Chapter 1: Sampling and Data

1.1: Definitions of Statistics, Probability, and Key Terms

Key Terms:

- **Statistics** - the science of planning studies and experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions based on the data.

- **Data** - collections of observations.

- **Descriptive Statistics** - organizing and summarizing data; by graphing and by numerical values (such as an average).

- **Inferential Statistics** - uses methods that take a result from a sample, extend it to the population, and measure the reliability of the result.

- **Probability** - the chance of an event occurring.

- **Population** - the complete collection of *all* individuals to be studied.
  Example:

- **Sample** - a subcollection of members selected from a population.
  Example:

- **Sampling** - selecting a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

- **Parameter** - a numerical measurement describing some characteristic of a *population*.

  Ex: The average age of students enrolled as an undergraduate at Texas Woman's University is 21.1 years.

- **Statistic** - a numerical measurement describing some characteristic of a *sample*.

  Ex: The average age of 500 TWU undergraduates selected at random is 20.3 years.

Example 1: (State Elections of Texas) The comptroller (management level position responsible for supervising the quality of accounting and financial reporting of an organization) of Texas wants to know how she will do in the upcoming State election. She hires a consulting firm and they polled 3,150 registered voters. The consulting firm finds that 57% of the polled voters will vote for her. It turns out that after the election she wins with 60% of the vote. Identify the parameter and the statistic in this problem.

- **Representative Sample** - the idea that the sample must contain the characteristics of the population. One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter.

- **Variable** - a characteristic or measurement that can be determined for each member of a population.

- **Mean** - or "average."

- **Proportion** - part out of the whole/total.

Example 2. For each research objective, identify the population and sample in the study.

(a) The research organization contacts 1,500 teenagers who are 13 to 17 years of age and live in the United States and asks whether or not they had been prescribed medications for any mental disorders, such as depression or anxiety.

(b) A quality-control manager randomly selects 50 bottles of soda that were filled on October 15 to assess the calibration of the filling machine.

(c) A farmer wanted to learn about the weight of his soybean crop. He randomly sampled 100 plants and weighed the soybeans on each plant.

(d) Every year the U.S. Census Bureau releases the *Current Population Report* based on a survey of 50,000 households. The goal of this report is to learn the demographic characteristics of all households within the United States, such as income.

(e) A large community college has noticed that an increasing number of full-time students are working while attending the school. The administration randomly selects 128 students and asks this question: How many hours per week do you work?

Example 3. Determine whether the underlined value is a parameter or a statistic.

(a) Following the 2006 national midterm election, 18% of the governors of the 50 United States were female.

(b) The average score for a class of 55 students taking a statistics midterm exam was 68%.

(c) In a national survey of high school students (grades 9 to 12), 25% of respondents reported that someone had offered, sold, or given them an illegal drug on school property.

(d) In a national survey on substance abuse, 66.4% of respondents who were full-time college students aged 18 to 22 reported using alcohol within the past month.

(e) Ty Cobb is one of Major League Baseball's greatest hitters of all time, with a career batting average of 0.366.

(f) Only 12 men have walked on the moon. The average age of these men at the time of their moonwalks was 39 years, 11 months, 15 days.

(g) A study of 5,984 adults in public rest rooms (in 4 major US cities) found that 21% did not wash their hands before exiting.

(h) Telephone interviews of 1,433 adults 18 years of age or older, conducted nationwide March 16-30, 2013, found that only 72% could identify the current vice-president.

Example 4. Determine whether each scenario is referring to a mean or a proportion.

(a) More than 30% of the registered voters in Denton County voted in the primary election.

(b) We want to compare the heights of men to the heights of women.

(c) In a recent poll, 73 out of 100 people said they access at least one social media account daily.

(d) Males and females were asked about what they would do if they received a $100 bill by mail, addressed to their neighbor, but wrongly delivered to them. Would they return it to their neighbor? Of the 69 males sampled, 52 said yes and of the 131 females sampled, 120 yes.

1.2: Data, Sampling, and Variation in Data and Sampling

More Terms

- **Quantitative** (or **numerical**) **data** - data that consists of numbers representing counts or measurements.

  Ex: height, weight, years of education, GPA, test scores.

- **Qualitative** (or **Categorical**) **data** - data that consists of names or labels that are not numbers representing counts or measurements.

  Ex: gender, hometown, ethnicity, zip code.

- **Discrete data** - quantitative data which results when the number of possible values is either a finite number or a countable number.

  Ex: The integers from 1 to 100.

- **Continuous data** - quantitative data which results when there are infinitely many possible values corresponding to some continuous scale that covers a range of values without gaps, interruptions, or jumps.

  Ex: All values from 0 to 1.

Sampling Methods

- **Simple random sample** - A sample of n subjects selected in such a way that every possible sample of the same size n has the same chance of being chosen.

- **Systematic sample** - A sample in which the researcher selects some starting point and then selects every kth element in the population.

- **Stratified sample** - A sample in which the researcher subdivides the population into at least two different subgroups (or strata), and then draws a sample from each subgroup.

- **Cluster sample** - A sample in which the researcher first divides the population into sections (or clusters), and then randomly selects all members from *some* of those clusters.

- **Convenience sample** - A sample in which the researcher simply uses results that are very easy to get. This is not a valid sampling method and will likely result in biased data.

Example 5: Taxpayers in a population are listed in order of increasing income. In each situation, determine which type of sampling method was used.

1. A sample is selected by first separating the taxpayers into 4 groups based on income, and then randomly sampling 50 people from each group.

2. A sample is selected by randomly choosing one of the first 100 names, then choosing every 100^{th} name from that point forward.

3. A random number generator in excel is used to select a sample of size 100.

4. A sample is selected by first separating the taxpayers into 4 groups based on income, and then sampling all members in two of the groups.

5. To save time a sample of the first 100 names is used.

Bias - if the results of the sample are not representative of the population.

Sources of Bias in Sampling

- **Sampling bias** - the technique used to obtain the individuals to be in the sample tends to favor one part of the population over another.

  Ex. Convenience sampling has bias - not chosen at random.

  Ex. Undercoverage - the proportion of one segment of the population is lower in the sample than it is in the population.

- **Nonresponse bias** - when individuals selected to be in the sample who do not respond to a survey have different opinions from those who do.

  Ex. They do not respond, or the interviewer can't contact them.

  How to Control nonresponse bias: callback, using rewards or incentives.

- **Response bias** - when answers on a survey do not reflect the true feelings of the respondent.

  Interview error: a trained interviewer is essential to obtain accurate information. They will have the skill necessary to elicit responses and make the interviewee feel comfortable.

  Misrepresented Answers: some survey questions result in responses that misrepresent facts or are flat-out lies.
  Ex. when asked salaries of graduates, reports are inflated.
  Ex. when asked how many push-ups one can do, then ask them to do it.

  Loaded Questions: The wording and presentation of questions plays a large role in the type of response given to the question. The way a question is worded can lead to response bias, so they must always be asked in balanced form.
  Ex. "Do you think the practice of aborting innocent unborn children should be made illegal?" vs. "Do you think the practice of aborting unborn children should be made illegal or remain legal?"

  Ordering of Questions/Words: Questions can be unintentionally loaded by the order of items being considered. Many surveys rearrange the order of the questions within a questionnaire so that responses are not affected by prior questions.

  Data-entry error: not technically a result of response bias, data-entry errors will lead to results not representative of the population.
  Ex. Most data gets input into computers, and it is easy to type 93 instead of 39.

Sampling Error vs. Nonsampling Error

- **Nonsampling error** - errors that result from undercoverage, nonresponse bias, response bias, or data-entry error.

- **Sampling error** - errors that result from using a sample to estimate information about a population. This type of error occurs because a sample gives incomplete information about a population.

Common problems to be aware of:

- Problems with samples: A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and invalid.

- Self-selected samples: Responses only by people who choose to respond, such as call-in surveys, are often unreliable.

- Sample size issues: Samples that are too small may be unreliable. Larger samples are better, if possible. In some situations, having small samples is unavoidable and can still be used to draw conclusions. Examples: crash testing cars or medical testing for rare conditions.

- Undue influence: collecting data or asking questions in a way that influences the response.

- Non-response or refusal of subject to participate: The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.

- Causality: A relationship between two variables does not mean that one causes the other to occur. They may be related (correlated) because of their relationship through a different variable.

- Self-funded or self-interest studies: A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good, but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.

- Misleading use of data: improperly displayed graphs, incomplete data, or lack of context.

- Confounding: When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

Example 6. Classify each variable as qualitative or quantitative.

(a) Nation of origin

(b) Number of siblings

(c) Grams of carbohydrates in a doughnut

(d) Number on a football player's jersey

(e) Number of unpopped kernels in a bag of ACT microwave popcorn

(f) Assessed value of a house

(g) Phone number

(h) Student ID number

Example 7. Determine whether the quantitative variable is discrete or continuous.

(a) Runs scored in a season by Derek Jeter

(b) Volume of water lost each day through a leaky facuet

(c) Length (in seconds) of a country song

(d) Number of sequoia trees in a randomly selected acre of Yosemite National Park

(e) Temperature on a randomly selected day in Denton, Texas

(f) Internet connection speed in kilobytes per second

(g) Points scored in an NCAA basketball game

(h) Air pressure in pounds per square inch in an automobile tire

Example 8. Identify the type of sampling method used.

(a) To estimate the percentage of defects in a recent manufacturing batch, a quality-control manager at Intel selects every 8th chip that comes off the assembly line starting with the 3rd until she obtains a sample of 140 chips.

(b) To determine the prevalence of human growth hormone (HGH) use among high school varsity baseball players, the State Athletic Commission randomly selects 50 high schools. All members of the selected high schools' varsity baseball teams are tested for HGH.

(c) A member of Congress wishes to determine her constituency's opinion regarding estate taxes. She divides her constituency into three income classes: low-income households, middle-income households, and upper-income households. She then takes a simple random sample of households from each income class.

(d) In an effort to identify if an advertising campaign has been effective, a marketing firm conducts a nationwide poll by randomly selecting individuals from a list of known users of the product.

(e) A farmer divides his orchard into 50 subsections, randomly selects 4, and samples all the trees within the 4 subsections to approximate the yield of his orchard.

(f) A school official divides the student population into five classes: freshman, sophomore, junior, senior, and graduate student. The official takes a simple random sample from each class and asks the members' opinions regarding student services.

Example 9. Determine the type of bias. If you choose 'response bias', provide a reason.

(a) A retail store manager wants to conduct a study regarding the shopping habits of its customers. He selects the first 60 customers who enter his store on a Saturday morning.

(b) An antigun advocate wants to estimate the percentage of people who favor stricter gun laws. He conducts a nationwide survey of 1,203 randomly selected adults 18 years old and older. The interviewer asks the respondents, "Do you favor harsher penalties for individuals who sell guns illegaly?"

(c) Suppose you are conducting a survey regarding the sleeping habits of students. From a list of registered students, you obtain a simple random sample of 150 students. One survey question is "How much sleep do you get?"

(d) A polling organization conducts a study to estimate the percentage of households that speaks a foreign language as the primary language. It mails a questionnaire to 1,023 randomly selected households throughout the United States and asks the head of household if a foreign language is the primary language spoken in the home. Of the 1,023 households selected, 12 responded.

1.3: Frequency, Frequency Tables, and Levels of Measurement

After collecting data it is often useful and informative to summarize the data by constructing a frequency distribution.

- **Frequency** - count of how many observations fall into a category.

- **Frequency distribution** - a listing of all categories along with their frequencies. Frequency distributions can be used with both qualitative and quantitative data. The definitions below are used in discussing and constructing frequency distributions for quantitative data.

- **Relative frequency** - the proportion or percentage of the count in a category relative to the total number of items in all categories.

$$\text{relative freq.} = \frac{\text{class frequency}}{\text{sum of all frequencies}}$$

- **Relative frequency distribution** - listing of all categories along with their relative frequencies. *The sum of all the relative frequencies must be 1.* (If rounding was used the sum may not be exactly 1, but should be close).

- **Cumulative frequency distribution** - the sum of the frequencies for that class and all previous classes.

Example 10: The table below lists frequencies of the amount of cash each student had in his or her pocket.

Money in $	Frequency
0-4	5
5-9	3
10-14	8
15-19	6
20-24	10
25-29	7

Use the frequency distribution table to construct the relative frequency distribution and the cumulative frequency distribution.

Money in $	Frequency	Rel. Freq.	Cum. Freq.
0-4	5		
5-9	3		
10-14	8		
15-19	6		
20-24	10		
25-29	7		

1.4: Experimental Design and Ethics

- **Explanatory variable** - The variable whose effect you want to study; the independent variable.

- **Response variable** - The variable that you suspect is affected by the other variable; the dependent variable.

- The response variable is often considered the outcome of interest, but we think those outcomes may be influenced by the explanatory variable. Often the explanatory variable comes first in a time sequence.

  For example, how hard a golf club is swung is used to determine how far the ball travels when hit. "Club head speed" is the explanatory variable and "distance the ball travels" is the response variable.

- **Experimental Unit** - a single object or individual to be measured.

Example 11. Determine the explanatory and response variable in each scenario.

(a) A researcher is curious if the number of hours spent doing homework has an effect on the grade a student earns on an exam.

- explanatory:

- response:

(b) In a study to determine whether surgery or chemotherapy results in higher survival rates for a certain type of cancer, whether or not the patient survived is one variable, and whether they received surgery or chemotherapy is the other.

- explanatory:

- response:

(c) A large industrial plant has seven divisions that do the same type of work. A safety inspector visits each division of 20 workers quarterly. The number of work-hours devoted to safety training and the number of work-hours lost due to industry-related accidents are recorded for each separate division.

- explanatory:
- response:

(d) Suppose you conducted an experiment to see how tall plants grew given different amounts of water.

- explanatory:
- response:

(e) Suppose a questionnaire was given to college students asking about their attitude, and they also indicated how far their college is from home. A researcher is interested in assessing the degree of homesickness of a student, based on the distance from home.

- explanatory:
- response: