Unit 6 Making statistical claims

6.1 Introduction

One of the most obvious advantages of using a corpus, as compared with intuition, is that a corpus can provide reliable quantitative data (cf. unit 1.5). In this unit we will consider what to do with the quantitative data that corpora provide. Our approach to this topic will be realistic – by and large most users of corpus data will not be capable of generating sophisticated statistical claims nor would they wish to do so. However, we want readers to be aware of what they should not claim on the basis of the simple descriptive statistics that they may use. Also, in this unit, we want to give guidance on how to interpret inferential statistics generated by the concordance programs used in Section C.

While statistics can be intimidating for many readers, a basic awareness of statistics is essential when adopting a corpus-based approach as 'the use of quantification in corpus linguistics typically goes well beyond simple counting' (McEnery and Wilson 2001: 81). This unit will cover the basic statistical concepts and techniques required to understand some excerpts in Section B and the case studies presented Section C of this book while trying to avoid a complex, technical treatment of statistics. Some of these concepts and techniques, though, will be further discussed in the case studies presented in Section C of this book.

6.2 Raw frequency and normalized frequency

In corpus linguistics *frequency* refers to the arithmetic count of the number of linguistic elements (i.e. tokens) within a corpus that belong to each classification (i.e. type) within a particular classification scheme (e.g. the CLAWS tagset, see unit 4.4.1). It is the most direct quantitative data a corpus can provide. Typically, frequency itself does not tell you much in terms of the validity of a hypothesis. Yet data of this type can be used in both descriptive statistics and inferential statistics (see unit 6.3).

Note, however, that frequency data must be interpreted with caution. It is possible to use *raw frequency* (i.e. the actual count) where no comparison between corpora is necessary. However, when comparing corpora (or segments in the same corpus) of markedly different sizes, raw frequencies extracted from those corpora often need to be *normalized* to a common base (see case studies 2, 4, 5 and 6). For example, the swear word *fucker(s)* occurs 25 times in the spoken section and 50 times in the written section of the BNC corpus (see case study 4). Can we say that the swear word is twice as frequent in writing as in speech? This is clearly not true, as writing accounts for around 90% of the BNC corpus whereas transcribed speech only takes up 10% of the corpus (see unit 7.2): there is nine times as much written data than spoken data. If we compare these frequencies on a common base, e.g. per million words, then we find the normalized frequency of *fucker(s)* in speech is 2.41 per million words whereas it is 0.56 in writing. Clearly this swear word occurs much more frequently (over four times as often) in speech than in writing. Is this difference statistically significant? We will leave this question unanswered till we introduce tests for statistical significance (see unit 6.4).

As the size of a sample may affect the level of statistical significance, the common base for normalization must be comparable to the sizes of the corpora (or corpus segments) under consideration (see case study 2). When we compare the

spoken section (10 million words) and the written section (90 million words) of the BNC corpus, for example, it would be inappropriate to normalize frequencies to a common base of 1,000 words, as the results obtained on an irrationally enlarged or reduced common base are distorted.

6.3 Descriptive and inferential statistics

Given that we have mentioned descriptive and inferential statistics above, what is the difference between the two types of statistics? Basically, descriptive statistics are used to describe a dataset, as the term suggests. Suppose a group of ten students took a test and their scores are as follows: 4, 5, 6, 6, 7, 7, 7, 9, 9 and 10. We might need to report the measure of *central tendency* of this group of test results using a single score.

There are different ways to do this. We can use the *mean*, the *mode* and the *median*. The mean is the arithmetic average, which can be calculated by adding all of the scores together and then dividing the sum by the number of scores. In our example, the mean is 7 (i.e. 70/10). The mean is the most common measure of central tendency. While the mean is a useful measure, unless one also knows how dispersed (i.e. spread out) the scores in a dataset are, the mean can be an uncertain guide. Under such circumstances other scores may help. For example, the mode is the most common score in a set of scores. In this example the mode is 7, because this score occurs more frequently than any other score. Another score one might use is the median. The median is the middle score of a set of scores ordered from lowest to the highest. For an odd number of scores the median is the central score while for an even number of scores, the median is the average of the two central scores. In the above example the median is 7 (i.e. (7+7)/2).

There are three important ways to measure the dispersion of a dataset: the *range*, the *variance* and the *standard deviation*. The range (i.e. the difference between the highest and lowest frequencies) is a simple way to measure the dispersion of a set of data. In the above example the range is 6 (i.e. 10 – 4). However, the range is only a poor measure of dispersion because an unusually high or low score in a dataset may make the range unreasonably large, thus giving a distorted picture of the dataset. The variance measures the distance of each score in the dataset from the mean. For example, in the test results above, the variance of the score 4 is 3 (i.e. 7–4) while the variance of the score 9 is 2. For the whole dataset, however, the sum of these differences is always zero as some scores will be above the mean while some will be below the mean. Hence, it is meaningless to use variance to measure the dispersion of a whole dataset. Standard deviation is a useful measure in such circumstances. Standard deviation is equal to the square root of the quantity of the sum of the deviation scores squared divided by the number of scores in a dataset. It can be expressed as:

$$\sigma = \sqrt{\frac{\sum (F - \mu)^2}{N}}.$$

In the formula F is a score in a dataset (i.e. any of the ten scores in the above example, μ is the mean score (i.e. 7) while N is the number of scores under consideration (i.e. 10). The standard deviation in our example of test results is 1.89. When one uses standard deviation to measure the dispersion of a normally distributed dataset, i.e. where most of the items are clustered towards the centre rather than the lower or

higher end of the scale, 68% of the scores lie within one standard deviation of the mean, 95% lie within two standard deviations of the mean, and 99.7% lie within three standard deviations of the mean. In this sense, the standard deviation is a more reasonable measure of the dispersion of a dataset.

While it may be time consuming to calculate these statistics manually, readers do not need to panic as they can be computed automatically using statistics packages such as SPSS, as shown in case study 5 of Section C. Note, however, that while descriptive statistics are useful in summarizing a dataset, it is inferential statistics that are typically used to formulate or test a hypothesis. Testing hypotheses in this way generally involves various statistical tests. These tests are typically used to test whether or not any differences observed are statistically significant. The sections that follow will briefly introduce the inferential statistical tests used in this book including the chi-square test, the log-likelihood (LL) test, Fisher's exact test, the MI (mutual information) test, the *t* test and the *z* test. The procedures for conducting each test are presented in the relevant case studies in Section C.

6.4 Tests of statistical significance

In testing a linguistic hypothesis, it would be nice to be 100% sure that the hypothesis can be accepted. Sadly, one can never be 100% sure. There is always the possibility that, for example, the differences observed between two corpora have arisen by chance due to inherent variability in the data (cf. Oakes 1998: 1). Hence, one must state the 'level of statistical significance' at which one will accept a given hypothesis. In short, how likely is it that what you are seeing is statistically significant and what tolerance do you have for uncertainty? While we cannot be 100% sure, the closer the likelihood is to 100%, the more confident we can be. By convention, the general practice is that a hypothesis can be accepted only when the level of significance is less than 0.05 (i.e. $p<0.05$). In other words, one must be more than 95% confident that the observed differences have not arisen by chance.

There are a number of techniques for testing statistical significance. The most commonly used statistical test in corpus linguistics is probably the chi-square test (also called the Pearson chi-square test). The chi-square test compares the difference between the observed values (e.g. the actual frequencies extracted from corpora) and the expected values (e.g. the frequencies that one would expect if no factor other than chance was affecting the frequencies - see case study 1 for further discussion). The greater the difference (absolute value) between the observed values and the expected values, the less likely it is that the difference is due to chance. Conversely, the closer the observed values are to the expected values, the more likely it is that the difference has arisen by chance.

Another commonly used statistical test is the log-likelihood test (also called the log-likelihood chi-square or *G*-square test). The log-likelihood (LL) test is preferred in this book, as it does not assume that data is normally distributed (cf. Dunning 1993; Oakes 1998). The log-likelihood statistic has a distribution similar to that of the chi-square, so the LL probability value (i.e. the *p* value) can be found in a statistical table for the distribution of the chi-square. To look up the *p* value, a further value is required, namely, the *degree of freedom* (or d.f.), which is computed by multiplying the number of rows less 1 with the number of columns less 1 in a frequency table (or contingency table) (see case study 1). For example, a contingency table with two rows and two columns has 1 degree of freedom. In both the chi-square and log-likelihood tests, the critical values with 1 d.f. are 3.83, 6.64 and 10.83 for the significance levels

of 0.05, 0.01 and 0.001 respectively. A probability value *p* close to 0 indicates that a difference is highly significant statistically, whereas a value close to 1 indicates that a difference is almost certainly due to chance. In the BNC example in unit 6.2, the calculated chi-square score is 42.664, and the log-likelihood score is 28.841 (1 d.f.), much greater than the critical value 10.83 for the significance level 0.001. Hence, we are more than 99.9% confident that the difference in the frequencies of *fucker(s)* observed in the spoken and written sections of the BNC corpus is statistically significant.

There are many web-based chi-square or log-likelihood calculators. Readers who use a standard statistics package like SPSS can even avoid the trouble of consulting a statistical table of distribution, as the program automatically gives (Pearson) chi-square and log-likelihood scores in addition to indicating the degree of freedom and statistical significance level. It should be noted, however, that proportional data (e.g. normalized scores) cannot be used in the chi-square or log-likelihood tests. The discrepancies in corpus sizes are unimportant here, as these tests automatically compare frequencies proportionally. Note also that the chi-square or log-likelihood test may not be reliable with very low frequencies. When the expected value in a cell of a contingency table is less than 5, Fisher's exact test is more reliable. SPSS computes Fisher's exact significance level automatically if at least one of the cells of the contingency table has an expected value less than 5 when the chi-square test is selected.

6.5 Tests for significant collocations

The term *collocation* refers to the characteristic co-occurrence patterns of words, i.e., which words typically co-occur in corpus data (see units 10.2 and 17). Collocates can be lexical words or grammatical words. Collocations are identified using a statistical approach. Three statistical formulae are most commonly used in corpus linguistics to identify significant collocations: the MI (mutual information), *t* and z scores. In this section, we will briefly introduce these tests. Other statistical measures for collocation will be introduced in case study 1 of Section C.

MI is a statistical formula borrowed from information theory. The MI score is computed by dividing the observed frequency of the co-occurring word in the defined span for the search string (so-called *node word*), e.g. a 4:4 window, namely four words to the left and four words to the right of the node word, by the expected frequency of the co-occurring word in that span and then taking the logarithm to the base 2 of the result. The MI score is a measure of collocational strength. The higher the MI score, the stronger the link between two items. The closer to 0 the MI score gets, the more likely it is that the two items co-occur by chance. The MI score can also be negative if two items tend to shun each other. Hunston (2002: 71) proposes an MI score of 3 or higher to be taken as evidence that two items are collocates.

However, as Hunston (2002: 72) suggests, collocational strength is not always reliable in identifying meaningful collocations. We also need to know the amount of evidence available for a collocation. This means that the corpus size is also important in identifying how certain a collocation is. In this regard, the *t* test is useful as it takes corpus size into account. As such, an MI score is not as dependent upon the corpus size as a *t* score is. The *t* score can be computed by subtracting the expected frequency from the observed frequency and then dividing the result by the standard deviation (see unit 6.3 for a discussion of standard deviation). A *t* score of 2 or higher is normally considered to be statistically significant, though the specific probability

level can be looked up in a table of distribution, using the computed *t* score and the number of degrees of freedom.

While the MI test measures the strength of collocations, the *t* test measures the confidence with which we can claim that there is some association (Church and Hanks 1990). Collocations with high MI scores tend to include low-frequency words whereas those with high *t*-scores tend to show high-frequency pairs. As such Church, Hanks and Moon (1994) suggest intersecting the two measures and looking at pairs that have high scores in both measures.

The *z* score is the number of standard deviations from the mean frequency. The *z* test compares the observed frequency with the frequency expected if only chance is affecting the distribution. In terms of the procedures of computation, the *z* score is quite similar to the *t* score whereas in terms of output, the *z* score is more akin to the MI score (see case study 1). A higher *z* score indicates a greater degree of collocability of an item with the node word. The *z* test is used relatively less frequently than the MI test in corpus linguistics, but it is worth mentioning as it is used in widely used corpus tools such as TACT (Text Analytic Computer Tools) and SARA/Xaira.

Readers may wish to avoid computing the MI, *t* or *z* scores manually by taking advantage of publicly available statistics packages or corpus tools. All of the three tests for collocation introduced in this section can be undertaken using computer programs. SPSS can compute *t* and *z* scores. WordSmith calculates the MI score, while SARA and Xaira allow users to choose from the *z* and MI scores as a measure of significant collocations. Case study 1 in Section C of this book will show readers how to compute the *z* and MI scores using BNCWeb.

6.6 Unit summary and looking ahead

In this unit we introduced the basic concepts and techniques needed to make statistical claims on the basis of the quantitative data provided by corpora. We first noted that the frequencies extracted from corpora need to be normalized to a rational common base if corpora (or subcorpora) of different sizes are compared using descriptive statistics. The chi-square and log-likelihood scores to test statistical significance were then introduced. If the expected frequency in a cell of a contingency table has a value less than 5, Fisher's exact test is recommended. Finally, three tests for significant collocations, namely the MI, *t* and *z* scores were introduced. Further tests for collocation will be discussed in case study 1 in Section C.

This unit serves only as a very minimal introduction to quantitative analysis in corpus linguistics. Nevertheless, some of the concepts introduced in this unit are essential in the case studies presented in Section C of this book, where we will also show readers how to carry out statistical tests using statistical package such as SPSS. It is our hope that this unit will raise readers' statistical awareness in taking a corpus-based approach to language studies. Readers who wish to further explore the use of statistics in corpus linguistics can find useful discussions in Barnbrook (1996), Oakes (1998) and McEnery and Wilson (2001).