

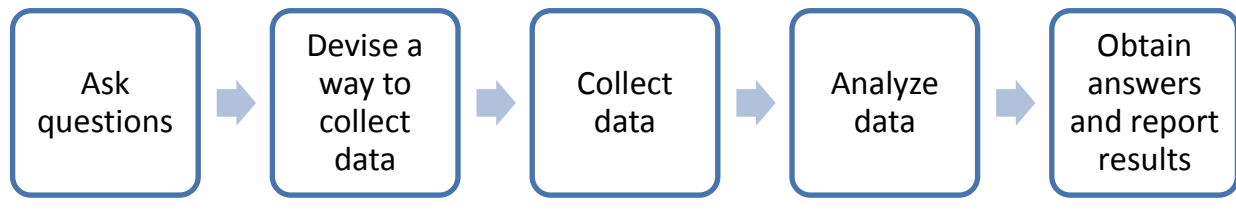
Chapter 1: Introduction

What is Statistics?

Statistics: The science of the collection, analysis and presentation of data.

Data: Information collected about people or things.

The investigative process:



Variable: The characteristic of interest for each person or thing in a population.

Types of Variables

- Qualitative variables can take on values that are names or labels.
- Quantitative variables can take on numeric values. There are two types of quantitative variables: discrete and continuous.
 - Discrete variables can take on a countable and/or finite number of values (e.g. counts)
 - Continuous variables can take on an infinite number of possible values in some interval (e.g. measurements)

Examples: Are the following qualitative, quantitative-discrete, or quantitative-continuous variables?

- Blood pressure in mm of mercury (*quantitative-continuous*)
- The number of heads in ten tosses of a coin (*quantitative-discrete*)
- Gender (*qualitative*)
- Whether or not someone is a smoker (*qualitative*)
- Height (*quantitative-continuous*)
- Weight (*quantitative-continuous*)
- Religion (*qualitative*)
- The number of tomatoes produced on a plant (*quantitative-discrete*)
- The head circumferences of standard poodles (*quantitative-continuous*)
- The length of time it takes to complete an obstacle course (*quantitative-continuous*)

- Color preference (1=Red, 2=Blue, 3=Green) (*qualitative*)

There are four main methods of data collection:

1. Census
2. Sample survey
3. Experiment
4. Observational study

Census: A census is a study that obtains data from every member of a population, which is a complete collection of all people or things under study. In most studies, a census is not practical, because of the cost and time required.

Sample Surveys: A sample survey is a study that obtains data from a subset of a population, called a sample, and generalizes results to the larger population.

Sampling Terminology:

- Parameter: a number that describes a population.
- Statistic: a number that describes a sample.

A well-designed sample survey can provide very precise estimates of population parameters. As long as members of a sample are randomly selected from a larger population, it is appropriate to generalize the results to the larger population.

Example: Suppose that a researcher wants to estimate the average amount of money that full-time college students spend on textbooks each year. A sample of 500 full-time students was selected, and for the sample, the average amount spent was \$841.

- a. What is the population?
All full-time college students.
- b. What is the sample?
The group of 500 college students.
- c. What is the variable?
The amount spent on textbooks.
- d. What is the parameter that is being estimated?
The average amount of money spent by all full-time college students.

- e. What is the value of the statistic?
\$841

Example: Before a recent election, a study was conducted to estimate the proportion of Massachusetts residents that would support a reduction in the state sales tax. Twenty-six hundred Massachusetts residents were surveyed and asked the question “Yes or no: Would you support a reduction in the state sales tax.” 65% of those surveyed stated that they supported a reduction.

- a. What is the population?
All Massachusetts residents.
- b. What is the sample?
The group of 2600 Massachusetts residents.
- c. What is the variable?
Whether or not a resident supported a reduction in the sales tax.
- d. What is the parameter that is being estimated?
The proportion of all Massachusetts residents that would support a reduction in the state sales tax.
- e. What is the value of the statistic?
65% (or 0.65)

Important notes about sampling:

- Ideally, a sample should be representative of the population. A sample that is not representative of the population is biased.
- The random error obtained by using part of the population to represent the whole population is called sampling error.
- The non-random error caused by improper data collection recording or sampling techniques, is called non-sampling error.

Random Sampling

With random sampling, each member of a population has an equal chance of being selected.

Sampling Methods

- Simple random sampling: Every possible sample of the same size has the same chance of being chosen.

- Systematic sampling: Select some starting point and then select every kth element in the population.
- Stratified sampling: Subdivide the population into at least two different subgroups (or strata) that share the same characteristics, then draw a simple random sample from each subgroup.
- Cluster sampling: Divide the population area into sections (or clusters), then randomly select some of those clusters, then choose all of the members from those selected clusters.
- Convenience sampling: Use results that are easy to get

Examples: What type of sampling procedure has been used?

1. Ask your family members for whom they plan to vote in the upcoming election. (*convenience*)
2. To conduct a study on malpractice rates, subdivide doctors into four groups—surgeons, family practitioners, obstetricians, and others. Take a sample of doctors within each group. (*stratified*)
3. Randomly select 10 neighborhoods in a city, and then ask all of the residents their opinion about the new mayor. (*cluster*)
4. Every 30th iPod is selected from an assembly line to check for defects. (*systematic*)
5. To gauge students' attitudes toward alcohol use, divide students into two groups, male and female, and sample 200 students within each group. (*stratified*)
6. Every 30th iPod is selected from an assembly line to check for defects. (*systematic*)

Experiments

Experiments deliberately impose some treatment on individuals in order to observe their response. This response is typically compared to that of a control group, which is a group of individuals who are given no treatment or a placebo.

Subjects are typically assigned to treatment and control groups using a random process. Properly designed and executed experiments can demonstrate a cause-and-effect relationship.

Observational Studies

Observational study: observe and measure specific characteristics but don't attempt to modify the subjects being studied. Observational studies cannot establish a cause-and-effect relationship.

Why are observational studies used?

- A randomized experiment would violate ethical standards.
- A randomized experiment may be impractical or impossible.

Example: Are the following experiments or observational studies?

1. Question: What are some trends regarding life jacket use by recreational boaters?
Study: 1,000 boaters were surveyed regarding type of boat, gender, age, water conditions (choppy/calm), and life jacket wear (yes/no). (*observational study*)
2. Question: Does math-tutoring software improve the algebra skills of high school students?
Study: After a six-month period, the algebra skills of 80 students who used the tutoring software were compared to the algebra skills of 80 students who did not use the software. (*experiment*)
3. Question: What are the cause of heart disease?
Study: Framingham Heart Study: 50 years of data involving thousands of Framingham residents is analyzed to determine which biologic and environmental led to heart disease and stroke. The study found that high cholesterol, high blood pressure, diabetes, and cigarette smoking lead to heart disease. (*observational study*)
4. Question: Is a new blood pressure medication effective?
5. Study: Two hundred people with high blood pressure were randomly assigned to two groups. One group received a new blood pressure medication, and the other group received a placebo. After a period of time, blood pressure measurements were recorded for each group. (*experiment*)

Problems with Statistical Studies

- Biased samples (A sample should be representative of the population.)
- Self-selected samples
- Sample size issues
- Collecting data or asking questions in a way that influences the response
- Non-response or refusal of subject to participate
- Causality: a relationship between two variables does not necessarily imply that one causes the other to occur; they may both be related to some other variable.
- Self-Funded or Self-Interest Studies
- Misleading Use of Data: improperly displayed graphs, incomplete data, lack of context
- Confounding: when the effects of multiple factors on a response cannot be separated. it becomes difficult or impossible to draw valid conclusions about the effect of each factor.