Descriptive Statistics: Numerical

Dr. Md. Israt Rayhan
Professor
Institute of Statistical Research and Training (ISRT)
University of Dhaka
Email: israt@isrt.ac.bd

Describing Data Numerically

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

Measures of Central Tendency

Overview

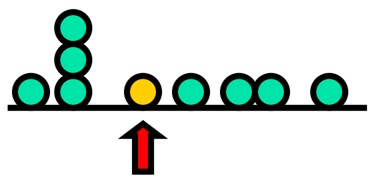
Central Tendency

Mean

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

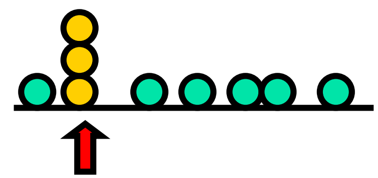
Arithmetic average

Median



Midpoint of ranked values

Mode



Most frequently observed value

Arithmetic Mean

- The arithmetic mean (mean) is the most common measure of central tendency
  - For a population of N values:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Population values

Population size

  - For a sample of size n:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

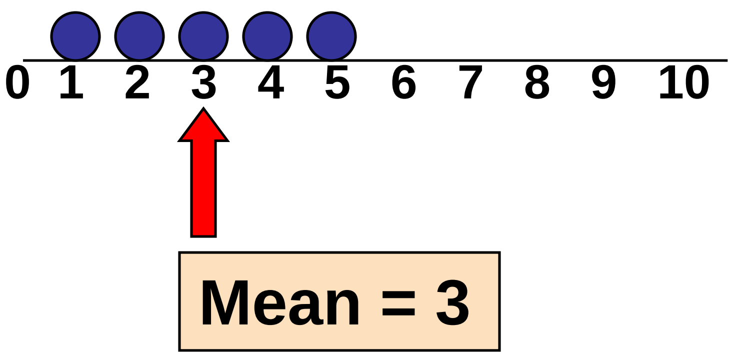
Observed values

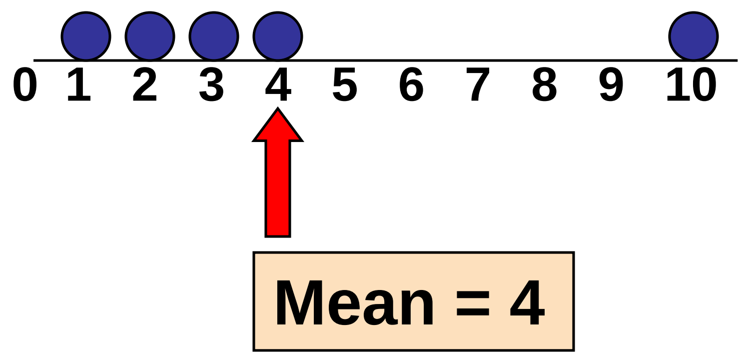
Sample size

Arithmetic Mean

(continued)

- The most common measure of central tendency
- Mean = sum of values divided by the number of values
- Affected by extreme values (outliers)



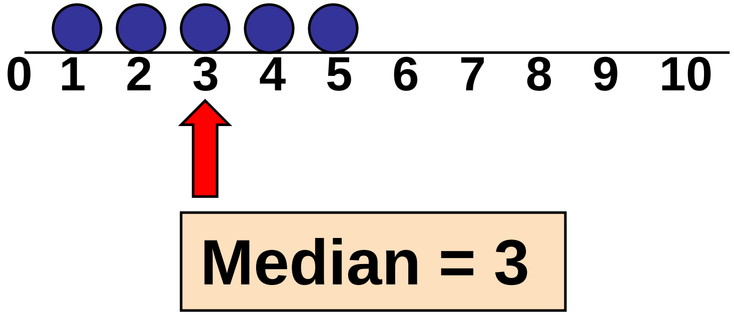


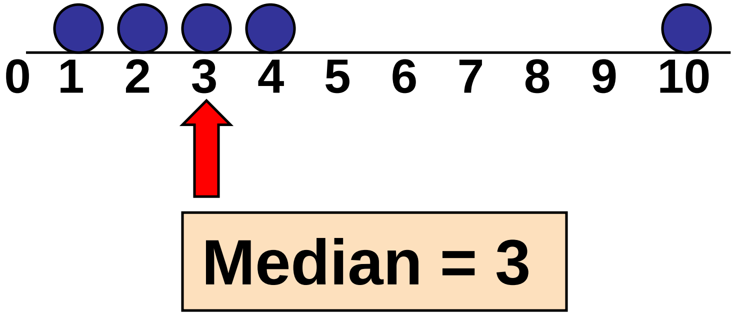
$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Median

- In an ordered list, the median is the “middle” number (50% above, 50% below)





- Not affected by extreme values

Finding the Median

- The location of the median:

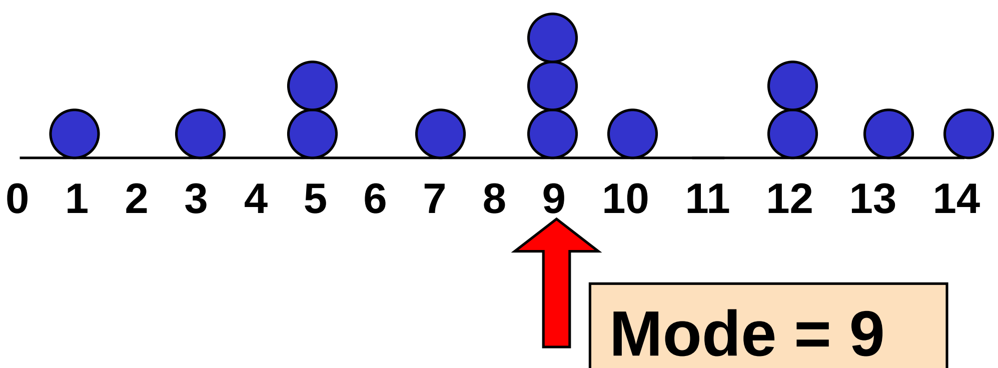
$$\text{Median position} = \frac{n+1}{2} \text{ position in the ordered data}$$

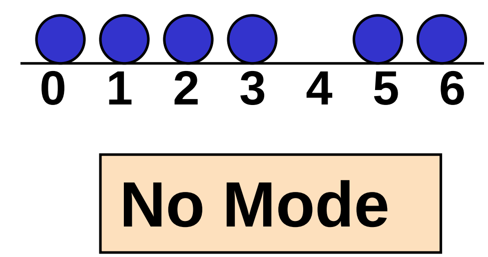
  - If the number of values is odd, the median is the middle number
  - If the number of values is even, the median is the average of the two middle numbers

- Note that $\frac{n+1}{2}$ is not the *value* of the median, only the *position* of the median in the ranked data

Mode

- A measure of central tendency
- Value that occurs most often
- Not affected by extreme values
- Used for either numerical or categorical data
- There may may be no mode
- There may be several modes





Which measure of location is the "best"?

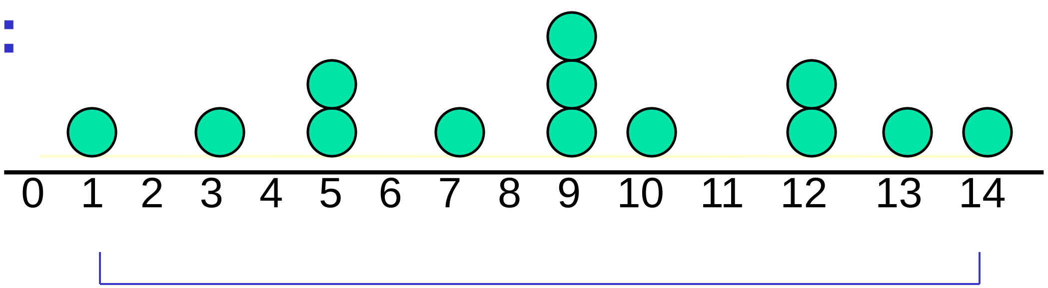
- **Mean** is generally used, unless extreme values (outliers) exist
- Then **median** is often used, since the median is not sensitive to extreme values.
  - Example: Median home prices may be reported for a region – less sensitive to outliers

Range

- Simplest measure of variation
- Difference between the largest and the smallest observations:

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

Example:



Range = 14 - 1 = 13

Population Variance

- Average of squared deviations of values from the mean
  - Population variance:

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

Where μ = population mean

N = population size

x_i = i^{th} value of the variable x

Sample Variance

- Average (approximately) of squared deviations of values from the mean
    - Sample variance:

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

Where $\bar{X}$ = arithmetic mean

n = sample size

X_i = i^{th} value of the variable X

Population Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

  - Population standard deviation:

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

Sample Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

  - Sample standard deviation:

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Calculation Example: Sample Standard Deviation

Sample Data (x_i) :

10	**12**	**14**	**15**	**17**	**18**	**18**	**24**

n = 8 Mean = $\bar{x}$ = 16

$$s = \sqrt{\frac{(10-\bar{X})^2 + (12-\bar{x})^2 + (14-\bar{x})^2 + \cdots + (24-\bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \cdots + (24-16)^2}{8-1}}$$

$$= \sqrt{\frac{126}{7}} = \boxed{4.2426}$$

A measure of the “average” scatter around the mean

Coefficient of Variation

- 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

$$CV = \left(\frac{s}{\bar{x}} \right) \times 100\%$$

Comparing Coefficient of Variation

- Stock A:
  - Average price last year = $50
  - Standard deviation = $5

$$CV_A = \left(\frac{s}{\bar{x}}\right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Stock B:
  - Average price last year = $100
  - Standard deviation = $5

$$CV_B = \left(\frac{s}{\bar{x}}\right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

Both stocks have the same standard deviation, but stock B is less variable relative to its price

Approximations for Grouped Data

Suppose a data set contains values $m_1, m_2, \ldots, m_k$, occurring with frequencies $f_1, f_2, \ldots f_K$

- For a population of N observations the mean is

$$\boxed{\mu = \frac{\sum_{i=1}^{K} f_i m_i}{N}} \qquad \text{where} \quad N = \sum_{i=1}^{K} f_i$$

- For a sample of n observations, the mean is

$$\boxed{\bar{x} = \frac{\sum_{i=1}^{K} f_i m_i}{n}} \qquad \text{where} \quad n = \sum_{i=1}^{K} f_i$$