

Sample, observations and variables

Statistics usually collects information from a group of subjects (which could be individuals, diseases, cars, or any other types of subjects). Statistical methods require data from many subjects in order to provide a summary about the whole group. This means that in statistics we have information (data) about subjects (individuals), samples and populations.

The population

The population about which we want to do a study. Population in statistics is not necessarily about individuals but it can be any collection of objects. For example a population could be all population of Sulaimani in one study, in another study it could be only the children, and yet in another it could be pregnant women in a particular quarter. In a study on contamination of drinking water, the population could be houses of the town; in a study on prevalence of hepatitis B, the population could be a set of blood samples from the blood bank. Before doing any study we have to define the target population and the study population.

The ***target population*** is the larger population on which you would like to generalize the results of your study and therefore which includes your study population. The ***study population*** is part of the target population from which we select people who participate in the study. The study population is more limited in number and more accessible for the research than the target population. For example in a study on anemia during pregnancy at a Primary Health Centre, the target population is all pregnant women at the catchment area of the PHC and the study population could be all pregnant women who are registered at the PHC.

The sample

In epidemiological studies, we are usually not able to study the whole population and collect information about them because this will be costly and beyond our resources. Instead we collect data from a much smaller group of the population which is called a sample. A sample therefore is part of the study population and of the target population. In order to be a good example of the population, a sample should be representative to the population from which it is drawn i.e. the characteristics of the subjects in the sample should reflect characteristics of the target population. If a sample is not representative to the target population, we cannot generalize findings of our study to the target population. There are many ways to select a representative sample, but the basic principle is chance. Chance should decide who will be included in the sample not the judgment of the investigator.

Observation (record, case)

A sample is composed of many individuals. Each single individual or object in the sample on which we do measurements is called an observation. One observation in the data contains all information we have collected about one individual. For example if we have done a study on blood donors, each blood donor will be one observation in the dataset and this will contain all information we have about this particular blood donor.

Variable

Any aspect (characteristic) of the individual on which measurement is done is called a variable. This characteristic is called a variable because it can take different (variable)

values. For each observation, we will have several variables. For example for one blood donor we will have variables for sex, age, blood group, weight etc. Usually we will have the same number of variables for all observations i.e. we collect the same information from all individuals.

Values

Variable contain data. A value is the result of the measurement of the variable. For example for age variable of the donors we might have values ranging from 20 to 50 years.

The relationship of population to data could be drawn in the following way:

Target population → Study population → sample → observations → variables → values

The following tables of an imaginary study on the students of the medical college shows the meaning of each of the above terms and their relationship with each other. All students comprise the *population*, the selected number of students represent the *sample*, each interviewed student will be an *observation*. The characteristics of that student will be the *variables* which will have *values*.

Population	Sample	Observation	Variables	Values	
All students of Medical College	100 students included in data collection	Student 1	Age	18, 19, 20, ...	
			Sex	Male/female	
			Weight	54, 54.2,...	
			Height	160, 161, ...	
		Student 2	father's education	None, primary, intermediate, secondary, university	
			Age	18, 19, 20, ...	
			Sex	Male/female	
			Weight	54, 54.2,...	
			Height	160, 161, ...	
		Students 3	father's education	None, primary, intermediate, secondary, university	
			...		
		... Student 100			