CHAPTER 1: SAMPLING AND DATA

OUTLINE

1.1 Definitions of Statistics, Probability, and Key Terms

1.2 Data, Sampling, and Variation in Data and Sampling

1.3 Frequency, Frequency Tables, and Levels of Measurement

1.4 Experimental Design and Ethics

SECTION 1.1
DEFINITIONS OF STATISTICS, PROBABILITY, AND KEY TERMS

SECTION 1-1 INTRODUCTION

Most people become familiar with probability and statistics through various media (radio, TV, Internet, newspapers, and magazines)

- Nearly one in seven US families are struggling with bills from medical expenses even though they have health insurance
- About 15% of men in the US are left-handed and 9% of women are left-handed
- The median age of couples who watch Jay Leno is 48.1 years
- Eating 10 grams of fiber a day reduces the risk of heart attack by 14%

STATISTICS IS EVERYWHERE

Statistics is used in almost ALL fields of human endeavor.

- Sports: a statistician may keep records of the number of yards a running back gains during the football game OR number of hits a baseball player gets in a season
- Public Health: an administrator might be concerned with the number of residents who contract a new strain of flu virus
- Education: a researcher might want to know if new teaching methods are better than old ones.
- Quality Control
- Prediction

WHY SHOULD WE STUDY STATISTICS?

To be able to read and understand various statistical studies performed in their fields—requires a knowledge of the vocabulary, symbols, concepts, and statistical procedures

To conduct research in their fields—requires ability to design experiments which involves collection, analysis, and summary of data

To become better consumers and citizens

IN THIS CHAPTER, WE WILL INTRODUCE THE BASIC CONCEPTS OF PROBABILITY AND STATISTICS BY ANSWERING THE FOLLOWING:

1. WHAT ARE THE BRANCHES OF STATISTICS?
2. WHAT ARE DATA?
3. HOW ARE SAMPLES SELECTED?

WHAT IS STATISTICS?

Statistics **IS** the science of gathering, describing, and analyzing data

OR

Statistics **ARE** the actual numerical descriptions of sample data

"LANGUAGE OF STATISTICS"

Variable: a characteristic or value that changes among members of the group.

Variables whose values are determined by chance are called **random variables**

Variables may be **numerical** or **categorical.** Numerical variables take on values with equal units such as weight in pounds and time in hours. Categorical variables place the person or thing into a category

Data: the counts, measurements, or observations gathered about a specific variable in a group in order to study it.

Datum: is a single value

POPULATION VS SAMPLE

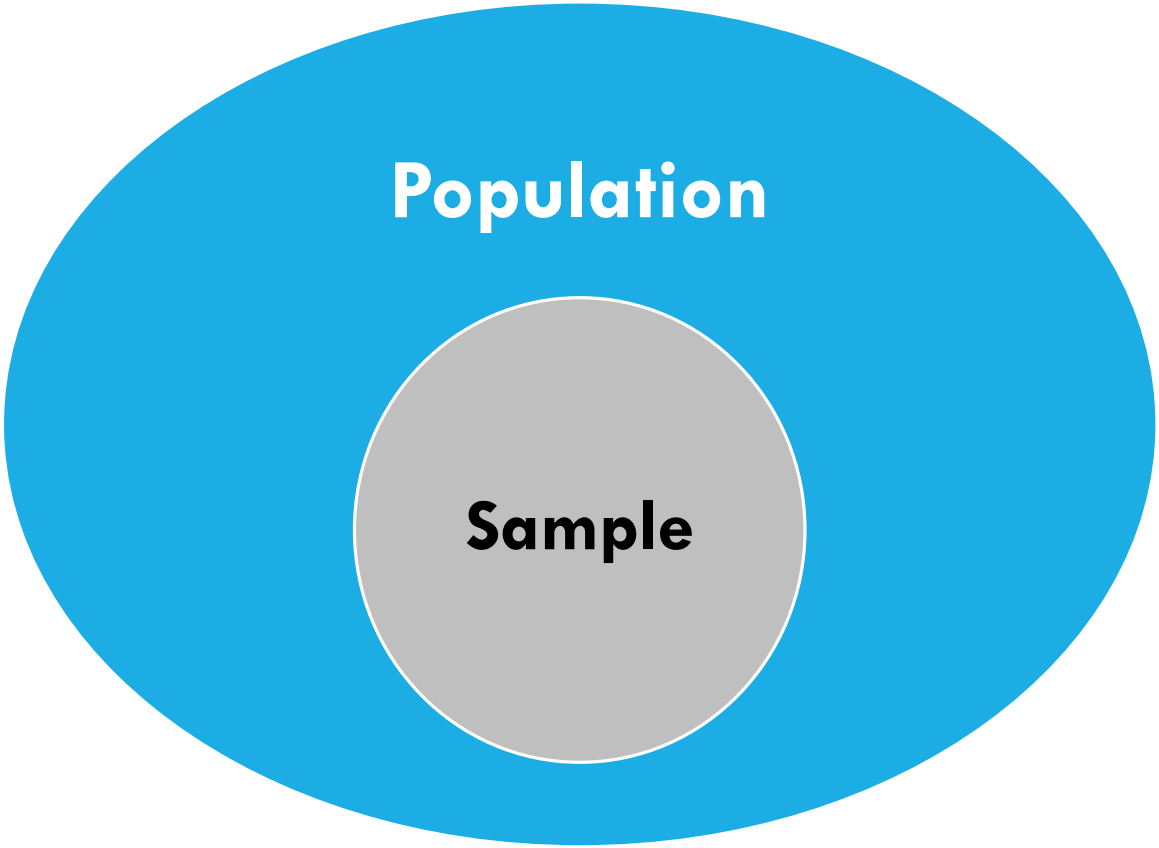
Population

- is the particular group of interest
- consists of ALL persons or things being studied
- Examples
  - All 202, 682,345 adult Americans
  - All 5257 students enrolled at GHC during Spring 2011
  - The governors of the 50 United States

Sample

- is a subset of the population from which data is collected
  - Care must be taken in choosing the sample so that the population is represented well and the results of the study are meaningful
- Examples
  - 1000 adult Americans surveyed to determine if he/she favors the legalization of marijuana
  - 28 GHC students in Mrs. Ralston's class surveyed to determine height

POPULATION VS SAMPLE



PARAMETER VS STATISTIC

Parameter

is a numerical description of a particular population characteristic

is a fixed number, BUT it is often impossible or impractical to determine it precisely (because of human limitations, usually the best we can do is estimate a population's true parameter)

Statistic

is the actual numerical description of a particular sample

can vary from sample to sample

The sample must contain the characteristics of the population in order to be a **representative sample.**

POPULATION VS SAMPLE

POPULATION	SAMPLE
Whole group	**Part** of the group
Group **I want to know** about	Group **I do know** about
Characteristics are called **parameters**	Characteristics are called **statistics**
Parameters are generally **unknown**	Statistics are always **known**
Parameter is **fixed**	Statistics **change** with the sample

EXAMPLES

Determine whether the statement describes a population or sample:

1. The ACT score of all students in Mrs. Ralston's MATH 2200 class
2. The television shows watched by 1045 families from across the US for the Nielsen ratings
3. The height of 15 out of 30 plants in a green house

EXAMPLES

For each scenario, identify the population being studied and the sample chosen

4. An education professor wants to know the hometowns of students attending Yale University. She obtains a list of registered students from the registrar's office and randomly chooses 300 students to survey.
5. A hotel chain wants to build a new facility. The board of directors uses a list of the top 100 vacation spots in America and visits 20 of the cities on the list to determine their feasibility as the new hotel's location.

EXAMPLES

Identify if the numerical value in each statement describes a parameter or statistic

6. Thirty-five percent of all graduating seniors from Kennesaw State University receive a degree in business.

7. The average height of a sample of men entering the armed forces is 6.1 feet.

8. The students in Mrs. Ralston's Math 1111 classes have an average of 2.5 siblings

TWO BRANCHES OF STATISTICS

Descriptive Statistics

As a science, involves the collection, organization, summarization, and presentation of data

Involves raw data, as well as graphs, tables, and numerical summaries

"Just the facts"

Refer to sample without making any assumptions about the population

Inferential Statistics

As a science, involves using descriptive statistics to estimate population parameters

Deals with interpretation of the information collected

Usually used in conjunction with descriptive statistics within a statistical study

EXAMPLES

Decide if the following statements are examples of descriptive or inferential statistics:

9. Eighty-two percent of the employees from a small local company attended the annual company picnic.
10. The average age of entering freshman at the University of Georgia is 20.8 years old, based on the information from the registrar's office.
11. The average number of vacationers spend in national parks during the summer months is 4.5 hours.

ANSWERS:

1. Population
2. Sample
3. Sample
4. P: All registered students
   S: 300 students selected
5. P: Top 100 Vacation Spots in America
   S: 20 cities selected
6. Parameter
7. Statistic
8. Parameter
9. Descriptive Statistics
10. Descriptive Statistics
11. Inferential Statistics

OTHER TERMS IN STATISTICS

Probability is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring.

Two words that come up often in statistics are **mean** and **proportion.**

If you were to take three exams in your math classes and obtain scores of 86, 75, and 92, you would calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). Also referred to as the **average**.

If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is 22/ 40 and the proportion of women students is 18 /40 .

SECTION 1.2
DATA, SAMPLING, AND VARIATION IN DATA AND SAMPLING

INTRODUCTION TO CLASSIFYING DATA

Just as animals can be classified into phylum and then further into species, data collected in a statistical study can be classified into different categories.

The different categories group data based on the type of statistical analysis that can be performed on the data. Therefore, knowing the classification of a set of data is the first step in any statistical process.

QUALITATIVE VS. QUANTITATIVE DATA

Qualitative Data (aka Categorical Data)

- Consist of labels or descriptions of traits
- Typically non-numeric, but not a requirement
- Examples:
  - Eye Color
  - Gender
  - Religious Preference
  - Yes/No
  - Hometown
  - Favorite Food
  - ID numbers (SS#, GHC#)

Quantitative Data

- Consist of counts or measurements
- Numerical
- Examples:
  - Heights
  - Weights
  - Pulse Rate
  - Age
  - Body Temperatures
  - Credit Hours
  - Test Scores
  - Average rainfall

EXAMPLES

Classify the following data as either qualitative or quantitative

1) The number of homes a bank repossesses in four randomly selected months
2) Five hundred people are asked the frequency with which they eat chocolate (never, seldom, occasionally, or frequently)
3) A McDonald's quality control inspector counts the number of fries in 40 individual servings
4) The license plate of a car

QUANTITATIVE VARIABLES CAN BE FURTHERED CLASSIFIED

Discrete Variables

Can be assigned values such as 0, 1, 2, 3

"Countable"

"Number of"

Examples:

- Number of children
- Number of credit cards
- Number of calls received by switchboard
- Number of students

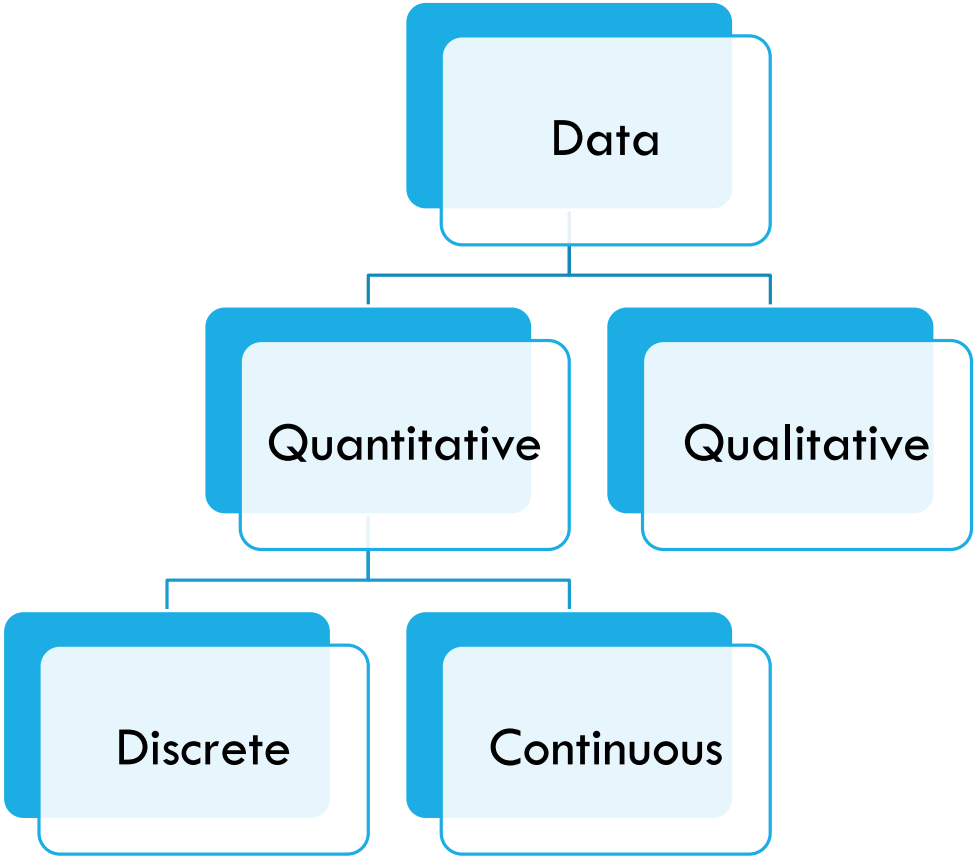
Continuous Variables

- Can assume an infinite number of values between any two specific values
- Obtained by measuring
- Often include fractions and decimals
- Examples:
  - Temperature
  - Height
  - Weight

EXAMPLES

Determine whether the following data are continuous or discrete:

5) The number on the uniform of a football player
6) The temperature in Celsius in Paris, France
7) The total weight of sugar imported by the United States each day
8) The prices of 50 randomly selected new cars


Data
Quantitative
Qualitative
Discrete
Continuous

QUALITATIVE DATA DISCUSSION

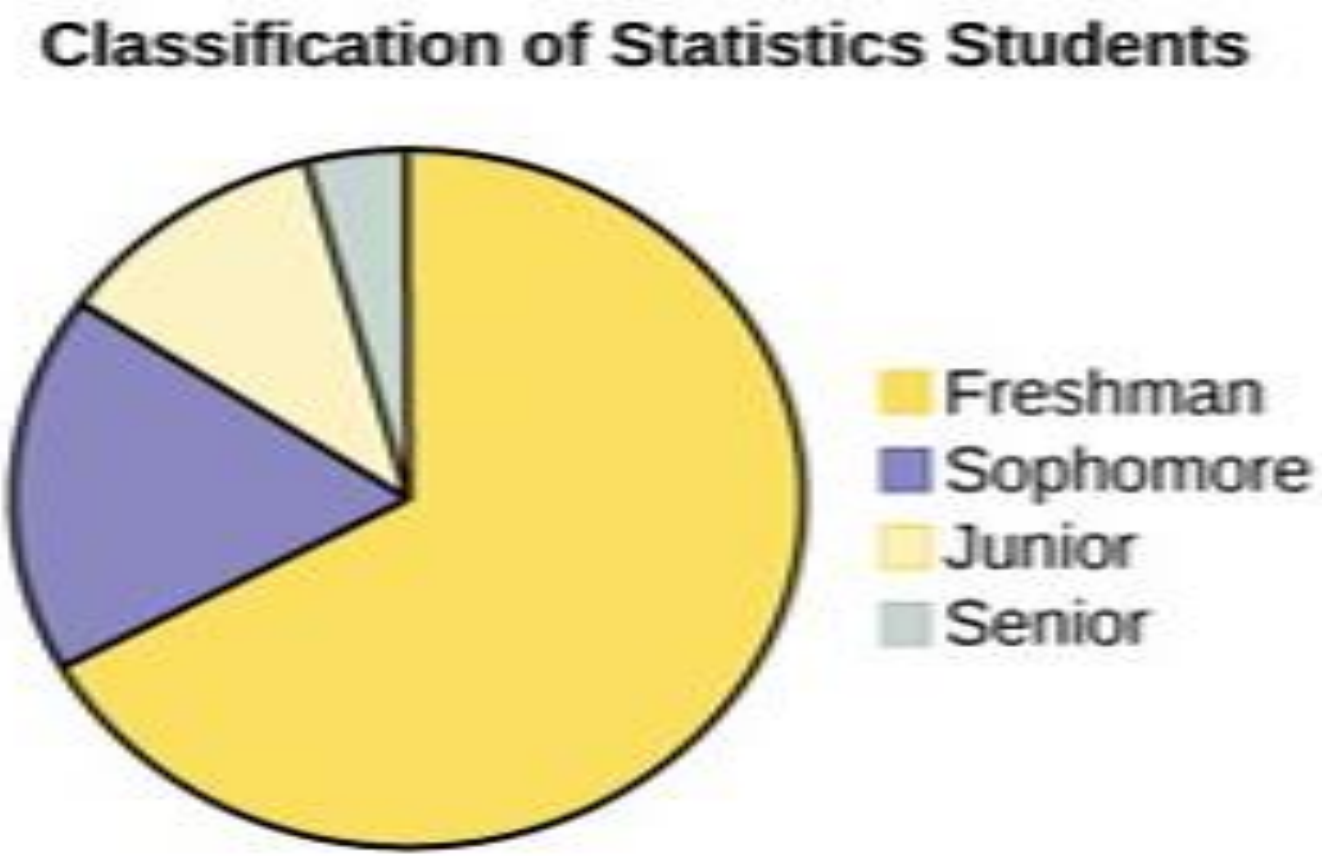
Tables are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data. There are no strict rules concerning which graphs to use.

Two graphs that are used to display qualitative data are **pie charts** and **bar graphs**.

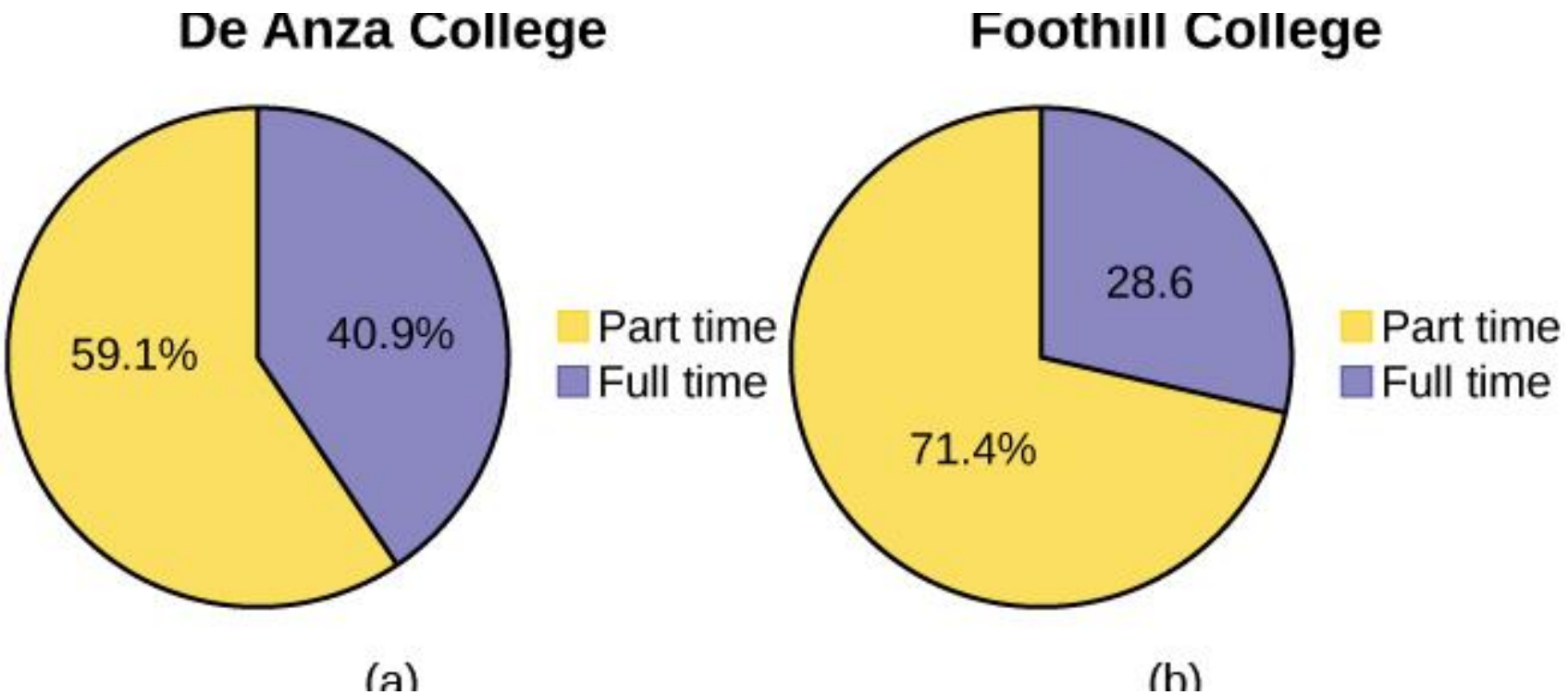
In a **pie chart**, categories of data are represented by wedges in a circle and are proportional in size to the percent of individuals in each category.

In a **bar graph**, the length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal. A **Pareto chart** consists of bars that are sorted into order by category size (largest to smallest).

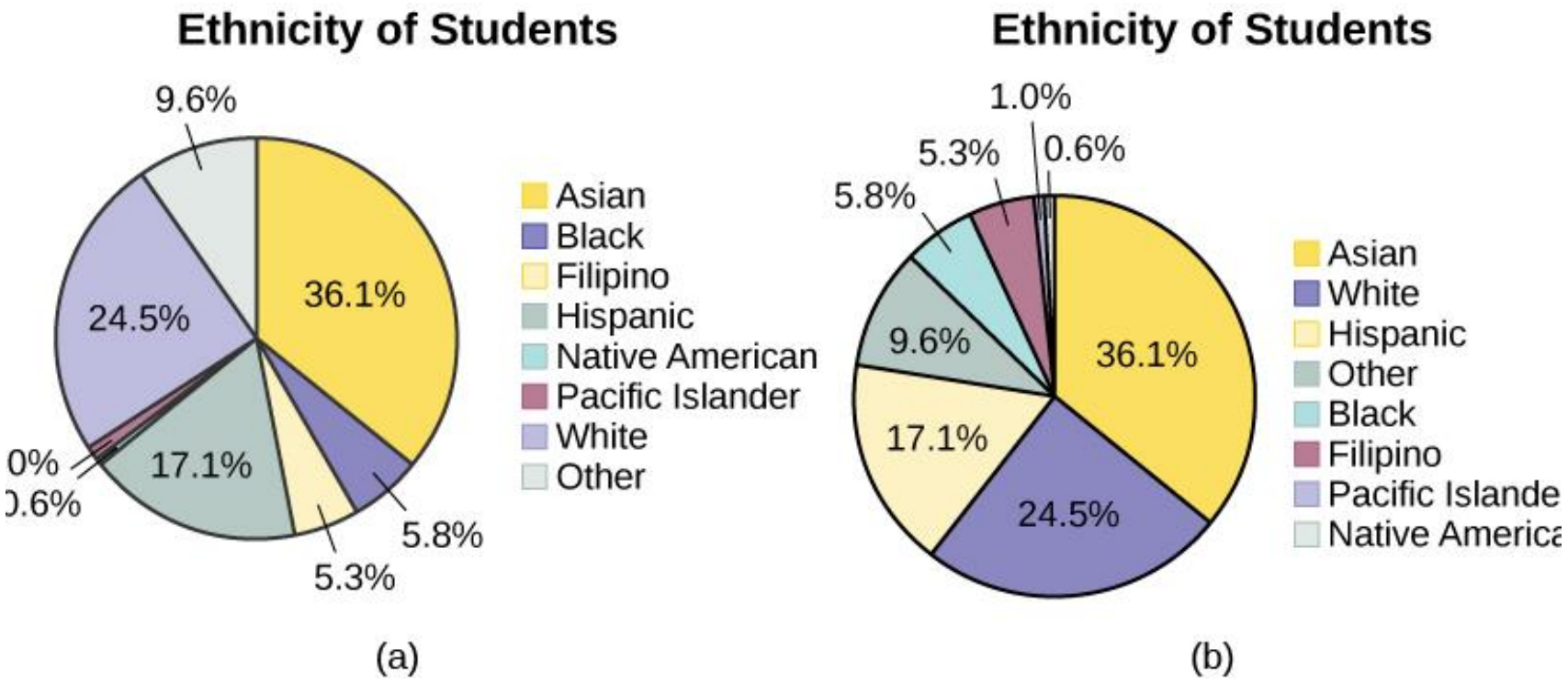
PIE CHART


Classification of Statistics Students
Freshman
Sophomore
Junior
Senior

SIDE BY SIDE PIE CHART

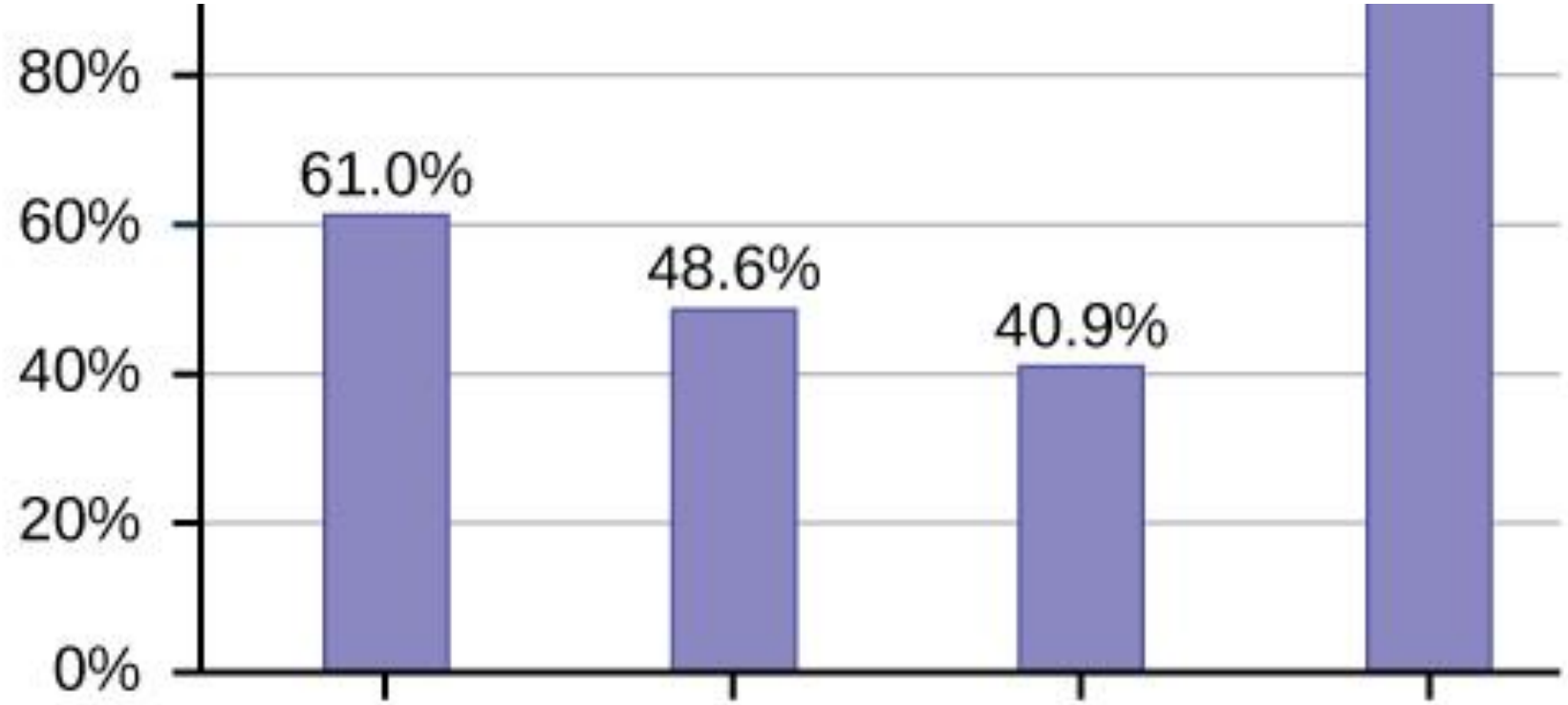


PIE CHARTS: NO MISSING DATA



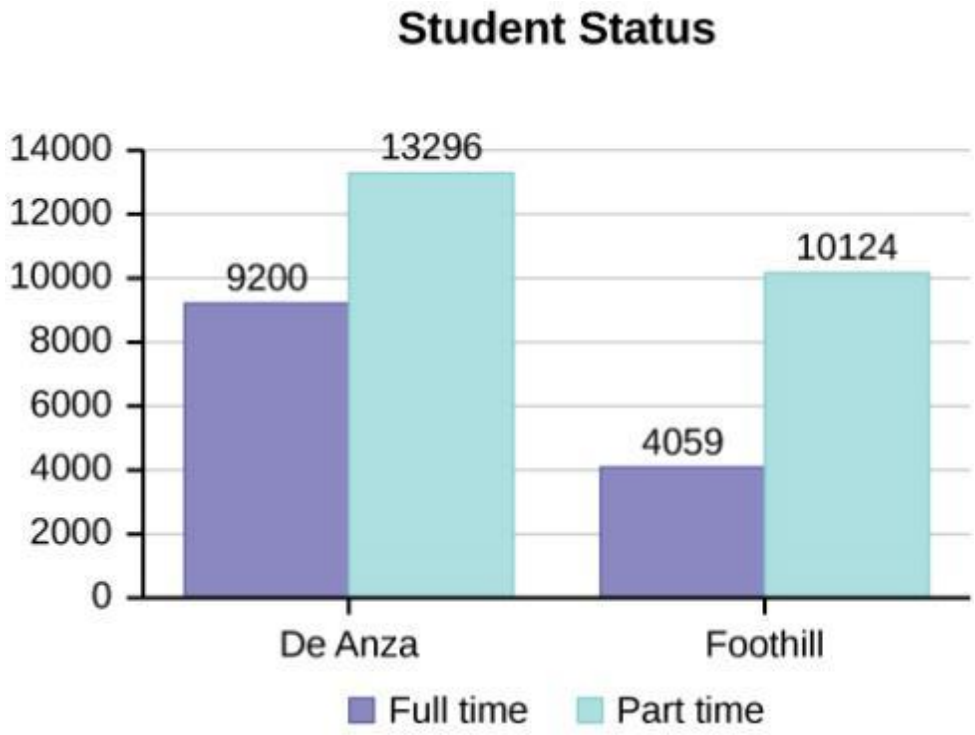
The following pie charts have the "Other/Unknown" category included (since the percentages must add to 100%). The chart in Figure 1.11b is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in Figure 1.11a.

BAR GRAPHS

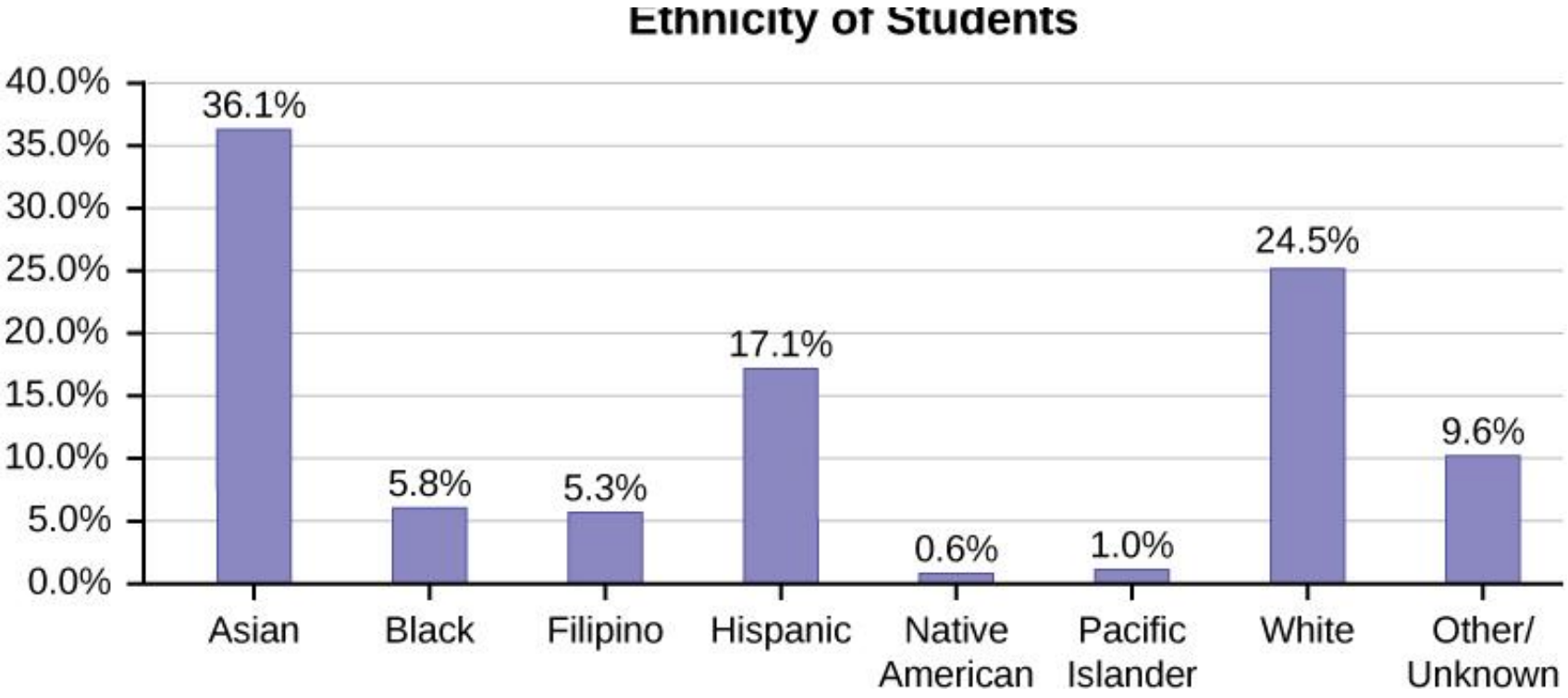


Sometimes percentages add up to be more than 100% (or less than 100%). In the graph, the percentages add to more than 100% because students can be in more than one category. A bar graph is appropriate to compare the relative size of the categories. A pie chart cannot be used. It also could not be used if the percentages added to less than 100%.

SIDE BY SIDE BAR GRAPH

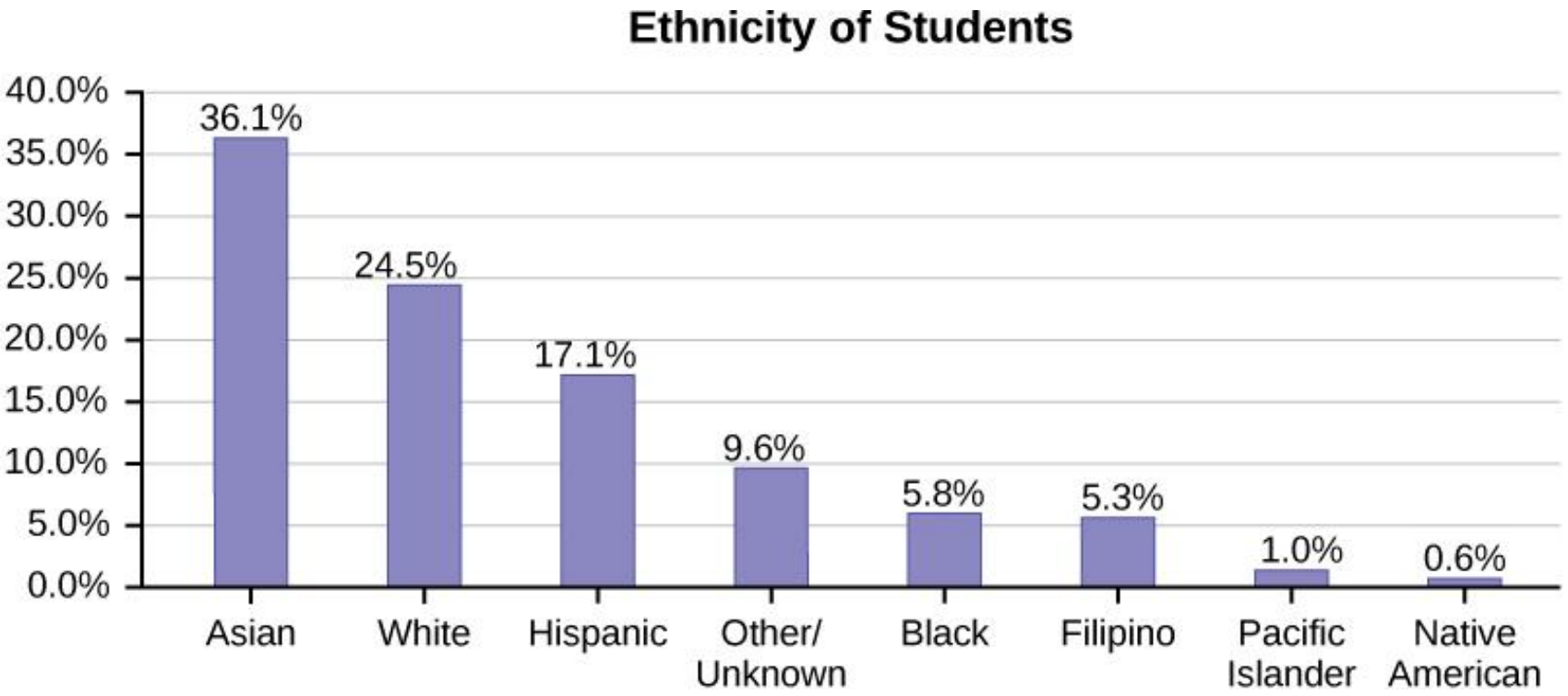


BAR GRAPH WITH UNKNOWN CATEGORY



Bar graph with Other/Unknown Category

PARETO CHART



Pareto Chart With Bars Sorted by Size

SELECTING A SAMPLE

Sample must be representative (Representative sample)

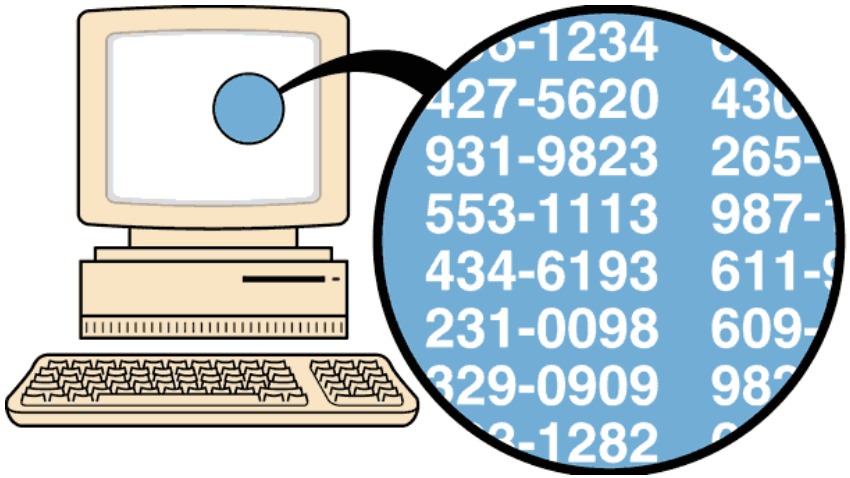
A representative sample:

- has the same relevant characteristics as the defined population and does not favor one group of the population over another.
- allows people to study a population without studying every single individual in that population.
- is a valuable research tool.

RANDOM SAMPLING

Random Sampling

- Selected by using chance or random numbers
- Each individual subject (human or otherwise) has an equal chance of being selected
- Examples:
  - Drawing names from a hat
  - Random Numbers
  - https://www.youtube.com/watch?v=aBJhaAF5GLs



SYSTEMATIC SAMPLING

Systematic Sampling

- Select a random starting point and then select every n^{th} subject in the population
- Simple to use so it is used often (choose a random starting point and then choose every n^{th} item

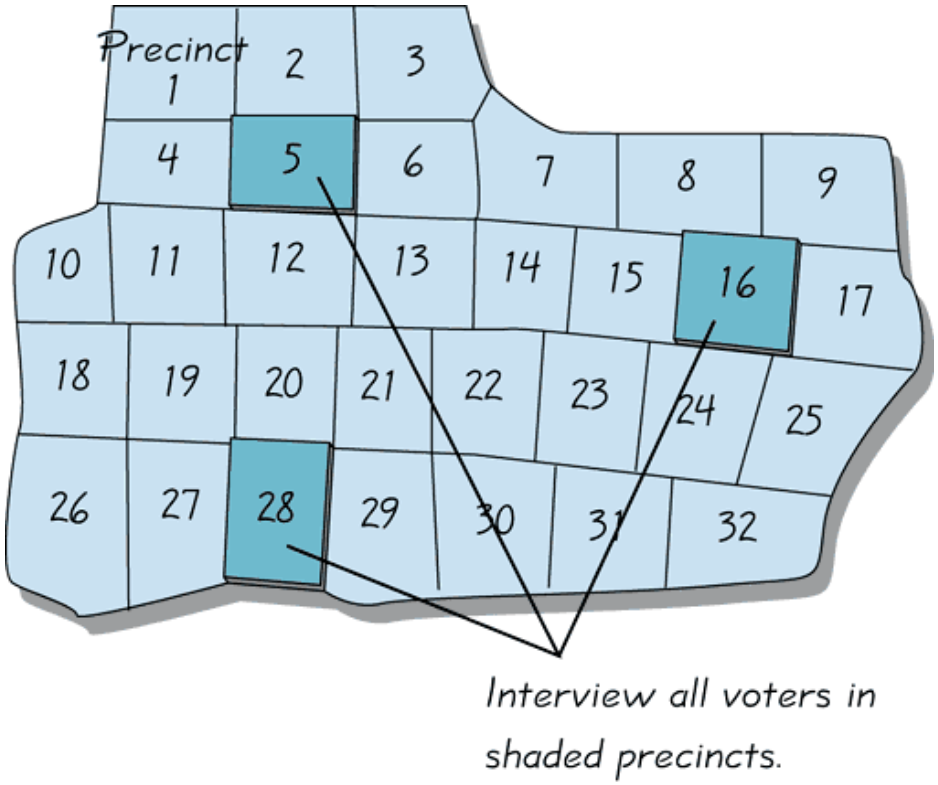
CONVENIENCE SAMPLING

- Convenience Sampling
  - Use subjects that are easily accessible
  - Prone to creating non-representative sample (member all have a similar characteristic)
  - Examples:
- Using family members or students in a classroom
- Mall shoppers

STRATIFIED SAMPLING

- Stratified Sampling
  - Divide the population into at least two different groups (**strata**) with common characteristic(s), then draw **SOME** subjects from each group
  - Results in a more representative sample
  - Helps preserve certain characteristics of the population

CLUSTER SAMPLING



- Cluster Sampling
  - Divide the population into groups (**clusters**), randomly select some of the groups, and then collect data from **ALL** members of the selected groups
  - Used extensively by government and private research organizations
  - Examples:
- Exit Polls

EXAMPLES

A study is done to determine the average tuition that San Jose State undergraduate students pay per semester. Each student in the following samples is asked how much tuition he or she paid for the Fall semester. What is the type of sampling in each case?

a. A sample of 100 undergraduate San Jose State students is taken by organizing the students' names by classification (freshman, sophomore, junior, or senior), and then selecting 25 students from each.

b. A random number generator is used to select a student from the alphabetical listing of all undergraduate students in the Fall semester. Starting with that student, every 50th student is chosen until 75 students are included in the sample.

c. A completely random method is used to select 75 students. Each undergraduate student in the fall semester has the same probability of being chosen at any stage of the sampling process.

d. The freshman, sophomore, junior, and senior years are numbered one, two, three, and four, respectively. A random number generator is used to pick two of those years. All students in those two years are in the sample.

e. An administrative assistant is asked to stand in front of the library one Wednesday and to ask the first 100 undergraduate students he encounters what they paid for tuition the Fall semester. Those 100 students are the sample.

ANSWERS

a. stratified;

b. systematic;

c. simple random;

d. cluster;

e. convenience

REPLACEMENT

True random sampling is done with **replacement**. That is, once a member is picked, that member goes back into the population and thus may be chosen more than once.

However for practical reasons, in most populations, simple random sampling is done **without replacement**. Surveys are typically done without replacement. That is, a member of the population may be chosen only once.

MORE EXAMPLES

Sampling without replacement instead of sampling with replacement becomes a mathematical issue only when the population is small.

For example, if the population is 25 people, the sample is ten, and you are sampling with replacement for any particular sample, then the chance of picking the first person is ten out of 25, and the chance of picking a different second person is nine out of 25 (you replace the first person).

If you sample without replacement, then the chance of picking the first person is ten out of 25, and then the chance of picking the second person (who is different) is nine out of 24 (you do not replace the first person).

Compare the fractions 9/25 and 9/24. To four decimal places, 9/25 = 0.3600 and 9/24 = 0.3750. To four decimal places, these numbers are not equivalent.

SAMPLING ERRORS

When you analyze data, it is important to be aware of **sampling errors** and **nonsampling errors**.

The actual process of sampling causes **sampling errors**.

For example, the sample may not be large enough.

Factors not related to the sampling process cause **nonsampling errors**. A defective counting device can cause a **nonsampling error**.

In reality, a sample will never be exactly representative of the population so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, a **sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen).

When a **sampling bias** happens, there can be incorrect conclusions drawn about the population that is being studied.

VARIATION

Variation is present in any set of data.

For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16 ounce cans were measured and produced the following amount (in ounces) of beverage: 15.8; 16.1; 15.2; 14.8; 15.8; 15.9; 16.0; 15.5

It was mentioned previously that two or more samples from the same population, taken randomly, and having close to the same characteristics of the population will likely be different from each other.

EXAMPLE OF VARIATION

Suppose Doreen and Jung both decide to study the average amount of time students at their college sleep each night.

Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the same sampling method, in all likelihood their samples would be different. Neither would be wrong, however.

Think about what contributes to making Doreen's and Jung's samples different. If Doreen and Jung took larger samples (i.e. the number of data values is increased), their sample results (the average amount of time a student sleeps) might be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other. This **variability** in samples cannot be stressed enough.

SAMPLE SIZE

The size of a sample (often called the number of observations) is important.

The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1,200 to 1,500 observations are considered large enough and good enough if the survey is random and is well done. You will learn why when you study confidence intervals.

Be aware that many large samples are biased. For example, call-in surveys are invariably biased, because people choose to respond or not.

SECTION 1.3
FREQUENCY, FREQUENCY TABLES, AND LEVEL OF MEASUREMENTS

ANSWERS AND ROUNDING OFF

A simple way to round off answers is to carry your final answer one more decimal place than was present in the original data.

Round off only the final answer.

Do not round off any intermediate results, if possible.

If it becomes necessary to round off intermediate results, carry them to at least twice as many decimal places as the final answer.

For example, the average of the three quiz scores four, six, and nine is 6.3, rounded off to the nearest tenth, because the data are whole numbers. Most answers will be rounded off in this manner.

LEVEL OF MEASUREMENT

Four levels of measurement

- Nominal
- Ordinal
- Interval
- Ratio

The higher the level of measurement, the more mathematical calculations that can be performed on that data.

MEASUREMENT SCALES

Nominal

- Classifies data into mutually exclusive (nonoverlapping) exhausting categories
- No order or ranking can be imposed
- Qualitative
- No calculations an be performed on Nominal data
- Examples:
  - Gender
  - Zip Codes
  - Political Affiliation
  - Religion

Ordinal

- Classifies data into categories
- Usually qualitative
- RANKING (natural order), but precise differences between ranks do not exist (addition or division do not make sense)
- Examples:
  - Letter grades (A, B, C, D, F)
  - Judging contest (1^{st}, 2^{nd} , 3^{rd})
  - Ratings (Above Avg, Avg, Below Avg, Poor)

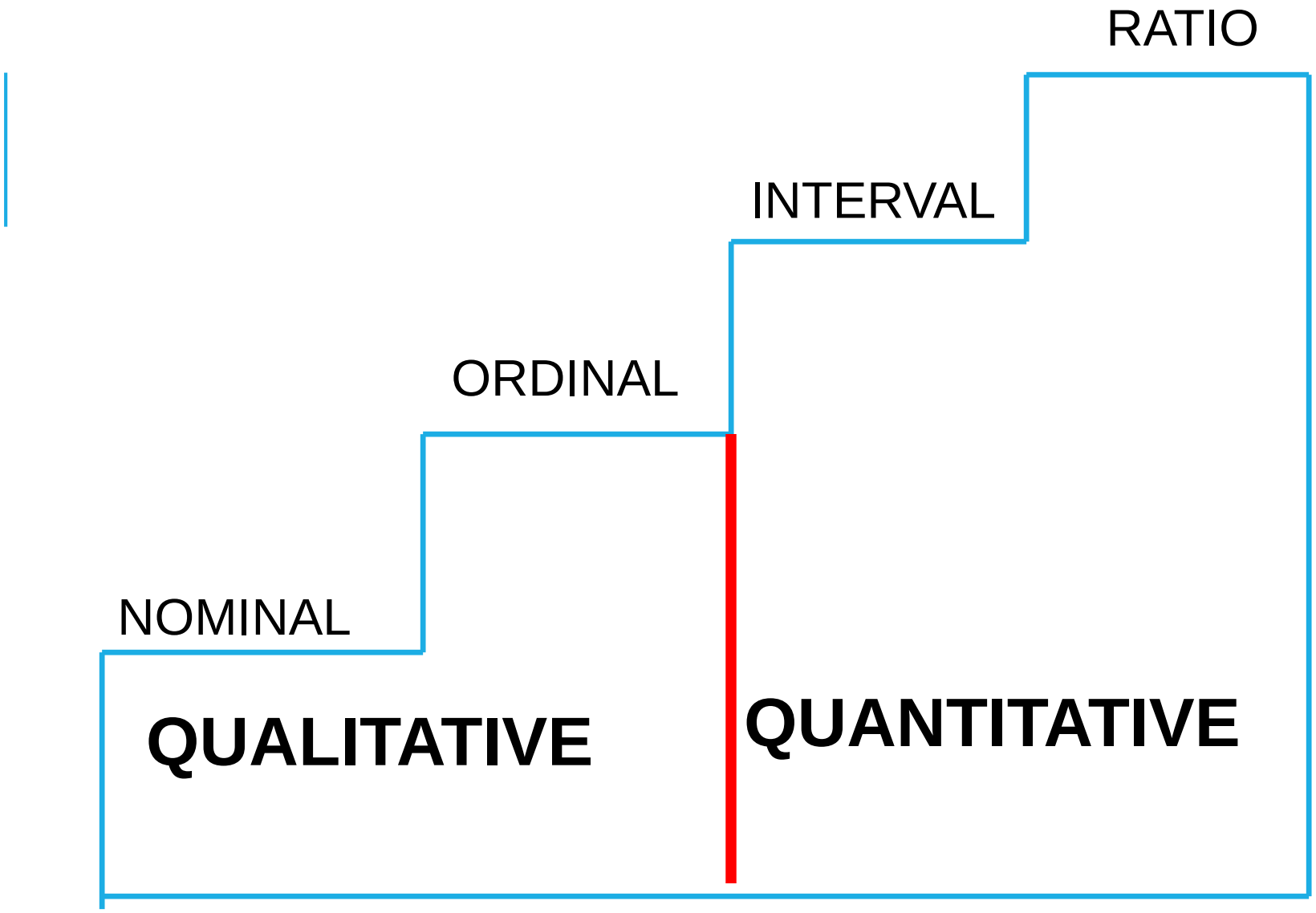
MEASUREMENT SCALES

Interval

- Quantitative data
- Ranks (orders) data
- PRECISE DIFFERENCES between units of measure do exist and are meaningful
- No meaningful zero (position on a scale, but does not mean absence of something) Zero is simply a placeholder
- Examples:
  - Temperature (0° does not mean no heat at all)
  - IQ Scores (0 does not imply no intelligence)
  - Calendar dates

Ratio

- Quantitative data
- Ranks (orders) data
- Precise differences exist and are meaningful
- TRUE ZERO exist (Zero means absence of something)
- Can add, subtract, multiply, and divide data values
- Examples:
  - Height
  - Weight
  - Area
  - Number of phone calls received
  - Salary


RATIO
INTERVAL
ORDINAL
NOMINAL
QUALITATIVE
QUANTITATIVE

EXAMPLES

Determine if the following qualitative or quantitative and then determine the level of measurement

9) The free throw shooting percentage of a basketball player
10) A survey response to "are you predominantly left-handed or right-handed?"
11) The temperature in Fahrenheit in Atlanta, Georgia
12) The individual page numbers at the bottom of each page in this book

EXAMPLES ANSWERS

Answers to examples:

1. Quantitative
2. qualitative
3. quantitative
4. qualitative
5. discrete
6. continuous
7. continuous
8. discrete
9. Quantitative, ratio
10. qualitative, nominal
11. quantitative, interval
12. qualitative, ordinal

WHAT IS A FREQUENCY DISTRIBUTION?

A frequency distribution is the organization of raw data in table form, using classes (groups) and frequencies

Class (group) is a quantitative or qualitative category

Frequency- the number of times a value appears in a certain data set.

Relative frequency – the number of times a value appears in a certain data set divided by the total number of values in the data set.

Cumulative relative frequency – Add all the previous relative frequencies to the relative frequency for the current row.

EXAMPLE

Below is the data set of 20 students' hours for studying.

2,3,2,8,2,3,4,5,1,7,0,1,2,3,2,2,3,4,5,1

To make a frequency table, you need to make groups in which to divide the data by.

I am going to start with 0 to 1 hours. These groups will be given to you so you will not have to make them.

EXAMPLE

Hours of Studying	Frequency	Relative Frequency	Cumulative Relative Frequency
0-1			
2-3			
4-5			
6-7			
8-9			

Next, we will need to add up the frequency for each group using the data: 2,3,2,8,2,3,4,5,1,7,0,1,2,3,2,2,3,4,5,1

For the first group, there is one 0 and three 1's so the frequency is 4.

For the second group, there are six 2's and four 3's so the frequency is 10.

Hours of Studying	Frequency	Relative Frequency	Cumulative Relative Frequency
0-1	4		
2-3	10		
4-5	4		
6-7	1		
8-9	1		

To get the relative frequency, I know the data set has 20 values so I will put the frequency over the total number in the data set.

Hours of Studying	Frequency	Relative Frequency	Cumulative Relative Frequency
0-1	4	4/20 = 0.20	
2-3	10	10/20 = 0.50	
4-5	4	4/20 = 0.20	
6-7	1	1/20 = 0.05	
8-9	1	1/20 = 0.05	

To get the cumulative relative frequency, I will add up the relative frequencies as I go down the table.

Hours	Frequency	Relative Frequency	Cumulative Relative Frequency
0-1	4	4/20 = 0.20	0.20
2-3	10	10/20 = 0.50	0.20 + 0.50 = 0.70
4-5	4	4/20 = 0.20	0.20+0.50+0.20 = 0.90
6-7	1	1/20 = 0.05	0.20+0.50+0.20+0.05 = 0.95
8-9	1	1/20 = 0.05	0.20+0.50+0.20+0.05+0.05 = 1.00

SECTION 1.4
EXPERIMENTAL DESIGN AND ETHICS

EXPERIMENTS AND VARIABLES

The purpose of an experiment is to investigate the relationship between two variables.

When one variable causes change in another, we call the first variable the **explanatory variable**.

The affected variable is called the **response variable**.

In a randomized experiment, the researcher manipulates values of the explanatory variable and measures the resulting changes in the response variable.

The different values of the explanatory variable are called **treatments**.

An experimental unit is a single object or individual to be measured.

EXPERIMENTS AND VARIABLES

Additional variables that can cloud a study are called **lurking variables**.

In order to prove that the explanatory variable is causing a change in the response variable, it is necessary to isolate the explanatory variable.

The researcher must design her experiment in such a way that there is only one difference between groups being compared: the planned treatments.

This is accomplished by the **random assignment** of experimental units to treatment groups. When subjects are assigned treatments randomly, all of the potential lurking variables are spread equally among the groups.

At this point the only difference between groups is the one imposed by the researcher. Different outcomes measured in the response variable, therefore, must be a direct result of the different treatments. In this way, an experiment can prove a cause-and-effect connection between the explanatory and response variables.

TREATMENTS

When participation in a study prompts a physical response from a participant, it is difficult to isolate the effects of the explanatory variable.

To counter the power of suggestion, researchers set aside one treatment group as a **control group**. This group is given a **placebo treatment**–a treatment that cannot influence the response variable. The control group helps researchers balance the effects of being in an experiment with the effects of the active treatments.

Of course, if you are participating in a study and you know that you are receiving a pill which contains no actual medication, then the power of suggestion is no longer a factor.

Blinding in a randomized experiment preserves the power of suggestion. When a person involved in a research study is blinded, he does not know who is receiving the active treatment(s) and who is receiving the placebo treatment. A **double-blind experiment** is one in which both the subjects and the researchers involved with the subjects are blinded.

ETHICS

The widespread misuse and misrepresentation of statistical information often gives the field a bad name. Some say that "numbers don't lie," but the people who use numbers to support their claims often do.

A recent investigation of famous social psychologist, Diederik Stapel, has led to the retraction of his articles from some of the world's top journals including Journal of Experimental Social Psychology, Social Psychology, Basic and Applied Social Psychology, British Journal of Social Psychology, and the magazine Science.

Diederik Stapel is a former professor at Tilburg University in the Netherlands. Over the past two years, an extensive investigation involving three universities where Stapel has worked concluded that the psychologist is guilty of fraud on a colossal scale. Falsified data taints over 55 papers he authored and 10 Ph.D. dissertations that he supervised.

ETHICS

Stapel did not deny that his deceit was driven by ambition. But it was more complicated than that, he told me. He insisted that he loved social psychology but had been frustrated by the messiness of experimental data, which rarely led to clear conclusions.

His lifelong obsession with elegance and order, he said, led him to concoct sexy results that journals found attractive. “It was a quest for aesthetics, for beauty—instead of the truth,” he said.

He described his behavior as an addiction that drove him to carry out acts of increasingly daring fraud, like a junkie seeking a bigger and better high.

ETHICS

The committee investigating Stapel concluded that he is guilty of several practices including:

- creating datasets, which largely confirmed the prior expectations,
- altering data in existing datasets,
- changing measuring instruments without reporting the change,
- misrepresenting the number of experimental subjects. Clearly, it is never acceptable to falsify data the way this researcher did. Sometimes, however, violations of ethics are not as easy to spot.

IRB

When a statistical study uses human participants, as in medical studies, both ethics and the law dictate that researchers should be mindful of the safety of their research subjects.

The U.S. Department of Health and Human Services oversees federal regulations of research studies with the aim of protecting participants. When a university or other research institution engages in research, it must ensure the safety of all human subjects.

For this reason, research institutions establish oversight committees known as **Institutional Review Boards (IRB)**.

INFORMED CONSENT

All planned studies must be approved in advance by the IRB. Key protections that are mandated by law include the following:

- Risks to participants must be minimized and reasonable with respect to projected benefits.

- Participants must give **informed consent**. This means that the risks of participation must be clearly explained to the subjects of the study. Subjects must consent in writing, and researchers are required to keep documentation of their consent.

- Data collected from individuals must be guarded carefully to protect their privacy