Preparing to analyse data

Manikandan S

Assistant Editor, JPP

Understanding biostatistics is important for a biomedical researcher because modern day research methodology is based on the principles of statistics. Determination of sample size, sampling techniques, bias eliminating techniques such as randomization, summarizing data, interpretation of data and many other procedures followed in research use statistical methods so that the predictions or conclusions drawn at the end of a study are valid (validity) and can be applied to the population (generalization) from which the sample is drawn. Not only students but also many budding biomedical researchers develop a dislike for statistics which stems from blindly following the steps of statistical methods for data analysis without understanding the principles behind the statistical procedures. The application of statistical methods is considered by many a ritual at the end of the study rather than an important aspect that has a role in every step of research from planning till publication.

This section hosts a series of articles that focus on the understanding of statistics as a research tool. It deals with the need for statistics in biomedical research, common statistical methods used, their applications and limitations. Once data collection is over, the next logical step the researcher undertakes is to analyze them to draw meaningful conclusions. Before the actual statistical analysis is carried out, one has to look hard at the data and prepare it to facilitate further analysis.

DATA CHECKING

During data collection, data are entered into spread sheets from which it can be imported into many statistical software packages for analysis. Errors can creep into the master chart while entering the data. So a logical check should initially be carried out to spot the errors, i.e. blood group C, age - 260 years, pregnant male, etc. Even though there are no shortcuts for checking, many spread sheet programs provide facilities to prevent logical errors at the time of entry. Microsoft Excel has a feature called 'data validation' which can be used for logical checks. An error alert can also be invoked. For example if the valid range for a variable is 1-5, any values outside this range will not be accepted by the spread sheet.

DERIVED DATA

Some variables of interest to us cannot be measured directly. They have to be calculated from other measured variables. Examples include calculation of body mass index / body surface area from the height and weight of subjects. Before starting the analysis, these parameters can be derived from the original variables using the formulas. One can automate the calculations using 'Insert function (fx)' command in the spread sheets.

If the variable is measured at baseline and after an intervention in two independent groups, then we might need to calculate the change or percentage change in the variable in the two groups for analysis. A classical example would be in an experiment / trial to evaluate the antidiabetic efficacy of a new drug, where one would have collected the blood glucose values before and after treatment in both control and new (experimental) drug group. The analysis is actually carried out by comparing the difference in blood glucose levels (before-after) in one group with that calculated for the other group. The most appropriate way is to compare the percentage change in blood glucose in both the groups and the percentage change is an example of derived data that is used for analysis.

If the values of a dependent variable are affected by any factor other than the independent variable, then every effort should be made to derive a variable by which the influence of the third factor is compromised. A study is conducted to find out whether nutritional status alters the serum levels of phenytoin in epileptics. Serum phenytoin levels are measured

Address for correspondence:
Manikandan S, Department of Pharmacology, Jawaharlal Institute of Postgraduate Medical Education and Research (JIPMER),
Pondicherry -605006, India. Email: drsmanikandan001@gmail.com

DOI:10.4103/0976-500X.64540

in a group (*n* = 50 in each group) of well nourished and malnourished epileptic patients. Since the dose of phenytoin taken by each patient might be different and the difference in dose can also contribute to the difference in serum phenytoin levels, plasma drug level per unit dose should be calculated and used for statistical analysis.

In a nutshell, derived data needs to be used for statistical analysis if that is the most appropriate measure that can answer the research question or address the objectives.

OUTLIERS

Individual observations that fall well outside the overall pattern of the data are called outliers or outriders. Outliers can be substantially smaller or greater than the other values in a data set. By looking at the measures of central tendency (mean, median) and range, one can get a general idea about the presence of outliers in a given data set. Outliers can also arise due to experimental / observational error or if the observation was made from a population other than from which the rest of data was collected.

A simple outlier can have a dramatic effect on some measures of central tendency and dispersion (eg. mean, range). It can also drastically alter the statistical analysis and its conclusion. So care should be taken to decide whether to reject the extremes of observations and consider only the data which agree with each other or accept it as it might represent the normal biological variability.

There are statistically accepted methods for rejecting the outliers. This can be done manually using formulas or by software. These methods calculate the probability of the extreme values (outliers) occurring by chance. If the calculated probability is high, we can reject it. There is also a limit for the number of observations that can be rejected as outliers. As a rule of thumb, not more than 5% of the total number of observations should be rejected.

DROP OUTS

One of the major problems in the conduct of clinical trials is drop outs. The decision whether to include or exclude drop outs from the analysis should be considered initially, at the start of the trial. This depends on the type of analysis. There are two types of analysis – intention to treat (ITT) and per protocol (PP).[1] In ITT analysis, drop outs are considered as treatment failures and included in the analysis.[2,3] But in PP analysis, only those who have completed the treatment schedule according to the protocol of the trial will be considered for analysis. The type of analysis depends upon the purpose of the clinical trial, which is decided *a priori*. ITT analysis is best suited for pragmatic trials of effectiveness (effect of treatment given to everyone), whereas PP analysis is done in explanatory trials for testing the efficacy (effect of treatment in those who completed the study protocol).

REFERENCES

1. Peduzzi P, Henderson W, Hartigan P, Lavori P. Analysis of randomized controlled trials. Epidemiol Rev 2002;24:26-38.
2. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ 1999;319:670-4.
3. Fergusson D, Aaron SD, Guyatt G, Hebert P. Post-randomisation exclusions: The intention to treat principle and excluding patients from analysis. BMJ 2002;325:652-4

Source of Support: Nil, **Conflict of Interest:** None declared