Versatile Mathematics

COMMON MATHEMATICAL APPLICATIONS

by FCC Math Faculty

Frederick Community College

2015

Attributions This book benefited tremendously from others who went before and freely shared their creative work. The following is a short list of those whom we have to thank for their work and their generosity in contributing to the free and open sharing of knowledge.

- David Lippman, author of *Math in Society.* This book uses sections derived from his chapters on Finance, Growth Models, and Statistics. He also administers MyOpenMath, the free online homework portal to which the problems in this text were added.
- The developers of onlinestatbook.com.
- OpenStax College (their book *Introductory Statistics* was used as a reference)
  OpenStax College, *Introductory Statistics.* OpenStax College. 19 September 2013. <http://cnx.org/content/col11562/latest/>
- The authors of OpenIntro Statistics, which was also used as a reference.
- The Saylor Foundation Statistics Textbook: http://www.saylor.org/site/textbooks/Introductory%20Statistics.pdf

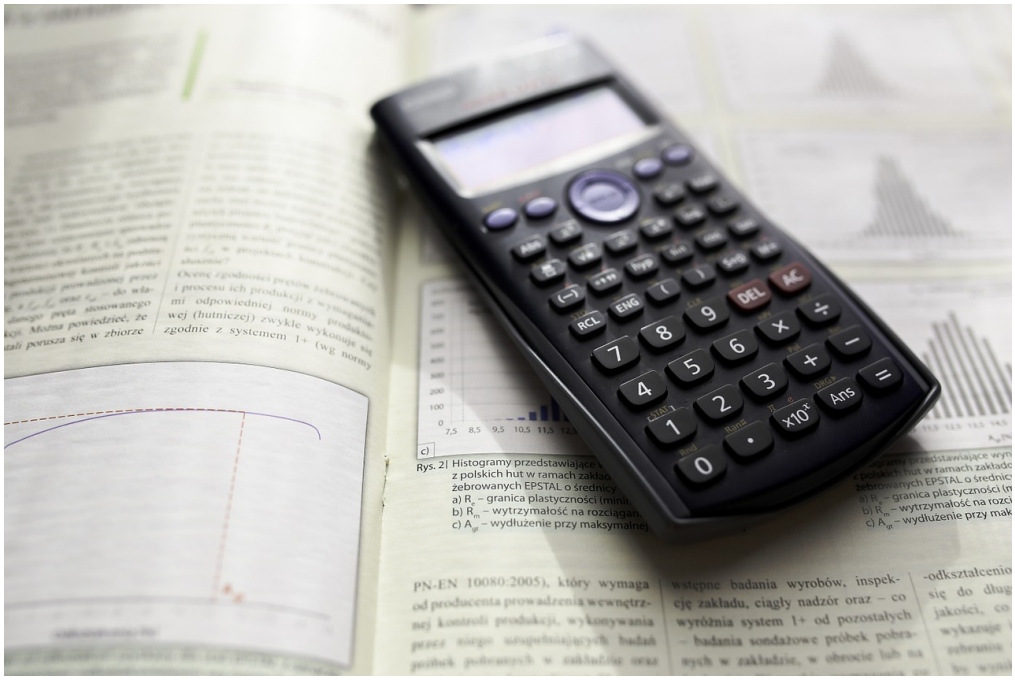
Thanks The following is a short list of those whom we wish to thank for their help and support with this project.

- The President's office at Frederick Community College, for providing a grant to write the first chapters.
- Gary Hull, who in his tenure as department chair gave us his full support and gave us the impetus to start the project, and generously shared his notes for MA 103.
- The entire FCC math department, who provided untold support and encouragement, as well as aid in reviewing and editing the text.

Contents

Chapter

3 Statistics



There's a popular joke among statisticians that 64.8% of all statistics are made up on the spot. How can you tell the difference between good and bad statistics? Where do the numbers come from? How is data collected? No other branch of mathematics has a more tremendous impact on our lives than the field of statistics. Statistics are everywhere, from crime rates in your city to weight percentiles for children on growth charts. When a research team is testing a new treatment for a disease, they can use statistics to make conclusions based on a relatively small trial and show that there is good evidence that their drug is effective. Statistics allowed prosecutors in the 1950's and 60's to demonstrate that racial bias existed in jury panels. In this chapter, you will get a glimpse into this important subject, understanding the essentials and learning to become a wise consumer of statistics.

SECTION 3.1 Sampling and Graphs

The 2010 census cost about $13 billion to administer

Let us start with the basics. What is statistics? The field of **statistics** encompasses collecting, organizing, analyzing and presenting data. **Data** is simply collected information. That information can be collected via surveys, polls, records, experiments, studies, or censuses, just to name a few. A **census** is when we collect data on the entire population, polling each and every individual. As you can imagine, that would take a lot of time and resources. That is why in the United States, a census occurs only every 10 years. In all other situations, we do what is known as sampling, where we assume that if we study a small portion of the total group, the results will be similar to what we would find if we polled the entire group.

Our goal in statistics, then, is to select a good sample, gather data from the sample, organize and summarize the data, and draw conclusions from the sample about the total population.

Population and Sample

A **population** is a collection of persons or objects under study. To study the population, we select a **sample**. The idea of **sampling** is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Sampling is a very practical technique. If you wished to find the average height of students at your school, you probably wouldn't collect data from every single student; this wouldn't be very feasible. It would make more sense to select a sample of students. The data collected would be the students' heights. In presidential elections, opinion polls sample 1000 to 2000 people. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of ice cream take samples to determine if a pint of ice cream actually contains a pint of ice cream.

EXAMPLE 1 **PET OWNERSHIP**

A sample of 2,000 households in the U.S. was selected and asked if they currently own at least one pet. The results show that 69% of households do own at least one pet. Identify the sample and population in this situation.

Solution

The sample is the 2,000 households and the population is all households in the U.S. Notice that even though the population is not explicitly stated, we can infer it from carefully reading the sentence.

TRY IT

A researcher wants to know how citizens of Frederick City felt about a voter initiative. To study this, she goes to the Francis Scott Key Mall in the city, randomly selects 500 shoppers and asks them their opinion. Sixty percent indicate they are supportive of the initiative. What is the sample and population?:

Let's go back and think about finding the average height of students at your school. How would you go about obtaining your sample? Would you sample all your friends, or all your classmates? How about asking people in the library on a Wednesday afternoon? Would any of these sampling techniques give you a good estimate of the average height of all students?

We have to make sure our sample is not **biased**, or leaning in a certain direction. Maybe you like to hang out with tall people and all your friends are tall. Maybe only 10 people are in the library on Wednesdays. Your sample must be random and representative in order to get a good estimate of the population data. A sample is **random** if everybody in the population has the same chance of being selected into the sample. A sample is **representative** if it contains the characteristics of the population.

Random Samples

The key to a random sample is that each member of the population is equally likely to be selected.

Examples of Random Samples

- If the population is students in a particular classroom, number the students in the classroom from 1 to n and use a random number generator to select numbers between 1 and n.
- If the population is FCC students, list their FCC email accounts and number them, then pick random numbers between 1 and n, where n is the number of students.

Examples of Biased Samples

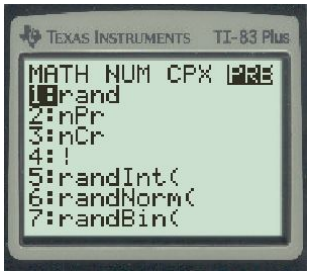
- If the population is residents of Frederick County, number the entries in the phone book and use a random number generator to select a sample.
- If the population is American citizens, go to the entrance of Yankee Stadium and poll everyone entering.
- If the population is FCC students, poll the students in one class.

Not every resident of the county has a phone, let alone a phone listed in the phone book

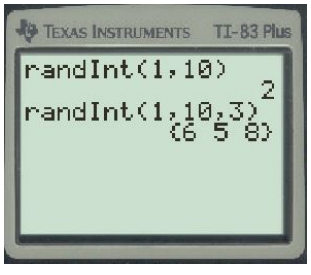
If you're ever unclear on whether or not a sampling method is random, simply ask whether or not every member of the population has an equal chance of being selected. A random sample gives better results than a biased sample because characteristics that could muddy the results of the study get averaged out by the randomness.

Picking Random Numbers with a TI-83

Your graphing calculator can select random numbers. To access this menu, press the MATH button and use the left and right arrows to navigate to the PRB tab (PRB for probability).



Selecting `rand` will select a "random" (technically pseudo-random, but close enough for us) number between 0 and 1. Selecting `randInt(` will allow you to choose a random integer between a given lower and upper bound, or several of those.



To select a single random integer, enter `randInt(`lower bound`,`upper bound`)`, with whatever numbers you want for the lower and upper bounds. To select n random integers, enter `randInt(`lower bound`,`upper bound`,`n`)`

EXAMPLE 2 **CARTWHEELS**

A coach is interested in how many cartwheels the average college freshman can do at his university. Eight volunteers from the freshman class step forward. After observing their performance, the coach concludes that college freshmen can do an average of 16 cartwheels in a row without stopping. Is this sample random and representative?

Solution

The population is the class of all freshmen at the coach's university. The sample is composed of all freshmen so that is good. However, the sample is poorly chosen because volunteers are more likely to be able to do cartwheels than the average freshman: people who cannot do cartwheels probably did not volunteer! Hence, this sample is not random. We are also not told of the gender of the volunteers. Were they all women, for example? That might affect the outcome (if the school is co-ed). We cannot be sure this sample is representative.

TRY IT

A substitute teacher wants to know how students in the class did on their last test. The teacher asks the 10 students sitting in the front row to state their latest test score. He concludes from their responses that the class did extremely well. What is the sample and population? Is the sample random and representative? Why or why not?

In general, self-selected samples (or volunteer samples) are not representative of the population. For this reason, surveys with voluntary responses are not reliable. People who volunteer their opinion for online reviews, for instance, tend to be strongly positive or negative; the voluntary sample misses everyone in the middle who doesn't have a strong opinion.

The most famous example of this comes from the 1936 presidential election, where the incumbent Democrat, Franklin D. Roosevelt, was challenged by the Republican governor of Kansas, Alf Landon. The *Literary Digest*, a weekly magazine, boasted that it had correctly predicted the results of the last 4 elections by sending out questionnaires to its huge sample of readers. In 1936, the *Digest* sent out 10 million questionnaires and received over 2 million responses, predicting that Landon would unseat Roosevelt with a handy victory. When Election Day came, though, Roosevelt received over 60% of the popular vote, carrying every state except for Maine and Vermont (including Landon's home state). It was one of the most lopsided victories in U.S. history. The reason for the failure of this poll was largely based on the voluntary response nature–those who responded were more likely to be those who were unhappy with the current administration; people who were happy with Roosevelt's programs had no incentive to fill out the questionnaire and send it in.

Largely due to this failure and embarrassment, the *Literary Digest* folded within a few years. In contrast, a young pollster named George Gallup (whose name is borne by the Gallup polls today) made his name in the 1936 election by correctly predicting the winner with a much smaller, carefully chosen sample.

Frequency Distributions

One of the most fundamental parts of statistics is presenting data clearly. After the data has been collected, it must be presented in a way so that it conveys all the necessary information and is easy for the viewer to understand. It is usually impractical and unhelpful to list all of the data that we gather for our audience; imagine gathering data on standardized test scores in a state and listing thousands of unsorted scores–there would be no way to make sense of the data. When we present data, we want to show a clear picture of what's going on, without overwhelming our audience with so much data that the reality gets lost in the noise.

The first way that we'll present data is with a **frequency distribution**, which simply counts how often each data value occurs. For instance, suppose twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows:

$$5, 6, 3, 3, 2, 4, 7, 5, 2, 3, 5, 6, 5, 4, 4, 3, 5, 2, 5, 3.$$

The table below lists the different data values in ascending order and their frequencies.

Data Value	Frequency
2	3
3	5
4	3
5	6
6	2
7	1

If one data value between 2 and 7 didn't appear in the data set–for example, if none of the students responded that they worked 4 hours per day–we would usually still include the row for 4 and just note that the frequency was 0 for that value.

The **frequency** is the number of times a data value occurs. According to the table above, there are three students who work two hours, five students who work three hours, and so on. The sum of the values in the frequency column, 20, represents the total number of students included in the sample.

We can also calculate the **relative frequency** of each value, which is the ratio of the number of times a value occurs to the total number of outcomes. We can include these in a third column on our frequency table. To find the relative frequencies, divide each frequency by the total number of students in the sample–in this case, 20. Relative frequencies can be written as fractions, decimals, or as in the table below, as percents.

Data Value	Frequency	Relative Frequency
2	3	3/20 or 15%
3	5	5/20 or 25%
4	3	3/20 or 15%
5	6	6/20 or 30%
6	2	2/20 or 10%
7	1	1/20 or 5%

The sum of all the values of the relative frequency column of the table above is 20/20, or 1. Note, because of rounding, the relative frequency column may not always sum to one. However, it should be close to one.

DAILY COMMUTE EXAMPLE 3

Nineteen people were asked how many miles, to the nearest mile, they commute to work each day. They responded as follows:

$$2, 5, 7, 3, 2, 10, 18, 15, 20, 7, 10, 18, 5, 12, 13, 12, 4, 5, 10.$$

This is summarized in the frequency table below:

Data Value	Frequency	Relative Frequency
3	3	3/19
4	1	1/19
5	3	3/19
7	2	2/19
10	3	4/19
12	2	2/19
13	1	1/19
15	1	1/19
18	1	1/19
20	1	1/19

- Is the table correct? If it is not correct, what is wrong?

  It is incorrect, because the frequency column sums to 18, not 19 as it should. One of the data values was left out. Besides, two people responded that they commute 2 miles, and that doesn't appear on the table at all.

- True or false: Three of the people surveyed commute three miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.

  False. The frequency for 3 miles should be 1. When building the table, the two that responded 2 miles got lumped into the 3 category.

- What fraction of the people surveyed commute five or seven miles?

  5/19

- What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between 5 and 13 miles (not including those who commute exactly 5 or exactly 13 miles)?

  7/19, 12/19, 7/19

Sometimes, it's more practical to make a **grouped frequency distribution**, where the frequencies are not single numbers, but groups of numbers.

For instance, suppose you gathered data on how long it took you to get ready in the morning. For 40 days, you measured the amount of time between when your alarm went off and you left the house, and you got the following results in minutes:

35.6	28.7	25.5	23.2	23.7	32.1	29.6	19.5	21.4	13.6
24.9	26.7	25.8	31.8	30.4	20.1	25.3	29.9	37.5	26.6
32.3	36.5	18.5	17.7	15.2	24.7	21.4	16.6	19.5	30.3
38.7	27.4	22.9	24.6	28.5	17.5	31.4	32.2	21.6	28.4

If we built a frequency distribution with one row for each distinct value, the table would not give any useful information; it wouldn't serve the purpose of a frequency distribution, which is to give a preliminary idea of where the data is clustered and where it is spread out.

Instead, we can make a grouped frequency distribution by grouping together data values in the same range. If we do this for the data set above by grouping in sets of 5, we get the grouped frequency distribution below.

Data Values	Frequency
10–less than 15	1
15–less than 20	7
20–less than 25	10
25–less than 30	11
30–less than 35	7
35–less than 40	4

This distribution has six categories, called **classes**. The starting values of the classes (10, 15, 20, etc.) are called the **lower class limits**, and the ending values are called the **upper class limits**. The **class width** is the difference between the lower class limits. Notice that if we subtract $20 - 15$, we get 5. Therefore, the class width in this example is 5.

Choosing Classes

When building a grouped frequency distribution, you'll usually have the freedom to choose how you want to separate the classes. Here are some guidelines you should follow:

- Each class should be the same width. Notice in the example above that each class was five units wide. If not, the grouped frequency distribution would not give a clear picture of how the data is arranged.
- Classes cannot overlap. In the example above, we avoided overlap by letting the first class go up to "less than 15" and having the second class start at 15, and similarly for the other classes. If the classes overlapped at 15, it wouldn't be clear into which category we should place a data value of 15.0.
- Avoid empty classes if possible. This can occur if you choose to have too many classes.
- Don't make open-ended classes. For instance, in the example above we didn't make the last class "35 and above," which would have been an open-ended class. The reason not to do this is that it violates the first guideline about having all classes have the same width.

To find the appropriate class width to use, you can start by deciding on the number of classes you want to use. In our example above, we chose 6. Then the class width is found by dividing the distance from the minimum to the maximum by the number of classes.

Class Width

$$\text{Class Width} = \frac{\text{Maximum - Minimum}}{\text{Number of Classes}}$$

Round UP to the next whole number.

Histograms

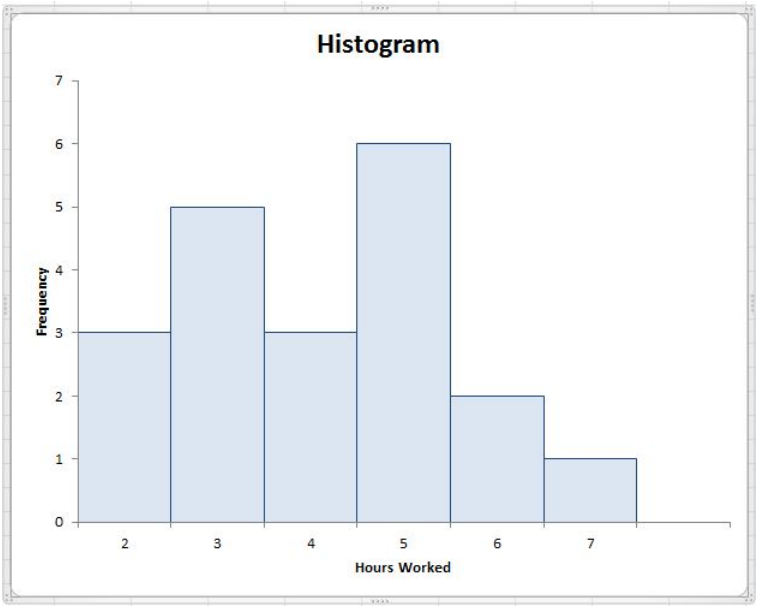
Let's revisit the example we looked at earlier of the twenty students and the number of hours worked each day.

Data Value	Frequency
2	3
3	5
4	3
5	6
6	2
7	1

We can represent a frequency distribution with a picture instead of a list of numbers, which is helpful to our visual-oriented brains. The graph we build from a frequency distribution is called a **histogram**. A histogram is a bar graph with a bar for each class in the frequency distribution, and the height of the bar represents the frequency listed for that class.

Note carefully: a **bar graph** has gaps between the bars. A **histogram** (what we do in this chapter) has no gaps.

The vertical axis is labeled either frequency or relative frequency, the horizontal axis is labeled with what the data represents (for instance, hours worked per day). The histogram from the working hours frequency distribution is below:



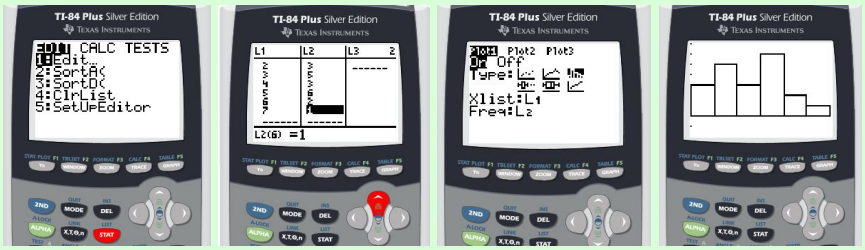
Using Your Calculator

The TI calculator will construct a histogram for you. You have two options: entering the frequency distribution into the calculator, or entering the raw data and letting the calculator find the frequency of each value.

Option 1: Enter the frequency distribution

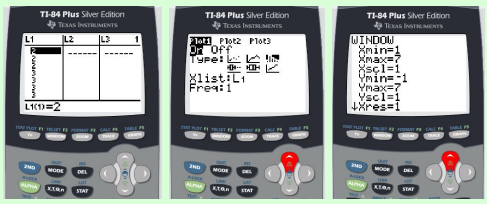
1. Press the **STAT** button to enter the statistics menu
2. Choose the **Edit** option to edit the data list
3. Enter the classes (or values if it isn't a grouped table) into L1 and enter the frequency into L2
4. Choose the **STATPLOT** option by clicking the **2nd** and **Y=** buttons
5. Turn Plot1 On and choose the histogram option, with the data (Xlist) in L1 and the frequency in L2
6. Press **ZOOM** and choose the "ZoomStat" option

By pressing the **TRACE** button, you can see the frequency of the bars.

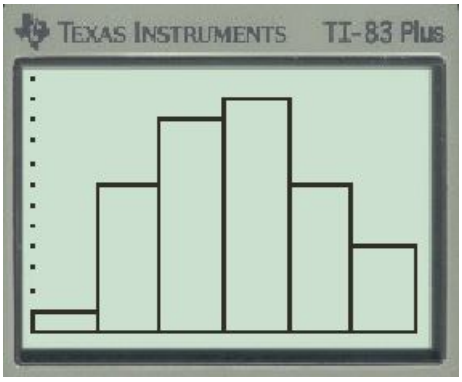


Option 2: Enter the raw data

Instead of using L1 and L2, simply enter the data into L1 and enter "1" for the Freq option when creating the histogram. You can change the width of the classes by changing Xscl under the **WINDOW** menu.



The graph below was created by the TI-83 using the data about getting ready in the morning, by entering the raw data in L1:



Stem-and-Leaf Plots

There is one major downside to grouped frequency distributions: some of the data gets lost in the summary. In other words, maybe in an example all we know is that there are 10 observations between 20 and 25, but we don't know exactly what all those observations are. This is an example of the trade-off between clarity and precision: often, the more precise we are, the less clear our summary will be.

To split the difference and display the data in a way that exhibits where it is clustered without losing any information about the data is to use a **stem-and-leaf** plot. Here, the data is grouped by tens; each tens value is a stem, and all the data points that have that tens value are listed as the leaves. We'll illustrate with an example, using the same data set we used to construct the grouped frequency distribution above, but this time without the decimal places:

35	28	25	23	23	32	29	19	21	13
24	26	25	31	30	20	25	29	37	26
32	36	18	17	15	24	21	16	19	30
38	27	22	24	28	17	31	32	21	28

The tens places are 1, 2, and 3, and each of them gets a category:

Stems	Leaves
1	
2	
3	

Finally we go through (carefully) and find each value that begins with a 1 and list the ones place of each of them under the first category, and similarly with the other two categories.

Stems	Leaves
1	3 5 6 7 7 8 9 9
2	0 1 1 1 2 3 3 4 4 4 5 5 5 6 6 7 8 8 8 9 9
3	0 0 1 1 2 2 2 5 6 7 8

Notice that we arranged the leaves in order; this isn't strictly necessary, but it makes the data a bit more orderly.

Once again, this stem-and-leaf plot illustrates where the data is clustered, as the length of each row of leaves is equivalent to the height of a bar on a histogram, but it does this without losing any information. In other words, if we were simply given the stem-and-leaf plot, we could completely recreate the data set.

For three-digit data values (or longer), the leaves are usually still the last digits (the unit digits), and the stems are everything before that. For instance, observe the data set below and the corresponding stem-and-leaf plot.

135	128	125	123	123	132	129	119	121	113
124	126	125	131	130	120	125	129	137	126

Stems	Leaves
11	3 9
12	0 1 3 3 4 5 5 5 6 6 8 9 9
13	0 1 2 5 7

Exercises 3.1

For problems 1–2, identify the population and sample.

1. A political scientist surveys 28 of the current 106 representatives in a state's congress. Of them, 14 said they were supporting a new education bill, 12 said they were not supporting the bill, and 2 were undecided.

2. The city of Frederick has 9500 registered voters. There are two candidates for the city council in an upcoming election: Marfani and Rahman. The day before the election, a telephone poll of 350 randomly selected registered voters was conducted. Of them, 112 said they would vote for Marfani, 207 said they would vote for Rahman, and 31 were undecided.

For exercises 3–6, identify the most relevant source of bias in the situation.

3. To determine opinions on voter support for a downtown renovation project, a surveyor randomly questions people working in downtown businesses.

4. To select a sample, a pollster calls every 100th name in the phone book.

5. Suppose pollsters call people at random, but once they have met their quota of 390 Democrats, they only gather people who do not identify themselves as a Democrat.

6. A survey seeks to investigate whether a new pain medication is safe to market to the public. They test by randomly selecting 300 men from a set of volunteers.

7. Fifty part-time students were asked how many courses they were taking this semester. The (incomplete) results are shown below. Fill in the blank cells to complete the table.

# of courses	Frequency	Relative Frequency
1	30	0.6
2	15	
3		

8. The following is the average daily temperature for Frederick, Maryland for the month of June:

$$74, 60, 58, 58, 64, 67, 64, 74, 72, 70,$$
$$78, 80, 80, 79, 80, 80, 70, 83, 76, 78,$$
$$81, 78, 81, 70, 70, 71, 66, 66, 68, 74.$$

(a) Construct a grouped frequency and relative frequency distribution using a class width of 5.

(b) Construct a histogram from the frequency distribution.

9. A researcher gathered data on hours of video games played by school-aged children and young adults. She collected the following data:

$$0, 0, 1, 1, 1, 2, 2, 3, 3, 3,$$
$$4, 4, 4, 4, 5, 5, 5, 6, 6, 7,$$
$$7, 7, 8, 8, 8, 8, 8, 9, 9, 9,$$
$$10, 10, 11, 12, 12, 12, 12, 13.$$

(a) Construct a grouped frequency and relative frequency distribution using 6 classes.

(b) Construct a histogram from the frequency distribution.

10. The following stem-and-leaf plots compare the ages of 30 actors and 30 actresses at the time they won the Oscar award for Best Actor or Actress.

Actors	Stems	Actresses
	2	146667
98753221	3	00113344455778
88776543322100	4	11129
6651	5	
210	6	011
6	7	4
	8	0

(a) What is the age of the youngest actor to win an Oscar?

(b) What is the age difference between the oldest and the youngest actress to win an Oscar?

(c) What is the oldest age shared by two actors to win an Oscar?

For exercises 11–14, use the frequency table below, which contains the total number of deaths worldwide as a result of earthquakes for the period from 2000 to 2012.

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	**823,356**

11. What is the frequency of deaths measured from 2006 through 2009?

12. What percentage of deaths occurred after 2009 (from 2010 onwards)?

13. What is the relative frequency of deaths that occurred in 2003 or earlier?

14. What is the percentage of deaths that occurred in 2004?

15. What is wrong with the following grouped frequency distribution?

Grades	Frequency
50–55	2
55–60	4
60–70	9
70–80	15
80–90	7
90 and above	4

(a) The classes do not all have the same width.

(b) The classes overlap.

(c) There are open-ended classes.

(d) All of the above.

SECTION 3.2 Measures of Center

What comes to mind when you think of the word "average?" You may think of situations where the average is used: average income, average height, average number of Facebook friends, average test score, etc. The average is a type of center. It tells us the middle of the data. We call the average a **measure of center**, because it gives a quick snapshot of what a *typical* data point is. But the average is not the only measure of center. In this section, we'll examine several measures of center, how to calculate them and when best to use them.

The Mean

The **mean** is what most people call the average. To find the mean, you add all the values and divide by the number of values. We can write this in mathematical notation. The Greek letter sigma Σ is the symbol for sum. When you see Σx, that stands for "the sum of x." The symbol for the mean is $\overline{x}$, read "x bar." Using this information, we can devise a formula for the mean.

Mean

The mean of a sample of size n is

$$\overline{x} = \frac{\sum x}{n}.$$

The mean of a population of size N is

$$\mu = \frac{\sum x}{N}.$$

We must note that the symbol for the sample mean is $\overline{x}$, whereas the symbol for population mean is μ (the lowercase Greek letter, read "mu" and pronounced "mew"). You can assume that the data sets given throughout this section are all samples; hence the mean will be $\overline{x}$.

EXAMPLE 1 **AIDS PATIENTS**

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest):

3	4	8	8	10	11	12	13	14	15
15	16	16	17	17	18	21	22	22	24
24	25	26	26	27	27	29	29	31	32
33	33	34	34	35	37	40	44	44	47

Calculate the mean.

Solution

The calculation for the mean is

$$\overline{x} = \frac{3+4+8+8+10+11+12+\ldots+40+44+44+47}{40} = 23.575 \text{ months.}$$

To take into account repeated values, we can rewrite the formula. Since the value 8 is repeated twice, instead of using 8+8, we can write (2)(8), and similarly for other repeated values. Hence, we would get:

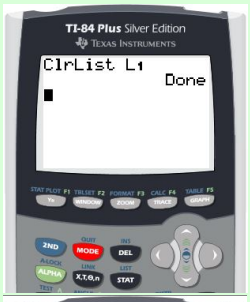
$$\overline{x} = \frac{3+4+(2)(8)+10+\ldots+40+(2)(44)+47}{40} \approx 23.6 \text{ months.}$$

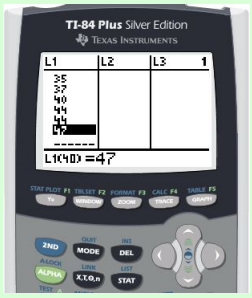
Notice that the answer in this alternate solution is rounded to one decimal place.

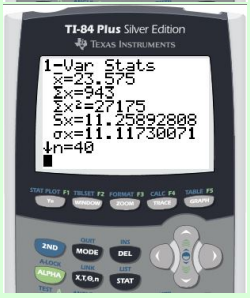
Using Your Calculator

The TI calculator can find the mean for you.

1. To enter data into the list editor, press **STAT** then choose option 1:Edit, then enter the data values into list L1.
2. Press **STAT** and use the arrow button to navigate to the right to the CALC menu.
3. Choose option 1:1-VarStats. Press the **2ND** button, then **1** for list L1. Press **ENTER**.







TRY IT

The following data show the number of months patients typically wait on a transplant list before getting surgery. Calculate the mean using your calculator.

3	4	5	7	7	7	7	8	8	9
9	10	10	10	10	11	12	12	12	13
14	14	15	15	17	17	18	19	19	19
21	21	22	22	23	24	24	24	24	

The mean can also be calculated from a frequency distribution (as long as it is not grouped). To illustrate this, consider the following frequency distribution of the data in the example from the previous section about how many hours students worked:

Data Value	Frequency
2	3
3	5
4	3
5	6
6	2
7	1

Recall that this means that there are three 2's, five 3's, and so on. To calculate the mean, we need to add them all up, but we can do this by multiplying each value by how often it occurs and adding up those results (note that this is exactly what we did at the end of the example about AIDS patients).

$$\overline{x} = \frac{2(3) + 3(5) + 4(3) + 5(6) + 6(2) + 7(1)}{20} = 4.1$$

The Median

We noted at the beginning of this section that we have several measures of center, or ways to describe what a typical data point is. Why is this necessary? Since the mean is easy to calculate, why don't we always just use that?

To illustrate why, let's look at the following data set, which shows the income for a sample of 5 people:

$30,000 $45,000 $50,000 $52,000 $1,000,000

Calculating the mean for this data set yields $235,400. Does the mean give a true picture for the center of the data? Can I rightfully say the average person in my data set earns roughly $235,000? Obviously, the answer is no, since 80%–or 4/5–of the people in this group earn less than $53,000. When we have an **outlier**–a number far removed from the majority of data values–in our data, we should use the median instead of the mean to give a measure of center.

The **median** is the middle value of a data set when the data is ordered. It is denoted with a capital letter M. To find the median, sort the data in order from smallest to largest. If there are an odd number of values in the data set, then the median is the middle value. If there are an even number of values in the data set, then the median is the average or mean of the two middle values.

EXAMPLE 2 **AIDS PATIENTS**

Let's return to our AIDS data set. The data has already been arranged in order from smallest to largest:

3	4	8	8	10	11	12	13	14	15
15	16	16	17	17	18	21	22	22	24
24	25	26	26	27	27	29	29	31	32
33	33	34	34	35	37	40	44	44	47

Let's find the median.

Solution Since there are an even number of values, we have to take the average of the two middle values. Because there are 40 values, the two middle values are the 20th and 21st values.

3	4	8	8	10	11	12	13	14	15
15	16	16	17	17	18	21	22	22	**24**
24	25	26	26	27	27	29	29	31	32
33	33	34	34	35	37	40	44	44	47

Here, the 20th and 21st values are both 24. The average of 24 and 24 is 24. So the median is

$$M = 24 \text{ months.}$$

Note that, like the mean, the units of the median are the same as the original data values.

You can also use the TI Calculator to find the median. The steps are the same as to find the mean; just use the arrow keys to scroll down to the median.

Where's the Median?

To find the median, we want to be able to quickly figure out what position to count to in the ordered data set. To do this, calculate

$$\frac{n+1}{2},$$

where n is the size of the data set. Notice that this is the average of 1 and n, so it makes sense that this would be halfway between them.

- If n is odd, $\frac{n+1}{2}$ is a whole number. Count to that position and there you'll find the median.
- If n is even, $\frac{n+1}{2}$ is halfway between two whole numbers. Find the average of the values at those two positions.

For instance, in the example above, n was 40, so we calculated $\frac{n+1}{2} = 20.5$, so we knew that the median would be between positions 20 and 21 (the average of the numbers in those positions).

TRY IT

The following data shows the number of months patients typically wait on a transplant list before getting surgery. Calculate the median either by hand or using your calculator.

3	4	5	7	7	7	7	8	8	9
9	10	10	10	10	11	12	12	12	13
14	14	15	15	17	17	18	19	19	19
21	21	22	22	23	24	24	24	24	

Just like we did with the mean, we can calculate the median from a frequency distribution. Let's use the same frequency distribution we did before.

Data Value	Frequency
2	3
3	5
4	3
5	6
6	2
7	1

To find the median, we calculate what position it will occupy in the data set. Since there are 20 data points, $\frac{n+1}{2} = 10.5$, so the median will be between the 10th and 11th positions.

If we count through the ordered data set, we go through three 2's, five 3's, and then three 4's, which brings us to the 11th position. The 10th and 11th positions are both occupied by 4's, so $M = 4$.

Outliers: Let's revisit our income data set:

$30,000 $45,000 $50,000 $52,000 $1,000,000

The median is $50,000, which is a more accurate measure of center than the mean. This shows the mean is *sensitive* to outliers, whereas the median is *resistant* to outliers.

Mean and Median

When outliers are present, the median is a better measure of center. When outliers are absent, the mean can be used.

EXAMPLE 3 **SIBLINGS**

A dozen people were asked how many siblings they have. The data is as follows:

0 0 1 1 2 2 4 4 5 5 6 6

Find the mean and median. Then write a sentence or two explaining why the data values result in those particular mean and median.

Solution The mean is $\overline{x} = 3$ siblings and the median is also $M = 3$ siblings. This is because there are an equal number of high and low values and they are evenly spaced out.

The sibling data above is called **symmetric**. Symmetric data is balanced around the mean. A data set is symmetric if the mean and median are roughly equal.

EXAMPLE 4 **SALARIES**

Suppose that in a small town of 50 people, one person earns $5,000,000 per year and the other 49 each earn $30,000. Which is the better measure of the "center": the mean or the median?

Solution Since an outlier ($5,000,000) is present, the median would be a better measure of center. This is why you'll often find the median salary in a region reported rather than the mean.

When trying to decide whether to use the mean or the median as the measure of the center of a data set, compare them to decide if they are drastically different. Of course, the best policy is to report both of them and compare them to determine whether the data set is symmetric or *skewed*.

The Mode

There is another, less used, measure of center: the mode. The **mode** is the most frequently occurring value. There can be more than one mode in a data set as long as those values have the same frequency and the frequency is the highest. A data set with two modes is called **bimodal**.

EXAMPLE 5 **EXAM SCORES**

Statistics exam scores for 20 students are as follows:

50 53 59 59 63 63 72 72 72 72
72 76 78 81 83 84 84 84 90 93

Find the mode.

Solution The most frequent score is 72, which occurs five times.

Mode = 72.

TRY IT

The number of books checked out from the library from 25 students are as follows:

0 0 0 1 2 3 3 4 4 5
5 7 7 7 7 8 8 8 8 9
10 10 11 11 12

Find the mode.

The mode can also be observed by looking at a frequency distribution (look for the category with the highest frequency) or a histogram (look for the tallest bar).

Exercises 3.2

1. Which is the greatest of the following data set: the mean, median or mode?

$$11, 11, 12, 12, 12, 12, 13, 15, 17, 22, 22, 22$$

2. Which is the greatest of the following data set: the mean, median or mode?

$$80, 83, 83, 87, 87, 87, 91, 97, 98$$

3. One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. The 12 change scores are as follows:

$$3, 8, -1, 2, 0, 5, -3, 1, -1, 6, 5, -2$$

(a) What is the mean change score?

(b) What is the median change score?

(c) What is the best measure of center for this data set: the mean or the median? Why?

4. The following are the amounts of total fat (in grams) in different kinds of sweet treats available at your local donut shop:

$$16, 17, 16, 13, 15, 17, 16, 14, 15, 17,$$
$$18, 18, 16, 16, 15, 20, 22, 19, 25, 15, 15.$$

(a) What is the mean amount of fat?

(b) What is the median amount of fat?

(c) What is the best measure of center for this data set: the mean or the median? Why?

5. In a neighborhood donut shop, one type of donut has 530 calories, three types of donuts have 330 calories, four types of donuts have 320 calories, seven types of donuts have 410 calories, and five types of donuts have 380 calories. Find the mean and median calories of the donuts.

6. In a recent issue of *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted 2 days, 36 lasted 3 days, 18 lasted 4 days, 19 lasted 5 days, 4 lasted 6 days, 1 lasted 7 days, 1 lasted 8 days, and 1 lasted 9 days. Find the mean and median length (in days) of an engineering conference.

7. A researcher gathered data on hours of video games played by school-aged children and young adults. She collected the following data:

$$0, 0, 1, 1, 1, 2, 2, 3, 3, 3,$$
$$4, 4, 4, 4, 5, 5, 5, 6, 6, 7,$$
$$7, 7, 8, 8, 8, 8, 8, 9, 9, 9,$$
$$10, 10, 11, 12, 12, 12, 12, 13.$$

Find the mean and median number of hours.

8. The following are the sizes in square feet of the seventeen new faculty offices in the mathematics department at Frederick Community College:

$$202, 120, 120, 120, 137, 167, 122, 111, 109,$$
$$108, 102, 108, 103, 103, 103, 106, 127.$$

Find the mean and median square footage.

In exercises 9–10, find the mean, median, and mode for each data set. If there is no mode, state so.

9. 83, 83, 87, 80, 91, 87, 97, 98

10. 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 5

In exercises 11–14, find the mean and median. Then write a sentence or two explaining why the data values result in those particular means and medians.

11. 0, 0, 1, 1, 2, 2, 6, 6, 7, 7, 8, 8

12. 0, 0, 1, 1, 2, 2, 6, 7, 8

13. 0, 0, 1, 1, 8

14. 1, 1, 1, 1, 1, 1, 8

15. Fifteen students took a statistics pre-test and post-test. The results are below. Find the mean and median pre-test and post-test scores. Then write a sentence or two about what the mean and median tell you in this case. A score of 28 is considered a perfect score.

Student	Pre-test	Post-test
1	9	22
2	11	28
3	9	18
4	4	24
5	10	25
6	11	16
7	9	19
8	9	20
9	7	18
10	8	14
11	5	16
12	15	26
13	12	21
14	0	14
15	9	13

16. Below are the ages and salaries (in thousands of dollars) for CEOs of small companies. Find the mean and median age and salary. Then write a sentence or two about what the mean and median tell you. Data from: http://lib.stat.cmu.edu/DASL/Datafiles/ceodat.html.

CEO	Age	Salary
1	53	145
2	43	621
3	33	262
4	45	208
5	46	362
6	55	424
7	37	300
8	41	339
9	55	736
10	36	291
11	45	58
12	55	498
13	50	643
14	49	390
15	47	332
16	69	750

17. In a fifth-grade class, the teacher was interested in the average age of her students. The following data are the ages of a sample of 20 fifth-graders. The ages are rounded to the nearest half-year. Find the average age for this sample:

$$9, 9.5, 9.5, 10, 10, 10, 10, 10.5, 10.5, 10.5, 10.5, 11, 11, 11, 11, 11, 11, 11.5, 11.5, 11.5.$$

18. An experiment compared the ability of three groups of participants to remember briefly-presented chess positions. The data are shown below. The numbers represent the number of pieces correctly remembered from three chess positions. Find the mean and median for each group. Then explain what the mean and median tell you about each group's ability to correctly remember chess positions.

Non-players:	22.1	22.3	26.2	29.6	31.7	33.5	38.9	39.7	43.2	43.2
Beginners:	32.5	37.1	39.1	40.5	45.5	51.3	52.6	55.7	55.9	57.7
Tournament players:	40.1	45.6	51.2	56.4	58.1	71.1	74.9	75.9	80.3	85.3

SECTION 3.3 Measures of Spread

What if I told you the average age for a group of 5 people was 25? Without giving you any more information, you would not be able to guess the five individual ages. One possible set of ages could be 10, 10, 25, 40, 40. Another possible set of ages could be 25, 25, 25, 25, 25. Yet another possible set of ages could be 5, 5, 5, 10, 100. The mean for each of these data sets is 25, yet the individual values are very different. This is where the spread of the data becomes very important. If we know how much spread there is in the data, we can have a much better idea of what the data set really looks like. In this section, you will learn the different measures of spread.

The Range

The **range** is the simplest measure of spread. It is simply the highest data value minus the lowest data value, or the maximum minus the minimum.

Range

The range is $Maximum - Minimum$

AGES **EXAMPLE 1**

Let's examine our group of five people again. Below are their possible ages. Calculate the range for each of the data sets below.

a. 10, 10, 25, 40, 40

b. 25, 25, 25, 25, 25

c. 5, 5, 5, 10, 100

Solution

a. $Range = Max - Min = 40 - 10 = 30$

b. $Range = Max - Min = 25 - 25 = 0$

c. $Range = Max - Min = 100 - 5 = 95$

As you can see from the example above, even though each data set has a mean of 25, the ranges are wildly different. The higher the range, the more spread out the data. The first data set has some spread, the second data set has no spread (hence a range of 0), and the third data set has a lot of spread.

TRY IT

The number of books checked out from the library from 25 students are as follows:

0	0	0	1	2	3	3	4	4	5
5	7	7	7	7	8	8	8	8	9
10	10	11	11	12					

Find the range.

Unfortunately, the TI Calculator will not calculate the range for you. However, the 1-VAR-STATS option will give you the maximum and minimum data values, from which you can easily get the range.

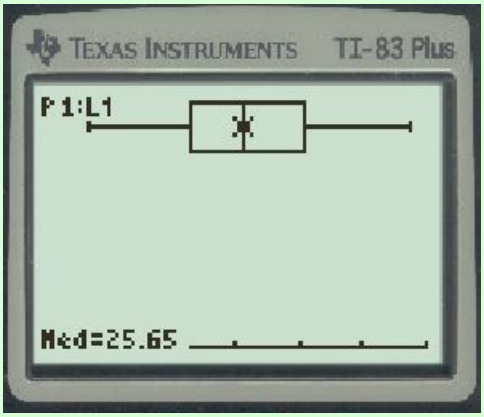
The Five Number Summary

Another way to observe how a data set is spread out is to calculate the **quartiles**. The quartiles are similar to the median: the 1st quartile is the data value that is a quarter of the way through the set; the 2nd quartile is the median, halfway through the set; and the 3rd quartile is three quarters of the way through the set. If you look at the 1-VAR-STATS on the TI calculator, you'll notice Q_1 (the 1st quartile) and Q_3 (the 3rd quartile) listed.

If you put the three quartiles together with the minimum and maximum of the data set, these form what is known as the **five number summary**. They split the data into quarters (Min $\to Q_1$, $Q_1 \to Q_2$, etc.) where each quarter contains a fourth of the data points. This can provide a good picture of whether the data is clustered on the lower end or on the higher end, or whether it is symmetric or clustered in the middle.

Boxplots

We can display the five number summary visually with a **boxplot**. A boxplot consists of a box drawn from the first to the third quartile, with a line at the median, or second quartile. Lines, or whiskers, extend from this box down to the minimum and up to the maximum (we can also use the quartiles to find outliers and mark them on a boxplot, but we'll omit that here).
The TI calculator can do this for you, if you choose the boxplot option when setting up the STAT PLOT.



By using the TRACE function, you can see the minimum, maximum, and quartiles.

The Standard Deviation

The **standard deviation** is a number that measures how far a typical data value is from the mean. The standard deviation is always positive or zero. A small standard deviation means less spread in the data; a large standard deviation means more spread in the data.

First of all, the **deviation** of each data point x is its difference from the mean $\overline{x}$:

$$x - \overline{x}.$$

Each value in the data set has a deviation associated with it. If we want to find how far, on average, each data point is from the mean, it would make sense to take the average of the deviations. However, there's a problem with doing that: if we add up the deviations, we'll always get 0, because of the way that $\overline{x}$ is calculated. Some of the deviations are positive, some are negative, and the positives cancel out the negatives.

To get around this, we square the deviations so that everything becomes positive. NOW, when we take an average[1] of these *squared* deviations, we get a meaningful number, instead of getting 0 every time. We call this "average" the variance:

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

It is not quite an average, since we divide by $n - 1$ instead of n. The reasons for this are complicated, but they have to do with making the sample variance be what is called an unbiased estimator for the population variance. You can ignore this note except to remember to divide by $n - 1$.

The only problem now is that we've got an average of these squared things, so the units of our answer are not the same units we started with. In other words, if the data is given in units

[1] *almost*

of inches, we've got a variance in square inches. To get an answer, we take the square root of the variance, and that's what we call the standard deviation.

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

To calculate the standard deviation,

1. Calculate the mean of the data values, $\overline{x}$.
2. Subtract the mean from each data value to find the deviations:

$$\text{deviation} = x - \overline{x}$$

3. Square each deviation: $(x-\overline{x})^2$
4. Take the sum of the squared deviations: $\Sigma(x-\overline{x})^2$
5. Divide that sum by n minus 1, where n is the number of data values: $\frac{\Sigma(x-\overline{x})^2}{n-1}$
6. Take the square root of this quotient: $\sqrt{\frac{\Sigma(x-\overline{x})^2}{n-1}}$

Standard Deviation

The standard deviation of a sample is given by

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

where n stands for the number of data values and x stands for each data value.

AT THE MALL EXAMPLE 2

Kari went shopping and bought five things. The prices are as follows:

\$20 \$4 \$15 \$9 \$3

Let's calculate the standard deviation for this data.

$$\overline{x} = (20+4+15+9+3)/5 = 10.20$$

The mean is \$10.20. We can use the table below to get the standard deviation.

Data Value x	**Deviations** $(x-\overline{x})$	**Deviations**^{2} $(x-\overline{x})^2$
20	$20-10.20=9.8$	$(9.8)^2=96.04$
4	$4-10.20=-6.2$	$(-6.2)^2=38.44$
15	$15-10.20=4.8$	$(4.8)^2=23.04$
9	$9-10.20=-1.2$	$(-1.2)^2=1.44$
3	$3-10.20=-7.2$	$(-7.2)^2=51.84$

Adding up all the values in the third column yields 210.8. The variance, s^2, is equal to this sum divided by the total number of data values minus one.

$$s^2 = \frac{210.8}{5-1} = \$52.7.$$

The standard deviaion s is equal to the square root of the variance.

$$s = \sqrt{52.7} = \$7.26$$

As you can see, the calculation for standard deviation is very tedious. For larger data sets, the calculations get even more tedious. Fortunately, your calculator can easily compute the standard deviation. Use the 1-VAR-STATS option on the TI calculator (the same we used to find the mean and median) and scroll down. The standard deviation is denoted by s_x.

Let's revisit our income data set that we saw in the previous section. We'll call it

"Income Data A:" \$30,000 \$45,000 \$50,000 \$52,000 \$1,000,000

Using your calculator, you should get a standard deviation of approximately

$$s_x = \$427,511.$$

That is a very large standard deviation. This tells us the data is VERY spread out, which makes sense given we have an extreme outlier of \$1,000,000.

Let's change that outlier to be closer to the other data values. Find the standard deviation for the following data set, which we'll call

"Income Data B:" \$30,000 \$45,000 \$50,000 \$52,000 \$55,000

You should have gotten an approximate standard deviation of

$$s_x = \$9,864.$$

Compared to the the previous standard deviation, this new data set is a LOT less spread out, so it has a smaller standard deviation. Another word for the "spread" is **variability**. So we can say that Income Data A has more variability than Income Data B.

The units for both the range and standard deviation are the same as the original data values.

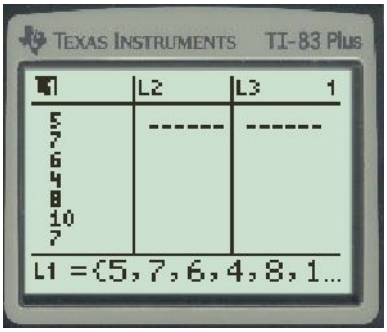
EXAMPLE 3 QUIZ SCORES

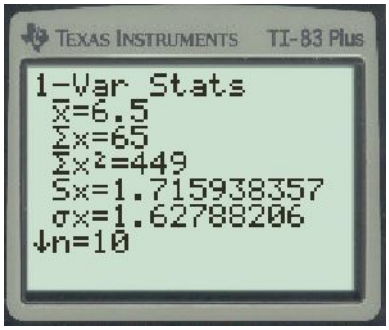
Find the standard deviation of the following quiz scores:

5	7	6	4	8
10	7	7	6	5

We could calculate s_x by hand by calculating the deviations, squaring them, "averaging" them, and taking the square root, but we can also do it more quickly and easily with our calculator.

We enter the data, have it calculate the 1-VAR-STATS, and scroll down to s_x.





The standard deviation is approximately 1.72.

TRY IT

The prices of a jar of peanut butter at five stores are shown below.

\$3.29 \$3.59 \$3.79 \$3.75 \$3.99

Calculate the standard deviation of this data set.

Exercises 3.3

1. Which is the greatest of the following data set: the range or the standard deviation?

$$11, 11, 12, 12, 12, 12, 13, 15, 17, 22, 22, 22$$

2. Which is the greatest of the following data set: the range or the standard deviation?

$$80, 83, 83, 87, 87, 87, 91, 97, 98$$

3. One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. A positive number for change in attitude indicates that a teacher's attitude toward math became more positive. The 12 change scores are as follows:

$$3, 8, -1, 2, 0, 5, -3, 1, -1, 6, 5, -2$$

(a) What is the range of the scores?

(b) What is the standard deviation of the scores?

4. The following are the amounts of total fat (in grams) in different kinds of sweet treats available at your local donut shop:

$$16, 17, 16, 13, 15, 17, 16, 14, 15, 17,$$
$$18, 18, 16, 16, 15, 20, 22, 19, 25, 15, 15.$$

(a) What is the range for this data set?

(b) What is the standard deviation?

5. In a neighborhood donut shop, one type of donut has 530 calories, three types of donuts have 330 calories, four types of donuts have 320 calories, seven types of donuts have 410 calories, and five types of donuts have 380 calories. Find the range and standard deviation of the calories of the donuts.

6. In a recent issue of *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted 2 days, 36 lasted 3 days, 18 lasted 4 days, 19 lasted 5 days, 4 lasted 6 days, 1 lasted 7 days, 1 lasted 8 days, and 1 lasted 9 days. Find the range and standard deviation of length (in days) of an engineering conference.

7. A researcher gathered data on hours of video games played by school-aged children and young adults. She collected the following data:

$$0, 0, 1, 1, 1, 2, 2, 3, 3, 3,$$
$$4, 4, 4, 4, 5, 5, 5, 6, 6, 7,$$
$$7, 7, 8, 8, 8, 8, 8, 9, 9, 9,$$
$$10, 10, 11, 12, 12, 12, 12, 13.$$

(a) Find the range.

(b) Find the standard deviation.

(c) Find the five-number summary.

8. The following are the sizes in square feet of the seventeen new faculty offices in the mathematics department at Frederick Community College:

$$202, 120, 120, 120, 137, 167, 122, 111, 109,$$
$$108, 102, 108, 103, 103, 103, 106, 127.$$

(a) Find the range.

(b) Find the standard deviation.

(c) Find the five-number summary.

In exercises 9–10, find the five-number summary for each data set.

9. 83, 83, 87, 80, 91, 87, 97, 98

10. 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 5

In exercises 11–14, find the range and standard deviation. Then write a sentence or two explaining what those values tell you about the spread of the data.

11. 0, 0, 1, 1, 2, 2, 6, 6, 7, 7, 8, 8

12. 0, 0, 1, 1, 2, 2, 6, 7, 8

13. 0, 0, 1, 1, 8

14. 1, 1, 1, 1, 1, 1, 8

15. Fifteen students took a statistics pre-test and post-test. The results are below. Find the range and standard deviation of the pre-test and post-test scores. Then write a sentence or two about what that tells you about the spread of the data. A score of 28 is considered a perfect score.

Student	Pre-test	Post-test
1	9	22
2	11	28
3	9	18
4	4	24
5	10	25
6	11	16
7	9	19
8	9	20
9	7	18
10	8	14
11	5	16
12	15	26
13	12	21
14	0	14
15	9	13

16. Below are the ages and salaries (in thousands of dollars) for CEOs of small companies. Find the range and standard deviation of age and salary. Then write a sentence or two about what this tells you about the spread of the data. Data from: http://lib.stat.cmu.edu/DASL/Datafiles/ceodat.html.

CEO	Age	Salary
1	53	145
2	43	621
3	33	262
4	45	208
5	46	362
6	55	424
7	37	300
8	41	339
9	55	736
10	36	291
11	45	58
12	55	498
13	50	643
14	49	390
15	47	332
16	69	750

17. In a fifth-grade class, the teacher was interested in the range of ages of her students. The following data are the ages of a sample of 20 fifth-graders. The ages are rounded to the nearest half-year.

$$9, 9.5, 9.5, 10, 10, 10, 10, 10.5, 10.5, 10.5, 10.5, 11, 11, 11, 11, 11, 11, 11.5, 11.5, 11.5.$$

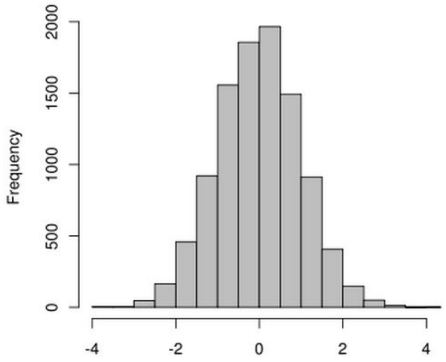
Find the range, the standard deviation, and the five-number summary.

18. An experiment compared the ability of three groups of participants to remember briefly-presented chess positions. The data are shown below. The numbers represent the number of pieces correctly remembered from three chess positions. Find the range and standard deviation for each group. Then explain what this tells you about each group's ability to correctly remember chess positions.

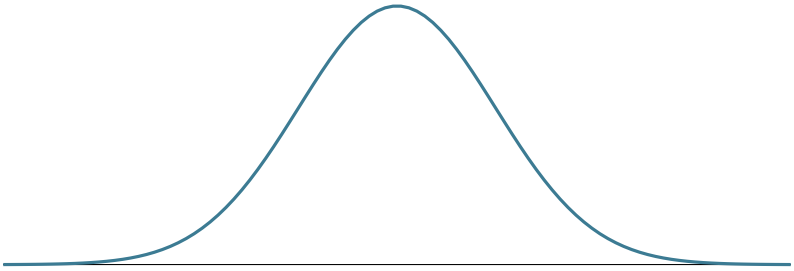
Non-players:	22.1	22.3	26.2	29.6	31.7	33.5	38.9	39.7	43.2	43.2
Beginners:	32.5	37.1	39.1	40.5	45.5	51.3	52.6	55.7	55.9	57.7
Tournament players:	40.1	45.6	51.2	56.4	58.1	71.1	74.9	75.9	80.3	85.3

SECTION 3.4 The Normal Distribution

Suppose you had the following histogram of a frequency distribution:



If we connected the bars via a smooth line, we would get something like this:



This is known as a **bell-shaped curve**. It is the most widely used and abused graph in many disciplines, from psychology, business and economics to science and medicine. This curve is also known as a "Gaussian Curve," named after Carl Friedrich Gauss, the famous mathematician. The bell-shaped curve is the graph of the **normal distribution**, the most important of all distributions in statistics.

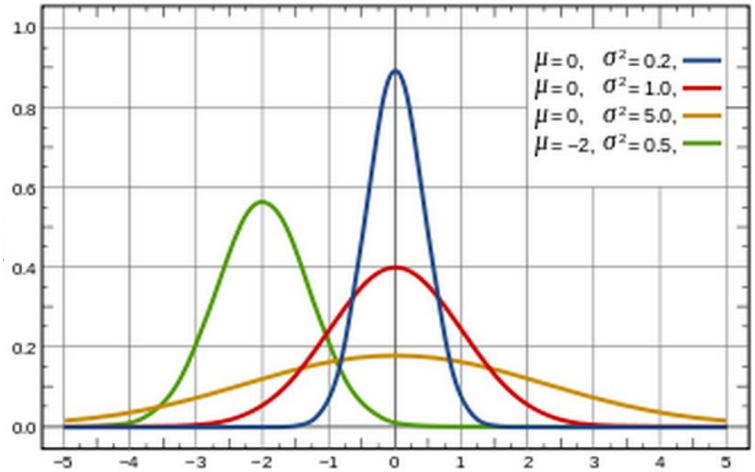
Notice how the bell-shaped curve is symmetric about a central line. The center of the curve is the mean. But it's also the median and the mode. Hence, for all bell-shaped curves, the mean is equal to the median which is equal to the mode.

C.F.Gauss

What kind of data does the bell-shaped curve describe? Data that has many middle or average values and fewer high or low values. Think of men's heights. There are many men of average height, yet few very tall and few very short men. The same applies to women's heights. Hence, height is a variable that is **normally distributed**.

IQ is also normally distributed. Most people have average IQs. There are very few people with extremely high IQs and very few with extremely low IQs.

There are many variables that are normally distributed. Any variable that has a bell-shaped distribution is said to be normally distributed. Different variables have different bell-shaped curves.



How the curve looks depends on the mean and standard deviation. Changing the mean shifts the curve right and left. Changing the standard deviation changes the shape of the curve: a large standard deviation results in a wide curve and a small standard deviation results in a narrow curve.

The blue curve in the above diagram has a mean of 0, whereas the green curve has a mean of -2. Since the green curve's mean is lower, it is to the left of the blue curve (think about where the numbers are on the number line).

The standard deviation of the blue curve is 0.45. The standard deviation of the green curve is 0.71. Since the standard deviation of the green curve is larger, it is wider compared to the blue curve.

On the same diagram, notice the red curve. It has a mean of 0 and a standard deviation of 1. This is a special curve known as the **standard normal curve**.

Z-Scores

The normal distribution is precisely defined enough that if we have some information about a particular normally distributed quantity–specifically its mean and standard deviation–we can tell precisely where a specific data point falls in that distribution. In other words, we can answer questions like: is it unusual for a man to be over 6'4"? Just *how* unusual?

We can describe the position of a data point in a normal distribution using its **z-score**, which measures by how many standard deviations a data point differs from the mean.

For instance, suppose a particular data set is normally distributed with a mean of 100 and a standard deviation of 10. Then a data point of 110 has a z-score of 1, a data point of 150 has a z-score of 5, and a data point of 70 has a z-score of -3 (negative because it is *below* the mean).

To find the z-score for a data point in a data set with a known mean and standard deviation, we just need to find its distance from the mean, and then divide that by the standard deviation to find out how many steps it will take from the mean to reach it.

Z-Scores

If x is a data value in a data set with mean $\overline{x}$ and standard deviation s, the z-score that corresponds to that data value is

$$z\text{-score} = \frac{\text{data value} - \text{mean}}{\text{standard deviation}}$$

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

EXAMPLE 1 **FEMALE HEIGHTS**

Female adult height is normally distributed with a mean of 65 in. and a standard deviation of 3.5 in.

Find the z-scores of the following heights:

(a) 58 in.

(b) 71 in.

Solution

(a) The z-score corresponding to 58 in. is

$$z = \frac{58 - 65}{3.5} = -2$$

(b) The z-score corresponding to 71 in. is

$$z = \frac{71 - 65}{3.5} = 1.71$$

Thus, a woman at 71 in. tall is 1.71 standard deviations above the mean, while a woman at 58 in. is 2 standard deviations below the mean.

TRY IT

Scores on the SAT and ACT are normally distributed:

Test	Mean	Std. Deviation
SAT	500	100
ACT	18	6

You score 550 on the SAT and 24 on the ACT. On which test did you have a better score, relative to everyone else who took the test?

EXAMPLE 2

WORKING BACKWARD FROM Z-SCORES

Scores on an IQ test are normally distributed with a mean of 100 and a standard deviation of 15. Find the IQ score that corresponds to each of the following z-scores.

(a) -1.5

(b) 2.05

Solution

Recall that $z = \dfrac{\text{data value} - \text{mean}}{\text{standard deviation}}$

(a) If the z-score is -1.5:

$$-1.5 = \frac{\text{IQ} - 100}{15} \longrightarrow -22.5 = \text{IQ} - 100 \longrightarrow \text{IQ} = 77.5$$

(b) If the z-score is 2.05:

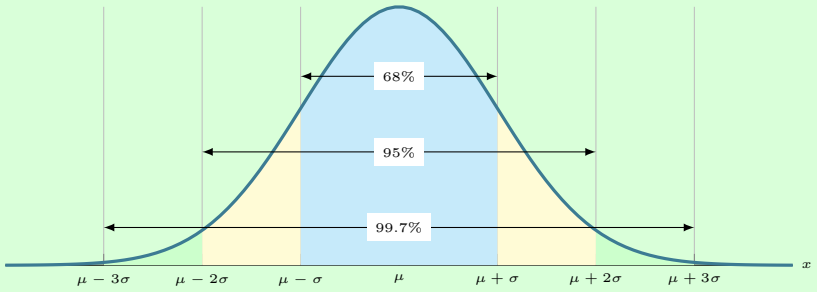
$$2.05 = \frac{\text{IQ} - 100}{15} \longrightarrow 30.75 = \text{IQ} - 100 \longrightarrow \text{IQ} = 130.75$$

The Empirical Rule

Not only do the mean and standard deviation define the normal distribution, they play a vital part in the Empirical Rule. This important rule allows us to determine whether a particular data point is unusual or not by comparing it to where most of the data falls. The Empirical Rule gives a precise prediction about where the data is distributed.

The Empirical Rule

Approximately 68% of the data is within **one** standard deviation of the mean. Approximately 95% of the data is within **two** standard deviations of the mean. Approximately 99.7% of the data is within **three** standard deviations of the mean.



Note that this diagram uses μ for the population mean (as opposed to $\overline{x}$ for the sample mean) and σ for the population standard deviation (as opposed to s for the sample standard deviation).

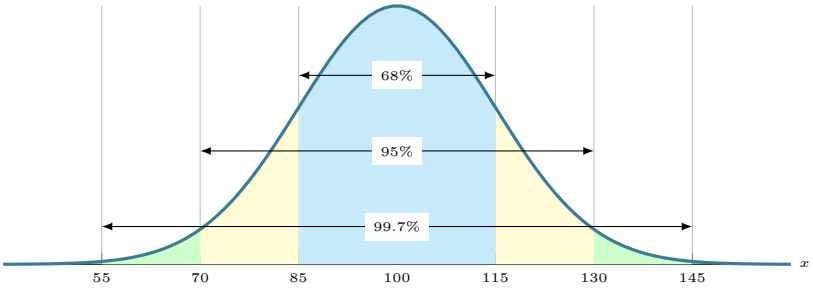
As you can see, another name for the Empirical Rule is the "68-95-99.7 Rule." Let's use this rule to look at IQ in the population.

EXAMPLE 3 THE INTELLIGENCE QUOTIENT

IQ is normally distributed with a mean of 100 and a standard deviation of 15. Use the Empirical Rule to the find the data that is within one, two, and three standard deviations of the mean.

Solution

1. 68% of the data is within one standard deviation of the mean.
   IQ = mean $\pm$(1· standard deviation) = $100 \pm (1 \cdot 15) = 100 \pm 15 = (85, 115)$
   Thus, 68% of people have an IQ between 85 and 115.
2. 95% of the data is within two standard deviations of the mean.
   IQ = mean $\pm$(2· standard deviation) = $100 \pm (2 \cdot 15) = 100 \pm 30 = (70, 130)$
   Thus, 95% of people have an IQ between 70 and 130.
3. 99.7% of the data is within three standard deviations of the mean.
   IQ = mean $\pm$(3· standard deviation) = $100 \pm (3 \cdot 15) = 100 \pm 45 = (55, 145)$
   Thus, 99.7% of people have an IQ between 55 and 145.



Again, this rule gives a way to decide whether a data point is unusual or not. An IQ of over 130 is very unusual, and an IQ of over 145 is even more so.

Since 99.7% have IQs in the range from 55 to 145, only 0.3% of people have IQs outside that range. Since the bell curve is symmetric, half of those, or 0.15% of people (15 people out of 1000) have IQs over 145.

TRY IT

The mean height of boys 15 to 18-years old from Chile is 170 cm with a standard deviation of 6 cm. Male heights are known to be normally distributed. Using the Empirical Rule, find the range of heights that contain approximately 68%, 95%, and 99.7% of the data.

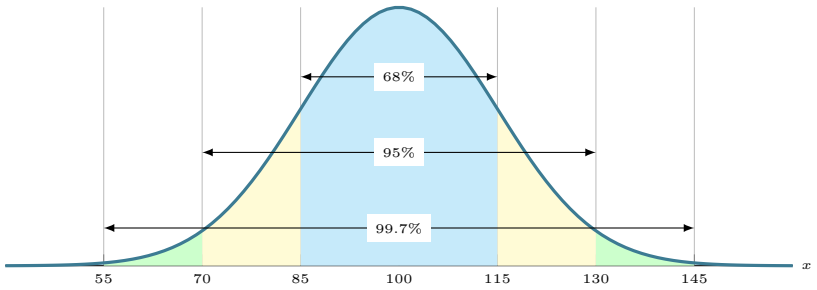
EXAMPLE 4 COLLEGE ENTRANCE EXAMS

The scores on a college entrance exam are normally distributed with a mean of 52 points and a standard deviation of 11 points. About 95% of the values lie between what two scores?

Solution

We know 95% of the data is within two standard deviations of the mean.
Scores = Mean $\pm$ 2· Standard Deviation $= 52 \pm 2 \cdot 11 = 52 \pm 22 = (30, 74)$.
Hence, 95% of the values fall between a score of 30 and a score of 74.

Let's go back to the figure from the first example:

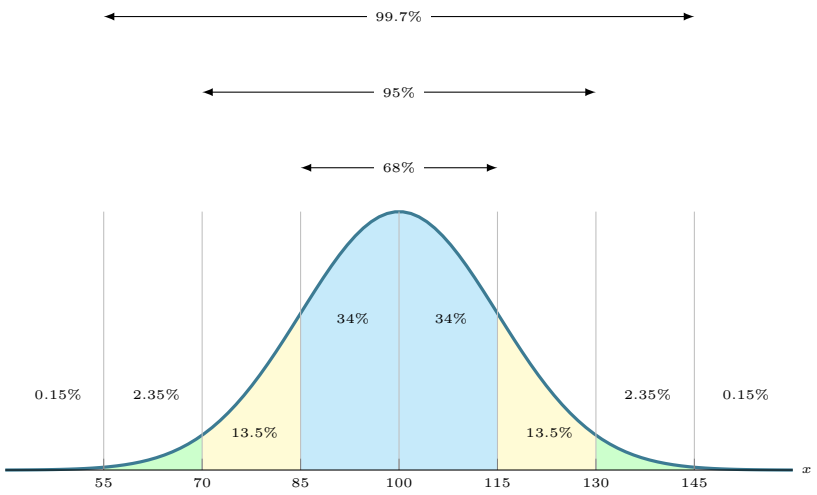


What if we want to find what percentage of the data falls in some other range? Like what about the percentage of IQs that fall between 100 and 115? Or above 85? Between 70 and 115?

All of this can be done with a little clever analysis of the figure above. We just need to divide it up into segments that are each one standard deviation (15 IQ points) wide and figure out what percentage of the data is in each slice.

First of all, notice that the center region (between 85 and 115) contains 68% of the data. Because the graph is symmetric, we can conclude that each half of that contains 34%.

Next, the two yellow regions together contain $95\% - 68\% = 27\%$, so each region contains half of that, or 13.5%. Similarly, the two green regions account for $99.7\% - 95\% = 4.7\%$, so each of them contains 2.35% of the data. Finally, the tails outside the green account for the remaining 0.3% of the data, so each side contains 0.15%.



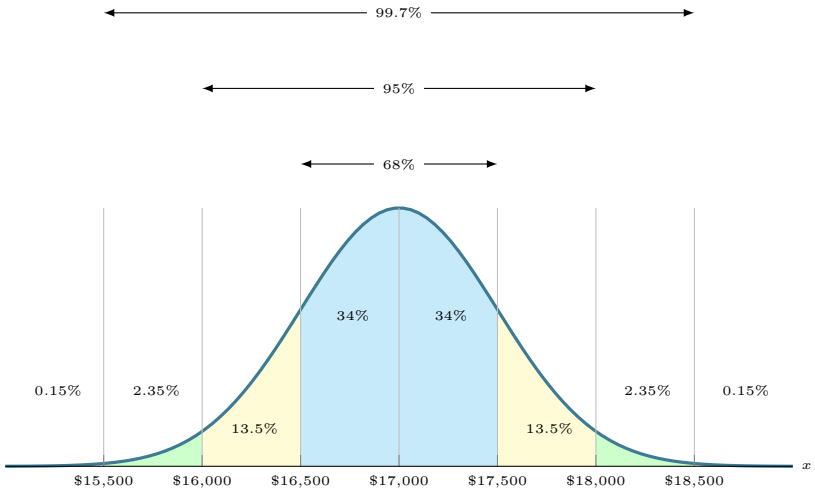
The important point is not to memorize these percentages, but rather to understand how we figured them out. If you can follow and recreate that process, all you'll have to memorize is the 68–95–99.7 part, and you can reproduce a picture like that one in a minute or two of quick thought. Once you can do that, you can answer questions like the following one.

EXAMPLE 5 CAR SALES

Suppose you know that the prices paid for cars are normally distributed with a mean of $17,000 and a standard deviation of $500. Use the 68–95–99.7 Rule to find the percentage of buyers who paid

(a) between $16,500 and $17,500
(b) between $17,500 and $18,000
(c) between $16,000 and $17,000
(d) between $16,500 and $18,000
(e) below $16,000
(f) above $18,500

Solution We can use the same process that was just described to build the following diagram, using the given mean and standard deviation.



You should be able to use the figure above to reason out that

(a) the percentage of buyers who spent between $16,500 and $17,500 was 68%.

(b) the percentage of buyers who spent between $17,500 and $18,000 was 13.5%.

(c) the percentage of buyers who spent between $16,000 and $17,000 was 47.5%.

(d) the percentage of buyers who spent between $16,500 and $18,000 was 81.5%.

(e) the percentage of buyers who spent below $16,000 was 2.5%.

(f) the percentage of buyers who spent above $18,500 was 0.15%.

TRY IT

The mean height of boys 15 to 18-years old from Chile is 170 cm with a standard deviation of 6 cm. Male heights are known to be normally distributed. Using the Empirical Rule, find

(a) the percentage of boys with heights between 158 and 176.

(b) the percentage of boys with heights above 188.

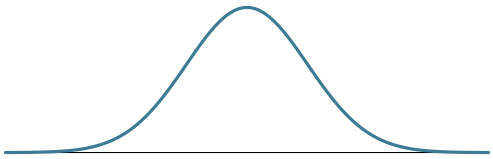
(c) the percentage of boys with heights below 164.

The Normal Distribution and Polls

Suppose you're tasked with conducting a straw poll to predict the victor in a close Senate race between Jonas Hawkins and Violeta Gass. You poll (randomly, because you're a good statistics student) 500 people and ask them who they plan to vote for, and 52% of them respond Hawkins and 48% Gass. Good so far, but you begin to wonder: is this really an accurate representation of the population? You picked a good sample, but is there any way to put a number on how certain you are that your results are a valid predictor of what the population will do?

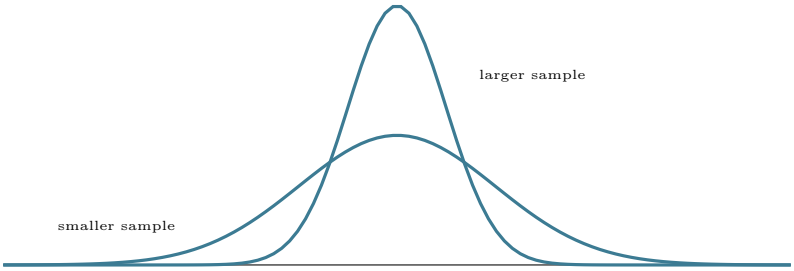
The answer is based on the Normal Distribution. The idea is this: if we took another sample and polled them, and then another sample, and another and another, and repeated this process

many times over, the results of our poll would begin to look like a normal distribution.



In other words, most of those polls we conducted would look similar to each other, and they would be grouped together. There would be a few polls that would have drastically lower percentages for Hawkins, and a few would have drastically higher percentages–simply due to the inherent variability of a sample that we can never fully eliminate–but most of them would be clustered around the true percentage of the population that plan to vote for Hawkins. In other words, the wrong polls would be rare, and the polls that are more right would be more common.

In fact, it turns out that we can specifically define this distribution as a normal distribution, which tells us that we're–for instance–95% confident that the results we got when we took the first poll were within two standard deviations of the mean. Now then, if we make our sample larger, it turns out that the standard deviation on this normal distribution gets smaller, so the results are more precise.



Margin of Error This allows us to define something called the margin of error. You may have seen or heard this term in the context of polls, especially political polls. The margin of error is an inevitable part of using a sample to predict what a larger population will do, and it only depends on n, the size of the sample (strangely enough, it doesn't depend on the size of the population). The larger the sample size, the smaller the margin of error will be, and thus the more precise the results of the poll will be.

Margin of Error

If the sample size of a poll is n, there is at least a 95% chance that the sample percentage lies within

$$\frac{1}{\sqrt{n}} \times 100\%$$

of the population percent. The margin of error with a 95% *confidence level* is $\pm\frac{1}{\sqrt{n}} \times 100\%$.

Beware, though, that you don't take for granted that the margin of error is the only thing to worry about; we've already seen that there are other sources of error, like poor sampling. Also, we didn't even talk about other sources of bias, like self-interest or word choice.

MARGIN OF ERROR EXAMPLE 6

What is the margin of error on a poll with a sample size of 1000 people?

Solution

The margin of error is

$$\pm\frac{1}{\sqrt{1000}} \times 100\% = \pm 3.16\%$$

A margin of error of about 3% (which is common for many political polls) corresponds to a sample size of 1000.

Exercises 3.4

1. The widths of platinum samples manufactured at a factory are normally distributed, with a mean of 1.1 cm and a standard deviation of 0.2 cm. Find the z-scores that correspond to each of the following widths.

(a) 1.5 cm

(b) 0.94 cm

2. The average resting heart rate of a population is 88 beats per minute, with a standard deviation of 12 bpm. Find the z-scores that correspond to each of the following heart rates.

(a) 120 bpm

(b) 71 bpm

3. The average height of American adult males is 177 cm, with a standard deviation of 7.4 cm. Meanwhile, the average height of Indian males is 165 cm, with a standard deviation of 6.7 cm. Which is taller relative to his nationality, a 175-cm American man or a 162-cm Indian man?

4. Kyle and Ryan take entrance exams at two different universities. Kyle scores a 430 on an exam with a mean of 385 and a standard deviation of 70, while Ryan scores a 31 on an exam with a mean of 28 and a standard deviation of 4.5. Which do you think is more likely to be accepted at their university of choice?

5. A doctor measured serum HDL levels in her patients, and found that they were normally distributed with a mean of 63.4 and a standard deviation of 3.8. Find the serum HDL levels that correspond to the following z-scores.

(a) $z = -0.85$

(b) $z = 1.33$

6. If the distribution of weight of newborn babies in Maryland is approximately normal, with a mean of 3.23 kilograms and a standard deviation of 0.87 kilograms, find the weights that correspond to the following z-scores.

(a) $z = 2.20$

(b) $z = -1.73$

7. Once again, the heights of American adult males are normally distributed with a mean of 177 cm and a standard deviation of 7.4 cm. Find the range of heights that contain approximately

(a) 68% of the data

(b) 95% of the data

(c) 99.7% of the data

8. Suppose again that babies' weights are normally distributed with a mean of 3.23 kg and a standard deviation of 0.87 kg. Find the range of weights that contain approximately

(a) 68% of the data

(b) 95% of the data

(c) 99.7% of the data

9. Suppose that the scores on a statewide standardized test are normally distributed with a mean of 72 and a standard deviation of 4. Estimate the percentage of scores that were

(a) between 68 and 76.

(b) above 76.

(c) below 64.

(d) between 68 and 84.

10. Water usages in American showers are normally distributed, with the average shower using 17.2 gallons, and a standard deviation of 2.5 gallons. Estimate the percentage of showers that used

(a) more than 22.2 gallons.

(b) less than 14.7 gallons.

(c) between 12.2 and 22.2 gallons.

(d) between 9.7 and 19.7 gallons.

11. GMAT scores are approximately normally distributed with a mean of 547 and a standard deviation of 95. Estimate the percentage of scores that were

(a) between 262 and 832.

(b) above 642.

(c) below 262.

(d) between 262 and 452.

12. Suppose that wedding costs in the Caribbean are normally distributed with a mean of \$7,500 and a standard deviation of \$975. Estimate the percentage of Caribbean weddings that cost

(a) between \$6525 and \$9450.

(b) above \$9450.

(c) below \$6525.

(d) between \$4575 and \$10,425.

13. What is the margin of error for a poll with a sample size of 2000 people?

14. What is the margin of error for a poll with a sample size of 150 people?

15. If you want a poll to have a margin of error of 2.5%, how large will your sample have to be?

16. If you want a poll to have a margin of error of 1%, how large will your sample have to be?