Basic Biostatistics for Non-Statisticians

Padmavathi. A, Karthika. A, Lions Aravind Institute of Community Ophthalmology, Madurai

Introduction

A great and extensive use of statistical terms and techniques in most medical journals has made the physicians and non statisticians to look back and search for statistical concepts. This article gives a very simple and basic definition of the simple terms used in statistics, though not discussed in detail. The purpose of the article is to kept as a ready reconer for those who begin with basics in statistics.

Statistics and biostatistics

Statistics is a science that deals with collection, analysis and interpretation of data about a population using only a limited number of observations. Application of statistics in medical and biological sciences is aptly called "Biostatistics". In other words, biostatistics is a sub discipline of applied statistics, which focuses on statistical support in the areas of clinical medicine, public health, environmental science and related fields.

What is data?

Data is nothing but a collection of facts and figures. For example a collection of the details on height, weight and blood from a group of students of a college forms data. A collection of type of ocular injuries of patients coming to an ophthalmic clinic is a data. The data may be in the form of numbers such as counts or percentages or in the form of simple precise statements that best reveals the content. The data may be collected either from the primary or secondary sources. Primary data is the first hand data collected for some specific purposes. For example, census data collection is a primary data collection. On the other hand, making use of such primary data, collected previously for some other purposes is secondary data collection.

Data collection plays a prominent role, since the collected data constitute the foundation on which the superstructure of statistical analysis is built, and the results are properly interpreted, which help us to take decisions. If the data collected is inaccurate or inadequate, then it might lead to faulty decisions thereby misleading the whole process.

Types of data

The types of data can be mainly classified as in (fig 1).

Quantitative data is amenable to mathematical calculations. It can be obtained by calculating or measuring using certain instruments. For example, blood pressure, temperature, intra ocular pressure etc.,

Data types that take only integers or whole numbers are referred to as discrete data types. For example, number of beds in a hospital. On the other hand, data

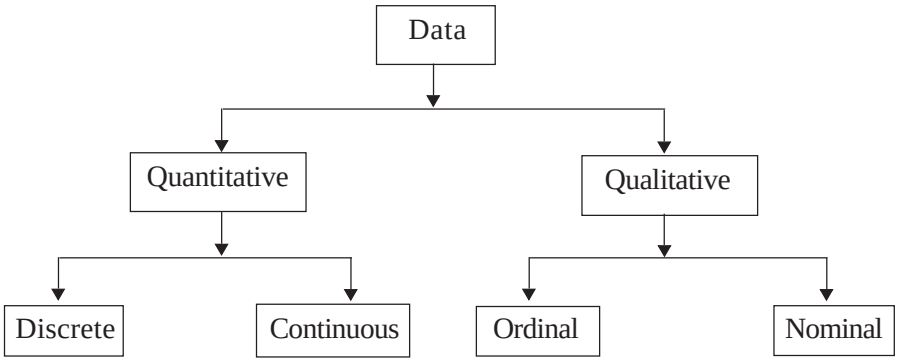

Fig 1: Data type

types that can take any values within a given range are called continuous data types. For example, there can be any number of observations within the height 146.7 to 146.8.

A nominal data type is a subclassification of qualitative data, which doesn't have a specified order.

For example, blood groups, gender etc. We cannot say that blood group 'B' should come in between "A" and "AB". As the name indicates, the ordinal data is the data that can be classified under some order. When we account for the amount of pain, we get answers like 'less pain', 'moderate pain', 'severe pain' etc. which could be ordered accordingly.

Presentation of data

After selecting a sample by applying a suitable sampling technique, the next step is to analyse and interpret the results. The first step in analysis is the "Frequency Table". A first look at a frequency table/ frequency distribution would give us an idea of how the values of the variable are spread across (Refer table1). It also gives the percent contribution of each value to that variable. This is the first and foremost step which enable us to further decide on the categorisation of the variable. A "Frequency" is just the display of the count against the values of the variable. For instance, in the above example, we can

Table 1. Frequency distribution of IOP level

IOP level (in mmHg)	Frequency	Percent (%)	Cumulative percentage
17.9	12	15.4	15.4
18.1	10	12.8	28.2
18.8	5	6.4	34.6
20.1	8	10.3	44.9
21.5	22	28.2	73.1
22.0	18	23.1	96.2
23.0	2	2.6	98.8
26.1	1	1.2	100.0
Total	78	100.0	

Table 2. Frequency distribution of classified IOP level

IOP level	Frequency	Percent (%)
< 22mmHg	57	73.1
≥ 22mmHg	21	26.9
Total	78	100.0

infer that maximum number of persons (28%) are in the group 21.5 mmHg and minimum number of people (1.2%) lie in the group 26.1 mmHg IOP. Now if we classify the IOP levels as low IOP and high IOP groups and find what happens? Thus, we see in table 2, 73% of the subjects are having IOP<22 mmHg and 27% are with higher IOP level (ie > 22 mm Hg)

Cross tabulation

In a similar fashion, a cross tabulation of more than 1 variable, their row and column percentage would give us an idea of the combination of the two variables involved. Consider the following hypothetical example,

Similarly in table 4., the row percentages add up to 100. With the row percentages, we can compare the awareness and unawareness in male and female **separately**. Thus the first row reveals that, '**among the males,** 64.4% are aware of eye donation and the remaining 35.6% are unaware of the same'. Similarly, the second row interpretation is '**among the females** only 25.5% are aware of the eye donation and the rest 74.5% are unaware of the eye donation'. Thus if we wish to **compare the row group, a column percent will help us** and if we wish to compare the column groups, then a row percent is good enough.

Table 3. Awareness of eye donation

	Aware		Unaware		Total	
	No	**(%)**	**No**	**(%)**	**No**	**(%)**
Male	152	(60.8)	84	(22.7)	236	(38.1)
Female	98	(39.2)	286	(77.3)	384	(61.9)
Total	250	(100.0)	370	(100.0)	620	(100.0)

Table 4. Awareness of eye donation

	Aware		Unaware		Total	
	No	**(%)**	**No**	**(%)**	**No**	**(%)**
Male	152	(64.4)	84	(35.6)	236	(100.0)
Female	98	(25.5)	286	(74.5)	384	(100.0)
Total	250	(40.3)	370	(59.7)	620	(100.0)

table 3 is an example for 'Column Percentage' in which each column percent add up to 100. This column percent enable us to compare the gender contribution in the 'Aware' and 'Unaware' groups individually. Here, the first column should be interpreted as '**among those who are aware** of eye donation, 60.8% are male and 39.2% are female. Whereas the second column interpretation is, '**among those who are unaware** of eye donation, 22.7% are male and 77.3% are female'. Thus the 'Total' column says that 'totally we have 38.1% male and 61.9% female in our example'. Let us see the interpretation of the same table, but with row percentage.

Measures of central tendency and dispersion

Sometimes we wish to have a summary of all the values of a variable in a single number that would best represent the whole values. Measures of central tendency or location helps us in finding a central value that would best represent all the values of the variable. Different measures are used to identify the central value for different situations based on the distribution of the data. Arithmetic mean, median, mode, geometric mean and harmonic mean are such measures of location. Though we have different measures of central tendency, the most frequently used simplest measure is an arithmetic mean.

Suppose we have two series of data with same measure of central tendency, say mean, it is not necessary that both the series are same. It is possible that the observations of one series may vary largely compared to the other. In such situations, the measures of dispersion or measures of deviation help us to identify the most varying series. As the measures of location focus on the central value, measures of dispersion focus on the distribution or deviation of all other values from the central value. range, quartile deviation, mean deviation and standard deviation are such measures. Among all, the standard deviation is the best and most frequently used deviation procedure. In most medical journals, we can see the results expressed as a combination of mean and standard deviation. A simple look at this combination will reveal to us the distribution of observations of that variable.

Suggested readings

1. *Sundar Rao PSS, Richard J, An Introduction to Biostatistics - A Manual for Students in Health Sciences, Prentice-Hall of India, New Delhi, 1997.*
2. *Swinscow TDV, Statistics at Square One, BMJ Publishing Group, 1997.*
3. *Bernard Rosner, Fundamentals of Biostatistics, Duxbury Thomson Learning, USA, 2000.*