ANALYSIS OF FREQUENCY DISTRIBUTIONS

V. Gardiner & G. Gardiner

ISBN 0 902246 98 4

V. Gardiner and G. Gardner

CATMOG

(Concepts and Techniques in Modern Geography)

CATMOG has been created to fill a teaching need in the field of quantitative methods in undergraduate geography courses. These texts are admirable guides for the teachers, yet cheap enough for student purchase as the basis of class-work. Each book is written by an author currently working with the technique or concept he describes.

1. An introduction to Markov chain analysis - L. Collins
2. Distance decay in spatial interactions - P.J. Taylor
3. Understanding canonical correlation analysis - D. Clark
4. Some theoretical and applied aspects of spatial interaction shopping models - S. Openshaw
5. An introduction to trend surface analysis - D. Unwin
6. Classification in geography - R.J. Johnston
7. An introduction to factor analytical techniques - J.B. Goddard & A. Kirby
8. Principal components analysis - S. Daultrey
9. Causal inferences from dichotomous variables - N. Davidson
10. Introduction to the use of logit models in geography - N. Wrigley
11. Linear programming: elementary geographical applications of the transportation problem - A. Hay
12. An introduction to quadrat analysis - R.W. Thomas
13. An introduction to time-geography - N.J. Thrift
14. An introduction to graph theoretical methods in geography - K.J. Tinkler
15. Linear regression in geography - R. Ferguson
16. Probability surface mapping. An introduction with examples and Fortran programs - N. Wrigley
17. Sampling methods for geographical research - C. Dixon & B. Leach
18. Questionnaires and interviews in geographical research - C. Dixon & B. Leach
19. Analysis of frequency distributions - V. Gardiner & G. Gardiner
20. Analysis of covariance and comparison of regression lines - J. Silk
21. An introduction to the use of simultaneous-equation regression analysis in geography - D. Todd

Other titles in preparation

This series, Concepts and Techniques in Modern Geography is produced by the Study Group in Quantitative Methods, of the Institute of British Geographers.
For details of membership of the Study Group, write to the Institute of British Geographers, I Kensington Gore, London, S.W.7.
The series is published by Geo Abstracts, University of East Anglia, Norwich, NR4 7TJ, to whom all other enquiries should be addressed.

ANALYSIS OF FREQUENCY DISTRIBUTIONS

by

V. Gardiner and G. Gardiner

(Leicester University and Formerly North Staffordshire Polytechnic)

CONTENTS

I INTRODUCTION

(i) Pre-requisites of reader

This monograph is intended to provide a convenient summary of methods which may be used to assess the distributional characteristics of data used in geographical work. We assume of the reader only an ability to perform simple arithmetic operations, although a basic understanding of the use and meaning of simple statistical tests would also be an advantage. If the techniques outlined are to be applied to large amounts of data then the ability to translate simple arithmetic operations into computer code is also necessary. Those fully conversant with the concept of a frequency distribution may wish to omit Sections I-III. A glossary of terms used (italics in text) is included on Page ; this also includes symbols used in formulae.

(ii) The concept of a frequency distribution

Data analysed in geography often consist of a series of figures which measure some attribute of the geographical *individuals* concerned. Geographical data may, for example, be the yield of wheat per unit area for a set of parishes, the average slope of a series of drainage basins, or the annual rainfall for a number of meteorological stations; the individuals in these examples are parishes, drainage basins and meteorological stations respectively. Geographical individuals may be either regions having an areal extent, such as administrative units or drainage basins, or may be regarded as being point locations, such as industrial enterprises or meteorological stations. Each figure in such a set of data is termed a *variate* whilst each measured characteristic is termed a *variable*. For example, the weather at a series of meteorological stations may be described by variables such as temperature, rainfall and humidity, each observation being a variate.

Some impression of the range and average value of a variable may be gained by merely looking through a list of variates but for the sake of simplicity or clarity information in grouped or generalised form is often more useful. Each separate value in a set of data has greater importance when it is considered in relation to the other values as part of a numerical distribution than as a single isolated observation. Presentation of grouped data is usually executed by preparation of a frequency diagram, usually in the form of a *histogram*. To prepare a histogram the variates are grouped into classes showing how many individual variates fall into each class. This is termed the frequency for each class, and a diagram showing these frequencies in ascending magnitude of class values in the form of a bar graph is called a histogram.

The above ideas are illustrated by the following example, using thirty years of rainfall data as the set of variates (Table I, Figure I). These are a subset or selection from all possible yearly rainfall totals. All such possible rainfall totals are the *population* of interest and small subsets selected are *samples*.

Table I. Annual Rainfall (mm), 1931-1960.

Year	Werrington Park	North Tamerton	Year	Werrington Park	North Tamerton
1931	1316	914	1946	1439	1008
1932	1266	889	1947	944	792
1933	869	682	1948	1181	843
1934	1150	849	1949	1037	688
1935	1195	927	1950	1369	994
1936	1278	854	1951	1246	882
1937	1247	1017	1952	1157	795
1938	1280	844	1953	820	637
1939	1344	981	1954	1268	1007
1940	1184	815	1955	987	926
1941	1036	737	1956	983	774
1942	1103	778	1957	1111	895
1943	1208	842	1958	1305	998
1944	1157	721	1959	1376	1034
1945	1159	726	1960	1568	1362

The variable concerned is annual rainfall and the thirty values of total rainfall, one for each year from 1931 to 1960 for the rainfall stations (Werrington Park and North Tamerton), provide thirty variates for each station. The thirty variates are each assigned to a class, the class limits being at intervals of 150mm. Histograms of these data are presented in Figure I.

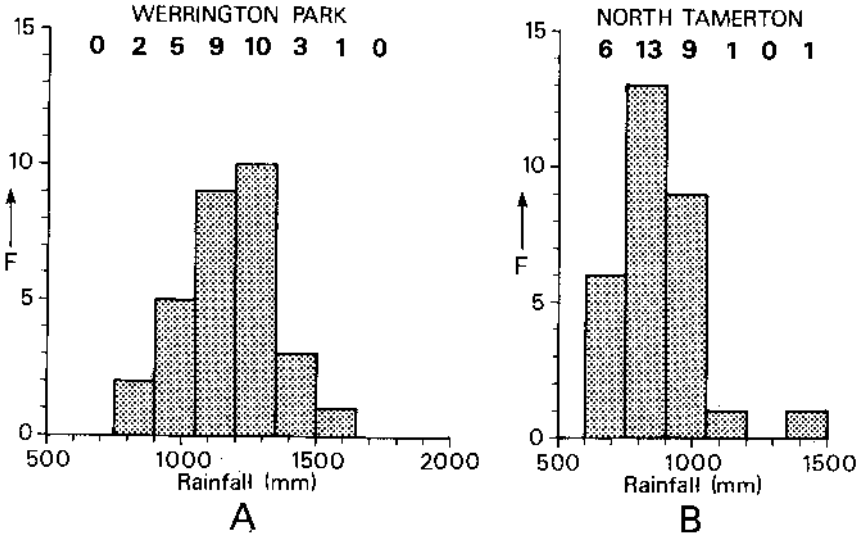

Figure I. Frequency histograms for the rainfall data of Table I. The vertical scale is one of frequency (F); the number of variates in each class is given above each bar, the heights of which are proportional to the frequency associated with each class.

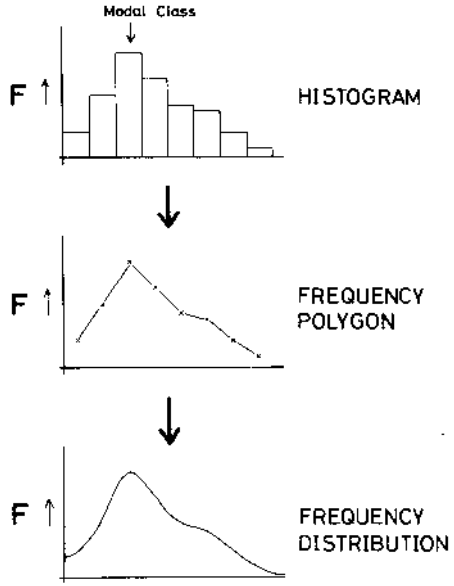

Figure 2. The transition from histogram to frequency distribution. Note however that the transitional step via the frequency polygon is often omitted.

The overall spread of frequencies between the classes is called the *frequency distribution* of the data, and it is with methods by which the nature of such distributions may be analysed that this monograph is concerned. Note that the range and distribution of variates as a concept is independent of the class intervals used in the preparation of the histogram, and although a frequency distribution is conventionally visualised as a histogram it is more correctly thought of as a generalised concept relating to the spread or concentration of values, independent of the classes used.

We have so far visualised frequency distribution in terms of histograms, but if we imagine a very large sample drawn from a very large population, and the variates are assigned to classes with very small class ranges of the variable, then the histogram produced would be in the form of a smooth curve. Histograms are often presented in this form, or as frequency polygons (Figure 2), as these more nearly accord with our concept of what the underlying frequency distribution should look like, uninfluenced by the sample taken or class size used in preparation of the histogram.

(iii) Types of data

It has so far been assumed that the data of concern represent values from a continuous range. However it is possible to recognise both *continuous* and *discrete* data. The former may take on any value within a given range and in general are measured values such as rainfall, area or length, or derived ratios such as population density or percentage of land use. Discrete data consist of counted values, such as numbers of buildings or drumlins, and are generally whole numbers. In practice the distinction may become somewhat blurred since continuous scale measurements have a

finite resolution, and arithmetic rounding off reduces the possible values of a continuous variable, so that continuous data may be considered discrete if rounded to the nearest whole number. Nevertheless the distinction is an important one as it determines the way in which data should be handled. The techniques described in this monograph are intended primarily for analysis of continuous data, although they may also be applied to discrete data in some cases; a full discussion of frequency distributions for use with discrete data is given by Gates and Ethridge (1972).

Further specific types of continuous data which need to be recognised include closed data, which may only take on values between particular limits, as for example 0 and 1 in the case of many shape measures or 0 and 100 as in the case of percentage data, and open ratios, which may range from 0 to infinity. This distinction demands detailed consideration when transformations are being considered (Section V).

II THE NEED FOR EXAMINATION OF FREQUENCY DISTRIBUTIONS IN GEOGRAPHICAL RESEARCH

The character of frequency distributions is important to all geographers who are likely to wish to use numerical methods in any branch of the subject; some of the more important applications are described below. The ways in which frequency distributions may be examined or used are first of all outlined and then some specific applications are described.

(i) Descriptive statistics

A set of variates may be visually summarised by a histogram, but this is not always convenient, in which case descriptive statistics may be used. The most common of these are the *mean* and the *standard deviation,* and these describe respectively the average value or central tendency of the variates and the dispersion of values to either side of the mean value. The calculation and use of these and other measures is described further below (Section IV).

Characteristics of frequency distributions may be succinctly summarised by these simple measures (Figure 3A). An even simpler form of statistical shorthand is the use of *modal class,* which is the class having the highest frequency. However these statistical shorthands may occasionally give rise to erroneous or misleading conclusions, as in the case of frequency distributions which are asymmetric or *skewed,* and frequency distributions which have more than one modal class (bimodal, trimodal etc; Figure 3B). The mean provides an entirely valid indication of central tendency only for symmetric unimodal frequency distributions, although its use is often tolerated for slightly skewed but nevertheless unimodal frequency distributions. It is therefore important that the nature of frequency distributions be examined before such descriptive statistics be employed.

Descriptive use of histograms is very common in geographic research; some examples from human geography include work of Boddy (1976), who used histograms to compare Building Societies and Local Authorities in terms of the purchase prices of property funded, and the age, income and socio-

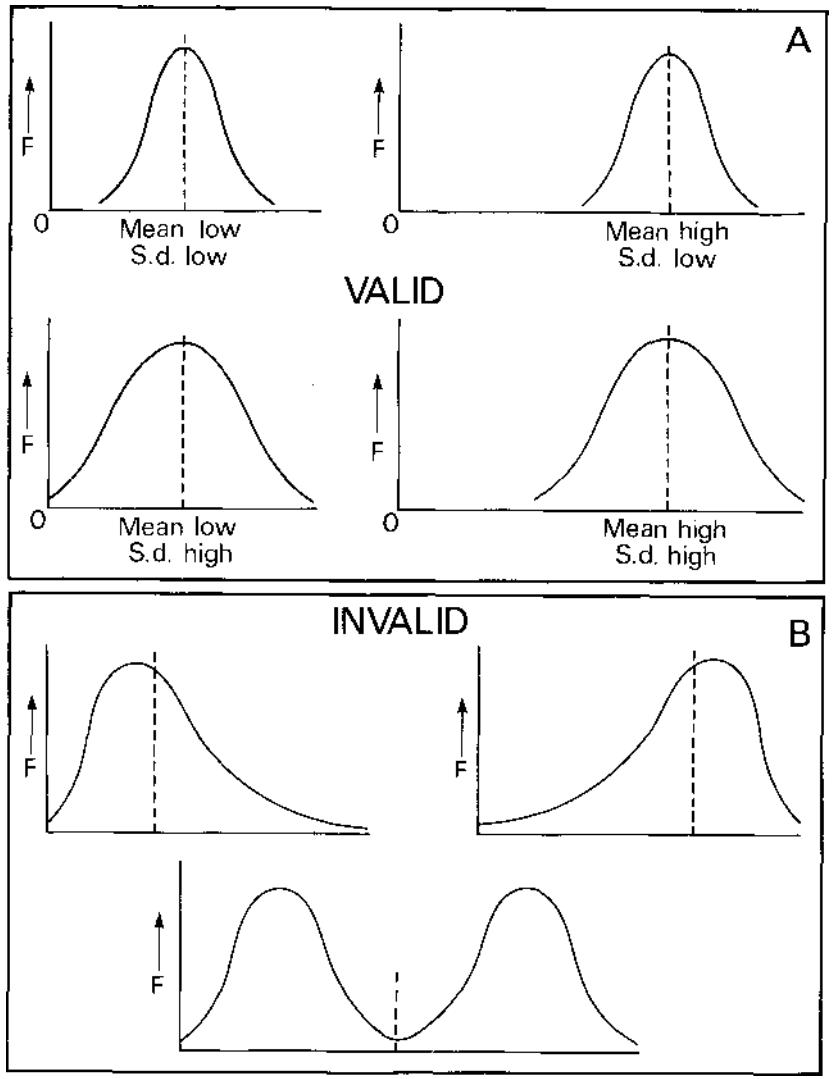

Figure 3. Mean and standard deviation as descriptions of frequency distributions.
A. Combinations of high and low means and standard deviations
B. Frequency distributions for which the mean is not a valid indication of central tendency of the distribution.

economic group of their borrowers; of Rowley (1975), who presented histograms of the population sizes of English constituencies in 1970 and 1974 in order to demonstrate the effects of electoral re-apportionment; and of Morrill and Symons (1977), who used histograms to describe aspects of time and distance to travel to a set of facilities in a study of optimum location. Some examples from physical geography include the work of Walling and Webb (1975), who presented histograms of solution load in samples of stream water from various rock and land use types; of Gardiner (1976), who used histograms of various land form parameters to examine differences between Land Systems; and of Hill (1973), who used histograms of various characteristics of drumlins to suggest possible modes of drumlin formation.

(ii) Inferential statistics.

Statistical tests are often employed either to establish whether the differences between two samples of variates are significant or whether they could have arisen by chance, or to test whether there is a significant co-variation between two variables for a given set of individuals. For example, one may wish to test the hypothesis that the mean values of rainfall at Herrington Park and North Tamerton (Table 1, Figure 1) differ significantly, or alternatively that there is some degree of correlation between them. The tests employed are of two types, parametric and non-parametric. *Parametric* tests are usually the more powerful but their underlying theory is based on the assumption that the variates considered have a frequency distribution of a particular type called a *normal* or *Gaussian* distribution. A normal frequency distribution has the properties of being symmetric and unimodal and a certain proportion of variates will occur within specified limits above and below the mean, as specified in Figure 4. It is often described as being 'bell-shaped'; a more rigorous definition is given in the Glossary.

Strictly, parametric tests should only be used on data which are normally distributed. No data set likely to be encountered will be distributed exactly as the normal distribution, but it is important that parametric tests should not be used on variates which are manifestly non-normally distributed, since the statistical procedure is based upon an assumption of normality and violation of this assumption invalidates any conclusions which may be drawn from the analysis. However non-parametric statistical tests may be used whatever the form of the frequency distribution, but results may be less precise than in the case of parametric tests. Non-parametric tests are however equally affected by violation of the assumption of independence (see Section V).

Again there are many instances of the use of inferential statistics relating to frequency distributions in geography. For example Gardiner (1971) established that the frequency distributions of estimated drainage density on three rock types were different, and Boots (1977) generated cellular networks by, three processes and compared their frequency distributions of contact numbers.

(iii) Causal explanation of spatial processes.

Geographical processes resulting in spatial distributions are difficult to observe and examine and recourse is often made to examination of spatial patterns as a surrogate for examination of their causative processes.

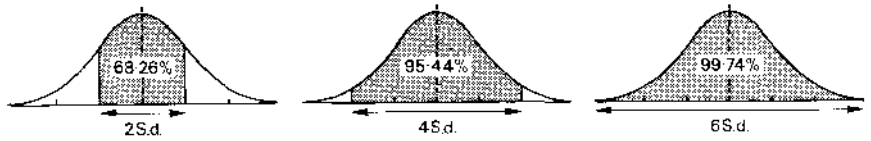

Figure 4. The percentage of variates within 1, 2 and 3 standard deviations of the mean, for the normal distribution. The mean is indicated by a dotted line, the ranges are in standard deviations and the shaded areas represent the appropriate area under the curve and hence the percentage of variates within the given limits, which are symmetric about the mean.

For example the spatial pattern of settlements may be studied as a substitute for and supplement to the more fundamental economic, social, political and physical processes governing their distribution. This may be performed by examining characteristics of the frequency distributions of certain aspects of the spatial pattern of settlements. For example it has been found that if regular quadrats are overlaid on an area and the number of settlements occurring in each cell is counted, then by regarding these counts as variates in a frequency distribution, the form of the resulting frequency distribution may give an indication of particular spatial processes such as diffusion or random placement. In a random pattern with a relatively small number of settlements per quadrats the frequency distribution of points per quadrat approximates to a particular type of distribution known as a Poisson distribution, and the closer the observed distribution approaches this the more reasonable it is to postulate a spatially random process underlying the generation of settlements. An example of the use of these techniques is work by Harvey (1966), in which the frequency distributions of counts of points in quadrats are used to test models of spatial diffusion. The techniques employed are discussed further in Harvey (1968) and a full account of them is given by Rogers (1974).

Spatial processes may also be inferred from frequency distributions in other ways. For example Afolabi Ojo (1973) has suggested a distance growth decay process to apply to the journey to work in Yorubaland from a consideration of the frequency distribution of distance travelled, and de Smith (1977) has considered a variety of uses to which properties of the frequency distributions of distances between all points in a region may be put.

(iv) Other uses of frequency distributions in geography.

Frequency distributions may be employed in many other areas of geography (Figure 5). For example Taylor (1971) has suggested a method for the measurement of the shape of geographical regions. This requires the shape to be overlain by a regular grid of points, either physically, or notionally by means of a computer program. Distances between all possible pairs of points within the shape are calculated or measured and regarded as variates in a frequency distribution. Figure 5A illustrates

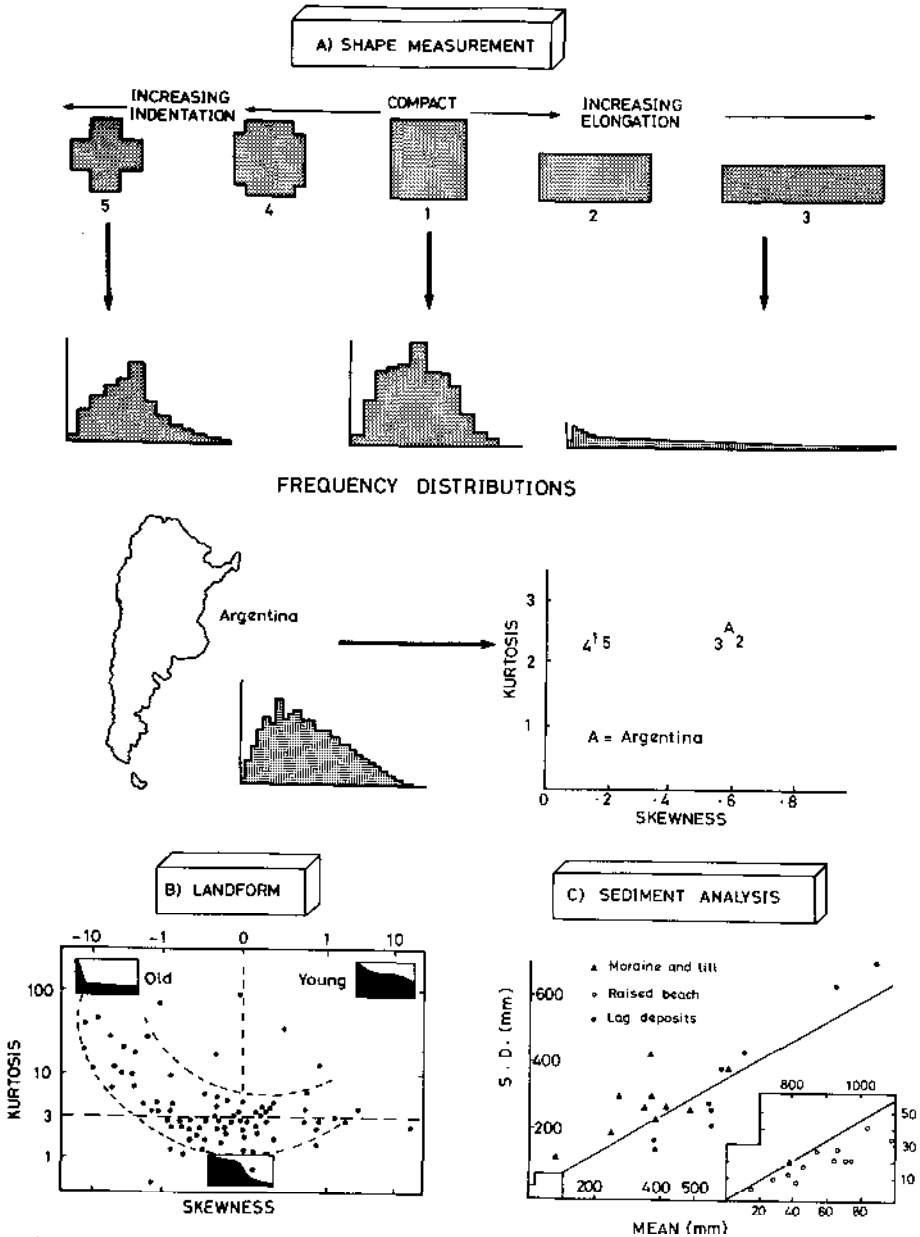

Figure 5. Geographical applications of frequency distributions. A. Shape measurement, employing the frequency distributions of distances between all pairs of points within shapes (after Taylor, 1971). B. Landform analysis, Each poind represents a sample of spot heights (after Tanner, 1960). C. Sediment analysis. Each point represents a sample of pebbles, the depositional environments of which are shown by types of symbol (after Doorn amp and King, 1971).

the way in which this may be used to assess how Argentina compares in shape with a set of geometric figures (after Taylor, 1971). A similar method has been proposed (Ongley, 1970) for measurement of drainage basin shape, and Boots (1977) uses the frequency distributions of contact numbers of cellular networks as an indication of their overall shape and formative processes. In geomorphology Tanner (1959; 1960) has suggested that the frequency distributions of samples of altitude may be diagnostic of certain stages of development of fluvial landforms (Figure 5B), and Evans (1972) has further investigated this field. Frequency distributions of characteristics of sediments such as grain size or particle shape have been widely applied in studies of depositional environments (Figure 5C), where geomorphologists have employed many geological techniques. Sampling designs for spatial data have been compared by examination of the frequency distribution of the resulting variables in simulation studies (e.g. Keyes et al., 1976), and finally studies of drainage network topology (e.g. Krumbein and James 1969) have often been able to use frequency distributions of network characteristics such as link lengths and basin areas to investigate network development.

III GRAPHICAL METHODS FOR EXAMINATION OF FREQUENCY DISTRIBUTIONS

The purpose of summarising the principal characteristics of a frequency distribution is to make it easier to interpret and to aid comparison with other distributions. In this way the complexity of the information may be considerably reduced, as individual values are generalised, whilst its usefulness is enhanced as the general pattern of the data is retained.

Frequency distributions may be regarded as simple summaries or generalisations of the original data. In the example of rainfall (Figure 1), the histograms afford a simple visual impression of annual rainfall at the two stations, and values of the two stations may usefully be compared to each other. Thus Werrington Park appears to have a higher mean value of rainfall, and a greater dispersion of values than North Tamerton.

(i) Construction of histograms

The procedure for constructing histograms first involves grouping the data into a number of classes and counting the number of variates which occur in each class. The frequency of occurrence is represented in the histogram by the height of the appropriate bar. Thus individual data are grouped and then illustrated graphically.

The number of classes used can vary according to the characteristics of the data, provided the whole range of values is covered. Obviously distributions which contain large numbers of variates normally justify use of more classes than those with few, but it is important to realise that the choice of class size may have a great effect on the mode and range obtained. One drawback of a grouping may be that too much information is lost; a careful balance has to exist between generalisation and amount of information retained. In Figure 6 the Werrington Park total annual rainfall data are used to illustrate a process of controlled generalisation.

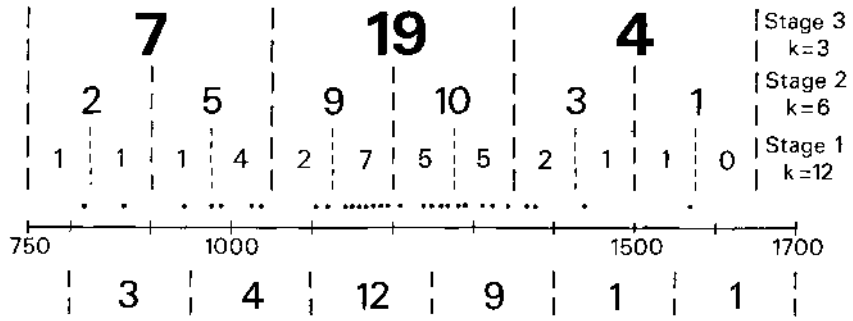

Figure 6. Grouping of variates into frequency classes (Werrington Park data).

In the centre of Figure 6 the 30 variates are plotted individually along a line representing the range of data. The 30 variates are then allocated to classes of equal width covering the whole range of the data, in the top of the figure. For any one of these classes there is now a frequency count of the number of variates in that class; these counts may then be represented graphically by a frequency histogram (e.g. Figures 1 and 7). Three sets of differently-sized classes have been used to derive frequency counts for these data, and these result in progressively more generalised histograms as the class width is increased. It is possible to assess the loss of information resulting from such class grouping by comparing the sum of grouped values at each stage, employing the group mean values as an indication of the variates within each class. For the ungrouped data the sum of the variates, $x_1, x_2, x_3 \ldots x_{30}$ is given exactly as:-

$$\sum_{i=1}^{i=30} x_i = 35{,}583$$

The '$\sum$' (sigma) sign implies summation, the values of i between which the x_i values are to be summed are indicated above and below the sigma.

In the generalised frequency histograms this sum can only be approximated since the magnitudes of the original values are concealed, and in Figure 7 instead of dealing with the original 30 variates it is necessary to deal with only 12 (Stage 1), 6 (Stage 2) or 3 (Stage 3) representative values for the classes. Assuming that the class midpoint values are representative of the variates contained in each class, then the sum of grouped variates may be approximated as:-

Stage 1. 12 clasess (11 if zero class discounted).

$$\sum_{j=1}^{j=12} C_j = 35{,}425$$

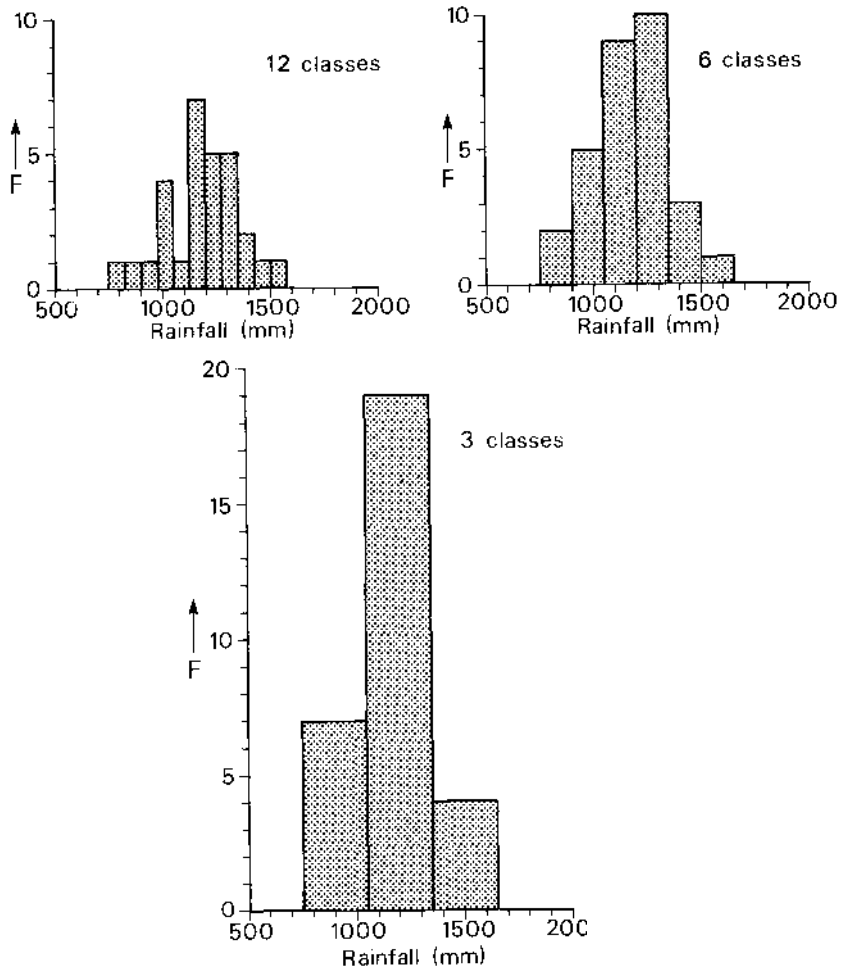

Figure 7. Histograms of Werrington Park data employing differently sized class intervals.

Stage 2. 6 classes.

$$\sum_{j=1}^{j=6} C_j = 33{,}775$$

Stage 3. 3 classes.

$$\sum_{j=1}^{j=3} C_j = 35,100$$

where C_1, $C_2 \ldots C_K$ refers to the product of class midpoint value and number of variates in that class, for the K classes in the appropriate stage of generalisation. This arithmetic procedure is illustrated in Table 2. The difference between these values and the corresponding sum for the ungrouped data is one illustration of the overall loss of information associated with grouping the data. It may be noted that there is unfortunately no simple relationship between loss of information and number of classes adopted, and that it is possible for the calculated sum to be the same as that for the raw data, despite the fact that some information loss must inevitably occur in any such grouping of data.

The number of classes selected is generally a subjective judgement on the part of the researcher but if the data are to be used for further analysis beyond the simple descriptive stage it is essential that in the construction of the histogram the class interval remains constant, and also that the whole range of original observations is covered. One way of arriving at a sound grouping is to experiment with different class intervals and to select that which appears to give the most satisfactory result. However there are more objective guidelines available. For example Huntsberger (1961) gives the following formula for estimating the number (K) of classes to be used

where n = the total number of variates.
Table 3 illustrates the use of this formula in constructing a frequency histogram of the rainfall data referred to above. Six classes are selected as optimum by this formula. For larger data sets the suggestion of Huntsberger obviously leads to an underestimate. For example a histogram of 10,000 variates would require about 14 classes according to the formula, yet intuitive judgement would suggest this to be an over-generalised representation of the data. An alternative approach is offered by Brooks and Carruthers (1953) who suggest:-

Application of this relationship is also illustrated in Table 3, resulting in a value of 7.4, i.e. 7 classes. A third procedure has been suggested by Norcliffe (1977) who advocated the use of

which enables a greater number of classes to be employed for larger datasets than in the' cases of the previous suggestions. Finally Croxton and Cowden (1948) suggest that most frequency histograms should have between six and sixteen classes. Clearly there can be no 'correct' number of classes since loss of detail will vary with different datasets, but care should always be taken to choose the number of classes so that a compromise is established between very few classes, showing relatively little, and too many classes, giving too detailed a classification. In the case of the rainfall data, with 30 variates, 6 or 7 classes would be appropriate, as used in Figure 1.

Table 2. Information loss for different class sizes, using Werrington Park data.

Ungrouped data. $\sum_{i=1}^{i=30} x_i = 35,583$

Grouped data.

12 groups.

(1) Class mean value	(2) Frequency in class	Product of (1) x (2)
1537.5	1	1537.5
1462.5	1	1462.5
1377.5	2	2755.
1312.5	5	6562.5
1237.5	5	6187.5
1162.5	7	8137.5
1077.5	2	2155.
1012.5	4	4050.
937.5	1	937.5
862.5	1	862.5
777.5	1	777.5
		35,425.0

Difference from ungrouped data, as % of latter 0.44%

6 groups.

1575	1	1575
1425	3	4275
1275	10	12750
1125	9	10125
975	5	4875
825	2	1750
		33775

Difference from ungrouped data, as % of latter 5.08%

3 groups.

1500	4	600
1200	19	22800
900	7	6300
		35,100

Difference from ungrouped data, as % of latter 1.36%

Table 3. Determination of number of classes in histogram.

Huntsberger method. No. of classes (K).

$K = 1 + 3.3 \log_{10} n \quad (n = 30).$

$K = 1 + (3.3 \times 1.47712)$

$K = 1 + 4.9 \quad = 5.9$ classes

Brooks and Carruthers method.

$K \leq 5. \log_{10} n$

$K \leq 5 \times (1.47712) = 7.4$ classes

Croxten and Cowden 6 to 16 classes

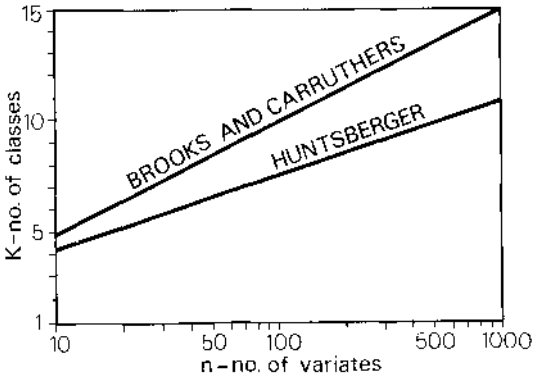

Figure 8. Graph to facilitate choice of number of histogram classes.

Two of the methods for calculation of number of classes are made simpler to apply by use of Figure 8. This plots the relationships of Huntsberger and Brooks and Carruthers between K, number of classes, and n, number of variates, for values of n from 10 to 1000, and an appropriate number of classes can be read off from this graph by looking at the relevant value of n. It should be emphasised that the upper line represents a suggested upper limit for K, and any convenient value of K lying between the two lines would normally be suitable for smaller datasets; K may need to be increased for larger datasets, when Norcliffe's (1977) formula is more appropriate.

The shape of the resulting histogram is also influenced by the position of the class boundaries as well as by the class range. For example, in the

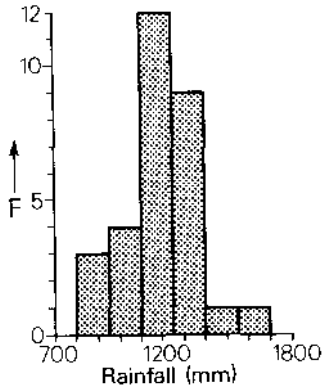

Figure 9. Histogram of Werrington Park rainfall data employing class limits shown on the bottom of Figure 6.

lower part of Figure 6 variates have been allocated to six classes the lowest of which commences 100mm higher than in the equivalent grouping, stage 2, of the upper part of the figure; this results in a histogram (Figure 9) which differs in detail, if not overall form, from that of Figure 1. No rigid rules can be given for selecting the positions of the class boundaries, but this should preferably be done with reference to the scattergram of values such as that in Figure 6, so that class boundaries do not pass through dense groups of points.

Evans (1977) has offered some suggested guidelines for selection of class intervals for maps, and this may be extended to the case of histogram preparation. He concluded that class intervals should be governed by the overall shape of the distribution. Rectangular distributions, in which all classes have approximately equal numbers of variates, require equal divisions of the overall range. Dominantly unimodal distributions require class intervals related to the standard deviation of the data, and skewed (in Evans' terminology J-shaped) distributions require class intervals which fall into a geometric progression, the base of which increases as skewness increases. This judgement does of course require some a priori appreciation of the overall shape of the distribution before the class intervals are selected and the histogram drawn; this would normally be gained from visual inspection of the scattergram of variates (e.g. Figure 6).

Specification of class interval limits also needs to be performed with respect to the distinction emphasised in Section I, between continuous and discrete data. The essential requirement is that all values should be assignable to classes without uncertainty. For example in Table 4 the classes are given as inclusive limits, and since annual rainfall is reported to the nearest millimetre this suffices to provide an unequivocal classification for the data. Since 0mm is theoretically possible the classes have been notionally commenced with 0 - 149mm; thus the first class for North Tamerton is 600-749mm rather than 601-750mm.

Further difficulties arise if it is wished to compare two samples with different numbers of variates in each, or with markedly different ranges. In these cases it is necessary to standardise the frequency or variable scale as appropriate. In the first case the frequencies may be converted into percentages of the total sample for each class in each

Table 4. Cumulative frequency data-groups

Werrington Park.

Class	Frequency	% Frequency	Cumulative % Frequency
750-899	2	6.67	6.67
900-1049	5	16.67	23.34
1050-1199	9	30.00	53.34
1200-1349	10	33.33	86.67
1350-1499	3	10.00	96.67
1500-1649	1	3.33	99.99

North Tamerton.

Class	Frequency	% Frequency	Cumulative % Frequency
600-749	6	20.00	20.00
750-899	13	43.33	63.33
900-1049	9	30.00	93.33
1050-1199	1	3.33	96.66
1200-1349	0	0.00	96.66
1350-1499	1	3.33	99.99

histogram, and percentage histograms employed. For the latter case the data values can be converted into standardised units such as standard deviations above and below the mean; this is described fully below (Section V(v)). Whilst it is preferable to use the same class intervals in each of two histograms to be compared (Mitchell, 1975) it may be necessary to ignore this principle, and it must be acknowledged that the majority of histograms represent a compromise between those features of the distribution it is wished to show, the desire to generalise the data, the need to retain as much information as possible, and the choice of easy to use and possibly meaningful class ranges.

(ii) Interpretation of histograms.

A histogram provides a simple visual impression of the distribution of a set of variates. For instance in the case of the frequency histogram of North Tamerton rainfall (Figure 1) there is a greater number of variates towards the lower end of the range and comparatively few at the other extreme. Asymmetrical distributions such as this are said to be skewed. If the tail

of the distribution is to the right, as in this case, the distribution is said to be *postively skewed*, with the modal class (that class which contains the largest number of variates) off-centre to the left (e.g. Figure 3B, left). Other distributions may be *negatively skewed*, with the modal class being towards the upper end of the range (e.g. Figure 3B, right). In general, frequency distributions encountered in geographic research show varying tendencies toward positive skewness, although exceptions do occur. Symmetric distributions (skewness = 0) may however still be non-normal owing to the existence of *kurtosis*. Kurtosis is often described as the 'peakedness' of a frequency distribution, although views have been expressed that it may alternatively be regarded as a measure of the extent of unimodality versus multimodality of the distribution (Darlington, 1970) and the length of its tails (Finucan, 1964). A *platykurtic* distribution, with low kurtosis, has less variates in the classes near the modal value and/or more substantial tails to the distribution (Figure 12). By contrast a *leptokurtic* distribution, with high values of kurtosis, has a larger than expected number of variates in the near-modal classes, and very long tails.

Skewness and kurtosis are usually considered as being separate and independent characteristics of frequency distributions. However some studies have shown that the two measures are related, and that the effect of *transformations* (Section V) on one of the measures is usually similar for the other measure. For example, Tanner (Figure 5B) has shown a series of distributions for which kurtosis has been plotted against skewness, revealing a generally U-shaped relationship between the two, and Gardiner (1973) demonstrated a similar U-shaped relationship between the two measures for various transformations of the same data, with minimum kurtosis occurring when skewness was nearest to zero.

The existence of skewness and/or kurtosis in geographical data is of significance when using inferential statistics in research procedures, as the more powerful parametric tests of statistical inference demand symmetry and particular assumptions of shape such that the variates concerned should be approximately *normally distributed*. The most important properties of the normal distribution are that it is symmetric, that inflections occur in the frequency distribution curve at one standard deviation above and below the mean and that the area under the curve (i.e. number of variates) within specified limits from the mean is as shown in Figure 4. Histogram shape, whether assessed visually or by means of indices such as those described in Section IV, is also of interest in that it may itself suggest useful information about the population concerned (Pringle, 1976). For example, Miller (1955) considered the positively skewed distribution of personal income in the United States and concluded that its skewness was due largely to merging several symmetrical distributions, namely those for men and women separately, and for individual occupational groups within the sexes. Skewed distributions for size-related variables also arise as a result of processes concerned with the packing together of individuals in space. For example, the frequency distributions of drainage basin area and related measures are invariably logarithmic or positively skewed (Gardiner, 1973) because the spatial packing of basins and the requirements of the random model of drainage composition (Smart, 1978) require many very small basins and declining numbers of large ones. Negatively skewed distributions are much rarer than positively skewed ones, although the size distribution of material on a boulder beach may follow this pattern, and Pringle (1976) noted that three of sixty-four census variables were also extremely negatively

skewed. Extremely positively skewed or J-shaped distributions, such as that of total population per grid square (Pringle, 1976), are commonly produced for count and density measurements for grid square data. Quotient measures for grid squares, such as percentage of total population in a particular age category, are less predictably distributed, however, and may be either approximately normal or skewed. Logarithmically distributed data are also very common. For example Desbarats (1976) reported the frequency of semantic response in a study of the perceived environment to follow a logarithmic distribution, and Gardiner (1973) concluded that many indices of drainage basin form were logarithmically distributed.

Bimodal distributions are particularly common in geomorphology; for example, the distribution of sediment particle size may be bimodal, as in the case of beaches which have much fine sand and larger boulders but little material of intermediate grade. Bimodality often implies a mixture of two populations, with similar frequency distribution shape characteristics but differing mean values. For example the frequency distribution of pebble shape in a river may be bimodal where a major tributary introduces material of a roundness differing from that of the trunk stream.

Kurtosis can also arise by the admixture of two populations, when the means are only slightly different but the standard deviations are the same. Thus a platykurtic unimodal frequency distribution is produced by mixing together two normal distributions with the same standard deviation but with means differing by twice the standard deviation (Bliss, 1967, p 141). Conversely a leptokurtic distribution is produced by mixing two populations with the same mean and differing standard deviations (Bliss, pp 141-2). Thus kurtosis in a distribution may also suggest that the sample is derived from two separate populations. Methods have been described for resolving suspected composite distributions into component parts; for example Mundry (1972) describes a method combining visual judgement and regression analysis, and gives references to other procedures based upon both graphical and numerical methods.

Visual examination of data in histogram form may provide an adequate preliminary indication of whether a distribution is approximately normal. For example in Figure 1 the distribution for Werrington Park is probably more normal than that for North Tamerton, but it must be emphasised that this is only a quick, simple and subjective visual assessment, and more rigorous tests of normality should usually be applied.

(iii) Cumulative frequency graphs and probability plots

An alternative graphical approach to the description and examination of geographical data is afforded by the *cumulative frequency graph.* Cumulative frequencies may be plotted either for ungrouped data which have been ranked by value or for grouped data. For example in Table 4 the grouped frequency data used in constructing the histograms of Figure 1 have been converted into percentages of the total number of variates in the distribution. The percentage frequencies are then cumulated upwards, giving percentage figures for the proportion of occurrence in a particular class plus all lower value classes. These cumulative percentages are then plotted against the upper class limit, producing the curve illustrated in Figure 10A. Such curves are often known as *ogives,* although a better term is distribution curve. Unimodal distributions result in S-shaped (sigmoidal) distribution

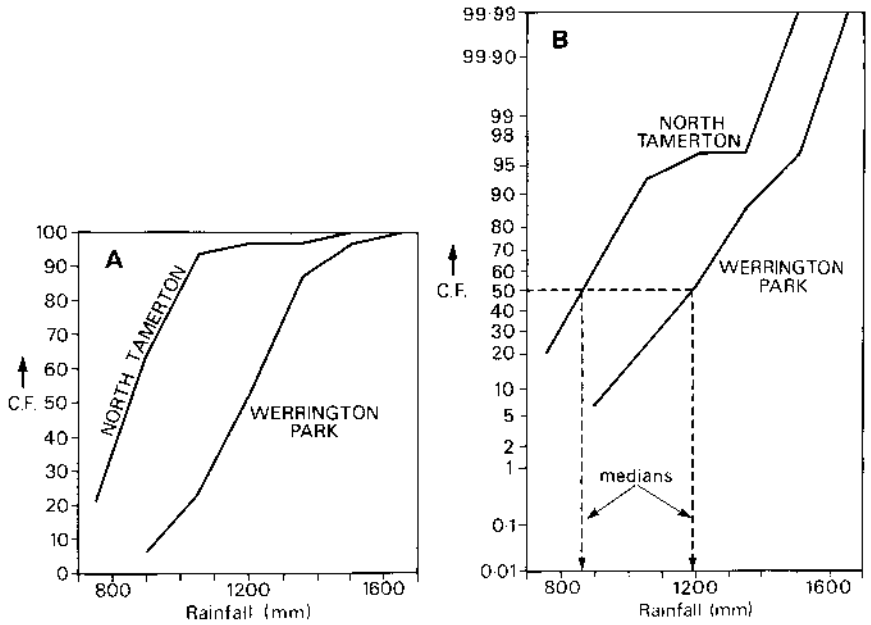

Figure 10. Cumulative frequency plots of rainfall data employing grouped data.
A. On ordinary scales.
B On probability scale.

curves; however departures from the normal are not easy to judge, and a further modification to the technique has been devised to facilitate comparison with the normal distribution. In this the cumulative percentages are plotted against class limits in the same way, but on graph paper known as *probability paper.* This has percentage probabilities set out along the vertical axis, and the horizontal axis is regularly divided and used to plot class upper boundary values. The paper is so designed that cumulative percentages representing a normal distribution, when plotted against upper class boundary values, will fall along a straight line. In Figure 10B the cumulative percentage frequencies of the Werrington Park data fall almost exactly on a straight line, indicating a distribution very close to normal. The other distribution diverges markedly from a straight line, and therefore from a normal distribution, thus confirming the visual impression gained from the histograms. Again, as with the construction of frequency histograms, a process of controlled generalisation applies, according to the size of frequency grouping selected. Obviously a more accurate indication of the nature of the distribution may be obtained by plotting many small groups; this leads eventually to the use of the original variates (Table 5). These may be ranked, from which cumulative percentages are calculated and plotted against the appropriate values. Again a straight line represents normality. Figure 11 shows the distribution of rainfall for North Tamerton plotted in

Table 5. Cumulative frequency data - individual variates.

North Tamerton

Year	Rank	Value	% Frequency	Cumulative % Frequency.
1953	30	637	3.33	3.33
1933	29	682	3.33	6.66
1949	28	688	3.33	9.99
1944	27	721	3.33	13.33
1945	26	726	3.33	16.66
1941	25	737	3.33	19.99
1956	24	774	3.33	23.33
1942	23	778	3.33	26.66
1947	22	792	3.33	29.99
1952	21	795	3.33	33.33
1940	20	815	3.33	36.66
1943	19	842	3.33	39.99
1948	18	843	3.33	43.33
1938	17	844	3.33	46.66
1934	16	849	3.33	49.99
1936	15	854	3.33	53.33
1951	14	882	3.33	56.66
1932	13	889	3.33	59.99
1957	12	895	3.33	63.33
1931	11	914	3.33	66.66
1955	10	926	3.33	69.99
1935	9	927	3.33	73.33
1939	8	981	3.33	76.66
1950	7	994	3.33	79.99
1958	6	998	3.33	83.33
1954	5	1007	3.33	86.66
1946	4	1008	3.33	89.99
1937	3	1017	3.33	93.33
1959	2	1034	3.33	96.66
1960	1	1362	3.33	99.99

this way, and although considerable non-normality is still evident, it is seen that the markedly non-normal appearance of Figure 10B is due partly to the grouping into artificial classes.

An advantage of this method of representation of the data is that it models the discrete data by a continuous function, the scatter of points being generalised by a line. From this intermediate values may be read off, to enable standardised comparisons to be made between a number of distributions. For example from Figure 10B it may be determined that a point on the line representing the Werrington Park distribution with a cumulative frequency of 50% also represents a value of the variable, rainfall, of about 1190mm. Thus, if this generalised line is representative of the variable, then fifty per cent of years will have greater than 1190mm. rainfall, and fifty per cent less. 1190mm. therefore represents a midway or *median* value, which can be compared from one distribution to another. By comparison the median of the North Tamerton data is 865 mm.

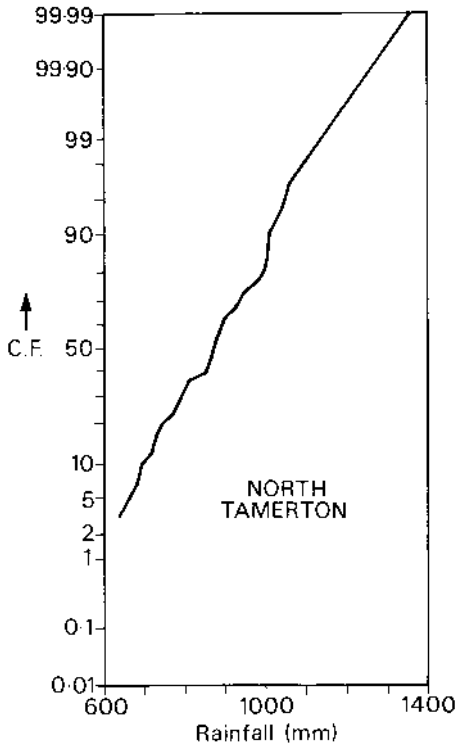

Figure 11. Cumulative frequency plot of North Tamerton data on probability paper, employing individual values.

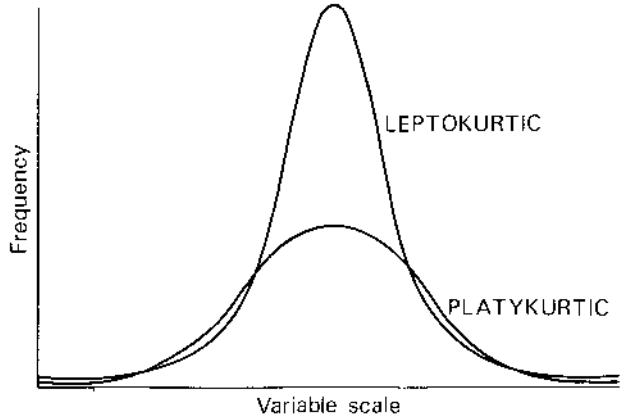

Figure 12. Kurtosis in frequency distributions.

This process need not be confined to determination of the median, since any other percentage frequency values may be read from the graph. Commonly used values are the 25% and 75% points (lower and upper quartiles).

The percentage cumulative frequency graph has been particularly widely used in sedimentology, for the portrayal of grain size distributions. In this field a wide variety of descriptive measures of size distributions have been based on various combinations of percentage points. These are described further below (Section IV).

(iv) Conclusions

Graphical representations of geographical variates, in histogram, cumulative frequency and probability plot form, provide simple and effective means for description and elementary analysis of the data. Of these methods the latter offers the most rigorous and useful approach and enables many distributions to be plotted on one graph for comparative purposes, unlike histograms which have to be compared separately. By the use of these graphical methods complex data may be summarised and compared to other distributions, and the complexity of the data is considerably reduced whilst the most important elements are retained. Some subjective impression of the statistical normality of the data may also be gained, especially if cumulative frequency graphs are employed: more sophisticated graphical methods also exist (e.g. Tukey, 1957).

However these methods are essentially descriptive tools, allowing only a subjective visual comparison of distributions. For more exact comparisons, either between distributions or with the normal distribution, it has been widely accepted that inferential statistics must be employed. These are reviewed in the succeeding sections. However before leaving graphical procedures it is worth noting that some statisticians (e.g. Gnanadesikan, 1977, pp 167-8; Wilk and Gnanadesikan, 1968) are now less sanguine about the inevitable need for inferential tests and suggest that in many ways graphical procedures such as probability plots offer the safest approach to the problem.

IV COMPARISON BETWEEN OBSERVED FREQUENCY DISTRIBUTIONS AND THE NORMAL DISTRIBUTION

Having established in sections I and II a need for the examination of frequency distributions and having suggested in section III that purely visual methods are not fully adequate for this purpose it is now necessary to examine numerical procedures for comparing an observed frequency distribution with the normal distribution. However it must be emphasised that the inferential statistics to be described must not be used blindly. If a sample of variates is such that it is not possible to reject the null hypothesis of no difference from normal, at the chosen significance level, the worker must not necessarily immediately accept the sample as being perfectly normal - it is merely of unproven non-normality, not proven normality. This consideration is particularly relevant to small samples, for which it is perhaps more appropriate to think in terms of degree of non-normality rather than the significance of non-normality. Many methods are available for comparing an observed frequency distribution with the normal

Table 6. Methods for comparing frequency distributions

Aspect considered	Graphical methods and descriptive statistics	Statistical tests	
Skewness	Histogram	Fisher's g_1	
	Pearson's skewness		
	Graphical skewness from cumulative probability plot		
	Momental skewness		
		Small samples	Large samples
Kurtosis	Histogram	Geary mean deviation ratio	Range/s.d.test Fisher's g_2
	Graphical kurtosis from cumulative probability plot		
	Momental kurtosis		
Overall goodness of fit	Histogram	Shapiro & Wilk's W	Kolmogorov-Smirnov test Chi-squared test Chi-squared using log-likelihood ratio *Joint Chi-squared of g_1 and g_2
	Cumulative probability plot		
Outliers	Histogram	Dixon's gap test	Range/s.d. test

* not strictly goodness of fit.

distribution (e.g. Gnanadesikan, 1977, pp 162-8) but those which are most likely to be met within geographical work are outlined in Table 6. Statistical tests based upon viewing the data as continuous data are first examined. Overall goodness-of-fit tests based upon non-parametric measures are then discussed; this latter procedure models the normal frequency distribution by discrete data in groups and is therefore less likely to be as precise as the exact measures. Exact measures are only able to consider certain aspects of the observed distribution, however, and not its overall shape in relation to the normal.

(i) Analysis based upon moment measures.

Moment measures of frequency distributions have been mentioned in section III in the context of their uses for the description of the shape of frequency distributions. Four moment measures are normally recognised and may be defined as:-

$$M_1 = \frac{[u^1]}{n} \qquad M_2 = \frac{[u^2]}{n}$$

$$M_3 = \frac{[u^3]}{n} \qquad M_4 = \frac{[u^4]}{n}$$

where u stands for the deviation of each value from the mean and [ ] the symbol implies summation over all n variates in the distribution concerned. Thus in more conventional, but unwieldy terminology:-

$$M_1 = \frac{\sum_{i=1}^{i=n} (x_i - \bar{x})}{n}$$

$$M_2 = \frac{\sum_{i=1}^{i=n} ((x_i - \bar{x})^2)}{n}$$

$$M_3 = \frac{\sum_{i=1}^{i=n} ((x_i - \bar{x})^3)}{n}$$

$$M_4 = \frac{\sum_{i=1}^{i=n} ((x_i - \bar{x})^4)}{n}$$

where x_i = the ith variate of a distribution with n variates and a mean $\bar{x}$.

The first moment, $[u^1]/n$, is given for completeness but is of course equal to zero. Moments are usually further developed into the standard deviation or variance for moment two, the skewness for moment three, and the kurtosis for moment four. Higher power moments also exist but are very seldom employed. Standard deviation, (s.d.) is given by:-

$$\text{s.d.} = \sqrt{M_2}$$

skewness (Sk) by:-

$$Sk = \frac{M_3}{M_2^{3/2}}$$

and kurtosis (KT) by:-

$$Kt = \frac{M_4}{M_2^2}$$

These are the usual descriptive momental measures employed and give dimensionless measures of distribution shape. For a normal distribution the skewness is obviously zero, as the distribution is symmetric, and the kurtosis value is three. Skewness may be positive or negative, indicating the direction of asymmetry of the distribution (Figure 3) ; kurtosis can only be positive in value.

To test the significance of departures of a frequency distribution from the normal in terms of asymmetry and kurtosis a series of measures based upon the moments described above has been proposed. Using the notation as above these may be calculated from a set of statistics known as Fisher's K statistics:-

$$K_2 = \frac{[u^2]}{n-1}$$

$$K_3 = \frac{[u^3]\ n}{(n-1)(n-2)}$$

$$K_4 = \frac{([u^4]\ n^2 + n))/(n-1) - 3[u^2]^2)}{(n-2)\ (n-3)}$$

From these K statistics two statistics, g_1 and g_2, are calculated as:-

$$g_1 = \frac{K_3}{\sqrt{K_2^3}}$$

$$g_2 = \frac{K_4}{K_2^2}$$

The significance of departures of these g statistics from normality are then

tested by two further values, calculated as:-

$$V_1 = \frac{6n(n-1)}{(n-2)(n+1)(n+3)}$$

$$V_2 = \frac{4(n^2-1)V_1}{(n-3)(n+5)}$$

These are the sampling variances of the statistics g_1 and g_2 for a normal distribution, and are obviously dependent only upon the sample size, n. From these variances the ratios:-

$$X_1 = \frac{g_1}{\sqrt{V_1}}$$

and $X_2 = \frac{g_2}{\sqrt{V_2}}$

are calculated and finally referred separately to a table of the *normal probability integral* (such as Table 4A in Bliss (1967)). In this table the calculated value is regarded as the standardised normal deviate, ignoring the sign, and the probability read from the table is doubled, because this is a two-tailed test. Probability values smaller than the selected significance level require rejection of the null hypothesis of no significant difference from the normal distribution.

If only one of the two test statistics indicates a significant difference from the normal distribution then a combined Chi-squared statistic is calculated from:-

$$\chi^2 = \frac{g_1^2}{V_1} + \frac{g_2^2}{V_2}$$

and is referred to a χ^2 table (such as Table C in Siegel (1956)), with two degrees of freedom. A value of χ^2 greater than the tabulated value at the chosen significance level requires rejection of the null hypothesis of no difference between observed and normal distributions. This rather lengthy procedure is illustrated in Table 7. For most applications the necessary arithmetic is best performed by computer. Methods are also available (e.g. Cole and King, 1968, pp 111-5; Bliss, 1967, pp 140-6) by which these measures may be calculated for grouped data.

(ii) Graphical estimates of moment measures

Approximations to moment measures may be obtained by graphical means, using information derived from the cumulative probability plots previously outlined. It was suggested in Section III that intermediate values of the variable scale could be read off for any percentage probability, representing a value of the variable below which the stated percentage of the sample lies. Thus in Figure 10 for Werrington Park 25% of the variates lie below

Table 7. Calculation of moment measures.

Werrington Park. (Calculations are shown in detail for only the first and last variates in the distribution; all summation signs are understood to apply to summation over all thirty variates.)

Mean = $\bar{x}$ = 1186.1 mm n = 30

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$	$(x_i - \bar{x})^4$
1316	129.9	16874.0	2191933.9	284732213.5
⋮	⋮	⋮	⋮	⋮
1568	381.9	145847.6	55699202.3	21271525342.7
35583	0.0	808626.7 = $[u^2]$	-23457671.04 = $[u^3]$	650175734l3.81 = $[u^4]$

$$M_2 = \frac{[u^2]}{n} = 269542.2$$

$$M_3 = \frac{[u^3]}{n} = -781922.3$$

$$M_4 = \frac{[u^4]}{n} = 2167252447.13$$

Standard deviation $= \sqrt{M_2} = 164.18$

$$\text{Skewness} = \frac{M_3}{M_2^{3/2}} = -0.18$$

$$\text{Kurtosis} = \frac{M_4}{M_2^2} = 2.98$$

Significance testing.

Fisher's k statistics

$$k_2 = \frac{[u^2]}{n-1} = 27883.7$$

$$k_3 = \frac{[u^3]\,n}{(n-1)(n-2)} = -866662.7$$

$$k_4 = \frac{([u^4]\,(n^2+n))/((n-1) - 3 \times [u^2]^2)}{(n-2)(n-3)} = 163247222.2$$

$$g_1 = \frac{k_3}{k_2^{3/2}} = -0.186$$

Table 7 (continued)

$$g_2 = \frac{k_4}{k_2^2} = \underline{0.21}$$

$$V_1 = \frac{6n(n-1)}{(n-2)(n+1)(n+3)} = \underline{0.182}$$

$$V_2 = \frac{4 \times (n^2-1) \times V_1}{(n-3)(n+5)} = \underline{0.694}$$

$$X_1 = \frac{g_1}{\sqrt{V_1}} = \underline{-0.436}$$

$$X_2 = \frac{g_2}{\sqrt{V_2}} = \underline{0.252}$$

From the table of the normal probability integral (e.g. Bliss, 1967; Table A4) the probabilities associated with these values are, for X_1, 0.33 x 2 = 0.66., and for X_2, 0.40 x 2 = 0.80. The null hypothesis of no difference between the observed and the normal distribution cannot therefore be rejected at, say 0.01, as the tabulated value is larger than this.

North Tamerton

Equivalent values for North Tamerton are:-

Mean = 873.7mm Standard deviation = 141.44mm

Skewness = 1.11 Kurtosis = 5.54

X_1 = 2.74 X_2 = 3.90

From the table the associated probabilities are:-

X_1 = 0.003 x 2 = 0.006

X_2 = 0.0048 x 2 = 0.00096

As both of these are less than, say 0.1, the null hypothesis may be rejected, and this distribution does appear to be significantly non-normal.

If only one of these statistics had indicated non-normality then the joint statistic could have been calculated as:-

$$\chi^2 = \frac{g_1^2}{V_1} = \frac{g_2^2}{V_2} + \frac{-0.186^2}{0.182} + \frac{0.21^2}{0.694} = \underline{22.73}$$

The tabulated value of χ^2 at 0.1 significance level and two degrees of freedom is 9.21, and the null hypothesis of no difference between the normal distribution and the North Tamerton data may therefore be rejected at this level.

1060 mm rainfall, therefore 1060mm rainfall may be said to represent the 25th *percentage* point of the distribution, it being understood that cumulation is carried out in the manner described. The terminology employed here for this will be P25. Therefore for the distributions in Table 1 P25 for Werrington Park is 1060mm and P25 for North Tamerton is 775 mm.

By selecting certain critical percentage values a series of procedures has been proposed by which estimates of the four moment measures may be derived from the cumulative probability plot. A similar procedure is widely used in sedimentology, where grain size data are usually measured on a logarithmic scale called the phi (0) scale.

The mean. Reading off the P50 value yields the median as an indication of central tendency. However this is seriously in error as an estimate of the mean for skewed distributions (Figure 3B), and Folk and Ward (1957) suggested that the tails of the distribution could be incorporated into the calculation of the mean by:-

$$\text{Mean} = \frac{P16 + P50 + P84}{3}$$

McCammon (1962) evaluated this as the most effective existing graphical measure of the mean but proposed a slightly better measure as:-

$$\text{Mean} = \frac{P5 + P15 + P25 + P35 + P45 + P55 + P65 + P75 + P85 + P95}{10}$$

Standard deviation. This has been the subject of much investigation by sedimentologists since when applied to grain size data it gives an indication of sorting, which is a key element in genetic interpretation.

Folk and Ward (1957) suggested the use of:-

$$\text{Standard deviation} = \frac{P84 - P16}{4} + \frac{P95 - P5}{6.6}$$

McCammon (1962) again concluded that this was the most effective existing measure but suggested as a possible improvement:-

$$\text{Standard deviation} = \frac{P70 + P80 + P90 + P97 - P3 - P10 - P20 - P30}{9.1}$$

Skewness.

Folk and Ward (1957) suggest an index which they termed Inclusive Graphic Skewness, as:-

$$\text{Skewness} = \frac{(P16 + P84 - 2P50)}{2(P84 - P16)} + \frac{(P5 + P95 - 2P50)}{2(P95 - P5)}$$

This ranges from -1 to +1, with most values lying between -0.8 and +0.8.

Because the phi scale is reversed in that large phi values represent small grain diameters and vice versa the skewness measure derived from sedimentology as described above has a sign which denotes the reverse of that conventionally expected. Thus positive skewness from this graphical measure is equivalent to conventional negative skewness, and vice versa.

Kurtosis.

Folk and Ward (1957) suggest:-

$$\text{Kurtosis} = \frac{P95 - P5}{2.44\ (P75-P25)}$$

This ranges from a mathematical minimum of 0.41 to an observed maximum of about 8.0, with the normal distribution having a value of 1.0. These graphic measures are illustrated in Table 8, which suggests the two distributions to be different in form, that of Werrington Park having kurtosis and skewness close to the normal distribution and North Tamerton's distribution having slight negative (i.e. positive in conventional terms) skewness and low kurtosis. These graphical measures have not been widely used outside sedimentology, although there would seem to be scope for their application in at least the exploratory phases of investigations in other fields. Further variants of the indices exist (e.g. Tanner, 1960), especially in sedimentological literature, to which geographers will find an introduction in C.A.M. King (1966, pp 274-91).

(iii) Other methods of assessing skewness and kurtosis

Estimates of skewness may be rapidly obtained from the relationship between the mean and other measures of central tendency, as standardised by the variability, using the formulae:-

$$\text{Skewness} = \frac{3(\text{mean}-\text{median})}{\text{Standard deviation}}$$

$$\text{or Skewness} = \frac{\text{mean} - \text{mode}}{\text{Standard deviation}}$$

These dimensionless ratios are known as the Pearsonian coefficients of skewness and obviously afford only a very generalised indication of distributional asymmetry. The first one, which is the most widely used, ranges from -3 to +3, with zero representing a symmetric distribution and values between -1 and +1 representing only moderate skewness. The second index does not have defined limits and is therefore of lesser value.

A simple procedure for testing kurtosis has been suggested by Geary (1936). This is based upon the statistic:-

$$Q = \frac{\bar{d}}{s.d.}$$

where s.d. is the standard deviation and $\bar{d}$ is the mean deviation, i.e.

Table 8. Graphical estimation of moment measures.

Werrington Park

From Figure *10*:-

Mean

$$\text{Mean} = \frac{P16 + P50 + P84}{3}$$

From Figure 10, P16 = 1005mm

P50 = 1190mm

P84 = 1340mm

$$\therefore \text{Mean} = \frac{1005 + 1190 + 1340}{3} = 1178.3\text{mm}$$

$$\text{Similarly McCammon's mean} = \frac{870+1000+1060+1115+1160+1210+1240+1295+1345+1460}{10}$$

$$= \frac{11755}{10} = 1175.5$$

Standard deviation

$$\frac{P84 - P16}{4} + \frac{P95 - P5}{6.6} = \frac{1340 - 1005}{4} + \frac{1460 - 870}{6.6}$$

$$= 173.09$$

$$\text{McCammon's standard deviation} = \frac{1260+1315+1395+1505-835-945-1030-1085}{9.1}$$

$$= 173.6$$

Skewness

$$\text{Skewness} = \frac{(P16 + P84 - 2P50)}{2(P84 - P16)} + \frac{(P5 + P95 - 2P50)}{2(P95 - P5)}$$

$$= \frac{(1005 + 1340 - 2 \times 1190)}{2(1340 - 1005)} + \frac{(870 + 1460 - 2 \times 1190)}{2(1460 - 870)}$$

$$= -.094$$

Kurtosis

$$\text{Kurtosis} = \frac{P95 - P5}{2.44\ (P75 - P25)} + \frac{1460 - 870}{2.44\ (1295 - 1060)} = 1.03$$

North Tamerton

From Figure 11:-

Mean = 860.5mm

Standard deviation = 261.54mm

Skewness = -0.048

Kurtosis = 0.89

$$\bar{d} = \frac{\sum_{i=1}^{i=n} |x_i - \bar{x}|}{n}$$

For a normal distribution this ratio has a value of 0.7979. The significance of departures from this value may be tested by reference to tables such as that contained in Geary (1936), using (n-1) degrees of freedom. For platykurtic distributions the ratio is larger than 0.7979, and for leptokurtic distributions is less. An advantage of this statistic is that its sampling distribution has been tabulated down to very small values of n by Geary (1936); an illustration of its calculation and use is given in Table 9.

Table 9. Calculation of approximate indications of moment measures.

Werrington Park

Pearsonian Skewness = $\frac{3(\text{mean} - \text{median}}{\text{standard deviation}} = \frac{3(1186.1 - 1190)}{164.18} = -0.07$

Geary kurtosis

$$\bar{d} = \frac{\sum_{i=1}^{i=n} |(xi - \bar{x})|}{n} = \frac{3827}{30} = 127.57$$

$$Q = \frac{\bar{d}}{s.d.} = \frac{127.57}{164.18} = 0.7770$$

North Tamerton

Pearsonian Skewness = $\frac{3(873.7 - 865)}{141.44} = 0.19$

Geary kurtosis

$$\bar{d} = \frac{3204.4}{30} = 106.81$$

$$Q = \frac{\bar{d}}{s.d.} = \frac{106.81}{141.44} = 0.7552$$

From the table in Geary (1936), or Bliss (1967), Table A8, the upper and lower critical levels of the Q statistic are at the 0.05 significance level, 0.8625 and 0.7404, with 30 degrees of freedom. Since both calculated values are within the range given by these values it is not possible to reject the null hypothesis of no difference between the observed distribution and the normal distribution, for either set of data.

A further simple estimate of kurtosis, David et al.'s (1954) range/standard deviation test (Pearson and Stephens, 1964), is considered below, (the effect of outliers, Section IV (vi)).

The moment-based measures, especially kurtosis, should not normally be used for less than about 40 or 50 variates. For datasets of less than approximately 40 variates the procedure of Shapiro and Wilk (1965) given below should preferably be employed. An exception to this restriction is the Geary ratio, which can be applied to distributions with as few as ten variates.

(iv) The W test of normality (Shapiro and Wilk, 1965)

This is an exact test of normality for small samples and is based upon the second moment of the distribution, employing the following procedure:-

(1) Rank the observations in ascending order from x_1 to x_n

(2) Calculate the mean of the variates, $\bar{x} = \sum_{i=1}^{i=n} \frac{x_i}{n}$

(3) Calculate the total squared deviation from the mean for the distribution (i.e. $[u^2]$ using the terminology employed earlier).

$$[u^2] \quad \sum_{i=1}^{i=n} (x_i - \bar{x})^2$$

(4) Calculate a statistic b, as:-

$$b = \sum_{i=1}^{i=m} a_m (x_{n-i+1} - x_i)$$

where $m = n/2$, and a_1 to a_m represents a series of coefficients which are given in Shapiro and Wilk (1965). If n is odd, then $m = (n+1)/2$, and the final a coefficient is always zero.

(5) Calculate the ratio W, as:-

$$W = \frac{b^2}{[u^2]}$$

(6) Refer this W value to the table in Shapiro and Wilk (1965), in which smaller values of W than those tabulated indicate significant non-normality at the selected significance level. The necessary coefficients (a) and percentage points for W have been tabulated by Shapiro and Wilk for values of n up to 50. The procedure is illustrated in Table 10, from which it is clear that the test is one which has the advantage of considering the absolute values of all of the variates, in paired combinations reflecting their rank order within the overall distribution.

Table 10. Calculation of Shapiro and Wilk's W statistic.

North Tamerton.

Data are ranked in Table 5. Mean = 873.7mm.

Total squared deviation from mean $\left[u^2\right]$ = 600186.3

$l = n/2 = 30/2 = 15$

l	a_l (coefficients) (from Shapiro and Wilk, 1965)	x_i	x_{n-i+1}	$x_{n-i+1}-x_i$	$a_k x(x_{n-i+1}-x_i)$
1	0.4254	1362	637	725	308.42
2	0.2944	1034	682	352	103.63
3	0.2487	1017	688	329	81.82
4	0.2148	1008	721	287	61.65
5	0.1870	1007	726	281	52.55
6	0.1630	998	737	261	42.54
7	0.1415	994	774	220	31.13
8	0.1219	981	778	203	24.75
9	0.1036	927	792	135	13.99
10	0.0862	926	795	131	11.29
11	0.0697	914	815	99	6.90
12	0.0537	895	842	53	2.85
13	0.0381	889	843	46	1.75
14	0.027	882	845	38	0.86
15	0.0076	854	849	5	0.04
					744.16 = b

$$W = \frac{b^2}{\left[u^2\right]} = \frac{744.16^2}{600186.3} = 0.923$$

This value is *less* than the tabulated value of 0.927 at a 0.05 significance level, and the null hypothesis of no difference between the observed and normal distributions may be rejected.

For the Werrington Park data a W statistic of 0.9841 results; since this is greater than the tabulated value at the required significance level the null hypothesis may not be rejected.

(v) Non-parametric overall goodness-of-fit tests

The procedures examined above have regarded the variates as samples from an underlying continuous distribution. In the remaining set of procedures a non-parametric test is used to compare the observed frequency distribution, in terms of frequencies in classes, with a discrete representation of the normal distribution using the same classes and having the same mean and standard deviation as the observed distribution. The tests are therefore estimating the difference between two distributions with the same mean and dispersion but possibly different shapes, in other words the observed distribution and an 'expected' normal distribution.

Two non-parametric tests are commonly employed in this approach, the Chi-square and Kolmogorov-Smirnov tests, but the overall procedure may be summarised by the following steps:-

(1) Calculate mean and standard deviation of the data.

(2) Group the data into classes, as if preparing to draw a histogram. For Chi-square certain restrictions on the number of classes apply; these are detailed below.

(3) Obtain the 'expected' frequencies in each of the chosen classes, under the assumption of a normal distribution with the same mean and standard deviation as the observed distribution. This procedure, which yields the standard against which the observed distribution may be compared by one of the non-parametric tests, is also detailed further below.

(4) Calculate the appropriate test statistics according to the test chosen.

Fitting the normal distribution

This is done by using a table of the normal probability integral (e.g. Table A4 in Bliss (1967)). Stages in the procedure, which is illustrated by Table 11, are:-

(1) Convert the class boundaries into units of standard deviation above and below the mean.

$$z_i = \frac{t_i - \text{mean}}{\text{Standard deviation}}$$

where t_i = the value of the class boundary. These standardised values are termed z scores or standard deviates, and enable a standardised set of tables to be applied for all distributions, irrespective of units employed or range of values.

(2) Consult a table of the normal probability integral to find the normal probability that a value will fall between the two class boundaries of concern. This is equivalent to the area under the normal distribution curve (Figure 4), and is given in the normal probability integral table in units arranged so that the area under the total curve is 1.0.

Table 11. Calculation of χ^2 goodness-of-fit tables.

North Tamerton data.

Class limits (mm)	Class limits (z-score)	Normal probability integral (from tables)	Probability in each class	Expected frequency	Observed frequency	$f(=(O-E)^2/E)$
600	-1.94	.026190				
			.054567	1.64	1	0.250
675	-1.40	.080757	.108673	3.26	5	0.929
750	-0.88	.18943	.17750	5.33	5	0.020
825	-0.34	.36693	.20842	6.25	8	0.490
900	+0.19	.42465	.18889	5.67	3	1.257
975	+0.72	.23576	.13011	3.90	7	2.464
1050	+1.25	.10565	.068112	2.04	0	2.04
1125	+1.78	.037538	.027094	0.81	0	0.0
1200	+2.31	.010444	.0081883	0.25	0	0.0
1275	+2.84	.0022557	.0018799	0.06 (1.13)	0 (1)	0.0
1350	+3.37	.00037584	.003277	0.01	1	0.015
1425	+3.90	.000048096	.0000432	0.001	0	0.0
1500	+4.42	.000004935				
						7.465 = χ^2

The calculated value of χ^2, of 7.465, does not exceed the tabulated value, of 11.05, at the 0.05 significance level, with 5 degrees of freedom. The null hypothesis of no difference between the observed distribution and the normal distribution may not therefore be rejected.

Werrington Park data.

The χ^2 value for this data is 3.522. Again, therefore, the null hypothesis may not be rejected.

Each number in the table gives the area under the curve between the mean and the value of standardised deviate for a particular value of deviate, and the area between two values of standardised deviate (i.e. class boundaries) is found by subtraction of the two area values. (An obvious exception to this is the class containing the mean, for which the tabulated values for the class limits either side of the class containing the mean should be added together and subtracted from 1.0 to give the area appropriate to the central class containing the mean.)

(3) Check that these areas sum to 1.0. When cumulated for all classes the standard deviate values should total 1.0 but may often be slightly less as the theoretical normal distribution has tails which extend to infinity in either direction and these cannot be fitted to the observed distribution.

(4) Multiply the class probabilities (i.e. areas under the curve) by the total frequency, n, to give the 'expected' number of variates in each class, assuming that a discrete representation of the normal distribution is fitted to the data. In Table 11 the product of class probability, derived as described above, and sample size yields the values shown in the column headed 'expected frequency'.

The observed and expected distributions may be compared by the Chi-Squared test or the Kolmogorov-Smirnov test. If the latter is to be used for comparison this last step, step 4, is omitted and the expected proportions produced in step (2) are used instead.

Chi-Squared test.

To allow the use of a Chi-Squared test for comparison between an observed and normal distributions it is necessary for the number of groups employed in the frequency distribution to be relatively large, as many as 18 or 20 groups being preferable. Watson (1957) pointed out that X^2 values in this application tend to be larger than their correct value, although the difference is insignificant for 10 or more groups; Chayes (1954) suggested that each class should be no larger than half the standard deviation, and should preferably be less than a quarter. It is also essential that the 'expected' values calculated for the end classes in the distribution are not less than one. As this is normally the case this requirement may be circumvented by combining the extreme classes in the distribution with their adjoining classes until the requirement is observed. Of course the observed frequencies must also be combined for the same classes. Stages in the test, as illustrated in Table 11, are then:-

(1) For each one of the k classes calculate:-

$$f_i = \frac{(O - E)^2}{E}$$

where O = observed frequency

and E = expected frequency

(2) Calculate the final χ^2 value as:-

$$\chi^2 \quad \sum_{i=1}^{i=k} f_i$$

by summing over all of the k groups.

(3) This χ^2 value should then be compared with standard χ^2 tables, using k-3 degrees of freedom. Values of χ^2 greater than the tabulated value at the chosen significance level require rejection of the null hypothesis of no difference between observed and normal distributions.

In some cases a major non-normality in the observed data may be concealed by combining the end categories in order to allow use of the test. When visual inspection of the data suggests the tails of the distribution to be critical in judging normality, χ^2 can be approximated by the log likelihood ratio, χ^2_*, which does not require classes to be amalgamated. This is calculated from:-

$$\chi^2_* = 4.60517 \sum_{i=1}^{i=k} \left(O_i \log_{10}\left(\frac{O_i}{E_i}\right)\right)$$

where O and E are defined as before, the subscript i refers to each of the k classes, and the summation extends over the k classes comprising the frequency distribution. The degrees of freedom, d, for testing the significance of χ^2_* depend upon the nature of the observed distribution. When there are no observed classes within the distribution with no occurrences then d = k - 2. If a number of observed zero frequencies occur in the tail of the distribution and their total expected frequency is greater that, or equal to 0.7, then d = k - 1. If one or more classes have an observed zero frequency within the distribution, rather than in the tails, and if their total expected frequency is greater than or equal to 2, then d = k. χ^2_* is tested using χ^2 tables in the manner outlined above (for examples, see Table 12).

The Kolmogorov-Smirnov test.

This test procedure has no requirements concerning the number of classes employed, although a reasonable number is required to allow a meaningful test, and the greater the number employed, the more effective is the procedure. Stages in the procedure, as illustrated in Table 13, are:-

(1) Cumulate the data into a cumulative frequency distribution as if preparing a cumulative frequency curve or probability plot (Section III).

(2) Convert the cumulative frequency for each class in the observed distribution into a cumulative proportion by dividing each by n, the sample size. Do the same for the expected distribution, i.e. use the information given by step 2 of fitting the normal distribution.

Table 12. Chi-squared test using log-likelihood ratio.

North Tamerton

Class (mm)	Observed frequency (O)	Expected frequency (E)	$f=(O \log_{10}(O/E))$
600- 675	1	1.64	- 0.2148
676- 750	5	3.26	0.9288
751- 825	5	5.33	- 0.1388
826- 900	8	6.25	0.8577
901- 975	3	5.67	- 0.8294
976-1050	7	3.90	1.7782
1051-1125	0	2.04	0.0
1126-1200	0	0.81	0.0
1201-1275	0	0.25	0.0
1276-1350	0	0.06	0.0
1351-1425	1	0.01	4.3817
1426-1500	0	0.001	0.0
			4.3817

$$= \sum (O \log_{10}(O/E))$$

$$\chi^2_* = 4.3817 \times 4.60517 = 20.178$$

The observed zero frequencies are in the tails of the distribution and their total expected frequency is greater than 0.7. Therefore K-1 degrees of freedom are appropriate, i.e. 11.

The tabulated value for 11 degrees of freedom at the 0.05 level is 19.675, and the null hypothesis of no difference between observed and expected distributions may therefore be rejected at this level.

Werrington Park

Similarly, for these data $\chi^2_* = 5.31$. Since this does not exceed the tabulated value the null hypothesis may not be rejected.

This analysis demonstrated the superiority of the log-likelihood method over the standard χ^2 procedure when the tails of the distribution are critical in judging normality, since the marked positive skewness of the North Tamerton data is camouflaged by the amalgamation of classes necessary to allow the standard procedure.

Table 13. Kolmogorov-Smirnov goodness-of-fit test.

North Tamerton data.

Class (mm)	Observed frequency	Cumulative observed frequency as number	Cumulative observed frequency as proportion of n	Expected probability (from Table 11)	Cumulative expected probability
600-675	1	1	.03	.054567	.054567
675-750	5	6	.20	.108673	.16324
750-825	5	11	.37	.17750	.34074
825-900	8	19	.63	.20842	.54916
900-975	3	22	.73	.18889	.73805
975-1050	7	29	.97	.13011	.86816
1050-1125	0	29	.97	.068112	.93627
1125-1200	0	29	.97	.027094	.963366
1200-1275	0	29	.97	.0081883	.9715541
1275-1350	0	29	.97	.0018799	.9734342
1350-1425	1	30	1.00	.003277	.9767112
1425-1500	0	30	1.00	.0000432	.9767544

Maximum difference between the expected (i.e. normal) and observed cumulative proportions is 0.1018 for class 975-1050mm.

From tables such as that in Siegel (1956) the critical level is 0.24 at 0.05 significance level. Since the calculated value does not exceed this the null hypothesis of no difference between the observed and normal distributions may not be rejected.

Werrington Park data.

Maximum class difference 0.051783. Again, therefore, the null hypothesis may not be rejected.

Table 14. Critical values of K_s in the Kolmogorov-Smirnov test, for sample size over 35.

Significance level	0.1	0.05	0.02	0.01
Critical value of K_s	1.22	1.36	1.51	1.63

(3) Calculate the difference between the observed cumulative proportion and the expected cumulative proportion as calculated for the normal distribution, for each class in the histogram.

(4) Find D, the largest of the differences calculated in step (3), ignoring the sign.

(5) The method for determining the significance of D depends upon n, the sample size. Where n is less than 36 reference is made to a table of critical values (e.g. Siegel, 1956, Table E). For n over 35 a statistic K_s is calculated as:-

$$K_s = D\sqrt{n}$$

This K_s statistic is then compared with a table of critical values (e.g. Table 14), and if larger than the tabulated K_s value the null hypothesis of no difference between the observed and normal distribution is rejected.

(vi) The effect of outliers and other disturbances to normal distribution of data

The graphical methods discussed in section III may sometimes indicate that a particular sample is substantially normal except for one or a few very extreme observations or *outliers* in the distribution. Such might, for example, be the case if individual tropical storm rainfall totals are examined, where the bulk of data represent normal cyclonic or convective storms but a few very high values are caused by the occasional passage of hurricanes over the station. These outliers therefore belong to a different population and, depending on the purpose of the investigation, it may be thought necessary to omit them from any further analysis. Two statistical procedures may be suggested to enable the decision to be made whether an extreme value represents a true outlier. Dixon's gap test is applicable to samples which have from 3 to 30 variates. These are ranked in order of size, with the suspected anomalous value, x_1, having the rank of 1. The difference between the value of this variate and its nearest neighbours is then judged in relation to the overall range of the distribution, according to one of the following formulae; chosen according to sample size:-

$$R_1 = \frac{x_2 - x_1}{x_n - x_1}$$ for n between 3 and 8 inclusive

$$R_2 = \frac{x_3 - x_1}{x_{n-1} - x_1}$$ for n between 8 and 13 inclusive

$$R_3 = \frac{x_3 \quad x_1}{x_{n-2} - x_1}$$ for n between 13 and 30 inclusive

For 8 and 13 variates either appropriate formula may be used. The significance of the appropriate value is then judged by reference to the table presented in Dixon (1951) and reproduced by Bliss (1967); this procedure is illustrated by Table 15.

Table 15. Calculation of statistics concerning outliers.

North Tamerton data.

Dixon's gap test. These have been ranked according to size in Table 5. The suspected anomalous variate is that for 1960, having the value of 1362 mm.

The test statistic used is:-.

$$R_3 = \frac{x_3 - x_1}{x_{n-2} - x_1} = \frac{1017 - 1362}{688 - 1362} = \underline{0.512}$$

The significance of this is judged by reference to the table in Dixon (1951) or Bliss (1967). As it exceeds the tabulated value of 0.457 at a significance level of 0.01 it is judged to represent a significant departure from normal, and an outlier therefore exists. If this outlier is rejected the statistic may then be recalculated for the remaining 29 variates, to establish whether a further outlier exists. In this case:-

$$R_3 = \frac{x_3 - x_1}{x_{n-2} - x_1} = \frac{1008 - 1034}{688 - 1034} = \underline{0.075}$$

This is less than the tabulated value of 0.381 at the 0.05 significance level and the null hypothesis may therefore not be rejected. Therefore only one variate has been identified as an outlier.

David et.al's statistic.

$$P_s = \frac{|x_n - x_1|}{\text{standard deviation}} = \frac{637 - 1362}{141.44} = \underline{5.13.}$$

The significance of this may be judged by reference to the table given by Pearson and Stephens (1964), in which the tabulated value at the 0.05 level is 4.89. Since the calculated value exceeds the tabulated value the null hypothesis of no difference from the normal distribution may be rejected at the stated significance level.

Werrington Park data. Corresponding values are:-

Dixon's gap test. $R_3 = \underline{0.31}$ therefore no outlier exists.

David et. al's statistic. Ps = 4.56 therefore not significantly different from normal.

For larger samples, David et al. (1954) suggest the use of the statistic P_s, as:-

$$P_s = \frac{|x_n - x_1|}{\text{Standard deviation}} = \frac{\text{range}}{\text{Standard deviation}}$$

using the ranked x notation as above. This is referred to a table such as

that given by Pearson and Stephens (1964) or Bliss (1967) (1967, Table A10), as illustrated by Table 15. Use of this latter procedure should be performed with caution as it may not only indicate departure from normality due to the presence of outliers but also due to the existence of kurtosis, as too small a ratio will be produced by a platykurtic distribution.

Other forms of disturbance may also affect the normality of particular samples. This may be due to the mixing of two populations, as, for example, in a markedly bimodal frequency distribution of total rainfall amount for single rainfall events which could result from small amounts of rainfall being produced by convectional mechanisms and larger amounts by cyclonic mechanisms. A compound distribution therefore results. *censoring* may also affect the nature of a frequency distribution. This occurs when a sample has been cut off at a given point, such as may occur when flood peaks above a given level are not available in a series of peak discharge values because of instrumental limitations or failures at high discharges.
A similar form of distributional disturbance may occur in census data when, in order to retain confidentiality, data are suppressed or are randomly varied for small areas of very low population density. If the number of values omitted from a distribution is not even known then the distribution is said to be *truncated* rather than censored. The procedure for analysis of such distributions is beyond the scope of this introductory monograph but the interested reader will find further detail in Bliss (1967, pp.152-83).

(vii) Choice of methods for comparing frequency distributions with the normal distribution

A wide range of methods for analysing and testing the nature of frequency distributions has been presented; these have been summarised in Table 6.

The graphical methods and the associated graphical indices of symmetry and kurtosis should be regarded only as quick indications of the nature of frequency distributions. More rigorous indicators of symmetry and kurtosis are afforded by the more laborious calculation of momental skewness and kurtosis. An exact statistical test of skewness is that developed from momental skewness in the form of Fisher's g1 statistic. A number of methods exists for testing kurtosis and choice between them is conditioned by the sample size. For large samples Fisher's g2 statistic, based on momental kurtosis, is most powerful, but David et al.'s ratio is easier to calculate. For small samples the Geary mean deviation ratio should be employed. No definite figure can be given for distinguishing large and small samples, but 40 provides a useful guideline.

If the distribution is to be tested against the normal by an overall test rather than by separate examination of skewness and kurtosis then Shapiro and Wilk's W test should be used for small samples and one of the non-parametric goodness-of-fit tests for large samples. If g_1 and g_2 statistics have been calculated and only one allows rejection of the null hypothesis of no difference, then a joint statistic may be calculated from these. This is however a joint skewness and kurtosis statistic rather than a true goodness-of-fit test.

If the distribution of concern seems to be non-normal only because of the existence of one or more outliers then a test may be used to assess whether such outliers do exist. Dixon's gap test may be used for samples of 30 or less variates, and David et al.'s test for larger samples. Use of the latter must be done with caution, however, as it compounds platy-kurtosis and the effect of outliers.

Table 16. Summary of statistical measures employed.

		North Tamerton	Werrington Park
Skewness	Pearson's skewness	0.19	-0.07
	Graphical skewness	-0.05	-0.09
	Momental skewness	1.11	-0.18
	Fisher's g_1	S	NS
Kurtosis	Graphical kurtosis	0.89	1.03
	Momental kurtosis	5.54	2.98
	Geary's ratio	0.755	0.777
	Fisher's g_2	S	NS
Overall	Shapiro and Wilk's W	S	NS
	χ^2 goodness of fit test	NS	NS
	K-S goodness of fit test	NS	NS
	χ^2 employing log-likelihood ratio	S	NS
Outliers	Dixon's gap test	S	NS
	David et.al's ratio	NS	NS

NS = Not significantly different from normal at at least 0.05 level.

S = Significantly different from normal at at least 0.05 level.

Note that the sign for graphical skewness is the reverse of that conventionally used (see text).

Comparison between the tests available may also be made in terms of their efficiency in rejecting the null hypothesis of no difference between the observed distribution and the normal when a real difference does exist (Table 16). Shapiro and Wilk (1965) compared a range of tests for detecting non-normality in a variety of different distributions. They concluded that the Kolmogorov-Smirnov test is generally inferior to the χ^2 test, although with a few exceptions; these exceptions they attributed to the latter's rather arbitrary group amalgamation procedure. Mitchell (1971) similarly points out that the Kolmogorov-Smirnov test facilitates testing when the χ^2 expected frequency requirements are not satisfied. David et.al's ratio is only effective for non-skewed distributions, and Geary (1935) has

suggested that the Geary mean deviation ratio is more effective in detecting long-tailedness. The moment-based measures are generally powerful, although it must be stressed that these, like the measures of Geary and David et al., apply only to two specific attributes of non-normality, namely asymmetry and kurtosis, rather than overall normality. The goodness-of-fit tests are more general, in that they are concerned with the overall nature of the distribution. The W test is probably the most effective procedure for small samples, but is somewhat laborious in calculation and its efficiency declines for larger samples, by comparison with the moment measures.

Use of the methods outlined in this chapter therefore offers an effective means by which frequency distributions may be described, examined, and tested against the normal distribution. The goodness-of-fit methods may also be extended to allow comparison between an observed distribution and *any* theoretical distribution, such as the Poisson, binomial or gamma distributions. In essence the procedure is the same as that described above, except that 'expected' frequencies are generated according to the distribution concerned rather than by consideration of the normal distribution. Details of this are beyond the scope of this monograph, but may be found in texts such as Keeping (1962, pp.28-94) and Croxton and Cowden (1948, pp.265-304).

The techniques described above therefore offer an indication as to the nature of a frequency distribution. The courses open to the worker when the distribution is non-normal or does not satisfy other distributional requirements are examined in Section V.

V STRATEGIES FOR USE WITH NON-NORMAL DATA

(i) Assumptions of the linear model

The preceding sections have examined and outlined methods by which the distributional characteristics of data may be assessed. Attention is now turned to the strategies which may be employed if the application of these methods suggests that the frequency distribution of concern is of a form that does not satisfy the required distribution for a particular use, such as the use of parametric statistical procedures. Following Mitchell (1974) and Hoyle (1973), it may be suggested that the statistical techniques associated with the linear model involve the following basic assumptions, in order of importance (Kurskal, 1968):-

Independence of observations. This implies that observations are made in a random manner so that each variate is independent of every other variate. This assumption is very often contravened in geographical work, and *autocorrelation* often exists. For example the value of a variable at a moment in time, such as a daily rainfall figure, is statistically dependent upon the value of the preceding day, because wet and dry periods tend to span more than one daily observation period. The variates are thus said to be autocorrelated in the time domain. Autocorrelation may also occur in the space domain where, for example, rainfall figures for stations in neighbouring locations will tend to be correlated with one another. Spatial autocorrelation in geographical research has been the subject of some recent work (e.g. Cliff and Ord, 1973; Hepple, 1974).

Additivity or linearity. This implies that the main effects examined in an analysis combine in a linear fashion without complicating interactions, to 'explain' the observations of a particular variable. For example Gregory and Gardiner (1975) attempted to examine the controls exerted upon drainage density by both rainfall and geology in Southwest England, but were unable to perform a full two-way analysis of both factors because rainfall is closely related to altitude which is in turn, related to rock type, the higher areas being underlain by more resistant rocks. Thus rock type and rainfall were found to be related to one another and their separate effects could not be readily identified.

Homogeneity of, or constant, variance. This implies that the variance of the observations is constant and is therefore independent of both the expected value of the observations and of the sample size. This assumption is usually made, as it greatly simplifies the estimation of the statistical parameters of concern, and is of particular importance in regression and in the Analysis of Variance when the test for statistical difference between group means assumes *homoscedasticity* or equality between the group variances.

Normality. This assumption is critically important in the testing of hypotheses as it allows the use of a large body of simple, standard testing procedures, for which distribution functions have been extensively tabulated.

(ii) Strategies if the assumptions of the linear model are not satisfied

The third and fourth of these four assumptions therefore concern distributional characteristics of the data. Mitchell (1974) has reviewed the strategies which may be used if the data do not satisfy all or some of the assumptions, and suggests that the options available include:-

Faith *in the robustness of parametric statistical tests.* A statistical test is said to be robust if violation of assumptions concerning the frequency distribution of the data does not affect the probability distribution of the test statistic. Considerable efforts have been made to examine the impact of violation of distributional assumptions on testing procedures, and much early work gave at least some support for the notion of some workers who prefer to use these powerful statistical tests despite some violation of the assumptions on which they rest. In 1953 Box introduced the term robustness to describe the ability of a statistical test to avoid a *Type 1 error* despite assumption violation, and in 1955 pointed out that many of the tests used to examine assumptions, such as that of homogeneity of variance, are themselves likely to lead to erroneous conclusions if little is known about the parent distribution. He argued strongly against the use of non-parametric tests, preferring to rely upon the robustness of parametric measures, particularly when group or sample sizes were equal.

However since about 1960 it has been possible to investigate the effect of non-normality by means of computer simulation techniques, with the result that less confidence is now placed in the robustness of parametric measures (Hogg, 1974). For example Kowalski (1972) concluded from an extensive simulation study that the product moment correlation coefficient is quite sensitive to the effects of non-normality, and recommended that its use be confined to nearly normal data. Norris and Hjelm (1961) used simulated samples of 10,000 variates to reach similar conclusions, and Hogg (1974)

has given several other examples of computerised simulation work casting doubt on the robustness of parametric tests.

The concept of robustness is still one which is relatively undefined; Huber (1972) and Bradley (1968) have emphasised that a continuum from robust to non-robust must exist, and the precise degree of robustness of a specific test in a particular application must depend not only upon the degree of assumption violation in the data but also upon the sample sizes, the nature of the violation, the sampling and test conditions, and the nature of the inferences which the researcher wishes to make. Robustness does not therefore offer a universal panacea and if its existence is to be relied upon in particular applications care should be taken to specify fully the conditions under which it is used (Mitchell, 1974).

Table 17. Some common non-parametric statistical tests and their relative power-efficiencies (after Siegel, 1956)

Parametric test.	Non-parametric equivalent	Relative power efficiency (%)
t test	Walsh test	> 87.5
	Median test	63 - 95
	Mann-Whitney U	95
	Kolmogorov-Smirnov	< 96
F test	Friedman two-way analysis of variance	100
	Kruskal Wallis one-way analysis of variance	95.5
Pearson r	Spearman rank correlation coefficient	91
	Kendall rank correlation coefficient	91

Power-efficiency is calculated as described by Siegel (1956, pp-20-1), using the assumption that all requirements of the appropriate parametric test are met.

The use of non-parametric statistical tests. Distribution-free or non-parametric tests offer a second alternative when the assumptions inherent in parametric statistical tests are violated. These tests are generally considered to be less powerful (Table 17), in that they are less likely to detect a significant difference when one exists, but in much geographical work this problem can be readily circumvented by taking a larger sample of individuals. Furthermore the oft-quoted differences in power-efficiency between the two types of test have in most cases been evaluated for situations where all the assumptions of the parametric test are fulfilled, but not for situations where some assumption violation occurs (Bradley, 1968), and Blalock (1960) has noted that *relative efficiency* of a test is a function of sample size, significance level and type of inferential hypothesis. Non-parametric tests also have a weakness that in their present state of development they cannot

accommodate complex research designs. They do however offer an alternative when parametric assumptions are violated, and fuller details of their use may be found in texts such as Siegel (1956) or Walsh (1962).

Transformation of the data. If data can be recast into a form satisfying the normal distribution then greater and more powerful resources exist for testing statistical hypotheses (Mueller, 1949).

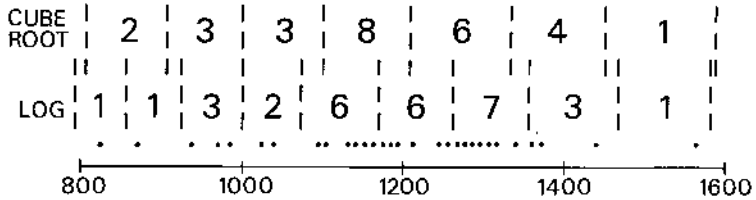

Figure 13. Grouping of variates into classes for transformed data. (Werrington Park data). This should be compared with Figure 6, for the untransformed data.

The principle underlying the transformation process is illustrated by Figure 13, which employs a similar approach to Figure 6, in that the Werrington Park rainfall data are plotted along the central line representing the variable scale. However unlike Figure 6 the histogram frequencies above the line are not derived for equal-sized class intervals but are for classes which have boundaries of unequal sizes. The size of these classes increases upwards, but if the logarithms (or cube roots) of the boundary values are examined (Table 18) then it will be seen that these do follow a regular sequence. In effect, therefore the variable scale is distorted by employing, say, logarithmic class boundaries. In practical terms this is more easily done by taking the logarithms of the data and plotting them on a linear scale divided into equal parts (Table 19; Figure 14), by which the same results are accomplished.

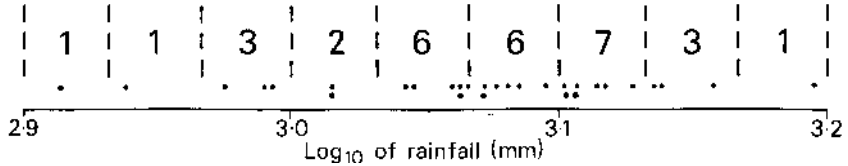

Figure 14. Logarithmically transformed variates plotted in the same manner as Figure 6 (Werrington Park data).

The pragmatic rationale for transformation may however lead to both interpretational and operational difficulties. For example transformation in a bivariate correlation by a logarithmic transformation of both sets of data implies the existence of an exponential relationship between the variables. Another problem is that if a transformation is carried out in order to correct one violated assumption then another may be violated to a greater extent than previously. For example in the Analysis of Variance a set of data which is homoscedastic between groups but which is skewed within groups may be made heteroscedastic by normalisation of individual group data. A final problem is that concerning interpretation of the resulting relationship between sets of transformed data, in that it may be very difficult to recognise causal mechanisms underlying relationships

Table 18. Class boundaries employed in Figures 13 and 14.

Logarithmic transformation		Cube root transformation	
Logarithm	*Equivalent original data value*	*Cube root*	*Equivalent original data value*
2.900	794.3	9.33	812
2.933	857.7	9.67	904
2.967	926.8	10.00	1000
3.000	1000.0	10.33	1102
2.033	1079.0	10.67	1215
3.067	1166.0	11.00	1331
3.100	1259.0	11.33	1454
3.133	1359.0	11.67	1589
3.167	1468.0		
3.200	1581.0		

between variables transformed in a complex manner. However, many transformations have been suggested and a brief review is now presented; a useful introductory survey is Gilbert (1973, pp.56-62) and fuller details may be found in Hoyle (1973).

(iii) Types of transformation

A great variety of types of transformation may be recognised; Hoyle (1973) examines nineteen transformations in general use, some of which are special cases of others. Transformations may be classified according to the assumption violation corrected, i.e. additivity, homogeneity or normality; alternatively they may be classified according to the mathematical operations involved in the transformation process. The latter approach is used here. Each of the major groups of transformation, according to these criteria, will now be briefly examined, with reference to particular transformations and to their effects on the violations of assumptions evident in the non-transformed data. The original variates are indicated by x, the transformed variates by y.

Power Transformations. These consist of raising all of the variates (x) to a given power, p

$$y = (x + w)^p$$

where w is a constant which is often zero, and p is a constant for a particular transformation. Particular values of p which have been commonly used are:-

Table 19. Transformed data and resulting moment measures for North Tamerton data.

Year	Untransformed	Transformation: $y=x^{0.5}$ (square root)	$y=x^{0.33}$ (cube-root)	$y=x^{0.67}$	log	loglog	chordal	optimal Dunlop & Duffy $y=x^{-0.83}$
1931	914	30.23	9.49	94.40	2.96	0.471	60.48	0.03487
1932	889	29.82	9.40	92.67	2.95	0.470	59.65	0.03568
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
1960	1362	36.91	10.82	123.17	3.13	0.496	73.82	0.02504
Skewness	1.11	0.77	0.66	0.88	0.45	0.37	0.77	0.00
Kurtosis	5.54	4.50	4.23	4.81	3.76	3.60	4.50	3.10

Power(p)	Transformation
- 1.0	Reciprocal
0.33	Cube root
0.5	Square root

The effect of these transformations is to reduce positive skew or increase negative skew. The reciprocal is the most powerful and the square root has the least effect. Negatively skewed data may also be transformed in the same way if they are first of all subtracted from a constant (h), i.e.

$$y = (h-x+w)^p$$

h must be larger than the largest observation. An alternative to this is to use a power (p) greater than one. Extremely positively skewed distributions may be made symmetric by the reciprocal transformation; this is most commonly employed when such a transformation possesses physical meaning and hence interpretability, as in Evans et al.'s (1975) use of it to transform the persons per household census variable to households per person. The cube-root transformation has been found useful for removal of moderate positive skewness and has been fairly widely applied in analysis of rainfall data (e.g. Stidd, 1953; Howell, 1965). The square root transformation removes relatively slight positive skewness and may be used where the data have a Poisson distribution, in which the variance increases linearly with the group means. This transformation has also found applications in rainfall studies. Related compound transformations include the chordal transformation (Freeman and Tukey, 1950):-

$$y = x^{0.5} + (x+1)^{0.5}$$

which is useful for distributions containing a mode at zero. Values of powers other than those given above may also be employed. For example Evans et al. (1975) tested the effect of eight transformations of this type with powers ranging from +2 to -3, and Anscombe (1953) found the transformation

$$y = x^{\frac{2}{3}}$$

to be better than $y = x^{0.5}$ for normalising variables having a Poisson distribution. If non-integer values of p are acceptable on interpretative grounds then an *iterative* procedure may be employed to test the effect of successively changing the power employed until a most nearly normal distribution is reached. This approach has been used by Box and Cox (1964) to develop a transformation

$$y = \frac{x^p - 1}{n}$$

in which a value of p may be found iteratively (Dunlop and Duffy, 1974) permitting transformation to a distribution with zero skewness (Hinkley, 1975, 1977). An an example of this procedure the values of p necessary to yield completely symmetric distributions for the North Tamerton and Werrington Park data are -0.83 and 1.43 respectively.

Tukey (1957; see also Mosteller and Tukey,(1977)) has also attempted to resolve power transformations into 'families' by presenting charts which portray the effect of transformations of the form:-

$$y = (x + c)^p$$

for all values of p less than 1.0, where c is a constant.

Logarithmic transformations. The logarithmic transformation (Aitchison and Brown, 1957) may be represented in the general case by:

$$y = \log (x + w)$$

where w is a constant which will often be zero. Many geographical datasets follow a log-normal distribution and their positive skewness can be removed by taking logarithms of the original variates. The effect of this is less severe than the reciprocal transformation, but more so than the cube-root transformation (Gardiner, 1973). The logarithmic transformation has been widely employed, particularly in Analysis of Variance when the variances of groups have been found to increase as the square of the group mean increases, and also in regression where logarithmic transformation of the original variables suggests a power function relationship between the original variables, of the form

$$y = wx^h,$$

w and h being constants.

Extreme values of skewness may be corrected by application of the logarithmic transformation to the logarithmically transformed data (Gardiner, 1973); this is termed the log log transformation and has been used in geographical studies (e.g. La Valle, 1967): it may be represented by:-

$$y = \log (\log x).$$

The logit transformation

$$y = \log \frac{x}{1 - x}$$

has also been used in geographical work (Wrigley, 1973). It requires the variates to be within the range from zero to one and transforms them to be between plus and minus infinity, in other words from a closed to a non-closed number system. Although some closure effect remains it is quite effective in reducing asymmetry; it is often employed for percentage data, suitably recast to lie between zero and one.

The two groups of transformation examined above are the most commonly used in geographical studies. However others are occasionally of use, and these will be briefly outlined. Some include trigonometric functions in their calculation; the most widely used is the angular or inverse sine transformation:-

$$y = \text{arc sin} \left(\frac{x + w}{n + h} \right)^{0.5}$$

where w and h are constants which may be zero, n is the total number and x is the number of 'successes' in a *binomial* variate. 'Arc sin' mean $\sin^{-1}$ i.e. reverse of the sine, *not* reciprocal of sine; the transformed data are therefore taken as being the angle in degrees which possesses the sine value calculated. The transformation is used to stabilise the variance of binomial variates, which have values from zero to one. A table of values of this transformation is given by Fisher and Yates (1948), and further

modifications are given by Hoyle (1973) and Keeping (1962), who also discuss the various values given to w and h in particular circumstances.

The angular transformation above should not be confused with the arc sin transformation

$$y = \text{arc sin} (x + w)$$

where w is a constant which may be zero; this transformation has also been used for binomial variates and for percentage data (Till, 1974) as an alternative to the logit transformation. It should be noted that considerable anarchy exists in the statistical literature concerning the terms angular, inverse sine and arc sin transformations, and they are often incorrectly used synonymously. The usage adopted here follows that of Hoyle (1973) and Kendall and Buckland (1971).

A further trigenometric transformation has been suggested by Cox (in Evans et al. 1975), namely:

$$y = \tan (\pi (x - 0.5))$$

This may be applied to variates which because of their definition, for example as shape ratio indices, are confined to the limited range 0 - 1.0. They cannot therefore be considered normally distributed without transformation, as the normal distribution has an infinitely small but finite frequency of occurrence at infinity on the variable scale. The transformation has no effect on the symmetry of the distribution, but gives it bounds of plus and minus infinity. Other transformation procedures less commonly used in geographical work are examined by Hoyle (1973).

The use of transformations therefore offers a third approach to the problem of assumption violation in the parametric statistical model; it cannot solve violation of the independence requirement but can allow the use of a powerful body of statistical theory in many cases, and a knowledge of transformational procedures is therefore essential in many geographical studies.

The use and calculation of the more important transformations outlined is illustrated by Table 19, which shows original data, transformed variates and resulting values of momental skewness and kurtosis for the rainfall data previously examined. The distributions of the transformed data are also illustrated by Figure 15 for the North Tamerton data.

VI CONCLUSIONS

From the outline presented in the preceding sections it is clear that there is a multiplicity of methods available for the analysis of frequency distributions and particularly for their comparison with the normal distribution. However Table 6 should suggest which technique is the most appropriate according to the criteria of the aims together with available computational power for a particular study, and the sample size available. It should also be borne in mind that, as Table 16 demonstrates, the available tests have differing degrees of effectiveness in rejecting the null hypothesis.

However it is worth re-emphasising that the normal distribution is only

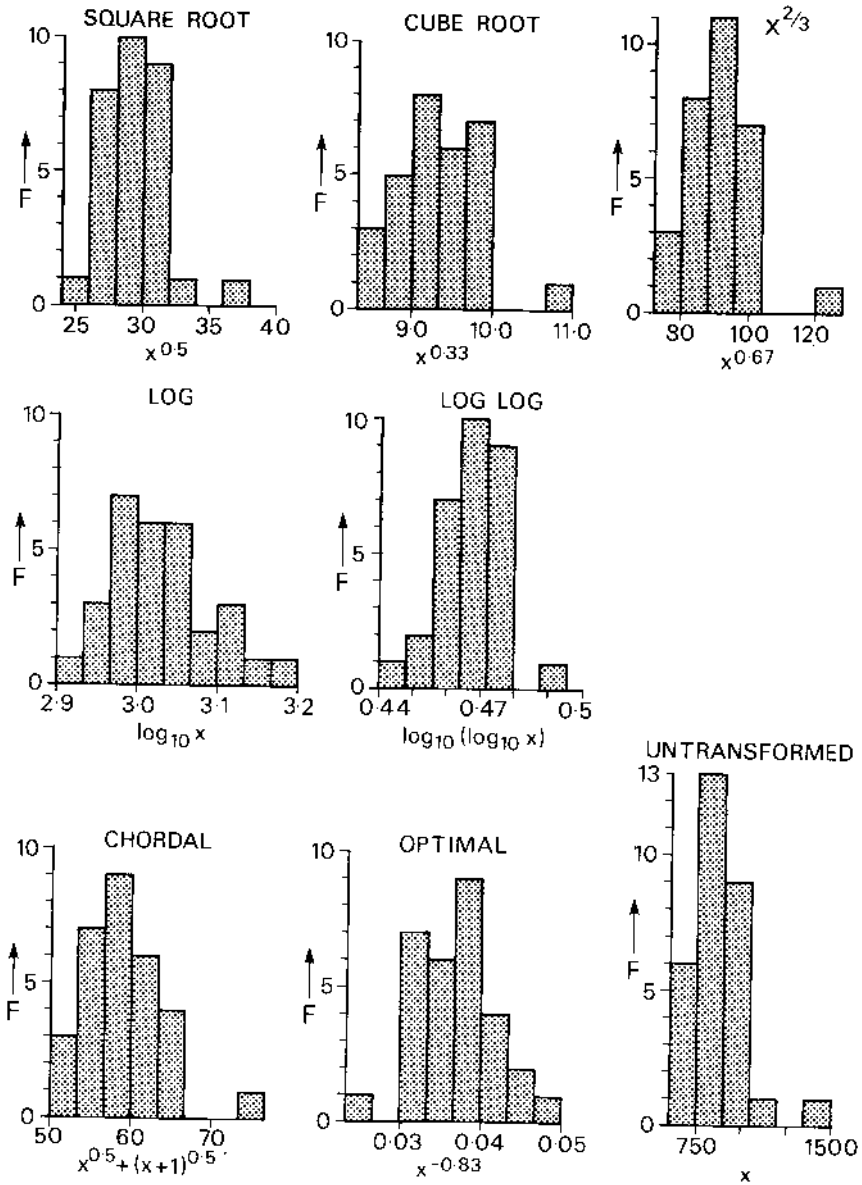

Figure 15. Histograms of transformed North Tamerton data employing some common transformations.

one of an infinite number of possible models, that there is nothing sacrosanct about normality, and that transformation represents only one possible solution to the problem of non-normality. Geographers are increasingly becoming familiar with different types of frequency distribution, and some guidance to the most accessible sources for an introduction to these is given in the bibliography.

A final point worthy of emphasis is that this monograph has been concerned solely with univariate distributional characteristics. However geographical problems are increasingly multivariate in nature, where one is concerned with relationships between two or more variables. Thus, if one is employing parametric techniques, one should consider whether data are multivariate-normal in their frequency distribution (Kowalski, 1970; Gnanedesikan, 1977, pp.168-195). This field has been little explored in geography, although a relatively accessible geological study, based upon work by Mardia (1970), is that of Reyment (1971).

Reasons and techniques for examination of the frequency distributions of geographical data have been explored in this introductory monograph; the necessary motivation should lie in the desire of the researcher, whether at undergraduate dissertation or more advanced level, to extract as much information as possible from his data without producing misleading conclusions by contravening the principles upon which the basic statistical tools are founded. The opportunity is generally present, with the almost universal availability of calculating or computing facilities, and it is to be hoped that future geographers will capitalise upon these opportunities.

GLOSSARY OF TERMS USED (italics in text)

Autocorrelation. The internal correlation between members of series of observations ordered in time or space. Spatial autocorrelation occurs because observations for locations in close proximity to one another will tend to be similar, and likewise temporal autocorrelation occurs because observations for moments in time close to one another will tend to be similar

Binomial variate. If a trial can have one of two mutually exclusive outcomes (e.g. yes/no, on/off, heads/tails) and if the probability of one outcome (p) is constant over a series of n independent trials then the probability, Pr, of obtaining r such outcomes is:-

$$Pr = \frac{n!}{r!\,(n-r)!}\, p^r\, q^{n-r}, \text{ where } q = 1 - p.$$

The distribution of r is termed the binomial distribution, and variables having this distribution are binomial variates.

Censoring. A censored distribution is one in which a known number of individuals have values larger or smaller than the extreme individuals for which data are available. (See also truncation)

Continuous variates are variates which can take any value within the overal range of each variable. They are usually derived by some form of measuremen or formed by making a ratio of other variates.

Cumulative frequency graph. A graph showing a frequency distribution in the form of the total frequency of occurrence for less than each class limit plotted against each class limit.

Discrete or discontinuous variates are variates which can have only a discontinuous set of values. Those most commonly met are variates derived b, counting.

Frequency distribution. A statement of how the frequencies of variates in a population are distributed according to the values of the variable with which it is concerned.

Gaussian. A Gaussian or normal distribution is a unimodal and symmetric frequency distribution having defined proportions of observations within specified values of the variable above and below the mean. (Figure 4). It is often described as 'bell-shaped', but is more rigorously defined by:-

$$f(x) = \frac{1}{s.d.\sqrt{2\pi}}\, e^{-(-0.5\ [(x - \bar{x})/s.d.]^2)}$$

which is fully explained in most standard statistical texts.

Heteroscedasticity. The property of unequal variances or standard deviations.

Histogram. A bar graph in which a bar is drawn for each class of a variable, with the height of each bar being proportional to the frequency of variates in each class.

Homoscedasticity. The property of equal variances or standard deviations.

Individual. An entity for which measurements are made.

Iterative. A computational procedure which relies upon the performance of the same sequence of arithmetic operations a number of times is said to be iterative. Such a process usually converges on the desired solution after a few iterations.

Kurtosis. An index describing the shape of a frequency distribution, being derived from the fourth moment of the distribution. It is usually interpreted as an indication of the peakedness of the distribution and the extent of its tails.

Leptokurtic. A distribution with a high value of kurtosis is said to be leptokurtic. Such distributions are more peaked than the normal distribution, or have longer tails (see Figure 12).

Mean. The arithmetic average of a set of variates; the usual indicator of central tendency.

Median. That value of a variable which divides the total frequency of variates into two halves when ranked in order. It may be derived as the value of the middle ranking observation when the sample size is odd, by averaging the two central-ranking observations when the sample size is even, or by reading off the fiftieth percentage point from a cumulative probability graph. The median offers an alternative to the mean as an indication of central tendency.

Modal class. See mode.

Mode. That class in a frequency distribution which has the highest frequency of occurrence is termed the modal class. For continuous distributions the class mid-value can be termed the mode or modal value of the distribution. Alternatively, for discrete data the mode is that single value having the highest frequency of occurrence.

Moment measures. Measures of the shape and location of a frequency distribution based upon the calculation of moments about the mean. (See mean, standard deviation, variance, skewness, kurtosis.)

Negatively skewed. Negatively skewed distributions are asymmetric distributions in which the 'tail' of the distribution is towards the lower end of the variable scale.

Non-parametric. Non-parametric statistical tests are those which do not incorporate assumptions concerning the frequency distributions of the data in their significance testing procedure.

Normal distribution. See Gaussian.

Normally distributed. See Gaussian.

Normal probability integral. A standardised function giving the area under the normal distribution curve for given class intervals from the mean. This area is proportional to the frequency of occurrence of variates in each class.

ogive. The curve showing the relationship of cumulated frequency against the variate value is commonly known as an ogive. However this term should strictly be confined to those cases where the curve is sigmoidal, and the term distribution curve is to be preferred in general applications.

Outliers. Variates which are beyond the limits of a frequency distribution, according to some particular definition, are termed outliers. Since many distributions, including the normal, have tails which extend to infinity, identification of outliers can usually only be performed according to some selected significance level.

Parametric. Parametric statistical tests are those which incorporate assumptions, usually of normality, concerning the frequency distributions of the data in their significance testing procedure.

Percentage point. A value of a variable below which a given percentage of observations lie is termed that particular percentage point of the distribution. This is most easily derived from the cumulative probability graph of the distribution.

Platykurtic. A distribution with a low value of kurtosis is said to be platykurtic. Such distributions are less peaked than the normal distribution, or have shorter tails (see Figure 12).

Population. The entire group of individuals of interest.

Positively skewed. Positively skewed distributions are asymmetric distributions in which the 'tail' of the distribution is towards the upper end of the variable scale.

Power of a statistical test is the probability that it rejects the alternative hypothesis when that alternative is false.

Probability paper. Graph paper with one axis so designed that a cumulative percentage frequency graph of any normal frequency distribution will plot as a straight line.

Probability plot. A cumulative frequency graph plotted with the frequencies on a probability scale, on which a normal distribution plots as a straight line.

Relative efficiency (or power-efficiency) of a test is the ratio of sample sizes concerned with two tests of a statistical hypothesis necessary to yield the same *power* against the same hypothesis.

Sample. A subset of the population.

Skewed. A distribution which is not symmetric about the mean is skewed.

Skewness. An index of the extent to which a distribution is not symmetric about the mean. Several indices exist but the usual one is that based upon the third moment of the distribution.

Standard deviation. An index of the variability of a frequency distribution. It is based upon the second moment of the frequency distribution and is equal to the square root of the variance.

Transformation. The process of data transformation is that by which some arithmetic operation is performed on all variates in a frequency distribution.

Truncated. A truncated distribution is one in which an unknown number of individuals have values of the variate of concern larger or smaller than the extreme individuals for which data are available (See also censoring)

Type I error. A Type I error in statistical testing procedure is to reject the null hypothesis (of no difference) when in fact it is true, and no genuine difference exists.

Variable. A quantity or index of which measurements are made.

Variance. In general terms variance is the degree of variability of a data set. In more precise terms the variance of a frequency distribution is equal to the standard deviation squared, and is similarly an index of the variability of the distribution, being calculated from its second moment. Analysis of Variance is a parametric statistical test which examines the relative extent of variation both within and between sets of variates.

variate. One observation of a variable, for one individual.

Symbols used

Alphabetic

a – coefficients in W test from table in Shapiro and Wilks (1965).

b – calculated value in W test.

C_j – class values in calculation of information loss for differently-sized classes of histograms.
C = product of variable class mid-point and frequency.

d – degrees of freedom in statistical testing.

$\bar{d}$ – mean deviation from mean.

D – largest difference between observed and expected cumulative frequency distributions in Kolmogorov-Smirnov goodness-of-fit test.

E – expected frequency in class in Chi-squared goodness-of-fit test

f – intermediate value of Chi-squared calculated for each class in Chi-squared goodness-of-fit test.

F – frequency in each class in figures.

g_1 and g_2 – statistics calculated in the significance testing of moment measures.

h – constant in transformation.

k_2, k_3, k_4 – statistics calculated in the significance testing of moment measures.

K – number of classes in a frequency grouping.

K_s – test statistic calculated in the Kolmogorov-Smirnov goodness-of-fit test when n > 35.

Kt – momental kurtosis.

$\log_{10}$ – logarithm to the base ten.

mm – millimetres.

M_1, M_2, M_3, M_4 – first to fourth moments of a frequency distribution.

n – number of variates.

O – observed frequencies in a distribution, in Chi-squared goodness-of-fit test.

p – power or exponent in transformation.

P – denotes a particular percentage point of a distribution.

P_s – test statistic calculated in David et al's statistical procedure.

R_1, R_2, R_3 – test statistics calculated in Dixon's gap test.

s.d. – standard deviation

sk – momental skewness.

ti – class boundary values in the Chi-squared goodness-of-fit test.

u^1, u^2, u^3, u^4 – the deviation of a variate from the mean, raised to the power indicated (see also [1).

V_1, V_2 - statistics calculated in the significance testing of moment measures.

w - constant in transformation.

$x_1, x_2 \ldots x_n$ - variates one to n forming the data of concern.

$\bar{x}$ - the mean of the data of concern.

X_1, X_2 - test statistics calculated in the significance testing of moment measures.

Zi - class limits standardised according to their deviation from the mean in relation to standard deviation. Also known as standardised normal deviates.

Other

$\sum_{i=1}^{i=n}$ - sigma. Indicates summation over the range given by the i values shown - in this case from one to n.

χ^2 - Chi-squared Test statistic with a well-documented distribution which can be used in a number of significance-tests.

χ^2_* - Chi-squared value derived by the log-likelihood method.

$|\ |$ - modulus, or absolute value of, irrespective of algebraic sign.

$\left[\ \right]_i$ - indicates summation over all of the variates of concern. e.g. $\left[u^2\right]$ indicates summation of squared deviations from mean, over all sample.

BIBLIOGRAPHY. (Most useful introductory material is asterisked

A. Texts and papers containing substantive discussion of relevant concepts.

Aitchison, J. and J.A.C. Brown (1957). *The Lognormal Distribution.* Cambridge.

Blalock, H.M. (1960). *Social Statistics.* New York.

*Bliss, C.I. (1967). *Statistics for Biologists,* Vol.1. New York.

Brooks, C.E.P. and N. Carruthers (1953). *Handbook of Statistical Methods in Meteorology.* London, H.M.S.O.

Cole, J.P. and C.A.M. King (1968). *Quantitative Geography.* London.

*Cormack, R.M. (1971). *The Statistical Argument.* Edinburgh.

Croxton, F.E. and D.J. Cowden (1948). *Applied General Statistics.* New York.

Davies, R.G. (1971). *Computer Programming in Quantitative Biology.* London.

Dixon, W.J. (1951). Ratios involving extreme values. *Annals of Mathematical Statistics,* 22, 68-78.

Doornkamp, J.C. and C.A.M. King (1971). *Numerical Analysis in Geomorphology.* London.

*Ebdon, D. (1977). *Statistics in Geography.* Oxford.

Evans, I.S. (1977). The selection of class intervals. *Transactions of the Institute of British Geographers, New Series,* 2, 98-124.

Evans, I.S. J. Catterall and D.W. Rhind (1975). *Specific transformations are necessary.* Census Research Unit, University of Durham, Working Paper 2.

Gates, C.G. and F.G. Ethridge (1972). A generalised set of discrete distributions with FORTRAN program. *Mathematical Geology,* 4, 1-24.

*Gilbert, N.E. (1973). *Biometrical Interpretation.* London.

Gnanedesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations.* New York.

*Gudgin, G. and J.B. Thornes (1975). Probability in geographical research: applications and problems. *The Statistician,* 23, 157-77.

Hammond, R. and P. McCullagh (1974). *Quantitative Techniques in Geography.* Oxford.

*Hoyle, M.H. (1973). Transformations - an introduction and a bibliography. *International Statistical Review,* 41, 203-223.

Keeping, E.S. (1962). *Introduction to Statistical Inference.* Princeton.

King, L.J. (1969). *Statistical Analysis in Geography.* Englewood Cliffs.

Krumbein, W.C. and F.A. Graybill (1965). *An Introduction to Statistical Models in Geology.* New York.

Lewis, P. (1977). *Maps and Statistics.* London.

*Mitchell, B. (1974). Three approaches to resolving problems arising from assumption violation during statistical analysis in geographical research. *Cahiers de Geographie de Quebec,* 18, 507-23.

Mosteller, F. and J.W. Tukey (1977). *Data Analysis and Regression.* Reading, Mass.

Mundry, E. (1972). On the resolution of mixed frequency distributions into normal components. *Mathematical Geology,* 4, 55-60.

*Norcliffe, G.B. (1977). *Inferential Statistics for Geographers.* London

Pearson, E.S. and M.A. Stephens (1964). The ratio of range to standard deviation in the same normal sample. *Biometrika,* 51, 484-7.

Reyment, R.A. (1971). Multivariate normality in morphometric analysis. *Journal of the International Association of Mathematical Geology,* 3, 357-68.

Shapiro, S.S. and M.B. Wilk (1965). An analysis of variance test for normality (complete samples). *Biometrika,* 52, 591-611.

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences.* New York.

*Smith, D.M. (1975). *Patterns in Human Geography.* Newton Abbot.

Taylor, P.J. (1977). *Quantitative methods in Geography: an Introduction to Spatial Analysis.* Boston.

Wilk, M.B. and R. Gnanadesikan (1968). Probability plotting methods for the analysis of data. *Biometrika,* 55, 1-17.

B. Examples of the use of frequency concepts in Geography.

Afolabi Ojo, G.J. (1973). Journey to agricultural work in Yorubaland. *Annals, Association of American Geographers,* 63, 85-96.

Boddy, M.J. (1976). The structure of mortgage finance: building societies and the British social formation. *Transactions of the Institute of British Geographers, New Series,* 1, 58-71.

Boots, B.N. (1977). Contact number properties in the study of cellular networks, *Geographical Analysis,* 9, 379-87.

Clark, D. (1973). Normality, transformation and the principal components solution: an empirical note. *Area,* 5, 110-3.

Desbarats, J.M. (1976). Semantic structure and perceived environment. *Geographical Analysis,* 8, 453-67.

de Smith, M.J. (1977). Distance distributions and trip behaviour in defined regions. *Geographical Analysis,* 9, 332-45.

Doornkamp, J.C. and C.A.M. King (1971). *Numerical Analysis in Geomorphology,* London.

Evans, I.S. (1972). General geomorphometry, derivatives of altitude and descriptive statistics. in *Spatial Analysis in Geomorphology,* ed R.J. Chorley, 17-90. London.

Evans, I.S., J. Catterall and D.W. Rhind (1975). *Specific transformations are necessary.* Census Research Unit, University of Durham, Working Paper 2.

Gardiner, V. (1971). A drainage density map of Dartmoor. *Transactions of the Devonshire Association,* 103, 167-80.

Gardiner, V. (1973). Univariate distributional characteristics of some morphometric variables. *Geografiska Annaler,* 54A, 147-53.

Gardiner, V. (1976). Land evaluation and the numerical delimitation of natural regions. *Geographia Polonica,* 34, 11-30.

Harvey, D.W. (1966). Geographical processes and point patterns: testing models of diffusion by quadrat sampling. *Transactions of the Institute of British Geographers,* 40, 81-95.

Harvey, D.W. (1968). Some methodological problems in the use of the Neyman Type A and negative binomial probability distributions in the analysis of spatial series. *Transactions of the Institute of British Geographers,* 44, 85-95.

Hill, A.R. (1973). The distribution of drumlins in County Down. Ireland. *Annals, Association of American Geographers,* 63, 226-40.

Kansky, K.J. (1963). *Structure of transportation networks: relationships between network geometry and regional characteristics.* Department of Geography, Res. Paper No. 84, University of Chicago.

Keyes, D.L., U. Basoglu, E.L. Kuhlmey and M.L. Rhyner (1976). Comparison of several sampling designs for geographical data, *Geographical Analysis,* 8, 295-303.

*Killen, J.E. (1973). Statistical techniques in geography: some applications and problems. *Geographical viewpoint,* 2, 383-96.

Krumbein, W.C. and W.R. James (1969). Frequency distributions of stream link lengths. *Journal of Geology,* 77, 544-65.

La Valle, P. (1967). Some aspects of linear karst depression development in in south central Kentucky. *Annals, Association of American Geographers,* 57, 49-71.

Morgan, R.P.C. (1970). Some examples of drainage component analysis. *Area,* 4, 37-41.

Morrill, R.L. and J. Symons (1977). Efficiency and equity aspects of optimum location. *Geographical Analysis,* 9, 215-25.

Ongley, E.D. (1970). Drainage basin axial and shape parameters from moment measures. *Canadian Geographer,* 14, 38-44.

Pringle, D. (1976). Normality, transformations, and grid square data. *Area,* 8, 42-5.

Rowley, G. (1975). The redistribution of parliamentary seats in the United Kingdom: themes and opinions. *Area,* 7, 16-21.

Stidd, C.K. (1953). Cube-root normal precipitation distributions. *Transactions, American Geophysical Union,* 34, 31-5.

Tanner, W.F. (1959). Examples of departure from the Gaussian in geomorphic analysis. *American Journal of Science,* 257, 458-60.

Tanner, W.F. (1960). Numerical comparison of geomorphic samples. *Science,* 131, 1525-6.

Taylor, P.J. (1971). Distances within shapes: an introduction to a family of finite frequency distributions. *Geografisk Annaler,* 53B, 40-53.

Till, R. (1974). *Statistical methods for the earth scientist: an introduction.* London.

Walling, D.E. and B.W. Webb. (1975). Spatial variation of river water quality: a survey of the River Exe. *Transactions of the Institute of British Geographers,* 65, 155-71.

C. Other works referred to in the text.

Andrews, D.F., P.J. Bickel, F.R. Hampel, P.J. Huber, W.H. Rogers and J.W. Tukey (1972). *Robust Estimates of Location.* Princeton.

Anscombe, F.J. (1953). Discussion following a paper by Hotelling. *Journal Royal Statistical Society, Series B,* 15, p. 229.

Bartlett, M.S. (1947). The use of transformations. *Biometrics,* 3, 39-52.

Box, G.E.P. (1953). Non-normality and tests on variances. *Biometrika,* 40, 318-35.

Box, G.E.P. and S.L. Andersen (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society, Series B,* 17, 1-34.

Box, G.E.P. and D.R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B,* 26, 211-52.

Bradley, J.V. (1968). *Distribution-free statistical tests.* Englewood Cliffs.

Chayes, F. (1954). The log-normal distribution of the elements; a discussion. *Geochim. Cosmochim. Acta,* 6, 119-20.

Cliff, A.D. and K. Ord (1973). *Spatial Autocorrelation.* London.

Darlington, R.B. (1970). Is kurtosis really 'peakedness'? *American Statistician,* 24, 19-22.

David, N.A., H.O. Hartley and E.S. Pearson. (1954). The distribution of the ratio, in a single normal sample, of range to standard deviation. *Biometrika,* 41, 482-93.

Dunlop, W.P. and J.A. Duffy (1974). A computer program for determining optimal data transformations minimising skew. *Behavioural Research Methods and Instruments,* 6, 46-8.

Finucan, H.M. (1964). A note on kurtosis. *Journal of the Royal Statistical Society,* B26, 111-2.

Fisher, R.A. and F. Yates (1948). *Statistical Tables,* Edinburgh.

Folk, R.L. and W.C. Ward (1957). Brazos River Bar: a study in the significance of grain size parameters. *Journal of Sedimentary Petrology,* 27, 3-26.

Freeman, M.F. and J.W. Tukey (1950). Transformations related to the angular and the square root. *Annals of Mathematical Statistics,* 21, 607-11.

Geary, R.C. (1935). The ratio of the mean deviation to the standard deviation as a test of normality. *Biometrika,* 27, 310-55.

Geary, R.C. (1936). Moments of the ratio of the mean deviation to the standard deviation for normal samples. *Biometrika,* 28, 295-307.

Gregory, K.J. and V. Gardiner (1975). Drainage density and climate. *Zeitschrift fur Geomorphologie,* 19, 287-298.

Hepple, L.W. (1974). The impact of stochastic process theory upon spatial analysis in human geography. *Progress in Geography,* 6, 91-142.

Hinkley, D. (1975). On power transformation to symmetry. *Biometrika,* 62, 101-11.

Hinkley, D. (1977). On quick choice of power transformation. *Applied Statistics,* 26, 67-9.

Hogg, R.V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory. *Journal of the American Statistical Association,* 69, 909-27.

Howell, W.E. (1965). Cloud seeding against the 1964 drought in the North-East. *Journal of Applied Meteorology,* 4, 553-9.

Huber, P.J. (1972). Robust statistics: a review. *Annals of Mathematical Statistics,* 43, 1041-67.

Huntsberger, D.V. (1961). *Elements of Statistical Inference.* Boston.

Kendall, M.G. and W.R. Buckland (1971). *A Dictionary of Statistical Terms. (3rd. Edition).* Edinburgh.

King, C.A.M. (1966). *Techniques in Geomorphology.* London.

Kowalski, C.J. (1970). The performance of some rough tests for bivariate normality before and after coordinate transformation to normality. *Technometrics,* 12, 517-44.

Kowalski, C.J. (1972). On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. *Applied Statistics,* 21, 1-12.

Kruskal, J.B. (1968). Statistical analysis: transformation of data. in *International Encyclopaedia of the Social Sciences.* Chicago.

Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika,* 57, 519-30.

McCammon, R.B. (1962). Efficiencies of percentile measures for describing the mean size and sorting of sedimentary particles. *Journal of Geology,* 70, 453-65.

Michell, R.L. (1975). What do successive frequency distributions show? *Agricultural Economics Research,* 27, 101-4.

Miller, H.P. (1955). Elements of symmetry in the skewed income curve. *Journal of the American Statistical Association,* 50, 55-71.

Mitchell, B. (1971). A comparison of Chi-Square and Kolmogorov-Smirnov tests. *Area,* 3, 237-41.

Mitchell, B. (1974). Three approaches to resolving problems arising from assumption violation during statistical analysis in geographical research. *Cahiers de Geographic de Quebec,* 18, 507-23.

Mueller, C.G. (1949). Numerical transformations in the analysis of experimental data. *Psychological Bulletin,* 46, 198-223.

Norris, R.C. and H.F. Hjelm (1961). Non-normality and product-moment correlation. *Journal of Experimental Education,* 29, 261-70.

Rogers, A. (1974). *Statistical analysis of spatial dispersion: the quadrat method.* London.

Smart, J.S. (1978). The analysis of drainage network composition. *Earth Surface Processes,* 3, 129-70.

Smith, E.J., Adderley, E.E. and F.D. Bethwaite (1965). A cloud seeding experiment in New England, Australia. *Journal of Applied Meteorology,* 4, 433-41.

Tukey, J.W. (1957). On the comparative anatomy of transformations. *Annals of Mathematical Statistics*, 28, 602-32.

Walsh, J.E. (1962). *Handbook of nonparametric statistics*. 3 vols. Princeton.

Watson, G.S. (1957). The X^2 goodness-of-fit test for normal distributions. *Biometrika*, 44, 336-48.

Wrigley, N. (1973). *An introduction to the use of logit models in geography*. CATMOG No. 10. Geo Abstracts, Norwich.