1.2 Describing Data

Once we have collected data from surveys or experiments, we need to summarize and present the data in a way that will be meaningful to the reader. We will begin with graphical presentations of data then explore numerical summaries of data.

1.2.1 Presenting Categorical Data Graphically

Categorical, or qualitative, data are pieces of information that allow us to classify the objects under investigation into various categories. We usually begin working with categorical data by summarizing the data into a **frequency table.**

Frequency Table
A frequency table is a table with two columns. One column lists the categories, and another for the frequencies with which the items in the categories occur (how many items fit into each category).

Example 1.2.1

An insurance company determines vehicle insurance premiums based on known risk factors. If a person is considered a higher risk, their premiums will be higher. One potential factor is the color of your car. The insurance company believes that people with some color cars are more likely to get in accidents. To research this, they examine police reports for recent total-loss collisions. The data is summarized in the frequency table below.

Color	Frequency
Blue	25
Green	52
Red	41
White	36
Black	39
Grey	23

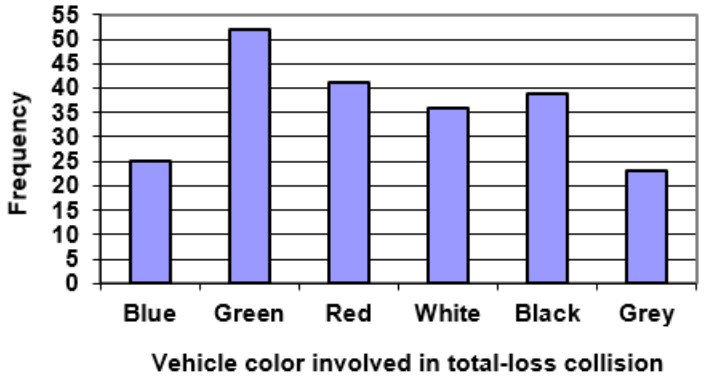
Sometimes we need an even more intuitive way of displaying data. This is where charts and graphs come in. There are many, many ways of displaying data graphically, but we will concentrate on one very useful type of graph called a bar graph. In this section we will work with bar graphs that display categorical data; the next section will be devoted to bar graphs that display quantitative data.

Bar graph
A **bar graph** is a graph that displays a bar for each category with the length of each bar indicating the frequency of that category.

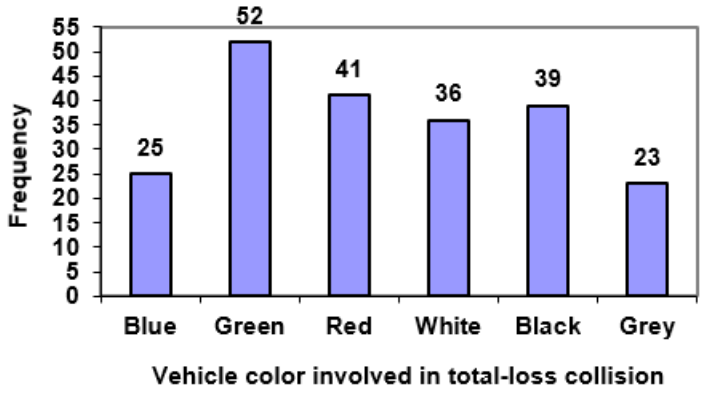
To construct a bar graph, we need to draw a vertical axis and a horizontal axis. The vertical direction will have a scale and measure the frequency of each category; the horizontal axis has no scale in this instance. The construction of a bar chart is most easily described by use of an example.

Example 1.2.2

Using our car data from above, note the highest frequency is 52, so our vertical axis needs to go from 0 to 52, but we might as well use 0 to 55, so that we can put a hash mark every 5 units:



Notice that the height of each bar is determined by the frequency of the corresponding color. The horizontal gridlines are a nice touch, but not necessary. In practice, you will find it useful to draw bar graphs using graph paper, so the gridlines will already be in place, or using technology. Instead of gridlines, we might also list the frequencies at the top of each bar, like this:

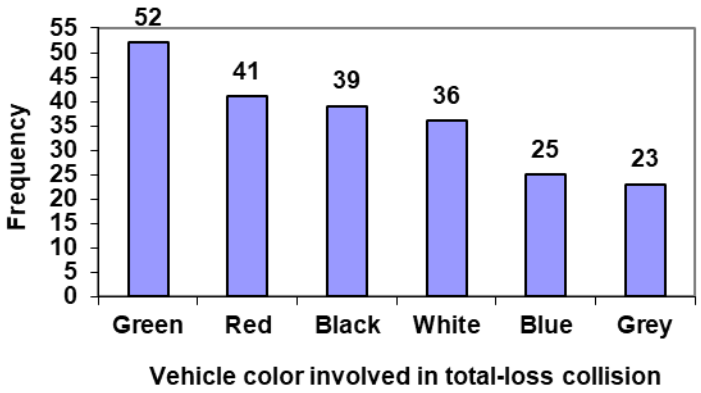


In this case, our chart might benefit from being reordered from largest to smallest frequency values. This arrangement can make it easier to compare similar values in the chart, even without gridlines. When we arrange the categories in decreasing frequency order like this, it is called a **Pareto chart**.

Pareto chart
A **Pareto chart** is a bar graph ordered from highest to lowest frequency

Example 1.2.3

Transforming our bar graph from earlier into a Pareto chart, we get:



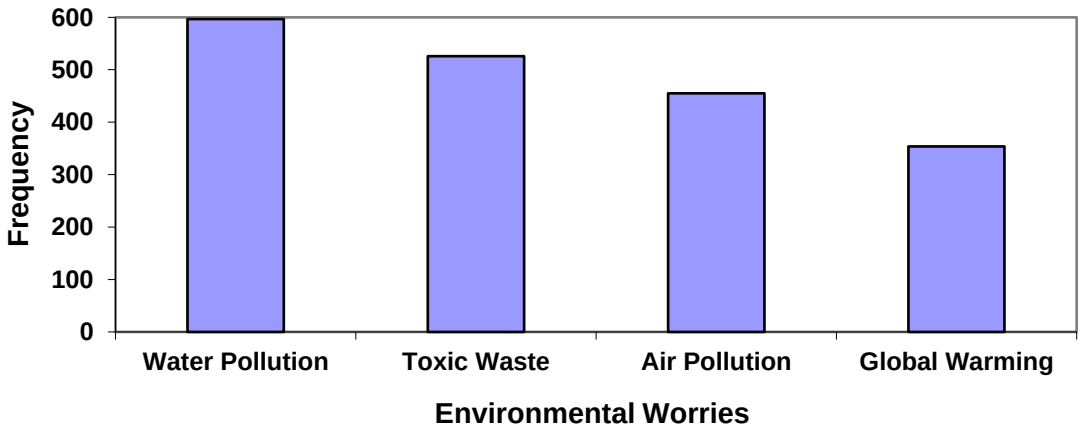
Example 1.2.4

In a survey[1], adults were asked whether they personally worried about a variety of environmental concerns. The numbers (out of 1012 surveyed) who indicated that they worried "a great deal" about some selected concerns are summarized below.

Environmental Issue	**Frequency**
Pollution of drinking water	597
Contamination of soil and water by toxic waste	526
Air pollution	455
Global warming	354

This data could be shown graphically in a bar graph:

[1] Gallup Poll. March 5-8, 2009. http://www.pollingreport.com/enviro.htm



To show relative sizes, it is common to use a pie chart.

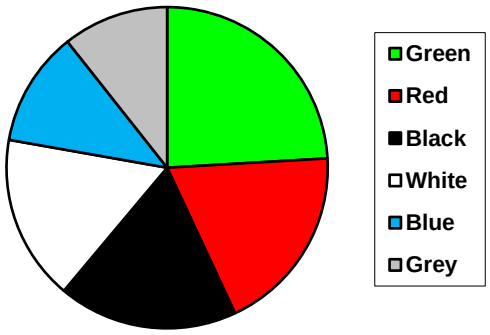
Pie Chart

A **pie chart** is a circle with wedges cut of varying sizes marked out like slices of pie or pizza. The relative sizes of the wedges correspond to the relative frequencies of the categories.

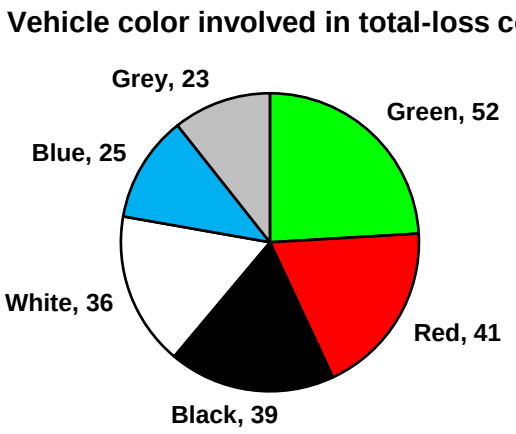
Example 1.2.5

For our vehicle color data, a pie chart might look like this:

Vehicle color involved in total-loss collisions

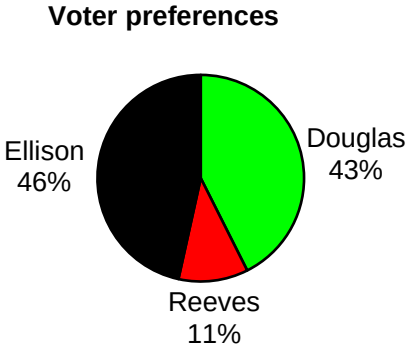


Pie charts can often benefit from including frequencies or relative frequencies (percents) in the chart next to the pie slices. Often having the category names next to the pie slices also makes the chart clearer.



Example 1.2.6

The pie chart below shows the percentage of voters supporting each candidate running for a local senate seat.



If there are 20,000 voters in the district, the pie chart shows that about 11% of those, about 2,200 voters, support Reeves.

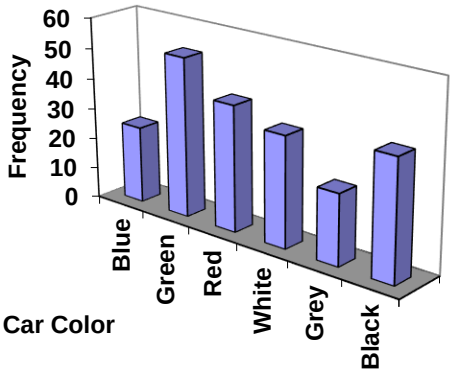
Pie charts look nice, but are harder to draw by hand than bar charts since to draw them accurately we would need to compute the angle each wedge cuts out of the circle, then measure the angle with a protractor. Computers are much better suited to drawing pie charts. Common software programs like Microsoft Word or Excel, OpenOffice.org Write or Calc, or Google Docs are able to create bar graphs, pie charts, and other graph types. There are also numerous online tools that can create graphs[2].

Try it Now 1.2.1

Create a bar graph and a pie chart to illustrate the grades on a history exam below.
A: 12 students, B: 19 students, C: 14 students, D: 4 students, F: 5 students

[2] For example: http://nces.ed.gov/nceskids/createAgraph/ or http://docs.google.com

Don't get fancy with graphs! People sometimes add features to graphs that don't help to convey their information. For example, 3-dimensional bar charts like the one shown below are usually not as effective as their two-dimensional counterparts.



Here is another way that fanciness can lead to trouble. Instead of plain bars, it is tempting to substitute meaningful images. This type of graph is called a **pictogram**.

Pictogram
A **pictogram** is a statistical graphic in which the size of the picture is intended to represent the frequencies or size of the values being represented.

Example 1.2.7

A labor union might produce the graph to the right to show the difference between the average manager salary and the average worker salary.

Looking at the picture, it would be reasonable to guess that the manager salaries is 4 times as large as the worker salaries – the area of the bag looks about 4 times as large. However, the manager salaries are in fact only twice as large as worker salaries, which were reflected in the picture by making the manager bag twice as tall.

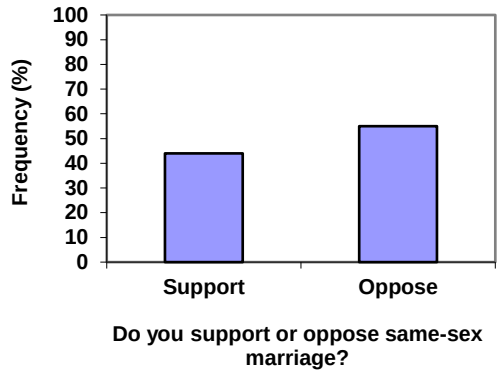
Manager Salaries **Worker Salaries**

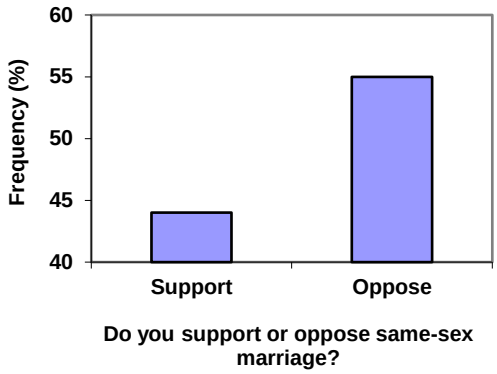
Another distortion in bar charts results from setting the baseline to a value other than zero. The baseline is the bottom of the vertical axis, representing the least number of cases that could have occurred in a category. Normally, this number should be zero.

Example 1.2.8

Compare the two graphs below showing support for same-sex marriage rights from a poll taken in December 2008[3]. The difference in the vertical scale on the first graph suggests a different story than the true differences in percentages; the second graph makes it look like twice as many people oppose marriage rights as support it.

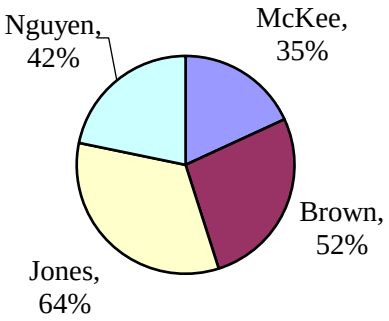
[3]CNN/Opinion Research Corporation Poll. Dec 19-21, 2008, from http://www.pollingreport.com/civil.htm





Try it Now 1.2.2

A poll was taken asking people if they agreed with the positions of the 4 candidates for a county office. Does the pie chart present a good representation of this data? Explain.



1.2.2 Presenting Quantitative Data Graphically

Quantitative, or numerical, data can also be summarized into frequency tables.

Example 1.2.9

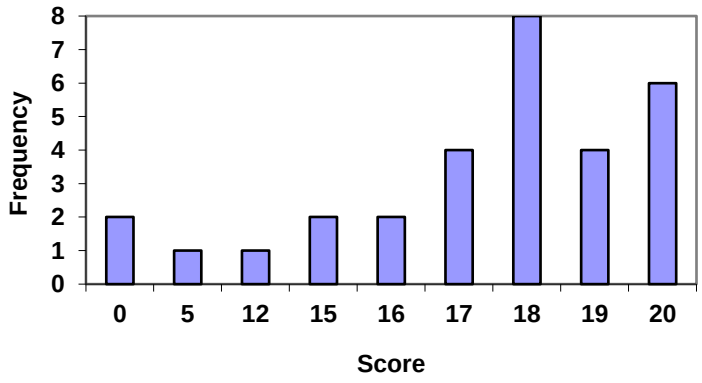
A teacher records scores on a 20-point quiz for the 30 students in his class. The scores are:

19 20 18 18 17 18 19 17 20 18 20 16 20 15 17 12 18 19 18 19 17 20 18 16 15 18 20 5 0 0

These scores could be summarized into a frequency table by grouping like values:

Score	Frequency
0	2
5	1
12	1
15	2
16	2
17	4
18	8
19	4
20	6

Using this table, it would be possible to create a standard bar chart from this summary, like we did for categorical data:

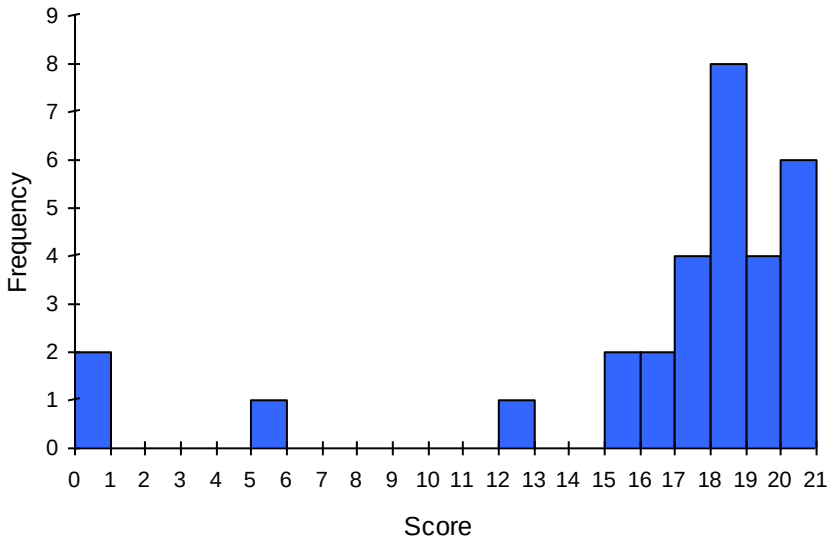


However, since the scores are numerical values, this chart doesn't really make sense; the first and second bars are five values apart, while the later bars are only one value apart. It would be more correct to treat the horizontal axis as a number line. This type of graph is called a **histogram**.

Histogram
A histogram is like a bar graph, but where the horizontal axis is a number line

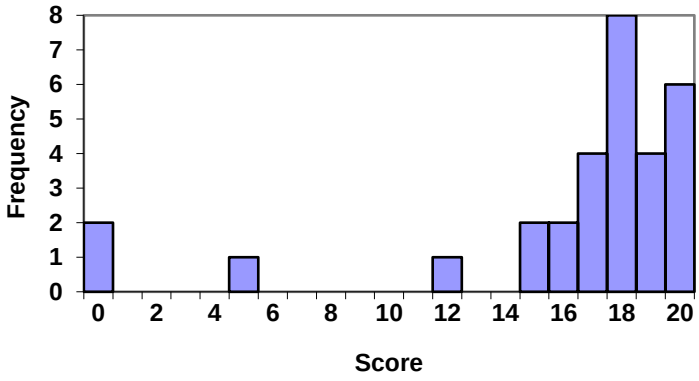
Example 1.2.10

For the values above, a histogram would look like:



Notice that in the histogram, a bar represents values on the horizontal axis from that on the left hand-side of the bar up to, but not including, the value on the right hand side of the bar. Some people choose to have bars start at ½ values to avoid this ambiguity.

Unfortunately, not a lot of common software packages can correctly graph a histogram. About the best you can do in Excel or Word is a bar graph with no gap between the bars and spacing added to simulate a numerical horizontal axis.



If we have a large number of widely varying data values, creating a frequency table that lists every possible value as a category would lead to an exceptionally long frequency table, and probably would not reveal any patterns. For this reason, it is common with quantitative data to group data into **class intervals**.

Class Intervals

Class intervals are groupings of the data. In general, we define class intervals so that:

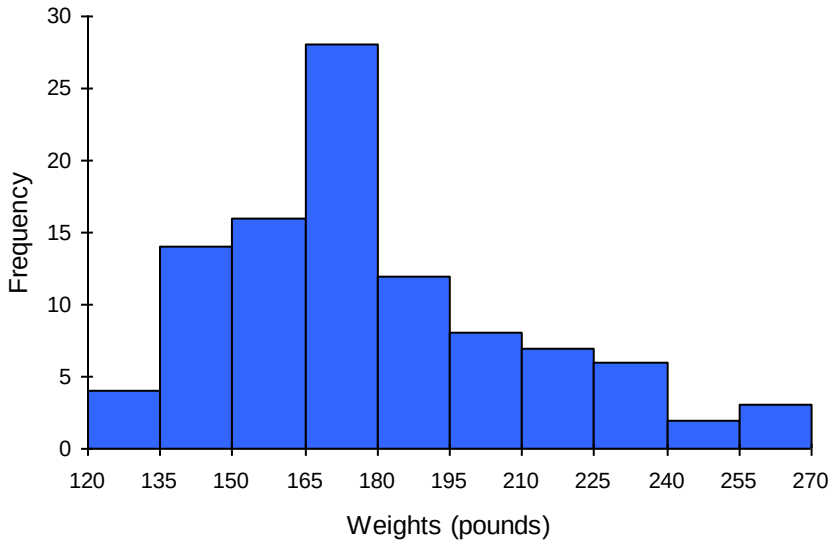
- Each interval is equal in size. For example, if the first class contains values from 120-129, the second class should include values from 130-139.
- We have somewhere between 5 and 20 classes, typically, depending upon the number of data we're working with.

Example 1.2.11

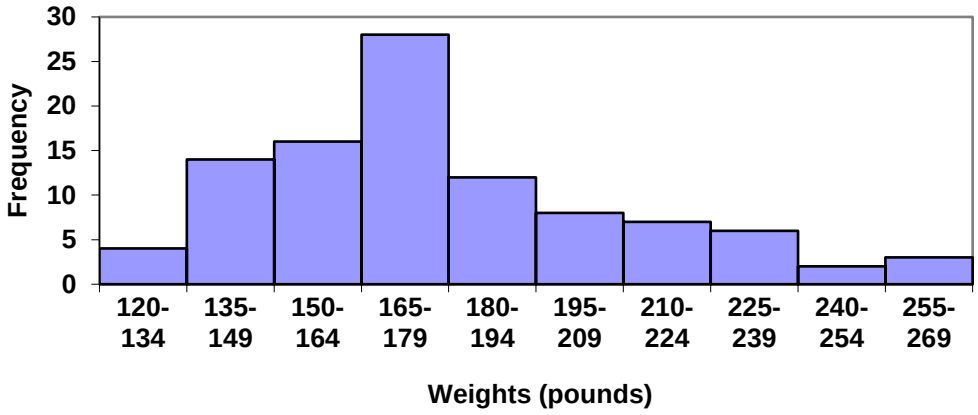
Suppose that we have collected weights from 100 male subjects as part of a nutrition study. For our weight data, we have values ranging from a low of 121 pounds to a high of 263 pounds, giving a total span of 263-121 = 142. We could create 7 intervals with a width of around 20, 14 intervals with a width of around 10, or somewhere in between. Often time we have to experiment with a few possibilities to find something that represents the data well. Let us try using an interval width of 15. We could start at 121, or at 120 since it is a nice round number.

Interval	Frequency
120 - 134	4
135 – 149	14
150 – 164	16
165 – 179	28
180 – 194	12
195 – 209	8
210 – 224	7
225 – 239	6
240 – 254	2
255 - 269	3

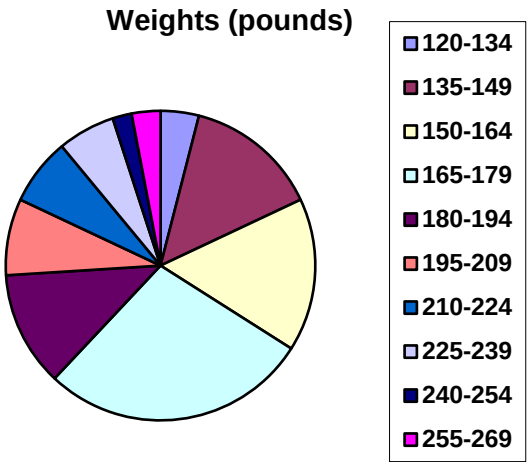
A histogram of this data would look like:



In many software packages, you can create a graph similar to a histogram by putting the class intervals as the labels on a bar chart.



Other graph types such as pie charts are possible for quantitative data. The usefulness of different graph types will vary depending upon the number of intervals and the type of data being represented. For example, a pie chart of our weight data is difficult to read because of the quantity of intervals we used.



Try it Now 1.2.3

The total cost of textbooks for the term was collected from 36 students. Create a histogram for this data.

$140	$160	$160	$165	$180	$220	$235	$240	$250	$260	$280	$285
$285	$285	$290	$300	$300	$305	$310	$310	$315	$315	$320	$320
$330	$340	$345	$350	$355	$360	$360	$380	$395	$420	$460	$460

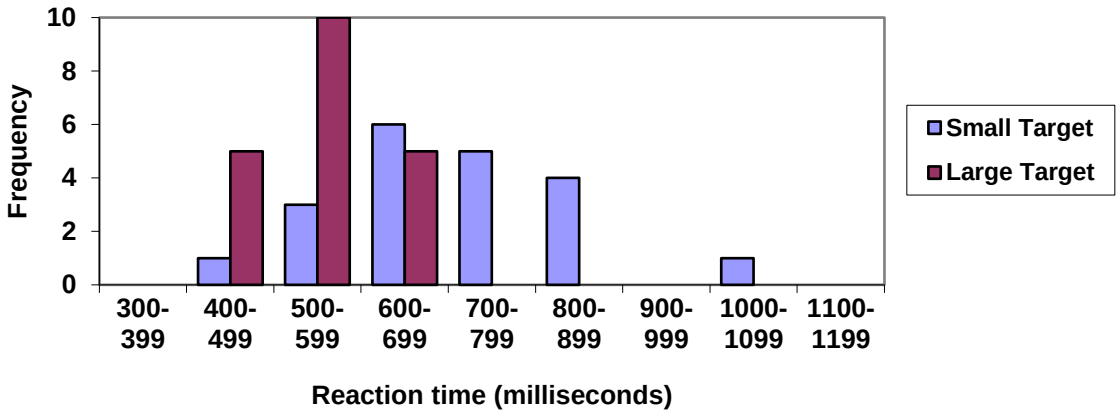
When collecting data to compare two groups, it is desirable to create a graph that compares quantities.

Example 1.2.12

The data below came from a task in which the goal is to move a computer mouse to a target on the screen as fast as possible. On 20 of the trials, the target was a small rectangle; on the other 20, the target was a large rectangle. Time to reach the target was recorded on each trial.

Interval (milliseconds)	Frequency small target	Frequency large target
300-399	0	0
400-499	1	5
500-599	3	10
600-699	6	5
700-799	5	0
800-899	4	0
900-999	0	0
1000-1099	1	0
1100-1199	0	0

One option to represent this data would be a comparative histogram or bar chart, in which bars for the small target group and large target group are placed next to each other.

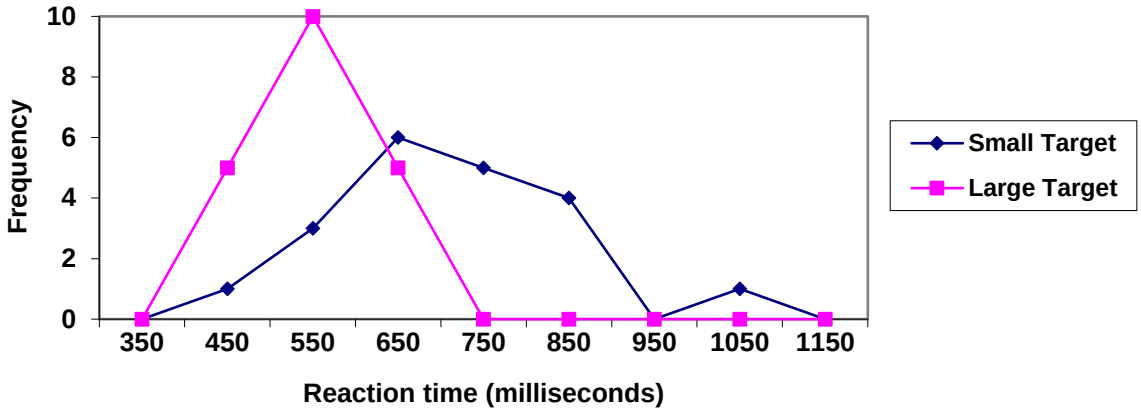


Frequency polygon

An alternative representation is a **frequency polygon**. A frequency polygon starts out like a histogram, but instead of drawing a bar, a point is placed in the midpoint of each interval at height equal to the frequency. Typically the points are connected with straight lines to emphasize the distribution of the data.

Example 1.2.13

This graph makes it easier to see that reaction times were generally shorter for the larger target, and that the reaction times for the smaller target were more spread out.



1.2.3 Numerical Summaries of Data

It is often desirable to use a few numbers to summarize a distribution. One important aspect of a distribution is where its center is located. Measures of central tendency are discussed first. A second aspect of a distribution is how spread out it is. In other words, how much the data in the distribution vary from one another. The second section describes measures of variability.

1.2.4 Measures of Central Tendency

Let's begin by trying to find the most "typical" value of a data set.

Note that we just used the word "typical" although in many cases you might think of using the word "average." We need to be careful with the word "average" as it means different things to different people in different contexts. One of the most common uses of the word "average" is what mathematicians and statisticians call the **arithmetic mean**, or just plain old **mean** for short. "Arithmetic mean" sounds rather fancy, but you have likely calculated a mean many times without realizing it; the mean is what most people think of when they use the word "average".

Mean

The **mean** of a set of data is the sum of the data values divided by the number of values.

Example 1.2.14

Marci's exam scores for her last math class were: 79, 86, 82, 94. The mean of these values would be:

$$\frac{79+86+82+94}{4}=85.25.$$

Example 1.2.15

The number of touchdown (TD) passes thrown by each of the 31 teams in the National Football League in the 2000 season are shown below.

37 33 33 32 29 28 28 23 22 22 22 21 21 21 20
20 19 19 18 18 18 18 16 15 14 14 14 12 12 9 6

Adding these values, we get 634 total TDs. Dividing by 31, the number of data values, we get 634/31 = 20.4516. It would be appropriate to round this to 20.5.

It would be most correct for us to report that "The mean number of touchdown passes thrown in the NFL in the 2000 season was 20.5 passes," but it is not uncommon to see the more casual word "average" used in place of "mean."

Try it Now 1.2.4

The price of a jar of peanut butter at 5 stores was: $3.29, $3.59, $3.79, $3.75, and $3.99. Find the mean price.

Example 1.2.16

The one hundred families in a particular neighborhood are asked their annual household income, to the nearest $5 thousand dollars. The results are summarized in a frequency table below.

Income (thousands of dollars)	Frequency
15	6
20	8
25	11
30	17
35	19
40	20
45	12
50	7

Calculating the mean by hand could get tricky if we try to type in all 100 values.

We could calculate this more easily by noticing that adding 15 to itself six times is the same as $15 \times 6 = 90$. Using this simplification, we get

$$\frac{15\times 6+20\times 8+25\times 11+30\times 17+35\times 19+40\times 20+45\times 12+50\times 7}{100}=\frac{3390}{100}$$

This simplifies to 33.9.

The mean household income of our sample is 33.9 thousand dollars ($33,900).

Median

The **median** of a set of data is the value in the middle when the data is in order

To find the median, begin by listing the data in order from smallest to largest, or largest to smallest.

Count the number of data values (N). Then find (N+1)/2.

If the number of data values, N, is odd, then (N+1)/2 is an integer. The integer is the position of the median in the data set.

If the number of data values is even, then (N+1)/2 is a decimal such as 8.5 (it will always have just one decimal place, with a 5 in the that position). If it were 8.5, as an example, that tells you that the median is the value in between the 8^{th} and 9^{th} data values.

Example 1.2.17

Returning to the football touchdown data, we would start by listing the data in order. Luckily, it was already in decreasing order, so we can work with it without needing to reorder it first.

37 33 33 32 29 28 28 23 22 22 22 21 21 21 20
20 19 19 18 18 18 18 16 15 14 14 14 12 12 9 6

Since there are 31 data values, an odd number, the median will be the middle number, the 16^{th} data value (31/2 = 15.5, round up to 16, leaving 15 values below and 15 above). The 16^{th} data value is 20, so the median number of touchdown passes in the 2000 season was 20 passes. Notice that for this data, the median is fairly close to the mean we calculated earlier, 20.5.

Example 1.2.18

Find the median of these quiz scores: 5 10 8 6 4 8 2 5 7 7

We start by listing the data in order: 2 4 5 5 6 7 7 8 8 10

Since there are 10 data values, an even number, there is no one middle number. So we find the mean of the two middle numbers, 6 and 7, and get (6+7)/2 = 6.5.

The median quiz score was 6.5.

Try it Now 1.2.5

The price of a jar of peanut butter at 5 stores were: \$3.29, \$3.59, \$3.79, \$3.75, and \$3.99. Find the median price.

Example 1.2.19

Let us return now to our original household income data

Income (thousands of dollars)	Frequency
15	6
20	8
25	11
30	17
35	19
40	20
45	12
50	7

Here we have 100 data values. If we didn't already know that, we could find it by adding the frequencies. Since 100 is an even number, we need to find the mean of the middle two data values - the 50^{th} and 51^{st} data values. To find these, we start counting up from the bottom:

There are 6 data values of \$15, so	Values 1 to 6 are \$15 thousand
The next 8 data values are \$20, so	Values 7 to (6+8)=14 are \$20 thousand
The next 11 data values are \$25, so	Values 15 to (14+11)=25 are \$25 thousand
The next 17 data values are \$30, so	Values 26 to (25+17)=42 are \$30 thousand
The next 19 data values are \$35, so	Values 43 to (42+19)=61 are \$35 thousand

From this we can tell that values 50 and 51 will be \$35 thousand, and the mean of these two values is \$35 thousand. The median income in this neighborhood is \$35 thousand.

In addition to the mean and the median, there is one other common measurement of the "typical" value of a data set: the **mode**.

Mode
The **mode** is the element of the data set that occurs most frequently.

The mode is fairly useless with data like weights or heights where there are a large number of possible values. The mode is most commonly used for categorical data, for which median and mean cannot be computed.

Example 1.2.20

In our vehicle color survey, we collected the data

Color	Frequency
Blue	3
Green	5
Red	4
White	3
Black	2
Grey	3

For this data, Green is the mode, since it is the data value that occurred the most frequently.

It is possible for a data set to have more than one mode if several categories have the same frequency, or no modes if each every category occurs only once. We will only use no mode, one mode, or two modes.

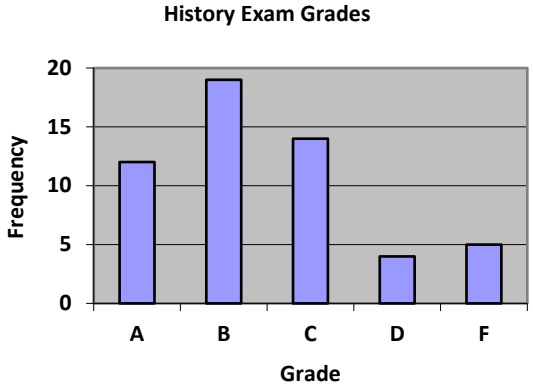
Try it Now 1.2.6

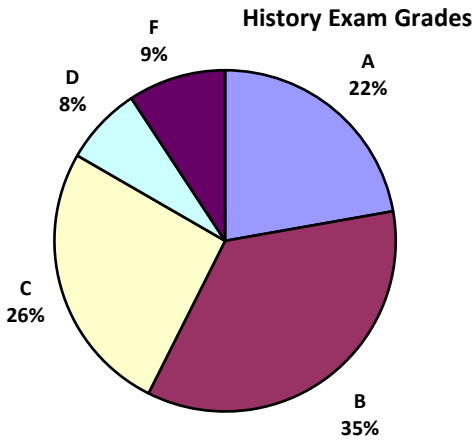
Reviewers were asked to rate a product on a scale of 1 to 5. Find
a. The mean rating
b. The median rating
c. The mode rating

Rating	Frequency
1	4
2	8
3	7
4	3
5	1

Try it Now Answers

1.2.1.





1.2.2. While the pie chart accurately depicts the relative size of the people agreeing with each candidate, the chart is confusing, since usually percents on a pie chart represent the percentage of the pie the slice represents.

1.2.3. Using a class intervals of size 55, we can group our data into six intervals:

Cost interval	Frequency
$140-194	5
$195-249	3
$250-304	9
$305-359	12
$360-414	4
$415-469	3

We can use the frequency distribution to generate the histogram

1.2.4. Adding the prices and dividing by 5 we get the mean price: $3.682

1.2.5. First we put the data in order: $3.29, $3.59, $3.75, $3.79, $3.99. Since there are an odd number of data, the median will be the middle value, $3.75.

1.2.6. There are 23 ratings.

a. The mean is $\frac{1\cdot 4+2\cdot 8+3\cdot 7+4\cdot 3+5\cdot 1}{23}\approx 2.5$

b. There are 23 data values, so the median will be the 12^{th} data value. Ratings of 1 are the first 4 values, while a rating of 2 are the next 8 values, so the 12^{th} value will be a rating of 2. The median is 2.

c. The mode is the most frequent rating. The mode rating is 2.

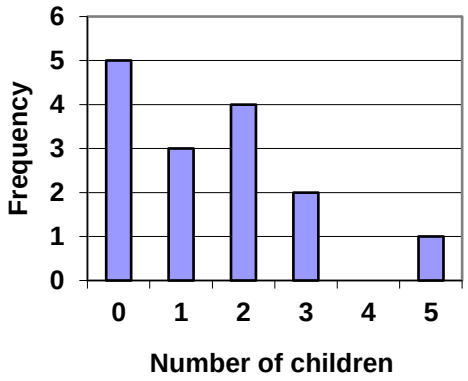
Section 1.2 Exercises

Skills

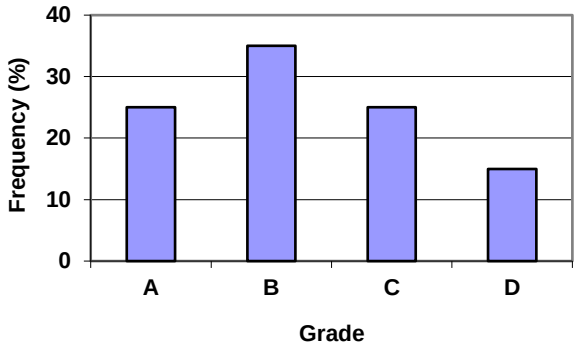
1. The table below shows scores on a Math test.
    a. Complete the frequency table for the Math test scores
    b. Construct a histogram of the data

80	50	50	90	70	70	100	60	70	80	70	50
90	100	80	70	30	80	80	70	100	60	60	50

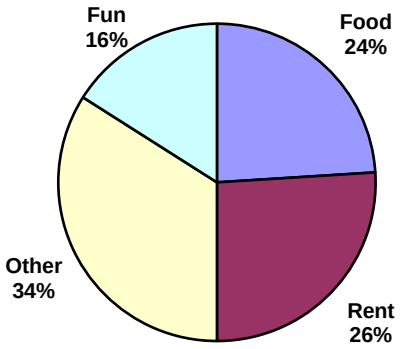
2. A group of adults were asked how many children they have in their families. The bar graph to the right shows the number of adults who indicated each number of children.
    a. How many adults where questioned?
    b. What percentage of the adults questioned had 0 children?



3. The bar graph below shows the *percentage* of students who received each letter grade on their last English paper. The class contains 20 students. What number of students earned an A on their paper?



4. Kori categorized her spending for this month into four categories: Rent, Food, Fun, and Other. The percents she spent in each category are pictured here. If she spent a total of $2600 this month, how much did she spend on rent?



5. Fifty individuals were surveyed and asked to identify the number of children they had living in their home. The responses are summarized in the following table. Determine the mean, median, and mode of the data. 5. Twenty individuals were surveyed and asked to identify the number of times they have undergone an intensive medical surgery. The responses are summarized in the following table. Determine the mean, median, and mode of the data.

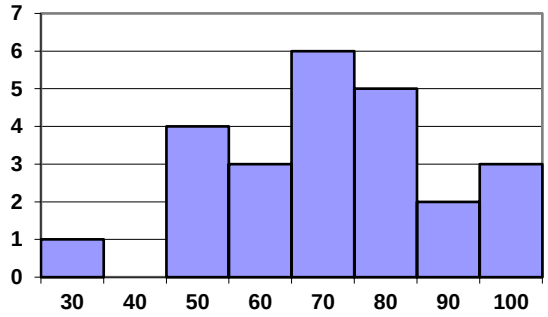
Number of Children	Frequency
0	21
1	6
2	12
3	9
5	2

Section 1.2 Exercises – Answer Key

1. a.

Score	Frequency
30	1
40	0
50	4
60	3
70	6
80	5
90	2
100	3

b. This is technically a bar graph, not a histogram:



2. a. 5+3+4+2+1 = 15
b. 5/15 = 0.3333 = 33.33%

3. Bar is at 25%. 25% of 20 = 5 students earned an A

4. 26% is the same as 0.26. $0.26 \times 2600 = \$676.00$

5. mean: 1.34 median: 1 mode: 0