

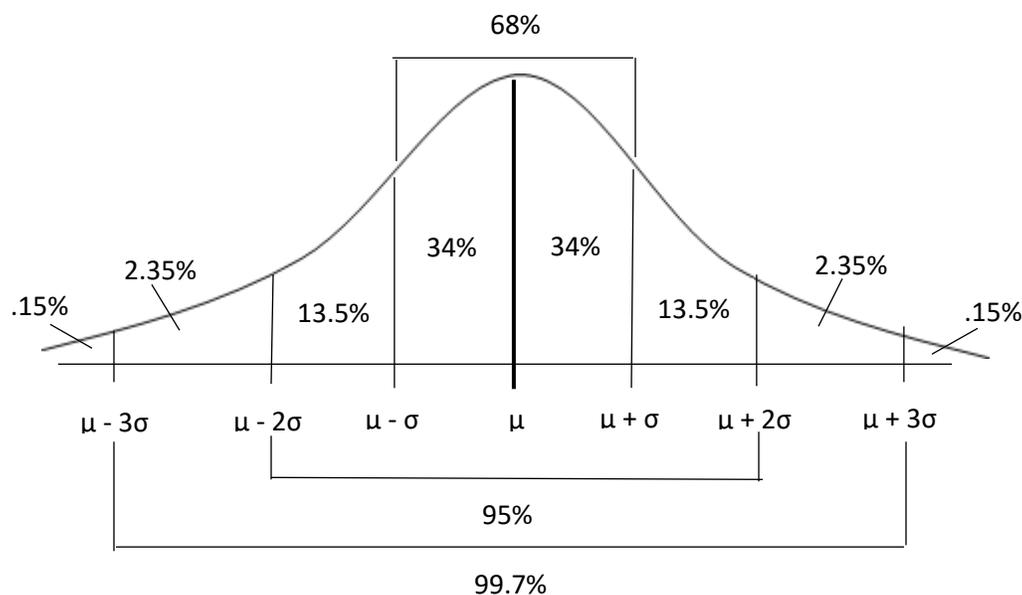
Applications of Data Dispersions

Key Definitions

- **Standard Deviation:** The standard deviation shows how far away each value is from the mean on average.
- **Z-Scores:** The distance between the mean and a given value, represented in number of standard deviations.
- **Percentile:** A percentage that represents a set of data less than or equal to the value of data given.
- **Quartiles:** Divide the data into four equal parts.
- **Interquartile Range:** The range of the middle 50% of the data set.
- **Fences:** The cutoff point when finding the outliers.
- **Outliers:** The outlier is a data value that is not close to or similar to the other data values. These are known as extreme values.

Empirical Rule

- **Empirical Rule Usage:** If our data has a bell-shaped distribution, we can utilize the empirical rule to find the percentage that lies within a certain number of standard deviations. We will use a bell-shaped curve (shown below) to look at this more closely.



- **Explaining the Empirical Rule's Bell-Shaped Curve:** The highest point of the graph is where the mean lies. The lines to the right and left of the mean are data value points that lie a certain number of standard deviations away from the mean. The percentages located within the different standard deviations is the percentage of data that lies within the region.

- **How to Interpret Percentage Using the Empirical Rule Graph:** There are two ways they will ask you to find the percentage using the empirical rule graph. The following will explain both:
 - **Version 1:** The first way is they will tell you that the data lies within “ k ” standard deviations. The “ k ” represents the number of standard deviations away the data is and not the actual value of the standard deviation. We mark the lines that have the “ k ” standard deviations before the “ σ ”. We then add up the percentages in-between.
 - **Version 2:** In this version, you will apply the steps from version 1 after you have applied some new steps. If you are given the value of data, the mean, and the standard deviation and asked to find the percentage, you must find the “ k ” standard deviations first before adding up the percentages. To do this, we use an equation to solve for the missing “ k ”. The following is the equation (x is the data value given):

$$\mu + k_1\sigma = x \text{ and } \mu - k_2\sigma = x$$

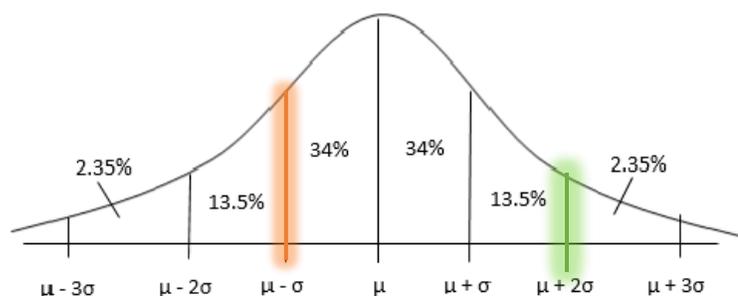
You plug in the information already known and solve for “ k ”. When you have found “ k ”, you mark the lines that have the “ k ” standard deviations before the “ σ ”, and add up the percentages in-between.

- **Example of Interpreting Percentage from Empirical Rule Graph:**
 - **Version 2:** You are looking for the percentage of kids who have taken swimming lessons at a facility that are between the ages 7 and 10. The mean of the data is 8 and the standard deviation is 1. We must plug into the equation to find the “ k ”.

$$8 + k_1 = 10 \text{ and } 8 - k_2 = 7$$

$$k_1 = 2 \text{ and } k_2 = 1$$

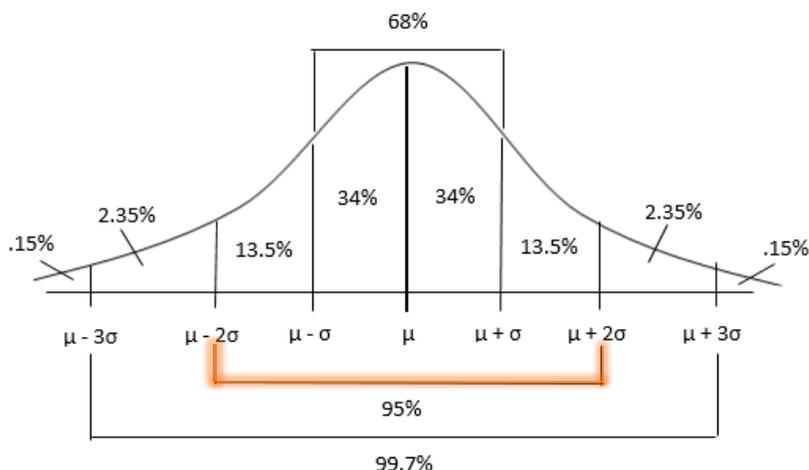
Since the k_1 is 2 and the k_2 is 1, we mark the lines on the graph with the proper “ k ” in front of the standard deviation.



We add together those numbers to find the answer. Then, we get the answer:

$$\text{Percentage} = 34\% + 34\% + 13.5\% = 81.5\%$$

- **How to Interpret Values Using the Empirical Rule Graph:** When you have a problem that gives you the mean, standard deviation, and the percentage, they want you to find the data values. To do this, you look to see which already known percentage it is (68%, 95%, and 99.7%). When you find the proper percentage on the graph, follow the lines to the proper equations and plug in your mean and standard deviations. From there, you will have to generate the data values.
- **Example of Interpreting Values Using the Empirical Rule Graph:** You are looking to find the ages of the kids who have taken swimming lessons at a facility. The mean age is 8 and the standard deviation is 2. What is 95% of the kids’ ages? We first look to find how many standard deviations away is the 95% mark on the graph.



When we mark the lines, we see the proper formulas. Now, we plug into the formulas to find the values:

$$\mu - 2\sigma \quad \text{and} \quad \mu + 2\sigma$$

$$8 - 2(2) = 8 - 4 = 4 \quad \text{and} \quad 8 + 2(2) = 8 + 4 = 12$$

95% of the kids in the swim lessons are between the ages 4 and 12.

Chebyshev's Inequality

- **How to Find Percentages from "k" Standard Deviations:** It is easy to find the percentage with "k" standard deviations with Chebyshev's Inequality. When $k > 1$, you square the "k", divide 1 by the squared "k", subtract that number from 1, and then multiply the subtraction by 100%. The following is the formula:

$$\left(1 - \frac{1}{k^2}\right) \cdot 100\%$$

- **Example of Finding the Percentages from "k" Standard Deviations:** You are looking for the percentage of kids who have taken swimming lessons at a facility. The mean of the data is 8 and the standard deviation is 2. What is the minimum percentage of the ages of kids within 2.5 standard deviations?

$$\left(1 - \frac{1}{2.5^2}\right) \cdot 100\% = \left(1 - \frac{1}{6.25}\right) \cdot 100\% = (1 - .16) \cdot 100\% = (.84) \cdot 100\% = 84\%$$

The minimum percentage of the ages of kids within 2.5 standard deviations is 84%.

- **How to Find Percentages from Values:** If you are given the mean, standard deviation, and two data values, you can find the minimum percentage using Chebyshev's inequality. You first must find the "k". To do so, you plug in your information into the following formulas:

$$\mu - k_1\sigma = x_1 \quad \text{and} \quad \mu + k_2\sigma = x_2 \quad \text{where} \quad x_1 < x_2$$

If k_1 and k_2 are the same, then you plug the "k" into Chebyshev's inequality to find the percentage. Follow the steps from the "How to Find Percentages from "k" Standard Deviations" above on how to use the inequality where you know the "k".

- **Example of Finding Percentages from Values:**
 - You are looking for the percentage of kids who have taken swimming lessons at a facility. The mean of the data is 8 and the standard deviation is 2. What is the minimum percentage of the ages of kids who are between the ages 5 and 11?

To find the percentage, we first must find the “ k ”. We must plug our data values into our formulas:

$$\mu - k_1\sigma = x_1 \quad \text{and} \quad \mu + k_2\sigma = x_2$$

$$8 - 2k_1 = 5 \quad \text{and} \quad 8 + 2k_2 = 11$$

$$k_1 = 1.5 \quad \text{and} \quad k_2 = 1.5$$

Since k_1 and k_2 are the same, we can now use Chebyshev’s inequality to find the percentage.

$$\left(1 - \frac{1}{1.5^2}\right) \cdot 100\% = \left(1 - \frac{1}{2.25}\right) \cdot 100\% = (1 - .44) \cdot 100\% = (.56) \cdot 100\% = 56\%$$

The percentage of kids who are in swimming class that are between the ages of 5 and 11 is 56%.

- **How to find Values from “ k ” Standard Deviations:** To find data values from the “ k ” standard deviations, you just plug into the following formulas:

$$\mu - k\sigma \quad \text{and} \quad \mu + k\sigma$$

- **Example of Finding Values from “ k ” Standard Deviations:**

You are looking for the ages of kids who have taken swimming lessons at a facility. The mean of the data is 8 and the standard deviation is 2. What are the ages of the kids within 2.5 standard deviations?

You plug into the formula with the information to find the values.

$$\begin{aligned} \mu - k\sigma \quad \text{and} \quad \mu + k\sigma \\ 8 - 2.5(2) = 8 - 5 = 3 \quad \text{and} \quad 8 + 2.5(2) = 8 + 5 = 13 \end{aligned}$$

Z-Scores

- **How to Find Z-Scores:** To find the z-score for a value of data, we take the value and subtract the mean from it. After we subtract the two, we divide the subtraction by the standard deviation. We use the same formula for sample and population data sets. The following is both formulas:

Sample Z-Score:

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

Population Z-Score:

$$z = \frac{x - \mu}{\sigma}$$

- **Example of Finding Z-Scores:**

You are looking for the ages of kids who have taken swimming lessons at a facility. The mean of the data is 8 and the standard deviation is 2. How many standard deviations away (what is the z-score) from the mean is a child who goes at the age of 12?

Since, we have our data value, mean, and standard deviation, we can plug into the formula above:

$$z = \frac{x - \mu}{\sigma} = \frac{12 - 8}{2} = \frac{4}{2} = 2$$

A child who is 12 is 2 standard deviations away from the mean of the children’s ages in the swimming lessons.

Percentiles

- **k th Percentile:** When we talk about the k th percentile, we are saying that $k\%$ percent of the data had the value of your exact data value or lower. It is represented by a P_k .

- **Example of How to Interpret the k th Percentile:**

On a math exam, a student's score of 91% makes them part of the 88th percentile. This means that 88% of people who took the math exam scored a 91% or lower on the exam. This, also, means that 12% of people scored higher than 91% on the math exam.

Quartiles and Interquartile Range

- **How to Find the 1st Quartile, 2nd Quartile, and 3rd Quartile:** To find the different quartiles, you first must find the 2nd quartile, also known as the median. We find the 2nd quartile the same way we have found the median in the past. From there, we divide the data set into halves. The data to the left side of the median will be the data used to find the 1st quartile, and the data to the right side of the median is used to find the 3rd quartile. The 1st quartile is the median of the first half of the data. So, we take the data to the left side of the 2nd quartile and mark one off from the left and the right until we are left with the data in the middle. The 3rd quartile is the median of the data to the right side of the 2nd quartile. To find the 3rd quartile, we mark one data off from the left and the right until we have the data in the middle left.
- **How to Find the Interquartile Range:** Find the interquartile range is easy once we have found the 1st quartile and the 3rd quartile. We follow the following formula:

$$IQR = Q_3 - Q_1$$

- **Example of Finding the 1st Quartile, 2nd Quartile, 3rd Quartile, and the Interquartile Range:**

Our data value set is the following:

72, 73, 75, 75, 78, 83, 85, 90, 94, 98

To find the 2nd quartile/median of the data, we mark one on the left and one on the right continuously until we are left with the values in the middle. Since it is an even, we have to add the two values together and then divide by two to find the 2nd Quartile:

$$Q_2 = \frac{78 + 83}{2} = 80.5$$

Now that we have found the 2nd quartile, we can find the 1st quartile. To do that, we take the data to the left side of the median. Because we have an even sample size, we must also take the data value 78. We will mark one off the left and the right until we have a value left in the middle.

72, 73, 75, 75, 78

So, we have now found that $Q_1 = 75$.

Now, we have to find the 3rd quartile. To do this, we take the data to the right side of the median. Because we have an even sample size, we must also take the data value 83. We will mark one off the left and the right until we have a value left in the middle.

83, 85, 90, 94, 98

So, we have found that $Q_3 = 90$.

Now, the last step is to find the interquartile range. To do this, we take the Q_1 and the Q_3 we just found and subtract them from one another.

$$IQR = Q_3 - Q_1 = 90 - 75 = 15$$

* For an easier way to find the Quartiles, you can use our Excel spreadsheet for an example and instructions on how to find the Quartiles in Excel.

Fences and Outliers

- **How to Determine an Outlier:** Determining what an outlier is in the data can be obvious or a little harder than expected. We have a formula that we use to find the cutoff points for data that is considered to not be an outlier, known as fences. To find the fences and therefore the outliers, you must have found the 1st quartile, the 3rd quartile, and the interquartile range. If you have these three already found, you are able to find the two fences. The following formulas are to find our fences:

$$\text{Lower Fence} = Q_1 - 1.5(IQR)$$

$$\text{Upper Fence} = Q_3 + 1.5(IQR)$$

Once we have found our lower fence and upper fence, we look to see if any of our data is below the lower fence or higher than the upper fence. If a data value is below the lower fence or above the upper fence, then that data value is an outlier. If it is not below the lower fence or higher than the upper fence, then there are no outliers in your data set. You can have multiple outliers.

- **Example of Determining an Outlier:**

Using the data set from the example from the “Quartiles and Interquartile Range” section, we already know our 1st quartile, 3rd quartile, and our interquartile range. The following is the data, the 1st quartile, the 3rd quartile, and the interquartile range:

72, 73, 75, 75, 78, 83, 85, 90, 94, 98

$Q_1 = 75, Q_3 = 90, \text{ and } IQR = 15$

With the information, we plug into the formulas for our fences:

$$\text{Lower Fence} = 75 - 1.5(15) = 75 - 22.5 = 52.5$$

$$\text{Upper Fence} = 90 + 1.5(15) = 90 + 22.5 = 112.5$$

Now, we look at our data set to see if any of the data is lower than 52.5 or higher than 112.5. Since no data is smaller or higher, there are no outliers in our data set.

Symbol Guide

Chapter Title Symbols		
Term	Symbol	Use
Population Mean	μ	Identify the population mean
Population Standard Deviation	σ	Identify the population standard deviation
Sample Mean	\bar{x}	Identify the sample mean
Sample Standard Deviation	s	Identify the sample standard deviation
Amount of Standard Deviations	k	Identify the amount of standard deviations used to reach value
Z-Score	z	Identify the z-score
<i>k</i> th Percentile	P_k	Identify the <i>k</i> th percentile
1 st Quartile	Q_1	Identify the first quartile
2 nd Quartile/Median	$Q_2 \text{ or } M$	Identify the second quartile/median
3 rd Quartile	Q_3	Identify the third quartile
Interquartile Range	IQR	Identify the interquartile range