Section 1: Organizing Data

Statistics is a branch of mathematics used to collect, describe, analyze and make predictions based on data. **Data** are measurements of an observation under study.

For example, weather data (such as temperature, humidity, wind velocity, and precipitation) are collected and reported daily. Scientists use these data not only to predict the weather for the next few days, but also to determine long-term trends in climate. If a scientists want to determine whether global warming exists, they must decide what data to include and how to analyze those data. Daily temperature readings have not been taken at every point on the earth's surface; they are taken mainly in cities and at scientific weather stations. Moreover, a continual, instantaneous temperature reading for an entire day may not be very informative for the purpose of judging climate change. Since we cannot include all the temperature data for each instant of the day at every point on the earth's surface, we must take a **sample** of the relevant data. (A sample is a subset of data that is used to represent the whole.) The whole group under study is called the **population.**

Samples can be collected in various ways, and the method of choosing a sample is very important in maintaining the integrity of a study. There are four basic sampling techniques: random, systematic; stratified, and cluster.

A **random sample** is a sample in which each member of the population has an equal chance of being selected. For example, if a researcher wanted to do a study about students at a particular university, the researcher could use a computer program that listed all the students' identification numbers and then chose some subset of those numbers in a random fashion.

A **systematic sample** is a sample generated by giving every item in the population a number and then choosing every *k*th member, where *k* is a natural number. For example, the researcher doing the university study above might choose every 100^{th} student, beginning with the student whose number is the 38^{th} in the list. (The beginning number, 38, must be chosen randomly.)

If a population is divided into groups of similar characteristics (a university study might divide students into groups of freshmen, sophomores, juniors, seniors, and graduate students), and a researcher chooses members from each group randomly, this sample is called a **stratified sample**.

If a researcher uses an existing group that represents the population, this sample is called a **cluster sample**. For example, if a researcher interviewed all the students who crossed in front of the administration building between the hours of 10 and 2 on a particular day, this would be a cluster sample. Similarly, testing the light bulbs in one box of an outgoing order would be a method of cluster sampling.

The data collected by taking measurements (or asking questions) are **raw data**. In order to describe the data or to draw conclusions, the researcher must organize the data. There are several methods of doing

this. Two such methods **are stem-and-leaf plots** and **frequency distributions**. Suppose that a researcher has recorded the following temperature readings at 3:00 pm for each day in the month of August:

Example 1:

87, 98,99,85,78, 94, 88, 85, 97, 104, 99, 95, 90, 88, 85, 82, 82, 79, 76, 88, 90, 92, 87, 95,99,88,87, 84, 77, 89, 102.

The researcher wants to order these data to see how many days had temperatures over 100, in the 90's and so on. The researcher might then use a stem-and-leaf plot, in which the stem would be the number of tens in the temperature reading, and the leaves would be the digits in the ones' place.

Table 1:

Stem	leaves
10	4, 2
9	8,9,4,7,9,5,0,0,2,5,9
8	7,5,8,5,8,5,2,2,8,7,8,7,4,9
7	8, 9,6,7

Since the temperatures range from the 70s to the 100s, the stems go from 7 to 10. To indicate the first temperature, 87, we place a "7" in the leaves column across from the stem "8." By arranging the data this way, we can see immediately that there were more days with temperature readings in the 90s than in any other group, and that there were fewer days in the 100s than in the other ranges. There were no days with a temperature recorded below 76.

Exercise 1: Find the average daily temperatures for your home city for the month of July 2015 and arrange them in a stem-and-leaf plot.

Stem-and-leaf plots can be useful when the data recorded are numbers. However, this is not always the case. For example, suppose you wish to record the air quality index in your city each day for a month. These indices are green, yellow, orange, red, purple, and maroon. Each morning, you listen to the weather forecast, and you write down the air quality index for the day on your calendar. At the end of the month, your calendar looks like this:

Example 2:

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
						1 green
2 green	3 yellow	4 yellow	5 orange	6 yellow	7 green	8 green
9 green	10 green	11 yellow	12 yellow	13 orange	14 yellow	15 yellow
16 green	17 green	18 yellow	19 orange	20 red	21 orange	22 yellow
23 yellow	24 green	25 green	26 green	27 yellow	28 green	29 green
30 green						

How would you arrange these data to give an overall view of the air quality for the month? You could use a frequency distribution. Because the data are not numerical, this frequency distribution is called a **categorical frequency distribution**.

Table 2.

Type	Tally	Frequency
Green	𝍸 𝍸 ||||	14
Yellow	𝍸 𝍸 |	11
Orange	||||	4
Red	|	1

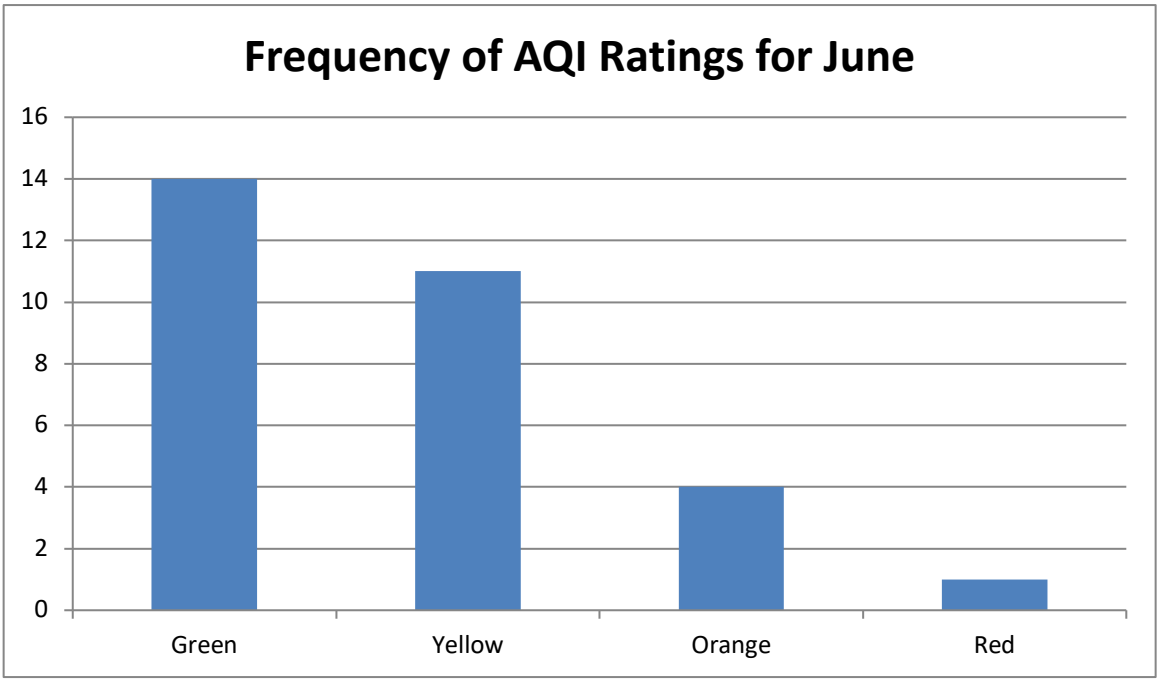
As you look at the calendar, you make a mark in the tally column for each day in the green range. When you see how many marks you have made, you list that number in the frequency column. Then you do the same for each of the other AQI codes. (Add the frequencies to make sure you get the right total!)

Exercise 2: Make a frequency table for the air quality index codes for Los Angeles in July 2015.

Section 2: Picturing Data

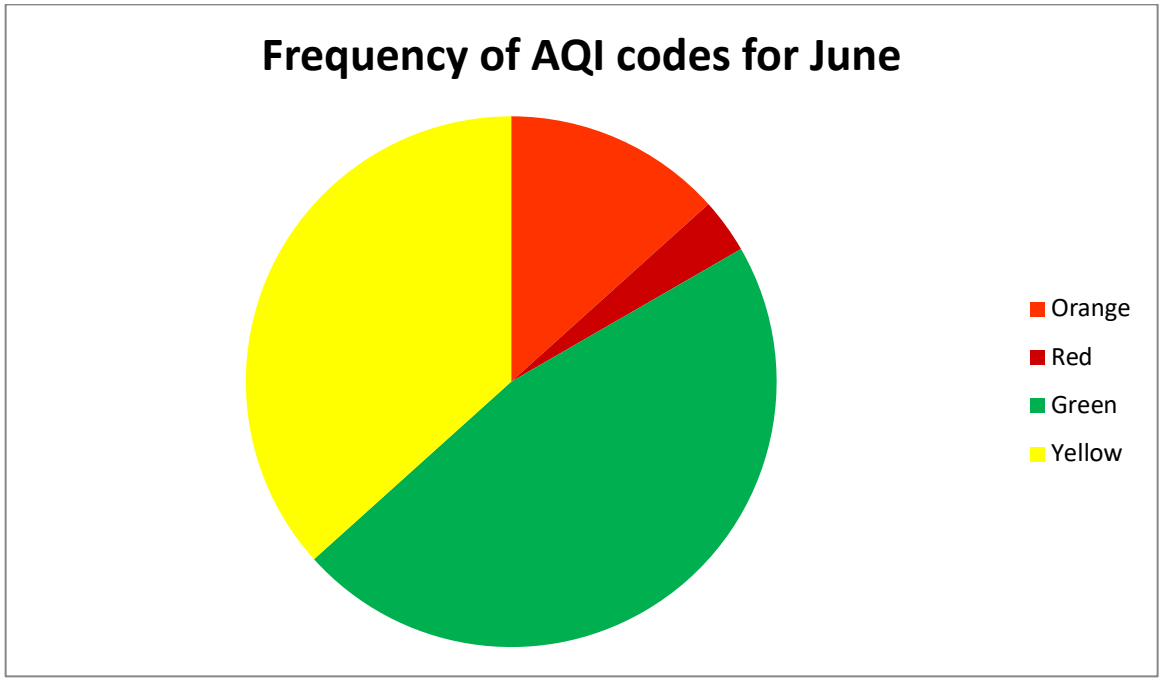
If you have ever looked at a *USA Today* newspaper, you have probably seen some pictures or diagrams describing data. Four common ways to represent data are with **bar graphs, pie charts, frequency polygons,** and **time series graphs.** Let us use the data from Example 2. We would like to represent our findings with a bar graph. We will have four bars, one for each air quality index recorded in the month. The bars appear along the horizontal axis, with the appropriate air quality index labeled below. Frequencies are indicated on the vertical axis.

Figure 1.



We could represent the same data with a pie chart. The full circle represents ALL the days of the month, and each section represents the fraction of those days that had the relevant AQI code.

Figure 2.



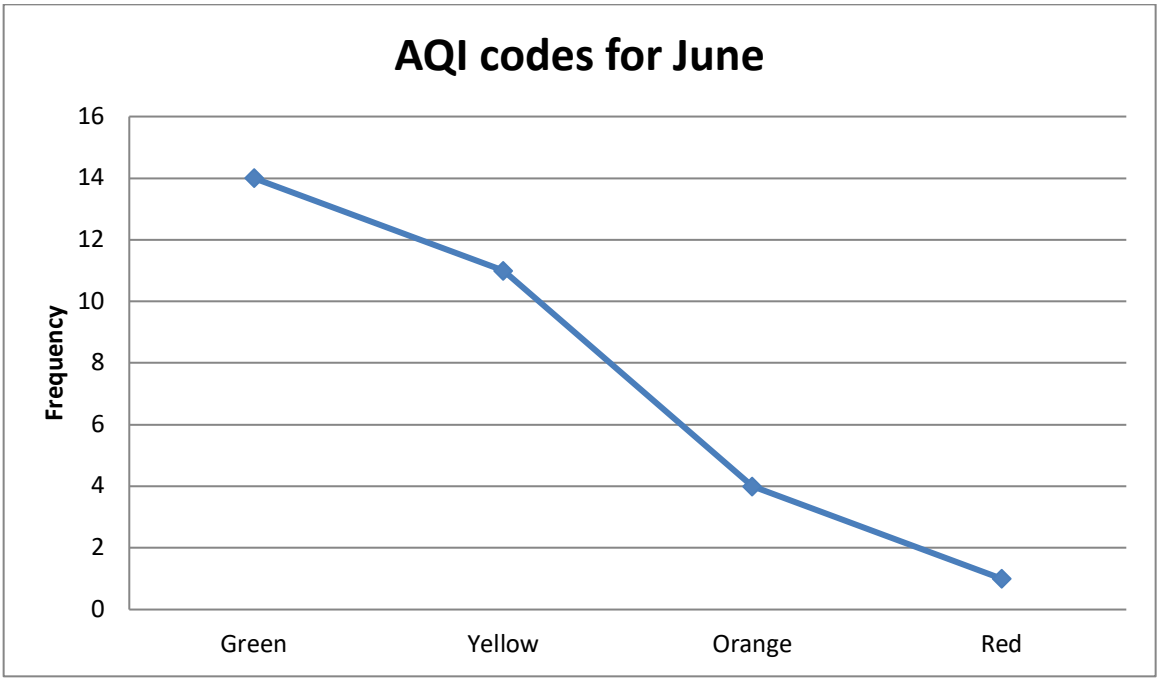
Suppose you were drawing this chart by hand, using a protractor. A circle has 360 degrees. The month of June has 30 days. How many degrees correspond to a single day?

$$\frac{360°}{30} = 12°$$

Therefore, since only one day had a reading of red for the air quality index, the red portion of the circle should be represented by a 12° section. Since four days had an orange rating, $4 \times 12° = 48°$ should be the size of the orange section, and so on. (These charts can be made relatively easily using a spreadsheet program.)

Finally, we can represent the same data with a frequency polygon. This chart is similar to the bar graph, but instead of using bars to represent the frequency of each type of data, a single point is sketched on the graph, and then the points are joined by straight line segments.

Figure 3.



Often, when we are studying data relating to climate, we want to see whether a particular measurement has changed over time. In this case, we will draw a **time series graph**. Let's suppose we are studying wildfires in California. We want to see whether the number and severity of the forest fires have changed over recent years.

Consider the following data:

Table 3.

Year	Fires	Acres
1998	5,227	92,456
1999	7,562	285,272
2000	5,177	72,718
2001	6,223	90,985
2002	5,759	112,810
2003	5,961	404,328
2004	5,574	168,134
2005	4,908	74,004
2006	4,805	222,896
2007	5,647	425,238
2008	4,923	347,310
2009	3,546	73,098
2010	2,961	23,191
2011	3,056	51,889
2012	2,922	94,510

2013	4,681	88,169
2014	5,620	90,606

http://cdfdata.fire.ca.gov/incidents/incidents_statsevents#2000

We could chart either the number of forest fires over the years in question, or the number of acres burned. We would have the following two time series graphs:

Figure 4.

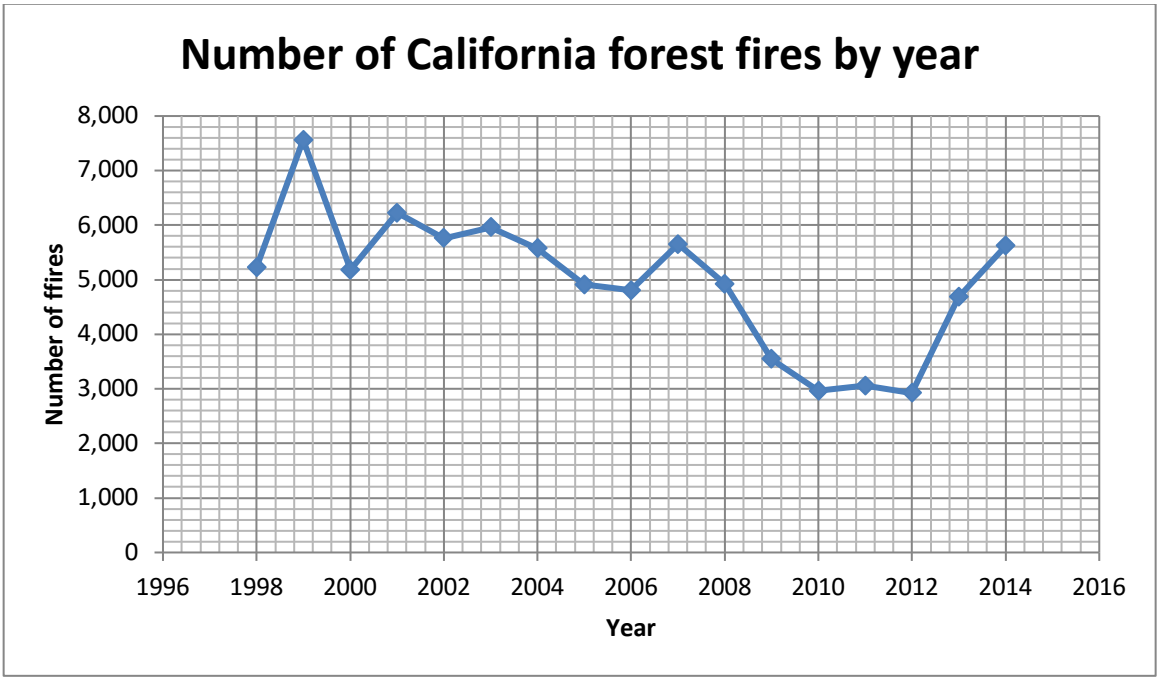
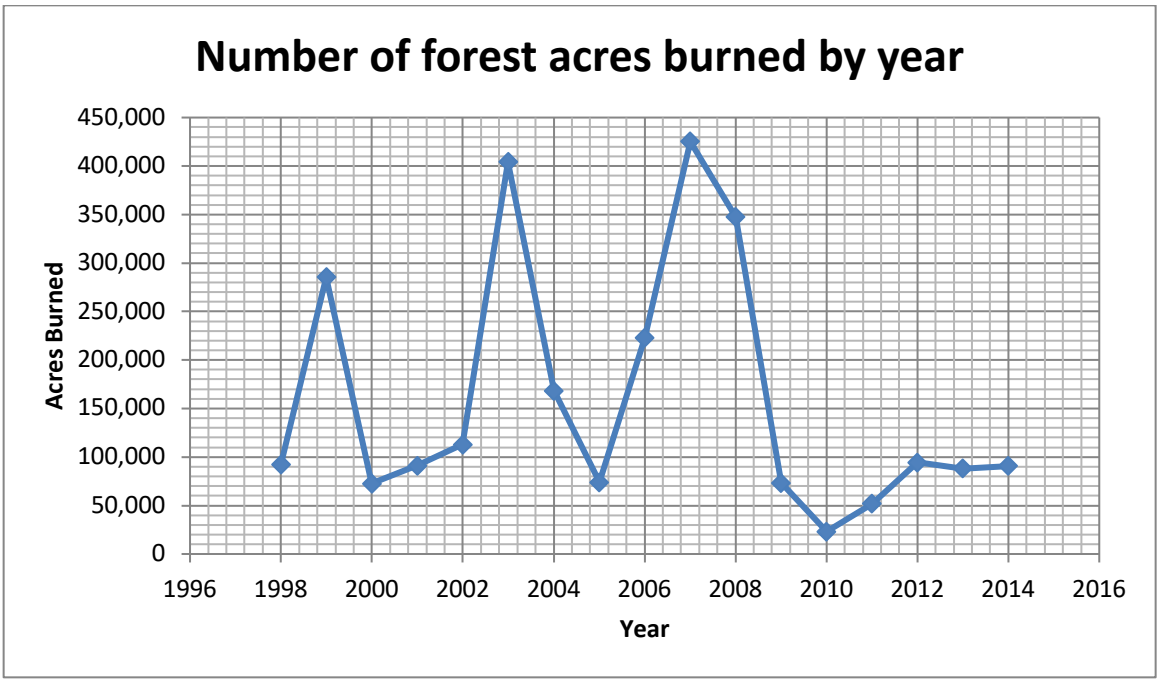


Figure 5.



Exercise 3: Make a time series graph showing the frequency of hurricanes in the eastern United States by year.

Section 3: Measures of Central Tendency

Once a researcher has gathered, organized, and pictured data, the researcher often wants to summarize the data by selecting a reading that best represents the whole group of data. That is, what lies at the center of these data? What do we mean by center? A statistician has three types of measures of the center of a collection of data: **mean, median,** and **mode**. The **mean**, or average of the data, is the number found by adding all the data readings together, and then dividing by the number of readings taken. For example, let suppose that a researcher wishes to record the average daily temperature for a particular day. The researcher decides to take the outdoor temperature reading six times in a day and average these temperatures together for the average daily temperature of the day. The researcher reads the thermometer at 2:00 AM, 6:00 AM, 10:00 AM, 2:00 PM, 6:00 PM, and 10:00 PM. (What type of sampling is this? Random, systematic, stratified, or cluster? Why might this be a good sampling technique for average daily temperature?) Suppose that the temperature readings at those times are given in the table below:

Table 4.

Time	Temperature
2 AM	68˚
6 AM	65˚

10 AM	75˚
2 PM	82˚
6 PM	80˚
10 PM	75˚

What would be the average daily temperature for that day?

$$\frac{68+65+75+82+80+75}{6}=\frac{445}{6}=74\frac{1}{6}\approx 74.1667$$

The researcher divides by 6 to find the average, since there are 6 readings taken. (*Exercise 4:* Contact NOAA or your local weather station and find out how they compute the average daily temperature. How many readings are taken?)

The median could vary widely from the average. Recall that the median is just the middle value when the data are written ascendingly. Average is the sum of all the values divided by their total number. The existence of a few very high or very low values will skew the average higher or lower. For example, according to US Census bureau, the average household income in 2014 was 75,738 while the median household income was 53,657. The average was higher because there are many very high income households. The median is lower means that there were lots of low income households. Half the population was making less than 53,657.

Another measure of central tendency is the mode. The mode is the reading that appears the most often. In the case of the daily temperature readings from *Table 4*, the mode for those data would be 75˚, since that is the only reading that appears more than once. (*Exercise 5:* Look at the data in *Table 1*, and try to determine the mode of those data. Can you think of any problems that might arise in the use of the mode as a measure of central tendency? Could be more than one – bimodal, trimodal, etc.,C) Mode has the advantage that it can be used to describe the central tendency of non-numerical data. For example, since green is the most frequent reading for the AQI codes in *Table 2,* that is the mode of those data.

If all the data are of same frequency there is no mode.

A third measure of central tendency is the median. The median is the middle value when all the values of a data set are arranged in order. For example, if the highest daily temperatures for a particular week were 81˚, 80˚, 95˚, 97˚, 88˚, 82˚, and 78˚, then when we arranged them in order, they would be

78˚ 80˚ 81˚ (82˚) 88˚ 95˚ 97˚

Since 82˚ is the middle value, it is the median high temperature for the week.

On the other hand, for the six readings in *Table 4*, there is no middle value, since there is an even number of readings. In this case, the median is found by taking the two middle readings, and finding their average.

65 68 75 75 80 82

Here, the two middle readings, 75 and 75, are the same, so their average is $\left(\frac{75+75}{2} = \frac{150}{2} = 75\right)$ again 75.

Exercise 6: Find the mean and median of the data in *Table 1.*

Exercise 7: Find the mean, median, and mode of the data you found in *Exercise 1*.

Sometimes data are given in categories. For example, suppose you read the following data from the California Department of Forestry and Fire Protection:

Table 5. Number of Fires by size, unit, and region – Northern Region 2012

Region	Total # fires	< 0.25 acres	0.26-9.99	10 to 99	100 to 299	300 to 999	1000 to 4999	5000 to 300,000
Almador-Eldorado	236	160	64	11	1	0	0	0
Butte	118	54	58	4	1	0	1	0
Humboldt-Del Norte	123	77	39	6	1	0	0	0
Lassen-Modoc	42	20	18	2	2	0	0	0
Mendocino	102	45	47	7	1	1	1	0
Nevada-Yuba-Placer	249	116	129	3	0	0	1	0
San Mateo-Santa Cruz	58	36	20	1	1	0	0	0
Santa Clara	84	34	42	6	2	0	0	0
Shasta-Trinity	193	58	125	5	4	0	1	0
Siskiyou	99	64	29	3	3	0	0	0
Sonoma-Lake-Napa	277	123	129	12	5	3	3	2
Tahama-Glen	82	33	37	8	2	1	0	1
Totals	1663	820	737	68	23	5	7	3

http://www.fire.ca.gov/downloads/redbooks/2012Redbook/2012_Redbook_Fires_bySize_byUnit_byCounty_CNR.pdf

We are not given the total exact number of acres burned for each fire, but only a range. How would we find the size of the average fire in northern California for this year, using this table? We treat each fire within a class as having a size equal to the midpoint of the class. Between 0 and 0.25, the midpoint is 0.125, between 0.26 and 9.99, the midpoint is $\frac{0.26+9.99}{2} = 5.125$, from 10 to 99, the midpoint is $\frac{10+99}{2} = 54.5$. The midpoints of the other categories are found similarly. They are 199.5, 649.5, 2999.5, and 152,500. Therefore, we would find the average area burned by wildfires in the Almador-Eldorado region by taking the 160 fires in the first category, and assuming that they all burned 0.125 acres, for a total of $160 \times 0.125 = 20$ acres. Similarly, if all the 64 fires in the second category burned 5.125 acres, they burned a total of $64 \times 5.125 = 328$ acres. In the third category, we have 11 fires each burning 54.5 acres, or $11 \times 54.5 = 599.5$ acres. In the fourth category , we have only 1 fire burning 199.5 acres. So the total acres estimated acres burned by these fires would be $20 + 328 + 599.5 + 199.5 =$

1147 acres. Since there are a total of 236 fires in the Almador-Eldorado region, the size of the average fire is estimated to be $\frac{1147}{236} \approx 4.86$ acres.

Exercise 8: Find the average size of the wildfires in each of the other regions of California.

Exercise 9: Use the totals at the bottom of *Table* 5 to find the average size of a wildfire in Nothern California for the year.

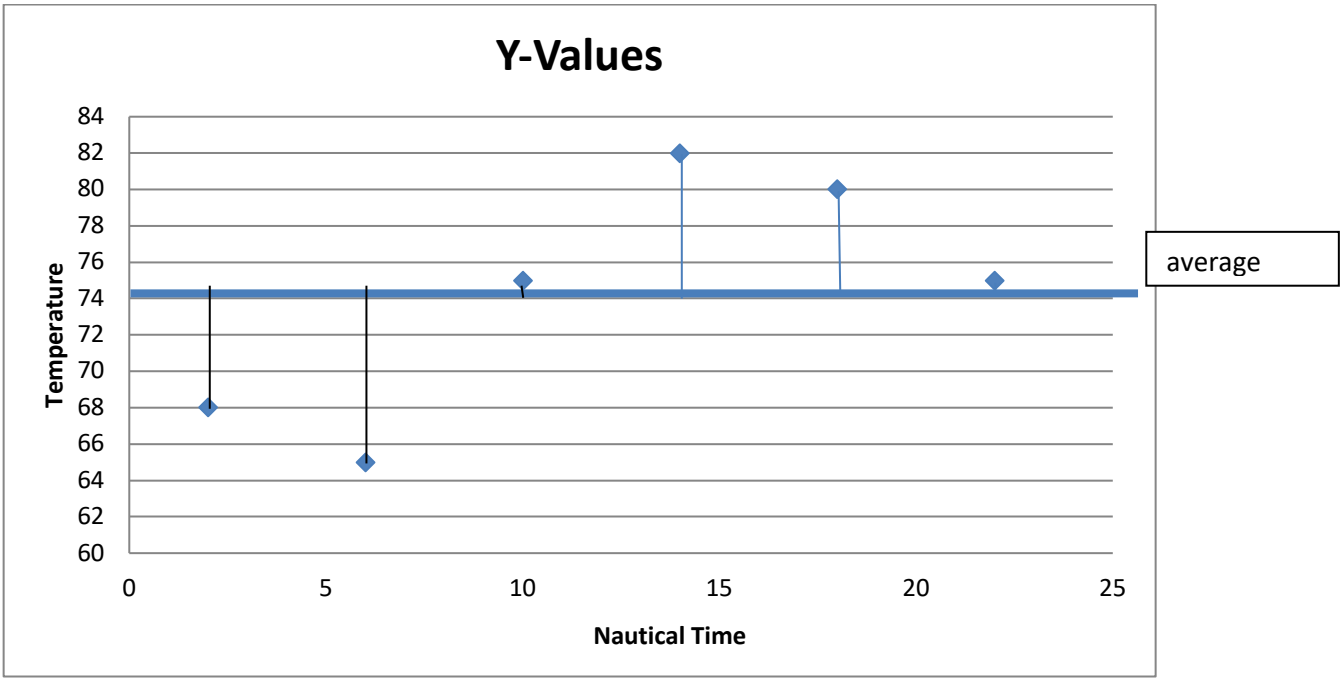
Section 4. Measures of Variation

When we want to know by how much the data in a set differ from one another, we are asking about the variation in the data. One measure of variation is the range. The **range** of a set of data is the difference between the highest and the lowest values in the set. For example, if we look again at Table 4, we find that the recorded temperatures are 65 68 75 75 80 82. The range in the temperatures taken for that day is $82 - 65 = 17$ degrees.

Two other measures of variation are **variance** and **standard deviation.** By variation, we mean how much do the readings differ from one another. Look at the temperatures in Table 4 again. Let's consider by how much each reading varies from the average. That is, let's find the difference between each temperature and the average temperature.

$$65 - 74\frac{1}{6} = -9\frac{1}{6}$$
$$68 - 74\frac{1}{6} = -6\frac{1}{6}$$
$$75 - 74\frac{1}{6} = \frac{5}{6}$$
$$75 - 74\frac{1}{6} = \frac{5}{6}$$
$$80 - 74\frac{1}{6} = 5\frac{5}{6}$$
$$82 - 74\frac{1}{6} = 7\frac{5}{6}$$

Figure 6.



Since we are interested in how much ALL the measurements differ from the average, let's add up all these differences:

$$-9\frac{1}{6}+-6\frac{1}{6}+\frac{5}{6}+\frac{5}{6}+5\frac{5}{6}+7\frac{5}{6}=0!!$$

Well, that does not give us a good idea of how spread out these numbers are, does it? In fact, this is not a coincidence. If we add the differences from the average for any set of numbers, we will always get zero. (Can you think why that would be true?) So we must find a way to use these numbers so that they do not cancel each other out. One way might be to add the absolute values of the differences. Another way is to square the differences and add their squares. In that case, we get

$$\left(-\frac{55}{6}\right)^2+\left(-\frac{37}{6}\right)^2+\left(\frac{5}{6}\right)^2+\left(\frac{5}{6}\right)^2+\left(\frac{35}{6}\right)^2+\left(\frac{47}{6}\right)^2=\frac{3025+1369+25+25+1225+2209}{36}$$
$$=\frac{7878}{36}=\frac{1313}{6}\approx 218.8333$$

(Here we changed all the mixed numbers to fractions for easier computation.) The only trouble with this number is that the more readings we take, the larger the number gets. It does not give us a good idea of how close together all the numbers may be. If every reading differed from the average by only ± 0.01, but there were thousands of readings, this sum could still be very large. Therefore, we divide by one less than the number of readings. (Why one less? It has to do with the fact that we are taking a sample of all the readings, rather than the entire group (or population) of all temperatures at every second of every day.)

So we divide the sum by one less than the number of readings:

$$\frac{1313}{6} \div 5 = \frac{1313}{30} \approx 43.7$$

This number is called the **variance** of the data. But we squared all those original differences, didn't we? So, it might make sense to take the square root of this sum:

$$\sqrt{\frac{1313}{30}} \approx 6.6156$$

This number, the square root of the variance, is called the **standard deviation** of the data, and it is used whenever statistics are employed to describe the spread within a data set.

To illustrate, let us consider the following sea level data, which is measured hourly. These data come from Darwin, Australia on the first of January, 2015.

Table 6.

Date & UTC Time	Sea Level
1/1/2015 0:00	2.323
1/1/2015 1:00	2.551
1/1/2015 2:00	3.152
1/1/2015 3:00	3.977
1/1/2015 4:00	4.823
1/1/2015 5:00	5.523
1/1/2015 6:00	5.89
1/1/2015 7:00	5.85
1/1/2015 8:00	5.457
1/1/2015 9:00	4.878
1/1/2015 10:00	4.305
1/1/2015 11:00	3.875
1/1/2015 12:00	3.757
1/1/2015 13:00	3.968

1/1/2015 14:00	4.407
1/1/2015 15:00	4.921
1/1/2015 16:00	5.367
1/1/2015 17:00	5.693
1/1/2015 18:00	5.817
1/1/2015 19:00	5.514
1/1/2015 20:00	4.765
1/1/2015 21:00	3.858
1/1/2015 22:00	3.115
1/1/2015 23:00	2.536

http://www.bom.gov.au/oceanography/projects/abslmp/data/

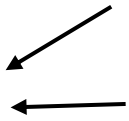
The measurements are taken in meters above tide gauge zero. Let us find the range, variance, and standard deviation of these data. Since the highest level measured that day is 5.89, and the lowest level is 2.323, the range is $5.89 - 2.323 = 3.567$

To find the variance and standard deviation, we must first find the average of these readings. So we add the readings, to find a sum of 106.322. Then we divide by the number of readings (i.e., 24), to find an average of 4.430083. Since we have several steps to complete to find variance and standard deviation, let's put all our steps into the table, so we can keep track.

Table 7.

Date & UTC Time	Sea Level	X - av	$(X - av)^2$
1/1/2015 0:00	2.323	-2.10708	4.4398
1/1/2015 1:00	2.551	-1.87908	3.530954
1/1/2015 2:00	3.152	-1.27808	1.633497
1/1/2015 3:00	3.977	-0.45308	0.205285
1/1/2015 4:00	4.823	0.392917	0.154384
1/1/2015 5:00	5.523	1.092917	1.194467
1/1/2015 6:00	5.89	1.459917	2.131357
1/1/2015 7:00	5.85	1.419917	2.016163
1/1/2015 8:00	5.457	1.026917	1.054558
1/1/2015 9:00	4.878	0.447917	0.200629
1/1/2015 10:00	4.305	-0.12508	0.015646
1/1/2015 11:00	3.875	-0.55508	0.308118
1/1/2015 12:00	3.757	-0.67308	0.453041
1/1/2015 13:00	3.968	-0.46208	0.213521
1/1/2015 14:00	4.407	-0.02308	0.000533
1/1/2015 15:00	4.921	0.490917	0.240999

1/1/2015 16:00	5.367	0.936917	0.877813
1/1/2015 17:00	5.693	1.262917	1.594959
1/1/2015 18:00	5.817	1.386917	1.923538
1/1/2015 19:00	5.514	1.083917	1.174875
1/1/2015 20:00	4.765	0.334917	0.112169
1/1/2015 21:00	3.858	-0.57208	0.327279
1/1/2015 22:00	3.115	-1.31508	1.729444
1/1/2015 23:00	2.536	-1.89408	3.587552
sum	106.322		29.12058
average	4.430083	div by 23	1.266112
		Square rt	1.125216



At the bottom of the second column, we have listed the sum and the average of the daily readings. In the third column, we have subtracted the average reading (4.430083) from each data point. In the fourth column, the squares of each of these differences have been listed. At the bottom of the fourth column, the sum of the squares has been computed as 29.12058. When we divide that number by one less than the number of data points (i.e., 23), we find the variance, 1.266112. To find the standard deviation, we then take the square root of the variance.

Exercise 10: Find the range, variance, and standard deviation for these sea level measurements from Darwin, Australia on March 22, 2015:

Table 8.

3/22/2015 0:00	6.104
3/22/2015 1:00	4.189
3/22/2015 2:00	2.563
3/22/2015 3:00	1.592
3/22/2015 4:00	1.153
3/22/2015 5:00	1.46
3/22/2015 6:00	2.644
3/22/2015 7:00	4.45
3/22/2015 8:00	6.201
3/22/2015 9:00	7.345
3/22/2015 10:00	7.861
3/22/2015 11:00	7.576
3/22/2015 12:00	6.309
3/22/2015 13:00	4.412
3/22/2015 14:00	2.649
3/22/2015 15:00	1.475
3/22/2015 16:00	0.769
3/22/2015 17:00	0.67

variance

Standard deviation

3/22/2015 18:00	1.522
3/22/2015 19:00	3.178
3/22/2015 20:00	5.077
3/22/2015 21:00	6.628
3/22/2015 22:00	7.522
3/22/2015 23:00	7.721

http://www.bom.gov.au/oceanography/projects/abslmp/data/

Section 5. Measures of Position

Recall that when we wanted to find the median of a set of data, it was expedient to list the data in order. The median value is also called the 50th **percentile** of the data. The percentile value of a particular data point is the percentage of data values that lie at or below that point.

Suppose that we take 10 daily temperature readings for a day, and we have the following:

65, 62, 70, 71, 75, 80, 85, 87, 76, 68

If we arrange these values in order, we have

62, 65, 68, 70, 71, 75, 76, 80, 85, 87

What percentile ranking is 71? Since there are four (out of 10) temperature readings at or below 71, 71 represents $\frac{4}{10} = 40\%$, or the 40th percentile.

Exercise 11: Find the percentile rank of the 80 degree temperature reading.

If you had a 65th percentile score on your SATs, you did better than 65% of the students taking the test.

The average daily temperatures for Washington DC in the month of July 2015 (according to Weather Underground), listed in order are given below:

Table 9.

order	av. Temp.
1	73
2	75
3	77
4	78
5	78
6	79
7	79
8	79
9	79
10	80
11	80
12	80
13	81
14	81
15	82
16	82
17	83
18	83
19	83
20	83
21	83
22	83
23	83
24	85
25	85
26	85
27	85
28	86
29	86
30	89
31	90

http://www.wunderground.com/history/airport/KDCA/2015/8/14/DailyHistory.html

Here, we see (unlike the previous list), there are many values that are repeated. For example, we see that 79 is listed as the 6th, 7th, 8th, and 9th temperatures on this list. How can we assign a percentile to this value? If we call it the 6th lowest value, then we would call it the $\frac{6}{31} \approx 19\%$, or 19th percentile; if we call it the 9th lowest temperature, it would be the $\frac{9}{31} \approx 29\%$, or 29th percentile. Instead, we "split the difference." That is, we count the number 79 as having $5+\frac{1}{2} \cdot 4 = 7$ scores "at or

below" 79. We count all 5 of the scores that are actually below 79, but only half of those that are equal to 79. Therefore, 79 would be equal to the $\frac{7}{31} \approx 23\%$, 23rd percentile.

Exercise 12: What percentile value do you assign to the 85 degree temperature reading?

Exercise 13: One incoming Howard University freshman was ranked 10th in a class of 125; another ranked 75th in a class of 620. Which has the higher percentile rank?

A concept related to percentile rank is the **quartile.** This divides the data into quarters. The second quartile is the same as the median; the first quartile is the median of the lower half of the data, and the third quartile is the median for the upper have of the data.

Look again at the data in Table 9. The median value (value #16) is 82. The median of the first 15 values is #8, which is 79, so that is the first quartile. The median of the last 15 values is #23, or 83 degrees. That is the third quartile.

Section 6. Normal Distribution

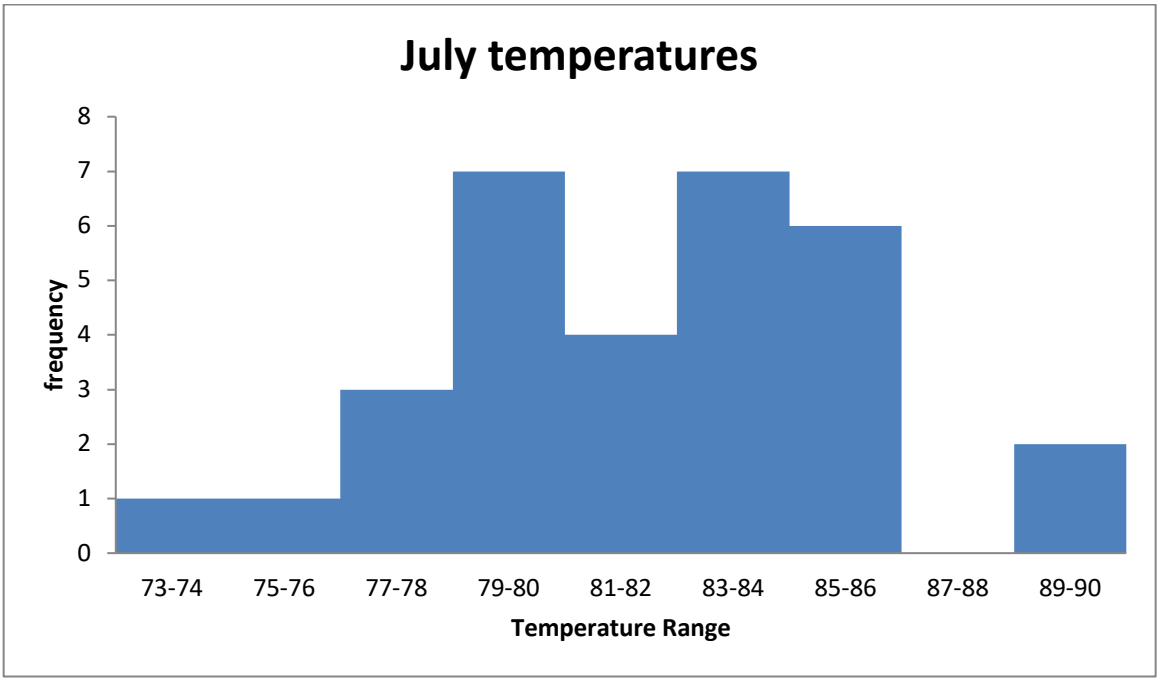
Let's look again at the data in Table 9. We want to break the data into classes and make a frequency distribution. Suppose we want about 10 classes. Since the low temperature is 73 and the high is 90, there is a range of 17 degrees. To have 10 classes, we should have about two degrees in a class.

Table 10.

Class	Count	Frequency
73 – 74	|	1
75 – 76	|	1
77 – 78	|||	3
79 – 80	𝍸 ||	7
81 – 82	||||	4
83 – 84	𝍸 ||	7
85 – 86	𝍸 |	6
87 – 88		0
89 – 90	||	2

We see that we will only have 9 classes, but we will make do. We draw a bar graph.

Figure 7.



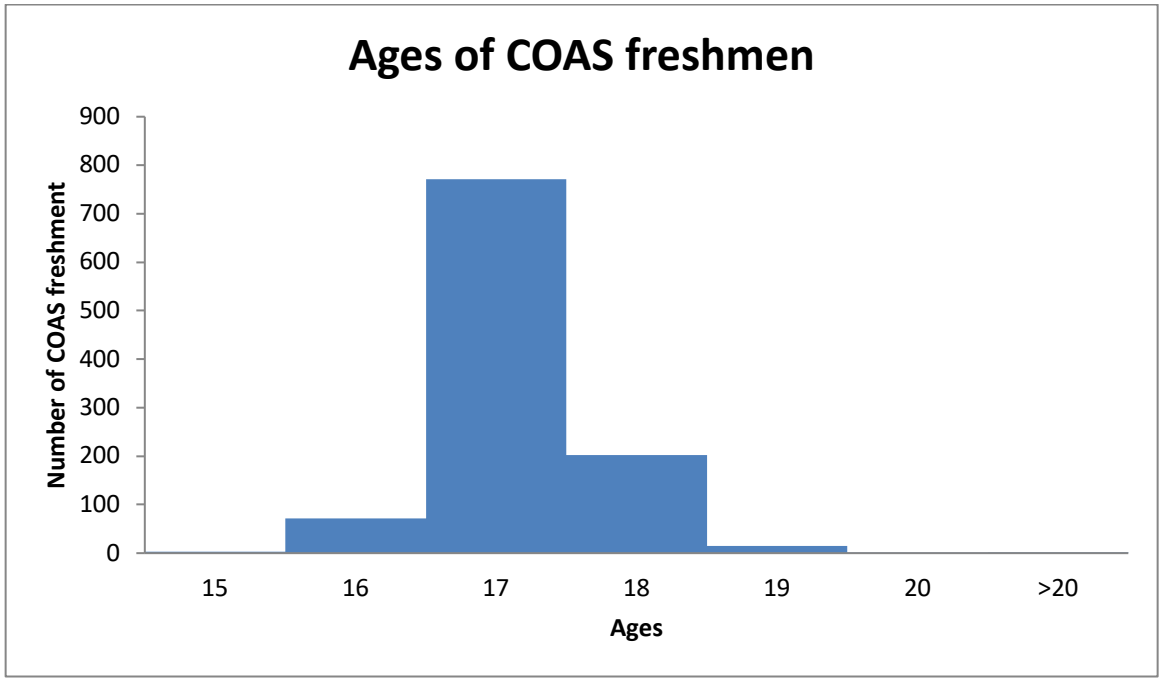
We see that the frequencies of the middle ranges tend to be higher, and those to the far left and right have lower frequencies. This is a very common pattern in data values, and it tends to become more pronounced as more data points are recorded.

For example, here is a list of the ages of incoming freshmen in the College of Arts and Sciences at Howard University in 2013.

Table 11.

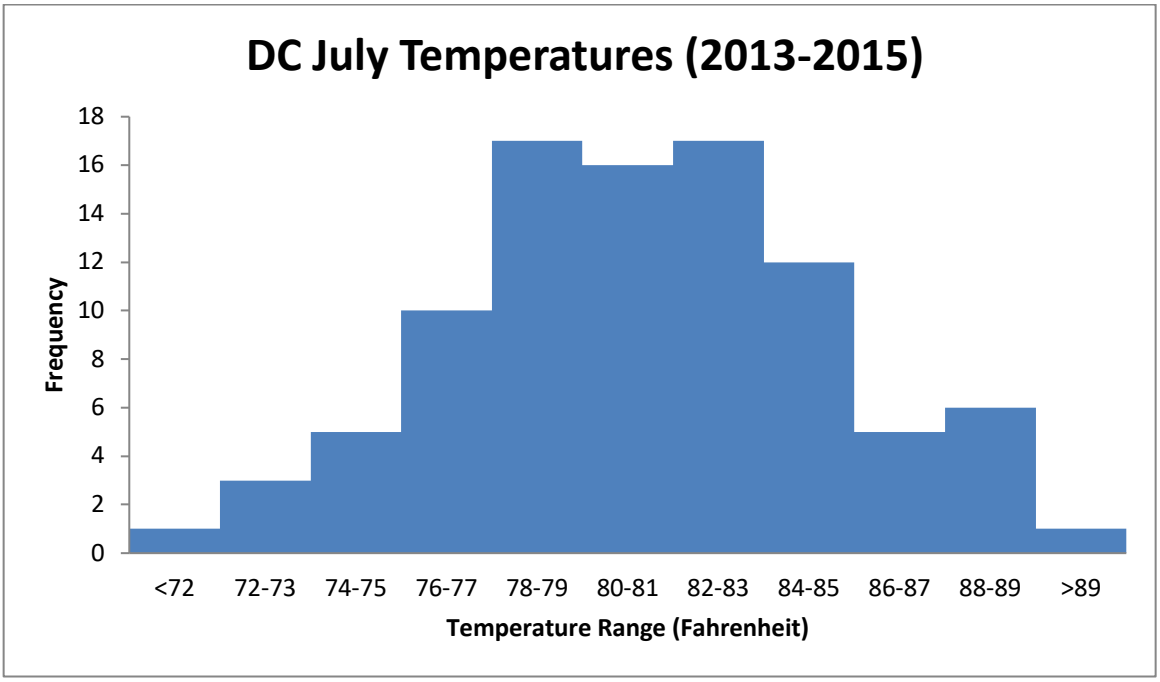
age	# fresh
15	3
16	72
17	771
18	202
19	15
20	1
>20	2

Figure 8.

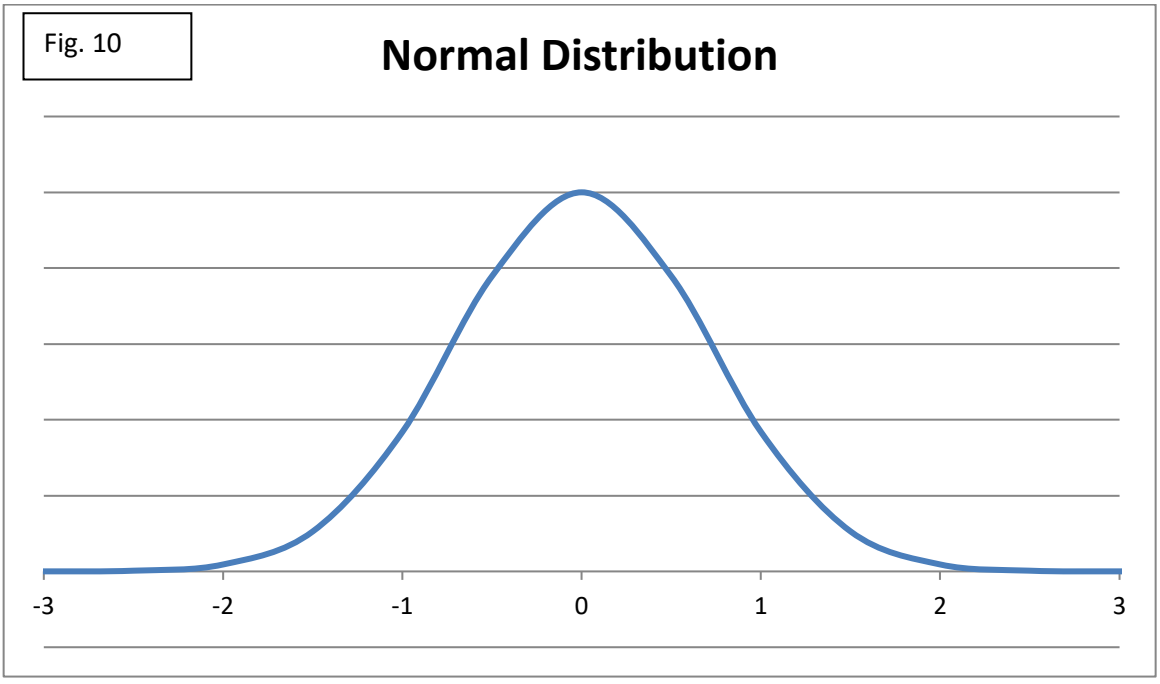


As we increase the sample size and decrease the class widths, the distribution looks more and more bell-shaped.

Figure 9.



As the sample size increases, and the class width decreases, the shape becomes more and more bell-shaped.



Some properties of the normal distribution are as follows:

1. It is bell-shaped.
2. It is continuous (no gaps, holes, or jumps)
3. It has only one mode.

4. The mean, median, and mode are all the same.
5. The function is symmetric about the mean.
6. The total area under the curve is 1.
7. The fraction of data that fall between two values is equal to the area under the curve between the two values.

Empirical Rule: When data are normally distributed, about 68% of the values are within 1 standard deviation of the mean, about 95% within 2 standard deviations of the mean, and about 99.7 are within 3 standard deviations of the mean.

A normal distribution is called the **standard normal distribution** if its mean is zero and the standard deviation is one. Many books contain tables showing the area under the curve of the standard normal distribution between zero and any given number. (And your graphing calculator will compute this value!) We can find the area under any other normal distribution using a **z-score**, which converts between a general normal distribution and the standard normal distribution. The z-score is computed by the formula

$$\boldsymbol{z} = \frac{\text{data value -mean}}{\text{standard deviation}}$$

Exercise 14: Our data set of July temperatures have a mean 80.96 and standard deviation 4.15. If we have a normal distribution with that mean and that standard deviation, what would be the z-score of 82?

Suppose again we have a normal distribution of July temperatures with mean 80.96 and standard deviation 4.15, and we want to find the percentage of our temperature data between 78 and 84 degrees. We need to find the z-score of each temperature.

$$z_1 = \frac{78 - 80.96}{4.15} = -0.71325$$
$$z_2 = \frac{84 - 80.96}{4.15} = 0.73253$$

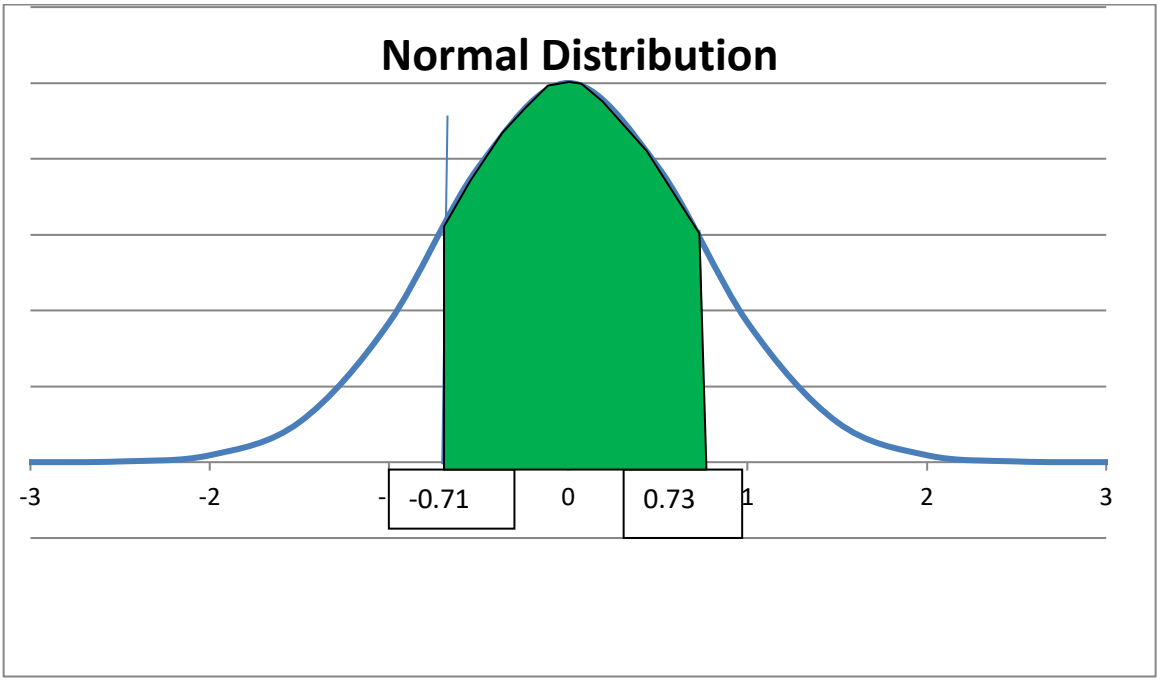
Now the area under the standard normal distribution between zero and 0.73 is given in the following table:
http://images.tutorvista.com/cms/images/67/full-Z-score-table.PNG

Table 12

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0	0.004	0.008	0.012	0.016	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.091	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.148	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.17	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.195	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.219	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.258	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.291	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.334	0.3365	0.3389
1	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.377	0.379	0.381	0.383
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.398	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.437	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.475	0.4756	0.4761	0.4767
2	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.483	0.4834	0.4838	0.4842	0.4846	0.485	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.489
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.492	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.494	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4855	0.4956	0.4957	0.4959	0.496	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.497	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.498	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.499	0.499

Look across the row labeled 0.7 and under the column 0.03. We see 0.2673. That is the area under the standard normal distribution between 0 and 0.73. We have the same area under the curve on the negative side, so between zero and -0.71, the area is 0.2611. So the area under the graph between -0.71 and 0.73 is the area on the left of zero plus the area on the right of zero.

Fig 11

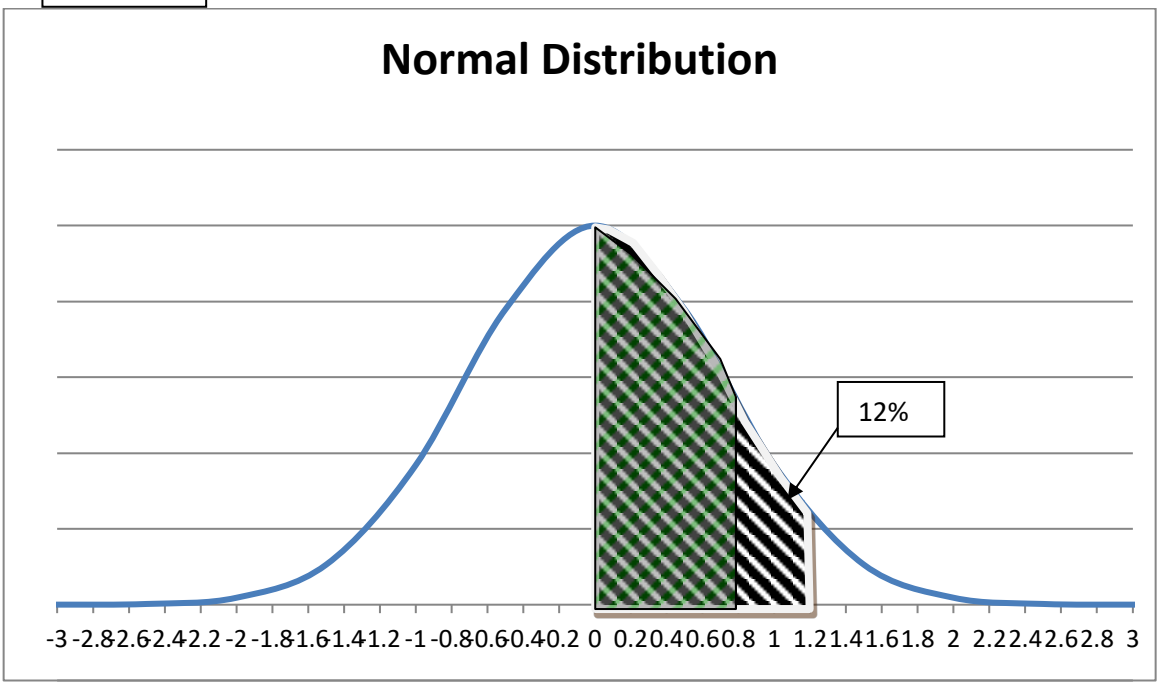


$$0.2673 + 0.2611 = 0.5284$$

So about 53% of the data are between 78 and 84 degrees.

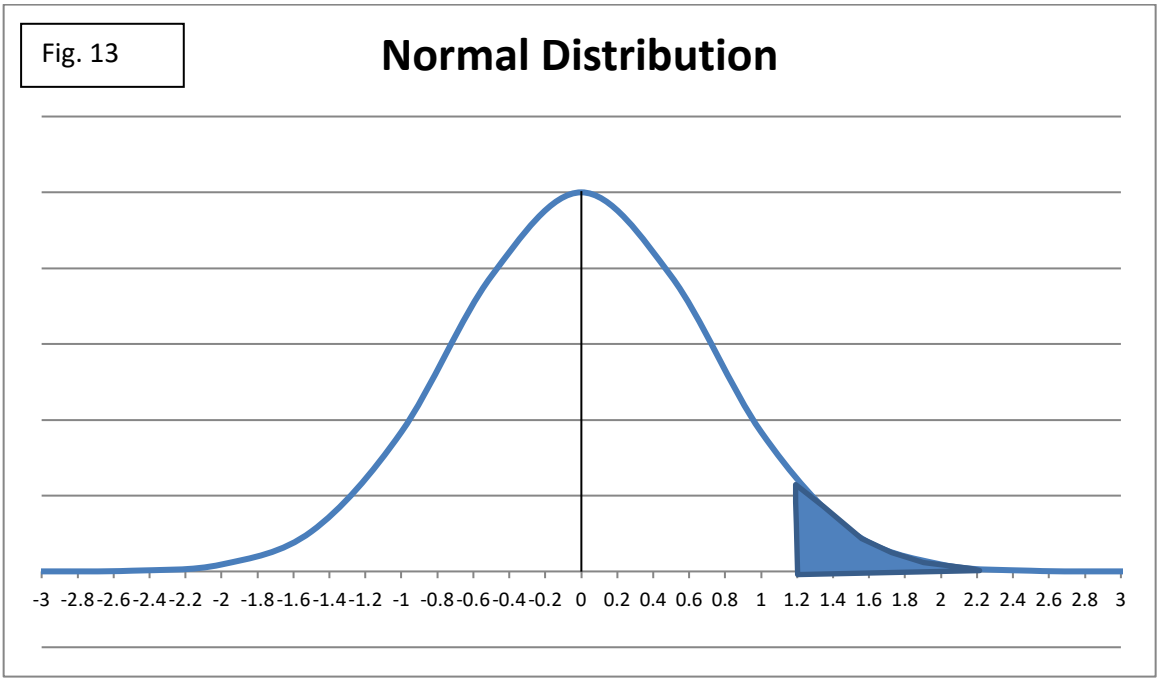
If our two z-scores were on the same side of zero, we would have to *subtract* the corresponding area values, since we would be counting the area between zero and the lower value twice. For example, suppose we want to find the percentage of temperature readings between 84 and 86. The z-score for 86 is $z = \frac{86-80.96}{4.15} = 1.214$. The associated area is (look across from 1.2 and under the column 0.01) 0.3869. We already found that the z-score for 84 is 0.73253. The area between 0 and 0.73 under the standard normal distribution curve is 0.2673. (Look on the 0.7 row, under the 0.03 column.) Since both lie on the same side of zero, we have to subtract the two areas to find the area between them. The area between these two values is $0.3869 - 0.2673 = .1196$, or about 12%.

Fig. 12



If we wanted to find the percentage of temperatures above 86°, we would need to find the area in a "tail" of the curve:

Fig. 13



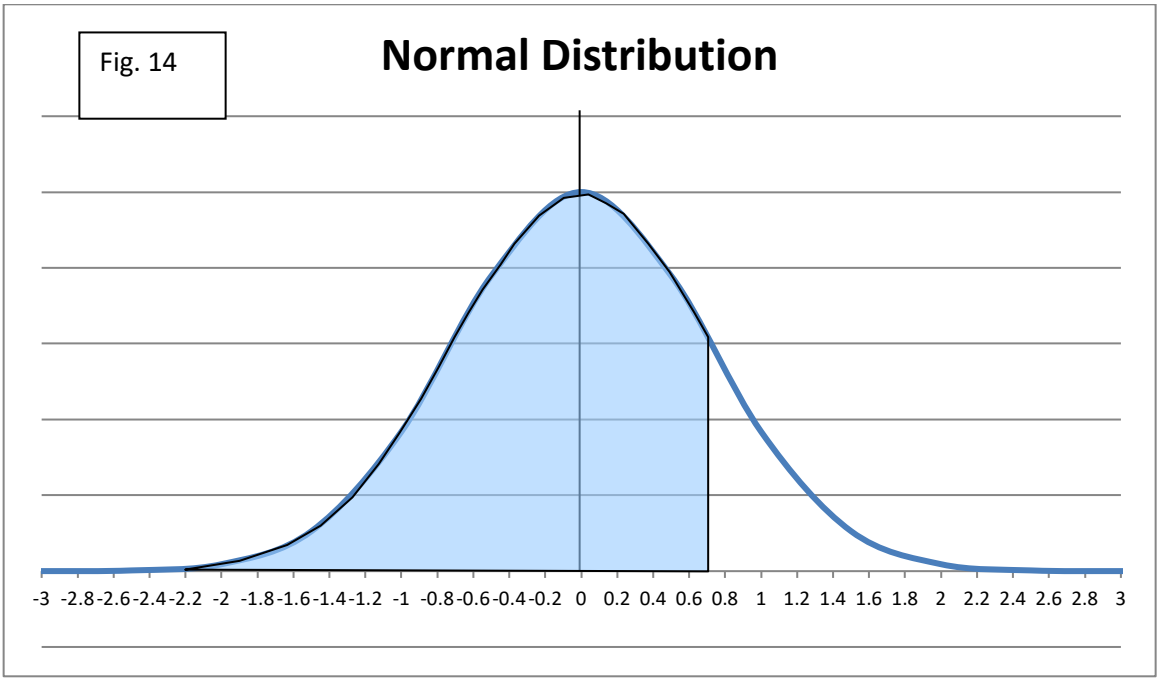
We already know that the z-score of 86° is 1.214, so on the standard normal distribution, we would be looking for how many values lie above 1.214, or the area of the shaded portion of the graph. Now, we know that the table of areas for associated z-values gives us the area under the graph between zero and the z-value. We also know that the area under the whole graph is equal to 1. (This is because if we add up all the values in the data, we get 100% of the values. 1=100%) We know that the graph is symmetric

about 0. That means that half of the values (or 50%) lie to the right, and half lie to the left. So the total area on the right-hand side of the graph, from zero on, is 0.5. We know that between 0 and 1.21, the area is 0.3869. So to find the area in the tail, we subtract 0.3869 from the total right hand side, .5:

$$0.5000 - 0.3869 = .1131$$

We find that the percentage of temperatures that fall above 86° is about 11%

What percentage of our temperatures are less than 84°? To answer this question, we want to find the shaded area in the graph below:



The area given in the table is from 0 to .73. So we need to add to that area to get all the extra area to the left of .73. Since the left half of the graph has area .5 (just as the right half does), we add the two areas: $0.5 + 0.2673 = 0.7673$, so about 77% of the temperatures are below 84°.

Exercise 15:

According to the statistical abstract of the United States (https://catalog.data.gov/dataset/statistical-abstract-of-the-united-states), in 2008, the U.S. consumed an average of 327 million BTUs per capita in energy. (http://www.census.gov/prod/2011pubs/12statab/energy.pdf) A BTU, or british thermal unit, is a traditional unit of work equal to about 1055 joules (or newton-meters). Suppose that the consumption of energy is normally distributed among the U.S. population, and that the standard deviation is 30 BTUs.

(a) Estimate the percentage of the U.S. population who are responsible for consuming more than 400 BTUs per year.

(b) Estimate the percentage of the U.S. population who consume between 320 and 350 BTUs per year.

Section 8: Correlation and Regression Analysis

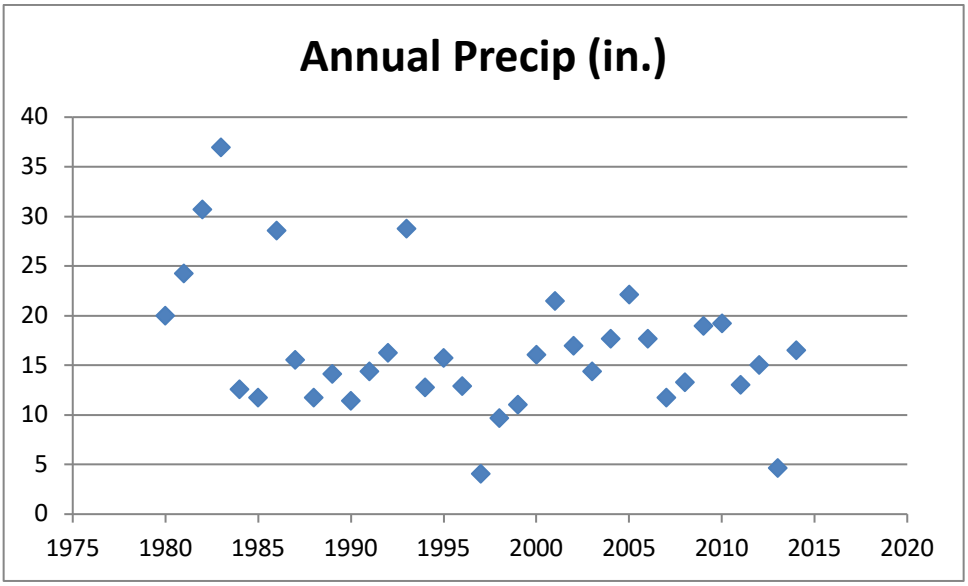
In our projects on environmental phenomena, we will research collected data and try to predict future data based on the patterns of the past. Suppose that a student finds these values for the annual rainfall in Sacramento, California (http://www.wunderground.com/history/airport/KSAC/2009/5/10/DailyHistory.html):
Table 13

Year	Annual Precip (in.)
1980	20.02
1981	24.29
1982	30.72
1983	36.95
1984	12.59
1985	11.75
1986	28.61
1987	15.55
1988	11.78
1989	14.16
1990	11.41
1991	14.37
1992	16.28
1993	28.77
1994	12.81
1995	15.78
1996	12.94
1997	4.07
1998	9.71
1999	11.04
2000	16.08
2001	21.47
2002	16.95
2003	14.37
2004	17.67
2005	22.11
2006	17.69
2007	11.73
2008	13.32
2009	19.01
2010	19.27
2011	13.07
2012	15.05

2013	4.65
2014	16.51

The student then makes a “scatter plot” of the data, graphing the year against the annual precipitation for that year.

Fig. 15



We can see that most of the very high values are in the earlier years, and more recent values seem lower, but we don’t see a very obvious pattern.

Below is a table of my Algebra 1 students’ scores on their homework program (ALEKS) versus their scores on the final exam (out of 200) for the fall of 2014.

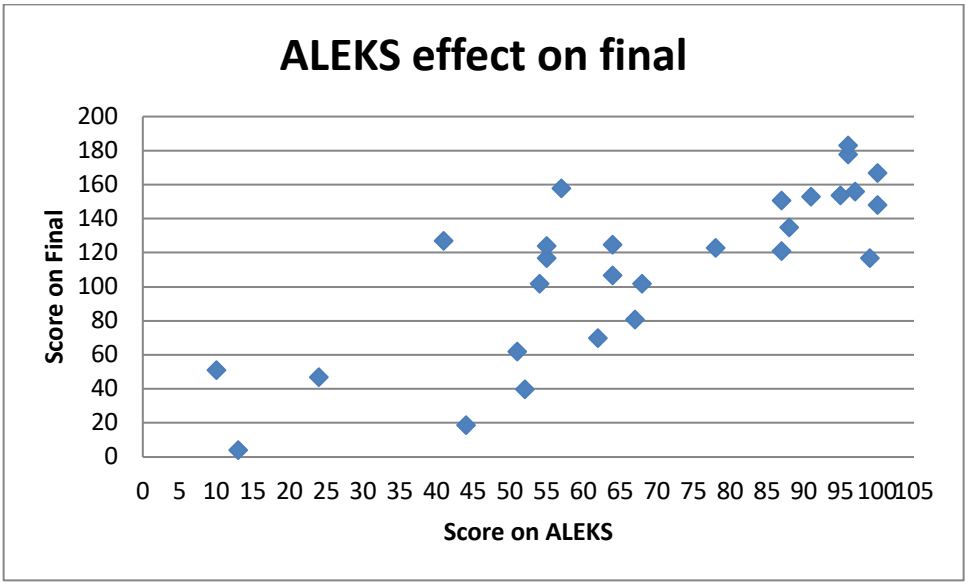
Table 14

ALEKS	Final
10	51
100	167
44	19
87	121
68	102
52	40
96	178
100	148
41	127
55	117
64	107
55	124

95	154
24	47
96	183
13	4
97	156
62	70
78	123
99	117
54	102
67	81
57	158
87	151
88	135
91	153
64	125
51	62

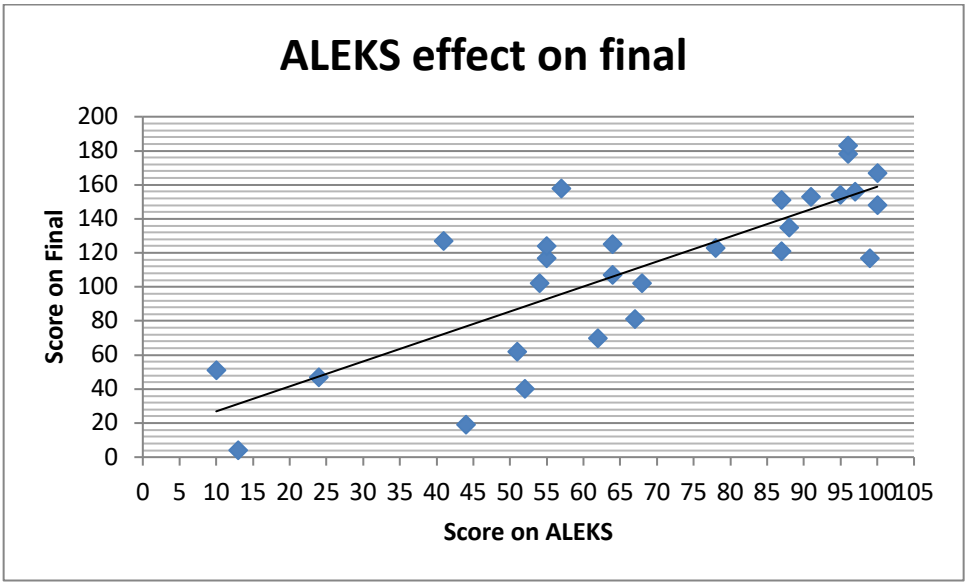
Here is a scatter plot of the same data:

Fig. 16



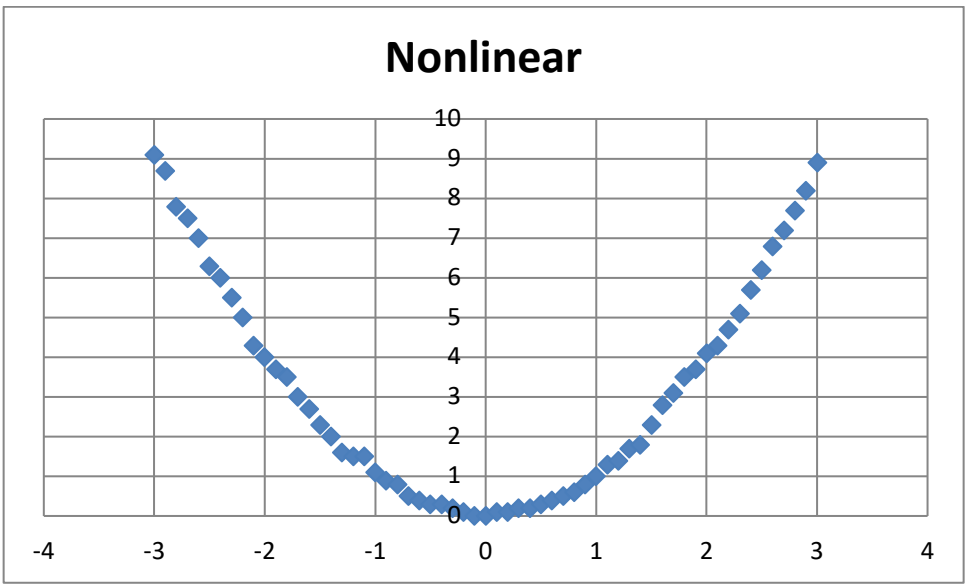
In this graph, we see a much more pronounced upward climb in the student's scores on the final as their score on the ALEKS program rose. Sometimes we want to quantify the relationship between two variables in a data set. We can estimate what line would fall the closest to most points in the data set.

Fig. 17



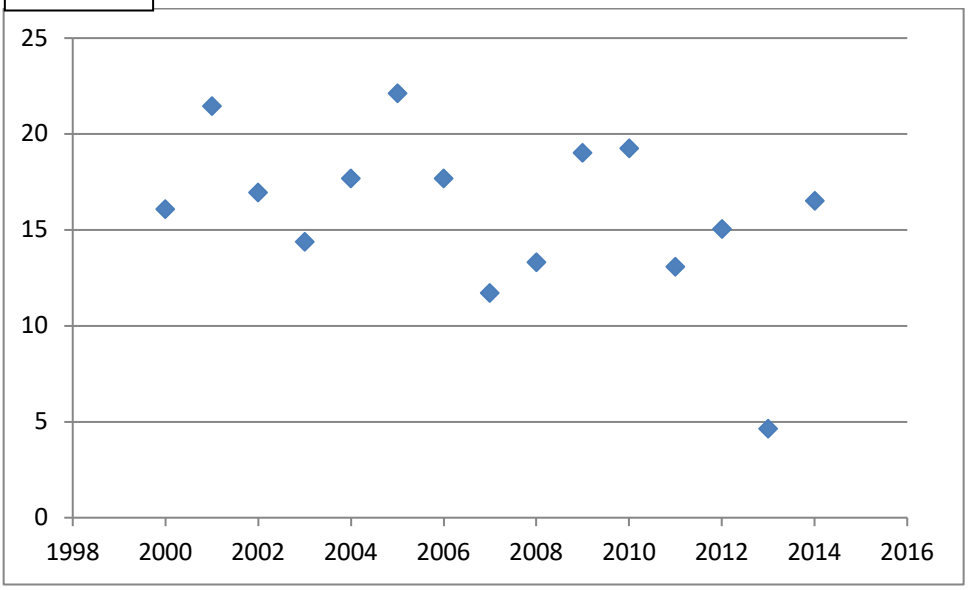
We see that the line has a positive slope. We say that the data have a **positive linear correlation**. A line with a negative slope would indicate a **negative linear correlation**. Data can also have a nonlinear correlation:

Fig. 18



Some data have no correlation. If we take just the last 15 years of the precipitation data, there does not seem to be any particular pattern.

Fig. 19



There is a coefficient that describes how closely two data variables are related. It is called the **correlation coefficient, *r*.** The formula for the correlation coefficient is

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

where $n = \text{number of data pairs}$, and $\sum w$ means that you add up all the *w*'s. So $\sum x$ means the sum of all n *x*'s, $\sum y$ means the sum of all *n y*'s, etc. And $\sum x^2$ means you add the squares of the *x*'s, but $(\sum x)^2$ means that you add up all the *x*'s, and then you square the answer.

If $r = 1$, then the variables are perfectly related by a linear equation (that is, $y = mx + b$ for some positive slope *m* and some *y*-intercept *b.*) If $r = -1$, then the variables are again perfectly related by a linear equation, but this time the slope would be negative. If the variables are not even approximately linearly related, then r is close to zero. The correlation coefficient for the data in Figures 18 and 19 would be close to zero.

When we use Excel to make graphs for our projects, we can ask it to compute r^2 for us. This is just the square of the correlation coefficient, and it gets closer to 1 as the correlation becomes more approximately linear, whether the slope is positive or negative. If r^2 is close to zero, again, there is not a linear correlation.

Let us look again at my data for students' ALEKS scores versus their final exam scores. We will use these data and compute the correlation coefficient.

Table 15

ALEKS (x)	Final (y)	xy	x^2	y^2
10	51	510	100	2601
100	167	16700	10000	27889
44	19	836	1936	361
87	121	10527	7569	14641
68	102	6936	4624	10404
52	40	2080	2704	1600
96	178	17088	9216	31684
100	148	14800	10000	21904
41	127	5207	1681	16129
55	117	6435	3025	13689
64	107	6848	4096	11449
55	124	6820	3025	15376
95	154	14630	9025	23716
24	47	1128	576	2209
96	183	17568	9216	33489
13	4	52	169	16
97	156	15132	9409	24336
62	70	4340	3844	4900
78	123	9594	6084	15129
99	117	11583	9801	13689
54	102	5508	2916	10404
67	81	5427	4489	6561
57	158	9006	3249	24964
87	151	13137	7569	22801
88	135	11880	7744	18225
91	153	13923	8281	23409
64	125	8000	4096	15625
51	62	3162	2601	3844
$\sum x$	$\sum y$	$\sum xy$	$\sum x^2$	$\sum y^2$
1895	3122	238857	147045	411044

We then use these values to compute *r*. We see that there are 28 pairs of values, so $n = 28$. Now we compute *r*.

$$\frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}} =$$

$$\frac{28(238857) - (1895)(3122)}{\sqrt{[28(147{,}045) - 1895^2][28(411044) - 3122^2]}} \approx$$

0.8014

We see that the correlation coefficient is about 0.8. This seems "close to" 1, but is it close enough to say that there is a correlation? For this question, we turn to a table for significance for the correlation coefficient.

http://www.gifted.uconn.edu/siegle/research/correlation/corrchrt.htm

Table 16

df = *n* -2	*n*				
Level of Significance (p) for Two-Tailed Test		**0.1**	**0.05**	**0.02**	**0.01**
df					
1	3	0.988	0.997	0.9995	0.9999
2	4	0.9	0.95	0.98	0.99
3	5	0.805	0.878	0.934	0.959
4	6	0.729	0.811	0.882	0.917
5	7	0.669	0.754	0.833	0.874
6	8	0.622	0.707	0.789	0.834
7	9	0.582	0.666	0.75	0.798
8	10	0.549	0.632	0.716	0.765
9	11	0.521	0.602	0.685	0.735
10	12	0.497	0.576	0.658	0.708
11	13	0.476	0.553	0.634	0.684
12	14	0.458	0.532	0.612	0.661
13	15	0.441	0.514	0.592	0.641
14	16	0.426	0.497	0.574	0.623
15	17	0.412	0.482	0.558	0.606
16	18	0.4	0.468	0.542	0.59
17	19	0.389	0.456	0.528	0.575
18	20	0.378	0.444	0.516	0.561
19	21	0.369	0.433	0.503	0.549
20	22	0.36	0.423	0.492	0.537
21	23	0.352	0.413	0.482	0.526

22	24	0.344	0.404	0.472	0.515
23	25	0.337	0.396	0.462	0.505
24	26	0.33	0.388	0.453	0.496
25	27	0.323	0.381	0.445	0.487
26	28	0.317	0.374	0.437	0.479
27	29	0.311	0.367	0.43	0.471
28	30	0.306	0.361	0.423	0.463
29	31	0.301	0.355	0.416	0.456
30	32	0.296	0.349	0.409	0.449
35	37	0.275	0.325	0.381	0.418
40	42	0.257	0.304	0.358	0.393
45	47	0.243	0.288	0.338	0.372
50	52	0.231	0.273	0.322	0.354
60	62	0.211	0.25	0.295	0.325
70	72	0.195	0.232	0.274	0.303
80	82	0.183	0.217	0.256	0.283
90	92	0.173	0.205	0.242	0.267
100	102	0.164	0.195	0.23	0.254

A 0.1 significance means that there is a 10% chance that the experimenter is wrong, and there is no correlation between the two variables. In general, researchers look for at least 5% significance level. If $|r|$ is greater than or equal to the value in the table under 0.05 and across from the appropriate "n" value, then we can be 95% sure that there is a correlation between the variables. In the case of the ALEKS data, when $n = 28$, we see that the 0.05 significance level is 0.374. Since our value for r is 0.8, we can be 95% sure that there is a correlation between the variables.