BHARATHIDASAN UNIVERSITY
Tiruchirappalli- 620024, Tamil Nadu, India

Programme: M.Sc., Biotechnology(Environment)

Course Title: Bioinformatics and Biostatistics
Course Code : EC02A

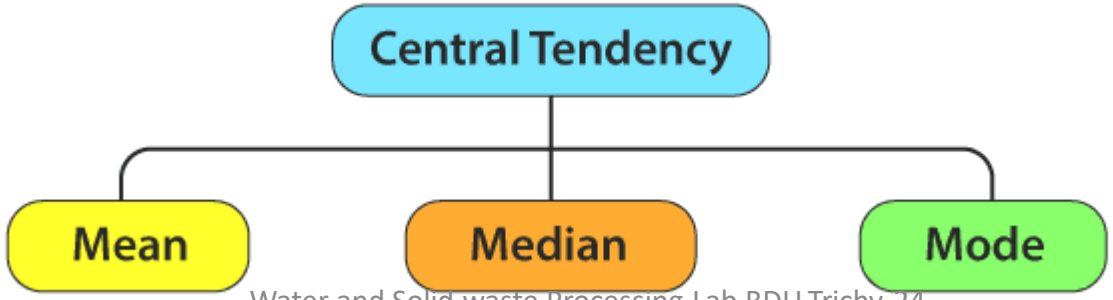
Unit-V
Biostatistics

Dr.M.VASANTHY
Professor
Department of Environmental Biotechnology

Measures of Central Tendency

Measures of Central Tendency

- In statistics, the **central tendency** is the descriptive summary of a data set.
- Through the single value from the dataset, it reflects the centre of the data distribution.
- Moreover, it does not provide information regarding individual data from the dataset, where it gives a summary of the dataset. Generally, the central tendency of a dataset can be defined using some of the measures in statistics.



Mean

- The mean represents the average value of the dataset.
- It can be calculated as the sum of all the values in the dataset divided by the number of values. In general, it is considered as the arithmetic mean.
- Some other measures of mean used to find the central tendency are as follows:
  - Geometric Mean (nth root of the product of n numbers)
  - Harmonic Mean (the reciprocal of the average of the reciprocals)
  - Weighted Mean (where some values contribute more than others)
- It is observed that if all the values in the dataset are the same, then all geometric, arithmetic and harmonic mean values are the same. If there is variability in the data, then the mean value differs.

Arithmetic Mean

Arithmetic mean represents a number that is obtained by dividing the sum of the elements of a set by the number of values in the set. So you can use the layman term Average. If any data set consisting of the values b_1, b_2, b_3, …., b_n then the arithmetic mean B is defined as:

B = (Sum of all observations)/ (Total number of observation)

Innings	1	2	3	4	5	6	7	8	9	10
Runs	50	59	90	8	106	117	59	91	7	74

The arithmetic mean of batting scores also called Batting Average is;

Sum of runs scored/Number of innings = 661/10

The arithmetic mean of his scores in the last 10 innings is 66.1.

Harmonic Mean

Harmonic mean is defined as the average of the reciprocal values of the given values. That is for finding the harmonic mean of the given data set we first take the reciprocal of the data set and then divide the number of the data set by the sum of the reciprocal values. The value so obtained is called the Harmonic mean.

$$\text{Harmonic Mean Formula} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \cdots\cdots + \frac{1}{x_n}}$$

- We can understand the concept of the harmonic mean by studying the example discussed below.
- For example, find the harmonic mean of the data set (2, 4, 8, 16).
- Solution:Given data set, (2, 4, 8, 16)
- Reciprocal of the data set, (1/2, 1/4, 1/8, 1/16)
- Finding the sum of reciprocal value = (1/2 + 1/4 + 1/8 + 1/16)
- ⇒ Harmonic Mean = 4/(8/16 + 4/16 + 2/16 + 1/16)
- ⇒ Harmonic Mean = 4/(15/16)
- ⇒ Harmonic Mean = 64/15

- Thus, the required harmonic mean is 64/15.

- Find the harmonic mean of the data set (3, 6, 9).
- Solution:Given data set, 3, 6, 9)
- Step 1: Here, n = 9
- Step 2: Reciprocal of the data set, (1/3, 1/6, 1/9)
- Step 3: Finding the sum of reciprocal value = (1/3 + 1/6 + 1/9) = (6/18 + 3/18 + 2/18) = 11/18
- Step 4: Finding Harmonic Mean i.e., H.M.= 3/(11/18)
- ⇒ H.M. = 54/11

- Tnus, the harmonic mean of the data set is 54/11.

Geometric Mean

- Geometric Mean is the nth root of the product of the given dataset. It gives the central measure of the data set. To find the geometric mean of various numbers we first multiply the given numbers and then take the nth root of the given number. Suppose we are given 3 numbers 3, 9, and 27 then the geometric mean of the given values is calculated by taking the third root of the product of the three given data. The calculation of Geometric Mean is shown below:
- $\sqrt[3]{(3\times9\times27)} = \sqrt[3]{(729)} = 9$
- Thus, geometric mean is the measure of the central tendency that is used to find the central value of the data set.

Geometric Mean Formula

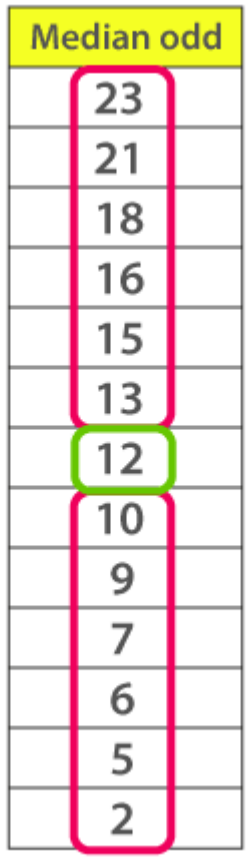
$$GM = (x_1 \times x_2 \times x_n \times .... \times x_n)^{1/n}$$

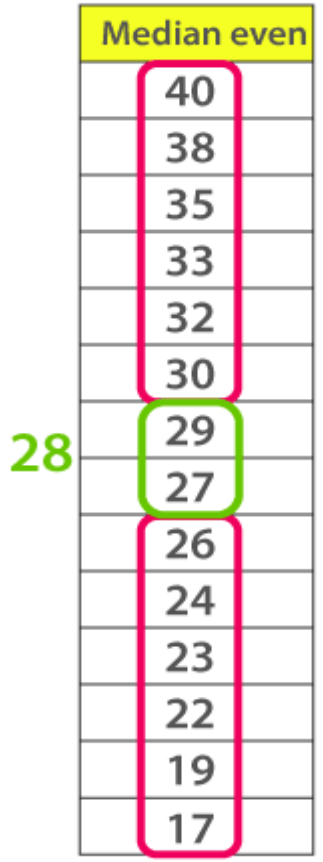
Formula

- Find the geometric mean of 4 and 16.
- Solution:
- Given Numbers = 4 and 16
- GM of 4 and 16 = $\sqrt{(4\times16)} = \sqrt{(64)} = 8$
- Thus, the GM of 4 and 16 is 8

Median

- Median is the middle value of the dataset in which the dataset is arranged in the ascending order or in descending order.
- When the dataset contains an even number of values, then the median value of the dataset can be found by taking the mean of the middle two values.
- If you have skewed distribution, the best measure of finding the central tendency is the median.
- The median is less sensitive to outliers (extreme scores) than the mean and thus a better measure than the mean for highly skewed distributions, e.g. family income. For example mean of 20, 30, 40, and 990 is (20+30+40+990)/4 =270. The median of these four observations is (30+40)/2 =35. Here 3 observations out of 4 lie between 20-40. So, the mean 270 really fails to give a realistic picture of the major part of the data. It is influenced by extreme value 990.





Mode

- The mode represents the frequently occurring value in the dataset.
- Sometimes the dataset may contain multiple modes and in some cases, it does not contain any mode at all.
- If you have categorical data, the mode is the best choice to find the central tendency.

Mode
5
5
5
4
4
3
2
2
1

Correlation coefficient

- It is necessary to know the relationship between two variables as we know about uni variables that with increase in height, weight also increases; if the demand of the commodity increases, the prices also increases; For such data we would like to find answer for the following questions.
- 1. Are the two variables related?
- 2. To what extend they are related?
- Correlation analysis is needed to answer these questions. Correlation analysis is concerned with measuring the strength or degree of relationship between variables. The measure of correlation is called the correlation co-efficient.

- If two variables vary together in the same direction or in opposite directions, they are said to be correlated. That is as x increases y increases consistently, we say that x and y are positively correlated. There are some variables, which are negatively correlated. Where, as x increases y decreases and as x decreases y increases.
- Ex. Price increases as the supply decreases. If the change in one variable is proportional to the change in the other, the two variables are said to be perfectly correlated.
- Correlation can be measured through three different methods; viz., Scatter Diagram, Karl Pearson's Coefficient of Correlation, and Spearman's Rank Correlation Coefficient.

Methods of predicting the correlation

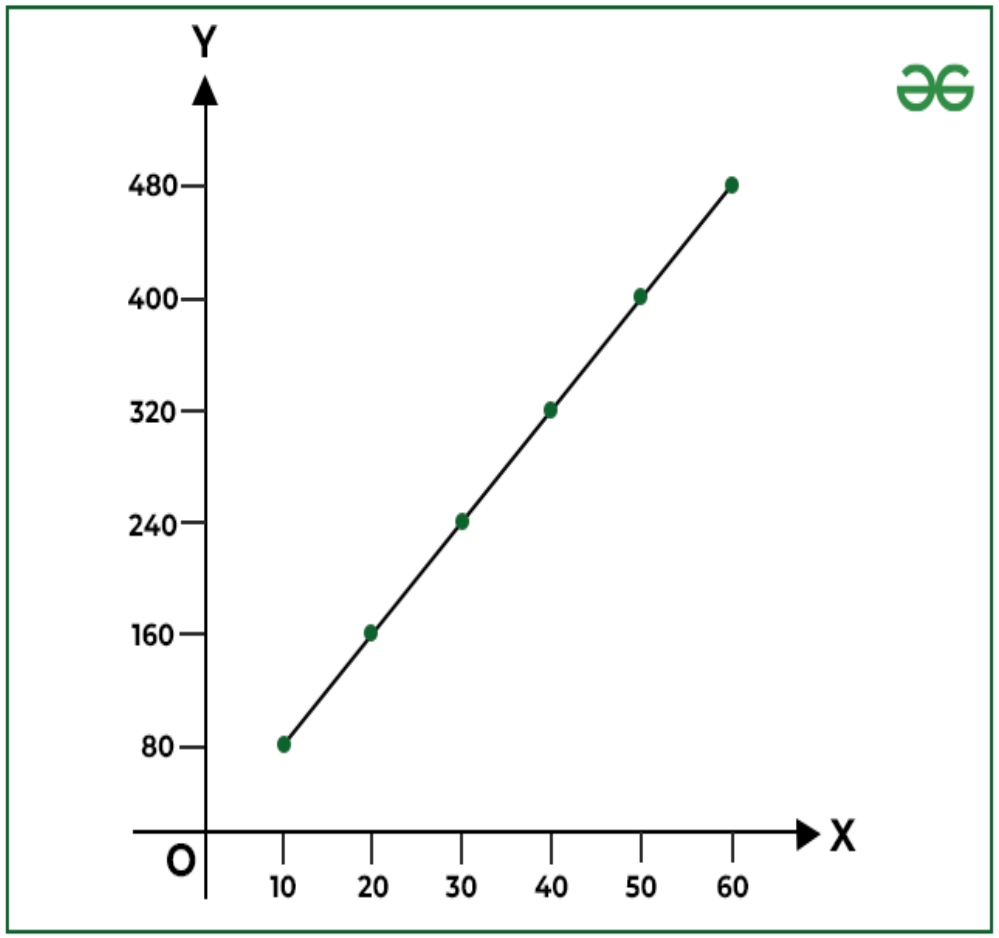
- There are 4 methods of finding whether two given variables are related or not.
- 1. Scatter diagram
- 2. Correlation table
- 3. Correlation graph
- 4. Coefficient of correlation.

- Draw a Scatter Diagram for the following data and state the type of correlation between the given two variables X and Y.
- Information Table
-

X	10	20	30	40	50	60
Y	80	160	240	320	400	480

- Solution:
- We will draw the scatter diagram by plotting the values of Series X on the X-axis and values of Series y on the Y-axis (10, 80), (20, 160),……(60, 480).
- Scatter Diagram
- We can see that all the points of the given two variables X and Y are plotted on a positively sloping straight line, which means that there is a Positive Correlation between the values of Series X and Y.

Scatter diagram



- We can see that all the points of the given two variables X and Y are plotted on a positively sloping straight line, which means that there is a Positive Correlation between the values of Series X and Y.

Karl Pearson's Coefficient of Correlation

- The first person to give a mathematical formula for the measurement of the degree of relationship between two variables in 1890 was Karl Pearson. Karl Pearson's Coefficient of Correlation is also known as Product Moment Correlation or Simple Correlation Coefficient. This method of measuring the coefficient of correlation is the most popular and is widely used. It is denoted by 'r', where r is a pure number which means that r has no unit.

- According to Karl Pearson, "Coefficient of Correlation is calculated by dividing the sum of products of deviations from their respective means by their number of pairs and their standard deviations."

Use Actual Mean Method and determine the coefficient of correlation for the following data:

X	12	16	20	24	28	32	36
Y	6	9	12	15	18	21	24

Calculation of Coefficient of Correlation (Actual Mean Method)

X Series			**Y Series**			**xy**
X	$x = X - \bar{X}$	$\mathbf{x^2}$	**Y**	$y = Y - \bar{Y}$	$\mathbf{y^2}$	
12	-12	144	6	-9	81	108
16	-8	64	9	-6	36	48
20	-4	16	12	-3	9	12
24	0	0	15	0	0	0
28	4	16	18	3	9	12
32	8	64	21	6	36	48
36	12	144	24	9	81	108
$\sum X = 168$		**$\sum x^2 = 448$**	**$\sum Y = 105$**		**$\sum y^2 = 252$**	**$\sum xy = 336$**

$$\overline{X} = \frac{\Sigma x}{N} = \frac{168}{7} = 24 \qquad \overline{Y} = \frac{\Sigma Y}{N} = \frac{105}{7} = 15$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$\Sigma xy = 336 \quad \Sigma x^2 = 448 \quad \Sigma y^2 = 252$$

$$r = \frac{336}{\sqrt{448 \times 252}} = \frac{336}{\sqrt{1,12,896}} = \frac{336}{336} = 1$$

Co-efficient of correlation = 1.

It means that there is a positive correlation between the values of series x and series Y.

Self Check Exercise - 2

1. The following data refers to the standard length and Head length of 10 male fish treated for binary.

Standard length	172	179	122	198	186	144	180	164	222	219
Head length	55	56	38	64	59	48	56	53	69	68

Compute the correlation coefficient between standard length and Head length of this fish.

2. The following data relates to the BOD and COD population per square meter on 8. Compute the correlation between the two.

BOD mg (Lit)	16	59	22	56	02	20	29	32
COD (Mg (L)	120	38	118	26	212	68	56	64

References

- Primer of Biostatistics, Stanton A & Glantz, (2012), McGraw Hill Inc., New York.
- Essential Bioinformatics, Jin xiong (2007). Cambridge University Press, New York.
- Introduction to Biostatistics, Gurumani (2005); MJP Publishers
- Modern statistics for Life Sciences, Alan Graphen Rosie Hails (2002);
- Introduction to Biostatistics and Research methods, Sundhar Rao., David Clark (2016), 5th Edn, PHI Learning Pvt Ltd.
- Biostatistics an Introduction, Mariappan (2013), Pearson Education;

- Thanks for your attention