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: XƎLATEX

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 iii

Acknowledgments v

1 Scientific Inference and Error 1

2 Guessing and Counting 3

2.1 Unexpected Outcomes . . . 3
2.2 Tea . . . 5
2.3 Inference: Fisher's Exact Test . . . 13
2.4 Probability Distributions . . . 17
2.5 Sample Size and Effect Size . . . 17
2.6 The Distribution of P-Values . . . 17

3 Describing Data 19

4 Sampling Accuracy and Confidence 21

5 Inferences About Means 23

5.1 Population Means and Sample Means . . . 23
5.1.1 Introducing the Logic . . . 23
5.1.2 Doing the Maths . . . 27
5.1.3 Interpreting the Results . . . 31
5.1.4 Differences Between the Fisher Test and the Z-Test . . . 32
5.1.5 The Distribution of P Values . . . 32
5.2 The Unknown Population . . . 32
5.3 Means of the Unknowns . . . 32

6 Differences in Means and Variances 33

7 Some Other Scenarios 35

8 Positive Inferences 37

9 Quantifiying Correlation Between Variables 39

10 Modelling Linear Relationships 41

11 Modelling Arbitrary Outcomes 43

12 Modelling With Groups 45

Index 47

Name index 47

Preface

Acknowledgments

1 Scientific Inference and Error

2 Guessing and Counting

Overview

In this chapter, [todo]

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, how surprised would you be to find ten passives of *sleep* in a corpus of English? You should really think about your answers to these questions before continuing on.

2.1 Unexpected Outcomes

Did you think about the questions from the Problem Statement? Here's one possible discussion. Most importantly, the questions from the Problem Statement cannot be answered properly, 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? Is it the Eurojackpot, where (to simplify matters a bit) you have to guess five numbers out of fifty 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. Potentially, you wondered whether the lottery was supposed to be fair or not. However, when presented with this example, people typically don't worry too much about how the lottery was conducted and whether it was fair. At least with big national lotteries, most people put trust in there being sufficient oversight and the draw being—here it comes—properly random. Above all, 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 very similar, but 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 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 twelve-digit phone numbers of 28 guests with the same accuracy, for example. Furthermore and most likely because it involves a psychic, this scenario usually 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 in order to track down relatives who lived as maids and servants at the court of Henry 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 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 can come 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 and that you saw the parallels to the first two scenarios.

[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.)

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 are vastly more complicated than in the case of a lottery. For example, is it a corpus of language produced by native speakers, children, L2 learners of English, state-of-the-art large language models, or even some cute language bot from 1998? 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? Maybe all ten sentences are instances of silliness such as (1). Would the result still count as unexpected, regardless of the quantitative evaluation?

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

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 applying statistical inference should be aware of it, even if they're not in Tea Studies.

2.2 Tea

What unites the examples in the Problem Statement is that they describe a confrontation with chance. Then, you're asked what kind of a result would be unexpected under the assumption that there is nothing going on: you're not prescient, the psychic isn't actually a psychic, *sleep* cannot be passivised. In this section, we formalise the notion of **unexpected outcome** in relation to 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 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 you were absolutely certain anyway that *sleep* cannot be passivised, you wouldn't bother to do a corpus search for passivised forms of that verb. In fact, that's what many self-described theoretical linguists do. Clearly, unexpected outcomes are not miracles where everything we know about the world can be negated.

What we usually mean when we deem an outcome *unexpected* is that it had a very slim chance 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 a winning one, and you draw one, most people know you have a chance of one in a thousand (or 1:1,000) to win. Usually, it is understood intuitively that this means that if you played this game over and over again with a fair urn, 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 is 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 are 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's essentially 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 even for that we can apply a smilar logic.

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 **Fisher1935a**, 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 some 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 above 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 get right 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 eight cups. 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 identifiably individually. There is one gray, one red, one blue, and one green cup for each of the conditions (milk-first or tea-first).

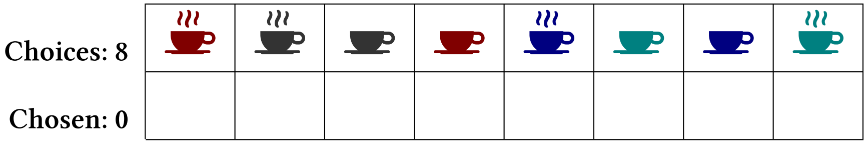

Figure 2.1: Four tea-first cups (steaming) and four milk-first cups (not steaming) for Muriel Bristow to choose from; in the actual experiment, they'd all look exactly the same (without colours).

When 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.

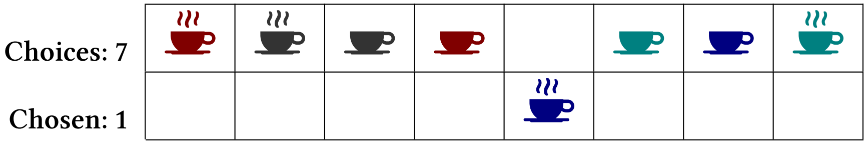

Figure 2.2: That's the first cup chosen! Choices left: 7.

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 visual example, she chose the fifth cup and has the first, second, third, fourth, sixth, seventh, and eighth still available. Had she chosen the leftmost cup (red and steaming) instead in the first step, she'd now be confronted with a different set of 7 options (all except the leftmost one). 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.[3] Put differently, she has taken 1 out of 56 possible decision paths to choose 2 out of 8 cups.

[3] 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?

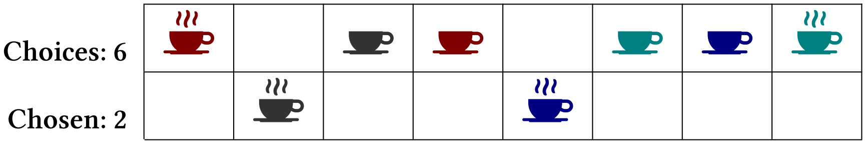

Figure 2.3: After another cup was chosen, there are now 6 choices left.

The story continues in a similar vein. Let's assume she chooses the red steaming cup as her third pick, as in Figure 2.4. (Notice that she's doing well so far in the example. All three choices were correct.) She has now chosen 1 of $8 \cdot 7 \cdot 6 = 336$ possible choices, since for each of the 7 options left over after her previous decision, she had 6 distinct choices available.

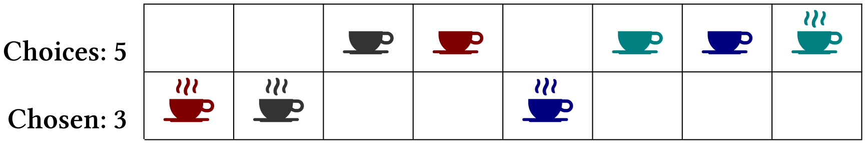

Figure 2.4: As Ms Bristow picks another cup, we're down to 5 choices.

To make things more interesting, she makes her first incorrect guess for the fourth cup. She picks the green non-steaming one (sixth from the left), which is a tea-first cup. As it was agreed upon that we would take her first answer, she has now decided on 1 specific configuration from $8 \cdot 7 \cdot 6 \cdot 5 = 1,680$ possible configurations. Or has she? As we mentioned in Footnote 3, there's a catch.

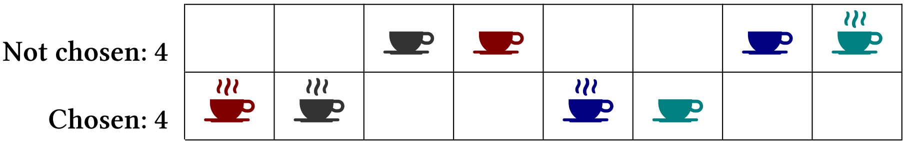

Figure 2.5: Ms Bristow has picked her last cup.

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 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 seven choices, etc. She could have chosen the red steaming one and still arrived at the same 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 staeming one and the green non-steaming one. Put in quantitative terms, there are groups within the set of $1,680$ 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, it is indeed irrelevant. In our example, she guessed three cups correctly, and the order of decisions that led to this choice would not change her success rate in our eyes.

How do we know how many of the $1,680$ decision paths lead to identical results? Well, how many different ways of ordering four cups are there? Image you had to arrange the 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

$$\frac{8 \cdot 7 \cdot 6 \cdot 5}{4 \cdot 3 \cdot 2 \cdot 1} = \frac{1{,}680}{24} = 70$$

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

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 four 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 an immaculate guess. Hence, 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. In other words, there is a 1:70 chance of guessing the four cups by mere luck. Put differently, 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%.[5] 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

Before proceeding to such delicate matters of scientific inference, let's calculate the probability of guessing three cups correctly, as in the example. To do

[4] And for those who are beginning to remember their elementary stochastics: results vary if the order is not irrelevant and if a cup can be chosen more than once.

[5] 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 real values between 0 and 1, and not as percentages.

factorial

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 $4 \cdot 3 \cdot 2 \cdot 1 = 4!$ if $n = 4$. To calculate the number of possible decision chains for four out of eight 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

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

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

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

binomial coefficient

To account for the decision paths that led to identical results, we divided this 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:

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

The binomial coefficient is usually read *n choose k*, but it's okay to read it as *k chosen from n* etc. Since the factorial results in very large numbers even for relatively low input numbers, it is often practically impossible 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 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:

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

There are thus 16 distinct sets of 3 correct and 1 incorrect cups. We already know that there are 70 ways of choosing any 4 cups out of 8, and hence $16 \div 70 \approx 0.23$ is the frequentist probability of achieving such a result by chance. Would this be a highly unexpected result? Most likely not, 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 tree 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:

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

The probability of obtaining 3 or 4 correct cups by chance is thus $17 \div 70 \approx 0.24$. Again, in the light of this number, choosing 3 or more correct cups is not at all an unexpected result. It provides 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. So, do you find this impressive?

What does this statistical anecdote show? In a first approximation,

[continue here]

Big Point: Unexpected Outcomes

The outcome of an experiment is unexpected if it had a low probability before the experiment was conducted. After that, the outcome is a fact and doesn't have a meaningful frequentist probability assigned to it. A low probability of a specific outcome means that it would be rare if the experiment were conducted very often.

[blabla]

There are two important aspects with respect to making inferences from unexpected outcomes. Can you be *sure* that Muriel Bristow had the capability of discerning tea-first cups from milk-first cups just because in the real experiment, she pointed out the four 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? After all:

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

Clearly, most readers would not make such an inference, regardless of how low the p value is. The interpretation of such a statistical result needs to be informed and made with great care. In this case, no known mechanism could account for ESP, which is why most people don't even bother to run experiments investigating ESP. Also, notice that we did not conduct an experiment about my grandfather's ESP capabilities, but I merely told a story about him winning the lottery. I could have told the same 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, you're bound to find it somewhere, and the low p value becomes utterly meaningless. Interestingly, Carl and his friends also repeated the "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 and got lucky once. Obviously, repeating experiments (so-called **replication**) is a very good way to further test any inferences made from unexpected outcomes.

replication

Astonishingly, researchers in soft sciences are often satisfied with $p < 0.05$ in order to proceed with a substantive inference from an experiment. Such thresholds are often called the α level, although we prefer **sig level**. 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 any 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

sig level

three, and there was a big prize behind one of the doors and a dud called *Zonk* 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 1 door is hiding 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 $sig := 0.05$. On the other hand, we have seen that even $p \approx 0.0000000715$ might be meaningless. Doing maths is easy, but making good inferences isn't. Therefore, two of the major themes in this book are that (i) you shouldn't take high risks in making inferences and (ii) there is no recipe-like procedure that leads to good inferences.[6]

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 the so-called Fisher Exact Test or just Fisher Test. You'll see that the logic of the Tea Tasting Experiment lies behind those omnipresent 2x2 tables that often pop up in the corpus literature, especially in research on collocations and collostructions (**StefanowitschGries2003**, **Evert2008**).

2.3 Inference: Fisher's Exact Test

Our examples from the Problem Statement are all about comparing counts of events. How many tea-first cups were chosen correctly compared to the number of tea-first cups in the setup of the experiment, and the same for milk-first cups. 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 shown as **contingency tables**, see Table 2.1.

contingency tables

A contingency table tabulates counts of events characterised by two variables, each having two or more discrete possible values. One variable is shown in columns, the other one in rows, and the table has the potential to show which values of the two variables co-occur very often or very rarely. In the case of the Tea Tasting Experiment, we tabulate numbers of cups (or events of choosing

[6] Anyone who thinks they're a Bayesian and just shouted "*Yeah, except in Bayesian inference!*" should stop reading and take a long, hard look in the mirror.

Table 2.1: A typical 2x2 contingency table with row sums, column sums, and a grand total

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

cups, to be more precise). In row 1, we put the 4 tea-first cups according to Muriel Bristow. In row 2, we put the 4 milk-first cups as assigned by her. The row sums in the last column show that there were indeed 4 cups for each condition. In the two columns, we count the real tea-first and milk-first cups. As you can see, of the three real tea-first cups, Muriel Bristow classified 3 as tea-first (cell 1,1) and 1 as milk-first (cell 2,1).[7] The opposite is true 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 same contingency table with the relevant components for Fisher's Exact Test highlighted

		Reality		
		Tea-First	Milk-First	
Bristow	Tea-First	3	1	4
	Milk-First	1	3	4
		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.

Finally, we turn to the corpus example from the Problem Statement.[8] As we pointed out in Section 2.1, the Problem Statement mentions 10 passives of the verb

[7] 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.

[8] 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.

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

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

sleep, but we need more information. Let's introduce that information and the argument that comes with it, roughly following the logic behind collostructional analysis.[9] 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**	
Verb	***sleep***	90	10	**100**
	Other	800	200	**1,000**
		890	**210**	**1,100**

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 25% of all other verbs are passives. Using the maths introduced in this chapter, we can attempt to quantify how unexpected this result is, 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:

$$\frac{\binom{890}{90}\cdot\binom{210}{10}}{\binom{1,100}{100}} \approx \frac{6.573\cdot10^{141}}{1.423\cdot10^{144}} \approx 0.005$$

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

[9] 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 a specific use of it in the process.

now inch our way towards making an inference about *sleep*. Our theory states that there should be no or at least very few passives of verbs like *sleep* (unaccusatives) compared to other verbs, many of which can be readily passivised (transitives, unergatives). If *sleep* behaved like the average verb, a sample of sentences containing it would be expected to resemble a random sample from the corpus with respect to the number of actives and passives. The more extreme the distribution of verbal voice in the *sleep* sample is, the more we are inclined to assume that *sleep* does not behave like the rest of the verbs with respect to passivisation. This is actually a similar logic as in the Tea Tasting Experiment. An unexpected result under the assumption of mere guessing on Muriel Bristow's part is conceptually very similar to an unexpected result regarding the distribution of verbal voice in a corpus sample under the assumption that the sample is not in some way different from the rest of the corpus. The Fisherian type of the frequentist logic of inference is based on this kind of argument: Since it is very difficult to come up with quantitative evidence in favour of research hypotheses (see Chapter 1), a **Null Hypothesis** (or simply *Null*, symbolised H_0) is constructed. The Null states in some way that the effect predicted by the theory is absent. If the result obtained in the experiment has a very low 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. Unexpected results are not in any way taken as a proof of the theory or part of the theory, and it's obvious why. First, we never know whether a rare (unexpected) 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 expectedness of the actual result under a Null (i. e., by chance). Second, take the abovementioned "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:

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

This is indeed the p value as calculated in Fisher's Exact Test.

[continue]

2.4 Probability Distributions

2.5 Sample Size and Effect Size

2.6 The Distribution of P-Values

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-

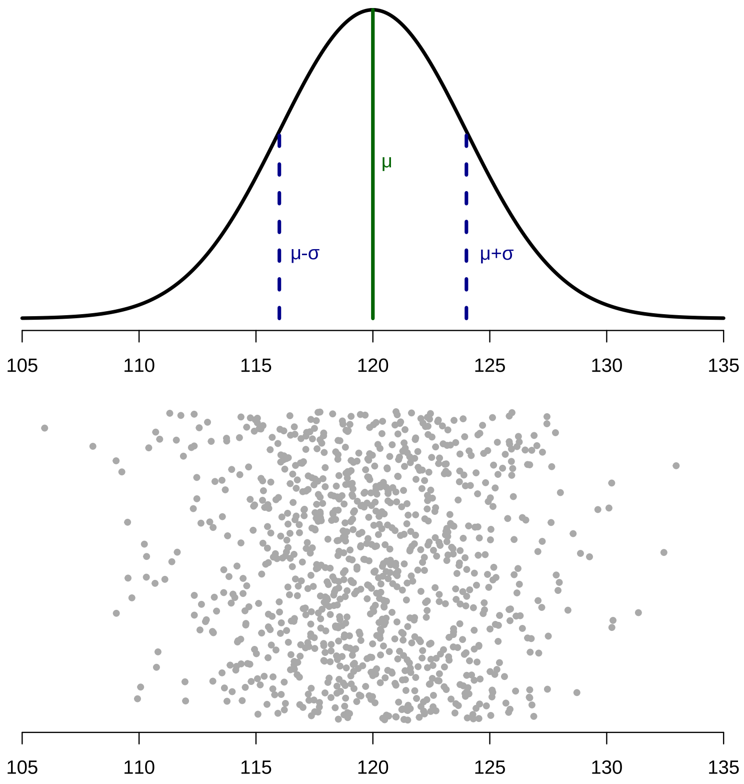

Figure 5.1: Theoretical population distribution for a normal distribution with μ=120 and σ=4 and a simulated random sample of n=1000 measurements from the population

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

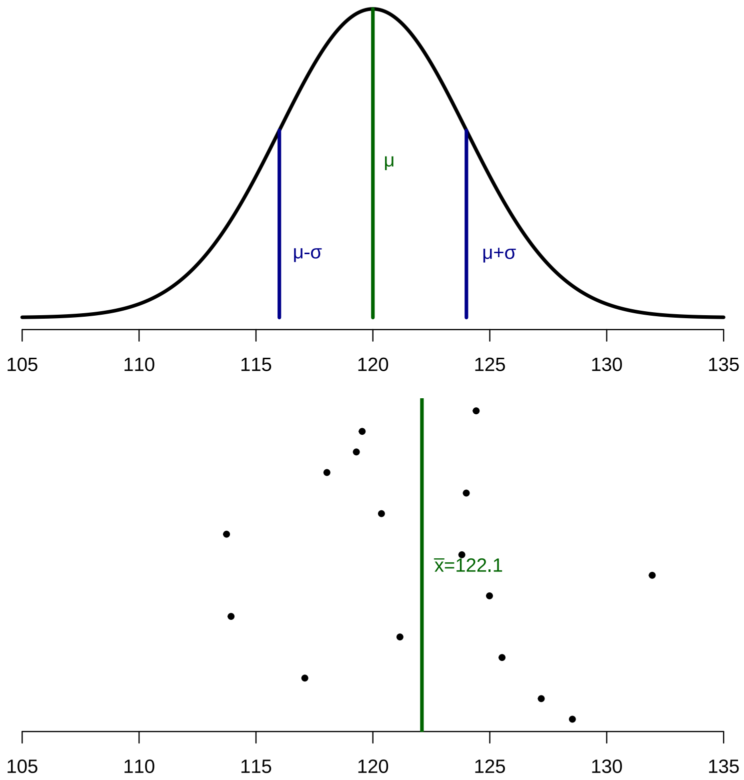

Figure 5.2: Theoretical population distribution for a normal distribution with μ=120 and σ=4 and a simulated random sample of n=16 measurements from some population

short) N: $\mu_1 \ngtr \mu$.[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: $\mu_1 \ngtr \mu$.[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.

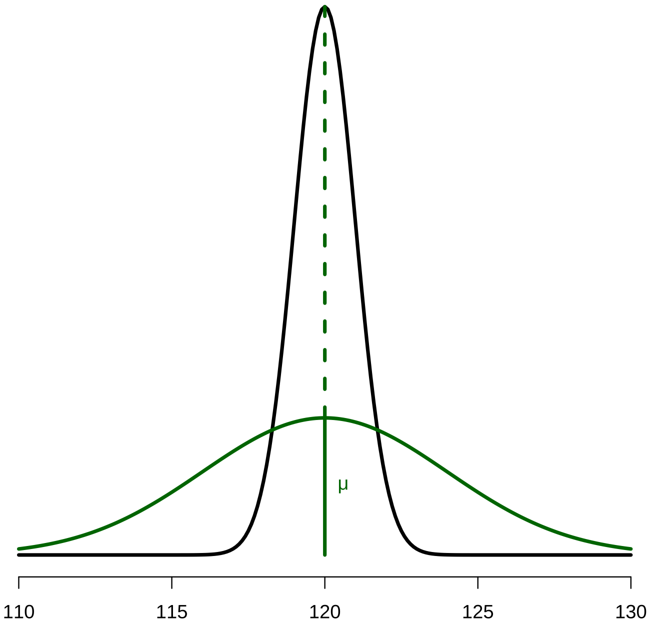

Figure 5.3: Theoretical population distribution for a normal distribution with μ=120 and σ=4 (green) the distribution of sample means for samples from this distribution with n=16 (black)

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. 23) 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 $n = 1, 2, \ldots$.

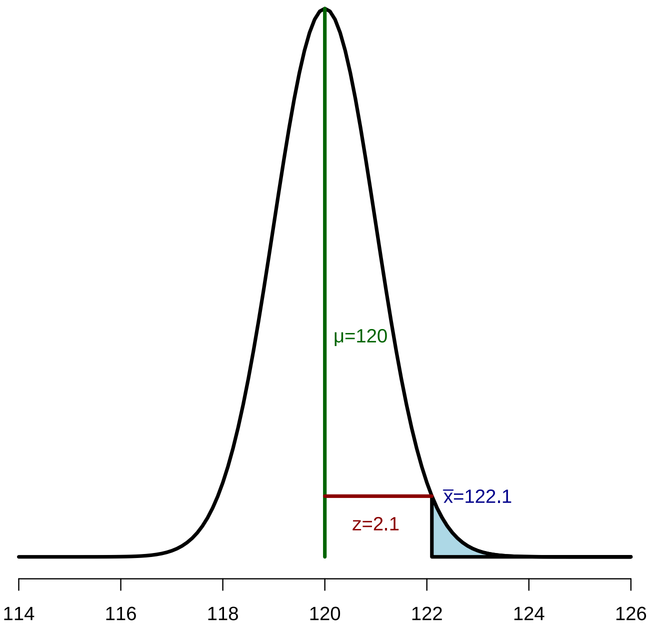

Figure 5.4: Distribution of sample means for mean μ and standard error S with obtained sample mean $\bar{x}$ and corresponding z-value; area under the curve for results equal to or greater than $\bar{x}$ is shaded

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: $\mu_1 \ngtr$

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 $p_{\textbf{Norm}}(2.1)$, 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

8 Positive Inferences

9 Quantifiying Correlation Between Variables

10 Modelling Linear Relationships

11 Modelling Arbitrary Outcomes

12 Modelling With Groups

Statistical Inference for Everybody and a Linguist

This book is good.