ELEMENTARY STATISTICS

For Math 1401

at The University of West Georgia

First Edition - 2022

by Jim Bellon

Contents

Introduction 1

1 Descriptive Statistics 3

1.1 Collecting Data . . . 3

1.1.1 Key Terms . . . 3

1.1.2 Sampling . . . 7

1.1.3 Exercises: Collecting Data . . . 12

1.2 Summarizing Data . . . 14

1.2.1 Distributions . . . 14

1.2.2 Graphs of Data . . . 18

1.2.3 Exercises: Summarizing Data . . . 28

1.3 Measuring Data Sets . . . 31

1.3.1 Measures of Center . . . 31

1.3.2 Measures of Variation . . . 36

1.3.3 Exercises: Measuring Data Sets . . . 40

1.4 Measures of Relative Standing 42
1.4.1 Z-scores 42
1.4.2 Percentiles 43
1.4.3 Boxplots 45
1.4.4 Exercises: Measures of Relative Standing 49
1.5 Data Sets with the TI-83 and Similar Calculators 51

2 Probability **58**

2.1 Probability Basics 58
2.1.1 Calculating Probability 58
2.1.2 Odds 64
2.1.3 Exercises: Probability Basics 67
2.2 Counting Rules 69
2.2.1 Combinations 69
2.2.2 Permutations 70
2.2.3 Probabilities and Counting 72
2.2.4 Exercises: Counting Rules 74

2.3 More Probability 75
2.3.1 The Addition Rule 75
2.3.2 Conditional Probability and the Multiplication Rule 77
2.3.3 Venn Diagrams 80
2.3.4 Exercises: More Probability 84

3 Probability Distributions **85**

3.1 Discrete Distributions 85
3.1.1 Discrete Random Variables 85
3.1.2 Expected Value 88
3.1.3 Binomial Distribution 90
3.1.4 Exercises: Discrete Distributions 95
3.2 Continuous Distributions 97
3.2.1 The Uniform Distribution 97
3.2.2 The Normal Distribution 99
3.2.3 Exercises: Continuous Distributions 117
3.3 Sampling Distributions 119
3.3.1 Sampling Distributions 119
3.3.2 The Central Limit Theorem 120
3.3.3 Exercises: Sampling Distributions 125

4 Inference: From Sample to Population **126**

4.1 Confidence Intervals . . . 126

4.1.1 Estimating the Population Mean . . . 126

4.1.2 The T-distribution . . . 129

4.1.3 Estimating the Population Proportion . . . 131

4.1.4 Required Minimum Sample Size . . . 133

4.1.5 Exercises: Confidence Intervals . . . 136

4.2 Hypothesis Testing . . . 137

4.2.1 Z-test for the mean . . . 139

4.2.2 T-test for the mean . . . 145

4.2.3 Z-test for the proportion . . . 148

4.2.4 Exercises: Hypothesis Testing . . . 152

5 Bivariate Data **154**

5.1 Correlation and Regression . . . 154

5.1.1 Correlation . . . 154

5.1.2 Regression . . . 160

5.1.3 Exercises: Correlation and Regression . . . 167

5.2 Joint Distributions . . . 169

5.2.1 Marginal Frequencies . . . 169

5.2.2 Conditional Frequencies . . . 172

5.2.3 Exercises: Joint Distributions . . . 173

6 Solutions **175**

Answers to Try This On Your Own Problems 175

Solutions to Exercises: Sec 1.1 Collecting Data 184

Solutions to Exercises: Sec 1.2 Summarizing Data 186

Solutions to Exercises: Sec 1.3 Measuring Data 191

Solutions to Exercises: Sec 1.4 Measuring of Relative Standing 193

Solutions to Exercises: Sec 2.1 Probability Basics 196

Solutions to Exercises: Sec 2.2 Counting Rules 198

Solutions to Exercises: Sec 2.3 More Probability 200

Solutions to Exercises: Sec 3.1 Discrete Distributions 202

Solutions to Exercises: Sec 3.2 Continuous Distributions 205

Solutions to Exercises: Sec 3.3 Sampling Distributions 213

Solutions to Exercises: Sec 4.1 Confidence Intervals 215

Solutions to Exercises: Sec 4.2 Hypothesis Tests 216

Solutions to Exercises: Sec 5.1 Correlation and Regression 219

Solutions to Exercises: Sec 5.2 Joint Distributions 223

Standard Normal Table **225**

Student's T Table **227**

Index **229**

Introduction

Welcome to Elementary Statistics! Just what is Statistics? The simple answer is "the analyzing of data", but you may want a more detailed explanation than that. Data Analysis is a term commonly used for the collection of fundamental concepts from statistics, probability, graphs, measurements, and converting units.

Memorizing formulas and just being able to do problems are not enough to prepare for advanced math classes and college. In order to succeed in mathematics, it is my belief that students must fully understand the concepts, be able to apply them to real world problem solving, and be able to make sense of the results. Here we will focus on not just the what and when, but also the why, how, and where is it used. Most students who do not succeed in math or dislike math, probably were told: "Just do it the way I told you, get through it quickly and move on. Don't worry about understanding it". I have seen many students with a fear and dislike of math, who have succeeded and actually started to like it, once they are shown the "big picture" and how it all comes together.

The main excuse I hear for failing is "I won't ever need to know this or use this, so why should I learn this?" The benefit of learning mathematics is not necessarily to obtain particular knowledge about certain math topics. Retaining the knowledge is important, if you have to go on to the next level. However, even for those who just need to pass one core math course, there is still plenty to gain. Learning mathematics trains your brain to think logically and develops analytical skills, which can be used for the rest of your life. Some examples include doing your taxes, running a business, managing people or projects, building a tree-house for your kids, etc.

Think of it this way: professional basketball players are mainly concerned with getting the ball into a basket, so why do they lift weights (the ball is not heavy), run laps (the court is only 94 feet long), and analyze film. Because these things help train them for being better

at what they want to do. The same can be said about learning math. It helps train your brain to be better at dealing with the real world, life, careers, etc. So even if you do not care to know the math itself, learning math can be a vehicle for increasing your brainpower and critical thinking skills.

Here are some tips on how to use this book and what else you can do to succeed in this course. These ideas are not my discovery and can be applied to almost any course.

- Read each section of the book BEFORE you cover it in class. Let your brain mull it over so class time will seem like review or at least let you get a better grip on the material.
- TAKE GOOD NOTES!! I have noticed that many students (especially home-schoolers) do not take notes. Just because you have a textbook, does not mean that you should not write out your thoughts in your own words.
- Do ALL of the exercises, practice makes perfect (or at least closer to perfect).
- Work the problems and think them over BEFORE you look at the solutions.
- Start working on assignments and studying early, this way you avoid "something else came up, I couldn't finish".
- Many test problems will be similar to exercises, but some will combine topics and/or be longer. Prepare very thoroughly for tests.
- Don't be afraid to ask questions. Most teachers/tutors take them seriously. Questions help them to assess where you are BEFORE you're tested.
- Search the internet for one of the many FREE online math help sites.

Chapter 1

Descriptive Statistics

1.1 Collecting Data

1.1.1 Key Terms

In the year 2022 there were several hot topics being reported in the news in the form of data. Some of the numbers out there were the following:

- The median price of houses sold in the US in May 2022 is $428,700. An increase of 20% from April 2021.
- The median household annual income in 2022 is $67,521.
- 65% of homes are owned by the people who live there.
- The average price of gasoline in January 2022 was $3.41, in May 2022 was $4.55 per gallon.
- The US population by race is 60% white, 18% hispanic, 13% black, 6% asian, 3% mixed.

- Total during the Covid-19 pandemic 2020 - June 2022: There were 85,921,461 cases in the US, 1,008,196 deaths, 83% of the population over age 5 have been vaccinated.
- President Biden's approval rating went from a high of 57% after he was elected, to a low of 36% in May 2022.
- From February to June 2022, there were almost 10,000 Ukrainian civilian casualties caused by the Russian Invasion.
- The probability of winning the Megamillions lottery for one drawing was 1 out of 302 million or 0.00000033%

Where do these numbers come from? What do they mean? How accurate are they? In this course, we will look at the basic concepts of probability and statistics and how they can affect our world. If we wish to analyze data, we need to understand what data is. People and things have characteristics which can be observed or measured. A **Variable** is a general type of characteristic (or type of status), which can be different for each person or thing. Some variables are: name, height, color, texture, mood, wingspan, density, anxiety level, etc. **Data** is the collection of all observations for a particular variable or variables, from one or more people or things.

The branch of mathematics that covers the methods and procedures in analyzing data is called **Statistics**. Statistics includes methods for planning studies and experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions based on data. **Descriptive Statistics** consists of collecting, organizing, summarizing, and presenting data. **Inferential Statistics**, involves analyzing, interpreting, and drawing conclusions based on data. We will do both in this course.

Below are some important terminology we will be using in data analysis.

A **Population** is the collection of all individuals or items under consideration in a study.

A **Census** is information (data) obtained from the entire population.

In reality, most large censuses (such as the US national census) are an attempt to collect from the entire population, but being so large, some data is never collected. There are just some people who do not wish to be found and others who are too busy to report their data. The results are often adjusted to be a good approximation of the population information. The US census is so large and takes so much time, money, and staff, that it is done only every ten years.

In many cases, it is usually easier and sufficient to collect data from only some of the elements in the population. A **Sample** is the part of a population from which information is actually collected.

A **Parameter** is a numerical measurement describing some characteristic of a population. Examples: The average starting salary of elementary school teachers in Georgia is $33,673. The average for the whole United States is $35,763.

A **Statistic** is a numerical measurement describing some characteristic of a sample. Example: A survey of ten job postings for elementary school teachers in the Atlanta area, had an average starting salary of $38,541.

****Try this on your own**: For the following scenario, describe the population of interest, describe the sample, state the parameter of interest, and the statistic that was calculated.

A farm wants to track the weight gain of their chickens after they switched to a new feed. The farm has over 10,000 chickens. They isolated 200 chickens and weighed them before the switch, then every week for the next 10 weeks. At the end of 10 weeks, the 200 isolated chickens gained an average of 1.2 pounds.

There are two main types of variables. In this book, we will focus mainly on data from numerical variables, since you can perform calculations with numbers more easily.

Qualitative Variables are variables which have values that are words, symbols, or categories. They can also be numbers that have no absolute measure, order or units. Examples: gender, job title, letter grade, phone number, numbers on football jerseys, etc. These are also referred to as **Categorical Variables**.

Quantitative Variables are variables which have values that are numerical values with a specific order and units. They represent counts or measures. Examples: height, weight, temperature, number of siblings, hours of sleep, etc.

Quantitative variables can be further broken down into two different types. **Discrete** variables have possible values that form a finite set of numbers or values (typically a count). **Continuous** variables have possible values that form an interval of numbers and can be almost anything between (typically a measurement).

Examples: Number of siblings is discrete since the only possible values are 0, 1, 2, 3, ... up to some realistic maximum. Adult height is continuous, since it can be ANY value between 21 inches (shortest ever recorded) to 8 feet 11 inches (tallest ever recorded), including fractional measurements.

Often the purpose of a statistical study is to investigate whether a relationship exists between 2 characteristics, such as smoking and lung cancer, age and income, or stock price and company revenues. We distinguish between 2 types of procedures.

In an **Observational study**, researchers observe characteristics and take measurements but don't attempt to modify the subjects being studied. An example would be a study of animals by simply hiding and watching what they do.

In a **Designed experiment**, researchers impose treatments and controls and then observe characteristics and take measurements. An Experiment can help establish causes. The

subjects are called experimental units. Experiments have **Treatments** , which are the variables that are controlled and changed in order to test the effects. Drug trials are experiments. Different drugs and different doses are given to different groups to see which one is the best.

In experiments, they often use a **Placebo**, which is an inert substance that is used in place of a treatment. It has no direct effect, unknown to the subject. The purpose of a placebo is to eliminate psychological effects of the subjects. In drug tests, they will have a control group who do not get any of the drug. If they know they are not getting drugs, they might feel negative and stressed about not being helped, so the researchers will typically give them a fake pill.

One problem that arises in studies is **Bias** . Bias is when the results systematically favor certain outcomes. To eliminate bias, treatments are assigned **Double Blind** where neither the subjects nor experimenters know who receives which treatment and who gets the placebo until the results are recorded.

1.1.2 Sampling

How do we make sure the design is truly random? Some methods for obtaining a random sample are to pick cards, roll dice, pick out of a hat, use random numbers from software. The method that is most useful in schools, is use of a random number generator from our calculators.

There are several ways to pick the subjects that will be in a sample. **Random Sampling** is using a random method to select the sample from population. This is better than human judgment, but no guarantee of getting a perfect sample. It is important to make sure you have selected a **Representative Sample**, which means that the sample has characteristics similar to the population being studied, in order to avoid having bias.

For example, if you want to know what students at a college think about proposed changes to graduation requirements, you should get a sample that has students from all class levels, different sex/gender, and not all from the same major.

Here is an example of a sample that is not representative. If you need 10 people for a survey, then ask the first ten people you come in contact with. This can often lead to extreme bias. The ten people might be related or friends, and have similar opinions.

****Try this on your own**: Would the following sample be representative of the population? A teacher would like to know how students feel about the new math curriculum. They selected a sample of students from the ones who are failing the class and come for extra help.

Simple random sampling is a sampling procedure for which each possible sample of a given size is equally likely. If the sample is chosen with replacement, then each member of population can be selected more than once. Without replacement means each member of population can be selected only once. We will assume simple random sampling without replacement (unless otherwise specified).

Systematic Sampling is a sampling method that follows a system (pattern) and so is not random, but is easy especially for computers. Divide the population size, N, by the desired sample size n and always round DOWN to the nearest whole number. We will refer to this number as m. Then systematically chose the members of the sample from the population list picking every "m-th" person. So positions are m, 2m, 3m, etc., until you obtain the sample size n.

Example: if we want to select a sample of $n = 15$ people from a population of $N = 1400$ people, then $\frac{N}{n} = \frac{1400}{15} = 93.33$, so $m = 93$. Now the sample members will be every 93rd person from the population list as follows: 93, 186, 279, 372, etc., ending with the 15th

selection, 1395. Notice that if we add 93 again by mistake, we will be over 1400 and there will not be anyone to choose.

In **Cluster Sampling**, the population is divided into clusters (usually based on location). Randomly choose a cluster and use the members to get the sample size. If you need more to complete the sample, randomly choose another cluster and use part or all as needed.

The method that usually has least amount of bias is **Stratified Sampling**, where the population is divided into sub-populations called strata (one is called a stratum). Members within a particular stratum should have common characteristics relative to the statistical study. Then a simple random sample is taken from each stratum in close proportion to the size of the stratum. The strata samples are combined to obtain the overall sample.

Example: If we wish to choose a stratified sample of 11 people, from a population of 78 males and 42 females, then the male sample should be about $\frac{78}{120} = 0.65 = 65\%$. For 11 people, 65% of 11 is 7.15, so randomly choose 7 of the 78 males, then the other 4 people will be randomly chosen from the females.

Convenience Sampling is when we just select the easiest elements to be in the sample. For example, if you need 10 people for a survey, then ask the first ten people you come in contact with. This can often lead to extreme bias. The ten people might be related or friends, and have similar opinions.

****Try this on your own**: What sampling methods would each description below be classified as?

1. A teacher selected a sample of students by selecting one row and picking all students in that row.

2. A researcher was conducting a survey where they selected a sample by going to every tenth neighborhood and surveying every tenth home from those neighborhoods

Whether conducting statistical analysis of data that we have collected, or analyzing a statistical analysis done by someone else, we should not rely on blind acceptance of mathematical calculation. We should consider these factors:

1. Context of the data: What do the values represent? Why were they collected? An understanding of the context will directly affect the statistical procedure used.

2. Source of the data: Is the source objective or biased? Is there something to gain or lose by distorting results? Be vigilant and skeptical of studies from sources that may be biased, such as a nutrition study done by a fast food company.

3. Sampling method: Is the method chosen appropriate and help eliminate bias? Voluntary response(self-chosen) samples often have bias (those with strong opinions are more likely to participate). These sample results are not necessarily valid. Other methods are more likely to produce good results.

4. Conclusions: Make statements that are clear to those without an understanding of statistics and its terminology. Avoid making statements not justified by the statistical analysis.

5. Practical implications: State practical implications of the results. The results may be valid and significant yet there may be NO practical significance. Does anyone even care about it? Common sense might suggest that the finding does not make enough of a difference to justify its use or to be practical.

6. Consider the likelihood of getting the results by chance. If results could easily occur by chance, then they are not statistically significant (you did not justify anything). If the likelihood of getting the results by chance is so small, then the results are statistically significant (you found strong evidence).

7. It is important to carefully plan the study and know what you are trying to show before you do any work. Improper planning can result in a poor or incomplete study. Make sure you have enough resources (time, money, people, supplies) to complete your study and report your results in a professional manner.

1.1.3 Exercises: Collecting Data

Solutions appear at the end of this textbook.

1. Identify the population, sample, parameters, and statistics for the following situation. A textbook company wants to know the average price of homeschool science textbooks in the United States. They obtain a list of 15 science books and compute the average price of the 15 books is $52.

2. A teenager put the following information on his myspace page. What are the variables, what are their values, which variables are qualitative, which are quantitative? Name: Sean Higgins, Ht: 5ft.10in., Wt: 185 lbs., Eyes: Green, Hair: Red, Page-hits: 142

3. Use systematic sampling to select 7 people out of a group of 6700. State the place numbers of the selected sample.

4. Does the following describe an experiment or just an observational study? Explain why. Sarah was on a field trip for her science class. At the beach, she saw a sand castle with 2 crabs crawling inside it. She timed the crabs to see how long they took to find their way out.

5. Which type of sampling best fits the following? Explain why. DJ Paulie D is looking for some songs to be the background beats for his new mix. He sorts his ipod collection into four categories: Rap, Instrumental, Dance, and Alternative. Then he randomly picks 5 songs from each category and listens to the beats.

6. What is a census? The US government does a census every ten years. Why don't they do one every year? Is the US census an actual census? Why or why not?

7. Which sampling methods tend to have bias? Explain how they have bias.

8. Why do experimenters use placebos? How do they use them?

9. Which samples are representative of their populations, which are not? Explain why.

   (a) A marketing firm wants to know how much time teenagers spend on youtube. They post ads to take their survey on the top ten youtube videos.

   (b) A large company wants to know how far their employees drive to work. They pick employees from several cities, several different job levels, and a mix of young and older employees.

1.2 Summarizing Data

1.2.1 Distributions

After data is collected, it is a good idea to summarize and display the data in ways that show the important characteristics. One of the best ways is to create a **Distribution** of each variable, which is information that tells us what values the variable takes on and how often. Large sets of data are often grouped according to **Classes**, which are categories or groupings. Numerical data should be grouped into consecutive intervals, where the starting value of each interval is called the **lower limit** and the ending value of each interval is called the **upper limit**.

For example, grades on a test are typically grouped into grade intervals such as 60-69, 70-79, 80-89, 90-99, etc. The lower limits would be 60, 70, 80, and 90 and the upper limits would be 69, 79, 89, and 99.

The classes should NEVER OVERLAP. If intervals went from 100-130 and then 130-160, which class would 130 belong too? That would lead to double counting values of the data. Each class has a **class midpoint** which is the average of the lower and upper limit. There are also **class boundaries** between each two classes. The boundary is the number halfway between the upper limit of one class and the lower limit of the next class.

The distribution has a **class width**. The width is the difference between consecutive lower limits, or consecutive upper limits, or consecutive midpoints, or consecutive boundaries.

Example: For the classes shown below, state the lower limits, upper limits, midpoints, and boundaries, then find the class width.

100-119	120-139	140-159	160-179	180-199

Solution: For the classes shown below, state the lower limits, upper limits, midpoints, and boundaries, then find the class width.

class	100-119	120-139	140-159	160-179	180-199
lower limit	100	120	140	160	180
upper limit	119	139	159	179	199
midpoint	109.5	129.5	149.5	169.5	189.5
boundary		119.5	139.5	159.5	179,5

The class width is 20. Notice we can get this by any of the following, $120 - 100 = 20$, $159 - 139 = 20$, $149.5 - 129.5 = 20$, or $179.5 - 159.5 = 20$.

The three guidelines for grouping data into classes are:

1. Small number of classes to be effective, but enough to show differences.
2. Each observation must belong to only one class, the classes MUST NOT overlap.
3. Whenever feasible, classes should have same width.

Then compute the **Frequency** of each class, which is the number of observations that fall into a class (count). A listing of all classes of the data and their frequencies is called a **Frequency distribution**.

When data sets are different sizes, it is hard to compare them. A good way to compare is to compute **Relative Frequency**, which is the ratio of the frequency of a class to the total number of observations. A listing of all classes and their relative frequencies is called a **Relative Frequency distribution**. Most distributions show frequencies as well as relative frequencies.

Example: Bradley worked a summer job to earn money for college. His weekly hours over a 12 week period were 25, 32, 36, 32, 18, 28, 30, 36, 12, 16, 35, 36. We can group the

hours into 3 classes in various ways, but one simple choice would be 10-19, 20-29, and 30-39.

The Distribution would be as follows:

Hours	Frequency	Relative Frequency
10-19	3	$\frac{3}{12} = 0.25 = 25\%$
20-29	2	$\frac{2}{12} = 0.167 = 17\%$
30-39	7	$\frac{7}{12} = 0.583 = 58\%$
Total	12	100%

The lower limits are 10, 20, and 30. The upper limits are 19, 29, and 39. The class midpoints would be 14.5, 24.5, and 34.5. The boundary between the first two classes is 19.5. The boundary between the last two classes is 29.5.

Notice that the frequencies add up to 12, and there are 12 weeks of data. Also the relative frequencies add up to 100%. These must always happen, or the distribution was done incorrectly. If the relative frequencies are rounded, then total percentage may be slightly off, between 99 to 101%. Any larger difference is not valid.

Sometimes it is useful to calculate **Cumulative Frequencies** which are a running total of the frequencies accumulated up through each class group. You can have either cumulative frequency or cumulative relative frequency or both. Just add the frequency of teach group with all of the frequencies before it.

The cumulative frequency of the first class is always just the frequency of that class, since there are no classes before it. The cumulative frequency of the last class is always the total, since there is no more data after that.

Example: Times for a high school track team running the 100 meter dash are grouped in the distribution below. Fill in the missing values.

Time (sec)	Frequency	Cumulative Frequency	Relative Frequency	Cum. Rel. Frequency
10-10.9	2	2	0.06	0.06
11-11.9	6	8		0.25
12-12.9	10	18	0.31	0.56
13-13.9		27	0.28	
14-14.9	5	32	0.16	1.00
Total			1.00	

Solution: The class 13-13.9 has a cumulative frequency of 27, so that frequency must be $27 - 18 = 9$. The last cumulative frequency is 32, so the total for frequencies is 32. The relative frequency for 11-11.9 is $\frac{6}{32} = 0.19$. The cumulative relative frequency for 13-13.9 seconds is $0.56 + 0.28 = 0.84$

Time (sec)	Frequency	Cumulative Frequency	Relative Frequency	Cum. Rel. Frequency
10-10.9	2	2	0.06	0.06
11-11.9	6	8	0.19	0.25
12-12.9	10	18	0.31	0.56
13-13.9	9	27	0.28	0.84
14-14.9	5	32	0.16	1.00
Total	32		1.00	

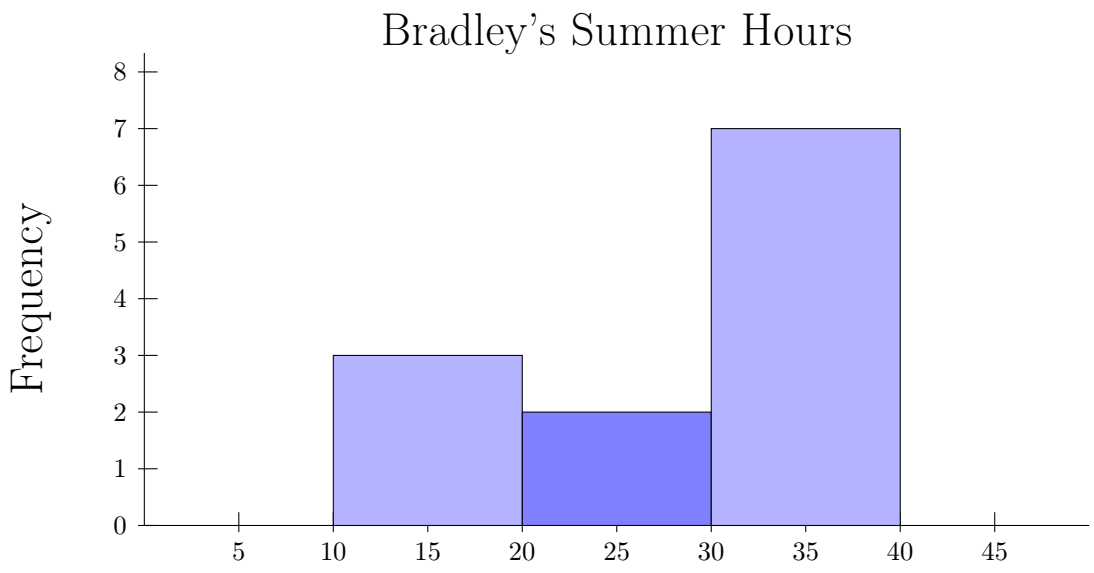
1.2.2 Graphs of Data

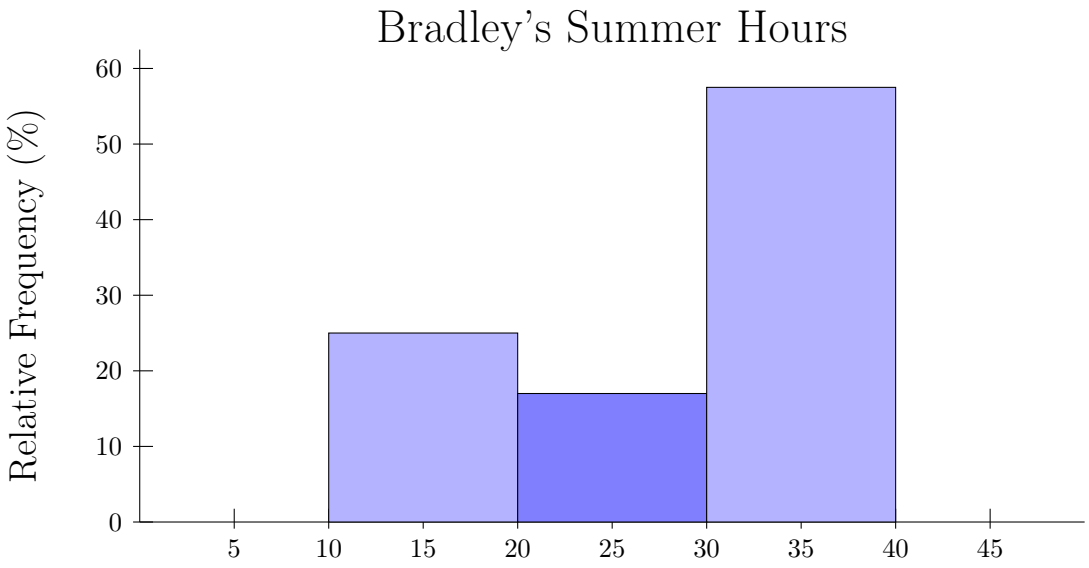
As the old saying goes, a picture is worth a thousand words. Data summaries can come in pictures or graphs. Here are some of the typical types of graphs to display distributions. They can give us a quick overview of the big picture and the characteristics of the data.

A **Frequency Histogram** is a graph that displays the classes on the horizontal axis and the frequencies on the vertical axis. It consists of vertical bars, whose height is equal to the frequency of the class(interval). The bars are drawn next to each other (without gaps), since they encompass the range of the data in numerical order. The left side of each bar starts at the lower limit of the class interval. The right side goes up to the lower limit of the next interval. A Histogram is only for quantitative data, not qualitative.

A **Relative Frequency Histogram** is the same as a frequency histogram, except it uses relative frequencies for the vertical axis and the bar heights.

Example: The frequency and relative frequency histograms for Bradley's summer job data are shown below.

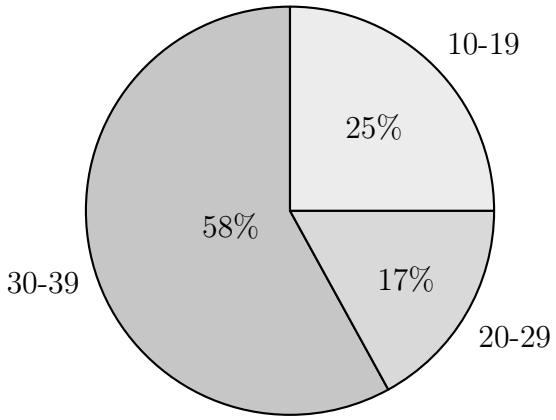




Notice that these graphs are the same shape, this is because the relative frequencies are based on the frequencies, so the same relationships between the classes are maintained.

A **Pie Chart** is a disk (circle) divided into pie-shaped pieces proportional to the relative frequencies. A pie chart should be labeled well, with class and the relative frequency for each slice. If a slice is very small, then the labels can go outside with an arrow pointing to the corresponding slice. The preferred way to sketch a pie chart is to start slices at 12 o'clock and rotate clockwise.

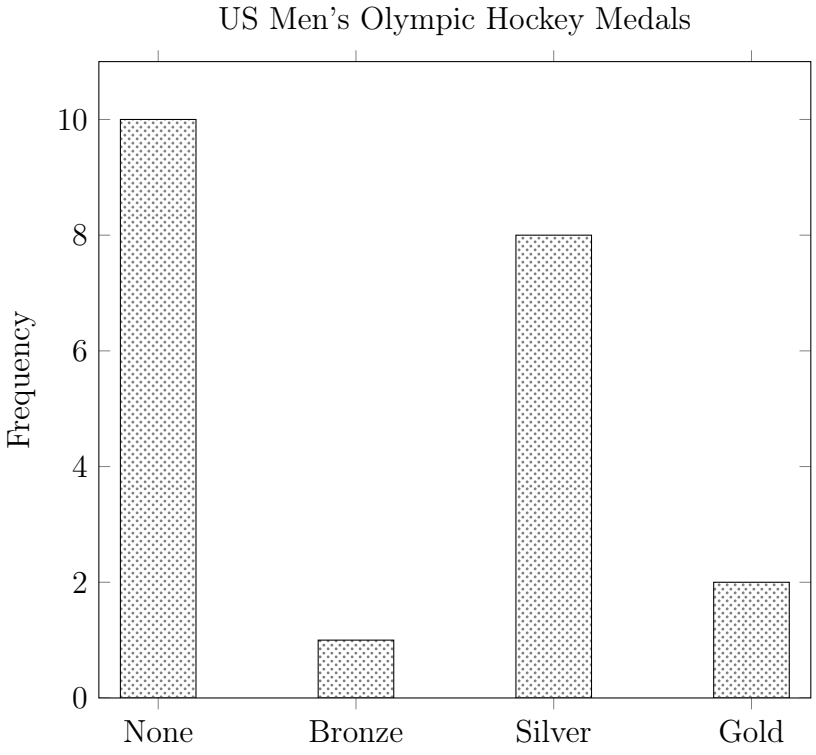
Example: The pie chart for Bradley's summer job data is shown below.

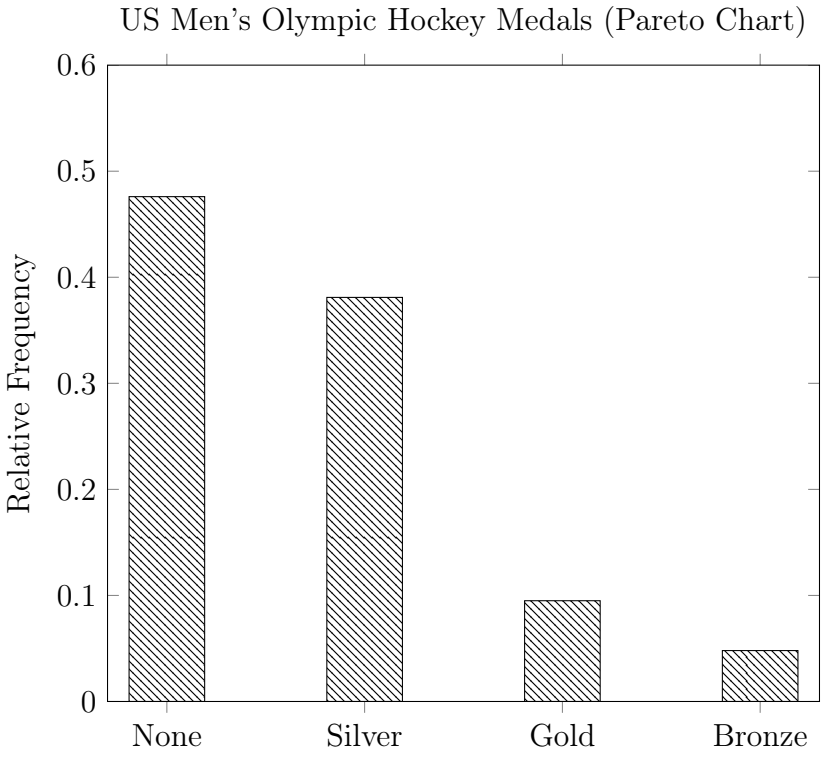


****Try this on your own**: The grades on a science final exam were 75, 83, 96, 82, 90, 78, 60, 76, 82, 71, 92, 86, 83, 88. Create a table with frequencies and relative frequencies using the intervals 60-69, 70-79, 80-89, 90-99. Then sketch a frequency histogram and a relative frequency pie chart.

Histograms are for quantitative data. There is a similar graph for qualitative data (categories), called a **Bar Graph**. In a bar graph, the width of the bars is arbitrary and the bars are not connected. Can show frequency or relative frequency. A **Pareto chart** is a specific type of bar graph where the classes are reordered so that the bars are in size order.

Example: The US Men's olympic hockey teams have played in 21 olympic games, winning 11 medals (2 gold, 8 silver, 1 bronze). Below are a frequency bar graph (ordered from worst to best finishes) and a relative frequency Pareto chart (ordered from highest to lowest).

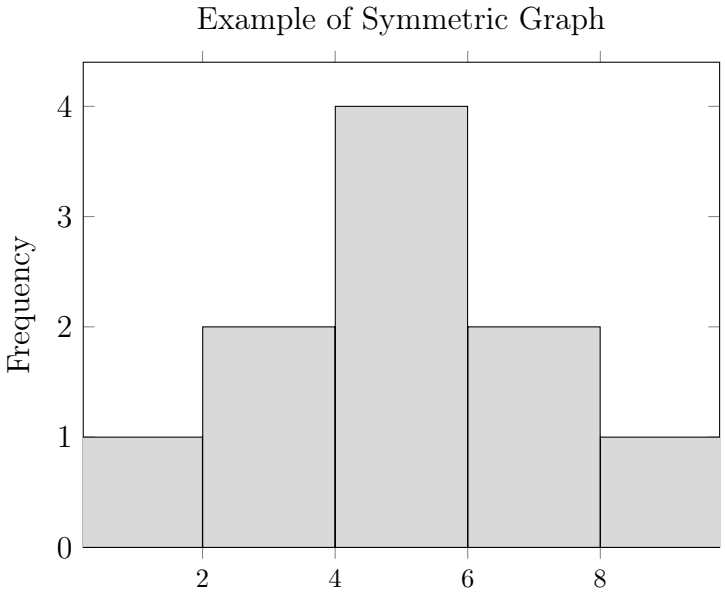


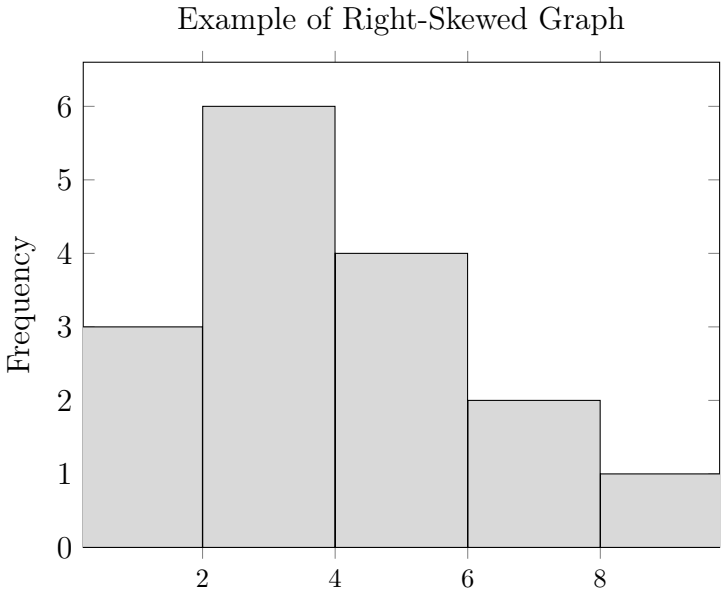


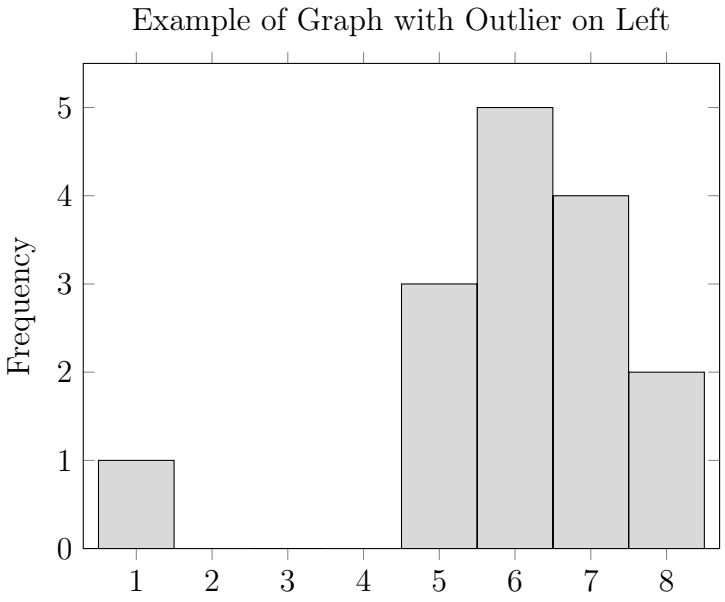
Graphs can help us get an overall view of the data set. When looking at a graph, pay attention to the following:

- **Center**: where is the middle of the graph, and the highest point.
- **Spread**: how are the parts of the graph spread out from each other?
- **Shape**: what shape does the graph have? Bell shape, straight across (uniform), repeatedly up and down, random?
- **Symmetry**: graph can be split in half with two mirror image parts, almost equal amount on both sides of the peak. **Uniform** graphs (level straight across) are also technically symmetric.
- **Skewness** A graph that extends more out to left side of the peak (high point) is called **Left-skewed**. A graph that extends more out to right side of the peak (high point) is called **Right-skewed**.

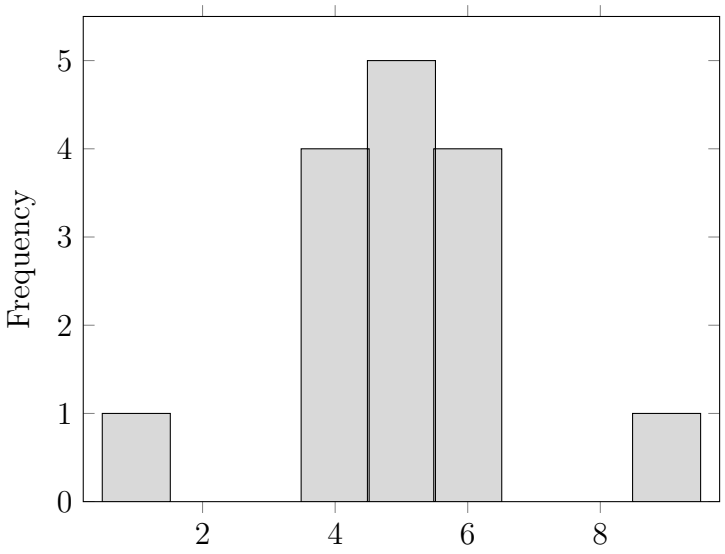
- **Outliers**: are data values (small parts of the graph) that are far from the other data (parts).







Example: what are the characteristics of the following graph? What does is suggest about the data?



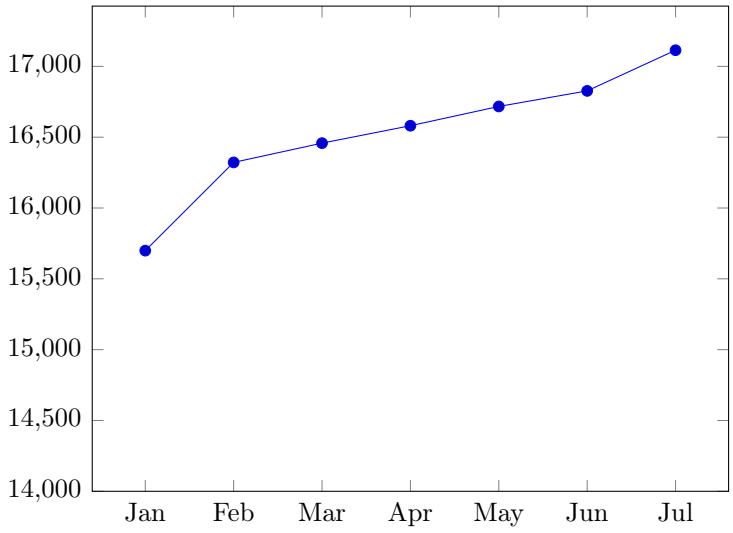
Solution: The center of the graph is at data value 5, also the highest point is at 5. Most of the data in concentrated near the middle (from 4 to 6) with outliers on both extremes (only one value at 1 and one value at 9). It is not spread out very

much. The graph is symmetric, mirror images on either side of 5. It is somewhat bell-shaped.

Another type of simple graph that plots the values of a single variable over time, is called a **Time Series**. The horizontal axis is in time units and the vertical axis is scaled for the values of the variable being graphed. It has values plotted as points and connected by lines, although the lines themselves do not represent data, they are just to show the pattern of the points. The vertical scale should start at zero, unless all of the data is very large, then it is better to start higher.

Example: Here is a time series graph for the Dow Jones stock market closing average for the first business day of each month of 2014.

Dow Jones closing value: 1st business day of month, 2014

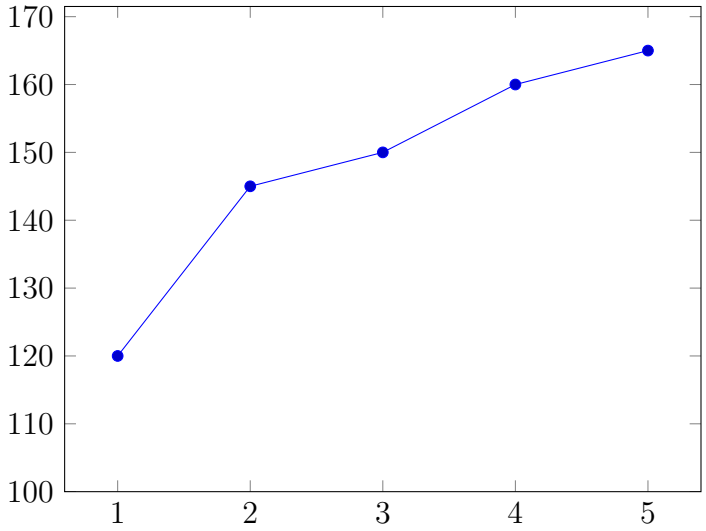


From this graph, we can see a general upward trend (increasing over time), with the sharpest increase from January to February. Notice the vertical axis for the price, does not start at zero. This is because the values are all very large, and starting at zero would collapse it into a small area at the top. It would be hard to see any details.

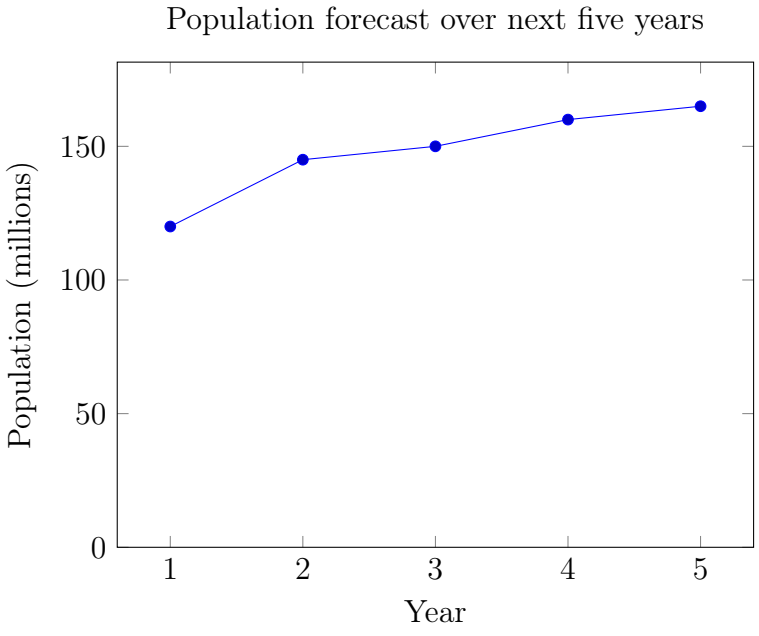
Beware of misleading or bad graphs!! Any data can be shown accurately but with different graphs and seem like it is showing very different results. There are several common ways that graphs can be misleading.

- Starting the vertical axis (values or frequencies) above zero. This chops the bars down and exaggerates the differences. The only exception is for line graphs (time series) with very large values for all of the data, to avoid having the graph squished into a small area.
- Using uneven scales on the axes.
- Using multiple dimensions when the data is just one dimension (using area or volume, when data is only the height for the bars).
- Unclear labelling.
- Too many cosmetic enhancements. This makes the graph hard to read, it is too busy!
- Poor choice of grouping.

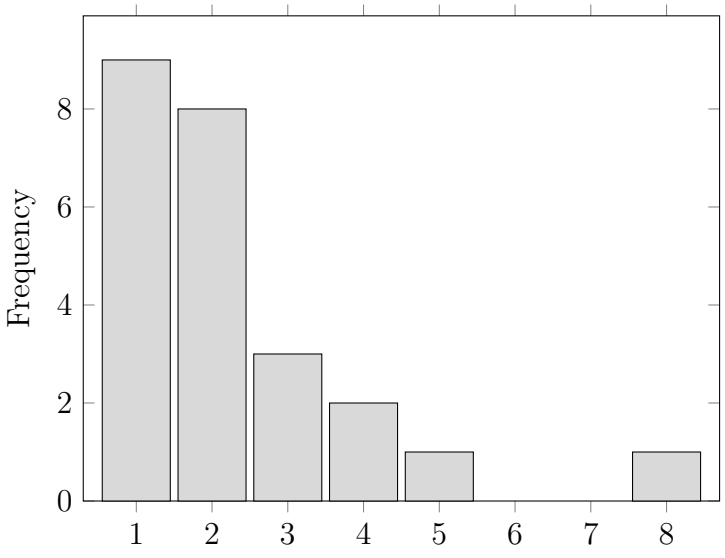
Here is an example of how to create a good or bad graph with the same data. The first graph is bad because it starts the vertical axis at 100, which makes the increase look very steep. It is also not labelled. The second graph shows a better view of the data starting at zero and labelled.



Notice the increase from 1 to 2, looks very steep above, due to the vertical axis being chopped down. Below, the axis starts at zero and shows a less drastic increase. Also the graph below is easier to read with labels.



****Try this on your own**: What are the characteristics of the following graph? Examine the spread, symmetry, and outliers.



1.2.3 Exercises: Summarizing Data

Solutions appear at the end of this textbook.

1. When grouping data into classes, why should we use only a small number of classes? If a data set has 1000 values, why not use 100 classes?

2. When computing relative frequencies, how do we know that we have calculated them correctly?

3. For the classes shown below, state the lower limits, upper limits, midpoints, and boundaries, then find the class width.

1-23	24-46	47-69	70-92	93-115

4. What is the difference between a bar graph and a Pareto chart? Can a Pareto chart be done from quantitative data? Why or why not?

5. Group the following data, calculate the frequencies and relative frequencies. Convert relative frequencies to nearest whole percent. A Farmer kept a log for one month of which days it rained. Here is the order. Tuesday, Saturday, Sunday, Tuesday, Friday, Sunday, Wednesday, Sunday, Friday, Tuesday

6. Make a pie-chart and a bar graph for the distribution you created from the farmer's observations in the previous exercise.

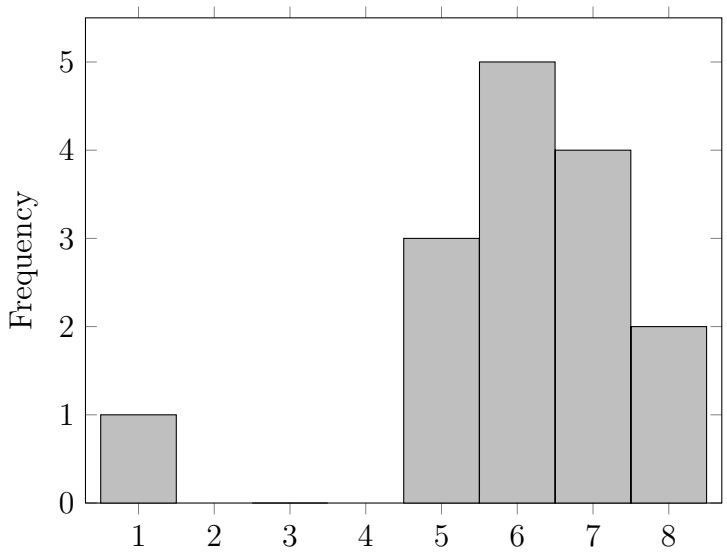
7. Group the following data into classes, calculate the frequencies and relative frequencies. Convert relative frequencies to nearest tenth of a percent. The grades on a history test were as follows: 67, 72, 99, 100, 82, 83, 94, 90, 80, 85, 85, 77, 48, 88, 75, 50, 75, 82, and 95. Use Classes of under 60, 60-69, 70-79, 80-89, and 90-100.

8. Make a pie-chart and a histogram for the distribution you created from the history grades in the previous exercise.

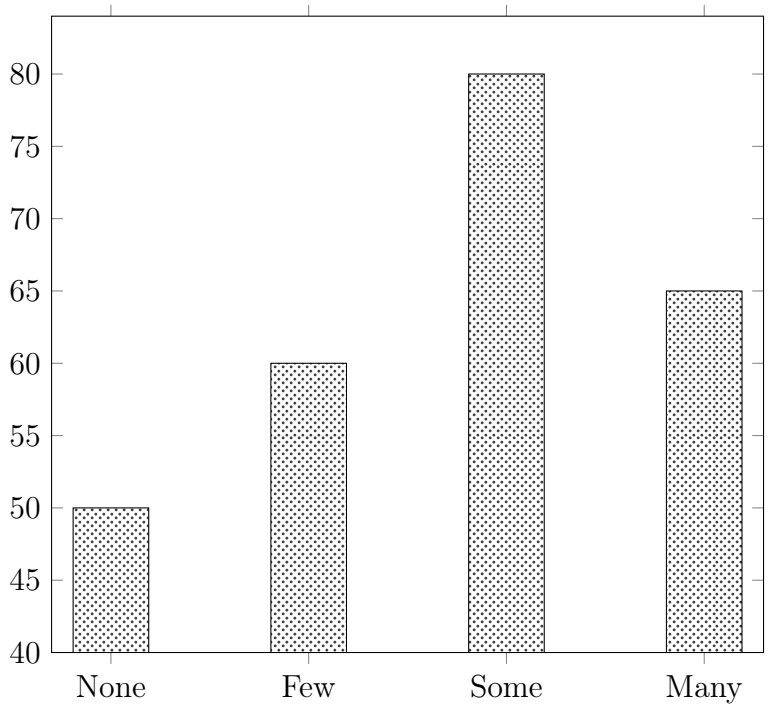
9. The Covid-19 fully vaccinated rates for the 50 US states are shown in the distribution below. Fill in the missing values. The data comes from www.usafacts.org as of May 15th, 2022. The lowest rate is 51% in Wyoming. The highest rate is 83% in Rhode Island.

Rate %	Frequency	Cumulative Frequency	Relative Frequency	Cum. Rel. Frequency
50-54.9	9		0.18	
55-59.9	8			
60-64.9	11		0.22	
65-69.9	7			
70-74.9	6		0.12	
75-79.9	7			
80-84.9	2		0.04	
Total	50		1.00	

10. Describe the characteristics of the following graph (center, spread, shape, symmetry, outliers).



11. Describe two issues that are bad or misleading about the following graph.



1.3 Measuring Data Sets

1.3.1 Measures of Center

In order to use mathematics to measure data sets, there will be several formulas that we will use. We will typically use letters to represent variables (typically X or Y or a letter that has some relevant meaning like S to represent the variable salary).

Particular values (observations) of a variable X can be denoted by subscripts x_1, x_2, etc. For summation of values, we use the Greek capital letter Sigma Σ, with a variable next to it to show which one is being summed. For example: $\sum x$ stands for the sum of the values of the variable X.

Descriptive measures are numbers used to describe data sets (average, min, max, etc.). They fall into two main categories: Measures of Center and Measures of Variation. We will use a common rule for rounding here.

The **Round-Off Rule**: Round all calculations to one more decimal place than is present in the data. Round only the final answer, not the steps along the way. For example, if the data set has values that are go out to 2 decimal places, then all calculated statistics should be reported showing 3 decimal places.

Measures of Center (or measures of central tendency) are descriptive measures that indicate where the center or most typical value of a data set lies. Some of the specific measures of center are shown below.

The **Mean** is sum of all the values of the observations, divided by the number of observations. The mean is also more commonly just called *average* .

The Sample mean is represented by the symbol $\overline{x}$, which is called 'x-bar': $\overline{x} = \dfrac{\sum x}{n}$, where n is the number of values in the sample.

The Population mean is represented by the symbol μ, which is called 'mu': $\mu = \frac{\sum x}{N}$, Where N is the number of values in the population.

Both of these are the same procedure and give the average of the values, the only difference is where the values come from, a sample, or the whole population.

Example: In the previous section, we looked at graphs for Bradley's weekly hours at a summer job: 25, 32, 36, 32, 18, 28, 30, 36, 12, 16, 35, 36. Find the mean (average) hours he worked in a week.

Solution: We add up the values and divide by 12 (the count of how many values). Since this is the entire set of his summer job hours, it is a population mean. $\mu = \frac{\sum x}{N} = \frac{336}{12} = 28$, which we will report, according to the round off rule, as 28.0 hours worked in an average week.

The **Median** is the value that divides the bottom 50% of data from top 50%. It is the middle value when the values are placed in size order. To find the median, first arrange the data in increasing order. If there are an odd number of observations, the median is the middle value in order. If there are an even number of observations, the median is the average of two middle values in order .

Example: Find the median of Bradley's summer weekly hours.

Solution: First we must put the values in order from lowest to highest: 12, 16, 18, 25, 28, 30, 32, 32, 35, 36, 36, 36. There are two values in the middle, 30 and 32, with five values below and above. The average of the two middle values is 31, which is the median. Notice that the median is not one of the original data values here. According to the round-off rule, we will report the median as 31.0 hours worked in a week. Half of the data is below this and other half above this.

The **Mode** is the value that has the most number of observations (frequency), but must occur more than once. There can be multiple modes (a tie for the most often).

Example: Find the mode of Bradley's summer weekly hours.

Solution: Having the values in order makes this easier: 12, 16, 18, 25, 28, 30, 32, 32, 35, 36, 36, 36. The values that occurs the most often is 36, which is the only mode here.

Here are some cautions about measures of center. The mean is sensitive to extreme values. If a company has 20 workers making $15,000 each and the owner makes $500,000, then the mean would be $38,095. This mean does not give a complete picture of the company salaries. Nobody makes close to that value, everyone except the owner is way below average.

When dealing with salaries or prices, the median is often a better measure of the data set. The median for the company would be $15,000 and a better representation of what potential employees could expect to be paid.

Most of the time, data that is left-skewed, will have a mean that is less than the median. For right-skewed, the mean is greater than the median. This is because the skewed data out to the extreme, pulls the mean closer to that extreme, but the middle values are still in place so the median is not affected.

If a large data set has already been grouped and you have only the frequency distribution (but not the actual data), the average can be estimated using the midpoint of each class (group) as the estimate of the typical value in each class and multiplying that by the frequency of that class. The formula is: $\frac{\sum \widehat{x}f}{\sum f}$, where $\widehat{x}f$ is the product of each class midpoint $\widehat{x}$, times the class frequency, f.

Example: Estimate the mean of the following frequency distribution.

Class	1-8	9-16	17-24	25-32	33-40
Frequency	3	5	7	2	1

Solution: The midpoints of the classes ($\widehat{x}$) are: $4.5, 12.5, 20.5, 28.5, 36.5$.
Then the formula would be $\frac{\sum \widehat{x}f}{\sum f} = \frac{4.5(3)+12.5(5)+20.5(7)+28.5(2)+36.5(1)}{3+5+7+2+1} = \frac{313}{18} = 17.3889$

According to the round-off rule, we should report this as 17.4. Let's think about the estimate of 17.4, does it make sense? Notice in the frequency distribution, that the class 17-24 occurs the most, with more in the lower classes than in the upper classes. Then the estimate of 17.4, in the beginning of the class 17-24, makes sense.

Another special case of finding average is when we know the values, but they are not all equally weighted. This idea is know as a **Weighted Mean**. The formula is $\overline{x} = \dfrac{\sum xw}{\sum w}$, where xw is the product of each data value x multiplied by its weight w. The weight could be a dollar amount or a credit amount (as in college classes), or other amounts.

Example: If an investment earns 6% interest on \$1,000 and 4% interest on \$250, what is the average interest rate for the entire investment.

Solution: here we are looking for the average interest rate, so the data values are the interest rates. The dollar amounts are the weights. More money, means more weight for its corresponding interest rate. Therefore, the average rate will be closer to 6%, since it corresponds to the biggest dollar amount invested. $\overline{x} = \frac{6(1000)+4(250)}{1000+250} = \frac{7000}{1250} = 5.6$, so the average interest rate on the entire investment is 5.6%.

Example: Rachel's class is graded in the following manner: Test average counts 30%, homework average counts 25%, participation counts 10%, project counts 15%, and final exam

counts 20%. What is her overall grade average if she has the following grades:
Test avg 86, HW avg 94, part. 100, project 80, final 82.

Solution: here we are looking for the average grade, so the data values are grades. The percentages are the weights. Test average has the most weight (30%). Class Avg $= \frac{86(30)+94(25)+100(10)+80(15)+82(20)}{30+25+10+15+20} = \frac{8770}{100} = 87.7$, so her overall grade in the class is about 88, a high B. Almost an A, but not quite.

Academic Grade Point Average (or GPA) is a very common concept, but one that most people don't understand how to compute. Letter grades earned in classes are assigned a numerical value called quality points, ranging from 0 to 4. Typical grade scale is $A = 4.0$, $B = 3.0$, $C = 2.0$, $D = 1.0$, and $F = 0$. Some schools have grades in between with a $\pm$, with fractional values. The weights of the grade points are the credit hours for the classes. A longer class for 4 credit hours has more weight than a short elective for 2 credits.

Example: Compute the overall semester GPA for a college student who earned the following grades. Use the standard letter points shown above, assume there are no $\pm$ grades in between.

Course	Grade	Credits
Physics	B=3.0	4
English	C=2.0	3
Math	C=2.0	3
Study Skills	B=3.0	1
History	A=4.0	3

Solution: here we are looking for the average grade points, so the data values are grade (quality) points. The credit amounts are the weights.
$GPA = \dfrac{3.0(4) + 2.0(3) + 2.0(3) + 3.0(1) + 4.0(3)}{4+3+3+1+3} = \dfrac{39}{14} = 2.79$, which is a high C,
close to an overall B. Not the greatest, but it is passing.

1.3.2 Measures of Variation

Measures of Variation (or measures of spread) are descriptive measures that indicate how much variation is in the data or how spread out the data values are from each other.

The **Minimum** (Min) is the lowest value in the data set.

The **Maximum** (Max) is the highest value in the data set.

The **Range** is the difference between Min and Max: ($Range = Max - Min$). To give a more detailed range, some studies will just say the range goes from MIN to MAX.

Example: Find the min, max, and range of Bradley's summer weekly hours.

Solution: Having the values in order makes this easier as well: 12, 16, 18, 25, 28, 30, 32, 32, 35, 36, 36, 36. Here $min = 12$, $max = 36$, and $Range = 36 - 12 = 24$ hours. We could also say the range is from 12 to 36 hours.

Perhaps the hardest measure to compute (without using special function on calculators or computers) is the next measure of spread.

Standard Deviation is the measure of how far, on average, the data is from the mean. Another related measure, is the **Variance** which is standard deviation squared. It will not be used much here, except as a step in the computation of standard deviation.

The standard deviation and variance for a SAMPLE are calculated by the following symbols and formulas:

Variance: $s^2 = \dfrac{\sum (x - \overline{x})^2}{n-1}$ Standard deviation: $s = \sqrt{\dfrac{\sum (x - \overline{x})^2}{n-1}}$

The variance and standard deviation for a POPULATION are calculated by the following symbols and formulas:

Variance: $\sigma^2 = \dfrac{\sum (x-\mu)^2}{N}$ Standard deviation: $\sigma = \sqrt{\dfrac{\sum (x-\mu)^2}{N}}$

Example: Find the standard deviation of Bradley's summer weekly hours.

Solution: To work out standard deviation yourself, it is very important to work out each step carefully and organized. The data set is the population of the entire summer hours for Bradley, so we will use the second set of formulas. First we use the mean $\mu = 28$ and the total count $N = 12$ from the previous examples. Then it helps to make a table like the one below. For repeated values, the calculation is only shown once (one row), but used multiple times in the formula.

x (hours)	$x - \mu$	$(x-\mu)^2$	Count
12	$12 - 28 = -16$	$(-16)^2 = 256$	once
16	$16 - 28 = -12$	$(-12)^2 = 144$	once
18	$18 - 28 = -10$	$(-10)^2 = 100$	once
25	$25 - 28 = -3$	$(-3)^2 = 9$	once
28	$28 - 28 = 0$	$0^2 = 0$	once
30	$30 - 28 = 2$	$2^2 = 4$	once
32	$32 - 28 = 4$	$4^2 = 16$	used twice
35	$35 - 28 = 7$	$7^2 = 49$	once
36	$36 - 28 = 8$	$8^2 = 64$	used 3 times
Sum		642	

The standard deviation, $\sigma = \sqrt{\dfrac{\sum (x-\mu)^2}{N}} = \sqrt{\dfrac{642}{12}} = \sqrt{53.5} = 7.31437$

Using the round-off rule, $\sigma = 7.3$ hours. Computing the standard deviation can be done very quickly and easily using special calculator functions or computers, but interpreting the meaning of it requires human understanding. In this example, it means that Bradley's

weekly hours are somewhat spread out, on average his weekly hours are 7.3 hours away from the mean of 28 hours. Sometimes they were close to 28 (one week exactly), and sometimes farther away (as much as 16 hours below).

If another employee worked with Bradley, but had a standard deviation of only 3 hours, then that would mean that the other employee worked weekly hours that were much more consistent, working about the same each week, and not as spread out as Bradley's hours were.

Example: Find the variance and standard deviation of the sample data from a science quiz: 0, 8, 5, 6, 8, 3, 10, 9, 8, 6, 3.

Solution: The mean is the sum of the data divided by 11, $\overline{x} = \frac{66}{11} = 6$. The deviations from the mean are -6, 2, -1, 0, 2, -3, 4, 3, 2, 0, -3. The squared deviations are 36, 4, 1, 0, 4, 9, 16, 9, 4, 0, 9. The sample variance is $s^2 = \dfrac{\sum (x - \overline{x})^2}{n-1} = \dfrac{92}{11-1} = 9.2$
The sample standard deviation is $s = \sqrt{9.2} = 3.03315$, and by the round-off rule $s = 3.0$.

If we wish to compare two data sets, to figure out which is spread out more, there are two cases to consider. First, if the data sets are from similar variables with similar sizes, then we can directly compare the standard deviations, since they are the same units. As an example, we already discussed Bradley's hours versus his fellow employee.

The other case is when comparing two very different data sets. For that we will use a special measure called the **Coefficient of Variation** (or CV). It is equal to the standard deviation divided by the mean, converted into a percent. It has no units, it is only a ratio as percent. The formulas are slightly different, depending upon the data set being from a sample or a population. The CV states how big the standard deviation is, relative to the average size of the data.

For a sample: $CV = \dfrac{s}{\overline{x}} \times 100\%$ For a population: $CV = \dfrac{\sigma}{\mu} \times 100\%$

Example: Which data set is more spread out, the weight of elephants in a herd: $s = 1,175$ pounds and $\overline{x} = 12,342$ pounds, or the price of regular unleaded gasoline in a US: $s = \$0.26$ and $\overline{x} = \$3.73$?

Solution: For the elephant weights, $CV = \dfrac{1,175}{12,342} \times 100\% = 9.5\%$. For the gas prices, $CV = \dfrac{0.26}{3.73} \times 100\% = 7.0\%$. The weights of elephants are a more spread out set of data than US gas prices. This is NOT because the values are larger. Another set of large values could have a lower CV than gas prices.

****Try this on your own**: The grades for a sample of a science final exam were 75, 83, 96, 82, 90, 78, 60, 76, 82, 71, 92, 86, 83, 88. Calculate the mean, median, mode, range and standard deviation.

1.3.3 Exercises: Measuring Data Sets

Solutions appear at the end of this textbook.

1. Compute the mean, median, and mode of the following data sets. Use the round-off rule.

   (a) 1, 2, 3, 4, 5, 6, 7

   (b) 1, -2, 0, 3, 4, 0

   (c) 123, 318, 222, 301, 188, 195, 253, 172, 230, 103, 155, 281

2. For the following data, compute the mean, median, and mode. Use the round-off rule. The grades on a history test were as follows: 67, 72, 99, 100, 82, 83, 94, 90, 80, 85, 85, 77, 48, 88, 75, 50, 75, 82, and 95

3. For the following data, compute the min, max, range, variance and standard deviation. One of the hottest selling concerts of 2013 was the Honda Civic Tour featuring Maroon 5, Kelly Clarkson, and Rozzi Crane. A sample of ticket prices for the Atlanta Lakewood show on August 1st were $44, $74, $94, and $116.

4. Explain what the standard deviation from the previous exercise tells you about the ticket prices.

5. How are the mean, median and mode affected by extreme values?

6. If a set of data is not known to be from a sample or a population, then there are two possible standard deviation formulas to use. Explain why the sample standard deviation is always larger than the population standard deviation.

7. College GPA is a weighted average of the grades earned in all of your courses. Letter grades are equated to a numeric value called quality points. The weights are the credit hours earned for the course. On a typical A-F system with no plus/minus, the points

are $A = 4.0$, $B = 3.0$, $C = 2.0$, $D = 1.0$, and $F = 0$. If a student had the following grades for their first semester, what would their GPA be for that semester? Grade B in College Algebra(3 credits), Grade B in Chemistry(4 credits), Grade A in Phys Ed(2 credits), Grade C in Writing(3 credits), Grade A in Economics(3 credits).

8. A fitness company did a study and found the following statistics for 1000 women in the Atlanta area. Mean weight = 162 pounds, median weight = 141 pounds, standard deviation = 45 pounds. What do these statistics suggest about the distribution of women's weights?

9. Estimate the mean of the following frequency distribution.

Class	10-14	15-19	20-24	25-29	30-34	35-39
Frequency	12	5	7	2	6	3

10. Which data set is more spread out, the shot put throws for a high school track team: $s = 5.5$ feet and $\overline{x} = 38$ feet, or the gymnastics scores for a college team: $s = 1.4$ points and $\overline{x} = 8.45$ points?

1.4 Measures of Relative Standing

1.4.1 Z-scores

In the last section, we looked at measures which described the data set as a whole. In this section we will look at measures that describe how particular data values compare to each other, called **Measures of Relative Standing**.

A **z-score** (or standardized score) is the number of standard deviations that a given value is above or below its mean. Whenever a value is below the mean, its corresponding z-score will be negative. Usual values are z-scores from -2 to $+2$. Unusual values are z-scores outside this range. More than 3 standard deviations away from the mean is very unusual for most data sets. Z-scores have no units.

Z-scores are found by the following formulas.

$$\text{For a sample: } z = \frac{x - \overline{x}}{s} \qquad \text{For a population: } z = \frac{x - \mu}{\sigma}$$

Example: A scientist took a sample of tree heights in a forest. His results were $s = 1.8$ meters and $\overline{x} = 9.1$ meters. Calculate the z-score for a tree that is 14 meters tall, and explain what the value means.

Solution: $z = \dfrac{14 - 9.1}{1.8} = 2.72$. This means that the tree is 2.72 standard deviations above average, it falls into the category of unusually tall (between -2 and 2), but not extreme (more than 3).

Example: Standardized IQ test scores, for the overall population, have $\sigma = 15$ points and $\mu = 100$ points. Andy Warhol was a famous artist and leader of the 70's pop art movement. It is reported that his IQ was only 86. Calculate his z-score and explain what the value means.

Solution: $z = \frac{86 - 100}{15} = -0.93$. This means that his IQ is about 1 standard deviation below average, it falls into the category of usual intelligence (between -2 and 2).

In the above examples, the z-scores were rounded to two decimal places. This does not follow the round-off rule used previously. Z-scores are almost always shown to two decimal places (hundredths), because there is a special use for z-scores that require looking them up in a table. That table uses two decimal places. This table and its uses will be in a later chapter, but for now, let's get used to z-scores having their own rule of two decimal places.

****Try this on your own**: Calculate the z-score of a woman who is 5 feet tall if the mean height is 65 inches and standard deviation is 3 inches. Is she unusually short or not? Round Z to two decimal places.

1.4.2 Percentiles

Many measurements, including large standardized test scores and children's height and weight, report how a data value ranks among the data set. Most are from a broad series of rankings called **Percentiles**. Percentiles are the values of increasing size, that divide the data into 100 parts, each with about the same number of values.

There are 99 percentiles: $P_1, P_2, P_3, ..., P_{98}, P_{99}$. P_1 is the low value that has only one percent of the data at or below it, and 99% (almost all) of the data above it. P_{80} is the high value that has 80% (most) of the data at or below it, and only 20% of the data above it.

There are three special percentiles, called **Quartiles** (every 25th percentile), which divide the data into four equal parts. There are 3 quartiles Q_1, Q_2, Q_3. The median is the same as the second quartile Q_2, which is also the 50th percentile P_{50}.

We calculate the median as described previously. Place the data is increasing size order, and find the value in the middle. Once you find the median, then the data set will be split into two halves (lower and upper). Q_1 is the median of the first half of the data. Q_3 is the median of the second half of the data.

To find other percentiles, there are a few accepted methods. They sometimes give slightly different values, but the concept is still the same, percentiles give a good estimate of how high a value is, to be equal to or above a certain percent of all of the data. The method which I present here is one that is simple and used in many other books.

First we need to put the data set in increasing order, then find the location (position) of the particular percentile we wish to find. Once we determine the position, we take the data value in that position as the percentile. The location is calculated by the formula $L = n\left(\frac{p}{100}\right)$, where n is the size of the data set and p is the percentile value we are looking for (20, 60, 75, etc.). If L is a whole number, then the percentile will be the average of the values in positions L and $L + 1$. If L is a fraction, bump L up to the next whole number and the percentile is the data value in that position.

Example: Find the 20th and 93rd percentile of the following data set:

45, 84, 61, 34, 94, 5, 97, 42, 34, 15, 18, 71, 8, 65, 22, 10, 71, 80

Solution: The data in order are: $5, 8, 10, 15, 18, 22, 34, 34, 42, 45, 61, 65, 71, 71, 80, 84, 94, 97$ There are 18 values. The location of the 20th percentile is $L = 18\left(\frac{20}{100}\right) = 3.6$, so we look in the 4th position. The 4th value in order is the data value 15, so $P_{20} = 15$. There are actually 22% of the values at or below the 4th position, but this is just an estimate for the 20th percentile. The location of the 93rd percentile is $L = 18\left(\frac{93}{100}\right) = 16.74$, so we look in the 17th position. The 17th value in order is the data value 94, so $P_{93} = 94$.

1.4.3 Boxplots

There is a special set of measurements which is used to make a graphical picture of data. It is called the **Five-Number Summary** and consists of the $Min, Q_1, Median, Q_3, Max$. The graph for this is called a **Boxplot** or Box and Whisker Diagram. The steps to create a boxplot are as follows:

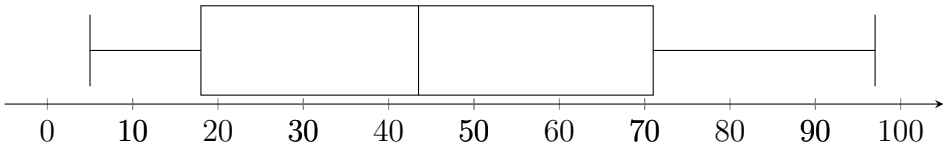
1. Compute the values in the 5 number summary.
2. Draw a horizontal numberline on which the 5 numbers can be located. Use a simple scale, like 0-50 or 25-100, etc.
3. Above the numberline, mark Q_1 and Q_3 with large vertical lines above their values and connect them to make the box.
4. Mark another vertical line for the median, above its value. This splits the box into two sections.
5. Mark the Min and Max with smaller vertical lines above their values.
6. Connect the centers of the sides of the box to the Min and Max with horizontal lines(whiskers).

Example: Find the five number summary of the following data set, and sketch the boxplot.

45, 84, 61, 34, 94, 5, 97, 42, 34, 15, 18, 71, 8, 65, 22, 10, 71, 80

Solution: The data in order are: $5, 8, 10, 15, 18, 22, 34, 34, 42, 45, 61, 65, 71, 71, 80, 84, 94, 97$ There are 18 values, so the median is the average of the 9th and 10th values: $Med = \frac{42+45}{2} = 43.5$. The first quartile is the median of the first nine values, so $Q_1 = 18$ (the 5th value). The third quartile is the median of the last nine values, so $Q_3 = 71$

(the 14th value). The five number summary is: $5, 18, 43.5, 71, 97$. A good numberline could be from 0-100, with marks every 10 units. The boxplot is shown below.



A useful interval related to the quartiles is the **Interquartile Range** (IQR), which is the range of the middle 50% of the data values, when they are in size order. $IQR = Q3 - Q1$.

****Try this on your own**: The grades on a science final exam were 75, 83, 96, 82, 90, 78, 60, 76, 82, 71, 92, 86, 83, 88. Calculate the 5-number summary, IQR, and sketch a regular boxplot.

Outliers were mentioned in the section on graphs. Here we will look at outliers as observations that fall well outside the overall pattern of data. Outliers are usually located far away from the mean. They may be errors in measurement, mistakenly part of the population, or just extreme values. A method for detecting outliers in a data set, is to compute upper and lower fences (described below). Any data values that lie outside the lower and upper fences, are considered to be outliers.

The **Lower Fence** (LF) of a data set is defined as $LF = Q1 - 1.5(IQR)$. It is also called the lower limit.

The **Upper Fence** (UF) of a data set is defined as $UF = Q3 + 1.5(IQR)$. It is also called the upper limit.

A **Modified Boxplot**, is a boxplot, with the addition of the fences and outliers shown as separate points. The steps to make a modified boxplot are:

1. Calculate the five number summary and the fences.

2. Determine if any data values fall outside the fences. These are outliers.

3. Determine the lowest and highest values that are NOT outliers.

4. Draw the box section of the boxplot using the quartiles.

5. Extend the whiskers out from the box, stopping at the low/high values from step 3.

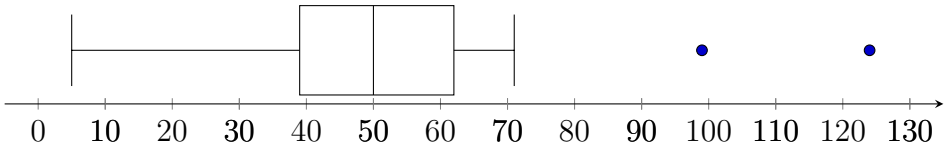
6. Plot each outlier as separate points, using asterisk, dot or other marker of choice.

Example: For the following data set, find the five number summary, IQR, fences, and determine if there are any outliers. Then sketch the modified boxplot.

48, 99, 60, 39, 5, 124, 47, 36, 71, 29, 62, 52, 50, 42, 57

Solution: Data in order are: $5, 29, 36, 39, 42, 47, 48, 50, 52, 57, 60, 62, 71, 99, 124$
There are 15 values, so the median is the 8th value 50. The first quartile is the median seven values below 50, so $Q_1 = 39$. The third quartile is the median of the seven values above 50, so $Q_3 = 62$. The five number summary is: $5, 39, 50, 62, 124$.

Then $IQR = 62 - 39 = 23$, $LF = 39 - 1.5(23) = 4.5$, and $UF = 62 + 1.5(23) = 96.5$. Only the low end, 5 is close to the lower fence, but not an outlier. On the high end, 99 and 124 are beyond the upper fence, so they are outliers and will be separate points. Therefore, the upper whisker will stop at the next data value down, which is 71. A good numberline could be from 0-130, with marks every 10 units. The modified boxplot is shown below.

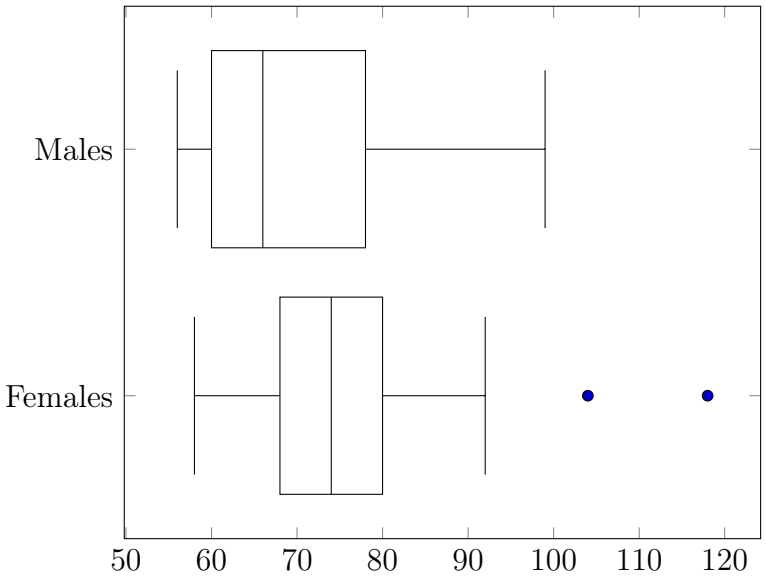


Boxplots can be used to compare two sets of data, by graphing them above one another using the same scaled axis. To tell the symmetry or shape of a data set, we generally look primarily at the box, but the whiskers can help as well.

A symmetric data set, will have a symmetric box (whiskers should be relatively symmetric). Left-skewed data will have a box that is wider on the left side. Right-skewed data will

have a box that is wider on the right side. The data set that is more spread out, will have a wider overall box and the whiskers generally span out farther. Ignore outliers for comparing the spread and shape.

Example: Based on the boxplots below, state the characteristics of the data sets. Which one is more spread out?



Solution: Ignoring the outliers, the female data set is symmetric and less spread out than the male data set. The male data set is right-skewed and more spread out.

1.4.4 Exercises: Measures of Relative Standing

Solutions appear at the end of this textbook.

1. IQ scores have a mean of 100 and a standard deviation of 15. Compute the z-scores of the famous people below and state which ones are unusual.

Name	IQ score
Garry Kasporov (chess champion)	190
Albert Einstein (scientist)	160
Arnold Schwarzenegger (actor)	135
Tim Tebow (football player)	104
Howard Stern (talk radio host)	99
George W Bush (43rd president)	125
Muhammad Ali (boxer)	78
Barack Obama (44th president)	130

2. Heights of adult males have a mean of 69 inches and a standard deviation of 3 inches. How tall must a man be to be considered unusually short or unusually tall?

3. For a data set with 38 values, compute the location (position) of the 45th percentile.

4. If a student scores at the 85th percentile on a standardized test, what does that mean?

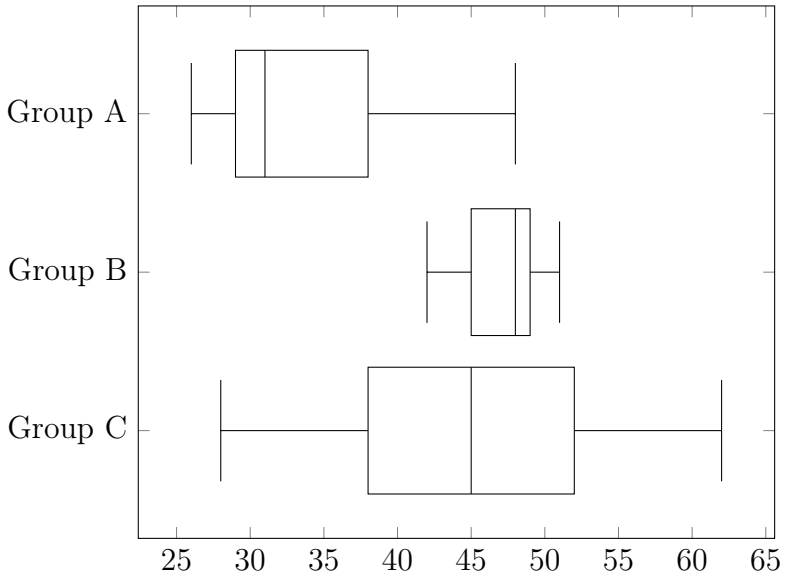
5. Find the Five-Number summary for 17, 7, 9, 10, 8, 7, 14, 20, 8, 5, 15, and 8.

6. For the following grades on a history test, compute the Five-Number summary.
   67, 72, 99, 100, 82, 83, 94, 90, 80, 85, 85, 77, 48, 88, 75, 50, 75, 82, and 95.
   Did the class do well on the test? Explain.

7. Create a regular box-plot for the data in the previous exercise.

8. Compute the lower and upper fences for the history test data and state which values, if any, are outliers. Then create a modified boxplot.

9. Based on the boxplots below, state the characteristics of the data sets. Which one is more spread out?



1.5 Data Sets with the TI-83 and Similar Calculators

So far, we have focused on understanding the material and working out the steps ourselves. This section will introduce the various calculator functions, which can help give quick and accurate information and graphs, especially for large data sets. It is very important to remember that a calculator is just a tool. It is quicker and more accurate than we are, but it CANNOT THINK!

If the average person were given professional tools, they could not build a house, without first learning how and gaining experience. So the average student should not be doing data analysis on a calculator, without first understanding the material and working problems out themselves (except using a calculator for large number arithmetic).

In the sampling section, you were shown how to get a random number. To get a random whole number between two values, we need to select the random integer function and input the lowest and highest values we wish to have the calculator select between. Press the MATH button, scroll right to the **PRB** menu, scroll down and select the *randInt* function and type the low/high values in parentheses, separated by a comma.

For example, to get a random whole number between 1 and 150, follow the process above and type randInt(1,150) and hit ENTER button. Try this yourself. Keep hitting ENTER and each time you will get another random number between 1 and 150. Occasionally numbers will repeat.

In order to have the calculator compute measures or sketch graphs, the data set must be entered into the calculator lists. The data lists can be found by pressing the STAT button and using the **EDIT** menu. The first function **Edit**, allows you to input and view data. The default setup for data lists is a set of columns labeled L_1 (for List #1), L_2 (for List #2), etc. This allows you to store more than one set of data. Each data set will have its own list.

Whenever you ask the calculator to perform a function with a list, it is strongly recommended that you specifically tell the calculator which list to work with. If you do not, the calculator will automatically do everything with L_1. You should get into the habit of always giving the calculator specific instructions. You need to be the master to make sure you are getting what you want. Do not let the calculator programming decide for you.

Before using the lists, it is a good idea to clear out any previous data, and start with blank lists. To do this, press the [STAT] button and under the **EDIT** menu, scroll down to the clear list function **ClrList** and press [ENTER]. This will bring up the command *ClrList* on the screen. This command must be followed by list names, so the calculator knows which lists to clear out. To put the list names on the screen, you need to press the keys [2nd] and [1] to type L_1 or [2nd] and [2] to type L_2 and so forth.

Try this: press [STAT] button and under the **EDIT** menu, scroll down to the clear list function **ClrList** and press [ENTER], then press [2nd] [1] followed by comma button [,] then [2nd] [2] [,] [2nd] [3] [,] [2nd] [4] [,] [2nd] [5] [,] [2nd] [6] then [ENTER]. This will clear L_1 through L_6.

Now we are ready to input data into a list. Press the [STAT] button and under the **EDIT** menu, select the **Edit** function and press [ENTER]. Use the arrow keys to move to the first blank under L_1. Now type the following data set in the list, by typing each value, then hit enter to move down to the next blank space. Continue to type each value and hit enter. The data is: $21, 22, 8, 19, 24, 2, 47, 30, 27, 28, 31, 40$.

The calculator functions can give us many of the computed measures from sections 1.3 and 1.4. The functions use data stored in a list. Press the [STAT] button and scroll over to the right to go under the **CALC** menu, select the **1-Var Stats** function and press [ENTER]. This will bring up the command *1-Var Stats* on the screen. Now we want to tell the calculator which list we want it to do the stats for. After the *1-Var Stats* command, type the appropriate list L_1, etc. Then hit [ENTER].

If you input the data into L_1 as shown in the previous paragraph, you should now have the statistics results for L_1. There are many values here, so you may have to use the arrow keys to scroll down to see them all. All of the stats here are for x. How does it know this is supposed to be x and not t for tests grades or other data? Well, it doesn't, calculators can't think! The calculator does not know what the data represents, so all lists are labelled as a generic unknown x.

The first piece of information is $\overline{x} = 24.91666667$. You should recognize this as the mean. The next stat is $\sum x = 299$, which is the sum of the data values in that list. Next is $\sum x^2 = 9113$, which is the sum of the squares of the data values. Then the calculator shows $S_x = 12.29529351$ and $\sigma_x = 11.7718473$. These are standard deviations. S is the sample standard deviation and σ is the population standard deviation. Why does the calculator give us both, doesn't it know which type of data we have? No, it does not. The calculator is just a machine. You need to choose the correct value for the type of data.

Scrolling down you will see $n = 12$ (we input 12 values), followed by the five number summary: $minX = 2$, $Q_1 = 20$, $Med = 25.5$, $Q_3 = 30.5$, and $maxX = 47$. You should try to work out and check all of these values yourself, to make sure you really understand the concepts from this chapter.

Going forward, I suggest that you work problems out yourself first, then use the calculator to quickly check your work. The real advantage of the calculator is when you are pressed for time on a major test. You can let the calculator do its job (giving quick, accurate information) and you can spend more time thinking, understanding, analyzing, applying, and evaluating the concepts. As you advance in higher mathematics and rigorous college math, you will be expected to understand and explain the concepts and use tools quickly.

The TI-83 (and similar calculators) can show three graphs from this chapter, histograms and boxplots. To display a graph, the data must first be input into a list. To get to the statistical graph menu, press 2nd Y=, which gets us to the STAT PLOTS menu. There

are several plots that can be activated separately or at the same time. Select the first item textbfPlot1 and hit ENTER. This will bring us to the controls for Plot1. Move the cursor over the **On** choice and press ENTER to make **On** highlighted.

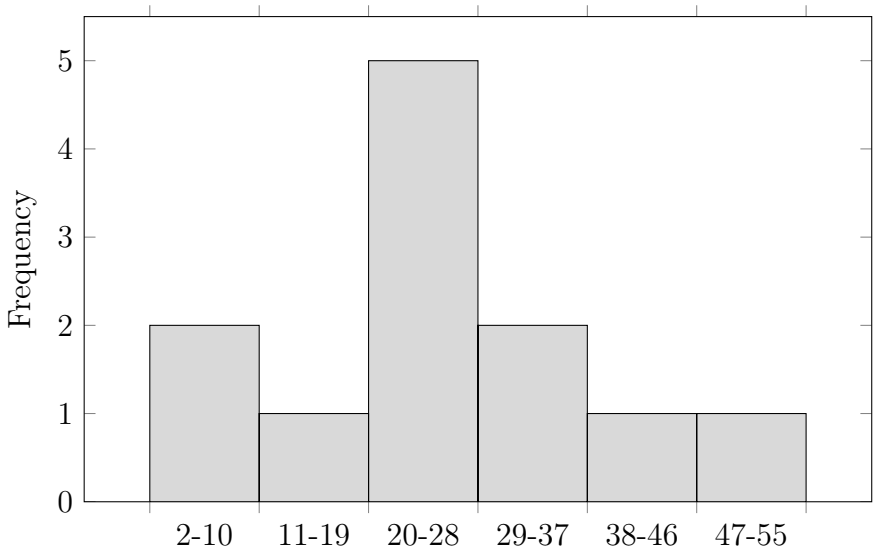
Use arrows to move down to the **Type** and scroll right to select the desired graph type. What is confusing, is that there appears to be two rows of icons for graphs, but this is just a continuation of the same row. If you try to arrow down to the second row, the cursor will move down to the next item on the menu. Just keep scrolling to the right instead, and the cursor will automatically move to the lower group of graph icons. As you may recognize, the third icon is a histogram, the fourth is a modified boxplot (shows outliers), and the fifth is a regular boxplot (no outliers shown). Select the modified boxplot.

The three graph types mentioned above, only use one data set (one list), so make sure the item **Xlist** is set to the appropriate list (L_1, etc.) that you want to use for that particular plot. The item **Mark**, is just a preference for the outlier symbol. Once you have all of the items set for the Plot1, you press the GRAPH button to see it. The GRAPH button is on the top row of calculator buttons, just below the screen on the right.

If you do not see your graph, do not worry. Most likely this is due to the data having larger values than what the screen is set for. To have the calculator automatically resize the screen to fit a stats plot, press the ZOOM button and scroll down to the item **ZoomStat** and press ENTER.

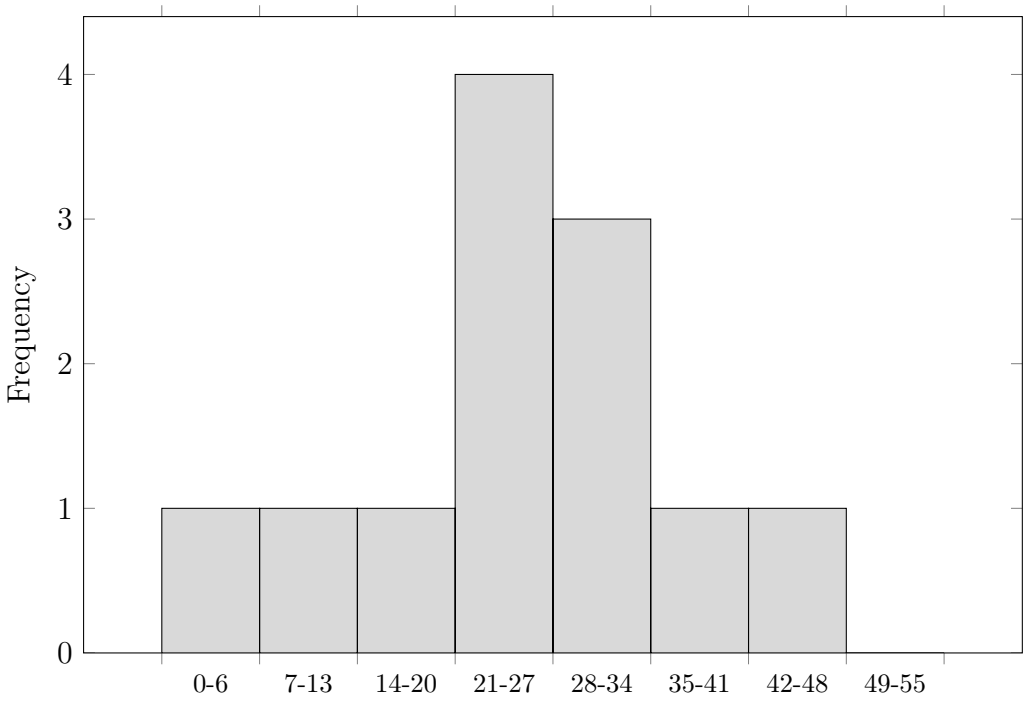
If you still have the data set $21, 22, 8, 19, 24, 2, 47, 30, 27, 28, 31, 40$ stored in L_1, try displaying the following graphs:

For histogram, press 2nd Y= (for stat plot menu), choose Plot1, turn it On, select type icon for histogram, set Xlist to L_1, then press GRAPH. If you do not see it, go to ZOOM and select ZoomStat, hit enter. You should see something like this:

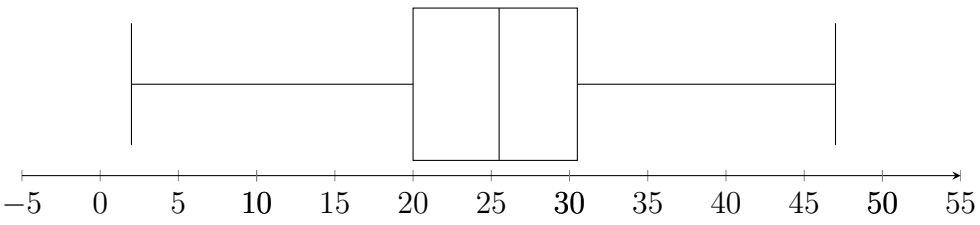


You can see graph values by pressing the TRACE button and using arrows to scroll across the graph. The values are displayed below the graph. The first bar has $min = 2$ and $max < 11$ and $n = 2$. This is the first class from 2 to 10 (less than 11) with a height of 2 (two values in this class).

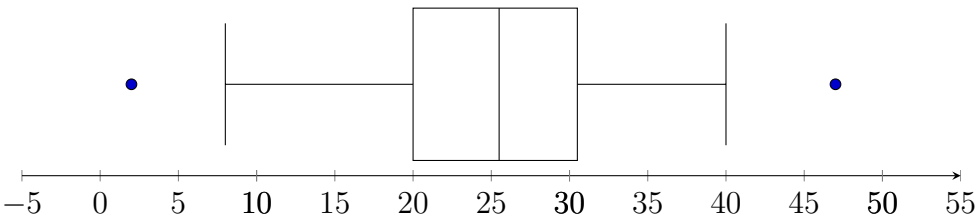
The calculator sets the number of bars and the classes (grouping intervals) according to some generic programming. If you wish to have more bars (more classes) and set the width of the classes, then press the WINDOW button and change the X-axis values. Let's try Xmin=0, Xmax=56, and Xscl=7. This sets intervals of every 7 units starting at 0 and going to 56, so there will be 8 bars (some may be zero height). The new histogram should look like this:


Frequency
0
1
2
3
4
0-6
7-13
14-20
21-27
28-34
35-41
42-48
49-55

For regular boxplot, press 2nd Y= (for stat plot menu), choose Plot1, select type icon for regular boxplot (no outliers), then press GRAPH. If you do not see it, go to ZOOM and select ZoomStat, hit enter. You should see something like this:



For modified boxplot, press 2nd Y= (for stat plot menu), choose Plot1, select type icon for modified boxplot (with outliers), then press GRAPH. If you do not see it, go to ZOOM and select ZoomStat, hit enter. You should see something like this:



To compare two boxplots, you can turn on two plots (for example: Plot1 and Plot2) and they will graph in the same window, using the same scale. This can help determine the similarities and differences between the data sets as we saw in the previous section.

Chapter 2

Probability

2.1 Probability Basics

2.1.1 Calculating Probability

In the first chapter, we learned about descriptive statistical methods for summarizing and displaying data, and descriptive measures (mean, variance, percentiles, etc.). In most studies, information about the population is the goal, but data is usually only collected from a sample, due to a census being too difficult to do. Even for single observations, we would like to know if and when something will happen. This leads us to the mathematics behind uncertainty. The science of uncertainty is called **Probability**. There are some important terms we need to know.

An **Experiment** is an action or procedure whose outcome cannot be predicted with certainty.

An **Event** is some specified result that may or may not occur when an action or procedure is performed. For example, a roll of a 6-sided die is an action. Getting a result of an even number is an event. An event is a collection of results or outcomes. In the die rolling

example, the event of "even" is the collection of outcomes of the numbers 2, 4, or 6. An outcome that cannot be further broken down into components is called a **Simple Event**.

The **Sample Space** is the set of all possible outcomes of an experiment. The number of possible outcomes for a sample space is denoted by the capital letter N. All events are subsets of the sample space.

Generally we assign capital letters to represent events (A, B, C, etc.). The letter can be something meaningful such as E for even. We might assign letters in the following way. Let J = the event that someone chosen from the class is a junior. Let F = the event that someone chosen from the class is a freshman. Now we can refer to J and F instead of writing out the events.

The probability value (or just probability) of an event, is the chance that the event will happen, relative to all of the possible outcomes. There are three basic properties for probability values:

1. The probability of an event is always between 0 and 1, inclusive (or 0% to 100%).
2. The probability of an event that cannot occur is equal to 0 (the event is said to be impossible).
3. The probability of an event that must occur is equal to 1 (the event is said to be certain).

When referring to the probability of an event, the capital letter P is used with the associated event (or its assigned letter) inside parentheses and placed next to the capital P. For example the probability of event being a junior would be shown as $P(J)$.

There are three commonly used ways to calculate a probability value. The particular situation and the level of detail desired, determines which way to use.

Rule #1 for calculating a probability is the **Empirical Probability**. The probability is based on the actual results observed for some number of trials of an experiment. This is similar to relative frequency. The formula for the empirical probability of event A is $P(A) = \frac{f}{n}$, where f is the number of times the event occurred (like frequency) and n is the total number of trials of the experiment.

Example: If a women makes 14 out of 20 free-throws in her WNBA basketball tryout, what is her empirical probability of making a free-throw?

Solution: Here $n = 20$ and the number of shots she made was $f = 14$. Let F be the event of making a free-throw. Then $P(F) = \frac{14}{20} = 0.7 = 70\%$. Assuming this is what she normally does, then for the near future, her probability of making a free-throw can be estimated as 70%. If she makes the team and practices, the probability will hopefully go up.

Rule #2 for calculating a probability is the **Theoretical Probability**. This is a logical approach, but only applies to equally likely outcomes. The formula for the theoretical probability of event A is $P(A) = \frac{e}{N}$, where e is the number of possible outcomes that fall under the event and N is the total number of possible outcomes in the sample space.

Example: A father buys two raffle tickets for his son's baseball team raffle. There were 80 tickets sold, and one will be picked out of a hat as the winner. What is the theoretical probability of the father winning the raffle?

Solution: Here the number of possible tickets that could be picked is $N = 80$ and the number of tickets the father has is $e = 2$. Let W be the event of the father winning. Then $P(W) = \frac{2}{80} = 0.025 = 2.5\%$. Notice this probability was calculated BEFORE the ticket was picked. Theoretical can be reasoned out based on logic, but only when outcomes are equally likely, such as picking a ticket out of a hat.

The empirical probability of the free-throws in the previous example could not be reasoned out before we had the data of the woman's attempts. Also, making or missing the free-throw are not equally likely for most people, so empirical probability was done for that and not theoretical.

Rule #3 for calculating a probability is the **Subjective Probability**. Subjective probability is estimated by using personal knowledge or experience. It is not scientific nor mathematical and rarely logical. For example, if a young teen sneaks out of his house twice, late at night, without getting caught, he might assume his chances of getting caught are very low. Another example of this is betting on your favorite team (or against the opposing team). Most people bet with emotion and not logic.

****Try this on your own**: In each situation below, calculate the probability, deciding whether to use empirical or theoretical probability.

1. 200 people are at a banquet and 8 people are at your table including you. What is the probability that someone at your table is chosen at random to win a prize out of the entire banquet?

2. Danny has played 20 tennis matches this season and has won 17 of them. What is the probability that he wins his next match?

Sometimes rule #1 and #2 give about the same value, especially in the long run. This is known as the **Law of Large Numbers**, which states that as an experiment is repeated again and again, the empirical (relative frequency) probability of an event TENDS to approach the theoretical probability. For example, the theoretical probability of flipping a coin and getting the result 'Tails' is $\frac{1}{2}$ or 50% (it is one out of two equal outcomes). If you flip a coin once, the outcome will either be 100% heads or 100% tails, never 50%, but as you flip a coin many times, most likely it will be close to even for number of heads and tails.

Here is an experiment for you to try, that will demonstrate the Law of Large Numbers. Find any coin that you can distinguish one side as heads and the other as tails. Make a table on paper with 5 columns: Flips, heads, P(H), tails, P(T). For each step, fill in the values in each column, along a row and then go to the next step/row.

In step one, flip the coin five times and in the first row, put 5 under flips. Under heads, write down however many heads came up for you and same for tails. Calculate $P(H) = \frac{heads}{5}$ and $P(T) = \frac{tails}{5}$, multiplying them by 100 and rounding to nearest whole percent. For example, if you get 3 heads and 2 tails, then $P(H) = \frac{3}{5} \times 100 = 60\%$, etc. With five flips, it will be impossible to get 50%, and you could easily get values far from it.

In step two flip the coin 25 times and in the second row, put 25 under flips. Under heads, write down however many heads came up for you and same for tails. Calculate $P(H) = \frac{heads}{25}$ and $P(T) = \frac{tails}{25}$, multiplying them by 100 and rounding to nearest whole percent. For example, if you get 11 heads and 14 tails, then $P(H) = \frac{11}{25} \times 100 = 44\%$, etc. With twenty-five flips, it will be impossible to get exactly 50%, but the values are most likely somewhat close. Since this is random, it could be farther away from 50% that in the first step, but unlikely.

In step three flip the coin 150 times and in the second row, put 150 under flips. Under heads, write down however many heads came up for you and same for tails. Calculate $P(H) = \frac{heads}{150}$ and $P(T) = \frac{tails}{150}$, multiplying them by 100 and rounding to nearest whole percent. For example, if you get 68 heads and 82 tails, then $P(H) = \frac{68}{150} \times 100 = 45\%$, etc. With 150 flips, it is possible to get exactly 50%, but the values are more likely to be somewhat close and probably closer that in the previous steps.

In rare cases, your values may have gotten further away instead of approaching 50%. That is why the Law states the values TEND to approach, but anything could happen. If many people did this experiment, most of them would see the values get closer to 50%. You may or

may not have seen this happen. If not, try it again and I'm pretty sure it will work this time.

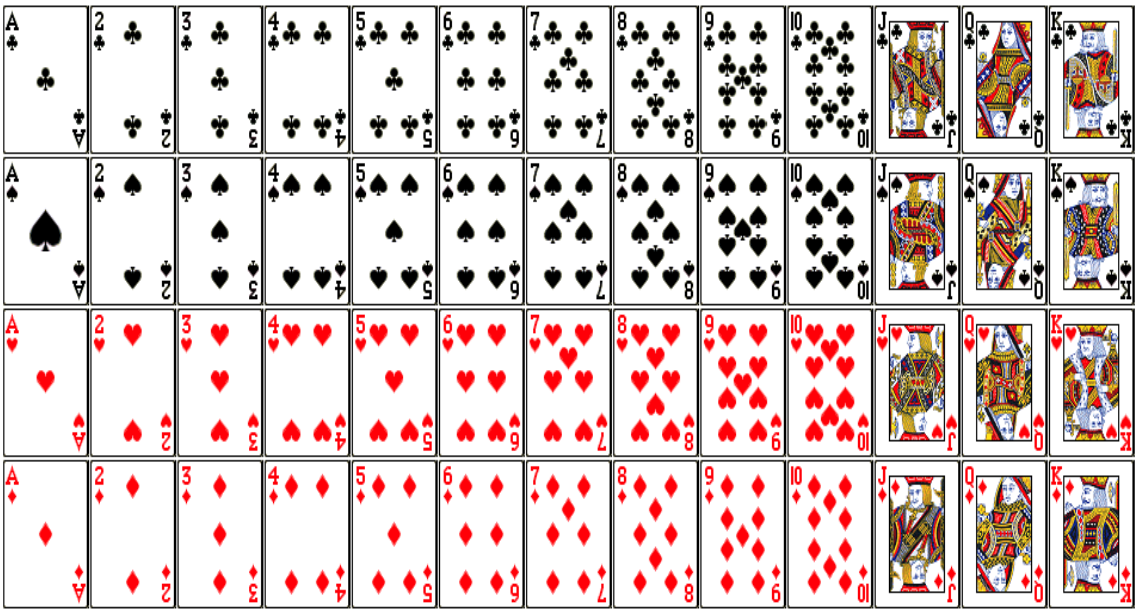
You may be thinking that events, outcomes, and sample spaces seem to be related to sets. Well, they are actually sets. Much of what we learned in the beginning of the chapter about sets, will apply to probability and events. The Complement of an event A consists of all outcomes in which event A does NOT occur. It is denoted by $\overline{A}$. Since every outcome must either belong to set A or its complement, there is a special relationship for the probabilities. The **Complement Rule** is stated as a formula: $P(A) + P(\overline{A}) = 1$ or 100%. So once you know the probability value of one, the other is automatically determined.

Example: A bag contains 29 marbles: 3 blue, 4 red, 6 yellow, 7 orange, 4 brown, and 5 green marbles. Calculate the theoretical probability of picking a marble that is not blue. Then use the complement rule to find the same probability and compare these.

Solution: $P(\text{not blue}) = \dfrac{\#\text{not blue}}{total} = \dfrac{4+6+7+4+5}{29} = 0.897 = 90\%$
Using complement rule $P(\overline{B}) = 1 - P(B) = 1 - \frac{3}{29} = \frac{26}{29} = 0.897 = 90\%$, which is the same value as it should be, since this is the same event.

Another classic example of simple probabilities are picking a random card from a standard deck of playing cards. A standard deck has 52 cards, half are black and half are red. See the image on the next page for what a standard deck might look like.

The cards are divided equally into 4 types of shapes shown on the cards, 13 cards have clubs, 13 have spades, 13 have hearts, and 13 have diamonds. The clubs and spades are the black cards. The hearts and diamonds are the red cards. With each shape, each card has a value on it from 1 to 13. The 1 card is called the "ace" and has letter A instead of the number 1. There are numbers on 2 through 10. The cards for 11,12, and 13 are special cards with faces and names Jack, Queen, and King. They have the letters J, Q, and K.



Example: For a random pick of one card from a standard deck, find the probability of picking a face card, then the probability of picking a 7 card. Reduce to fractions in lowest terms.

Solution: There are 12 face cards (the 4 jacks, 4 queens, and 4 kings). $P(face) = \frac{12}{52} = \frac{3}{13}$. There are 4 sevens. $P(7) = \frac{4}{52} = \frac{1}{13}$.

2.1.2 Odds

In everyday life, most people do not fully understand probability and often get it confused with a related idea called **Odds**. Odds are **NOT** direct probabilities themselves. They are the ratio of the probability that an event occurs to the probability that the event does not occur. So odds can be (and often are) greater than one. Odds are usually written as a ratio of two whole numbers and not shown as a single number, a decimal, and never a percent. Percent makes no sense for odds. Odds are commonly shown as a ratio with a colon between the values, and the values as whole numbers in lowest ratio.

The formulas for odds are:

Odds in favor of event A $= \dfrac{P(A)}{P\left(\overline{A}\right)}$ Odds Against event A $= \dfrac{P\left(\overline{A}\right)}{P(A)}$

Odds can be shown is three forms. During the computation the best form is a fraction $\dfrac{A}{B}$, which should be reduced so that A and B are whole numbers with no common factors. If B reduces to 1, DO NOT leave it out. Odds of $\dfrac{3}{1}$ should NEVER be shown simply as 3. Remember that odds are a relationship between to probabilities, so should always have two numbers. The other two forms are as a relation with the word "to", written as A to B, or with a colon ":", written as $A : B$. Both of these show A and B as reduced whole numbers with no common factors.

For example odds against of $\dfrac{8}{2}$ would be reduced to $\dfrac{4}{1}$ or 4 to 1 or $4 : 1$.

Example: Calculate the odds against, and the odds in favor of, picking a blue marble from the marble bag in the previous example.

Solution: Odds in favor of blue $= \dfrac{P(B)}{P\left(\overline{B}\right)} = \dfrac{\frac{3}{29}}{\frac{26}{29}} = \dfrac{3}{26}$ or $3 : 26$ odds in favor. We can state this as 3 to 26 odds in favor of blue (not very good odds). This means that out of every 29 picks, only 3 are likely to be blue and 26 not blue.

Odds against blue $= \dfrac{P\left(\overline{B}\right)}{P(B)} = \dfrac{\frac{26}{29}}{\frac{3}{29}} = \dfrac{26}{3}$ or $26 : 3$ odds against. We can state this as 26 to 3 odds against blue. Notice that odds against is just the reciprocal of the odds in favor (flip the fraction over). This makes it easy to calculate odds once you have one of them.

We can also find the probabilities from the odds. When odds in favor are shown as $A : B$ then the probability for the event occurring is $P(for) = \dfrac{A}{A+B}$ and the probability of the event not occurring is $P(not) = \dfrac{B}{A+B}$. The two probabilities are complements of each other and should total to 1 or 100%.

Example: If the odds in favor of a person surviving a risky surgery are $3 : 8$, calculate probability of the person surviving and the probability they do not survive.

Solution: We can think of the numbers 3 : 8 and out of 11 such surgeries, the patient will survive in 3 of them, but not survive in the other 8. The probability that a patient survives is $P(\text{survive}) = \dfrac{A}{A+B} = \dfrac{3}{3+8} = \dfrac{3}{11} = 27.3\%$. The probability that a patient does not survive is $P(\text{not survive}) = \dfrac{B}{A+B} = \dfrac{8}{3+8} = \dfrac{8}{11} = 72.7\%$. Notice the two probabilities do add to 1 or 100%.

****Try this on your own**: If a team has a 65% chance of winning, find the odds for and against winning.

2.1.3 Exercises: Probability Basics

Solutions appear at the end of this textbook.

1. List the sample spaces for the following experiments:

   (a) Flipping a coin once

   (b) Flipping a coin 3 times

   (c) Rolling a 6-sided die once

   (d) Rolling two 6-sided dice and computing the sum of the dice

   (e) Randomly picking a color of the rainbow and a season of the year

2. Which of the following are valid values for a probability?

   $$0.35,\ 0.004,\ 1.23,\ 213\%,\ -0.25,\ \tfrac{3}{8},\ \tfrac{8}{3}$$

3. A student is about to roll a 6-sided die. Find the following theoretical probabilities. P(even), P(3), P(>2).

4. John is trying out for the basketball team. He shoots from the foul line 12 times. His results are: make, miss, miss, make, miss, make, miss, miss, miss, make, make, miss. What is the empirical probability of John making a shot from the foul line? What is your subjective probability of John making it onto the team?

5. If the probabilities for the types of precipitation are as follows, what is the probability of no precipitation? P(rain) = 0.20, P(snow) = 0.10, P(sleet) = 0.15.

6. For a standard deck of playing cards, find the following probabilities for picking one card at random.

   a) P(heart)

   b) P(ace)

   c) P(red)

   d) P(# from 2 to 10)

   e) P(black king)

   f) P(red club)

7. If the probability of rain today is 20%, find the odds in favor of and against rain.

8. If the odds against a team winning are $9 : 5$, calculate probability of the team winning and the probability they do not win.

9. Find a 6-sided die. Roll it 15 times and compute the empirical probabilities of rolling each number (1 to 6). Then compute the theoretical probabilities of rolling each number. Compare the empirical to the theoretical. How different are they? Explain why or not.

2.2 Counting Rules

2.2.1 Combinations

You have been counting since you were a toddler. It is pretty easy to do, or so you thought! Counting can get very complicated, especially if you had to count how many different combinations of clothing you could put together from 12 shirts, 8 pants, 3 belts, and 6 pair of shoes. Believe it or not, there are actually $1,728$ different combinations of clothing (choosing one of each type). The long way to reach that value is to count each combination one at a time. The short way is to use some really cool math rules for counting.

The first one is called the **Fundamental Counting Principle**. It states that when picking one item each, from several groups of items, the total number of combinations of those items is equal to the product of how many items are in each type. In simple terms, multiply the number of items of the first type by the number of items of the second type, etc.

For the clothing example, the answer was calculated using the Fundamental Counting Principle, $12 \times 8 \times 3 \times 6 = 1,728$.

For any number of items selected from a total, the number of possible **Combinations** for a selection of r elements drawn from a population of N elements, is found using the formula ${}_NC_r = \dfrac{N!}{r!(N-r)!}$, where the exclamation symbol, "!", is the factorial symbol. To make sure you understand the factorial notation, $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$. Factorial is a whole number multiplied by all whole numbers going down to 1.

Example: Compute ${}_{14}C_6$

Solution: ${}_{14}C_6 = \dfrac{14!}{6!(14-6)!} = \dfrac{14!}{6!(8!)} = \dfrac{14 \cdot 13 \cdot 12 \cdot 11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1(8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)}$
Now notice that we can reduce and simplify this. The numbers 8 down to 1 on the top, will cancel off the 8 to 1 on the bottom. So the combinations $= \dfrac{14 \cdot 13 \cdot 12 \cdot 11 \cdot 10 \cdot 9}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}$

To even further simplify, we could cancel 10 and 12 on top, using the 4,3,5, and 2 on bottom. This leaves $\frac{14 \cdot 13 \cdot 11 \cdot 9}{6 \cdot 1} = \frac{18018}{6} = 3003$. So there are 3,003 different combinations of 6 elements chosen from just 14. I don't know about you, but this seems crazy even though I know it is the correct value. It is amazing how large combinations can get, just picking from a small amount like 14.

Example: A student is considering taking the following subjects this semester: Math, Physics, Literature, Economics, Psychology, and Music. How many different 4 course combinations can this student possibly take?

Solution: The student must pick 4 out of 6 courses.

$$_6C_4 = \frac{6!}{4!(6-4)!} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{4 \cdot 3 \cdot 2 \cdot 1(2 \cdot 1)} = 15 \text{ possible schedules.}$$

2.2.2 Permutations

A **Permutation** is an ordering of r elements selected from a set of N distinct elements. The elements selected will be in r positions. For example selecting first, second, third in a talent contest. The number of permutations of r objects chosen from a possible N is found using the formula $_NP_r = \frac{N!}{(N-r)!}$. Here, a different ordering of the same picks, is considered a different permutation.

Example: How many different permutations can be selected from a group of 10 paintings, if the judges must select a first, second, and third place award?

Solution: There are three awards, so $_{10}P_3 = \frac{10!}{(10-3)!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 720$ possible ordering of three paintings for the awards.

There are some special cases of the counting rules.

1. It is necessary for mathematicians to define $0! = 1$. This makes sense in the context of the next special case, or else the formula would lead to no answer.

2. There is only one way to select no items from a group of items, that is to not select at all. As formulas: ${}_NP_0 = 1$ and ${}_NC_0 = 1$. Since nothing is selected, there is no difference between combination or permutation (no ordering possible).

3. There are as many ways to select one item, as there are items, ${}_NP_1 = N$ and ${}_NC_1 = N$.

4. The number of ways to select all items from a group of items is just to take them all, so only one combination is possible ${}_NC_N = 1$.

5. There are many ways to create an ordering of all items, ${}_NP_N = N!$.

6. Selecting some items, is the same as leaving behind the others, so ${}_NC_r = {}_NC_{N-r}$. For example, selecting 5 out of 12 items is the same as leaving behind the other 7. Either way you have separated the items into two groups, one with 5 and the other 7. Notice that ${}_{12}C_5 = {}_{12}C_7 = 792$

Sometimes when working on word problems, it can be difficult to know if it involves combinations or permutations. The way to tell is if there is any wording that denotes some specific order of the elements. Such as award place, rank of officers (president, VP, secretary), or order based on timing (eat a donut first, then a muffin next, etc.).

If you forget which formula to use, simply calculate them both. There are always more permutations than combinations, so the larger answer comes from permutation formula. This is easy to see for a simple case of picking 3 letters out of A, B, C, D. There are only 4 combinations (leave any one of the letters out). However, there are 24 permutations, since ACB and ABC are the same combination, but different ordering.

****Try this on your own**: For a lottery in which you pick five numbers from 1 to 50, how many different sets can you pick if they can be in any order, and if they must be in a specific order?

2.2.3 Probabilities and Counting

Counting rules are useful for finding probabilities when there are a large number of outcomes. A common situation is a lottery game.

Example: There is a popular lottery ticket game called Lotto. In each play, you choose 6 different numbers from 1 to 59. To win the big jackpot, all 6 of your numbers must match the winning combination. How many different combinations of 6 numbers can be played? What is the probability of winning the jackpot with each play? What are the odds against winning?

Solution: The number of different plays of 6 numbers out of 59, is ${}_{59}C_6 = 45,057,474$ (over 45 million!!). Only one combination is a winner. The probability of winning $P(Win) = \frac{1}{45057474} = 0.0000000222$ (extremely small!). By the complement rule, probability of not winning the jackpot is 0.9999998778 (very large!). The odds against winning are the ratio $\frac{\frac{45057473}{45057474}}{\frac{1}{45057474}} = 45{,}057{,}473$ to 1.

NOTE: most lottery rules will incorrectly use the word odds when they are stating the probability and vice-versa. Since the probabilities are so small, there is not much difference, and most people don't understand the difference, so they don't bother to be mathematically correct.

When dealing with combinations that are made up of different types, we need to get a count of how many combinations of each type, and how many combinations there are in total, then divide them.

For example, If you are dealt a hand of 5 cards in poker, what is the probability of being dealt a full house of three kings and two fives? By the Fundamental Counting Principle, we multiply the number of ways to pick 3 out of 4 kings, by the number of ways to pick 2 out of 4 fives. Each of these is a combination, so our numerator is ${}_4C_3 \times {}_4C_2$. The total number of sets of 5 is ${}_{52}C_5$. Now the probability is $\frac{{}_4C_3 \times {}_4C_2}{{}_{52}C_5} = \frac{4(6)}{2598960} = 0.000923\%$

****Try this on your own**: You and two friends entered a contest that will randomly pick 3 people to go to a concert. What is the probability that the 3 winners are you and your two friends?

2.2.4 Exercises: Counting Rules

Solutions appear at the end of this textbook.

1. How many different combo meals can you buy, if you get to pick one of five entrees, one of four sides, and one of three desserts?

2. How many different outfits can you put together, if you pick one of 12 shirts, one of 6 skirts, one of 8 pairs of shoes, and one of 3 belts?

3. Compute the following: ${}_8C_3$, ${}_{11}C_9$, ${}_7P_4$, ${}_8P_8$, ${}_5C_1$

4. How many different movie sequences can three friends watch, if they have 10 movies and only enough time to watch 3 of them?

5. The Fantasy 5 Lottery game consists of picking 5 different numbers from 1 to 39. How many sets of 5 numbers can be played?

6. What is the probability of winning the Fantasy 5 lottery jackpot (matching all 5 numbers), playing only once? What are the odds against winning?

7. A game pays \$1,000 if you match the exact 3 digit number from 1 to 999. What is the probability of not winning? Show as a percent rounded to 4 decimal places.

8. Why is it impossible to compute ${}_5C_9$?

9. If a lottery game awards a \$100 prize for matching any 3 out of 5 numbers, chosen from 1 to 29, what is the probability of winning?

10. If you are dealt a hand of 5 cards, what is the probability of being dealt four aces?

2.3 More Probability

2.3.1 The Addition Rule

What chance did you think you had of being done with probability? Well sorry, there is plenty more. Some people actually get a PhD doctoral degree in probability theory. We won't go that far here, but we will look at the next level of concepts.

Events that cannot occur at the same time for one outcome of an experiment, are called **Disjoint Events** or **Mutually Exclusive**. For example, when rolling a 6-sided die, the events are rolling a 3 and rolling a 4 are mutually exclusive, because you cannot roll two numbers on one roll. You could roll a 3, then roll a 4 on the next roll, but they cannot both occur for the same roll.

Since an event and its complement never have any outcomes in common, it should be clear that complementary events are mutually exclusive. When you take a test, you either pass or fail, you can't do both at the same time. An example of events that can happen at the same time are passing a test and getting an A on the test. They are not the same, but they do share outcomes (scores of 90+). Actually, getting an A is a subset of passing.

Here is where the concepts from sets and Venn diagrams will come into play. Since events are sets of outcomes, we can combine events to get compound events with intersections and unions.

The Intersection of two events, is the set of outcomes that are part of the first event AND part of the second event at the same time. It is the set of outcomes they share in common. The symbol for intersection is $\bigcap$.

The Union of two events, is the set of outcomes that are part of the first event OR part of the second event (or both). It is the set of outcomes from both combined into one larger set. The symbol for union is $\bigcup$.

First let's look at probabilities of compound events, from a logical or reasoning perspective. We can find $P(A\ or\ B)$ if we know the individual outcomes in each event (not just the probability values). We find the sum of the number of outcomes from A, and the number of outcomes from B, in such a way that every outcome is counted only once. Then divide this sum by the total number of outcomes in the sample space.

In a similar way, we can find $P(A\ and\ B)$ by finding the number of outcomes that A and B both share in common. Then divide this sum by the total number of outcomes in the sample space.

Example: A class consists of 14 boys (8 are juniors, 6 are seniors) and 12 girls (8 are juniors, 4 are seniors). If one student is to be selected at random to come up to the board, find the following probabilities: P(boy $\bigcup$ junior) and P(girl $\bigcap$ senior).

Solution: The event boy $\bigcup$ junior is the same as boy OR junior. There are 26 students total. There are 14 boys and 16 juniors which equals 30??? How can that be? Remember that 8 of the juniors are also boys, so when we count the students for our compound event, we should only count the 14 boys (which include 8 juniors) and then the other 8 juniors (girls) to get 22 students who are boys or juniors. Now probability is $\frac{22}{26} = \frac{11}{13} = 0.846 = 84.6\%$.

The event girl $\bigcap$ senior is the same as girl AND senior. There are 26 students total. There are 4 girls who are seniors. The probability P(girl $\bigcap$ senior)$= \frac{4}{26} = \frac{2}{13} = 0.154 = 15.4\%$.

If we have the probabilities of each event, then we can find the probabilities of compound events using formulas.

The **Addition Rule** states $P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$ or $P(A \bigcup B) = P(A) + P(B) - P(A \bigcap B)$, where $P(A\ or\ B)$ is the probability at least one of the events occur in an outcome, and $P(A\ and\ B)$ is the probability that both A and B occur at the same time in an outcome.

Example: If the probability of rain today is 0.7, the probability you forget your umbrella is 0.4, and the probability they both happen together is 0.3, what is the probability that it rains or you forget your umbrella?

Solution: P(rain or forget) $= P(rain) + P(forget) - P(both) = 0.7 + 0.4 - 0.3 = 0.8$ or 80%.

***Try this on your own**: A football team has 42 players. There are 18 players who play offense, 20 players who play defense, and 10 players who play on special teams. Six of the offensive players play both offense and special teams. Find the probability that a player is on the offense or special teams.

2.3.2 Conditional Probability and the Multiplication Rule

The addition rule requires that we know the intersection probability at the end of the formula. There is a rule for calculating the intersection probability directly, but before we work with that formula, we need to define **Conditional Probability**. The conditional probability of an event is the probability that results after another event has already happened and could affect the new event.

The logical way to compute conditional probability is to take into account what has already happened and adjust the possible outcomes accordingly. For example, your probability of passing a test depends upon certain conditions. If you studied well, the probability will likely increase. If you didn't realize there was a test and didn't study at all, then the probability will likely decrease. Another example could be, the probability of rolling a 5 on a 6-sided die is $\frac{1}{6}$, but the probability of rolling a 5 on a 6-sided die, after you know the roll is an even number, is 0, since 5 is odd.

If two events affect the occurrence of each other, they are said to be **Dependent**. If two events do not affect the occurrence of each other, they are said to be **Independent**.

When two events are dependent, their probabilities must be calculated using conditional probability. As a formula, the probability of the intersection of events A and B is given by $P(A \bigcap B) = P(A) \cdot P(B|A)$. This is known as the **Multiplication Rule**. The notation $B|A$ is read as "B given A". The formula states that the probability of both A and B, is equal to the probability of event A (considered to happen first) times the probability of event B, given that event A has already happened. P(B|A) is the conditional probability of event B, given A.

When two events are independent, then the condition of one happening does not matter, and the multiplication rule simplifies to $P(A \bigcap B) = P(A) \cdot P(B)$. Remember, this only happens for independent events.

Many probability problems deal with picking cards from a standard deck of playing cards. Here is description. A standard deck has 52 cards split into four symbols (called suits). The two red symbols are hearts and diamonds. The two black symbols are clubs and spades. Each suit has 13 cards with a rank (or value). The ranks are $2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A$. 'J' stands for jack, 'Q' for queen, 'K' for king, and 'A' for ace. The jack, queen and king are called face cards, because they usually have faces of people on them. There are 4 of each rank card, one from each suit (symbol).

Example: If two cards selected at random from a standard deck of playing cards, what is the probability of picking two aces?

Solution: The probability of picking the first ace is $P(ace) = \frac{4}{52} = \frac{1}{13}$. The probability of the second ace being picked, depends upon the condition of an ace already being picked. P(2nd ace|1st ace) $= \frac{3}{51}$, since there would be 3 aces out 51 remaining cards. Therefore, $P(ace \bigcap ace) = \frac{1}{13} \cdot \frac{3}{51} = \frac{3}{663} = 0.005 = 0.5\%$.

Example: If two 6-sided dice are rolled, what is the probability of rolling two ones?

Solution: Here the two dice have no affect on each other, they are independent rolls. $P(one \bigcap one) = P(one) \cdot P(one) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} = 0.028 = 2.8\%$.

In some situations, the conditional probability may be unknown and you wish to compute it. If the intersection probability is known, then we can rearrange the multiplication rule to find the probability of event B under the condition that event A had already happened, by $P(B|A) = \dfrac{P(A \bigcap B)}{P(A)}$. The formula states that the probability of event B, given that event A has already happened, is equal to the probability both events happen divided by the probability of event A.

Example: If one card is selected at random from a standard deck of playing cards, what is the probability of picking a Jack, given that the card picked is a face card?

Solution: Without any conditions being known, the probability of picking a jack would simply be $\frac{4}{52}$. Under these conditions, $P(jack|face) = \dfrac{P(jack \bigcap face)}{P(face)} = \dfrac{\frac{4}{52}}{\frac{12}{52}} = \dfrac{4}{12} = \dfrac{1}{3}$. Once you know it is a face card, it is more likely to be a jack, that just getting a jack out of all cards.

It is helpful to know all of the outcomes in a sample space for an experiment and their corresponding probabilities. A table which lists all outcomes and the probabilities is know as a **Probability Distribution**. There are two requirements for a valid probability distribution. Each probability must be between 0 and 1 (0% and 100%). The sum of all the probabilities must equal 1 or 100%. That way you know every outcome has been included properly.

Example: Which of the two tables are valid probability distributions (if any)? Why or why not?

Day	Monday	Tuesday	Wednesday	Thursday	Friday
Probability	0.22	0.13	0.34	0.24	0.07

Value	1	2	3	4	5	6
Probability	-0.3	0.3	0.4	0.4	0.1	0.1

Solution: The first table is a valid probability distribution, since each probability value is between 0 and 1, and they add up to 1. The second table is not. The probabilities do add up to 1, but one of them is negative.

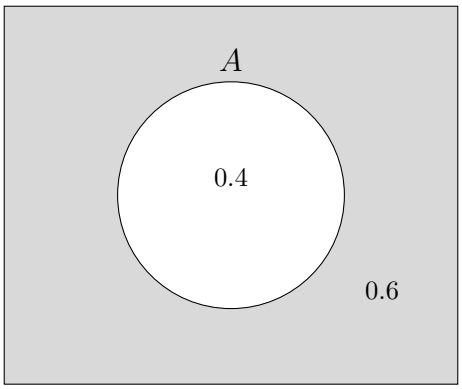
2.3.3 Venn Diagrams

There are special diagrams used in probability to represent events, called **Venn Diagrams**, named after the British mathematician John Venn. They can be shown in many ways, but most common is to show the sample space as a large rectangle, with the events you want to feature shown as circles inside the rectangle. The compound events will be the parts of the circles, or outside the circles, or where they overlap, etc. Even if there are many events in a problem, the events included as circles should only be the ones that are relevant to the particular problem we are trying to answer. Never draw complements as circles, they will be the outside of the events they are complements of.

Instead of showing all of the elements in each set, Venn diagrams usually just have the probability values shown in the diagram, placed inside or outside the appropriate event circles (or parts of circles). The corresponding area which represents any compound event being focused on, is usually shaded. The entire diagram should have a total probability value of 1 (or 100%).

Example: Show the Venn diagram for the event A and its complement, where $P(A) = 0.4$.

Solution: For the complement of A, we draw a rectangle with one circle that represents event A. We place the probability value of A inside the circle, and the complement value outside the circle. $P(\overline{A}) = 1 - P(A) = 1 - 0.4 = 0.6$. Then shade outside the circle so we know we are referring to the complement of A.



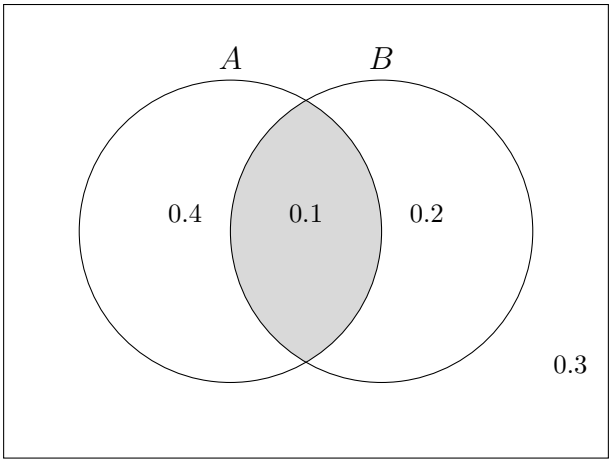
$\overline{A}$ or complement of A

When we have two events, we draw the rectangle with overlapping circles for the events, unless we know that they are mutually exclusive. We will then have an overlapping area for the intersection in the center and partial areas (crescent moon shapes) for the section of each event that is exclusive to each event. We must split the probabilities into each part of the diagram. Always start with the intersection and then subtract that value from each event probability to find the value for each exclusive part (the crescent moons).

Example: Show the Venn diagram for the compound A *and* B, where $P(A) = 0.5$, $P(B) = 0.3$, and $P(A \bigcap B) = 0.1$.

Solution: Here we draw a rectangle with two overlapping circles that represent events A and B. We place the probability value of $P(A \bigcap B) = 0.1$ inside the intersection piece in the center. Then event A has a total probability of 0.5, but 0.1 is already in circle A. The remaining probability $0.5 - 0.1 = 0.4$ goes inside the extra part of A (the crescent moon).

We do the same for the other part of B with probability $0.3 - 0.1 = 0.2$. Then shade the intersection only so we know we are referring to compound event $A \bigcap B$.



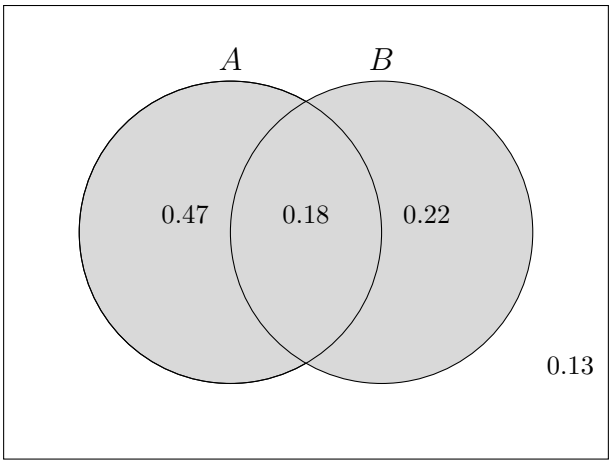
$A \bigcap B$ or intersection of A and B

The three probabilities inside the circles add to 0.7, so the remaining probability of $1 - 0.7 = 0.3$ goes in the outer area of the rectangle. Notice this will be the probability of neither event A or B. Shown in symbols as $P(\overline{A \text{ or } B}) = 0.3$

To see how Venn diagrams can be used to solve for probabilities, see the example below.

Example: Use a Venn diagram to calculate $P(A \bigcup B)$, where $P(A) = 0.65$, $P(B) = 0.4$, and $P(A \bigcap B) = 0.18$.

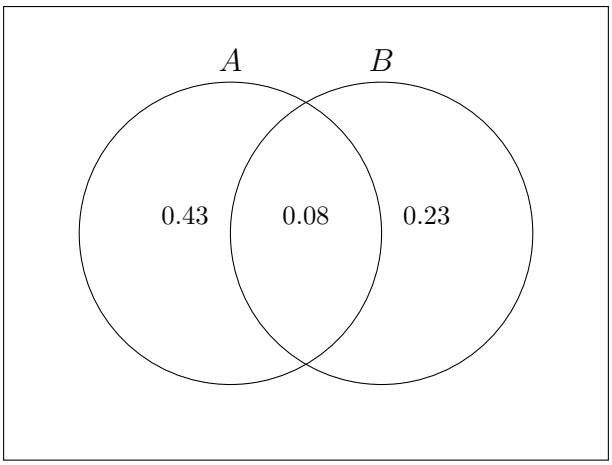
Solution: For the union, we draw same setup as the intersection, with the partial probabilities. We place the probability value of $P(A \bigcap B) = 0.18$ inside the intersection piece in the center. Then event A has a total probability of 0.65, but 0.18 is already in circle A. The remaining probability $0.65 - 0.18 = 0.47$ goes inside the extra part of A (the crescent moon). We do the same for the other part of B with probability $0.4 - 0.18 = 0.22$. The remaining probability of $1 - 0.47 - 0.18 - 0.22 = 0.13$ goes in the outer area of the rectangle and we shade both circles (the union).



$A \bigcup B$ or union of A or B

The probability of A or B can be found by adding all pieces inside the shaded area. $P(A \bigcup B) = 0.47 + 0.18 + 0.22 = 0.87$. We can double check this answer by using the addition rule $P(A \bigcup B) = P(A)+P(B)-P(A \bigcap B) = 0.65+0.4-0.18 = 0.87$, which matches.

***Try this on your own**: Use the Venn diagram below to calculate $P(B)$, as well as the probability of neither A nor B, $P(\overline{A \text{ or } B})$.



2.3.4 Exercises: More Probability

Solutions appear at the end of this textbook.

1. Give an example of two events that are mutually exclusive.

2. At Mega University there are 32 physics majors, 49 math majors, and 112 engineering majors. Out of these, 8 are double majors in physics and math, and 14 are double majors in physics and engineering. Find the probability that one of these students selected at random is a physics or engineering major.

3. Given $P(A) = 0.5$, $P(B) = 0.7$, and $P(A \textit{ and } B) = 0.3$. Find $P(A \textit{ or } B)$.

4. Given $P(A) = 0.65$, $P(A \textit{ or } B) = 0.85$, and $P(A \textit{ and } B) = 0.25$. Find $P(B)$.

5. If two cards are picked at random from a deck of cards, what is the probability of picking two red sixes?

6. If one card is picked at random from a deck of cards, what is the probability of picking the ten of hearts, given that you know the card is red?

7. Give an example of two events that are independent.

8. Can two events be both mutually exclusive and independent? Explain.

9. Is this a valid probability distribution? Why or why not?

Meal	Pizza	Chicken	Steak	Pasta	Fish
Probability	0.48	0.12	0.12	0.20	0.05

10. Use a Venn diagram to calculate $P(B)$, where $P(A) = 0.72$, $P(A \bigcup B) = 0.91$,, and $P(A \bigcap B) = 0.12$.

Chapter 3

Probability Distributions

3.1 Discrete Distributions

3.1.1 Discrete Random Variables

Instead of having a data set of observations, sometimes we only know a probability distribution for the possible values of a variable. The probabilities could be based on prior data or on some pattern or formula from theory. Either way, a variable is unknown until it takes on a specific value. So most variables are random in the sense that we don't know its value until it occurs. This is known as a **Random Variable**.

For example, in a college classroom, each student has a class level such as freshman, sophomore, junior, or senior. However until we choose a specific student and ask them, we don't know which level they are. Even after we survey all students in the room, we would only know how many of each level, but until we look at a specific student, we don't know which level a random student would have. The variable for a student's class level in this case is a qualitative random variable, since the variable takes on category values and not numbers.

A college class of 30 students could have the distribution of class level as follows.

Level - X	Fresh	Soph	Junior	Senior
prob - P(x)	0.47	0.33	0.17	0.03

The most frequent value of class level is freshman, so that is the mode. We cannot calculate other statistics like mean or median, since the data is qualitative (categories). However, the sum of all the probabilities must total to 1 or 100%, so we can find a missing probability.

Example: The distribution of blood types for people in the U.S. as of 2021 is shown in the table below. Find the probability that a person has A-negative blood type.

Type	O+	A+	B+	AB+	O-	A-	B-	AB-
P(x)	38%	34%	9%	3%	7%	?	2%	1%

Solution: simply add the percentages and subtract from 100 to find the missing probability. P(A-)=100-94=6%

When data is quantitative (numbers) and discrete, the variable is known as a **Discrete Random Variable**. We can then calculate statistics as follows.

1. The mode will be the value with the greatest probability.
2. To find the quartiles (and median) look at cumulative probabilities and where the cumulative probabilities reach 0.25, 0.50, and 0.75. Those values will be the first quartile Q_1, median, and third quartile Q_3.
3. The mean, μ, is the weighted mean $\mu = \sum x * P(x)$
4. The standard deviation $\sigma = \sqrt{(\sum x^2 * P(x)) - \mu^2}$

Example: A survey was taken asking 835 college students how many classes they were taking this semester. Find the mode, median, and quartiles of the number of classes taken.

x	1	2	3	4	5	6	7
P(x)	0.04	0.08	0.15	0.34	0.29	0.08	0.02

Solution: We can add a new row to show the cumulative probabilities.

x # classes	1	2	3	4	5	6	7
P(x)	0.04	0.08	0.15	0.34	0.29	0.08	0.02
cum P(x)	0.04	0.12	0.27	0.61	0.90	0.98	1.00

The greatest probability is 0.34, so the mode is 4 classes. The cumulative probability reaches 0.25 within the x-value of 3, so $Q_1 = 3$. The cumulative probability reaches 0.50 within the x-value of 4, so $med = 4$. The cumulative probability reaches 0.75 within the x-value of 5, so $Q_3 = 5$.

Example: Major League Baseball has the World Series to decide the champion team. It is a best of seven game series, so the first team to win four games, wins the title. Below is the distribution for the number of total games played to have a winner, based on the results from 1950-2021. Find the mean and standard deviation of the number of games in a world series.

x - # games	4	5	6	7
P(x)	0.18	0.18	0.23	0.41

Solution: We can add new rows to show the weighted values and take their sums.

x - # games	4	5	6	7
P(x)	0.18	0.18	0.23	0.41
x*P(x)	0.72	0.90	1.38	2.87
$x^2 * P(x)$	2.88	4.50	8.28	20.09

$\mu = \sum x * P(x) = 0.72 + 0.90 + 1.38 + 2.87 = 5.87$, so the average number of games played in the world series is about 6 games. The standard deviation of the number of games played in each world series is $\sigma = \sqrt{(\sum x^2 * P(x)) - \mu^2} = \sqrt{35.75 - 5.87^2} = 1.14$.

****Try this on your own**: Below is the distribution for the shoe size of the players on a college basketball team. Find the missing probability and then the mean and standard deviation of the shoe size.

shoe size	10	10.5	11	11.5	12	12.5	13	14
P(x)	.063	.063	.189	.124	.250	.124	.063	?

3.1.2 Expected Value

For a quantitative variable, we can use the probability distribution to find the **Expected Value**. The expected value is like an average, weighted by the probabilities. It gives the typical value of the variable over many observations. It is used everyday in many ways in the real world: to set prices for insurance, to choose investments, and in most of the sciences. The formula for the expected value of a variable x is $E(x) = \sum x \cdot p(x)$, where $p(x)$ is the probability of a value x.

Example: A particular insurance policy has a claim distribution shown below. Find the expected value of a claim for a one year period. Then compute the price of the monthly

insurance premium, if the company will charge 10% profit margin and the premium is paid over 12 months.

Claim Amount	$0	$500	$1,500	$5,000
Probability	0.75	0.15	0.07	0.03

Solution: $E(x) = 0(.75) + 500(.15) + 1500(.07) + 5000(.03) = \$330$, so the insurance company expects to pay out \$330 on average each year, for every policy it sells. Some policies will pay out more (large claim of \$5,000), most less (\$0). In order to stay in business, the company must charge more than \$330. The annual profit margin they charge will be $10\%(330) = \$33$, so the annual premium is \$363. Then the monthly premium is $\frac{\$363}{12} = \$30.25$. This is a simplified example of insurance, but the concept is the same as what insurance companies use everyday.

Another situation where expected value is used, is for games of chance where the players can win money, such as casino games and lottery tickets. The mean is then the value that a player can expect to win on average over a large number of attempts. The casinos and lottery creators know this value and set the cost to play at a value greater than the expected value. They have to do this, in order for them to make a profit and continue to offer the games.

Example: The table below shows the prizes and probabilities for the lottery game called "Georgia FIVE". Find the expected value.

Prize	$10,000	$225	$21	$20	$11	$10	$2	$1	$0
P(x)	.00001	.00018	.00018	.00171	.00162	.01613	.00813	.16667	.80537

Solution: The expected value $\mu = \sum x * P(x) = 0.54$, so when a person plays many times, they can expect to win on average \$0.54 each game. The game costs \$1.00 to play, therefore the Georgia lottery gains about \$0.46 for each play. Some of that goes to adminis-

trative costs. Most of the profits go to fund pre-K, and college HOPE scholarships in Georgia.

****Try this on your own**: A particular game has the prize distribution shown below. Find the expected value of a prize.

Prize Amount	$0	$25	$100	$500
Probability	0.7	0.2	0.09	0.01

3.1.3 Binomial Distribution

Some discrete random variables have probabilities that follow a pattern, so they can be calculated by formulas and are usually easier to deal with. One such variable is a **Binomial Variable**, which is the count of successes of a **Binomial Experiment**. Probabilities for binomial variables follow a pattern called the **Binomial Distribution**. The conditions for a binomial experiment are the following.

1. There are a fixed number of trials (attempts) of the experiment, represented by the letter n.

2. Each trial (attempt) is independent of each other. The outcome of one attempt has no affect on the outcome of any other attempt.

3. Each attempt can be classified into two types of outcomes, success or failure. Success can be any result, with all other possibilities considered as failure.

4. The probability of success on each attempt is a constant, represented by the lower case letter p. Then the probability of failure is the complement, typically represented by the letter q. So $q = 1 - p$.

5. Then the binomial random variable X is the number of successes in n attempts of the experiment.

Example: Which of the 5 situations would be considered binomial experiments?

1. A kid throws water balloons at a friend until they land a direct shot.
2. A teenager goes to a carnival with $20 and plays twenty $1 games at various booths, trying to win a prize.
3. A mother buys three raffle tickets for her sons track team fundraiser. There are three prizes to win.
4. A poker player is dealt five cards, hoping to get as many face cards as possible.
5. Someone rolls a die 10 times and counts how many even numbers they roll.

1. **Solution**: This is not binomial. The number of attempts is a variable, it could be any number.
2. **Solution**: This is not binomial. The probabilities are likely different from game booth to game booth.
3. **Solution**: This is basically binomial. The number of attempts is $n = 3$. The probability of winning for each of her 3 tickets is (essentially) the same, assuming there are a large number of raffle tickets sold. However, if the number of tickets is small, then the probabilities would change significantly and this would not be binomial.
4. **Solution**: This is not binomial. The probability of each card being a face card changes based on the conditions of the previous cards dealt. The attempts are not independent.
5. **Solution**: This is binomial. The number of attempts is $n = 10$. Each roll is independent, with p = P(even) = $\frac{1}{2}$.

Binomial probabilities follow a simple pattern and can be calculated by the formula $P(X = k) = {}_nC_k \ p^k q^{n-k}$, where X is the random variable count of successes, k is the specific number of successes for the situation, ${}_nC_k$ is the combination for k successes out of n attempts, p and q are the probabilities of success and failure for each attempt, and $n - k$ is the number of failures. The possible values of X are all of the whole numbers from 0 to n.

Example: For four flips of a fair coin, calculate the probabilities for the possible number of tails in the four flips.

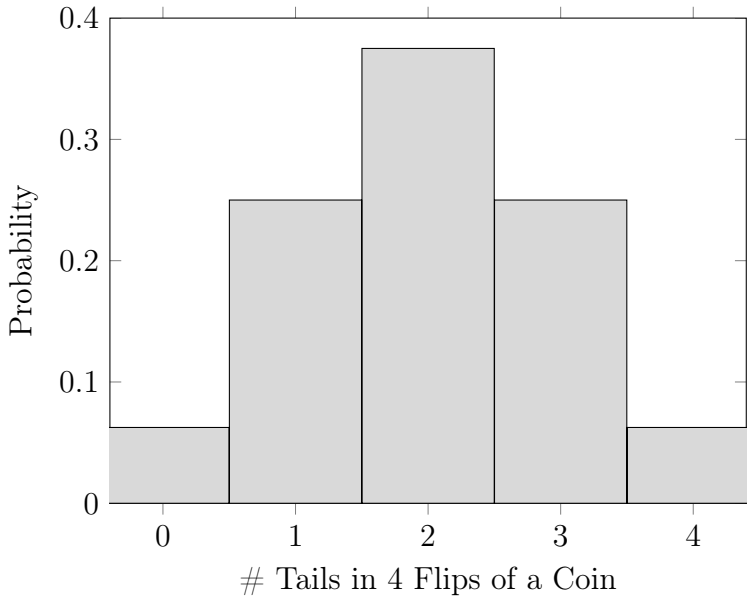
Solution: Here $n = 4$, $p = \frac{1}{2} = q$ also. The probability calculations are as follows.

$P(X = 0) = {}_4C_0 \ (\frac{1}{2})^0(\frac{1}{2})^4 = 1(1)(\frac{1}{16}) = \frac{1}{16}$ $\quad$ $P(X = 1) = {}_4C_1 \ (\frac{1}{2})^1(\frac{1}{2})^3 = 4(\frac{1}{2})(\frac{1}{8}) = \frac{4}{16}$

$P(X = 2) = {}_4C_2 \ (\frac{1}{2})^2(\frac{1}{2})^2 = 6(\frac{1}{4})(\frac{1}{4}) = \frac{6}{16}$ $\quad$ $P(X = 3) = {}_4C_3 \ (\frac{1}{2})^3(\frac{1}{2})^1 = 4(\frac{1}{8})(\frac{1}{2}) = \frac{4}{16}$

$P(X = 4) = {}_4C_4 \ (\frac{1}{2})^4(\frac{1}{2})^0 = 1(\frac{1}{16})(1) = \frac{1}{16}$

We can sketch a **Probability Histogram** which shows all possible outcomes with their probabilities as bars. Notice the symmetry of the probability values and the bars. This happens when p is close to $\frac{1}{2}$.



Example: In the general human population, 10% of people are naturally left-handed. Calculate the probabilities for the possible number of left-handed people in a random selection of 5 people.

Solution: Here $n = 5$, $p = 0.1$, $q = 0.9$. The probability calculations are as follows.

$P(X = 0) = {}_5C_0\ (0.1)^0(0.9)^5 = 1(1)(.590) = 0.590$

$P(X = 1) = {}_5C_1\ (0.1)^1(0.9)^4 = 5(.1)(.656) = 0.328$

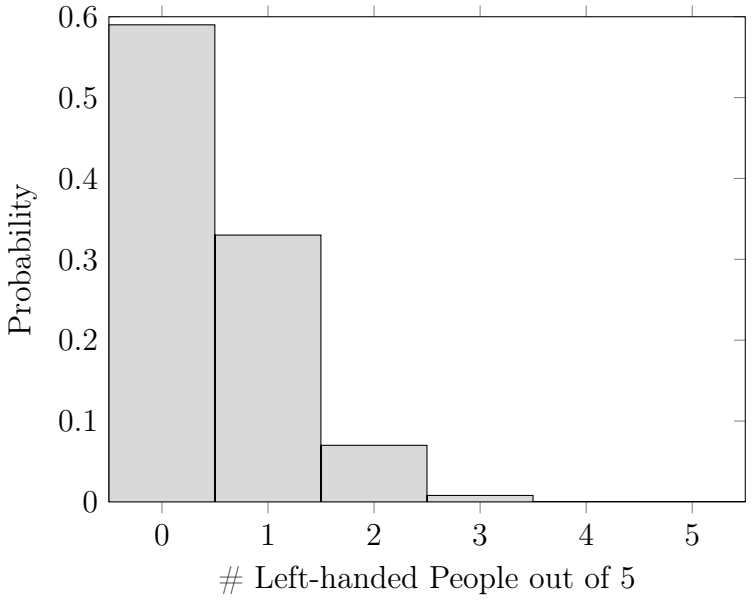
$P(X = 2) = {}_5C_2\ (0.1)^2(0.9)^3 = 10(.01)(.729) = 0.073$

$P(X = 3) = {}_5C_3\ (0.1)^3(0.9)^2 = 10(.001)(.81) = 0.008$

$P(X = 4) = {}_5C_4\ (0.1)^4(0.9)^1 = 5(.0001)(.9) = 0.0005$

$P(X = 5) = {}_5C_5\ (0.1)^5(0.9)^0 = 1(.00001)(1) = 0.00001$

The probability histogram is below. Notice the graph is right skewed (high probability on left) since p is small.



The mean and standard deviation of a binomial variable can be found using the discrete random variable formulas previously stated. However, since binomial probabilities follow a pattern, the formulas can be simplified to $\mu = np$ and $\sigma = \sqrt{npq}$.

Example: For four flips of a fair coin, calculate the mean and standard deviation of the number of tails in the four flips.

Solution: $\mu = np = 4 * (\frac{1}{2}) = 2$ and $\sigma = \sqrt{npq} = \sqrt{4(\frac{1}{2})(\frac{1}{2})} = 1$. This should make sense, we expect half of the flips to be tails and half heads on average.

Example: Calculate the mean and standard deviation of the number of left-handed people in a random selection of 5 people.

Solution: $\mu = np = 5 * (0.1) = 0.5$ and $\sigma = \sqrt{npq} = \sqrt{5(0.1)(0.9} = 0.67$. This should seem a bit odd, since this says we expect half of a person to be left-handed. Remember that this is an average. Most of time there would be one or none of the five people being left-handed, and only rarely more than that.

****Try this on your own**: According to a survey of US adults done by the Pew research center in 2021, 85% of adults own a smartphone. For a random sample of 50 adults, find the mean, standard deviation and the probability that 45 of the 50 adults own a smartphone.

3.1.4 Exercises: Discrete Distributions

Solutions appear at the end of this textbook.

1. The distribution below shows the percentage of people who have received Covid-19 vaccines in Georgia as of June 16, 2022. Find the missing probability and the mode.

Vax dose	None	1 dose	2 doses	3+ doses
prob - P(x)	?	7%	33%	24%

2. Based on a study by the National Student Clearinghouse Research Center in 2016, the table below shows the percent of students and the number of elapsed years it took them to earn a Bachelor's degree. Find the mode, median, and quartiles of the number of years. Treat the last group of 9+ as 9 years.

Years	4	5	6	7	8	9+
P(x)	0.38	0.26	0.12	0.07	0.03	0.14

3. Based on a study by The Demographic and Health Surveys (DHS) Program in 2017, the table below shows the percent of women who stated the age at which they first had sexual intercourse. Find the mean and standard deviation of the age. Treat the last group of 20+ as 20 years.

Age	10	11	12	13	14	15	16	17	18	19	20+
P(x)	0.01	0.01	0.02	0.04	0.12	0.21	0.15	0.11	0.11	0.04	0.18

4. A game consists of throwing a dart at a target. The target has 3 rings surrounding a bullseye. The prizes and probability for landing the dart in each area is shown below. Find the expected value of the prize.

Area	outer ring	middle ring	inner ring	bullseye
Prize	\$0	\$5	\$10	\$50
prob - P(x)	0.42	0.30	0.21	0.07

5. Find the expected value of the cash prize for a lottery game. How much should they charge to play, if they want to make some profit?

Prize	\$0	\$5	\$100	\$2,500	\$20,000	\$100,000
Probability	$\frac{752{,}944}{800{,}000}$	$\frac{45{,}000}{800{,}000}$	$\frac{2{,}000}{800{,}000}$	$\frac{50}{800{,}000}$	$\frac{5}{800{,}000}$	$\frac{1}{800{,}000}$

6. Calculate the binomial probabilities for the given values. Round to 4 places.

   a) $n = 40$, $p = \frac{1}{4}$, $x = 8$

   b) $n = 10$, $p = 0.75$, $x = 8$

   c) $n = 4$, $p = \frac{1}{2}$, $x = 1$

   d) $n = 16$, $p = 0.03$, $x = 0$

7. For five rolls of a standard 6-sided die, calculate the probabilities for the possible number of times a 5 shows on the roll.

8. Sketch the probability histogram for the previous exercise for the possible number of times a 5 shows on five rolls of a die.

9. According to the National Institute on Drug Abuse, in 2015, 17% of US men and 13% of US women smoke cigarettes. For a random sample of 100 men and 100 women, find the mean and standard deviation for each sex.

3.2 Continuous Distributions

3.2.1 The Uniform Distribution

In the beginning of the book, we learned that quantitative variables can be discrete or continuous. Probability distributions apply to discrete variables. There are only so many values and each one has a specific probability. Continuous variables must be dealt with differently. There is a continuous interval of infinitely many values or effectively infinite due to so many values in a small interval. This makes it almost impossible to calculate a probability for one particular value. Fortunately, there are formulas and tools which can give us information about probabilities of continuous variables.

Intervals of values for continuous variables have a probability weight (or density). Some intervals can carry more weight than others, and as an interval increases, the weight increases. When the values are graphed relative to the probability density, we can see patterns or lack of a pattern.

The simplest pattern is one where every value in a continuous interval has equal weight, although there are effectively an infinite number of possible values. When every value is equal, this is called the **Uniform Distribution**. For an interval that goes from $x = a$ to $x = b$, the probability weight (density) of all values is equal to 1 over the length of the interval shown as $\frac{1}{b-a}$.

The probability density graph is just a straight horizontal line above all the values in the interval. The shape is a simple rectangle with total area equal to 1 or 100%. See figure 3.1 on the next page.

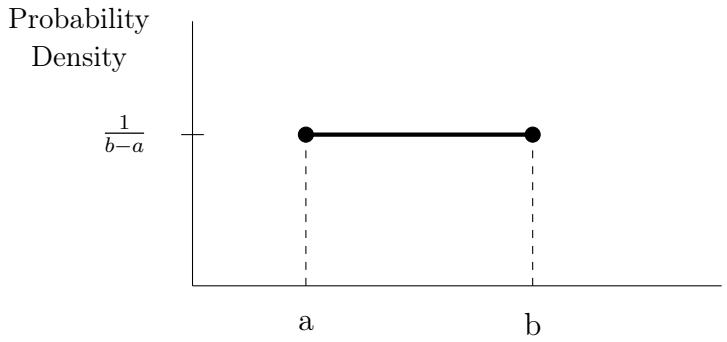
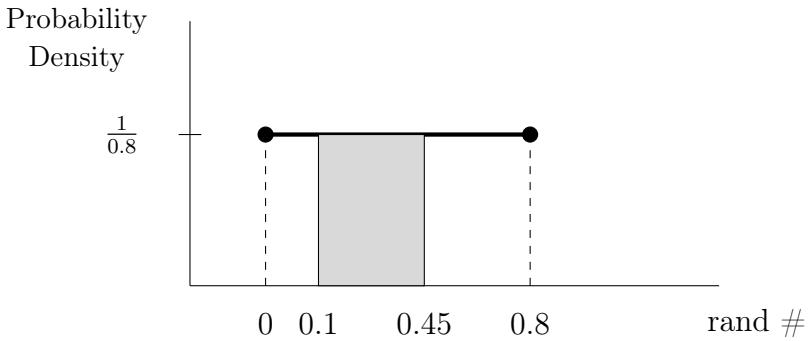


Figure 3.1: Uniform Probability Density over the interval $x = a$ to $x = b$

The base width of the rectangle shape is the distance $b - a$, the height is the density $\frac{1}{b-a}$. The area is base times height, area $= (b-a)\frac{1}{b-a} = 1$. The probability for any subinterval between two values c and d is the area of the partial rectangle. The height will always be $\frac{1}{b-a}$, but the width would be the length of the subinterval $d - c$.

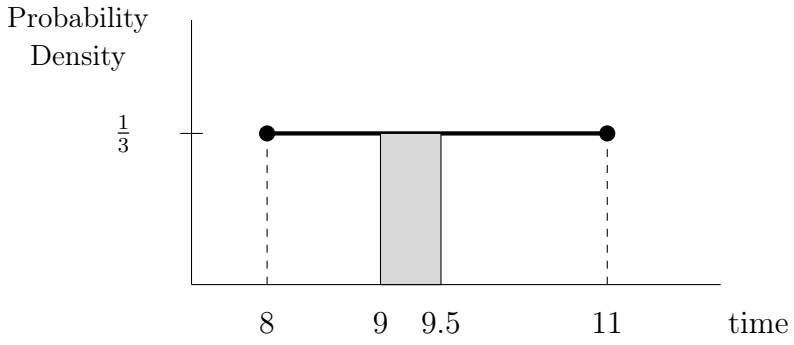
Example: A scientist is running a simulation in which they use a computer to pick random numbers out to six decimal places. The numbers can be between 0 and 0.8. Find the probability of picking a number between 0.1 and 0.45

Solution: Here the full interval is from $a = 0$ to $b = 0.8$, so the probability density is $\frac{1}{b-a} = \frac{1}{0.8} = 1.25$. The probability of picking a number between 0.1 and 0.45 is the area of that rectangle. $P(0.1 < x < 0.45) = (0.45 - 0.1) * 1.25 = 0.4375$. See figure below. The shaded area is 0.4375=43.75%, just under half of the entire interval.



Example: When there are no accidents or construction, a campus bus takes anywhere between 8 and 11 minutes to travel the full route around campus. Assume the time is uniformly distributed. Find the probability that the bus takes between 9 and $9\frac{1}{2}$ minutes to complete the full route.

Solution: Here the full interval is from $a = 8$ to $b = 11$, so the probability density is $\frac{1}{b-a} = \frac{1}{3}$. The probability of that the bus takes between 9 and $9\frac{1}{2}$ is the area of that rectangle. $P(9 < x < 9.5) = (9.5 - 9) * \frac{1}{3} = 0.167 = 16.7\%$. See figure below.



****Try this on your own**: A computer will pick random numbers out to ten decimal places. The numbers can be between 1 and 5.5. Find the probability of picking a number between 1.4 and 2.8

3.2.2 The Normal Distribution

If a continuous variable has a graph that is symmetric and bell shaped, it can usually be described by a pattern and a formula. One variable that has this bell pattern is called the **Normal Distribution**.

The formula itself is very complicated and requires some seriously advanced math. Lucky for us, there are tables and calculator functions which can do the math for us. A normal

distribution is completely determined by two parameters, the mean μ and the standard deviation σ.

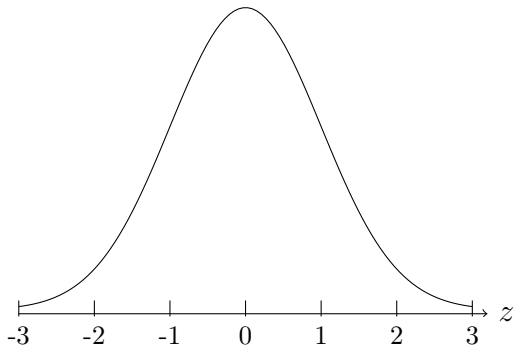
Just to show you how difficult it is to work with the formula directly, the normal distribution probability density is given by the formula:

$$N(x) = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$$

There are many real world phenomena that follow a normal distribution, leading to many different variations. However, if you look closely at the exponent of the numerator, you will see a familiar ratio $\frac{x-\mu}{\sigma}$, which is the standardized z-score of a variable x. This allows all of the different normal variables to be converted into one universal bell shape distribution called the **Standard Normal Distribution**.

The standard normal distribution has mean $\mu = 0$ and standard deviation $\sigma = 1$, which makes it very easy to work with. There is also a table that can be used to calculate values and probabilities. The table can be found at the end of this book. Many calculators have a function that can give values from the table as well.

What makes the standard normal distribution so nice to work with, is that probabilities directly correspond to area under the curve. The total area under the curve is equal to 1 or 100%. The standard normal distribution graph is shown below.



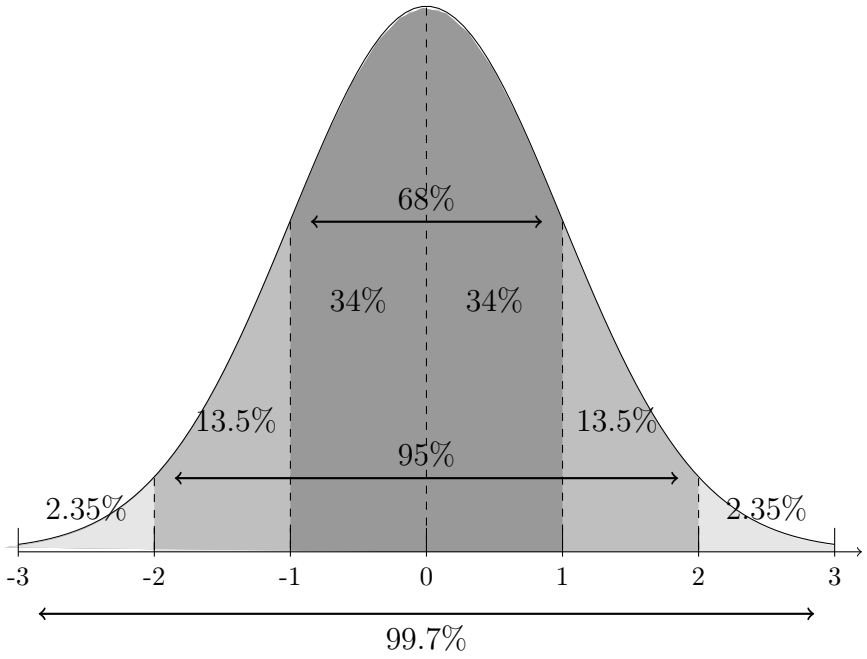
Notice that a large part of the graph in concentrated between $z = -1$ and $z = +1$, with most of the graph between $z = -2$ and $z = +2$ (usual values), and just about all of the graph between $z = -3$ and $z = +3$. The graph technically goes out forever, but it gets so low that effectively there is not much beyond ± 3.

There are two methods to calculating probabilities for areas under the bell curve. One is very quick and easy, but limited to intervals between the whole values of z-scores and only requires the use of a general rule. The other is more detailed but requires formulas, tables, and/or technology.

The quick and easy method is to use the **The Empirical Rule** which is stated below.

- 68% of the data will be located within one standard deviation to either side of the mean
- 95% of the data will be located within two standard deviations to either side of the mean
- 99.7% of the data will be located within three standard deviations to either side of the mean

The Empirical Rule is also known as the **68-95-99.7 Rule**. This scope of this rule can be best shown with the following diagram.



Since the bell curve is symmetric, each side is a mirror image of equal size. The percentage for one side, from 0 on up or up to 0, is exactly half of the graph, so 50%. Each of the intervals mentioned above between $\pm$ whole units, are split down the middle. That is how we know that the two sections in the middle (darkest) are 34% each, half of 68.

The next tier on either side is 13.5% each. This is found by taking difference between $95 - 68 = 27\%$ and dividing by two, since there is a section on either side. The two small slices on the ends (between 2 and 3) are 2.35% each. This is found by taking difference between $99.7 - 95 = 4.7\%$ and dividing by two.

To solve problems about area/probability of certain intervals, just add up or subtract the appropriate sections.

Example: What percent of the bell curve lies between $z = -1$ and $z = +2$?

Solution: Between -1 and 1 is 68%, with additional area of 13.5% between 1 and 2. This results in $68 + 13.5 = 81.5\%$

Example: What percent of the bell curve lies between $z = -2$ and $z = 0$?

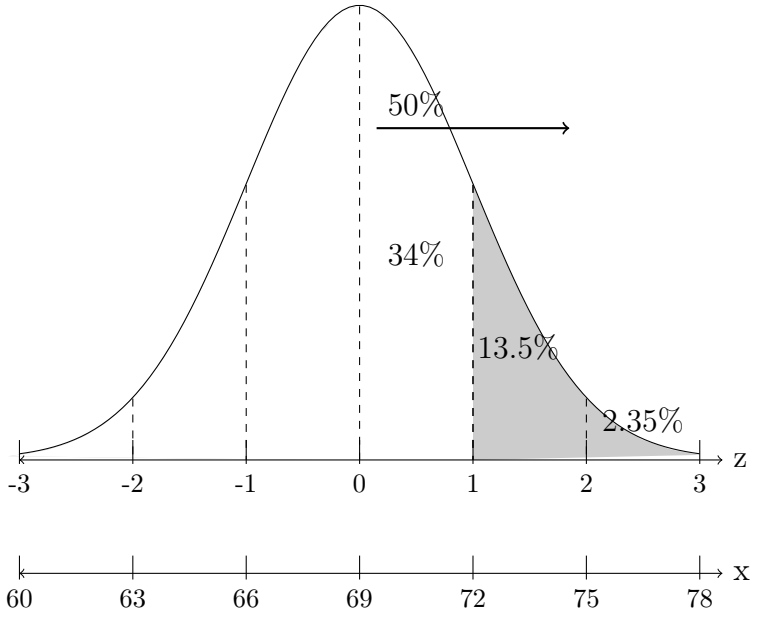
Solution: Between -2 and -1 is 13.5%, with additional area of 34% between -1 and 0. This results in $34 + 13.5 = 47.5\%$

This graph and the **68-95-99.7 Rule** can be applied to real world data that follow a bell curve. When working problems applied to specific data (such as heights, IQ scores, etc.), we can match the data values with z-scores on the graph. Since the z-scores are actually the number of standard deviations from the mean, then we can lineup the data mean below $z = 0$, then add/subtract the data standard deviation to put data values in line with the z-score units from -3 to 3.

Example: Adult male heights are normally distributed (follow bell curve), with a mean of $\mu = 69$ inches and a standard deviation of $\sigma = 3$ inches. What percent of men are taller than 6 feet?

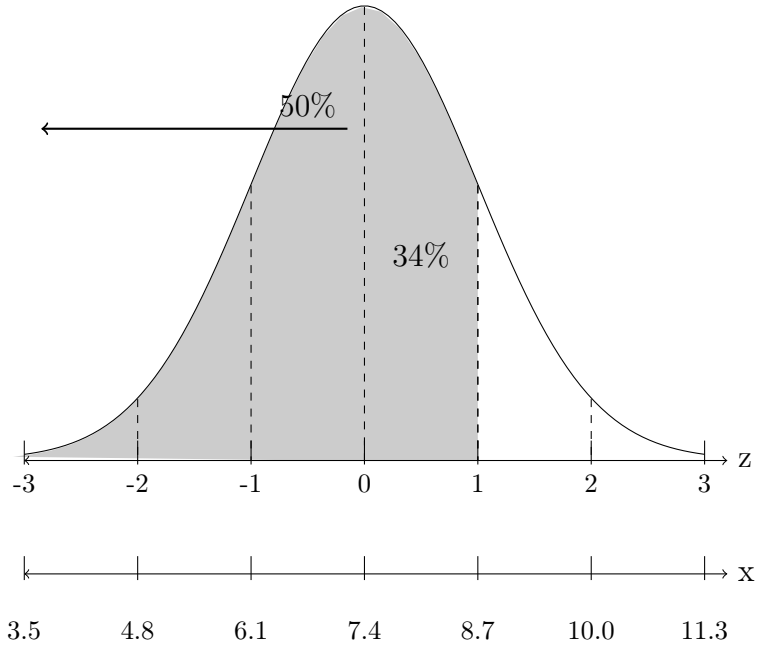
Solution: Draw a bell curve with standard z-axis from -3 to 3 and below that, an x-axis with heights that correspond to the z marks. The mean height of 69 will go below $z = 0$, one standard deviation higher ($69+3 = 72$ inches) will go below $z = 1$, two deviations higher, 75, goes below $z = 2$, and 78 below $z = 3$. Do similar process on left side, subtracting standard deviation to go under the negative z-values.

Change 6 feet into 72 inches. So we are looking for the slice of the graph above 72, which is above $z = 1$. The area is the upper half (50%) minus the section from $z = 0$ to 1 of 34%. The answer is $50 - 34 = 16\%$. Notice that this is also the same as adding the pieces to the right $13.5 + 2.35 = 15.85\%$, where the extra bit to make the full 16% is the tiny slice after $z = 3$. The graph is shown below.



Example: The birth weights of babies in the USA are normally distributed, with a mean of $\mu = 7.4$ pounds and a standard deviation of $\sigma = 1.3$ pounds. What percentage of babies are born with a weight less than 8.7 pounds?

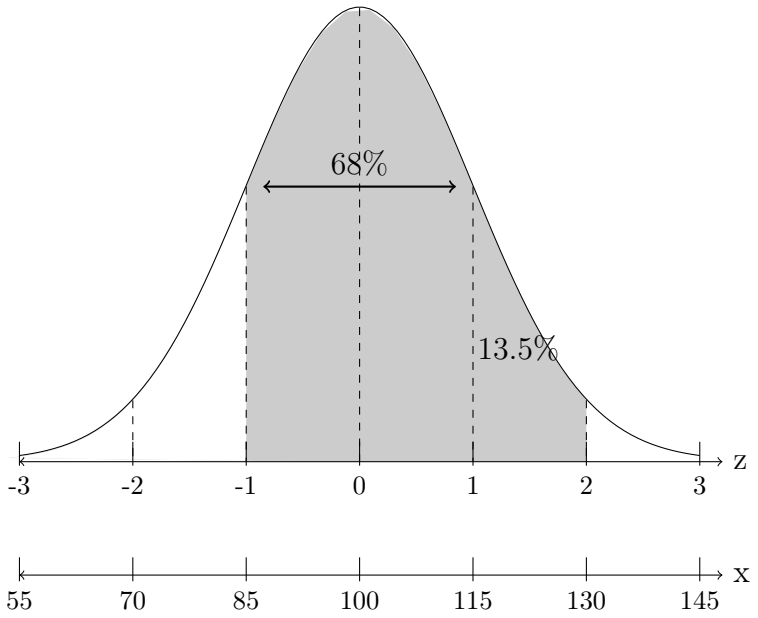
Solution: Draw a bell curve with standard z-axis from -3 to 3 and below that, an x-axis with weights that correspond to the z marks. The mean weight of 7.4 will go below $z = 0$, one standard deviation higher (8.7 pounds) will go below $z = 1$, etc. So we are looking for the slice of the graph below 8.7, which is above $z = 1$. The area is the lower half (50%) plus the section from $z = 0$ to 1 of 34%. The answer is $50 + 34 = 84\%$. The graph is shown below.



Example: IQ test scores are normally distributed, with a mean of $\mu = 100$ and a standard deviation of $\sigma = 15$. What percent of the population have IQ scores between 85 and 130?

Solution: Draw a bell curve with standard z-axis from -3 to 3 and below that, an x-axis with IQ scores that correspond to the z marks. The mean of 100 will go below $z = 0$, one standard deviation higher ($100 + 15 = 115$ inches) will go below $z = 1$, etc. Do similar process on left side, subtracting standard deviation to go under the negative z-values.

We are looking for the slice of the graph from 85 to 130, which are z-scores from -1 to $+2$. The area is the middle (68%) plus the section from $z = 1$ to 2 of 13.5%. The answer is $68 + 13.5 = 81.5\%$. The graph is shown below.



****Try this on your own** The birth weights of babies in South America are normally distributed, with a mean of $\mu = 3,100$ grams and a standard deviation of $\sigma = 400$ grams. Use the 68-95-99.7% rule to find the percentage of babies born with a weight more than $2,300$ grams.

Recall from the first chapter, that a z-score is the number of standard deviations that a given value is above or below its mean. Whenever a value is below the mean, its corresponding z-score will be negative. Usual values are z-scores from -2 to $+2$. Unusual values are z-scores outside this range. Z-scores have no units. The formulas is $z = \frac{x - \mu}{\sigma}$. To match the table, z-scores are rounded to two decimal places.

When working problems applied to specific data (such as heights, IQ scores, etc.), we can convert that data into z-scores and use the standard normal distribution and table to calculate probabilities. Before we work with applied data, we need practice the procedures for the standard normal values.

The first type of problem is finding the probability corresponding to a range of z-scores. The probability is equal to the corresponding area under the bell curve, which is above the range of z-scores. The steps are listed below.

1. Draw a standard normal curve like the one shown above.
2. At the z-scores mentioned in the problem, draw vertical lines that slice the graph into sections.
3. Shade the section above the corresponding range of z-scores mentioned in the problem.
4. Look on the Standard Normal Z-table (at end of book) for the areas below (to left) of the z-scores.

On the table, the first two digits (whole and tenths) of the z-scores are listed along the left side row headings, and the hundredths are listed across the top column headings. Find the row/column that matches the z-score from the problem. Then look where that row/column meet in the body of the table to find the area (probability) to the left (below) the z-score. The areas are shown to four decimal places (the thousandths).

For example, the area to left of $z = -2.36$ is found on the first page of the table (negative z-scores), scrolling down along the left to the row for -2.3 and across under column for 0.06. The area value in this location is 0.0091. This means that the probability of a z-score being less than -2.36 is $0.0091 = 0.91\%$.

If you are looking for the probability below a z-score (to left), the answer is simply the area from table, $P(Z < \#) = area$.

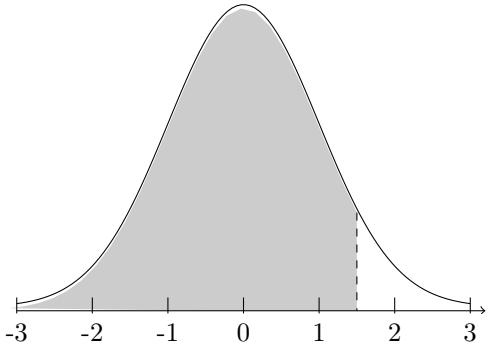
If you are looking for the probability above a z-score (to right), then use complement rule, $P(Z > \#) = 1 - area$. This is the area to the right side.

If you are looking for the probability between two z-scores, then look up both table areas and subtract. $P(\#_1 < Z < \#_2) = area2 - area1$. This is the area between.

Example: Find the probability of a standard normal z-score being less than 1.5. In symbols, this is $P(Z < 1.5)$.

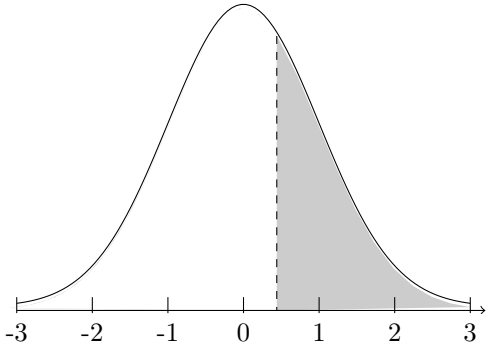
Solution: Draw a bell curve, make a line to slice the graph at $z = 1.5$, shade below (to left), then go to the table at end of the book. Look on the second page (positive z-scores) and go down to the row for 1.5 and across to the first column 0.00 (since the z-score is really 1.50). There we find the area of 0.9332, which is our answer, $P(Z < 1.5) = 0.9332 = 93.32\%$.

The graph is shown below. Notice that the probability is large (93.32%) and the shaded area is very large, so this makes sense.



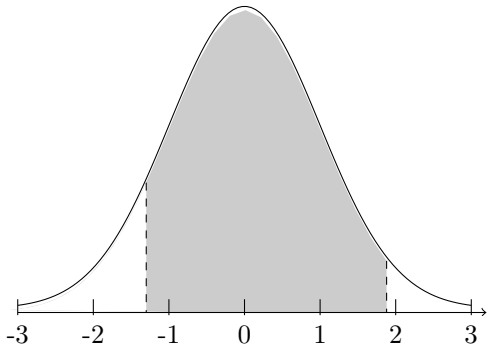
Example: Find the probability of a standard normal z-score being greater than 0.44. In symbols, this is $P(Z > 0.44)$.

Solution: Draw a bell curve, make a line to slice the graph at $z = 0.44$, shade above (to right), then go to the table at end of the book. Look on the second page (positive z-scores) and go down to the row for 0.4 and across to the fifth column 0.04. There we find the area of 0.6700, which is the area to the left, but we want the area to the right. Therefore, $P(Z > 0.44) = 1 - P(Z < 0.44) = 1 - 0.6700 = 0.3300 = 33.00\%$. The graph is shown below. Notice that the probability is somewhat small (33%) and the shaded area is somewhat small, so this makes sense.



Example: Find the probability of a standard normal z-score being between -1.3 and $+1.88$. In symbols, this is $P(-1.3 < Z < 1.88)$.

Solution: Draw a bell curve, make lines to slice the graph at $z = -1.3$ and $z = 1.88$, shade above that range, then go to the table at end of the book. Look on the first page (negative z-scores) and go down to the row for -1.3 and across to the first column 0.00 (since the z-score is really -1.30). There we find the area of 0.0968, which is the area to the left of $z = -1.3$. Look on the second page (positive z-scores) and go down to the row for 1.8 and across to the column under 0.08. There we find the area of 0.9699, which is the area to the left of $z = 1.88$. Now we want the area between these, so we subtract these areas. Therefore, $P(-1.3 < Z < 1.88) = P(Z < 1.88) - P(Z < -1.30) = 0.9699 - 0.0968 = 0.8731 = 87.31\%$. The graph is shown below. Notice that the size of the probability matches the size of the shaded area, so this makes sense.



****Try this on your own**: Compute $P(-0.5 < Z < 1.35)$

The second type of problem is finding the z-scores which are the cutoffs for a particular area (probability). The steps are listed below. The steps are somewhat reversed from the previous type, and there are a few issues to be careful of.

1. Draw a standard normal curve (without the numbers along the axis).

2. Draw vertical lines that slice the graph into sections, to give a rough approximation of where the z-score cutoffs would have to be to match the given area or probability.

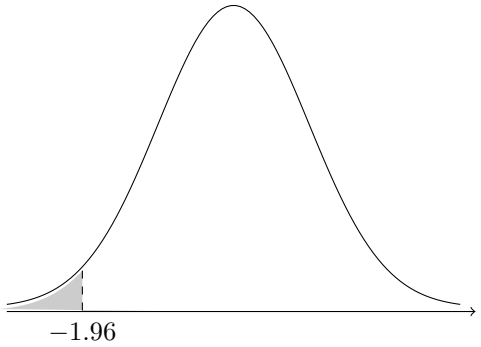
3. Shade the appropriate section of the graph (to left, to right, or between).

4. Figure out the size of other areas on the graph, so that you get a value for an area that is from a right cutoff and goes down all the way to the left end. We must do this, since the table only shows areas of this form and we need to match them.

5. Look on the Standard Normal Z-table for the left side areas we figured out in the previous step. The areas are inside the body of the table (4 decimal place numbers). Once we locate the appropriate area, then look to the edges to find the digits of the corresponding z-score. One issue that might arise, is that the table does not have every possible area value. If the area (probability) we are looking for is not there, then look for the closest area you can find, and use its z-score. If the area is exactly in the middle of two areas, then use both and average their z-scores.

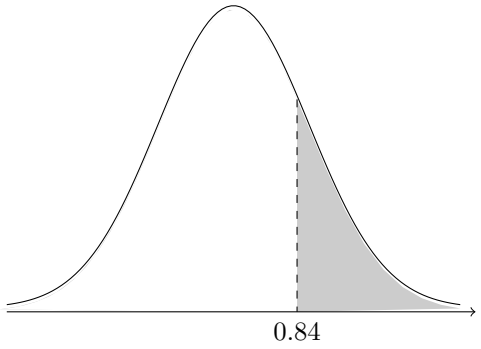
Example: Find the standard normal z-score such that the probability of finding a z-score less than it, is 2.5% . In symbols, we want the z, such that $P(Z < z) = 0.025$.

Solution: Draw a bell curve, draw a line to slice the graph into a small slice way over to the left, shade below (to left), then go to the table at end of the book. Look on the first page, negative z-scores, since the cutoff is over on the negative side of the graph. Look in the body of the table for 0.0250. Notice it is in the row for -1.9 and below the column 0.06. Therefore, the z-score is -1.96. The graph is shown on the next page.



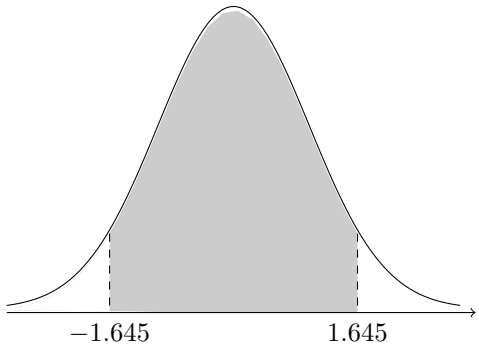
Example: Find the standard normal z-score such that the area to its right is 20%. In symbols, we want the z, such that $P(Z > z) = 0.2$.

Solution: Draw a bell curve, draw a line to slice the graph into a somewhat small slice over to the right, shade above (to right), then go to the table at end of the book. Since the table shows areas to the left (below), and our given area is to the right (above), we need to use the complement rule to convert, in order to match the table. The area to the left is $1 - 0.2 = 0.8$ (or 80%). Look on the second page, positive z-scores, since the cutoff is over on the positive side of the graph. Look in the body of the table for 0.8000. This exact value does not appear in the table, but the closest value is 0.7995. It is in the row for 0.8 and below the column 0.04. Therefore, the z-score is 0.84. The graph is shown below.



Example: Find the z-scores that cutoff the middle 90% of the graph. In symbols, we want z_1 and z_2, such that $P(z_1 < Z < z_2) = 0.9$.

Solution: Draw a bell curve, draw a line to slice the graph into mirror image small slices, one over to the right and one over to left, shade between them, then go to the table at end of the book. If 90% is between, then the other 10% is on the edges, with 5% in each tail. Then the area to the left of the low edge is 0.05. Look on the first page, negative z-scores. Look in the body of the table for 0.0500. This exact value does not appear in the table, but is exactly between 0.0505 and 0.0495. So we get both corresponding z-scores and average them. The row is -1.6 and the columns are 0.04 and 0.05. Therefore, the z-score is the average of -1.64 and -1.65. There fore $z = \frac{-1.64+-1.65}{2} = -1.645$. The graph is shown below.

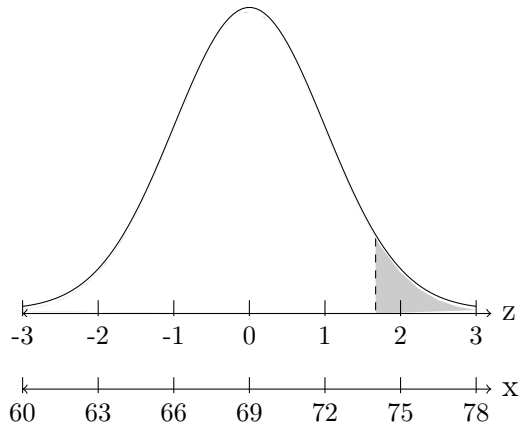


When working with actual data, such as heights or test scores, we will still use the z-table procedures, but with two important steps to add. We will use the z-score formula to convert between the data values x and the standard normal values z, as well as draw another axis below the graph to show how the x data values line up with the z-scores.

Since the z-scores are actually the number of standard deviations from the mean, then we can lineup the data mean below $z = 0$, then add/subtract the data standard deviation to put data values in line with the z-score units from -3 to 3.

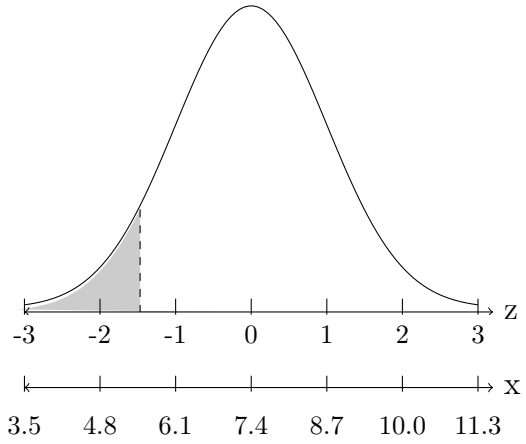
Example: Adult male heights are normally distributed (follow bell curve), with a mean of $\mu = 69$ inches and a standard deviation of $\sigma = 3$ inches. What percent of men are taller than 6 feet 2 inches?

Solution: Draw a bell curve with standard z-axis from -3 to 3 and below that, an x-axis with heights that correspond to the z marks. The mean height of 69 will go below $z = 0$, one standard deviation higher (72 inches) will go below $z = 1$, two deviations higher (75) goes below $z = 2$, and 78 below $z = 3$. Do similar process on left side, subtracting standard deviation to go under the negative z-values. Change 6 feet 2 inches into 74 inches. Then convert 74 into a z-score, $z = \frac{74-69}{3} = 1.67$. Make a line to slice the graph at about $z = 1.67$, shade above (to right), then go to the table at end of the book. Look on the second page (positive z-scores) and go down to the row for 1.6 and across to the column 0.07. There we find the area of 0.9525, which is the area to the left, but we want the area to the right. Therefore, $P(X > 74) = P(Z > 1.67) = 1 - P(Z < 1.67) = 1 - 0.9525 = 0.0475 = 4.75\%$. Just less than 5% of men are taller than 6 foot 2 inches. The graph is shown below.



Example: The birth weights of babies in the USA are normally distributed, with a mean of $\mu = 7.4$ pounds and a standard deviation of $\sigma = 1.3$ pounds. Find the probability of a baby being born with a weight less than 5.5 pounds for a single full-term birth, which is considered to be an unhealthy birth weight.

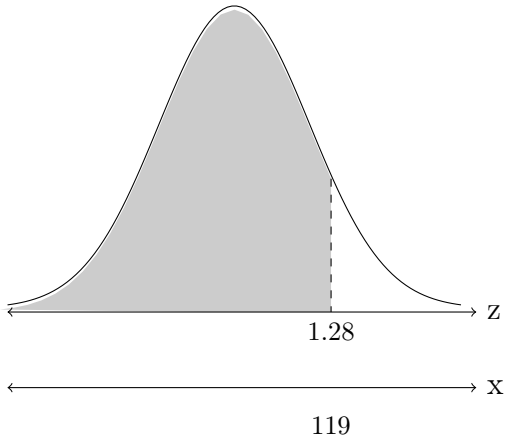
Solution: Draw a bell curve with standard z-axis from -3 to 3 and below that, an x-axis with weights that correspond to the z marks. The mean weight of 7.4 will go below $z = 0$, one standard deviation higher (8.7 pounds) will go below $z = 1$, etc. Convert 5.5 into a z-score, $z = \frac{5.5-7.4}{1.3} = -1.46$. Make a line to slice the graph at about $z = -1.46$, shade below (to left), then go to the table at end of the book. Look on the first page (negative z-scores) and go down to the row for -1.4 and across to the column 0.06. There we find the area of 0.0721, which is the area we need. Therefore, $P(X < 5.5) = P(Z < -1.46) = 0.0721 = 7.21\%$. So 7.21% of babies in the U.S. are born too small. The graph is shown on the next page.



Example: IQ test scores are normally distributed, with a mean of $\mu = 100$ and a standard deviation of $\sigma = 15$. What IQ score would be at the 90th percentile?

Solution: Draw a bell curve, draw a line to slice the graph into a small slice way over to the right, shade below (to left), then go to the table at end of the book. Look on the positive z-score page, since the cutoff is over on the positive side of the graph. Look in the body of the table for 0.9000. This exact value does not appear in the table, but the closest

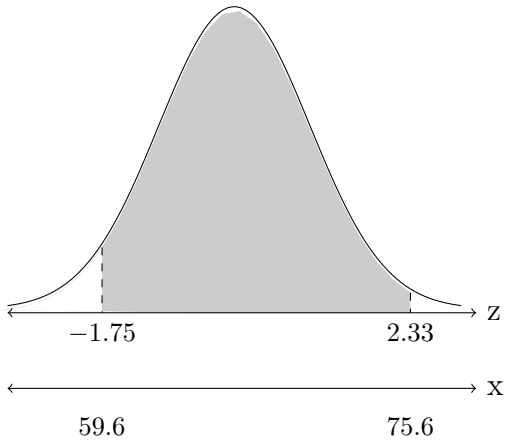
value is 0.8997. It is in the row for 1.2 and below the column 0.08. Therefore, the z-score is 1.28. Using the formula $z = \frac{x-\mu}{\sigma}$, we can solve for the unknown IQ x. Formula setup is $1.28 = \frac{x-100}{15}$. After multiplying both sides by 15 and adding 100, we get $x = 119.2$. Rounded to whole number, the 90th percentile 119. This means that 90% of all people have an IQ lower than 119. The graph is shown below.



Example: A company wants to design a new specialty mountain bike. Based on research and limitations of technology, they will make a bike that is suitable for people of almost every height, except for the shortest 4% of women and the tallest 1% of men. Both groups have heights that are normally distributed. Women with a mean of $\mu = 64$ inches and a standard deviation of $\sigma = 2.5$ inches. Men with a mean of $\mu = 69$ inches and a standard deviation of $\sigma = 3$ inches. What are the cutoff heights for those percentages?

Solution: Draw a bell curve, draw a line to slice the graph into small slices way over to the right and left and shade those end tails, then go to the table at end of the book. Look on the negative z-score page in the body of the table for area of 0.0400. This exact value does not appear in the table, but the closest value is 0.0401. The z-score we will use is -1.75.

For the men, 1% above the cutoff is same as 99% below. Go to positive z-score page, look in the body of the table for area of 0.9900. This exact value does not appear in the table, but the closest value is 0.9901. The z-score we will use is 2.33. Using the z-score formula for each, we get cutoff of $x = 59.6$ inches for women and 75.6 inches for men. Therefore the bike will be too big for women under 4 feet 11.6 inches and too small for men above 6 feet 3.6 inches. The combined graph is shown below.



****Try this on your own** The birth weights of babies in Brazil are normally distributed, with a mean of $\mu = 3,110$ grams and a standard deviation of $\sigma = 463$ grams. Find the probability of a baby being born with a weight more than $3,000$ grams.

3.2.3 Exercises: Continuous Distributions

Solutions appear at the end of this textbook.

1. If a variable X is uniformly distributed over the values 10 to 36, find $P(12 < X < 19.5)$. Round to nearest 0.1%.

2. The time of day at which a baby is born, is uniformly distributed over the 24 hours on the clock. What is the probability that a baby is born between 4 a.m. and 7 a.m.? Show answer as reduced fraction.

3. Using the 68-95-99.7 rule, compute the area of a bell curve between $z = -3$ and $z = -1$.

4. SAT test scores for English are normally distributed, with a mean of $\mu = 500$ and a standard deviation of $\sigma = 100$. Sketch a bell curve and match the whole z-scores on the bell curve with SAT test scores.

5. Adult female heights are normally distributed, with a mean of $\mu = 64$ inches and a standard deviation of $\sigma = 2.5$ inches. Using the Empirical Rule, what percent of women are shorter than 5 feet 1-1/2 inches?

6. Find the probabilities below of the standard normal z-score.

   a) $P(Z < 2.73)$

   b) $P(Z < -1.04)$.

   c) $P(Z > -0.45)$.

   d) $P(Z > 1.20)$.

   e) $P(-1.40 < Z < 1.40)$.

   f) $P(0.5 < Z < 1.85)$.

7. Find the standard normal z-score with the areas described below.

   a) area to its left is 0.0129

   b) area to its left is 0.6950

   c) area to its right is 0.0700

   d) area to its right is 0.5910

   e) area between $-z$ and z is 0.8740

8. Find the z-scores that cutoff outer edge tails of 10% on either side of the bell curve.

9. Explain what causes $P(Z > -n) = P(Z < n)$, for all values of n.

10. Adult female heights are normally distributed, with a mean of $\mu = 64$ inches and a standard deviation of $\sigma = 2.5$ inches. What percent of women are shorter than 5 feet 6 inches?

11. The birth weights of hospital born babies in Pakistan are normally distributed, with a mean of $\mu = 2.9$ kg and a standard deviation of $\sigma = 0.5$ kg. Find the probability of a baby being born with a weight greater than 3.5 kg.

12. IQ scores are normally distributed with a mean of 100 and standard deviation of 15. What percent of the population has IQ scores between 90 and 120?

13. SAT test scores are normally distributed, with a mean of $\mu = 500$ and a standard deviation of $\sigma = 100$. What score would be at the 75th percentile?

14. Tinytown College offers three different scholarships. The silver scholarships go to applicants who score above the 85th percentile on the math portion of the SAT, but still within the 95th percentile (above that qualifies for the gold scholarship). What scores will qualify applicants for the silver scholarship?

3.3 Sampling Distributions

3.3.1 Sampling Distributions

If we are looking at a data observation that comes from a normal distribution, then we can use the Empirical Rule or the Normal z-table or calculator functions. What happens if the data is from a different distribution? Under certain conditions, when we gather data from a sample and look at the average (mean) or percentage (proportion) from the sample, those measures are variables that will follow a bell curve or very close to bell curve. This allows us to then use the normal distribution anyway.

To understand how the sample mean can be a variable, let's look at a small set of data as a population. An upper level college class could have just six students and their grades on the final exam could be $62, 72, 78, 89, 91, 91$. This data is skewed left. The mean $\mu = 80.5$ and standard deviation $\sigma = 10.9$.

If we randomly select two grades, there are only so many different combinations. There are exactly 15 different sets of two grades that can be picked. Each pair of grades is a sample of size $n = 2$. Each of the 15 possible samples are shown below with their corresponding sample means (average of the two grades). This is known as a **Sampling Distribution**, which shows all possible samples and one or more of their measures.

Grades	$62, 72$	$62, 78$	$62, 89$	$62, 91$	$62, 91$	$72, 78$	$72, 89$	$72, 91$
mean $\overline{x}$	67	70	75.5	76.5	76.5	75	80.5	81.5
Grades	$72, 91$	$78, 89$	$78, 91$	$78, 91$	$89, 91$	$89, 91$	$91, 91$	
mean $\overline{x}$	81.5	83.5	84.5	84.5	90	90	91	

Notice that as the combination of grades changes, the sample mean changes, so it is a variable. If we take the mean and standard deviation of all 15 values of $\overline{x}$ we get $\mu_{\overline{x}} = 80.5$ and $\sigma_{\overline{x}} = 6.9$. The mean of $\overline{x}$ is the same as the population mean of x itself, but the standard deviation has decreased significantly. The 15 values of $\overline{x}$ are also spread out more evenly, it is not skewed as much as the original set of data.

Now let's look at samples of size $n = 5$ grades. There are only six possible samples of size $n = 5$, which are show in the table below.

Grades	mean $\overline{x}$
$62, 72, 78, 89, 91$	78.4
$62, 72, 78, 89, 91$	78.4
$62, 72, 78, 91, 91$	78.8
$62, 72, 89, 91, 91$	81
$62, 78, 89, 91, 91$	82.2
$72, 78, 89, 91, 91$	84.2

The mean of $\overline{x}$ is again $\mu_{\overline{x}} = 80.5$, the same as the population mean of x itself, but the standard deviation has again decreased significantly to $\sigma_{\overline{x}} = 2.2$. The six values of $\overline{x}$ are also not spread out that much and very symmetric.

3.3.2 The Central Limit Theorem

So what have we learned from all of this, that even if a population does not follow a normal distribution individually, the sample mean will become closer and closer to following a normal distribution, but with a smaller standard deviation. This idea is stated more formally in the following theorem.

(3.3.1) Theorem. *The **Central Limit Theorem**: For a population with mean μ and standard deviation σ, the sampling distribution of the sample means $\overline{x}$ of size n becomes ap-*

proximately normally distributed as the sample size n gets large, no matter what distribution the population has.

If the population is already normal to start with, then the sample mean will follow normal distribution as well. If the population is not normal, then the accepted sample size is at least 30 to get a good normal distribution for the sample mean $\overline{x}$. The normal distribution for $\overline{x}$ will have the same mean as the population $\mu_{\overline{x}} = \mu$, but the standard deviation will decrease to $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$.

Example: The birth weights of babies in the USA are normally distributed, with a mean of $\mu = 7.4$ pounds and a standard deviation of $\sigma = 1.3$ pounds. What is known about the distribution of the sample mean of a random group of ten babies?

Solution: Since individual birth weights are already normally distributed, the sample size does not matter. The sample mean will also follow a normal distribution with $\mu_{\overline{x}} = 7.4$ and $\sigma_{\overline{x}} = \dfrac{1.3}{\sqrt{10}} = 0.41$.

Example: The number of homicides in U.S. States during 2018 was NOT normally distributed, but the mean was $\mu = 324$ murders per state and a standard deviation of $\sigma = 438$ murders. What is known about the distribution of the sample mean of a random group of 32 states?

Solution: X is NOT normally distributed, but the sample size is greater than 30. The sample mean will now follow an approximately normal distribution with $\mu_{\overline{x}} = 324$ and $\sigma_{\overline{x}} = \dfrac{438}{\sqrt{32}} = 77$.

Example: The time spent studying for a test by students at a college is skewed with mean of $\mu = 3$ hours and a standard deviation of $\sigma = 2$ hours. What is known about the distribution of the sample mean of a random group of ten students?

Solution: Since individual student time is NOT normally distributed, the sample size DOES matter. Since the sample size is only 10, the new distribution is unknown.

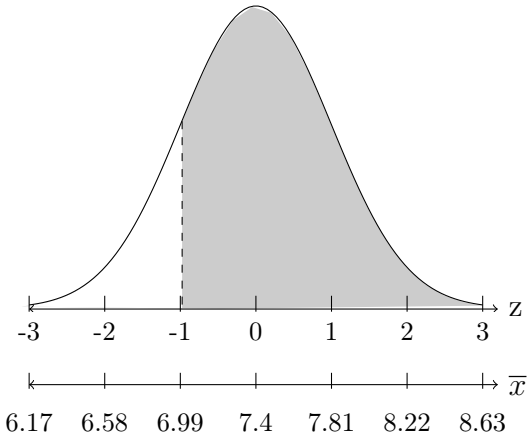
For the first two examples, we can use the normal distribution to answer questions about probability of a sample mean being within intervals, using the same techniques as the previous section. The only difference is to use the adjusted standard deviation $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$. The normal distribution cannot be used with the last example for time spent studying.

Now that you are hopefully getting the hang of the normal probability problems using the z-table, we can switch to using the calculator functions to get answers much quicker. However, we must make sure we understand the problem and set it up correctly. The calculator function that gives the area (probability) of a normal variable over an interval of data values is **normalcdf** found under distribution menu on TI calculators. The four inputs are the mean, standard deviation, the low end of the interval and the high end of the interval. The order of the inputs depends upon the model of the calculator. The data does NOT have to be converted into z-scores, as long as we use all values in the same units of the problem.

Example: The birth weights of babies in the USA are normally distributed, with a mean of $\mu = 7.4$ pounds and a standard deviation of $\sigma = 1.3$ pounds. What is probability that a sample mean of a random group of ten babies, is greater than 7 pounds?

Solution: Here the mean $\mu_{\overline{x}} = 7.4$, new standard deviation $\sigma_{\overline{x}} = \dfrac{1.3}{\sqrt{10}} = 0.41$, low end of interval starts at 7 pounds, the high end goes forever. Since we need a value for the high end, just use any relatively large value. A value of 99999 usually works for most problems. On the TI-83/84 calculators the function would be

$normalcdf(7, 99999, 7.4, 0.41) = 0.8354 = 83.54\%$. This can be interpreted as there is a good chance, 83.54%, that a sample of ten babies will have an average weight greater than 7 pounds. See graph of this situation below.



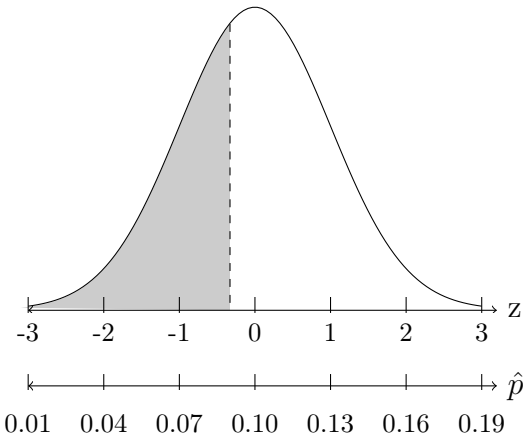
****Try this on your own**: The birth weights of babies in Brazil are normally distributed, with a mean of $\mu = 3,110$ grams and a standard deviation of $\sigma = 463$ grams. Find the probability of a sample of 25 babies being born with an average weight between $3,200$ and $3,400$ grams.

Sometimes the data being examined is the proportion (or percent) of a set of data that meets some condition. For example, what proportion of people would vote for a candidate in an election, or what percent of products are defected in manufacturing.

In these cases we are dealing with some population proportion p, a random sample of size n with a sample proportion $\hat{p}$ (known as "p-hat"). The sampling distribution for $\hat{p}$ will be approximately normal with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$ as long as the sample size n is relatively large. In most problems, we will assume it is large enough.

Example: In 2018 about 10% of babies in the US are born premature (less than 37 weeks). What is the probability that less than 9% of the babies in a random group of 100 babies are born premature?

Solution: Here the mean $\mu = 0.10$ and standard deviation $\sigma = \sqrt{\frac{.10(1-.10)}{100}} = 0.03$, low end of interval starts at negative infinity, the high end stops at 0.09. Since we need a value for the low end, just use any relatively large negative value. A value of -99999 usually works for most problems. Obviously we can't have negative percent, but we still need to use the low values down very far. On the TI-83/84 calculators the function would be $normalcdf(-99999, 0.09, 0.10, 0.03) = 0.3694 = 36.94\%$. This can be interpreted as there is reasonable chance, 36.94%, that less than 9% out of a sample of 100 babies will be born premature. See graph of this situation below.



3.3.3 Exercises: Sampling Distributions

Solutions appear at the end of this textbook.

1. For the data $4, 1, 2, 4, 6$ find the mean μ of the entire data set, then show the sampling distribution of samples of size three and the sample mean $\overline{x}$ for each sample.

2. Find the overall mean of the sample means from the previous exercise. Compare that to the population mean μ.

3. State and explain the Central Limit Theorem.

4. For each situation, find the mean $\mu_{\overline{x}}$ and standard deviation $\sigma_{\overline{x}}$ for the sampling distribution of the sample mean $\overline{x}$. Then state whether the sampling distribution will be normal, approximately normal, or unknown.

   a) SAT math scores are normally distributed, with mean $\mu = 500$ and standard deviation $\sigma = 100$. Samples are taken of size 12 students.

   b) Staring salary for engineers is not normally distributed. The mean $\mu = \$68,000$ and $\sigma = \$3,100$. Samples are taken of size 10 engineers.

   c) Variable X follows an unknown distribution, with mean $\mu = 80.5$ and standard deviation $\sigma = 11$. Samples are taken of size 55.

5. The birth weights of hospital born babies in Pakistan are normally distributed, with a mean of $\mu = 2.9$ kg and a standard deviation of $\sigma = 0.5$ kg. Find the probability of a sample of 36 babies having an average weight less than 2.8 kg.

6. In 2016 about 82% of households in the US have internet access. What is the probability that more than 80% of the households in a random group of 200 have internet?

Chapter 4

Inference: From Sample to Population

4.1 Confidence Intervals

4.1.1 Estimating the Population Mean

In the real world, the mean, standard deviation, or proportion from a population are rarely known. We can take a sample and use it to estimate the population value using the concepts from this course. We use the fact that when a sampling distribution is normal (or approximately), then most of the time the sample value will be within a few standard deviations or the population value. That also means that the population value will be within a few standard deviations of the sample value.

So we start with the sample value and build an interval (range) of values that will most likely contain the population value. The interval will be a reasonable estimate of the real value. This is called a **Confidence Interval**.

The confidence interval to estimate an unknown population mean is given by the formula $\overline{x} \pm E$ where E is the **margin of error** and $\overline{x}$ is the sample mean from a sample of some size n. The confidence interval to estimate an unknown population proportion (percentage)

is given by the formula $\hat{p} \pm E$, where $\hat{p}$ is the sample proportion from a sample of some size n.

As the sample size n gets larger, the confidence interval will get smaller (become more narrow range of values). This should make sense, that more data gives a better estimate. Sometimes we can be given an assumed value for the margin or error. To calculate a margin error directly can be challenging depending upon what we know from the population and/or the sample data.

Typically we want to have a particular level of confidence (probability) that the interval will estimate the population value well. The confidence level C determines how many standard deviations to go to either side of the sample value to create the confidence interval. The number of standard deviations will be a critical value corresponding to an area of $C\%$ in the center of a bell curve. The appropriate standard deviation for a sample mean is $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$, where σ is the population standard deviation.

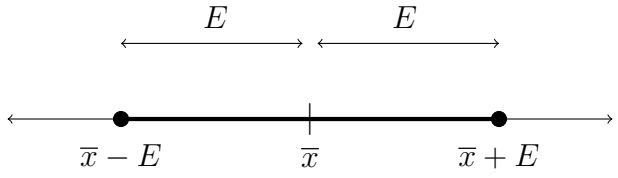
In situations where we don't know the population mean, it is rare that we would know the population standard deviation. So we can either be given some assumption of the population standard deviation σ, or we can use the sample standard deviation s in its place.

In the situation where we have some assumption for the population standard deviation σ, the full margin of error would be $E = z\left(\dfrac{\sigma}{\sqrt{n}}\right)$, where the z-score is the critical value of the standard normal distribution that has $C\%$ in between $-z$ and $+z$ of the bell curve. Then the confidence interval to estimate the population mean μ would be $\overline{x} \pm z\left(\dfrac{\sigma}{\sqrt{n}}\right)$.

Example: According to the CDC, the average height of men in the US in 2016 was 69 inches with a standard deviation of 3 inches. We would like to estimate the average height in 2022, using a sample of 800 men. The average height of the sample is 69.5 inches. Compute a 95% confidence interval for the mean height of all men in the US in 2022, assuming that the standard deviation is the same 3 inches.

Solution: The interval is $\overline{x} \pm z\left(\frac{\sigma}{\sqrt{n}}\right)$. We have the sample mean $\overline{x} = 69.5$ and sample size $n = 800$. Now we just need the z-score that splits bell curve into 95% center piece. If 95% is in the center, then the other 5% is on the edges, split in half, so the lower tail is 2.5% = 0.025 in area. Use the table or calculator function **invNorm(0.025)** to get the correct z-score of -1.96. Now the confidence interval is $69.5 \pm 1.96\left(\frac{3}{\sqrt{800}}\right) = 69.5 \pm 0.2$ or the interval of values from 69.3 to 69.7. To state this in words, we can be 95% confident that the average height of men in the US in 2022 falls somewhere in the interval 69.3 to 69.7 inches.

Now that you know how a confidence interval is formed, you can use that process to deconstruct a given interval and determine the sample mean and margin of error that were used to create it. To fully understand how that works and see that it is very straight forward, notice how the pieces fit into the numberline diagram below.



The sample mean is always right smack in the middle of the interval, since that is where the interval starts and the margin or error is the same distance out on both sides. The sample mean is the average of the lower and upper bounds of the interval, and the margin of error is the distance from the sample mean to either the lower or upper bounds.

Example: If a confidence interval for the mean price of particular model of car is from $\$22,000$ to $\$26,800$, find the sample mean and margin of error used to form this interval.

Solution: The sample mean is in the middle at $\frac{22,000 + 26,800}{2} = \$24,400$ and the margin of error is $26,800 - 24,400 = \$2,400$.

4.1.2 The T-distribution

In the situation where we have no idea what the population standard deviation σ might be, we can use the sample standard deviation s in its place. The problem with that is every sample could have a different value of s, leading to a distribution for the sample mean that varies much more and does not fit the Standard Normal distribution. So does that mean we cannot calculate a confidence interval?

We cannot use z-scores, but luckily statisticians analyzed the new distribution and it turns out to be a flattened bell curve, but the shape changes as the sample size changes. When the sample size is small, the distribution is a very flat and wide somewhat bell curve. As the sample size increases, the distribution becomes more like the Standard Normal distribution. This new flattened bell curve distribution is called the **Student's T-distribution** or just the T-distribution.

Since each sample size has different values for the T-distribution, there is not just one standard T-distribution. To get critical values requires not only the area, but the sample size. The new parameter of the T-distribution is called **Degrees of Freedom** or df. The degrees of freedom is one less than the sample size, $df = n - 1$.

In the back of this book is a T-table, showing T critical values that correspond to confidence levels and their tail areas for various degrees of freedom. It is setup differently that the Standard Normal Z-table. Some calculators have the function **invT(tail area, df)** which gives the negative of the t-value. Once we have the t-value, the margin of error for a confidence interval becomes $E = t\left(\frac{s}{\sqrt{n}}\right)$ and the confidence interval estimate for the population mean becomes $\overline{x} \pm t\left(\frac{s}{\sqrt{n}}\right)$.

Example: City officials did a survey of 50 businesses and their average revenue was $\$189,250$ with a standard deviation of $\$48,500$. They would like to estimate the average revenue for all businesses in the city. Compute a 95% confidence interval.

Solution: The interval is $\overline{x} \pm t\left(\frac{s}{\sqrt{n}}\right)$. We have the sample mean $\overline{x} = 189,250$ and sample size $n = 50$. We do not have the population standard deviation so we will use the sample value $s = 48,500$. Now we just need the t-score that splits bell curve into 95% center portion. If 95% is in the center, then the other 5% is on the edges, split in half, so the lower tail is 2.5% = 0.025 in area. The degrees of freedom is $df = 50 - 1 = 49$. Use the table or calculator function **invT(0.025,49)** to get the correct t-score of -2.010. Now the confidence interval is $189,250 \pm 2.010\left(\frac{48,500}{\sqrt{50}}\right) = 189,250 \pm 13,786$ or the interval of values from \$175,464 to \$203,036. To state this in words, we can be 95% confident that the average revenue for all businesses in the city falls somewhere in the interval \$175,464 to \$203,036, but there is a 5% chance it does not.

One limitation of using the table, is that tables will only t-values for limited values of degrees of freedom and confidence levels. The table at the end of this book only covers the five most often used confidence levels and df from 1 to 50, then 60, 70, 80, 90, 100, 200, 300, 500, and 1000. If a problem requires df = 187, then just use the closet entry of df = 200.

Example: A farm wishes to sell their eggs as large size. The US Department of Agriculture requires a minimum weight of 2 oz. on average for the dozen in order for the eggs to be labeled as "large" size. They weigh a sample of 84 eggs and find they have mean weight of 1.95 oz. with a standard deviation of 0.25 oz. They would like to estimate the average weight of their eggs. If the interval includes 2 oz., then they can label them as large. Compute a 98% confidence interval to 2 decimal places and state whether they can use the "large" label.

Solution: The interval is $\overline{x} \pm t\left(\frac{s}{\sqrt{n}}\right)$. We have the sample mean $\overline{x} = 1.95$ and sample size $n = 84$. The sample standard deviation is $s = 0.25$. Now we just need the t-score that splits bell curve into 98% in the center portion. If 98% is in the center, then the other 2%

is on the edges, split in half, so the lower tail is $1\% = 0.01$ in area. The degrees of freedom is $df = 84 - 1 = 83$. Use the table at closet df = 80 giving approximate t-score of 2.374 or calculator function **invT(0.01,83)** to get the correct t-score of -2.372. Now the confidence interval is $1.95 \pm 2.372\left(\frac{0.25}{\sqrt{84}}\right) = 1.95 \pm 0.06$ or the interval of values from 1.89 to 2.01 oz. Since this interval includes the possibility of a mean of 2 oz., they can use the large label on their egg cartons.

****Try this on your own**: A 2012 study with 543 participants with type 1 diabetes, had their average life expectancy to be 69 years with a standard deviation of 21 years. Find the 90% confidence interval to estimate the average life expectancy of all people with type 1 diabetes.

4.1.3 Estimating the Population Proportion

Another parameter that is very important in today's society and frequently estimated with confidence intervals is the population proportion of some characteristic of focus. Some examples include percentage of votes for candidates or issues, demographic proportions such as percent of people under poverty level, out of work, or own a home, as well as health issues such as percent of homes with special needs children or percent of people with mental illness.

In all of these situations, the true population proportion (or percentage) is often difficult to measure, so a survey or poll is taken. The population proportion is shown by the symbol p. The sample proportion measured is shown by the symbol $\hat{p}$ and is called "p-hat". For most sample sizes, the sample proportion $\hat{p}$ varies by a normal distribution around the true population proportion p.

The $C\%$ confidence interval to estimate the population proportion p is given by the formula $\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where the margin of error $E = z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, n is the size of the sample used, and the z-score is the critical value of the standard normal distribution that has $C\%$ in between $-z$ and $+z$ of the bell curve. The sample proportion is not given, it is found by the ratio of how many subjects fit a characteristic divided by the sample size, $\hat{p} = \dfrac{x}{n}$.

Example: Darrell is running for class president. His friends took a poll from 80 students and 45% said they would vote for Darrell. They estimated the margin of error to be 10%. Find the interval that likely contains the proportion of students who would vote for him.

Solution: The interval is $\hat{p} \pm E = 45 \pm 10$ which results in an interval from 35% to 55%. This means that Darrell is not guaranteed to win, so he needs to do some more campaigning to ensure more votes.

Example: A poll of 400 people asked if they felt anxiety concerning their job security. There were 280 people who said they felt anxiety. Find the 90% confidence interval to estimate the proportion of people in the U.S. who feel anxiety concerning their job security.

Solution: First we need to calculate the sample proportion $\hat{p} = \frac{280}{400} = 0.7$. The z-score is **invNorm(0.05)** which is -1.645. The interval is $\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.7 \pm 1.645\sqrt{\frac{0.7(1-0.7)}{400}} = 0.7 \pm 0.038$ which results in an interval from 0.662 to 0.738 or 66.2% to 73.8%. This shows that there is likely a large majority of people who feel anxiety concerning their job security.

****Try this on your own**: A company sampled their new product and found that 4 out of 80 items were defective. Find the 95% confidence interval to estimate the proportion of items that might be defective for the entire product line.

4.1.4 Required Minimum Sample Size

When gathering sample data in order to do a confidence interval, the confidence level C can be chosen. The most common level is $C = 95\%$. Higher confidence levels make much wider intervals which can be too broad to give meaningful results. Lower levels, lack the confidence.

The margin of error depends upon the sample size, so cannot be chosen so easily. The margin can be decreased by selecting a larger sample size, but there are often limits on how much data can be obtained. For a desired margin of error, the minimum sample size can be derived that will give at most that amount of margin. The process for finding the required minimum sample size depends upon the parameter being estimated, the mean or the proportion, since the formulas for those intervals are very different.

For population means, the margin of error is $E = z\left(\frac{\sigma}{\sqrt{n}}\right)$. In order to find the sample size, we need to know the desired error E, the z for the desired confidence level, and the population standard deviation σ. It was already mentioned that σ is often unknown. To do t-intervals, we used the sample standard deviation s in its place. We cannot do that here since we are trying to determine the sample size to use, so we don't have a sample yet. What can be done is to use some assumed value for σ.

The formula can be rearranged to solve for $n = \left(\frac{z\sigma}{E}\right)^2$. If the value is a decimal number, we must round UP to the next whole number, since sample size is a count and cannot go below the minimum.

Example: In 2019, several studies have concluded that the average student loan debt in the US was approximately $\$30,000$ per student who has taken out loans. The standard deviation was approximately $\$5,450$. In order to estimate the current average student loan debt, find the minimum sample size required to have a maximum margin of error of \$800 for a confidence level of 93%.

Solution: The z-score is given by invNorm(0.035) = 1.81 and the minimum required sample size is $n = \left(\frac{z\sigma}{E}\right)^2 = \left(\frac{1.81*5450}{800}\right)^2 = 152.04$. Since this is the minimum, we must complete each person and go up to 153. To get the desired results, the survey must have at least 153 people to have the error be less that \$800.

For population proportions, the margin of error is $E = z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Rearranging the formula gives $n = \left(\frac{z}{E}\right)^2 \hat{p}(1-\hat{p})$. Here we must know the sample proportion $\hat{p}$, which is a problem. We don't have the sample yet.

There are two ways around this. The first way is to use a previous sample proportion value, provided that value is assumed to be close to what might occur now. The second way is simply to use a proportion of 50%. A proportion split of 50-50 gives the widest interval and largest sample size for the conditions. This is the more conservative approach. Using a proportion of 50% is recommended unless the researchers feel confident about the previous sample proportion still being valid.

Example: In May 2022 a global poll done by US News across 17 nations found that 85% of people surveyed said they had an unfavorable attitude toward Russia. In order to estimate the current proportion, find the minimum sample size required to have a maximum margin of error of 2.5% for a confidence level of 95%. Do the calculation both ways, using the previous value of 85% as well as using 50%.

Solution: The proportions must be converted to decimal form as 0.85 and 0.50. The margin is 0.025. The z-score is given by invNorm(0.025) = 1.96. Using the previous proportion, the minimum required sample size is $n = \left(\frac{z}{E}\right)^2 \hat{p}(1-\hat{p}) = \left(\frac{1.96}{0.025}\right)^2 (0.85)(1-0.85) = 783.69$. The minimum sample size by this method is 784.

Using 50%., the minimum required sample size is $n = \left(\frac{z}{E}\right)^2 \hat{p}(1 - \hat{p}) = \left(\frac{1.96}{0.025}\right)^2 (0.50)(1 - 0.50) = 1536.64$. The minimum sample size by this method is 1,537.

****Try this on your own**: Find the minimum required sample size to estimate the mean time that teenagers spend streaming video content per day, with a maximum error of 0.5 hours in a 90% confidence interval. Use the assumed value $\sigma = 2.8$ hours.

4.1.5 Exercises: Confidence Intervals

Solutions appear at the end of this textbook.

1. A confidence interval for estimating the population mean was calculated to be from 88 to 128. What are the sample mean and margin of error that were used to calculate the interval?

2. The mean SAT exam score is 1,000 with a standard deviation of 200. A city would like to estimate the average SAT score for their students, using a sample of 350 students. The average SAT score of the sample is 1,085. Compute a 95% confidence interval for the mean SAT score for all students in that city, assuming that the standard deviation is the same 200.

3. According to a study done at Montreal's Royal Victoria Hospital in 1990, the average weight of newborn babies born to Chinese immigrants was 3,195 grams with a standard deviation of 493 grams from a sample of 1,597 babies. Compute a 99% confidence interval for the mean height of all babies born to Chinese immigrants in Canada.

4. In a poll of 800 people, 58% said they would vote for the current governor and re-elect her. Find the 95% confidence interval to estimate the percent of voters who would likely vote for her in the election.

5. In order to estimate a population proportion, find the minimum sample size required to have a maximum margin of error of 3% for a confidence level of 96%. Assume $\hat{p} = 50\%$.

4.2 Hypothesis Testing

Imagine a person has been arrested for murder. Their ex-roommate was found dead a few hours after they went to confront them about leaving them with the lease and stealing some of their stuff. The person knows they are innocent, but they were home alone after confronting the ex-roommate, so have no alibi. With evidence against them and possible motive and opportunity, they are worried. They find a good lawyer and the case goes to trial.

The prosecutor will try to show that there is plenty of strong evidence against them and convince the jury that there is no doubt that they are guilty and they should reject their claim of not-guilty. Their lawyer will defend them and try to show that although there is some evidence suggesting they could have done it, it is not strong enough, so the jury cannot reject their claim of not-guilty.

If the jury reaches a verdict of not-guilty, they go free. However, they are not saying they believe then person to be a perfect innocent person, just that there was not enough evidence to have a strong sense of their guilt. There is a similar process in statistics where we examine evidence and see how it matches two opposing claims. That process is called **Hypothesis Testing**.

All hypothesis tests are setup the same way. There will be two opposing claims that state something about certain population parameters. The evidence will be similar statistics from one or more samples. The evidence will compared to the what the claims state. Whichever claim is more likely based on the evidence will remain and the other rejected.

The two opposing claims are called Hypotheses. There will be the default claim called the **Null Hypothesis** and an opposing claim called the **Alternative Hypothesis**.

Just as a person on trial is assumed innocent until proven guilty, in statistical hypothesis tests, the null hypothesis will be assumed to be true until we can find strong enough evidence

against it. If the sample evidence is strongly against it, the null hypothesis will be rejected and the alternative will be assumed as likely. If the sample evidence is not strongly against the null hypothesis, then the null hypothesis will be considered as plausible (not absolutely true) and the alternative hypothesis will be rejected.

Notice that the null hypothesis cannot be absolutely proven as fact, since we will only have a sample as evidence. This is similar to court, where non-guilty does not prove the person innocent. It only means they could not find enough evidence to say guilty.

The particular details for a hypothesis test depends upon the population parameter involved and the type of sample evidence collected. In this book , we will only look at evidence from one sample to compare to a claim about one population mean or population proportion.

The symbols for the null and alternative hypotheses are H_o and H_a. The null hypothesis will be setup in the form H_o : parameter $= \#$ and the alternative as H_a : parameter $< \#$ or parameter $> \#$ or parameter $\neq \#$. The three options for the alternative are referred to as a left-tail test, right-tail test, or two-tail test. These labels will become clear when we show the full details in examples.

As we have already seen, most sample statistics follow a bell curve centered around the true population parameter, so just like in the section on confidence intervals we will use the normal and t-distributions. In order to use one of these distributions we have to have the conditions that the population variable follows a somewhat bell curve, or by the central limit theorem, the sample size must be 30+.

4.2.1 Z-test for the mean

A company that manufactures electric cars just came out with a new model they claim has a higher maximum driving range than the old model. The old model had an average max range of 210 miles with a standard deviation of 22 miles. Can we test their claim?

We can do a hypothesis test with a sample of 30+ cars, since we don't know if the range follows a bell curve itself. With sample size at least 30, the distribution of the sample mean will be approximately bell shape. Assuming the population standard deviation of $\sigma = 22$ miles remains the same, we can use the normal distribution.

First let's setup the two hypotheses. Whichever claim has an inequality involved is always the alternative. The null hypothesis must have equality, so we have a specific population parameter value for our assumption and the calculations in our process.

The hypotheses here will be $H_o : \mu = 210$ and $H_a : \mu > 210$. The alternative is what the company is claiming, that the new model has an improved max average range. The null is stating the default, that the new model is the same as the old model in max average range.

In a court of law, the jury cannot weigh the evidence unless they have some benchmark or criteria that sets the rules for how to know if the evidence fits with the claim or not, and how close/far it has to be in order to switch between guilty or not guilty. The criteria are the laws and the court procedures.

Now comes the trial in the court of statistics. The criteria used for hypothesis testing is having a specific value that the sample evidence must differ from the assumption by, in order to determine that the sample evidence is far enough away and strong enough to go against the default assumption (null hypothesis).

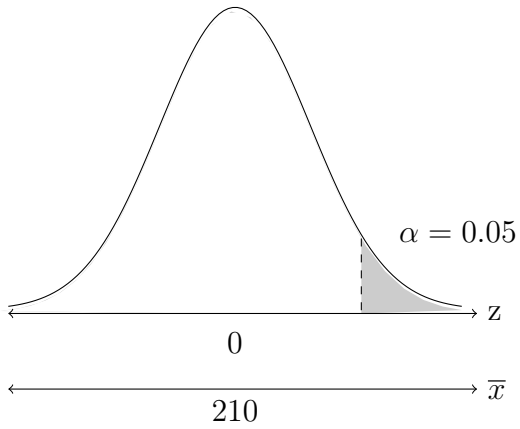
Basically, if the sample evidence is near the claimed value in the null hypothesis, then the claim seems reasonable and cannot be rejected. Only when the sample evidence is far

enough away, does it put too much doubt on the claimed value, and leads to rejecting the null hypothesis claim and replacing it with the likely alternative claim.

In math we need specific criteria, something more than just saying "too far". How far is too far? The person doing the hypothesis test will decide a cutoff for how far away by specifying a size of the tail area of the bell curve that will be considered too far. This value is called the **level of significance** or **significance level**, and is represented by the lower case greek alpha α. The most common value is 5% as a decimal $\alpha = 0.05$. Some tests are done with slightly lower or higher values.

The setup of a hypothesis test can be done without having the sample evidence yet, just as the court system and laws are in place before a person goes to trial and evidence is gathered. In our example of the electric car company, if sample means follow a bell curve, the assumed value of $\mu = 210$ will be in the center. With the assumed value $\sigma = 22$, this will be a z normal distribution.

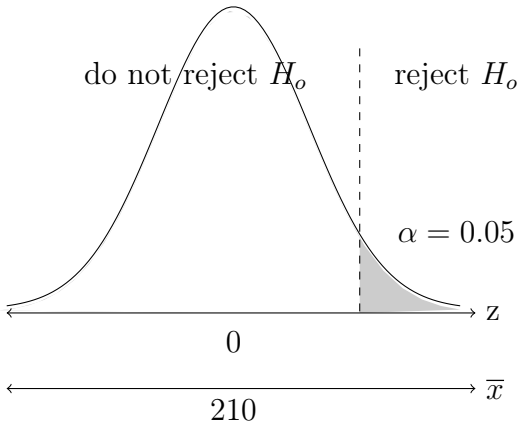
The alternative hypothesis has a greater than symbol, meaning that the sample mean (evidence) must be significantly greater than 210 in order to reject the null claim and go with the alternative. This places our cutoff on the right side tail of the bell curve. See diagram below.



Now that we have a map with a cutoff, we can relate any sample evidence to these and see if the data crosses that line. If the sample data is out beyond the cutoff and in the right tail, it will lead to rejecting the claim in the null hypothesis. If the sample mean is far enough away from the assumed population mean, that puts doubt on the assumption.

However, if the sample data does not go beyond the cutoff, then even if it does not match the claim, it could just be due to the variability of the sampling distribution and the assumed population mean is reasonable, so cannot be rejected. That does not result in the claim being fact.

Our "map" can be updated to show the two regions and the results. We then look at sample data and see where it falls. The right tail side is called the **Rejection Region** and everything else to the left is called the **Non-rejection Region**.



So we can now continue our example by looking at some evidence. A car magazine decided to test the company's claim by recording the max range on 40 cars. This sample had a mean of 218 miles. In order to compare "apples to apples" we cannot just look at 218 compared to 210. The value 218 is certainly above 210, but how far above? We have no reference on our map using just the data of miles itself.

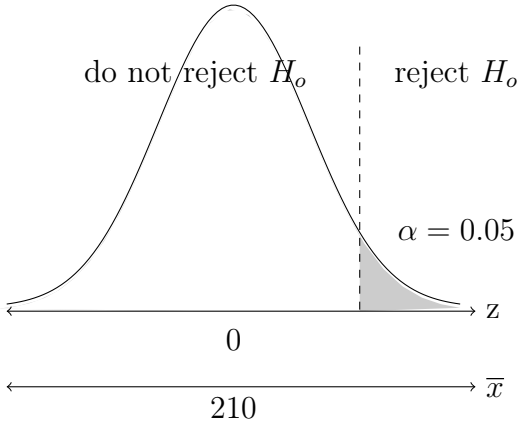
We can convert the sample mean into a z-score by the formula $z = \dfrac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

This formula gives what is called the **Test Statistic** and for the Z test it is referred to as Z_{data}.

For the sample data here, $z = \dfrac{218 - 210}{\frac{22}{\sqrt{40}}} = 2.30$, but we still cannot compare yet. This z-score is above zero, but how far, and where does it fall on our map?

There are two methods to bring everything together and be able to compare "apples to apples". The **Critical Value Method** is done by converting the tail area into z critical values and comparing those to the Z_{data} test statistic. The **P-Value Method** is done by converting the Z_{data} into an area or probability for the normal bell curve and comparing directly to α.

Now let's bring everything together and complete our example of the electric cars. The two hypotheses and the map we will use to decide are $H_o : \mu = 210 \quad H_a : \mu > 210$



For the critical value method, the z-score that cuts off the significance level α in the right tail is found by InvNorm(0.95) = 1.645. Remember that the inverse normal function uses the area to the left. With $\alpha = 0.05$ in the right tail, 0.95 is to the left. Now the test statistic $Z_{data} = 2.30$ is greater than the critical value 1.645, so the data evidence falls out into the right tail rejection region.

Therefore, we can say that there is enough evidence to reject the claim that the average max range for the new model is still 210 miles. Alternatively, it is likely greater than 210

miles. Notice we cannot say it is exactly 218, that was only for that sample, but 218 is far enough away from 210 to reject the 210 and go with the population mean being something greater than 210.

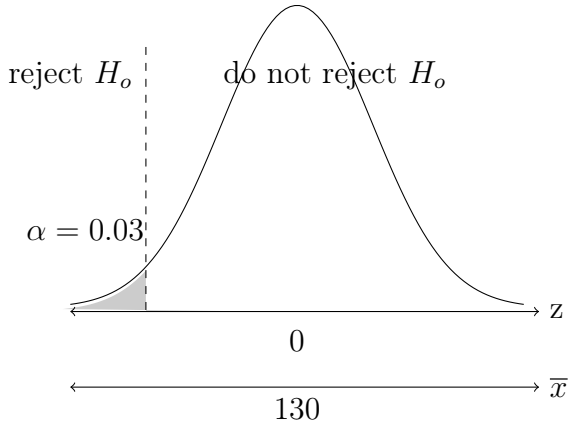
For the p-value method, the test statistic $Z_{data} = 2.30$ is converted into a right tail standard normal probability by normalcdf(2.30,99999,1,0) = 0.011 and this is less than $\alpha = 0.05$. Being less than alpha, shows that the data is out further into the right tail which is the rejection region. We reach the same conclusion here, that there is enough evidence to reject the claim that the average max range for the new model is still 210 miles. Alternatively, it is likely greater than 210 miles.

To remember what to do in the p-value method there is a saying to help you.

When the p-value is low (lower than alpha), the null hypothesis has to go.

Example: Your friend goes to a fancy private college and claims that the students there have a much higher IQ than average college students. They claim that the average IQ at their school is 130, which is considered extremely bright. You don't believe it, and think the average IQ at his school is closer to the mean for college students in general (which is about 110). IQ scores follow a normal distribution with $\sigma = 15$. You survey 8 students at that college and find they have an average IQ of 116. Setup and perform a hypothesis test at the $\alpha = 0.03$ significance level using the p-value method.

Solution: Their claim of 130 is an exact value, so that will be the null hypothesis $H_o : \mu = 130$. You think the average is less, so the alternative is $H_a : \mu < 130$. With less than, we are doing a left-tail test. The setup for the bell curve is

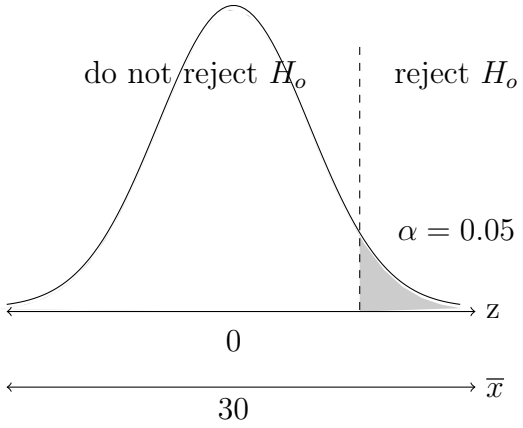


$Z_{data} = \dfrac{116 - 130}{\frac{15}{\sqrt{8}}} = -2.64$, the p-value = normalcdf(-99999,-2.64,0,1) = 0.004

Since the p-value $<$ alpha, the data is far out into the left tail, so we reject the null hypothesis. Therefore, there is enough evidence to reject the claim that the mean IQ at the fancy college is 130. It is most likely something less than 130.

Example: A collector of Sci-fi action figures has been following the trends in the value of popular character figures in mint condition. The mean value of mint condition new action figures released last year was \$30 with a standard deviation of \$8 and followed a normal distribution. They believe that the demand will increase over time and those same action figures will be worth more on average. Two years later, they look at a sample of 10 of those action figures find they have an average price of \$33. Setup and perform a hypothesis test at the $\alpha = 0.05$ significance level using the critical value method.

Solution: Their claim of increase in value, gives the alternative hypothesis $H_a : \mu > 30$. The null is then $H_o : \mu = 30$. With greater than, we are doing a right-tail test. The setup for the bell curve is



$Z_{data} = \dfrac{33 - 30}{\frac{8}{\sqrt{10}}} = 1.19,$ $Z_{crit} = \text{InvNorm}(0.95) = 1.645$ or from the z-table.

Since the test statistic 1.19 is less than the critical value, it is to the left of the cutoff and falls in the non-rejection region. Even though the sample mean is greater than 30, that is not enough evidence to reject the claim of the overall mean price of the action figures being \$30. It is plausible that this greater sample mean is just a fluctuation and the true mean of all the action figures is still \$30.

****Try this on your own**: The mean body temperature of healthy adults is typically reported as 98.6° F with standard deviation of 0.7° F. A nurse at a clinic has experienced several lower body temperatures and wants to test the hypothesis that the mean body temperature of healthy adults is less than 98.6° F. She records temperatures from a sample of 50 people and obtains a sample mean of 98.0° F. Perform the hypothesis test using the p-value method with a significance level of $\alpha = 0.02$.

4.2.2 T-test for the mean

In situations where the population standard deviation is unknown and there is no assumed value to use, we can substitute the sample standard deviation s in its place. Just as was

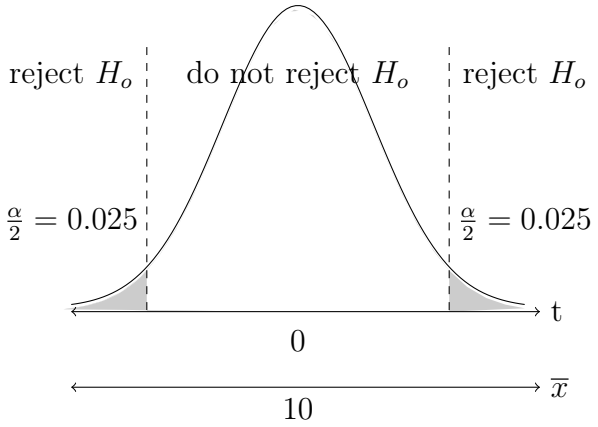
done for confidence intervals, when using the sample statistics, we have to switch from the normal distribution z to the student's T-distribution. Otherwise, hypothesis testing is the same procedure whether we use z or t.

From the data, the sample mean is converted into the test statistic t-score by the formula $t_{data} = \dfrac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$. Sometimes only the data observations are given, so the sample mean and standard deviation must be calculated. Most T-tests are done by the p-value method, since the tables only have limited critical values.

Example: Mr. Fit the diet guru claims that people who follow his diet will lose an average of 10 pounds in a week. You have a fitness blog and before you post his information, you want to see some results to see if the average weight loss is 10 pounds or something different either way. Assume weight loss follows a bell shape curve. The observations for 12 people's weight loss is show below. Setup and perform a hypothesis test at the $\alpha = 0.05$ significance level using the p-value method.

Weight loss values: 11, 13, 9, 9, 10, 11, 6, 14, 12, 10, 8, 10

Solution: The claim of 10 pounds is an exact value, so that will be the null hypothesis $H_o : \mu = 10$. The average might be something other than 10, so the alternative is $H_a : \mu \neq 10$. With not equal, we are doing a two-tail test. Alpha is split in half, and so is the p-value to match. The setup for the bell curve is

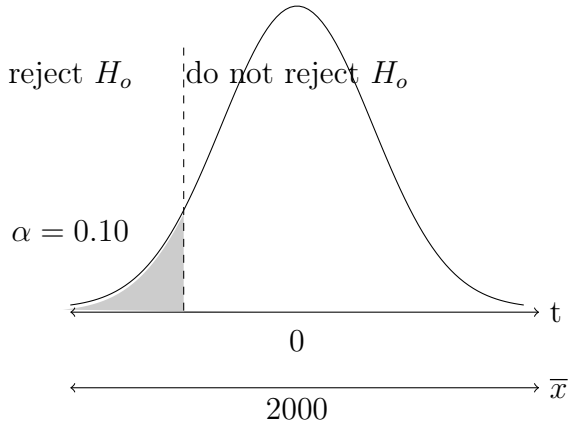


Sample mean $\overline{x} = 10.25$, sample standard deviation $s = 2.18$, $df = 12 - 1 = 11$,
$t_{data} = \frac{10.25-10}{\frac{2.18}{\sqrt{12}}} = 0.397$, right tail half of p-value = tcdf(0.397,9999,11) = 0.3495,
total p-value = $2 * 0.3495 = 0.06990$

Since the p-value > alpha, the data is close in near the null claim, so we cannot reject the null hypothesis. Therefore, there is not enough evidence to reject the claim that the mean weight loss is 10 pounds. Mr. Fit knows his process well.

Example: Alfonso is interviewing for a retail computer sales job and tells the company that he can close an average \$2,000 of sales per day. They hire him conditionally and require him to prove himself over the next 30 days. The manager is skeptical and thinks he will fall short. Over that time period, the sample statistics for Alfonso's sales are $\overline{x} = 1,948$, sample standard deviation $s = 209$. Perform a hypothesis test at the $\alpha = 0.10$ significance level using the p-value method.

Solution: The claim of \$2,0000 is an exact value, so that will be the null hypothesis $H_o : \mu = 2000$. The manager thinks he will fall short, so the alternative is $H_a : \mu < 2000$. With less than, we are doing a left-tail test. The setup for the bell curve is



$t_{data} = \frac{1948-2000}{\frac{209}{\sqrt{30}}} = -1.362$, p-value = tcdf(-9999,1.362,29) = 0.0918

Since the p-value $<$ alpha, the data is far enough from the null claim, so we reject the null hypothesis. Therefore, there is enough evidence to reject the claim that the mean sales per day is \$2,000. His sales are likely less than that and he might not keep this job.

****Try this on your own**: Based on a 2013 research study, the FDA has proposed a maximum level of nicotine in cigarettes of 0.5 milligrams. A tobacco company attempts to make a new cigarette below the proposed level. A health researcher doubts their ability and thinks they probably have more than the 0.5 level. They test a sample of 100 cigarettes. The nicotine levels had $\overline{x} = 0.52$, $s = 0.02$. Perform a hypothesis test at the $\alpha = 0.05$ significance level using the p-value method.

4.2.3 Z-test for the proportion

The last type of test we will cover in this course is testing a claim about a population proportion using the sample proportion (percentage) as evidence. These can be confusing due to the fact that three of the variables in these tests are proportions/probabilities and are represented by the letter p. There will be some claimed population proportion p, the sample proportion $\hat{p}$, and the p-value associated with the likelihood of the data being as far away as observed.

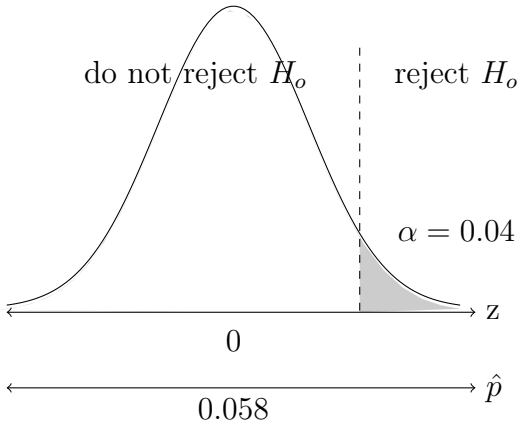
Sample proportions for the most part follow a normal distribution centered about the population proportion, so we will go back to using z-scores. The sample size generally has to be about 20+ to reasonably assume normal distribution.

The hypotheses are setup the same way, just with p instead of μ and the test statistic is given by the formula $Z_{data} = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$ where $\hat{p} = \dfrac{x}{n}$ with x successes out of n observations.

Example: A 2015 survey by the Rand corporation of almost 200,000 military personnel, found that 5.8% of the US military identify as lesbian, gay, or bisexual. If a 2021 survey of 11,000 military personnel found 695 identifying as LGB, would that be enough evidence

to suggest that the proportion has increased? Setup and perform a hypothesis test at the $\alpha = 0.04$ significance level using the critical value method.

Solution: The question of increase in value, gives the alternative hypothesis $H_a : p > 0.058$. The null is then $H_o : p = 0.58$. With greater than, we are doing a right-tail test. The setup for the bell curve is



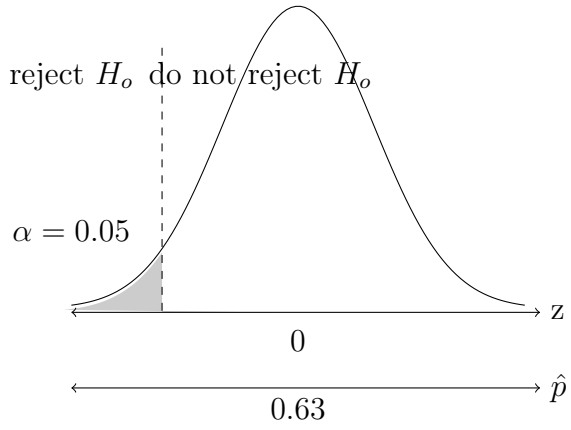
$$\hat{p} = \frac{x}{n} = \frac{695}{11000} = 0.063 \qquad Z_{data} = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.063-0.058}{\sqrt{\frac{0.058(1-0.058)}{11000}}} = 2.24$$

$Z_{crit} = \text{InvNorm}(0.96) = 1.75$ or from the z-table.

Since the test statistic 2.24 is greater than the critical value, it is to the right of the cutoff and falls in the rejection region. The sample proportion of 6.3% is enough evidence to reject the claim of the population proportion still being 5.8%. It is likely that the proportion has increased over the past 6 years.

Example: In a study done by the American Geriatrics Society from 2014-2016, 63% of medicare patients in hospice care are prescribed opioid drugs. In 2022, if a survey finds 1,282 out of 2,105 medicare patients in hospice care were prescribed opioid drugs, would that be enough evidence to suggest that the proportion has decreased? Setup and perform a hypothesis test at the $\alpha = 0.05$ significance level using the p-value method.

Solution: The question of decrease in value, gives the alternative hypothesis $H_a : p < 0.63$. The null is then $H_o : p = 0.63$. With less than, we are doing a left-tail test. The setup for the bell curve is



$\hat{p} = \dfrac{x}{n} = \dfrac{1282}{2105} = 0.61$ $\qquad$ $Z_{data} = \dfrac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} = \dfrac{0.61-0.63}{\sqrt{\frac{0.63(1-0.63)}{2105}}} = -1.99$

p-value = normalcdf(-99999,-1.99,0,1) = 0.023

Since the p-value $<$ alpha, the data is far out into the left tail, so we reject the null hypothesis. Therefore, there is enough evidence to reject the claim that 63% of medicare patients in hospice care are prescribed opioid drugs. It is most likely lower than 63%.

****Try this on your own**: According to a survey by the Federal Reserve 12% of the 11,000 US investors surveyed said they have invested in some type of cryptocurrency in 2021. In March 2022 Quinnipiac University did a survey of 1,936 adults and 16% said they own cryptocurrency. Would that be enough evidence to suggest that the proportion has increased? Setup and perform a hypothesis test at the $\alpha = 0.05$ significance level using the p-value method.

There are two ways that a hypothesis test can reach a wrong conclusion.

1. Rejecting the null hypothesis when it was actually true, which is known as a **Type-1 Error**. This is like rejecting a person's claim of not-guilty and putting them in jail when they are actually innocent. The probability of a type-1 error is equal to the significance level alpha α.

2. Not rejecting the null hypothesis when it was actually false, which is known as a **Type-2 Error**. This is like not finding a person guilty and setting them free when they actually are guilty.

So how can these errors occur? In a court of law, even if a person committed a crime, if the police or prosecution do not do a good job of collecting or presenting enough evidence, they might not get a guilty verdict. Unfortunately, some innocent people are found guilty due to falsified evidence or mistaken identity.

In statistical hypothesis testing, we use the bell curve property that MOST of the values fall near the middle. Sometimes a sample could just be randomly far away, when most other samples would not. That leads to the sample data falling out into the rejection region even though the null hypothesis might still be true. This is rare, so the conclusion usually goes with the more likely result and the evidence not supporting the null hypothesis puts doubt on the null hypothesis.

4.2.4 Exercises: Hypothesis Testing

Solutions appear at the end of this textbook.

1. For each hypothesis setup, state whether it will be a left, right, or two-tail test.

   a) $H_o : \mu = 87$ and $H_a : \mu < 87$

   b) $H_o : p = 0.18$ and $H_a : p \neq 0.18$

   c) $H_o : \mu = 900$ and $H_a : \mu > 900$

   d) $H_o : p \leq 0.5$ and $H_a : p > 0.5$

2. For each hypothesis setup, state whether it will be a z-test for a mean, a t-test for a mean, or a z-test for a proportion, and whether it will be a left, right, or two-tail test.

   a) A bank wants to know if the average household income of their customers is above the national average? They have sample data for 500 customers, but no information about the national distribution.

   b) A senator wants to know if they are still leading in the election polls after a smear advertisement from their opponent was released in an attempt to steal votes.

   c) A new type of treatment for high cholesterol is in trials. The drug manufacturer wants to know if the treatment works well or poorly and will make any difference in the amount of cholesterol in patients. They know that similar treatments have a normal distribution for the effects with a standard deviation of 25.

   d) A new basketball coach wants to know if a his technique will help the players decrease the number of fouls committed. He only knows the average from last year, but will look at upcoming data.

3. For each set of hypothesis test results, state the conclusion as reject or do not reject the null claim. Explain what you compare to make that decision.

   a) $H_o : \mu = 87$ and $H_a : \mu \neq 87$, $\alpha = 0.05$, $Z_{data} = 2.06$, p-value $= 0.04$

   b) $H_o : p = 0.18$ and $H_a : p < 0.18$, $\alpha = 0.025$, $Z_{data} = -1.65$, $Z_{crit} = -1.96$

   c) $H_o : \mu = 900$ and $H_a : \mu > 900$, $\alpha = 0.02$, $t_{data} = 0.789$, p-value $= 0.27$

4. If a population last year had a mean of 400 and a standard deviation of 25, does the following sample data for this year suggest that the mean is now different? Do a hypothesis test at the 5% significance level. Use the critical value method. The sample mean is 403 with a sample size of 50.

5. According to the National Center for Health Statistics, in 2014 the average age of a woman when she has her first child was 26.3 years old. If a survey of 101 Latina mothers finds that their average age when they had their first child is 25.9 years with a standard deviation of 2 years, perform a hypothesis test at the 5% significance level to test whether the Latina mother mean age is lower than the national average. Use the p-value method.

6. Based on the CDC's National Health Interview Survey in 2020, 37% of US adults do not have a landline phone in their home. A survey of 380 college students finds that 147 of them do not have a landline. Is there enough evidence to assume that the proportion of college students without a landline is greater than the CDC survey results? Use 3% level of significance and the critical value method.

Chapter 5

Bivariate Data

5.1 Correlation and Regression

5.1.1 Correlation

In previous sections, the data sets we looked at were for one variable. Many studies are done to examine the relationship between variables. Two variables measured on the same subjects are said to have a **Correlation**, when certain values of one variable tend to occur more often with certain values of the other variable. For example, people with larger heights tend to have larger weights and those with smaller heights tend to have smaller weights. This does not mean that every tall person weighs more than every short person, just that more often it happens than not. Sometimes when two variables are related, it can be that one tends to cause the other, but many times they are just related to another variable we don't know about.

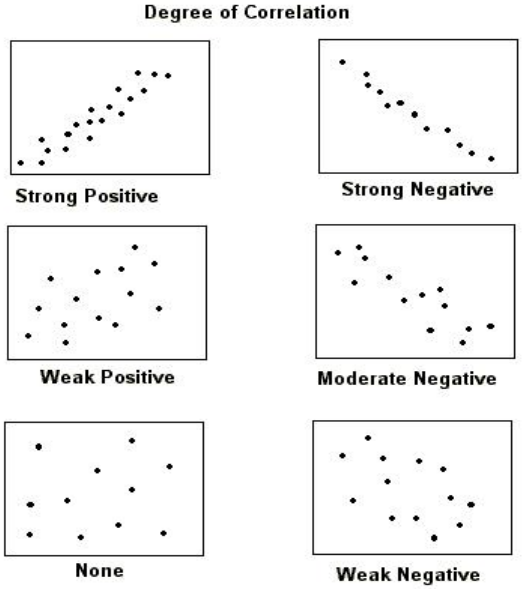
Two variables are said have **Positive Correlation**, when higher values of one variable are paired with higher values of the other (and lower with lower). Two variables are said to

have **Negative Correlation**, when higher values of one variable are paired with LOWER values of the other (and lower of one with HIGHER of the other).

When the variables are quantitative (numerical) we can get a picture of the relationship between them by creating a **Scatterplot**, which is a set of axes with values of one variable along the horizontal axis and values of the other along the vertical axis. Each pair of measurements for each subject are represented by a single point.

When looking at a scatterplot, we should examine the pattern of the points. We can describe the shape, strength and direction of the pattern. We should also examine the deviation from the main pattern by looking for outliers. The typical shape we look for is a line (linear relationship), where the strength would be how close to a perfect line the pattern is. Often when examining relationships, we can look at additional variables that are categorical and show them by different symbols or colors on the scatterplot. This can help distinguish different patterns for different groups.

The chart below shows examples of strong, weak, and no linear patterns.

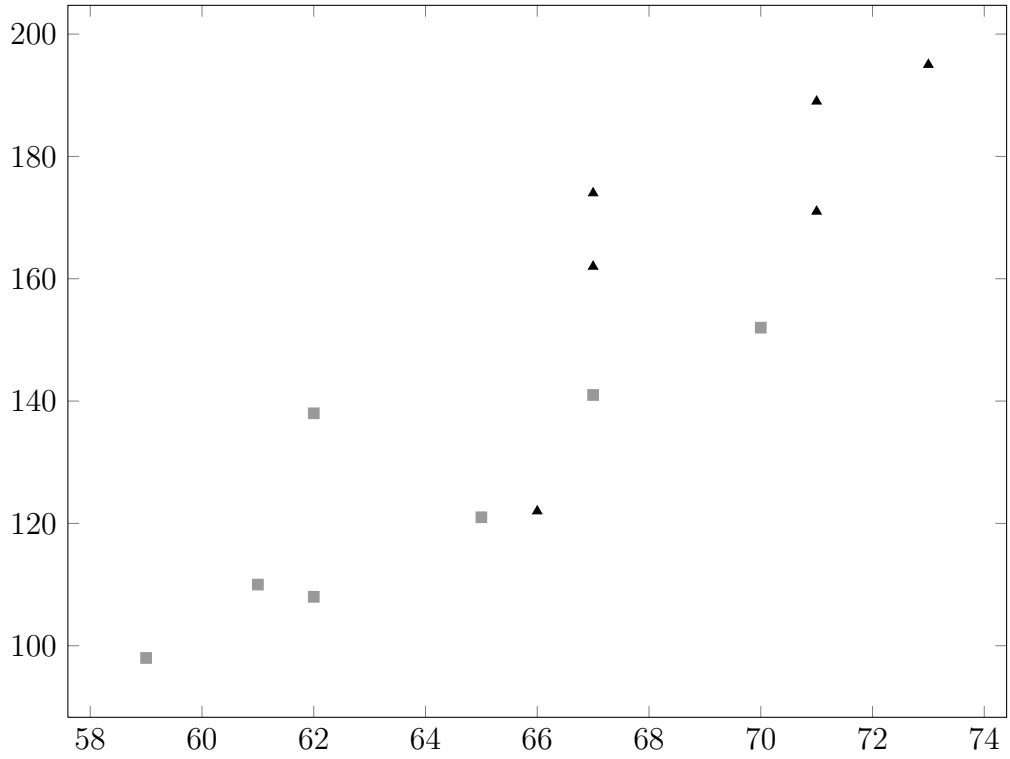


Example: The following data was collected from an actual sample of college students in a statistics class at the University of West Georgia. Create a scatterplot of the data, using

different symbols for each gender (the categorical variable). What type of patterns do you see? What shape, direction and how strong are the patterns?

Gender	F	F	F	F	F	F	F	M	M	M	M	M	M
Ht (in)	59	61	62	62	65	67	70	66	67	67	71	71	73
Wt (lbs)	98	110	108	138	121	141	152	122	174	162	171	189	195

Solution: The scatterplot is below, using gray squares for the females and black triangles for the males. Overall, there is a fairly strong linear pattern in a positive direction, with no outliers. For the females, there is a strong positive linear pattern with one outlier $(62, 138)$. The male points are similar, strong positive linear pattern with one outlier $(66, 122)$.



The **Linear Correlation Coefficient** measures the strength and direction of a linear relationship between two numerical variables. It is represented by the letter r.

Properties of the correlation coefficient, r:

1. It does not matter which variable is X and which is Y, the value of r is the same.

2. Changing the units of measurement does not change the value of r.

3. Positive (negative) value of r indicates a positive (negative) association.

4. r is always a value between -1 and $+1$.

5. A value of r close to zero indicates a weak (or no) relationship.

6. A value of r close to -1 or $+1$, indicates a strong linear relationship.

7. r is affected by outliers. Use caution if many outliers appear.

If we have two variables which we label as X and Y, there are several versions of the correlation formula. Two versions are shown here, but we will typically get r from the calculator, since even a small set of data can be time consuming and allow many chances to make mistakes. The first version of the formula uses the data values with the mean and standard deviation of each variable. The second one uses the sums of the variables and their products, which can be found from organizing the data in a table and using columns fro each type of calculation or the 1-VarStats calculator function.

$$r = \frac{\sum (x - \overline{x})(y - \overline{y})}{(n-1)s_x s_y}$$

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

So what value of r is close enough to -1 or $+1$ to say the correlation is strong? That depends upon the sample size. The table below shows the minimum absolute value of r required to say that there is a strong correlation between the variables. For the Ht/Wt data

of the statistics students, $r = +0.889$, so we would be justified in saying the relationship between Ht/Wt is a strong positive one, lower heights with lower weights, and higher heights with higher weights. We can say this because the minimum value required from the table below is 0.553, and our value of 0.889 is greater than that minimum.

Critical Values for the Linear Correlation Coefficient

sample size n	3	4	5	6	7	8	9	10	11	12	13	14	15	16
minimum $\|r\|$	.997	.950	.878	.811	.754	.707	.666	.632	.602	.576	.553	.532	.514	.497
sample size n	17	18	19	20	21	22	23	24	25	26	27	28	29	30
minimum $\|r\|$	.482	.468	.456	.444	.433	.423	.413	.404	.396	.388	.381	.374	.367	.361

To show a scatterplot and the value of r on the calculator, we need to put the data into a list. Press STAT button and under EDIT menu, select the Edit function and press ENTER. Use the arrow keys to move to the first blank under L_1. Now enter the height data from the example with the stats students. Make sure you type the values in order. Then move over to L_2 and input the weight data exactly in order, so that each height matches with its correct weight. If you mixup any values, it will mess up your calculations.

Press STAT, go over to CALC menu, scroll down to item LinReg(ax+b) and press ENTER. This will put the command onto the main screen, but we also have to tell the calculator which two lists to use (or else it will give error message or even worse, it will choose for you!). Type L_1, L_2 then hit enter again. You should see the following output:

LinReg

$y = ax + b$

$a = 6.575445816$

$b = -290.8045267$

$r^2 = .7908716696$

$r = .889309659$

This is the same value $r = +0.889$, that was mentioned previously. The other values will be discussed later. If you do not see the values for r and r^2, it is just an issue of settings

on the calculator. To change the settings, press 2nd 0 (for catalog menu), scroll down and select "diagnostics on", press enter until it says DONE. Now go back to STAT, CALC menu, LinReg(ax+b) and you will now see the full output.

To get the scatterplot, press 2nd Y= (for stat plot menu), choose Plot1, select first type icon for scatterplot. Set the Xlist to L_1 and Ylist to L_2, then press GRAPH. If you do not see it, go to ZOOM and select ZoomStat, hit enter. You should see a graph similar to the one shown with the ht/wt data previously in this section.

Many websites or spreadsheets can also compute the correlation.

Here is an easy example of how to calculate r using the formula. The data set is small and has whole numbers. It uses the bigger version of the formula which is actually easier to work with. The terms come directly from the data and can be organized into a table of columns. The formula uses the sums of each column along with the number of data points $n = 6$.

Example: For the data below, compute the correlation r. Is the r value large enough to state that the correlation is strong?

X	Y
1	10
2	6
3	9
4	4
5	6
6	5

Solution: We fill in the values under each column, for the corresponding X and Y values, then add up each column at the bottom.

X	Y	X^2	Y^2	XY
1	10	1	100	10
2	6	4	36	12
3	9	9	81	27
4	4	16	16	16
5	6	25	36	30
6	5	36	25	30
$\sum x = 21$	$\sum y = 40$	$\sum x^2 = 91$	$\sum y^2 = 294$	$\sum xy = 125$

Now we input the value $n = 6$ and all of the sums into the formula to get:

$$r = \frac{6(125) - 21(40)}{\sqrt{6(91) - 21^2}\sqrt{6(294) - 40^2}} = \frac{-90}{131.225} = -0.686$$

We compare this to the minimum of 0.811 from the table on page 97. The value -0.686 is lower (in absolute value) than the minimum, so it is not enough to say there is a strong correlation between the variables. The correlation is a weak negative correlation.

5.1.2 Regression

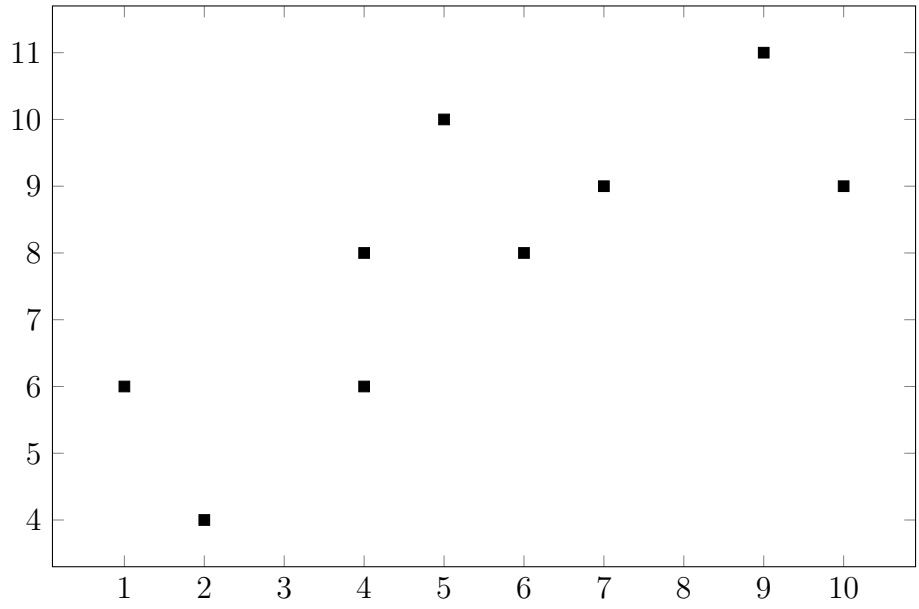
Regression is the procedure for finding an equation (and graph) which is the best fit for a set of data. If we find that two variables have a strong linear correlation, then we would like to know the best fitting line that the scatterplot follows along. This line is called the **Regression Equation**. The formulas are very complex, so we will just use the LinReg function on the calculator to get the equation.

Fitting a line to a data set is called **Linear Regression**, but if the data follows another pattern, there are other types of regression to use, which give an appropriate type of an

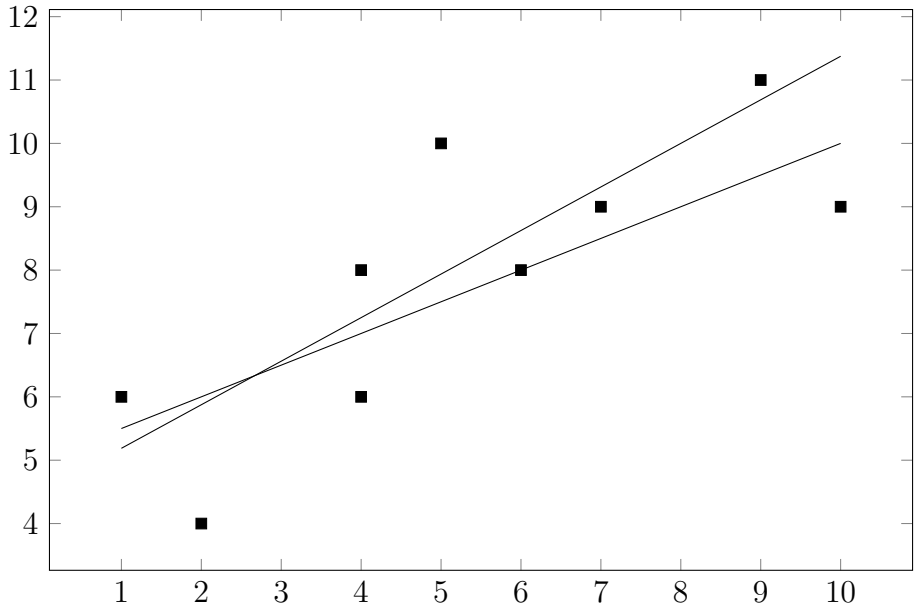
equation. Some other types of regression are quadratic, cubic, exponential, logarithmic, logistic, and sinusoidal. This book will only deal with linear regression.

We can look at a scatterplot, and if it seems to follow a linear pattern, we can sketch a reasonable line along the middle of the data to estimate where the regression line might fit. It should follow the direction of the pattern and go between the values, possibly touching some of them. There should be some values above the line and others below.

Example: Sketch a reasonable line that could fit the pattern of the scatterplot below.



Solution: The graph on the next page shows two reasonable lines. Either one could be a good fit.



To get the best fitting linear regression equation, we use the LinReg function, which we used to get r. The equation is in the form $\hat{y} = ax + b$. The value of a is the slope of the line. The value of b is the y-intercept of the line. there are also formulas which use the data values or many websites and spreadsheets can give you the regression line.

The formulas are shown here.

$$a = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

$$b = \overline{y} - a\left(\overline{x}\right)$$

In the context of the data, a represents the amount of increase(+) or decrease(-) in the output $\hat{y}$ for each one unit increase of the input x, and b represents the amount of the output $\hat{y}$ when the input x is equal to zero, if the same pattern was to continue all the way down to zero. In most cases, the pattern does not continue that far, so the intercept may not be realistic, but it is needed to use the formula to compute values within the range of the data.

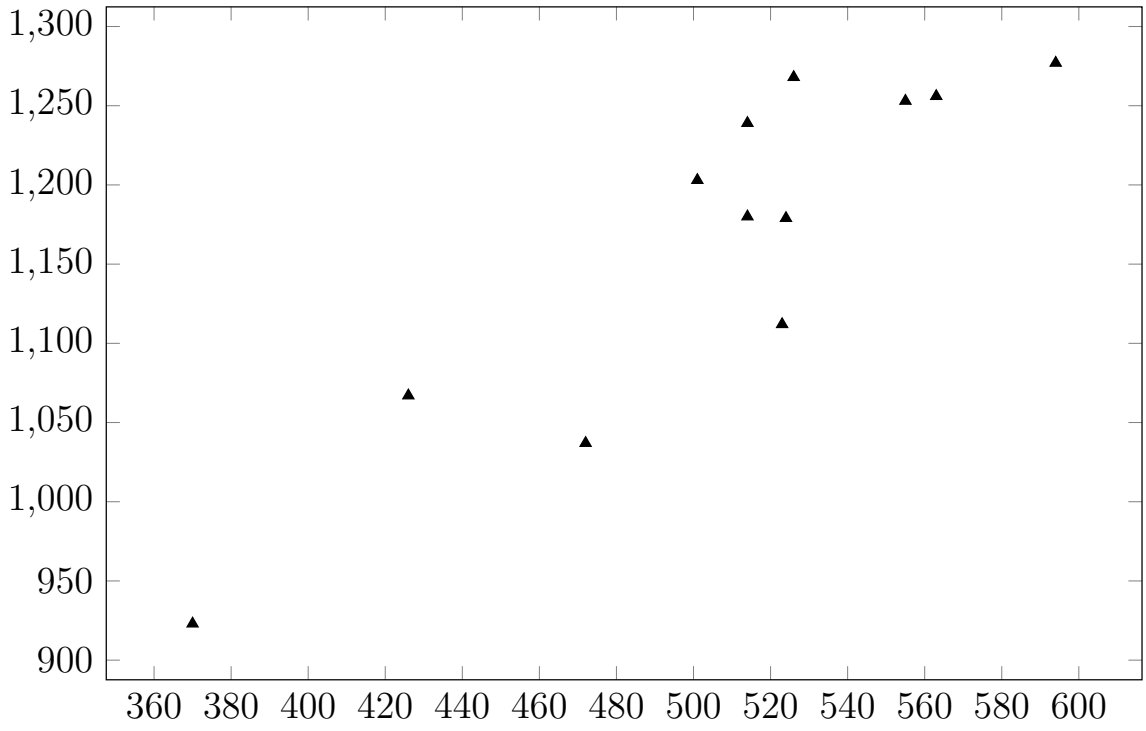
Many of the data points can be away from the line, sometimes very far away if the pattern is weak. The differences between the actual data values y and the estimated regression values $\hat{y}$ are called the **Errors** or **Residuals**. They are found by the formula $e = y - \hat{y}$. The regression line equation is the line that has the smallest possible amount of errors overall.

The regression equation can be used to predict or forecast output values for given input values. If the input value falls in the range of the original data, it is called **Interpolation**. Interpolation usually gives reasonable and realistic predictions, since the inputs are within the data range. If the input value falls outside the range of the original data, it is called **Extrapolation**. Beware of using predictions with extrapolation. For example, if you have sales data from $1999 - 2009$, predicting sales for 2024 is too far into the future to be reasonable.

Example: The table below shows data for average daily sales of ice cream at a shop over a twelve month period, as well as the average number of monthly crimes in the neighborhood over the same period. Create a scatterplot and find the correlation coefficient r. Look on the table to determine if the value of r is large enough to say there is a strong linear correlation between the variables. Then find the regression equation and use it to forecast the number of crimes output for the sales values \$450 and \$650. Are the forecasts interpolation or extrapolation? Do you notice a pattern between sales and crime? What could explain the pattern?

Sales \$	472	426	523	514	524	501	563	526	555	594	514	370
Crimes	1037	1067	1112	1180	1179	1203	1256	1268	1253	1277	1239	923

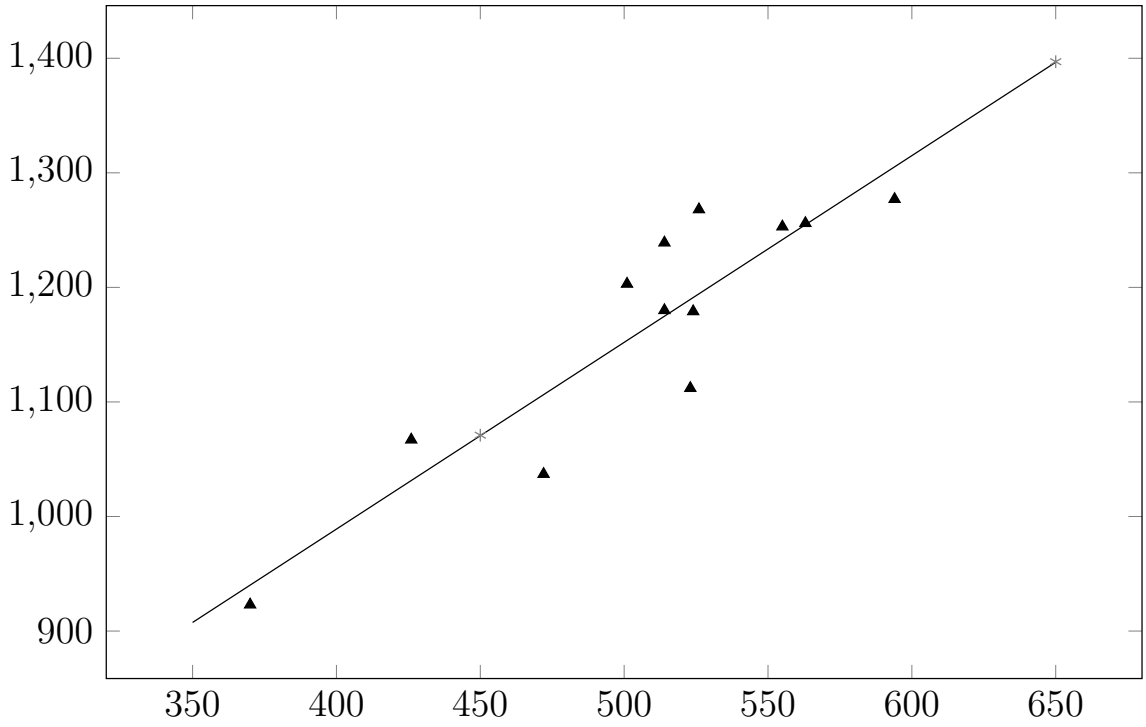
Solution: The scatterplot is below. Overall there seems to be a fairly strong linear pattern in a positive direction. This is confirmed by $r = +0.900$ which is greater than the table value 0.576.



The output from the calculator shows the regression equation $y = 1.63x + 337$. This equation means that on average the number of crimes goes up by 1.63 for every extra dollar in ice cream sales, and with no ice cream sales, there would be a minimum of 337 crimes in that month.

The predictions are $y = 1.63(450) + 337 = 1071$ crimes in a month with \$450 average daily sales in ice cream, and $y = 1.63(650) + 337 = 1397$ crimes in a month with \$650 average daily sales in ice cream. The first one is interpolation, since 450 is in the range of the data (370 to 594). The second is extrapolation, since 650 is outside the range.

The scatterplot is shown again, along with the best fit regression line and the two forecast values as asterisks. It is easy to graph the line, just connect the prediction points, since they came from the line equation.



So what does all this imply about ice cream and crime? There is a strong correlation between them, so does ice cream cause more crime, or more crime make people desire ice cream? Certainly not! Remember, strong correlation does NOT mean the variables cause one another (although it is possible they could).

During the research into this data, it was found that there is actually a hidden variable that causes both. It is hot weather. When the temperature rises, people want ice cream. There are also more people on the streets and they are hot and irritated, causing more crime. The one value that did not follow this pattern was the lowest point, which was during 100 degrees. It was too hot for most people to commit crimes or go out for ice cream.

****Try this on your own**: The table below shows data from ten people of their average monthly spending on fast food, as well as their average number of days of exercise each month. Create a scatterplot and find the correlation coefficient r. Then find the regression equation and use it to forecast the number of days of exercise output for the fast food value \$70. Is that prediction interpolation or extrapolation? Do you notice a pattern between fast food spending and exercise? What could explain the pattern?

fast food \$	20	40	58	50	140	30	90	45	100	120
exercise days	20	15	13	11	3	26	7	18	12	1

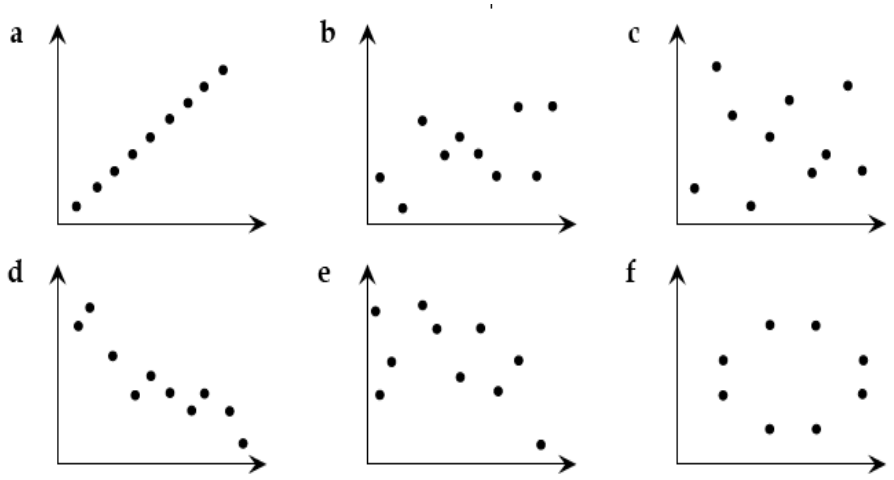
5.1.3 Exercises: Correlation and Regression

Solutions appear at the end of this textbook.

1. Explain what correlation is, and the difference between positive and negative correlation.

2. Match the most likely linear correlation values to the graphs below.

   $r = +0.7 \quad r = +0.99 \quad r = -0.4 \quad r = +0.15 \quad r = -0.86 \quad r = 0$



3. For the data below, use the formula to calculate the correlation r.

X	2	5	7	10	12	14
Y	2	6	7	9	11	14

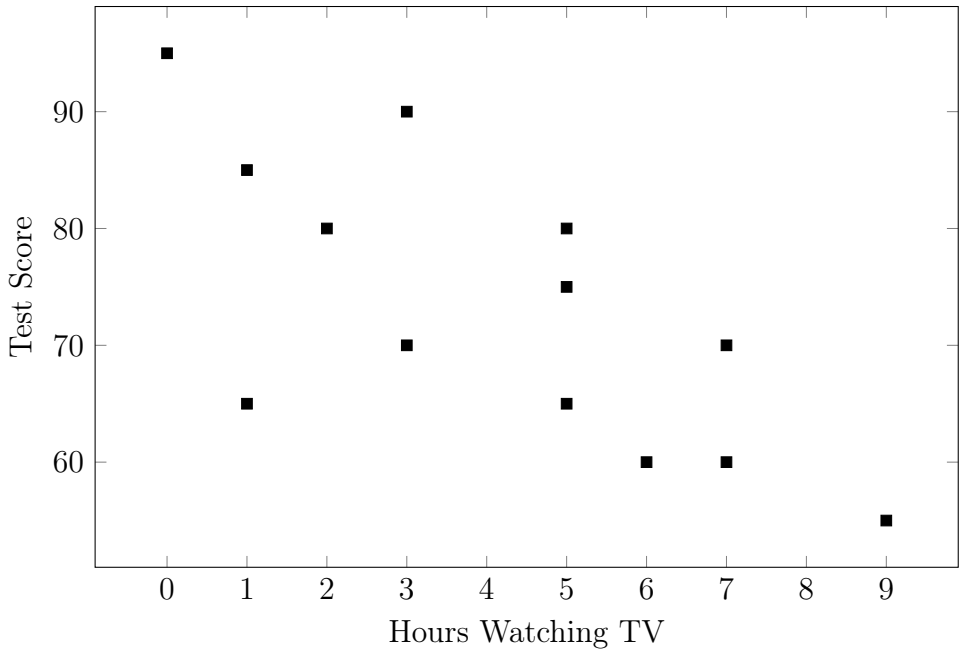
4. For the data below, use the calculator to find the correlation r. Is the r value large enough to state that the correlation is strong? Create the scatterplot on the calculator.

SAT math	643	558	703	512	552	430	605
College GPA	3.52	2.91	3.63	2.21	3.02	2.80	3.18

5. Find the regression equation for the SAT/GPA data from the previous problem. Interpret the equation values. What do they say in the context of SAT scores and GPA? Add the regression line to the scatterplot on the calculator.

6. Use the regression line from the previous problem to predict GPA for math scores of 760 and 500. Are these predictions interpolation or extrapolation?

7. Draw a good estimate of the regression line for the scatterplot below. What relationship do you notice between test scores and TV watching? Which one do you think causes the other, or could there be a hidden variable causing both?



8. What is the relationship between a linear correlation coefficient r and the slope of the corresponding regression line?

5.2 Joint Distributions

5.2.1 Marginal Frequencies

A **two-way frequency chart**, or bivariate chart, can be used with data summarized into two categories. Each category is a qualitative variable, and the category values are the labels in the chart. Two-way frequency charts are made up of cells in rows and columns.

The value in each cell is the frequency of things that belong to both the row and column categories for the cell. The values in each cell are called **joint frequencies**, since each cell is where a row and column join (meet). These are a type of joint distribution, showing how two or more variables are jointly distributed.

Example: The table below shows the breakdown of the members of the 116th Senate of the United States as of May 2020. There are 100 senators broken down by two categories, gender (male or female) and party affiliation (republican, democrat, or independent). The two rows of data represent the two gender categories and the three columns of data represent the three party categories.

	Democrat	Republican	Independent
Female	17	9	0
Male	28	44	2

The value of 17 in the first data cell is jointly in the female row and the democrat column, so the joint frequency of the female democrats is 17. This means that the US Senate consists of 17 senators who are female democrats. Likewise, the 44 is jointly in the male row and the republican column, so the joint frequency of the male republicans is 44. This means that the US Senate consists of 44 senators who are male republicans.

Each row and column can be totaled. The row totals would be over in the right margin and the column totals along the bottom margin. For this reason, the row and column totals

are called **marginal frequencies**. These frequencies are for one category only. The table is shown below with the marginal frequencies added.

	Democrat	Republican	Independent	Totals
Female	17	9	0	26
Male	28	44	2	74
Totals	45	53	2	100

For this example, the frequencies at the bottom represent counts of senators in each party. There are 45 democrats, 53 republicans, and 2 independents. These add up to the last cell in the table of 100 senators. The marginal frequencies on the right represent counts of females and males in the Senate. There are 26 females and 74 males, with grand total of 100.

****Notice that the row marginal frequencies and column marginal frequencies must add up to the same grand total. If they do not, there is something wrong in the chart.**

The joint and marginal values can also be shown as relative frequencies, usually as percentages. To calculate the relative frequencies for each cell and margin, just divide the frequency by the grand total and multiply by 100 to make percent. A common practice is to round the percentages to nearest tenth of a percent.

Example: The table below shows the breakdown of students at a private college based on class level and scholarship level. Calculate the marginal frequencies and the grand total. Then rewrite the chart with relative frequencies.

	Freshman	Sophomore	Junior	Senior
Full Scholarship	38	19	12	6
Partial Scholarship	143	72	47	21
No Scholarship	131	121	101	128

Solution: Just add down each column to get the bottom marginal frequencies, add across each row to get the right marginal frequencies. Then add each marginal row or column to check the grand total is the same. See completed table below.

	Freshman	Sophomore	Junior	Senior	Totals
Full Scholarship	38	19	12	6	75
Partial Scholarship	143	72	47	21	283
No Scholarship	131	121	101	128	481
Totals	312	212	160	155	839

To get the relative frequencies, divide each value by the grand total. For example, the relative frequency for sophomores with partial scholarship would be $\frac{72}{839} \times 100\% = 8.6\%$ and the marginal relative frequency for all freshman would be $\frac{312}{839} \times 100\% = 37.2\%$. The relative frequency table is shown below. Notice that the row marginal relative frequencies and column marginal relative frequencies both add up to 100%.

	Freshman	Sophomore	Junior	Senior	Totals
Full Scholarship	4.5%	2.3%	1.4%	0.7%	8.9%
Partial Scholarship	17.0%	8.6%	5.6%	2.5%	33.7%
No Scholarship	15.6%	14.4%	12.0%	15.3%	57.3%
Totals	37.2%	25.3%	19.1%	18.5%	100%

5.2.2 Conditional Frequencies

The last topic in this section is **conditional frequency**, which is the relative frequency of a particular cell (or cells) compared to its row only or column only. These can also be considered as conditional probabilities, so they are often show in percent form.

Example: Use the previous set of data showing the breakdown of students at a private college based on class level and scholarship level. What percent of the freshmen got a full scholarship? What is the probability of a full scholarship recipient being an upperclassman (juniors or seniors)?

Solution: We look at the frequency (count) in the cell for freshman/full scholarship and divide that by the marginal frequency of all freshmen. This gives $\frac{38}{312} \times 100\% = 12.2\%$. For the second part, we take the sum of the two cells for full scholarship juniors and full scholarship seniors and divide that by the marginal total of all full scholarship recipients. $\frac{12+6}{75} \times 100\% = 24.0\%$

****Try this on your own**: The table below shows the breakdown of coronavirus cases in the USA, based on age group and sex. Calculate the marginal frequencies and the grand total. Then calculate the conditional frequency of age 60+ who are female rounded to tenth of a percent. The data comes from the CDC as of May 30th, 2020.

	age 0-19	20-39	40-59	60+
Male	35,045	194,232	228,897	188,184
Female	34,658	203,086	226,016	210,370

5.2.3 Exercises: Joint Distributions

Solutions appear at the end of this textbook.

1. The table below shows the breakdown of employed nurses by type and race they identify with. The data comes from the US Department of Health and Human Services and are rounded to nearest 100. Calculate the marginal frequencies and the grand total.

	RN	LPN
Asian	234,600	24,800
African American	279,600	162,800
Hispanic/Latino	135,600	51,800
Native American	11,300	4,100
White	2,164,100	446,500

2. The table below shows the breakdown of 2020-2021 UWG students by status and level. Calculate the marginal frequencies and the grand total, then rewrite the chart with relative frequencies.

	Full-time	Part-time
Undergraduate	7,528	2,803
Graduate	652	2,436

3. The table below shows partial data for the breakdown of US adults by group and opinion about guns. Fill in the missing frequencies.

	Married Men	Unmarried Men	Married Women	Unmarried Women	totals
Protect gun rights	568	396		332	1,744
Have more gun control		311	527	565	
totals	887		975	897	

4. The table below shows the breakdown of US adults by sex and body condition, from a 2017-1018 study by the CDC national center for Health Statistics. Find the conditional frequencies for the percent of men who are obese, and then the percent of severely obese people who are women.

	Healthy Weight	Overweight	Obese	Severely Obese
Men	906	1,572	2,171	471
Women	832	1,499	2,344	651

5. A University collected data on their academic faculty in the categories of highest degree and job rank. Some faculty have a Master's degree and other have a Doctoral (PhD) degree. The job ranks are Instructor, Lecturer, Assistant Professor, and Professor. There are 120 Instructors in total. If the probability that an Instructor has a Master's degree is 90%, how many Instructors have a Master's degree?

6. A University collected data on their academic faculty in the categories of highest degree and job rank. Some faculty have a Master's degree and other have a Doctoral (PhD) degree. The job ranks are Instructor, Lecturer, Assistant Professor, and Professor. There are 8 Lecturers who have a Doctoral (PhD) degree. If the probability that a Lecturer has a Doctoral (PhD) degree is 16%, how many Lecturers are there in total?

Chapter 6

Solutions

Answers to Try This On Your Own Problems

Section 1.1: For the following scenario, describe the population of interest, describe the sample, state the parameter of interest, and the statistic that was calculated. A farm wants to track the weight gain of their chickens after they switched to a new feed. The farm has over 10,000 chickens. They isolated 200 chickens and weighed them before the switch, then every week for the next 10 weeks. At the end of 10 weeks, the 200 isolated chickens gained an average of 1.2 pounds. **ANSWERS: Population is all 10,000 chickens, sample is the 200 isolated chickens, parameter is weight gain, the statistic calculated is 1.2 average weigh gain**

Would the following sample be representative of the population? A teacher would like to know how students feel about the new math curriculum. They selected a sample of students from the ones who are failing the class and come for extra help. **ANSWER: Not representative, since they leave out students who are doing well.**

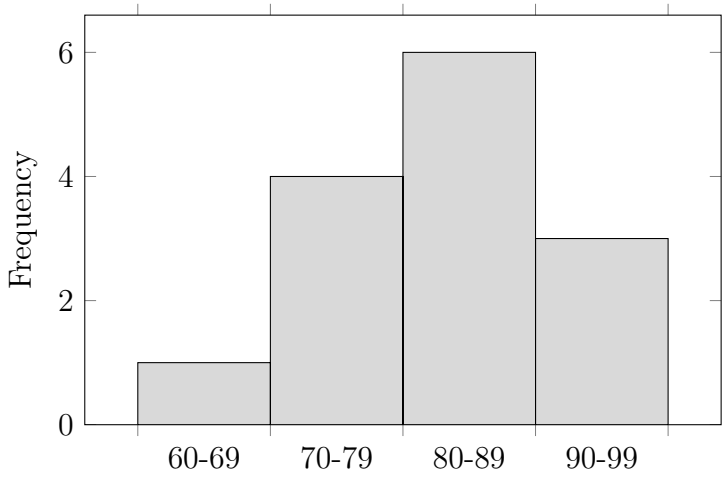
What sampling methods would each description below be classified as?

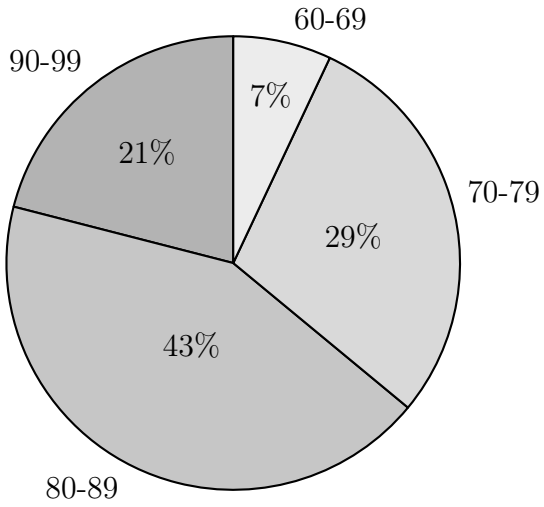
1. A teacher selected a sample of students by selecting one row and picking all students in that row. **ANSWER: Cluster sampling**

2. A researcher was conducting a survey where they selected a sample by going to every tenth neighborhood and surveying every tenth home from those neighborhoods. **ANSWER: Systematic sampling**

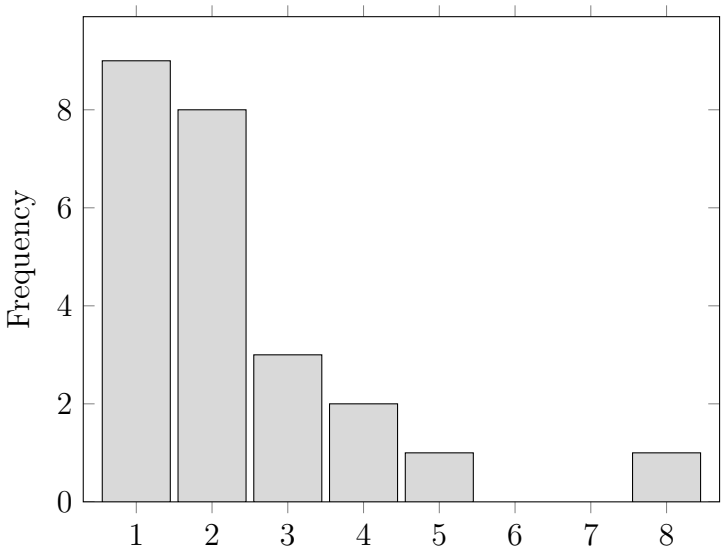
Section 1.2: The grades on a science final exam were 75, 83, 96, 82, 90, 78, 60, 76, 82, 71, 92, 86, 83, 88. Create a table with frequencies and relative frequencies using the intervals 60-69, 70-79, 80-89, 90-99. Then sketch a frequency histogram and a relative frequency pie chart. **ANSWERS:**

Grades	Frequency	Relative Frequency
60-69	1	7%
70-79	4	29%
80-89	6	43%
90-99	3	21%
Total	14	100%





What are the characteristics of the following graph? Examine the spread, symmetry, and outliers.



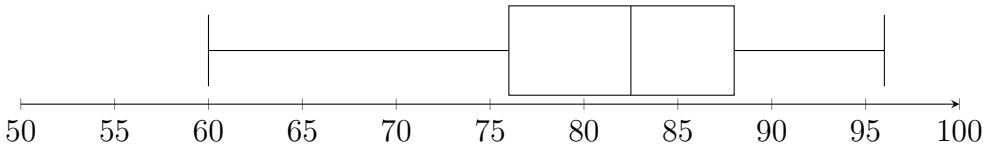
ANSWERS: The peak of the graph is at data values 1 and 2. The outlier is the one value at 8. It is not spread out very much, since most of the data is concentrated near left peak. The graph is right skewed.

Section 1.3: The grades from a sample of a science final exam were 75, 83, 96, 82, 90, 78, 60, 76, 82, 71, 92, 86, 83, 88. Calculate the mean, median, mode, range and standard

deviation. **ANSWERS: mean = 81.6, med = 82.5, mode = 82 and 83, range = 36, stddev = 9.3**

Section 1.4: Calculate the z-score of a woman who is 5 feet tall if the mean height is 65 inches and standard deviation is 3 inches. Is she unusually short or not? Round Z to two decimal places. **ANSWER:** $z = \frac{60-65}{3} = -1.67$**, she is a bit short but within the usual values of -2 to 2.**

The grades on a science final exam were 75, 83, 96, 82, 90, 78, 60, 76, 82, 71, 92, 86, 83, 88. Calculate the 5-number summary, IQR, and sketch a regular boxplot. **ANSWERS: min = 60, Q1 = 76, med = 82.5, Q3 = 88, max = 96, IQR = 12**



Section 2.1: In each situation below, calculate the probability, deciding whether to use empirical or theoretical probability.

1. 200 people are at a banquet and 8 people are at your table including you. What is the probability that someone at your table is chosen at random to win a prize out of the entire banquet? **ANSWER: Theoretical** $P(win) = \frac{8}{200} = 0.04 = 4\%$

2. Danny has played 20 tennis matches this season and has won 17 of them. What is the probability that he wins his next match? **ANSWER: Empirical** $P(win) = \frac{17}{20} = 0.85 = 85\%$

If a team has a 65% chance of winning, find the odds for and against winning.

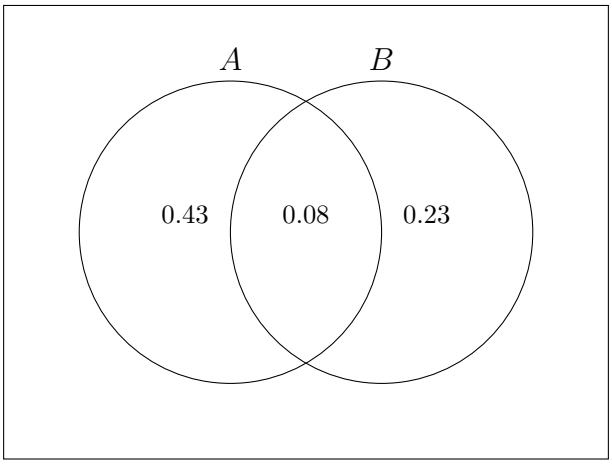
ANSWERS: P(not win)=35%. Odds for win $= \frac{65}{35} = \frac{13}{7}$ **= 13 to 7. Odds against win** $= \frac{35}{65}$ **= 7 to 13.**

Section 2.2: For a lottery in which you pick five numbers from 1 to 50, how many different sets can you pick if they can be in any order, and if they must be in a specific order? **ANSWERS:** ${}_{50}C_5 = 2,118,760$ **and** ${}_{50}P_5 = 254,251,200$

You and two friends entered a contest that will randomly pick 3 people to go to a concert. There are 200 people entered. What is the probability that the 3 winners are you and your two friends? **ANSWER:** $\dfrac{1}{{}_{200}C_3} = \dfrac{1}{1313400} = 0.000076\%$

Section 2.3: A football team has 42 players. There are 18 players who play offense, 20 players who play defense, and 10 players who play on special teams. Six of the offensive players play both offense and special teams. Find the probability that a player is on the offense or special teams. **ANSWER: P(O or S) = P(O) + P(S) - P(O and S)** $= \frac{22}{42} = 0.52 = 52\%$

Use the Venn diagram below to calculate $P(B)$, as well as the probability of neither A nor B, $P(\overline{A \text{ or } B})$.



ANSWERS: $P(B) = 0.08 + 0.23 = 0.31$, $P(\overline{A \text{ or } B}) = 1 - 0.43 - 0.08 - 0.23 = 0.26$

Section 3.1: Below is the distribution for the shoe size of the players on a college basketball team. Find the missing probability and then the mean and standard deviation of the shoe size. **ANSWERS: 0.124, 11.90, and 1.11**

shoe size	10	10.5	11	11.5	12	12.5	13	14
P(x)	.063	.063	.189	.124	.250	.124	.063	?

A particular game has the prize distribution shown below. Find the expected value of a prize. **ANSWER: $19**

Prize Amount	$0	$25	$100	$500
Probability	0.7	0.2	0.09	0.01

According to a survey of US adults done by the Pew research center in 2021, 85% of adults own a smartphone. For a random sample of 50 adults, find the mean, standard deviation and the probability that 45 of the 50 adults own a smartphone.
ANSWERS: $\mu = np = 42.5$**,** $\sigma = \sqrt{npq} = 2.52$**, and** $P(X = 45) = 0.107 = 10.7\%$

Section 3.2: A computer will pick random numbers out to ten decimal places. The numbers can be between 1 and 5.5. Find the probability of picking a number between 1.4 and 2.8. **ANSWER: 0.311 or 31.1%**

The birth weights of babies in South America are normally distributed, with a mean of $\mu = 3,100$ grams and a standard deviation of $\sigma = 400$ grams. Use the 68-95-99.7% rule to find the percentage of babies born with a weight more than $2,300$ grams. **ANSWER:** $47.5 + 50 = 97.5\%$

Compute $P(-0.5 < Z < 1.35)$ **ANSWER: 0.603 or 60.3%**

The birth weights of babies in Brazil are normally distributed, with a mean of $\mu = 3,110$ grams and a standard deviation of $\sigma = 463$ grams. Find the probability of a baby being born with a weight more than $3,000$ grams. **ANSWER: 0.595 or 59.5%**

Section 3.3: The birth weights of babies in Brazil are normally distributed, with a mean of $\mu = 3,110$ grams and a standard deviation of $\sigma = 463$ grams. Find the probability of a sample of 25 babies being born with an average weight between $3,200$ and $3,400$ grams. **ANSWER: 0.165 or 16.5%**

Section 4.1: A 2012 study with 543 participants with type 1 diabetes, had their average life expectancy to be 69 years with a standard deviation of 21 years. Find the 90% confidence interval to estimate the average life expectancy of all people with type 1 diabetes. **ANSWER: 67.5 to 70.5**

A company sampled their new product and found that 4 out of 80 items were defective. Find the 95% confidence interval to estimate the proportion of items that might be defective for the entire product line. **ANSWER: 0.002 to 0.098 or 0.2% to 9.8%**

Find the minimum required sample size to estimate the mean time that teenagers spend streaming video content per day, with a maximum error of 0.5 hours in a 90% confidence interval. Use the assumed value $\sigma = 2.8$ hours. **ANSWER: 85 people**

Section 4.2: The mean body temperature of healthy adults is typically reported as 98.6° F with standard deviation of 0.7° F. A nurse at a clinic has experienced several lower body temperatures and wants to test the hypothesis that the mean body temperature of healthy adults is less than 98.6° F. She records temperatures from a sample of 50 people and obtains a sample mean of 98.0° F. Perform the hypothesis test using the p-value method with a significance level of $\alpha = 0.02$. ANSWERS: $H_o : \mu = 98.6$, $H_a : \mu < 98.6$, $Z_{data} = \frac{98-98.6}{\frac{0.7}{\sqrt{50}}} = -6.06$, $p - value = normalcdf(-99999, -6.06, 0, 1) = 0.00000$, reject H_o. The p-value exists but is so small, it is effectively zero.

Based on a 2013 research study, the FDA has proposed a maximum level of nicotine in cigarettes of 0.5 milligrams. A tobacco company attempts to make a new cigarette below the proposed level. A health researcher doubts their ability and thinks they probably have more

than the 0.5 level. They test a sample of 40 cigarettes. The nicotine levels had $\overline{x} = 0.51$, $s = 0.04$. Perform a hypothesis test at the $\alpha = 0.05$ significance level using the p-value method. ANSWERS: $H_o : \mu = 0.50$, $H_a : \mu > 0.50$, $df = 39$, $t_{data} = \frac{0.51-0.50}{\frac{0.04}{\sqrt{40}}} = 1.581$, $p - value = normalcdf(1.581, 99999, 39) = 0.0610$, do not reject H_o. It is plausible their cigarettes have 0.5 mg of nicotine, even though the sample has more. This could easily be a random fluctuation in the sampling distribution.

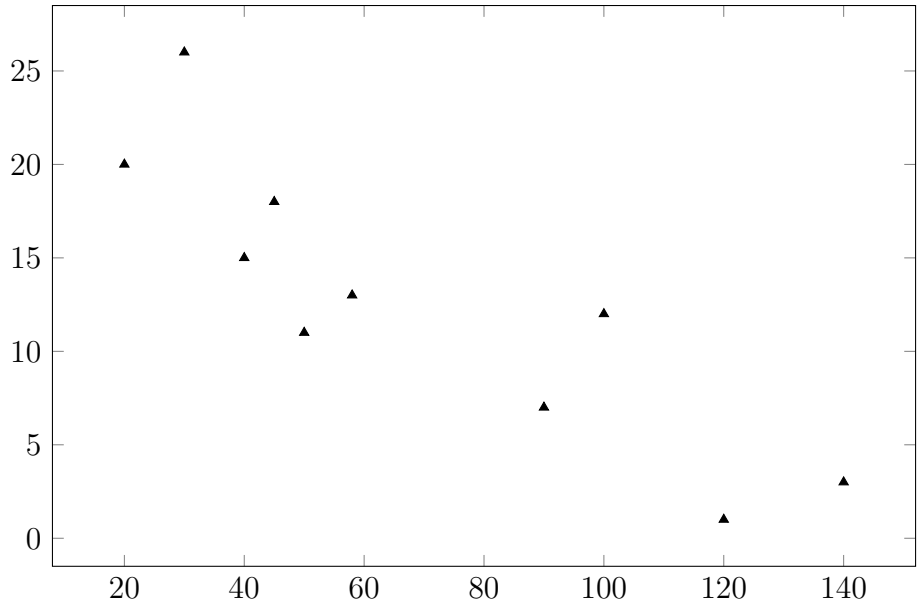
According to a survey by the Federal Reserve 12% of the 11,000 US investors surveyed said they have invested in some type of cryptocurrency in 2021. In March 2022 Quinnipiac University did a survey of 1,936 adults and 16% said they own cryptocurrency. Would that be enough evidence to suggest that the proportion has increased? Setup and perform a hypothesis test at the $\alpha = 0.05$ significance level using the p-value method. ANSWERS: $H_o : p = 0.12$, $H_a : p > 0.12$, $df = 1935$, $\hat{p} = 0.16$ $Z_{data} = \frac{0.16-0.12}{\sqrt{\frac{0.12(1-0.12)}{1936}}} = 5.42$
p-value = normalcdf(5.42,99999,0,1) = 0.00000, reject H_o the percentage has likely increased.

Section 5.1: The table below shows data from ten people of their average monthly spending on fast food, as well as their average number of days of exercise each month. Create a scatterplot and find the correlation coefficient r. Then find the regression equation and use it to forecast the number of days of exercise output for the fast food value \$70. Is that prediction interpolation or extrapolation? Do you notice a pattern between fast food spending and exercise? What could explain the pattern?

fast food \$	20	40	58	50	140	30	90	45	100	120
exercise days	20	15	13	11	3	26	7	18	12	1

ANSWER: The scatterplot is below. Overall there seems to be a mildly strong linear pattern in a negative direction. This is confirmed by $r = -0.879$**. The regression equation is** $y = -0.166x + 24.1$**. The prediction is** $y = -0.166(70) + 24.1 =$ 12.5 **days of exercise in a month with** \$70 **fast food spending. This is interpolation, since** 70 **is in the range of the data (20 to 140). It seems that as fast food**

spending increases, the number of days of exercise goes down. This could be because people who eat a lot of fast food are not as health conscious as those who eat better, and they would then not exercise as much on average.



Section 5.2: The table below shows the breakdown of coronavirus cases in the USA, based on age group and sex. Calculate the marginal frequencies and the grand total. Then calculate the conditional frequency of age 60+ who are female rounded to tenth of a percent. The data comes from the CDC as of May 30th, 2020. **ANSWER: The complete frequency table is shown below. The percentage of the age 60+ who are female is** $\frac{210370}{398554} \times 100\% = 52.8\%$

	age 0-19	20-39	40-59	60+	Totals
Male	35,045	194,232	228,897	188,184	646,358
Female	34,658	203,086	226,016	210,370	674,130
Totals	69,703	397,318	454,913	398,554	1,320,488

Solutions to Exercises: Sec 1.1 Collecting Data

1. The population is all homeschool science textbooks in the United States. The sample is the 15 science books obtained. The (unknown) parameter is the population average price of all the books, which they hope to determine. The statistic measured is the average price of the 15 books = \$52.

2. The variables are: name, height, weight, eye color, hair color, and page-hits. The corresponding values are: Sean Higgins, 5ft.10in., 185 lbs., Green, Red, and 142. Name, eye color, and hair color, are qualitative (categories). Height, weight, and number of page-hits are quantitative (measures or counts).

3. Step 1: $m = \frac{6700}{7} = 957.14$, round DOWN to $m = 957$. Step2: start with # 957 Step 3: keep adding m to list the place values of the sample selections. Sample is the set of people in places 957, 1914, 2871, 3828, 4785, 5742, 6699.

4. This is an observational study, since she just observed the crabs and their time. She did not impose treatments or try to control anything.

5. This is stratified sampling, since the items are grouped (strata) and a few from each group are selected. This is not cluster. In cluster, the items are grouped, but entire groups are used.

6. A census is a gathering of information from the entire population of people or things that is being studied. The US government only does a census every ten years, because it takes so much time, money, and staff to complete. Technically, the US census is not a true complete census. It is impossible to keep track of everyone in the country at one exact moment in time. There are people being born and dying every day, criminals or illegal aliens hiding who don't want to be found, and some people who ignore requests or lie about their information.

7. No method is 100% bias free. Simple random is not biased in how it selects, but the sample you get could be biased. Convenience is biased, since it leaves out most of the population. Systematic is biased, because it does not allow most of the combinations to be picked. Cluster can be biased, if the people/items that are together have the same characteristics. Stratified tends to have the least amount of bias. It selects some of each type and the sample is representative of the population.

8. Experimenters use placebos to prevent psychological effects and bias from the experimental units. People who know which treatment they get, can change their stress level and affect the results.

9. Which samples are representative of their populations, which are not? Explain why.

   (a) Not representative since the people watching the top ten youtube videos will likely watch more than most teenagers. They need to include teenagers who do a variety of activities, have jobs, etc.

   (b) Yes representative since they pick many different types of employees from several different locations.

Solutions to Exercises: Sec 1.2 Summarizing Data

1. Since we are trying to summarize, a small number of groups makes it easy to see the big picture. If a data set has 1000 values, using 100 groups would be so large and cumbersome, it would not be a summary and difficult to see anything.

2. When computing relative frequencies, the total should equal 100% or very close. 100% means all the data has been accounted for.

3. The values are shown in the chart below, the class width is 23.

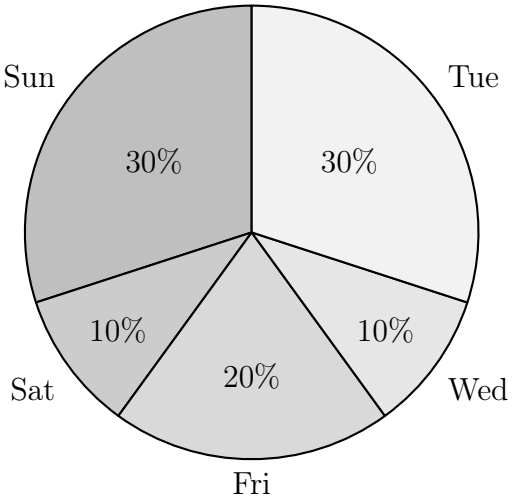
class	1-23	24-46	47-69	70-92	93-115
lower limit	1	24	47	70	93
upper limit	23	46	69	92	115
midpoint	12	35	58	81	104
boundary		23.5	46.5	69.5	92,5

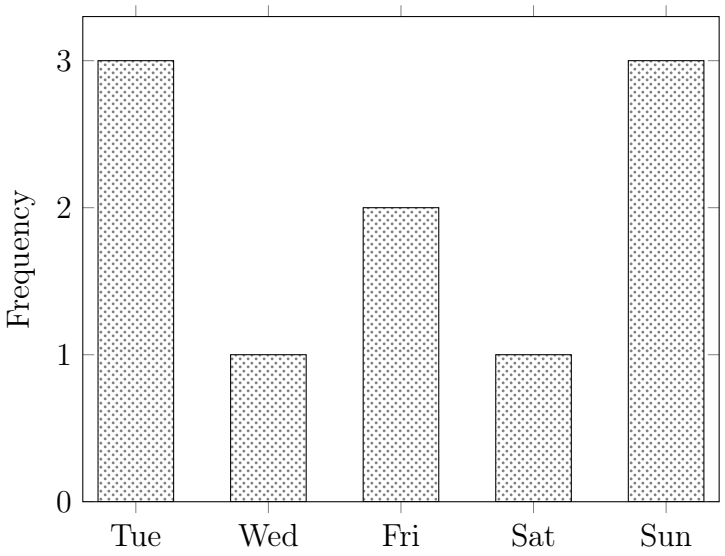
4. A bar graph can be in any order, but a Pareto chart has the bars shown in size order. A Pareto chart cannot be done from quantitative data, since numbers must go in numerical order of the classes (intervals) and the graph is a histogram.

5. It rained ten days in this month. The distribution of which days it rained would be as follows:

Day	Frequency	Relative Frequency
Tuesday	3	$\frac{3}{10} = 0.3 = 30\%$
Wednesday	1	$\frac{1}{10} = 0.1 = 10\%$
Friday	2	$\frac{2}{10} = 0.2 = 20\%$
Saturday	1	$\frac{1}{10} = 0.1 = 10\%$
Sunday	3	$\frac{3}{10} = 0.3 = 30\%$
Total	10	100%

6. The pie chart and bar graph for which days it rained are:

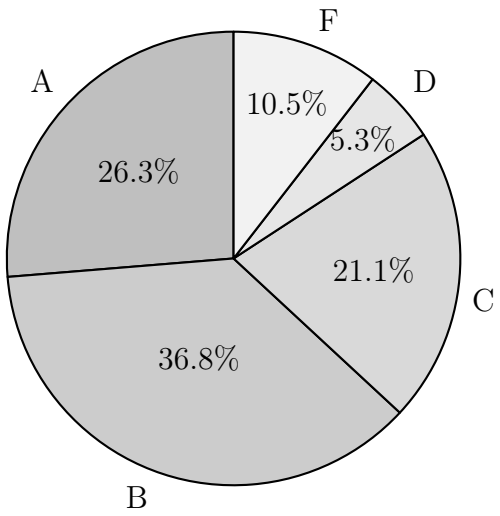


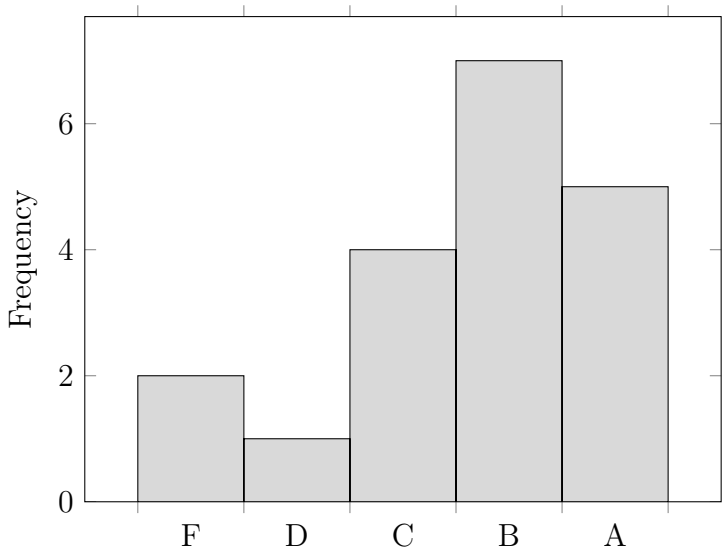


7. For the classes, we can use the common grading scale F (<60), D (60-69), C (70-79), B (80-89), and A (90+). If your school has a different scale, that would be fine also. The distribution is shown below.

Grade	Frequency	Relative Frequency
F (<60)	2	$\frac{2}{19} = 0.10526 = 10.5\%$
D (60-69)	1	$\frac{1}{19} = 0.05263 = 5.3\%$
C (70-79)	4	$\frac{4}{19} = 0.21053 = 21.1\%$
B (80-89)	7	$\frac{7}{19} = 0.36842 = 36.8\%$
A (90+)	5	$\frac{5}{19} = 0.26316 = 26.3\%$
Total	19	100.0%

8. The pie chart and histogram for the history grades are:





9. For the cumulative frequencies, simply add all of the frequencies up through that class. For the missing relative frequencies, divide the frequency by 50. For the cumulative relative frequencies, add all of the relative frequencies up through that class.

Rate %	Frequency	Cumulative Frequency	Relative Frequency	Cum. Rel. Frequency
50-54.9	9	9	0.18	0.18
55-59.9	8	17	0.16	0.34
60-64.9	11	28	0.22	0.56
65-69.9	7	35	0.14	0.70
70-74.9	6	41	0.12	0.82
75-79.9	7	48	0.14	0.96
80-84.9	2	50	0.04	1.00
Total	50		1.00	

10. The center is at 6. The graph is not spread out that much, it is concentrated from 5-8, with a single outlier at 1. The graph is left-skewed semi-bell shape.

11. The graph has a truncated vertical axis, making the first bar appear much smaller and the third bar much larger than the others. The first bar height is around 50, the second bar around 60, not really that much of a difference. The second bar appears to be twice the size of the first, this is misleading. Also there is no title and the categories are vague. This is a very bad graph.

Solutions to Exercises: Sec 1.3 Measuring Data

1. Compute the mean, median, and mode of the following data sets. Use the round-off rule.

   (a) median = 4, mean = 4, there is no mode.

   (b) in order: -2, 0, 0, 1, 3, 4, median = 0.5, mean = 1, mode = 0.

   (c) in order: 103, 123, 155, 172, 188, 195, 222, 230, 253, 281, 301, 318
   median = 208.5, mean = 221.8, there is no mode.

2. Mean $= \frac{\sum x}{N} = \frac{1527}{19} = 80.3684$, rounded is 80.4. After data is put in order, the median is the 10th value, 82. There are three values which are repeated twice, so the three modes are: 75, 82, and 85.

3. The min is \$44, the max is \$116, so the range is $116 - 44 = \$72$. We can setup a table to help organize the calculations. The mean is 82.

Price (x)	$x - \overline{x}$	$(x - \overline{x})^2$
\$44	$44 - 82 = -38$	$(-38)^2 = 1444$
\$74	$74 - 82 = -8$	$(-8)^2 = 64$
\$94	$94 - 82 = 12$	$12^2 = 144$
\$116	$116 - 82 = 34$	$34^2 = 1156$
Sum		2808

   The variance $s^2 = \dfrac{\sum (x - \overline{x})^2}{n - 1} = \dfrac{2808}{4 - 1} = 936$.
   The standard deviation $s = \sqrt{936} = 30.59412$. Using round-off rule, $s = \$30.6$.

4. Notice some are more than \$30.6 away from the average of \$82, and others are less. The ticket prices are spread out by \$30.60 on average away from the mean \$82. This means that the ticket prices vary quite a bit.

5. The median and mode are usually unaffected by extreme values, since the median is in the middle (not at extremes) and the extreme values usually don't occur often. The mean is found from the sum of all values, one extreme value can affect the mean drastically.

6. The two formulas are almost the same, but the sample formula is divided by $n-1$ instead of just n. The sample standard deviation is always larger than the population standard deviation, because its denominator is less, making the fraction more.

7. $GPA = \dfrac{3.0(3)+3.0(4)+4.0(2)+2.0(3)+4.0(3)}{3+4+2+3+3} = \dfrac{47}{15} = 3.13$, which is a low B, not bad for a first semester in college.

8. Since the mean is greater than the median, this data set is probably right-skewed. The top half of the weights are spread far out into very large values. This matches with the large standard deviation. There must be some women well more than 45 pounds over average.

9. The midpoints of the classes ($\widehat{x}$) are: $12, 17, 22, 27, 32, 37$. Then the formula would be $\frac{\sum \widehat{x}f}{\sum f} = \frac{12(12)+17(5)+22(7)+27(2)+32(6)+37(3)}{12+5+7+2+6+3} = \frac{740}{35} = 21.14286$, rounded to 21.1.

10. For the shot put, $CV = \dfrac{5.5}{38} \times 100\% = 14.5\%$.
    For the gymnastics, $CV = \dfrac{1.4}{8.45} \times 100\% = 16.6\%$.
    The gymnastics scores are a more spread out set of data than the shot put throws.

Solutions to Exercises: Sec 1.4 Measuring of Relative Standing

1. The z-scores are shown in the table below.

Name	IQ score	z-score
Garry Kasporov	190	$z = \frac{190-100}{15} = +6.00$
Albert Einstein	160	$z = \frac{160-100}{15} = +4.00$
Arnold Schwarzenegger	135	$z = \frac{135-100}{15} = +2.33$
Tim Tebow	104	$z = \frac{104-100}{15} = +0.27$
Howard Stern	99	$z = \frac{99-100}{15} = -0.07$
George W Bush	125	$z = \frac{125-100}{15} = +1.67$
Muhammad Ali	78	$z = \frac{78-100}{15} = -1.47$
Barack Obama	130	$z = \frac{130-100}{15} = +2.00$

   Usual values are between -2 and $+2$, so the only ones that are unusual are Kasporov, Einstein, and Schwarzenegger. Note: Kasporov is way off the charts, among the smartest humans ever.

2. Unusually short would be a z-score below -2, so we can setup the formula and solve for $x < 63$ inches (or 5ft 3in).

$$z = \frac{x-\mu}{\sigma}$$
$$-2 = \frac{x-69}{3}$$
$$3(-2) = \left(\frac{x-69}{\cancel{3}}\right)\cancel{3}$$
$$-6 = x - 69$$
$$69 - 6 = x$$

Unusually tall would be a z-score above +2, so we can setup the formula and solve for $x > 75$ inches (or 6ft 3in).

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ 2 &= \frac{x-69}{3} \\ 3(2) &= \left(\frac{x-69}{\cancel{3}}\right)\cancel{3} \\ 6 &= x-69 \\ 69+6 &= x \end{aligned}$$

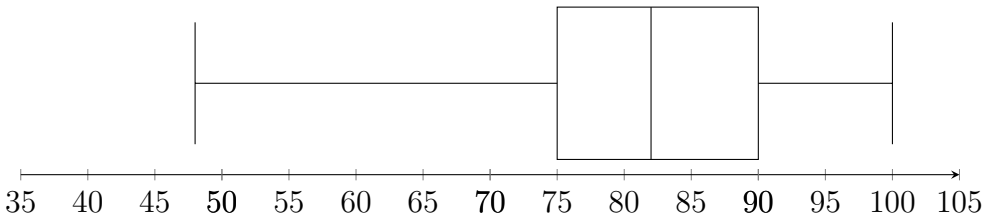
3. $L = 38\left(\frac{45}{100}\right) = 17.1$, so the percentile is the data value in the 17th position. We don't have the data set here, so we cannot state the actual value, it would just be whatever value is in that location of 38 values put in order.

4. It means that their score on the test was greater than (or equal to) 85% of all the scores. They beat 85% of the test takers.

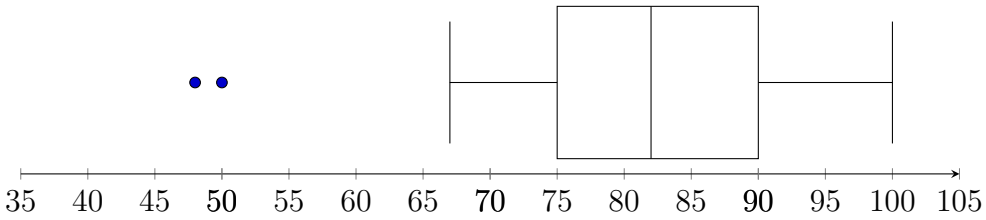
5. First put the data values in size order. The 6th and 7th values are tied for the middle, so median is their average 8.5. Q_1 is the median of the first 6 values and Q_3 is the median of the last 6 values. The five-number summary is $\text{Min} = 5, Q_1 = 7.5, \text{Med} = 8.5, Q_3 = 14.5, \text{Max} = 20$.

6. First put the data values in size order. The tenth value is in the middle (median). Q_1 is the median of the first 9 values and Q_3 is the median of the last 9 values. The five-number summary is $\text{Min} = 48, Q_1 = 75, \text{Med} = 82, Q_3 = 90, \text{Max} = 100$. The class did relatively well. $Q_1 = 75$ means only 25% of the class got below 75, the median of 82 means half the class scored above 82, $Q_3 = 90$ means 25% of the class got above 90.

7. The boxplot looks like:



8. $IQR = 90 - 75 = 15$, so $LF = 75 - 1.5(15) = 52.5$ and $UF = 90 + 1.5(15) = 112.5$. There are two outliers (48 and 50) on the low end (below 52.5), but no outliers on the high end (above 112.5). Then modified boxplot looks like:



9. Group C is the most spread out, the box is wider and the whiskers extend out farther that the other boxplots. Group C is symmetric, Group B is left-skewed, and Group A is right-skewed.

Solutions to Exercises: Sec 2.1 Probability Basics

1. List the sample spaces for the following experiments:

   (a) SS = { H, T }, $N = 2$.

   (b) We need to look at what happens for each flip and combine them.

   SS = { HHH, HHT, HTH, HTT, THH, THT, TTH, TTT }, $N = 8$.

   (c) $SS = \{1, 2, 3, 4, 5, 6\}$, $N = 6$.

   (d) The sums range from 2 (rolling a 1+1) up to 12 (rolling 6+6)

   $SS = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$, $N = 11$.

   (e) Using letter abbreviations for the 7 rainbow colors and seasons,

   SS = { Rw, Ow, Yw, Bw, Gw, Iw, Vw, Rsp, ..., Rsm, ..., Rf, Of, Yf, Bf, Gf, If, Vf }, $N = 28$.

2. Probabilities must be between 0 and 1 (or 0% and 100%), so the valid values are 0.35, 0.004, and $\frac{3}{8}$. All of the others are either negative or greater than 1.

3. The event 'even' consists of the outcomes 2, 4, and 6. This is three out of the 6 possible outcomes, so $P(even) = \frac{3}{6} = 0.5 = 50\%$. $P(3) = \frac{1}{6} = 0.167 = 17\%$. P(>2) = P(3 or 4 or 5 or 6) = $\frac{4}{6} = 0.667 = 67\%$.

4. Probability of making a shot, based on his data, is $\frac{5}{12} = 0.417 = 42\%$. Subjectively, you might believe this to be low percentage and he might not make the team.

5. The probability of precipitation equals 0.45 (the sum of all three given). By the complement rule, the probability of no precipitation = 1 - probability of precipitation $= 1 - 0.45 = 0.55 = 55\%$.

6. For a standard deck of playing cards, find the following probabilities for picking one card at random.

   a) $\text{P(heart)} = \frac{13}{52} = \frac{1}{4}$ b) $\text{P(ace)} = \frac{4}{52} = \frac{1}{13}$

c) $\text{P(red)} = \dfrac{26}{52} = \dfrac{1}{2}$

d) $\text{P(\# from 2 to 10)} = \dfrac{36}{52} = \dfrac{9}{13}$

e) $\text{P(black king)} = \dfrac{2}{52} = \dfrac{1}{26}$

f) P(red club) = 0 the clubs are black

7. The probability of no rain today is $100\% - 20\% = 80\%$. Odds in favor $= \dfrac{P(rain)}{P(no\ rain)} = \dfrac{20}{80} = \dfrac{1}{4}$ or 1 to 4 odds in favor. Odds against rain $= \dfrac{P(no\ rain)}{P(rain)} = \dfrac{80}{20} = \dfrac{4}{1}$ or 4 to 1 odds against.

8. Given odds against, we can think of the numbers $9 : 5$ as $B : A$. Out of 14 games, the team will not win 9 of them, but will win the other 5. The probability that they win is $P(\text{win}) = \dfrac{A}{A+B} = \dfrac{5}{5+9} = \dfrac{5}{14} = 35.7\%$. The probability that they do not win is $P(\text{not win}) = \dfrac{B}{A+B} = \dfrac{9}{5+9} = \dfrac{9}{14} = 64.3\%$.

9. The theoretical probabilities of rolling each number are all equal to $\frac{1}{6}$ or about 17%. Your empirical probabilities are computed by dividing the count of how often a number was rolled, by the total of 15. Your values most likely vary, some smaller than 17% and some larger. If they are close to the theoretical, you got lucky. If they are not, that is because your rolls are random and with only 15 rolls, the law of large numbers does not work very well.

Solutions to Exercises: Sec 2.2 Counting Rules

1. Using the Fundamental Counting Principle, the number of combo meals is $5 \times 4 \times 3 = 60$.

2. The number of different outfits is $12 \times 6 \times 8 \times 3 = 1,728$

3. ${}_8C_3 = \frac{8!}{3!(8-3)!} = \frac{8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1(5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)} = \frac{8 \cdot 7 \cdot 6}{3 \cdot 2 \cdot 1} = 56$
   ${}_{11}C_9 = \frac{11!}{9!(11-9)!} = \frac{11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1(2 \cdot 1)} = \frac{11 \cdot 10}{2 \cdot 1} = 55$
   ${}_7P_4 = \frac{7!}{(7-4)!} = \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1} = 7 \cdot 6 \cdot 5 \cdot 4 = 840$
   ${}_8P_8 = \frac{8!}{(8-8)!} = \frac{8!}{0!} = \frac{8!}{1} = 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 40,320$
   ${}_5C_1 = \frac{5!}{1!(5-1)!} = \frac{5!}{4!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{4 \cdot 3 \cdot 2 \cdot 1} = 5$

4. If we consider sequences of the 3 movies, that implies a specific order, so we want permutation here. ${}_{10}P_3 = \frac{10!}{(10-3)!} = \frac{10!}{7!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 10 \cdot 9 \cdot 8 = 720$

5. Since we are just looking at the numbers selected, and not in any order, this is a combination. ${}_{39}C_5 = \frac{39!}{5!(39-5)!} = \frac{39 \cdot 38 \cdot 37 \cdot 36 \cdot 35}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = \frac{69090840}{120} = 575,757$ different sets of 5 numbers to play.

6. The probability of winning the Fantasy 5 lottery jackpot (matching all 5 numbers) is $\frac{1}{{}_{39}C_5} = \frac{1}{575757} = 0.00000174$, pretty small chance. There is one jackpot winning set of numbers and 575,756 losing sets, so the odds against winning are 575,756 to 1, very much stacked against you!

7. There are 999 numbers from 1 to 999. Only one is the exact match. $P(not - win) = \frac{998}{999} = 99.8999\%$

8. It is impossible to compute ${}_5C_9$, since this means to select 9 items from a group of 5. This makes no sense. Also the formula would have a negative value, and factorial is only for positive values.

9. We need to get a count of how many combinations with 3 matching numbers, and how many combinations there are in total, then divide them. If we select 3 winning numbers, that means we also have selected 2 losing numbers. By the Fundamental Counting Principle, we multiply the number of ways to pick 3 out of 5 winning numbers, by the number of ways to pick 2 out of the 24 losing numbers. Each of these is a combination, so our numerator is ${}_5C_3 \times {}_{24}C_2$. The total number of sets of 5 is ${}_{29}C_5$. Now the probability is $\frac{{}_5C_3 \times {}_{24}C_2}{{}_{29}C_5} = \frac{10(276)}{118755} = 0.023 = 2.3\%$

10. By the Fundamental Counting Principle, we multiply the number of ways to pick 4 out of 4 aces, by the number of ways to pick 1 out of any other card. Each of these is a combination, so our numerator is ${}_4C_4 \times {}_{48}C_1$. The total number of sets of 5 is ${}_{52}C_5$. Now the probability is $\frac{{}_4C_4 \times {}_{48}C_1}{{}_{52}C_5} = \frac{1(48)}{2598960} = 0.00185\%$

Solutions to Exercises: Sec 2.3 More Probability

1. Mutually exclusive means that the events cannot happen at the same time. One example is rolling a 3 on a die and a 5 on a die. Another example would be the experiment picking a name for a raffle winner, with events picking a student and picking a teacher.

2. We add up all of the physics or engineering majors, but make sure not to double count the 14 double majors. There are $32+(112-14) = 130$ in that group. For total students we add all three majors without double counting, $32 + (49 - 8) + (112 - 14) = 171$. Now P(Phys or Eng)$= \frac{130}{171} = 0.760 = 76\%$.

3. Using addition rule, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = 0.5 + 0.7 - 0.3 = 0.9$.

4. Setting up the addition rule, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$, we get $0.85 = 0.65 + P(B) - 0.25$. By simplifying and solving, we get $P(B) = 0.45$

5. These are dependent events, the first card picked affects what cards are left and so affects the second pick. We use the multiplication rule $P(\text{red6 and red6}) = P(\text{1st red6}) \cdot P(\text{2nd red6} \mid \text{1st red6}) = \frac{2}{52} \cdot \frac{1}{51} = \frac{1}{1326} = 0.00075 = 0.075\%$, which is a very small chance.

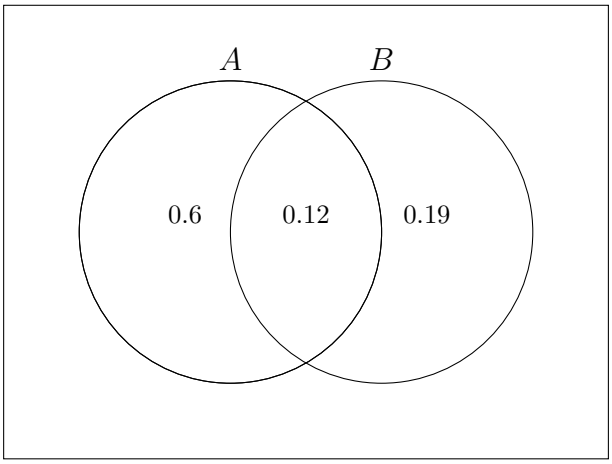
6. By reasoning, we know the ten of hearts is one out of 26 red cards. By formula, $P(\text{10hearts}|\text{red}) = \dfrac{P(\text{10hearts and red})}{P(red)} = \dfrac{\frac{1}{52}}{\frac{26}{52}} = \dfrac{1}{26} = 0.038 = 3.8\%$

7. One example is rolling two dice and getting a 3 on one die and a 5 on the other die. Another example is randomly picking a person and determining that they work at Home Depot and like vanilla ice cream. Where you work has no affect on what ice cream you like.

8. Two events can never be both mutually exclusive and independent. If they are mutually exclusive, the occurrence of one automatically affects the other (it prohibits the other). Once you know a coin lands heads up, then tails has no chance (until the next flip).

9. This is not a valid probability distribution. The probabilities themselves are valid, but the total only sums to 0.97 or 97%.

10. Draw two overlapping circles inside a rectangle. We place the probability value of $P(A \bigcap B) = 0.12$ inside the intersection piece in the center. Then event A has a total probability of 0.72, with 0.12 is already in circle A. The remaining probability $0.72 - 0.12 = 0.6$ goes inside the extra part of A (the crescent moon). The union $P(A \bigcup B) = 0.91$ which is all the circle parts added together. We can calculate the missing piece for the exclusive part of B as $0.91 - 0.6 - 0.12 = 0.19$ which goes on the outer part of B. Then the full probability $P(B) = 0.12 + 0.19 = 0.31$



Solutions to Exercises: Sec 3.1 Discrete Distributions

1. The distribution below shows the percentage of people who have received Covid-19 vaccines in Georgia as of June 16, 2022. Find the missing probability and the mode.

Vax dose	None	1 dose	2 doses	3+ doses
prob - P(x)	?	7%	33%	24%

The missing probability is 100% minus the values shown. P(none) = 100 - 7 - 33 - 24 = 36%, which is the greatest probability so the mode is zero vaccinations.

2. We add a cumulative row. The greatest probability is 0.38, so the mode is 4 years. The cumulative probability reaches 0.25 within the x-value of 4, so $Q_1 = 4$. The cumulative probability reaches 0.50 within the x-value of 5, so $med = 5$. The cumulative probability reaches 0.75 within the x-value of 6, so $Q_3 = 6$.

Years	4	5	6	7	8	9+
P(x)	0.38	0.26	0.12	0.07	0.03	0.14
cum P(x)	0.38	0.64	0.76	0.83	0.86	1.00

3. We add new rows for the weighted value of X and X^2, then $\mu = \sum x * P(x) = 16.4$, so the average age that a women first has sexual intercourse is during age 16. The standard deviation is $\sigma = \sqrt{(\sum x^2 * P(x)) - \mu^2} = \sqrt{274.89 - 16.4^2} = 2.4$.

Age	10	11	12	13	14	15	16	17	18	19	20+
P(x)	0.01	0.01	0.02	0.04	0.12	0.21	0.15	0.11	0.11	0.04	0.18
x*P(x)	0.1	0.11	0.24	0.52	1.68	3.15	2.40	1.87	1.98	0.76	3.60
x^2*P(x)	1	1.21	2.88	6.76	23.52	47.25	38.4	31.79	35.64	14.44	72

4. The new row shows the products of each $X * P(x)$. The expected value is $\sum X * P(x) = 7.1$. The expected prize is \$7.10 for each throw (on average).

Area	outer ring	middle ring	inner ring	bullseye
Prize X	\$0	\$5	\$10	\$50
prob - P(x)	0.42	0.30	0.21	0.07
X*P(x)	0	1.5	2.1	3.5

5. $E(x) = \frac{752{,}944}{800{,}000}(0) + \frac{45{,}000}{800{,}000}(5) + \frac{2{,}000}{800{,}000}(100) + \frac{50}{800{,}000}(2500) + \frac{5}{800{,}000}(20000) + \frac{1}{800{,}000}(100000) = \$0.94$, so the lottery expects to pay out \$0.94 on average for each play of the game. In order to make a profit, they need to charge more than that, maybe \$1.00 or \$2.00.

6. Calculate the binomial probabilities for the given values.

   a) $P(X = 8) = {}_{40}C_8\ (\frac{1}{4})^8(\frac{3}{4})^{32} = 0.118$

   b) $P(X = 8) = {}_{10}C_8\ (0.75)^8(0.25)^2 = 0.282$

   c) $P(X = 1) = {}_4C_1\ (\frac{1}{2})^1(\frac{1}{2})^3 = 0.25$

   d) $P(X = 0) = {}_{16}C_0\ (0.03)^0(0.97)^{16} = 0.614$

7. Here $n = 5$, $p = \frac{1}{6}$, $q = \frac{5}{6}$, and $X = 0, 1, 2, 3, 4, 5$. The probabilities are

   $P(X = 0) = {}_5C_0\ (\frac{1}{6})^0(\frac{5}{6})^5 = 1(1)(\frac{3125}{7776}) = 0.402$

   $P(X = 1) = {}_5C_1\ (\frac{1}{6})^1(\frac{5}{6})^4 = 5(\frac{1}{6})(\frac{625}{1296}) = 0.402$

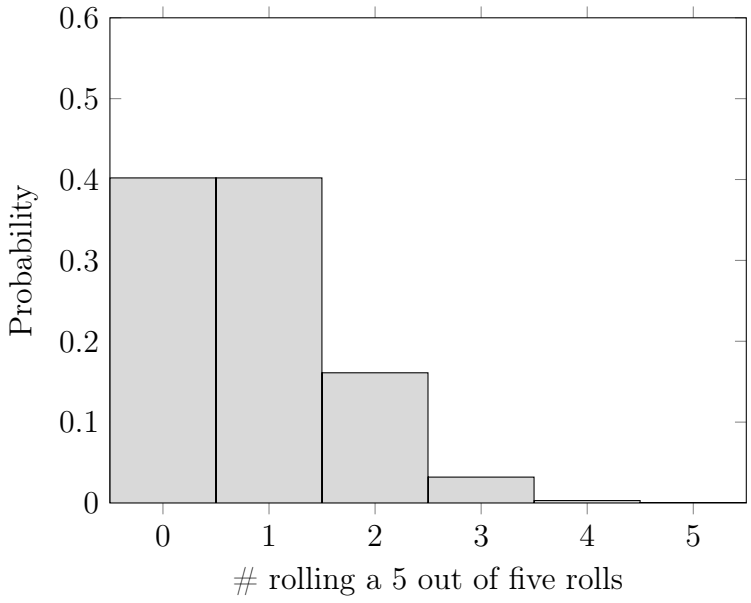
   $P(X = 2) = {}_5C_2\ (\frac{1}{6})^2(\frac{5}{6})^3 = 10(\frac{1}{36})(\frac{125}{216}) = 0.161$

   $P(X = 3) = {}_5C_3\ (\frac{1}{6})^3(\frac{5}{6})^2 = 10(\frac{1}{216})(\frac{25}{36}) = 0.032$

   $P(X = 4) = {}_5C_4\ (\frac{1}{6})^4(\frac{5}{6})^1 = 5(\frac{1}{1296})(\frac{5}{6}) = 0.003$

   $P(X = 5) = {}_5C_5\ (\frac{1}{6})^5(\frac{5}{6})^0 = 1(\frac{1}{7776})(1) = 0.0001$

8. Sketch the probability histogram for the previous exercise for the possible number of times a 5 shows on five rolls of a die.



9. For men, $\mu = np = 100 * (0.17) = 17$ men and $\sigma = \sqrt{npq} = \sqrt{100(0.17)(0.83} = 3.76$. For women, $\mu = np = 100 * (0.13) = 13$ women and $\sigma = \sqrt{npq} = \sqrt{100(0.13)(0.87} = 3.36$.

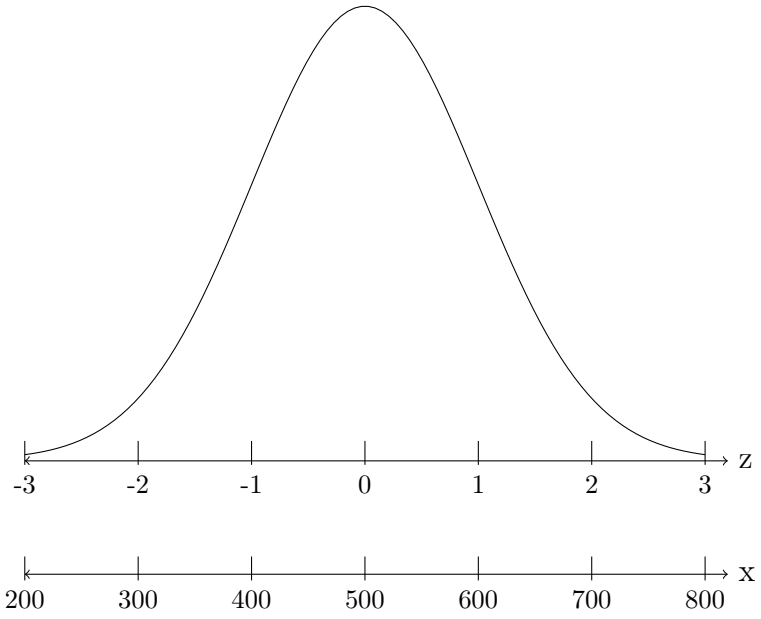
Solutions to Exercises: Sec 3.2 Continuous Distributions

1. We need to figure out what proportion the given interval is compared to the entire set. From 10 to 36 is $36 - 10 = 26$ units. The interval from 12 to 19.5 is $19.5 - 12 = 7.5$ units. Therefore, the probability is $P(12 < X < 19.5) = \dfrac{7.5}{26} = 0.288 = 28.8\%$

2. 4 a.m. to 7 a.m. is 3 hours out of the 24 hours of the day, so the probability is $\dfrac{3}{24} = \dfrac{1}{8}$, which is also 12.5%.

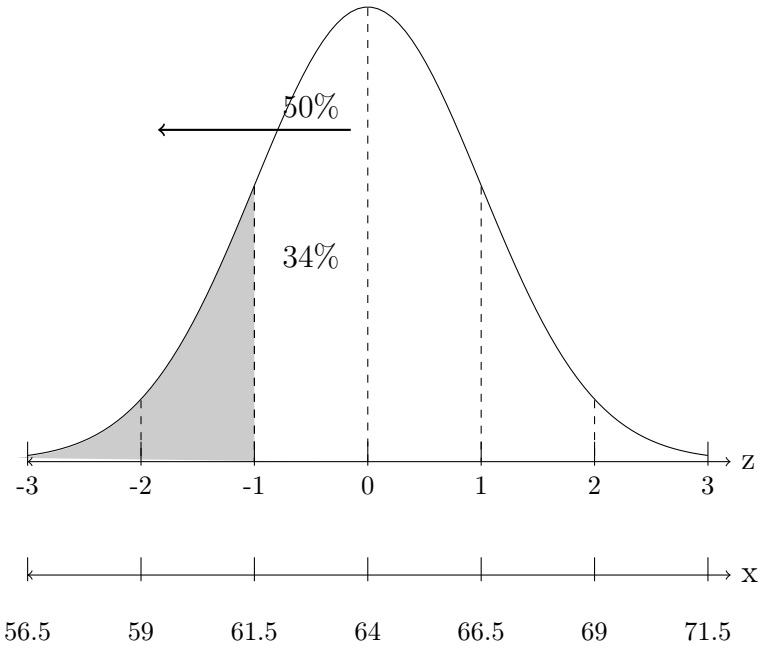
3. From $z = -3$ and $z = -2$ is 2.35% and from $z = -1$ and $z = -2$ is another 13.5%. Added together these result in $13.5 + 2.35 = 15.85\%$

4. SAT test scores for English are shown on the x-axis below the standard normal distribution bell curve.



5. Draw a bell curve with standard z-axis from -3 to 3 and below that, an x-axis with heights that correspond to the z marks. The mean height of 64 will go below $z = 0$, one standard deviation higher ($64 + 2.5 = 66.5$ inches) will go below $z = 1$, etc. Do similar process on left side, subtracting standard deviation to go under the negative z-values.

Change 5 feet 1-1/2 inches into 61.5 inches. So we are looking for the slice of the graph below 61.5, which is to the left $z = -1$. The area is the lower half (50%) minus the section from $z = -1$ to 0 of 34%. The answer is $50 - 34 = 16\%$. The graph is shown below.



6. Find the probabilities below of the standard normal z-score.

   a) Look on the Z table, second page (positive z-scores) and go down to the row for 2.7 and across to the column 0.03. There we find the area of 0.9968, which is our answer, $P(Z < 2.73) = 0.9968 = 99.68\%$.

   b) Look on the Z table, first page (negative z-scores) and go down to the row for -1.0 and across to the column 0.04. There we find the area of 0.1492, which is our answer, $P(Z < -1.04) = 0.1492 = 14.92\%$.

   c) Look on the first page (negative z-scores) and go down to the row for -0.4 and across to the column 0.05. There we find the area of 0.3264, which is the area to the left, but we want the area to the right. Therefore, $P(Z > -0.45) = 1 - P(Z < -0.45) = 1 - 0.3264 = 0.6736 = 67.36\%$.

d) Look on the Z table, second page (positive z-scores) and go down to the row for 1.2 and across to the column 0.00. There we find the area of 0.8894, which is the area to the left, but we want the area to the right. Therefore, $P(Z > 1.20) = 1 - P(Z < 1.20) = 1 - 0.8894 = 0.1106 = 11.06\%$.

e) Look on the first page (negative z-scores) and go down to the row for -1.4 and across to the first column 0.00. There we find the area of 0.0808, which is the area to the left of $z = -1.40$. Look on the second page (positive z-scores) and go down to the row for 1.4 and across to the first column under 0.00. There we find the area of 0.9192, which is the area to the left of $z = 1.40$. Now we want the area between these, so we subtract these areas. Therefore, $P(-1.40 < Z < 1.40) = P(Z < 1.40) - P(Z < -1.40) = 0.9192 - 0.0808 = 0.8384 = 83.84\%$.

f) Look on the second page (positive z-scores) and go down to the rows for 0.0 and 1.8, then respectively across to the columns under 0.00 and 0.05. There we find the areas of 0.6915 and 0.9678. Now we want the area between these, so we subtract these areas. Therefore, $P(0.5 < Z < 1.85) = P(Z < 1.85) - P(Z < 0.5) = 0.9678 - 0.6915 = 0.2763 = 27.63\%$.

7. a) We want the z, such that $P(Z < z) = 0.0129$. A small area to left will be a negative z-score, so look on first page of table to find the area .0129 in the body of the table. We find it in row for -2.2 and column for 0.03. So $z = -2.23$ to match $P(Z < -2.23) = 0.0129$.

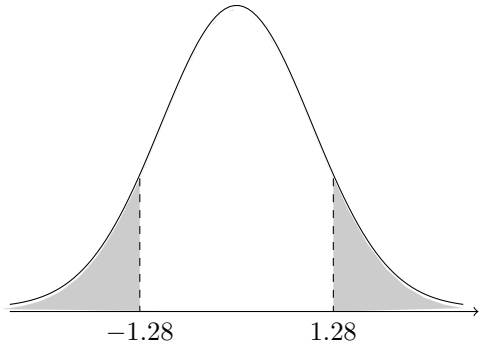
b) We want the z, such that $P(Z > z) = 0.6950$. A large area to left will be a positive z-score, so look on second page of table to find the area .6950 in the body of the table. We find it in row for 0.5 and column for 0.01. So $z = 0.51$ to match $P(Z < 0.51) = 0.6950$.

c) Since the table shows areas to the left (below), and our given area is to the right (above), we need to use the complement rule to convert, in order to match the table. The area to the left is $1 - 0.0700 = 0.9300$ (or 93%). Look on the second page with positive z-scores. Look in the body of the table for 0.9300. This exact value does not appear in the table, but the closest value is 0.9306. It is in the row for 1.4 and below the column 0.08. Therefore, the z-score is 1.48.

d) The area to the left is $1 - 0.5910 = 0.4090$. Look on the first page with negative z-scores. Look in the body of the table for 0.4090. It is in the row for -0.2 and below the column 0.03. Therefore, the z-score is -0.23.

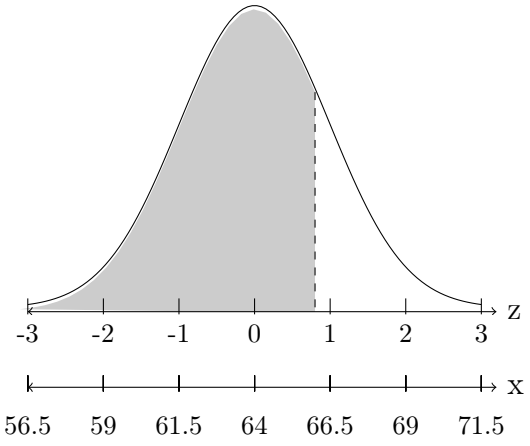
e) The area between is 0.8740, so the area in the two outer tails is $1 - 0.8740 = 0.1260$. Due to the perfect symmetry, the area in each tail is half or 0.0630. Look on the first page with negative z-scores. Look in the body of the table for 0.0630. It is in the row for -1.5 and below the column 0.03. Therefore, the z-scores are -1.53 and $+1.53$.

8. With 10% on each edge, there is 80% in the middle section. We want z_1 and z_2, such that $P(z_1 < Z < z_2) = 0.80$. Draw a bell curve, draw a line to slice the graph into mirror image small slices, one over to the right and one over to left, shade the outer edge slices beyond them, then go to the table at end of the book. The area to the left of the low edge is 0.10. Look on the first page, negative z-scores. Look in the body of the table for 0.1000. This exact value does not appear in the table, but the closest value is 0.1003. It is in the row for -1.2 and below the column 0.08. Therefore, the z-score is $z_1 = -1.28$. Because of the perfect symmetry, the positive cutoff on the other edge is $z_2 = 1.28$. The graph is shown below.

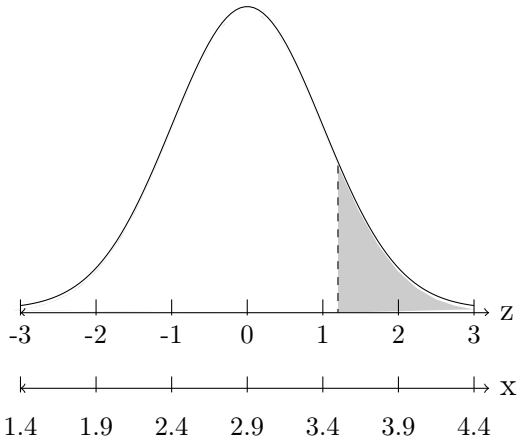


9. This is due to the symmetry of the graph. Going from a certain value below zero and over to the right, is a mirror image of its absolute value and over to the left.

10. Draw a bell curve with standard z-axis from -3 to 3 and below that, an x-axis with heights that correspond to the z marks. The mean height of 64 will go below $z = 0$, one standard deviation higher (66.5 inches) will go below $z = 1$, two deviations higher (69) goes below $z = 2$, and 71.5 below $z = 3$. Do similar process on left side, subtracting standard deviation to go under the negative z-values. Change 5 feet 6 inches into 66 inches. Then convert 66 into a z-score, $z = \frac{66-64}{2.5} = 0.80$. Make a line to slice the graph at about $z = 0.80$, shade below (to left), then go to the table at end of the book. Look on the second page (positive z-scores) and go down to the row for 0.8 and across to the first column 0.00. There we find the area of 0.7881, which is the area to the left, exactly what we need. Therefore, $P(X < 66) = P(Z < 0.80) = 0.7881 = 78.81\%$. A bit more than $\frac{3}{4}$ of women are shorter than 5 foot 6 inches. The graph is shown below.

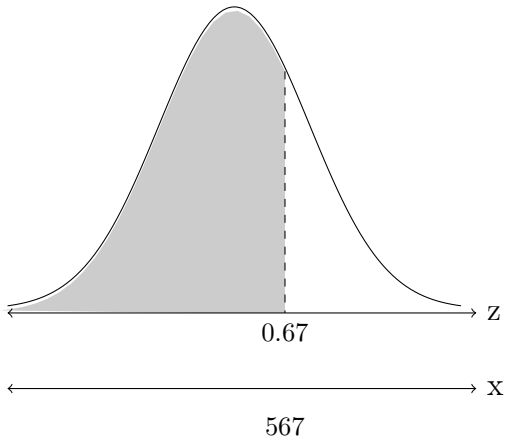


11. Draw a bell curve with standard z-axis from -3 to 3 and below that, an x-axis with weights that correspond to the z marks. The mean weight of 2.9 will go below $z = 0$, one standard deviation higher (3.4 kg) will go below $z = 1$, etc. Convert 3.5 into a z-score, $z = \frac{3.5-2.9}{0.5} = 1.20$. Make a line to slice the graph at about $z = 1.20$, shade above (to right), then go to the table at end of the book. Look on the positive z-scores page and go down to the row for 1.2 and across to the first column 0.00. There we find the area of 0.8849, which is the area to the left, but we want the area to the right. Therefore, $P(X > 3.5) = P(Z > 1.2) = 1 - P(Z < 1.2) = 1 - 0.8849 = 0.1151 = 11.51\%$. A bit more than 11% of babies in the Pakistan are born very large. The graph is shown below.

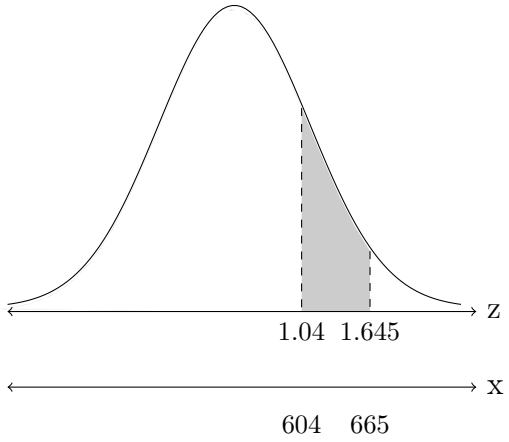


12. Draw a bell curve with standard z-axis from -3 to 3 and below that, an x-axis with IQ scores that correspond to the z marks. The mean score of 100 will go below $z = 0$, one standard deviation higher, 115, will go below $z = 1$, etc. Convert 90 and 120 into z-scores, $z_1 = \frac{90-100}{15} = -0.67$ and $z_2 = \frac{120-100}{15} = 1.33$. Make lines to slice the graph at about $z = 0.67$ and 1.33, shade between. Then go to the table at end of the book. Look on up those z-scores . There we find the areas of 0.2514 and 0.9082, and we want the area between. Therefore, $P(90 < X < 120) = P(-0.67 < Z < 1.33) = P(Z < 1.33) - P(Z < -0.67) = 0.9082 - 0.2514 = 0.6568 = 65.68\%$. Almost 66% of the population has an IQ between 90 and 120.

13. Draw a bell curve, draw a line to slice the graph into a small slice way over to the right, shade below (to left), then go to the table at end of the book. Look on the positive z-score page and look in the body of the table for 0.7500. This exact value does not appear in the table, but the closest value is 0.7486. It is in the row for 0.6 and below the column 0.07. Therefore, the z-score is 0.67. Using the formula $z = \frac{x-\mu}{\sigma}$, we can solve for the unknown score x. Formula setup is $0.67 = \frac{x-500}{100}$. After multiplying both sides by 100 and adding 500, we get $x = 567$ as the 90th percentile. This means that 75% of all people score lower than 567 on any section of the SAT. The graph is shown below.



14. Draw a bell curve, draw a line to slice the graph into a small slices way over to the right, shade between, then go to the table at end of the book. Look on the positive z-score page and look in the body of the table for 0.8500 and 0.9500. These exact values do not appear in the table, but the closest values are 0.8508 and 0.9505 or 0.9495. The first one is in the row for 1.0 and below the column 0.04. Therefore, the z-score is 1.04. The other is exactly halfway between, so the average of the corresponding z-scores is $z = 1.645$. Using the z-score formula for each, we get $x = 604$ as the 85th percentile and $x = 665$ as the 95th percentile. This means that applicants who score between 604 and 664 on the math section of the SAT will qualify. The graph is shown below.


z
1.04
1.645
x
604
665

Solutions to Exercises: Sec 3.3 Sampling Distributions

1. The mean $\mu = \frac{17}{5} = 3.4$. The sampling distribution is

Values	$4,1$	$4,2$	$4,4$	$4,6$	$1,2$	$1,4$	$1,6$	$2,4$	$2,6$	$4,6$
mean $\overline{x}$	2.5	3	4	5	1.5	2.5	3.5	3	4	5

2. There are ten samples. Add up their sample means and divide by 10. $\mu_{\overline{x}} = \frac{34}{10} = 3.4$. This is equal to the population mean itself, which always happens when you have the complete sampling distribution.

3. The Central Limit Theorem says "For a population with mean μ and standard deviation σ, the sampling distribution of the sample means $\overline{x}$ of size n becomes approximately normally distributed as the sample size n gets large, no matter what distribution the population has."

   This means that when individual values follow a normal distribution, the sample means of a particular sample size will also follow a normal distribution. If the individual values do not follow a normal distribution, or the distribution is unknown, the sample means of a particular sample size will only follow an approximately normal distribution when the sample size gets large (about 30+). The larger the sample size, the better the approximation.

4. For each situation, find the mean $\mu_{\overline{x}}$ and standard deviation $\sigma_{\overline{x}}$ for the sampling distribution of the sample mean $\overline{x}$. Then state whether the sampling distribution will be normal, approximately normal, or unknown.

   a) Since individual scores are normally distributed, the sampling distribution will be normal as well with $\mu_{\overline{x}} = 500$ and $\sigma_{\overline{x}} = \dfrac{100}{\sqrt{12}} = 28.9$.

   b) Since individual salaries are not normally distributed and the sample size is under 30, the sampling distribution will be unknown with $\mu_{\overline{x}} = \$68,000$ and $\sigma_{\overline{x}} = \dfrac{3,100}{\sqrt{10}} = \$980.31$.

c) Since individual values are unknown but the sample size is over 30, the sampling distribution will be approximately normal with $\mu_{\overline{x}} = 80.5$ and $\sigma_{\overline{x}} = \dfrac{11}{\sqrt{55}} = 1.5$.

5. Draw a bell curve with standard z-axis from -3 to 3 and below that, an x-axis with weights that correspond to the z marks. The mean weight of 2.9 will go below $z = 0$, one standard deviation higher (3.4 kg) will go below $z = 1$, etc. Make a line to slice the graph just before the mean, shade above (to left). Use the normal function on the calculator **Normalcdf(-9999, 2.8, 2.9, 0.5)** which results in area of $0.2743 = 27.43\%$. This makes sense, since a group of 36 babies should have an average very close to the overall mean of 2.9, so the probability of being less than that is not too large.

6. Here the mean $\mu = 0.82$ and standard deviation $\sigma = \sqrt{\frac{.82(1-.82)}{200}} = 0.027$, low end of interval starts at 80, the high end goes to infinity. Since we need a value for the high end, just use any relatively large negative value. A value of 99999 usually works for most problems. On the TI-83/84 calculators the function would be $normalcdf(.80, 99999, 0.82, 0.027) = 0.7706 = 77.06\%$. This can be interpreted as there is pretty high chance, 77.06%, that more than 80% out of a sample of 200 households will have internet.

Solutions to Exercises: Sec 4.1 Confidence Intervals

1. The sample mean is the average of the lower and upper bounds, $\overline{x} = \dfrac{88 + 128}{2} = 108$. The margin of error is the distance from one of the bounds to the sample mean, $E = 108 - 88 = 20$.

2. The interval is $\overline{x} \pm z\left(\dfrac{\sigma}{\sqrt{n}}\right)$. We have the sample mean $\overline{x} = 1085$ and sample size $n = 350$. Now we just need the z-score that splits bell curve into 95% center piece. Use the table or calculator function **invNorm(0.025)** to get the correct z-score of -1.96. Now the confidence interval is $1085 \pm 1.96\left(\dfrac{200}{\sqrt{350}}\right) = 1085 \pm 21$ or the interval of values from 1,064 to 1,106.

3. The interval is $\overline{x} \pm t\left(\dfrac{s}{\sqrt{n}}\right)$. We have the sample mean $\overline{x} = 3195$ and sample size $n = 1597$. We do not have the population standard deviation so we will use the sample value $s = 493$. Now we just need the t-score that splits bell curve into 99% center piece. If 99% is in the center, then the other 1% is on the edges, split in half, so the lower tail is 0.5% = 0.005 in area. The degrees of freedom is $df = 1597 - 1 = 1596$. Use the table at df = 1000 to get the approximate t-score of 2.581. Now the confidence interval is $3195 \pm 2.581\left(\dfrac{493}{\sqrt{1597}}\right) = 3195 \pm 31.8$ or the interval of values from 3,163.2 to 3,226.8 grams.

4. We already have the sample proportion $\hat{p} = 0.58$. The z-score is **invNorm(0.025)** which is -1.96. The interval is $\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.58 \pm 1.96\sqrt{\frac{0.58(1-0.58)}{800}} = 0.58 \pm 0.034$ which results in an interval from 0.546 to 0.614 or 54.6% to 61.4%. This shows that the governor will likely be re-elected.

5. In order to estimate a population proportion, find the minimum sample size required to have a maximum margin of error of 3% for a confidence level of 96%. Assume $\hat{p} = 50\%$. The formula gives $n = \left(\dfrac{z}{E}\right)^2 \hat{p}(1-\hat{p}) = \left(\frac{2.05}{0.03}\right)^2 (0.50)(1-0.50) = 1167.36$. The minimum sample size is 1,168.

Solutions to Exercises: Sec 4.2 Hypothesis Tests

1. For each hypothesis setup, state whether it will be a left, right, or two-tail test.

   a) Since the alternative hypothesis has less than, this is a left tail test.

   b) Since the alternative hypothesis has not equal, this is a two tail test.

   c) Since the alternative hypothesis has greater than, this is a right tail test.

   d) Since the alternative hypothesis has greater than, this is a right tail test.

2. For each hypothesis setup, state whether it will be a z-test for a mean, a t-test for a mean, or a z-test for a proportion, and whether it will be a left, right, or two-tail test.

   a) With the word above, this is a right tail test. The word average implies this deals with the population mean. Since there is no population standard deviation, the sample value will be used, which makes this a t-test for the mean.

   b) Stealing votes implies they are worried the poll numbers will go down, so this is a left tail test. Since polls are based on percentages, this will be a z-test for proportions.

   c) Cholesterol is a measurement, for populations and samples that would lead to averages. The statement "any difference in the amount" results in either direction, so this is a two tail test and having an assumed population standard deviation, this would be a z-test for a mean.

   d) With the word decrease, this is a left tail test. Since there is no population standard deviation, the sample value will be used, which makes this a t-test for the mean.

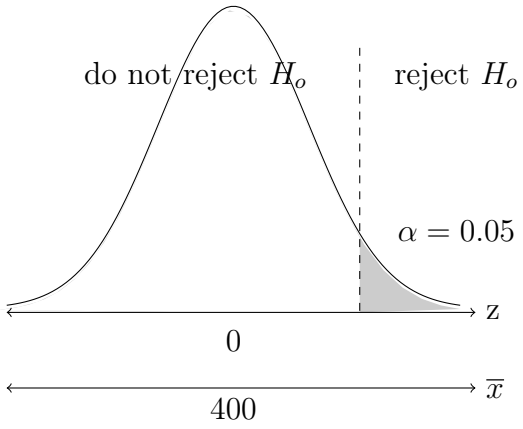
3. For each set of hypothesis test results, state the conclusion as reject or do not reject the null claim. Explain what you compare to make that decision.

a) The p-value $<$ alpha, reject null claim that $\mu = 87$, go with $H_a : \mu \neq 87$ instead.

b) $Z_{data} = -1.65$ does not go past the left tail $Z_{crit} = -1.96$, so do not reject the null hypothesis. The population proportion could very well be 0.18.

c) The p-value $>$ alpha, do not reject null claim that $\mu = 900$, keep it, it is reasonable.

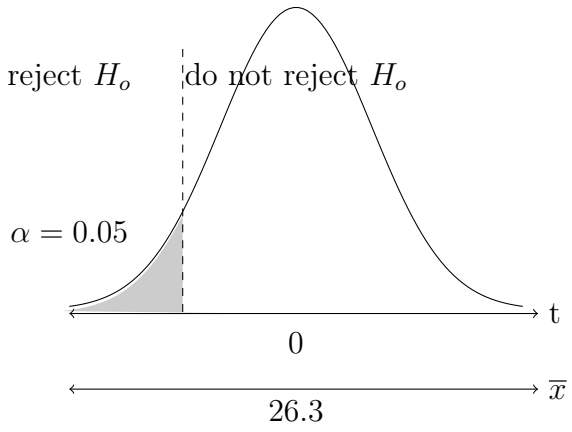
4. $H_o : \mu = 400 \quad H_a : \mu > 400$



$Z_{data} = \dfrac{403 - 400}{\frac{25}{\sqrt{50}}} = 0.84,$ $\qquad$ $Z_{crit} = \text{InvNorm}(0.95) = 1.645$ or from the z-table.

Z_{data} is less than the critical value 1.645, so the data evidence falls closer to center in the do not reject region. Therefore, we can say that there is not enough evidence to reject the claim that the mean is 400. Notice the sample mean was greater, but only slightly. Not far enough to put doubt on the null claim value.

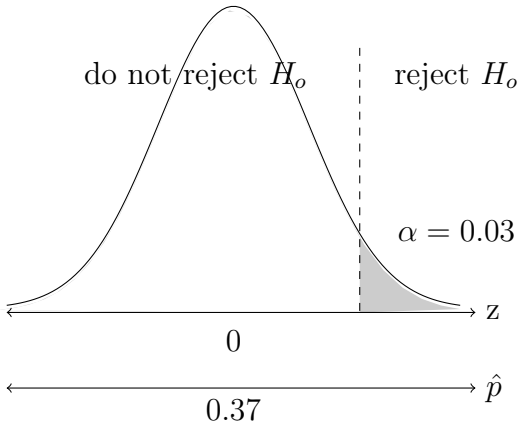
5. $H_o : \mu = 26.3 \quad H_a : \mu < 26.3$. With less than, we are doing a left-tail test. The setup for the bell curve is



$t_{data} = \frac{25.9-26.3}{\frac{2}{\sqrt{101}}} = -2.010,$ p-value = tcdf(-9999,-2.010,100) = 0.024

Since the p-value $<$ alpha, the data is far enough from the null claim, so we reject the null hypothesis. Therefore, there is enough evidence to reject the claim that the mean age for Latina women is same as the population. Their age at first birth is likely less than that 26.3 years.

6. $H_o : p = 0.37$ $H_a : p > 0.37$. The setup for the bell curve is



$\hat{p} = \dfrac{x}{n} = \dfrac{147}{380} = 0.387$ $Z_{data} = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.387-0.37}{\sqrt{\frac{0.37(1-0.37)}{380}}} = 0.68$

Z_{crit} = InvNorm(0.97) = 1.88 or from the z-table. Since the test statistic does not fall out into the right tail rejection region, there is not enough evidence to suggest that college students have greater proportion of not having landline than adults in general.

Solutions to Exercises: Sec 5.1 Correlation and Regression

1. Correlation is when two variables tend to be related. Certain values of one variable tend to be paired with certain values of the other. For positive correlation, as one goes up (increases), the other tends to go up. For negative correlation, as one goes up, the other tends to go down (decreases).

2. Match the most likely linear correlation values to the graphs below.

   $r = +0.7 \Rightarrow b \quad r = +0.99 \Rightarrow a \quad r = -0.4 \Rightarrow c \quad r = +0.15 \Rightarrow e \quad r = -0.86 \Rightarrow d$

   $r = 0 \Rightarrow f$

3. We can add extra rows to organize the calculations.

x	2	5	7	10	12	14	$\sum x = 50$
y	2	6	7	9	11	14	$\sum y = 49$
x^2	4	25	49	100	144	196	$\sum x^2 = 518$
y^2	4	36	49	81	121	196	$\sum y^2 = 487$
xy	4	30	49	90	132	196	$\sum xy = 501$

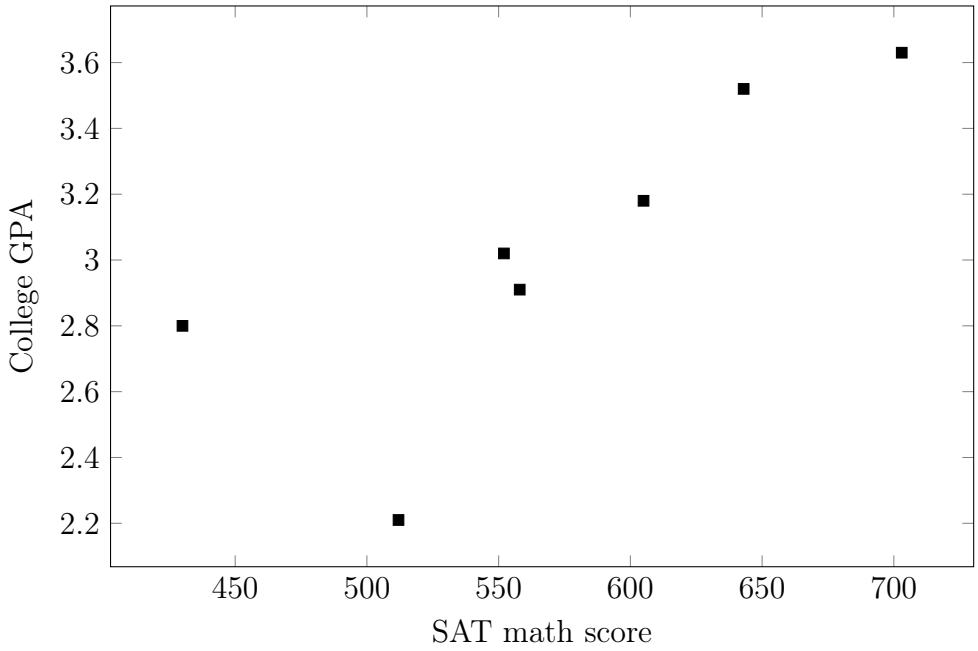
   Now we input the value $n = 6$ and all of the sums into the formula to get:

$$r = \frac{501 - \frac{50(49)}{6}}{\sqrt{\left[518 - \frac{(50)^2}{6}\right]\left[487 - \frac{(49)^2}{6}\right]}} = \frac{501 - 408.33}{\sqrt{[518 - 416.67][487 - 400.17]}} = \frac{92.67}{93.80} = +0.988$$

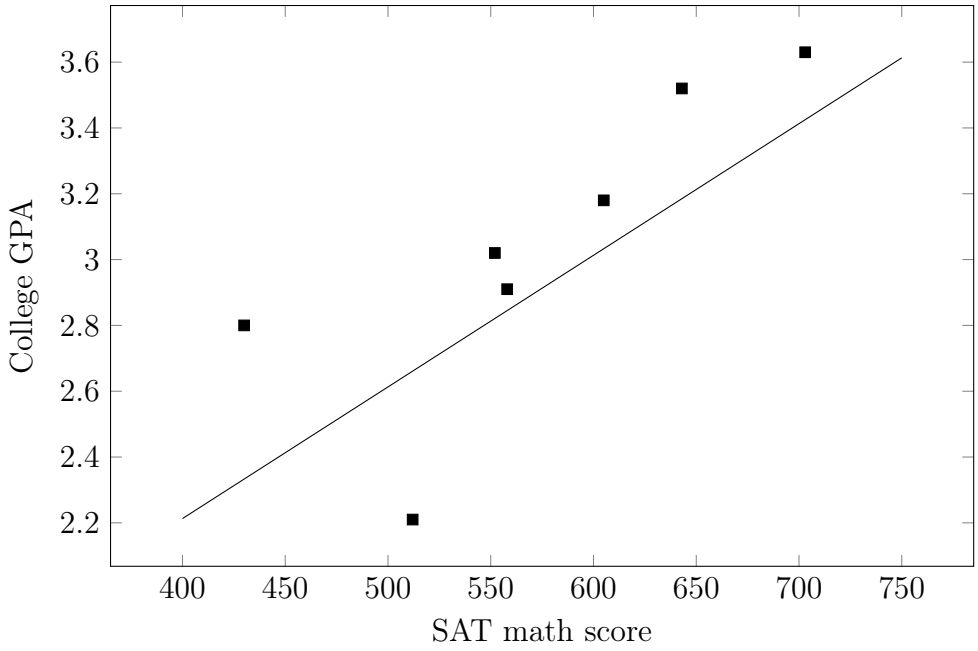
4. Go to STAT EDIT menu, enter the SAT data into L_1. Then move over to L_2 and input the GPA data exactly in order. Press STAT, CALC menu, scroll down to item LinReg(ax+b) and press ENTER. Type L_1, L_2 then hit enter again. The results you should see are $r = +0.793$ (rounded). This is greater than the minimum value from the table of 0.754, therefore, the value $r = +0.793$ is large enough to state that there is a strong correlation between SAT math score and college GPA. It is probably that

the students who get higher scores, work harder, study well, and are more interested in learning, so they do well in college.

To get the scatterplot, press 2nd Y= (for stat plot menu), choose Plot1, select first type icon for scatterplot. Set the Xlist to L_1 and Ylist to L_2, then press GRAPH. If you do not see it, go to ZOOM and select ZoomStat, hit enter. The scatterplot will look like this:

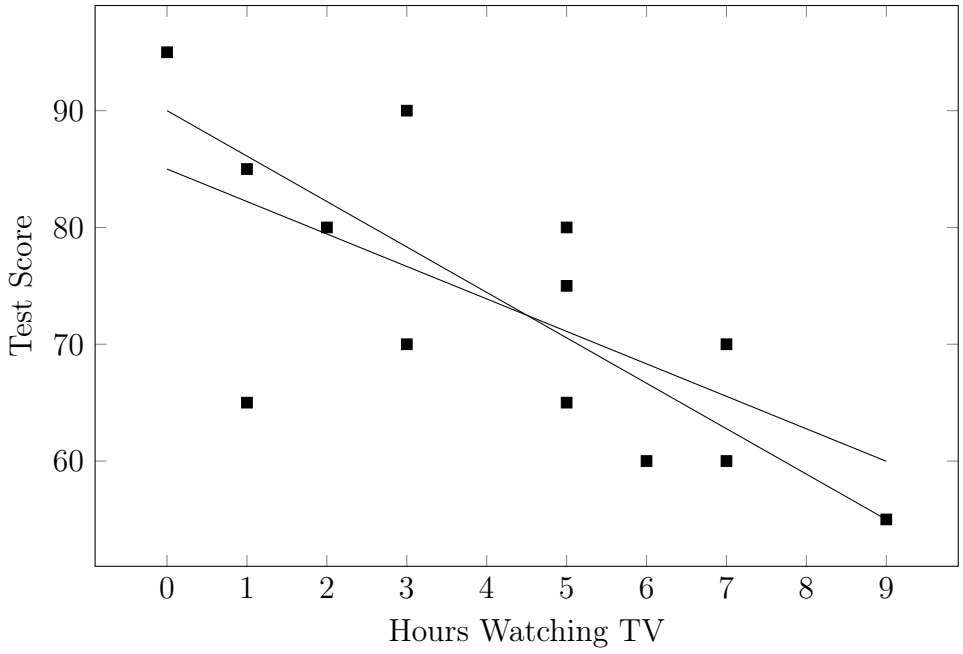


5. Press STAT, CALC menu, scroll down to item LinReg(ax+b) and press ENTER. Type L_1, L_2 then hit enter again. The results you should see are $y = 0.004x + 0.613$ (rounded). These values imply that for every point your SAT math score increases, your college GPA will increase by 0.004, and that with a zero on SAT math, you should still be able to get a 0.613 GPA. In reality, the lowest SAT score is 200, but the equation follows that pattern if we could project down to zero. The regression line is shown along the scatterplot points below.



6. The predictions are $y = 0.004(760) + 0.613 = 3.653$ GPA with an SAT math score of 760, and $y = 0.004(500) + 0.613 = 2.613$ GPA with an SAT math score of 500. The first one is extrapolation, since 760 is outside the range of the data (430 to 703). The second is interpolation, since 500 is in the range.

7. Two possible regression lines are shown below. Other similar lines would be good fits as well. There appears to be a somewhat strong negative correlation between the variables. More TV watching tends to match with lower scores. It is reasonable to assume that watching a lot of TV, takes time away from studying, and therefore causes lower scores (on average). There are exceptions, but this is a general fact backed up by much research.



8. A positive correlation coefficient r corresponds with a positive slope. A negative correlation coefficient r corresponds with a negative slope. The values have no relation, only the sign, $\pm$.

Solutions to Exercises: Sec 5.2 Joint Distributions

1. The table below shows the breakdown of employed nurses by type and race they identify with. The data comes from the US Department of Health and Human Services and are rounded to nearest 100. Calculate the marginal frequencies and the grand total.

	RN	LPN	totals
Asian	234,600	24,800	258,800
African American	279,600	162,800	442,400
Hispanic/Latino	135,600	51,800	187,400
Native American	11,300	4,100	15,400
White	2,164,100	446,500	2,610,600
totals	2,824,600	690,000	3,514,600

2. The table below shows the breakdown of 2020-2021 UWG students by status and level. Calculate the marginal frequencies and the grand total, then rewrite the chart with relative frequencies.

	Full-time	Part-time	totals
Undergraduate	7,528	2,803	10,331
Graduate	652	2,436	3,088
totals	8,180	5,239	13,419

	Full-time	Part-time	totals
Undergraduate	56.1%	20.9%	77.0%
Graduate	4.9%	18.2%	23.0%
totals	61.0%	39.0%	100%

3. The table below shows partial data for the breakdown of US adults by group and opinion about guns. Fill in the missing frequencies. By adding or subtracting to complete each row and column, we get the completed table below.

	Married Men	Unmarried Men	Married Women	Unmarried Women	totals
Protect gun rights	568	396	448	332	1,744
Have more gun control	319	311	527	565	1,722
totals	887	707	975	897	3,466

4. The percent of men who are obese is found by dividing the number of obese men by the total number of men $\frac{2171}{5120} = 42.4\%$. The percent of severely obese people who are women is found by dividing the number of severely obese women by the total number of severely obese people $\frac{651}{1122} = 58.0\%$

5. The probability would be calculated as the number of Instructors with a Master's divided by the total number of Instructors. We can setup an equation for that as $\frac{x}{120} = 0.90$, multiply both sides by 120 to get $x = 0.90 * 120 = 108$, so there must be 108 Instructors who have a Master's degree.

6. The probability would be calculated as the number of Lecturers with a PhD divided by the total number of Lecturers. We can setup an equation for that as $\frac{8}{x} = 0.16$, multiply both sides by x, then divide by 0.16 to get $x = \frac{8}{0.16} = 50$, so there must be a total of 50 Lecturers.

Standard Normal Table: Cumulative Areas to Left of Z (negative)

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1432	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Standard Normal Table: Cumulative Areas to Left of Z (positive)

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

T Table: Critical Values of the Student's T Distribution

	Confidence level C				
	99%	98%	95%	90%	80%
	Area in One Tail				
	0.005	0.01	0.025	0.05	0.10
df	**Critical t-values**				
2	9.925	6.965	4.303	2.920	1.886
3	5.841	4.541	3.182	2.353	1.638
4	4.604	3.747	2.776	2.132	1.533
5	4.032	3.365	2.571	2.015	1.476
6	3.707	3.143	2.447	1.943	1.440
7	3.499	2.998	2.365	1.895	1.415
8	3.355	2.896	2.306	1.860	1.397
9	3.250	2.821	2.262	1.833	1.383
10	3.169	2.764	2.228	1.812	1.372
11	3.106	2.718	2.201	1.796	1.363
12	3.055	2.681	2.179	1.782	1.356
13	3.012	2.650	2.160	1.771	1.350
14	2.977	2.624	2.145	1.761	1.345
15	2.947	2.602	2.131	1.753	1.341
16	2.921	2.583	2.120	1.746	1.337
17	2.898	2.567	2.110	1.740	1.333
18	2.878	2.552	2.101	1.734	1.330
19	2.861	2.539	2.093	1.729	1.328
20	2.845	2.528	2.086	1.725	1.325
21	2.831	2.518	2.080	1.721	1.323
22	2.819	2.508	2.074	1.717	1.321
23	2.807	2.500	2.069	1.714	1.319
24	2.797	2.492	2.064	1.711	1.318
25	2.787	2.485	2.060	1.708	1.316
26	2.779	2.479	2.056	1.706	1.315
27	2.771	2.473	2.052	1.703	1.314
28	2.763	2.467	2.048	1.701	1.313
29	2.756	2.462	2.045	1.699	1.311
30	2.750	2.457	2.042	1.697	1.310

T Table: Critical Values of the Student's T Distribution

	Confidence level C				
	99%	98%	95%	90%	80%
	Area in One Tail				
	0.005	0.01	0.025	0.05	0.10
df	**Critical t-values**				
31	2.744	2.453	2.040	1.696	1.309
32	2.738	2.449	2.037	1.694	1.309
33	2.733	2.445	2.035	1.692	1.308
34	2.728	2.441	2.032	1.691	1.307
35	2.724	2.438	2.030	1.690	1.306
36	2.719	2.434	2.028	1.688	1.306
37	2.715	2.431	2.026	1.687	1.305
38	2.712	2.429	2.024	1.686	1.304
39	2.708	2.426	2.023	1.685	1.304
40	2.704	2.423	2.021	1.684	1.303
41	2.701	2.421	2.020	1.683	1.303
42	2.698	2.418	2.018	1.682	1.302
43	2.695	2.416	2.017	1.681	1.302
44	2.692	2.414	2.015	1.680	1.301
45	2.690	2.412	2.014	1.679	1.301
46	2.687	2.410	2.013	1.679	1.300
47	2.685	2.408	2.012	1.678	1.300
48	2.682	2.407	2.011	1.677	1.299
49	2.680	2.405	2.010	1.677	1.299
50	2.678	2.403	2.009	1.676	1.299
60	2.660	2.390	2.000	1.671	1.296
70	2.648	2.381	1.994	1.667	1.294
80	2.639	2.374	1.990	1.664	1.292
90	2.632	2.368	1.987	1.662	1.291
100	2.626	2.364	1.984	1.660	1.290
200	2.601	2.345	1.972	1.653	1.286
300	2.592	2.339	1.968	1.650	1.284
500	2.586	2.334	1.965	1.648	1.283
1000	2.581	2.330	1.962	1.646	1.282

Index

68-95-99.7 Rule, 101

Addition Rule, 76

Alternative Hypothesis, 137

Average, 31

Bad graphs, 25

Bar Graph, 20

Bias, 7

Binomial Distribution, 90

Binomial Experiment, 90

Binomial probability formula, 92

Binomial Variable, 90

Bivariate data, 169

Boxplot, 45

Calculator graphing, 53

Categorical data, 169

Categorical Variables, 6

Census, 5

Center, 21

Central Limit Theorem, 120

Class boundaries, 14

Class midpoint, 14

Class width, 14

Classes, 14

Cluster Sampling, 9

Coefficient of Variation, 38

Combination, 69

Complement Rule, 63

Conditional frequency, 172

Conditional Probability, 77

Confidence Interval, 126

Confidence level, 127

Continuous, 6

Convenience Sampling, 9

Correlation, 154

Critical Value Method, 142

Cumulative Frequencies, 16

Degrees of Freedom, 129

Dependent events, 78

Designed experiment, 6

Discrete, 6

Discrete Random Variable, 86

Disjoint Events, 75

Distribution, 14

Double Blind, 7

Empirical Probability, 60
Empirical Rule, 101
Errors, 163
Estimation, 126
Event, 58
Expected Value, 88
Experiment, 58
Extrapolation, 163

Factorial notation, 69
Five-Number Summary, 45
Frequency, 15
Frequency distribution, 15
Frequency Histogram, 18
Fundamental Counting Principle, 69

Hypothesis Testing, 137

Independent events, 78
Interpolation, 163
Interquartile Range, 46
Intersection, 75
invNorm calculator function, 128, 215
invT calculator function, 129

Joint distribution, 169
Joint frequencies, 169

Law of Large Numbers, 61
Left-skewed, 21
Left-tail test, 138
level of significance, 140
Linear Correlation Coefficient, 156
Linear Regression, 160
Lower Fence, 46
Lower limit of a class, 14

Margin or error, 126
Marginal frequencies, 170
Maximum, 36
Mean, 31
Measures of center, 31
Measures of Relative Standing, 42
Measures of spread, 36
Measures of variation, 36
Median, 32
Minimum, 36
Mode, 33
Modified Boxplot, 46
Multiplication Rule, 78
Mutually Exclusive, 75

Negative Correlation, 155
Normal Distribution, 99
normalcdf calculator function, 122
Null Hypothesis, 137

Observational study, 6
Odds, 64

Outlier, 22

P-Value Method, 142

Parameter, 5

Pareto chart, 20

Percentiles, 43

Permutation, 70

Pie Chart, 19

Placebo, 7

Population, 4

Positive Correlation, 154

Probability, 58

Probability Distribution, 79

Probability Histogram, 92

Qualitative Variables, 6

Quantitative Variables, 6

Quartiles, 43

randInt: calculator function, 51

Random Sampling, 7

Random Variable, 85

Range, 36

Regression, 160

Regression Equation, 160, 162

Rejection Region, 141

Relative Frequency, 15

Relative Frequency distribution, 15

Relative Frequency Histogram, 18

Representative Sample, 7

Residuals, 163

Right-skewed, 21

Right-tail test, 138

Round-Off Rule, 31

Sample, 5

Sample Space, 59

Sampling Distribution, 119

Scatterplot, 155

Shape, 21

significance level, 140

Simple Event, 59

Simple random sampling, 8

Skewness, 21

Spread, 21

Standard deviation, 36

Standard Normal Distribution, 100

STAT PLOTS menu, 53

Statistic, 5

Statistics, 4

Stratified Sampling, 9

Student's T-distribution, 129

Subjective Probability, 61

Symmetry, 21

Systematic Sampling, 8

T-distribution, 129

Test Statistic, 142

Theoretical Probability, 60

Time series, 24

Treatments, 7

Two-tail test, 138

Two-way frequency chart, 169

Type-1 Error, 151

Type-2 Error, 151

Uniform, 21

Uniform Distribution, 97

Union, 75

Upper Fence, 46

Upper limit of a class , 14

Variable, 4

Variance, 36

Venn diagrams, 80

Weighted Mean, 34

z-score, 42

ZoomStat, 54