International Council for the Exploration of the Sea

CM 1980/D:10
Statistics Committee
Ref: Hydrography Committee

SOME STATISTICAL TECHNIQUES FOR LOOKING AT LARGE DATA SETS

by

J A Pope
DAFS Marine Laboratory, Aberdeen, Scotland, UK

Abstract

Some statistical methods for studying and analysing extensive sets of data are described. Methods suitable for calculation by hand, as well as ones requiring the use of a computer, are presented.

Résumè

On décrit quelques méthodes statistiques pour étudier et traiter des séries d'informations approfondies. On présente des méthodes pour lesquelles une évaluation non mécanique suffit ainsi que celles qui nécessitent l'emploi d'un ordinateur.

1 Introduction

Nowadays the collection of large data sets, especially by automatic means, is particularly easy and very widespread. The simplicity with which this can be done often results in inadequate thought being given to a proper consideration of the relevance of the data to the purpose in hand and to the sort of questions the data will be expected to answer. Indeed, there is a widely held attitude that if the cost of collecting observations on some phenomenon is relatively small one might as well collect the data as not. All too often, once the data have been collected, this attitude changes to a belief that, since there are a lot of data, they must contain valuable information on something which will be magically revealed by the application of some statistical method, usually by way of a computer package. Proper consideration at the planning stages of any study entailing the collection of data and the application of sound statistical methods of analysis to the data collected are essential.

A useful discussion of some of the principles to be followed in the collection of data, particularly in the context of observational studies and experimental design, is to be found in the Reports of the Working Group on Standardisation of Scientific Methods for Comparing the Catching Performance of Different Fishing Gear (Anon, 1974, 1977). The present paper is restricted to a description of some procedures which are useful for the preliminary study of data. The procedures selected have been chosen because of their relevance to situations where the amount of data available for analysis is large. Some of the procedures described require little more than paper and pencil but others can only be attempted realistically with the aid of an electronic computer. None of the methods are new but most are sufficiently new for them not to be widely known in fisheries and possibly oceanographic research.

2 Characterization of Univariate Data Sets

On being presented with a large, single, unclassified set of data the first task will usually be to display its main features by means of some simple descriptive measures. Most proceedures for summarizing data require the observations to be sorted or partially sorted. If $x_1, x_2, \ldots\ldots, x_n$ are the elements of the data set, the (fully) sorted array is the permutation $x(j)$ such that

$$x(1) \leq x(2) \leq \ldots\ldots \leq x(n)$$

The element $x(j)$ is called the jth order statistic. The partially sorted array is $x^*(j)$ where

$$(a)\quad x^*(j) = x(j) \qquad j = j(2), j(3), \ldots\ldots, j(k)$$

$$(b)\quad x^*(j(l)) \leq x^*(j) \leq x^*(j(l+1)) \qquad j(l) \leq j \leq j(l+1)$$

where $0 = j(1), j(2), \ldots\ldots, j(k), j(k+1) = n + 1$ are $(k + 2)$ ordered integers. Good sorting algorithms already exist and special sorting programs have been developed. Where a large number of observations are to be sorted such programs should be used. A full account of sorting procedures is given by Knuth (1973).

Order statistics are required in short-cut and robust estimation procedures. For instance the median of a set of n observations is the $(n + 1/2)$th order statistic if n is odd or the mean of the $(n/2)$th and $(n/2 + 1)$th order statistic if n is even. Other useful short-cut estimators of the centre of location of the data which employ order statistics are

$$\{ x(\tfrac{1}{4}) + 2x(\tfrac{1}{2}) + x(\tfrac{3}{4}) \}/4$$

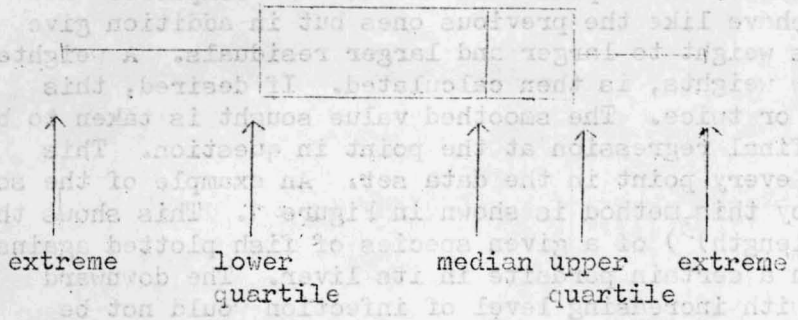
$$\{ x(1/16) + x(\tfrac{1}{4}) + 2x(\tfrac{1}{2}) + x(\tfrac{3}{4}) + x(15/16) \}/6$$

where $x(f)$ is the value of x exceeded by fn of the observations. These estimators are very efficient and, once the data have been sorted, very easy to calculate when no computing aid is available. When a computer is available, calculations of the ordinary arithmetic mean of a large set of data is trivial and will usually provide the best estimate of the centre of location. When the data are classified into several categories resulting in relatively small numbers (n_j) of observations in each category (say $10 \leq n_j \leq 50$), estimators of the central value in each category may require the use of robust estimation proceedures (Andrews et al., 1972).

Besides a measure of the centre of location of the data set some measure of the extent of variability will be required. Several quick methods, based on order statistics are available for estimating the standard deviation of the data. For example,

$$\{x(1/16) + 3x(\tfrac{4}{4})/4 - 3x(\tfrac{3}{4})/4 - x(15/16)\}/4$$

However, it is almost always much more informative to describe the variability in the data in visual form and for this the "box-and-whisker" plot suggested by Tukey (1977) is eminently suitable. This is obtained by first finding the median, the upper and lower quartiles and the extremes. A long box is then drawn between the quartiles with a bar denoting the relative position of the median. From the ends of the box two "whiskers" are drawn to the extremes. An example is shown below.



A box-and-whisker plot

This type of display is particularly easy to understand and very informative when several batches of data are to be compared. Use of a one-way analysis of variance, a proceedure often resorted to, for the comparison of several batches of large data sets is not recommended. The analysis of variance is a technique for testing hypothesis about the equality of mean values only and, when large numbers of observations are involved, will simply state the obvious. Elaborations of the "box-and-whisker" plot described here are given by McGill et al., (1978). These plots are very simple to draw by hand and relatively easy to produce on a line printer or visual display screen.

3 Looking for Relationships

When observations are available on two or more variates it will often be of interest to determine whether or not statistical relationships exist among them and, if so, to estimate these. Under conditions where all factors which may significantly influence the values assumed by the variates are carefully controlled, any relationships which exist should be easy to detect and estimate. Provided the statistical error distributions of the data are normal (Gaussian) or nearly normal, ordinary least squares will usually provide the best method of estimation. In general practice such text-book conditions are rarely met. The existence and form of any relationships are usually not known a priori and, particularly in studies involving extensive data sets, there may well be a high "noise to signal" ratio or outlying values which may well make relationship detection and estimation difficult.

The first step in dealing with multivariate data is usually to produce two-dimensional scatter-plots. Whilst this may be, and still often is, done manually a number of graphical software systems are available which provide visual displays

or hard-copy prints suitable for most purposes. A good graph plotter is now a fairly standard requirement for most scientific computer systems. This step will often produce sufficient information about any structure in the data to permit a formal analysis to be commenced.

If the plot shows a great deal of variability it may not be easy to detect any tendency for the two variates to be associated. A smoothing proceedure which the author has found very useful for revealing any association in such cases is one due to Cleveland (1979). This method operates as follows. Each y-value is initially smoothed by identifying its m nearest x-neighbours. Here, typically, $n/3 \leq m \leq n/2$. A weighted regression of y on x is fitted, using only the m points identified. The weight function used at this initial stage has its maximum value at the one point within the set to be smoothed and falls off symmetrically about this point. The residuals of the observed y-values from the fitted regression at each of the m points in the set are computed. New weights are formed which behave like the previous ones but in addition give progressively less and less weight to larger and larger residuals. A weighted regression, using these new weights, is then calculated. If desired, this stage may be repeated once or twice. The smoothed value sought is taken to be the fitted value from the final regression at the point in question. This proceedure is repeated for every point in the data set. An example of the sort of smoothed plot obtained by this method is shown in Figure 1. This shows the condition factor (weight/(length)^3) of a given species of fish plotted against the level of infection with a certain parasite in its liver. The downward trend in condition factor with increasing level of infection would not be obvious without the smoothing. Application of this smoothing proceedure is not feasible without the use of a computer.

A particularly interesting method for displaying multivariate data has been proposed by Andrews (1972). This method consists of mapping each multivariate observation into a function, f(t), of a single variable t. The function proposed by Andrews is a linear combination of orthonormal functions of t, the coefficients being the observed values of each variate. Specifically, the ith multivariate observation is mapped according to the relation

$$f_i(t) = x_{1i}/\sqrt{2} + x_{2i}\sin(t) + x_{3i}\cos(t) + x_{4i}\sin(2t) + x_{5i}\cos(2t) + \ldots\ldots\ldots$$

where i = 1, 2, ..., n (n being the number of multivariate observations) and x_1, x_2, ... denote the first, second, ... variate. The functions $f_1(t)$, $f_2(t)$, ... may then be plotted simultaneously against t. The values of these functions at different values of t correspond to different linear combinations of the variates.

Function plots are useful for detecting clusters or outliers in a set of multivariate data. If there are distinct clusters in the data this will be manifested by the functions separating into groups, curves corresponding to observations in the same cluster being closer together than curves corresponding to observations from different clusters.

The specific ordinal labelling of the variates may be important as, clearly, the function is not invariant to different orderings. Different permutations of the variates will, therefore, produce different plots and, in practice, it may prove useful to try several.

Functional plotting may also prove useful in detecting the existence of multicollinearity in multiple regression. Multicollinearity arises when there exist functional, or near functional, relationships between some of the dependent variates. Such a relationship will produce a linear combination with zero variance and the functional curves will all pass through one point.

4 Final Remarks

In a paper as brief as this it is impossible to cover all the methods currently available for exploring and analysing large data sets. Many useful techniques, such as probability plotting and adaptive estimation, are not mentioned. The author has simply picked out a few of his own particular favourites. Had this paper been written by someone else, quite a different, and possibly better, set of techniques might have been presented.

References

Andrews, D.F. 1972 Plots of high-dimensional data. Biometrics 28, 125-136.

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. and Tukey, J.W. 1972 Robust Estimates of Location. Survey and Advances. Princeton University Press, Princeton, N.J.

Anon. 1974 Report of the Working Group on Standardization of Scientific Methods for Comparing the Catching Performance of Different Fishing Gear. ICES Coop. Res. Rep. 38.

Anon. 1977 Report of the Working Group on Standarization of Scientific Methods for Comparing the Catching Performance of Different Fishing Gear. ICES Coop. Res. Rep. 66.

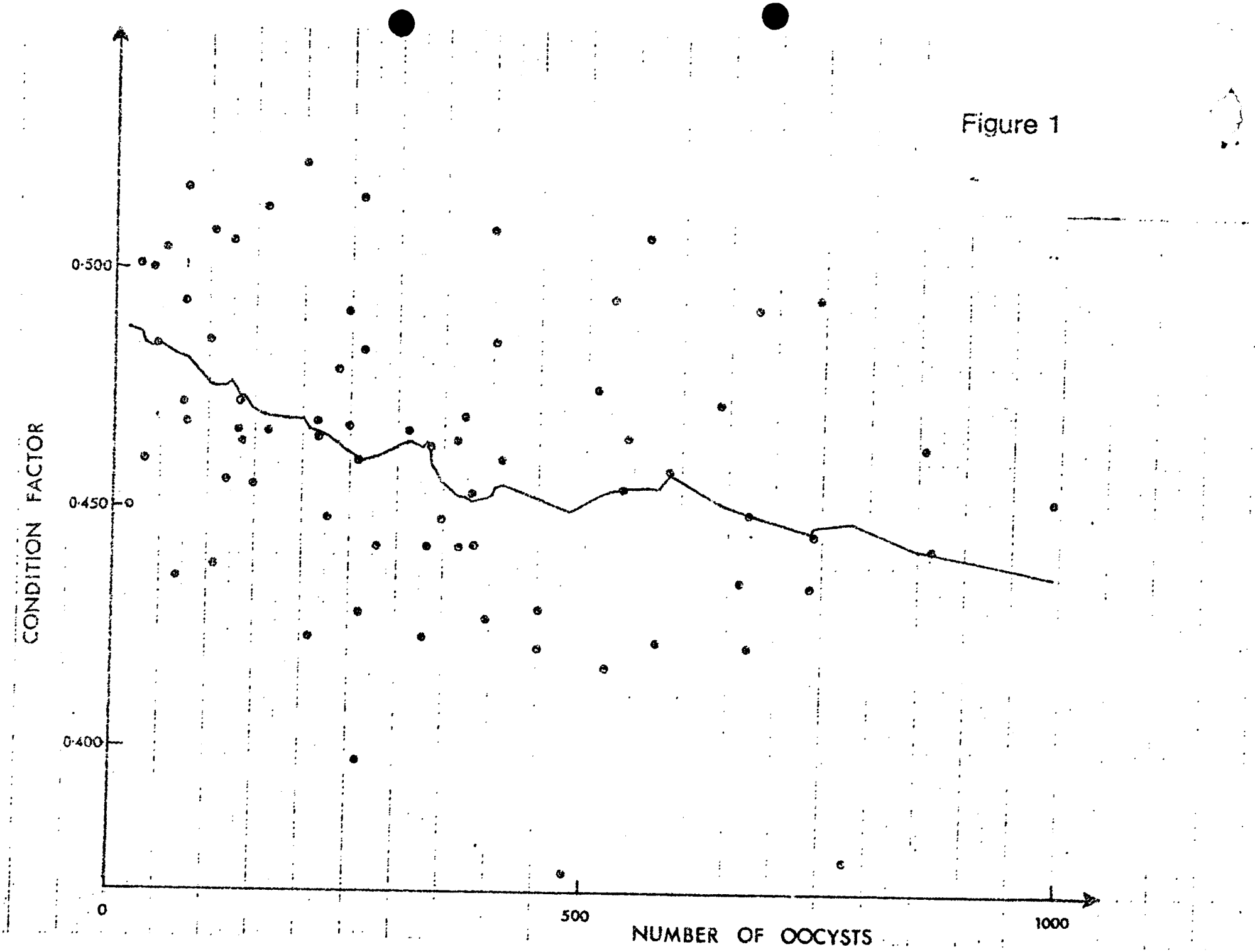
Cleveland, W.S. 1979 Robust locally weighted regression and smoothing scatterplots. Technometrics 74, 829 - 836.

Knuth, D.E. 1973 The Art of Computer Programming 3. Sorting and Searching. Addison-Wesley, Reading MA.

McGill, R., Tukey, J.W., and Larsen, W.A. 1978 Variations of box plots. Amer. Statistician. 32, 12 - 16.

Tukey, J.W. 1977 Exploratory Data Analysis. Addison-Wesley, Reading MA.

Figure 1


0·500
0·450
0·400
CONDITION FACTOR
0
500
1000
NUMBER OF OOCYSTS