MAT 142 College Mathematics Module #3

*Statistics*

**Terri Miller**

Spring 2009

*revised March 24, 2009*

---

## 1. Population, Sample, and Data

### 1.1. Basic Terms.

A **population** is the set of all objects under study, a **sample** is any subset of a popultion, and a **data point** is an element of a set of data.

**Example 1.** Population, sample, data point
Population: all ASU students
Sample: 1000 radomly selected ASU students

data: $10, 15, 13, 25, 22, 53, 47$
data point: 13
data point: 53

The **frequency** is the number of times a particular data point occurs in the set of data. A **frequency distribution** is a table that list each data point and its frequency. The **relative frequency** is the frequency of a data point expressed as a percentage of the total number of data points.

**Example 2.** Frequency, relative frequency, frequency distribution
data: $1, 3, 6, 4, 5, 6, 3, 4, 6, 3, 6$
frequency of the data point 1 is 1
frequency of the data point 6 is 4
the relative frequency of the data point 6 is $(4/11) \times 100\% \approx 36.35\%$
the frequency distribution for this set of data is: (where $x$ is a data point and $f$ is the frequency for that point)

| x | f |
|---|---|
| 1 | 1 |
| 3 | 3 |
| 4 | 4 |
| 2 | 2 |
| 5 | 1 |

Data is often described as ungrouped or grouped. **Ungrouped** data is data given as individual data points. **Grouped** data is data given in intervals.

**Example 3.** Ungrouped data without a frequency distribution.

1, 3, 6, 4, 5, 6, 3, 4, 6, 3, 6

**Example 4.** Ungrouped data with a frequency distribution.

| **Number of television sets** | **Frequency** |
|---|---|
| 0 | 2 |
| 1 | 13 |
| 2 | 18 |
| 3 | 0 |
| 4 | 10 |
| 5 | 2 |
| Total | 45 |

**Example 5.** Grouped data.

| **Exam score** | **Frequency** |
|---|---|
| 90-99 | 7 |
| 80-89 | 5 |
| 70-79 | 15 |
| 60-69 | 4 |
| 50-59 | 5 |
| 40-49 | 0 |
| 30-39 | 1 |
| Total | 37 |

## 1.2. **Organizing Data.**

Given a set of data, it is helpful to organize it. This is usually done by creating a frequency distribution.

**Example 6.** Ungrouped data.
Given the following set of data, we would like to create a frequency distribution.

1 5 7 8 2
3 7 2 8 7

To do this we will count up the data by making a tally (a tick mark in the tally column for each occurance of the data point. As before, we will designate the data points by $x$.

| **x** | **tally** |
|---|---|
| 1 | \| |
| 2 | \|\| |
| 3 | \| |
| 4 | |
| 5 | \| |
| 6 | |
| 7 | \|\|\| |
| 8 | \|\| |


© 2009 ASU School of Mathematical & Statistical Sciences and Terri L. Miller

Now we add a column for the frequency, this will simply be the number of tick marks for each data point. We will also total the number of data points. As we have done previously, we will represent the frequency with $f$.

| **x** | **tally** | **f** |
|---|---|---|
| 1 | \| | 1 |
| 2 | \|\| | 2 |
| 3 | \| | 1 |
| 4 | | 0 |
| 5 | \| | 1 |
| 6 | | 0 |
| 7 | \|\|\| | 3 |
| 8 | \|\| | 2 |
| | Total | 10 |

This is a frequency distribution for the data given. We could also include a column for the relative frequency as part of the frequency distribution. We will use rel $f$ to indicate the relative frequency.

| **x** | **tally** | **f** | **rel f** |
|---|---|---|---|
| 1 | \| | 1 | $\frac{1}{10} * 100\% = 10\%$ |
| 2 | \|\| | 2 | $\frac{2}{10} * 100\% = 20\%$ |
| 3 | \| | 1 | 10% |
| 4 | | 0 | 0% |
| 5 | \| | 1 | 10% |
| 6 | | 0 | 0% |
| 7 | \|\|\| | 3 | $\frac{3}{10} * 100\% = 30\%$ |
| 8 | \|\| | 2 | 20% |
| | Totals | 10 | 100% |

For our next example, we will use the data to create groups (or categories) for the data and then make a frequency distribution.

**Example 7.** Grouped Data.
Given the following set of data, we want to organize the data into groups. We have decided that we want to have 5 intervals.

| | | | | |
|---|---|---|---|---|
| 26 | 18 | 21 | 34 | 18 |
| 38 | 22 | 27 | 22 | 30 |
| 25 | 25 | 38 | 29 | 20 |
| 24 | 28 | 32 | 33 | 18 |

Since we want to group the data, we will need to find out the size of each interval. To do this we must first identify the highest and the lowest data point. In our data the highest data point is 38 and the lowest is 18. Since we want 5 intervals, we make the compution

$$\frac{\text{highest} - \text{lowest}}{\text{number of intervals}} = \frac{38 - 18}{5} = \frac{20}{5} = 4$$


© 2009 ASU School of Mathematical & Statistical Sciences and Terri L. Miller

Since we need to include all points, we always take the next highest integer from that which was computed to get the length of our interval. Since we computed 4, the length of our intervals will be 5. Now we set up the first interval

$$\text{lowest} \leq x < \text{lowest} + 5 \qquad \text{which results in} \quad 18 \leq x < 23.$$

Our next interval is obtained by adding 5 to each end of the first one:

$$18 + 5 \leq x < 23 + 5 \qquad \text{which results in} \quad 23 \leq x < 28.$$

We continue in this manner to get all of our intervals:

$$\begin{gathered} 18 \leq x < 23 \\ 23 \leq x < 28 \\ 28 \leq x < 33 \\ 33 \leq x < 38 \\ 38 \leq x < 43. \end{gathered}$$

Now we are ready to tally the data and make the frequency distribution. Be careful to make sure that a data point that is the same number as the end of the interval is placed in the correct interval. This means that the data point 33 is counted in the interval $33 \leq x < 38$ and NOT in the interval $28 \leq x < 33$.

| **x** | **tally** | **f** | **rel f** |
|---|---|---|---|
| $18 \leq x < 23$ | \|\|\|\|\|\|\| | 7 | $\frac{7}{20} * 100\% = 35\%$ |
| $23 \leq x < 28$ | \|\|\|\|\| | 5 | $\frac{5}{20} * 100\% = 25\%$ |
| $28 \leq x < 33$ | \|\|\|\| | 4 | $\frac{4}{20} * 100\% = 20\%$ |
| $33 \leq x < 38$ | \|\| | 2 | $\frac{2}{20} * 100\% = 10\%$ |
| $38 \leq x < 43$ | \|\| | 2 | 10% |
| | Totals | 20 | 100% |

## 1.3. **Histogram.**

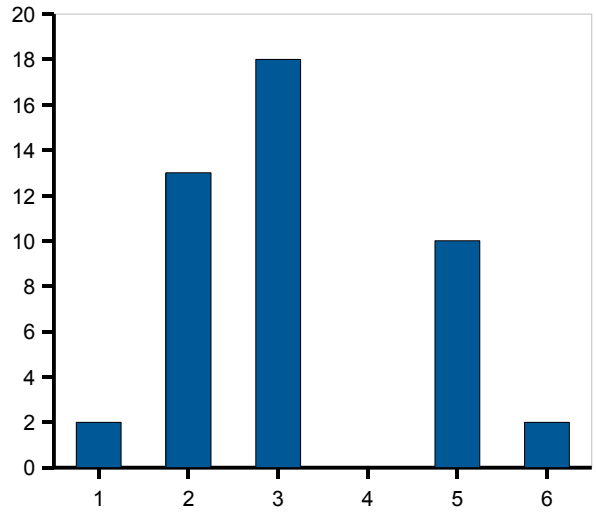
Now that we have the data organized, we want a way to display the data. One such display is a **histogram** which is a bar chart that shows how the data are distrubuted among each data point (ungrouped) or in each interval (grouped)

**Example 8.** Histogram for ungrouped data.
Given the following frequency distribution:

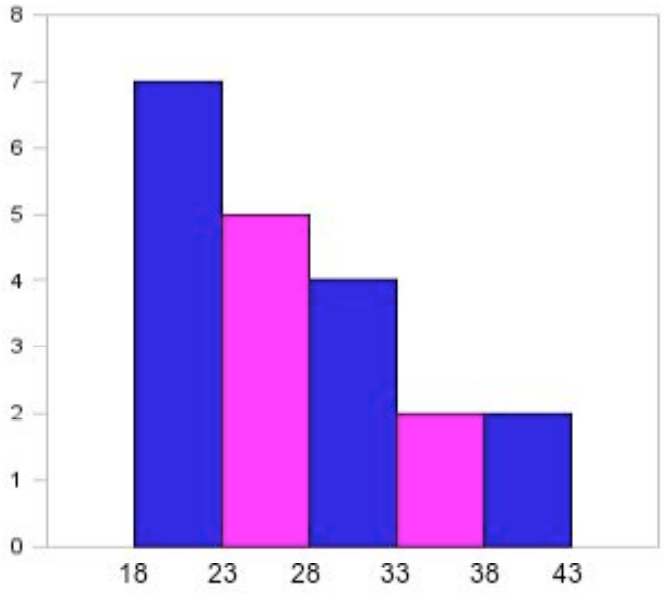
| **Number of television sets** | **Frequency** |
|---|---|
| 0 | 2 |
| 1 | 13 |
| 2 | 18 |
| 3 | 0 |
| 4 | 10 |
| 5 | 2 |

The histogram would look as follows:


© 2009 ASU School of Mathematical & Statistical Sciences and Terri L. Miller




**Example 9.** Histogram example for grouped data. We will use the data from example 7. The histogram would look as follows.




## 2. Measures of Central Tendency

### 2.1. **Mode.**

The **mode** is the data point which occurs most frequently. It is possible to have more than one mode, if there are two modes the data is said to be **bimodal**. It is also possible for a set of data to not have any mode, this situation occurs if the number of modes gets to be "too large". It it not really possible to define "too large" but one should exercise good judgement. A resaonable, though very generous, rule of thumb is that if the number of data points accounted for in the list of modes is half or more of the data points, then there is no mode.


© 2009 ASU School of Mathematical & Statistical Sciences and Terri L. Miller

Note: if the data is given as a list of data points, it is often easiest to find the mode by creaating a frequency distribution. This is certainly the most organized method for finding it. In our examples we will use frequency distributions.

**Example 10.** A data set with a single mode.
Consider the data from example 8:

| **Number of television sets** | **Frequency** |
|---|---|
| 0 | 2 |
| 1 | 13 |
| 2 | 18 |
| 3 | 0 |
| 4 | 10 |
| 5 | 2 |

You can see from the table that the data point which occurs most frequently is 2 as it has a frequency of 18. So the mode is 2.

**Example 11.** A data set with two modes.
Consider the data:

| **Number of hours of television** | **Frequency** |
|---|---|
| 0 | 1 |
| 0.5 | 4 |
| 1 | 8 |
| 1.5 | 9 |
| 2 | 13 |
| 2.5 | 10 |
| 3 | 11 |
| 3.5 | 13 |
| 4 | 5 |
| 4.5 | 3 |

You can see from the table that the data points 2 and 3.5 both occur with the highest frequency of 13. So the modes are 2 and 3.5.

**Example 12.** A data set with no mode.
Consider the data:


© 2009 ASU School of Mathematical & Statistical Sciences and Terri L. Miller

| Age | Frequency |
|---|---|
| 18 | 12 |
| 19 | 5 |
| 20 | 3 |
| 21 | 9 |
| 22 | 1 |
| 23 | 8 |
| 24 | 12 |
| 25 | 12 |
| 26 | 5 |
| 27 | 3 |
| Total | 71 |

You can see from the table that the data points 18, 24 and 25 all occur with the highest frequency of 12. Since this would account for 36 of the 71 data points, this would qualify as "too large" a number of data points taken accounted for. In this case, we would say there is no mode.

### 2.2. **Median.**

The **median** is the data point in the middle when all of the data points are arranged in order (high to low or low to high). To find where it is, we take into account the number of data points. If the number of data points is odd, divide the number of data points by 2 and then round up to the next integer; the resulting integer is the location of the median. If the number of data points is even, there are two middle values. We take the number of data points and divide by 2, this integer is the first of the two middles, the next one is also a middle. Now we average these two middle values to get the median.

**Example 13.** An odd number of data points with no frequency distribution.

$$3, 4.5, 7, 8.5, 9, 10, 15$$

There are 7 data points and 7/2=3.5 so the median is the 4th number, 8.5.

**Example 14.** An odd number of data points with a frequency distribution.

| Age | Frequency |
|---|---|
| 18 | 12 |
| 19 | 5 |
| 20 | 3 |
| 21 | 9 |
| 22 | 2 |
| Total | 31 |

There are 31 data points and 31/2=15.5 so the median is the 16th number. Start counting, 18 occurs 12 times, then 19 occurs 5 times getting us up to entry 17 (12+5); so the 16th entry must be a 19. This data set has a median of 19.

**Example 15.** An even number of data points with no frequency distribution.

$$3, 4.5, 7, 8.5, 9, 10, 15, 15.5$$


© 2009 ASU School of Mathematical & Statistical Sciences and Terri L. Miller

There are 8 data points and 8/2=4 so the median is the average of the 4th and 5th data point, (8.5+9)/2=8.75. This data set has a median of 8.75.

**Example 16.** An even number of data points with a frequency distribution.

| Age | Frequency |
|---|---|
| 18 | 11 |
| 19 | 5 |
| 20 | 3 |
| 21 | 9 |
| 22 | 2 |
| Total | 30 |

There are 30 data points and 30/2=15 so the median is the average of the 15th and 16th number. Start counting, 18 occurs 11 times, then 19 occurs 5 times getting us up to entry 16 (12+5); so both the 15th and 16th entries must be a 19. This data set has a median of 19.

**Example 17.** A second case of an even number of data points with a frequency distribution.

| Age | Frequency |
|---|---|
| 18 | 10 |
| 19 | 5 |
| 20 | 4 |
| 21 | 9 |
| 22 | 2 |
| Total | 30 |

There are 30 data points and 30/2=15 so the median is the average of the 15th and 16th number. Start counting, 18 occurs 10 times, then 19 occurs 5 times getting us up to entry 15 (10+5); so both the 15th entries is a 19 and the 16th entry must be 20, so the median is the average of these two datapoints, (19+20)/2=19.5. This data set has a median of 19.5.

### 2.3. **Mean.**

The **mean** is the average of the data points, it is denoted $\overline{x}$. There are three types of data for which we would like to compute the mean, ungrouped of frequency 1, ungrouped with a frequency distribution, and grouped.

Starting with the first type, ungrouped of frequency 1, is when data is given to you as a list and it is not organized into a frequency distribution. When this happens, we compute the average as we have always done, add up all of the data points and divide by the number of data points. To write a formula for this, we use the capital greek letter sigma, $\Sigma x$. This just means to add up all of the data points. We will use $n$ to represent the number of data points.

$$\text{mean:} \quad \overline{x} = \frac{\Sigma x}{n}$$

This corresponds to the left hand column of your calculator instructions.


© 2009 ASU School of Mathematical & Statistical Sciences and Terri L. Miller

**Example 18.** Given the ungrouped data list below:

$$10 \quad 15 \quad 13 \quad 25 \quad 22 \quad 53 \quad 47$$

We would enter the data into the calculator following the instructions in the left hand column, the result is $\overline{x} = 26.4285714$.

When have a frequency distribution for the data, we have to take the average as before but remember that the frequency gives the number of times that the data point occurs.

$$\text{mean:} \quad \overline{x} = \frac{\Sigma(fx)}{n}$$

This corresponds to the right hand column of your calculator instructions.

**Example 19.** Given the frequency distribution for ungrouped data below:

| **Number of television sets** | **Frequency** |
|---|---|
| 0 | 2 |
| 1 | 13 |
| 2 | 18 |
| 3 | 0 |
| 4 | 10 |
| 5 | 2 |

We would enter the data into our calculators following the directions in the right hand column. The result is $\overline{x} = 2.2$.

Our final type of data is grouped data. This requires a computation before we can begin. Since we cannot enter the entire interval as a data point, we use a representative for each interval, $x_i$. This representative is the midpoint of the interval, to find the midpoint of an interval you add the two endpoints and divide by 2. These are the numbers that you use as data points for computing the mean.

**Example 20.** Mean for Grouped data.
We will use the data from example 7 again:

| **x** | **f** |
|---|---|
| $18 \leq x < 23$ | 7 |
| $23 \leq x < 28$ | 5 |
| $28 \leq x < 33$ | 4 |
| $33 \leq x < 38$ | 2 |
| $38 \leq x < 43$ | 2 |
| Total | 20 |

Start by calculating the representative for each interval.

$$\frac{23 + 18}{2} = \frac{41}{2} = 20.5$$

Since this is the midpoint of the first interval and the intervals have length 5, we find the rest by adding 5 to this one.


© 2009 ASU School of Mathematical & Statistical Sciences and Terri L. Miller

| x | $x_i$ | f |
|---|---|---|
| $18 \leq x < 23$ | 20.5 | 7 |
| $23 \leq x < 28$ | 25.5 | 5 |
| $28 \leq x < 33$ | 30.5 | 4 |
| $33 \leq x < 38$ | 35.5 | 2 |
| $38 \leq x < 43$ | 40.5 | 2 |
| Totals | | 20 |

Now enter the data as directed using the right hand column of the calculator directions. You use $x_i$ as the data point (list 1) and the frequency as usual in list 2. The result should be $\overline{x} = 27.25$.

## 3. Standard Deviation

### 3.1. Definition and Computation of Standard Deviation.

The **standard deviation** is a measurement of how much the data varies from the mean. It is a measure of dispersion, the more dispersed the data, the less consistent the data is. A lower standard deviation means that the data is more clustered around the mean and hence the data set is more consistent. We will denote the standard deviation with the symbol *sd*.

The formulae for computing the standard deviation are:

$$\text{data with frequency 1:} \quad sd = \sqrt{\frac{\Sigma(x-\overline{x})^2}{n-1}}$$

$$\text{data with frequency distribution:} \quad sd = \sqrt{\frac{\Sigma f(x-\overline{x})^2}{n-1}}.$$

You need to read your calculator instructions to see what notation your calculator uses for the standard deviation.

**Example 21.** Standard deviation for a data set with frequency 1.
Using the data from example 18:

10 15 13 25 22 53 47

We found the mean to be $\overline{x} = 26.4285714$. You should also see from the same calculation that the standard deviation is $sd = 16.98879182$.

**Example 22.** Standard deviation for a data set with a frequency distribution.
Using the data from example 19:

| Number of television sets | Frequency |
|---|---|
| 0 | 2 |
| 1 | 13 |
| 2 | 18 |
| 3 | 0 |
| 4 | 10 |
| 5 | 2 |


© 2009 ASU School of Mathematical & Statistical Sciences and Terri L. Miller

We found the mean to be $\overline{x} = 2.2$. You should also see from the same calculation that the standard deviation is $sd = 1.324592562$.

### 3.2. **Interpretation of Standard Deviation.**

Our first use of the standard deviation is to compare two data sets. This uses the standard deviation to interpret how consistent the data is. As mentioned in the previous section, the lower the standard deviation, the more consistent the data is.

**Example 23.** Two bowlers have the scores given below:

| | | | | | |
|---|---|---|---|---|---|
| **Katie's Scores** | 189 | 146 | 200 | 241 | 231 |
| **Mike's Scores** | 235 | 201 | 217 | 168 | 186 |

Both sets of data have a mean of $\overline{x} = 201.4$. Does this mean they are equivalent bowlers? No, consider the standard deviations. Katie has a standard deviation of $sd = 37.6470$ and Mike has a standard deviation of $sd = 26.1017$. Since Mike has a smaller standard deviation, he is a more consisetent bowler than Katie, *i.e.* Mike is more likely to get a score of 201.4.

The other use we will make of the standard deviation is to determine how much of our data is near the mean. To do this we set up some intervals using the mean and the standard deviation.

The interval (or range of data) that corresponds to one standard deviation from the mean is found by subtracting and adding the standard deviation from/to the mean. The interval that corresponds to two standard deviations from the mean is found by subtracting and adding twice the standard deviation from/to the mean. The interval that corresponds to three standard deviations from the mean is found by subtracting and adding three times the standard deviation from/to the mean.

**Example 24.** Intervals around the mean for data with frequency 1.
Using the data from example 3:

$$1, \quad 3, \quad 6, \quad 4, \quad 5, \quad 6, \quad 3, \quad 4, \quad 6, \quad 3, \quad 6$$

This data has a mean of 4.2727 and a standard deviation of 1.6787.

The interval that corresponds to one standard deviation from the mean is:

$$\begin{array}{rcl} \overline{x} - sd & \text{to} & \overline{x} + sd \\ \text{4.2727-1.6787} & \text{to} & \text{4.2727+1.6787} \\ \text{2.594} & \text{to} & \text{5.9514} \end{array}$$

This means that any data point that is larger than 2.594 and smaller than 5.9514 is within one standard deviation of the mean. From our data set, those data points that are within one standard deviation of the mean are 3, 4, 5, 3, 4, 3.

The interval that corresponds to two standard deviations from the mean is:

$$\begin{array}{rcl} \overline{x} - 2sd & \text{to} & \overline{x} + 2sd \\ 4.2727 - 2 \times 1.6787 & \text{to} & 4.2727 + 2 \times 1.6787 \\ \text{4.2727-3.3574} & \text{to} & \text{4.2727+3.3574} \\ \text{0.9153} & \text{to} & \text{7.6301} \end{array}$$


© 2009 ASU School of Mathematical & Statistical Sciences and Terri L. Miller

This means that any data point that is larger than 0.9153 and smaller than 7.6301 is within two standard deviation sof the mean. From our data set this happens to be all of the data points.

The interval that corresponds to three standard deviations from the mean is:

$$\begin{array}{rcl} \overline{x} - 3sd & \text{to} & \overline{x} + 3sd \\ 4.2727 - 3 \times 1.6787 & \text{to} & 4.2727 + 3 \times 1.6787 \\ 4.2727\text{-}5.0361 & \text{to} & 4.2727\text{+}5.0361 \\ -0.7634 & \text{to} & 9.3088 \end{array}$$

This means that any data point that is larger than $-0.7634$ and smaller than 9.3088 is within three standard deviations of the mean. Again, for our data, all data points are within three standard deviations of the mean.

Now that we can find the intervals, we can use them to discuss how dispersed the data is.

**Example 25.** A set of data with frequency 1.
*Consumer Reports* magazine gave the following data for the number of calories in a meat hot dog of 17 major brands.

$$173, 191, 182, 190, 172, 147, 146, 138, 175, 136, 179, 153, 107, 195, 135, 140, 138$$

The mean is $\overline{x} = 158.6471$ and the standard deviation is $sd = 25.2857$.

What percent of the data lies within one standard deviation of the mean?

Since the interval corresponding to one standard deviation from the mean is 133.3613 to 183.9328, we find that 13 of the 17 data points are in that interval. So, $(13/17) \times 100\% \approx 76.5\%$ of the data is within one standard deviation of the mean.

What percent of the data lies within two standard deviations of the mean?

Since the interval corresponding to two standard deviations from the mean is 108.0756 to 209.2185, we find that 16 of the 17 data points are in that interval. So, $(16/17) \times 100\% \approx 94.1\%$ of the data is within two standard deviations of the mean.

What percent of the data lies within three standard deviations of the mean?

Since the interval corresponding to three standard deviations from the mean is 82.7899 to 234.5, we see that all of the data points are in that interval. So, 100% of the data is within three standard deviations of the mean.

**Example 26.** A set of data with a frequency distribution.
The following data gives a frequency distribution for the number of social interactions of longer than 10 minutes in one week for a group of college students.

| **x** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 | 47 |
|---|---|---|---|---|---|---|---|---|---|---|
| **f** | 12 | 16 | 16 | 16 | 10 | 11 | 4 | 3 | 3 | 3 |

The mean is $\overline{x} = 17.2660$ and the standard deviation is $sd = 11.6943$.

What percent of the data lies within one standard deviation of the mean?

Since the interval corresponding to one standard deviation from the mean is 5.5717 to 28.96023, we find that 69 of the 94 data points are in that interval (all of the 7's, 12's,


© 2009 ASU School of Mathematical & Statistical Sciences and Terri L. Miller

17's, 22's, and 27's). So, $(69/94) \times 100\% \approx 73.4\%$ of the data is within one standard deviation of the mean.

What percent of the data lies within two standard deviations of the mean?

Since the interval corresponding to two standard deviations from the mean is $-6.1226$ to 40.6545, we find that 88 of the 94 data points are in that interval (in addition to the previous 69, all of the 2's, 32's, and 37's). So, $(88/94) \times 100\% \approx 93.6\%$ of the data is within two standard deviations of the mean.

What percent of the data lies within three standard deviations of the mean?

Since the interval corresponding to three standard deviations from the mean is $-17.8169$ to 52.3488, we see that all of the data points are in that interval. So, 100% of the data is within three standard deviations of the mean.


© 2009 ASU School of Mathematical & Statistical Sciences and Terri L. Miller