Module 3: Describing Data

Jon Starkweather, PhD
jonathan.starkweather@unt.edu
Consultant
Research and **S**tatistical **S**upport

UNT UNIVERSITY OF NORTH TEXAS
Discover the power of ideas.

Introduction to Statistics for the Social Sciences

RSS
Research and Statistical Support

The RSS short courses

The Research and Statistical Support (RSS) office at the University of North Texas hosts a number of "Short Courses". A list of them is available at:

`http://www.unt.edu/rss/Instructional.htm`

RSS
Research and Statistical Support

Outline

1. Introduction
   - Context of example data
   - Descriptive Statistics
   - Notation
   - Summation

Outline

1. Introduction
   - Context of example data
   - Descriptive Statistics
   - Notation
   - Summation
2. Classes of Descriptive Statistics
   - Central Tendency
   - Dispersion
   - Shape
   - Relationship

RSS
Research and Statistical Support

Outline

Outline

RSS
Research and Statistical Support

Fictional Extraction Example

Suppose a man, we'll call him Bob Prentice, from England; owns an oil company which operates 1000 deep water drilling rigs...

Fictional Extraction Example

Suppose a man, we'll call him Bob Prentice, from England; owns an oil company which operates 1000 deep water drilling rigs...
Bob wants to maximize his profits (barrels of oil extracted) while minimizing his expenditures (operating costs).

Fictional Extraction Example

Suppose a man, we'll call him Bob Prentice, from England; owns an oil company which operates 1000 deep water drilling rigs...

Bob wants to maximize his profits (barrels of oil extracted) while minimizing his expenditures (operating costs).

So Bob gathers data on how many barrels of oil each of his 1000 rigs extracted in 1 month because, Bob might want to shut down rigs which do not extract much oil, but still cause him to pay the *substantial* operating costs.

Fictional Extraction Example

Suppose a man, we'll call him Bob Prentice, from England; owns an oil company which operates 1000 deep water drilling rigs...
Bob wants to maximize his profits (barrels of oil extracted) while minimizing his expenditures (operating costs).
So Bob gathers data on how many barrels of oil each of his 1000 rigs extracted in 1 month because, Bob might want to shut down rigs which do not extract much oil, but still cause him to pay the *substantial* operating costs.

- The resulting data file is available at the following link: `http://www.unt.edu/rss/class/Jon/ISSS/Module003/M3_BigOilData.txt`

Fictional Extraction Example

Suppose a man, we'll call him Bob Prentice, from England; owns an oil company which operates 1000 deep water drilling rigs...
Bob wants to maximize his profits (barrels of oil extracted) while minimizing his expenditures (operating costs).
So Bob gathers data on how many barrels of oil each of his 1000 rigs extracted in 1 month because, Bob might want to shut down rigs which do not extract much oil, but still cause him to pay the *substantial* operating costs.

- The resulting data file is available at the following link: `http://www.unt.edu/rss/class/Jon/ISSS/Module003/M3_BigOilData.txt`

How will Bob **describe** his data?

RSS
Research and Statistical Support

Describing Data

Bob is interested in describing the extraction and cost of his rigs. But, he does not want to look at all 1000 rigs' data (*population*) and instead decides to draw at random 10 rigs' data (*sample*) which is available on the next slide (Table 1).

RSS
Research and Statistical Support

Describing Data

Bob is interested in describing the extraction and cost of his rigs. But, he does not want to look at all 1000 rigs' data (*population*) and instead decides to draw at random 10 rigs' data (*sample*) which is available on the next slide (Table 1).

- After displaying his data (as covered in Module 2), with a variety of tables and figures (graphs), Bob will likely use **Descriptive Statistics** to describe his data.

RSS
Research and Statistical Support

Describing Data

Bob is interested in describing the extraction and cost of his rigs. But, he does not want to look at all 1000 rigs' data (*population*) and instead decides to draw at random 10 rigs' data (*sample*) which is available on the next slide (Table 1).

- After displaying his data (as covered in Module 2), with a variety of tables and figures (graphs), Bob will likely use **Descriptive Statistics** to describe his data.
- Recall the goals of science and how we achieve them from Module 1:

Describing Data

Bob is interested in describing the extraction and cost of his rigs. But, he does not want to look at all 1000 rigs' data (*population*) and instead decides to draw at random 10 rigs' data (*sample*) which is available on the next slide (Table 1).

- After displaying his data (as covered in Module 2), with a variety of tables and figures (graphs), Bob will likely use **Descriptive Statistics** to describe his data.
- Recall the goals of science and how we achieve them from Module 1:
  - Observation yields **Description**

RSS
Research and Statistical Support

Describing Data

Bob is interested in describing the extraction and cost of his rigs. But, he does not want to look at all 1000 rigs' data (*population*) and instead decides to draw at random 10 rigs' data (*sample*) which is available on the next slide (Table 1).

- After displaying his data (as covered in Module 2), with a variety of tables and figures (graphs), Bob will likely use **Descriptive Statistics** to describe his data.
- Recall the goals of science and how we achieve them from Module 1:
  - Observation yields **Description**
  - Experimentation yields **Explanation**

Describing Data

Bob is interested in describing the extraction and cost of his rigs. But, he does not want to look at all 1000 rigs' data (*population*) and instead decides to draw at random 10 rigs' data (*sample*) which is available on the next slide (Table 1).

- After displaying his data (as covered in Module 2), with a variety of tables and figures (graphs), Bob will likely use **Descriptive Statistics** to describe his data.
- Recall the goals of science and how we achieve them from Module 1:
  - Observation yields **Description**
  - Experimentation yields **Explanation**
  - Modeling yields **Prediction**

Describing Data

Bob is interested in describing the extraction and cost of his rigs. But, he does not want to look at all 1000 rigs' data (*population*) and instead decides to draw at random 10 rigs' data (*sample*) which is available on the next slide (Table 1).

- After displaying his data (as covered in Module 2), with a variety of tables and figures (graphs), Bob will likely use **Descriptive Statistics** to describe his data.
- Recall the goals of science and how we achieve them from Module 1:
  - Observation yields **Description**
  - Experimentation yields **Explanation**
  - Modeling yields **Prediction**
- Descriptive Statistics only allow us to *describe* the data.

RSS
Research and Statistical Support

Data: Extraction & Cost Sample (n = 10)

Table 1: Raw Data for One Month

rig	barrels	costs
065	166	570
142	185	560
198	159	520
277	207	580
408	194	530
533	191	560
621	176	510
788	216	550
796	199	560
915	228	560

“rig” = identification number; “barrels” = 1,000 barrels extracted; “costs” = 1,000 USD

A note about Notation

- Generally, variables are given capital letters, commonly X and Y.

A note about Notation

- Generally, variables are given capital letters, commonly X and Y.
- Subscripts are used to identify each score on a variable.

RSS
Research and Statistical Support

A note about Notation

- Generally, variables are given capital letters, commonly X and Y.
- Subscripts are used to identify each score on a variable.
    - A set of scores $[35, 22, 15, 96, 84, 77]$ on Variable X

RSS
Research and Statistical Support

A note about Notation

- Generally, variables are given capital letters, commonly X and Y.
- Subscripts are used to identify each score on a variable.
  - A set of scores $[35, 22, 15, 96, 84, 77]$ on Variable X
  - So that scores can be 'called' $X_1 = 35$; $X_2 = 22$ ... $X_6 = 77$

A note about Notation

- Generally, variables are given capital letters, commonly X and Y.
- Subscripts are used to identify each score on a variable.
  - A set of scores $[35, 22, 15, 96, 84, 77]$ on Variable X
  - So that scores can be ‘called’ $X_1 = 35$; $X_2 = 22$ ... $X_6 = 77$
- Multiple subscripts can be used to identify a score located at a particular row (i) and column (j).

RSS
Research and Statistical Support

A note about Notation

- Generally, variables are given capital letters, commonly X and Y.
- Subscripts are used to identify each score on a variable.
  - A set of scores $[35, 22, 15, 96, 84, 77]$ on Variable X
  - So that scores can be 'called' $X_1 = 35$; $X_2 = 22$ ... $X_6 = 77$
- Multiple subscripts can be used to identify a score located at a particular row (i) and column (j).
  - So, we can use the convention X_{ij} to identify a particular score, such as $X_{23} = 51$

RSS
Research and Statistical Support

A note about Notation

- Generally, variables are given capital letters, commonly X and Y.
- Subscripts are used to identify each score on a variable.
  - A set of scores $[35, 22, 15, 96, 84, 77]$ on Variable X
  - So that scores can be 'called' $X_1 = 35$; $X_2 = 22$ ... $X_6 = 77$
- Multiple subscripts can be used to identify a score located at a particular row (i) and column (j).
  - So, we can use the convention X_{ij} to identify a particular score, such as $X_{23} = 51$
  - Which indicates that score 51 is located in the 2nd row and in the 3rd column of a table of data.

RSS
Research and Statistical Support

Use and rules of Sigma as Summation

Please note that:

Use and rules of Sigma as Summation

Please note that:

- The $\sum X$ is read as the sum of the values of X

Use and rules of Sigma as Summation

Please note that:

- The $\sum X$ is read as the sum of the values of X
  - Such as $X_1 + X_2 + X_3 + \ldots X_n$
  - Where n is the number of scores of variable X.

RSS
Research and Statistical Support

Use and rules of Sigma as Summation

Please note that:

- The $\sum X$ is read as the sum of the values of X
  - Such as $X_1 + X_2 + X_3 + \ldots X_n$
  - Where n is the number of scores of variable X.
- The $\sum X^2$ is read as the sum of the squared values of X.

Use and rules of Sigma as Summation

Please note that:

- The $\sum X$ is read as the sum of the values of X
  - Such as $X_1 + X_2 + X_3 + \ldots X_n$
  - Where n is the number of scores of variable X.
- The $\sum X^2$ is read as the sum of the squared values of X.
  - Such as $X_1^2 + X_2^2 + X_3^2 + \ldots X_n^2$

RSS
Research and Statistical Support

Use and rules of Sigma as Summation

Please note that:

- The $\sum X$ is read as the sum of the values of X
  - Such as $X_1 + X_2 + X_3 + \ldots X_n$
  - Where n is the number of scores of variable X.
- The $\sum X^2$ is read as the sum of the squared values of X.
  - Such as $X_1^2 + X_2^2 + X_3^2 + \ldots X_n^2$
- **Which is very different from the following.**

RSS
Research and Statistical Support

Use and rules of Sigma as Summation

Please note that:

- The $\sum X$ is read as the sum of the values of X
  - Such as $X_1 + X_2 + X_3 + \ldots X_n$
  - Where n is the number of scores of variable X.
- The $\sum X^2$ is read as the sum of the squared values of X.
  - Such as $X_1^2 + X_2^2 + X_3^2 + \ldots X_n^2$
- **Which is very different from the following.**
- The $(\sum X)^2$ is read as the *square* of the sum of the values of X

RSS
Research and Statistical Support

Use and rules of Sigma as Summation

Please note that:

- The $\sum X$ is read as the sum of the values of X
  - Such as $X_1 + X_2 + X_3 + \ldots X_n$
  - Where n is the number of scores of variable X.
- The $\sum X^2$ is read as the sum of the squared values of X.
  - Such as $X_1^2 + X_2^2 + X_3^2 + \ldots X_n^2$
- **Which is very different from the following.**
- The $(\sum X)^2$ is read as the *square* of the sum of the values of X
  - Such as $(X_1 + X_2 + X_3 + \ldots X_n)^2$

RSS
Research and Statistical Support

The 4 Classes of Descriptive Statistics

There are many ways to describe data because, there are many types of descriptive statistics and many individual descriptive statistics.

The 4 Classes of Descriptive Statistics

There are many ways to describe data because, there are many types of descriptive statistics and many individual descriptive statistics.

- There are four *classes* of descriptive statistics.

The 4 Classes of Descriptive Statistics

There are many ways to describe data because, there are many types of descriptive statistics and many individual descriptive statistics.

- There are four *classes* of descriptive statistics.
  - Central Tendency

The 4 Classes of Descriptive Statistics

There are many ways to describe data because, there are many types of descriptive statistics and many individual descriptive statistics.

- There are four *classes* of descriptive statistics.
  - Central Tendency
  - Dispersion

The 4 Classes of Descriptive Statistics

There are many ways to describe data because, there are many types of descriptive statistics and many individual descriptive statistics.

- There are four *classes* of descriptive statistics.
  - Central Tendency
  - Dispersion
  - Shape

RSS
Research and Statistical Support

The 4 Classes of Descriptive Statistics

There are many ways to describe data because, there are many types of descriptive statistics and many individual descriptive statistics.

- There are four *classes* of descriptive statistics.
  - Central Tendency
  - Dispersion
  - Shape
  - Relationship

The 4 Classes of Descriptive Statistics

There are many ways to describe data because, there are many types of descriptive statistics and many individual descriptive statistics.

- There are four *classes* of descriptive statistics.
  - Central Tendency
  - Dispersion
  - Shape
  - Relationship
- Each class tells Bob something about how his rigs are producing.

RSS
Research and Statistical Support

The 4 Classes of Descriptive Statistics

There are many ways to describe data because, there are many types of descriptive statistics and many individual descriptive statistics.

- There are four *classes* of descriptive statistics.
  - Central Tendency
  - Dispersion
  - Shape
  - Relationship
- Each class tells Bob something about how his rigs are producing.
- Each class of Descriptive Statistics tells us something about how scores are distributed.

RSS
Research and Statistical Support

Central Tendency

- There are 3 *primary* measures of central tendency; each has pros and cons, but all attempt to describe the center point of a distribution of scores.

Central Tendency

- There are 3 *primary* measures of central tendency; each has pros and cons, but all attempt to describe the center point of a distribution of scores.
    - Mode

Central Tendency

- There are 3 *primary* measures of central tendency; each has pros and cons, but all attempt to describe the center point of a distribution of scores.
  - Mode
  - Median

Central Tendency

- There are 3 *primary* measures of central tendency; each has pros and cons, but all attempt to describe the center point of a distribution of scores.
  - Mode
  - Median
  - Mean

Research and Statistical Support

Central Tendency

- There are 3 *primary* measures of central tendency; each has pros and cons, but all attempt to describe the center point of a distribution of scores.
  - Mode
  - Median
  - Mean
- Measures of central tendency offer us a “point” estimate, or single number, which we can use as a summary of a distribution of scores.

Central Tendency

- There are 3 *primary* measures of central tendency; each has pros and cons, but all attempt to describe the center point of a distribution of scores.
  - Mode
  - Median
  - Mean
- Measures of central tendency offer us a “point” estimate, or single number, which we can use as a summary of a distribution of scores.
  - One number which represents or characterizes the entire distribution (as best as one number can).

Central Tendency

- There are 3 *primary* measures of central tendency; each has pros and cons, but all attempt to describe the center point of a distribution of scores.
  - Mode
  - Median
  - Mean
- Measures of central tendency offer us a "point" estimate, or single number, which we can use as a summary of a distribution of scores.
  - One number which represents or characterizes the entire distribution (as best as one number can).
- Keep in mind, the center point of scores in a distribution may not be in the middle of the scale of those scores (more on this later).

The Mode: symbol = *Mo*

The mode is the most frequently occurring score in a distribution.

[1] Multi-modal distributions have multiple scores which occur most frequently; meaning multiple scores occur the same (most frequent) number of times.

The Mode: symbol = *Mo*

The mode is the most frequently occurring score in a distribution.

- Commonly used with categorical variables.

[1] Multi-modal distributions have multiple scores which occur most frequently; meaning multiple scores occur the same (most frequent) number of times.

The Mode: symbol = *Mo*

The mode is the most frequently occurring score in a distribution.

- Commonly used with categorical variables.
- Pro: Easy to compute: simply observe and report the most frequent score(s).

[1] Multi-modal distributions have multiple scores which occur most frequently; meaning multiple scores occur the same (most frequent) number of times.

The Mode: symbol = *Mo*

The mode is the most frequently occurring score in a distribution.

- Commonly used with categorical variables.
- Pro: Easy to compute: simply observe and report the most frequent score(s).
- Pro: Not affected by outliers

[1] Multi-modal distributions have multiple scores which occur most frequently; meaning multiple scores occur the same (most frequent) number of times.

The Mode: symbol = *Mo*

The mode is the most frequently occurring score in a distribution.

- Commonly used with categorical variables.
- Pro: Easy to compute: simply observe and report the most frequent score(s).
- Pro: Not affected by outliers
- Con: Usually only reflects one *actual* score[1].

[1] Multi-modal distributions have multiple scores which occur most frequently; meaning multiple scores occur the same (most frequent) number of times.

The Mode: symbol = *Mo*

The mode is the most frequently occurring score in a distribution.

- Commonly used with categorical variables.
- Pro: Easy to compute: simply observe and report the most frequent score(s).
- Pro: Not affected by outliers
- Con: Usually only reflects one *actual* score[1].
- Con: Different samples almost always produce different modes for the same variable.

[1] Multi-modal distributions have multiple scores which occur most frequently; meaning multiple scores occur the same (most frequent) number of times.

RSS
Research and Statistical Support

Oil Sample (n = 10) Example showing the Mode

rig	barrels	costs
065	166	570
142	185	560
198	159	520
277	207	580
408	194	530
533	191	560
621	176	510
788	216	550
796	199	560
915	228	560

Barrels:
Since no score on this variable occurs more than once, there are multiple modes. Meaning, all of the scores are the Mode.

Costs:
Since this is a very small data set (only 10 cases) and 560 is the only score which occurs more than once, it is very easy to identify it as the Mode.

RSS
Research and Statistical Support

The Median: symbol = *Mdn*

The median is the “middle” score of a distribution.

The Median: symbol = *Mdn*

The median is the "middle" score of a distribution.

- More precisely, it is the point that lies in the middle of a distribution.

The Median: symbol = *Mdn*

The median is the "middle" score of a distribution.

- More precisely, it is the point that lies in the middle of a distribution.
  - Sometimes referred to as the 50th percentile because, there are as many scores above it as there are below it.

The Median: symbol = *Mdn*

The median is the “middle” score of a distribution.

- More precisely, it is the point that lies in the middle of a distribution.
  - Sometimes referred to as the 50th percentile because, there are as many scores above it as there are below it.
- Pro: Not affected by outliers (extreme scores).

RSS
Research and Statistical Support

The Median: symbol = *Mdn*

The median is the “middle” score of a distribution.

- More precisely, it is the point that lies in the middle of a distribution.
  - Sometimes referred to as the 50th percentile because, there are as many scores above it as there are below it.
- Pro: Not affected by outliers (extreme scores).
- Con: Ignores all but the middle of a distribution.

RSS
Research and Statistical Support

The Median: symbol = *Mdn*

The median is the "middle" score of a distribution.

- More precisely, it is the point that lies in the middle of a distribution.
  - Sometimes referred to as the 50th percentile because, there are as many scores above it as there are below it.
- Pro: Not affected by outliers (extreme scores).
- Con: Ignores all but the middle of a distribution.

To calculate the median, first the scores must be arranged in sequential order (e.g., smallest to largest).

The Median: symbol = *Mdn*

The median is the “middle” score of a distribution.

- More precisely, it is the point that lies in the middle of a distribution.
  - Sometimes referred to as the 50th percentile because, there are as many scores above it as there are below it.
- Pro: Not affected by outliers (extreme scores).
- Con: Ignores all but the middle of a distribution.

To calculate the median, first the scores must be arranged in sequential order (e.g., smallest to largest).

- Then;

RSS
Research and Statistical Support

The Median: symbol = *Mdn*

The median is the “middle” score of a distribution.

- More precisely, it is the point that lies in the middle of a distribution.
  - Sometimes referred to as the 50th percentile because, there are as many scores above it as there are below it.
- Pro: Not affected by outliers (extreme scores).
- Con: Ignores all but the middle of a distribution.

To calculate the median, first the scores must be arranged in sequential order (e.g., smallest to largest).

- Then;
  - For an odd number of scores, the middle score is the Median.

RSS
Research and Statistical Support

The Median: symbol = *Mdn*

The median is the "middle" score of a distribution.

- More precisely, it is the point that lies in the middle of a distribution.
  - Sometimes referred to as the 50th percentile because, there are as many scores above it as there are below it.
- Pro: Not affected by outliers (extreme scores).
- Con: Ignores all but the middle of a distribution.

To calculate the median, first the scores must be arranged in sequential order (e.g., smallest to largest).

- Then;
  - For an odd number of scores, the middle score is the Median.
  - For an even number of scores, the average of the two middle scores is the median.

RSS
Research and Statistical Support

Oil Sample (n = 10) Example showing the Median

Table 2: Sorted Sample Data

barrels	costs
159	510
166	520
176	530
185	550
191	560
194	560
199	560
207	560
216	570
228	580

Barrels: $Mdn = 192.5$
Costs: $Mdn = 560$

With an even number of scores, we must take the midpoint of the middle two scores.

Barrels: 191, 194

Costs: 560, 560

RSS Research and Statistical Support

The Mean: sample symbol = $\overline{X}$, population symbol = μ

The mean is the arithmetic average of the scores of a distribution.

RSS
Research and Statistical Support

The Mean: sample symbol = $\overline{X}$, population symbol = μ

The mean is the arithmetic average of the scores of a distribution.

- Mean is the most popular measure of central tendency.

RSS
Research and Statistical Support

The Mean: sample symbol = $\overline{X}$, population symbol = μ

The mean is the arithmetic average of the scores of a distribution.

- Mean is the most popular measure of central tendency.
- Pro: Generally the best measure of central tendency because, it utilizes all the scores.

RSS
Research and Statistical Support

The Mean: sample symbol = $\overline{X}$, population symbol = μ

The mean is the arithmetic average of the scores of a distribution.

- Mean is the most popular measure of central tendency.
- Pro: Generally the best measure of central tendency because, it utilizes all the scores.
- Con: Very sensitive to outliers (extreme scores).

The Mean: sample symbol = $\overline{X}$, population symbol = μ

The mean is the arithmetic average of the scores of a distribution.

- Mean is the most popular measure of central tendency.
- Pro: Generally the best measure of central tendency because, it utilizes all the scores.
- Con: Very sensitive to outliers (extreme scores).

To calculate the **sample** mean, simply sum all of the scores and divide by the number of scores:

RSS
Research and Statistical Support

The Mean: sample symbol = $\overline{X}$, population symbol = μ

The mean is the arithmetic average of the scores of a distribution.

- Mean is the most popular measure of central tendency.
- Pro: Generally the best measure of central tendency because, it utilizes all the scores.
- Con: Very sensitive to outliers (extreme scores).

To calculate the **sample** mean, simply sum all of the scores and divide by the number of scores:

$$\overline{X} = \frac{\sum X}{n}$$

RSS
Research and Statistical Support

Oil Sample (n = 10) Example showing the Mean

General formula for Mean:

[2] Recall, the capital letter we chose to represent a variable is arbitrary, X and Y are commonly used examples.

Oil Sample (n = 10) Example showing the Mean

General formula for Mean:

$$\overline{X} = \frac{\sum X}{n} \qquad \text{or}^2\text{:} \qquad \overline{Y} = \frac{\sum Y}{n}$$

[2] Recall, the capital letter we chose to represent a variable is arbitrary, X and Y are commonly used examples.

Oil Sample (n = 10) Example showing the Mean

General formula for Mean:

$$\overline{X} = \frac{\sum X}{n} \qquad \text{or}^2\text{:} \qquad \overline{Y} = \frac{\sum Y}{n}$$

Barrels: $\overline{X} = \frac{159+166+176+185+191+194+199+207+216+228}{10}$

[2] Recall, the capital letter we chose to represent a variable is arbitrary, X and Y are commonly used examples.

RSS
Research and Statistical Support

Oil Sample (n = 10) Example showing the Mean

General formula for Mean:

$$\overline{X} = \frac{\sum X}{n} \qquad \text{or}^2\text{:} \qquad \overline{Y} = \frac{\sum Y}{n}$$

Barrels: $\overline{X} = \frac{159+166+176+185+191+194+199+207+216+228}{10}$

- Barrels: $\overline{X} = \frac{1921}{10} = 192.1$

[2] Recall, the capital letter we chose to represent a variable is arbitrary, X and Y are commonly used examples.

Oil Sample ($n = 10$) Example showing the Mean

General formula for Mean:

$$\overline{X} = \frac{\sum X}{n} \qquad \text{or}^2: \qquad \overline{Y} = \frac{\sum Y}{n}$$

Barrels: $\overline{X} = \frac{159+166+176+185+191+194+199+207+216+228}{10}$

- Barrels: $\overline{X} = \frac{1921}{10} = 192.1$

Costs: $\overline{Y} = \frac{510+520+530+550+560+560+560+560+570+580}{10}$

[2] Recall, the capital letter we chose to represent a variable is arbitrary, X and Y are commonly used examples.

RSS
Research and Statistical Support

Oil Sample ($n = 10$) Example showing the Mean

General formula for Mean:

$$\overline{X} = \frac{\sum X}{n} \qquad \text{or}^2: \qquad \overline{Y} = \frac{\sum Y}{n}$$

Barrels: $\overline{X} = \frac{159+166+176+185+191+194+199+207+216+228}{10}$

- Barrels: $\overline{X} = \frac{1921}{10} = 192.1$

Costs: $\overline{Y} = \frac{510+520+530+550+560+560+560+560+570+580}{10}$

- Costs: $\overline{Y} = \frac{5500}{10} = 550$

[2] Recall, the capital letter we chose to represent a variable is arbitrary, X and Y are commonly used examples.

RSS Research and Statistical Support

Trimmed Mean & M-estimators

Because mean is very sensitive to outliers, alternatives have been proposed which attempt to correct for this problem.

Trimmed Mean & M-estimators

Because mean is very sensitive to outliers, alternatives have been proposed which attempt to correct for this problem.

- Trimmed mean simply refers to a mean calculated after "trimming" a certain percentage of extreme scores.

Trimmed Mean & M-estimators

Because mean is very sensitive to outliers, alternatives have been proposed which attempt to correct for this problem.

- Trimmed mean simply refers to a mean calculated after "trimming" a certain percentage of extreme scores.
  - The median is an extreme example of a trimmed mean; the median trims all but the middle score or middle two scores.

Trimmed Mean & M-estimators

Because mean is very sensitive to outliers, alternatives have been proposed which attempt to correct for this problem.

- Trimmed mean simply refers to a mean calculated after "trimming" a certain percentage of extreme scores.
  - The median is an extreme example of a trimmed mean; the median trims all but the middle score or middle two scores.
  - Common examples are 10% and 20% trimmed means; where the 10 or 20% of the most extreme scores (high & low) are trimmed.

Trimmed Mean & M-estimators

Because mean is very sensitive to outliers, alternatives have been proposed which attempt to correct for this problem.

- Trimmed mean simply refers to a mean calculated after "trimming" a certain percentage of extreme scores.
    - The median is an extreme example of a trimmed mean; the median trims all but the middle score or middle two scores.
    - Common examples are 10% and 20% trimmed means; where the 10 or 20% of the most extreme scores (high & low) are trimmed.
- M-estimators are weighted means; meaning scores near the middle are given more weight and scores at the extremes are given less weight.

Dispersion

Measures of dispersion offer us an idea of how spread out the scores are, or how wide is the distribution of scores.

Dispersion

Measures of dispersion offer us an idea of how spread out the scores are, or how wide is the distribution of scores.

- There are 5 primary measures of dispersion; 3 of which will be used repeatedly during the rest of this course.

Dispersion

Measures of dispersion offer us an idea of how spread out the scores are, or how wide is the distribution of scores.

- There are 5 primary measures of dispersion; 3 of which will be used repeatedly during the rest of this course.
  - Range

RSS
Research and Statistical Support

Dispersion

Measures of dispersion offer us an idea of how spread out the scores are, or how wide is the distribution of scores.

- There are 5 primary measures of dispersion; 3 of which will be used repeatedly during the rest of this course.
  - Range
  - Sums of Squares

RSS
Research and Statistical Support

Dispersion

Measures of dispersion offer us an idea of how spread out the scores are, or how wide is the distribution of scores.

- There are 5 primary measures of dispersion; 3 of which will be used repeatedly during the rest of this course.
  - Range
  - Sums of Squares
  - Variance

RSS
Research and Statistical Support

Dispersion

Measures of dispersion offer us an idea of how spread out the scores are, or how wide is the distribution of scores.

- There are 5 primary measures of dispersion; 3 of which will be used repeatedly during the rest of this course.
  - Range
  - Sums of Squares
  - Variance
  - Standard Deviation

RSS
Research and Statistical Support

Dispersion

Measures of dispersion offer us an idea of how spread out the scores are, or how wide is the distribution of scores.

- There are 5 primary measures of dispersion; 3 of which will be used repeatedly during the rest of this course.
  - Range
  - Sums of Squares
  - Variance
  - Standard Deviation
  - Coefficient of Variation

Dispersion

Measures of dispersion offer us an idea of how spread out the scores are, or how wide is the distribution of scores.

- There are 5 primary measures of dispersion; 3 of which will be used repeatedly during the rest of this course.
  - Range
  - Sums of Squares
  - Variance
  - Standard Deviation
  - Coefficient of Variation
- All measures of dispersion must **not** be zero.

RSS
Research and Statistical Support

Dispersion

Measures of dispersion offer us an idea of how spread out the scores are, or how wide is the distribution of scores.

- There are 5 primary measures of dispersion; 3 of which will be used repeatedly during the rest of this course.
  - Range
  - Sums of Squares
  - Variance
  - Standard Deviation
  - Coefficient of Variation
- All measures of dispersion must **not** be zero.
  - If a measure of dispersion is zero, then you do not have a *variable*, you have a *constant*.

RSS
Research and Statistical Support

Dispersion

Measures of dispersion offer us an idea of how spread out the scores are, or how wide is the distribution of scores.

- There are 5 primary measures of dispersion; 3 of which will be used repeatedly during the rest of this course.
  - Range
  - Sums of Squares
  - Variance
  - Standard Deviation
  - Coefficient of Variation
- All measures of dispersion must **not** be zero.
  - If a measure of dispersion is zero, then you do not have a *variable*, you have a *constant*.
  - If our scores are: (5, 5, 5, 5, 5) then dispersion is zero and this is a constant.

The Range

The range is simply the maximum score, minus the minimum score.

The Range

The range is simply the maximum score, minus the minimum score. Examples from our oil data:

The Range

The range is simply the maximum score, minus the minimum score. Examples from our oil data:

- Barrels: $228 - 159 = 69$

The Range

The range is simply the maximum score, minus the minimum score. Examples from our oil data:

- Barrels: $228 - 159 = 69$
- Costs: $580 - 510 = 70$

The Range

The range is simply the maximum score, minus the minimum score. Examples from our oil data:

- Barrels: $228 - 159 = 69$
- Costs: $580 - 510 = 70$

Disadvantages:

The Range

The range is simply the maximum score, minus the minimum score. Examples from our oil data:

- Barrels: $228 - 159 = 69$
- Costs: $580 - 510 = 70$

Disadvantages:

- It is calculated from only 2 scores.

The Range

The range is simply the maximum score, minus the minimum score. Examples from our oil data:

- Barrels: $228 - 159 = 69$
- Costs: $580 - 510 = 70$

Disadvantages:

- It is calculated from only 2 scores.
- Those two values are the most extreme in the distribution (obviously sensitive to outliers).

The Range

The range is simply the maximum score, minus the minimum score. Examples from our oil data:

- Barrels: $228 - 159 = 69$
- Costs: $580 - 510 = 70$

Disadvantages:

- It is calculated from only 2 scores.
- Those two values are the most extreme in the distribution (obviously sensitive to outliers).
- The range can change dramatically from sample to sample (of the same variable).

RSS
Research and Statistical Support

The Range

The range is simply the maximum score, minus the minimum score. Examples from our oil data:

- Barrels: $228 - 159 = 69$
- Costs: $580 - 510 = 70$

Disadvantages:

- It is calculated from only 2 scores.
- Those two values are the most extreme in the distribution (obviously sensitive to outliers).
- The range can change dramatically from sample to sample (of the same variable).
- The range is not terribly informative.

RSS
Research and Statistical Support

Sums of Squares: symbol = *SoS*

The Sums of Squares are the sum of the squared deviations from the mean for a distribution of scores.

Sums of Squares: symbol = *SoS*

The Sums of Squares are the sum of the squared deviations from the mean for a distribution of scores.

- Though not informative or used as a measure of dispersion, it is **very** frequently used in the calculation of other statistics.

RSS
Research and Statistical Support

Sums of Squares: symbol = *SoS*

The Sums of Squares are the sum of the squared deviations from the mean for a distribution of scores.

- Though not informative or used as a measure of dispersion, it is **very** frequently used in the calculation of other statistics.

The general formula for calculating a variable's *SoS* is:

RSS
Research and Statistical Support

Sums of Squares: symbol = *SoS*

The Sums of Squares are the sum of the squared deviations from the mean for a distribution of scores.

- Though not informative or used as a measure of dispersion, it is **very** frequently used in the calculation of other statistics.

The general formula for calculating a variable's *SoS* is:

$$SoS = \sum \left(X - \overline{X}\right)^2$$

RSS
Research and Statistical Support

Variance: sample symbol = S^2, population symbol = σ^2

The variance is the average of each score's squared difference from the mean.

Variance: sample symbol = S^2, population symbol = σ^2

The variance is the average of each score's squared difference from the mean.

- The general formula for calculating a variable's **sample** variance is:

Research and Statistical Support

Variance: sample symbol = S^2, population symbol = σ^2

The variance is the average of each score's squared difference from the mean.

- The general formula for calculating a variable's **sample** variance is:

$$S^2 = \frac{\sum(X - \overline{X})^2}{n-1} \qquad \text{or} \qquad S^2 = \frac{SoS}{n-1}$$

Variance: sample symbol = S^2, population symbol = σ^2

The variance is the average of each score's squared difference from the mean.

- The general formula for calculating a variable's **sample** variance is:

$$S^2 = \frac{\sum (X - \overline{X})^2}{n-1} \qquad \text{or} \qquad S^2 = \frac{SoS}{n-1}$$

- The formula for calculating a variable's **population** variance is:

Variance: sample symbol = S^2, population symbol = σ^2

The variance is the average of each score's squared difference from the mean.

- The general formula for calculating a variable's **sample** variance is:

$$S^2 = \frac{\sum(X - \overline{X})^2}{n-1} \qquad \text{or} \qquad S^2 = \frac{SoS}{n-1}$$

- The formula for calculating a variable's **population** variance is:

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N}$$

RSS
Research and Statistical Support

Variance: sample symbol = S^2, population symbol = σ^2

The variance is the average of each score's squared difference from the mean.

- The general formula for calculating a variable's **sample** variance is:

$$S^2 = \frac{\sum (X - \overline{X})^2}{n-1} \qquad \text{or} \qquad S^2 = \frac{SoS}{n-1}$$

- The formula for calculating a variable's **population** variance is:

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

- Note: with a sample, we divide by $n - 1$; if we divided by n, our variance statistic would be less representative of the variance parameter (i.e., the sample value would be systematically smaller than the population value).

RSS Research and Statistical Support

Variance: sample symbol = S^2, population symbol = σ^2

The variance is the average of each score's squared difference from the mean.

- The general formula for calculating a variable's **sample** variance is:

$$S^2 = \frac{\sum (X - \overline{X})^2}{n-1} \qquad \text{or} \qquad S^2 = \frac{SoS}{n-1}$$

- The formula for calculating a variable's **population** variance is:

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

- Note: with a sample, we divide by $n - 1$; if we divided by n, our variance statistic would be less representative of the variance parameter (i.e., the sample value would be systematically smaller than the population value).
- Also note: when referring to total number of scores in a *population* we use N, in a *sample* we use n.

Standard Deviation: sample symbol = S, population symbol = σ

The Standard Deviation is the square root of the variance and allows us to compare the dispersion of one distribution to another.

[3] The American Psychological Association (APA) Publication Manual requires that mean and standard deviation be reported whenever one is referring to a group of scores.

RSS
Research and Statistical Support

Standard Deviation: sample symbol = S, population symbol = σ

The Standard Deviation is the square root of the variance and allows us to compare the dispersion of one distribution to another.

- It is the most commonly reported measure of dispersion[3].

[3] The American Psychological Association (APA) Publication Manual requires that mean and standard deviation be reported whenever one is referring to a group of scores.

Standard Deviation: sample symbol = S, population symbol = σ

The Standard Deviation is the square root of the variance and allows us to compare the dispersion of one distribution to another.

- It is the most commonly reported measure of dispersion[3].
- It is very easy to calculate...just take the square root of the variance.

[3] The American Psychological Association (APA) Publication Manual requires that mean and standard deviation be reported whenever one is referring to a group of scores.

RSS
Research and Statistical Support

Standard Deviation: sample symbol = S, population symbol = σ

The Standard Deviation is the square root of the variance and allows us to compare the dispersion of one distribution to another.

- It is the most commonly reported measure of dispersion[3].
- It is very easy to calculate...just take the square root of the variance.
    - Sample Formula: $S = \sqrt{S^2}$
    - Population Formula: $\sigma = \sqrt{\sigma^2}$

[3] The American Psychological Association (APA) Publication Manual requires that mean and standard deviation be reported whenever one is referring to a group of scores.

RSS
Research and Statistical Support

Standard Deviation: sample symbol = S, population symbol = σ

The Standard Deviation is the square root of the variance and allows us to compare the dispersion of one distribution to another.

- It is the most commonly reported measure of dispersion[3].
- It is very easy to calculate...just take the square root of the variance.
  - Sample Formula: $S = \sqrt{S^2}$
  - Population Formula: $\sigma = \sqrt{\sigma^2}$
- Note the use of the word “Standard” which you will see often; it refers to standardization, which tends to allow us to compare statistics from different variables or distributions (i.e., apples & oranges).

[3] The American Psychological Association (APA) Publication Manual requires that mean and standard deviation be reported whenever one is referring to a group of scores.

RSS
Research and Statistical Support

Formula *Smormula*: Computational formulas vs. Definitional formulas

Computational: designed to make computing by hand easier.

Formula *Smormula*: Computational formulas vs. Definitional formulas

Computational: designed to make computing by hand easier.

- A matter of opinion these days...

Research and Statistical Support

Formula *Smormula*: Computational formulas vs. Definitional formulas

Computational: designed to make computing by hand easier.

- A matter of opinion these days...

Definitional: designed to make understanding the concept easier, formula follows the definition of the concepts.

Formula *Smormula*: Computational formulas vs. Definitional formulas

Computational: designed to make computing by hand easier.

- A matter of opinion these days...

Definitional: designed to make understanding the concept easier, formula follows the definition of the concepts.

- Here are both for the **standard deviation** of a **sample**.

Formula *Smormula*: Computational formulas vs. Definitional formulas

Computational: designed to make computing by hand easier.

- A matter of opinion these days...

Definitional: designed to make understanding the concept easier, formula follows the definition of the concepts.

- Here are both for the **standard deviation** of a **sample**.

Definitional

$$S = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}}$$

Computational

$$S = \sqrt{\frac{\sum X^2 - [(\sum X)^2 / n]}{n-1}}$$

RSS Research and Statistical Support

Formula *Smormula*: Computational formulas vs. Definitional formulas

Computational: designed to make computing by hand easier.

- A matter of opinion these days...

Definitional: designed to make understanding the concept easier, formula follows the definition of the concepts.

- Here are both for the **standard deviation** of a **sample.**

Definitional

$$S = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}}$$

Computational

$$S = \sqrt{\frac{\sum X^2 - [(\sum X)^2 / n]}{n-1}}$$

- Either can be used; both types provide the same answer.

Calculating *SoS* using example data (X = barrels)

X_i	X	$\overline{X}$	$\left(X - \overline{X}\right)$	$\left(X - \overline{X}\right)^2$
1	159	192.1	-33.1	1095.61
2	166	192.1	-26.1	681.21
3	176	192.1	-16.1	259.21
4	185	192.1	-7.1	50.41
5	191	192.1	-1.1	1.21
6	194	192.1	1.9	3.61
7	199	192.1	6.9	47.61
8	207	192.1	14.9	222.01
9	216	192.1	23.9	571.21
10	228	192.1	35.9	1288.81
	$\sum X = 1921$	$SoS =$	$\sum \left(X - \overline{X}\right)^2$	$= 4220.90$

Sample mean $= \overline{X} = \sum X / n = 1921/10 = 192.1$

Calculating variance & standard deviation using example data (X = barrels)

Taking the information from the last slide...

Calculating variance & standard deviation using example data (X = barrels)

Taking the information from the last slide...

- Sample Variance for 'Barrels' is:

$$S^2 = \frac{\sum(X - \overline{X})^2}{n-1} = \frac{SoS}{n-1} = \frac{4220.90}{10-1} = \frac{4220.90}{9} = 468.99$$

Research and Statistical Support

Calculating variance & standard deviation using example data (X = barrels)

Taking the information from the last slide...

- Sample Variance for 'Barrels' is:

$$S^2 = \frac{\sum (X - \overline{X})^2}{n-1} = \frac{SoS}{n-1} = \frac{4220.90}{10-1} = \frac{4220.90}{9} = 468.99$$

- Sample Standard Deviation for 'Barrels' is:

$$S = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}} = \sqrt{\frac{SoS}{n-1}} = \sqrt{S^2} = \sqrt{468.99} = 21.66$$

RSS Research and Statistical Support

Coefficient of Variation

The Coefficient of Variation (CV) is calculated by dividing the standard deviation by the mean, then multiply the result times 100 to express it as a percentage.

Coefficient of Variation

The Coefficient of Variation (CV) is calculated by dividing the standard deviation by the mean, then multiply the result times 100 to express it as a percentage.
The CV allows us to compare the standard deviation of one distribution to another.

Coefficient of Variation

The Coefficient of Variation (CV) is calculated by dividing the standard deviation by the mean, then multiply the result times 100 to express it as a percentage.
The CV allows us to compare the standard deviation of one distribution to another.

$$CV = \frac{S}{\overline{X}} \times 100 = \frac{21.66}{192.1} \times 100 = 0.1128 \times 100 = 11.28$$

RSS
Research and Statistical Support

Coefficient of Variation

The Coefficient of Variation (CV) is calculated by dividing the standard deviation by the mean, then multiply the result times 100 to express it as a percentage.
The CV allows us to compare the standard deviation of one distribution to another.

$$CV = \frac{S}{\overline{X}} \times 100 = \frac{21.66}{192.1} \times 100 = 0.1128 \times 100 = 11.28$$

The CV for 'Barrels' tells us that the standard deviation is 11.28% of the mean.

RSS
Research and Statistical Support

Coefficient of Variation

The Coefficient of Variation (CV) is calculated by dividing the standard deviation by the mean, then multiply the result times 100 to express it as a percentage.
The CV allows us to compare the standard deviation of one distribution to another.

$$CV = \frac{S}{\overline{X}} \times 100 = \frac{21.66}{192.1} \times 100 = 0.1128 \times 100 = 11.28$$

The CV for 'Barrels' tells us that the standard deviation is 11.28% of the mean.
In contrast, the CV of 'Costs' was 4.11% of the mean; the mean was 550.

RSS
Research and Statistical Support

Coefficient of Variation

The Coefficient of Variation (CV) is calculated by dividing the standard deviation by the mean, then multiply the result times 100 to express it as a percentage.
The CV allows us to compare the standard deviation of one distribution to another.

$$CV = \frac{S}{\bar{X}} \times 100 = \frac{21.66}{192.1} \times 100 = 0.1128 \times 100 = 11.28$$

The CV for 'Barrels' tells us that the standard deviation is 11.28% of the mean.
In contrast, the CV of 'Costs' was 4.11% of the mean; the mean was 550.

- You should be able to work backwards from the information in the lines directly above to get the standard deviation, variance, & sums of squares for 'Costs'.

Shape

Measures of shape offer us an idea of what the distribution of scores looks like when plotted.

Shape

Measures of shape offer us an idea of what the distribution of scores looks like when plotted.

- Unimodal: one peak

RSS
Research and Statistical Support

Shape

Measures of shape offer us an idea of what the distribution of scores looks like when plotted.

- Unimodal: one peak
- Bimodal: two peaks

RSS
Research and Statistical Support

Shape

Measures of shape offer us an idea of what the distribution of scores looks like when plotted.

- Unimodal: one peak
- Bimodal: two peaks
- Multimodal: multiple peaks

Shape

Measures of shape offer us an idea of what the distribution of scores looks like when plotted.

- Unimodal: one peak
- Bimodal: two peaks
- Multimodal: multiple peaks
- Rectangular distributions: multiple peaks of the same magnitude

RSS
Research and Statistical Support

Shape

Measures of shape offer us an idea of what the distribution of scores looks like when plotted.

- Unimodal: one peak
- Bimodal: two peaks
- Multimodal: multiple peaks
- Rectangular distributions: multiple peaks of the same magnitude

There are two measures of shape we commonly use:

Research and Statistical Support

Shape

Measures of shape offer us an idea of what the distribution of scores looks like when plotted.

- Unimodal: one peak
- Bimodal: two peaks
- Multimodal: multiple peaks
- Rectangular distributions: multiple peaks of the same magnitude

There are two measures of shape we commonly use:

- Skewness (or simply Skew)

RSS
Research and Statistical Support

Shape

Measures of shape offer us an idea of what the distribution of scores looks like when plotted.

- Unimodal: one peak
- Bimodal: two peaks
- Multimodal: multiple peaks
- Rectangular distributions: multiple peaks of the same magnitude

There are two measures of shape we commonly use:

- Skewness (or simply Skew)
- Kurtosis

Research and Statistical Support

Skewness

The Skewness refers to the amount of non-symmetry a distribution of scores contains.

Skewness

The Skewness refers to the amount of non-symmetry a distribution of scores contains.

- **Negative skew** is when the tail points to the smaller values and most scores are located at the larger values.

RSS
Research and Statistical Support

Skewness

The Skewness refers to the amount of non-symmetry a distribution of scores contains.

- **Negative skew** is when the tail points to the smaller values and most scores are located at the larger values.
- **Positive skew** is when the tail points to the larger values and most scores are located at the smaller values.

RSS
Research and Statistical Support

Skewness

The Skewness refers to the amount of non-symmetry a distribution of scores contains.

- **Negative skew** is when the tail points to the smaller values and most scores are located at the larger values.
- **Positive skew** is when the tail points to the larger values and most scores are located at the smaller values.
- **Zero skew** indicates symmetry.

Skewness

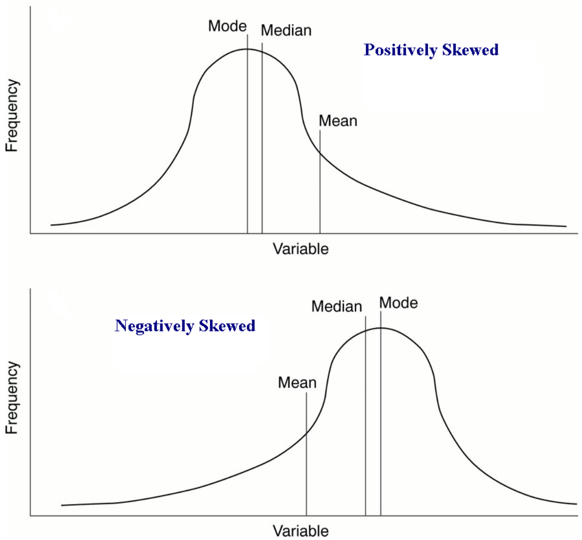
The Skewness refers to the amount of non-symmetry a distribution of scores contains.

- **Negative skew** is when the tail points to the smaller values and most scores are located at the larger values.
- **Positive skew** is when the tail points to the larger values and most scores are located at the smaller values.
- **Zero skew** indicates symmetry.

The farther from zero the skewness, the less symmetric the distribution of scores.

RSS
Research and Statistical Support

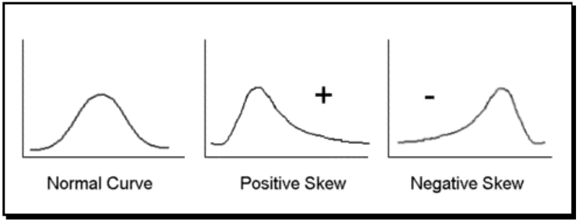
Recognizing Skewness



RSS
Research and Statistical Support

Recognizing Skewness

Zero Skewness Positive Skewness Negative Skewness



Kurtosis

The Kurtosis measures the amount of tail magnitude, commonly referred to as "peak-ness" or "flatness" of a distribution of scores.

Kurtosis

The Kurtosis measures the amount of tail magnitude, commonly referred to as “peak-ness” or “flatness” of a distribution of scores.

- Kurtosis is based on the size of a distribution’s tails.

Kurtosis

The Kurtosis measures the amount of tail magnitude, commonly referred to as “peak-ness” or “flatness” of a distribution of scores.

- Kurtosis is based on the size of a distribution’s tails.
  - A distribution with a large, positive kurtosis has thin tails and the distribution *looks* peaked.

RSS
Research and Statistical Support

Kurtosis

The Kurtosis measures the amount of tail magnitude, commonly referred to as “peak-ness” or “flatness” of a distribution of scores.

- Kurtosis is based on the size of a distribution’s tails.
  - A distribution with a large, positive kurtosis has thin tails and the distribution *looks* peaked.
    - This is known as Leptokurtic.

Kurtosis

The Kurtosis measures the amount of tail magnitude, commonly referred to as “peak-ness” or “flatness” of a distribution of scores.

- Kurtosis is based on the size of a distribution’s tails.
  - A distribution with a large, positive kurtosis has thin tails and the distribution *looks* peaked.
    - This is known as Leptokurtic.
  - A distribution with a large, negative kurtosis has thick tails and the distribution *looks* flat.

RSS
Research and Statistical Support

Kurtosis

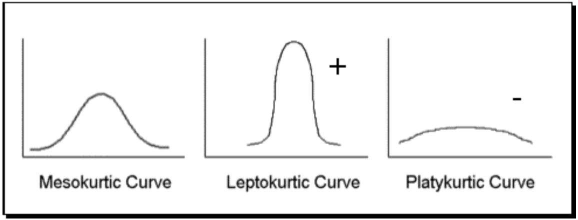
The Kurtosis measures the amount of tail magnitude, commonly referred to as “peak-ness” or “flatness” of a distribution of scores.

- Kurtosis is based on the size of a distribution’s tails.
  - A distribution with a large, positive kurtosis has thin tails and the distribution *looks* peaked.
    - This is known as Leptokurtic.
  - A distribution with a large, negative kurtosis has thick tails and the distribution *looks* flat.
    - This is known as Platykurtic (like a plateau).

RSS
Research and Statistical Support

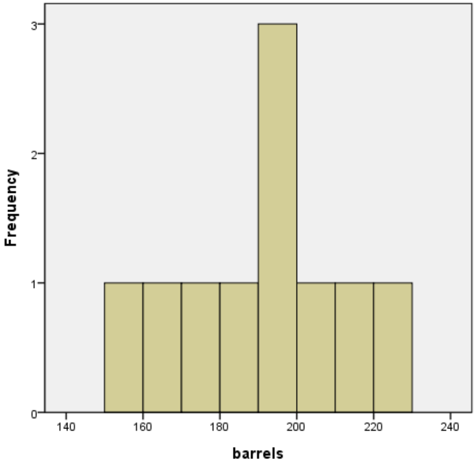
Recognizing Kurtosis

Zero Kurtosis Positive Kurtosis Negative Kurtosis



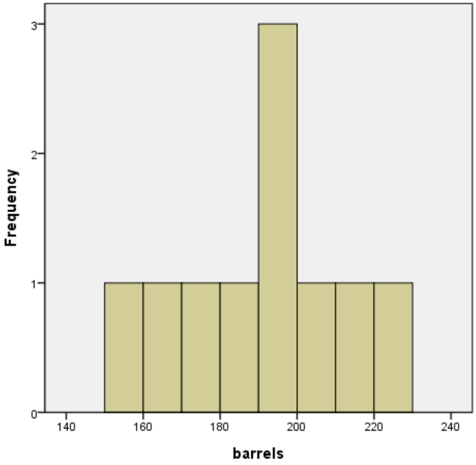
Describing a Distribution’s Shape

How can we describe the shape of the distribution of our sample’s Barrels variable?



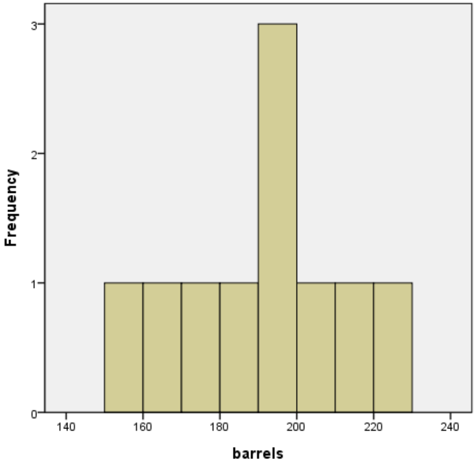
Describing a Distribution’s Shape

How can we describe the shape of the distribution of our sample’s Barrels variable? Unimodal.



Describing a Distribution's Shape

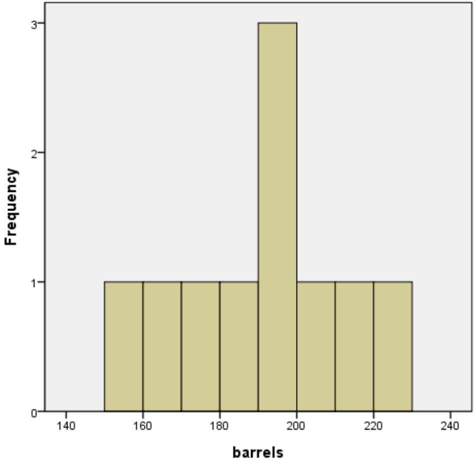
How can we describe the shape of the distribution of our sample's Barrels variable? Unimodal. Slightly Negatively Skewed?



RSS
Research and Statistical Support

Describing a Distribution's Shape

How can we describe the shape of the distribution of our sample's Barrels variable? Unimodal. Slightly Negatively Skewed? Leptokurtic (positive kurtosis)?



Deceptive Sample Leads to Poor Judgment

- When we have such a small sample size (n = 10), we must be careful when eyeballing the distribution.

RSS
Research and Statistical Support

Deceptive Sample Leads to Poor Judgment

- When we have such a small sample size ($n = 10$), we must be careful when eyeballing the distribution.
    - Actually, the Skewness for Barrels is .068

Deceptive Sample Leads to Poor Judgment

- When we have such a small sample size (n = 10), we must be careful when eyeballing the distribution.
  - Actually, the Skewness for Barrels is .068
  - And, the Kurtosis for Barrels is -.595

RSS
Research and Statistical Support

Deceptive Sample Leads to Poor Judgment

- When we have such a small sample size (n = 10), we must be careful when eyeballing the distribution.
  - Actually, the Skewness for Barrels is .068
  - And, the Kurtosis for Barrels is -.595
- Generally, in the social sciences; we expect variables to have skewness and kurtosis between +1 and -1.

RSS
Research and Statistical Support

Deceptive Sample Leads to Poor Judgment

- When we have such a small sample size (n = 10), we must be careful when eyeballing the distribution.
  - Actually, the Skewness for Barrels is .068
  - And, the Kurtosis for Barrels is -.595
- Generally, in the social sciences; we expect variables to have skewness and kurtosis between +1 and -1.
- When a variable displays a skewness or kurtosis larger than +1 or -1, then we say the variable is not symmetrical and/or does not have well proportioned tails.

RSS Research and Statistical Support

Relationship

- Measures of relationship offer us an idea of how two sets of scores, or two variables, are related.

Relationship

- Measures of relationship offer us an idea of how two sets of scores, or two variables, are related.
- How much variance do the two variables share.

Relationship

- Measures of relationship offer us an idea of how two sets of scores, or two variables, are related.
- How much variance do the two variables share.
- How much does one variable *overlap* another.

RSS
Research and Statistical Support

Relationship

- Measures of relationship offer us an idea of how two sets of scores, or two variables, are related.
- How much variance do the two variables share.
- How much does one variable *overlap* another.
  - Measures of Association: When at least one of the two variables is ordinal or nominal in scale.

Relationship

- Measures of relationship offer us an idea of how two sets of scores, or two variables, are related.
- How much variance do the two variables share.
- How much does one variable *overlap* another.
  - Measures of Association: When at least one of the two variables is ordinal or nominal in scale.
  - Correlational Measures: When both variables are interval or ratio scaled (continuous or nearly so).

Relationship

- Measures of relationship offer us an idea of how two sets of scores, or two variables, are related.
- How much variance do the two variables share.
- How much does one variable *overlap* another.
  - Measures of Association: When at least one of the two variables is ordinal or nominal in scale.
  - Correlational Measures: When both variables are interval or ratio scaled (continuous or nearly so).
- For the time being, we will do a quick overview of the Measures of Association and focus more attention on the Correlational Measures.

Measures of Association

There are several Measures of Association.

Measures of Association

There are several Measures of Association.

- Point-Biserial Correlation (r_{pb}) when one variable is dichotomous.

RSS
Research and Statistical Support

Measures of Association

There are several Measures of Association.

- Point-Biserial Correlation (r_{pb}) when one variable is dichotomous.
- Phi Coefficient (ϕ) when both variables are dichotomous.

Measures of Association

There are several Measures of Association.

- Point-Biserial Correlation (r_{pb}) when one variable is dichotomous.
- Phi Coefficient (ϕ) when both variables are dichotomous.
- Spearman's rho (ρ) or (r_s) and Kendall's tau (τ) when one or both variables are ranked (ordinal).

RSS
Research and Statistical Support

Correlational Measures

There are three key Correlational Measures we will cover here.

Correlational Measures

There are three key Correlational Measures we will cover here.

- Covariance (*COV*)
  - COV_{XY} where x and y are the two variables we are using.

Correlational Measures

There are three key Correlational Measures we will cover here.

- Covariance (*COV*)
  - COV_{XY} where x and y are the two variables we are using.
- Correlation; the Pearson Product-Moment Correlation Coefficient (r)

Correlational Measures

There are three key Correlational Measures we will cover here.

- Covariance (*COV*)
  - COV_{XY} where x and y are the two variables we are using.
- Correlation; the Pearson Product-Moment Correlation Coefficient (r)
- Adjusted Correlation Coefficient (r_{adj})

RSS
Research and Statistical Support

Covariance

The Covariance is a non-standardized measure of relationship; meaning it can not be used to compare the relationship of two variables to the relationship of two other variables.

Covariance

The Covariance is a non-standardized measure of relationship; meaning it can not be used to compare the relationship of two variables to the relationship of two other variables.

- Covariance is not terribly meaningful by itself, but it is used in calculating other statistics (e.g., correlation).
  - It is not terribly meaningful because, its scale or metric is determined by the two specific variables on which it is calculated.

RSS
Research and Statistical Support

Covariance

The Covariance is a non-standardized measure of relationship; meaning it can not be used to compare the relationship of two variables to the relationship of two other variables.

- Covariance is not terribly meaningful by itself, but it is used in calculating other statistics (e.g., correlation).
  - It is not terribly meaningful because, its scale or metric is determined by the two specific variables on which it is calculated.
  - For this reason, it is not comparable across different pairs of variables.

What is Covariance then?

- Calculating the covariance gives us a numeric measure of the degree or strength of relationship between two variables.

What is Covariance then?

- Calculating the covariance gives us a numeric measure of the degree or strength of relationship between two variables.
  - Covariance can range between $-\infty$ and $+\infty$

Research and Statistical Support

What is Covariance then?

- Calculating the covariance gives us a numeric measure of the degree or strength of relationship between two variables.
  - Covariance can range between $-\infty$ and $+\infty$
  - The larger the number (negatively or positively), the greater or stronger the relationship.

What is Covariance then?

- Calculating the covariance gives us a numeric measure of the degree or strength of relationship between two variables.
  - Covariance can range between $-\infty$ and $+\infty$
  - The larger the number (negatively or positively), the greater or stronger the relationship.
  - When covariance is zero, there is **no** relationship between the variables; virtually never happens.

What is Covariance then?

- Calculating the covariance gives us a numeric measure of the degree or strength of relationship between two variables.
  - Covariance can range between $-\infty$ and $+\infty$
  - The larger the number (negatively or positively), the greater or stronger the relationship.
  - When covariance is zero, there is **no** relationship between the variables; virtually never happens.
- The sign associated with a covariance tells us the *direction* of the relationship between the two variables.

What is Covariance then?

- Calculating the covariance gives us a numeric measure of the degree or strength of relationship between two variables.
    - Covariance can range between $-\infty$ and $+\infty$
    - The larger the number (negatively or positively), the greater or stronger the relationship.
    - When covariance is zero, there is **no** relationship between the variables; virtually never happens.
- The sign associated with a covariance tells us the *direction* of the relationship between the two variables.
    - If the sign is negative, then high scores on one variable are associated with low scores on the other variable.

RSS
Research and Statistical Support

What is Covariance then?

- Calculating the covariance gives us a numeric measure of the degree or strength of relationship between two variables.
  - Covariance can range between $-\infty$ and $+\infty$
  - The larger the number (negatively or positively), the greater or stronger the relationship.
  - When covariance is zero, there is **no** relationship between the variables; virtually never happens.
- The sign associated with a covariance tells us the *direction* of the relationship between the two variables.
  - If the sign is negative, then high scores on one variable are associated with low scores on the other variable.
  - If the sign is positive, then high scores on one variable are associated with high scores on the other variable.

Calculating Covariance

The definitional formula for calculating the covariance of two *sample* variables (X, Y) is:

$$COV_{xy} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{n - 1}$$

RSS
Research and Statistical Support

Calculating Covariance

The definitional formula for calculating the covariance of two *sample* variables (X, Y) is:

$$COV_{xy} = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{n - 1}$$

This formula is very similar to the variance formula; for instance if we swap out all the Y's in the above formula for more X's, we get the variance of X:

Calculating Covariance

The definitional formula for calculating the covariance of two *sample* variables (X, Y) is:

$$COV_{xy} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{n-1}$$

This formula is very similar to the variance formula; for instance if we swap out all the Y's in the above formula for more X's, we get the variance of X:

$$S_X^2 = \frac{\sum(X-\overline{X})(X-\overline{X})}{n-1} = \frac{\sum(X-\overline{X})^2}{n-1}$$

RSS
Research and Statistical Support

Computational formula for Covariance

- The computational formula is generally considered more manageable when calculating by hand.

Computational formula for Covariance

- The computational formula is generally considered more manageable when calculating by hand.
- But; as mentioned previously with standard deviation, both the computational and definitional formulas provide the same answer.

Computational formula for Covariance

- The computational formula is generally considered more manageable when calculating by hand.
- But; as mentioned previously with standard deviation, both the computational and definitional formulas provide the same answer.

$$COV_{XY} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{n-1}$$

Oil Example sample data: Covariance Calculation

XY_i	Barrels (X)	Costs (Y)	XY
1	159	520	82680
2	166	570	94620
3	176	510	89760
4	185	560	103600
5	191	560	106960
6	194	530	102820
7	199	560	111440
8	207	580	120060
9	216	550	118800
10	228	560	127680
	$\sum X = 1921$	$\sum Y = 5500$	$\sum XY = 1058420$

$$n = 10$$

Example Calculation continued

Taking the sums and n from the previous slide, we can use the computational formula to complete the calculation of covariance.

Example Calculation continued

Taking the sums and *n* from the previous slide, we can use the computational formula to complete the calculation of covariance.

$$COV_{XY} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{n-1} = \frac{1058420 - \frac{(1921)(5500)}{10}}{10-1} \ldots$$

RSS Research and Statistical Support

Example Calculation continued

Taking the sums and *n* from the previous slide, we can use the computational formula to complete the calculation of covariance.

$$COV_{XY} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{n-1} = \frac{1058420 - \frac{(1921)(5500)}{10}}{10-1} \ldots$$

$$COV_{XY} = \frac{1058420 - 1056550}{9} = \frac{1870}{9} = 207.78$$

Example Calculation continued

Taking the sums and n from the previous slide, we can use the computational formula to complete the calculation of covariance.

$$COV_{XY} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{n-1} = \frac{1058420 - \frac{(1921)(5500)}{10}}{10-1} \ldots$$

$$COV_{XY} = \frac{1058420 - 1056550}{9} = \frac{1870}{9} = 207.78$$

So, the covariance of X and Y is 207.78; which does not seem terribly meaningful.

RSS
Research and Statistical Support

Example Calculation continued

Taking the sums and n from the previous slide, we can use the computational formula to complete the calculation of covariance.

$$COV_{XY} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{n-1} = \frac{1058420 - \frac{(1921)(5500)}{10}}{10-1} \ldots$$

$$COV_{XY} = \frac{1058420 - 1056550}{9} = \frac{1870}{9} = 207.78$$

So, the covariance of X and Y is 207.78; which does not seem terribly meaningful.

- Positive number, indicates that high scores on X are associated with high scores on Y (and vice versa).

Research and Statistical Support

Example Calculation continued

Taking the sums and n from the previous slide, we can use the computational formula to complete the calculation of covariance.

$$COV_{XY} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{n-1} = \frac{1058420 - \frac{(1921)(5500)}{10}}{10-1} \ldots$$

$$COV_{XY} = \frac{1058420 - 1056550}{9} = \frac{1870}{9} = 207.78$$

So, the covariance of X and Y is 207.78; which does not seem terribly meaningful.

- Positive number, indicates that high scores on X are associated with high scores on Y (and vice versa).
- Large number (i.e., far from zero), so the two variables are likely to be fairly well related.

RSS
Research and Statistical Support

Example Calculation continued

Taking the sums and n from the previous slide, we can use the computational formula to complete the calculation of covariance.

$$COV_{XY} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{n-1} = \frac{1058420 - \frac{(1921)(5500)}{10}}{10-1} \ldots$$

$$COV_{XY} = \frac{1058420 - 1056550}{9} = \frac{1870}{9} = 207.78$$

So, the covariance of X and Y is 207.78; which does not seem terribly meaningful.

- Positive number, indicates that high scores on X are associated with high scores on Y (and vice versa).
- Large number (i.e., far from zero), so the two variables are likely to be fairly well related.
- Beyond that, not much can be said.

RSS
Research and Statistical Support

Correlation

The Correlation (r) is a standardized measure of relationship; meaning it can be compared across multiple pairs of variables, regardless of scale.

[4] There will be a great deal more discussion of correlation later in the course.

Correlation

The Correlation (r) is a standardized measure of relationship; meaning it can be compared across multiple pairs of variables, regardless of scale.

- Correlation is the most frequently used statistic for assessing the relationship between two variables[4].

[4] There will be a great deal more discussion of correlation later in the course.

Correlation

The Correlation (r) is a standardized measure of relationship; meaning it can be compared across multiple pairs of variables, regardless of scale.

- Correlation is the most frequently used statistic for assessing the relationship between two variables[4].
- Correlation allows us to describe the direction and magnitude of a relationship between two variables.

[4] There will be a great deal more discussion of correlation later in the course.

Correlation

The Correlation (r) is a standardized measure of relationship; meaning it can be compared across multiple pairs of variables, regardless of scale.

- Correlation is the most frequently used statistic for assessing the relationship between two variables[4].
- Correlation allows us to describe the direction and magnitude of a relationship between two variables.
- Correlation is very similar to covariance, indeed we use the covariance to calculate correlation.

[4] There will be a great deal more discussion of correlation later in the course.

Correlation

The Correlation (r) is a standardized measure of relationship; meaning it can be compared across multiple pairs of variables, regardless of scale.

- Correlation is the most frequently used statistic for assessing the relationship between two variables[4].
- Correlation allows us to describe the direction and magnitude of a relationship between two variables.
- Correlation is very similar to covariance, indeed we use the covariance to calculate correlation.
- Correlation is used in many inferential statistics and often a matrix of correlations is the input data used to calculate them.

[4] There will be a great deal more discussion of correlation later in the course.

Interpretation of Correlation

Once calculated:

Interpretation of Correlation

Once calculated:

- Correlation (r) can range between -1 and $+1$.

Interpretation of Correlation

Once calculated:

- Correlation (r) can range between -1 and $+1$.
- The larger the value (positive or negative), the stronger the relationship between the variables.

RSS
Research and Statistical Support

Interpretation of Correlation

Once calculated:

- Correlation (r) can range between -1 and $+1$.
- The larger the value (positive or negative), the stronger the relationship between the variables.
- If r is negative, then high scores on one variable are associated with low scores on the other variable.

RSS
Research and Statistical Support

Interpretation of Correlation

Once calculated:

- Correlation (r) can range between -1 and $+1$.
- The larger the value (positive or negative), the stronger the relationship between the variables.
- If r is negative, then high scores on one variable are associated with low scores on the other variable.
- If r is positive, then high scores on one variable are associated with high scores on the other variable.

RSS
Research and Statistical Support

Interpretation of Correlation

Once calculated:

- Correlation (r) can range between -1 and $+1$.
- The larger the value (positive or negative), the stronger the relationship between the variables.
- If r is negative, then high scores on one variable are associated with low scores on the other variable.
- If r is positive, then high scores on one variable are associated with high scores on the other variable.
- If $r = 0$, then there is no relationship between the variables (virtually never occurs).

Interpretation of Correlation

Once calculated:

- Correlation (r) can range between -1 and $+1$.
- The larger the value (positive or negative), the stronger the relationship between the variables.
- If r is negative, then high scores on one variable are associated with low scores on the other variable.
- If r is positive, then high scores on one variable are associated with high scores on the other variable.
- If $r = 0$, then there is no relationship between the variables (virtually never occurs).

The size of r indicates the *strength* of the relationship and the sign (positive or negative) indicates the *direction* of the relationship.

RSS
Research and Statistical Support

Calculating Correlation

Calculating correlation is quite easy, once you have the covariance.

Calculating Correlation

Calculating correlation is quite easy, once you have the covariance.

$$r_{XY} = \frac{cov_{XY}}{s_X s_Y}$$

Calculating Correlation

Calculating correlation is quite easy, once you have the covariance.

$$r_{XY} = \frac{COV_{XY}}{S_X S_Y}$$

- So, given the descriptive statistics from previous slides:

RSS Research and Statistical Support

Calculating Correlation

Calculating correlation is quite easy, once you have the covariance.

$$r_{XY} = \frac{COV_{XY}}{S_X S_Y}$$

- So, given the descriptive statistics from previous slides:
  - Barrels (X): $S_X = 21.66$
  - Costs (Y): $S_Y = 22.61$
  - and: $COV_{XY} = 207.78$

RSS
Research and Statistical Support

Calculating Correlation

Calculating correlation is quite easy, once you have the covariance.

$$r_{XY} = \frac{COV_{XY}}{S_X S_Y}$$

- So, given the descriptive statistics from previous slides:
  - Barrels (X): $S_X = 21.66$
  - Costs (Y): $S_Y = 22.61$
  - and: $COV_{XY} = 207.78$

$$r_{XY} = \frac{207.78}{(21.66)(22.61)} = \frac{207.78}{489.73} = 0.424$$

RSS Research and Statistical Support

Calculating Correlation

Calculating correlation is quite easy, once you have the covariance.

$$r_{XY} = \frac{COV_{XY}}{S_X S_Y}$$

- So, given the descriptive statistics from previous slides:
  - Barrels (X): $S_X = 21.66$
  - Costs (Y): $S_Y = 22.61$
  - and: $COV_{XY} = 207.78$

$$r_{XY} = \frac{207.78}{(21.66)(22.61)} = \frac{207.78}{489.73} = 0.424$$

- The correlation is .424 between Barrels and Costs.

RSS
Research and Statistical Support

What does that *mean*?

The correlation between Barrels and Costs is 0.424...so what?

What does that *mean*?

The correlation between Barrels and Costs is 0.424...so what?

- We can say that Barrels and Costs are positively related.

What does that *mean*?

The correlation between Barrels and Costs is 0.424...so what?

- We can say that Barrels and Costs are positively related.
  - Meaning; high scores on one tend to be associated with high scores on the other.
  - And, low scores on one tend to be associated with low scores on the other.

RSS
Research and Statistical Support

What does that *mean*?

The correlation between Barrels and Costs is 0.424...so what?

- We can say that Barrels and Costs are positively related.
  - Meaning; high scores on one tend to be associated with high scores on the other.
  - And, low scores on one tend to be associated with low scores on the other.
- But what of the magnitude?

RSS Research and Statistical Support

What does that *mean*?

The correlation between Barrels and Costs is 0.424...so what?

- We can say that Barrels and Costs are positively related.
  - Meaning; high scores on one tend to be associated with high scores on the other.
  - And, low scores on one tend to be associated with low scores on the other.
- But what of the magnitude? This is a bit more tricky.

RSS
Research and Statistical Support

What does that *mean*?

The correlation between Barrels and Costs is 0.424...so what?

- We can say that Barrels and Costs are positively related.
  - Meaning; high scores on one tend to be associated with high scores on the other.
  - And, low scores on one tend to be associated with low scores on the other.
- But what of the magnitude? This is a bit more tricky.
  - Generally, familiarity with recent, similar research will guide your interpretation of magnitude.

RSS
Research and Statistical Support

What does that *mean*?

The correlation between Barrels and Costs is 0.424...so what?

- We can say that Barrels and Costs are positively related.
  - Meaning; high scores on one tend to be associated with high scores on the other.
  - And, low scores on one tend to be associated with low scores on the other.
- But what of the magnitude? This is a bit more tricky.
  - Generally, familiarity with recent, similar research will guide your interpretation of magnitude.
    - Same field, topic, variables, etc.

RSS
Research and Statistical Support

What does that *mean*?

The correlation between Barrels and Costs is 0.424...so what?

- We can say that Barrels and Costs are positively related.
  - Meaning; high scores on one tend to be associated with high scores on the other.
  - And, low scores on one tend to be associated with low scores on the other.
- But what of the magnitude? This is a bit more tricky.
  - Generally, familiarity with recent, similar research will guide your interpretation of magnitude.
    - Same field, topic, variables, etc.
  - In the social sciences, it is common to find correlations around .400 to .600 referred to as ‘moderate’, ‘good’, or even ‘strong’ (in the case of .600).

RSS
Research and Statistical Support

Taking Correlation a step further.

One very good way of helping yourself to interpret correlation is to *square* it.

RSS
Research and Statistical Support

Taking Correlation a step further.

One very good way of helping yourself to interpret correlation is to *square* it.

- By squaring the correlation (r^2), we can interpret it as the amount of variance shared between the two variables.

RSS
Research and Statistical Support

Taking Correlation a step further.

One very good way of helping yourself to interpret correlation is to *square* it.

- By squaring the correlation (r^2), we can interpret it as the amount of variance shared between the two variables.

$$r^2 = .424^2 = .1798$$

RSS
Research and Statistical Support

Taking Correlation a step further.

One very good way of helping yourself to interpret correlation is to *square* it.

- By squaring the correlation (r^2), we can interpret it as the amount of variance shared between the two variables.

$$r^2 = .424^2 = .1798$$

- So, we can say barrels extracted and rig operating costs share 17.98% of their variance.

RSS
Research and Statistical Support

Taking Correlation a step further.

One very good way of helping yourself to interpret correlation is to *square* it.

- By squaring the correlation (r^2), we can interpret it as the amount of variance shared between the two variables.

$$r^2 = .424^2 = .1798$$

- So, we can say barrels extracted and rig operating costs share 17.98% of their variance.
- Now we have a better understanding of the relationship between the two variables.

Taking Correlation a step further.

One very good way of helping yourself to interpret correlation is to *square* it.

- By squaring the correlation (r^2), we can interpret it as the amount of variance shared between the two variables.

$$r^2 = .424^2 = .1798$$

- So, we can say barrels extracted and rig operating costs share 17.98% of their variance.
- Now we have a better understanding of the relationship between the two variables.
- Keep in mind:

RSS
Research and Statistical Support

Taking Correlation a step further.

One very good way of helping yourself to interpret correlation is to *square* it.

- By squaring the correlation (r^2), we can interpret it as the amount of variance shared between the two variables.

$$r^2 = .424^2 = .1798$$

- So, we can say barrels extracted and rig operating costs share 17.98% of their variance.
- Now we have a better understanding of the relationship between the two variables.
- Keep in mind:
  - Squaring any correlation coefficient makes it smaller (r is always between -1 and +1).

RSS
Research and Statistical Support

Adjusted Correlation

When sample sizes are small, as they are here ($n = 10$), the sample correlation will tend to overestimate the population correlation.

Adjusted Correlation

When sample sizes are small, as they are here ($n = 10$), the sample correlation will tend to overestimate the population correlation.

- Meaning, r and r^2 tend to be larger than they truly are in the population.
- The relationship appears stronger than it actually is in the population.

RSS
Research and Statistical Support

Adjusted Correlation

When sample sizes are small, as they are here ($n = 10$), the sample correlation will tend to overestimate the population correlation.

- Meaning, r and r^2 tend to be larger than they truly are in the population.
- The relationship appears stronger than it actually is in the population.
- So, we generally correct for this problem by *adjusting r*.

Adjusted Correlation

When sample sizes are small, as they are here ($n = 10$), the sample correlation will tend to overestimate the population correlation.

- Meaning, r and r^2 tend to be larger than they truly are in the population.
- The relationship appears stronger than it actually is in the population.
- So, we generally correct for this problem by *adjusting r*.

$$r_{adj} = \sqrt{1 - \frac{(1-r^2)(n-1)}{n-2}}$$

Adjusting our example correlation

Adjusting our Oil example correlation:

Adjusting our example correlation

Adjusting our Oil example correlation:

$$r_{adj} = \sqrt{1 - \frac{(1-r^2)(n-1)}{n-2}} = \sqrt{1 - \frac{(1-.424^2)(10-1)}{10-2}} = \ldots$$

RSS
Research and Statistical Support

Adjusting our example correlation

Adjusting our Oil example correlation:

$$r_{adj} = \sqrt{1 - \frac{(1-r^2)(n-1)}{n-2}} = \sqrt{1 - \frac{(1-.424^2)(10-1)}{10-2}} = \ldots$$

$$\sqrt{1 - \frac{(1-.1798)(9)}{8}} = \sqrt{1 - \frac{(.8202)(9)}{8}} = \ldots$$

Adjusting our example correlation

Adjusting our Oil example correlation:

$$r_{adj} = \sqrt{1 - \frac{(1-r^2)(n-1)}{n-2}} = \sqrt{1 - \frac{(1-.424^2)(10-1)}{10-2}} = \ldots$$

$$\sqrt{1 - \frac{(1-.1798)(9)}{8}} = \sqrt{1 - \frac{(.8202)(9)}{8}} = \ldots$$

$$\sqrt{1 - \frac{7.3818}{8}} = \sqrt{1 - .9227} = \ldots$$

RSS
Research and Statistical Support

Adjusting our example correlation

Adjusting our Oil example correlation:

$$r_{adj} = \sqrt{1 - \frac{(1-r^2)(n-1)}{n-2}} = \sqrt{1 - \frac{(1-.424^2)(10-1)}{10-2}} = \ldots$$

$$\sqrt{1 - \frac{(1-.1798)(9)}{8}} = \sqrt{1 - \frac{(.8202)(9)}{8}} = \ldots$$

$$\sqrt{1 - \frac{7.3818}{8}} = \sqrt{1 - .9227} = \ldots$$

$$\sqrt{.0773} = .2780$$

RSS Research and Statistical Support

Adjusting our example correlation

Adjusting our Oil example correlation:

$$r_{adj} = \sqrt{1 - \frac{(1-r^2)(n-1)}{n-2}} = \sqrt{1 - \frac{(1-.424^2)(10-1)}{10-2}} = \ldots$$

$$\sqrt{1 - \frac{(1-.1798)(9)}{8}} = \sqrt{1 - \frac{(.8202)(9)}{8}} = \ldots$$

$$\sqrt{1 - \frac{7.3818}{8}} = \sqrt{1 - .9227} = \ldots$$

$$\sqrt{.0773} = .2780$$

- Our correlation shrank from $r = .424$ to $r_{adj} = .278$

Adjusting our example correlation

Adjusting our Oil example correlation:

$$r_{adj} = \sqrt{1 - \frac{(1-r^2)(n-1)}{n-2}} = \sqrt{1 - \frac{(1-.424^2)(10-1)}{10-2}} = \ldots$$

$$\sqrt{1 - \frac{(1-.1798)(9)}{8}} = \sqrt{1 - \frac{(.8202)(9)}{8}} = \ldots$$

$$\sqrt{1 - \frac{7.3818}{8}} = \sqrt{1 - .9227} = \ldots$$

$$\sqrt{.0773} = .2780$$

- Our correlation shrank from $r = .424$ to $r_{adj} = .278$
- Shared variance shrank from $r^2 = .1789$ to $r^2_{adj} = .0773$.

Adjusting our example correlation

Adjusting our Oil example correlation:

$$r_{adj} = \sqrt{1 - \frac{(1-r^2)(n-1)}{n-2}} = \sqrt{1 - \frac{(1-.424^2)(10-1)}{10-2}} = \ldots$$

$$\sqrt{1 - \frac{(1-.1798)(9)}{8}} = \sqrt{1 - \frac{(.8202)(9)}{8}} = \ldots$$

$$\sqrt{1 - \frac{7.3818}{8}} = \sqrt{1 - .9227} = \ldots$$

$$\sqrt{.0773} = .2780$$

- Our correlation shrank from $r = .424$ to $r_{adj} = .278$
- Shared variance shrank from $r^2 = .1789$ to $r^2_{adj} = .0773$.
- We now have a *more* accurate sample estimate of the relationship between Barrels and Costs.

RSS
Research and Statistical Support

Adjusting our example correlation

Adjusting our Oil example correlation:

$$r_{adj} = \sqrt{1 - \frac{(1-r^2)(n-1)}{n-2}} = \sqrt{1 - \frac{(1-.424^2)(10-1)}{10-2}} = \ldots$$

$$\sqrt{1 - \frac{(1-.1798)(9)}{8}} = \sqrt{1 - \frac{(.8202)(9)}{8}} = \ldots$$

$$\sqrt{1 - \frac{7.3818}{8}} = \sqrt{1 - .9227} = \ldots$$

$$\sqrt{.0773} = .2780$$

- Our correlation shrank from $r = .424$ to $r_{adj} = .278$
- Shared variance shrank from $r^2 = .1789$ to $r^2_{adj} = .0773$.
- We now have a *more* accurate sample estimate of the relationship between Barrels and Costs.
- They share 7.73% of their variance (i.e., clearly a weak relationship).

RSS
Research and Statistical Support

Additional Considerations with Measures of Relationship

Measures of relationship tell us something about whether or not two (or more) variables share variance.

Additional Considerations with Measures of Relationship

Measures of relationship tell us something about whether or not two (or more) variables share variance.
They do **NOT** tell us what causes the relationship!

Additional Considerations with Measures of Relationship

Measures of relationship tell us something about whether or not two (or more) variables share variance.

They do **NOT** tell us what causes the relationship!

Nor do they tell us if one variable causes another!

Additional Considerations with Measures of Relationship

Measures of relationship tell us something about whether or not two (or more) variables share variance.

They do **NOT** tell us what causes the relationship!

Nor do they tell us if one variable causes another!

- You will often hear this: “Correlation does **not** equal causation!”

RSS Research and Statistical Support

Additional Considerations with Measures of Relationship

Measures of relationship tell us something about whether or not two (or more) variables share variance.

They do **NOT** tell us what causes the relationship!

Nor do they tell us if one variable causes another!

- You will often hear this: “Correlation does **not** equal causation!”
  - X may cause Y
  - Y may cause X
  - Z may be causing the relationship between X and Y

Additional Considerations with Measures of Relationship

Measures of relationship tell us something about whether or not two (or more) variables share variance.
They do **NOT** tell us what causes the relationship!
Nor do they tell us if one variable causes another!

- You will often hear this: “Correlation does **not** equal causation!”
  - X may cause Y
  - Y may cause X
  - Z may be causing the relationship between X and Y
- Although, we do tend to use correlation (and other measures of relationship) in the process of *investigating* causal relationships.

RSS
Research and Statistical Support

Properties of Statistics

There are 4 properties we use to evaluate statistics.

Properties of Statistics

There are 4 properties we use to evaluate statistics.

- Sufficiency

Properties of Statistics

There are 4 properties we use to evaluate statistics.

- Sufficiency
- Unbiasedness

Properties of Statistics

There are 4 properties we use to evaluate statistics.

- Sufficiency
- Unbiasedness
- Efficiency

Properties of Statistics

There are 4 properties we use to evaluate statistics.

- Sufficiency
- Unbiasedness
- Efficiency
- Resistance

Properties of Statistics

There are 4 properties we use to evaluate statistics.

- Sufficiency
- Unbiasedness
- Efficiency
- Resistance

Some of them you are already familiar with...

Sufficiency

The Sufficiency of a statistic refers to whether or not it makes use of all the information contained in a sample to estimate its corresponding parameter.

Sufficiency

The Sufficiency of a statistic refers to whether or not it makes use of all the information contained in a sample to estimate its corresponding parameter.

- As an example, consider measures of central tendency:

Sufficiency

The Sufficiency of a statistic refers to whether or not it makes use of all the information contained in a sample to estimate its corresponding parameter.

- As an example, consider measures of central tendency:
  - The mean is very *sufficient* because, it uses all the scores when being calculated.

RSS
Research and Statistical Support

Sufficiency

The Sufficiency of a statistic refers to whether or not it makes use of all the information contained in a sample to estimate its corresponding parameter.

- As an example, consider measures of central tendency:
  - The mean is very *sufficient* because, it uses all the scores when being calculated.
  - The median and mode are not very sufficient because, they only use one or two scores.

RSS
Research and Statistical Support

Unbiasedness

The Unbiasedness refers to how well a sample statistic represents its associated population parameter.

Unbiasedness

The Unbiasedness refers to how well a sample statistic represents its associated population parameter.

- As we saw with correlation, some statistics (r) are more *biased* than others (r_{adj}).

RSS
Research and Statistical Support

Unbiasedness

The Unbiasedness refers to how well a sample statistic represents its associated population parameter.

- As we saw with correlation, some statistics (r) are more *biased* than others (r_{adj}).
- Recall how we calculated:

Unbiasedness

The Unbiasedness refers to how well a sample statistic represents its associated population parameter.

- As we saw with correlation, some statistics (r) are more *biased* than others (r_{adj}).
- Recall how we calculated:
  - Sample variance: $S^2 = \sum (X - \overline{X})^2 / n - 1$

Unbiasedness

The Unbiasedness refers to how well a sample statistic represents its associated population parameter.

- As we saw with correlation, some statistics (r) are more *biased* than others (r_{adj}).
- Recall how we calculated:
  - Sample variance: $S^2 = \sum (X - \overline{X})^2 / n - 1$
  - Population variance: $\sigma^2 = \sum (X - \mu)^2 / N$

Unbiasedness

The Unbiasedness refers to how well a sample statistic represents its associated population parameter.

- As we saw with correlation, some statistics (r) are more *biased* than others (r_{adj}).
- Recall how we calculated:
  - Sample variance: $S^2 = \sum(X - \overline{X})^2/n - 1$
  - Population variance: $\sigma^2 = \sum(X - \mu)^2/N$
- This is due to something we will discuss more later, *degrees of freedom (df)*.

Unbiasedness

The Unbiasedness refers to how well a sample statistic represents its associated population parameter.

- As we saw with correlation, some statistics (r) are more *biased* than others (r_{adj}).
- Recall how we calculated:
  - Sample variance: $S^2 = \sum (X - \overline{X})^2 / n - 1$
  - Population variance: $\sigma^2 = \sum (X - \mu)^2 / N$
- This is due to something we will discuss more later, *degrees of freedom (df)*.
- For now, consider this: we use N in the population formula because, we have **all** of the scores.

Unbiasedness

The Unbiasedness refers to how well a sample statistic represents its associated population parameter.

- As we saw with correlation, some statistics (r) are more *biased* than others (r_{adj}).
- Recall how we calculated:
  - Sample variance: $S^2 = \sum(X - \overline{X})^2/n - 1$
  - Population variance: $\sigma^2 = \sum(X - \mu)^2/N$
- This is due to something we will discuss more later, *degrees of freedom (df)*.
- For now, consider this: we use N in the population formula because, we have **all** of the scores.
- When dealing with samples, we do not have all the scores (of the defined population) and we make an adjustment, dividing by $n - 1$.

RSS
Research and Statistical Support

Efficiency

The Efficiency refers to how much a statistic can change from sample to sample. An efficient statistic does not change.

Efficiency

The Efficiency refers to how much a statistic can change from sample to sample. An efficient statistic does not change.

- Consider the sample mean: $\overline{X} = \sum X/n$

Efficiency

The Efficiency refers to how much a statistic can change from sample to sample. An efficient statistic does not change.

- Consider the sample mean: $\overline{X} = \sum X/n$
- If we took an infinite number of repeated samples from a symmetrical population distribution with μ in the center:

Efficiency

The Efficiency refers to how much a statistic can change from sample to sample. An efficient statistic does not change.

- Consider the sample mean: $\overline{X} = \sum X/n$
- If we took an infinite number of repeated samples from a symmetrical population distribution with μ in the center:
  - The mean of each sample would be fairly close to μ and the mean of all those sample means would **be** μ.

Research and Statistical Support

Efficiency

The Efficiency refers to how much a statistic can change from sample to sample. An efficient statistic does not change.

- Consider the sample mean: $\overline{X} = \sum X/n$
- If we took an infinite number of repeated samples from a symmetrical population distribution with μ in the center:
  - The mean of each sample would be fairly close to μ and the mean of all those sample means would **be** μ.
- The key to that statement being true is "symmetrical population distribution with μ in the center".

Research and Statistical Support

Efficiency

The Efficiency refers to how much a statistic can change from sample to sample. An efficient statistic does not change.

- Consider the sample mean: $\overline{X} = \sum X/n$
- If we took an infinite number of repeated samples from a symmetrical population distribution with μ in the center:
  - The mean of each sample would be fairly close to μ and the mean of all those sample means would **be** μ.
- The key to that statement being true is "symmetrical population distribution with μ in the center".
  - Extremely high and low scores (those farthest from μ) are rare when compared to the number of scores near μ.

RSS
Research and Statistical Support

Efficiency

The Efficiency refers to how much a statistic can change from sample to sample. An efficient statistic does not change.

- Consider the sample mean: $\overline{X} = \sum X/n$
- If we took an infinite number of repeated samples from a symmetrical population distribution with μ in the center:
  - The mean of each sample would be fairly close to μ and the mean of all those sample means would **be** μ.
- The key to that statement being true is "symmetrical population distribution with μ in the center".
  - Extremely high and low scores (those farthest from μ) are rare when compared to the number of scores near μ.
  - Therefore, we can expect most of those repeated samples to have a mean close to μ, because most of the scores in general (in the population) are close to μ.

RSS
Research and Statistical Support

Resistance

The Resistance refers to how *resistant* a statistic is to outliers (extreme scores).

Resistance

The Resistance refers to how *resistant* a statistic is to outliers (extreme scores).

- If extreme scores do **not** influence the statistic, then the statistic is *resistant*.

RSS
Research and Statistical Support

Resistance

The Resistance refers to how *resistant* a statistic is to outliers (extreme scores).

- If extreme scores do **not** influence the statistic, then the statistic is *resistant*.
- Again, consider measures of central tendency: Mo, Mdn, $\overline{X}$

Research and Statistical Support

Resistance

The Resistance refers to how *resistant* a statistic is to outliers (extreme scores).

- If extreme scores do **not** influence the statistic, then the statistic is *resistant*.
- Again, consider measures of central tendency: Mo, Mdn, $\overline{X}$
- Both Mo and Mdn only consider the very center of a distribution, so they are very *resistant*.

RSS
Research and Statistical Support

Resistance

The Resistance refers to how *resistant* a statistic is to outliers (extreme scores).

- If extreme scores do **not** influence the statistic, then the statistic is *resistant*.
- Again, consider measures of central tendency: Mo, Mdn, $\overline{X}$
- Both Mo and Mdn only consider the very center of a distribution, so they are very *resistant*.
- $\overline{X}$ is very sensitive to outliers, they *pull* the mean toward them thus making the mean not very resistant.

Summary of Module 3 (continued on next slide)

- Introduced:

Summary of Module 3 (continued on next slide)

- Introduced:
  - The context of an example study

Summary of Module 3 (continued on next slide)

- Introduced:
  - The context of an example study
  - Descriptive Statistics

Summary of Module 3 (continued on next slide)

- Introduced:
  - The context of an example study
  - Descriptive Statistics
- Classes of Descriptive Statistics

Summary of Module 3 (continued on next slide)

- Introduced:
  - The context of an example study
  - Descriptive Statistics
- Classes of Descriptive Statistics
  - Central Tendency
    - Mode
    - Median
    - Mean

Summary of Module 3 (continued on next slide)

- Introduced:
  - The context of an example study
  - Descriptive Statistics
- Classes of Descriptive Statistics
  - Central Tendency
    - Mode
    - Median
    - Mean
  - Dispersion
    - Range
    - Sums of Squares
    - Variance
    - Standard Deviation
    - Coefficient of Variation

Summary of Module 3 (continued on next slide)

- Introduced:
  - The context of an example study
  - Descriptive Statistics
- Classes of Descriptive Statistics
  - Central Tendency
    - Mode
    - Median
    - Mean
  - Dispersion
    - Range
    - Sums of Squares
    - Variance
    - Standard Deviation
    - Coefficient of Variation
  - Shape
    - Skewness
    - Kurtosis

RSS Research and Statistical Support

Summary continued

- Classes of Descriptive Statistics (continued)

Summary continued

- Classes of Descriptive Statistics (continued)
  - Relationship
    - Covariance
    - Correlation
    - Adjusted Correlation

RSS
Research and Statistical Support

Summary continued

- Classes of Descriptive Statistics (continued)
  - Relationship
    - Covariance
    - Correlation
    - Adjusted Correlation
- Properties of statistics

RSS
Research and Statistical Support

Summary continued

- Classes of Descriptive Statistics (continued)
  - Relationship
    - Covariance
    - Correlation
    - Adjusted Correlation
- Properties of statistics
  - Sufficiency

Summary continued

- Classes of Descriptive Statistics (continued)
  - Relationship
    - Covariance
    - Correlation
    - Adjusted Correlation
- Properties of statistics
  - Sufficiency
  - Unbiasedness

Summary continued

- Classes of Descriptive Statistics (continued)
  - Relationship
    - Covariance
    - Correlation
    - Adjusted Correlation
- Properties of statistics
  - Sufficiency
  - Unbiasedness
  - Efficiency

Summary continued

- Classes of Descriptive Statistics (continued)
    - Relationship
        - Covariance
        - Correlation
        - Adjusted Correlation
- Properties of statistics
    - Sufficiency
    - Unbiasedness
    - Efficiency
    - Resistance

This concludes Module 3

Next time Module 4.

- Next time we'll begin covering "The Normal Curve".
- Until next time; have a nice day.

These slides initially created on: September 16, 2010
These slides last updated on: September 23, 2010

- The bottom date shown is the date this Adobe.pdf file was created; LaTeX[5] has a command for automatically inserting the date of a document's creation.

RSS
Research and Statistical Support

[5]This document was created in LaTeX using the Beamer package