac{2.1}{1} = 2.1$$

By dividing the distance by the standard error and calculating the z-score, we normalise it and make its interpretation independent of the concrete slope (σ) of the distribution. The z-score is the distance from the mean in a Standard Normal Distribution (with $\mu = 0$ and $\sigma = 1$) that corresponds to the measured distance from the known mean. Hence, $z = 2.1$ can be interpreted independently of the concrete sample mean and population mean, and it establishes a direct link to the probability we're looking for: another *p-value*. We can use tables to look up

p-value

p-values corresponding to z-scores, and for $z = 2.1$, we get $Pr = 0.02$, which in our application is the p-value $p = 0.02$. Table 5.1 shows such a table, and it's the reverse of tables such as Table 4.1.By the way, the values are those for the Standard Normal Distribution with $\mu = 0$ and $\sigma = 1$. Do you see why? We give two columns, one for the single-sided test and one for the two-sided test (see p. 2.3.2.1, immediately below, and Section 5.1.3).

Table 5.1: Probabilities for Some z-Values

Z	Pr (one-sided)	Pr (two-sided)
1.5	0.07	0.13
1.6	0.05	0.11
1.7	0.04	0.09
1.8	0.04	0.07
1.9	0.03	0.06
2.0	0.02	0.05
2.1	0.02	0.04
2.2	0.01	0.03
2.3	0.01	0.02
2.4	0.01	0.02
2.5	0.01	0.01

Clearly, either the Null is false or the Null is true and a relatively rare event has occurred. Whether a chance of 1 in 50 (equivalent to $p = 0.02$) is rare or unexpected enough to make an inference can only be decided by you based on your knowledge of the field you're working in. How precise are your measurements? What magnitude does your theory predict the difference should be?[6] How does the measured difference in reading times compare to differences observed for similar processing penalties? In corpus linguistics, we fail to see how $p = 0.02$ would be surprising enough in any situation to warrant substantive inferences.

one-sided test
two-sided test

One final word on the *one-sided test* and the *two-sided test*, also called single-tailed and two-tailed. In the example in this section, the substantive hypothesis was $\mu_1 > \mu$, and the Null was consequently $H_0 : \mu_1 \leq \mu$. We were interested in the probability of obtaining a mean as high as 2.1 or higher if $\mu_1 \leq \mu$, and hence values that deviate negatively (to the left of) $\mu = 120$ wouldn't be unexpected at all and aren't counted. There are two other possible scenarios we could have wished to test. Our substantive hypothesis could have been $\mu_1 < \mu$. Alternatively,

[6] It doesn't make numeric predictions? Blimey! Then back to the drawing board.

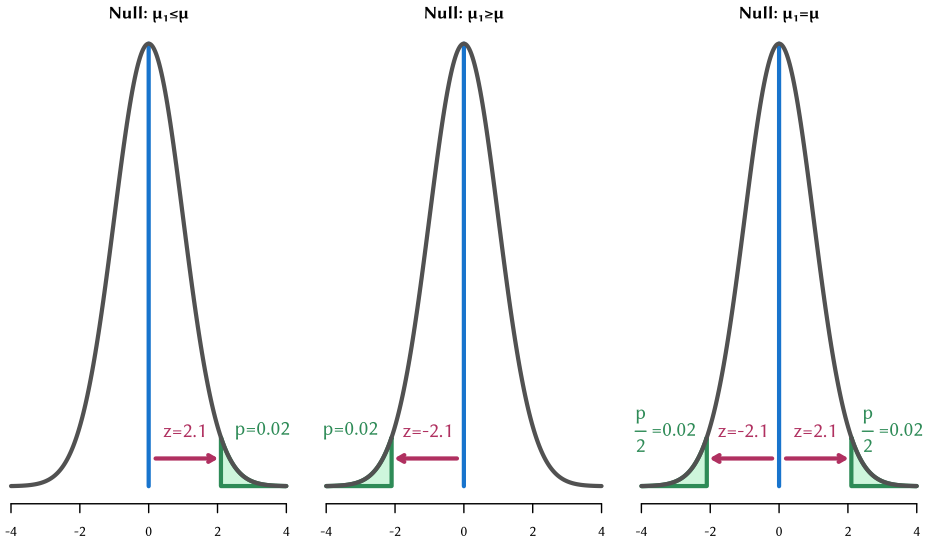

Figure 5.5: Single-sided and two-sided tests

it could have been $\mu_1 \neq \mu$. The corresponding Nulls are $H_0 : \mu_1 \geq \mu$ and $H_0 : \mu_1 = \mu$.

Each of these cases corresponds to a slightly different test, illustrated in Figure 5.5 for the Standard Normal Distribution as the distribution of sample means, but with the z-value from the example above, i. e., $z = 2.1$. On the left, the test as conducted above is illustrated once again. We deviate by $z = 2.1$ from the known mean μ, and that defines 2% of the area under the curve, hence $p = 0.02$. In the middle, the opposite is shown, which is applicable if the Null is: *the unknown mean μ_1 is equal to or larger than the known mean μ*. In this case, we're only interested in negative deviations from the known mean μ. The probability corresponding to a negative z-statistic in the one-sided case is the same as the one corresponding to a positive z-statistic, hence $p = 0.02$ according to Table 5.1. Finally, if your hypothesis is an *undirected hypothesis* and we only suspect the sample mean $\bar{\mathbf{x}}$ represents a mean μ_1 that is different from μ, we have to take into account both tails of the distribution, which simply doubles the p-value to $p = 0.04$. The right panel in Figure 5.5 clearly shows that values as extreme of more extreme than $z = 2.1$ in either direction of μ correspond to the double area under the curve.

undirected hypothesis

If you think this is trivial, you're lucky. However, don't be too hard on yourself if you find this difficult to fiddle apart and to memorise. This is completely normal, and it has nothing to do with you or frequentist statistics. Fort some reason,

many human brains have difficulties of keeping track of positive and negative signs of z-statistics, greater-than and smaller-than relation, while having to flip them around in formulating and evaluating hypotheses. Just give it time.

5.1.3 The Difference Between Error Intervals and the z-Test

The z-test allows us to check whether a mean measured in an experiment was an unexpected outcome under the Null. We attach a specific statistic—the p-value—to the outcome. It's a matter of debate whether the statistic itself should be interpreted directly numerically. However, you've probably heard of people using the p-value to make decisions when it's smaller than a certain threshold called α-level or—as we prefer—*sig*-level They call results *significant* when they reach a certain level such as $sig = 0.05$. If $p < sig$, *we've reached significance*. However, when $p = 0.009$ in the next experiment, suddenly that result might be called *even more significant*. While we consider it vital to report the actual p-value and not just $p < sig$, we advise against cascaded conceptions of significance. Whether an outcome would be unexpected under the Null depends on the phenomenon at hand, the precision of your theory and of your measurement, whether you've ensured that the conditions for the test were met properly, and probably many other factors. Nobody can tell you or your reseqarch community what your threshold needs to be. For example, you might have seen reports from nuclear physics (such as results from experiments with the Large Hadron Collider of CERN) where it is claimed that 5σ was reached. That's just 5 standard deviations from the expected value under the Null ($z = 5$), which corresponds to approximately $p = 0.00000029$. In physics, this is not interpreted as a test but rather as a requirement for measurements to be taken seriously at all, to be accepted as measurements of something. That's because there is large uncertainty in the measurements, and the phenomena of interest are tiny beyond microscopic and can only observed very indirectly. Do you need that level of unexpectedness? Think about it. If you're uncertain and your field has just gotten started (such as linguistics, cognitive science, or social psychology), we recommend you don't dare to be any more lenient than $sig = 0.001$. Honestly, $sig = 0.05$ is borderline ridiculous!

In any case, if you reach some sig-level under a testing approach, you might proceed to an inference, a decision of some sort. Usually, it should be something like this:

> *We're not going to discard our substantive hypothesis just yet, but we'll submit it to further error probing.*

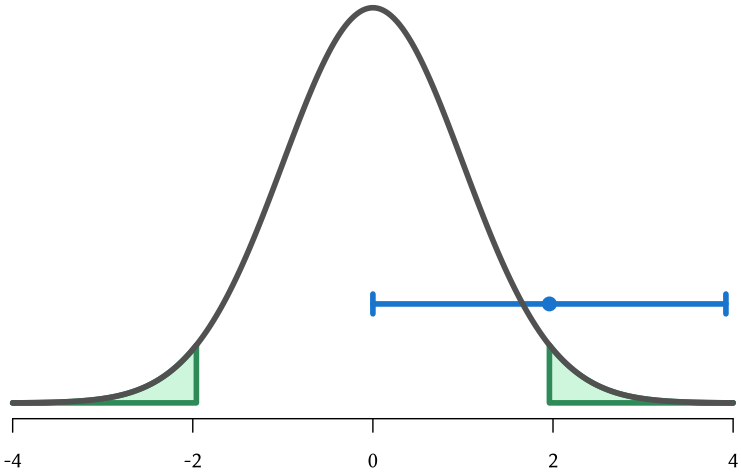

Figure 5.6: Convergence of z-test and error interval: two-sided test

Do you see how this is different from the following statement?

> *We've statistically proven that our substantive hypothesis is correct with a probability of 95%.*

If you don't see this, go back to Chapter 2 and start all over again.

While tests and error intervals are related mathematically (see below), they're entirely different in their interpretation. You can estimate a value without having any hypothesis at all. For example, if you just want to find out the average rate at which a manufacturing machine produces defective parts, you might take a sample of 1000 parts, count the proportion of defective parts, calculate a safe error interval (at, say, $a = 0.99$), and see whether the resulting interval looks acceptable from an economic point of view. If you do this with all your machines over many years and decades, your estimate will included the true value in close to 99% of all times. Depending on the value of the individual parts that you're working with in your business, this might be acceptable. In Chapter 7, we'll return to this example from toy manufacturing and move it closer to a testing framework.

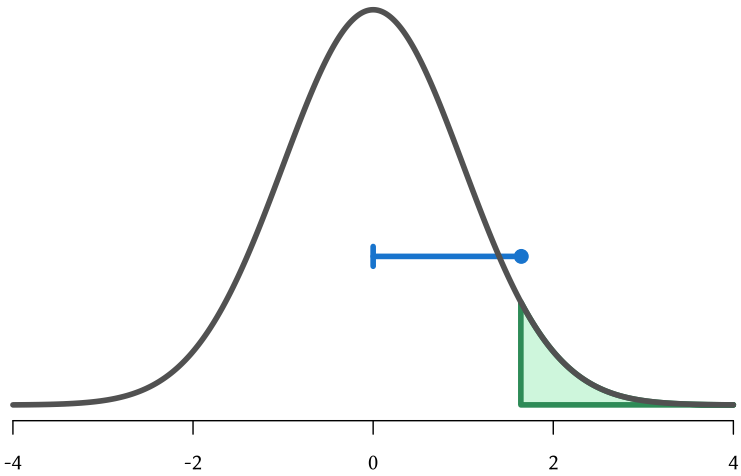

Figure 5.7: Convergence of z-test and error interval: single-sided test

Their interpretations differ, but frequentist error intervals and frequentist tests are related mathematically. To illustrate, let's use an even simpler example where the sampling distribution is the Standard Normal Distribution. Hence, $\mu = 0$, $\sigma = SE_\mu = 1$, and we assume that the measured mean was $\bar{\mathbf{x}} = 1.96$. Since the standard error is 1, all calculations are quite simple. First, for the test, $z = 1.96 \div 1 = 1.96$. This gives almost exactly $p = 0.05$ for the two-sided test, i. e., the test which merely checks whether the measured mean lies in the outermost 5% of the probability mass. In Figure 5.6, the shaded areas correspond to those five percent, and they begin at a distance of 1.96 to both sides of the mean. If we calculate an error interval with $a = 0.95$ around the measurement of $\bar{\mathbf{x}} = 1.96$, it is exactly 1.96 wide to both sides of the measured mean. Hence, it includes 0, as you can see in Figure 5.6, where the error interval is plotted in blue.

The two-sided z-test at some sig-level (in this case $sig = 0.05$) is equivalent to checking whether the error interval for $a = 1 - sig$ contains the population mean μ. This extends to a single-sided test and a single-sided error interval, which only extends to one side from the measured mean. See Figure 5.7, where we assume $\mu = 0$, $\sigma = SE_\mu = 1$, and $\bar{\mathbf{x}} = 1.65$. The error interval is left-sided, and it corre-

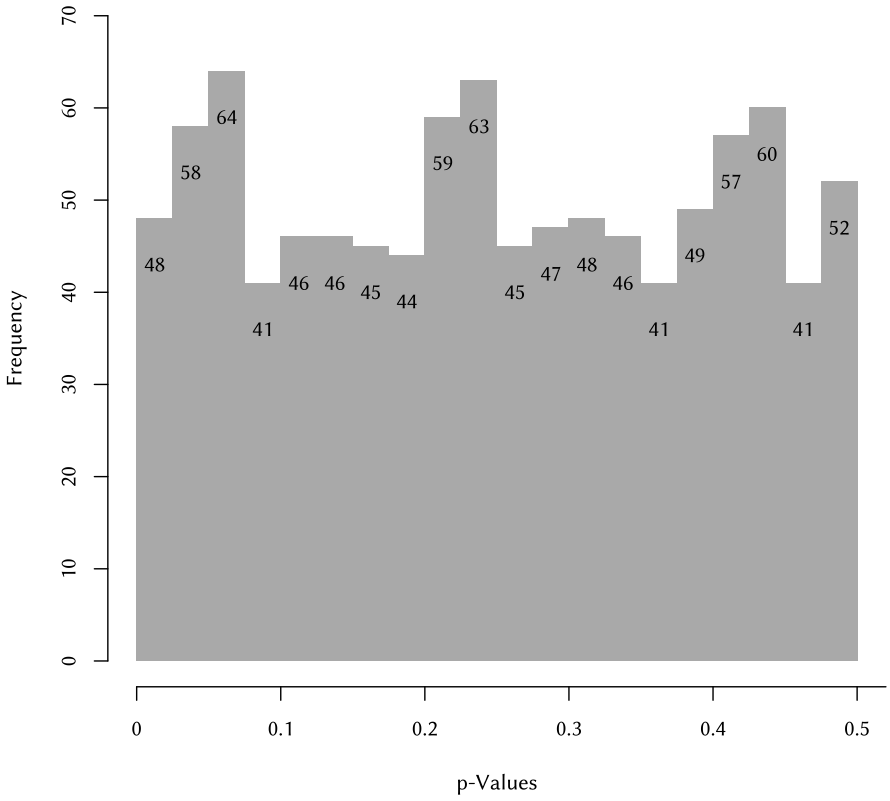

Figure 5.8: Histograms of p-values from z-tests when the Null is true; it's irrelevant what the parameters were, as any p-values have a uniformly random distribution under any true Null

sponds to the right-sided test. As $\bar{\mathbf{x}} = 1.65$ was chosen as the z-value that reaches $sig = 0.05$ for the single-sided z-test, the interval again includes 0.

5.1.4 The Distribution of p-Values

In Section 2.5 we showed how simulated p-values of Fisher's Exact Test are distributed when the Null is true and when it it's false. In this section, we'll do the same for the z-test. First, let's assume the Null is true. We simulate a two-sided

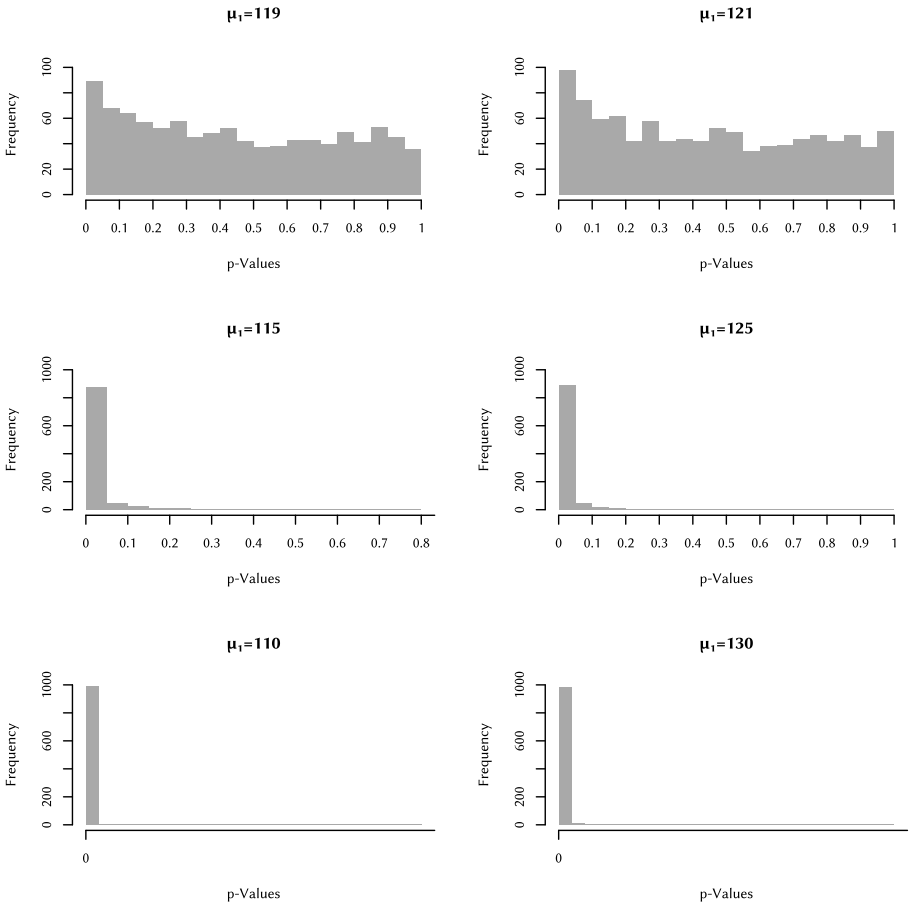

Figure 5.9: Histograms of p-values from z-tests when the Null is false with different effect strengths

test with $H_0 : \mu_1 = \mu$. The measurements are normally distributed with $\mu = 120$ and $\sigma = 16$. We simulate $n = 100$ data points per sample in 1000 replications. The standard error of the mean is $SE_\mu = 16 \div \sqrt{100} = 1.6$. Figure 5.8 shows that each p-value has (minor fluctuations aside) the same probability if the Null is true. If the mean within the population from which our samples are drawn is $\mu_1 = \mu = 120$, any p-value between 0 and 1 has an equal chance of resulting from the experiment.

We then simulated the same for $\mu_1 \neq \mu$. Clearly, there are many ways to satisfy $\mu_1 \neq \mu$. The difference between the two means could be -0.1, $+29$, or

what have you. All these different possible concrete numerical inequalities between μ_1 and μ make the Null false. The larger the difference, the larger the effect strength. For example, if reading times under a marked condition differ from those in the unmarked condition by 200 ms, the effect is much larger than if they differ by only 2 ms. We simulated 1000 samples each for the true means $\mu_1 \in \{119, 121, 115, 125, 110, 130\}$. Figure 5.9 shows histograms of the results. Even if the difference between the means is only 5 ($\mu_1 = 115$ or $\mu = 125$), the p-values already collapse close to 0 (middle left and middle right panel). Once again (remember Section 2.5), larger effect strengths make it easier to detect the effect by rejecting the Null. As in the case of Fisher's Exact Test, a large sample alone does not increase the chance of obtaining a low p-value. However, as soon as there is even a small true effect, and we draw a relatively large sample, p-values go down rapidly. If the effect is tiny and irrelevant from a theoretical point of view, we can detect it easily with a large sample, it's our duty to have a good enough understanding of the statistical methods we use to be able to interpret that. This behaviour of the test itself is, of course, by design. Frequentist statistics per se is not responsible for misinterpretations and misuse by lazy or corrupt practitioners.

5.2 The Undiscovered Population

5.2.1 Accounting for Unknown Variance

The z-test allowed us to test whether some mean μ_1 is different (smaller, larger, or any of the two) from a known population mean if the population variance is unknown. Some readers might have wondered when on Earth that is actually the case. Since the t-test doesn't account for any uncertainties in the knwon mean and variance, they must be known beyond any need for further testing. This is where the *t-test* comes into play. It allows us to test for mean differences if the variance is unknown (this section) or both means and the variance are unknown (Section 5.2.2).

t-test

The logic of the test is perfectly identical to the z-test. We simply substitute the sample variance s^2 for the population variance σ^2 and go on with the calculations but call the test statistic t instead of z (hence *t-test*). To remind us all of Equation 3.4:

$$s^2(\mathbf{x}) = \frac{\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2}{n-1} = \frac{SQ(\mathbf{x})}{n-1} \tag{5.1}$$

Do you see why this can't be all? While the sample variance might be the best approximation of the population variance we have, it's not going to be a perfect approximation in most (virtually all) cases. There might be cases where s^2 *overestimates* σ. Imagine what this would do to our statistic t, which is calculated exactly the same way as z:

$$t = \frac{\bar{\mathbf{x}} - \mu}{SE_\mu} \tag{5.2}$$

Except SE_μ is calculated from the sample variance:

$$SE_\mu = \sqrt{\frac{s^2}{n}}$$

With increasing s^2, SE increases, and t decreases as the denominator of its formula is SE. However, smaller values of the z or t statistics leads to larger p-values. Look at Figure 5.4 to convince yourself that this is true. Whichever criterion we adopt to reject the Null based on an obtained p-value, we'd reject less Nulls.[7] Since we attach no inferential interpretation to a rejected Null, rejecting more Nulls than we should because the sample variance overestimates the population variance is not a big deal.[8]

If, on the other hand, the sample variance s^2 underestimates the population variance σ^2, we might end up rejecting less Nulls than we should. Convince yourself that this is true by going through the formulas above. This is potentially harmful as rejections of a Null have a limited inferential interpretation: *Either the Null is false or a rare event has occurred.* To compensate for this, a statistician by the name of William Sealy Gosset, who called himself *Student*, came up with a distribution similar to the Standard Normal Distribution: *Student's t-Distribution*. It accounts for the fact that in the long run a larger sample provides a better estimate of the population variance than a smaller sample. The distribution has only one parameter: the sample size n, or rather the *degrees of freedom* ν with $\nu = n - 1$. We'll return to the idea of degrees of freedom repeatedly. For now, we just contemplate why the t-Distribution has this one parameter and don't worry too much about its name.

Student's t-Distribution

degrees of freedom

First, the t-Distribution corresponds to the *Standard* Normal Distribution, not the more general Normal Distribution. Hence, its variance is fixed at 1 and its

[7] Once again, we encourage you to go through this argument step by step multiple times until you're satisfied that you've understood and memorised the argument.

[8] If you find this a very unsatisfying claim, read on until Chapter 7.

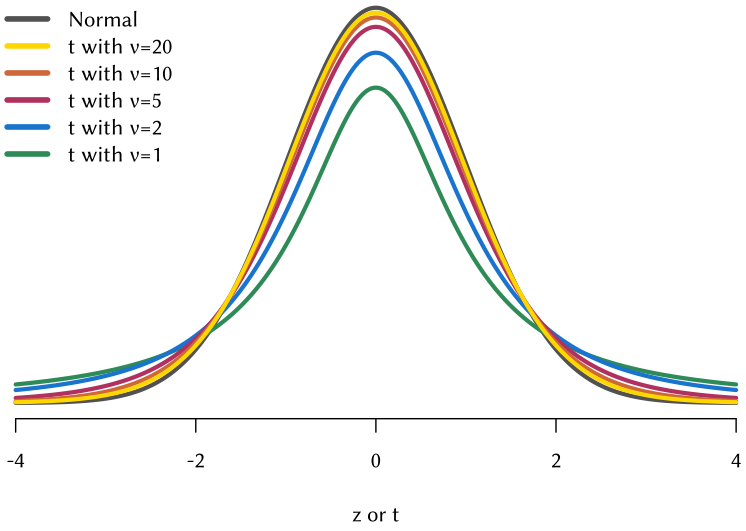

Figure 5.10: Densities of the Standard Normal Distribution and a selection of t-Distributions for increasing degrees of freedom ν

mean is fixed at 0. At the same time, the effect of the expected mismatch described above between the sample variance and the true population variance is small when the sample is large and vice versa. A larger sample approximates any population parameter more accurately than a smaller sample. Now, see Figure 5.10 to grasp the effect of the ν parameter. A lower ν, which corresponds to a lower sample size n, flattens the peak of the curve around the mean while lifting both tails up. However, even at $\nu = 20$, the density of the t-Distribution is already almost identical to that of the Standard Normal Distribution.

Figure 5.11 shows how two-sided p-values from the t-Distribution for a fixed t-value of 1.96 but with increasing sample size and thus increasing degrees of freedom. The t-value of 1.96 was chosen because we know that it corresponds to $p = 0.05$ in a z-Test with a Standard Normal Distribution. The Normal Distribution has no parameter for degrees of freedom, and the p-value is therefore always $p = 0.05$ for $z = 1.96$. With ever larger samples, the p-value from a t-test approximates that from a z-test very quickly. With a slight oversimplification, we can say that the t-test makes it harder to reach any standard for an unexpected

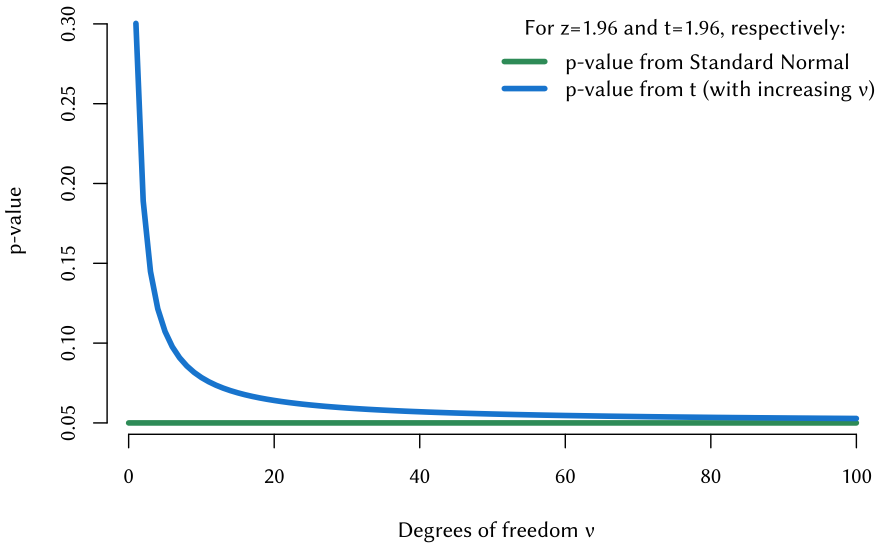

Figure 5.11: How the p-value from the t-Distribution at t=1.96 (two-sided) approaches the p-value from the Standard Normal Distribution at z=1.96 (two-sided) with increasing degrees of freedom ν

result compared to the z-test, thus accounting for the added uncertainty from the variance estimated from the sample. With a larger sample (higher degrees of freedom), that added uncertainty disappears more and more as the estimate becomes more and more accurate.

Finally, we simulate t-tests and compare them to incorrectly conducted z-tests. The simulation answers the question of what happens if we disregard the fact that the variance was merely estimated from a sample and was not known, and we decide to perform a z-Test. For simplicity's sake, let's say the known mean is $\mu = 0$, and the true but unknown standard deviation is $\sigma = 1$. Figure 5.12 shows the distributions of p-values for inappropriate z-Tests based on the sample variance and the corresponding appropriate t-Tests.[9] The only difference is that for

[9]These plots show mildly smoothed densities from a histogram, made using `hist(breaks = 40)`, the resulting densities, and `spline.smooth(spar = 0.6)` from R. They're not traditional kernel-smoothed density estimates as these would give a false impression by showing the density going down towards 0 and 1. Please ignore this information if you don't know what it means.

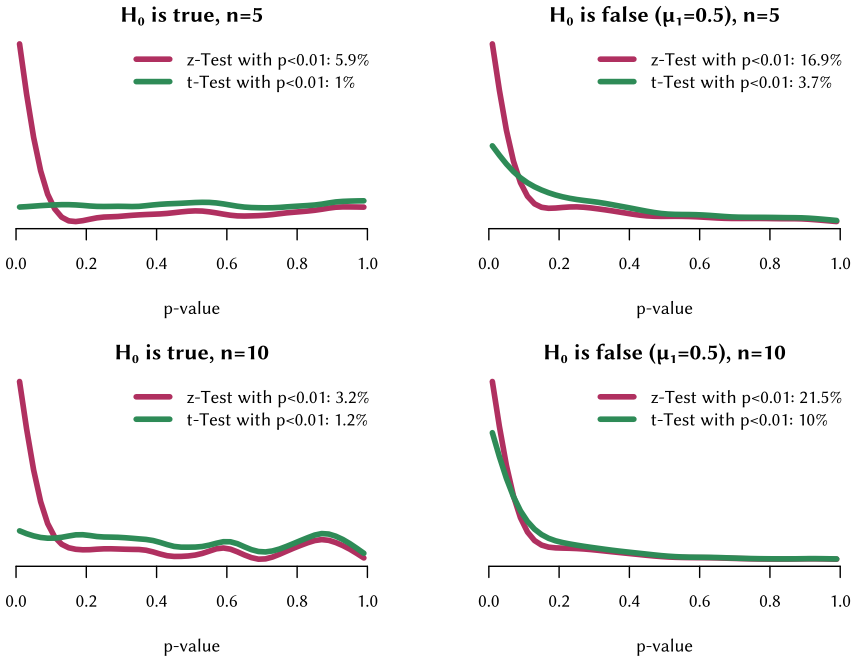

Figure 5.12: Distributions (histogram densities) of p-values from simulations of z-Tests and t-Tests with variances estimated from samples; the z-Tests are inappropriate in this case

the z-Test, the cumulative distribution function of the Standard Normal Distribution was used, and for the t-Test, the cumulative distribution function of the t-Distribution (with the respective degrees of freedom) was used. The z-value and the t-value are numerically identical in this case, and they both use the standard error calculated from the sample variance, which is, again, only correct for the t-Test:

$$SE_{\mathbf{x}} = \sqrt{\frac{s^2}{n}}$$

Each curve in Figure 5.12 plots the estimated density of p-values from 10000 random experiments with the accompanying two-sided tests. In the two left panels, the Null is true, and the data are generated from a Standard Normal Distribution. Hence, the the z-values and the t-values should be centred around 0, and all p-values should have the same long-run frequency, which is always assumed under the Null. We see that for $n = 5$, the z-Test produces incorrect results if the

variance is not known but merely estimated. Under the Null (upper left panel) and with $n = 5$ ($\nu = 4$), the t-Tests correctly land at $p < 0.01$ in 1% of all simulations, but the z-Tests produce many false significant results. Instead of the nominal 1% at $sig = 0.01$, the Null is rejected in 5.9% of the simulations.[10]

With more degrees of freedom, this problem is alleviated as the variance estimate gets more precise. In the lower-left panel, the simulations for $n = 10$ ($\nu = 9$) are shown. The t-Test is still almost spot-on and reaches $p < 0.01$ in 1.2% of all simulations. The z-Tests now produce 3.2% reaching the threshold, which means at least many fewer false significant results. Still, this would lead to false rejections of the Null and false positive inferences regarding the substantive hypothesis much more often than it should. The error rate increases sharply, as it should be 1% at $sig = 0.01$, 0.1% at $sig = 0.001$, etc.

In the panels on the right-hand side, p-values from simulations are shown where the Null is false. The samples now come from a DGP with $\mu_1 = 0.5$. We haven't discussed yet how this effect strength affects the tilt of the p-values numerically. Therefore, we have no precise expectation of how many simulations we expect to achieve $p < 0.01$. However, we clearly see that the inappropriate z-Tests are shifted much more heavily towards 0. Under the assumption (corroborated in the simulations with a true Null) that the t-Test is appropriate and the z-Test isn't under the given circumstances, the z-Test will lead to many more wrong inferences than we expect from correct z-Tests. In sum, using the z-Test where a t-Test would be appropriate leads to a considerable and incorrect increase in rejections of the Null. As this would mar our inferences, the t-Test is a highly useful tool when the true variance is unknown. But, we hear you say, population means are also largely unknown, which makes the t-Test much less useful than we claim. Hold that thought! The solution is presented in the next section.

5.2.2 Accounting for Two Unknown Means

5.2.2.1 Extending the Logic of the t-Test to Two Samples

We now turn to a much more realistic scenario: You haven't got any precise idea about the population mean to which you could compare a sample mean. Also, the population variance isn't known with sufficient precision. In such situations, researchers usually compare two samples $\mathbf{x}$, $\mathbf{y}$ taken under two conditions, asking whether the sample means differ strongly enough to warrant the inference that they come from two populations where the population means differ. The logic of

[10] We arbitrarily chose $sig = 0.01$ for convenience. Results would be similar for other sig-levels.

the inference is, as usual, frequentist: The Null states that the population means are identical $H_0 : \mu = \mu_1$. The measured sample means differ by a certain amount, a distance $d = \bar{\mathbf{x}} - \bar{\mathbf{y}}$. The distance is converted to a test statistic (in this case t by dividing it by the appropriate standard error). We then ask how often (given the sample sizes and the variance estimated from the samples) a statistic as extreme as (or more extreme than) t would be expected if the Null were true. If such an event is sufficiently rare, we conclude that the Null is false or a rare event has occurred.

For example, as a socio-phonetician you might be interested in comparing the mean frequency of the F1 formant in articulations of self-described members of US coastal elites to the mean frequency of the same formant in articulations of self-described rural folk from the US Midwest in the long [iː] in *weakling*.[11] Some deep sociolinguistic theory predicts that the formant frequency should be lower for rural folk.

Per group, $n = 10$ participants read a short text where the word *weakling* occurs once. Hence, we have 20 data points in total. The results are (where $\mathbf{x}$ is from the elites and $\mathbf{y}$ from the rural folk):

$$\mathbf{x} = \langle 236, 236, 236, 244, 236, 229, 230, 249, 230, 251 \rangle$$
$$\mathbf{y} = \langle 238, 240, 227, 233, 230, 220, 242, 242, 234, 234 \rangle$$

The means are $\bar{\mathbf{x}} = 240$ and $\bar{\mathbf{y}} = 235$. The standard deviations are $s_{\mathbf{x}} = 9$ and $s_{\mathbf{y}} = 9$. Indeed, the formant frequency differs by 5 Hz (which is the distance d mentioned above). How can we apply a t-Test under these circumstances? What are the logic and the maths of the *Two-Sample t-Test*?

Two-Sample t-Test

First of all, what is the substantive hypothesis? It's that the true means between the DGPs producing the formant frequencies differ. More precisely, it's that $\mu > \mu_1$: The mean frequency in the coastal population is higher than in the rural population. In other words, by hypothesis there are two DGPs and thus two populations, one corresponding to coastal speakers (with the unknown true mean μ) and one corresponding to speakers from the Midwest (with the unknown true mean μ_1). Under the Null, the mean frequencies in the two populations are identical $H_0 : \mu = \mu_1$.[12]

[11] As usual in this book, the examples are completely made up. To experts in the respective subfields they might seem nonsensical. This approach stems from our conviction that the examples shouldn't matter much for a sound understanding of statistics. Whether it's mean formant frequencies measured in hertz or mean daily milk yields of cows measured in pints doesn't really matter.

[12] Logically, the Null is that the mean frequencies are identical *or the mean frequency in the rural*

This leads us nicely to the maths. The t-Test, like the z-Test, is based on subtracting a population mean from a sample mean. Remember Equation 5.2 for a single sample mean $\bar{\mathbf{x}}$ and a known true mean μ, repeated here for convenience:

$$t = \frac{\bar{\mathbf{x}} - \mu}{SE_{\mu}}$$

Mathematically, the population mean μ is a constant. It's a known and fixed parameter of the population in question. Furthermore, we assume that the sample means are normally distributed, and hence $\bar{\mathbf{x}} - \mu$ should also be normally distributed.[13] How can we extend this to a two-sample case? The t-Test is only applicable when we can subtract a *known population parameter* (mathematically a constant) from a normally distributed sample statistic. If we knew the *difference between the population means* (mathematically also a constant), we could compare it to (in other words, subtract it from) the difference of the sample means.[14] The corresponding equation is this:[15]

$$t = \frac{(\bar{\mathbf{x}} - \bar{\mathbf{y}}) - (\mu - \mu_1)}{SE_{\mu-\mu_1}} \tag{5.3}$$

As indicated above, Equation 5.3 might not seem to be very useful as we clearly haven't got any idea what the true means μ and μ_1 are. But this isn't a dead end. To see why, keep in mind that the test is always performed under the assumption that the Null is true. But what is $\mu - \mu_1$ under the Null? As both true means are identical under the Null, the difference between them is $\mu - \mu_1 = 0$. Hence:

$$t = \frac{(\bar{\mathbf{x}} - \bar{\mathbf{y}}) - (\mu - \mu_1)}{SE_{\mu-\mu_1}} = \frac{(\bar{\mathbf{x}} - \bar{\mathbf{y}}) - 0}{SE_{\mu-\mu_1}}$$

population is lower $H_0 : \mu \leq \mu_1$. For developing the maths of the t-statistic for two samples, we assume that the Null is just $H_0 : \mu = \mu_1$. This is because the maths only define the PDF (probability density function) of t-values if there is no effect. Whether we perform a single-sided test or a two-sided test is just a matter of calculating the correct area under the PDF (centre, left, or right).

[13] Notice that the means themselves are always normally distributed, even when we're performing a t-Test. Only the t-score is t-distributed because it adds an estimated variance to the calculation.

[14] As we'll demonstrate in Section 5.2.3, differences of normally distributed random variables are themselves normally distributed random variables.

[15] We'll return to the question of how to calculate the standard error $SE_{\mu-\mu_1}$ presently. For the moment, please believe us that there is an appropriate standard error for mean differences.

The final t-statistic for two samples is thus:

$$t = \frac{\bar{\mathbf{x}} - \bar{\mathbf{y}}}{SE_{\mu-\mu_1}} \tag{5.4}$$

There's still one open question, though. What is the appropriate standard error, which we referred to above simply as $SE_{\mu-\mu_1}$? There's also a minor catch: The refreshingly simple logic and maths described above only work as expected if the population variances σ and σ_1 are equal. However, if that is the case *and the two samples have the same size*, the estimated standard error for the mean differences $SE_{\bar{\mathbf{x}}-\bar{\mathbf{y}}}$ is calculated by just adding the variances of the two samples. Why do we just add the two? The variance of the difference score $\bar{\mathbf{x}} - \bar{\mathbf{y}}$ obviously comes from *two* means (estimated from two samples). We subtract the sample means in order to make an inference about a difference between two potential population means. But as both estimates contribute their own amount of uncertainty to the calculation, we have two sources of error. Hence, the standard error adds one error component for each sample (see also Section 5.2.3). In the One-Sample t-Test, there is only one source of error and therefore only one error component.

$$SE_{\bar{\mathbf{x}}-\bar{\mathbf{y}}} = \sqrt{\frac{s_{\mathbf{x}}^2}{n_{\mathbf{x}}} + \frac{s_{\mathbf{y}}^2}{n_{\mathbf{y}}}} \tag{5.5}$$

With this standard error, we can calculate the t-statistic using Equation 5.4. While the sample means and their difference are normally distributed, the unreliability of the standard error with low samples sizes means that the t-statistic is t-distributed and not normal. This unreliability gets lower when the total number of data points from either sample or from both samples gets larger. To account for his, the degrees of freedom for sample differences $\nu_{\bar{\mathbf{x}}-\bar{\mathbf{y}}}$ are simply the added degrees of freedom of the individual samples:

$$\nu_{\bar{\mathbf{x}}-\bar{\mathbf{y}}} = \nu_{\bar{\mathbf{x}}} + \nu_{\bar{\mathbf{y}}} = (n_{\mathbf{x}} - 1) + (n_{\mathbf{y}} - 1) \tag{5.6}$$

5.2.2.2 A Sample Calculation

Let's apply all of this to our example:

$$SE_{\bar{\mathbf{x}}-\bar{\mathbf{y}}} = \sqrt{\frac{81}{10} + \frac{81}{10}} = \sqrt{8.1 + 8.1} = 3.32$$

and:

$$\nu_{\bar{x}-\bar{y}} = (10-1) + (10-1) = 18$$

The t-statistic is calculated as:

$$t = \frac{240 - 235}{3.32} = \frac{5}{3.32} = 1.51$$

By looking up the statistic $t = 1.51$ in a table for the cumulative density of the t-Distribution ($\nu = 18$, single-sided lookup), we get $p = 0.07$.[16] This doesn't even meet the lowest standard for a unexpected result under the Null ($sig = 0.05$), and hence we do not conclude anything from this result.

5.2.2.3 Standard Errors with Unequal Sample Sizes

We pointed out that the regular t-Test and its standard error calculated as in Equation 5.5 are only appropriate if both (potential) populations or DGPs have identical variances, and if the samples have the same size. If the variances differ (even potentially), please refer to *Welch's t-Test*, a variant of the t-Test with modified calculations to account for unequal variances. We don't cover Welch's Test in this book for reasons of space and focus. However, if the two samples are of different sizes but the population variances can be safely assumed to be equal there's a relatively simple solution. Let's first consider why different sample sizes are a problem at all for calculating the standard error.

Welch's t-Test

We know by now that larger samples represent (on average) any population (or DGP) better than smaller samples. It is never the case that smaller samples lead to better estimates of any population parameter (such as mean and variance) than larger samples. Remember that the two populations have to have identical variances for the Two-Sample t-Test to be applicable. If there is a considerable difference in sample size, one sample—the larger one—is thus a better approximation of the common variance $\sigma^2 = \sigma_1^2$ of μ and μ_1. By simply adding error components based in the individual samples' variance estimates, both are given the same weight in the calculation of the total error inherent in the estimation of the mean difference. This cannot be the optimal solution, because smaller samples will inevitable drag down larger samples, at least on average.

At the same time, keep in mind that the amount of error contributed by the two samples depends on their respective sample sizes. Our strategy to deal with these facts is as follows: First, we calculate the best possible estimate of the common variance $\sigma^2 = \sigma_1^2$ using information from *both* samples. Then, we use this

[16] Convenient tables can be found in the appendices.

variance estimate to introduce two error components into the standard error, which depend on their respective sample sizes. The best possible estimate of the identical population variances $\sigma^2 = \sigma_1^2$ is the so-called *pooled variance* $s^2_{\mathbf{x},\mathbf{y}}$ of the samples, given by:

pooled variance

$$s^2_{\mathbf{x},\mathbf{y}} = \frac{SQ_{\mathbf{x}} + SQ_{\mathbf{y}}}{(n_{\mathbf{x}} - 1) + (n_{\mathbf{x}} - 1)} \tag{5.7}$$

SQ stands for *sum of squares*. Please revise Section 3.2.3 if you don't remember exactly what this means. For comparison, the variance $s^2_{\mathbf{x}}$ of a single sample $\mathbf{x}$ was given in Equation 3.4, repeated here in condensed form for convenience:

$$s^2_{\mathbf{x}} = \frac{SQ_{\mathbf{x}}}{n - 1}$$

The pooled variance is what the name suggests: We pool both samples to calculate a maximally precise estimate of the population variance. If one sample is smaller, its sum of squares is smaller, but so is its sample size. Hence, Equation 5.7 intrinsically weighs the samples' contribution to the common variance. With $s^2_{\mathbf{x},\mathbf{y}}$ (the variance estimated from both samples), the standard error is given as in Equation 5.8. Notice that we couldn't simply pool the samples at this stage, as each error component depends specifically on the size of one of the two samples.[17]

$$SE_{\bar{\mathbf{x}}-\bar{\mathbf{y}}} = \sqrt{\frac{s^2_{\mathbf{x},\mathbf{y}}}{n_{\mathbf{x}}} + \frac{s^2_{\mathbf{x},\mathbf{y}}}{n_{\mathbf{y}}}} \tag{5.8}$$

Figure 5.13 presents a plot of simulated estimates of standard errors from two samples. In each of the 100 replications, one sample of size 5 and one sample of size 20 were drawn from a DGP with $s = 5$ ($s^2 = 25$). We know (as it's a simulation) that the true standard error is:

$$SE_{\mu-\mu_1} = \sqrt{\frac{25}{5} + \frac{25}{20}} = 2.5$$

[17] For further enlightenment: Because the variance must be identical between the two populations (even if the means differ), we can fully *pool* the samples in the calculation of the variance. However, the error components in the calculation of the standard error are *not* identical if the samples have different sizes. Hence, there can't be full pooling.

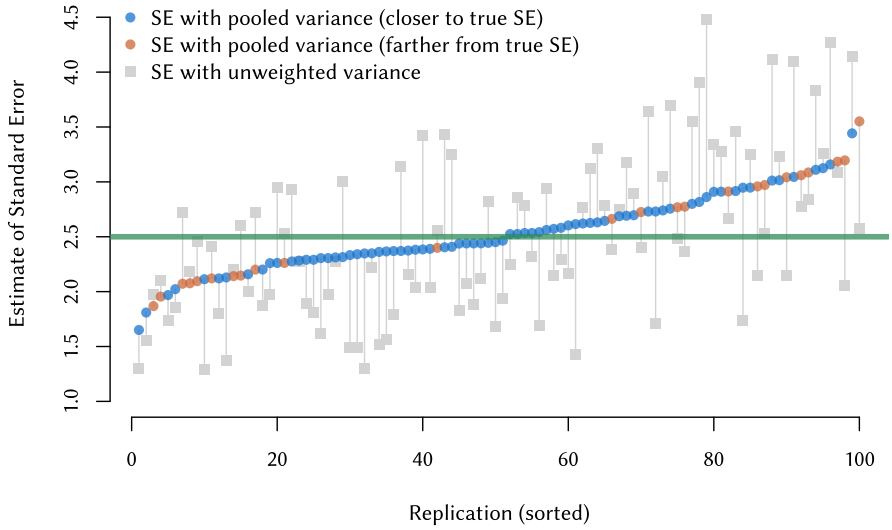

Figure 5.13: 100 simulations of standard errors for mean differences: calculated with simply added variances (gray squares) and calculated with pooled variances (blue dots and orange dots); sample sizes are 5 and 20; the true SE is 2.5 (green line)

In each replication, we calculated the standard error both *incorrectly* by just adding the error components with individually computed variances (as if the samples were of the same size, Equation 5.5) and *correctly* using the pooled variance (Equation 5.8).[18] The standard error using the pooled variance is closer to the true standard error in 76% of the replications. This illustrates that with the pooled variance, the (on average) less accurate variance estimate from the smaller sample gets fewer chances to drag down the precision of the (on average) more precise variance estimate from the larger sample.

5.2.2.4 A Sample Calculation

Returning to our fictitious experiment on fine phonetic details, let's assume that the last two data points from the elite sample **x** had to be removed because it be-

[18] To make the plot less cluttered, the replications were sorted by the standard error calculated from the pooled variance.

came known after the experiment that the two speakers were actually Canadian. The updated sample size is $n = 8$, the mean is $\bar{\mathbf{x}} = 237$, the sample variance is $s^2_{\mathbf{x}} = 44.29$ ($s_{\mathbf{x}} = 6.65$). The pooled variance is calculated like so:

$$s^2_{\mathbf{x},\mathbf{y}} = \frac{310 + 452}{9 + 10} = 47.62$$

The updated standard error is:

$$SE_{\bar{\mathbf{x}}-\bar{\mathbf{y}}} = \sqrt{\frac{47.62}{8} + \frac{47.62}{10}} = \sqrt{5.95 + 4.76} = 3.27$$

and:

$$\nu_{\bar{\mathbf{x}}-\bar{\mathbf{y}}} = (8 - 1) + (10 - 1) = 16$$

The t-statistic is calculated as:

$$t = \frac{237 - 235}{3.27} = \frac{2}{3.27} = 0.61$$

By looking up the statistic $t = 0.61$ in a table for the cumulative density of the t-Distribution ($\nu = 16$, single-sided lookup), we get $p = 0.28$. Without the data points from the two Canadians the mean distance has shrunken considerably. Also, the standard error went up slightly, leading to a result that is not unexpected at all under the Null. Please keep in mind that the high p-value still is *not* in any way *evidence in favour of the Null*! Under the Null, all p-values have the same probability.

5.2.2.5 Conditions for Testing Mean Differences

We hope that we could convince you that these tests are (as most people think) quite basic, but that there are lots of small catches and details to take into account. While the Two-Sample t-Test is applicable to many more situations than the z-Test, we'd like to point out that it's not a test you can apply to any sample means whatsoever. Table 5.2 summarises the conditions that must be met for the tests discussed in this chapter to be applicable. If they are not met, inferences will be marred: The error rates that you hope to control will *not* be at the level you think they are. Please revise Section 5.2.1 (especially Figure 5.12) in order to make absolutely sure that you understand this.

Table 5.2: Necessary conditions for z-Tests and t-Tests

		One-Sample	Two-Sample
	z-Test	t-Test	
The data are numerical.		①	
The means are normally distributed.		②	
The data are normally distributed.		③	
The measurements are independent.		④	
The population mean is known.		⑤	
The population variance is known.	⑥		
The population variances are equal.			⑦
The samples are independent.			⑧

Some of these requirements have not been mentioned very prominently. So, let's go through them.

① **The data are numerical.** For any of the tests of mean differences (such as z-Tests and t-Tests) the mean has to be computable, which is the case only for numerical measurements. In principle, this is all there is to say. However, we'd like to tell a story: We know of a paper in linguistics (which we gracefully refuse to reference) where a theoretical syntactician and an empiricist demonstrated something something very convenient for experiments where acceptability ratings are gathered.[19] The authors showed that instead of numerical ratings of sentence acceptability (such as split-100 ratings or magnitude estimation) one could also use binary ratings (acceptable/not acceptable). Calculating a t-Test on the numerical ratings led to the same conclusion as calculating a t-Test on proportions. The idea was something along the following lines: When participants rate a sentence as 60% acceptable in the numerical measurement, 60% of the participants in the binary measurement rated the sentence as fully acceptable. So, when you have 10 sentences for condition A and 10 sentences for condition B, then you can take the mean of the 10 proportions of *acceptable* ratings for condition A and B and perform a t-Test. Mind-boggling! Besides the legions of devils that lie in the details with this approach, consider this: They tried this out in a single instance, and the

[19] For non-linguists, let us explain: Linguists sometimes ask naive speakers to rate sentences for their acceptability. Participants are presented with sentences as stimuli, and they're asked to press one button if the sentence sounds *acceptable* to them in their native language and another button if it doesn't, which results in a binary measurement. Alternatively, they are asked to rate sentences on an absolute or relative scale, which results in (something like) numerical measurements.

inferences were the same for the mishandled binary ratings as for the numerical ratings. You're familiar enough with frequentist thinking to know that you can't test whether you can break the rules of a test by trying it out *once*. The true error rate can only be found in the limit, which requires either mathematical proof or a long—very long—series of experiments. Please revise Figure 5.12. Frequentism is all about the control of error rates, and you need to obey the rules (meet the conditions of your tests) to ensure that your error rates are what you think they are. *We tried it out once, and it worked!* isn't acceptable.

② **The means are normally distributed.** The normality of means underlies our entire logic of inferences about them. It's a typically frequentist condition, and it's quite difficult to check whether it holds in a specific instance. We usually have only a single sample (or at least not more than a few) from which we calculate the mean. It's impossible to check whether this one sample comes from a normally distributed population of means, but at the same time the frequentist machinery requires it to be normal. There is no easy solution, but there is one, which we'll discuss immediately.

③ **The data are normally distributed**. Technically speaking, the data themselves (reaction times, formant frequencies, sentence lengths, etc.) don't need to be normally distributed. Only the means themselves must fulfil this requirement. Interestingly, there is a fundamental result (which at first appears very counterintuitive to many) which roughly speaking states that sample means drawn from any underlying distribution will be normal at sufficiently large sample sizes. This fundamental result is called the *Central Limit Theorem* (CLT). If you watched the videos recommended in Section 4.4 you've already got a basic idea of what it is. In any case, if samples are large enough, the data themselves need not be normally distributed, but the mean will be. Unfortunately, as nobody knows in advance what the distribution of measurements in your data sets is going to be, nobody can tell you in advance what is *large enough*. The safest way to meet the condition of normality of means is having normal data, which entails that means will always be normal, regardless of the sample size. There are specific tests to check whether data are normally distributed, but we believe they do not land anywhere near the perfect spot between rejecting too many data sets as non-normal and accepting too many as normal. Try a raw plot, a histogram, or a density plot to get a rough idea if your data's normality.

Central Limit Theorem

In linguistics—and especially in corpus linguistics—, many data are known to not be normally distributed. Because many types of data from corpora are counts or frequencies, their distributions are discrete anyway, but they're often skewed or diverge from the Normal Distribution in other ways, too. If you look at sentence lengths measured in words, for example, it could never be truly normal be-

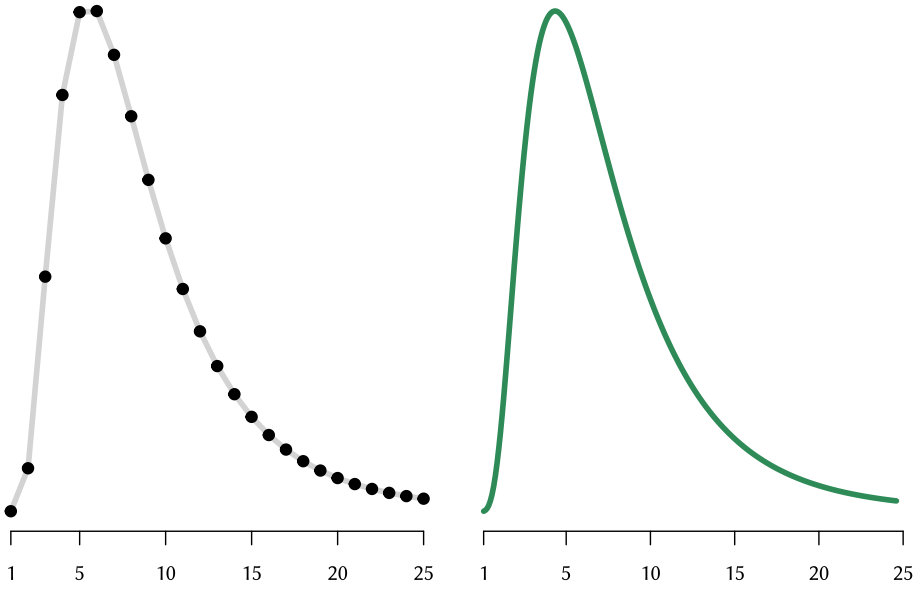

Figure 5.14: Left: idealised discrete distribution of sentence lengths measured in words; Right: a continuous Lognormal Distribution corresponding to the left panel

cause sentence lengths are discrete. A sentence can consist of 1, 2, 3, ... words, but not 3.812 words, for example. Additionally, the distribution of sentence lengths usually looks approximately as in the left panel of Figure 5.14.[20]

According to Figure 5.14, there are very few very short sentences (of 1 and 2 words), but there are a lot of sentences between roughly 3 and 10 words long. For longer sentence lengths, the frequency decreases sharply. The distribution clearly isn't normal, but it looks similar to the continuous *Log-Normal Distribution* (maybe with some scaling and offsetting). The Log-Normal Distribution has the welcome property of being defined for positive values only, which is plausible for things like sentence lengths, as sentences cannot be 0 or −1 words long. A density function of the corresponding Log-Normal Distribution is shown in the right panel of Figure 5.14. Interestingly, the Log-Normal Distribution is called

Log-Normal Distribution

[20] Many experts will probably scream at us because distribution of sentence lengths varies greatly by factors such as style, and the distribution as shown is not a good model of sentence lengths, etc. This is all true, but for the purpose of a mere illustration of basic differences to the Normal Distribution, Figure 5.14 is quite accurate.

Log-Normal because it turns into a Normal Distribution if you logarithmise the values.

Hence, in order to ensure that your sample means are normal, there are at least two strategies. You can logarithmise the underlying values and thus indirectly force the means to be normal. Or you can increase the sample size and rely on the CLT to ensure that the sample means are normally distributed. Increasing the sample size has some side-effects to be taken into account (see Chapter 7), but it's relatively unproblematic if you're aware of the side-effects. Logarithmising the values is an instance of a *data transformation*. If the data don't have the properties they should have, a transformation is any mathematical operation that brings the data closer to the expected distribution in order to make correct frequentist inferences possible. There is a considerable downside, however. Assume you want to show that mean sentence lengths in forewords to English novels from two decades differ. Because the data aren't normal and you've got a limited amount of data, you logarithmise them and make an inference, rejecting the Null which says that the means are equal. You also observe that the mean difference is 0.04. But what does this mean? What do logarithmised sentence lengths mean within your theory, and how do you interpret the result? You've not made an inference about sentence lengths but about logarithmised sentence lengths. Such questions should be considered with great care, and they are theoretical questions, not mathematical ones.

data transformation

④ **The measurements are independent.** We'll keep this one short as we'll return to it later in the book anyway. For inferences to be valid, each measurement in your sample(s) must come from a random event *independent* of all the other events (see Section 2.2.3). This condition is violated very easily. However, it's (once again) statistics can't decide whether it's violated or not. It's your job when you design an experiment and analyse the data.

Many situations can potentially lead to non-independence of measurements: asking the same person more than once for an acceptability rating; taking measurements in a temporal sequence where one event influences the probabilities in the subsequent event(s); using the same verb more than once in different stimulus sentences; using corpus data from texts written in two very different styles of writing without accounting for those stylistic differences. Each of these situations introduces *grouping* into the data. You get groups of ratings per participant; chains of measurements where the previous one affects the subsequent ones; groups defined by lexical items; groups defined by styles of writing. We'll discuss advanced techniques of taking care of grouping (but not temporal sequences) in Chapter 12. However, in simple tests like the t-Test, such grouping can lead to a situation where the distribution of variance in the data is more

grouping

complex than the test assumes, which (again) means that your actual error rates might be off from the nominal error rates. For example, the Two-Sample t-Test is applicable if the variance between the two groups is homogeneous and equal, and if the only difference between the groups is the mean difference, which is causally related to the substantive distinction the researcher is interested in.
⑤ **The population mean is known.** This is straightforward. The z-Test and the One-Sample t-Test require that the population mean is known beyond doubt. If it's an estimate, you're introducing variability which is not accounted for in the maths, and your error rates will be off. There's no way to check for this. You—the researcher—need to be aware of what you're doing.
⑥ **The population variance is known.** Again, this is a very straightforward matter. Go back to Figure 5.12 to see what happens if you pretend you know the population variance (and perform a z-Test) when in fact you just estimated it (and should perform a t-Test).
⑦ **The population variances are equal.** For the Two-Sample t-Test, the variances in the two populations must be equal beyond doubt. If in doubt, use Welch's variant of the t-Test.
⑧ **The samples are independent.** This is a variation on the theme from ④. In a Two-Sample t-Test, there must be no groupings between data points from one group and the other. For example, if you measure the same participants' reactions under two conditions, there is a pairwise grouping between reactions from individual participants under condition A and condition B. If you have data that is structured like this, use the t-Test for Repeated Measures. It's just a small amendment, but it will greatly improve your inferences.

5.2.3 In-Depth Mean Differences Are Normally Distributed

In Section 5.2.2.1, we assumed that subtracting two normally distributed random variables results in another normally distributed random variable. In this section, we illustrate this fundamental result further. The claim was that if $\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$ are normally distributed, then $\bar{\mathbf{x}} - \bar{\mathbf{y}}$ is also normally distributed. We show this in Figure 5.15. The left panel provides a density plot of 10000 random draws (extremely simple simulations) of two normally distributed random variables: $\mathbf{x}$ with $\mu = 3$ and $\mathbf{y}$ with $\mu_1 = 1$, both with $\sigma = \sigma_1 = 1$.

Also, we plotted $\mathbf{x} - \mathbf{y}$, a subtraction of the randomly paired values from $\mathbf{x}$ and $\mathbf{y}$. As expected, the mean is $\mu - \mu_1 = 3 - 1 = 2$. Furthermore, as the subtraction of the two random variables introduces *two* sources of uncertainty, the variance of $\mathbf{x} - \mathbf{y}$ is larger than that of the two independent variables. The standard deviation

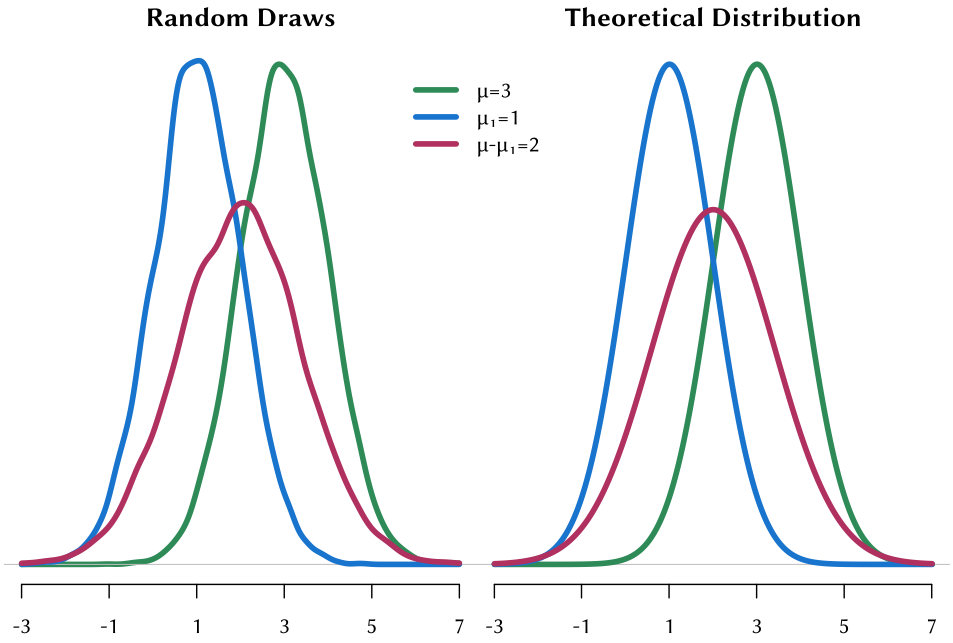

Figure 5.15: Left: Density plots of 10000 random draws each from a Normal Distribution with mean μ=3 and σ=1 (blue), from another Normal Distribution with mean μ_1=1 and σ_1=1 (blue) and the subtraction of random pairs from the second from the first; Right: the theoretical distributions corresponding to the left panel

is around 1.4.[21] This directly illustrates what we said in Section 5.2.2.1 about adding two error components to calculate the standard error for mean differences (Equations 5.5 and 5.8).

To show how simulations and random draws converge with theoretically known facts, we added the right panel of Figure 5.15. It's the theoretical expectation for the random draws shown in the left panel. It's a theoretical *expectation* inasmuch as we can safely expect the left panel to look more and more like the right panel as the sample sizes approach infinity. And on that bombshell, we move on to the next chapter.

[21]How to get to this value analytically is a more complex matter which is better reserved for textbooks focussing on the mathematical side of things.

Exercises for Chapter 5

(1) **x** as given below is a series of numeric measurements. The known population variance is σ =. Under the substantive hypothesis that the mean is smaller than the know population mean μ =, formulate a Null and perform a z-Test. All required statistics should be calculated by hand with a pocket calculator. Interpret the result, and make an inference if possible.

(2) **x** as given below is a series of numeric measurements. The known population variance is σ =. Under the substantive hypothesis that the mean is greater than the know population mean μ =, formulate a Null and perform a z-Test. All required statistics should be calculated by hand with a pocket calculator. Interpret the result, and make an inference if possible.

(3) **x** as given below is a series of numeric measurements. The known population variance is σ =. Under the substantive hypothesis that the mean is different from the know population mean μ =, formulate a Null and perform a z-Test. All required statistics should be calculated by hand with a pocket calculator. Interpret the result, and make an inference if possible.

(4) Assuming the population variance is unknown, perform a t-Test corresponding to Exercise 5.2.3. Interpret the differences in the results between the z-Test and the t-Test.

(5) For the samples **x** and **y** given below, perform a t-Test for differences in sample means under the substantive hypothesis that $\bar{\mathbf{y}}$ is smaller than $\bar{\mathbf{x}}$. Formulate a Null and perform the test. All required statistics should be calculated by hand with a pocket calculator. Interpret the result, and make an inference if possible.

(6) For the samples **x** and **y** given below, perform a t-Test for differences in sample means under the substantive hypothesis that $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are not equal. Formulate a Null and perform the test. All required statistics should be calculated by hand with a pocket calculator. Interpret the result, and make an inference if possible.

(7) We have argued in Section 5.1.4 that larger effect sizes (increasing actual differences between the known population mean and the measured means) tilt the distribution of p-values increasingly towards 0. There is a tradition where researchers don't look at the concrete p-value but just perform a mechanical significance test at a pre-set sig-level such as $sig = 0.05$. This means that tests where $p \leq sig$ are called *significant*, all others *not significant*. Under such an approach, do you get in the long run: (i) more significant results, (ii) less significant results, or (iii) a constant number of significant results when the actual effect strength increases (all other things being equal)? Still under this approach, is your *error rate* constant or does it change with increased effect strength? The error rate is the percentage of samples which make you erroneously reject the Null when it is true.

(8) A fictitious big data study in sociolinguistics has found through a popular social media app (and only from users that consented to be part of the experiment) that the mean time it took self-identified men to read a specific message about tomato ketchup was 8.9 seconds ($n_x = 45231$, $s_x^2 = 0.3$), whereas the mean reading time for self-identified women was 9.2 seconds ($n_y = 21231$, $s_y^2 = 0.4$). The theory-driven substantive hypothesis was that men should read the message faster. Formulate the Null and perform a t-Test for means from two samples. Calculate all required statistics using a pocket calculator, including the appropriate variance. Interpret and evaluate the result.

(9) The co-supervisor of the dissertation that led to the study described in Exercise 8 reads the report and quickly writes an email to the author, cc'ing the supervisor. They say that the study is invalid anyway because the sample wasn't *balanced*. It should have contained an equal number of data points from men and from women in order to be *properly representative of all relevant groups in the study*. What do you – the main supervisor – reply in order to defend your candidate? (We experienced this exact situation at some university some time ago, except in a different sub-field of linguistics and with real data.)

(10) The fictitious study on formants in Section 5.2.2 didn't produce surprising (i. e., *significant*) results, and the authors point out in the publica-

tion that the experiment did not lend support to their substantive hypothesis, discussing potential reasons for the failure as well as possible future experiments. The publication was pre-registered (see Chapter 1) and was published despite reporting only negative results. Swiftly, a proponent of a rival theory submits a short reply to the *Bulletin of the Sociolinguistic Society*. The rival theory explicitly predicts that rural folk should *not* produce higher F1 frequencies in the word *weakling*. The author of the short reply argues that the failed experiment indirectly counts as evidence for the rival theory and demands that the original paper be retracted because the theory that motivated the reported experiment is clearly inferior, maybe even utterly false. You are known as an expert on statistics and asked to review the short reply. Do you recommend to accept it or to rejected it? What do you think about the demand that the original paper be retracted?

6 Inference: Three or More Means

6.0.1 In-Depth The F Distribution

$$\sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}} \tag{6.1}$$

$$\frac{1}{x \, \mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \tag{6.2}$$

7 Inference: Making Positive Inferences

8 Inference: This and That

$$\binom{1000}{100} \approx 6.385 \cdot 10^{139}$$

$$\binom{1000}{10} \approx 2.634 \cdot 10^{23}$$

8.0.1 In-Depth The Binomial Distribution

$$\binom{n}{k} p^k q^{n-k} \tag{8.1}$$

8.0.2 In-Depth The χ^2 Distribution

$$x^{\frac{k}{2-1}} e^{-\frac{x}{1}} \tag{8.2}$$

$$\frac{1}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})} \tag{8.3}$$

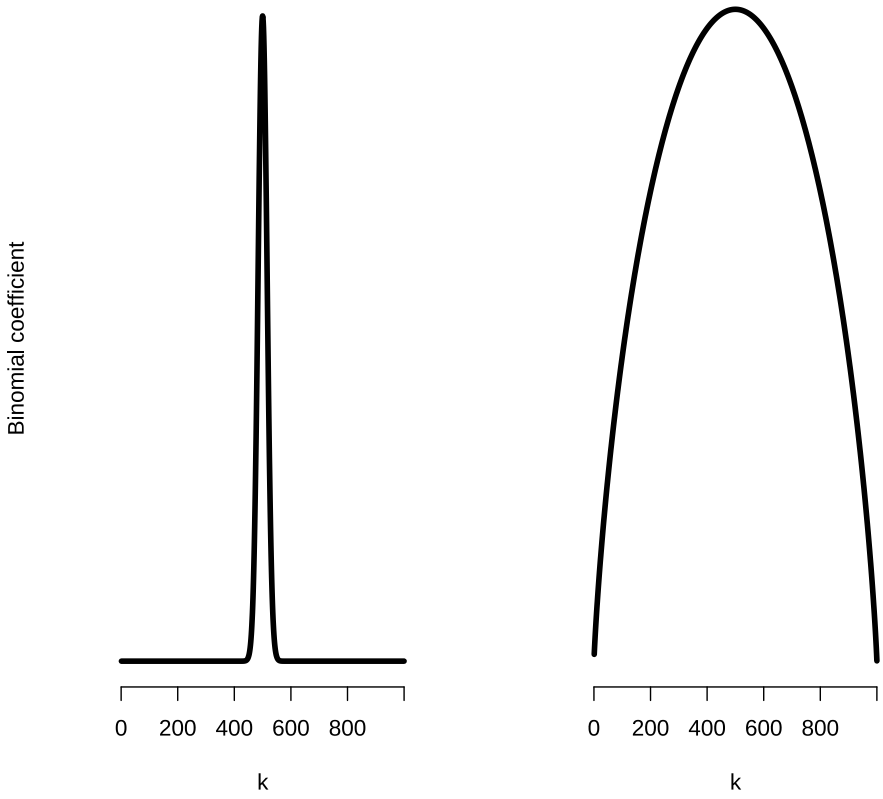

Figure 8.1: Number of ways of choosing k elements from $n = 1000$ elements; Left: regular y-axis, Right: logarithmic y-axis

9 Data: Co-Varying Variables

10 Modelling: Linear Relationships

10.1 In-Depth The Equivalence of the ANOVA and the Linear Model

11 Modelling: Arbitrary Outcomes

12 Modelling: Grouped Data

Appendix

Greek Alphabet

The pronunciation is given as British / American in case of substantial differences. The pronunciations follow the Wikipedia entries for the respective letters. Our preferred pronunciation is highlighted.

UC	LC	Name	Pronunciation
Α	α	Alpha	**ɑlfə** / ˈælfə
Β	β	Beta	ˈ**biːtə** / ˈbeɪtə
Γ	γ	Gamma	ˈ**gæmə**
Δ	δ	Delta	ˈ**dɛltə**
Ε	ε	Epsilon	**ɛp**ˈ**saɪlən** / ˈɛpsɪlɒn
Ζ	ζ	Zeta	ˈ**ziːtə** / ˈzeɪtə
Η	η	Eta	ˈ**iːtə**
Θ	θ	Theta	ˈ**θiːtə** / ˈθeɪtə
Ι	ι	Iota	**aɪ**ˈ**oʊtə**
Κ	κ	Kappa	ˈ**kæpə**
Λ	λ	Lambda	ˈ**læmdə**
Μ	μ	Mu	ˈ**mjuː**
Ν	ν	Nu	ˈ**njuː**
Ξ	ξ	Xi	**zaɪ**
Ο	ο	Omicron	**oʊ**ˈ**maɪkrɒn** / ˈoʊmɪkrɒn, ˈɒmɪkrɒn
Π	π	Pi	ˈ**paɪ**
Ρ	ρ	Rho	ˈ**roʊ**
Σ	σ, ς	Sigma	ˈ**sɪgmə**
Τ	τ	Tau	ˈ**taʊ**
Υ	υ	Upsilon	**ʊp**ˈ**saɪlɒn**, (j)uːpˈsaɪlən / ˈʌpsɪˌlɒn, ˈ(j)uːpsɪˌlən
Φ	φ	Phi	**faɪ**
Χ	χ	Chi	**kaɪ**
Ψ	ψ	Psi	**saɪ**
Ω	ω	Omega	ˈ**oʊmɪgə** / oʊˈmeɪgə, oʊˈmɛgə, oʊˈmiːgə

Symbols used in this book

	Name	Definition and Reference
n	*sample size*	Number of observations/measurements in a sample
N	*population size*	Mostly hypothetical number of objects in a population
p	*p-value*	Frequentist pre-experiment probability of an obtained result under the Null
q	*proportion*	Proportion of measurements with a specific value within a sample or population
s	*variance*	Mean deviation of measurements from the mean in a sample
s^2	*standard deviation*	Mean squared deviation of measurements from the mean in a sample
$\mathbf{x}$	*sample*	A sample of measurements as defined in the text in the form of a tuple (also called vector or array)
$\bar{\mathbf{x}}$	*mean*	The (arithmetic) mean of the values in a sample $\mathbf{x}$
$\tilde{\mathbf{x}}$	*median*	The middle value of a sorted sample $\mathbf{x}$
$\dot{\mathbf{x}}$	*mode*	The most frequent value in $\mathbf{x}$(binary, nominal, ordinal) or the maximum of a distribution (numeric)
$\mathbf{y}$	*sample*	See $\mathbf{x}$, also for modifiers such as $\bar{\mathbf{y}}$, $\tilde{\mathbf{y}}$, and $\dot{\mathbf{y}}$
$\mathbf{z}$	*sample*	See $\mathbf{x}$, also for modifiers such as $\bar{\mathbf{z}}$, $\tilde{\mathbf{z}}$, and $\dot{\mathbf{z}}$
μ	*mean*	The same as $\bar{\mathbf{x}}$, except for a population
σ	*standard deviation*	The same as s but for populations
σ^2	*variance*	The same as s^2 but for populations

References

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4). 243–257.

Carsey, Thomas M. & Jeffrey J. Harden. 2014. *Monte Carlo simulation and resampling methods for social science*. Thousand Oaks: Sage Publications.

Eisenberg, Peter. 2020. *Grundriss der deutschen Grammatik: Das Wort*. 5th edn. Stuttgart: Metzler.

Evert, Stefan. 2008. Corpora and collocations. In Anke Lüdeling & Maria Kytö (eds.), *Corpus linguistics: An international handbook*, vol. 2, 1212–1248. Berlin: Mouton.

Fisher, Ronald A. 1935. *The design of experiments*. London: Macmillan.

Gries, Stefan Th. 2014. Frequencies, probabilities, and association measures in usage-/exemplar-based linguistics – some necessary clarifications. In Nikolas B. Gisborne & Willem Hollmann (eds.), *Theory and data in cognitive linguistics*, 15–48. Amsterdam: Benjamins.

Gries, Stefan Th. 2015. More (old and new) misunderstandings of collostructional analysis: On Schmid and Küchenhoff (2013). *Cognitive Linguistics* 26(3). 505–536.

Gries, Stefan Th. 2022. What do (some of) our association measures measure (most)? association? *Journal of Second Language Studies* 5(1). 1–33.

Küchenhoff, Helmut & Hans-Jörg Schmid. 2015. Reply to "More (old and new) misunderstandings of collostructional analysis: On Schmid & Küchenhoff" by Stefan Th. Gries. *Cognitive Linguistics* 26(3). 537–547.

Schmid, Hans-Jörg & Helmut Küchenhoff. 2013. Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics* 24(3). 531–577.

Senn, Stepen J. 2011. You may believe you are a Bayesian but you are probably wrong. *Rationality, Markets, and Morals* 2. 48–66.

Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243.

Vasishth, Shravan & Michael Broe. 2011. *The foundations of statistics: A simulation-based approach.* Berlin: Springer.

Statistical Inference for Everybody (and a Linguist)

This book is good.