Egyptian Journal of Surgery Vol. (22), No. (3), July 2003

PROFESSIONAL DEVELOPMENT

Handling Surgical Data – Exploratory Data Analysis

By
Egyptian Group for Surgical Science and Research
Said Rateb, EGSSR Moderator
Nabil Dowidar, EGSSR Secretary General
Mohamed Farid
Ahmed Hussein
Ahmed Hazem

The process of data analysis should be thought of as taking place in two phases:

1. Exploration and description of the data
2. Confirmatory statistical analysis (will be covered in the next issue of the EJS)

EXPLORATORY DATA ANALYSIS

The need to explore data

To proceed immediately to the statistical analysis of your surgical data is a decidedly risky practice. There are two main reasons for this caution. First, to proceed immediately to statistical analysis may make you miss the most illuminating features of your data or miss the presence of unlogical data indicating errors in measurement or data recording. Second, the performance of a statistical test always presupposes that certain assumptions about the data are correct (e.g. data show a normal distribution). Should these assumptions be false, the results of statistical tests may be misleading.

Types of data

The first step before any exploratory data analysis is performed, is to decide what type of data one is dealing with. A useful typology is given in (Table 1). The basic distinction is between quantitative variables "how much?" and categorical variables "what type?"

Table (1): *Types of Data*

Quantitative	
Continuous	**Discrete**
Blood pressure, temperature, weight, age	Number of patients, number of attacks
Categorical	
Ordinal	**Nominal**
Grade of disease, degree of pain	Sex, alive or dead, blood group

Quantitative variables can be continuous or discrete. Continuous variables, such as weight, can in theory take any value within a given range. Examples of discrete variables are: number of patients in a ward, number of attacks of appendicitis per week.

Categorical variables are either nominal (unordered) or ordinal (ordered). Examples of nominal variables are male/female, alive or dead, blood group O, A, B, AB. For nominal variables with more than two categories the order does not matter. For example, one cannot say that people in blood group B lie between those in A and those in AB. However, with ordinal data

such as grade of disease and degree of pain “mild”, “moderate”, or “severe” the order does matter and it is usually important to account for it.

Variables can be converted from one type to the other by using "cut off points". For example, blood pressure can be turned into a nominal variable by defining "hypertension" as a diastolic blood pressure greater than 90 mmHg, and "normotension" as blood pressure less than or equal to 90 mmHg. Weight (continuous) can be converted into "normal", “overweight”, or "obese" (ordinal).

In general it is easier to summarise categorical variables, and so quantitative variables are often converted to categorical ones for descriptive purposes. To make a clinical decision on someone, one does not need to know the exact serum potassium level (continuous) but whether it is within the normal range (nominal). However, categorising a continuous variable reduces the amount of information available and reduces the sensitivity (power) of the statistical tests used with such type of data. Categorising data is therefore useful for summarising results, but not for statistical analysis.

DATA SUMMARY

Frequencies

Suitable for categorical data and can be presented in a tabular fashion. Data can also be divided into subgroups.

Measures of Central Tendency

There are three methods of describing the central tendency of a group of data. These measures give us an idea where the core of the data lies.

The mean

The mean is the arithmetic average of a set of numbers. This is simply the sum of the values divided by the number of values. It represents the centre value of the data. When the number of observations in the data set is small, the mean becomes very sensitive to extreme values. For this reason, the mean is misleading when applied to unsymmetric data.

$$\text{Mean, } \bar{x} = \frac{\sum x}{n}$$

The median

When a sample of observations is arranged in order of magnitude, the median is the middle value (for an odd number of observations) or the average of the two middle values (for an even number of observations). The median is useful for summarizing unsymmetric data because it is insensitive to extreme values.

$$\text{Median } = \frac{(n+1)}{2} \text{ th value of ordered observations}$$

NB Symmetric distribution is one in which frequencies are equal at points on either side of the central value. For symmetric data, the median and mean have the same values.

The mode

The mode is simply the value that occurs most often in any set of data. It is seldom used.

MEASURES OF DISPERSION

As the name suggests, measures of dispersion indicate the width of spread of the values in a particular set of data. Taken together with one of the measures of central tendency already described, a useful description of the data distribution is obtained. Dispersion is also the key to comparison of results between samples.

The range

The range is the difference between the largest and the smallest values in a set of data.

The standard deviation and the variance

The standard deviation is a number that indicates how much, on average, each of the values in a sample differs from the mean of the sample. The variance is the average of the squared deviation from the mean.

$$\text{Standard deviation, } s = \sqrt{\left[\frac{\sum (x-\bar{x})^2}{(n-1)}\right]} \qquad \text{Variance, } s^2 = \frac{\sum (x-\bar{x})^2}{(n-1)}$$

The confidence interval

The confidence interval (CI) can be defined as a range of values that is likely to cover the true population. For instance, with the 95% CI, if the study was repeated 100 times, the confidence interval would be expected to include the true value on roughly 95 occasions.

DATA DISPLAY

Graphical displays of data have an immediate visual impact that is always absent from numerical tables.

Bar chart

This graph is suitable for categorical data (Fig. 1). In a bar chart, the bars are separated to clarify the fact that the horizontal axis contains no scale of measurement.

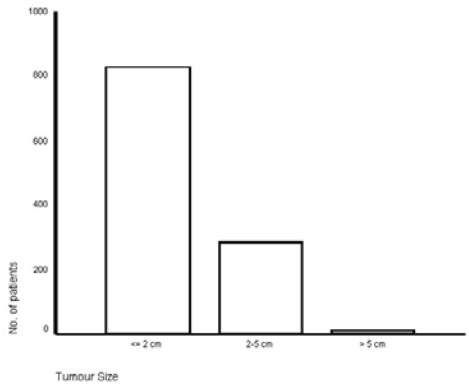

Fig (1): ***Size of tumour of a group of patients with breast cancer1***

Histogram

A histogram, on the other hand is appropriate for continuous data (Fig. 2). A histogram can also be used to see whether the data has a normal distribution (bell shape) or not. In a histogram, in contradistinction to a bar chart, the bars touch each one another: there are no spaces.

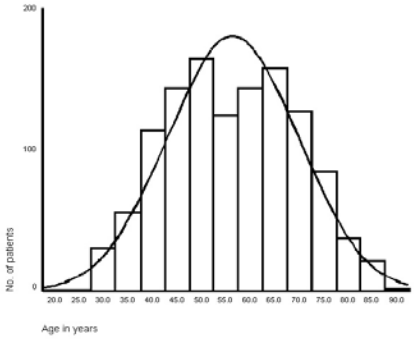

Fig. (2): ***Age distribution in a group of cancer patients***.

Dot Plot

The simplest way to show data is a dot plot (Fig. 3). All observations are displayed by symbols or dots in a vertical array alongside an appropriate scale with a transverse line indicating the location of the median. This plot gives a visual indication of the central tendency and dispersion of the data.

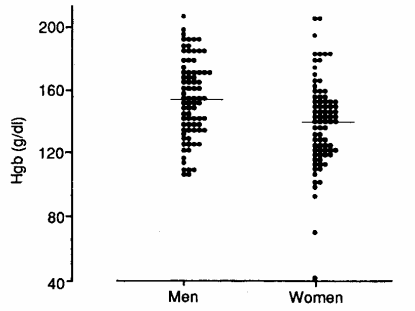

Fig. (3): ***Haemoglobin concentration in men and women.***

Box-Whisker Plot

When the data sets are large, plotting individual points can be cumbersome. An alternative is a box-whisker plot (Fig. 4). The box is marked by the first, third quartile and the median, and the whiskers extend to the range 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 (box length), and the outlying data are plotted individually as dots.

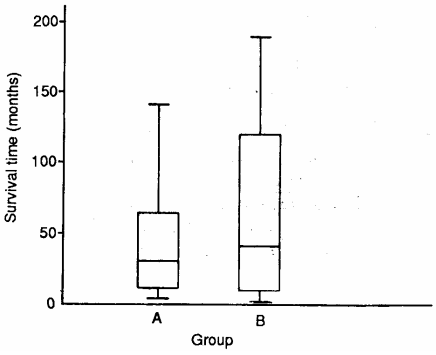

Fig.(4)*: Survival of two groups of patients with cancer*.

DATA INTERPRETATION

The normal distribution

Many standard statistical analyses of continuous data are based on the assumption that the spread of values across the population can be described by the so called "normal distribution" (Fig. 5). This form of distribution was originally selected as the basis of much statistical work because it is unimodal and symmetrical, with a high proportion of values relatively close to the mean and decreasing concentration of values in the "tails" of the distribution, similar to many distributions observed in practice. As normality is an important assumption for several statistical methods it is important to be able to assess whether a particular set of data shows evidence of serious non-normality.

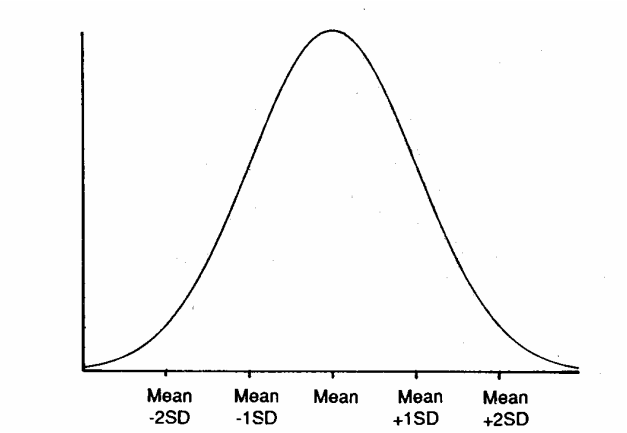

Fig. (5). Normal distribution of a set of data showing a symmetrical and unimodal character.

Symmetry

This can be assessed from a box-whisker plot. Assymetrical data can be converted into a more or less symmetrical form by "transformation" of the raw results by taking logarithms, square roots, or other simple mathematical functions (Fig. 6). Care should be taken if any values are less than zero because logarithms or square roots of negative numbers do not exist. However, this can be overcome by adding a suitable number to all observations to remove negative numbers. Transformation of data is a useful approach as it permits the use of statistical methods that require symmetrical distributions on data that are clearly asymmetrical in their raw form.

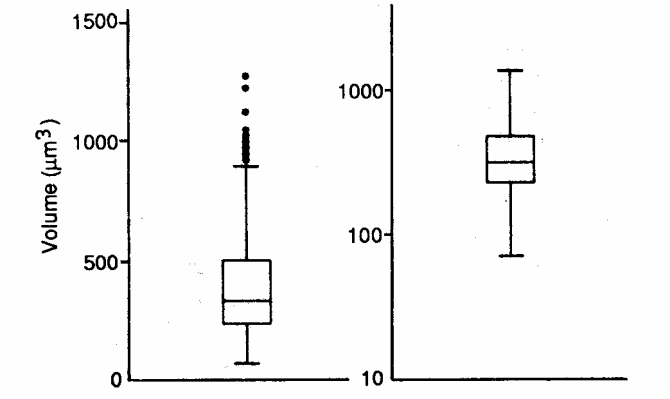

Fig. (6): Box and whisker plot showing on the left an asymmetrical distribution of a set of data with several outlier values. On the right is the plot of the logarithms of the same set of data showing a symmetrical distribution with no outlier values.

Outliers

Occasional values that are considerably larger or smaller than the main group are termed outliers (Fig. 6). Such observations are commoner than is popularly supposed, but they are unlikely to be appreciated unless the data are displayed using either a dot or box and whisker plot format. Sometimes an outlier is an artefact due to an error in transcription or instrument reading, sometimes it is the result of inclusion of a result that is inappropriate. 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 uses 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. Perhaps, the safest way to deal with outliers is to analyses the data with and without the inclusion of the suspect values. When the results remain substantially unchanged, then you can be reassured that the conclusions drawn from the statistical analysis have not been unduly influenced by the outlier values. When the analysis disagree, it is then obvious that the outlier values are influencing the analysis and at this point it is advisable to consult a professional statistician.