Open Educational Resources

Queensborough Community College

2015

Mathematics in Contemporary Society Chapter 4

Patrick J. Wallach
Queensborough Community College

Chapter 4

Correlation and Causality

Correlation exists between two different data sets (or two different variables) when an increase in the values in one data set tend to correspond with increases or decreases in the other data set.

Positive correlation exists when an **increase** in the values in one data set tends to correspond with **increases** in corresponding values in the other data set. Inversely, a **decrease** in the values in one data set tends to correspond with **decreases** in corresponding values in the other data set. In positive correlation, the data sets tend to do the same thing together (either both increase or both decrease).

For example:

SAT Scores and IQ Scores
Weight of turkey and cost of same turkey
Height and Weight

We would probably believe that, for any given person (or any given turkey), that as the left value increases the corresponding value on the right tends to increase as well.

Question 1: Give an example of a pair of data sets with positive correlation.

Negative correlation exists when an **increase** in the values in one data set tends to correspond with **decreases** in corresponding values in the other data set. In negative correlation, the data set tend to do the **opposite** (one increases, one decreases) of the other.

For example:

Store price of can of soda and Number of cans sold at the store
Number of drinks a person has at a party and Person's grade on exam the next day
Pitcher's ERA and Number of Innings Pitched

We would probably believe that, for any given person or thing, that as the left value increases the corresponding value on the right tends to decrease as well.

Question 2: Give an example of a pair of data sets with negative correlation.

No correlation exists when an **increase** in the values in one data set tends have **no corresponding effect** on the value in the other data set.

For example:

Height and SAT Scores

Cost of can of soda and cost of gallon of gas
IQ Scores and Weight

We would probably believe that there is no relationship between the values on the left and the values on the right.

But who knows? If we want to be sure, we can test it. For instance, to test the first example, we could gather the height and SAT score values of 50 students. By examining the Height-SAT Score pairs (either by themselves or with a graph), we could determine for sure whether or not correlation exists. For example, if it seems like the taller students tend to have higher SAT scores, there is positive correlation between height and SAT scores.

<u>Question 3:</u> Give an example of a pair of data sets with no correlation.

We can explore correlation with scatter plots (or scatter diagrams). You will actually create your own in Lab #4. We associate certain results with positive, negative and no correlation.

Suppose we compare the values in Data Set A with corresponding values in Data Set B.

For example, suppose Data Set A is a person's age and Data Set B is the person's corresponding score in a test of skill. Here are a few possibilities:

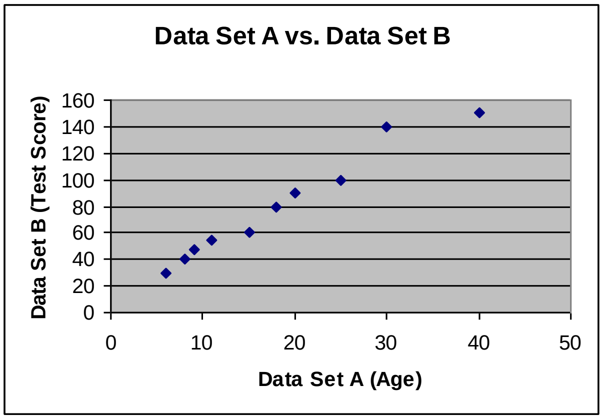
Example 1:

We can create a table listing the age and test score (it doesn't matter what kind of test) of a set of people. Suppose we examine 10 people and create the following table:

Name	Age (Data Set A)	Test Score (Data Set B)
Sue	15	60
Bob	18	80
Amy	20	90
Joe	25	100
Lynn	30	140
Jose	40	150
Lisa	11	55
John	9	48
Lori	8	40
Patrick	6	30

The names become unimportant. We can look at each person as a pair of scores. There is person with an age of 15 and a score of 60. There is a person with an age of 18 and a score of 80. And so on, until the last person with an age of 6 and a score of 30.

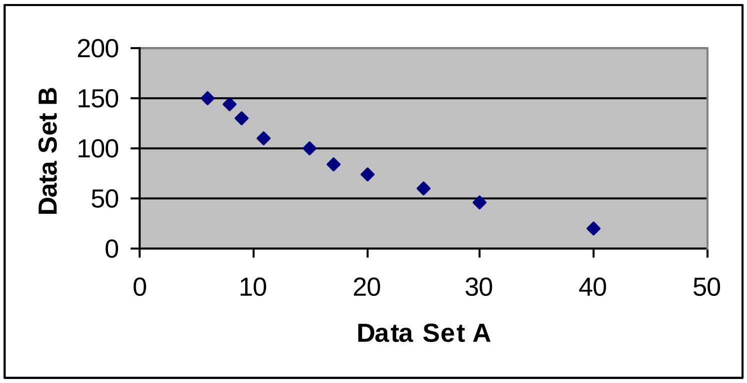
Each person becomes a point of the scatter plot. If you look below, the highest point (on the right) represents the 40 year old with a score of 150. Nearby is the 30 year-old with a score of 140. The lowest point on the left represents the 6 year-old with a score of 30.



(In positive correlation, as values in Data Set A increase, so do the corresponding values in Data Set B. This graph suggests that, for this test, as a person gets older, their corresponding test score increases.)

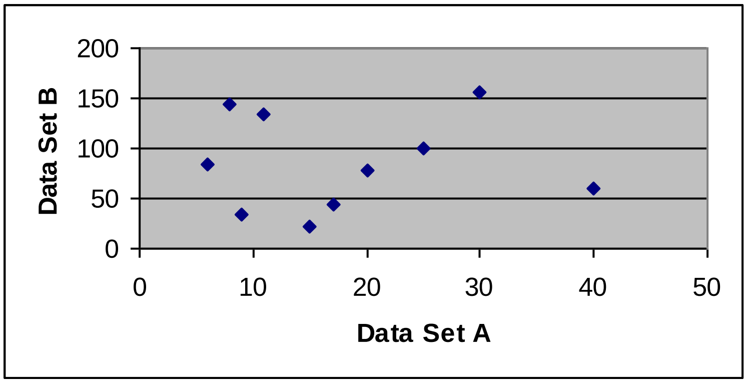
We will now look at other scatter plot examples independent of the tables that generated them.

Example 2:



(In negative correlation, as values in Data Set A increase, the corresponding values in Data Set B tend to decrease. This graph suggests that, for this test, as a person gets older, their corresponding test score decreases.)

Example 3:



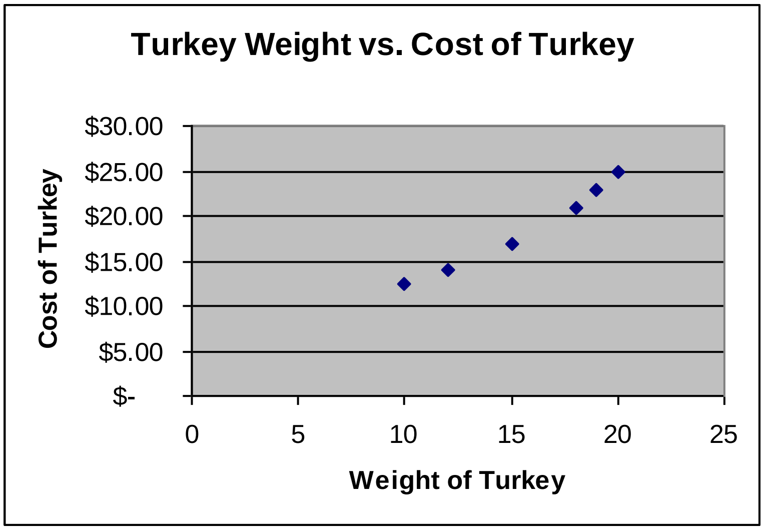
(In no correlation, there seems to be no pattern. There is no apparent relationship between age and test scores here.)

After determining that correlation (positive or negative) exists, we are also interested in the level of correlation. This is easier to determine with scatter plots or the **correlation coefficient**.

(Don't worry too much about correlation coefficient just yet. In Lab #4, you will see that any set of paired data has a corresponding correlation coefficient. The correlation coefficient is a value between –1 and +1 that measures correlation.)

With **strong positive correlation**, the scatter plot points more closely resemble a straight line (going up left to right). The correlation coefficient is near +1.

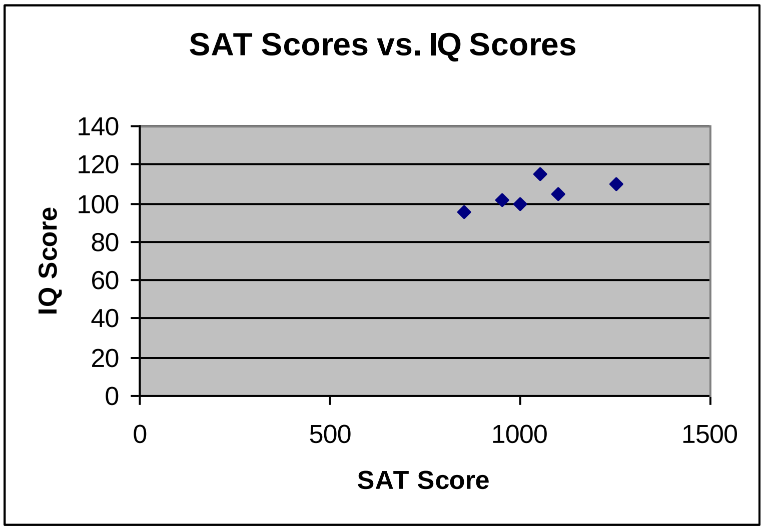
Example 4:



We can see the points of the scatter plot could almost be connected with a straight line. Obviously, as the weight of a turkey goes up, it will almost certainly cost more. There is a strong positive correlation between turkey weight and turkey cost.

With **weak positive correlation**, there is a general trend of increase (left to right) in the scatter plot and a correlation coefficient of about 0.6 to 0.8.

Example 5:

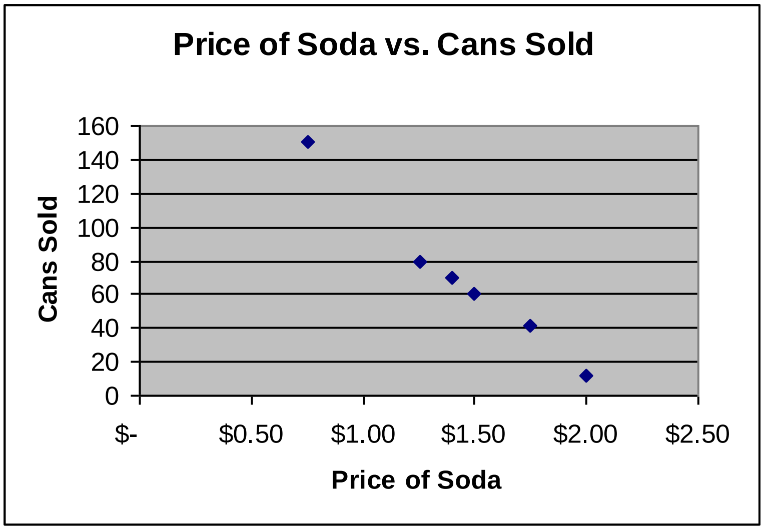


We can see the points of the scatter plot tend to rise as we move from left to right, but not so much that we could connect the dots with a simple straight line. But we see a general trend that IQ scores tend the rise as SAT Scores do.

Question 4: Now that we know more about correlation, give an example of a pair of data sets with strong positive correlation. Give an example of a pair of data sets with weak positive correlation.

With **strong negative correlation**, the scatter plot points more closely resemble a straight line (going down left to right). The correlation coefficient is near -1.

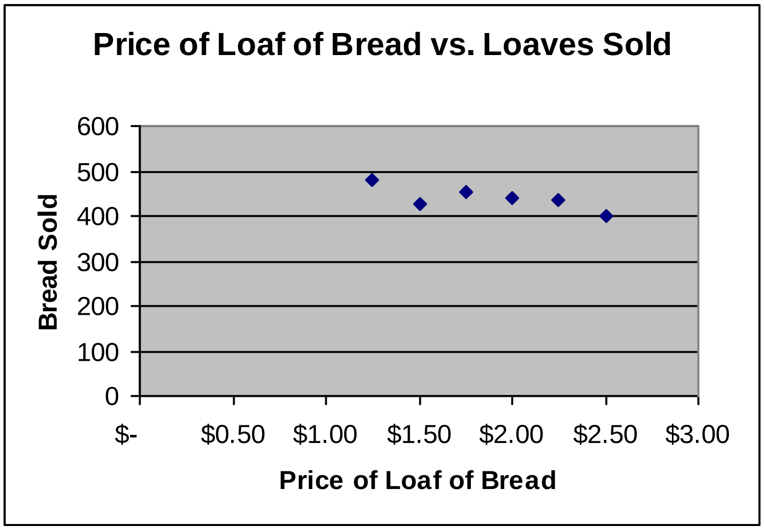
Example 6:



We can see the points of the scatter plot could almost be connected with a straight line. Obviously, as the price of a can of soda goes up, less soda will be sold.

With **weak negative correlation**, there is a general trend of decrease (points go down left to right) in the scatter plot and a correlation coefficient of about -0.8 to -0.6.

Example 7:



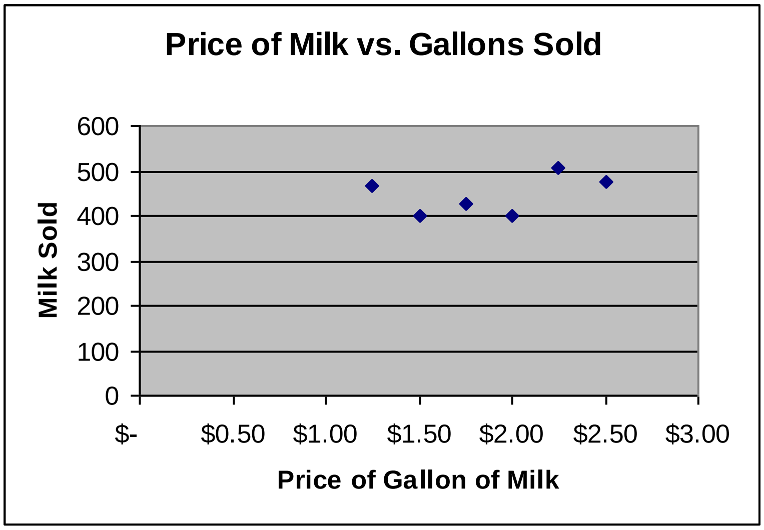
We can see the points of the scatter plot tend to drop as we move from left to right, but not so much that we could connect the dots with a simple straight line. But we see a general trend that bread sales tend to fall as the price rises.

Question 5: Why do you think the negative correlation is stronger in the first example? (Hint: Think about the nature of the items.)

Question 6: Give an example of a pair of data sets with strong negative correlation. Give an example of a pair of data sets with weak negative correlation.

With **no correlation**, the scatter plot points are scattered about the graph, and do not resemble a straight line or an upward or downward trend. The correlation coefficient is generally between –0.5 and +0.5.

Example 8:



In this example, there doesn't seem to be any relationship between the price of milk and how much milk is sold. This suggests that people will buy milk no matter what the price.

How to Explain Correlation

Finding correlation doesn't necessarily mean that correlation actually exists between two data sets (or two variables), and there are different ways to explain it:

1) The correlation may be a coincidence. (If I happen to find a set of ten students such that the five tallest students have the highest GPAs, is that enough to say that height and GPA are positively correlated? I would probably be more comfortable making that conclusion with a larger set of students, say 50 or 100.)
2) It may be that both data sets are influenced by some other underlying cause. (You may find that the price of airline tickets and the price of shipping packages tend to increase at the same time, but they may both by affected by the rising costs of fuel.)
3) It may be that one of the variables is the **direct cause** of the values of the other variable. (We refer to this as **causality,** which we will explore tomorrow.)

Question 7: Give an example of a correlation you may encounter by studying college students. Explain this correlation using one of the explanations above.

Lab Assignment #4 – Correlation
Due _________________

If we have paired numerical data, we can use the **Charts** window to create scatter plots (or scatter diagrams).

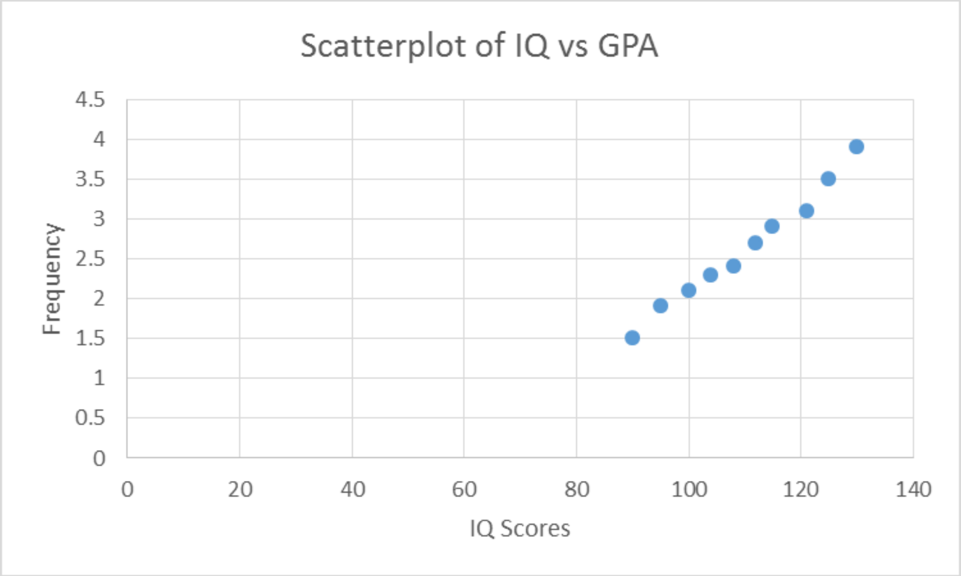
Correlation exists between pairs of values if increases in one value tend to correspond with increases (or decreases) in the other value. The following problems explore correlation graphically.

1) Is IQ positively correlated to GPA? Create your own test by creating 10 student results (each as an IQ-GPA pair). For example

IQ Score	GPA
90	1.5
95	1.9
100	2.1
etc.	etc.

Each line represents the IQ Score and GPA of one student. Add 7 more examples.

1) You can create a scatter plot in Excel using the **Charts** Window. Select the **Scatter** option. You should have a graph that looks something like this:



<u>Question 1:</u> Does there seem to be a positive correlation? Why or why not? If the correlation is positive, does it seem to be weak or strong?

Now let's look at another example:

2) Zippy Car Rental wants to know if the price of renting a car is correlated to car rentals per day. Create your own test by examining the data from the last 10 days:

Price	Rentals
$30	153

$40	121
$45	108
etc.	etc.

(Put in your own set of 10 paired values here.)

From the table, you can create another scatter plot.

Question 2: Does there seem to be a negative correlation? Why or why not? If the correlation is negative, does it seem to be weak or strong?

3) Create your own analysis of two different kinds of values that you think have no correlation between them. (For example, there is probably no correlation between shoe size and IQ score. You should make up your own example, though.) Repeat the process of 1) & 2).

Question 3: Does there seem to be any correlation? Why or why not?

4) The **correlation coefficient** measures correlation. If it is near +1, strong positive correlation is indicated; if it is near –1, strong negative correlation is indicated. **Weak positive correlation** tends to be between 0.6 and 0.8. **Weak negative correlation** tends to be between -0.6 and -0.8. No correlation is indicated generally for values between –0.5 and +0.5.

You can calculate the correlation coefficient by doing the following:

a) If you want to compare values in cells A2 to A11 to the cells in B2 to B11, click on a cell below the values (like B13) and then click on f_x (**Insert Function**, next to the Formula Bar).
b) Under **Select a Category**, select **Statistical**
c) For **Select a Function**, select CORREL
d) For **Array1**, select the first set of values (A2:A11)
e) For **Array2**, select the second set of values (B2:B11)
f) The number placed in the cell below is the correlation coefficient!

Question 4:

a) Calculate the correlation coefficient for examples 1)-3). You should get a number between -1 and +1. (For example, a correlation coefficient of 0.972 indicates strong positive correlation.)
b) How do the results you obtained correspond to the answers you gave in 1)-3) concerning correlation?

How to Explain Correlation

Let’s talk about causality.

Causality

Causality is the existence of a cause-and-effect relationship between two different variables. To show causality (not just correlation), we need to satisfy a stronger set of standards to be certain beyond a reasonable doubt.

For example, we may want to show that **cigarette smoking causes lung cancer**. This is something we may suspect is true, but we want to come to this conclusion through careful study. Here are some examples of what can be done to establish causality in this case:

1) **Determine correlation first.** We may start with a study of lung cancer victims. We could compare the amount of cigarettes the person smoked per day (nonsmokers would be 0) to their age when the cancer was diagnosed. If we could show negative correlation (as the number of cigarettes smoked increases, the age at which cancer is diagnosed decreases), that’s a good start.
2) **Account for other factors.** We would also want to explain why other factors are not the cause of lung cancer. For example, we cannot say air pollution is the cause of lung cancer, because cigarette smokers in rural, unpolluted areas still have a high incidence of lung cancer. We could also try to eliminate a genetic explanation for lung cancer by showing that only the smokers among siblings developed lung cancer.
3) **Show that when the cause is present the effect is present, and when the cause is absent the effect is absent.** If possible, we would want to examine the incidence of lung cancer in nonsmokers. If we cannot find any lung cancer victims who aren’t smokers, or we find very few nonsmoking lung cancer victims and we can explain the reason (a nonsmoking lung cancer victim has a smoking spouse), we would tend to believe smoking is the cause.
4) **Try an experiment if possible.** If can conduct an experiment (perhaps on human or animal subjects, which may be unethical) that shows a definite relationship between cigarette smoke and lung cancer, we would tend to believe smoking is the cause. Perhaps scientists could expose laboratory rats in a treatment group to cigarette smoke and compare the results to a control group of rats who are not exposed. (Of course, many people may view such an experiment as cruel and unnecessary.)
5) **Determine the scientific reason why the cause produces the effect.** If we can scientifically explain the physical changes in the lung of a smoker which increase the incidence of lung cancer, we would again tend to believe smoking is the cause. (I imagine most of you have already seen pictures comparing a smoker’s lung to a nonsmoker’s lung—there is an obvious difference.)

These are just some general guidelines. What works in one example may not make as much sense in others.

With this information in hand we can decide if smoking is the cause of lung cancer, either:

1) As a **possible cause**—we have found a correlation, but do not have enough information to believe in causality. (In the criminal justice system, possible cause is probably enough for a person to be picked up by the police and questioned about a crime.)
2) As a **probable cause**—we have good reason to believe in causality, but there may be some unexplained or unclear issues. (In the criminal justice system, probable cause is probably enough for a person to be charged with a crime and sent to trial.)
3) As a **cause beyond a reasonable doubt**—we are almost certain that causality exists, we have eliminated other explanations, and we have a scientific basis for our belief. (In the criminal justice system, cause beyond a reasonable doubt is the standard used to convict a suspect of a crime and send him or her to jail.)

Question 8: Which of the above (1-3) is your level of confidence in causality between cigarette smoking and lung cancer? Explain why.

Question 9: Explain the difference between correlation and causality.

Question 10: The following are examples of correlation. For each situation, determine if the correlation can be explained as one of the following:

a) Simple coincidence
b) Effects related to another underlying cause
c) Causality (a cause and effect relationship between the two variables)

Defend your choice.

I) The smarter students in Dr. Wallach's MA-240 class sit in the back of the room.
II) Cirrhosis of the liver seems to be most common among heavy drinkers.
III) 30% of the people who have used the medication Zipporill in the last year have had heart attacks sometime during the year.
IV) Florida was hit by a record number of hurricanes this year, and the summer temperatures were below normal throughout the U.S.
V) The average annual temperature of the world has slowly increased over the last 50 years; so has the level of air pollution in the atmosphere.

Writing Assignment: Conduct your own study to establish causality. Choose **one** of the following as an effect:

Global warming
Skin Cancer
Breast Cancer
Road Rage

Underage drinking
Decline of the sea turtle population

For the effect situation you have chosen, do the following:

1) Identify a possible cause.
2) What correlation do you suspect between the cause and the effect? Describe the correlation.
3) How would you try to establish correlation and causality? (Refer to the guidelines from the earlier example.)
4) What are other possible causes?
5) What is your level of confidence in causality?

Characteristics of Data Sets

When you take an exam, the first thing you want to know is how you did. OK, you get your exam back and you got an 80. What's the second thing you want to know? You'll probably start looking around—how did everyone else do?

You may see some scores above 80 and some scores below 80. But how did the class do overall? We want a way to measure the performance of the "average" student and compare our exam results to him or her.

Measures of center of a data set describe the "average" value of the data set. But what is "average"? There are different measures we can use:

A) The **mean** of a data set is the numerical average.

Example: Suppose a professor gives a test to 10 students. The scores are:
89 95 82 78 97 68 99 83 84 71

We can calculate the mean with the formula:

$$\text{Mean} = \frac{\text{Sum of Values}}{\text{Number of Values}}$$

$$= \frac{(89+95+82+78+97+68+99+83+84+71)}{10}$$

$$= \frac{846}{10}$$

$$= 84.6$$

The mean is easy to obtain with a small data set. It is the measure of average we are most familiar with and it is usually reliable.

<u>Question 11:</u> Calculate the mean of the following set of GPAs
3.2 2.9 3.1 2.7 3.3 3.6 2.1 4.0 2.2 1.7 3.9

B) The **median** of a data set is the middle term (or average of the 2 middle terms) when the numbers of the data set are arranged in ascending order.

Example: Suppose we look at the high temperature of the last seven days. The temperatures are: 72 78 68 79 73 81 65

To calculate the median, we arrange the numbers from lowest to highest:
65 68 72 **73** 78 79 81

The number in the middle (with three values above and three values below) is 73. So 73 is the median.

Example: Suppose a professor looks at class attendance for the last eight days. The number of students in class on each day was: 24 21 19 18 14 28 25 22

To calculate the median, we arrange the numbers from lowest to highest:
14 18 19 **21** **22** 24 25 28

The numbers in the middle (with three values above and three values below) are 21 and 22. The average of 21 and 22 is:

$$= \frac{(21+22)}{2}$$
$$= \frac{43}{2}$$
$$= 21.5$$

So 21.5 is the median.

<u>Question 12:</u> Calculate the median of the following set of IQ scores
105 107 115 118 121 99 117 119 125 95 103

When the numbers of a data set are arranged in order, finding the median is relatively easy. Suppose you had a table of 1,000 SAT scores arranged in order:

Number	**SAT Score**
1	400
2	410
3	410
etc.	etc.
etc.	etc.

1000	1600

Where can the median be found? You look for the middle values. Is the 500th value in the middle? Yes, that's one of the middle values. The 501st value is also in the middle. There are 499 scores below the 500th value and 499 scores above the 501st value. So that's what we look for:

Number	***SAT Score***
1	400
2	410
3	410
etc.	etc.
500	1000
501	1010
etc.	etc.
1000	1600

If the two values in the middle are 1000 and 1010, we would take the average of the two numbers and find that 1005 is the median.

One of the useful characteristics of the median is that it divides the data set in half. If the median of 1,000 SAT scores is 1005, then roughly 50% of the scores are below 1005 and roughly 50% of the scores are above 1005. (I say "roughly" because in some cases the median may equal some values in the data set.)

Question 13: If you had 5543 ACT scores arranged in order, where could you find the median?

C) The **mode** of a data set is the value (or values) that appear with the greatest frequency.

Example: Suppose a professor gives a test to 12 students. The scores are:
88 93 82 **88** 97 68 99 83 82 71 85 **88**

Since 88 appears the most often, 88 is the mode.

Example: Suppose we look at how many runs the Mets scored during their last 10 games. The values are: **3** 5 **2** 8 **2** **3** 7 1 0 9

Since 2 and 3 appear the most often (twice), there are two modes, 2 and 3. (Don't try to average them!)

We see that one limitation on the mode is that it not necessarily unique. In fact, it may not exist at all! Consider the next example:

Example: Suppose the GPAs of the students in a class are:
2.3 3.1 3.6 2.9 3.7 2.8 2.2 3.4 3.8

There is **no mode** here because each value appears only once.

So what's so great about the mode? If a test is given to 25 students and it is announced that the mode for the test was 82, what does that tell you? It may mean that two students received an 82 and all other scores were unique. (That's not that useful. Who knows what all the other scores were?)

But what if I ask 25 students to vote for their favorite ice cream flavor and get the following results:

Vanilla Chocolate Vanilla Chocolate Strawberry Chocolate
Strawberry Chocolate Chocolate Vanilla Chocolate Vanilla
Vanilla Chocolate Vanilla Chocolate Chocolate Chocolate
Vanilla Strawberry Chocolate Vanilla Chocolate Vanilla
Chocolate

The final tally is:

Vanilla:	9 votes
Chocolate:	**13 votes**
Strawberry:	3 votes

Since Chocolate received the most votes, Chocolate is the mode and the winner!

We see then that the mode can be extremely useful for nonnumerical data. Although it makes no sense to find the mean or median ice cream flavor, I can certainly find the mode.

<u>Question 14:</u> Give examples of three data sets for which it would be most helpful to know that mode instead of the mean or median.

Comparing Measures of Center

Which measure of center is the best? It depends. To help answer the question, let's talk about outliers.

Outliers are data values that are unusually higher or lower than the rest of the values in a data set.

Example: Suppose a professor gives a test to 7 students. The scores are:
86 97 88 77 80 86 12

Obviously, 12 is an outlier. Perhaps this represents a student who came in an hour late and only answered one question. Certainly, the 6 others scores are very different.

How does the score of 12 affect the mean, median and mode? Let’s see:

$$\text{Mean} = \frac{(86+97+88+77+80+86+12)}{7} = \frac{526}{7} = 75.14$$

This is interesting, because 75.14 is below 6 of the 7 scores. The outlier has caused the mean to be significantly lower.

The values in order are: 12 77 80 **86** 86 88 97

Median = 86

The median is not really affected by the outlier.

Mode = 86

The mode is not really affected by the outlier.

We should note here that if we changed the value of 12 to 22, or 32, or 52, or 72, it would change the value of the mean but not the value of median and mode!

For additional consideration, if we drop 12 from the data set, we would get the following results:

Mean = 85.67 (a much more reasonable result)
Median = 86
Mode = 86

We see, from this example, that the mean tends to be the most affected by outliers. Outliers can cause the mean to be a poorer representation of average than the median and mode.

Here is summary chart about the three measures:

Mean	Median	Mode
Easy to calculate (with calculator or small set)	Easy to calculate (when data set is ordered)	Easiest to calculate by visual inspection

Takes all values into account	Takes order only into account	Takes repetition only into account
Most affected by outliers	Less affected by outliers	Least affected by outliers
Most reliable	Mostly reliable	Least Reliable
Most frequently used measure of center	Splits data in half	Can be used for nonnumeric data

Question 15: Consider the table. Why do you think the mean is considered the most reliable, the median mostly reliable, and the mode is the least reliable?