

Preface

In language study, as in the natural sciences, sociology or psychology, many kinds of work require the collection of quantitative data. The literary stylistician may wish to count the relative numbers of various colour terms, tense forms, alliterative sounds or some other linguistic feature of the texts in which he is interested. The language teacher or course designer may wish to obtain and compare measures of students' performances under two teaching methods. The theoretician may wish to count how many words in a corpus of texts occur once, how many twice and so on, and to compare these observed data with those predicted by a theoretical model of vocabulary distribution in texts. These are just a few examples of the many possible kinds of quantitative investigation into language. In all of them, we need ways of making sense of the data, and this is the purpose of statistical methods.

In many quantitative studies, we cannot investigate every possible example of the phenomenon we are interested in. In some cases exhaustive investigation is *theoretically* impossible; for example, if we were studying the time taken by informants to utter a particular sentence, the number of possible readings is infinite. In other cases, exhaustive examination is theoretically possible but impracticable; for instance, if we were examining some phonological feature of the English spoken in Birmingham, we could in theory obtain data from every Birmingham resident (or, better, from every resident satisfying a set of predetermined criteria for qualifying as a 'Birmingham speaker'); but this would be extremely time-consuming and difficult to organise, so that we should almost certainly be content with a *sample* from the total population we are concerned with. One important part of

statistics is concerned with methods of sampling, and with the relationships between measurements made on samples, and the properties of the populations these samples are intended to represent.

Once we have a set of data, either for every occurrence of our chosen phenomenon or for a sample of it, we usually need to summarise it in such a way that we can discern its general characteristics. The tools available for this task constitute *descriptive statistics*. Presented with a long list of numbers representing our observations, it is often not easy to see, at a glance, any general trends in the data. Such trends become more obvious when we look at the distribution of the data. For instance, in a language proficiency test on 100 learners, we might record marks out of 20. We can determine how many learners score 0, how many score 1, how many 2 and so on, up to 20, thereby drawing up a *frequency distribution* for the data, which may be converted to graphical form, and which gives an indication of the most typical score as well as the spread of marks. More precise measures of these properties can be obtained by performing certain statistical calculations on the data.

Very often, we are concerned not with the characteristics of just one set of data, but with the comparison of two (or more) sets. For example, we might be interested in testing the hypothesis that the performance of two groups of learners, taught by different methods, will differ in a language proficiency test; or we may wish to investigate whether the proportions of two pronunciations differ in the casual and formal speech of informants. In such cases we face the problem of designing our study in such a way that it will isolate just those phenomena we wish to test. The samples we use must be chosen so as to minimise variation arising from unwanted complicating factors, so that we can be reasonably confident that any effects owing to our chosen phenomenon are not swamped by other, 'irrelevant', effects. Experimental design is, or should be, inseparable from statistical work: no amount of sophisticated statistics can compensate for a badly designed investigation.

Where comparisons are involved, we need to know not only the general characteristics of each sample (such as the most typical value and the spread of values) but also whether the characteristics of the two samples are sufficiently different for us to conclude that there is a real effect that is due to the factor we are investigating. We can never be absolutely sure that a difference between

two sets of observations has not arisen 'by chance', owing to inherent variability in our material. We can, however, carry out tests which may allow us to claim 'real' differences with a specifiable margin of error, say 5 per cent, or 1 per cent, or 0.1 per cent. That is, we may, as a result of our calculations, claim to be 95 per cent sure, or 99 per cent sure, or even 99.9 per cent sure, that we have found a 'real' difference. This area, known as *hypothesis testing*, is an important part of *inferential statistics*.

In summary, then, whenever we wish to collect quantitative data on language, we need to pay careful attention to the design of our study, and to the selection of appropriate statistical methods for summarising the data, and for testing hypotheses concerning differences between sets of data. All these aspects of statistics are discussed in this book. However, since the book is introductory in scope, some techniques of interest to linguists, such as multiple correlation and regression, cluster analysis, and analysis of variance with more than one independent variable, are excluded. In order to deal adequately with these more advanced techniques, at least one further volume would be required.

Many courses on applications of statistics concentrate far too heavily on the methods themselves, and do not pay sufficient attention to the reasoning behind the choice of particular methods. I have tried to avoid this pitfall by discussing the 'why' as well as the 'how' of statistics. A difficult problem for the writer of any text on mathematical topics for non-mathematicians is how far to go into the derivation of formulae. While recognising that most linguists (including myself) will have neither an interest in the more theoretical side of the subject nor the mathematical background necessary for a full discussion, I feel that it is highly unsatisfactory for readers or students simply to be presented with a formula, with no explanation whatever of how it is arrived at. Where I thought it appropriate, I have attempted to give an idea of the rationale behind the various methods discussed in the book. Nevertheless, readers should find that their school arithmetic and algebra will see them through quite easily.

I should like to express my thanks to the various groups of students who have worked through the material presented here, and to Tim Gibson, who checked many of the exercises. My thanks go also to the following, for permission to use copyright or unpublished material:

Statistical tables in appendix 1: Dr H. R. Neave and his publishers George Allen & Unwin, for original or adapted versions of tables

2.1(a), 3.1, 3.2, 3.3, 5.1, 5.3, 6.2, 6.4 and 7.1 from *Statistics Tables for Mathematicians, Engineers, Economists and the Behavioural and Management Sciences* (1978); question 7 of chapter 9 exercises: Dr J. Connolly, for data from his article, 'Quantitative analysis of syntactic change', *Nottingham Linguistic Circular* 8/2, 108-18 (1979); question 6 of chapter 9 exercises: Dr J. Coates and Professor G. Leech, for data from their article, 'The meanings of the modals in modern British and American English', *York Papers in Linguistics* 8, 23-34 (1980); question 4 of chapter 11 exercises: Professor G. Wells and the publishers of *Research in Education*, for data from the article, 'Language use and educational success: a response to Joan Tough's *The Development of Meaning* (1977)', *Research in Education* 18, 9-34; questions 2 and 3 of chapter 2 exercises, question 3 of chapter 3 exercises, question 5 of chapter 9 exercises, questions 1 and 3 of chapter 10 exercises: Dr A. S. Crompton, for data from his work on stress and pause in French.

I should also like to thank Professor D. Crystal for advice on the production of the book.