Resident Corner www.pjkd.com.pk

Statistics and Biostatistics-1 (Basic Concepts)

Afshan Khanum, M Phil (Statistics)
Fatima Memorial College of Medicine & Dentistry, Lahore, Pakistan

Email: Afshankhanam16@hotmail.com

Statistics is a branch of applied mathematics which contains mainly four aspects of data:

- Collection
- Summarization
- Analysis
- Interpretation

Basically, data is a collection of raw facts and figures and the purpose of statistics is to convert those into meaningful information. Statistics is being used in many fields like biology, sociology, economics and business etc. When statistical tools are employed to the data related to biological sciences like biology, medicine or public health, then the term used is known as Biostatistics.

There are two major areas of statistics:

- Descriptive statistics.
- Inferential statistics.

Descriptive Statistics:

This area of statistics deals with the summarization and description of data. **This is the most basic form of statistics on which all statistical knowledge is based.** Descriptive statistics mainly contains measure of central tendency and dispersion (scattering of data).

a) Measure of central tendency:

Measure of central tendency is a measure which identifies the central or average value of the data. It includes mean median, mode skewness, kurtosis etc.

Mean: It is a well-known measure of central tendency which can be calculated for both continuous and discrete data. **In Mathematical terms mean is defined as sum of observations divided by the total number of observations.** It is a mathematical average. It is suitable in case of symmetrical data (evenly distributed data). One main disadvantage of mean is that it is affected by extreme values. In case of extreme values or skewed data it loses its ability to provide the best central location of the data because skewness is dragging it away from the typical central value. Mean uses all the information contained in the sample whereas other measures like median and mode lack this ability.

Example: Suppose following are the ages of the patients with severe COVID 19 infection.

44, 77, 42, 63, 59, 53, 77, 43, 42 59, 59, 67

Mean can be calculated by

1. Adding all the above ages
   44+ 77+ 42+ 63, 59+ 53+ 77+ 43+ 42 +59+ 59+ 67=685
2. Dividing their sum to total number of observations
   685/12=57.08

Hence, in this hypothetical data the average age of the patients with severe COVID 19 infection is almost 57 years.

Median: central value in arranged data is median, which is more appropriate to calculate in case of skewed data. Skewed data means if values are more towards one extreme eg. Height of children: among 30 children if 20 are around 3 feet tall and very few around 2 feet, then there are more number of tall children then shorter in the group and thus the data is skewed towards one side. **Median is a positional average and it is not affected by extreme** values so the point when the mean is unable or inappropriate to provide suitable measure of central tendency, median is available there. One main disadvantage of median is that it does not uses all the information contained in the sample.

Example: Suppose following are the ages of the patients with severe COVID 19 infection.

44, 77, 42, 63, 59, 53, 77, 43, 42 59, 59, 67

Median can be calculated by

1. Arranging them in ascending or descending order:

42, 42, 43, 44, 53, 59, 59, 59, 63, 67, 77, 77

2. Finding the middle position of the data by the formula (n+1)/2
   Where n is the total number of patients.
   - If total number of observation is odd, then median falls on single observation otherwise it falls between two middle observations. So, in case of even number of observation median is the average of two middle values.

   Here in this example n is 12
   Hence (12+1)/2= 6.5th term is the median
   It means that median of the data lies between 6th and 7th term
3. Identification of the value at the middle position
   The 6th and 7th term in data is 59 so the median is 59

Mode: Most frequently occurring value in data. This is the least used measure of central tendency. **This is the best and the only measure of central tendency in case of dealing with the nominal data.** A data can have one, two or multiple modes. Hence the distribution of a data can be unimodal, bimodal or multimodal.

In the presence of symmetrical (evenly distributed) data the mean median and mode will be same. Where data is skewed, the median and mode are different.

Example: Consider the above example in median most frequent value in data is age 59 hence is the mode.

b) Measure of variability or dispersion:

Measure of variability or dispersion indicates the scatter of the data. **Simply the dispersion measures the extent to which the observations tend to differ from each other and their central mean value.** Low variability means data points are more clustered around the central value (this indicate data is more consistent) and a high variability indicates that the data points tend to fall away from the central measure. **Measure of central tendency provides a typical central value only; measure of variability tells how far the values are lying from that central value,** hence, reporting measure of central tendency only is not meaningful unless it is being reported with its particular measure of variability. Range, variance, standard deviation, Interquartile range etc. are the examples of measure of dispersion.

Range: This is the simplest and easy to calculate form of measure of variability. **Range of a data set is the difference between the smallest and highest value.** The larger the difference the larger the variability is. As range depends on two values of the data so it is susceptible of outliers. If one those number is very high or low, it changes the entire variability of the data which is definitely not a true representative of that particular data. Range is suitable only in case of small sample size.

Example: Consider the above mentioned example, largest value in data is 77 and smallest value is 42. Range can be calculated as

Range=largest value – smallest value

Range=77-42

Range=32=5

Interquartile Range (IQR): IQR is the middle half of the data based on dividing the data into quartiles. Quartiles divide the rank ordered data into four equal parts by the three points known as Q1, Q2, Q3. Q1 is called lower quartile, Q3 is called upper quartile and Q2 median of the data. IQR is the difference between the upper and the lower quartile. It is appropriate measure of variability in case of skewed data which is reported with median.

Formula:

$Q1=(n+1)/4^{th}$ term

$Q3=3(n+1)/4^{th}$ term

Example: Consider data in above mentioned example:

44, 77, 42, 63, 59, 53, 77, 43, 42 59, 59, 67

Quartile 1 can be calculated by

Arranging them in ascending or descending order:

42, 42, 43, 44, 53, 59, 59, 69, 63, 67, 77, 77

Finding the First quartile position of the data by the formula (n+1)/4

Where n is the total number of patients.

Here in this example n is 12

Hence (12+1)/4= 3.25th term is the Q1
Q1= 43+0.25*(44-43) =43.25
Similarly, Q3 can be calculated as
Q3=3(n+1)/4= 3(12+1)/4= 9.75
Q3=63+0.75*(67-63) =66
So IQR= Q3-Q1
IQR= 66-43.25=22.75

Standard deviation:

Standard deviation is a measure of dispersion that tells how the data points are spread out around the mean. The smaller standard deviation means all the data points are lying near mean of the data and there is less spread in data. A larger standard deviation indicates larger spread in data. Standard deviation is reported with the mean of the data.

Formula: $\Sigma(y-\bar{y})^2/n$

Where y is the variable and $\bar{y}$ is the mean of that variable.

Descriptive statistics also help us to understand the shape of the data with the help of charts and graphs.

Inferential Statistics

It is the branch of statistics that deals with making inferences about the characteristics of the population on the basis of the information contained in the sample. **It is used to test statistical hypothesis and draw conclusions by computing confidence interval or statistical significance.** It estimates population parameters using information obtained from a sample of that population.

There are two major areas of inferential statistics:

- Estimation
- Hypothesis testing.

Estimation:

There are two main types of estimation: point estimation and interval estimation.

Point Estimation: It is the process of estimating a single value of parameter. In which sample data is used to calculate statistic (for example sample mean) and that statistic is used to say something about the population parameter (for example population mean). Sample is the point estimate of population mean.

Interval estimation: It is a process which provides a range of values in which population parameter tend to lie.

Hypothesis testing:

It is a process of making inferences about the population parameter on the basis of sample data.