Basic Statistics 01

Describing Data

Describing Data

1. Numerical Measures
   - Measures of Location
   - Measures of Dispersion
   - Correlation Analysis
2. Frequency Distributions
   - (Relative) Frequency Distribution
   - Histogram

Measures of Location

- Arithmetic mean
- Weighted mean

$$\overline{X}_w = \frac{(w_1 X_1 + w_2 X_2 + \ldots + w_n X_n)}{(w_1 + w_2 + \ldots w_n)}$$

- Geometric mean

$$GM = \sqrt[n]{(X_1)(X_2)(X_3)\ldots(X_n)}$$

- Median
  - The midpoint of the values after they have been ordered from the smallest to the largest
- Mode
  - The value of the observation that appears most frequently

Population Mean

- The ***population mean*** is the sum of all the population values divided by the total number of values:

$$\mu = \frac{\sum X_i}{N}$$

where μ is the population mean.

- N is the total number of observations.
- X_i is a particular value.
- Σ indicates the operation of adding.

Sample Mean

- The ***sample mean*** is the sum of all the sample values divided by the number of sample values:

$$\overline{X} = \frac{\Sigma X_i}{n}$$

Where ***n*** is the total number of values in the sample.

Properties of the Arithmetic Mean

- Every data set has a mean.
- All the values are included in computing the mean.
- A set of data has a unique mean.
- The mean is affected by unusually large or small data values.
- The arithmetic mean is the only measure of central tendency where the sum of the deviations of each value from the mean is zero.

Measures of Dispersion

- Range
  - The difference between the largest and the smallest values
- Mean deviation
  - The arithmetic mean of the absolute values of the deviations from the arithmetic mean. The formula is:

$$MD = \frac{\Sigma \left| X_i - \overline{X} \right|}{n}$$

- Variance
- Standard deviation (*s.d.*)

Variance

- The ***population variance*** is the arithmetic mean of the squared deviations from the population mean. The formula is:

$$\sigma^2 = \frac{\Sigma(X_i - \mu)^2}{N}$$

- The formula for the ***sample variance*** is:

$$s^2 = \frac{\Sigma(X_i - \overline{X})^2}{n-1}$$

Properties of Variance & S.D.

- The major characteristics of variance are:
  - All values are used in the calculation.
  - Not influenced by extreme values.
  - The units are the square of the original units.
- The population (sample) ***standard deviation*** σ (s) is the square root of the population (sample) variance.
  - The units are the same as the original ones.

The Coefficient of Correlation

- The (sample) ***Coefficient of Correlation*** (r) is a measure of the strength of the ***linear*** relationship between two variables.

$$r = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\sqrt{\Sigma(X - \overline{X})^2 \Sigma(Y - \overline{Y})^2}}$$

  - It can range from -1 to 1.
  - Values of -1 or 1 indicate perfect & strong correlation.
  - Values close to 0 indicate weak correlation.
  - Negative values indicate an inverse relationship and positive values indicate a direct relationship.

Frequency Distribution

- A ***frequency distribution*** is a grouping of data into mutually exclusive categories showing the number of observations in each class.
- A ***relative frequency*** distribution shows the percent of observations in each class.

EXAMPLE 1

Dr. K is a professor of ABC University. He wishes to prepare a report showing the number of hours per week students spend studying. About his 30 students, he determines the number of hours each student studied last week.

15.0, 23.7, 19.7, 15.4, 18.3, 23.0, 14.2, 20.8, 13.5, 20.7, 17.4, 18.6, 12.9, 20.3, 13.7, 21.4, 18.3, 29.8, 17.1, 18.9, 10.3, 26.1, 15.7, 14.0, 17.8, 33.8, 23.2, 12.9, 27.1, 16.6.

EXAMPLE 1 *continued*

Hours	Frequency f	Relative Frequency
7.5 up to 12.5	1	1/30=.0333
12.5 up to 17.5	12	12/30=.400
17.5 up to 22.5	10	10/30=.333
22.5 up to 27.5	5	5/30=.1667
27.5 up to 32.5	1	1/30=.0333
32.5 up to 37.5	1	1/30=.0333
TOTAL	30	30/30=1

Histogram for Studying-Hours