

# 1 Introduction

## Statistics Meets Corpus Linguistics

### 1.1 What Is This Chapter About?

This chapter introduces basic principles of statistical thinking that are necessary for informed application of statistical procedures to corpus data. It starts with an explanation of the role of statistics in scientific research in general and corpus linguistics in particular. After that, more specific topics such as the creation of corpora, types of research design, basic statistical terminology, as well as data exploration and visualization are discussed. The chapter ends with a case study demonstrating the use of statistics in corpus research.

In particular, we'll be exploring answers to five questions:

- What is the role of statistics in science and corpus research? (Section 1.2)
- What are the key terms in corpus statistics? (Section 1.3)
- How do we build and analyse corpora? (Section 1.4)
- How can we explore and visualize data? (Section 1.5)
- How can statistics be used in corpus research? (Section 1.6)

### 1.2 What Is Statistics? Science, Corpus Linguistics and Statistics

#### Think about . . .

Before reading this section, think about the following questions:

1. What is science? What are the basic features of scientific enquiry?
2. Which of these statements about language are scientific statements?
  - (a) Women's speech seems in general to contain more instances of 'well', 'y'know', 'kinda', and so forth . . .
  - (b) Words are easy, like the wind.
  - (c) Passives are most common by far in academic prose [compared to other registers], occurring about 18,500 times per million words.
  - (d) The faculty of language can reasonably be regarded as a 'language organ'.

*(See next page for one more example)*

- (e) Our results show that there were significant changes in at least one formant<sup>1</sup> for 10 of 11 vowel sounds and in both formants for 5 of 11 vowel sounds from the 1950s to the 1980s Christmas broadcasts . . . We conclude that the Queen no longer speaks the Queen’s English of the 1950s . . .

Unlike other sources of information such as mythology, philosophy or art, **science** relies on the systematic collection of empirical data and testing of theories and hypotheses. One of the most influential theoreticians of science, Karl Popper, defined a scientific statement or theory as something that can in principle be falsified (Popper, 2005 [1935]). In other words, we can call a statement or theory scientific only if it can be tested empirically. This means that we need to collect data and evaluate if the data is consistent with our theory. If not, we can say that the available evidence contradicts the theory. When we look at the statements in question 2 of the ‘Think about’ task we can see that they vary considerably in whether they can be put to a test by collecting empirical evidence: clearly, statements (c) and (e) can be considered scientific.<sup>2</sup> Not only can they be empirically tested, these statements are already accompanied by empirical evidence. On the other hand, the poetic statement (b) expresses a metaphor, which despite its power would be difficult to test by collecting data. Statement (a), which is taken from Lakoff’s (1975) book *Language and Woman’s Place*, can be empirically tested (indeed numerous researchers have tested it), yet the author herself offers little empirical evidence in her book beyond anecdotes. Statement (d), which comes from Chomsky (2000), proposes a view of language that relies more on philosophical (pre-empirical) understanding of the human language faculty and does not necessarily seek empirical confirmation. In sum, statements about language which provide direct reference to systematically<sup>3</sup> collected empirical evidence can be considered scientific.

**Corpus linguistics** is a scientific method of language analysis. It requires the analyst to provide empirical evidence in the form of data drawn from language corpora in support of any statement made about language. Another scientific requirement corpus linguists follow in principle is replicability of results. This means that researchers need to be able to confirm the findings of one study in follow-up studies (see Section 8.3). In order for the results to be replicable, corpus linguists need to make their choice of corpora and analytical techniques transparent. It is also good practice in corpus linguistics to make corpora available to other researchers who can explore the same dataset further and thus advance knowledge in the field.<sup>4</sup>

<sup>1</sup> Formant is a component of a vowel sound in phonetic research.

<sup>2</sup> Sources of statements: (a) Lakoff 1975: 53; (b) Shakespeare? 1992 [1599]: 269; (c) Biber et al. 1999: 476; (d) Chomsky 2000: 4; (e) Harrington et al. 2000: 927.

<sup>3</sup> Empirical research can be both qualitative (descriptive and interpretational) and quantitative (using numbers). These areas of research are complementary.

<sup>4</sup> Unfortunately, sometimes corpora are ‘locked’ behind corporate walls with unclear principles of how these corpora were constructed. This makes their use difficult for any serious scientific

In essence, corpus linguistics is a quantitative methodology; this means that corpus linguistics typically works with numbers which reflect the frequencies of words and phrases in corpora (McEnery & Hardie 2011.) For this reason, statistics is crucial for corpus linguists because it helps us work effectively with quantitative information. There are many different understandings of what statistics is. In this book, we will be working with the following definition: **statistics** is a discipline which helps us make sense of quantitative data; in other words, statistics is a ‘science of collecting and interpreting data’ (Diggle & Chetwynd 2011: vii) that can be counted, measured or quantified in some way. One of the important tools in statistics is the use of mathematical expressions – that’s why we’ll be looking at various equations in this book. Mathematical expressions help us understand complex and fuzzy reality through capturing important features of the data by means of numbers and symbols, making it possible to handle the data easily in the process of analysis.

Let us have a look at two examples that illustrate this point. First, imagine that we are interested in the number of adjectives different British fiction writers use in their texts. We might hypothesize that using more adjectives leads to more colourful descriptions in novels. We have randomly selected 11 fiction texts by different authors from the *British National Corpus* (BNC) and counted the number of adjectives in each text; this is their absolute frequency (see Section 2.3). We have then normalized the absolute frequency for comparability.<sup>5</sup> In statistics, we call our 11 texts a **sample**. The following are the relative frequencies of adjectives per 10,000 words:

508, 542, 552, 553, 565, 567, 570, 599, 656, 695, 699

However, showing a long list of results is not a very efficient way of dealing with quantitative data – imagine what would happen if we had to list 100 or 1,000 different results. Instead, we can use a very simple statistical measure to summarize our findings. This measure is called the **mean** and gives us an average value which represents a whole range of values. The mean for the numbers above is 591.45.

The mean is calculated in the following way:

$$\text{mean} = \frac{\text{sum of all values}}{\text{number of cases}}$$

exploration and must lead to doubt being cast on claims made using such corpora. If corpus linguistics wants to retain its scientific status, it should not be content with statements such as ‘this feature was found in a large corpus that is, however, not available’.

<sup>5</sup> Because the texts are of different length, we have taken the relative frequencies per 10,000 words to show how many adjectives on average each author uses in 10,000 words (see Section 2.3 for the explanation of relative frequency). The relative frequencies have been rounded to the nearest integer.

Applied to the dataset above:

$$\begin{aligned} \text{mean} &= \frac{508 + 542 + 552 + 553 + 565 + 567 + 570 + 599 + 656 + 695 + 699}{11} \\ &= 591.45 \end{aligned} \quad (1.1)$$

Because the mean describes our sample, it is part of what we call **descriptive statistics**. Another example of a mathematical representation of complex linguistic reality is a line, in statistics called a **regression line** or **line of the best fit** (see Chapter 4 for an explanation of regression models). Imagine that we are interested in whether the authors that use more adjectives also use more verbs. We can list the frequencies of verbs<sup>6</sup> just below the frequencies of adjectives to see whether there is any relationship between these two linguistic features:

508, 542, 552, 553, 565, 567, 570, 599, 656, 695, 699  
 2339, 2089, 2056, 2276, 2233, 2056, 2241, 1995, 2043, 1976, 2062

However, a better way of finding out whether there is a relationship between the use of adjectives and verbs is to display these numbers in a graph (see Section 1.5 on how to create graphs).

The graph in Figure 1.1 shows a clear tendency marked by the regression line. The regression line points to the fact that the number of verbs and adjectives in the sample is in an inversely proportional relationship – the more adjectives

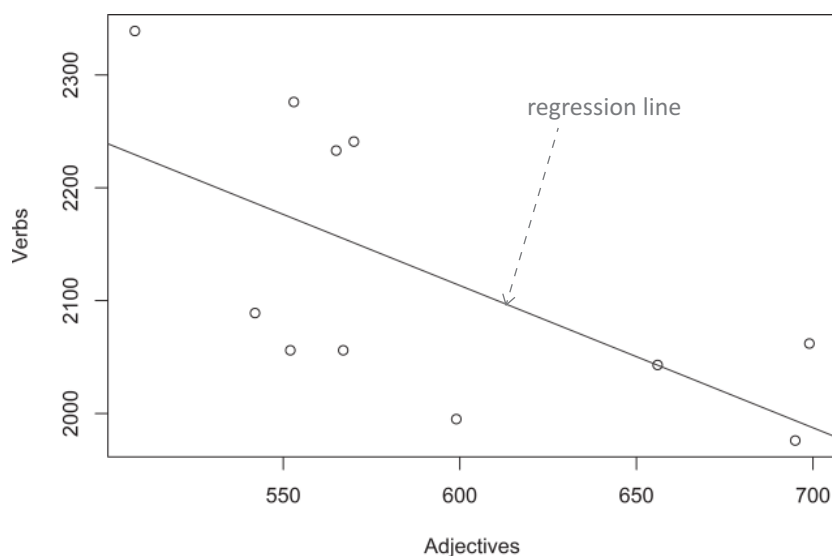


Figure 1.1 *The relationship between the relative frequency of adjectives and verbs*

<sup>6</sup> Again, shown as relative frequencies per 10,000 words rounded to the nearest integer.

authors use, the fewer the verbs they employ and vice versa. The line is plotted in such a way as to find the best fit for all the individual data points (marked as dots in the graph). It is by sheer coincidence that one of the points actually lies exactly on the line; often the line does not go through any of the actual data points because it is a mathematical abstraction representing the dataset as a whole. The purpose of this mathematical model of reality is to tell us something interesting about the data that we wouldn't necessarily notice if we looked at the individual data points in isolation. These two examples demonstrate the main point of statistical thinking that will appear in various forms throughout the book: statistics in corpus linguistics is about mathematical modelling of a complex linguistic reality. It can help us discover and elucidate patterns and tendencies in the data that might otherwise remain hidden.

### 1.3 Basic Statistical Terminology

#### Think about . . .

Before reading this section, think about the meaning of the following terms. Have you heard them before? If so, in what context? Would you be able to define them?

- assumption
- case
- confidence interval
- dataset
- dispersion
- distribution
- effect size
- normal distribution
- null hypothesis
- outlier
- p-value
- robust
- rogue value
- statistical measure
- statistical test
- standard deviation
- variable

The following is an overview of basic statistical terminology used in this book. It includes key terms with examples from corpus research and is ordered from basic concepts to more complex ones which rely on the understanding of the previous terminology. Mastering these terms will make reading of the rest of the book, and many papers in corpus linguistics, much easier.

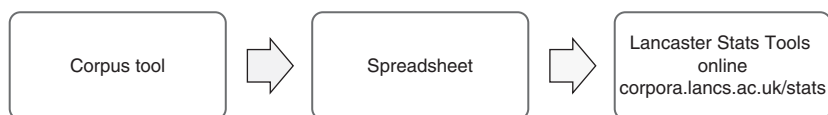


Figure 1.2 *Process of statistical analysis*

**Corpus** (pl. corpora) is a specific form of linguistic data. It is a collection of written texts or transcripts of spoken language that can be searched by a computer using specialized software. A corpus usually represents a **sample** of language, i.e. a (small) subset of the language production of interest; in some limited cases of very specialized corpora, a corpus can include the whole **population**, i.e. all language of interest to the researcher (see Section 1.4). The software that is used to search a corpus usually implements basic types of statistical analysis such as the statistical identification of collocations and keywords (see Chapter 3). For more sophisticated statistical analyses, however, we usually need to use appropriate statistical packages. This book uses Lancaster Stats Tools online, free statistical tools available from the companion website. Figure 1.2 outlines the process of analysis with Lancaster Stats Tools online.

Note that preparing the spreadsheet in the appropriate format is as important as the statistical analysis that follows. The book offers many examples of datasets based on different corpora, which are suitable for different types of analysis. It is always useful to compare your data to the model examples provided (full datasets are available from the companion website) to see if your data is in the appropriate format.

**Dataset** is a series of corpus-based findings that can be statistically analysed. It is a systematic collection of individual results that can be stored in the form of a table in a spreadsheet program (e.g. Excel, Calc etc.), each line representing an individual **data point** or **case** and each column representing a separate **variable**. Figure 1.3 provides an example of a dataset with five variables and multiple cases, each case representing one speaker. Note that example datasets used in this book are available at the companion website. It is important to study them for the particular ‘shape’ of data that lends itself to particular types of statistical analyses.

**Variable**, as the name suggests, is something that can vary and take on different values. For example, speaker’s age is a variable that can take on different **values** from about one year (when children typically learn their first words) to over 100. Much corpus research can be characterized as searching for variables in corpora and analysing the relationship between them. An important distinction needs to be made between linguistic variables and explanatory variables. **Linguistic variables** capture frequencies of linguistic features of interest in the corpus. **Explanatory variables** (sometimes called ‘independent variables’) capture contexts in which the linguistic features appear. For instance, an

	explanatory variables		linguistic variables				
	gender	proficiency	I	you	pers_pronouns_all		
2	6_SP_51	0	1	38.75969	9.302326	80.62015504	
3	6_SL_7	0	1	33.46856	19.26978	100.4056795	
4	8_ME_24	1	2	39.10112	38.65169	129.8876404	case
5	8_IT_28	1	2	51.98181	11.04613	122.8070175	
6	8_IT_14	0	2	33.41584	8.663366	108.9108911	value
7	IT_65	1	3	37.127	19.43095	100.9715475	
8	7_CH_17	0	2	58.64198	23.91975	100.308642	
9	7_ME_6	1	2	42.48573	10.14585	119.2136969	
10	6_CH_15	0	1	56.12245	22.95918	145.4081633	
11	IT_54	1	3	25.81369	19.01969	101.010101	
12	6_ME_2	1	1	34.90401	33.15881	108.2024433	
13	6_CH_25	1	1	47.82147	11.68969	145.5897981	
14	CH_6	1	3	52.44601	25.1212	121.6394888	
15	6_IN_3	1	1	29.83599	26.74807	131.6872428	

gender: 0... male, 1... female; English proficiency: 1...pre-intermediate, 2...intermediate, 3... advanced

Figure 1.3 Example of a dataset

explanatory variable can be the genre/register or date of publication of a text as well as speaker's age, gender and language proficiency, to name only a few. The dataset in Figure 1.3, which comes from the *Trinity Lancaster Corpus* of spoken L2 production (Gablasova et al. 2017), contains two explanatory (gender and language proficiency) and three linguistic variables (relative frequencies of *I*, *you* and all personal pronouns together).

Variables (both linguistic and explanatory) can be either nominal, ordinal or scale variables. A **nominal variable** has values that represent different categories into which the cases in a dataset can be grouped; there is no order or hierarchy between the categories. Speaker's gender is an example of a nominal variable because we can assign speakers in the dataset to one of two groups: (1) male speakers and (2) female speakers. There is no hierarchy in this classification. For convenience, we often use numbers to indicate the group membership. In the dataset in Figure 1.3, 0 stands for 'male speaker' and 1 for 'female speaker' but these numbers have no inherent value; they are just a shorthand for longer labels. We could just as well have used 1 (or any unique number) for indicating the male speakers and 0 (or any unique number) for indicating the female speakers. An **ordinal variable** is similar to the nominal variable in that it groups cases into distinct categories; the categories, however, can be ordered according to some inherent hierarchy. For example, speaker's proficiency in a foreign language is an ordinal variable because we can rank speakers according to their proficiency from beginners to advanced speakers. In the dataset in Figure 1.3, 1 indicates a lower proficiency than 2 and 2 indicates a lower proficiency than 3. Finally, a **scale variable** is a quantitative variable because it can take on any value on a scale showing the quantity of a particular feature; this also means that values taken on such


scales can be added, subtracted, multiplied and divided, because they represent measurable quantities, not just rank orders.<sup>7</sup> In the case of linguistic variables, the scale shows the relative frequencies of a linguistic feature in different texts, speakers or subcorpora in a corpus. For example, the numbers indicating the relative frequencies per 1,000 words of the first-person pronoun *I* in Figure 1.3 are values of a scale variable. In fact, all three linguistic variables in the dataset in Figure 1.3 are of this type.

The dataset in Figure 1.3 can be used to explore different types of research questions. For instance:

- Is there a relationship between speaker's gender (a nominal explanatory variable) and the use of personal pronouns (a scale linguistic variable)?
- Does a speaker's English proficiency (an ordinal explanatory variable) have an effect on the use of the first-person pronoun (a scale linguistic variable)?
- Is there a relationship between the use of the first-person and the second-person pronouns (both of which are scale linguistic variables)?

The **frequency distribution** of a variable provides information about the values a variable takes and their frequencies. Distributions of scale variables can be shown in a histogram (see Section 1.5). Figure 1.4 displays the distribution of the first-person pronoun from the dataset in Figure 1.3. The x-axis lists different frequency bands of the linguistic variable, in this case the first-person pronoun, per 1,000 words, while the y-axis shows the number of cases in the dataset for each frequency band. Thus, for example, the graph shows that in the corpus there were 19 texts (speakers) where the first-person pronoun was used 10 times or less per 1,000 words (this information is indicated by the first bar from the left), 88 texts where it appeared 11–20 times (second bar from the left), 214 where it occurred between 21 and 30 times (third bar from the left) etc.

As a benchmark in statistics, one of the common distributions – the **normal distribution**<sup>8</sup> – is often used. The shape of the normal distribution is a symmetrical bell as shown in Figure 1.5.

Although a lot of data in the natural and social world follows the normal distribution, most linguistic data is positively skewed () , i.e. there is more data to the left of the distribution than the right, as we saw, for example, in Figure 1.4. Distributions in statistics are crucial because they provide an overview of the data, which indicates what statistical techniques are appropriate to use. The shape of the distribution thus plays an important role in the assumptions of different statistical procedures (see 'assumptions' below).

<sup>7</sup> The label 'scale variable' subsumes interval (without a meaningful zero point) and ratio (with a meaningful zero point) variables, which are distinguished for some purposes; the distinction is not essential for corpus analysis.

<sup>8</sup> 'Normal' here is used in a technical sense as a label introduced by Pearson (1920: 25) for a specific statistically important distribution; there is nothing abnormal about other types of distributions.



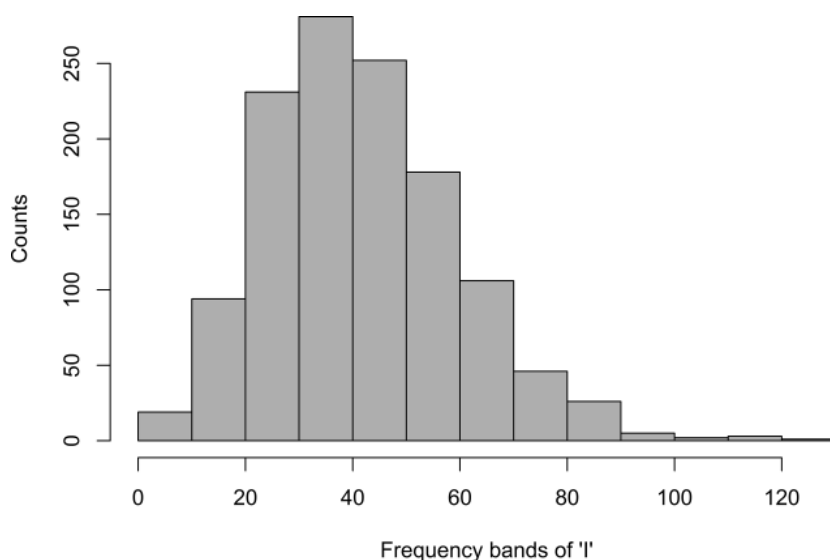


Figure 1.4 *The distribution of the first-person pronoun in the Trinity Lancaster Corpus*

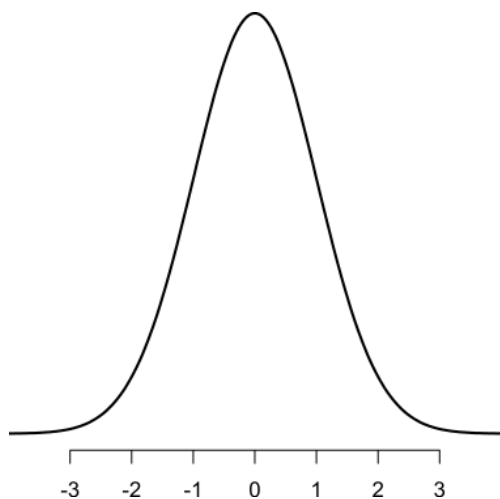


Figure 1.5 *Standard normal distribution*

**Outlier or rogue value?** When we look at distributions we often check for outliers. **Outliers** are extreme values, i.e. values that are very far from the other values. Section 1.5 will introduce boxplots, a useful means of identifying outliers. When we find an outlier we need to check if the outlier is a genuine value or a measurement error – a so-called **rogue value**. A rogue value can be caused, for instance, by mistyping data in a spreadsheet or by a tagging error in the corpus.

An outlier, instead, is a valid data point, which for some reason stands out from others. While outliers are not in themselves ‘errors’, they present problems for statistical models because they may obscure the general tendency (see ‘measure of central tendency’ below) in the data and the researcher must decide how to go about the analysis of data which includes outliers. If there is a good reason, outliers can be excluded (bracketed out) from part of the analysis that focuses on the central tendency in the data.

The **measure of central tendency** or ‘average’ provides one summary value for a series of values of a scale variable. It is a simple statistical model that is usefully paired with dispersion (see below) to complete the summary description of the data. Different types of average can be used. In corpus linguistics the most useful ones are: mean, median and 20% trimmed mean. **Mean (M or  $\bar{x}$ )**, as we have already seen, is the sum of all values divided by the number of cases (see Section 1.2). The mean is a useful measure in distributions which do not have extreme values (outliers) that sway the mean towards them; in distributions with outliers, the mean might represent the outlier more than the rest of the values, which leads to the mean failing to be a useful model. Take for instance the frequency of adjectives in 11 fiction texts taken from the *British National Corpus* (BNC) used as an example to calculate the mean in Section 1.2.

508, 542, 552, 553, 565, 567, 570, 599, 656, 695, 699

The mean of these 11 values is 591.45. However, imagine what would happen if the last value in the series was 6,990 instead of 699. In this case, the mean would be pulled towards the extreme value and we would get 1,163.36; this number is a poor model for the data because only one out of 11 values is above 1,000. One way around this problem of the sensitivity of the mean to outliers is to use the median instead. The **median (mdn)** is the middle value in a series of values ordered from the smallest to the largest. For our 11 values the median is 567, as can be seen from the illustration below.

508, 542, 552, 553, 565, **567**, 570, 599, 656, 695, 699

The median will always stay in the middle of the distribution regardless of what happens at the periphery i.e. whether we have 699 or 6,990 or even 69,900 as the maximum.

If we had ten instead of eleven values, which is an even number, the median would lie half way between the two central values 565 and 567, as demonstrated below.

508, 542, 552, 553, **566**, **567**, 570, 599, 656, 695

The general rule for the median is this: the median is the middle value in the case of an odd number of values; in the case of an even number of values, the