Statistics Notes

St. Joseph's University

Asher Roberts

Contents

Chapter 1

Getting Started

1.1 What is Statistics?

Definition 1.1.1. Statistics is the study of how to collect, organize, analyze, and interpret numerical information from data.

Definition 1.1.2. Individuals are the people or objects included in the study. A variable is a characteristic of the individual to be measured or observed.

Definition 1.1.3. A quantitative variable has a value or numerical measurement for which operations such as addition or averaging make sense. A qualitative variable describes an individual by placing the individual into a category or group, such as left-handed or right-handed.

Definition 1.1.4. In population data, the data are from *every* individual of interest.
In sample data, the data are from *only some* of the individuals of interest.

Definition 1.1.5. A population parameter is a numerical measure that describes an aspect of a population.
A sample statistic is a numerical measure that describes an aspect of a sample.

Example 1. The Hawaii Department of Tropical Agriculture is conducting a study of ready-to-harvest pineapples in an experimental field.

(a) Assume the researchers are interested in the individual weights of pineapples in the field. What are the objects (individuals) and variable of the study?

(b) Suppose the researchers also want data on taste. What is the variable of the study?

Example 2. How important is music education in school (K–12)? The Harris Poll did an online survey of 2286 adults (aged 18 and older) within the United States. Among the many questions, the survey asked if the respondents agreed or disagreed with the statement, "Learning and habits from music education equip people to be better team players in their careers." In the most recent survey, 71% of the study participants agreed with the statement.

(a) Identify the individuals of the study and the variable.

(b) Do the data comprise a sample? If so, what is the underlying population?

(c) Is the variable qualitative or quantitative?

(d) Identify a quantitative variable that might be of interest.

(e) Is the proportion of respondents in the sample who agree with the statement regarding music education and effect on careers a statistic or a parameter?

Definition 1.1.6 (Levels of Measurement)**.**
The nominal level of measurement applies to data that consist of names, labels, or categories. There are no implied criteria by which the data can be ordered from smallest to largest.
The ordinal level of measurement applies to data that can be arranged in order. However, differences between data values either cannot be determined or are meaningless.
The interval level of measurement applies to data that can be arranged in order. In addition, differences between data values are meaningful.
The ratio level of measurement applies to data that can be arranged in order. In addition, both differences between data values and ratios of data values are meaningful. Data at the ratio level have a true zero.

Example 3. Identify the type of data.

(a) Taos, Acoma, Zuni, and Cochiti are the names of four Native American pueblos from the population of names of all Native American pueblos in Arizona and New Mexico.

(b) In a high school graduating class of 319 students, Tatum ranked 25th, Nia ranked 19th, Elias ranked 10th, and Imani ranked 4th, where 1 is the highest rank.

(c) Body temperatures (in degrees Celsius) of trout in the Yellowstone River.

(d) Length of trout swimming in the Yellowstone River.

Example 4. The following describe different data associated with a state senator. For each data entry, indicate the corresponding *level of measurement*.

(a) The senator's name is Hollis Wilson.

(b) The senator is 58 years old.

(c) The years in which the senator was elected to the Senate are 2000, 2006, and 2012.

(d) The senator's total taxable income last year was $878,314.

(e) The senator surveyed her constituents regarding her proposed water protection bill. The choices for response were strong support, support, neutral, against, or strongly against.

(f) The senator's marital status is "married."

(g) A leading news magazine claims the senator is ranked seventh for her voting record on bills regarding public education.

Definition 1.1.7. Descriptive statistics involves methods of organizing, picturing, and summarizing information from samples or populations.
Inferential statistics involves methods of using information from a sample to draw conclusions regarding the population.

1.2 Random Samples

Definition 1.2.1. A simple random sample of n measurements from a population is a subset of the population selected in such a manner that every sample of size n from the population has an equal chance of being selected.

Example 1. Is open space around metropolitan areas important? Players of the Colorado Lottery might think so, since some of the proceeds of the game go to fund open space and outdoor recreational space. To play the game, you pay \$1 and choose any six different numbers from the group of numbers 1 through 42. If your group of six numbers matches the winning group of six numbers selected by simple random sampling, then you are a winner of a grand prize of at least \$1.5 million.

(a) Is the number 25 as likely to be selected in the winning group of six numbers as the number 5?

(b) Could all the winning numbers be even?

(c) Your friend always plays the numbers

1 2 3 4 5 6

Could they ever win?

Example 2. Use a random-number table to pick a random sample of 30 cars from a population of 500 cars.

Definition 1.2.2. A simulation is a numerical facsimile or representation of a real-world phenomenon.

Example 3. Use a random-number table to simulate the outcomes of tossing a balanced (that is, fair) penny 10 times.

(a) How many outcomes are possible when you toss a coin once?

(b) There are several ways to assign numbers to the two outcomes. Because we assume a fair coin, we can assign an even digit to the outcome "heads" and an odd digit to the outcome "tails." Then, starting at block 3 of row 2 of Table 1 in Appendix II, list the first 10 single digits.

(c) What are the outcomes associated with the 10 digits?

(d) If you start in a different block and row of Table 1 in Appendix II, will you get the same sequence of outcomes?

Definition 1.2.3 (Sampling Techniques)**.**
Random sampling: Use a simple random sample from the entire population.
Stratified sampling: Divide the entire population into distinct subgroups called strata. The strata are based on a specific characteristic such as age, income, education level, and so on. All members of a stratum share the specific characteristic. Draw random samples from each stratum.
Systematic sampling: Number all members of the population sequentially. Then, from a starting point selected at random, include every kth member of the population in the sample.
Cluster sampling: Divide the entire population into pre-existing segments or

clusters. The clusters are often geographic. Make a random selection of clusters. Include every member of each selected cluster in the sample.
Multistage sampling: Use a variety of sampling methods to create successively smaller groups at each stage. The final sample consists of clusters.
Convenience sampling: Create a sample by using data from population members that are readily available.

Definition 1.2.4. A sampling frame is a list of individuals from which a sample is actually selected.
Undercoverage results from omitting population members from the sample frame.

Definition 1.2.5. A sampling error is the difference between measurements from a sample and corresponding measurements from the respective population. It is caused by the fact that the sample does not perfectly represent the population.
A nonsampling error is the result of poor sample design, sloppy data collection, faulty measuring instruments, bias in questionnaires, and so on.

1.3 Introduction to Experimental Design

Definition 1.3.1. In a census, measurements or observations from the *entire* population are used.

Definition 1.3.2. In a sample, measurements or observations from *part* of the population are used.

Definition 1.3.3. In an observational study, observations and measurements of individuals are conducted in a way that doesn't change the response or the variable being measured.
In an experiment, a *treatment* is deliberately imposed on the individuals in order to observe a possible change in the response or variable being measured.

Example 1. In 1778, Captain James Cook landed in what we now call the Hawaiian Islands. He gave the islanders a present of several goats, and over the years these animals multiplied into wild herds totaling several thousand. They eat almost anything, including the famous silver sword plant, which was once unique to Hawaii. At one time, the silver sword grew abundantly on the island of Maui (in Haleakala, a national park on that island, the silver sword can still be found), but each year there seemed to be fewer and fewer plants. Biologists suspected that the goats were partially responsible for the decline in the number of plants and conducted a statistical study that verified their theory.

(a) To test the theory, park biologists set up stations in remote areas of Haleakala. At each station two plots of land similar in soil conditions, climate, and plant count were selected. One plot was fenced to keep out the goats, while the other was not. At regular intervals a plant count was made in each plot. What type of study was this?

(b) What part of the study was the plot that was not fenced in?

Definition 1.3.4. The placebo effect occurs when a subject receives no treatment but (incorrectly) believes they are in fact receiving treatment and responds favorably.

Definition 1.3.5. A completely randomized experiment is one in which a random process is used to assign each individual to one of the treatments.

Example 2. Can chest pain be relieved by drilling holes in the heart? For more than a decade, surgeons have been using a laser procedure to drill holes in the heart. Many patients report a lasting and dramatic decrease in angina (chest pain) symptoms. Is the relief due to the procedure, or is it a placebo effect? A recent research project at Lenox Hill Hospital in New York City provided some information about this issue by using a completely randomized experiment. The laser treatment was applied through a less invasive (catheter laser) process. A group of 298 volunteers with severe, untreatable chest pain were randomly assigned to get the laser or not. The patients were sedated but awake. They could hear the doctors discuss the laser process. Each patient thought they were receiving the treatment.
Draw a flow chart that represents this experimental design.

Definition 1.3.6. A block is a group of individuals sharing some common features that might affect the treatment.
In a randomized block experiment, individuals are first sorted into blocks, and then a random process is used to assign each individual in the block to one of the treatments.

Definition 1.3.7. In good experimental design, there is a control group. This group receives a dummy treatment, enabling the researchers to control for the placebo effect. In general, a control group is used to account for the influence of other known or unknown variables that might be an underlying cause of a change in response in the experimental group. Such variables are called lurking or confounding variables.
Randomization is used to assign individuals to the two treatment groups. This helps prevent bias in selecting members for each group.
Replication of the experiment on many patients reduces the possibility that the differences in pain relief for the two groups occurred by chance alone.

Example 3. Which technique for gathering data (sampling, experiment, simulation, or census) do you think might be the most appropriate for the following studies?

(a) Study of the effect of stopping the cooling process of a nuclear reactor.

(b) Study of the amount of time college students taking a full course load spend exercising each week in the student recreation center.

(c) Study of the effect on bone mass of a calcium supplement given to young female participants.

(d) Study of the credit hour load of *each* student enrolled at your college at the end of the drop/add period this semester.

Definition 1.3.8 (Some Potential Pitfalls of a Survey)**.**
Nonresponse: Individuals either cannot be contacted or refuse to participate. Nonresponse can result in significant undercoverage of a population.
Truthfulness of response: Respondents may lie intentionally or inadvertently.
Faulty recall: Respondents may not accurately remember when or whether an event took place.
Hidden bias: The question may be worded in such a way as to elicit a specific response. The order of questions might lead to biased responses. Also, the number of responses on a Likert scale may force responses that do not reflect the respondent's feelings or experience.
Vague wording: Words such as "often," "seldom," and "occasionally" mean different things to different people.
Interviewer influence: Factors such as tone of voice, body language, dress, gender, authority, and ethnicity of the interviewer might influence responses.
Voluntary response: Individuals with strong feelings about a subject are more likely than others to respond. Such a study is interesting but not reflective of the population.

Definition 1.3.9. A lurking variable is one for which no data have been collected but that nevertheless has influence on other variables in the study.

Two variables are confounded when the effects of one cannot be distinguished from the effects of the other. Confounding variables may be part of the study, or they may be outside lurking variables.

Example 4. Comment on the usefulness of the data collected as described.

(a) A uniformed police officer interviews a group of 20 college freshmen. The officer asks each one their name and then if they have used an illegal drug in the last month.

(b) Frankie saw some data that show that cities with more low-income housing have more people who are homeless. Does building low-income housing cause homelessness?

(c) A survey about food in the student cafeteria was conducted by having forms available for customers to pick up at the cash register. A drop box for completed forms was available outside the cafeteria.

(d) Extensive studies on coronary problems were conducted using male participants over age 50 as the subjects.

Chapter 2

Organizing Data

2.1 Frequency Distributions, Histograms, and Related Topics

Definition 2.1.1. A frequency table partitions data into classes or intervals of equal width and shows how many data values are in each class. The classes or intervals are constructed so that each data value falls into exactly one class.

Example 1. A task force to encourage car pooling conducted a study of one-way commuting distances of workers in the downtown Dallas area. A random sample of 60 of these workers was taken. The commuting distances of the workers in the sample are given in Table 2.1. Make a frequency table for these data.

Table 2.1: One-Way Commuting Distances (in Miles) for 60 Workers in Downtown Dallas

13	47	10	3	16	20	17	40	4	2
7	25	8	21	19	15	3	17	14	6
12	45	1	8	4	16	11	18	23	12
6	2	14	13	7	15	46	12	9	18
34	13	41	28	36	17	24	27	29	9
14	26	10	24	37	31	8	16	12	16

Definition 2.1.2. The lower class limit is the lowest data value that can fit in a class. The upper class limit is the highest data value that can fit in a class. The class width is the difference between the lower class limit of one class and the lower class limit of the next class.

Example 2. Make a histogram and a relative-frequency histogram with six bars for the data in Table 2.1 showing one-way commuting distances.

Example 3. An irate customer called an online website's customer services 40 times during the last two weeks to see why their order had not arrived. Each time they called, the customer recorded the length of time they were put "on hold" before being allowed to talk to a customer service representative. See Table 2.2.

Table 2.2: Length of Time on Hold, in Minutes

1	5	5	6	7	4	8	7	6	5
5	6	7	6	6	5	8	9	9	10
7	8	11	2	4	6	5	12	13	6
3	7	8	8	9	9	10	9	8	9

(a) What are the largest and smallest values in Table 2.2? If we want five classes in a frequency table, what should the class width be?

(b) Complete the table, Table 2.3.

Table 2.3: Time on Hold

Class Limits Lower-Upper	Tally	Frequency	Midpoint
1–3			
4–			
–9			
–			
–			

(c) Recall that the class boundary is halfway between the upper limit of one class and the lower limit of the next. Use this fact to find the class boundaries in Table 2.4 and to make a histogram.

Table 2.4: Class Boundaries

Class Limits	Class Boundaries
1–3	0.5–3.5
4–6	3.5–6.5
7–9	6.5–
10–12	–
13–15	–

(d) Compute the relative class frequency f/n for each class in Table 2.5 and make a relative-frequency histogram.

Table 2.5: Relative Class Frequency

Class	f/n
1–3	$3/40 = 0.075$
4–6	$15/40 = 0.375$
7–9	
10–12	
13–15	

(e) According to the graphs, about what percentage of the hold times are between 7 and 9 minutes? Would it be unusual for the hold times to be less than 4 minutes or greater than 12?

Definition 2.1.3 (Distribution Shapes)**.**
Mound-shaped symmetric: This term refers to a mound-shaped histogram in which both sides are (more or less) the same when the graph is folded vertically down the middle.
Uniform or rectangular: These terms refer to a histogram in which every class has equal frequency. From one point of view, a uniform distribution is symmetric with the added property that the bars are of the same height.

Skewed left or skewed right: These terms refer to a histogram in which one tail is stretched out longer than the other. The direction of skewness is on the side of the longer tail where outliers of the data would be present. So, if the longer tail is on the left, we say the histogram is skewed to the left.
Bimodal: This term refers to a histogram in which the two classes with the largest frequencies are separated by at least one class. The top two frequencies of these classes may have slightly different values. This type of situation sometimes indicates that we are sampling from two different populations.

Definition 2.1.4. Outliers in a data set are the data values that are very different from other measurements in the data set.

Definition 2.1.5. The cumulative frequency for a class is the sum of the frequencies for that class and all previous classes.

Example 4. Aspen, Colorado, is a world-famous ski area. If the daily high temperature is above 40°F, the surface of the snow tends to melt. It then freezes again at night. This can result in a snow crust that is icy. It also can increase avalanche danger.
Table 2.6 gives a summary of daily high temperatures (°F) in Aspen during the 151-day ski season.

Table 2.6: High Temperatures During the Peak Aspen Ski Season (°F)

Class Limits		Frequency	Cumulative Frequency
Lower	Upper		
10.5	20.5	23	
20.5	30.5	43	
30.5	40.5	51	
40.5	50.5	27	
50.5	60.5	7	

(a) Complete Table 2.6 by filling in the cumulative frequencies.

(b) Draw the corresponding ogive.

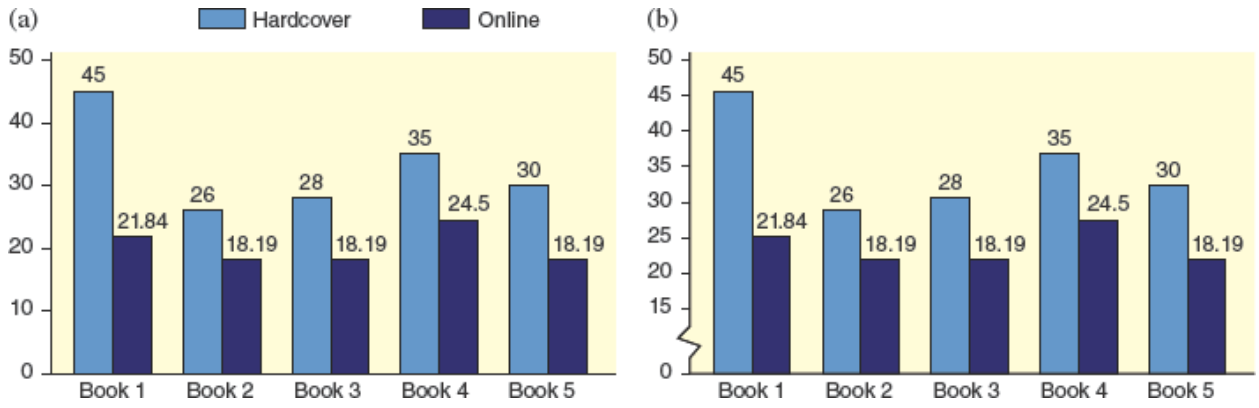
(c) Estimate the total number of days with a high temperature lower than or equal to 40°F.

(d) Find the cumulative proportion of days with a high temperature lower than or equal to 40°F.

2.2 Bar Graphs, Circle Graphs, and Time-Series Graphs

Example 1. Figure 2.1 shows two bar graphs depicting the cost of hardcover books and their digital editions for the top five bestsellers from an online shopping website. Discuss the features of these graphs.

Figure 2.1: Cost of Books



Definition 2.2.1. A Pareto chart is a bar graph in which the bar height represents frequency of an event. In addition, the bars are arranged from left to right according to decreasing height.

Example 2. This exercise is adapted from *The Deming Management Method* by Mary Walton. Suppose you want to arrive at college 15 minutes before your first class so that you can feel relaxed when you walk into class. An early arrival time also allows room for unexpected delays. However, you always find yourself arriving "just in time" or slightly late. What causes you to be late? Charlotte made a list of possible causes and then kept a checklist for 2 months (Table 2.7). On some days more than one item was checked because several events occurred that caused her to be late.

Table 2.7: Causes for Lateness (September–October)

Cause	Frequency
Snoozing after alarm goes off	15
Car trouble	5
Too long over breakfast	13
Last-minute studying	20
Finding something to wear	8
Talking too long with roommate	9
Other	3

(a) Make a Pareto chart using the information provided. Be sure to label the causes and draw the bars using the same vertical scale.

(b) Compare the Pareto chart to a bar graph of the data. What are the advantages of having the data presented as a Pareto chart?

(c) Looking at the Pareto chart, what recommendations do you have for Charlotte?

Definition 2.2.2. In a circle graph or pie chart, wedges of a circle visually display proportional parts of the total population that share a common characteristic.

Example 3. How much does social media impact the daily lives of teens? The results are taken from a 2018 survey of 736 teens (as reported in Pew Research Center) and shown in Table 2.8. We'll make a circle graph to display these data.

Table 2.8: Impact of Social Media on the Lives of Teens

Time	Number	Fractional Part	Percentage
A **mostly positive** effect	225	225/736	30.6
A **mostly negative** effect	179	179/736	24.3
Neither positive nor negative effect	332		

(a) Fill in the missing parts of the table. Do the percentages total 100% (within rounding error)?

(b) Draw a circle graph. Divide the circle into pieces. Label each piece, and show the percentage corresponding to each piece.

(c) Based on the information shown on the circle graph, what would you say would be the effects of social media on teens?

Definition 2.2.3. In a time-series graph, data are plotted in order of occurrence at regular intervals over a period of time.

Example 4. Suppose you have been in the walking/jogging exercise program for 20 weeks, and for each week you have recorded the distance you covered in 30 minutes. Your data log is shown in Table 2.9.

Table 2.9: Distance (in Miles) Walked/Jogged in 30 Minutes

Week	1	2	3	4	5	6	7	8	9	10
Distance	1.5	1.4	1.7	1.6	1.9	2.0	1.8	2.0	1.9	2.0
Week	11	12	13	14	15	16	17	18	19	20
Distance	2.1	2.1	2.3	2.3	2.2	2.4	2.5	2.6	2.4	2.7

(a) Make a time-series graph.

(b) From looking at your graph, can you detect any patterns?

Definition 2.2.4. Time-series data consist of measurements of the same variable for the same subject taken at regular intervals over a period of time.

2.3 Stem-and-Leaf Displays

Definition 2.3.1. A stem-and-leaf display is a method of exploratory data analysis that is used to rank order and arrange data into groups.

Example 1. Many airline passengers seem weighted down by their carry-on luggage. Just how much weight are they carrying? The carry-on luggage weights in pounds for a random sample of 40 passengers returning from a vacation to Hawaii were recorded (see Table 2.10).

Table 2.10: Weights of Carry-On Luggage in Pounds

30	27	12	42	35	47	38	36	27	35
22	17	29	3	21	0	38	32	41	33
26	45	18	43	18	32	31	32	19	21
33	31	28	29	51	12	32	18	21	26

Make a stem-and-leaf display with the data.

Example 2. What does it take to win at sports? If you're talking about basketball, one sportswriter gave the answer. He listed the winning scores of the conference championship games over the last 35 years. The scores for those games follow.

132	118	124	109	104	101	125	83	99
131	98	125	97	106	112	92	120	103
111	117	135	143	112	112	116	106	117
119	110	105	128	112	126	105	102	

To make a stem-and-leaf display, we will use the first *two digits* as the stems since it would not be feasible to use a *single digit* stem to represent values in the data set above 99.

(a) Use the first *two* digits as the stem. Then order the leaves. Provide a label that shows the meaning and units of the first stem and first leaf.

(b) Based on the stem-and-leaf display, what shape do you think is the distribution of the data? Explain.

Chapter 3

Averages and Variation

3.1 Measures of Central Tendency: Mode, Median, and Mean

Definition 3.1.1. The mode of a data set is the value that occurs most frequently. Note: If a data set has no single value that occurs more frequently than any other, then that data set has no mode.

Example 1. Count the letters in each word of this sentence and give the mode.

Definition 3.1.2. The median is the central value of an ordered distribution.

Example 2. The local School of Nursing held a panel featuring recent graduates as part of an Admissions event for prospective new students. The graduates on the panel shared their starting salaries (in thousands of dollars per year):

60.6 71.1 59.9 35.1 64.8 71.1

(a) Find the median salary of the graduates.

(b) Five of the graduates on the panel started working as registered nurses. The low salary reported was from a graduate who went to work for a non-profit and was not working as a nurse. Find the median starting salaries of registered nurses on the panel.

Example 3. Belleview College must make a report to the budget committee about the average credit hour load a full-time student carries. The budget allocated to the college will depend on the value they report. (A 12-credit-hour load is the minimum requirement for full-time status. For the same tuition, students may take up to 20 credit hours.) A random sample of 40 students yielded the following information (in credit hours):

17	12	14	17	13	16	18	20	13	12
12	17	16	15	14	12	12	13	17	14
15	12	15	16	12	18	20	19	12	15
18	14	16	17	15	19	12	13	12	15

(a) Organize the data from smallest to largest number of credit hours.

(b) Since there are an (odd, even) number of values, we add the two middle values and divide by 2 to get the median. What is the median credit hour load?

(c) What is the mode of this distribution? Is it different from the median?

(d) If the budget committee is going to fund the college according to the average student credit hour load (more money for higher loads), which of these two averages do you think the college will report?

Remark 1. For an ordered data set of size n,

$$\text{Position of the middle value} = \frac{n+1}{2}.$$

Definition 3.1.3. An average that uses the exact value of each entry is the mean (sometimes called the arithmetic mean). To compute the mean, we add the values of all the entries and then divide by the number of entries.

Example 4. To graduate, Linda needs at least a B in biology. She did not do very well on her first three tests; however, she did well on the last four. Here are her scores:

58 67 60 84 93 98 100

Compute the mean and determine if Linda's grade will be a B (80 to 89 average) or a C (70 to 79 average).

Remark 2. When we compute the mean, we sum the given data. There is a convenient notation to indicate the sum. Let x represent any value in the data set. Then the notation Σx represents the sum of all given data values.
In other words, we are to sum all the entries in the distribution. The *summation symbol* Σ means *sum the following* and is denoted by capital sigma, the S of the Greek alphabet.
The symbol for the mean of a *sample* distribution of x values is denoted by $\bar{x}$ (read "x bar"). If your data comprise the entire *population*, we use the symbol μ (lowercase Greek letter mu, pronounced "mew") to represent the mean.

Example 5. *Barron's Profiles of American Colleges*, 19th edition, lists average class size for introductory lecture courses at each of the profiled institutions. A sample of 20 colleges and universities in California showed class sizes for introductory lecture courses to be

14 20 20 20 20 23 25 30 30 30
35 35 35 40 40 42 50 50 80 80

(a) Compute a 5% trimmed mean for the sample.

(b) Find the median of the original data set.

(c) Find the median of the 5% trimmed data set. Does the median change when you trim the data?

Remark 3. Sometimes we wish to average numbers, but we want to assign more importance, or weight, to some of the numbers. In this case the weighted average is given by

$$\text{Weighted average} = \frac{\sum xw}{\sum w}$$

where x is a data value and w is the weight assigned to that data value. The sum is taken over all data values.

Example 6. Suppose that we have the following mean monthly rents for five different apartment complexes with the given number of units.

Number of Units	Ave. Rent
200	\$820
25	\$940
35	\$1030
100	\$1190
15	\$920

Find the average monthly rent for all 375 apartments in these complexes.

3.2 Measures of Variation

Definition 3.2.1. The range is the difference between the largest and smallest values of a data distribution.

Example 1. A large bakery regularly orders cartons of Maine blueberries. The average weight of the cartons is supposed to be 22 ounces. Random samples of cartons from two suppliers were weighed. The weights in ounces of the cartons were

Supplier I:	17	22	22	22	27
Supplier II:	17	19	20	27	27

(a) Compute the range of carton weights from each supplier.

(b) Compute the mean weight of cartons from each supplier.

(c) Look at the two samples again. The samples have the same range and mean. How do they differ? The bakery uses one carton of blueberries in each blueberry muffin recipe. It is important that the cartons be of consistent weight so that the muffins turn out right.

Definition 3.2.2 (Defining Formulas (Sample Statistic))**.**

$$\text{Sample variance} = s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

$$\text{Sample standard deviation} = s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

where x is a member of the data set, $\bar{x}$ is the mean, and n is the number of data values. The sum is taken over all data values.

Definition 3.2.3 (Computation Formulas (Sample Statistic))**.**

$$\text{Sample variance} = s^2 = \frac{\sum x^2 - (\sum x)^2/n}{n-1}$$

$$\text{Sample standard deviation} = s = \sqrt{\frac{\sum x^2 - (\sum x)^2/n}{n-1}}$$

where x is a member of the data set and n is the number of data values. The sum is taken over all data values.

Example 2. Big Blossom Greenhouse was commissioned to develop an extra large rose for the Rose Bowl Parade. A random sample of blossoms from Hybrid A bushes yielded the following diameters (in inches) for mature peak blooms.

2 3 3 8 10 10

Use the defining formula to find the sample variance and standard deviation.

Example 3. Big Blossom Greenhouse gathered another random sample of mature peak blooms from Hybrid B. The six blossoms had the following diameters (in inches):

5 5 5 6 7 8

(a) Construct a table so that we can find the mean, variance, and standard deviation more easily. In this case, what is the value of n? Find Σx and compute the mean.

(b) What is the value of $n - 1$? Use the computation formula to find the sample variance s^2.

(c) Use a calculator to find the square root of the variance. Is this the standard deviation?

Definition 3.2.4 (Population Parameters)**.**

$$\text{Population mean} = \mu = \frac{\sum x}{N}$$

$$\text{Population variance} = \sigma^2 = \frac{\sum (x-\mu)^2}{N}$$

$$\text{Population standard deviation} = \sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

where N is the number of data values in the population and x represents the individual data values of the population.

Remark 1. To estimate the mean from a frequency table or histogram, we can compute a weighted average using the midpoints of each class as the data values and the frequencies as the weights. If n is the total number of data values represented, then $\Sigma f = n$, and

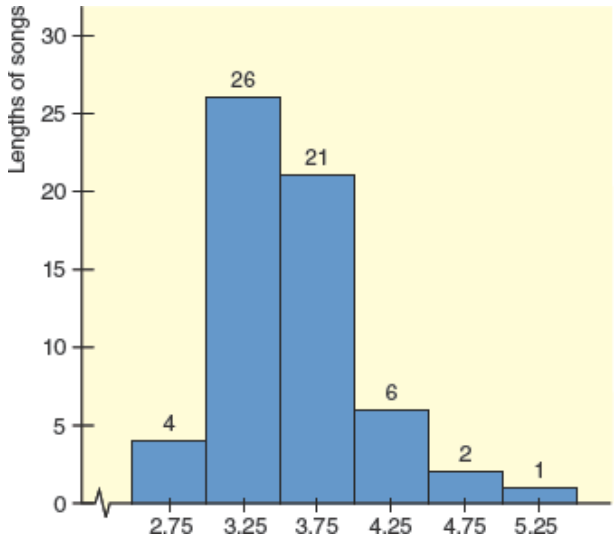
$$\bar{x} \approx \frac{\sum xf}{n}.$$

To estimate the sample standard deviation for grouped data, the corresponding formula is

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

Example 4. The faculty advisor of the local college radio station is interested in analyzing the lengths of songs played. The station manager produced the following histogram (Figure 3.1) showing the lengths of songs played during a 4-hour block of programming.

Figure 3.1: Lengths of Songs



(a) Estimate the mean length of songs played in this 4-hour block.

(b) Estimate the standard deviation in the length of songs played.

Definition 3.2.5. If $\bar{x}$ and s represent the sample mean and sample standard deviation, respectively, then the sample coefficient of variation CV is defined to be

$$CV = \frac{s}{\bar{x}} \cdot 100\%.$$

If μ and σ represent the population mean and population standard deviation, respectively, then the population coefficient of variation CV is defined to be

$$CV = \frac{\sigma}{\mu} \cdot 100\%.$$

Example 5. Javier runs a small aviary in Florida that is home to eight adult scarlet macaws, rescued from pet owners that could no longer take care of them. He is interested in how the size of birds in his small population compares to the population of scarlet macaws in the wild. The tip-to-tail lengths of the birds in his population (in inches) were found to be

36.3 38.6 35.4 34.4 29.6 33.3 31.7 34.5

(a) Use a calculator with appropriate statistics keys to verify that, for Javier's macaw population, $\mu = 34.2$ inches and $\sigma = 2.59$ inches.

(b) Compute the CV of Javier's length data and comment on the meaning of the result.

Example 6. Javier found data from a sample of wild scarlet macaws from a region of the Amazon forest in northwest Peru. The lengths of birds in the sample (measured in centimeters) was

80.0 74.4 75.9 82.6 78.7 96.6 95.1 83.1 58.5 74.2

(a) Use a calculator with sample mean and sample standard deviation keys to compute $\bar{x}$ and σ.

(b) Compute the CV for the lengths of macaws in this sample.

(c) Compare the mean, standard deviation, and CV for the population of macaws at Javier's aviary in the previous example and the sample of macaws from Peru.

Theorem 3.2.1. *For any set of data (either population or sample) and for any constant k greater than 1, the proportion of the data that must lie within k standard deviations on either side of the mean is at least*

$$1 - \frac{1}{k^2}$$

Put another way: For sample data with mean $\bar{x}$ and standard deviation s, at least $1 - 1/k^2$ (fractional part) of data must fall between $\bar{x} - ks$ and $\bar{x} + ks$.

Remark 2. For any set of data:

- at least 75% of the data fall in the interval from $\mu - 2\sigma$ to $\mu + 2\sigma$.
- at least 88.9% of the data fall in the interval from $\mu - 3\sigma$ to $\mu + 3\sigma$.
- at least 93.8% of the data fall in the interval from $\mu - 4\sigma$ to $\mu + 4\sigma$.

Example 7. Students Who Care is a student volunteer program in which college students donate work time to various community projects such as planting trees. Professor Gill is the faculty sponsor for this student volunteer program. For several years, Dr. Gill has kept a careful record of x = total number of work hours volunteered by a student in the program each semester. For a random sample of students in the program, the mean number of hours was $\bar{x} = 29.1$ hours each semester, with a standard deviation of $s = 1.7$ hours each semester. Find an interval A to B for the number of hours volunteered into which at least 75% of the students in this program would fit.

Example 8. Free games on smart phones are often supported by ads. Developers and advertisers track the number of times users interact with ads in order to determine the fees that advertisers pay the developers. A popular game measured an average of $\bar{x} = 1123$ interactions per day of running an ad with a standard deviation of $s = 55$. Determine a Chebyshev interval about the mean in which at least 88.9% of the data fall.

3.3 Percentiles and Box-and-Whisker Plots

Definition 3.3.1. For whole numbers P (where $1 \leq P \leq 99$), the Pth percentile of a distribution is a value such that $P\%$ of the data fall at or below it and $(100 - P)\%$ of the data fall at or above it.

Example 1. You took the English achievement test to obtain college credit in freshman English by examination.

(a) If your score is at the 89th percentile, what percentage of scores are at or below yours?

(b) If the scores ranged from 1 to 100 and your raw score is 95, does this necessarily mean that your score is at the 95th percentile?

Definition 3.3.2. Quartiles are those percentiles that divide the data into fourths. The first quartile Q_1 is the 25th percentile, the second quartile Q_2 is the median, and the third quartile Q_3 is the 75th percentile.
The interquartile range (IQR) is the difference between the third and first quartiles.

$$\text{Interquartile range} = Q_3 - Q - 1$$

Example 2. In a hurry? On the run? Hungry as well? How about an ice cream bar as a snack? Ice cream bars are popular among all age groups. Consumer Reports did a study of ice cream bars. Twenty-seven bars with taste ratings of at least "fair" were listed, and cost per bar was included in the report. Just how much does an ice cream bar cost? The inflation adjusted data, expressed in dollars, appear in Table 3.1. As you can see, the cost varies quite a bit, partly because the bars are not of uniform size.

Table 3.1: Ordered Cost of Ice Cream Bars (in Dollars)

0.46	0.47	0.48	0.50	0.55	0.63	0.63	0.65
0.67	0.68	0.70	0.77	0.80	0.80	0.80	0.93
1.14	1.27	1.27	1.29	1.30	1.33	1.37	1.37
1.37	1.38	1.53					

(a) Find the quartiles.

(b) Find the interquartile range.

Example 3. Many people consider the number of calories in an ice cream bar as important as, if not more important than, the cost. The Consumer Reports article also included the calorie count of the rated ice cream bars (Table 3-5). There were 22 vanilla-flavored bars rated. Again, the bars varied in size, and some of the smaller bars had fewer calories. The calorie counts for the vanilla bars are in Table 3.2.

Table 3.2: Calories in Vanilla-Flavored Ice Cream Bars

342	377	319	353	295
234	294	286	377	182
310	439	111	201	182
197	209	147	190	151
131	151			

(a) Our first step is to order the data. Do so.

(b) There are 22 data values. Find the median.

(c) How many values are below the median position? Find Q_1.

(d) There are the same number of data above as below the median. Use this fact to find Q_3.

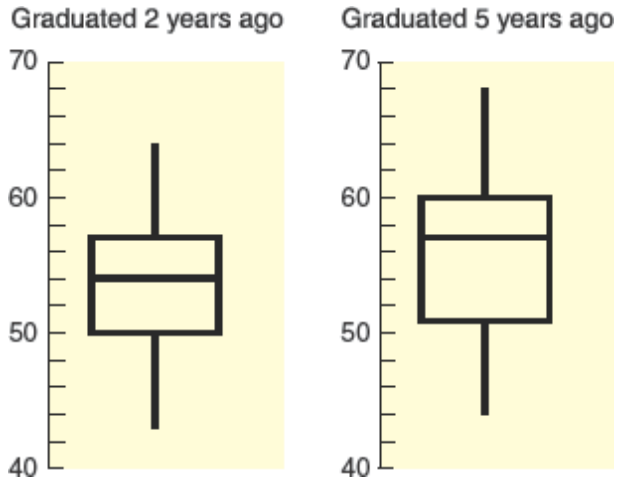
(e) Find the interquartile range and comment on its meaning.

Definition 3.3.3 (Five Number Summary)**.** The five number summary is the Lowest value, Q_1, median, Q_3, and highest value.

Example 4. Using the data from the previous example, make a box-and-whisker plot showing the calories in vanilla-flavored ice cream bars. Use the plot to make observations about the distribution of calories.

Example 5. The Renata College Development Office sent salary surveys to alumni who graduated 2 and 5 years ago. The voluntary responses received are summarized in the box-and-whisker plots shown in Figure 3.2.

Figure 3.2: Box-and-Whisker Plots for Alumni Salaries (in Thousands of Dollars)



(a) From Figure 3.2, estimate the median and extreme values of salaries of alumni graduating 2 years ago. In what range are the middle half of the salaries?

(b) From Figure 3.2, estimate the median and the extreme values of salaries of alumni graduating 5 years ago. What is the location of the middle half of the salaries?

(c) Compare the two box-and-whisker plots and make comments about the salaries of alumni graduating 2 and 5 years ago.

Chapter 4

Correlation and Regression

4.1 Scatter Diagrams and Linear Correlation

Definition 4.1.1. A scatter diagram is a graph in which data pairs (x, y) are plotted as individual points on a grid with horizontal axis x and vertical axis y. We call x the explanatory variable and y the response variable.

Example 1. Phosphorous is a chemical used in many household and industrial cleaning compounds. Unfortunately, phosphorous tends to find its way into surface water, where it can kill fish, plants, and other wetland creatures. Phosphorous-reduction programs are required by law and are monitored by the Environmental Protection Agency (EPA) (Reference: *EPA Case Study 832-R-93-005*).
A random sample of eight sites in a California wetlands study gave the following information about phosphorous reduction in drainage water. In this study, x is a random variable that represents phosphorous concentration (in 100 mg/L) at the inlet of a passive biotreatment facility, and y is a random variable that represents total phosphorous concentration (in 100 mg/L) at the outlet of the passive biotreatment facility.

x	5.2	7.3	6.7	5.9	6.1	8.3	5.5	7.0
y	3.3	5.9	4.8	4.5	4.0	7.1	3.6	6.1

(a) Make a scatter diagram for these data.

(b) Comment on the relationship between x and y in your scatter diagram.

Example 2. A large industrial plant has seven divisions that do the same type of work. A safety inspector visits each division of 20 workers quarterly. The number x of work-hours devoted to safety training and the number y of work-hours lost due to industry-related accidents are recorded for each separate division in Table 4.1.

Table 4.1: Safety Report

Division	x	y
1	10.0	80
2	19.5	65
3	30.0	68
4	45.0	55
5	50.0	35
6	65.0	10
7	80.0	12

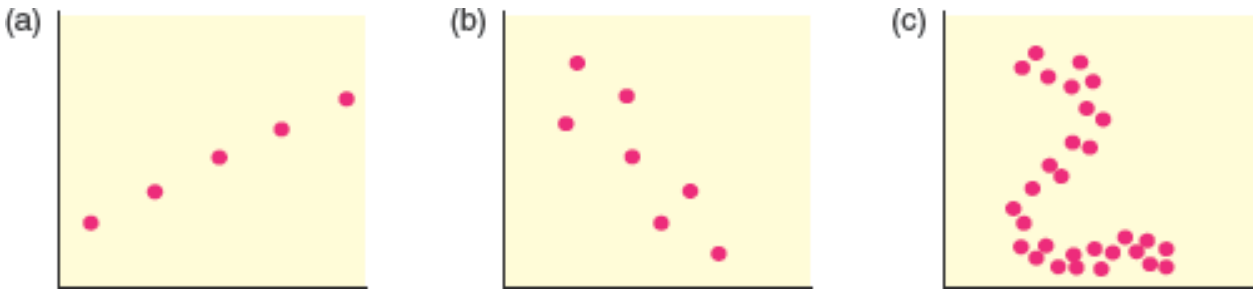
(a) Make a scatter diagram for these pairs. Place the x values on the horizontal axis and the y values on the vertical axis.

(b) Does a line fit the data reasonably well? Draw a line that you think "fits best."

(c) As the number of hours spent on safety training increases, what happens to the number of hours lost due to industry-related accidents?

Example 3. Examine the scatter diagrams in Figure 4.1 and then answer the following questions.

Figure 4.1: Scatter Diagrams



(a) Which diagram has no linear correlation?

(b) Which has perfect linear correlation?

(c) Which can be reasonably fitted by a straight line?

Definition 4.1.2. The sample correlation coefficient r is a numerical measurement that assesses the strength of a linear relationship between two variables x and y from sample data.

1. r is a unitless measurement between -1 and 1. In symbols, $-1 \leq r \leq 1$. If $r = 1$, there is perfect positive linear correlation. If $r = -1$, there is perfect negative linear correlation. If $r = 0$, there is no linear correlation. The closer r is to 1 or -1, the better a line describes the relationship between the two variables x and y.

2. Positive values of r imply that as x increases, y tends to increase. Negative values of r imply that as x increases, y tends to decrease.

3. The value of r is the same regardless of which variable is the explanatory variable and which is the response variable. In other words, the value of r is the same for the pairs (x, y) and the corresponding pairs (y, x).

4. The value of r does not change when either variable is converted to different units.

Remark 1. For a random sample of n data pairs (x, y), r can be computed using the formula

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum xy)^2}}.$$

Example 4. Sand driven by wind creates large, beautiful dunes at the Great Sand Dunes National Park, Colorado. Of course, the same natural forces also create large dunes in the Great Sahara and Arabia. Is there a linear correlation between wind velocity and sand drift rate? Let x be a random variable representing wind velocity (in 10 cm/sec) and let y be a random variable representing drift rate of sand (in 100 gm/cm/sec). A test site at the Great Sand Dunes National Park gave the following information about x and y (Reference: *Hydrologic, Geologic, and Biologic Research at Great Sand Dunes National Monument*, Proceedings of the National Park Service Research Symposium).

x	70	115	105	82	93	125	88
y	3	45	21	7	16	62	12

(a) Construct a scatter diagram. Do you expect r to be positive?

(b) Compute r using the computation formula.

(c) What does the value of r tell you?

Example 5. Parks with large grassy areas around ponds or lakes often have problems with geese. Geese feeding activity can lead to erosion, and geese feces can be a health hazard. The City Park staff has been experimenting with coyote decoys to encourage groups of geese (called gaggles) to find other places to congregate. Over several years they have tested the placements of different numbers of decoys during geese migration season, and counted the number of gaggles of geese that landed in a given day. For each day, x represents the number of coyote decoys in the park and y represents the number of gaggles of geese observed on that day. (A group of geese was counted as a gaggle if there were more than 10 geese that stayed for more than 10 minutes.)

x	10	15	16	1	4	6	18	12	14	7
y	5	2	1	9	7	8	1	5	3	6

(a) Construct a scatter diagram of x and y values.

(b) From the scatter diagram, do you think the computed value of r will be positive, negative, or zero? Explain.

(c) Compute r.

(d) What does the value of r tell you about the relationship between the number of coyote decoys and the number of unwanted groups of geese in the park?

Definition 4.1.3. In the ordered pair (x, y) we call the variable x the explanatory variable and we call y the response variable. We will say that some of the changes in the response variable *are explained by* changes in the explanatory variable, but this should *not* be interpreted as "changes in x *cause* changes in y." Even though we call y the response variable, the changes in y might or might not be in direct response to changes in x. There can always be other variables lurking in the background that are causing both x and y to change. Despite this terminology, *changes in the explanatory variable do not necessarily cause the corresponding changes in the response variable.* Thus we call variables that are neither explanatory nor response variables lurking variables.

Example 6. Over a period of years, the population of a certain town increased. It was observed that during this period the correlation between x, the number of people attending church, and y, the number of people in the city jail, was $r = 0.90$. Does going to church cause people to go to jail? Is there a lurking variable that might cause both variables x and y to increase?

4.2 Linear Regression and the Coefficient of Determination

Remark 1. The least-squares line $\hat{y} = a + bx$ is the line such that the sum of the squares of the vertical distances from the points of a scatter diagram to the line are made as small as possible. The slope and intercept of the line can be obtained using the computation formulas

$$\text{Slope: } b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2},$$
$$\text{Intercept: } a = \bar{y} - b\bar{x}.$$

Example 1. Let's find the least-squares equation relating the variables $x =$ size of caribou population (in hundreds) and $y =$ size of wolf population in Denali National Park. Use x as the explanatory variable and y as the response variable.

x	y	x^2	y^2	xy
30	66	900	4356	1980
34	79	1156	6241	2686
27	70	729	4900	1890
25	60	625	3600	1500
17	48	289	2304	816
23	55	529	3025	1265
20	60	400	3600	1200
$\sum x = 176$	$\sum y = 438$	$\sum x^2 = 4628$	$\sum y^2$ =28,026	xy =11,337

(a) Use the computation formulas to find the slope b of the least-squares line and the y intercept a.

(b) Use the values of a and b (either computed or obtained from a calculator) to find the equation of the least-squares line.

(c) Graph the equation of the least-squares line on a scatter diagram.

Definition 4.2.1. The slope of the least-squares line tells us how many units the response variable is expected to change for each unit change in the explanatory variable. The number of units change in the response variable for each unit change in the explanatory variable is called the marginal change of the response variable.

Definition 4.2.2. A data pair is influential if removing it would substantially change the equation of the least-squares line or other calculations associated with linear regression. An influential point often has an x-value near the extreme high or low value of the data set.

Definition 4.2.3. The residual is the difference between the y value in a specified data pair (x, y) and the value $\hat{y} = a + bx$ predicted by the least-squares line for the same x. In other words, the residual is

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

Definition 4.2.4. The prediction for an x value *between* observed x values in the data set is interpolation. The prediction for an x value *beyond* the observed x values in the data set is called extrapolation. Interpolation is generally good at producing reliable predictions. Extrapolation may produce unrealistic predictions.

Example 2. We continue with the previous example regarding the size of the wolf population as it relates to the size of the caribou population. Suppose you want to predict the size of the wolf population when the size of the caribou population is 21 (hundred).

(a) In the least-squares model developed in the previous example, which is the explanatory variable and which is the response variable? Can you use the equation to predict the size of the wolf population for a specified size of caribou population?

(b) The sample data pairs have x values ranging from 17 (hundred) to 34 (hundred) for the size of the caribou population. To predict the size of the wolf population when the size of the caribou population is 21 (hundred), will you be interpolating or extrapolating?

(c) Predict the size of the wolf population when the caribou population is 21 (hundred).

Example 3. The Quick Sell car dealership has been using in-app social media ads showing the different models and price ranges of cars on the lot that week. During a 10-week period, a Quick Sell dealer kept a weekly record of the number x of social media ads versus the number y of cars sold. The results are given in Table 4.2.

Table 4.2: Social Media Ads Versus Cars Sold

x	y
6	15
20	31
0	10
14	16
25	28
16	20
28	40
18	25
10	12
8	15

The manager decided that Quick Sell can afford only 12 ads per week. At that level of advertisement, how many cars can Quick Sell expect to sell each week? We'll answer this question in several steps.

(a) Draw a scatter diagram for the data.

(b) What quantities are needed to find the slope of the least-squares line?

(c) Add columns for x^2 and xy to Table 4.2 and sum the columns.

(d) Compute the sample means $\bar{x}$ and $\bar{y}$.

(e) Compute a and b for the equation $\hat{y} = a + bx$ of the least-squares line.

(f) What is the equation of the least-squares line $\hat{y} = a + bx$?

(g) Plot the least-squares line on your scatter diagram.

(h) Read the $\hat{y}$ value for $x = 12$ from your graph. Then use the equation of the least-squares line to calculate $\hat{y}$ when $x = 12$. How many cars can the manager expect to sell if 12 ads per week are purchased on their chosen social media platform?

(i) How reliable do you think the prediction is? Explain.

Definition 4.2.5. The coefficient of determination r^2 is the square of the sample correlation coefficient r. r^2 is a measure of the proportion of variation in y that is explained by the regression line, using x as the explanatory variable.

Example 4. In the previous example, we looked at the relationship between x = number of in-app social media ads showing different models of cars and y = number of cars sold each week by the sponsoring car dealership.

(a) Using the sums found in the previous example, compute the sample correlation coefficient r.

(b) Compute the coefficient of determination r^2.

(c) What percentage of the variation in the number of car sales can be explained by the ads and the least-squares line?

(d) What percentage of the variation in the number of car sales is not explained by the ads and the least-squares line?

Chapter 5

Elementary Probability Theory

5.1 What Is Probability?

Definition 5.1.1. Probability is a numerical measure between 0 and 1 that describes the likelihood that an event will occur. Probabilities closer to 1 indicate that the event is more likely to occur. Probabilities closer to 0 indicate that the event is less likely to occur.

Definition 5.1.2. $P(A)$, read "P of A," denotes the probability of event A.
If $P(A) = 1$, event A is certain to occur.
If $P(A) = 0$, event A is certain not to occur.

Definition 5.1.3 (Probability Assignments)**.**

1. A probability assignment based on intuition incorporates past experience, judgment, or opinion to estimate the likelihood of an event.

2. A probability assignment based on relative frequency uses the formula

$$\text{Probability of event} = \text{relative frequency} = \frac{f}{n}$$

   where f is the frequency of the event occurrence in a sample of n observations.

3. A probability assignment based on equally likely outcomes uses the formula

$$\text{Probability of event} = \frac{\text{Number of outcomes favorable to event}}{\text{Total number of outcomes}}.$$

Example 1. Consider each of the following events, and determine how the probability is assigned.

(a) A sports announcer claims that Sheila has a 90% chance of breaking the world record in the 100-yard dash.

(b) Henry figures that if he guesses on a true–false question, the probability of getting it right is 0.50.

(c) A pharmaceutical company claimed that the new flu vaccine developed by their scientists has an efficacy rate of 0.91. These results are based on a random sample of 1000 patients, of which 910 of them showed immunity to the flu.

Example 2. Assign a probability to the indicated event on the basis of the information provided. Indicate the technique you used: intuition, relative frequency, or the formula for equally likely outcomes.

(a) A random sample of 500 students at Hudson College were surveyed and it was determined that 375 wear glasses or contact lenses. Estimate the probability that a Hudson College student selected at random wears corrective lenses.

(b) The Friends of the Library hosts a fundraising barbecue. George is on the cleanup committee. There are four members on this committee, and they draw lots to see who will clean the grills. Assuming that each member is equally likely to be drawn, what is the probability that George will be assigned the grill-cleaning job?

(c) Joanna photographs whales for Sea Life Adventure Films. On her next expedition, she is to film blue whales feeding. Based on her knowledge of the habits of blue whales, she is almost certain she will be successful. What specific number do you suppose she estimates for the probability of success?

Theorem 5.1.1 (Law of Large Numbers)**.** *The law of large numbers states that when you repeat an experiment a large number of times (i.e., the sample size, n, increases) then the mean of the results would get closer to the theoretical mean of the experiment.*
One of the consequences of the law of large numbers is that in the long run, as the sample size increases and increases, the relative frequencies of outcomes get closer and closer to the theoretical (or actual) probability value.

Definition 5.1.4. A statistical experiment or statistical observation can be thought of as any random activity that results in a definite outcome.
An event is a collection of one or more outcomes of a statistical experiment or observation.
A simple event is one particular outcome of a statistical experiment.
The set of all simple events constitutes the sample space of an experiment.

Example 3. Human eye color is controlled by a single pair of genes (one from the father and one from the mother) called a genotype. Brown eye color, B, is dominant over blue eye color, ℓ. Therefore, in the genotype Bℓ, consisting of one brown gene B and one blue gene ℓ, the brown gene dominates. A person with a Bℓ genotype has brown eyes.
If both parents have brown eyes and have genotype Bℓ, what is the probability that their child will have blue eyes? What is the probability the child will have brown eyes?

Remark 1. The sum of the probabilities of all simple events in a sample space must equal 1.

Example 4. Professor Gutierrez is making up a final exam for a course in literature of the southwest. He wants the last three questions to be of the true–false type. To guarantee that the answers do not follow his favorite pattern, he lists all possible true–false combinations for three questions on slips of paper and then picks one at random from a hat.

(a) List all of the outcomes in the sample space.

(b) What is the probability that all three items will be false?Use the formula

$$P(\text{all } F) = \frac{\text{No. of favorable outcomes}}{\text{Total no. of outcomes}}.$$

(c) What is the probability that exactly two items will be true?

(d) Do you think it would be likely to get all three true–false questions correct if a student was to blindly guess on the final exam?

Definition 5.1.5. The complement of event A is the event that A does not occur. A^c designates the complement of event A. Furthermore,

1. $P(A) + P(A^c) = 1$.
2. $P(\text{event } A \text{ does } \textit{not} \text{ occur}) = P(A^c) = 1 - P(A)$.

Example 5. The probability that a college student will get the flu is 0.45. What is the probability that a college student will not get the flu?

Example 6. A veterinarian tells you that if you breed two cream-colored guinea pigs, the probability that an offspring will be pure white is 0.25. What is the probability that an offspring will not be pure white?

5.2 Some Probability Rules–Compound Events

Definition 5.2.1. Two events are independent if the occurrence or nonoccurrence of one event does *not* change the probability that the other event will occur.

Theorem 5.2.1 (Multiplication Rule for Independent Events)**.** *If two events A and B are independent, then*

$$P(A \text{ and } B) = P(A) \cdot P(B).$$

Theorem 5.2.2 (General Multiplication Rule for Any Events)**.**

$$P(A \text{ and } B) = P(A) \cdot P(B \mid A),$$
$$P(A \text{ and } B) = P(B) \cdot P(A \mid B).$$

Theorem 5.2.3 (Conditional Probability (When $P(B) \neq 0$))**.**

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}.$$

Example 1. Suppose you are going to throw two fair dice. What is the probability of getting a 5 on each die?

Example 2. Consider a collection of 6 balls that are identical except in color. There are 3 green balls, 2 blue balls, and 1 red ball. Compute the probability of drawing 2 green balls from the collection if the first ball is *not replaced* before the second ball is drawn.

Example 3. Kamal is 55, and the probability that he will be alive in 10 years is 0.72. Elaina is 35, and the probability that she will be alive in 10 years is 0.92. Assuming that the life span of one will have no effect on the life span of the other, what is the probability they will both be alive in 10 years?

(a) Are these events dependent or independent?

(b) Use the appropriate multiplication rule to find P(Kamal alive in 10 years *and* Elaina alive in 10 years).

Example 4. A quality-control procedure for testing digital cameras consists of drawing two cameras at random from each lot of 100 without replacing the first camera before drawing the second. If both are defective, the entire lot is rejected. Find the probability that both cameras are defective if the lot contains 10 defective cameras. Since we are drawing the cameras at random, assume that each camera in the lot has an equal chance of being drawn.

(a) What is the probability of getting a defective camera on the first draw?

(b) The first camera drawn is not replaced, so there are only 99 cameras for the second draw. What is the probability of getting a defective camera on the second draw if the first camera was defective?

(c) Does drawing a defective camera on the first draw change the probability of getting a defective camera on the second draw? Are the events dependent?

(d) Use the formula for dependent events to compute P(1st camera defective *and* 2nd camera defective).

Example 5. Indicate how each of the following pairs of events are combined. Use either the *and* combination or the *or* combination.

(a) Satisfying the humanities requirement by taking a course in the history of Japan or by taking a course in classical literature

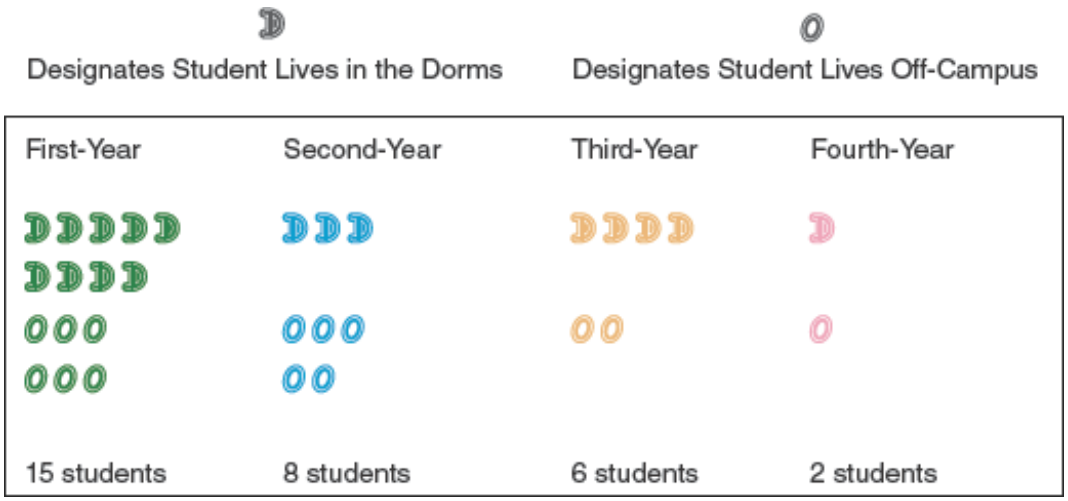
(b) Buying new tires and aligning the tires

(c) Getting an A not only in psychology but also in biology

(d) Having at least one of these pets: cat, dog, bird, rabbit

Example 6. Consider an introductory statistics class with 31 students. The students range from first- to fourth-year college students. Some students live in the dorms while others live off-campus. Figure 5.1 shows the sample space of the class.

Figure 5.1: Sample Space for Statistics Class



(a) Suppose we select one student at random from the class. Find the probability that the student is either a first-year student or a second-year student.

(b) Select one student at random from the class. What is the probability that the student is either a student who lives off-campus or a second-year?

Definition 5.2.2. Two events are mutually exclusive or disjoint if they cannot occur together. In particular, events A and B are mutually exclusive if $P(A \text{ and } B) = 0$.

Theorem 5.2.4 (Addition Rule for *Mutually Exclusive* Events A and B)**.** *If A and B are two mutually exclusive events, then*

$$P(A \text{ or } B) = P(A) + P(B).$$

Theorem 5.2.5 (Addition Rule for Any Events A and B)**.** *If A and B are any events, then*

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$

Example 7. The Cost Less Clothing Store carries pants at bargain prices. If you buy a pair of pants in your regular waist size without trying them on, the probability that the waist will be too tight is 0.30 and the probability that it will be too loose is 0.10.

(a) Are the events "too tight" and "too loose" mutually exclusive?

(b) If you choose a pair of pants at random in your regular waist size, what is the probability that the waist will be too tight or too loose?

Example 8. Professor Jackson is in charge of a program to prepare people for a high school equivalency exam. Records show that 80% of the students need work in math, 70% need work in English, and 55% need work in both areas.

(a) Are the events "needs math" and "needs English" mutually exclusive?

(b) Use the appropriate formula to compute the probability that a student selected at random needs math or needs English.

Example 9. Lara is playing a game with friends involving a pair of 6-sided dice. On her next move she needs to throw a sum bigger than 8 on the two dice. What is the probability that Lara will roll a sum bigger than 8?

Example 10. At Hopewell Electronics, all 140 employees were asked about their political affiliations. The employees were grouped by type of work, as executives or production workers. The results with row and column totals are shown in Table 5.1.

Table 5.1: Employee Type and Political Affiliation

	Political Affiliation			
Employee Type	Democrat (D)	Republican (R)	Independent (I)	Row Total
Executive (E)	5	34	9	48
Production worker (PW)	63	21	8	92
Column Total	68	55	17	140 Grand Total

Suppose an employee is selected at random from the 140 Hopewell employees. Let us use the following notation to represent different events of choosing: $E =$ executive; $PW =$ production worker; $D =$ Democrat; $R =$ Republican; $I =$ Independent.

(a) Compute $P(D)$ and $P(E)$.

(b) Compute $P(D \mid E)$.

(c) Are the events D and E independent?

(d) Compute $P(D \text{ and } E)$.

(e) Compute $P(D \text{ or } E)$.

Example 11. A common trend for finding a potential partner is through the use of online dating. Several online dating websites use information to match their clients with individuals who share similar interests. However, just because individuals are matched does not necessarily mean that their dates go well. Table 5.2 is survey data of 100 clients from an online dating site that evaluated whether they enjoyed their dates based on the type of dating activity. Let us use the following notation to represent different events of choosing: $I =$ Indoor Activity; $O =$ Outdoor Activity; $G =$ Good Date; $B =$ Bad Date; $N =$ Neutral Date.

Table 5.2: Date Results Based on Dating Activity

	Evaluation of the Date			
Type of Activity	Good Date (G)	Bad Date (B)	Neutral Date (N)	Row Total
Indoor Activity (I)	23	24	10	57
Outdoor Activity (O)	26	13	4	43
Column Total	49	37	14	100

(a) Compute $P(G)$ and $P(I)$.

(b) Compute $P(B \mid I)$.

(c) Compute $P(G \text{ and } O)$ by considering information given in the table as a sample space.

(d) Compute $P(G \text{ and } O)$ using the multiplication rule. Check to see if the results match those in part (c).

(e) Compute $P(G \text{ or } O)$

5.3 Trees and Counting Techniques

Theorem 5.3.1 (Multiplication Rule of Counting)**.** *Consider the series of events E_1 through E_m, where n_1 is the number of possible outcomes for event E_1, n_2 is the number of possible outcomes for event E_2, and n_m designates the number of possible outcomes for event E_m. Then the product*

$$n_1 \times n_2 \times \cdots \times n_m$$

gives the total number of possible outcomes for the series of events E_1, followed by E_2, up through event E_m.

Example 1. Jacqueline is in a nursing program and is required to take a course in psychology and one in physiology (A and P) next semester. She also wants to take Spanish II. If there are two sections of psychology, two of A and P, and three of Spanish II, how many different class schedules can Jacqueline choose from? (Assume that the times of the sections do not conflict.)

Example 2. Using the information from the previous example, make a tree diagram that shows all the possible course schedules for Jacqueline.

Example 3. Louis plays three tennis matches. Use a tree diagram to list the possible win and loss sequences Louis can experience for the set of three matches.

(a) On the first match Louis can win or lose. From Start, indicate these two branches.

(b) Regardless of whether Louis wins or loses the first match, he plays the second and can again win or lose. Attach branches representing these two outcomes to each of the first match results.

(c) Louis may win or lose the third match. Attach branches representing these two outcomes to each of the second match results.

(d) How many possible win–lose sequences are there for the three matches?

(e) Complete a list of win–lose sequences starting with winning the first match.

(f) Use the multiplication rule to compute the total number of outcomes for the three matches.

Example 4. Suppose there are five balls in an urn. They are identical except for color. Three of the balls are red and two are blue. You are instructed to draw out one ball, note its color, and set it aside. Then you are to draw out another ball and note its color. What are the outcomes of the experiment? What is the probability of each outcome?

Definition 5.3.1 (Factorial Notation)**.** For a counting number n,

$$0! = 1$$
$$1! = 1$$
$$n! = n(n-1)(n-2)\cdots 1.$$

Example 5.

(a) Evaluate 3!.

(b) In how many different ways can three objects be arranged in order?

Theorem 5.3.2 (Counting Rule for Permutations)**.** *The number of ways to arrange in order n distinct objects, taking them r at a time, is*

$$P_{n,r} = \frac{n!}{(n-r)!}$$

where n and r are whole numbers $n \geq r$. Another commonly used notation for permutations is nPr.

Example 6. Compute the number of possible ordered seating arrangements for eight people in five chairs.

Theorem 5.3.3 (Counting Rule for Combinations)**.** *The number of combinations of n objects taken r at a time is*

$$C_{n,r} = \frac{n!}{r!(n-r)!}$$

where n and r are whole numbers and $n \geq r$. Other commonly used notations for combinations include nCr and $\binom{n}{r}$.

Example 7. In your political science class, you are assigned to read any four books from a list of 10 books. How many different groups of four are available from the list of 10?

Example 8. The board of directors at Belford Community Hospital has 12 members.

(i) Three officers—president, vice president, and treasurer—must be elected from among the members. How many different slates of officers are possible? We will view a slate of officers as a list of three people, with the president listed first, the vice president listed second, and the treasurer listed third. For instance, if Felix, Jamie, and Leia wish to be on a slate together, there are several different slates possible, depending on the person listed for each office. Not only are we asking for the number of different groups of three names for a slate, we are also concerned about order.

(a) Do we use the permutations rule or the combinations rule? What is the value of n? What is the value of r?

(b) Use the permutations rule with $n = 12$ and $r = 3$ to compute $P_{12,3}$.

(ii) Three members from the group of 12 on the board of directors at Belford Community Hospital will be selected to go to a convention (all expenses paid) in Hawaii. How many different groups of three are there?

(c) Do we use the permutations rule or the combinations rule? What is the value of n? What is the value of r?

(d) Use the combinations rule with $n = 12$ and $r = 3$ to compute $C_{12,3}$.

Chapter 6

The Binomial Probability Distribution and Related Topics

6.1 Introduction to Random Variables and Probability Distributions

Definition 6.1.1. A quantitative variable x is a random variable if the value that x takes on in a given experiment or observation is a chance or random outcome.
A discrete random variable can take on only a finite number of values or a countable number of values.
A continuous random variable can take on any of the countless number of values in a line interval.

Example 1. Which of the following random variables are discrete and which are continuous?

(a) The time it takes a selected student to register for the Fall term.

(b) The number of text messages received by a selected student on a randomly chosen day.

(c) The number of miles an electric vehicle can drive on a full charge.

(d) Pick a random sample of 50 registered voters in a district and find the number who voted in the last county election.

Definition 6.1.2. A probability distribution is an assignment of probabilities to each distinct value of a discrete random variable or to each interval of values of a continuous random variable.

Example 2. Dr. Mendoza developed a test to measure agility. He administered it to a group of 20,000 adults between the ages of 25 and 35. The possible scores were 0, 1, 2, 3, 4, 5, and 6, with 6 indicating the most agile. The test results for this group are shown in Table 6.1.

Table 6.1: Agility Test Scores for 20,000 Subjects

Score	Number of Subjects
0	1400
1	2600
2	3600
3	6000
4	4400
5	1600
6	400

(a) If a subject is chosen at random from this group, what is the probability that they will have a score of 3?

(b) Make a graph of this distribution using a relative-frequency histogram.

(c) The Topnotch Clothing Company needs to hire someone with a score on the agility test of 5 or 6 to operate the fabric press machine. What is the probability someone in the group scored a 5 or 6?

Example 3. A tool of cryptanalysis (science of code breaking) is to use relative frequencies of occurrence of letters to help break codes. Creators of word games also use the relative frequencies of letters when designing puzzles. Suppose we take a random sample of 1000 words occurring in crossword puzzles. Table 6.2 shows the relative frequency of letters occurring in the sample.

Table 6.2: Frequencies of Letters in a 1000-Letter Sample

Letter	Freq.	Prob.	Letter	Freq.	Prob.
A	85		N	66	0.066
B	21	0.021	O	72	
C	45	0.045	P	32	0.032
D	34	0.034	Q	2	0.002
E	112		R	76	0.076
F	18	0.018	S	57	0.057
G	25	0.025	T	69	0.069
H	30	0.030	U	36	
I	75		V	10	0.010
J	2	0.002	W	13	0.013
K	11	0.011	X	3	0.003
L	55	0.055	Y	18	0.018
M	30	0.030	Z	3	0.003

(a) Use the relative frequencies to compute the omitted probabilities in Table 6.2.

(b) Do the probabilities of all the individual letters add up to 1?

(c) If a letter is selected at random from a crossword puzzle, what is the probability the letter will be a vowel?

Definition 6.1.3. The mean and the standard deviation of a discrete population probability distribution are found by using these formulas:

$$\mu = \sum xP(x);\ \mu \text{ is called the expected value of } x$$
$$\sigma = \sqrt{\sum (x-\mu)^2 P(x)};\ \sigma \text{ is called the standard deviation of } x$$

where x is the value of a random variable, $P(x)$ is the probability of that variable, and the sum Σ is taken for all the values of the random variable.
Note: μ is the *population mean*, and σ is the underlying *population standard deviation* because the sum Σ is taken over *all* values of the random variable (i.e., the entire sample space).

Example 4. Do podcast ads work? One marketing firm determined the number of times buyers of a product had been exposed to a podcast ad for that product before purchase. The results are shown here:

Number of Times Buyers Saw Infomercial	1	2	3	4	5*
Percentage of Buyers	27%	31%	18%	9%	15%

*This category was 5 or more, but will be treated as 5 in this example.

We can treat the information shown as an estimate of the probability distribution because the events are mutually exclusive and the sum of the percentages is 100%. Compute the mean and standard deviation of the distribution.

Example 5. At a carnival, you pay \$2.00 to play a coin-flipping game with three fair coins. On each coin one side has the number 0 and the other side has the number 1. You flip the three coins at one time and you win \$1.00 for every 1 that appears on top. Are your expected earnings equal to the cost to play? We'll answer this question in several steps.

(a) In this game, the random variable of interest counts the number of 1s that show. What is the sample space for the values of this random variable?

(b) What are the eight equally likely outcomes for throwing three coins?

(c) Create a frequency table for this random variable. Include columns for $P(x)$ and $xP(x)$.

(d) Sum the appropriate column of your frequency table to find the expected value μ. Are your expected earnings less than, equal to, or more than the cost of the game?

6.2 Binomial Probabilities

Remark 1 (Features of a Binomial Experiment).

1. There is a *fixed number of trials*. We denote this number by the letter n.
2. The n trials are *independent* and repeated under identical conditions.
3. Each trial has only *two outcomes*: success, denoted by S, and failure, denoted by F.
4. For each individual trial, the *probability of success is the same*. We denote the probability of success by p and that of failure by q. Since each trial results in either success or failure, $p + q = 1$ and $q = 1 - p$.
5. The central problem of a binomial experiment is to find the *probability of r successes out of n trials*.

Example 1. On a TV game show, each contestant has a try at the wheel of fortune. The wheel of fortune is a roulette wheel with 36 equal slots, one of which is gold. If the ball lands in the gold slot, the contestant wins \$50,000. No other slot pays. If a binomial experiment is used to find the probability that the game show will have to pay the fortune to 3 contestants out of 100, what are p, q, n, and r for the experiment?

Example 2. Let's analyze the following binomial experiment to determine p, q, n, and r:
According to the *Textbook of Medical Physiology*, 5th edition, by Arthur Guyton, 9% of the population has blood type B. Suppose we choose 18 people at random from the population and test the blood type of each. What is the probability that 3 of these people have blood type B? *Note:* Independence is approximated because 18 people is an extremely small sample with respect to the entire population.

(a) In this experiment, we are observing whether or not a person has type B blood. We will say we have a success if the person has type B blood. What is failure?

(b) The probability of success is 0.09, since 9% of the population has type B blood. What is the probability of failure, q?

(c) How many trials are in this experiment?

(d) We wish to compute the probability of 3 successes out of 18 trials. What is r in this case?

Definition 6.2.1 (Formula for the Binomial Probability Distribution)**.**

$$P(r) = \frac{n!}{r!(n-r)!}p^r q^{n-r} = C_{n,r}p^r q^{n-r}$$

where

n = number of binomial trials
$q = 1 - p$ = probability of failure on each trial
r = random variable representing the number of successes out of n trials $(0 \leq r \leq n)$.

Example 3. Privacy of online information is a concern for many people. One survey showed that 59% of people are concerned about the confidentiality of their personal information online. Based on this information, what is the probability that for a random sample of 10 people, 6 are concerned about the privacy of their personal information online?

Example 4. A biologist is studying a new hybrid tomato. It is known that the seeds of this hybrid tomato have probability 0.70 of germinating. The biologist plants six seeds.

(a) What is the probability that *exactly* four seeds will germinate?

(b) What is the probability that *at least* four seeds will germinate?

Example 5. A rarely performed and somewhat risky eye operation is known to be successful in restoring the eyesight of 30% of the patients who undergo the operation. A team of surgeons has developed a new technique for this operation that has been successful in four of six operations. Does it seem likely that the new technique is much better than the old? We'll use the binomial probability distribution to answer this question. We'll compute the probability of at least four successes in six trials for the old technique.

(a) Each operation is a binomial trial. In this case, what are n, p, q, and r?

(b) Use your values of n, p, and q, as well as the formula or your calculator, to compute $P(4)$.

(c) Compute the probability of at least four successes out of the six trials.

(d) Under the older operation technique, what is the probability that at least four patients out of six regain their eyesight? Does it seem that the new technique is better than the old? Would you encourage the surgeon team to do more work on the new technique?

6.3 Additional Properties of the Binomial Distribution

Example 1. A waiter at the Green Spot Restaurant has learned from long experience that the probability that a lone diner will leave a tip is only 0.5. During one lunch hour, the waiter serves six people who are dining by themselves. Make a graph of the binomial probability distribution that shows the probabilities that 0, 1, 2, 3, 4, 5, or all 6 lone diners leave tips.

Example 2. Tashika enjoys playing basketball. She figures that she makes about 80% of the free throws she attempts during a game. Make a histogram showing the probability that Tashika will make 0, 1, 2, 3, 4, 5, or 6 shots out of six attempted free throws.

(a) This is a binomial experiment. In this situation, we'll say success occurs when Tashika makes an attempted free throw. What are the values of n and p?

(b) Construct a table of values for this binomial experiment.

(c) Create a histogram for the table of values.

Remark 1.

$\mu = np$ is the expected number of successes for the random variable r
$\sigma = \sqrt{npq}$ is the standard deivation for the random variable r

where

r is a random variable representing the number of successes in a binomial distribution,
n is the number of trials
p is the probability of success on a single trial, and
$q = 1 - p$ is the probability of failure on a single trial.

Example 3. Let's compute the mean and standard deviation for the distribution of the previous example that describes that probabilities of lone diners leaving tips at the Green Spot Restaurant.

Example 4. When Tashika shoots free throws in basketball games, the probability that she makes a shot is 0.8.

(a) The mean of the binomial distribution is the expected value of r successes out of n trials. Out of six throws, what is the expected number of free throws Tashika will make?

(b) For six trials, what is the standard deviation of the binomial distribution of the number of successful free throws Tashika makes?

Remark 2. For a binomial distribution, it is unusual for the number of successes r to be higher than $\mu + 2.5\sigma$ or lower than $\mu - 2.5\sigma$.

Chapter 7

Normal Curves and Sampling Distributions

7.1 Graphs of Normal Probability Distributions

Remark 1 (Important Properties of a Normal Curve).

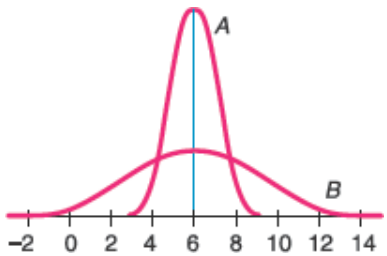
1. The curve is bell-shaped, with the highest point over the mean μ.
2. The curve is symmetric about a vertical line through μ.
3. The curve approaches the horizontal axis but never touches or crosses it.
4. The inflection (transition) points between concave upward and downward occur above $\mu + \sigma$ and $\mu - \sigma$.
5. The area under the entire curve is 1.

Example 1. Look at the normal curves in Figure 7.1.

(a) Do these distributions have the same mean? If so, what is it?

(b) One of the curves corresponds to a normal distribution with $\sigma = 3$ and the other to one with $\sigma = 1$. Which curve has which σ?

Figure 7.1: P-value



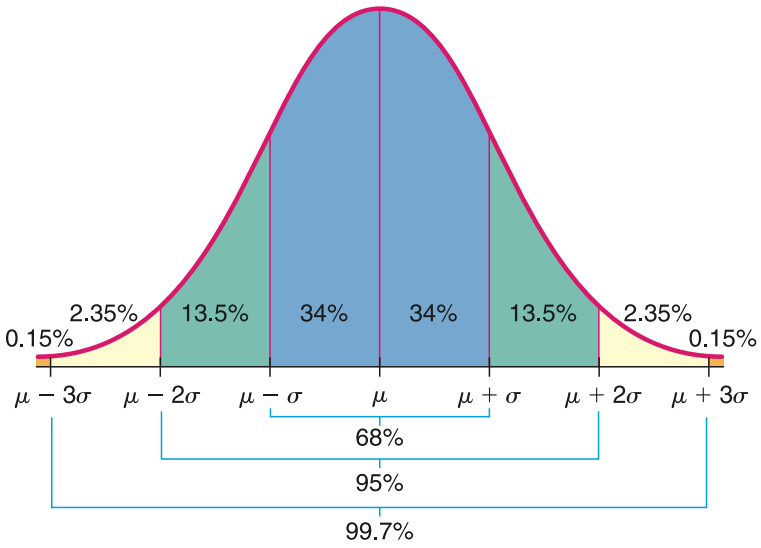
Remark 2 (Empirical Rule). For a distribution that is symmetric and bell-shaped (in particular, for a normal distribution):

Approximately 68% of the data values will lie within 1 standard deviation on each side of the mean.

Approximately 95% of the data values will lie within 2 standard deviations on each side of the mean.

Approximately 99.7% (or almost all) of the data values will lie within 3 standard deviations on each side of the mean.

Figure 7.2: Area Under a Normal Curve



Example 2. The logon time for an application is the time between entering a user name and password and the app becoming active. Tomas is testing an app with login times that are normally distributed with a mean $\mu = 5$ seconds and a standard deviation $\sigma = 1.5$ seconds. What is the probability that a logon event selected at random takes from 5 to 6.5 seconds?

Example 3. The yearly wheat yield per acre on a particular farm is normally distributed with mean $\mu = 35$ bushels and standard deviation $\sigma = 8$ bushels. (A bushel is a measurement of volume used for dry goods.)

(a) Draw a normal curve and shade the area under the curve that represents the probability that an acre will yield between 19 and 35 bushels.

(b) Is the area the same as the area between $\mu - 2\sigma$ and μ?

(c) Use Figure 7.2 to find the percentage of area over the interval between 19 and 35.

(d) What is the probability that the yield will be between 19 and 35 bushels per acre?

7.2 Standard Units and Areas under the Standard Normal

Definition 7.2.1. The z value or z score (also known as standard score) gives the number of standard deviations between the original measurement x and the mean μ of the x distribution:

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

Example 1. A pizza parlor franchise specifies that the average (mean) amount of cheese on a large pizza should be 8 ounces and the standard deviation only 0.5 ounce. An inspector picks out a large pizza at random in one of the pizza parlors and finds that it is made with 6.9 ounces of cheese. Assume that the amount of cheese on a pizza follows a normal distribution. If the amount of cheese is below the mean by more than 3 standard deviations, the parlor will be in danger of losing its franchise.

How many standard deviations from the mean is 6.9? Is the pizza parlor in danger of losing its franchise?

Definition 7.2.2. Given an x distribution with mean μ and standard deviation σ, the raw score x corresponding to a z score is

$$x = z\sigma + \mu.$$

Example 2. Ezra figures that it takes an average (mean) of 17 minutes with a standard deviation of 3 minutes to drive from home, park the car, and walk to an early morning class.

(a) One day it took Ezra 21 minutes to get to class. How many standard deviations from the average is that? Is the z value positive or negative? Explain why it should be either positive or negative.

(b) What commuting time corresponds to a standard score of $z = -2.5$? Could Ezra count on making it to class in this amount of time or less?

Definition 7.2.3. The standard normal distribution is a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

Example 3. Use Table 3 of the Appendix to find the described areas under the standard normal curve.

(a) Find the area under the standard normal curve to the left of $z = -1.00$.

(b) Find the area to the left of $z = 1.18$.

Example 4. Table 3, Areas of a Standard Normal Distribution, is located in the Appendix. Spend a little time studying the table, and then answer these questions.

(a) As z values increase, do the areas to the left of z increase?

(b) If a z value is negative, is the area to the left of z less than 0.5000?

(c) If a z value is positive, is the area to the left of z greater than 0.5000?

Example 5. Use Table 3 of the Appendix to find the specified areas.

(a) Find the area between $z = 1.00$ and $z = 2.10$.

(b) Find the area to the right of $z = 0.94$.

Example 6. Let z be a random variable with a standard normal distribution.

(a) $P(z \geq 1.15)$ refers to the probability that z values lie to the right of 1.15. Shade the corresponding area under the standard normal curve and find $P(z \geq 1.15)$.

(b) Find $P(-1.78 \leq z \leq 0.35)$. First, sketch the area under the standard normal curve corresponding to the area.

7.3 Areas Under Any Normal Curve

Example 1. Let x have a normal distribution with $\mu = 10$ and $\sigma = 2$. Find the probability that an x value selected at random from this distribution is between 11 and 14. In symbols, find $P(11 \leq x \leq 14)$.

Example 2. The life span of a rechargeable battery is the time before the battery must be replaced because it no longer holds a charge. One tablet computer model has a battery with a life span that is normally distributed with a mean of 2.3 years and a standard deviation of 0.4 year. What is the probability that the battery will have to be replaced during the guarantee period of 2 years?

(a) Let x represent the battery life span. The statement that the battery needs to be replaced during the 2-year guarantee period means the life span is less than 2 years, or $x \leq 2$. Convert this statement to a statement about z.

(b) Sketch the area to be found under the standard normal curve. Does this area correspond to the probability that $z \leq -0.75$?

(c) Use Table 3 of the Appendix to find $P(z \leq -0.75)$.

(d) What is the probability that the battery will fail before the end of the guarantee period?

Example 3. Magic Video Games, Inc., sells an expensive video games package. Because the package is so expensive, the company wants to advertise an impressive guarantee for the life expectancy of its computer control system. The guarantee policy will refund the full purchase price if the control system fails during the guarantee period. The research department has done tests that show that the mean life for the computer is 30 months, with standard deviation of 4 months. The computer life is normally distributed. How long can the guarantee period be if management does not want to refund the purchase price on more than 7% of the Magic Video packages?

Example 4. Find the z value such that 90% of the area under the standard normal curve lies between $-z$ and z.

Example 5. Find the z value such that 3% of the area under the standard normal curve lies to the right of z.

(a) Draw a sketch of the standard normal distribution showing the described area.

(b) Find the area to the left of z.

(c) Look up the area in Table 3 of the Appendix and find the corresponding z.

(d) Suppose the time to complete a test is normally distributed with $\mu = 40$ minutes and $\sigma = 5$ minutes. After how many minutes can we expect all but about 3% of the tests to be completed?

(e) Use Table 3 of the Appendix to find a z value such that 3% of the area under the standard normal curve lies to the left of z.

(f) Compare the z value of part (c) with the z value of part (e). Is there any relationship between the z values?

Example 6. Consider the following data, which are rounded to the nearest integer.

19	19	19	16	21	14	23	17	19	20	18	24	20	13	16
17	19	18	19	17	21	24	18	23	19	21	22	20	20	20
24	17	20	22	19	22	21	18	20	22	16	15	21	23	21
18	18	20	15	25										

(a) Look at a histogram and box-and-whisker plot of the data and comment about normality of the data from these indicators.

(b) Check for skewness using

$$\text{Pearon's index} = \frac{3(\bar{x} - \text{median})}{s}.$$

(c) Look at a normal quantile plot and comment on normality.

(d) Interpret the results.

7.4 Sampling Distributions

Definition 7.4.1. A statistic is a numerical descriptive measure of a *sample*. A parameter is a numerical descriptive measure of a *population*.

Definition 7.4.2. A sampling distribution is a probability distribution of a sample statistic based on all possible simple random samples of the same size from the same population.

Example 1. Pinedale, Wisconsin, is a rural community with a children's fishing pond. Posted rules state that all fish under 6 inches must be returned to the pond, only children under 12 years old may fish, and a limit of five fish may be kept per day. Jasmine is a college student who was hired by the community last summer to make sure the rules were obeyed and to see that the children were safe from accidents. The pond contains only rainbow trout and has been well stocked for many years. Each child has no difficulty catching his or her limit of five trout.
As a project for her biometrics class, Jasmine kept a record of the lengths (to the nearest inch) of all trout caught last summer. Hundreds of children visited the pond and caught their limit of five trout, so Jasmine has a lot of data. To make Table 7-8, Jasmine selected 100 children at random and listed the lengths of each of the five trout caught by that child. Then, for each child, she listed the mean length of the five trout that child caught.

Figure 7.3: Length Measurements of Trout Caught by a Random Sample of 100 Children at the Pinedale Children's Pond

Sample	Length (to nearest inch)					$\bar{x}$ = Sample Mean	Sample	Length (to nearest inch)					$\bar{x}$ = Sample Mean
1	11	10	10	12	11	10.8	51	9	10	12	10	9	10.0
2	11	11	9	9	9	9.8	52	7	11	10	11	10	9.8
3	12	9	10	11	10	10.4	53	9	11	9	11	12	10.4
4	11	10	13	11	8	10.6	54	12	9	8	10	11	10.0
5	10	10	13	11	12	11.2	55	8	11	10	9	10	9.6
6	12	7	10	9	11	9.8	56	10	10	9	9	13	10.2
7	7	10	13	10	10	10.0	57	9	8	10	10	12	9.8
8	10	9	9	9	10	9.4	58	10	11	9	8	9	9.4
9	10	10	11	12	8	10.2	59	10	8	9	10	12	9.8
10	10	11	10	7	9	9.4	60	11	9	9	11	11	10.2
11	12	11	11	11	13	11.6	61	11	10	11	10	11	10.6
12	10	11	10	12	13	11.2	62	12	10	10	9	11	10.4
13	11	10	10	9	11	10.2	63	10	10	9	11	7	9.4
14	10	10	13	8	11	10.4	64	11	11	12	10	11	11.0
15	9	11	9	10	10	9.8	65	10	10	11	10	9	10.0
16	13	9	11	12	10	11.0	66	8	9	10	11	11	9.8
17	8	9	7	10	11	9.0	67	9	11	11	9	8	9.6
18	12	12	8	12	12	11.2	68	10	9	10	9	11	9.8
19	10	8	9	10	10	9.4	69	9	9	11	11	11	10.2
20	10	11	10	10	10	10.2	70	13	11	11	9	11	11.0
21	11	10	11	9	12	10.6	71	12	10	8	8	9	9.4
22	9	12	9	10	9	9.8	72	13	7	12	9	10	10.2
23	8	11	10	11	10	10.0	73	9	10	9	8	9	9.0
24	9	12	10	9	11	10.2	74	11	11	10	9	10	10.2
25	9	9	8	9	10	9.0	75	9	11	14	9	11	10.8
26	11	11	12	11	11	11.2	76	14	10	11	12	12	11.8
27	10	10	10	11	13	10.8	77	8	12	10	10	9	9.8
28	8	7	9	10	8	8.4	78	8	10	13	9	8	9.6
29	11	11	8	10	11	10.2	79	11	11	11	13	10	11.2
30	8	11	11	9	12	10.2	80	12	10	11	12	9	10.8
31	11	9	12	10	10	10.4	81	10	9	10	10	13	10.4
32	10	11	10	11	12	10.8	82	11	10	9	9	12	10.2
33	12	11	8	8	11	10.0	83	11	11	10	10	10	10.4
34	8	10	10	9	10	9.4	84	11	10	11	9	9	10.0
35	10	10	10	10	11	10.2	85	10	11	10	9	7	9.4
36	10	8	10	11	13	10.4	86	7	11	10	9	11	9.6
37	11	10	11	11	10	10.6	87	10	11	10	10	10	10.2
38	7	13	9	12	11	10.4	88	9	8	11	10	12	10.0
39	11	11	8	11	11	10.4	89	14	9	12	10	9	10.8
40	11	10	11	12	9	10.6	90	9	12	9	10	10	10.0
41	11	10	9	11	12	10.6	91	10	10	8	6	11	9.0
42	11	13	10	12	9	11.0	92	8	9	11	9	10	9.4
43	10	9	11	10	11	10.2	93	8	10	9	9	11	9.4
44	10	9	11	10	9	9.8	94	12	11	12	13	10	11.6
45	12	11	9	11	12	11.0	95	11	11	9	9	9	9.8
46	13	9	11	8	8	9.8	96	8	12	8	11	10	9.8
47	10	11	11	11	10	10.6	97	13	11	11	12	8	11.0
48	9	9	10	11	11	10.0	98	10	11	8	10	11	10.0
49	10	9	9	10	10	9.6	99	13	10	7	11	9	10.0
50	10	10	6	9	10	9.0	100	9	9	10	12	12	10.4

Now let us turn our attention to the following question: What is the average (mean) length of a trout taken from the Pinedale children's pond last summer?

Example 2.

(a) What is a population parameter? Give an example.

(b) What is a sample statistic? Give an example.

(c) What is a sampling distribution?

(d) In Table 7.3, what makes up the members of the sample? What is the sample statistic corresponding to each sample? What is the sampling distribution? To which population parameter does this sampling distribution correspond?

(e) Where will sampling distributions be used in our study of statistics?

7.5 The Central Limit Theorem

Theorem 7.5.1 (Sampling Distribution for $\bar{x}$ from a Normal Distribution)**.** *Let x be a random variable with a normal distribution whose mean is μ and whose standard deviation is σ. Let $\bar{x}$ be the sample mean corresponding to random samples of size n taken from the x distribution. Then the following are true:*

(a) The $\bar{x}$ distribution is a normal distribution.

(b) The mean of the $\bar{x}$ distribution is μ.

(c) The standard deviation of the $\bar{x}$ distribution is $\sigma/\sqrt{n}$.

Remark 1. Theorem 7.5.1 implies that we can convert the $\bar{x}$ distribution to the standard normal z distribution using the following formulas.

$$\mu_{\bar{x}} = \mu$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$
$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

where n is the sample size, μ is the mean of the x distribution, and σ is the standard deviation of the x distribution.

Example 1. Suppose a team of biologists has been studying the Pinedale children's fishing pond. Let x represent the length of a single trout taken at random from the pond. This group of biologists has determined that x has a normal distribution with mean $\mu = 10.2$ inches and standard deviation $\sigma = 1.4$ inches.

(a) What is the probability that a single trout taken at random from the pond is between 8 and 12 inches long?

(b) What is the probability that the mean length $\bar{x}$ of five trout taken at random is between 8 and 12 inches?

(c) Looking at the results of parts (a) and (b), we see that the probabilities (0.8433 and 0.9977) are quite different. Why is this the case?

Definition 7.5.1. The standard error is the standard deviation of a sampling distribution. For the $\bar{x}$ sampling distribution,

$$\text{standard error} = \sigma_{\bar{x}} = \sigma/\sqrt{n}.$$

Theorem 7.5.2 (The Central Limit Theorem for Any Probability Distribution)**.** *If x possesses any distribution with mean μ and standard deviation σ, then the sample mean $\bar{x}$ based on a random sample of size n will have a distribution that approaches the distribution of a normal random variable with mean μ and standard deviation $\sigma/\sqrt{n}$ as n increases without limit.*

Example 2.

(a) Suppose x has a normal distribution with mean $\mu = 18$ and standard deviation $\sigma = 13$. If you draw random samples of size 5 from the x distribution and $\bar{x}$ represents the sample mean, what can you say about the $\bar{x}$ distribution? How could you standardize the $\bar{x}$ distribution?

(b) Suppose you know that the x distribution has mean $\mu = 75$ and standard deviation $\sigma = 12$, but you have no information as to whether or not the x distribution is normal. If you draw samples of size 30 from the x distribution and $\bar{x}$ represents the sample mean, what can you say about the $\bar{x}$ distribution? How could you standardize the $\bar{x}$ distribution?

(c) Suppose you did not know that x had a normal distribution. Would you be justified in saying that the $\bar{x}$ distribution is approximately normal if the sample size were $n = 8$?

Example 3. A certain strain of bacteria occurs in all raw milk. Let x be the bacteria count per milliliter of milk. The health department has found that if the milk is not contaminated, then x has a distribution that is more or less mound-shaped and symmetric. The mean of the x distribution is $\mu = 2500$, and the standard deviation is $\sigma = 300$. In a large commercial dairy, the health inspector takes 42 random samples of the milk produced each day. At the end of the day, the bacteria count in each of the 42 samples is averaged to obtain the sample mean bacteria count $\bar{x}$.

(a) Assuming the milk is not contaminated, what is the distribution of $\bar{x}$?

(b) Assuming the milk is not contaminated, what is the probability that the average bacteria count $\bar{x}$ for one day is between 2350 and 2650 bacteria per milliliter?

(c) At the end of each day, the inspector must decide to accept or reject the accumulated milk that has been held in cold storage awaiting shipment. Suppose the 42 samples taken by the inspector have a mean bacteria count $\bar{x}$ that is not between 2350 and 2650. If you were the inspector, what would be your comment on this situation?

Example 4. In mountain country, major highways sometimes use tunnels instead of long, winding roads over high passes. However, too many vehicles in a tunnel at the same time can cause a hazardous situation. Traffic engineers are studying a long tunnel in Colorado. If x represents the time for a vehicle to go through the tunnel, it is known that the x distribution has mean $\mu =$ 12.1 minutes and standard deviation $\sigma = 3.8$ minutes under ordinary traffic conditions. From a histogram of x values, it was found that the x distribution is mound-shaped with some symmetry about the mean.
Engineers have calculated that, on average, vehicles should spend from 11 to 13 minutes in the tunnel. If the time is less than 11 minutes, traffic is moving too fast for safe travel in the tunnel. If the time is more than 13 minutes, there is a problem of bad air quality (too much carbon monoxide and other pollutants).
Under ordinary conditions, there are about 50 vehicles in the tunnel at one time. What is the probability that the mean time for 50 vehicles in the tunnel will be from 11 to 13 minutes?
We will answer this question in steps.

(a) Let $\bar{x}$ represent the sample mean based on samples of size 50. Describe the $\bar{x}$ distribution.

(b) Find $P(11 < \bar{x} < 13)$.

(c) Interpret your answer to part (b).

Definition 7.5.2. A sample statistic is unbiased if the mean of its sampling distribution equals the value of the parameter being estimated.
The spread of the sampling distribution indicates the variability of the statistic. The spread is affected by the sampling method and the sample size. Statistics from larger random samples have spreads that are smaller.

7.6 Normal Approximation to the Binomial and $\hat{p}$ Distributions

Remark 1 (Normal Approximation to the Binomial Distribution). Consider a binomial distribution where

n = number of trials
r = number of successes
p = probability of success on a single trial
$q = 1 - p$ = probability of failure on a single trial.

If $np > 5$ and $nq > 5$, then r has a binomial distribution that is approximated by a normal distribution with $\mu = np$ and $\sigma = \sqrt{npq}$.
Note: As n increases, the approximation becomes better.

Example 1. Graph the binomial distributions for which $p = 0.25$, $q = 0.75$, and the number of trials is first $n = 3$, then $n = 10$, then $n = 25$, and finally $n = 50$.

Example 2. The owner of a new apartment building must install 25 water heaters. From past experience in other apartment buildings, she knows that Quick Hot is a good brand. A Quick Hot heater is guaranteed for 5 years only, but from the owner's past experience, she knows that the probability it will last 10 years is 0.25.

(a) What is the probability that 8 or more of the 25 water heaters will last at least 10 years? Define success to mean a water heater that lasts at least 10 years.

(b) How does this result compare with the result we can obtain by using the formula for the binomial probability distribution with $n = 25$ and $p = 0.25$?

(c) How do the results of parts (a) and (b) compare?

Remark 2. Remember that when we use the normal distribution to approximate the binomial, we are computing the areas under bars. The bar over the discrete variable r extends from $r - 0.5$ to $r + 0.5$. This means that the corresponding continuous normal variable x extends from $r-0.5$ to $r+0.5$. Adjusting the values of discrete random variables to obtain a corresponding range for a continuous random variable is called making a *continuity correction*.

Example 3. From many years of observation, a biologist knows that the probability is only 0.65 that any given Arctic tern will survive the migration from its summer nesting area to its winter feeding grounds. A random sample of 500 Arctic terns were banded at their summer nesting area. Use the normal approximation to the binomial and the following steps to find the probability that between 310 and 340 of the banded Arctic terns will survive the migration. Let r be the number of surviving terns.

(a) Find the μ and σ needed to approximate $P(310 \leq r \leq 340)$.

(b) Make a continuity correction for $P(310 \leq r \leq 340)$.

(c) Convert the condition $309.5 \leq x \leq 340.5$ to a condition in standard units.

(d) Find $P(-1.45 \leq z \leq 1.45)$.

(e) Will the normal distribution make a good approximation to the binomial for this problem? Explain your answer.

Remark 3 (Sampling Distribution for the Proportion $\hat{p} = \frac{r}{n}$). Given

$$\begin{aligned} n &= \text{number of binomial trials (fixed constant)} \\ r &= \text{number of successes} \\ p &= \text{probability of success on each trial} \\ q &= 1 - p = \text{probability of failure on each trial.} \end{aligned}$$

If $np > 5$ and $nq > 5$, then the random variable $\hat{p} = r/n$ can be approximated by a normal random variable (x) with mean and standard deviation

$$\mu_{\hat{p}} = p \text{ and } \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}.$$

Example 4. The annual crime rate in the Capitol Hill neighborhood of Denver is 111 victims per 1000 residents. This means that 111 out of 1000 residents have been the victim of at least one crime (Source: Neighborhood Facts, Piton Foundation). For more information, visit the Piton Foundation website. These crimes range from relatively minor crimes (stolen hubcaps or purse snatching) to major crimes (murder). The Arms is an apartment building in this neighborhood that has 50 year-round residents. Suppose we view each of the $n = 50$ residents as a binomial trial. The random variable r (which takes on values $0, 1, 2, \ldots, 50$) represents the number of victims of at least one crime in the next year.

(a) What is the population probability p that a resident in the Capital Hill neighborhood will be the victim of a crime next year? What is the probability q that a resident will not be a victim?

(b) Conider the random variable

$$\hat{p} = \frac{r}{n} = \frac{r}{50}$$

Can we approximate the $\hat{p}$ distribution with a normal distribution? Explain.

(c) What are the mean and standard deviation for the $\hat{p}$ distribution?

Chapter 8

Estimation

8.1 Estimating μ When σ Is Known

Definition 8.1.1. A point estimate of a population parameter is an estimate of the parameter using a single number.
$\bar{x}$ is the point estimate for μ.
When using $\bar{x}$ as a point estimate for μ, the margin of error is the magnitude of $\bar{x} - \mu$ or $|\bar{x} - \mu|$.

Definition 8.1.2. For a confidence level c, the critical value z_c is the number such that the area under the standard normal curve between $-z_c$ and z_c equals c.

Example 1. Use Table 3 of the Appendix to find a number $z_{0.99}$ such that 99% of the area under the standard normal curve lies between $-z_{0.99}$ and $z_{0.99}$. That is, find $z_{0.99}$ such that

$$P(-z_{0.99} < z < z_{0.99}) = 0.99.$$

Definition 8.1.3. A c confidence interval for μ is an interval computed from sample data in such a way that c is the probability of generating an interval containing the actual value of μ. In other words, c is the proportion of confidence intervals, based on random samples of size n, that actually contain μ.

Remark 1. Let x be a random variable. Obtain a simple random sample (of size n) of x values and compute the sample mean $\bar{x}$. If x has an approximately normal distribution and the value of σ is already known, then the confidence interval for μ is

$$\bar{x} - E < \mu < \bar{x} + E$$

where $\bar{x}$ = sample mean of a simple random sample and

$$\begin{aligned} E &= z_c \frac{\sigma}{\sqrt{n}} \\ c &= \text{confidence level } (0 < c < 1) \\ z_c &= \text{critical value for confidence level } c. \end{aligned}$$

Example 2. Cameron enjoys jogging. They have been jogging over a period of several years, during which time their physical condition has remained constantly good. Usually, they jog 2 miles per day. The standard deviation of their times is $\sigma = 1.80$ minutes. During the past year, Cameron has recorded their times to run 2 miles. They have a random sample of 90 of these times. For these 90 times, the mean was $\bar{x} = 15.60$ minutes. Let μ be the mean jogging time for the entire distribution of Cameron's 2-mile running times (taken over the past year). Find a 0.95 confidence interval for μ.

Example 3. Emmet usually meets Cameron at the track. Emmet prefers to jog 3 miles. From long experience, Emmet knows that $\sigma = 2.40$ minutes for jogging times. For a random sample of 90 jogging sessions, the mean time was $\bar{x} = 22.50$ minutes. Let μ be the mean jogging time for the entire distribution of Emmet's 3-mile running times over the past several years. Find a 0.99 confidence interval for μ.

(a) Is the $\bar{x}$ distribution approximately normal? Do we know σ?

(b) What is the value of $z_{0.99}$?

(c) What is the value of E?

(d) What are the endpoints for a 0.99 confidence interval for μ?

(e) Explain what the confidence interval tells us.

Example 4. A wildlife study is designed to find the mean weight of salmon caught by an Alaskan fishing company. A preliminary study of a random sample of 50 salmon showed $s = 2.15$ pounds. How large a sample should be taken to be 99% confident that the sample mean $\bar{x}$ is within 0.20 pound of the true mean weight μ?

8.2 Estimating μ When σ Is Unknown

Definition 8.2.1. Assume that x has a normal distribution with mean μ. For samples of size n with sample mean $\bar{x}$ and sample standard deviation s, the t variable

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

has a Student's t distribution with degrees of freedom $d.f. = n - 1$.

Remark 1 (Convention for Using a Student's t Distribution Table). If the degrees of freedom $d.f.$ you need are not in the table, use the closest $d.f.$ in the table that is smaller. This procedure results in a critical value t_c that is more conservative, in the sense that it is larger. The resulting confidence interval will be longer and have a probability that is slightly higher than c.

Example 1. Use Table 4 of the Appendix to find the critical value t_c for a 0.99 confidence level for a t distribution with sample size $n = 5$.

Example 2. Use Table 4 of the Appendix to find t_c for a 0.90 confidence level for a t distribution with sample size $n = 9$.

Remark 2. Let x be a random variable. Obtain a simple random sample (of size n) of x values and compute the sample mean $\bar{x}$ and the sample standard deviation s. If x has an approximately normal distribution or simply a mound-shaped symmetric distribution, then the confidence interval for μ is

$$\bar{x} - E < \mu < \bar{x} + E$$

where $\bar{x}$ = sample mean of a simple random sample and

$E = t_c \frac{s}{\sqrt{n}}$

c = confidence level $(0 < c < 1)$

t_c = critical value for confidence level c and degrees of freedom $d.f. = n - 1$.

Example 3. Suppose an archaeologist discovers seven fossil skeletons from a previously unknown species of miniature horse. Reconstructions of the skeletons of these seven miniature horses show the shoulder heights (in centimeters) to be

45.3 47.1 44.2 46.8 46.5 45.5 47.6

For these sample data, the mean is $\bar{x} = 46.14$ and the sample standard deviation is $s \approx 1.19$. Let μ be the mean shoulder height (in centimeters) for this entire species of miniature horse, and assume that the population of shoulder heights is approximately normal.

Find a 99% confidence interval for μ, the mean shoulder height of the entire population of such horses.

Example 4. A company has a new process for manufacturing large artificial sapphires. In a trial run, 37 sapphires are produced. The distribution of weights is mound-shaped and symmetric. The mean weight for these 37 gems is $\bar{x} = 6.75$ carats, and the sample standard deviation is $s = 0.33$ carats. Let μ be the mean weight for the distribution of all sapphires produced by the new process.

(a) Is it appropriate to use a Student's t distribution to compute a confidence interval for μ?

(b) What is $d.f.$ for this setting?

(c) Use Table 4 of the Appendix to find $t_{0.95}$. Note that $d.f. = 36$ is not in the table. Use the $d.f.$ closest to 36 that is smaller than 36.

(d) Find E.

(e) Find a 95% confidence interval for μ.

(f) What does the confidence interval tell us in the context of the problem?

8.3 Estimating p in the Binomial Distribution

Definition 8.3.1. The point estimates for p and q are

$$\hat{p} = \frac{r}{n}$$
$$\hat{q} = 1 - \hat{p}$$

where n = number of trials and r = number of successes.

Remark 1. Consider a binomial experiment with n trials, where p represents the population probability of success on a single trial and $q = 1 - p$ represents the population probability of failure. Let r be a random variable that represents the number of successes out of the n binomial trials. The number of trials n should be sufficiently large so that both $n\hat{p} > 5$ and $n\hat{q} > 5$. Then the confidence interval for p is

$$\hat{p} - E < p < \hat{p} + E$$

where

$$E \approx z_c\sqrt{\frac{\hat{p}\hat{q}}{n}} = z_c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

c = confidence level $(0 < c < 1)$

$z_c =$ critical value for confidence level c based on the standard normal distribution.

Example 1. Let's return to our flu shot experiment described at the beginning of this section. Suppose that 800 students were selected at random from a student body of 20,000 and given shots to prevent a certain type of flu. All 800 students were exposed to the flu, and 600 of them did not get the flu. Let p represent the probability that the shot will be successful for any single student selected at random from the entire population of 20,000. Let q be the probability that the shot is not successful.

(a) What is the number of trials n? What is the value of r?

(b) What are the point estimates for p and q?

(c) Would it seem that the number of trials is large enough to justify a normal approximation to the binomial?

(d) Find a 99% confidence interval for p.

Example 2. A random sample of 188 students at a large dormitory cafeteria showed that 66 chose the vegetarian option as their meal plan. Let p represent the proportion of students selecting the vegetarian option as their meal plan.

(a) What is a point estimate for p?

(b) Find a 90% confidence interval for p.

(c) What does the confidence interval you just computed mean in the context of this application?

(d) To compute the confidence interval, we used a normal approximation. Does this seem justified?

Remark 2 (General Interpretation of Poll Results).

1. When a poll states the results of a survey, the proportion reported to respond in the designated manner is $\hat{p}$, the sample estimate of the population proportion.

2. The margin of error is the maximal error E of a 95% confidence interval for p.

3. A 95% confidence interval for the population proportion p is

poll report $\hat{p}$ $-$ margin of error $E < p <$ poll report $\hat{p}$ $+$ margin of error E.

Example 3. An online website published a news poll based on nationwide interviews of 2108 adults conducted over the weekend by a reputable polling organization.

The sample was drawn from 315 randomly selected geographic points in the continental United States. Each region was represented in proportion to its population. Households were selected by a method that gave all citizens an equal chance of being included.

One adult, 18 years or older, was selected from each household by a procedure to provide an appropriate number of respondents to represent the population.

Chances are 19 of 20 that if all adults in the United States had been surveyed, the findings would differ from these poll results by no more than 2.6 percentage points in either direction.

(a) What confidence level corresponds to the phrase "chances are 19 out of 20 that if . . . "?

(b) The complete article indicates that everyone in the sample was asked the question "Which party, the Democratic Party or the Republican Party, do you think would do a better job handling ... climate change?" Possible responses were "Democrats," "neither," "both," or "Republicans." The poll reported that 56% of the respondents said, "Democrats." Does 56% represent the sample statistic $\hat{p}$ or the population parameter p for the proportion of adults responding, "Democrats"?

(c) Continue reading the last paragraph of the article. It goes on to state, "...if all adults in the U.S. had been surveyed, the findings would differ from these poll results by no more than 2.6 percentage points in either direction." Use this information, together with parts (a) and (b), to find a 95% confidence interval for the proportion p of the specified population who responded, "Democrats" to the question.

Remark 3. To find the sample size n for estimating a proportion p take

$$n = p(1-p)\left(\frac{z_c}{E}\right)^2 \text{ if you have a preliminary estimate for } p,$$

$$n = \frac{1}{4}\left(\frac{z_c}{E}\right)^2 \text{ if you do } \textit{not} \text{ have a preliminary estimate for } p,$$

where

$E =$ specified maximal error of estimate

$z_c =$ critical value from the normal distribution for the desired confidence level c.

If n is not a whole number, increase n to the next higher whole number. Also, if necessary, increase the sample size n to ensure that both $np > 5$ and $nq > 5$ to satisfy the requirements for a confidence interval. Note that n is the minimal sample size for a specified confidence level and maximal error of estimate.

Example 4. A company is in the business of selling wholesale popcorn to grocery stores. The company buys directly from farmers. A buyer for the company is examining a large amount of corn from a certain farmer. Before the purchase is made, the buyer wants to estimate p, the probability that a kernel will pop.

Suppose a random sample of n kernels is taken and r of these kernels pop. The buyer wants to be 95% sure that the point estimate $\hat{p} = r/n$ for p will be in error either way by less than 0.01.

(a) If no preliminary study is made to estimate p, how large a sample should the buyer use?

(b) A preliminary study showed that p was approximately 0.86. If the buyer uses the results of the preliminary study, how large a sample should they use?

Chapter 9

Hypothesis Testing

9.1 Introduction to Statistical Tests

Definition 9.1.1.
Null hypothesis H_0: This is the statement that is under investigation or being tested. When conducting hypothesis testing, this is assumed to be true until proven otherwise by the data. Usually the null hypothesis represents a statement of "no effect," "no difference," or, put another way, "things haven't changed."
Alternate hypothesis H_1: This is the statement you will adopt in the situation in which the evidence (data) is so strong that you reject H_0. A statistical test is designed to assess the strength of the evidence (data) against the null hypothesis.

Example 1. A car manufacturer advertises that its new hybrid models get 57 miles per gallon (mpg). Let μ be the mean of the mileage distribution for these cars. You assume that the manufacturer will not underrate the car, but you suspect that the mileage might be overrated.

(a) What shall we use for H_0?

(b) What shall we use for H_1?

Example 2. A cell phone manufacturer claimed that the average battery life on their new smart phone on a full charge can last 12.75 hours. To check whether the claim about the average battery life is correct, the company formulates a statistical test.

(a) What should be used for H_0?

(b) What should be used for H_1?

Definition 9.1.2. A statistical test is:

left-tailed if H_1 states that the parameter is less than the value claimed in H_0.

right-tailed if H_1 states that the parameter is greater than the value claimed in H_0.

two-tailed if H_1 states that the parameter is different from (or not equal to) the value claimed in H_0.

Example 3. Rosie is an aging sheep dog in Montana who gets regular check-ups from her owner, the local veterinarian. Let x be a random variable that represents Rosie's resting heart rate (in beats per minute). From past experience, the vet knows that x has a normal distribution with $\sigma = 12$. The vet checked the Merck Veterinary Manual and found that for dogs of this breed, $\mu = 115$ beats per minute.

Over the past 6 weeks, Rosie's heart rate (beats/min) measured

93 109 110 89 112 117

The sample mean is $\bar{x} = 105$. The vet is concerned that Rosie's heart rate may be slowing. Do the data indicate that this is the case?

(a) Establish the null and alternate hypotheses.

(b) Are the observed sample data compatible with the null hypothesis?

(c) How do we compute the probability in part (b)?

(d) What conclusion can be drawn about Rosie's average heart rate?

(e) Have we proved $H_0 : \mu = 115$ to be false and $H_1 : \mu < 115$ to be true?

Definition 9.1.3. Assuming H_0 is true, the probability that the test statistic will take on values as extreme as, or more extreme than, the observed test statistic (computed from sample data) is called the P-value of the test. The smaller the P-value computed from sample data, the stronger the evidence against H_0.

Definition 9.1.4. The level of significance α is the probability of rejecting H_0 when it is true. This is the probability of a Type I error.
The *probability of making a Type II error* is denoted by β.
The quantity $1 - \beta$ is called the power of a test and represents the probability of rejecting H_0 when it is, in fact, false.

Example 4. Let's consider Example 2, in which we were considering the battery life on a new smart phone. The hypotheses were

$$H_0 : \mu = 12.75 \text{ hours} \qquad H_1 : \mu \neq 12.75 \text{ hours}$$

(a) Suppose the manufacturer requires a 1% level of significance. Describe a Type I error, its consequence, and its probability.

(b) Discuss a Type II error and its consequences.

Remark 1.
If P-value $\leq \alpha$, we reject the null hypothesis and say the data are statistically significant at the level α.
If P-value $> \alpha$, we do not reject the null hypothesis.

Example 5. The Environmental Protection Agency has been studying Miller Creek regarding ammonia nitrogen concentration. For many years, the concentration has been 2.3 mg/L. However, a new golf course and new housing developments are raising concern that the concentration may have changed because of lawn fertilizer. Any change (either an increase or a decrease) in the ammonia nitrogen concentration can affect plant and animal life in and around the creek (Reference: EPA Report 832-R-93-005). Let x be a random variable representing ammonia nitrogen concentration (in mg/L). Based on recent studies of Miller Creek, we may assume that x has a normal distribution with $\sigma = 0.30$. Recently, a random sample of eight water tests from the creek gave the following x values.

2.1 2.5 2.2 2.8 3.0 2.2 2.4 2.9

The sample mean is $\bar{x} = 2.51$.

Let us construct a statistical test to examine the claim that the concentration of ammonia nitrogen has changed from 2.3 mg/L. Use level of significance $\alpha = 0.01$.

(a) What is the null hypothesis? What is the alternate hypothesis? What is the level of significance α?

(b) Is this a right-tailed, left-tailed, or two-tailed test?

(c) What sampling distribution shall we use? Note that the value of μ is given in the null hypothesis, H_0.

(d) What is the value of the sample test statistic? Convert the sample mean $\bar{x}$ to a standard z value.

(e) Draw a sketch showing the P-value area on the standard normal distribution. Find the P-value.

(f) Compare the level of significance α and the P-value. What is your conclusion?

(g) Interpret your results in the context of this problem.

9.2 Testing the Mean μ

Example 1. Sunspots have been observed for many centuries. Records of sunspots from ancient Persian and Chinese astronomers go back thousands of years. Some archaeologists think sunspot activity may somehow be related to prolonged periods of drought in the southwestern United States. Let x be a random variable representing the average number of sunspots observed in a 4-week period. A random sample of 40 such periods from Spanish colonial times gave the following data (Reference: M. Waldmeir, *Sun Spot Activity*, International Astronomical Union Bulletin).

12.5	14.1	37.6	48.3	67.3	70.0	43.8	56.5	59.7	24.0
12.0	27.4	53.5	73.9	104.0	54.6	4.4	177.3	70.1	54.0
28.9	13.0	6.5	134.7	114.0	72.7	81.2	24.1	20.4	13.3
9.4	25.7	47.8	50.0	45.3	61.0	39.0	12.0	7.2	11.3

The sample mean is $\bar{x} \approx 47.0$. Previous studies of sunspot activity during this period indicate that $\sigma = 35$. It is thought that for thousands of years, the mean number of sunspots per 4-week period was about $\mu = 41$. Sunspot activity above this level may (or may not) be linked to gradual climate change. Do the data indicate that the mean sunspot activity during the Spanish colonial period was higher than 41? Use $\alpha = 0.05$.

(a) Establish the null and alternate hypotheses.

(b) What distribution do we use for the sample test statistic? Compute the z value of the sample test statistic $\bar{x}$.

(c) Find the P-value of the test statistic.

(d) Conclude the test.

(e) Interpret the results in the context of the problem.

Example 2. The drug 6-mP (6-mercaptopurine) is used to treat leukemia. The following data represent the remission times (in weeks) for a random sample of 21 patients using 6-mP (Reference: E. A. Gehan, University of Texas Cancer Center).

10	7	32	23	22	6	16	34	32	25	11
20	19	6	17	35	6	13	9	6	10	

The sample mean is $\bar{x} \approx 17.1$ weeks, with sample standard deviation $s \approx 10.0$. Let x be a random variable representing the remission time (in weeks) for all patients using 6-mP. Assume the x distribution is mound-shaped and symmetric. A previously used drug treatment had a mean remission time of $\mu = 12.5$ weeks. Do the data indicate that the mean remission time using the drug 6-mP is different (either way) from 12.5 weeks? Use $\alpha = 0.01$.

(a) Establish the null and alternate hypotheses.

(b) What distribution do we use for the sample test statistic $\bar{x}$? Compute the sample test statistic $\bar{x}$ and the corresponding t value.

(c) Find the P-value or the interval containing the P-value.

(d) Conclude the test.

(e) Interpret the results in the context of the problem.

Example 3. Archaeologists become excited when they find an anomaly in discovered artifacts. The anomaly may (or may not) indicate a new trading region or a new method of craftsmanship. Suppose the lengths of projectile points (arrowheads) at a certain archaeological site have mean length $\mu = 2.6$ cm. A random sample of 61 recently discovered projectile points in an adjacent cliff dwelling gave the following lengths (in cm) (Reference: A. Woosley and A. Mclntyre, *Mimbres Mogollon Archaeology,* University of New Mexico Press).

3.1	4.1	1.8	2.1	2.2	1.3	1.7	3.0	3.7	2.3	2.6	2.2	2.8	3.0
3.2	3.3	2.4	2.8	2.8	2.9	2.9	2.2	2.4	2.1	3.4	3.1	1.6	3.1
3.5	2.3	3.1	2.7	2.1	2.0	4.8	1.9	3.9	2.0	5.2	2.2	2.6	1.9
4.0	3.0	3.4	4.2	2.4	3.5	3.1	3.7	3.7	2.9	2.6	3.6	3.9	3.5
1.9	4.0	4.0	4.6	1.9									

The sample mean is $\bar{x} \approx 2.92$ cm and the sample standard deviation is $s \approx 0.85$, where x is a random variable that represents the lengths (in cm) of all projectile points found at the adjacent cliff dwelling site. Do these data indicate that the mean length of projectile points in the adjacent cliff dwelling is longer than 2.6 cm? Use a 1% level of significance.

(a) State H_0, H_1, and α.

(b) What sampling distribution should you use for $\bar{x}$? What is the t value of the *sample test statistic*?

(c) When you use Table 4, of the Appendix, to find an interval containing the P-value, do you use one-tail or two-tail areas? Why? Sketch a figure showing the P-value. Find an interval containing the P-value.

(d) Do we reject or fail to reject H_0?

(e) Interpret your results in the context of the application.

Example 4. Consider Example 1 regarding sunspots. Let x be a random variable representing the number of sunspots observed in a 4-week period. A random sample of 40 such periods from Spanish colonial times gave the number of sunspots per period. The raw data are given in Example 1. The sample mean is $\bar{x} \approx 47.0$. Previous studies indicate that for this period, $\sigma = 35$. It is thought that for thousands of years, the mean number of sunspots per 4-week period was about $\mu = 41$. Do the data indicate that the mean sunspot activity during the Spanish colonial period was higher than 41? Use $\alpha = 0.05$.

(a) Set the null and alternate hypotheses.

(b) Compute the z value of the sample test statistic.

(c) Determine the critical region and critical value based on H_1 and $\alpha = 0.05$.

(d) Conclude the test.

(e) Interpret the results in the context of the application.

(f) How do results of the critical region method compare to the results of the P-value method for a 5% level of significance?

9.3 Testing a Proportion p

Remark 1. For tests of proportions, we convert the sample test statistic $\hat{p}$ to a z value using

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

where

$\hat{p} = r/n$ is the sample test statistic and r is the number of successes
$n =$ number of trials
$p =$ proportion specified in H_0
$q = 1 - p$
$np > 5$ and $nq > 5$.

Example 1. A team of eye surgeons has developed a new technique for a risky eye operation to restore the sight of people blinded from a certain disease. Under the old method, it is known that only 30% of the patients who undergo this operation recover their eyesight.
Suppose that surgeons in various hospitals have performed a total of 225 operations using the new method and that 88 have been successful (i.e., the patients fully recovered their sight). Can we justify the claim that the new method is better than the old one? (Use a 1% level of significance.)

(a) Establish H_0 and H_1 and note the level of significance.

(b) Is the sample sufficiently large to justify use of the normal distribution for $\hat{p}$? Find the sample test statistic $\hat{p}$ and convert it to a z value, if appropriate.

(c) Find the P-value of the test statistic.

(d) Conclude the test.

(e) Interpret the results in the context of the problem.

Example 2. A botanist has produced a new variety of hybrid wheat that is better able to withstand drought than other varieties. The botanist knows that for the parent plants, the proportion of seeds germinating is 80%. The proportion of seeds germinating for the hybrid variety is unknown, but the botanist claims that it is 80%. To test this claim, 400 seeds from the hybrid plant are tested, and it is found that 312 germinate. Use a 5% level of significance to test the claim that the proportion germinating for the hybrid is 80%.

(a) Let p be the proportion of hybrid seeds that will germinate. Notice that we have no prior knowledge about the germination proportion for the hybrid plant. State H_0 and H_1. What is the required level of significance?

(b) Using the value of p in H_0, are both $np > 5$ and $nq > 5$? Can we use the normal distribution for $\hat{p}$?

(c) Calculate the sample test statistic $\hat{p}$.

(d) Next, we convert the sample test statistic $\hat{p} = 0.78$ to a z value. Based on our choice for H_0, what value should we use for p in our formula? Since $q = 1 - p$, what value should we use for q? Using these values for p and q, convert $\hat{p}$ to a z value.

(e) Is the test right-tailed, left-tailed, or two-tailed? Find the P-value of the sample test statistic and sketch a standard normal curve showing the P-value.

(f) Do we reject or fail to reject H_0?

(g) Interpret your conclusion in the context of the application.

Example 3. Let's solve Example 2 using the critical region approach. In that problem, 312 of 400 seeds from a hybrid wheat variety germinated. For the parent plants, the proportion of germinating seeds is 80%. Use a 5% level of significance to test the claim that the population proportion of germinating seeds from the hybrid wheat is different from that of the parent plants.

(a) Find the sample test statistic $\hat{p}$ and the corresponding z value.

(b) Find the critical value z_0.

(c) Find the critical regions and the location of the sample test statistic.

(d) Conclude the test and compare the results to Example 2.

Chapter 10

Inferences about Differences

10.1 Tests Involving Paired Differences (Dependent Samples)

Example 1. A shoe manufacturer claims that among the general population of adults in the United States, the average length of the left foot is longer than that of the right. To compare the average length of the left foot with that of the right, we can take a random sample of 15 U.S. adults and measure the length of the left foot and then the length of the right foot for each person in the sample. Is there a natural way of pairing the measurements? How many pairs will we have?

Example 2. The Instrumental Enrichment Program is a systematic approach to learning that was developed to enhance the cognitive functions necessary for academic achievement. It is based on the belief that intelligence is dynamic and changeable, rather than fixed. To test the program, extensive statistical tests were conducted. In one experiment, a random sample of 10-year-old students with IQ scores below 80 was selected. An IQ test was given to these students before they spent 2 years in an IE Program, and an IQ test was given to the same students after the program.

(a) On what basis can you pair the IQ scores?

(b) If there were 20 students in the sample, how many data pairs would you have?

Theorem 10.1.1. *Consider a random sample of n data pairs. Suppose the differences d between the first and second members of each data pair are (approximately) normally distributed, with population mean μ_d. Then the t values*

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$$

where $\bar{d}$ is the sample mean of the d values, n is the number of data pairs, and

$$s_d = \sqrt{\frac{\sum(d - \bar{d})^2}{n-1}}$$

is the sample standard deviation of the d values, follow a Student's t distribution with degrees of freedom $d.f. = n - 1$.

Example 3. A team of heart surgeons at Saint Ann's Hospital knows that many patients who undergo corrective heart surgery have a dangerous buildup of anxiety before their scheduled operations. The staff psychiatrist at the hospital has started a new counseling program intended to reduce this anxiety. A test of anxiety is given to patients who know they must undergo heart surgery. Then each patient participates in a series of counseling sessions with the staff psychiatrist. At the end of the counseling sessions, each patient is retested to determine anxiety level. Table 10.1, indicates the results for a random sample of nine patients. Higher scores mean higher levels of anxiety. Assume the distribution of differences is mound-shaped and symmetric.

From the given data, can we conclude that the counseling sessions reduce anxiety? Use a 0.01 level of significance.

(a) Note the level of significance and set the hypotheses.

(b) Is it appropriate to use a Student's t distribution for the sample test statistic? Explain. What degrees of freedom are used?

Table 10.1

	B	A	
Patient	Score Before Counseling	Score After Counseling	$d = B - A$ Difference
Jan	121	76	45
Tom	93	93	0
Diane	105	64	41
Barbara	115	117	-2
Mike	130	82	48
Bill	98	80	18
Frank	142	79	63
Carol	118	67	51
Alice	125	89	36

(c) Find the sample test statistic $\bar{d}$ and convert it to a corresponding test statistic t.

(d) Find the P-value for the test statistic and sketch the P-value on the t distribution.

(e) Conclude the test.

(f) Interpret the results in the context of the application.

Example 4. Do educational toys make a difference in the age at which a child learns to read? To study this question, researchers designed an experiment in which one group of preschool children spent 2 hours each day (for 6 months) in a room well supplied with "educational" toys such as alphabet blocks, puzzles, ABC readers, coloring books featuring letters, and so forth. A control group of children spent 2 hours a day for 6 months in a "noneducational" toy room. It was anticipated that IQ differences and home environment might be uncontrollable factors unless identical twins could be used. Therefore, six pairs of identical twins of preschool age were randomly selected. From each pair, one member was randomly selected to participate in the experimental (i.e., educational toy room) group and the other in the control (i.e., noneducational toy room) group. For each twin, the data item recorded is the age in months at which the child began reading at the primary level (Table 10.2). Assume the distribution of differences is mound-shaped and symmetric.

Table 10.2: Reading Ages for Identical Twins (in Months)

Twin Pair	Experimental Group B = Reading Age	Control Group A = Reading Age	Difference $d = B - A$
1	58	60	
2	61	64	
3	53	52	
4	60	65	
5	71	75	
6	62	63	

(a) Compute the entries in the $d = B - A$ column of Table 10.2. Using formulas for the mean and sample standard deviation or a calculator with mean and sample standard deviation keys, compute $\bar{d}$ and s_d.

(b) What is the null hypothesis?

(c) To test the claim that the experimental group learned to read at a *different age* (either younger or older), what should the alternate hypothesis be?

(d) What distribution does the sample test statistic $\bar{d}$ follow? Find the degrees of freedom.

(e) Convert the sample test statistic $\bar{d}$ to a t value.

(f) When we use Table 4 of the Appendix to find an interval containing the P-value, do we use one-tail or two-tail areas? Why? Sketch a figure showing the P-value. Find an interval containing the P-value.

(g) Using $\alpha = 0.05$, do we reject or fail to reject H_0?

(h) What do the results mean in the context of this application?

Example 5. Let's revisit Example 4 regarding educational toys and reading age and conclude the test using the critical region method. Recall that there were six pairs of twins. One twin of each set was given educational toys and the other was not. The difference d in reading ages for each pair of twins was measured, and $\alpha = 0.05$.

(a) Find the critical values for $\alpha = 0.05$.

(b) Sketch the critical regions and place the t value of the sample test statistic $\bar{d}$ on the sketch. Conclude the test. Compare the result to the result given by the P-value method of Example 4.

10.2 Inferences about the Difference of Two Means $\mu_1 - \mu_2$

Definition 10.2.1. Two samples are dependent if each data value in one sample can be paired with a corresponding data value in the other sample.
Two samples are independent if the selection of sample data from one population is completely unrelated to the selection of sample data from the other population.

Example 1. For each experiment, categorize the sampling as independent or dependent, and explain your choice.

1. In many medical experiments, a sample of subjects is randomly divided into two groups. One group is given a specific treatment, and the other group is given a placebo. After a certain period of time, both groups are measured for the same condition. Do the measurements from these two groups constitute independent or dependent samples?

2. In an accountability study, a group of students in an English composition course is given a pretest. After the course, the same students are given a posttest covering similar material. Are the two groups of scores independent or dependent?

Theorem 10.2.1. *Let x_1 and x_2 have normal distributions with means μ_1 and μ_2 and standard deviations σ_1 and σ_2, respectively. If we take independent random samples of size n_1 from the x_1 distribution and of size n_2 from the x_2 distribution, then the variable $\bar{x}_1 - \bar{x}_2$ has*

1. *a normal distribution*
2. *a mean $\mu_1 - \mu_2$*
3. *a standard deviation $\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$*

Remark 1. When testing the difference of means, it is customary to use the null hypothesis

$$H_0 : \mu_1 - \mu_2 = 0 \text{ or, equivalently, } H_0 : \mu_1 = \mu_2.$$

Example 2. A consumer group is testing camp stoves. To test the heating capacity of a stove, it measures the time required to bring 2 quarts of water from 50°F to boiling (at sea level). Two competing models are under consideration. Ten stoves of the first model and 12 stoves of the second model are selected at random and tested. The following results are obtained:

Model 1: Mean time $\bar{x}_1 = 11.4$ min; $\sigma_1 = 2.5$ min; $n_1 = 10$
Model 2: Mean time $\bar{x}_2 = 9.9$ min; $\sigma_2 = 3.0$ min; $n_2 = 12$.

Assume that the time required to bring water to a boil is normally distributed for each stove.
Hypothesis test: Is there any difference (either way) between the performances of these two models? Use a 5% level of significance.

(a) State the null and alternate hypotheses and note the value of α.

(b) What distribution does the sample test statistic follow?

(c) Compute the sample test statistic $\bar{x}_1 - \bar{x}_2$ and then convert it to a z value.

(d) Find the P-value and sketch the area on the standard normal curve.

(e) Conclude the test.

(f) Interpret the results.

Confidence interval: Find a 95% confidence interval for the population difference $\mu_1 - \mu_2$ of mean times to boil water for the two stoves. Interpret the results.

Example 3.

(a) Suppose a study was conducted to compare the difference in average income (in millions) between Hollywood stars (μ_1) and YouTube stars (μ_2). A 90% confidence interval for the difference of means is

$$10 < \mu_1 - \mu_2 < 20.$$

For this interval, what can we conclude about the respective values of μ_1 and μ_2?

(b) Suppose a study was conducted to compare the difference in average income (in millions) between TikTok stars (μ_1) and YouTube stars (μ_2). A 95% confidence interval for the difference of means is

$$-0.32 < \mu_1 - \mu_2 < 0.16.$$

For this interval, what can we conclude about the respective values of μ_1 and μ_2?

Example 4. Two competing headache remedies claim to give fast-acting relief. An experiment was performed to compare the mean lengths of time required for bodily absorption of brand A and brand B headache remedies.

Twelve people were randomly selected and given an oral dose of brand A. Another 12 were randomly selected and given an equal dose of brand B. The lengths of time in minutes for the drugs to reach a specified level in the blood were recorded. The means, standard deviations, and sizes of the two samples follow.

$$\text{Brand A: } \bar{x}_1 = 21.8 \text{ min}; \; s_1 = 8.7 \text{ min}; \; n_1 = 12$$
$$\text{Brand B: } \bar{x}_2 = 18.9 \text{ min}; \; s_2 = 7.5 \text{ min}; \; n_2 = 12.$$

Past experience with the drug composition of the two remedies permits researchers to assume that both distributions are approximately normal.
Hypothesis test: Use a 5% level of significance to test the claim that there is no difference in the mean time required for bodily absorption. Also, find or estimate the P-value of the sample test statistic.

(a) State the null and alternate hypotheses and note the value of α.

(b) What distribution does the sample test statistic follow?

(c) Compute the t value of the sample test statistic.

(d) Estimate the P-value and sketch the area on a t graph.

(e) Conclude the test.

(f) Interpret the results.

Confidence interval: Find a 95% confidence interval for the population difference $\mu_1 - \mu_2$ of mean times for the competing drugs to reach a specified level in the blood. Interpret the results.

Example 5. Suppose the experiment to measure the time in minutes for the headache remedies to enter the bloodstream (Example 4) yielded sample means, sample standard deviations, and sample sizes as follows:

$$\text{Brand A: } \bar{x}_1 = 20.1 \text{ min; } s_1 = 8.7 \text{ min; } n_1 = 12$$
$$\text{Brand B: } \bar{x}_2 = 11.2 \text{ min; } s_2 = 7.5 \text{ min; } n_2 = 8.$$

Brand B claims to be faster.
Hypothesis test: Is this claim justified at the 5% level of significance?

(a) What is α? State H_0 and H_1.

(b) What distribution does the sample test statistic follow? Explain.

(c) Compute the sample test statistic $\bar{x}_1 - \bar{x_2}$ and convert it to a t value.

(d) What degrees of freedom do you use? To find an interval containing the P-value, do you use one-tail or two-tail areas in Table 4 of the Appendix? Sketch a figure showing the P-value. Find an interval for the P-value.

(e) Do we reject or fail to reject H_0?

(f) Interpret the results in the context of the application.

Confidence interval: Find a 90% confidence interval for the difference of population means $\mu_1 - \mu_2$ of times for the two competing headache remedies to enter the bloodstream.

(g) Find the degrees of freedom and the critical value $t_{0.90}$.

(h) Find the maximal error of estimate E for a 90% confidence interval.

(i) Find the sample difference $\bar{x}_1 - \bar{x}_2$ and a 90% confidence interval for $\mu_1 - \mu_2$.

(j) Interpret the results in the context of the application.

Example 6. Use the critical region method to solve the application in Example 4 (test $\mu_1 - \mu_2$ when σ_1 and σ_2 are unknown).

10.3 Inferences about the Difference of Two Proportions $p_1 - p_2$

Theorem 10.3.1. *Suppose we have two independent binomial experiments—that is, outcomes from one binomial experiment are in no way paired with outcomes from the other. We use the notation*

Binomial Experiment 1	*Binomial Experiment 2*
$n_1 =$ *number of trials*	$n_2 =$ *number of trials*
$r_1 =$ *number of successes*	$r_2 =$ *number of successes*
$p_1 =$ *population probability of success on a single trial*	$p_2 =$ *population probability of success on a single trial*

For large values of n_1 and n_2, the distribution of sample differences

$$\hat{p}_1 - \hat{p}_2 = \frac{r_1}{n_1} - \frac{r_2}{n_2}$$

is closely approximated by a normal distribution with mean μ and standard deviation σ as shown:

$$\mu = p_1 - p_2 \qquad \sigma = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

where $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$.

Remark 1. The pooled estimate of proportion $\bar{p}$ given by

$$\bar{p} = \frac{r_1 + r_2}{n_1 + n_2}$$

gives the best sample estimate for p_1 and p_2 under the assumption that $p_1 = p_2$. Also, $\bar{q} = 1 - \bar{p}$.

Example 1. The county clerk in your area wishes to improve voter registration. One method under consideration is to send reminders in the mail to all citizens in the county who are eligible to register. As part of a pilot study to determine if this method will actually improve voter registration, a random sample of 1250 potential voters was taken. This sample was then randomly divided into two groups.

Group 1: There were 625 people in this group. No reminders to register were sent to them. The number of potential voters from this group who registered was 295.

> *Group 2:* This group also contained 625 people. Reminders were sent in the mail to each member in the group, and the number who registered to vote was 350.

The county clerk claims that the proportion of people who registered was significantly greater in group 2. On the basis of this claim, the clerk recommends that the project be funded for the entire population. Use a 5% level of significance to test the claim that the proportion of potential voters who registered was greater in group 2, the group that received reminders.

(a) What is α? State H_0 and H_1.

(b) Calculate the *pooled estimates* $\bar{p}$ and $\bar{q}$.

(c) What distribution does the sample test statistic follow?

(d) Compute the sample test statistic $\hat{p}_1 - \hat{p}_2$, and convert it to a z value.

(e) Find the P-value and sketch the area on the standard normal curve.

(f) Conclude the test.

(g) Interpret the results in the context of the application.

(h) Use the critical region method to conclude the test at the 5% level of significance. Compare your results with the P-value method.

Example 2. In Example 1 about voter registration, suppose that a random sample of 1100 potential voters was randomly divided into two groups.

> *Group 1:* 500 potential voters; no registration reminders sent; 248 registered to vote
>
> *Group 2:* 600 potential voters; registration reminders sent; 332 registered to vote

Do these data support the claim that the proportion of voters who registered was greater in the group that received reminders than in the group that did not? Use a 1% level of significance.

(a) What is α? State H_0 and H_1.

(b) Under the null hypothesis $p_1 = p_2$, calculate the *pooled estimates* $\bar{p}$ and $\bar{q}$.

(c) What distribution does the sample test statistic follow?

(d) What is the sample test statistic $\hat{p}_1 - \hat{p}_2$?

(e) Convert sample test statistic $\hat{p}_1 - \hat{p}_2 = -0.057$ to a z value.

(f) Find the P-value and sketch the area on the standard normal curve.

(g) Conclude the test.

(h) Interpret the results in the context of the application.

Example 3. During normal sleep, there is a phase known as *REM* (rapid eye movement). For most people, REM sleep occurs about every 90 minutes or so, and it is thought that dreams occur just before or during the REM phase. Using electronic equipment in the Sleep Laboratory, it is possible to detect the REM phase in a sleeping person. If a person is wakened immediately after the REM phase, they usually can describe a dream that has just taken place. Based on a study of over 650 people in the Zurich Sleep Laboratory, it was found that about one-third of all dream reports contain feelings of fear, anxiety, or aggression. There is a conjecture that if a person is in a good mood when going to sleep, the proportion of "bad" dreams (fear, anxiety, aggression) might be reduced.

Suppose that two groups of subjects were randomly chosen for a sleep study. In group I, before going to sleep, the subjects spent 1 hour watching a comedy movie. In this group, there were a total of $n_1 = 175$ dreams recorded, of which $r_1 = 49$ were dreams with feelings of anxiety, fear, or aggression. In

group II, the subjects did not watch a movie but simply went to sleep. In this group, there were a total of $n_2 = 180$ dreams recorded, of which $r_2 = 63$ were dreams with feelings of anxiety, fear, or aggression.

(a) Why could groups I and II be considered independent binomial distributions? Why do we have a "large-sample" situation?

(b) What is $p_1 - p_2$? Compute a 95% confidence interval for $p_1 - p_2$.

(c) Explain the meaning of the confidence interval that you constructed in part (b).

Index

Bibliography

[1] Charles Henry Brase and Corrinne Pellillo Brase. *Understanding Basic Statistics*. Cengage Learning, 2023. Print.