Occasional article

Statistics on Microcomputers

A non-algebraic guide to their appropriate use in biomedical research and pathology laboratory practice. A series of six articles

1 Data handling and preliminary analysis

R A BROWN, J SWANSON BECK

From the Departments of Mathematical Sciences and Pathology, University of Dundee, Scotland

The recent developments in microcomputer technology, particularly the introduction of the IBM personal computer and the mass production of inexpensive "clones", have made available relatively powerful machines (such as the XT and AT models) within the budgets of most laboratories. In response to growing demand computer software has been written to perform a wide range of tasks, and very sophisticated programs are now available at modest prices. Consequently, most workers in pathology laboratories are likely to have access to computer power that could not have been imagined a decade ago. As a result the users of statistical techniques need no longer be concerned with the arithmetical and algebraic details of various statistical methods and can concentrate instead on understanding the underlying ideas and basic principles of statistical analysis. This series of articles has been written to show how microcomputers can be used to facilitate the analysis of data. It is our intention:

- (i) to explain the rational basis of widely applicable statistical methods without requiring the reader to understand the underlying algebra; and
- (ii) to summarise the limitations of the various statistical procedures so that the user can make an informed choice when analysing data.

We do not intend to present an extensive account of the design of clinical trials as the techniques and approaches applicable to this type of work have been recently discussed fairly exhaustively.^{1,2} We shall concentrate on statistical methods applicable to most laboratory investigations. Details of the algebraic formulae used for particular statistical methods may be obtained from standard texts.³⁻⁵

Accepted for publication 21 April 1988

Organisation of information in the microcomputer

In certain specialised applications data are transferred directly from analytical apparatus to a microcomputer via a connecting cable. Diverse observations made on each individual patient or specimen, however, will more commonly be keyed in manually before any statistical analysis can be done. The microcomputer can store data in files on a floppy or hard disc, depending on the configuration of the machine. The way in which files are stored need not concern us here as the disc operating system (DOS) takes care of this.

To put the data into the file system the user has three options: the data can be entered by using (a) a proprietary Spreadsheet program, (b) a proprietary Database Management System (DBMS), or (c) the data entry facilities provided within the statistical analysis package.

Conceptually, a spreadsheet is a table arranged in rows and columns, each row corresponding to an individual patient or sample, with named columns for the values of specific observations. This corresponds exactly with the table one might construct manually during data collection. The spreadsheet program makes it easy to insert the data into appropriate rows and columns and to correct errors. New columns may be defined in terms of existing ones and the program will automatically fill the new columns. The information put into a spreadsheet will be stored on disc as a data file which can be extended as additional results come in.

An alternative manual system of record keeping is to have a record card for each individual with labelled spaces for the data. This results in a stack of record cards rather than a table. The computer

implementation of this is a DBMS. The user designs a master form which contains labelled blank spaces for the data, and which is stored on disc. To facilitate data entry it is a good idea to make the screen format look like the corresponding manual record card. To store a series of records the DBMS presents the user with a fresh proforma on screen for each record: each item can then be inserted into the appropriate blank space. This method is preferable when a large number of records have to be filed; the use of the master as a prompt to the user greatly reduces the incidence of errors.

Good statistical packages have data entry facilities which usually operate on something like the spreadsheet principle so it is not essential to use one of the spreadsheet or DBMS programs, but many workers find such programs convenient. It must be emphasised, however, that individual data management systems may store data on disc in different formats. If a proprietary program is used then care must be taken to ensure that its format is compatible with that of the chosen statistical package. Good statistical packages can read a limited number of common formats (some important compatibilities are given in the appendix).

VARIABILITY IN RESULTS

Whatever characters are being observed in biomedical research or hospital practice some variability between observations will be apparent. Such variability in results may arise in a variety of ways. It will be partly methodological, due to imperfect function of the measuring instrument—for example, voltage fluctuations and electronic noise—or sample variation—for example, biological variation between successive histological sections. Biological variation due to differences between patients (in severity of disease, environmental conditions, or a host of other predictable or unpredictable factors) or experimental conditions (such as sampling inaccuracies, temperature fluctuations, differing diffusion conditions, etc), however, may give rise to even more variability. For some investigations the observed variability is small but in others it may be great and so statistical methods may be required to allow objective assessment of differences or associations within the study. By appropriate use of statistical methods it is often possible to draw conclusions from observations made in a well designed investigation that are applicable generally to clinical or pathological practice in other centres or throughout the world.

TYPES OF INVESTIGATION

Statisticians distinguish between experimental investigations and observational studies, and the distinction is important as there are more problems with the

interpretation of the results of statistical analysis applied to the latter.

Experimental investigations aim to compare two or more *treatments* (such as the administration of various drugs), applied to groups of subjects, or to investigate the consequences of deliberate variation of *factors* which are controlled by the investigator. A typical example of the study of this type would be the investigation of the effect of storage temperature on the recovery of bone marrow cells for transplantation. The sample of bone marrow was split into aliquots each of which was stored at one of a number of predetermined temperatures and viability of the cells measured at the end of the in vitro storage. The aliquots to be stored at each temperature were chosen at random, and identical methods were used to measure viability in all samples so that the effect of other variables was minimised.

In this way the investigator has complete freedom to decide which experimental subject should be assigned to each treatment. There will inevitably be unsuspected differences between people even within a supposedly homogeneous group, and to ensure that these are evened out and so do not bias observed mean responses it is essential to allocate patients randomly to the treatment groups. This is known as *randomisation* and the process ensures the validity of the inference that observed differences between groups are caused by the different treatment they have received.

In summary, an experimental investigation entails:

- (i) *replication*, several patients allocated to each group;
- (ii) allocation under the control of the investigator;
- (iii) *randomisation*.

In contrast, an *observational study* usually lacks the second and third elements listed above. For example, an observational study of the prognostic value of microscopical and other features of primary tumours in breast cancer might proceed by extracting all available case records and specimens. Measurements or gradings of potentially important factors, such as tumour size, and of other possible prognostic indicators such as extent of lymph node disease would be recorded. The investigator might then divide the cases into two groups—poor and good survival—and look for some subset of the recorded features, a discriminating subset, which differs substantially between the two groups. If this is possible then some criterion possibly determined by combining values of several features in the discriminating subset to yield a score for each patient can be selected so as to give a predictor of survival in new cases. Thus the composition of the groups is not controlled by the investigator; it is clearly impossible to allocate patients randomly to the poor and good survival groups. Consequently, one must be very cautious in ascribing the observed differences between the prognostic features to the

*STATGRAPHICS Statistical Graphics System***DATA MANAGEMENT AND SYSTEM UTILITIES**

- A. Data Management
- B. System Environment
- C. Report Writer and Graphics Replay
- D. Plotter Interface

PLOTTING AND DESCRIPTIVE STATISTICS

- E. Plotting Functions
- F. Descriptive Methods
- G. Estimation and Testing
- H. Distribution Functions
- I. Exploratory Data Analysis

ANOVA AND REGRESSION ANALYSIS

- J. Analysis of Variance
- K. Regression Analysis

TIME SERIES PROCEDURES

- L. Forecasting
- M. Quality Control
- N. Smoothing
- O. Time Series Analysis

ADVANCED PROCEDURES

- P. Categorical Data Analysis
- Q. Multivariate Methods
- R. Nonparametric Methods
- S. Sampling
- T. Experimental Design

MATHEMATICAL AND USER PROCEDURES

- U. Mathematical Functions
 - V. Supplementary Operations
- Use cursor keys to highlight desired section.
Then press ENTER.

Fig 1 *The Statgraphics main menu allows the user to choose from a wide range of options.*

overt difference between survival experience. There could well be some unsuspected factors which were seriously unbalanced between the two groups which could equally be the cause of the observed differences. It is therefore essential to verify the results of potentially important observational studies by accumulation of further data in follow up studies in which the discriminating criteria are applied to new cases.

Statistical packages for microcomputers

There are several good quality statistical packages available which, when used in a knowledgeable way, allow the experimenter to use many more methods than is practical using traditional paper calculations. This series of articles will be illustrated by reference to the working of two well known packages—Minitab and Statgraphics. These packages have a good reputation within the field of statistics and have been evaluated by bodies such as the American Statistical Association, which runs a continuous program of package review and evaluation. Both have clearly written, unambiguous manuals. There are other less expensive packages but we have not used them; the principles that we explain in this series should be generally applicable.

The data can be entered directly into the statistical package from the keyboard, or indirectly from a spreadsheet or database. Thereafter the statistical package is driven by the operator selecting procedures from a menu (fig 1) or by a system of brief but memorable commands according to the design of the package or the convenience of the user. At various points in certain analyses the user is allowed to select

from a range of options to see plots or perform supplementary analyses. While all of the common methods of display, summary, and analysis are literally at the user's fingertips, current statistical packages, being essentially "unintelligent", are incapable of advising on the most appropriate method for a particular data set and this remains the prerogative of the user.

The statistical packages arrange the data in spreadsheet form, so it is important to identify variables unambiguously and to identify missing values. The conventions of the package must be rigidly adhered to if malfunction is to be avoided.

TYPES OF DATA

Data obtained by observation or measurement in the laboratory or clinic can be of three different types. These must be distinguished clearly as different methods of analysis are required for the various types of data.

Nominal data The observations are classified into categories that are different in character and cannot be measured or ordered—for example, hair colour or race—consequently, the named character can be only present (+) or absent (−).

Ordinal data Here the observations refer to a common character and they can be grouped into a limited number of categories which can be ordered in an ascending series or rank, such as staging of cancer: commonly such data are presented as +, ++, or +++, but numbers can also be used to describe the categories so that there is a danger of confusion of this form of ranking with true measurements. It is essential to remember that in ordinal data the grouping is only semiquantitative and that + is not half as big as ++:

moreover, the intervals between successive categories will probably be different.

Continuous data This is the term used to describe observations made on a scale which has a natural zero and a well defined unit of measurement, such as length: the observations will usually be expressed as decimal numbers.

Both nominal and ordinal data give rise to counts of the incidence of occurrence of the various categories within the groups of subjects under investigation and these counts are naturally whole numbers. They may be converted into rates, proportions, or percentages. When expressed in percentages they appear to be decimal numbers but they are essentially different in character from continuous data and need appropriate statistical treatment. (The first three articles will deal with the analysis of continuous data; the special methods for other data types will be discussed in the subsequent papers).

Data displays and statistical summaries

It is important that any conclusion to be drawn from an experiment is intuitively accepted as "common sense" and that statistical methods are used to facilitate objective assessment of results. It is therefore essential that the experimenter examines the "raw" data from every experiment. Graphical displays have an immediate visual impact and a compactness which is always absent from numerical tables. Differences between groups, variability within groups, time trends or other forms of associations between sets of measurements can be investigated and communicated by simple well designed graphical displays. Such displays are a valuable feature of all good statistical packages—Statgraphics is particularly good in this respect. There are no rigid rules for the selection of the most appropriate method of display to show the magnitude and variability of the observations in any particular experiment. The choice is guided by informed common sense, but the chosen format should, in the first instance, present all the original observations (not just condensed summaries) to allow the observer to "eyeball" the results and judge the plausibility of any conclusions presented.

A basic assumption of several widely used statistical methods is that the groups of results being compared have similar variability and this can often best be assessed by visual inspection of their distributions. Such preliminary examination of the data can usually be performed conveniently on the visual display screen of the microcomputer. When the data sets are large some summarised form of display may be needed because of the confusion of overlapping points on the screen, but, whatever the method, it must present a fair and unbiased view of the variability of the data.

Fig 2 Various displays of transcutaneous measurements of $p\text{CO}_2$ in forearm skin of normal subjects. (a) A dot plot of the data. (b) The interquartile range spans the central 50% of the data. (c) The box and whisker plot suppresses details but indicates spread and symmetry.

The dot plot All the observations are displayed by symbols in a vertical array alongside an appropriate scale (fig 2a). This plot gives a visual indication of the magnitude and spread of each set of values, but comparison of sets of values is more difficult than it appears: it is therefore useful to supplement the display with indicators of the group means at least.

The histogram is an alternative display that is useful when each group consists of a large number of measurements (100 or more). The interval of the measurement scale which contains all of the data is divided into equal subintervals and the number of values falling in each subinterval is represented by a rectangle with height proportional to this number. Care is needed in selection of the subintervals as too many will reduce the number of values in each one, emphasising minor fluctuations, too few will obscure the variability. The Statgraphics package is capable of generating histogram displays automatically and the user can override the inbuilt criteria for selection of interval size.

The fact that the individual observations are not shown and that the number of intervals is chosen subjectively makes this a poor choice for final presentation of results. This forms the basis of the policy of the *Journal of Clinical Pathology* not to publish data in histogram format.

SUMMARY STATISTICS TO INDICATE LOCATION, SPREAD AND SYMMETRY

There are a few simple statistics which can convey a lot of information about a set of observations. Most standard statistical methods for analysing two or more sets of measurements assume that:

(i) each set has more or less the same variability; (ii) the distribution of values around the mean is more or less symmetrical (for "parametric" analyses), or has a

similar shape in each group; (iii) there are no outliers (values far removed from the main body) in the data. **Magnitude** An indication of the “middle” of the data gives useful information when attempting to determine whether groups of data are similar or different. Two statistics commonly used for this purpose are the *mean* (arithmetical average) and the *median* (middle value when the observations are ranked in order). Either value will be equally informative when the distribution is symmetrical but the mean forms the basis of certain standard methods for assessing differences between groups. If the mean and the median are quite different this indicates serious asymmetry, and methods of statistical analysis based on medians and on the rank order of data values will be necessary (details are given in subsequent articles). Outliers do not affect the median but they may make the mean unrepresentative.

Spread Again there is a choice of summary statistics. The *standard deviation* (SD) is widely used, but it is valid only when the data are more or less symmetrical: it is the square root of the average of the squared deviations of the data values from the mean. An alternative is the *interquartile range* (fig 2b): when the data are ordered the upper and lower quartiles are the values that separate off the upper and lower 25% of values; thus the interquartile range measures the spread of the middle 50% of the data and this is valid whether or not the data are symmetrical. The *range*—the difference between the largest and smallest values—is much less informative because it is so sensitive to outliers. All these summary statistics are readily generated in microcomputer packages.

Symmetry This can be assessed from a “box and whisker plot” which indicates graphically the median, quartiles, and range of a set of data and can easily be

Fig 3 (a) Box and whisker plot of volumes of 300 lymphocytes. The distribution is asymmetrical: the package calculates the length of the upper whisker as one and a half times the interquartile range (length of rectangle) and the outlying points are plotted individually. (b) Box and whisker plot of the logarithms of the same data: the transformation had made the distribution symmetrical.

generated in both packages (fig 2c). The central 50% of the data is represented by a rectangle extending from the lower to the upper quartile, the median being indicated by a bar across the rectangle. The “whiskers” extend to the extremes unless there are values which are far away from the central 50% (outliers), in which case the length of the corresponding whisker is set at one and a half times the interquartile range and the outlying points are plotted individually.

Sometimes asymmetrical data sets can be converted into more or less symmetrical form by “*transformation*” of the raw results by taking logarithms, square roots, or other simple mathematical functions (fig 3), and such procedures can be performed automatically with most packages. Care should always be taken if any values are less than zero as logarithms or square roots of negative numbers do not exist! Most packages will insert a “missing value code” in place of an impossible value when transforming. Adding a suitable number to all observations to remove negative numbers is also a valid option. Transformation of data is a valuable approach; if it is successful it permits the use of statistical methods which require symmetrical distributions on data which are clearly asymmetrical in their raw form. Care is of course necessary in interpretation of analyses performed on transformed data.

DISPLAYS TO BE AVOIDED

One of the commonest forms of data presentation seen in articles published in biomedical journals is the “mean and error bar” plot. Here the original data are not shown, but the mean of each set is plotted with two equal bars above and below (fig 4). This type of plot has little to recommend it: it is misleading in several ways.

Firstly, the number of observations in each category is not clearly apparent and so disparity between size of groups being compared or the presence of a very small

Fig 4 Transcutaneous measurements of $p\text{CO}_2$. The mean and error bar plot (larger bars) understates the variability. Using the SEM (smaller bars) gives a very misleading impression of variability.

group that may not be representative can be obscured.

Secondly, the presentation of the equal sized error bars above and below the mean can imply that the data are symmetrically distributed, but the reader cannot assess this for him/herself.

Thirdly, the basis of the calculation of the length of the error bars may be misleading as the indicated range will contain only about 70% of the observations if the standard deviation is chosen. Choice of the standard error of the mean (SEM)—SD divided by square root of number of observations—to indicate variability is even more misleading as the real extent of variability will be seriously understated, particularly for large samples.

OUTLIERS

Occasional values that are considerably larger or smaller than the main group are termed *outliers*. Such observations are commoner than is popularly supposed, but they will be unlikely to be appreciated unless the data are displayed using either the dot plot or box and whisker plot format. Sometimes an outlier is an artefact due to an error in transcription or instrument reading, and sometimes it has resulted from inclusion of a result that is inappropriate—for example, erroneous selection of patient or misclassification of observation. It is good practice to check the validity of outlier observations, but each observation must be accepted unless there is clear evidence that it is artefactual.

Most of the standard methods that use means and standard deviations are adversely affected by the presence of outliers. Methods which use medians, quartiles, and ranks are generally much less vulnerable to distortion by outliers. It is not possible to give guidelines on how to deal with outliers. Perhaps the safest policy is to be constantly aware of the potential hazards and when the outliers are encountered to analyse the data with and without the inclusion of the suspect values. When the results remain substantially unchanged, then one is reassured that the conclusions drawn from the statistical analysis are not unduly influenced by the outlier values. When the analyses disagree, however, then it is obvious that the outlier values are influencing the analysis. In this situation advice should be sought from a professional statistician who may advise the use of the more valid method of analysis with statistical procedures that are less influenced by outliers.

Appendix

MISSING VALUES

If certain observations are not available then this must be indicated during data entry by using a "missing value code" to represent them. In Statgraphics a missing value is indicated by leaving a blank in the corresponding position in the worksheet; in Minitab an asterisk is used instead of a blank.

When the experimenter is aware that there are missing data the manual for the package should be consulted to verify that there is some method of specifying missing values otherwise every analysis will have to be preceded by a data editing session to weed out the cases with missing values—a very tedious business.

IMPORTING SPREADSHEET AND DATABASE FILES

Importing is the name given to the process for transforming data files from proprietary spreadsheets and DBMS into a statistical package. Statgraphics will cope with files from Lotus 1-2-3, Symphony and dBase III. Minitab can read files from Lotus 1-2-3 alone, but certain "Lotus-clones" may be acceptable.

COMMAND LINES

The Minitab package uses commands in preference to a menu system; the user specifies the task to be done by typing a command line. Minitab uses command lines which are reasonably memorable: for example, the sequence of commands

```
NAME C1 "height"
NAME C2 "weight"
DOTPLOT "height"
PLOT "height" v "weight"
```

names the first and second columns of the worksheet "height" and "weight", produces a dotplot of "height" and a plot of "height" against "weight".

References

- 1 Pocock SJ. *Clinical trials: a practical approach*. New York: John Wiley and Sons, 1983.
- 2 Gore S, Altman DG. *Statistics in practice*. London: British Medical Association, 1982.
- 3 Armitage P. *Statistical methods in medical research*. Oxford: Blackwell Scientific, 1971.
- 4 Bailey NTJ. *Statistical methods in biology*. London: Hodder and Stoughton, 1981.
- 5 Ingelfinger JA, Mosteller F, Thibodeau LA, Ware JH. *Biostatistics in clinical medicine* 2nd ed. New York: Macmillan, 1987.

Requests for reprints to: Professor J Swanson Beck, Department of Pathology, Ninewells Hospital and Medical School, PO Box 120, Dundee DD1 9SY, Scotland.