Statistical Inference for Everybody and a Linguist

It wasn't me!

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

It wasn't me!

language
science
press

It wasn't me! 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

Freie Universität Berlin

Contents

Preface

1 The Need for A New Introduction to Statistics

2 Methods and the Progression of Topics

Readers should be aware that Chapter 2 progresses very slowly, is very repetitive, and contains a lot of detail and material for illustrative purposes. This changes in later chapters, and we speed up the narrative significantly. However, a slow start is clearly desirable given what we said in Section 1.

Acknowledgments

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 Guessing and Counting

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 unergative 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, (ii) the psychic isn't actually a psychic, (iii) *sleep* cannot be passivised. 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 you saw no chance of winning the lottery, you wouldn't play it. If you absolutely knew for certain that your psychic friend couldn't guess phone numbers, you wouldn't humiliate them and ask them to guess numbers at your party, except maybe if there were others who didn't know for certain that the psychic didn't have the ability in question. 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 which 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 gray, 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 gray steaming cup, the gray 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 cup and has the other 7 cups still available. Had she chosen the red steaming cup

[3] We switch to writing numerals using arabic numbers for clarity.

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 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, she now makes her first incorrect guess. She picks the green non-steaming one, which is a tea-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.

First, she chose the blue steaming cup, then the gray steaming cup, then the red steaming cup, and finally the the green non-steaming cup. But is this really

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

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 gray 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 capability), 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 unexpected result? So unexpected maybe that you'd doubt that Muriel Bristow merely got lucky? Well, you tell me!

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.

2.2.3 Interlude: 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 reflect on the notion of *probability*. Probability Theory is a complex field comprising complex formal aspects (the axiomatisation of the maths of probability) as well as philosophical aspects (the interpretation of the maths). There is neither just one unique axiomatisation nor one unique interpretation of the concept. We present one view that is not very technical but useful for practitioners who want to use statistical inference in the analysis of empirical data.

We only contemplate situations that are **experiments** or that resemble experiments (like a lottery). An real and well-designed experiment is characterised by the fact that we know in advance what might happen. We just calculated the number of possible outcomes in a classical Tea Tasting Experiment (70), and we could list all 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 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, there is always a sequence of 4 events of choosing cups. If there aren't 4 such events, the experiment wasn't actually conducted properly—it failed. In a proper execution of the experiment, each of these events (such as choosing the blue tea-first cup in the first step) has a given **probability**. The probability of the event is just what we introduced in the previous section: There are 8 possible events available for the first actual event, and none of these events has any special preference attached to it (if the person choosing the cups is just guessing). Hence, if we repeated the first step over and over again, each cup would be selected equally often. The notion of probability advocated in this book draws its interpretation from this concept of a hypothetically infinite sequence of repeated events. While it's hypothetical, it's very much useful and related to everyday experience. It's the reason why you can't rely on a fair die to win any money in the long run. There is no way to predict the number of pips, 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.

probability

While the probability of each event doesn't have to be equal as in the case of

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 the example.

To do that, we first introduce a convenient general notation 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 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 (running the experiment over and over again), even a random guesser would classify 3 cups correctly with a probability 0, 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:

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

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

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

Again, 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 then 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. 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 only has a probability of 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 capability 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 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 exper-*

significance threshold

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

The purpose of this section is to reconstruct and critique such inferences.[7] 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 just about the interpretation of the resulting number.

There are two important aspects with respect to making inferences from unexpected outcomes. Can we be *sure* that Muriel Bristow had the capability 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, Carl. Carl 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. Hypotheses about why ESP should be real simply don't exist. Also, notice that we did not, in fact, conduct an experiment about my grandfather's ESP capabilities, but I merely told a story about him winning the lottery. At some point I just 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 something,

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

you're bound to find it somewhere, and a low p-value becomes utterly meaningless. Interestingly, Carl 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 ever 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 with a substantive inference from an experiment. 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 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 will 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 often was a correct cup of tea chosen relative to the maximal number of events of choosing a cup of tea in the experiment? 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 *even more extreme result*

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

variables

A contingency table tabulates counts of events characterised by two **variables**, each having two or more discrete possible values. Here, one variable (called *Reality* in Table 2.1) characterises the event as involving a real tea-first or a milk-first cup, and the other variable (called *Bristow* in Table 2.1) characterises the event as involving a cup 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 events involving tea-first cups according to Muriel Bristow. In row 2, we put the 4 events involving milk-first cups as assigned by her. The row sums in the last column reflect the fact that there were indeed 4 events (and cups) for each condition. In the two columns, we count the event involving real tea-first and milk-first cups. As you can see, of the three real tea-first cups, Muriel Bristow classified 3 as tea-first (cell 1,1) and 1 as milk-first (cell 2,1).[8] The opposite is true

[8] 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.

for the real 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 either the rows or the columns and the corresponding sums. In Table 2.2, the cells shaded in blue and the cells shaded in green suffice to calculate the probability of drawing 3 from 4 and independently 1 from 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

	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.[9] 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

[9]Be warned that people actually use the test like this, but that this is a slightly incorrect use. There will be ample discussion of this caveat throughout the book.

analysis.[10] Assume that we drew 90 active sentences containing *sleep* in addition to the 10 passives. Furthermore, assume that the corpus contains 1,100 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:

$$\frac{\binom{890}{90}\binom{210}{10}}{\binom{1,100}{100}} \approx \frac{6.573 \cdot 10^{141}}{1.423 \cdot 10^{144}} \approx 0.005$$

This probability is useful 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

[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.

(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 frequentist 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 capabilities 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:

$$p = \frac{\binom{890}{90}\binom{210}{10}}{\binom{1,100}{100}} + \frac{\binom{890}{91}\binom{210}{9}}{\binom{1,100}{100}} + \cdots + \frac{\binom{890}{100}\binom{210}{0}}{\binom{1,100}{100}} \approx 0.008$$

This is indeed a possible p-value as calculated in Fisher's Exact Test. It's not the only possible p-value, as we will 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 capabilities, 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 a specific substantive hypothesis generated by many theories of syntax and the lexicon in mind, 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 was 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).

single-sided test

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

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:

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

There might be verbs for which the same theory predicts that they passivise more often than the average verb. 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:

$$p = \frac{\binom{869}{69}\binom{231}{31}}{\binom{1,100}{100}} + \frac{\binom{869}{68}\binom{231}{32}}{\binom{1,100}{100}} + \cdots + \frac{\binom{869}{0}\binom{231}{100}}{\binom{1,100}{100}} \approx 0.009$$

Since we are limited by the corpus size of 1,100 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 8, 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 will return repeatedly in later chapters, very crucially in Chapter 8.

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:[12]

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

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's just guessing whether cups are tea-first or milk-first, this will show over time, i. e., with an increasing 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 much less winners. To test his claim that the urn contains 30% winning lots, you agree to draw a sample of 10 lots. If you end 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.[13] 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

[12] Many of these calculations are virtually impossible to do by hand. Don't worry: We made most of them using software and double-checked everything.

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

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.[14] 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 place. If she claimed she had 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 better than merely rejecting a Null, thus providing weak evidence in favour of a hypothesis, etc. 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

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

(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. Thus, 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 potential p-values. Different effect strengths, on the other hand, correlate with different distributions of potential p-values. Again, Section 2.5 will shed some more light on this, but it's highly important to understand the difference. A large effect is easier to detect per se. The better Ms Bristow's tea-first detection, the easier it is 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.

effect strength

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.[15] 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.

[15] 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.

A p-value clearly isn't a 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.[16] Therefore, we can use them to simulate processes that generate data, where a *process* can be anything from Ms Bristow labelling cups of 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 (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.[17] Based on this uninformed guesser,

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

[17] 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

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

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:[18]

> *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! (Wow, I feel so sophisticated.)*

Unfortunately, this is fundamentally incorrect, as we will demonstrate using simulations now.[19]

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.

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

[19] 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.

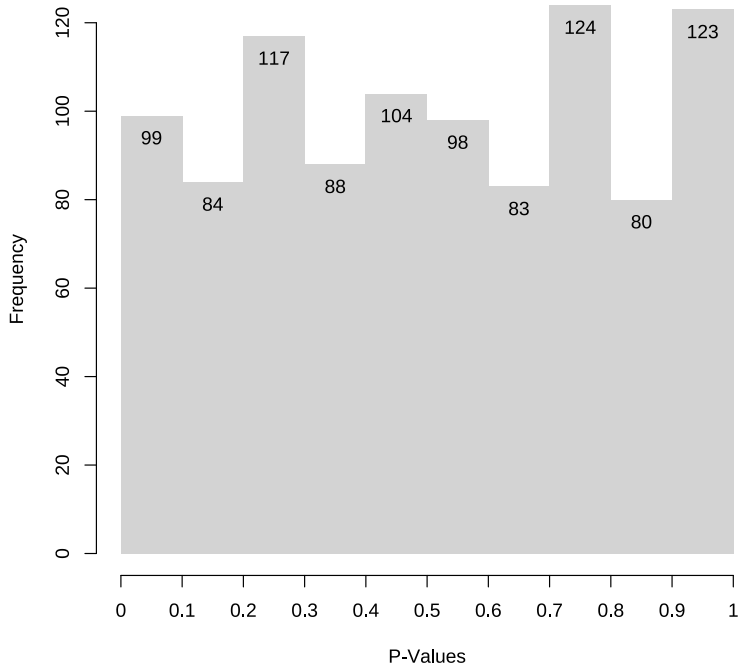

Figure 2.6: Histogram of p-values of 1000 simulated experiments with 1000 cups where the participant is merely guessing

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) correspond 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.[20]

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. This point is mostly neglected in introductions for practitioners, and it's one of the most dangerously 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-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 guesses tea-first cups slightly better than chance. We should warn you that the results might shock you, especially if you use corpora and statistical methods in your own work. Figure 2.7 shows the distributions of p-values for 1000 experiments with 1000 cups. The different panels plots the p-values for simulated participants with

[20] 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.

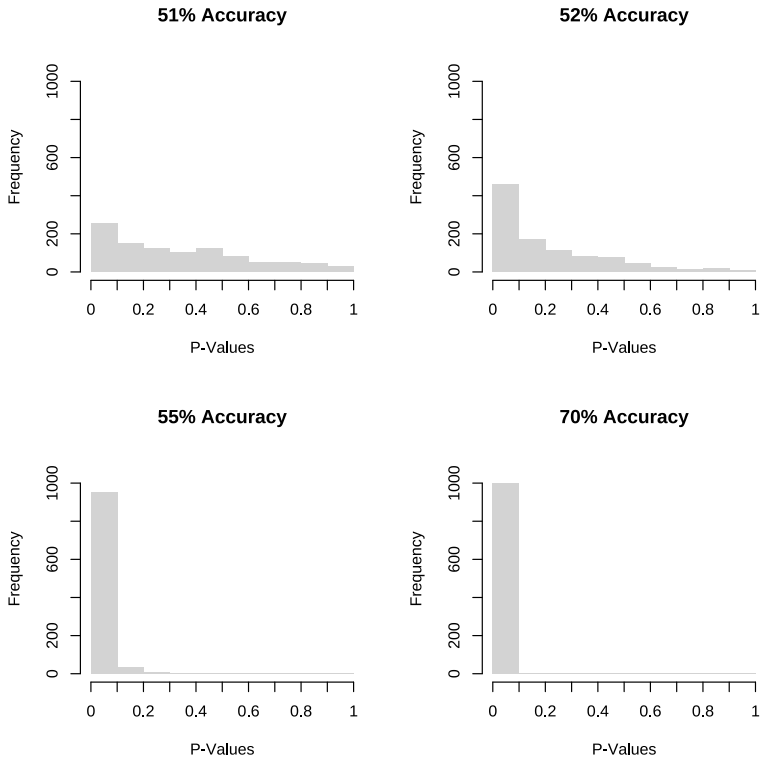

Figure 2.7: Histograms of p-values of 1000 simulated experiments with 1000 cups where the participant has increasing prediction accuracy (increasing from upper-left to lower-right); all y-axes are scaled to 1000 for better comparability

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, too, is an expected result. If the Null is not true, lower 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 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 $:= 0.05$ 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 capabilities 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.

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. With a large sample, the distribution of the p-values collapses close to 0 very quickly, which means that mechanical inferences based on pre-set sig-levels are almost always successful, even if the effect is 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. However, 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.[21] This is because Fisher's Exact Test was designed for small samples such as 8 cups, 4 types of wheat, etc. Also, the established sig-level of 0.05 was set for experiments with necessarily small samples. In analyses of corpus data with $n = 1000$, $n = 100000$ (or even $n = 200$) Fisher's Exact Test (and virtually all other frequentist tests) 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 8.

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, etc. 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 20) in order to make our point. This is because

[21] 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.

Fisher's Exact Test is extremely well-suited to introduce the logic of frequentist 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 8. 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 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 significantly 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.[22] 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, 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. Here is what they are: In statistics, a variable is a measurement of some kind of property of events. In the

[22] 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.

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 find out and formalise the probabilities for certain values of those variables. 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? And so on.

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 $N = 8$ choices in total (the overall number of cups). This effort culminated in Equation 2.4 on page 13, 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:

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

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), 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 gray 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 specific probability for each outcome that can be calculated with arbitrary precision using the discrete **density function** in Equation 2.6.[23] Some terminological clarifications are in order. We used the term **parameter** above to describe N, K, and n. Informally speaking, there is

Hypergeometri Distribution

density function parameter

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

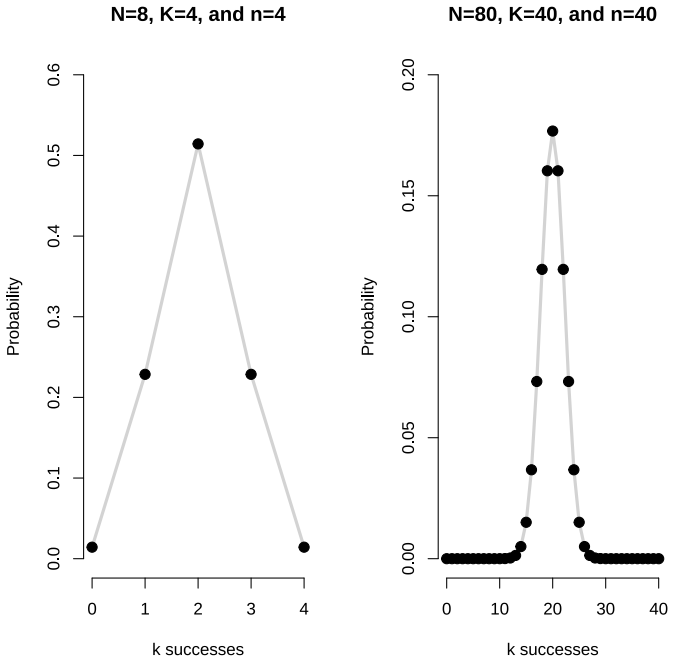

Figure 2.8: Probability density for k successes in Tea Tasting Experiments; the density function of the Hypergeometric Distribution

not one Hypergeometric Distribution but a large (infinite) class of distributions, one for each permutation of N, K, and n. The parameters define the concrete function and determine the exact shape of the curve for the admissible values of k, 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 correctly. We'll encounter other distributions later that are quite different.

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. 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.).[24] The exact maths aside, as exactly one

[24]Notice that the distribution for given parameters N, K, and n only describes properly termi-

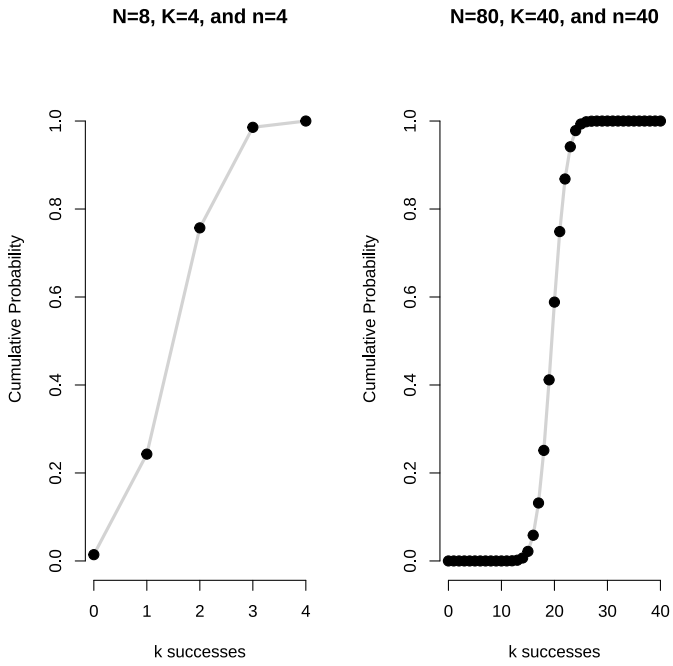

Figure 2.9: Cumulative distribution function for k successes in Tea Tasting Experiments; the distribution function of the Hypergeometric Distribution

outcome has to occur, the probability of the joint outcomes must be 1.

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.[25] 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

cumulative distribution function

nated 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.

[25] Unfortunately, cumulative distribution functions are often much harder to disentangle mathematically, and we do not provide their mathematical formulations. They are all readily available on Wikipedia.

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 $k = 10$.[26]

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

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

3 Describing Data

4 Sampling Accuracy and Confidence

5 Inferences About Means

5.1 Population Means and Sample Means

Problem Statement: Z-Tests

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?

5.1.1 Introducing the Logic

The Problem Statement exemplifies a common question: Given a known mean value, do means 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 tackle such questions. 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 truly known, which is why the Problem Statement stresses that the mean was predcited by a robust theory and that the prediction was tested in a long series of experiments. If these conditions are not met, other tests apply, and we're going to introduce such tests as we go along.

z-Test

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 corre-

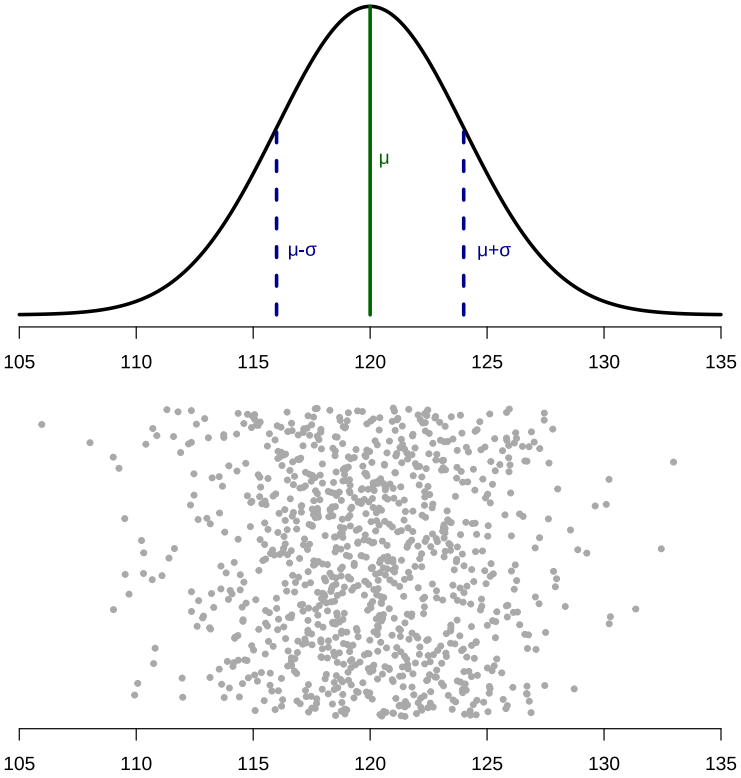

Figure 5.1: Theoretical population distribution for a normal distribution with μ=120 and σ=4 and a simulated random sample of n=1000 measurements from the population

sponds to a standard deviation of $\sigma = 4$. If the population values are generated according to a normal distribution, values are distributed according to the bell curve in Figure 5.1.

To recapitulate, this curve plots the probability density for the known population (for example, of reaction times). In the simplest terms, it shows for each measurement (x-axis) the probability with which it occurs (y-axis).[1] 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 dashed lines show

[1]Technically, the probability of each point measurement is 0, and non-zero probabilities are only defined as integrals of the density curve for intervals. This mathematical detail is mostly irrelevant for practical applications, but it should be kept in mind.

the standard deviation in each direction from the mean. As a result, a very much expected sample of $n = 1000$ measurements is shown in the form of the point cloud below the curve. 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 higher (for example because we're measuring reaction times under a condition that is more difficult to process) we expect samples to turn out differently and have a higher sample mean compared to the known population mean. However, this 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 by being very much unexpected (albeit not impossible) under the assumption that the desired inference is not correct. In the case of the reaction times described in the Problem Statement, the desired inference 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, as will be shown. Figure 5.2 shows a possible outcome with $n = 16$.

Let's call the sample plotted in Figure 5.2 x, a vector of 16 measurements x_1 through x_{16}. The mean of x is $\bar{x} = 122.1$. 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 H, where H: $\mu_1 > \mu$. For several reasons, we cannot gather evidence that supports this hypothesis directly. First, μ_1 is obviously not observable. It's a hypothesised mean in a population that exists as separate from the known population if H is correct. If that population is not substantially different from the known one, then we have $\mu_1 = \mu$. Given that μ_1 is a non-observable, all we've got are 16 data points from x and the sample mean $\bar{x}$ calculated from them. While we certainly hope that $\bar{x}$ is a good indicator of the true value μ_1, we have no guarantees whatsoever that it actually is. Second, as we're still Fisherians on this page of this book, we have no formal method of gathering positive evidence. Only Neyman-Pearson philosophy and the Severity approach will give us this power (pun intended) later in Chapter 8, but with some important caveats. Therefore, we can only check whether the data x are in accord with a **Null Hypothesis** (*Null* for

Null Hypothesis

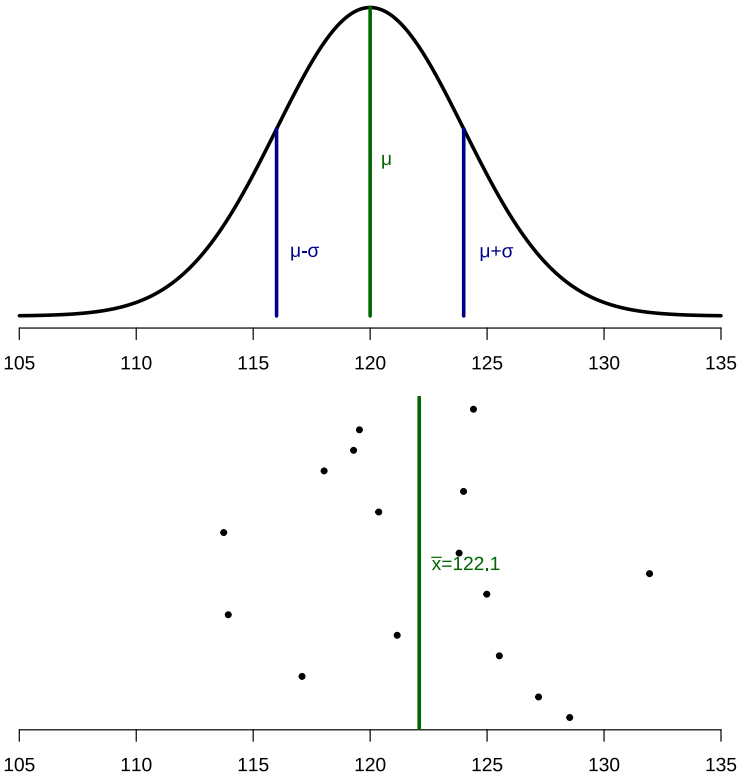

Figure 5.2: Theoretical population distribution for a normal distribution with μ=120 and σ=4 and a simulated random sample of n=16 measurements from some population

short) N: $\mu_1 \not\succ \mu$.[2]

To test this hypothesis, the Fisherian framework assigns a certain well-defined kind of probability to (obtaining) the data x given that N is correct, formally $p(x|N)$ or *the probability of* x *given* N. This probability can be used to assess whether the data x are in accord with the Null N. Before moving on to the calculations, it is vital to consider the question of which inferences are warranted in case x is or isn't in accord with N. First, what do we infer from the data x (in other words, from our experiment) if they are compatible with N? The answer

[2]The Null Hypothesis N is sometimes designated as H_0, in which case H is often called H_1. This nomenclature is—in our view—where the confusion between Fisher and Neyman-Pearson begins. Furtermore, as the Fisherian Null Hypothesis is not a proper hypothesis anyway but rather a non-hypothesis, we call it Null or N.

is: absolutely nothing! If the data are in accord with N, 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 Baeysians with Dutch names will come to haunt you (or at least unfollow you on the messaging platform of your choice)—and rightly so. Second, what do we infer from the data x if they are not compatible with N? This is the much more interesting case, but it's difficult to define the admissible inerences without creating false ideas if it's done without the maths and without a deeper look at the way H and N were set up. Let's say rather informally that in such a case we've found some evidence in support of H because N and H partition the range of possible values of μ_1: either it's greater than μ (H) or it is not greater than μ (N). Finding no accordance with N despite serious attempts to do so (see Chapter 8) provides at least some indication that H might be correct. If you're looking for 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 Doing the Maths

We have argued that in Fisherian inference, we have to asses whether x is compatible or in accord with N. But how do we do this? The most naive but not at all wrong thing to do is calculate the difference between the known population mean μ and the mean of the obtained sample $\bar{x}$. 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: $\mu_1 \not\succ \mu$.[4]

However, solid empirical inference requires us to evaluate how significant this difference actually is. This evaluation follows the same logic as in our introduction to Fisher's philosophy (Chapter 2). In the analysis of the Tea Tasting experiment, we asked how often someone who merely guesses would classify one, two, three, or four cups correctly by chance. In the case at hand, simply counting events is not informative as the events are occurrences of specific values, namely individual reaction times. Hence, the question becomes: how often would we expect to see a sample of 16 measurements with a sample mean of 122.1 or larger

[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.*

[4] This way of putting it is slightly sloppy and informal. We will return to this notion and make it more precise. 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.

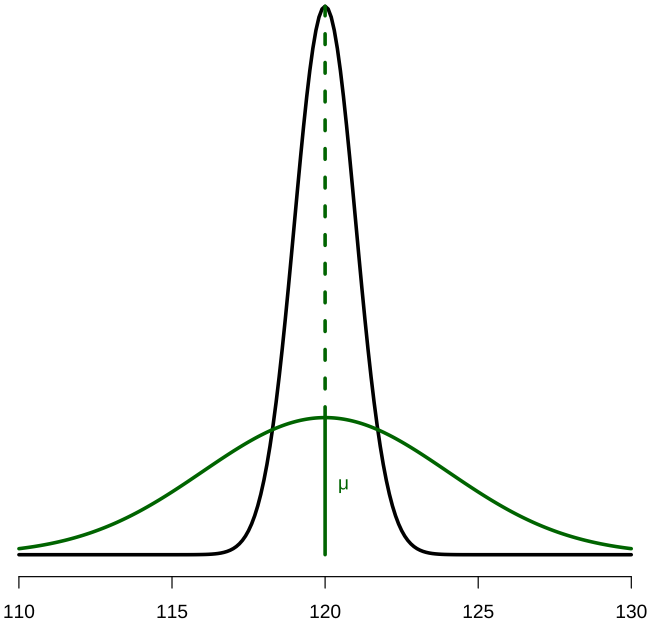

Figure 5.3: Theoretical population distribution for a normal distribution with μ=120 and σ=4 (green) the distribution of sample means for samples from this distribution with n=16 (black)

standard error

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. The standard error for $n = 16$ and the known variance $s = 4$ (see p. 43) tells us how much samples of this size differ from the true mean on average. The standard error of the mean is:

$$S(n, \sigma) = \frac{4}{\sqrt{16}} = 1$$

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 $\bar{x}_i$ of repeated samples of size n are themselves normally distributed with the standard error S being the standard deviation of that normal distribution. Furthermore,

keep in mind that we're talking about the distribution of the 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]

Since (i) the distribution of sample means is normal, (ii) we know its mean, (iii) we know its variance/standard deviation, we can calculate how many samples of infinitely many samples (or at least a lot of samples) drawn from the known population have a mean of 122.1 or larger. In other words, we can calculate how many sample means would deviate by 2.1 from the population mean anyway due to expected sampling error if we took a lot of samples of size $n = 16$. This is exactly parallel to the argument regarding the Tea Tasting experiment where we calculated how often we could expect certain outcomes anyway, even if the person performing the Tea Tasting task had no ability to detect which liquid was poured into the cup first. It gives us a very precise and well-defined measure of how surprising the obtained result would be were the Null true.

There are many ways of calculating the number of interest. Since the area under the normal curve sums up to 1 (as should be the case with probability density functions), we could integrate it over the interval $[122.1, \infty]$. Alternatively, we could use the so-called cumulative density function, to which we will return later. To make things much simpler in practice, the most widely used way in applied statistics is based on counting the distance between the distribution mean and the sample mean as multiples of the standard error, which gives us the **z-score**:

z-score

$$z = \frac{\bar{x}-\mu}{S(n,\sigma)} = \frac{122.1-120}{1} = \frac{2.1}{1} = 2.1$$

Figure 5.4 shows the distribution of sample means for the known population, the true mean μ, the obtained sample value $\bar{x}$, and the area under the normal curve which defines how many samples of size $n = 16$ (in the limit) have means of $\bar{x}$ or greater. In addition, the red line measures the distance from μ to $\bar{x}$, which corresponds to the z-score. By dividing the the distance between the sample mean and the population mean with the standard error, the z-score normalises said

[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 size $n = 1, 2, \ldots$.

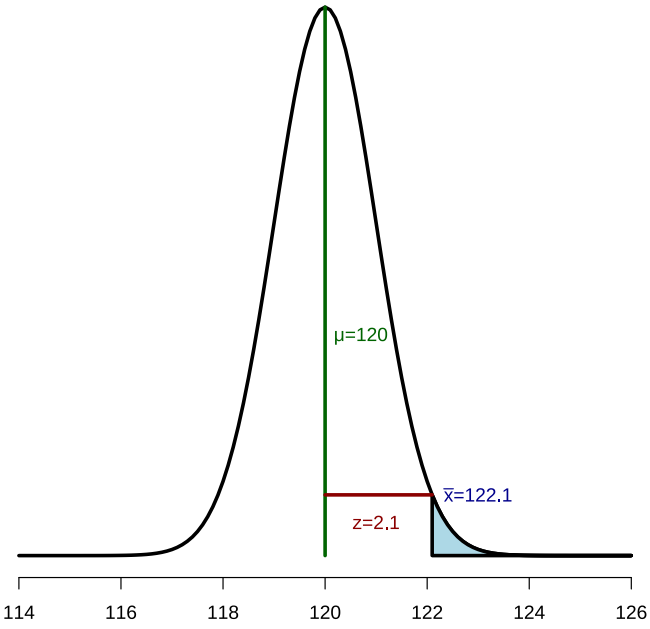

Figure 5.4: Distribution of sample means for mean μ and standard error S with obtained sample mean $\bar{x}$ and corresponding z-value; area under the curve for results equal to or greater than $\bar{x}$ is shaded

distance, which itself varies with the standard deviation in the population and subsequently with the standard error. Hence, regardless of the slope of the concrete distribution, $z = 2.1$ tells us how (im)probable the sample mean (or a more extreme sample mean) is under the Null. It gives us another infamous **p-value**. In the olden days (and in this book), tables were used to look up p-values corresponding to z-scores, and modern statistics software has functions to achieve the same. We will discuss those tables later in this chapter, but for the time being, let's take it for granted that

p-value

$$Pr(\bar{x}|N) = p_{\mathbf{Norm}}(2.1) = 0.02$$

Let's go through this step by step. First, $Pr(\bar{x}|N)$ reads *the probability of the sample mean $\bar{x}$ given the Null N*. Remember that the Null was specified as N: $\mu_1 \succ$

μ, which reas *the mean under the condition of interest* (the mean reaction time in the emergent type of relative clause) *is not greater than the population mean* (the reaction time in other relative clauses). Second, this probability is equated to $p_{\text{Norm}}(2.1)$, which is the p-value corresponding to the z-value of 2.1 as calculated above, which is 0.02.

5.1.3 Interpreting the Results

At this point, a little exercise is in order. From the following statements, choose the one which is a correct interpretation of the calculations above given the scenario described in the Problem Statement. There is no pressure. Nobody can read your mind, nobody even cares whether you pick a wrong statement, and you can only gain by actually *thinking* about each statement thoroughly. Do not decide on one statement because it is intuitively correct, but because you know why it is correct.

1. The probability that hypothesis H (reaction times are longer in the emergent type of relative clause) is true is 0.02.
2. The probability that hypothesis H (reaction times are longer in the emergent type of relative clause) is true is 0.98.
3. The results prove that the emergent type of relative clause incurs longer reaction times than other relative clauses.
4. The p-value is very small, which indicates that mean reaction times in the emergent type of relative clause are substantially longer than in other relative clauses.
5. The probability of obtaining another sample with a mean of 122.1 or greater in an exact replication of the experiment is 0.02.
6. Based on this outcome, we can reject the possibility that reaction times under the condition of interest are actually *smaller* than in the population with a certainty of $1 - 0.02 = 0.98$ (or 98%).
7. The probability that we actually drew a sample with a mean of 122.1 is 0.02.
8. The experiment has shown that reaction times are normally distributed.
9. The experiment provides evidence in favour of the underlying theory of linguistic processing.

[TODO continue here]

Big Point: Interpretation of the P-Value in Z-Tests

The p-value in a z-test is the frequentist probability of drawing a sample x with a mean as extreme as or more extreme than the one that was actually obtained if the Null were true. The frequentist probability is the probability of an event occurring before it has actually occurred. After the sample has been drawn, the probability that it was drawn is 1, regardless of how extreme its mean is.

5.1.4 Differences Between the Fisher Test and the Z-Test

5.1.5 The Distribution of P Values

5.2 The Unknown Population

5.3 Means of the Unknowns

6 Differences in Means and Variances

7 Some Other Scenarios

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

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

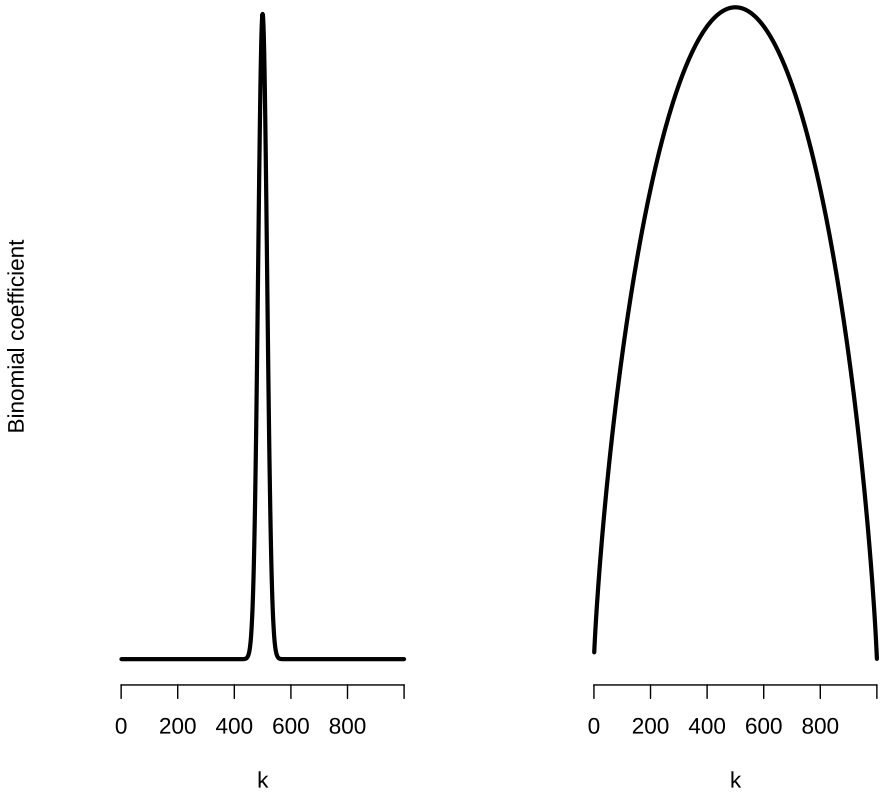

Figure 7.1: Number of ways of choosing k elements from $n = 1000$ elements; Left: regular y-axis, Right: logarithmic y-axis

8 Positive Inferences

9 Quantifiying Correlation Between Variables

10 Modelling Linear Relationships

11 Modelling Arbitrary Outcomes

12 Modelling With Groups

References

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

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.