Statistics are around us both seen and in ways that affect our lives without us knowing it. We have seen data organized into charts in magazines, books and newspapers. That's descriptive statistics! We can also measure the spread of the data set or the data set's most "typical" value. Comparing pairs of related values helps us in prediction. Understanding the type and strength of the relation between two different variables is the first step.

Course Outcomes:

- Efficiently use relevant technology
- Demonstrate proficiency in basic concepts and procedures related to descriptive statistics
- Calculate probabilities, and use and apply the normal distribution

8.1 Frequency Distribution, Frequency Polygon, Histogram

Data is organized into frequency distributions. Students learn to present the results of the frequency distributions as frequency polygons and histograms.

8.2 Measure of Central Tendency

Measures of central tendency all arrive at one value to represent the whole data set. Several different measures of central tendency are found. The most important measures of central tendency are mean, median, and mode.

8.3 Measures of Dispersion

The spread of the data is measured and represented with a single value. Standard deviation measures the spread of the data around the mean. Range is the difference between the largest and the smallest data values.

8.4 Percentiles and Normal Distribution

Percentiles are defined in order to use the standard normal distribution chart. Percentages and probabilities are found using the normal distribution.

8.5 Correlation

The relationship between two variables is examined. Students learn scatter plots and types of correlation – positive, negative, and no correlation. The strength of the correlation is introduced.

When we ask the question what are statistics, we may come up with various answers. One common notion is the various facts that are involved in a situation. For example, we may talk about the technical capabilities of a computer: amount of memory, speed of the central processing unit, power input, etc. Here we are really describing the computer.

There are two types of statistics:
1. Descriptive statistics are used to collect, organize and present data
2. Inferential statistics uses information about a sample to say something about a population.

The population is all the values of interest. A sample is a subset of the population. The connection between a sample and a population is based on probabilities. For example, we may look at how 100 students feel about the new educational resources to find out how all students feel about the new educational resources. In the class *Introductory Statistics* we study descriptive statistics, probabilities, and inferential statistics.

If we study 100 students that take MATH 103, the 100 students is the sample. All students taking MATH 103 is the population.
If observe 25 cars that come out of a factory. The population is all the cars that come out of the factory. The sample is the 25 cars that we observe.

Data refers to a set of observations or possible outcomes. There are two types of data:

1. Qualitative: non-numerical data which refers to a characteristic of the data being observed
2. Quantitative: numerical data, which can be of two types
    a. Quantitative Discrete: most likely the result of counting how many
    b. Quantitative Continuous: the data can take any of a arrange of values; for quantitative continuous for any two values there is always another possible value between them

Qualitative data includes:
Students' hair color, American's religious background, ethnic background of students taking a standardized test

Quantitative discrete data includes:
The number of students entering the tenth grade, the number of red cars, the number of high schools in a town

Quantitative continuous data includes:
The distance from home to work, the weight of candy bars
Note that for any two distances 5.2 km and 5.3 km there is always another value in between such as 5.25 km. The same is true for weight.

Here we are going to focus on descriptive statistics. If we have a list of data, we may want to start by organizing it into a frequency distribution. A frequency distribution contains a column of classes and their frequencies.

Consider the years of education of a group of young adults:

12	10	15	12	15	13	11	14	11	13
13	14	13	13	11	14	12	16	15	16

The frequency distribution is as follows:

Years of education	Number of young adults
10	1
11	3
12	3
13	5
14	3
15	3
16	2

The first column contains the classes and the second column contains the frequencies. The frequency distribution lets us interpret the data more easily:

- 10 years of education is the value that occurs the least (only once)
- 13 years of education is the most frequently occurring value (called the mode)
- The data is fairly equally spread around the value of 13 years of education, equally spread data is said to be symmetric.
- There are 8 young adults that have at least 14 years of education. At least 14 years is 14 years or more (14 years or 15 years or 16 years). 3+3+2=8 young adults

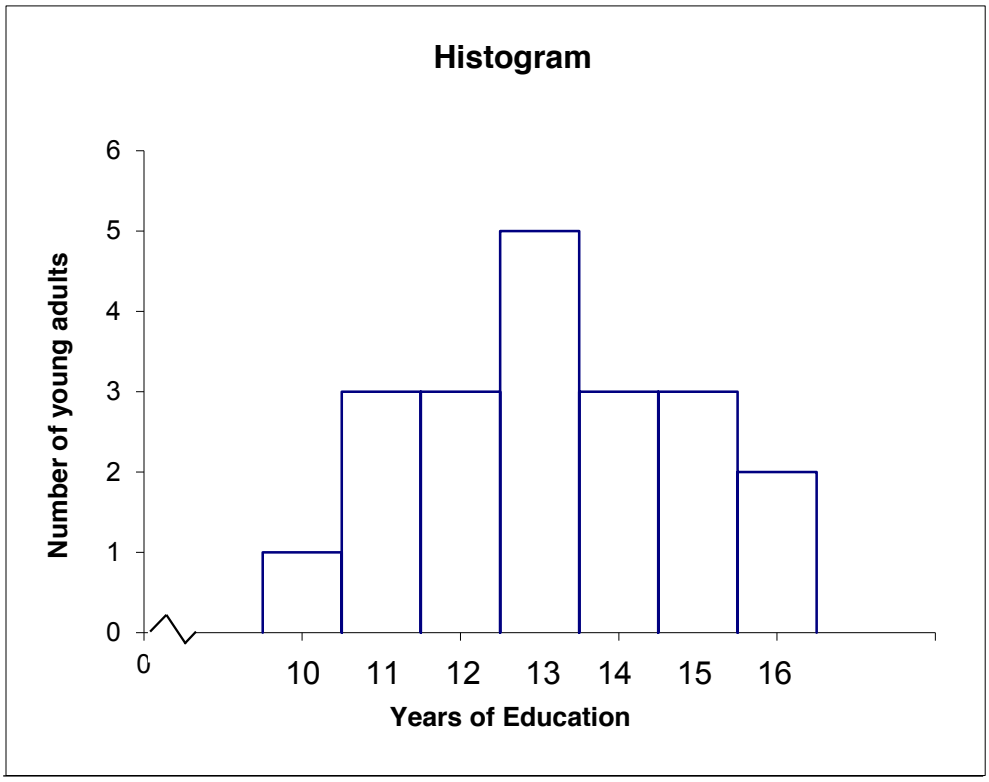
We create a frequency distribution by making classes starting with the smallest value and ending with the highest value. Then we tally all of the values to get the frequency for each class.

Some characteristics of frequency distributions:

- Classes are exhaustive – all values fall in some class
- Mutually exclusive classes – the classes do not overlap
- Avoid open-ended classes like more than 12 years
- 5 to 20 classes is typical for a frequency distribution
- Equal size classes

Histograms are bar charts where the vertical axis is for the frequency and the horizontal axis is for the classes. The bars will touch.

Using the frequency distribution from above:

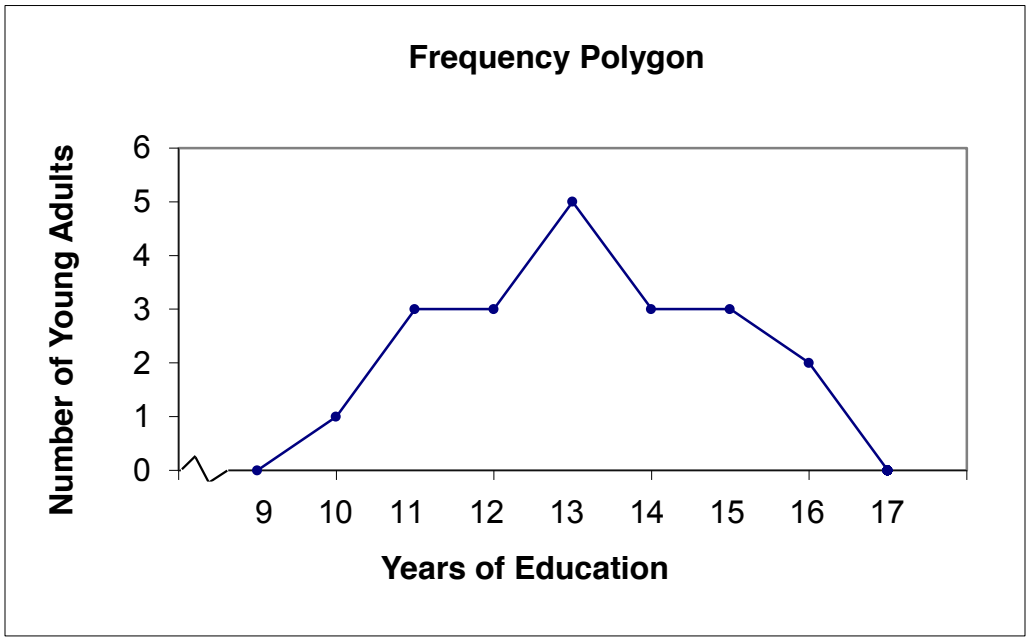


Here we have a histogram. The ⩘ symbol on the horizontal axis indicates that we are skipping out to a higher value

A frequency polygon is a line graph with the vertical axis representing the frequencies and horizontal axis representing the class (or class midpoint). To make a polygon, we can draw the class prior to the frequency distribution, which is 9 and the class after the distribution, which is 17 with frequencies of zero.

The frequency distribution was as follows:

Years of education	Number of young adults
10	1
11	3
12	3
13	5
14	3
15	3
16	2



Both the histogram and the frequency polygon give us similar types of information. We see a picture of the shape of the graph. In both cases we can see that our data (years of education) is somewhat bell shaped.

Exercises

Say whether the following data is qualitative, quantitative discrete, or quantitative continuous.

1. The number of fish in an aquarium

2. Types of species found on a remote island

3. Speed of an Internet connection

4. The number of students in a statistics class

5. Color of a dog

6. Percent of alcohol in a popular adult beverage

7. Thirty families were asked how many electronic devices (tablets, cell phones, and computers) they have in their home. The number of devices for each family is listed below:

3	5	2	6	1	3	3	6	4	5
2	3	5	4	6	3	5	5	2	4
8	4	8	8	3	6	5	6	5	4

   Make a frequency distribution.
   How many families had at least 3 electronic devices in their home?

8. Forty office workers were asked how many cups of coffee that they drink in a day. The number of cups of coffee for each worker is listed below:

2	0	2	1	3	4	2	0	2	5
3	2	2	3	2	5	2	2	3	3
2	0	3	2	3	1	0	2	0	6
4	2	1	0	0	3	2	4	2	3

Make a frequency distribution.
How many of the office workers drink at most three cups of coffee in a day?

9. Twenty students were asked how many pairs of sunglasses that they lost over the last year. The number of sunglasses that each student lost is listed below:

1	0	2	1	0	0	1	0	3	0
0	1	2	0	1	3	1	0	2	1

Make a frequency distribution.
How many students did not lose any sunglasses over the last year?

10. The number of gas stations in eighteen towns was counted. The number of gas stations in the towns is listed below:

7	10	9	10	11	7	8	12	10
11	8	10	7	9	11	10	11	9

Make a frequency distribution.
How many of the towns have between 9 and 11 gas stations inclusive?

11. Make a histogram for number 7. Label the axes. How many families had at most 3 electronic devices in their home?

12. Make a histogram for number 8. Label the axes. How many of the office workers drink at least three cups of coffee in a day?

13. Make a histogram for number 9. Label the axes. How many students lost at least 1 pair of sunglasses over the last year?

14. Make a histogram for number 10. Label the axes. How many of the towns have between 7 and 9 gas stations inclusive?

15. Make a frequency polygon for number 7. Label the axes. How many families had between 2 and 4 electronic devices inclusive in their home?

16. Make a frequency polygon for number 8. Label the axes. How many of the office workers drink between 1 and 3 cups of coffee inclusive in a day?

17. Make a frequency polygon for number 9. Label the axes. How many students lost less than 4 pairs of sunglasses over the last year?

18. Make a frequency polygon for number 10. Label the axes. How many of the towns have at least 10 gas stations?

19. How are the frequency distribution, histogram, and frequency polygon similar? How are they different?

20. Do you prefer a histogram or a frequency polygon? Why?

For a data set, we may want to come up with one value to represent the whole data set. For instance a student over her college career will have various different grades for a multitude of courses. So, schools will assign 4.0 to an A, 3.0 to a B, 2.0 to a C, 1.0 to D, and 0.0 to a F and then take the average of all the student's grades on this 4, 3, 2, 1, 0 scale. The result is the grade point average (GPA). A student's GPA is a quick way to determine how a student did throughout her college career by looking at only one value.

Mean (also called arithmetic mean) for individual data:

$$\bar{x} = \frac{\sum x}{n}$$

$\bar{x}$ is the mean
Σ means add all the values; it is the Greek letter "sigma" which is used for sums
x are the individual data values
n is the number of data values

This notation is new for many of us but the idea of adding all the values and dividing by the number of values is the "average" that we have seen throughout our lives.

Example:

1. Five people consume the following number of calories in a day:
2000 2500 2000 3200 2300
Find the mean.

$$\bar{x} = \frac{\sum x}{n} = \frac{2000+2500+2000+3200+2300}{5} = \frac{12{,}000}{5} = 2400 \text{ calories}$$

The mode is the most frequently occurring value. Another non-statistical way to think of mode is as the prevailing fashion. That may help us remember that mode is the most popular or most frequently occurring data value.

Example:

2. Five people consume the following number of calories in a day:
2000 2500 2000 3200 2300
Find the mode.

The mode is 2000 calories because it occurs most often.

The median is another measure of central tendency. When the data is ordered small to large (or large to small) the median is the data value in the middle. For an even number of data values the median may lie between two data values in which case we take the mean of those two data values.

Examples:

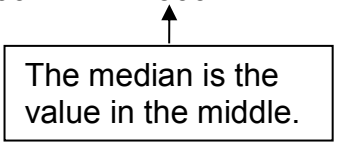
3. Five people consume the following number of calories in a day:

2000 2500 2000 3200 2300

Find the median.

First order the data small to large:

2000 2000 2300 2500 3200



The median is 2300 calories.

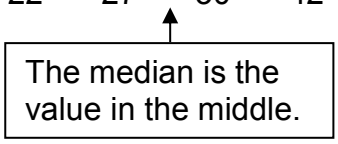
4. Six students graduate from college at the following ages:

22 30 27 42 22 47

Find the median.

First order the data small to large:

22 22 27 30 42 47



Since the median is between two data values, we take the mean of the 27 and 30. $\frac{27+30}{2} = 28.5$. The median age for graduation is 28.5 years old.

The midrange is the data value in the middle of the lowest and highest values.

midrange = $\frac{\text{lowest data value + highest data value}}{2}$ We can also say the midrange is the mean of the lowest and highest data values.

Example:

5. Five people consume the following number of calories in a day:

2000 2500 2000 3200 2300

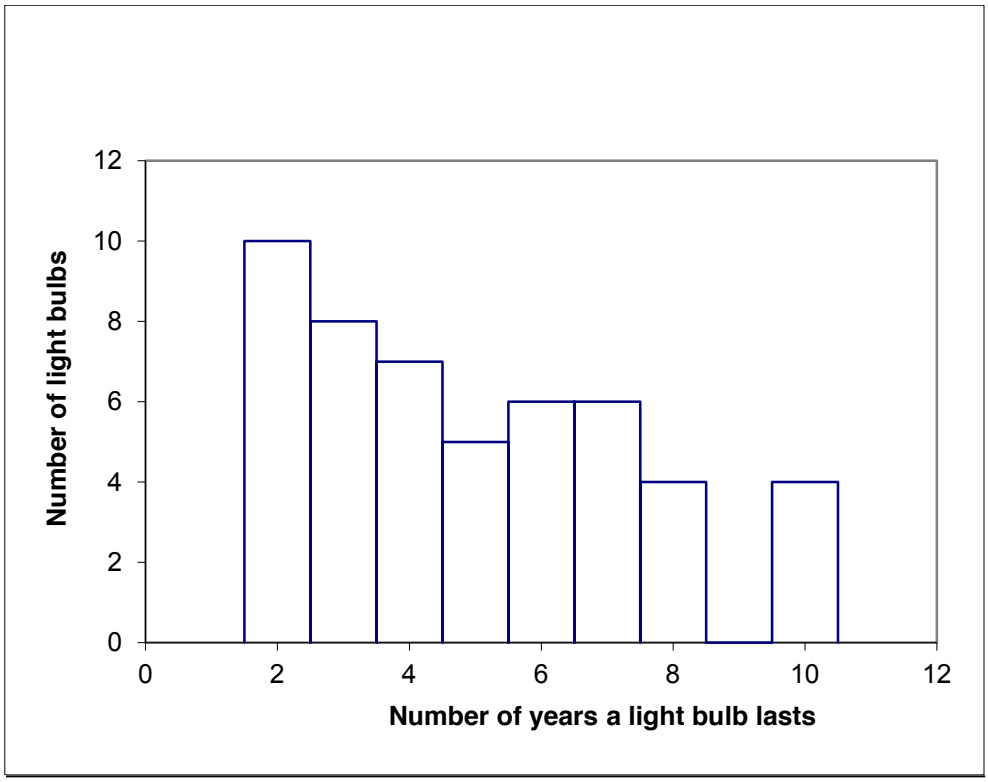
Find the midrange.

midrange = $\frac{\text{lowest data value + highest data value}}{2} = \frac{2000+3200}{2} = 2600$ calories

All four measures of central tendency will have the same units as the original, individual data values. However, we may get four different values for the four measures of central tendency (mean, mode, median, midrange), as in the calorie example above. So, which measure of central tendency is the best? That will certainly depend. The midrange and the mean are affected by extreme values. In fact the midrange is only dependent on the lowest and highest values, which may not be representative. The mode is only dependent on the one value that is repeated the most. So, the mode is easy to calculate, but it does not take into account all of the data values. The median is the middle of the data when it is ordered. So, extreme values do not affect it. The mean turns out to be very important in large part because all of the data values are used in its calculation.

We can also calculate the measures of central tendency from data that is grouped into a frequency distribution, histogram, or frequency polygon. A good first step will often be to write the histograms or frequency polygons as a frequency distribution.

Consider the actual life span of 50 light bulbs:



We can translate the histogram into a frequency distribution. Remember the classes are along the horizontal axis and the frequencies are along the vertical axis.

Lifespan of light bulb in years	Number of light bulbs
2	10
3	8
4	7
5	5
6	6
7	6
8	4
9	0
10	4

The mode from a frequency distribution is easy. The mode is the data value with the highest frequency. In the light bulb example the mode is a lifespan of 2 years because it has the highest frequency, which is 10. For a histogram or frequency polygon it is the class with the highest bar or point.

The midrange is just the $\frac{\text{lowest data value + highest data value}}{2} = \frac{2+10}{2} = 6$ years. Since the grouped data displays lowest and highest values we can just look at the chart or graph to get the needed values.

The mean gets a little more complicated. To find the mean we add up all the data values. Since there are 10 light bulbs with a 2 year lifespan we would be adding 2+2+2+2+2+2+2+2+2+2=20, but it is much easier to just multiply the data value by the frequency 2·10=20. The same holds true for all the other classes. There are 8 light bulbs with a 3 year life span. Rather than adding 3+3+3+3+3+3+3+3=24, it is easier to multiply data value by frequency 3·8=24.
By grouped data we mean a frequency distribution, frequency polygon, or histogram.

Formula for mean of grouped data:

$$\bar{x} = \frac{\sum f \cdot x_m}{n}$$

$\bar{x}$ is the mean
Σ means add all the values. It is called a summation.
f is the frequency for each class
x_m= class (or class midpoint)
n is the number of data values, which is found by adding up all of the frequencies

Use the frequency distribution to make a table:

x_m	f	$f \cdot x_m$
2	10	20
3	8	24
4	7	28
5	5	25
6	6	36
7	6	42
8	4	32
9	0	0
10	4	40
	50	247

The sum of the frequencies is total number of data values n.

Adding the $f \cdot x_m$ gives us the sum of all the individual data values or $\sum f \cdot x_m$.

The mean is
$\bar{x} = \frac{\sum f \cdot x_m}{n} = \frac{247}{50} = 4.94$ years

Finally, we have the median. Remember the median is the middle of the data listed from low to high. It is a little tricky at first because we tend to overlook that there are 10 values of 2 years, 8 values of 3 years, 7 values of 4 years, so on and so forth.

2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 5 5 5 5 5 6 6 6 6 6 6 7 7 7 7 7 7 8 8 8 8 10 10 10 10

The median is between 4 and 5.
So, we say the median is 4.5 years

x_m	f	Cumulative frequency
2	10	10
3	8	18
4	7	25
5	5	30
6	6	
7	6	
8	4	
9	0	
10	4	

There are 25 data values in the first three classes, which is halfway to the total of 50 data values. The median must be between 4 years and 5 years, which is 4.5 years.

These four measures of central tendency can be calculated from a list of individual data or when the data is grouped into a frequency distribution, frequency polygon, or histogram. The above light bulb example shows how to find the measures of central tendency for a histogram or frequency distribution. To find the measures of central tendency for a frequency polygon is similar.

Exercises

A group of five students received the following grades on a quiz:

100 70 90 80 95

1. Find the mean.
2. Find the mode.
3. Find the median.
4. Find the midrange.

A group of seven students received the following grades on a quiz:

70 85 70 100 95 75 65

5. Find the mean.
6. Find the mode.
7. Find the median.
8. Find the midrange.
9. Which do you think is the best measure of central tendency and why?

A class of thirty students had the following number of absences per week:

0 5 2 6 0 3 4 1

10. Find the mean.
11. Find the mode.

12. Find the median.

13. Find the midrange.

Instructors are evaluated on a five point scale for many different attributes and then given an overall score. The overall score for a group of ten instructors is listed below:

4.5 3.9 4.4 4.5 3.7 4.8 4.7 4.5 3.5 4.1

14. Find the mean.

15. Find the mode.

16. Find the median.

17. Find the midrange.

18. Which do you think is the best measure of central tendency and why?

In summertime children often get lost on the beach and then the parents have to go pick the children up at the Beach Public Services Office. The number of children taken to the Beach Public Services Office for ten days is listed below:

7 4 6 6 2 5 6 3 2 3

19. Find the mean.

20. Find the mode.

21. Find the median.

22. Find the midrange.

At a supermarket when people drop a box of eggs on the floor and make a mess they usually do not pay for the eggs or pick them up. Over a week the number of

cartons of broken eggs can mount up. At a large supermarket the store manager decided to keep track of the number of cartons of eggs that were dropped and cleaned up on a weekly basis. Over twelve weeks the number of cartons of eggs dropped and then cleaned up by store employees is listed below:

11 5 15 17 12 15 3 8 19 15 7 10

23. Find the mean.

24. Find the mode.

25. Find the median.

26. Find the midrange.

Forty office workers were asked how many cups of coffee that they drink in a day with the following frequency distribution:

Cups of Coffee	Number of workers
0	7
1	3
2	15
3	9
4	3
5	2
6	1

27. Find the mean.

28. Find the mode.

29. Find the median.

30. Find the midrange.

31. Which do you think is the best measure of central tendency and why?

Twenty students were asked how many pairs of sunglasses that they lost over the last year. The number of sunglasses that each student lost is listed below:

Pairs of Sunglasses	Number of students
0	8
1	7
2	3
3	2

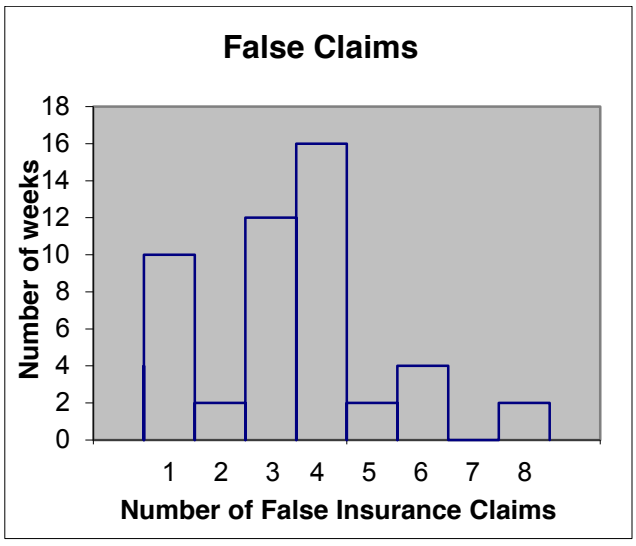
32. Find the mean.

33. Find the mode.

34. Find the median.

35. Find the midrange.

An insurance company receives a lot of false claims over a year. The number of false claims is counted and organized into the following histogram:



36. Find the mean.

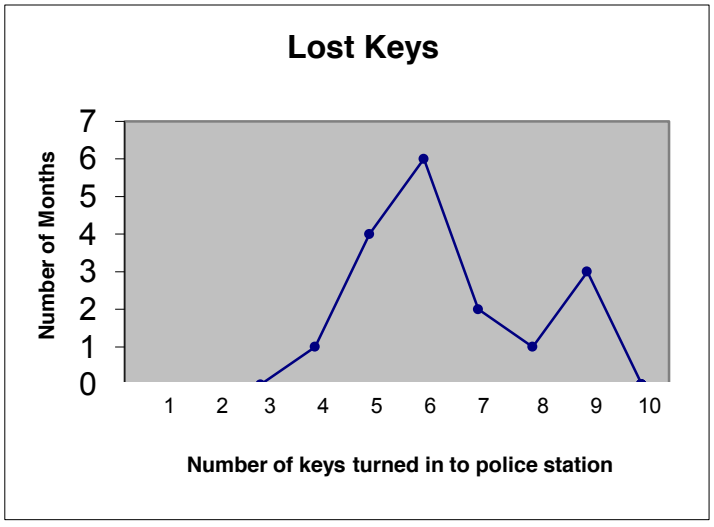
37. Find the mode.

38. Find the median.

39. Find the midrange.

40. Which do you think is the best measure of central tendency and why?

When people lose their keys, the keys are often turned into the local police station. Consider the following frequency polygon for the number of keys turned into local police station per month:



41. Find the mean.

42. Find the mode.

43. Find the median.

44. Find the midrange.

In the last section we saw how to find one data value to represent data, but we may also want to know whether the data is close together or spread out. Consider the following two sets of data:

Data Set 1:
20 20 20 50 80 80 80

Data Set 2:
40 40 40 50 60 60 60

The mean, median, and midrange are 50 for both data sets, but the first data set is more spread out in than the second. We will be concerned with two measure of dispersion – the range and the standard deviation.

The range is the highest value minus the lowest value. The range for Data Set 1 is 80 – 20 = 60. The range for Data Set 2 is 60 – 40 = 20.

To measure the spread of the data (or dispersion), we might try to compare each data value to the mean by subtracting the mean from each data value as in the table below. Remember $\bar{x}$ = mean and x= data values .
For Data Set 1:

x	$x - \bar{x}$	
20	-30	for 20 – 50
20	-30	for 20 – 50
20	-30	for 20 – 50
50	0	for 50 – 50
80	30	for 80 – 50
80	30	for 80 – 50
80	30	for 80 – 50

Here we have compared each data value to the mean, but when we add the column $x - \bar{x}$ the sum is zero and it will be zero every time. There are two choices at this point. Firstly, we could take the absolute value of the $x - \bar{x}$ column. Then all the values would be positive and we could find some average distance to the mean, which is called the mean deviation:

x	$x - \bar{x}$	$\lvert x - \bar{x} \rvert$
20	-30	30
20	-30	30
20	-30	30
50	0	0
80	30	30
80	30	30
80	30	30
		180

$$\sum |x - \bar{x}| = 180$$

$$\text{Mean Deviation} = \frac{\sum |x - \bar{x}|}{n} = \frac{180}{7} = 25.71$$

This mean deviation is an average distance from the mean for all the data values. It is easy enough to understand, but it turns out that another measure of dispersion is far more important. Let's consider the data values minus the mean again, but this time we will square those differences instead of taking the absolute value.

x	$x - \bar{x}$	$(x - \bar{x})^2$
20	-30	900
20	-30	900
20	-30	900
50	0	0
80	30	900
80	30	900
80	30	900
		5400

Not to worry too much about why, but squaring the $x - \bar{x}$ is another way to get positive values that will not cancel when we add them.

Formula for standard deviation:

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

s = standard deviation
x =data values
$\bar{x}$ = mean
$\sum$ = add all the values
n = the number of data values

The units for standard deviation are the same as the units for the original data values.

The standard deviation measure the spread of the data around the mean.

From our above chart for Data Set 1, we have $\sum(x - \bar{x})^2 = 5400$. Because there are 7 data values $n = 7$. So, the standard deviation for Data Set 1 is

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{5400}{7-1}} = 30$$

Example:

Calculate the standard deviation for the following ages of students
25 21 37 18 32 27 45 19

X	$x - \bar{x}$	$(x - \bar{x})^2$
25	-3	9
21	-7	49
37	9	81
18	-10	100
32	4	16
27	-1	1
45	17	289
19	-9	81
224		626

First find the mean:

$\bar{x} = \frac{\sum x}{n} = \frac{224}{8} = 28$ years old

To find the standard deviation:

Subtract the mean from each data value, square those differences, and then take the sum of those squared differences:

$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{626}{8-1}} = 9.46$ years

Finding the standard deviation with the formula has some disadvantages:

1. If the mean is a rounded-off decimal value, then there will be a lot of rounding error in the final answer.
2. If we have a frequency distribution, frequency polygon, or histogram, then it may be very difficult to calculate the standard deviation. Imagine that there are several hundred data values that are grouped into a histogram.

The good news is that a modern scientific calculator will let you calculate the mean and standard deviation of individual data (a list of values) or grouped data (frequency distribution, frequency polygon, or histogram). We just load the data into the calculator and it does all the work for us.

So, what does the standard deviation measure?

The standard deviation measures the spread of the data around the mean.

Exercises

A group of five students received the following grades on a quiz:

100 70 90 80 95

1. Find the range.

2. Find the standard deviation.

A group of seven students received the following grades on a quiz:

70 85 70 100 95 75 65

3. Find the range.

4. Find the standard deviation.

Consider the following two sets of test scores:
Set 1:
60 60 80 100 100

Set 2:
60 75 80 85 100

5. Find both means.

6. Find both ranges.

7. Find both standard deviations.

8. Do the means and ranges indicate that there is a difference in the two data sets?

9. What do the standard deviations show about the two data sets?

A class of thirty students had the following number of absences per week:

0 5 2 6 0 3 4 1

10. Find the range.

11. Find the standard deviation.

Instructors are evaluated on a five point scale for many different attributes and then given an overall score. The overall score for a group of ten instructors is listed below:

4.5 3.9 4.4 4.5 3.7 4.8 4.7 4.5 3.5 4.1

12. Find the range.

13. Find the standard deviation.

In summertime children often get lost on the beach and then the parents have to go pick the children up at the Beach Public Services Office. The number of children taken to the Beach Public Services Office for ten days is listed below:

7 4 6 6 2 5 6 3 2 3

14. Find the range.

15. Find the standard deviation.

16. What does the standard deviation measure?

At a supermarket when people drop a box of eggs on the floor and make a mess they usually do not pay for the eggs or pick them up. Over a week the number of cartons of broken eggs can mount up. At a large supermarket the store manager decided to keep track of the number of cartons of eggs that were dropped and cleaned up on a weekly basis. Over twelve weeks the number of cartons of eggs dropped and then cleaned up by store employees is listed below:

11 5 15 17 12 15 3 8 19 15 7 10

17. Find the range.

18. Find the standard deviation.

Forty office workers were asked how many cups of coffee that they drink in a day with the following frequency distribution:

Cups of Coffee	Number of workers
0	7
1	3
2	15
3	9
4	3
5	2
6	1

19. Find the standard deviation.

20. Find the range.

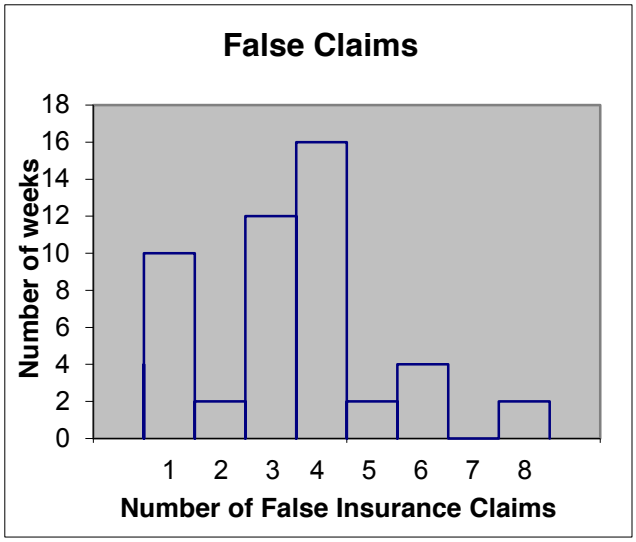
Twenty students were asked how many pairs of sunglasses that they lost over the last year. The number of sunglasses that each student lost is listed below:

Pairs of Sunglasses	Number of students
0	8
1	7
2	3
3	2

21. Find the standard deviation.

22. Find the range.

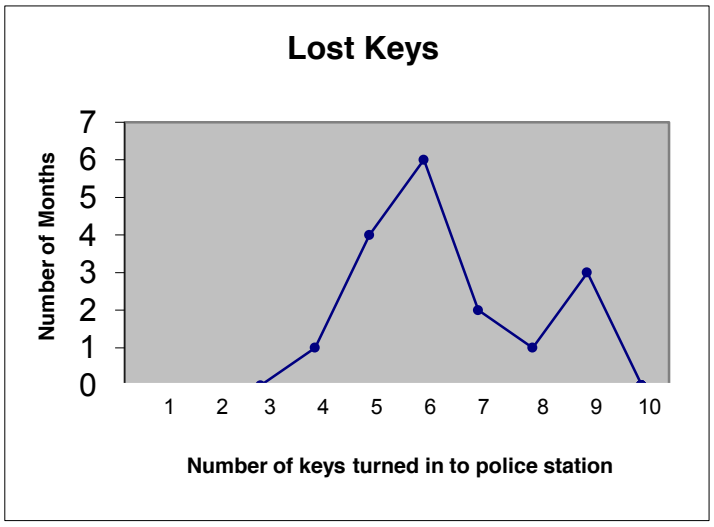
An insurance company receives a lot of false claims over a year. The number of false claims is counted and organized into the following histogram:



23. Find the standard deviation.

24. Find the range.

When people lose their keys, the keys are often turned into the local police station. Consider the following frequency polygon for the number of keys turned into local police station per month:



25. Find the standard deviation.

26. Find the range.

Standardized tests such as Navy Advancement Exams, Scholarship Aptitude Test (SAT), and IQ tests may be based on scores of 80, 1600, 100 or some other score. These scores may or may not mean much too most people depending on the individual's experience with the specific exam. So, along with these various scores, the results will often include a percentile score.

Percentile refers to the value that is above a given percent of the data. For example a score of 700 on the Math SATs may put an individual in the 93rd percentile. That means that the score of 700 is higher than 93% of the individuals that took the test.

Percentiles are also used when talking about peoples physical characteristics like height and weight. A height of 57 inches is the 81st percentile for a 10 year old boy, which means a 57 inch tall 10 year old boy is taller than 81% of all 10 year old boys.

The normal distribution, which is bell-shaped distribution, is found throughout nature. Characteristics of people like intelligence, height, and weight follow the normal distribution. The normal distribution is based on the normal curve.

Characteristics of the normal curve:
1. bell-shaped
2. symmetric
3. area under the curve is 1 or 100%
4. mean, mode, median are in the middle
5. the curve approaches the horizontal axis
6. the shape is completely determined by the mean and standard deviation
7. the empirical rule applies

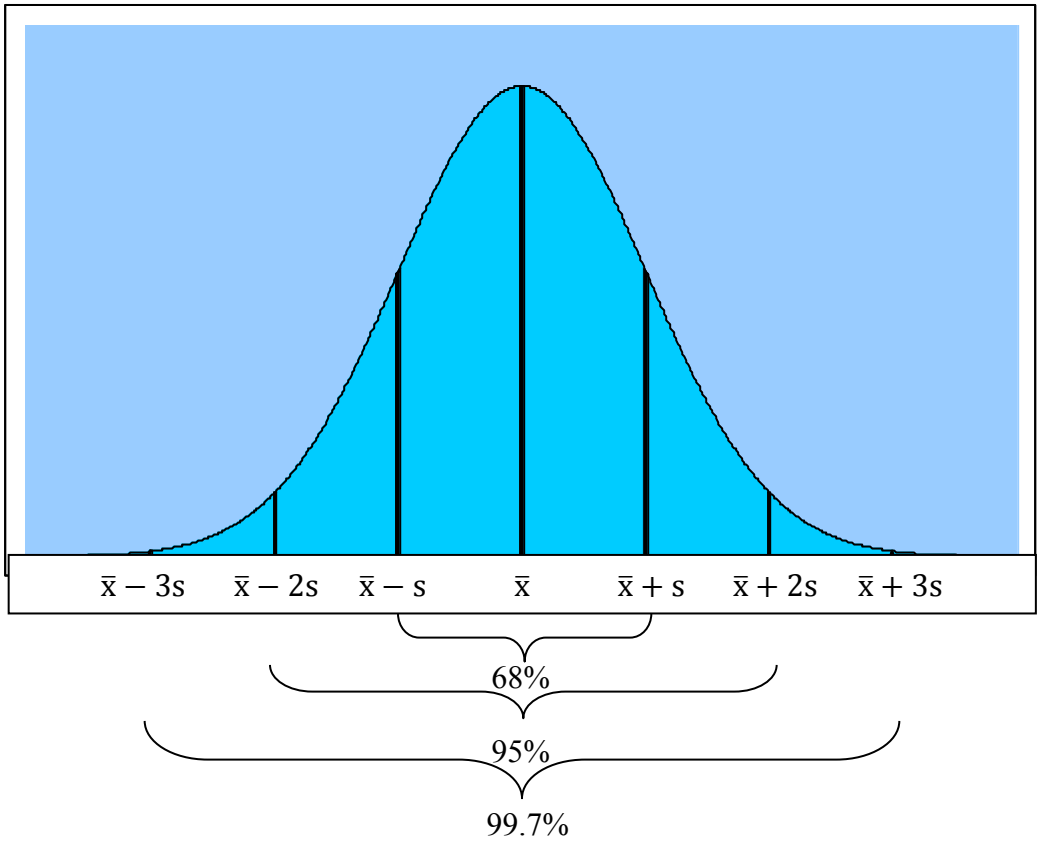
The empirical rule states that for normally distributed data:

68% of the data is within 1 standard deviation of the mean
95% of the data is within 2 standard deviations of the mean
99.7% of the data is within 3 standard deviation of the mean

Below is a picture of the empirical rule and normal curve. Remember that

$\bar{x}$ is the mean
s is the standard deviation
$\bar{x} - 2s$ is the mean minus to standard deviations. So, between $\bar{x} - 2s$ and $\bar{x} + 2s$ is within 2 standard deviations of the mean.

Notice the shape of the normal curve as well as the empirical rule.

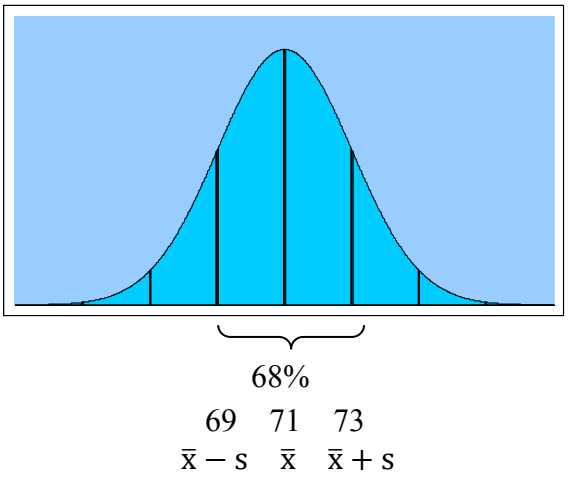


Examples:

1. Heights for adult males are normally distributed with a mean of 71 inches and a standard deviation of 2 inches.

a. What percentage of adult males are between 69 and 73 inches tall?

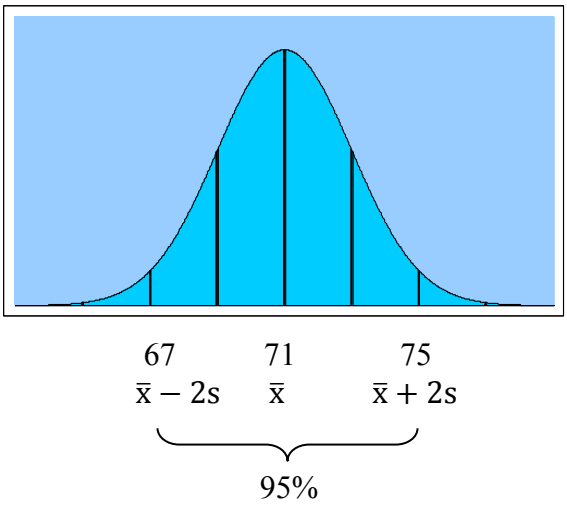
At this point we only know the empirical rule. As it turns out if we add and subtract one standard deviation (2 inches) to the mean we get
71 – 2 = 69 inches, which is the mean minus 1 standard deviation
71 + 2 = 73 inches, which is the mean plus 1 standard deviation



Since the heights of 69 inches to 73 inches are within 1 standard deviation of the mean and heights are normally distributed, 68% of adult males have heights between 69 and 73 inches.

b. What percentage of adult males are more than 75 inches tall?

71 + 2 + 2 = 75 inches 75 inches is the value 2 standard deviations above the mean.



The empirical rule tells us that 95% of the data is within 2 standard deviations of the mean. So, 100 – 95 = 5% of the data is outside of 2 standard deviations from the mean (more than 75 and less than 67). We only want the heights more than 75 inches, which is half of the 5% or 2.5%.

2.5% of adult males are taller than 75 inches.

Using the empirical rule is not only awkward, but we are limited to talking about values that are 1, 2, or 3 standard deviations from the mean. Also, different types of data yield different normal distributions because they have different means and different standard deviations. Next we will standardize normal distributions so that we can use one chart to determine percents of data and probabilities.

Z-scores:
The z-score measure the number of standard deviations between a data value and the mean.
z-score formula:

$$z = \frac{x - \bar{x}}{s}$$

x = data value we are interested in
$\bar{x}$ = mean
s = standard deviation
z = z-score

Examples:

2. For normally distributed data with a mean of 50 and standard deviation of 10, find the z-score for the following values:

a. value of 60

Steps	Reasons
$z = \frac{x - \bar{x}}{s}$	Write the formula for the normal distribution from the formula sheet.
$z = \frac{60 - 50}{10}$	Plug in the data value, mean, and standard deviation. Calculate the z-score.
$z = 1$	

Because z is positive the value of 60 is 1 standard deviation more than the mean.

b. value of 32

Steps	Reasons
$z = \frac{x - \bar{x}}{s}$	Write the formula for the normal distribution from the formula sheet.
$z = \frac{32 - 50}{10}$	Plug in the data value, mean, and standard deviation. Calculate the z-score.
$z = -1.8$	

Because z is negative the value of 32 is 1.8 standard deviations less than the mean.

Not only do z-scores tell us the number of standard deviations that a value is from the mean, it also lets us use one chart to calculate the percent of data less than a value, more than a value, or between two values when we know the data is normally distributed.

The normal distribution chart that comes with this course is what is known as a percentile chart. It will tell us the percent of data that is less than a given z-value.

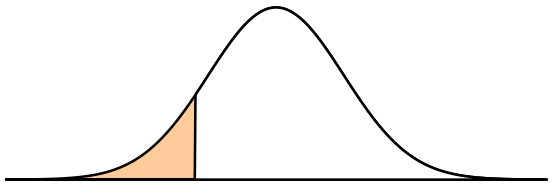
Examples:

3. To find the percent of data less than z = -2.47, go down the chart on the left to the z-score -2.4. Then go across to .07 to

Standard Normal Distribution Cumulative Probabilities (Percentiles)
Table Values represent area to the left of z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	**0.07**	0.08	0.09
-3.00	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	**0.0011**	0.0010	0.0010
-2.90	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	**0.0015**	0.0014	0.0014
-2.80	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	**0.0021**	0.0020	0.0019
-2.70	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	**0.0028**	0.0027	0.0026
-2.60	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	**0.0038**	0.0037	0.0036
-2.50	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	**0.0051**	0.0049	0.0048
-2.40	**0.0082**	**0.0080**	**0.0078**	**0.0075**	**0.0073**	**0.0071**	**0.0069**	**0.0068**	0.0066	0.0064

So, 0.0068 or 0.68% of the data is less than z=-2.47. That percentage is represented by the shaded region below. The chart always gives us the percent less or shaded area to the left of z.



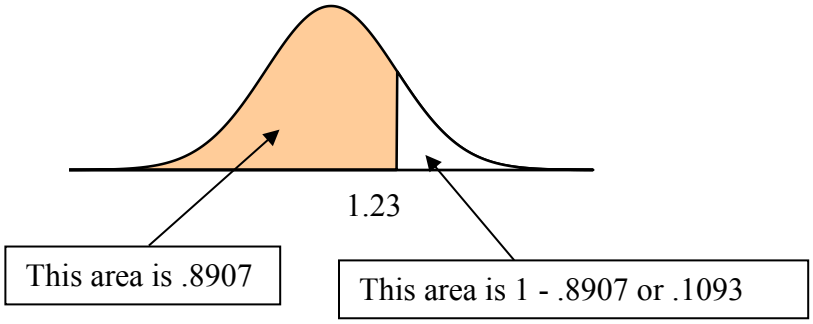
-2.47

4. To find the percent of values that is more than z = 1.23, we still use the chart:

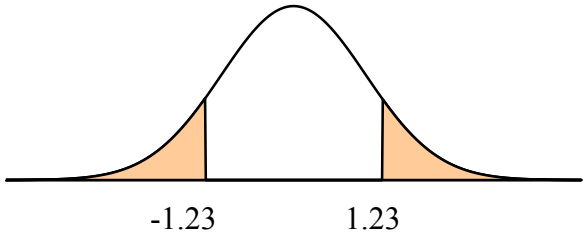
Standard Normal Distribution Cumulative Probabilities (Percentiles)
Table Values represent area to the left of z

Z	0.00	0.01	0.02	**0.03**	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.5000	0.5040	0.5080	**0.5120**	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
1.00	0.8413	0.8438	0.8461	**0.8485**	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.10	0.8643	0.8665	0.8686	**0.8708**	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.20	**0.8849**	**0.8869**	**0.8888**	**0.8907**	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015

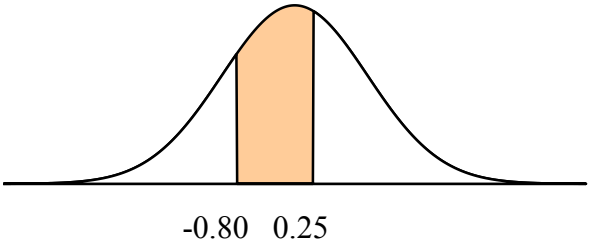
So, 0.8907 of the data is less than z = 1.23, but we are asked for the percent of data more than 1.23. Look at the picture of the normal distribution below



For the area more than a z-score, we just take 1 – value from the chart. So, .1093 or 10.93% of the data is more than a z-score of 1.23. We should also note that a score of more than 1.23 is the same as a z-score of less than 1.23. This is true because the normal distribution is symmetric as in the picture below.



5. Find the percent of values that is between z = -0.80 and z = 0.25.



Here we want to think of the area, which is also equal to the percent. The shaded area is all the area to left of z=0.25 minus the area to the left of z=-0.80. So, we will use the chart to look up both percentiles and find the difference.

Z	0.00	0.01	0.02	0.03	0.04	**0.05**	0.06	0.07	0.08	0.09
0.10	0.5398	0.5438	0.5478	0.5517	0.5557	**0.5596**	0.5636	0.5675	0.5714	0.5753
0.20	**0.5793**	**0.5832**	**0.5871**	**0.5910**	**0.5948**	**0.5987**	0.6026	0.6064	0.6103	0.6141
0.30	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0.90	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.80	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.70	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148

The percentile (or area) for z = .25 is 0.5987 and the percentile (or area) for z = -0.80 is 0.2119. We subtract the smaller area from the larger area to get the area between the two values. 0.5987 – 0.2119 = 0.3868

38.68% of the values are between z = -0.80 and z = 0.25

These last three examples really show us the three possibilities from the chart:

1. Percent less than a z-score: read it off the chart
2. Percent more than a z-score: 1 – values from the chart
3. Percent between two z-scores: look up both z-scores and find the difference

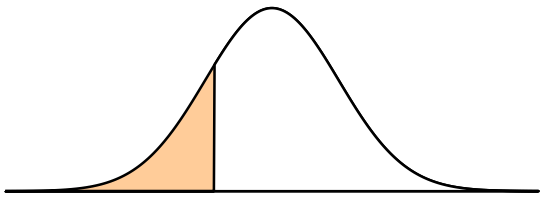
It is best to draw the bell curve when working with the normal distribution, but these three situations really do not change.

Since z-scores and the chart can be used to find the percent of data, they can also be used to find the probability of selecting a single value that is more than, less than, or between values when the data is normally distributed. The next step is to put everything together. We will take normally distributed values, calculate the z-scores, and use the chart to calculate percents or probabilities.

Examples:

For examples 6, 7, and 8, assume that life expectancy is normally distributed with a mean life expectancy of 75 years and a standard deviation of 12 years.

6. What percent of the population live less than 65 years?



x-values	65	75
z-score	-0.83	0

Draw the bell-curve and find the z-score: $z = \frac{x-\bar{x}}{s} = \frac{65-75}{12} - 0.8333$ ...
Shade to the left for values less than. Use z = – 0.83 rounded off to the hundredth because the chart has values to the hundredth.

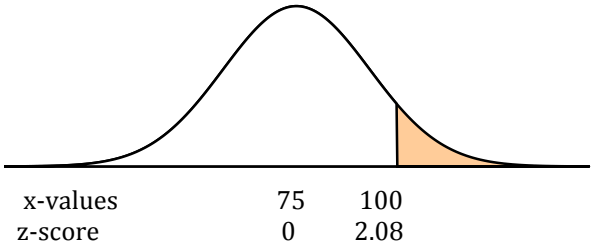
Z	0.00	0.01	0.02	**0.03**	0.04	0.05	0.06	0.07	0.08	0.09
-0.90	0.1841	0.1814	0.1788	**0.1762**	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.80	**0.2119**	**0.2090**	**0.2061**	**0.2033**	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.70	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148

The chart gives the percent of values less than. So, the answer is read from the chart.

0.2033 or 20.3% of the population live less than 65 years.

7. What is the probability that a randomly selected individual lives over 100 years?

First we need to note that the percent of people living over 100 years is the same as randomly selecting an individual that lives over 100 years. We continue as we have been with z-scores and the normal distribution chart.



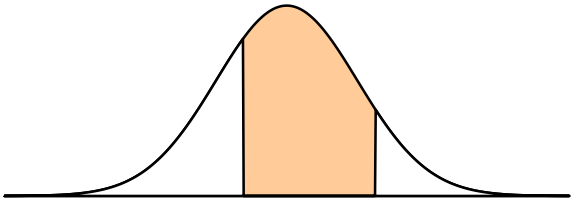
The chart always gives us area, percents, and probability for less than. To get the probability for more than 100 years we take 1 – the chart value.

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	**0.08**	0.09
1.90	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	**0.9761**	0.9767
2.00	**0.9772**	**0.9778**	**0.9783**	**0.9788**	**0.9793**	**0.9798**	**0.9803**	**0.9808**	**0.9812**	0.9817
2.10	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857

1 - 0.9812 = 0.0188

There is a 0.0188 or 1.88% probability that an individual will live more than 100 years.

8. What is the probability of selecting an individual that will live between 70 and 90 years?



When we use the normal distribution to find the percent or probability between two values, we find to z-scores, find two percentages, and take the difference. Looking at picture above, we see that the area of the shaded region is the area to the left of z=1.25 minus the unshaded area to the left of z=-0.42.

$z = \frac{x-\bar{x}}{s} = \frac{70-75}{12} = -0.4166 \ldots \text{use } z = -0.42$ Chart is exact to the hundredth.

$z = \frac{x-\bar{x}}{s} = \frac{90-75}{12} = 1.25$

Z	0.00	0.01	0.02	0.03	0.04	**0.05**	0.06	0.07	0.08	0.09
1.10	0.8643	0.8665	0.8686	0.8708	0.8729	**0.8749**	0.8770	0.8790	0.8810	0.8830
1.20	**0.8849**	**0.8869**	**0.8888**	**0.8907**	**0.8925**	**0.8944**	0.8962	0.8980	0.8997	0.9015
1.30	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177

Z	0.00	0.01	**0.02**	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0.50	0.3085	0.3050	**0.3015**	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.40	**0.3446**	**0.3409**	**0.3372**	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.30	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483

0.8944 – 0.3372 = 0.5572

There is a 0.5572 or 55.72% chance that an individual will live between 70 and 90 years.

Exercises

1. What does it mean that a height of 74 inches is the 95^{th} percentile for men?

2. What are the characteristics of the normal distribution?

3. What does it mean that a height of 69 inches is the 92^{nd} percentile for women?

4. What does it mean that a test score of 60 is the 85^{th} percentile?

5. What does it mean that a test score of 52 is the 98^{th} percentile?

Blood pressure is normally distributed. The units millimeter of mercury are represented by mm Hg. In a particular country the mean systolic blood pressure is 130 mm Hg with a standard deviation of 15 mm Hg. Use the empirical rule to answer the following questions about this particular country:

6. What percent of the population has a systolic blood pressure between 115 and 145 mm Hg?

7. What percent of the population has a systolic blood pressure between 100 and 160 mm Hg?

8. What percent of the population has a systolic blood pressure between 85 and 175 mm Hg?

9. What percent of the population has a systolic blood pressure more than 145 mm Hg?

10. What percent of the population has a systolic blood pressure less than 115 mm Hg?

11. What percent of the population has a systolic blood pressure less than 100 mm Hg?

12. What percent of the population has a systolic blood pressure more than 160 mm Hg?

Serum cholesterol levels are normally distributed. The mean serum cholesterol level for adult males is 210 mg/dL with a standard deviation of 45 mg/dL.

13. What percent of adult males have a serum cholesterol level between 165 mg/dL and 255 mg/dL?

14. What percent of adult males have a serum cholesterol level between 120 mg/dL and 300 mg/dL?

15. What percent of adult males have a serum cholesterol level less than 120 mg/dL?

16. What percent of adult males have a serum cholesterol level less than 165 mg/dL?

17. What percent of adult males have a serum cholesterol level more than 255 mg/dL?

18. What percent of adult males have a serum cholesterol level more than 255 mg/dL?

For normally distributed data with a mean of 72 and standard deviation of 6, find the z-score for the following data values. Round answers to the nearest hundredth.

19. 84

20. 66

21. 63

22. 87

23. 83

24. 68

25. 72

26. 58

27. 70

28. 91

For the standard normal distribution find the percent of z-scores for z:

29. less than z = 2.10

30. less than z = 1.57

31. less than z = -1.34

32. less than z = -2.31

33. more than z = 0.27

34. more than z = -0.35

35. more than z = -2.14

36. more than z = 0.59

37. more than z = -0.29

38. more than z = -3.04

39. between z = -1.36 and z = 2.08

40. between z = 0.17 and z = 1.23

41. between z = -0.09 and z = 3.01

42. between z = -2.12 and z = -1.08

The life span for light bulbs is normally distributed. The mean life span for a particular brand of light bulb is 975 hours with a standard deviation of 45 hours. For this particular brand of light bulbs, find the percent of light bulbs that have a life span of

43. less than 1047 hours

44. less than 1029 hours

45. more than 894 hours

46. more than 1074 hours

47. less than 890 hours

48. less than 920 hours

49. more than 1000 hours

50. more than 900 hours

51. between 975 and 1100 hours

52. between 925 and 975 hours

53. between 950 and 1000 hours

54. between 850 and 1050 hours

55. between 860 and 1035 hours

Assume that women's heights are normally distributed with a mean of 64 inches and a standard deviation of 2.5 inches. Find the probability of randomly selecting a women who is

56. less than 69 inches tall.

57. more than 67 inches tall.

58. between 61 and 70 inches tall.

59. less than 68 inches tall.

60. more than 65.8 inches tall.

61. between 62.4 and 68.3 inches tall.

62. less than 64 inches tall.

63. more than 71.6 inches tall.

64. between 56.4 and 70.4 inches tall.

65. less than 56.4 inches tall.

66. more than 62.3 inches tall.

67. between 59.7 and 63.1 inches tall.

When we were graphing, we were studying the relationship between two variables. Those graphs could take on many shapes such as lines and parabolas. When we study the relationship between two variables, one of the first steps is to draw the relationship on the coordinate plane.

With a scatter plot we draw the points of two variables to see if there is a relation. The independent variable is represented by the x-coordinates along the horizontal axis. The dependent variable is represented by the y-coordinates along the vertical axis. The independent variable is used to predict the dependent variable. With correlation and regression, we do not show that the independent variable (x) causes the dependent variable (y) to act in a certain way. We are only showing whether or not the dependent variable can be used to predict the value of the independent variable.

Example:

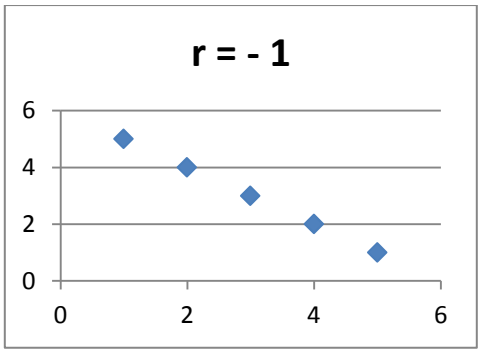
1. Different people were studied to see if there is a relationship between age and cholesterol level. Cholesterol measures the milligrams (mg) of cholesterol per deciliter (dL) of blood. Use the data below to draw a scatter diagram to see if there is a relationship between age and cholesterol.

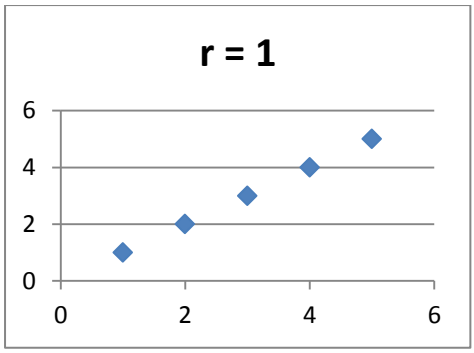
Age (x)	35	12	65	47	21	32	52	15	57
Cholesterol (y)	210	160	285	255	180	220	190	170	275

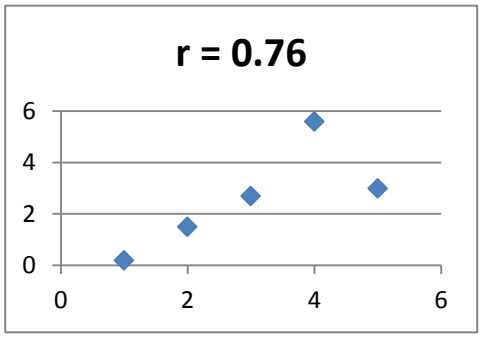
Graph the points (x,y) on a set of axes.

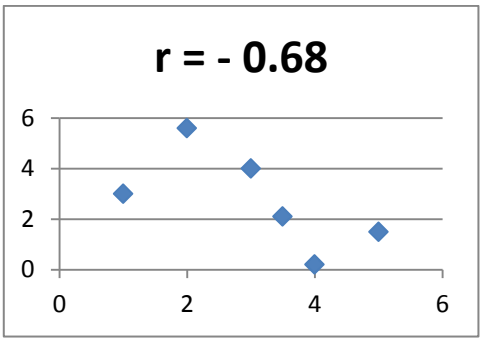
It is a good idea to label the axes with age and cholesterol level so that the reader knows what the variables are. Here we would be using the age (independent variable) to predict the cholesterol level (dependent variable). We are not saying that age is the cause or only cause of cholesterol level. It very well may be a factor, but there are clearly other factors such as heredity, diet, exercise, etc.

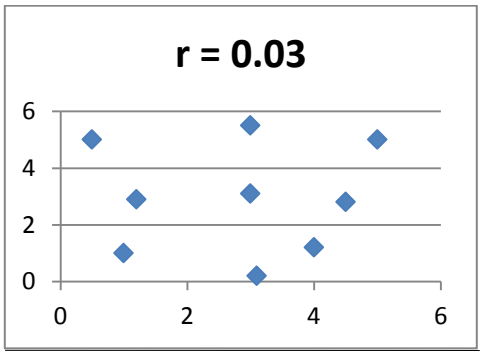
Once we have graphed the scatter plot, we can ask ourselves the question what is the shape of the graph. While not perfect the graph does look somewhat like a line, especially if we do not include the point (52,190). Linear correlation refers to the linear relationship between two variables. A strong linear correlation means that the scatter diagram points are close to a line. A weak linear correlation means that the scatter diagram points follow a line but not closely. The linear correlation coefficient or Pearson's coefficient (also called the correlation coefficient) ranges from r = -1 to 1. A value of r = 1 indicates there is perfect positive linear correlation, which is to say the scatter diagram points fall exactly on a line with positive slope ↗. A linear correlation coefficient of r= -1 indicates that the scatter diagram points fall exactly on a line with a negative slope ↘. If r = 0, there is no correlation.

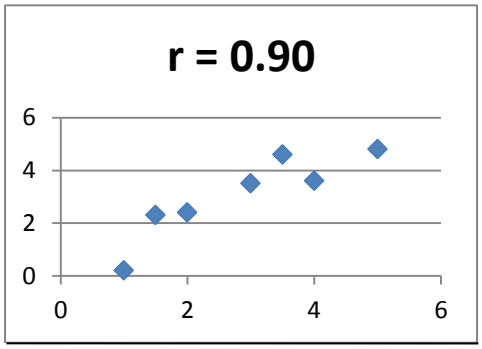












The formula for the linear correlation coefficient is the following:

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

r is the linear correlation coefficient
n is the number of data values
x are the independent variables (x)
y are the dependent variables (y)

There is one example of how to find the correlation coefficient below, but it is much easier to use a scientific calculator to find the correlation coefficient.

Examples:

2. Using the same data from example 1, different people were studied to see if there is a relationship between age and cholesterol level. Cholesterol measures the milligrams (mg) of cholesterol per deciliter (dL) of blood. Use the data below to find the correlation coefficient.

Age (x)	35	12	65	47	21	32	52	15	57
Cholesterol (y)	210	160	285	255	180	220	190	170	275

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

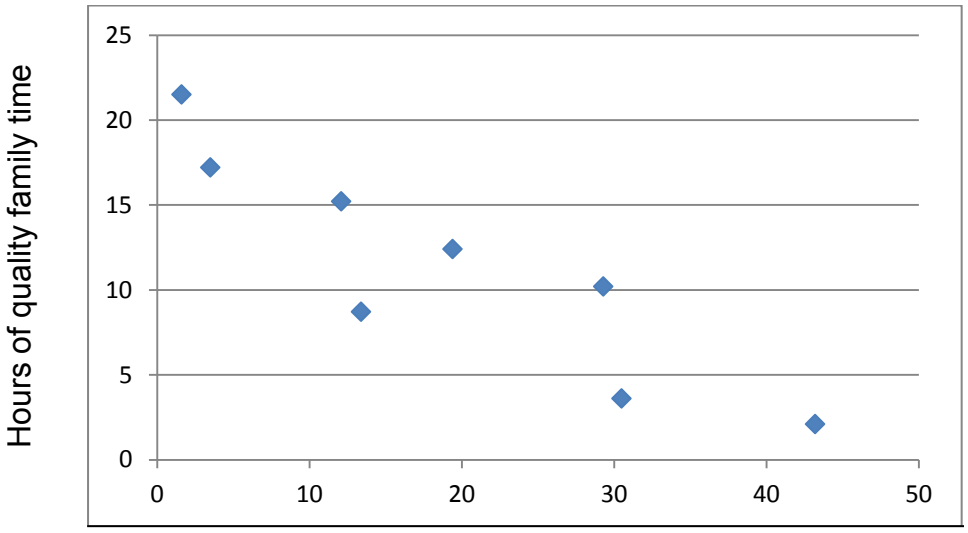
x	y	x^2	y^2	xy
35	210	1225	44100	7350
12	160	144	25600	1920
65	285	4225	81225	18525
47	255	2209	65025	11985
21	180	441	32400	3780
32	220	1024	48400	7040
52	190	2704	36100	9880
15	170	225	28900	2550
57	275	3249	75625	15675
336	1945	15446	437375	78705

$$r = \frac{9(78{,}705) - (336)(1945)}{\sqrt{\left[9(15{,}446) - (336)^2\right]\left[9(437{,}375) - (1945)^2\right]}} = \frac{54825}{63286.61233} = 0.866297$$

$r = 0.87$ is an indication that there is a strong positive correlation.

3. An experimenter believes that there is a relationship between the number of hours children spend with electronic devices with screens such as televisions, tablets, computers, or telephones and the amount of quality time spent with family. In the chart below the independent variable is the number of hours spent with screens and the dependent variable is the amount of quality time spent with family. Make a scatter plot, find the correlation coefficient, and comment on the correlation.

Hours with screens (x)	30.5	43.2	3.5	12.1	1.6	29.3	13.4	19.4
Hours of quality family time (y)	3.6	2.1	17.2	15.2	21.5	10.2	8.7	12.4



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

x	y	x^2	y^2	xy
30.5	3.6	930.25	12.96	109.8
43.2	2.1	1866.24	4.41	90.72
3.5	17.2	12.25	295.84	60.2
12.1	15.2	146.41	231.04	183.92
1.6	21.5	2.56	462.25	34.4
29.3	10.2	858.49	104.04	298.86
13.4	8.7	179.56	75.69	116.58
19.4	12.4	376.36	153.76	240.56
153	90.9	4372.12	1339.99	1135.04

$$r = \frac{8(1135.04) - (153)(90.9)}{\sqrt{\left[8(4372.12) - (153)^2\right]\left[8(1339.99) - (90.9)^2\right]}}$$

$$r = \frac{-4827.38}{5331.3929}$$

$$r = -0.90546$$

Both the scatter plot and the correlation coefficient of r = – 0.905 indicate that there is a strong negative correlation. Remember it is much easier to calculate the correlation coefficient using a scientific calculator.

Exercises

For the following sets of data, draw a scatter diagram. From the scatter diagram decide whether there is positive linear correlation, negative linear correlation, or no linear correlation.

1. A city wants to see if there is a linear relationship between property tax and annual income of its residents. A random sample of seven residents was selected with the following results:

Property Tax in thousands of dollars (X)	3.5	5.2	9.4	7.1	1.2	15.8	6.3
Annual Income thousands of dollars (Y)	55	65	125	102	23	205	85

2. A county relief organization wants to determine if there is a linear relationship between the number of children and annual household income in thousands of dollars.

Number of children (X)	3	6	2	5	1	2	4
Annual Income thousands of dollars (Y)	47	21	75	30	94	81	35

3. A fraternity at a large university wanted to determine if there is a linear relationship between the number of alcoholic beverages consumed per week not including weekends and the grade point average (GPA) of its members.

Number of Alcoholic Beverages Consumed (X)	4	18	7	0	10	6	12	3	15
Grade Point Average (Y)	3.8	2.3	3.5	3.7	2.8	4.0	3.2	3.6	2.9

4. A psychologist wants to find out if there is a linear relationship between the number of sunny days in a year and positive attitude. The psychologist measures positive attitude on a 100 point scale where a score of 100 indicates the highest possible positive attitude and 0 indicates a completely negative attitude (or the lowest possible positive attitude).

Number of Sunny Days (X)	240	315	142	85	330	220	123	98	57
Psychologists Attitude Test (Y)	73	89	65	25	68	78	81	58	32

5. A high school soccer coach wants to determine if there is a linear relationship between the number of days of practice and the number of wins in a season. She compares wins and days of practice for her last six seasons.

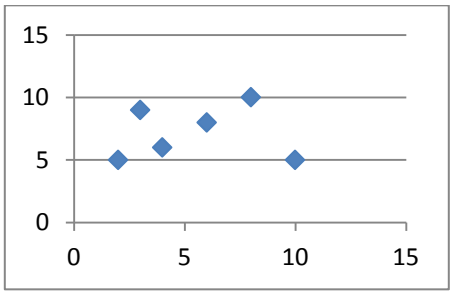
Number of Days of Practice (X)	45	56	48	60	52	58
Number of Wins (Y)	5	8	9	5	7	8

6. A regional manager for McDonald's believes that there is a linear relationship between a town's population and the number of successful McDonald's franchises in a town.

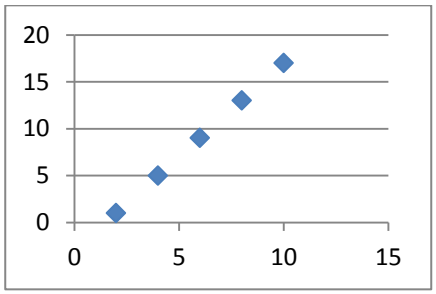
Population in tens of thousands (X)	25	37	6	15	11	9	53	3	28
Number of successful franchises (Y)	12	17	3	14	7	3	21	4	17

For the following problems, consider the following scatter plots:

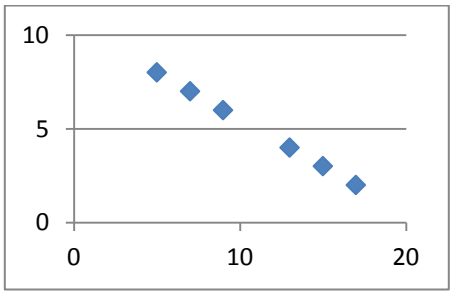
A



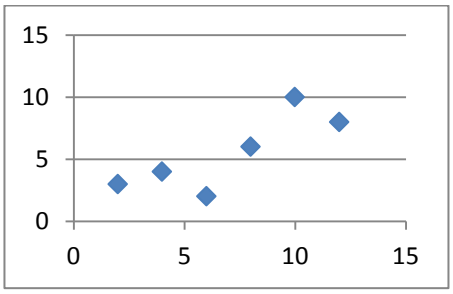
B



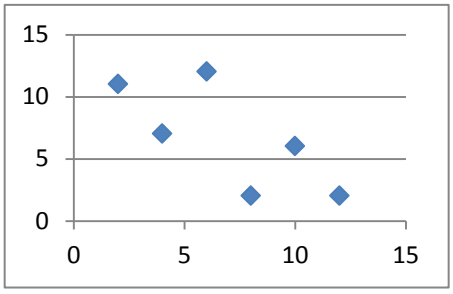
C



D



E



State which graph has the following correlation coefficient.

7. r = 0.82

8. r = -1.00

9. r = - 0.72

10. r = 1.00

11. r = 0.08

For the following problems you may use a calculator that finds the correlation coefficient or the formula. It is easier and quicker to use a calculator with the correlation coefficient feature.

12. Use the data from problem number 1 to find the correlation coefficient. According to the correlation coefficient is the correlation strong, moderate, or weak? Does the correlation coefficient support what you said about whether there is positive linear correlation, negative linear correlation, or no linear correlation?

13. Use the data from problem number 2 to find the correlation coefficient. According to the correlation coefficient is the correlation strong, moderate, or weak? Does the correlation coefficient support what you said about whether there is positive linear correlation, negative linear correlation, or no linear correlation?

14. Use the data from problem number 3 to find the correlation coefficient. According to the correlation coefficient is the correlation strong, moderate, or weak? Does the correlation coefficient support what you said about whether there is positive linear correlation, negative linear correlation, or no linear correlation?

15. Use the data from problem number 4 to find the correlation coefficient. According to the correlation coefficient is the correlation strong, moderate, or weak? Does the correlation coefficient support what you said about whether there is positive linear correlation, negative linear correlation, or no linear correlation?

16. Use the data from problem number 5 to find the correlation coefficient. According to the correlation coefficient is the correlation strong, moderate, or weak? Does the correlation coefficient support what you said about whether there is positive linear correlation, negative linear correlation, or no linear correlation?

17. Use the data from problem number 6 to find the correlation coefficient. According to the correlation coefficient is the correlation strong, moderate, or weak? Does the correlation coefficient support what you said about whether there is positive linear correlation, negative linear correlation, or no linear correlation?

Exercise Set 8.1

1. quantitative discrete

3. quantitative continuous

5. qualitative

7.

Number of Devices	Number of families
1	1
2	3
3	6
4	5
5	7
6	5
7	0
8	3

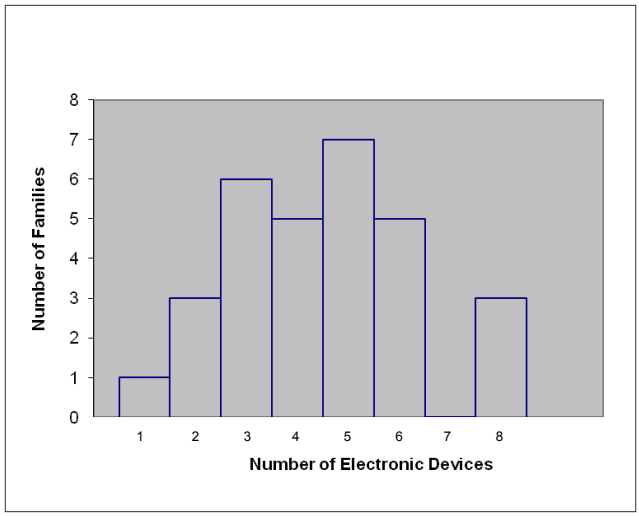
26 families had at least 3 electronic devices in their home.

9.

Number of Sunglasses	Number of Students
0	8
1	7
2	3
3	2

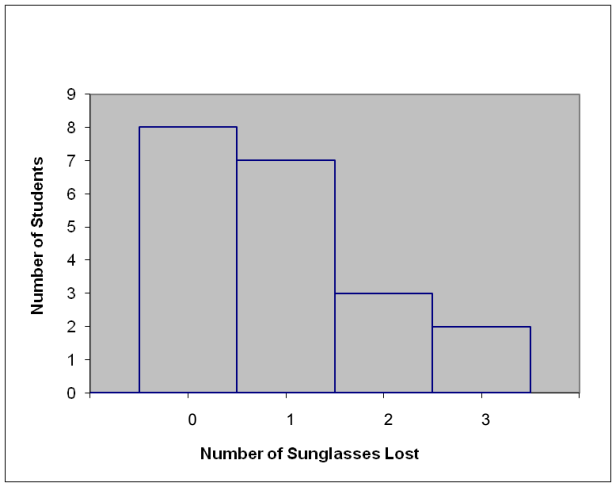
8 students did not lose any sunglasses over the last year.

11.



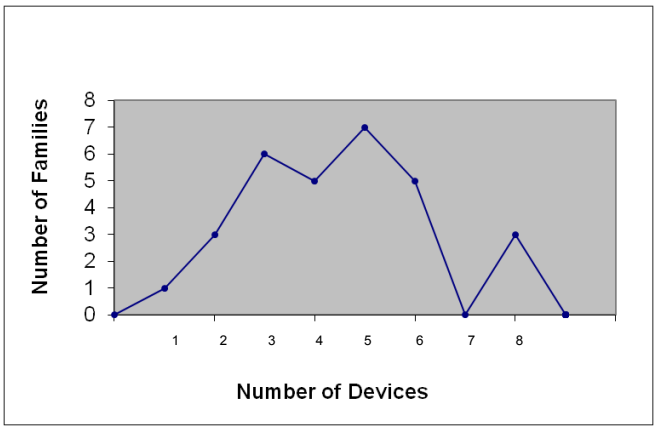
10 families had at most 3 electronic devices in their home.

13.



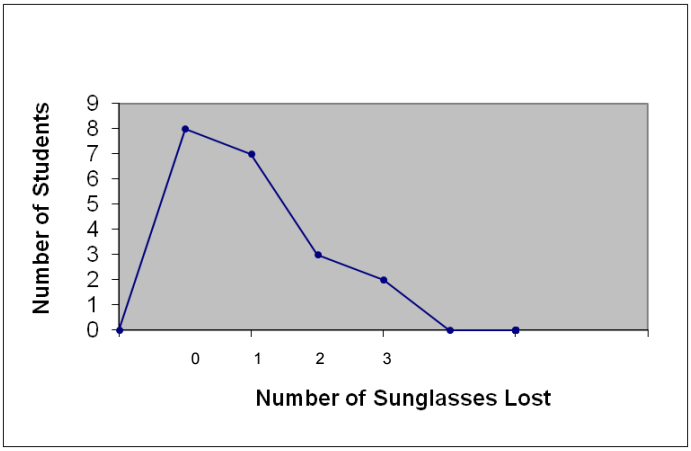
12 students lost at least 1 pair of sunglasses over the last year?

15.



14 families had between 2 and 4 electronic devices inclusive in their home.

17.



All 20 students lost less than 4 pairs of sunglasses over the last year.

19. Frequency distributions, frequency polygons, and histograms all group the raw data into different classes and display the frequency of the different classes. Frequency polygons and histograms display the data as pictures whereas frequency polygons use a table with numbers to show the size of the classes. Histograms use the height of bars to represent the size of the different classes.

Exercise Set 8.2

1. 87 is the mean grade.

3. 90 is the median grade.

5. A score of 80 is the mean grade.

7. A score of 75 is the median grade.

9. Answers will vary. The mean takes the size of all the data values into account in its calculation. The mode is the most frequently occurring value. The median is the "middle" value for all of the values. The midrange is the "middle" value for the lowest and highest values only.

11. 0 absences

13. 3 absences

15. 4.5 is the mode of the scores.

17. 4.15 is the midrange of the scores.

19. 4.4 children

21. 4.5 children

23. 11.4 cartoons of eggs

25. 11.5 cartoons of eggs

27. 2.2 cups of coffee

29. 2 cups of coffee

31. Answers will vary. The mean takes the size of all the data values into account in its calculation. The mode is the most frequently occurring value. The median is the "middle" value for all of the values. The midrange is the "middle" value for the lowest and highest values only.

33. 0 lost pairs of sunglasses

35. 1.5 lost pairs of sunglasses

37. 4 false claims

39. 4.5 false claims

41. 6.4 lost keys

43. 6 lost keys

Exercise Set 8.3

1. The range for the grades is 30.

3. The range for the grades is 35.

5. Set 1: The mean score is 80.
Set 2: The mean score is 80.

7. Set 1: The standard deviation of the test scores is 20.
Set 2: The standard deviation of the test scores is 14.6.

9. The standard deviations show that the data for set 1 is more spread out than the data for set 2 even though the means and ranges are the same..

11. 2.3 absences

13. The standard deviation of the scores is 0.44 .

15. 1.8 lost children

17. 16 eggs

19. 1.5 cups of coffee

21. 0.999 lost pairs of sunglasses

23. 1.75 false claims

25. 1.5 lost keys

Exercise Set 8.4

1. It means that 95% of men are less than 74 inches tall.

3. It means that 92% of women are less than 69 inches tall.

5. It means that 98% of the scores on the test are less than 52.

7. 95%

9. 16%

11. 2.5%

13. 68%

15. 2.5%

17. 16%

19. 2

21. -1.5

23. 1.83

25. 0

27. -0.33

29. 0.9821 or 98.21%

31. 0.0901 or 9.01%

33. 0.3936 or 39.36%

35. 0.9838 or 98.38%

37. 0.6141 or 61.41%

39. 0.8943 or 89.43%

41. 0.5346 or 53.46%

43. 0.9452 or 94.52%

45. 0.9641 or 96.41%

47. 0.0294 or 2.94%

49. 0.2877 or 28.77%

51. 0.4973 or 49.73%

53. 0.4246 or 42.46%

55. 0.903 or 90.3%

57. 0.1151 or 11.51%

59. 0.9452 or 94.52%

61. 0.6962 or 69.62%

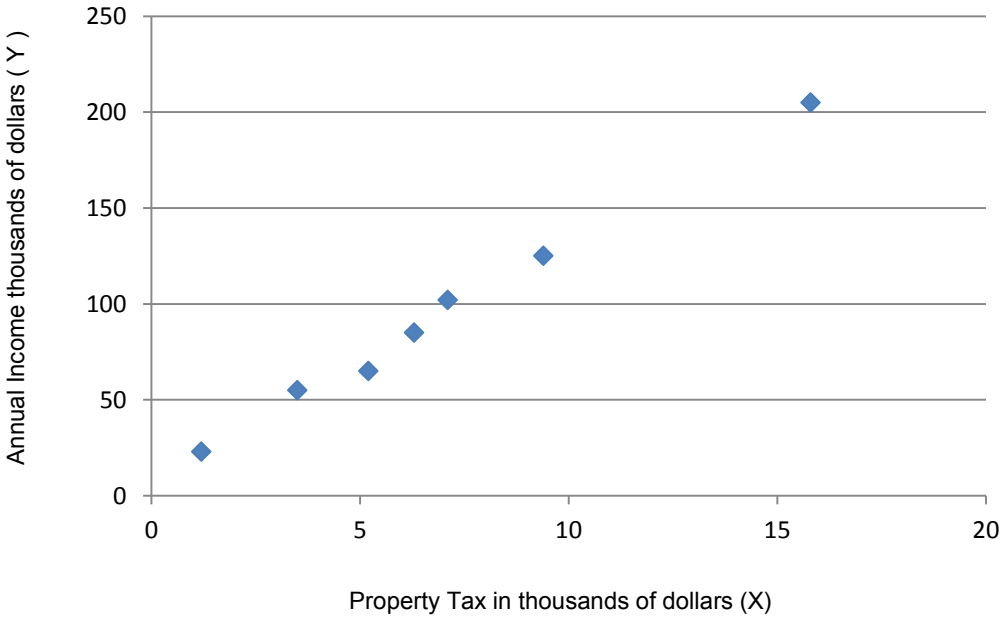
63. 0.0012 or 0.12%

65. 0.0012 or 0.12%

67. 0.3167 or 31.67%

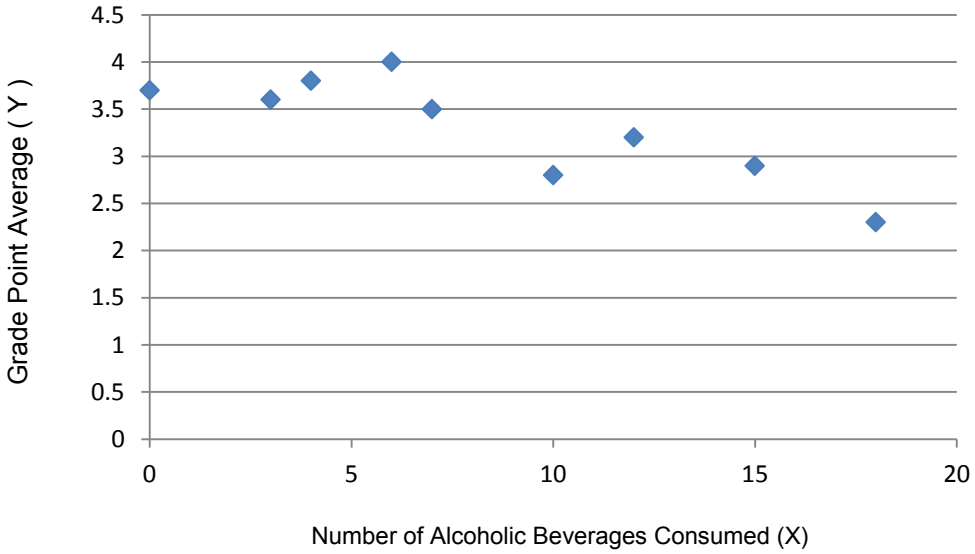
Exercise Set 8.5

1.



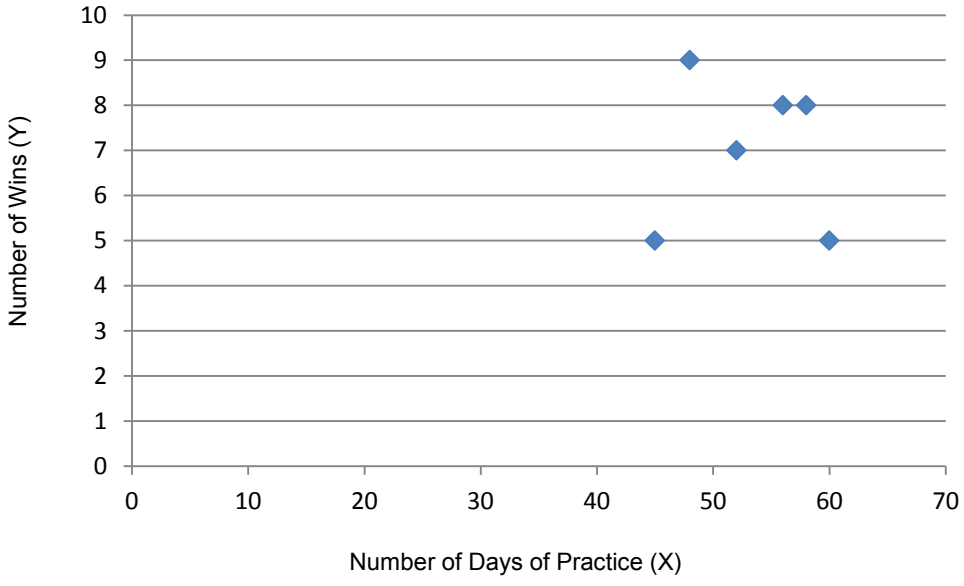
There is a positive linear correlation between property tax and annual income.

3.



There is a negative linear correlation between number of alcoholic beverages consumed and grade point average.

5.



There is no linear correlation between number of days of practice and number of wins.

7. D

9. E

11. A

13. r = -0.96 ; strong negative correlation ; supported by the scatter plot

15. r = 0.69 ; moderate positive linear correlation ; supported by the scatter plot

17. r = 0.915 ; strong positive linear correlation ; support by the scatter plot

Math 103 Formula Sheet

Financial Management

Simple Interest: $\text{Int} = \text{Prt}$

Future Value for Simple Interest: $\text{FV} = \text{P}(1 + \text{rt})$

Future Value for Compound Interest: $\text{FV} = \text{P}\left(1 + \frac{\text{r}}{\text{n}}\right)^{\text{nt}}$

Present Value for Compound Interest: $\text{P} = \frac{\text{FV}}{\left(1 + \frac{\text{r}}{\text{n}}\right)^{\text{nt}}}$

Future Value for continuous compounding: $\text{FV} = \text{Pe}^{\text{r}\cdot\text{t}}$

Effective Annual Yield: $\text{EAY} = \left(1 + \frac{\text{r}}{\text{n}}\right)^{\text{n}} - 1$

Future Value of an Annuity (Pmt is the amount of each deposit): $\text{FV} = \frac{\text{Pmt}\left[\left(1 + \frac{\text{r}}{\text{n}}\right)^{\text{nt}} - 1\right]}{\left(\frac{\text{r}}{\text{n}}\right)}$

Periodic deposits for an Annuity (FV is the future value of the annuity): $\text{Pmt} = \frac{\text{FV}\left(\frac{\text{r}}{\text{n}}\right)}{\left[\left(1 + \frac{\text{r}}{\text{n}}\right)^{\text{nt}} - 1\right]}$

Periodic Mortgage Payments (B is the amount of mortgage): $\text{Pmt} = \frac{\text{B}\left(\frac{\text{r}}{\text{n}}\right)}{\left[1 - \left(1 + \frac{\text{r}}{\text{n}}\right)^{-\text{nt}}\right]}$

Probability and Counting Rules

Permutation rule: ${}_{\text{n}}\text{P}_{\text{k}} = \frac{\text{n}!}{(\text{n} - \text{k})!}$

Combination rule: ${}_{\text{n}}\text{C}_{\text{k}} = \frac{\text{n}!}{(\text{n} - \text{k})!\,\text{k}!}$

$\text{P}(\overline{\text{E}}) = 1 - \text{P}(\text{E})$

$\text{P}(\text{E}) = 1 - \text{P}(\overline{\text{E}})$

$\text{P}(\text{A or B}) = \text{P}(\text{A}) + \text{P}(\text{B})$

$\text{P}(\text{A or B}) = \text{P}(\text{A}) + \text{P}(\text{B}) - \text{P}(\text{A and B})$

$\text{P}(\text{A and B}) = \text{P}(\text{A}) \cdot \text{P}(\text{B})$

$\text{P(B given A)} = \frac{\text{number of common outcomes for B and A}}{\text{number of outcomes within A}}$

Statistics

Mean for individual data: $\bar{\text{x}} = \frac{\sum \text{x}}{\text{n}}$

Mean for grouped data: $\bar{\text{x}} = \frac{\sum \text{f} \cdot \text{x}_{\text{m}}}{\text{n}}$

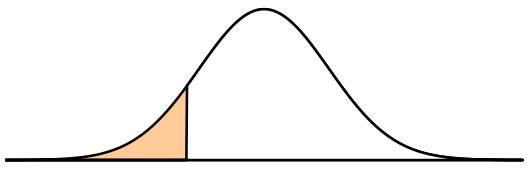
Standard Deviation: $\text{s} = \sqrt{\frac{\sum(\text{x} - \bar{\text{x}})^2}{\text{n} - 1}}$

Z-score: $\text{z} = \frac{\text{x} - \bar{\text{x}}}{\text{s}}$

$\bar{\text{x}}$ = mean x= data values $\sum$ = add all the values f = frequency x_{m}= class or class midpoint s = standard deviation

Standard Normal Distribution Cumulative Probabilities (Percentiles)
Table Values represent area to the left of z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.00	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.90	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.80	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.70	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.60	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.50	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.40	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.30	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.20	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.10	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.00	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.90	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.80	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.70	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.60	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.50	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.40	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.30	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.20	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.10	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.00	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.90	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.80	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.70	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.60	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.50	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.40	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.30	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.20	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.10	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.00	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

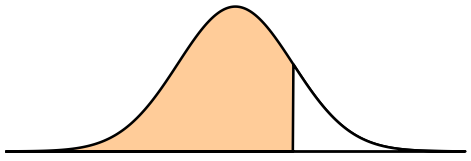


z

For values of z less than -3.09 use 0.0010

Standard Normal Distribution Cumulative Probabilities (Percentiles)
Table Values represent area to the left of z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.10	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.20	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.30	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.40	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.50	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.60	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.70	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.80	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.90	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.00	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.10	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.20	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.30	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.40	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.50	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.60	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.70	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.80	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.90	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.00	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.10	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.20	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.30	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.40	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.50	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.60	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.70	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.80	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.90	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.00	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990



For values of z more than 3.09 use 0.9990

z