Module 4 Working with Statistics

Section 4.1 Data Types

Looking Back 4.1

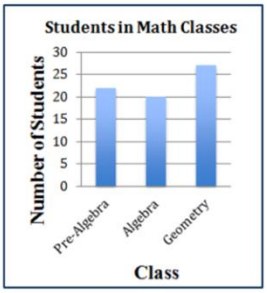
We have previously investigated probable events and expected outcomes. We used Pascal's Triangle to solve real-world problems. Blaise Pascal knew the patterns in the triangle existed because God is a God of order and predictability. In this module, we will investigate and analyze data. Again, we will find order and predictability.

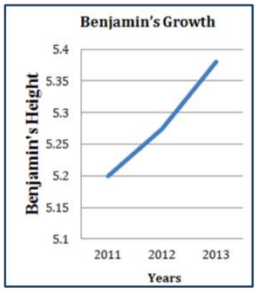
Looking Ahead 4.1

We will first look into discrete data and continuous data. Discrete data is a count, like the number of students in a classroom. Discrete data may be represented by a bar graph. These data figures are certain values.

Continuous data is a measure, like the growth of a human. Continuous data may be represented by a line graph. These data figures are within a range of values.

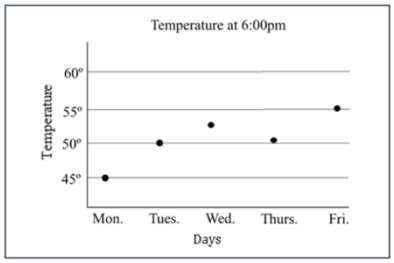
Example 1:	Which graph below represents discrete data and which represents continuous data?



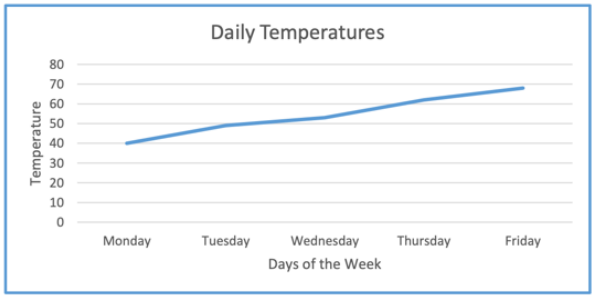


Example 2:	Give three examples of discrete data and three examples of continuous data.

Example 3:	Below is a graph of temperatures at 6 P.M. each day. Is this discrete data or continuous data?



Example 4:	Below is a graph of temperatures throughout the week. Is this discrete data or continuous data?



We will also investigate categorical data and numerical data in this module. Categorical data can be sorted by categories (or groups). Bar graphs and/or pie charts are typically used to represent categorical data.

Numerical data includes values that are observed and can be measured. Scatterplots and/or line graphs are typically used to represent numerical data.

Example 5:	Give three examples of categorical data and three examples of numerical data.

Example 6:	There are two graphs in Example 1. Is the data in the graphs categorical or numerical?

Section 4.2 Univariate Data

Looking Back 4.2

Just as there is discrete data and continuous data, and categorical data and numerical data, there is univariate data and bivariate data. The prefix "uni" means *one*. The prefix "bi" means *two*. These prefixes come from Greek and Latin roots.

Differences between univariate data and bivariate data are listed below:

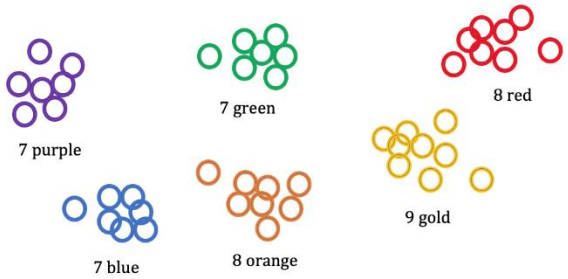
Univariate Data	**Bivariate Data**
• One variable • Does not deal with causes and relationships • Purpose is to describe • Measures of central tendency are analyzed in terms of the following: mean; median; mode; range of data Graphs used: • Bar graphs • Histograms • Pie charts • Line plots or Dot Plots • Box-and-Whisker plots	• Two variables • Does deal with causes and relationships • Purpose is to explain • Dependent and independent variables are analyzed in terms of the following: correlations/comparisons; causes/relationships; explanations Graphs used: • Scatterplots (demonstrate the relationship between two variables) • Line Graphs made from connected scatterplots

An example of univariate data would be the number of students in 9th grade taking Algebra that are females. An example of bivariate data would be the relationship between freshman females in Algebra and their scores on college entrance examinations.

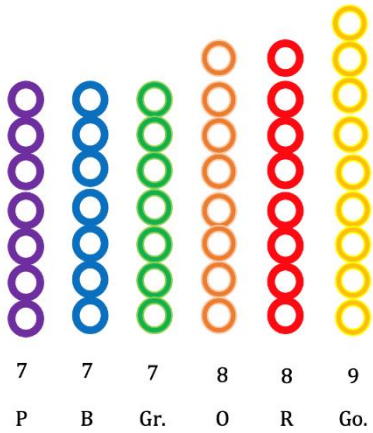
Looking Ahead 4.2

We have experience with measures of central tendency that are used to show variability; these include mean, median, and mode. These are also called measures of center. The median is in the middle of the data when ordered least to greatest or greatest to least and the mode is what occurs most often; we will go over the mean later in this section.

Example 1: If we pour out $\frac{1}{4}$ cup of Fruity Loops on a table, and then group them by color, what is the mode?



Example 2: What is the mode of the group shown below?



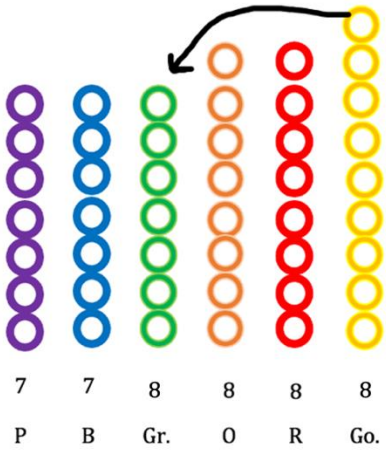
Range is also a measure of variability. It is the difference between the largest and smallest numbers in a set of data.

What is the range of the Fruity Loops? What can you say about the spread of the Fruity Loops?

Example 3: The median is the middle of numerical data after the data is put in numerical order. What is the median of the Fruity Loop data from Example 2?

The mean is the average or the number that gives an equal share in the distribution. To find average, we want to make each column about the same height. To make each column about the same height in the Fruity Loop example, we can move one gold Fruity Loop over to the green column. Now we can see that the average height of the columns is between 7 and 8, but closer to 8; let us suppose it is 7.8. Notice we are looking at average height, which is the width of the shape the data makes, which is a rectangle. We know length multiplied by width gives us area and area divided by length gives us width. The area is all of the Fruity Loops and the length is the number of columns (number of colors). This is a geometric representation of an algebraic problem.

Example 4:	Find the average of the Fruity Loop data.



Section 4.3 Mean Distribution

Looking Back 4.3

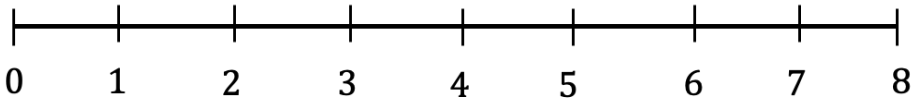
There is another way to look at average. We have already learned that mean is equal distribution (fair share). Think about a balance scale; the scale is balanced when all of the data is spread out evenly on the scale.

Looking Ahead 4.3

The table below shows boxes of crackers and grams of fat per box. We will be using it in an experiment later in this module. For now, let us look specifically at measures of central tendency.

Cracker	Fat (g.)
Zips	7
Snoodles	7
Quigley's	3
Munchables	7
Crunchies	2
Snips	7
Cheezies	2
Holts	2
PB&J	7
Crackles	7
Kringles	3
Ripp's	7
Zoodles	7

Example 1: Using the table above, make a line plot to represent the distribution of the grams of fat in crackers. Put an "X" above a number each time it appears on the table.



a) Find the range of the grams of fat per box for all the boxes of crackers.

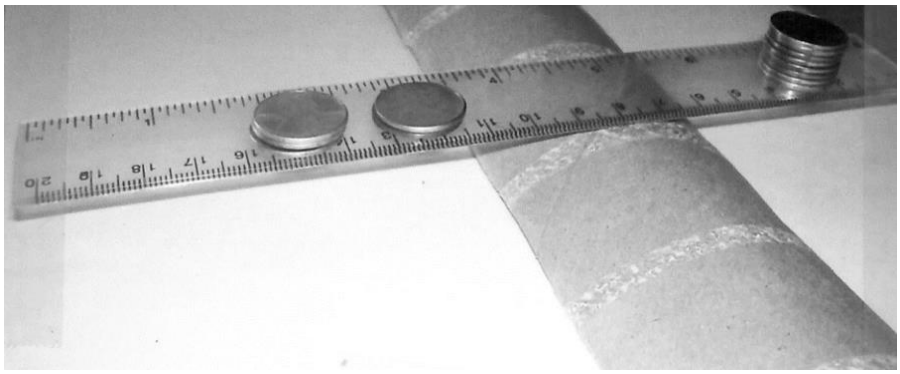
b) Find the median of the grams of fat per box for all the boxes of crackers.

c) Find the mode of the grams of fat per box for all the boxes of crackers.

d) Find the mean of the grams of fat per box for all the boxes of crackers.

We are now going to set up a geometric representation of this using an 8-inch ruler, pennies, and a paper towel tube. The paper towel tube will be cut in half lengthwise to act as the fulcrum. The ruler is placed on the paper towel tube perpendicular to it. The 4-inch mark on the ruler will be placed in the middle of the paper towel tube to keep it balanced. There is no resistance on either side.

Given the numbers on the ruler represent the grams of fat and pennies represent the boxes of crackers, three pennies can be placed on the 2-inch mark, two pennies can be placed on the 3-inch mark, and eight pennies can be placed on the 7-inch mark. Now the ruler should be unbalanced. To balance it again, the fulcrum will be moved to the mean, which is at the $5\frac{3}{13}$-inch mark. Now we have a geometric representation of average (mean).



Section 4.4 Bar Graphs or Histograms

Looking Back 4.4

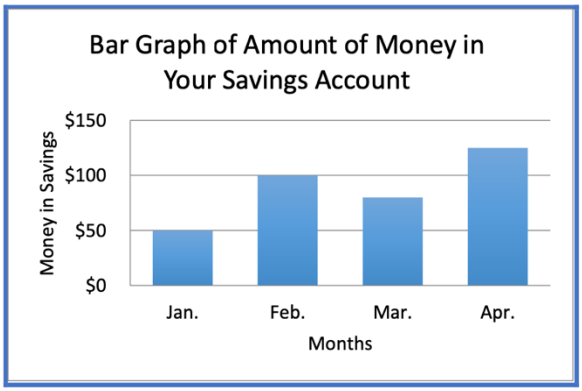
When we look at raw data, we see individual values. However, when we look at summarized data, it is presented in a way that allows us to analyze it.

Remember, the Fruity Loops on the bar graph in Section 4.2 Univariate Data were arranged (categorized) by color. The color of the Fruity Loops did not correspond with any numerical value other than how many of each color there were in $\frac{1}{4}$ cup of Fruity Loops. The different colors were ordered from least to greatest by number to make it easier to identify the median, mode, range, and mean of the group.

Looking Ahead 4.4

Bar graphs divide data into categories and show the total amount in each category. Bar graphs have a visual effect. The bars on the graph generally have spaces between them, but these are not necessary. The width of the bar does not matter because no interval is associated with it. However, the bars must all be the same width, whatever that width may be. Order does not always matter with bar graphs.

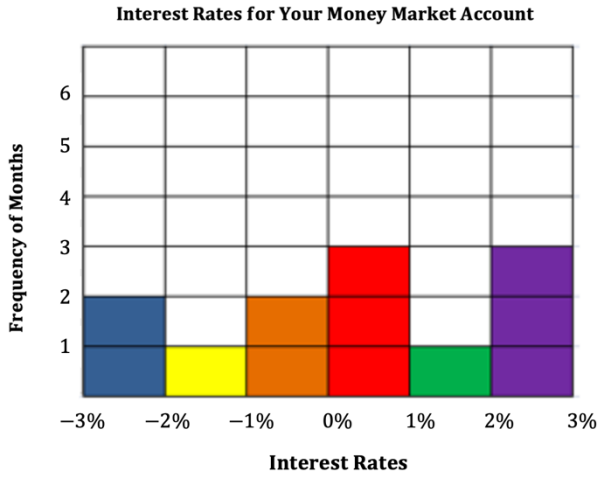
Example 1: What are the categories for the graph shown below? What can we say about the totals in February and April? What can we not say about the totals in February and April?



A histogram is a special bar graph in which the bars are next to one another because the data is frequencies of different numerical values within a group. Frequencies refer to how often something occurs. Order matters in these cases and the scale is within a specific interval.

Example 2: Below is an example of a histogram. There is no space between the bars.

In the histogram, what do you notice about the intervals? What is the scale of each interval? List the intervals below the graph. How many frequencies are there according to the data?



Section 4.5 Normal Distribution

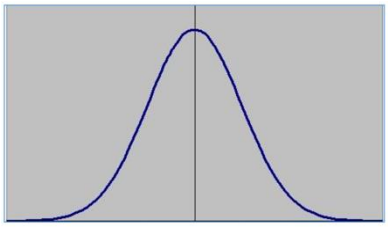
Looking Back 4.5

In the previous few sections, we learned data is "spread out" or "distributed" in different ways. We call this variability. Now, we want to talk about normal distribution and the bell-shaped curve.

In the bell-shaped curve, data is "spread out" or "distributed" rather evenly. The distribution takes a shape that looks like the Liberty Bell, which is where it gets the name *bell-shaped curve*. In this normal distribution, the data is clustered in the center and spread out on either end. It increases from the left, reaches a peak, and then decreases to the right.

Looking Ahead 4.5

Example 1: Using the bell-shaped curve (shown below), describe the distribution of data and the measures of central tendency.

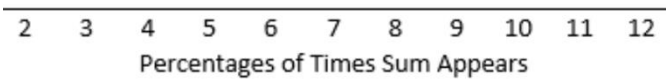


Example 2: If you roll two dice and add the numbers you land on, the chance of rolling each outcome or sum will fall close to the percentages below. Make a line plot for the data. Round to the nearest one's place.

2: 2.78% 3: 5.56% 4: 8.33% 5: 11.11%

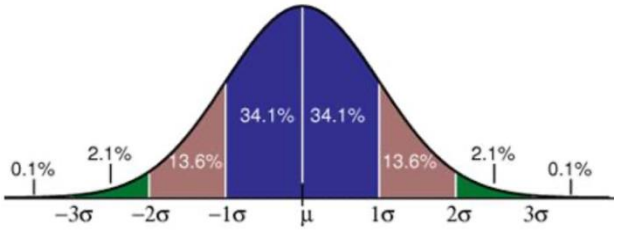
6: 13.89% 7: 16.67% 8: 13.89% 9: 11.11%

10: 8.33%. 11: 5.56% 12: 2.27%



What do you notice about the data and line plot?

Now, let us look at standard deviation for the normal distribution below.



The Greek symbol μ (pronounced: "Mu") is the symbol that represents average. The Greek symbol σ (Sigma) is the symbol that represents standard deviation; standard deviation shows us how much variance or spread there is from the mean (average), which is the expected value given everything is equally distributed.

Therefore, 1σ is 1 standard deviation, which means 68.2% of the population sampled in the diagram above falls within 1 standard deviation of the mean.

The lower the standard deviation is, the closer the data is to the mean. The higher the standard deviation is, the farther the data is from the mean.

Example 3: What percent of the data from the diagram above falls 2σ from the mean? What percent of this data falls 3σ from the mean?

Section 4.6 Pie Charts

Looking Back 4.6

Line plots are good to use when the data includes fewer than twenty-five numbers. Just like a line plot, a bar graph is a quick display of the maximum, minimum, and range of data. Not much additional information can be gathered from a bar graph or line plot. Pie charts are simple as well.

Pie charts contain less data, but what is important can be seen very easily. We usually focus our energy on the biggest piece of the pie. If we want to graph change over time, such as in unemployment rates, then line graphs are preferrable to bar graphs and pie charts. However, if we want to show unemployment rates in a group of different cities or among different age groups, bar graphs and pie charts are preferrable to line plots.

Looking Ahead 4.6

Pie charts are especially useful to see the percentages of a whole. To make a pie chart, we must perform the following steps:

1. Find a total number of data items.
2. Divide the number of frequencies in each category by the total number of frequencies. That number can be converted to a percentage by multiplying it by 100.
3. Multiply the fraction in Step 2 by 360° (the degrees in a circle) to find the number of degrees that represent each category on the circle.
4. Use your protractor to draw a circle.
5. Mark off the degrees on your circle from Step 3.
6. Shade each piece (percentage) of the pie (circle) with a different color. Label each piece with the percent from Step 2 and title of the category.

Example 1: Use the steps above to create a pie chart for the data below.

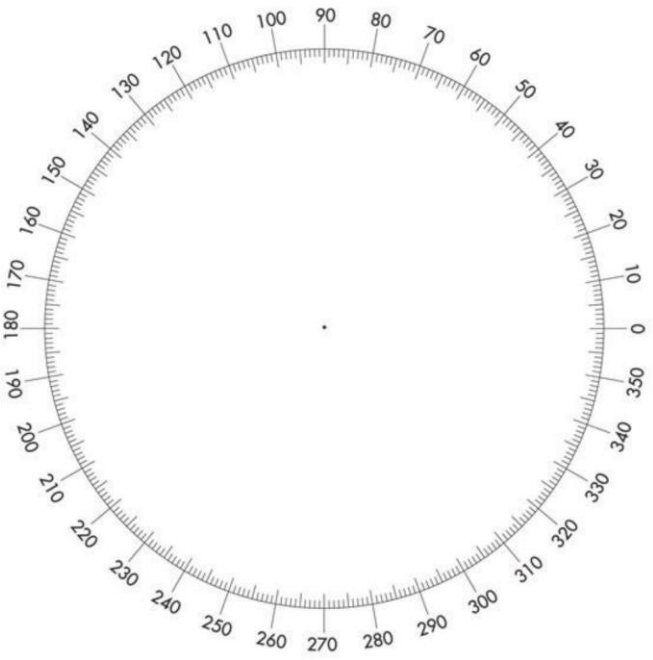
Number of Students in Mountain Home School District by Grade Level	
Homeschool Grade	**Students**
9^{th} grade	30
10^{th} grade	26
11^{th} grade	33
12^{th} grade	42

1. Find the total number of data items.

2. Divide the number of frequencies in each category by the total number of frequencies. That number can be converted to a percentage by multiplying it by 100.

3. Multiply the fraction in Step 2 by 360° (the degrees in a circle) to find the number of degrees that represent each category on the circle.

4. Use your protractor to draw a circle.



5. Mark off the degrees on your circle from Step 3.

6. Shade each piece (percentage) of the pie (circle) with a different color. Label each piece with the percent from Step 2 and title of the category.

Line Graph

- Tracks smaller changes over time

Bar Graph

- Tracks larger changes over time

Line Graph/Bar Graph

- Shows comparisons or trends for several groups over the same time period(s) using different colors to represent each group

Pie Chart

- Does not show changes over time
- Used to compare the parts to the whole

Section 4.7 Stem-and-Leaf Plots

Looking Back 4.7

In the previous section, we looked at grades on a pre-algebra test. The histogram for the data of these grades showed us an example of the normal bell curve, but the frequencies got lost in the outer regions.

A Stem-and-Leaf plot is another way to display a frequency distribution. While the histogram focuses on the intervals, the Stem-and-Leaf plot focuses on the specific numerical values of the distribution.

Looking Ahead 4.7

There are two parts to a Stem-and-Leaf plot. For two-digit numbers, the STEM part is the tens place, and the LEAF part is the ones place. The distribution is easier to see because the values are listed out individually.

Example 1: Let us look at a completed Stem-and-Leaf plot of pre-algebra test scores before we learn how to make them. List all the scores on this pre-algebra test.

Pre-Algebra Test Scores

STEM	LEAF
6	1 2 4
7	1 2 4
8	3 3 4 4
9	2 7 9

Example 2: Make a Stem-and-Leaf plot given the information and data below.

Do you remember the gopher game from the Practice Problems in which gophers' heads popped up out of holes and Isaiah tried to tap them with a rubber mallet as fast as he could? Below are his reaction times in seconds. What would be the key for this data?

5.2	5.8	5.9
6.4	4.8	5.2
4.8	4.9	6.2

Example 3: Make a Stem-and-Leaf plot given the information and data below.

There are two numbers that have three decimal places after the decimal point. Therefore, you will have to make all the leaves of the Stem-and-Leaf plot three decimal places by adding zeroes.

2.6	2.7	2.88	2.9	3.085
3.2	3.2	3.3	3.3	3.334
3.4	3.56	3.6	3.75	3.8

Section 4.8 Box-and-Whisker Plots

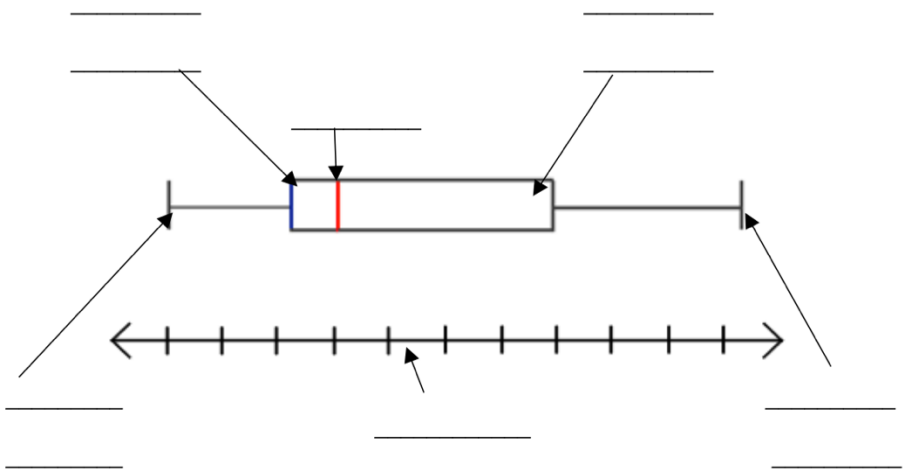
Looking Back 4.8

Stem-and-Leaf plots are a visual way of looking at specific data as a frequency distribution rather than as grouped by intervals.

In this section, we will learn about Box-and-Whisker plots. A Box-and-Whisker plot gets its name from its parts: a "box" and its "whiskers;" a rectangle (box) being in the center above a scale with lines on both ends (whiskers). Box-and-Whisker plots give us a visual picture of the center, spread, and overall range of the distribution of data. These plots are useful for handling many data values, unlike Stem-and-Leaf plots. With Box-and-Whisker plots, we can explore data and draw informal conclusions.

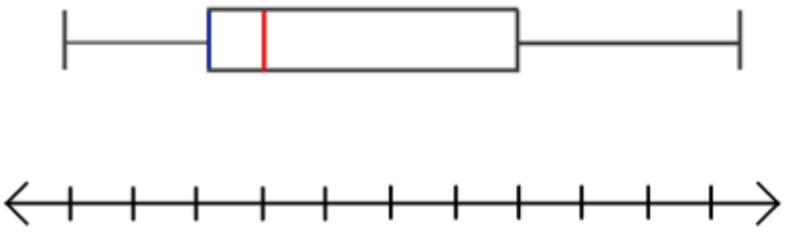
Looking Ahead 4.8

Example 1:	Identify the parts of the Box-and-Whisker plot below.



The data is broken into four groups called Quartiles: the lower quartile, Q_1 (read: Q sub one), is the median of the lower half of the data. The upper quartile, Q_3 (read: Q sub three), is the median of the upper half of the data. The second quartile, Q_2 (read: Q sub two), divides the data into two parts, the lower half and the upper half.

Example 2:	Label the quartiles of the Box-and-Whisker plot below.



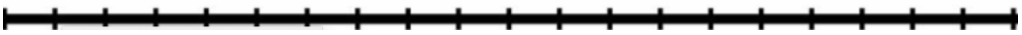
The maximum or minimum can be an outlier. An outlier is a(n) _____ score that can really skew the measures of central tendency and range. It is represented an X on the Box-and-Whisker plot.

Example 3:	Find the median of fourteen data points given the middle two are 12 and 13.

Example 4:	Follow the steps to make a Box-and-Whisker plot using the data below.

The following scores are those for the Gopher Game from a sample of seventeen children at a birthday party listed in order from least to greatest. Each child used their dominant hand to tap the gophers.

3.9, 4.1, 4.1, 4.2, 4.3, 4.3, 4.3, 4.4, 4.4, 4.4, 4.5, 4.7, 5.0, 5.1, 5.1, 5.2, 5.7



Median: Upper Extreme: Lower Extreme:

Step 1: Find the median of all the data. This is Q2.
Step 2: Find the median of the lower half of the data. This is Q1.
Step 3: Find the median of the upper half of the data. This is Q3.
Step 4: Draw the box from Q1 to Q2 and Q2 to Q3.
Step 5: Locate the greatest value and least value and mark them with vertical lines.
Step 6: Draw lines to represent whiskers from the Q1 to the least value and to Q3 to the greatest value.

Example 5:	Complete the five-point summary for the Gopher Game data above.

The Median:

The Lower Quartile: The Upper Quartile:

The Lower Extreme: The Upper Extreme:

We also find the inter-quartile range (IQR) of the data. This is the difference between the upper quartile range and the lower quartile range. In this case, it is _____ − _____ = _____. This is a very useful measure because it is not influenced by the extreme measures (an extremely fast time or an extremely slow time).

Section 4.9 Scatterplots

Looking Back 4.9

All the graphs we have looked at contain univariate data. In this section, we will move on to bivariate data. Scatterplots are graphic displays of bivariate data. They may demonstrate correlations or cause-and-effect relationships between the two types of data.

In science, scatterplots are often used to make conclusions based on given hypotheses using the scientific method with experimentation. Sometimes there is no cause-and-effect relationship between two sets of data and sometimes a relationship may be assumed when it does not exist.

Bivariate data contains one variable that is independent. The other variable is the dependent variable because it depends on the independent variable.

Looking Ahead 4.9

Example 5: Let us look at the relationship between temperature and ice cream sales.

Temperature in F°	Ice Cream Sales in Dollars
57.56	215
61.52	325
53.42	185
59.36	332
65.3	406
71.78	522
66.92	412
77.18	614
74.12	544
64.58	421
72.68	445
62.96	408

Follow the steps below to make a scatterplot of the data:

Step 1: _____ is the independent variable. Create an appropriate scale for the x-axis. The temperature ranges from $53.42°$ to $77.18°$.

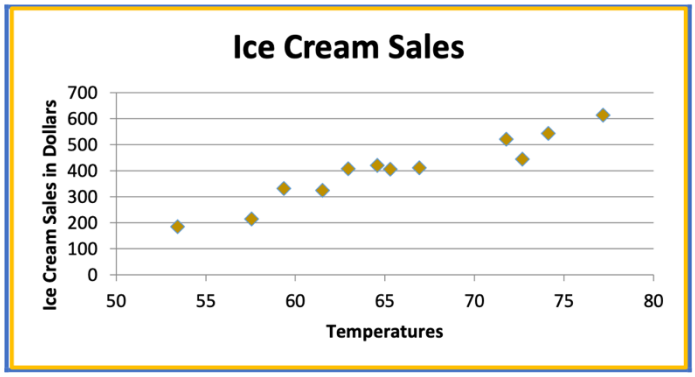
A good scale would be _____ with increments of _____.

Step 2: _____ is the dependent variable. Create an appropriate scale for the y-axis. The sales range from $\$185.00$ to $\$614.00$.

A good scale would be _____ with increments of _____.

Step 3: Draw the x-axis and y-axis with the scaled increments the same distance apart. Plot your data. Do this for each set of data points until twelve dots are on the scatterplot.

Your scatterplot will look something like this:



Section 4.10 Bivariate Data

Looking Back 4.10

Bivariate data involves two variables. One is called the independent variable and the other is called the dependent variable. Bivariate data also involves correlations or relationships between two sets of data. Sometimes one is the cause of the other, but not always.

Looking Ahead 4.10

In bivariate data, we graph the data and then look for trends and/or correlations. We make a scatterplot of the data, then draw a line of best fit that is as close as we can get to keeping half the data on each side of the line. To accomplish this, the line can go through any of the data points; however, sometimes the trend line will not go through any of the data points.

The independent variable goes on the x-axis and the dependent variable goes on the y-axis. If all values are positive, the data is displayed in the first quadrant.

If the data goes uphill from left to right, the correlation is positive.

If the data goes downhill from left to right, the correlation is negative.

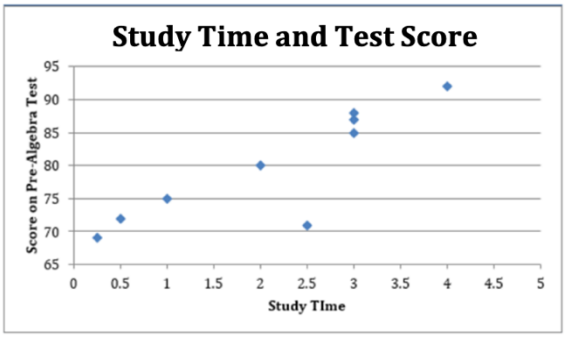
If the data is clustered close to an imaginary line, then it is a strong correlation, whether it is positive or negative.

If the data is not clustered close to an imaginary line, then it is a weak correlation, whether it is positive or negative.

If the data is all over the place, then there is no correlation.

Example 1: Let us look at the correlation between test scores and study time. The data from the table on the left is graphed on the right.

Hours of Study	Scores on Pre-Algebra Test
2.00	80
3.00	88
4.00	92
1.00	75
0.50	72
2.50	71
3.00	85
3.00	87
0.25	69



Draw a trend line, or line of best fit, to represent the general cluster of the data.

a) Is the correlation positive or negative?

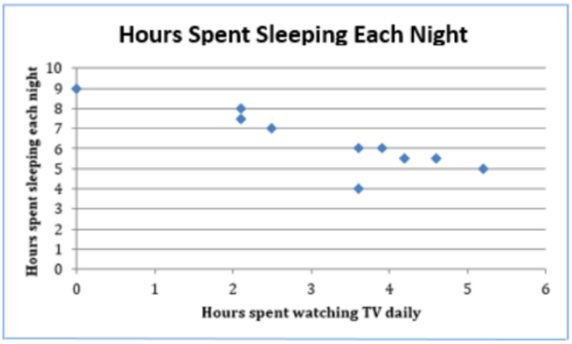
b) Is the correlation strong or weak?

c) What is the relationship between study time and test scores?

d) Can you predict the test score for a student who studies for 4.5 hours?

Example 2: Now, let us investigate the relationship between the time spent watching television daily and the hours spent sleeping at night to see if there is a strong correlation or weak correlation.

Hours Spent Watching TV Daily	Hours Spent Sleeping Each Night
3.6	4.0
2.5	7.0
2.1	8.0
5.2	5.0
0	9.0
4.6	5.5
3.9	6.0
3.6	6.0
4.2	5.5
2.1	7.5



Draw a trend line, or line of best fit, to represent the general cluster of the data.

a) Is the correlation positive or negative?

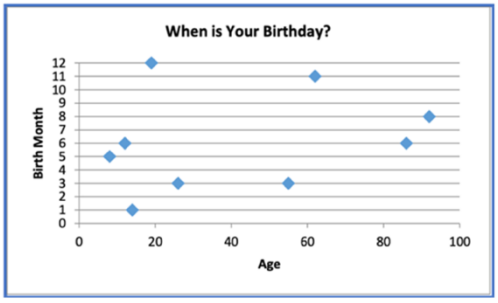
b) Is the correlation strong or weak?

c) What is the relationship between hours spent watching TV and hours spent sleeping each night?

d) Can you predict the hours spent sleeping for a person who spends 4.5 hours watching TV?

Example 3: Finally, let us explore the birth month and age of nine people to see if there is a correlation between the two.

Age	Birth Month
12	6
19	12
55	3
86	6
92	8
14	1
26	3
8	5
62	11



a) Is there a correlation between age and birth month?

b) What does this say about the relationship between age and birth month?

Section 4.11 Probability Sampling

Looking Back 4.11

We have been investigating various ways to graph and analyze data. Now, we will look at methods of gathering data. When a researcher does a study, he/she must clearly decide the group he/she wants to know something about. We call this a target population. To find out information about a target population, a researcher selects a sample group from which data will be gathered.

Let us suppose that a diaper company wants to improve their product. Their target population is mothers of newborns. If the target audience is small, they can all be included in the sample. If the target audience is large, the researcher must decide upon a method to gather a smaller sample of the larger population. The sample must reflect the target population. Sampling methods are classified as either *probability sampling* or *non-probability sampling*.

In this section, we will be investigating probability sampling; probability sampling methods include random sampling, systematic sampling, and stratified sampling.

Looking Ahead 4.11

Random sampling is the most probable method of probability sampling because each member has an equal chance of being selected and they are aware of this probability.

Systematic sampling is an often-used sampling method. In systematic sampling, every n^{th} person is selected from a list of population members. This is as good as random sampling as long as the sample is not previously ordered in some specific way.

Stratified sampling is a commonly used probability method. It is better than random sampling because it reduces the probability of error as the group is a smaller portion, which shares something in common, taken from the larger group.

Example 1:	Name the type of probability sampling performed for each given scenario below.

a) A group is randomly chosen in a work setting. All the female and male managers from the group are chosen to complete a survey.

b) All female swim team members at swim clubs in a given city count off from 1 one to 6. Every swimmer who counted a 3 is chosen to complete a survey.

c) Three dice are rolled and the numbers the dice land on form an area code. The part of the state with that area code will be selected as the sample to complete a survey.

Section 4.12 Non-Probability Sampling

Looking Back 4.12

We have learned that random sampling, systematic sampling, and stratified sampling are three methods of *probability sampling*. The advantage of these methods is that the sampling error, which is the degree to which the sample differs from the population, can be calculated. Remember, the population is the entire group from which the sample is taken. An example of a population could be all residents of Ohio. A sample is a smaller group that represents the population. An example of a sample could be all the residents of a specific three-digit area code in Ohio.

Now that we have looked at the *probability sampling* methods, we will explore *non-probability sampling* methods. These include convenience sampling, judgment sampling, voluntary-response sampling, and quota sampling.

Looking Ahead 4.12

Convenience sampling is just like it sounds: convenient. In this type of survey, the researcher just surveys whoever is available. For example, a teacher may survey her whole first period class simply because they are all there. This does not involve the time or cost of a random sample.

Judgment sampling is also like it sounds. The researcher judges whether or not the sample represents the population. For example, the researcher may decide that sampling a given city is representative of the whole state.

Voluntary-response sampling is like it sounds as well. Volunteers are requested to reply to mailings or emails, etc. For example, a magazine may send out a survey to all its female readers and ask them to please respond.

Quota sampling for *non-probability sampling* is like stratified sampling for *probability sampling*. This method relies on the sample population being a good representation of the whole population. For example, a researcher may investigate the fast-food dining habits of one-hundred teenagers between the ages of sixteen and eighteen.

Example 1:	Name the type of *non-probability sampling* performed for each given scenario below.

a) In voting, the sample population must be a fair representation of the entire population.

b) A bus driver surveys the tour group on board once all of them are present.

c) A church mails a letter to all of its members asking them to please fill out and return a response card.

d) A coach decides that a sample of the football team is representative of the school athletes.

Section 4.13 Sampling Bias

Looking Back 4.13

There are *probability* and *non-probability* methods in sampling. No matter which method we use, there is always bias in sampling methods; some are more biased than others.

Bias means certain outcomes are more favored than others. This can be intentional or unintentional, but being aware of it can help us avoid commonly made mistakes in statistics by identifying them initially. Random samples are considered the least biased. Voluntary-response samples are considered the most biased.

Life is unfair, but God is just. Some bias is unavoidable and inherent in research, but God tells us in James 2:1-9 to try to "treat others as fairly as possible, not to show partiality or favoritism to others, and to love our neighbors as we love ourselves." This is as fair and unbiased as we can possibly be in our treatment of others.

Looking Ahead 4.13

Identify the bias in each of the following examples:

Example 1:

Random Sampling

A researcher goes through a phone directory and randomly calls computer-generated numbers and asks the receivers to take a survey

Example 2:

Systematic Sampling

A researcher picks the n^{th} person that comes through a door to survey

Example 3:

Stratified Sampling

A researcher studies the grades of students at a high school by investigating a smaller sample of the larger population

Example 4:

Voluntary-response Sampling

A researcher asks people to take a survey about the quality of food in a grocery store and gathers results from the filled-out surveys

Example 5:

Convenience Sampling

A researcher asks fellow bus passengers questions on the bus ride home from school

Example 6:

Judgment Sampling

A researcher decides who to survey about the time it takes to get ready for work in the morning based on their own thought and feelings

God says in Matthew 7:1, "Judge not, lest you be judged." He tells us in Hebrews 4:12 that His Word is "able to judge the thoughts and intentions of the heart."

Example 7:

Quota Sampling

A researcher figures out which age group prefers NIV Bibles by breaking up the age groups they survey and then taking data from small samples of the larger populations