J Indian Prosthodont Soc (July-Sept 2013) 13(3):384–386
DOI 10.1007/s13191-013-0300-8

ABSTRACT ON BIOSTATISTICS

Know Your Data Before You Undertake Research

Nagesh Lakshminarayan

Published online: 13 July 2013

Data is a piece of information collected in the research (by interaction with patients, participants and research subjects or direct observation of subjects or specimens) [1]. Primary data is obtained directly from the source but secondary data is collected by going through the existing records like past patient records or census records. A researcher should know and understand the nature of data that needs to be handled before embarking on conducting research. The nature of data will have an ultimate say about how the observations are going to be described and analyzed.

I. Data can be broadly classified into two types [2]:

a. Qualitative data: Includes those which are defined by some characteristic or quality.

E.g.: Satisfaction level of patient after receiving complete denture.
Gender distribution of patients visiting prosthodontics clinic.

b. Quantitative data: It includes such data which are measured on a numerical scale.

E.g.: Surface hardness of different heat cure acrylic resins.
Tensile strength of different alloys used for casting in fabrication of full crowns.

II. Data can also be classified into:

Categorical Data

a. Nominal data
Is a type of qualitative data where observations are listed under multiple groups.

E.g.: Gender—Male and female
Blood groups: A, B, AB, and O
Rh^+ and Rh^-

When there is a possibility of only two categories such a data is called Dichotomous or Binary data.

E.g.: Denture wearer/not denture wearer
Edentulous/non edentulous

The number of observations in each category is then represented as absolute counts, percentages, rates or proportions.

b. Ordinal data
It is also a categorical data. The observations are categorized or listed under categories just like the nominal data but there is an inherent rank order among the categories.

E.g.: Pain score
0 = No pain
1 = Mild pain
2 = Moderate pain
3 = Severe pain
4 = Very severe pain

In this example the scores from 0 to 4 in order represent pain in increasing order. Score 2 represents certainly more pain than score 1, but how much more is not known since intervals are not defined as it is not an interval scale.

E.g.: Patient satisfaction
0 = Not satisfied
1 = Mild satisfaction
2 = Moderate satisfaction
3 = High satisfaction

N. Lakshminarayan (✉)
Bapuji Dental College and Hospital, Davangere, India
e-mail: drlnagesh72@gmail.com

These observations are described or represented by absolute counts, percentages, rates or proportions, ordinal data can also be summarized by the median value range.
Note Nominal and ordinal data do not have well defined intervals between the anchors on the scale, hence they should be analyzed using *non-parametric statistics.*

Numerical Data

It is quantitative data and characterized by well defined intervals on interval scale. To say that the data is on interval scale we must be sure that equal intervals on scale represent equal differences in the property being measured. E.g.: The difference between 2 and 3 kgs weight is same as the difference between 10 and 11 kgs, because the intervals are meaningful and uniform. This kind of data can be subdivided into two groups:

a. Discrete data
b. Continuous data

Discrete data can be recorded as whole numbers (integers) whereas continuous data can assume any value between the whole numbers. In simple terms the observations that are measured are continuous.
E.g.:

1. Episodes of denture fracture after delivery of the denture (discrete)
2. Number of visits given by patient to clinic before denture insertion is done (discrete)
3. Weight of denture (continuous)
4. Modulus of resilience of denture base (continuous)
   Continuous data can be further divided into interval scale and ratio scale

c. Interval scale: Is such a scale that the differences or intervals on the scale are uniform but they are not meaningful because the scale lacks true zero.

   E.g., Temperature as measured on absolute scale The difference or interval between 20 and 30 °C is same as difference between 60 and 70 °C but the important issue is 20 °C is not double that of 10 or 10 °C is not half that of 20 °C, this is because the scale lacks true zero and values <0 are possible.

d. Ratio scale: Is interval scale but has true zero. The intervals are not only uniform but also meaningful.

   E.g., Weight as measured in any unit. 20 kgs is double that of 10 and 10 kgs is half that of 20 kgs. This has meaningful intervals because it has true zero.

A majority of numerical data that are used in dentistry is of ratio scale. Numerical data is measured and represented as mean and standard deviation or median and range. When numerical data follows normal distribution, parametric statistics can be applied (Scheme 1).

Selection of Data Scales

Precision is less for categorical scales and they provide limited information. Numerical data have better precision and they also provide more meaningful information.
Thumb rule: When it is possible, use numerical data. Never reduce numerical data into categorical because precision of data is lost. It is not advisable to convert numerical data into categorical data since there is loss of vital information.

E.g.: Smoking status of a population recorded in following scales

Categorical

Nominal (dichotomous, binary): smokers/non smokers
Ordinal data: Heavy smoker/light smokers/ex smokers/non smokers

Numerical

Discrete data: By the number of cigarettes smoked/day
Continuous data (ratio scale): By recording serum cotinine levels

Presentation of Data

In any research, data has to be scrupulously collected and presented. The presentation of data is a hard science and a fine art. Data can be presented verbally or graphically. There are specific guidelines about how data needs to be presented. Before going into details of data presentation, it is better to have an overview of statistical averages.

Descriptive statistics summarize a collection of data from sample or population. Data consisting of many values should always be condensed summarized and presented as statistical averages, like *mean, median and mode.* These averages are measures of central tendency. A measure of central tendency represents a central value in the data distribution around which all other values are distributed.
Mean

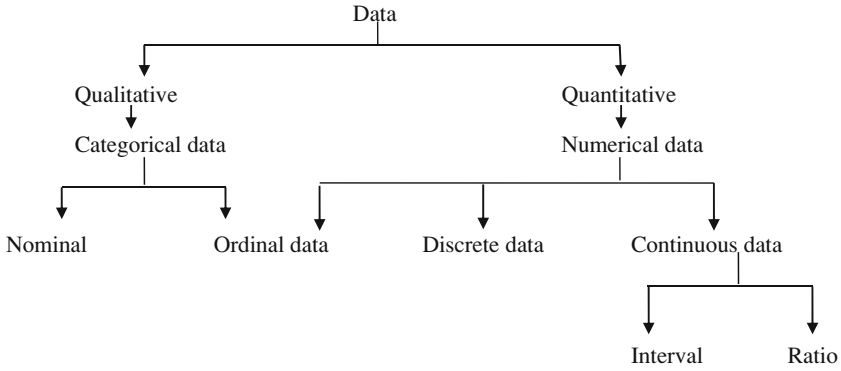
More correctly called the arithmetic mean. It is calculated as the sum of the observations, divide by the number of observations. The formula for finding out the mean is:

$$\bar{X} = \frac{\sum X}{n}$$

where $\bar{X}$ is the mean, $\sum X$ is the sum of all individual observations, n is the number of observations

Mean is the most commonly used statistical average. Although it is reliable, few extreme values in the data

Scheme 1 Schematic diagram to represent different types of data



distribution may pull the mean apart from the centre. It is a good average when data shows normal distribution, hence it is used in parametric statistical analyses.

Median

It is the middle value when the data is systematically arranged in either ascending or descending order. If the number of observations are odd, there is one middle value when the data is orderly arranged and that is the median. If the number of observations is in even, there are two middle values when data is orderly arranged, then median is the average of two middle values.

$$\text{Median} = \frac{m_1 + m_2}{2}$$

Median is a better average when the data shows non normal distribution or when the data is skewed. It is used in non-parametric statistical analyses.

Mode

It is the most frequent observation in a data distribution. Data may be unimodal or bimodal or trimodal depending on how many values act as modes in a data distribution.

E.g.,

(1) Mean DMFT of 10 children is given as 2,1,6,2,4,3,2,5,8,7

Mode is 2, because it is the most frequently observed value. It is 'Unimodal'

(2) Mean DMFT of 10 children is given as 3,4,6,3,2,5,2,8,7,0

There are two Modes, they are 3 and 2, because both are most frequently observed in the data, hence the data is said to be 'Bimodal'.

References

1. Park K (2011) Park's textbook of preventive and social medicine, 21st edn. M/s Banarsidas Bhanot Publishers, Jabalpur
2. Myles PS, Gin T (2000) Statistical methods for anaesthesia and intensive care. Butterworth-Heinemann, Waltham, Massachusetts