DESCRIPTIVE ANALYSIS

AJ

What is Descriptive Analysis

A descriptive analysis is an important first step for conducting statistical analyses. It gives you an idea of the distribution of your data, helps you detect outliers and typos, and enable you identify associations among variables, thus making you ready to conduct further statistical analyses.

Descriptive statistics can be useful for two purposes:

1) to provide basic information about variables in a dataset

2) to highlight potential relationships between variables.

There are variety of descriptive statistics:

- **Measures of central tendency – mean, median, mode**
- **Measures of dispersion – range, variance, standard deviation**
- **Measures of shape – skewness, kurtosis**

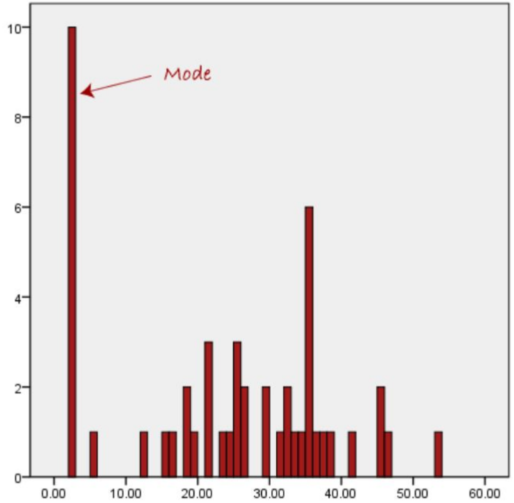
Measures of central tendency

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data.

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others.

$$\text{Mean}(X) = \frac{1}{m}\sum_{i=1}^{m} Xi$$

$$Median(X) = \begin{cases} X_{(r+1)} & \text{If } m \text{ is odd, i.e., r = (m-1)/2} \\ \frac{1}{2}\left(X_{(r)} + X_{(r+1)}\right) & \text{If m is even, i.e., r = m/2} \end{cases}$$



Mean, Median & Mode

Mean

- The mean is the simple mathematical average of a set of two or more numbers
- The mean is the most common measure of the location of a set of points However, the mean is very sensitive to outliers.
- Mean can only be used with numeric data

Median

The middle number; found by ordering all data points and picking out the one in the middle(or if there are two middle numbers, taking the mean of those two numbers).

- It may be thought of as the "middle" value of a data set.

And m=Total number in a dataset. r = Position of the middle value

Mode

The mode is the most frequent score in our data set.

Measures of dispersion

Dispersion is the extent to which a distribution is stretched or squeezed

Summary statistics can also be used to understand variation or dispersion in the data.

Dispersion is a set of measures that helps one to determine the quality of data in an objectively quantifiable manner.

The measure of dispersion contains almost the same unit as the quantity being measured.

There are many Measures of Dispersion found which help us to get more insights into the data:

Range ,Variance ,Standard Deviation ,Skewness , IQR

Range, Variance, Standard Deviation

Range

Range is the difference between a highest and a lowest observation

Range = Highest observation - lowest observation

Example: In {2, 3, 4, 6, 9, 3, 7, 16, 21 } the lowest value is 2, and the highest is 21

Range: 21 – 2 = 19

Variance

Variance is a measurement of the spread between numbers in a data set.

It measures how far each number in the set is from the mean.

Standard Deviation

Standard deviation is a measure of the dispersion of a set of data from its mean

Standard deviation s (or σ) is just the square root of variance s2 (or σ 2)

- When you have "N" data values:

✓ The Population: divide by N

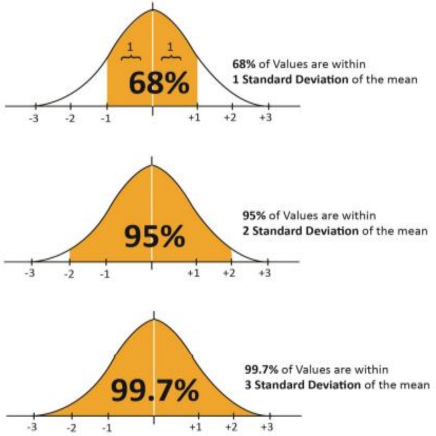
✓ A Sample: divide by N-1

For Sample it is $$s^2 = \frac{1}{(N-1)} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

For Population it is $$\sigma^2 = \frac{1}{N} \sum_{i=1}^{n} (x_i - \mu)^2$$

- When we calculate the standard deviation of normal distribution we find that (generally):

Standard Deviations



Percentile, IQR

Example: Find the 60th percentile in an ordered array of 19 values.

$$i = \frac{p}{100}(n) = \frac{60}{100}(19) = 11.4$$

➡ So use value in the i = 12th position

Percentile

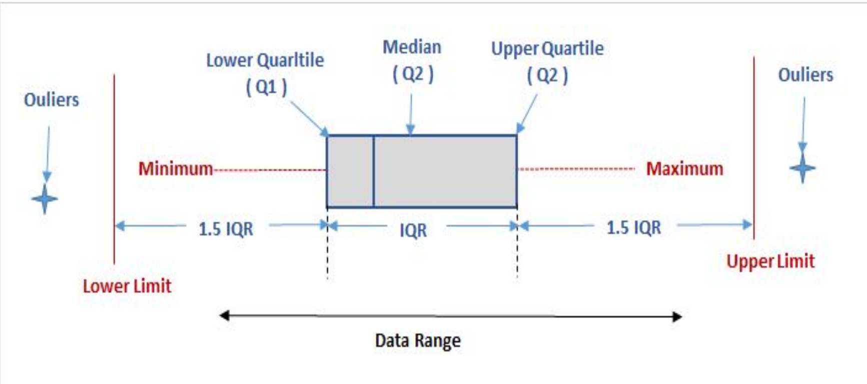
A percentile (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall. For example, the 20th percentile is the value (or score) below which 20% of the observations may be found. The p th percentile in an ordered array of n values is the value in i th position,

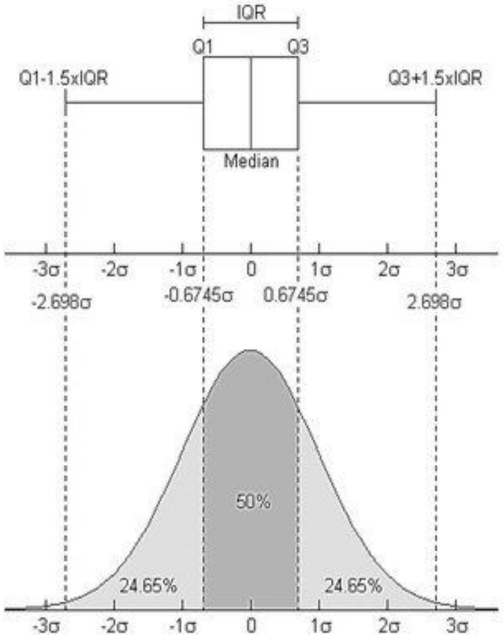
Where,

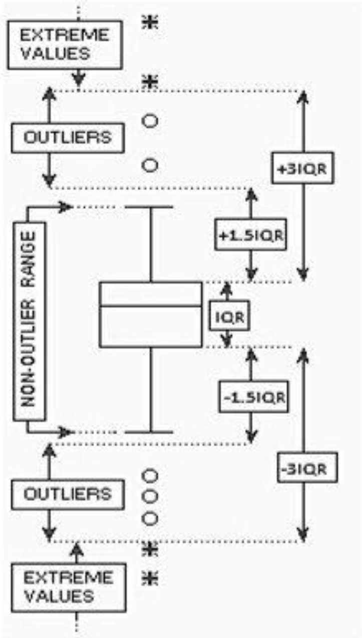
$$i = \frac{p}{100}(n)$$

If i is not an integer, round up to the next higher integer value

Interquartile Range (IQR) : IQR = Q3 – Q1.



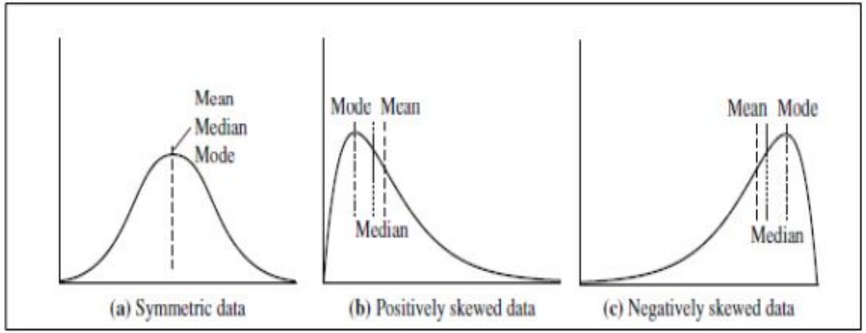




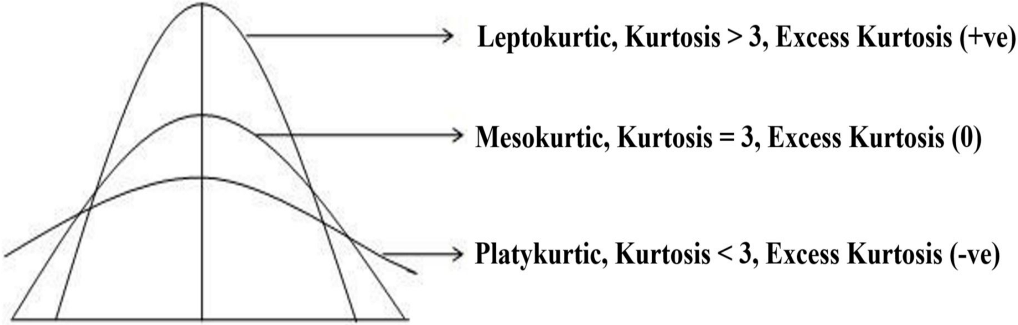
Measures of shape

Measures of shape describes the distribution or pattern of the data in a set. The distribution shape of the quantitative data can be described as there is a logical order to the values and the low and high end values on the horizontal axis of the histogram The distribution shape of the qualitative data cannot be described.

Measures of shape are as follows: Degree of Skewness and Kurtosis



(a) Symmetric data (b) Positively skewed data (c) Negatively skewed data


Leptokurtic, Kurtosis > 3, Excess Kurtosis (+ve)
Mesokurtic, Kurtosis = 3, Excess Kurtosis (0)
Platykurtic, Kurtosis < 3, Excess Kurtosis (-ve)

Skewness

$$\text{Skewness} = \frac{\sum_i^N (X_i - \bar{X})^3}{(N - 1) * \sigma^3}$$

Kurtosis Formula

$$\text{Kurtosis} = n * \frac{\sum_i^n (Y_i - \bar{Y})^4}{\sum_i^n (Y_i - \bar{Y}^2)^2}$$

Degree of Skewness:

Skewness is the tendency for the values to be more frequent around the high or low ends of the x axis

• Skewness is a measure of symmetry

• Symmetric data – The data is symmetrically distributed on both side of medium mean = median = mode

• Positively skewed - Tail on the right side is longer than the left side. mode < median < mean

• Negatively skewed - Tail on the left side is longer than the right side. mode > median > mean

Skewness

If skewness is less than -1 or greater than 1, the distribution is highly skewed.

If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed.

If skewness is between -0.5 and 0.5, the distribution is approximately symmetric, close to Normal Distribution

Kurtosis:

Kurtosis is the sharpness of the peak of a frequency-distribution curve.

• It describes the shape of the distribution of the tail's in relation to its shape

Types of Kurtosis

Mesokurtic – It has flatter tail than standard normal distribution and slightly lower peak

Leptokurtic – It has extremely thick tail and a very thin and tall peak

Platykurtic – It has slender tail and a peak that's smaller than Mesokurtic distribution

Kurtosis - Measure of the relative peak of a distribution.

K = 3 indicates a normal "bell-shaped" distribution (mesokurtic).

K < 3 indicates a platykurtic distribution (flatter than a normal distribution with shorter tails).

K > 3 indicates a leptokurtic distribution (more peaked than a normal distribution with longer tails)