Report WN/90/16

The statistical analysis and summarisation of geotechnical databases

J.R.Hallam

October 1990

Engineering Geology Group,
British Geological Survey,
Keyworth,
Nottingham,
NG12 5GG

CONTENTS

1. INTRODUCTION

Geotechnical databases are compilations of mainly numerical results of tests carried out in the field or laboratory on small volumes or quantities of ground materials. These values are regarded as 'samples' from the relatively infinite volumes of actual ground present.

Most of the databases compiled to date by the Engineering Geology Group have been as components of the applied geology mapping projects, in which they are used to provide broad geotechnical assessments of the principal geological formations encountered. It is a prerequisite of these databases that each numerical item of data is allocated to an engineering geological 'unit' that can be taken to possess some degree of coherence in its origin, occurrence, properties etc. and which can be meaningfully distinguished from other such units. The source for such databases, of necessity, has been limited to the existing available site investigation reports. To a much lesser extent, relatively small databases have been used in in-house research projects, where the data acquisition can usually be planned and closely controlled. This report is concerned primarily with the former type of database, although many of the aspects discussed will be relevant to the latter.

On completion of a database all the available data values for the various geotechnical properties are abstracted for each engineering geological 'unit'. It is the analysis of these individual 'batches' of data which forms the subject of this report. The approach taken here attempts to follow that strongly advocated in a recent major work on data analysis (Hoaglin et al. 1983):

> "Look at the data and think what you are doing."

To this could well be added:

> "Consider where the data came from, what conclusions can realistically be expected and how these can be presented most effectively."

2. NATURE OF THE SOURCE DATA

2.1 Data Accuracy

Each data 'batch' to be analysed consists of all the individual recorded data values of a given geotechnical parameter for a given engineering geological unit. The number of values, if any, in a batch may range from one to several thousands, but will typically be in the tens and less often the hundreds. Each value is taken to be that of the given parameter at a specified point in space.

The great range and variability of the values within many batches can be attributed to the combined effect of several factors:

a) Inherent soil variability. The in-situ composition and state of soils is the net result of a great number of processes, including the supply of source material, the environment of deposition, consolidation, lithifaction, stress regimes, weathering processes and many others, which will all, to a greater or lesser extent, vary in time and space. As a consequence, soils will vary in composition, structure and fabric, and thereby in geotechnical properties, on all scales from microscopic to regional, even within a given unit.

Many writers, and particularly those with an engineering background, treat this variability simplistically in a statistical sense, as purely random.

b) Soil sampling and handling. The procedures for sampling, packaging and transporting soil samples from the in-situ location to the laboratory will inevitably induce changes in the soil. Whilst some of the changes could be purely random in nature, others clearly will not be. For example, stress relief will be greater for those samples from greater depths. As most soils have a high degree of saturation, changes in moisture content will tend to be those of reduction. In some soil types sample size may contribute to variability.

c) Soil testing. Testing procedures, both in the field and laboratory, are subject to many human influences which can introduce both systematic and random errors. For many simple tests, repeatability is usually only moderate. Where the testing is destructive, it becomes even more difficult to establish accuracy or detect errors.

d) Data transfer. The numeric data derived from a test will have to be transcribed several times before it is in place in the data bank. At each stage errors may be introduced.

e) Unit definition and data allocation. For data banking and analysis, the test results must be allocated to an engineering geological unit. This is unlikely to be achieved without some degree of error. To

minimise this, data should only be entered when there is reasonable confidence with regard to its allocation to a unit.

Whilst the variability which one wishes to analyse is solely that described in (a) above (the 'true' or inherent soil variability), one cannot escape the probability that it will be contaminated by the other factors discussed above. It is difficult to imagine many other types of data for which the statisticians term 'dirty' could be more appropriate.

2.2 Data Validation

In its broadest sense the concept of 'data validation' is a misnomer in a geotechnical context. No procedure can confirm any data value as totally valid, i.e. the 'real' value of the parameter at the given field location. The possible sources of error, as discussed above, are simply too many and too complex.

In the narrower sense that is conventional in computing, the expression is taken to mean that the data is keyed in a second time and preferably by a different person. The two data sets are then compared automatically and any discrepancies are rejected. In the present context this procedure is of limited value. It can only address one of the many sources of error.

The statistical approach which is advocated in this report recognises the potentially contaminated nature of the source data. As discussed later, errors are essentially accommodated, although gross errors can be highlighted and rejected by examination of the data distribution (see section 6.3.1).

2.3 Spatial Distribution of Data

The source of geotechnical data for the applied geology project databases to date has been limited to the existing, and available, site investigation reports from the areas concerned. Predominantly these fall into two categories. Firstly, investigations for some of the more significant structures within those parts of the project areas that have already been developed. Secondly, investigations of a few narrow corridors for possible major road construction. Almost by definition those areas that will be of most interest for future development are virtually devoid of any data.

Therefore the spatial distribution of the data is generally in the form of clusters, concentrated in and around urban centres, and occasional lines or ribbons. The only available means of achieving a more even distribution is to delete the great majority of this concentrated data, and thereby virtually the whole database.

3 STATISTICAL APPROACH AND OBJECTIVES

3.1 Classical Statistics

Before addressing the subject of geotechnical data, it may be helpful to illustrate the use of classical statistics in a simpler context. Suppose that during topographic surveying an angle is to be measured by theodolite. The instrument is set up, as are two clear, precise targets. Each in turn is intersected on the crosswires and the bearing recorded, to the maximum precision available from the instrument. The difference between these is the required angle. The process is repeated to give perhaps 4, 8, 16 or even more values, on different faces and portions of the calibrated scale. The recorded angles should all be very similar but will not be identical. The slight differences are attributable to the finite accuracy of the instrument's construction and calibration and the ability and concentration of the operator. None of the individual values can be presumed to be correct. A true value does however exist, as the instrument and targets occupy three precise and actual points in space. The observing procedure is designed to eliminate systematic errors, whilst gross errors (mistakes) will be self evident. The remaining errors are taken to be random in origin and, according to classical statistics, should fall approximately on a 'normal' distribution. The mean of the recorded values is the best estimate (most probable value) of the true angle. However, of equal importance is the determination of the standard deviation (spread) of the values, as from this is calculated the standard error of the mean, i.e. the accuracy of the estimate. The statistical premise is that if an infinite number of observations were made, the standard error of the mean would reduce to zero and the mean would equal the true value.

This classical approach is applicable to data sets in a great many fields. However, it does depend on a number of assumptions, which can easily be forgotten. One is that all the values are 'good' and of equal validity. Another is that the variations between the values are truly random and can be described by the Gaussian frequency distribution. A third is that a single 'true' value exists to be predicted.

Classical statistics can certainly address more complex problems, where some of these assumptions may be modified, although usually at the expense of increasing mathematical complexity. Statistical probability, in which the parameters of a whole 'population' are predicted, to a specified degree of confidence, from a set of samples, invariably depends on factors such as the consistent quality of these samples and an even and unbiased distribution of the samples within the population.

3.2 Objectives

The objectives which can be set in the statistical analysis of data batches from geotechnical databases must take account of the source and nature of the data. The overall objective, of course, is to predict the properties of each engineering geological unit, or population in statistical terms, to a quantifiable degree of confidence. An analysis of the actual data values available should be the means of achieving this, and not an end in itself. Inherent in the data are several aspects which inhibit this broad objective:

a) Data distribution. Spatial distribution of the data values, as discussed in section 2.2, is most often far from ideal. To predict a parameter for a whole engineering geological unit would require a statistically valid distribution of the sample locations, which is rarely available. Therefore, each analysis or data summary which is produced for a unit must be assessed for its applicability throughout the unit. As this can only be subjective, and non-mathematical, it is concluded that attempts to analyse the existing data in great detail, or to a high precision, are quite pointless.

b) Validity of engineering geological units. The allocation of individual data values to these units is fundamental to the database and its analysis. Some of these units may indeed be well-defined, distinctive and spatially consistent geological formations. Others may be, of necessity, almost dustbins of variable materials (e.g. fill and some glacial deposits), some of which should ideally be classified as units in themselves. It cannot be assumed that an engineering geological unit necessarily possesses sufficient consistency to constitute a 'population' in the statistical sense. Hence the statistical approach should avoid any prior assumption that the data values will fall within a mathematically definable distribution, such as the Gaussian.

c) Data accuracy. As the data is likely to be highly variable in accuracy, and may possibly contain gross errors, the statistical method, as far as possible, should accommodate these defects.

Figure 1, from Hampel et al. (1986), epitomises the different approaches of 'robust' statistics, as will be advocated in this report, and conventional or classical statistics. With the reservations noted in (a) above, it is clear that even a 'robust' fit must be used with caution.

4 LITERATURE REVIEW

An extensive literature search has produced about 100 references which relate to the statistical analysis of geotechnical property data. Of these the great majority are concerned with probabilistic analyses for geotechnical prediction. Statistical probability invariably requires that the frequency distribution(s) of the input data be expressed as some mathematical function. In almost all cases, as in most classical statistics, the distributions are taken to be Gaussian, either explicitly or implicitly. The validity of this underlying assumption is only rarely considered.

A few authors do discuss geotechnical frequency distributions, the most notable being Biernatowski (1985), Corotis et al. (1975), Ejezie and Harrop-Williams (1984), Fredlund and Dahlman (1971), Lumb (1966, 1969), McGuffey et al. (1980), Rethati (1983, 1988) and Schultze (1971). The most often quoted of these is Lumb (1966), who concluded from a study of four 'typical' soil formations that the Gaussian or a closely related (e.g. log-normal) distribution could adequately describe natural soils. However, the same author, Lumb (1970), subsequently states that this premise is false and that the family of beta distributions, being more versatile, will usually provide better fits. This conclusion is also reached by some of the other authors quoted. A point often made is that a statistical analysis can only be carried out on statistically homogeneous materials, the test for which is implied to be the ability to fit an acceptable mathematical distribution. Hence the argument can be somewhat circular. The amount of data on which these papers are based is usually small and, one might suspect, rather selective.

Only one instance has been found of a well documented geotechnical database, that for the State of Indiana (USA). This is described in five papers, Goldberg et al. (1978, 1980), Lo and Lovell (1982), Lovell and Lo (1983) and Lo and McCabe (1984), with an emphasis on discussing the data banking and simple statistical analysis of actual geotechnical data. The basis used for data classification is essentially physiographic and pedologic rather than lithostratigraphic. The major point of interest is that in the first two papers of this series the data are summarised by conventional statistical parameters, whereas 'nonparametric' methods are substituted in the second two papers. The authors of these four papers are all geotechnical engineers. However, in the final paper the second author, McCabe, is a professor of statistics. The statistical conclusion to this paper is that:

> "nonparametric robust statistical methods are preferred to conventional statistical methods for soil data analysis".

During the 1970's Steve Henley was a member of the I.G.S. Computer Unit. His subsequent book (Henley 1981), although concerned mainly with the spatial variation of geological data, discusses, in a very readable and non-mathematical manner, the difficulties in applying conventional

statistical methods:

> "... all of the geological modelling and resource estimation studies I participated in had data that were non-ideal in one respect or another (or just plain 'dirty'): the standard ways of handling the data with simpler parametric methods gave reasonable results, but always there were nagging doubts and some lack of confidence because of the corners that had to be cut in generating a model" (p. vii).

> "... many nonparametric methods are of direct application to a wide range of problems in geology and the other natural sciences. Some have already been used, but few such methods have been readily accepted as the standard, preferred method, even though it has become increasingly obvious that the traditional methods, straitjacketed by normal distribution theory, and many similar assumptions, are not altogether appropriate to sciences in which simple linear interactions are exceedingly rare occurrences" (p. 64 - 65).

There appears to have been some confusion and overlap a few years ago in applying the terms 'nonparametric' and 'robust', which may have arisen because some common statistics, such as the median, are used in both. From the standard texts on nonparametric statistical methods (Conover [1980]), Gibbons [1985] and Hollander and Wolfe [1973]), it can be seen that most of the available methods are various types of tests. Whilst the underlying assumptions are very few and weak, they are nevertheless still very strict. Only specific aspects of a data set are considered, not its structure as a whole. As with conventional parametric statistics, no allowance is specifically made for variation in the quality of the raw data.

Robust statistics, together with the closely related exploratory data analysis, emerged during the 1970's and appears to be a very active area of current research. Its origins have been credited to J. W. Tukey and particularly his paper of 1962 on the future of data analysis, from which Henley takes the following quotation as the theme for his book:

> "Far better an approximate answer to the **right** question which is often vague, than an exact answer to the **wrong** question, which can always be made precise".

An excellent introduction to the subject is given by Hoaglin, Mosteller and Tukey (1983) who stress the concepts involved rather than the higher mathematics. More specialised methods of exploratory analysis are examined in their companion volume (Hoaglin et al. 1985).

Some of the simpler techniques of exploratory analysis, complete with the necessary Basic and Fortran programmes, are given by Velleman and Hoaglin (1981). Hampel et al. (1986) provide a rigorous and highly mathematical treatment of robust statistics.

5 DISTRIBUTION STATISTICS

5.1 Geotechnical Data Variability

Each batch consists of all the values recorded within the database for a given parameter from an engineering geological unit. For any batch, the values will be variable to a greater or lesser extent. Although some of this variability can doubtless be attributed to errors arising from the soil sampling, laboratory testing and other unwanted factors (section 2.1), it must be presumed that the larger proportion of it represents true in-situ variability. In contrast to the simple example of a theodolite angle (section 3.1), there is no single 'true' value underlying the observed variability. Geotechnical parameters are inevitably variable at different points in space, even if measured with complete accuracy. Although it should be obvious, it is worth remembering that the output sought, and not just the input data, is a numerical distribution.

5.2 Methods of Distribution Statistics

The numerical distribution of a parameter, based on a large number of samples, is often represented as a frequency distribution curve (Figure 2). The horizontal axis gives the measurement scale for the given parameter and the vertical axis gives the frequency of occurrence. Very typically the curve is bell-shaped, with the greatest concentration of data in the central area and decreasing amounts laterally to each tail. Such curves may be symmetrical or skewed to one side, peaked or flat, regular or irregular in general form.

There are two conventional means of describing or summarising a frequency distribution in numerical terms: parametrically and nonparametrically.

5.2.1 Parametric Statistics

The parametric approach is to measure several essentially separate attributes or parameters of the whole distribution. Firstly the centre or 'location' of the distribution is determined by the arithmetic mean. Secondly the spread or dispersion, by the standard deviation. Much less frequently measured are the skewness and the kurtosis, the latter being essentially a measure of data in the tails, not the peak. In each case all the data values contribute to the given parameter, according to the square of their distance from the mean for the standard deviation, to the cube for the skewness and the fourth power for the kurtosis. The calculations are made in accordance with the method of moments and therefore the parameters, from the mean to the kurtosis are sometimes referred to as the first, second, third and fourth moments of the distribution.

A great part of conventional or classical statistics is concerned with the situation where data variation can be regarded as purely random. Here there is a vast amount of empirical evidence to show that the distribution of the variations will generally follow the Gaussian distribution, which therefore is more commonly, if unfortunately, referred to as the 'normal' distribution. In fact, there is no fundemental mathematical basis for this distribution, nor is it borne out by very large data sets, which commonly have larger tails than predicted. However, it is a very useful working tool in many fields.

All Gaussian distributions have the same identical shape or proportions, the only variables being the distribution centre (mean) and spread (standard deviation). Given these two, the complete distribution curve is always defined precisely. Figure 3 illustrates four Gaussian distributions having the same mean but differing standard deviations.

Therefore, it is always tempting either to assume that the actual data have been derived from a Gaussian distribution or to carry out a simple test to show, given the size of the data set, that it could have been so derived. If one is prepared to make this assumption, or believe the test, then the benefits can be very great. To a quantifiable level of confidence the frequency distribution of the whole population, and not just the samples, can be predicted.

For a Gaussian distribution, the skewness and kurtosis, by definition, have values of 0 and 3 respectively. These figures are rarely, if ever, achieved for real data distributions. Certain levels of discrepancy are allowable on the premise that data sets of any finite size will never follow a Gaussian distribution exactly.

Where the data distribution can be taken to be Gaussian, the two parameters, the mean and standard deviation, will completely define the distribution and therefore provide the best summary. Where this assumption cannot be made, it would be very difficult, and probably impossible, to reconstruct the frequency distribution from its parameters.

If the data distribution cannot be assumed to be Gaussian, there are a wide variety of other mathematical distributions which could be substituted. The benefit in doing so is that the means of predicting the population distribution is retained. The penalty is that the defining parameters become more obscure and difficult to visualise. They also become more numerous and the distributions more difficult to compare.

5.2.2 Nonparametric Statistics

The alternative approach is to dispense with the concept of parameters (hence the term 'nonparametric') and to use order or rank statistics (the two terms are essentially synonymous). Here the data values must first be rearranged into ascending numerical order. The order statistics are then simply the numerical data values at given levels in this ascending

order. These levels may be referred to in several ways, but in each case as proportions of the complete order sequence, e.g.

i) 0.1 order statistic, i.e. the data value one-tenth up the data sequence
ii) 20th percentile, i.e. the value 20% up the sequence
iii) deciles, quartiles, etc.

The lower quartile could alternatively be described as the 0.25 order statistic or the 25th percentile. Likewise the upper quartile as the 0.75 order statistic or 75th percentile. The 0.5 order statistic is more commonly known as the Median or 50th percentile.

Problems obviously arise where there are two or more identical data values (ties). In such instances there are conventional rules to apply, as also where a required percentile, for example, falls between two data values.

Therefore, the function of an order statistic is simply to determine a defined point in the distribution sequence. The greater the number of statistics used, the more precisely can the data distribution be mapped out.

The advantage of nonparametric or order statistics is that they remain equally valid whatever the nature of the underlying data distribution. There are, for instance, always an equal number of data values above and below the median. The information provided by these statistics is always clear, unambiguous, but necessarily simple. In contrast each parametric statistic summarises a given aspect of the entire distribution in one value. The two simplest of these statistics will completely define a distribution, but only where stringent assumptions can be met.

5.3 Robust Statistics

Whilst both parametric and nonparametric statistics have their advantages, they also both depend on all the data values being 'good' and equally reliable.

Robust statistics essentially discards the totally rigid approaches of both parametric and nonparametric statistics. It provides a flexible approach or attitude to the data, rather than any specific set of mathematical rules. Amongst its aims, two are particularly relevant in the present context.

Firstly, it attempts to allow for the fact that most, if not all, real data sets do contain a proportion of poor or bad data values. This is particularly true of many modern computerised databases, where the sheer volume of data would make any rigorous validation procedure uneconomic.

Secondly, it recognises that virtually all data sets do have some underlying structure. It can be regarded as intermediate between parametric and nonparametric statistics. In its simpler aspects it is closer to the latter, but with increasing sophistication it approaches the former.

It is often allied with exploratory data analysis, usually on the premise that the data should be examined before the most appropriate statistics can be selected.

6 EXPLORATORY DATA ANALYSIS

The objective in an exploratory analysis of the data, in this case a distribution, is to reveal both the general and detailed structure of the data. To achieve this effectively a graphical approach is almost essential. For a distribution the graphical presentation should reveal such features as the shape, spread and symmetry of the data and the presence of gaps or concentrations in the data.

One of the simplest, and perhaps the best-known, representations of a data distribution is the histogram. In reality good histograms can be difficult to produce. Their success increases with the size of the data set, but even with large sets they can be very susceptible to the choice of class intervals. Where the interval is too coarse the result is uninformative (e.g. Figure 4, where some 240 data values are allocated to just 3 classes). By reducing the interval the shape of the distribution can be more fully displayed (Figure 5). However, if the interval is too fine, the histogram is likely to have a 'noisy' appearance (Figure 6). In the extreme, where the interval is finer than the data precision, regular gaps will be artificially introduced in an essentially continuous distribution (Figure 7). Where data are recorded only to a low level of precision, care is needed to avoid an unequal loading of the class intervals (Figure 8).

As the size of the data set is reduced, the histogram will become unstable and sensitive to class limits as well as intervals. With discrete distributions, histograms can be particularly misleading where many, or even all, of the values fall exactly on the class limits.

Despite these difficulties histograms can offer one of the best means of portraying a frequency distribution. However, they do need to be formulated individually, and are arguably best kept to illustrate the occasional distribution which cannot easily be summarised in another form.

For an initial exploration of the data an alternative method of presentation, called the stem-and-leaf display, has been devised (Hoaglin et al. 1983) which reduces many of these difficulties and provides features not available in histograms.

6.1 Stem-and-Leaf Displays

The stem-and-leaf display is a very simple and easily understood technique where the most significant digits of the data are themselves used to sort the data into numerical order and displayed in a form which is very similar to a histogram. Before commencing, it is helpful to establish the number of values present in the data batch, together with the maximum and minimum values (and thereby the range).

The individual data values are each split into two parts, at a consistent point with respect to the decimal point. This split will usually be located such that either one or two of the most significant (leading) digits are separated from the remainder to form the STEM. A separate line in the display is then allocated to each possible value of the stem between the minimum and maximum data values. The first trailing digit (the LEAF) of each data value is then entered on the line corresponding to its leading digits. When all the data values have been entered, it is then customary to sort the LEAF values on each STEM into ascending numerical order.

The simplest form of the stem and leaf display is shown in Figure 9(a). The stems are listed to the left of the bold line, with the leaf values on the appropriate lines to the right. In this case the stem values (to the first decimal place) can be reunited with their leaves (e.g. 11/3 = 1.13) to give values in units of 0.01. The size of these units is stated at the head of the display as a reminder. It is an essential feature of this technique that the second and further trailing digits are truncated. Data values are never rounded off.

To the left, the standard display gives 'depths' for each stem. Starting from both ends of the display, these depths give the cumulative total of the number of data values for each stem line. For the 'middle' line, where there would be an overlap of the depths calculated from each end, the number of leaves on this stem is shown in parentheses. This feature is not required where the total number of data values is even and the median falls between two stems.

With a display constructed in the above manner, it would often be the case that a large number of leaf values fell on a relatively small number of stems. Alternatively, if the stem/leaf split is made after the next digit, there would be too many stems, with too few leaf values on each. The analogy would be histograms with either too few or too many classes. In these situations two lines can be used for each stem, with leaf values of 0-4 allocated to the first lines and values of 5-9 to the second lines. The convention with such a presentation is to denote each first line with an asterisk and each second line with a dot or small circle. An example is shown in Figure 9(b).

Where neither of these display formats gives the desired result, a third format is available, namely to use five lines per stem, as shown in Figure 9(c). Here the convention is to denote the intermediate lines with a 't' (for Two's and Three's), an 'f' (for Four's and Five's), or an 's' (for Six's and Seven's).

Although the stem-and-leaf technique was devised for the manual processing and display of data, it can be particularly successful with a good statistical graphics software package. An obvious advantage with a printer output is that a consistent width is used for each digit in a leaf. When viewed side-on, the effect is essentially that of a histogram, with

the length (height) of each stem line proportional to the number of leaf values. The software will invariably incorporate a number of desirable refinements.

One such refinement is to incorporate a degree of resistance to markedly anomalous or outlying data values. The whole data set is initially analysed to see whether any such values exist, and if so these are separated and listed at the relevant end of the display. Examples of this are illustrated in Figures 9(a) and 9(b), where 'LO' and 'HI' outliers identified respectively.

Another useful refinement is the automatic selection of the stems and the number of lines per stem. In addition to the outlier resistance just discussed, this algorithm takes account of the total number of data values. As this number increases, so, also, will the number of stem lines, and thereby the detail of the display.

Apart from an aesthetic roughness, which is of little relevance at the exploratory stage, and could be largely overcome with familiarisation, the stem-and-leaf display provides all the usual information which can be deduced from a histogram, e.g.:

a) the symmetry of the data
b) the spread of the data
c) the isolation of a few values from the main body of the data
d) local concentrations within the data
e) gaps in the data

Two particular advantages of the stem-and-leaf display are usually cited. Any patterns or peculiarities in the digits in any line can be seen, e.g. if '0's predominated it might infer that part of the raw data had been rounded off more than the rest. As the display is composed of actual data values it becomes much easier to trace particular values of interest back to the individual raw data.

For use with the geotechnical databases there are further advantages over histograms:

a) There are no problems or doubts with values at or close to class limits. It is quite clear into which class or line any value will fall.

b) The automatic selection of the stems and leaves ensures that an at least tolerable display or "histogram" will be produced. The technique is resistant to the major pitfalls which can arise in producing conventional histograms.

6.2 Histograms

The major difficulties in producing histograms, e.g. the class intervals, class limits and the influence of discrete data, have been mentioned earlier. To formulate a good display, it is essential to have:

a) the total number of data values
b) an outline of the distribution in the tails.

The latter is required in order to select the upper and lower limits for the histogram. These can be set slightly outside the extreme data values. However, if these are isolated outliers, one might wish to exclude them from the display. A more reliable method is to determine realistic limits from an examination of a preliminary stem and leaf display.

The number of classes or intervals in the histogram should be determined from the total number of data values. Several simple rules have been proposed for this, which are given, with their results, in Table 1 (after Hoaglin et al. 1983), where 'n' denotes the total number of data values.

TABLE 1 Number of lines for a stem-and-leaf display or number of intervals for a histogram, as suggested by three rules.

	Rule (Integer Part of)		
n	$10 \log_{10} n$	$2\sqrt{n}$	$1 + \log_2 n$
10	10.0	6.3	4.3
20	13.0	8.9	5.3
30	14.7	10.9	5.9
40	16.0	12.6	6.3
50	16.9	14.1	6.6
75	18.7	17.3	7.2
100	20.0	20.0	7.6
150	21.7	24.4	8.2
200	23.0	28.2	8.6
300	24.7	34.6	9.2
16	12.0	8.0	5
32	15.1	11.3	6
64	18.1	16.0	7
128	21.1	22.6	8
256	24.1	32.0	9
512	27.1	45.3	10

It is suggested that the lower of the results given by the $10 \log_{10} n$ and $2\sqrt{n}$ rules would probably be superior to the exclusive use of either.

With approximate figures for the histogram limits and the number of classes, it is then necessary (or at least highly desirable) to refine these so that the class interval is equal to a discrete rounded number of data units. The class limits will then fall at discrete rounded values.

Care is required where the data is coarsely discrete, as, for example, where Atterberg limits have been rounded to whole percentages. The class interval must then be made equal to, or an integral multiple of, the discrete data precision.

A further difficulty can arise with some graphical software, where the labelling is applied at the class intervals. With plastic limits (rounded to whole percentages) a class might have limits of, and be labelled as, 18 to 20. This interval might in fact record the number of values which are greater than 18 and less than or equal to 20, i.e. values of 19 and 20. The labelled display will in this case have a bias or shift of 0.5 units. To overcome this problem the limits could be set at 18.5 and 20.5, so that it becomes quite clear into which interval the values of 20 will fall. Ideally the labelling would be applied at rounded values, e.g. 20, and not at the class limits.

6.3 Probability Plots

Whilst histograms and stem-and-leaf plots can demonstrate the general structure of a data distribution and reveal at least some of the anomalies within it, these displays suffer from one major drawback as means of data exploration. In the great majority of practical cases their visual impact is little more than a statement of the obvious. The data are predominantly concentrated near the centre of the distribution, with decreasing proportions of the data to either side, i.e. towards the tails. This general 'bell' shape, albeit somewhat distorted and coarsely stepped, is the dominant feature apparent to the viewer. Subtler aspects of the distribution will tend to be masked rather than revealed.

This bell shape is the major trend underlying many actual data distributions. By removing it, or at least the greater part of it, other aspects of a distribution will become more readily apparent. This can usually be achieved by presenting the data as a 'normal' probability plot. The x-axis is scaled to the data units, whilst the y-axis has the cumulative percentages of the data plotted on the Gaussian probability scale. The result is that a Gaussian or 'normal' frequency distribution will be portrayed not as a bell-shape but as a straight line. Such a distribution can be totally summarised in two parameters, the mean and the standard deviation. On the probability plot the mean is the centre point of the line, i.e. the data value corresponding to 50% on the y-axis. The standard deviation is proportional to the slope of the line. Thus if various batches of the same geotechnical property had different distributions, which were all Gaussian, they would plot as straight but distinct lines.

Occasionally batches of geotechnical data do plot virtually as straight lines, as, for example, in Figure 10. However, most commonly the probability plots will depart in one or more respects from such a straight line. For instance, a somewhat irregular or 'noisy' plot may be

encountered, particularly where the data batch is of limited size (Figure 11). Nevertheless, a plot can almost invariably be expected to exhibit some simple continuous and consistent pattern or structure, provided the data batch has a coherent distribution.

Probability plots can be very valuable for exploratory analysis, in several respects:

6.3.1 Validation

As a result of the normal (Gaussian) probability scale used on the y-axis, these plots will inevitably show the data points as most concentrated at the centre and increasingly more separated towards the tails. With this greater separation the tails will usually appear more irregular. Nevertheless they should follow a pattern consistent with the main bulk of the plot.

In Figure 12 the lowest data value is clearly seen to be inconsistent with the remainder and therefore can be identified as an outlier which does not belong with this data batch. Several values may show a coherent pattern, but at variance with the majority of the data, as in Figure 13. Whilst a single outlier may be erroneous for several reasons, a grouping will probably be easier to explain. Usually the data values will have been mis-coded with respect to either the geological formation or the geotechnical property. Where time permits it may be worthwhile comparing the plots for several parameters. If the same sample appears as an outlier in more than one plot, it is probable that the data values are good but that the geological formation has been incorrectly identified. Conversely where a given sample is seen as an outlier on only one plot, that individual data value is likely to be suspect.

A reasonably consistent probability plot cannot guarantee, of course, that all the data values are in fact derived from one coherent distribution. Where two distributions have similar data values, and particularly where they are of similar size, they may be virtually impossible to distinguish within a single combined plot. A sharp change of gradient in a plot would probably indicate such a situation.

Where one or more values have been deleted from a batch as gross outliers, it is usually wise to replot the remainder, to ensure that they remain consistent and do not contain any lesser outliers.

6.3.2 Skewness and Transformation

After the removal of any evident outliers, the probability plot should exhibit some consistent shape or structure, albeit somewhat irregular at the tail ends. Where this shape is essentially a straight line, the data follow a Gaussian distribution. More usually the plot will be curved, and to a first-approximation follow one of the four patterns in Figure 14. In (a) the data is lighter tailed (deficient in tail values) than a Gaussian

distribution, whilst that in (b) is heavier tailed. The shape in (c) is left skewed, in that the higher data values have a steeper slope and hence less spread, whereas the opposite is true for the lower data values. In contrast the shape of (d) is right skewed.

Differences of kurtosis [(a) and (b) above] are of no particular importance here, unless there is a requirement to identify the underlying frequency distribution. The sinuous plots that it produces could only be straightened by rescaleing the y-axis to a non-Gaussian distribution.

Skewness, however, is more significant, as it is dependant on the scaling of the x-axis, i.e. the data scale. Differing skewness will result from changes in this scaling. Plots in which the measurement units of the recorded data values are scaled arithmetically are obviously convenient, both for plotting and subsequent reading. However quite often it may be argued that this convenience is outweighed by statistical and/or geotechnical considerations.

Some of the geotechnical parameters are not fundamental and independent, but derived from, or otherwise dependent upon, others. In some cases the parameters are simply the conventional, but not the only, manner in which to measure a property of the material. As an example, specific gravity, void ratio, degree of saturation, moisture content, dry and bulk densities are all interdependent. Liquid and plastic limits, liquidity and plasticity indices and moisture content comprise another such group. Some parameters could be expressed with equal logic by their inverses (strength/weakness, stiffness/compressibility, etc.). For the same batch of samples, the distribution of their properties may appear quite different, depending on the parameters selected and their graphical scaling. Their order may remain constant but their apparent concentration or spread can change dramatically.

An approximately Gaussian distribution is to be expected when both the data variability is random and the measurement scale is effectively unlimited. The specific gravity, at least of most rock-forming minerals, is a good example of the latter. It would be most unusual to find specific gravities which fell outside the range 2.5 - 2.8. In a distribution of this parameter the standard deviation would be very small in relation to the mean.

For SPT's the measurement scale is severely limited. Whilst there is no physical upper bound, 'N' values of zero or less are impossible (any recorded values of zero would have been rounded down from some fractional, although unmeasured, value). Standard deviations for this parameter are usually very large in relation to the mean. As a measure of penetration it would be equally logical to record the inverse, i.e. penetration/blows rather than blows/penetration. Materials which gave high and broadly spread values in one parameter would give low and concentrated values in the other.

It is the proportional change between data values that is usually significant, not the change in absolute measurement units. Thus cohesion values varying between 1 and 10 kPa denote a far greater variability than values varying between 501 and 510 kPa. As one of the primary objectives in analysing and then summarising the geotechnical data distributions is to examine the variability within and between batches, it is important that constant variability should be recognizable as such. This will generally be achieved when the data axis is scaled logarithmically.

In this respect the probability plot is very useful. It enables the skewness of a distribution to be assessed for the bulk of the data, whereas the standard calculation of skewness is based on all the data with particular emphasis on the more extreme values. If a plot displays skewness, then the gradient of the plot, i.e. the spread, will vary along its length. Figure 15 shows a batch of SPT results to be right-skewed, with the data scaled arithmetically. This batch, in fact, is the same as that shown in Figure 10, where the data axis is scaled logarithmically.

Skewness should be assessed not just for individual plots, but for all the plots of the same parameter from the different formations. The data scaling which then gives the least overall skewness should generally be adopted. As arithmetic scaling has the advantage of familiarity, it should be retained where logarithmic or other scaling is not clearly superior. Other scalings which could be considered include power transformations, such as the square and square root.

6.3.3 Discrete Data Precision

In the previous examples the probability plot lines have been essentially continuous. Figure 16 illustrates a case where the plot line is coarsely stepped. This situation will arise where the precision of the recorded data values is not much less than the spread of the values. Those parameters which are calculated as moisture contents (e.g. Atterberg limits) very often suffer from this problem as they are recorded to the nearest whole percentage. Such stepping is purely a function of the data recording. Were the moisture contents, for instance, to be recorded to a precision of 0.1%, the coarseness of the steps would be reduced by a factor of 10 and in most cases would then be hardly noticeable.

This stepping degrades the appearance of probability plots, and may occasionally make the assessment of skewness or potential outliers more difficult. However, its most serious effect is that data values read against percentages on the y-axis will be insufficiently accurate. For example, the same data value might be plotted for all percentages from 4% to 12%. If the 5th and 10th percentiles (i.e. the data values at 5% and 10%) are now required, they would appear to be identical, which in reality they are not. The result is that any further statistical operations which are dependent on such percentiles will suffer in their accuracy.

The only remedy that appears to be available is to construct probability plots based on frequency distribution 'bins'. For example if 17 plastic limits were recorded as having values of 12% moisture, these would now be taken as 17 values falling between 11.5% and 12.5% (which had been rounded to 12% when recorded). Thus every data value is allocated to a bin, the limits of which are 0.5 data units above and below the recorded value. The required frequency distribution can readily be produced by statistical software packages, as a listing of cumulative frequencies for each of the bin boundaries. From this listing the probability plot would have to be drawn manually on standard probability paper. The required percentiles would then have to be read off and entered into the database. This process could be time-consuming, but would greatly enhance the value of any statistics based on low precision data.

Some probability plots have an irregular or 'noisy' appearance, even when the precision of the recorded values is adequate. This can be particularly true where the size of the data batch is small. It is arguable that such plots should be manually smoothed and the required percentiles then entered into the database, as above.

6.3.4 Small Data Batches

Many of the data batches encountered in the geotechnical databases consist of only a few data values. Quite often only a single value is available. Probability plots can provide a suitable context in which to consider the minimum size of a data batch for which any statistical analysis or summary is worthwhile.

Each data value is regarded as a sample from the overall population of values which could be derived for the specified parameter of the geological formation in question. Given a large number of such samples, their values will form a distribution which should approximate on a probability plot to a straight or gently curved line. With a lesser number of samples, this line can be expected to have a locally more irregular or 'noisy' appearance, but its general form should still be apparent. For statistics to have any validity, there must be sufficient data values to broadly establish such a line.

One data value achieves nothing. A straight line can always be drawn exactly through two points, and a simple curve through three points. To define a distribution curve of non-standard kurtosis [e.g. Figure 14(a)] requires a minimum of four points. Five values is arguably the absolute minimum number to provide even one 'redundant' point to give a check on the shape of the line. Five is the smallest number used by Hoaglin et al. (1983) in their consideration of very small data batches.

Although five values might give a rough indication of the centre of a distribution, the apparent spread (the slope on the probability plot) must be treated with caution. The outer values will inevitably plot at only 10%

and 90%. At best only the spread of the central part of a distribution is even indicated. A data set of around 10 data values is probably the smallest to have any statistical worth. With 30-40 values the essence of a distribution should be clearly discernable.

There can be no firm rules for the size of the smallest batch that should be used. A subjective assessment of the regularity of the probability plot is a useful guide. An obvious problem with very small data batches is that their geographical distribution is likely to be very limited. Often they will have been derived from just one or two site investigation reports or even individual boreholes. As such they may be particularly poor indicators for a whole engineering geology unit.

7 NUMERICAL SUMMARY STATISTICS

The graphic displays that have been described earlier, and particularly the probability plot, provide detailed representations of individual frequency distributions. These distributions must then be summarised so that they may be compared and their most significant aspects recorded. Summaries may be presented in numerical or graphical form, of which the latter is usually to be preferred. Numerical summaries are more readily available, as for instance from statistical software packages. Their attributes and particularly their limitations are discussed in this section.

The common summary statistics can be divided into three groups:

1) Sample size (number of data values)
   Minimum
   Maximum
   Range
   Mode

2) Mean (arithmetic)
   Geometric mean
   Variance
   Standard deviation
   Skewness
   Kurtosis

3) Median
   Lower quartile
   Upper quartile
   Interquartile range

The first group do not summarise the data so much as give its limits. The minimum and maximum (and the resulting range) simply record the most extreme values which have been included (quite possibly erroneously) in the data set. These have very little use other than in setting limits for graphical displays. The mode records, at least in theory, the peak on the frequency distribution curve. In practice it can give an unreliable and misleading impression of the distribution and has poor repeatability between batches from the same data source.

The second group are the conventional or classical summary statistics. The mean and the variance (or its square root, the standard deviation) are by far the most widely quoted and used of summary statistics. Where the distribution is normal, i.e. Gaussian, they are excellent. In such cases they not only summarize the distribution, they completely define it and are therefore referred to as parameters of the distribution. The distribution can be reconstructed in its entirety from just these two numbers. They are appropriate in many fields of statistics, especially when there is a 'true' value of some quantity and statistics are applied to

repeated physical measurements. Examples of this would be a series of weighings of a single sample. In other instances, of which the geotechnical data being considered here are an excellent example, they can be inappropriate and misleading; there are no single 'true' values underlying each distribution; the distributions themselves may be far from Gaussian; and the data set may include values which are grossly erroneous or belong elsewhere.

The method of moments, by which these statistics are conventionally calculated, is increasingly significant for the higher orders of the moments involved, from the second (variance), to the third (skewness) and fourth (kurtosis). Each data value is included in the calculation, with equal weight, as the second (square), third (cube) or fourth power, respectively, of its distance from the centre (mean) of the distribution. To a successively greater extent these higher order statistics reflect the behaviour of the tails rather than the main body of the distribution. They can be totally distorted by a few extreme values.

The third group of summary statistics, unlike the second, are not derived from comprehensive calculations on the whole data batch. Although their broad objective is similar, their values are determined simply by taking specific points on a distribution. They are referred to as nonparametric statistics. The distribution can never be reconstructed from these, only specific points on it. But although the information they give may be much less complete than for their classical equivalents, they can offer very significant advantages.

The first and most important statistic that is sought for a distribution is the location of its 'centre', of which the mean and median are two possible measures. These are discussed in the next section, together with the advantages in seeking 'robust' measures which are in a sense intermediate between the two. In a fairly simple way this will illustrate how the robust approach attempts to reach the optimum between parametric and nonparametric statistics when applied to 'real' rather than ideal data. The second statistic that is usually required, a measure of the distribution spread, is examined in the following section.

7.1 Statistics for the Distribution Centre

Of the many possible statistics for the centre of a distribution, the most commonly used are the mean and the median. The (arithmetic) mean is simply the average of the values in the data batch, i.e. the sum of the data values divided by the number of values. It has the property of being the centre of gravity of the distribution and hence is its first moment. The median is the central data value in a distribution, in that the numbers of greater and lesser data values will be equal.

7.1.1 Resistance

A major drawback of the mean is its lack of resistance to gross errors. Resistance is quantified as the minimum 'breakdown value', which is the maximum proportion of the data values which can be taken from the least favourable locations in the data set and reassigned with extreme values, without rendering the statistic unstable.

For example, suppose there are 99 data values in a data set, with a median of 40 and extreme values of 15 and 60, and all the 49 values below the median are then reassigned erroneous values of $+\infty$. The original data value of 60 will now become the median, with 49 greater values, all $+\infty$. The median has admittedly moved, but only to the end of the original distribution, where it still has a 'sensible' value. If this exercise is repeated, but with the mean in place of the median, then the reassignment of any **single** data value to $+\infty$ will cause the mean to follow to $+\infty$ also. The resistance of the mean is zero, whilst that of the median is slightly less than 50%.

The resistance of the median derives from its 'safe' position with an equal number of data values above and below (leaving aside the problem of ties). It has very little regard for the actual data values, however extreme any of these may be, apart from the one or two central values which determine it.

7.1.2 Local Shift Sensitivity

The major weakness of the median is its very high sensitivity to local shifts in values at the centre of the distribution. As it is determined by only one or two data values, it strongly reflects any rounding or grouping of values, particularly where these are systematic. Again, an extreme example will most readily illustrate the problem.

Take a series of data sets, with distributions from strongly right skewed through to left skewed. Each data set has 100 data values falling in the range 6-24. The raw values are rounded to the nearest 10, with the result that the data set now consists of values of either 10 or 20. Suppose that at one skewed extreme there are 95 values each of 10 and 5 values of 20. At the other extreme the opposite proportions apply. As one progresses through the distributions, with gradually less 10 values and more 20's, the mean would increase slowly from 10.5 to 19.5. In contrast the median will start with a value of 10 and maintain this value until there are 49 10's and 51 20's in the distribution. With an equal number of 10's and 20's, the median will change to 15. Once the 20's predominate, the median will change to and maintain a value of 20. Where there are an odd rather than even number of data values, the change in median from 10 to 20 would be instantaneous. With such data the median gives a very poor measure of the centre of the distributions.

Using the original, unrounded, data the median would perform much better, but still in discrete steps.

7.1.3 Intermediate Central Statistics

Local sensitivity can be reduced if, instead of just the central one or two data values, the central 3 or 5 values (or 4 or 6 where the data set is even) are abstracted and the mean of these are taken. Such a measure is termed a broadened median. The greater the number of central values that are included, the lower will the local sensitivity of the broadened median become.

If the mean is used, all the data values are included, including any gross errors with their markedly disturbing effects. Instead of examining the extreme values, and possibly applying any rejection rules, a fixed proportion of data values, say 5%, could be eliminated from each end of the data distribution, and the mean of the central 90% taken. This would almost certainly remove any gross errors, except in a very small data set.

The mean and the median are, in fact, the two ends of a series of central measures termed the α-trimmed means, where α indicates the proportion of data trimmed from each end. The example quoted above is of a 5% trimmed mean - T(5%). The mean itself is of course 0% trimmed, whilst the median is a trimmed mean with a trimming of slightly below 50% (the exact percentage depends on the size of the data set).

Trimmed means overcome the major drawbacks of both the median and the pure mean. However, in certain 'ideal' situations, they will not prove as efficient as these measures (particularly the mean). The optimum percentage of trimming will vary, depending on the range of distribution types and the degree of contamination encountered. For a general wide applicability, Hoaglin et al. (1983) advocate the use of a 25% trimmed mean. This can be viewed as the even compromise between the mean and median, as it is the mean of the mid 50% of the data, and hence is generally termed the midmean. Hoaglin et al. (1983) recommend use of the midmean for data sets of 8 or more data values. For 7 values they would take the mean of the 3 central values. The median should be used for smaller data sets.

7.1.4 Robust Central Statistics

As measures of the centre of a distribution, the mean and median have strongly contrasting properties, which follow from the quite different manner in which they are derived from the data. Figure 17(a) illustrates the 'weight' which each attributes to data values at varying distances 'x' away from the respective statistics. For the mean, this weight is constant, irrespective of the distance of the value from the centre, however extreme. In contrast, the median attributes virtually all the total weight to the central one or two data values. The remaining small,

and decreasing, weight given to values away from the centre reflects the contribution which they make in actually determining, or rather limiting the possible range of, the centre values. For real data the desirable approach would lie between these two extremes.

For data values very remote from the body of the data, one can conclude that they are physically impossible or totally implausible in the circumstances. These can be rejected, or given zero weight. As one retreats from the extreme, a point is reached, albeit fuzzy, where one can say "this value could just be possible, but I still feel certain that it is a gross error". This outer limit is still given zero, or at least negligible, weight. Moving inwards, the probability that a data value is good will gradually increase and be attributed an increasing weight.

At the centre of a distribution, the data values would be full or maximum weight. Possible arguments in support of this might be:

a) If we don't believe these values, what faith do we have in any of the data?

b) The maximum probability that a data value is good must be at the centre of the data set.

c) The possibility that some of these values could be slightly erroneous, or poor, is accepted. However, even if a number of purely fictitious values were added at or near the centre of the distribution, their net effect on the determination of the centre would be negligible.

d) There are likely to be very few gross errors which fall near the centre, as their true values would have to be in the extremes of the distribution, where the probability of occurrence is by definition very low.

Moving away from the centre, through the bulk of the data, the confidence level remains high and the weight is maintained at or near maximum. As the tail area is entered, the probability of erroneous data values gradually increases.

Thus the weight given to data values should ideally vary in relation to their position in the distribution. Such a variation is illustrated in Figure 17(b). Here the weight is essentially the same as that accorded to the mean in the central area. Beyond a certain point (the rejection point) the weight is zero and data values, whilst remaining in the data set, are effectively ignored. On the intervening slope one attempts to relate weight to the plausibility of the data values.

Many possibilities exist for the shape and size of such curves. A major part of the recent research in robust statistics has been directed, not only to devising the mathematical functions which produce these

curves, but to evaluating which of the possibilities are the most useful and efficient. The concept of efficiency is essentially concerned with the accuracy or reliability of the calculated parameter or statistic for a given size of data set. It is used as a relative measure between the possible options rather than in an absolute sense. This rather complex topic is discussed fully by Hoaglin et al. (1983) and Hampel et al. (1986). As an example, the mean is the most efficient estimator for locating the centre of a pure Gaussian distribution and in this case is accorded 100% efficiency. However, if this Gaussian distribution is now contaminated with gross errors, or if the distribution is non-Gaussian and much heavier-tailed, the efficiency of the mean may drop drastically to 1%. For the same situations the median would have a much more consistent efficiency of perhaps 40-60% throughout. Some of the more sophisticated functions alluded to above would in the same cases maintain efficiencies of 80-90% or greater, whilst still having resistance close to 50%, as for the median.

Whilst these highly robust (efficient) statistics for a distribution centre may be desirable, their computation can be lengthy and require iteration. If relatively subtle differences between very similar distributions are being sought, then the effort in obtaining these statistics may be necessary. Where a central statistic is simply required as a component in summarizing a distribution a less rigorous approach may be quite adequate.

7.2 Statistics for the Distribution Spread

Having located the centre of the distribution the next statistic usually sought is some measure of its spread or dispersion. The many problems and considerations in deciding on statistics for the centre also apply to those for spread. In the case of the centre, there is at least the concept of a central value, the problem being to locate it. However for the spread there is no single value, even as a concept. A frequency distribution curve will obviously have different widths or spreads at different levels.

The conventional standard deviation is only adequate where the distribution is Gaussian. In this case the proportional shape of the curve is already defined and from a single statistic the width of the distribution at any level can be calculated. Without a defined shape the spread of the distribution at one level cannot be determined from that at another.

The usual nonparametric measure is the interquartile range (IQR), that is the range, or spread, between the lower and upper quartiles. Thus it is the spread of the central half of the distribution (with one quarter lying beyond each end). As with the median, this statistic has a high resistance, in this case 25%. An alternative, which gives an almost identical numerical value, is the Median Absolute Deviation from the median (MAD), which has a resistance of almost 50%.

There are a few robustly efficient measures of spread, but these are not frequently used. In fact there is little to be gained by using highly sophisticated measures of spread. Without a defined distribution shape, it is probably more useful to find a succinct way of presenting the spreads at a selection of different levels. The extended boxplot, to be discussed later, attempts to do this.

For a single figure statistic, the interquartile range (IQR) or MAD is in common use. An alternative could be a 'pseudo standard deviation', calculated as half the spread between the 16th and 84th percentiles. With a Gaussian distribution this would be virtually identical to the conventional standard deviation. With this same assumption the statistic can also be calculated as 0.6745 x IQR.

7.3 Higher Order Statistics

Hoaglin et al. (1983, 1985) and Hampel et al. (1986) discuss possible resistant and even robust measures for skewness and kurtosis. However these soon become very complex and would be of little benefit in the present context. If these aspects of a distribution are required, adequate information should be found in the probability plots and, to a lesser extent, the extended box plots.

8 BOX PLOTS

The box plot is a simple compact graphical method of summarising a frequency distribution, based on the resistant median and quartiles. The alternative term 'box and whisker' plot is also in use.

In Figure 18, a box plot is shown for a batch of liquid limit values. The ends of the box are drawn at the lower and upper quartiles (25 and 75 percentiles) with an internal division at the median value. The side bars or whiskers are drawn from the ends of the box to the lowest and highest data values that are not outliers. The outliers are represented by individual crosses beyond the whisker ends. The outline frequency distributions of several batches of data may be compared by drawing parallel box plots to a common data scale.

With a box plot it is possible to grasp the major aspects of a distribution at a glance. The centre of the distribution is shown by the median crossbar within the box. For an indication of spread, the interquartile range is shown by the length of the box. The whiskers illustrate the tail lengths of the distribution. The relative position of the median crossbar within the box and the relative lengths of the whiskers indicate the skewness of the distribution.

The simple box plot portrays the skeleton of the distribution of the actual data. A common modification is the notched box plot, which indicates the extent to which the total population distribution can be inferred from the actual data distribution. The width of the notch (to be read against the same data scale) is usually calculated such that there is a minimum 95% probability that the population median will lie within the limits of the notch.

To a first approximation the confidence with which the parameters of an actual distribution can be used to infer those of the total population increases as the square root of the number of data values. Thus if the height of the boxes is drawn in proportion to the square root of the size of each data set, the relative significance of each can be compared.

As a summary of a geotechnical property distribution the box plot has two particular limitations. There is a simple convention to determine whether a value will fall within a tail whisker or be classed as an outlier. The lower and upper cutoffs are 1.5 x IQR below the lower and above the upper quartiles respectively. However this approach is rather too simplistic where the distribution is appreciably skewed or has a particularly non-Gaussian kurtosis. In these cases reasonable tail values will be classed as outliers and vice versa. It would be preferable to determine the two cutoffs separately, with regard to the distribution in each tail area. This would also help in determining a realistic 'effective range' within which the great bulk of the distribution falls.

The second limitation is again concerned with the tail areas. By far the greatest distinction between the many distributions encountered is to be found in the tail areas. The mid part of a distribution is usually very well defined by just the median and quartiles (in fact, Winsor's principle [Hoaglin et al. 1983, p.363] states that "all distributions are normal in the middle"). The conventional box plot gives no information between the quartiles and the ends of the tail whiskers.

It is proposed that a refinement of the notched box plot should be used for geotechnical data distributions. This could be referred to as an 'extended notched box plot' or 'chequered notched box plot'.

As the first stage in constructing the plot, a set of simple percentiles are calculated, such as those at 1, 2, 5, 10, 25, 50, 75, 90, 95, 98 and 99 percentages. The 25, 50 and 75 percentiles are used to construct the central box with median crossbar, as for the usual plot. The remaining percentages are used to define a series of subsidiary boxes to either side of the central box (Figure 19).

The heights of the various boxes are again scaled in proportion to the square root of the number of samples 'contained' in each. Thus for a data set of 500 values, these heights would be calculated as follows:

Box limits	Total %	No values	Height
25-75	50	250	15.8
10-25 , 75-90	30	150	12.2
5-10 , 90-95	10	50	7.1
2- 5 , 95-98	6	30	5.5
1- 2 , 98-99	2	10	3.2

In order to distinguish between, and readily recognise, the successive boxes, they are shaded alternately. Thus the outer limits of the shaded boxes will fall at the 1, 5, 25, 75, 95 and 99th percentiles.

Typically, most actual data batches will be of insufficient size to calculate the outer percentiles and will have only perhaps one or two data values contained within the outermost boxes. Therefore, to ensure that the plot is reasonably meaningful, it is necessary to limit the number of subsidiary boxes with regard to the size of the data set. It is proposed that the outermost box, at each end, should contain a minimum of 3 values and that at least 2 further values should fall beyond this box. Using this rule, the plot format would be as follows:

Data values	Outer box limits
10 - 19	25, 75%
20 - 59	10, 90
60 - 99	5, 95
100 -299	2, 98
300+	1, 99

Within the second boxes from the centre a short bar is drawn at the 16th and 84th percentiles. These bars serve two purposes. Firstly, the distance of each from the median would be equal for a Gaussian distribution and almost identical to the standard deviation. Secondly, if an outline frequency distribution is reconstructed from the box plot, these values fall within what would otherwise be considerable gaps.

The only data values not considered so far would be those represented by the whisker lines and any outliers. Whilst a rule, albeit rather complex, probably could be devised for determining suitable outlier cutoff limits in sympathy with the frequency distribution, it would be preferable to use the probability plots to determine these limits and/or eliminate the outliers.

The software required to calculate and display these modified boxplots should not be unduly complex. Such plots would offer the following advantages:

a) Compact graphical displays are used to compare the distributions of several data sets (Figure 20).

b) The distribution centre and several measures of its spread are shown.

c) The height of each display and the number of subsidiary boxes indicate the significance which should be given to each data set.

d) The width of the notch gives conservative 95% confidence limits for the population centre. This is the narrowest confidence interval for any percentile, as the distribution is densest at the centre. For box limits (percentiles) further from the centre, the confidence intervals would rapidly widen, at least for near-Gaussian data.

e) Being based on percentiles, the box plot is resistant to any major disturbance by gross outliers and is not dependant on any underlying frequency distribution. It emphasises the structure of the bulk of the data.

f) The outermost boxes indicate the rough limits to which any statements concerning the distribution tails are justified by the data as at all meaningful or worthwhile (the outermost limit shown should always be treated with considerable caution). Thus it may be of practical use to say that 90% of the actual data have values above 'x', or 98% below 'y', whereas the conventional range (between the most extreme values) is virtually worthless.

9 SUMMARY AND CONCLUSIONS

a) It is almost inevitable that the data compiled in a geotechnical database will be 'dirty' in a statistical sense. There are many potential sources of error in the numerical values, the spatial distribution of the data is usually poor, and the allocation to geological units cannot be achieved with consistent reliability.

b) For data of this nature it is much more appropriate to take a 'robust' rather than classical approach to statistics. By placing emphasis on the structure exhibited by the bulk of the data, a higher level of confidence can be placed in the reliability of the resultant statistics.

c) Graphical rather than purely numerical displays are much to be preferred, both for analysis of the data and its summarisation.

d) Histograms should generally be avoided as great care may be required in their formulation. The stem-and-leaf display is more reliable where a bar display is required in data analysis.

e) The most valuable tool for data analysis is the probability plot. It will reveal the structure and coherence of a data batch and provide a basis for identifying possibly erroneous data values.

f) The probability plot is the best means for assessing the skewness of a distribution. Unless there are good reasons for not doing so, the data axis should be rescaled if the skewness can thereby be significantly reduced. Often this will entail the use of a logarithmic rather than an arithmetic scale. Skewness is undesirable in that the apparent spread or dispersion should generally be independent of the data level.

g) The probability plot will indicate the degree to which the distribution is degraded by the finite precision of the data values. Where the plot is significantly stepped rather than continuous, it would be preferable to replot it manually from the cumulative frequency of 'bin' values.

h) The classical parametric statistics, such as the mean and standard deviation, should be avoided as a means of summarising a distribution. They place undue weight on the tail values and can be seriously misleading where the distribution is non-Gaussian. The range is a particularly poor statistic, as it takes account of only the most extreme values and will generally increase with the size of the data batch.

i) For a numerical summary it is preferable to use a selection of percentiles, including the median and quartiles. These are highly resistant to erroneous values in the data batch.

j) A fuller and more informative summary of a distribution is provided by the extended box plot. This graphical display will emphasise the essential structure within a distribution and the significant differences between distributions. It also indicates the degree of confidence that may be placed in the summary. To fully utilise the 'robust' approach, the required percentiles for a box plot should be abstracted from a manually smoothed probability plot.

10 REFERENCES

Biernatowski, K. 1985. Statistical characteristic of subsoil. Proc. XI Int. Conf. Soil Mech. Found. Engng. San Francisco, USA, August 1985; Rotterdam-Boston: A. A. Balkema 1985, Vol. 2, pp 799-802.

Conover, W.J. 1980. Practical nonparametric statistics 2nd Edn. Wiley 493 pp.

Corotis, R.B., Azzouz, A.S. and Krizek, R.J. 1975. Statistical evaluation of soil index properties and constrained modulus. ICASP 2, Aachen, pp 273-292.

Ejezie, S.U. and Harrop-Williams, K. 1984. Probabilistic characterization of Nigerian soils. ASCE Symposium, Probabilistic characterization of soil properties, Atlanta 1984, pp 140-156.

Fredlund, D.G. and Dahlman, A.E. 1971. Statistical geotechnical properties of glacial Lake Edmonton sediments. ICASP I, Hong Kong, pp 203-228.

Gibbons, J.D. 1985. Non parametric methods for quantitative analysis. 2nd Edn. Amer. Sci. Press.

Goldberg, G.D., Lovell, C.W. and Miles, R.D. 1978. Use the geotechnical data bank! Transp. Res. Rec. 702, pp 140-146.

Goldberg, G.D., Lovell, C.W. and Miles, R.D. 1980. Computerized information system for Indiana soils. Trans. Res. Rec. No. 733, pp 74-82.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. 1986. Robust statistics - the approach based on influence functions. Wiley 502 pp.

Henley, S. 1981. Nonparametric geostatistics. Applied Science Publishers 145 pp.

Hoaglin, D.C., Mosteller, F., and Tukey, J.W. 1983. Understanding robust & exploratory data analysis. Wiley 447 p.

Hoaglin, D.C., Mosteller, F. and Tukey, J.W. 1985. Exploring Data Tables, Trends & Shapes. Wiley, 527 pp.

Hollander, M. and Wolfe, D.A. 1973. Nonparametric statistical methods. Wiley, 503 pp.

Lo, Y.K.T. and Lovell, C.W. 1982. Prediction of soil properties from simple indices. Transp. Res. Rec. No. 873, pp 43-49.

Lo, Y.K.T. and McCabe, G.P. 1984. Characteristics of Indiana soil properties. ASCE Symposium, Probabilistic characterization of soil properties, Atlanta, pp 106-118.

Lovell, C.W. and Lo, Y.K.T. 1983. Experience with a state-wide geotechnical data bank. Proc. 20th Boise Symp. 'Engng Geol. a. Soils Engng', April 1983, Idaho St Univ. pp 193-203.

Lumb, P. 1966. The variability of natural soils. Can. Geotech. Jnl., Vol. 3, No. 2, pp 74-97.

Lumb, P. 1970. Safety Factors and the probability distribution of soil strength. Can. Geotech. Jnl., Vol. 7, No. 3, pp 225-242.

McGill, R., Tukey, J.W. and Larsen, W.A. 1978. Variations of Box Plots. The American Statistician, Vol. 32, No. 1, pp 12-16.

McGuffey, V., Iori, J., Kyfor, Z and Athanaziou-Grivas, D. 1980. Statistical Geotechnical Properties of Lockport Clays. Transp. Res. Rec., No. 809, pp 54-60.

Rethati, L. 1983. Distribution functions of the soil physical characteristics. Proc. 8th Europ. Conf. Soil Mech. Found. Engng., Helsinki, pp 405-410.

Rethati, L. 1988. Probabilistic Solutions in Geotechnics. Developments in Geotech. Engng. Vol. 46, Elsevier 451pp.

Schultze, E. 1971. Frequency Distributions & Correlations of Soil Properties. ICASP I, Hong Kong, pp 371-388.

Tukey, J.W. 1962. The future of data analysis. Ann. Math. Stats., Vol. 33, pp 1-67.

Velleman, P.F. and Hoaglin, D.C. 1981. Applications, Basics & Computing of Exploratory Data Analysis. Duxbury Press 354 pp.

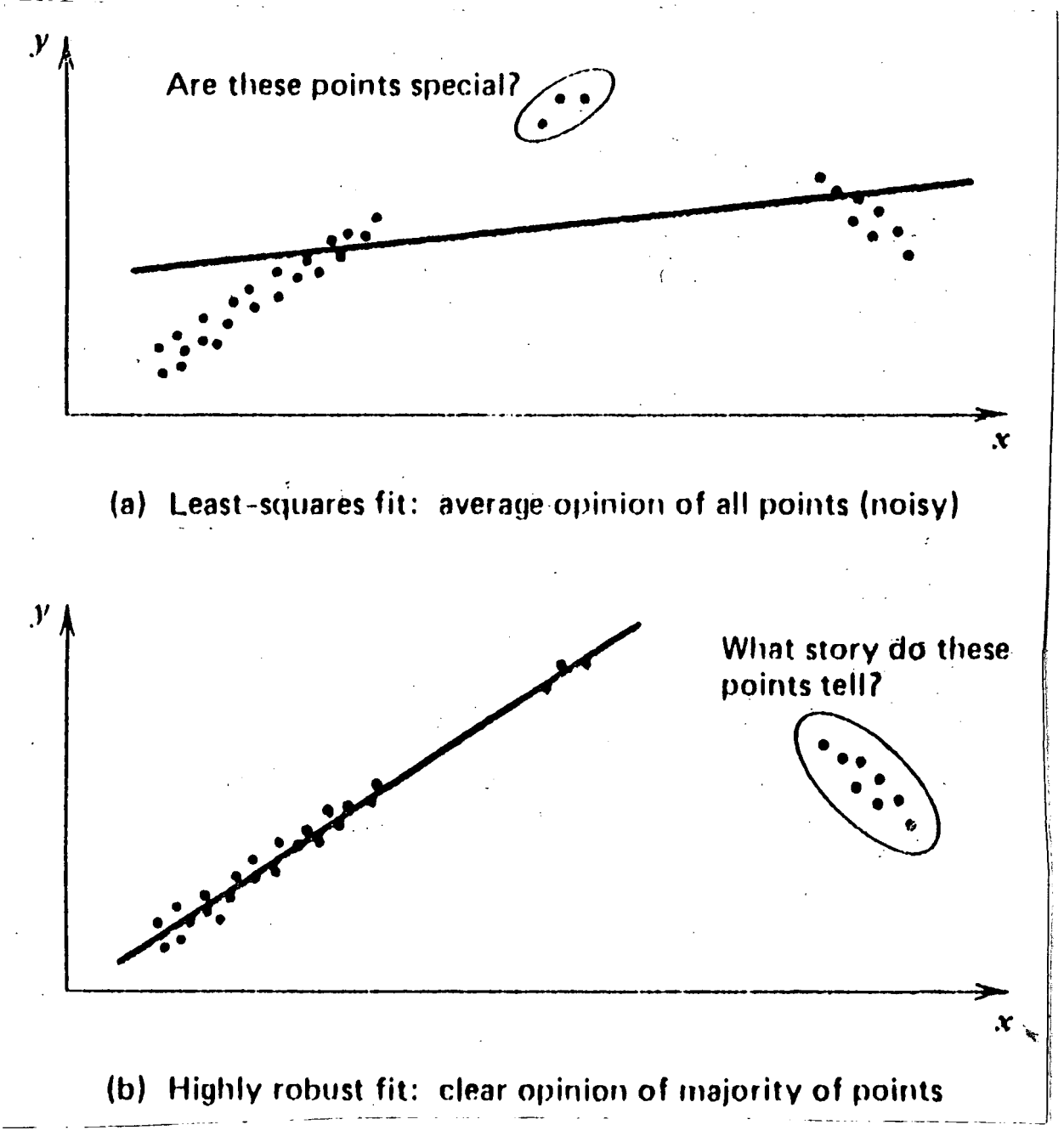

Figure 1. Classical and robust approaches to statistics

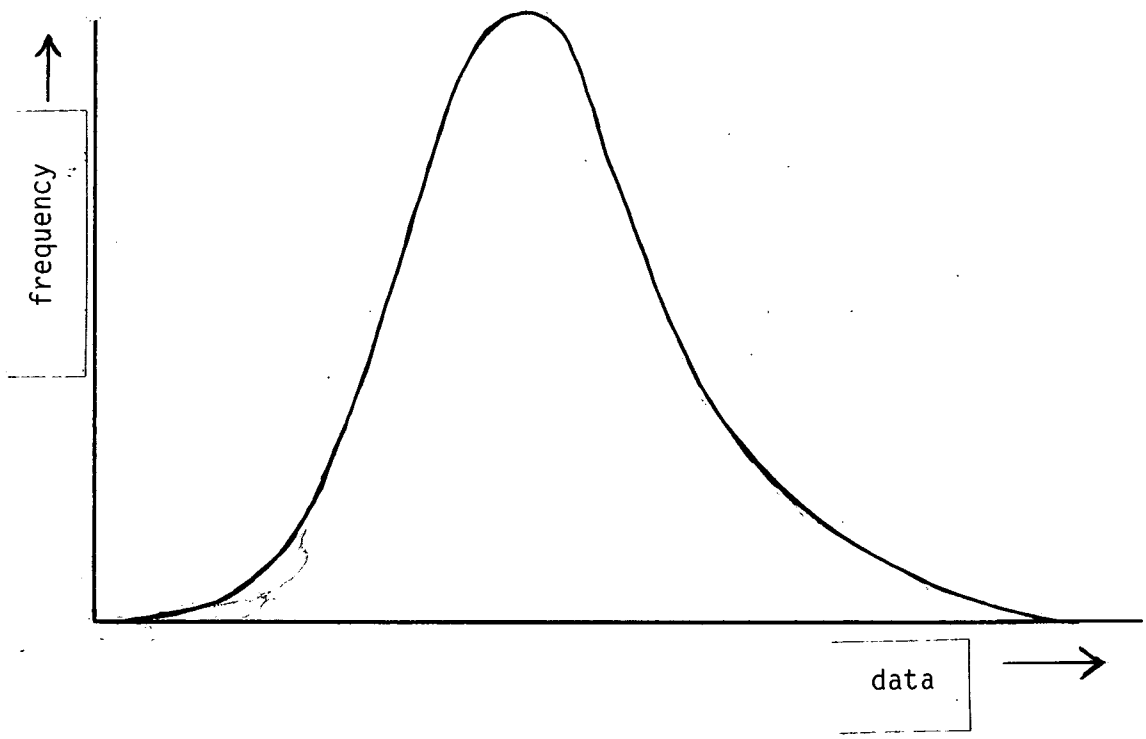

Figure 2. Frequency distribution curve

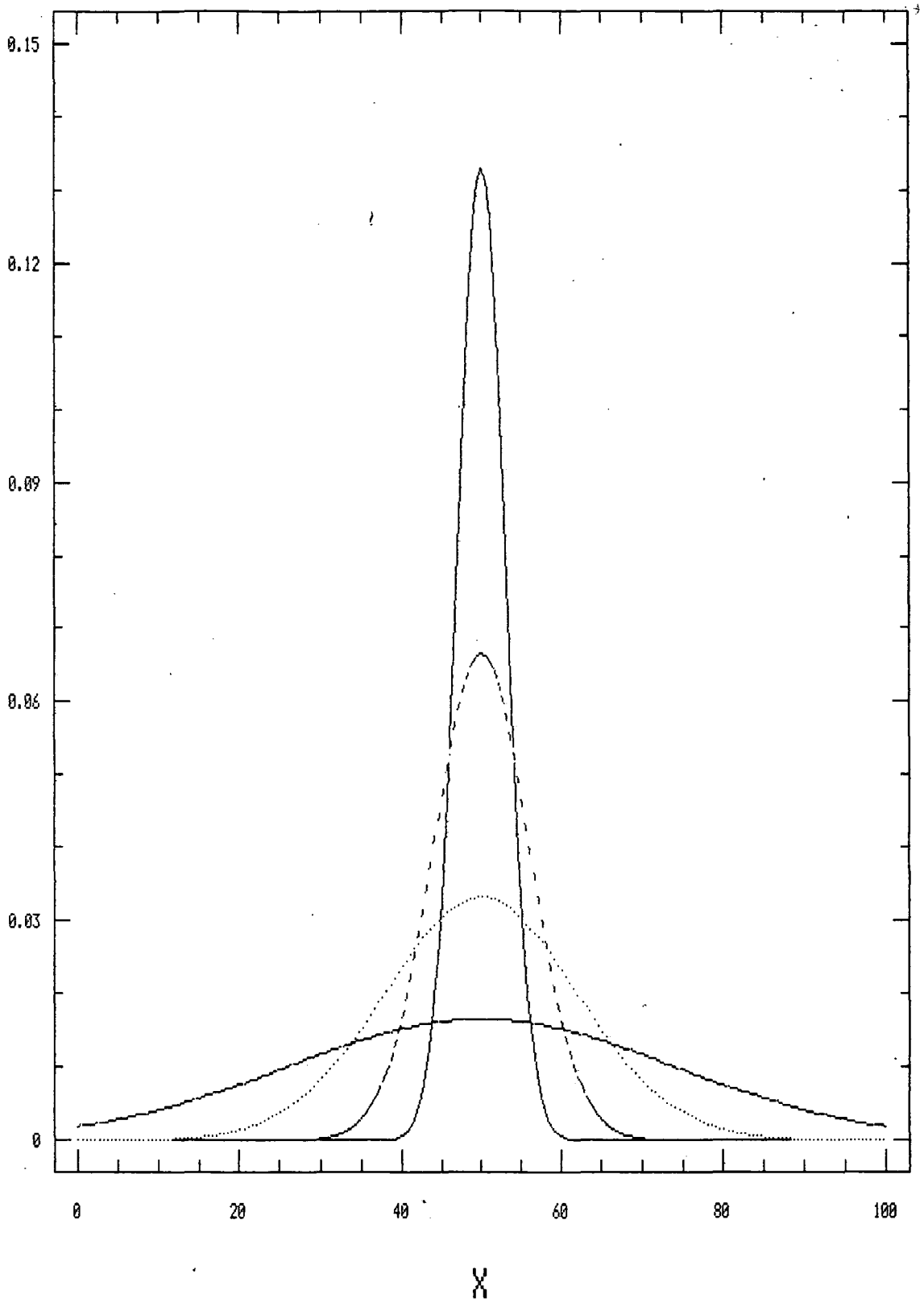

Figure 3. Four Gaussian distributions with the same mean but different standard deviations

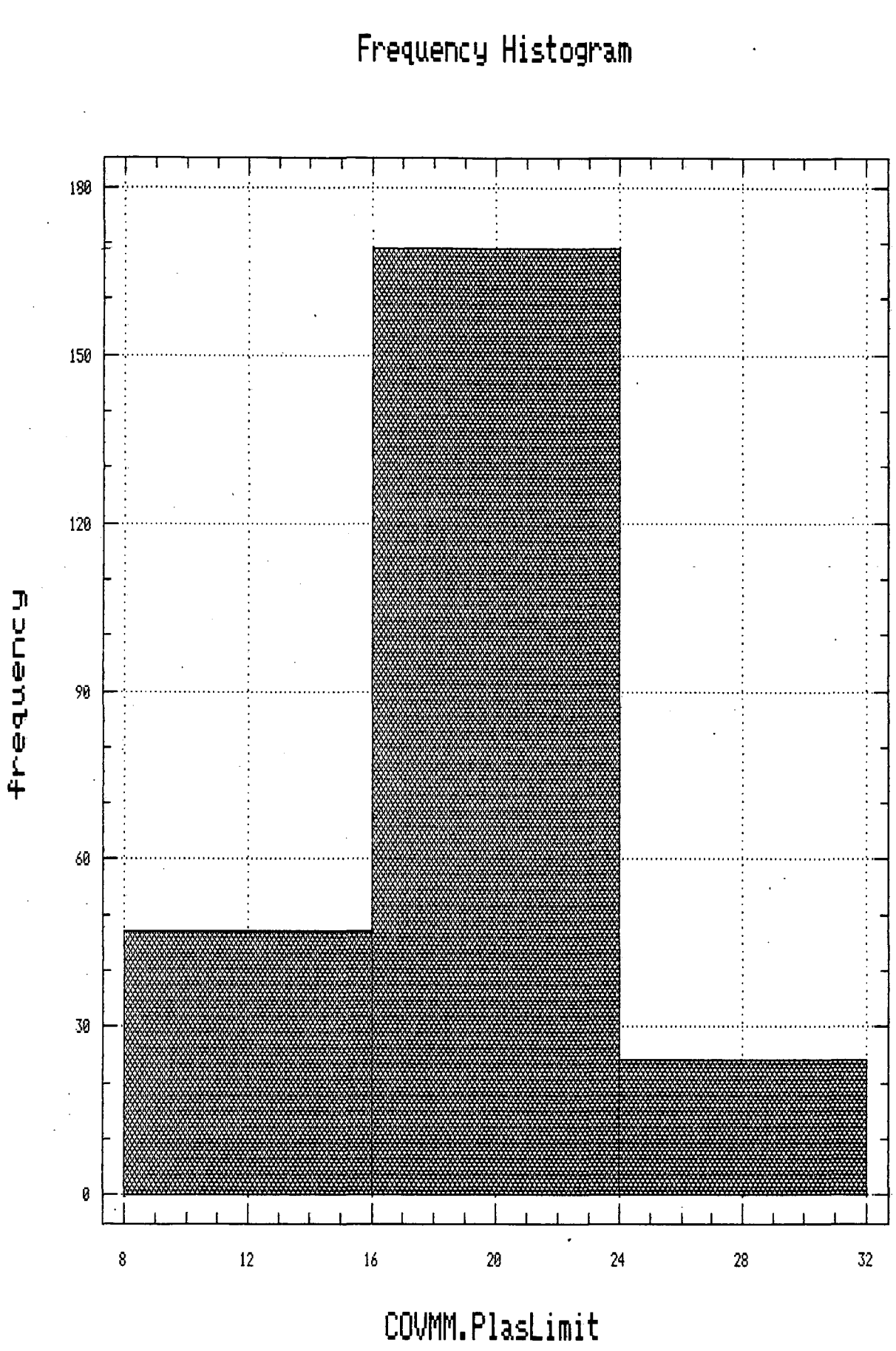

Figure 4. Histogram with excessively broad class intervals

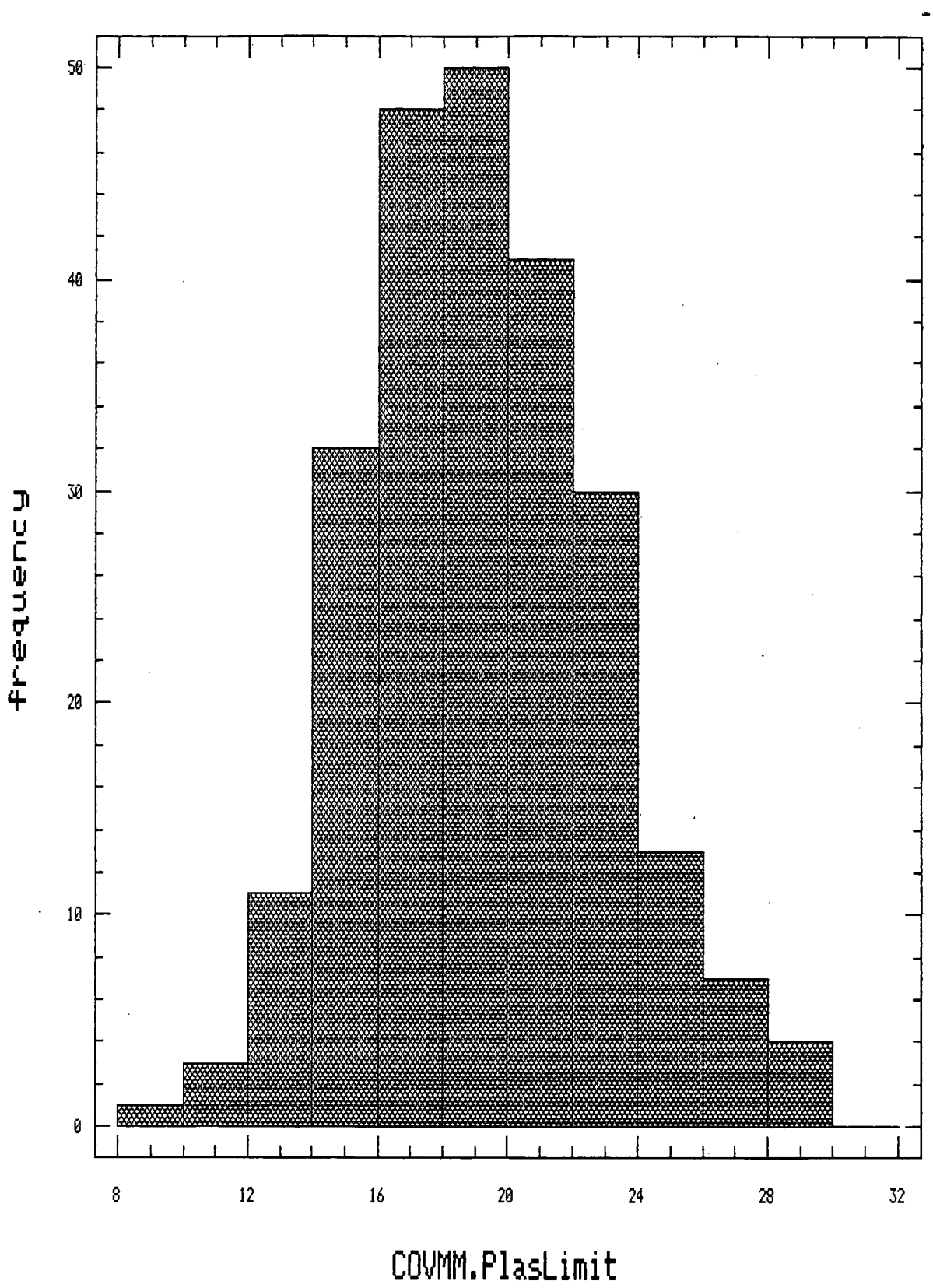

Figure 5. Well constructed histogram

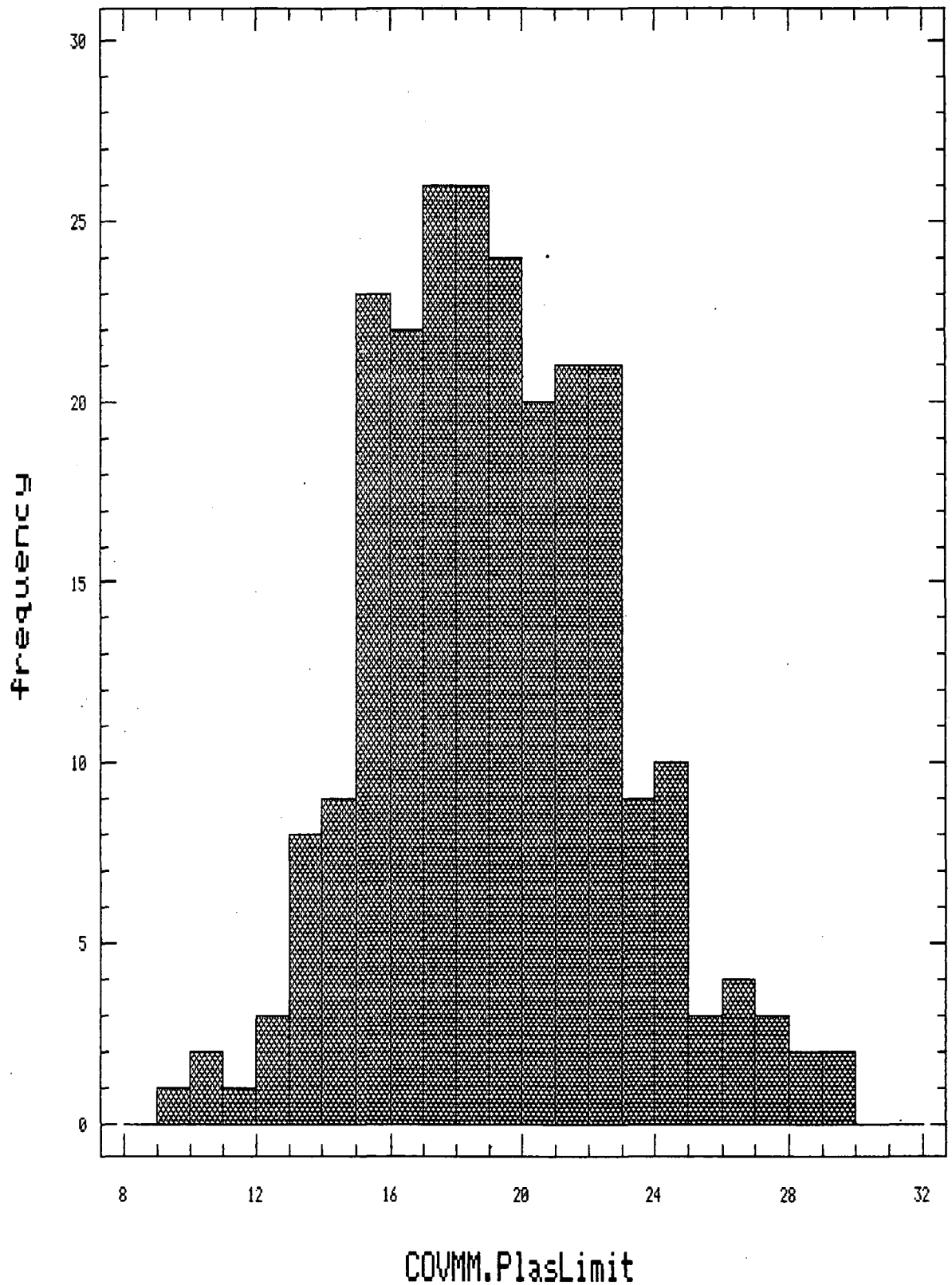

Figure 6. Histogram with too many classes

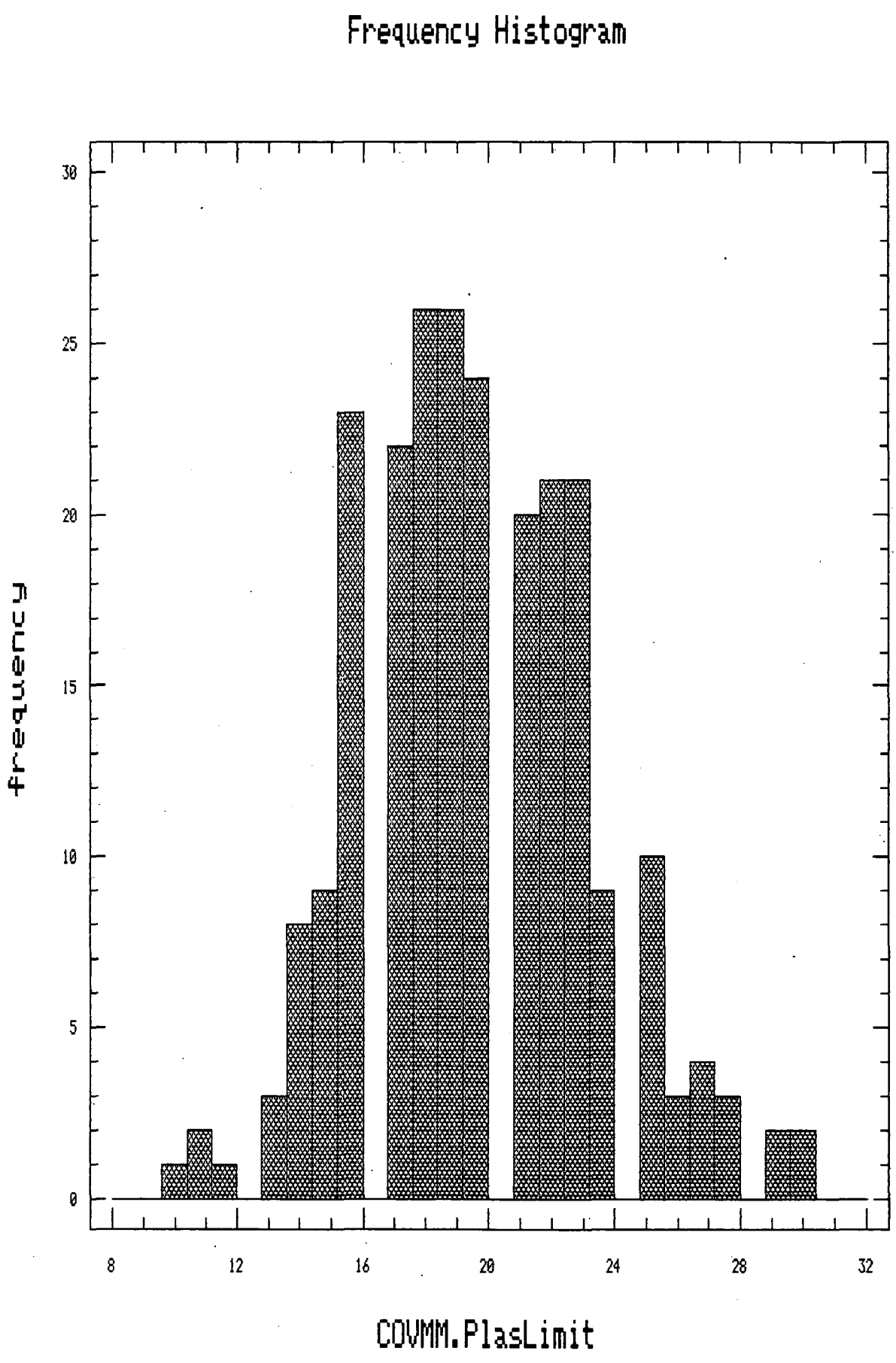

Figure 7. Misleading histogram, where the gaps are merely an artifact of the poorly chosen class intervals

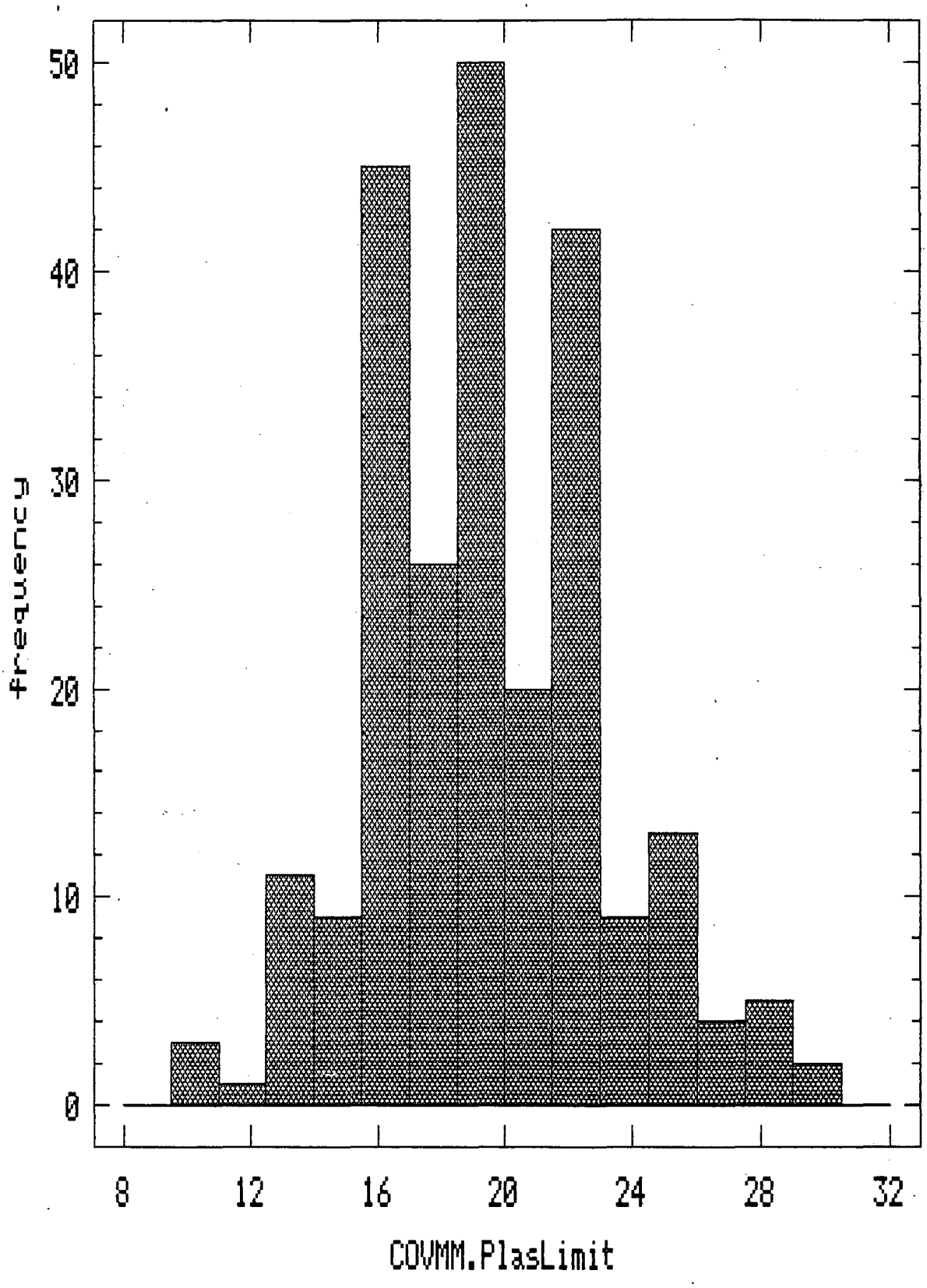

Figure 8. Another misleading histogram resulting from inappropriate class intervals

(a) Stem-and-leaf display for COVMM.DryDens: unit = 0.01 1|2 represents 0.12

```
  1   11|3
  2   12|8
  7   13|03569
 15   14|34779999
 51   15|011111122222244444555555577777779999
 72   16|002222222234677777888
(25)  17|000011122333334667788899
 48   18|001111111334444445555667799
 21   19|1222233344445555667
  2   20|3
  1   21|0
```

(b) Stem-and-leaf display for COVWC.BulkDens: unit = 0.01 1|2 represents 0.12

```
      LO|138,154,158,163,165,166

 10   19*|1244
 28   19o|555666777888889999
 51   20*|00222222222233333444444
(29)  20o|55555555666666667778888889999
 28   21*|000000000111122444
 10   21o|66889
  5   22*|11224
```

(c) Stem-and-leaf display for COVMM.PlasLimit: unit = 1 1|2 represents 12

```
  3    1*|011
  7    1T|2333
 24    1F|44444444555555555
 69    1S|666666666666666666666667777777777777777777777
(52)   1o|88888888888888888888888889999999999999999999999999
120    2*|000000000000000000000000111111111111111111111
 76    2T|22222222222222222222233333333333333333333
 34    2F|4444444445555555555
 15    2S|6667777
  8    2o|88899

      HI|30,30,35
```

Figure 9. Stem-and-leaf displays, with 1, 2 and 5 lines per stem

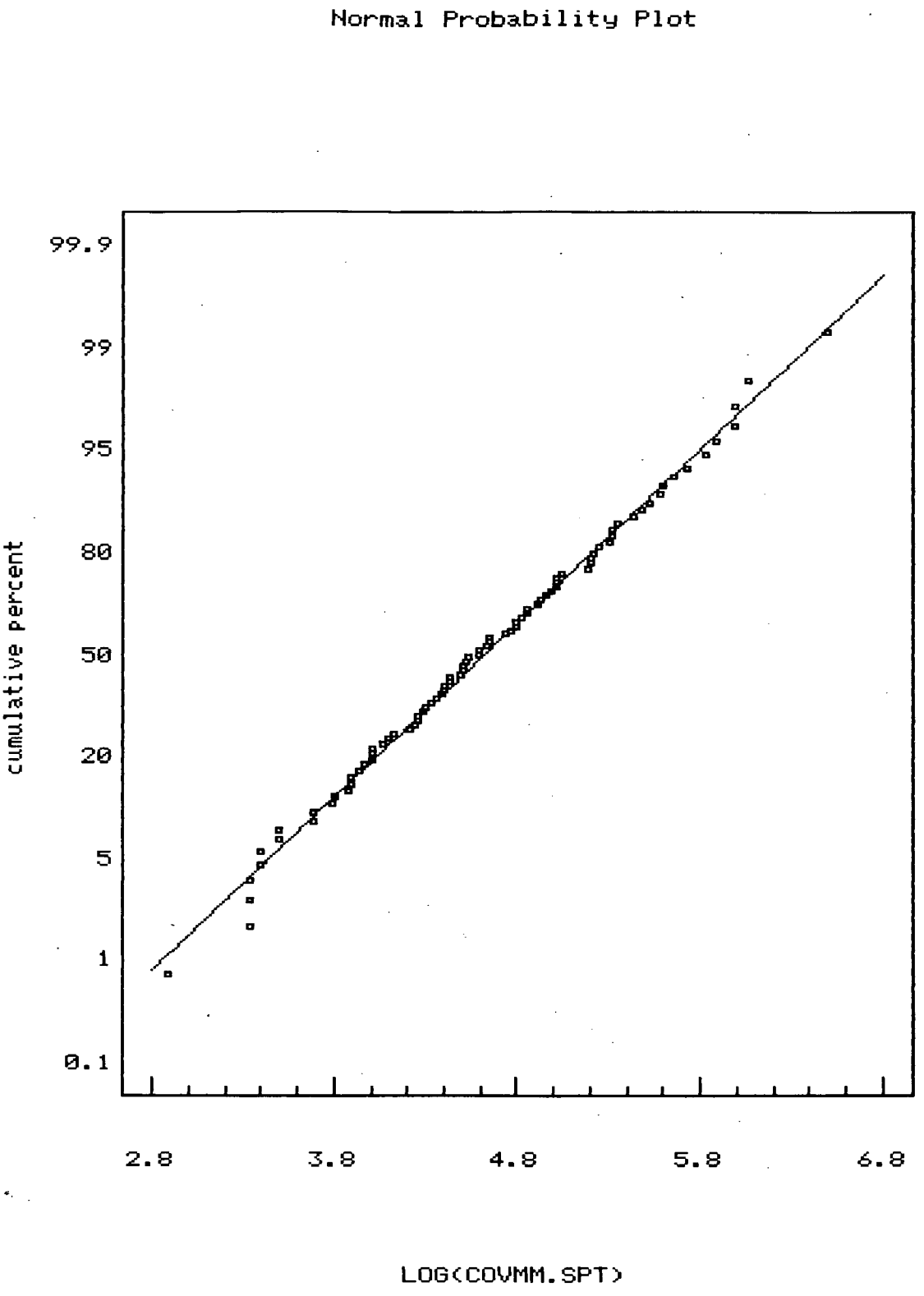

Figure 10. Probability plot for a near-Gaussian distribution

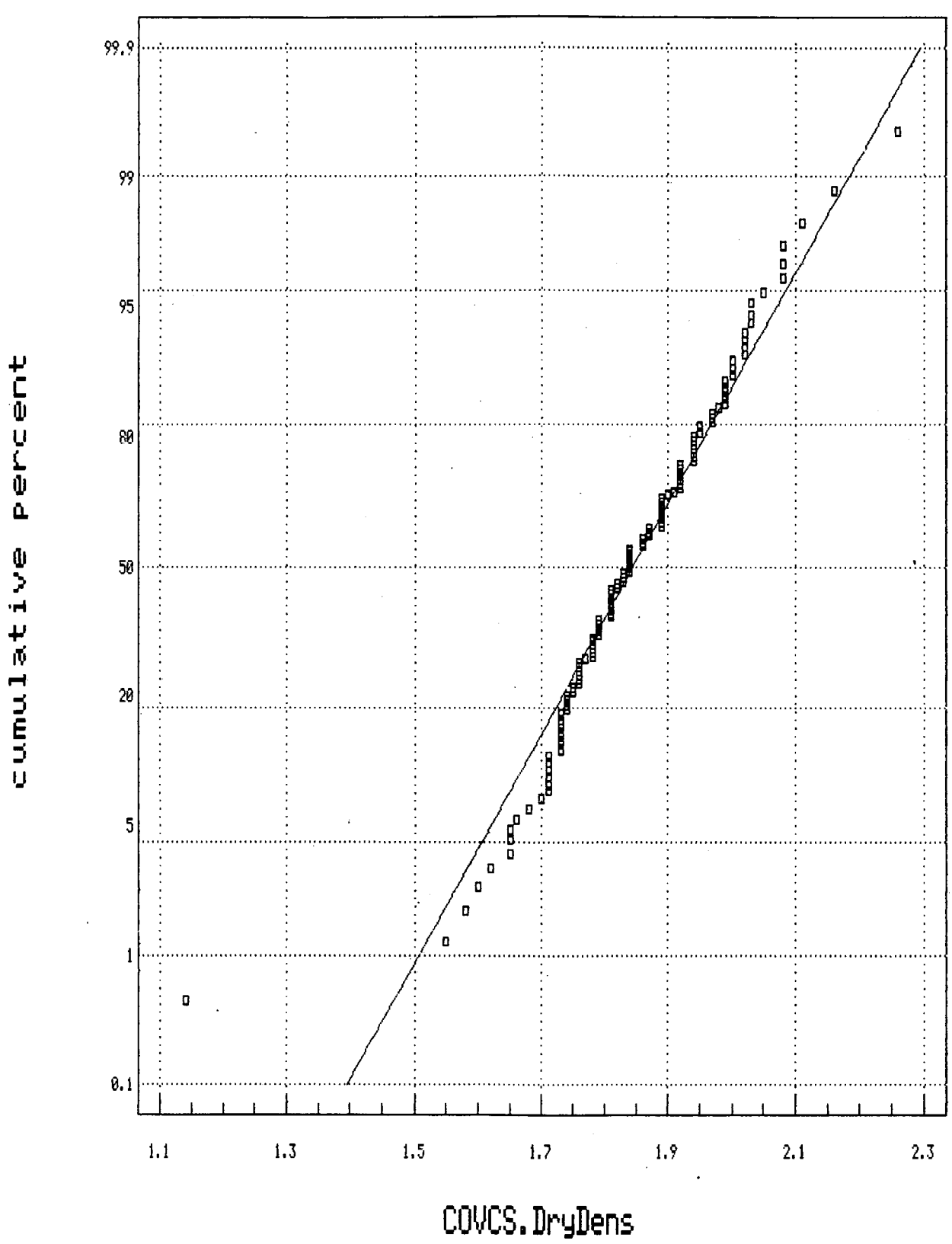

Figure 12. Probability plot with one anomalous low value

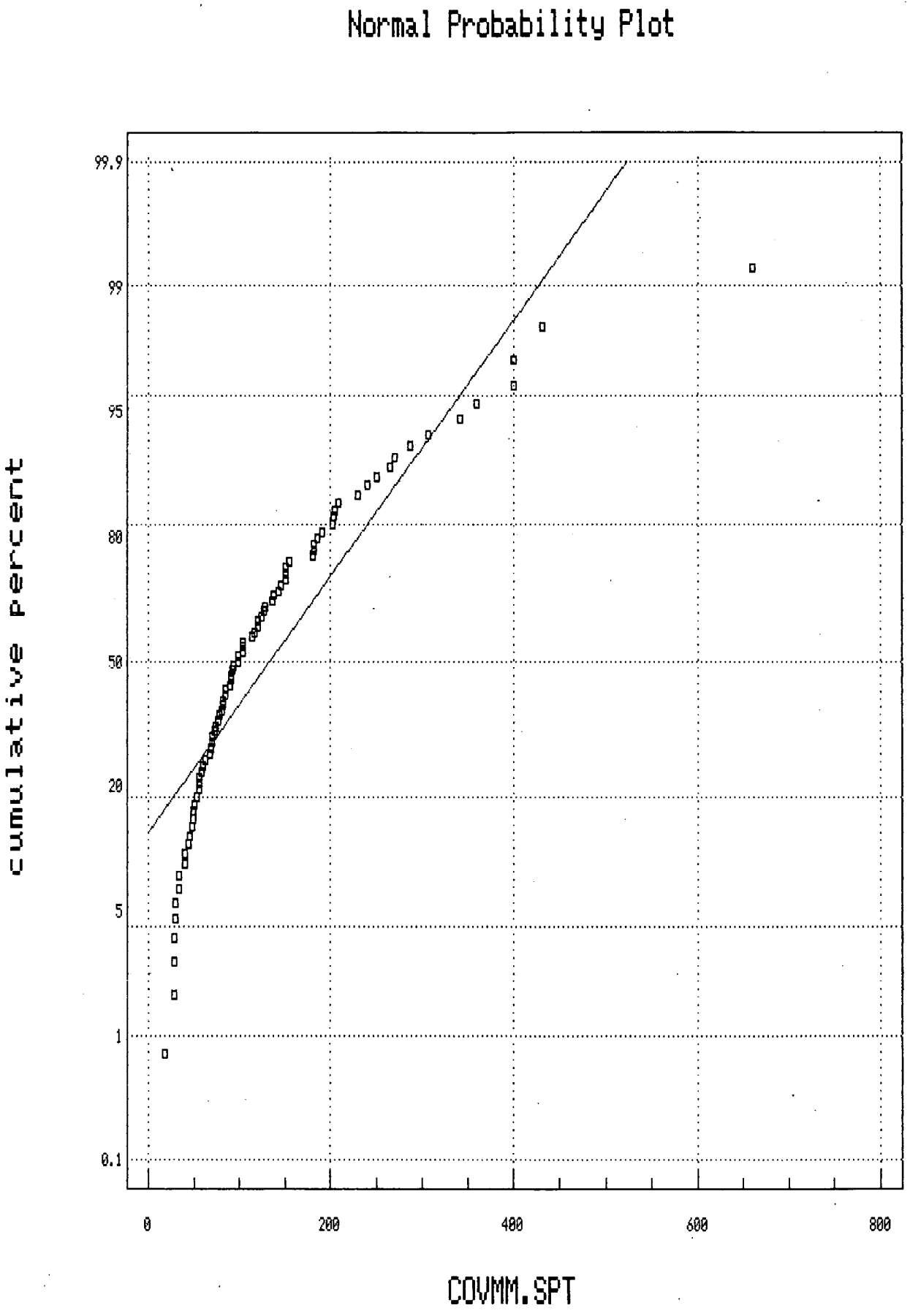

Figure 15. Right-skewed probability plot (cf. Fig. 10)

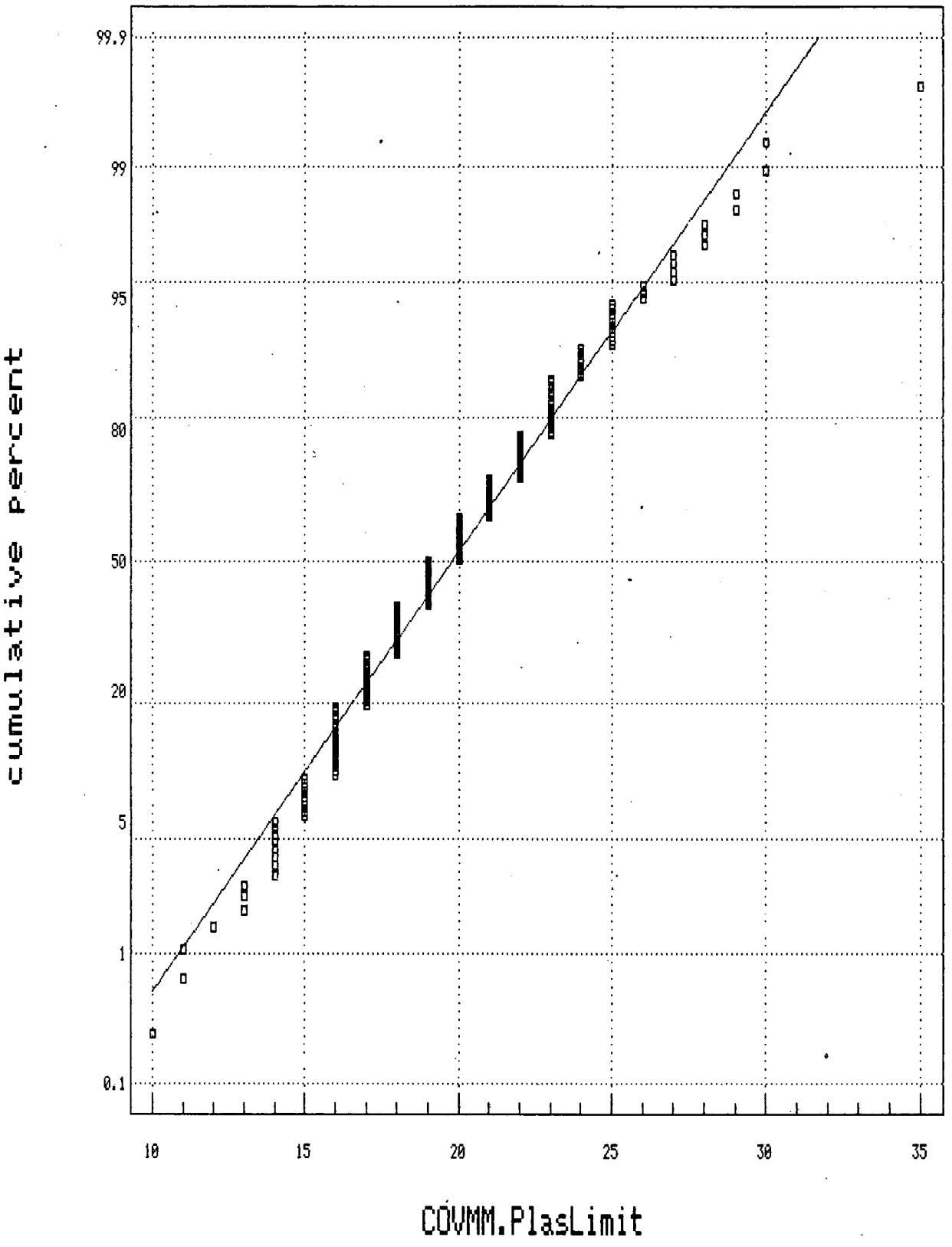

Figure 16. Coarsely stepped probability plot, due to low data precision

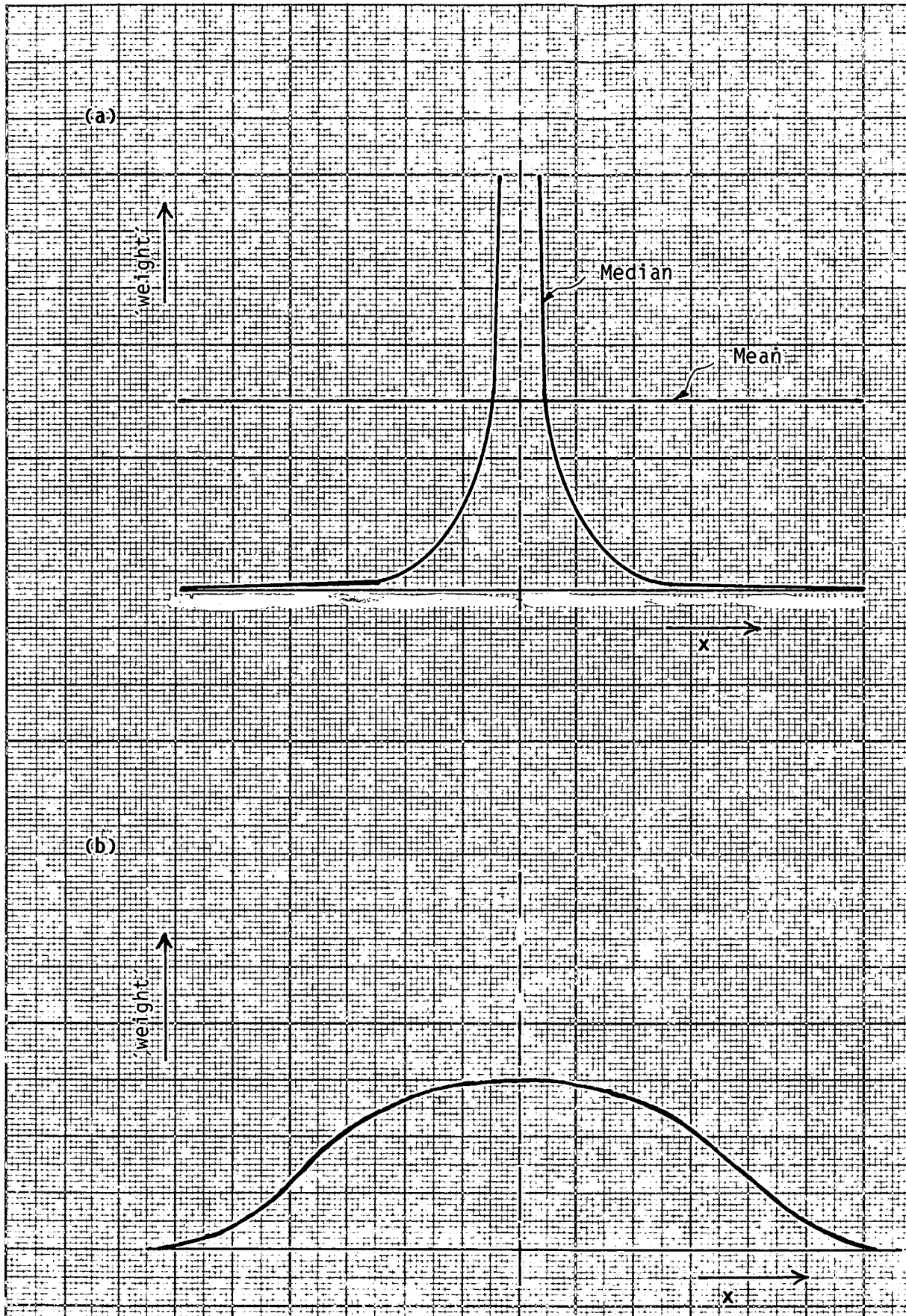

Figure 17. 'Weights' effectively utilised in deriving various statistics for central location, (a) mean and median, (b) robust estimator.

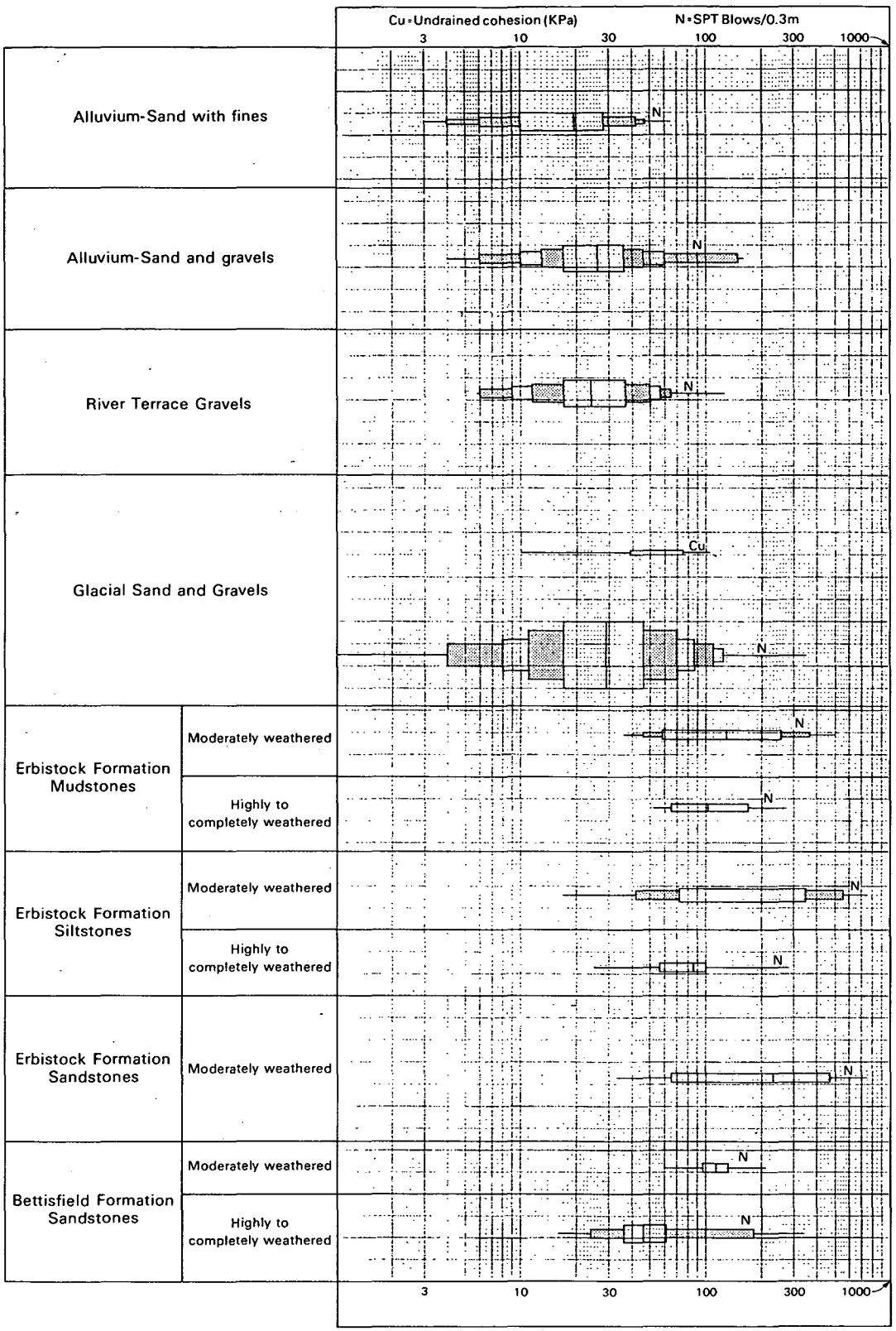

Figure 20- Example of Extended Box Plots for summarising and comparing geotechnical data

This folder contains negatives