STATISTICS : Informed Decisions Using Data

Chapter - 1 : Data Collection

Shivam Agrawal

13 March 2019

Defination 1

Statistics is the science of collecting, organizing, summarizing, and analyzing information to draw conclusions or answer questions. In addition, statistics is about providing a measure of confidence in any conclusions.

Defination 2

The entire group to be studied is called the **population**. An **individual** is a person or object that is a member of the population being studied. A **sample** is a subset of the population that is being studied.

Defination 3

A **statistic** is a numerical summary of a sample. **Descriptive statistics** consist of organizing and summarizing data. Descriptive statistics describe data through numerical summaries, tables, and graphs.

Defination 4

Inferential statistics uses methods that take a result from a sample, extend it to the population, and measure the reliability of the result.

Defination 5

A **parameter** is a numerical summary of a population.

Defination 6

Qualitative, or categorical, variables allow for classification of individuals based on some attribute or characteristic.

Quantitative variables provide numerical measures of individuals. The values of a quantitative variable can be added or subtracted and provide meaningful results.

Defination 7

A **discrete variable** is a quantitative variable that has either a finite number of possible values or a countable number of possible values. The term *countable* means that the values result from counting, such as 0,1,2,3, and so on. A discrete variable cannot take on every possible value between any two possible values.

A **continuous variable** is a quantitative variable that has an infinite number of possible values that are not countable. A continuous variable may take on every possible value between any two values.

Defination 8

A variable is at the **nominal level of measurement** if the values of the variable name, label, or categorize. In addition, the naming scheme does not allow for the values of the variable to be arranged in a ranked or specific order.

A variable is at the **ordinal level of measurement** if it has the properties of the nominal level of measurement, however the naming scheme allows for the values of the variable to be arranged in a ranked or specific order.

A variable is at the **interval level of measurement** if it has the properties of the ordinal level of measurement and the differences in the values of the variable have meaning. A value of zero does not mean the absence of the quantity. Arithmetic operations such as addition and subtraction can be performed on values of the variable.

A variable is at the **ratio level of measurement** if it has the properties of the interval level of measurement and the ratios of the values of the variable have meaning. A value of zero means the absence of the quantity. Arithmetic operations such as multiplication and division can be performed on the values of the variable.

Defination 9

An **observational study** measures the value of the response variable without attempting to influence the value of either the response or explanatory variables. That is, in an observational study, the researcher observes the behavior of the individuals without trying to influence the outcome of the study.

Defination 10

If a researcher assigns the individuals in a study to a certain group, intentionally changes the value of an explanatory variable, and then records the value of the variable for each group, the study is a **designed experiment**.

Defination 11

Confounding in a study occurs when the effects of two or more explanatory variables are not separated. Therefore, any relation that may exist between an explanatory variable and the response variable may be due to some other variable or variables not accounted for in the study.

Defination 12

A **lurking variable** is an explanatory variable that was not considered in a study, but that affects the value of the response variable in the study. In addition, lurking variables are typically related to explanatory variables considered in the study.

Defination 13

A **confounding variable** is an explanatory variable that was considered in a study whose effect cannot be distinguished from a second explanatory variable in the study.

Defination 14

A **census** is a list of all individuals in a population along with certain characteristics of each individual.

Defination 15

Random sampling is the process of using chance to select individuals from a population to be included in the sample.

Defination 16

A sample of size n from a population of size N is obtained through **simple random sampling** if every possible sample of size n has an equally likely chance of occurring. The sample is then called a **simple random sample**.

Example 1

Find the simple random sample of 5 numbers out of 1 to 30.

```
set.seed(34)   # Setting the random seed
sample(30, 5, replace = F)
```

```
## [1] 14 29 25  7  6
```

Defination 17

A **stratified sample** is obtained by separating the population into nonoverlapping groups called *strata* and then obtaining a simple random sample from each stratum. The individuals within each stratum should be homogeneous (or similar) in some way.

Defination 18

A **systematic sample** is obtained by selecting every *k*th individual from the population. The first individual selected corresponds to a random number between 1 and k.

Defination 19

A **cluster sample** is obtained by selecting all individuals within a randomly selected collection or group of individuals.

Defination 20

A **convenience sample** is a sample in which the individuals are easily obtained and not based on randomness.

Defination 21

If the results of the sample are not representative of the population, then the sample has **bias**.

Defination 22

Nonsampling errors result from undercoverage, nonresponse bias, response bias, or data-entry error. Such errors could also be present in a complete census of the population. **Sampling error** results from using a sample to estimate information about a population. This type of error occurs because a sample gives incomplete information about a population.

Defination 23

An **experiment** is a controlled study conducted to determine the effect varying one or more explanatory variables or **factors** has on a response variable. Any combination of the values of the factors is called a **treatment**.

Defination 24

In **single-blind** experiments, the experimental unit (or subject) does not know which treatment he or she is receiving. In **double-blind** experiments, neither the experimental unit nor the researcher in contact with the experimental unit knows which treatment the experimental unit is receiving.

Defination 25

A **completely randomized design** is one in which each experimental unit is randomly assigned to a treatment.

Defination 26

A **matched-pairs design** is an experimental design in which the experimental units are paired up. The pairs are selected so that they are related in some way (that is, the same person before and after a treatment, twins, husband and wife, same geographical location, and so on). There are only two levels of treatment in a matched-pairs design.

Defination 27

Grouping together similar (homogeneous) experimental units and then randomly assigning the experimental units within each group to a treatment is called **blocking**. Each group of homogeneous individuals is called a **block**.

Defination 28

A **randomized block design** is used when the experimental units are divided into homogeneous groups called blocks. Within each block, the experimental units are randomly assigned to treatments.