**Continuous Medical Education**
pISSN 0976 3325 | eISSN 2229 6816
Open Access Article
www.njcmindia.org

# Medical Statistics Series: Type of Data, Presentation of Data & Summarization of Data

**Swati Patel**[1]

[1]Statastician cum Assistant Professor, Department of Community Medicine, SMIMER, Surat

**ABSTRACT**

Bio Statistics can define as the application of mathematical tools used in statistics to field of biological science and medicine statistics. Collection, organization, analysis ,interpretation and presentation of data very important part of bio medical research, which is mainly used in research studies .It deal with aspect of this , including the planning of data collection in term of the design of survey and experiments. Now a day, doing various statistical tests has been made easy by sophisticated computer software. But mainly two things are very important to researcher/ investigator is to choose the appropriate statistical test for the computer to perform based on the nature of data derived from one's own research. The second is to understand if an analysis was performed appropriately during review and interpretation of others' research. A basic understanding of biostatistics is needed to understand and interpret the medical literature. The basic step of research is defining the research question, review of literature, formulate hypothesis, preparing study design, data collection, analysis and interpretation of it.

Bio statistics is mainly concern with descriptive statistics and inferential statistics. The objective of this article is that to understand the type of data, choose the correct method of presentation and summarization of the data according to nature of data.

**Keywords:** Data type, Biostatistics, Medical Analysis, Presentation of data

## INTRODUCTION

Biostatistics has played a very significant role in modern medicine .Statistical methods are important to draw valid conclusion from obtained data. So, Understanding the fundamentals of biostatistics is essential in order to plan a scientific study and interpret its results.

In the first article of this series, we look types of data, Presentation of data and the measures used to describe or summarize data.

Data can be defined as a systematic record of a particular quantity It is the different values of that quantity represented together in a set. It is a collection of facts and figures to be used for a specific purpose such as a survey or analysis. Source of data (primary or secondary) is also an important factor.

**Primary data:** These are the data that are ***collected for the first time*** by an investigator for a specific purpose. It is **data that is collected by a researcher from first-hand sources**, using methods like surveys, interviews, or experiments. It is collected with the research project in mind, directly from primary sources. An example of primary data is the Census of India, NFHS survey

**Secondary data:** They are the data that are ***sourced from someplace*** that has originally collected it. This means that this kind of data has already been collected by some researchers or investigators in the past and is available either in published or unpublished form. This information is impure as statistical operations may have been performed on them already. It is the data gathered from studies, surveys, or experiments that have been run by other people or for other research.

**How to cite this article:** Patel S. Medical Statistics Series: Type of Data, Presentation of Data & Summarization of Data. Natl J Community Med 2021;12(2):40-44

**Financial Support:** None declared **Conflict of Interest:** None declared

**Date of Submission**: 07-02-2021; **Date of Acceptance**: 24-02-2021; **Date of Publication**: 28-02-2021

**Correspondence:** Dr. Swati Patel (E-mail: Swati84patel@gmail.com)

**TYPE OF DATA**

There are two types of data (i) Qualitative data (ii) Quantitative data

**(i) Qualitative Data**

The data which cannot be measurable, but it can be defined the characteristic of it is called qualitative data, which is divided in two parts, i.e. (a) Binary and (b) Categorical data.

**(a) Binary Data: -** If qualitative data has only two possible levels, it is referred as binary or dichotomous data.

Example: Gender of new born baby (M/F) Disease (present/absent), Smoking (yes/no), Outcome of patients (Death/Survive)

**(b) Categorical Data: -** The data which belongs to the more than two categorical values is known as categorical data, which is divided in two parts i.e (b.1) Ordinal data and (b.2) Nominal data.

**(b.1) Ordinal Data:-**The data which implied ranking or ordering of observation, is known as ordinal data .It is classification of set of observation. Each observation expressed as relative position in group as first, second, etc. It is also known as ranked data.

Example: Level of pain associated with the medical condition (pain may be grated as mild, moderate, sever), Level of satisfaction of patients toward treatments, Degree of malnutrition (It may be grated as mild, moderate, sever and no malnutrition)

**(b.2) Nominal Data**: - It is provided simple categorization of observation.

Example:-Blood Group, Type of Delivery, Co-morbid conditions

**(ii) Quantitative Data**

The data which can be measurable is known as Quantitative Data. It is usually referred as enumeration or counting type data .It is classified in two parts i.e. (a) Discrete and (b) Continuous.

**(a) Discrete data**: - Discrete data *is* data that can only be counted in whole numbers. e.g., number of hospital visits, Glasgow Coma Score, Richmond Agitation Sedation Scale. This type of data cannot be represented in decimals – so Glasgow Coma Scale score can be 7 or 8 not 7.5.

**(b) Continuous data**: - Continuous data is data that theoretically has no gap between data points can be counted in decimals (depending on the precision of the measuring instrument) e.g., blood glucose, haemoglobin, serum lipids. In summary, quantitative data deals with objective measurements.

Raw data is arranged in the order in which they are collected. Data in this form is difficult to understand and interpret. To get information the raw data, the data must be organized in some orderly fashion i.e. data must be presented either in tabular or graphical form. The presentation of data is as kill which help to draw conclusion based on statistical analysis. The wrong presentation may lead to wrong impression about findings.

**PRESENTATION OF DATA**

There are three ways of representation of data; we can use any one out of it.

i.e. As text, Tabular form and Graphically form

**1) Text Presentation**

Text is the main method of conveying information as it is used to explain results and trends, and provide contextual information. Data are fundamentally presented in paragraphs or sentences. Text can be used to provide interpretation or emphasize certain data. If quantitative information to be conveyed consists of one or two numbers, it is more appropriate to use written language than tables or graphs. Example) "A total of 330 patients were initially selected for the study No differences were observed in mean baseline BMI between hypo and hyperthyroid patients. Overweight or obesity was observed in 76.5% and 58.8% of hypothyroid and hyperthyroid patients, respectively, which is statistically significant.

.If this information was to be presented in a graph or a table, it would occupy an unnecessarily large space on the page, without enhancing the readers understanding of the data. If more data are to be presented or other information such as that regarding data trends are to be conveyed, a table or a graph would be more appropriate. By nature, data take longer to read when presented as texts and when the main text includes a long list of information, readers and reviewers may have difficulties in understanding the information.

**2) Tabular Form**

A table helps representation of even large amount of data in an engaging, easy to read and coordinated manner. The data is arranged in rows and columns. This is one of the most popularly used forms of presentation of data as data tables are simple to prepare and read.

**Tabular representation of Qualitative data:-** Qualitative tabulation is simple because it is categorical data, which is tabulated by listing all possible categories in the given data set. Example: Details of co-morbidities in TB patients. (Table.1)

**Table 1: Co-morbidities in TB patients**

| Co-morbidities | Numbers |
|---|---|
| Hypertension | 120 |
| Asthma | 69 |
| Diabetes | 137 |
| Hypertension+ Diabetes | 89 |
| Total | 415 |

**Tabular representation of Quantitative data**:- Classification is the first step in the tabular representation of quantitative data. It means separating the total group of observations in to smaller groups according to similarity or dissimilarity of the items with respect to characteristic under study. Example: Haemoglobin level of 843 pregnant women.

**Table 2: Haemoglobin level of 843 pregnant women**

| Hb level | Numbers |
|---|---|
| <7 | 49 |
| 7-9 | 129 |
| 9-11 | 231 |
| 11-13 | 267 |
| 13-15 | 167 |
| Total | 843 |

**Graphical Presentation of Data:-**

Graphics are powerful tools to communicate research results and to gain information from data. However, researchers should be careful when deciding which data to plot and the type of graphic to use, as well as other details.

**1) Bar Diagram**:-It is commonly used for the presentation of qualitative data (i.e mainly Nominal )It may consist of either horizontal or vertical columns. The greater the length of the bars, the greater the value. They are used to compare a single variable value between several groups, such as the Gender wise mortality due to different vector bone disease.

**Simple bar diagram (figure.1)**

Simple bar diagram is used to represent data involving only one qualitative variable classification (figure.1)

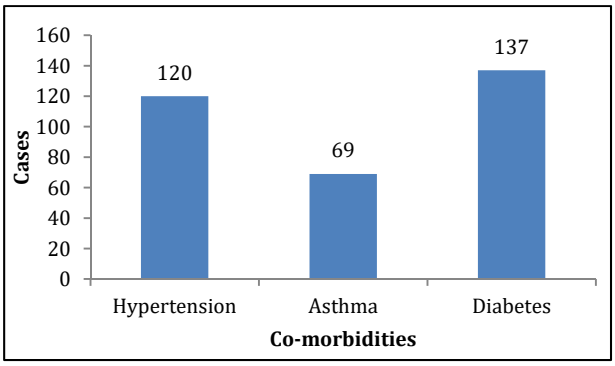


**Figure 1: Co-morbidities in TB patients**

**Multiple bar diagram:**

Multiple bar diagram is used to represent data involving more than one qualitative variable classification, which shows the relationship between different set of data.

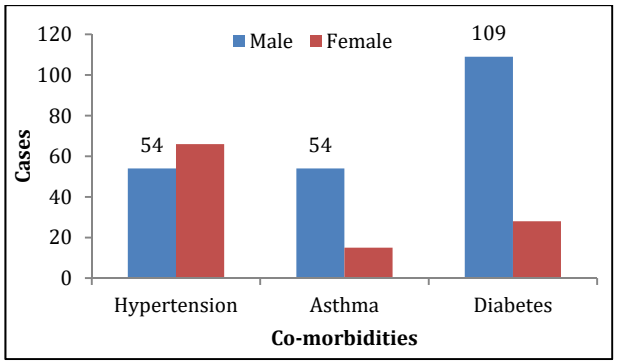


**Figure 2: Gender wise Distribution of co-morbidities in TB patients**

Component Bar diagram:- In component bar diagram the bars will be divided two of more parts, each part represents a certain item and proportional to the magnitude of that particular item. Example : frequency distribution of awareness about HIV/AIDS by socioeconomic status

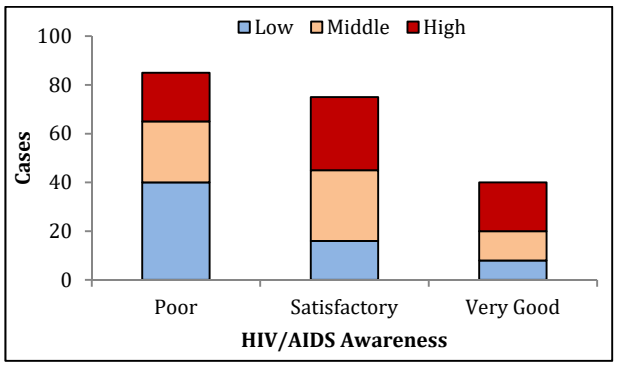


**Figure 3: Socioeconomic status & HIV/AIDS awareness**

**Pie Chart**:-

A pie chart is best used when trying to work out the composition of something. If you have Qualitative-categorical data then using a pie chart would work really well as each slice can represent a proportion of different category. Example –percentage of co-morbidities in T patients.

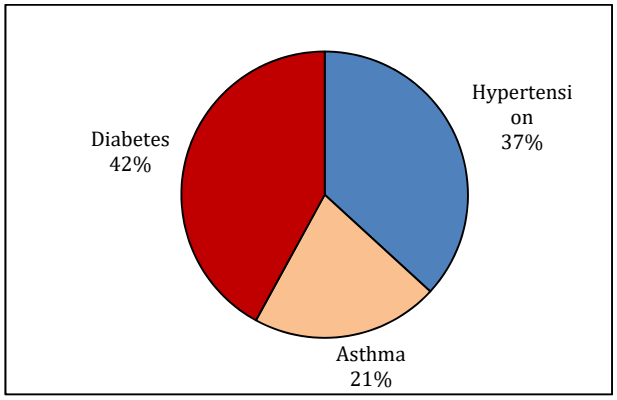


**Figure 4: Co-morbidities in TB patients**

**Line Chart**

Line graphs are used to track changes over short and long periods of time. When smaller changes exist, line graphs are better to use than bar graphs, in short when periodic data`s are given (i.e growth of bacteria per second/min/hour, cases of covid19 per day/week/months, yearly case of malaria) line chart is the best graphical presentation than others. X- axis of graph consider as time period (i.e sec/ min/ hours/ day/ week/ month/ year)and Y-axis of graph consider frequency/numbers.

Line graphs used to compare changes over the same period of time for more than one group is called multiple line graph.

Figure 5 shows trend of malaria Cases in city A (when periodic data is given with single group).

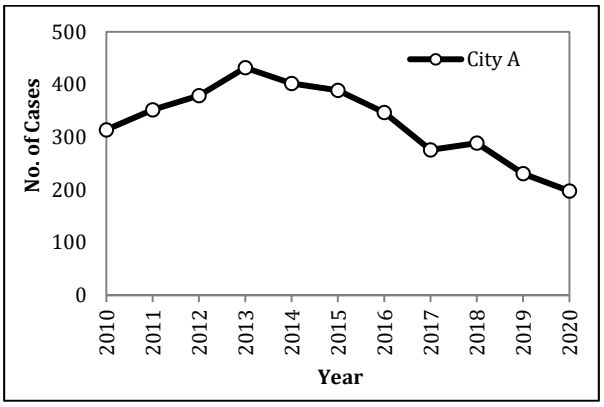


**Figure 5: Trend of malaria Cases in City A**

Trend of malaria Cases in City A and City B (when periodic data is given with more than one group of same variable)

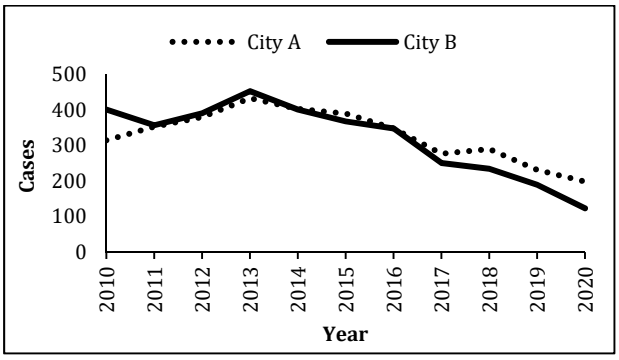


**Figure 6: Trend of malaria Cases in City A and B**

**Histogram**

Histograms are useful if you are trying to graph a large set of quantitative data. A Histogram displays a range of values of a variable that have been broken into groups or intervals. To make a Histogram, divide the range of data into intervals of equal length, count the number of observations in each interval, and represent each interval with a bar indicating the number of observations.

In histogram the joining of the midpoint of column with smooth curve is called frequency polygon, which mainly use to know the type of distribution of data.

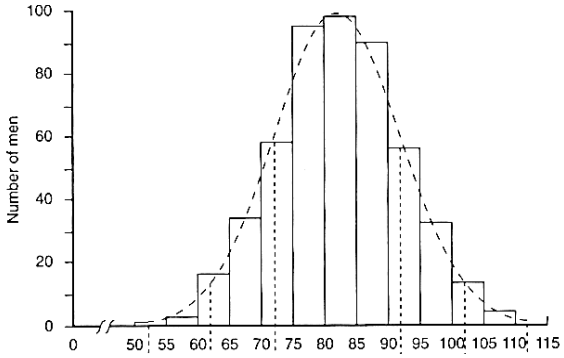


**Figure 7: Classification of Diastolic Blood pressure in men**

Summarization of data:-

Summarization of raw data set (qualitative or quantitative) using percentage /proportion, measure of central tendency and measure of dispersion can be defined as descriptive statistics.

Measure of central Tendency:-

The mean (or more correctly, the arithmetic mean) is calculated as the sum of the individual values in a data series, divided by the number of observations. The mean is the most commonly used measure of central tendency to summarize a set of numerical observations. It is not reliable or good measure of summarization when extreme values (outlines) present in a data series. It should not, generally be used, in describing categorical variables because of the arbitrary nature of category scoring. It may, however, be used to summarize category counts.

For data sets with extreme values, the median is a more appropriate measure of central tendency. If the values in a data series are arranged in order, the median denotes the middle value (for an odd number of observations) or the average of the two middle values (for an even number of observations). The median denotes the point in a data series at which half the observations are larger and half are smaller. As such it is identical to the 50$^{th}$ percentile value. If the distribution of the data is perfectly symmetrical (as in the case of a normal distribution that we discuss later), the values of the median and mean coincide. If the distribution has a long tail to the right (a positive skew), the mean exceeds the median; if the long tail is to the left (a negative skew), the median exceeds the mean. Thus, the relationship of the two gives an idea of the symmetry or asymmetry (skewness) of the distribution of data.

Mode is the most frequently occurring value in a data series. It is not often used, for the simple reason that it is difficult to pinpoint a mode if no value occurs with a frequency markedly greater than the rest. Furthermore, two or more values may occur with equal frequency, making the data series bimodal or multimodal.

**Measure of Dispersion:-**

The measures of central tendency are not enough to describe data. Two data sets can have the same mean but they can be completely different. Thus to describe data, one needs to know the extent of variability. This is given by the measures of dispersion. Range, interquartile range, and standard deviation are the three commonly used measures of dispersion.

Range:-The numerical distance between the largest value (maximum) and smallest value (minimum), it tells us about the variation in scores we have in our data, or it tells us the width of our data set. Range is not reliable measure of dispersion; it gives only rough idea about the spread of data.

**Inter quartile range (IQR):**

It is a measure of variability, a ranked –based data set divided into four equal parts (25%). The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, Q3, respectively. The second quartile ($Q_2$) is the median. Inter quartile rage shows the variation between first and third Quartile. Mainly

**Standard Deviation**:- It is a very important measure of variability to summarize the quantitative (numeric) data, which shows that the observations are how much far /scatter from the mean values. The lower value of standard deviation shows less variation within the observations form mean value, whereas the large value of standard deviation show the more variation within the observation.

Any one measure (either measure of central tendency or dispersion) is not enough to summarization of data, as when two or more sets of data have the same mean but the variation among the observations may be not same so need to use both measure to summarization of data. Quantitative data (numeric) can be summarised using Mean ±SD, if no extreme value is present in the given data series. But when some extreme values present in the given data series (quantitative – numeric) for that Median (IQR) is best measure of summarization instead of Mean ±SD. The summarization of data set is depend on the type of data, as qualitative data (binary or categorical) can be represented as percentage/proportion.

| Type of data | Presentation of data | Graphically presentation | Summarization | Example |
|---|---|---|---|---|
| **Qualitative** | | | | |
| Binary | Frequency table | Pie chart | Percentage/ proportion | Do you have fever?(Yes/No); Outcome of patients (Death/Alive); Gender(Male /Female) |
| Categorical-Nominal | Frequency table | Pie chart Bar chat | Percentage/ proportion | Co-morbidities in covid19 +ve cases (i.e Hypertension, Asthma, Diabetes...etc); Blood group of patients (O+, O-,....) |
| Categorical-Ordinal | Frequency table | Bar chat | Percentage/ proportion | Pain of patient(mild, moderate, sever); Psychometric scale (i.e. Strongly Agree, Agree, Disagree,....) |
| **Quantitative** | | | | |
| Discrete | Frequency table | Bar chat | Percentage/ proportion | Number of lived children delivered by a women; Number of pregnancy |
| Continuous-(when no extreme value present in the data series ) | Frequency table* | Histogram, Frequency polygon. | Mean ±SD | Blood pressure(SBP/DBP); Blood Glucose; Hb |
| Continuous-(when extreme value present in the data series) | | | Median (IQR) | Age of patient of Malaria (2yrs, 1yrs, 13yrs, 21yrs, 45yrs, 78yrs, 90yrs……..); Hospitalization of patients (in days) (3,2,5,1,15,45……..) |

*when quantitative data converted in to qualitative data ex. Blood pressure (SBP/DBP) values convert to slight above normal(Systolic 140-159 mm of Hg and/or Diastolic 90-99 mm of Hg), Moderately high (Systolic 160-179 mm of Hg and/or Diastolic 100-109 mm of Hg), Very high (Systolic ≥180 mm of Hg and/or Diastolic ≥110 mm of Hg)

## REFERENCES


1. K. visweskar Rao, Edition (2009) Biostatistics- AManual of Statistical Methods for Use in Health, Nutrition and Anthropology; Jaypee Brothers.
2. Avijit Hazra[1], Nithya Gogtay[2], Biostatistics Series Module 1: Basics of Biostatistics, Indian J Dermetol.2016 Jan-Feb;61(1);10-20. https://www.ncbi.nlm.nih.gov/pmc/ articles/PMC4763618/
3. Priya Ranganathan[1] and Nithya J Gogtay[2] ,An Introduction to Statistics – Data Types, Distributions and Summarizing Data, Indian J Crit Care Med. 2019 Jun; 23 (Suppl 2): S 169–S170. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6707495/