SECTION 3.1: MEASURES OF CENTER

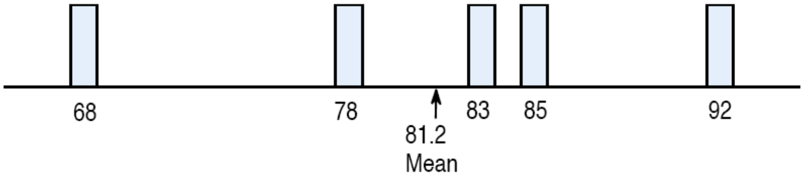
(central measure of tendency)

OBJECTIVES

1. Compute the mean of a data set
2. Compute the median of a data set
3. Compare the properties of the mean and median
4. Find the mode of a data set
5. Approximate the mean with grouped data

OBJECTIVE 1
COMPUTE THE MEAN OF A DATA SET

The ___mean____________ of a data set is a measure of center. If we imagine each data value to be a weight, then the mean is the point at which the data set balances.



NOTATION – POPULATION VERSUS SAMPLE

Recall that a population consists of an entire collection of individuals about which information is sought, and a sample consists of a smaller group drawn from the population. The method for calculating the mean is the same for both samples and populations, except for the notation.

NOTATION:

Population Mean: μ (mu)
Sample Mean: $\bar{x}$ (x bar)

Computing the Mean

A list of n numbers is denoted by $x_1, x_2, \ \dots, \ x_n$
$\sum x_i$ represents the sum of these numbers: $\sum x_i = x_1 + x_2 + x_3 + \ \dots + x_n$ Σ sigma summation

If $x_1, x_2, \ \dots, \ x_n$ is a sample, then the **sample mean** is given by $\bar{x} = \dfrac{\Sigma x_i}{n}$

If $x_1, x_2, \ \dots, \ x_N$ is a population, then the **population mean** is given by $\mu = \dfrac{\Sigma x_i}{N}$

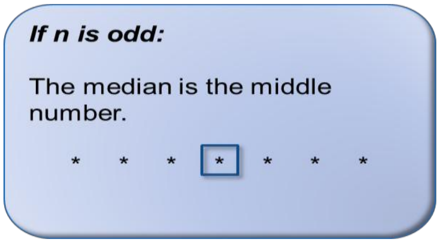
EXAMPLE: During a semester, a student took five exams. The population of exam scores is 78, 83, 92, 68, and 85. Find the mean.

SOLUTION: The population mean is given by

$$\mu = \frac{\sum x_i}{N} = \frac{(78 + 83 + 92 + 68 + 85)}{5} = 81.2$$

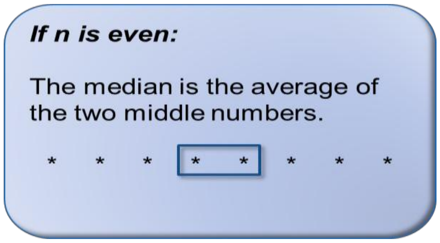
OBJECTIVE 2
COMPUTE THE MEDIAN OF A DATA SET

The median is another measure of center. The median is a number that splits the data set in half, so that half the data values are less than the median and half of the data values are greater than the median. The procedure for computing the median differs, depending on whether the number of observations in the data set is even or odd.



If n is odd:

The median is the middle number.



If n is even:

The median is the average of the two middle numbers.

EXAMPLE: During a semester, a student took five exams. The population of exam scores is 78, 83, 92, 68, and 85. Find the median of the exam scores.

SOLUTION: First arrange the data values in increasing order

68, 78, 83, 85, 92

The median is 83

EXAMPLE: Eight patients undergo a new surgical procedure and the number of days spent in recovery for each is as follows. Find the median number of days in recovery.

20 15 12 27 13 19 13 21

SOLUTION: First arrange the data values in increasing order

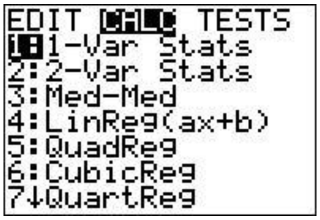
12, 13, 13, 15, 19, 20, 21, 27

The median is the average of the two middle numbers.

$$\text{median} = \frac{(15 + 19)}{2} = 17$$

MEAN AND MEDIAN ON THE TI-84 PLUS

The **1-Var Stats** command in the TI-84 PLUS Calculator displays a list of the most common parameters and statistics for a given data set. This command is accessed by pressing **STAT** and then highlighting the **CALC** menu.



The 1-Var Stats command returns the following quantities:

$\bar{x}$	The mean	minX	The minimum data value
$\sum x$	The sum of all data values	Q_1	The first quartile
$\sum x^2$	The sum of the square of all data values	Med	The median
Sx	The sample standard deviation	Q_3	The third quartile
σx	The population standard deviation	maxX	The maximum data value
n	The number of data values		

EXAMPLE:

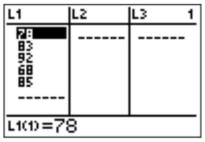
During a semester, a student took five exams. The population of exam scores is 78, 83, 92, 68, and 85. Find the mean and median using the TI-84 PLUS.

Step 1: Enter the data in **L1**.

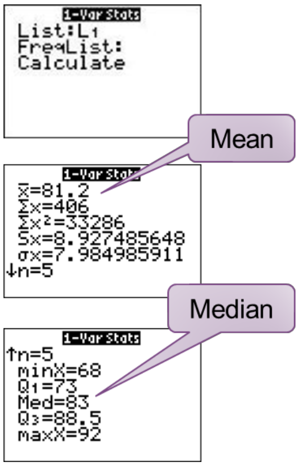
Step 2: Press **STAT** and highlight the **CALC** menu.

Step 3: Select **1-Var Stats** and press **ENTER**. Enter **L1** in the **List** field and run the command.

Note: If your calculator does not support Stat Wizards, enter L1 next to the 1-Var Stats command on the home screen and press enter to run the command



```
L1     L2     L3     1
78     ------ ------
83
92
68
85
------
L1(1)=78
```



```
1-Var Stats
List:L1
FreqList:
Calculate
```

```
1-Var Stats
x̄=81.2
Σx=406
Σx²=33286
Sx=8.927485648
σx=7.984985911
↓n=5
```

```
1-Var Stats
↑n=5
 minX=68
 Q1=73
 Med=83
 Q3=88.5
 maxX=92
```

OBJECTIVE 3
COMPARE THE PROPERTIES OF THE MEAN AND MEDIAN

A statistic is resistant if its value is not affected much by extreme values (large or small) in the data set. The median is resistant, but the mean is not.

EXAMPLE: Five families have annual incomes of $25,000, $31,000, $34,000, $44,000 and $56,000. One family, whose income is $25,000, wins a million dollar lottery, so their income increases to $1,025,000.

Before lottery win

mean = 38000

median = 34000

After lottery win

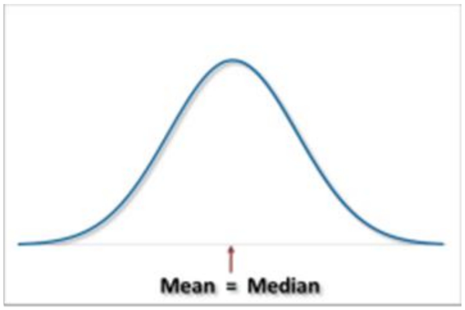
mean = 238000

median = 44000

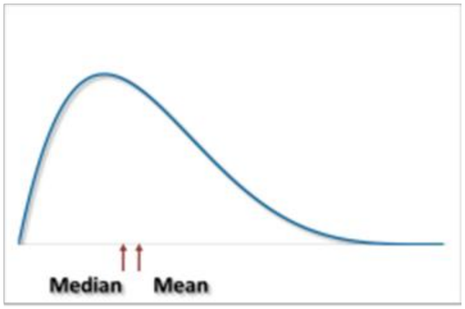
The extreme value of $1025000 influences the mean quite a lot; increasing it from $38000 to $ 238000. In comparison, the median has been influenced much less increasing from $ 34000 to $ 44000. This shows that the median is resistant and the mean is not.

MEAN, MEDIAN, AND THE SHAPE OF A DATA SET

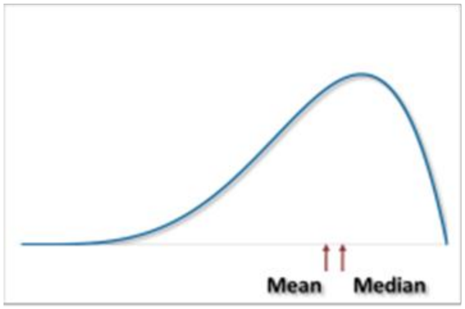
The mean and median measure the center of a data set in different ways. When a data set is **symmetric**, the mean and median are equal.



When a data set is **skewed to the right**, there are large values in the right tail. Because the median is resistant while the mean is not, the mean is generally more affected by these large values. Therefore for a data set that is skewed to the right, the mean is often greater than the median.



Similarly, when a data set is **skewed to the left**, the mean is often less than the median.



OBJECTIVE 4
FIND THE MODE OF A DATA SET

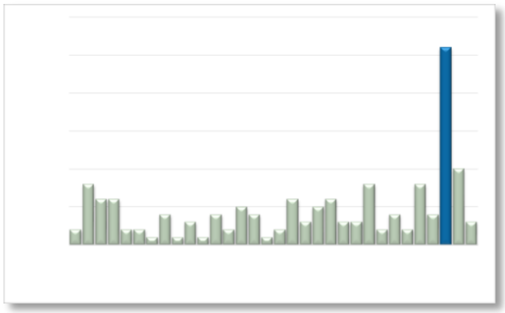
Another value that is sometimes classified as a measure of center is the mode.

- The mode of a data set is the value that appear most frequently.
- If two or more values are tied for the most frequent, they are all considered to be modes.
- If the values all have the same frequency, we say that the data set has no mode.

EXAMPLE: Ten students were asked how many siblings they had. The results, arranged in order, were 0 1 1 1 1 2 2 3 3 6. Find the mode of this data set.

SOLUTION: The mode is 1.

The mode is sometimes classified as a measure of center. However, this isn't really accurate. The mode can be the largest value in a data set, or the smallest, or anywhere in between.



MODE FOR QUALITATIVE DATA

The mean and median can be computed only for quantitative data. The mode, on the other hand, can be computed for qualitative data as well. For qualitative data, the mode is the most frequently appearing category.

EXAMPLE: Following is a list of the makes of all the cars rented by an automobile rental company on a particular day. Which make of car is the mode?

Honda	Toyota	Toyota	Honda	Ford
Chevrolet	Nissan	Ford	Chevrolet	Chevrolet
Honda	Dodge	Ford	Ford	Toyota
Chevrolet	Toyota	Toyota	Toyota	Nissan

SOLUTION: The mode is "Toyota"

OBJECTIVE 5
APPROXIMATE THE MEAN USING GROUPED DATA

Sometimes we don't have access to the raw data in a data set, but we are given a frequency distribution. In these cases we can approximate the mean using the following steps.

Step 1: Compute the midpoint of each class. The midpoint of a class is found by taking the average of the lower class limit and the lower limit of the next larger class.

Step 2: For each class, multiply the class midpoint by the class frequency.

Step 3: Add the products (Midpoint)x(Frequency) over all classes.

Step 4: Divide the sum obtained in Step 3 by the sum of the frequencies.

$$\bar{x} = \frac{\Sigma\,(\text{midpoint})(\text{frequency})}{\Sigma\ \text{frequency}}$$

EXAMPLE: The following table presents the number of text messages sent via cell phone by a sample of 50 high school students. Approximate the mean number of messages sent.

Number of Text Messages Sent	Frequency
0 – 49	10
50 – 99	5
100 – 149	13
150 – 199	11
200 – 249	7
250 – 299	4

SOLUTION:

midpoint	frequency
$\frac{(0+50)}{2} = 25$	10
$= 75$	5
125	13
175	11
225	7
275	4

To approximate the mean, we use

$$\bar{x} \approx \frac{\sum (\text{midpoint})(\text{freq})}{\sum \text{freq}}$$

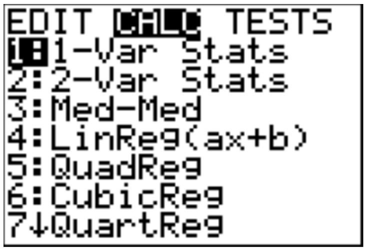
$$= \frac{25(10) + 75(5) + 125(13) + 175(11) + 225(7) + 275(4)}{50}$$

$$= \frac{6850}{50} = 137$$

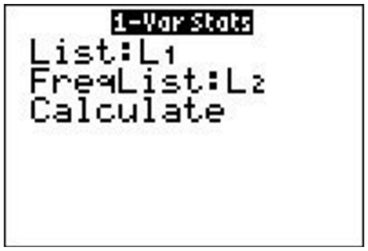
The mean number of messages sent is approximately 137.

GROUPED DATA ON THE TI-84 PLUS

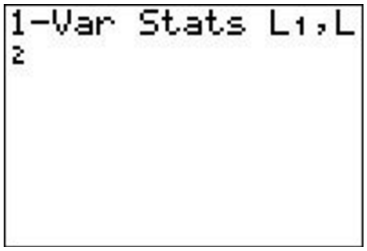
To compute the mean for grouped data in a frequency distribution, enter the midpoint for each class into **L1** and the corresponding frequencies in **L2**. Next, select the **1-Var Stats** command and enter **L1** in the **List** field and **L2** in the **FreqList** field, if using Stats Wizards. If you are not using Stats Wizards, you may run the**1-Var Stats** command followed by **L1**, **comma**, **L2**.



```
EDIT CALC TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7↓QuartReg
```



```
1-Var Stats
List:L1
FreqList:L2
Calculate
```



```
1-Var Stats L1,L
2
```

YOU SHOULD KNOW …

- How to compute and interpret the mean of a data set
- The notation for a population mean and sample mean
- How to compute the median
- How to use the TI-84 PLUS calculator to compute the mean and median
- The definition of *resistant* and which measure of center is resistant
- How the mean and median are related to the shape of a data set including
  - Skewed to the left
  - Skewed to the right
  - Approximately symmetric
- How to identify the mode of a data set
- How to approximate the mean for grouped data

SECTION 3.2: MEASURES OF SPREAD

OBJECTIVES

1. Compute the range of a data set
2. Compute the variance of a population and a sample
3. Compute the standard deviation of a population and a sample
4. Approximate the standard deviation with grouped data
5. Use the Empirical Rule to summarize data that are unimodal and approximately symmetric
6. Use Chebyshev's Inequality to describe a data set

OBJECTIVE 1
COMPUTE THE RANGE OF A DATA SET

The ___range___ of a data set is the difference between the largest value and the smallest value.

EXAMPLE: The average monthly temperatures, in degrees Fahrenheit, for San Francisco are

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
San Francisco	51	54	55	56	58	60	60	61	63	62	58	52

The range of temperatures is: ___63 − 51 = 12___.

Although the range is easy to compute, it is not often used in practice. The reason is that the range involves only two values from the data set: the largest and smallest.

OBJECTIVE 2
COMPUTE THE VARIANCE OF A POPULATION AND A SAMPLE

VARIANCE

When a data set has a small amount of spread, like the San Francisco temperatures, most of the values will be close to the mean. When a data set has a larger amount of spread, more of the data values will be far from the mean. The variance is a measure of how far the values in a data set are from the mean, on the average. The variance is computed *slightly differently* for populations and samples. The population variance is presented first.

Let $x_1, x_2, x_3, \dots, x_N$ denote the values in a population of size N. Let μ denote the population mean. The population variance, denoted by σ^2, is

sigma

POPULATION VARIANCE:

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N}$$

EXAMPLE: Compute the population variance for the San Francisco temperatures:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
San Francisco	51	54	55	56	58	60	60	61	63	62	58	52

↑ x

SOLUTION: Compute the population mean μ: $\mu = \frac{\sum x_i}{N} = \frac{(51+54 \ldots +52)}{12} = 57.5$

x_i	$(x_i-\mu)^2$
51	42.25
54	12.25
55	6.25
56	2.25
58	0.25
60	6.25
60	6.25
61	12.25
63	30.25
62	20.25
58	0.25
52	30.25

total $= \frac{\sum (x_i-\mu)^2}{N} = 169$

$$\sigma^2 = \frac{\sum (x_i-\mu)^2}{N} = \frac{169}{12} = 14.083$$

The population varriance for the San Francisco temperatures is 14.083

Sample Variance

When the data values come from a *sample* rather than a population, the variance is called the sample variance. The procedure for computing the sample variance is a bit different from the one used to compute a population variance. In the formula, the mean μ is replaced by the sample mean $\bar{x}$ and the denominator is $n-1$ instead of N. The sample variance is denoted by s^2.

SAMPLE VARIANCE:

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

WHY DIVIDE BY $n-1$?

$(x_i - \bar{x})$

When computing the sample variance, we use the sample mean to compute the deviations. For the population variance we use the population mean for the deviations. It turns out that the deviations using the sample mean tend to be a *bit smaller* than the deviations using the population mean. If we were to divide by n when computing a sample variance, the value would tend to be a bit smaller than the population variance. It can be shown mathematically that the appropriate correction is to divide the sum of the squared deviations by $n-1$ rather than n.

EXAMPLE: A company that manufactures batteries is testing a new type of battery designed for laptop computers. They measure the lifetimes, in hours, of six batteries, and the results are 3, 4, 6, 5, 4, 2. Find the sample variance of the lifetimes.

SOLUTION: The sample mean is $\bar{x} = \frac{(3+4+6+5+4+2)}{6} = 4$

The sample variance of the lifetime is

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{(3-4)^2 + (4-4)^2 + (6-4)^2 + (5-4)^2 + (4-4)^2 + (2-4)^2}{(6-1)}$$

$= \frac{10}{5} = 2$ The sample variance of the liftetime is 2

OBJECTIVE 3
COMPUTE THE STANDARD DEVIATION OF A POPULATION AND A SAMPLE

STANDARD DEVIATION

Because the variance is computed using squared deviations, the units of the variance are the squared units of the data. For example, in the Battery Lifetime example, the units of the data are hours, and the units of variance are squared hours. In most situations, it is better to use a measure of spread that has the same units as the data.

We do this simply by taking the square root of the variance. This quantity is called the standard deviation. The standard deviation of a sample is denoted s, and the standard deviation of a population is denoted by σ.

SAMPLE STANDARD DEVIATION:

$$s = \sqrt{s^2}$$

POPULATION STANDARD DEVIATION:

$$\sigma = \sqrt{\sigma^2}$$

EXAMPLE: The population variance of temperatures in San Francisco is $\sigma^2 = 14.083$. Find the population standard deviation.

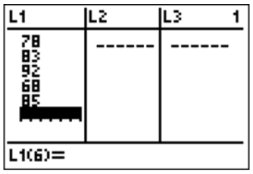
SOLUTION: The population standard deviation is $\sigma = \sqrt{\sigma^2} = \sqrt{14.083} = 3.753$

EXAMPLE: The variance of the lifetimes for a sample of six batteries $s^2 = 2$. Find the sample standard deviation.

SOLUTION: The sample standard deviation is $s = \sqrt{s^2} = \sqrt{2} = 1.414$

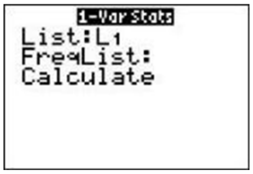
STANDARD DEVIATION ON THE TI-84 PLUS

The following steps will compute the standard deviation for both sample data and population data on the TI-84 PLUS Calculator:



```
L1      |L2      |L3     1
 78     | ------ | ------
 83     |        |
 92     |        |
 68     |        |
 85     |        |
L1(6)=
```

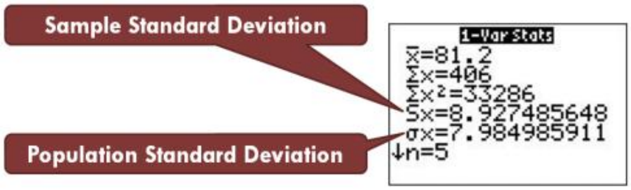
Step 1: Enter the data in **L1**.



```
1-Var Stats
List:L1
FreqList:
Calculate
```

Step 2: Run the **1-Var Stats** command (the same command used for means and medians), selecting **L1** as the location of the data.

Note: If your calculator does not support Stat Wizards, enter L1 next to the 1-Var Stats command on the home screen and press enter to run the command



```
1-Var Stats
x̄=81.2
Σx=406
Σx²=33286
Sx=8.927485648
σx=7.984985911
↓n=5
```

STANDARD DEVIATION AND RESISTANCE

Recall that a statistic is **resistant** if its value is not affected much by extreme values (large or small) in the data set. The standard deviation is not resistant.

That is, the standard deviation is affected by extreme values.

OBJECTIVE 4
APPROXIMATE THE STANDARD DEVIATION USING GROUPED DATA

STANDARD DEVIATION

Sometimes we don't have access to the raw data in a data set, but we are given a frequency distribution. In these cases we can approximate the standard deviation using the following steps.

Step 1: Compute the midpoint of each class and approximate the mean of the frequency distribution.

Step 2: For each class, subtract the mean from the class midpoint to obtain (Midpoint – Mean).

Step 3: For each class square the difference obtained in Step 2 to obtain $(\text{Midpoint} - \text{Mean})^2$, and multiply by the frequency to obtain
$(\text{Midpoint} - \text{Mean})^2$ x (Frequency).

Step 4: Add the products $(\text{Midpoint} - \text{Mean})^2$ x (Frequency) over all classes.

Step 5: To compute the *population* variance, divide the sum obtained in Step 4 by n. To compute the *sample* variance, divide the sum obtained in Step 4 by $n - 1$.

Step 6: Take the square root of the variance obtained in Step 5. The result is the standard deviation.

Population variance:

$$\sigma^2 = \frac{\Sigma\,(\text{midpoint} - \text{mean})^2 \cdot \text{freq}}{n}$$

sample variance:

$$s^2 = \frac{\Sigma\,(\text{midpoint} - \text{mean})^2 \cdot \text{freq}}{n-1}$$

EXAMPLE: The following table presents the number of text messages sent via cell phone by a sample of 50 high school students. Approximate the sample standard deviation number of messages sent.

Number of Text Messages Sent	Frequency
0 – 49	10
50 – 99	5
100 – 149	13
150 – 199	11
200 – 249	7
250 – 299	4

SOLUTION: Recall from the last section that the sample mean $\bar{x} = 137$

L_3 = freq

L_1	$L_1 - 137$	$L_2^2 \cdot L_3$
midpoints	midpoint − mean	$(\text{midpoint} - \text{mean})^2 \cdot \text{freq}$
25	−112	$(-112)^2 \cdot 10 = 125440$
75	−62	19220
125	−12	1872
175	38	15884
225	88	54208
275	138	76176

$$\Sigma\,(\text{midpoint} - \text{mean})^2 \cdot \text{freq} = 292800$$

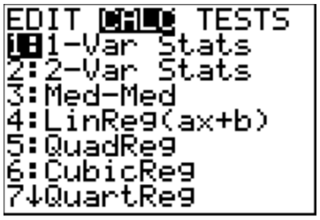
The sample standard deviation is approximately

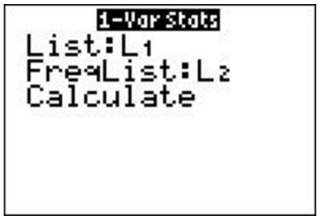
$$s^2 = \frac{\Sigma\,(\text{midpoint} - \text{mean})^2 \cdot \text{freq}}{n-1} = \frac{292800}{(50-1)} = 5975.510204$$

$$s = \sqrt{s^2} = \sqrt{5975.510204} = 77.30142$$

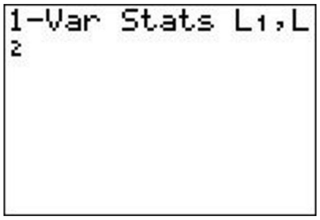
GROUPED DATA ON THE TI-84 PLUS

The same procedure used to compute the mean for grouped data in a frequency distribution may be used to compute the standard deviation. Enter the midpoint for each class into **L1** and the corresponding frequencies in **L2**. Next, select the **1-Var stats** command and enter **L1** in the **List** field and **L2** in the **FreqList** field, if using Stats Wizards. If you are not using Stats Wizards, you may run the**1-Var Stats** command followed by **L1**, **comma**, **L2.**





```
1-Var Stats
List:L1
FreqList:L2
Calculate
```

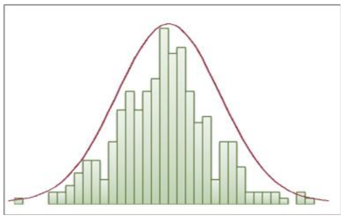


```
1-Var Stats L1,L
2
```

OBJECTIVE 5
USE THE EMPIRICAL RULE TO SUMMARIZE DATA THAT ARE UNIMODAL AND APPROXIMATELY SYMMETRIC

BELL-SHAPED HISTOGRAMS

Many histograms have a single mode near the center of the data, and are approximately symmetric. Such histograms are often referred to as bell-shaped.



THE EMPIRICAL RULE

When a data set has a bell-shaped histogram, it is often possible to use the standard deviation to provide an approximate description of the data using a rule known as **The Empirical Rule**.

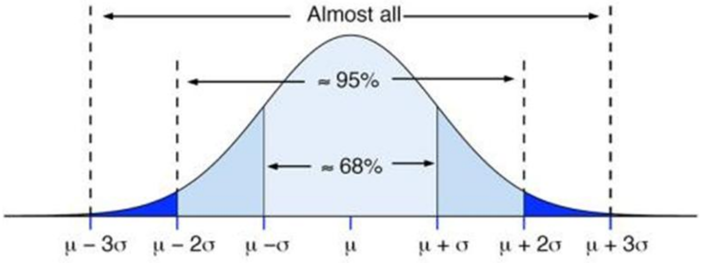
THE EMPIRICAL RULE

When a population has a histogram that is approximately bell-shaped, then:

Approximately 68 % of the data will be within one standard deviation of. the mean

Approximately 95 % of the data will be within two standard deviation of the mean

All, or almost all of the data will be within three standard deviation of. the mean



EXAMPLE: The following table presents the U.S. Census Bureau projection for the percentage of the population aged 65 and over for each state and the District of Columbia. Use the Empirical Rule to describe the data.

14.1	14.3	14.4	17.8	12.0	14.9	12.6	13.7	12.8	13.8	13.7	12.4	13.8
14.1	13.3	14.3	16.0	8.1	11.5	14.1	10.2	12.4	13.4	15.6	12.8	13.9
12.3	14.1	15.3	13.0	13.6	10.5	12.4	13.5	13.9	10.7	11.5	14.3	12.7
13.1	12.2	12.4	15.0	12.6	13.6	13.7	15.5	14.6	9.0	12.2	14.0	

SOLUTION: First, notice if we construct the histogram, this is approximately bell-shaped. Thus we can use the Empirical Rule to describe the data.

From TI-84, we get $\mu = 13.249$ and $\sigma_x = 1.6827$

$\mu - \sigma = 13.249 - 1.6827 \approx 11.57$ } approximately 68% of the data values are between 11.57 and 14.93
$\mu + \sigma = 13.249 + 1.6827 = 14.93$

$\mu - 2\sigma \approx 13.249 - 2(1.6827) = 9.88$ } approximately 95% of the data values are between 9.88 and 16.61
$\mu + 2\sigma \approx 13.249 + 2(1.6827) = 16.61$

$\mu - 3\sigma = 8.2$ } almost all of the data values are between 8.2 and 18.3
$\mu - 3\sigma = 18.3$

OBJECTIVE 6
USE CHEBYSHEV'S INEQUALITY TO DESCRIBE A DATA SET

When a distribution is bell-shaped, we use __the Empirical Rule__ to approximate the proportion of data within one or two standard deviations. Another rule called

__Chebyshev's Inequality__ holds for *any* data set.

CHEBYSHEV'S INEQUALITY:

In *any* data set, the proportion of the data that is within *K* standard deviations of the mean is at least $\mathbf{1 - 1/K^2}$. Specifically, by setting *K* = 2 or *K* = 3, we obtain the following results. $K > 1$

__At least $1 - 1/2^2 = 3/4$ or 75%__, of the data are within __two__ standard deviations of the mean.

__At least $1 - 1/3^2 = 8/9$ or 89%__, of the data are within __three__ standard deviations of the mean.

EXAMPLE: As part of a public health study, systolic blood pressure was measured for a large group of people. The mean was 120 and the standard deviation was 10. What information does Chebyshev's Inequality provide about these data? Notice that the shape of distribution is not given

SOLUTION: $\bar{x} = 120$ and $s = 10$

for $K = 2$, $1 - 1/K^2 = 1 - 1/2^2 = 3/4 = 0.75$. We also know $\bar{x} - 2(s) = 120 - 2(10) = 100$ and $\bar{x} + 2(s) = 120 + 2(10) = 140$. Thus, at least 75 % of the people had systolic blood pressure between 100 and 140

for $K = 3$, $1 - 1/3^2 = 8/9 = 0.889 \approx 0.89$, we also know $\bar{x} - 3(s) = 120 - 3(10) = 90$ and $\bar{x} + 3(s) = 120 + 3(10) = 150$. Thus, at least 89 % of the people had systolic blood pressure between 90 and 150

YOU SHOULD KNOW …

- How to compute the range of a data set
- The notation for population variance, population standard deviation, sample variance, and sample standard deviation
- How to compute the variance and the standard deviation for populations and samples
- How to use the TI-84 PLUS calculator to compute the variance and standard deviation for populations and samples
- How to approximate the standard deviation for grouped data
- How to use The Empirical Rule to describe a bell-shaped data set
- How to use Chebyshev's Inequality to describe any data set
- How to compute and interpret the coefficient of variation

SECTION 3.3: MEASURES OF POSITION

OBJECTIVES

1. Compute and interpret z-scores
2. Compute the quartiles of a data set
3. Compute the percentiles of a data set
4. Compute the five-number summary for a data set
5. Understand the effects of outliers
6. Construct boxplots to visualize the five-number summary and outliers

OBJECTIVE 1
COMPUTE AND INTERPRET z-SCORES

Z-SCORE

Who is taller, a man 73 inches tall or a woman 68 inches tall? The obvious answer is that the man is taller. However, men are taller than women on the average. Suppose the question is asked this way: Who is taller relative to their gender, a man 73 inches tall or a woman 68 inches tall? One way to answer this question is with a z-score.

The z-score of an individual data value tells how many standard deviations that value is from its population mean. For example, a value one standard deviation above the mean has a z-score of $z = 1$ and a value two standard deviations below the mean has a z-score of $z = -2$.

Let x be a value from a population with mean μ and standard deviation σ.

The z-score for x is $z = \frac{x - \mu}{\sigma}$. $\frac{\text{value} - \text{mean}}{\text{standard deviation}}$

EXAMPLE: A National Center for Health Statistics study states that the mean height for adult men in the U.S. is μ = 69.4 inches, with a standard deviation of σ = 3.1 inches. The mean height for adult women is μ = 63.8 inches, with a standard deviation of σ = 2.8 inches. Who is taller relative to their gender, a man 73 inches tall, or a woman 68 inches tall?

SOLUTION: First, we standardize each height by using z-scores and then compare them.

z man $= \frac{x - \mu}{\sigma} = \frac{(73 - 69.4)}{3,1} = 1.16$

z woman $= \frac{x - \mu}{\sigma} = \frac{(68 - 63.8)}{2,8} = 1,5$

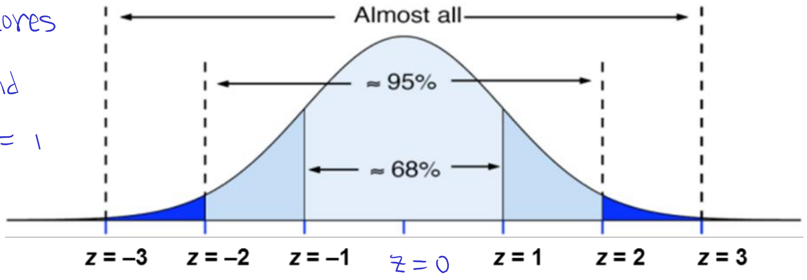
The woman is relatively taller than the man because her z score is higher

Z-SCORES & THE EMPIRICAL RULE

Since the z-score is the number of standard deviations from the mean, we can easily interpret the z-score for bell-shaped populations using The Empirical Rule.
When a population has a histogram that is approximately bell-shaped, then

- Approximately 68% of the data will have z-scores between –1 and 1.
- Approximately 95% of the data will have z-scores between –2 and 2.
- All, or almost all of the data will have z-scores between –3 and 3.

Note that for z scores the mean $\mu = 0$ and standard deviation $\sigma = 1$



OBJECTIVE 2
COMPUTE THE QUARTILES OF A DATA SET

QUARTILES

In a previous section, we learned how to compute the mean and median of a data set as measures of the center. Sometimes, it is useful to compute measures of position other than the center to get a more detailed description of the distribution. **Quartiles** divide a data set into four approximately equal pieces.

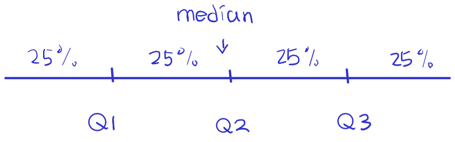
QUARTILES:

Every data set has three quartiles:

The first quartile, denoted Q_1 separates the lowest 25% of the data from the highest 75%.

The second quartile, denoted Q_2 separates the lowest 50% of the data from the highest 50%. Q_2 is the same as the median.

The third quartile, denoted Q_3 separates the lowest 75% of the data from the highest 25%.



COMPUTING QUARTILES

There are several methods for computing quartiles, all of which give similar results. The following procedure is one fairly straightforward method:

Step 1: Arrange the data in increasing order.

Step 2: Let n be the number of values in the data set. To compute the second quartile, simply compute the median. For the first or third quartiles, proceed as follows:

For the first quartile, compute $L = 0.25n$

For the third quartile, compute $L = 0.75n$

Step 3: If L is a whole number, the quartile is the average of the number in position L and the number in position $L + 1$.

If L is not a whole number, round it *up* to the next higher whole number. The quartile is the number in the position corresponding to the rounded-up value.

EXAMPLE: The following table presents the annual rainfall, in inches, in Los Angeles during the month of February from 1969 to 2013. Compute the quartiles for the data.

0.00	0.08	0.13	0.14	0.16	0.17	0.20	0.29	0.56	0.67	0.70	0.92	1.22	1.30	1.48
1.64	1.72	1.90	2.37	2.58	2.84	3.06	3.12	3.21	3.29	3.54	3.57	3.71	4.13	4.27
4.37	4.64	4.89	4.94	5.54	6.10	6.61	7.89	7.96	8.03	8.87	8.91	11.02	12.75	13.68

SOLUTION: Since $n = 45$, for the first quartile Q_1, compute $L = 0.25n = 0.25(45) = 11.25$ Since 11.25 is not a whole number, we round it up to 12. Thus Q_1 is the number in the 12th position, $Q_1 = 0.92$. We find Q_2 by using the methods previously presented. Q_2 or median is 3.12

for Q_3, we compute $L = 0.75(45) = 33.75$, so we round it up to 34

QUARTILES ON THE TI-84 PLUS

The **1-Var Stats** command in the TI-84 PLUS Calculator displays a list of the most common parameters and statistics for a given data set. This command is accessed by pressing **STAT** and then highlighting the **CALC** menu.

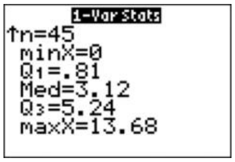
EXAMPLE: QUARTILES ON THE TI-84 PLUS

Step 1: Enter the data in **L1**.

Step 2: Press **STAT** and highlight the **CALC** menu.

Step 3: Select **1-Var Stats** and press **ENTER**. Enter **L1** in the **List** field and run the command.

Note: If your calculator does not support Stat Wizards, enter L1 next to the 1-Var Stats command on the home screen and press enter to run the command

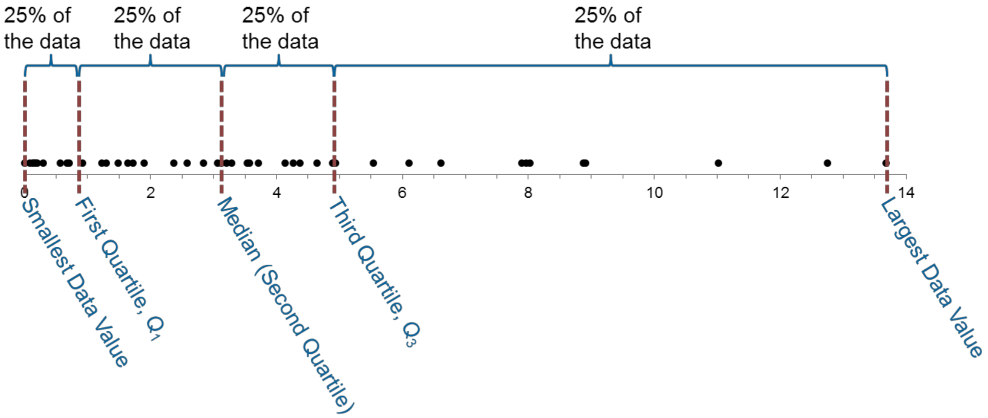


```
1-Var Stats
↑n=45
 minX=0
 Q1=.81
 Med=3.12
 Q3=5.24
 maxX=13.68
```

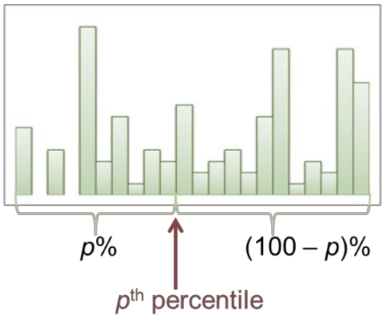
The quartile values produced by the TI-84 PLUS may differ from results obtained by hand because it uses a slightly different procedure than the one described in the text.

VISUALIZING THE QUARTILES

Following is a dotplot of the Los Angeles rainfall data with the quartiles indicated. The quartiles divide the data set into four parts, with approximately 25% of the data in each part.



OBJECTIVE 3
COMPUTE THE PERCENTILES OF A DATA SET



Quartiles describe the shape of a distribution by dividing it into fourths.

Sometimes it is useful to divide a data set into a greater number of pieces to get a more detailed description of the distribution.

___________________________ divide a data set into hundredths.

PERCENTILES:

For a number *p* between 1 and 99, the *p*th percentile separates the lowest *p*% of the data from the highest (1–*p*)%.

COMPUTING PERCENTILES

The following procedure computes the *p*th percentile of a data set:

Step 1: Arrange the data in increasing order.

Step 2: Let n be the number of values in the data set. For the *p*th percentile, compute $L = \left(\frac{p}{100}\right)n$.

Step 3: If L is a whole number, the *p*th percentile is the average of the number in position L and the number in position $L + 1$.
If L is not a whole number, round it *up* to the next higher whole number. The *p*th percentile is the number in the position corresponding to the rounded-up value.

EXAMPLE: The following table presents the annual rainfall, in inches, in Los Angeles during the month of February from 1969 to 2013. Compute the 60th percentile for the data.

0.00	0.08	0.13	0.14	0.16	0.17	0.20	0.29	0.56	0.67	0.70	0.92	1.22	1.30	1.48
1.64	1.72	1.90	2.37	2.58	2.84	3.06	3.12	3.21	3.29	3.54	3.57	3.71	4.13	4.27
4.37	4.64	4.89	4.94	5.54	6.10	6.61	7.89	7.96	8.03	8.87	8.91	11.02	12.75	13.68

SOLUTION: For the 60 th percentile, we compute $L = (\frac{p}{100})n = (\frac{60}{100})45 = 27.$

since 27 is a whole number, the 60 th percentile is the average of the numbers in the 27 th and 28 th position

That is, $P_{60} = \frac{(3,57 + 3,71)}{2} = 3,64$

Interpretation: 60 % of the numbers in the data set are below 3.64

COMPUTING A PERCENTILE FROM A GIVEN DATA VALUE

Sometimes we are given a value from a data set and wish to compute the percentile corresponding to that value. Following is the procedure for doing this:

Step 1: Arrange the data in increasing order.

Step 2: Let x be the data value whose percentile is to be computed. Use the following formula to compute the percentile:

$$\text{Percentile} = 100 \cdot \frac{(Number\ of\ values\ less\ than\ x) + 0.5}{Number\ of\ values\ in\ the\ data\ set}$$

Round the result to the nearest whole number. This is the percentile corresponding to the value x.

EXAMPLE: The following table presents the annual rainfall in Los Angeles during February from 1969 to 2013. In 1989, the rainfall was 1.90. What percentile does this correspond to?

0.00	0.08	0.13	0.14	0.16	0.17	0.20	0.29	0.56	0.67	0.70	0.92	1.22	1.30	1.48
1.64	1.72	1.90	2.37	2.58	2.84	3.06	3.12	3.21	3.29	3.54	3.57	3.71	4.13	4.27
4.37	4.64	4.89	4.94	5.54	6.10	6.61	7.89	7.96	8.03	8.87	8.91	11.02	12.75	13.68

SOLUTION: There are 17 values less than 1.90

$$\text{Percentile} = 100\left[\frac{((\#\text{ of values less than }1.90) + 0.5)}{\#\text{ of values in the data set}}\right]$$

$$= 100\left[\frac{(17+0.5)}{45}\right] = 38.9$$

we round up to 39. The value 1.90 corresponds to the 39th percentile

OBJECTIVE 4
COMPUTE THE FIVE-NUMBER SUMMARY FOR A DATA SET

The five-number summary__________ of a data set consists of the median, the first quartile, the third quartile, the smallest value, and the largest value. These values are generally arranged in order.

EXAMPLE: Recall the Los Angeles annual rainfall data. Compute the five-number summary.

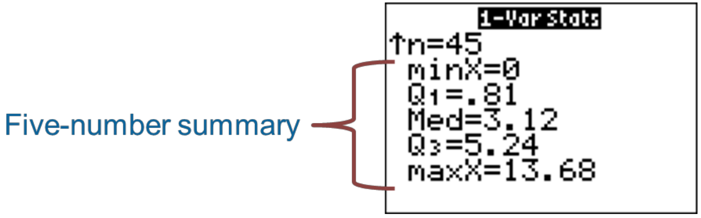
0.00	0.08	0.13	0.14	0.16	0.17	0.20	0.29	0.56	0.67	0.70	0.92	1.22	1.30	1.48
1.64	1.72	1.90	2.37	2.58	2.84	3.06	3.12	3.21	3.29	3.54	3.57	3.71	4.13	4.27
4.37	4.64	4.89	4.94	5.54	6.10	6.61	7.89	7.96	8.03	8.87	8.91	11.02	12.75	13.68

SOLUTION: from the previous example, $Q = 0.92$, med $= 3.12$, $Q_3 = 4.94$

The five-number summary is given by

0.00	0.92	3.12	4.94	13.68
min	Q_1	med	Q_2	max

When using the TI-84 PLUS Calculator, the five-number summary is given by the 1-Var Stats command.



```
1-Var Stats
↑n=45
 minX=0
 Q1=.81
 Med=3.12
 Q3=5.24
 maxX=13.68
```

OBJECTIVE 5
UNDERSTAND THE EFFECTS OF OUTLIERS

An outlier is a value that is considerably larger or considerably smaller than most of the values in a data set. Some outliers result from errors; for example a misplaced decimal point may cause a number to be much larger or smaller than the other values in a data set. Some outliers are correct values, and simply reflect the fact that the population contains some extreme values.

EXAMPLE: The temperature in a downtown location is measured for eight consecutive days during the summer. The readings, in Fahrenheit, are

81.2 85.6 89.3 91.0 83.2 8.45 79.5 87.8

Which reading is an outlier? Is the outlier an error or is it possible that it is correct?

SOLUTION: The outlier is 8.45. It is most likely resulting from a misplaced decimal point.

INTERQUARTILE RANGE

One method for detecting outliers involves a measure called the **Interquartile Range**.

INTERQUARTILE RANGE

The interquartile range is found by subtracting the first quartile from the third quartile.

IQR = $Q_3 - Q_1$

IQR METHOD FOR DETECTING OUTLIERS

The most frequent method used to detect outliers in a data set is the **IQR Method**. The procedure for the IQR Method is:

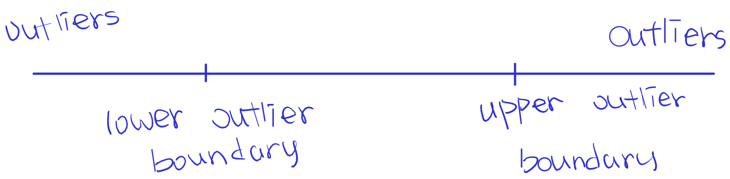
Step 1: Find the first quartile Q_1, and the third quartile Q_3.

Step 2: Compute the interquartile range: IQR = $Q_3 - Q_1$.

Step 3: Compute the outlier boundaries. These boundaries are the cutoff points for determining outliers:

Lower Outlier Boundary = Q_1 – 1.5(IQR)
Upper Outlier Boundary = Q_3 + 1.5(IQR)

Step 4: Any data value that is less than the lower outlier boundary or greater than the upper outlier boundary is considered to be an outlier.



EXAMPLE: The following table presents the number of students absent in a middle school in northwestern Montana for each school day in January. Identify any outliers.

65 67 71 57 51 49 44 41 59 49 42 56
45 77 44 42 45 46 100 59 53 51

SOLUTION: using the TI-84, $Q_1 = 45$ and $Q_3 = 59$

$IQR = Q_3 - Q_1 = 59 - 45 = 14$

The oulier boundaries :

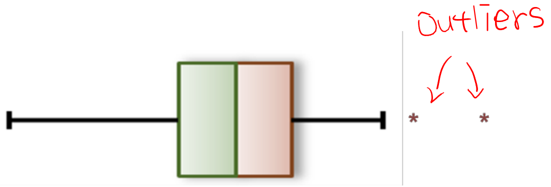
lower outlier boundary $= Q_1 - 1.5(IQR) = 45 - 1.5(14) = 24$

upper outlier boundary $= Q_3 + 1.5(IQR) = 59 + 1.5(14) = 80$

since the value 100 is greater than the upper boundary 80, 100 is an outlier.

OBJECTIVE 6
CONSTRUCT BOXPLOTS TO VISUALIZE THE FIVE-NUMBER SUMMARY AND OUTLIERS

A boxplot is a graph that presents the five-number summary along with some additional information about a data set. There are several different kinds of boxplots. The one we describe here is sometimes called a modified boxplot.



Procedure for Constructing a Boxplot

- **Step 1:** Compute the first quartile, the median, and the third quartile.
- **Step 2:** Draw vertical lines at the first quartile, the median, and the third quartile. Draw horizontal lines between the first and third quartiles to complete the box.
- **Step 3:** Compute the lower and upper outlier boundaries.
- **Step 4:** Find the largest data value that is less than the upper outlier boundary. Draw a horizontal line from the third quartile to this value. This horizontal line is called a **whisker**.
- **Step 5:** Find the smallest data value that is greater than the lower outlier boundary. Draw a horizontal line (whisker) from the first quartile to this value.
- **Step 6:** Determine which values, if any, are outliers. Plot each outlier separately.

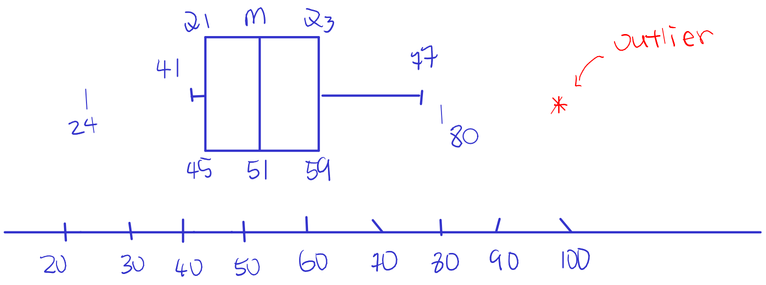
EXAMPLE: The following table presents the number of students absent in a middle school in northwestern Montana for each school day in January. Identify any outliers.

65	67	71	57	51	49	44	41	59	49	42	56
45	77	44	42	45	46	100	59	53	51		

SOLUTION: The five-number summary:

41	45	51	59	100
min	Q1	med	Q3	max

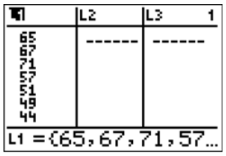
lower outlier boundary = 24 and upper outlier boundary = 80



BOXPLOTS ON THE TI-84 PLUS

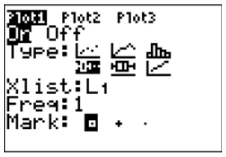
The following steps will create a boxplot for the student absences data on the TI-84 PLUS.

Step 1: Enter the data in **L1**.



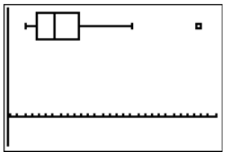
```
L1    L2      L3      1
65    ------  ------
67
71
57
51
49
44
L1 ={65,67,71,57…
```

Step 2: Press **2nd,Y=**, then **1** to access the Plot1 menu. Select **On** and the boxplot type.



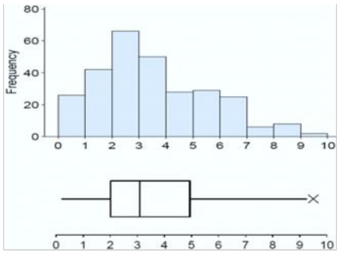
```
Plot1  Plot2  Plot3
On  Off
Type:
Xlist:L1
Freq:1
Mark: ▪  +  ·
```

Step 3: Press **Zoom, 9** to view the plot.



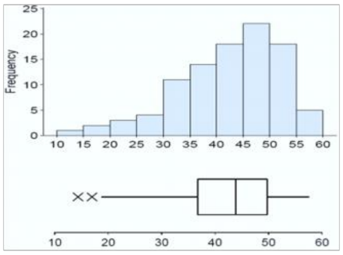
DETERMINING THE SHAPE OF A DATA SET FROM A BOXPLOT

Boxplots can be used to determine skewness in a data set.

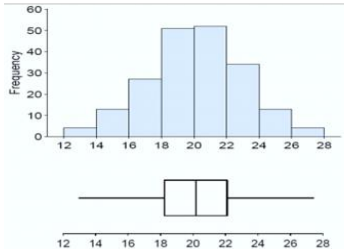


If the median is closer to the first quartile than to the third quartile, or the upper whisker is longer than the lower whisker, the data are skewed to the right.

If the median is closer to the third quartile than to the first quartile, or the lower whisker is longer than the upper whisker, the data are skewed to the left.



If the median is approximately halfway between the first and third quartiles, and the two whiskers are approximately equal in length, the data are approximately symmetric



YOU SHOULD KNOW …

- How to compute and interpret z-scores
- How to compute the quartiles of a data set
- How to compute a percentile of a data set
- How to compute the percentile corresponding to a given data value
- How to find the five-number summary for a data set
- How to determine outliers using the IQR method
- How to construct a boxplot and use it to determine skewness