CHAPTER 1

What is statistics?

Main statistical notions and principles

What you will learn from this chapter:

> What is statistics? What can and cannot statistics do for you? How to formulate and test research hypotheses? What kind of statistical tests are there? These and many other questions are discussed in this chapter. In addition, you will also learn about different types of variables, parametric and non-parametric tests, *p*-values and many other things which you will need in order to understand explanations provided in the following chapters.

1.1 Statistics and statistics

Not many people know that the English word **statistics** is in fact ambiguous. This can cause some confusion. On the one hand, it is a mass noun in the singular, which designates a discipline, like mathematics, physics or economics. Broadly speaking, statistics is a set of techniques and tools for describing and analysing data. On the other hand, the word is often used in the plural to refer to values derived from a sample, as opposed to the entire population. A **population** is a group that represents all objects of interest. For example, if we want to compare the speech rate of speakers of Dutch spoken in Belgium and in the Netherlands, the corresponding population will include all Dutch speakers in the two countries. The values obtained from a population are called **parameters**. An example of a parameter is the mean speech rate of all Belgian Dutch speakers. Of course, it would be too time-consuming, expensive and tedious to measure the speech rate of every Belgian Dutch and Netherlandic Dutch speaker. This is why linguists normally deal with **samples** which, as they hope, are representative of the population. So, **statistics** in the plural are measures obtained from samples. An example of a statistic is the average speech rate obtained for a sample of one hundred Belgian Dutch speakers.

Statistics (both in the singular and in the plural) can be subdivided into descriptive and inferential. **Descriptive statistics** can be used to describe the characteristics of a sample. An example is the above-mentioned average speech rate in the sample of one hundred Belgian Dutch speakers. **Inferential statistics** allows the researcher to use the characteristics of a sample in order to make conclusions about the population in general. If we compare the average speech rates of Belgian Dutch and Netherlandic Dutch speakers,

inferential statistics can tell us if the difference is statistically significant or it can be merely attributed to chance. Inferential statistics allows us to make an amazing leap from the sample to the entire population and infer to all Dutch speakers in the two countries, given the results based on a small sample. If not for inferential statistics, one would have to measure the speech rate of all Dutch speakers. And now imagine how much time and money you could save if you had to carry out a similar study on speakers of English or Chinese.

Certainly, in order to make this leap from a sample to the population, one has to be sure that the sample is representative of the population. The difference between a sample statistic and the corresponding population parameter is called the **sampling error**. The smaller the sampling error, the closer the sample represents the characteristics of the population. The higher the sampling error, the more difficult it will be to extend the results of your study to the population. This is why it is very important that your sample should represent the population as closely as possible. The best sampling method is random sampling. This means that every member of the population has equal chances to be selected. In the speech rate example, we would have to make a random selection from the list of all Belgian and Netherlandic Dutch speakers. Of course, this approach is not always feasible, although it is considered the most reliable. Alternatively, one can use so-called representative and convenience sampling. Representative sampling means that the researcher draws a sample in such a way that it matches the population on certain characteristics. For example, one may create a sample of Belgian Dutch speakers by reproducing the same ratios of men and women, older and younger speakers, different ethnic groups and dialects, etc. This method is slightly inferior to random sampling because there is risk of omitting some important variables. Finally, there is convenience sampling, the least reliable, but probably the most widely used method. One can simply make recordings of different speakers of Dutch only in a few easily accessible cities. Of course, the less random the sampling procedure, the higher the risk of a bias.

1.2 How to formulate and test your hypotheses

1.2.1 Null and alternative hypotheses

Before beginning any statistical analysis, it is necessary to formulate a research hypothesis. It is an 'educated guess', which usually posits some difference between groups or relationship between variables. Consider the example with the speech rate. You might have heard several Netherlandic and Belgian Dutch speakers, and you might feel that the former speak a bit faster than the latter. The hypothesis would be then that Dutch speakers in the Netherlands speak on average faster than Belgian Dutch speakers (in fact, this is exactly what was found by Verhoeven et al. [2004], who also controlled for other demographic factors). The research hypothesis, which is also called, probably counterintuitively, the **alternative hypothesis**, always goes together with the **null hypothesis**, which says that there is

no difference between different groups, or no association between different variables, etc. In our example, the null hypothesis is that there is no difference in the average speech rates between the Dutch speakers in these two countries. Below are a few other examples:

A. H_0 (the null hypothesis): There is no difference in the number of lexemes that denote snow in Eskimo and Yucatec Maya.
   H_1 (the alternative hypothesis): There are more lexemes that denote snow in Eskimo than in Yucatec Maya.
B. H_0 (the null hypothesis): there is no relationship between the frequency of a word and how fast it is recognized in a lexical decision task.
   H_1 (the alternative hypothesis): the more frequent a word, the faster it is recognized in a lexical decision task.
C. H_0 (the null hypothesis): there is no difference in the relative frequencies of metaphoric expressions used by men and women when they speak about sex.
   H_1 (the alternative hypothesis): there is a difference in the relative frequencies of metaphoric expressions used by men and women when they speak about sex.

The alternative hypotheses in A and B are called **directional**. They posit a direction to the inequality assumed by the researcher because they contain expressions like 'X is more than Y' and 'the greater X, the greater Y'. In contrast, the alternative hypothesis in C is **non-directional**. The researcher has no expectations about the frequencies of metaphoric expressions used by men and women. She simply expects to find a difference between the sexes. It can be expressed as 'X does not equal Y'.

Why do we need the null and alternative hypotheses? This is because contemporary science is based on the logic of falsification. It is impossible to prove that something is right, but it is possible to reject the opposite. Consider an example. According to Universal Grammar, all human languages have recursion, i.e. one can embed clauses within sentences endlessly, e.g. *This is the cat that killed the rat that ate the malt that lay in the house that Jack built*… It is impossible to prove that all human languages have recursion. After all, even if one checks all living languages, there is no way to analyse the extinct ones and future ones. However, it is possible to falsify the hypothesis if you find at least one language that does not have recursion. In fact, it seems that such a language has been found. It is Pirahã, a Brazilian Amazon language. It is claimed that cultural factors have made linguistic recursion in that language unnecessary (Everett 2005). Therefore, you should not try to prove that your research (alternative) hypothesis is right. Instead, you should try to reject its opposite, that is, the null hypothesis.

1.2.2 Those mysterious *p*-values…

You know now that one should try to reject the null hypothesis, but how can one do it in practice? In statistics, rejection or non-rejection of the null hypothesis depends on the

value of the corresponding test statistic, such as t, χ^2, W, F., etc. The test statistic is computed on the basis of a sample. If the test statistic is beyond some critical value, one can suspect that this may not be due to mere chance. Therefore, the null hypothesis should be wrong.

How can we decide whether the test statistic is extreme enough to suggest that something is going on? This is possible because statisticians know how test statistics are distributed. A **distribution** is a collection of scores, or values, on a variable. It is common to represent distributions as scores arranged from the smallest to the largest ones with varying likelihood of occurrence. Let us imagine that we have collected the heights of all adult hobbits[1] in a hobbit village. The heights can be represented then as points on the horizontal axis ranged from the smallest to the largest ones, as shown in Figure 1.1. The vertical axis represents the so-called probability density. To put it simply, the plot shows that heights around 110 cm are much more likely to occur than heights around 60 or 160 cm.

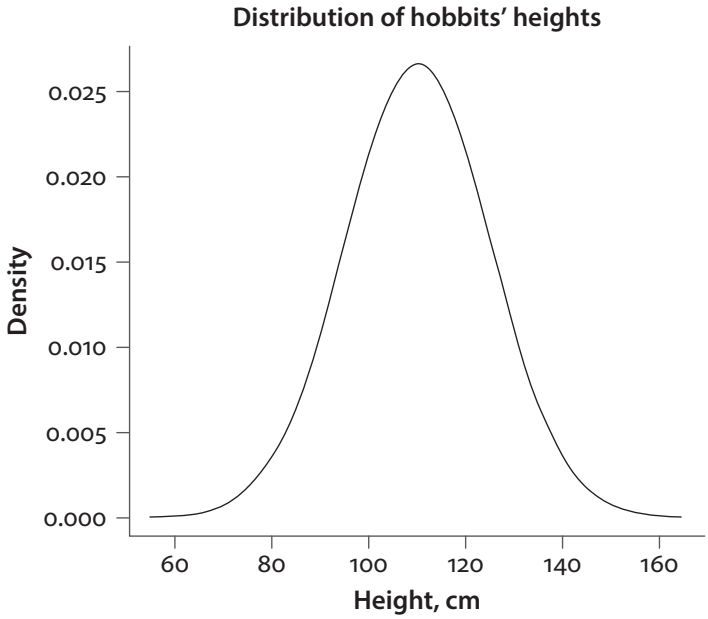

Figure 1.1. Imaginary probability distribution of hobbits' heights

It is worth mentioning that this kind of nicely symmetric bell-shaped distribution is called normal, or Gaussian (after the name of Carl Friedrich Gauss, a great German mathematician). It is a very important concept in statistics because many tests, which are called parametric, assume that the data are distributed normally (see Section 1.4). Of course,

1. Hobbits are small human-like creatures with hairy feet from Tolkien's novels.

there exist a variety of other important distributions with a different shape, such as the F-distribution or χ^2-distribution.

A very useful thing about knowing the shape of a distribution is that one can compute the exact probabilities for a range of x. The entire area under the curve corresponds to the probability of 1 (or 100%). That is, if you measure the height of any random hobbit, there is 100% probability that his or her height will be somewhere under the curve. One can also say that the probability that a random hobbit's height is under 110 cm (the left shaded area on the plot in Figure 1.2) is 0.5, or 50%, because this is a half of the entire area under the curve. Of course, this will hold only if the distribution is symmetric. It is also possible to estimate the probability of meeting a hobbit who is 150 cm tall or taller (the small shaded area on the right in Figure 1.2), around 0.038, or 3.8%.

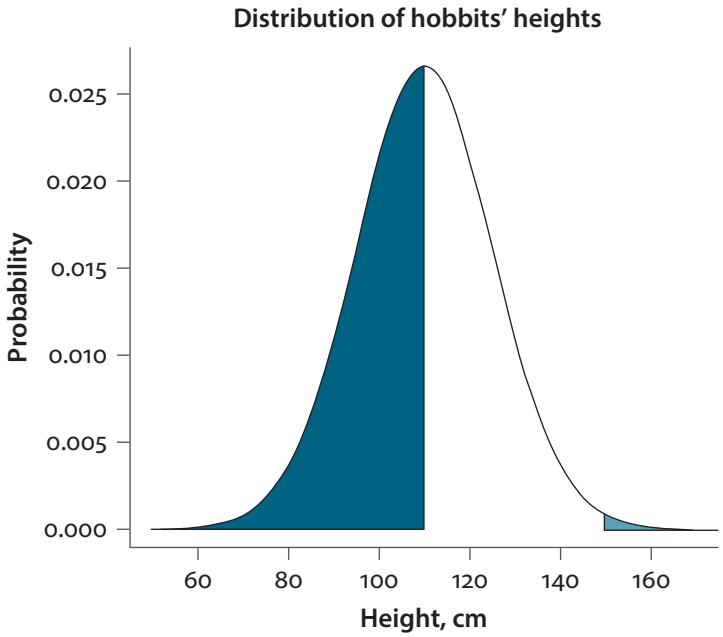

Figure 1.2. Probability distribution of hobbits' heights and different probabilities of observing particular values. Large shaded area under the curve on the left: probability of finding a hobbit shorter than 110 cm (50%); small shaded area under the curve on the right: probability of finding a hobbit taller than 150 cm (3.8%)

Therefore, if one knows what distribution the data come from, it is possible to obtain the probability of observing the actual statistic (and more extreme values) by chance. This is crucial for hypothesis testing. You can compute a test statistic and see the probability of obtaining it and more extreme results by chance alone, that is, under the assumption that the null hypothesis is true. This is what the p-value means.

The p-value shows the probability of obtaining a given test statistic value or more extreme values if the null hypothesis is true.

If a *p*-value is smaller than some conventional level (usually 0.05 or 0.01), then the null hypothesis is rejected, and one has grounds to believe that the result is not due to chance. Therefore, one can conclude that there is a true difference between the groups, association between the variables, etc., depending on the research hypothesis and statistical test. If the *p*-value is larger than this conventional value, then the null hypothesis cannot be rejected, and you can conclude that there is no sufficient evidence that the groups are different (or the variables are correlated, associated, etc). The values 0.05 or 0.01 are called the **significance level**. This is the degree of risk you are willing to take that you will reject a null hypothesis that is actually true. It is crucial that the significance level is decided on *before* the statistical analysis, not after it.

Degrees of freedom

In addition to the test statistic value, one also has to know the number of degrees of freedom (often designated as *df*) in order to compute the *p*-value. In a nutshell, this is the number of values that are free to vary. For many basic statistical tests, it is the sample size minus one. For example, you have three observations with the scores of 3, 4 and 8, and you are interested in their average score, which is 5. You can freely change only two scores without changing the average. The third value is not free. In this example, if you replace the first score with 1, and the second score with 10, your third score can only be 4. Any other third score will change the average. For five scores, you have four degrees of freedom; for ten scores, you will have nine, and so on. For a 2-by-2 table with fixed marginal totals (i.e. the totals of each row and each column), the number of degrees of freedom is $(2 - 1) \times (2 - 1) = 1$. This means that only one cell is free. You can calculate all other cells from the marginal totals.

Degrees of freedom are crucial because they tell us what the distribution of a test statistic ought to look like. For different numbers of degrees of freedom, the **critical values** of the test statistic (that is, the values that correspond to a given significance level, e.g. 0.05 or 0.01) will be different. The number of degrees of freedom is usually reported along with the test statistic and the *p*-value, e.g. χ^2 (1) = 7.47, p = 0.006. The number in brackets displays the number of degrees of freedom (1).

1.2.3 Type I and Type II errors

The significance level says how much risk you agree to take that you discard a null hypothesis that is in fact true. This is called the Type I error, 'false alarm' or 'false positive'. An example is when a doctor diagnoses a disease, but the patient is healthy. Or you think you have found difference between two experimental conditions, but the result is due to sampling error and there is no real difference. If the level of significance is 0.05, it means that there is a 5% chance of rejecting the null hypothesis when it is in fact true.

A Type II error, also called 'false negative', is committed when the researcher accepts a null hypothesis which is in fact false and there is true difference between groups (or association between variables, etc.). An example is when the patient is sick, but the doctor fails to identify the disease. A Type II error is related to the notion of statistical **power.** The more powerful (in the statistical sense) the test, the higher the likelihood that it will reject the null hypothesis when it is false. If you commit a Type II error, this means that your statistical analysis lacked power, which may have to do, for example, with an insufficient sample size.

Note that decreasing the significance level (e.g. from 0.05 to 0.01) will decrease the chances of a Type I error and increase the chances of a Type II error, other things being constant. In contrast, if you raise the significance level (e.g. from 0.05 to 0.1), you will increase the chances of a Type I error and decrease the chances of a Type II error. In general, most linguists and other researchers use the 0.05 level as a trade-off, and one should have very good reasons for changing it.

1.2.4 One-tailed and two-tailed statistical tests

It has been mentioned above that an alternative hypothesis can be directional (e.g. 'X is greater than Y') or non-directional (e.g. 'X is different from Y'). This distinction is very important when one chooses an appropriate statistical test. Most tests (not all!)[2] come in two flavours: one-tailed and two-tailed. If the alternative hypothesis is directional, it is correct to use a one-tailed test. If it is non-directional, one should normally use a two-tailed test. Why is that important? As discussed above, hypothesis testing involves computing a test statistic (e.g. t, T, W or F) and deciding whether it is extreme enough to be expected by pure chance. The reason for distinguishing between one-tailed and two-tailed tests is that you will need different minimum or maximum test statistics in order to obtain a significant result. Recall that the probability of observing a test statistic value under the null hypothesis should be 0.05 or less if the result is significant. In case of a directional one-tailed test, we should look only at one tail of the test statistic

2. For example, the F-test in ANOVA is always one-tailed.

distribution, as shown in the left panel of Figure 1.3. If your alternative hypothesis is 'X is greater than Y', the test statistic should be somewhere in the shaded area (the region of rejection). If your hypothesis is 'X is smaller than Y', the test statistics should be located on the left, in a region of the same size.

In contrast, if your hypothesis is non-directional, that is, 'X is different from Y', you can observe an extreme result either in the left or right tail. This is why the 0.05 value will be split into 0.025 (for the left tail) and 0.025 (for the right tail), as shown in the right panel of Figure 1.3. The shaded areas correspond to the critical regions of the values of the test statistic.

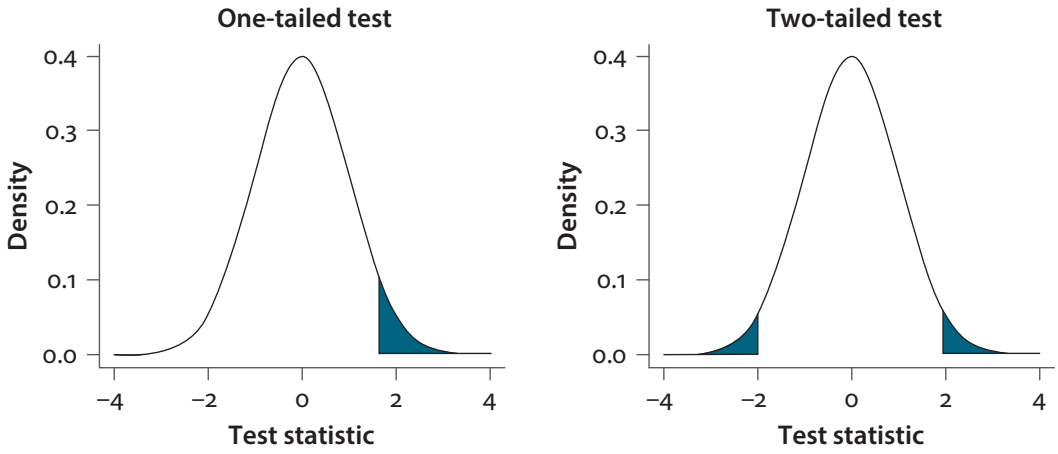

Figure 1.3. One-tailed and two-tailed tests. Left: one-tailed; right: two-tailed. Shaded: the critical region(s) of the test statistic values, which lead(s) us to reject the null hypothesis

Decide on your tails first!

The difference in the size of critical regions for one-tailed and two-tailed tests has important consequences. If you have a directional alternative hypothesis, you need a less extreme value of your test statistic to discard the null hypothesis than if you have a non-directional one. One may be tempted to change the original non-directional alternative hypothesis to a directional one when the one-tailed test yields a significant result, and the two-tailed test does not. Doing such things is scientifically unethical! It is crucial that you formulate your alternative hypothesis and make your choice between the one- and two-tailed tests **before** you compute any test statistic.

1.3 What statistics cannot do for you

Although statistics can help you test your hypotheses, describe your data and do many other things, you cannot let your statistical software do all research for you. Let us imagine that you test a hypothesis, and it is supported by the data. Well done! However, that is not always the case. If we find, as in our example, no significant difference between the speech rates in two national varieties of Dutch, should we give up, or change anything about our research design, data, methods, and so on? Such a decision is not always an easy one. A famous case is Michelson-Morley's series of experiments in the end of the 19th century, which were meant to prove the existence of aether, or the so-called fifth element, which was believed to fill the regions of the universe and conduct light (see an interesting discussion in Geeraerts 1999). The first experiment in 1881 did not bring the expected results. Michelson and Morley decided that the lack of results was due to flaws in the experimental settings and repeated the experiment some years later with a very sophisticated design, all in vain. The second opportunity, however, would be to reject the initial hypothesis, as was done later by Einstein. As this example shows, the ultimate decision is always the researcher's responsibility.

Another thing that one cannot expect from statistics is that it will answer the question 'why'. Causality is always imposed by the researcher on the basis of her theoretical considerations, empirical data and common sense. There is a well-known saying, 'correlation does not imply causation'. Sometimes even a strong and statistically significant correlation may be entirely spurious. For example, according to the website *Spurious Correlations*,[3] the number of people who drowned by falling into a swimming pool in the USA from 1999 to 2009 correlates with the number of films Nicholas Cage appeared in during the same period. Another example is a correlation between the number of world-wide non-commercial space launches and the number of sociology doctorates awarded in the USA from 1997 to 2009.

In some cases there may be a third factor that influences both phenomena in question. For example, if one finds a correlation between the number of infants and storks in a neighbourhood, this does not mean that one causes the other. Both variables can be related to a third factor, which can be described as the degree of urbanization. Birth rates and the populations of storks are usually higher in the country than in urban areas. Consider another example. Ember and Ember (2007) found a positive correlation between sonority of sounds in a language and the frequency of extramarital sex in the corresponding linguistic community. Does it mean that louder speech sounds encourage intimate contacts? Or the other way round, can sexual freedom influence the free flow of voice during

3. See http://www.tylervigen.com/ (last access 11.06.2015).

articulation? Maybe these variables are manifestations of some other underlying cultural and ecological factors? These are questions that statistics alone cannot answer.

1.4 Types of variables

To use statistical methods correctly, one has to know which type of data they are dealing with. At this point it is necessary to introduce the notion of **variable**, i.e. some property of the objects that can vary and that can be measured or described. For instance, if you have a set of words, you can compare them according to the following characteristics: length, grammatical word class, semantic class, frequency of use, register, origin, morphological complexity, and so on. All those characteristics can become your variables. The simplest way is to think of your dataset as a spreadsheet like in Microsoft Excel or OpenOffice Calc with individual subjects or observations as rows and variables as columns. Depending on the number of variables per each item, your data can be univariate (one variable), bivariate (two variables) and multivariate (three and more variables).

Variables in a study may have different status, which depends on how you model the relationships between them. The outcome variable, or the one which changes as a function of some other parameters of interest, is called the **response**, or **dependent variable**. The variables that influence the outcome are called **explanatory**, or **independent variables**.[4] If we carry out an experiment to find out whether the frequency of a word influences how quickly it is recognized in a lexical decision task, the explanatory variable will be the frequency, and the response variable will be the reaction time. Remember that the relationship of causality is something that is imposed by the researcher based on theoretical considerations. Such a relationship is possible, but not necessary. For example, in multivariate exploratory methods (see Chapters 15–19) the notions of dependent or independent variables are irrelevant.

Moreover, all variables vary along different scales of measurement: nominal, ordinal, interval and ratio. This classification is crucial for a correct choice of statistical tests.

– **Nominal** variables are two or more categories that are mutually exclusive. If the number of categories is only two, one speaks of a **binary** variable. For example, a speaker of a language can be male or female, native or non-native; possession may be expressed by the Saxon genitive (*the people's voice*) or the Norman genitive (*the voice of people*). Examples of a greater number of categories are numerous, as well. For instance, a simple clause in many languages may be intransitive, transitive or ditransitive;

4. Some statisticians avoid using the term 'independent variable' because the variables that we consider independent in an experiment or corpus analysis may in fact depend on other variables. Moreover, most corpus-based analyses are correlational and do not assume any (direct) causal relationship between variables.

a nominal phrase in German may be in the nominative, genitive, dative or accusative case; languages can be accusative, ergative, neutral or mixed, displaying split ergativity. Nominal variables represent the least precise and informative level of measurement, in comparison with the ones that follow.

- The categories may be ordered. In that case, we deal with **ordinal** variables. An example is answers in a questionnaire on a five-point Likert scale, e.g. 'strongly disagree' – 'disagree' – 'neither agree nor disagree' – 'agree' – 'strongly agree'. The categories thus differ in order, but we do not know yet by how much. We cannot say, for example, that the difference between 'disagree' and 'neither agree nor disagree' is the same as the difference between 'strongly disagree' and 'disagree'. Different subjects can be using different internal scales. Although sometimes you can see numbers that represent the responses (from 1 to 5), equal intervals on the scale do not represent equal differences between the responses.
- If equal intervals on the scale represent equal differences between the points on the scale, we deal with an **interval** variable. A common example is temperature on the Celsius or Fahrenheit scale. The difference between 20 and 25 degrees is the same as the difference between 25 and 30 degrees. However, it is important that interval variables do not include a zero point, or, if they do, it is arbitrary.[5] That is, the temperature of 0 degrees Celsius does not mean that there is no temperature. This is why it does not make sense to say that twenty degrees Celsius is twice as warm as ten degrees.
- **Ratio** variables are very similar to interval ones, but they include zero on the scale, and the zero point is meaningful, not arbitrary. For example, consider the frequency of the word *aardvark* in a text. One can also very well imagine that a text contains no occurrences of *aardvark*. In that case, zero is perfectly meaningful. As a result, ratio variables allow for multiplication. For example, if the word *aardvark* occurs a hundred times in text A and only ten times in text B, it means that the word is ten times more frequent in text A than in text B.

These scales of measurements are displayed in Figure 1.4 as a set of steps, with the most informative ratio-scaled variables at the top, and the least informative nominal variables at the bottom. It is always possible to go down the ladder from a higher to a lower level of measurement. For example, word frequencies in a corpus can be subdivided into high, medium and low. In that case, we would go down from the ratio scale represented by the frequencies to the nominal level. However, one should go down the ladder only when this is absolutely necessary because every step down means a loss of information. In addition, ratio- and interval-scaled variables can be subdivided into **continuous**,

5. Celsius chose the point at which water freezes as the zero, whereas Fahrenheit used the temperature of a mixture of ice, salt and water, or, according to another version, the coldest temperature he could observe in his home town Gdańsk (Danzig).

when the scale of measurement is meaningful at all points between the numbers given (e.g. height in centimeters/feet or reaction times in milliseconds) and **discrete,** when the units of measurement cannot be split up (e.g. corpus frequencies or the number of phonemes in a language). Ratio- and interval-scaled variables are also often called **quantitative**, or **numerical**, whereas nominal and ordinal variables are called **qualitative**, or **categorical**.

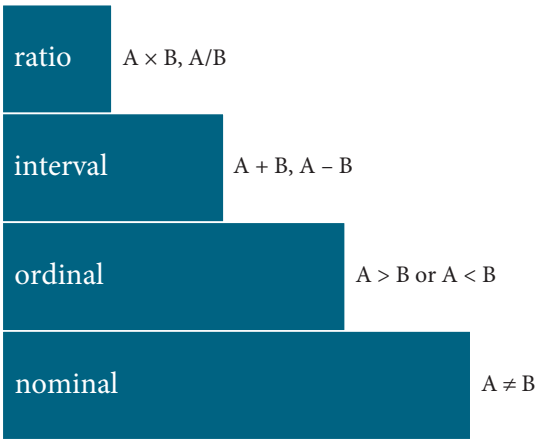

Figure 1.4. Scales of measurement. Every level above adds new information about values A and B, and includes all information at the levels below

If you have ratio or interval data, and they meet some assumptions, such as normality, homogeneity of variance, etc., you will normally be able to use a **parametric test**. Parametric tests are based on some assumptions about the underlying distribution. However, if these assumptions are violated, the resulting test statistic may be meaningless. In such cases, a **non-parametric** test is a more appropriate option. Non-parametric tests do not assume any specific distribution of the data. Since linguistic data are frequently non-normal and small-scale, this book pays special attention to non-parametric tests and situations when one should use them. In this book you will also learn some methods of statistical inference based on resampling (e.g. bootstrap and permutation), which allow one to validate their conclusions without collecting new data.

1.5 Summary

This chapter discussed the most important statistical notions and terms. It began from the distinction between a sample and a population and continued to show how data from a sample can be used to make inferences about the properties of the entire population. The distinction between descriptive and inferential statistics was introduced. Next, you learnt about the logic of hypothesis testing, and got acquainted with the notions of distribution,

p-value, significance level and Type I and Type II errors. Finally, different types of variables and measurement scales were discussed, and the distinction between parametric and non-parametric statistics was mentioned. These notions will be crucial in the next chapters. There are also two important lessons to learn: first, do not let statistics do the thinking for you; second, formulate your hypothesis and decide on the type of test (one- or two-tailed) before you begin your statistical analysis.

More on basic statistical notions

You can find more information about the notions introduced in this chapter in a variety of introductory statistical textbooks, for example, N. J. Salkind's *Statistics for People Who* (*Think They*) *Hate Statistics* (2011) or T.C. Urdan's *Statistics in Plain English* (2010). See also Chapter 1 in Gries' *Statistics for Linguistics with R* (2013).