Chapters 1-4

Stat 128: Elementary Statistics

A Document By:
Nathan Warner

July 03, 2023
Computer Science
Joliet Junior College
United States

Contents

1 Learning Outcomes

Chapter 1:

1. Define data collection techniques including observational studies and design of experiments.
2. Identify appropriate sampling methods.

Chapter 2:

1. Differentiate qualitative and quantitative data graphically.
   This includes graphs such as bar plots, histograms, and dot plots.

Chapter 3:

1. Calculate measures of central tendency for data.
2. Explain the concept of resistance.
3. Decide which measure of central tendency to report for various data sets.
4. Determine measures of dispersion for data.
5. Determine standard scores, percentiles, and quartiles.
6. Identify outliers using quartiles.
7. Interpret boxplots.

Chapter 4:

1. Evaluate the linear correlation coefficient for bivariate quantitative data.
2. Evaluate whether the coefficient is significant at a given level.
3. Explain the difference between correlation and causation.
4. Determine the least-squares regression equation for a given set of bivariate data.
5. Predict values of the dependent variable using the least-squares regression equation.
6. Interpret the slope and intercept of the least-squares regression equation.
7. Test the requirements of the least-squares regression model using residual analysis.
8. Determine and interpret the coefficient of determination.
9. Graphically analyze bivariate quantitative data for outliers and influential observations.
10. Describe the association between two qualitative variables using conditional distributions.
11. Explain Simpson's Paradox.

2 Chapter 1:

2.1 1.1: Introduction to the Practice of Statistics

Objectives for this section.

1. Define Statistics and Statistical Thinking
2. Explain the Process of Statistics
3. Distinguish between Qualitative and Quantitative Variables
4. Distinguish Between Discrete and Continuous Variables.

Define Statistics and Statistical Thinking:

Statistics is the science of collecting, organizing, summarizing, and analyzing information to draw conclusions or answer questions. In addition, statistics is about providing a measure of confidence in any conclusions.

Note: We must report a measure of our confidence in our results because we do not have 100% certainty our answers are correct.

The information referred to in the definition above is *data*. **Data** are a "fact or proposition used to draw a conclusion or make a decision." Data describes characteristics of an individual.

One crucial thing to understand about **data**, is that is **varies**. One thing that makes an interesting study is the fact that the data within the study varies. A study about number of hearts a human has is not only uninteresting but not worth doing. This is because the data does not vary.

Two Major Goals:

1. Describe Variability.
2. Understand sources of Variability.

In Statistics, the same approach to solving a problem can still lead to different results. This does not happen in a math class.

Explain the Process of Statistics.

First lets define some vocabulary:

- **Population:** The entire group to be studied is called the population.
- **Sample:** In statistics, it is often impractical or impossible to get access to the entire **population**, which is why we only look at a **sample.** A sample is a **subset** of the population being studied.
- **Individual:** An individual is a person or object that is a member of the population being studied.
- **Statistic:** A statistic is a numerical summary of a sample.
- **Descriptive Statistics:** Descriptive statistics consist of organizing and summarizing data. Descriptive statistics describe data through numerical summaries, tables, and graphs.
- **Inferential Statistics:** inferential Statistics uses methods that take a result from a sample, extend it to the population, and measure the reliability of the result.
- **Parameter:** A parameter is a numerical summary of a population.

The process of statistics:

1. Identify the problem to be solved. It is important to clearly lay out the questions that the researcher wants answered, along with clearly specifying which population the study applies.
2. Collect the data.
3. Describe the data.
4. Preform inference.

Example: The AP - National Constitution Center conducted a national poll to learn how adult Americans feel existing gun-control laws infringe on the second amendment to the U.S Constitution

The Following statistical process allowed the researchers to conduct their study.

1. **Identify the research objective.**: The researchers wished to determine the percentage of adult Americans who believe gun-control laws infringe on the public's right to bear arms.
2. **Collect the information needed to answer the question posed in (1).**: It is unreasonable to expect to survey the more than 200 million adult Americans to determine how they feel about gun-control laws. So the researchers surveyed a sample of 1007 adult Americans. Of those surveyed, 514 stated they believe existing gun-control laws infringe on the public's right to bear arms.
3. **Describe the data.**: Of the 1007 individuals in the survey, 51% believe existing gun-control laws inferring on the public's right to bear arms. This is a descriptive statistic, because its value is determined from a sample.
4. **Preform inference.**: The researchers at the AP - National Constitution Center wanted to extend the results of the survey to **all** adult Americans. When generalizing results from a sample to a population, the results are **uncertain**. To account for this uncertainty, researchers reported a 3% *margin of error.* This means that the researchers feel fairly certain (in fact, 95% certain) that the percentage of *all* adult Americans who believe existing gun-control laws infringe on the public's right to bear arms is somewhere between 48% and 54%

Distinguish between Qualitative and Quantitative Variables

First let's define some vocab:

- **Variables:** The characteristics of the individuals in a study. Variables vary, which means they can take on different values.
- **Constants:** Variables that do not vary. Inferential statistics is not necessary with constants.

One goal of research is to learn the causes of variability.

Variables can be classified into two groups: qualitative and quantitative.

- **Qualitative, or categorical variables** allow for the classification of individuals base on some attribute or characteristic.
- **Quantitative variables** provide numerical measures of individuals. The values of a quantitative variable can be added or subtracted and provide meaningful results.

Example: Determine whether the following variables are qualitative or quantitative.

a.) **Gender.**: Qualitative

b.) **Temperature.**: Quantitative

c.) **Number of days during the past week that a college student studied.**: Quantitative

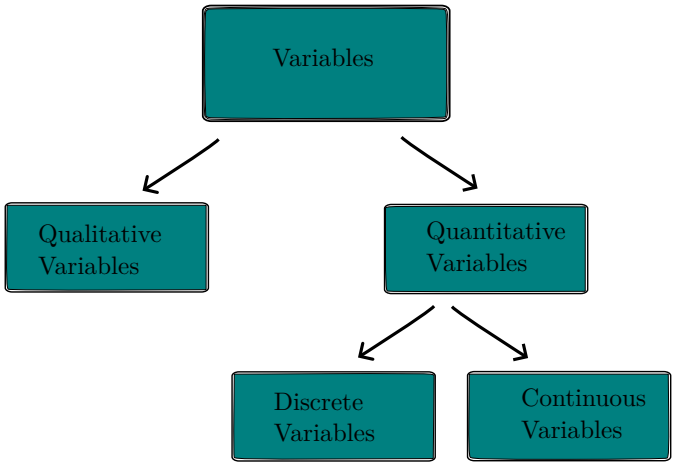
d.) **Zip Code.** Qualitative

Caution: A numeric value does not automatically suggest a variable is quantitative.

Distinguish between Discrete and Continuous Variables.

- A **discrete variable** is a quantitative variable that has either a finite number of possible values or a countable number of possible values. A discrete variable cannot take on every possible value between any two possible values.
- A **continuous variable** is a quantitative variable that has an infinite number of possible values that are not countable. A continuous variable may take on every possible value between any two values. Continuous variables typically result from measurement. Continuous variables are often rounded. If a certain make of car gets 24 miles per gallon (mpg) of gasoline, its miles per gallon must be greater than or equal to 23.5 and less than 24.5, or $23.5 \leqslant mpg \leqslant 24.5$

This Figure illustrates the relationship among qualitative, quantitative, discrete, and continuous variables.



Example: Distinguish whether the quantitative variables are discrete or continuous.

a.) **The number of heads obtained after flipping a coin five times.**: Discrete

b.) **The number of cars that arrive at a McDonald's drive through between 12:00 PM and 1:00 PM**: Discrete

c.) **The Distance a 2011 Toyota Prius can travel in city driving conditions with a full tank of gas.**: Continuous

Vocab:

- The list of observed values for a variable is **data.**
- **Qualitative data** are observations corresponding to a **qualitative variable.**
- **Quantitative data** are observations corresponding to a quantitative variable.
- **Discrete data** are observations corresponding to a discrete variable.
- **Continuous data** are observations corresponding to a continuous variable.

Example: Distinguish between Variables and Data

The following table presents a group of selected countries and information regarding these countries.

Identify the individuals, variables, and data.

Country	Government Type	Life Expectancy (Years)	Population (in millions)
Australia	Federal parliamentary democracy	81.63	21.3
Canada	Constitutional monarchy	81.23	33.5
France	Republic	80.98	64.4
Morocco	Constitutional monarchy	75.47	31.3
Poland	Republic	75.63	38.5
Sri Lanka	Republic	75.14	21.3
United States	Federal Republic	78.11	307.2

Qualitative: Government Type
Quantitative: Life Expectancy and Population
Continuous: Life Expectancy
Discrete: Population
Data: Everything under Government Type, Life Expectancy, and Population.

All Vocab / Concepts From Section 1.1

- **Population:** The entire group to be studied is called the population.
- **Sample:** In statistics, it is often impractical or impossible to get access to the entire **population**, which is why we only look at a **sample.** A sample is a **subset** of the population being studied.
- **Individual:** An individual is a person or object that is a member of the population being studied.
- **Statistic:** A statistic is a numerical summary of a sample.
- **Descriptive Statistics:** Descriptive statistics consist of organizing and summarizing data. Descriptive statistics describe data through numerical summaries, tables, and graphs.
- **Inferential Statistics:** inferential Statistics uses methods that take a result from a sample, extend it to the population, and measure the reliability of the result.
- **Parameter:** A parameter is a numerical summary of a population.
- **Variables:** The characteristics of the individuals in a study. Variables vary, which means they can take on different values.
- **Constants:** Variables that do not vary. Inferential statistics is not necessary with constants.
- **Data:** The list of observed values for a variable.
- **Qualitative data** are observations corresponding to a **qualitative variable.**
- **Quantitative data** are observations corresponding to a quantitative variable.
- **Discrete data** are observations corresponding to a discrete variable.
- **Continuous data** are observations corresponding to a continuous variable.

Concepts:

- Statistics and Statistical Thinking.
- Describe Variability
- Understand Sources of variability
- Statistical studies are concerned with both describing the variability in the data and understanding the sources of variability in data. Understanding the sources allows researchers to control it and reach better conclusions.
- The process of statistics
- Inferential/Descriptive Statistics
- Variables
  - Qualitative (Categorical) / Quantitative
  - Discrete / Continuous
- Data
  - Qualitative (Categorical) / Quantitative
  - Discrete / Continuous

2.2 1.2: Observational Studies versus Designed Experiments

Objectives for this section.

1. Distinguish between an Observational Study and a Designed Experiment
2. Explain the Various Types of Observational Studies

Distinguish between an Observational Study and a Designed Experiment

- **Observational studies:** Observational studies involve observing and analyzing data collected from real-world settings without any intervention or manipulation by the researcher. Researchers passively observe and record information to identify correlations or associations between variables.
- **Designed experiments:** Designed experiments, also known as randomized controlled trials (RCTs), involve researchers actively manipulating variables and randomly assigning participants to different groups. This allows researchers to establish cause-and-effect relationships by comparing the effects of different interventions or treatments on the outcome of interest.

Vocab:

- **Explanatory Variable:** An explanatory variable, also known as an independent variable or predictor variable, is a variable that is manipulated or controlled by researchers in an experiment or study. It is the variable that is hypothesized to have an impact on the outcome or dependent variable.
- **Lurking variable**: An explanatory variable that was not considered in a study, but that affects the value of the response variable.
- **Response Variable**: The response variable, also known as the dependent variable or outcome variable, is the variable that is measured or observed to determine the effect or response of the explanatory variable(s). It is the variable that researchers are interested in studying or predicting.
- **Confounding:** Occurs when the effects of two or more explanatory variables are not separated. Therefore, any relation that may exist between an explanatory variable and the response variable may be due to some other variable or variables not accounted for in the study.
- **Census:** List of individuals in a population along with certain characteristics of each individual.

Note:-

Is observational studies, we **are not** allowed to make statements of *causality*, meaning we cannot say that changes in the explanatory variable *cause* changes in the response variable. We can only say changes in the explanatory variable are associated with changes in the response variable.

Why would we ever conduct an observational study if we cannot claim causation? Because it is often unethical to conduct a designed experiment.

Consider the link between smoking and lung cancer. In a designed experiment (on humans) to determine if smoking causes lung cancer, a researcher would divide a group of volunteers into two groups—Group 1 would smoke a pack of cigarettes every day for the next 10 years, and Group 2 would not smoke. Eating habits, sleeping habits, and exercise would be controlled so that the only difference between the two groups would be smoking. After 10 years, the experiment's researcher would compare the proportion of participants in the study who contract lung cancer in the smoking group with the nonsmoking group. If the two proportions differ significantly, it could be said that smoking causes lung cancer. This designed experiment controls many potential cancer-causing factors that would not be controlled in an observational study. However, it is an unethical experiment. Do you see why?

Other reasons exist for conducting observational studies over designed experiments. An article in support of observational studies states, "Observational studies have several advantages over designed experiments, including lower cost, greater timeliness, and a broader range of patients." From Kjell Benson, BA, and Arthur J. Hartz, MD, PhD. "A Comparison of Observational Studies and Randomized Controlled Trials."

In designed experiments, it is possible to have two explanatory variables in a study that are related to each other and related to the response variable. For example, suppose Professor Egner wanted to conduct an experiment in which she compared student success using online homework versus traditional textbook homework. To do the study, she taught her morning statistics class using the online homework and her afternoon class using traditional textbook homework. At the end of the semester, she compared the final exam scores for the online section to the textbook section. If the morning section had higher scores, could Professor Egner conclude that online homework is the cause of higher exam scores? Not necessarily. It is possible that the morning class had students who were more motivated. It is impossible to know whether the outcome was due to the online homework or to the time at which the class was taught. In this sense, we say that the time of day the class is taught is a confounding variable.

Lurking Vs Confounding Variables:

The big difference between lurking variables and confounding variables is that lurking variables are not considered in the study (for example, we did not consider lifestyle in the pneumonia study) whereas confounding variables are measured in the study (for example, we measured morning versus afternoon classes).

So lurking variables are related to both the explanatory and response variables, and this relation is what creates the apparent association between the explanatory variable and response variable in the study. For example, lifestyle (healthy or not) is associated with the likelihood of getting an influenza shot as well as the likelihood of contracting pneumonia or influenza.

A confounding variable is a variable in a study that does not necessarily have any association with the other explanatory variable but does have an effect on the response variable. Perhaps morning students are more motivated, and this is what led to the higher final exam scores, not the homework delivery system.

The bottom line is that both lurking variables and confounding variables can confound the results of a study, so a researcher should be mindful of their potential existence.

Explain the Various Types of Observational Studies

- **Cross-sectional Studies:** Observational studies that collect information about individuals at a specific point in time, or over a very short period of time.
- **case-control Studies:** These studies are **retrospective,** meaning that they require individuals to look back in time or require the researcher to look at existing records. In case-control studies, individuals that have certain characteristics are matched with those that do not.
  - **Positive:** Control group allows for a comparison
  - **Negative:** Individuals must remember details
  - **Negative:** Records might not exist
- **Cohort Studies:** A cohort study first identifies a group of individuals to participate in the study (cohort). The cohort is then observed over a period of time. Over this time period, characteristics about the individuals are recorded. Because the data is collected over time, cohort studies are **prospective.**
  - **Advantage:** Researcher does not need to rely on memory of study participants or existing records.
  - **Disadvantage:** Requires a lot of time.
  - **Disadvantage:** Could be expensive.

Is a designed experiment superior to an observational study? Not necessarily.

- Because cross-sectional and case-control observational studies are relatively inexpensive, they allow researchers to explore possible associations prior to undertaking large cohort studies or designed experiments.
- It is not always possible to conduct an experiment. For example, we could not conduct an experiment to investigate the perceived link between high-tension wires and leukemia (on humans). Do you see why?

2.3 1.3: Simple Random Sampling

Learning Objectives for this section.

1. Obtain a simple random sample

Vocab:

- **Random Sampling:** The process of using chance to select individuals from a population to be included in the sample.
- **Simple Random Sampling:** A sample of size n from a population of size N is obtained through simple random sampling if every possible sample of size n has an equal chance of occurring. The sample is then called a simple random sample.
  - $n < N$
- **frame:** a list of all the individuals within the population.

Note:-

For the results of a survey to be reliable, the characteristics of the individuals in the sample must be representative of the characteristics of the individuals in the population. The key to obtaining a sample representative of a population is to let chance or randomness, rather than convenience, play a role in dictating which individuals are in the sample. If convenience is used to obtain a sample, the results of the survey are meaningless.

Recognizing a Convenience Sample and Its Limitations:

Suppose that Gallup wants to know the proportion of adult Americans who consider themselves to be baseball fans. If Gallup obtained a sample by standing outside Fenway Park (home of the Boston Red Sox professional baseball team), the survey results are not likely to be reliable. Why? Clearly, the individuals in the sample do not accurately reflect the makeup of the entire population.

Suppose you wanted to learn the proportion of students on your campus who work. It might be convenient to survey the students in your statistics class, but do these students represent the overall student body? Does the proportion of freshmen, sophomores, juniors, and seniors in your class mirror the proportion of freshmen, sophomores, juniors, and seniors on campus? Does the proportion of males and females in your class resemble the proportion of males and females across campus? Probably not. What about evening (or day) students? For these (and many other) reasons, the convenient sample is not representative of the population, which means that any results reported from your survey are misleading.

Effective Sampling Techniques:

1. **Simple random sampling**
2. **Stratified sampling**
3. **Systematic sampling**
4. **Cluster sampling**

These sampling methods are designed so that any selection biases the surveyor introduced (knowingly or unknowingly) during the selection process are eliminated. In other words, the surveyor does not have a choice as to which individuals are in the study. We will discuss simple random sampling in this section and the remaining three types of sampling in the next section.

Bonus: Consider a set of 5 possibilities A-E, and we want to determine the total number of combinations of selecting 3 letters:

We can use the formula:

$$nCk = \frac{n!}{(k!(n-k)!)}.$$

Where n is the total number of classes in the course list and k is the number of classes to be chosen.

So we have $n = 5$ and $k = 3$:

$$\begin{aligned}
&\frac{5!}{(3!(5-3)!)} \\
&= \frac{5!}{3! \cdot 2!} \\
&= \frac{5 \cdot 4 \cdot 3!}{3! \cdot 2 \cdot 1} \\
&= \frac{5 \cdot 4}{2 \cdot 1} \\
&= \frac{20}{2} \\
&= 10.
\end{aligned}$$

And we can calculate the chance over a certain event happening with:

$$Probablility = \frac{NumberofOccurences}{TotalNumberofOccurences}.$$

How do we select the individuals in a simple random sample?

We could write the names of the individuals in the population on different pieces of paper and then select names from a hat. Often, however, the size of the population is so large that performing simple random sampling in this fashion is not practical.

Typically, each individual in the population is assigned a unique number between 1 and N, Where N is the size of the population. Then n distinct random numbers are selected, where n is the size of the population

To number the individuals in the population, we need a **frame:** a list of all the individuals within the population.

Obtaining a simple random sample with calculator (ti-84)

The accounting firm of Senese and Associates has grown. To make sure their clients are still satisfied with the services they are receiving, the company decides to send a survey out to a simple random sample of 5 of its 30 clients.

So we need 5 unique random numbers from a range of 1-30. To do this in our ti-84 calculator:

1. Math → Prod → randIntNorep
2. Syntax: randIntNoRep(*lowerbound*, *upperbound*, n), where n is the number of unique random numbers we must generate.
3. Lower: 1
4. Upper: 30
5. n: 5
6. Select Paste.

2.4 1.4: Other Effective Sampling Methods

Learning Objectives For This Section.

1. **Obtain a Stratified Sample**
2. **Obtain a Systematic Sample**
3. **Obtain a Cluster Sample**

Vocab:

- **Stratified sample**: is obtained by dividing the population into nonoverlapping groups called strata and then obtaining a simple random sample from each stratum. The individuals within each stratum should be homogenous (similar) in some way.
  - Within Stratified samples, the number of individuals sampled from each stratum should be proportional to the size of the strata in the population.
- **Systematic sample** is obtained by selecting every kth individual from the population. The first individual selected corresponds to a number between 1 and k
- **Cluster sample** is obtained by selecting all individuals within a randomly selected collection or group of individuals.
- **Convenience sample:** the individuals are easily obtained and not based on randomness.

Obtaining a Stratified Sample:

Suppose Congress was considering a bill that abolishes estate taxes. In an effort to determine the opinion of her constituency, a senator asks a pollster to conduct a survey within her state.

The pollster may divide the population of registered voters within the state into three strata: Republican, Democrat, and Independent. This grouping makes sense because the members within each of the three parties may have similar opinions regarding estate taxes, but opinions among parties may differ. The main criterion in performing a stratified sample is that each group (stratum) must have a common attribute that results in the individuals being similar within the stratum.

An advantage of stratified sampling over simple random sampling is that it may allow fewer individuals to be surveyed while it obtains the same or more information. This result occurs because individuals within each subgroup have similar characteristics, so opinions within the group are not as likely to vary much from one individual to the next. In addition, a stratified sample guarantees that each stratum is represented in the sample.

Obtaining a Systematic Sample:

For example, to learn about the outcome of an election, a pollster might survey every tenth individual that leaves a polling place.

Because systematic sampling does not require a frame, it is a useful technique when you cannot gather a list of the individuals in the population. Also, systematic samples typically provide more information for a given cost than does simple random sampling. In addition, systematic sampling is easier to employ; so there is less likelihood of interviewer error occurring, such as selecting the wrong individual to be surveyed.

Choosing a value for k:

If the size of the population is unknown, there is no mathematical way to determine k

The value of k must be small enough to achieve our desired sample size.

The value of k must be large enough to obtain a sample that is representative of the population.

Example: Suppose we have a scenario where $k = 30$, and we want to start with individual 3 and select 40 individuals.

Firstly, we can observe that the sequence follows an arithmetic progression.

We can represent the sequence as:

$$a_{40} = \{3, 33, 63, \ldots, a_{40}\}$$

To find the 40th term, a_{40}, we can use the formula:

$$a_n = a + (n-1)d, \quad \text{where } n = 40,\ d = 30$$

Substituting the values into the formula:

$$a_{40} = 3 + (40 - 1) \cdot 30 = 1173$$

Therefore, our sequence is:

$$a_{40} = \{3, 33, 63, \ldots, 1173\}$$

Note: It's important to note that if our population does not have 1173 individuals, the desired sample size will not be achieved.

Example: Now Suppose $k = 4$

$$\begin{aligned} a_{40} &= 3 + (40 - 1) \cdot 4 \\ &= 159. \end{aligned}$$

So we have the sequence:

$$a_{40} = \{3, 7, 11, ..., 159\}.$$

Note: Suppose our population is Kroger shoppers, The 159^{th} shopper might leave the store at 3pm, so our survey would not include any of the evening shoppers.

Note:-

An estimate of the size of the population would help to determine an appropriate value of k

Systematic Sampling Determining k when N is Known:

Steps to deduce k:

1. If possible, approximate the population size N.
2. Determine the sample size desired, n.
3. Divide N by n $\left(\frac{N}{n},\right)$ and round down to the nearest integer. This value is k.
4. Randomly select a number between 1 *and* k, call this number a (starting point)
5. The sample will consist of the following individuals:

$$a_n = a, a+k, a+2k, ..., a+(n-1)k.$$

Example: Suppose $N = 20,325$ and we desire a sample size of $n = 100$.

$$K = \frac{20,325}{100}$$
$$K = 203.$$

Now let's further suppose that we start with the 90^{th} individual.

To compute our arithmetic sequence with these parameters, first lets find a_{100}:

$$a_{100} = 90 + (100-1) \cdot 203$$
$$a_{100} = 19890.$$

So we have the sequence:

$$a_{100} = \{90, 293, 496, ..., 20187\}.$$

Obtain a Cluster Sample.

Suppose a school administrator wants to learn the characteristics of students enrolled in online classes. Rather than obtaining a simple random sample based on the frame of all students enrolled in online classes, the administrator treats each online class as a cluster and then finds a simple random sample of these clusters. The administrator then surveys all students in the selected clusters. This reduces the number of classes that get surveyed.

The following questions arise in cluster sampling:

- How do I cluster the population?
- How many clusters do I sample?
- How many individuals should be in each cluster?

First, we must determine whether the individuals within the proposed cluster are homogeneous or heterogeneous.

city blocks tend to have similar households. Survey responses from houses on the same city block are likely to be similar. This results in duplicate information. We conclude that if the clusters have homogeneous individuals, it is better to have more clusters with fewer individuals in each cluster.

What if the cluster is heterogeneous? Under this circumstance, the heterogeneity of the cluster likely resembles the heterogeneity of the population. In other words, each cluster is a scaled-down representation of the overall population.

For example, a quality control manager might use shipping boxes that contain 100 light bulbs as a cluster. The manager does this because the rate of defects within the cluster resembles the rate of defects in the population, assuming that the bulbs are randomly placed in the box. Thus, when each cluster is heterogeneous, fewer clusters with more individuals in each cluster are appropriate.

Convenience Sampling:

In the four sampling techniques just presented (simple random sampling, stratified sampling, systematic sampling, and cluster sampling), the individuals are selected randomly. Often, however, inappropriate sampling methods are used in which the individuals are not randomly selected.

Have you ever been stopped in the mall by someone holding a clipboard? These folks are responsible for gathering information, but their methods of data collection are inappropriate, and the results of their analysis are suspect because they collect data using a convenience sample.

The most popular convenience samples are those in which the individuals in the sample are self-selected, meaning the individuals themselves decide to participate in the survey. Self-selected surveys are also called voluntary response samples. One example of self-selected sampling is phone-in polling—a radio personality will ask his or her listeners to phone or text the station to submit their opinions. Another example is the use of the Internet to conduct surveys. For example, a TV news show will present a story regarding a certain topic and ask its viewers to ”tell us what you think” by completing an online questionnaire or tweeting an opinion with a hashtag.

Both of these samples are poor designs because the individuals who decide to be in the sample generally have strong opinions about the topic. A more typical individual in the population will not bother phoning, texting, or tweeting to complete a survey. Any inference made regarding the population from this type of sample should be made with extreme caution.

Multistage Sampling

In practice, most large-scale surveys obtain samples using a combination of the techniques just presented.

As an example of multistage sampling, consider Nielsen Media Research. Nielsen randomly selects households and, through a People Meter, monitors the television programs these households are watching. The meter is an electronic box connected to each TV within the household. The People Meter measures what program is being watched and who is watching it. Nielsen selects the households with the use of a two-stage sampling process.

1. U.S. Census data, Nielsen divides the country into geographic areas (strata). The strata are typically city blocks in urban areas and geographic regions in rural areas. About
2. sends representatives to the selected strata and lists households within the strata. The households are then randomly selected through a simple random sample.

Nielsen sells the information obtained to television stations and companies. These results are used to help determine prices for commercials.

Sample Size Considerations:

Throughout our discussion of sampling, we did not mention how to determine the sample size. Researchers need to know how many individuals they must survey to draw conclusions about the population within some predetermined margin of error.

Researchers must find a balance between the reliability of the results and the cost of obtaining these results. Time and money determine the level of confidence researchers will place on the conclusions drawn from the sample data. The more time and money researchers have available, the more accurate the results of the statistical inference will be.

Later in the course, we will discuss techniques for determining the sample size required to estimate characteristics regarding the population within some margin of error. (For a detailed discussion of sample size considerations, consult a text on sampling techniques such as Elements of Sampling Theory and Methods by Z. Govindarajulu, Pearson, 1999.)

2.5 1.5: Bias in Sampling

Learning Objectives For This Section:

1. **Explain the Sources of Bias in Sampling**

Vocab:

- **Bias:** If the results of the sample are not representative of the population. Sampling bias means that the technique used to obtain the sample's individuals tends to favor one part of the population over another. Any convenience sample has sampling bias because the individuals are not chosen through a random sample.
- **Undercoverage:** Occurs when the proportion of one segment of the population is lower in a sample than it is in the population. This can result if the frame used to obtain the sample is incomplete or not representative of the population.
- **Sampling bias:** sampling bias is a bias in which a sample is collected in such a way that some members of the intended population have a lower or higher sampling probability than others. It results in a biased sample of a population in which all individuals, or instances, were not equally likely to have been selected
- **Nonresponse bias:** exists when individuals selected to be in the sample who do not respond to the survey have different opinions from those who do
  - This can be controlled with **callbacks**.
  - This can also be controlled with **rewards or incentives**
- **Response bias:** Exists when the answers on a survey do not reflect the true feelings of the respondent.
- **Open Question:** Allows the respondent to choose his or her response
- **Closed Question:** requires the respondent to choose from a list of predetermined responses
- **Nonsampling errors:** result from undercoverage, nonresponse bias, response bias, or data-entry error. Such errors could also be present in a census.
- **Sampling error:** results from using a sample to estimate information about a population. This type of error occurs because a sample gives incomplete information about a population.

Explain the sources of Bias in Sampling:

There are three sources of bias in sampling:

- **Sampling bias**
- **Nonresponse bias**
- **Response bias**

Interviewer Error.

Do not be quick to trust surveys conducted by poorly trained interviewers.

Do not trust survey results if the sponsor has a vested interest in the results of the survey.

Misrepresented Errors

Some survey questions result in responses that misrepresent facts or are flat-out lies.

Wording of Questions.

The way a question is worded can lead to response bias in a survey, so questions must always be asked in balanced form.

Ordering of Questions or Words

Many surveys will rearrange the order of the questions within a questionnaire so that responses are not affected by prior questions.

Type of Question.

One of the first considerations in designing a question is determining whether the question should be *open* or *closed*

In closed questions, the possible responses should be rearranged because respondents are likely to choose early choices in a list rather than later choices.

Closed questions limit the number of respondent choices and, therefore, the results are mush easier to analyze. The limited choices, however, do not always include a respondent's desired choice.

An open question should be phrased so that the responses are similar. This allows for easy analysis of the responses

Note:-

Survey designers recommend conducting pretest surveys with open questions and then using the most popular answers as the choices on closed-question surveys
The goal is to limit the number of choices in a closed question without forcing respondents to choose an option they do not prefer, which would make the survey have response bias.

Can a census have bias?

Yes.

A question on a census form could be misunderstood, thereby leading to response bias in the results.

We also mentioned that it is often difficult to contact each individual in a population. For example, the U.S. Census Bureau faces challenges in counting each homeless person in the country, so the census data published by the U.S. government likely suffers from nonresponse bias.

Sampling Error versus Nonsampling Error:

Nonresponse bias, response bias, and data-entry errors are types of nonsampling error.

However, when a sample is used to learn information about a population, sampling error is also likely to occur.

2.6 1.6: The Design of Experiments

Learning Objectives For This Section:

1. **Describe the Characteristics of an Experiment (vocab)**
2. **Explain the Steps in Designing an Experiment**
3. **Explain the Completely Randomized Design**
4. **Explain the Matched-Pairs Design**

Vocab:

- **Experiment:** is a controlled study conducted to determine the effect of varying one or more explanatory variables or **factors** has on a response variable.
- **Factor:** A variable whose effect on the response variable is to be assessed by the experimenter
- **Treatment:** Any combination of the values of the factors is called a treatment
- **Experimental Unit (or subject)** is a person, object or some other well-defined item upon which a treatment is applied
- **Control Group:** Serves as a baseline treatment that can be used to compare to other treatments.
- **Placebo:** is an innocuous medication, such as a sugar tablet, that looks, tastes, and smells like the experimental medication.
- **Blinding:** refers to nondisclosure of the treatment an experimental unit is receiving.
- **Single-blind** experiment is one in which the experimental unit (or subject) does not know which treatment he or she is receiving.
- **Double-blind** experiment is one in which neither the experimental unit nor the researcher in contact with the experimental unit knows which treatment the experimental unit is receiving.
- **Design:** To design an experiment means to describe the overall plan in conducting the experiment. Conducting an experiment requires a series of steps.
- **completely randomized design:** is one in which each experimental unit is randomly assigned to a treatment.
- **matched-pairs design:** is an experimental design in which the experimental units are paired up. The pairs are selected so that they are related in some way (that is, the same person before and after a treatment, twins, husband and wife, same geographical location, and so on). There are only two levels of treatment in a matched-pairs design.

Steps in Designing an Experiment:

1. Identify the Problem to Be Solved. The statement of the problem should be as explicit as possible and should provide the experimenter with direction. The statement must also identify the response variable and the population to be studied. Often, the statement is referred to as the claim.
2. Determine the Factors That Affect the Response Variable. The factors are usually identified by an expert in the field of study. In identifying the factors, ask, "What things affect the value of the response variable?" After the factors are identified, determine which factors to fix at some predetermined level, which to manipulate, and which to leave uncontrolled.
3. Determine the Number of Experimental Units. As a general rule, choose as many experimental units as time and money allow. Techniques exist for determining sample size, provided certain information is available.
4. Determine the Level of Each Factor. There are two ways to deal with the factors, control or randomize.

(a) Control: There are two ways to control the factors.

a.) Set the level of a factor at one value throughout the experiment (if you are not interested in its effect on the response variable).

b.) Set the level of a factor at various levels (if you are interested in its effect on the response variable). The combinations of the levels of all varied factors constitute the treatments in the experiment.

(b) Randomize: Randomly assign the experimental units to treatment groups. Because it is difficult, if not impossible, to identify all factors in an experiment, randomly assigning experimental units to treatment groups reduces the effect of variation attributable to factors (explanatory variables) not controlled. That is, randomly assigning experimental units to treatment groups tends to "even out" any uncontrolled factors.

5. Conduct the Experiment.

a.) Replication occurs when each treatment is applied to more than one experimental unit. Using more than one experimental unit for each treatment ensures the effect of a treatment is not due to some characteristic of a single experimental unit. It is a good idea to assign an equal number of experimental units to each treatment.

b.) Collect and process the data. Measure the value of the response variable for each replication. Then organize the results. The idea is that the value of the response variable for each treatment group is the same before the experiment because of randomization. Then any difference in the value of the response variable among the different treatment groups is a result of differences in the level of the treatment.

6. Test the Claim. This is the subject of inferential statistics. Inferential statistics is a process in which generalizations about a population are made on the basis of results obtained from a sample. Provide a statement regarding the level of confidence in the generalization.

Explain a Matched-Pairs Design:

In matched-pairs design, one matched individual will receive one treatment and the other receives a different treatment. The matched pair is randomly assigned to the treatment using a coin flip or a random-number generator. We then look at the difference in the results of each matched pair. One common type of matched-pairs design is to measure a response variable on an experimental unit before and after a treatment is applied. In this case, the individual is matched against itself. These experiments are sometimes called before–after or pretest–posttest experiments.

3 Chapter 2:

3.1 2.1: Organizing Qualitative Data

Learning Objectives for This Section:

1. **Organize Qualitative Data in Tables**
2. **Construct Bar Graphs**
3. **Construct Pie Charts**

Vocab:

- **A frequency distribution** lists each category of data and the number of occurrences for each category of data.
- **The relative frequency** is the proportion (or percent) of observations within a category and is found using the formula

$$Relative\ frequency = \frac{Frequency}{Sum\ of\ all\ frequency}.$$

- **A relative frequency distribution** lists each category of data together with the relative frequency.
- **A bar graph** is constructed by labeling each category of data on either the horizontal or vertical axis and the frequency or relative frequency of the category on the other axis. Rectangles of equal width are drawn for each category. The height of each rectangle represents the category's frequency or relative frequency.
- **A Pareto chart** is a bar graph whose bars are drawn in decreasing order of frequency or relative frequency.
- **A pie chart** is a circle divided into sectors. Each sector represents a category of data. The area of each sector is proportional to the frequency of the category.

Organize Qualitative Data in Tables

When qualitative data are collected, we often first determine the number of occurrences within each category.

Frequency Distribution Chart

To construct a frequency distribution, create a list of the body parts (categories) and tally each occurrence. Then, add up the number of tallies (observations) to determine the frequency.

Note:-

In any frequency distribution, it is a good idea to add up the frequency column to make sure that it equals the number of observations.

Relative Frequency Distribution of qualitative data:

Similar to the frequency distribution chart, However in this distribution we will add all the frequencies and then use:

$$Relative\ frequency = \frac{Frequency}{Sum\ of\ all\ frequency}.$$

to compute the relative frequency of each category of data.

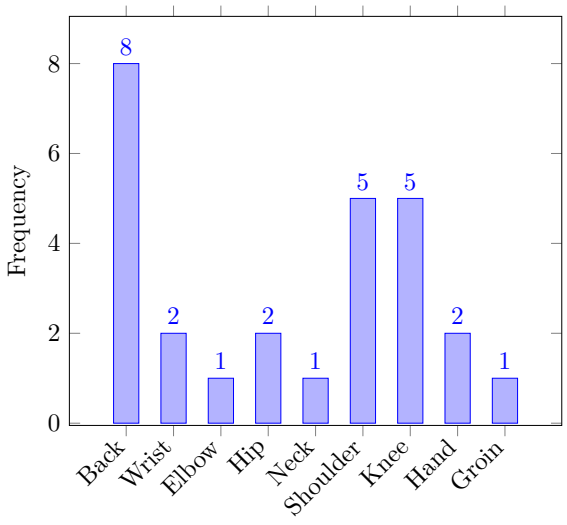
Note:-

It is a good idea to add up the relative frequencies to be sure they sum to 1. It is a good idea to add up the relative frequencies to be sure they sum to 1. In decimal form, the sum may differ slightly from 1 due to rounding

Construct Bar Graphs:

Use a horizontal axis to indicate the categories of the data and a vertical axis to represent the frequency or relative frequency. Draw rectangles of equal width to the height that is the frequency or relative frequency for each category. The bars do not touch each other.

Figure:



Comparing Two Sets of Data

First, determine the relative frequencies of each category for each year. To construct side-by-side bar graphs, draw two bars for each category of data.

Horizontal Bars

Bar graphs may also be drawn with horizontal bars. Horizontal bars are preferable when category names are lengthy. For example, Figure 4 uses horizontal bars to display the same data as in Figure 3.

Construct Pie-Graphs

Pie charts are typically used to present the relative frequency of qualitative data. In most cases, the data are nominal, but ordinal data can also be displayed in a pie chart.

When should a Bar Graph or Pie Chart be Used?

Pie chart are useful for showing the division of all possible values of a qualitative variable into its parts.

However, because angles are often hard to judge in pie charts, they are not as useful in comparing two specific values of the qualitative variable.

Instead, the emphasis is on comparing the part to the whole.

Bar graphs are useful are useful when we want to compare the different parts, not necessarily the parts to the whole.

Since bars are easier to draw and compare, some forgo pie charts in favor of Pareto charts when comparing parts to the whole.

3.2 2.2: Organizing Quantitative Data: The Popular Displays

Learning Objectives For This Section:

1. **Organize Discrete Data in Tables**
2. **Construct Histograms of Discrete Data**
3. **Organize Continuous Data in Tables**
4. **Construct Histograms of Continuous Data**
5. **Draw Dot Plots**
6. **Identify the Shape of a Distribution**

Vocab:

- A **histogram** is constructed by drawing rectangles for each class of data. The height of each rectangle is the frequency or relative frequency of the class. The width of each rectangle is the same, and the rectangles touch each other.
- **Classes:** The Categories in which data is grouped
- **lower class limit:** the smallest value within the class
- **upper class limit:** the largest value within the class
- **Class Width:** is the difference between consecutive lower class limits.
- A table is **open ended** if the first class has no lower class limit or the last class has no upper class limit
- We draw a **dot plot** by placing each observation horizontally in increasing order and placing a dot above the observation each time it is observed.
- **uniform distribution:** frequency of each value of the variable is evenly spread across the values of the variable.
- **bell-shaped distribution:** highest frequency occurs in the middle and frequencies tail off to the left and right of the middle.
- **skewed right:** the tail to the right of the peak is longer than the tail to the left of the peak
- **skewed left:** tail to the left of the peak is longer than the tail to the right of the peak.

Organize Discrete Data in Tables

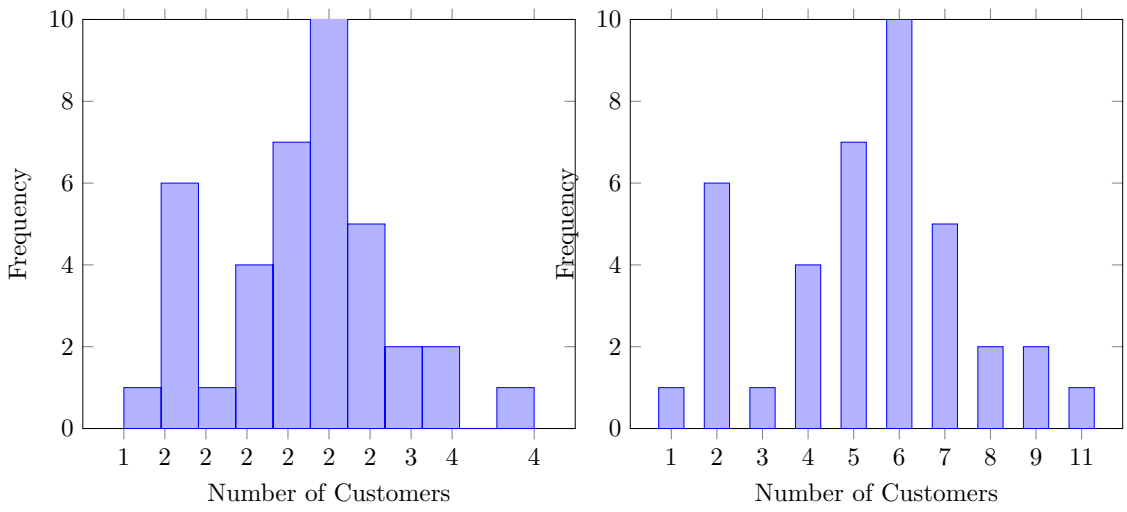
Use the values of the discrete variable to create the classes when the number of distinct data values is small. The approach to summarizing the data is similar to that of constructing frequency or relative frequency distributions from qualitative data where the categories of data are determined by the actual observations.

Construct Histograms of Discrete Data

The *histogram*, a graph used to present quantitative data, is similar to the bar graph.

The value of each category of data (number of customers) is on the horizontal axis and the frequency or relative frequency is on the vertical axis. Draw rectangles of equal width centered at the value of each category.

Histogram (Left) vs Bar Graph (Right)



Organize Data Into Tables:

When a data set consists of a large number of different discrete data values or when a data set consists of continuous data, create classes by using intervals of numbers.

StatCrunch Steps:

- Data > Bin
- Select the column containing the data
- Choose a starting point and a binwidth (or automatic)
- Stat > Tables > Frequency (With newly generated data)

Draw Histograms of Continuous Data:

StatCrunch Steps:

- Graph > Histogram
- Select Data
- Input Bins

Constructing Histograms Is Somewhat of an Art Form

There is no one correct frequency distribution for a particular set of data. However, some frequency distributions better illustrate patterns within the data than others. So constructing frequency distributions is somewhat of an art form. Use the distribution that seems to provide the best overall summary of the data.

When the data set is small, we usually want fewer classes. When the data set is large, we usually want more classes. The larger the class width, the fewer the classes in a frequency distribution. Use the following guidelines to help determine an appropriate lower class limit of the first class and class width.

Guidelines for Determining the Lower Class Limit of the First Class and Class Width

Choosing the Lower Class Limit of the First Class:

Choose the smallest observation in the data set or a convenient number slightly smaller than the smallest observation in the data set.

Determining the Class Width:

- Decide on the number of classes. Generally, there should be between 5 and 20 classes. The smaller the data set, the fewer the classes.
- Determine the class width by computing

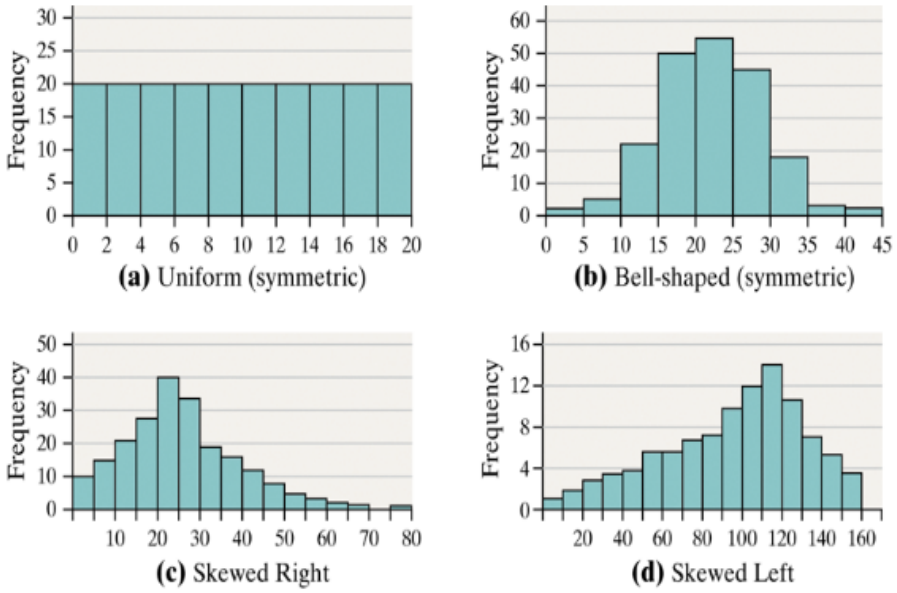
$$Class\ Width \approx \frac{Largest\ data\ value - smallest\ data\ value}{number\ of\ classes}.$$

- Round the value to a convenient number. Rounding up may result in fewer classes than were originally intended, while rounding down may result in more class than originally intended.

Drawing Dot Plots:

We draw a dot plot by placing each observation horizontally in increasing order and placing a dot above the observation each time it is observed.

Identify the Shape of a Distribution:



3.3 2.4: Graphical Misrepresentations of Data

Learning Objectives For This Section:

1. **Describe What Can Make a Graph Misleading or Deceptive**

Describe What Can Make a Graph Misleading or Deceptive

Statistics is the only science that enables different experts using the same figures to draw different conclusions.

Graphs that mislead unintentionally create an incorrect impression

Graphs that are deceptive purposely create an incorrect impression

Most common graphical misrepresentations of data:

- **Scale**
- **Inconsistent Scale**
- **Misplaced Origin**

Guidelines for Constructing Good Graphics

- Label and name the axes clearly, providing explanations if needed. Include units of measurement and a data source when appropriate.
- Include a meaningful title on the graph.
- Avoid distortion. Never lie about the data.
- Minimize the amount of white space in the graph. Use the available space to let the data stand out. If you truncate the scales, clearly indicate this to the reader.
- Avoid clutter, such as excessive gridlines and unnecessary backgrounds or pictures. Don't distract the reader from the data.
- Avoid three dimensions. Three-dimensional charts may look nice, but they distract the reader and often lead to misinterpretation of the graphic.
- Do not use more than one design in the same graphic. Sometimes graphs use a different design in one portion to draw attention to that area. Don't try to force the reader to a specific part of the graph. Let the data speak for themselves.
- Avoid relative graphs that do not contain data or scales.
- One final point to make. When reading graphs, look at the source of the data represented in the graphic. Often, a group with an agenda will conduct allegedly unbiased studies and report the results that support their position. Always "consider the source" and any possible hidden agendas they may have when reading graphics.

4 Chapter 3:

4.1 3.1: Measures of Central Tendency

Learning Objectives For This Section:

1. **Determine the Arithmetic Mean of a Variable from Raw Data**
2. **Determine the Median of a Variable from Raw Data**
3. **Explain What It Means for a Statistic to be Resistant**
4. **Determine the Mode of a Variable from Raw Data**

Vocab:

- **The arithmetic mean** of a variable is computed by adding all the values of the variable in the data set and dividing by the number of observations.
- **The population arithmetic mean**, μ, (pronounced "mew"), is a parameter that is computed using data from all the individuals in a population.

$$\mu = \frac{x_1 + x_2 + x_N}{N} = \frac{\sum x_i}{N}.$$

- **The sample arithmetic mean**, $\bar{x}$ (pronounced x-bar"), is a statistic that is computed using data from individuals in a sample.

$$\bar{x} = \frac{x_1 + x_2 + x_n}{n} = \frac{\sum x_i}{n}.$$

- **The median** of a variable is the value that lies in the middle of the data when arranged in ascending order. We use M to represent the median.
- A numerical summary of data is said to be **resistant** if observations that are extreme (very large or small) relative to the data do not affect its value substantially.
  - So the median is resistant, but the mean is not resistant.
- **The mode** of a variable is the observation of the variable that occurs most frequently in the data set.
  - If no observation occurs more than once, we say that the data have **no mode**.
- **Bimodal:** If the data has two modes
- **Multimodal:** If the data has more than two modes
- A numerical summary of data is said to be **resistant** if observations that are extreme (very large or small) relative to the data do not affect its value substantially.
  - So the median is resistant, but the mean is not resistant.

Determine the Arithmetic Mean of a Variable from Raw Data:

Throughout this course, we agree to round the mean to one more decimal place than that in the raw data.

If $x_1, x_2, ..., x_N$ are the N observations of a variable from a population, then the population mean, μ (pronounced "mew"), is:

$$\mu = \frac{x_1 + x_2 + ... + x_N}{N} = \frac{\sum x_i}{N}.$$

If $x_1, x_2, ..., x_n$ are the n observations of a variable from a sample, then the sample mean, $\overline{x}$ (pronounced "x-bar"), is:

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum x_i}{n}.$$

Example: The table shows the first exam scores of the ten students enrolled in Introductory Statistics.

Student	Score
1. Michelle	82
2. Ryanne	77
3. Bilal	90
4. Pam	71
5. Jennifer	62
6. Dave	68
7. Joel	74
8. Sam	84
9. Justine	94
10. Juan	88

a.) **Compute the population mean,** μ

b.) **Find a simple random sample of size** $n = 4$ **students**

c.) **compute the sample mean,** $\overline{x}$

a.) To compute the population mean, μ, add all the data values (test scores) and then divide by the number of individuals in the population.

$$\frac{\sum x_i}{N}.$$

Since we know $N = 10$:

$$\sum x_i = 82 + 77 + 90 + 71 + 62 + 68 + 74 + 84 + 94 + 88$$

$$= 790.$$

Now:

$$\frac{790}{10} = 79.$$

Note: Although it was not necessary in this problem, we will agree to round the mean to one more decimal place than that in the raw data.

b.) To find a simple random sample of size $n = 4$, we will use a ti-84 calculator.

$$rand(1, 10, 4) = \{5, 10, 2, 6\}$$
$$= Jennifer,\ Juan,\ Ryanne,\ Dave.$$

c.)

$$\frac{x_5 + x_{10} + x_2 + x_6}{4}$$
$$= \frac{62 + 88 + 77 + 68}{4}$$
$$= 73.8.$$

Note: Notice that we rounded the sample mean to the nearest tenth (which is one more decimal point than the original data).

Determine the Median of a Variable from Raw Data

A second measure of central tendency is the median. To compute the median of a set of data, the data must be quantitative.

Steps in Finding the Median of a Data Set:

1. Arrange the data in ascending order.
2. Determine the number of observations, n
3. Determine the observation in the middle of the data set.
   - If the number of observations is odd, then the median is the data value exactly in the middle of the data set. That is, the median is the observation that lies in the $\frac{n+1}{2}$ position.
   - If the number of observations is even, then the median is the mean of the two middle observations in the data set. That is, the median is the mean of the observations that lie in the $\frac{n}{2}$ and the $\frac{n}{2} + 1$ position.

Example: Determining the Median of a Data Set (Odd Number of Observations)

Song Name	Length
"Sister Golden Hair"	201
"Black Water"	257
"Free Bird"	284
"The Hustle"	208
"Southern Nights"	179
"Stayin' Alive"	222
"We Are Family"	217
"Heart of Glass"	206
"My Sharona"	240

Step 1. Arrange the data in ascending order:

$$179, 201, 206, 208, 217, 222, 240, 257, 284.$$

Step 2. Notice that there are $n = 9$ observations

Step 3. Because n is odd, the median is the observation exactly in the middle of the data set with the data

written in ascending order. This value lies in the 5^{th} position.

$$M = 217.$$

Example: Determining the Median of a Data Set (Even Number of Observations)

Student	Score
1. Michelle	82
2. Ryanne	77
3. Bilal	90
4. Pam	71
5. Jennifer	62
6. Dave	68
7. Joel	74
8. Sam	84
9. Justine	94
10. Juan	88

Step 1. Arrange the data in ascending order:

$$62, 68, 71, 74, 77, 82, 84, 88, 90, 94.$$

Step 2. Notice there are $n = 10$ observations

Step 3. Because n is even, the median is the mean of the two middle observations, the fifth $\left(\frac{n}{2} = \frac{10}{2} = 5\right)$ and sixth $\left(\frac{n}{2} + 1 = \frac{10}{2} + 1 = 6\right)$ observations with the data written in ascending order. So the median is the mean of 77 and 82

$$\begin{aligned} M &= \frac{77 + 82}{2} \\ &= 79.5. \end{aligned}$$

Explain What It Means for a Statistic to be Resistant:

Which measure of central tendency is better to use—the mean or the median? It depends.

Note:-

The value of the mean can be impacted by a single observation.

You may be asking yourself, "Why would I ever compute the mean?" After all, the mean and median are close in value for symmetric data, and the median is the better measure of central tendency for skewed data. The reason we compute the mean is that much of statistical inference is based on the mean.

Relation among the Mean, Median, and Distribution Shape

Distribution Shape	Mean versus Median
Skewed left	Mean substantially smaller than median
Symmetric	Mean roughly equal to median
Skewed right	Mean substantially larger than median

Note:-

A word of caution is in order. The relation between the mean, median, and skewness are guidelines. These guidelines tend to hold up well for continuous data, but when the data are discrete the rules can be easily violated.
When the data set is skewed the median is the preferred measure of central tendency.
When the data set is symmetric the mode is the preferred measure of central tendency.

Determine the Mode of a Variable from Raw Data

A third measure of central tendency is the **mode**, which can be computed for either quantitative or qualitative data.

- To compute the mode, tally the number of observations that occur for each data value.
- The data value that occurs most often is the mode.
- If no observation occurs more than once, we say that the data have no mode.
- A set of data can have no mode, one mode, or more than one mode.

Summary:

Measure of Central Tendency	Computation	Interpretation	When to Use
Mean	Population mean: $\mu = \frac{\Sigma x_i}{N}$ Sample mean: $\overline{x} = \frac{\Sigma x_i}{n}$	Center of gravity	When data are quantitative and the frequency distribution is roughly symmetric
Median	Arrange the data in ascending order and find the observation in the middle	Divides the bottom 50% of the data from the top 50%	When the data are quantitative and the frequency distribution is skewed left or skewed right
Mode	Tally data to determine most frequent observation	Most frequent observation	When the most frequent observation is the desired measure of central tendency or the data are qualitative

4.2 3.2: Measures of Dispersion

Learning Objectives For This Section:

1. **Determine the Range of a Variable from Raw Data**
2. **Determine the Standard Deviation of a Variable from Raw Data**
3. **Determine the Variance of a Variable from Raw Data**
4. **Use the Empirical Rule to Describe Data That Are Bell-Shaped**

Vocab:

- **Dispersion:** Degree to which the data are spread out.
- **Range:** The range, r, of a variable is the difference between the largest and smallest data value. That is,

  $$range = R = Largest\ data\ value - smallest\ data\ value.$$

  Note: Range is **not** resistant
- **Deviation:** a deviation refers to the difference between an individual data point and a central value, such as the mean or median. It represents how much a particular data point varies or deviates from the average or typical value in a data set. When can comopute a deviation with:

  $$Individual\ data\ point - mean.$$

  We call this calculation, "deviation about the mean"

  Note: The sum of the deviations about the mean always equals zero
- **The population standard deviation** of a variable is the square root of the sum of squared deviations about the population mean divided by the number of observations in the population, N. The population standard deviation is symbolically represented by σ (lowercase Greek sigma). The formula is given by:

  $$\sigma = \sqrt{\frac{(x_1-\mu)^2 + (x_2-\mu)^2 + ... + (x_N-\mu)^2}{N}} = \sqrt{\frac{\sum(x_i-\mu)^2}{N}}.$$

  Note: Standard Deviation is **not** resistant
- **The sample standard deviation**, s, of a variable is the square root of the sum of squared deviations about the sample mean divided by $n-1$, where n is the sample size. The formula is given as

  $$s = \sqrt{\frac{(x_1-\overline{x})^2 + (x_2-\overline{x})^2 + ... + (x_n-\overline{x})^2}{n-1}} = \sqrt{\frac{\sum(x_i-\overline{x})^2}{n-1}}.$$

  Note: Standard Deviation is **not** resistant
- we call $n-1$ the **degrees of freedom** because the first $n-1$ observations have freedom to be any value, but the n^{th} observation has no freedom. It must be whatever value forces the sum of the deviations about the mean to equal zero.
- The variance of a variable is the square of the standard deviation.
  - **The population variance** is σ^2
  - **The Sample Variance** is s^2

  Note: The units of measure in variance are squared values. So if the variable is measured in dollars, the variance is measured in dollars squared. This makes interpreting the variance difficult.

measures of center (such as the mean) are not sufficient in describing distributions of data. We determine numerical measures of dispersion to quantify the spread of data. This section discusses three numerical measures of dispersion, or spread, of data: the range, standard deviation, and variance.

Determine the Range of a Variable from Raw Data:

$$range = R = Largest\ data\ value - smallest\ data\ value.$$

Note: Range is **not** resistant

Determine the Standard Deviation of a Variable from Raw Data:

Example: Population Standard Deviation

Consider the data:

Student	**Score**
1. Michelle	82
2. Ryanne	77
3. Bilal	90
4. Pam	71
5. Jennifer	62
6. Dave	68
7. Joel	74
8. Sam	84
9. Justine	94
10. Juan	88

To use the formula:

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}.$$

We must first compute μ:

$$\begin{aligned} \mu &= \frac{\sum x_i}{N} \\ &= \frac{82 + 77 + 90 + 71 + 62 + 68 + 74 + 84 + 94 + 88}{10} \\ &= 79. \end{aligned}$$

Now:

$$\begin{aligned} \sigma &= \sqrt{\frac{(82 - 79)^2 + (77 - 79)^2 + ... + (88 - 79)^2}{10}} \\ &= \sqrt{96.4\ Points^2} \\ &= 9.8\ Points. \end{aligned}$$

Example: Calculating Standard Deviation (Sample Standard Deviation)

Consider the data:

Name	Exam Scores
Jennifer	62
Juan	88
Ryanne	77
Dave	68

If:

$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}.$$

First, we must find $\overline{x}$:

$$\begin{aligned} \overline{x} &= \frac{\sum x_i}{n} \\ &= \frac{62 + 88 + 77 + 68}{4} \\ &= 73.75. \end{aligned}$$

Now that we have computed $\overline{x}$, we can use the standard deviation formula:

$$\begin{aligned} s &= \sqrt{\frac{(62 - 73.25)^2 + (88 - 73.25)^2 + (77 - 73.25)^2 + (68 - 73.25)^2}{4-1}} \\ &= \sqrt{128.25\ points^2} \\ &\approx 11.3. \end{aligned}$$

Remember, round the sample standard deviation to one more decimal place than the raw data.

Interpretations of the Standard Deviation

How does the value of the standard deviation relate to the spread of the distribution?

Standard deviation represents the "typical" deviation from the mean. As such, the standard deviation may be used to judge whether a particular observation is "far away" from the mean of a data set. For example, is a measure of 31 cm far from 25 cm? It depends. If the standard deviation of the data is 6cm cm, then the answer is no because 31 cm would be only 1 standard deviation from 25 cm. However, if the standard deviation of the data is 2 cm, then the answer is yes because 31 cm would be 3 standard deviations from 25 cm. **A good rule of thumb is to consider an observation "far away" if it is more than 2 standard deviations from the other observation (such as the mean).**

So, when judging the unusualness of an observation, it is vital that you consider the underlying variation in the data as measured by the standard deviation.

When comparing two populations, the larger the standard deviation, the greater the dispersion, or spread, of the distribution provided the variable of interest from the two populations has the same unit of measure. The units of measure must be the same so that we are comparing "apples with apples." For example, \$100 is not the same as 100 japanese yen (because recently was equivalent to about 114 yen) yen). So a standard deviation of \$100 yen). So a standard deviation of 100 yen

The standard deviation is used to describe the spread in symmetric distributions (while the mean is used to describe the center of the distribution).

Note:-

Higher Standard Deviation = more dispersion.

Determine the Variance of a Variable from Raw Data

The units of measure in variance are squared values. So if the variable is measured in dollars, the variance is measured in dollars squared. This makes interpreting the variance difficult.

We can compute the variance in one of two ways.

1. Compute the std. dev and then square the result
2. The expression under the radical in the formula for standard deviation is the formula for variance. Therefore if we drop the radical in the standard deviation formula, we can compute the variance.

Note:-

Using a rounded value of the standard deviation to obtain the variance results in a round-off error

Bias in the Variance and Standard Deviation

The sample variance is obtained using the formula:

$$s^2 = \frac{\sum_{n}^{i=1}(x_i - \overline{x})^2}{n-1}.$$

What if we divided by n instead of $n-1$? to obtain the sample variance, as one might expect? Then the sample variance would consistently underestimate the population variance. Whenever a statistic consistently underestimates a parameter, it is said to be biased. To obtain an unbiased estimate of the population variance, divide the sum of the squared deviations about the sample mean by $n-1$

Unfortunately, the sample standard deviation given by the formula:

$$s = \sqrt{\frac{\sum_{n}^{i=1}(x_i - \overline{x})^2}{n-1}}.$$

is not an unbiased estimate of the population standard deviation. In fact, it is not possible to provide an unbiased estimator of the population standard deviation for all distributions. The explanation is beyond the scope of this class (it has to do with the shape of the square root function). However, for the applications in this text, the bias is minor and does not impact results.

Use the Empirical Rule to Describe Data That Are Bell-Shaped

If data have a distribution that is bell-shaped, the Empirical Rule can be used to determine the percentage of data that will lie within k standard deviations of the mean.

The Empirical Rule:

If a distribution is roughly bell shaped, then

- approximately 68% of the data within 1 standard deviation of the mean. That is, approximately 68% of the data will lie between $\mu - 1\sigma$ and $\mu + 1\sigma$
- approximately 95% of the data within 2 standard deviation of the mean. That is, approximately 95% of the data will lie between $\mu - 2\sigma$ and $\mu + 2\sigma$
- approximately 99.7% of the data within 3 standard deviation of the mean. That is, approximately 99.7% of the data will lie between $\mu - 3\sigma$ and $\mu + 3\sigma$

Note:-

The Empirical Rule can also be used based on sample data with $\overline{x}$ in place of μ and s in place of σ

4.3 3.3: Measures of Central Tendency and Dispersion from Grouped Data

Learning Objectives For This Section:

1. **Approximate the Mean of a Variable from Grouped Data**
2. **Compute the Weighted Mean**
3. **Approximate the Standard Deviation from a Frequency Distribution**

Vocab/Formulas:

- **Class Midpoint:** The class midpoint is the sum of consecutive lower class limits divided by 2
- **Approximate Population Mean (if we do not have access to the original data, ie data has been grouped (classed) and frequency chart has already been made)**

$$\mu = \frac{\sum x_i f_i}{\sum f_i}$$
$$= \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{f_1 + f_2 + \cdots + f_n}.$$

  where: x_i is the midpoint or value of the i th class
  f_i is the frequency of the i th class
  n is the number of classes

- **Approximate Sample Mean (if we do not have access to the original data, ie data has been grouped (classed) and frequency chart has already been made)**

$$\overline{x} = \frac{\sum x_i f_i}{\sum f_i}$$
$$= \frac{x_1 f_1 + x_2 f_2 + ... + x_n f_n}{f_1 + f_2 + ... + f_n}.$$

  where: x_i is the midpoint or value of the i th class
  f_i is the frequency of the i th class
  n is the number of classes

- **The weighted mean**, $\overline{x}_w$, of a variable is found by multiplying each value of the variable by its corresponding weight, adding these products, and dividing this sum by the sum of the weights. It can be expressed using the formula

$$\overline{x}_w = \frac{\sum w_i x_i}{w_i} = \frac{w_1 + x_1 + w_2 + x_2 + ... + w_n + x_n}{w_1 + w_2 + ... + w_n}.$$

  Where: w_i is the weight of the i^{th} observation
  x_i is the value of the i^{th} observation.

- **Approximate Population Standard Deviation (if we do not have access to the original data, ie data has been grouped (classed) and frequency chart has already been made)**

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2 f_i}{\sum f_i}}.$$

  Where: x_i is the midpoint or value of the ith class
  f_i is the frequency of the i^{th} class

- **Approximate Sample Standard Deviation (if we do not have access to the original data, ie data has been grouped (classed) and frequency chart has already been made)**

$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2 f_i}{\sum f_i - 1}}.$$

  Where: x_i is the midpoint or value of the ith class
  f_i is the frequency of the i^{th} class

Approximate the Mean of a Variable from Grouped Data

Because raw data cannot be retrieved from a frequency table, we assume that within each class, the mean of the data values is equal to the class midpoint. Then multiply the class midpoint by the frequency. This product is expected to be close to the sum of the data that lie within the class. Repeat the process for each class and add the results. This sum approximates the sum of all the data.

In each formula, $x_1\ f_1$ approximates the sum of all the data values in the first class, $x_2 f_2$ approximates the sum of all the data values in the second class, and so on. Notice that the formulas for the population mean and sample mean are essentially identical, just as they were for computing the mean from raw data.

Example: Approximating the Mean for Continuous Quantitative Data from a Frequency Distribution

Consider the data:

Class	Frequency
8–8.99	2
9–9.99	2
10–10.99	4
11–11.99	1
12–12.99	6
13–13.99	13
14–14.99	7
15–15.99	3
16–16.99	1
17–17.99	0
18–18.99	0
19–19.99	1

First let's find x_i (Midpoints)

Interval	Frequency	Midpoint (Full Expression)	Midpoint (Simplified)
8–8.99	2	$\frac{8+9}{2}$	8.5
9–9.99	2	$\frac{9+10}{2}$	9.5
10–10.99	4	$\frac{10+11}{2}$	10.5
11–11.99	1	$\frac{11+12}{2}$	11.5
12–12.99	6	$\frac{12+13}{2}$	12.5
13–13.99	13	$\frac{13+14}{2}$	13.5
14–14.99	7	$\frac{14+15}{2}$	14.5
15–15.99	3	$\frac{15+16}{2}$	15.5
16–16.99	1	$\frac{16+17}{2}$	16.5
17–17.99	0	$\frac{17+18}{2}$	17.5
18–18.99	0	$\frac{18+19}{2}$	18.5
19–19.99	1	$\frac{19+20}{2}$	19.5

Now we can proceed with the formula:

$$\overline{x} = \frac{\sum x_i f_i}{\sum f_i}.$$

So:

$$\begin{aligned}\sum x_i f_i &= (8.5 \times 2) + (9.5 \times 2) + (10.5 \times 4) + (11.5 \times 1) + (12.5 \times 6) \\ &\quad + (13.5 \times 13) + (14.5 \times 7) + (15.5 \times 3) + (16.5 \times 1) + (17.5 \times 0) \\ &\quad + (18.5 \times 0) + (19.5 \times 1) = 524 \\ f_i &= 2 + 2 + 4 + 1 + 6 + 13 + 7 + 3 + 1 + 0 + 0 + 1 = 40\end{aligned}$$

Therefore:

$$\begin{aligned}\overline{x} &= \frac{524}{40} \\ &= 13.1.\end{aligned}$$

Thus we can conclude that approximate mean is 13.1%

Steps for StatCrunch:

- Stat > Summary Stats > Grouped/Binned data
- **Midpoints Defined By The Average Of:** →Consecutive Lower Limits

Compute the Weighted Mean

When data values have different importance, or weights, associated with them, we compute the weighted mean. For example, grade point average is a weighted mean, with weights equal to the number of credit hours in each course. The value of the variable is equal to the grade converted to a point value.

Example: Computing the Weighted Mean

Consider the data:

Class	Credit Hours	Grade
Statistics	4	A (4 Points)
Sociology	3	B (3 Points)
Psychology	3	A (4 Points)
Computer Programming	5	C (2 Points)
Drama	1	A (4 Points)

So:

$$\begin{aligned}\overline{x}_w = \frac{\sum w_i x_i}{\sum w_i} &= \frac{(4 \cdot 4) + (3 \cdot 3) + (3 \cdot 4) + (5 \cdot 2) + (1 + 4)}{4 + 3 + 3 + 5 + 1} \\ &= \frac{51}{16} = 3.19.\end{aligned}$$

Therefore:

$$\begin{aligned}&\frac{51}{16} \\ &= 3.19.\end{aligned}$$

Thus, Marissa's grade-point average for her first semester is 3.19

Steps for finding weighted mean in statcrunch:

1. Stat > Summary Stats > Grouped/Binned Data

Approximate the Standard Deviation from a Frequency Distribution

The procedure for approximating the standard deviation from grouped data is similar to that of finding the mean from grouped data. Because we do not have access to the original data, the standard deviation is approximate.

Example: Approximating the Standard Deviation from a Frequency Distribution

Consider the data:

Interval	Frequency	Midpoint (Full Expression)	Midpoint (Simplified)
8–8.99	2	$\frac{8+9}{2}$	8.5
9–9.99	2	$\frac{9+10}{2}$	9.5
10–10.99	4	$\frac{10+11}{2}$	10.5
11–11.99	1	$\frac{11+12}{2}$	11.5
12–12.99	6	$\frac{12+13}{2}$	12.5
13–13.99	13	$\frac{13+14}{2}$	13.5
14–14.99	7	$\frac{14+15}{2}$	14.5
15–15.99	3	$\frac{15+16}{2}$	15.5
16–16.99	1	$\frac{16+17}{2}$	16.5
17–17.99	0	$\frac{17+18}{2}$	17.5
18–18.99	0	$\frac{18+19}{2}$	18.5
19–19.99	1	$\frac{19+20}{2}$	19.5

Next, we can find $\overline{x}$:

$$\begin{aligned}\sum x_i f_i &= (8.5 \cdot 2) + (9.5 \cdot 2) + (10.5 \cdot 4) + (11.5 \cdot 1) + (12.5 \cdot 6) \\ &+ (13.5 \cdot 13) + (14.5 \cdot 7) + (15.5 \cdot 3) + (16.5 \cdot 1) + (19.5 \cdot 1) \\ &= 524\end{aligned}$$

.

$$\begin{aligned}\sum f_i &= 2 + 2 + 4 + 1 + 6 + 13 + 7 + 3 + 1 + 1 \\ &= 40 \\ \overline{x} = \frac{\sum x_i f_i}{\sum f_i} &= \frac{524}{40} \\ &= 13.1.\end{aligned}$$

Now:

$$\begin{aligned}\sum (x_i - \overline{x}_i)^2 f_i &= ((8.5 - 13.1)^2 \cdot 2) + ((9.5 - 10.33)^2 \cdot 2) + ((10.5 - 10.33)^2 \cdot 4) \\ &+ ((11.5 - 13.1)^2 \cdot 1) + ((12.5 - 10.33)^2 \cdot 6) + ((13.5 - 10.33)^2 \cdot 13) \\ &+ ((14.5 - 13.1)^2 \cdot 7) + ((15.5 - 10.33)^2 \cdot 3) + ((16.5 - 10.33)^2 \cdot 1) \\ &+ ((19.5 - 13.1)^2 \cdot 1) \\ &= 185.6.\end{aligned}$$

And:

$$\begin{aligned}\sum f_i &= 2+2+4+1+6+13+7+3+1+1 \\ &= 40.\end{aligned}$$

Finally:

$$\begin{aligned}\sigma &= \sqrt{\frac{185.6}{40-1}} \\ &= 2.182.\end{aligned}$$

Thus, The approximate standard deviation of the five-year rate of return is 2.182%

4.4 3.4: Measures of Position

Learning Objectives For This Section:

1. **Determine and Interpret z-Scores**
2. **Interpret Percentiles**
3. **Determine and Interpret Quartiles**
4. **Determine and Interpret the Interquartile Range**
5. **Check a Set of Data for Outliers**

Vocab:

- **The z-score** represents the distance that a data value is from the mean in terms of the number of standard deviations. We find it by subtracting the mean from the data value and dividing this result by the standard deviation.
  - **Population Z-score**

$$z = \frac{x - \mu}{\sigma}.$$

  - **Sample Z-score**

$$z = \frac{x - \overline{x}}{s}.$$

  Note: The Z-score is unitless. It has mean 0 and a standard deviation of 1
  Round z-scores to the nearest hundredth

- The median is a special case of a general concept called the **percentile.**
- **the k^{th} percentile,** denoted P_k, of a set of data is a value such that k percent of the observations are less than or equal to the value.
- The most common percentiles are **quartiles**, which divide data sets into fourths, or four equal parts.
- The **interquartile range, IQR,** is the range of the middle 50% of the observations in a data set. That is, the IQR is the difference between the first and third quartiles and is found using this formula

$$IQR = Q_3 - Q_1.$$

- **Outliers:** When analyzing data, we must check for extreme observations, called outliers. Outliers can occur by chance, because of errors in the measurement of a variable, during data entry, or from errors in sampling.
- **Fences** serve as cutoff points for determining outliers.

$$\begin{aligned} Lower\ Fence &= Q_1 - 1.5 \cdot IQR \\ Upper\ Fence &= Q_3 + 1.5 \cdot IQR. \end{aligned}$$

Determine and Interpret z-scores:

If a data value is larger than the mean, the z-score is positive. If a data value is smaller than the mean, the z-score is negative. If the data value equals the mean, the -score is zero. A z-score measures the number of standard deviations an observation is above or below the mean. For example, a z-score of 1.24 means the data value is 1.24 standard deviations above the mean. A z-score of -2.31 means the data value is 2.31 standard deviations below the mean.

Example: Z-score

In a certain city, the average 20- to 29-year old man is 69.8 inches tall, with a standard deviation of 3.1 inches, while the average 20- to 29-year old woman is 64.5 inches tall, with a standard deviation of 3.8 inches. Who is relatively taller, a 75-inch man or a 70-inch woman?

First, let's calculate the z-score of the man. We know that $\overline{x} = 69.8$ and $s = 3.1$, so:

$$\begin{aligned} z &= \frac{75 - 69.8}{3.1} \\ &= 1.68. \end{aligned}$$

Now let's calculate the z-score of the women.

$$\begin{aligned} z &= \frac{70 - 64.5}{3.8} \\ &= 1.45. \end{aligned}$$

Thus, we can conclude: The z-score for the 1.68, is larger than the z-score for the woman, 1.45, so he is relatively taller.

Interpret Percentiles

Percentiles divide a set of data, written in ascending order, into 100 parts such that 99 percentiles can be determined. For example, P_1 divides the bottom 1% divides the bottom 99%, P_2 divides the bottom 2% of the observations from the top 98%, and so on.

Percentiles are used to give the relative standing of an observation. Many standardized exams, such as the SAT, use percentiles to let students know how they scored on the exam in relation to all others who took the exam.

Interpret Quartiles:

Steps for finding Quartiles

1. Arrange the data in ascending order.
2. Determine the median, M, Determine the median, Q_2
3. Divide the data set into two halves: the observations less than M and the observations greater than M The first quartile, Q_1 is the median of the bottom half, and the third quartile, Q_3 is the median of the top half. Do not include M in these halves.

Example: Find the quartiles

Consider the data:

$$2,4,6,8,10,12,14,16,18$$

So we find Q_2 by finding the median of the entire data set:

$$\begin{aligned} Q_2 = M = \frac{n+1}{2} = \frac{10}{2} = 5 \\ = 10. \end{aligned}$$

Now we can find Q_1 and Q_3 by dividing the data set into two sets at Q_2:

$$L_1 = 2, 4, 6, 8, 10 \quad L_2 = 10, 12, 14, 16, 18.$$

Now:

$$\begin{aligned} Q_1 = M_{L_1} = \frac{n+1}{2} = \frac{6}{2} = 3 \\ = 6. \end{aligned}$$

$$\begin{aligned} Q_3 = M_{L_2} = \frac{n+1}{2} = \frac{6}{2} = 3 \\ = 14. \end{aligned}$$

Thus:

$$\begin{aligned} Q_1 = 6 \\ Q_2 = 10 \\ Q_3 = 14. \end{aligned}$$

Which means we can conclude:

$$\begin{aligned} 25\% \leqslant 6 \quad 75\% \geqslant 6 \\ 50\% \leqslant 10 \quad 50\% \geqslant 10 \\ 75\% \leqslant 14 \quad 25\% \geqslant 14. \end{aligned}$$

Example: Find the quartiles:

consider the data:

$$d_{18} = \$\ 180, \$189, \$370, \$618, \$735, \$802, \$1185, \$1414, \$1657,$$
$$\$1953, \$2332, \$2336, \$3461, \$4668, \$6751, \$9908, \$10{,}034, \$21{,}147$$

First we find $Q_2 = M$:

$$Q_2 = M = \frac{d_{\frac{n}{2}} + d_{\frac{n}{2}+1}}{2} = \frac{1657 + 1953}{2} = \$1805.$$

Now we split the data into two halves:

$$L_1 = \$180, \$189, \$370, \$618, \$735, \$802, \$1185, \$1414, \$1657$$
$$L_2 = \$1953, \$2332, \$2336, \$3461, \$4668, \$6751, \$9908, \$10,034, \$21,147.$$

Thus:

$$Q_1 = M_{L_1} = \frac{n+1}{2} = \frac{10}{2} = 5$$
$$= \$735.$$

$$Q_2 = M_{L_2} = \frac{n+1}{2} = \frac{10}{2} = 5$$
$$= \$4668.$$

Steps for finding quartiles in StatCrunch:

1. Stat > Summary Stats > Columns
2. Select Column
3. Select Median, Q_1, and Q_3

Determine and Interpret the Interquartile Range

So far we have discussed three measures of dispersion: range, standard deviation, and variance. None of these measures of dispersion are resistant . Quartiles, however, are resistant. For this reason, quartiles are used to define a resistant measure of dispersion.

The interpretation of the interquartile range is the range of the middle 50% of the data. The more spread a set of data has, the higher the interquartile range will be. The interquartile range, IQR, is a resistant measure of dispersion.

Deciding Which Measure of Central Tendency and Dispersion to Report:

Shape of Distribution	Measure of Central Tendency	Measure of Dispersion
Symmetric	Mean	Standard Deviation
Skewed Left or Skewed Right	Median	Interquartile Range

Note:-

For the remainder of this course, the phrase "describe the distribution" will mean to describe its shape (skewed left, skewed right, or symmetric), its center (mean or median), and its spread (standard deviation or interquartile range).

Resistant Measures of Central Tendency:

- **Median**

Non-Resistant Measures of Central Tendency:

- **Mean**
- **Mode**

Resistant Measures of Dispersion:

- **Quartiles**

Non-Resistant Measures of Dispersion:

- **Range.**
- **Standard Deviation.**
- **Variance.**

Check a Set of Data for Outliers

When analyzing data, we must check for extreme observations, called outliers. Outliers can occur by chance, because of errors in the measurement of a variable, during data entry, or from errors in sampling.

Outliers aren't always due to error or chance. Sometimes extreme observations are common within a population. For example, suppose we wanted to estimate the mean price of a European car. We might take a random sample of size 5 from the population of all European cars. If our sample included a Ferrari F430 Spider (approximately $175,000)), it probably would be an outlier because this car costs much more than the typical European car. The value of this car would be considered unusual because it is not a typical value from the data set.

Steps for finding outliers:

1. Determine the first and third quartiles of the data.
2. Compute the interquartile range.
3. Determine the fences. Fences serve as cutoff points for determining outliers.

$$\begin{aligned} Lower\ Fence &= Q_1 - 1.5 \cdot IQR \\ Upper\ Fence &= Q_3 + 1.5 \cdot IQR \end{aligned}$$

.

4. If a data value is less than the lower fence or greater than the upper fence, it is considered an outlier.

4.5 3.5: The Five-Number Summary and Boxplots

Learning Objectives For This Section:

1. **Determine the Five-Number Summary**
2. **Draw and Interpret Boxplots**

Vocab:

- The **five-number summary** of a set of data consists of the smallest data value, Q_1 the median, Q_3 and the largest data value. We use the five-number summary to learn information about the extremes of the data set. The summary is organized as follows:

$$\text{Minimum } Q_1 \ M \ Q_3 \text{ Maximum}$$

Determine the Five-Number Summary

Remember that the median is resistant to extreme values, so it is the preferred measure of central tendency when data are skewed right or skewed left.

The three measures of dispersion that are not resistant are the range, standard deviation, and variance. The interquartile range is resistant. However, the median, Q_1 and Q_3 do not provide information about the extremes of the data. For this, we need the smallest and largest values in the data set.

Draw and Interpret Boxplots

The five-number summary can be used to create a graph called a boxplot.

Steps for drawing boxplot:

1. Determine the lower and upper fences:
2. Draw a number line long enough to include the maximum and minimum values. Insert vertical lines at $Q_1, M, \ and \ Q_3$ Enclose these vertical lines in a box.
3. Label the lower and upper fences with a temporary mark.
4. Draw a line from Q_1 to the smallest data value that is larger than the lower fence. Draw a line from Q_3 to the largest data value that is smaller than the upper fence. These lines are called whiskers.
5. Plot any data values less than the lower fence or greater than the upper fence as outliers. Outliers are marked with an asterisk (*). Remove the temporary marks labeling the fences.

Example: Draw a boxplot

In Example 1,
we found that $Q_1 = 26.06$, $M = 30.95$, and $Q_3 = 37.24$ Therefore, the IQR $= 37.24 - 26.06 = 11.18$ minutes. From this, we find that the lower and upper fence are:

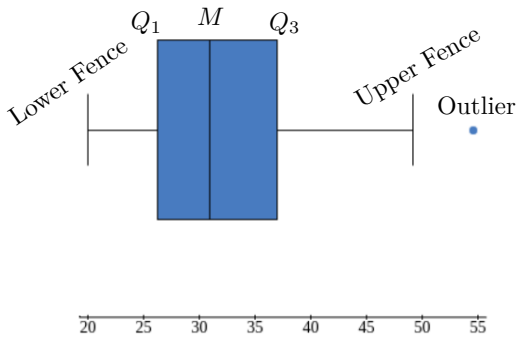
$$^{L}ower\ Fence\ = Q_1 - 1.5(IQR) = 26.06 - 1.5(11.18) = 9.29$$
$$Upper\ Fence\ = Q_3 + 1.5(IQR) = 37.24 + 1.5(11.18) = 54.01.$$

So, Draw a horizontal number line with a scale that will accommodate our graph. Draw vertical lines at $Q_1 = 26.06$, $M = 30.96$, and $Q_3 = 37.24$. Enclose these lines in a box.

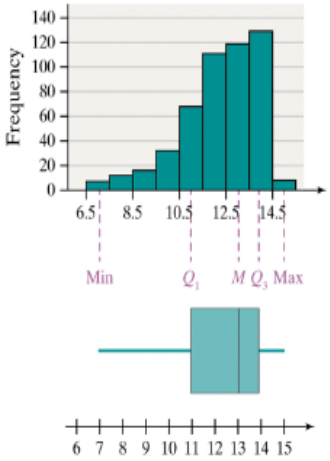
Draw a horizontal line from Q_1 to 19.95 the smallest data value that is larger than 9.29 (the lower fence). Draw a horizontal line from Q_3 to 49.17 the largest data value that is smaller than 54.01 (the upper fence).

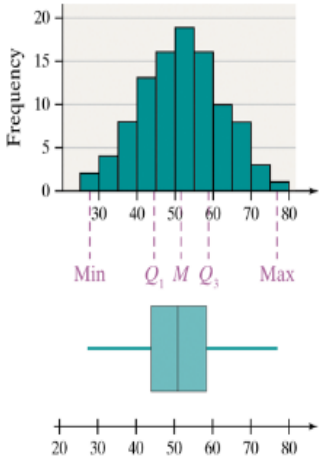
Plot any data values less than the lower fence or greater than the upper fence as outliers. Outliers are marked with an asterisk (*). Remove the temporary marks labeling the fences. See Figure 10(d).

Figure:



Skewed left	Symmetric	Skewed right
• Median is right of center in the box • Left whisker is longer than right whisker	• Median is in the center of the box • Left and right whiskers are roughly the same length	• Median is left of center in the box • Left whisker is shorter than right whisker





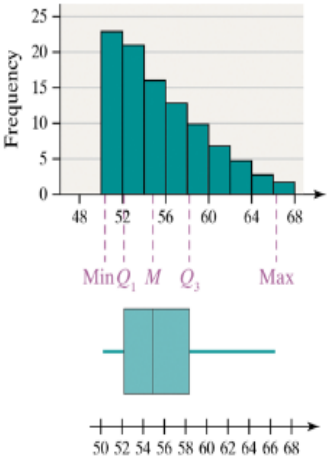


Figure 11

5 Chapter 4:

5.1 4.1: Scatter Diagrams and Correlation

Learning Objectives For This Section:

1. **Draw and Interpret Scatter Diagrams**
2. **Describe the Properties of the Linear Correlation Coefficient**
3. **Compute and Interpret the Linear Correlation Coefficient**
4. **4Determine Whether a Linear Relation Exists between Two Variables**
5. **Explain the Difference between Correlation and Causation**

Vocab:

- **bivariate data:** data in which two variables are measured on an individual. For example, we might want to know whether the amount of cola consumed per week is related to a person's bone density. The individuals would be the people in the study, and the two variables would be the amount of cola consumed weekly and bone density.
- **The response (dependent) variable** is the variable whose value can be explained by the value of the explanatory (or predictor or independent) variable.
- **A scatter diagram** is a graph that shows the relationship between two quantitative variables measured on the same individual. Each individual in the data set is represented by a point in the scatter diagram. The explanatory variable is plotted on the horizontal axis, and the response variable is plotted on the vertical axis.
- Two variables that are linearly related are **positively associated** when above-average values of one variable are associated with above-average values of the other variable (or below-average values of one variable are associated with below-average values of the other variable). That is, two variables are positively associated if, whenever the value of one variable increases, the value of the other variable also increases.
- Two variables that are linearly related are **negatively associated** when above-average values of one variable are associated with below-average values of the other variable. That is, two variables are negatively associated if, whenever the value of one variable increases, the value of the other variable decreases.
- The **linear correlation coefficient**, or Pearson product moment correlation coefficient, is a measure of the strength and direction of the linear relation between two quantitative variables. The Greek letter ρ (rho) represents the population correlation coefficient, and r represents the sample correlation coefficient. We present only the formula for the sample correlation coefficient.

$$r = \frac{\sum \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{n-1}.$$

Where:

x_i is the i th observation of the explanatory variable
$\overline{x}$ is the sample mean of the explanatory variable
$s_x$is the sample standard deviation of the explanatory variable
y_i is the i th observation of the response variable
$\overline{y}$ is the sample mean of the response variable
$\mathbf{S}_y$ is the sample standard deviation of the response variable
n is the number of individuals in the sample

Draw and Interpret Scatter Diagrams

Steps to draw scatter diagram in StatCrunch:

1. Graph > Scatter Plot
2. Select x and y columns

In a scatter diagram, the explanatory variable is plotted on the horizontal axis and the response variable is plotted on the vertical axis.

Deciding Which Variable Is the Explanatory Variable and the Response Variable

It is not always clear which variable should be considered the response variable and which the explanatory variable. For example, does high school GPA predict a student's SAT score or can the SAT score predict GPA? The researcher must determine which variable plays the role of explanatory variable based on the questions he or she wants answered. For example, if the researcher wants to predict SAT scores based on high school GPA, then high school GPA is the explanatory variable.

What does it mean to say that two variables are positively associated?

There is a linear relationship between the variables, and whenever the value of one variable increases, the value of the other variable increases.

What does it mean to say that two variables are negatively associated?

There is a linear relationship between the variables, and whenever the value of one variable increases, the value of the other variable decreases.

Describe the Properties of the Linear Correlation Coefficient

It is dangerous to use only a scatter diagram to determine if two variables are linearly related.

Just as we can manipulate the scale of graphs of univariate data, we can also manipulate the scale of graphs of bivariate data, possibly resulting in incorrect conclusions. Therefore, numerical summaries of bivariate data should be used in addition to graphs to determine any relation that exists between two variables.

Propreties of the Linear Correlation Coefficient:

- The linear correlation coefficient is always between -1 and 1, inclusive. That is, $-1 \leqslant r \leqslant 1$.
- If $r = +1$, then a perfect positive linear relation exists between the two variables.
- If $r = 1$, then a perfect negative linear relation exists between the two variables.
- The closer r is to $+1$, the stronger is the evidence of positive association between the two variables.
- The closer r is to -1, the stronger is the evidence of negative association between the two variables.
- If r is close to 0, then little or no evidence exists of a linear relation between the two variables. So a value of r close to 0 does not imply no relation, just no linear relation.
- The linear correlation coefficient is a unitless measure of association. So the unit of measure for x and y plays no role in the interpretation of r.
- The correlation coefficient is not resistant. Therefore, an observation that does not follow the overall pattern of the data could affect the value of the linear correlation coefficient.

Compute and Interpret the Linear Correlation Coefficient

Steps for finding Correlation cofficient in StatCrunch:

1. Stat > Summary Stats > Correlation
2. Select Both Columns
3. Compute!

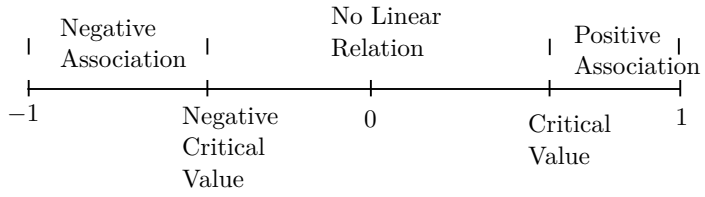
Determine Whether a Linear Relation Exists between Two Variables

How do we know the correlation between two variables is strong enough to conclude that a linear relation exists between them? Although rigorous tests can answer this question, for now, we will use a simple comparison test.

Testing for a Linear Relation:

1. Determine the absolute value of the correlation coefficient.
2. Find the critical value in Table II for the given sample size.
3. If the absolute value of the correlation coefficient is greater than the critical value, we say that a linear relation exists between the two variables. Otherwise, no linear relation exists.

Another way to think about the procedure is to consider Figure 6. If the correlation coefficient is positive and greater than the critical value, then the variables are positively associated. If the correlation coefficient is negative and less than the opposite of the critical value, then the variables are negatively associated.



Explain the Difference between Correlation and Causation:

We have discussed two types of studies: observational studies and designed experiments.

The club-head speed data examined in Examples 1, 3, and 4 are the result of an experiment. Therefore, we can claim that a faster club-head speed causes the golf ball to travel a longer distance. That is, there is a causal relationship between club-head speed and distance.

5.2 4.2: Least-Squares Regression

Learning Objectives For This Section:

1. **Find the Least-Squares Regression Line and Use the Line to Make Predictions**
2. **Interpret the Slope and the y-Intercept of the Least-Squares Regression Line**
3. **Compute the Sum of Squared Residuals**

Vocab:

- **The least-squares regression line** minimizes the sum of the squared errors (or residuals). This line minimizes the sum of the squared vertical distance between the observed values of y and those predicted by the line, $\hat{y}$ (read "y-hat"). We represent this as $\sum residuals^2$

  $$\hat{y} = b_1 x + b_0.$$

  Where:

  $$b_1 = r \cdot \frac{s_y}{s_x} \text{ is the slope of the least-squares regression line.}$$

  And:

  $$b_0 = \overline{y} - b_1 \overline{x} \text{ is the y-Intercept of the least-squares regression line.}$$

- The observed distance for this club-head speed is 274 yards. The difference between the observed and predicted values of y is the error, or **residual**.

  $$Residual \ = observed - predicted.$$

Find the Least-Squares Regression Line and Use the Line to Make Predictions

Key Ideas about the Least-Squares Regression Line:

- The least-squares regression line, $\hat{y} = b_1 x + b_0$, always contains the point $(\overline{x}, \overline{y})$.
- Because s_y and s_x must both be positive, the sign of the linear correlation coefficient, r, and the sign of the slope of the least-squares regression line, x_1, are the same.
- The predicted value of y,$\hat{y}$, is an estimate of the mean value of the response variable for any value of the explanatory variable.
- The sign of the linear correlation coefficient, r, and the sign of the slope of the least-squares regression line, b_1 , are the same.

Throughout the course, we agree to round the slope and y-intercept to four decimal places

Finding the least-squares regression line using statcrunch:

1. Stat > Regression > simple linear
2. Select x and y variables
3. Prediction of y: enter x value(s)
4. Compute

Interpret the Slope and the y-Intercept of the Least-Squares Regression Line

Interpretation of Slope

Interpreting slope for least-squares regression lines has a minor twist. Statistical models such as a least-squares regression equation are probabilistic. This means that any predictions or interpretations made as a result of the model are based on uncertainty. Therefore, when we interpret the slope of a least-squares regression equation, we do not want to imply that there is 100% certainty behind the interpretation. For example, the slope of the least-squares regression line from Example 3 is 3.1661 yards per mph. In algebra, we would interpret the slope to mean "if x increases by 1 mph, then y will increase by 3.1661 yards." In statistics, this interpretation is close but not quite accurate because increasing the club-head speed by 1 mph does not guarantee that the distance the ball will travel will increase by 3.1661 yards. Instead, for the range of data for which we have observations of the explanatory variable, an increase in club-head speed of 1 mph will increase the distance by 3.1661 yards, on average–sometimes the ball will travel a shorter additional distance, sometimes a longer additional distance, but on average, this is the change in distance. So two interpretations of slope are acceptable:

- If club-head speed increases by 1 mile per hour, the distance the golf ball will travel increases by 3.1661 yards, on average.
- If club-head speed increases by 1 mile per hour, the expected distance the golf ball will travel increases by 3.1661 yards.

Interpretation of the y-Intercept

the y-intercept of any line is the point where the graph intersects the vertical axis. In general, we interpret a y-intercept as being the value of the response variable when the value of the explanatory variable is 0. It is found by letting $x = 0$ in an equation and solving for y. To interpret the y-intercept, we must first ask two questions

- Is 0 a reasonable value for the explanatory variable?
- Do any observations near x=0 exist in the data set?

If the answer to either of those questions is no, then we do not interpret the y-intercept. In the regression equation of Example 3, a swing speed of 0 miles per hour does not make sense; so we do not interpret the y-intercept.

The second condition for interpreting the y-intercept is especially important because we should not use the regression model to make predictions outside the scope of the model, meaning that we should not use the regression model to make predictions for values of the explanatory variable that are much larger or much smaller than those observed. This is a dangerous practice because we cannot be certain of the behavior of data for which we have no observations.

Predictions When There Is No Linear Relation

When the correlation coefficient indicates no linear relation between the explanatory and response variables and the scatter diagram indicates no relation between the variables, then we use the mean value of the response variable as the predicted value so that $\hat{y} = \overline{y}$

Compute the Sum of Squared Residuals:

Recall that the least-squares regression line minimizes the sum of the squared residuals. This means that the sum of the squared residuals, $\sum residuals^2$, is smaller for the least-squares line than for any other line that may describe the relation between the two variables. In particular, the sum of the squared residuals is smaller for the least-squares regression line in Example 3 than for the line obtained in Example 1. It is worthwhile to verify this result.

5.3 4.3: Diagnostics on the Least-Squares Regression Line

Learning Objectives For This Section:

1. **Compute and Interpret the Coefficient of Determination**
2. **Perform Residual Analysis on a Regression Model**
3. **Identify Influential Observations**

Vocab:

- **The coefficient of determination**, R^2, measures the proportion of total variation in the response variable that is explained by the least-squares regression line.

$$R^2 = r^2.$$

- **An influential observation** significantly affects the least-squares regression line's slope and/or y-intercept. (It also affects the value of the correlation coefficient.) Methods exist for determining whether a particular observation is influential; however, they are beyond the scope of this course. Nonetheless, we can still get a sense as to whether a particular observation is influential right now.
- the difference in our predicted value, and our actual value, is due to factors (variables) other than the club-head speed (wind speed and position of the ball on the club face, for example) and to random error. The differences just discussed are called **deviations**.
- **Total Deviation:** The deviation between the observed value, y, and mean value, $\overline{y}$, of the response variable.

$$y - \overline{y}$$
$$Or : \text{Explained Deviation} + \text{Unexplained Deviation}.$$

- **Explained Deviation:** The deviation between the predicted value, $\hat{y}$, and mean value, $\overline{y}$, of the response variable.

$$\hat{y} - \overline{y}.$$

- **Unexplained Deviation:** The deviation between the observed value, y, and predicted value, $\hat{y}$, of the response variable

$$y - \hat{y}.$$

- If a plot of the residuals against the explanatory variable shows the spread of the residuals increasing or decreasing as the explanatory variable increases, then a strict requirement of the linear model is violated. This requirement is called **constant error variance**. The statistical term for constant error variance is **homoscedasticity**

Compute and Interpret the Coefficient of Determination:

The proportion of variation in the response variable that is explained by the least-squares regression line is called the coefficient of determination.

The coefficient of determination is a number between 0 and 1, inclusive. That is, $0 \leqslant R^2 \leqslant 1$. If $R^2 = 0$, the least-squares regression line has no explanatory value. If $R^2 = 1$, the least-squares regression line explains 100% of the variation in the response variable.

Perform Residual Analysis on a Regression Model

Recall that a residual is the difference between the observed value of y and the predicted value, $\hat{y}$. Residuals play an important role in determining the adequacy of a linear model. We will analyze residuals for the following purposes:

- To determine whether a linear model is appropriate to describe the relation between the explanatory and response variables
- To determine whether the variance of the residuals is constant
- To check for outliers

Is a linear model appropriate?

if a plot of the residuals against the explanatory variable shows a discernable pattern. Such as curved, then the response and explanatory variable may not be linearly related.

Is the variance of the residuals constant?

If a plot of the residuals against the explanatory variable shows the spread of the residuals increasing or decreasing as the explanatory variable increases, then a strict requirement of the linear model is violated.

This requirement is called constant error variance. The statistical term for constant error variance is **homoscedasticity**

Are there any outliers?

A plot of residuals against the explanatory variable may also reveal outliers. These values will be easy to identify because the residual will lie far from the rest of the plot.

Graphing residual plot in statcrunch:

1. Stats > Regression > simple linear
2. select x and y
3. Graphs: > residuals vs x-values

Identify Influential Observations

An influential observation significantly affects the least-squares regression line's slope and/or y-intercept. (It also affects the value of the correlation coefficient.) Methods exist for determining whether a particular observation is influential; however, they are beyond the scope of this course. Nonetheless, we can still get a sense as to whether a particular observation is influential right now.

Influence is affected by two factors: (1) the relative vertical position of the observation (residuals) and (2) the relative horizontal position of the observation (leverage). Leverage is a measure that depends on how much the observation's value of the explanatory variable differs from the mean value of the explanatory variable. Using these terms, Case 1 has low leverage and a large residual; Case 2 has high leverage and a small residual; Case 3 has high leverage and a large residual. From the previous activity, you should conclude that observations such as Case 3 (high leverage with a large residual) tend to be influential.

Deciding What To Do about Influential Observations

As with outliers, influential observations should be removed only if there is justification to do so. When an influential observation occurs in a data set and its removal is not warranted, two possible courses of action are to (1) collect more data so that additional points near the influential observation are obtained or (2) use techniques that reduce the influence of the influential observation. (These techniques are beyond the scope of this text.)

5.4 4.4: Contingency Tables and Association

Learning Objectives For This Section:

1. **Compute the Marginal Distribution of a Variable**
2. **Use the Conditional Distribution to Identify Association Among Categorical Data**
3. **Explain Simpson's Paradox**

Vocab:

- **A marginal distribution** of a variable is a frequency or relative frequency distribution of either the row or column variable in the contingency table.
- **A conditional distribution** lists the relative frequency of each category of the response variable, given a specific value of the explanatory variable in the contingency table.
- **Simpson's Paradox**, which describes a situation in which an association between two variables inverts or goes away when a third variable is introduced to the analysis.

a professor at a community college in New Mexico conducted a study to assess the effectiveness of delivering an introductory statistics course via traditional lecture base method, online delivery (no classroom instruction), and hybrid instruction (online course with weekly meetings) method s, the grades students received in each of the courses were tallied.

	Traditional	Online	Hybrid
A	36	39	24
B	52	55	66
C	57	68	90
D	46	38	41
F	46	54	31

This table is referred to as a **Contingency Table**, or **Two-Way table**, because it relates two categories of data. The **row variable** is grad, because each row in the table describes the grad received for each group. The **column variable** is delivery method. Each box inside the table is referred to as a **cell**.

Compute the Marginal Distribution of a Variable

The first step in summarizing data in a contingency table is to determine the distribution of each variable separately. To do so, we create marginal distributions.

A marginal distribution removes the effect of either the row variable or the column variable in the contingency table.

To create a marginal distribution for a variable, calculate the row and column totals for each category of the variable. The row totals represent the distribution of the row variable. The column totals represent the distribution of the column variable.

Create Contingency table in StatCrunch

1. Stats > Tables > Contingency (with summary)
2. Select all columns that have data
3. Select Row Label column
4. Compute

Create Relative Frequency Contingency table in StatCrunch

1. Stats > Tables > Contingency (with summary)
2. Select all columns that have data
3. Select Row Label column
4. Select Row Percent and Column Percent
5. Compute

Use the Conditional Distribution to Identify Association Among Categorical Data

As we look at the information in Table 9 and Table 10, we might ask whether a higher level of education is associated with a higher likelihood of being employed.

If level of education does not play a role in employment status, we would expect the relative frequencies for employment status at each level of education to be close to the relative frequency marginal distribution for employment status given in blue in Table 10. So we would expect 61.7% of individuals who did not finish high school, 61.7% of individuals who finished high school, 61.7% of individuals with some college, and 61.7% of individuals with at least a Bachelor's degree to be employed. If the relative frequencies for these various levels of education are different, we might associate this difference with the level of education.

The marginal distributions in Tables 9 and 10 allow us to see the distribution of either the row variable (Employment Status) or the column variable (Level of Education), but we do not get a sense of association between employment status and level of education from these tables.

To learn about any association that may exist, we need a different table.

Constructing a Conditional Distribution:

1. Stats > Tables > Contingency (with summary)
2. Select all columns that have data
3. Select Row Label column
4. Select and Column Percent
5. Compute

Drawing a Bar Graph of a Conditional Distribution:

1. Chart > Columns
2. Select all columns that have data
3. Select Row Label column
4. Plot: Vertical Bar Split

Explain Simpson's Paradox

At this point, we know how a lurking variable can cause two quantitative variables to be correlated even though they are unrelated. This same phenomenon exists when we explore the relation between two qualitative variables.