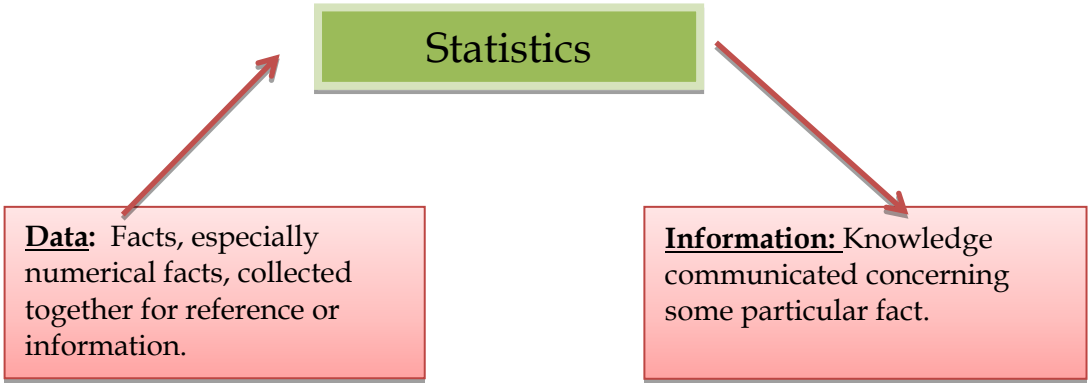
Basic Statistics

T.V.Sathianandan*, Safeena P. K and Rameez Rahman
Principal Scientist, FRAD, CMFRI
Email: tvsedpl@gmail.com

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data.



1. Key Statistical Concepts

Population: In statistics, the term population is used to mean the totality of cases (items) in an investigation.

Sample: Any subset or subgroup of a population.

Parameter: A descriptive measure of a population.

Statistic: A descriptive measure of a sample.

Variable: Some characteristic of a population or sample.

Any characteristic of an individual or entity. A variable can take different values for different individuals. Typically denoted with a capital letter: X, Y, Z… Eg: - student grades. Following are the different types of variables.

- ***Quantitative and Qualitative :***

Quantative Variables that have are measured on a numeric or quantitative scale Eg : country's population, a person's shoe size, or a car's speed. Qualitative variables that is not numerical. It describes data that fits into categories. Eg: Eye colors (variables include: blue, green, brown, hazel).

- ***Discrete and Continuous :***

Discrete variable is a variable whose value is obtained by counting. Continuous variable is a variable whose value is obtained by measuring.

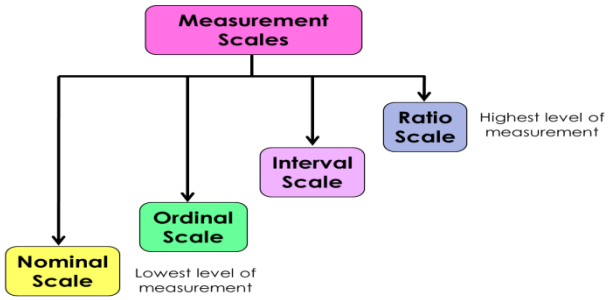
- ***Dependent and independent :***

A variable whose value depends on that of another. A variable whose variation does not depend on that of another.

Data: Observed *values* of a variable.

Following are the different Data types.

- ***Nominal*:** Categorical variables with no inherent order or ranking sequence such as names or classes (e.g., gender). Value may be a numerical, but without numerical value (e.g., I, II, III). The only operation that can be applied to Nominal variables is enumeration.
- ***Ordinal:*** Variables with an inherent rank or order, e.g. mild, moderate, severe. Can be compared for equality, or greater or less, but not *how much* greater or less.



- ***Interval:*** Values of the variable are ordered as in Ordinal, and additionally, differences between values are meaningful, however, the scale is not absolutely anchored.
  Eg : Calendar dates and temperatures on the Fahrenheit scale
- ***Ratio:*** Variables with all properties of Interval plus an absolute, non-arbitrary zero point.
  Eg : age, weight, temperature (Kelvin).

2. Descriptive Statistics

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.

2.1 Data collection

Information you gather can come from a range of sources. Likewise, there are a variety of techniques to use when gathering primary data. Listed below are some of the most common data collection techniques used for collecting data.

2.1.1 Interviews

- Interviews can be conducted in person or over the telephone
- Interviews can be done formally (structured), semi-structured, or informally
- Questions should be focused, clear, and encourage open-ended responses
- Interviews are mainly qualitative in nature
  Ex: - One-on-one conversations with parent of at-risk youth who can help you understand the issue.

2.1.2 Questionnaires and Surveys

- Responses can be analyzed with quantitative methods by assigning numerical values to Likert-type scales
- Results are generally easier (than qualitative techniques) to analyze
- Pre-test/Post-test can be compared and analyzed

  Ex: - Results of a satisfaction survey or opinion survey

2.1.3 Documents and Records

- Consists of examining existing data in the form of databases, meeting minutes, reports, attendance logs, financial records, newsletters, etc.
- This can be an inexpensive way to gather information, but may be an incomplete data source

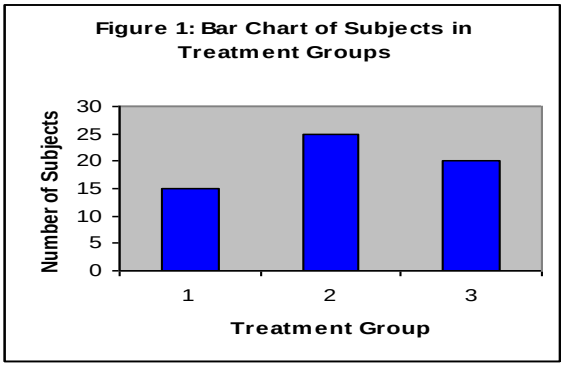
  Ex: - To understand the primary reasons students miss school, records on student absences are collected and analyzed

2.2 Data Presentation

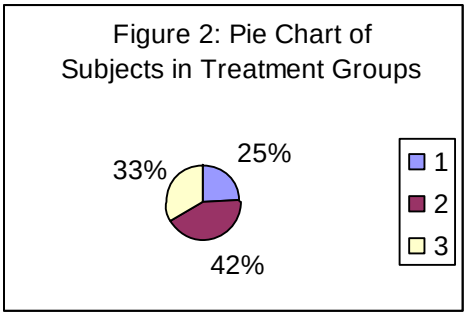
The visual representation of data may be used not only to present results/findings in the data, but may also be used to learn about the data.

2.2.1 Graphical Method

***A) Bar Diagram*: A** bar graph is a chart that uses bars to show comparisons between categories of data.



B) Pie Diagram :A pie chart displays data, information, and statistics in an easy-to-read 'pie-slice' format with varying slice sizes telling you how much of one data element exists. The bigger the slice, the more of that particular data was gathered. The main use of a pie chart is to show comparison.



C) Line chart **:**A line chart or line graph is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments.

Age	1	2	3	4	5	6
Frequency	5	3	7	5	4	2

2.2.2 *Tabular method*

Age	1	2	3	4	5	6
Frequency	5	3	7	5	4	2
Cumulative Frequency	5	8	15	20	24	26

Frequency table

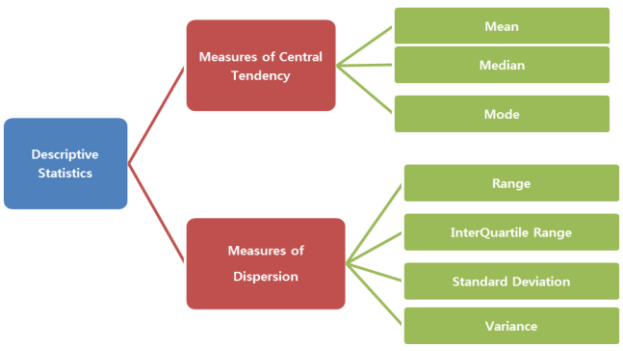
The only allowable calculation on nominal data is to count the frequency of each value of the variable. We can summarize the data in a table that presents the categories and their counts called a frequency distribution. A relative frequency distribution lists the categories and the proportion with which each occurs.

Cumulative frequency

It defined as a running total of frequencies. The frequency of an element in a set refers to how many of that element there are in the set. Cumulative frequency can also defined as the sum of all previous frequencies up to the current point.

2.3 Characterize Data/Summary Measures

Summary measures summarizeand provide information about your sample data. It tells you something about the values in your data set. This includes where the average lies and whether your data is skewed. Summary measures fall into two main categories. They are Measures of location (also



called central tendency or central measurement) and Measures of spread (also called Measures of variation) .*2.3.1 Methods of central Measurement*

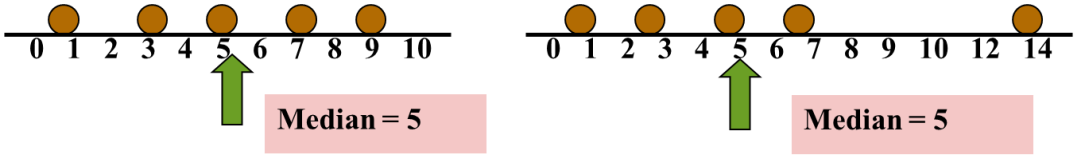
A) Mean

Mean is the average of all values of a variable and is computed by summing all the scores and dividing it by the number of scores.it does not have to be an observed value.it tends to be the most stable estimate of the population mean from sample mean. Mean of 20, 30, 40 is (20+30+40)/3 = 30.

Notation : Let $x_1, x_2, \ldots x_n$ are n observations of a variable x. Then the mean of this variable,

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

B) Median

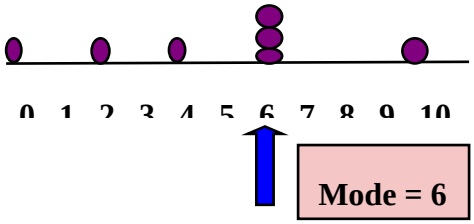


The median is the "middle" observation when the data are arranged in ascending or descending order. The median does not have to be an observable value.

- If number of observation is odd, the median is the middle number
- If number of observation is even, the median is the average of the two middle numbers

C) Mode

- The value which is most frequent.it represents the most common response.
  - Used for either numerical or categorical data
  - There may be no mode
  - There may be several modes



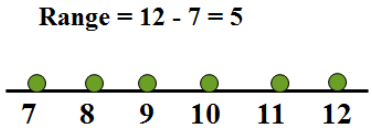
2.3.2 Measures of variation

Variability refers to how spread out a group of data is. In other words, variability measures how much your scores differ from each other. Variability is also referred to as dispersion or spread. Data sets with similar values are said to have little variability, while data sets that have values that are spread out have high variability. Following are the common measures of variation

A) Range

The range is the simplest measure of variability to calculate. The range is simply the highest score minus the lowest score.

$$\text{Range} = X_{\text{Largest}} - X_{\text{Smallest}}$$



B) Variance

The variance of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of the n observations $x_1, x_2,\ldots,x_n$ is

$$S^2 = \frac{(x_1-\bar{x})^2 + \ldots + (x_n-\bar{x})^2}{n-1}$$

Eg: Variance of 5, 7, 3? Mean is (5+7+3)/3 = 5 and the variance is $\frac{(5-5)^2+(3-5)^2+(7-5)^2}{3-1} = 4$

C) Standard Deviation

Square root of the variance. The standard deviation of the above example is 2.

Workout example

Example - The following sample consists of the number of jobs six randomly selected students applied for: 17, 15, 23, 7, 9, 13.Find the sample mean, variance and standard deviation.

Sample mean

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{17+15+23+7+9+13}{6} = \frac{84}{6} = 14\ jobs$$

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{1}{6-1}\left[(17-14)^2+(15-14)^2+\ldots(13-14)^2\right] = 33.2$$

Sample Variance (Short-cut method)

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right] = \frac{1}{6-1}\left[\left(17^2+15^2+\ldots+13^2\right) - \frac{(17+15+\ldots+13)^2}{6}\right] = 33.2$$

Standard Deviation

S= √33.2