

LECTURE 5

DESCRIPTIVE STATISTICS: MEASURES OF VARIABILITY

PSY2002

Hye Won Suk

LIMITATIONS OF CENTRAL TENDENCY MEASURES

- A central tendency measure is a single value (a center) that is most typical or most representative of the entire set of scores.
- Central tendency measures effectively summarize the scores in a set.
- However, central tendency measures might not capture all the important features of the scores.

LIMITATIONS OF CENTRAL TENDENCY MEASURES

- Let's consider the following two samples.

- Sample 1:
3, 3, 3, 3, 3

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} =$$

- Sample 2:
0, 1, 4, 5, 5

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} =$$

Although the two samples have the same mean, the two samples have very different characteristics. In sample 1, all the scores are identical (no variation). On the other hand, in sample 2 the scores vary across individuals.

VARIABILITY

- Variability (변동성/변산성) describes the spread of scores in a distribution.
- While a central tendency measure indicates the location of the center of a distribution, a variability measure indicates the difference (or distance) of the scores in the distribution from each other.
- A larger variability indicates that the scores are more spread out and more variable across individuals.
- A variability measure also indicates how well a central tendency measure represents the entire distribution. A smaller value indicates that the central tendency measure represents the entire distribution better.

VARIABILITY MEASURES

- We will discuss three different variability measures:
 - Range
 - Variance
 - Standard Deviation

I. RANGE

- Definition of the range
 - The range (범위) is the difference between the largest and the smallest scores.
 - What is the range of the following scores?

5, 6, 1, 8, 12, 3, 4

- Max =
- Min =
- Range = Max – Min =

I. RANGE

- The range is a crude measure of variability. It does not consider the distribution of scores between the two extremes.

<u>Set 1:</u> 1, 1, 1, 1, 5, 5, 5, 5, 5, 5	<u>Set 2:</u> 1, 1, 2, 2, 3, 3, 4, 4, 5, 5
<ul style="list-style-type: none">Max =Min =Range = Max – Min =	<ul style="list-style-type: none">Max =Min =Range = Max – Min =

2.VARIANCE

- The variance (분산) is one of the most commonly used measures of variability.
- The variance is the average squared distance from the mean.
- The definition of the variance is the same for both populations and samples, but the calculations (formulas) may differ.

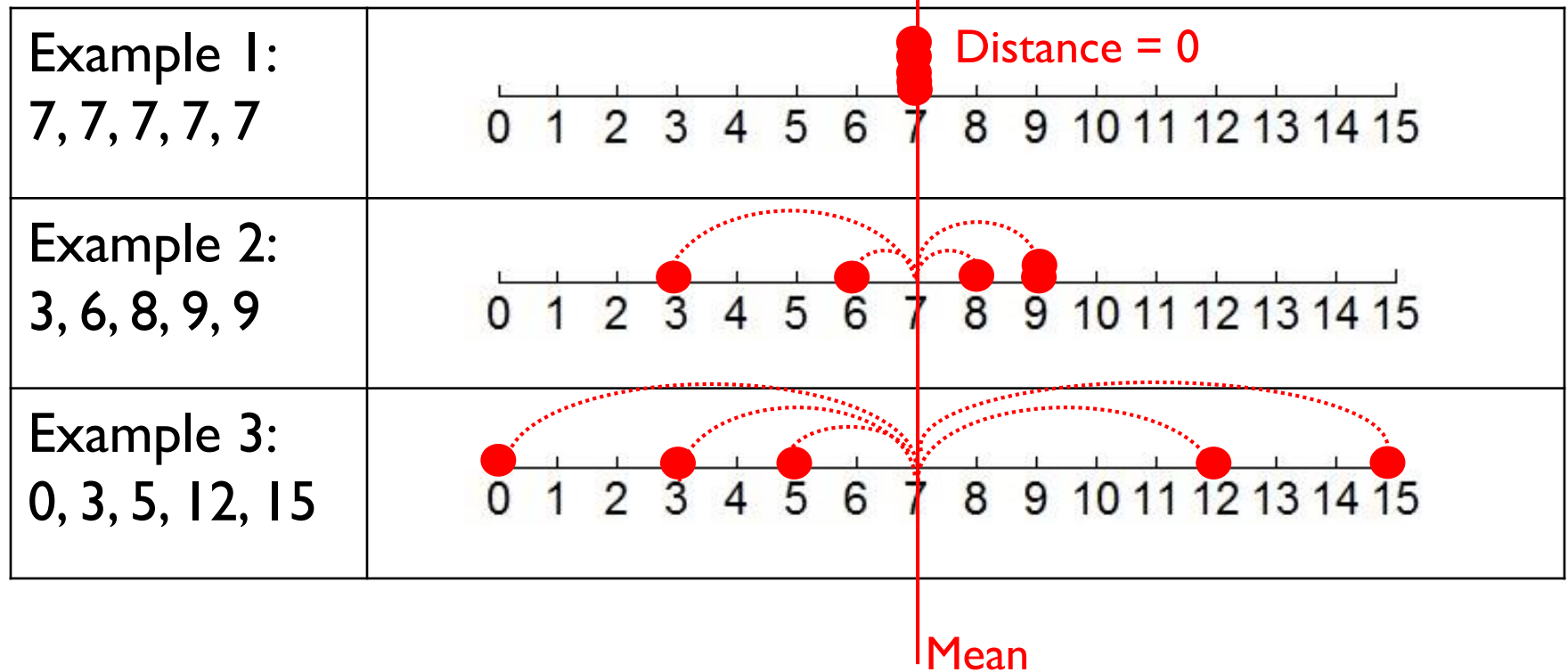
2.VARIANCE

- The variance is based on the idea of using the distance (거리) from each score in the distribution to the center (i.e., mean) of the distribution.

Example 1: 7, 7, 7, 7, 7	$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{7 + 7 + 7 + 7 + 7}{5} = \frac{35}{5} = 7$
Example 2: 3, 6, 8, 9, 9	$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{3 + 6 + 8 + 9 + 9}{5} = \frac{35}{5} = 7$
Example 3: 0, 3, 5, 12, 15	$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{0 + 3 + 5 + 12 + 15}{5} = \frac{35}{5} = 7$

2.VARIANCE

- The variance is based on the idea of using the distance (거리) from each score in the distribution to the center (i.e., mean) of the distribution.



2.VARIANCE

- Population variance

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

- Sample variance

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

2-1. POPULATION VARIANCE

- Let's assume that a researcher measured a population of 6 persons and obtained the following data.
- Find the population variance.

X
3
6
7
10
0
4

2-1. POPULATION VARIANCE

- Step 1: Obtain the population mean.

X
3
6
7
10
0
4

$$\mu = \frac{\sum_{i=1}^N X_i}{N}$$

=

2-1. POPULATION VARIANCE

- Step 2: Obtain the deviation scores ($X - \mu$).

X	$X - \mu$
3	
6	
7	
10	
0	
4	

- Deviation scores indicate how much each score deviates from the mean.
- So an aggregation of the deviation scores will provide a number that measures how much variability exists in the entire set of scores.
- What about the sum of the deviation scores?

2-1. POPULATION VARIANCE

- Step 2: Obtain the deviation scores ($X - \mu$).

X	$X - \mu$
3	$3 - 5 = -2$
6	$6 - 5 = 1$
7	$7 - 5 = 2$
10	$10 - 5 = 5$
0	$0 - 5 = -5$
4	$4 - 5 = -1$

- We already learned that the sum of the deviation scores is always 0.

$$\sum_{i=1}^N (X_i - \mu) =$$

- The sum of the deviation scores is not a good way to aggregate the deviation scores so as to measure variability.

2-1. POPULATION VARIANCE

- Step 3: Obtain the squared deviation scores $(X - \mu)^2$.

X	$X - \mu$	$(X - \mu)^2$
3	$3 - 5 = -2$	
6	$6 - 5 = 1$	
7	$7 - 5 = 2$	
10	$10 - 5 = 5$	
0	$0 - 5 = -5$	
4	$4 - 5 = -1$	

2-1. POPULATION VARIANCE

- Step 4: Calculate the sum of squared deviations (SS).

X	$X - \mu$	$(X - \mu)^2$
3	$3 - 5 = -2$	4
6	$6 - 5 = 1$	1
7	$7 - 5 = 2$	4
10	$10 - 5 = 5$	25
0	$0 - 5 = -5$	25
4	$4 - 5 = -1$	1

$$SS = \sum_{i=1}^N (X_i - \mu)^2 =$$

2-1. POPULATION VARIANCE

- Step 5: Obtain the population variance.

$$\sigma^2 = \frac{SS}{N} = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N} =$$

- σ^2 is read “sigma squared” and indicates the population variance.
- The population variance is the sum of squared deviations divided by the number of observations in the population.
- That is, the population variance is the average squared deviation.

2-2. SAMPLE VARIANCE

- Let's assume that a researcher measured a sample of 5 persons and obtained the following data.
- Find the sample variance.

X
2
6
8
9
5

2-2. SAMPLE VARIANCE

- Step 1: Obtain the sample mean.

X
2
6
8
9
5

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

=

2-2. SAMPLE VARIANCE

- Step 2: Obtain the deviation scores ($X - \bar{X}$).

X	$X - \bar{X}$
2	
6	
8	
9	
5	

2-2. SAMPLE VARIANCE

- Step 3: Obtain the squared deviation scores $(X - \bar{X})^2$.

X	$X - \bar{X}$	$(X - \bar{X})^2$
2	$2 - 6 = -4$	
6	$6 - 6 = 0$	
8	$8 - 6 = 2$	
9	$9 - 6 = 3$	
5	$5 - 6 = -1$	

2-2. SAMPLE VARIANCE

- Step 4: Calculate the sum of squared deviations (SS).

X	$X - \bar{X}$	$(X - \bar{X})^2$
2	$2 - 6 = -4$	16
6	$6 - 6 = 0$	0
8	$8 - 6 = 2$	4
9	$9 - 6 = 3$	9
5	$5 - 6 = -1$	1

$$SS = \sum_{i=1}^n (X_i - \bar{X})^2 =$$

2-2. SAMPLE VARIANCE

- Step 5: Obtain the sample variance.

$$s^2 = \frac{SS}{n} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} =$$

- The sample variance is the sum of squared deviations divided by the number of observations in the sample.
- That is, the sample variance is the average squared deviation.

INTERPRETATION OF VARIANCE

- The variance (for a population or a sample) indicates the average squared distance (or deviation) of the scores from the mean.
- For example, if the variance is 6, it means that, on average, the squared distance of the scores from the mean is 6 squared units (e.g., squared centimeters, squared kilograms, squared points, etc.).
- A larger variance indicates a more variability in the scores (if the score are measured in the same unit).

PROPERTIES OF VARIANCE

- We will discuss important properties of the variance. In the following slides, we will consider the *sample* variance calculated using s^2 only.
- However, these properties will hold for both population variance and sample variance.

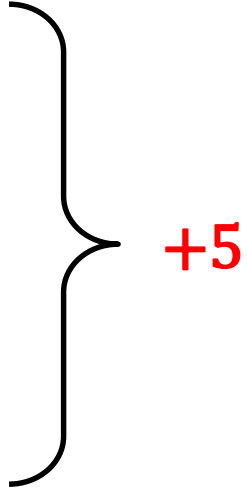
PROPERTIES OF VARIANCE

- I. Adding a constant to each and every score does not affect the variance (for a population or a sample).
- Let's consider the same example that we used for the sample variance.

2, 6, 8, 9, 5

- Sample mean: $\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{2+6+8+9+5}{5} = \frac{30}{5} = 6$
- Sample variance: $s^2 = \frac{SS}{n} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} = \frac{30}{5} = 6$

- Let's assume that we add a constant 5 to each and every score for some reason.

- Original scores: 2, 6, 8, 9, 5
 - New scores: 7, 11, 13, 14, 10
- 

- The mean of the newly obtained scores will increase by 5.

- Original scores: 2, 6, 8, 9, 5

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{2 + 6 + 8 + 9 + 5}{5} = \frac{30}{5} = 6$$

- New scores: 7, 11, 13, 14, 10

$$\overline{X + 5} = \frac{\sum_{i=1}^n (X_i + 5)}{n} =$$

- However, the deviation scores and thus SS do not change.

- Original scores: 2, 6, 8, 9, 5

$$\begin{aligned}SS &= \sum_{i=1}^n (X_i - \bar{X})^2 \\&= (2 - 6)^2 + (6 - 6)^2 + (8 - 6)^2 + (9 - 6)^2 + (5 - 6)^2 \\&= (-4)^2 + 0^2 + 2^2 + 3^2 + (-1)^2 = 16 + 0 + 4 + 9 + 1 = 30\end{aligned}$$

- New scores: 7, 11, 13, 14, 10

$$\begin{aligned}SS &= \sum_{i=1}^n (X_i + 5 - \overline{X + 5})^2 \\&= \end{aligned}$$

- Therefore, the variance does not change.

- Original scores: 2, 6, 8, 9, 5

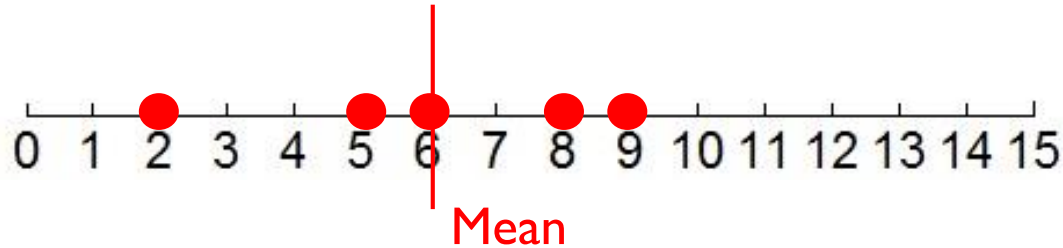
$$s^2 = \frac{SS}{n} = \frac{30}{5} = 6$$

- New scores: 7, 11, 13, 14, 10

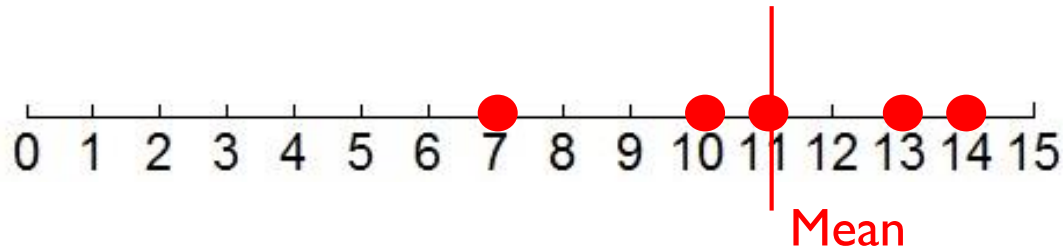
$$s^2 = \frac{SS}{n} =$$

GRAPHICAL REPRESENTATION

- Original scores:



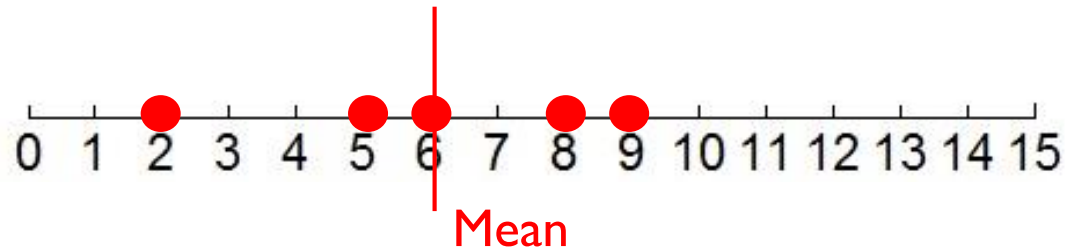
- New scores:



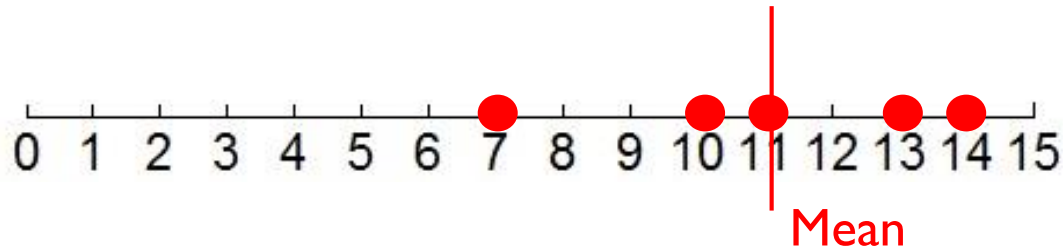
- The scores and the mean became bigger (increased by 5).

GRAPHICAL REPRESENTATION

- Original scores:



- New scores:



- The scores and the mean became bigger (increased by 5).
- However, the variability (how much the scores vary around the mean) did not change.

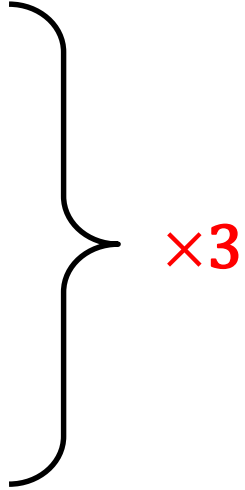
PROPERTIES OF VARIANCE

2. Multiplying each and every score by a constant (c) changes the variance by c^2 times.
- Let's consider the same example again that we used for the sample variance.

2, 6, 8, 9, 5

- Sample mean: $\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{2+6+8+9+5}{5} = \frac{30}{5} = 6$
- Sample variance: $s^2 = \frac{SS}{n} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} = \frac{30}{5} = 6$

- Let's assume that we multiply each and every score by 3 for some reason.

- Original scores: 2, 6, 8, 9, 5
 - New scores: 6, 18, 24, 27, 15
- 
- $\times 3$

- The mean of the newly obtained scores is tripled.

- Original scores: 2, 6, 8, 9, 5

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{2 + 6 + 8 + 9 + 5}{5} = \frac{30}{5} = 6$$

- New scores: 6, 18, 24, 27, 15

$$\overline{3X} = \frac{\sum_{i=1}^n 3X_i}{n} =$$

- The deviation scores are tripled and thus SS is 9 times bigger.

- Original scores: 2, 6, 8, 9, 5

$$\begin{aligned}SS &= \sum_{i=1}^n (X_i - \bar{X})^2 \\&= (2 - 6)^2 + (6 - 6)^2 + (8 - 6)^2 + (9 - 6)^2 + (5 - 6)^2 \\&= (-4)^2 + 0^2 + 2^2 + 3^2 + (-1)^2 = 16 + 0 + 4 + 9 + 1 = 30\end{aligned}$$

- New scores: 6, 18, 24, 27, 15

$$\begin{aligned}SS &= \sum_{i=1}^n (3X_i - \overline{3X})^2 \\&= \end{aligned}$$

- Therefore, the variance is also 9 times bigger.

- Original scores: 2, 6, 8, 9, 5

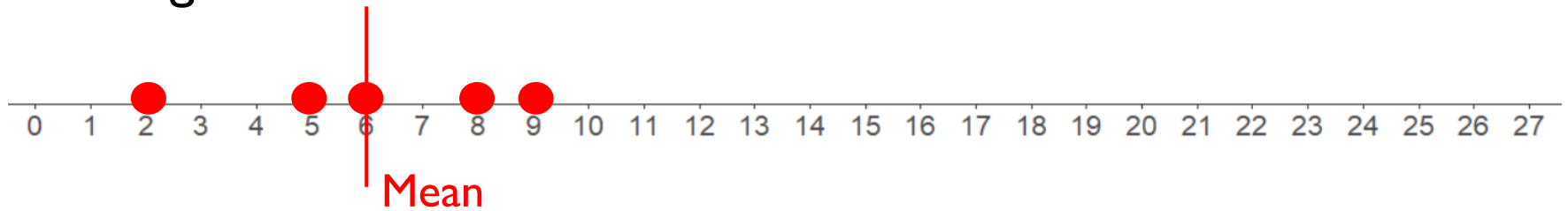
$$s^2 = \frac{SS}{n} = \frac{30}{5} = 6$$

- New scores: 6, 18, 24, 27, 15

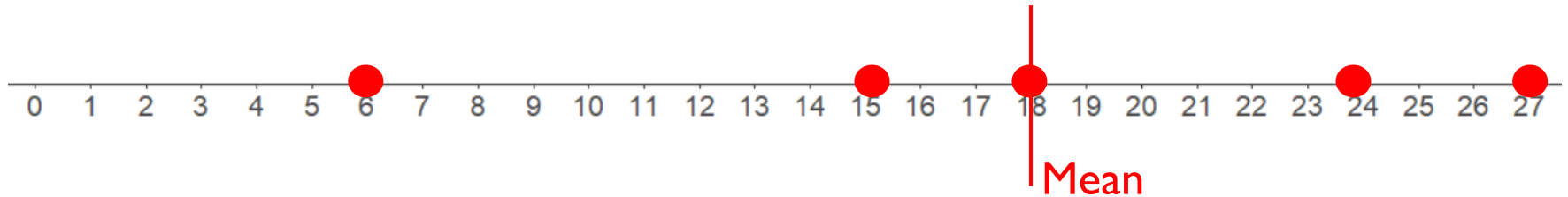
$$s^2 = \frac{SS}{n} =$$

GRAPHICAL REPRESENTATION

- Original scores:



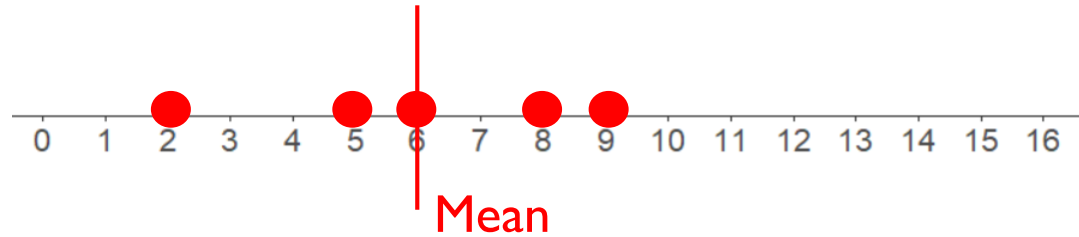
- New scores:



- The scores and the mean are tripled.

GRAPHICAL REPRESENTATION

- Original scores:



- New scores:



- The scores and the mean are tripled.
- The variance (how much the scores vary around the mean in squared units) is 9 times as large as the original one.

POP QUIZ

- The variance of midterm scores of a class is 15. The professor of the class transformed the original scores as follows:

$$X_i \longrightarrow 4X_i - 20$$

Find the variance of the transformed scores.

AN IMPLICATION

- This property of variance implies that comparing two variances obtained from the scores measured in different units makes no sense.
 - 2, 6, 8, 9, 5 (measured in cm)
 - Variance = 6
 - 20, 60, 80, 90, 50 (measured in mm)
 - Variance = 600
 - Can we conclude that the second set of data (in mm) have more variability (100 times larger variability) than the first set of data (in cm)?

3. STANDARD DEVIATION

- Variances are widely used to describe the variability of the scores. However, the concept of squared distance (or squared unit) might not be intuitive.
- The standard deviation (표준 편차) is the square root of the variance, i.e., the square root of the mean squared deviation.
- It provides a measure of variability in the distance (not squared distance) in the original unit.

3. STANDARD DEVIATION

- Population standard deviation (σ is read “sigma.”)

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}}$$

- Sample standard deviation

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}}$$

EXAMPLES

- For a set of scores, the variance is 25. Find the value of the standard deviation for this set of scores.
- For a set of scores, the standard deviation is 4. Find the value of the variance for this set of scores.

INTERPRETATION OF STANDARD DEVIATION

- The standard deviation (for a population or a sample) roughly indicates the average distance (or deviation) of the scores from the mean.
- For example, if the standard deviation is 3, it means that, on average, the scores are distant from the mean by 3 units (e.g., centimeters, kilograms, points, etc.).
- A larger standard deviation indicates a more variability in the scores (if the score are measured in the same unit).

PROPERTIES OF STANDARD DEVIATION

1. Adding a constant to each and every score does not affect the **standard deviation** (for a population or a sample).
2. Multiplying each and every score by a constant (c) changes the **standard deviation** by c times.

POP QUIZ

- A researcher measured students' height in centimeters and obtained the standard deviation of 12. Later he found that there was a systematic error in measurement, and the measured height was taller than the actual height by 3 cm. The researcher wants to correct the measurement error and change the measurement unit to millimeters by transforming the original scores as follows:

$$X_i \longrightarrow 10X_i - 30$$

Find the standard deviation of the transformed scores.

POP QUIZ

- It is possible to obtain a negative range?
- It is possible to obtain a negative variance?
- It is possible to obtain a negative standard deviation?

TWO VERSIONS OF SAMPLE VARIANCE

- In fact, there are two different formulas for sample variance.

- n formula

$$s_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

- (n-1) formula

$$s_{n-1}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

- Why do we have two different formulas for sample variance?

TERMINOLOGY (I)

- To answer this question, let's go beyond descriptive statistics and take a brief look at inferential statistics.
- Estimation (추정)
 - Estimation is a process of using the value of a statistic derived from a sample to make an educated guess about the value of a corresponding population parameter.
 - For example, we estimate the population mean (μ) by using the sample mean (\bar{X}).

TERMINOLOGY (2)

- We need to distinguish the following two terms in estimation.
 - Estimator vs. estimate
 - An estimator is a statistic (a rule or formula) that you apply to data in order to obtain the estimate.
 - An estimate the end product (a specific value) of an estimator on data.
 - For example, the sample mean $\left(\frac{\sum_{i=1}^n X_i}{n}\right)$ is an estimator of the population mean. $\bar{X} = 35$ is an estimate of the population mean derived from a sample.

PROPERTIES OF A GOOD ESTIMATOR

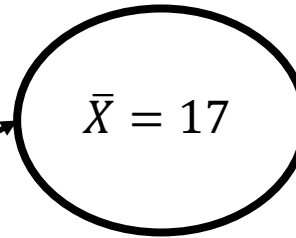
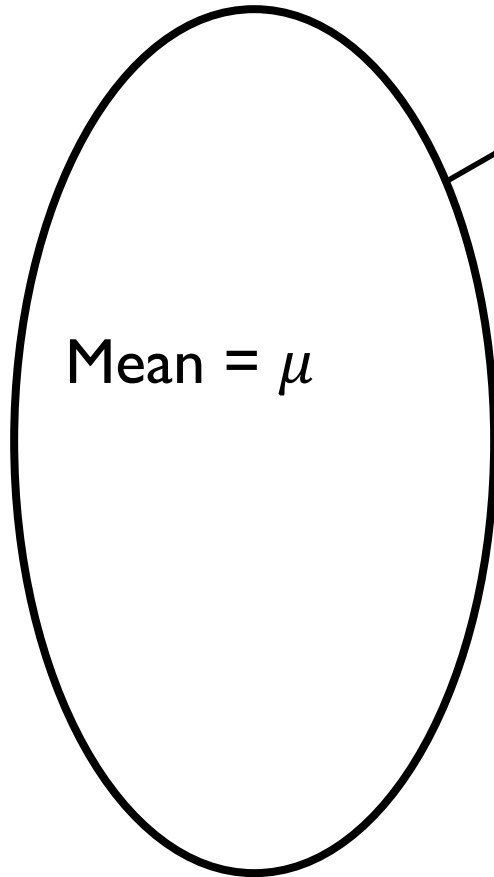
- How do we know that an estimator is a good estimator?
- For example, when we estimate the population mean, how do we know that the sample mean is a better estimator than the sample median, sample mode, or any other potential estimators?

PROPERTIES OF A GOOD ESTIMATOR

- There are several properties that a good estimator should have.
- One of the properties is called ‘unbiasedness.’ That is, an unbiased estimator is considered a good estimator.
- What is an unbiased estimator? Let’s consider the following example in which we estimate the population mean using the sample mean.

Sample

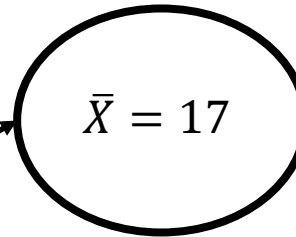
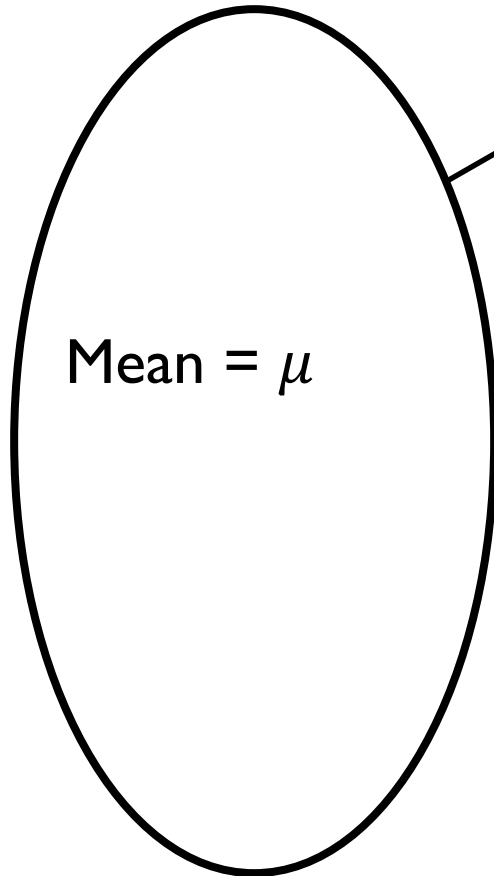
Population



- Let's say that we collect a sample of size 100 from the population.
- The sample mean is 17.
- We estimate the population mean by using the sample mean and say that the estimate of the population mean is 17.

Sample

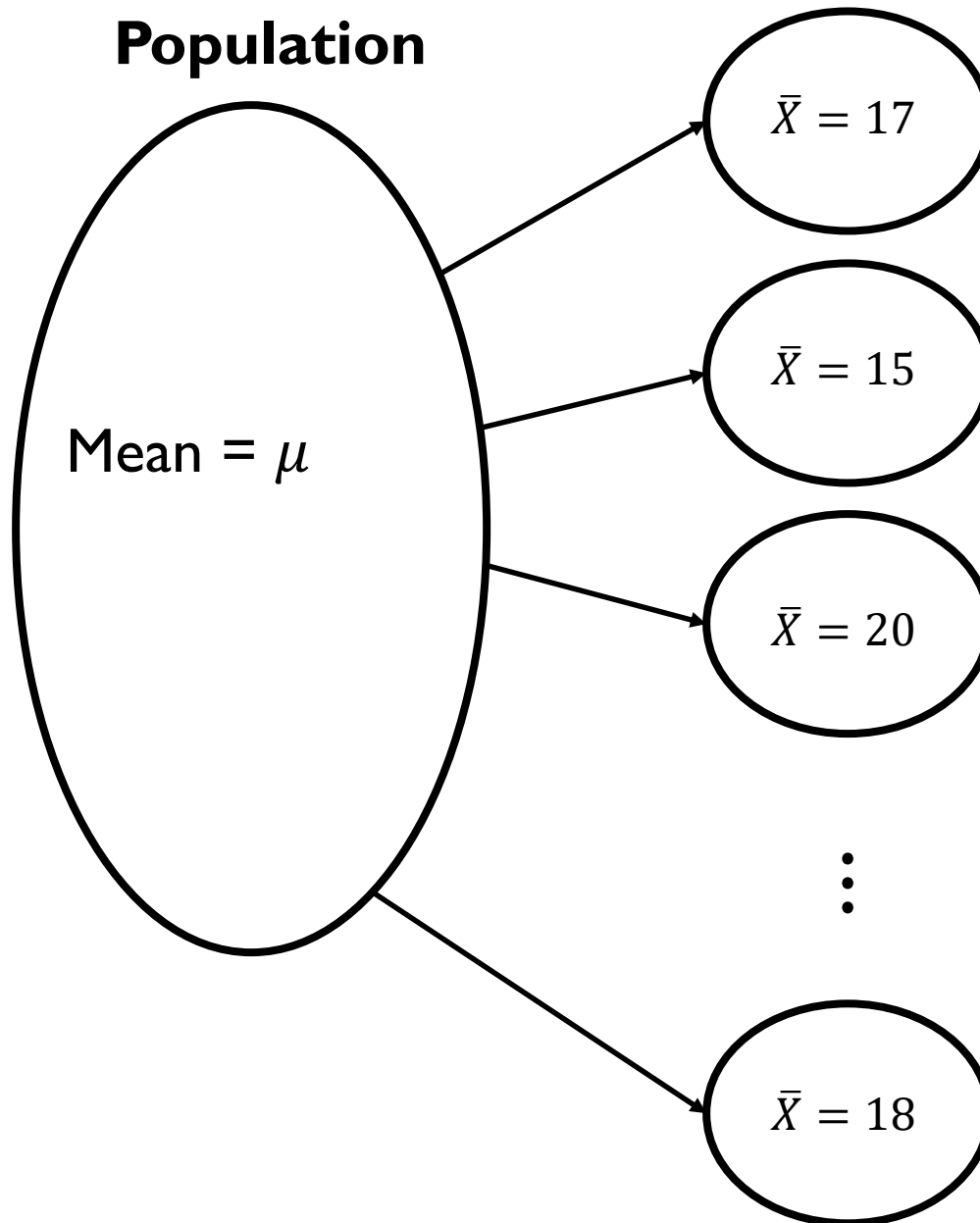
Population



- However, this estimate (=17) will be different from the population mean.
- This estimate (=17) is from a single sample out of all other potential samples that we could have obtained.
- How do we know that this is a good estimate?

Sample (of size n)

Population



- If we were able to collect all different possible samples **of the same size** from the population, it has been proven that the mean of the sample means equals the population mean.

UNBIASED ESTIMATOR

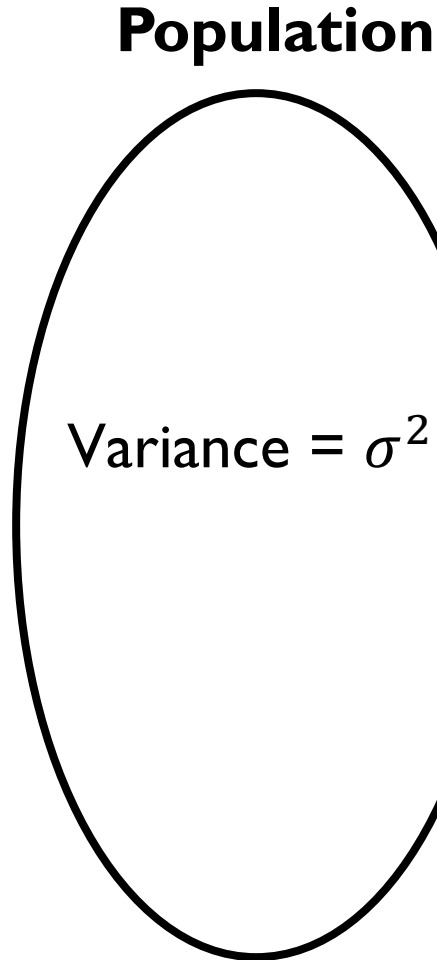
- Like in the previous example, when the population mean of a sample statistic equals the population parameter that it aims to estimate, it is called an unbiased estimator (불편향 추정치).
- In statistics, the population mean is also called the expected value and denoted as follows.
 - $E(X)$: Expected value of a variable X
 - $E(\bar{X})$: Expected value of the estimator \bar{X}
- Therefore, an unbiased estimator satisfies:
 - $E(\text{Estimator}) = \text{Population parameter}$
 - That is, an unbiased estimator equals the population parameter **on average**.

UNBIASED ESTIMATOR

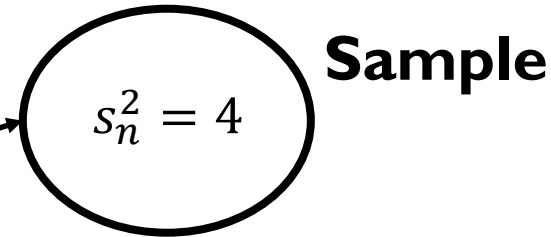
- The sample mean (\bar{X}) satisfies the following property:
 - $E(\bar{X}) = \mu$
 - Therefore, the sample mean is an unbiased estimator of the population mean and considered a good estimator.
 - This is why people use the sample mean to estimate the population mean.

BIASED ESTIMATOR

- When the population mean of the sample statistic does not equal the population parameter that it aims to estimate, it is called a biased estimator (편향 추정치).
- In general, a biased estimator is not considered a good estimator (even though there are some exceptions).
- It is known that s_n^2 is a biased estimator of the population variance. Let's see the following slides.



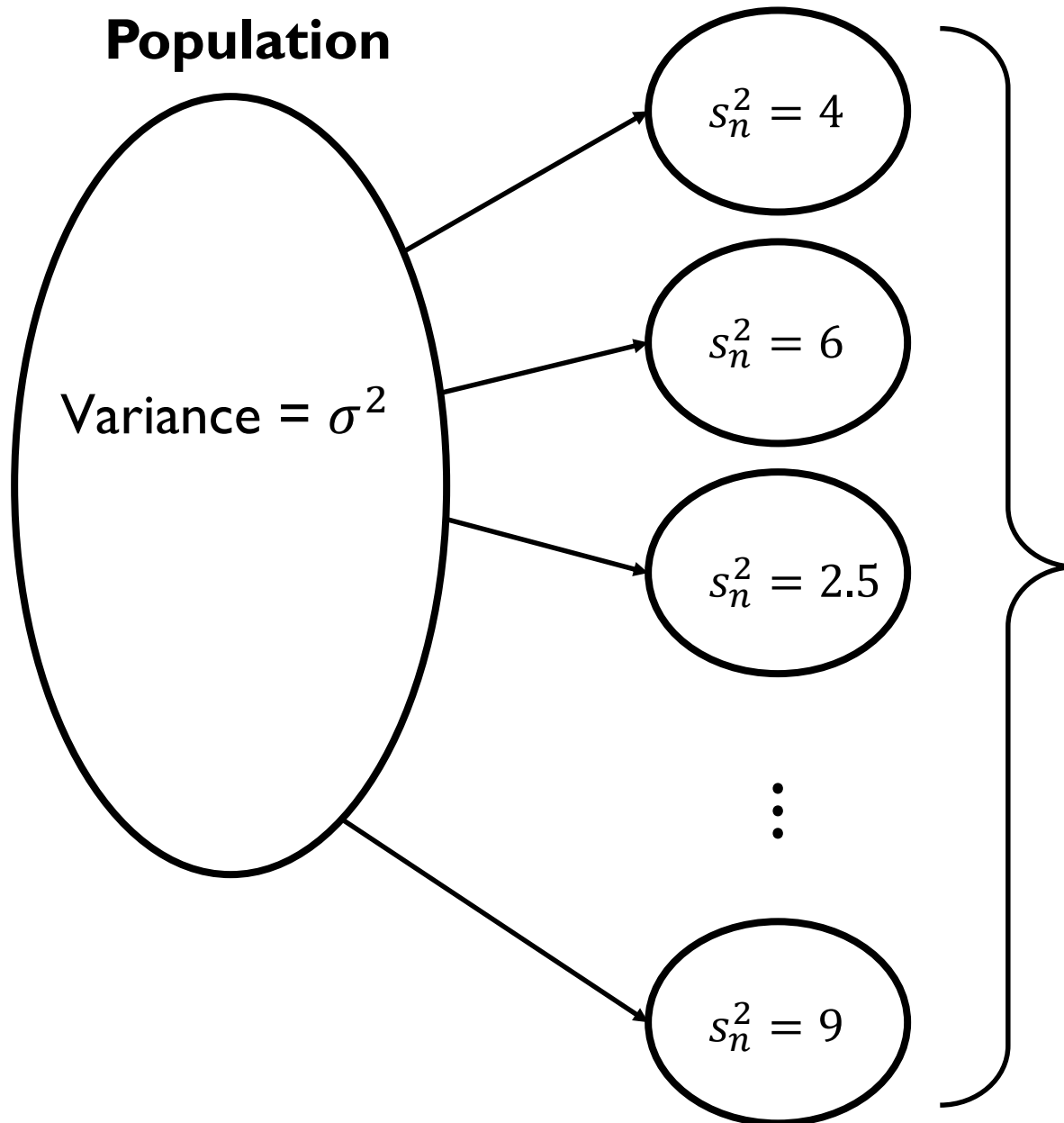
- Now, let's assume that we want to estimate the population variance (σ^2) using the sample variance (s_n^2).



- Let's say that we collect a sample of size 100 from the population and the sample variance is 4.
- We estimate the population variance by using the sample variance and say that the estimate of the population variance is 4.
- In fact, the estimate is biased.

Sample (of size n)

Population



- If we were able to collect all different possible samples of the same size from the population, it has been proven that the mean of the sample variances (s_n^2) does not equal the population variance.

TWO DIFFERENT VERSIONS

- n formula

$$s_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

- $E(s_n^2) \neq \sigma^2$
- On average, the estimator (s_n^2) does not equal the population variance.
- It is biased, and not a good estimator of the population variance.

TWO DIFFERENT VERSIONS

- (n-1) formula

$$s_{n-1}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

- $E(s_{n-1}^2) = \sigma^2$
- On average, the estimator (s_{n-1}^2) equals the population variance.
- It is unbiased, and a good estimator of the population variance.
- Proof: <https://www.youtube.com/watch?v=DlhgiAla3KI>

TWO DIFFERENT VERSIONS

- If you are interested in just describing the distribution of a given sample only and not interested in estimating the population variance, you can use either the unbiased (s_{n-1}^2) or biased (s_n^2) estimator.
- However, if you are interested in estimating the population variance using the sample variance, it is recommended to use the unbiased estimator (s_{n-1}^2).

SUMMARY

- Measures of variability
 - Range
 - Variance
 - Standard deviation
- Unbiased and biased estimators of the population variance