Chapter One

A REVIEW OF STATISTICAL TECHNIQUES IN THE ANALYSIS OF LINGUISTIC DATA*

WHY LINGUISTS NEED TO KNOW ABOUT STATISTICAL TECHNIQUES

A glance at any major linguistics journal will reveal that a considerable amount of research in the area of language studies is quantitative in nature. A phonetician may wish to study voice onset times; a sociolinguist may be interested in the variation in frequency of linguistic features with social class, age, sex, etc.; a stylistician may find it useful to compare sentence lengths or vocabulary profiles for two authors or texts; the language teacher may want to compare the performance of two groups of learners taught by different methods -- the list could be extended indefinitely. As soon as we start to examine any phenomenon quantitatively, we come up against the problem of variability in our data, and this is what often gives rise to the need for statistical treatment.

One type of statistical procedure is concerned with the need to summarize complex data sets, for instance by calculating measures of typicality and of variability. Such methods are part of what is known as descriptive statistics. Often, however, we are concerned not just with a single set of data, but with comparisons of two or more sets. We may wish to ask, for example, in which of two phonological environments a particular vowel is longer, or whether two groups of learners differ significantly in

---

* Christopher Butler, Department of Linguistics, University of Nottingham, UK

their performance on a language test. We may also need to know how far we can extrapolate from the properties of a sample of linguistic data to the whole population from which the samples were drawn. In such cases, inferential statistical techniques are required.

The aim of the present chapter is to give an overview of the techniques available, rather than details of the procedures involved. A more detailled treatment of some of these techniques can be found in Chapter Four of this book, and in the accounts by Butler (1985) and by Woods et al. (1986).

SUMMARISING THE DATA: SOME SIMPLE DESCRIPTIVE STATISTICAL TECHNIQUES

Frequency distributions

In investigations of a quantitative nature, we measure the values of a variable quantity, such as scores on a language test. A useful initial step in examining such a data set is to construct a frequency table showing how many times each particular value (or range of values -- see below) occurs. For instance, if our language test is scored from 0 to 20, we could tabulate the numbers of students who scored marks of 0, 1, 2,3, etc., up to 20. Such a table often gives a good idea of the way in which the data are distributed. An ever clearer view can be obtained by showing the distribution graphically, either as a histogram in which frequencies are represented by the height of a box drawn over the particular value of the variable, or as a frequency polygon, in which dots or crosses are placed at appropriate coordinates on the graph and then joined up. Figure 1 shows a histogram for a hypothetical set of language test scores, for which a frequency table is given in Table 1. Figure 2 shows a frequency polygon for word lengths in a sample of text. It may also be useful, for some purposes (eg. the calculation of a median -- see below), to draw a graph showing the cumulative frequencies of observations, for instance, the number of students who obtained marks of 1 or less, 2 or less, 3 or less, etc. A cumulative distribution graph for the language test scores is given in Figure 3.

Table 1

Hypothetical data on language test scores

Score	Frequency	Score	Frequency
0	0	6	6
1	1	7	3
2	2	8	2
3	5	9	1
4	10	10	0
5	12		
10	10	10	N = 42

If the number of values which a variable can take is large, we may obtain a distibution where the frequency of many values is very low. In such cases, it is often best to combine values into groups. For instance, if we were measuring sentence lengths in a text, we might find a range of lengths from, say, 5 to 40 words. We could then group the data into units of 5 words, so that our frequency table would record how many sentences had between 1 and 5 words, how many had 6-10 words, and so on. The frequency table could then be converted to a histogram or frequency polygon.

If we have measured the values of a variable for two or more experimental conditions (eg. two groups of learners taught in different ways, or several different environments for a particular vowel), we can present the frequencies for each condition side by side in the same table, to facilitate comparison. We may also need, at times, to measure the frequencies of items classified according to more than one set of criteria: for instance, occurrences of verb forms classified (i) according to whether they are present or past in tense, and (ii) according to whether they are simple or progressive in aspect. In such cases, a contingency table such as that in Table 2, may be used.

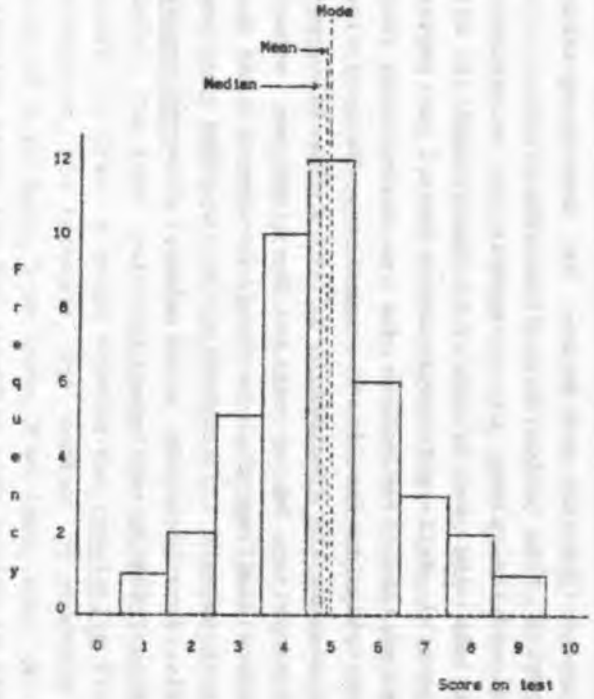

Fig. 1.
Hypothetical data on language test scores.

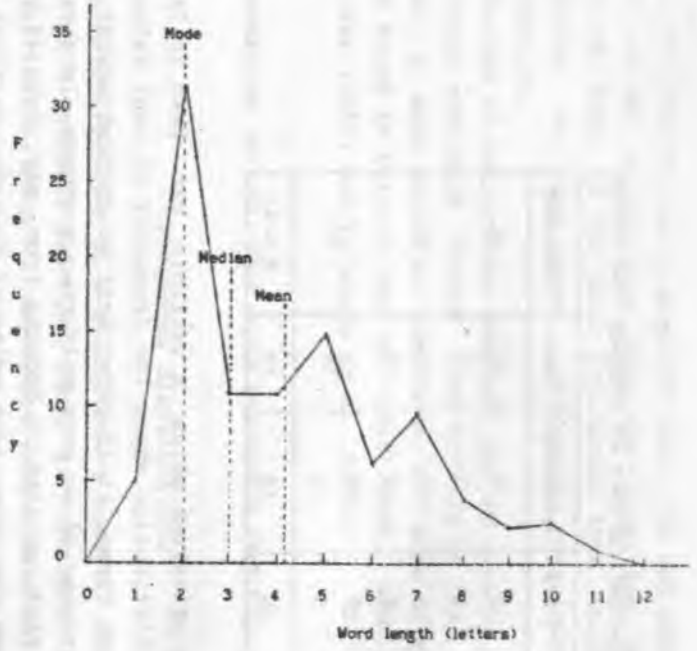

Fig. 2.
Lengths of first 100 words in Proust's *Du Cote de chez Swann*.

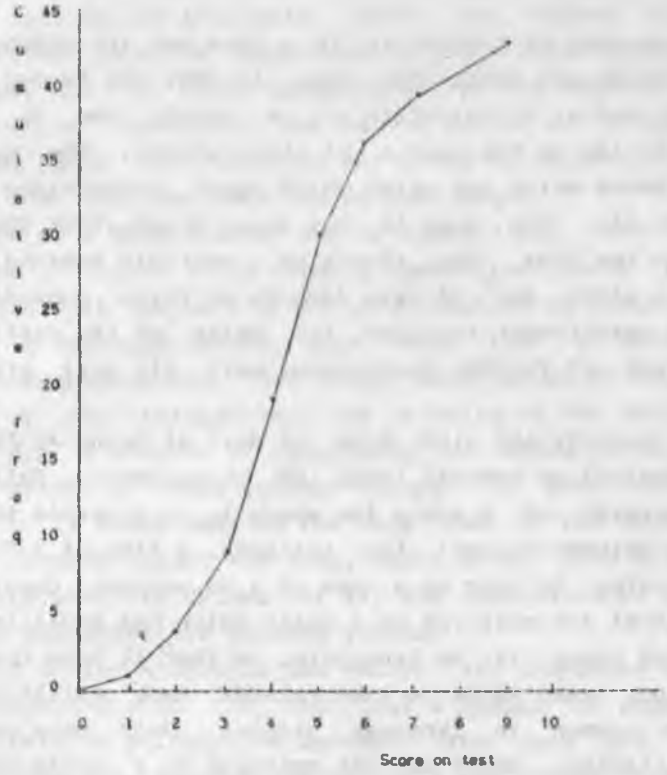

Fig. 3.

Cumulative frequency distribution for language test data in Table 1.

T a b l e 2

Contingency table for hypothetical data on frequencies of tense and aspect forms

Tense	Simple	Progressive	Total
Present	43	87	130
Past	35	26	61
Total	78	113	191

Measures of typicality or 'central tendency'

Three measures of typicality in a data set are commonly used: the mean, median and mode. The mean is what the layman means by an 'average' value: to calculate it, we simply add up all the values and divide by the number of observations. The median is that value above which and below which equal proportions of the observations lie. The mode is that value which shows the highest frequency in the data. The choice of a suitable measure of typicality for a given set of data depends on three considerations: the kind of measurement involved, the shape of the distribution, and what kinds of further statistical work (if any) are to be done.

We may classify any given type of data as being at the ratio, interval, ordinal or nominal level of measurement. Ratio level data are measured on a scale for which it is sensible to calculate ratios between values: for instance, a time of 2.00 seconds is exactly twice as long as a time of 1.00 seconds. Quantities of interval level are measured on a scale which has equal intervals, but for which there is no true zero, so that it does not, strictly speaking, make sense to take ratios. Such variables are, however, not common in language studies. Much more common are ordinal variables, which are not measured on a scale with equal intervals, but which can be ranked in order of magnitude. For example, in assessing the politeness of a particular form of request in English, on a scale from 1 to 7, we would not want to claim that a form with a rating of 3 was exactly 1,5 times as polite as one with a rating of 2, though we could safely say that it was rated as more polite. Finally, nominal variables are of a yes-or-no type: for instance, either a verb in English is simple past in tense or it is not; there is no reasonable sense in which one verb can be 'more simple past tense' than another. Returning now to the measures of central tendency, we see that the calculation of a mean assumes that it makes sense to add up values and divide by the number of observations: in other words, it is most suitable for ratio or interval data such as times, scores on a test, etc. On the other hand, since the median is based on ranking, it is suitable for ordinal data. We can, of course, use

the median also for ratio or interval data, but we thereby lose some of the information available, since the median takes account only of the relative magnitude of the observations, whereas the mean makes use of their absolute sizes. The mode is a rather crude measure of typicality, but may be useful as a rough guide on occasions. Some distributions may prove to be bi- or even polymodal, in that they may show two or more peaks.

The shape of a distribution may be symmetrical (for instance the histogram in Figure 1 is roughly symmetrical about the mode) or skewed (as with the word length distribution in Figure 2). If the pattern is strongly skewed, this means that a small proportion of the observations has either much larger or much smaller values of the variable than the majority of the data. If we are looking for a measure of typicality, we would do well to minimise the effect of these extreme values. In such cases, the median, which is based only on the rank order of the observations, is a better measure than the mean, which as we have seen takes account of the absolute values of all the observations, and so is sensitive to the effect of extreme values.

The third factor affecting the choice of a measure of typicality is concerned with the fact that a number of powerful and well-known tests of differences between data sets are based on the mean. There are, however, as we shall see, other tests which are based on ranking, and so are related to the median.

Measures of variability

The most obvious and easily calculated measure of variability is the range of a set of observations, which is simply the difference between the highest and lowest values in the data set. The range is, however, very susceptible to the effects of extreme values (see discussion of the median and mean above). A better measure for ordinal data, or data with strongly skewed distributions, is the interquartile range, which is the difference between the values below which and above which a quarter of the observations lie (ie. between the first and third 'quartiles', quartiles being the values which divide the number of observations into four just as the median divides them into two). A further

easily interpretable measure is the mean deviation, which is the mean of the differences between each observation and the mean. Unfortunately, the mean deviation does not have the mathematical properties required for further statistical work, and so is little used. The most common measure of variability is the standard deviation, which is the square root of a quantity known as the variance. To obtain the variance of a set of data which is a sample from a population, we calculate the sum of the squares of the deviations of each observation from the mean, and then divide by (N - 1), where N is the number of observations. The standard deviation is unfortunately difficult to interpret in any common-sense way. Furthermore, since it involves treating the values as numbers to be added up, divided, etc., it is really more suitable for ratio or interval data than for ordinal data.

INFERENTIAL STATISTICS: HYPOTHESIS TESTING

G e n e r a l p r i n c i p l e s. As was remarked earlier, quantitative investigations in linguistics are frequently comparative. In other words, we are often interested in studying the values of a particular variable (the so-called dependent variable) in two or more sets of data which reflect differences in one or more other variables (the independent variable(s)). We might, for instance, want to investigate the effect of sex (the independent variable) on the frequency of use of a particular intonation pattern (the dependent variable).

If our aim is to find out whether the data obtained under two or more conditions reflect significantly different properties in the populations of observations from which the samples have been drawn (for discussion of the concept of 'significance', see below), we must attempt to ensure that there is no systematic difference in the conditions under which the data are collected, except for that which is related to the variable we are interested in. This means that we must be very careful about the design of our investigations. It cannot be emphasised too strongly that the results of an investigation can be totally vitiated by lack of attention to the details of design: no amount of sophistica-

ted statistical juggling can remedy the faults of a badly set up project.

We may recognise three basic types of experimental design. Perhaps the most common is the independent groups type, in which the sources of the sets of data are quite independent. For instance, we might take two quite separate groups of learners of English as a foreign language, and test them on particular language skills. We can, if we wish, take some steps to make the groups similar: for instance, we might ensure that they contained equal proportions of male and female students, of students from different types of language background, and so on. Within such constraints, however, we would want to make sure that each available subject (that is, each of the persons being studied) had an equal chance of appearing in each of the two groups (see the discussion of random sampling techniques in Chapter Two. Rather more control over extraneous variables can be exercised in a matched pairs design, where each member of a sample is closely matched with a member of the other sample(s) on criteria which are considered likely to affect the results. For instance, we might try to find pairs of language learners who were matched for sex, age, language background, motivation and aptitude. As with the independent groups design, we would want to ensure that members of a pair were allocated randomly to the two groups under test. Unless large numbers of subjects are available, it is often hard to implement the matched pairs design. In some kinds of study, it is possible to achieve maximal control over unwanted sources of variation by using a repeated measures design, in which the same subjects are involved in all conditions of the test. For instance, we could ask the same set of subjects to pronounce a series of words in which a given vowel was present in different phonological environments. Such a technique is clearly not suitable for certain types of study, such as the language testing situation discussed earlier.

For the sake of simplicity, it will be assumed for the rest of this section that we are dealing with just two sets of observations, and are interested in testing whether the populations from which the samples were drawn can justifiably be claimed to differ significantly in their measures of central tendency; later,

other types of test will be discussed. Corresponding to any such situation, we can set up a null hypothesis: an example would be the hypothesis that populations from which two groups of EFL learners are drawn would not differ significantly in their mean scores on a particular test. The alternative hypothesis in this case could be of either of two kinds. We could predict that there is a significant difference between the means for the two populations, but not which will be higher, in which case we have a nondirectional hypothesis (sometimes said to be two-tailed, because it involves both ends, or 'tails' of the distribution curve for the test statistic concerned). Or we might be able, on the basis of previous evidence, to predict not only that there will be a significant difference, but also which population would have the higher mean score, in which case we have a directional or one-tailed hypothesis.

So far, the term 'significant' has been used without explanation. It is important to realise that in assessing differences between samples, no conclusion can ever be 100% certain. However different the measures on the samples may be, there is always a chance (which may be very small) that the differences have arisen through the variation inherent in sampling procedures. For any given investigation, therefore, the analyst must decide what level of possible error can be tolerated. Often, in linguistic work, a possible error level of 5% is taken as sufficient; that is, a difference is accepted as significant if there is a 5% chance, or less, that the difference could have arisen 'by chance'. We are then operating at the 5% or 0.05 significance level, often shown as $p \leq 0.05$, where p stands for 'probability that the observed difference occurs by chance variation'. In cases where the validity of the results is more orucial (for instance, in an attempt to refute someone else's claims) we might want to operate with a more stringent significance level, say 1% ($p \leq 0.01$) or even 0.1% ($p < 0.001$).

HOW TO CHOOSE A TEST OF THE SIGNIFICANCE OF THE DIFFERENCE IN CENTRAL TENDENCY BETWEEN TWO SETS OF DATA

The choice of an appropriate test depends on several factors. Firstly, we must make a distinction between parametric and non-parametric tests. Parametric tests assume an interval or ratio level of measurement, and also a reasonable degree of conformity to the symmetrical, bell-shaped distribution known as the 'normal' distribution. Some also require that the variances of the populations from which the samples are derived (usually estimated from the variances of the samples themselves) are similar. Further tests are available for checking the validity of these assumptions for particular sets of data, but we shall not go into these here. Non-parametric tests make no assumptions about the shape of the distribution or about the variances, and are available for the analysis of nominal and ordinal, as well as interval and ratio, levels of data. Figure 4 shows in flowchart form the steps involved in deciding which test to use.

We shall consider briefly three examples of the choice of tests. Firstly, let us imagine that we have measured scores on a language test for two independent groups of learners. Here, we have a ratio level of measurement. Provided that rough checks show the distribution of the data to be roughly normal, and the variances of the samples to be similar, we may use the t-test for independent samples. If these assumptions are not met, we may use the Mann-Whitney U-test. Secondly, let us take the measurement of vowel lengths in two different phonological environments, in utterances produced by the same set of subjects. Again, we have a ratio level of measurement, but clearly we need the t-test for related samples, rather than that for independent samples. Thirdly, consider a set of acceptability ratings for two sentences in a particular social context, as given by a single set of informants. In this case the data are not ratio or interval, but ordinal, and we have a repeated measures design, so that the appropriate test is the sign test.

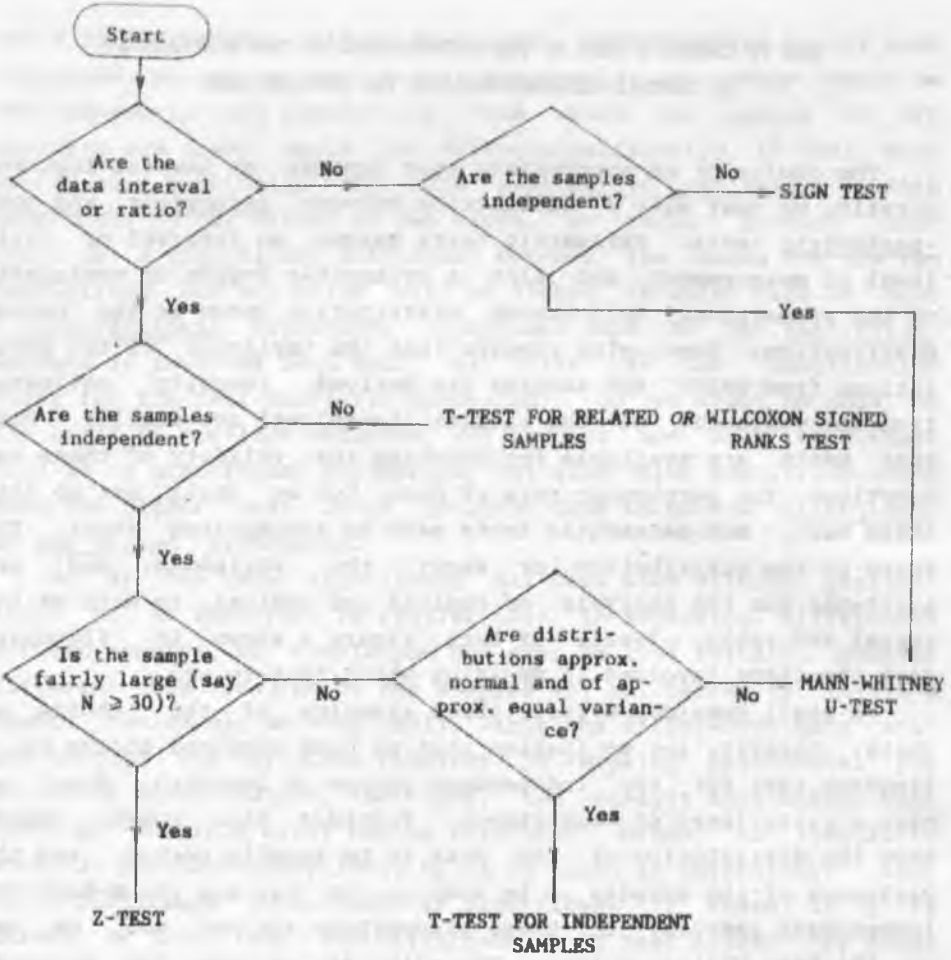

Fig. 4. How to choose a test for the significance of the difference in central tendency between two data sets

Comparing more than two sets of data

It has so far been assumed that we are dealing with just two sets of data. We may, however, wish to compare more than two sets, to see if there are any significant overall differences. For instance, we might want to compare scores on a language test taken by groups of French, German, Spanish, Italian, Japanese and Arab students. Unfortunately, it can be shown that if we perform, say, a number of separate t-tests on all possible pairs of data sets, we may be misled about the significance of the observed differences. There is, however, a technique known as the analysis of variance (ANOVA) which will allow us to examine all the data sets at the same time. ANOVA works by partitioning the variability of the data into two kinds: that which occurs within each data set and that which is found between the data sets. The ratio of these measures of variability is known as the F-ratio, and its value allows us to say whether there are significant differences between the data sets. ANOVA techniques can also be used for situations where the effects of more than one independent variable are being investigated. For example, we may be interested not only in the effect of native language background in the performance of learners on a language test, but also in the effect of two different teaching methods. In such a case, a 2-way ANOVA will allow us to study the effects of each of the independent variables (language background and teaching method), and also any interactions between these variables (ie. the difference in response to the two teaching methods may itself depend on language background).

TESTING FOR INDEPENDENCE OR ASSOCIATION BETWEEN VARIABLES

The chi-square (χ^2) test

This very useful test may be employed where we have measured frequencies of occurrence of entities classified according to one or more nominal (yes/no) variables. Its function is to compare an observed distribution of data with a predicted distribution. The test is used in two main ways in linguistic work. Firstly, it

can be used to test the 'goodness of fit' of a set of observed data to a theoretical model. For instance, we may attempt to construct a mathematical model of the distribution of vocabulary in a text, which will take the total number of running words, the number of different words, and the number of words occurring once only, and use these figures to predict the numbers of words occurring twice, three times, four times, etc., in the text. We may then use the chi-square test to look for significant differences between this theoretical distribution of frequencies and that actually found in the text. The second, and more common, use of the test is to look for independence or association between different variables. For example, imagine that we have taken a random sample of sentences from the early, middle and late novels of a particular author, and have counted the number of sentences in each sample which can be classified as short, medium length or long, according to a specific set of definitions of these terms. We may set these observations out in the form of a 3 x 3 contingency table, then use chi-square to test the null hypothesis that there is no association between the length of sentences and the period in the author's output from which the sentences are taken. If significant differences are found, we may interpret these by looking again at the frequency table.

Correlation and regression techniques

By means of simple correlational techniques, we can answer questions about whether high values of one variable tend to be associated in a linear manner with high, or low, values of a second variable. For instance, we might want to know whether the size of a child's vocabulary rises in a linear fashion with age, or whether in a class of language learners scores on a comprehension test rise linearly as scores on a verbal production test increase. The relationship between the two variables can be expressed graphically in the form of a scattergram in which values of one variable are plotted against values of the other. We may also calculate a correlation coefficient, which takes values from +1 (perfect positive correlation: high values of variable A are associated in an exactly linear manner with high values of

variable B) to -1 (perfect negative correlation: high values of A go with low values of B, and low values of A with high values of B). The most commonly used measures are the Pearson product moment correlation coefficient (r), suitable for ratio and interval variables such as are involved in the examples just given and the Spearman correlation coefficient (rho), which is based on ranking and therefore appropriate for ordinal data (eg. measurements of politeness, grammaticality, acceptability, etc.) We may set up the null hypothesis that there is no correlation between values of the two variables, and then test whether the value of the coefficient is significantly different from zero.

A useful elaboration on correlational analysis is linear regression analysis, which allows us to predict values of one variable if we know it is significantly correlated with a second variable about which more is known. For example, we may be able, on the basis of measurements on a sample of children of varying ages, to construct a simple equation describing the relationship between the age of a child in months and his or her vocabulary size, and then use this to predict the vocabulary size for a child of a particular age who was not investigated as part of the study.

MULTIVARIATE ANALYSIS

So far, we have considered only those cases where measurements are made on a single dependent variable. In some kinds of study, however, we may wish to examine the relationships among a number of variables, so that multivariate analysis is required. As an example, let us take the situation where we have a set of texts, for each of which we have measured a number of properties, such as median sentence length, median word length, type-token ratio (i.e., the ratio of the number of different words to the total number of running words), the proportion of words occurring only one, etc. We may now calculate a measure of similarity or dissimilarity between each pair of texts, a common way of doing this being to calculate a correlation coefficient (r) for each pair across the whole range of measures and then either to use

this as a measure of similarity, or to express the dissimilarity as $(1 - r^2)$. This will allow us to construct either a similarity matrix or a dissimilarity matrix showing the values for each possible pair of texts. This matrix may now be used as input to a number of techniques for discovering groupings of the texts.

Hierarchical cluster analysis will look for the tightest clusterings of texts in terms of similarity over the various measures, and will then attempt to combine these tight clusters into looser configurations. Multidimensional scaling techniques give a pictorial representation of the relationships concerned. Linear discriminant analysis can then be used to test whether the clusters isolated are in some sense 'real', while factor analysis and principal components analysis are techniques for reducing the multiple dimensions of variation to a smaller number of factors. It should be clear even from this very brief summary that multivariate techniques are essentially descriptive, allowing the user to explore the structure of complex relationships between variables. Because they involve lengthy and complex calculations, they are invariably performed by computer programs. For more detail about these methods, see Woods et al. (1986, Chapters 14 and 15), and for an example of the applications of hierarchical cluster analysis and multidimensional scaling see Chapter Four of the present book.

CALCULATORS AND COMPUTERS IN LINGUISTIC STATISTICS

Today, it is unnecessary to carry out manually the often complex calculations involved in statistical work. Scientific calculators normally have facilities for the calculation of means and standard deviations, and often also correlation coefficients and linear regression statistics. Pre-written computer 'packages' are available for a wide range of statistical techniques, two common ones being SPSS (the Statistical Package for the Social Sciences) and Minitab, both of which are now available for the IBM PC and compatible machines. A very powerful package, Genstat, is also about to be released in a version for such machines. Less comprehensive but still very useful packages are available for smaller computers such as the BBC machines.

Useful as such computational aids are, they bring their own dangers, in that they allow the user to produce large quantities of statistical information without requiring a knowledge of the underlying principles. It cannot be too strongly emphasised that in order to use statistical techniques sensibly, linguists should make the effort to understand the reasons for the choice of particular methods, the assumptions being made, and the limitations of the techniques available.