A PowerPoint Presentation Package to Accompany

Applied Statistics in Business & Economics, 5th edition

David P. Doane and Lori E. Seward

Prepared by Lloyd R. Jaisingh

Data Collection

Chapter Contents

Data Collection

Chapter Learning Objectives

LO2-1: Use basic terminology for describing data and samples.

LO2-2: Explain the difference between numerical and categorical data.

LO2-3: Explain the difference between time series and cross-sectional data.

LO2-4: Recognize levels of measurement in data and ways of coding data.

LO2-5: Recognize a Likert scale and know how to use it.

Data Collection

Chapter Learning Objectives

LO2-6: Use the correct terminology for samples and populations.

LO2-7: Explain the common sampling methods and how to implement them.

LO2-8: Find everyday print or electronic data sources.

LO2-9: Describe basic elements of survey types, survey designs, and response scales.

LO2-1

2.1 Variables and Data

LO2-1: Use basic terminology for describing data and samples.

Data Terminology: ***Observations, Variables, Data Sets***

- Observation: a single member of a collection of items that we want to study, such as a person, firm, or region.
- Variable: a characteristic of the subject or individual, such as an employee's income or an invoice amount
- Data Set: consists of all the values of all of the variables for all of the observations we have chosen to observe.

2.1 Variables and Data

Table 2.2: Number of Variables and Typical Tasks

TABLE 2.2

Number of Variables and Typical Tasks

Data Set	*Variables*	*Example*	*Typical Tasks*
Univariate	One	Income	Histograms, basic statistics
Bivariate	Two	Income, Age	Scatter plots, correlation
Multivariate	More than two	Income, Age, Gender	Regression modeling

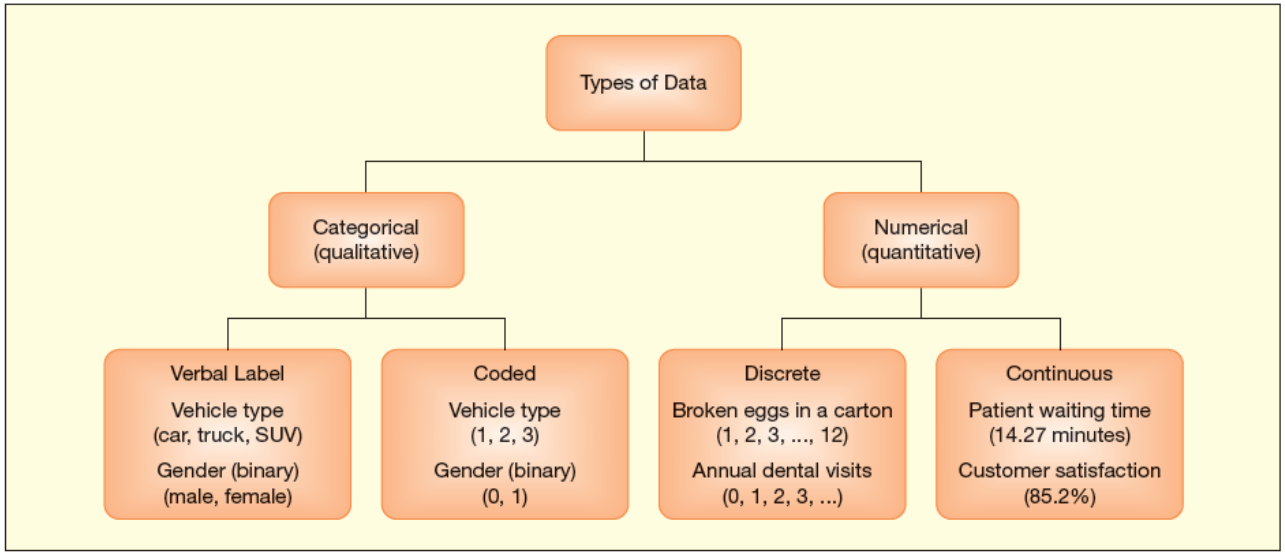
LO2-2

Variables and Data

LO2-2: Explain the difference between numerical and categorical data.

FIGURE 2.1

Data Types and Examples



- ***Note:*** Ambiguity is introduced when continuous data are rounded to whole numbers. Be cautious.

LO2-3

Variables and Data

LO2-3: Explain the difference between time series and cross-sectional data.

Time Series Data and Cross-Sectional Data

- Each observation in the sample represents a different equally spaced point in time (e.g., years, months, days).
- *Periodicity* may be annual, quarterly, monthly, weekly, daily, hourly, etc.
- We are interested in *trends and patterns over time* (e.g., personal bankruptcies from 1980 to 2008 as shown in Figure 2.2).

LO2-3

Variables and Data

Cross Sectional Data

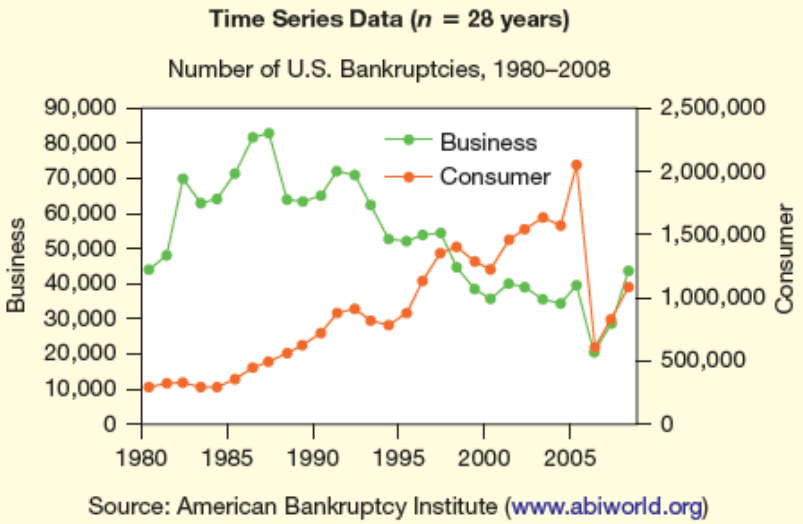
- Each observation represents a different individual unit (e.g., person) at the same point in time (e.g., monthly VISA balances).
- We are interested in:
  - *variation among observations* (e.g. accounts receivable in 20 Subway franchises) or in
  - *relationships (e.g. whether accounts receivable are related to sales volume in 20 Subway franchises as shown in Figure 2.2). .*
- We can combine the two data types to get *pooled cross-sectional and time series data.*

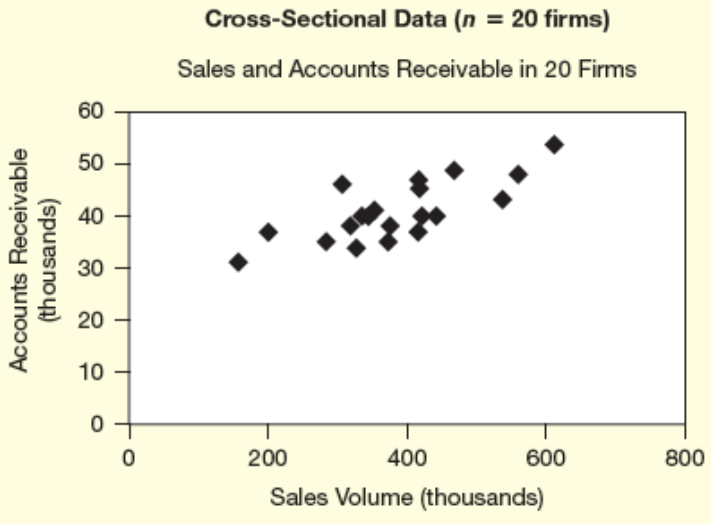
LO2-3

Variables and Data

FIGURE 2.2

Examples of Time Series versus Cross-Sectional Data





LO2-4

2.2 Level of Measurement

LO2-4: Recognize levels of measurement in data and ways of coding data.

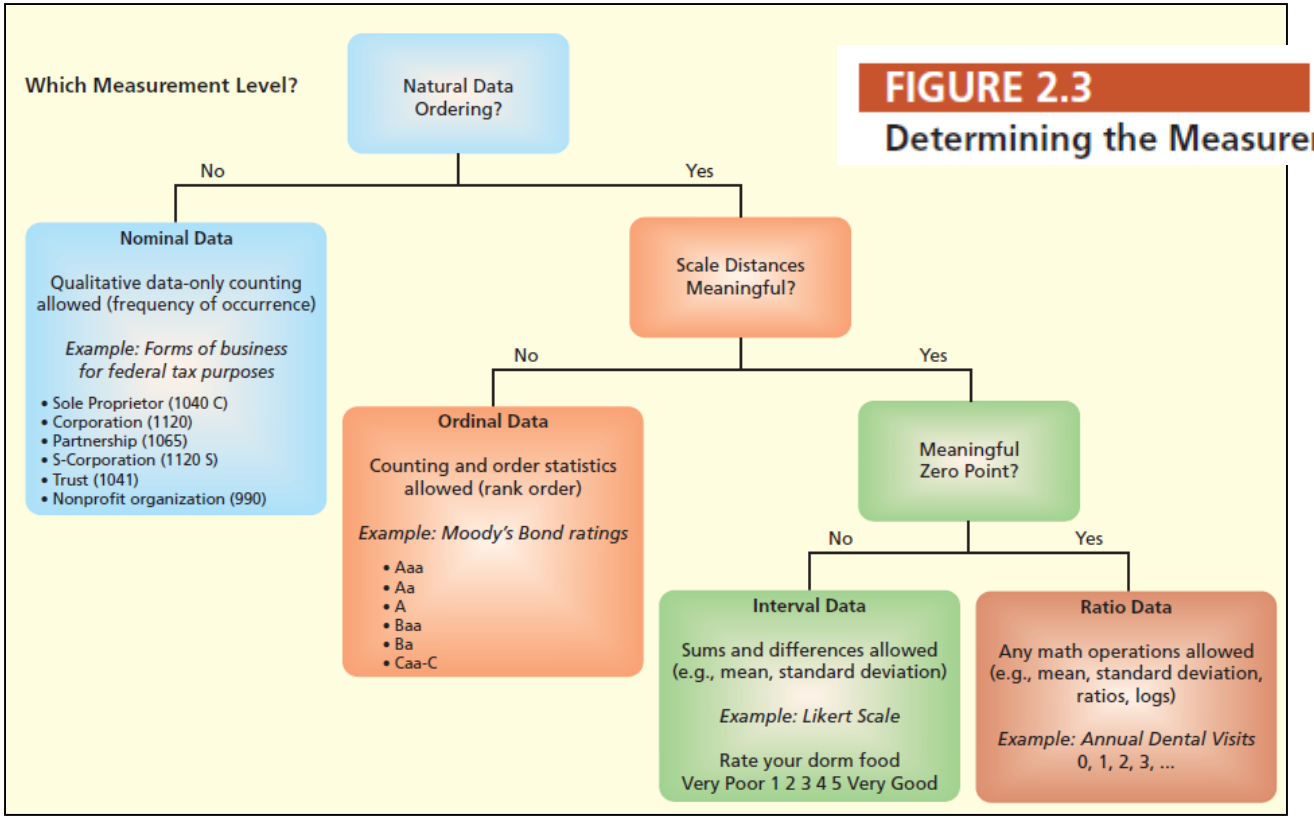


FIGURE 2.3
Determining the Measurement Level

LO2-4

2.2 Level of Measurement

LO2-4: Recognize levels of measurement in data and ways of coding data.

Levels of Measurement

Level of Measurement	Characteristics	Example
Nominal	**Categories only**	**Eye color (blue, brown, green, etc.)**
Ordinal	**Rank has meaning. No clear meaning to distance**	**Rarely, never**
Interval	**Distance has meaning**	**Temperature (57^{o} Celsius)**
Ratio	**Meaningful zero exists**	**Accounts payable ($21.7 million)**

LO2-4

2.2 Level of Measurement

Nominal Measurement

- Nominal data merely identify a ***category***.
- Nominal data are qualitative, attribute, categorical or classification data and can be coded numerically (e.g., 1 = Apple, 2 = Compaq, 3 = Dell, 4 = HP).
- Only mathematical operations are counting (e.g., frequencies) and simple statistics.

Ordinal Measurement

- Ordinal data codes can be *ranked* (e.g., 1 = Frequently, 2 = Sometimes, 3 = Rarely, 4 = Never).

LO2-4

2.2 Level of Measurement

Ordinal Measurement

- *Distance* between codes is not meaningful (e.g., distance between 1 and 2, or between 2 and 3, or between 3 and 4 lacks meaning).
- Many useful statistical tests exist for ordinal data. Especially useful in social science, marketing and human resource research.

Interval Measurement

- Data can not only be ranked, but also have meaningful intervals between scale points (e.g., difference between 60°F and 70°F is same as difference between 20°F and 30°F).

LO2-4

2.2 Level of Measurement

Interval Measurement

- Since intervals between numbers represent *distances*, mathematical operations can be performed (e.g., average).
- Zero point of interval scales is arbitrary, so ratios are not meaningful (e.g., 60°F *is not* twice as warm as 30°F).

Ratio Measurement

- *Ratio data* have all properties of nominal, ordinal and interval data types and also possess a ***meaningful zero*** (absence of quantity being measured).

LO2-4

2.2 Level of Measurement

Ratio Measurement

- Because of this zero point, ratios of data values are meaningful (e.g., $20 million profit is twice as much as $10 million).
- Zero does not have to be observable in the data; it is an absolute reference point.

LO2-5

2.2 Level of Measurement

LO2-5: Recognize a Likert scale and know how to use it.

Likert Scales

- A special case of interval data frequently used in survey research.
- The *coarseness* of a Likert scale refers to the number of scale points (typically 5 or 7).

LO2-5

2.2 Level of Measurement

Likert Scales (examples)

"College-bound high school students should be required to study a foreign language." (check one)

❑	❑	❑	❑	❑
Strongly Agree	Somewhat Agree	Neither Agree Nor Disagree	Somewhat Disagree	Strongly Disagree

"How would you rate your Internet service provider?" (check one)

❑ Terrible ❑ Poor ❑ Adequate ❑ Good ❑ Excellent

LO2-4

2.2 Level of Measurement

Use the following procedure to recognize data types:

Question	If "Yes"
Q1. Is there a meaningful zero point?	Ratio data (statistical operations are allowed)
Q2. Are intervals between scale points meaningful?	Interval data (common statistics allowed, e.g., means and standard deviations)
Q3. Do scale points represent rankings?	Ordinal data (restricted to certain types of nonparametric statistical tests)
Q4. Are there discrete categories?	Nominal data (only counting allowed, e.g., finding the mode)

LO2-4

2.2 Level of Measurement

Changing Data By Recoding

- In order to simplify data or when exact data magnitude is of little interest, ratio data can be recoded *downward* into ordinal or nominal measurements (but not conversely).
- For example, recode systolic blood pressure as “normal” (under 130), “elevated” (130 to 140), or “high” (over 140).
- The above recoded data are ordinal (ranking is preserved), but intervals are unequal and some information is lost.

LO2-6

2.3 Sampling Concepts

LO2-6: Use the correct terminology for samples and populations

Sample or Census

- A ***sample*** involves looking only at some items selected from the population.
- A ***census*** is an examination of all items in a defined population.
- Why can't the United States Census survey every person in the population? – mobility, un-documented workers, budget constraints, incomplete responses, etc.

LO2-6

2.3 Sampling Concepts

TABLE 2.4 Sample or Census?

Situations Where a Sample May Be Preferred	*Situations Where a Census May Be Preferred*
Infinite Population No census is possible if the population is of indefinite size (an assembly line can keep producing bolts, a doctor can keep seeing more patients).	**Small Population** If the population is small, there is little reason to sample, for the effort of data collection may be only a small part of the total cost.
Destructive Testing The act of measurement may destroy or devalue the item (battery life, vehicle crash tests).	**Large Sample Size** If the required sample size approaches the population size, we might as well go ahead and take a census.
Timely Results Sampling may yield more timely results (checking wheat samples for moisture content, checking peanut butter for salmonella contamination).	**Database Exists** If the data are on disk, we can examine 100% of the cases. But auditing or validating data against physical records may raise the cost.
Accuracy Instead of spreading resources thinly to attempt a census, budget might be better spent to improve training of field interviewers and improve data safeguards.	**Legal Requirements** Banks must count *all* the cash in bank teller drawers at the end of each business day. The U.S. Congress forbade sampling in the 2000 decennial population census.
Cost Even if a census is feasible, the cost, in either time or money, may exceed our budget.	
Sensitive Information A trained interviewer might learn more about sexual harassment in an organization through confidential interviews of a small sample of employees.	

LO2-6

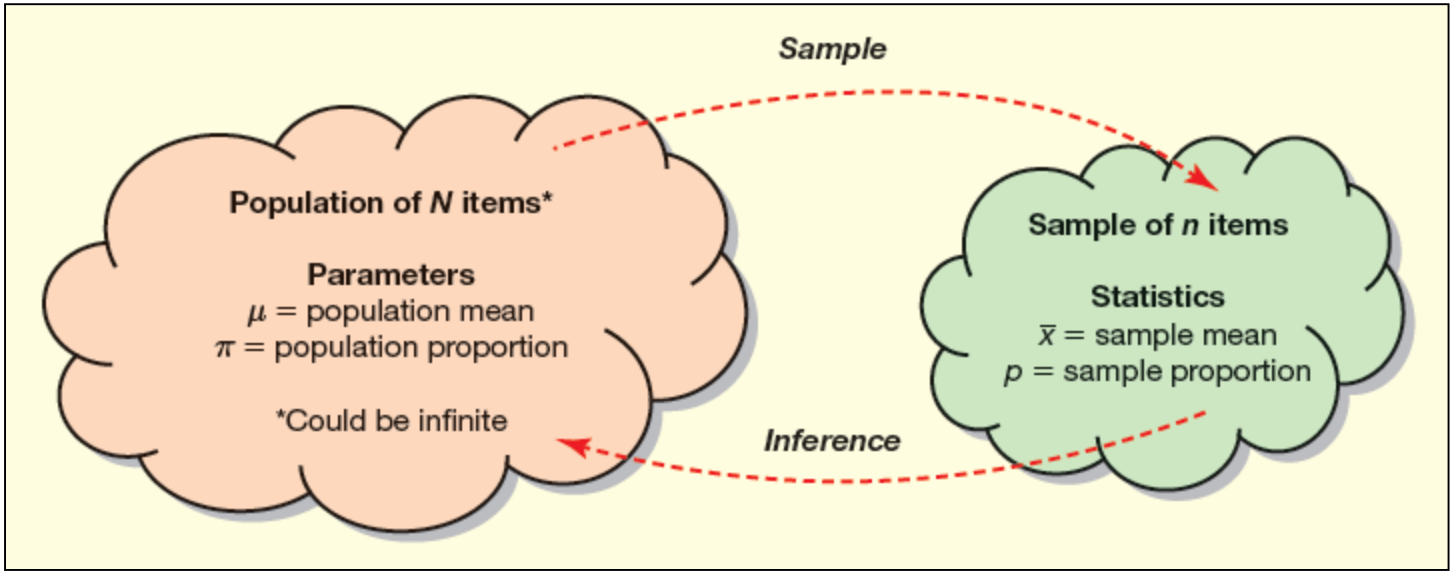
2.3 Sampling Concepts

Parameters and Statistics

- ***Statistics*** are computed from a sample of *n* items, chosen from a population of *N* items.
- Statistics can be used as estimates of ***parameters*** found in the population.
- Symbols are used to represent population parameters and sample statistics.

LO2-6

2.3 Sampling Concepts



Rule of Thumb: A population may be treated as infinite when N is at least 20 times n (i.e., when $N/n \geq 20$).

LO2-6

2.3 Sampling Concepts

Target Population

- The population must be carefully specified and the sample must be drawn scientifically so that the sample is representative.
- The ***target population*** is the population we are interested in (e.g., U.S. gasoline prices).
- The ***sampling frame*** is the group from which we take the sample (e.g., 115,000 stations).
- The frame should not differ from the target population.

LO2-7

2.4 Sampling Methods

LO2-7: Explain the common sampling methods and how to implement them

Random Sampling Methods

Simple Random Sample	Use random numbers to select items from a list (e.g., Visa cardholders).
Systematic Sample	Select every *k*th item from a list or sequence (e.g., restaurant customers).
Stratified Sample	Select randomly within defined strata (e.g., by age, occupation, gender).
Cluster Sample	Select random geographical regions (e.g., zip codes) that represent the population.

LO2-7

2.4 Sampling Methods

Non-random Sampling Methods

Judgment Sample	Use expert knowledge to choose "typical" items (e.g., which employees to interview).
Convenience Sample	Use a sample that happens to be available (e.g., ask co-workers' opinions at lunch).
Focus Groups	In-depth dialog with a representative panel of individuals (e.g., iPod users).

LO2-7

2.4 Sampling Methods

With or Without Replacement

- If we allow duplicates when sampling, then we are sampling ***with replacement***.
- Duplicates are unlikely when *n* is much smaller than large *N*.
- If we do not allow duplicates when sampling, then we are sampling ***without replacement***.

LO2-7

2.4 Sampling Methods

Computer Methods: Examples of alternative ways to choose 10 integers between 1 and 875.

Excel—Option A	Enter the Excel function =RANDBETWEEN(1,875) into 10 spreadsheet cells. Press F9 to get a new sample.
Excel—Option B	Enter the function =INT(1+875*RAND()) into 10 spreadsheet cells. Press F9 to get a new sample.
Internet	The website www.random.org will give you many kinds of excellent random numbers (integers, decimals, etc).
MINITAB	Use MINITAB's Random Data menu with the Integer option.
Pocket Calculator	Press the RAND key to get a random number in the interval [0,1], multiply by 875, then round up to the next integer.

These are *pseudo-random* generators because even the best algorithms eventually repeat themselves.

LO2-7

2.4 Sampling Methods

Row – Column Data Arrays

- When the data are arranged in a rectangular array, an item can be chosen at random by selecting a row and column.
- For example, in the 4 x 3 array, select a random column between 1 and 3 and a random row between 1 and 4.
- This way, each item has an equal chance of being selected.

LO2-7

2.4 Sampling Methods

Randomizing a List

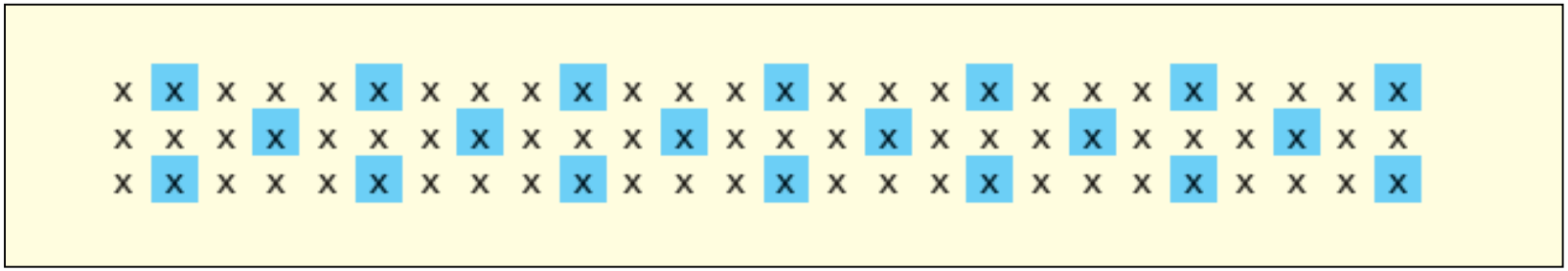
- In Excel, use function =RAND() beside each row to create a column of random numbers between 0 and 1.
- Copy and paste these numbers into the same column using *Paste Special > Values* in order to paste only the values and not the formulas.
- Sort the spreadsheet on the random number column.

LO2-7

2.4 Sampling Methods

Systematic Sampling

- Sample by choosing every *kth* item from a list, starting from a randomly chosen entry on the list.
- For example, starting at item 2 (see below), we sample every 4 items to obtain a sample of $n = 20$ items from a list of $N = 78$ items.



Note that $N/n = 78/20 \approx 4$ (periodicity).

LO2-7

2.4 Sampling Methods

Stratified Sampling

- Utilizes prior information about the population.
- Applicable when the population can be divided into relatively homogeneous subgroups of known size (***strata***).
- A simple random sample of the desired size is taken within each ***stratum***.
- For example, from a population containing 55% males and 45% females, randomly sample from 110 males and 90 females (n = 200).

LO2-7

2.4 Sampling Methods

Cluster Sample

- Strata consist of geographical regions.
- ***One-stage*** cluster sampling – sample consists of all elements in each of *k* randomly chosen subregions (clusters).
- ***Two-stage*** cluster sampling, first choose *k* subregions (clusters), then choose a random sample of elements within each cluster.

LO2-7

2.4 Sampling Methods

Cluster Sample

- Here is an example of 4 elements sampled from each of 3 randomly chosen clusters (two-stage cluster sampling).

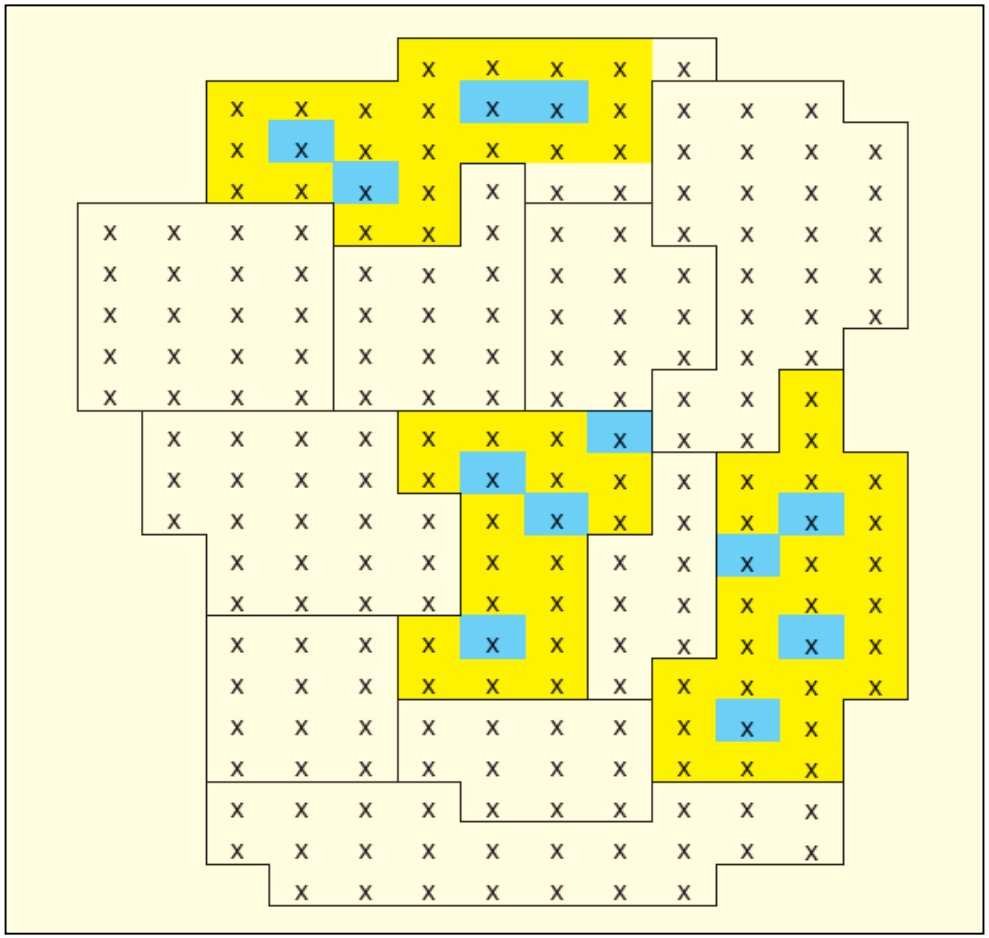


FIGURE 2.7

Two-Stage Cluster Sampling: Randomly choose three clusters, then randomly choose four items in each cluster

LO2-7

2.4 Sampling Methods

Judgment Sample

- A non-probability sampling method that relies on the expertise of the sampler to choose items that are representative of the population.
- Can be affected by subconscious bias

  (i.e., *non-randomness* in the choice).
- *Quota sampling* is a special kind of judgment sampling, in which the interviewer chooses a certain number of people in each category.

LO2-7

2.4 Sampling Methods

Convenience Sample

- Take advantage of whatever sample is available at that moment. A quick way to sample.

Focus Groups

- A panel of individuals chosen to be representative of a wider population, formed for open-ended discussion and idea gathering.

LO2-8

2.5 Data Sources

LO2-8: Find everyday print or electronic data sources.

- One goal of a statistics course is to help you learn where to find data that might be needed. Fortunately, many excellent sources are widely available. Some sources are given in the following table.

TABLE 2.10

Useful Data Sources

Type of Data	*Examples*
U.S. job-related data	*U.S. Bureau of Labor Statistics*
U.S. economic data	*Economic Report of the President*
Almanacs	*World Almanac, Time Almanac*
Periodicals	*Economist, Bloomberg Businessweek, Fortune, Forbes*
Indexes	*The New York Times, The Wall Street Journal*
Databases	Compustat, Citibase, U.S. Census
World data	*CIA World Factbook*
Web	Google, Yahoo!, MSN

LO2-9

2.6 Surveys

LO2-9: Describe basic elements of survey types, survey designs, and response scales.

Basic Steps of Survey Research

- Step 1: State the goals of the research.
- Step 2: Develop the budget (time, money, staff).
- Step 3: Create a research design (target population, frame, sample size).
- Step 4: Choose a survey type and method of administration.

LO2-9

2.6 Surveys

Basic Steps of Survey Research

- Step 5: Design a data collection instrument (questionnaire).
- Step 6: Pretest the survey instrument and revise as needed.
- Step 7: Administer the survey (follow up if needed).
- Step 8: Code the data and analyze it.

LO2-9

2.6 Surveys

Survey Types

- Mail
- Telephone
- Interviews
- Web
- Direct observation

Survey Guidelines

- Planning
- Design
- Quality
- Pilot test
- Buy-in
- Expertise

LO2-9

2.6 Surveys

Questionnaire Design

- Use a lot of white space in layout.
- Begin with short, clear instructions.
- State the survey purpose
- Assure anonymity.
- Instruct on how to submit the completed survey.
- Break survey into naturally occurring sections.
- Let respondents bypass sections that are not applicable (e.g., “if you answered no to question 7, skip directly to Question 15”).

LO2-9

2.6 Surveys

Questionnaire Design

- Pretest and revise as needed.
- Keep as short as possible.

Types of Questions

Open-ended
Fill-in-the-blank
Check boxes
Ranked choices
Pictograms
Likert scale

LO2-9

2.6 Surveys

Question Wording

- The way a question is asked has a profound influence on the response. For example,

1. Shall state taxes be cut?

2. Shall state taxes be cut, if it means reducing highway maintenance?

3. Shall state taxes be cut, if it means firing teachers and police?

LO2-9

2.6 Surveys

Question Wording

- Make sure you have covered all the possibilities. For example,

  Are you married? ❑ Yes ❑ No

- Overlapping classes or unclear categories are a problem. What if your father is deceased or is 45 years old.

  How old is your father?
  - ❑ 35 – 45
  - ❑ 45 – 55
  - ❑ 55 – 65
  - ❑ 65 or older

LO2-9

2.6 Surveys

Data Screening

- Responses are usually coded numerically (e.g., 1 = male, 2 = female).
- Missing values are typically denoted by special characters (e.g., blank, “.” or “*”).
- Discard questionnaires that are flawed or missing many responses.
- Watch for multiple responses, outrageous or inconsistent replies or out-of-range answers.
- Followup if necessary and always document your data-coding decisions.