Statistics- An Overview
Narasimha Murthy N S¹, Raghuram V²

Before we start our journey down the lane of statistics, let us learn about the relationship between two interrelated disciplines of research and statistics.

Definition of Research
To start with, it is important to consider several definitions of research. Research means “Search for knowledge” - which is true, we are searching for the knowledge.

Another definition of research is “systematized effort to gain knowledge”- of course, to gain knowledge systematized effort is required like observational studies, correlation studies, analytical studies and experimental studies. Yet another definition is “A movement for known to unknown”- This is also true because we start from known knowledge to gain new knowledge about unknown.

However, all these definitions sound well but lack measurement. To overcome this difficulty one definition which is based on the function of research has been defined by Clifford Woody which says “Research is to define problem, formulate the hypotheses, collect the data, organize the data, analyse the data and draw the conclusions “.

Functions of Research
Let us discuss each one of these six functions.

Define the problem: the researcher should have clear-cut idea about the problem, which he/she is interested. It may be to find the prevalence, incidence, or distribution of a disease or to find out the causes of a disease or may be to find out the effectiveness of vaccine or drug etc.

Formulation of hypotheses: Hypothesis means assumption about population value called parameter. The researcher must know exactly what he/she wants to prove. That means researcher should know the objectives clearly. In other words, what exactly he/she wants to prove must be clearly known. For example: prevalence of a particular disease is 4 per 1000 population or drug ‘A’ is more effective than drug ‘B’ etc.

Collecting the data: Based on the problem and objectives of the study, the investigator will collect data. The data thus collected should satisfy the objectives of the study.

Organizing the data: The data collected is raw data, which has to be reduced to a meaningful form. This is an important function to understand the data.

Analysis of the data: Here the data is dissected in detail to get the answer for the hypotheses.

Drawing Conclusion: Based on the analysis of data, the conclusions are drawn.

So far we have looked at research which is having six functions. Now let us look at the definition of statistics.

Definition of statistics and bio statistics
There are several definitions of statistics, but the authors consider the definition based on its function, which is universally accepted.

Statistics is defined as “a science, which deals with collection, tabulation, analysis, and inference of data”

When statistical methods are applied to biological sciences, it is called Biostatistics.

Research and statistics: Is there an association?
Research has six functions and statistics has four functions. On examination, four functions are common between research and statistics between these two. In research, the researcher should know his/her problems and the objectives. It varies from study to study. It can be a descriptive study or analytical study like case-control/ cohort or an experimental study. Once the problem of study is different, what the researcher wants to prove also becomes different. Looking at the next function, collection of data is the same in both research and statistics.

Organizing data is a research terminology, whereas, tabulation is a statistical terminology, they are the same. Analysis of data, Conclusions, and Inference are the same both in research and in statistics.

Hence, these four functions are common between research and statistics namely: Collection, Tabulation, Analysis, and Inference of data. It implies that to perform research, statistics is essential. The only variation is study problems and its objectives.

Important Developments in Statistics
Two important developments that took place in statistics are:

¹Department of Community Medicine, Saptagiri Institute of Medical Sciences, and Research centre, Bangalore. ²Department of Community Medicine, Raja Rajeswari Medical College and Hospital, Bangalore.
Correspondence to Dr Raghuram V (raghu3873@gmail.com)
(1) Statistical methods used to generalize about a large population with relatively small sample and

(2) Making prediction about an unknown parameter.

To generalize about large population with relatively small sample the following criteria has to be satisfied:

(1) Sample is representative of the same population

(2) Randomization to ensure that the sample is representative of the population. Randomization is a process of elimination of bias.

(3) Optimum sample size is required to make a valid inference. Sample size depends upon required precision. Precision consists of significance level and allowable error.

If a given sample follows all the above criteria then only inference is true.

The second important development is making a prediction. Prediction is possible by using what is known as regression analysis. For example, by knowing the height we can predict the normal weight. We can also make prediction by using several other factors known as multiple regression analysis. For example, the survival of particular type of cancer patients can be predicted based on age, gender, stage of cancer and associated diseases like hypertension, diabetes.

**Statistical methods**

Now let us start with a definition of statistics to arrive at an inference. The first function of statistics starts with collection of data. *Data is a recorded fact*. Collection of data depends upon the objective of the study. The data that is collected is an unorganized data called raw data. This data has to be organized in a meaningful form, which is known as tabulation. Tabulation is done by using frequency distribution tables. Distribution of frequency over the class interval is known as frequency table. Frequency table is nothing but keeping the like things in one place on the table. For example: males and females, arranging participants in different age groups, arranging the data based on Hb values, arranging the data based on birth weight etc. If the tabulation is done properly, the meaning of the data will be known very clearly. This is a function of utmost importance.

To understand the data exhaustively, it may not be sufficient only by tabulation, we require an indicator called “measures of central tendency” which gives the central value of the distribution. There are several measures of central tendency namely, arithmetic mean, median, mode etc. This can be compared to an institution where one individual representing the organization namely a principal, or a chairperson, or a general manager etc. who are the center of the organization. Similarly, there is a value in a set of observations called measures of central tendency.

It is not sufficient to understand the meaning of the data only by measures of central tendency. We require another indicator called “measures of dispersion” which measures variation in a set of observations. Measures of dispersion gives us how are values scattered around the central value. They are range, mean deviation, standard deviation, and coefficient of variation. This can be explained in terms of an organization where the head can be compared to a central value and the other staff members are like other observations.

Based on the objectives of the study, we collect the data, tabulate; find the central values and its dispersion. Still it is not sufficient to understand the data completely. We require a probability of the event happening in any situation. Probability can be calculated for a success or failure of an event. For example, we can calculate probability of survival, probability of getting a particular disease if a particular habit is present in an individual.

We can test if the hypothesis is true or not, based on the objectives of the study. In an experimental study, we can test whether new drug is effective or not. In a case control study or cohort study, we can test whether there is any association between cause and effect. Any hypotheses can be tested by using appropriate statistical methods.

Based on the objectives, the investigator collects the data, tabulate, and start analysing it by using measures of central tendency and dispersion. Later calculates the probability of an event and tests whether it is true or not. In this way inference is drawn.

Depending on the problem, we can find out the relationship between several variables by using correlation. We can find out the exact relationship called regression. Based on the studies we can find out the association between risk factor and disease.

**Conclusion**

Statistics teaches us how to quantify numerically. Statistics is the only subject which tells us how much wrong we may be. The statistical inferences are true 95% times or 99% times. Hence, statistical methods can be used to the requirement and benefit of the researcher. No research is complete and perfect without the use of statistical methods.

*This article is a primer for a series of articles on biostatistics, which will*
be published in subsequent issues of the journal

References