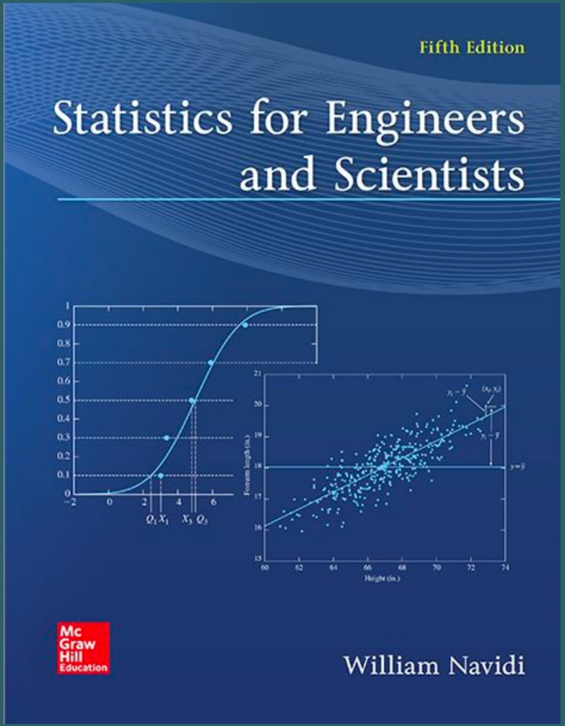
Mc Graw Hill Education



Fifth Edition

Statistics for Engineers and Scientists

William Navidi

Chapter 1

Sampling and Descriptive Statistics (part 1)

Outline of Chapter 1 (p.1-47)

1-1 Sampling

1-2 Summary Statistics

1-3 Graphical Summaries

Introduction

Statistics is the science of conducting studies to

- collect,
- organize,
- summarize,
- analyze, and
- draw conclusions from data.

Introduction (cont.)

- Statistics is used in science, engineering and almost all fields of human life. Examples:

1. In sports, a statistician may keep records of the number of points a basketball player gets in a season.
2. In public health, an administrator might be concerned with the number of residents who contract a new strain of flu virus during a certain year.
3. In education, a researcher might want to know if new methods of teaching are better than old ones.

Introduction

- Statistics is used to analyze the results of surveys and as a tool in scientific research to make decisions based on controlled experiments.
- Other uses of statistics include operations research, quality control, estimation, and prediction.

Descriptive and Inferential Statistics

- **Descriptive statistics** consists of the collection, organization, summarization, and presentation of data (Ch 1-6 in the textbook).
- **Inferential statistics** consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions (Ch 7-10 in the textbook).

Example 1: Descriptive and Inferential Statistics

The average price of a 30-second ad for the Academy Awards show in a recent year was 1.90 million dollars.

Example 2: Descriptive and Inferential Statistics

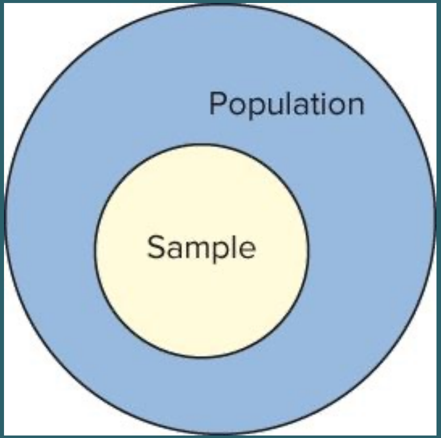
The Department of Economic and Social Affairs predicts that the population of Mexico City, Mexico, in 2030 will be 238,647,000 people.

Example 3: Descriptive and Inferential Statistics

A survey of 2234 people conducted by the Harris Poll found that 55% of the respondents said that excessive complaining by adults was the most annoying social media habit.

Basic Statistical Terms

- A **variable** is a characteristic or attribute that can assume different values.
- Variables whose values are determined by chance are called random variables.
- The values that a variable can assume are called **data**.
- A **population** consists of **ALL** subjects (human or otherwise) that are studied.
- A **sample** is a subset of the population.



Census versus Sample

Census

- Collection of data from ***every*** member of a population

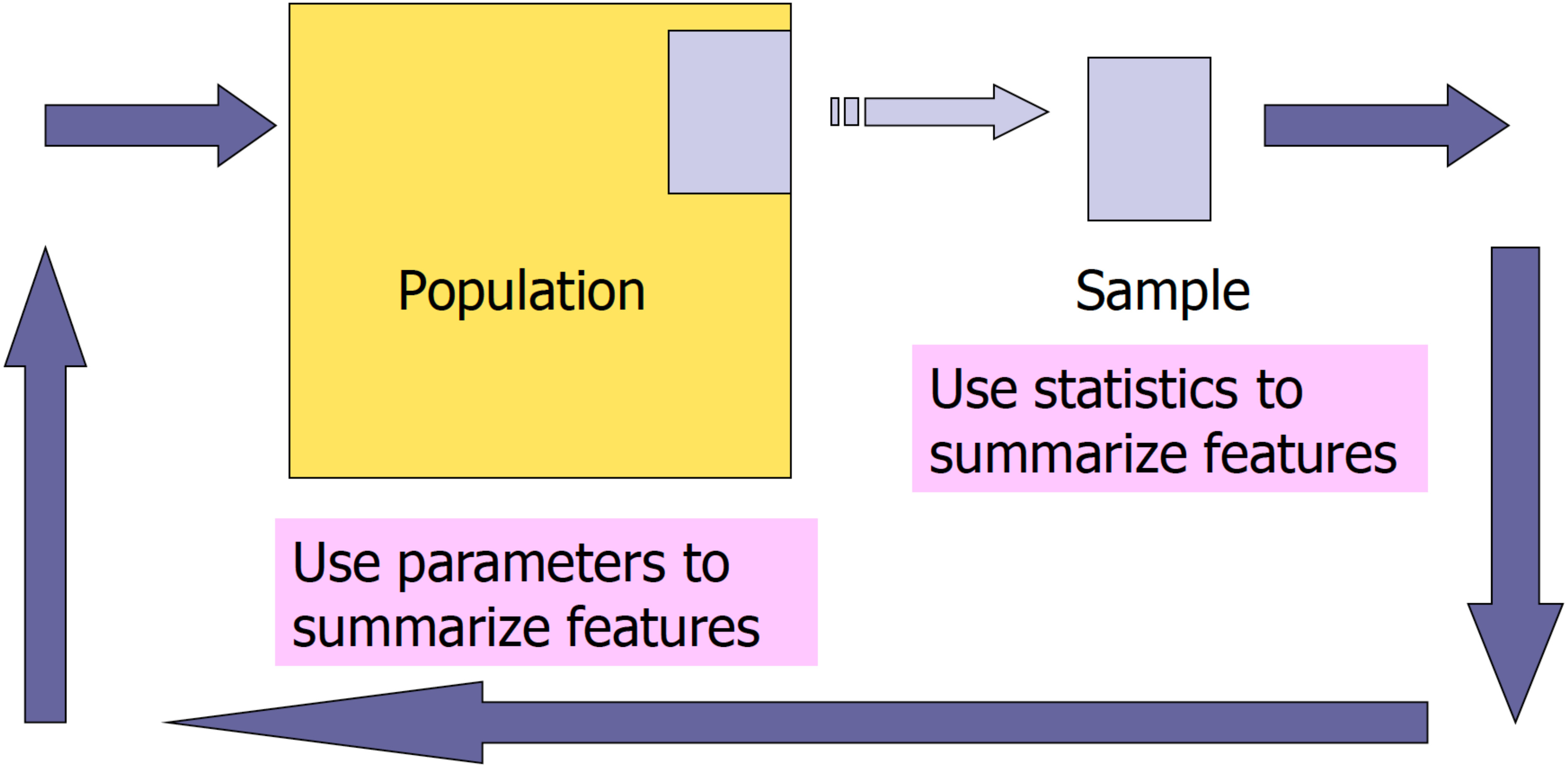
Sample

- *Subcollection* of members selected from a population

The Basic Idea of Statistics

- The basic idea behind all statistical methods of data analysis is to make inferences about a population by studying a relatively small sample chosen from it.

Population and Sample



Use statistics to summarize features

Use parameters to summarize features

Inference on the population from the sample

Example

- A machine makes steel rods for use in optical storage devices. The specification for the diameter of the rods is 0.45 ± 0.02 cm. During the last hour, the machine has made 1000 rods.
- The quality engineer wants to know approximately how many of these rods meet the specification.
- He does not have time to measure all 1000 rods. So, he draws a random sample of 50 rods, measures them, and finds that 46 of them (92%) meet the diameter specification.

Example (cont.)

- Now, it is unlikely that the sample of 50 rods represents the population of 1000 perfectly. The proportion of good rods in the population is likely to differ somewhat from the sample proportion of 92%.
- What the engineer needs to know is just how large that difference is likely to be.
- For example, is it plausible that the population percentage could be as high as 95%? 98%? or as low as 90%? 85%?

Example (cont.)

Some specific questions that the engineer might need to answer based on these sample data:

- The engineer needs to compute a rough estimate of the likely size of the difference between the sample proportion and the population proportion. How large is a typical difference for this kind of sample?

Example (cont.)

- The quality engineer needs to note in a logbook the percentage of acceptable rods manufactured in the last hour.
- Having observed that 92% of the sample rods were good, he will indicate the percentage of acceptable rods in the population as an interval of the form **92% ± x%**, where x is a number calculated to provide reasonable certainty that the true population percentage is in the interval.
- How should *x* be calculated?

Example (cont.)

- The engineer wants to be fairly certain that the percentage of good rods is at least 90%
- Otherwise, he will shut down the process for recalibration.
- How certain can he be that at least 90% of the 1000 rods are good?

Sampling

- Since statistical methods are based on the idea of analyzing a sample drawn from a population, the sample must be chosen in an appropriate way to make sure that the sample represents the population without bias.
- Information obtained from a statistical sample is said to be biased if the results from the sample are radically different from the results of a census of the population.

Example of Bias

- Suppose an employee in a polling firm wishes to gage the public attitude towards a specific public policy.
- Each member of the population could potentially have an opinion about the policy.
- However, it would be impossible from a time and expense standpoint to poll each member of the population.

Example of Bias (cont.)

For instance, if a pre-election poll suggests that a particular candidate for a public office will receive approximately 62% of the vote and then loses the election when his or her opposition receives more than 50% of the vote then the **polling results contained bias**.

Methods of Sampling

In order to eliminate bias, statisticians use four basic methods of sampling that are designed to ensure that each member of a population has an equal probability of being selected for the sample. These four sampling techniques are called:

- **Random Sampling**
- Systematic Sampling
- Stratified Sampling
- Cluster Sampling

Random Sampling

A **random sample** is a sample in which each member of the population has an equal probability of being selected.

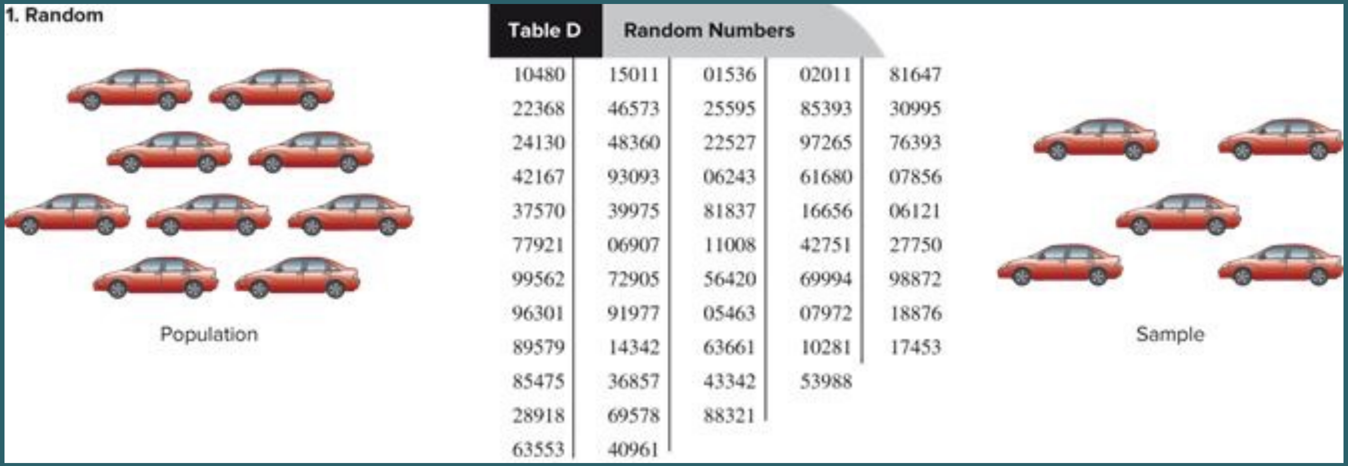


Table D Random Numbers

10480	15011	01536	02011	81647
22368	46573	25595	85393	30995
24130	48360	22527	97265	76393
42167	93093	06243	61680	07856
37570	39975	81837	16656	06121
77921	06907	11008	42751	27750
99562	72905	56420	69994	98872
96301	91977	05463	07972	18876
89579	14342	63661	10281	17453
85475	36857	43342	53988	
28918	69578	88321		
63553	40961			

This method is also called a **simple random sampling**

Simple Random Sample Example

- Suppose an executive of a company that employs thousands of people all over the country would like to include her employees in shaping the vision for the future of the company.
- This would require a considerable amount of time and effort for each employee who is included in the process. Therefore, it would be impossible to include the population in this process.
- She wishes for the opinions of the included employees to be representative of those of the entire company.
- She decides that there should be 30 employees on this committee.

Simple Random Sample Example (cont.)

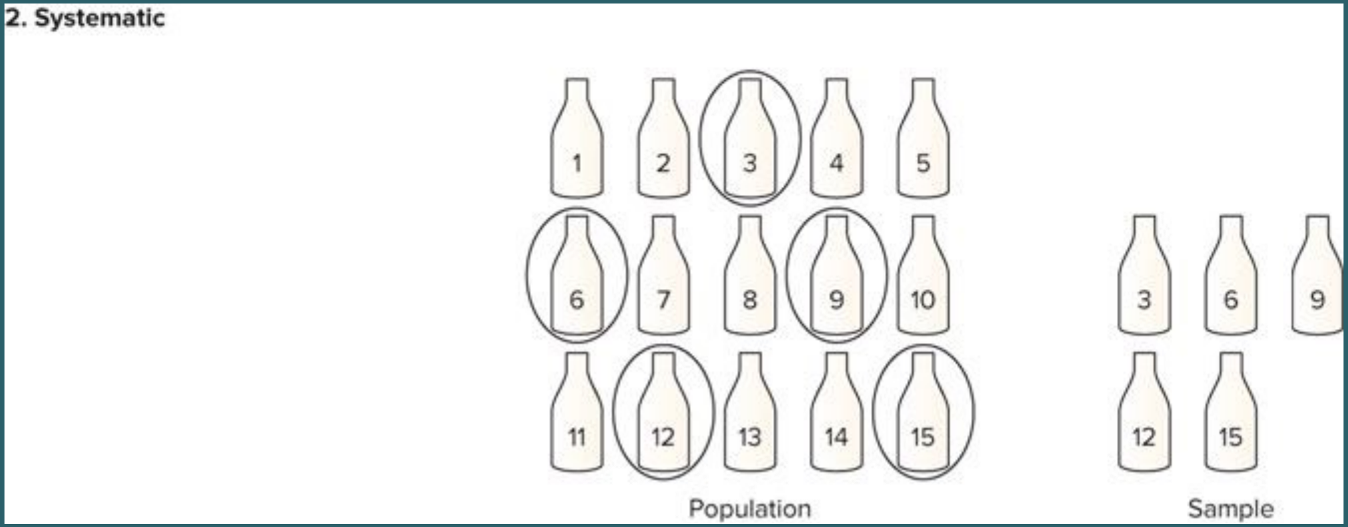
- She decides that there should be 30 employees on this committee.
- She coordinates with the human resources office to create a database that assigns a unique number to each of her employees.
- She then uses a computer program to generate a list of 30 unique random numbers and uses them to select the employees for the committee.
- Since the numbers were generated randomly, each employee had an equal probability of being selected for the committee.

Simple Random Sample Example

- Suppose an executive of a company that employs thousands of people all over the country would like to include her employees in shaping the vision for the future of the company.
- This would require a considerable amount of time and effort for each employee who is included in the process. Therefore, it would be impossible to include the population in this process.
- She wishes for the opinions of the included employees to be representative of those of the entire company.
- She decides that there should be 30 employees on this committee.

Systematic Sampling

A **systematic sample** is a sample obtained by selecting every kth member of the population where k is a counting number.



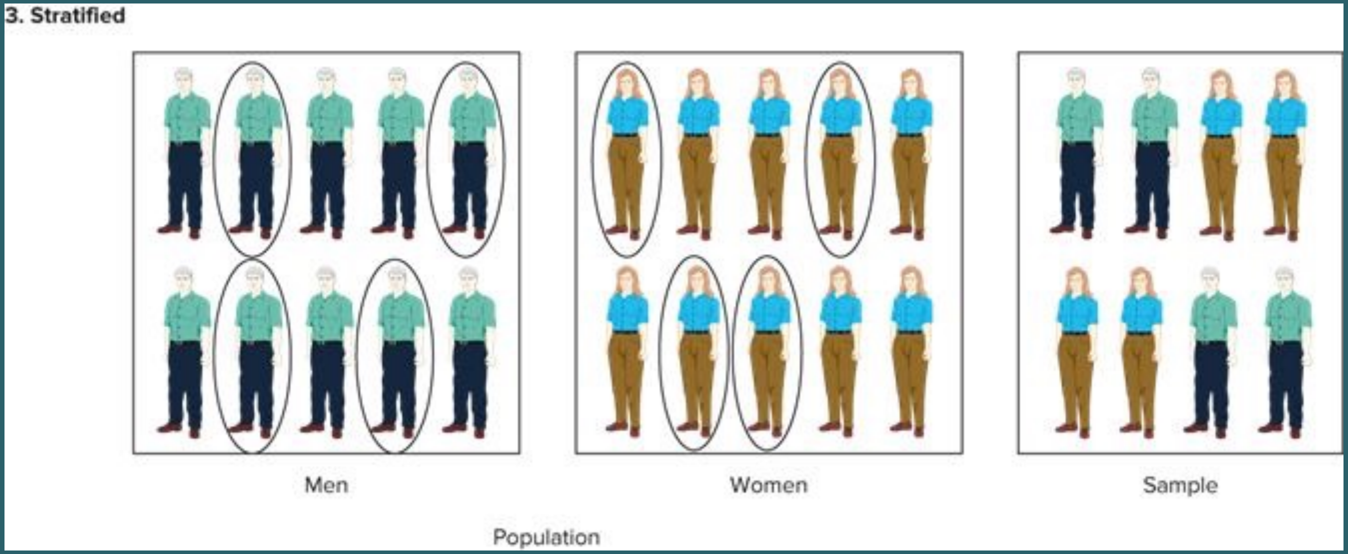
Systematic Sample Example

- Suppose a bottling company would like to test the machines that are filling the bottles by selecting a sample of filled bottles and measuring the amount of product that the machine is putting in the bottles.
- The company statistician goes to the end of the bottling line and selects every 20th bottle and removes it for testing.
- This "system" has thus generated a systematic sample.

Stratified Sampling

A **stratified sample** is obtained by dividing the population into subgroups or strata according to some characteristic relevant to the study. Then subjects are selected from each subgroup.

Note: There can be several subgroups, if required.



Example of Stratified Sampling

Suppose a marketing executive is testing a product and measuring the sales potential within age groups. Opinions are gathered from people who fall into the following age groups:

18 to 29

30 to 41

42 to 53

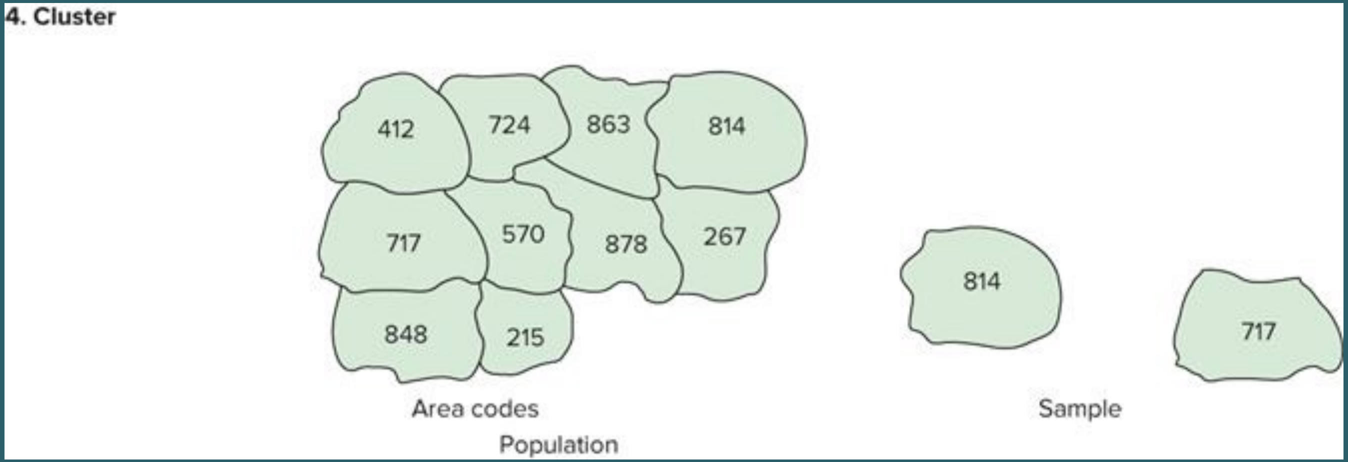
54 to 65

66 and older

The data can then be used to determine the sales potential within each group.

Cluster Sample

A **cluster sample** is obtained by dividing the population into sections or clusters and then selecting one or more clusters at random and using all members in the clusters as members of the sample.

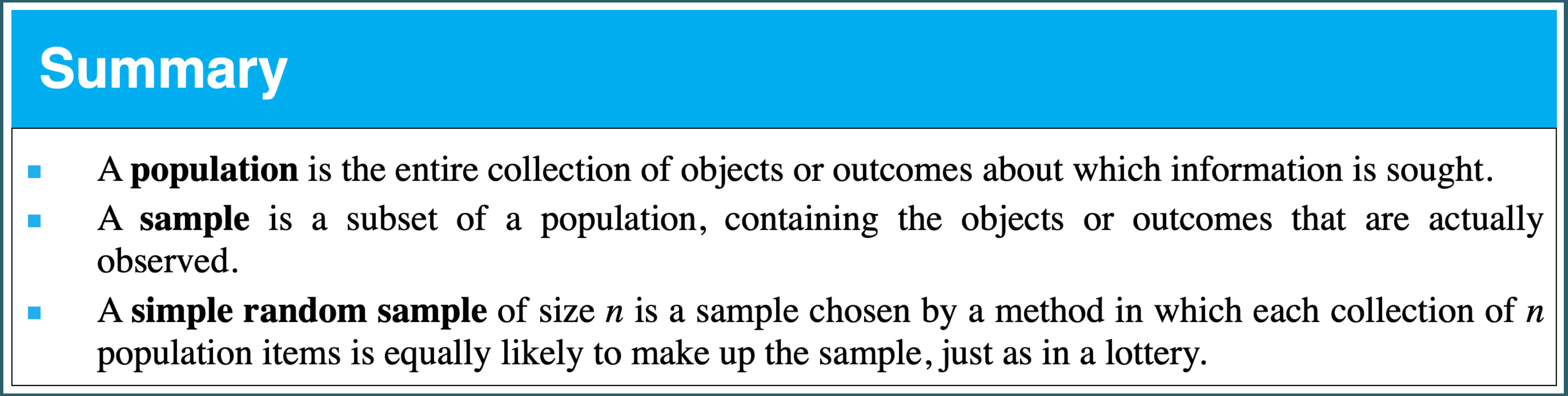


Example of a Cluster Sample

- Suppose the governor of a province would like to find out what the citizens think about a certain budgetary item that will require extra taxation.

- He is also concerned with the opinions of those who live in the western part and the central part of the province.

- The governor will randomly select citizens from these parts of the province to see if there are differences of opinion between the regions.



Summary

- A **population** is the entire collection of objects or outcomes about which information is sought.
- A **sample** is a subset of a population, containing the objects or outcomes that are actually observed.
- A **simple random sample** of size *n* is a sample chosen by a method in which each collection of *n* population items is equally likely to make up the sample, just as in a lottery.

Example 1.1

A physical education professor wants to study the physical fitness levels of students at her university. There are 20,000 students enrolled at the university, and she wants to draw a sample of size 100 to take a physical fitness test. She obtains a list of all 20,000 students, numbered from 1 to 20,000. She uses a computer random number generator to generate 100 random integers between 1 and 20,000 and then invites the 100 students corresponding to those numbers to participate in the study. Is this a simple random sample?

Solution

Yes, this is a simple random sample. Note that it is analogous to a lottery in which each student has a ticket and 100 tickets are drawn.

Example 1.2

A quality engineer wants to inspect rolls of wallpaper in order to obtain information on the rate at which flaws in the printing are occurring. She decides to draw a sample of 50 rolls of wallpaper from a day's production. Each hour for 5 hours, she takes the 10 most recently produced rolls and counts the number of flaws on each. Is this a simple random sample?

Solution

No. Not every subset of 50 rolls of wallpaper is equally likely to make up the sample. To construct a simple random sample, the engineer would need to assign a number to each roll produced during the day and then generate random numbers to determine which rolls make up the sample.

Sample of Convenience

- In some cases, it is difficult or impossible to draw a sample in a truly random way.
- In these cases, the best one can do is to sample items by some convenient method.
- A **sample of convenience** is a sample that is obtained in some convenient way, and not drawn by a well-defined random method.

Sample of Convenience (cont.)

- For example, imagine that a construction engineer has just received a shipment of 1000 concrete blocks, each weighing approximately 50 pounds.
- The blocks have been delivered in a large pile. The engineer wishes to investigate the crushing strength of the blocks by measuring the strengths in a sample of 10 blocks.
- To draw a simple random sample would require removing blocks from the center and bottom of the pile, which might be quite difficult.
- For this reason, the engineer might construct a sample simply by taking 10 blocks off the top of the pile.

Sampling Variation

- Note that simple random samples always differ from their populations in some ways, and occasionally may be substantially different.
- Two different samples from the same population will differ from each other as well.
- This phenomenon is known as **sampling variation**.
- Sampling variation is one of the reasons that scientific experiments produce somewhat different results when repeated, even when the conditions appear to be identical.

Example 1.3

A quality inspector draws a simple random sample of 40 bolts from a large shipment and measures the length of each. He finds that 34 of them, or 85%, meet a length specification. He concludes that exactly 85% of the bolts in the shipment meet the specification. The inspector's supervisor concludes that the proportion of good bolts is likely to be close to, but not exactly equal to, 85%. Which conclusion is appropriate?

Solution

Because of sampling variation, simple random samples don't reflect the population perfectly. They are often fairly close, however. It is therefore appropriate to infer that the proportion of good bolts in the lot is likely to be close to the sample proportion, which is 85%. It is not likely that the population proportion is equal to 85%, however.

Example 1.4

Continuing Example 1.3, another inspector repeats the study with a different simple random sample of 40 bolts. She finds that 36 of them, or 90%, are good. The first inspector claims that she must have done something wrong, since his results showed that 85%, not 90%, of bolts are good. Is he right?

Solution

No, he is not right. This is sampling variation at work. Two different samples from the same population will differ from each other and from the population.

Tangible and Conceptual Populations

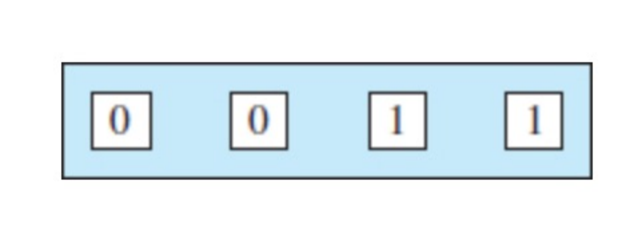
- Populations that consist of actual physical objects (e.g., the students at a university, the concrete blocks in a pile) are called tangible populations.
- Tangible populations are always finite. After an item is sampled, the population size decreases by 1.
- A simple random sample may consist of values obtained from a process under identical experimental conditions. In this case, the sample comes from a population that consists of all the values that might possibly have been observed. Such a population is called a conceptual population.

Sample Independence

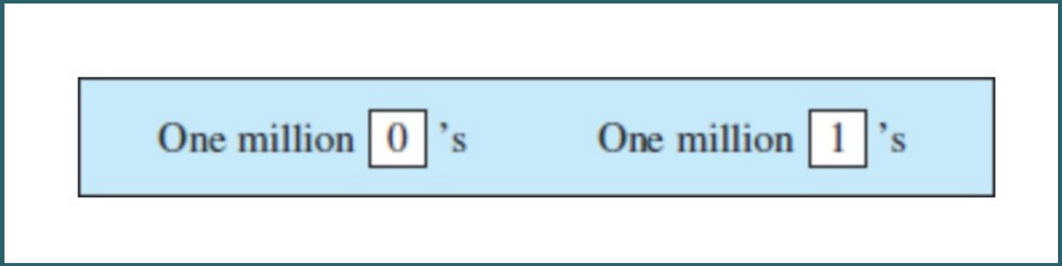
- The items in a sample are said to be independent if knowing the values of some of them does not help to predict the values of the others.
- With a finite, tangible population, the items in a simple random sample are not strictly independent, because as each item is drawn, the population changes.
- This change can be substantial when the population is small.
- When the population is very large, this change is negligible, and the items can be treated as if they were independent.

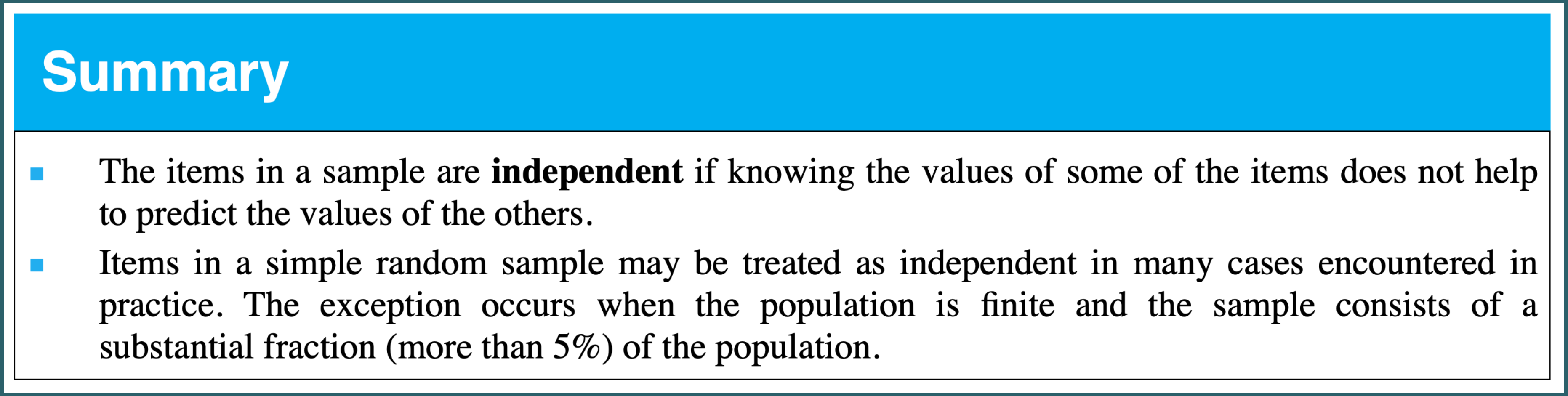
Sample Independence (cont.)

- Imagine that we draw a simple random sample of 2 items from the population



- How about now?

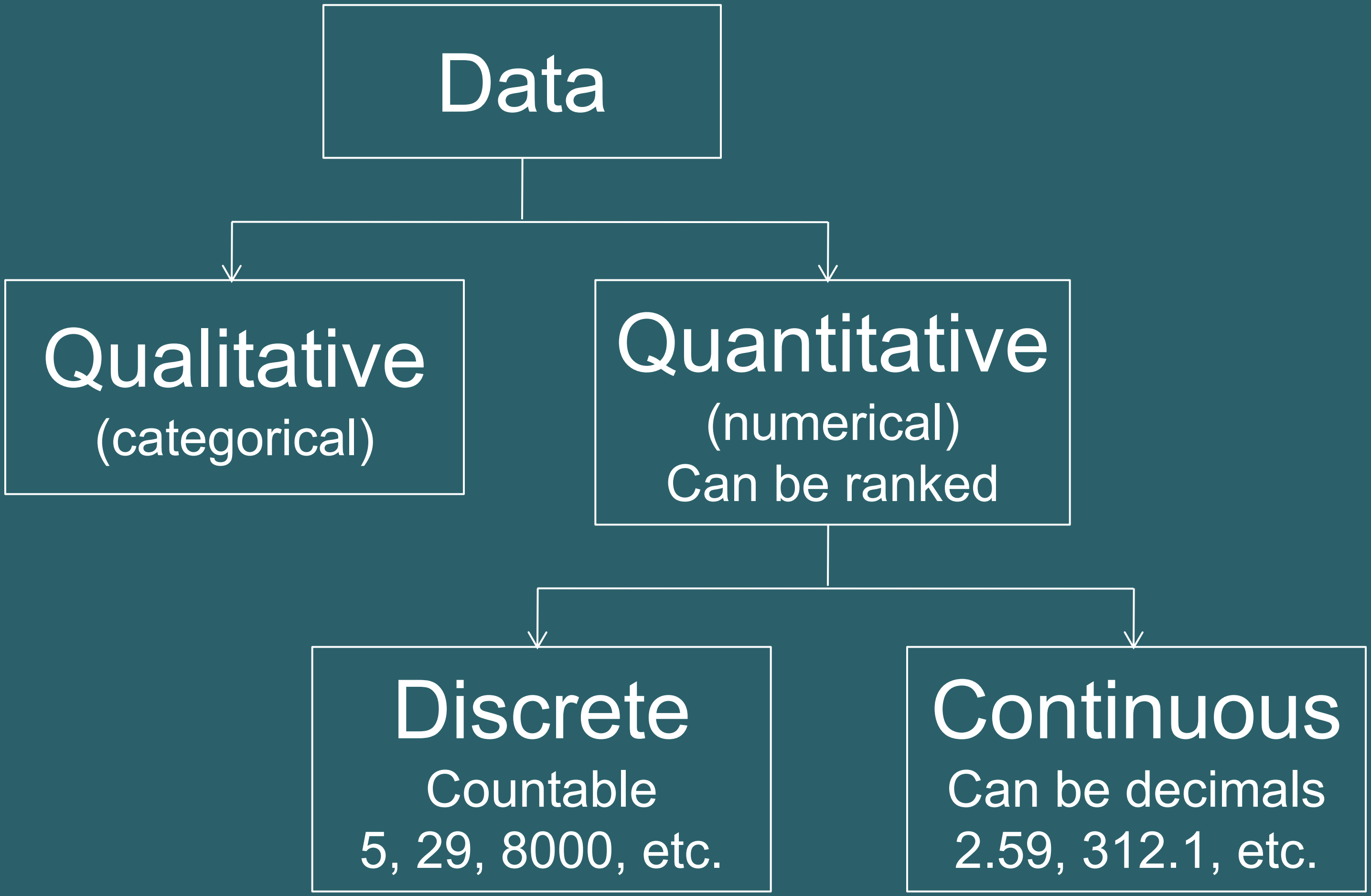




Summary

- The items in a sample are **independent** if knowing the values of some of the items does not help to predict the values of the others.
- Items in a simple random sample may be treated as independent in many cases encountered in practice. The exception occurs when the population is finite and the sample consists of a substantial fraction (more than 5%) of the population.

Types of Data (Variables)



Data

Qualitative
(categorical)

Quantitative
(numerical)
Can be ranked

Discrete
Countable
5, 29, 8000, etc.

Continuous
Can be decimals
2.59, 312.1, etc.

Qualitative Variables

Qualitative Variables are variables that have distinct categories according to some characteristic or attribute.

Examples of qualitative variables

- Hair color
- Soft drink brand
- Jersey number

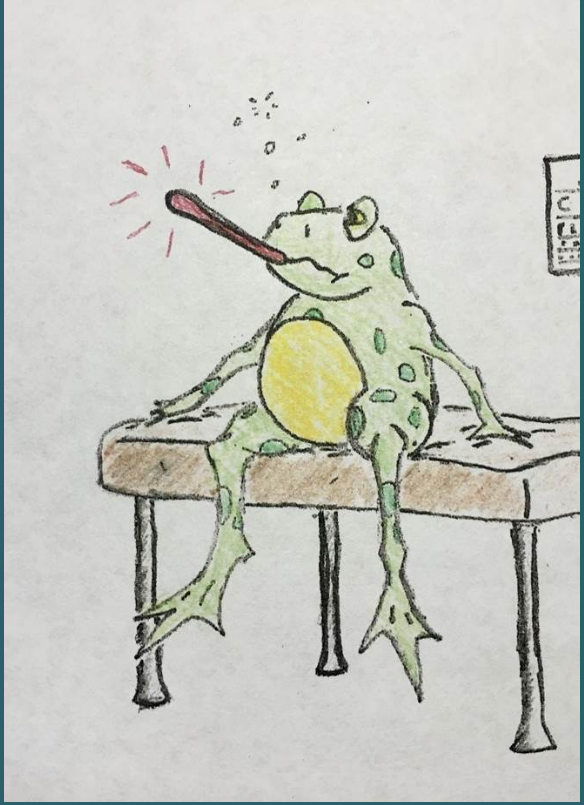
Although the variable jersey number is numerical. It bears no associated countable or measureable quantity.

Quantitative Variables

Quantitative variables are variables that can be counted or measured.

Examples of quantitative variables

- Number of frogs in a jumping contest
- Distance a frog can jump
- Temperature of a frog



Discrete Variables

Discrete variables assume values that can be counted.

Examples of discrete variables

- Number of frogs in a jumping contest
- Number of basketball points scored during a game
- Number of tomato plants in a garden

This variable can only take on counting numbers like 5, 6, 10, or 1000.

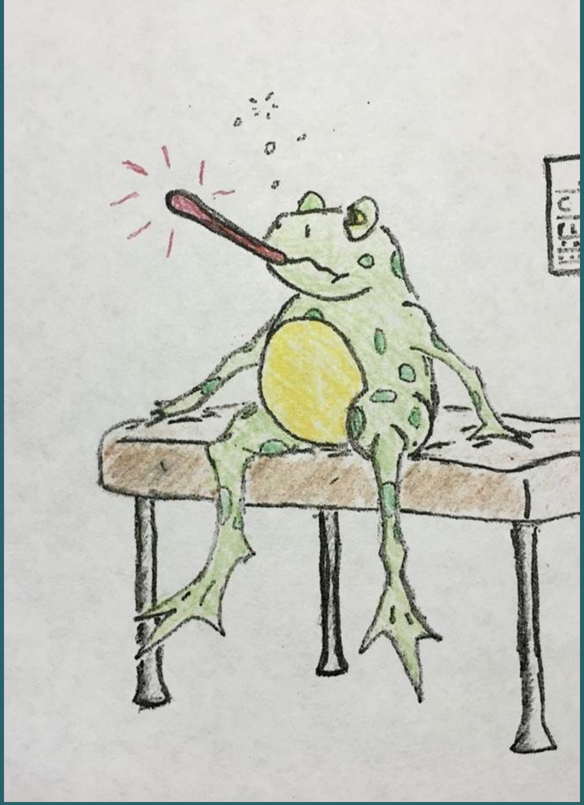
It could not take on a value like 5.5, 10.6, or 99.9.

Continuous Variables

Continuous variables can assume an infinite number of values between any two specific values. They are obtained by measuring.

Examples of continuous variables

- Distance a frog jumps in a contest
- Temperature of a frog
- Weight of a frog



Example 1.8

The article “Wind-Uplift Capacity of Residential Wood Roof-Sheathing Panels Retrofitted with Insulating Foam Adhesive” (P. Datin, D. Prevatt, and W. Pang, *Journal of Architectural Engineering*, 2011:144–154) presents tests in which air pressure was applied to roof-sheathing panels until failure. The pressure at failure for each panel was recorded, along with the type of sheathing, sheathing thickness, and wood species. The following table presents results of four tests.

Sheathing Type	Failure Pressure (kPa)	Thickness (mm)	Wood Species
5-ply plywood	2.63	11.9	Douglas Fir
Oriental Strand Board	3.69	15.1	Spruce-Pine-Fir
Cox plywood	5.26	12.7	Southern Yellow Pine
4-ply plywood	5.03	15.9	Douglas Fir

Which data are numerical and which are categorical?

Experimental and Observational Studies

We can divide scientific experiments into two types:

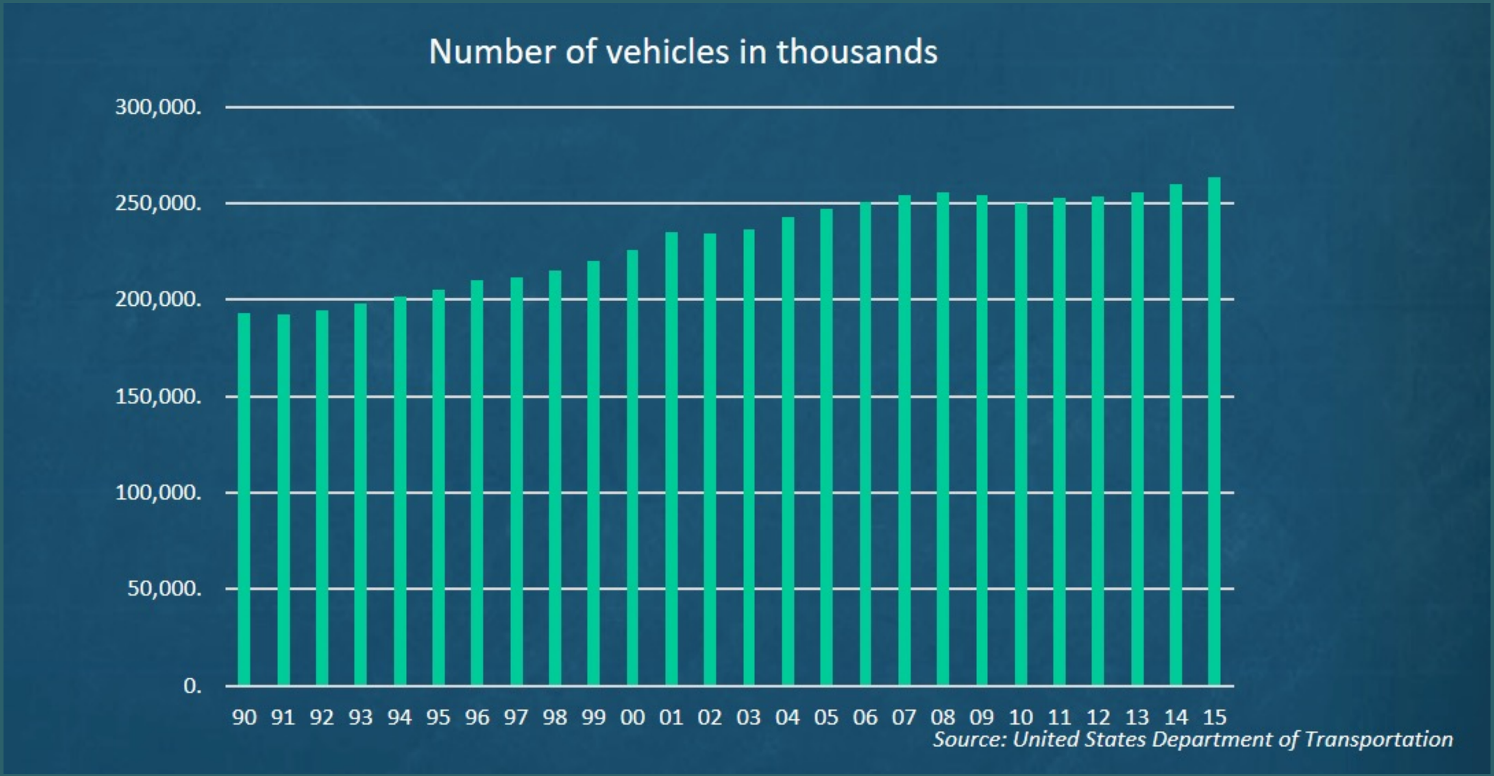
- Observational study
- Experimental study, also called a controlled experiment

Observational Study

The researcher merely **observes** what is happening or what has happened in the past and tries to draw conclusions based on these observations.

Example of Observational Study

If a researcher looks at the number of registered vehicles in the US from the years 1990 through 2015. Then the researcher is merely looking at a historical data set. There is **no intervention** by the researcher in the process of gathering and reporting this data.



Source: United States Department of Transportation

Advantages of Observational Study

Observational studies are typically carried out in a natural setting.

Observational studies can be carried out in situations where intervention by the researcher would be considered unethical or even dangerous such as in the case of crime statistics.

Observational studies can be done using variables where manipulation is impossible such as studies involving height, age, and race.

Disadvantages of Observational Study

A definite cause and effect relationship cannot be shown since the researcher cannot control other influencing variables.

The research is subject to the inaccuracies of other data gatherers such as in the case of historical data like crime statistics from the 1800s or health statistics from another country.

Experimental Study (Controlled Experiment)

The researcher manipulates one of the variables and tries to determine how the manipulation influences other variables.

- The **independent variable** in an experimental study is the one that is being manipulated by the researcher. The independent variable is also called the *explanatory* variable.

- The **resultant variable** is called the *dependent* variable or the *outcome* variable.

Example of Experimental Study

When a pharmaceutical company tests the side effects for experimental medications, they will administer a *placebo* to a control group and the actual medical therapy to the treatment group. They would then compare to see if a statistical significance exists between the incidents of a side effect between the two groups.

	Control Group	Treatment Group
Cranial Diminution	2.2%	5.3%

Advantages of Experimental Study

The researcher can **decide how to select subjects** and assign them to groups such as control and treatment groups.

The researcher can **manipulate variables** such as dosages in medical studies.

Disadvantages of Experimental Study

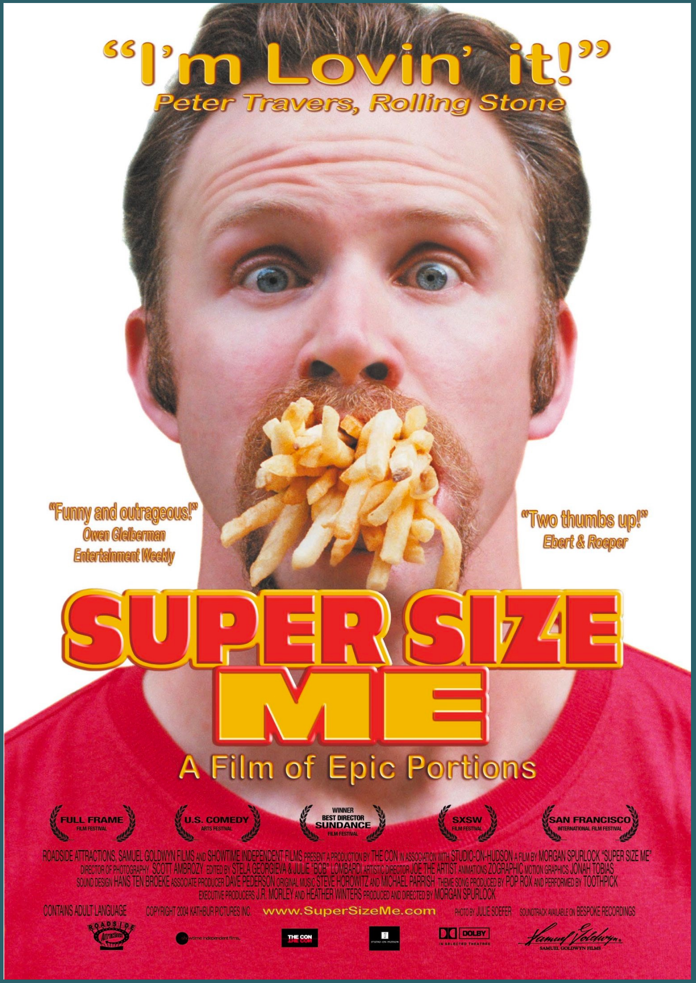
They may occur in **unnatural settings**, such as laboratories and special classrooms. This can lead to several problems. One such problem is that the results might not apply to the natural setting.

The age-old question then is, *‘This mouthwash may kill 10,000 germs in a test tube, but how many germs will it kill in my mouth?*

Disadvantages of Experimental Study

Another disadvantage with an experimental study is the Hawthorne effect.

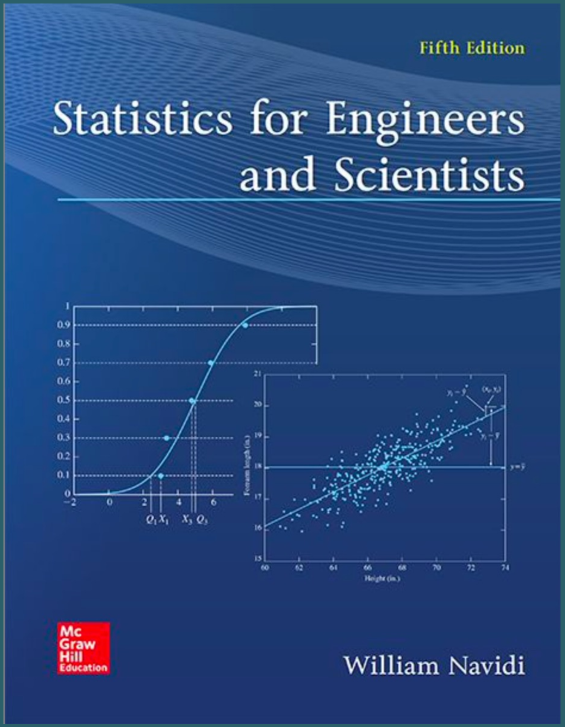
The subjects who knew they were participating in an experiment actually changed their behavior in ways that affected the results of the study.


"I'm Lovin' it!"
Peter Travers, Rolling Stone
"Funny and outrageous!"
Owen Gleiberman
Entertainment Weekly
"Two thumbs up!"
Ebert & Roeper
SUPER SIZE ME
A Film of Epic Portions
FULL FRAME FILM FESTIVAL
U.S. COMEDY ARTS FESTIVAL
WINNER BEST DIRECTOR SUNDANCE FILM FESTIVAL
SXSW FILM FESTIVAL
SAN FRANCISCO INTERNATIONAL FILM FESTIVAL
ROADSIDE ATTRACTIONS, SAMUEL GOLDWYN FILMS AND SHOWTIME INDEPENDENT FILMS PRESENT A PRODUCTION BY THE CON IN ASSOCIATION WITH STUDIO-ON-HUDSON A FILM BY MORGAN SPURLOCK "SUPER SIZE ME"
DIRECTOR OF PHOTOGRAPHY SCOTT AMBROZY EDITED BY STELA GEORGIEVA & JULIE "BOB" LOMBARDI ARTISTIC DIRECTOR JOE THE ARTIST ANIMATIONS ZOGRAPHIC MOTION GRAPHICS JONAH TOBIAS
SOUND DESIGN HANS TEN BROEKE ASSOCIATE PRODUCER DAVE PEDERSON ORIGINAL MUSIC STEVE HOROWITZ AND MICHAEL PARRISH THEME SONG PRODUCED BY POP ROX AND PERFORMED BY TOOTHPICK
EXECUTIVE PRODUCERS J.R. MORLEY AND HEATHER WINTERS PRODUCED AND DIRECTED BY MORGAN SPURLOCK
CONTAINS ADULT LANGUAGE
COPYRIGHT 2004 KATHBUR PICTURES INC.
www.SuperSizeMe.com
PHOTO BY JULIE SOEFER
SOUNDTRACK AVAILABLE ON BESPOKE RECORDINGS
ROADSIDE Attractions
showtime independent films.
THE CON
STUDIO-ON-HUDSON
DOLBY IN SELECTED THEATRES
SAMUEL GOLDWYN FILMS

Summary

In this Section we learned the following:

- Recognize, and understand the advantages and disadvantages of, an **observational study**.
- Recognize, and understand the advantages and disadvantages of, an **experimental study**.



Fifth Edition

Statistics for Engineers and Scientists

William Navidi

Chapter 1

Sampling and Descriptive Statistics

(End of part 1)