

Statistics

What is Statistics?

Statistics: The science of collecting, describing, and interpreting data.

Two areas of statistics:

Descriptive Statistics: collection, presentation, and description of sample data.

Inferential Statistics: making decisions and drawing conclusions about populations.

Example: A recent study examined the math and verbal SAT scores of high school seniors across the country. Which of the following statements are descriptive in nature and which are inferential.

- The mean math SAT score was 492.
- The mean verbal SAT score was 475.
- Students in the Northeast scored higher in math but lower in verbal.
- 80% of all students taking the exam were headed for college.
- 32% of the students scored above 610 on the verbal SAT.
- The math SAT scores are higher than they were 10 years ago.

Introduction to Basic Terms

Population: A collection, or set, of individuals or objects or events whose properties are to be analyzed.

Two kinds of populations: *finite* or *infinite*.

Sample: A subset of the population.

Variable: A characteristic about each individual element of a population or sample.

Data (singular): The value of the variable associated with one element of a population or sample. This value may be a number, a word, or a symbol.

Data (plural): The set of values collected for the variable from each of the elements belonging to the sample.

Experiment: A planned activity whose results yield a set of data.

Parameter: A numerical value summarizing all the data of an entire population.

Statistic: A numerical value summarizing the sample data.

Example: A college dean is interested in learning about the average age of faculty. Identify the basic terms in this situation.

The *population* is the age of all faculty members at the college.

A *sample* is any subset of that population. For example, we might select 10 faculty members and determine their age.

The *variable* is the “age” of each faculty member.

One *data* would be the age of a specific faculty member.

The *data* would be the set of values in the sample.

The *experiment* would be the method used to select the ages forming the sample and determining the actual age of each faculty member in the sample.

The *parameter* of interest is the “average” age of all faculty at the college.

The *statistic* is the “average” age for all faculty in the sample.

Two kinds of variables:

Qualitative, or Attribute, or Categorical,

Variable: A variable that categorizes or describes an element of a population.

Note: Arithmetic operations, such as addition and averaging, are *not* meaningful for data resulting from a qualitative variable.

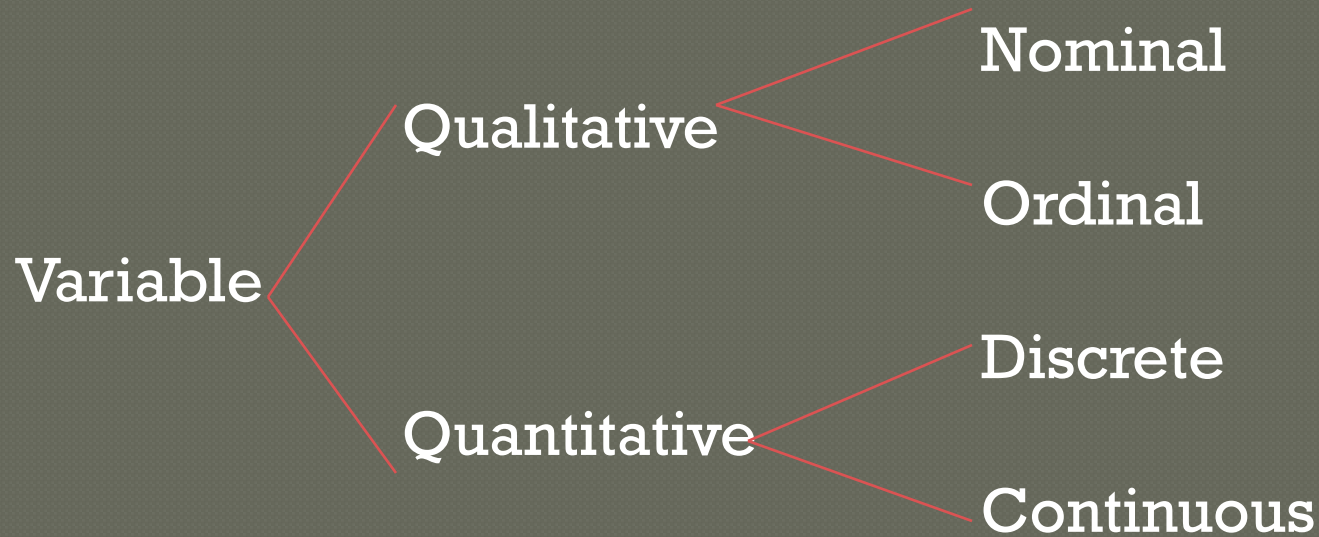
Quantitative, or Numerical, Variable: A variable that quantifies an element of a population.

Note: Arithmetic operations such as addition and averaging, are meaningful for data resulting from a quantitative variable.

Example: Identify each of the following examples as attribute (qualitative) or numerical (quantitative) variables.

1. The residence hall for each student in a statistics class. (Attribute)
2. The amount of gasoline pumped by the next 10 customers at the local Unimart. (Numerical)
3. The amount of radon in the basement of each of 25 homes in a new development. (Numerical)
4. The color of the baseball cap worn by each of 20 students. (Attribute)
5. The length of time to complete a mathematics homework assignment. (Numerical)
6. The state in which each truck is registered when stopped and inspected at a weigh station. (Attribute)

Qualitative and quantitative variables may be further subdivided:



Nominal Variable: A qualitative variable that categorizes (or describes, or names) an element of a population.

Ordinal Variable: A qualitative variable that incorporates an ordered position, or ranking.

Discrete Variable: A quantitative variable that can assume a countable number of values. Intuitively, a discrete variable can assume values corresponding to isolated points along a line interval. That is, there is a gap between any two values.

Continuous Variable: A quantitative variable that can assume an uncountable number of values. Intuitively, a continuous variable can assume any value along a line interval, including every possible value between any two values.

Note:

1. In many cases, a discrete and continuous variable may be distinguished by determining whether the variables are related to a count or a measurement.
2. Discrete variables are usually associated with counting. If the variable cannot be further subdivided, it is a clue that you are probably dealing with a discrete variable.
3. Continuous variables are usually associated with measurements. The values of discrete variables are only limited by your ability to measure them.

Example: Identify each of the following as examples of qualitative or numerical variables:

1. The temperature in Barrow, Alaska at 12:00 pm on any given day.
2. The make of automobile driven by each faculty member.
3. Whether or not a 6 volt lantern battery is defective.
4. The weight of a lead pencil.
5. The length of time billed for a long distance telephone call.
6. The brand of cereal children eat for breakfast.
7. The type of book taken out of the library by an adult.

Example: Identify each of the following as examples of (1) nominal, (2) ordinal, (3) discrete, or (4) continuous variables:

1. The length of time until a pain reliever begins to work.
2. The number of chocolate chips in a cookie.
3. The number of colors used in a statistics textbook.
4. The brand of refrigerator in a home.
5. The overall satisfaction rating of a new car.
6. The number of files on a computer's hard disk.
7. The pH level of the water in a swimming pool.
8. The number of staples in a stapler.

Measure and Variability

- No matter what the response variable: there will always be **variability** in the data.
- One of the primary objectives of statistics: measuring and characterizing variability.
- Controlling (or reducing) variability in a manufacturing process: statistical process control.

Example: A supplier fills cans of soda marked 12 ounces. How much soda does each can really contain?

- It is very *unlikely* any one can contains exactly 12 ounces of soda.
- There is variability in any process.
- Some cans contain a little more than 12 ounces, and some cans contain a little less.
- On the average, there are 12 ounces in each can.
- The supplier hopes there is little variability in the process, that most cans contain *close* to 12 ounces of soda.

Data Collection

- First problem a statistician faces: how to obtain the data.
- It is important to obtain *good*, or *representative*, data.
- Inferences are made based on statistics obtained from the data.
- Inferences can only be as good as the data.

Biased Sampling Method: A sampling method that produces data which systematically differs from the sampled population. An **unbiased sampling method** is one that is not biased.

Sampling methods that often result in biased samples:

1. **Convenience sample:** sample selected from elements of a population that are easily accessible.
2. **Volunteer sample:** sample collected from those elements of the population which chose to contribute the needed information on their own initiative.

Process of data collection:

1. Define the objectives of the survey or experiment.

Example: Estimate the average life of an electronic component.

2. Define the variable and population of interest.

Example: Length of time for anesthesia to wear off after surgery.

3. Defining the data-collection and data-measuring schemes. This includes sampling procedures, sample size, and the data-measuring device (questionnaire, scale, ruler, etc.).

4. Determine the appropriate descriptive or inferential data-analysis techniques.

Methods used to collect data:

Experiment: The investigator controls or modifies the environment and observes the effect on the variable under study.

Survey: Data are obtained by sampling some of the population of interest. The investigator does not modify the environment.

Census: A 100% survey. Every element of the population is listed. Seldom used: difficult and time-consuming to compile, and expensive.

Sampling Frame: A list of the elements belonging to the population from which the sample will be drawn.

Note: It is important that the sampling frame be representative of the population.

Sample Design: The process of selecting sample elements from the sampling frame.

Note: There are many different types of sample designs. Usually they all fit into two categories: judgment samples and probability samples.

Judgment Samples: Samples that are selected on the basis of being “typical.”

Items are selected that are representative of the population. The validity of the results from a judgment sample reflects the soundness of the collector’s judgment.

Probability Samples: Samples in which the elements to be selected are drawn on the basis of probability. Each element in a population has a certain probability of being selected as part of the sample.

Random Samples: A sample selected in such a way that every element in the population has an equal probability of being chosen. Equivalently, all samples of size n have an equal chance of being selected. Random samples are obtained either by sampling with replacement from a finite population or by sampling without replacement from an infinite population.

Note:

1. Inherent in the concept of randomness: the next result (or occurrence) is not predictable.
2. Proper procedure for selecting a random sample: use a random number generator or a table of random numbers.

Example: An employer is interested in the time it takes each employee to commute to work each morning. A random sample of 35 employees will be selected and their commuting time will be recorded.

There are 2712 employees.

Each employee is numbered: 0001, 0002, 0003, etc. up to 2712.

Using four-digit random numbers, a sample is identified: 1315, 0987, 1125, etc.

Systematic Sample: A sample in which every k th item of the sampling frame is selected, starting from the first element which is randomly selected from the first k elements.

Note: The systematic technique is easy to execute. However, it has some inherent dangers when the sampling frame is repetitive or cyclical in nature. In these situations the results may not approximate a simple random sample.

Stratified Random Sample: A sample obtained by stratifying the sampling frame and then selecting a fixed number of items from each of the strata by means of a simple random sampling technique.

Proportional Sample (or Quota Sample): A sample obtained by stratifying the sampling frame and then selecting a number of items in proportion to the size of the strata (or by quota) from each strata by means of a simple random sampling technique.

Cluster Sample: A sample obtained by stratifying the sampling frame and then selecting some or all of the items from some of, but not all, the strata.

Comparison of Probability and Statistics

Probability: Properties of the population are assumed known. Answer questions about the sample based on these properties.

Statistics: Use information in the sample to draw a conclusion about the population.

Example: A jar of M&M's contains 100 candy pieces, 15 are red. A handful of 10 is selected.

Probability question: What is the probability that 3 of the 10 selected are red?

Example: A handful of 10 M&M's is selected from a jar containing 1000 candy pieces. Three M&M's in the handful are red.

Statistics question: What is the proportion of red M&M's in the entire jar?

Statistics and the Technology

- The electronic technology has had a tremendous effect on the field of statistics.
- Many statistical techniques are repetitive in nature: computers and calculators are good at this.
- Lots of statistical software packages: MINITAB, SYSTAT, STATA, SAS, Statgraphics, SPSS, and calculators.