Unit-II: Descriptive Statistical Measures

Population and Sample

◦ Population is the set of all possible observations (often called cases, records, subjects or data points) for a given context of the problem. The size of the population can be very large in many cases.

◦ Population (also known as universal set) is the set of all possible data for a given context whereas sample is the subset taken from a population.

Measures of Location

Mean (Arithmetic Mean)

The most commonly used measure of location is the **mean (arithmetic mean)**, or average value, for a variable. The mean provides a measure of central location for the data. If the data are for a sample (typically the case), the mean is denoted by x. The sample mean is a point estimate of the (typically unknown) population mean for the variable of interest. If the data for the entire population are available, the population mean is computed in the same manner, but denoted by the Greek letter μ.

SAMPLE MEAN

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Population Mean

$$\mu = \frac{\sum x_i}{N} = \frac{x_1 + x_2 + x_3 + \cdots + x_N}{N}$$

Median

The **median** is *the middle value in an ordered array of numbers*. For an array with an odd number of terms, the median is the middle number. For an array with an even number of terms, the median is the average of the two middle numbers. The following steps are used to determine the median.

Step 1 Arrange the observations in an ordered data array.

Step 2 For an odd number of terms, find the middle term of the ordered array. This is the median.

Step 3 For an even number of terms, find the average of the middle two terms. This average is the median.

- For odd number = middle term
- For Even number = (n+1)/2 th term

Mode

- The **mode** is *the most frequently occurring value in a set of data*. Organizing the data into an ordered array (an ordering of the numbers from smallest to largest) helps to locate the mode.

- In the case of a tie for the most frequently occurring value, two modes are listed. Then the data are said to be **bimodal**. Data sets with more than two modes are referred to as **multimodal**.

Measures of Dispersion

Dispersion refers to the degree of variation in the data, that is, the numerical spread of the data. Several statistical measures characterize dispersion: the *range*, *variance*, and *standard deviation.*

Range

The range is *the difference between the largest value of a data set and the smallest value of a set*. Although it is usually a single numeric value, some business analysts define the range of data as the ordered pair of smallest and largest numbers (smallest, largest). It is a crude measure of variability, describing the distance to the outer bounds of the data set.

- Range = Highest – Lowest

Interquartile Range

The difference between the first and third quartiles, *Q*3 - *Q*1, is often called the **interquartile range (IQR)**, or the **midspread**. This includes only the middle 50% of the data and, therefore, is not influenced by extreme values. Thus, it is sometimes used as an alternative measure of dispersion.

Variance

The **variance** is a measure of variability that utilizes all the data. The variance is based on the *deviation about the mean*, which is the difference between the value of each observation (*xi*) and the mean.

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

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

Standard Deviation

The **standard deviation** is defined to be the positive square root of the variance. We use *s* to denote the sample standard deviation and σ to denote the population standard deviation.

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

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

Coefficient of Variation

The **coefficient of variation** is a statistic that *is the ratio of the standard deviation to the mean expressed in percentage* and is denoted CV. The coefficient of variation essentially is a relative comparison of a standard deviation to its mean. The coefficient of variation can be useful in comparing standard deviations that have been computed from data with different means.

COEFFICIENT OF VARIATION

$$\left(\frac{\text{Standard deviation}}{\text{Mean}} \times 100\right)\%$$

Measures of Association

Covariance

Covariance indicates the direction of the linear relationship between variables. Positive values indicate a positive relationship; negative values indicate a negative relationship. For a sample of size *n* with the observations (*x*1, *y*1), (*x*2, *y*2), and so on, the sample covariance is defined as follows:

SAMPLE COVARIANCE

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Correlation

Correlation is a statistical measure of an association relationship between two random variables. Correlations measure association by measuring how when one variable changes with respective to another variable, which means it is measuring dynamic relationships (change).

Examples:

Family income and expenditure on luxury items.

Yield of a crop and quantity of fertilizer used.

Sales revenue and expenses incurred on advertising.

Frequency of smoking and lung damage.

Weight and height of individuals.

Age and hours of TV viewing per day.

The correlation coefficient can take only values between –1 and +1. Correlation coefficient values near 0 indicate no linear relationship between the two variables. Correlation coefficients greater than 0 indicate a positive linear relationship between the two variables. The closer the correlation coefficient is to +1, the closer the two values are to

forming a straight line that trends upward to the right (positive slope). Correlation coefficients less than 0 indicate a negative linear relationship between the Two variables. The closer the correlation coefficient is to –1, the closer the two values are to forming a straight line with negative slope.

$$r = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

.

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \times \sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

SIGNIFICANCE OF MEASURING CORRELATION

1. Correlation analysis contributes to the understanding of economic behaviour, aids in locating the critically important variables on which others depend, may reveal to the economist the connections by which disturbances spread and suggest to him the paths through which stabilizing forces may become effective.

2.The effect of correlation is to reduce the range of uncertainty of our prediction. The prediction based on correlation analysis will be more reliable and near to reality.

3. In economic theory we come across several types of variables which show some kind of relationship. For example, there exists a relationship between price, supply, and quantity demanded;

4. Correlations are useful in the areas of healthcare such as determining the validity and reliability of clinical measures or in expressing how health problems are related to certain biological or environmental factors.