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.

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** which is just a table with two columns, one for categories and another for the frequencies with which the items in the categories occur.

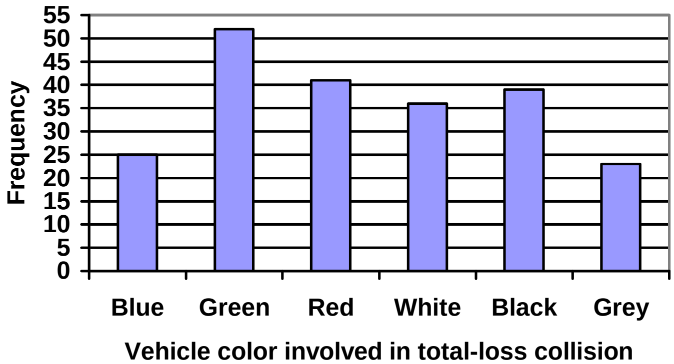
Example: An insurance company determines vehicle insurance premiums based on known risk factors. If a person is considered a higher risk, their premiums will be higher. One potential factor is the color of your car. The insurance company believes that people with some color cars are more likely to get 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

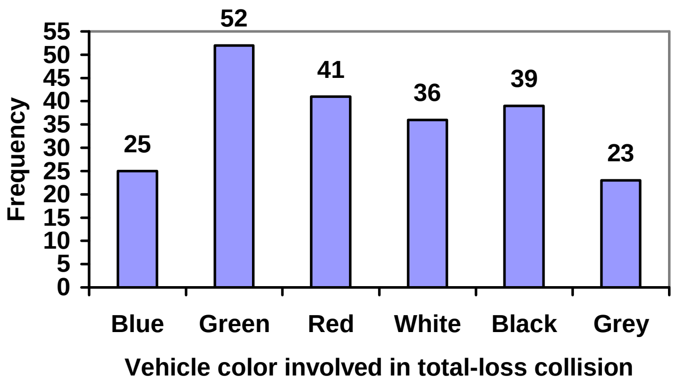
Sometimes we need an even more intuitive way of displaying data. This is where charts and graphs come in. There are many, many ways of displaying data graphically, but we will concentrate on one very useful type of graph called a bar graph. In this section we will work with bar graphs that display categorical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

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.

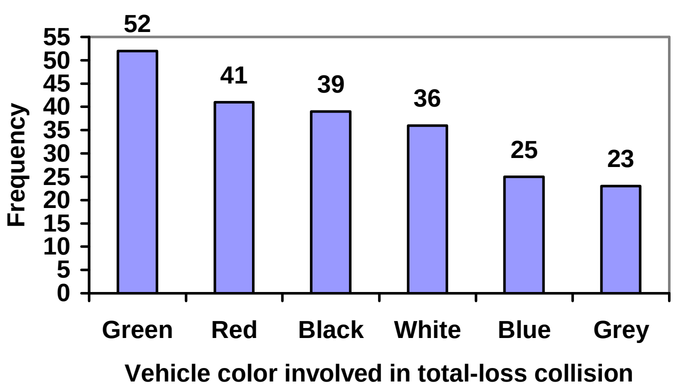
Using our car data from above, note the highest frequency is 52, so our vertical axis needs to go from 0 to 52, but we might as well use 0 to 55, so that we can put a hash mark every 5 units:



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. Instead of gridlines, we might also list the frequencies at the top of each bar, like this:



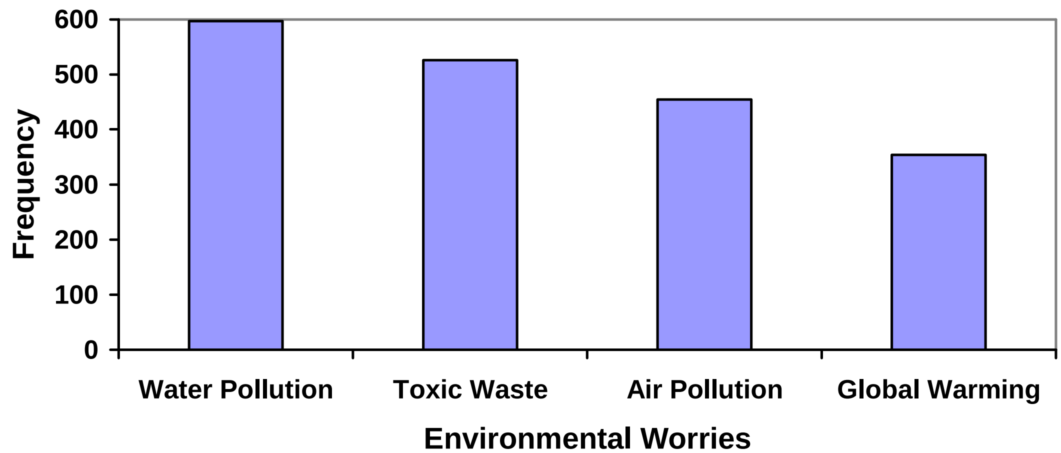
In this case, our chart might benefit from being reordered from largest to smallest frequency values. This arrangement can make it easier to compare similar values in the chart, even without gridlines. When we arrange the categories in decreasing frequency order like this, it is called a **Pareto chart**.



Example: In a recent survey[1], adults were asked whether they personally worried about a variety of environmental concerns. The number (out of 1012 surveyed) indicating that they worried "a great deal" about some selected concerns is summarized below.

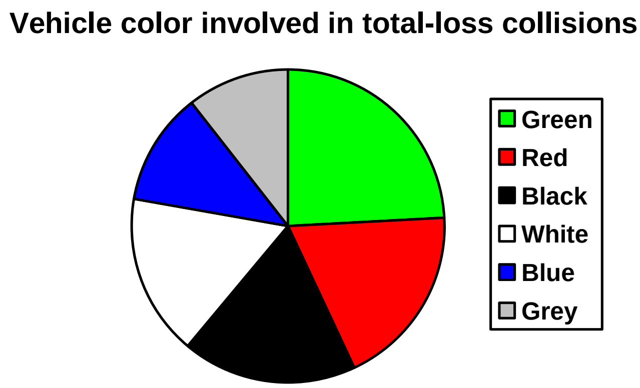
Environmental Issue	Frequency
Pollution of drinking water	597
Contamination of soil and water by toxic waste	526
Air pollution	455
Global warming	354

This data could be shown graphically in a bar graph:



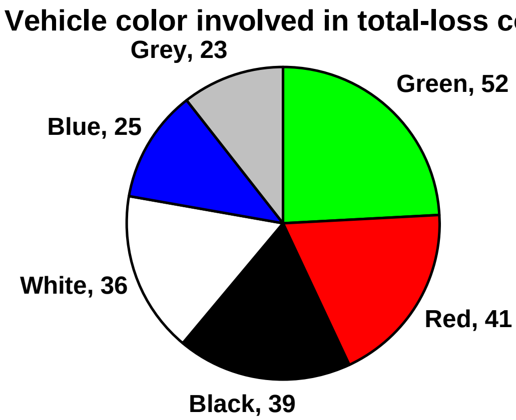
Another type of chart in frequent use is a **pie chart**, which is basically a circle with wedges cut of varying sizes marked out like slices of pie or pizza. The relative sizes of the wedges correspond to the relative frequencies of the categories.

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



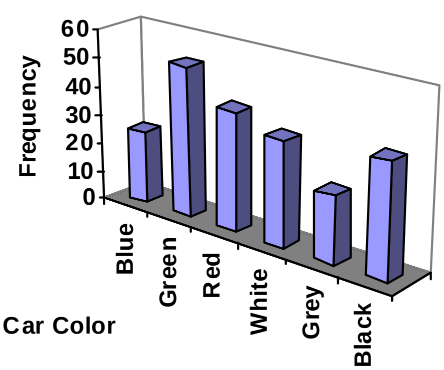
Pie charts can often benefit from including frequencies or relative frequencies (percents) in the chart next to the pie slices. Often having the category names next to the pie slices also makes the chart clearer.

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



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, or OpenOffice.org Write or Calc are able to create bar graphs, pie charts, and other graph types. There are also numerous online tools that can create graphs[2].

Don't get fancy! 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**.

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

Manager Salaries **Worker Salaries**

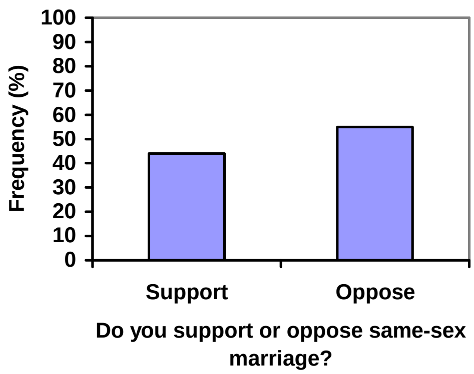
Looking at the picture, it would be reasonable to guess that the manager salaries is 4 times as large as the worker salaries – the area of the bag looks about 4 times as large. However, the manager salaries are in fact only twice as large as worker salaries, which were reflected in the picture by making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

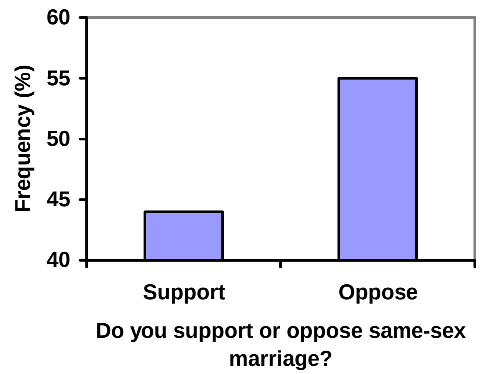
---

[2] For example: http://nces.ed.gov/nceskids/createAgraph/

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: Compare the two graphs below showing support for same-sex marriage rights from a poll taken in December 2008[3]. Once again, the difference in areas 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.





Presenting Quantitative Data Graphically

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

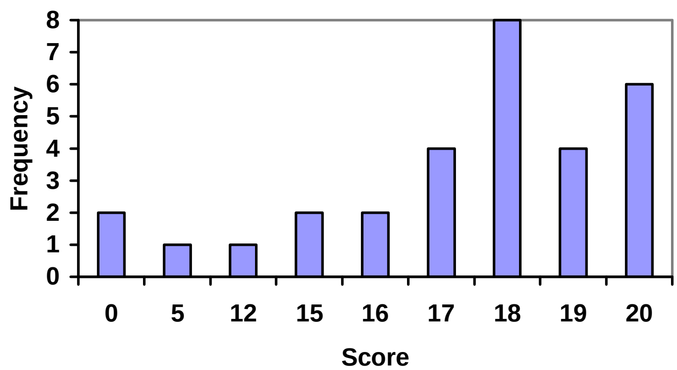
Example: A teacher records scores on a 20-point quiz for the 30 students in his class. The scores are:

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

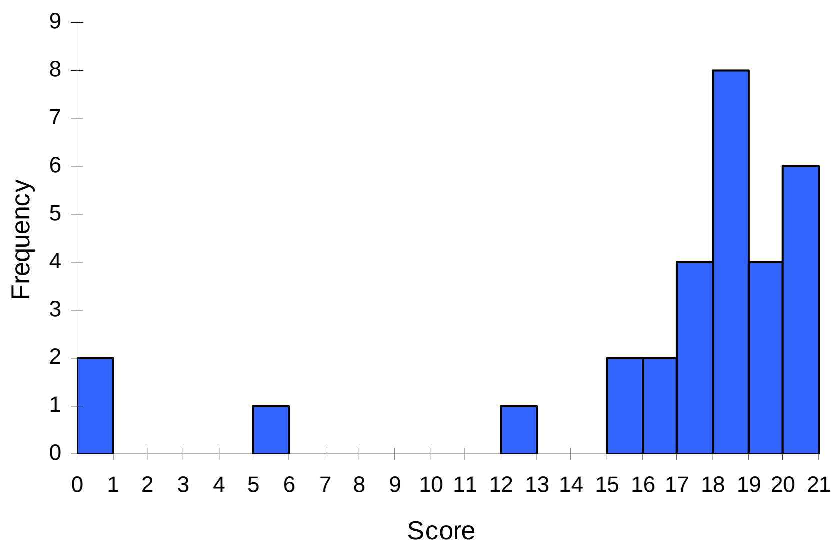
Score	Frequency
0	2
5	1
12	1
15	2
16	2
17	4
18	8
19	4
20	6

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

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

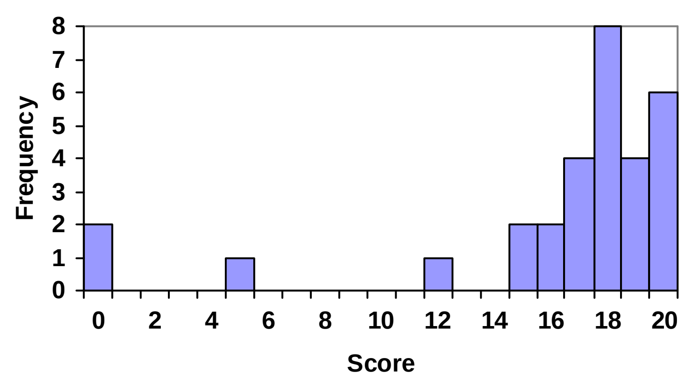


However, since the scores are numerical values, it would be more correct to treat the horizontal axis as a number line. This type of graph is called a **histogram**.



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.



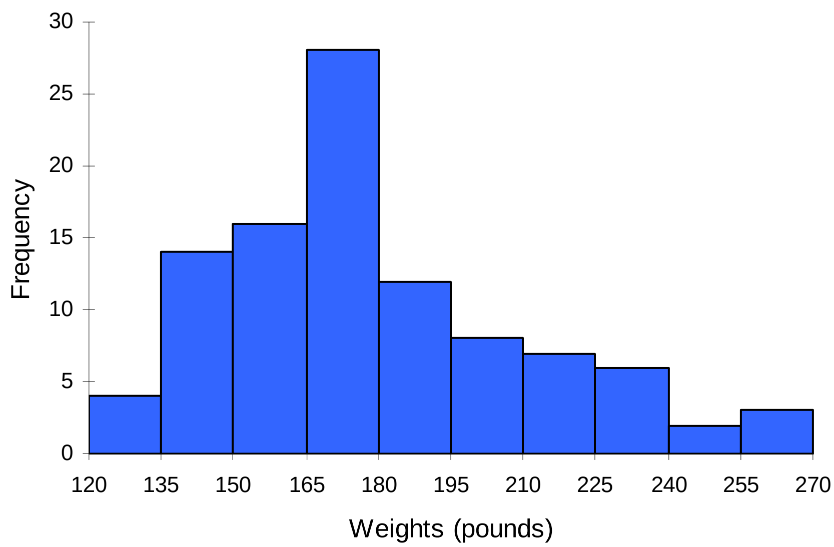
Now, suppose that we have collected weights from 100 male subjects as part of a nutrition study. Listing every possible weight with a frequency would create an exceptionally long frequency table, and probably would not reveal any patterns. For this reason, it is common with quantitative data to group data into **class intervals**. In general, we define class intervals so that:

1) 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.
2) We have somewhere between 5 and 20 classes, typically, depending upon the number of data we're working with.

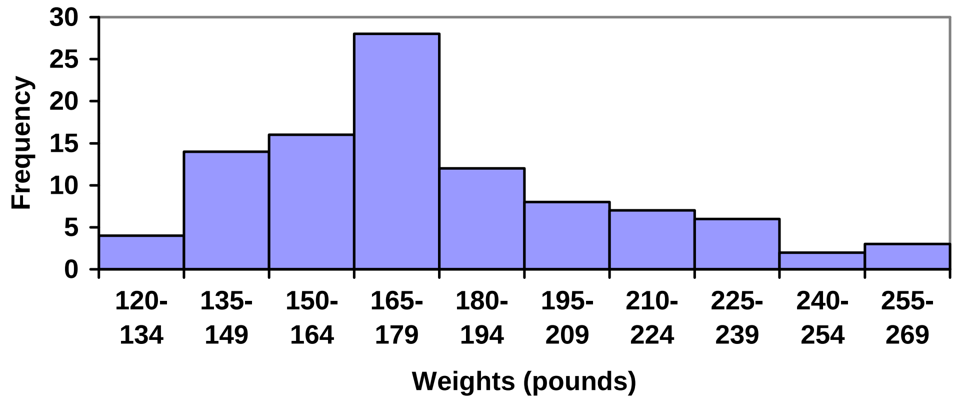
Example: For our weight data, we have values ranging from a low of 121 pounds to a high of 263 pounds, given a total span of 263-121 = 142. We could create 7 intervals with a width of around 20, 14 intervals with a width of around 10, or somewhere in between. Often time we have to experiment with a few 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

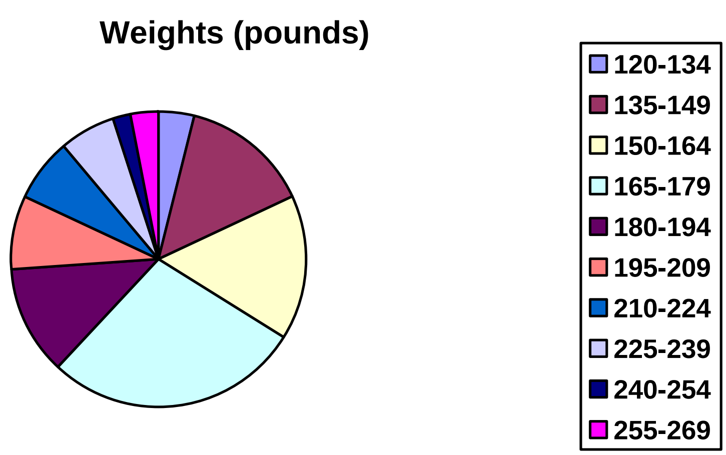
A histogram of this data would look like:



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.

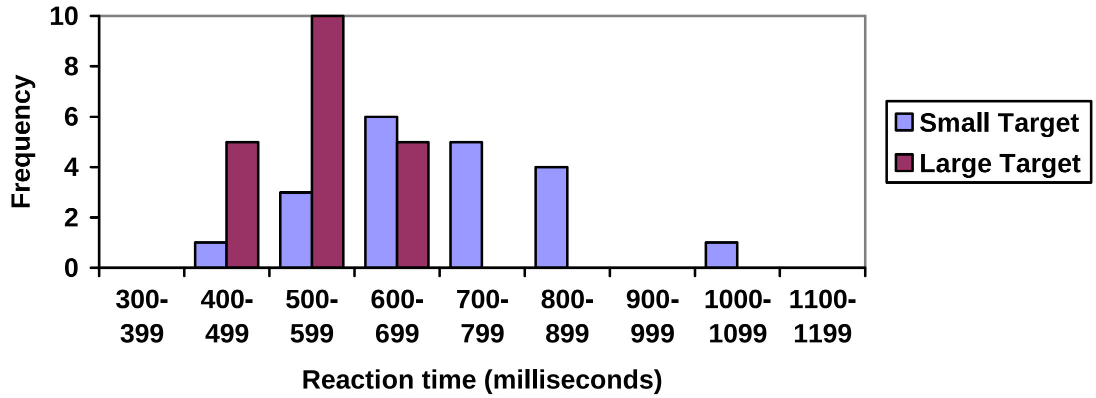


When collecting data to compare two groups, it is desirable to create a graph that compares quantities.

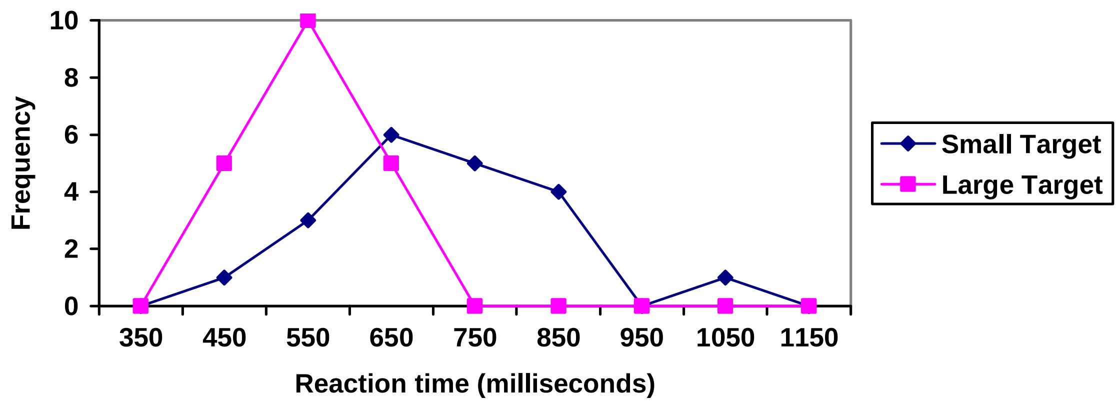
Example: The data below came from a task in which the goal is to move a computer mouse to a target on the screen as fast as possible. On 20 of the trials, the target was a small rectangle; on the other 20, the target was a large rectangle. Time to reach the target was recorded on each trial.

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.



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



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.

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 case 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": the sum of the data values divided by the number of values.

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

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

Typically we round means to one more decimal place than the original data had. In this case, that would 85.3.

Example: 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".

Example: 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:

$$\frac{\overbrace{15+\cdots+15}^{\text{6 terms}}+\overbrace{20+\cdots+20}^{\text{8 terms}}+\overbrace{25+\cdots+25}^{\text{11 terms}}+\cdots}{100}$$

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

$$\frac{15\cdot 6+20\cdot 8+25\cdot 11+30\cdot 17+35\cdot 19+40\cdot 20+45\cdot 12+50\cdot 7}{100}=\frac{3390}{100}=33.9$$

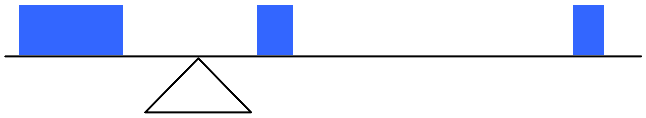
So the mean household income of our sample is 33.9 thousand dollars ($33,900).

Now, suppose a new family moves into the neighborhood that has a household income of $5 million ($5000 thousand). Adding this to our sample, our mean is now

$$\frac{15\cdot 6+20\cdot 8+25\cdot 11+30\cdot 17+35\cdot 19+40\cdot 20+45\cdot 12+50\cdot 7+5000\cdot 1}{101}=\frac{8390}{101}=83.069$$

While 83.1 thousand dollars ($83,069) is the correct mean household income, it no longer represents a "typical" value.

Imagine the data values on a see-saw or balance scale. The mean is the value that keeps the data in balance, like in the picture below.



If we graph our household data, the $5 million data value is so far out to the right that the mean has to adjust up to keep things in balance

For this reason, when working with data that have **outliers** – values far outside the primary grouping – it is common to use a different measure of center, the **median**. Just like the median of a freeway is the strip that runs down the middle, the median of a set of numbers is the value in the middle.

To find the median, begin by listing the data in order from smallest to largest. If the number of data values, N, is odd, then the median is the middle data value. This value can be found by rounding $N/2$ up to the next whole number. If the number of data values is even, there is no one middle value, so we find the mean of the two middle values (values $N/2$ and $N/2 + 1$)

Example: 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.

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: 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. Finding the mean: (6+7)/2 = 6.5. So the median quiz score was 6.5.

Example: 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:

Values 1 to 6 are $15
Values 7 to (6+8)=14 are $20
Values 15 to (14+11)=25 are $25
Values 26 to (25+17)=42 are $30
Values 43 to (42+19)=61 are $35

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

Note that if we add in the new neighbor with a $5 million household income, it will not affect the median in this case. The median is not swayed as much by outliers.

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

Measures of Variation

Consider these three sets of quiz scores:

Section A: 5 5 5 5 5 5 5 5 5 5

Section B: 0 0 0 0 0 10 10 10 10 10

Section C: 4 4 4 5 5 5 5 6 6 6

All three of these sets of data have a mean of 5 and median of 5, yet the sets of scores are clearly quite different. In section A, everyone had the same score; in section B half the class got no points and the other half got a perfect score, assuming this was a 10-point quiz. Section C was not as consistent as section A, but not as widely varied as section B.

In addition to the mean and median, which are measures of the "typical" or "middle" value, we also need a measure of how "spread out" or varied each data set is.

There are several ways to measure this "spread" of the data. The first is the simplest and is called the **range**. The range is simply the difference between the maximum value and the minimum value.

Example: Using the quiz scores from above,

For section A, the range is 0 since both maximum and minimum are 5 and 5 – 5 = 0
For section B, the range is 10 since 10 – 0 = 10
For section C, the range is 2 since 6 – 4 = 2

So far range seems to be revealing how spread out the data is. However, if we add a fourth section:

Section D: 0 5 5 5 5 5 5 5 5 10

This section also has a mean and median of 5. The range is 10, yet this data set is quite different than Section B. To better illuminate the differences, we'll have to turn to more sophisticated measures of variation.

The **standard deviation** is a measure of variation based on measuring how far each data value deviates, or is different, from the mean. A few important characteristics:

- Standard deviation is always positive. Standard deviation will be zero if all the data values are equal, and will get larger as the data spreads out.
- Standard deviation has the same units as the original data.
- Standard deviation, like the mean, can be highly influenced by outliers.

Using the data from section D, we could compute for each data value the difference between the data value and the mean:

data value	deviation: data value - mean
0	0-5 = -5
5	5-5 = 0
5	5-5 = 0
5	5-5 = 0
5	5-5 = 0
5	5-5 = 0
5	5-5 = 0
5	5-5 = 0
5	5-5 = 0
10	10-5 = 5

We would like to get an idea of the "average" deviation from the mean, but if we find the average of the values in the second column the negative and positive values cancel each other out (this will always happen), so to prevent this we square every value in the second column:

data value	deviation: data value - mean	deviation squared
0	0-5 = -5	$(-5)^2 = 25$
5	5-5 = 0	$0^2 = 0$
5	5-5 = 0	$0^2 = 0$
5	5-5 = 0	$0^2 = 0$
5	5-5 = 0	$0^2 = 0$
5	5-5 = 0	$0^2 = 0$
5	5-5 = 0	$0^2 = 0$
5	5-5 = 0	$0^2 = 0$
5	5-5 = 0	$0^2 = 0$
10	10-5 = 5	$(5)^2 = 25$

We then add the squared deviations up to get 25 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 25 = 50. Ordinarily we would then divide by the number of scores, *n*, (in this case, 10) to find the mean of the deviations. But we only do this if the data set represents a population; if the data set represents a sample (as it almost always does), we instead divide by *n* - 1 (in this case, 10 - 1 = 9)[4]

So in our example, we would have 50/10 = 5 if section D represents a population and 50/9 = about 5.56 if section D represents a sample. These values (5 and 5.56) are called, respectively, the **population variance** and the **sample variance** for section D.

Variance can be a useful statistical concept, but note that the units of variance in this instance would be points-squared since we squared all of the deviations. What are points-squared? Good question. We would rather deal with the units we started with (points in this case), so to convert back we take the square root and get:

$$\text{population standard deviation} = \sqrt{\frac{50}{10}} = \sqrt{5} \approx 2.2$$

or

$$\text{sample standard deviation} = \sqrt{\frac{50}{9}} \approx 2.4$$

If we are unsure whether the data set is a sample
or a population, we will usually assume it is a sample, and we will round answers to one more decimal place than the original data, as we have done above.

[4] The reason we do this is highly technical, but we can see how it might be useful by considering the case of a small sample from a population that contains an outlier, which would increase the average deviation: the outlier very likely won't be included in the sample, so the mean deviation of the sample would underestimate the mean deviation of the population; thus we divide by a slightly smaller number to get a slightly bigger average deviation.

To compute standard deviation:

1) Find the deviation of each data from the mean. In other words, subtract the mean from the data value.
2) Square each deviation.
3) Add the squared deviations.
4) Divide by *n*, the number of data values, if the data represents a whole population; divide by *n*-1 if the data is from a sample.
5) Compute the square root of the result.

Example: Computing the standard deviation for Section B above, we first calculate that the mean is 5. Using a table can help keep track of your computations for the standard deviation:

data value	deviation: data value - mean	deviation squared
0	0-5 = -5	$(-5)^2 = 25$
0	0-5 = -5	$(-5)^2 = 25$
0	0-5 = -5	$(-5)^2 = 25$
0	0-5 = -5	$(-5)^2 = 25$
0	0-5 = -5	$(-5)^2 = 25$
10	10-5 = 5	$(5)^2 = 25$
10	10-5 = 5	$(5)^2 = 25$
10	10-5 = 5	$(5)^2 = 25$
10	10-5 = 5	$(5)^2 = 25$
10	10-5 = 5	$(5)^2 = 25$

Assuming this data represents a population, we will add the squared deviations, divide by 10, the number of data values, and compute the square root:

$$\sqrt{\frac{25+25+25+25+25+25+25+25+25+25}{10}} = \sqrt{\frac{250}{10}} = 5$$

Notice that the standard deviation of this data set is much larger than that of section D since the data in this set is more spread out.

For comparison, the standard deviations of all four sections are:

Section A: 5 5 5 5 5 5 5 5 5 5	Standard deviation: 0
Section B: 0 0 0 0 0 10 10 10 10 10	Standard deviation: 5
Section C: 4 4 4 5 5 5 5 6 6 6	Standard deviation: 0.8
Section D: 0 5 5 5 5 5 5 5 5 10	Standard deviation: 2.2

Where standard deviation is based on the mean, **quartiles** are based on the median. The median is the value so that 50% of the data values are below it. The first quartile (Q_1) is the value so that 25% of the data values are below it; the third quartile (Q_3) is the value so that 75% of the data values are below it. You may have guessed that the second quartile is the same as the median.

While quartiles are not a 1-number summary of variation like standard deviation, the quartiles are used with the median, minimum, and maximum values to form a **5 number summary** of the data looking like: Minimum, Q_1, Median, Q_3, Maximum

To find the first quartile, we need to find the data value so that 25% of the data is below it. If *n* is the number of data values, we compute a locator by finding 25% of *n*. If this locator is a decimal value, we round up, and find the data value in that position. If the locator is a whole number, we find the mean of the data value in that position and the next data value.

To find the first quartile, Q_1:

Compute the locator: $L = 0.25n$

If L is a decimal value:

Round up to $L+$

Use the data value in the $L+^{th}$ position

If L is a whole number:

Find the mean of the data values in the L^{th} and $L+1^{th}$ positions.

To find the third quartile, Q_3:

Use the same procedure, with locator: $L = 0.75n$

Examples should help make this clearer.

Example: Suppose we have measured 9 females and their heights (in inches), sorted from smallest to largest are:

59 60 62 64 66 67 69 70 72

To find the first quartile we first compute the locator: 25% of 9 is 0.25(9) = 2.25. Since this value is not a whole number, we round up to 3. The first quartile will be the third data value: 62 inches.

Example: Suppose we had measured 8 females and their heights (in inches), sorted from smallest to largest are:

59 60 62 64 66 67 69 70

To find the first quartile we first compute the locator: 25% of 8 is 0.25(8) = 2. Since this value *is* a whole number, we will find the mean of the 2nd and 3rd data values: (60+62)/2 = 61, so the first quartile is 61 inches.

The third quartile is computed similarly, using 75% instead of 25%. Note that the median could be computed the same way, using 50%.

Example: Using our 9 female sample of heights, to find the third quartile, we first compute the locator: 75% of 9 is 0.75(9) = 6.75. Since this value is not a whole number, we round up to 7. The third quartile will be the seventh data value: 69 inches.

Example: Using our 8 female sample of heights, the locator would be: 75% of 8 is 0.75(8) = 6. Since this value *is* a whole number, we will find the mean of the 6th and 7th data values: (67+69)/2 = 68, so the third quartile is 68 inches.

The 5-number summary combines the first and third quartile with the minimum, median, and maximum values.

Example: For the 9 female sample, the median is 66, the minimum is 59, and the maximum is 72. The 5 number summary is: 59, 62, 66, 69, 72.

For the 8 female sample, the median is 65, the minimum is 59, and the maximum is 70, so the 5 number summary would be: 59, 61, 65, 68, 70.

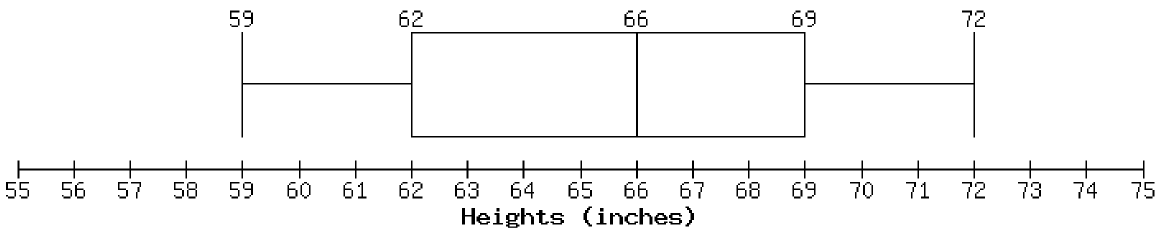
Example: Returning to our quiz score data. In each case, the first quartile locator is 0.25(10) = 2.5, so the first quartile will be the 3rd data value, and the third quartile will be the 8th data value. Creating the 5-number summaries:

Section and data	5-number summary
Section A: 5 5 5 5 5 5 5 5 5 5	5, 5, 5, 5, 5
Section B: 0 0 0 0 0 10 10 10 10 10	0, 0, 5, 10, 10
Section C: 4 4 4 5 5 5 5 6 6 6	4, 4, 5, 6, 6
Section D: 0 5 5 5 5 5 5 5 5 10	0, 5, 5, 5, 10

Note that the 5 number summary divides the data into four intervals, each of which will contain about 25% of the data.

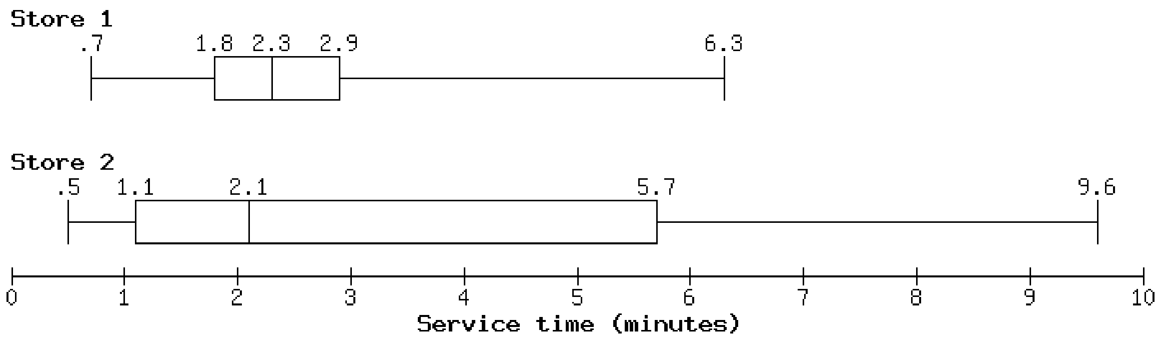
For visualizing data, there is a graphical representation of a 5-number summary called a **box plot**, or box and whisker graph. In a box plot, a number line is first drawn. A box is drawn from the first quartile to the third quartile, and a line is drawn through the box at the median. "Whiskers" are extended out to the minimum and maximum values.

Example: The box plot below is based on the 9 female height data with 5 number summary: 59, 62, 66, 69, 72.



Box plots are particularly useful for comparing data from two populations.

Example: The box plot of service times for two fast-food restaurants is shown below.



While store 2 had shorter median service time (2.1 minutes vs. 2.3 minutes), store 2 is less consistent, with wider spread of the data. At store 1, 75% of customers were served within 2.9 minutes, while at store 2, 75% of customers were served within 5.7 minutes. Which store should you go to in a hurry? That depends upon your opinions about luck – 25% of customers at store 2 had to wait between 5.7 and 9.6 minutes.

Exercises

Skills

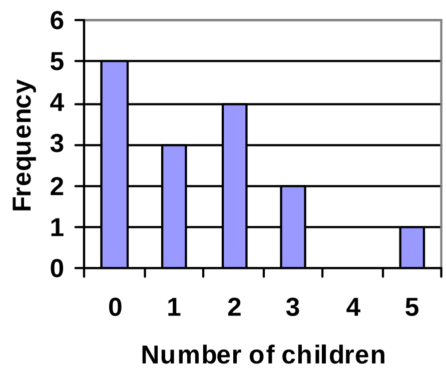
1. The table below shows scores on a Math test.
   a. Complete the frequency table for the Math test scores
   b. Construct a histogram of the data
   c. Construct a pie chart of the data

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

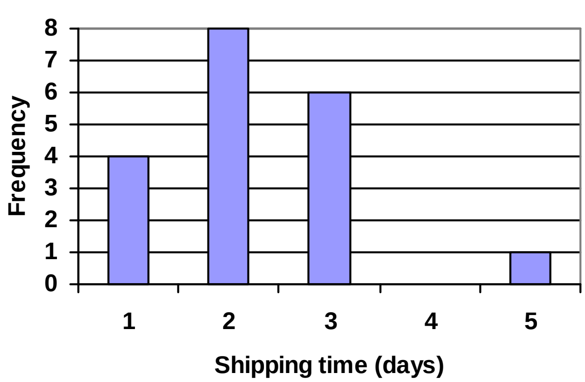
2. A group of adults where asked how many cars they had in their household
   a. Complete the frequency table for the car number data
   b. Construct a histogram of the data
   c. Construct a pie chart of the data

1	4	2	2	1	2	3	3	1	4	2	2
1	2	1	3	2	2	1	2	1	1	1	2

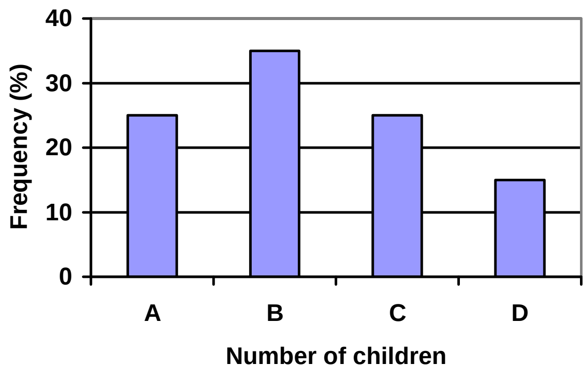
3. A group of adults were asked how many children they have in their families. The bar graph to the right shows the number of adults who indicated each number of children.
   a. How many adults where questioned?
   b. What percentage of the adults questioned had 0 children?



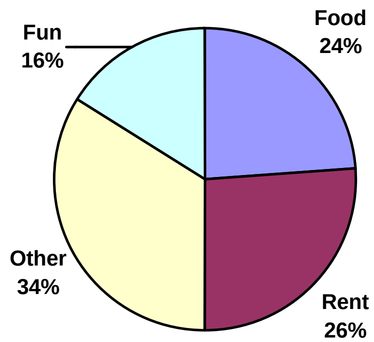
4. Jasmine was interested in how many days it would take an order from Netflix to arrive at her door. The graph below shows the data she collected.
   a. How many movies did she order?
   b. What percentage of the movies arrived in one day?



5. 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?



6. 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?



7. A group of diners were asked how much they would pay for a meal. Their responses were: $7.50, $8.25, $9.00, $8.00, $7.25, $7.50, $8.00, $7.00.
   a. Find the mean
   b. Find the median
   c. Write the 5-number summary for this data

8. You recorded the time in seconds it took for 8 participants to solve a puzzle. The times were: 15.2, 18.8, 19.3, 19.7, 20.2, 21.8, 22.1, 29.4.
   a. Find the mean
   b. Find the median
   c. Write the 5-number summary for this data

9. Refer back to the histogram from question #3.
   a. Compute the mean number of children for the group surveyed
   b. Compute the median number of children for the group surveyed
   c. Write the 5-number summary for this data.
   d. Create box plot.

10. Refer back to the histogram from question #4.
    a. Compute the mean number of shipping days
    b. Compute the median number of shipping days
    c. Write the 5-number summary for this data.
    d. Create box plot.

Concepts

11. The box plot below shows salaries for Actuaries and CPAs. Kendra makes the median salary for an Actuary. Kelsey makes the first quartile salary for a CPA. Who makes more money? How much more?

12. Referring to the boxplot above, what percentage of actuaries makes more than the median salary of a CPA?

Exploration

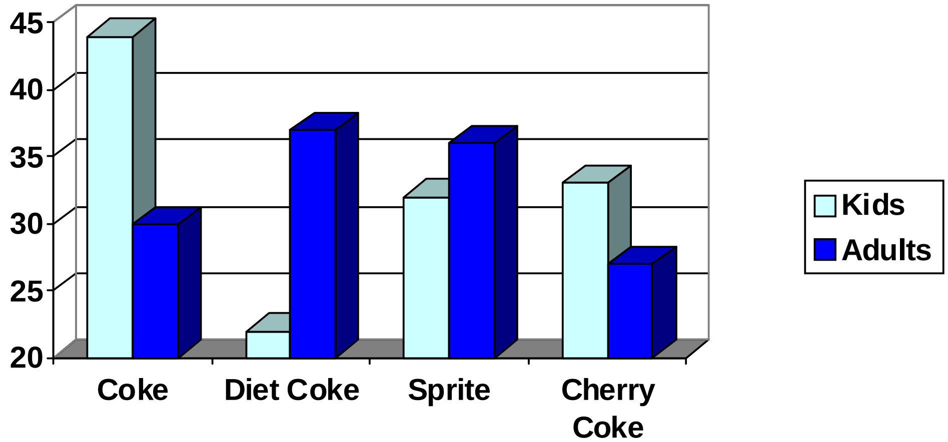
13. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 AIDS patients from the start of treatment until their deaths. The following data (in months) are collected.

    Researcher 1: 3; 4; 11; 15; 16; 17; 22; 44; 37; 16; 14; 24; 25; 15; 26; 27; 33; 29; 35; 44; 13; 21; 22; 10; 12; 8; 40; 32; 26; 27; 31; 34; 29; 17; 8; 24; 18; 47; 33; 34

    Researcher 2: 3; 14; 11; 5; 16; 17; 28; 41; 31; 18; 14; 14; 26; 25; 21; 22; 31; 2; 35; 44; 23; 21; 21; 16; 12; 18; 41; 22; 16; 25; 33; 34; 29; 13; 18; 24; 23; 42; 33; 29

    a. Create comparative histograms of the data
    b. Create comparative boxplots of the data

14. A graph appears below showing the number of adults and children who prefer each type of soda. There were 130 adults and kids surveyed. Discuss some ways in which the graph below could be improved



15. Make up three data sets with 5 numbers each that have:
    a. the same mean but different standard deviations.
    b. the same mean but different medians.
    c. the same median but different means.

16. A sample of 30 distance scores measured in yards has a mean of 7, a variance of 16, and a standard deviation of 4.
    a. You want to convert all your distances from yards to feet, so you multiply each score in the sample by 3. What are the new mean, median, variance, and standard deviation?
    b. You then decide that you only want to look at the distance past a certain point. Thus, after multiplying the original scores by 3, you decide to subtract 4 feet from each of the scores. Now what are the new mean, median, variance, and standard deviation?