Chapter 2 - Describing Data: Numerical

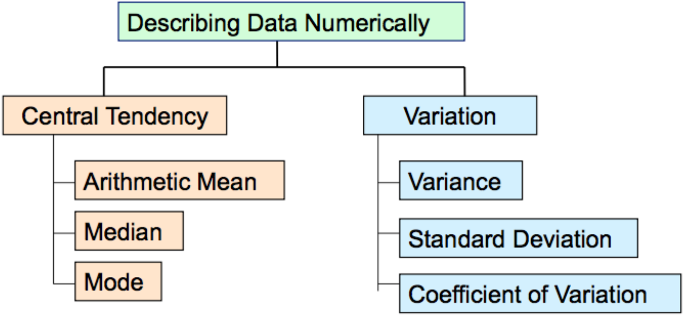
Ercan Karadas

New York University

ercan@nyu.edu

Summer, 2016

Content


Describing Data Numerically
Central Tendency
Arithmetic Mean
Median
Mode
Variation
Variance
Standard Deviation
Coefficient of Variation

Arithmetic Mean

- The arithmetic mean (or simply mean) is the most commonly used measure of central tendency.
- Let $\{x_1, x_2, \ldots, x_N\}$ be a set of values obtained (or observed) from a **population**, then population mean (average) is given by

Arithmetic Mean

- The arithmetic mean (or simply mean) is the most commonly used measure of central tendency.
- Let $\{x_1, x_2, \ldots, x_N\}$ be a set of values obtained (or observed) from a **population**, then population mean (average) is given by

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

Arithmetic Mean

- The arithmetic mean (or simply mean) is the most commonly used measure of central tendency.
- Let $\{x_1, x_2, \ldots, x_N\}$ be a set of values obtained (or observed) from a **population**, then population mean (average) is given by

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

  - Population average is denoted by μ, and N denotes the population size.

Arithmetic Mean

- The arithmetic mean (or simply mean) is the most commonly used measure of central tendency.
- Let $\{x_1, x_2, \ldots, x_N\}$ be a set of values obtained (or observed) from a **population**, then population mean (average) is given by

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

  - Population average is denoted by μ, and N denotes the population size.
- Similarly, for a **sample** that is composed of the values $\{x_1, x_2, \ldots, x_n\}$, sample mean (average) can be calculated as

Arithmetic Mean

- The arithmetic mean (or simply mean) is the most commonly used measure of central tendency.
- Let $\{x_1, x_2, \ldots, x_N\}$ be a set of values obtained (or observed) from a **population**, then population mean (average) is given by

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

  - Population average is denoted by μ, and N denotes the population size.
- Similarly, for a **sample** that is composed of the values $\{x_1, x_2, \ldots, x_n\}$, sample mean (average) can be calculated as

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Arithmetic Mean

- The arithmetic mean (or simply mean) is the most commonly used measure of central tendency.
- Let $\{x_1, x_2, \ldots, x_N\}$ be a set of values obtained (or observed) from a **population**, then population mean (average) is given by

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

  - Population average is denoted by μ, and N denotes the population size.
- Similarly, for a **sample** that is composed of the values $\{x_1, x_2, \ldots, x_n\}$, sample mean (average) can be calculated as

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

  - Sample average is denoted by $\overline{X}$, and n is the sample size.

Arithmetic Mean

- The arithmetic mean (or simply mean) is the most commonly used measure of central tendency.

Arithmetic Mean

- The arithmetic mean (or simply mean) is the most commonly used measure of central tendency.
- Pros: Easy to calculate,

Arithmetic Mean

- The arithmetic mean (or simply mean) is the most commonly used measure of central tendency.
- Pros: Easy to calculate,

$$\text{Arith. Mean} = \text{sum of values over number of elements}$$

Arithmetic Mean

- The arithmetic mean (or simply mean) is the most commonly used measure of central tendency.
- Pros: Easy to calculate,

$$\text{Arith. Mean} = \text{sum of values over number of elements}$$

- Cons: It is affected by extreme values (outliers). Therefore, if there are some extreme values among the sample, then arithmetic mean might not be a good measure of central tendency as the following example demonstrates.

Example.

Suppose a sample consists of the values $\{1, 2, 3, 4, 5\}$. Then

Example.

Suppose a sample consists of the values $\{1, 2, 3, 4, 5\}$. Then

- Arithmetic mean

$$\overline{X} = \frac{1 + 2 + 3 + 4 + 5}{5} = 3$$

Example.

Suppose a sample consists of the values $\{1, 2, 3, 4, 5\}$. Then

- Arithmetic mean

$$\overline{X} = \frac{1+2+3+4+5}{5} = 3$$

  - Is 3 a good measure of central tendency?

Example.

Suppose a sample consists of the values $\{1, 2, 3, 4, 5\}$. Then

- Arithmetic mean

$$\overline{X} = \frac{1+2+3+4+5}{5} = 3$$

  - Is 3 a good measure of central tendency? YES.

Example.

Suppose a sample consists of the values $\{1, 2, 3, 4, 5\}$. Then

- Arithmetic mean

$$\overline{X} = \frac{1+2+3+4+5}{5} = 3$$

  - Is 3 a good measure of central tendency? YES.

- Now, let us change only one of the values in the sample

Example.

Suppose a sample consists of the values $\{1, 2, 3, 4, 5\}$. Then

- Arithmetic mean

$$\overline{X} = \frac{1+2+3+4+5}{5} = 3$$

  - Is 3 a good measure of central tendency? YES.
- Now, let us change only one of the values in the sample
- For instance, changing $5 \rightarrow 15$ gives a new sample $\{1, 2, 3, 4, 15\}$.

Example.

Suppose a sample consists of the values $\{1, 2, 3, 4, 5\}$. Then

- Arithmetic mean

$$\overline{X} = \frac{1+2+3+4+5}{5} = 3$$

  - Is 3 a good measure of central tendency? YES.
- Now, let us change only one of the values in the sample
- For instance, changing $5 \to 15$ gives a new sample $\{1, 2, 3, 4, 15\}$.
- New arithmetic mean:

$$\overline{X} = \frac{1+2+3+4+15}{5} = 5$$

Example.

Suppose a sample consists of the values $\{1, 2, 3, 4, 5\}$. Then

- Arithmetic mean

$$\overline{X} = \frac{1+2+3+4+5}{5} = 3$$

  - Is 3 a good measure of central tendency? YES.
- Now, let us change only one of the values in the sample
- For instance, changing $5 \to 15$ gives a new sample $\{1, 2, 3, 4, 15\}$.
- New arithmetic mean:

$$\overline{X} = \frac{1+2+3+4+15}{5} = 5$$

  - Is 5 a good measure of central tendency?

Example.

Suppose a sample consists of the values $\{1, 2, 3, 4, 5\}$. Then

- Arithmetic mean

$$\overline{X} = \frac{1+2+3+4+5}{5} = 3$$

  - Is 3 a good measure of central tendency? YES.
- Now, let us change only one of the values in the sample
- For instance, changing $5 \to 15$ gives a new sample $\{1, 2, 3, 4, 15\}$.
- New arithmetic mean:

$$\overline{X} = \frac{1+2+3+4+15}{5} = 5$$

  - Is 5 a good measure of central tendency? NO.

Example.

Suppose a sample consists of the values $\{1, 2, 3, 4, 5\}$. Then

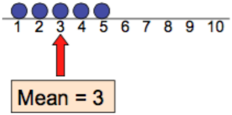
- Arithmetic mean

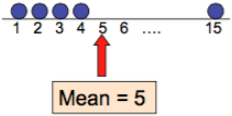
$$\overline{X} = \frac{1+2+3+4+5}{5} = 3$$

  - Is 3 a good measure of central tendency? YES.
- Now, let us change only one of the values in the sample
- For instance, changing $5 \to 15$ gives a new sample $\{1, 2, 3, 4, 15\}$.
- New arithmetic mean:

$$\overline{X} = \frac{1+2+3+4+15}{5} = 5$$

  - Is 5 a good measure of central tendency? NO.
  - Actually, 5 is not even among the values in the sample!





Median

Median: In an ordered data set, the median is the observation right in the middle.

Median

Median: In an ordered data set, the median is the observation right in the middle.

- If there are an odd number of observations, the median leaves exactly 50% above and 50% below.

Median

Median: In an ordered data set, the median is the observation right in the middle.

- If there are an odd number of observations, the median leaves exactly 50% above and 50% below.
- If there are an even number of observations, the median is the average of the two middle values.

Median

Median: In an ordered data set, the median is the observation right in the middle.

- If there are an odd number of observations, the median leaves exactly 50% above and 50% below.
- If there are an even number of observations, the median is the average of the two middle values.
- Not affected by extreme values

Median

Median: In an ordered data set, the median is the observation right in the middle.

- If there are an odd number of observations, the median leaves exactly 50% above and 50% below.
- If there are an even number of observations, the median is the average of the two middle values.
- Not affected by extreme values

Example

Find the median for the following data set
$13, 3, 13, 10, 9, 11, 7, 6, 12, 7, 1, 10, 10, 5$

Median

Median: In an ordered data set, the median is the observation right in the middle.

- If there are an odd number of observations, the median leaves exactly 50% above and 50% below.
- If there are an even number of observations, the median is the average of the two middle values.
- Not affected by extreme values

Example

Find the median for the following data set
$13, 3, 13, 10, 9, 11, 7, 6, 12, 7, 1, 10, 10, 5$

- Data set is not ordered, first sort the data
  $1, 3, 5, 6, 7, 7, 9, 10, 10, 10, 11, 12, 13, 13$

Median

Median: In an ordered data set, the median is the observation right in the middle.

- If there are an odd number of observations, the median leaves exactly 50% above and 50% below.
- If there are an even number of observations, the median is the average of the two middle values.
- Not affected by extreme values

Example

Find the median for the following data set
$13, 3, 13, 10, 9, 11, 7, 6, 12, 7, 1, 10, 10, 5$

- Data set is not ordered, first sort the data
  $1, 3, 5, 6, 7, 7, 9, 10, 10, 10, 11, 12, 13, 13$
- There are 14 observations (even number of obs.), so the median is the average of the two middle values:

Median

Median: In an ordered data set, the median is the observation right in the middle.

- If there are an odd number of observations, the median leaves exactly 50% above and 50% below.
- If there are an even number of observations, the median is the average of the two middle values.
- Not affected by extreme values

Example

Find the median for the following data set
$13, 3, 13, 10, 9, 11, 7, 6, 12, 7, 1, 10, 10, 5$

- Data set is not ordered, first sort the data
  $1, 3, 5, 6, 7, 7, 9, 10, 10, 10, 11, 12, 13, 13$
- There are 14 observations (even number of obs.), so the median is the average of the two middle values:

$$\underbrace{1, 3, 5, 6, 7, 7,}_{6\text{ obs.}} \quad \underbrace{\underbrace{9, 10}_{(9+10)/2}}_{Median=9.5} \quad , \underbrace{10, 10, 11, 12, 13, 13}_{6\text{ obs.}}$$

Mode

Mode: is the value that occurs most often.

Mode

Mode: is the value that occurs most often.

- Not affected by extreme values

Mode

Mode: is the value that occurs most often.

- Not affected by extreme values
- Used for either numerical or categorical data

Mode

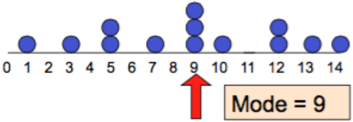
Mode: is the value that occurs most often.

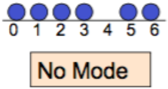
- Not affected by extreme values
- Used for either numerical or categorical data
- There may may be no mode or several modes

Mode

Mode: is the value that occurs most often.

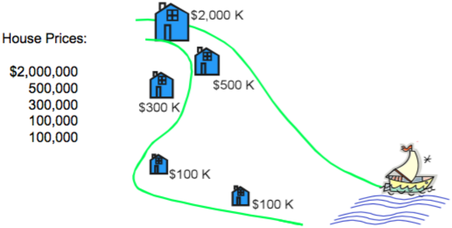
- Not affected by extreme values
- Used for either numerical or categorical data
- There may may be no mode or several modes





Review Example

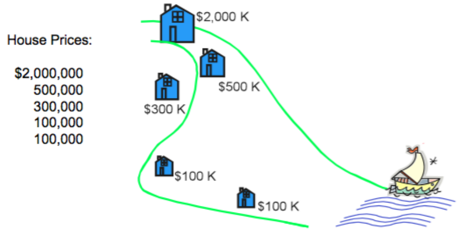
Consider the following five houses on a hill by the beach



Mean $= \$600{,}000$
Median $= \$300{,}000$
Mod $= \$100{,}000$

Review Example

Consider the following five houses on a hill by the beach

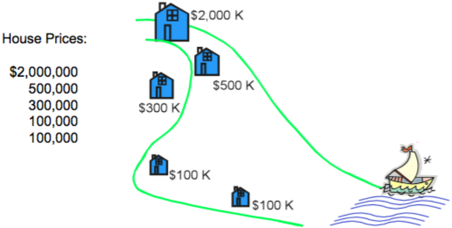


Mean = \$600,000
Median = \$300,000
Mod = \$100,000

- Median home prices should be reported since there is an outlier in the data set

Review Example

Consider the following five houses on a hill by the beach



Mean = \$600,000
Median = \$300,000
Mod = \$100,000

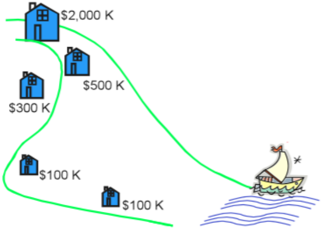
- Median home prices should be reported since there is an outlier in the data set
- Mean is generally used, unless extreme values (outliers) exist

Review Example

Consider the following five houses on a hill by the beach

House Prices:

$2,000,000
500,000
300,000
100,000
100,000

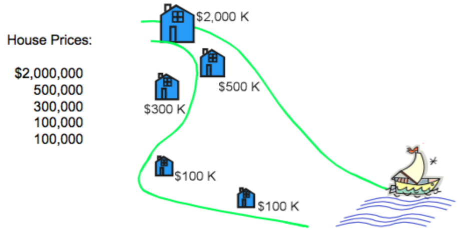


Mean $=$ \$600,000
Median $=$ \$300,000
Mod $=$ \$100,000

- Median home prices should be reported since there is an outlier in the data set
- Mean is generally used, unless extreme values (outliers) exist
- When the variable is categorical, mode has to be used. For example,

Review Example

Consider the following five houses on a hill by the beach



Mean = \$600,000
Median = \$300,000
Mod = \$100,000

- Median home prices should be reported since there is an outlier in the data set
- Mean is generally used, unless extreme values (outliers) exist
- When the variable is categorical, mode has to be used. For example,suppose 100 people will choose among three candidates $\{A, B, C\}$. In this case, only the mode is relevant measure of tendency.

Geometric Mean

Example

An investment of \$100,000 rose to \$150,000 at the end of year one and increased to \$180,000 at end of year two. What is average annual increase?

Geometric Mean

Example

An investment of \$100,000 rose to \$150,000 at the end of year one and increased to \$180,000 at end of year two. What is average annual increase?

- Temptation: use arithmetic mean?

Geometric Mean

Example

An investment of \$100,000 rose to \$150,000 at the end of year one and increased to \$180,000 at end of year two. What is average annual increase?

- Temptation: use arithmetic mean?
  $100 \underbrace{\longrightarrow}_{50\%} 150 \underbrace{\longrightarrow}_{20\%} 180 \Longrightarrow 35\% \ ?$

Geometric Mean

Example

An investment of \$100,000 rose to \$150,000 at the end of year one and increased to \$180,000 at end of year two. What is average annual increase?

- Temptation: use arithmetic mean?
  $100 \underbrace{\longrightarrow}_{50\%} 150 \underbrace{\longrightarrow}_{20\%} 180 \Longrightarrow 35\% \ ?$
- If 35% was the correct, then starting from \$100,000 and increasing the value by 35% each year we should be able to obtain \$180,000.

Geometric Mean

Example

An investment of \$100,000 rose to \$150,000 at the end of year one and increased to \$180,000 at end of year two. What is average annual increase?

- Temptation: use arithmetic mean?
  $100 \underbrace{\longrightarrow}_{50\%} 150 \underbrace{\longrightarrow}_{20\%} 180 \Longrightarrow 35\%$?
- If 35% was the correct, then starting from \$100,000 and increasing the value by 35% each year we should be able to obtain \$180,000.
- But instead we get:
  $100 \underbrace{\longrightarrow}_{35\%} 135 \underbrace{\longrightarrow}_{35\%} 182.25$

Geometric Mean

Example

An investment of \$100,000 rose to \$150,000 at the end of year one and increased to \$180,000 at end of year two. What is average annual increase?

- Temptation: use arithmetic mean?
  $100 \underbrace{\longrightarrow}_{50\%} 150 \underbrace{\longrightarrow}_{20\%} 180 \Longrightarrow 35\%$?
- If 35% was the correct, then starting from \$100,000 and increasing the value by 35% each year we should be able to obtain \$180,000.
- But instead we get:
  $100 \underbrace{\longrightarrow}_{35\%} 135 \underbrace{\longrightarrow}_{35\%} 182.25$
- So why arithmetic mean doesn't work?

Geometric Mean

Example

An investment of \$100,000 rose to \$150,000 at the end of year one and increased to \$180,000 at end of year two. What is average annual increase?

- Temptation: use arithmetic mean?
  $100 \underbrace{\longrightarrow}_{50\%} 150 \underbrace{\longrightarrow}_{20\%} 180 \Longrightarrow 35\%$?
- If 35% was the correct, then starting from \$100,000 and increasing the value by 35% each year we should be able to obtain \$180,000.
- But instead we get:
  $100 \underbrace{\longrightarrow}_{35\%} 135 \underbrace{\longrightarrow}_{35\%} 182.25$
- So why arithmetic mean doesn't work? Because it doesn't take into account the compound effect of growth!
- Use geometric mean!

Geometric Mean Formula

If growth rate is $r_t\%$ in period t (i.e., the amount from the previous period increases by $r_t\%$), where $t = 1, 2, 3, \ldots, n$, then average growth rate over n years is

$$r = \left[(1 + r_1)(1 + r_2)\ldots(1 + r_n)\right]^{1/n} - 1$$

Geometric Mean Formula

If growth rate is r_t% in period t (i.e., the amount from the previous period increases by r_t%), where $t = 1, 2, 3, \ldots, n$, then average growth rate over n years is

$$r = \left[(1 + r_1)(1 + r_2)\ldots(1 + r_n)\right]^{1/n} - 1$$

Going back to our example:

Geometric Mean Formula

If growth rate is $r_t\%$ in period t (i.e., the amount from the previous period increases by $r_t\%$), where $t = 1, 2, 3, \ldots, n$, then average growth rate over n years is

$$r = \left[(1 + r_1)(1 + r_2)\ldots(1 + r_n)\right]^{1/n} - 1$$

Going back to our example:

- $r_1 = 0.5$ and $r_2 = 0.2$

Geometric Mean Formula

If growth rate is $r_t\%$ in period t (i.e., the amount from the previous period increases by $r_t\%$), where $t = 1, 2, 3, \ldots, n$, then average growth rate over n years is

$$r = \left[(1 + r_1)(1 + r_2)\ldots(1 + r_n)\right]^{1/n} - 1$$

Going back to our example:

- $r_1 = 0.5$ and $r_2 = 0.2$
- Therefore average increase is

$$r = \left[(1 + 0.5)(1 + 0.2)\right]^{1/2} - 1 = 0.341$$

Geometric Mean Formula

If growth rate is $r_t\%$ in period t (i.e., the amount from the previous period increases by $r_t\%$), where $t = 1, 2, 3, \ldots, n$, then average growth rate over n years is

$$r = \left[(1 + r_1)(1 + r_2) \ldots (1 + r_n)\right]^{1/n} - 1$$

Going back to our example:

- $r_1 = 0.5$ and $r_2 = 0.2$
- Therefore average increase is

$$r = \left[(1 + 0.5)(1 + 0.2)\right]^{1/2} - 1 = 0.341$$

- Indeed
  $100 \underbrace{\longrightarrow}_{34.1\%} 134.1 \underbrace{\longrightarrow}_{34.1\%} 180$

What is the average growth rate of sales if sales have grown 25% over the last 5 years?

Justifying Geometric Mean Formula

- Suppose that initially we start with the quantity S.

Justifying Geometric Mean Formula

- Suppose that initially we start with the quantity S.
- On the LHS we will increase the amount from the previous period by $r\%$ in every period.

Justifying Geometric Mean Formula

- Suppose that initially we start with the quantity S.
- On the LHS we will increase the amount from the previous period by $r\%$ in every period.
- On the RHS we will increase the amount from the previous period by $r_t\%$ at time t, $t = 1, 2, \ldots, n$.

Justifying Geometric Mean Formula

- Suppose that initially we start with the quantity S.
- On the LHS we will increase the amount from the previous period by $r\%$ in every period.
- On the RHS we will increase the amount from the previous period by $r_t\%$ at time t, $t = 1, 2, \ldots, n$.

$S_0 = S$	$S_0 = S$
$S_1 = S + rS = (1 + r)S$	$S_1 = S + r_1 S = (1 + r_1)S$
$S_2 = S_1 + rS_1 = (1 + r)S_1$	$S_2 = S_1 + r_2 S_1 = (1 + r_2)S_1$
$= (1 + r)(1 + r)S$	$= (1 + r_1)(1 + r_2)S$
$= (1 + r)^2 S$	$= \ldots$
$\vdots$	$\vdots$
$S_n = (1 + r)^n S$	$S_n = (1 + r_1)(1 + r_2)\ldots(1 + r_n)S$

Justifying Geometric Mean Formula

- Suppose that initially we start with the quantity S.
- On the LHS we will increase the amount from the previous period by $r\%$ in every period.
- On the RHS we will increase the amount from the previous period by $r_t\%$ at time t, $t = 1, 2, \dots, n$.

$$\begin{aligned}
S_0 &= S \\
S_1 &= S + rS = (1+r)S \\
S_2 &= S_1 + rS_1 = (1+r)S_1 \\
&= (1+r)(1+r)S \\
&= (1+r)^2 S \\
&\vdots \\
S_n &= (1+r)^n S
\end{aligned}$$

$$\begin{aligned}
S_0 &= S \\
S_1 &= S + r_1 S = (1+r_1)S \\
S_2 &= S_1 + r_2 S_1 = (1+r_2)S_1 \\
&= (1+r_1)(1+r_2)S \\
&= \dots \\
&\vdots \\
S_n &= (1+r_1)(1+r_2)\dots(1+r_n)S
\end{aligned}$$

- From these two equations, solving for r yields the geometric mean growth rate formula.

Measures of Variability

- Measures of variation give information on the spread or variability of the data values.

Measures of Variability

- Measures of variation give information on the spread or variability of the data values.
- Simplest measure of variation is Range

Measures of Variability

- Measures of variation give information on the spread or variability of the data values.
- Simplest measure of variation is Range
- **Range =** Difference between the largest and the smallest observations:

Measures of Variability

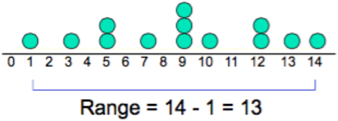
- Measures of variation give information on the spread or variability of the data values.
- Simplest measure of variation is Range
- **Range =** Difference between the largest and the smallest observations:

$$Range = X_{largest} - X_{smallest}$$

Measures of Variability

- Measures of variation give information on the spread or variability of the data values.
- Simplest measure of variation is Range
- **Range =** Difference between the largest and the smallest observations:

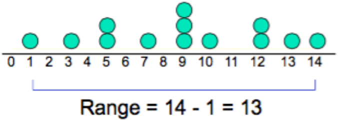
$$Range = X_{largest} - X_{smallest}$$



Measures of Variability

- Measures of variation give information on the spread or variability of the data values.
- Simplest measure of variation is Range
- **Range =** Difference between the largest and the smallest observations:

$$Range = X_{largest} - X_{smallest}$$

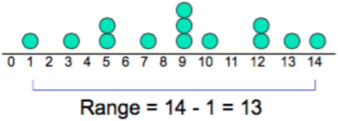


- However, range is not a good measure of variation because

Measures of Variability

- Measures of variation give information on the spread or variability of the data values.
- Simplest measure of variation is Range
- **Range =** Difference between the largest and the smallest observations:

$$Range = X_{largest} - X_{smallest}$$

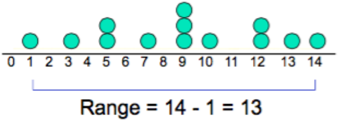


- However, range is not a good measure of variation because
  - Does not exploit the information in all data points

Measures of Variability

- Measures of variation give information on the spread or variability of the data values.
- Simplest measure of variation is Range
- **Range =** Difference between the largest and the smallest observations:

$$Range = X_{largest} - X_{smallest}$$

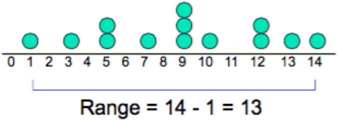


- However, range is not a good measure of variation because
  - Does not exploit the information in all data points
  - Ignores the way in which data are distributed:

Measures of Variability

- Measures of variation give information on the spread or variability of the data values.
- Simplest measure of variation is Range
- **Range =** Difference between the largest and the smallest observations:

$$Range = X_{largest} - X_{smallest}$$

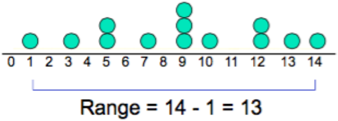


- However, range is not a good measure of variation because
  - Does not exploit the information in all data points
  - Ignores the way in which data are distributed: $\{7, 8, 9, 10, 11, 12\} \Longrightarrow$ range $= 5$

Measures of Variability

- Measures of variation give information on the spread or variability of the data values.
- Simplest measure of variation is Range
- **Range =** Difference between the largest and the smallest observations:

$$Range = X_{largest} - X_{smallest}$$

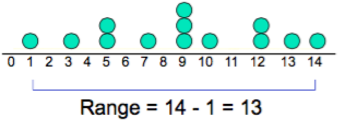


- However, range is not a good measure of variation because
  - Does not exploit the information in all data points
  - Ignores the way in which data are distributed:
    $\{7, 8, 9, 10, 11, 12\} \Longrightarrow$ range $= 5$
    $\{7, 11, 11, 12, 12, 12\} \Longrightarrow$ range $= 5$

Measures of Variability

- Measures of variation give information on the spread or variability of the data values.
- Simplest measure of variation is Range
- **Range =** Difference between the largest and the smallest observations:

$$Range = X_{largest} - X_{smallest}$$

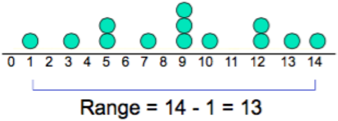


- However, range is not a good measure of variation because
  - Does not exploit the information in all data points
  - Ignores the way in which data are distributed:
    $\{7, 8, 9, 10, 11, 12\} \Longrightarrow \text{range} = 5$
    $\{7, 11, 11, 12, 12, 12\} \Longrightarrow \text{range} = 5$
  - Sensitive to outliers

Measures of Variability

- Measures of variation give information on the spread or variability of the data values.
- Simplest measure of variation is Range
- **Range =** Difference between the largest and the smallest observations:

$$Range = X_{largest} - X_{smallest}$$

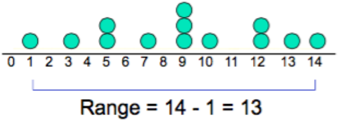


- However, range is not a good measure of variation because
  - Does not exploit the information in all data points
  - Ignores the way in which data are distributed:
    $\{7, 8, 9, 10, 11, 12\} \implies \text{range} = 5$
    $\{7, 11, 11, 12, 12, 12\} \implies \text{range} = 5$
  - Sensitive to outliers
    $\{1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 5\} \implies \text{range} = 4$

Measures of Variability

- Measures of variation give information on the spread or variability of the data values.
- Simplest measure of variation is Range
- **Range =** Difference between the largest and the smallest observations:

$$Range = X_{largest} - X_{smallest}$$

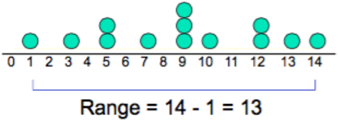


- However, range is not a good measure of variation because
  - Does not exploit the information in all data points
  - Ignores the way in which data are distributed:
    $\{7, 8, 9, 10, 11, 12\} \Longrightarrow$ range $= 5$
    $\{7, 11, 11, 12, 12, 12\} \Longrightarrow$ range $= 5$
  - Sensitive to outliers
    $\{1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 5\} \Longrightarrow$ range $= 4$
    $\{1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 120\}$

Measures of Variability

- Measures of variation give information on the spread or variability of the data values.
- Simplest measure of variation is Range
- **Range =** Difference between the largest and the smallest observations:

$$Range = X_{largest} - X_{smallest}$$



- However, range is not a good measure of variation because
  - Does not exploit the information in all data points
  - Ignores the way in which data are distributed:
    $\{7, 8, 9, 10, 11, 12\} \Longrightarrow$ range $= 5$
    $\{7, 11, 11, 12, 12, 12\} \Longrightarrow$ range $= 5$
  - Sensitive to outliers
    $\{1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 5\} \Longrightarrow$ range $= 4$
    $\{1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 120\} \Longrightarrow$ range $= 119$

Variance

Variance: Average of squared deviations of values *from the mean*:

Variance

Variance: Average of squared deviations of values *from the mean*:

$$\text{Variance} = \frac{\text{sum of squared deviations of values from the mean}}{\#\text{ of } (\textit{independent}) \text{ observations}}$$

Variance

Variance: Average of squared deviations of values *from the mean*:

$$\text{Variance} = \frac{\text{sum of squared deviations of values from the mean}}{\#\text{ of }(\textit{independent})\text{ observations}}$$

- Why *deviations*...?

Variance

Variance: Average of squared deviations of values *from the mean*:

$$\text{Variance} = \frac{\text{sum of squared deviations of values from the mean}}{\#\text{ of } (\textit{independent}) \text{ observations}}$$

- Why *deviations*...?
- Why *sum of*... ?

Variance

Variance: Average of squared deviations of values *from the mean*:

$$\text{Variance} = \frac{\text{sum of squared deviations of values from the mean}}{\#\text{ of }(\textit{independent})\text{ observations}}$$

- Why *deviations*...?
- Why *sum of*... ?
- Why *squared*... ?

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1} \qquad \text{(Sample Variance)}$$

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1} \qquad \text{(Sample Variance)}$$

where

n: sample size

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1} \qquad \text{(Sample Variance)}$$

where

n: sample size

x_i: i^{th} observation (or i^{th} value of variable) in the sample

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1} \qquad \text{(Sample Variance)}$$

where

n: sample size

x_i: i^{th} observation (or i^{th} value of variable) in the sample

$\overline{X}$: arithmetic mean, ($\overline{X} = \left(\sum_{i=1}^{n} x_i\right)/n$).

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1} \qquad \text{(Sample Variance)}$$

where

n: sample size

x_i: i^{th} observation (or i^{th} value of variable) in the sample

$\overline{X}$: arithmetic mean, $(\overline{X} = \left(\sum_{i=1}^{n} x_i\right)/n)$.

Example:

Calculate the sample variance for the following sample data:

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1} \qquad \text{(Sample Variance)}$$

where

n: sample size

x_i: i^{th} observation (or i^{th} value of variable) in the sample

$\overline{X}$: arithmetic mean, $(\overline{X} = \left(\sum_{i=1}^{n} x_i\right)/n)$.

Example:

Calculate the sample variance for the following sample data:

$$\{10, 12, 14, 15, 17, 18, 18, 24\}$$

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1} \qquad \text{(Sample Variance)}$$

where

n: sample size

x_i: i^{th} observation (or i^{th} value of variable) in the sample

$\overline{X}$: arithmetic mean, $(\overline{X} = \left(\sum_{i=1}^{n} x_i\right)/n)$.

Example:

Calculate the sample variance for the following sample data:

$$\{10, 12, 14, 15, 17, 18, 18, 24\}$$

- $n = 8$, and $\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{10+12+14+15+17+18+18+24}{8} = 16$

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1} \qquad \text{(Sample Variance)}$$

where

n: sample size

x_i: i^{th} observation (or i^{th} value of variable) in the sample

$\overline{X}$: arithmetic mean, ($\overline{X} = \left(\sum_{i=1}^{n} x_i\right)/n$).

Example:

Calculate the sample variance for the following sample data:

$$\{10, 12, 14, 15, 17, 18, 18, 24\}$$

- $n = 8$, and $\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{10+12+14+15+17+18+18+24}{8} = 16$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1} = \frac{(x_1 - \overline{X})^2 + (x_2 - \overline{X})^2 + \ldots + (x_n - \overline{X})^2}{n-1}$$

$$= \frac{(10-16)^2 + (12-16)^2 + \ldots + (24-16)^2}{8-1}$$

$$= \frac{130}{7} = 18.6$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}} \qquad \text{(Sample Standard Deviation)}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}} \qquad \text{(Sample Standard Deviation)}$$

We take square root of variance to obtain a measure of variation that has the same units as the original data.

Sample Standard Deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}} \qquad \text{(Sample Standard Deviation)}$$

We take square root of variance to obtain a measure of variation that has the same units as the original data.

Example cont'd:

Sample Standard Deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}} \qquad \text{(Sample Standard Deviation)}$$

We take square root of variance to obtain a measure of variation that has the same units as the original data.

Example cont'd:

- Notice that in the previous example, we found $s^2 = 18.6$ which is big number considering the individual data values.

Sample Standard Deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}} \qquad \text{(Sample Standard Deviation)}$$

We take square root of variance to obtain a measure of variation that has the same units as the original data.

Example cont'd:

- Notice that in the previous example, we found $s^2 = 18.6$ which is big number considering the individual data values.
- However, when we calculate the sample standard deviation we get

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}} = \sqrt{18.6} = 4.3$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}} \qquad \text{(Sample Standard Deviation)}$$

We take square root of variance to obtain a measure of variation that has the same units as the original data.

Example cont'd:

- Notice that in the previous example, we found $s^2 = 18.6$ which is big number considering the individual data values.
- However, when we calculate the sample standard deviation we get

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}} = \sqrt{18.6} = 4.3$$

which has the same units as the original data.

Sample Standard Deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}} \qquad \text{(Sample Standard Deviation)}$$

We take square root of variance to obtain a measure of variation that has the same units as the original data.

Example cont'd:

- Notice that in the previous example, we found $s^2 = 18.6$ which is big number considering the individual data values.
- However, when we calculate the sample standard deviation we get

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}} = \sqrt{18.6} = 4.3$$

  which has the same units as the original data.
- For this reason, ssd is the most commonly used measure of variation.

Comparing Sample Std Deviations

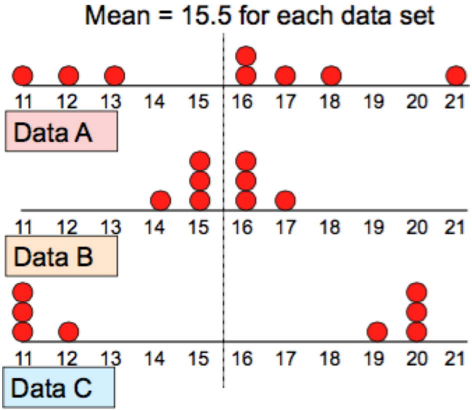
- Consider the following three data sets, A, B, and C.

Comparing Sample Std Deviations

- Consider the following three data sets, A, B, and C.
- Which one has the highest/lowest sample std deviation?

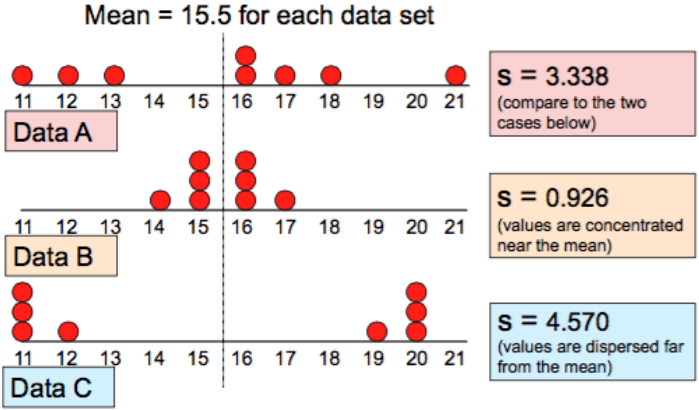
Comparing Sample Std Deviations

- Consider the following three data sets, A, B, and C.
- Which one has the highest/lowest sample std deviation?



Comparing Sample Std Deviations

- Consider the following three data sets, A, B, and C.
- Which one has the highest/lowest sample std deviation?



Population Variance and Standard Deviation

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \qquad \text{(Population Variance)}$$

Population Variance and Standard Deviation

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \qquad \text{(Population Variance)}$$

where
N: population size

Population Variance and Standard Deviation

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \qquad \text{(Population Variance)}$$

where
N: population size
x_i: i^{th} observation (or i^{th} value of variable) in the population,

Population Variance and Standard Deviation

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \qquad \text{(Population Variance)}$$

where
N: population size
x_i: i^{th} observation (or i^{th} value of variable) in the population,
μ : population mean, $(\mu = \left(\sum_{i=1}^{N} x_i\right)/N)$.

Population Variance and Standard Deviation

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \qquad \text{(Population Variance)}$$

where
N: population size
x_i: i^{th} observation (or i^{th} value of variable) in the population,
μ : population mean, $(\mu = \left(\sum_{i=1}^{N} x_i\right)/N)$.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \qquad \text{(Population Standard Deviation)}$$

Why there is (n-1) in the denominator of Sample formulas?

- The reason is theoretical $\rightarrow$ later chapters ...

Why there is (n-1) in the denominator of Sample formulas?

- The reason is theoretical $\rightarrow$ later chapters ...
- But basic intuition is the following:

Why there is (n-1) in the denominator of Sample formulas?

- The reason is theoretical $\rightarrow$ later chapters ...
- But basic intuition is the following:
- Recall that main goal of inferential statistics was to exploit sample data, which are *available*, to predict population parameters, which are (*unavailable or not observable*).

Why there is (n-1) in the denominator of Sample formulas?

- The reason is theoretical $\rightarrow$ later chapters ...
- But basic intuition is the following:
- Recall that main goal of inferential statistics was to exploit sample data, which are *available*, to predict population parameters, which are (*unavailable or not observable*).
- For example, we would like to go from

Why there is (n-1) in the denominator of Sample formulas?

- The reason is theoretical $\rightarrow$ later chapters ...
- But basic intuition is the following:
- Recall that main goal of inferential statistics was to exploit sample data, which are *available*, to predict population parameters, which are (*unavailable or not observable*).
- For example, we would like to go from
  - $\overline{X} \longrightarrow \mu$

Why there is (n-1) in the denominator of Sample formulas?

- The reason is theoretical $\rightarrow$ later chapters ...
- But basic intuition is the following:
- Recall that main goal of inferential statistics was to exploit sample data, which are *available*, to predict population parameters, which are (*unavailable or not observable*).
- For example, we would like to go from
  - $\overline{X} \longrightarrow \mu$
  - $s^2 \longrightarrow \sigma^2$

Why there is (n-1) in the denominator of Sample formulas?

- The reason is theoretical $\rightarrow$ later chapters ...
- But basic intuition is the following:
- Recall that main goal of inferential statistics was to exploit sample data, which are *available*, to predict population parameters, which are (*unavailable or not observable*).
- For example, we would like to go from
  - $\overline{X} \longrightarrow \mu$
  - $s^2 \longrightarrow \sigma^2$
- But we also want to make sure that these predictions are accurate *on average*

Why there is (n-1) in the denominator of Sample formulas?

- The reason is theoretical $\rightarrow$ later chapters ...
- But basic intuition is the following:
- Recall that main goal of inferential statistics was to exploit sample data, which are *available*, to predict population parameters, which are (*unavailable or not observable*).
- For example, we would like to go from
  - $\overline{X} \longrightarrow \mu$
  - $s^2 \longrightarrow \sigma^2$
- But we also want to make sure that these predictions are accurate *on average*
  - (n-1) in the sample variance formula assures that is the case

Why there is (n-1) in the denominator of Sample formulas?

- The reason is theoretical $\rightarrow$ later chapters ...
- But basic intuition is the following:
- Recall that main goal of inferential statistics was to exploit sample data, which are *available*, to predict population parameters, which are (*unavailable or not observable*).
- For example, we would like to go from
  - $\overline{X} \longrightarrow \mu$
  - $s^2 \longrightarrow \sigma^2$
- But we also want to make sure that these predictions are accurate *on average*
  - (n-1) in the sample variance formula assures that is the case
  - because once you fixed $\overline{X}$ there are only (n-1) independent observations

Why there is (n-1) in the denominator of Sample formulas?

- The reason is theoretical $\rightarrow$ later chapters ...
- But basic intuition is the following:
- Recall that main goal of inferential statistics was to exploit sample data, which are *available*, to predict population parameters, which are (*unavailable or not observable*).
- For example, we would like to go from
  - $\overline{X} \longrightarrow \mu$
  - $s^2 \longrightarrow \sigma^2$
- But we also want to make sure that these predictions are accurate *on average*
  - (n-1) in the sample variance formula assures that is the case
  - because once you fixed $\overline{X}$ there are only (n-1) independent observations
  - For example: suppose $n = 3$ and fix $\overline{X} = 7$.

Why there is (n-1) in the denominator of Sample formulas?

- The reason is theoretical $\rightarrow$ later chapters ...
- But basic intuition is the following:
- Recall that main goal of inferential statistics was to exploit sample data, which are *available*, to predict population parameters, which are (*unavailable or not observable*).
- For example, we would like to go from
  - $\overline{X} \longrightarrow \mu$
  - $s^2 \longrightarrow \sigma^2$
- But we also want to make sure that these predictions are accurate *on average*
  - (n-1) in the sample variance formula assures that is the case
  - because once you fixed $\overline{X}$ there are only (n-1) independent observations
  - For example: suppose $n = 3$ and fix $\overline{X} = 7$.
  - Can you determine x_1, x_2 and x_3 independently?

Why there is (n-1) in the denominator of Sample formulas?

- The reason is theoretical $\rightarrow$ later chapters ...
- But basic intuition is the following:
- Recall that main goal of inferential statistics was to exploit sample data, which are *available*, to predict population parameters, which are (*unavailable or not observable*).
- For example, we would like to go from
  - $\overline{X} \longrightarrow \mu$
  - $s^2 \longrightarrow \sigma^2$
- But we also want to make sure that these predictions are accurate *on average*
  - (n-1) in the sample variance formula assures that is the case
  - because once you fixed $\overline{X}$ there are only (n-1) independent observations
  - For example: suppose $n = 3$ and fix $\overline{X} = 7$.
  - Can you determine x_1, x_2 and x_3 independently?
  - No!

Why there is (n-1) in the denominator of Sample formulas?

- The reason is theoretical $\rightarrow$ later chapters ...
- But basic intuition is the following:
- Recall that main goal of inferential statistics was to exploit sample data, which are *available*, to predict population parameters, which are (*unavailable or not observable*).
- For example, we would like to go from
  - $\overline{X} \longrightarrow \mu$
  - $s^2 \longrightarrow \sigma^2$
- But we also want to make sure that these predictions are accurate *on average*
  - (n-1) in the sample variance formula assures that is the case
  - because once you fixed $\overline{X}$ there are only (n-1) independent observations
  - For example: suppose $n = 3$ and fix $\overline{X} = 7$.
  - Can you determine x_1, x_2 and x_3 independently?
  - No!
  - Example: if $x_1 = 6$ and $x_2 = 10$, then x_3

Why there is (n-1) in the denominator of Sample formulas?

- The reason is theoretical $\rightarrow$ later chapters ...
- But basic intuition is the following:
- Recall that main goal of inferential statistics was to exploit sample data, which are *available*, to predict population parameters, which are (*unavailable or not observable*).
- For example, we would like to go from
  - $\overline{X} \longrightarrow \mu$
  - $s^2 \longrightarrow \sigma^2$
- But we also want to make sure that these predictions are accurate *on average*
  - (n-1) in the sample variance formula assures that is the case
  - because once you fixed $\overline{X}$ there are only (n-1) independent observations
  - For example: suppose $n = 3$ and fix $\overline{X} = 7$.
  - Can you determine x_1, x_2 and x_3 independently?
  - No!
  - Example: if $x_1 = 6$ and $x_2 = 10$, then x_3 *must be* 5 in order to have $\overline{X} = 7$.

Why there is (n-1) in the denominator of Sample formulas?

- The reason is theoretical $\rightarrow$ later chapters ...
- But basic intuition is the following:
- Recall that main goal of inferential statistics was to exploit sample data, which are *available*, to predict population parameters, which are (*unavailable or not observable*).
- For example, we would like to go from
  - $\overline{X} \longrightarrow \mu$
  - $s^2 \longrightarrow \sigma^2$
- But we also want to make sure that these predictions are accurate *on average*
  - (n-1) in the sample variance formula assures that is the case
  - because once you fixed $\overline{X}$ there are only (n-1) independent observations
  - For example: suppose $n = 3$ and fix $\overline{X} = 7$.
  - Can you determine x_1, x_2 and x_3 independently?
  - No!
  - Example: if $x_1 = 6$ and $x_2 = 10$, then x_3 *must be* 5 in order to have $\overline{X} = 7$.
  - (n-1) is called **degrees of freedom (d.o.f)**.

Why there is (n-1) in the denominator of Sample formulas?

- The reason is theoretical $\rightarrow$ later chapters ...
- But basic intuition is the following:
- Recall that main goal of inferential statistics was to exploit sample data, which are *available*, to predict population parameters, which are (*unavailable or not observable*).
- For example, we would like to go from
  - $\overline{X} \longrightarrow \mu$
  - $s^2 \longrightarrow \sigma^2$
- But we also want to make sure that these predictions are accurate *on average*
  - (n-1) in the sample variance formula assures that is the case
  - because once you fixed $\overline{X}$ there are only (n-1) independent observations
  - For example: suppose $n = 3$ and fix $\overline{X} = 7$.
  - Can you determine x_1, x_2 and x_3 independently?
  - No!
  - Example: if $x_1 = 6$ and $x_2 = 10$, then x_3 *must be* 5 in order to have $\overline{X} = 7$.
  - (n-1) is called **degrees of freedom (d.o.f)**.

Generating new data from the old

- Consider **Data B** in the previous example:

Generating new data from the old

- Consider **Data B** in the previous example:
  Data B$= \{14, 15, 15, 15, 16, 16, 16, 17\}$

Generating new data from the old

- Consider **Data B** in the previous example:
  Data B$= \{14, 15, 15, 15, 16, 16, 16, 17\}$
- If we multiply each data value in the data by, say, 10 we obtain a new data set

Generating new data from the old

- Consider **Data B** in the previous example:
  Data B$= \{14, 15, 15, 15, 16, 16, 16, 17\}$
- If we multiply each data value in the data by, say, 10 we obtain a new data set
  Data B new$= \{140, 150, 150, 150, 160, 160, 160, 170\}$

Generating new data from the old

- Consider **Data B** in the previous example:
  Data B$= \{14, 15, 15, 15, 16, 16, 16, 17\}$
- If we multiply each data value in the data by, say, 10 we obtain a new data set
  Data B new$= \{140, 150, 150, 150, 160, 160, 160, 170\}$
- Let us generalize this: Suppose we have a data set X:

$$X = \{x_1, x_2, \ldots, x_n\}$$

Generating new data from the old

- Consider **Data B** in the previous example:
  Data B$= \{14, 15, 15, 15, 16, 16, 16, 17\}$
- If we multiply each data value in the data by, say, 10 we obtain a new data set
  Data B new$= \{140, 150, 150, 150, 160, 160, 160, 170\}$
- Let us generalize this: Suppose we have a data set X:

$$X = \{x_1, x_2, \ldots, x_n\}$$

  Multiplying each data point in X by a and then adding b we can generate a new data set, say Y, as:

Generating new data from the old

- Consider **Data B** in the previous example:
  Data B$= \{14, 15, 15, 15, 16, 16, 16, 17\}$
- If we multiply each data value in the data by, say, 10 we obtain a new data set
  Data B new$= \{140, 150, 150, 150, 160, 160, 160, 170\}$
- Let us generalize this: Suppose we have a data set X:

$$X = \{x_1, x_2, \ldots, x_n\}$$

Multiplying each data point in X by a and then adding b we can generate a new data set, say Y, as:

$$Y = \{ax_1 + b, ax_2 + b, \ldots, ax_n + b\}$$

Generating new data from the old

- Consider **Data B** in the previous example:
  Data B$= \{14, 15, 15, 15, 16, 16, 16, 17\}$
- If we multiply each data value in the data by, say, 10 we obtain a new data set
  Data B new$= \{140, 150, 150, 150, 160, 160, 160, 170\}$
- Let us generalize this: Suppose we have a data set X:

$$X = \{x_1, x_2, \ldots, x_n\}$$

Multiplying each data point in X by a and then adding b we can generate a new data set, say Y, as:

$$Y = \{ax_1 + b, ax_2 + b, \ldots, ax_n + b\}$$

or we could simply write

Generating new data from the old

- Consider **Data B** in the previous example:
  Data B$= \{14, 15, 15, 15, 16, 16, 16, 17\}$
- If we multiply each data value in the data by, say, 10 we obtain a new data set
  Data B new$= \{140, 150, 150, 150, 160, 160, 160, 170\}$
- Let us generalize this: Suppose we have a data set X:

$$X = \{x_1, x_2, \ldots, x_n\}$$

Multiplying each data point in X by a and then adding b we can generate a new data set, say Y, as:

$$Y = \{ax_1 + b, ax_2 + b, \ldots, ax_n + b\}$$

or we could simply write

$$Y = \{y_1, y_2, \ldots, y_n\} \quad \text{where} \quad y_i = ax_i + b$$

Generating new data from the old

- Consider **Data B** in the previous example:
  Data B$= \{14, 15, 15, 15, 16, 16, 16, 17\}$
- If we multiply each data value in the data by, say, 10 we obtain a new data set
  Data B new$= \{140, 150, 150, 150, 160, 160, 160, 170\}$
- Let us generalize this: Suppose we have a data set X:

$$X = \{x_1, x_2, \ldots, x_n\}$$

  Multiplying each data point in X by a and then adding b we can generate a new data set, say Y, as:

$$Y = \{ax_1 + b, ax_2 + b, \ldots, ax_n + b\}$$

  or we could simply write

$$Y = \{y_1, y_2, \ldots, y_n\} \quad \text{where} \quad y_i = ax_i + b$$

- This process is also called **data transformation:**

Generating new data from the old

- Consider **Data B** in the previous example:
  Data B$= \{14, 15, 15, 15, 16, 16, 16, 17\}$
- If we multiply each data value in the data by, say, 10 we obtain a new data set
  Data B new$= \{140, 150, 150, 150, 160, 160, 160, 170\}$
- Let us generalize this: Suppose we have a data set X:

$$X = \{x_1, x_2, \dots, x_n\}$$

  Multiplying each data point in X by a and then adding b we can generate a new data set, say Y, as:

$$Y = \{ax_1 + b, ax_2 + b, \dots, ax_n + b\}$$

  or we could simply write

$$Y = \{y_1, y_2, \dots, y_n\} \quad \text{where} \quad y_i = ax_i + b$$

- This process is also called **data transformation:**

$$X \longrightarrow Y$$

$$x_i \longrightarrow ax_i + b$$

Generating new data from the old

- Consider **Data B** in the previous example:
  Data B$= \{14, 15, 15, 15, 16, 16, 16, 17\}$
- If we multiply each data value in the data by, say, 10 we obtain a new data set
  Data B new$= \{140, 150, 150, 150, 160, 160, 160, 170\}$
- Let us generalize this: Suppose we have a data set X:

$$X = \{x_1, x_2, \ldots, x_n\}$$

  Multiplying each data point in X by a and then adding b we can generate a new data set, say Y, as:

$$Y = \{ax_1 + b, ax_2 + b, \ldots, ax_n + b\}$$

  or we could simply write

$$Y = \{y_1, y_2, \ldots, y_n\} \quad \text{where} \quad y_i = ax_i + b$$

- This process is also called **data transformation:**

$$X \longrightarrow Y$$

$$x_i \longrightarrow ax_i + b$$

Comparing Std Deviations cont'd

- Consider again the data sets A, B, and C.

Comparing Std Deviations cont'd

- Consider again the data sets A, B, and C.
- Suppose we would like to generate three new data sets through the following transformations:

Comparing Std Deviations cont'd

- Consider again the data sets A, B, and C.
- Suppose we would like to generate three new data sets through the following transformations:

$$A \longrightarrow A_{new}$$

Comparing Std Deviations cont'd

- Consider again the data sets A, B, and C.
- Suppose we would like to generate three new data sets through the following transformations:

$$A \longrightarrow A_{new}$$
$$x_i \longrightarrow x_i$$

Comparing Std Deviations cont'd

- Consider again the data sets A, B, and C.
- Suppose we would like to generate three new data sets through the following transformations:

$A \longrightarrow A_{new}$

$x_i \longrightarrow x_i$

$(a=1, b=0)$

Comparing Std Deviations cont'd

- Consider again the data sets A, B, and C.
- Suppose we would like to generate three new data sets through the following transformations:

$$A \longrightarrow A_{new}$$
$$x_i \longrightarrow x_i$$
$$(a{=}1, b{=}0)$$

Comparing Std Deviations cont'd

- Consider again the data sets A, B, and C.
- Suppose we would like to generate three new data sets through the following transformations:

$A \longrightarrow A_{new}$ $\qquad$ $B \longrightarrow B_{new}$

$x_i \longrightarrow x_i$

$(a=1, b=0)$

Comparing Std Deviations cont'd

- Consider again the data sets A, B, and C.
- Suppose we would like to generate three new data sets through the following transformations:

$A \longrightarrow A_{new}$	$B \longrightarrow B_{new}$
$x_i \longrightarrow x_i$	$x_i \longrightarrow 10x_i$
$(a=1,b=0)$	

Comparing Std Deviations cont'd

- Consider again the data sets A, B, and C.
- Suppose we would like to generate three new data sets through the following transformations:

$A \longrightarrow A_{new}$	$B \longrightarrow B_{new}$
$x_i \longrightarrow x_i$	$x_i \longrightarrow 10x_i$
$(a{=}1, b{=}0)$	$(a{=}10, b{=}0)$

Comparing Std Deviations cont'd

- Consider again the data sets A, B, and C.
- Suppose we would like to generate three new data sets through the following transformations:

$A \longrightarrow A_{new}$	$B \longrightarrow B_{new}$	$C \longrightarrow C_{new}$
$x_i \longrightarrow x_i$	$x_i \longrightarrow 10x_i$	
$(a{=}1, b{=}0)$	$(a{=}10, b{=}0)$	

Comparing Std Deviations cont'd

- Consider again the data sets A, B, and C.
- Suppose we would like to generate three new data sets through the following transformations:

$A \longrightarrow A_{new}$	$B \longrightarrow B_{new}$	$C \longrightarrow C_{new}$
$x_i \longrightarrow x_i$	$x_i \longrightarrow 10x_i$	$x_i \longrightarrow x_i + 10$
$(a=1, b=0)$	$(a=10, b=0)$	

Comparing Std Deviations cont'd

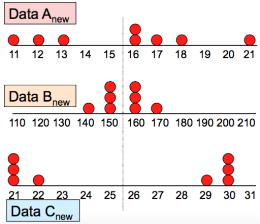
- Consider again the data sets A, B, and C.
- Suppose we would like to generate three new data sets through the following transformations:

$A \longrightarrow A_{new}$	$B \longrightarrow B_{new}$	$C \longrightarrow C_{new}$
$x_i \longrightarrow x_i$	$x_i \longrightarrow 10x_i$	$x_i \longrightarrow x_i + 10$
$(a{=}1, b{=}0)$	$(a{=}10, b{=}0)$	$(a{=}1, b{=}10)$

Comparing Std Deviations cont'd

- Consider again the data sets A, B, and C.
- Suppose we would like to generate three new data sets through the following transformations:

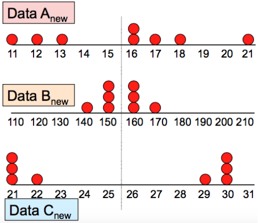
$A \longrightarrow A_{new}$	$B \longrightarrow B_{new}$	$C \longrightarrow C_{new}$
$x_i \longrightarrow x_i$	$x_i \longrightarrow 10x_i$	$x_i \longrightarrow x_i + 10$
$(a{=}1, b{=}0)$	$(a{=}10, b{=}0)$	$(a{=}1, b{=}10)$



Comparing Std Deviations cont'd

- Consider again the data sets A, B, and C.
- Suppose we would like to generate three new data sets through the following transformations:

$A \longrightarrow A_{new}$	$B \longrightarrow B_{new}$	$C \longrightarrow C_{new}$
$x_i \longrightarrow x_i$	$x_i \longrightarrow 10x_i$	$x_i \longrightarrow x_i + 10$
$(a{=}1, b{=}0)$	$(a{=}10, b{=}0)$	$(a{=}1, b{=}10)$



$s_{new} = 3.33$

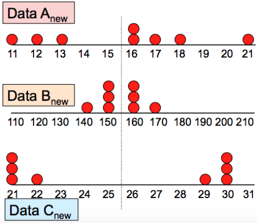
$s_{new} = 9.2$ (!)

$s_{new} = 4.57$

Comparing Std Deviations cont'd

- Consider again the data sets A, B, and C.
- Suppose we would like to generate three new data sets through the following transformations:

$A \longrightarrow A_{new}$	$B \longrightarrow B_{new}$	$C \longrightarrow C_{new}$
$x_i \longrightarrow x_i$	$x_i \longrightarrow 10x_i$	$x_i \longrightarrow x_i + 10$
$(a=1, b=0)$	$(a=10, b=0)$	$(a=1, b=10)$



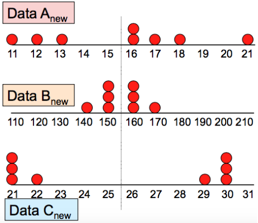
$s_{new} = 3.33$	s = 3.338 (compare to the two cases below)
$s_{new} = 9.2$ (!)	s = 0.926 (values are concentrated near the mean)
$s_{new} = 4.57$	s = 4.570 (values are dispersed far from the mean)

- Only std of data B has changed!
  - **Scaling effects the standard deviation!!!**

Comparing Std Deviations cont'd

- Consider again the data sets A, B, and C.
- Suppose we would like to generate three new data sets through the following transformations:

$A \longrightarrow A_{new}$	$B \longrightarrow B_{new}$	$C \longrightarrow C_{new}$
$x_i \longrightarrow x_i$	$x_i \longrightarrow 10x_i$	$x_i \longrightarrow x_i + 10$
$(a=1, b=0)$	$(a=10, b=0)$	$(a=1, b=10)$



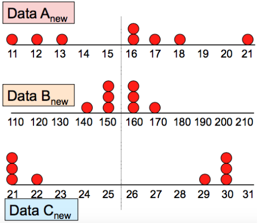
$s_{new} = 3.33$	s = 3.338 (compare to the two cases below)
$s_{new} = 9.2$ (!)	s = 0.926 (values are concentrated near the mean)
$s_{new} = 4.57$	s = 4.570 (values are dispersed far from the mean)

- Only std of data B has changed!
  - **Scaling effects the standard deviation!!!**
  - Isn't it true that Data B still *relatively* the least dispersed?

Comparing Std Deviations cont'd

- Consider again the data sets A, B, and C.
- Suppose we would like to generate three new data sets through the following transformations:

$A \longrightarrow A_{new}$	$B \longrightarrow B_{new}$	$C \longrightarrow C_{new}$
$x_i \longrightarrow x_i$	$x_i \longrightarrow 10x_i$	$x_i \longrightarrow x_i + 10$
$(a{=}1, b{=}0)$	$(a{=}10, b{=}0)$	$(a{=}1, b{=}10)$



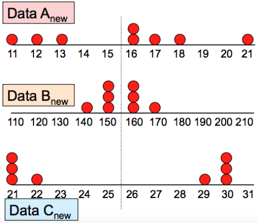
$s_{new} = 3.33$	s = 3.338 (compare to the two cases below)
$s_{new} = 9.2$ (!)	s = 0.926 (values are concentrated near the mean)
$s_{new} = 4.57$	s = 4.570 (values are dispersed far from the mean)

- Only std of data B has changed!
  - **Scaling effects the standard deviation!!!**
  - Isn't it true that Data B still *relatively* the least dispersed?
  - But it has the highest standard deviation!

Comparing Std Deviations cont'd

- Consider again the data sets A, B, and C.
- Suppose we would like to generate three new data sets through the following transformations:

$A \longrightarrow A_{new}$	$B \longrightarrow B_{new}$	$C \longrightarrow C_{new}$
$x_i \longrightarrow x_i$	$x_i \longrightarrow 10x_i$	$x_i \longrightarrow x_i + 10$
$(a{=}1, b{=}0)$	$(a{=}10, b{=}0)$	$(a{=}1, b{=}10)$



$s_{new} = 3.33$ | s = 3.338 (compare to the two cases below)

$s_{new} = 9.2$ (!) | s = 0.926 (values are concentrated near the mean)

$s_{new} = 4.57$ | s = 4.570 (values are dispersed far from the mean)

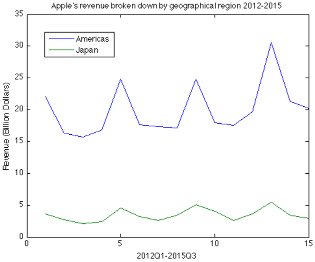
- Only std of data B has changed!
  - **Scaling effects the standard deviation!!!**
  - Isn't it true that Data B still *relatively* the least dispersed?
  - But it has the highest standard deviation!

Comparing Std: A time-series example

- Apple's revenue broken down by geographical region from 2012-Q1 to 2015-Q3 is depicted in the following figure

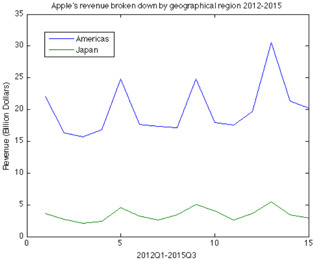
Comparing Std: A time-series example

- Apple's revenue broken down by geographical region from 2012-Q1 to 2015-Q3 is depicted in the following figure



Comparing Std: A time-series example

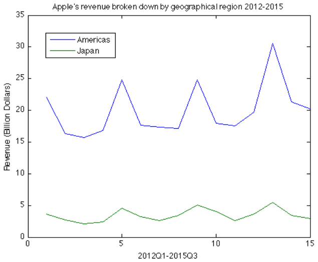
- Apple's revenue broken down by geographical region from 2012-Q1 to 2015-Q3 is depicted in the following figure



- Is Apple's revenue more volatile in Americas or in Japan?

Comparing Std: A time-series example

- Apple's revenue broken down by geographical region from 2012-Q1 to 2015-Q3 is depicted in the following figure

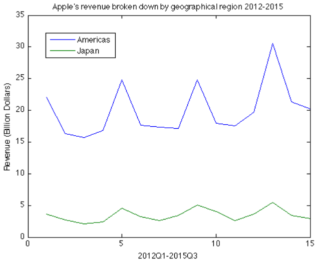


- Is Apple's revenue more volatile in Americas or in Japan?

$$s_{Amer.} = 4.10$$
$$s_{Jap.} = 0.98$$

Comparing Std: A time-series example

- Apple's revenue broken down by geographical region from 2012-Q1 to 2015-Q3 is depicted in the following figure



- Is Apple's revenue more volatile in Americas or in Japan?

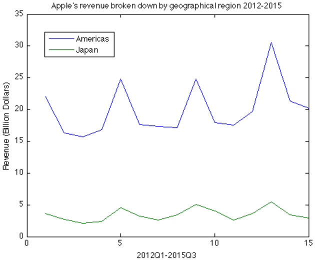
$$s_{Amer.} = 4.10 \qquad \overline{X}_{Amer.} = 20.2 \qquad \frac{s_{Amer.}}{\overline{X}_{Amer.}} \times 100 = 20\%$$

$$s_{Jap.} = 0.98 \qquad \overline{X}_{Jap.} = 3.5 \qquad \frac{s_{Jap.}}{\overline{X}_{Jap.}} \times 100 = 28\%$$

- Since we care about *relative volatility* Apple's revenue in Japan is more volatile than in Americas.

Comparing Std: A time-series example

- Apple's revenue broken down by geographical region from 2012-Q1 to 2015-Q3 is depicted in the following figure



- Is Apple's revenue more volatile in Americas or in Japan?

$$s_{Amer.} = 4.10 \qquad \overline{X}_{Amer.} = 20.2 \qquad \frac{s_{Amer.}}{\overline{X}_{Amer.}} \times 100 = 20\%$$

$$s_{Jap.} = 0.98 \qquad \overline{X}_{Jap.} = 3.5 \qquad \frac{s_{Jap.}}{\overline{X}_{Jap.}} \times 100 = 28\%$$

- Since we care about *relative volatility* Apple's revenue in Japan is more volatile than in Americas.
- Next, we will give a special name to this ratio.

Coefficient of Variation

Sample coefficient of variation:

$$CV = \left(\frac{s}{\overline{X}}\right) \times 100\%$$

Coefficient of Variation

Sample coefficient of variation:

$$CV = \left(\frac{s}{\overline{X}}\right) \times 100\%$$

Population coefficient of variation:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100\%$$

Coefficient of Variation

Sample coefficient of variation:

$$CV = (\frac{s}{\overline{X}}) \times 100\%$$

- Measures relative variation

Population coefficient of variation:

$$CV = (\frac{\sigma}{\mu}) \times 100\%$$

Coefficient of Variation

Sample coefficient of variation:

$$CV = \left(\frac{s}{\overline{X}}\right) \times 100\%$$

- Measures relative variation
- Always in percentage (%)

Population coefficient of variation:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100\%$$

Coefficient of Variation

Sample coefficient of variation:

$$CV = \left(\frac{s}{\overline{X}}\right) \times 100\%$$

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean

Population coefficient of variation:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100\%$$

Coefficient of Variation

Sample coefficient of variation:

$$CV = \left(\frac{s}{\overline{X}}\right) \times 100\%$$

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare two or more sets of data measured in different units

Population coefficient of variation:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100\%$$

Coefficient of Variation

Sample coefficient of variation:

$$CV = \left(\frac{s}{\overline{X}}\right) \times 100\%$$

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare two or more sets of data measured in different units

Population coefficient of variation:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100\%$$

Example:

Coefficient of Variation

Sample coefficient of variation:

$$CV = \left(\frac{s}{\overline{\overline{X}}}\right) \times 100\%$$

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare two or more sets of data measured in different units

Population coefficient of variation:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100\%$$

Example:

Stock A

Coefficient of Variation

Sample coefficient of variation:

$$CV = \left(\frac{s}{\overline{X}}\right) \times 100\%$$

Population coefficient of variation:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100\%$$

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare two or more sets of data measured in different units

Example:

Stock A

- Average price last year = \$50

Coefficient of Variation

Sample coefficient of variation:

$$CV = \left(\frac{s}{\overline{X}}\right) \times 100\%$$

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare two or more sets of data measured in different units

Population coefficient of variation:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100\%$$

Example:

Stock A

- Average price last year $= \$50$
- Standard deviation $= \$5$

Coefficient of Variation

Sample coefficient of variation:

$$CV = \left(\frac{s}{\overline{X}}\right) \times 100\%$$

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare two or more sets of data measured in different units

Population coefficient of variation:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100\%$$

Example:

Stock A

- Average price last year = \$50
- Standard deviation = \$5

$$\Longrightarrow CV_A = \left(\frac{s_A}{\overline{X}_A}\right) \times 100\%$$

$$= \frac{\$5}{\$50} 100\% = 10\%$$

Coefficient of Variation

Sample coefficient of variation:

$$CV = \left(\frac{s}{\overline{X}}\right) \times 100\%$$

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare two or more sets of data measured in different units

Population coefficient of variation:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100\%$$

Example:

Stock A

- Average price last year = \$50
- Standard deviation = \$5

$$\Longrightarrow CV_A = \left(\frac{s_A}{\overline{X}_A}\right) \times 100\%$$

$$= \frac{\$5}{\$50} 100\% = 10\%$$

Coefficient of Variation

Sample coefficient of variation:

$$CV = \left(\frac{s}{\overline{X}}\right) \times 100\%$$

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare two or more sets of data measured in different units

Population coefficient of variation:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100\%$$

Example:

Stock A

- Average price last year = \$50
- Standard deviation = \$5

$$\Longrightarrow CV_A = \left(\frac{s_A}{\overline{X}_A}\right) \times 100\%$$

$$= \frac{\$5}{\$50} 100\% = 10\%$$

Stock B

Coefficient of Variation

Sample coefficient of variation:

$$CV = \left(\frac{s}{\overline{X}}\right) \times 100\%$$

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare two or more sets of data measured in different units

Population coefficient of variation:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100\%$$

Example:

Stock A

- Average price last year = \$50
- Standard deviation = \$5

$\Longrightarrow CV_A = \left(\frac{s_A}{\overline{X}_A}\right) \times 100\%$

$= \frac{\$5}{\$50} 100\% = 10\%$

Stock B

- Average price last year=\$100

Coefficient of Variation

Sample coefficient of variation:

$$CV = \left(\frac{s}{\overline{X}}\right) \times 100\%$$

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare two or more sets of data measured in different units

Population coefficient of variation:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100\%$$

Example:

Stock A

- Average price last year = \$50
- Standard deviation = \$5

$\Longrightarrow CV_A = \left(\frac{s_A}{\overline{X}_A}\right) \times 100\%$

$= \frac{\$5}{\$50} 100\% = 10\%$

Stock B

- Average price last year=\$100
- Standard deviation = \$5

Coefficient of Variation

Sample coefficient of variation:

$$CV = \left(\frac{s}{\overline{X}}\right) \times 100\%$$

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare two or more sets of data measured in different units

Population coefficient of variation:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100\%$$

Example:

Stock A

- Average price last year = \$50
- Standard deviation = \$5
  $\Longrightarrow CV_A = \left(\frac{s_A}{\overline{X}_A}\right) \times 100\%$
  $= \frac{\$5}{\$50} 100\% = 10\%$

Stock B

- Average price last year=\$100
- Standard deviation = \$5
  $\Longrightarrow CV_B = \left(\frac{s_B}{\overline{X}_B}\right) \times 100\%$
  $= \frac{\$5}{\$100} 100\% = 5\%$

Coefficient of Variation

Sample coefficient of variation:

$$CV = \left(\frac{s}{\overline{X}}\right) \times 100\%$$

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare two or more sets of data measured in different units

Population coefficient of variation:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100\%$$

Example:

Stock A

- Average price last year = \$50
- Standard deviation = \$5 $\Longrightarrow CV_A = \left(\frac{s_A}{\overline{X}_A}\right) \times 100\% = \frac{\$5}{\$50} 100\% = 10\%$

Stock B

- Average price last year=\$100
- Standard deviation = \$5 $\Longrightarrow CV_B = \left(\frac{s_B}{\overline{X}_B}\right) \times 100\% = \frac{\$5}{\$100} 100\% = 5\%$

Coefficient of Variation

Sample coefficient of variation:

$$CV = \left(\frac{s}{\overline{X}}\right) \times 100\%$$

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare two or more sets of data measured in different units

Population coefficient of variation:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100\%$$

Example:

Stock A

- Average price last year = \$50
- Standard deviation = \$5 $\Longrightarrow CV_A = \left(\frac{s_A}{\overline{X}_A}\right) \times 100\% = \frac{\$5}{\$50} 100\% = 10\%$

Stock B

- Average price last year=\$100
- Standard deviation = \$5 $\Longrightarrow CV_B = \left(\frac{s_B}{\overline{X}_B}\right) \times 100\% = \frac{\$5}{\$100} 100\% = 5\%$

They have the same standard deviation but Stock A is relatively more volatile.

Measures of Relationships Between Two Variables

- So far we have dealt with only sample statistics based on a single variable.

Measures of Relationships Between Two Variables

- So far we have dealt with only sample statistics based on a single variable.
- But we are also often interested in the relationship between two variables.

Measures of Relationships Between Two Variables

- So far we have dealt with only sample statistics based on a single variable.
- But we are also often interested in the relationship between two variables.
- For example, consider the following two scatter diagrams for two different pairs of variables of interest, per state in US

Measures of Relationships Between Two Variables

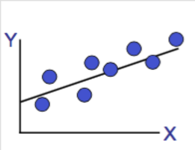
- So far we have dealt with only sample statistics based on a single variable.
- But we are also often interested in the relationship between two variables.
- For example, consider the following two scatter diagrams for two different pairs of variables of interest, per state in US

SAT Math vs. Verbal Scores
(in 2014)

Measures of Relationships Between Two Variables

- So far we have dealt with only sample statistics based on a single variable.
- But we are also often interested in the relationship between two variables.
- For example, consider the following two scatter diagrams for two different pairs of variables of interest, per state in US

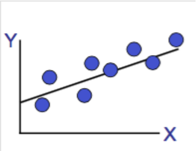
SAT Math vs. Verbal Scores
(in 2014)



Measures of Relationships Between Two Variables

- So far we have dealt with only sample statistics based on a single variable.
- But we are also often interested in the relationship between two variables.
- For example, consider the following two scatter diagrams for two different pairs of variables of interest, per state in US

SAT Math vs. Verbal Scores *(in 2014)*

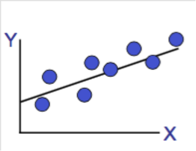


# of cars vs. # of accidents *(both measured per-person)*

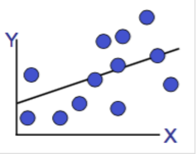
Measures of Relationships Between Two Variables

- So far we have dealt with only sample statistics based on a single variable.
- But we are also often interested in the relationship between two variables.
- For example, consider the following two scatter diagrams for two different pairs of variables of interest, per state in US

SAT Math vs. Verbal Scores *(in 2014)*



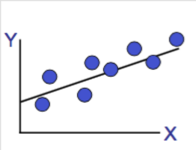
# of cars vs. # of accidents *(both measured per-person)*



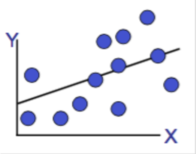
Measures of Relationships Between Two Variables

- So far we have dealt with only sample statistics based on a single variable.
- But we are also often interested in the relationship between two variables.
- For example, consider the following two scatter diagrams for two different pairs of variables of interest, per state in US

SAT Math vs. Verbal Scores *(in 2014)*



# of cars vs. # of accidents *(both measured per-person)*



- Which of these two pairs of variables shows a stronger relationship?

Covariance and Correlation Coefficient

- Two measures of the relationship between variables

Covariance and Correlation Coefficient

- Two measures of the relationship between variables
- **Covariance:** measures the *direction* of a linear relationship between two variables.

Covariance and Correlation Coefficient

- Two measures of the relationship between variables
- **Covariance:** measures the *direction* of a linear relationship between two variables.
- **Correlation Coefficient:** measures *both the direction and the strength* of a *linear* relationship between two variables.

Covariance and Correlation Coefficient

- Two measures of the relationship between variables
- **Covariance:** measures the *direction* of a linear relationship between two variables.
- **Correlation Coefficient:** measures *both the direction and the strength* of a *linear* relationship between two variables.

Covariance

- **Sample Covariance:**

Covariance

- **Sample Covariance:**

$$Cov(X, Y) = s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{n-1}$$

Covariance

- **Sample Covariance:**

$$Cov(X,Y) = s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{n-1}$$

- **Population Covariance:**

Covariance

- **Sample Covariance:**

$$Cov(X, Y) = s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{n-1}$$

- **Population Covariance:**

$$Cov(X, Y) = \sigma_{XY} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{N}$$

Covariance

- **Sample Covariance:**

$$Cov(X,Y) = s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{n-1}$$

- **Population Covariance:**

$$Cov(X,Y) = \sigma_{XY} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{N}$$

- Only concerned with the direction of the relationship

Covariance

- **Sample Covariance:**

$$Cov(X, Y) = s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{n-1}$$

- **Population Covariance:**

$$Cov(X, Y) = \sigma_{XY} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{N}$$

- Only concerned with the direction of the relationship
  - $Cov(X, Y) > 0 \Longrightarrow$ X and Y tend to move in the same direction

Covariance

- **Sample Covariance:**

$$Cov(X, Y) = s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{n-1}$$

- **Population Covariance:**

$$Cov(X, Y) = \sigma_{XY} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{N}$$

- Only concerned with the direction of the relationship
  - $Cov(X, Y) > 0 \implies$ X and Y tend to move in the same direction
  - $Cov(X, Y) < 0 \implies$ X and Y tend to move in the opposite direction

Covariance

- **Sample Covariance:**

$$Cov(X,Y) = s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{n-1}$$

- **Population Covariance:**

$$Cov(X,Y) = \sigma_{XY} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{N}$$

- Only concerned with the direction of the relationship
    - $Cov(X,Y) > 0 \implies$ X and Y tend to move in the same direction
    - $Cov(X,Y) < 0 \implies$ X and Y tend to move in the opposite direction
    - $Cov(X,Y) = 0 \implies$ there is no linear relationship between X and Y.

Covariance

- **Sample Covariance:**

$$Cov(X, Y) = s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{n-1}$$

- **Population Covariance:**

$$Cov(X, Y) = \sigma_{XY} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{N}$$

- Only concerned with the direction of the relationship
  - $Cov(X, Y) > 0 \implies$ X and Y tend to move in the same direction
  - $Cov(X, Y) < 0 \implies$ X and Y tend to move in the opposite direction
  - $Cov(X, Y) = 0 \implies$ there is no linear relationship between X and Y. Why?

Covariance

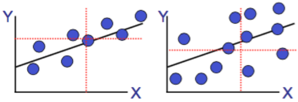
- **Sample Covariance:**

$$Cov(X,Y) = s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{n-1}$$

- **Population Covariance:**

$$Cov(X,Y) = \sigma_{XY} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{N}$$

- Only concerned with the direction of the relationship
  - $Cov(X,Y) > 0 \implies$ X and Y tend to move in the same direction
  - $Cov(X,Y) < 0 \implies$ X and Y tend to move in the opposite direction
  - $Cov(X,Y) = 0 \implies$ there is no linear relationship between X and Y. Why?



Covariance

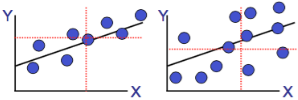
- **Sample Covariance:**

$$Cov(X,Y) = s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{n-1}$$

- **Population Covariance:**

$$Cov(X,Y) = \sigma_{XY} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{N}$$

- Only concerned with the direction of the relationship
  - $Cov(X,Y) > 0 \Longrightarrow$ X and Y tend to move in the same direction
  - $Cov(X,Y) < 0 \Longrightarrow$ X and Y tend to move in the opposite direction
  - $Cov(X,Y) = 0 \Longrightarrow$ there is no linear relationship between X and Y. Why?



- No causal effect is implied

Covariance

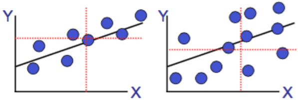
- **Sample Covariance:**

$$Cov(X, Y) = s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{n-1}$$

- **Population Covariance:**

$$Cov(X, Y) = \sigma_{XY} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{N}$$

- Only concerned with the direction of the relationship
  - $Cov(X, Y) > 0 \implies$ X and Y tend to move in the same direction
  - $Cov(X, Y) < 0 \implies$ X and Y tend to move in the opposite direction
  - $Cov(X, Y) = 0 \implies$ there is no linear relationship between X and Y. Why?



- No causal effect is implied
- Unit sensitive $\implies$ does NOT tell anything about the strength of the relationship

Covariance

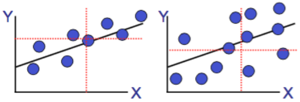
- **Sample Covariance:**

$$Cov(X,Y) = s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{n-1}$$

- **Population Covariance:**

$$Cov(X,Y) = \sigma_{XY} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{N}$$

- Only concerned with the direction of the relationship
  - $Cov(X,Y) > 0 \implies$ X and Y tend to move in the same direction
  - $Cov(X,Y) < 0 \implies$ X and Y tend to move in the opposite direction
  - $Cov(X,Y) = 0 \implies$ there is no linear relationship between X and Y. Why?



- No causal effect is implied
- Unit sensitive $\implies$ does NOT tell anything about the strength of the relationship

Coefficient of Correlation

Measures the direction and relative strength of the linear relationship between two variables

- **Sample correlation coefficient:**

Coefficient of Correlation

Measures the direction and relative strength of the linear relationship between two variables

- **Sample correlation coefficient:**

$$r_{XY} = \frac{Cov(X, Y)}{s_x s_Y}$$

Coefficient of Correlation

Measures the direction and relative strength of the linear relationship between two variables

- **Sample correlation coefficient:**

$$r_{XY} = \frac{Cov(X, Y)}{s_x s_Y}$$

- **Population correlation coefficient:**

Coefficient of Correlation

Measures the direction and relative strength of the linear relationship between two variables

- **Sample correlation coefficient:**

$$r_{XY} = \frac{Cov(X, Y)}{s_X s_Y}$$

- **Population correlation coefficient:**

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Coefficient of Correlation

Measures the direction and relative strength of the linear relationship between two variables

- **Sample correlation coefficient:**

$$r_{XY} = \frac{Cov(X, Y)}{s_x s_Y}$$

- **Population correlation coefficient:**

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_x \sigma_Y}$$

- Unit free

Coefficient of Correlation

Measures the direction and relative strength of the linear relationship between two variables

- **Sample correlation coefficient:**

$$r_{XY} = \frac{Cov(X,Y)}{s_x s_Y}$$

- **Population correlation coefficient:**

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_x \sigma_Y}$$

- Unit free
- Ranges between -1 and 1

Coefficient of Correlation

Measures the direction and relative strength of the linear relationship between two variables

- **Sample correlation coefficient:**

$$r_{XY} = \frac{Cov(X, Y)}{s_x s_Y}$$

- **Population correlation coefficient:**

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_x \sigma_Y}$$

- Unit free
- Ranges between -1 and 1
- The closer to -1, the stronger the negative linear relationship

Coefficient of Correlation

Measures the direction and relative strength of the linear relationship between two variables

- **Sample correlation coefficient:**

$$r_{XY} = \frac{Cov(X,Y)}{s_X s_Y}$$

- **Population correlation coefficient:**

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- Unit free
- Ranges between -1 and 1
- The closer to -1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship

Coefficient of Correlation

Measures the direction and relative strength of the linear relationship between two variables

- **Sample correlation coefficient:**

$$r_{XY} = \frac{Cov(X, Y)}{s_x s_Y}$$

- **Population correlation coefficient:**

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_x \sigma_Y}$$

- Unit free
- Ranges between -1 and 1
- The closer to -1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any linear relationship

Coefficient of Correlation

Measures the direction and relative strength of the linear relationship between two variables

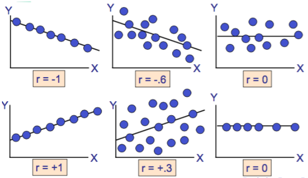
- **Sample correlation coefficient:**

$$r_{XY} = \frac{Cov(X,Y)}{s_X s_Y}$$

- **Population correlation coefficient:**

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- Unit free
- Ranges between -1 and 1
- The closer to -1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any linear relationship



Interpreting the Result

Example

- Suppose that we obtained the scatter plot of 10 students on the first test and second test as

Interpreting the Result

Example

- Suppose that we obtained the scatter plot of 10 students on the first test and second test as

Interpreting the Result

Example

- Suppose that we obtained the scatter plot of 10 students on the first test and second test as

- Furthermore, suppose that we calculated the sample correlation coefficient as $r = 0.73$.

Interpreting the Result

Example

- Suppose that we obtained the scatter plot of 10 students on the first test and second test as

- Furthermore, suppose that we calculated the sample correlation coefficient as $r = 0.73$.
  - There is a relatively strong positive linear relationship between test score #1 and test score #2.

Interpreting the Result

Example

- Suppose that we obtained the scatter plot of 10 students on the first test and second test as

- Furthermore, suppose that we calculated the sample correlation coefficient as $r = 0.73$.
  - There is a relatively strong positive linear relationship between test score #1 and test score #2.
  - Students who scored high on the first test tended to score high on second test