Chapter 1

An introduction to Statistics

Outline

1.1 Populations and Samples

1.2 Selecting a Random Sample

1.3 Ratio, Interval, Ordinal, and Nominative Scales of Measurement (Optional)

1.4 An Introduction to Survey Sampling (Optional)

1.5 More About Data Acquisition and Survey Sampling (Optional)

1.1 Populations and Samples

Data

- **Data**: facts and figures from which conclusions can be drawn
- **Data set**: the data that are collected for a particular study
- **Elements**: may be people, objects, events, or other entries
- **Variable**: any characteristic of an element
- **Measurement**: A way to assign a value of a variable to the element
- **Example:** A bank might measure the time it takes for a credit card-holder's bill to be paid to the nearest day.

Data

- The variable is said to be **quantitative**: Measurements that represent quantities are numbers (for example, “how much” or “how many”). For example, **annual starting salary** is quantitative, **age and number of children** is also quantitative
- The variable is said to be **qualitative** or **categorical**: Measurements that represent quantities fall into several categories. For example, **a person’s gender**, **the make of an automobile** and **whether a person who purchases a product is satisfied with the product** are qualitative.

Two types of qualitative variables:

- Nominative
  - Unranked categorization
    - Example: gender, car color
- Ordinal
  - Rank-order categories
  - Ranks are relative to each other
    - Example: Low (1), moderate (2), or high (3) risk

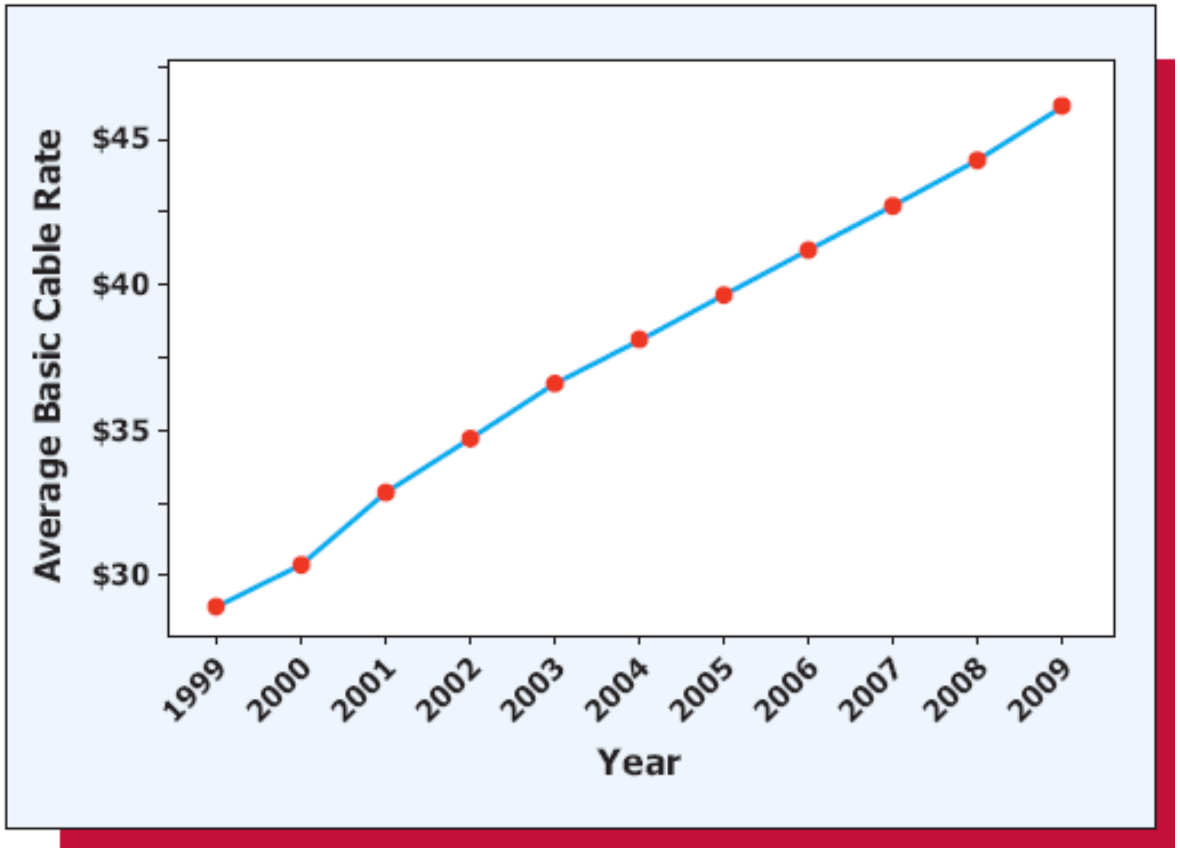
Cross-Sectional Data

- **Cross-sectional data**: Data collected at the same or approximately the same point in time
- **Time series data**: data collected over different time periods

Time Series Data

Year	Average Basic Cable Rate
1999	$ 28.92
2000	30.37
2001	32.87
2002	34.71
2003	36.59
2004	38.14
2005	39.63
2006	41.17
2007	42.72
2008	44.28
2009	46.13

Source: U.S. Energy Information Administration, http://www.eia.gov/



Populations and Samples

Population

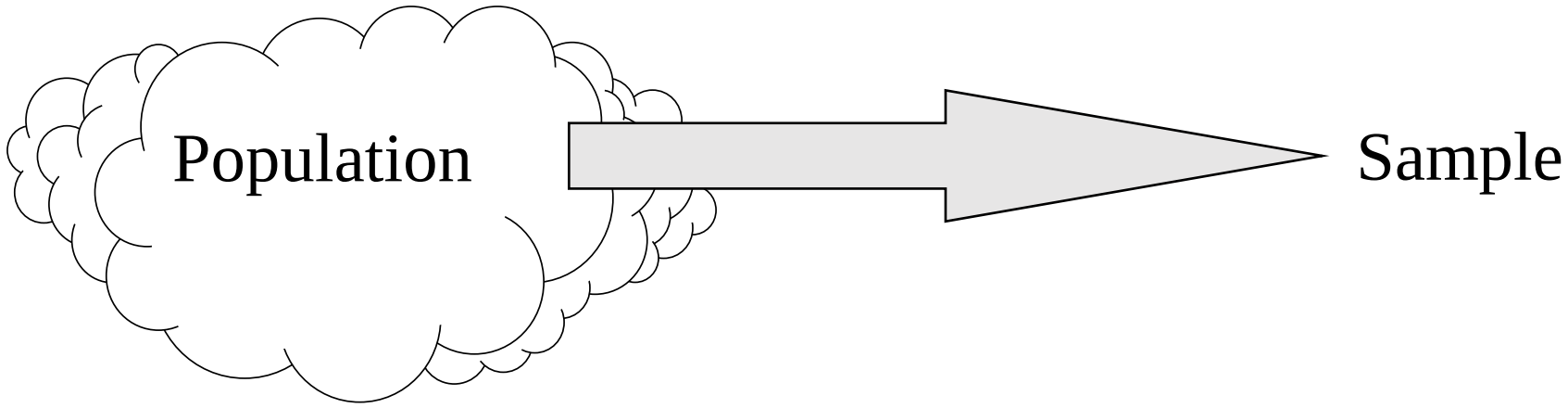
The set of all elements about which we wish to draw conclusions (people, objects or events)

- All of the last year's graduates of Dartmouth College's Master of Business Administration program.
- All Lincoln Town Cars that were produced last year.
- All accounts receivable invoices accumulated last year by The Procter & Gamble Company.
- All fire reported last month to the Tulsa, Oklahoma, fire department.

Census An examination of the entire population of measurements.

Note: Census usually too expensive, too time consuming, and too much effort for a large population.

Sample A selected subset of the units of a population.



Example

A university graduated 8,742 students

a. This is too large for a census.

b. So, we select a sample of these graduates and learn their annual starting salaries.

Sample of measurements

- Measured values of the variable of interest for the sample units.
- For example, the actual annual starting salaries of the sampled graduates.

Population of Measurements

- Measurement of the variable of interest for each and every population unit
  - For example, annual starting salaries of all graduates from last year's MBA program
- Sometimes called *observations*
- If population too large, will analyze a subset

Descriptive statistics

The science of describing the important aspects of a set of measurements

- For example, for a set of annual starting salaries, we want to know:
  - How much to expect
  - What is a high versus low salary
  - How much the salaries differ from each other
- If the population is small enough, could take a census and not have to sample and make any statistical inferences
- But if the population is too large, then …….

Statistical Inference

The science of using a sample of measurements to make generalizations about the important aspects of a population of measurements.

- For example, use a sample of starting salaries to estimate the important aspects of the population of starting salaries

There is a criteria on how to choose a sample:

the information contained in a sample is to accurately reflect the population under study.

1.2 Selecting a Random Sample

Random sample

A random sample is a sample selected from a population so that:

- Each population unit has the same chance of being selected as every other unit
  - Each possible sample (of the same size) has the same chance of being selected
- For example, randomly pick two different people from a group of 15:
  - Number the people from 1 to 15; and write their numbers on 15 different slips of paper
  - Thoroughly mix the papers and randomly pick two of them
  - The numbers on the slips identifies the people for the sample

Sample with replacement

Replace each sampled unit before picking next unit

- The unit is placed back into the population for possible reselection
- However, the same unit in the sample does not contribute new information

Sample without replacement

A sampled unit is withheld from possibly being selected again in the same sample

- Guarantees a sample of different units
  - Each sampled unit contributes different information
  - Sampling without replacement is the usual and customary sampling method

Approximately Random Samples

Sometimes it is not possible to list and thus number all the units in a population. In such a situation we often select a systematic sample, which approximates a random sample.

A Systematic Sample

Randomly enter the population and systematically sample every *k*th unit.

Three Case Studies that Illustrate Sampling and Statistical Inference

1. The Cell Phone Case: Estimating Cell Phone Costs
2. The Marketing Research Case: Rating a New Bottle Design
3. The Car Mileage Case: Estimating Mileage

Example 1.1: The Cell Phone Case: Estimating Cell Phone Costs

- Considering using a company to manage their cellular resources
- Random sample of 100 employees on 500-minute plan
- Many overages and underage

EXAMPLE 1.1 The Cell Phone Case: Reducing Cellular Phone Costs

Part 1: The cost of company cell phone use Rising cell phone costs have forced companies having large numbers of cellular users to hire services to r other wireless resources. These cellular management services use sc mathematical models to choose cost-efficient cell phone plans for the mindWireless of Austin, Texas, specializes in automated wireless cost to Kevin Whitehurst, co-founder of mindWireless, cell phone carriers more minutes than one's plan allows—and *underage*—using fewer m paid for—to deliver almost half of their revenues.[3] As a result, a comp phone use can be excessive—18 cents per minute or more. However, Mr by using mindWireless automated cost management to select calling plans, this cost can be reduced to 12 cents per minute or less.

In this case we consider a bank that wishes to decide whether to hire a cellular management service to choose its employees' calling plans. While the bank has over 10,000 employees on many different types of calling plans, a cellular management service suggests that by studying the calling patterns of cellular users on 500-minute-per-month plans, the bank can accurately assess whether its cell phone costs can be substantially reduced. The bank has 2,136 employees on a variety of 500-minute-per-month plans with different basic monthly rates, different overage charges, and different additional charges for long distance and roaming. It would be extremely time consuming to analyze in detail the cell phone bills of all 2,136 employees. Therefore, the bank will estimate its cellular costs for the 500-minute plans by analyzing last month's cell phone bills for a *random sample* of 100 employees on these plans.[4]

The Cell Phone Case: The Data

Table 1.2

A Sample of Cellular Usages (in minutes) for 100 Randomly Selected Employees									
75	485	37	547	753	93	897	694	797	477
654	578	504	670	490	225	509	247	597	173
496	553	0	198	507	157	672	296	774	479
0	822	705	814	20	513	546	801	721	273
879	433	420	521	648	41	528	359	367	948
511	704	535	585	341	530	216	512	491	0
542	562	49	505	461	496	241	624	885	259
571	338	503	529	737	444	372	555	290	830
719	120	468	730	853	18	479	144	24	513
482	683	212	418	399	376	323	173	669	611

Example 1.2: The Marketing Research Case: Rating a New Bottle Design

- Studying to see if changes should be made in the bottle design for a popular soft drink
- Using “mall intercept method”
- Sample size of 60

EXAMPLE 1.2 The Marketing Research Case: Rating a Bottle Design

Part 1: Rating a bottle design The design of a package or bottle can have an important effect on a company’s bottom line. In this case a brand group wishes to research consumer reaction to a new bottle design for a popular soft drink. To do this, the brand group will show consumers the new bottle and ask them to rate the bottle image. For each consumer interviewed, a bottle image **composite score** will be found by adding the consumer’s numerical responses to the five questions shown in Figure 1.2. It follows that the minimum possible bottle image composite

The Marketing Research Case: The Form and the Data

Figure 1.1 and Table 1.3

The Bottle Design Survey Instrument

Please circle the response that most accurately describes whether you agree or disagree with each statement about the bottle you have examined.

Statement	Strongly Disagree						Strongly Agree
The size of this bottle is convenient.	1	2	3	4	5	6	7
The contoured shape of this bottle easy to handle.	1	2	3	4	5	6	7
The label on this bottle is easy to read.	1	2	3	4	5	6	7
This bottle is easy to open.	1	2	3	4	5	6	7
Based on its overall appeal, I like this bottle design.	1	2	3	4	5	6	7

A Sample of Bottle Design Ratings (Composite Scores for a Systematic Sample of 60 Shoppers)

34	33	33	29	26	33	28	25	32	33
32	25	27	33	22	27	32	33	32	29
24	30	20	34	31	32	30	35	33	31
32	28	30	31	31	33	29	27	34	31
31	28	33	31	32	28	26	29	32	34
32	30	34	32	30	30	32	31	29	33

Terms

- **Process**: a sequence of operations that takes inputs and turns them into outputs
- **Finite population**: a population of limited size
- **Infinite population**: a population of unlimited size

Sampling a Process

Process

A sequence of operations that takes *inputs* (labor, raw materials, methods, machines, and so on) and turns them into *outputs* (products, services, and the like)

Processes produce output over time

- The “population” from a process is all output produced in the past, present, and the yet-to-occur future.
- For example, all automobiles of a particular make and model, for instance, the Lincoln Town Car
  - Cars will continue to be made over time

Example

The Coffee Temperature Case: Monitoring Coffee Temperatures

This case concerns coffee temperatures at a fast-food restaurant. To do this, the restaurant personnel measure the temperature of the coffee being dispensed (in degrees F) at half-hour intervals from 10 A.M. to 9:30 P.M. on a given day. Data is list on P15, Exercises 1.12.

- A process is in statistical control if it does not exhibit any unusual process variations.
- To determine if a process is in control or not, sample the process often enough to detect unusual variations
- A *runs plot* is a graph of individual process measurements over time. Figure 1.6 shows a runs plot of the temperature data.

THE COFFEE TEMPERATURE CASE DS Coffee

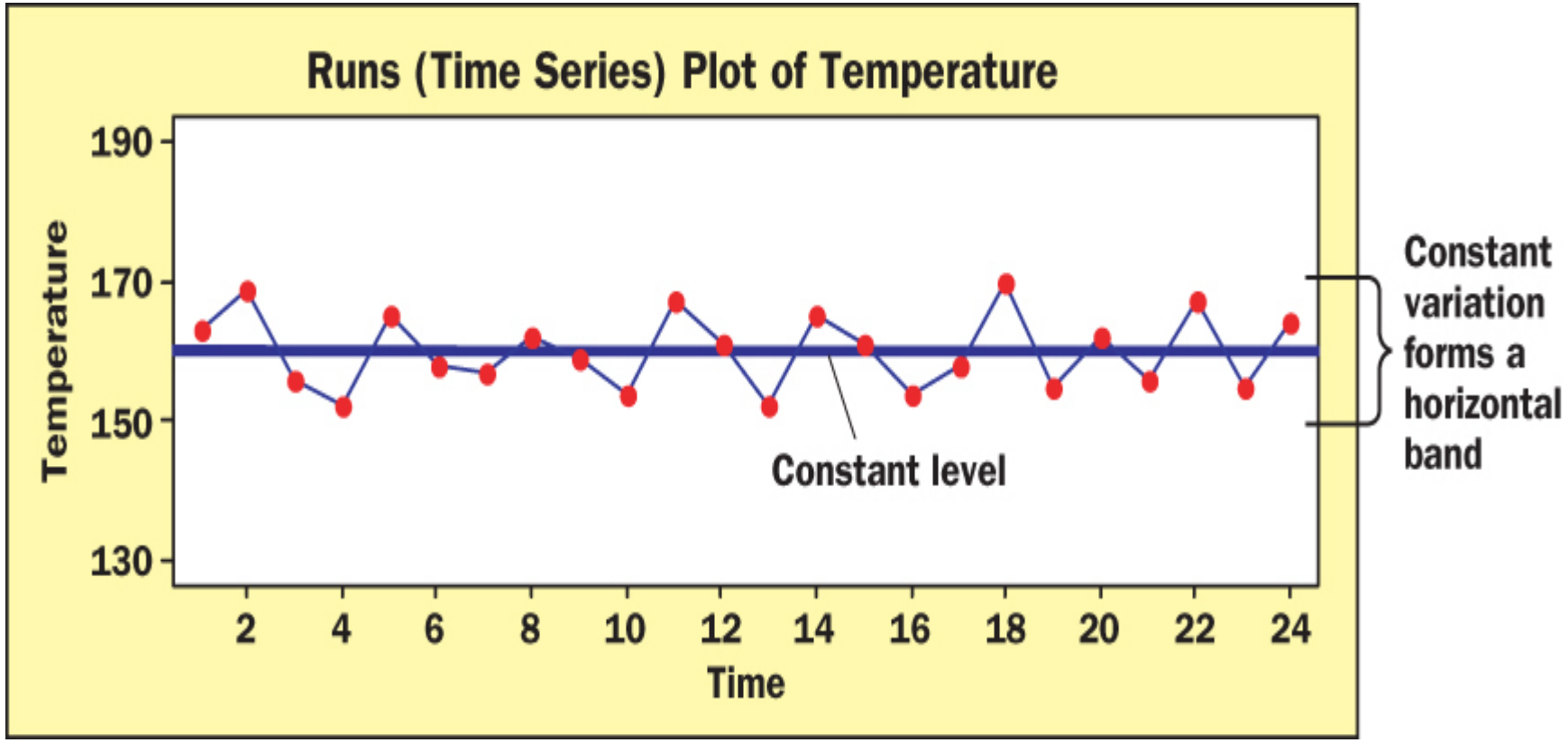
According to the website of the American Association for Justice[11] Stella Liebeck of Albuquerque, New Mexico, was severely burned by McDonald's coffee in February 1992. Liebeck, who received third-degree burns over 6 percent of her body, was awarded $160,000 in compensatory damages and $480,000 in punitive damages. A postverdict investigation revealed that the coffee temperature at the local Albuquerque McDonald's had dropped from about 185°F before the trial to about 158° after the trial.

This case concerns coffee temperatures at a fast-food restaurant. Because of the possibility of future litigation and to possibly improve the coffee's taste, the restaurant wishes to study the temperature of the coffee it serves. To do this, the restaurant personnel measure the temperature of the coffee being dispensed (in degrees Fahrenheit) at a randomly selected time during each of the 24 half-hour periods from 8 A.M. to 7:30 P.M. on a given day. This is then repeated on a second day, giving the 48 coffee temperatures in Table 1.10. Make a time series plot of the coffee temperatures, and assuming process consistency, estimate limits between which most of the coffee temperatures at the restaurant would fall.

TABLE 1.7 24 Coffee Temperatures Observed in Time Order (°F) Coffee

Time		Coffee Temperature	Time		Coffee Temperature	Time		Coffee Temperature
(10:00 A.M.)	1	163°F	(2:00 P.M.)	9	159°F	(6:00 P.M.)	17	158°F
	2	169		10	154		18	170
	3	156		11	167		19	155
	4	152		12	161		20	162
(12:00 noon)	5	165	(4:00 P.M.)	13	152	(8:00 P.M.)	21	156
	6	158		14	165		22	167
	7	157		15	161		23	155
	8	162		16	154		24	164

Runs Plot of Coffee Temperatures: The Process is in Statistical Control.



Results

- Over time, temperatures appear to have a fairly constant amount of variation around a fairly constant level
  - The temperature is expected to be at the constant level shown by the horizontal blue line
    - Sometimes the temperature is higher and sometimes lower than the constant level
  - About the same amount of spread of the values (data points) around the constant level
    - The points are as far above the line as below it
    - The data points appear to form a horizontal band
- So, the process is in statistical control
  - Coffee-making process is operating “consistently”

Example 1.3: The Care Mileage Case: Estimating Mileage

- Study of tax credit offered by the federal government for improving fuel economy
- Automaker has introduced a new model and wishes to demonstrate it qualifies for the tax credit
- Sample of 50 cars

The Care Mileage Case: The Data

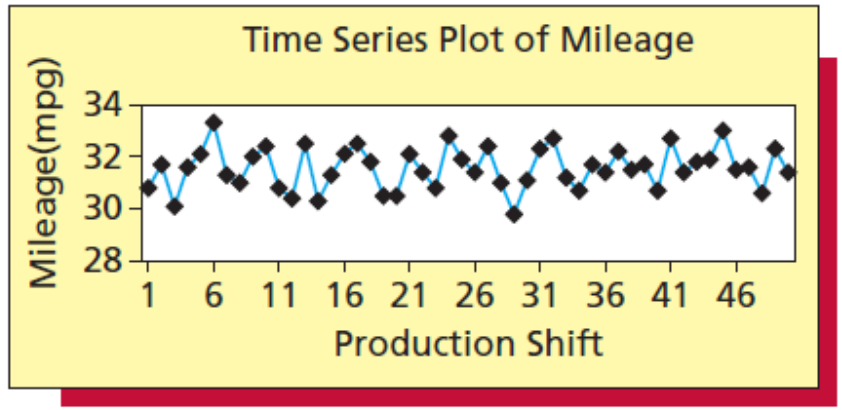
Table 1.4 and Figure 1.2

A Sample of 50 Mileages

30.8	30.8	32.1	32.3	32.7
31.7	30.4	31.4	32.7	31.4
30.1	32.5	30.8	31.2	31.8
31.6	30.3	32.8	30.7	31.9
32.1	31.3	31.9	31.7	33.0
33.3	32.1	31.4	31.4	31.5
31.3	32.5	32.4	32.2	31.6
31.0	31.8	31.0	31.5	30.6
32.0	30.5	29.8	31.7	32.3
32.4	30.5	31.1	30.7	31.4

Note: Time order is given by reading down the columns from left to right.

A Time Series Plot of the 50 Mileages



1.3 Ratio, Interval, Ordinal, and Nominative Scales of Measurement (Optional)

- Nominative
- Ordinal
- Interval
- Ratio

Qualitative Variables

- **Nominative**: A qualitative variable for which there is no meaningful ordering, or ranking, of the categories
  - Example: gender, car color
- **Ordinal**: A qualitative variable for which there is a meaningful ordering, or ranking, of the categories
  - Example: teaching effectiveness

Interval Variable

- All of the characteristics of ordinal
- Measurements are on a numerical scale with an arbitrary zero point
  - The “zero” is assigned: it is nonphysical and not meaningful
  - Zero does not mean the absence of the quantity that we are trying to measure

Interval Variable

- Can only meaningfully compare values by the interval between them
  - Cannot compare values by taking their ratios
  - “Interval” is the arithmetic difference between the values
- Example: temperature
  - 0° F means “cold,” not “no heat”
  - 60° F is ***not*** twice as warm as 30° F

Ratio Variable

- Measurements are on a numerical scale with a meaningful zero point
  - Zero means “none” or “nothing”
- Values can be compared in terms of their interval and ratio
  - $30 is $20 more than $10
  - $0 means no money

Ratio Variable

- In business and finance, most quantitative variables are ratio variables, such as anything to do with money
  - Examples: Earnings, profit, loss, age, distance, height, weight

1.4 An Introduction to Survey Sampling (Optional)

- Methods for obtaining a sample are called sampling designs. The sample we take is sometimes called a sample survey.
- Stratified random sampling, cluster sampling, and systematic sampling
- In order to select a stratified random sample, we divide the population into nonoverlapping groups of similar units (people, objects, etc.). These groups are called strata. Then a random sample is selected from each stratum, and these samples are combined to form the full sample.

- Multi-stage cluster sampling
- Systematic sampling

1.5 More about Data Acquisition and Survey Sampling (Optional)

- Existing sources: data already gathered by public or private sources
  - Internet
  - Library
  - Private data sources
- Experimental and observational studies: data we collect ourselves for a specific purpose
  - Response variable: variable of interest
  - Factors: other variables related to response variable

  Example: A designed experiment

  A survey

  An observational study

Initiating a Study

- First, define the variable of interest, called a **response variable**
- Next, define other variables that may be related to the variable of interest and will be measured, called **independent variables**
- If we manipulate the independent variables, we have an **experimental study**
- If unable to control independent variables, the study is **observational**

Summary

Chapter Summary

We began this chapter by discussing **data.** We learned that the data that are collected for a particular study are referred to as a **data set,** and we learned that **elements** are the entities described by a data set. In order to determine what information we need about a group of elements, we define important **variables,** or characteristics, describing the elements. **Quantitative variables** are variables that use numbers to measure quantities (that is, "how much" or "how many") and **qualitative, or categorical, variables** simply record into which of several categories an element falls.

We next discussed the difference between cross-sectional data and time series data. **Cross-sectional data** are data collected at the same or approximately the same point in time. **Time series data** are data collected over different time periods. There are various **sources of data.** Specifically, we can obtain data from **existing sources** or from **experimental or observational studies** done in-house or by paid outsiders.

We often collect data to study a **population,** which is the set of all elements about which we wish to draw conclusions. We saw that, since many populations are too large to examine in their entirety, we frequently study a population by selecting a **sample,** which is a subset of the population elements. Next we learned that, if the information contained in a sample is to accurately represent the population, then the sample should be **randomly selected** from the population.

We concluded this chapter with optional Section 1.5, which considered different types of quantitative and qualitative variables. We learned that there are two types of **quantitative variables—ratio variables,** which are measured on a scale such that ratios of its values are meaningful and there is an inherently defined zero value, and **interval variables,** for which ratios are not meaningful and there is no inherently defined zero value. We also saw that there are two types of **qualitative variables—ordinal variables,** for which there is a meaningful ordering of the categories, and **nominative variables,** for which there is no meaningful ordering of the categories.

Thank you!