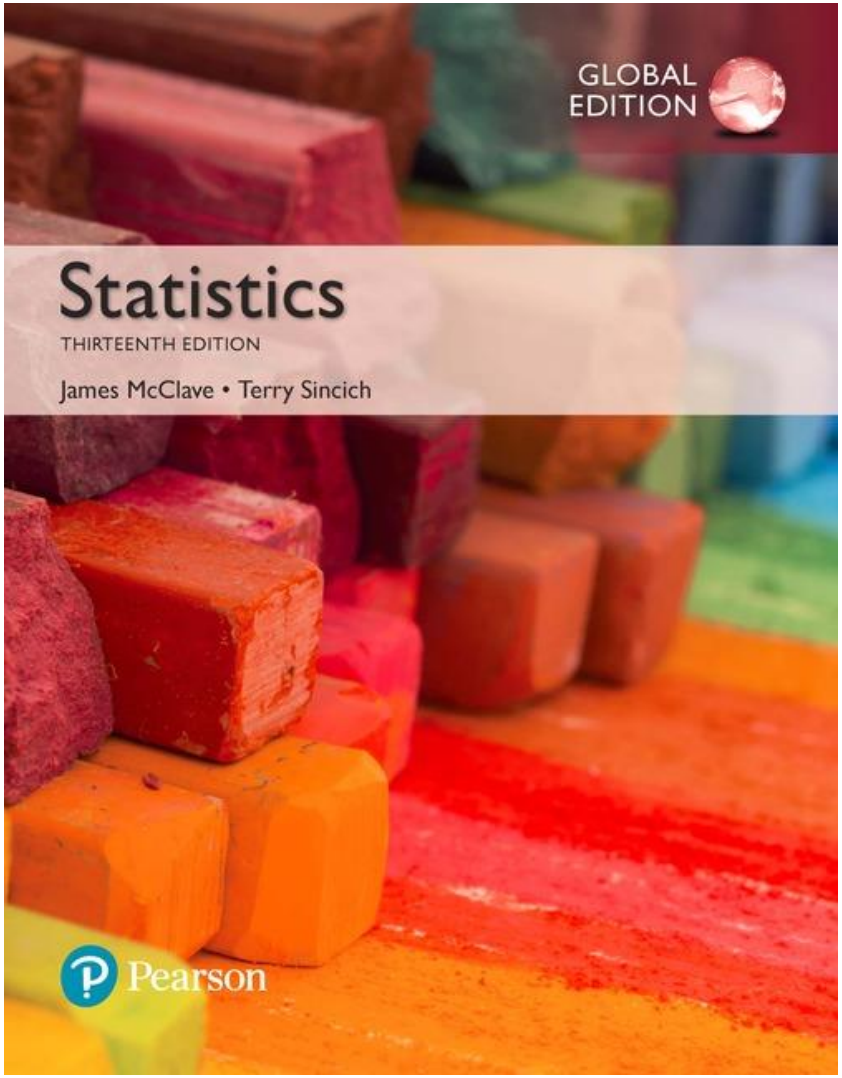


Chapter 1

Statistics, Data, and Statistical Thinking

1.1

The Science of Statistics

Definition

Statistics is the science of data. This involves collecting, classifying, summarizing, organizing, analyzing, presenting, and interpreting numerical and categorical information.

1.2

Types of Statistical Applications

Definition

faydalanmak

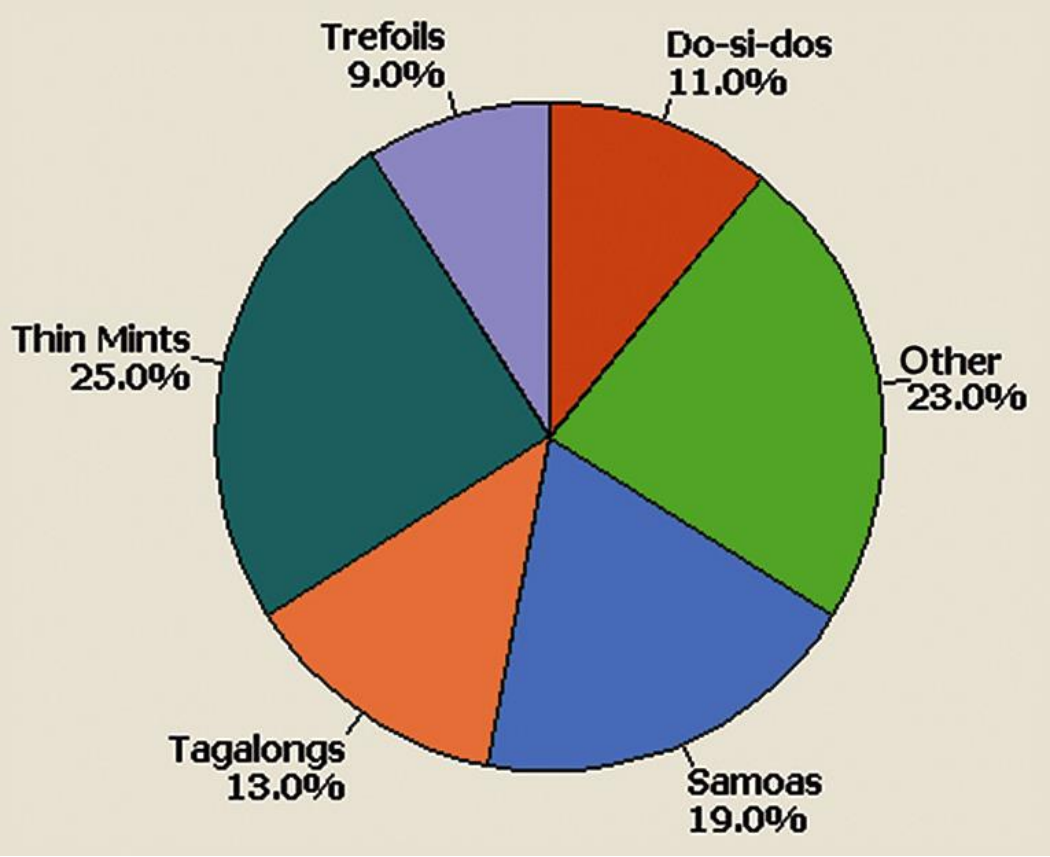
Descriptive statistics utilizes numerical and graphical methods to look for patterns in a data set, to summarize the information revealed in a data set, and to present that information in a convenient form.

uygun

Definition

Inferential statistics utilizes sample data to make estimates, decisions, predictions, or other generalizations about a larger set of data.

Figure 1.1 MINITAB graph of the best-selling Girl Scout Cookies



1.3

Fundamental Elements of Statistics

Definition

Statistical methods are particularly useful for studying, analyzing, and learning about **populations** of **experimental units**.

> An **experimental** (or **observational**) **unit** is an object (e.g., person, thing, transaction, or event) about which we collect data.

Definition

A **population** is a set of all units (usually people, objects, transactions, or events) that we are interested in studying.

populations may include
(1) *all* employed workers in the United States,
(2) *all* registered voters in California,
(3) *everyone* who is afflicted with AIDS,
(4) *all* the cars produced last year by a particular assembly line,
(5) the *entire* stock of spare parts available at Southwest Airlines' maintenance facility,
(6) *all* sales made at the drive-in
window of a McDonald's restaurant during a given year, or
(7) the set of *all* accidents
occurring on a particular stretch of interstate highway during a holiday period

Definition

A **variable** is a characteristic or property of an individual experimental (or observational) unit in the population.

we may be interested in the variables **age**, **gender**, and **number of years of education** of the people currently unemployed in the United States.

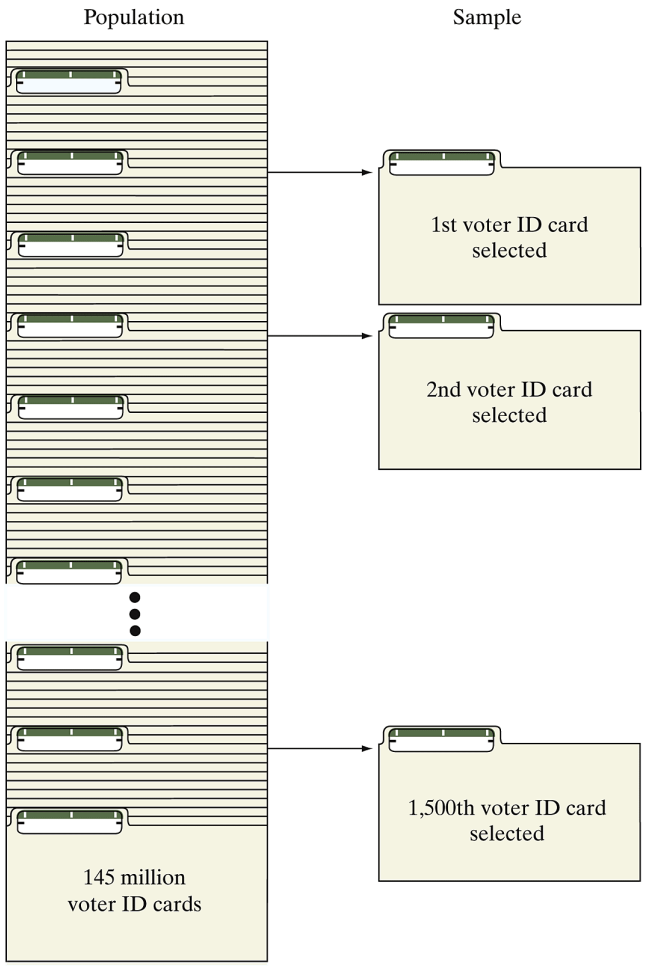
Definition

When we measure a variable for every unit of a population, it is called a **census** of the population.

A **sample** is a subset of the units of a population.

For example, instead of polling all 145 million registered voters in the United States during a presidential election year, a pollster might select and question a sample of just 1,500 voters.

Figure 1.2 A sample of voter registration cards for all registered voters



Definition

A **statistical inference** is an estimate, prediction, or some other generalization about a population based on information contained in a sample.

That is, *we use the information contained in the smaller sample to learn about the larger population.*

Example

Problem According to *Variety* (Jan. 10, 2014), the average age of Broadway ticketbuyers is 42.5 years. Suppose a Broadway theatre executive hypothesizes that the average age of ticketbuyers to her theatre's plays is less than 42.5 years. To test her hypothesis, she samples 200 ticketbuyers to her theatre's plays and determines the age of each.

- Describe the population.
- Describe the variable of interest.
- Describe the sample.
- Describe the inference.

Solution

- Describe the population.
- Describe the variable of interest.
- Describe the sample.
- Describe the inference.

- The population is the set of all units of interest to the theatre executive, which is the set of all ticketbuyers to her theatre's plays.
- The age (in years) of each ticketbuyer is the variable of interest.
- The sample must be a subset of the population. In this case, it is the 200 ticketbuyers selected by the executive.
- The inference of interest involves the *generalization* of the information contained in the sample of 200 ticketbuyers to the population of all her theatre's ticketbuyers. In particular, the executive wants to *estimate* the average age of the ticketbuyers to her theatre's plays in order to determine whether it is less than 42.5 years. She might accomplish this by calculating the average age of the sample and using that average to estimate the average age of the population.

Definition

cikarim

making the inference is only part of the story; we also need to know its **reliability**—that is, how good the inference is.
Reliability, then, is the fifth element of inferential statistical problems.

nicel

A **measure of reliability** is a statement (usually quantitative) about the degree of uncertainty associated with a statistical inference.

For now, we simply want you to realize that an inference is incomplete without a measure of its reliability.

Summary

Four Elements of Descriptive Statistical Problems

1. The population or sample of interest
2. One or more variables (characteristics of the population or sample units) that are to be investigated
3. Tables, graphs, or numerical summary tools
4. Identification of patterns in the data

Summary

Five Elements of Inferential Statistical Problems

1. The population of interest
2. One or more variables (characteristics of the population units) that are to be investigated
3. The sample of population units
4. The inference about the population based on information contained in the sample
5. A measure of the reliability of the inference

1.4

Types of Data

Definition

Quantitative data are measurements that are recorded on a naturally occurring numerical scale.

1. The temperature (in degrees Celsius) at which each piece in a sample of 20 pieces of heat-resistant plastic begins to melt
2. The current unemployment rate (measured as a percentage) in each of the 50 states
3. The scores of a sample of 150 law school applicants on the LSAT, a standardized law school entrance exam administered nationwide
4. The number of convicted murderers who receive the death penalty each year over a 10-year period

Definition

In contrast, qualitative data cannot be measured on a natural numerical scale; they can only be classified into categories.
For this reason, this type of data is also called. *categorical data*.)

Qualitative (or **categorical**) **data** are measurements that cannot be measured on a natural numerical scale; they can only be classified into one of a group of categories.

1. The political party affiliation (Democrat, Republican, or Independent) in a sample of 50 voters
2. The defective status (defective or not) of each of 100 computer chips manufactured by Intel
3. The size of a car (subcompact, compact, midsize, or full size) rented by each of a sample of 30 business travelers
4. A taste tester's ranking (best, worst, etc.) of four brands of barbecue sauce for a panel of 10 testers

1.5

Collecting Data:
Sampling and Related Issues

Collecting Data: Sampling and Related issues

Generally, you can obtain data in three different ways:

1. From a **published source**

- Turkish Statistical Institute (TÜİK)
- The Wall Street Journal (financial data)
- kaggle.com (datasets for machine learning projects)

Collecting Data: Sampling and Related issues

Generally, you can obtain data in three different ways:

2. From a **designed experiment**

> A **designed experiment** is a data collection method where the researcher exerts full control over the characteristics of the experimental units sampled. These experiments typically involve a group of experimental units that are assigned the *treatment* and an untreated (or *control*) group.

Collecting Data: Sampling and Related issues

Generally, you can obtain data in three different ways:

3. From an observational study (e.g., a survey)

An **observational study** is a data collection method where the experimental units sampled are observed in their natural setting. No attempt is made to control the characteristics of the experimental units sampled. (Examples include *opinion polls* and *surveys*.)

Definition

Regardless of which data collection method is employed, it is likely that the data will be a sample from some population. And if we wish to apply inferential statistics, we must obtain a **representative sample**.

temsili örnek

A **representative sample** exhibits characteristics typical of those possessed by the target population.

sergilemek

Definition

A **simple random sample** of n experimental units is a sample selected from the population in such a way that every different sample of size n has an equal chance of selection.

If the pollster samples 1,500 of the 145 million voters in the population so that every subset of 1,500 voters has an equal chance of being selected, she has devised a random sample.

Figure 1.3 Random Selection of 20 Households Using MINITAB

Problem Suppose you wish to assess the feasibility of building a new high school. As part of your study, you would like to gauge the opinions of people living close to the proposed building site. The neighborhood adjacent to the site has 711 homes. Use a random number generator to select a simple random sample of 20 households from the neighborhood to participate in the study.

Solution In this study, your population of interest consists of the 711 households in the adjacent neighborhood. To ensure that every possible sample of 20 households selected from the 711 has an equal chance of selection (i.e., to ensure a simple random sample), first assign a number from 1 to 711 to each of the households in the population. These numbers were entered into MINITAB. Now, apply the random number generator of MINITAB, requesting that 20 households be selected without replacement. Figure 1.3 shows the output from MINITAB. You can see that households numbered 78, 152, 157, 177, 216, . . . , 690 are the households to be included in your sample.

↓	C1	C2
	HouseNum	Sample20
1	1	78
2	2	152
3	3	157
4	4	177
5	5	216
6	6	234
7	7	242
8	8	255
9	9	262
10	10	280
11	11	323
12	12	344
13	13	406
14	14	408
15	15	480
16	16	495
17	17	549
18	18	610
19	19	645
20	20	690
21	21	
22	22	
23	23	
24	24	
25	25	

More on sampling

In addition to simple random samples, there are more complex random sampling designs that can be employed. These include (but are not limited to)

- stratified random sampling,
- cluster sampling,
- systematic sampling, and
- randomized response sampling.

Stratified random sampling

Typically used when the experimental units associated with the population can be separated into two or more groups of units, called *strata*, where the characteristics of the experimental units are more similar within strata than across strata.

- Representative samples of both Republicans and Democrats (in proportion to the number of Republicans and Democrats in the voting population)

Cluster sampling

Sometimes it is more convenient and logical to sample natural groupings (clusters) of experimental units first, then collect data from all experimental units within each cluster.

kümeler

- sample 10 of the 150 restaurant locations (clusters), then interview all customers eating at each of the 10 locations

Systematic sampling

Involves systematically selecting every *k*th experimental unit from a list of all experimental units.

- every fifth person who walks into a shopping mall could be asked whether s/he owns a smart phone
- quality control engineer at a manufacturing plant may select every 10th item produced on an assembly line for inspection.

Randomized response sampling

This design is particularly useful when the questions of the pollsters are likely to elicit false answers.

ortaya çkarmak

- Suppose each person in a sample of wage earners is asked whether he or she cheated on an income tax return. A cheater might lie, thus biasing an estimate of the true likelihood of someone cheating on his or her tax return. To circumvent this problem, each person is presented with two questions, one being the object of the survey and the other an innocuous question, such as:
- 1. Did you ever cheat on your federal income tax return?
- 2. Did you drink coffee this morning?

Definition

onyargi

Selection bias results when a subset of experimental units in the population has little or no chance of being selected for the sample.

This results in samples that are not representative of the population.

Definition

Nonresponse bias is a type of selection bias that results when data on all experimental units in a sample are not obtained.

Consider an opinion poll that employs either a telephone survey or mail survey. After collecting a random sample of phone numbers or mailing addresses, each person in the sample is contacted via telephone or the mail and a survey conducted. Unfortunately, these types of surveys often suffer from selection bias due to *nonresponse*. Some individuals may not be home when the phone rings, or others may refuse to answer the questions or mail back the questionnaire.

Definition

Measurement error refers to inaccuracies in the values of the data collected. In surveys, the error may be due to ambiguous or leading questions and the interviewer's effect on the respondent.

- "How often did you change the oil in your car last year?" (ambiguous)
- "Does the new health plan offer more comprehensive medical services at less cost than the old one?" (leading to answer «**yes**»)
  - "Which health plan offers more comprehensive medical services at less cost, the old one or the new one?" (not leading version)

Example

- **Problem** What percentage of Web users are addicted to the Internet? To find out, a psychologist designed a series of 10 questions based on a widely used set of criteria for gambling addiction and distributed them through the Web site ABCNews.com. (A sample question: “Do you use the Internet to escape problems?”) A total of 17,251 Web users responded to the questionnaire. If participants answered “yes” to at least half of the questions, they were viewed as addicted. The findings, released at an annual meeting of the American Psychological Association, revealed that 990 respondents, or 5.7%, are addicted to the Internet.
  - Identify the data collection method.
  - Identify the target population.
  - Are the sample data representative of the population?

Solution

- The data collection method is a survey: 17,251 Internet users responded to the questions posed at the ABCNews.com Web site.
- Since the Web site can be accessed by anyone surfing the Internet, presumably the target population is all Internet users.
- Because the 17,251 respondents clearly make up a subset of the target population, they do form a sample. Whether or not the sample is representative is unclear, since we are given no information on the 17,251 respondents. However, a survey like this one in which the respondents are self-selected (i.e., each Internet user who saw the survey chose whether to respond to it) often suffers from nonresponse bias. It is possible that many Internet users who chose not to respond (or who never saw the survey) would have answered the questions differently, leading to a higher (or lower) percentage of affirmative answers.

1.6

The Role of Statistics in Critical Thinking and Ethics

Definition

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." H. G. Wells

Statistical thinking involves applying rational thought and the science of statistics to critically assess data and inferences. Fundamental to the thought process is that variation exists in populations of data.

Quantitative literacy can help you make intelligent decisions, inferences, and generalizations; that is, it helps you *think critically* using statistics.

Example

Problem An article in the New York Times considered the question of whether motorcyclists should be required by law to wear helmets. In supporting his argument for no helmets, the editor of a magazine for Harley-Davidson bikers presented the results of one study that claimed “nine states without helmet laws had a lower fatality rate (3.05 deaths per 10,000 motorcycles) than those that mandated helmets (3.38)” and a survey that found “of 2,500 bikers at a rally, 98% of the respondents opposed such laws.” Based on this information, do you think it is safer to ride a motorcycle without a helmet? What further statistical information would you like?

Solution

- use “statistical thinking” to help you critically evaluate the study
  - how reliably 2,500 bikers were selected (random from the target population of all bikers?)
  - comparing the motorcycle fatality rate
    - Were the data obtained from a published source?
    - Were all 50 states included in the study, or were only certain states selected? That is, are you seeing sample data or population data?
    - Do the helmet laws vary among states?
    - If so, can you really compare the fatality rates?