Biostatistics in Relation to Research

The learning objective would be to understand the concept of nature of data, scales of measurement, descriptive statistics and the requirement of sample size for various study designs.

Slide 3 -6:

Data refers to a set of values, which are usually organized by variables (what is being measured) and observational units (members of the sample/population). Variable can be that something you measure, example measuring weight, height, BP or the condition that you change, eg studying effect of 2 types of diet on Overweight / Obese subjects.

Variables are of different types and can be classified in many ways, for example as numerical and categorical variables. Numerical variables are measured by some (usually existing) measures, whereas categorical variables are qualitative

The continuous variable is numerical and it can take, in theory, an infinite amount of values. An example of such a variable is length in centimeters or inches. The discrete variable is also numerical, but differs from a continuous variable in that it takes a finite number of values. An example of a discrete variable is a performance score from 1 to 10.

Categorical variables, in turn, can be nominal, in which case there is no order at all: each category has its unique meaning ("Blood group 1= A +, 2 = B+, 3 = O+, 4 = AB+"). If there is a sense of order there, the variables are called ordinal. A Likert-type scale represents an ordinal measurement: 1 = "not like me at all", 2 = "not like me", 3 = "not sure", 4 = "somewhat like me", 5 = "very much like me". A special type of variable is the dichotomous one. It can have only two values (e.g. gender).

Slide 8 - 16:

Descriptive statistics are used to describe the basic features of the data in a study. Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way such that, for example, patterns might emerge from the data. Descriptive statistics do not, however, allow us to make conclusions beyond the data we have analysed or

reach conclusions regarding any hypotheses we might have made. They are simply a way to describe our data in a logical, meaningful, and efficient way.

Typically, there are two general types of statistic that are used to describe data:

Measures of central tendency: The central tendency of a distribution is an estimate of the "center" of a distribution of values. these are ways of describing the central position of a frequency distribution for a group of data. These are three different kinds of "averages" and certainly the most popular ones. Central tendency determines the tendency for the values of your data to cluster around its mean, mode, or median.

The mean is simply the average and considered the most reliable measure of central tendency for making assumptions about a population from a single sample. The mean is computed by the sum of all values, divided by the number of values. The median is the "middle" value or midpoint in your data and is also called the "50th percentile". Note that the median is much less affected by outliers and skewed data than the mean. The median will not be heavily affected by these outliers since it is only the "middle" value of all data points. Therefore the median is a much more suited statistic, to report about your data. The mode is the value or category that occurs most often within the data. Therefore a dataset has no mode, if no number is repeated or if no category is the same. The mode is also the only measure of central tendency that can be used for categorical variables. In a normal distribution, these measures all fall at the same midline point. This means that the mean, mode and median are all equal.

Measures of spread: these are ways of summarizing a group of data by describing how spread out the data points are. Measure of Spread refers to the idea of variability within your data. To describe this spread, a number of statistics are available to us, including the range, quartiles, absolute deviation, variance and standard deviation.

The range is the difference between the largest and the smallest observation in the data. The prime advantage of this measure of dispersion is that it is easy to calculate. It is very sensitive to outliers and does not use all the observations in a data set.

Interquartile range is defined as the difference between the 25^{th} and 75^{th} percentile (also called the first and third quartile). Hence the interquartile range describes the middle 50% of

observations. If the interquartile range is large it means that the middle 50% of observations are spaced wide apart. The important advantage of interquartile range is that it can be used as a measure of variability if the extreme values are not being recorded exactly (as in case of open-ended class intervals in the frequency distribution).

Standard deviation is the measurement of average distance between each quantity and mean. It is a measure of spread of data about the mean. SD is the square root of sum of squared deviation from the mean divided by the number of observations. A low standard deviation indicates that the data points tend to be close to the mean of the data set, while a high standard deviation indicates that the data points are spread out over a wider range of values. The reason why SD is a very useful measure of dispersion is that, if the observations are from a normal distribution, then 68% of observations lie between mean ± 1 SD 95% of observations lie between mean ± 2 SD and 99.7% of observations lie between mean ± 3 SD. The other advantage of SD is that along with mean it can be used to detect skewness.

Slide 17 -18:

A frequency distribution tells how frequencies are distributed over values. Frequency distributions are mostly used for summarizing categorical variables. The frequency of an observation tells you the number of times the observation occurs in the data. It is a summary of the data occurrence in a collection of non-overlapping categories. These result in tables and charts that give insight into your data. Histogram are the visualize frequencies for *intervals* of values rather than each distinct value. Generally, frequency distribution can be associated with the charting of a normal distribution. Probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment. A probability distribution is a statistical function that describes all the possible values and likelihoods that a random variable can take within a given range. This range will be bounded between the minimum and maximum possible values, but precisely where the possible value is likely to be plotted on the probability distribution depends on several factors. These factors include the distribution's mean (average), standard deviation, skewness, and kurtosis.

Slide 19:

Choice of statistical tests:

Some of the commonly used statistical tests:

Independent t test:

When to use: When the objective is to compare the mean values between two independent study groups; Example when we compare the mean haemoglobin levels between rural and urban population. One variable is categorical (Rural/Urban), and another variable is continuous (Hb level) in nature.

Paired t test:

When to use: When the objective is to compare the mean values of paired samples (before and after experiment measures); Example when we compare the mean haemoglobin levels within the same group in study subjects before and after treatment. One variable is categorical (Pre /Post), and another variable is continuous (Hb level) in nature.

Chi-square Test:

When to use: When the objective is to test the association between two categorical variables. Example test the proportion of anaemia (One categorical - Anaemia – present / absent) between rural and urban (Other categorical - rural / urban) population.

Slide 21:

Sample size estimation is based on the primary objective and primary outcome of the study. Slide 21 gives the requirement for sample size calculations. Sample size estimation depends on the number of study groups, primary outcome and clinically expected variation and meaning difference that we expect between study groups.

Slide 22 – 43 gives sample size requirement and the formula for various study designs.

Sample Size Estimation:

A research can be conducted for various objectives. It may be done to establish a difference between two treatment regimens in terms of predefined parameters like beneficial effect of these regimens. Sometimes, the purpose may be to achieve certain estimation in the population, such as the prevalence of a disease. Whatever be the aim, one can draw a precise and accurate conclusion only with an appropriate sample size. The three main factors which must be considered are α-error, β-error and clinically significant difference or the effect size. Type I error or α-error is failure to accept the null hypothesis when it is actually true. Usually it is set at 5%. The sample size has to be increased if this value has to be lowered. Type II error or β-error is failure to reject the null hypothesis when it is not true. By convention, it can be set at 20%, 10% or 5%. Power of the study is equal to 1-type II error; hence any study should be at least 80% powered. The sample size increases when the power of study is increased from 80% to 90% or 95%. The third factor is the effect size. A small clinically significant difference is difficult to identify and needs a larger sample size as compared to a study with a larger clinically significant difference. The other factors which need to be considered are standard deviation for quantitative measurements, margin of error and attrition rate. These values are either known from literature or can be decided by a pilot study or by reasonable guess work.