CHAPTER 2 How to Look at Data

Chapter Outline

2.1 Frequency Tables

Learning Objectives

- Read and make frequency tables for a data set.
- Construct relative frequencies for a data set.

Introduction

Charts and graphs, when created carefully, can provide instantaneous information about a data set without having to calculate or even have knowledge of statistical measures. This chapter will concentrate on frequency tables, a precursor to graphing. When we look at all the values in a sample or a dataset, we are looking at a **distribution**. Often it is the distribution of data – and not the individual data points themselves - that are of interest to us. You may have earned a B- on your most recent philosophy exam, for example, but you want to know how you performed relative to your classmates. To answer that question, your first need to know what the distribution of exam scores looks like.

Frequency Tables

As an example of how organizing data can help us better understand the world around us, let's take a look at the impact of limited resources and recycling challenges on our planet. The earth has seemed so large in scope for thousands of years that it is only recently that many people have begun to take seriously the idea that we live on a planet of limited and dwindling resources. This is something that residents of the Galapagos Islands are also beginning to understand. Because of its isolation and lack of resources to support large, modernized populations of humans, the problems that we face on a global level are magnified in the Galapagos. Basic human resources such as water, food, fuel, and building materials must all be brought in to the islands. As the human population grows exponentially, the Islands are confronted with the problem of what to do with all the waste.

Let's look specifically at one resource: bottled water. Bottled water consumption worldwide has grown and continues to grow at a phenomenal rate. According to the Earth Policy Institute, 154 billion gallons were produced in 2004. While there are places in the world where safe water supplies are unavailable, most of the growth in consumption has been due to other reasons. The largest consumer of bottled water is the United States, which arguably could be the country with the best access to safe, convenient, and reliable sources of tap water. Only a small fraction of the plastic that is recycled. In addition, huge volumes of carbon emissions are created when these bottles are manufactured using oil and transported great distances by oil-burning vehicles.

Our Bottled Water Usage

Imagine you were to take an informal poll of your class. You ask each member of the class, on average, how many bottled waters they use in a week. Once you collect this data, you want to organize it so it is easier to understand. A frequency table is a common starting point. **Frequency tables** simply display each value of the variable, and the number of occurrences (the frequency) of each of those values. In this example, the variable is the number of plastic bottles of water consumed each week.

Consider the following raw data:

$6, 4, 7, 7, 8, 5, 3, 6, 8, 6, 5, 7, 7, 5, 2, 6, 1, 3, 5, 4, 7, 4, 6, 7, 6, 6, 7, 5, 4, 6, 5, 3$

Here is a frequency table to display the information collected above:

TABLE 2.1: Completed Frequency Table for Water Bottle Data

Number of Plastic Beverage Bottles per Week	Frequency
1	1
2	1
3	3
4	4
5	6
6	8
7	7
8	2

When creating a frequency table, it is often helpful to use tally marks as a running total to avoid missing a value or over-representing another. But those tally marks typically don't appear in the final table.

TABLE 2.2: Frequency Table using Tally Marks

Number of Plastic Beverage Bottles per Week	Tally	Frequency
1	|	1
2	|	1
3	|||	3
4	||||	4
5	$\not{\text{IIII}}$ |	6
6	$\not{\text{IIII}}$ |||	8
7	$\not{\text{IIII}}$ ||	7
8	||	2

Grouped Frequency Distributions

In the table below, we can see the raw data for how much bottled water is consumed by different countries in the world. Suppose we want to create a frequency distribution for this data. It would make more sense to report a frequency for a range of values (e.g., 80-90, 90-100, etc.). We will use a process called **binning** to separate the raw data into categories of a given, consistent size.

TABLE 2.3:

Country	Liters of Bottled Water Consumed per Person per Year
Italy	183.6
Mexico	168.5
United Arab Emirates	163.5
Belgium and Luxembourg	148.0
France	141.6

TABLE 2.3: (continued)

Country	Liters of Bottled Water Consumed per Person per Year
Spain	136.7
Germany	124.9
Lebanon	101.4
Switzerland	99.6
Cyprus	92.0
United States	90.5
Saudi Arabia	87.8
Czech Republic	87.1
Austria	82.1
Portugal	80.3

Below is a grouped frequency distribution for the water-consumption data. A bracket, '[' or ']', indicates that the endpoint of the interval is included in the class. A parenthesis, '(' or ')', indicates that the endpoint is not included. It is common practice in statistics to include a number that borders two classes as the larger of the two numbers in an interval. For example, $[80-90)$ means this classification includes everything from 80 and gets infinitely close to, but not equal to, 90. 90 is included in the next class, $[90-100)$.

TABLE 2.4:

Liters per Person	Frequency
$[80-90)$	4
$[90-100)$	3
$[100-110)$	1
$[110-120)$	0
$[120-130)$	1
$[130-140)$	1
$[140-150)$	2
$[150-160)$	0
$[160-170)$	2
$[170-180)$	0
$[180-190)$	1

Relative Frequencies

If you were evaluating a set of data describing the numbers of "A's", "B's", "C's", and "D's" that students earned on a particular test, and needed to display the data on a relative frequency table, how would you go about it?

A **relative frequency table** is specifically designed to display the ratio of each individual frequency to the total frequency of the data. Sometimes these may be represented as percentages.

Example A

The students in a class were asked what kind of music they liked. 18 liked rock, 11 liked pop, 5 liked hip hop, and 8 liked country. Create a frequency and relative frequency table using this information.

To create the frequency table, we just need one column for each category:

TABLE 2.5:

Rock	Pop	Hip Hop	Country
18	11	5	8

To convert to a relative frequency table, just divide each frequency by the total:

TABLE 2.6:

Rock	Pop	Hip Hop	Country
$\frac{18}{42} = .43$	$\frac{11}{42} = .26$	$\frac{5}{42} = .12$	$\frac{8}{42} = .19$

To build a relative frequency table, start by grouping the values into categories or bins, depending on the type of data you have. You should try to limit the number of total groups to less than a dozen in most cases. Once you have all of your data separated into groups, tally the number of values in each group.

To calculate the **relative frequency** of each category, divide the number of values in a group by the overall frequency. The decimal you get will represent the part of the entire sample that is represented by that category. Once you have calculated all of the relative frequencies for every category, add them up to make sure they total 1.0.

NOTE: If you are graphing the relative frequencies of a **continuous variable**, you will need to specify how to handle any values that fall right on the edge of a bin. Here are a couple of ways to do this:

- You can specify on your table that values equal to lower class limits are included in the bin, but values equal to upper class limits are not (this is the conventional method). This means that a value of 5 would be considered part of a 5-10 class, but not part of a 1-5 class.
- You can also define each category so that there are no overlapping values:

$$1 - 4.995 - 9.9910 - 14.9915 - 20$$

Example B

You are given a bag of marbles in multiple colors, if there are 25 red, 22 yellow, 17 green, and 28 blue marbles, what are the relative frequencies of each color?

Solution

Start by totaling the number of marbles: $25 + 22 + 17 + 28 = 92$ total marbles

Divide the number of each color by the total number of marbles:

$$\frac{25\ red\ marbles}{92\ total\ marbles} = .272$$

$$\frac{22\ yellow\ marbles}{92\ totalmarbles} = .239$$

$$\frac{17\ green\ marbles}{92\ total\ marbles} = .185$$

$$\frac{28\ blue\ marbles}{92\ total\ marbles} = .304$$

Add your totals together to verify that they equal 1:

$$.272 + .239 + .185 + .304 = 1.0$$

Note that each of the relative frequencies can also be understood as percentages:

.272 = 27.2% red marbles

.239 = 23.9% yellow marbles

.185 = 18.5% green marbles

.304 = 30.4% blue marbles

$$27.2\% + 23.9\% + 18.5\% + 30.4\% = 100\%$$

Example C

A police officer is reviewing accident statistics for her city. She notes that there were a total of 23 incidents involving teen drivers between ages sixteen and twenty-one, 19 incidents involving drivers aged twenty-two through twenty-six, 19 involving twenty-seven to forty-year-olds, and 18 for ages above forty-one.

What are the relative frequencies for each age range?

Solution

The total number of accidents is:

$$23 + 19 + 19 + 18 = 79 \text{ total accidents}$$

The relative frequencies are:

$$\frac{23 \textit{ in age range } 16-21}{79 \textit{ total}} = .291$$
$$\frac{19 \textit{ in age range } 22-26}{79 \textit{ total}} = .241$$
$$\frac{19 \textit{ in age range } 27-40}{79 \textit{ total}} = .241$$
$$\frac{18 \textit{ in age range } 41+}{79 \textit{ total}} = .228$$

Verify that the relative frequencies total 1.0:

$$.291 + .241 + .241 + .228 = 1.001 (\textit{due to rounding})$$

Lesson Summary

A frequency table is useful to organize data into classes according to the number of occurrences, or frequency, of each class. Relative frequency shows the percentage of data in each class.

Vocabulary

A **relative frequency table** compares the number of entries in each of several categories to the number of entries in the total sample size.

Binning is the common term for the process of dividing data up into multiple categories, classes, or intervals in preparation for graphing.

A **continuous variable** is a variable that can represent *any* value between a given minimum and maximum. Age is a common continuous variable, since age can be measured in infinitely small increments. By contrast,

a **discrete variable** can only represent *specific* values in a given range. The number rolled on a standard die is a discrete variable since it can only be one of the numbers 1 –6, no partials or fractions.

Review Questions

1. Lois was gathering data on the plastic beverage bottle consumption habits of her classmates, but she ran out of time as class was ending. When she arrived home, something had spilled in her backpack and smudged the data for the 2's. Fortunately, none of the other values was affected, and she knew there were 30 total students in the class. Complete her frequency table.

TABLE 2.7:

Number of Plastic Beverage Bottles per Week	Tally	Frequency
1				
2		
3					
4				
5					
6								
7							
8			

2. The following frequency table contains exactly one data value that is a positive multiple of ten. What must that value be?

TABLE 2.8:

Class	Frequency
$[0-5)$	4
$[5-10)$	0
$[10-15)$	2
$[15-20)$	1
$[20-25)$	0
$[25-30)$	3
$[30-35)$	0
$[35-40)$	1

(a) 10

(b) 20

(c) 30

(d) 40

(e) There is not enough information to determine the answer.

3. The following table includes the data from the same group of countries from the earlier bottled water consumption example, but is for the year 1999, instead. Create a frequency table for this data set.

TABLE 2.9:

Country	Liters of Bottled Water Consumed per Person per Year
Italy	154.8
Mexico	117.0
United Arab Emirates	109.8
Belgium and Luxembourg	121.9
France	117.3
Spain	101.8
Germany	100.7
Lebanon	67.8
Switzerland	90.1
Cyprus	67.4
United States	63.6
Saudi Arabia	75.3
Czech Republic	62.1
Austria	74.6
Portugal	70.4

4. The following table shows the potential energy that could be saved by manufacturing each type of material using the maximum percentage of recycled materials, as opposed to using all new materials. Construct a frequency table, including the actual frequency, the relative frequency (round to the nearest tenth of a percent), and the relative cumulative frequency. Assume a bin width of 25 million BTUs

TABLE 2.10:

Manufactured Material	Energy Saved (millions of BTU's per ton)
Aluminum Cans	206
Copper Wire	83
Steel Cans	20
LDPE Plastics (e.g., trash bags)	56
PET Plastics (e.g., beverage bottles)	53
HDPE Plastics (e.g., household cleaner bottles)	51
Personal Computers	43
Carpet	106
Glass	2
Corrugated Cardboard	15
Newspaper	16
Phone Books	11
Magazines	11
Office Paper	10

Guided Practice

1. The Sackmore and Headbut village football teams have played each other 50 times. Sackmore has won 10 times, Headbut has won 35 times, and the teams have drawn 5 times. Based on past performance, what is the probability that Sackmore will win the next match?
2. Tony estimates that the probability that there will be an empty space in the car park when he arrives at work is $\frac{4}{5}$. His estimate is based on 50 observations. On how many of these 50 days was he *unable* to find an empty space in the car park?
3. A pair of dice (one red, one green) is cast 30 times, and on 4 of these occasions, the sum of the numbers facing up is 7. What is the relative frequency that the sum is 7?
4. In 1990, there were approximately 10,000 fast food outlets in the US that specialized in Mexican food. Of these, the largest were Taco Bell with 4809 outlets, Taco John's with 430 outlets and Del Taco with 275 outlets. The relative frequency that a fast food outlet that specializes in Mexican food is none of the above is:

Solutions

1. So far, Sackmore hase won 10 out of the 50 matches. We can write this as a fraction, which (reduced) is: $\frac{1}{5}$. This fraction isn't really the probability of Sackmore winning, but it is an *estimate* of that probability. We say that the *relative frequency* of Sackmore winning is $\frac{1}{5}$.
2. If Tony has figured that he *is* able to find a space 4 of every 5 times he arrives, then he *is not* able to find a space 1 in every 5 times. If we set the ratio: $\frac{1}{5} = \frac{x}{50}$, we can solve for x to find that he did not have a space 10 times.
3. Out of thirty throws, four of them were 7's. The relative frequency is $\frac{4}{30}$ *or* $\frac{2}{15}$.
4. The likelihood that a restaurant is *not* one of the top three would equal the number of Mexican fast food restaurants that are not one of the three: $10,000 - 4809 - 430 - 275 = 4486$, divided by the total number of Mexican fast food restaurants, 10,000:

$$\frac{4,486}{10,000} = .4486 \text{ or } 44.86\%$$

More Practice

30 Students in a class surveyed each other to find out their favorite movie series, and recorded the results in a table like the one shown below.

TABLE 2.11:

Movie Series	Number of Likes
Twilight	7
Lord of the Rings	5
Pirates of the Caribbean	9
Harry Potter	6
Narnia	2
High School Musical	1

1. What was the relative frequency for Narnia?
2. What was the relative frequency for Pirates of the Caribbean?

3. 100 people were asked whether they were left-handed. 8 people answered yes. What is the relative frequency of left-handed people in the survey?
4. The relative frequency of getting a white candy from a particular bag is 0.3. If the bag contains 100 candies, estimate the number of whites.
5. Kyle observed 80 cars as they drove by his bedroom window. 24 of them were red. What is the relative frequency of red cars?
6. The relative frequency of rain in April is .6. There are 30 days in April. Estimate the number of days of rain expected in April.

Use the table below listing the heights of 100 male semiprofessional soccer players.

TABLE 2.12:

Heights (Inches)	Frequency of Students	Relative Frequency
59.95-61.95	5	
61.95-63.95	3	3/100 = 0.03
63.95-65.95		15/100 = 0.15
65.95-67.95	40	40/100 = 0.40
67.95-69.95	17	
69.95-71.95	12	12/100 = 0.12
71.95-73.95		7/100 = 0.07
73.95-75.95	1	1/100 = 0.01
	Total = 100	Total = 1.00

7. Fill in the blanks and check your answers.
8. The percentage of heights that are from 67.95 to 71.95 inches is:
9. The percentage of heights that are from 67.95 to 73.95 inches is:
10. The percentage of heights that are more than 65.95 inches is:
11. The number of players in the sample who are between 61.95 and 71.95 inches tall is:
12. What kind of data does this chart highlight, qualitative or quantitative?

2.2 Graphs for Categorical Data

Learning Objective

- Identify and translate data sets to and from a bar graph and a pie graph.

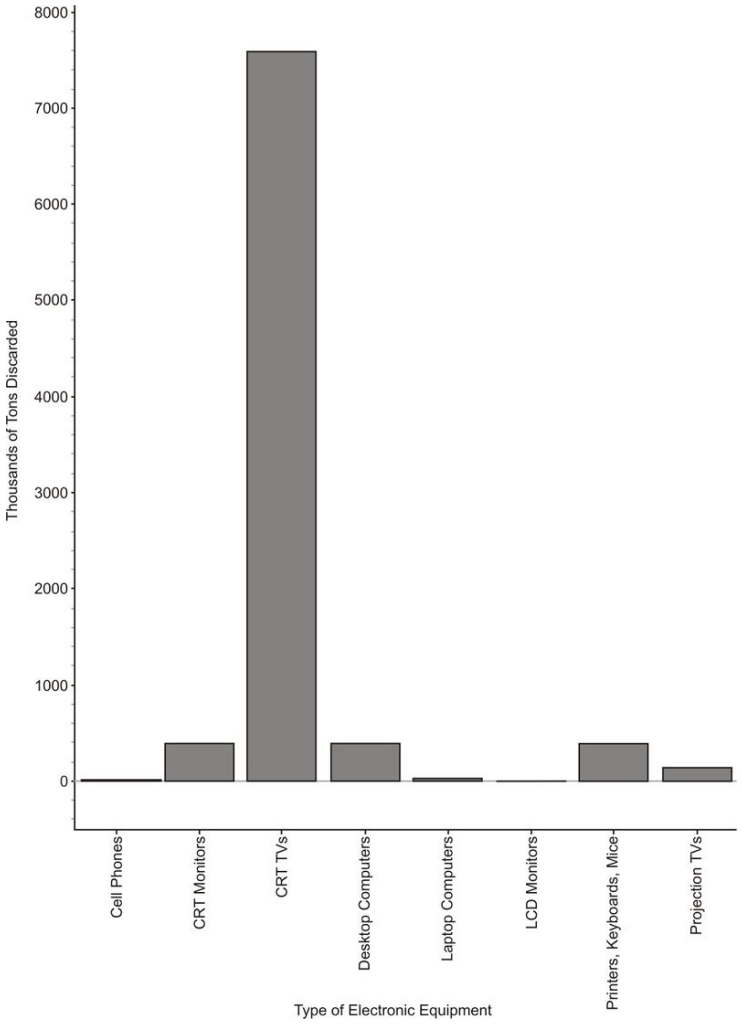
Categorical Variables: Bar Graphs and Pie Graphs

We live in an age of unprecedented access to increasingly sophisticated and affordable personal technology. Cell phones, computers, and televisions now improve so rapidly that, while they may still be in working condition, the drive to make use of the latest technological breakthroughs leads many to discard usable electronic equipment. Much of that ends up in a landfill, where the chemicals from batteries and other electronics add toxins to the environment. Approximately 80% of the electronics discarded in the United States is also exported to third world countries, where it is disposed of under generally hazardous conditions by unprotected workers[1]. The following table shows the amount of tonnage of the most common types of electronic equipment discarded in the United States in 2005.

TABLE 2.13:

Electronic Equipment	Thousands of Tons Discarded
Cathode Ray Tube (CRT) TV's	7591.1
CRT Monitors	389.8
Printers, Keyboards, Mice	324.9
Desktop Computers	259.5
Laptop Computers	30.8
Projection TV's	132.8
Cell Phones	11.7
LCD Monitors	4.9

The type of electronic equipment is a categorical variable, and therefore, this data can easily be represented using the **bar graph** below:



The bars in a bar graph usually are separated slightly. The graph is just a series of disjoint categories, all represented along the same axis. The height of each bar tells you the frequency of that particular value in the data set. It doesn't make sense to talk about the shape of this distribution of values. If we rearranged the categories in a different order, the same data set could be made to look different. Do not try to infer shape from a bar graph!

Pie Graphs

Usually, data that can be represented in a bar graph can also be shown using a **pie graph** (also commonly called a circle graph or pie chart). In this representation, we convert the count into a percentage so we can show each category relative to the total. Each percentage is then converted into a proportionate sector of the circle. To make this conversion, simply multiply the percentage by 3.6, which represents 360 (the total number of degrees in a circle) divided by 100% (the total percentage available).

Here is a table with the percentages and the approximate angle measure of each sector:

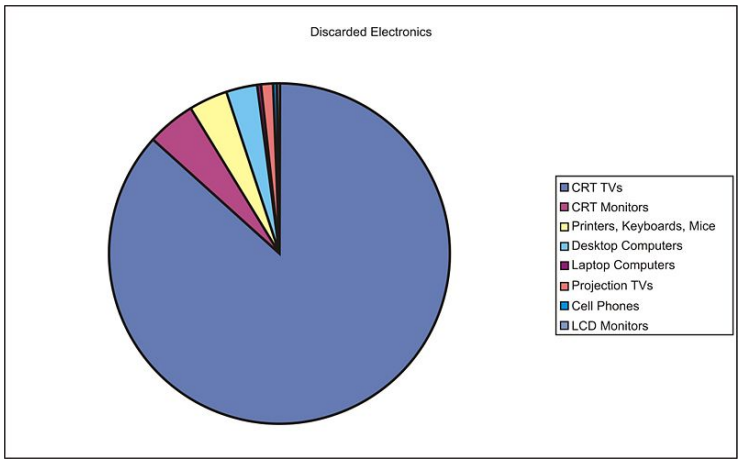
TABLE 2.14:

Electronic Equipment	Thousands of Tons Discarded	Percentage of Total Discarded	Angle Measure of Circle Sector
Cathode Ray Tube (CRT) TV's	7591.1	86.8	312.5
CRT Monitors	389.8	4.5	16.2

TABLE 2.14: (continued)

Electronic Equipment	Thousands of Tons Discarded	Percentage of Total Discarded	Angle Measure of Circle Sector
Printers, Keyboards, Mice	324.9	3.7	13.4
Desktop Computers	259.5	3.0	10.7
Laptop Computers	30.8	0.4	1.3
Projection TV's	132.8	1.5	5.5
Cell Phones	11.7	0.1	0.5
LCD Monitors	4.9	~0	0.2

And here is the completed pie graph:



Lesson Summary

Bar graphs are used to represent categorical data. Pie (or circle) graphs are also useful ways to display categorical variables, especially when it is important to show how percentages of an entire data set fit into individual categories.

Points to Consider

- What characteristics of quantitative data make it easier or harder to graph than categorical data?

Review Questions

1. Computer equipment contains many elements and chemicals that are either hazardous, or potentially valuable when recycled. The following data set shows the contents of a typical desktop computer weighing approximately 27kg. Some of the more hazardous substances, like Mercury, have been included in the 'other' category, because they occur in relatively small amounts that are still dangerous and toxic.

TABLE 2.15:

Material	Kilograms
Plastics	6.21
Lead	1.71
Aluminum	3.83

TABLE 2.15: (continued)

Material	Kilograms
Iron	5.54
Copper	2.12
Tin	0.27
Zinc	0.60
Nickel	0.23
Barium	0.05
Other elements and chemicals	6.44

a. Create a bar graph for this data.
b. Complete the chart below to show the approximate percentage of the total weight for each material.

TABLE 2.16:

Material	Kilograms	Approximate Percentage of Total Weight
Plastics	6.21	
Lead	1.71	
Aluminum	3.83	
Iron	5.54	
Copper	2.12	
Tin	0.27	
Zinc	0.60	
Nickel	0.23	
Barium	0.05	
Other elements and chemicals	6.44	

c. Create a pie graph for this data.

2.3 Graphs for Quantitative Data

Learning Objectives

- Identify and translate data sets to and from a histogram, a relative frequency histogram, and a frequency polygon.
- Identify histogram distribution shapes as skewed or symmetric and understand the basic implications of these shapes.

Displaying Univariate Data

Dot Plots

A **dot plot** is one of the simplest ways to represent numerical data. After choosing an appropriate scale on the axes, each data point is plotted as a single dot. Multiple points at the same value are stacked on top of each other using equal spacing to help convey the shape and center.

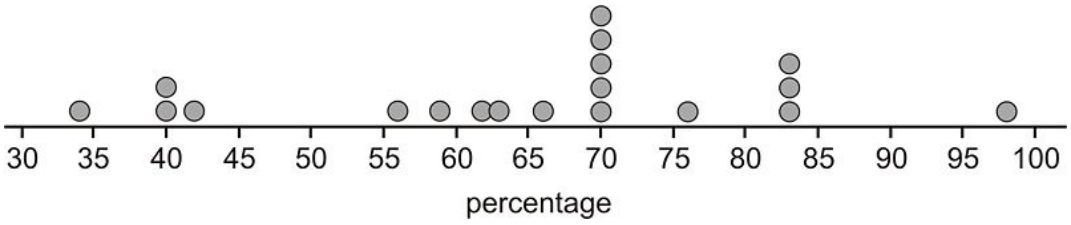
Example

The following is a data set representing the percentage of paper packaging manufactured from recycled materials for a select group of countries.

TABLE 2.17:

Country	% of Paper Packaging Recycled
Estonia	34
New Zealand	40
Poland	40
Cyprus	42
Portugal	56
United States	59
Italy	62
Spain	63
Australia	66
Greece	70
Finland	70
Ireland	70
Netherlands	70
Sweden	70
France	76
Germany	83
Austria	83
Belgium	83
Japan	98

The dot plot for this data would look like this:



Notice that this data set is centered at a manufacturing rate for using recycled materials of between 65% and 70%. It is spread from 34% to 98%, and appears very roughly symmetric, perhaps even slightly skewed left. Dot plots have the advantage of showing all the data points and giving a quick and easy snapshot of the shape, center, and spread. Dot plots are not much help when there is little repetition in the data. They can also be very tedious if you are creating them by hand with large data sets, though computer software can make quick and easy work of creating dot plots from such data sets.

Stem-and-Leaf Plots

One of the shortcomings of dot plots is that they do not show the actual values of the data. You have to read or infer them from the graph. From the previous example, you might have been able to guess that the lowest value is 34%, but you would have to look in the data table itself to know for sure. A **stem-and-leaf plot** is a similar plot in which it is much easier to read the actual data values. In a stem-and-leaf plot, each data value is represented by two digits: the stem and the leaf. In this example, it makes sense to use the ten's digits for the stems and the one's digits for the leaves. The stems are on the left of a dividing line as follows:

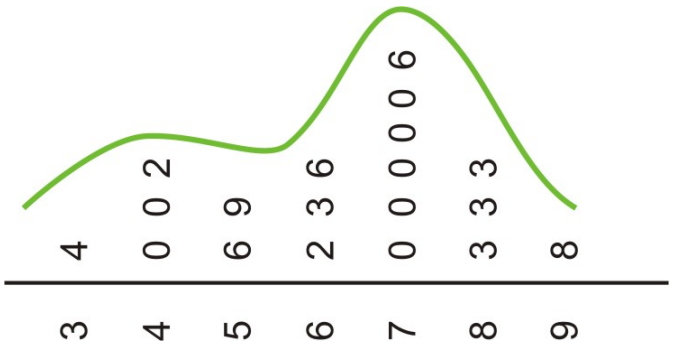
Once the stems are decided, the leaves representing the one's digits are listed in numerical order from left to right:

3	4
4	0 0 2
5	6 9
6	2 3 6
7	0 0 0 0 0 6
8	3 3 3
9	8

It is important to explain the meaning of the data in the plot for someone who is viewing it without seeing the original data. For example, you could place the following sentence at the bottom of the chart:

NOTE: 5|69 means 56% and 59% are the two values in the 50's.

If you could rotate this plot on its side, you would see the similarities with the dot plot. The general shape and center of the plot is easily found, and we know exactly what each point represents. This plot also shows the slight skewing to the left that we suspected from the dot plot. Stem plots can be difficult to create, depending on the numerical qualities and the spread of the data. If the data values contain more than two digits, you will need to remove some of the information by rounding. A data set that has large gaps between values can also make the stem plot hard to create and less useful when interpreting the data.



Back-to-Back Stem Plots

Stem plots can also be a useful tool for comparing two distributions when placed next to each other. These are commonly called **back-to-back stem plots**.

In a previous example, we looked at recycling in paper packaging. Here are the same countries and their percentages of recycled material used to manufacture glass packaging:

TABLE 2.18:

Country	% of Glass Packaging Recycled
Cyprus	4

TABLE 2.18: (continued)

Country	% of Glass Packaging Recycled
United States	21
Poland	27
Greece	34
Portugal	39
Spain	41
Australia	44
Ireland	56
Italy	56
Finland	56
France	59
Estonia	64
New Zealand	72
Netherlands	76
Germany	81
Austria	86
Japan	96
Belgium	98
Sweden	100

In a back-to-back stem plot, one of the distributions simply works off the left side of the stems. In this case, the spread of the glass distribution is wider, so we will have to add a few extra stems. Even if there are no data values in a stem, you must include it to preserve the spacing, or you will not get an accurate picture of the shape and spread.

Glass		Paper
4	0	
	1	
71	2	
94	3	4
41	4	002
9666	5	69
4	6	236
62	7	000006
61	8	333
86	9	8
0	10	

We have already mentioned that the spread was larger in the glass distribution, and it is easy to see this in the comparison plot. You can also see that the glass distribution is more symmetric and is centered lower (around the

mid-50's), which seems to indicate that overall, these countries manufacture a smaller percentage of glass from recycled material than they do paper. It is interesting to note in this data set that Sweden actually imports glass from other countries for recycling, so its effective percentage is actually more than 100.

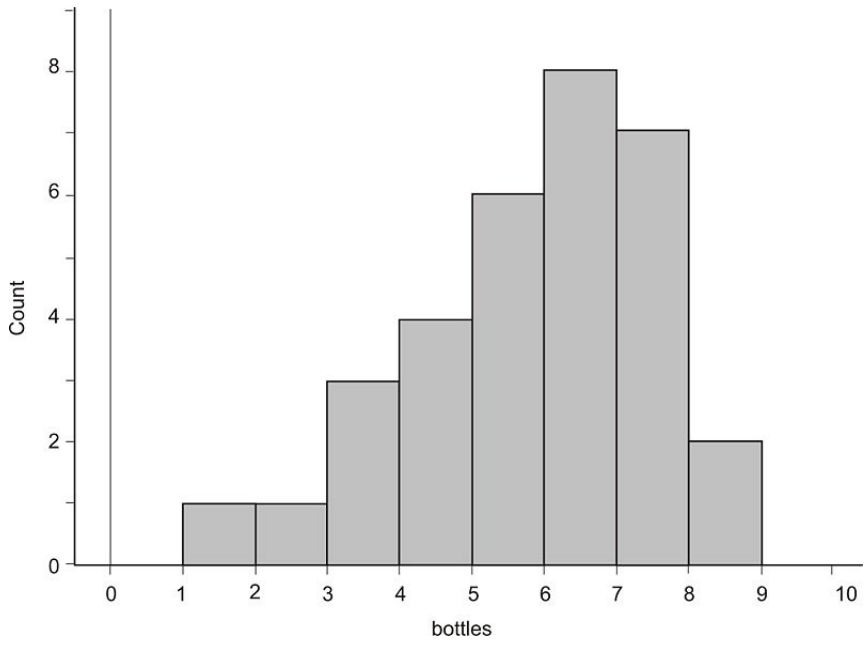
Histograms

Once you create a frequency table, you are ready to create a graphical representation called a **histogram**. Let's revisit our data about student bottled beverage habits.

TABLE 2.19: Completed Frequency Table for Water Bottle Data

Number of Plastic Beverage Bottles per Week	Frequency
1	1
2	1
3	3
4	4
5	6
6	8
7	7
8	2

Here is the same data in a histogram:



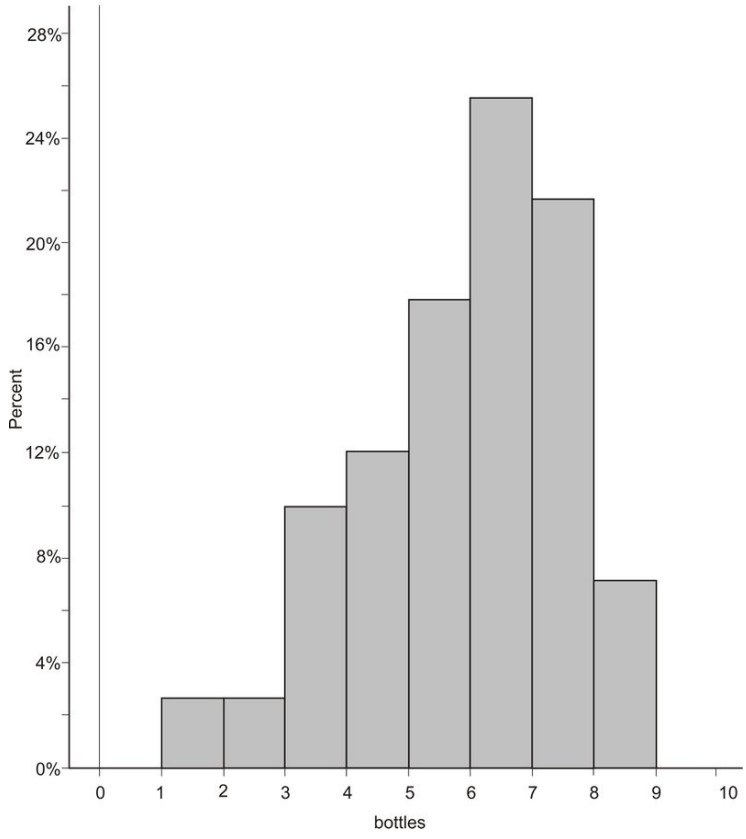
In this case, the horizontal axis represents the variable (number of plastic bottles of water consumed), and the vertical axis is the frequency, or count. Each vertical bar represents the number of people in each class of ranges of bottles. For example, in the range of consuming $[1-2)$ bottles, there is only one person, so the height of the bar is at 1. We can see from the graph that the most common class of bottles used by people each week is the $[6-7)$ range, or six bottles per week.

A histogram is for numerical data. With histograms, the different sections are referred to as **bins**. Think of a column, or bin, as a vertical container that collects all the data for that range of values. If a value occurs on the border between two bins, it is commonly agreed that this value will go in the larger class, or the bin to the right. It is important when

drawing a histogram to be certain that there are enough bins so that the last data value is included. Often this means you have to extend the horizontal axis beyond the value of the last data point. In this example, if we had stopped the graph at 8, we would have missed that data, because the 8's actually appear in the bin between 8 and 9. Very often, when you see histograms in newspapers, magazines, or online, they may instead label the midpoint of each bin. Some graphing software will also label the midpoint of each bin, unless you specify otherwise.

Relative Frequency Histogram

A **relative frequency histogram** is just like a regular histogram, but instead of labeling the frequencies on the vertical axis, we use the percentage of the total data that is present in that bin. For example, there is only one data value in the first bin. This represents $\frac{1}{32}$, or approximately 3%, of the total data. Thus, the vertical bar for the bin extends upward to 3%.

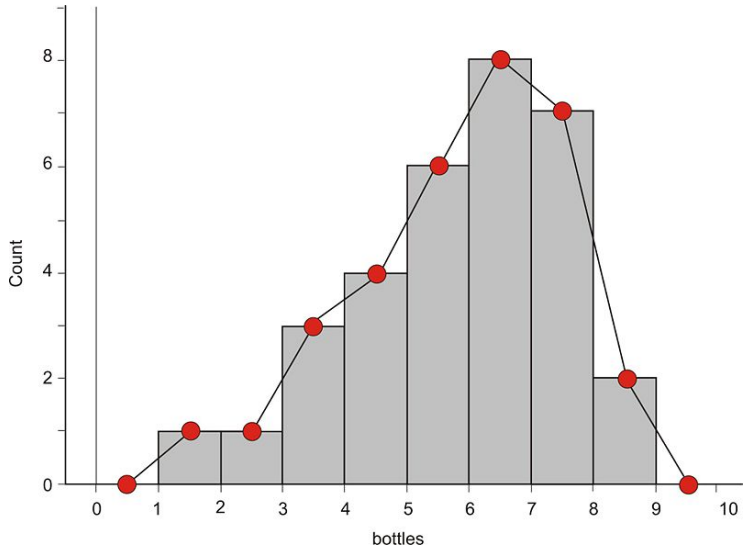


Frequency Polygons

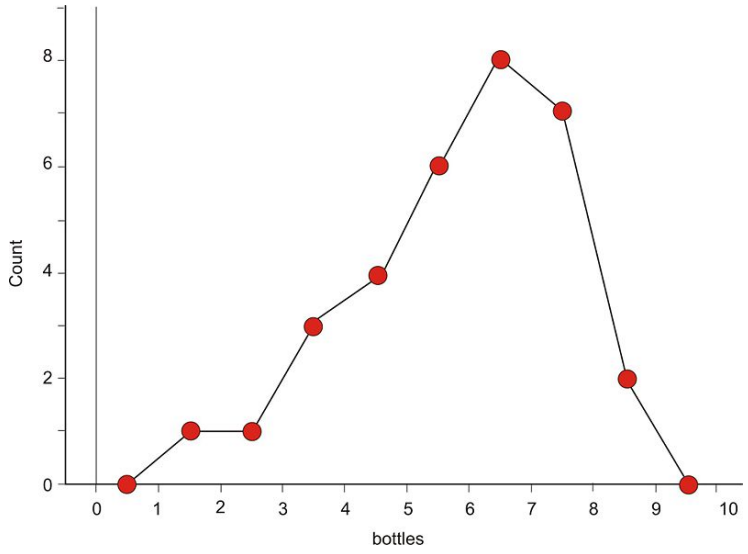
A **frequency polygon** is similar to a histogram, but instead of using bins, a polygon is created by plotting the frequencies and connecting those points with a series of line segments.

To create a frequency polygon for the bottle data, we first find the midpoints of each classification, plot a point at the frequency for each bin at the midpoint, and then connect the points with line segments. To make a polygon with the horizontal axis, plot the midpoint for the class one greater than the maximum for the data, and one less than the minimum.

Here is a frequency polygon constructed directly from the previously-shown histogram:



Here is the frequency polygon in finished form:



Frequency polygons are helpful in showing the general overall shape of a distribution of data. They can also be useful for comparing two sets of data. Imagine how confusing two histograms would look graphed on top of each other!

Shape, Center, Spread

Center and spread are important descriptors of a data set. The shape of a distribution of data is very important as well. Shape, center, and spread should always be your starting point when describing a data set.

Referring to our imaginary student poll on using plastic beverage containers, we notice that the data are spread out from 0 to 9. The graph for the data illustrates this concept, and the range quantifies it. Look back at the graph and notice that there is a large concentration of students in the 5, 6, and 7 region. This would lead us to believe that the center of this data set is somewhere in this area. It is also important that you *see* that the center of the distribution is near the large concentration of data. This is done with shape.

Shape is harder to describe with a single statistical measure, so we will describe it in less quantitative terms. A very important feature of this data set, as well as many that you will encounter, is that it has a single large concentration

of data that appears like a mountain. A data set that is shaped in this way is typically referred to as *mound-shaped* . Mound-shaped data will usually look like one of the following three pictures:

Density Curve 1 Density Curve 2 Density Curve 3

Think of these graphs as frequency polygons that have been smoothed into curves. In statistics, we refer to these graphs as **density curves**. The most important feature of a density curve is symmetry. The first density curve above is *symmetric* and mound-shaped. Notice the second curve is mound-shaped, but the center of the data is concentrated on the left side of the distribution. The right side of the data is spread out across a wider area. This type of distribution is referred to as **skewed right**. It is the direction of the long, spread out section of data, called the *tail*, that determines the direction of the skewing. For example, in the 3^{rd} curve, the left tail of the distribution is stretched out, so this distribution is **skewed left**. Our student bottle data set has this skewed-left shape.

Lesson Summary

A frequency table is useful to organize data into classes according to the number of occurrences, or frequency, of each class. Relative frequency shows the percentage of data in each class. A histogram is a graphical representation of a frequency table (either actual or relative frequency). A frequency polygon is created by plotting the midpoint of each bin at its frequency and connecting the points with line segments. Frequency polygons are useful for viewing the overall shape of a distribution of data, as well as comparing multiple data sets. For any distribution of data, you should always be able to describe the shape, center, and spread. A data set that is mound shaped can be classified as either symmetric or skewed. Distributions that are skewed left have the bulk of the data concentrated on the higher end of the distribution, and the lower end, or tail, of the distribution is spread out to the left. A skewed-right distribution has a large portion of the data concentrated in the lower values of the variable, with the tail spread out to the right.

Points to Consider

- What characteristics of a data set make it easier or harder to represent it using frequency tables, histograms, or frequency polygons?
- What characteristics of a data set make representing it using frequency tables, histograms, or frequency polygons, more or less useful?
- What effects does the shape of a data set have on the statistical measures of center and spread?
- How do you determine the most appropriate classification to use for a frequency table or the bin width to use for a histogram?

Review Questions

1. The following table is from our earlier bottled water example.

TABLE 2.20:

Country	Liters of Bottled Water Consumed per Person per Year
Italy	154.8

TABLE 2.20: (continued)

Country	Liters of Bottled Water Consumed per Person per Year
Mexico	117.0
United Arab Emirates	109.8
Belgium and Luxembourg	121.9
France	117.3
Spain	101.8
Germany	100.7
Lebanon	67.8
Switzerland	90.1
Cyprus	67.4
United States	63.6
Saudi Arabia	75.3
Czech Republic	62.1
Austria	74.6
Portugal	70.4

a. a. Create a frequency table for this data set.
   b. Create the histogram for this data set.
   c. How would you describe the shape of this data set?

2. The following table shows the potential energy that could be saved by manufacturing each type of material using the maximum percentage of recycled materials, as opposed to using all new materials.

TABLE 2.21:

Manufactured Material	Energy Saved (millions of BTU's per ton)
Aluminum Cans	206
Copper Wire	83
Steel Cans	20
LDPE Plastics (e.g., trash bags)	56
PET Plastics (e.g., beverage bottles)	53
HDPE Plastics (e.g., household cleaner bottles)	51
Personal Computers	43
Carpet	106
Glass	2
Corrugated Cardboard	15
Newspaper	16
Phone Books	11
Magazines	11
Office Paper	10

a. a. Construct a frequency table, including the actual frequency, the relative frequency (round to the nearest tenth of a percent), and the relative cumulative frequency. Assume a bin width of 25 million BTUs.

b. Create a relative frequency histogram from your table in part (a).
c. Draw the corresponding frequency polygon.
d. Comment on the shape, center, and spread of this distribution as it relates to the original data. (Do not actually calculate any specific statistics).

2.4 Shapes of Distributions

Learning Objective

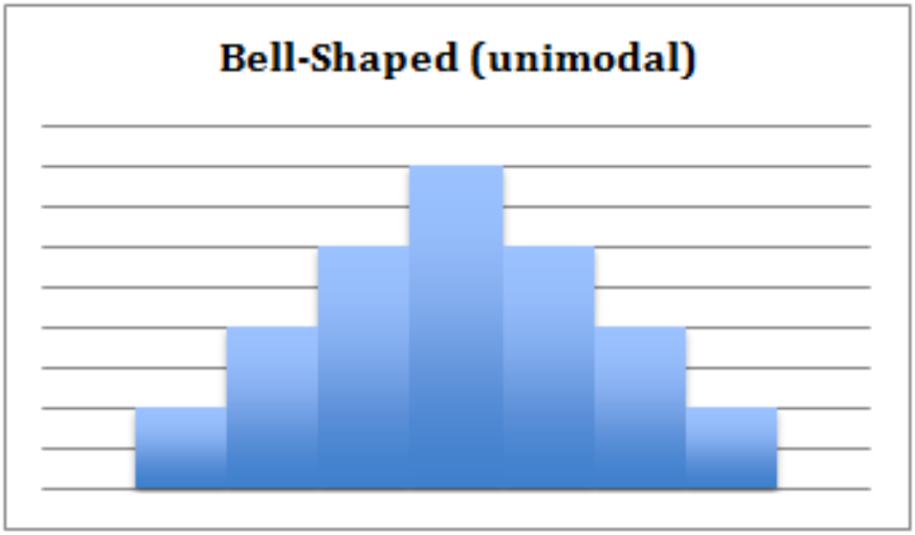
- Learn how to recognize the shape of a distribution from a histogram.
- Learn how to draw appropriate conclusions about the data from its shape.

Distribution Shapes

Histograms are a very common method of visualizing quantitative data, and that means that understanding how to interpret histograms is a valuable and important skill in virtually any career. There are a number of things to pay particular attention to when reading a histogram , including the range of the data and the size of the bins. It is particularly useful to recognize the shape of a histogram because that understanding can lead to valuable conclusions about the nature of the data. In this section, we focus on naming common shapes of distributions and exploring what we can say about the data that have these shapes.

Bell-Shaped

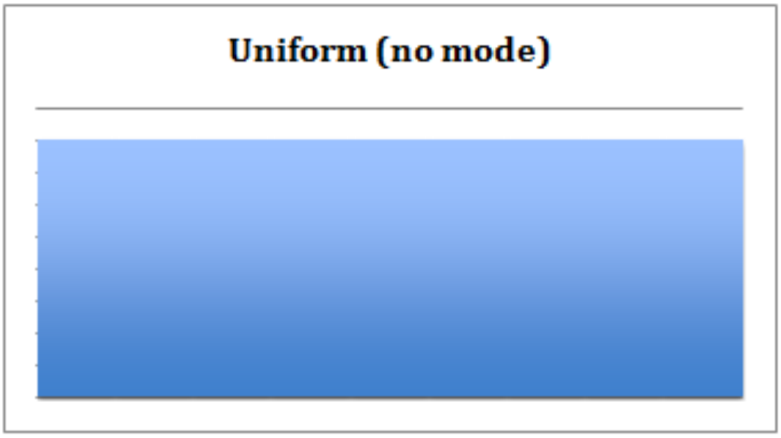
A histogram with a prominent 'mound' in the center and similar tapering to the left and right. One indication of this shape is that the data is **unimodal** –meaning that the data has a single mode, identified by the 'peak' of the curve. Note that a **normally distributed** data set creates a symmetric histogram that looks like a bell, leading to the common term for a normal distribution: a **bell curve**.



Uniform

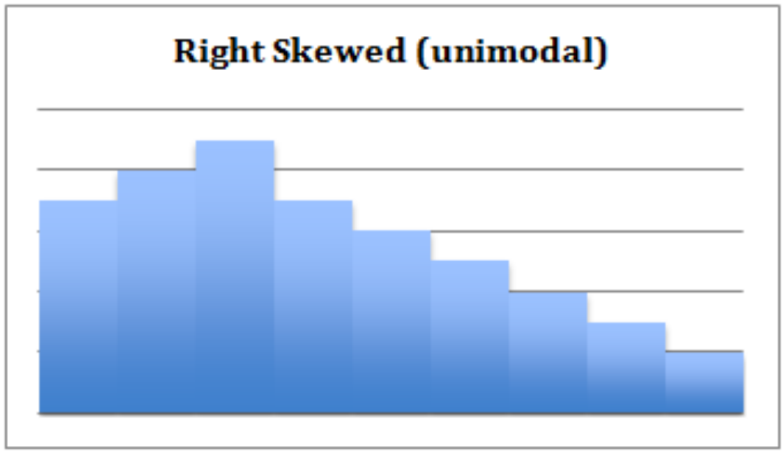
A uniform shaped histogram indicates data that is very consistent; the frequency of each class is very similar to that of the others. A data set with a uniform-shaped histogram may be **multimodal** –the having multiple intervals with the maximum frequency. One indication of a uniform distribution is that the data may not be split into enough

separate intervals or classes. Another possibility is that the scale of the histogram may need to be adjusted in order to offer meaningful observations.



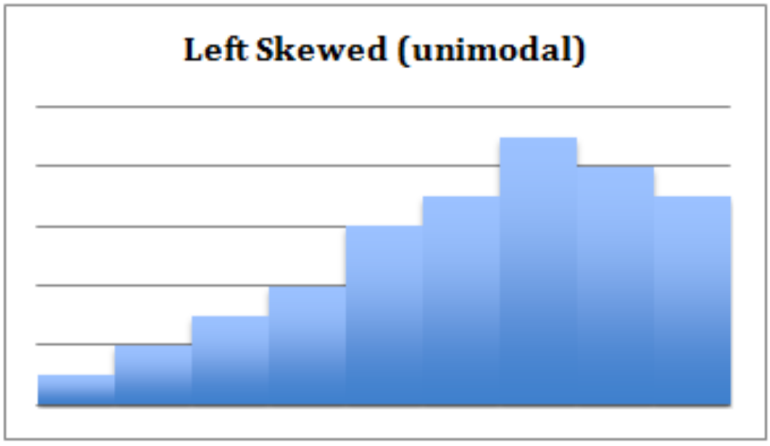
Right-Skewed

A right-skewed histogram has a peak that is left of center and a more gradual tapering to the right side of the graph. This is a unimodal data set. This shape indicates that there are a number of data points, perhaps **outliers**, that are greater than the mode.



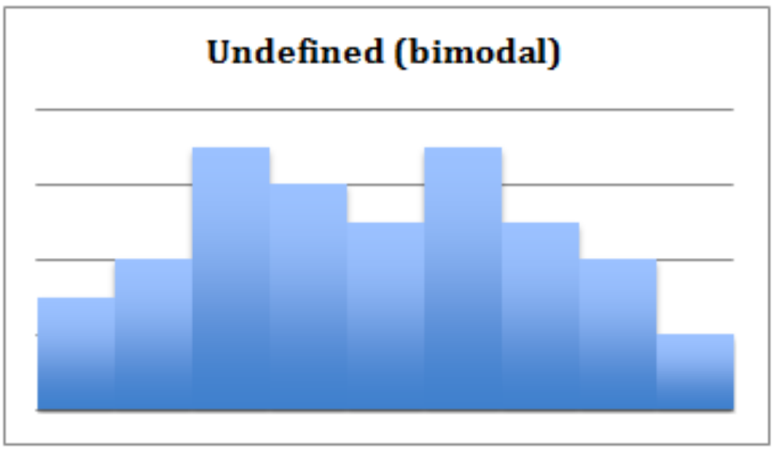
Left-Skewed

A left-skewed histogram has a peak to the right of center, more gradually tapering to the left side. It is unimodal also. This shape indicates that the outliers may be smaller in value that the cluster of more typical values around the mode.



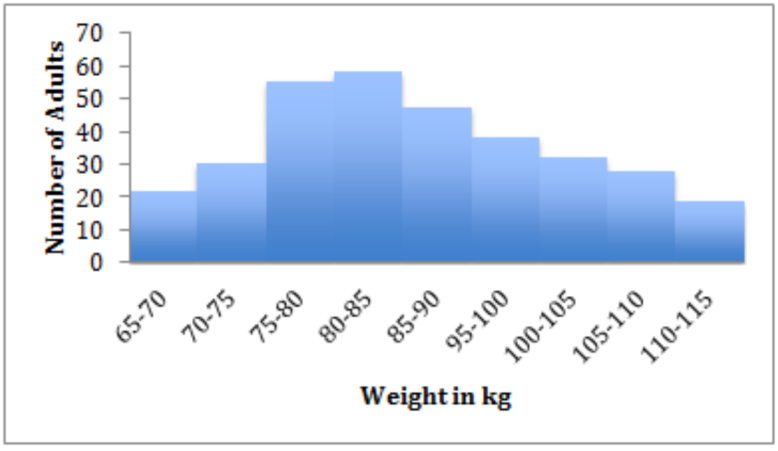
Undefined Bimodal

This shape is not specifically defined, but we can note regardless that it is bi-modal, having two separated classes or intervals equally representing the maximum frequency of the distribution.



Example A

Describe the shape of the histogram and state a few notable characteristics:

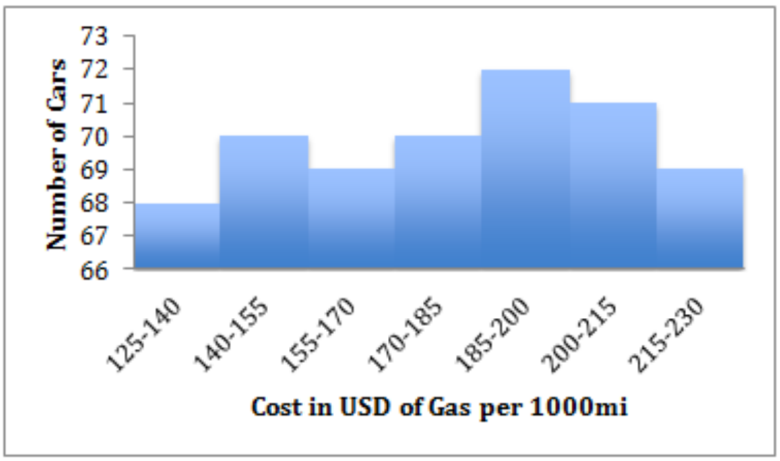


Solution

This is a right-skewed distribution. If the modal class of 80-85kg represents a healthy normal weight, this graph would suggest a sample that tended toward being overweight.

Example B

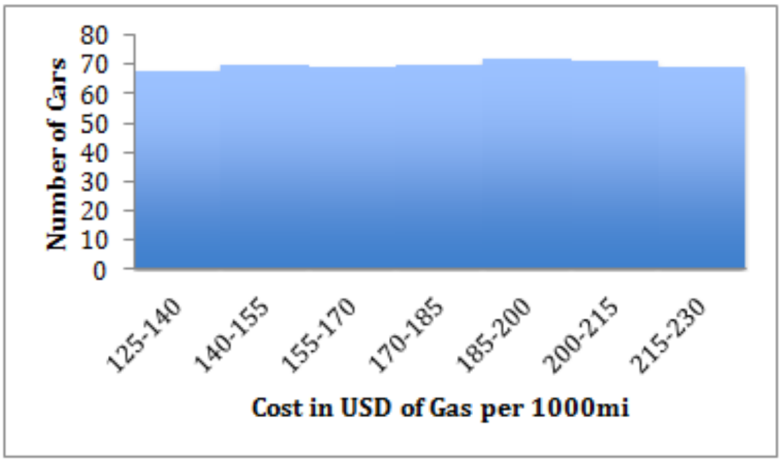
Identify the general shape of the histogram and what the shape indicates about the data:



Solution

This is a slightly tricky one. The overall shape appears somewhat left-skewed and obviously unimodal at first glance. However a closer look tells a different story. The shape is deceiving in large part because the vertical axis does not start at 0, which exaggerates the differences between the classes.

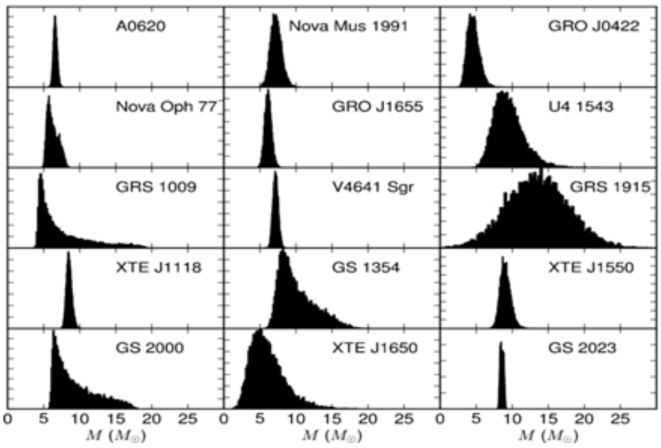
Look what happens if we re-draw the histogram *with the same data* but with the vertical axis at 0:



Pretty huge difference, isn't it? Now it is apparent that this is really a pretty uniform distribution, and that there is not a very meaningful difference in frequency between the classes.

Example C

The image below represents data on the relative masses of a number of sampled black holes.



Data source: http://iopscience.iop.org

Evaluate the group of histograms as a whole; identify the common shapes and any notable features.

Solution

Most of the individual histograms are clearly unimodal, and all are clustered rather closely around a single peak, with the exception of GRS 1915. Most of the graphs appear largely symmetrical, with the others being right-shifted. The sharp and narrow peaks in most of the plots suggest that the mass measurements are generally consistent. The location of the majority of the peaks at the same general location on the scale would suggest that the masses of the different black holes appear similar at this scale. The tendency of the non-symmetrical plots to be right-shifted suggests that it would be more reasonable to favor slightly greater mass estimates than slightly lesser ones.

The GRS 1915 plot is notably different, and the broad peak suggests that perhaps clear data on the mass of that particular black hole is difficult to come by.

Vocabulary

Multimodal histograms have more than one 'peak' in the data. Recall that the mode is the most common value, so a multimodal histogram represents data with multiple classes that have a frequency equal to the greatest single frequency in the data.

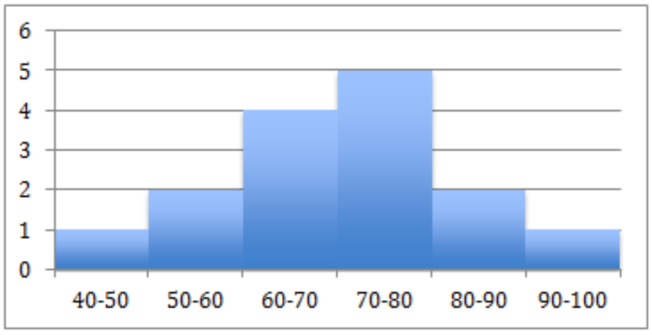
Unimodal histograms have a single peak, and represent data with a single most common frequency.

Outliers are uncommon frequencies occurring some distance from the peak. We will learn how to identify these in later units.

A **normal distribution** creates a histogram in the shape of a bell. This **bell curve** makes it clear that the majority of the data lies close to the mean.

Guided Practice

1. What type of distribution does the histogram below display?



(a) Symmetric, single peaked (unimodal) distribution

(b) Symmetric, double peaked (bimodal) distribution

(c) Skewed left distribution

(d) Skewed right distribution

2. Determine the spread and any outliers of this graph.

Solutions

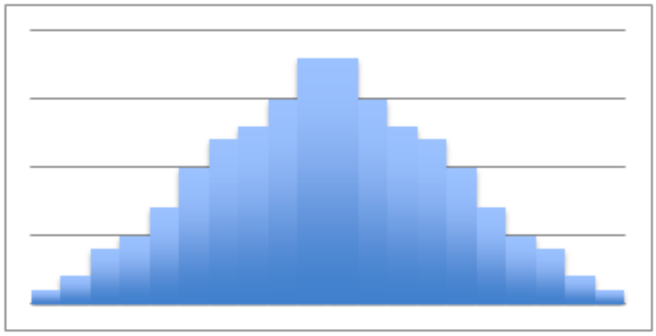
1. This is a symmetric, single peaked (unimodal) distribution.
2. No outliers.

TABLE 2.22:

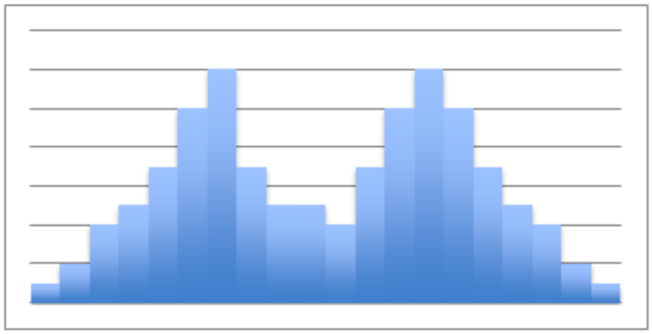
approximate min:	45 (the middle of the lowest interval of scores)
approximate max:	95 (the middle of the hightest interval of scores)
approximate range:	$95 - 45 = 50$

More Practice

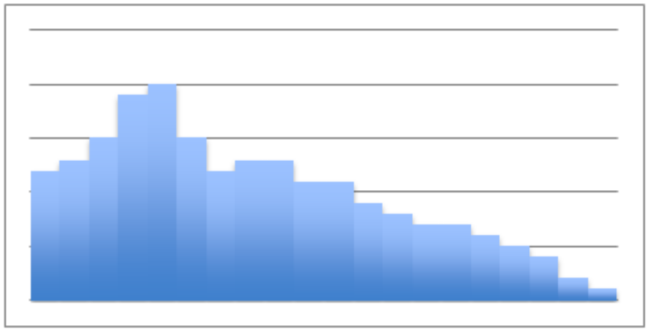
Identify which images show symmetric distributions and which show skewed distributions. Identify what type of symmetric or skewed distributions are displayed.

1. 

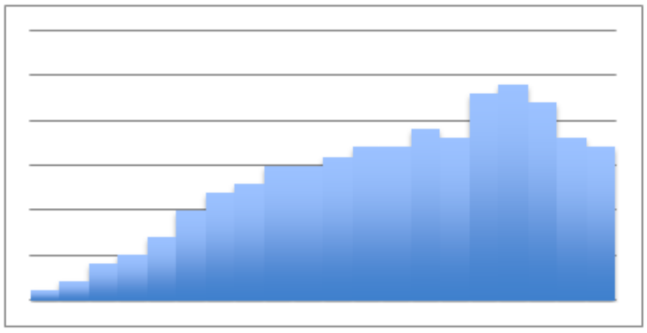
2.



3.



4.



5. What do you think is the shape of the distribution of the age at which a child takes its first steps? Why?

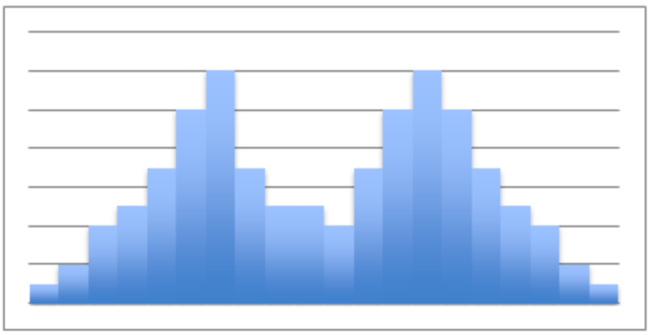
(a) Symetric –Uniform
(b) Skewed left
(c) Skewed right
(d) Symmetric –Unimodal
(e) Symmetric –Bimodal

6. What do you think is the shape of the distribution of rolling a 6-sided die 1,000 times is? Why?

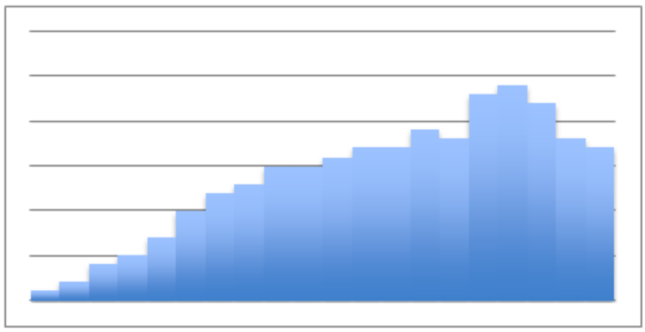
(a) Symmetric –Uniform
(b) Skewed left
(c) Skewed right
(d) Symmetric –Unimodal
(e) Symmetric Bimodal?

7. Match the graph with the data it most likely displays.

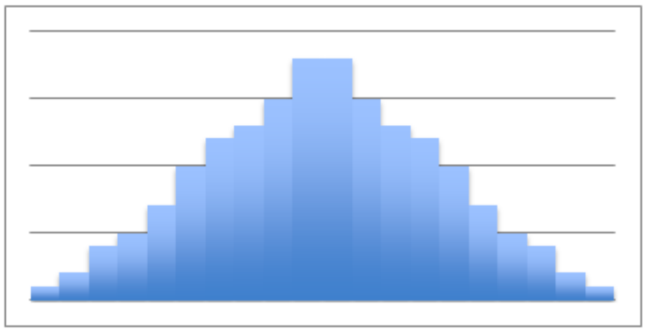
(a)



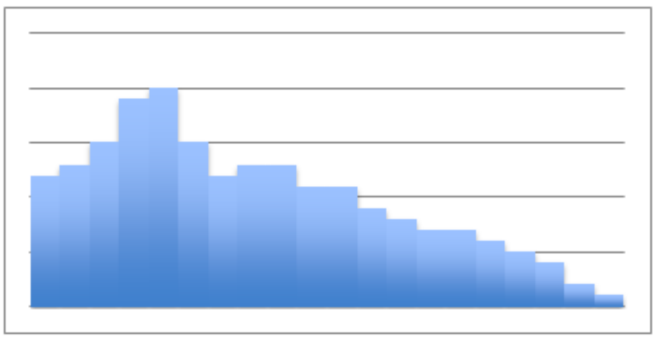
(b)



(c)



(d)



(e) SAT Math Scores of future doctors and engineers.

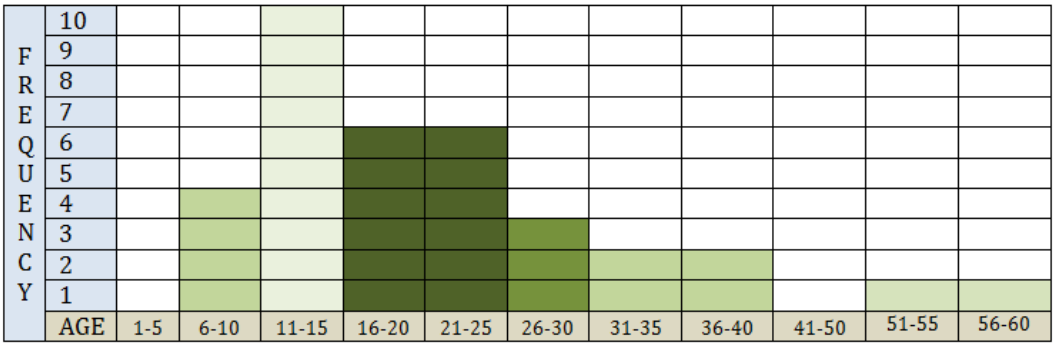
(f) Prices of 1,000 homes within a given geographical area.

(g) Cholesterol levels of 1000 adults.

(h) Men's women's clothing sizes.

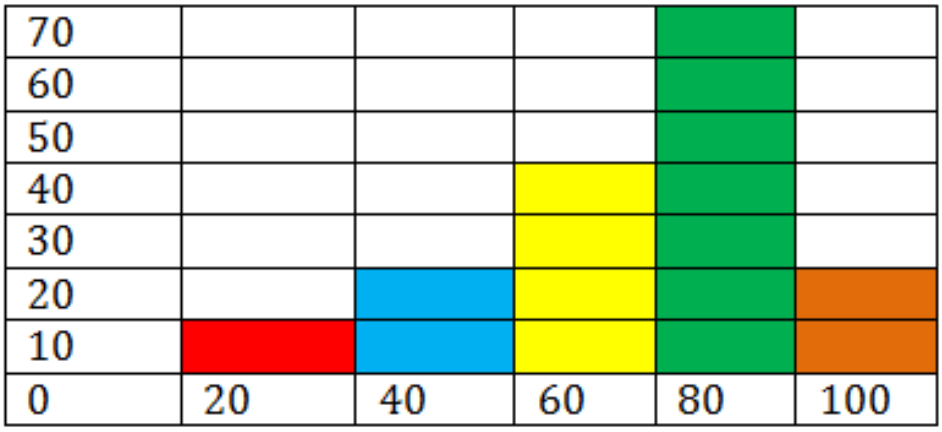
(i) The data below shows the number of surveyed people, and their respective ages, who enjoy riding roller coasters.

Use the histogram below to answer questions 8-11.



8. What is the shape of this histogram?
9. What is the center of this histogram?
10. What is the Spread of this histogram?
11. What are the outliers of this histogram?

Use the histogram below to answer questions 12-15.



12. What is the shape of this histogram?
13. What is the center of this histogram?
14. What is the spread of this histogram?