مـركـز الإحـصـاء
STATISTICS CENTRE

Statistical analysis Principles Handbook

Methodology and Quality Guides - Guide No. (10)

Table of Contents

Introduction

As part of its role to provide statistical handbooks on methedologies of all kinds of staistical work, including surveys and polls, data processing and validation, and quality measurement and control, Statistics Centre - Abu Dhabi issued this handbook, featuring principles of the descriptive and inferential statistical analysis, to acquaint users with the most important ways of data analysis and presentation, statistical standards development, statistical estimates process, statistical assumption testing, correlations and linear regression between two variables.

According to the statistical theory, statistical analysis methods are various and intertwined, depending on the number and types of variables and the types of their relations. This handbook examines the fundamental principles of data descriptive analysis methods. Analysts can also resort to more detailed statistical resources and theories in case they want to use other methods.

This handbook was issued in four chapters to cover all fundamental principles of the statistical analysis. The first chapter tackles the presentation methods of various data, including individual or grouped data over intervals, and methods of building relative and cumulative frequency tables. The second chapter covers various statistical measures, including the central tendency measures of median, mean, and mode, and the dispersion measures, such as variance, standard deviation, mean deviation, etc.

The third chapter addresses statistical inference topics of statistical estimates, including point estimates and confidence intervals estimates. This chapter also elaborates on the development and testing of simple statistical hypothesis about the population ratio or mean.

The fourth chapter provides explanation on the correlation coefficients, namely the Pearson's correlation coefficient that measures correlation between two continuous variables, the Spearman's correlation that measures discrete variables, as well as the partial correlation coefficients that measures different variables. This chapter also covers the development of simple linear regression model by calculating regression coefficients estimates, developing the statistical model and testing its accuracy and efficiency, and establishing predictions for the dependent variable values, assuming that independent variable values were also given.

Chapter I

Presentation of statistical data

1.1. Introduction

Statistics is used to collect, organize, summarize, present and analyze data to draw acceptable results and make sound decisions based upon this analysis. Moreover, statistics has a descriptive part, descriptive statistics is the methods used to organize and summarize informaion to be understood. Accordingly, it is necessary to cover data presentation in this chapter, as a way used to organize and provide data to the user; to be understand, compare relevant terms, and draw preliminary results easily. The way data is presented in many fields is crucial, there are methods that encourage the recipient to interact more. Data may be used in business sector, economics, research, statistics, etc. Therefore, there are many methods used to present data in the best way, and in a way that serves the purpose completely.

1.2. Summarization and presentation of data

Having colleted data, it may be difficult to be studied and understood without being organized and tabulated. The researcher often aims, while presenting data, to attract the reader's attention towards the relationship between the variables being studied or comparison of data sets; accordingly, the researcher simplifies data by presenting data in expressive and meaningful forms. Accordingly, it is necessary to present such data clearly and accurately. Data presentation methods include the following:

1.2.1. Tabulation

Tabulation includes many categorization methods, most importantly:

1. Frequency table

The first step to present statistical data is designing a frequency distribution table, which organizes, summarizes and divides statistical data into two columns and a set of rows; the first column represents the category of quantitative data, and the second column represents the frequency of the category or charactaristic, and shows the number of observations of data for each category.

Table 1: Frequency distribution of bachelor's degree holders in an educational institution by gender.

Gender	Frequency
Male	23
Female	26

Table 2: Frequency distribution of average wages of of employees in an institution

Wage (in AED)	Frequency
1500	1
5000	3
6500	5
10500	1
11200	3
14000	3
16500	2
18000	1
21500	1
Sum	**20**

Intervals Formation in the frequency distribution table

The purpose of creating regular (equal-length) intervals for data is to reduce the volume of row data by grouping close values. There are no specific rules to determine the lengths or numbers of intervals. Accordingly, it is desirable that the number of intervals shall neither be small to avoid loss of many details about the distribution of the data, nor too large, otherwise the wisdom of grouping is lost in intervals . Determination of the numbers and length of each interval relies upon data range (the difference between the highest and lowest value of data). To clarify how intervalsare constructed, the data of Table 2 are used and the steps are as follows:

- Calculate the range of the data R; where R = 21500 - 1500 = 20000
- Choose number of intervals, to be for example 5 intervals
- Calculate the interval length (L) by dividing the data range by Number of intervales: L = 20000/ 5 = 4000
- Choose the smallest value of data values to be the first rounded interval, with the interval length added thereto to get the start point of the second interval and so on. For example, the start points of the first interval, as stated in Table 2, is: 1500 + 4000 = 5500
- To determine the endpoint of any interval, add to the start point the length of interval and subtract 1 from the result. E.g., the endpoint of the first interval is 5499.

The frequency table is made up of two columns: the first column represents the wage interval boundarys, and the second column represents the frequency. Accordingly, data are grouped int intervals as shown in the table below.

Table 3: Intervals boundarys and frequency for the data in (Table 2)

Interval boundarys	Frequency
5499 - 1500	4
9499 - 5500	5
13499 - 9500	4
17499 - 13500	5
21499 - 17500	2
Sum	**20**

2. Relative and percentage frequency distribution tables

The frequency distribution table can be used to form two other types of tables; relative and percentage frequency distribution, each table is made up of two columns, the relative frequency distribution includes the relative frequency; which is the frequency of any interval divided by the sum of frequencies, and the relative frequency sum of all intervals will equal one. The percentage frequency distribution includes the percentage frequency, which is obtained by multiplying the relative frequency by 100. The sum of the percentage frequencies equals 100, as shown in Table 4 below, using data stated in Table 2.

Table 4: Relative and percentage frequency distribution table

Interval boundarys	Relative frequency	Percentage frequency (%)
5499 - 1500	0.20	20
9499 - 5500	0.25	25
13499 - 9500	0.20	20
17499 - 13500	0.25	25
21499 - 17500	0.10	10
Sum	**1.00**	**100**

3. Exact interval boundarys

According to the aforementioned, it can be seen that intervals are not connected together, and there are some uncovered values among the intervals. Hence, it is necessary to identify upper- and lower-interval boundarys and interval midpoints to be used later while presenting data. They can be calculated through the following formulas:

- Exact upper-interval boundary = upper interval boundary + 1/2X length unit
- Exact lower-interval boundary = lower interval boundary - 1/2 X length unit
- Interval centre (interval midpoint) = (upper interval boundary + lower interval boundary) ÷ 2

The table below shows exact interval boundarys and midpoints, calculated based on the data stated in Table 2.

Table 5: Exact interval boundarys and medpoints calculated based on the data in Table 2:

Exact interval boundarys	Interval mid point	Frequency
5499.5 - 1499.5	3499.5	4
9499.5 - 5499.5	7499.5	5
13499.5 - 9499.5	11499.5	4
17499.5 - 13499.5	15499.5	5
21499.5 - 17499.5	19499.5	2

4. Cumulative frequency table

We often focus on the frequencies that are less than or equal specific data values. Accordingly, the cumulative frequency distribution is obtained by adding successively the frequencies of all the previous intervals along with the interval against. This type of distribution is called cumulative frequency. The cumulative frequency distribution table is made up of two columns, the first is the exact upper-interval boundarys for each interval, and the secod is the frequencies less the the exact upper-interval boundary, as shown in Table 6.

Table 6: Cumulative frequency distribution of wages of employees as per data stated in Table 2

Rounded interval boundarys	Exact upper interval boundarys	Cumulative Frequency
————	1499.5>	0
5499 - 1500	5499.5>	4
9499 - 5500	9499.5>	9
13499 – 9500	13499.5>	13
17499 - 13500	17499.5>	18
21499 - 17500	21499.5>	20

1.2.2. Graphical Representation

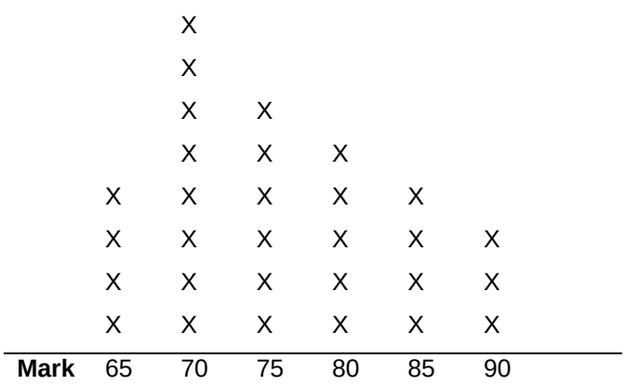
Statisticians have used graphical representations to better describe data, since graphic representation provides a quick description of a particular phenomenon by observation, without going into details; allows comprisons and draws some indicators and interpretations. The major graphical representation types are as follows:

1. Dot plot

A method used to present, summarize and represent data using dots, each dot on the vertical axis represents the frequency of the variable values. This representation tool is used in analysis to identify the characteristics of statistical data distribution and outliers or gaps in a data set.

Example: the following figure shows a dot plot of the marks of 30 students in science course, showing that the most frequent mark is 70 and the distribution of other marks as well, starting with 65, the frequency of marks is getting increased at 70, and decreased after 70 moving towards 90, which is least frequent.

Figure 1: Dot plot of the marks of students

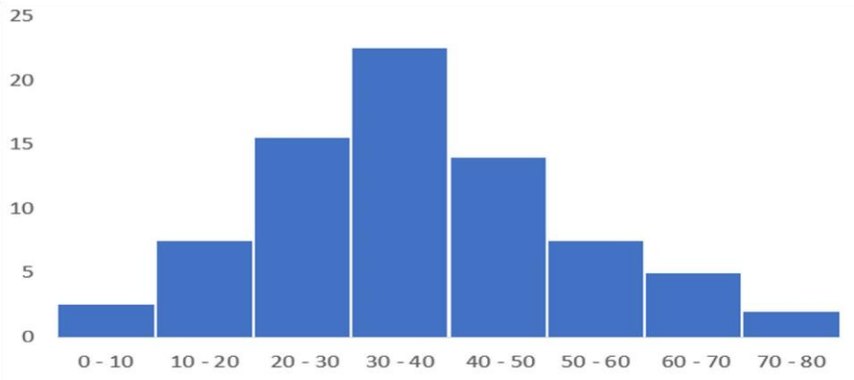


2. Histogram

A method used to present and summarize continuous data and identify the type and characteristics of the probability distribution of data, and depends on grouping data range, each group is represented by a column, where the column width represents the interval length, and the height represents the frequency of data values in this interval.

Example: the following graph shows the histogram of the ages of a population, where the column height represents the number of individuals in thousands, while the width of the column represents the interval of the age, e.g., individuals with the ages (30 – 40) are about 22000 in the population.

Figure 2: Histogram of ages



3. Stem and leaf plot

A method used to present quantitative data, and is widely used for data analysis, when available data are relatively few. Stem and leaf plot is somewhat similar to the histogram, and usually attached to the frequency table of data.

Example: the following figure presents the stem and leaf plot of the marks of 25 students. The stem represents the first two digits, and the leaves represent the other third digit, e.g. (the lowest) marks 55, 55, 56 and 59 are represented on the first stem by number (05); the leaves contained the numbers 9, 6, 5 and 5; while 100 is the top mark in the group and represented by three columns, the first two digits 10 are for stem and the last column is represented by the leaf.

Figure 3: Stem and leaf plot of the marks of a group of students

Stem	Leaves					
05	5	5	6	9		
06	2	3	5	6		
07	2	5	6	8	9	9
08	1	5	7	7	9	
09	2	3	5	6		
10	0	0				

4. Box and dot plot

A chart that shows distribution and spread of data, is used to detect outliers or inconsistent data. Box and dot plot start with determining the first quartile i.e. the data value under which 25% of data is found; the third quartile i.e. the data value under which 75% of data is found; the minimum value showing the first side of the plot; and the maximum value showing the other side of the plot.

Mid-quartile range is calculated by subtracting the third from the first quartile and dividing the result by 2.

Accordingly, the following is calculated as follows:

Minimum outliers = third quartile + (mid-quartile range × 1.5)

Maximum outliers = first quartile + (mid-quartile range × 1.5)

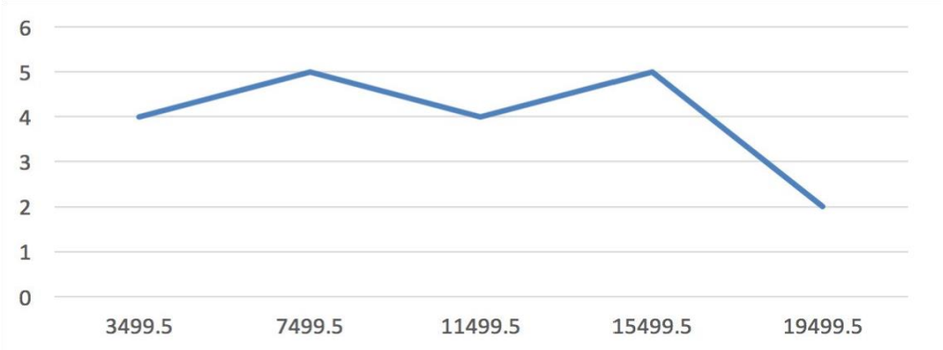
Data falling beyond upper and lower boundarys of outliers are called inconsistent data or -sometimes - outliers.

Figure 4: Box and dot plot

5. Polygon

Straight polygonal lines that represents the scale of the phenomenon in vertical and horizontal axes, A polygon can be drawn using Excel, where the horizontal axis represents the interval med points of the phenomenon under strudy and the vertical axis represents the interval frequency. Below the polygon for data given in the table (5) above.

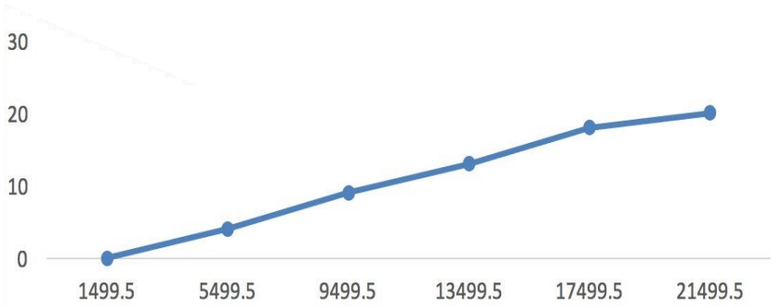
Figure 5: Polygon



6. The cumulative frequency curves

A curve describes a specific variable or phenomenon and can be drawn on a plane axes; the horizontal axis represents the exact interval boundaries, and the vertical axis represents the cumulative frequencies. Points shall be drawn above the exact lower interval boundarys so that the height represents the cumulative frequency. This is shown in the following chart, which represents the data of Table 6.

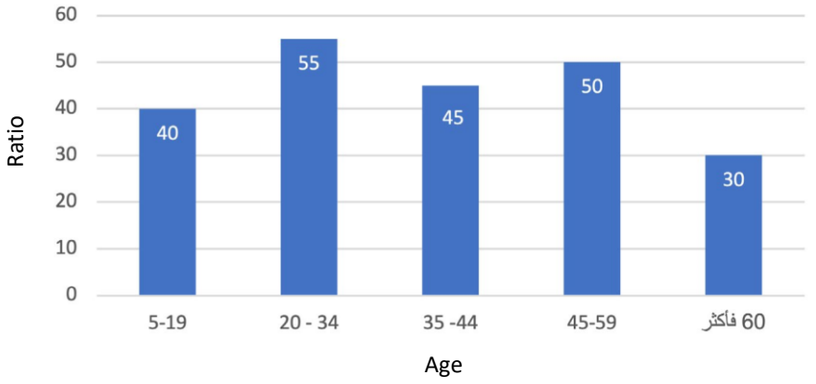
Figure 6: Cumulative frequency curve



7. Column charts

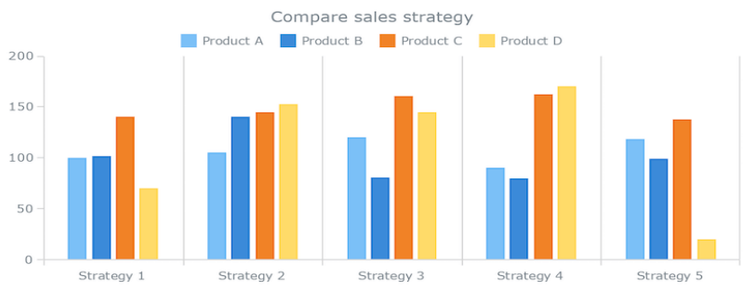
This method makes enabling read data and make comparisons between different values very easily, which also facilitates making different decisions based on observation. In this method, numbers are represented by columns whose length is proportional to the relevant value, so that the longest column represents the number of the higher value and vice versa. The variable to be represented by columns may include one dimension, represented by a column; this is called column charts. The following is a column chart showing the population distribution for each age group.

Figure 7: Distribution of individuals by age



The variable to be represented by columns may include more than one classification; each one is represented by a separate column within an interval, e.g. the number of products (A, B, C,D) in a strategy (1,2,3,4,5), the following is a column chart, showing the sales for products based on different strategies.

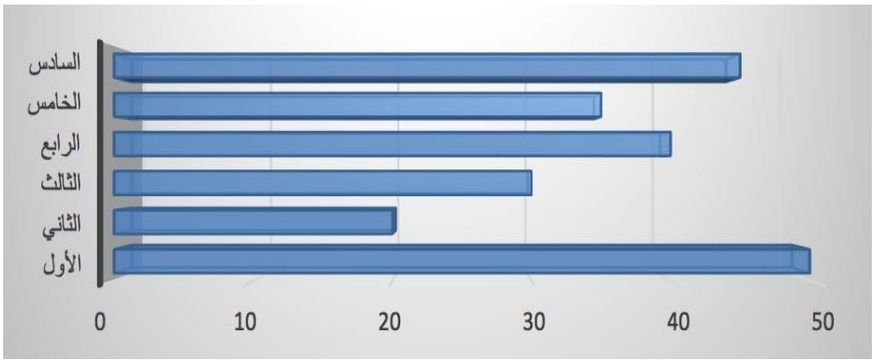
Figure 8: Sales of products by strategy



8. Bar charts

It is horizontal lines arrangerd in a specific order, easy to read and less confusing than column charts. The following is a bar chart, showing the number of students at a school for a year from Grade 1 to 6.

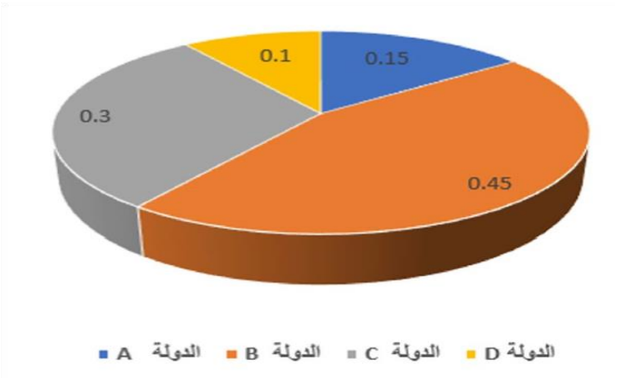
Figure 9: Number of students by grade



9. Pie chart

It is a circle divided into segments or sectors, used to show and place the relative importance of the population within different groups of the qualitative variable, widely used to represent data, because it is very easy to read. The following is a pie chart, showing the exports of a country, by country of destination.

Figure 10: Percentage distribution of exports by country



1.2.3. Images

One of the most common data representation methods, where the user enjoys interaction with the presented data, and is known for its high ability to enable the user to memorize the represented data for as long as possible, as people often prefer this method to receive data, which relies mainly on representing data visually preserving their connotation.

Chapter II

Statistical measures

The previous chapter has outlined the methods usd to present and summerize statistical data through frequency distribution tables or charts; to have some characteristics of the study population. However, such methods are not sufficient to describe data. Hence, numerical measures shall be provided to describe these data. This chapter covers two types of statistical measures: measures of central tendency and measures of dispersion. In this chapter, we will discuss the advantages and limitations of these measures in detail, which depend on the nature and the purpose of using data.

2.1. Measures of central tendency

Measures of central tendency, location measures or averages are numerical measures that localize data distribution; accordingly, the value that typically represents a specific data set and tends to fall in the center, hence averages are called measures of central tendency. They are important when comparing different distributions of data, and more useful in distributions similar in nature and shape, but different in locations. For example, while studying spending for a sample of rural and urban households, we can compare them using these measures. We will review the advantages and limitations of the most important measures of central tendency below.

2.1.1. Mean

A value around which a set of data gather, one of the most important measures of central tendency, the most widely used in statistics and practice, and usually used in many comparisons between different phenomena.

Mathematically, the simple mean is calculated by adding together all of the numbers in a data set and then dividing the sum by the total count of numbers, calculated by the following formula:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example 1: if the wages of 5 employees in a company are 250, 280, 320, 450 and 370 (USD), the simple mean is calculated as follows:

$$\bar{\mathbf{x}} = \frac{250 + 280 + 320 + 450 + 370}{5} = \mathbf{334}$$

2.1.1.1. Mean of grouped data (frequency tables)

If we have a number k of intervals midpoints (x_1, x_2, …, x_k) and frequencies (f_1, f_2,…, f_k), respectively, mean is calculated as follows:

$$\bar{\mathbf{x}} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_k x_k}{f_1 + f_2 + \cdots + f_k} = 1/n \sum_{i=1}^{k} f_i x_i$$

Example 2: calculate the mean age of students in the table below:

Age	5-6	7-8	9-10	11-12
No. of students	4	6	5	9

Solution: to simplify solving this problem, create the following table:

Age	Interval midpoint (x)	Frequency (f)	fi xi
5 - 6	5.5	4	22
7 - 8	7.5	6	45
9 - 10	9.5	5	47.5
11 - 12	11.5	9	103.5
Sum		**24**	**218**

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{n} = \frac{\sum_{i=1}^{n} x_i f_i}{n}$$

Hence the simple mean is equal to 218/24 = 9.1

2.1.1.2. Major charactaristics of the mean

- Easily calculated, subject to Algebraic operation, and it is the most common measure in statistics.
- Considers all values under study.
- Always between the lowest and largest value in the sample.
- The sum of the data deviations from the mean is always zero.

2.1.1.3. Some limitations of the mean

- Affected by outliers, extremely large or small values.
- Approximated in case of open frequency distribution tables, requiring identifying the medpoint of each interval.
- Cannot be calculated in case of nominals and catagorial data.

2.1.2. Median

The median defined as a value among the sorted data set where half count of the data values is lower than the median and the remaining data are larger than the median value, or in other words, it is the measure that equally separates the sorted data set in half.

2.1.2.1 Median of ungrouped data

To calculate the median for data in raw form (ungrouped) the following steps shall be taken:

- Sorting the data (observations) in ascending order.
- Determining the rank of the median, (n+1)/ 2, where n is the count of data values. If the count of data is odd, the median is the order observation in the middle, and if the count of data is even, the median is the average of the two orderd observations in the middle.

Example 3: calculate the median for the following data: 52, 15, 102, 68 and 44.

The data is arranged in an ascending order and ranked as follows:

value	15	44	52	68	102
rank	1	2	3	4	5

Using the data above, we can determine the median rank is (5+1)/2, it is 3. Accordingly, the median is the observation value with rank 3, it is 52.

Example 4: calculate the median of the following data: 52, 15, 102, 68, 44 and 72

The data is arranged in an ascending order and ranked as follows:

value	15	44	52	60	68	72	102
rank	1	2	3	3.5	4	5	6

Using the data above, we can determine the median rank as (6+1)/2, it is 3.5, between (3 - 4), so the median is the average of the two-observation ranked with 3 and 4 respectively, it is as follows:

$$\text{Median} = \frac{68 + 52}{2} = 60.0$$

2.1.2.2. Median for grouped data

If we have a number k of intervals with med points (x_1, x_2, …, x_k) and frequencies (f_1, f_2,…, f_k), respectively, median is calculated as follows:

- Create the cumulative frequency table using exact interval boundaries.
- Find the median rank.
- Find the median interval, whose frequency is higher or equivalent to the median rank.

Find the frequency (f1) of interval that preceeding the median interval, the length (L) of the median interval, and the exact lower boundry of the median interval (A).

The median is calculated as follows:

$$\text{Median} = A + \frac{(n/2 - f_1)}{f_2 - f_1} L$$

Example 5: calculate the median of student ages using example above:

Create the cumulative distribution table (as shown in the previous chapter):

Cumulative frequency (f)	Cumulative ages
0	<=4.5
4	<=6.5
10	<=8.5
15	<=10.5
24	<=12.5

Calculate Median rank as: (n/2) = 12, in the cumulative frequency column this value is between 10 and 15, hence:

$$A = 8.5, \quad f_1 = 10, \quad f_2 = 15, \quad L = 15 - 10 = 3$$

Apply the median formula:

$$\text{Med} = 8.5 + \frac{(12 - 10)}{15 - 10} \times 3 = 9.7 \quad \text{year}$$

2.1.2.3. Charactaristics of the median

- Not affected by outliers and can be found in case of catagorial data that can be sorted.
- The sum of absolute deviations is minimum when taken around the median.

2.1.2.4. Limitations of the median

- Does not take all values into account when calculated.
- Difficult to use in statistical and mathematical analyses.

2.1.3. Mode

The most frequent value in a data set, widely used in catagorial data to identify the most common pattern (level). A data set may have one mode and is, hence, called unimodal data set, or have more than one mode and is, hence, called multimodal data set. When a data set has no mode, it is called no mode data set.

Example 6: calculate the mode of 8, 6, 4, 2, 8 and 15
data has one mode = 8.

Example 7: calculate the mode of 8, 6, 4, 2, 8, 15, 9 , 4,and 12
data has two modes, are 4 and 8.

In case of grouped data or frequency distribution tables, one may not assume that a specific value is the most frequent, since values are integrated into various sets. Hence, there are modal interval that has the highest frequency.

2.1.3.1. Charactaristics of the mode

- Easily calculated and not affected by outliers.
- Can be found for catagrial data and open frequency distributions.

2.1.3.2. Limitations of the mode

When calculating the mode, not all data values are considered. Some data may have more than one mode; hence, a mode may not have a single value.

2.1.4. Geometric mean

It is a value that measures the central tendency of a data set, is calculated by taking the *n*th root of the product of the data set, where n is the number of data. GM of the values (x1, x2, …, xn) is:

$$GM = \sqrt[n]{x_1 \times x_2 . x_3 \dots x_n}$$

Example 8: calculate the geometric mean of 2 and 8
Square root of their product:

$$GM = \sqrt[2]{2 \times 8} = 4$$

Example 9: calculate the geometric mean of 1, 2, 4
Cubic root of their product:

$$GM = \sqrt[3]{1 \times 2 \times 4} = 2$$

2.1.4.1. Charactaristics and Limitations of the geometric mean

- Not affected by outliers.
- Cannot be used with data including negative or zero values.

2.1.5. Harmonic mean

It is usually used if data values are in ratios, or when the reciprocal of the data value is significant. The harmonic mean (H) of a set of values is the reciprocal of the mean of such values:

$$H = \frac{1}{\bar{x}}$$

Example 10: calculate the harmonic mean of 10, 7, 8, 6, 14, 9

$$\bar{x}_{\text{Non-tabulated}} = \frac{10+7+8+6+14+9}{6} = 9$$

$$H = \frac{1}{\bar{x}} = \frac{1}{9}$$

This example is for ungrouped data, however, the same formula above applies to grouped data, after calculating the mean for grouped data.

2.2. Measures of dispersion

Numerical measures used to measure the degree of homogeneity (closeness) or dispersion (divergence) of data items. Measures of dispersion are used to describe a data set and compare different data sets, as measures of central tendency alone are not sufficient to describe a data set or compare different data sets. Among the most popular measures of dispersion are:

2.2.1. Range

One of the simplestt measures to define and calculate, it provides a quick idea of data dispersion, and has the symbol (R). Range of a set of data is calculated by the following formulas:

2.2.1.1. In case of ungrouped data:

Range (R) = largest value - smallest value

Example 11: calculate the range of 54, 89, 65, 70, 95, 47

R = 95 - 47 = 84

2.2.1.2. In case of grouped data, 2 methods are available:

Range (R) = upper interval boundary - lower interval boundary.

Range (R) = highest interval midpoint - lowest interval midpoint.

Example 12: calculate the range of ages in the table below:

Find interval boundarys and mid points as shown in the previous chapter covering data tables.

Age	15 - 6	25 - 16	35 - 26	45 - 36	55 - 46	65 - 56
Exact interval boundarys	15.5 - 5.5	25.5 - 15.5	35.5 - 25.5	45.5 - 35.5	55.5 - 45.5	65.5 - 55.5
Interval medpoints	10.5	20.5	30.5	40.5	50.5	60.5
Frequency (f)	10	16	14	6	9	5

As per the first method:

(R) = 65.5 - 5.5 = 60

As per the second method:

(R) = 60.5 - 10.5 = 50

Noting that both methods are different, the first method is often used to calculate range.

2.2.1.3. Charactaristics of the range

- Simple to define and calculate.
- Highlights the nature of data and is widely used in various life phenomena, e.g. quality control and weather forecasting.

2.2.1.4. Limitations of the range

- Calculated based on 2 values of data and does not consider the other values.
- Affected by outliers, hence, does not reflect an accurate view of the data nature which makes it an approximate measure.

2.2.1.5. Mid-quartile range

The mid-range between the first quartile (Q_1) and the third quartile (Q_3), has the symbol Q and calculated by the following formula:

$$Q = \frac{Q_3 \mp Q_1}{2}$$

Where Q_1 is the value under which 25% of data points are found when they are sorted in increasing order, and Q_3 the value under which 75% of data points are found when sorted in increasing order.

2.2.1.6. Calculation of mid-quartile range for ungrouped data

- Arrange data in an ascending order.
- Find Q1.
- Find Q3.
- Apply the formula above.

Example 13: calculate the mid-quartile range of 53, 89, 65, 70, 95, 47, 74, 86

Arrange data in an ascending order: 47, 53, 65, 70, 74, 86, 89, 95

Q_1 of data above is the 2nd value, as the quartile rank is the product of n values and the quartile percentage 25%, 50% or 75%, accordingly:

$$Q_1 = X_{(2)} = 53 \quad , \quad Q_3 = X_{(6)} = 86 \quad , \quad Q = \frac{86+53}{2} = 69.5$$

2.2.1.7. Mid- quartile range for grouped data

The mid-quartile range for such data is calculated using difference method. The first and third quartiles are calculated by the formulas below previously explained to calculate the median, with a slight difference by adding n/2 when calculating Q_3:

$$Q_1 = A_1 + \frac{(n/4 - f_1)}{f_2 - f_1} L$$

$$Q_3 = A_3 + \frac{(3n/4 - f_3)}{f_4 - f_3} L$$

Where A_1 is the exact boundary of the interval preceeding Q_1 interval, A_3 is the exact boundary of the interval preceeding Q_3 interval, L is the interval length = upper interval boundray - lower interval boundary, f_1 is the cumulative frequency preceding Q_1 ranks, f_3 is the cumulative frequency preceding Q_3 ranks f_2 is the cumulative frequency following Q_1 rank and f4 is the cumulative frequency following Q_3 rank

Example 14: find the mid-quartile range (Q) for the ages in example 12.

Create the cumulative frequency distribution as shown below, then calculate Q using the formulas above.

Less than cumulative frequency distribution	Interval boundarys
0	5.5
10	15.5
26	25.5
40	35.5
46	45.5
55	55.5
60	65.5

n = 60, $\frac{n}{4} = 15$, $\frac{3n}{4} = 45$, L = 10

$$Q_1 = 15.5 + \frac{(15 - 10)}{26 - 10} 10 = 18.6$$

$$Q_3 = 35.5 + \frac{(45 - 40)}{46 - 40} 10 = 43.8$$

$$Q = \frac{18.6 + 43.8}{2} = 31.2$$

2.2.1.8. Charactaristics of mid-quartile range

- Not affected by outliers.
- Can be calculated based upon open-ended frequency distributions.

2.2.1.9. Limitations of mid-quartile range

- Does not take all values into account when calculated.
- Hard to explain it in statistical analysis.

2.2.2. Mean deviation

The average of absolute deviations of data from the mean, has the symbol of MD.

2.2.2.1. Mean deviation for ungrouped data

Calculated by applying the following formula:

$$MD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

Where x_i is observations and n is the number. of observations.

Example 15: Calculate the mean deviation of 5, 9, 7, 14, 11, 8, 12, 6.

After calculating the simple mean, $\bar{x} = 1/n\sum x = 72/8 = 9$

create the following table:

x	$x - \bar{x}$	$\lvert x - \bar{x} \rvert$
5	-4	4
9	0	0
7	-2	2
14	5	5
11	2	2
8	-1	1
12	3	3
6	-3	3
72	0	20

Using the data of the table above and the ungrouped mean deviation formula:

MD = 20/9 = 2.22

2.2.2.2. Mean deviation for grouped data

Calculated by applying the following formula:

$$MD = 1/n\ \sum_{i=1}^{n} f_i|x_i - \bar{x}|$$

Where X_i is the interval medpoint, f_i is interval frequency, and n is sum of frequencies.

Example 16: Calculate the mean deviation of ages in example 15.

After calculating the mean $\bar{x} = 1/n \sum_{i=1}^{n} fx = 1860/60 = 31$

create the following table:

Interval mark (x)	Frequency (f)	fx	x - $\bar{x}$	| x - $\bar{x}$ |	f | x - $\bar{x}$ |
10.5	10	105	-20.5	20.5	205
20.5	16	328	-10.5	10.5	168
30.5	14	427	-0.5	0.5	7
40.5	6	243	9.5	9.5	57
50.5	9	454.5	19.5	19.5	175.5
60.5	5	302.5	29.5	29.5	147.5
Sum	**60**	**1860**			**760**

Using the data of the table above and the grouped data mean deviation formula:

MD **=** 760/60 **=** 12.7

2.2.3. Standard deviation

One of the most important and best measures of dispersion and the most common and widely used method in statistical analysis. As standard deviation is the square root of data variance. Variance is the mean of the squares of the deviations from the simple mean and has the symbol of V. Variance depends on the dispersion or divergence of data from the simple mean and is large if data is far from the mean and vice versa. Standard deviation can be calculated using the following formula:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_{i-\bar{x}})^2}{N}}$$

Where x - $\bar{x}$ is data deviations from the mean and N is the number of data items.

Standard deviation is the square root of variance and has the symbol of σ. As in case of variance, a value increase indicates a significant degree of dispersion or fluctuation and divergence of data, and the contrary is true in case of a value decrease. Using variance, standard deviation of a statistical sample can be calculated using the following formula, where n is the sample size.

$$s^2 = \frac{\sum_{i=1}^{n}(x_{i-\bar{x}})^2}{n-1}$$

2.2.3.1. Standard deviation for ungrouped data

In case of a sample, whose size is n, taken from a population, deviation has the symbol of S^2 and is calculated using the following formula:

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

Example 17: calculate the standard deviation of ages of 6 primary education students (x): 5, 8, 6, 9, 7, 10.

Create the table below after calculating the mean:

$\bar{x} = {}^{1}/_{n} \sum x = {}^{45}/_{6} = 7.5$

(x)	(x - $\bar{x}$)	(x - $\bar{x}$) ^{2}
5	-2.5	6.25
8	0.5	0.25
6	-1.5	2.25
7	-0.5	0.25
9	1.5	2.25
10	2.5	6.25
Sum	**0**	**17.5**

Using the data of the table above and the ungrouped data variance formula:

$$s^2 = \frac{17.5}{6-1} = 3.5$$

Standard deviation is calculated using variance and the following formula:

$$s = \sqrt{3.5} = 1.871$$

2.2.3.2. Standard deviation for grouped data

If we have a number k of interval medpoints (x_1, x_2, …, x_k) and frequencies (f_1, f_2,…, f_k), respectively, sample variance is calculated as follows:

$$s^2 = \frac{\sum_{i=1}^{n} f_i(x_{i-\bar{X}})^2}{n-1}$$

Standard deviation is calculated using variance and the following formula:

$$s = \sqrt{\frac{\sum_{i=1}^{n} f_i(x_{i-\bar{X}})^2}{n-1}}$$

Example 18: calculate standard deviation for ages in example 10.

Create the table below after calculating the mean:

$$\bar{x} = 1/n \sum_{i=1}^{n} fx = 1/60 \times (1860) = 31$$

Interval medpoint (x)	Frequency (f)	fx	$x - \bar{x}$	$\lvert x - \bar{x} \rvert$	$f \lvert x - \bar{x} \rvert$
10.5	10	105	-20.5	420.25	4202.5
20.5	16	328	-10.5	110.25	1764
30.5	14	427	-0.5	0.25	3.5
40.5	6	243	9.5	90.25	541.5
50.5	9	454.5	19.5	380.25	3422.25
60.5	5	302.5	29.5	870.25	4351.25
Sum	**60**	**1860**	**-**	**-**	**14285**

Using the data of the table above and the grouped data variance formula:

$$s^2 = 242.12$$

Standard deviation is calculated as a square roat of variance and given by:

$$s = 15.56$$

2.2.4. Standard value

Measure deviations from the mean using standard deviation units and has the symbol of Z. if we have variable X having values (x_1, x_2, ..., x_n), mean $\bar{x}$, and standard deviation S, Z is calculated using the following formula for the standard value of xi :

$$Z_i = \frac{x_i - \bar{x}}{S}$$

Example 19: if a student gets agrade 86 in accounting, where mean is 77, standard deviation is 11, and he gets agrade 96 in Economics, where mean is 84, standard deviation is 17, in which course did the student perform better? Standard value of both courses is calculated using the formula above:

Standard value of accounting:

$$Z_i = \frac{87 - 77}{11} = 0.82$$

Standard value of Economics:

$$Z_i = \frac{96 - 84}{17} = 0.7$$

The results show that the student did better in accounting than in Economics, although the mark of accounting is less.

2.3. Skewness

The degree of asymmetry or deviation from the symmetry of a distribution. If the data distribution curve has a longer tail to the right of the central maximum than to the left, the distribution is said to be skewed to the right or to have positively skewed. If the contrary is true, the distribution is said to be skewed to the left or to have negatively skewed.

There are many methods to measure skewness of frequency distribution or a data set, such as the following formulas:

$$\mathrm{Sk} = \frac{3(\bar{x} - \mathrm{M})}{\mathrm{S}}$$

$$\mathrm{Sk} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{S^3\,(N-1)}$$

Where $\bar{x}$ is mean, M is median, S is standard deviation and X_i is variable values.

This relative measure is negative if skewed to the left, and positive if skewed to the right. Distribution extends from (-3) if negatively skewed to (+3) if positively skewed and becomes zero when the mean and median are equal, when distribution is normal.

2.4. Kurtosis

A measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution, and a symmetric curve centered around the mean. If distribution is heavy tailed (greater than normal distribution), it is said to be leptokurtic. If distribution is flat, it is said to be platykurtic. If distribution is semi-heavy tailed (not leptokurtic or platykurtic), it is said to be mesokurtic. Kurtosis is not related to the distribution mean, there may be many distributions sharing the same mean but differ in terms of leptokurtic or flat curves.

Since the height of the peak of the normal distribution is approximately 3, distribution is platykurtic when the kurtosis factor is less than 3, while distribution is leptokurtic when the kurtosis factor is more than 3. The kurtosis factor is calculated using the following formula:

$$\mathrm{SK} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{S^4\,(N-1)}$$

Chapter III

Statistical estimation and hypotheses testing

The study of the characteristics of any statistical population depends on the nature and method used to deal with its members. When a census of population is done, characteristics of the population are studied by identifying statistical indicators of the distribution of the population in the light of the values and characteristics of its parameters, including the mean, median, average, standard deviation, etc. A population can be studied by taking a sample from the population members. The characteristics of a population are studied by conducting a statistical estimation for the parameters from the data of the selected sample. Each estimator is called statistic. However, the decision to accept and adopt estimators to study the characteristics of the population from which a sample has been taken is related to an evaluation of such estimators, because estimation of the statistical indicator based on the sample is not equal the parameter of the population. Differences between these two indicators refer to the estimation errors based on the selected sample.

Statistical estimation aims to find the best estimator for the parameters of a population. Testing statistical hypotheses involve building methods that depend on data under study to decide on a hypothesis formulated before dealing with the sample data. However, the distinction between estimation and testing is not reflected by separation of such two processes, they are interrelated, which necessitates a presentation thereof.

3.1. Estimation

Estimation is associated with a group of statistical problems, dealt with using inference that leads to accurate perceptions, as much as possible, to study one or more values of the population parameters. Estimation is either by seeking to obtain a specific point estimate derived from data of a population sample, trying to make it as close as possible to the real value of the parameter, or by calculating boundarys within which the real value of the parameter is likely expected to fall. The higher the probability, the greater the reliability in obtaining the true value of the parameter within confidence range.

3.1.1. Point estimation

A procedure pursuant to which the estimator $\hat{\theta}$ is adopted for the population parameter Θ, the subtraction, expressed by $(\hat{\theta} - \theta)$, is not adopted to judge the accuracy of estimation, sometimes replaced by $(\hat{\theta} - \theta)^2$ to get rid of the effect of the sign in subtractions, or an absolute value, where an estimator - making the expected value of the subtraction squares between the value of the parameter and the estimator as low as possible – is selected, the estimator is then called MSE estimator.

Example 1: If we have Abu Dhabi population, the target is to find the indicator of per capita expenditure in the Emirate, which requires comprehensive data for all members of the population of Abu Dhabi, who are to be asked about per capita expenditure, which means conducting a comprehensive survey of all members, leading to high costs and long time. The result of the indicator, per capita expenditure, will not be as accurate as required, due to the large size of data collected, the large number of field teams, and different chances of errors that will ultimately be reflected in the value of the population parameter.

A proper alternative to a comprehensive survey is a sampling survey by selecting a sample of families from the emirate, collecting the expenditure data for each member of the sample, calculating and

considering per capita expenditure, $\hat{\theta}$, as an estimation of per capita expenditure of the emirate's population, Θ.

The question is how to measure the accuracy of the estimated value of the indicator is. The estimator includes a specific error percentage, from two main sources; the sampling error and nonsample errors which cannot be measured but can be minimized by adjusting data collection and processing procedures.

Sample error may be measured based on standard deviation value of the data of a simple random sample, including n number of units, by calculating the so-called sampling error:

$$SE = \frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N}},$$

where N is the overall size of a population.

Example 2: If the population parameter in the household average size to be estimated. A sample of 5,000 households is taken and data thereof is obtained. The household average size is calculated based on the sample data; the estimate value is 6.4.

Standard deviation of the sample data is calculated, accordingly, S = 2.121. Sampling error is calculated by dividing standard deviation by the square root of the sample size, where $\left(\frac{N-n}{N}\right)$ is ignored, beign too small when the size of a sample is large.

Accordingly, sampling error is:

$$SE = \frac{2.121}{\sqrt{5000}} = 0.03$$

3.1.2. Interval estimation (confidence interval)

As discussed, interval estimation is done by determining interval boundaries for the confidence interval, depending on the sample data, where the population parameter is expected to be included in the interval at a predetermined level of confidence, e.g. 95%, 90%, or otherwise.

Determining both lower and upper boundaries for confidence interval requires assuming that data are normally distributed and/or the size of the sample is relatively large.

Accordingly, the confidence interval boundaries are determined by adding and subtracting the value “w “as:

$$w = z_{(1-\alpha/2)}\frac{S}{\sqrt{n}}$$

Where $Z_{1-\alpha/2}$is standard normal distribution value of the previously set confidence level. For example, if confidence level is 95%, $\alpha = 5\%$, and the constant $z\ (1-\alpha/2)$ using normal distribution, is 1.96. If confidence level is 90%, the value is 1.654. While if confidence level is 85%, the value is 1.28, where w is the margin error in confidence interval.

Accordingly, confidence interval is calculated by:

$$\left[\hat{\theta} - z_{(1-\alpha/2)}\frac{S}{\sqrt{n}}\ ,\ \hat{\theta} + z_{(1-\alpha/2)}\frac{S}{\sqrt{n}}\right]$$

It can be said that the value of the population parameter is expected to be within this interval with confidence level of $(1-\alpha/2)\%$.

Example3: based on the example above, if the average size of household is required at a confidence level of 95%, the bound of error is calculated as follows:

$$w = \frac{1.96 \times 2.121}{\sqrt{5000}} = 0.0588$$

Then the confidence interval is calculated as follows:

$$[6.4 - 0.0588\ ,6.4 + 0.0588]$$

Or

$$[6.34\ ,6.46]$$

Example 4: A food company produces a kind of juice, the weight of the bottle is 125 gm. If the production control manager takes a random sample of 36 bottles, measures the quantity of carbohydrates in gm, and finds the average amount of carbohydrates is 12 gm and standard deviation is 2.4 gm, if the production control department wants to estimate a confidence interval of 95% for the average amount of carbohydrates in the bottles, and the quantity of carbohydrates is normally distributed, the margin error is:

$$\mathrm{w} = \frac{\mathrm{z}_{1-\alpha/2}\mathrm{s}}{\sqrt{\mathrm{n}}} = \frac{1.96 \times 2.4}{\sqrt{36}} = 0.784$$

Hence, confidence interval is: [11.22, 12.78]
Accordingly, we are 95% confident that carbohydrates quantity ranges between (11.22 and 12.78).

Note: If the population data is not subject to normal distribution or is small (less than 30 observations), to calculate confidence interval, we assume that data is subject to student distribution (t) with n-1 degrees of freedom based on the table values of distribution (t) at a certain confidence level, (t) value can be obtained from the relevant statistical table and change z by t- value at the predetermined confidence level (1- α).
in the formula above,

$$\left[\hat{\theta} - \mathrm{t}_{(1-\alpha/2)\frac{\mathrm{s}}{\sqrt{\mathrm{n}}}}\ ,\ \hat{\theta} + \mathrm{t}_{(1-\alpha/2)\frac{\mathrm{s}}{\sqrt{\mathrm{n}}}}\right]$$

3.1.3. Statistical estimation of unweighted data

In this section, unweighted data are any data that does not all have the relative or the same importance in the population to which it belongs, the difference is a result of the difference of the phenomenon the data represents within the same population.

When data differs in importance in the same population, the indicator value cannot be directly calculated using this data, i.e. by dealing with all data as having the same importance.

For example, if the observation values of X_i of a population with weighte values w_i respectivly, the weighted indicator Ɵ, whether a mean or percentage, is:

$$\theta = \frac{\sum \mathrm{w_i x_i}}{\sum \mathrm{w_i}}$$

For example, if a sample of households in a number of cities is taken to fully identify the average household expenditure. Not all households have the same weight or importance, and therefore to find the average expenditure of a household, the average expenditure data of all households in all cities is not simple average. However, a variable size or weight of the city, where the household is, plays an influential role in the size of expenditure. Accordingly, in this case the averge is calculated as follows:

$$\overline{X} = \frac{\sum w_i \overline{x_i}}{\sum w_i}$$

Where:

$\overline{x_i}$ is the average observation in a population group (i).

w_i is the relative weight or importance of the population group (i).

If the number and average size of households in a number of cities are as in the table below, how can we calculate the overall weighted household's average size?

City	The household average size (x)	No. of households (w)	X. W
A	5.5	1800	9900
B	6.3	2500	15750
C	4.7	3000	14100
D	5.8	800	4640
E	7.0	1200	8400
Sum		**9300**	**52790**

the overall household average size is gevin by:

$$\overline{X} = \frac{\sum w_i x_i}{\sum w_i} = \frac{52790}{9300} = 5.7$$

The foregoing also applies to the percentage or average indicator. If we have a number of partial populations and the variable percentage in each part is different, then relative weights or importance of data will be different, and the estimated percentage as a statistical indicator at the level of the population as a whole is:

$$p = \frac{\sum w_i p_i}{\sum w_i}$$

Where p is the percentage or ratio at the overall population level, p_i is the percentage or ratio of variable values in part (i) of a population.

Example 5: If the percentage of infection with a particular disease among population groups differs by gender as in the following table, then the total percentage of infection with the disease at the population level as a whole is calculated as follows:

Gender	Percentage of infection (p)	No. of individuals	X. W
Males	20.4	2800	57120
Females	8.7	5200	45240
Sum		**8000**	**102360**

In this example, the total percentage of infection with the disease in the population is:

$$p = \frac{\sum w_i p_i}{\sum w_i} = \frac{102360}{8000} = 12.8$$

The method used to estimate the indicator of the sum τ instead of the average or percentage is based on the indicators of the sum in different parts of the population (i).

$$\tau = \sum \tau_i = \sum N_i \bar{x}_i$$

Where, τ_i is the total indicator values of part (i) of a population
N is the overall number of a population unit.

Example 6: A sample of economic establishments is taken from each city to estimate total revenues at the population level, noting that the relative importance represented by the number of establishments in cities is not the same; there shall be relative weights at the level of each city.

City	The size of the sample of establishments (n)	Total revenues (x)	Total number of establishments N	Average revenues per establishment	Total revenues
a	30	15000	150	500	75000
b	20	10000	90	500	45000
c	10	8500	60	850	51000
Sum	**60**	**33500**	**300**		**171000**

Based on the table and using the formula above, total revenues is the sum of the last column in the table.

$$\tau = \sum \tau_i = \sum N_i \bar{x}_i = 150 \text{ X } 500 + 90 \text{ X } 500 + 60 \text{ X } 850 = 1710$$

3.2. Hypotheses testing

3.2.1. Hypotheses testing concept

The methods used to study unknown population parameters have been initially addressed by using confidence intervals as supporting information in decision-making, as data of a random sample of the population whose parameters are to be estimated are used to calculate the required confidence interval at $(1-\alpha)$ confidence level. Accordingly, the result obtained by confidence intervals can often be drafted as follows:

We are $(1-\propto)$ confident that the confidence interval will contain the unknown true value of the population parameter.

It is noted that the confidence interval is calculated based on the data of a random sample, to be used in statistical inference about the true value of a population parameter. However, in practice, there is often a previous claim of the value of the unknown parameter, which does not necessarily have to be related to a specific value and can have a mathematical –relation such as the parameter value is less than or greater than a specific value. In this case, the aim of the statistical inference is more specific than in the calculation of a confidence interval, focusing on verifying the credibility thus making a decision to accept or reject the claim.

Dealing with and judging the credibility of hypotheses is called hypotheses testing. There is a relationship between calculating a confidence interval and hypotheses testing, as it can be said that hypotheses testing gives information more used in decision-making than information obtained from calculating confidence intervals. However, confidence intervals may be relied on in some cases to decide on whether a hypothesis is valid.

If a hypothesis is tested, a claim or assumption is to be tested. At first, the claim is assumed to be invalid, then the study data is used to prove the opposite; the claim is vaild. This mechanism gives hypotheses testing a strength stemming from avoiding bias and inaccuracies, as the poor study and data collection is in the interest of reversing the claim, then a claim can only be accepted if there is a strong statistical indication or evedence thereof.

3.2.2. Types of statistical hypotheses

Statistical hypotheses include two types:

- Alternative hypothesis (H_a)

  What a researcher needs to prove valid, recommended in many cases.

- Null hypothesis (H_0)

  What a researcher needs to prove invalid.

There are three ways to formulate alternative hypotheses:

Alternative Hypothesis (H_a)	Null Hypothesis (H_0)
Ha :μ ≠μo	*H*o :μ=μo
Ha :μ > μo	*H*o :μ≤μo
Ha :μ<μo	*H*o :μ≥μo

Where μo is the population parameter based on null hypothesis.

It is noted that null hypothesis is always accompanied by =, accordingly it can be written as follows:

Ho :μ=μo

There are 4 possibilities of decisions which may be made on null hypothesis:

Decision	Valid H_o	Invalid H_o
Rejecting H_o	α	(1 – β)
Accepting H_o	(**1 –** α)	β

Possibilities:

1. P (reject H0 when H0 is correct)=α (the probability of a type I error).
2. P(accept Ho when Ho is correct) = (1−α).
3. P(accept Ho when Ho is incorrect) =β.
4. P(reject Ho when Ho is incorrect) = (1−β).

Type I error is controllable and detected by the researcher prior to the test, its probability is called the significance level (α), most values used are 0.05, 0.01.

As a result, it can be said that in order to set up a null and an alternative hypothesis mathematically, several conditions shall be met:

- The unknown parameter to be tested shall firstly be identified as it may be a population average, the difference between two averages, the probability of an event in a population, the difference between two percentages, a population variance or the ratio of 2 variations.
- The value of the unknown parameter related to the claim to be tested shall be found.
- The relation between the parameter and relevant value shall be determined and in 1 of 3 forms; >, < or =, and the alternative hypothesis representing the claim.
- Null hypothesis shall be formulated, as it includes the parts of the alternative hypothesis with the alteration of the mathematical relation between the parameter and the relevant value with changing the sign of the inequality to reflect the corresponding case of the alternative hypothesis, and thus to represent the opposite of the claim.

Example 1: find input of the claim (the parameter, the relevant value, and the mathematical relation), then formulate the null and the alternative hypotheses.

- The claim made by the director of an economic department that the average time required for the maintenance of any machine is less than 12 hours.
- The claim made by a national battery factory that the average battery age produced by the factor is more than 1.5 years.
- The claim made by a researcher that the percentage of students receiving academic warnings at Khalifa University is less than 0.30 out of the total number of students.
- The claim made by an investor that the average profit generated by investment at Abu Dhabi Securities Exchange is not (different than) 0.10.

Solution (1):
Parameter: average time needed for the maintenance of a machine (μ)
Relevant value: 12 days
Mathematical relation: less than (<)
Hypotheses:
H_0 :μ ≥ 12
H_a :μ < 12

Solution (2):
Parameter: average battery age (μ)
Relevant value: 1.5 years
Mathematical relation: more than (>)
Hypotheses:
H_0 :μ≤ 1.5
H_a :μ > 1.5

Solution (3):
Parameter: percentage of student receiving academic warnings (P)
Relevant value: (0.30)
Mathematical relation: Less than (<)
Hypotheses:
H_0 : P ≥ 0.3
H_a : P < 0.3

Solution (4):
Parameter: the percentage profit generated by stock investment (P)
Relevant value: 10%
Mathematical relation: not equal (≠)
Hypotheses:
H_0 : P = 0.1
H_a : P ≠ 0.1

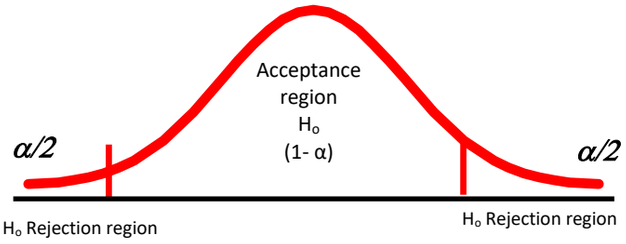
The a forementioned may be summarized as follows:

- Type I error is the rejection of an incorrect hypothesis, its probability is called the significance level (α).
- Type II error is the acceptance of an incorrect null hypothesis, its probability has the symbol β.
- Confidence level (1- α-1) is the probability of accepting a correct null hypothesis.
- Power of a test (1 - β) is the probability of rejecting an incorrect null hypothesis.

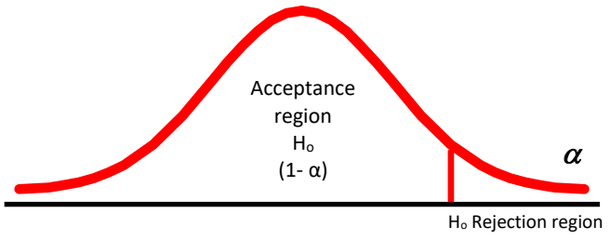
3.2.3. Types of hypothesis tests

There are two types of hypothesis tests determined based upon the alternative hypothsis type as follows:

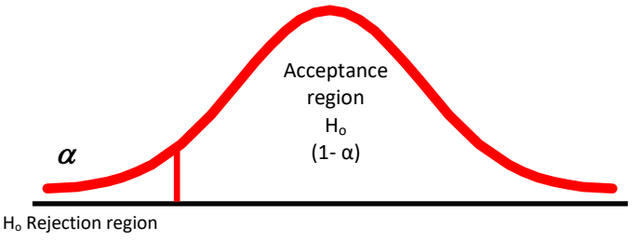
- Two-tailed test (if H_a: $\mu \neq \mu o$), the rejection region is at both ends of the curve.



- One-tailed test, all rejection regions α are at the end of the right or left curve:
- If H_a: $\mu > \mu_o$, the rejection region is at the right end of the curve, as shown below:



- If H_a :$\mu < \mu_o$, the rejection region is at the left end of the curve, as shown below:

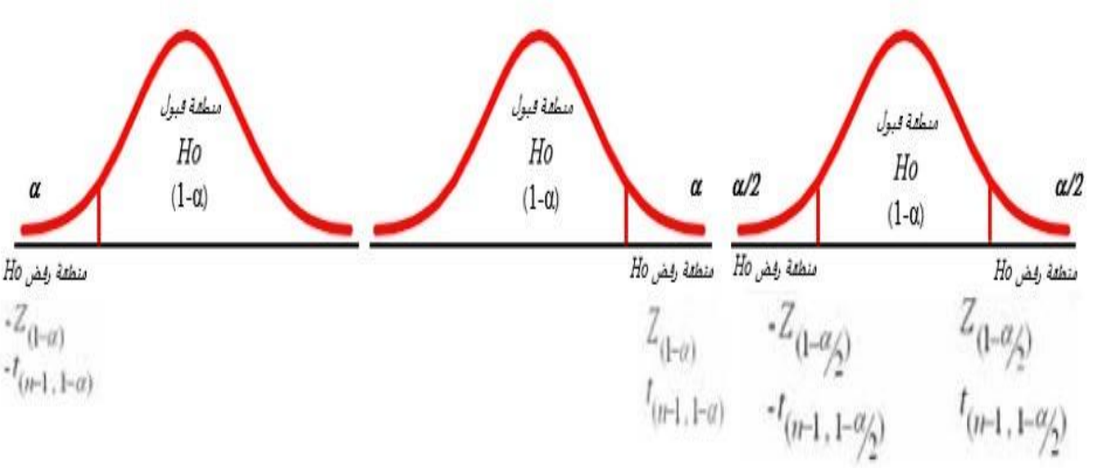


Steps of hypothses testing:

1. Formulation of hypotheses:

Alternative hypothesis H_a	Nul hypothesis H_o
H_a :$\mu \neq \mu o$	H_o :$\mu = \mu o$
H_a :$\mu > \mu o$	H_o :$\mu \leq \mu o$
H_a :$\mu < \mu o$	H_o :$\mu \geq \mu o$

2. Identification of significance level α, sample distribution, and acceptance and rejection regions: Sampling distribution, either a standard normal or (t) distribution with (n-1) degrees of freedom. Critical values, which define acceptance or rejection regions, are extracted from tables as shown in the following figure:



3. Calculation of test statistic:

   Using the sample data and the population average if H_o: μ = μo, a value called "test statistic" or the calculated value can be calculated and found according to whether the population variance is known as shown in the table below

Sample size	Test statistic (calculated value)	population variance
No specific sample size is required	$Z^* = \frac{(\bar{x} - \mu_o)}{\frac{\sigma}{\sqrt{n}}}$	Known
n < 30	$t^* = \frac{(\bar{x} - \mu_o)}{\frac{S}{\sqrt{n}}}$	Unknown
n >30	$t^* \sim Z^* = \frac{(\bar{x} - \mu_o)}{\frac{S}{\sqrt{n}}}$	Unknown

4. Deciding on null hypothesis:

If the calculated value (step No. 3) is in the rejection region, the null hypothesis is rejected, i.e. the alternative hypothesis is accepted.

Chapter IV
Correlation and simple linear regression

4.1. Correlation

Correlation analysis describes the strength of an association between two or more variables; correlation measures how much the variable values change in a regular manner. Correlation is a quantitative indicator used to determine the degree of dependence on one or more variables to predict the values of another variable. It is important to know what correlation analysis can and cannot provide. Correlation analysis neither provides any information to predict the values of a variable, nor any indication of a casual relationship between variables. However, the analysis can only determine if the degree of covariance is significant. Therefore, the relationship between the two phenomena or variables is called correlation. The correlation may be direct; the two phenomena change in the same direction, so that if a phenomenon increases, the other tends to increase, and vice versa. Correlation may be inverse; the two phenomena change in opposite directions, so that if a phenomenon increases, the other tends to decrease, and vice versa.

It is noted that the value of the correlation coefficient is a relative numerical value between +1 and -1, this value is not +1 and -1 unless correlation is complete.

4.1.1. Correlation types

Correlation is divided into various types which differ by the type of variable to be measured, as there are quantitative and qualitative variables are to be measured.

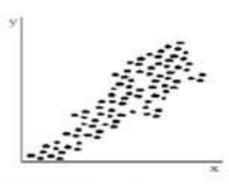
4.1.1.1. Correlation coefficient of in quantitative variables

Includes studying the relationship between measured phenomena, which can be measured quantitatively or numerically, including all phenomena that can be expressed numerically, such as height, income, production quantity, etc, and is divided into several types.

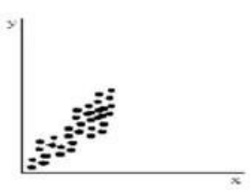
Simple correlation coefficient (Pearson coefficient)

A correlation coefficient that determines the amount, extent and direction of the relationship between two variables, in cases or phenomena in which the study is limited to two variables, e.g. if the extent and direction of the relationship between heights and weights of a group of people, or between the amount of monthly income and spending of households in a given population is required.

Correlation has several types that can be identified through both the amount of the correlation coefficient and the direction of the relationship between the two variables based on the scatter plot.

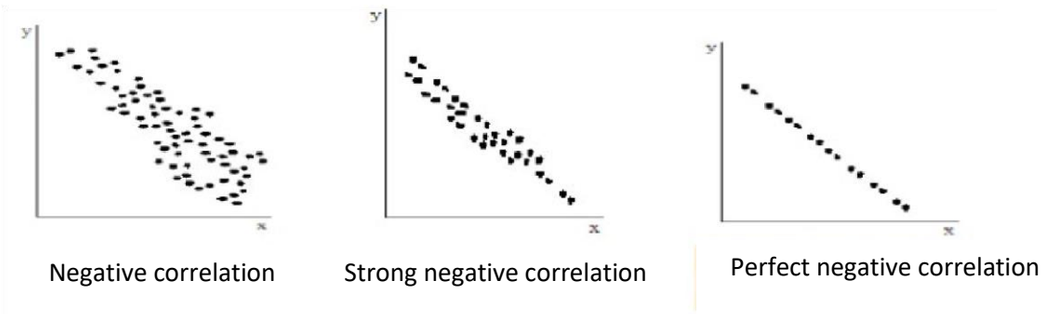


Positive correlation



Strong positive correlation

Perfect positive correlation



Negative correlation | Strong negative correlation | Perfect negative correlation

The following table summarizes correlation types and relationship directions between two variables:

Correlation coefficient value	Correlation relationship type
Perfect positive correlation	+1
Strong positive correlation	From 0.70 to 0.99
Moderate positive correlation	From 0.50 to 0.69
Weak positive correlation	from 0.01 to 0.49
Non-linear correlation	0

Likewise, at the same level, the correlation relationship is inverse if correlation coefficient is negative. The correlation coefficient between the values of two variables X and Y can be calculated using the following formula:

$$r_{xy}^2 \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{((n\sum x^2 - (\sum x)^2)((n\sum y^2 - (\sum y)^2))}}$$

Where:

$\sum xy$ is the product of the values x and y.

$\sum x$ is the total values variable X.

$\sum y$ is the total values variable Y.

$\sum x^2$ is the sum of squares of variable X.

$\sum y^2$ is the sum of squares of variable Y.

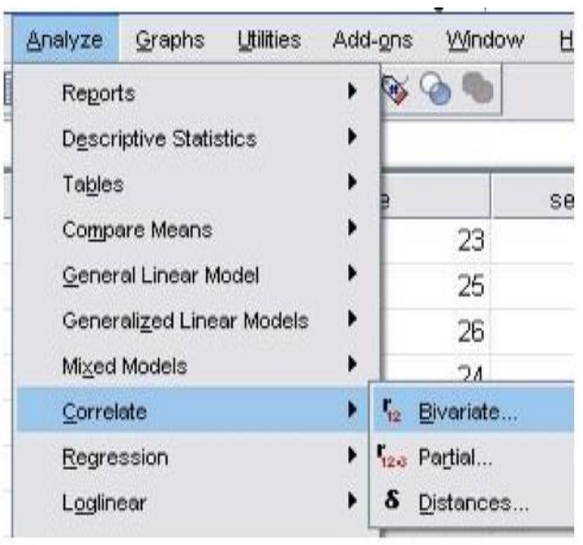
Example 1: Approximate readings of the volume of production (x) and the volume of exports (y) over several years are recorded as follows:

y_Square	x_Square	xy	Exports (y) in millions	Production (x) in millions
4	9	6	2	3
4	16	8	2	4
4	4	4	2	2
1	4	2	1	2
1	4	2	1	2
1	4	2	1	2
$\sum y^2 = 15$	$\sum x^2 = 41$	$\sum xy = 24$	$\sum y = 9$	$\sum x = 15$

Accordingly, Pearson correlation coefficient is calculated as follows:

$$r_{xy}^2 = \frac{6(24) - (15)9}{\sqrt{((6 \times 41) - 15^2)((6 \times 15) - 9^2)}} = 0.65$$

Since the correlation coefficient is 65.0, the relationship between volumes of production and exports is a moderate positive correlation. Statistical software may be used to calculate the correlation coefficient very easily, e.g. SPSS may be used, as shown on the screenshot below, to calculate the correlation coefficient between two variables.



Multiple correlation

A correlation coefficient that describes the relationship between a dependent variable and a number of independent variables. For example, this coefficient is used to identify the type of correlation between the volume of production of a dunum of wheat, the amount of rain and fertilizer, and the temperature. In this case, this coefficient measures correlation between the volume of production as a dependent variable and a set of other independent variables on which this variable depends.

The coefficient of multiple correlation is calculated using the following formula:

$$R^2_{1.23} = \sqrt{\frac{R^2_{12} + R^2_{13} - 2R_{12}R_{13}R_{23}}{1 - R^2_{23}}}$$

Where:

R^2_{12} is the square of the simple correlation coefficient of variables 1 and 2.

R^2_{13} is the square of the simple correlation coefficient of variables 1 and 3.

R_{12} is the simple correlation coefficient of variables 1 and 2.

R_{23} is the simple correlation coefficient of variables 2 and 3.

Example 2: A swimming coach wanted to know the relationship between the time of (100) meter freestyle swimming (dependent variable), and stretch (independent variable1) and cardiovascular and respiratory reflexes (independent variable 2), the simple correlation coefficient between the variables is:

R_{12} → performance time and stretch reflexex (0.82)

R^2_{13} → performance time and cardiovascular and respiratory reflexes (0.86)

R_{23} → stretch, cardiovascular and respiratory reflexes (0.80)

Using the formula above, the coefficient of multiple correlation $R^2_{1.23}$ = 0.88.

Partial correlation coefficient

A correlation coefficient that describes the relationship between two variables assuming that the third variable has a constant impact on both variables, $\rho_{12.3}$, used, for example, to find the strength or volume of the relationship between blood pressure variable and blood sugar levels, assuming that cholesterol levels are constant. The partial correlation coefficient of variable Y is calculated using the following formula:

$$\rho_{y2.1} = \frac{r^2_{y2} - r^2_{y1}r^2_{12}}{\sqrt{(1 - (R^2_{y1})^2)(1 - (R^2_{12})^2)}}$$

Where:

r^2_{y2} is the simple correlation coefficient of variables y and 2.

r^2_{12} is the simple correlation coefficient of variables 1 and 2.

r^2_{y1} is the simple correlation coefficient of variables Y and 1.

Example 3: An advertising agency wants to describe the relationship between the number of respondents to advertisements y, the size of the advertisement published in the newspaper x_1, and the number of distributed newspapers x_2. The agency has obtained the following data:

Number of respondents in hundreds (y_i), the size of the advertisement in inches (x_1) and number of distributed newspapers in thousands (x_2).

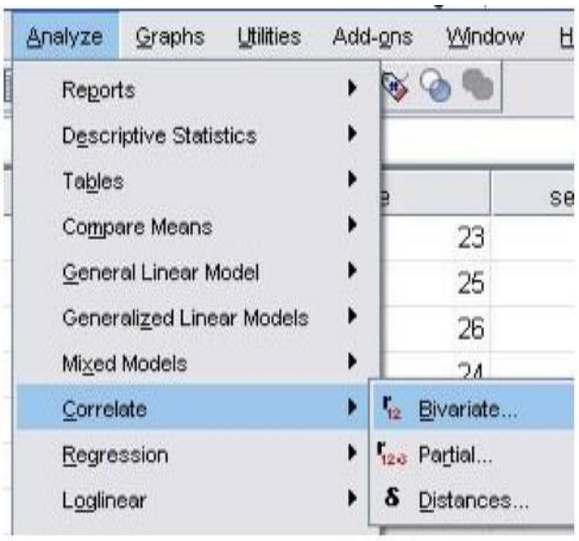
Accordingly, the following results have been obtained:

$$r^2_{12} = 0.741, \qquad r^2_{y2} = 0.931, \qquad r^2_{y1} = 0.909$$

using the formula above, the partial correlation coefficient $\rho_{y2.1}$ is calculated as follows:

$$\rho_{y2.1} = \frac{0.931 - 0.909 \, x \, 0.741}{\sqrt{(1 - 0.826)(1 - 0.549)}} = 0.92$$

There are also many statistical software packages used to calculate the partial correlation coefficient. In SPSS package, the partial correlation coefficient of variables is, after entering necessary data, calculated as follows:



4.1.1.2. Correlation coefficient for qualtitive phenomena (Spearman's rank coefficient of correlation)

There are some phenomena that can't be expressed in numbers, may be in the form of characteristics or ranks, including health of individuals and smoking, there is no quantitative measure of health or smoking habit. All that we can do is to classify health of individuals into graded types from bad to good or excellent, and so on. The same applies to the remaining similar phenomena such as ranks; converting numbers into ranks, divided into several types, the most common of which is Spearman's rank correlation coefficient. If we have two variables X and Y, each has n value, then Spearman's rank correlation coefficient is:

$$R = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

Where d_i is the difference between the value ranks of X and Y.
n is the number of value pairs of X and Y.

Example 4: The table below shows the grades of a group of students in a test done twice in a row on the same students. Calculate Spearman's rank correlation coefficient of the grades of both tests.

Grade of test 1 (X)	2	5	9	8	2
Grade of test 2 (Y)	4	6	7	4	3

Variable Y has two equal numbers (4, 4) and their order is (2, 3); each has the average rank (3 + 2)/ 2 = 2/5 = 2.5.

X	Y	**Ranks of X**	**Ranks of Y**	**d_i**	**di_Square**
3	4	2	2.5	-0.5	0.25
5	6	3	4	-1	1
9	7	5	5	0	0
8	4	4	2.5	1.5	2.25
2	3	1	1	0	0
Sum					**3.5**

Using the formula above, Spearman's rank correlation coefficient is:

$$\mathrm{R} = 1 - \frac{6(3.5)}{5(24)} = 0.825$$

4.2. Regression

Regression analysis is an analysis used to find a mathematical formula to correlate a dependent variable and one or more independent variable. For example, regression analysis is used to study factors that affect increased demand for the product and finding a mathematical relationship (formula) for this correlation, which not only enables us to understand the nature and determine the influencing factors of the relation, but also to predict the impact of changing any independent variable on the dependent variable.

This regression is variously and widely used. An engineer needs to study the factors affecting increased temperature of gases used in a process, may need to know the real effect of many factors. Using regression, an engineer can identify influencing and neglect non-influencing factors, and predict any change to the temperature of gases based on a specific change in any influencing variable. An HR manager needs to identify the factors affecting the performance of new employees including age, GPA, university, etc. Using regression analysis, An HR manager may identify influencing and non-influencing factors of the performance of new employees and have a mathematical relationship to predict and understand how much these factors influence performance.

As mentioned above, correlation coefficient describes the relationship among phenomena. However, we often need to understand the nature of relationships to study phenomena, as they may be in the form of a line or a curve. A regression curve is a graph representing the relationship between two

variables or a graph representing the relationship among phenomena, regression is used to estimate the value of the dependent variable, knowing the value of the other independent variables.

Graphical representation of variables

The first step to study the regression between two variables is graphical analysis, as visual inspection of data provides the following information:

- Covariance, an indicator of the degree of correlation between the two variables.
- The range and distribution of data sample points.
- Whether there are outliers.
- The type or shape of the relationship between the two variables.

Regression analysis is based on the relationship between two or more variables. Analysis here is based on having a dependent and an independent variable. Once the mathematical relationship between the two variables is identified, it is easy to identify the dependent variable based on the data of the independent variable.

If, for example, imports are affected by the national income, then by quantifying this relationship, imports may be predicted once the expected national income is determined. Mathematically, if the value of dependent variable Y depends on the amount of change to the value of independent variable X, then Y is expressed as a function of X, which is called regression. Regression coefficient is the indicator that shows the extent of change to a dependent variable based on a change to a unit of an independent variable.

4.2.1. Types of regression analysis

There are two types of regression analysis; linear regression - the most common - that means studying linear relationships, and non-linear regression that means studying relationships in the form of a curve. Linear regression is the most common and has two types; simple and multiple. Simple regression seeks to predict the relationship between a variable and a factor affecting the variable, while multiple regression seeks to predict the relationship between a variable and several factors affecting the variable. The first type, simple linear regression, is reviewed herein. To analyize using multiple linear or non-linear regression, please refer to relevant statistics books.

Simple linear regression

Simple linear regression analysis method is used to study and analyze the effect of a quantitative variable on another quantitative variable, for example:

- The effect of the amount of fertilizer on the production of a dunam.
- The effect of cost on production.
- The effect of the amount of protein eaten by cows on weight gain.
- The effect of income on consumer spending.

Thus, there are examples in many economic, agricultural, commercial, behavioral sciences, and other fields.

Linear regression model

In simple regression analysis, the researcher seeks to study the effect of the independent or predictor variable on the dependent or response variable, then the linear regression model may be presented as a linear equation, where the dependent variable is a function of the independent variable:

$$y_i = \beta_0 + \beta_1 x_i + e_i , \quad i = 1, 2, 3, \dots , n$$

Where:

x_i is the values of independent variable x.

β_1 is the estimated value of the regression coefficient.

β_0 is the intersection of the regression line with the vertical axis.

e is random error, the difference between true and estimated values of y.

Simple linear regression model estimation

Regression factors (β_1 and β_0) of the model can be estimated using the method of least squares, and this estimate minimizes the sum of the squares of random errors and is calculated using the following formula:

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right)}{\left(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Where $\bar{x}$ is the mean of x values, $\bar{y}$ is the mean of y values, and the estimated value of the dependent variable is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x ,$$

this estimate is called the regression equation Y on X.

Coefficient of determination R2

The real standard of the strength the regression relationship represents the analytical model is the coefficient of determination; it is the square of correlation coefficient, (r^2), is always positive and describes the strength of correlation between any two variablesa. For example, if correlation coefficient is (r=0.08), then there is a positive regression relationship and the strength of the relationship is (r^2=0.64), we can get the percentage of correlation using (r^2×100) to get the percentage of the strength of correlation.

Example 5: The table below shows the daily protein intake in gm and weight gain in kg for a sample of 10 individuals.

Protein intake	10	11	14	15	20	25	46	50	59	70
Weight gain	10	10	12	12	13	13	19	15	16	20

- Estimate the weight regression equation on protein intake.
- Explain the regression equation.
- Mention how much weight is gained when a person is given 50 gm of protein and random error.
- Draw the regression equation on the scatter plot.

Solution:

Estimation of the regression equation.
Assuming that x is protein intake and y is weight gain, the above two formulas can be used to calculate the following sums:

Protein intake x	Weight gain Y	xy	x_2
10	10	100	100
11	10	110	121
14	12	168	196
15	12	180	225
20	13	260	400
25	13	325	625
46	19	874	2116
50	15	750	2500
59	16	944	3481
70	20	1400	4900
320	140	5111	14664

Required sums
$\sum x = 320$
$\sum y = 140$
$\sum xy = 511$
$\sum x^2 = 14664$
Mean:
$\bar{x} = \frac{\sum x}{n} = \frac{320}{10} = 32$
$\bar{y} = \frac{\sum y}{n} = \frac{140}{10} = 14$

$\hat{\beta}_1$ is calculated using the first formula above:

$$\widehat{\beta_1} = \frac{6310}{44240} = 0.1426$$

$\hat{\beta}_0$ is calculated using the second formula above:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 14 - (0.1426)(32) = 9.4368$$

The estimated regression equation is:

$$\hat{y} = 9.44 + 0.143x$$

Explanation of the equation

- The constant: $\hat{\beta}_0 = 9.44$ shows that in case of no protein diet, weight will increase by 9.44 Kg.

- Regression coefficient ($\hat{\beta}_1 = 0.143$) shows that if protein intake is increased by 1gm, weight increases by 0.143 kg (143 gm).
- Weight gain observed by taking 50 gm of protein:

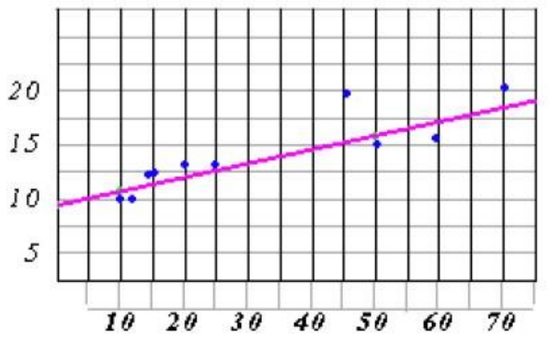
$$\hat{y} = 9.44 + 0.143(50) = 16.59$$

Random error is:

$$\hat{e}_{x=50} = y_{x=50} - \hat{y}_{x=50} = 15 - 16.59 = -1.59$$

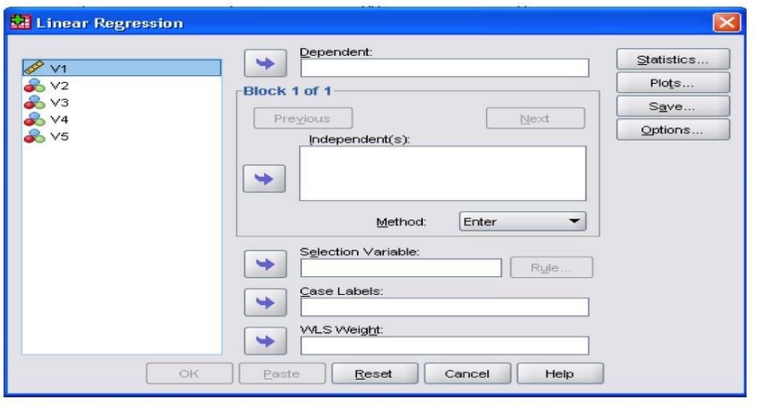
- Drawing the regression equation on the scatter plot.

The equation of a straight line can be drawn through two known points on the straight line.
Then, the regression equation line:



Coefficient of determination, in the example above, is (0.83 x 100 = 83%). Correlation can be explained that 83% of the two variables are affected by each other, or that 83% of the change in a dependent variable is due to the change in the other independent variable.

On the other hand, a statistical package, e.g. SPSS, may be used to make a linear regression analysis easily, as shown below:



References:

- Statistical Data Analysis, (www.pitt.edu/~super1/ResearchMethods/Arabic/statistical-ar.pdf).
- Principles of Statstics, https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnx0b3BkZXZvaXJzfGd4OjdiMTg4YWQ2ZTdhNjcxOA
- Methods of presenting and analysis of Statstical Data, (https://sites.google.com/site/drmohama/statand-control)
- Design and iImplementation of Statstical Surveys, SCAD.

مـركـز الإحـصـاء
STATISTICS CENTRE

الرؤية: بياناتنا نمضي نحو غدٍ أفضل
Vision: Driven by data for a better tomorrow

www.scad.gov.ae
adstatistics