ABC OF STATISTICS

Statistics in Medical Research

Dinesh Kumar

This article is third in the series on 'Statistics in Medical Research' and deals with Presentation of Data.

Not everyone is a statistician, but everyone, almost without a choice, must be a consumer of statistics. In either case, knowledge of the basic methods of manipulating data is important to intelligent behavior in a quantitative world.

Presentation of Data : Researchers are collectors of numbers and they seek out ways to put order into the masses of numbers they collect. Raw data emanating from any survey or study is usually in the form of large set of unorganized numerical values. One of the most fundamental techniques for putting order into a disarray of data is the frequency distribution. Basically it is a systematic process of arranging individual observations from least to most in relation to some quantifiable characteristic. These numerical values can be qualitative or quantitative or a mixture of both in any research project. It serves two purposes :

1) Put data in order so that visual analysis can be made of results of measurements.
2) Provide a convenient structure for simple computations.

Frequencies : Summarizing qualitative data and presenting it is very straightforward. The main task is to count the number of observations in each category. These counts are called frequencies. They can also be presented as percentages of total number of individual units as relative frequencies. Simple bar diagram or one of its variants (discussed later) or a pie chart is used commonly to illustrate these.

Frequency Distribution : If the number of observations is more (say more than 20), it is useful to present quantitative data in the form of frequency distribution.

The presentation of data can take one of the two following forms

A) **Tabular Presentation :** is undertaken to organize data into compact and readily comprehensible form. A table may show one or more than one variable.

Single Variable tables e.g., a table showing frequency of one variable like age, sex, religion, duration of private practice etc.

Cross-Tabulation

Tables based on two variables (Two-dimensional tables)

Tables based on three variables (Three-dimensional tables)

B) **Diagrammatic Presentation :** As impressed upon earlier in the series, the nature of variables[1] under study decides which diagram appropriately reflects the data.

For qualitative variables it is either a

Simple Bar diagram if the frequency of only one variable is to be shown OR

Multiple Bar Diagram if the frequency of more than one variable is to be shown OR

Component Bar diagram when the variable in question can be partitioned into component parts

For Quantitative data it is either a

Histogram or Frequency polygon representing frequency distribution of one data set OR

Frequency polygon representing frequency distribution of two data sets OR

Line Diagrams representing frequency distribution when time trends are to be shown.

From the Postgraduate Department of Preventive & Social Medicine, Govt. Medical College, Jammu 180001 (J&K) India.
Correspondence to : Dr. Dinesh Kumar, Editor Statistics, JK Science, Journal of Medical Education and Research.

Tabel 1 : Appropriate use of tabular data presentation

Data situation	Tabular method
Small data sets of say less than 20 in number	Ordered array
Individual observations, many in number, involving only one variable	Frequency table
Individual observations involving two or more variables	Cross-tabulation

Table 2 : Diagrammatic methods appropriate to Tabular Data

Tabular data	Diagrammatic method
Frequency table, quantitative variable, One set of data	Histogram or frequency polygon
Frequency table, quantitative variable, two sets of data	Frequency polygon
Frequency table, categorical data	Bar or pie chart

Before different methods of presentation of data are discussed, few general points need to be emphasized. They are :

a) It is important to start by using basic summary.
b) The analysis should then progress from simple to complex.
c) The method chosen should be simplest consistent with the requirement of the data.
d) Statistical reasoning should be applied hand in hand with common sense.
e) Wherever possible frequency distribution should be given in full.
f) All diagrams and tables should be clearly labeled and self-explanatory i.e., there should not be any need to refer to text to understand what the table/ diagram means. Lack of sufficient heading might limit the usefulness of the table/diagram e.g., age as heading should be in completed months or years i.e., the unit of measurement should be included. At the same time, it should not be cluttered with too many details.
g) In case of rates, ratios and proportions, the denominator should be clearly stated.
h) Graphical techniques though are strongly recommended for presentation of results since relationships, trends and comparisons are often more readily appreciated from a diagram than a table. BUT
i) Graphs should always be regarded as subsidiary aid to the tables. It follows that graphs are an unsatisfactory substitutes for tables. Without basic data the reader cannot adequately infer the validity of author's deductions and as such it is not always possible to translate the diagram back to figures accurately.
j) Breaks and discontinuities in the scales (X-axis and Y-axis) should be clearly marked.
k) Horizontal and vertical scales should start from zero point and should neither be compressed nor inflated.

Steps in tabular presentation

1) Count the total number of observations in raw data.
2) Identify the lowest and highest values
3) Decide how many grouping intervals to be used.

The decision regard step 3 is arbitrary. Most researchers consider 10-20 groups as adequate. If the grouping interval chosen is too wide, much detail is lost while if it is too narrow the table is as good as presenting individual observations. Also too many groups are not desirable as they fail to reveal salient features. As a general rule intervals should be drawn on smaller units to start with. It is easier to combine smaller units into bigger without surfing the whole data. On the other hand, if the original grouping is too broad, the subdivision is impossible without re-tabulating much of the material. It needs to be emphasized further that all grouping intervals should be of the same size with no gap between them thus retaining the quantitative nature of the data.

Suggested Reading

1. Kumar D. Statistics in Medical Research. *JK science* 2006 ; 8(1) : 62-63
2. Lwanga SK, Tye Cho-Yook, Ayeni O. Teaching health statistics. Lesson and seminar outlines.2nd ed. Geneva. World Health Organization, 1999.
3. Bradford Hill A. A short textbook of medical statistics. 11th ed. Hodder and Stoughton, London, 1984.