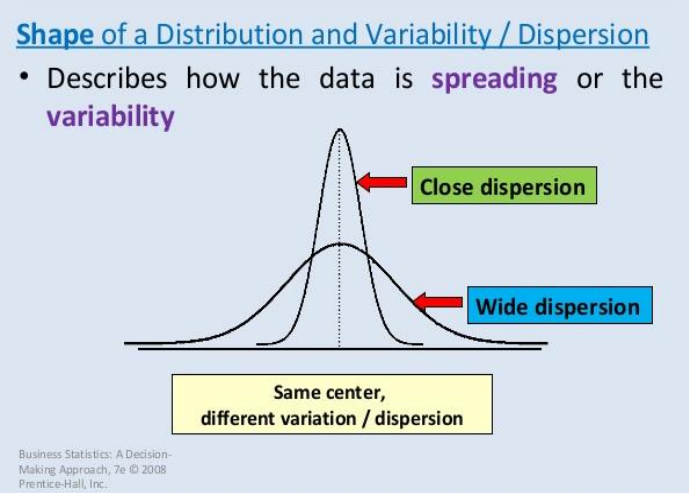
Measures of Variability

The terms **variability, spread, and dispersion** are synonyms, and refer to how spread out a distribution is. Just as in the previous lecture on central tendency where we discussed measures of the center of a distribution of scores, in this lecture we will discuss measures of the variability of a distribution.

Measures of dispersion describe the spread of scores in a distribution. The more spread out the scores are, the higher the dispersion or spread. In Figure bellow, the y-axis is frequency and the x-axis represents values for a variable. There are two distributions, labeled as **small and large**. You can see both are normally distributed (unimodal, symmetrical), and the mean, median, and mode for both fall on the same point. What is different between the two is the spread or dispersion of the scores. The taller-looking distribution shows a smaller dispersion while the wider distribution shows a larger dispersion. For the "small" distribution in the Figure, the data values are concentrated closely near the mean; in the "large" distribution, the data values are more widely spread out from the mean.



Example: Imagine that students in two different groups of statistics take first exam and the mean score in both classrooms is a 75. If that is the only descriptive statistic, I report you might assume that both classes are identical (similar) – but that is not necessarily true. Let's examine the scores for each group.

Group A Scores = 70, 70, 70, 70, 85, 85 mean of Group A = 75

Group B Scores = 70, 72, 73, 75, 75, 85 mean of Group B = 75

Comparing both groups you can see that the scores for group A very few scores are represented (e.g., 70 and 85) and they are very far from the mean, while in group B more scores are represented and clustered close to the mean.

We would say that the spread of scores for group A is greater than group B.

Objectives of measures of variation

- To judge the reliability of MCT
- To control variability itself
- To compare two or more groups of numbers in terms of their variability
- To make further statistical analysis.

1. **Range (R)** is the simplest measure of variability and is really easy to calculate by subtracting the smallest score from the largest score in the data set.

R for GA = 85-70= 15, **R for GB** = 85-70=15

You can see in our statistics course example above that Group A scores have a range of 15 and Group B scores have a range of 15.

That means all the other scores are not included and may not give an unbiased description of the data.

The simplicity of calculating range is appealing but it can be a very unreliable measure of variability. We noticed earlier that the spread of score for each group was very different for each group.

The problem with using range is that it is extremely sensitive to outliers, and one number far away from the rest of the data will greatly alter the value of the range. For example, in the set of numbers 1, 3, 4, 4, 5, 8, and 9, the range is 8 (9 – 1). However, if we add a single person whose score is nowhere close to the rest of the scores, say, 20, the range more than doubles from 8 to 19.

2. **The Mean Deviation (M.D)** is defined as a statistical measure that is used to calculate the average deviation from the mean value of the given data set.

Steps to calculate the mean deviation.

- Step 1: Find the mean value for the given data values
- Step 2: Now, subtract the mean value from each of the data values given (**Note: Ignore the minus symbol**)
- Step 3: Now, find the mean of those values obtained in step 2.

Mean deviation can be found for

a- raw data and $$M.D(\bar{X}) = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n}$$

Σ represents the addition of values
Xi represents each value in the data set
x̄ represents the mean of the data set

$$\Rightarrow \bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i f_i$$

n represents the number of data values

b- frequency distribution

$$M.D(\bar{X}) = \frac{\sum_{i=1}^{k} f_i |X_i - \bar{X}|}{n}$$

In case of median

$$M.A.D(M) = \frac{\sum_{i=1}^{n} f_i |x_i - M|}{N}$$

Example 1: Determine the mean deviation for the data values 5, 3, 7, 8, 4, 9.

Solution:
First, find the mean for the given data:
Mean, $\bar{x}$ = (5+3+7+8+4+9)/6
$\bar{x}$ = 36/6
$\bar{x}$ = 6
Therefore, the mean value is 6.
Now, subtract each mean from the data value, and ignore the minus symbol if any (Ignore"-")

5 – 6 = 1
3 – 6 = 3
7 – 6 = 1
8 – 6 = 2
4 – 6 = 2
9 – 6 = 3
Now, the obtained data set is 1, 3, 1, 2, 2, 3.
Finally, find the mean value for the obtained data set
Therefore, the mean deviation is
= (1+3 + 1+ 2+ 2+3) /6
= 12/6
= 2
Hence, the mean deviation for 5, 3,7, 8, 4, 9 is 2.

Example 2:

In a foreign language class, there are 4 languages, and the frequencies of students learning the language and the frequency of lectures per week are given as:

Language	**Arabic**	**Spanish**	**French**	**English**
No. of students(x_i)	**6**	**5**	**9**	**12**
Frequency of lectures(f_i)	**5**	**7**	**4**	**9**

Calculate the mean deviation about the mean for the given data.

x_i	f_i	$x_i f_i$	$\lvert x_i - \bar{x}\rvert$	$f_i \lvert x_i - \bar{x}\rvert$
6	5	30	2.36	11.8
5	7	35	3.36	23.52
9	4	36	0.64	2.56
12	9	108	3.64	32.76
	$\sum f_i = 25$	$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i f_i = 8.36$		$\sum_{i=1}^{n} f_i \lvert x_i - \bar{x}\rvert = 70.64$

Moving toward variance: sum of squares deviations

3. **Variance (S^2)** is defined as the average squared difference of the scores from the mean. The mathematical definition of the variance is the sum of the squared deviations (distances) of each score from the mean divided by the number of scores in the data set.

To Calculate Variance, we find the **Sum of Squares (SS)** of the deviations from the Sample mean and divide by the **Degrees of Freedom (df = n-1)**

a- (Hand) Direct Method:

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \text{ or } = \frac{SS}{df}$$

b- (Machine) Square Method:

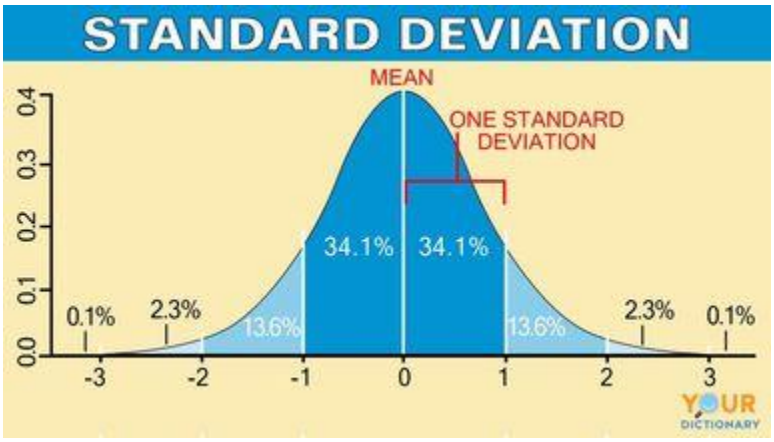
$$S^2 = \frac{\sum xi^2 - \frac{(\sum xi)^2}{n}}{n-1} = \frac{SS}{df}$$

$$SS = sum\ of\ squares = \sum xi^2 - \frac{(\sum xi)^2}{n}$$

For frequency distribution $S^2 x = [\ \Sigma Fi\ (Xi - \bar{X})^2\ ] / (\Sigma Fi - 1)$

4. **Standard Deviation** Is the square root of the variance, denoted by σ for a population and S for a sample. Thus: $\sigma = \sqrt{\sigma^2}$ and $s = \sqrt{S^2}$

The standard deviation is a measure of how dispersed the data is in relation to the mean.
Low standard deviation means data are clustered around the mean, and high standard deviation indicates data are more spread out.



Example: Find the variance and standard deviation of the following

1- sample data 5, 17, 12, 10.

xi	(xi- $\bar{x}$)	(xi- $\bar{x}$)^2
5	5-11= -6	36
17	17-11= 6	36
12	12-11= 1	1
10	10-11= -1	1
$\bar{x}$= 11		Total= 74

$S^2 = \frac{SS}{df}$ ⟶ 74/3 = 24.67

$s = \sqrt{S^2}$ ⟶ $\sqrt{24.67}$ = 4.97

2- The data is given in the form of frequency distribution

Class	Frequency	xi	(xi- $\bar{x}$)	(xi- $\bar{x}$)^2	F_i(xi- $\bar{x}$)^2
40-44	7	42	(42-57)= -15	225	7 X 225= 1575
45-49	10	47	(47-57)= -10	100	10 X 100= 1000
50-54	22	52	(52-57)= -5	25	22 X 25= 550
55-59	15	57	(57-57)= 0	0	15 X 0= 0
60-64	12	62	(62-57)= 5	25	12 X 25= 300
65-69	6	67	(67-57)= 10	100	6 X 100= 1000
70-74	3	72	(72-57) = 15	225	3 X 225= 675
	N= 75	$\bar{x}$ = 57			Total= 5100

S^2 x= [Σ Fi (Xi - $\bar{X}$)^2] / (Σ Fi -1) ⟶ 1/74 (5100) = 69

$s = \sqrt{S^2}$ ⟶ $\sqrt{69}$ = 8.3

5. **Coefficient of Variation (C.V)** is the ratio of the standard deviation to the mean and shows the extent of variability in relation to the mean of the population. The higher the CV, the greater the dispersion.

$$C.V.\ \% = \frac{S}{\bar{x}} * 100$$

The most common use of the coefficient of variation is to assess the precision of a technique. For example:

1. For testing the accuracy of data in Field Experiments
2. If the value of C.V.% is more than 20% it means the data is not accurate.
3. If the value of C.V.% is equal or less than 20% it means the data is accurate. C.V for the above sample is 8.24/57 * 100 = 14.45

6. **Standard Error (SE or $S_{\bar{x}}$)** The standard error is a statistical term that measures the accuracy with which a sample distribution represents a population by using standard deviation. It can be found by dividing standard deviation to square root of the sample size.

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} = \sqrt{\frac{S^2}{n}}$$

Standard error is used to estimate the efficiency, accuracy, and consistency of a sample for example if we want to know testing the accuracy of data of Laboratory Experiments.

1. If the value of SE is **equal or less than the 5% of mean** it means the data is **accurate.**
2. If the value of SE is **more than 5% of mean**, it means the data is **not accurate.**

Examples about Measures of variability

Example: The following data represent hours sleep of 10 students: **8, 7, 7, 6, 8, 9, 5, 7**

1) Calculate measures of variability.
2) Test the accuracy of data.

xi	Xi^2	$(xi- \bar{x})$	$(x- \bar{x})^2$
8	64	1	1
7	49	0	0
7	49	0	0
6	36	-1	1
8	64	1	1
9	81	2	4
5	25	-2	4
6	36	-1	1
$\sum x_i = 56$	$\sum x_i^2 = 404$		SS = 12

$\bar{x}$ = (8+7+7+6+8+9+5+6)/8 = 7

Range = Max – Min ⟹ 9-5= 4

first wat to find $S^2 = \frac{SS}{df}$ = 12/7 1.71

second way to find $S^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$ $S^2 = = \frac{404 - \frac{(56)^2}{8}}{8-1} = \frac{12}{7} = 1.71$

$S = \sqrt{S^2} = \sqrt{1.71} = 1.3$

$SEorS_{\bar{x}} = \sqrt{\frac{S^2}{n}} = \sqrt{\frac{1.71}{8}} = 0.46$

$\%C.V. = \frac{S}{\bar{x}} * 100 = \frac{1.3}{7} * 100 = 18.5\%$ its accurate

Ex - 3 When you are given :

$n - 1 = 5, SS = 392, \bar{x} = 24$

calculate C.V.% then test the acuuracy of data

$n - 1 = 5 \Rightarrow n = 6$ $S = \sqrt{392/5} = 8.85$

$S_{\bar{x}} = \frac{S}{\sqrt{6}} = \frac{8.85}{2.45} = 3.61 > 24 * 0.05 = 1.2$ not accurate

$C.V.\% = \frac{S}{\bar{x}} * 100 = \frac{8.85}{24} * 100 = 36.875\% > 20\%$ data not accurate

Ex - 4 You have $\bar{x} = 116, n = 30$ and $S_{\bar{x}} = 0.9$

$S_{\bar{x}} = \frac{S}{\sqrt{n}} \Rightarrow 0.9 = \frac{S}{5.48} \Rightarrow S = 4.93$

$C.V.\% = \frac{S}{\bar{x}} * 100 = \frac{4.93}{116} * 100 = 4.25\% < 20\%$ data accurate

Example: The following data represent the organic matter% of 10 forest soils

3 4 5 3 6 7 8 6 7 1

1-Calculate measures of variability.

2- Calculate measures of central tendency.

3-Test the accuracy of data.

Example: The following data represent the amount of rainfall (ml) for 20 years: 350, 400, 290, 250, 220, 200, 800, 730, 580, 200, 120, 450, 600, 90, 80, 100, 75, 110, 109, 105

Calculate C.V.% then Test the accuracy of data.