Chapter 1
Why do linguists need statistics?

James Myers
2022/2/6 draft

1. Introduction

If you're reading this book, you probably love language. Maybe you love teaching your native language to foreigners, or love teaching kids about how to read and write their own language, or love helping children overcome their speech impediments or aphasic patients cope with their communication deficits, love studying how people talk or write spontaneously, or how babies discover their mother tongue, or how Deaf people sign, or how language is processed in the mind or brain, or how language varies across a society or changes across generations, or how languages differ so much across the world while still sharing many things in common, or how language is structured grammatically, in phonology, morphology, syntax, or semantics.

What you might not love so much is math, and statistics is math. But everywhere you look, you see people using statistics to help study the things you love: educators comparing teaching methods, corpus linguists counting word usages, psycholinguists measuring response times, neurolinguists making colorful brain maps, phoneticians quantifying tone changes, even grammarians plotting how acceptable or unacceptable different kinds of sentences are. Maybe even your friends like to show off their knowledge of mysterious things like "*p* values" and "*t* tests". You may not love math, but you want know how those guys do it. You want to learn the secrets of statistics.

This book will reveal many of these secrets. It won't tell you everything that you might want to know about statistics, of course: reading an introductory statistics textbook may be better than nothing, but reading two textbooks would be even better, and there are a huge number of other statistics textbooks out there, including several focused on language (e.g., Baayen, 2008; Brown, 1988; Eddington, 2015; Gomez, 2013; Gries, 2013; Hatch & Lazaraton, 1991; Jockers, 2014; Johnson, 2008; Larson-Hall, 2015; Levshina, 2015; Plonsky, 2015; Rasinger, 2014; Winter 2019; Woods et al., 1986). There are even more useful resources on the Web, including many tutorials in Chinese.

I wasn't born knowing statistics either, and when I'm faced with something I don't understand, I search the Web for papers, books, Wikipedia articles, and discussion boards for help. You learn statistics the way you learn anything: expose yourself to a variety of viewpoints on the topic in order to form your own viewpoint (or as a linguist might say, your own mental representation), and then in order to solidify the concepts, you practice practice practice. Inevitably you'll forget something important, but no problem - just look it up again.

In other words, the most important secret to statistics is one you already know: even in a class with a teacher and a textbook, you still have to learn how to learn on your own.

Hopefully this textbook will be a helpful part of that process. In this first chapter, I'll just focus on the most basic question of all: How on earth could statistics help somebody who loves language more than math? I'll start by comforting those of you who don't like math, then explain why I think learning statistics is good for you anyway, and finally give a quick preview of the rest of this book.

2. Numbers and words

A lot of people are afraid of math, especially people interested in the humanities, like history, philosophy, and of course the many fields that involve language, from literature to language teaching to grammatical analysis. Why? I'm not an expert on math anxiety (see, e.g., Ashcraft, 2002), but the central problem seems to be that math doesn't seem intuitive. "Real" human life is illogical (there's no equation for love), but math requires you to use cold, unfeeling logic. Math also seems too abstract, with simple concepts building up into more complex ones, in an ever-growing tower that quickly rises out of sight high in the sky.

These criticisms of math have some truth to them, but there's a more positive way to see things. Math can't be entirely counterintuitive, since it's something that real people actually do and think about with those squishy biological organs called brains. Both mathematicians and cognitive scientists have long been interested in the interface between math and the human mind (see, e.g., Dehaene, 1997; Hersh, 1999; Lakoff and Núñez, 2000). Math must be related to the real world too, since a lot of mathematical concepts were inspired by trying to solve practical problems or playing with puzzles based on real-life observations. These problems and puzzles don't even need to involve numbers directly: geometry is about shapes, game theory is about games. There's even something called **formal language theory**, which is crucial in computational linguistics (Jurafsky & Martin, 2009, give a pretty clear introduction).

As a quick example illustrating these points, consider a puzzle inspired by real language: What is the longest possible Chinese sentence? Forget about how long a speaker of the sentence could actually live or how much paper we would need to write it down, and just consider: what is the longest logically possible sentence? The answer, of course, is that there's no upper limit at all. But how can we know this? How can we have such certain knowledge about something that we could never directly observe? I think the intuition is based on a simple proof (i.e., a kind of math, within formal language theory): Imagine that somebody claimed that some sentence S is indeed the longest possible Chinese sentence. But if S is a sentence, then isn't S'

= 我說S (“I say S”) also a Chinese sentence, and isn’t S’ longer than S? There you go: there’s no longest Chinese sentence, and you just did some abstract, but very intuitive, mathematics.

Anybody can do any kind of math, as long as they take it step by step. Geniuses might be able to jump over many steps at once; ordinary people might need the steps to be very very tiny, and maybe repeated a few times in a few different ways. But I believe anybody can climb that mathematical tower as high into the sky as they want.

Fortunately, for this book you don’t need to climb very high, since most of the time our focus will be very practical: how should we analyze this or that type of linguistic data? Nevertheless, along with the practice practice practice, you also need to develop intuitions about the mathematical concepts too, or else you won’t know when to apply what type of analysis to which kind of data. Good cooks aren’t slaves to their cookbooks; they develop their own intuitions about what works and what doesn’t. My job is to try to make it clear how the statistical concepts build on each other, step by step, in a logical way (or at least in a way that isn’t entirely crazy), so you can develop intuitions that you can apply to your real data, even if they don’t look exactly like the data I’ll analyze for you in this book.

Moreover, no matter how important intuitions are, cold, unfeeling logic can actually be good for you too. Intuitions are often wrong, and without taking a bit of extra effort you will never find out what’s really going on around you. The Earth looks flat; we only know it’s actually round because of the effort people took to think and look a bit more carefully.

Consider the following classic arithmetical puzzle (see, e.g., Kahneman, 2011), which I’ve modified from the original (about the prices of a bat and baseball) to turn it into a linguistic puzzle:

> If the consonant cluster [st] is 110 milliseconds (ms) long, and the [s] is 100 ms longer than the [t], how long is the [t]?

Even though I was a math major in college and have taught statistics for years, I still want to say the answer is 10 ms, but that’s wrong. To get the right answer, I have to slow down and think a bit more carefully. For example, we can use algebra (代數學):

(a) $s + t = 110$ (first statement)

(b) $s = t + 100$ (second statement)

(c) $s - t = 100$ (subtracting t from both sides of (b))

(d) $2s = 210$ (adding (a) and (c))

(e) $s = 105$ (dividing both sides of (d) by 2)

(f) $105 + t = 110$ (putting (e) into (a))

(g) $t = 5$ (subtracting 105 from both sides)

Let's see: 105 ms + 5 ms = 110 ms, and 105 ms is 100 ms longer than 5 ms. Correct! That's an extremely short [t] (but then no real baseball costs 5 cents either).

Maybe you're smarter than me, and were able to solve this puzzle instantly in your head. In that case, reading through (a)-(g) was probably kind of annoying, since the algebra works differently from your mental images, but at least the slowed down, purely logical algebra confirms that your intuitions did indeed give you the right answer.

Images - hm. What if we represent the puzzle visually? We can draw the [st] cluster and indicate what we know about the total length, and then to show the length difference, we can draw the [s] again, lined up on the right this time, so it can be compared with the [t], as in Figure 1:

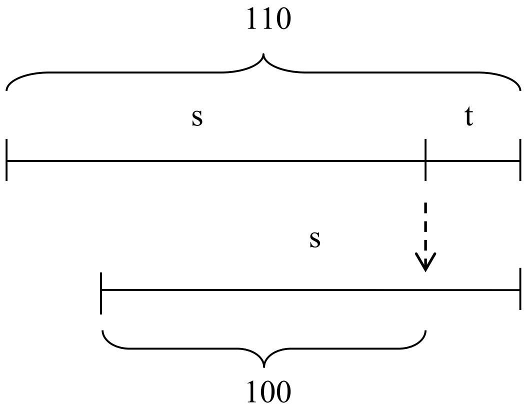

Figure 1. One way to visualize the [st] puzzle

Aha! If lining up the [s] on the left leaves a [t]-length bit on the right, then lining up [s] on the right must leave a [t]-length bit on the left... so the 100 ms bit must be exactly in the middle... so if the left and right [t]-length bits, which are the same size, add up to 10 ms, that means... [t] is 5 ms!

Did that help? If not, spend a few more minutes thinking about this silly problem until you develop your own intuitions about why the right answer is right. Then as you read the rest of this book, do the same thing with every equation and figure that I show you: slow down and think carefully, and maybe do it again another way, and hopefully you will be able to develop your own intuitions.

By the way, note how I was able to change the original example, which concerned the prices of a bat and baseball, into an example involving consonant durations. You should practice doing this kind of thing too, since it's very unlikely that your real data will look exactly like the examples discussed in this book. So another part of your intuition practice is learning how to see abstract (mathematical or logical) connections between situations that superficially seem very different.

3. Quantification and statistics

Here I'll mainly focus on two practical questions: When do linguists need **quantification**, and when do they need **statistics** in particular? But I'll also ask a bigger, more philosophical question: How does learning statistics affect the way you see the world?

3.1 Quantification

As we saw, math goes beyond mere numbers, and indeed there are times when numbers are entirely irrelevant to making a convincing argument. We already saw that with the proof that there is no longest Chinese sentence. Statistics wouldn't help here, since there's no way we can list all possible Chinese sentences in search of the longest one. Instead we must rely on pure logic.

Logic can also be decisive for more realistic questions as well. Suppose a linguist claims that it is absolutely impossible for the human mind to learn any language with an object-subject-verb (OSV) order, kind of like the weird-sounding language spoken by the alien Yoda in the *Star Wars* movies ("Your father he is"). This claim predicts that there should be no OSV language here on earth. However, such languages actually do exist (e.g., Tobati, a language spoken in Indonesia; Donohue, 2002). Since all we need is one language to kill this hypothesis, we don't need to do any quantification, let alone any statistics.

Other times we can't use numbers because there's no clear way to quantify the issue we want to study. If there isn't enough data, or the relevant concepts are not yet precise enough, then it would be more appropriate to do a **qualitative** study than a quantitative one (see, e.g., Denzin & Lincoln, 2011; Hasko, 2012; Heigham & Croker, 2009). For example, if the last speaker of an unrecorded language is a very old woman, it would make a lot more sense simply to consult with her to find out anything you can about her language, rather than annoying her with a long multiple-choice survey (for quantification), or trying to design an experiment where you don't know enough about the language to decide what the crucial variables might be. Even when you do plan to run a real experiment eventually, like testing two different language teaching methods, it's wise to first run a small-scale **pilot experiment**, just to find out if there are any obvious problems with your hypotheses or unexpected difficulties with your methods. If in the pilot your new teaching method immediately puts all the students to sleep, there's no point going on to a real experiment, let alone quantifying anything.

But in linguistics, quantification is possible, and even necessary, much more often than you might think. Let's go back to our linguist's hypothesis about OSV order. Suppose the linguist says that OSV order is not really impossible to learn, just very hard, so even if such languages do exist, they should still be very rare. It's not as easy to test this claim as the one about absolute impossibility, since it's vaguer: what do "very hard" and "very rare" mean, exactly? To get anywhere, we need to quantify it to make it more precise. Let's try.

Well, the **WALS database** (World Atlas of Language Structure; Haspelmath et al., 2005; http://wals.info) includes only four OSV languages out of a collection of 1,377. That seems quite rare, and indeed it is the rarest word order type in WALS. But is it rare enough to support this linguist's claim? There are six possible orders of S, V, and O (3! = 3×2×1 = 6), plus an additional language type with no dominant order, so if all were equally common, there should be about 1,377/7 ≈196.7 of each type in WALS. This is a lot more than the four observed OSV languages, consistent with the claim, but now that we're looking at numbers, here's another one: 189, the number of languages with no dominant order, is also slightly lower than 196.7. A full quantitative analysis would have to say something about this point. We'll say a bit more about this later in this chapter, and even more later in this book (for even more about the quantitative problems that arise in the study of cross-linguistic word order patterns, see Piantados & Gibson, 2014).

Numbers, numbers, numbers! We don't need them if our claims involve pure logic, or we don't have enough data, or we don't have clear concepts to test. But as soon as we are willing and able to make our empirical claims precise, we need numbers.

This is a general point about science (see, e.g., Chalmers, 1999), and assuming linguistics is also a science, it's true about linguistics too. Linguistic numbers are everywhere:

- *Psycholinguistics:* word frequency, reaction times, number of times a subject does something, accuracy in some task, etc
- *Child language:* age, mean length of utterance, number of mistakes, etc
- *Sociolinguistics:* age, proportion of times a speaker uses a variable rule, degree of similarity between dialects, etc
- *Phonetics:* fundamental frequency, duration, voice onset time, etc
- *Corpus linguistics:* word frequency, likelihood of collocations, etc
- *Typology:*number of languages of each type, etc
- *Language teaching:* test scores, degree of improvement, etc
- *Syntax:* relative sentence acceptability, number of people who accept vs. reject a sentence, etc
- *Phonology:* number of patterned vs. exceptional words, probability of one phoneme combining with another, etc

Notice that I include syntax and phonology in this list. There is a common myth that statistics is only relevant to what Chomsky (1965) calls **performance** (language use), not **competence** (grammar). According to this myth, a sentence is either grammatical or it isn't, and a consonant cluster is either allowed in a phonological system or it isn't, so where are the numbers? Woods et al. (1986, p. 1) even start their linguistics statistics textbook by claiming that in "grammar... [t]here appears to be no place ... for statistics", and then go on to focus on other types of linguistics.

In actual practice, however, numbers are just as crucial to grammatical research as to psycholinguistics, phonetics, and the rest of linguistics. After all, there are still only two ways to test a scientific (including linguistic) hypothesis: either you run an **experiment** to generate new data, or else you carry out an **observational** study looking at pre-existing data (i.e., do a kind of **corpus analysis**). Syntacticians mostly do the former: they make up novel sentences and test whether they sound good or bad. Phonologists mostly do the latter: they look at words in a dictionary (a kind of corpus) to see what patterns there are. Both methods involve numbers. Sentences are rarely perfect or impossible, but usually fall somewhere in between, and even sharp judgments can be sharply different across people, so you have to measure the degree of acceptability and cross-speaker acceptance rates. Phonological patterns often have exceptions, so you need to count them to make sure there aren't "too many"; even more than syntacticians, phonologists also increasingly recognize the theoretical importance of quantitative observations (e.g., the consonant cluster /st/ is much more common in English than /sf/, even though /sf/ is not totally banned, as in *sphere*). Thus grammarians are constantly performing informal statistical analyses, even when they don't show their numbers explicitly (see Myers, 2012, for an overview).

3.2 What statistics is for

Quantification is only the first step in doing statistics, however. Consistent with the above discussion about the real-world inspirations for math, the English word "statistics" is related to the word "state": originally statistics was about keeping government records. It still has that record-keeping meaning today, as in "baseball statistics". I don't know the history of the Chinese word 統計學, but from the characters we can see that it also relates to collecting and measuring (and 棒球統計 even has a Wikipedia entry). The closely related field of probability (機率) also has a real-world origin, namely in gambling (both as an intriguing intellectual puzzle and as a practical problem for starving mathematicians). Statistics didn't really take off, however, until the 19th century, with the growth of quantitative biology and especially psychology, with its adoption of laboratory experimentation and surveys. It had another growth spurt in the first half of the 20th century, when biologists started investigating the complexities of genetics and statisticians were drafted in World War II to break codes and predict what targets might be bombed next. Since World War II, the development of ever more powerful computers has spurred the development of ever more powerful statistical techniques, including some that were dreamt of much earlier but had been too complex to use. You can learn a lot more about the history of statistics in Salsburg (2001), a short and easy-to-read book, but the crucial point is that statistics has never been "pure math": it has always been closely linked with real-world problems and concerns.

This brief history also highlights the four main jobs that statistics performs (see Johnson, 2008, p. 3, for a similar list). First, statistics ***summarizes***: it simplifies masses of raw facts so that a general pattern can be understood better. Summarizing methods discussed in this book include making tables, finding averages, measuring variation, and plotting **graphs** - lots and lots of graphs! Graphs are probably the very most useful tool in statistics, since they translate those cold hard numbers into things that the brain understands much better: pretty shapes and colors.

As an example of summarizing, Table 1shows the reaction times (in milliseconds) for fifteen words produced by two participants in one of the psycholinguistic experiments reported in Myers et al. (2006).

Table 1. Some reaction times

Participant 4	445	471	483	449	446	479	569	443	449	568	659	457	674	627	630
Participant 5	941	794	583	588	801	642	632	904	1060	1097	620	831	671	652	849

Are there any patterns in those numbers? It's hard to tell. But suppose that I tell you that the average reaction time for Participant 4 is about 532 ms, and for Participant 5 it is about 778 ms, and that the distributions of the reaction times looks like Figure 2.

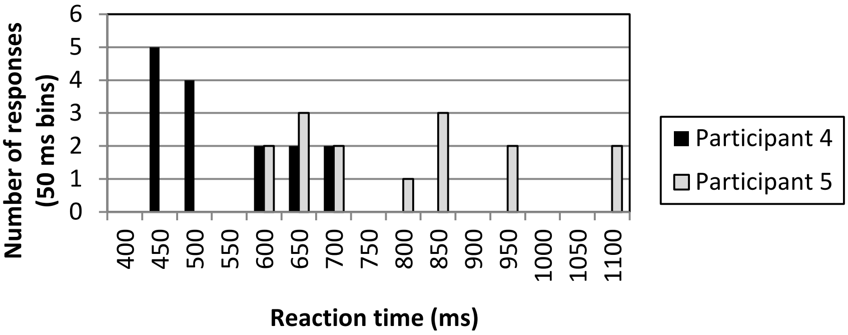

Figure 2. Plotting reaction times in bins (groupings) of 50 ms

Now you can easily see that Participant 4 was generally faster than Participant 5, and was also more consistent in response speed. The patterns are much clearer! (By the way, this graph was made using **Excel**, one of the two statistics programs that you will be using as you work through this book.)

The second job that statistics does for you is to ***compute probabilities***. This helps you determine whether a pattern that you observe in your data is so unlikely to arise by chance that you are justified in accepting the pattern as “real” (at least until you collect more data that changes your mind). In this book we will discuss probability-related methods like randomization, sampling, statistical significance, and adjusting your confidence in a claim as you collect more data. This is where the famous ***p*** **values** come in: *p* stands for *probability*.

As an example of using probabilities, let’s go back to that linguist’s hypothesis about different word orders across the world’s languages. We addressed this issue by saying that, all else being equal, there should be about 1,377/7 ≈196.7 of each of the seven types of language in the WALS **sample**. Why did we divide the total by seven? Because we have the intuition that if there is no bias favoring some language types over others, then the distribution should be even. This is a kind of probability claim: we expect 196.7 languages of each type to appear “by chance”. Of course we don’t really expect there to be exactly 196.7 (what would 0.7 of a language look like anyway?), so we’re willing to overlook slight differences. Indeed, you might have the intuition that 189 is “basically the same” as 196.7, and since there are 189 no-order languages in WALS, this seems consistent with what is expected “by chance” for this language type. By contrast, there are only four OSV languages, and that seems “much rarer” than we would expect “by chance”, just as the linguist claims. Probability theory simply allows us to make these kinds of intuitions more objective.

The third job statistics does is ***modeling***. A model is a “picture” of what you think is going on in your data. Modeling methods that we’ll learn include computing correlations, drawing trend lines in data plots, making predictions, comparing models to see which fits the data best, testing which part of a model is truly necessary to explain the data, and dealing with partially confounded influences. This may sound quite technical now, but in a sense modeling lies at the heart of this whole book, since the only reason we bother with quantification and summarizing and probability is that we want to learn what is “really” going on in our data, and that’s what the model is supposed to represent. Moreover, as we will see, the syntax of writing statistical analyses in **R** (the other statistics program that we will be using) is organized around the logic of modeling. Once you get the syntax figured out for simpler types of models in R, you can make pretty good guesses about how to encode more complex models.

As an example of modeling, consider Figure 3 below, from Myers & Tsay (2015, p. 368). We made this graph by first counting how many times speakers in our Taiwan Southern Min (Taiwanese) corpus utter the discourse marker 著 /tiəʔ˧/ “right”: once (著), twice (著著), three times (著著著), and so on. We did this because we had the hunch that people tend to repeat this word an odd number of times (which is phonologically interesting, at least to us, because it implies that speakers are unconsciously counting with disyllabic feet). To make these counts clearer in the plot, we squished together the higher numbers (the maximum count is 5,774, for one-syllable forms) and stretched out the lower numbers (the minimum count is 1, for eight- and ten-syllable forms), by computing the base-10 **logarithms** (對數: this basically converts a number into how many digits it has, e.g. $\log_{10}(5) \approx 1$, $\log_{10}(50) \approx 2$, etc). Then we plotted these transformed, but real, counts with the black dots and thick line.

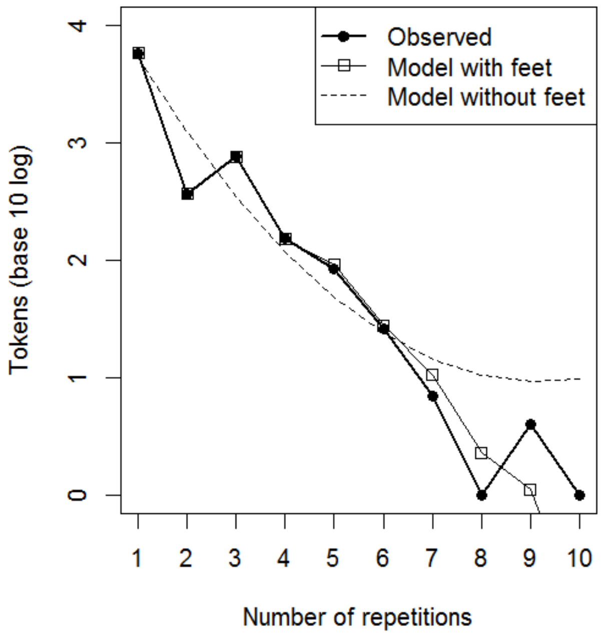

Figure 3. Repetitions of a Southern Min discourse marker

Next we modeled the counts in two different ways. The dashed line shows a model that only assumes that speakers prefer shorter utterances (which is also true, but we suspected that it wasn't the whole story). The white squares show a model that also assumes that speakers prefer odd-numbered repetitions, consistent with our hunch. Just by looking at this graph (made in R, by the way), it's obvious that the odd-numbered model fits the data much better. And indeed, when we did a probability analysis (computing *p* values), we found that this improved fit is unlikely to be due to chance alone. Our hunch seems to be right!

The fourth job that statistics does is ***data exploration***. Logically, this should come before the other three, since it's something you can do the very first time you look at your data, even before you summarize it, compute the probabilities, or try to model it. For example, suppose you have a hunch that the foreign students in your Chinese class may have different learning styles, but you're not sure what the difference is or what might cause it. So you give them a couple tests and look for patterns. One tool you could use is called **cluster analysis**, which helps to visualize which numbers in a set tend to group closer together than others, generating a kind of tree showing the major cluster divisions, and then all the smaller clusters within them. Looking at the cluster tree for the first test, you notice a pattern: most (though not all) of the students in the top-scoring cluster are ... I'm just making up this example, so let's say ... they're mostly from Sweden. But when you look at the second test, which probes a very different skill, you find that the top-scoring cluster mostly consists of the students from ... Zimbabwe. In fact, the Swedish and Zimbabwean students almost switch places in the two tests. Hm, you never noticed that before. What could explain this pattern...? I have no idea, this is a fake example. But now at least you have somewhere to start hypothesizing.

Crucially, by itself a cluster analysis cannot show that the clusters are likely to be "real" rather than "mere chance", or what might predict the clusters even if they are "real". That's why cluster analysis is just for data exploration; to finalize an analysis, you still need to use the three other powers of statistics: summarizing, computing probabilities, and modeling.

The current version of this textbook doesn't have a chapter on cluster analysis and related methods, but hopefully I'll get around to writing one eventually. But we will do plenty of data exploration in a more general sense, using data summaries, particularly in graphs.

3.3 Statistics as a way of looking at the world.

Some people have argued (and I agree) that statistics is not only a useful scientific tool, but also offers a useful philosophical tool for understanding the world more generally. For example, Salsburg (2001) (that statistics history book I mentioned earlier) suggests that statistics revolutionized science by making it possible to study messy things. Early modern scientists, like Newton, believed in a "clockwork universe": the moon moves through the sky just like the hands move around a clock. But look again at the graphs above: any clock that showed so much variation should be thrown out!

The reality is that the world has a lot of **variability** in it, and if scientists want to study reality, they had better know how to study variability. We know today that variability is everywhere, even in the moon's movements (and in the atoms and subatomic particles that make it up). But variability is especially obvious in complex systems like the human brain: that's why biologists and psychologists were the first to get excited by statistics and drive forward its development.

Another important "life lesson" that statistics teaches concerns probability: Accidents happen. The human brain is designed to look for patterns everywhere, but are all of them real? Is that shadow a ghost, or just a shadow? This is just a special case of the point I made before: You can't always trust your intuitions, and sometimes you need to work harder to discover the truth.

As a real-life example of chasing linguistic shadows, consider a case discussed in a book on statistical mistakes by Smith (2014) (look at the title of his book in the reference list: it was inspired by the classic book by Huff, 1954). The case involves the fact that the Chinese word for four (四 *sì*) sounds close to the word for death (死 *sǐ*), leading to a superstition (borrowed by nearby cultures like the Japanese) that four is an unlucky number. Some medical researchers (Phillips et al, 2001) thought that even if the superstition isn't true (and of course it isn't!), Chinese and Japanese Americans might still believe it enough to get stressed out every time the fourth day of the month rolls around. Indeed, they reported that Chinese and Japanese Americans had "significantly" higher cardiovascular (心血管) death rates on that date compared with other dates.

The problem is that their sampling was quite biased, as first pointed out by Smith (2002). There are many different kinds of diseases of the heart and blood vessels, but Phillips et al. (2001) only considered a subset of these; when other types are included, the fourth-of-the-month pattern disappears. There are also different ways that stress might possibly kill you besides heart disease, such as driving you to suicide, but those don't show the pattern either. Worst of all, the original study only looked at a narrow range of years (1989-1998); when more years are included (1969-1988, 1999-2001), the pattern again disappears.

Even more important, perhaps, Smith (2014) points out that the hypothesis itself isn't very plausible. As a Westerner, I always thought a fear of the number four was a rather impractical superstition, since you encounter the number four many times every day. Our Western superstition about "unlucky" 13 is equally ridiculous, but at least you almost never have to worry about running into 13. So why aren't Chinese and Japanese Americans keeling over with fright every day at 4 pm, or when they drive their cars (four wheels), or pet their dogs (four legs)? Of course it's not impossible for a purely empirical study to reveal a hitherto unknown causal force in the universe, but if your proposed mechanism seems implausible from the start, you'd better make sure your evidence is extremely solid.

My final word of wisdom on statistics is that you shouldn't think of it as a kind of decoration that you add to your paper at the last minute, just to conform to some sort of social convention in scientific writing. Instead you should take statistics into consideration from the very beginning of your research.

Why? First, as we saw above, sometimes statistics isn't even relevant: you've got to make sure that quantification even makes sense for your research question. If you're counting or measuring things, then statistics is relevant, but you still have to make sure you know what those "things" are, and why counting or measuring them is going to tell you anything related to your research interests. Don't throw in numbers just to show off.

Second, there's no magic computer program to turn bad research into good. You still need to collect as much good data as you can, since common sense says that more data is more convincing than not enough (and this intuition is borne out by the logic of probability as well). But even a huge pile of data can be useless if it's badly collected. For example, if you hypothesize that Chomsky's (1957) famous sentence "Colorless green ideas sleep furiously" is grammatical, and then ask a bunch of ordinary English speakers to judge it on a scale from 1 (worst) to 7 (best), maybe you'll get an average score significantly above 3.5 (the midpoint). But even if this happens, all you would know is that your participants aren't flipping a coin to make their choice. They may still be giving higher-than-chance scores only because they're trying to be nice; maybe they would tend to accept *any* sentence, even the pure word salad of

"Furiously sleep ideas green colorless"! This is a realistic risk with experiments on children, who tend to say "yes" to scary teacher-like people (including child language researchers), no matter what you ask them (see e.g. Crain &Thornton, 1998). So at the very least, you need to include an ungrammatical sentence in your experiment for comparison, and then you can run statistical tests to see how different the two acceptability scores are. Such a **control condition** is always necessary for proper **experimental design**. So if you study mistakes made by foreigners in a second language study, you still need to test native speakers as well, since you can't be sure that your textbook or your own intuitions reflect how ordinary native speakers actually talk. Similarly, child language studies need to test adults too.

Finally, you also need to know how to choose the right kind of statistics for your particular kind of data. For example, the model shown in the Southern Min graph above uses something called Poisson regression (after a guy named Poisson): **regression analysis** (迴歸分析; don't worry about the weird name yet) is a general class of model, and Poisson regression relates to count data, which is what our data were. It turns out that if we instead try to model the same data with **linear regression**, which is designed for continuous values (like reaction times or consonant durations), the fit is much worse: this is the wrong kind of model to use here. Of course, you could just try to memorize rules like "test A is for data type B", but in the messy world of real life, it is wiser to form deeper intuitions about *why* different tests work better for different types of data.

4. Preview of the rest of the book

Now that we got the comforting and the philosophy out of the way, what happens next? In the rest of this book I try to help you climb up the very practical but somewhat abstract tower of statistics, one step at a time. This means that just like a murder mystery, you can't skip to the end of this book, not because you'll find out who the killer is too soon, but because you probably won't understand the plot at all. We have no choice but to start with the fundamental logic of statistics, then learn some basic statistical methods, and finally learn some more complicated methods. Each step builds on the previous one.

But don't worry, I'll keep it concrete and language-related all the way through. Each chapter section starts with a concrete linguistic example, showing how to actually do it using Excel and/or R. Only after that will I tell you the mathematical secrets of how it works, and hopefully help you to develop your own mathematical intuitions. I'll also give you a "cheat sheet" summarizing all of the Excel and R commands used in the book (plus more!), so you can try to apply the techniques to the homework and your own data (currently it's kept at http://personal.ccu.edu.tw/~lngmyers/StatsFunctions.pdf and in the E-Course system, but my homepage is currently not stable so maybe the E-Course system is a better place to look).

The first five chapters give you the fundamentals. You're almost done with this one (chapter 1). Chapter 2 introduces you to Excel and R, the two programs we will be using in this book. Chapter 3 reminds you of the familiar concepts of averages and variability, and shows how to use them to analyze real data and formalize them mathematically. Chapter 4 reviews basic probability theory and shows how to use it to test hypotheses about real data (this is where I finally explain *p* values). Chapter 5 introduces the idea of modeling, by building on familiar ideas like correlation and linear equations (yes, you probably forgot all about linear equations, but I'll review them for you).

The middle of the book explains all the famous tests that you've seen in papers and heard about from your friends. So chapter 6 is about *t* tests, chapter 7 is about chi-squared tests, and chapters 8 and 9 are about ANOVA. I won't go into details here, but all of these methods turn out to be closely related to each other (among other things, they all build on another concept you may have heard of: **the bell curve**), and most importantly, they all apply the concepts of summarizing, probability, and modeling.

The remainder of the book is about more advanced topics, though you may have already seen some of them in linguistic studies too. Chapter 10 is about linear regression ("ordinary" regression), and chapter 11 is about logistic regression (commonly used in sociolinguistics, among other places; I'll also explain the much-less-used Poisson regression here). Chapter 12 is about a relatively new kind of regression modeling called mixed-effects modeling, which has recently become very popular (including in linguistics) because it finally makes it possible to deal with some annoying real-world data problems (often encountered in linguistics), and because home computers are finally powerful enough to compute it.

Finally, chapter 13 covers something called Bayesian statistics (after a guy named Bayes). This is an alternative approach to statistics, different in many ways from the traditional approach explained in chapters 6 through 12, and so it is almost never taught in introductory statistics textbooks. Nevertheless, I think it's important even for beginners to know about it because more and more scientists (including linguists) are using it, even preferring it to traditional statistics. Bayesian statistics is growing in popularity not only because computer power is finally available to process it, but also because it enables you to quantify how new data can change your mind, rather than merely comparing your results with chance, as traditional statistics does.

So that's what you have to look forward to. Let's get going!

References

Ashcraft, M.H. (2002). Math anxiety: Personal, educational, and cognitive consequences. *Directions in Psychological Science, 11*, 181-185.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.*

Cambridge, UK: Cambridge University Press.

Brown, J. D. (1988). *Understanding research in second language acquisition: A teacher's guide to statistics and research design*. Cambridge, UK: Cambridge University Press.

Chalmers, A. F. (1999). *What is this thing called science?* (Third edition). University of Queensland Press.

Chomsky, N. (1957). *Syntactic structures*. Mouton de Gruyter.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Crain, S., & Thornton, R. (1998). *Investigations in Universal Grammar: A guide to experiments in the acquisition of syntax and semantics*. Cambridge, MA: MIT Press.

Dehaene, S. (1997).*The number sense: How the mind creates mathematics*. New York: Oxford University Press.

Denzin, N. K., & Lincoln, Y. S. (2011). *The Sage handbook of qualitative research*, 4th edition. Los Angeles: Sage Publications.

Donohue, M. (2002). Tobati. In J. Lynch, M. Ross, & T. Crowley (eds.), *The Oceanic languages* (pp. 186-203). Richmond: Curzon.

Eddington,D. (2015). *Statistics for linguists: A step-by-step guide for novices*. Cambridge Scholars Publishing.

Gomez, P. C. (2013). *Statistical methods in language and linguistic research*. Equinox.

Gries, S. T. (2013). *Statistics for linguistics with R: A practical introduction* (2nd edition). Berlin: De Gruyter.

Hasko, V. (2012). Qualitative corpus analysis. In C. A. Chapelle (Ed.) *The encyclopedia of applied linguistics*. Wiley.

Haspelmath, M., Dryer, M.S., Gil, D., & Comrie, B. (Eds.) (2005). *The world atlas of language structure*. Oxford: Oxford University Press. Database at http://wals.info.

Hatch, E. and Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics.* Newbury House Publishers.

Heigham, J., & Croker, R. A. (2009). *Qualitative research in applied linguistics: A practical introduction*. Palgrave Macmillan.

Hersh, R. (1999). *What is mathematics, really?* Oxford: Oxford University Press.

Huff, D. (1954).*How to lie with statistics*. Norton, New York. Updated edition published in 1991 by Penguin.

Jockers, M. (2014). *Text analysis with R for students of literature*. Springer.

Johnson, K. (2008). *Quantitative methods in linguistics*. Wiley.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing, second edition*. Upper Saddle River, NJ: Pearson.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Lakoff, G., & Núñez, R. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being.* New York: Basic Books.

Larson-Hall, J. (2015). *A guide to doing statistics in second language research using SPSS and R* (second edition). Routledge.

Levshina, N. (2015).*How to do linguistics with R: Data exploration and statistical analysis*. John Benjamins.

Myers, J. (Ed.). (2012). *In search of grammar: Empirical methods in linguistics*. Language and Linguistics Monograph Series 48. Taipei: Institute of Linguistics, Academia Sinica.

Myers, J., & Tsay, J. (2015). Trochaic feet in spontaneous spoken Southern Min. In Hongyin Tao, Yu-Hui Lee, Danjie Su, Keiko Tsurumi, Wei Wang, & Ying Yang (Eds.), *Proceedings of the 27th North American Conference on Chinese Linguistics, Vol. 2*, 368-387. Los, Angeles: UCLA.

Myers, J., Huang, Y.-C., & Wang, W. (2006). Frequency effects in the processing of Chinese inflection. *Journal of Memory and Language, 54* (3), 300-323.

Phillips, D. P., Liu, G. C., Kwok, K., Jarvinen, J. R., Zhang, W., & Abramson, I. S. (2001). The Hound of the Baskervilles effect: Natural experiment on the influence of psychological stress on timing of death. *British Medical Journal, 323*(7327), 1443-1446.

Piantadosi, S. T., & Gibson, E. (2014). Quantitative standards for absolute linguistic universals. *Cognitive Science, 38*(4), 736-756.

Plonsky, L. (2015). *Advancing quantitative methods in second language research*. Routledge.

Rasinger,S. M. (2014). *Quantitative research in linguistics: An introduction* (2nd edition). Bloomsbury Academic.

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. New York: W. H. Freeman & Company. 葉偉文譯 (2001) 《統計，改變了世界》天下文化。

Smith, G. (2002). Scared to death? *British Medical Journal, 325*(7378), 1442-1443.

Smith, G. (2014). *Standard deviations: Flimsy theories, tortured data, and other ways to lie with statistics*. New York: Overlook Duckworth.

Winter, B. (2019). *Statistics for linguists: An introduction using R*. Routledge.

Woods, A., Fletcher, P., & Hughes, A. (1986). *Statistics in language studies.* Cambridge University Press.