Lecture 2:Descriptive Statistics

1. Measures of Central Tendency

Measures of central tendency describe the **center** or **middle point** of a data set. They provide a single value that summarizes an entire distribution and indicates where most values in the data set tend to cluster.

A. Mean (Arithmetic Average):

The **mean** is the most common measure of central tendency. It is calculated by adding up all the values and dividing by the number of values.

Formula:

$$\text{Mean} = \frac{\sum X}{N}$$

- X: each value in the dataset.
- N: total number of values.

Example:

If the ages of five students are:
16, 18, 17, 19, 20
The mean is:

$$(16 + 18 + 17 + 19 + 20)/5 = 90/5 = 18$$

Advantages:

- Easy to calculate and understand.
- Uses all data points in its calculation.

Disadvantages:

- Sensitive to **extreme values (outliers)**, which can distort the mean.

B. Median:

The **median** is the middle value when the data are arranged in **ascending** or **descending** order.
It divides the data into two equal halves.

How to Calculate:

- If the number of values (**N**) is **odd**, the median is the **middle** value.
- If **N** is **even**, the median is the **average** of the two middle values.

Example:

Dataset: 12, 15, 17, 19, 21

- N=5 (odd), so the median is the 3rd value → **17**.

Another example:
Dataset: 10, 12, 15, 18

- N=4 (even), so the median is:

$$(12+15)/2=13.5$$

Advantages:

- Not affected by outliers or extreme values.
- Useful for skewed distributions.

C. Mode:

The **mode** is the value that appears **most frequently** in a data set.
There can be **no mode**, **one mode (unimodal)**, **two modes (bimodal)**, or **multiple modes**.

Example:

Dataset: 15, 16, 16, 17, 18

- The mode = **16**, since it appears twice.

Advantages:

- Easy to identify.
- Useful for **categorical data**.
- Not affected by extreme values.

---

2. Measures of Dispersion (Variation)

While measures of central tendency summarize the center of a data set, **measures of dispersion** describe **how spread out the data are** around that center.

A. Range:

The **range** is the simplest measure of dispersion.
It shows the difference between the **largest** and **smallest** values in the dataset.

Formula:

$$\text{Range} = \text{Maximum value} - \text{Minimum value}$$

Example:

Dataset: 10, 15, 20, 25
Range = 25 - 10 = **15**

Advantages:

- Easy to calculate.

Disadvantages:

- Considers only **two values** (the extremes).
- Sensitive to outliers.

---

2. Standard Deviation (SD):

The **standard deviation** measures how much each data point **deviates** from the **mean** on average. It gives insight into the **spread** or **variability** of the dataset.

Formula:

$$SD = \sqrt{\frac{\sum(X - \overline{X})^2}{N}}$$

- X: each value in the data set.
- $\overline{X}$: the mean.
- N: number of values.

Example:

Dataset: 10, 12, 14
Mean = 12
Squared deviations:

$(10 - 12)^2 = 4$
$(12 - 12)^2 = 0$
$(14 - 12)^2 = 4$
Sum = 8
Variance = 8 / 3 ≈ 2.67
SD = √2.67 ≈ **1.63**

Advantages:

- Takes into account **all data points**.
- More reliable than the range.

---

3. Variance:

Variance is the **square** of the standard deviation.
It measures the **average squared deviation** from the mean.

Formula:

$$Variance = \frac{\sum(X - \overline{X})^2}{N}$$

Example (same as SD):

Variance ≈ **2.67**

Advantages:

- Important in inferential statistics.

Disadvantages:

- Units are **squared**, making it less interpretable in the original context compared to SD.

Exam: for the following dataset find **Central Tendency** AND **Measures of Dispersion**

Student	Hours Studied (X)	Test Score (Y)
A	2	55
B	3	60
C	4	65
D	5	70
E	5	75
F	6	80
G	7	85
H	7	85
I	8	90
J	9	95

Solution:

	Hours Studied (X)	**Test Score (Y)**
Mean		
Median		
Mode		
Range		
Standard Deviation		