

NORMAL DISTRIBUTIONS

MEASURES OF VARIATION

In statistics, it is important to measure the **spread** of data. A simple way to measure spread is to find the range. But statisticians want to know if the data are relatively close to, or far from the mean, so measures of variation are used to determine how far observations are from the mean. The common measures of variation are **mean deviation**, **variance**, and **standard deviation**.

Mean Deviation

The average distance each piece of data is from the mean.

To calculate the **mean deviation**:

1. Calculate the mean of the data.
2. Subtract the mean from each piece of data, and take the absolute value of the result.
3. Add all of the results.
4. Divide by the number of values in the data set.

Example 1

Consider these two sets of data:

Set A: 10, 30, 39, 40, 43, 48, 70

Set B: 10, 20, 30, 40, 50, 60, 70

The mean for both sets is 40 and the range for both is $70 - 10 = 60$. To decide which set has the greater spread, calculate the mean deviation.

1. Calculate the mean: $280/7=40$
2. Find the difference of each data point from the mean, then take the absolute value.

Set A:

$$|40 - 10| = 30$$

$$|40 - 30| = 10$$

$$|40 - 39| = 1$$

$$|40 - 40| = 0$$

$$|40 - 43| = 3$$

$$|40 - 48| = 8$$

$$|40 - 70| = 30$$

Set B:

$$|40 - 10| = 30$$

$$|40 - 20| = 20$$

$$|40 - 30| = 10$$

$$|40 - 40| = 0$$

$$|40 - 50| = 10$$

$$|40 - 60| = 20$$

$$|40 - 70| = 30$$

3. Add all of the results from Step 2.

$$\text{Set A: } 30 + 10 + 1 + 0 + 3 + 8 + 30 = 82$$

$$\text{Set B: } 30 + 20 + 10 + 0 + 10 + 20 + 30 = 120$$

4. Divide by the number of values: Set A: $82 \div 7 \approx 11.714$

$$\text{Set B: } 120 \div 7 \approx 17.142$$

Based on the mean deviation, Set B has the greater spread of data because the average distance of each piece of data from the mean is greater.

Another way to measure the spread of a set of data is called the **variance**. To find the **variance** of a set of data:

1. Calculate the mean of the data.
2. Subtract the mean from each piece of data then square the difference (this ensures that all values are positive, just as using the absolute value did for mean deviation).
3. Add all the squared differences.
4. Divide by the one less than the number of values in the data set.

The symbol for variance of a sample is s^2 .

Example 2

Find the variance for Sets A and B in Example 2.

1. Calculate the mean of each set: 40
2. Subtract the mean, then square the difference in each set:

Set A:

$$(40 - 10)^2 = 900$$

$$(40 - 30)^2 = 100$$

$$(40 - 39)^2 = 1$$

$$(40 - 40)^2 = 0$$

$$(40 - 43)^2 = 9$$

$$(40 - 48)^2 = 64$$

$$(40 - 70)^2 = 900$$

Set B:

$$(40 - 10)^2 = 900$$

$$(40 - 20)^2 = 400$$

$$(40 - 30)^2 = 100$$

$$(40 - 40)^2 = 0$$

$$(40 - 50)^2 = 100$$

$$(40 - 60)^2 = 400$$

$$(40 - 70)^2 = 900$$

3. Add all of the squared differences in each set:

$$\text{Set A: } 900 + 100 + 1 + 0 + 9 + 64 + 900 = 1974$$

$$\text{Set B: } 900 + 400 + 100 + 0 + 100 + 400 + 900 = 2800$$

4. Divide by one less than the number of values in each set:

$$\text{Set A: } 1974 \div (7 - 1) = 329$$

$$\text{Set B: } 2800 \div (7 - 1) = 466.67$$

The variance (s^2) is 329 for Set A and $s^2 = 466.67$ for Set B. Based on the variance, set B has a greater spread.

Finally, a more common measure of spread is the **standard deviation**. The **standard deviation** is the square root of the variance. The symbol for the standard deviation is s .

$$\text{standard deviation} = \sqrt{s^2} = s$$

If you know the variance (s^2) of a data set and want to find the standard deviation (s), take the square root of the variance. If the variance is not known, then it must first be calculated to find the standard deviation.

Example 3

Find the standard deviations for Set A and Set B in Example 3.

To find the standard deviation, take the square root of the variance.

$$\text{Set A: } s = \sqrt{329} \approx 18.14$$

$$\text{Set B: } s = \sqrt{466.67} \approx 21.60$$

While it is probably easier to understand and calculate the mean deviation, the standard deviation is most used because of its relationship to the normal distribution, which will be developed in the next example.

Example 4

Find the standard deviation of:
11, 19, 30, 22, 26

1. Mean: $108/5 = 21.6$
2. Sum of squared differences from the mean: $(11 - 21.6)^2 + (19 - 21.6)^2 + (30 - 21.6)^2 + (22 - 21.6)^2 + (26 - 21.6)^2 = 209.2$
3. Divide by one less than the total number of numbers: $209.2/4 = 52.3$
4. Square root: $\sqrt{52.3} = 7.232$

$$s = 7.232$$

Problems

1. Find the mean deviation of the data: 5, 2, 7, 8, 1, 5, 4, 6, 10, 9.
2. Find the variance of the data: 5, 2, 7, 8, 1, 5, 4, 6, 10, 9.
3. Find the variance of the data: 24, 39, 41, 32, 56, 23.
4. If the variance of a data set is 3125.87, find the standard deviation.
5. Find the standard deviation of the data set: 41, 58, 32, 56, 67, 49, 45, and 46.

Answers

1. Mean deviation is 2.3.

2. $s^2 = 8.456$

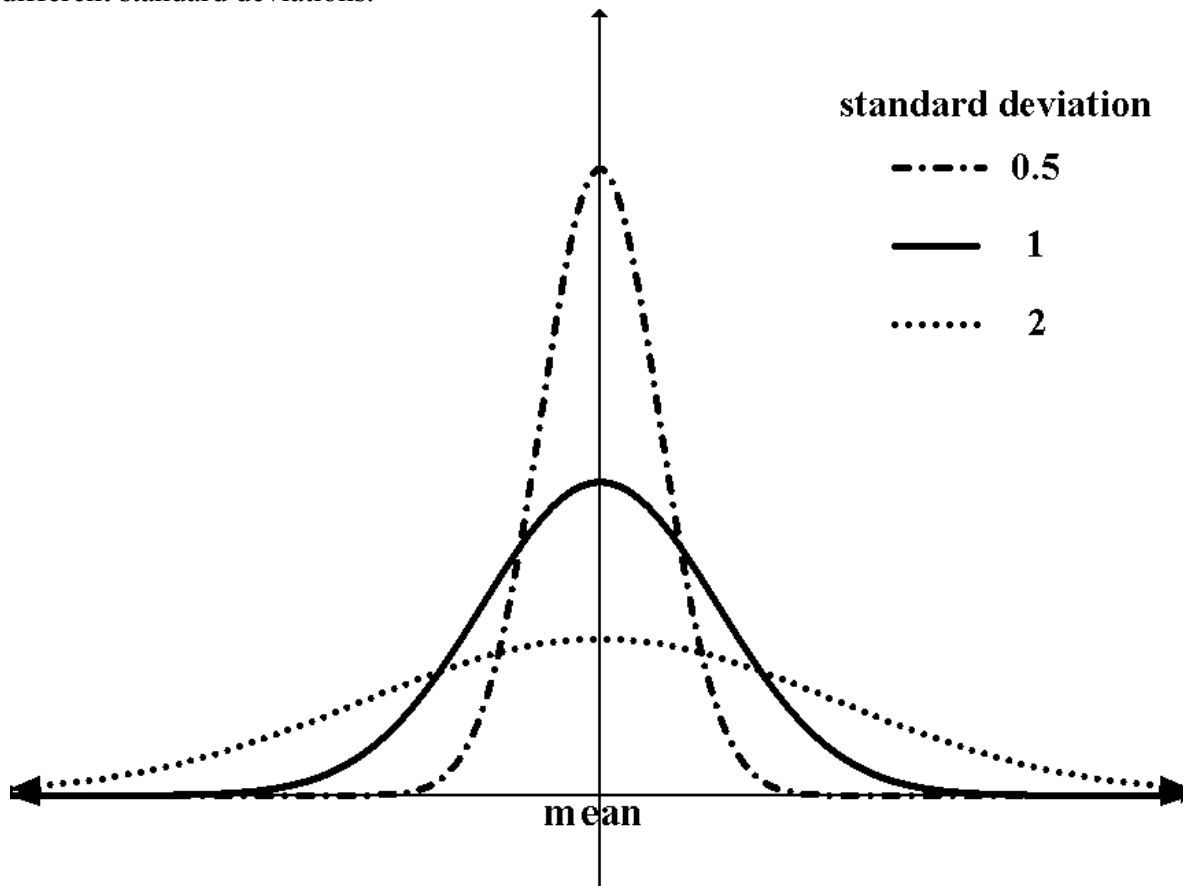
3. $s^2 = 152.567$

4. $s = 55.909$

5. $s = 10.899$

PREDICTIONS BASED ON NORMAL DISTRIBUTIONS

A **normal distribution** is an important theoretical idea used in statistics to model data applicable to large populations. The graph of a normal distribution is a symmetrical bell-shaped curve based on the mean and standard deviation of a sample, where the mean, median, and mode are the same. Below are the graphs of normal distributions with the same mean, but different standard deviations.



The mean is located at the center of the graph. Since this is also the median, half of the values lie above the mean and the other half lie below the mean. The shape of the curve is determined by the standard deviation, or spread of the data.

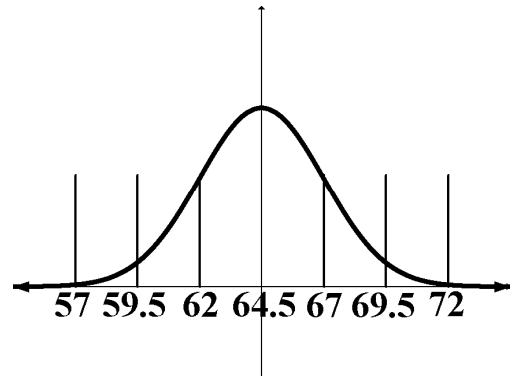
The **normal distribution** can describe many things in the world such as scores on tests like the SAT. Characteristics in nature also fall along the normal curve, such as height or weight for adult men and women.

When the mean and standard deviation of a normal distribution are known, then several other values of the distribution can be calculated. It is important to be able to draw a curve and label the mean, as well as three standard deviations above and below the mean. Theoretically, almost all of the data should fall within 3 standard deviations (on each side) of the mean.

Example 1

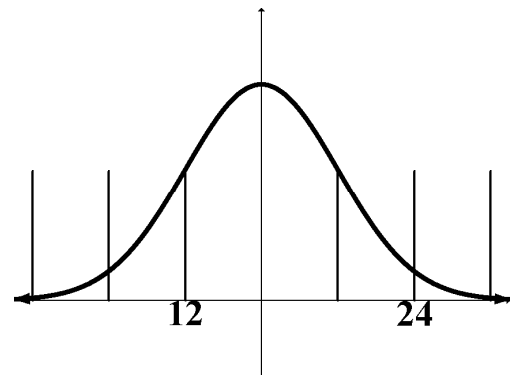
Suppose the average height of women is normally distributed and has a mean of 64.5 inches and a standard deviation of 2.5 inches. Draw a normal curve out to three standard deviations on either side of the mean.

Answer: Draw a bell-shaped curve with the mean of 64.5 in the center. Draw 3 equally spaced vertical lines both above and below the mean. Each vertical line represents one standard deviation. In this case the standard deviation is 2.5, so to label each vertical line, add or subtract 2.5 from the previous line to find the value.



Example 2

Based on the information given, label the remaining values, state the mean and standard deviation.



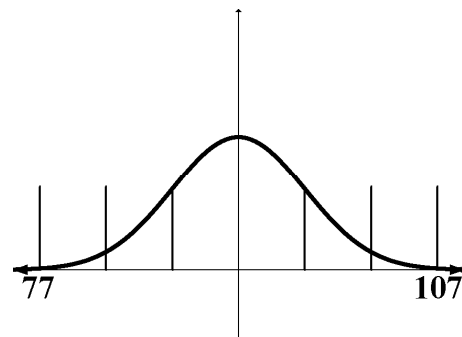
Numerically, the distance from 12 to 24 is 12. Since 24 is on the normal curve above is three standard deviations above 12, each standard deviation must be 4.

Redraw the curve and label from left to right: 4, 8, 12, 16, 20, 24, 28

The mean of the curve is 16 with a standard deviation of 4.

Problems

1. Draw and label a curve (out to three standard deviations) that has a mean of 50 and a standard deviation of 15.
2. Finish labeling the curve at right. State the mean and standard deviation.



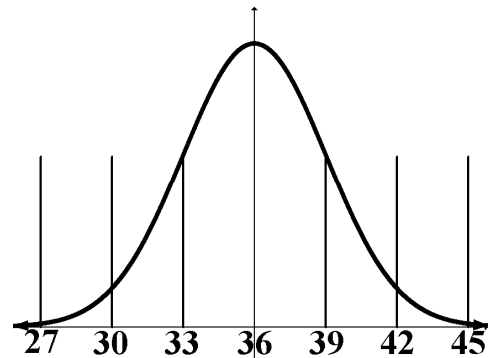
Standard deviations are used to measure distances on the normal distribution curve. There is a pattern in normal distributions known as the

68-95-99.7 Rule. The rule states that in any normal distribution:

- 68% of all observations fall within one standard deviation of the mean;
- 95% of all observations fall within two standard deviations of the mean; and
- 99.7% of all observations fall within three standard deviations of the mean.

Example 3

Suppose the ages of adults taking a business course for the last 5 years at a local college are normally distributed as shown in the graph at right. Based on the 68-95-99.7 rule we know:



68% of all adults taking the business class are between the ages of ____ and ____.
95% of all adults taking the business class are between the ages of ____ and ____.
99.7% of all adults taking the business class are between the ages of ____ and ____.

Answers:

68% of all adults taking the business class are between the ages of 33 and 39.
95% of all adults taking the business class are between the ages of 30 and 42
99.7% of all adults taking the business class are between the ages of 27 and 45.

Example 4

Based on the information in Example 3, what percentage of people taking the course are:

- a. Older than 36?
- b. Younger than 30?
- c. Between the ages of 36 and 42?

Answers:

- a. In a normal distribution, half the values lie above the mean, and the other half below the mean. So 50% are older than the mean age of 36.
- b. We know (from the 68-95-99.7 rule) that about 95% of the data lie between 30 and 42, so that must mean that the remaining 5% lie outside that interval. Since normal curves are symmetric, half of the 5% must lie below 30 and the other half above 42. So 2.5% of the people are younger than 30.
- c. As stated in parts (a) and (b). 50% of the data lie below the mean of 36, and 2.5% of the data lie above 42. The remaining 47.5% of the data must lie between 36 and 42.

Problems

3. Sketch a normal curve that has a mean of 30 and a standard deviation of 5. Then complete the statements:

68% of the data are between ___ and ___.

95% of the data are between ___ and ___.

99.7% of the data are between ___ and ___.

4. Using the graph you made in problem #3, answer the following questions:
 - a. What percent of the data lie above 15?
 - b. What percent of the data lie between 20 and 25?
 - c. What percent of the data lie below 35?

Answers

1. The curve is centered at 50, and the labels from left to right are: 5, 20, 35, 50, 65, 80, 95.
2. The curve is labeled from left to right: 77, 82, 87, 92, 97, 102, 107. The mean is 92 and the standard deviation is 5.
3. 68% of the data are between 25 and 35. 95% of the data are between 20 and 40. 99.7% of the data are between 15 and 45.
4. a. 99.85%
b. 13.5%
c. 84%