HMEC

Highway Materials Engineering Course

Lesson 6: Analyzing Data

Quality Assurance

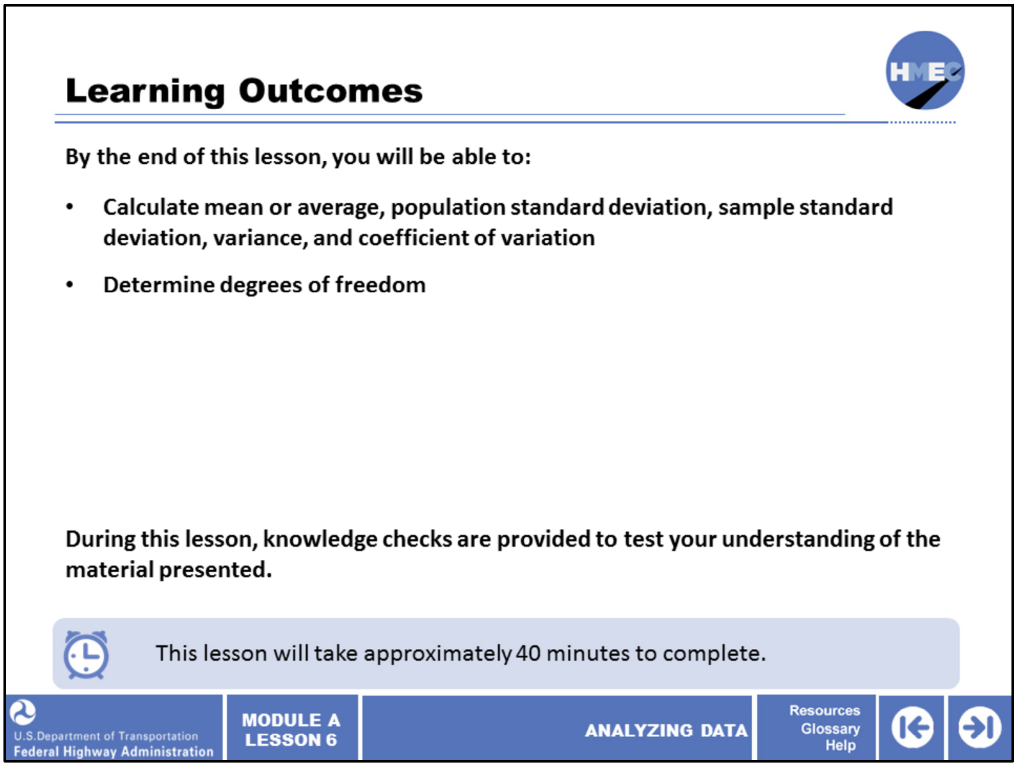
U.S.Department of Transportation
Federal Highway Administration

MODULE A

Welcome to the Highway Materials Engineering Course Module A, Lesson 6: Analyzing Data. This lesson provides an understanding of the basic elements of a statistically-based quality assurance (QA) program and includes an introduction to quality assurance as well as techniques for analyzing data.

A printer-friendly version of the lesson materials can be downloaded by selecting the paperclip icon. A copy of the slides and narration are provided for download.

If you need technical assistance during the training, please select the Help link in the upper right-hand corner of the screen.


HMEC
U.S.Department of Transportation
Federal Highway Administration
MODULE A
LESSON 6
ANALYZING DATA
Resources
Glossary
Help

Learning Outcomes

By the end of this lesson, you will be able to:

- Calculate mean or average, population standard deviation, sample standard deviation, variance, and coefficient of variation
- Determine degrees of freedom

During this lesson, knowledge checks are provided to test your understanding of the material presented.

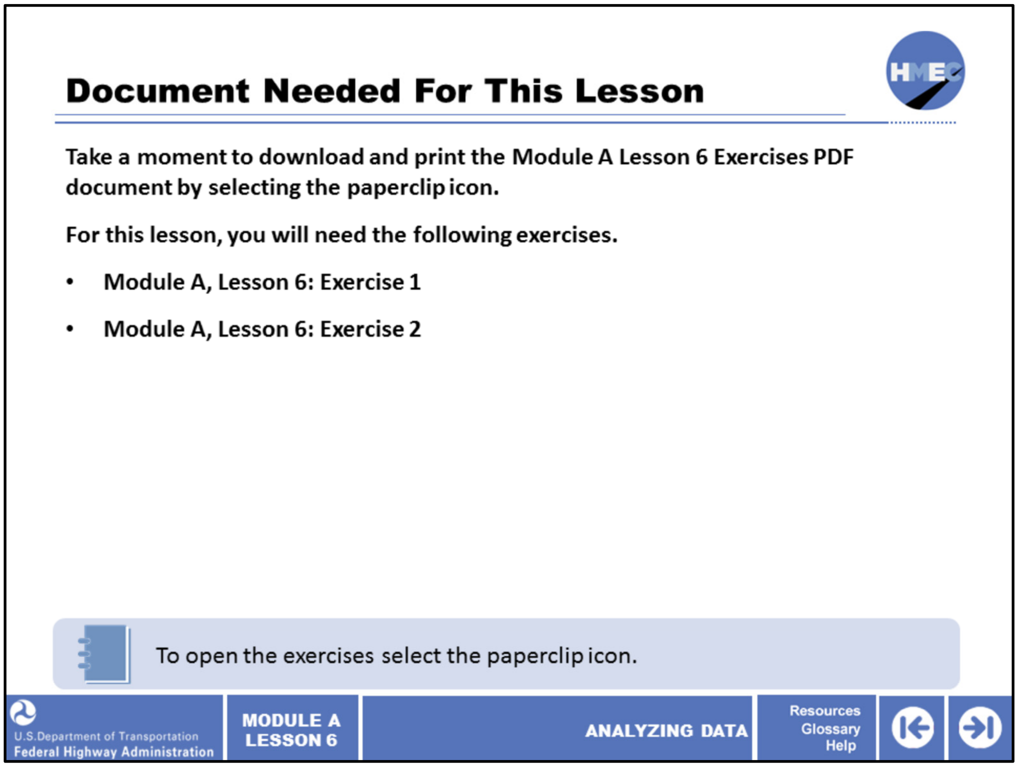
This lesson will take approximately 40 minutes to complete.

By the end of this lesson, you will be able to:

• Calculate mean or average, population standard deviation, sample standard deviation, variance, and coefficient of variation; and
• Determine degrees of freedom.

During this lesson, knowledge checks are provided to test your understanding of the material presented.

This lesson will take approximately 40 minutes to complete.



Document Needed For This Lesson

Take a moment to download and print the Module A Lesson 6 Exercises PDF document by selecting the paperclip icon.

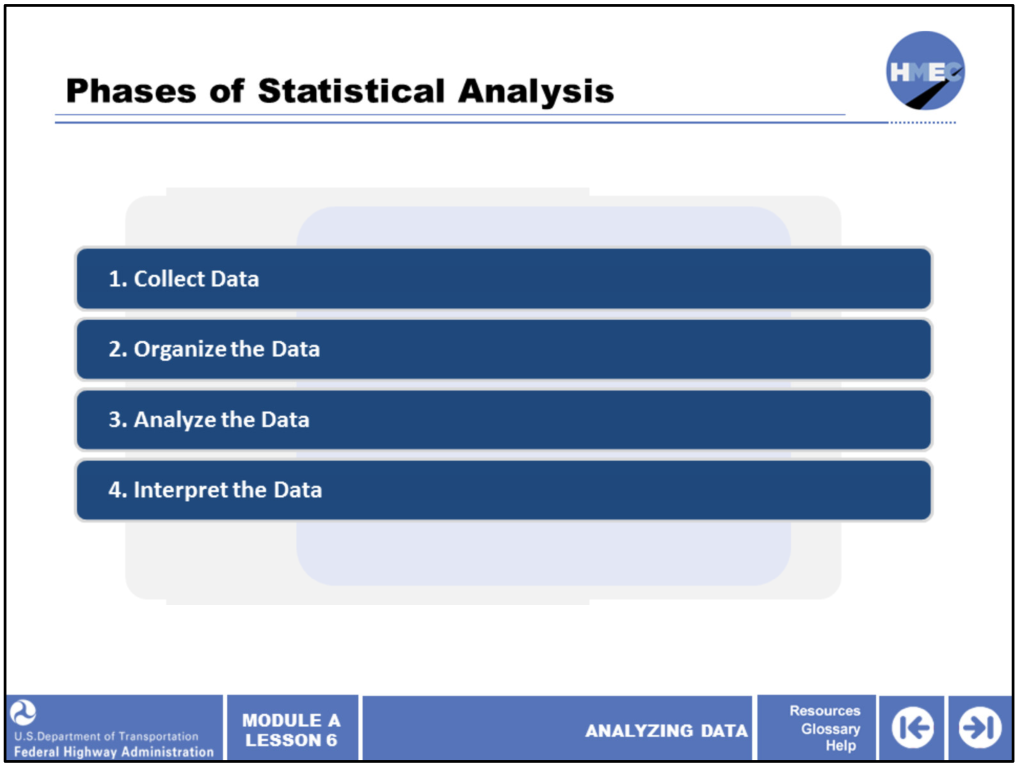
For this lesson, you will need the following exercises.

- **Module A, Lesson 6: Exercise 1**
- **Module A, Lesson 6: Exercise 2**

To open the exercises select the paperclip icon.

HMEC
U.S.Department of Transportation
Federal Highway Administration
MODULE A
LESSON 6
ANALYZING DATA
Resources
Glossary
Help

During this lesson, you will be prompted to reference the lesson exercises document. The referenced document is attached to the lesson in the paperclip icon. Please take a moment to open and print the document.



Phases of Statistical Analysis

1. Collect Data
2. Organize the Data
3. Analyze the Data
4. Interpret the Data

U.S.Department of Transportation
Federal Highway Administration

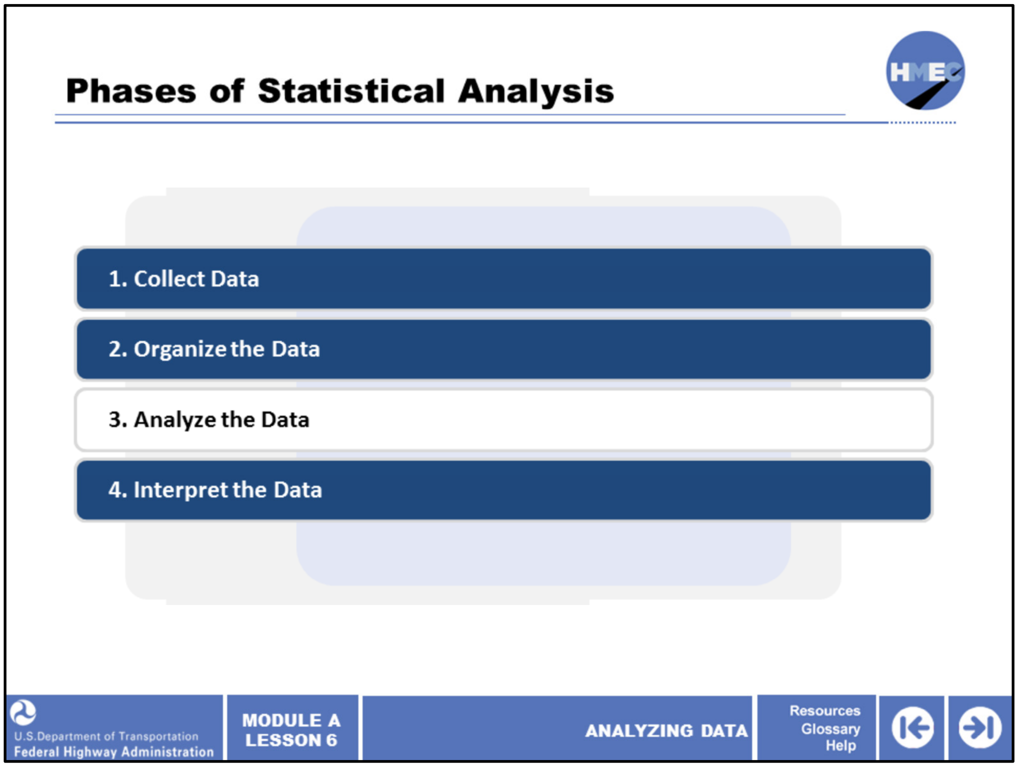
MODULE A
LESSON 6

ANALYZING DATA

Resources
Glossary
Help

Now let’s get started. To review, the four phases of statistical analysis are:

1. Collect data;
2. Organize the data;
3. Analyze the data; and
4. Interpret the data.



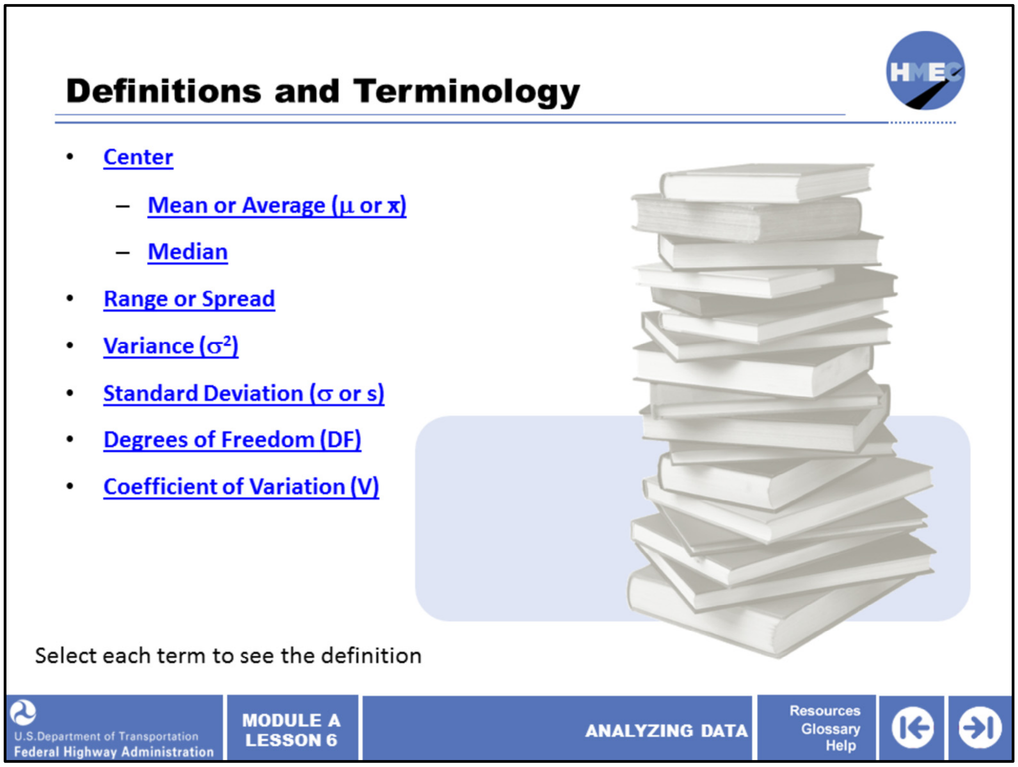
Phases of Statistical Analysis

1. Collect Data
2. Organize the Data
3. Analyze the Data
4. Interpret the Data

HMEC
U.S.Department of Transportation
Federal Highway Administration
MODULE A
LESSON 6
ANALYZING DATA
Resources
Glossary
Help

In this lesson, we'll be focusing on the third phase of statistical analysis: analyzing the data or the quantitative analysis of data. All four phases of statistical analysis are important, but this phase provides the basic information that will be used to write a comprehensible quality assurance (QA) specification.

This phase is a numerical determination of statistical measures that describe the important characteristics of the data.



Definitions and Terminology

- Center
  - Mean or Average (μ or $\bar{x}$)
  - Median
- Range or Spread
- Variance (σ^2)
- Standard Deviation (σ or s)
- Degrees of Freedom (DF)
- Coefficient of Variation (V)

Select each term to see the definition

U.S.Department of Transportation
Federal Highway Administration

MODULE A
LESSON 6

ANALYZING DATA

Resources
Glossary
Help

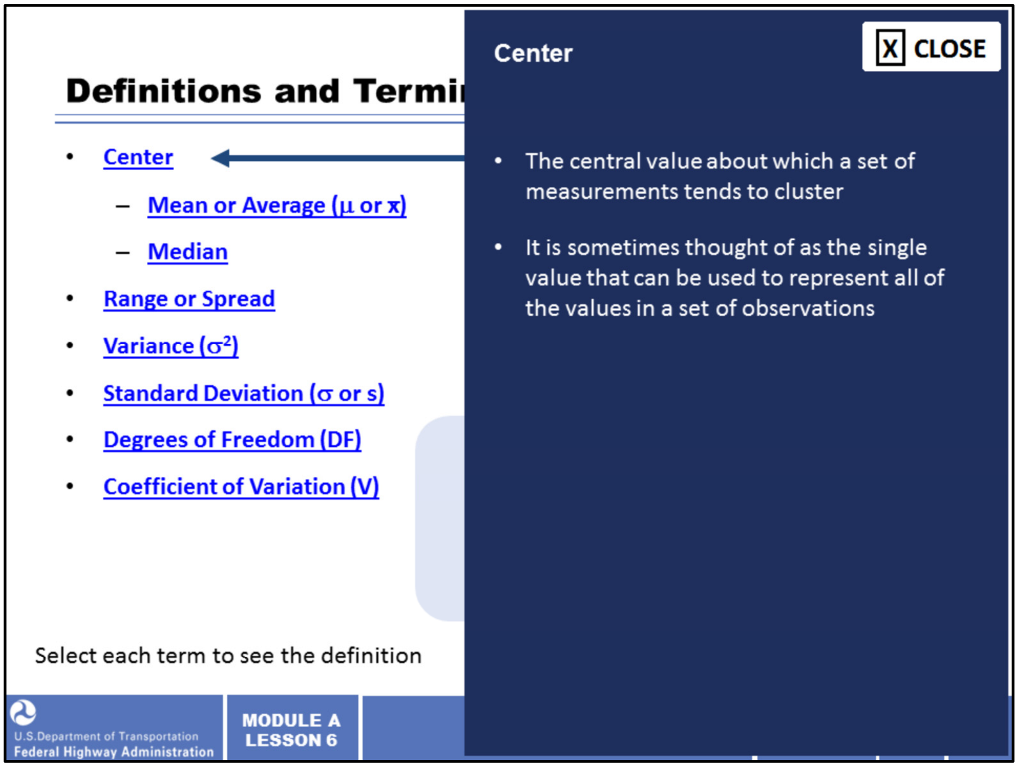
The definitions used in this lesson are important because we rely on them throughout this module. They can also be found in the Glossary.

They are:

- Center;
- Mean or Average (μ *or* $\bar{x}$);
- Median;
- Range or Spread;
- Variance (σ^2);
- Standard Deviation (σ or s);
- Degrees of Freedom (DF); and
- Coefficient of Variation (V).

Select each term to see its definition.

Image Description: Stack of books.



Definitions and Termi

- Center
  - Mean or Average (μ or $\bar{x}$)
  - Median
- Range or Spread
- Variance (σ^2)
- Standard Deviation (σ or s)
- Degrees of Freedom (DF)
- Coefficient of Variation (V)

Select each term to see the definition

Center

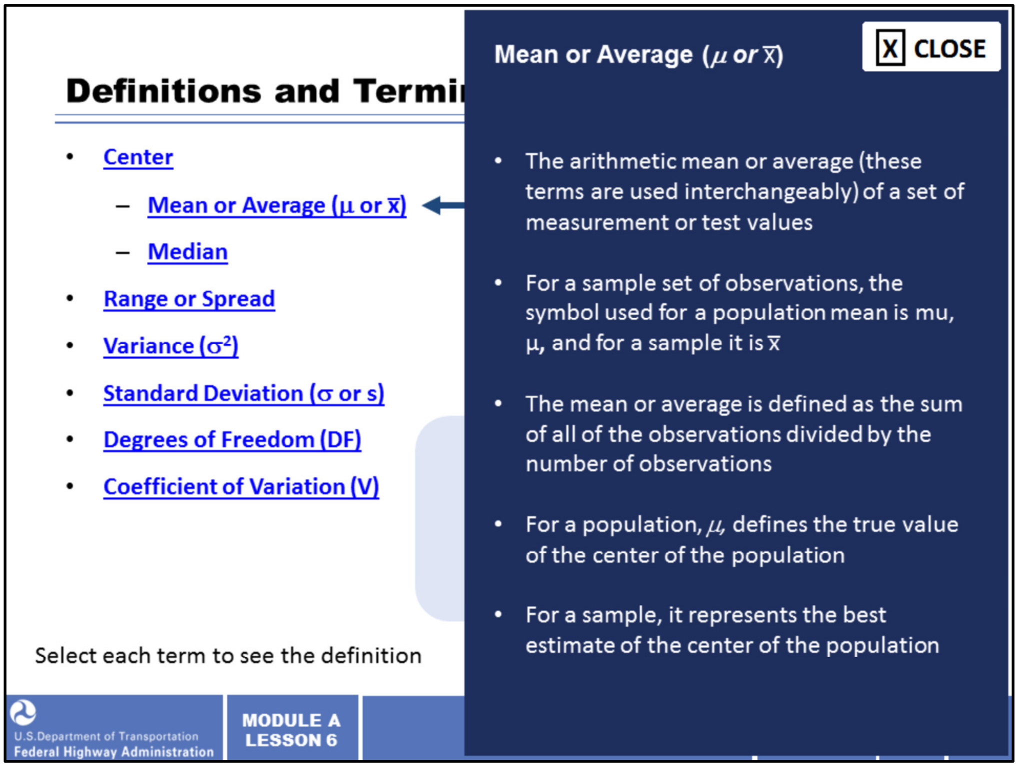
☒ CLOSE

- The central value about which a set of measurements tends to cluster
- It is sometimes thought of as the single value that can be used to represent all of the values in a set of observations

U.S.Department of Transportation
Federal Highway Administration

MODULE A
LESSON 6

Center is the central value about which a set of measurements tends to cluster. It is sometimes thought of as the single value that can be used to represent all of the values in a set of observations. In this module, two measures of the center are used: the mean (or average) and the median.



Definitions and Termi

- Center
  - Mean or Average (μ or $\bar{x}$)
  - Median
- Range or Spread
- Variance (σ^2)
- Standard Deviation (σ or s)
- Degrees of Freedom (DF)
- Coefficient of Variation (V)

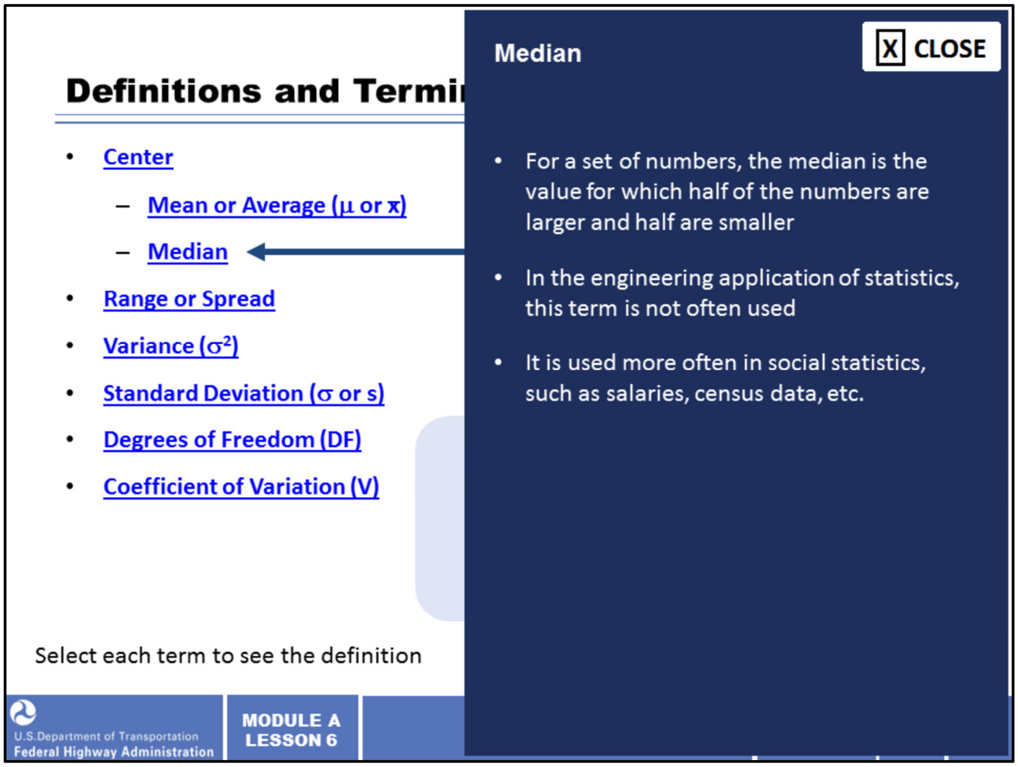
Select each term to see the definition

Mean or Average (μ or $\bar{x}$)

- The arithmetic mean or average (these terms are used interchangeably) of a set of measurement or test values
- For a sample set of observations, the symbol used for a population mean is mu, μ, and for a sample it is $\bar{x}$
- The mean or average is defined as the sum of all of the observations divided by the number of observations
- For a population, μ, defines the true value of the center of the population
- For a sample, it represents the best estimate of the center of the population

☒ CLOSE
U.S.Department of Transportation
Federal Highway Administration
MODULE A
LESSON 6

The mean or average (μ or $\bar{x}$) refers to the arithmetic mean or average (these terms are used interchangeably) of a set of measurement or test values. For a sample set of observations, the symbol used for a population mean is mu, μ, and for a sample it is $\bar{x}$. The mean or average is defined as the sum of all of the observations divided by the number of observations. For a population, μ, defines the true value of the center of the population. For a sample, it represents the best estimate of the center of the population.



Definitions and Termi

- Center
  - Mean or Average (μ or $\bar{x}$)
  - Median
- Range or Spread
- Variance (σ^2)
- Standard Deviation (σ or s)
- Degrees of Freedom (DF)
- Coefficient of Variation (V)

Select each term to see the definition

Median

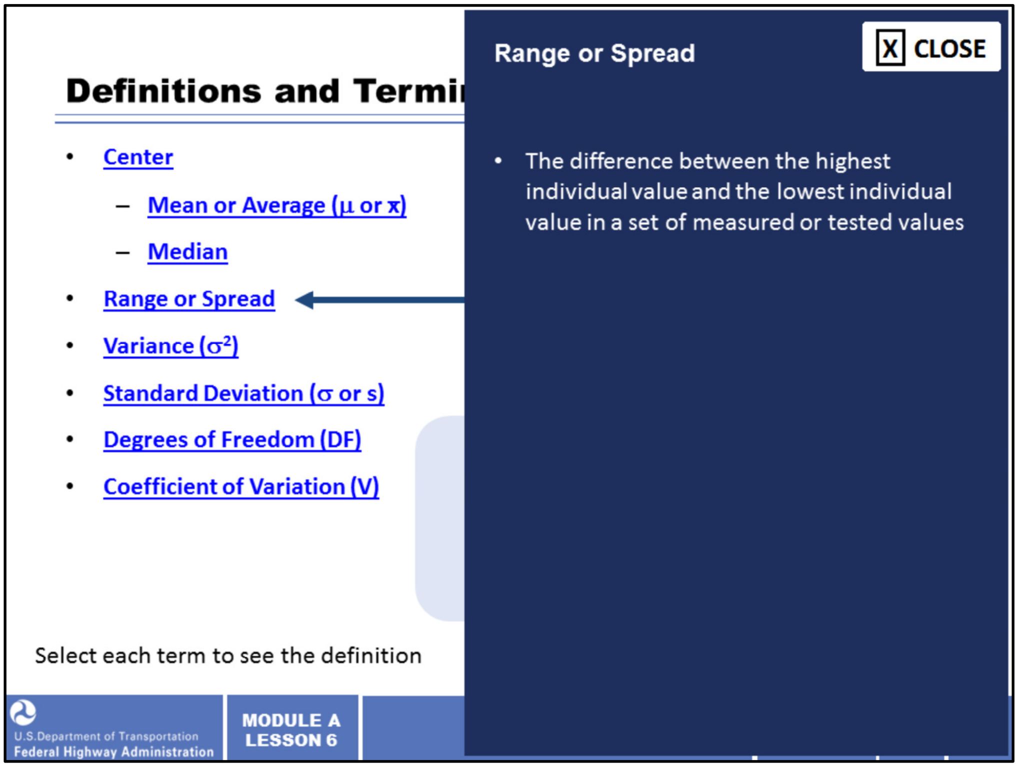
☒ CLOSE

- For a set of numbers, the median is the value for which half of the numbers are larger and half are smaller
- In the engineering application of statistics, this term is not often used
- It is used more often in social statistics, such as salaries, census data, etc.

U.S.Department of Transportation
Federal Highway Administration

MODULE A
LESSON 6

For a set of numbers, the median is the value for which half of the numbers are larger and half are smaller. In the engineering application of statistics, this term is not often used. It is used more often in social statistics, such as salaries, census data, etc.



Definitions and Termi

- Center
  - Mean or Average (μ or $\bar{x}$)
  - Median
- Range or Spread
- Variance (σ^2)
- Standard Deviation (σ or s)
- Degrees of Freedom (DF)
- Coefficient of Variation (V)

Select each term to see the definition

Range or Spread

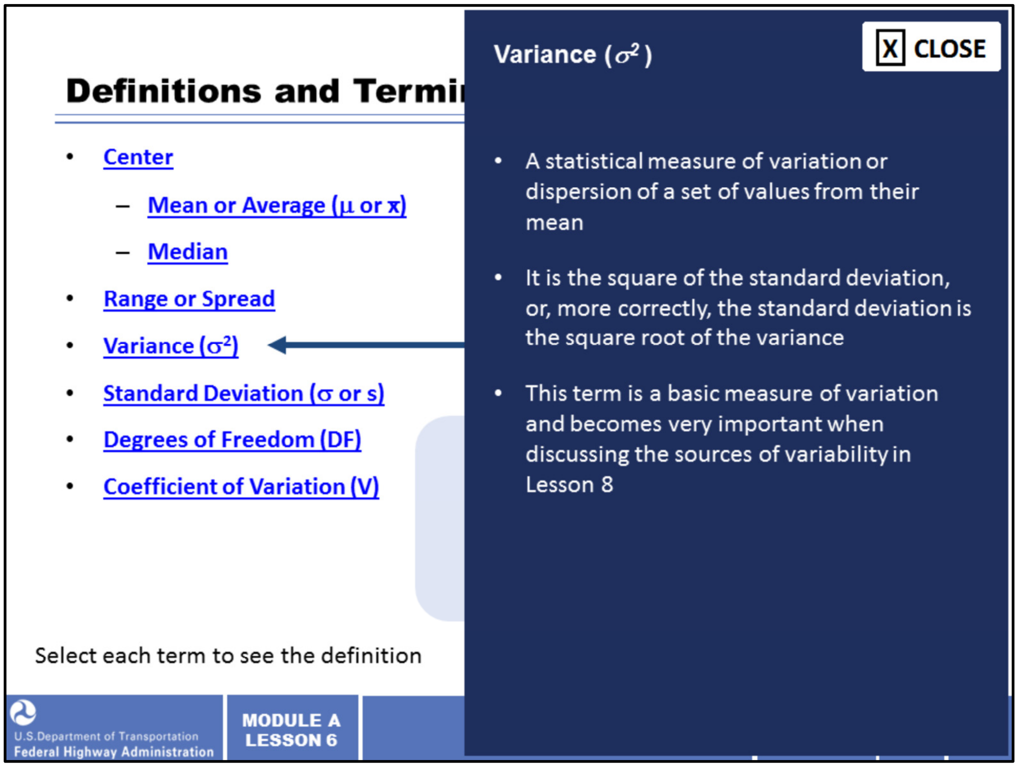
☒ CLOSE

- The difference between the highest individual value and the lowest individual value in a set of measured or tested values

U.S.Department of Transportation
Federal Highway Administration

MODULE A
LESSON 6

Range or spread is the difference between the highest individual value and the lowest individual value in a set of measured or tested values.



Definitions and Termi

- Center
  - Mean or Average (μ or $\bar{x}$)
  - Median
- Range or Spread
- Variance (σ^2)
- Standard Deviation (σ or s)
- Degrees of Freedom (DF)
- Coefficient of Variation (V)

Select each term to see the definition

Variance (σ^2)

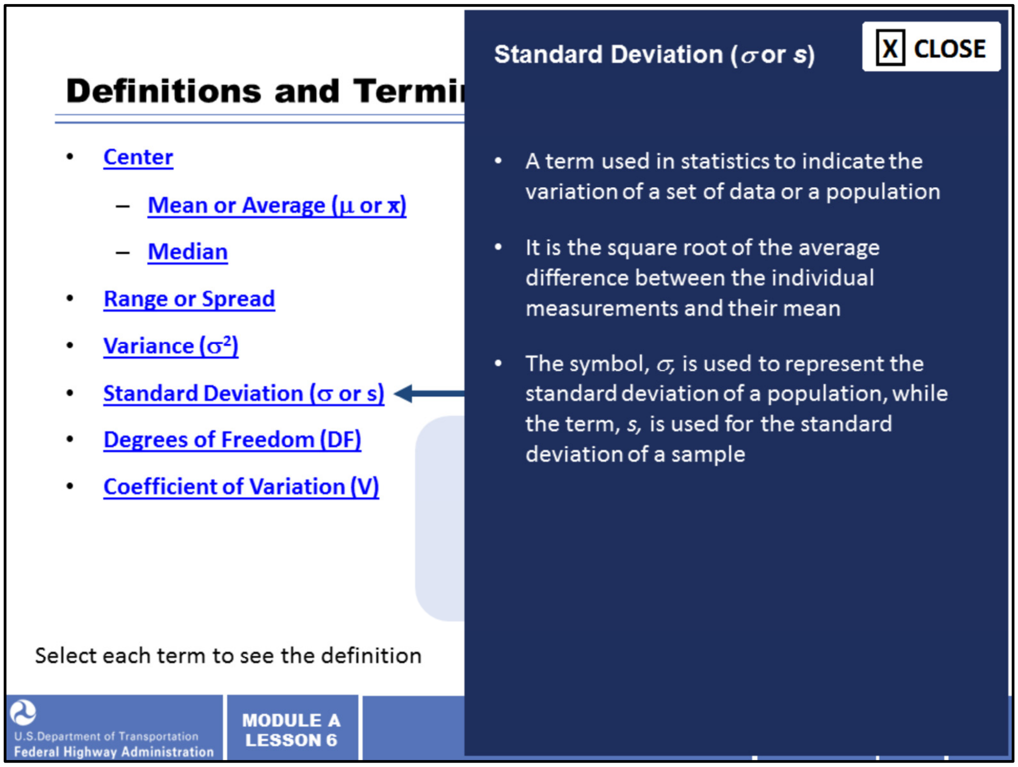
☒ CLOSE

- A statistical measure of variation or dispersion of a set of values from their mean
- It is the square of the standard deviation, or, more correctly, the standard deviation is the square root of the variance
- This term is a basic measure of variation and becomes very important when discussing the sources of variability in Lesson 8

U.S.Department of Transportation
Federal Highway Administration

MODULE A
LESSON 6

Variance (σ^2) is a statistical measure of variation or dispersion of a set of values from their mean. It is the square of the standard deviation, or, more correctly, the standard deviation is the square root of the variance. This term is a basic measure of variation and becomes very important when discussing the sources of variability in Lesson 8.


X CLOSE
U.S.Department of Transportation
Federal Highway Administration
MODULE A
LESSON 6

Definitions and Termi

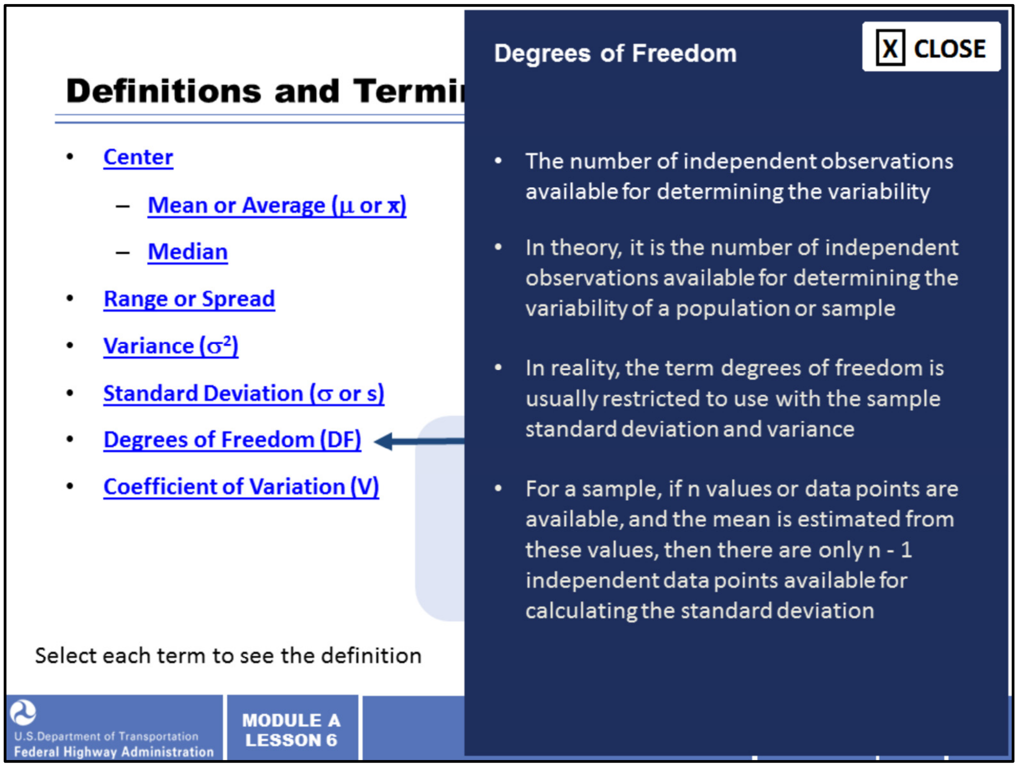
- Center
  - Mean or Average (µ or $\bar{x}$)
  - Median
- Range or Spread
- Variance (σ^2)
- Standard Deviation (σ or s)
- Degrees of Freedom (DF)
- Coefficient of Variation (V)

Select each term to see the definition

Standard Deviation (*σ* or *s*)

- A term used in statistics to indicate the variation of a set of data or a population
- It is the square root of the average difference between the individual measurements and their mean
- The symbol, *σ,* is used to represent the standard deviation of a population, while the term, *s,* is used for the standard deviation of a sample

Standard deviation (*σ* or *s*) is a term used in statistics to indicate the variation of a set of data or a population. It is the square root of the average difference between the individual measurements and their mean. The symbol, *σ,* is used to represent the standard deviation of a population, while the term, *s,* is used for the standard deviation of a sample.



Definitions and Termi

- Center
  - Mean or Average (μ or $\bar{x}$)
  - Median
- Range or Spread
- Variance (σ^2)
- Standard Deviation (σ or s)
- Degrees of Freedom (DF)
- Coefficient of Variation (V)

Select each term to see the definition

Degrees of Freedom

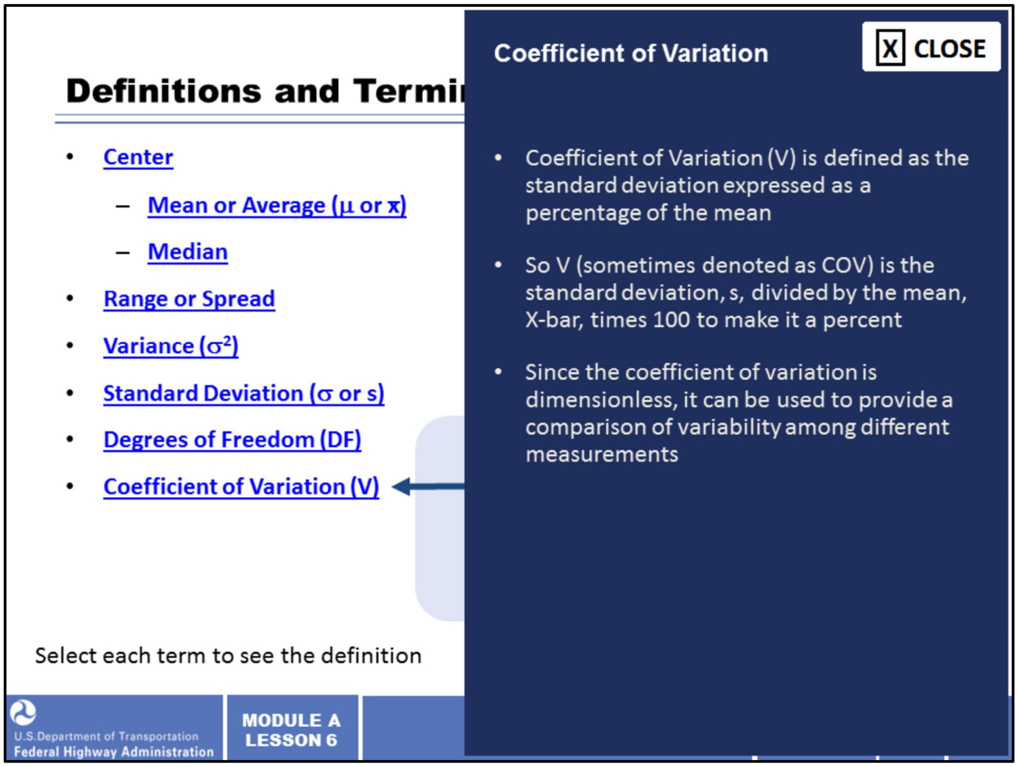
☒ CLOSE

- The number of independent observations available for determining the variability
- In theory, it is the number of independent observations available for determining the variability of a population or sample
- In reality, the term degrees of freedom is usually restricted to use with the sample standard deviation and variance
- For a sample, if n values or data points are available, and the mean is estimated from these values, then there are only n - 1 independent data points available for calculating the standard deviation

U.S.Department of Transportation
Federal Highway Administration

MODULE A
LESSON 6

Degrees of freedom is the number of independent observations available for determining the variability. In theory, it is the number of independent observations available for determining the variability of a population or sample. In reality, the term degrees of freedom is usually restricted to use with the sample standard deviation and variance. For a sample, if n values or data points are available, and the mean is estimated from these values, then there are only n - 1 independent data points available for calculating the standard deviation. For a population, since the true mean can be calculated, not estimated, there are still N independent values available for calculating σ.



Definitions and Termi

- Center
  - Mean or Average (μ or $\bar{x}$)
  - Median
- Range or Spread
- Variance (σ^2)
- Standard Deviation (σ or s)
- Degrees of Freedom (DF)
- Coefficient of Variation (V)

Select each term to see the definition

Coefficient of Variation

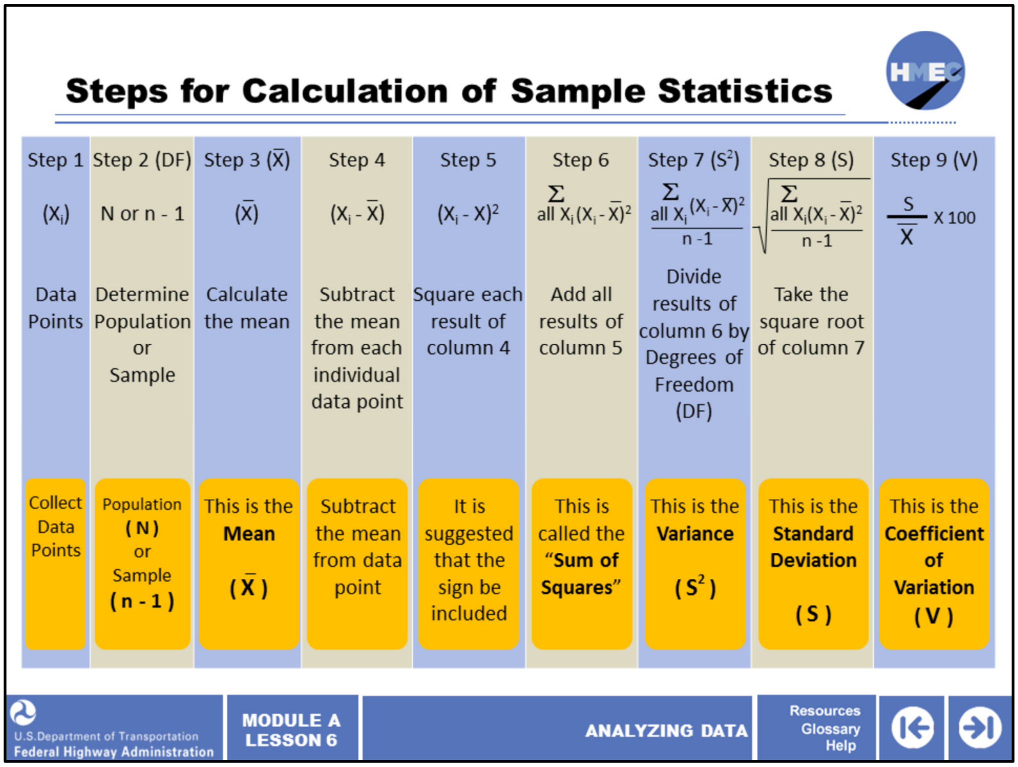
☒ CLOSE

- Coefficient of Variation (V) is defined as the standard deviation expressed as a percentage of the mean
- So V (sometimes denoted as COV) is the standard deviation, s, divided by the mean, X-bar, times 100 to make it a percent
- Since the coefficient of variation is dimensionless, it can be used to provide a comparison of variability among different measurements

U.S.Department of Transportation
Federal Highway Administration

MODULE A
LESSON 6

Coefficient of variation (V) is defined as the standard deviation expressed as a percentage of the mean. So V (sometimes denoted as COV) is the standard deviation, s, divided by the mean, X-bar, times 100 to make it a percent. Since the coefficient of variation is dimensionless, it can be used to provide a comparison of variability among different measurements.



Steps for Calculation of Sample Statistics

Step 1	Step 2 (DF)	Step 3 $(\bar{X})$	Step 4	Step 5	Step 6	Step 7 (S^2)	Step 8 (S)	Step 9 (V)
(X_i)	N or n - 1	$(\bar{X})$	$(X_i - \bar{X})$	$(X_i - X)^2$	$\sum_{\text{all } X_i} (X_i - \bar{X})^2$	$\frac{\sum_{\text{all } X_i} (X_i - \bar{X})^2}{n - 1}$	$\sqrt{\frac{\sum_{\text{all } X_i} (X_i - \bar{X})^2}{n - 1}}$	$\frac{S}{\bar{X}}$ X 100
Data Points	Determine Population or Sample	Calculate the mean	Subtract the mean from each individual data point	Square each result of column 4	Add all results of column 5	Divide results of column 6 by Degrees of Freedom (DF)	Take the square root of column 7	
Collect Data Points	Population **(N)** or Sample **(n - 1)**	This is the **Mean** **$(\bar{X})$**	Subtract the mean from data point	It is suggested that the sign be included	This is called the **“Sum of Squares”**	This is the **Variance** **(S^2)**	This is the **Standard Deviation** **(S)**	This is the **Coefficient of Variation** **(V)**

U.S.Department of Transportation
Federal Highway Administration

MODULE A
LESSON 6

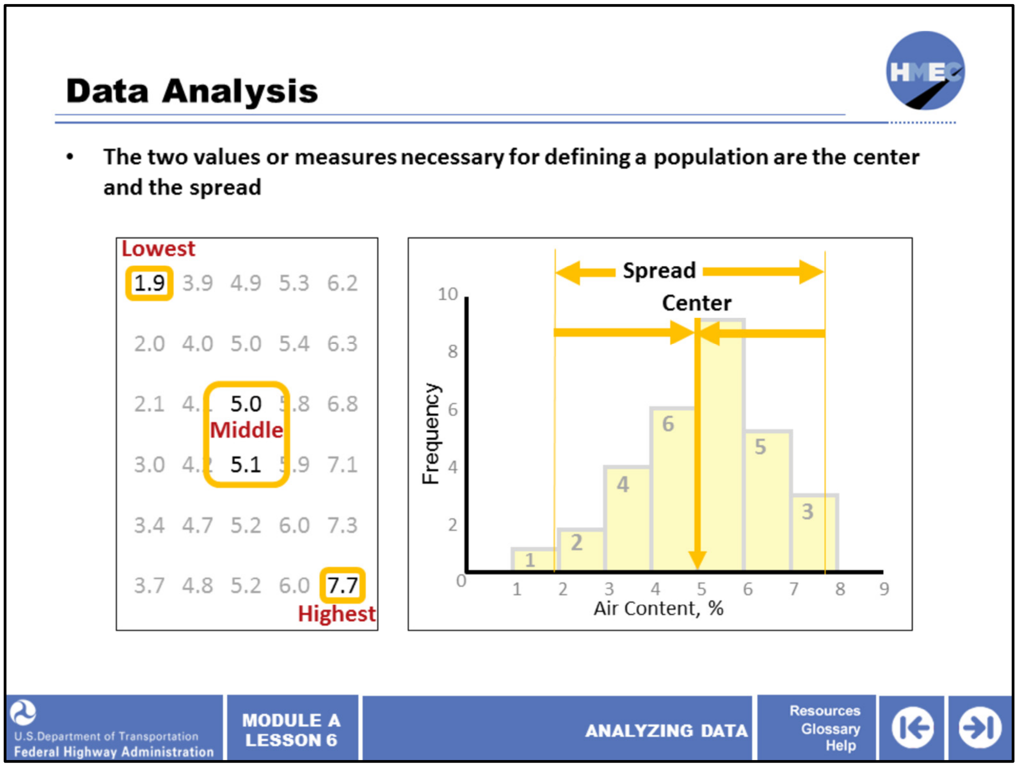
ANALYZING DATA

Resources
Glossary
Help

Now that we have reviewed the definitions, let’s move onto the steps for calculation of sample statistics. They are:

1. Collect data points;
2. Determine if degrees of freedom (DF) is a population or sample standard deviation. If this is a sample statistic, the denominator is subtracted by one. If this was a population statistic, the denominator would not be subtracted by one;
3. Calculate the mean of the four test results;
4. Subtract the mean from the individual result;
5. Square the result from the previous column. It is suggested that the sign be included to allow a check to be made that positive and negative values tend to be equal;
6. Total Step 5. This is called the “sum of squares” and it is the numerator in the equation;
7. Divide the results of the last column and divide by degrees of freedom;
8. The final step is to take the square root of the sum of the last column. This is the standard deviation; and
9. Determine the coefficient of variation.

Advance to the next slide and let’s learn more about each step.



Data Analysis

- The two values or measures necessary for defining a population are the center and the spread

Lowest

1.9	3.9	4.9	5.3	6.2
2.0	4.0	5.0	5.4	6.3
2.1	4.	5.0	.8	6.8
3.0	4.	5.1	.9	7.1
3.4	4.7	5.2	6.0	7.3
3.7	4.8	5.2	6.0	7.7

Middle

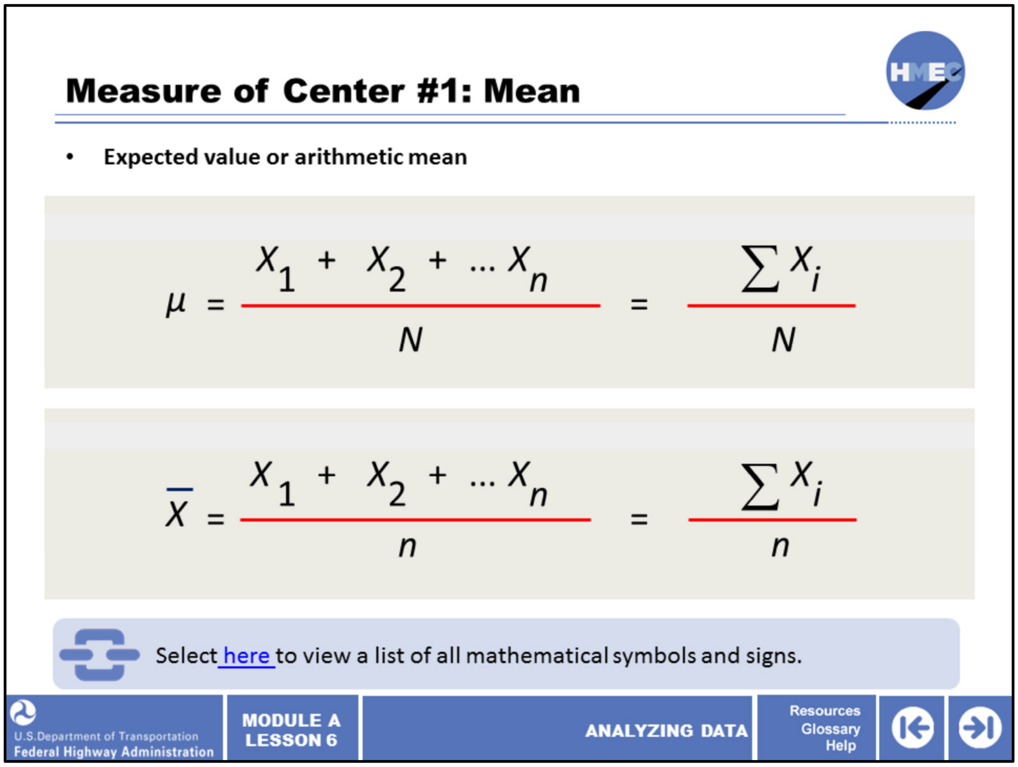
Highest

There are two values or measures necessary for defining a population. They are the center and the spread. In Lesson 5, we saw that the center of the frequency table and histogram were in the middle and tended to separate the data in half. And the spread was the range of data from lowest to highest. Because it is so important, it is repeated that the properties of the population (called population parameters) are needed to describe the product being produced—that is the roadway, stockpile, bridge deck, etc.

But for most analyses, only the properties of the sample (called sample statistics) are available. Thus, in these cases, the sample statistics are used to estimate the properties of the population. The two previous phases of statistical analysis must be done correctly to assure that the phase, “analysis of data” is meaningful. For example, if the data is not collected randomly, when we get to the present phase of analysis, the sample, in all likelihood, will not be a good estimate of the population.

Image description: Data set of thirty numbers displaying lowest 1.9, highest 7.7 and middle 5.0 and 5.1.

Image description: Histogram displaying spread and center.



Measure of Center #1: Mean

- Expected value or arithmetic mean

$$\mu = \frac{X_1 + X_2 + \dots X_n}{N} = \frac{\sum X_i}{N}$$

$$\bar{X} = \frac{X_1 + X_2 + \dots X_n}{n} = \frac{\sum X_i}{n}$$

Select here to view a list of all mathematical symbols and signs.

U.S.Department of Transportation
Federal Highway Administration

MODULE A
LESSON 6

ANALYZING DATA

Resources
Glossary
Help

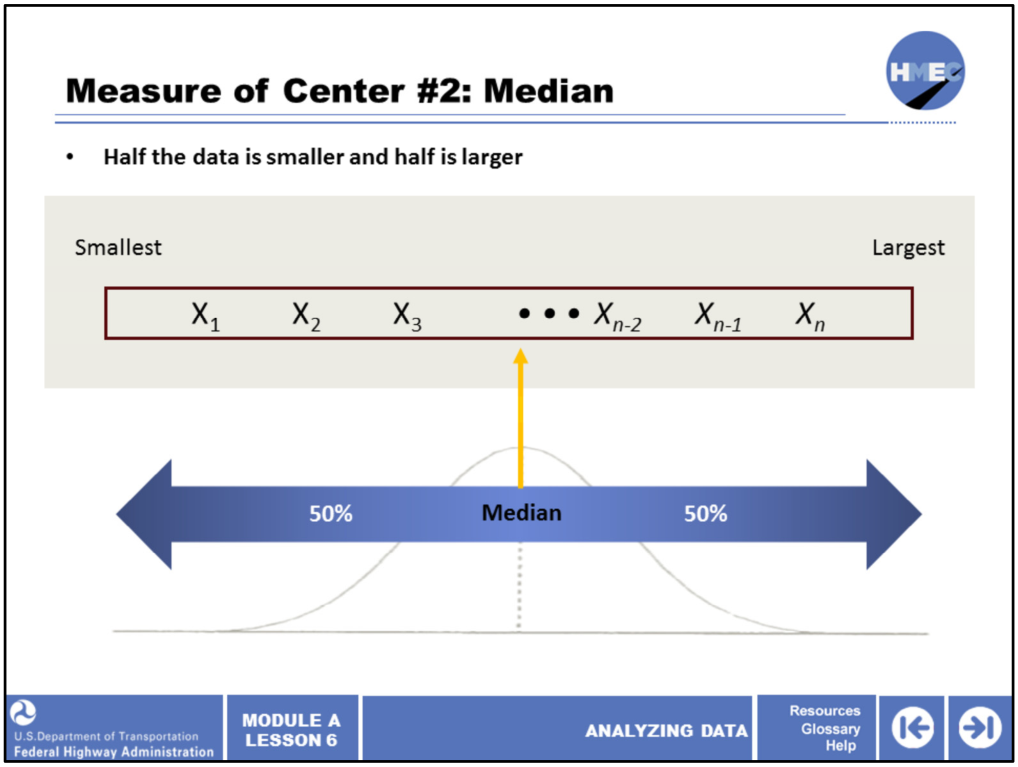
There are several measures of center. The most common, sometimes called the expected value, is the arithmetic mean. The mean, or average, is determined as shown in the slide. The mean of a population is usually denoted as μ and is obtained by adding each individual value, X_1, $X_{2,}$ $X_{3,}$ etc., and dividing by N (the number of values in the population). It is a convention in statistics to use capital N when describing the number of values or data points in a population.

A sample mean is often denoted as an “$\bar{x}$” with a bar over it and called X-bar. The mean is obtained by adding each individual value, X_1, $X_{2,}$ $X_{3,}$ etc. in the sample and dividing by the sample size. Again, the statistical convention is to denote the sample size by the lower case n.

Image description: Equation.

Image description: Equation.

Hyperlink: http://www.rapidtables.com/math/symbols/Basic_Math_Symbols.htm



Measure of Center #2: Median

- Half the data is smaller and half is larger

Smallest

Largest

X_1 X_2 X_3 • • • X_{n-2} X_{n-1} X_n

50% Median 50%

U.S.Department of Transportation
Federal Highway Administration

MODULE A
LESSON 6

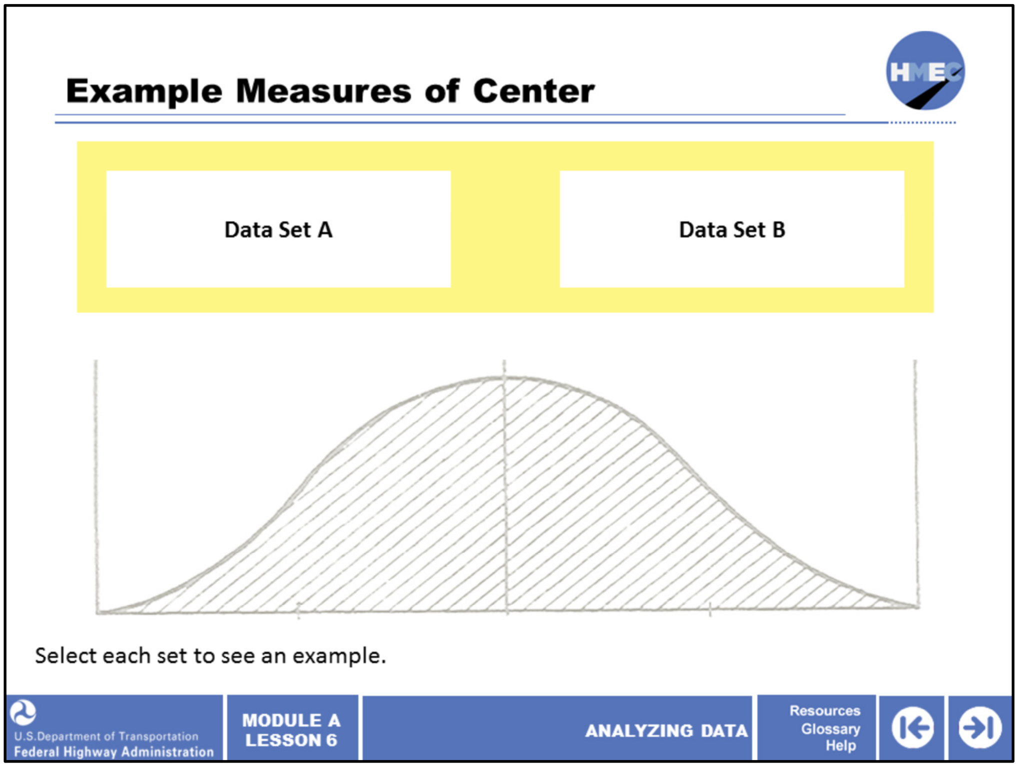
ANALYZING DATA

Resources
Glossary
Help

Another measure of the center of a set of data is the median. The median is the value for which half the data is smaller and half of the data is larger. One advantage of the median is that it is not affected by extreme values (sometimes called outliers) as the mean may be.

Image description: Data from smallest to largest of X_1 , X_2 , X_3 , $X_{n-2,}$, $X_{n-1,}$ $X_{n.}$

Image description: Two side arrow with 50% on each side of median over a distribution curve.



Example Measures of Center

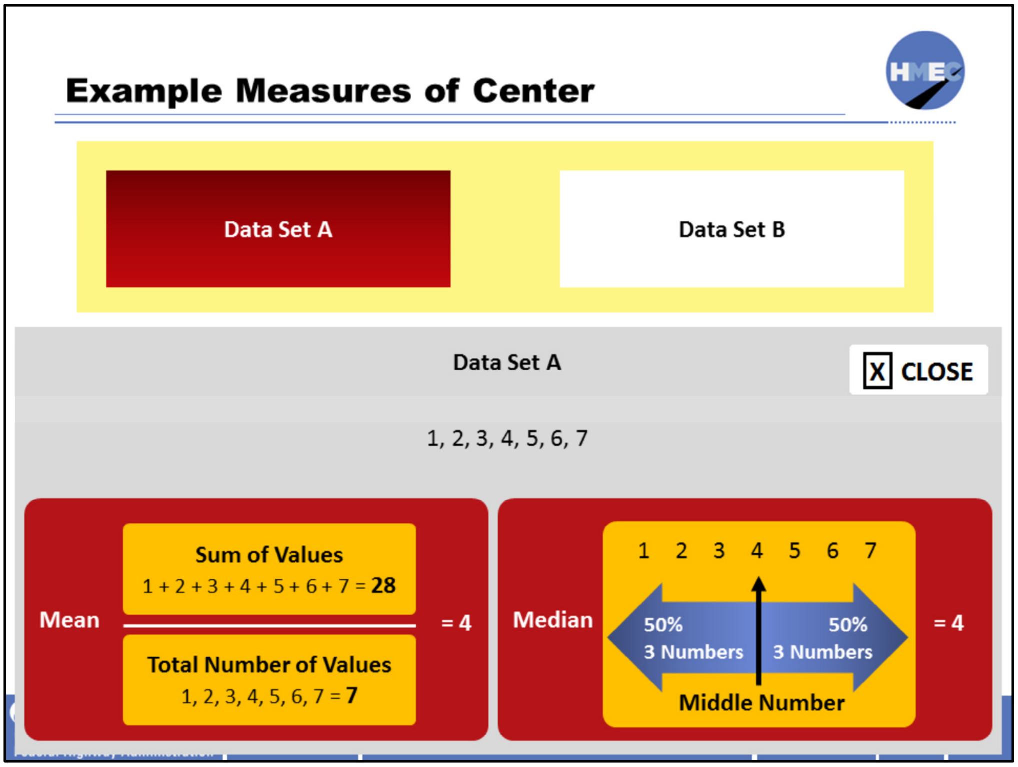
Data Set A

Data Set B

Select each set to see an example.

Here we have two data sets. Select each set for an example.

Image description: Distribution curve.



Example Measures of Center

Data Set A

Data Set B

Data Set A

☒ CLOSE

1, 2, 3, 4, 5, 6, 7

Mean

Sum of Values

1 + 2 + 3 + 4 + 5 + 6 + 7 = **28**

Total Number of Values

1, 2, 3, 4, 5, 6, 7 = **7**

= 4

Median

1 2 3 4 5 6 7

50%
3 Numbers

50%
3 Numbers

Middle Number

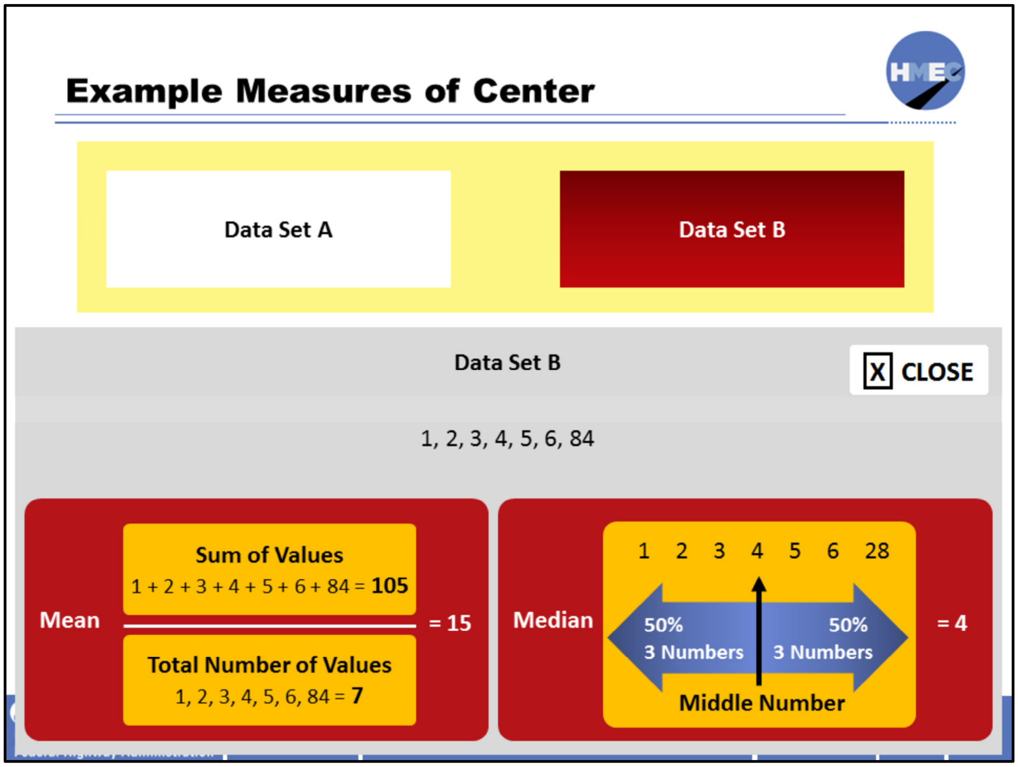
= 4

Data Set A has consecutive numbers from 1 to 7. The mean is the sum of the seven values, 28, divided by the number of values, 7, and is 4. The median, or middle number is also 4, so the mean and median are the same.

Image description: Data set A.

Image description: The sum of the seven values, 28, divided by the number of values, 7, and is 4.

Image description: The median, or middle number is also 4.



Example Measures of Center

Data Set B

1, 2, 3, 4, 5, 6, 84

Mean

$$\frac{\text{Sum of Values} \quad 1+2+3+4+5+6+84 = \mathbf{105}}{\text{Total Number of Values} \quad 1, 2, 3, 4, 5, 6, 84 = \mathbf{7}} = 15$$

Median

1 2 3 4 5 6 28

= 4

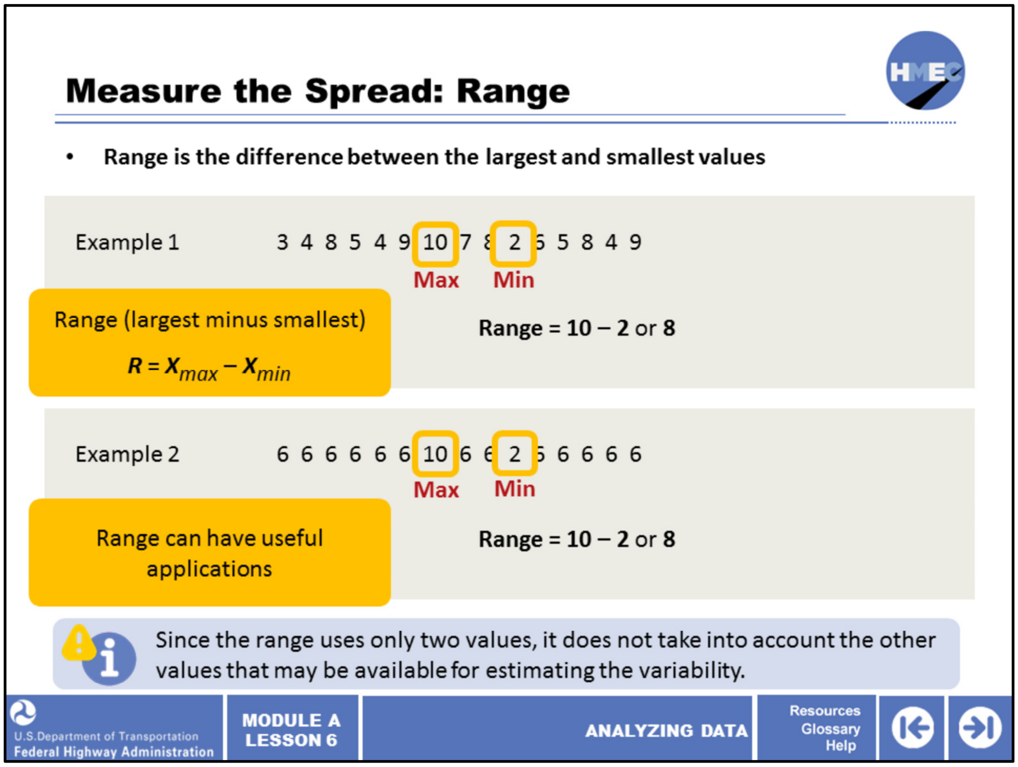
Data Set B has consecutive numbers from 1 to 6 and then the number 84. The mean is the sum of the seven values, 105, divided by the number of values, 7, and is 15. The median, or middle number is still 4, so the highest number, 84 had no effect on the median.

For the two sets of data, the extreme value in Data Set B has a large impact on the mean, but has no effect on the median. As mentioned earlier, the median is used more often in social statistics than in engineering applications. However, for most engineering applications, the influence of extreme values is important. The highest value, 84, is called an “outlier” and this term will be discussed in more detail later.

Image description: Data set B.

Image description: The sum of the seven values, 105, divided by the number of values, 7.

Image description: The median, or middle number is also 4.



Measure the Spread: Range

- Range is the difference between the largest and smallest values

Example 1 3 4 8 5 4 9 10 7 8 2 5 5 8 4 9

Max Min

Range (largest minus smallest)

$R = X_{max} - X_{min}$

Range = 10 – 2 or 8

Example 2 6 6 6 6 6 6 10 6 6 2 6 6 6 6 6

Max Min

Range can have useful applications

Range = 10 – 2 or 8

Since the range uses only two values, it does not take into account the other values that may be available for estimating the variability.

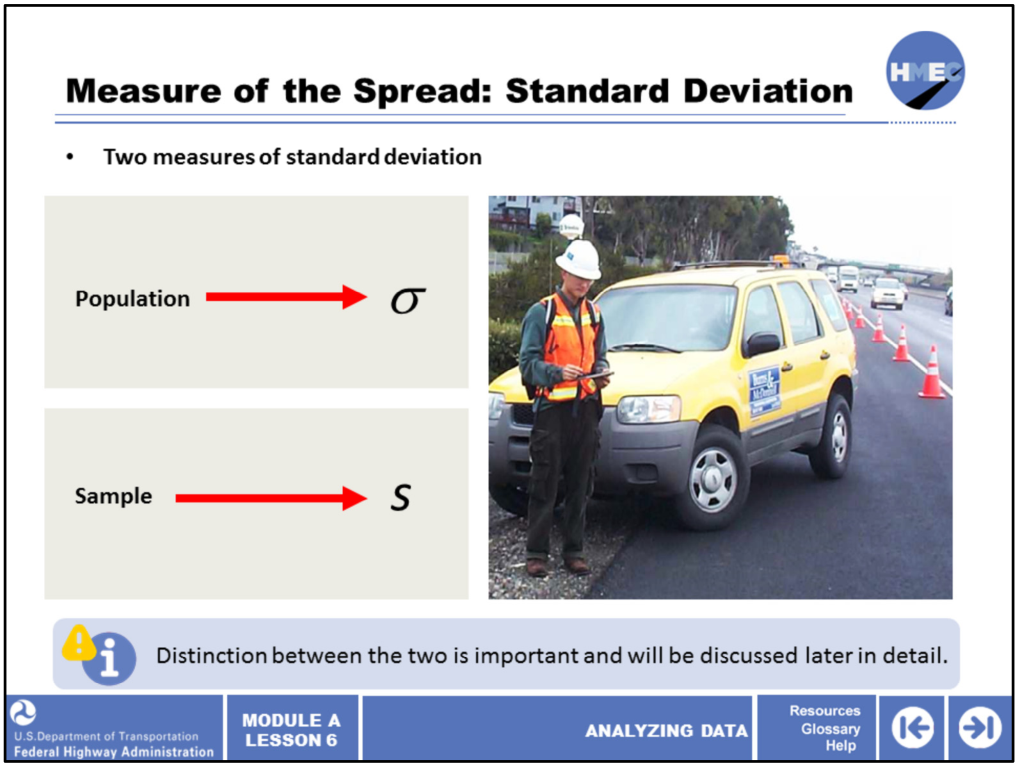
HMEC
U.S.Department of Transportation
Federal Highway Administration
MODULE A
LESSON 6
ANALYZING DATA
Resources
Glossary
Help

The second parameter necessary to define the population is a measure of spread or variability. One common and simple measure is the range, *R*. The range is the difference between the largest and smallest values. We used this parameter in Lesson 5 with the frequency table and histogram.

In this slide, we see two data sets that both have 15 values or data points. The first has a series of values with the highest 10 and the lowest 2, but with many different values in between. The range, *R*, is 10 - 2 = 8. The second data set has 13 values of 6, and a high value of 10 and a low value of 2, so the range is the same as the first set, 8. But it is obvious that the two data sets are from much different populations. So due to its simplistic nature, the range does not provide the best estimate of the variability. Since it uses only two values, it does not take into account the other values that may be available for estimating the variability. It is therefore not as an efficient use of the data as another measure of variability. As we will see in Lesson 9: Quality Control, the range can have useful applications.

Image description: Example 1.

Image description: Example 2.



Measure of the Spread: Standard Deviation

- Two measures of standard deviation

Population → σ

Sample → s

Distinction between the two is important and will be discussed later in detail.

U.S.Department of Transportation
Federal Highway Administration

MODULE A
LESSON 6

ANALYZING DATA

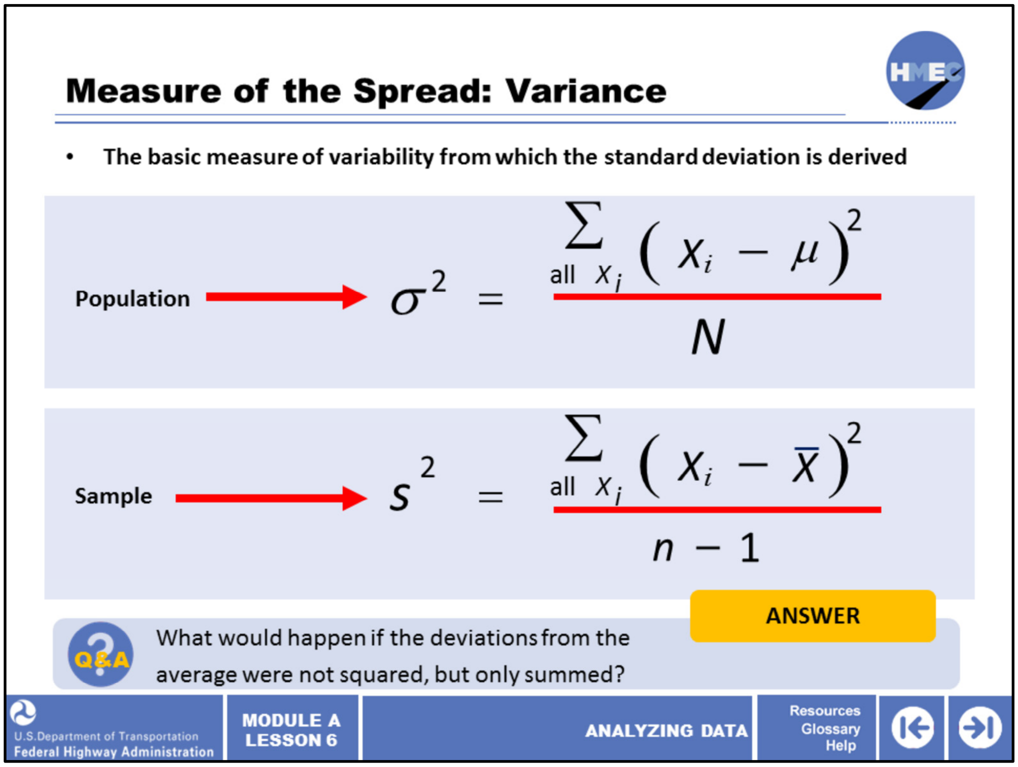
Resources
Glossary
Help

A better measure of the spread or variability is the standard deviation, σ or s. This slide shows the two measures of the standard deviation. It is common practice when dealing with a population standard deviation to use the symbol sigma, σ and when dealing with a sample standard deviation to use a lowercase “s”. This distinction is important and will be discussed in more detail later in this lesson.

Image description: A stretch of road with traffic cones and a construction worker standing in the closed lane working on a tablet.

Image description: Population with an arrow pointing to σ.

Image description: Sample with an arrow pointing to s.



Measure of the Spread: Variance

- The basic measure of variability from which the standard deviation is derived

Population → $$\sigma^2 = \frac{\sum_{\text{all } X_i} (X_i - \mu)^2}{N}$$

Sample → $$s^2 = \frac{\sum_{\text{all } X_i} (X_i - \overline{X})^2}{n - 1}$$

Q&A What would happen if the deviations from the average were not squared, but only summed?

ANSWER

U.S.Department of Transportation Federal Highway Administration | MODULE A LESSON 6 | ANALYZING DATA | Resources Glossary Help

Before discussing the standard deviation as a measure of the spread in more detail, another term is introduced. It is the variance. This is the basic measure of variability from which the standard deviation is derived. This parameter is seen in the equation at the top of the slide. It is the squared deviation of the individual values subtracted from the mean. The symbol for the population variance is sigma squared, σ^2. It is found by first calculating the average of the data set, next subtracting each individual value from the mean and squaring the result, then summing these values—this is the numerator. The last step is to divide the result by the number of data points, N.

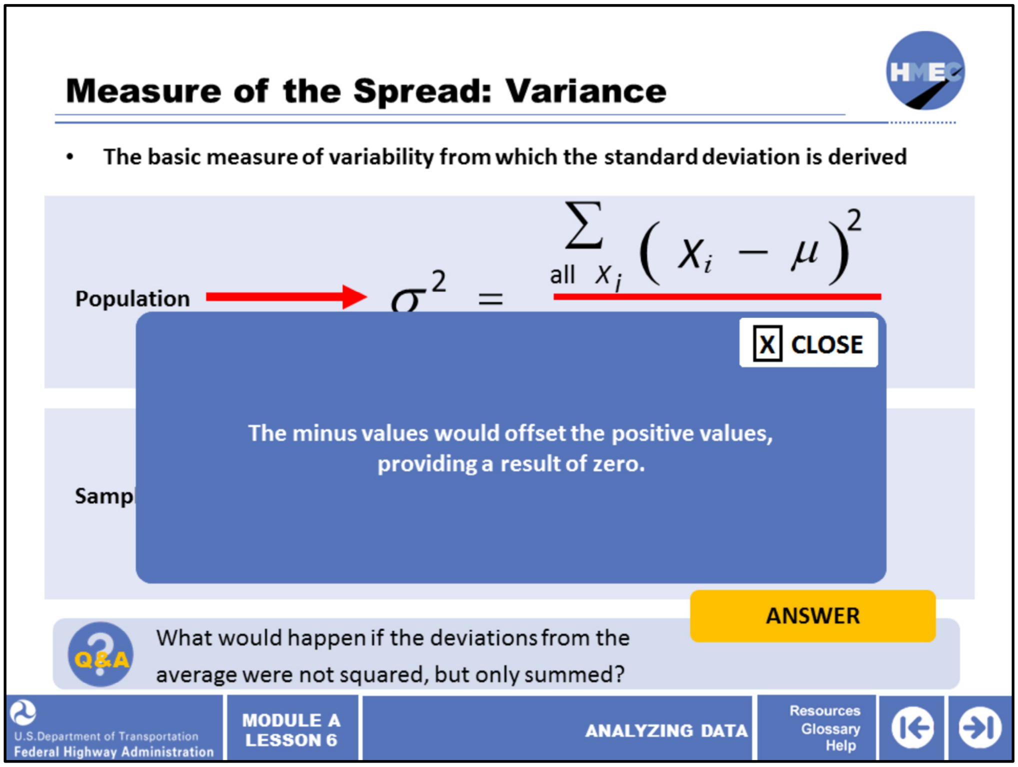
The bottom equation shows how to obtain the sample variance, s^2. It is similar to that of the population variance but has one important difference. The last step, in this case, is to divide the numerator by n - 1, that is one less than the number of data points, which is called the “degrees of freedom.” This term may sound funny, but is extremely important and will be discussed in detail in the next slide. It is obvious that since the value in the numerator is obtained by squaring the differences from the mean, the units obtained will also be squared. Many calculators have function keys for both the population variance and the sample variance. But it is imperative to use the proper calculator function as will be seen later in this lesson.

Since the variance as a measure of variability produces the basic units of measure in units squared, it is sometimes difficult to use. Thus, the standard deviation, which is the square root of the variance, allows the use of the basic unit of measure.

Select the box to answer the question, what would happen if the deviations from the average were not squared, but only summed?

Image description: Population equation.

Image description: Sample equation.



Measure of the Spread: Variance

- The basic measure of variability from which the standard deviation is derived

Population

$$\sigma^2 = \frac{\sum_{\text{all } X_i} (X_i - \mu)^2}{}$$

Samp

☒ CLOSE

The minus values would offset the positive values, providing a result of zero.

ANSWER

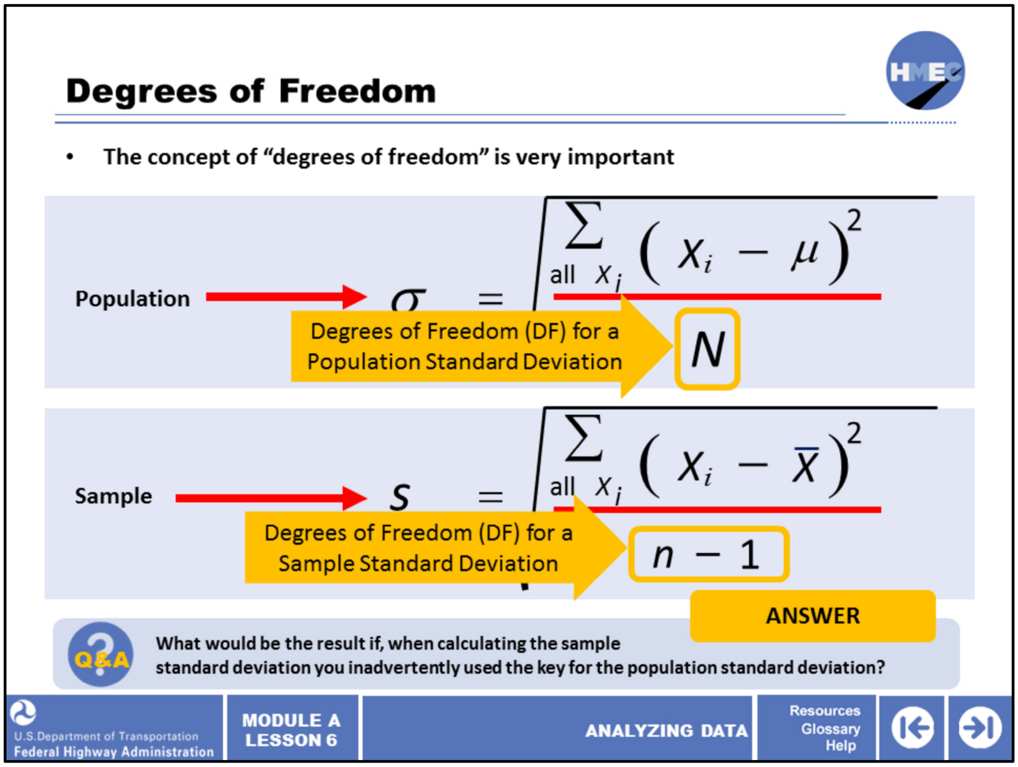
Q&A What would happen if the deviations from the average were not squared, but only summed?

U.S.Department of Transportation
Federal Highway Administration
MODULE A
LESSON 6
ANALYZING DATA
Resources
Glossary
Help

The minus values would offset the positive values, providing a result of zero.

Image description: Population equation.

Image description: Sample equation.



Degrees of Freedom

- The concept of "degrees of freedom" is very important

Population

$$\sigma = \sqrt{\frac{\sum_{\text{all } X_i} \left(X_i - \mu \right)^2}{N}}$$

Degrees of Freedom (DF) for a Population Standard Deviation

Sample

$$s = \sqrt{\frac{\sum_{\text{all } X_i} \left(X_i - \overline{X} \right)^2}{n - 1}}$$

Degrees of Freedom (DF) for a Sample Standard Deviation

ANSWER

Q&A What would be the result if, when calculating the sample standard deviation you inadvertently used the key for the population standard deviation?

U.S.Department of Transportation
Federal Highway Administration
MODULE A
LESSON 6
ANALYZING DATA
Resources
Glossary
Help

The concept of degrees of freedom is a very important one. In this slide, we have taken the equations in the previous slide, for the variance, and taken the square root of both sides. Notice that the denominators used to calculate the sample variance, s^2, and the sample standard deviation, *s*, is n - 1, while the denominators used for the population variance, σ^2, and population standard deviation, σ, is N.

In theory, the denominator in both of these equations is known as the degrees of freedom and is the number of independent observations available for determining the variability. In reality, the term degrees of freedom is usually restricted to use with the sample standard deviation and variance.

For a sample, if n values or data points are available, and the mean is estimated from these values, then there are only n - 1 independent data points available for calculating the standard deviation. In other words, one of the independent data points is lost in calculating the sample variance or sample standard deviation since an estimated value, X-bar, is used in the calculation. A simple analogy for remembering the relationship between n and n - 1 is that with a single value, n, an estimate of the average (although a poor one) can be obtained. However, for an estimate of variability, at least two values are needed; thus n - 1 in this case is 2 - 1 = 1. Again, this would be a poor estimate, but mathematically it is possible.

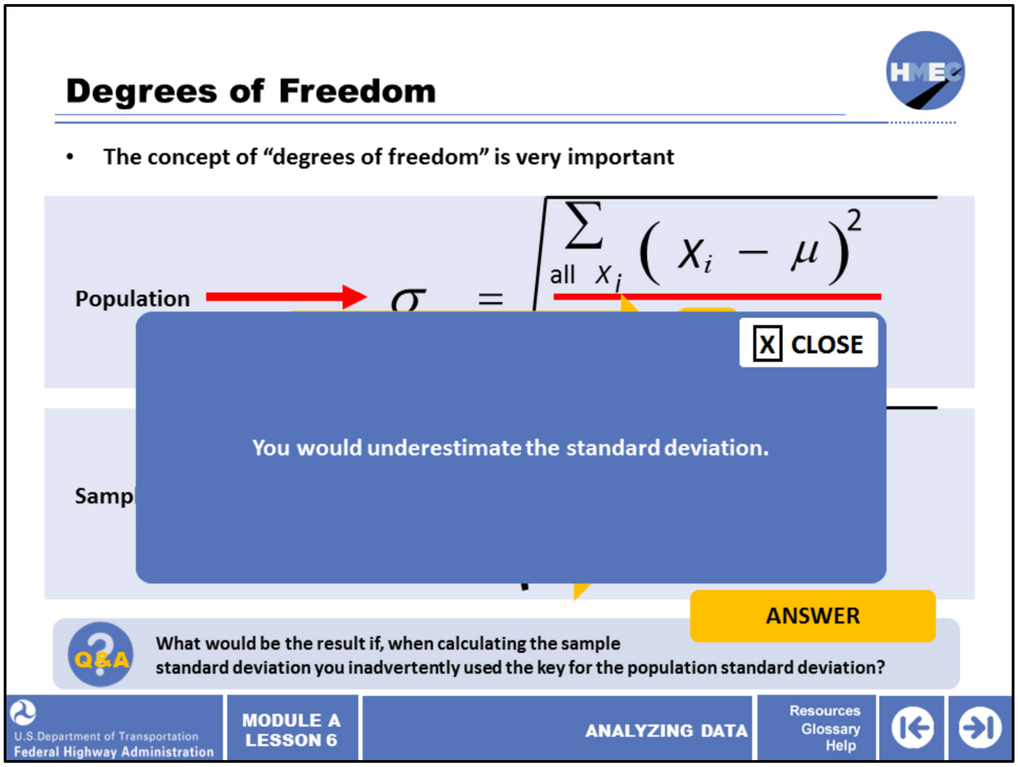
When calculating the standard deviation of a population the degrees of freedom is N because in a population all the data is included and the standard deviation is known, not estimated.

Calculators are available that have keys for the population standard deviation and the sample standard deviation.

Select the box to answer the question: What would be the result if, when calculating the sample standard deviation, you inadvertently used the key for the population standard deviation?

Image description: Population equation.

Image description: Sample equation.



Degrees of Freedom

- The concept of “degrees of freedom” is very important

Population

$$\sigma = \sqrt{\sum_{\text{all } x_i} (x_i - \mu)^2}$$

Samp

☒ CLOSE

You would underestimate the standard deviation.

ANSWER

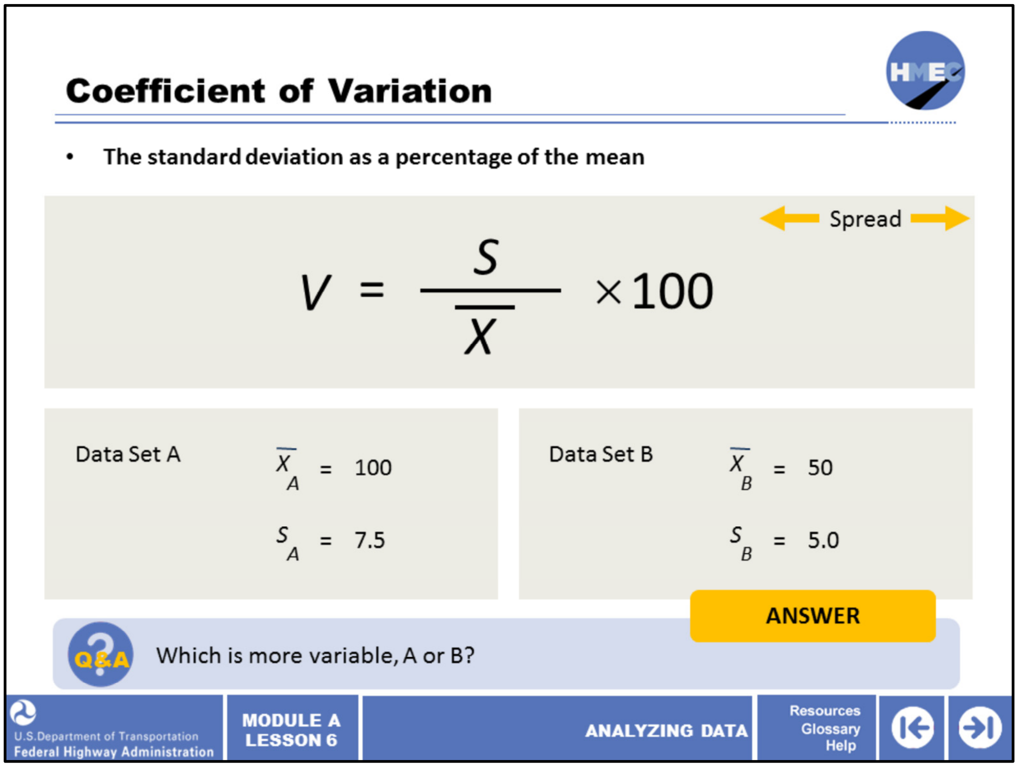
Q&A What would be the result if, when calculating the sample standard deviation you inadvertently used the key for the population standard deviation?

U.S.Department of Transportation
Federal Highway Administration
MODULE A
LESSON 6
ANALYZING DATA
Resources
Glossary
Help

You would underestimate the standard deviation.

Image description: Population equation.

Image description: Sample equation.



Coefficient of Variation

- The standard deviation as a percentage of the mean

Spread

$$V = \frac{s}{\overline{X}} \times 100$$

Data Set A $\overline{X}_A = 100$

$s_A = 7.5$

Data Set B $\overline{X}_B = 50$

$s_B = 5.0$

ANSWER

Q&A Which is more variable, A or B?

U.S.Department of Transportation Federal Highway Administration

MODULE A LESSON 6

ANALYZING DATA

Resources Glossary Help

One additional measure of variability is the coefficient of variation, V or COV. This is defined as the standard deviation as a percentage of the mean. So V or COV is the standard deviation, s, divided by the mean, X-bar, times 100 to make it a percent. Since the coefficient of variation is dimensionless, it can be used to provide a comparison of variability among different measurements. As difficult as quality is to measure, there is some feeling that a higher degree of consistency (lower coefficient of variation) is an indicator of better quality, other factors being the same. The examples given here are:

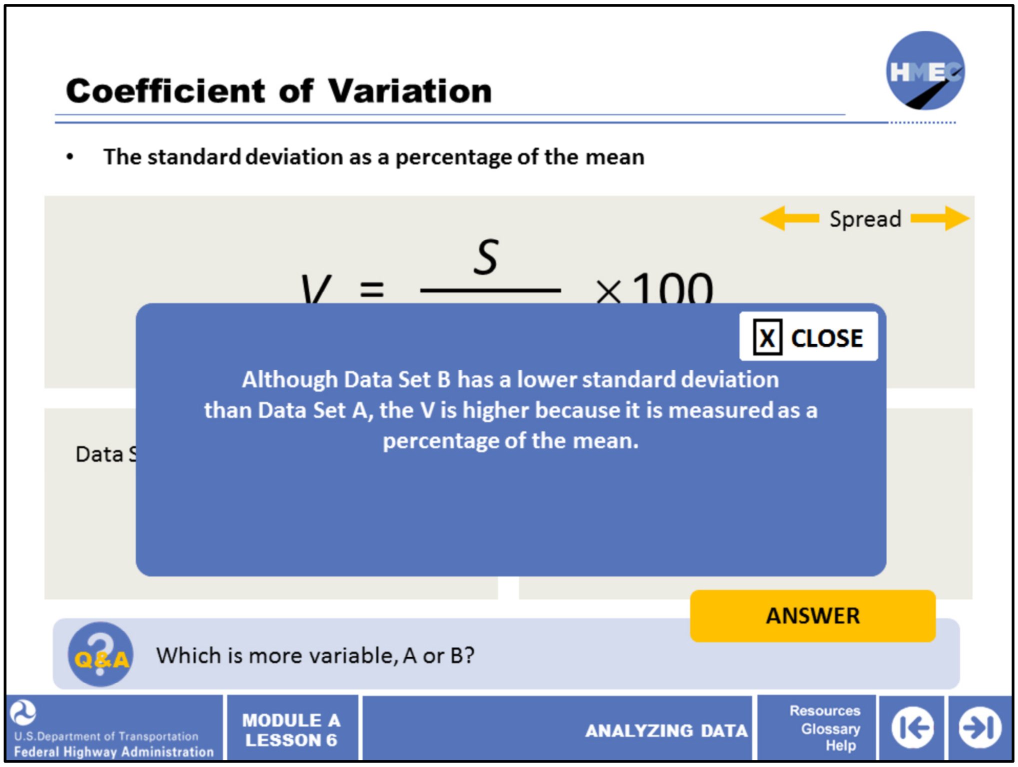
- For Data Set A, the mean X-bar is 100 and the standard deviation is 7.5. Thus, V_A = 7.5/100 x 100 = 7.5%; and
- For Data Set B, the mean X-bar is 50 and the standard deviation is 5.0. Thus, V_B = 5.0/50 x 100 = 10.0%.

Select the box to answer the question, which is more variable, A or B?

Image description: The coefficient of variation, V or COV.

Image description: Data set A.

Image description: Data set B.



Coefficient of Variation

- The standard deviation as a percentage of the mean

$$V = \frac{S}{} \times 100$$

☒ CLOSE

Although Data Set B has a lower standard deviation than Data Set A, the V is higher because it is measured as a percentage of the mean.

Data S

ANSWER

Q&A Which is more variable, A or B?

U.S.Department of Transportation
Federal Highway Administration

MODULE A
LESSON 6

ANALYZING DATA

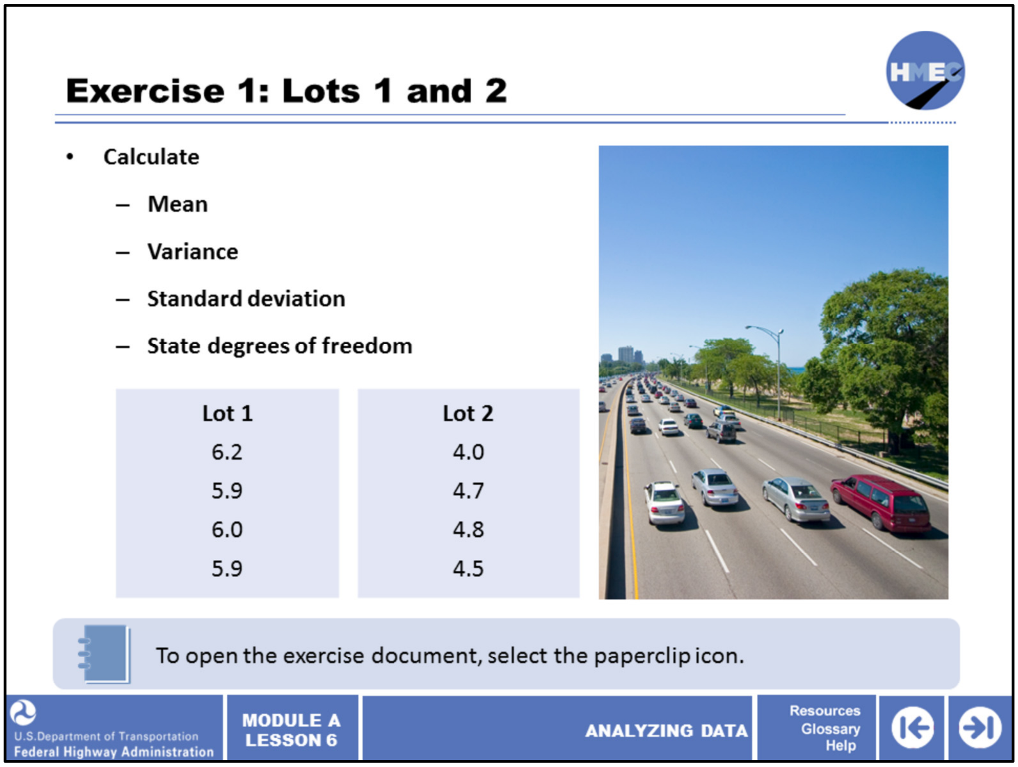
Resources
Glossary
Help

Although Data Set B has a lower standard deviation than Data Set A, the V is higher because it is measured as a percentage of the mean. This measure of variability is often used in Portland cement concrete construction.

Image description: The coefficient of variation, V or COV.

Image description: Data set A.

Image description: Data set B.



Exercise 1: Lots 1 and 2

- Calculate
  - Mean
  - Variance
  - Standard deviation
  - State degrees of freedom

Lot 1	Lot 2
6.2	4.0
5.9	4.7
6.0	4.8
5.9	4.5

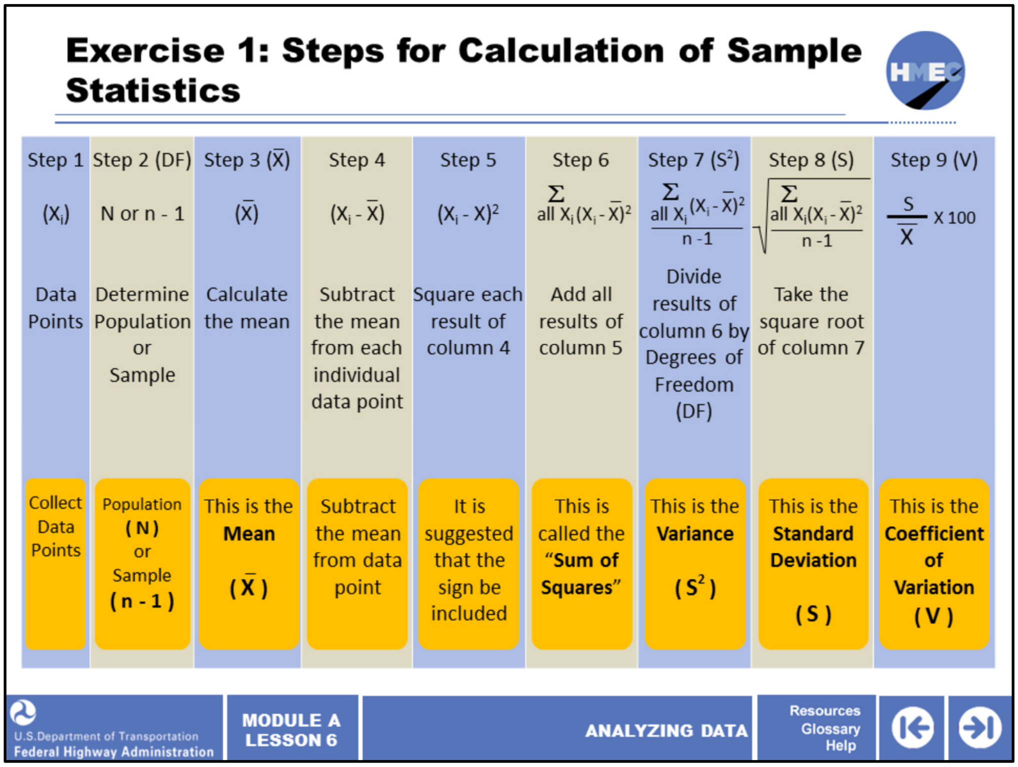
To open the exercise document, select the paperclip icon.

HMEC
U.S.Department of Transportation
Federal Highway Administration
MODULE A
LESSON 6
ANALYZING DATA
Resources
Glossary
Help

Now let’s take a moment for an exercise. Use the exercises document that is provided from the paperclip icon. Please allow 10 minutes to conduct this exercise.

To become familiar with the calculation of sample statistics by long-hand, we have a simple exercise to complete. Here we have four asphalt contents (%) from two lots of asphalt concrete. We want to calculate the mean, variance, and standard deviation of each lot and state the degrees of freedom. The tables allow you to proceed through the steps necessary to do the calculations, understanding that in reality you will use your calculator or other electronic means of accomplishing this task. Advance to the next slide to see the necessary steps.

Image Description: Photo of a stretch of highway.



Exercise 1: Steps for Calculation of Sample Statistics

Step 1	Step 2 (DF)	Step 3 ($\bar{X}$)	Step 4	Step 5	Step 6	Step 7 (S^2)	Step 8 (S)	Step 9 (V)
(X_i)	N or n - 1	$(\bar{X})$	$(X_i - \bar{X})$	$(X_i - X)^2$	$\sum_{\text{all } X_i} (X_i - \bar{X})^2$	$\frac{\sum_{\text{all } X_i} (X_i - \bar{X})^2}{n - 1}$	$\sqrt{\frac{\sum_{\text{all } X_i} (X_i - \bar{X})^2}{n - 1}}$	$\frac{S}{\bar{X}} \times 100$
Data Points	Determine Population or Sample	Calculate the mean	Subtract the mean from each individual data point	Square each result of column 4	Add all results of column 5	Divide results of column 6 by Degrees of Freedom (DF)	Take the square root of column 7	
Collect Data Points	Population **(N)** or Sample **(n - 1)**	This is the **Mean** **($\bar{X}$)**	Subtract the mean from data point	It is suggested that the sign be included	This is called the **"Sum of Squares"**	This is the **Variance** **(S^2)**	This is the **Standard Deviation** **(S)**	This is the **Coefficient of Variation** **(V)**

U.S.Department of Transportation
Federal Highway Administration

MODULE A
LESSON 6

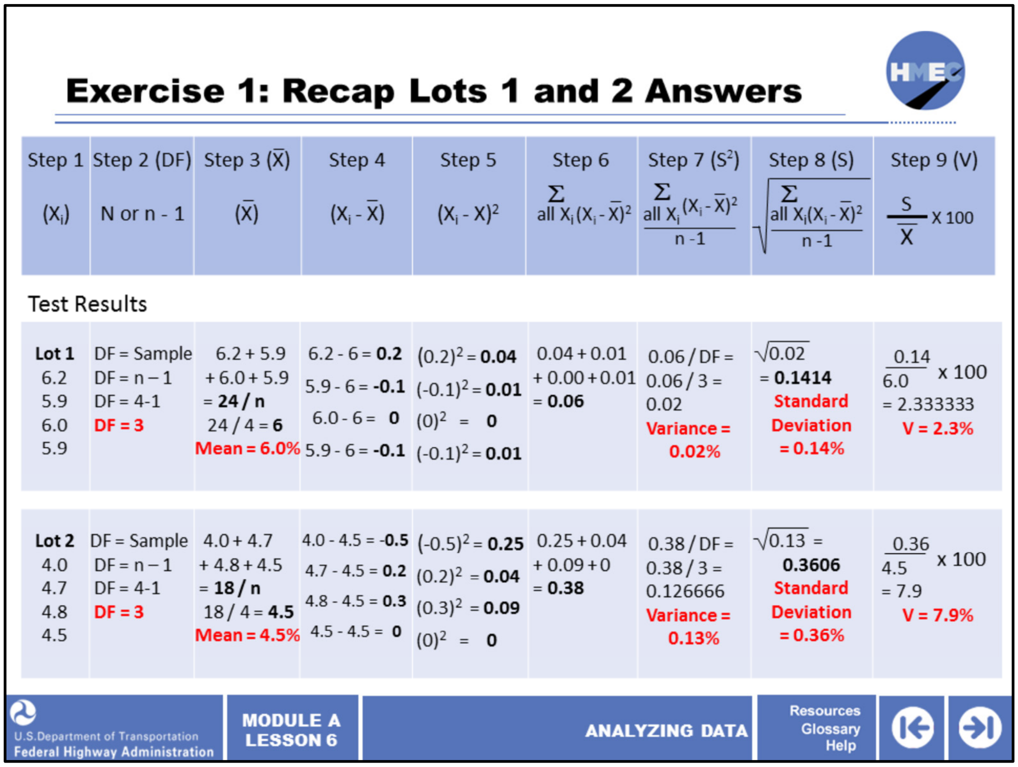
ANALYZING DATA

Resources
Glossary
Help

Again, let's review the steps for calculation of sample statistics. They are:

1. Collect data points;
2. Determine if degrees of freedom (DF) is a population or sample standard deviation. If this is a sample statistic, the denominator is subtracted by one. If this was a population statistic, the denominator would not be subtracted by one;
3. Calculate the mean of the four test results;
4. Subtract the mean from the individual result;
5. Square the result from the previous column. It is suggested that the sign be included to allow a check to be made that positive and negative values tend to be equal;
6. Total Step 5. This is called the "sum of squares" and is the numerator in the equation;
7. Divide the results of the last column and divide by degrees of freedom;
8. The final step is to take the square root of the sum of the last column. This is the standard deviation; and
9. Determine the coefficient of variation.

Advance to the next slide to see the test results and answers.



Exercise 1: Recap Lots 1 and 2 Answers

Step 1	Step 2 (DF)	Step 3 ($\bar{X}$)	Step 4	Step 5	Step 6	Step 7 (S^2)	Step 8 (S)	Step 9 (V)
(X_i)	N or n - 1	$(\bar{X})$	$(X_i - \bar{X})$	$(X_i - X)^2$	$\sum_{\text{all } X_i} (X_i - \bar{X})^2$	$\frac{\sum_{\text{all } X_i} (X_i - \bar{X})^2}{n - 1}$	$\sqrt{\frac{\sum_{\text{all } X_i} (X_i - \bar{X})^2}{n - 1}}$	$\frac{S}{\bar{X}} \times 100$

Test Results

Lot 1 6.2 5.9 6.0 5.9	DF = Sample DF = n – 1 DF = 4-1 **DF = 3**	6.2 + 5.9 + 6.0 + 5.9 = **24 / n** 24 / 4 = **6** **Mean = 6.0%**	6.2 - 6 = **0.2** 5.9 - 6 = **-0.1** 6.0 - 6 = **0** 5.9 - 6 = **-0.1**	$(0.2)^2 = \mathbf{0.04}$ $(-0.1)^2 = \mathbf{0.01}$ $(0)^2 = \mathbf{0}$ $(-0.1)^2 = \mathbf{0.01}$	0.04 + 0.01 + 0.00 + 0.01 = **0.06**	0.06 / DF = 0.06 / 3 = 0.02 **Variance = 0.02%**	$\sqrt{0.02}$ = **0.1414** **Standard Deviation = 0.14%**	$\frac{0.14}{6.0} \times 100$ = 2.333333 **V = 2.3%**
Lot 2 4.0 4.7 4.8 4.5	DF = Sample DF = n – 1 DF = 4-1 **DF = 3**	4.0 + 4.7 + 4.8 + 4.5 = **18 / n** 18 / 4 = **4.5** **Mean = 4.5%**	4.0 - 4.5 = **-0.5** 4.7 - 4.5 = **0.2** 4.8 - 4.5 = **0.3** 4.5 - 4.5 = **0**	$(-0.5)^2 = \mathbf{0.25}$ $(0.2)^2 = \mathbf{0.04}$ $(0.3)^2 = \mathbf{0.09}$ $(0)^2 = \mathbf{0}$	0.25 + 0.04 + 0.09 + 0 = **0.38**	0.38 / DF = 0.38 / 3 = 0.126666 **Variance = 0.13%**	$\sqrt{0.13}$ = **0.3606** **Standard Deviation = 0.36%**	$\frac{0.36}{4.5} \times 100$ = 7.9 **V = 7.9%**

U.S.Department of Transportation
Federal Highway Administration

MODULE A
LESSON 6

ANALYZING DATA

Resources
Glossary
Help

It is common statistical practice to report the mean to the same number of decimals as the raw data and the standard deviation to one more decimal place.

The test results for the lots are:

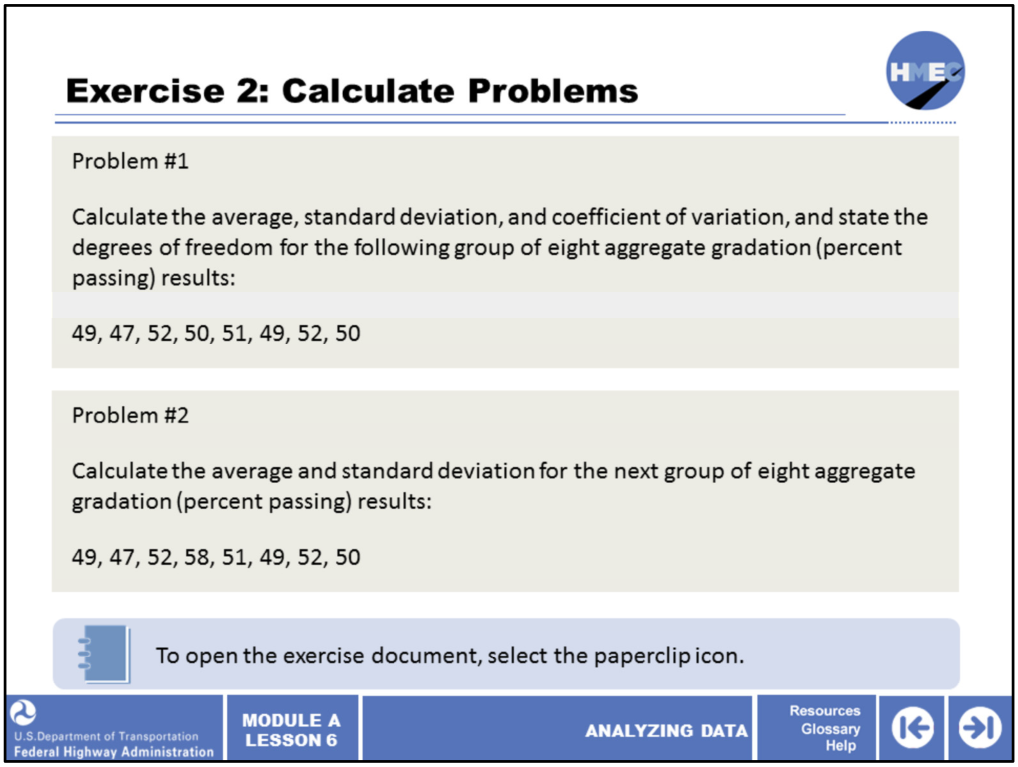
- Lot 1: 6.2, 5.9, 6.0, 5.9; and
- Lot 2: 4.0, 4.7, 4.8, 4.5.

The answers for Lot 1 are:

- Degrees of Freedom = 3;
- Mean = 6.0%;
- Variance = 0.02%;
- Standard Deviation = 0.14%; and
- Coefficient of Variation = 2.3%

The answers for Lot 2 are:

- Degrees of Freedom = 3;
- Mean = 4.5%;
- Variance = 0.13%;
- Standard Deviation = 0.36%; and
- Coefficient of Variation = 7.9%



Exercise 2: Calculate Problems

Problem #1

Calculate the average, standard deviation, and coefficient of variation, and state the degrees of freedom for the following group of eight aggregate gradation (percent passing) results:

49, 47, 52, 50, 51, 49, 52, 50

Problem #2

Calculate the average and standard deviation for the next group of eight aggregate gradation (percent passing) results:

49, 47, 52, 58, 51, 49, 52, 50

To open the exercise document, select the paperclip icon.

U.S.Department of Transportation
Federal Highway Administration

MODULE A
LESSON 6

ANALYZING DATA

Resources
Glossary
Help

Now let’s take a moment for another exercise. Use the exercises document that is provided from the paperclip icon. Please allow 10 minutes to conduct this exercise.

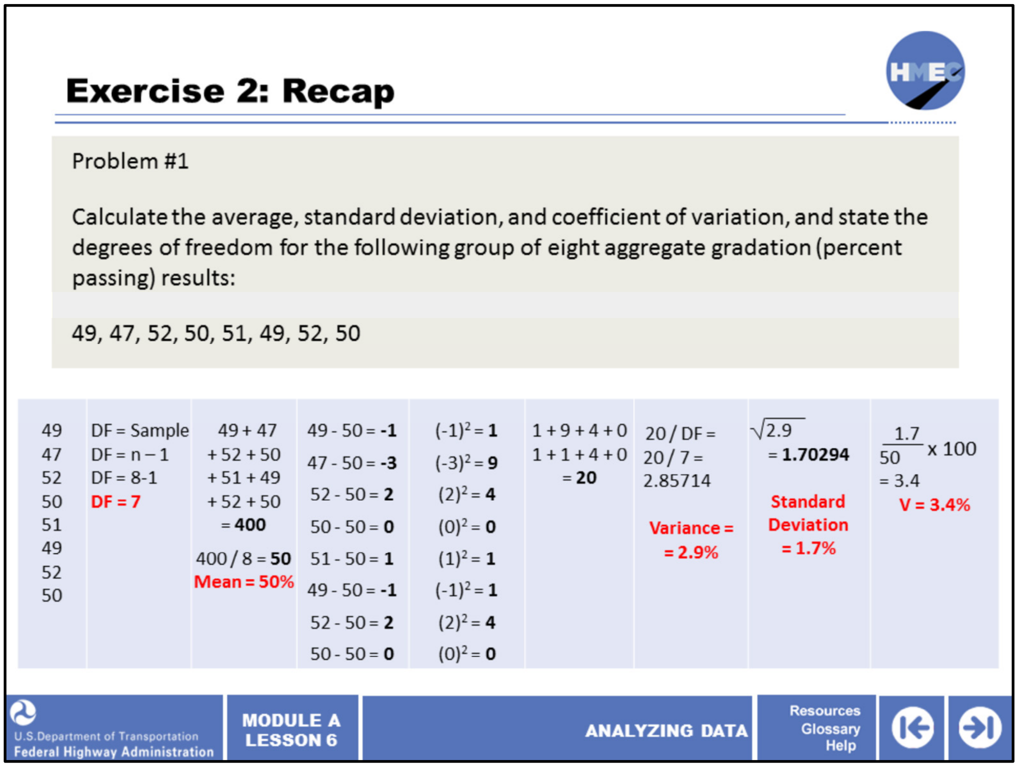
Here we have two calculation problems. This activity will allow you to use your calculator or other electronic device to become familiar with calculating the mean and standard deviation and determining the degrees of freedom.

Problem #1:
Calculate the average, standard deviation, and coefficient of variation and state the degrees of freedom for the following group of eight aggregate gradation (percent passing) results:
49, 47, 52, 50, 51, 49, 52, 50

Problem #2:
Calculate the average and standard deviation, for the next group of eight aggregate gradation (percent passing) results:
49, 47, 52, 58, 51, 49, 52, 50

Image description: Problem #1.

Image description: Problem #2.



Exercise 2: Recap

Problem #1

Calculate the average, standard deviation, and coefficient of variation, and state the degrees of freedom for the following group of eight aggregate gradation (percent passing) results:

49, 47, 52, 50, 51, 49, 52, 50

49 47 52 50 51 49 52 50	DF = Sample DF = n – 1 DF = 8-1 **DF = 7**	49 + 47 + 52 + 50 + 51 + 49 + 52 + 50 = **400** 400 / 8 = **50** **Mean = 50%**	49 - 50 = **-1** 47 - 50 = **-3** 52 - 50 = **2** 50 - 50 = **0** 51 - 50 = **1** 49 - 50 = **-1** 52 - 50 = **2** 50 - 50 = **0**	$(-1)^2$ = **1** $(-3)^2$ = **9** $(2)^2$ = **4** $(0)^2$ = **0** $(1)^2$ = **1** $(-1)^2$ = **1** $(2)^2$ = **4** $(0)^2$ = **0**	1 + 9 + 4 + 0 1 + 1 + 4 + 0 = **20**	20 / DF = 20 / 7 = 2.85714 **Variance = = 2.9%**	$\sqrt{2.9}$ = **1.70294** **Standard Deviation = 1.7%**	$\frac{1.7}{50}$ x 100 = 3.4 **V = 3.4%**

U.S.Department of Transportation
Federal Highway Administration

MODULE A LESSON 6

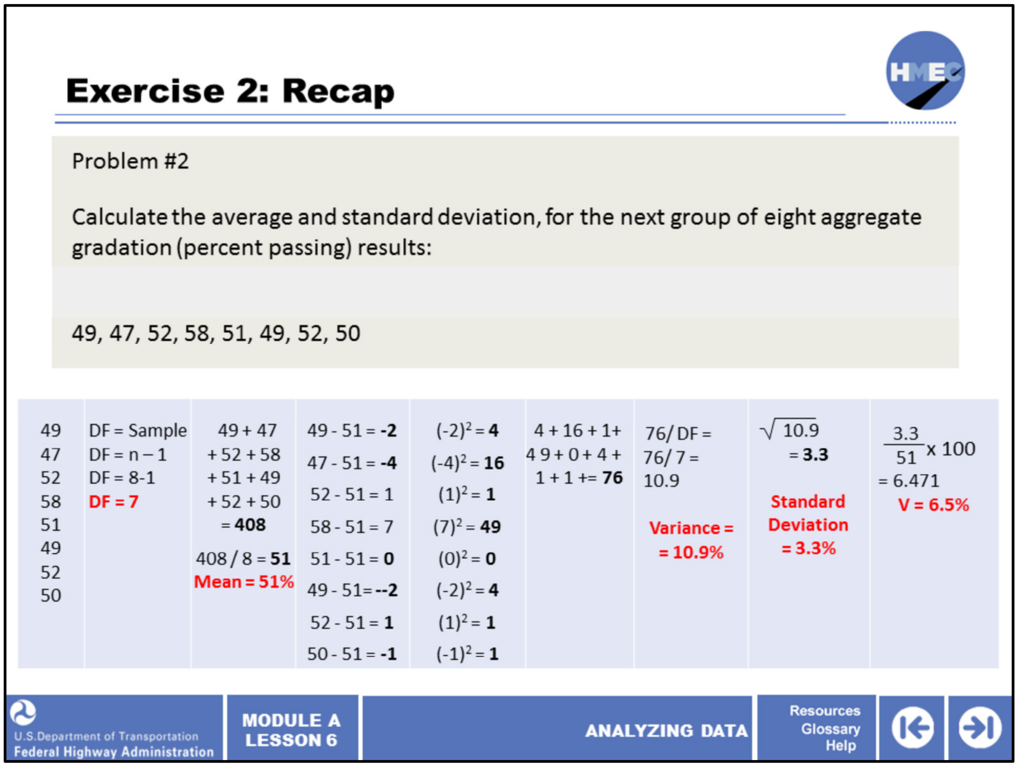
ANALYZING DATA

Resources
Glossary
Help

For Problem #1 you should have:

- Degrees of Freedom = 7;
- Mean = 50%;
- Variance = 2.9%;
- Standard Deviation = 1.7%; and
- Coefficient of Variation = 3.4%.

Image description: Problem #1.



Exercise 2: Recap

Problem #2

Calculate the average and standard deviation, for the next group of eight aggregate gradation (percent passing) results:

49, 47, 52, 58, 51, 49, 52, 50

49 47 52 58 51 49 52 50	DF = Sample DF = n – 1 DF = 8-1 **DF = 7**	49 + 47 + 52 + 58 + 51 + 49 + 52 + 50 = **408** 408 / 8 = **51** **Mean = 51%**	49 - 51 = **-2** 47 - 51 = **-4** 52 - 51 = 1 58 - 51 = 7 51 - 51 = **0** 49 - 51= **--2** 52 - 51 = **1** 50 - 51 = **-1**	$(-2)^2$ = **4** $(-4)^2$ = **16** $(1)^2$ = **1** $(7)^2$ = **49** $(0)^2$ = **0** $(-2)^2$ = **4** $(1)^2$ = **1** $(-1)^2$ = **1**	4 + 16 + 1+ 4 9 + 0 + 4 + 1 + 1 += **76**	76/ DF = 76/ 7 = 10.9 **Variance =** **= 10.9%**	$\sqrt{10.9}$ = **3.3** **Standard Deviation** **= 3.3%**	$\frac{3.3}{51}$ x 100 = 6.471 **V = 6.5%**

U.S.Department of Transportation Federal Highway Administration

MODULE A LESSON 6

ANALYZING DATA

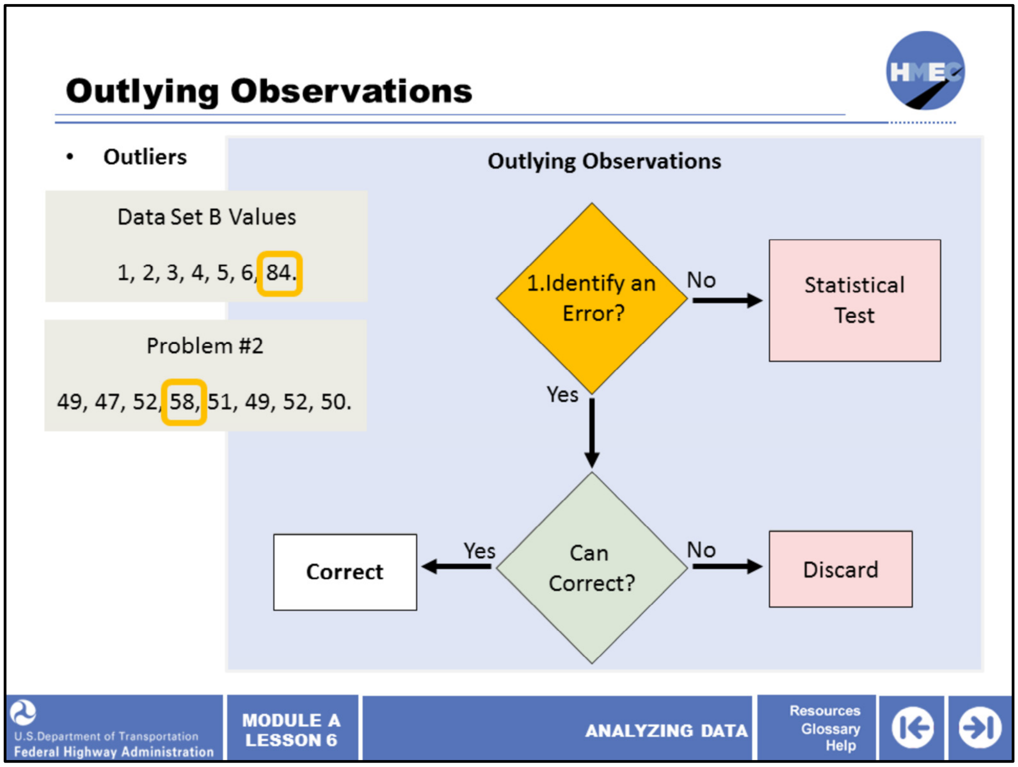
Resources Glossary Help

For Problem #2 you should have:

- Degrees of Freedom = 7;
- Mean = 51%;
- Variance = 10.9%;
- Standard deviation = 3.3%; and
- Coefficient of variation = 6.5%.

Observe that the only difference in the two sets of data is that a 50 in data set 1 has become a 58 in data set 2. This increased the mean by 1 percentage point but essentially doubled the standard deviation. So one data point far from the average can have a large effect on the variability.

Image description: Problem #2.



Outlying Observations

- Outliers

Data Set B Values

1, 2, 3, 4, 5, 6, 84.

Problem #2

49, 47, 52, 58, 51, 49, 52, 50.

Outlying Observations

The question of outlying observations or “outliers” often arises, particularly when it impacts a pay factor. As shown previously, Data Set B has values of 1, 2, 3, 4, 5, 6, and 84. Clearly, the 84 is an outlier. But suppose a value is not that far out.

For example Exercise #2 had values of 49, 47, 52, 58, 51, 49, 52, and 50. Would you consider the 58 to be an outlier? How would you know? And if you think it is, what would you do with this test value? These are important questions and the answer to them is important in specification enforcement. We do not test enough samples to arbitrarily discard data. So we need to have a system for testing for outliers.

There is an ASTM Standard Practice, E 178 *Dealing With Outlying Observations*, that addresses the subject. Several DOTs have adopted or simplified the ASTM procedure to test for outlying observations. Briefly, the procedure is:

- If a mistake or procedural error has been made in obtaining the value, it should be discarded;
- If a calculation or transposition error has been made and can be corrected, the correction should be made.

If it cannot be corrected, the value should be discarded; and

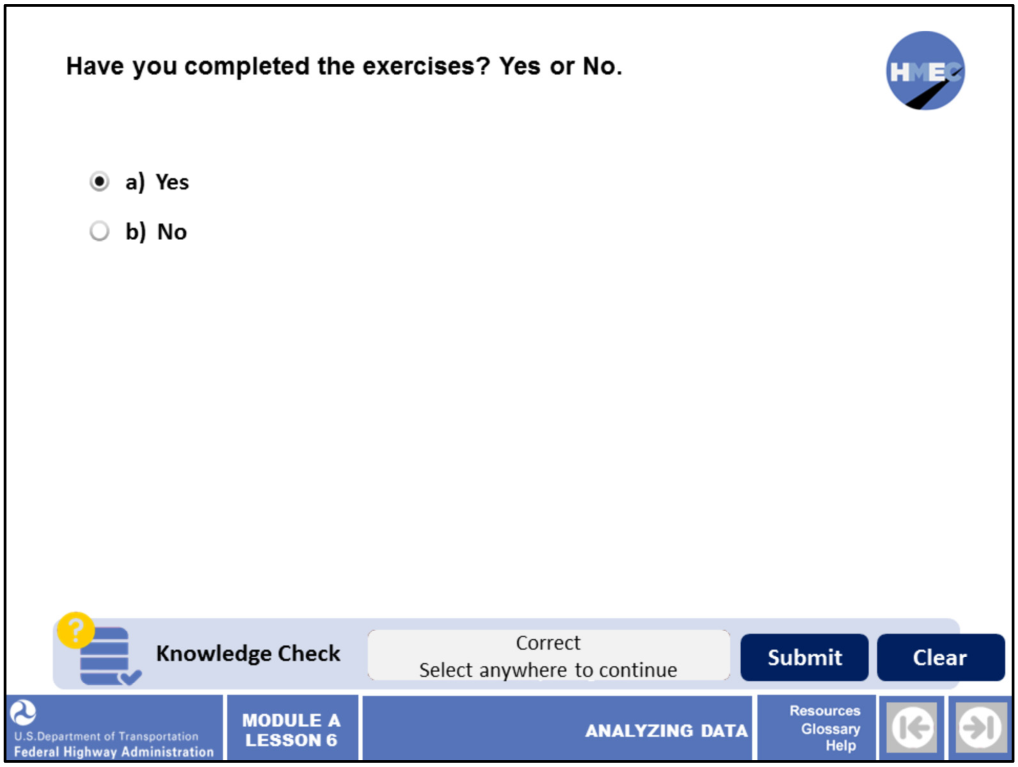
- If no reason for the potential outlier exists, one of several statistical tests should be run to determine the probability that it is an outlier. These will be mentioned in Lesson 7.

The answer to the question about the value of 58 being an outlier is that the probability is that it is not, and that is it is within the normal variability of this set of data points.

Image description: Data set B values.

Image description: Problem #2.

Image description: Outlying observations demonstration.



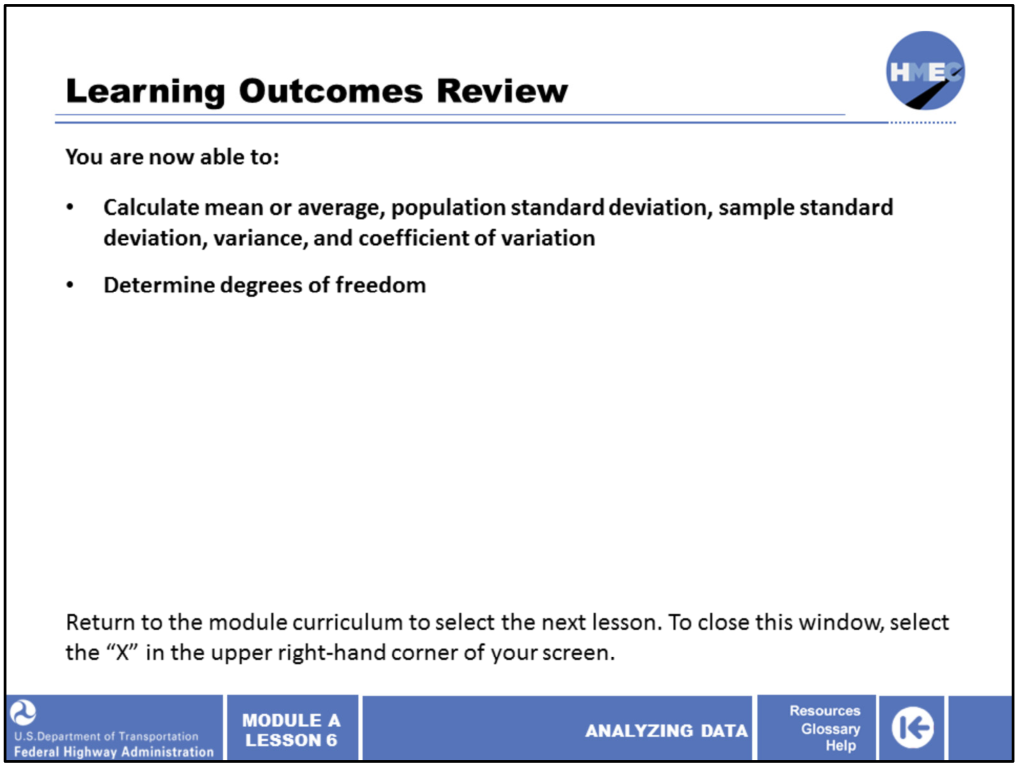
Have you completed the exercises? Yes or No.

● a) Yes

○ b) No

Knowledge Check
Correct
Select anywhere to continue
Submit
Clear
U.S.Department of Transportation
Federal Highway Administration
MODULE A
LESSON 6
ANALYZING DATA
Resources
Glossary
Help
HMEC

Have you completed the exercises? Yes or No.



Learning Outcomes Review

You are now able to:

- Calculate mean or average, population standard deviation, sample standard deviation, variance, and coefficient of variation
- Determine degrees of freedom

Return to the module curriculum to select the next lesson. To close this window, select the “X” in the upper right-hand corner of your screen.

U.S.Department of Transportation
Federal Highway Administration
MODULE A LESSON 6
ANALYZING DATA
Resources
Glossary
Help

You have completed Module A, Lesson 6: Organizing Data. You are now able to:

- Calculate mean or average, population standard deviation, sample standard deviation, variance, and coefficient of variation; and
- Determine degrees of freedom.

Close this lesson, and return to the module curriculum to select the next lesson. To close this window, select the “X” in the upper right-hand corner of your screen.