2.5.3 Frequency Distributions

A frequency distribution refers to data classified on the basis of some variable that can be measured such as price, wages, age, number of units produced or consumed. The term variable refers to the characteristics that vary in amount or magnitude in a frequency distribution. A variable may be either continuous or discrete.

2.5.3.1 Ungrouped Data Distributions

Ungrouped data distribution is a table, which shows the each value of the variable together with number of times it occurs. Number of times that a value occurs is called frequency of that value.

For example, let's suppose that you are collecting data on how many hours of sleep college students get each night. After conducting a survey of 30 of your classmates, you are left with the following set of scores: 7, 5, 8, 9, 4, 10, 7, 9, 9, 6, 5, 11, 6, 5, 9, 10, 8, 6, 9, 7, 9, 8, 4, 7, 8, 7, 6, 10, 4, 8

Because above data are not arranged in any systematic way they are referred to as raw data. One way of organizing the raw data is to arrange them into an array. An array is an arrangement of raw data in ascending or descending order of magnitude. Ascending array of above data is given below.

Hours of sleep each night	Frequency
4	3
5	3
6	4
7	5
8	5
9	6
10	3
11	1

2.5.3.2 Grouped Data Distribution

A tabular arrangement of data by classes together with the corresponding class frequencies is called a grouped data distribution. Following table gives a frequency distribution of the intelligence test scores for 200 candidates.

Marks	No.of candidates
0-09	13
10-19	20
20-29	30
30-39	17
40-49	18
50-59	12
60-69	40
70-79	13
80-89	29
90-99	08
$\sum f$	200

Although the grouping process generally destroys much of the original detail of the data, an important advantage is gained in the clear 'overall' picture that is obtained and in the vital relationships that are thereby made evident.

2.5.3.3 Class limits, Class boundaries, Class marks and Class width

A symbol defining a class, such as 60-69 in above table, is called a class interval. The end numbers, 60 and 69 are called class limits. The smaller number (60) is the lower class limit and the larger number (69) is the upper class limit.

A class interval that at least theoretically has either no upper class limit or no lower limit indicated is called an open class interval. For example, referring to age group of individuals, the class interval '50 year and above' is an open class interval.

Class Boundaries

Numbers, indicated briefly by the exact numbers 49.5 and 59.5 are called class boundaries, or true class limits, the smaller number (49.5) is the lower class boundary and the larger number (59.5) is the upper class boundary.

The Size or width of a Class Interval

The size or width of a class interval is the difference between class boundaries and is also referred to as the class width, class size or class length. If all class intervals of a frequency distribution have equal widths, this common width is denoted by c.

The Class Mark

The mark is the midpoint of the class interval and is obtained by adding the lower and upper class limits, and dividing it by 2. In purpose of further mathematical analysis, all observations belong to a given class interval are assumed to coincide with the class mark.

2.5.3.4 General rules for Forming Frequency Distributions

1. Determine the largest and smallest numbers in the raw data and thus find the range (the difference between the largest and smallest numbers).

2. Divide the range into a convenient number of class intervals having the same size. If this is not feasible, use class intervals of different sizes or open class intervals. The number of class intervals is usually taken between 5 and 20, depending on the data. Class intervals are also chosen so that the class marks (or midpoints) coincide with the actually observed data.

One can get the square root of n as the number of classes and get the class width depending on that.

3. Determine the number of observations falling into each class interval; that is find tile class frequencies. This can be done by using tally marks.

Example: Make a frequency distribution with 7 class intervals from the following data.

40 36 43, 57, 5I, 90, 92, 74, 66, 85, 41, 57, 30, 63, 84, 93, 71, 53, 36, 63, 39 44, 59, 43, 90, 82, 88, 72, 73, 45. 53, 64. 79, 85, 99 68. 65. 69, 83, 80

2.5.3.5 Cumulative Frequency Distributions

There are two types of cumulative distributions.

1. Less than type cumulative frequency distribution
2. More than type cumulative frequency distribution

Less than type cumulative frequency distribution

Less than cumulative frequency of a class interval is the total number of observations less than to the upper boundary of the class interval. A table, which shows the class intervals together with the less than cumulative frequencies, is called less than type cumulative frequency distribution.

More than type cumulative frequency distribution

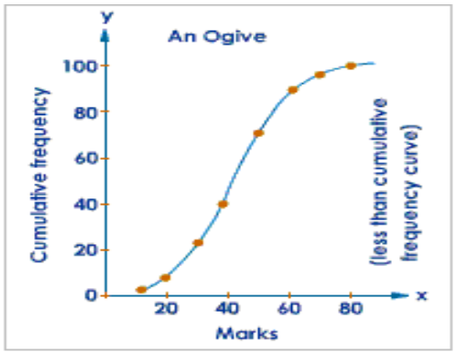
More than cumulative frequency of a class interval is the total number of observations equal or more than to the lower boundary of the class interval. A table, which shows the class intervals together with the more than cumulative frequencies, is called more than type's cumulative frequency distribution

Example : Calculate the both cumulative distribution for the following data.

Marks (class)	f	Less than Cumulative Frequency	More than Cumulative Frequency
30-39	3	3	40
40-49	6	9	37
50-59	6	15	31
60-69	7	22	25
70-79	5	27	18
80-89	8	35	13
90-99	5	40	5

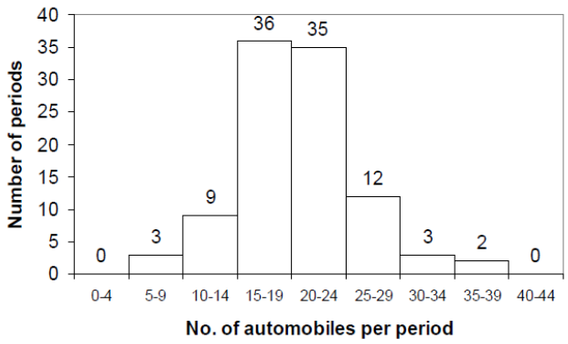
Ogives

An ogive is a graph in which a point is plotted above each class boundary at a height equal to the cumulative frequency corresponding to that boundary. Ogives can also be constructed for a cumulative relative frequency distribution as well as a cumulative percentage distribution. (Draw using above)



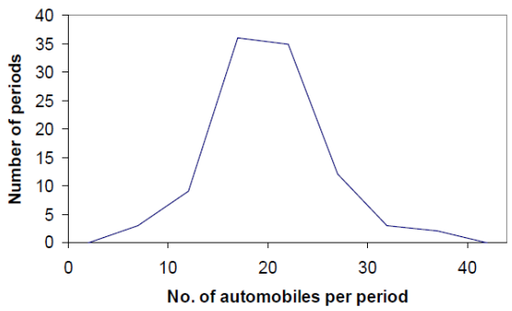
2.5.3.6 Histograms

A histogram is a graph that displays the classes on the horizontal axis and the frequencies of the classes on the vertical axis. The frequency of each class is represented by a vertical bar whose height is equal to the frequency of the class. A histogram is similar to a bar graph. However, a histogram utilizes classes or intervals and frequencies while a bar graph utilizes categories and frequencies. (Draw using above)



2.5.3.7 Frequency Polygon and Frequency Curve

As in the histogram, the base line is divided into sections corresponding to the class-interval, but instead of the rectangles, the points of successive class marks are being connected. The frequency polygon is particularly useful when two or more distributions are to be presented for comparison on the same graph. A frequency curve can be obtained by smoothing the frequency polygon.



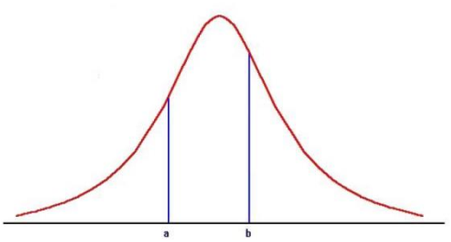
2.5.3.8 Relative Frequency

Relative frequency of a class is defined as:

$$\frac{\text{Frequency of the Class}}{\text{Total Frequency}}$$

If the frequencies are changed to relative frequencies, then a relative frequency histogram, a relative frequency polygon and a relative frequency curve can similarly be constructed.

Relative frequency curve can be considered as probability curve if the total area under the curve be set to 1. Hence the area under the relative frequency curve between a and b is the probability between interval a and b.



2.6 Summary Measures

2.6.1 Importance of Summary Measures

Summary measures are used to summarize the basic features of the data in a study. They provide simple summaries about the sample and the measures. In a research study we may have lots of measures. Or we may measure a large number of people on any measure. Descriptive statistics help us to simply large amounts of data in a sensible way. Each descriptive statistic reduces lots of data into a simpler summary.

2.6.2 Measures of Central Tendency

There is always a tendency of individual observations contained in any set of data to cluster or center around a specific value, which is the central value. This peculiar characteristic of the data is referred to as central tendency. A measure of central tendency lies in obtaining the central value so that it may be used as representative of the entire set of data to which it belongs. The following are the important types of measures of central tendencies.

1. Mean 2. Median 3. Mode

1.1 Simple Arithmetic Mean
1.2 Weighted Arithmetic Mean
1.3 Combined Mean
1.4 Geometric Mean

2.6.2 Simple Arithmetic Mean

The simple arithmetic mean of a series of values is defined as sum of values divided by number of values in the series. $\bar{X} = \frac{\sum X_i}{n}$ is used to calculate where n is the sample size and X_i denotes the i^{th} observation. $\bar{X} = \frac{\sum f_i X_i}{\sum f_i}$ is used for grouped data where X_i and f_i are the class mark and frequency of i^{th} group respectively.

Example: Find the mean of the following data set.

20, 30, 40, 50, 29, 33, 47, 51, 41, 59 35.5

2.6.3 Combined arithmetic mean

If we know the means and the number of items in two or more related groups, the combined or composite mean can be computed with the help of the following formula.

$\mu_x = \frac{\sum n_i \mu_{xi}}{\sum n_i}$; μ_{xi} is the arithmetic mean of i^{th} group and n_i is the number of observations in i^{th} group.

Example: There are two branches of a company, employing 100 and 80 persons respectively. If the arithmetic means of the monthly salaries paid by the two companies are Rs. 45000 and Rs. 57500 respectively, find the arithmetic mean of the salaries of the employees of the companies as a whole.

2.6.4 Geometric mean

Geometric mean is defined the n^{th} root of the product of n items. The geometric mean is never larger than the arithmetic mean: on occasion it may turn out to be the same as the arithmetic mean, but usually it is smaller. If there are zeros or negative values in the series, the geometric mean cannot be used.

$$Geometric\ mean = \sqrt[n]{\prod_{i=1}^{n} X_i} \quad \text{Where; } X_1, X_2, \dots \text{ are observed values.}$$

Example: The price of a commodity increased by 30% from 1997 to 1998 by 20% 1998 to 1999, and by 5% from 1999 to 2000. Find yearly average percentage increase from 1997 to 2000.

Uses of Geometric Mean

Geometric mean is highly useful in averaging ratios, percentages and rate of increase between two periods. Geometric mean is important in the construction of index numbers. In economic and social sciences, where we want to give more weight to smaller items and smaller weight to larger items, geometric mean is appropriate.

2.6.5 Median

Median is the point that divides the data sample into equal parts. Simply it is the middle point of the data set. Hence the sample has to be arranged in ascending or descending order before finding the median.

Calculation of Median for ungrouped data

1. Arrange the data in ascending order or descending order.
2. Apply the formula

$$Median\ (Q_2) = \left(\frac{n+1}{2}\right)^{th}\ observation$$

Example: Find the median of the following data set
1, 5, 6 , 9, 8, 7, 3 ,2, 4, 3, 1

Calculation of Median for grouped data

1. Find the median class using $\frac{n}{2}$

2. Apply the formula

$$median = \frac{\frac{n}{2} - cf}{f}(w) + L_m$$

L_m - Lower boundary of the median class
cf - Cumulative frequency of the class prior to median class
w - Width of the median class
f - Frequency of the median class

2.6.6 Mode

Mode is the most common item of a data set. Mode is defined which occurs most frequently in a distribution.

Calculation of mode for ungrouped data

Mode can often be found out by mere inspection in case of individual observations. The data have to be arranged in the form of an array so that the value which has the highest frequency can be known.

Example: Find the mode
1 2 5 8 9 4 6 7 6 6 5 6 3

Calculation of mode for grouped data

1. Identify the modal class(class with the highest frequency)

2. Apply the following equation

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

where $l =$ lower limit of the modal class,
$h =$ size of the class interval (assuming all class sizes to be equal),
$f_1 =$ frequency of the modal class,
$f_0 =$ frequency of the class preceding the modal class,
$f_2 =$ frequency of the class succeeding the modal class.

2.6.7 Quartiles

Values of the variable, which divide the distribution into four equal parts, are known as quartiles. Each portion contains equal number of items. The first, second and third points are termed as first quartile (Q_1), second quartile (Q_2 — Better named as median) and third quartile (Q_3). The first quartile (Q1), or lower quartile, has 25% of the items of the distribution below it and 75 % items are greater than it.

Calculation of Quartiles for ungrouped data

The method for locating the quartiles is the same as that for median. The following steps may be noted:

1. Arrange the data in ascending order or descending order

2. Find out the the quartile.

$$ith\ quartile\ (Q_i) = \ i * \left(\frac{n+1}{4}\right)^{th}\ observation$$

Example: Calculate the first quartile and the third quartile from the following data
1, 5, 6 , 9, 8, 7, 3 ,2, 4, 3, 1

2.7 Measures of Dispersion

By comparing several data sets, the average may be the same, but variables may highly differ in magnitudes and therefore the central tendency calculated from such variables may not be the most typical or representative, in many cases. To know the extent of spread about these averages or the variations of items, we have to resort to some other measures. One such measure is dispersion.

The following are the important methods of measuring dispersion.

1. Range
2. Semi Inter-quartile Range
3. Variance and Standard Deviation

2.7.1 Range

The range is the simplest measure of dispersion. It is defined as the difference between the smallest value and the largest value in the distribution. It is a rough measure or dispersion. Its measure depends upon the extreme items and not on all the items. Further, we use relative percentage range by dividing range from mean of the sample. And it is a unit less measure.

2.7.2 Semi Inter-Quartile Range

By eliminating the lowest 25% and the highest 25% of items in a series, we are left with the central 50%, which are ordinarily free of extreme values. Inter quartile range is computed by deducting the value of first quartile from the value of-third quartile. We can calculate Semi Inter-Quartile range or Quartile Deviation is defined as half the distance between third quartile and first quartile.

Inter quartile range = $Q_3 - Q_1$ Semi Inter-Quartile Range = $(Q_3 - Q_1)/2$

2.7.2 Variance and Standard Deviation

It is the most important measure of dispersion and is widely used in many statistical formulas. Standard deviation is also called Root-Mean Square Deviation or Mean Error. In this method the drawback of ignoring the algebraic sign (in mean deviation) is overcome by taking the square of deviations, thereby making all the deviations as positive. It is defined as positive square root of the arithmetic mean of the squares of the deviations of the given observation from their arithmetic mean. The population standard deviation is denoted by the Greek letter σ (sigma.) and the Variation is denoted by σ^2(for samples we use s and S^2). Standard deviation is the best measure of dispersion. It is widely used in statistics because it possesses most of the characteristics of an ideal measure of dispersion. It is used in sampling theory and biologists. It is used in coefficient of correlation and in the study of symmetrical frequency distribution.

Calculation of variance for ungrouped data

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$ Sample Variance $$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$ Population Variance

To get the unbiased sample variance we use 'n-1' at the denominator, but using 'n' is also a possible way to calculate the variance.

Calculation of variance for grouped data

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

2.7.3 Coefficient of variation (CV)

If we want to compare the variability of two or more samples then we can use CV. A greater CV value shows a greater the variation, less stability, less uniformity, less consistency or less homogeneity in the sample.

CV for a population:

$$CV = \frac{\sigma}{\mu} * 100\%$$

CV for a sample:

$$CV = \frac{s}{\bar{x}} * 100\%$$

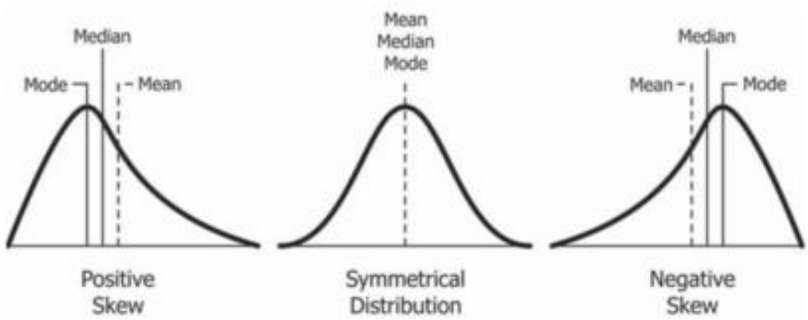
Example: Two cricketers scored following runs in last 10 innings. Who is the most consistent?

A; 42 17 83 59 72 76 64 45 40 32
B; 28 70 31 10 59 108 82 14 03 95

2.7.4 Skewness

The measure of central tendency or dispersion do not show, directly, whether the distribution is symmetric or not. They are unable to characterize the distribution completely. Thus Skewness is used.

Skewness is the lack of symmetry. When a frequency distribution is plotted, presence of skewness tends the series to be dispersed more on one side of the mean than on the other. The measure of skewness helps us to determine the nature and extent of concentration of the observations towards higher or lower values of the variable. Thus it gives the amount as well as the direction. The possible shapes are symmetric, positively skewed and negatively skewed.



Measures of Skewness

Karl Pearson's coefficients of skewness is calculated for unimodal and by modal as below.

$$S_{kp_1} = \frac{\text{mean - mode}}{\text{standard deviation}} \quad \text{or} \quad S_{kp_2} = \frac{3(\text{mean - median})}{\text{standard deviation}}$$

This value varies between the limits of ±3. Values as extreme as ±1 are rare. It is 0 for symmetric, positive and negative values denote the direction respectively. It is independent from the scale but affected by extreme values.

Bowley's coefficient of Skewness is calculated using quantiles as below.

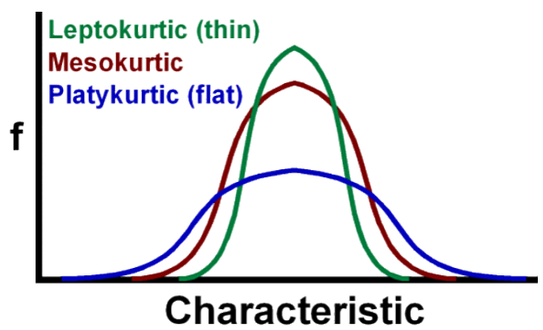
$$S_{kq} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

This measure depends on quartiles and varies between ±1 . And it is not affected by extreme values. But it does not utilize the data fully. For a symmetrical distribution, it is seen that Q_1, and Q_3 are equidistant from median (Q_2).

2.7.5 Kurtosis

Kurtosis is a measure of the "tailedness" of the probability distribution. A standard normal distribution has kurtosis of 3 and is recognized as mesokurtic. An increased kurtosis (>3) can be visualized as a thin "bell" with a high peak whereas a decreased kurtosis corresponds to a broadening of the peak and "thickening" of the tails. Kurtosis >3 is recognized as leptokurtic and <3 as platykurtic.

$$\text{Kurtosis} = \frac{m_4}{m_2^2}, \text{ where: } m_2 = 2^{\text{nd}} \text{ moment} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n} \text{ and: } m_4 = 4^{\text{th}} \text{ moment} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^4}{n}$$



Example Find the mean, median, mode, semi-IQR and skewness of the data set.

Marks (class)	f
30-39	3
40-49	6
50-59	6
60-69	7
70-79	5
80-89	8
90-99	5